Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών &
Μηχανικών Ηλεκτρονικών Υπολογιστών

Πανεπιστήμιο Πειραιά
Τμήμα Βιομηχανικής Διοίκησης&
Τεχνολογίας

ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΤΕΧΝΟ-ΟΙΚΟΝΟΜΙΚΑ ΣΥΣΤΗΜΑΤΑ»

Επισκόπηση Στοιχείων Μηχανικής, Αρχιτεκτονικών και Εργαλείων σε Συστήματα Μεγάλων Δεδομένων

Λυκοθανάσης Χρήστος
AM:03202947

Επιβλέποντες:

Δημήτριος Ασκούνης
Καθηγητής Ε.Μ.Π

Δρ.Κωνσταντίνος Τζαμαλούκας
Ειδικό Επιστημονικό Διδακτικό Προσωπικό Ε.Μ.Π

Αθήνα, Σεπτέμβριος 2020

Ευχαριστίες

Ευχαριστώ τους καθηγητές μου, κ. Κωνσταντίνο Τζαμαλούκα και κ.Δημήτριο Ασκούνη, για την απόδοχή και την εμπιστοσύνη τους στην ανάθεση αυτού του θέματος διπλωματικης που είναι πολύ κοντά στα ενδιαφεροντά μου. Η ανάπτυξη του με βοήθησε να αποκτήσω μια καλή βάση γνώσης για το επόμενο επαγγελματικό μου βήμα.

Περιεχόμενα

Κατάλογος Εικόνων & Διαγραμμάτων

Περίληψη

Η μεγάλη αυξηση της χρήσης των πληροφοριακών συστημάτων σε όλο και μεγαλύτερο εύρος δραστηριοτήτων και από συνεχώς αυξανόμενο πλήθος επιχειρήσεων καθώς και η διαρκής αλληλεπίδραση των επιχειρήσεων μεταξύ τους αλλά και με τους πελάτες τους μέσω βελτιωμένης διασύνδεσης, διαλειτουργικότας και ανοιχτότητας είναι ένα γεγονός ορατό παντού στις μέρες μας. Επίσης εύκολα κατανοητή μπορεί να γίνει και η επίπτωση που έχει επιφέρει η δραστηριότητα αυτή στον όγκο και την πολυπλοκότητα των δεδομένων που παράγονται , καταγράφονται και καταχωρούνται από τα IT συστήματα. Η εκμετάλλευση όλων αυτών των δεδομένων με σκοπό την εξαγωγή επιπλέον αξίας -πέραν δηλαδή της καθ'αυτού λειτουργικής χρήσης τους- έχει γίνει βασικό θέμα συζήτησης στις επιχειρήσεις και πλέον αρχίζει και καταλαμβάνει μεγάλο μέρος του μακροπρόθεσμου σχεδιασμού τους.

Τα λεγόμενα "Συστήματα Μεγάλων Δεδομένων" λοιπον αποτελούν ένα σύγχρονο πεδίο έρευνας και πρακτικής του ευρύτερου τομέα των επιχειρησιακών πληροφοριακών συστημάτων που έχει αναδυθεί μόλις τα τελευταία χρόνια ως αποτέλεσμα της προσπάθειας ικανοποίησης ακριβώς αυτής της ανάγκης: της συμπερίληψης των Big Data στην διαδικασία υποστήριξης επιχειρηματικών αποφάσεων. Η παρούσα εργασία στοχεύει -μέσω της βιβλιογραφικής ανασκόπησης- τον εντοπισμό των πιο αποδεκτών απόψεων και την σκιαγράφηση των βασικότερων συστατικών στοιχείων που συνθέτουν τα συστήματα από την σκοπιά της Τεχνολογίας της Πληροφορίας.Η ανάλυση εστιάζεται στους κύριους άξονες των θεωρητικών αρχών λειτουργίας τους, την σχετικής τεχνολογίας λογισμικού που έχει αναπτυχθεί για αυτά, τις πρότυπες θεωρητικές και εφαρμοσμένες αρχιτεκτονικές αναφοράς που τα υλοποιούν καθώς και τα διαθέσιμα έτοιμα εργαλεία λογισμικού.

Μελετώντας λοιπόν το πεδίο αυτό, καταρχάς θα κάνουμε μια ανασκόπηση στον ορισμό των Big Data , των βασικών χαρακτηριστικών τους, τα προβλήματα στα οποία μπορεί να δόσει λύσεις η υιοθέτηση εφαρμογών που διαχειρίζονται Big Data καθώς και πλεονεκτήματα που μπορούν να επιφέρουν στις επιχειρήσεις.

Έπειτα παρατίθενται οι κύριες προκλήσεις που συναντούν οι μηχανικοί λογισμικού κατά τον σχεδιασμό και την υλοποίηση συστημάτων Big Data, οι βασικές αρχές της τεχνολογίας λογισμικού που τα διέπουν καθώς και τα νέα πλαίσια που έχουν αναπτυχθεί για αυτά και αποτελούν πλέον την εφαρμοσμένη πρακτική του άμεσα συσχετιζόμενο υποκλάδου της μηχανικής που τα πραγματευέται,το Big Data Engineering.

Ακόμα, για τα παραπάνω στοιχεία θα κάνουμε μια ανασκόπηση στα πιο αποδεκτά και εφαρμοσμένα πρώτυπα αρχιτεκτονικών που οργανώνουν τα επιμέρους τμήματα και λειτουργικότητες σε μια ενιαία λύση επιχειρησιακής κλίμακας.

Τέλος, προχωρήσαμε και σε μια επικαιροποιημένη καταγραφή των ολοκληρωμένων λύσεων και εργαλείων που προσφέρονται τόσο ως εμπορικές εφαρμογές της αγοράς όσο και σαν ανοιχτού κώδικα έργα υποστηριζόμενα από αντίστοιχες κοινότητες χρηστών και συνθέτουν το IT οικοσύστημα των Big Data.

Μέσω της ανάλυσης αυτής ευελπιστούμε να αποκομίσουμε μια πολύ καλή σφαιρική γνώση των τρέχουσων τεχνολογιών που αφορούν το πεδίο των Μεγάλα Δεδομένων, το οποίο όπως όλα δείχνουν πιθανότατα θα αποτελέσει έναν από τους βασικότερους ρυθμιστές των τεχνολογικών και επιχειρηματικών εξελίξεων τα αμέσως επόμενα χρόνια.

Λέξεις Κλειδιά

Συστήματα Μεγάλων Δεδομένων, Μηχανική Μεγαλών Δεδομένων, Πλαίσια Αρχιτεκτονικής Μεγάλων Δεδομένων,Λύσεις Λογισμικού Μεγάλων Δεδομένων

Abstract

The large increase in the use of Information Systems in a growing range of activities and by a growing number of businesses as well as the constant interaction of businesses with each other and with their customers through improved interconnection, interoperability and openness is a common phenomenon nowadays. It is also easy to understand the impact that this activity has on the volume and complexity of the data produced, recorded and saved by IT systems. The exploitation of all this data in order to extract additional value – above their own functional use - has become a major topic of discussion in businesses and now begins occuping much of their long-term planning.

The so-called "**Big Data Systems**" are therefore a modern field of research and practice in the broader field of Business Information Systems that has emerged only in recent years as a result of trying to meet exactly this need: the inclusion of Big Data in businesses decision support process. This paper aims - through bibliographic review - to identify the most acceptable views and to outline the key components that make up these systems from the Information Technology point of view. Analysis focuses on the main axes of **theoretical principles**, the special **software engineering** topics developed for them, the standard theoretical and applied **reference architectures** that implement them as well as the available **software tools**.

So by studying this field, we will first look at the definition of Big Data, their key features, the problems that can be solved by adopting applications that manage Big Data as well as the benefits they can bring to businesses.

The following are the main challenges encountered by software engineers in designing and implementing Big Data systems, the basic principles of software technology that govern them as well as the new frameworks that have been developed for them and are now the applied practice of the directly related Engineering subsidiary that deals with them: the **Big Data Engineering**.

Also, we will review the most accepted and applied standards of architectures that organizes these individual components and functionalities above into a single solution of operational scale.

Finally, we proceeded with an updated recording of integrated solutions and tools that are offered both as commercial applications of the market and as open source projects supported by respective user communities and compose the IT ecosystem of Big Data.

Through this analysis we hope to gain a very good comprehensive knowledge of current technologies in the field of Big Data, which as everything shows will probably be one of the main regulators of technological and business development in the coming years.

Key-Words

Big Data Systems, Big Data Engineering, Big Data Architecture Frameworks,Big Data Software Solutions

1.Εισαγωγικές Έννοιες

1.1 Γενικά

Ως Big Data ορίζονται τα εκτεταμένα σύνολα δεδομένων των οποίων τα κυριάρχα χαρακτηριστικά του όγκου της ποικιλίας, της ταχύτητας μετάδοσης και η ευμεταβλητότητας καθιστούν απαραίτητη την εφαρμογή ευρείας κλίμακας αρχιτεκτονικών για την αποτελεσματική αποθήκευση, χειρισμό και ανάλυση τους.

Η σύνοψη των γενικά χαρακτηριστικών και τα ειδικών συστατικών του οικοσυστήματος των Big Data παρουσιάζεται στον παρακάτω ορισμο του **Gartner**:

"Οι τεχνολογίες Big Data (Data Intensive Tecnologies) στοχεύουν στη διεκπεραίωση υψηλού όγκου, υψηλής ταχύτητας, μεγάλης ποικιλίας δεδομένων (ως σύνολα ή διαθέσιμα πάγια) για την εξαγωγή θεμιτής αξίας με τρόπο που θα εξασφαλίζει παράλληλα υψηλή αυθεντικότητα μεταξύ των αρχικών δεδομένων και των εξαγώμενων πληροφοριών. Για το σκοπό αυτό απαιτούνται οικονομικά αποδοτικές και καινοτόμες μορφές επεξεργασίας δεδομένων (analytics) που θα στοχεύουν στην βελτίωση της εξόρυξη γνώσης,τη λήψη αποφάσεων και τον έλεγχο των διαδικασιών. Όλα αυτά απαιτούν με την σειρά τους νέα μοντέλα δεδομένων (που θα υποστηρίζουν όλες τις καταστάσεις και στάδια των δεδομένων κατά τη διάρκεια ολόκληρου του κύκλου ζωής τους) και νέες υπηρεσίες υποδομής και εργαλεία που θα επιτρέπουν την απόκτηση (και επεξεργασία) δεδομένων από διάφορες πηγές (συμπεριλαμβανομένων των δικτύων αισθητήρων) και την παροχή δεδομένων σε διάφορες μορφές αλλά και διάφορους καταναλωτές ή συσκευές αποτύπωσης πληροφοριών".

Όπως και πολλοί άλλοι όροι που έχουν έρθει σε ευρεία χρήση στην τρέχουσα εποχή της πληροφορίας, τα “Big Data” έχουν πολλές εναλλακτικές σημασίες ανάλογα με το πλαίσιο στο οποίο χρησιμοποιούνται. Οι συζητήσεις δε των Big Data είναι πολύπλοκες λόγω της έλλειψης αποδεκτών ορισμών, ταξινομήσης και κοινών απόψεων αναφοράς.

Οπότε, ο όρος κατα καιρούς έχει χρησιμοποιηθεί για να περιγράψει μονομερώς έναν αριθμό θεματικών πεδίων, συμπεριλαμβανομένων των εξής:

• Χαρακτηριστικά δεδομένων
• Περισσότερα δεδομένα σε σχέση τα προηγούμενα χρονια
• Μη δομημένα δεδομένα
• Οι νέες κλιμακούμενες τεχνικές επεξεργασίας δεδομένων (scalable process)
• Η ανάλυση εκτεταμένων συνόλων δεδομένων (massive datasets analysis)
• Τη δημιουργία αξίας
• Η απώλεια ιδιωτικότητας
• Ο αντίκτυπος στην κοινωνία

1.2 Εννοιολογικές Προσεγγίσεις και Ερμηνείες του όρου “Big Data”

[4.NIST]

Όγκος

"Παρόλο που τα μεγάλα δεδομένα δεν αναφέρονται σε κάποια συγκεκριμένη ποσότητα, ο όρος χρησιμοποιείται συχνά όταν μιλάμε για επίπεδα petabytes και exabytes δεδομένων." **Techtarget**

"Ο όρος Big Data αφορά την υπολογιστική επί δεδομένων πολύ μεγάλου μεγέθους, συνήθως στο βαθμό που η επέκταση και διαχείρισή τους δημιουργεί σημαντικές διοικητικές προκλήσεις. Επίσης αναφέρεται στον κλάδο της υπολογιστικής επιστήμης που περιλαμβάνει τέτοια δεδομένα.” **Αγγλικό Λεξικό της Οξφόρδης**

"Μεγαλύτερα" Δεδομένα

"Τα Big Data είναι δεδομένα που περιέχουν παρατηρήσεις ικανές να απαιτήσουν μη συνηθισμένο χειρισμό λόγω του τεράστιου μεγέθους τους και επιπλέον παρουσιάζουν ασυνήθιστες αλλαγές στην πάροδο του χρόνου οι οποιές ποικίλλουν από τη μια περιοχή στην άλλη" **Annette Greiner**

Όχι μόνο Όγκος

"Αυτό που είναι μεγάλο στα Big Data δεν είναι κατ 'ανάγκη το μέγεθος των βάσεων δεδομένων, είναι ο μεγάλος αριθμός πηγών δεδομένων που έχουμε, καθώς οι ψηφιακοί αισθητήρες και οι δείκτες συμπεριφοράς που διαχεόνται σε όλο τον κόσμο". **Quentin Hardy**

"*... ο αρχικός μας ορισμός ήταν ένα σύστημα που ήταν σε θέση να αποθηκεύσει 10 TB δεδομένων ή περισσότερα ... Με τον καιρό, η ποικιλομορφία των δεδομένων άρχισε να γίνεται πιο διαδεδομένη στα συστήματα αυτά (ιδιαίτερα η ανάγκη να αναμειχθούν δομημένα και μη δομημένα δεδομένα), γεγονός που οδήγησε σε πιο εκτεταμένη υιοθέτηση των "3 Vs" (όγκος, ταχύτητα και ποικιλία) ως ορισμό των Big Data* "

Chris Neumann

Big Data Engineering

"Οι τεχνολογίες Big Data περιγράφουν μια νέα γενιά τεχνολογιών και αρχιτεκτονικών, που αποσκοπούν στην αποτελεσματική εξαγωγή αξίας από πολύ μεγάλους όγκους δεδομένων που ποικιλουν ευρύτατα, επιτρέποντας την υψηλών ταχυτήτων εξαγωγη , διερεύνηση και ανάλυση" **IDC**

"Τα Big Data σημαίνουν δεδομένα που δεν μπορούν να προσαρμοστούν εύκολα σε μια τυπική σχεσιακή βάση δεδομένων" **Hal Varian**

"Τα Big Data αναφέρονται σε σύνολα δεδομένων, το μέγεθος των οποίων υπερβαίνει την ικανότητα των τυπικών εργαλείων λογισμικού βάσης δεδομένων να τα συλλέγουν, να τα αποθηκεύουν, να τα διαχειρίζονται και να τα αναλύουν" **McKinsey**

Λιγότερες Δειγματοληψίες

"Τα Big Data είναι όταν η επιχείρησή σας επιθυμεί να χρησιμοποιήσει τα δεδομένα για να λύσει ένα πρόβλημα, να απαντήσει σε μια ερώτηση, να παράγει ένα προϊόν κλπ., Δημιουργώντας δηλαδή μια λύση στο πρόβλημα που αξιοποιεί πλήρως τα δεδομένα όχι με απλή δειγματοληψία ή ανασκόπηση αρχείων" **John Foreman**

"Τα Big Data περιγράφονται αρχικά ως μια πρακτική στον κλάδο του Διαδικτύου για την εφαρμογή αλγορίθμων σε όλο και μεγαλύτερες ποσότητες ανόμοιων δεδομένων με σκοπό την βέλτιστη επίλυση προβλημάτων που προηγούμενως είχαν μη-βέλτιστες λύσεις βασιζόμενες σε προσομοιώσεις επί μικρότερων σύνολων δεδομένων" **Peter Skomoroch**

Νέοι Τύποι Δεδομένων

"Το ευρύ φάσμα των νέων και μαζικών τύπων δεδομένων που εμφανίστηκαν την τελευταία δεκαετία περίπου"

Tom Davenport

"Τα Big Data δεν αφορούν μόνο τον όγκο, αλλά είναι περισσότερο ο συνδυασμός διαφορετικών συνόλων δεδομένων και η ανάλυσή τους σε πραγματικό χρόνο με σκοπό την εξόρυξη πληροφορίας για τον οργανισμό σας. Επομένως, ο σωστός ορισμός των μεγάλων δεδομένων πρέπει στην πραγματικότητα να είναι: μικτά δεδομένα **Mark van Rijmenam**

Analytics

"Τα Bid Data σήμαιναν δεδομένα που μια μόνο μηχανή δεν μπορεί να χειριστεί. Τώρα τα Bid Data έχουν γίνει μια λέξη-κλειδί που σημαίνει οτιδήποτε σχετίζεται με την ανάλυση δεδομένων ή την απεικόνιση. "

Ryan Swanstrom

Επιστήμη των Δεδομένων

"Τα Big Data περιγράφουν σύνολα δεδομένων που είναι τόσο μεγάλα, περίπλοκα ή ταχέως μεταβαλλόμενα ώστε να αγγίζουν τα όρια της ίδιας της αναλυτικής μας ικανότητας" **Joel Gurin**

"Τα Big Data είναι ο ευρύς όρος που δίνεται στις προκλήσεις και τις ευκαιρίες που έχουμε, καθώς γίνονται διαθέσιμα δεδομένα για κάθε πτυχή της ζωής μας. Ωστόσο δεν πρόκειται μόνο για δεδομένα. Περιλαμβάνει επίσης τους ανθρώπους, τις διαδικασίες και την ανάλυση που μετατρέπει τα δεδομένα σε απτό νόημα"
Josh Ferguson

Αξία
"Για μένα, τα" Big Data "είναι η κατάσταση όπου ένας οργανισμός μπορεί (αναμφισβήτητα) να πει ότι έχει πρόσβαση σε αυτά που χρειάζεται για να ανακατασκευάσει, να κατανοήσει και να μοντελοποιήσει το μέρος του κόσμου που τον ενδιαφέρει». **Harlan Harris**

"Big Data αφορούν τη χρήση σύνθετων συνόλων δεδομένων που αποσκοπούν στο χειρισμό του προσανατολισμού, της κατεύθυνσης και τη λήψης αποφάσεων σε μια εταιρεία ή οργανισμό."
Jessica Kirkpatrick

"Τα μεγάλα δεδομένα είναι απλώς η δυνατότητα συλλογής πληροφοριών και στην συνέχεια η προσπέλαση τους με τέτοιο τρόπο ώστε να μπορούμε να μάθουμε πράγματα για τον κόσμο που προηγουμένως δεν μας ήταν προσβάσιμα". **Hilary Mason**

"Ο καλύτερος ορισμός που έχω δει είναι: Τα δεδομένα είναι μεγάλα όταν το μέγεθος των δεδομένων γίνεται μέρος του προβλήματος. Ωστόσο, αυτό αναφέρεται μόνο στο μέγεθος. Πλέον, η λέξη "Big Data" αναφέρεται στο νέο μοντέλο των επιχειρήσεων που καθοδηγούνται από τα δεδομένα, της επιστήμης και της τεχνολογίας, όπου το τεράστιο μέγεθος και εύρος δεδομένων επιτρέπουν καλύτερες και καινοτόμες υπηρεσίες, προϊόντα και πλατφόρμες" **Gregory Piatetsky-Shapiro**

Αλλαγή Εταιρικής Κουλτούρας
"Τα Big Data , τα οποία άρχισαν ως τεχνολογική καινοτομία της κατανεμημένης υπολογιστικής, είναι πλέον ένα πολιτιστικό κίνημα με το οποίο συνεχίζουμε να ανακαλύπτουμε πώς αλληλεπιδρά η ανθρωπότητα με τον κόσμο - και ο ένας με τον άλλο - σε μεγάλη κλίμακα". **Drew Conway**

"Τα Big Data αντιπροσωπεύουν μια μεταβολή κουλτουρας κατά την οποία όλο και περισσότερες αποφάσεις γίνονται με αλγόριθμους διαφανούς λογικής που λειτουργούν με τεκμηριωμένα αμετάβλητα στοιχεία. Νομίζω ότι το «μεγάλο» αναφέρεται περισσότερο στη διάχυτη φύση αυτής της αλλαγής παρά σε κάποιο συγκεκριμένο όγκο δεδομένων» **Daniel Gillick**

Ο όρος μεγάλα δεδομένα είναι περισσότερα από ένα πράγμα, αλλά μια σημαντική πτυχή είναι η χρήση τους ως ρητορικό εργαλείο, κάτι δηλαδη που μπορεί να χρησιμοποιηθεί για να εξαπατήσει ή να παραπλανήσει ή να υπεκφυγει». **Cathy O'Neil**

1.3 Αποδεκτοί Ορισμοί και Θεωρήσεις

1.3.1 Τα Καθολικά Χαρακτηριστικά των Big Data - 4Vs

[4.NIST]

Ο καθορισμός καθολικών κριτηρίων προσδιορισμού των χαρακτηριστικών των δεδομένων που απαιτούν λύσεις ευρείας κλίμακας είναι δύσκολη, καθώς η επιλογή συγκεκριμένης αρχιτεκτονικής μεταξύ διαφόρων εναλλακτικών βασίζεται πάντοτε σε μια προσπάθεια εξισορρόπισης μεταξύ των επιδόσεων, του κόστους και των χρονικών περιορισμών που ανακύπτουν κατά την πλήρη λειτουργία του. Γενικά όμως είναι ευρέως αποδεκτό πως ο χαρακτηρισμός ενός προβλήματος ως προβλήματος μεγάλων δεδομένων εξαρτάται από την ανάλυση των απαιτήσεων της εφαρμογής και οι βασικοί οδηγοί ως προς αυτό είναι ο όρος που συναντάται στην βιβλιογραφία ώς “αξίωμα των 4Vs” , τα οποία και αναλύονται παρακάτω :

Volume | Όγκος

Το πιο αναγνωρισμένο χαρακτηριστικό των Big Data είναι η ύπαρξη εκτεταμένων συνόλων δεδομένων(**massive datasets**) που αντιπροσωπεύουν την υπερπληθώρα δεδομένων που είναι διαθέσιμα για ανάλυση και από την αποία δύναται να εξαχθούν πολύτιμες πληροφορίες. Ωστόσο σε αυτό το σημείο υποβόσκει πάντοτε η παραδοχή ότι όσο περισσότερα είναι τα προς επεξεργασία δεδομένα τόσο μεγαλύτερη και η αξία που δύναται να προκύψει απο την ανάλυση τους. Υπάρχουν πολλά παραδείγματα επί αυτού, γνωστά και ως φαινόμενο δικτύου (**network effect**), όπου τα μοντέλα δεδομένων βελτιώνονται με μεγαλύτερους όγκους δεδομένων. Επίσης , μεγάλο μέρος της εξέλιξης στο **machine learning** προκύπτει από τις τεχνικές του εκείνες που επεξεργάζονται δεδομένα. Για παράδειγμα, η αναγνώριση αντικειμένων σε εικόνες βελτιώθηκε σημαντικά όταν το πλήθος των εικόνων που έγιναν διαθέσιμα για ανάλυση αυξήθηκαν από χιλιάδες σε εκατομμύρια μέσω της χρήσης τεχνικών κλιμάκωσης. Ο χρόνος και το κόστος που απαιτούνται για τη επεξεργασία massive datasets ήταν ένας από τους παράγοντες που οδήγησαν τελικά στην κατανεμημένη επεξεργασία. Ο όγκος οδηγεί στον παραλληλισμό κατά την επεξεργασία και την αποθήκευση καθώς και στη συνεχή διαχείριση αυτού.

Velocity | Ταχύτητα

Αποτελεί ένα μέτρο του ρυθμού ροής δεδομένων. Παραδοσιακά, τα συστήματα υψηλών ταχύτητων έχουν περιγραφεί ως **streaming data** . Παρόλο που αυτές οι έννοιες είναι νέες για ορισμένες βιομηχανίες-κλάδους , σε άλλες (οπως π.χ. τις τηλεπικοινωνίες και τις συναλλαγές μέσω πιστωτικών καρτών) εδώ και χρόνια εφαρμόζεται η επεξεργασία μεγάλων όγκων δεδομένων που παράγονται εντός πολύ σύντομων χρονικών διαστημάτων. Τα δεδομένα υποβάλλονται σε επεξεργασία και αναλύονται σε πραγματικό χρόνο (ή σχεδόν σε πραγματικό χρόνο) καθόσο κινούνται (δηλ.μεταβάλλονται διαρκώς) και πρέπει να αντιμετωπίζονται με πολύ διαφορετικό τρόπο από ό, τι τα μη μεταβαλλόμενα δεδομένα (**data in rest** δηλαδή αποθηκευμένα δεδομένα). Τα δεδομένα σε κίνηση (**data in motion**) τείνουν να μοιάζουν με αρχιτεκτονικές επεξεργασίας γεγονότων(**event processing**) και επικεντρώνονται σε εφαρμογές υποστήριξης λειτουργιών και πληροφόρησης πραγματικού χρόνου (**real-time or operational applications**) . Η ανάγκη επεξεργασίας δεδομένων σε πραγματικό χρόνο με ταυτόχρονη παρουσία μεγάλου όγκου δεδομένων οδηγεί σε διαφορετικούς τύπους αρχιτεκτονικής, όπου τα δεδομένα δεν αποθηκεύονται, αλλά συνήθως επεξεργάζονται στη μνήμη. Βέβαια, οι χρονικοί περιορισμοί για την επεξεργασία σε πραγματικό χρόνο μπορούν να δημιουργήσουν την ανάγκη για κατανεμημένη επεξεργασία ακόμα και όταν τα σύνολα δεδομένων είναι σχετικά μικρά – σενάριο συνηθες πχ στο Διαδίκτυο των Πραγμάτων (IoT).

Variety | Ποικιλία

Το χαρακτηριστικό της ποικιλίας αντιπροσωπεύει την ανάγκη να αναλύονται δεδομένα με διαφορετικές προελεύσεις, αντικείμενο ή χαρακτηριστικά. Η ποικιλία των δεδομένων αντιμετωπίστηκε τα προηγούμενα χρόνια απο τις βιομηχανίες κυρίως μέσω της προσπάθειας εντοπισμού γνωρισμάτων που θα επέτρεπαν την ευθυγράμμιση όλων των δεδομένων και τη σύντηξη τους σε μια κεντρική αποθήκη δεδομένων (**Enterprise Data Warehouse**). Η αυτοματοποιημένη σύντηξη των δεδομένων δηλαδή βασιζόταν σε σημασιολογικά μεταδεδομένα, όπου η κατανόηση των δεδομένων μέσω των μεταδεδομένων επέτρέπε την διασύνδεση τους. Σήμερα όμως το μεγάλο εύρος τύπων δεδομένων, προελεύσεων, λογικών μοντέλων, χρονοδιαγραμμάτων και σημασιολογίας περιπλέκει την διαδικασία ανάπτυξης αναλύσεων που θα μπορούν να καλύψουν την τόσο μεγάλη ποικιλία δεδομένων. Η χρήση τεχνικών κατανεμημένης επεξεργασίας όμως επιτρέπει την δημιουργία διακριτών προ-αναλυτικών στοιχείων για τους διαφορετικούς τύπους δεδομένων, τα οποία και ακολουθούνται από διαφορετικά στοιχεία συγκεντρωτικής ανάλυσης που θα συγκεκράσουν αυτά τα ενδιάμεσα αποτελέσματα. Παρόλο που ο όγκος και η ταχύτητα επιτρέπουν βελτιστοποίηση τους κόστους και του χρόνου παραγωγής analytics, η ποικιλία των δεδομένων είνα αυτή που επιτρέπει να εξαχθούν αποτελέσματα που δεν ήταν δυνατά ποτέ αλλοτε. Κοινή πεπόιθηση σχετικά με την ποικιλία των δεδομένων είναι: "Τα επιχειρηματικά οφέλη είναι συχνά υψηλότερα όταν αντιμετωπίζουμε την ποικιλία των δεδομένων από ό, τι όταν αντιμετωπίζουμε τον όγκο"

Variability | Μεταβλητότητα

Η μεταβλητότητα είναι ένα ελαφρώς διαφορετικό χαρακτηριστικό από τον όγκο, την ταχύτητα και την ποικιλία, δεδομένου ότι αναφέρεται σε μια αλλαγή σε ένα σύνολο δεδομένων αντί για το ίδιο σύνολο δεδομένων ή τη ροή του. Η μεταβλητότητα αφορά σε αλλαγές στην ταχύτητα ροής δεδομένων, τη μορφή / δομή και / ή τον όγκο, σε βαθμό που να επηρεάζεται η επεξεργασία του συνόλου δεδομένων. Οι επιπτώσεις μπορεί να περιλαμβάνουν την ανάγκη για επαναπροσδιορισμό των αρχιτεκτονικών, διεπαφών, των αλγορίθμων επεξεργασίας, τη διασύνδεση ή την αποθήκευση. Η μεταβλητότητα σε όγκο δεδομένων υποδηλώνει την ανάγκη για μεγέθυνση ή μείωση των εικονικών πόρων για την αποτελεσματικότερη διαχείριση του επιπλέον φορτίου επεξεργασίας, δυνατότητα που αποτελει πλεονέκτημα του cloud computing. Συγκεκριμένες πρακτικές ανάλυσης που χρησιμοποιούνται για την επεξεργασία δεδομένων μπορούν να βρεθούν σε δημοσιεύσεις του κλάδου που επικεντρώνεται σε operatinal

clouds ή virtualized αρχιτεκτονικές. Η δυνατότητα κατ'επιλογήν κλιμάκωσης διατηρεί τα συστήματα αποτελεσματικά, αντί να χρειάζεται να προσχεδιάζουν και να αξιοποιούν την εκτιμόμενη μέγιστη χωρητικότητα (όπου στις περισσότερες φορές το σύστημα παραμένει αδρανές). Πρέπει να σημειωθεί ότι αυτή η μεταβλητότητα αφορά μεταβολές στα χαρακτηριστικά όλου του συνόλου δεδομένων και όχι τις μεταβαλλόμενες τιμές των επιμέρους μεμονωμένων στοιχείων ή υποσυνόλων του.

1.3.2 Ορισμός 5 Μερών

[20. Demchenko, Membre]

1 - Ιδιότητες (5vs)

Επιπλεόν των 4Vs:Δυναμικότητα Δεδομένων και Καταγραφή του Κύκλου Ζωής της Εγγραφής

2 - Νέα Μοντέλα Δεδομένων

- Διασύνδεση Δεδομένων, Προέλευση και Ακεραιότητα Παραπομπών (referral integrity)
- Κύκλος ζωής των Δεδομένων και Διαχρονική Εξέλιξη

3 - Νέα Analytics

- Αναλύσεις σε πραγματικό χρόνο / streaming,
- Διαδραστικές Αναλύσεις (Interactive Analytics)
- Μηχανική Μάθηση

4 - Νέα Υποδομή και Εργαλεία

- Τεχνολογίες Υπολογιστικού Νέφους στην αποθήκευση (Cloud Storage), Δικτύωση, Υπολογιστική Υψηλών Επιδόσεων (High Performance Computing)
- Ενοποίηση Ετερογενών Υπηρεσιών από πολλαπλούς παρόχους
- Νέες Μοντέλα Υπηρεσιών προσανατολισμένα στα δεδομένα (απο πολλαπλούς φορείς)
- Νέα Μοντέλα Ασφαλείας προσανατολισμένα στα δεδομένα για αξιόπιστη υποδομές, επεξεργασία και αποθήκευση

5 - Πηγή και Στόχος

Αποτελούν σημαντικές πτυχες που ορισμένες φορές ορίζουν τύπους και δομές δεδομένων, π.χ. πρωτογενή δεδομένα, ροές δεδομένων, συσχετισμένα δεδομένα

- Καταγραφή δεδομένων υψηλής ταχύτητας από ποικιλία αισθητήρων και πηγών δεδομένων
- Παράδοση δεδομένων σε διαφορετικά συστήματα οπτικής απεικόνισης και αλληλεπίδρασης
- Ψηφιοποιημένη είσοδος και εξόδος, πανταχού παρόντες διασυνδεδεμένοι αισθητήρες και πλήρης ψηφιακός έλεγχος

1.3.3 Θεώρημα HACE

[3. Wu, Zhu, Ding]

Τα Big Data ξεκινούν με μεγάλου όγκου, ετερογενείς (**H**eterogenus), αυτόνομες πηγές (**A**utonomus sources) υπό κατανεμημένο και αποκεντρωμένο έλεγχο και επιδιώκουν να διερευνήσουν σύνθετες (**C**omplex) και εξελισσόμενες (**E**volving) εξαρτήσεις μεταξύ των δεδομένων.

Αυτά τα χαρακτηριστικά δημιουργούν μια εξαιρετική πρόκληση για ανακάλυψη χρήσιμης γνώσης από τα δεδομένα. Σαν ένα απλοικό παράδειγμα, μπορούμε να φανταστούμε ότι ένας αριθμός τυφλών προσπαθεί να εκτιμήσεί/μαντέψει ένα γιγαντιαίο ελέφαντα, ο οποίος θεωρούμε ότι είναι τα Big Data. Ο στόχος κάθε τυφλού είναι να σχεδιάσει μια εικόνα (ή συμπέρασμα) του ελέφαντα σύμφωνα με το μέρος των πληροφοριών που συλλέγει κατά τη διάρκεια της διαδικασίας συλλογής δεδομένων. Επειδή η άποψη κάθε ατόμου περιορίζεται στην τοπική του περιοχή, δεν αποτελεί έκπληξη το γεγονός ότι ανεξάρτητα ο κάθε τυφλός καταλήγει ότι ο ελέφαντας μοιάζει παλαμάρι,κολώνα ή τοίχο, ανάλογα με την περιοχή στην οποία είχε περιοριστεί. Για να γίνει το πρόβλημα περισσότερο περίπλοκο ας υποθέσουμε ότι 1) ο ελέφαντας αυξάνεται γρήγορα και η στάση του σώματος του μεταβάλλεται συνεχώς, και 2) κάθε τυφλός μπορεί να έχει τις δικές του (πιθανόν αναξιόπιστες και ανακριβείς) πηγές πληροφοριών που προκατειλημμένα τον ενημερώνουν σχετικά με τον ελέφαντα (π.χ. ένας τυφλός μπορεί να συζητήσει την αίσθηση του για τον ελέφαντα με τον άλλο τυφλό, όπου προφανώς η ανταλλαγή γνώσεων μεταξύ τους έχει εγγενής μεροληψία). Η εξερεύνηση Big Data σε αυτό το σενάριο είναι ισοδύναμη με τη συγκέντρωση ετερογενών πληροφοριών από διαφορετικές πηγές (τυφλοί) για να βοηθήσουν στην σκιαγράφηση της εικόνας ώστε πλησιάσει κατά το δυνατόν περισσότερο την αυθεντική μορφή του ελέφαντα και όλα αυτά μέσα από μια διαδικασία πραγματικού χρόνου. Πράγματι, αυτό το έργο δεν είναι τόσο απλό όσο να ζητάς από κάθε τυφλό να περιγράψει τα συναισθήματά του σχετικά με το ελέφαντα και στη συνέχεια ένας ειδικός να σχεδιάζει ένα ενιαίο σχέδιο από τον συνδυασμό των απόψεων, καθώς είναι πιθανό κάθε άτομο να μιλάει σε διαφορετική γλώσσα (ετερογενείς και ποικίλες πηγές πληροφόρησης) και μπορεί ακόμη και να έχουν ενδοιασμούς σχετικά με την προστασία της ιδιωτικής ζωής μέσα στα μηνύματα της διαδικασίας ανταλλαγής πληροφοριών.

Τεράστιοι Όγκοι Ετερογενών Δεδομένων Ποικίλων Διαστάσεων

Ένα από τα βασικά χαρακτηριστικά των Big Data είναι ο τεράστιος όγκος δεδομένων που αντιπροσωπεύουν ετερογενείς και ποικίλες διαστάσεις. Αυτό συμβαίνει επειδή διαφορετικοί συλλέκτες πληροφοριών προτιμούν τα δικά τους σχήματα ή πρωτόκολλα για την καταγραφή δεδομένων καθώς και τη φύση των διαφόρων εφαρμογών που επίσης οδηγεί σε διάφορετικές αναπαραστάσεις δεδομένων. Για παράδειγμα, ο κάθε άνθρωπος από βιολογικής-βιοιατρικής άποψης μπορεί να αναγνωριστεί χρησιμοποιώντας απλά δημογραφικά χαρακτηριστικά ,όπως το φύλο, την ηλικία, την ιστορία των οικογενειακών ασθενειών και ούτω καθεξής. Οι ακτινονογραφίες και αξονική τομογραφία για κάθε άτομο, σε μορφή εικόνας ή βίντεο χρησιμοποιούνται για να αναπαραστήσουν τα αποτελέσματα, καθώς παρέχουν οπτικές πληροφορίες στους ιατρούς βοηθώντας τους να εκτελέσουν πιο λεπτομερείς εξετάσεις. Για μια εξέταση DNA ή γονιδιώματος, χρησιμοποιούνται εικόνες και αλληλουχίες έκφρασης μικροσυστοιχιών για να αναπαραστήσουν τις πληροφορίες του γενετικού κώδικα επειδή αυτός είναι τρόπος που οι υπάρχουσες έως σήμερα τεχνικές ανακτούν τα δεδομένα αυτά. Υπό αυτές τις συνθήκες, τα ετερογενή χαρακτηριστικά αναφέρονται σε διαφορετικούς τύπους αναπαραστάσεων βιολογικών γνωρισμάτων για τα ίδια άτομα, και τα ποικίλα χαρακτηριστικά αναφέρονται στην πληθώρα διαθέσιμων χαρακτηριστικών που συμμετέχουν για να αντιπροσωπευθεί η κάθε παρατήρηση. Σε περιτώσεις δε που διαφορετικές οργανώσεις (ή επαγγελματίες υγείας) πιθανόν να έχουν τα δικά τους σχήματα για να πειργράψουν τον κάθε ασθενή, η ετερογένεια των δεδομένων και η ποικιλία διαστάσεων γίνονται μείζονες προκλήσεις εάν προσπαθήσουμε να αναλύσουμε συγκέντρωτικά συνδυάζοντας δεδομένα από όλες τις πηγές.

Αυτόνομες Πηγές με Κατανεμημένο και Αποκεντρωμένο Έλεγχο

Οι αυτόνομες πηγές δεδομένων με κατανεμημένους και αποκεντρωμένους ελέγχους αποτελούν βασικό χαρακτηριστικό του Big Data εφαρμογών. Όντας αυτόνομη, κάθε πηγή δεδομένων είναι σε θέση να δημιουργήσει και να συλλέξει πληροφορίες χωρίς να επηρεάζεται (ή βασίζεται) σε οποιονδήποτε κεντρικό έλεγχο. Αυτό είναι παρόμοιο με την λειτουργία του World Wide Web (WWW) όπου κάθε διακομιστής ιστού παρέχει ένα ορισμένο αριθμό πληροφοριών και είναι ικανος να λειτουργήσει πλήρως χωρίς απαραίτητα να βασίζεται σε άλλους διακομιστές. Από την άλλη πλευρά, οι τεράστιοι όγκοι των δεδομένων καθιστούν την εφαρμογή ευάλωτη σε επιθέσεις ή δυσλειτουργίες, εάν το σύνολο του συστήματος πρέπει να βασιστεί σε οποιοδήποτε κεντρική μονάδα ελέγχου. Για τις μεγάλες εφαρμογές που σχετίζονται με τα μεγάλα δεδομένα, όπως το Google, το Flickr, το

Facebook και η Walmart, μεγάλος αριθμός συστοιχιών διακομιστών (server farms) αναπτύσσονται σε όλο το στον κόσμο για να εξασφαλίστούν υπηρεσίες χωρίς διακοπές και γρήγορες απαντήσεις στις κατά τόπους αγορές. Αυτές οι αυτόνομες πηγές δεν είναι μόνο οι λύσεις που προέκυψαν απο τεχνικές αναλύσεις, αλλά και τα αποτελέσματα της νομοθεσίας και των κανονιστικών πλαισίων που ισχύουν σε διάφορες χώρες / περιοχές. Για παράδειγμα, οι ασιατική αγορά της Walmart είναι εγγενώς διαφορετική από τις αγορές της Βόρειας Αμερικής σε όρους εποχιακών προωθήσεων, κορυφαίων προϊόντων πώλησης και καταναλωτικών συμπεριφορών. Πιο συγκεκριμένα, οι κανονισμοί τοπικής αυτοδιοίκησης επηρεάζουν επίσης τη διαδικασία διαχείρισης χονδρικής και θα έχει ως αποτέλεσμα αναδιαρθρωμένες αναπαραστάσεις και αποθήκες δεδομένων για τις τοπικές αγορές αυτές.

Πολύπλοκες και Εξελισσόμενες Αλληλεξαρτήσεις

Ενώ αυξάνεται ο όγκος των Big Data, το ίδιο συμβαινει και με την πολυπλοκότητα και τις κρυφές σχέσεις μέσα στα δεδομένα. Στα πρώιμα στάδια της εποχής των κεντρικοποιημένων πληροφοριακών συστημάτων δεδομένων, το ζητούμενο ήταν η εύρεση των καλύτερων τιμών χαρακτηριστικών που αντιπροσωπεύουν τη κάθε παρατήρηση. Αυτό είναι ανάλογο με τη χρήση πολλών πεδίων δεδομένων, όπως η ηλικία, το φύλο, το εισόδημα, το επίπεδο της εκπαίδευσης και ούτω καθεξής για τον χαρακτηρισμό του κάθε άτομο. Σε αυτό το είδος δειγματοληψίας η αναπαράσταση αντιμετωπίζει εγγενώς το κάθε άτομο ως μια ανεξάρτητη οντότητα χωρίς να λαμβάνει υπόψη την και τις κοινωνικές τους διασυνδέσεις, που είναι ένας από τους σημαντικότερους παράγοντες της ανθρώπινης φύσης και κοινωνίας. Οι κύκλοι φίλων μας μπορούν να σχηματιστούν με βάση τα κοινά χόμπι ή τους άνθρωπους που μας συνδέουν βιολογικές σχέσεις. Τέτοιες κοινωνικές συνδέσεις συχνά υπάρχουν όχι μόνο στις καθημερινότητα μας, αλλά και σε πολύ δημοφιλή cyberworlds. Για παράδειγμα, μεγάλες ιστοσελίδες κοινωνικής δικτύωσης όπως το Facebook ή το Twitter, χαρακτηρίζονται κυρίως από κοινωνικές λειτουργίες όπως συνδέσεις φίλων και οπαδούς. Οι συσχετισμοί μεταξύ των ατόμων μπορούν εγγενώς να περιπλέξουν ολόκληρη την αναπαράσταση των δεδομένων και κάθε λογική επεξεργασία . Στην αναπαράσταση δείγματος χαρακτηριστικών, τα άτομα θεωρούνται παρόμοια αν έχουν κάποιο παρόμοιο χαρακτηριστικό ενώ στην αναπαράσταση δείγματος χαρακτηριστικού-σχέσης, δύο άτομα μπορούν να συνδεθούν μεταξύ τους (μέσω των κοινωνικών τους συνδέσεων) παρόλο που μπορεί να μην έχουν τίποτα κοινό σε όλους τους τομείς χαρακτηριστικών. Σε ένα δυναμικό κόσμο, τα χαρακτηριστικά που χρησιμοποιούνται για να αντιπροσωπεύουν τα άτομα και οι κοινωνικοί δεσμοί που παριστάνουν τις σχέσεις μας μπορούν επίσης να εξελίσσονται βάσει χρονικούς, χωρικούς ή άλλους παράγοντες. Μια τέτοια επιπλοκή γίνεται μέρος της πραγματικότητας των εφαρμογών Big Data, όπου το κλειδί είναι να δωθεί σημασία στις σύνθετες (πολλαπλές και μη γραμμικές) σχέσεις δεδομένων, μαζί με τις εξελισσόμενες αλλαγές, με σκοπό την ανακαλύψη χρήσιμων μοτίβων από τις συλλογές μεγάλων όγκου δεδομένων.

1.4 Τεχνικά Χαρακτηριστικά

1.4.1 Βασικές Αρχές Λειτουργίας

[13. Deloitte]

CAP Theorem

Στη θεωρητική πληροφορική, το θεώρημα της CAP, δηλώνει ότι είναι αδύνατο για ένα κατανεμημένο σύστημα αποθήκευσης δεδομένων να εγγυηθεί ταυτόχρονα περισσότερα από δύο εκ των τριών παρακάτω χαρακτηριστικών:

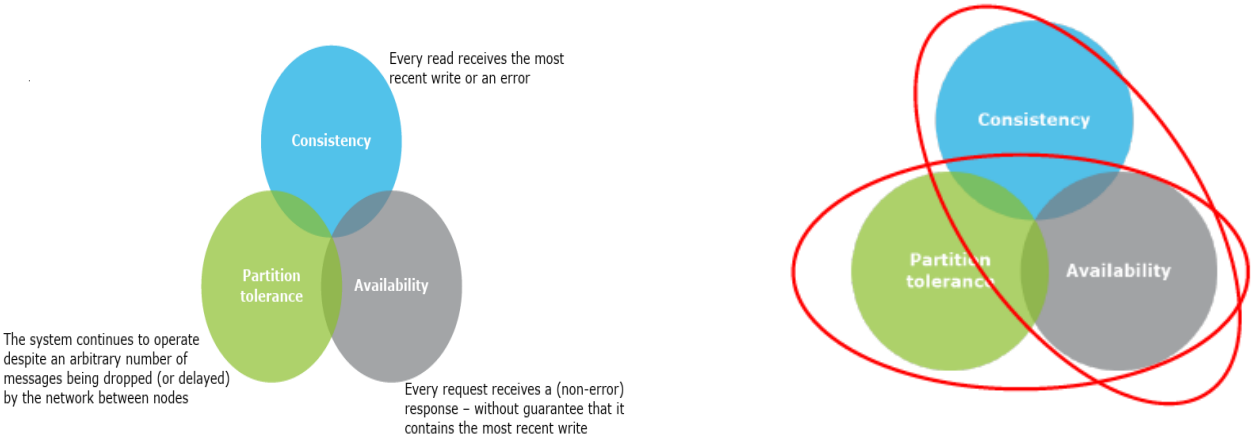
Consistency (Συνέπεια): Κάθε ανάγνωση δεδομένων επιστρέφει την πιο πρόσφατη εγγραφή ή μήνυμα σφάλματος

Availiability (Διαθεσιμότητα): Κάθε αίτημα λαμβάνει πάντοτε μια απάντηση (μη-εσφαλμένη), χωρίς όμως την εγγύηση ότι αυτή θα είναι και η πιο πρόσφατη εγγραφή

Partition Tolerance (Αντοχή Κατατμήσεων): Το σύστημα συνεχίζει να λειτουργεί παρόλη την ύπαρξη ενός αριθμού μηνυμάτων που παραπέφτουν (ή καθυστερούν) κατα την μεταφορά τους μεταξύ κόμβων του δικτύου.Όταν συμβαίνει μια αποτυχία κατάτμησης δικτύου, πρέπει να αποφασίσουμε

- Ακύρωση τη λειτουργίας και μειώση τη διαθεσιμότητας, αλλά εξασφαλίση με αυτό τον τρόπο της συνέπειας

- Συνέχιση τη λειτουργίας και παροχη της διαθεσιμότητας, αλλά αυξάνεται ο κίνδυνος ασυνεπειών



Εικόνα 0.Θεώρημα CAP

BASE Theorem

Οι υπηρεσίες θεωρούνται τύπου κατά περίπτωση συνέπεις(**Eventual Consistency**) συχνά θεωρείται ότι παρέχουν τις συμβάσεις BASE (Basically Available, Soft State, Eventual consistency), σε αντίθεση με τις παραδοσιακές εγγυήσεις ACID (Atomicity, Consistency, Isolation, Durability).Οι πρωτεγενείς ορισμοί του κάθε στοιχείου του BASE έχουν ως εξής:

(**B**) Basical (**A**) Availiabilty: Οι λειτουργίες ανάγνωσης και γραφής είναι διαθέσιμες όσο το δυνατόν περισσότερο (χρησιμοποιώντας όλους τους κόμβους της συστοιχίας της βάσης δεδομένων), αλλά χωρίς καμία εγγύηση συνέπειας (η εγγραφή(write) τελικά μπορεί να μην καταχωρηθεί μετά την διευθέτηση των διενέξεων ή να μην είναι εφικτό να αναγνωστεί η τελευταία εγγραφή)

(**S**) Soft State : Χωρίς εγγυήσεις συνέπειας, μετά από κάποιο χρονικό διάστημα, έχουμε μόνο μια κάποια πιθανότητα να γνωρίζουμε την ακριβή κατάσταση, αφού ίσως δεν έχει συγκλίνει ακόμα

(**E**) Eventual Consistency: Σε σύστημα που λειτουργεί, μπορέσουμε τελικά να γνωρίζουμε ποια είναι η κατάσταση της βάσης δεδομένων με την προυπόθεση ότι υπάρχει ανοχή στο περιμένουμε αρκετά μεγάλο χρονικό διάστημα μετά από κάθε εσιαγωγή ενός νέου συνόλου δεδομένων. Από κει και πέρα, οποιαδήποτε νεότερη ανάγνωση θα συμβαδίζει με τα αναμενόμενα στοιχεία.

1.4.2 Είδη Ανάπτυξης των Δεδομένων

[17.Doshi, Zhong, Lu ,Tang ,Lou, Deng]

Χρονολογική

Ορίζουμε την αύξηση λόγω της συσσώρευσης δεδομένων σε ημέρες, μήνες και έτη. Ένα τυπικό παράδειγμα θα μπορούσε να είναι τα λεπτομερή αρχεία στα συστήματα χρέωσης των εταιριών παροχής τηλεφωνικών υπηρεσιών. Για μια επαρχία μεσαίου μεγέθους, τα αρχεία κλήσεων για ένα μηνα είναι δισεκατομμύρια σε αριθμό και αρκετά σε terabytes (TBs) σε μέγεθος. Η συσσώρευση αυτή είναι σπάνια σημαντική για τις χρηματικές συναλλαγές που επιφέρουν, αλλά αποτελούν ένα ισχυρό ορυχείο δεδομένων που υποβοηθά τον στρατηγικό σχεδιασμό και την επέκταση.

Οριζόντια

Ονομάζεται αυτή που προκαλείται ως αποτέλεσμα εισαγωγής νέων υπηρεσιών ή διαδικασίων σε μια επιχείρηση ώστε να ανταποκρίνεται στις νέες επιχειρηματικές ανάγκες της. Για παράδειγμα σε πληροφοριακό σύστημα ενός νοσοκομείου μπορεί να εισαχθεί μια νέα διαδικασία που θα ψηφιοποιεί όλες τις ιατρικές εικόνες και θα τις

ενσωματώνει στην κλινική ροή εργασίας. Το αποτέλεσμα θα είναι η δραματική αύξηση των δεδομένων για τα οποία ένα παραδοσιακό RDBMS θεωρείται ανεπαρκές οπότε οι νέες διαδικασίες ενδέχεται να αποσταθεροποιήσουν τα υπάρχοντα συστήματα. Αντι αυτου, οι νέες ποικιλίες ημιδομημένων πληροφοριών μπορούν να στεγαστούν και και επεξεργάζονται μέσω τεχνολογιών NewSQL.

Κάθετη

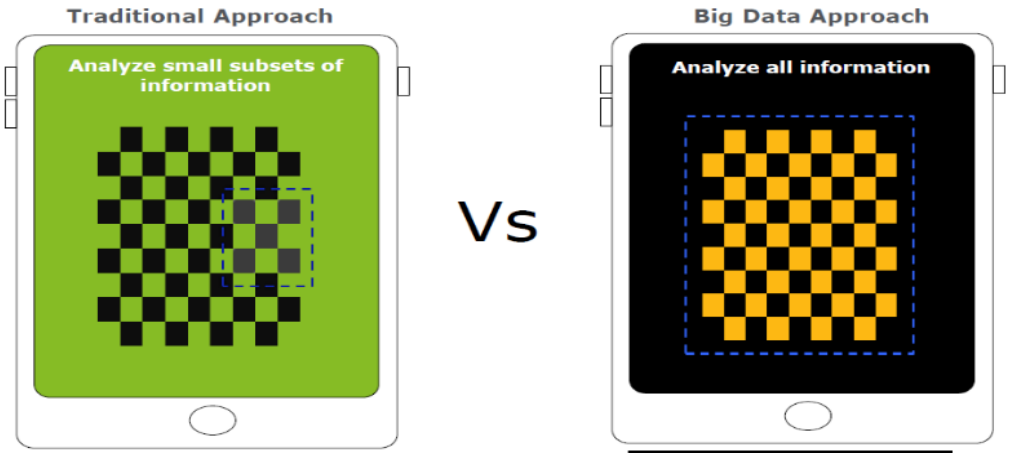
Η κάθετη ανάπτυξη αναφέρεται στην αύξηση των δεδομένων λόγω αντίστοιχης αύξησης της γενικότερης επιχειρηματικής δραστηριότητας και της πολυπλοκότητας. Για παράδειγμα, η μεγαλύτερη εστίαση στο περιβάλλον που αναγκάζει τις μονάδες παραγωγής ενέργειας σε βελτίωση της μετάδοσης ισχύος και της αποδοτικότητας της κατανάλωσης -οδηγώντας σε περισσότερους αισθητήρες, συχνή δειγματοληψία κλπ. Έτσι η ανάγκη που προκύπτει για υψηλή και ταυτόχρονη επεξεργασία σε πραγματικό χρόνο μπορεί εύκολα πνίξει ένα RDBMS και απαιτεί νέες προσεγγίσεις για το χειρισμό των δεδομένων των έξυπνων μετρητών με τρόπο διαφορετικό απ'οτι άλλα σχεσιακά δεδομένα.

1.4.4 Η παραδοσιακή Προσέγγιση και οι Νέες Δυνατότητες

[14. Deloitte]

Χρήση Δεδομένων

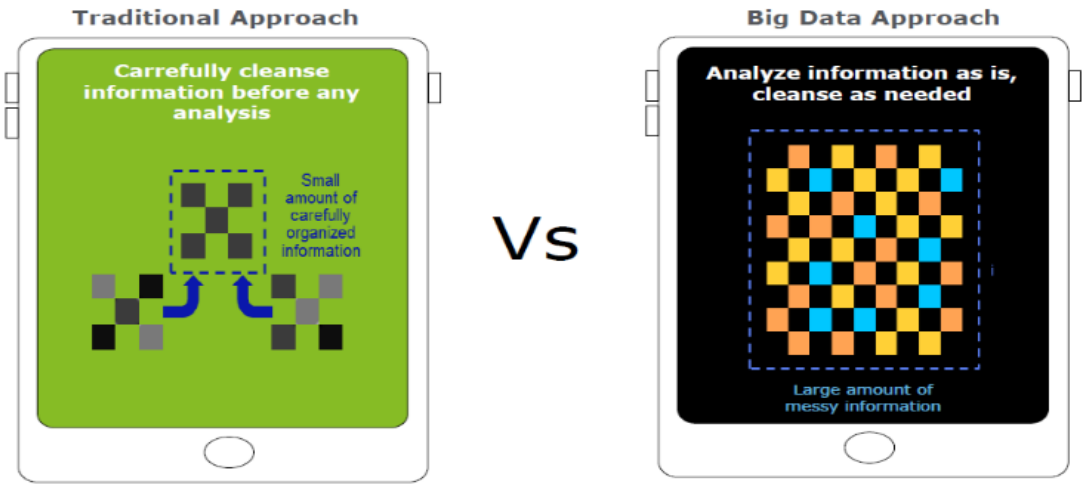
Ανάλυση Ενός Μοντελοποιημένου Υποσυνόλου | Ανάλυση Όλων των Διάθεσιμων Συνόλων



Εικόνα 1.Χρήση Δεδομένων

Απαιτήσεις Χρήσης

Καθαρισμός της Πληροφορίας πριν από οποιαδήποτε ανάλυση | Αναλύση τις Πληροφορίας ως έχει, καθαρισμός μόνο όπου απαιτείται



Εικόνα 2.Απαιτήσεις Χρήσης

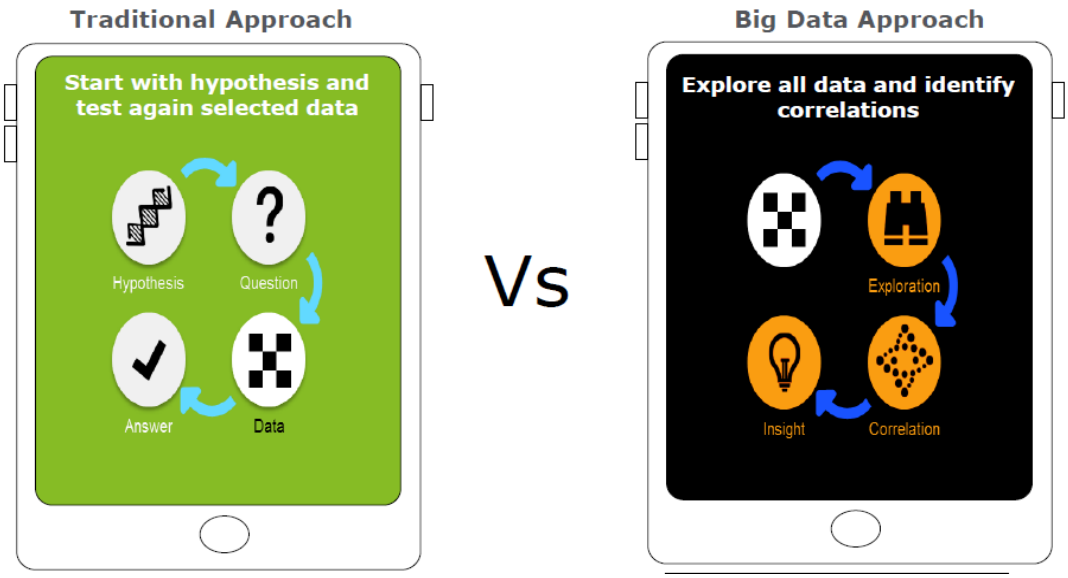
Διαδικασία Ανακάλυψης Πληροφορίας/Εξόρυξης Γνώσης

Παράθεση μιας υπόθεση και δοκιμή/επιβεβαίωση της επι συνόλου προσεκτικά επιλεγμένων δεδομένων

Υπόθεση→Ερώτηση→Δεδομένα→Απάντηση

Εξετάστε όλα των δεδομένων ανεξαιρέτως και προσπάθεια εντοπισμού των συσχετισμών

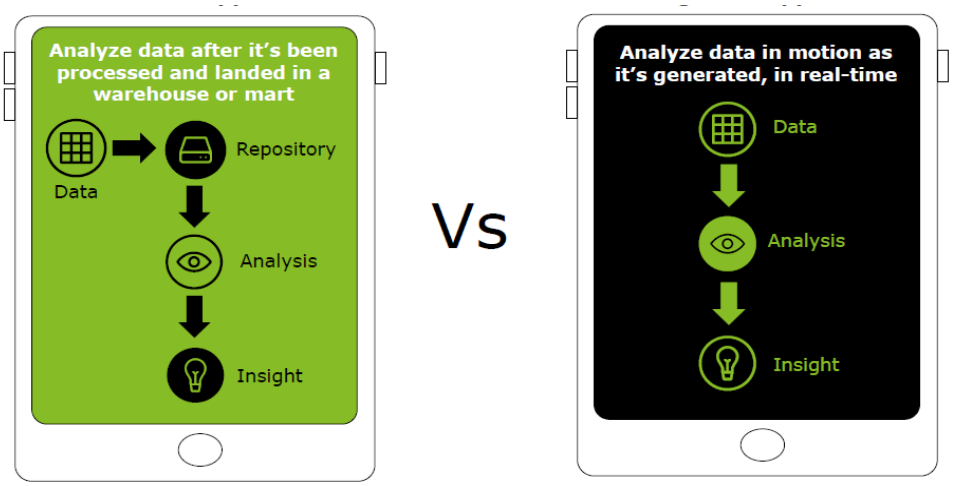
Δεδομένα→Εξερεύνηση→Συσχετισμοί→Γνώση



Εικόνα 3.Ανακάλυψη Πληροφορίας

Ανάλυση Κινούμενων Δεδομένων

Αναλύση των δεδομένων μετά την επεξεργασία και την φόρτωσή τους στο datawarehouse ή datamart /Αναλύση των δεδομένων σε κίνηση καθώς παράγονται, σε πραγματικό χρόνο



Εικόνα 4.Ανάλυση Κινούμενων Δεδομένων

1.4.5 Ελλείψεις Τεχνολογικών Προτύπων

[4.NIST]

Η ανάπτυξη συστημάτων Big Data παρουσιάζει σημαντικές δυσκολίες που προκύπτουν από την ίδια την φύση τους .Ωστόσο, παρατηρούνται και σημαντικά κενά στην προτυποποίηση των τεχνολογιών που εμπλέκονται σε αυτήν. Οι δραστηριότητες τυποποίησης που σχετίζονται αφοριούν στους ακόλουθους τομείς:

1.Περιπτώσεις Χρήσης Big Data, ορισμοί, κοινό λεξιλόγιο και αρχιτεκτονικές αναφοράς (π.χ. σύστηματα, δεδομένα, πλατφόρμες, σε απευθείας σύνδεση / εκτός σύνδεσης).

2.Προδιαγραφές και τυποποίηση των μεταδεδομένων, συμπεριλαμβανομένης της περιγραφής προέλευσης των δεδομένων (data provenance)

3.Μοντέλα Εφαρμογών (π.χ. batch, stream).

4.Γλώσσες Ερωτημάτων (Queries), συμπεριλαμβανομένων των μη-σχεσιακών ερωτημάτων για την υποστήριξη διαφόρων τύπων δεδομένων (π.χ. XML, Resource Description Framework [RDF], JSON, πολυμέσων) και λειτουργίες επί Big Data (π.χ. matrix operarions σε πίνακες)

5.Γλώσσες εξειδικευμένες για συγκεκριμένους τομείς

6.Σημασιολογία της ενδεχόμενης συνέπειας.

7.Προηγμένα πρωτόκολλα δικτύων για αποτελεσματική μεταφορά δεδομένων.

8.Γενικές οντολογίες και ταξινομίες για την περιγραφή της σημασιολογίας δεδομένων, συμπεριλαμβανομένης της διαλειτουργικότητας μεταξύ οντολογιών.

9.Έλεγχος της ασφάλειας και πρόσβασης ιδιωτικού απορρήτου.

10.Απομακρυσμένα (remote), κατανεμημένα και καθολικά (federated) analytics , συμπεριλαμβανομένων της ερευνας ανακάλυψης των πηγών των δεδομένων αυτών και των πόρων επεξεργασίας τους

11.Κοινή Χρήση και ανταλλαγή δεδομένων.

12.Αποθήκευση δεδομένων (δηλ. σύστημα Αποθήκευσης στην Μνήμη (in-memory), κατανεμημένο σύστημα αρχείων, αποθήκη δεδομένων).

13.Ανθρώπινη κατανάλωση και χρήση των αποτελεσμάτων της ανάλυσης των Big Data (δηλ. Οπτικοποίηση-Visualization).

14.Μέτρηση των αναγκών ενέργειας των συστημάτων Big Data.

15.Διασύνδεση μεταξύ σχεσιακών βάσεων δεδομένων (δηλ. SQL) και μη σχεσιακών (δηλ. NoSQL)

16.Περιγραφή και διαχείριση της ποιότητας και της αξιοπιστίας των Big Data (περιλαμβάνει και το master data management).

1.4.6 Ειδικά Θέματα Μηχανικής Λογισμικού | Software Engineering Issues

[2. Karakaya]

Επεκτασιμότητα | Scalability

Στα συστήματα Big Data υπάρχουν κατά βάση δύο μεγάλα ζητήματα επεκτασιμότητας: Των **υπολογιστικών λειτουργιών** και της **αποθήκευσης** δεδομένων.Συνήθως αυτές γίνονται με την προσθήκη πόρων στην συστοιχία υπολογιστών (cluster).Το λογισμικό επομένως που λειτουργεί στις υποδομές αυτές ίσως χρειαστεί να χειριστεί απο δεκάδες έως χιλιάδες υπολογιστές στην ίδια συστοιχία. Εδώ λοιπόν είναι που γεννιέται η πρόκληση:Πως μπορούμε να είμαστε σίγουροι πως η επέκταση των υπολογιστικών πόρων χωρίς αλλαγή στις εκάστοτε εφαρμογές που τους χρησιμοποιούν δεν θα δημιουργήσει πρόβλημα στην λειτουργία τους σε συνθήκες φόρτου εργασίας και μεγάλης εντάσεως δεδομένων (data-intensive workloads); Σε αυτές τις περιπτώσεις λοιπόν ο σχεδιασμός των εφαρμογών και των αλγορίθμων γίνεται κρίσιμο κομμάτι της διαδικασίας ανάπτυξης λογισμικού. Παρόλο λοιπόν που τα πλαίσια αρχιτεκτονικής των συστηματων Big Data είναι φτιαγμένα ώστε να μπορούν να επεκτείνονται κατ'απαίτηση, το λογισμικό που αναπτύσσεται πρέπει επίσης να είναι επεκτάσιμο. Και αντιστρόφως, για να μπορεί το λογισμικό να χρησιμοποιεί τα οφέλη της επεκτάσιμότητας, ο μηχανικός θα πρέπει να αναπτύσσει κατάλληλα των κώδικα και τους αλγορίθμους.

Διαθεσιμότητα | Availiability

Η ανάπτυξη συστημάτων Big Data χωρίς την χρήση κατανεμημένων τεχνολογιών υπολογισμού και αποθήκευσης θεωρείται σχεδόν αδύνατη.Οι τεχνολογίες αυτές όμως ως γνωστόν εμπεριέχουν πάντα τον κίνδυνο της κατάρευσσης ενός στοιχείου μιας συστοιχίας. Ο εντοπισμός και η αποκάτασταση ενός σφάλματος είναι επίσης αδύνατος χωρίς την χρήση ενός πλαισίου και απο την άλλη σε οποιοδήποτε πρόβλημα του συστήματος το λογισμικό θα πρέπει να ολοκληρώνει την διεργασία του χωρίς σφάλματα. Αυτά τα είδη προβλημάτων επιλύονται συνήθως απο τον διαχειριστή πόρων και τον χρονοπρογαμματιστή του πλαισίου των Big Data. Όμως τι θα συμβεί στην περίπτωση που η αιτία του σφάλματος είναι άγνωστη και παρουσιαστεί κατά την προσπάθεια επέκτασης; Σε αυτή την κλίμακα η χρήση εργαλείων αποσφαλμάτωσης ή δειγματοληψίας είναι σχεδόν αδύνατη. Επομένως γεννιέται η ανάγκη των λειτουργιών ανάλυσης loggs, το οποίο αποτελει από μόνο του μια ξεχωριστή πρόκληση για τα συστήματα Big Data. Ας πάρουμε το πάραδειγμα όπου έχει δημιουργηθεί ένα σύστημα που δέχεται δεδομένα από μία έξυπνη πόλη πχ.αισθητήρες,βίντεο κτλπ και παρατηρούμε μια απρόσμενη αποτυχία που προέρχεται από ένα στοιχείο που δεν μπορούμε να εντοπίσουμε, με αποτέλεσμα να χάνεται μεγάλο ποσό πληροφόρησης από της πηγές. Μπορούμε εκείνη τη στιγμή ως διορθωτική ενέργεια να διακόψουμε την λειτουργία του συστήματος; Φυσικά και όχι, σε αυτά τα συστήματα η διαθεσιμότητα είναι κρίσιμος παράγοντας και αυτό απαιτεί απο τους μηχανικούς λογισμικού να σκέφτονται βαθυτερα γύρω από την αρχιτεκτονική και τις μεθόδους αλλαγής των συστημάτων.

Ένταση Μνήμης

Η προσπέλαση τεραστιων όγκων δεδομένων απαιτεί αρκετό χρόνο ώστε να αναγνωστούν τα δεδομενα από τις μονάδες αποθήκευσης. Για την επιτάχυνση αυτής τη διαδικασίας θα πρέπει να είναι εφικτή η παράλληλη ανάγνωση ο χειρισμός της οποία αποτελεί μια μεγάλη πρόκληση για τους προγραμματιστές. Η αποθήκευση δηλαδή σε ένα μόνο σημείο πρέπει να αποφεύγεται και σχετικά και με την επεκτασιμότητα που έχει ήδη αναλυθεί, απαιτείται λοιπόν η χρήση κατανεμημένων συστημάτων αποθήκευσης στα οποία το πλάισιο διαχείρισης επιτρέπει την οριζόντια προσθήκη νέων στοιχείων αποθήκευσης. Η μεγάλη πρόκληση που γεννιέται όμως είναι ότι οι εφαρμογές πρέπει να έχουν γνώση των αναπαραγωγών των δεδομένων και να χειρίζονται τυχών ασυνέπειες που μπορεί να δημιουργηθούν απο αλληλοσυγκρουόμενες ενημερώσεις που θα λάβουν χώρα στα κατά τόπους αντίγραφα.

Έλεγχος και Αξιοπιστία Περιβάλλοντος Δοκιμών

Τις περισσότερες φορές είναι σχεδόν αδύνατο να δημιουργηθεί ένα αντιπροσωπευτικό υποσύνολο δεδομένων για τις δοκιμές σε μια εφαρμογή Big Data καθώς αυτή διαχειρίζεται τεράστιους όγκους διαφοροποιημένων δεδομένων.Επίσης, μπορεί να μην είναι τεχνικά εφικτή η προσπέλαση πραγματικών δεδομένων για σκοπούς ελέγχου. Κατά μια προσέγγιση αυτό θα μπορούσε να επιλυθεί με την χρήση προτυπων πολυσύνθετων δεδομένων, όμως και πάλι μπορούμε να καταλήξουμε σε περιττή ανασκόπηση όλων των αγνώστων σεναρίων ακόμα και αν αυτά δεν έχουν καμία πιθανότητα εμφάνισης στην δική μας περίπτωση.

Επίσης η αξιοπιστία ενός αναπτυχθεντος συστήματος δεν είναι εξασφαλίσμενη απλά με την μεταφορά της εγκατάστασης από το περιβάλλον δοκιμών στην παραγωγή καθώς η συμπεριφορά της κατανεμημένης εφαρμογής μπορεί να διαφοροποιείται σημαντικά σε πραγματική λειτουργία και αριθμό συστοιχιών επεξεργασιας. Αυτό σημαίνει πως οι μηχανικοί λογισμικού πρέπει να δοκιμάζουν το λογισμικό που αναπτύσσουν σε πραγματικές συνθήκες, διαδικασία που μπορεί να χρειαστεί ολόκληρες εβδομάδες.

Διαχείριση Δεδομένων Καταγραφής Δραστηριότητας/Παρακολούθησης Συστήματος

Η λειτουργία των συστημάτων Big Data παράγει τεραστιες ποσότητες αρχείων καταγραφής τα οποία μπορεί να χρησιμοποιούνται και από την ίδια την εφαρμογή.Οι μηχανικοί πρέπει να μεριμνούν για την αποθήκευση τους, την ανάλυση, καθώς και την λήψη προληπτικών ενεργειών βάσει αυτών.

Ανάκαμψη Σφαλμάτων

Όταν ένας κόμβος σε ένα σύστημα Big Data καταρρέυσει ή διακοψει προσωρινά την λειτουργία του δεν χάνουμε μόνο την εκτέλεση της τρέχουσας διεργασίας που είχε αναλάβει αλλά επιπλέον είναι πιθανόν να ανακύψει ζήτημα επαναταξινόμισης των δεδομένων η ακόμα και απώλειας τους.Για αυτό το λόγο, τα σύνολα δεδομένων θα πρέπει να είναι πολλαπλώς αντεγραμμένα σε παραπάνω κόμβους τις ίδιας συστοιχίας. Οι δυνατότητες των πλαισίων, ιδιαίτερα αυτών τύπου DFS, είναι σε θέση να προσπερνούν αυτό το πρόβλημα.

Ασφάλεια

Τα συστήματα Big Data προορίζονται για την χρήση πολλαπλών συμμετεχόντων που σημαίνει πως ο καθένας από αυτούς μπορεί να προσθέσει την εφαρμογή του στην ίδια συστοιχία.Υπό αυτήν την έννοια οι μηχανικόι πρέπει να μεριμνούν για την ασφάλεια και την απομόνωση των εφαρμογών που λειτουργούν επί των ίδιων πόρων

1.5 Προκλήσεις

1.5.1 Προκλήσεις στο Κύκλο Ζωής της Διαδικασίας Ανάπτυξης Λογισμικού

[15. Hummel, Eichelberger, Giloj, Werle, Schmid]

1.5.1.1 Διαχείριση Έργου και Προδιαγραφών

Ασαφείς Προδιαγραφές | Unclear Requirements

Κατά την ανάπτυξη των Big Data συστήματων, τα επιθυμητά αποτελέσματα είναι συχνά άγνωστα στην αρχή καθώς οι ενδιαφερόμενοι δεν μπορούν να φανταστούν τις δυνατότητες και τη δυναμική των αναλύσεων ούτε οι μελλοντικές τους επιθυμίες μπορούν να προβλεφθούν από την χρήση ενός υπό εξέλιξη σύστηματος. Αυτό δημιουργεί προκλήσεις για τα εμπλεκόμενα μέρη καθώς και για τις εφαρμοζόμενες μέθοδους. Σε κάποιο βαθμό, αυτό αλλάζει την καταγραφή των απαιτήσεων δραματικά, αφού οι αναλυτές δεδομένων πρέπει να εξηγήσουν το δυνατόν συντομότερο ποιές αναλύσεις είναι ρεαλιστικές και εφικτές και ποιές θα ήταν καλύτερο να αποφευχθούν π.χ. λόγω θεμάτων με στατιστική πιστότητα ή προστασία προσωπικών δεδομένων. Επιπλέον, όπως περιγράφεται παρακάτω, ενδέχεται να συνεχίσουν να εμφανίζονται νέες απαιτήσεις και ιδέες κατά τη διάρκεια της υλοποίησης, των δοκιμών και της λειτουργίας ενός συστήματος καθιστώντας ως μονόδρομο την ευέλικτη και διερευνητική προσέγγιση στη διαχείριση και την ανάπτυξη.

Ανάδυση Νέων Απαιτήσεων απο τα Δεδομένα | Emergence of New Requirements from Data

Οι παρατηρήσεις από τα δεδομένα μπορούν να οδηγήσουν σε νέες απαιτήσεις ή να επαναπροσδιορίσουν τις υπάρχουσες απαιτήσεις. Αυτό μπορεί να ανοίξει το δρόμο για νέες λειτουργίες ή διαφορετική υλοποίηση κάποιας υφιστάμενης λειτουργικότητας. Μερικές φορές αυτό μπορεί να να γίνει αντιληπτό μόνο κατά τη λειτουργία ενός συστήματος, γεγονός που οδηγεί σε ετεροχρονισμένες μεταβολές των απαιτήσεων. Ως εκ τούτου, μπορεί βασικά να μην είναι δυνατόν να οριστούν όλες οι απαιτήσεις πριν το τελικό σύστημα ξεκινήσει να εκτελείται. Ακόμη πιο δύσκολο, από την άποψη της διαχείρισης απαιτήσεων, είναι το γεγονός ότι τα δεδομένα μπορούν να γίνουν από μόνα τους ένας παράγοντας λήψης αποφάσεων, καθώς τα νέα δεδομένα θα μπορούσαν να διευκολύνουν νέες ευκαιρίες ανάλυσης. Ωστόσο, οι υπάρχουσες μέθοδοι ανάλυσης τεχνικών απαιτήσεων (Requirements Engineering) που επικεντρώνονται κατά κύριο λόγο στην εκ των προτέρων αντικειμενοστρεφή ανάλυση ενός πεπερασμένου συνόλου δεδομένων με γνωστή δομή δεν είναι σε θέση να αντιμετωπίσουν αυτή την κατάσταση.

Ετεροκλητες Ομάδες Εργασίας | Highly Interdisciplinary Teams

Η δημιουργία νέων και επιτυχημένων συστημάτων ανάλυσης δεδομένων απαιτεί το σχηματισμό και διαχείριση διεπιστημονικών ομάδων. Αυτό περιλαμβάνει την σύνθεση τεχνικών ικανότητων, ικανότητες ανάλυσης δεδομένων καθώς και ικανότητες σε διαθεματικές ενότητες π.χ. στην κοινωνιολογία, την εθνογραφία, την ψυχολογία, νομικά πλαίσια κ.α. Επίσης η συμμετοχή εξωτερικών συμβούλων είναι χρήσιμη , ιδίως κατά την κατάρτιση των πρώτων έργων, ωστόσο, για να εξασφαλιστεί η μακροπρόθεσμη επιτυχία, η ανάληψη αρμοδιοτήτων εσωτερικά είναι απαραίτητη. Επιπλέον, η εξυπηρέτηση καθηκόντων ανάλυσης δεδομένων είναι ιδιαίτερα δύσκολη για χρήστες που δεν είναι επαγγελματίες ανάλυσης δεδομένων και οι οποίοι είναι δεν είναι εξοικιωμένοι με τους κοινές έννοιες και αλγορίθμους ανάλυσης δεδομένων, επομένως αυτό απαιτεί νέους τρόπους επικοινωνίας και διαχείρισης και των ανθρώπων.

Διασύνδεση Στοιχείων Υλικού & Λογισμικού | Integration of Hardware &Software Components

Η εγκατάσταση συστημάτων Big Data απαιτεί συχνά την ενσωμάτωση διαφορετικών frameworks π.χ. για την εφαρμογή της Αρχιτεκτονικής Lambda. Επίσης, η επεξεργασία μεγάλων δεδομένων συχνά δεν μπορεί να επιτευχθεί μόνο μέσω εξειδικευμένου λογισμικού. Ειδικό υλικό επεξεργασίας όπως είναι οι μονάδες επεξεργασίας γραφικών (GPUs) χρησιμοποιείται ευρέως ως οικονομικά αποδοτική, ισχυρή και κλιμακούμενη πλατφόρμα για διάφορες εργασίες που απαιτούν υψηλών απαιτήσεων υπολογισμούς. Ο συνδυασμός επεξεργασίας με βάση το υλικό και της επεξεργασίας με βάση το λογισμικό σε ενιαία υβριδικά συστήματα είναι δύσκολη, καθώς απαιτεί ομάδες με διευρυμένες τεχνικές γνώσεις που μπορεί να κυμαίνονται από επιλογή στοιχείων υλικού, στον σχεδιασμό μέσω γρήγορων τεχνολογιών δικτύωσης έως και τις ικανότητες του λογισμικού και τις δυνατότητες διαλειτουργικότητας. Αυτό προσθέτει ακόμα μεγαλύτερη πίεση για διεπιστημονική συνεργασία όπως αναφέρεται στην προηγούμενη παράγραφο

Επιπτώσεις στην Ιδιωτικότητα | Privacy Implications

Αναλύσεις δεδομένων με χρήση και σύνθεση των προσωπικά δεδομένων πρέπει να συμμορφώνονται με τις αντίστοιχες κατευθυντήριες γραμμές προστασίας, τις άδειες χρήσης (εφόσον υπάρχουν) και την μεταβαλόμενη νομοθεσία, π.χ., χρησιμοποιώντας κρυπτογράφηση, διαμέρισεις, ανώνυμοποίηση, τεχνικές μη-προσωποποιησης ή ψευδονυμων, καθώς μπορεί επίσης να πρόκειται για θέματα διαπραγμάτευσης στα συμβούλια επιχειρήσεων. Έως ένα βαθμό, μια στρατηγική ασφαλείας στην περίπτωση μας μπορεί να είναι υπερβολικά περιοριστική, καθώς δεν είναι πάντα προφανές στους εξουσιοδοτημένους χρήστες όπως οι προγραμματιστές για το ποιά είδη αναλύσεων επιτρέπονται με τα δεδομένα που εχουν συλλεχθεί. Η προστασία των δεδομένων είναι επίσης ένας ταχέως αναπτυσσόμενος τομέας, δηλ. δημιουργεί ανάγκη για νομικούς συμβούλους ή ακόμη απαιτεί δημιουργία ειδικών θέσεων για υπαλλήλους που θα είναι υπεύθυνοι για την προληπτική παρακολούθηση της συμμόρφωσης προς τις νόμικες αλλαγές και θα αναγνωρίζουν τις απαραίτητες αλλαγές που αυτές επιφέρουν στα υπάρχοντα συστήματα.

Πολύπλοκες Ισορροπίες Μεταξύ Ποιότητας και Απόδοσης | Complex Trade-offs between Quality and Performance

Η προσεκτική εφαρμογή ευρέως αποδεκτών σχεδιαστικών τακτικών επιτρέπει συχνά την αναβάθμιση των παραδοσιακών πληροφοριακών συστημάτων με έναν σχετικά αναίμακτο τρόπο. Για τα συστήματα Big Data, ωστόσο, ο συνδυασμός των τεράστιων συνόλων δεδομένων με μεγάλης-κλίμακας κατανεμημένες λογικές και τους συμβιβασμούς μεταξύ της συνέπειας, της διαθεσιμότητας και της κατάμησης (γνωστό ως θεώρημα CAP) συνήθως δημιουργεί περισσότερες δυσκολίες. Πολύ συχνά, αυτό απαιτεί από τους ενδιαφερόμενους να εγκαταλείψουν τις προσδοκίες τους για τέλεια ποιότητα προς την προσεγγιστική, αλλά και ταχύτερους αλγόριθμους με υψηλότερο όμως ποσοστό λάθους. Είναι σαφές ότι αυτό το κάνει πολύ δύσκολο να προβλέψουμε πώς συμπεριφέρεται ένα σύστημα από άποψη της απόδοσης ή καθυστέρησεων. Αυτό, με τη σειρά του, κάνει την εκτίμηση των μεγεθών και το σχεδιασμό κατά την ανάπτυξη συστημάτων Big Data πιο δύσκολη ή ακόμη και απρόβλεπτη και συνεπώς μπορεί να απαιτούνται πρόσθετα πρωτότυπα και δοκιμές, π.χ., με πολύ υψηλά ή μεταβαλλόμενα φορτία.

Επιπλέον, στα συστήματα Big Data ο τρόπος εφαρμογής έχει σημαντικό αντίκτυπο στην απόδοση, τις καθυστέρησεις ή την ποιότητα των αποτελέσματων. Για παράδειγμα, ένας αλγόριθμος batch περιορίζει τους βαθμούς της ελευθερίας στην εφαρμογή, έρχεται με υψηλό λανθάνων χρόνο και είναι δύσκολο να χρησιμοποιηθεί με τις τρέχουσες προσεγγίσεις για ανάλυση απόδοσης (όπως πχ την Palladio). Τα τελευταία συχνά δεν υποστηρίζουν τα πλαίσια αρχιτεκτονικής Big Data και τις ιδιότητές τους ως έννοιες μοντελοποίησης, ιδίως δε για την εκτίμηση των επιπτώσεων σε περίπτωση ανταλλαγή μεταξύ παρόμοιων frameworks ή την κλιμάκωσή τους. Τα δε συστήματα δεδομένων που βασίζονται σε cloud εξακολουθούν να αντιμετωπίζουν τυπικά προβλήματα, όπως σύνθετη δημιουργία μοντέλου για middlewares ή απαιτήσεις πόρων για γεγονότα σπάνιας εμφάνισης.

Πρόληψη του Φαινομένου της Αυτοεκπληρούμενης Προφητείας | Prevention of Self-Fulfilling Prophecies

Για την αποτροπή παρερμηνειών των αποτελεσμάτων ή ακόμη και καθοδήγηση της ανάπτυξης κατα τρόπο που απλά αποσκοπεί να επιβεβαιώσει την αρχική εκτίμηση (αυτο-εκπληρούμενη προφητεία), ο σχεδιασμός και η ανάλυση απαιτεί να ακολουθούνται καθιερωμένες επιστημονικές μέθοδοι. Αυτό απαιτεί πειθαρχία από όλα τα εμπλεκόμενα μέρη προς την πρόληψη της υπερβολικής εμβάθυνσης στα δεδομένα και την ορθή εφαρμογή των στατιστικών μεθόδων, π.χ. προσδιορίζοντας υποθέσεις, συλλέγοντας δεδομένα, πειράματα, και την τεκμηρίωση μέσω αποδεικτικών στοιχείων .

1.5.1.2 Αρχιτεκτονική και Ανάπτυξη

Κατανομή Πόρων και Παραλληλία | Distribution and Concurrency

Τα μεγάλα συστήματα δεδομένων είναι ευρέως κατανεμημένα / παράλληλα καθώς ανήκουν στην τάξη των συστημάτων που δεν μπορούν να επεξεργαστούν τον φόρτο δεδομένων σε μία μόνο μηχανή. Κατά συνέπεια, κληρονομούν όλα τα ήδη γνωστά προβλήματα της ανάπτυξης τέτοιων συστημάτων , όπως δυσκολίες στην καθιέρωση μιας συνεκτικής κατανόησης της κατάστασης του συστήματος ή αντιμετώπιση βλαβών του συστήματος.

Μεγάλες Ανάγκες για Εκπαίδευση |Steep Learning Curves due to Novelty of the Field

Ενώ τα παραδοσιακά συστήματα πληροφοριών χρησιμοποιούν συνήθως εφαρμοσμένες ομάδες/λύσεις τεχνολογίας, τα συστήματα Big Data διέπονται πάρα πολύ τις χειροποίητες λύσεις και ως εκ τούτου συχνά αναγκάζουν τους προγραμματιστές να εμβαθύνουν σε ένα νέο συνδυασμό πλαισίων και τεχνολογιών. Κάθε ένα από αυτά μπορεί να οδηγήσει σε απότομη αύξηση της μάθησης, αλλά η διασύνδεση και ενορχήστρωση της αλληλεπίδρασής τους είναι μια νέα πρόκληση κάθε φορά. Αν και οι αρχιτεκτονικές αναφοράς, όπως η αρχιτεκτονική Lambda, και η ευρεία κατανόηση έχει αρχίσει να ωριμάζει, χρειάζεται περισσότερη δουλειά καθώς οι τεχνολογίες εφαρμογής εξακολουθούν να υποφέρουν από έλλειψη τυποποίησης, απότομα βήματα εξέλιξης που ενσωματώνονται στα frameworks , ή ακόμη και την εμφάνιση των εντελώς νέων τεχνολογιών. Συνεπώς, υπάρχει έλλειψη ως προς την βαθειά κατανόηση των τακτικών και μοτίβων της αρχιτεκτονικής Big Data, καθώς επίσης και για τις μεθόδους υλοποίησης και στην συνέχεια την ανάπτυξη τους σε κλιμακούμενα περιβάλλοντα οπως π.χ. το cloud. Αυτό προκαλεί σταθερή πίεση στους αρχιτέκτονες να πειραματιστούν και να μάθουν νέες προσεγγίσεις υλοποίησης. Παρόλο που η κοινότητα των επαγγελματιών αυτό το χρονικό διάστημα έχει ως στόχο τη συλλογή πρότυπων για κάποιο χρονικό διάστημα , η έρευνα γενικά έχει παραμένει σε μια αρκετά ανώριμη κατάσταση και συχνά βασίζεται σε λίγες ανέκδοτες αναφορές (όπως πχ αναρτήσεις ιστολογίων) αφού δεν υπάρχει επαρκής διαθέσιμη βιβλιογραφία σχετικά με τον τρόπο που μπορούμε να οικοδομήσουμε με επιτυχία τα συστήματα μεγάλων δεδομένων. Σε αυτή τη βάση, είναι σαφώς δύσκολο να προβλέψουμε πώς πρέπει να οριστεί μια αρχιτεκτονική η οποία θα επιτύχει καίριους στόχους έναντι απόδοσης/επεκτάσιμότητας κτλπ και πως να αλληλεπιδράσει με διαφορετικά προτυπα επεξεργασίας.

Εξελικτικός Σχεδιασμός βασισμένος σε ερωτήματα | Query-Driven and Evolutionary Design

Στα παραδοσιακά πληροφοριακά συστήματα, όπου εκτελούνται μόνο απλές λειτουργίες CREATE-READ-UPDATE-DELETE, προσεγγίσεις όπως η χαρτογράφηση O/R (Object/Relation mapping) είναι ώριμες και το να σχεδιάζουν το μόνιμο επίπεδο ενός συστήματος αποτελεί πλέον ρουτίνα παρά πρόκληση. Στα συστήματα Big Data, ωστόσο, η απόκτηση πρόσβασης στα δεδομένα και η επακόλουθη επεξεργασία τους συχνά αποτελούν τον πυρήνα της αναμενόμενης αξίας. Καθώς αυτά τα συστήματα τείνουν να αγγίζουν τεχνολογικά όρια, ο χειρισμός των δεδομένων έχει πολύ πιο βαθύ αντίκτυπο στην αρχιτεκτονική και χρειάζεται προσέγγιση που να βασίζεται στον τύπο ερωτήματων(queries) που πρόκειται να εκτελεστούν. Ακόμη χειρότερα: καθώς η επιθυμητή αξία συχνά δεν είναι σαφής εκ των προτέρων, ο πυρήνας του αρχιτεκτονικού σχεδιασμού συχνά καθοδηγείται από τα ερωτήματα που ανακαλύπτονται μετά την ανάλυση των απαιτήσεων, δηλαδή κατά την φάση δημιουργίας πρωτοτύπων για μεγάλα τμήματα της αρχιτεκτονικής. Έτσι, οι μεγάλες αποφάσεις αρχιτεκτονικής συχνά δεν είναι δυνατές εγκαίρως ή χρειάζεται να γίνουν με εξαιρετικά εξελικτικό(agile) τρόπο. Οι τρέχουσες δυνατότητες της μηχανικής λογισμικου έχουν αρκετές ελλείψεις στην υποστήριξη της δημιουργίας τέτοιου είδους "ευέλικτων αρχιτεκτονικών που βασίζονται σε ερωτήματα- **Query Driven Agile Architectures**".

Ελλειψη Υποστήριξης Ενιαίου Τρόπου Μοντελοποίησης | Lack of Unified Modelling Support

Οι κοινές γλώσσες μοντελοποίησης όπως η **UML** ή η σημειολογία διάγραμματων **E/R** (Entity-Relationship Diagrams) παρέχουν πολύ μικρή υποστήριξη για χρήση σε Big Data έννοιες. Αν και η UML μπορεί να επεκταθεί μέσω προσαρμοσμένων μεταμοντέλων, εξακολουθεί να υπάρχει έλλειψη τυποποιημένων επεκτάσεων, πόσο μάλλον εμπειρία με στην εφαρμογή τους. Ταυτόχρονα, αν και υπάρχουν διάφορες σημειολογίες μοντελοποίησης για τα εργαλεία και τα frameworks Big Data (π.χ. pipe diagrams για MapReduce, pipelines για τα ρεύματα δεδομένων ή τα προφίλ DICE UML), δεν διαθέτουν ισχυρή τυποποίηση και υποστήριξη. Επιπλέον, αναπτύσσονται συχνά εφαρμογές Big Data από πολύ διεπιστημονικές θεωρήσεις οπότε αποτελεί σαφώς μια άλλη πρόκληση η δημιουργία κοινώς αποδεκτών προσεγγίσεων μοντελοποίησης για την αποφυγή σύγχυσης με σημειολογίες από άλλες περιοχές όπως πχ το Data Science

Ιδιοσυγκρασία των χρησιμοποιούμενων Αλγορίθμων | Algorithmic Idiosyncrasies

Τα frameworks των Big Data συχνά απαιτούν την εφαρμογή παραδειγμάτων προγραμματισμού που είναι άγνωστα σε πολλούς προγραμματιστές, π.χ. MapReduce σε μη μετάβλητα δεδομένα. Αυτά τα (νέα) παραδείγματα χρειάζονται έναν διαφορετικό τρόπο σχεδιαστικής αντίληψης των λύσεων που αναπτύσσονται και συχνά επιβάλλουν περιορισμούς στο βαθμό που μπορεί να υπάρξει ελευθερία σε θέματα αρχιτεκτονικής. Επιπλέον, οι αλγόριθμοι Big Data πρέπει να είναι οριζόντιως κλιμακούμενοι , γεγονός που γενικά απαγορεύει ή περιορίζει σοβαρά τη χρήση αλγορίθμων με γραμμικη πολυπλοκότητα χρόνου ή χώρου. Για παράδειγμα, ο περιορισμένος διαθεσίμος χρόνος για επεξεργασία μπορεί να συνεπάγεται τη χρήση ειδικών εκδόσεων αλγορίθμων (π.χ., HyperLogLog), γεγονός που μπορεί να οδηγήσει σε απώλεια της ποιότητας των αποτελεσμάτων. Εδώ, η πρόκληση είναι να εξισορροπηθεί η κατανόηση, η ποιότητα και χρόνος επεξεργασίας των αποτελεσμάτων κατά τους απαιτούμενους σχεδιαστικούς συμβιβασμούς.

Διπλές Υλοποιήσεις | Duplicated Implementations

Καθώς δεν είναι πάντα δυνατό να θυσιάστει η ποιότητα για χάρη της απόδοσης, η πιο δημοφιλής αρχιτεκτονική αναφοράς για Big Data -η αρχιτεκτονική Lambda- προτείνει να υποστηριχθούν και οι δύο κατά τρόπο ευδιάκριτο. Αυτό έρχεται μαζί με το μειονέκτημα της εφαρμογής της λειτουργικότητας δύο φορές: μία φορά για το επονομαζόμενο (streaming) επίπεδο υψηλής ταχύτητας, όπου ο χρόνος είναι πολύ πιο σημαντικός από την ποιότητα, και μία φορά για το λεγόμενο batch επίπεδο, όπου η ποιότητα είναι πιο σημαντική από το χρόνο. Πολύ συχνά, οι αλγόριθμοι είναι διαφορετικοί και τα δύο επίπεδα χρειάζονται τη δική τους εξειδικευμένη συντήρηση, με αποτέλεσμα να μην είναι σπάνιο τα ίδια δεδομένα να αποθηκεύονται δύο φορές. Μερικές φορές δε,που ειδικές και πολυσύνθετες αναλύσεις πρέπει να εκτελεστούν, μπορεί να γίνει ακόμη περισσότερο δύσκολο ναι ικανοποιηθούν όλες οι απαιτήσεις απόδοσης. Εδώ, η χρήση της αποκαλούμενης πολυγλωσσικής συντήρησης (**Polyglot Persistence**) έχει γίνει συνήθης πρακτική. Ωστόσο, έρχεται με το τίμημα της διπλής ή πολλαπλής αποθηκευσης δεδομένων και κώδικα. Μια άλλη εναλλακτική σχετικά με την ταχεία επεξεργασία δεδομένων μπορεί να είναι οι λωρίδες προτεραιότητας, όπου τα σημαντικά δεδομένα θα πρέπει να προωθούνται προς την μονάδα επεξεργασίας γρήγορα, αλλά και πάλι με αυτόν τον τρόπο προστίθεται πολυπλοκότητα στον σχεδιασμό. Έτσι η δυσκολία εστιάζεται η κατανόηση διαφορετικών εκδόσεων του ίδιου αλγόριθμου ως προς την συνέπεια τους π.χ. βασισμένοι σε προσεγγίσεις καταλληλες για το μοντέλο.

Συνέπεια και Διαθεσιμότητα των Δεδομένων | Data Consistency and Availability

Παρότι η εφαρμογή Polyglot Persistence από μόνη της δεν είναι αρκετά δύσκολη, συνήθως έρχεται μαζί με προβλήματα συνέπειας μεταξύ διαφορετικών συστημάτων αποθηκευσης της επιχειρησιακής πληροφορίας.Οι έντονες συζητήσεις γύρω από το θεώρημα CAP και τη χρήση "Soft State" βάσεων δεδομένων τύπου NoSQL υπογραμμίζουν τη σημασία αυτού του θέματος. Ωστόσο, οι συζητήσεις αυτές περιστρέφονται κυρίως στην επίλυση της συνέπειας και της διαθεσιμότητας σε ένα μόνο συντηρούμενο σύστημα που δουλεύει κάτω από το βαρύ φόρτου και όχι μέσα σε ένα δίκτυο πολλών διασυνδεδομένων και επικοινωνούντων συστημάτων με διαφορετικά χαρακτηριστικά.

Αναπτυξη Συστήματος έναντι Ανάπτυξης Πλατφόρμας | System vs. Platform Development

Ενώ συνήθως μιλάμε για ολοκληρωμένα συστήματα Big Data, η ανάπτυξή τους μάλλον μοιάζει περισσότερο με μια πλατφόρμα που στήνεται πάνω από ενα οικοσύστημα λογισμικού. Αυτό μπορεί να έχει δύο αιτίες:

(α) **Την προγραμματισμένη ανάπτυξη ως πλατφόρμα**: η σημαντική επένδυση για συστήματα Big Data μπορεί συχνά να δικαιολογηθεί μόνο εάν μπορούν να επωφεληθούν μέσω αυτού μια σειρά απο υφιστάμενες εφαρμογές. Το σύστημα δηλαδή σχεδιάζεται ως πλατφόρμα από το αρχή

(β) **Ανάπτυξη ως πλατφόρμα λόγω αβεβαιότητας**: Εδώ ο στόχος αποτελεί ένα συγκεκριμένο σύστημα, αλλά λόγω υψηλης αβεβαιότητας, η ανάπτυξη χωρίζεται στην γενική πλατφόρμα και τα συγκεκριμένα λειτουργικά στοιχεία στην κορυφή αυτού. Μερικές φορές, εκεί υπάρχει επίσης ένας υβριδικός τρόπος: πλατφόρμες όπως το Twitter και το Facebook αρχικά σχεδιάστηκαν ως κλειστά συστήματα, αλλά μετασχηματίστηκαν σε πλατφορμες με την πάροδο του χρόνου. Έτσι, για επιτυχημένα συστήματα Big Data πρέπει να αντιμετωπίστουν ως προκλήσεις που έχουν να κάνουν με τη διαχείριση της μεταβλητότητας και το σχεδιασμό μακρόπνοων συστημάτων.

Προσαρμογή Χρόνου Εκτέλεσης και Ανεξαρτησία Πλατφόρμας | Runtime Adaptability and Platform Independence

Η ανάπτυξη με τρόπο που θα είναι ανεξάρτητος από την πλατφόρμα αποτελεί από καιρό μια πρόκληση στον τομέα της τεχνολογίας λογισμικού, αλλά ακόμα και μετά από τόσα χρόνια από την πρόταση του **Model-Driven Architecture** δεν υπάρχει ικανοποιητική λύση. Δεν υπάρχει βελτίωση δηλαδή για περιβάλλοντα Cloud, όπου συνήθως αναπτύσσονται οι εφαρμογές Big Data. Αντίθετως μάλιστα, παρουσιάζονται μεμονωμένα εργαλεία από κάθε vendor και ο φόβος δέσμευσης σε έναν συγκεκριμένο vendor αποτελεί κοινό πρόβλημα. Έτσι, πχ το έργο DICE (http://www.dice-h2020.eu) προβλέπει μοντέλα εφαρμογών Big Data που είναι ανεξάρτητα της τεχνολογίας (**technology agnostic**). Το marketing στον τομέα του cloud computing και η ανάγκη αντιμετώπισης διαφορετικών φορτίων δεδομένων έχουν οδηγήσει και σε μια άλλη επιθυμία: Οι εφαρμογές να είναι απρόσκοπτα επεκτάσιμες ή "ελαστικές". Αυτό από άποψη αρχιτεκτονικής απαιτεί πρόσθετες λειτουργικότητες παρακολούθησης, μια μηχανή οπτικοποίησης ικανή να παράγει γνώση από τα δεδομένα παρακολούθησης, διακόπτες ρύθμισης ή αυτόματη προσαρμογής όπως π.χ. προσομοιωτές προσθήκης νέου υλικού κατά την εκτέλεση.

1.5.1.3 Διασφάλιση Ποιότητας

Οπτικοποίηση και Τεκμηρίωση των Αποτελεσμάτων | Challenging Visualization and Explainability of Results

Η απεικόνιση των Big Data, ιδίως όσων έχουν πολλές διάστασεις (προέρχονται δηλαδή από σύνθετα ή ετερογενή σύνολα δεδομένων), είναι γνωστό πως είναι μια δύσκολη υπόθεση. Εδώ είναι σημαντικό να βρεθεί ο σωστός συνδυασμός διαστάσεων και επιθυμητού επιπέδου ανάλυσης που απαιτείται για την απεικόνιση, ώστε ο χρήστης να μπορέσει να αντλήσει πληροφορίες και, με τη σειρά του, να αξιολογήσει την εγκυρότητα των αποτελεσμάτων. Η αλληλεπίδραση με οπτικοποιήσεις σε πραγματικό χρόνο – με την δυνατότητα πχ παραμετροποίησης του χρονικού εύρους των εμφανιζόμενων δεδομένων ή του πλήθους των εμφανιζόμενων διαστάσεων -μπορούν να βοηθήσουν τους χρήστες να κατανοήσουν τα δεδομένα, αλλά και να μπορέσουν επίσης να παρέχουν ανατροφοδότηση σχετικά με θέματα επιδόσεων ή να αυξήσουν την πολυπλοκότητα του συστήματος . Ακόμη και αν τα αποτελέσματα της ανάλυσης παρουσιάζονται με τρόπο κατανοητό, η λεπτομερής διαδικασία για τον τρόπο με τον οποίο μπορεί να προκύψει αυτό το αποτέλεσμα ούτε είναι εύκολα κατανοητό ούτε δυνατόν να εντοπιστεί καν ακόμα και από τους ίδιους τους ειδικούς. Το πρόβλημα γίνεται χειρότερο εάν οι αποφάσεις παιρνονται, για παράδειγμα, από μοντέλα τεχνητής νοημοσύνης όπως πχ το Deep Learning. Ως εκ τούτου, η αξιοπιστία, η κατανόηση των δεδομένων, η επεξεργασία και τα αποτελέσματα της ανάλυσης είναι μια ιδιαίτερη πρόκληση που συνδέεται άμεσα την κατασκευή του συστήματος από άποψη μηχανικής λογισμικού.

Μη Διαισθητική Αντίληψη της Συνέπειας | Non-Intuitive Notion of Consistency

Επειδή τα συστήματα Big Data μπορούν να ανταλλάξουν την συνέπεια για χάρη της διαθεσιμότητας σε ένα συγκεκριμένο τμήμα του δικτύου ή για την απόδοση γενικά, η eventual consistency μπορεί να είναι αποδεκτή σε γενικές γραμμές. Ωστόσο, η ασυνέπεια μπορεί να προκαλέσει σύγχυση στους χρήστες και στους μηχανικούς ποιότητας, όπως για παράδειγμα όταν κατά την στιγμή ενημέρωσης των δεδομένων αυτά δεν είναι ορατά στα αιτήματα που ακολουθούν. Εδώ η πρόκληση είναι ο χειρισμός της συνέπειας με ένα κατανοητό τρόπο, π.χ., εξασφαλίζοντας ότι ο χρήστης βλέπει τουλάχιστον τις δικές του πρόσφατες αλλαγές. Ωστόσο, αυτό μπορεί και πάλι να οδηγήσει σε προβλήματα επιδόσης ή αυξημένη πολυπλοκότητα του συστήματος

Πολύπλοκη Επεξεργασία και Διαφορετικές Κατανοήσεις της Συνέπειας | Complex Data Processing and Different Notions of Correctness

Η επεξεργασία μεγάλων δεδομένων είναι συνήθως πολύπλοκη, π.χ. λόγω πολλών αλληλεπιδράσεων μεταξύ των φαινομενικά διακριτών βημάτων επεξεργασίας. Ως εκ τούτου, είναι πολύ δύσκολο να προσδιοριστεί το πότε μια λειτουργία είναι σωστή, καθως η επίδραση μιας μεμονωμένης ενέργειας στο σύνολο των αποτέλεσματων μπορεί να είναι σχετικά μικρή. Λόγω της πολυπλοκότητας της επεξεργασίας δεδομένων και των υπολογισμών επίσης ίσως είναι δύσκολο να προσδιοριστούν κατάλληλα τα αποτελέσματα δοκιμών. Αυτό οδηγεί στο παράδοξο του να μην έχουμε ακριβή δεδομένα για δοκιμές. Κατά συνέπεια, ο έλεγχος της αξιοπιστίας γίνεται εντελώς τυπικά. Η πρόκληση είναι να σχεδιαστούν απλοί μηχανισμοί ελέγχου οι οποίοι θα μπορούν να εντοπίζουν συγκεκριμένες πιθανότητες λάθους σε καταστάσεις ασάφειας ενώ παράλληλα θα μπορούν ξεκάθαρα να αναγνωρίζουν -όπου είναι εφικτό- λανθασμένες εφαρμογές βάσει των σημερινών ορθολογικών δυνατότητων και μεθόδων ελέγχου.

Υψηλές Απαιτήσεις Υλικού για τις Δοκιμές | High Hardware Requirements for Testing

Βλέπωντας τα κύρια χαρακτηριστικά των συστημάτων Big Data (όγκος, ταχύτητα και ποικιλία) καταλαβαίνουμε πως η κατανεμημένη επεξεργασία και ο υψηλός φόρτος εργασίας είναι αναπόστασστο κομμάτι τους. Στην πράξη, αυτό οδηγεί σε συστήματα που αποτυγχάνουν συχνά λόγω απροσδόκητων μικρών προβλημάτων όπως πχ η έλλειψη αποθήκευτικού χώρου. Επομένως, η διεξοδική δοκιμή συστημάτων Big Data απαιτεί ένα παρόμοιο φόρτο εργασίας με αυτό που θα χρησιμοποιηθεί στο πραγματικό σύστημα καθως και δοκιμές σε θέματα όπως ο παραλληλισμός, απόδοση, κλιμάκωση, κ.λπ. Για το σκοπό αυτό, ένα κατάλληλο σύστημα ελέγχου συγκρίσιμο με το σύστημα παραγωγής πρέπει να είναι διαθέσιμο. Ωστόσο, η πρακτική αυτή συχνά δεν είναι εφικτή γιατί το κόστος μπορεί να είναι απαγορευτικό ή γιατί το μόνο ικανοποιητικά μεγάλο σύνολο μηχανών είναι απλά διαθέσιμο μόνο για την παραγωγή. Κατά συνέπεια, κάποιο μέρος των δοκιμών μπορεί να είναι δυνατό μόνο με την χρήση υλικού παραγωγής, γεγονός το οποίο μπορεί να καθυστερήσει την συνολική ανάπτυξη του συστήματος. Ακόμη χειρότερα, η δοκιμή πραγματικό φόρτου εργασίας μπορεί να χρειαστεί να γίνει κατά τη διάρκεια της πραγματικής λειτουργίας και γι 'αυτό απαιτούνται και ειδικές προφυλάξεις και διαχείριση κινδύνου για αποφυγή επιπτώσεων στο σύστημα παραγωγής. Επιπλέον, δοκιμαζωντας ένα δυναμικά κλιμακούμενο, αλλά πιθανότητα ανώριμο σύστημα δημιουργούνται σοβαροί οικονομικοί κίνδυνοι για τις εγκαταστάσεις που βασίζονται σε υπηρεσίες cloud κατά παραγγελία. Έτσι, η δοκιμή συστημάτων Big Data σε ρεαλιστικές συνθήκες είναι πραγματικό πρόβλημα για την ποιότητα των προϊόντων και υπηρεσιών.

Δυσκολία Δημιουργίας Επαρκών Δεδομένων Υψηλής Ποιότητας | Difficult Generation of Adequate, High-Quality Data

Η δοκιμή μεγάλων συστημάτων δεδομένων απαιτεί ρεαλιστικά σύνολα δεδομένων υψηλής ποιότητας. Ενώ μεγάλες ποσότητες δεδομένων μπορεί να δημιουργηθούν απλά πολλαπλασιάζοντας μικρότερα αντιπροσοπευτικά υποσύνολα δεδομένων, μπορεί και πάλι να απαιτούνται terabyte αποθήκευσης για την αναπαραγωγή των λιγότερο ομαλών σύνολων . Επιπλέον, το να βασιζόμαστε σε δοκιμές που σχετίζονται μόνο με τον όγκο είναι υπερβολικά περιοριστικό στο να καλύφθουν και τα υπόλοιπα συναφή φαινόμενα δηλ. Και τα άλλα "Big Data Vs" όπως η ποικιλία και η ακρίβεια πρέπει επίσης να παρέχονται (αν είναι απαραίτητο, συμπεριλαμβάνωντας διαφόρους τύπους δεδομένων όπως έγγραφα ή βίντεο). Συνοπτικά, η δημιουργία και ο χειρισμός ρεαλιστικών σύνολων δεδομένων για δοκιμές σε συγκεκριμένες εφαρμογές ώστε να καλύπτουν όλα τα σχετικά χαρακτηριστικά αποτελεί μια πρακτική και μεθοδολογική δυσκολία.

Ελλειψη Μεθόδων Αποσφαλμάτωσης, Καταγραφής Λειτουργίας Και Σφαλμάτων | Lack of Debugging, Logging, and Error-Tracing Methods

Λόγω της κατανεμημένης φύσης τους, τα συστήματα Big Data πρέπει τελικά δοκιμάζονται σε ένα κατανεμημένο περιβάλλον. Ωστόσο, οι περιορισμένες σήμερα δυνατότητες στα εργαλεία κατανεμημένης ανάπτυξης και εντοπισμού σφαλμάτων δημιουργει μια άλλη δυσκολία: το ότι οι προγραμματιστες που χρησιμοποιούν τα frameworks Big Data συχνά έχουν βασιστέι σε κατανεμημένα αρχεία καταγραφής. Αυτό περιπλέκει την επεξεργασία καθώς η κατανόηση των πληροφοριών μπορεί να απαιτεί τη συγχώνευση των αρχείων καταγραφής και τελικά να απαιτεί προηγμένες προσεγγίσεις κατανεμημένης αποσφαλμάτωσης και οπτικοποίησης τους.

State Explosion in Verification

Η χρήση αναπτυσσόμενων clusters για την επεξεργασία Big Data αυξάνει ιδιαίτερα την πολυπλοκότητα στην εφαρμογή προσεγγίσεων επαλήθευσης λόγω εκθετικής υπολογιστικής έκρηξης. Παρότι η χρήση συμβολικών μοντέλων καταστάσεων ή οι τεχνικές μερικής μείωσης έχουν συμβάλει σημαντικά στην πρακτική επαλήθευση των υποθέσων δοκιμών, εξακολουθούν να χρησιμοποιούνται εξαντλητικά μοντέλα καταστασεων λόγω των καλύτερων δυνατότητων επαλήθευσης που προσφέρουν. Επομένως απαιτείται η εφαρμογή προσεγγίσεων επαλήθευσης για κατανεμημένους υπολογισμούς και, επιπλέον, για την υβριδική επεξεργασία δεδομένων.

Διασφάλιση Ποιότητας Δεδομένων | Ensuring Data Quality

Στα Big Data, η μεταφορά ότι "τα δεδομένα είναι το αργό πετρέλαιο του μέλλοντος" χρησιμοποιείται συχνά. Ωστόσο, τα δεδομένα από μόνα τους είναι περιορισμένης σημασίας. Η αξία προκύπτει μόνο όταν η ποιότητά τους (π.χ. από την άποψη της πληρότητας ή συνέπεια) αποδεικνύεται. Ωστόσο, η αξιολόγηση της ποιότητας των δεδομένων στα πλαίσια των Big Data δεν είναι ούτε σημασιολογικα ούτε υπολογιστικά τετριμένη, ιδίως εάν τα δεδομένα (διαφορετικών τύπων) έχουν ομαδοποιηθεί ή συγχωνεύονται κατά τη διάρκεια της επεξεργασίας.

1.5.1.4 Ανάπτυξη και Λειτουργία

Πολύπλοκη Ελαστική Τροφοδότηση | Complex “Elastic” Provisioning

Καθώς τα συστήματα Big Data πρέπει να είναι κλιμακούμενα κατά απαίτηση, συνήθως εγκαθιστάνται σε περιβάλλον Cloud. Ωστόσο, ζητήματα ιδιωτικότητας, νομικής συμμόρφωσης και αδειοδότησης για εμπορικών στοιχεία ή δεδομένα ενδέχεται να περιορίζουν τα πιθανά διαθέσιμα περιβάλλοντα. Στο QualiMaster για παράδειγμα οι άδειες απαιτούν η επεξεργασία των δεδομένων να γίνεται στις τοπικές υποδομές. Οι ευκίνητοι "ελαστικοί" μηχανισμοί ανάπτυξης είναι ένα κομβικό σημείο όπου πρέπει να αντιμετωπιστούν θέματα όπως η πολυπλοκότητα της κατανομής, οι παραμετροποιήσεις των στοιχείων, εξαρτήματα που είναι παρόμοια αλλά όχι πλήρως υποκατάστατα, όπως στην polyglot persistence. Επιπλέον, οι τεχνολογίες virtual machines και containers αυξάνουν την πολυπλοκότητα, δεδομένου ότι απαιτούν στενή συνεργασία μεταξύ της ανάπτυξης, της εγκατάστασης, και της λειτουργίας (**DevOps**). Επιπλέον, τα συστήματα μακρόπνοης λειτουργίας ενδέχεται απαιτούν τη μεταφορά του συστήματος μεταξύ εναλλακτικών cloud πλατφόρμων με την πάροδο του χρόνου, τα οποία συχνά είναι περιορισμένα ή και πλήρως παραμετροποιημένα απο τον vendor για συγκεκριμένα μόνο περιβάλλοντα. Εκτός από τις επιπτώσεις της μετακίνησης δεδομένων κατά τη διάρκεια της μετεγκατάστασης, ενδέχεται να υπάρχουν καταστάσεις δέσμευσης στον προμηθευτή (vendor lock-in) οι οποίες πρέπει να προληφθούν μέσω προσπάθειας από όλα τα εμπλεκόμενα μέρη για τυποποίηση και μοντέλοποιήση που επιτρέπουν τη δημιουργία και την μεταφορά της απαιτούμενη λειτουργικότητα βασιζόμενη σε μια αφηρημένη προδιαγραφή.

DO2 Complex Monitoring

Παρακολούθηση των επιχειρησιακών συστημάτων Big Data πρέπει να συμπεριλαμβάνει όλους τους εμπλεκόμενους πόρους, την επεξεργασία και τα επεξεργασμένα δεδομένα. Σε κάποιο βαθμό,τα Big Data Frameworks περιλαμβάνουν ήδη μηχανισμούς παρακολούθησης και να παρέχουν πίνακες ελέγχου που επιτρέπουν στο μηχανικό δεδομένων να επίβλεπει και να βελτιστοποιεί τις λειτουργίες. Τυπικά, οι υπάρχουσες λύσεις δεν είναι πλήρεις. π.χ.στο Apache Storm ενσωματώθηκαν περαιτέρω μηχανισμοί για να δώσουν μια γενική εικόνα της κατάσταση του συστήματος, συμπεριλαμβανομένου του χρησιμοποιούμενου υλικού και του λειτουργικού σύστηματος. Επιπλέον, οι χρησιμοποιούμενες μετρικές είναι συνήθως μη - τυποποιημένες και επομένως μη-συγκρίσιμες μεταξύ των διαφορετικών πλαίσιων ενός cluster. Έτσι, η ευέλικτη, και η χαμηλού κόστους παρακολούθηση των συστημάτων Big Data είναι ενα ανοιχτό θέμα αλλά και εμπόδιο στις προσεγγίσεις προηγμένης επεξεργασίας που περιλαμβάνουν π.χ. αυτοπροσαρμογή

1.5.2 Προκλήσεις στην Μηχανική Δεδομένων | Data Engineering Challenges

[1. Yang, Huang, Li, Liu, Hu]

Αποθήκευση Δεδομένων | Data Storage

Οι προκλήσεις αποθήκευσης δημιουργούνται από τον όγκο, την ταχύτητα και την ποικιλία των μεγάλων δεδομένων. Αποθήκευση μεγάλων δεδομένων στις παραδοσιακές τεχνολογίες φυσικής αποθήκευσης είναι προβληματική καθώς οι σκληροί δίσκοι (HDD) συχνά αποτυγχάνουν και οι παραδοσιακοί μηχανισμοί προστασίας δεδομένων (π.χ. RAID ή περιττές σειρές ανεξάρτητων δίσκων) δεν είναι αποτελεσματικοί για αποθήκευση σε κλίμακα Petabytes PB. Επιπλέον, η ταχύτητα των μεγάλων δεδομένων απαιτεί η αποθήκευση να γίνεται με τέτοιο τρόπο ώστε να είναι δυνατή η ταχεία κλιμάκωση που είναι δύσκολο να επιτευχθεί με τα παραδοσιακά συστήματα αποθήκευσης. Οι υπηρεσίες αποθήκευσης του Cloud (π.χ. Amazon S3, Elastic Block Store ή EBS) προσφέρουν σχεδόν απεριόριστη αποθήκευση με υψηλή ανοχή σφάλματων που παρέχει πιθανές λύσεις για την αντιμετώπιση προκλήσεων αποθήκευσης των Big Data. Ωστόσο, η μεταφορά και φιλοξενία δεδομένων στο cloud θεωρείται δαπανηρή δεδομένου του όγκου δεδομένων. Επομένως υπάρχει ανάγκη ανάπτυξης προτύπων και αλγόριθμων που θα λαμβάνουν υπόψη τα χωροχρονικά πρότυπα της χρήση δεδομένων και θα προσδιορίζουν την αναλυτική αξία των δεδομένων και της συντήρησής τους μέσω της εξισορρόπησης του κόστους αποθήκευσης και της μετάδοσης δεδομένων σε σχέση με τη γρήγορη συσσώρευση τους.

Μεταφορά Δεδομένων | Data Transmission

Η μετάδοση δεδομένων λαμβάνει χώρα σε διάφορα στάδια του κύκλου ζωής των δεδομένων ως εξής:

(i) συλλογή δεδομένων από αισθητήρες

(ii) διασύνδεση δεδομένων μεταξύ πολλαπλών κέντρων δεδομένων

(iii) διαχείριση δεδομένων για τη μεταφορά στις πλατφόρμες επεξεργασίας (π.χ. πλατφόρμες νέφους)

(iv) ανάλυση για μετακίνηση δεδομένων από την αποθήκευση στον κεντρικό αναλυτικό υπολογιστή (π.χ. συστοιχία υπολογιστών υψηλής απόδοσης (High Performance Computing Clusters-HPC).

Η μεταφορά μεγάλων όγκων δεδομένων δημιουργεί προφανείς προκλήσεις σε κάθε ένα από αυτά τα στάδια. Ως εκ τούτου, έξυπνες τεχνικές προεπεξεργασίας και οι αλγόριθμοι συμπίεσης δεδομένων απαιτούνται για την αποτελεσματική μείωση του μεγέθους των δεδομένων προς μεταφορά. Για παράδειγμα, ο Li et al. (2015) πρότεινε ένα αποδοτικό μοντέλο μετάδοσης δικτύου με ένα σύνολο τεχνικών συμπίεσης δεδομένων για μετάδοση γεωχωρικών δεδομένων σε περιβάλλον εικονικοποιημένων υποδομών (Virtualized Infastructure-CyberInfastructure). Επιπλέον, κατά τη μεταφορά Big Data σε cloud πλατφόρμες από τοπικά data center, απαιτούνται αποτελεσματικοί αλγόριθμοι που θα προτείνουν αυτόματα την καταλληλότερη υπηρεσία νέφους (ως προς την τοποθεσία) βασιζόμενοι στις χωροχρονικές μεταβλητές που θα μεγιστοποιήσουν την ταχύτητα μεταφοράς δεδομένων ενώ παράλληλα θα ελαχιστοποιούν το κόστος.

Διαχείριση Δεδομένων | Data Management

Είναι δύσκολο για τους υπολογιστές να διαχειρίζονται αποτελεσματικά, να αναλύουν και να απεικονίζουν μεγάλα, αδόμητα και ετερογενεη δεδομένα. Οι προυποθέσεις της ποικιλίας και της αξιοπιστίας των Big Data επαναπροσδιορίζουν το μοντέλο διαχείρισης απαιτώντας νέες τεχνολογίες (π.χ. Hadoop, NoSQL) που να καθαρίζουν, να αποθηκεύουν και να οργανώνουν μη δομημένα δεδομένα. Ενώ τα μεταδεδομένα είναι απαραίτητα για την ακεραιότητα των δεδομένων προέλευσης, παραμένει πρόκληση η αυτόματη δημιουργία μεταδεδομένων που θα περιγράφουν τα Big Data και τις σχετικές διαδικασίες. Η δημιουργία μεταδεδομένων για γεωχωρικά δεδομένα είναι ακόμη δύσκολη λόγω των εγγενών χαρακτηριστικά των πολλαπλών διαστάσεων(τρισδιάστατος χώρος και μονοδιάστατος χρόνος) και πολυπλοκότητα (π.χ. συσχετισμός χώρου-χρόνου και εξάρτηση). Εκτός από τη δημιουργία μεταδεδομένων, τα Big Data δημιουργουν επίσης προκλήσεις για τη βάση δεδομένων (DBMS), επειδή τα παραδοσιακά RBDMS δεν διαθέτουν δυνατότητα κλιμάκωσης για τη διαχείριση και την αποθήκευση σε μη δομημένα δεδομένων. Ενώ οι μη σχεσιακές (NoSQL) βάσεις δεδομένων όπως τα MongoDB και HBase έχουν σχεδιαστεί για Big Data παραμένει πρόκληση η προσαρμογή αυτών στο χειρισμό μεγάλης κλίμακας δεδομένων αναπτύσσοντας αποτελεσματικούς χωροχρονικούς αλγόριθμους ευρετηρίασης και διερεύνησης

Επεξεργασία Δεδομένων | Data Proccessing

Η επεξεργασία μεγάλων όγκων δεδομένων απαιτεί ειδικούς υπολογιστικούς πόρους και αυτό είναι εν μέρει διαχειρίσιμο μέσω της αυξανόμενη ταχύτητα των CPU, του δικτύου και της αποθήκευσης. Ωστόσο, οι υπολογιστικοί πόροι που απαιτούνται για την επεξεργασία των Big Data υπερβαίνουν κατά πολύ την ισχύ επεξεργασίας που προσφέρουν τα παραδοσιακά παραδείγματα επικοινωνίας συστημάτων. Το Cloud computing προσφέρει εικονικά απεριόριστη και κατά παραγγελία ισχύ επεξεργασίας ως μερική λύση. Ωστόσο, η μετακίνηση στο Cloud ανοίγει μια σειρά νέων θεμάτων. Πρώτον, ο περιορισμός του εύρους ζώνης δικτύου του cloud computing επηρεάζει την απόδοση υπολογισμών σε μεγάλους όγκους δεδομένων . Δεύτερο είναι η τοποθεσία που επιλέγεται για την επεξεργασία. Ενώ η "**moving computation to data**" μεταφορά των υπολογισμών στα ίδια τα δεδομένα είναι μια αρχή σχεδιασμού που ακολουθείται από πολλές πλατφόρμες επεξεργασίας όπως π.χ. το Hadoop, η εικονικοποίηση και η διαμοιραζόμενη σύνθεση του cloud computing καθιστούν μια πρόκληση την παρακολούθηση και τη διασφάλιση της τοπικότητας δεδομένων και την υποστήριξη της επεξεργασία δεδομένων που θα περιλαμβάνει εντατική ανταλλαγή δεδομένων και επικοινωνία.

Επιπλέον, η αξιοπιστία των Big Data απαιτεί προεπεξεργασία πριν από τη διεξαγωγή analytics και mining (π.χ. cluster analysis, classification, machine learning) για καλύτερη ποιότητα. Δεν είναι δυνατή η διαχείριση μεγάλων διαστάσεων χωροχρονικών δεδομένων από υπάρχοντες αλγόριθμους μείωσης δεδομένων εντός ενός ανεκτού χρονικού πλαισίου και αποδεκτής ποιότητας. Παραδείγματος χάρην, οι παραδοσιακοί αλγόριθμοι δεν είναι σε θέση να προεπεξεργαστούν σε πραγματικό χρόνο τους μαζικούς όγκους συνεχώς εισερχόμενων δεδομένα απο έξυπνους αισθητήρες ή ακόμα και απλούς αισθητήρες επιτήρησης. Εξαιρετικά αποδοτικοί και κλιμακούμενοι αλγόριθμοι μείωσης δεδομένων είναι απαραίτητοι για την απομάκρυνση του ενδεχομένως άσχετου, περιττού, θορυβώδους και παραπλανητικού περιεχομένου και αυτό είναι μια από τις πιο σημαντικές προκλήσεις στην έρευνα για τα Big Data

Ανάλυση Δεδομένων | Data Analysis

Η ανάλυση δεδομένων είναι μια σημαντική φάση στην αλυσίδα αξίας των Big Data για την εξαγωγή πληροφοριών και προβλέψεων. Ωστόσο, η ανάλυση των Big Data προκαλεί την πολυπλοκότητα και την κλιμάκωση των υποκείμενων αλγορίθμων. Απαιτείται ανάλυση μεγάλων δεδομένων και αντιμετωπίζεται με την κατάρτιση προγράμματων ανάλυσης σε πλατφόρμες παράλληλης επεξεργασίας (π.χ. Hadoop) για την αξιοποίηση των δυνατοτήτων της κατανεμημένης επεξεργασίας. Ωστόσο, αυτή η στρατηγική «διαίρεσης και κατάκτησης» δεν λειτουργεί με βαθιές και πολλαπλές επαναλήψεις που απαιτούνται από αρκετούς αλγορίθμους εξόρυξης δεδομένων. Επιπλέον, οι περισσότεροι υπάρχοντες αναλυτικοί αλγόριθμοι απαιτούν δομημένα ομοιογενή δεδομένα και αντιμετωπίζουν δυσκολία στην επεξεργασία της ετερογένειας των Big Data . Αυτό το κενό απαιτεί είτε νέους αλγόριθμους αμιγώς σχεδιασμένους να αντιμετωπίζουν ετερογενή δεδομένα ή νέα εργαλεία προεπεξεργασίας δεδομένων που θα τα κάνουν δομημένα και ταιριαστά με υπάρχοντες αλγόριθμους.

Οπτικοποίηση | Data Visualization

Η οπτικοποίηση μεγάλων δεδομένων αποκαλύπτει κρυμμένα μοτίβα και άγνωστες συσχετίσεις που βελτιώνουν τη λήψη αποφάσεων. Δεδομένου ότι τα μεγάλα δεδομένα είναι συχνά ετερογενών τύπων, δομής και η σημασιολογίας, η ορθη οπτικοποίηση είναι κρίσιμη για την κατανόηση των δεδομένων. Είναι όμως δύσκολο να παρέχετε οπτική απεικόνιση σε πραγματικό χρόνο και ανθρώπινη αλληλεπίδραση, διερεύνηση και ανάλυση. Η SAS (2012) συνοψίζει πέντε βασικές λειτουργίες για την οπτικοποίηση Big Data ως εξής:

(i) εξαιρετικά διαδραστικά γραφικά που ενσωματώνουν βέλτιστες πρακτικές οπτικοποίησης δεδομένων.

(ii) ολοκληρωμένες, διαισθητικές και προσιτές οπτικές αναλύσεις.

(iii) διαδραστικές διεπαφές μέσω διαδίκτυου για προεπισκόπηση, φιλτράρισμα ή δειγματοληψία δεδομένων πριν από οπτικοποιήσεις

(iv) επεξεργασία εντός της μνήμης και

(v) Απαντήσεις και προβλέψεις εύκολα διανομούμενες και σε κινητές συσκευές και δικτυακές πύλες.

Διασύνδεση Δεδομένων | Data Integration

Η διασύνδεση των δεδομένων (data integration) είναι κρίσιμη για την επίτευξη του 5ου V (Value-Αξίας) των Big Data μέσω της ενοποιημένης ανάλυσης δεδομένων και των διακλαδικών συνεργασιών. Οι Dong και Divesh (2015) συγκέντρωσαν τις προκλήσεις που παρουσιάζει το data integration όπως η αντιστοίχιση σχημάτων (schema mapping), την καταγραφή κύκλου ζωής των καταχωρήσεων (record linkage) και τη διασύνδεση δεδομένων (data fusion). Τα μεταδεδομένα θεωρούνται απαραίτητα για την αυτοματοποιημένη ενοποίηση των πηγών και στην διευκόλυνση αναλύσεων μεγάλης κλίμακας. Ωστόσο, αποτελεσματική και αυτόματη δημιουργία μεταδεδομένων απευθείας από τα Big Data εξακολουθεί να αποτελεί δύσκολο έργο.

Αρχιτεκτονική Δεδομένων | Data Architecture

Τα Big Data μετασχηματίζουν σταδιακά τον τρόπο με τον οποίο διεξάγεται η επιστημονική έρευνα, όπως αποδεικνύει το συνεχώς αυξανόμενο ποσοστό ερευνών που βασίζονται σε ανάλυση δεδομένων (**data-driven research**) και την προσεγγίση ανοιχτών επιστημών (**open-science**). Τέτοιες τάσεις όμως δημιουργούν προκλήσεις στις αρχιτεκτονικές των συστημάτων. Για παράδειγμα, η αποτελεσματική διασύνδεση διαφορετικών εργαλείων και υπηρεσιών παραμένουν υψηλή προτεραιότητα. Πρόσθετα θέματα προτεραιότητας περιλαμβάνουν την ενσωμάτωση αυτών των εργαλείων σε επαναχρησιμοποιήσιμες ροές εργασίας , διασυνδέοντας δεδομένα με τα εργαλεία ώστε να δημιουργηθούν ολοκληρωμένες λειτουργικότητες και την ανταλλαγή αναλύσεων μεταξύ των κοινοτήτων. Η ιδανική αρχιτεκτονική θα πρέπει να μπορεί άψογα να συνθέτει και μοιράζεται δεδομένα, υπολογιστικούς πόρους, δίκτυο, εργαλεία, μοντέλα και το σημαντικότερο, τους ανθρώπους.

Ασφάλεια Δεδομένων | Data Security

Η αυξανόμενη εξάρτηση από τους υπολογιστές και το Διαδίκτυο τις τελευταίες δεκαετίες καθιστά τις επιχειρήσεις και τα άτομα ευάλωτα στην παραβίαση και κατάχρηση δεδομένων. Τα Big Data δημιουργούν νέες προκλήσεις ασφάλειας για τα παραδοσιακά πρότυπα κρυπτογράφησης δεδομένων, τις μεθοδολογίες και τους αλγόριθμους. Προηγούμενες μελέτες στην κρυπτογράφηση δεδομένων επικεντρώνονταν σε δεδομένα μικρού έως μέσου μεγέθους, και δεν λειτουργούν αποτελεσματικά για μεγάλα δεδομένα λόγω προβλημάτων απόδοσης και επεκτασιμότητας. Επιπλέον, οι πολιτικές και τα συστήματα ασφάλειας για τα δομημένα δεδομένα που είναι αποθηκευμένα σε συμβατικά Συστήματα Βάσεων Δεδομένων (ΣΔΒΔ) επίσης δεν είναι αποτελεσματικά για τον

χειρισμό άκρως μη δομημένων και ετερογενών δεδομένων. Έτσι, αποτελεσματικές πολιτικές για τον έλεγχο των προσβάσεων και τη διαχείριση της ασφάλειας πρέπει να διερευνηθούν στο πεδίο των Big Data και πρέπει να ενσωματωθούν νέα συστήματα διαχείρισης δεδομένων και δομών αποθήκευσης. Στην περιοχή της υπολογιστικής νέφους (Cloud), από την στιγμή που οι κάτοχοι των δεδομένων (data owners) έχουν περιορισμένο έλεγχο στην εικονικοποιημένη αποθήκευση (virtualized storage), η εξασφάλιση της εμπιστευτικότητας των δεδομένων, της ακεραιότητας και της διαθεσιμότητας τους γίνεται πρώτο μέλημα.

Ιδιωτικότητα Δεδομένων | Data Privacy

Η άνευ προηγουμένου δικτύωση μεταξύ των έξυπνων συσκευών και υπολογιστικών πλατφορμών συμβάλλουν τα μέγιστα στην δίαδοση των Big Data αλλά ταυτόχρονα θέτουν ανησυχίες σχετικά με το πόσο προστατευμένη είναι η φηφιακή αποθήκευση δεδομένων που άπτονται της ιδιωτικής ζωής των ατόμων , όπως πχ η τοποθεσία που βρίσκεται ή διαμένει, η κοινωνική/καταναλωτική συμπεριφορά και οι συναλλαγές παντώς είδους. Για παράδειγμα, τα μέσα κοινωνικής δικτύωσης και τα μεμονωμένα ιατρικά αρχεία περιέχουν προσωπικές πληροφορίες για την υγεία. Ένα άλλο παράδειγμα είναι ότι οι εταιρείες χρησιμοποιούν Big Data για την παρακολούθηση της απόδοσης του εργατικού δυναμικού καταγράφοντας τις φυσικές κινήσεις και την παραγωγικότητα των εργαζομένων. Αυτά τα ζητήματα ιδιωτικότητας αναδεικνύουν το κενό που υπάρχει μεταξύ των συμβατικών πολιτικών / κανονισμών και των Big Data και σίγουρα απαιτούν την δημιουργία νέων πολιτικών προστασίας των προσωπικών δεδομένων .

Ποιότητα Δεδομένων | Data Quality

Η ποιότητα των δεδομένων περιλαμβάνει τέσσερις πτυχές: ακρίβεια, πληρότητα, πλεονασμό και συνέπεια (Chen et al.2014). Η εγγενής φύση της πολυπλοκότητας και της ετερογένειας των Big Data καθιστά την ακρίβεια και την πληρότητα των δεδομένων δύσκολο να εντοπιστεί και να καταγραφεί, αυξάνοντας έτσι τον κίνδυνο «εσφαλμένων ανακαλύψεων» (Lohr 2012). Για παράδειγμα, τα δεδομένα των μέσων κοινωνικών δικτύων είναι εξαιρετικά διαφεροποιημένα αναλογα με τον τόπο, το χρόνο και τα δημογραφικά στοιχεία. Ειδικά στην καταγραφή της τοποθεσίας η ακρίβεια μπορεί να κυμαίνεται από μέτρα έως εκατοντάδες χιλιόμετρα ανάλογα με τους αισθητήρες καταγραφής. Επιπλέον, ο έλεγχος του εμπλουτισμού των δεδομένων καθώς και το φιλτράρισμα τους πρέπει να διεξάγεται στο σημείο συλλογής δεδομένων σε πραγματικό χρόνο (π.χ. με δίκτυα αισθητήρων, Cuzzocrea, Fortino και Rana 2013. Chen et αϊ. 2014). Τέλος, η διασφάλιση της συνοχής και της ακεραιότητας των δεδομένων είναι πρόκληση στα Big Data ειδικά όταν τα δεδομένα αλλάζουν συχνά και διαμοιράζονται με πολλαπλούς συμμετέχοντες. (Khan et al., 2014).

1.5.3 Προκλήσεις στην Διοίκηση Επιχειρήσεων

[26. McAfee, Brynjolfsson]

Οι εταιρείες δεν πρόκειται να επωφεληθούν πλήρως από τη μετάβαση στη υιοθέτηση τεχνολογιών Big Data εκτός αν είναι σε θέση να διαχειριστούν αποτελεσματικά της αλλαγές που αυτή αποφέρει. Πέντε περιοχές είναι ιδιαίτερα σημαντικές σε αυτή τη διαδικασία είναι:

Ηγεσία

Οι εταιρείες επιτυγχάνουν στην Big Data εποχή όχι μόνο επειδή έχουν περισσότερα ή καλύτερα δεδομένα, αλλά επειδή έχουν ομάδες ηγεσίας ικανές να θέσουν σαφείς στόχους, να σκιαγραφήσουν το τι αποτελεί επιτυχία και θέσουν τα κατάλληλα ερωτήματα. Η ισχύς των Big Data δεν εξαλείφει την ανάγκη για όραμα και ανθρώπινη διορατικότητα. Αντιθέτως, πρέπει να έχουμε ηγετικά στελέχη επιχειρήσεων που μπορούν να εντοπίσουν μια μεγάλη ευκαιρία, να καταλάβουν πώς αναπτύσσεται η αγορά, να σκέφτονται δημιουργικά και προτείνουν πραγματικά νέες προσφορές, να αρθρώσουν ένα συναρπαστικό όραμα, να πείθουν τους άνθρωπους να το αγκαλιάσουν, να δουλέψουν σκληρά ώστε να τους δώσουν να το καταλάβουν και επιπλέον να χειριστούν αποτελεσματικά τους πελάτες, τους εργαζόμενους, τους μετόχους και άλλους ενδιαφερόμενους. Οι επιτυχημένες εταιρείες της επόμενης δεκαετίας θα είναι αυτές που θα έχουν οι ηγέτες που μπορούν να κάνουν όλα τα παραπάνω την στιγμή που οι ίδιες θα αλλάζουν τον τρόπο που λαμβάνουν αποφάσεις.

Διαχείριση Ταλέντων

Δεδομένου ότι τα δεδομένα γίνονται φθηνότερα, τα ακολουθήματα της διαχερίσης αυτών των δεδομένων γίνονται πιο πολύτιμα. Μερικά από τα πιο σημαντικά είναι οι επιστήμονες δεδομένων και γενικότερα οι ειδικευμένοι

επαγγελματίες που εργάζονται με μεγάλες ποσότητες πληροφοριών. Οι στατιστική επιστήμη είναι σημαντική, αλλά πολλές από τις βασικές τεχνικές για τη χρήση Big Data σπάνια διδάσκονται στα παραδοσιακά ακαδημαϊκά μαθήματα στατιστικής. Ίσως ακόμα πιο σημαντικές είναι οι δεξιότητες στον καθαρισμό και την οργάνωση μεγάλων συνόλων δεδομένων καθώς τα νέα είδη δεδομένων σπάνια έρχονται σε δομημένες μορφές. Τα εργαλεία και οι τεχνικές οπτικοποίησης αυξάνουν επίσης την αξία τους. Μαζί με τους επιστήμονες των δεδομένων, μια νέα γενιά επιστήμονων υπολογιστών φέρνει μαζί της τεχνικές εργασίας σε πολύ μεγάλα σύνολα δεδομένων. Η εμπειρία στον σχεδιασμό των πειραμάτων μπορεί να συμβάλει στη μείωση του χάσματος μεταξύ συσχέτισης και αιτιώδους συνάφειας. Οι καλύτεροι επιστήμονες δεδομένων είναι επίσης άνετοι στο να μιλούν τη γλώσσα των επιχειρησιακών συμβούλων και να βοηθούν τους ηγέτες να αναδιατυπώνουν τις προκλήσεις με τρόπους που μπορούν να αντιμετωπίστούν ως πρόβλημματα Big Data. Δεν αποτελεί έκπληξη, ότι άτομα με αυτές τις δεξιότητες είναι δύσκολο να βρεθούν διαθέσιμα εν μέσω τόσο μεγάλης ζήτησης.

Τεχνολογία

Τα εργαλεία που διατίθενται για τη διαχείριση του όγκου, της ταχύτητας και της ποικιλίας των μεγάλων δεδομένων έχουν βελτιωθεί τα τελευταία χρόνια. Γενικά, αυτές οι τεχνολογίες δεν είναι απαγορευτικά δαπανηρές, και πολλες υπάρχουν διαθέσιμες και ως λύσεις λογισμικού ανοικτού κώδικα. Το Hadoop πχ το πιο ευρέως χρησιμοποιούμενο πλαίσιο, συνδυάζει ιδιόκτητο υλικό με λογισμικό ανοιχτού κώδικα. Παίρνει εισερχόμενες ροές δεδομένων και τις διανέμει σε φτηνούς δίσκους. Παρέχει επίσης εργαλεία για την ανάλυση των δεδομένων. Ωστόσο, αυτές οι τεχνολογίες απαιτούν μια δεξιότητα που είναι νέα για τα περισσότερα τμήματα πληροφορικής, τα οποία για να την υιοθετήσουν πρέπει να εργαστούν σκληρά για την ενσωμάτωση όλων των σχετικών εσωτερικών και εξωτερικών πηγών δεδομένων. Η τεχνολογία δηλαδή είναι πάντα απαραίτητη συνιστώσα μιας στρατηγικής μεγάλης δεδομένων.

Λήψη Αποφάσεων

Μια αποτελεσματική οργάνωση δίνει την ίδια βαρυτητα στις πληροφορίες και τα σχετικά δικαιώματα απόφασης επί αυτών. Στη εποχή των Big Data, η πληροφορία δημιουργείται και μεταφέρεται ενω η τεχνογνωσία συχνά δεν εντοπίζεται στα συνηθισμένα μέρη. Ο επιδέξιος ηγέτης θα δημιουργήσει μια οργάνωτικη δομή που θα είναι αρκετά ευέλικτη ώστε να ελαχιστοποιεί το σύνδρομο του "δεν δημιουργήθηκε εδώ" και θα μεγιστοποιεί κατά το δυνατόν περισσοτερο τις διατμηματικές συνεργασίες. Σε άνθρωπους που γνωρίζουν τα προβλήματα πρέπει να δίνονται δικαίωμα στα καταλληλα data αλλά και στους ανθρώπους που γνωρίζουν τεχνικές αντιμετώπισης προβλημάτων να μπορούν να τους τα παρεχουν αποτελεσματικά.

Εταιρική Κουλτούρα

Η πρώτη ερώτηση που αναρωτιέται ένας data-driven οργανισμός δεν είναι "Τι πιστεύουμε;” αλλά "Τι γνωρίζουμε;". Αυτό απαιτεί την απομάκρυνση από το να ενεργεί αποκλειστικά με γνώμονα τις γνώσεις και το ένστικτο. Πρέπει επίσης να αλλάξει η κακή συνήθεια που έχει παρατηρηθει σε πολλούς οργανισμούς: Να προσποιούνται ότι είναι περισσότερο data-driven απ’ ότι είναι στην πραγματικότητα. Πολύ συχνά, τα στελέχη που φτιάχνουν αναφορές και εκθέσεις τις οποίες εμπλουτίζουν με πολλά δεδομένα που δείχνουν πως υποστήριζαν τις αποφάσεις που όμως είχαν ήδη ληφθεί χρησιμοποιώντας παραδοσιακές προσέγγισεις λήψης αποφάσεων (πχ HiPPO-Highest Paid Person Opinion). Δηλαδή τα στελέχη εκ των υστέρων καλούνται να βρουν το αριθμούς που θα δικαιολογούσαν την απόφαση αυτή, με αποτέλεσμα πολύ συχνά να γίνονται εσφαλμένες συσχέτισεις αιτίας-αποτελεσμάτος καθώς και να χρησιμοποιούνται παραπλανητικά πρότυπα στα δεδομένα. Οι προκλήσεις λοιπόν στον τομέα της κουλτούρας είναι τεράστιες και, φυσικά, η ιδιωτικότητα θα τείνει να γίνει μια από τις πιο σημαντικές.

Οι επικρατούσες βασικές τάσεις, τόσο στην τεχνολογία όσο και στην επιχειρησιακή ανάπτυξη, είναι αδιαμφισβήτητες. Τα αποδεικτικά στοιχεία είναι σαφή: Οι αποφάσεις που βασίζονται σε δεδομένα συνήθως είναι και καλύτερες αποφάσεις. Οι ηγέτες είτε θα αγκαλιάσουν αυτό το γεγονός ή θα αντικατασταθούν από άλλους που το κάνουν. Εντός ίδιων κλάδων της οικονομίας , εταιρείες που δεν μπορούν να να συνδυάσουν την εξειδίκευση στον κλάδο με την επιστήμη των δεδομένων θα εκτοπιστούν από τους αντιπάλους τους. Δεν μπορούμε να πούμε ότι οι νικητές θα είναι μόνο αυτοί που θα έχουν αξιοποιήσει τα Big Data στην λήψη αποφάσεων, αλλά τα στοιχεία μας λένε ότι είναι το πιο πιθανό ενδεχόμενο.

1.6 Πεδιά Εφαρμογής και Επιπτώσεις των Big Data

1.6.1 Αλλαγή Μοντέλου στην Επιστήμη και την Τεχνολογία | e-Science

Οι πρόσφατες εξελίξεις γενικώς στις Τεχνολογίες Πληροφορικής & Επικοινωνιών, Cloud Computing και Big Data διευκολύνουν την αλλαγή προς τα νέα μοντέλα του σύγχρονου κλάδου της ηλεκτρονικής επιστήμης (e-Science) που χαρακτηρίζονται από τα ακόλουθα χαρακτηριστικά:

→Μετασχηματισμός όλων των διαδικασιών, συμβάντων και προϊόντων σε ψηφιακή μορφή μέσω πολυδιάστατης πολυπαραγοντικης μέτρησης, παρακολούθησης και ελέγχου.Ψηφιοποίηση του υφιστάμενου και άλλου περιεχομένου.

→Αυτοματοποίηση όλων των διαδικασιών παραγωγής, κατανάλωσης και διαχείρισης δεδομένων, συμπεριλαμβανομένης της συλλογής δεδομένων, αποθήκευσης, ταξινόμησης, ευρετηρίασης και άλλων στοιχείων της γενικής επεξεργασίας και ανακτησης δεδομένων.

→Δυνατότητα επαναχρησιμοποίησης και επανακαθορισμού των αρχικών συνόλων δεδομένων για ανάλυση νέων και δευτερογενών αναλύσεων βασισμένων σε κύκλους επαναθεώρησης και βελτίωσης των αρχικών μοντέλων

→Παγκόσμια διαθεσιμότητα δεδομένων και πρόσβαση στο διαδίκτυο για τις συνεργαζόμενες ομάδες ερευνητών και τεχνολόγων, συμπεριλαμβανομένης της ευρείας πρόσβασης του κοινού σε επιστημονικά ή παραγωγικά δεδομένα.

→Η ύπαρξη των αναγκαίων δομικών στοιχείων και εργαλείων διαχείρισης που επιτρέπουν τη γρήγορη συνένωση, υιοθέτηση και προσαρμογή των υποδομών και των υπηρεσιών, κατόπιν ζήτησης, για συγκεκριμένα ερευνητικά έργα και καθήκοντα.

→Προηγμένες τεχνολογίες ασφάλειας και ελέγχου πρόσβασης που διασφαλίζουν την ασφαλή λειτουργία των σύνθετων υποδομών έρευνας και παραγωγής και επιτρέπουν τη δημιουργία εμπιστευμένου ασφαλούς περιβάλλοντος για συνεργαζόμενες ομάδες ερευνητών και ειδικών τεχνολογίας.

Επιπλέον, οι παρακάτω παράγοντες που θα δημιουργήσουν νέες προκλήσεις και θα δώσουν κίνητρα τόσο για την αλλαγή γενικών παραμέτρων όσο και για την αλλαγή μοντέλων ασφάλειας στο οικοσύστημα Big Data:

- Virtualization: μπορεί να βελτιώσει την ασφάλεια του περιβάλλοντος επεξεργασίας δεδομένων, αλλά δεν μπορεί να λύσει την ασφάλεια των δεδομένων "σε ηρεμία" (**data in rest**).
- Κινητικότητα των διαφόρων συνιστωσών της τυπικής υποδομής όπως αισθητήρες ή πηγές δεδομένων, χρήστες δεδομένων καθώς και τα ίδια δεδομένα (αρχικά δεδομένα και ενδιάμεσα/βοηθητικά δεδομένα). Αυτό έχει ως αποτέλεσμα τα ακόλουθα προβλήματα:
  - -Παροχή υπηρεσιών υποδομής κατά παραγγελία
  - -Διαμοιρασμός περιεχομένου μεταξύ χρηστών του ίδιου τομέα
- Απαιτήσεις για ομοιογενοποιήση σε επίπεδο Big Data τα οποία μπορεί να περιλαμβάνουν δεδομένα από διαφορετικές διοικητικές / λογικές περιοχές και μεταβαλλόμενες δομές δεδομένων (που συνήθως διαφέρουν και σημασιολογικά).
- Λεπτομερής Πολιτική: Τα μεγάλα δεδομένα ενδέχεται να έχουν περίπλοκη δομή και να απαιτούν διαφορετικές και υψηλής λεπτομέρειας πολιτικές για τον έλεγχο της πρόσβασης και των δικαιώματων χειρισμού τους.

1.6.2 Τα Big Data Ανά Τομέα Δραστηριότητας

[25. Chen, Chiang, Storey]

Παρακάτω ακολουθεί μια καταγραφή των εφαρμογών, των πηγών δεδομένων, των χαρακτηριστικών των δεδομένων, των σχετικών analytics και των επιπτώσεων που επιφέρουν τα Big Data στις εξής πέντε ομάδες οικονομικής δραστηριότητας:

- Ηλεκτρονικό Εμπόριο & Marketing
- Ηλεκτρονική Διακυβέρνηση
- Επιστήμη & Τεχνολογία
- Εφαρμογές Υγείας & Επίπεδο Ζωής
- Ασφάλεια & Δημόσια Τάξη

	Ηλεκρονικό Εμπόριο & Marketing	**Ηλεκτρονική Διακυβέρνηση**	**Επιστήμη & Τεχνολογία**	**Εφαρμογές Υγείας & Επιπεδο Ζωής**	**Ασφάλεια & Δημόσια Ταξη**
Εφαρμογή	-Συστήματα Καμπανιών -Παρακολούθηση & Ανάλυση Social Media -Συστήματα Χρηματοδότησης -Κοινωνικά και Εικονικά Παιχνίδια	-Παροχή Κυβερνητικών Υπηρεσιών από Παντού -Ιση πρόσβαση στις δημοσιες υπηρεσίες -Ευαισθητοποίηση και συμμετοχή πολιτών -Πολιτικές καμπάνιες και ηλεκτρονικές ψηφοφορίες	-Καινοτομία -Έλεγχος Υποθέσων και Σεναρίων -Ανακαλυψη Νέας Γνώσης	-Γονιδιωματική ανθρώπιων και φυτών -Υποστήριξη αποφάσεων στον τομέα της πρόνοιας -Αναλυση επιδημιολογικών στοιχίεων	-Ανάλυση εγκλημάτων -Υπολογιστική της εγκληματολογίας -Πληροφορηση πρόληψης της Τρομοκρατίας -Ανοιχτού Κώδικα Υπηρεσίες Πληροφοριών Κυβερνοασφάλεια
Πηγές Δεδομένων	-Αρχεία Καταγραφής -Εγγραφές Συναλλαγών Πελατών -Περιεχόμενο που δημιουργούν από τους πελάτες	-Κυβερνητικές πληροφορίες και Υπηρεσίες -Νόμοι και κανονισμοί -Επανάδραση και σχόλια από τους πολίτες	-Δεδομένα που γεννούν τεχνολογικές συσκευές -Περιεχόμενο από αισθητήρες και δίκτυα	-Γονιδιωματική και ακολουθιακά δεδομένα -Ηλεκτρονικά ιατρικά δεδομένα -Κοινωνικά Δίκτυα Ασθενών	-Ποινικά Μητρώα -Χάρτες κατανομής εγκλημάτων -Εγκληματικές Οργανώσεις στο Διαδίκτυο -Ειδήσεις και πληροφορίες από τον παγκόσμιο ιστό -Στοιχεία από τρομοκρατικές Ενέργειες -Ιοί, κυβερνοεπιθέσεις, και botnets
Χαρακτηριστικά Δεδομένων	-Δομημένα web-based περιεχόμενο που δημιουργείται από πελάτες, πλούσιες πληροφορίες δικτύου, άτυπες γνωμοδοτήσεις πελατών	-Κλειστά πληροφοριακά συστήματα παλιάς τεχνολογίας -Πλουσιο περιεχόμενο κειμένου -Μη δομημένες και ανεπίσημες συζητήσεις πολιτών	-Υψηλή Απόδοση -Συλλογή δεδομένων βασισμένη σε συγκεκριμένο μέσο -Μεγάλης κλίμακας,πολλα πλως διαμοιρασμένα -Εξειδικευμένες μορφοποιήσεις	-Ανεξάρτητα καταχωρημένα αλλά με εξαιρετικά διασυνδεδεμένο περιεχόμενο -Περιεχόμενο εξιδεικευμένο σε συγκεκριμένο άτομο, -HIPPA -Θέματα Δεοντολογίας	-Προσωπικές πληροφορίες ταυτοποίησης, -Μη αυστηρό και παραπλανητικό περιεχόμενο -Πλούσια πληροφορία από ομάδες και διασυνδεδεμένα δίκτυα, -Πολύγλωσσο περιεχόμενο
Analytics	-Εξόρυξη Κανόνων Συσχέτισης -Κατηγοριοποίηση και Συσταδοποίηση Βάσης Δεδομένων -Εντοπισμός Ανωμαλιών -Εξόρυξη Γράφων	-Διασύνδεση Πληροφορίας -Ανάλυση περιεχομένου και κειμένου -Σημασιολογικές τεχνολογίες και περιεχόμενο	-Εξειδικευμένα μαθηματικά και αναλυτικά μοντέλα	-Γονιδιωματική, ανάλυση αλληλουχίας και οπτικοποίηση -Εξόρυξη συσχετίσεων απο ηλεκτρονικές	-Εντοπισμός συσχετίσεων και κατηγοριοποίηση των εγκλημάτων -Ανάλυση εγκληματικών δικτύων -Χωροχρονική

	-Ανάλυση Κοινωνικών Δικτύων -Analytics Κειμένου και Διαδικτύου -Ανάλυση Συναισθήματος και Κινήτρου	-Παρακολούθηση και ανάλυση κοινωνικών μέσων -Ανάλυση κοινωνικών δικτύων -Ανάλυση Συναισθήματος και Κινήτρου		εγγραφές υγείας και συσταδοποίηση -Οντολογίες Υγείας -Ανάλυση Παρενεργειών Φαρμάκων -Ιχνηλάτιση επαφών ασθενών -Αναλυση ιατρικού κειμένου -Εξόρυξη γνωσης με τήρηση της προστασίας των προσωπικών δεδομένων	ανάλυση και οπτικοποίηση -Αναλυση πολυγλωσσικού κειμένου -Ανάλυση κυβερνοεπιθέσεων
Επιπτώσεις	-Μάρκετινγκ “μακριάς ουράς” -Στοχοθετημένες και προσωποποιημένες καμπάνιες -Αύξηση της ικανοποίησης των πελατών	-Μετασχηματισμός των Κυβερνήσεων -Ενδυνάμωση Ρόλου των Πολιτών -Ενίσχυση της διαφάνειας,συμμετοχής και της ισότητας	-Αντίκτυπο στην επιστημονική έρευνα, πρόοδος	-Βελτιωμένη ποιότητα υπηρεσιών πρόνοιας -Βελτιωμένες μακροχρόνιες υπηρεσίες υγείας -Ενδηνάμωση της ενημέρωσης του ασθενούς	-Βελτίωση δημόσιας τάξης και ασφάλειας

1.6.3 Επιστήμη Δεδομένων | Data Science

[4. NIST]

Στην καθαρότερη μορφή της, η επιστήμη των δεδομένων θεωρείται το τέταρτο παράδειγμα της επιστήμης, ακολουθώντας τις πειραματικές, τη θεωρητικές και υπολογιστικές επιστήμες. Το τέταρτο πρότυπο είναι ένας όρος που εισήγαγε ο Δρ. Jim Gray το 2007. Η επιστήμη δηλαδή που είναι εξαρτημένη από τα δεδομένα, συντομευμένο υπό τον όρο στην επιστήμη των δεδομένων, αναφέρεται στη διεξαγωγή ανάλυσης στα δεδομένα ως εμπειρική επιστήμη, στην άμεση εκμάθηση δηλαδή από τα ίδια τα δεδομένα. Αυτό συνήθως λαμβάνει τη μορφή συλλογής δεδομένων, ακολουθούμενη από ανοιχτή ανάλυση χωρίς να έχουν οριστεί εκ των προτέρων οι υποθέσεις εργασίες (μερικές φορές αναφέρεται ως ανακάλυψη ή εξερεύνηση δεδομένων). Η δεύτερη εμπειρική μέθοδος αναφέρεται στην διατύπωση μιας υπόθεσης, τη συλλογή των δεδομένων - νέων ή προϋπάρχοντων – με σκοπό την διερεύνηση της υπόθεσης και την αναλυτική επιβεβαίωση ή άρνηση της (ή στην εξαγωγή του συμπεράσματος ότι υπάρχει ανάγκη για πρόσθετες πληροφορίες ή μελέτη). Και στις δύο μεθόδους, τα συμπεράσματα βασίζονται στα δεδομένα. Σε πολλά έργα επιστήμης δεδομένων, αρχικά γίνεται περιήγηση στα δεδομένα, η οποία διαμορφώνει μια υπόθεση, η οποία υπόθεση στη συνέχεια διερευνάται. Όπως σε κάθε πειραματική επιστήμη, το αποτέλεσμα θα μπορούσε να είναι ότι η ίδια η αρχική υπόθεση πρέπει να αναδιατυπωθεί. Η βασική ιδέα λοιπόν είναι ότι η επιστήμη των δεδομένων είναι μια εμπειρική επιστήμη, που πραγματοποιεί την παραδοσιακή επιστημονική διαδικασία απευθείας σε δεδομένα. Σημειώστε ότι η υπόθεση μπορεί να καθοδηγείται από μια ανάγκη, ή μπορεί να είναι η αναδιατύπωση μιας ανάγκης σε όρους τεχνικής υπόθεσης.

Ο ορισμός επομένως του όρου διατυπώνεται ως εξης:

Η επιστήμη των δεδομένων είναι η εξαγωγή χρήσιμης γνώσης απευθείας από τα δεδομένα μέσω μιας διαδικασίας ανακάλυψης ή εκπόνησης υποθέσεων και δοκιμασιών των υποθέσεων αυτων.

Η επιστήμη των δεδομένων είναι στενά συνδεδεμένη με την ανάλυση των Big Data και αναφέρεται στη διαχείριση και την εκτέλεση end-to-end διεργασιών δεδομένων, συμπεριλαμβανομένων των συμπεριφορών των στοιχείων του ίδιου συστήματος δεδομένων. Ως εκ τούτου, η επιστήμη των δεδομένων περιλαμβάνει όλα τα είδη αναλύσεων

(analytics), αλλά τα analytics ποτέ δεν περιλαμβάνουν όλη την επιστήμη των δεδομένων. Όπως αναφέρθηκε, η επιστήμη των δεδομένων περιέχει διαφορετικές προσεγγίσεις χρήσης των δεδομένων για την επίλυση των προβλημάτων. Ενώ ο όρος επιστήμη των δεδομένων μπορεί να γίνει αντιληπτός ως οι απαραίτητες δραστηριότητες σε κάθε ροή εργασιών αναλύσεων που παράγουν γνώση από δεδομένα, ο όρος χρησιμοποιείται πολύ συχνά μαζί με το οικοσύστημα των Big Data.

Η επιστήμη των δεδομένων έγινε ένας κοινώς χρησιμοποιούμενος όρος στα μέσα της δεκαετίας του 2000 ως νέες τεχνικές για το χειρισμό Big Data άρχισαν να αναδύονται. Αρχικά εφαρμόστηκε στο πλαίσιο των συστημάτων Big Data για την επεξεργασία πολύ μεγάλων συνόλων δεδομένων, όπου το μέγεθος των δεδομένων δηλαδή αποτελεί από μόνο του πρόβλημα. Αυτό η πρόσθετη πολυπλοκότητα απαιτούσε την χρησιμοποιήση δεξιοτήτων επιστήμης των υπολογιστών προκειμένου να κατανοήσουμε πώς μπορούμε να αναπτύξουμε τα σύνολα δεδομένων μεγάλου όγκου σε πολλαπλούς κόμβους δεδομένων (data nodes) και το πώς πρέπει να μεταλλαχθεί ο τρόπος αναζήτησης και ανάλυσης για τον χειρισμό κατανεμημένων δεδομένων.

Επομένως, η επιστήμη των δεδομένων είναι ένα υπερσύνολο τομέων της στατιστικής, της εξόρυξης δεδομένων και της μηχανικής μάθησης για να καταστεί εφικτή η ανάλυση μεγάλων δεδομένων.

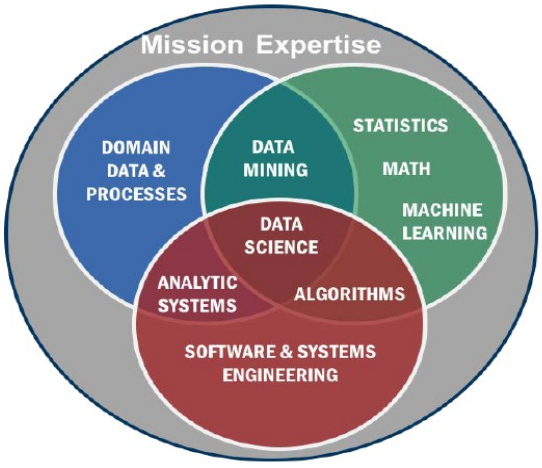
• Τόσο η στατιστική ανάλυση όσο και η ανάλυση για σκοπούς εξόρυξης δεδομένων απαιτούν προσεκτική δειγματοληψία των δεδομένων για να διασφαλιστεί ότι ο πληθυσμός των δεδομένων είναι αντιπροσωπευτικός. Στην επιστήμη των δεδομένων, συνήθως όλα τα δεδομένα επεξεργάζονται και αναλύονται μέσω τεχνικών κλιμάκωσης.

• Σε μερικά προβλήματα, που υποτίθεται ότι αφορούν τεράστιο όγκο δεδομένων, τα μικρά σφάλματα τείνουν να χάνουν την σημασία τους ,μειώνοντας ή εξαλείφοντας έτσι την ανάγκη καθαρισμού των δεδομένων.

• Έχωντας στην κατοχή μας μεγάλα συνόλα δεδομένων, πολύ συχνά οι απλούστεροι αλγόριθμοι μπορούν να επιφέρουν αποδεκτά αποτελέσματα. Αυτό έχει ανοίξει την συζήτηση για το αν σε ορισμένες περιπτώσεις απλώς τα περισσότερα δεδομένα είναι ανώτερα από τους βελτιστους αλγορίθμους.

• Πολλές υποθέσεις είναι δύσκολο να αναλυθούν, επομένως η επιστήμη των δεδομένων επικεντρώνεται επίσης στον καθορισμό ενός υποκατάστατου ερωτήματος που δεν εξετάζει την αρχική υπόθεση, αλλά του οποίου το αναλυτικό αποτέλεσμα μπορεί να εφαρμοστεί στην αρχική αποστολή.

• Ο πλούτος των πηγών δεδομένων αύξησε την ανάγκη διερεύνησης δεδομένων για να προσδιοριστεί το τι μπορεί να ενδιαφέρει προς ανάλυση.Σε αντίθεση με τη συλλογή δεδομένων που διενεργούν η στατιστική και η εξόρυξη δεδομένων, η ευρύτερη κατανόηση των δεδομένων οδηγεί είτε στην ανακάλυψη γνώσης είτε τη διατύπωση υποθέσεων για δοκιμές.



Εικόνα 5.Πεδία που συνθέτουν στο Data Science

Διαδικασία Επιστήμης Δεδομένων | Data Science

Η επιστήμη των δεδομένων επικεντρώνεται στον κύκλο ζωής επεξεργασίας δεδομένων από άκρο σε άκρο για να συμπεριλάβει τα μεγάλα δεδομένα.

> *Ο κύκλος ζωής της επιστήμης δεδομένων data science lifecycle είναι το σύνολο των διαδικασιών σε μια εφαρμογή που μετατρέπει τα δεδομένα σε χρήσιμα η γνώση.*

Ο κύκλος ζωής της επιστήμης δεδομένων από άκρο σε άκρο αποτελείται από πέντε βασικά βήματα:

1.Σύλληψη: συλλογή και αποθήκευση δεδομένων, συνήθως στην αρχική τους μορφή (δηλαδή, ακατέργαστα δεδομένα).

2.Προετοιμασία: διαδικασίες που μετατρέπουν τα ακατέργαστα δεδομένα σε καθαρισμένες, οργανωμένες πληροφορίες.

3.Ανάλυση: Τεχνικές που παράγουν συνδυαστική γνώση από τις οργανωμένες πληροφορίες.

4.Οπτικοποίηση: Παρουσίαση δεδομένων ή αναλυτικών αποτελεσμάτων κατά τρόπο που να επικοινωνεί με άλλους

5.Δράση: Δαδικασίες που χρησιμοποιούν τη συνδυαστική αυτή γνώση προκειμένου να προσδωθεί αξία στην επιχείρηση.

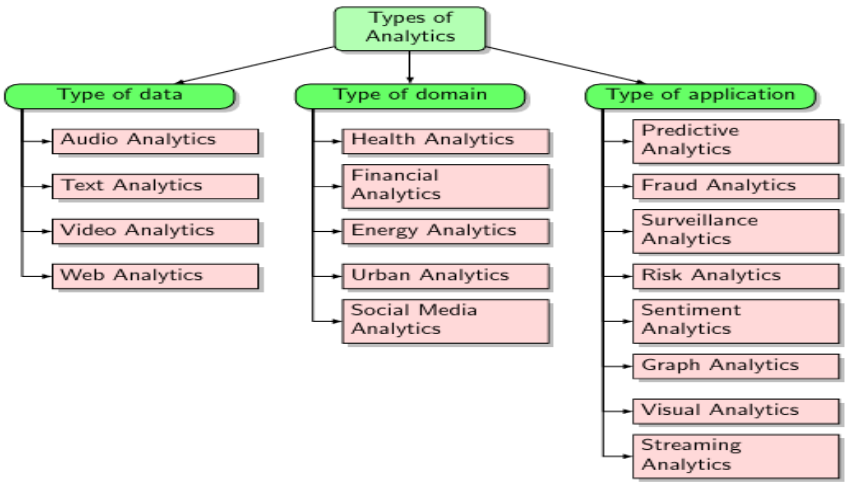
Analytics

Τα Analytics αναφέρονται σε ένα συγκεκριμένο βήμα του κύκλου ζωής των data analytics, ενώ η επιστήμη των δεδομένων περιλαμβάνει όλα τα βήματα αυτού του κύκλου ζωής συμπεριλαμβανομένης της επεξεργασίας Big Data. Ο κύκλος ζωής της επιστήμης των δεδομένων περιλαμβάνει πολλές ακόμη δραστηριότητες πέρα από αυτά τα πέντε θεμελιώδη βήματα, όπως την νομική/κανονιστική συμμόρφωση,τη διακυβέρνηση, τις λειτουργίες, την ασφάλεια των δεδομένων, τη διαχείριση βασικών δεδομένων(master data management), τη διαχείριση μετα-δεδομένων και τη συντήρηση / καταστροφή.

Οι αναλυτικές διεργασίες χαρακτηρίζονται συχνά ως **discovery** για την σχηματοποίηση της αρχικής διατύπωσης, **development** για την εγκαθίδρυση της διαδικασίας ανάλυσης για μια συγκεκριμένη υπόθεση και την **application** για την ενθυλάκωση της ανάλυσης σε σύστημα. Ενώ τα Big Data εμπλέκονται και στους τρεις τύπους αναλυτικών διαδικασιών, το μεγαλύτερο μέρος των αλλαγών παρατηρούνται στην ανάπτυξη και στις εφαρμοζόμενες αναλύσεις. Οι νέες τεχνολογίες στην μηχανική των Big Data αλλάζουν τους τύπους αναλύσεων που είναι εφικτές, αλλά αυτό δεν οδηγεί σε εντελώς νέους τύπους αναλύσεων. Ωστόσο, δεδομένης της ταχύτητας ανάκτησης, οι αναλυτές μπορούν να αλληλεπιδρούν με τα δεδομένα τους με τρόπους που προηγουμένως δεν ήταν δυνατοί. Οι παραδοσιακές στατιστικές αναλυτικές τεχνικές μειώνουν, δειγματοληπτούν ή συναθροίζουν τα δεδομένα πριν από την ανάλυση. Αυτό έγινε για να γίνει ανάλυση των μεγάλων συνόλων δεδομένων εφικτή σε υλικό που δεν θα μπορούσε να κλιμακωθεί. Τα Analytics επικεντρώνονται στην στατιστική και την εξόρυξη δεδομένων και στοχεύουν στην αιτιώδη συνάφεια,δηλαδή είναι σε θέση να περιγράψουν γιατί συμβαίνει κάτι. Ανακαλύπτοντας δηλαδή την βαθύτερη αιτία υποστηρίζει τους αποφασίζοντες στο να αλλάξουν μια τάση ή ένα αποτέλεσμα. Η επιστήμη των Big Data δίνει έμφαση στην αξία που πηγάζει από τους υπολογισμούς που εφαρμόζονται σε ολόκληρο το σύνολο δεδομένων. Ο προσδιορισμός των συσχετίσεων (και όχι κατ 'ανάγκη της αιτιώδους συνάφεια) μπορεί να είναι χρήσιμος όταν είναι γνωστές και η κατεύθυνση ή η τάση είναι αρκετές για να αναληφθούν σχετικές δράσεις.

Τύποι Analytics

[24. Rao , Mitra, Bhatt, Goswami]



Εικόνα 6.Τύποι Analytics

Τεχνικές & Εφαρμογές ανά είδος Analytics

[24. Rao , Mitra, Bhatt, Goswami]

Τύπος	Τεχνικές	Εφαρμογές
Κειμένου	-Διερευνητική Ανάλυση -Εξαγωγή Όρων -Διαχείριση Οντολογιών -Εξαγωγή Συνόψεων -Ερωταπαντήσεων -Εξαγωγής Άποψης -Επεξεργασίας Φυσικής Γλώσσας	-Blogs, ροές ενημέρωσης social media,emails, online συνόψεις ειδήσεων και κριτικών προιόντων από πελάτες
Ήχου	-Large Vocabulary Continuous Speech Recognition (LVCSR) -Προσεγγιση βασισμένη στην φωνή	-Χρήση τηλεφωνικών κέντρων για εξωτερική ανάθεση επιχειρηματικών διεργασιών -Οργανισμοί Υγείας
Εικόνας/Video	-Προσέγγιση βασισμένη στον εξυπηρετητη/server -Προσεγγιση βασισμένη στην τελική συσκευή	-Καταγραφές Κλειστών Κυκλωμάτων Τηλεόρασης -Κάμερες σε Εμπορικά Κέντρα -Ιστοθέσεις διαμοιρασμού περιεχομένου βίντεο χρηστών
Social Media	-Εντοπισμός σχέσεων τύπου γραμμικής συσχέτισης ή ανεξάρτητης αλληλεπικάλυψης -Εντοπισμός κεντρικών κόμβων(opinion leaders) και εγγύτητας	-Κοινωνικά δίκτυα, blogs,microblogs, -Ανάλυση εγκληματικών συμπεριφορών -Marketing -Επιδημιολογία -Εικονικοποίηση βασισμένη στην πραγματική εμπειρία του χρήστη
Προγνωστικά/Predictive	-Παλινδρόμιση -Μηχανική Μάθηση -Εξόρυξη Γνώσης	-Συσχετίσεις πελατών -Βελτιστοποίηση τιμής -Εντοπισμός απάτης στις ασφαλιστικές υπηρεσίες -Πρόγνωση και πρόληψη εγκλημάτων και τρομοκρατίας
Οικονομικά	-Ομαδοποίηση Πελατών -Ανάλυση Αισθημάτων -Συσταδοποίηση	-Μειώση Ρίσκου -Ανίχνευση Απάτης -Δείκτες Ικανοποίησης Πελάτη -Κατηγοριοποίηση Πελατειακής Βάσης
Υγείας	-Εξαγωγή Πληροφορίας -Προγνωστική Ανάλυση -Παλινδρόμιση	-Ανάκτηση βασισμένη στο περιστατικό -Συπτώματα και ενδείξεις καρδιακών επεισοδίων -Ομοιότητες μεταξύ ασθενών
Γράφων	-Εξαγωγή συχνά εμφανιζόμενων υπογράφων -Δείκτες κετρικότητας και ομοιότητας -Συσταδοποίηση και ευρετηρίαση γράφων -Επεξεργασία γράφων μεγάλης κλίμακας	-Εντοπισμός Ανωμαλιών -Γράφοι σε συνεχόμενα δεδομένα -Εξαγωγή ομάδων\κοινοτήτων σε κοινωνικά δίκτυα
Διαδικτύου	-Εξαγωγή Όρων -PageRank & Hits Αλγόριθμοι	-Συμπεριφορά Χρηστών κατά την Πλοήγηση

	-Εντοπισμός Συχνών Χρηστών -Αλυσίδες Markov	-Ανάλυση Γράφου Διαδικτύου

1.6.4 Big Data & Business Intelligence

[33. EDUCBA]

Παρακάτω είναι η λίστα των στοιχείων, που εξηγεί τις διαφορές μεταξύ του Business Intelligence και των Big Data

Μερικά καίρια σημεία

- Τόσο ο BI όσο και τα Big Data στοχεύουν στην υποστήριξη της επιχείρησης ωστε να λαμβάνει σωστές αποφάσεις αναλύοντας τα τεράστια σύνολα δεδομένων για να επεκτείνει την δραστηριότητα της και να βελτιστοποιήσει το κόστος.
- Αυτή η ανάλυση δεδομένων όχι μόνο καθιστά δυνατή τη λήψη αποφάσεων αλλά και συμμετέχει ενεργά στην ανάπτυξη στρατηγικών και μεθόδων που διασφαλίζουν την επιτυχία των οργανισμών. Αυτή η ανάλυση δεδομένων μπορεί να ονομαστεί "Business Intelligence", ενώ το "Big Data" είναι ένας σχετικά νέος όρος στο Business Intelligence.
- Από την εποχή της πρώτης ανάπτυξης του BI, οι όγκοι των συνόλων δεδομένων έχουν γίνει απίστευτα μεγάλοι, με το καλύτερο παράδειγμα που μπορούμε να εξετάσουμε είναι στα μέσα κοινωνικής δικτύωσης. Ως αποτέλεσμα, πρέπει να εφαρμοστούν περισσότερες προσπάθειες και στρατηγικές για να αντιμετωπιστούν και να γίνουν χρήσιμες στην επιτυχία των επιχειρήσεων.
- Το Business Intelligence βοηθά στην εύρεση των απαντήσεων στα επιχειρηματικά ερωτήματα που γνωρίζουμε, ενώ το Big Data μας βοηθά να ανακαλύψουμε ερωτήματα και τις απαντήσεις που δεν γνωρίζαμε πριν.
- Αν και το Business Intelligence και τα Big Data είναι δύο τεχνολογίες που χρησιμοποιούνται για την ανάλυση συνόλων δεδομένων για να βοηθήσουν τους οργανισμούς στη διαδικασία λήψης αποφάσεων, υπάρχουν διαφορές μεταξύ τους. Διαφέρουν στον τρόπο με τον οποίο αναλύουν τα δεδομένα.
- Η Business Intelligence βασίζεται στην αρχή του συνδυασμού όλων των συνόλων επιχειρηματικών δεδομένων σε έναν κεντρικό εξυπηρετητή, τα οποία θα αναλυθούν σε λειτουργία εκτός σύνδεσης, αφού αποθηκευτούν σε μια πλατφόρμα ή ένα περιβάλλον που ονομάζεται Data Warehouse. Τα σύνολα δεδομένων εκεί είναι δομημένα σε μια σχεσιακή βάση δεδομένων με πρόσθετα ευρετήρια και μορφές πρόσβασης στους πίνακες της αποθήκης.
- Αντιθέτως, στο περιβάλλον Big Data, τα δεδομένα αποθηκεύονται σε ένα κατανεμημένο σύστημα αρχείων (π.χ. HDFS) με σκοπό διανεμηθούν σε υπολογιστικούς κόμβους για ευκολότερη επεξεργασία. Το Distributed File System είναι πολύ πιο ασφαλές και ευέλικτο.
- Οι λύσεις BI μεταφέρουν τα δεδομένα στις λειτουργίες-διαδικασίες επεξεργασίας, ενώ οι λύσεις Big Data μεταφέρουν τις λειτουργίες επεξεργασίας στα σύνολα δεδομένων. Δεδομένου ότι η ανάλυση είναι εξειδικευμένη γύρω από τις πληροφορίες (Δεδομένα), είναι απλούστερος ο χειρισμός μεγαλυτέρων ποσοτήτων.
- Οι λύσεις BI αφορούν περισσότερο τα δομημένα δεδομένα, ενώ τα εργαλεία Big Data μπορούν να επεξεργάζονται και να αναλύουν δεδομένα σε διαφορετικές μορφές, τόσο δομημένες όσο και μη δομημένες.
- Οι λύσεις Big Data μπορούν να επεξεργαστούν ταυτόχρονα τα ιστορικά δεδομένα σε συνδυασμό με τα δεδομένα που προέρχονται από πηγές πραγματικού χρόνου, ενώ στο Business Intelligence επεξεργάζεται μόνο τα σύνολα ιστορικών δεδομένων.
- Η τεχνολογία Big Data χρησιμοποιεί έννοιες παράλληλης επεξεργασίας (Map Reduce), η οποία βελτιώνει την ταχύτητα ανάλυσης και επεξεργασίας των συνόλων δεδομένων με τη διανομή εργασιών σε διάφορες διαδικασίες παράλληλης εκτέλεσης, ενώ στο τέλος τα αποτελέσματα συνδυάζονται και εμφανίζονται, γεγονός που καθιστά ευκολότερη την ανάλυση των μεγάλων όγκων.

Πεδία Σύγκρισης	**Business Intelligence**	**Big Data**
Σκοπός	Ο σκοπός του Business Intelligence είναι να βοηθήσει την επιχείρηση να λάβει	Ο κύριος σκοπός τωβ Big Data είναι να συλλέγουν, να επεξεργάζονται και να

	καλύτερες αποφάσεις. Το Business Intelligence βοηθά στην παροχή αναφορών υψηλής ακρίβειας, εξάγοντας πληροφορίες απευθείας από την πηγή δεδομένων.	αναλύουν τα δεδομένα, τόσο δομημένα όσο και αδόμητα, για τη βελτίωση των αποτελεσμάτων των πελατών τους
Στοιχεία/ Οικοσύστημα	Operation Systems,ERP,Databases,Data Warehouse, Dashboard etc.	Hadoop, Spark, R Server, Hive, HDFS etc.
Εργαλεία	Παρακάτω είναι η λίστα με τα εργαλεία που χρησιμοποιούνται για επιχειρηματική ευφυΐα.Αυτά τα εργαλεία επιτρέπουν σε μια επιχείρηση να συγκεντρώνει, να αναλύει και να οπτικοποιεί δεδομένα, τα οποία μπορούν να χρησιμοποιηθούν για τη λήψη καλύτερων επιχειρηματικών αποφάσεων και για την εκπόνηση καλών στρατηγικών σχεδίων. • Tableau • Qlik Sense • Online analytical processing (OLAP) • Sisense • Data Warehousing • Digital Dashboards and Data mining • Microsoft Power BI • Google Analytics etc	Παρακάτω είναι η λίστα των εργαλείων που χρησιμοποιούνται στα Big Data. Αυτά τα εργαλεία ή πλαίσια αποθηκεύουν ένα μεγάλο όγκο δεδομένων και τα επεξεργάζονται για να αποκτήσουν στοιχεία από τα δεδομένα με σκοπό να βοηθήσουν στη λήψη καλών αποφάσεις για την επιχείρηση • Hadoop • Spark • Hive • Polybase • Presto • Cassandra • Plotly • Cloudera • Storm
Χαρακτηριστικά/ Ιδιότητες	Παρακάτω είναι τα έξι χαρακτηριστικά του Business Intelligence: Location Intelligence, Executive Dashboards, “what if” Analysis, Interactive Reports, Metadata Layer, and Ranking Reports	Τα μεγάλα δεδομένα μπορούν να περιγραφούν από ορισμένα χαρακτηριστικά όπως ο όγκος, η ποικιλία, η μεταβλητότητα και η ταχύτητα
Πλεονεκτημα	Παρακάτω είναι ο κατάλογος των πλεονεκτημάτων του Business Intelligence • Βοηθά στην λήψη καλύτερων επιχειρηματικών αποφάσεων • Ταχύτερες και ακριβέστερες αναφορές και ανάλυσεις • Βελτιωμένη ποιότητα δεδομένων • Μειωμένο κόστος • Αύξηση εσόδων • Βελτιωμένη λειτουργική αποδοτικότητα κ.λπ.	Παρακάτω είναι ο κατάλογος των πλεονεκτημάτων του Big Data • Καλύτερη Λήψη Αποφάσεων • Ανίχνευση Απάτης • Αποθήκευση, Εξόρυξη και Ανάλυση Δεδομένων • Αναλυση Αγοράς και Προβλέψεις • Βελτίωση των Υπηρεσίων • Βοηθά στην Εφαρμογή των νέων Στρατηγικών • Παρακολούθηση των τάσεων των πελατών • Εξοικονόμηση Κόστους • Καλύτερες πληροφορίες πωλήσεων, οι οποίες βοηθούν στην αύξηση των εσόδων κλπ
Πεδία Εφαρμογής	Ο τραπεζικός τομέας, ψυχαγωγία και μέσα μαζικής ενημέρωσης, υγειονομική περίθαλψη, λιανική και χονδρική πώληση κλπ	Μέσα μαζικής ενημέρωσης, υγειονομική περίθαλψη, βιομηχανία τυχερών παιχνιδιών, βιομηχανία τροφίμων κλπ

1.6.5 Εφαρμογή των Big Data σε Επιχειρηματικούς Κλάδους: Το παράδειγμα των Χρηματοπιστωτικών Υπηρεσίων

[11.ORACLE]

Οι εταιρείες στον τομέα της λιανικής τραπεζικής και των χρηματοπιστωτικών υπηρεσιών κατά κανόνα διατηρούν κεντρικές αποθήκες δεδομένων και εργαλεία επιχειρηματικής ευφυΐας για το reporting και την ανάλυση της συμπεριφοράς των πελατών ώστε να προβλέπουν καλύτερα τις ανάγκες και να βελτιστοποιούν τις λειτουργίες τους. Με την ανάπτυξη συστημάτων διαχείρισης μεγάλων δεδομένων που περιλαμβάνουν δεξαμενές δεδομένων-data lakes (πχ με Hadoop ή / και NoSQL), μπορούν να επιτευχθούν μεγαλύτερα οφέλη καθώς η επιχείρηση κερδίζει περισσότερες δυνατότητες πρόβλεψης και γίνεται πιο ευέλικτη. Η προσθήκη των συστημάτων Big Data επιτρέπει στους οργανισμούς αυτούς να αποκτούν γρηγορότερα insights και καθιστούν πιο αποτελεσματική τη λήψη αποφάσεων. Παρακάτω αναλύονται οι βασικότερες τάσεις οι οποίες οδηγούν τις επιχειρήσεις του συγκεκριμένου κλάδου στην υιοθέτηση λύσεων Big Data:

1. Μεγαλύτερα σύνολα δεδομένων της αγοράς που θα περιέχουν ιστορικά δεδομένα σε μεγαλύτερες χρονικές περιόδους και αυξημένη λεπτομέρεια τα οιποία και απαιτούνται για την τροφοδοσία προγνωστικών μοντέλων, προβλέψεων και εμπορικών επιπτώσεων κατά τη διάρκεια της ημέρας.

2. Οι νέες κανονιστικές απαιτήσεις και απαιτήσεις συμμόρφωσης δίνουν μεγαλύτερη έμφαση στη διακυβέρνηση και την υποβολή εκθέσεων σχετικά με τον κίνδυνο, δημιουργώντας την ανάγκη για βαθύτερη και πιο διαφανή αναλύση σε παγκόσμιους οργανισμούς.

3. Τα χρηματοπιστωτικά ιδρύματα επιταχύνουν τα επιχειρηματικά τους πλαίσια διαχείρισης κινδύνων, που βασίζονται σε στρατηγικές διαχείρισης δεδομένων για τη βελτίωση της διαφάνειας των επιχειρήσεων, την ικανότητα ελέγχου και την εκτελεστική εποπτεία του κινδύνου.

4. Οι εταιρείες χρηματοπιστωτικών υπηρεσιών προσπαθούν να εκμεταλλευτούν μεγάλες ποσότητες δεδομένων καταναλωτών από πολλαπλά κανάλια παροχής υπηρεσιών (υποκατάστημα, διαδίκτυο, κινητά) για την υποστήριξη νέων προγνωστικών μοντέλων ανάλυσης και ανακάλυψη της συμπεριφοράς των καταναλωτών καθώς και πρότυπα ώστε να αυξήσουν τα κέρδη προς τα επιθυμητά ποσοστά.

5. Στις μετα-αναδυόμενες αγορές όπως η Βραζιλία, η Κίνα και η Ινδία, οικονομικές και επιχειρηματικές ευκαιρίες ανάπτυξης ξεπερνούν την Ευρώπη και την Αμερική ως σημαντικές επενδύσεις που πραγματοποιούνται τοπικά σε υποδομές δεδομένων και σε cloud.

6. Η πρόοδος στα πλαίσια αποθήκευσης και επεξεργασίας Big Data θα βοηθήσουν οικονομικά τις επιχειρήσεις παροχής υπηρεσιών να ξεκλειδώνουν την αξία των δεδομένων που βρίσκονται στα επιμέρους τμήματα λειτουργίας τους προκειμένου να μειωθεί το κόστος της επιχειρηματικής δραστηριότητας και να ανακαλυφθούν νέες ευκαιρίες για arbitrage.

7. Σε μεγάλο πλήθος κεντρικών συστημάτων αποθήκευσης δεδομένων θα απαιτηθεί οι παραδοσιακές ETL διαδικασίες να επανασχεδιαστούν με πλαίσια Big Data ως απόκριση στην ανάπτυξη του όγκου πληροφοριών.

8. Προγνωστικά μοντέλα πιστωτικού κινδύνου που χρησιμοποιούν μεγάλες ποσότητες δεδομένων που αποτελούνται από ιστορική συμπεριφορά πληρωμών που υιοθετείται σε καταναλωτικές και εμπορικές συναλλαγές απαιτούν πρακτικές συλλογής για να βοηθήσουν στον προσδιορισμό της τάσης για παραβατικότητα ή πληρωμή.

9. Κινητές εφαρμογές και συσκευές συνδεδεμένες στο διαδίκτυο, όπως tablet και smartphone δημιουργούν μεγαλύτερη πίεση στην ικανότητα των τεχνολογικών υποδομών και των δίκτυων για κατανάλωση, ευρετηρίαση και ενσωμάτωση δομημένων και αδόμητων δεδομένων από μια ποικιλία πηγών.

10. Οι μεγάλες πρωτοβουλίες για δεδομένα οδηγούν σε αυξημένη ζήτηση αλγορίθμων για την επεξεργασία δεδομένων, καθώς και την έμφαση στις προκλήσεις όσον αφορά την ασφάλεια των δεδομένων και τον έλεγχο πρόσβασης καθώς και την ελαχιστοποίηση των επιπτώσεων στα υπάρχοντα συστήματα.

Πλατφόρμα Επιχειρησιακής Μοντελοποίησης

Οι οργανισμοί χρηματοπιστωτικών υπηρεσιών επανεξετάζουν τον τρόπο με τον οποίο μοντελοποιούν τις επιχειρήσεις τους. Αυτή η νέα θεώρηση κινητοποιείται εν μέρει και από τις νέες κανονιστικές απαιτήσεις. Επιπλέον, τα χρηματοπιστωτικά ιδρύματα ενσωματώνουν όλο και περισσότερο την αναλυτική δεδομένων στις διαδικασίες λήψης αποφάσεων. Η στατιστική μοντελοποίηση πλεόν αναλαμβάνει έναν ευρύτερο ρόλο μέσα στην επιχείρηση, καθώς τα ιδρύματα τείνουν να χρησιμοποιούν μοντέλα πρόβλεψεων και βελτιστοποίησης σε όλο και μεγαλύτερο φασμα επιχειρήσιακών λειτουργιών.

Με την αύξηση της υιοθέτησης όμως δημιουργούνται νέες προκλήσεις. Καθώς η παραγωγή των μοντέλων γίνεται μέρος των κανονιστικών και άλλων διαδικασιών επιχειρησιακής νοημοσύνης, η διαχείριση του επιχειρησιακού μοντέλου (παρόμοια με τη διαχείριση των επιχειρηματικών δεδομένων) πρέπει να αποτελέσει προτεραιότητα. Όταν η μοντελοποίηση γίνεται πιο διαδεδομένη, τα μοντέλα αναπτύσσονται συχνά μέσω κεντρικά διαχειριζόμενων πλατφόρμων IT που ευθυγραμμίζονται με τις γραμμές των επιμέρους επιχειρηματικών μονάδων. Ωστόσο, μπορεί να υπάρχει ένα χάσμα μεταξύ των θεωρήσεων της μοντελοποίησης και της τεχνολογίας της πληροφορίας.

Οι πλατφόρμες μοντελοποίησης περιέχουν συχνά αντίγραφα επιχειρησιακών δεδομένων. Ενώ μια τράπεζα μπορεί να έχει εφαρμόσει πολύπλοκες πολιτικές διακυβέρνησης δεδομένων γύρω από δεδομένα στην αποθήκη δεδομένων της επιχείρησης, συχνά χρησιμοποιούνται και άλλα που συχνά δεν εμπίπτουν στην αρμοδιότητα αυτών των συστημάτων διακυβέρνησης. Το πρόβλημα επιδεινώνεται από τις νέες πηγές δεδομένων που οι επαγγελματίες της μοντελοποιήσης (modelers) επιθυμούν να αποκτήσουν πρόσβαση. Η συχνά επαναλαμβανόμενη φράση ότι "το πρόβλημα της ανάλυσης είναι ένα πρόβλημα δεδομένων" υπογραμμίζει την ανάγκη στενής σύνδεσης των analytics και της διαχείρισης δεδομένων(data management). Ωστόσο, ενώ οι τράπεζες έχουν επενδύσει πόρους σε προγράμματα διαχείρισης και διακυβέρνησης δεδομένων, η επιχειρησιακής κλίμακας μοντελοποίηση δεν φαίνεται να έχει προσελκύσει το ίδιο επίπεδο προσοχής. Ακριβώς όπως οι κανονιστικές απαιτήσεις διαμορφώνουν την προσέγγιση διαχείρισης των δεδομένων των χρηματοπιστωτικών ιδρυμάτων, φαίνεται ότι στο άμεσο μέλλον τα αιτήματα των ρυθμιστικών αρχών για τη διαχείριση επιχειρησιακών μοντέλων θα είναι πολύ παρόμοιες.

Διαχείριση Κινδύνων και Κεφαλαίου

Οι παραδοσιακές επιχειρησιακές αρχιτεκτονικές εξυπηρετούσαν καλά τις τράπεζες και τις εταιρείες παροχής χρηματοοικονομικών υπηρεσιών εδώ και χρόνια. Οι αρχιτεκτονικές αυτές επέτρεπαν στα εν λόγω ιδρύματα να διαχειριστούν την πίστωση, τη ρευστότητα της αγοράς και τον λειτουργικό κίνδυνο. Επιπλέον, επέτρεψαν στα θεσμικά όργανα να διαχειρίζονται τα κεφάλαια τους και να πληρούν τις κανονιστικές απαιτήσεις της Βασιλείας.

Η βαθμολόγηση του πιστοληπτικού κινδύνου και της συμπεριφοράς για την ταξινόμηση νέων ή υφιστάμενων πελατών ως προς την πιστοληπτική ικανότητα απαιτούσε σημαντική ανάλυση των δεδομένων των αιτήσεων δανείων και των δεδομένων από γραφεία πιστωτικών αξιολογήσεων και εμπειρογνώμονες πιστώσεων. Δεδομένης της ώθησης των τραπεζών σε μικροπιστώσεις και της επέκτασης στις αναδυόμενες αγορές, η έλλειψη διαθέσιμων πιστωτικών στοιχείων είναι εξαιρετικά προβληματική. Αυτή η έλλειψη μπορεί να ξεπεραστεί μέσω της πρότυπης μοντελοποίησης χρησιμοποιώντας μη παραδοσιακές εισροές από ομότιμες ομάδες, τα δεδομένα πληρωμών P2P από κινητές συσκευές, την κατανάλωση, τα δεδομένα αγοράς προπληρωμένων κινητών υπηρεσιών επικοινωνίας και άλλες πηγές.

Η αποτίμηση σύνθετων και μη ρευστοποιήσιμων επενδυτικών στοιχείων (instruments) και χαρτοφυλακίων απαιτεί προσομοίωση χιλιάδων παραγόντων κινδύνου με χρήση στοχαστικών μοντέλων. Οι προσομοιώσεις Monte Carlo χρησιμοποιούνται συχνά. Καθώς όμως ο αριθμός των παραγόντων κινδύνου αυξάνεται, η απαιτούμενη υπολογιστική ισχύς για αυτές οι προσομοιώσεις αυξάνεται (εκθετικά με τον αριθμό των παραγόντων κινδύνου). Τα υπολογιστικά δίκτυα (computer grids) ή άλλες δαπανηρές και εξειδικευμένες λύσεις χρησιμοποιήθηκαν προηγουμένως για τέτοιες προσομοιώσεις, αλλά οι προσομοιώσεις αυτές συχνά σταμάτησαν μετά από ένα χρονικό διάστημα λόγω του κόστους τους. Οι τεχνολογίες Big Data επιτρέπουν στις προσομοιώσεις αυτές να ολοκληρωθούν με χαμηλό κόστος, εισάγοντας κοινόχρηστο υλικό και σε μεγάλο βαθμό λογισμικό ανοιχτού κώδικα. Οι τραπεζικοί αναλυτές κατασκευάζουν μοντέλα δεδομένων τα οποία μπορούν να προβλέψουν τα δάνεια που ενδέχεται να γίνουν παραβατικά προκειμένου να ξεκινήσουν προληπτικές ενέργειες. Οι μεγάλες τράπεζες στις Ηνωμένες Πολιτείες προβαίνουν σε μεταφορά των δικαιωμάτων εξυπηρέτησης σε οργανισμούς διαχείρισης απαιτήσεων για συγκεκριμένα δάνεια που βρίσκονται σε κατάσταση αθέτησης υποχρεώσεων. Ο προσδιορισμός των δανείων που πρόκειται να μεταφερθούν σε αυτούς τους οργανισμούς απαιτεί λεπτομερή ανάλυση διαφόρων στοιχείων δεδομένων, όπως είναι το ιστορικό του δανείου, ο δανειολήπτης και τα έγγραφα δανείου. Οι τράπεζες χρησιμοποιούν μερικές φορές αναλυτικές προβλέψεις για να δημιουργήσουν χάρτες που προσδιορίζουν περιοχές που είναι γνωστές για απάτες ακινήτων, παρουσιάζοντας λεπτομέρειες σε επίπεδο ταχυδρομικού κώδικα και ενδεχομένως σε ατομικό επίπεδο(ιδιοκτήτη). Αυτό καθιστά δυνατή την αποτελεσματική ανάλυση όταν διεξάγεται η διαδικασία αξιολόγησης για μια νέα αίτηση δανείου και μπορεί να βοηθήσει στην αποτροπή πιθανής υποτίμησης ακινήτων, κατοχής και απάτης βραγχυχρόνιων μεταπώλησεων.

Οι τεχνολογίες Big Data μπορούν να βοηθήσουν στην ταυτοποίηση των προτύπων δαπανών των νοικοκυριών δημιουργώντας μια πλουσιότερη εικόνα των πελατών. Οι τράπεζες μπορούν να αναλύσουν τα αρχεία καταγραφής συναλλαγών για όλα τα προϊόντα τους και να προσδιορίσουν τις τάσεις των δαπανών σε επίπεδο νοικοκυριού. Αυτό παρέχει μια καλύτερη εικόνα της πραγματικής ικανότητας του πελάτη να επιστρέφει τα δάνεια και βοηθά επίσης να εντοπίσει τις μελλοντικές ευκαιρίες cross-selling και up-selling.

Διαχείριση Περιουσίας

Καθώς τα χρηματοπιστωτικά ιδρύματα επιδιώκουν να επεκταθούν σε νέες γεωγραφικές περιοχές και να παρέχουν διαφοροποιημένες υπηρεσίες για την ενίσχυση ενός χαρτοφυλακίου διαχείρισης περιουσιακών στοιχείων, χρειάζονται μια ολοκληρωμένη εικόνα των πελατών, των νοικοκυριών και των δικτύων πελατών και πρέπει να αναπτύξουν εξατομικευμένες λύσεις που να υποστηρίζουν την ανάπτυξη αυτή.

Οι διαχειριστές χαρτοφυλακίων λαμβάνουν συνήθως ειδοποιήσεις ειδήσεων σχετικά με τις εταιρείες που κατέχουν. Προκειμένου να ανταποκριθεί, ο διαχειριστής χαρτοφυλακίου πρέπει στη συνέχεια να πλοηγηθεί εν μέσω τεράστιων ποσοτήτων δεδομένων για να καθορίσει εάν δικαιολογείται μια μετατόπιση των χορηγήσεων. Θα επιθυμούσαν να μπορούν να:

-Ανιχνεύσουν αλλαγές στις οικονομικές συνομιλίες για τις εταιρείες του χαρτοφυλακίου

-Διερευνήσουν τις συνομιλίες με σκοπό τον προσδιορισμό της πηγής πληροφοριών

-Συγκρίνουν τις κοινωνικές πληροφορίες με τις εσωτερικές πληροφορίες

-Εντοπίσουν όλα τα κεφάλαια που περιέχουν τις εταιρείες και επίσης να προσδιορίσει τις χορηγήσεις.

Αυτοί οι τύποι αναλυτικών δυνατοτήτων μπορούν να αποκτηθούν γρήγορα μέσω Big Data και σχετικών λύσεων.

Μέσω της βελτίωσης της γνώσης για τους πελάτες οι τράπεζες και οι εταιρείες παροχής χρηματοοικονομικών υπηρεσιών επιδιώκουν να διαφοροποιηθούν αναπτύσσοντας και παρέχοντας μοναδικά προϊόντα και υπηρεσίες στους πελάτες τους. Ωστόσο, σε αυτήν την πολύ ανταγωνιστική βιομηχανία, τα επιτυχημένα προϊόντα συχνά αντιγράφονται και οι ενδοιασμοί του πελάτη στην αναζήτηση εναλλακτικών είναι πολύ χαμηλοί. Πριν από δέκα χρόνια, ένα άτομο ήταν πιο πιθανό να έχει μια μακροχρόνια σχέση με μια τράπεζα, επιτρέποντας στην τράπεζα να υπαγορεύει τους όρους για τους τρεχούμενους λογαριασμούς, τους λογαριασμούς ταμιευτηρίου και τα ενυπόθηκα δάνεια. Το πρόσωπο αυτό μπορεί να είχε άλλη σχέση με μια μεσιτεία έκπτωσεων, όπου ο μεσίτης θα είχε ελέγξει τη διάρθρωση των τελών, τις απαιτήσεις περιθωρίου και τα ποσοστά CD. Οι προμηθευτές βρισκόταν στο επίκεντρο αυτών των σχέσεων. Σήμερα, το άτομο αυτό έχει πιθανώς πολλαπλές παροδικές σχέσεις με πολλές τράπεζες, συμπεριλαμβανομένου ενός λογαριασμού σε μια τράπεζα που δεν χρεώνει τέλη, λογαριασμών σε τράπεζες που προσφέρουν το υψηλότερο επιτόκιο αποταμίευσης και στεγαστικών δανείων από τράπεζες που προσφέρουν το χαμηλότερο επιτόκιο στεγαστικών δανείων. Ο πελάτης είναι τώρα το κέντρο της προσοχής, με τα χρηματοπιστωτικά ιδρύματα να είναι παροδικοί κόμβοι διελευσής του. Η αύξηση της πλεονεκτικής θέσης του πελάτη αναγνωρίζεται από τους νεοεισερχόμενους στην αγορά - ιδρύματα που έχουν χτιστεί γύρω από ένα πλήρες ψηφιακό αποτύπωμα. Αυτά οι οργανισμοί αύξησαν τις προσδοκίες των μεμονωμένων καταναλωτών. Ο καταναλωτής τώρα αναμένει να έχει απαντήσεις σε πραγματικό χρόνο στα ερωτήματά του και θεωρεί δεδομένη τη δυνατότητα να έχει αυτή τη συζήτηση δημόσια. Ο καταναλωτής πλεόν αναμένει πλήρη διαφάνεια σχετικά με τα προϊόντα και τις υπηρεσίες που προσφέρονται. Για τις τράπεζες και τις εταιρείες χρηματοπιστωτικών υπηρεσιών το να κρατούν τους πελάτες μακροπρόθεσμα, σημαίνει το να τους πλησιάσουν περισσότερο. Πρέπει να προβλέπουν τις ανάγκες τους και να είναι σε θέση να τοποθετούν τα προϊόντα τους πριν ακόμα η ανάγκη εδραιωθεί στους πελάτες. Αν αποτύχουν σε αυτό, οι πελάτες θα επιλέξουν να πάρουν χρηματοπιστωτικές υπηρεσίες από αλλού. Με την πάροδο του χρόνου, οι πελάτες μπορούν να εγκαταλείψουν εξ ολοκλήρου το ίδρυμα.

Το Ίντερνετ των Πραγμάτων (IoT) εισάγει νέες επιλογές καταναλωτών που συνδέουν πελάτες με άλλους παρόχους υπηρεσιών όπως οι λιανοπωλητές, οι αεροπορικές εταιρείες και τα ξενοδοχεία. Οι εταιρείες χρηματοπιστωτικών υπηρεσιών αναπτύσσουν εταιρικές σχέσεις με πολλά από αυτά για να επεκτείνουν την εμβέλειά τους και να ενσωματώνουν τα προϊόντα τους σε όλους τους τομείς της ζωής των πελατών τους. Η δημιουργία μιας απρόσκοπτης, συνεπούς εμπειρίας σε πολλαπλά κανάλια μπορεί να οδηγήσει σε άριστη εμπειρία πελατών και να οδηγήσει σε αυξημένες ευκαιρίες εσόδων.

Βελτίωση Ανίχνευσης Απάτης

Η ανίχνευση και η πρόληψη της απάτης διευκολύνεται από την ανάλυση δεδομένων συναλλαγών και το φιλτραρισμα εισερχόμενου ρεύματος συναλλαγών σε πραγματικό χρόνο έναντι ενός πολύ γνωστού συνόλου προτύπων. Οι τεχνολογίες Big Data επιτρέπουν τη συσχέτιση δεδομένων από πολλαπλές πηγές ή περιστατικά για τον προσδιορισμό της απάτης. Τα ατομικά περιστατικά από μόνα τους μπορεί να μην σηματοδοτούν ένα δόλιο γεγονός. Ένα παράδειγμα ύποπτης δραστηριότητας μπορεί να συμβεί όταν ένας έμπορος στέλνει με συνέπεια ένα μήνυμα ηλεκτρονικού ταχυδρομείου ή καλεί έναν αριθμό τηλεφώνου μέσα σε λίγα λεπτά από την πραγματοποίηση μιας μεγάλης εμπορικής συμφωνίας. Η προσθήκη νέων δεδομένων, όπως είναι τα δεδομένα γεωγραφικής θέσης, μπορεί να ενισχύσει την πρόληψη της απάτης - μια κάρτα ATM που χρησιμοποιείται στη Νέα Υόρκη για να αποσύρει μετρητά ενώ η κινητή συσκευή του πελάτη είναι ενεργή στο Λονδίνο αποτελεί ένδειξη πιθανής δόλιας κατάληψης. Οι σύγχρονοι αυτο-προσαρμοστικοί αλγόριθμοι εκμάθησης μηχανών μπορούν να

μάθουν και να παρακολουθήσουν τις συμπεριφορές των πελατών και των συσκευών που καθιστούν δυνατή την έγκαιρη αναγνώριση της απάτης

2.Big Data Engineering | Μηχανική των Μεγάλων Δεδομένων

Το Big Data Public Working Group του NIST (National Institute of Standards & Technology) υιοθέτησε τον ακόλουθο ορισμό της μηχανικής Big Data:

Το Big Data Engineering περιλαμβάνει προηγμένες τεχνικές που αξιοποιούν ανεξάρτητους πόρους για την οικοδόμηση κλιμακούμενων συστημάτων δεδομένων.

Η μηχανική μεγάλων δεδομένων χρησιμοποιείται όταν τα χαρακτηριστικά όγκου, ταχύτητας, ποικιλίας ή μεταβλητότητας των δεδομένων (δλδ τα 4Vs των Big Data) απαιτούν κλιμάκωση για άυξηση της αποδοτικότητας επεξεργασίας ή μείωσης του κόστους. Οι νέες τεχνικές σε επίπεδο αποθήκευσης καθοδηγούνται από την αυξανόμενη ανάκυψη συνόλων δεδομένων που δεν μπορούν να αντιμετωπιστούν αποτελεσματικά με ένα παραδοσιακό σχεσιακό μοντέλο (π.χ. τα μη δομημένα κείμενα και βίντεο). Η ανάγκη για κλιμακούμενη πρόσβαση σε δομημένα δεδομένα οδήγησε, για παράδειγμα, σε λύσεις λογισμικού που βασίζονται στο πρότυπο αποθήκευσης ζεύγαριών κλειδιού-τιμής (key-value pair). Η αύξηση της σπουδαιότητας της ανάλυσης εγγράφων δημιούργήσε το μοντέλο βάσης δεδομένων βασισμένης σε έγγραφα (document-oriented database) και η αυξανόμενη σημασία των σχέσεων εντός των δεδομένων έχει οδηγήσει σε βελτίωση της αποτελεσματικότητας διαμέσου της χρήση της αποθήκευσης δεδομένων με την μορφή γράφων (graph-oriented data storage).

2.1 Απαιτήσεις Συστημάτων Μεγάλων Δεδομένων -Big Data Requierements-

[7. NIST]

ΑΠΑΙΤΗΣΕΙΣ ΠΗΓΗΣ ΔΕΔΟΜΕΝΩΝ (Data Source Requierements - DSR)

• DSR-1: Αξιόπιστη, σε πραγματικό χρόνο, ασύγχρονη, συνεχούς ροής και παρτίδων(batch) επεξεργασία για τη συλλογή δεδομένων από κεντρικοποιημένες, κατανεμημένες πηγές η cloud πηγές δεδομένων, αισθητήρες ή όργανα καταμέτρησης
• DSR-2: Αργή, κατά ριπές(batch) ή υψηλής απόδοσης μετάδοση δεδομένων μεταξύ των πηγών και των υπολογιστικών συστοιχιών
• DSR-3: Διαφοροποιημένο περιεχόμενο των δεδομένων που κυμαίνεται από δομημένα και αδόμητα κείμενα, έγγραφα, γράφους, ιστοσελίδες, χωροταξικά, συμπιεσμένα, χρονομαρκαρισμένα, χωρικά, πολυμέσων, προσομοίωσης και αισθητήρων (πχ. από συστημάτα διαχείρισης και παρακολούθησης)

ΑΠΑΙΤΗΣΕΙΣ ΠΑΡΟΧΟΥ ΜΕΤΑΤΡΟΠΗΣ (Transformation Provider Requierements TPR)

• TPR-1: Διαφοροποιημένες τεχνικές υπολογιστικής πολυπλοκότητας, στατιστικής,ανάλυσης γράφων και μηχανικής μάθησης
• TPR-2: Επεξεργασία σε επίπεδο παρτίδας(batch) και πραγματικου χρόνου(stream)
• TPR-3: Επεξεργασία και μοντελοποιήση δεδομένων πολύ διαφοροποιημένου περιεχομένου
• TPR-4: Επεξεργασία δεδομένων σε κίνηση (π.χ. ροής, λήψης ανανεώσιμου περιεχομένου, παρακολούθησης, ιχνηλασιμότητας, διαχείριση αλλαγών δεδομένων και οριοθέτηση δεδομένων)

ΑΠΑΙΤΗΣΕΙΣ ΠΑΡΟΧΟΥ ΔΥΝΑΤΟΤΗΤΩΝ (Capability Provider Requierements-CPR)

• CPR-1: Παλαιό λογισμικό και συγχρονες προηγμένες λίστες λογισμικού
• CPR-2: Παλαιότερες και σύγχρονες υπολογιστικές πλατφόρμες
• CPR-3: Παλαιότερες και προηγμένες συστοιχίες κατανεμημένων υπολογιστών, συν-επεξεργαστές, επεξεργασία εισόδου/εξόδου (I/O)
• CPR-4: Προηγμένα δίκτυα (π.χ. Εικονικά δίκτυα καθορισμένα μέσω λογισμικού SDN-Software Defined Network) και ευμετάβλητη μετάδοση δεδομένων, συμπεριλαμβανομένων των δικτύων οπτικών ινών, καλωδίων και ασύρματων δικτύων (π.χ. τοπικό δίκτυο, δίκτυο ευρείας περιοχής, δίκτυο μητροπολιτικής περιοχής, Wi-Fi)
• CPR-5: Παλαιότερη, μεγάλη, εικονικοποιημένη και προηγμένη κατανεμημένη αποθήκευση δεδομένων
• CPR-6: Παλαιότερα και σύγχρονα προγραμματιστικά εκτελέσιμα, εφαρμογές, εργαλεία, βοηθητικά προγράμματα και βιβλιοθήκες

ΑΠΑΙΤΗΣΕΙΣ ΚΑΤΑΝΑΛΩΤΩΝ ΔΕΔΟΜΕΝΩΝ (Data Consumer Requierements DCR)

• DCR-1: Γρήγορες αναζητήσεις σε επεξεργασμένα δεδομένα με υψηλή σχετικότητα, ακρίβεια και ανάκληση
• DCR-2: Διαφοροποιημένες μορφές αρχείων εξόδου για απεικόνισεις, ερμηνειες και reporting
• DCR-3: Οπτικοποιημένες διατάξεις παρουσίασης αποτελεσμάτων
• DCR-4: Πλούσιο περιβάλλον χρήστη για πρόσβαση μέσω προγραμμάτων περιήγησης και εργαλείων οπτικοποίησης
• DCR-5: Πολυδιάστατο επίπεδο απεικόνισης δεδομένων υψηλής λεπτομέριας
• DCR-6: Τα αποτελέσματα συνεχούς ροής στους πελάτες

ΑΠΑΙΤΗΣΕΙΣ ΑΣΦΑΛΕΙΑΣ ΚΑΙ ΙΔΙΩΤΙΚΟΤΗΤΑΣ (Security & Privacy Requierements)

• SPR-1: Προστασία και διατήρηση της ασφάλεια και του απόρρητου των ευαίσθητων δεδομένων.
• SPR-2: Υποστήριξη ελέγχου πρόσβασης πολυσυμμετοχική και πολυεπίπεδη, αυθεντικοποίηση κατευθυνόμενη μέσω πολτικών στα προστατευόμενα δεδομένα και διασφάλιση ότι αυτά είναι εναρμονισμένα με τις αποδεκτές βέλτιστες πρακτικές διακυβέρνησης, αντιμετώπισης κίνδυνων,κανονιστικής συμμόρφωσης, εμπιστευτικότητας, ακεραιότητας και τη διαθεσιμότητα.

ΑΠΑΙΤΗΣΕΙΣ ΔΙΑΧΕΙΡΙΣΗΣ ΤΟΥ ΚΥΚΛΟΥ ΖΩΗΣ (Lifecycle Management Requierements)

• LMR-1: Επιδιόρθωση ποιότητας δεδομένων, συμπεριλαμβανομένης της προεπεξεργασίας, της ομαδοποίησης δεδομένων, της ταξινόμησης, της απομείωσης και της μορφοποίησης
• LMR-2: Δυναμικές ενημερώσεις για δεδομένα, τα προφίλ χρηστών και τους συνδέσμους
• LMR-3: Κύκλος ζωής δεδομένων και μακροπρόθεσμη πολιτική διατήρησης, συμπεριλαμβανομένης της καταγραφής προέλευσης των δεδομένων
• LMR-4: Επικύρωση δεδομένων
• LMR-5: Ανθρώπινο σχολιασμό και παρέμβαση κατά την επικύρωση δεδομένων
• LMR-6: Πρόληψη της ολοκληρωτικής απώλειας ή μερικής φθοράς των δεδομένων
• LMR-7: Πολλαπλώς διαμοιρασμένα αντίγραφα ασφαλείας (συμπεριλαμβανομένων των διασυνοριακών, γεωγραφικά διασκορπισμένων)
• LMR-8: Μόνιμη ταυτοποίηση και ιχνηλασιμότητα δεδομένων
• LMR-9: Τυποποίηση, συνδυασμός και ομαλοποίηση δεδομένων από διαφορετικές πηγές

ΑΛΛΕΣ ΑΠΑΙΤΗΣΕΙΣ

• OR-1: Πλούσιο περιβάλλον χρήστη από κινητές πλατφόρμες για πρόσβαση στα επεξεργασμένα αποτελέσματα
• OR-2: Παρακολούθηση επιδόσεων της αναλυτικής επεξεργασία από κινητές πλατφόρμες
• OR-3: Πλούσια αναζήτηση οπτικού περιεχομένου και δρομολόγηση προς κινητές πλατφόρμες
• OR-4: Ανάκτηση και διαχείριση δεδομένων κινητών συσκευών
• OR-5: Ασφάλεια σε κινητές συσκευές και άλλες έξυπνες συσκευές, όπως αισθητήρες

2.2 Έννοιες & Πλαίσια του Υλικού των Υποδομών -Infastructure Frameworks-

[7. NIST]

2.2.1 Hypervisors

Η εικονικοποίηση χρησιμοποιείται συχνά για την επίτευξη ελαστικότητας και ευελιξίας στην κατανομή των φυσικών πόρων και συχνά αναφέρεται ως Υποδομή Ως Υπηρεσία (Infastructure As a Service-IaaS) σε όρους cloud computing.Η εικονικοποίηση υλοποιείται μέσω των hypervisors που στην αρχιτεκτονική Big Data συνήθως υπάρχουν στις παρακάτω τρεις βασικές μορφές:

• **Native:** Σε αυτή τη μορφή, ένας hypervisor τρέχει εγγενώς σε φυσικό μηχάνημα και διαχειρίζεται πολλαπλές εικονικές μηχανές που αποτελούνται από λειτουργικά συστήματα (OS) και εφαρμογές.

• **Hosted:** Σε αυτή το πρότυπο, ένα λειτουργικό σύστημα τρέχει εγγενώς σε φυσικό μηχάνημα και ένας εικονικοποιητης τρέχει πάνω από αυτό για να φιλοξενήσει το λειτουργικό σύστημα και τις εφαρμογές πελάτη. Αυτό το μοντέλο δεν παρατηρείται συχνά στις αρχιτεκτονικές Big Data λόγω της αυξημένης επιβάρυνσης του επιπέδου του επιπλέον OS.

• **Containerized:** Σε αυτή τη μορφή, οι λειτουργίες του hypervisor είναι ενσωματωμένες στο λειτουργικό σύστημα, το οποίο τρέχει σε φυσικό μηχάνημα. Οι εφαρμογές εκτελούνται μέσα σε υποδοχεις, οι οποία ελέγχουν ή περιορίζουν την πρόσβαση στο λειτουργικό και τους φυσικους πόρους της μηχανής. Αυτή η προσέγγιση έχει

κερδίσει δημοτικότητα για τις αρχιτεκτονικές Big Data γιατί μειώνει περαιτέρω τα γενικά έξοδα, καθώς οι περισσότερες λειτουργίες του λειτουργικού συστήματος αποτελούν κοινόχρηστο πόρο. Μπορεί να μην θεωρείται εντελώς ασφαλές ή σταθερό, διότι σε περίπτωση που οι έλεγχοι / τα όρια του υποδοχέα αποτύχουν, μια εφαρμογή μπορεί να καταλύσει κάθε εφαρμογή που μοιράζεται τους ίδιους φυσικούς πόρους.

2.2.2 Φυσική και Εικονική Δικτύωση

Η συνδεσιμότητα αποτελεί ένα ζήτημα της αρχιτεκτονικής υποδομής που πρέπει να διευθετηθεί καθώς επηρεάζει το χαρακτηριστικό της ταχύτητας του συνολικού συστήματος Big Data. Ενώ μερικές εγκαταστάσεις Big Data αχολούνται αποκλειστικά με δεδομένα που βρίσκονται ήδη στο κέντρο δεδομένων και δεν χρειάζεται να φύγουν από τα όρια του τοπικού δικτύου, για άλλα ίσως χρειαστεί να σχεδιαστεί και να ληφθή υπόψην η κίνηση των Big Data είτε μέσα είτε έξω από το κέντρο δεδομένων. Η χωροταξια ενός συστήματος Big Data που απαιτεί μεταφορά μπορεί να εξαρτηθεί από τη διαθεσιμότητα εξωτερικών σύνδεσεων δικτύων (δηλαδή το εύρος ζώνης τους) και τους περιορισμούς του πρωτοκόλλου TCP-Transmission Control Protocol. Προκειμένου να αντιμετωπιστούν οι περιορισμοί του TCP, οι αρχιτέκτονες συστημάτων Big Data ίσως χρειαστεί να εξετάσουν ορισμένα προηγμένα πρωτοκόλλων επικοινωνίας που δεν βασίζονται στο TCP και είναι ειδικά σχεδιασμένα για την μεταφορά μεγάλων αρχείων, όπως βίντεο και εικόνες.

Η συνολική διαθεσιμότητα των εξωτερικών συνδέσεων είναι μια άλλη πτυχή της υποδομής που σχετίζεται με την ταχύτητα των Big Data που θα πρέπει να ληφθούν υπόψη κατά τον σχεδιασμό της εξωτερικής συνδεσιμότητας. Μια διασύνδεση μπορεί να είναι σε θέση να χειρίζεται εύκολα την ταχύτητα των δεδομένων όταν λειτουργεί σωστά. Ωστόσο,σε περίπτωση απομείωσης της ποιότητας της υπηρεσίας της σύνδεσης ή καθολικής αποτυχία της, τα δεδομένα ενδέχεται να χαθούν ή απλά να αποθηκευτούν αλλά σε σημείο που τα καθιστά μη ανακτήσιμα. Υπάρχουν υποθέσεις χρήσης όπου στην πρόβλεψη καταστάσεων έκτακτων αναγκών για το δίκτυο οι διακοπές λειτουργίας συνεπάγονται την αντιγραφή δεδομένων σε φυσικά μέσα και τη μεταφορά του υλικού στον επιθυμητό προορισμό.

Τα χαρακτηριστικά όγκου και ταχύτητας των Big Data συχνά αποτελούν παράγοντες που καθοδηγούν την ανάπτυξη της εσωτερικής υποδομής δικτύου. Για παράδειγμα, αν η εφαρμογή απαιτεί συχνές μεταφορές μεγάλα αρχεία πολλών gigabyte μεταξύ των κόμβων της συστοιχίας, τότε απαιτούνται σύνδεσμοι υψηλής ταχύτητας και χαμηλής καθυστέρησης που θα διατηρούν τη συνδεσιμότητα προς όλους τους κόμβους του δικτύου. Προβλέψεις για Δυναμική Ποιότητα Υπηρεσίας (QoS) και προτεραιότητας της υπηρεσίας μπορεί να είναι απαραίτητες προκειμένου να επιτρέπεται στους αποτυχημένους ή αποσυνδεδεμένους κόμβους να συγχρονίσουν ξανά ώστε αποκαθίσταται η συνδεσιμότητα. Ανάλογα με τις απαιτήσεις διαθεσιμότητας, πλεονάζοντες σύνδεσμοι και συνδέσεις ανθεκτικές σε σφάλματα μπορεί να απαιτηθούν. Άλλες πτυχές της υποδομής δικτύου περιλαμβάνουν την ανάλυση διευθυνσοδότησης (π.χ. Domain Name Server-DNS) και η κρυπτογράφηση μαζί με τείχη προστασίας και άλλες περιφερειακές δυνατότητες ελέγχου πρόσβασεων. Τέλος, η υποδομή δικτύου μπορεί επίσης να περιλαμβάνει αυτοματοποιημένη ανάπτυξη, δυνατότητες πρόβλεψης απαιτήσεων σε πόρους ή πράκτορες παρακολούθησης της υποδομής που αξιοποιούν τη διαχείριση / επικοινωνία στοιχεία για την εφαρμογή ενός συγκεκριμένου μοντέλου.

Για την υποστήριξη κλιμακούμενων δικτύων και συστηματων που τα χρησιμοποιούν έχουν αναπτυχθεί πρόσφατα δύο έννοιες, το Software Defined Networks (SDN) και το Network Function Virtualization (NFV)

Δίκτυα Καθορισμένα από Λογισμικό| Software Defined Networks SDN

Συχνά αγνοείται, αλλά κρίσιμη για την απόδοση των κατανεμημένων συστημάτων και πλαισίων, και ιδιαίτερα κρίσιμη για τις υλοποιήσεις Big Data, είναι η αποδοτική και αποτελεσματική διαχείριση των πόρων δικτύωσης. Όπως τα πλαίσια εικονικοποίησης, διαχειρίζονται κοινές ομάδες CPU / μνήμης / δίσκου, έτσι και SDN (ή εικονικά δίκτυα) διαχειρίζονται ομάδες φυσικών πόρων δικτύου. Σε αντίθεση με τις παραδοσιακές προσεγγίσεις της αποκλειστικής ανάθεσης φυσικού δικτύων, οι SDN περιέχουν πολλαπλούς φυσικούς πόρους (συμπεριλαμβανομένων των συνδέσεων και πραγματικών στοιχείων μεταγωγής) που συγκεντρώνονται και κατανέμονται ανάλογα με τις ανάγκες συγκεκριμένων λειτουργικότητων ή συγκεκριμένων εφαρμογων. Αυτή η κατανομή μπορεί να αποτελείται από ακατέργαστο εύρος ζώνης, προτεραιότητα στην ποιότητα της υπηρεσίας ή ακόμη και τις πραγματικές διαδρομές δεδομένων.

Εικονικοποίηση Λειτουργιών Δικτύου | Network Function Virtualization

Με την εμφάνιση της εικονικοποίησης, οι εικονικές συσκευές μπορούν τώρα να υποστηρίξουν αξιόπιστα ένα μεγάλο αριθμό λειτουργιών δικτύου που παραδοσιακά εκτελούνται από δεσμευμένες φυσικές συσκευές.

Λειτουργίες δικτύου που μπορούν να υλοποιούνται με αυτόν τον τρόπο συμπεριλαμβάνουν δρομολόγηση/δρομολογητές, υπεράσπιση περιμέτρου (π.χ. firewalls), εξουσιοδότηση απομακρυσμένης πρόσβασης και την παρακολούθηση της κυκλοφορίας/ φόρτου του δικτύου. Ορισμένα βασικά πλεονεκτήματα του NFV είναι η ελαστικότητα, η ανοχή στο σφάλμα και η ευκολότερη διαχείριση των πόρων. Για παράδειγμα, η δυνατότητα αυτόματης ανάπτυξης / πρόβλεψης πρόσθετων τείχων προστασίας ως απόκριση της αύξησης των συνδέσεων χρήστων ή διεπαφών δεδομένων και στη συνέχεια η απελευθερώση τους όταν η ζήτηση μειωθεί μπορεί να είναι κρίσιμο στοιχείο για τον χειρισμό των όγκων που σχετίζονται με τα Big Data

2.2.3 Φυσική και Εικονική Υπολογιστική

Η λογική κατανομή της υποδομής συστοιχιών/υπολογιστών μπορεί να περιλαμβάνει από κλασσικα cluster HPC-High Performance Computing, πυκνά πλέγματα φυσικών μηχανημάτων σε στήλους(racks), σύνολα εικονικών μηχανών που εκτελούνται σε έναν φορέα παροχής υπηρεσιών νέφους (Cloud Service Provider) ή σε ένα χαλαρά συνδεδεμένο σύνολο μηχανών διεσπαρμένων σε όλο τον πλανήτη που παρέχουν πρόσβαση στους αχρησιμοποίητους υπολογιστικούς τους πόρους. Η υπολογιστική υποδομή επίσης συχνά περιλαμβάνει τα υποκείμενα λειτουργικά συστήματα και τις συναφείς υπηρεσίες που χρησιμοποιούνται για τη διασύνδεση των πόρων της συστοιχίας μέσω των στοιχείων δικτύωσης. Οι υπολογιστικοί πόροι μπορούν επίσης να περιλαμβάνουν επιταχυντές υπολογισμών, όπως μονάδες επεξεργασίας γραφικών (Graphic Processing Units - GPU) και προγραμματιζόμενες σειρές πύλης (Field Programmable Gate Arrays FPGA), οι οποίες μπορούν παρέχουν δυναμικά προγραμματισμένες δυνατότητες μαζικής παράλληλης επεξεργασίας σε μεμονωμένους κόμβους της υποδομής.

2.2.4 Αποθήκευση

Η υποδομή αποθήκευσης μπορεί να περιλαμβάνει οποιονδήποτε πόρο από απομονωμένους τοπικούς δίσκους σε δίκτυα αποθήκευσης (Storage Area Network - SAN) ή αποθήκευση συνδεδεμένη στο δίκτυο (Network Attached Storage-NAS).Δύο πτυχές στην τεχνολογία της υποδομής αποθήκευσης που επηρεάζουν άμεσα την καταλληλότητά τους για Big Data είναι η χωρητικότητα και το εύρος ζώνης μεταφοράς. Οι τοπικοί δίσκοι / συστήματα αρχείων περιορίζονται ειδικά από το μέγεθος των διαθέσιμων μέσων. Υλικό ή λογισμικό λύσεων περιττής μήτρας ανεξάρτητων δίσκων (RAID-Redunant Array of Independent Disks) - σε αυτή την περίπτωση τοπικά εγκατεστημένη σε ένα κόμβο επεξεργασίας – βοηθά την κλιμάκωση, επιτρέποντας πολλαπλά τεμάχια μέσων επεξεργασίας να λειτουργούν ως ενιαία συσκευή. Ωστόσο, αυτή η προσέγγιση περιορίζεται από τις φυσικές διαστάσεις των μέσων και τον μέγιστο αριθμό των συσκευών που μπορεί να δεχθεί ο κάθε κόμβος. Οι υλοποιήσεις SAN και NAS - συχνά γνωστές ως λύσεις διαμοιράζομενων δίσκων - καταργούν αυτό το όριο ενοποιώντας την αποθήκευση σε μια συγκεκριμένη συσκευή. Με την ενοποίηση της αποθήκευσης το δεύτερο ζήτημα (δλδ το εύρος ζώνης μεταφοράς δεδομένων) μπορεί να αποτελέσει πρόβλημα. Ενώ οι διεπαφές δικτύου και I/O γίνονται ταχύτερες και πολλές εφαρμογές υποστήριζουν πολλαπλά καναλιών μεταφοράς, το εύρος ζώνης I/O μπορεί να παραμείνει ένας περιοριστικός παράγοντας. Επιπλέον, παρά τα πολλαπλά οφέλη που παρέχει το RAID θέματα όπως η υπερθέρμανση των ανταλλακτικών, η πολλαπλή τροφοδοσία ενέργειας,οι πολλοί ελεγκτές κτλπ μπορούν συχνά να καταστήσουν τις δομές αυτές κέντρα συμφόρηση I / O ή μεμονωμένα σημεία αποτυχίας σε μια επιχείρηση. Πολλές λύσεις Big Data δηλαδή αντιμετωπίζουν αυτά τα ζητήματα προτιμόντας την χρήση κατανεμημένων συστήματων αρχείων εντός του πλαισίου πλατφόρμας.

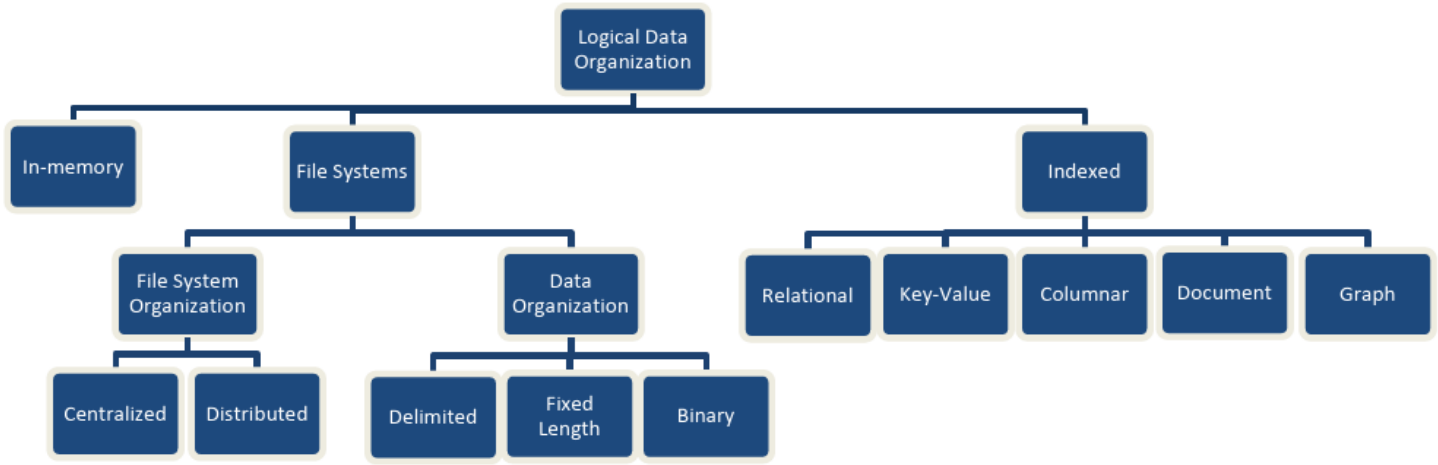
2.2.5 Φυσική Χωροθέτηση/Εγκατάσταση

Περιβαλλοντικοί πόροι, όπως η ηλεκτρική ενέργεια, η θέρμανση, ο εξαερισμός και ο κλιματισμός που παρέχονται στα σημεία φυσικής εγκατάστασης, είναι κρίσιμα για την λειτουργία του συστήματος. Οι επιχειρησιακές απαιτήσεις της αρχιτεκτονικής της υποδομής κυμαίνονται από τη βασική ισχύ και την ψύξη έως το εύρος ζώνης εξωτερικών σύνδεσεων. Μια βασική εξέλιξη που καθοδηγήθηκε από τις ανάγκες των συστηματων Big Data είναι η αύξηση του μεγέθους των διακομιστών (δηλ. περισσότερη CPU / μνήμη / δίσκους ανά μονάδα rack). Ωστόσο, με την αύξηση του μεγέθους καθίσταται πιο δύσκολη και η επαρκής,σταθερή και ισοκατανεμημένη τροφοδότηση ενέργειας και ψύξης εντός του κέντρου δεδομένων. Επιπλέον, με την άυξηση κόστος διαχείρισης της κατανάλωσης ενέργειας στα κέντρα δεδομένων, αναπτύχθηκαν τεχνολογίες που μειώνουν γενικά τους ενεργειακούς πόρους ή τους αχρησιμοποίητους πόρους με σκοπό την γενικότερη ή την κατά τις περιόδους αιχμής εξοικονόμηση ενέργειας.

Επίσης σημαντικό σε αυτό το θέμα είναι η φυσική ασφάλεια των εγκαταστάσεων και των βοηθητικών υποδομών (π.χ.υποσταθμών). Συγκεκριμένα περιμετρική ασφάλεια για την επαλήθευση διαπιστευτηρίων (π.χ. ταυτοποίηση/ βιομετρία), συνεχής επιτήρηση και περιμετρικοί συναγερμοί είναι απαραίτητοι.

2.3 Πλαίσια για τις Πλατφόρμες Δεδομένων -Data Platform Frameworks-

Τα Πλαίσια Πλατφόρμας Δεδομένων προβλέπουν την οργάνωση και τη διανομή λογικών δεδομένων σε συνδυασμό με τις συνδεδεμένες διεπαφές προγραμματισμού εφαρμογών πρόσβασης (API) ή μεθόδων. Τα πλαίσια μπορούν επίσης περιλαμβάνουν υπηρεσίες μητρώου δεδομένων και μεταδεδομένων μαζί με σημασιολογικές περιγραφές δεδομένων όπως τυπικές οντολογίες ή ταξινομίες. Η οργάνωση των λογικών δεδομένων μπορεί να κυμαίνεται από απλά οριοθετημένα επίπεδα αρχείων έως πλήρως κατανεμημένα αποθηκευτικά δεδομένα ή στήλες.



Εικόνα 7.Λογική Οργάνωση των Δεδομένων

2.3.1 Εντός Μνήμης | In-Memory

[9. Yu , Guo]

Οι βάσεις δεδομένων στη μνήμη (In-Memory Databases) είναι τα συστήματα δεδομένων που χρησιμοποιούν κατά βάση την κύρια μνήμη(RAM) ενός κόμβου υλικού για αποθήκευση και επεξεργασία δεδομένων σε αντίθεση με τα περισσότερα συστήματα διαχείρισης βάσεων δεδομένων που χρησιμοποιούν τους δίσκους. Αυτές οι βάσεις δεδομένων είναι πολύ γρήγορες με προβλέψιμη απόδοση κατά την εκτέλεση των ερωτημάτων καθώς οι εσωτερικοί αλγόριθμοι παραμετροποίησης τους είναι απλούστεροι και εκτελούν λιγότερες εντολές στην CPU. Για παράδειγμα,ενα από τα βασικά χαρακτηριστικά σχεδίασης των In-Memory Databases σε Java είναι η χρήση μνήμης εκτός σωρού.Επίσης δεν επηρεάζονται από συχνές παύσεις του εσωτερικού ελεγκτη μνήμης (Garbage Collection) του Java Virtual Machines. Μερικές φορές παρέχουν ακόμα και τη δυνατότητα αποθήκευσης των δεδομένων στο δίσκο με ασύγχρονο τρόπο.

Ένα πιθανό τεχνικό εμπόδιο με την αποθήκευση δεδομένων στη μνήμη είναι η αστάθεια της μνήμης RAM. Συγκεκριμένα, σε περίπτωση απώλειας ισχύος, εκ προθέσεως ή κατά τύχη, τα δεδομένα που είναι αποθηκευμένα σε πτητική μνήμη RAM χάνονται. Με την εισαγωγή μη-πτητικής τεχνολογίας μνήμης τυχαίας προσπέλασης, οι In-Memory Databases θα μπορούν να λειτουργούν με πλήρη ταχύτητα και να διατηρούν δεδομένα σε περίπτωση διακοπής ρεύματος.

Όπως και οι λεγόμενες βάσεις δεδομένων NoSQL, οι In-Memory databases μπορούν επίσης να παρέχουν πολύ υψηλές τιμές απόδοσης read / write και μπορούν να χρησιμοποιηθούν για διαδικτυακή επεξεργασία συναλλαγών ως προσωρινή αποθήκευση ενδιάμεσων δεδομένων πριν από την αποθήκευση των δεδομένων σε μονάδες αποθήκευσης μακράς διάρκειας (data stores).Μπορούν επίσης να χρησιμοποιηθούν για την επεξεργασία αναλυτικών στοιχείων καθώς μπορούν να διαθέσουν πολύ μεγάλο όγκο μνήμης στις εφαρμογές του πελάτη. Σαν πρόσφατη τάση, οι υβριδικές εφαρμογές συναλλαγών και ανάλυσης (HTAP) βασίζονται σε μεγάλο βαθμό σε In-Memory databases.

Ακόμα,πολλές δημοφιλείς βάσεις δεδομένων NoSQL μπορούν να ρυθμιστούν ώστε να αποθηκεύουν πλήρως και εξυπηρετούν τα δεδομένα από τη μνήμη με την αναπαραγωγή δεδομένων σε πολλούς κόμβους και ασύγχρονη διατήρηση στο δίσκο (eventual consistency).Από την άλλη πλευρά, πολλές in-memory databases μπορούν να

υποστηρίξουν columnar ή document δομές αποθήκευσης NoSQL .Αυτός είναι ο λόγος που πολλές φορές οι τεχνολογίες βάσης δεδομένων μνήμης είναι ομαδοποιούνται μαζί με τεχνολογίες NoSQL.

2.3.2 Συστηματα Αρχείων | File Systems

[7. NIST]

Πολλά πλαίσια και εφαρμογές επεξεργασίας Big Data έχουν πρόσβαση στα δεδομένα τους απευθείας από το υποκείμενο σύστημα αρχείων. Σε όλες σχεδόν τις περιπτώσεις, τα συστήματα αρχείων εφαρμόζουν σε κάποιο επίπεδο τα πρότυπα της διεπαφής φορητού λειτουργικού συστήματος (POSIX) για δικαιώματα και σχετικές λειτουργίες αρχείων. Αυτό τους επιτρέπει να λειτουργούν με σχετική διαφάνεια ως προς το κατά πόσον το υποκείμενο σύστημα αρχείων είναι τοπικό ή πλήρως κατανεμημένο. Οι προσεγγίσεις που βασίζονται σε αρχεία είναι δύο ειδών:

Οργάνωση Διακριτών Αρχείων | File System Organization

Τα συστήματα αρχείων μπορούν να είναι κεντρικοποιημένα ή κατανεμημένα. Τα κεντρικοποιήμενα συστήματα αρχείων είναι βασικά υλοποιήσεις τοπικών συστημάτων αρχείων που βρίσκονται σε μία μεγάλη πλατφόρμα αποθήκευσης (π.χ. SAN ή NAS) και είναι προσβάσιμα μέσω κάποιων δυνατοτήτων δικτύου. Σε ένα εικονικοποιημένο περιβάλλον, πολλαπλά φυσικά κεντρικοποιημένα συστήματα αρχείων μπορούν να συνδυαστούν, να διαχωριστούν ή να διανεμηθούν για τη δημιουργία πολλαπλών λογικών συστημάτων αρχείων.

Τα κατανεμημένα συστήματα αρχείων επιδιώκουν να ξεπεράσουν τα ζητήματα απόδοσης που παρουσιάζονται από τα χαρακτηριστικά όγκου και ταχύτητας στα Big Data, συνδυάζοντας τη διακίνηση εισόδου / εξόδου σε πολλαπλες συσκευές για κάθε κόμβο δεδομένων, με πολλαπλασιασμό ή αναπαραγωγή δεδομένων επίπεδου μπλοκ σε πολλαπλούς κόμβους. Πολλές από αυτές τις εφαρμογές αναπτύχθηκαν για την υποστήριξη υπολογιστικές λύσεων HPC που απαιτούν υψηλή απόδοση και δυνατότητοτητες κλιμάκωσης. Η απόδοση σε πολλές τέτοιες οι υλοποιήσεις επιτυγχάνεται συχνά μέσω αποκλειστικών κόμβων αποθήκευσης χρησιμοποιώντας ιδιόκτητες μορφές και διατάξεις αποθήκευσης

Τα **Distibuted Object Stores** (επίσης γνωστά ως global object stores) αποτελούν ένα μοναδικό παράδειγμα κατανεμημένου συστήματος οργάνωσης αρχείου . Σε αντίθεση με άλλες προσεγγίσεις που εφαρμόζουν ένα παραδοσιακό σύστημα αρχείων για την ιεράρχιση του χώρου ονομάτων, τα global object stores παρουσιάζουν ένα επίπεδο χώρο ονομάτων με ένα καθολικά μοναδικό αναγνωριστικό (GUID) για κάθε κομμάτι δεδομένων. Γενικά, τα δεδομένα στο object stores βρίσκονται μέσω ερωτήματος σε έναν κατάλογο μεταδεδομένων που επιστρέφει τα αντίστοιχα GUID. Το GUID παρέχει την υποκείμενη υλοποίηση λογισμικού με τη θέση αποθήκευσης των δεδομένων που ενδιαφέρουν. Τέτοια συστήματα αναπτύσσονται και διατίθενται στην αγορά για την αποθήκευση πολύ μεγάλων αντικειμένων δεδομένων, από πλήρη σύνολα δεδομένων έως μεγάλα μεμονωμένα αντικείμενα (π.χ. εικόνες υψηλής ανάλυσης μεγέθους δεκάδων gigabytes GB). Ο μεγαλύτερος περιορισμός αυτών global object stores στα Big Data είναι η συμφόρηση του δικτύου (δηλ. ταχύτητα), επειδή πολλοί μπορεί να απαιτούν ένα αντικείμενο στο οποίο υπάρχει καθολική πρόσβαση. Ωστόσο, οι μελλοντικές τάσεις δείχνουν την επικράτηση της έννοια της μεταφοράς του υπολογισμού / εφαρμογής προς στα δεδομένα έναντι της ανάγκης να έρθουν τα δεδομένα στην εφαρμογή.

Από την άποψη της ωριμότητας, δύο βασικοί τομείς στους οποίους είναι πιθανό να βελτιωθούν τα κατανεμημένα συστήματα αρχείων είναι (1) τυχαία εγγραφή εισόδου/εξόδου και η συνέπεια, και (2) την παραγωγή de facto προτύπων σε ένα παρόμοιο ή υψηλότερο επίπεδο, όπως οι σειρές των εγγράφων του Internet Engineering Task Force που είναι διαθέσιμα πχ για το πρωτόκολλο του συστήματος αρχείων δικτύου (NFS).

Οργάνωση εντός των Δεδομένων του Αρχείου | In File Data Organization

Πολύ λίγα συστήματα διαφοροποιούνται στην λεγόμενη οργάνωση δεδομένων εντός αρχείων. Τα δεδομένα που βασίζονται σε αρχεία μπορούν να είναι κείμενο, δυαδικά δεδομένα,εγγραφες σταθερού μήκους ή κάποιας οριοθετημένης δομής (π.χ. τιμές διαχωρισμένες με κόμματα CSV, XML). Η αποθήκευση με προσανατολισμό στην εγγραφή (είτε οριοθετημένου είτε σταθερού μήκους) γενικά δεν αποτελεί ζήτημα Big Data εκτός αν υπάρχουν εγγραφές του μπορούν να υπερβουν σε απαιτήσεις αποθήκευσης το μέγεθος ενός μπλοκ. Ορισμένα κατανεμημένα συστήματα αρχείων οι εφαρμογές παρέχουν συμπίεση στο επίπεδο τόμου ή καταλόγου και την εφαρμόζουν κάτω από το λογικό επίπεδο (π.χ. όταν ένα μπλοκ διαβάζεται από το σύστημα αρχείων, αποσυμπιέζεται / αποκρυπτογραφείται πριν επιστραφεί ως απάντηση). Λόγω της απλότητας, της εξοικείωσης και της φορητότητάς τους, τα οριοθετημένα αρχεία είναι συχνά η προεπιλεγμένη μορφή αποθήκευσης σε πολλές υλοποιήσεις Big Data. Το αντιστάθμισμα είναι αποδοτικότητα του I/O (εισόδου/εξόδου δηλ. η ταχύτητα).Παρόλο

που τα μεμονωμένα μπλοκ σε ένα κατανεμημένο σύστημα αρχείων μπορούν να προσπελαστούν παράλληλα, κάθε μπλοκ πρέπει απαιτείται να διαβαστεί διαδοχικά. Σε περίπτωση οριοθετημενοι αρχείου, αν ενδιαφέρει μόνο το τελευταίο πεδίο απο ένα σύνολο εκατοντάδες πεδία, πολύ I/O και εύρος επεξεργασίας χάνεται.

Η δυαδική μορφοποίηση (binary format) τείνει να είναι προσανατολισμένη σε συγκεκριμένες εφαρμογές ή υλοποίησεις. Ενώ μπορούν να προσφέρουν πολύ περισσότερο αποτελεσματική πρόσβαση λόγω μικρότερων μεγεθών δεδομένων (δηλ., οι ακέραιοι αριθμοί είναι δυο έως τέσσερα byte σε δυαδικό, ενώ ένα byte καταλαμβάνεται ανά ψηφίο αρχείου σε ASCII), προσφέρουν περιορισμένη συμβατότητα μεταξύ διαφορετικών εφαρμογών.Κάθε δημοφιλές κατανεμημένο σύστημα αρχείων παρέχει το δικό του ιδιόκτητο πρότυπο δυαδικό σχήμα, το οποίο επιτρέπει στα δεδομένα να είναι μεταφέρσιμα μεταξύ πολλαπλών εφαρμογών χωρίς πρόσθετο λογισμικό. Ωστόσο, ο κύριος όγκος των προσεγγίσεων οργάνωσης ευρετηρίων που αναφέρονται παρακάτω χειρίζονται τις δυαδικές μορφές για σκοπούς καλύτερης απόδοσης.

2.3.3 Πλατφόρμες Αποθήκευσης Οργάνωμένες Βάσει Ευρετηρίων | Indexed Storage Organization

Η ίδια η φύση των Big Data (κυρίως τα χαρακτηριστικά όγκου και ταχύτητας) οδηγουν στην πράξη σε απαιτήσεις για την εφαρμογή κάποιας μορφής ευρετηρίασης. Ο μεγάλος όγκος δεδομένων απαιτεί να υπάρχουν συγκεκριμένα στοιχεία δεδομένων που θα εντοπίζονται γρήγορα χωρίς σάρωση σε όλο το σύνολο δεδομένων. Μεγάλη ταχύτητα απαιτείται ακόμα και για απλή αντιστοίχιση (π.χ., εισερχόμενα δεδομένα που ταιριάζουν με αυτά ενός υπάρχοντος σύνολου δεδομένων) είτε να με το να γνωρίζει άμεσα το σύστημα πού να γράψει/ενημερώσει τα νέα δεδομένα.

Η επιλογή μιας συγκεκριμένης μεθόδου ή μεθόδων ευρετηρίασης εξαρτάται κυρίως από την φύση των ίδιων των δεδομένων και τη εφαρμογής. Για παράδειγμα, τα δεδομένα γράφων (δηλ. κορυφές, ακμές και ιδιότητες) μπορούν εύκολα να αναπαρασταθούν σε αρχεία κειμένου ως ζεύγη κορυφών-ακμής, τριπλετών ακμής-κορυφής ή καταγραφές λίστας κορυφών-ακμών. Ωστόσο, η επεξεργασία αυτών των δεδομένων αποτελεσματικά θα απαιτούσε ενδεχομένως τη φόρτωση ολόκληρου του συνόλου δεδομένων στη μνήμη ή να είναι σε θέση να κατανέμει την εφαρμογή και το σύνολο δεδομένων σε πολλαπλούς κόμβους έτσι ώστε να είναι μέρος του γραφήματος να τοποθετηθεί στην μνήμη κάθε κόμβου.

Οι προσεγγίσεις ευρετηρίασης τείνουν να ταξινομούνται βάσει των χαρακτηριστικών που παρέχονται στην υλοποίηση, και συγκεκριμένα:

-Την πολυπλοκότητα των δομών δεδομένων που μπορούν να αποθηκευτούν

-Πόσο καλά μπορούν να επεξεργαστούν οι διασυνδέσεις εντός των δεδομένων

-Πόσο εύκολα υποστηρίζουν μοτίβα πολλαπλών προσβάσεων.

Δεδομένου ότι οποιοδήποτε από αυτά τα χαρακτηριστικά μπορεί να υλοποιηθούν στον κώδικα της εφαρμογής, οι τιμές απεικονίζονται με τρόπο που ακολουθεί συγκεκριμένα πρότυπα. Για παράδειγμα, η αποθήκευση κλειδιών-τιμών λειτουργούν καλά για δεδομένα που πρέπει να προσπελαστούν μόνο μέσω ενός μόνο κλειδιού, των οποίων οι τιμές μπορούν να είναι εκφράζονται σε μια ενιαία επίπεδη δομή όπου δεν χρειάζεται να συσχετίζονται μεταξύ τους. Από την άλλη οι αποθηκεύσεις βάσει εγγράφων μπορούν να υποστηρίξουν πολύ σύνθετες δομές αυθαίρετου πλάτους και τείνουν να είναι ευρετηριασμένες για πρόσβαση μέσω διαφόρων ιδιότητων του εγγράφου, και επομένως δεν μπορούν να υποστηρίζουν καλά τις σχέσεις μεταξύ εγγραφών. Σημειώνεται ότι οι συγκεκριμένες εφαρμογές για κάθε προσέγγιση αποθήκευσης ποικίλουν σημαντικά. Η αποθήκευση βάσει ευρετηρίων γενικά διακρίνεται στα εξής είδη:

- Σχεσιακές | Relational
- Σε Ζευγη Κλειδιού-Τιμής |Key-Value
- Διευρυμένων Στηλών | Wide-Columns
- Γράφων | Graph

2.4 Πλαίσια Επεξεργασίας -Processing Frameworks-

[7. NIST]

Τα πλαίσια επεξεργασίας καθορίζουν τον τρόπο οργάνωσης των υπολογισμών και της επεξεργασίας των δεδομένων. Οι εφαρμογές Big Data βασίζονται σε διάφορες πλατφόρμες και τεχνολογίες για να αντιμετωπίσουν τις προκλήσεις της κλιμακούμενης ανάλυσης δεδομένων και των αδιάλλειπτων λειτουργιών.

Τα πλαίσια επεξεργασίας λοιπόν γενικά επικεντρώνονται στον χειρισμό των δεδομένων, ο οποίος μπορεί να κυμαίνεται μεταξύ της επεξεργασίας τύπου παρτίδας(batch) και επεξεργασίας ροής χωρίς τα μεταξύ τους όρια να είναι πολύ σαφή. Συνήθως, τα σύγχρονα συστήματα συμπεριλαμβάνουν πολλαπλά είδη πλαισίων για την υποστήριξη ευρούς φάσματος απαιτήσεων.

Τυπικά, τα πλαίσια επεξεργασίας κατηγοριοποιούνται με βάση το αν υποστηρίζουν επεξεργασία παρτίδας ή ροής. Αυτή η κατηγοριοποίηση αναφέρεται γενικά στην οπτική του χρήστη (π.χ., πόσο γρήγορα παίρνει ένας χρήστης απάντηση σε αίτημα). Ωστόσο, τα πλαίσια επεξεργασίας Big Data στην πραγματικότητα περιλαμβάνουν τρεις φάσεις επεξεργασίας:

Εισαγωγή των δεδομένων, την **ανάλυση** δεδομένων και **διάδοση** δεδομένων, οι οποίες δομούνται σε στενές ροές δεδομένων μέσω της αρχιτεκτονικής. Οι δραστηριότητες του Παροχου της Εφαρμογής Big Data ελέγχουν της επιμέρους ειδικές λειτουργίες του πλαισίου σε καθεμία από αυτές τις φάσεις επεξεργασίας.Για παράδειγμα, τα δεδομένα ενδέχεται να εισέλθουν στο σύστημα με μεγάλη ταχύτητα και ο τελικός χρήστης να πρέπει γρήγορα να παραλάβει μια σύνοπτική αναφορά των δεδομένων της προηγούμενης ημέρας. Σε αυτή την περίπτωση, η λήψη των δεδομένων στο σύστημα πρέπει να γίνει να είναι τύπου NRT-near real time και να συμβαδίζει με την τρέχουσα ροή δεδομένων. Το τμήμα ανάλυσης μπορεί να είναι αυξητικό (π.χ.η προσαύξηση η οποία θα εκτελείται κατα την διάρκεια της εισαγωγής των δεδομένων) ή θα μπορούσε να είναι μια διαδικασία batch που εκτελείται σε συγκεκριμένη χρονική στιγμή, ενώ η ανάκτηση (δλδ η οπτικοποίηση) των δεδομένων θα μπορούσε να είναι διαδραστική. Ανάλογα με την περίπτωση χρήσης, μπορεί να γίνει μετασχηματισμός δεδομένων σε οποιοδήποτε σημείο κατά τη διάρκεια της μετακίνησης εν μέσω του συστήματος. Για παράδειγμα, η φάση της εισαγωγής μπορεί να γράψει μόνο τα δεδομένα όσο το δυνατόν γρηγορότερα ή μπορεί να εκτελεστεί κάποια βασική ανάλυση για την σταδιακής παρακολούθηση μετρήσεων όπως ελάχιστο, μέγιστο, μέσο όρο κλπ. Η κεντρική εργασία επεξεργασίας μπορεί να εκτελέσει μόνο το αναλυτικά στοιχεία που απαιτούνται από τον Παροχο της Εφαρμογής Big Data και υπολογίζουν έναν πίνακα δεδομένων ή μπορεί δημιουργούν υποστήρικτικές δομές για την απεικόνιση των στοιχείων. Για να είναι η προβολή όσο το δυνατόν ταχύτερη, η φάση διάδοσης των δεδομένων σχεδόν πάντα εκτελεί κάποιες μεταλλάξεις, αλλά η έκταση τους εξαρτάται από τη φύση των δεδομένων και της επιθυμητής οπτικοποίησης.

Batch Πλαίσια

Τα πλαίσια παρτίδας, των οποίων οι ρίζες προέρχονται από την εποχή των mainframes, είναι μερικά από τα πιο διαδεδομένα και ώριμα συστατικά της αρχιτεκτονικής Big Data επειδή οι ιστορικά μεγάλοι χρόνοι επεξεργασίας αφορούν μεγάλους όγκους δεδομένων. Τα πλαίσια παρτίδας ιδανικά δεν συνδέονται με έναν συγκεκριμένο αλγόριθμο ή ακόμη και με έναν τύπο αλγορίθμου, αλλά παρέχουν ένα μοντέλο προγραμματισμού όπου μπορούν να εφαρμοστούν πολλαπλές κατηγορίες αλγορίθμων. Επίσης, ειδικά για τα Big Data, αυτά τα μοντέλα επεξεργασίας κατανεμονται συχνά σε πολλαπλά κόμβους μιας συστοιχίας. Διακρίνονται συνήθως από την ποσότητα της ανταλλαγής δεδομένων μεταξύ διαδικασιών / δραστηριότητων εντός του μοντέλου.

Streaming Πλαίσια

Τα πλαίσια ροής είναι κατασκευασμένα για να αντιμετωπίζουν δεδομένα που απαιτούν επεξεργασία το ίδιο η ακόμα πιο γρήγορα από την ταχύτητα με την οποία καταφτάνουν στο σύστημα Big Data. Ο πρωταρχικός στόχος της ροής πλαισίων είναι να να μειωθεί η καθυστέρηση μεταξύ της άφιξης δεδομένων στο σύστημα και της δημιουργίας, αποθήκευσης ή παρουσίασης των αποτελέσματικών δεδομένων.

Το CEP-Complex Event Processing είναι ένας από τους προβληματικούς τομείς που συχνά αντιμετωπίζονται από τα streaming πλαίσια. Το CEP χρησιμοποιεί δεδομένα από ένα ή περισσότερα ρεύματα / πηγές για την εξαγωγή ή αναγνώριση συμβάντων ή μοτίβων στο NRT.

Σχεδόν όλα τα πλαίσια ροής που είναι διαθέσιμα σήμερα εφαρμόζουν κάποια μορφή βασικής ροής επεξεργασίας. Αυτές οι ροές εργασίας χρησιμοποιούν πλαίσια μηνυμάτων/επικοινωνιών για τη μετάδοση αντικείμενων δεδομένων (που συχνά αναφέρονται ως συμβάντα events) μεταξύ των βημάτων στη ροή εργασίας. Αυτό συχνά παίρνει τη μορφή κατευθυνόμενου γράφου εκτέλεσης. Ο διαχωρισμός των πλαισίων ροής είναι τυπικά οργανωμένη γύρω από τα ακόλουθα τρία χαρακτηριστικά: την **προτεραιοποίηση των συμβάντων** και τις

εγγυήσεις στην διαδικασία επεξεργασίας, την **διαχείριση κατάστασης** και τον **διαχωρισμό/παραλληλισμό**. Αυτά τα τρία χαρακτηριστικά περιγράφονται παρακάτω.

Εγγύηση Ακολουθίας και Επεξεργασίας Συμβάντων | Event Ordering and Processing Guarantees

Αυτό το χαρακτηριστικό αναφέρεται στο αν τα στοιχεία επεξεργασίας ροής είναι φτιαγμένα ώστε εγγυημένα να βλέπουν μηνύματα ή συμβάντα ακριβώς με τη σειρά που εισέρχονται το Big Data System, καθώς και πόσο συχνά ένα μήνυμα ή ένα συμβάν μπορεί ή δεν μπορεί να μην υποστεί επεξεργασία. Σε μια μη κατανεμημένη και ροής μοναδικού νήματος λειτουργία, αυτές οι εγγύησεις είναι σχετικά δεδομένες. Μόλις όμως προστίθενται στο σύστημα κατανεμημένες και/ή πολλαπλές ροές, η εγγύηση γίνεται περισσότερο περίπλοκη. Με την κατανεμημένη επεξεργασία, οι εγγυήσεις πρέπει να εφαρμόζονται για κάθε διαμέριση των δεδομένων. Επιπλοκές προκύπτουν όταν η αλληλεπίδραση μιας διαδικασία/εργασίας και ενός διαμερίσματος επεξεργασίας καταρρεύσει. Οι εγγυήσεις επεξεργασίας χωρίζονται συνήθως στις ακόλουθες τρεις κατηγορίες:

• **Το πολύ μια παράδοση:** Αυτή είναι η απλούστερη μορφή εγγύησης και επιτρέπει μηνύματα ή συμβάντα να αποσυρθούν σε περίπτωση αποτυχίας στην επεξεργασία ή την επικοινωνία ή αν φτάσουν εκτός σειράς. Ισχύει για δεδομένα για τα οποία δεν υπάρχει εξάρτηση από την κατάσταση γεγονότων που δημιουργήθηκαν από προηγούμενα γεγονότα.

• **Παράδοση τουλάχιστον μία φορά**: Σε αυτή την ομάδα, τα πλαίσια θα παρακολουθούν κάθε μήνυμα ή συμβάν (και οποιαδήποτε μεταγενέστερα μηνύματα ή συμβάντα δημιουργηθούν) για να επαληθεύσει ότι γίνεται επεξεργασία εντός ενός προκαθορισμένου χρονικού πλαίσιου. Τα μηνύματα ή τα γεγονότα που δεν επεξεργάζονται κατά το επιτρεπόμενο χρονικό διάστημα εισάγονται εκ νέου στο ρεύμα. Αυτή η δυνατότητα απαιτεί εκτεταμένη διαχείριση της κατάστασης από το πλαίσιο (και μερικές φορές ακόμα και από την συσχετισμένη εφαρμογή) για τον εντοπισμό των γεγονότων που έχουν επεξεργαστεί από κάποια στάδια της ροής εργασίας. Ωστόσο, στην κατηγορία αυτή, μηνύματα ή συμβάντα μπορούν να υποβάλλονται σε επεξεργασία περισσότερες από μία φορές ή να καταφτάνουν εκτός τάξης. Αυτή η κατηγορία εγγύησεων είναι κατάλληλη για συστήματα όπου όλα μήνυματα ή συμβάντα πρέπει να υποβάλλονται σε επεξεργασία ανεξάρτητα από τη σειρά (π.χ., καμία εξάρτηση από προηγούμενα συμβάντα),και η εφαρμογή είτε δεν επηρεάζεται από την διπλή επεξεργασία των γεγονότων είτε έχει τη δυνατότητα για διπλά γεγονότα.

• **Ακριβώς μία παράδοση:** Αυτή η κλάση επεξεργασίας πλαισίου απαιτεί την ίδια κατάσταση ανώτερου επιπέδου τουλάχιστον μιας παράδοσης αλλά ενσωματώνει μηχανισμούς εντός του πλαισίου για την ανίχνευση και την παράκαμψη των διπλών. Αυτή η κατηγορία συχνά εγγυάται την σειρά των αφίξεων των συμβάντων και απαιτείται απο εφαρμογές όπου η επεξεργασία οποιουδήποτε γεγονότος εξαρτάται από την επεξεργασία προηγούμενων γεγονότων. Σημειώνεται ότι αυτές οι εγγυήσεις ισχύουν μόνο για το χειρισμό δεδομένων εντός του πλαίσιο. Εάν υπάρχουν δεδομένα που περάσουν έξω από την τοπολογία, κατόπιν η εφαρμογή πρέπει να διασφαλίσει ότι η ορθή κατάσταση επεξεργασίας διατηρείται.

Διαχείριση Κατάστασης | State Management

Ένα κρίσιμο χαρακτηριστικό των πλαισίων επεξεργασίας ρευμάτων είναι η ικανότητά τους να ανακάμπτουν και να μην χάνουν κρίσιμα δεδομένων σε περίπτωση βλάβης διαδικασίας ή κόμβου. Τα πλαίσια συνήθως παρέχουν αυτή την διαχείριση κατάστασης μέσω της διατήρηση των δεδομένων σε κάποια μορφή αποθήκευσης. Αυτή η διατηρεί μπορεί να είναι:

-τοπική, επιτρέποντας την επανεκκίνηση της αποτυχημένης διαδικασίας στον ίδιο κόμβο

-ένα απομακρυσμένο ή κατανεμημένο στοιχείο αποθήκευσης δεδομένων, επιτρέποντας τη διαδικασία να ξαναρχίσει σε οποιονδήποτε κόμβο

-τοπική αποθήκευση που αναπαράγεται σε άλλους κόμβους.

Ο συμβιβασμός μεταξύ αυτών των μεθόδων αποθήκευσης είναι η καθυστέρηση που επιφέρει από την διατήρηση. Τόσο ο όγκος των δεδομένων κατάστασης που θα διατηρηθούν όσο και ο χρόνος που απαιτείται για να διασφαλιστεί ότι τα δεδομένα αποθηκεύονται ασφαλώς συμβάλλει στην καθυστέρηση. Στην περίπτωση της απομακρυσμένης ή κατανεμημένης αποθήκευσης δεδομένων, η απαιτούμενη καθυστέρηση εξαρτάται γενικά από το βαθμό στον οποίο στην αποθήκευση εφαρμόζονται οι συνθήκες του ACID (Ατομικότητα, Συνέπεια, Απομόνωση, Αντοχή) ή του BASE ως στυλ συνέπειας. Με την αναπαραγωγή της τοπικής αποθήκευσης, η αξιοπιστία της διαχείρισης κατάστασης εξαρτάται εξ ολοκλήρου από την ικανότητα της αναπαραγωγής να ανακάμψει σε περίπτωση αποτυχία μιας διεργασίας ή κόμβου. Μερικές φορές αυτή η αναπαραγωγή της κατάστασης εφαρμόζεται πραγματικά χρησιμοποιώντας το ίδιο πλαίσιο μηνυμάτων/επικοινωνίας που χρησιμοποιείται για την επικοινωνία μεταξύ των επεξεργαστών ροής. Ορισμένα πλαίσια υποστηρίζουν την πλήρη σημασιολογία των συναλλαγών, συμπεριλαμβανομένων πολυεπίπεδων κατοκυρωσέων (multistage commits) και την επαναφορά συναλλαγών(transaction rollback). Το αντιστάθμισμα είναι το ίδιο που υπάρχει για οποιοδήποτε

σύστημα συναλλαγών όπου οι εγγύησεις τύπου ACID εισαγάγουν καθυστέρηση. Η υπερβολική καθυστέρηση σε οποιοδήποτε σημείο της ροής μπορεί αποτελέσει σημείο συμφόρησης και μπορεί να οδηγήσει σε καταστάσεις αδιέξοδης παύσης(deadlock) ή ατερμονος βρόχου - ειδικά όταν είναι ανεκτό κάποιο επίπεδο αποτυχιών.

Κατατμήση και Παραλληλισμός | Partitioning & Parallelism

Αυτό το χαρακτηριστικό των πλαισίων ροής σχετίζεται με την κατανομή των δεδομένων μεταξύ των κόμβων και των πρακτόρων εργασίας (**workers**) να παρέχει την οριζόντια επεκτασιμότητα που χρειάζεται για την αντιμετώπιση του όγκου και της ταχύτητας των Big Data. Αυτό το σύστημα κατάτμησης πρέπει να αλληλεπιδρά με το πλαίσιο διαχείρισης πόρων για τους σκοπούς της κατανομής πόρων. Η ομοιόμορφη κατανομή των δεδομένων μεταξύ των κατατμήσεων είναι απαραίτητη για την ομοιόμορφη κατανομή της σχετικής εργασίας. Η ομαλή κατανομή δεδομένων σχετίζεται άμεσα με την επιλογή ενός κλειδιού (π.χ., αναγνωριστικό χρήστη, όνομα κεντρικού υπολογιστή) που είναι ομοιόμορφα κατανεμημένο. Η απλούστερη μορφή μπορεί να χρησιμοποιεί έναν αριθμό που αυξάνει κατά ένα και τότε επεξεργάζεται με βάση μια υπολογιστική συνάρτηση επιλογής από τις διαθέσιμες εργασίες/πράκτορες. Εάν εξαρτήσεις εντός των δεδομένων απαιτούν όλες οι εγγραφές με το ίδιο κλειδί να υποβάλλονται σε επεξεργασία από τον πράκτορα τότε η διασφάλιση της ομοιόμορφη κατανομή των δεδομένων κατά τη διάρκεια της ζωής του ρεύματος μπορεί να είναι δύσκολη. Ορισμένα πλαίσια ροής αντιμετωπίζουν αυτό το ζήτημα υποστηρίζοντας δυναμική κατάτμηση όπου οι πράκτορες που έχουν υπερφορτωθεί διαιρούνται και μετατίθενται σε άλλους ελεύθερους υφιστάμενους ή νεοσύστατους πράκτορες. Για την επίτευξη επιτυχίας - ειδικά όταν η εξάρτηση των δεδομένων ή κατάστασης που σχετίζεται με το κλειδί- αποτελεί κρίσιμο σημείο το πλαίσιο να έχει έχει διαχείριση κατάστασης, η οποία επιτρέπει τα συναφή δεδομένα κατάστασης να μεταφερθούν στο νέο πράκτορα.

Ροές Εργασίας

[21. BARIKA, GARG, ZOMAYA, WANG, VAN MOORSEL, RAJIV, RANJAN]

Partitioning

Η εκκίνηση της ροής εργασιών στοχεύει να χωρίσει (με ή χωρίς περιορισμούς) μια ροή εργασίας σε τμήματα με σκοπό την παράλληλη εκτέλεση αυτών των τμημάτων αυτών σε διαφορετικούς υπολογιστικούς πόρους.

Non-Constraint-based Partitioning

Αυτή η προσέγγιση αποσυνθέτει μια ροή εργασίας σε μικρότερα τμήματα για να είναι εφικτή η κατανομή αυτών των τμημάτων για παράλληλη εκτέλεση σε διάφορους υπολογιστικούς πόρους. Λαμβάνει υπόψη τις εξαρτήσεις των εργασιών και των δεδομένων στη ροή εργασίας και αποφεύγει την αλληλεξάρτηση/σύγκρουση.Η απόφαση δεν λαμβάνει καθόλου υπόψην τις δυνατότητες των υπολογιστικών πόρων ή το κόστος μετακίνησης των δεδομένων.

Constraint-based Partitioning

Αυτή η προσέγγιση χωρίζει μια ροή εργασίας σε μικρότερα τμήματα, λαμβάνοντας υπόψη τους προκαθορισμένους περιορισμούς ώστε να επιτρέπεται την κατανομή αυτών των τμημάτων σε υπολογιστικούς πόρους για παράλληλη εκτέλεση. Υπάρχουν πέντε ακόλουθες τεχνικές για την υποστήριξη διαμέρισης βάσει περιορισμών:

- ***Data Transfer Constrained Partitioning*** ,στοχεύει στην ελαχιστοποίηση του όγκου των δεδομένων που θα μετακινηθούν μεταξύ των τμημάτων μιας ροής εργασίας
- *Security and Privacy Constrained Partitioning*, διαχωρισμό μιας ροής εργασίας σε τμήματα υπό περιορισμούς ασφάλειας και απορρήτου. Για παράδειγμα, μια ροή εργασίας μπορεί να περιέχει μια κρίσιμη δραστηριότητα που απαιτεί την εκτέλεση σε έναν αξιόπιστο ιστότοπο cloud
- ***Compute Capacity Constrained Partitioning*****,** σύμφωνα με την διαμόρφωση των υπολογιστικών πόρων. Οι διαφορετικές διαμορφώσεις των υπολογιστικών πόρων στο cloud ή ετερογενείς διαμορφώσεις πολλαπλών cloud τοποθεσιών μπορούν να χρησιμοποιηθούν για την προσαρμογή της διαμέρισης της ροής εργασίας
- ***Storage Constrained Partitioning***, σεβασμό των περιορισμών αποθήκευσης κατά τη διάρκεια της κατάτμησης μιας ροής εργασίας σε τμήματα
- **Multi-Constraints Partitioning***, σέβεται πολλούς παράγοντες ή περιορισμούς στη διαδικασία* διαμέρισης μιας ροής εργασίας

Parallelization

Coarse-grained Parallelization

Αυτή η προσέγγιση επιτυγχάνει παραλληλισμό στο επίπεδο της ροής εργασίας. Είναι ζωτικής σημασίας για την εκτέλεση ροής εργασιών μετα-ροής ή εκτέλεσης ροής εργασιών παραμέτρων. Για **μετα-ροή εργασίας**,αυτή η τεχνική παραλληλίζει την εκτέλεση ανεξάρτητων υπο-ροών εργασίας υποβάλλοντας τις σε αντίστοιχες μηχανές ροής εργασίας. Σε μια εκτέλεση ροής εργασίας **σάρωσης παραμέτρων**, κάθε σύνολο τιμών παραμέτρων εισόδου οδηγεί σε ανεξάρτητη υπο-ροή εργασίας.

Fine-grained Parallelization

Αυτή η προσέγγιση επιτυγχάνει παραλληλισμό σε επίπεδο δραστηριότητας μέσα σε μια ροή εργασίας ή μια υπο-ροή εργασίας, όπου διαφορετικές δραστηριότητες θα εκτελούνται παράλληλα. Σε αυτό το επίπεδο, υπάρχουν διαφορετικοί τύποι παραλληλισμού

- ***Data Parallelization:***Αυτός ο τύπος χειρίζεται τον παραλληλισμό μέσα στην δραστηριότητα. Για να επιτευχθεί πρέπει να έχει διάφορες νήματα να εκτελούν την ίδια δραστηριότητα και καθένα από αυτά να επεξεργάζεται διαφορετικό τμήμα των δεδομένων εισόδου σε διάφορετικό υπολογιστικό κόμβο. Έτσι, δεδομένου ότι τα δεδομένα εισόδου είναι διαχωρισμένα ,τα δεδομένα εξόδου είναι εξίσου διαχωρισμένα. Αυτά τα διαχωρισμένα αποτελέσματα λοιπόν θα μπορούσαν να είναι δεδομένα εισαγωγής για παραλληλισμό δεδομένων για τις επόμενες δραστηριότητες ή να συνδυαστούν για να παράγουν ένα και μόνο αποτέλεσμα. Μπορεί να είναι **στατικός**, όπου ο αριθμός των τμήματων δεδομένων είναι σταθερά και καθορισμένα πριν από την εκτέλεση **δυναμικός** όπου ο αριθμός των τμήματων δεδομένων αναγνωρίζονται κατά το χρόνο εκτέλεσης αλλά και προσαρμόσιμος, όπου είναι ο αριθμός τμημάτων δεδομένων αυτόματα τροποποιείται από περιβάλλον εκτέλεσης
- ***Independent Parallelism:***Χειρίζεται τον παραλληλισμό μεταξύ ανεξάρτητων δραστηριοτήτων μιας ροή εργασίας. Για να επιτευχθεί ,η ροή εργασίας θα πρέπει να έχει τουλάχιστον δύο ή περισσότερα ανεξάρτητα τμήματα δραστηριοτήτων και οι δραστηριότητες κάθε τμήματος να μην έχει εξαρτήσεις δεδομένων με δραστηριότητες άλλων τμημάτων. Επιπλέον, αυτές οι ανεξάρτητες δραστηριότητες πρέπει να προσδιοριστούν προκειμένου να εκτελεστούν παράλληλα.
- ***Pipeline Parallelization****:*Αυτός ο τύπος χειρίζεται τον παραλληλισμό μεταξύ εξαρτημένων δραστηριοτήτων. Οι δραστηριότητες διατηρούν κάποιον συνηθισμένο τύπο σχέσης μεταξύ τους (δηλ. σχέση παραγωγού-καταναλωτή) μπορούν να εκτελεστούν με παράλληλο τρόπο, όπου η έξοδος ενός τμήματος δεδομένων μιας δραστηριότητας είναι η είσοδος για τις ακόλουθες εξαρτημένες δραστηριοτήτες. Αξιοποιώντας αυτόν τον τύπο παραλληλισμού, η κατανάλωση τμημάτων δεδομένων πραγματοποιείται μόλις αυτά τα είναι έτοιμα.
- ***Hybrid Parallelism:***Αυτός ο τύπος συνδυάζει και τους τρεις παραπάνω τύπους παραλληλισμού προκειμένου να επιτευχθεί υψηλότερος βαθμός τελικού παραλληλισμού.

2.5 Έννοιες στην Υπολογιστική και Ανάκτηση των Big Data -Big Data Computing & Retrieval-

Η υπολογιστική των Big Data επικεντρώνεται σε τρείς τομείς:
-Επεξεργασία μεγάλου όγκου δεδομένων σε ηρεμία
-Επεξεργασία μεγάλου όγκου δεδομένων συνεχούς ροής
-Τυχαία προσπέλαση σε δεδομένα μεγάλου όγκου για ανάγνωση ή εγγραφή από/σε μεγάλο όγκο δεδομένων

Για την αντιμετώπιση των παραπάνω, στις αρχιτεκτονικές Big Data συστημάτων ο υπόλογιστικός πυρήνας δομείται έχωντας ακολουθήσει μια από τις εξής δύο σχολές σκέψης:
-**Distributed Proccessing**, το οποίο με την σειρά του διακρίνεται στις εξής δύο υποκατηγορίες λύσεων:
-Κατανεμημένες Συστοιχίες (Distributed Cluster Systems)
-Μαζική Παραλληλη Επεξεργασία (MPP)
-**In-Memory Proccessing**

Στις παρακάτω ενότητες αναλύονται τα είδη αυτά.

2.5.1 Κατανεμημένα Συστήματα Συστοιχίας | Distributed Cluster Systems

Οι βασικές αφαίρεσεις που χρησιμοποιούνται για την αντιμετώπιση αυτών των διαφόρετικών τύπων επεξεργαστικών αναγκών είναι:
1.Οι Κατανεμημένες Μηχανές Επεξεργασίας | Distributed Processing Engines
2.Τα Συστατικά Στοιχεία Εφαρμογών | Application Components
3.Οι Διεπαφές Πρόσβασης στα Δεδομένα | Data Access Interfaces
4.Η Ασφάλεια Δεδομένων | Data Security

Παρακάτω ακολουθεί η ανάλυση τους.

2.5.1.1 Distributed Processing Engine | Κατανεμημένη Μηχανή Επεξεργασίας

[9. Yu , Guo]

Η κατανεμημένη μηχανή επεξεργασίας είναι η βασική αφαίρεση που χρησιμοποιείται στις τεχνολογίες Big Data. Στις περισσότερες περιπτώσεις οι Μηχανές Κατανεμημένης Επεξεργασίας αναπτύσσονται χρησιμοποιώντας τις έννοιες της αρχιτεκτονικής Μαζικής Παράλληλης Επεξεργασίας(Massive Parallel Processing) αλλά με ελαφρώς διαφοροποιημένο τρόπο υλοποιησης. Οι Κατανεμημένες Μηχανές Επεξεργασίας αντιμετωπίζουν τις ανάγκες απόκτησης,επεξεργασίας, φιλτράρισματος, αναζήτησης, μοντελοποίησης, εξαγωγής και αρχειοθέτησης μεγάλου όγκου δεδομένων στις υποδομές Big Data. Πολλές από τις σύγχρονες εφαρμογές Κατανεμημένων Μηχανών Επεξεργασίας προσπαθούν να εκμεταλλευτούν το υπολογιστικό μοντέλο SIMD (Single Instruction Multiple Datasets) σε επίπεδο υλικού στοχεύοντας στην εκτέλεση της ίδιας ακολουθίας εντολών ταυτόχρονα σε μεγάλο αριθμό διακεκριμένων συνόλων δεδομένων. Αλλο βασικό χαρακτηριστικό των Μηχανών Κατανεμημένης Επεξεργασίας είναι η δυνατότητα επανεκκίνησης μιας διαδικασίας στον ίδιο ή σε διαφορετικό κόμβο όταν μια συγκεκριμένη εκτελούμενη διαδικασία αποτυγχάνει για οποιοδήποτε λόγο.

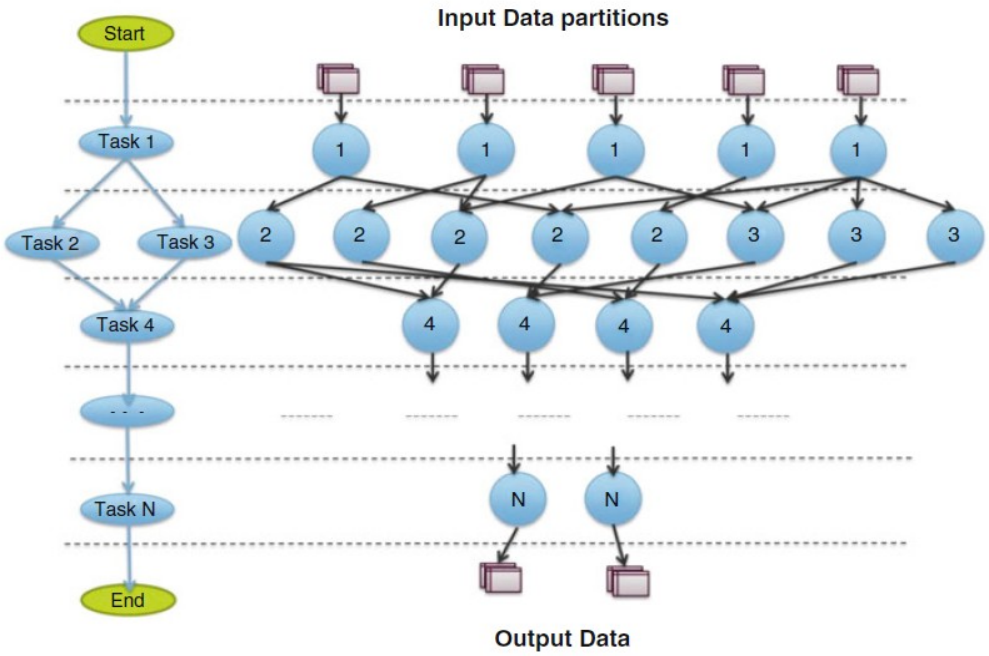
Παρακάτω ακολουθούν τα τυπικά μοτίβα που χρησιμοποιούνται ως Κατανεμημένες Μηχανές Επεξεργασίας από διαφορετικές τεχνολογίες Big Data.

Κατευθυνόμενου Ακυκλικού Γράφου
Directed Acyclic Graph (DAG) Distributed Processing Engines

Στην προσέγγιση Κατευθυνόμενης Κατανεμημένης Επεξεργασίας Κατευθυνόμενου Ακυκλικού Γράφου (DAG) κάθε εργασία χωρίζεται σε ένα αυθαίρετο σύνολο καθηκόντων. Κάθε κορυφή αντιπροσωπεύει μια εργασία που πρέπει να εκτελεστεί στα δεδομένα και κάθε ακμή αντιπροσωπεύει τη ροή δεδομένων μεταξύ των συνδεδεμένων κορυφών. Οι κορυφές μπορούν να έχουν οποιοδήποτε αυθαίρετο αριθμό εισόδων και εξόδων από ακμές. Πολλές

κορυφές μπορούν να εκτελέσουν την ίδια εργασία αλλά σε διαφορετικά μέρη του ίδιου υποσύνολου δεδομένων. Στο χρόνο εκτέλεσης, οι κορυφές γίνονται διαδικασίες που εκτελούν την εργασία και οι ακμές χρησιμοποιούνται για τη μεταφορά μιας πεπερασμένης ακολουθίας εγγραφών μεταξύ των κορυφών. Οι φυσικές υλοποιήσεις των ακμών τυπικά πραγματοποιούνται με κοινόχρηστη μνήμη, διάυλους TCP ή δίσκους. Κάθε εργασία εκτελείται παράλληλα στα δεδομένα που είναι αποθηκευμένα σε διάφορους κόμβους μεταφέροντας τις απαραίτητες λειτουργίες στους αντίστοιχους κόμβους δεδομένων.

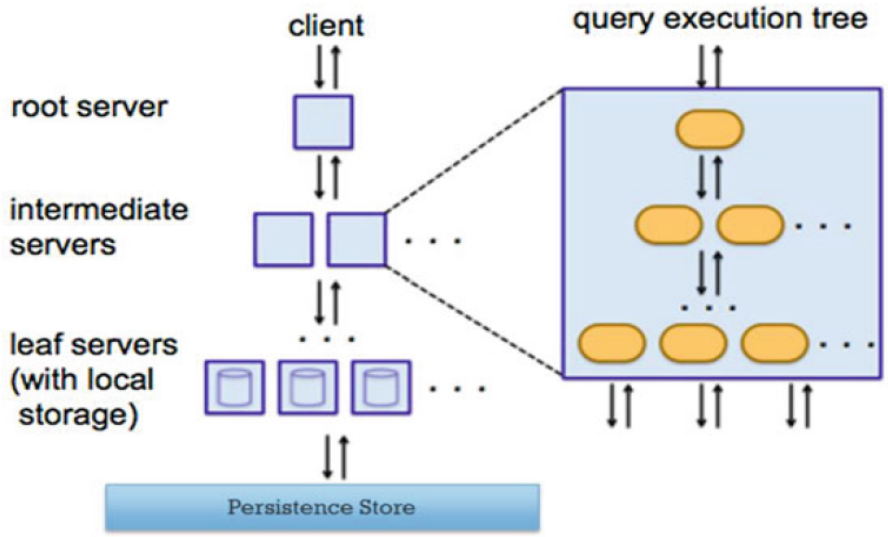
Παρακάτω παρατίθεται ένα παράδειγμα που περιγράφει λεπτομερώς τα βήματα εκτέλεσης που εμπλέκονται σε αυτή την προσεγγιση. Εδώ, τα δεδομένα εισόδου χωρίζονται σε πέντε κόμβους δεδομένων. Σε κάθε έναν από αυτούς τους κόμβους δεδομένων εκτελείται η εργασία υπ'αριθμ.1 και οι έξοδοί της από διανεμονται σε διαφορετικούς κόμβους. Στη συνέχεια εκτελούνται οι εργασίες υπ'αριθμ.2 και υπ'αριθμ.3 σε 8 κόμβους δεδομένων και οι εξόδοι από αυτές κατανέμονται περαιτέρω σε 4 κόμβους δεδομένων.Η εργασία υπ'αριθμ.4 εκτελείται τώρα σε τέσσερις κόμβους δεδομένων. Αυτή η διαδικασία συνεχίζεται μέχρι τη τελευταία εργασία, Εργασία Ν.



Εικόνα 8.Σχηματική Αναπαράσταση Επεξεργασίας DAG

Επεξεργασία βάσει Πολυεπίπεδων Δέντρων Εξυπηρέτησης MLST-Multi Level Serving Tree Based Distributed Processing

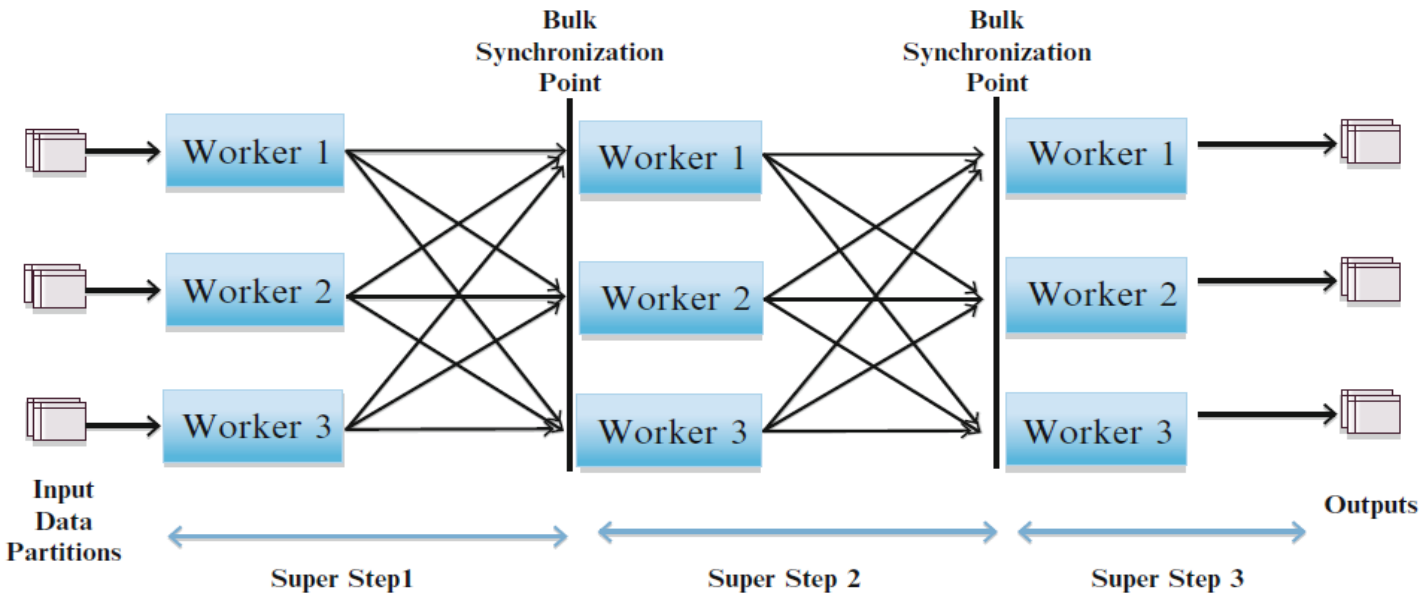
Η προσέγγιση που βασίζεται σε δέντρο πολλαπλών επιπέδων (δημοφιλές στο Google Dremel) χρησιμοποιεί την έννοια ενός δένδρου εξυπηρέτησης με πολλαπλά επίπεδα για την εκτέλεση μιας εργασίας. Όπως φαίνεται στην εικόνα όταν ένας ριζικός διακομιστής λαμβάνει ένα εισερχόμενο ερώτημα από έναν πελάτη, ξαναγράφει το ερώτημα σε κατάλληλες υποερωτήσεις με βάση πληροφορίες μεταδεδομένων και στη συνέχεια δρομολογεί-υποβιβάζει στο επόμενο επίπεδο του δέντρο εξυπηρέτησης. Κάθε επίπεδο εξυπηρέτησης εκτελεί μια παρόμοια επανεγγραφή και αλλαγή δρομολόγησης. Τελικά, τα υποερωτήματα φτάνουν στους διακομιστές φύλλων,που επικοινωνούν με το επίπεδο αποθήκευσης ή έχουν απευθείας πρόσβαση στα δεδομένα μόνιμης αποθήκευσης.Στο δρόμο προς τα πάνω, οι ενδιάμεσοι διακομιστές εκτελούν μια παράλληλη συνάθροιση των ενδιάμεσων αποτελέσματων έως ότου το επιθυμητό αποτέλεσμα του ερωτήματος να προωθηθεί ξανά στον διακομιστή ρίζας.



Εικόνα 9.Σχηματική Αναπαράσταση Επεξεργασίας MLST

Μαζική Σύγχρονη Παράλληλη Κατανεμημένη Επεξεργασία
Bulk Synchronous Parallel (BSP) Based Distributed Processing

Η προσέγγιση βασισμένη στην Μαζική Συγχρονισμένη Παράλληλη (BSP) Επεξεργασία χρησιμοποιεί την πιο γενική μορφή κατευθυνόμενου γράφου με κύκλους. Τα δεδομένα εισόδου είναι τα πρώτα που χωρίζονται χρησιμοποιώντας κατάλληλες τεχνικές κατανομής γραφημάτων σε πολλαπλούς κόμβους δεδομένων(**Data Nodes**). Στη συνέχεια, ολόκληρος ο απαιτούμενος φόρτος επεξεργασίας για τη μετατροπή των κατατμήσεων εισόδου σε τελικές εξόδους χωρίζονται σε ένα σύνολο **Supersteps** (ή επαναλήψεις). Σε ένα Superstep έκαστος εργαζόμενος(**Worker)**, που τρέχει σε έναν συγκεκριμένο κόμβο δεδομένων και αντιπροσωπεύει μια κορυφή, εκτελεί μια δεδομένη εργασία (μέρος ενός συνολικού αλγορίθμου) για την διαμέρηση δεδομένων που είναι διαθέσιμη σε αυτόν τον κόμβο.Αφού τελειώσουν όλοι οι εργαζόμενοι με το έργο τους, λαμβάνει χώρα μαζίκός συγχρονισμός των εξόδων που συλλέχθηκαν από κάθε worker. Στο τέλος του Superstep μια κορυφή μπορεί να τροποποιήσει την κατάσταση της ή εκείνη των εξερχόμενων στις ακμές του. Μια κορυφή μπορεί επίσης να λάβει μηνύματα που της έχουν αποστέλλεί από το προηγούμενο Superstep, μπορεί επίσης να στείλει μηνύματα σε άλλες κορυφές (για να ληφθούν στο επόμενο Superstep), ή ακόμα και ολόκληρη η τοπολογία του γράφου μπορεί να μεταβληθεί. Αυτή η διαδικασία επαναλαμβάνεται για τα υπόλοιπα Supersteps ως και το τέλος της παραγωγής.

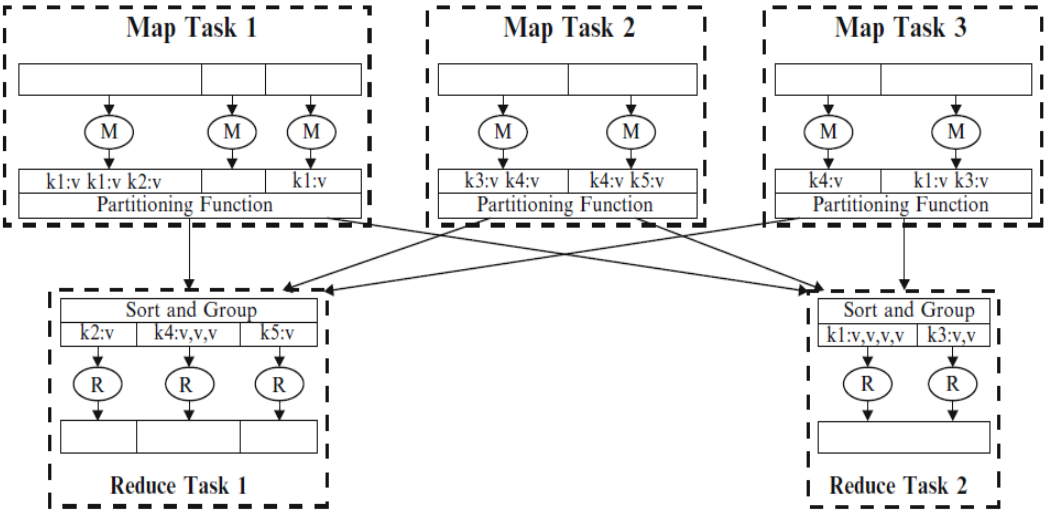


Εικόνα 10.Σχηματική Αναπαράσταση Επεξεργασίας BSP

Map Reduce (MR) Based Distributed Processing Engines

Είναι μέχρι στιγμής μία από τις πιο δημοφιλείς και επιτυχημένες προσεγγίσεις στην Κατανεμημένη Επεξεργασία. Η Google παρουσίασε το Map Reduce στην εφαρμογή επεξεργασίας κατανεμημένων δεδομένων που ήταν εμπνευσμένη από το Map Reduce ως βασική υπολογιστική λειτουργία.Προσφέρει δύο λειτουργίες, συγκεκριμένα την Αντιστοίχηση (**Map**) και την Απομείωση(**Reduce**), οι οποίες μπορούν να υλοποιηθούν ως γενικός τρόπος αλλά και ως αφηρημενη λογική επεξεργασίας και να χρησιμοποιηθούν για επεξεργασία οποιουδήποτε μεγάλου όγκου δεδομένων. Η διεπαφή του Map ,δέχεται ως είσοδο οποιαδήποτε σύνολο δεδομένων που αντιπροσωπεύεται ως ζεύγος κλειδιών/τιμών στα οποία συνήθως εφαρμόζονται κλιμακωτοί μετασχηματισμοί (σε επίπεδο εγγραφής). Οι εξόδοι από ένα Map ομαδοποιούνται ανά κλειδί. Έπειτα τα κλειδιά αυτά και ο προκύπτων πίνακας τιμών που συνδέεται με κάθε κλειδί ταξινομούνται, διαμερίζονται (σε partitions) και στη συνέχεια στέλνωνται στους απομειωτές(Reducers) ως λίστα των κλειδιών προς επεξεργασία. Η διεπαφή του Reducer υλοποιείται συνήθως για να επιτευχθούν πράξεις συνόλων επι της λίστας τιμών του κάθε κλειδιού. Προαιρετικά μπορεί να υπάρχει και μια διεπαφή Συνδυαστή (**Combiner**) που θα κάνει την σύνδεση των ίδιων κλειδιών μιας διαδικασίας Map προκειμένου να εκτελεστεί στο πρωταρχικό επίπεδο μείωσης.

Το σχήμα απεικονίζει τα βήματα που σχετίζονται με την τυπική διαδικασία Map-Reduce. Πρέπει να σημειωθεί ότι το Map Reduce μπορεί να θεωρηθεί ως μια συγκεκριμένη εφαρμογή της DAG κατανεμημένης επεξεργασίας. Απλώς στο Map Reduce η εργασία που εκτελείται από κάθε κορυφή σε κάθε στάδιο του DAG είναι πάντοτε μια λογική Map ή Reduce αντί για οποιαδήποτε αυθαίρετη λειτουργία. Οι περισσότερες πλατφόρμες Big Data παρέχουν ένα Map Reduce interface όπου ο τελικός χρήστης μπορεί να εκφράσει μια δεδομένη απαίτηση επεξεργασίας δεδομένων. Ωστόσο, τα περισσότερα από αυτά χρησιμοποιούν επιπλέον εσωτερικά τις προσεγγίσεις DAG ή BSP υλοποιώντας ως Map η Reduce της εργασίες των κορυφών ενός DAG ή ενός Superstep στο BSP.



Εικόνα 11.Σχηματική Αναπαράσταση Επεξεργασίας Map-Reduce

Νηματικές Διαδικασίες Μακράς Διάρκειας
Long Running Shard Processes Based Distributed Processing Engine

Χρησιμοποιείται για την εξυπηρέτηση κυρίως εφαρμογών υψηλής συχνότητας προσπελάσεων Read/Write και αναζήτησης. Αυτό αποτελεί μια ειδική περίπτωση προσέγγισης βασισμένη στην MLST όπου αντί για πολλαπλά επίπεδα υπάρχουν μόνο δύο επίπεδα. Σε αυτό το μοντέλο συνήθως πολλαπλές διαδικασίες τύπου “δαίμονα” (**slave daemons**) εκτελούνται σε διάφορα Data Nodes (που ισοδυναμούν με τους κόμβους των φύλλων). Κάθε υποτελής διαδικασία παίρνει στην ιδιοκτησία μιας διαμέριση από την ολότητα των δεδομένων και την υποχρέωση για εξυπηρέτηση των αιτημάτων-ερωτημάτων όταν απευθύνονται προς αυτά. Η διαδικασία master (διακομιστής root) εκτελείται στον διαχείριστικό κόμβο ο οποίος συνήθως αυτό που κάνει είναι καταγράφει τον σωστό κόμβο δεδομένων όπου βρίσκονται τα κάθε φορά επιθυμητά δεδομένα. Όταν ένα αίτημα καταχωρείται από τον πελάτη, η κύρια διαδικασία αναστέλλει πρώτα το ίδιο και στη συνέχεια στέλνει το αίτημα στο κατάλληλο νήμα. Κάποια στιγμή, μόλις το κατάλληλο νήμα εντοπιστεί η κύρια διαδικασία σταδιακά οδηγεί τον πελάτη προς σύνδεση με το νημα και καθώς περνάει η επεξεργασία αργότερα όλη η αλληλεπίδραση συμβαίνει απευθείας μεταξύ του πελάτη και του shard.

Δίκτυο Πρακτόρων τύπου Παραγωγών – Καταναλωτών
Producer-Consumer Agent Network Based Distributed Processing Engines

Χρησιμοποιούνται κυρίως στην επεξεργασία δεδομένων ροής. Σε αυτό το μοντέλο αναπτύσσονται πολλαπλοί πράκτορες όπου κάθε ένας από αυτούς ταυτίζεται με ένα ρεύμα γεγονότων/δεδομένων, το επεξεργάζεται και τελικά προωθεί σε άλλους πράκτορες. Αυτοί οι πράκτορες είναι συνήθως διασυνδεδεμένοι μεταξύ τους για μπορέσουν να εφαρμόσουν τα πολλαπλά στάδια που απαιτούνται. Τυπικά, το σημείο εκκίνησης του δικτύου πρακτόρων είναι μια πηγή δεδομένων (όπως μια θύρα tcp ή θύρα http ή ουρά μηνυμάτων ή το σύστημα αρχείων κ.λπ.) και το τελικό σημείο είναι μια άλλη τεχνολογία Big Data(όπως Hadoop, NoSQL Databases, Berkeley Data Analytics Stack κ.λπ.) . Πολλαπλοί κλώνοι των πρακτόρων εφαρμόζονται στα διάφορα Data Nodes για να χειριστούν την επεκτασιμότητα και τη διαθεσιμότητα. Αυτό το μοντέλο είναι ουσιαστικά μια παραλλαγή του DAG με τις ακόλουθες διαφορές:

-Στην περίπτωση του Δικτύου Πρακτόρων Παραγωγών-Καταναλωτων, η DAG για τα βήματα επεξεργασίας αποφασίζεται κατά το σχεδιασμό από τον κύριο του έργου (σχεδιαστή)

-Η δεύτερη διαφορά είναι ότι η μετακίνηση δεδομένων από μία εργασία σε άλλη συμβαίνει άμεσα στα πλάισια των αρμοδιοτήτων του παραγωγού-καταναλωτή χωρίς να χρειάζεται μια κύρια διαδικασία διαιτησίας

2.5.1.2 Application Components | Συστατικά Στοιχεία Εφαρμογών

Χρησιμοποιούν μία ή περισσότερες Μηχανές Κατανεμημένης Επεξεργασίας για να εξυπηρετησουν ένα αίτημα που προέρχεται από τον πελάτη. Είναι συνήθως η γέφυρα μεταξύ της πραγματικής εφαρμογής του χρήστη και της κατανεμημένης μηχανής. Συνήθως υποστηρίζει μια συγκεκριμένη διεπαφή πρόσβασης δεδομένων (που αναλύεται παρακάτω). Η εφαρμογή πελάτη χρησιμοποιεί μια Διασύνδεση Πρόσβασης Δεδομένων που υποστηρίζει το Στοιχείο Εφαρμογής προκειμένου να έχει πρόσβαση ή να επεξεργαστεί δεδομένα. Το στοιχείο εφαρμογής μεταφράζει το αίτημα του πελάτη σε κατάλληλες κλήσεις υπηρεσιών στην κατανεμημένη μηχανή επεξεργασίας.Επίσης συνήθως διαχειρίζονται τις ταυτόχρονες αιτήσεις από τον πελάτη, την ασφάλεια, την διαθέσιμη χωριτικότητα σε συνδέσεις κλπ. Αυτά τα στοιχεία συνήθως υλοποιούνται υπό την μορφή μακροχρόνιων διεργασιών (daemons) έτσι ώστε να εξυπηρετούν διαρκώς τα εισερχόμενα αιτήματα. Σε ορισμένες περιπτώσεις δε, η υλοποίηση μπορεί να είναι μια προγραμματισμένη διαδικασία που τρέχει σε συγκεκριμένες χρονικές περιόδους.

2.5.1.3 Data Access Interfaces | Διεπαφές Πρόσβασης Δεδομένων

Η γλώσσα SQL είναι αναμφισβήτητα το ευρύτερα χρησιμοποιούμενο εργαλείο πρόσβασης δεδομένων και το πιο δημοφιλές στον κλάδο τις τελευταίες δεκαετίες. Συνεπώς, παραμένει η προτιμώμενη διεπαφή πρόσβασης δεδομένων ακόμη και για τις τεχνολογίες Big Data. Ωστόσο, η υποστήριξης της SQL ποικίλλει από την μια τεχνολογία σε άλλη. Εκτός από τις διεπαφές βασισμένες σε SQL, οι τεχνολογίες Big Data συνήθως υποστηρίζουν λειτουργίες read/write μέσω API σε διάφορες γλώσσες προγραμματισμού όπως Java, Python, C++, Scala κ.λπ.

Πολλές φορές οι τεχνολογίες Big Data με κύρια υποστήριξη διασύνδεσης την SQL, παρέχεουν καθορισμένες από το χρήστη συναρτήσεις (User Defined Functions -UDF) ωστε να επιτύχουν επεξεργασία μη σχεσιακή. Αυτά τα UDF τυπικά γράφονται σε γλώσσες προγραμματισμού όπως C, C++ Java κ.λπ. και διατίθενται ως εκτελέσιμη βιβλιοθήκη εντός του μηχανισμού επεξεργασίας SQL. Ο χρήστης ενσωματώνει αυτά τα UDF μέσα στο ίδιο το ερώτημα SQL και το SQL query engine εκτελεί τις συγκεκριμένες βιβλιοθήκες του UDF σε κατάλληλα σημεία εντός του πλάνου εκτέλεσης.Επίσης συνηθισμένη είναι και η αντίστροφη προσεγγιση δηλαδή η κλήση SQL τμημάτων εντός του σώματος του API κώδικα. Μερικές όμως απο τις τεχνολογίες Big Data όπως οι βάσεις δεδομένων NoSQL, οι τεχνολογίες επεξεργασίας συμβάντων ροής και οι τεχνολογίες αναζήτησης δεν υποστηρίζουν καθόλου διεπαφή SQL για πρόσβαση ή / και επεξεργασία δεδομένων. Υποστηρίζουν μόνο συγκεκριμένο API σε γλώσσες προγραμματισμού για σκοπούς την ανάγνωση / εγγραφή / επεξεργασία δεδομένων.

Τέλος ,πολλές από τις τεχνολογίες Big Data υποστηρίζουν τη δημιουργία και τη χρήση διαφόρων μοντέλων προβλέψεων (Predictive Models). Αυτά τα πιθανοτικά μοντέλα (όπου τα δεδομένα προσαρμόζονται σε ένα μοντέλο που αντιπροσωπεύει κάποια φυσική συμπεριφορά) ή μοντέλα Machine Learning (όπου το μοντέλο αναδύεται από τα δεδομένα). Τυπικά η δημιουργία αυτών των μοντέλων γίνεται με χρήση API που διατίθενται σε διαφορετικές γλώσσες προγραμματισμού όπως το C++, Java, Python, Scala, R, etc.

Οι περισσότερες από τις τεχνολογίες Big Data κάνουν εφικτή την ανάγνωση/εγγράφή μέσα από τις γενικευμένες(δηλαδή ανεξάρτητες από πλατφόρμα) διεπαφές, συγκεκριμένα το Apache Thrift και το REST. Η διασύνδεση Apache Thrift συνδυάζει μια στοίβα λογισμικού με μια μηχανή δημιουργίας κώδικα για την κατασκευή υπηρεσιών που λειτουργούν το ίδιο αποτελεσματικά και απρόσκοπτα σε διάφορες γλώσσες προγραμματισμού όπως C++, Java, Python, PHP, Ruby, Erlang, Perl κ.ά. Η διασύνδεση REST παρέχει πρόσβαση σε δεδομένα

ανεξάρτητα από κάθε τεχνολογία μέσω του πρωτοκόλλου http. Οι διασυνδέσεις Thrift και REST παρέχουν τεράστια ευελιξία στα Big Data ώστε να διασυνδέονται με άλλα εργαλεία και τεχνολογίες μιας επιχείρησης.

2.5.1.4 Data Privacy | Ασφαλεια & Ιδιωτικότητα Δεδομένων

Όλες οι τεχνολογίες Big Data παρέχουν μέτρα για την προστασία της ιδιωτικότητας και της ασφάλειας των δεδομένων σε μια ποικιλία θεμάτων. Τα κυριότερα χαρακτηριστικά που καλύπτονται είναι η αυθεντικοποίηση, ο έλεγχος εξουσιοδότησης ανά ρόλους, κρυπτογράφηση δεδομένων όταν μεταφέρονται καθώς και όταν βρίσκονται σε κατάσταση ηρεμίας, καθως και δραστηριοτήτες ελέγχου που σχετίζονται με την πρόσβαση σε δεδομένα.

Για τους σκοπούς του ελέγχου ταυτότητας και την εξουσιοδότηση βάσει ρόλων συνήθως γίνεται διασύνδεση με το κεντρικό σύστημα καταλόγου της επιχείρησης (LDAP) . Για κρυπτογράφηση δεδομένων που μεταφέρονται μεταξύ του πελάτη και του κορμού της Big Data εφαρμογής τοποθετείται στο ενδιάμεσο στρώμα ασφαλούς υποδοχής στοιχείων. Το πρωτόκολλο Kerberos χρησιμοποιείται επίσης σε πολλές περιπτώσεις (το οποίο λειτουργει βάσει «εισιτηρίων» συνόδου) ώστε να επιτραπεί στους κόμβους που επικοινωνούν μέσω ενός μη ασφαλούς δικτύου να αποδείξουν την ταυτότητά αναμεταξύ τους με πλήρως διασφαλισμένο τρόπο.Οι διάφορες Τεχνολογίες Big Data τυπικά παρέχουν επίσης εξειδικευμένη μεθοδολογία για την κρυπτογράφηση δεδομένων η οποία είναι συμβατή με οποιοδήποτε πρότυπο κρυπτογράφησης (όπως πχ κρυπτογράφηση 128 bit ή 256 bit). Η κρυπτογράφηση των τοπικά αποθηκευμένων δεδομένων είναι απαραίτητη για τους κανονισμούς ιδιωτικότητας και ασφάλειας που αφορούν διάφορους κλάδους (HIPAA για Υγειονομική Περίθαλψη, PCI DSS για πληρωμες με κάρτα κ.λπ.). Για τις δυνατότητες ελέγχου υπάρχουν συγκεκριμένες προσεγγίσεις τεχνολογίας οι οποίες συνήθως ακολουθούν τις έννοιες του Aspect Oriented Programming - AOP

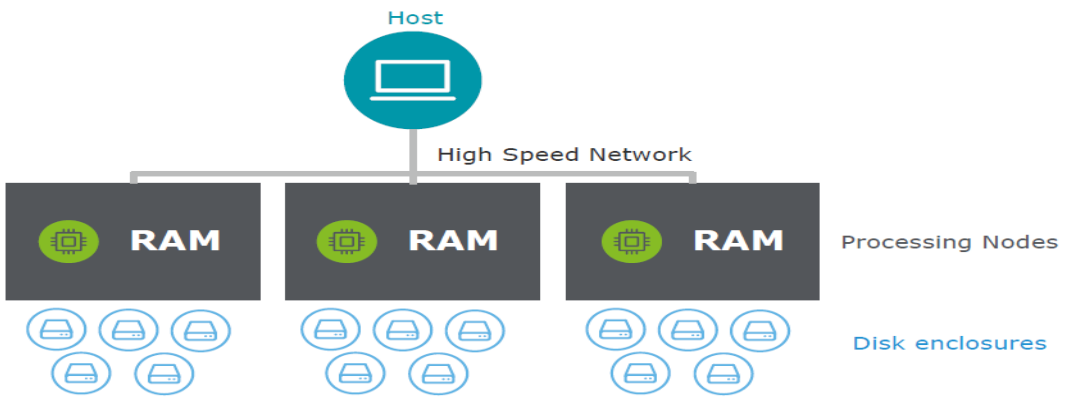
2.5.2 Μαζική Παράλληλη Επεξεργασία | Massively Parallel Processing

[36. Bartosz Konieczny]

Ένα σύστημα μαζικής παράλληλης επεξεργασίας (MPP) αποτελείται από μεγάλο αριθμό μικρών ομοιογενών κόμβων επεξεργασίας που διασυνδέονται μέσω ενός δικτύου υψηλής ταχύτητας. Οι κόμβοι επεξεργασίας σε ένα μηχάνημα MPP είναι ανεξάρτητοι - συνήθως δεν μοιράζονται μνήμη και συνήθως κάθε επεξεργαστής μπορεί να κατέχει τη δική του παρουσία ενός λειτουργικού συστήματος.Συχνά υπάρχουν εφαρμογές συστημικού ελεγκτή που φιλοξενούνται σε κόμβους επεξεργασίας με ρόλο ηγέτη που έχουν ως αποστολή να καθοδηγούν τους μεμονωμένους κόμβους επεξεργασίας όλης της δομής MPP σχετικά με τις εργασίες που πρέπει να εκτελεστούν.

Οι κόμβοι σε μηχανήματα MPP μπορούν επίσης να συνδεθούν απευθείας με τις δικές τους συσκευές I/O ή το I/ O μπορεί να διοχετεύεται σε ολόκληρο το σύστημα μέσω διασυνδέσεων υψηλής ταχύτητας. Η επικοινωνία μεταξύ κόμβων είναι πιθανό να πραγματοποιηθεί με συντονισμένο τρόπο, όπου όλοι οι κόμβοι σταματούν την επεξεργασία και συμμετέχουν σε ανταλλαγή δεδομένων σε ολόκληρο το δίκτυο ή με μη συντονισμένο τρόπο, με μηνύματα που στοχεύουν συγκεκριμένους παραλήπτες να εγχέονται ανεξάρτητα στο δίκτυο.

Επειδή τα δεδομένα μπορούν να μεταδοθούν μέσω του δικτύου και να στοχεύουν συγκεκριμένους κόμβους, μια μηχανή MPP είναι κατάλληλη για εφαρμογές παράλληλης επεξεργασίας δεδομένων. Σε αυτήν την περίπτωση, όλοι οι επεξεργαστές εκτελούν το ίδιο πρόγραμμα σε διαφορετικές ροές δεδομένων. Επιπλέον, επειδή οι μεμονωμένοι διαφορετικοί επεξεργαστές μπορούν να εκτελουν διαφορετικά προγράμματα, μια μηχανή MPP είναι κατάλληλη για χονδροειδείς παραλληλισμούς(coarse-grained parallelism) και μπορεί επίσης να παραμετροποιηθεί για εκτέλεση pipelined ροών.



Εικόνα 12.Σχηματική Αναπαράσταση Δομής MPP

2.5.3 Επεξεργασία στην Μνήμη | In-Memory Processing Solutions

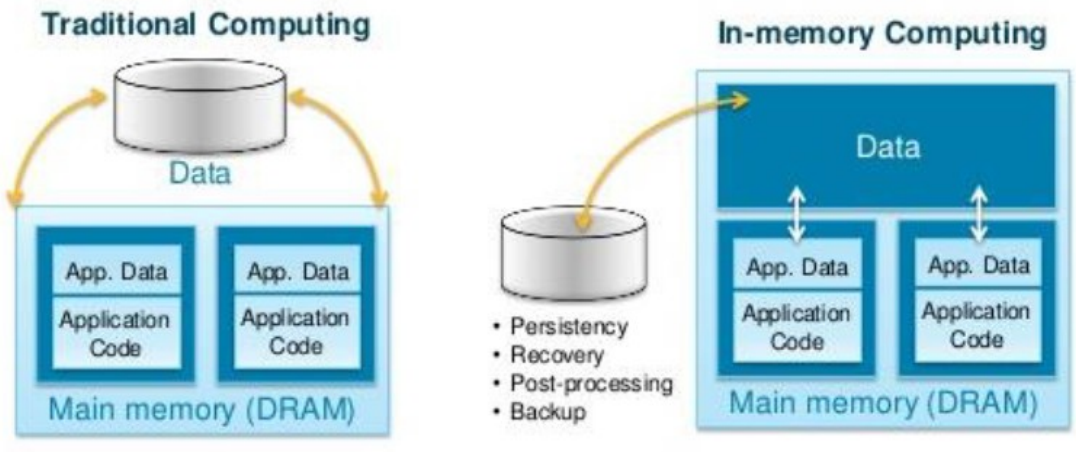
Οι τεχνολογίες αποθήκευσης εντός μνήμης, πολλές από τις οποίες αναπτύχθηκαν για να υποστηρίξουν τον επιστημονικό τομέα HPC-High Performance Computing, χρησιμοποιούνται ολοένα και περισσότερο λόγω της σημαντικής μείωσης του κόστους της μνήμης και των αυξημένων δυνατοτήτων επέκτασης στους σύγχρονους διακομιστές και λειτουργικά συστήματα. Ωστόσο,για κάθε στοιχείο μιας υποδομής που είναι προσανατολισμένη στην ταχύτητα απαιτείται κάτι περισσότερο από απλά διευρημένη Μνήμη Τυχαίας Προσπέλασης (**RAM**). Δηλαδή απαιτούνται επίσης βελτιστοποιημένες δομές δεδομένων και αλγορίθμων πρόσβασης στην μνήμη ωστε να είναι δυνατή η πλήρης εκμετάλλευση της απόδοσης RAM. Οι τρέχουσες προσφορές βάσεων δεδομένων στη μνήμη (**in-memory databases**) έχουν αρχίσει να αντιμετωπίζουν αυτό το ζήτημα. Οι λύσεις διαμοιραζόμενης μνήμης (**shared memory**) που είναι συνηθισμένες σε περιβάλλοντα HPC συνήθως εφαρμόζονται για την αντιμετώπιση των αναγκών ενδοεπικοινωνίας του συστήματος και των απαιτήσεων συγχρονισμού.

Οι παραδοσιακές αρχιτεκτονικές διαχείρισης βάσεων δεδομένων έχουν σχεδιαστεί για τη χρήση των περιστρεφόμενων δίσκων ως πρωτεύουσα αποθήκευση με την κύρια μνήμη του υπολογιστικού περιβάλλοντος να υποβιβάζεται στο ρόλο της απλής παροχής προσωρινής αποθήκευσης δεδομένων και ευρετήριων. Πολλοί από αυτούς τους μηχανισμούς αποθήκευσης εντός της μνήμης έχουν τις ρίζες τους στις μεθόδους τις Μαζικής Παράλληλης Επεξεργασίας(Massive Parallel Processing) που είναι δημοφιλείς στην επιστημονική κοινότητα.

Αυτές οι προσεγγίσεις δεν πρέπει να συγχέονται με δίσκους σταθερής κατάστασης (π.χ. flash) ή συστήματα κλιμακωτής αποθήκευσης (tiered storage) που υλοποιούν αποθήκευση βασιζόμενη στην μνήμη η οποία απλώς αναπαράγει τις διεπαφές όμοιες με απλού δίσκου αλλά σε ταχύτερο μέσο αποθήκευσης. Τα πραγματικά συστήματα αποθήκευσης εντός της μνήμης συνήθως αποφεύγουν το κόστος της σημασιολογίας των αρχείων συστήματος και βελτιστοποιούν τις δομές αποθήκευσης δεδομένων, ώστε να ελαχιστοποιηθεί η χρήση μνήμης και να μεγιστοποιηθεί το ποσοστό πρόσβασης στα δεδομένα. Αυτά τα συστήματα μνήμης μπορούν να υλοποιήσουν σχεσιακή οργάνωση γενικού σκοπού αλλα και NoSQL οργάνωση και διεπαφές ή να είναι πλήρως βελτιστοποιημένα σε συγκεκριμένο πρόβλημα και δομή δεδομένων.

Όπως τα παραδοσιακά συστήματα Big Data που βασίζονται σε δίσκους, αυτές οι υλοποιήσεις υποστηρίζουν συχνά οριζόντια κατανομή δεδομένων και επεξεργασία σε πολλαπλούς αυτόνομους κόμβους αν και οι τεχνολογίες διαμοιραζόμενης μνήμης εξακολουθούν να επικρατούν σε εξειδικευμένες εφαρμογές. Σε αντίθεση με τις παραδοσιακές μεθόδους δίσκου, οι in-memory λύσεις και οι υποστηριζόμενες εφαρμογές πρέπει να χειρίζονται και το γεγονός της μη διατήρησης των δεδομένων σε περιπτώσεις αποτυχίας του συστήματος. Ορισμένες εφαρμογές αξιοποιούν μια υβριδική προσέγγιση που περιλαμβάνει την σποραδική μόνιμη εγγραφή-αποθήκευση για υποβοήθηση στην αντιμετώπιση του προβλήματος.

Τα πλεονεκτήματα των προσεγγίσεων in-memory περιλαμβάνουν ταχύτερη επεξεργασία σε φορτία εργασίας απαιτητικής ανάλυσης και reporting. Είναι ιδιαίτερα χρήσιμα για την ανάλυση δεδομένων σε πραγματικό χρόνο, όπως για παράδειγμα κάποια περίπλοκη επεξεργασία συμβάντων (CEP) σε ρεύματα δεδομένων. Για φόρτο εργασίας που αφορά reporting η βελτίωση της απόδοσης μπορεί συχνά να είναι της τάξης αρκετών εκατοντάδων φορών πιο γρήγορη – ειδικά για αναλύσεις αραιής μήτρας και προσομοίωσεων.



Εικόνα 13.Σχηματική Αναπαράσταση Επεξεργασίας In-Memory

2.6 Διαχειριστικά Πλαίσια | Management Frameworks

[7. NIST]

2.6.1 Πλαίσια Μηνυμάτων / Επικοινωνιών | Messaging Frameworks

Τα πλαίσια μηνύματων και επικοινωνίας έχουν τις ρίζες τους στα περιβάλλοντα HPC που ως γνωστόν είναι πολύ δημοφιλή στις επιστημονικές και ερευνητικές κοινότητες. Αναπτύχθηκαν για να παρέχουν API για αξιόπιστο έλεγχο ουρών αναμονής (queuing), μεταφορά και λήψη δεδομένων μεταξύ κόμβων σε συστοιχίες οριζόντιας κλιμάκωσης. Αυτά τα πλαίσια στην αρχιτεκτονική τους τυπικά εφαρμόζουν είτε ένα μοντέλο μεταφοράς από σημείο σε σημείο (point-to-point) είτε ένα μοντέλο μεταφοράς προσωρινής αποθήκευσης και προώθησης (store-and-forward)

Σε ένα μοντέλο από σημείο σε σημείο **(point-to-point)**, τα δεδομένα μεταφέρονται απευθείας από αποστολέα στους δέκτες. Η πλειονότητα των υλοποιήσεων από σημείο σε σημείο δεν περιέχει καμία μορφή πρόβλεψης για ανάκτηση μηνυμάτων σε περιπτώσεις κατάρρευσης προγράμματος ή διακοπής της σύνδεσης μεταξύ αποστολέα και δέκτη.Τυπικά εφαρμόζουν όλη την λογική εντός του χώρου προγραμμάτων του αποστολέα και του δέκτη, συμπεριλαμβανομένων των εγγυήσεων παράδοσης ή των δυνατοτήτων επαναποστολής των μηνυμάτων. Μια κοινή παραλλαγή αυτού του μοντέλου είναι η πολυεκπομπή (δηλαδή, η μετάδοση από ένα σε πολλά ή από πολλά σε πολλά), η οποία επιτρέπει στον αποστολέα να μεταδίδει τα μηνύματα μέσω ενός καναλιού και οι δέκτες με τη σειρά του ακούν τα κανάλια που τους ενδιαφέρουν. Συνήθως, τα μηνύματα πολλαπλής διανομής δεν εφαρμόζουν καμία μορφή διασφάλισης της παράδοσης.

Με το μοντέλο προσωρινής αποθήκευσης και προώθησης **store-and-forward**, ο αποστολέας θα απευθύνει το μήνυμα σε ένα ή περισσότερους δέκτες αλλά θα το αποστείλει μέσο ενός παρεμβαλόμενου μεσίτη, ο οποίος θα αποθηκεύσει το μήνυμα και στη συνέχεια θα το προωθήσει στους δέκτες. Πολλές από αυτές τις εφαρμογές υποστηρίζουν κάποια μορφή διατήρησης για μηνύματα που δεν έχουν ακόμη παραδοθεί, παρέχοντας έτσι και την ανάκτησης σε περίπτωση βλάβης της διαδικασίας ή του συστήματος. Τα μηνύματα πολλαπλής διανομής (**multicast messaging**) μπορεί επίσης να εφαρμοστεί σε αυτό το μοντέλο και συχνά αναφέρεται ως **pub/sub** μοντέλο.

2.6.2 Πλαίσια Διαχείρισης Πόρων | Resource Management Frameworks

Καθώς τα συστήματα Big Data έχουν εξελιχθεί και τείνουν να γίνουν πιο περίπλοκα,οι επιχειρήσεις εργάζονται για να χειριστούν την περιορισμένη διαθεσιμότητα σε υπολογιστικά και αποθηκευτικά μέσα που είναι ικανά να αντιμετώπισουν ένα ευρύτερο φάσμα εφαρμογών και επιχειρηματικών προκλήσεων,η απαίτηση της αποτελεσματικής διαχείρισης των πόρων αυτών έχει αυξηθεί σημαντικά. Ενώ όμως τα εργαλεία διαχείρισης πόρων και ο ελαστικής υπολογιστικής “elastic computing” έχουν επεκταθεί και ωριμάσει ανταποκρινόμενα στις ανάγκες των παρόχων λύσεων cloud και των τεχνολογιών εικονικοποίησης, τα Big Data έχουν εισάγει σημαντικά διαφοροποιημένες απαιτήσεις για αυτά τα εργαλεία καθώς απαιτούν περισσότερο το κατανεμημένο υπολογιστικό πρότυπο, το οποίο παρουσιάζει πρόσθετες δυσκολίες.

Τα χαρακτηριστικά μεγάλου όγκου δεδομένων και της ταχύτητας οδηγούν τις απαιτήσεις σε σχέση με τη διαχείριση πόρων στα συστήματα Big Data. Η ελαστική υπολογιστική είναι η πιο κοινή προσέγγιση για την αντιμετώπιση της επέκτασης του όγκου ή ταχύτητας των δεδομένων που εισέρχονται στο σύστημα.Η CPU και η μνήμη είναι οι δύο πόροι που θεωρούνται οι πιο σημαντικοί για τη διαχείριση καταστάσεων Big Data και οι ελλείψεις ή η υπερβολική κατανομή σε οποιοδήποτε από αυτα τα δύο θα έχουν σημαντικές επιπτώσεις στην απόδοση του συστήματος, ειδικά η ακατάλληλη ή αναποτελεσματική διαχείριση της μνήμης είναι συχνά καταστροφική. Τα Big Data διαφέρουν και γίνονται πιο περίπλοκα στην κατανομή των υπολογιστικών πόρων σε διαφορετικά πλαίσια αποθήκευσης ή επεξεργασίας που έχουν παραμετροποιηθεί για συγκεκριμένες εφαρμογές και δομές δεδομένων. Ως εκ τούτου, τα πλαίσια διαχείρισης πόρων χρησιμοποιούν συχνά την τοπικότητα των δεδομένων ως μία από τις μεταβλητές εισόδου για να αποφασίσουν που ακριβώς θα πρέπει να λάβουν χώρα νέα στοιχεία (π.χ.master nodes, processing nodes, job slots) που χρειάζεται να προστεθούν στο πλαίσιο επεξεργασίας . Ειδικά, επειδή τα δεδομένα είναι μεγάλα (δηλ.σε όγκο), δεν είναι γενικά αποδοτική η μεταφορά δεδομένων εντός των πλαίσια επεξεργασίας. Επιπλέον, ενώ σχεδόν όλα τα πλαίσια επεξεργασίας μεγάλων δεδομένων μπορούν να εκτελεστούν σε εικονικοποιημένα περιβάλλοντα, τα περισσότερα έχουν σχεδιαστεί για να λειτουργούν σε αποκλειστικής χρήσης εξειδικευμένο υλικό ικανό να παρέχει αποτελεσματική είσοδο/έξοδο.

Δύο ξεχωριστές προσεγγίσεις στη διαχείριση πόρων στα πλαίσια των Big Data ανακύπτουν. Το πρώτο είναι **εντός του πλαισίου διαχείριση πόρων**, όπου το ίδιο το πλαίσιο διαχειρίζεται την κατανομή των πόρων μεταξύ των διαφορετικών στοιχείων/τμημάτων του συστήματος. Αυτή η ανακατανομή καθοδηγείται συνήθως από το

φόρτο εργασίας του πλαισίου και συχνά επιδιώκει να απενεργοποιεί τους περιττούς πόρους ώστε να ελαχιστοποιηθούν οι συνολικές απαιτήσεις του πλαισίου στο σύστημα ή να ελαχιστοποιήσει το λειτουργικό κόστος του συστήματος μειώνοντας την κατανάλωση ενέργειας. Με αυτήν την προσέγγιση, οι εφαρμογές μπορούν να προγραμματίζουν και να αιτούνται πόρους που-όπως και τα λειτουργικά συστήματα των mainframe του παρελθόντος – διαχειρίζονται μέσω προκαθορισμένων ουρών και ομάδων εργασίας.

Η **δεύτερη προσέγγιση είναι η διαχείριση πόρων εκτός των διαφόρων πλαισίων**, η οποία έχει σχεδιαστεί για να ανταποκρίνεται στις ανάγκες πολλών συστημάτων Big Data για την υποστήριξη πολλαπλών πλαισίων αποθήκευσης και επεξεργασίας που μπορούν να παρουσιαστούν καθώς και να είναι βελτιστοποιημένο για ευρύ φάσμα εφαρμογών. Με αυτή την προσέγγιση, το πλαίσιο διαχείρισης των πόρων λειτουργεί ως υπηρεσία που υποστηρίζει και διαχειρίζεται αιτήματα πόρων από πλαίσια, παρακολουθεί τη χρήση πόρων πλαισίου και σε ορισμένες περιπτώσεις διαχειρίζεται τις ουρές των ίδιων εφαρμογών. Από πολλές απόψεις, αυτή η προσέγγιση μοιάζει με τα συνηθισμένα επίπεδα διαχείρισης πόρων που συναντούνται στα περιβάλλοντα cloud/virtualization. Παράλληλα γίνονται προσπάθειες προς τη δημιουργία υβριδικών πλαισίων διαχείρισης πόρων που θα είναι ικανά να χειρίζονται ταυτόχρονα τόσο τους φυσικούς όσο και τους εικονικούς πόρους.

Η περαιτέρω υιοθέτηση αυτών των εννοιών και ο συνδυασμός τους οδηγεί στις αναδυόμενες τεχνολογίες που αναπτύσσονται γύρω αυτό που ονομάζεται **Software Defined Data Center - SDDCs**. Αυτή η επέκταση σε ευέλικτη υπολογιστική και υπολογιστική νέφους αναπτύσσει περαιτέρω τη διαχείριση συγκεκριμένων ομάδων φυσικών υπολογιστικών πόρων ως εικονικών πόρων ώστε να περιλαμβάνει την αυτοματοποιημένη ανάπτυξη και πρόβλεψη χαρακτηριστικών και δυνατοτήτων σε φυσικούς πόρους. Για παράδειγμα, αυτοματοποιημένα εργαλεία ανάπτυξης που διασυνδέονται με virtualization ή άλλα APIs του πλαισίου μπορούν να χρησιμοποιηθούν για την αυτόματη ανάρτηση ολόκληρων clusters ή για την προσθήκη επιπλέον φυσικών πόρων σε φυσικά ή εικονικά clusters

Χειρισμός της Μεταβλητότητας Πόρων | Resource Volatility

[21. BARIKA, GARG, ZOMAYA, WANG, VAN MOORSEL, RAJIV, RANJAN]

Σε οποιοδήποτε περιβάλλον, υπάρχει πιθανότητα απώλειας αυτών πόρων ή της κατάστασης της διεργασίας ανάλυσης που εκτελείται από το πλαίσιο επεξεργασίας Big Data ανά πάσα στιγμή λόγω διάφορων αστοχιών. Ο μετριασμός αυτών των αποτυχιών πρέπει να πραγματοποιηθεί σε διαφορετικά επίπεδα όπως περιγράφονται παρακάτω:

Σε επίπεδο VM

Αυτή η προσέγγιση στοχεύει στον μετριασμό της αποτυχίας και της απώλειας της κατάσταση των εικονικών μηχανών σε όρους δεδομένων που είναι αποθηκευμένα στη μνήμη RAM και / ή τη μη μόνιμη αποθήκευση.

Τεχνικές:COLO, VM workload consolidation-based fault-tolerance technique, Hybrid adaptive checkpointing technique

Σε επίπεδο Πλαισίου Επεξεργασίας Big Data

Αυτή η προσέγγιση στοχεύει στον μετριασμό της αποτυχία και την απώλεια της κατάστασης των υπολογιστικών μονάδων / διαδικασιών κατά την επεξεργασία

Σε επίπεδο Ροής Εργασίας

Αυτή η προσέγγιση στοχεύει στον μετριασμό της αποτυχίας και της απώλειας εργασιών της ροής εργασίας, συμπεριλαμβανομένης της απώλειας του αναλυτικών υπολογισμών που έχει ολοκληρωθεί μέχρι τώρα η οποία ενδέχεται να επιφέρει πρόσθετο κόστος ή καθυστέρηση στην εκτέλεση.

2.6.3 Πλαίσια Παρακολούθησης | Monitoring Frameworks

Για την παρακολούθηση της κατανεμημένης και σύνθετης φύσης της υποδομής Big Data, η διαχείριση του συστήματος βασίζεται στα εξής:

• Τυποποιημένα Πρωτόκολλα(**standard protocols**), όπως το Simple Network Management Protocol (SNMP), το οποίο χρησιμοποιείται για την παρακολούθηση της κατάσταση μετάδοσης όσων αφορά τους πόρους και τις πληροφορίες σφάλματων

• Εκτελέσιμους Πράκτορες (**agents**) ή Συνδέσμους Διαχείρισης (**connectors**) που επιτρέπουν και στο διαχειριστικό πλαίσιο να παρακολουθεί και να ελέγχει στοιχεία του συστήματος.

Αυτά τα δύο στοιχεία βοηθούν στην παρακολούθηση της υγείας των διαφόρων τύπων υπολογιστικών πόρων και επιδόσης και στην αντιμετώπιση αποτυχιών, διατηρώντας παράλληλα την ποιότητα των υπηρεσιών που απαιτούνται από τον Πάροχο Εφαρμογής Big Data. Οι σύνδεσμοι διαχείρισης είναι απαραίτητοι για σενάρια όπου οι πάροχοι υπηρεσίας νέφους δίνουν δυνατότητα διαχείρισης μέσω API. Είναι λογικό τα στοιχεία της υποδομής να

περιέχουν αυτόνομες, αυτορυθμιζόμενες και αυτοθεραπευτικές ικανότητες, μειώνοντας έτσι το συγκεντρωτικό μοντέλο παρακολούθησης του συστήματος.

2.6.4 Πλαίσια Πρόβλεψης / Παραμετροποίησης

Σε μεγάλες υποδομές με πολλές χιλιάδες κόμβους υπολογιστών και αποθήκευσης, η προμήθεια εργαλείων και εφαρμογών πρέπει να είναι όσο το δυνατόν πιο αυτοματοποιημένες. Η εγκατάσταση λογισμικού, η παραμετροποίησης εφαρμογής και η τακτική συντήρηση των ενημερώσεων θα πρέπει να ωθηθεί και να επαναληφθεί προς όλους τους κόμβους με αυτοματοποιημένο τρόπο, η οποία μπορεί να γίνει με πλήρης γνώση της τοπολογίας της υποδομής. Με την έλευση του virtualization, η αξιοποίηση εικονικών στιγμιότυπων μπορεί να επιταχύνει τη διαδικασία αποκατάστασης και να παράσχει αποτελεσματικό patching που μπορεί να ελαχιστοποιήσει το χρόνο διακοπής λόγω της προγραμματισμένης συντήρησης. Αυτά τα πλαίσια αλληλεπιδρούν επίσης με το Ασφάλεια και Ιδιωτικό Απόρρητο για να διασφαλιστεί ότι η διαμόρφωση του συστήματος πληρεί τις απαιτήσεις ασφάλειας που περιγράφονται στις πολιτικές που καθορίζει ο ενορχηστρωτής τους συστήματος. Υπάρχουν δύο σχετικές προσεγγίσεις:

Στατική

Αυτή η προσέγγιση λαμβάνει την απόφαση πρόβλεψης εικονικών πόρων που απαιτούνται για την εκτέλεση τμημάτων ροής εργασίας ή εργασιών πριν από την εκτέλεση.Δεν είναι σε θέση να κλιμακώσει δυναμικά με την χρήση εσωτερικών ή εξωτερικών πόρων. Οι προβλεπόμενοι πόροι είναι σταθεροί, και είναι οι μόνοι διαθέσιμοι πόροι καθ 'όλη τη διάρκεια της εκτέλεσης της ροής εργασίας. Ετσι, μια τέτοια προσέγγιση είναι κατάλληλη για χρήση σε περιπτώσεις όπου η ζήτηση της ροής εργασίας μπορεί να προβλεφθεί και είναι πλήρως προκαθορισμένη ως προς τους πόρους.

Δυναμική

Αντίθετα, αυτή η προσέγγιση λαμβάνει την απόφαση παροχής πόρων κατά την εκτέλεση της ροής εργασίας ή κατά το χρόνο εκτέλεσης. Αποφασίζει ποιοί πόροι ,τύποι και διαμορφώσεις είναι οι πλέον κατάλληλοι και πότε πρέπει να προστεθούν ή να αφαιρεθούν σύμφωνα με τις απαιτήσεις της ροής εργασίας. Με άλλα λόγια, αυτή η προσέγγιση λαμβάνει όλες τις αρχικές αποφάσεις ή τις καθορίζει κατά το χρόνο εκτέλεσης προσδιορίζοντας τους εικονικούς πόρους που πρέπει να συνεχίσουν να λειτουργούν, ποιοι πόροι πρέπει να διατεθούν και ποιοι πόροι πρέπει να καταργηθούν καθώς προχωρά η εκτέλεση της ροής εργασίας. Αυτή η προσέγγιση στοχεύει στην αποφυγή υποπρομήθειας λόγω της επίπτωσής της στην απόδοση (χαμηλότερη απόδοση) και την υπερβολική παροχή υπηρεσιών λόγω των επιπτώσεών στο κόστος και τη χρήση του συστήματος (αύξηση του κόστους και μείωση χρήση του συστήματος).

2.6.5 Διαχειριστές Πακέτων

Τα στοιχεία διαχείρισης πακέτων υποστηρίζουν την εγκατάσταση και τις ενημερώσεις άλλων στοιχείων του πλαισίου εντός του συστήματος Big Data.Τα συστατικά μέρη αυτής της κατηγορίας γενικά χρησιμοποιούν ένα κεντρικό δικτυακό αποθετήριο για να εξασφαλιστεί ότι η σωστή έκδοση του στοιχείου εφαρμόζεται με συνέπεια σε όλο το cluster.

2.6.6 Διαχειριστές Κύκλου Ζωής Δεδομένων

Τα στοιχεία διαχείρισης του Κύκλου Ζωής είναι απαραίτητα για τη διαχείριση του κύκλου ζωής των δεδομένων που εισάγονται στο σύστημα, αποθηκεύονται,διατηρούνται και είναι προσβάσιμα για σκοπούς επεξεργασίας ή διάδοσης.

Ο **κατάλογος μεταδεδομένων** είναι ο κατάλογος όλων των συνόλων δεδομένων του σύστηματος. Θα πρέπει να περιέχει το μοντέλο για τη θεμελιώδη έννοια της "μονάδας" δεδομένων, είτε πρόκειται για μια εγγραφή βάσης δεδομένων (π.χ. ζεύγος κλειδιού-τιμής είτε σχεσιακή σειρά πίνακα) ή ένα σύνολο δεδομένων (π.χ. αρχείο εξαχθέν από βάση δεδομένων). Κάθε μονάδα δεδομένων έχει χαρακτηριστικά που περιγράφονται στα σχετικά μεταδεδομένα, τα οποία πρέπει να περιλαμβάνουν τουλάχιστον ένα μοναδικό αναγνωριστικό και χρονική σήμανση που υποδεικνύει πότε το τα δεδομένα δημιουργήθηκαν ή / και λήφθηκαν. Αυτά τα χρονικά σήματα θα βοηθήσουν τον Διαχειριστή Κύκλου Ζωής Δεδομένων να παρακολουθεί την "ωρίμανση" των δεδομένων εντός του συστήματος. Επιπλέον, ο κατάλογος μεταδεδομένων θα πρέπει να υποστηρίζει την ανεύρεση των δεδομένων που είναι απαραίτητη σε θέματα εξουσιοδοτημένης πρόσβασης και διακυβέρνησης. Υπάρχουν πολυάριθμα διεθνή και εθνικά πρότυπα που διέπουν το περιεχόμενο, το μοντέλο και τις διεπαφές για τους καταλόγους μεταδεδομένων.

Ο Tracker δεδομένων παρακολουθεί την κίνηση δεδομένων σε όλο το σύστημα, από το σημείο λήψης έως το διάδοσης ή καταστροφής.Υπάρχουν δύο είδη:

• **Κίνησεις εισόδου και εξόδου:** παρακολουθεί δεδομένα που εισέρχονται και εξέρχονται από το σύστημα. Η έξοδος μπορεί να προέρχεται είτε από ανάγκες της εφαρμογής ή της πολιτικής διατήρησης μεγάλων δεδομένων. Πράγματι, ορισμένες εφαρμογές ενδέχεται να απαιτούν φρέσκα δεδομένα για συγκεκριμένους αναλυτικούς σκοπούς. Ο βαθμός ανανέωσης εξαρτάται από τις ειδικές απαιτήσεις των επιχειρηματικών εφαρμογών και μπορεί να επηρεαστεί από την πολιτική και τους κανονισμούς. Για παράδειγμα, ενώ μια εφαρμογή visual analytics παρακολουθεί τον βαθμό έγκρισης ή απόρριψης κατά τη διάρκεια ενός τηλεοπτικού debate για την προεδρική εκλογή απαιτεί δεδομένα σε πραγματικό χρόνο και τα πιο πρόσφατα δεδομένα από tweet και blog, η μελέτη της τάσης του εισοδήματος των νοικοκυριών κατά τα τελευταία 50 χρόνια χρειάζεται τόσο τα πρόσφατα όσο και τα ιστορικά απογραφικά δεδομένα.Από την άλλη, οι νόμοι και οι πολιτικές διαχείρισης αρχείων ενδέχεται να υπαγορεύουν στην επειχείρηση τον χρόνο διατήρησης των δεδομένων και συνεπώς να επηρεάσουν τη λογική διατήρησης που θα ακολουθηθεί.

• **Μετακίνηση εντός συστήματος:** Λόγω του μεγάλου όγκου των δεδομένων, ο διαχειριστής συστήματος δεδομένων πιθανότατα θα χρησιμοποιεί πολυεπίπεδη αποθήκευση για σκοπους αποδοτικότητας κόστους και ευκολίας κλιμάκωσης. Μέσα σε αυτό το περιβάλλον αποθήκευσης, τα δεδομένα γίνονται διαθέσιμα στις διαδικασίες ανάλυσης που διαχειρίζονται οι Πάροχοι Εφαρμογών Big Data. Οι πάροχοι εμπορικών υποδομών προσφέρουν διαφορετικές κατηγορίες αποθήκευσης με αντίστοιχα διαφορετικά μοντέλα τιμολόγησης. Η διάθεσης δεδομένων σε διεργασίες και εφαρμογές μπορεί να πραγματοποιηθεί με τη φυσική μεταφορά των δεδομένων στην μονάδα αποθήκευσης όπου επι αυτής μπορεί να λειτουργήσει το λογισμικό επεξεργασίας. Ωστόσο, ένα πρόσφατο παράδειγμα είναι η μετακίνηση των δυνατότητων υπολογισμού και επεξεργασίας στο σημείο που βρίσκονται τα δεδομένα για να παρακάμψει η μεγάλη μεταφορά δεδομένων μεταξύ των επιπέδων αποθήκευσης.

Ο Tracker δεδομένων συνήθως διασυνδέεται με το στοιχείο διατήρησης δεδομένων για την εφαρμογή των πολιτικών μακροπρόθεσμης αποθήκευσης.Το στοιχείο διατήρησης δεδομένων Data Preservation component εφαρμόζεται τόσο σε μόνιμα όσο και σε προσωρινά δεδομένα. Η ευθύνη του είναι να ελέγχουν συνεχώς την "ωρίμανση" των δεδομένων στο σύστημα και να κάνουν ενέργειες στα δεδομένα βασιζόμενοι στη πολιτική διατήρησης. Για τα μόνιμα δεδομένα, η διατήρηση δεδομένων θα πραγματοποιήσει το ορισμένο Σχέδιο Διατήρησης, το οποίο μπορεί να αποτελείται από μετεγκατάσταση δεδομένων σε μορφή κατάλληλη για μακροπρόθεσμη διατήρηση, περιοδική ανανέωση του υλικού αποθήκευσης ή συντήρηση περιβαλλοντος προσομοίωσης/εξομοίωσης που χρησιμοποιείται για την ανάγνωση των αρχειοθετημένων δεδομένων. Η Διατήρηση Δεδομένων θα αξιοποιήσει την πολυεπίπεδη αποθήκευση που πληρεί τις απαιτήσεις ανθεκτικότητας των δεδομένων και επιτυγχάνει αποδοτικότητα κόστους. Εαν τα δεδομένα θεωρείται ότι έχουν περιορισμένη διάρκεια ζωής, τότε η Διατήρηση Δεδομένων θα εφαρμόσει τις κατάλληλες μεθόδους για να τα εξαλείψει από το σύστημα. Οι μέθοδοι καθαρισμού θα εξαρτηθούν από την πολιτική ασφαλείας ώστε να εξασφάλιστεί και η εμπιστευτικότητα των δεδομένων

2.7 Αποθήκευση Big Data

[9. Yu , Guo]

Η αποθήκευση των δεδομένων είναι η καρδιά των Big Data εργαλείων και πλατφορμών. Εάν ο μεγάλος όγκος δεδομένων διαφόρων μορφών δεν αποθηκευτεί κατάλληλα, οι υπολογισμοί και η ανάκτηση των δεδομένων δεν μπορούν να γίνουν αποτελεσματικά.

2.7.1 Μοντέλα Δεδομένων | Data Models

Οι τεχνολογίες Big Data υποστηρίζουν τυπικά διάφορους τύπους μοντέλων δεδομένων που αντιπροσωπεύουν τα δεδομένα για πρόσβαση και διαχείριση. Το πιο δημοφιλές είναι το Σχεσιακό Μοντέλο, το οποίο είναι αναμφισβήτητα το ευρύτερα χρησιμοποιούμενο μοντέλο στην βιομηχανία τις τελευταίες δεκαετίες. Προτάθηκε το 1969, από τον Edgar F. Codd, και μέσω αυτού όλα τα δεδομένα αντιπροσωπεύονται με όρους πλειάδων (ή αρχείων), ομαδοποιημένων σχέσεων (ή πινάκων) και σχετικών αρχείων και συνδέονται μεταξύ τους με ένα κλειδί. Στα συστήματα βάσεων δεδομένων με σχεσιακό μοντέλο χρησιμοποιείται η δομημένη γλώσσα ερωτημάτων (SQL) για τον τον ορισμό και την πρόσβαση στα δεδομένα. Οι διάφορες επεκτάσεις της υποστήριξης του **Σχεσιακό Μοντέλου** και της **SQL** ωστε να παρέχονται και από τις τεχνολογίες Big Data ποικίλλουν σημαντικά από το ένα σύστημα στο άλλο.

Ωστόσο, υπάρχουν τεχνολογίες Big Data που δεν το υποστηρίζουν. Είναι γνωστές ως Βάσεις Δεδομένων **NoSQL**. Υποστηρίζουν μοντέλο βασισμένο σε ζεύγη κλειδιού\τιμής, όπου σε κάθε αυθαίρετο κλειδί η τιμή μπορεί να είναι οτιδήποτε ξεκινώντας από τους συνηθισμένους τύπους δεδομένων (όπως ακέραιος, χαρακτήρας, byte κ.λπ.), blob, πολυδιάστατη αντιστοίχιση δομής σε πολύπλοκα αντικείμενα εγγράφων όπως XML, JSON κ.λπ.

Το άλλο ενδιαφέρον μοντέλο δεδομένων που υποστηρίζεται από τις τεχνολογίες Big Data είναι το μοντέλο γράφων όπου τα δεδομένα αντιπροσωπεύονται ως κόμβοι και συνδέσεις (ή ακμές). Το μοντέλο Graph μπορεί να αντιπροσωπεύει διάφορα προβλήματα, όπως υπολογιστικές αποστάσεις ή πολυπλοκότητα, κυκλικότητα στις σχέσεις, προσδιορισμό συνδεσιμότητας κλπ με τρόπο κατα πολύ ανώτερο των άλλων μοντέλων δεδομένων που προαναφέρθηκαν. Εχει εφαρμογή σε πολλές πραγματικές περιπτώσεις χρήσης που σχετίζονται με τα Social Media, το Δίκτυα Πληρωμών, Δίκτυα Πελατών, κλπ.

Ανεξαρτήτως των μοντέλων, οι περισσότερες Big Data τεχνολογίες βασίζονται στην έννοια του ορισμού του μοντέλου κατα τον χρόνου εκτέλεσης (“**schema on read**”) . Αυτό βοηθά στην εκ των προτέρων κατανόηση των δεδομένων χωρίς να έχει προαποφασίστέι το μοντέλο δεδομένων που θα ενσωματωθούν. Μόλις τα δεδομένα γίνουν διαθέσιμα στην δεξαμενη δεδομένων(**data lake**), ορίζεται το σχήμα (μοντέλο δεδομένων) κατά την ανάγνωση των δεδομένων. Αυτή η προσέγγιση είναι ένας σημαντικός παράγοντας μεταστροφής του παιχνιδιού σε σύγκριση με την προσέγγιση που χρησιμοποιείται στις παραδοσιακές τεχνολογίες βάσεων δεδομένων όπου πρέπει να οριστεί το σχήμα πριν γίνει οποιαδήποτε μετακίνηση δεδομένα στην πλατφόρμα. Αυτή η ιδέα συμβάλλει στην προώθηση του σκεπτικού όπου πρώτα φέρνουμε τα δεδομένα σε μια κοινή πλατφόρμα, χωρίς να ξοδεύεται μεγάλο μέρος του χρονου στην ανάλυση για τον ορισμό του νέου ή του ενημερωμένου σχήματος (ενημέρωση που συμβαίνει πολύ πιο συχνά στις μέρες μας λόγω των συνεχώς μεταβαλλόμενων επιχειρηματικών οικοσυστημάτων και αναγκών). Μόλις τα δεδομένα λοιπόν γίνουν διαθέσιμα στην δεξαμενη δεδομένων φιλτράρονται για τις επιθυμητές πληροφορίες, επεξεργάζονται για να παράγουν πληροφορία προστιθέμενης αξίας και τελικά καταναλώνονται από τις επιχειρηματικές διαδικασίες. Το αχρησιμοποίητο τμήμα των αρχικών δεδομένων εξακολουθεί να διατηρείται για μελλοντικές ανάγκες, αποτελώντας τα «γνωστά άγνωστα» και «άγνωστα άγνωστα» data assets του οργανισμού.

2.7.2 Κατατμήσεις Δεδομένων | Data Partioning

Κάθε τεχνολογία Big Data πρέπει να ακολουθήσει κάποια προσέγγιση για τη κατάτμηση των δεδομένων μεταξύ των διαφόρων κόμβων δεδομένων (Data Nodes). Αυτό οφείλεται στο γεγονός ότι όλα τα δεδομένα δεν μπορούν να αποθηκευτούν σε μία μοναδική διακριτη μηχανή και επίσης για να διασφαλιστεί ότι η επεξεργασία δεδομένων θα εκτελείται παράλληλα με τη χρήση διαθέσιμων πόρων από όλες τις μηχανές.

Υπάρχουν πολλαπλές προσεγγίσεις που ακολουθούνται για την κατάτμηση δεδομένων.

1)Για δεδομένα που έχουν μοντελοποιηθεί στην μορφή κλειδιού / τιμής ή κλειδιού/πλειάδας, η κατανομή γίνεται με βάση το κλειδί. Μπορούν να ακολουθήσουν διάφοροι τύποι σχημάτων για να επιτευχθεί ο απαιτούμενος διαχωρισμός όπως:

- **Καταμερισμός Εύρους | Range partitioning** - όπου τα δεδομένα διαχωρίζονται με βάση το εύρος ενός κλειδιού χαρακτηριστικών, όπου τα κλειδιά με διαφορετικές τιμές βρίσκονται σε διαφορετικούς κόμβους.

- **Διαίρεση Hash** - όπου τα κλειδιά έχουν εκχωρηθεί σε έναν κόμβο με βάση το αποτέλεσμα μιας συνάρτησης κατακερματισμού που εφαρμόζεται σε ένα ή περισσότερα χαρακτηριστικά.
- **Λίστα Κατανομής** - όπου οι μοναδικές τιμές των κλειδιών σε κάθε διαμέρισμα καθορίζονται ως απλή λίστα.
- **Τυχαία Κατανομή** - όπου τα κλειδιά εκχωρούνται σε κόμβους με τυχαίο τρόπο
- **Εκ περιτροπής Κατανομή | Round Robin partitioning** - όπου τα κλειδιά εκχωρούνται σε κόμβους χρησιμοποιώντας μηχανισμό με αλγόριθμο χρονοπρογραμματισμού Round-Robin
- **Διαίρεση βασιζόμενη σε Ετικέτες | Tag-based partitioning** -όπου βασίζομενα σε κάποια ετικέτα στα δεδομένα, τα κλειδιά ομαδοποιούνται σε ένα λογικό διαμέρισμα.

2)Για τα δεδομένα που είναι αποθηκευμένα και προσπελάζονται μαζικά χωρίς κάποιο προκαθορισμένο σχήμα, χρησιμοποιείται μέθοδος κατάτμησης που βασίζεται σε μπλοκ. Σε αυτή την προσέγγιση κάθε διαδοχικό μπλοκ από bytes, βασιζόμενο σε ένα συγκεκριμένα προκαθορισμένο μέγεθος μπλοκ, ταξινομείται σε διαφορετικούς κόμβους.

3)Για διαχωρισμό δεδομένων που έχουν μοντελοποιηθεί σαν γράφοι τυπικά χρησιμοποιούν δύο προσεγγίσεις – Μείωση(κλάδεμα) κορυφών ή άκρων ανάλογα με την φύση του προβληματος. Ορισμένοι αλγόριθμοι επεξεργασίας χρησιμοποιούν επίσης υβριδική προσέγγιση η οποία κατανέμει ομοιόμορφα κορυφές χαμηλού βαθμού μαζί με τις ακμές τους όπως κοπή ακμών και κατανέμει ομοιόμορφα άκμές υψηλού βαθμού κορυφές όπως κόψιμο κορυφής.

Η αναγνώριση ενός κατάλληλου Data Node για την απόθεση μιας κατάτμησης γίνεται συνήθως με τους ακόλουθους τρόπους.

Η **πρώτη προσέγγιση** είναι να τοποθετείται στον κόμβο που είναι τοπικός του πρόγραμματος -πελάτη που γράφει τα δεδομένα. Αυτό συνήθως γίνεται κατά το Block Partitioning.

Στη **δεύτερη προσέγγιση** (που ισχύει στην μέθοδο Hash,Round Robin και στην τυχαίας κατάτμησης), το κλειδί κατακερματισμού, το τυχαίο κλειδί ή το κλειδί του Round Robin μηχανισμού χρησιμοποιείται για τον προσδιορισμό του κόμβου.

Στην **τρίτη προσέγγιση** (που ισχύει για την κατάτμηση Εύρους, Λίστας ή Ετικετών) οι κόμβοι δεδομένων μπορούν συνήθως να επισημανθούν εκ των προτέρων βάσει του αριθμού εύρους ή λίστας ή ετικέτας.

Υπάρχει και μια **τέταρτη προσέγγιση**, όπου τα δεδομένα αλληλοσυνδέονται με άλλα δεδομένα στον ίδιο κόμβο βάσει κάποιο κλειδιού αναφοράς / εξωτερικού κλειδιού.

Η κατάτμηση δεδομένων συνεπάγεται επίσης την τακτική εξισορρόπηση/αναδιοργάνωση των δεδομένων εντός των πολλαπλών κόμβων για να διασφαλιστεί ότι δεν υπάρχουν κατακερματισμένα ή υπερφορτωμένα σημεία (η κατάσταση κατά την οποία τα ερωτήματα πελάτη ή οι διεργασίες καταλήγουν πάντα μόνο σε έναν ή πολύ λίγους κόμβους δεδομένων). Αυτό επιτυγχάνεται με τους εξής τρόπους:

Πρώτον, χωρίζοντας εκ νέου μια κατάτμηση όταν μεγαλώσει περισσότερο από ένα συγκεκριμένο κατώφλι και, Δεύτερον, επαναναδιανέμοντας ολόκληρο το αρχείο / πίνακα της συστοιχίας με προγραμματισμένο τρόπο σε ημερήσια, εβδομαδιαία ή μηνιαία βάση. Αυτή η προσέγγιση χρησιμοποιείται επίσης όταν προστίθενται ή αφαιρούνται ολόκληροι κόμβοι δεδομένων.

2.7.3 Πολλαπλή Αντιγραφή Δεδομένων | Data Replication

Η πολλαπλή αναπαραγωγή δεδομένων είναι ένα κοινό χαρακτηριστικό για όλους τους τύπους τεχνολογιών Big Data. Η αναπαραγωγή παρέχει πλεονασμό και συνεπώς αυξάνει τη διαθεσιμότητα των δεδομένων και την τοπικότητα. Με πολλαπλά αντίγραφα των δεδομένων σε διαφορετικά Data Nodes, το replication προστατεύει τα αποθηκευμένα δεδομένα από την απώλεια ενός διακομιστή εξαιτίας αποτυχίας υλικού, διακοπών υπηρεσίας κ.λπ. Με πολλαπλά αντίγραφα των δεδομένων καθένα απο αυτά μπορεί να αφιερωθεί στην ανάκτηση από καταστροφές (disaster ecovery), αναφορά(reporting), δημιουργία αντιγράφων ασφαλείας(backup) κ.λπ. Το replication επίσης βοηθά στην κατανεμημένη επεξεργασία ερωτημάτων(distributed query proccessing). Η τοπικότητα των δεδομένων βοηθάει στη μείωση του χρόνου επεξεργασίας. Στην περίπτωση μεγάλου όγκου ταυτόχρονων αιτημάτων ανάγνωσης συμβάλλει επίσης στην αύξηση της διαθεσιμότητας.

Τυπικά, τρία αντίγραφα ενός data partition μπορούν να ανταπεξέλθουν στις περισσότερες καταστάσεις αποτυχίας. Σε περιβάλλον συστοιχίας (cluster), το πρώτο αντίγραφο τυπικά τοποθετείται σε διαφορετικό rack από αυτό που έχει πρωτεύον αντίγραφο του ίδιου partition. Το δεύτερο αντίγραφο μπορεί να τεθεί σε διαφορετικό(χωρικά) data center για την περίπτωση που απαιτείται ανάκτηση λόγω καταστροφής. Για την επιλογή των σωστών Data Nodes σε διαφορετικά racks, χρησιμοποιούνται τεχνικές που είναι παρόμοιες με αυτές του data partitioning. Πολλές φορές τα διαφορετικά αντίγραφα των δεδομένων δημιουργούνται με τέτοιο τρόπο ώστε κάθε

αντίγραφο να αντιπροσωπεύει διαφορετική απο-κανονικοποιημένη όψη(denormalized view) των ίδιων δεδομένων.

Οι στρατηγικές του replication των δεδομένων στα Big Data επηρεάζονται πολλές φορές από το θεώρημα CAP. Όπως έχει ήδη αναλυθεί,το θεώρημα CAP λέει ότι στην περίπτωση του κατανεμημένου συστήματος, δεν μπορούν να επιτευχθούν ταυτόχρονα η συνέπεια, διαθεσιμότητα και κατάτμηση. Κατά τη λήψη αποφάσεων σχετικά με το που θα διατηρούνται τα αντίγραφα, τα Big Data τυπικά είτε υποστηρίζουν την Συνέπεια και την Διαθεσιμότητα σε βάρος της Κατάτμησης(Αυτό το μοντέλο είναι γνωστό ως CA) είτε Διαθεσιμότητα και Διαχωρισιμότητα σε βάρος της Συνέπειας(Αυτό το μοντέλο είναι γνωστό ως AP) .Για μερικά συστήματα Big Data που είναι προσανατολισμένα στην χρήση μνήμης αντί replication δεδομένων, χρησιμοποιείται η τεχνική καταγραφής προέλευσης Provenance (Lineage). Τα δεδομένα δηλαδή μπορούν να δημιουργηθούν εκ νέου χρησιμοποιώντας τις πληροφορίες που συντηρεί η lineage σε περίπτωση που ένας συγκεκριμένος κόμβος καταρρεύσει

2.7.4 Συμπίεση Δεδομένων | Data Compressing

Τα προβλήματα που απαιτούν λύσεις Big Data σχετίζονται κυρίως με δεδομένα που απαιτούν την αποθήκευση και την επεξεργασία μεγάλων όγκων. Ως εκ τούτου, οι ανάγκες για συμπίεση των δεδομένων αυτών είναι πιο σημαντικές από ποτέ. Η συμπίεση δεδομένων βοηθά με πολλούς τρόπους. Πρώτον, μειώνει το μέγεθος των δεδομένων. Αυτό σημαίνει ότι, εκτός από το μικρότερο χώρο αποθήκευσης, υπάρχει λιγότερο κόστος κατά την ανάγνωση και τη γραφή των δεδομένων από και προς τους δίσκους. Δεύτερον, εξασφαλίζει μικρότερη χρήση του εύρους ζώνης του δικτύου όταν η επεξεργασία δεδομένων απαιτεί την μετακίνηση των δεδομένων από έναν κόμβο σε άλλο. Το προφανές μειονέκτημα της συμπίεσης δεδομένων είναι ο χρόνος επεξεργασίας που απαιτείται για τη συμπίεση και αποσυμπίεση των δεδομένων κατά την εγγραφή και ανάγνωση των δεδομένων αντίστοιχα. Μια άλλη δύσκολη πτυχή της συμπίεσης δεδομένων είναι η ικανότητα διάσπασης των συμπιεσμένων δεδομένων για λόγους partitioning. Πολλές τεχνικές συμπίεσης δηλαδή μειωνεκτούν απο τον περιορισμό ότι τα συμπιεσμένα δεδομένα που παράγουν δεν μπορούν να χωριστούν σε διάφορους κόμβους.

Οι τεχνικές συμπίεσης παραδοσιακά ήταν εγγενής και κάθε Big Data τεχνολογία χρησιμοποίησε το δικό της σχήμα συμπίεσης σχεδιασμένο για να επιτύγχάνει συγκεκριμένο στόχο ως προς την απόδοση και το χρόνο απόκρισης. Ωστόσο, κατά την τελευταία δεκαετία ικανοποιητικός αριθμός των γενικών τεχνικών συμπίεσης εμφανίστηκε στην αγορά που μπορεί να χρησιμοποιηθεί από οποιαδήποτε τεχνολογία. Ο πίνακας παρακάτω παρέχει μια συγκριτική μελέτη ορισμένων από τις γενικές τεχνικές συμπίεσης. Ο λόγος συμπίεσης και ο αριθμός συμπιέσεων προκύπτουν από τη σύγκριση που έγινε στο Yahoo

Table 2.1 Comparative study of some of the common compression techniques

Tools	Algorithm	Strategy	Compression performance	Decompression performance	Compression ratio	Splitability
Gzip	Based on the DEFLATE algorithm, which is a combination of LZ77 and Huffman Coding	Dictionary based compression strategy	Low	Low	High (~60 %)	N
LZO	Uses PPM family of statistical compressors, a variant of LZ77	Dictionary based and block oriented	High	High	Less (~50 %)	Yes if indexed. It is possible to index LZO compressed files to determine split points so that LZO files can be processed efficiently in subsequent processing.
Snappy	LZ77 based	Block oriented	Highest	Highest	Low (~40 %)	Yes if used in a container format like Avro or sequence file
bzip2	Uses Burrows-Wheeler transform	Transformation based and block oriented	Lowest	Lowest	Highest (~70 %)	Yes

Εικόνα 14.Βασικότερες Τεχνικές Συμπίεσης

2.7.5 Είδη/Μορφοποίηση Δεδομένων | Data Format

Οι διαφορετικές μορφοποιήσεις που χρησιμοποιούνται για την αποθήκευση και την επεξεργασία δεδομένων στις τεχνολογίες Big Data χρειάζονται ειδική ανάλυση καθώς είναι το θεμέλιο και θέτουν τα όρια της επεκτασιμότητας και της ευελιξίας τουν συνόλου της όποιας λύσης Big Data. Οι μορφές που είναι ευρέως δημοφιλείς σήμερα είναι:

Text File Format | Αρχεία Κειμένου - Αυτά είναι τα αρχεία που έχουν δεδομένα σε ανθρώπινη αναγνώσιμη μορφή όπου τα αρχεία χωρίζονται χρησιμοποιώντας τον χαρακτήρα νέας γραμμής '\ n' και τα πεδία συνήθως

διαχωρίζονται με χαρακτήρες όπως ',' '\ t', κ.λπ. Τα παραδείγματα των δημοφιλών αρχειοθετημένων αρχείων είναι τα αρχεία CSV (τιμές διαχωρισμένες με κόμματα), τα αρχεία TSV (τιμές που χωρίζονται από καρτέλες) κλπ.

Parquet - είναι ένας τρόπος αποθήκευσης δεδομένων σε στήλες όπου αποθηκεύονται τα δεδομένα όπου όλες οι τιμές μιας στήλης τοποθετούνται μαζί με όλες τις τιμές μιας σειράς. Τα parquets που κατασκευάστηκαν με σκοπό να υποστηρίξουν τα πολυσύνθετα ενθηλακωμένα δεδομένα (nested data) με χρήση των αλγορίθμων γλώσσας χαμηλού επιπέδου (assembly) για πρόσβαση σε επίπεδο εγγραφής (record shredding). Μπορεί να χρησιμοποιηθεί από οποιοδήποτε πλαίσιο επεξεργασίας δεδομένων για τη βελτίωση της ανάγνωσης, της γραφής και της επεξεργασίας δεδομένων. Για την διαχείριση δεδομένων τύπου υπάρχει έχει αντίστοιχο έργο ανοικτού κώδικα ανώτερου επιπέδου στο Apache

Optimized Row Columnar Files or ORC – Αποτελεί αποθήκευση σε στήλες με τη φιλοσοφία της αποθήκευσης όλων των τιμών μιας στήλης μαζί αντί για όλες τις τιμές μιας γραμμής. Αυτή η μορφή αποθήκευσης χρησιμοποιείται ευρέως από το Hive . Αποθηκεύει τα δεδομένα σε λωρίδες και διατηρεί πρόσθετες πληροφορίες (όπως ευρετήρια, συναθροίσεις κτλ.) στα ίδια μπλοκ δεδομένων που αποθηκεύουν τις λωρίδες.

Το Avro -είναι μια μορφή αποθήκευσης δεδομένων προσανατολισμένη στην γραμμή(record) σε αντίθεση με το Parquet ή το ORC. Η Avro βασίζεται στα σχήματα. Τα δεδομένα Avro αποθηκεύονται σε ένα αρχείο μαζί με το σχήμα τους του έτσι ώστε να μπορεί να επεξεργαστεί αργότερα από οποιοδήποτε πρόγραμμα.

Sequence Files - Τα αρχεία αλληλουχίας είναι αρχεία που αποτελούνται από δυαδικά κλειδιά/τιμές . Υπάρχουν τρεις τύποι: τα μη συμπιεσμένα κλειδιά / τιμές , αρχεία όπου συμπιέζονται μόνο οι τιμές και το τρίτο όπου τα κλειδιά και οι τιμές συλλέγονται ξεχωριστά και συμπιεσμένα σε «μπλοκ». Τα αρχεία ακολουθίας χρησιμοποιούνται ευρέως για την αποθήκευση των δεδομένων που είναι δύσκολο να χωριστούν όπως το XML, JSON κ.α

Ειδικές Μορφές - Όπως και στην περίπτωση της συμπίεσης, υπάρχουν και πολλές άλλες μορφές αποθήκευσης που είναι συγκεκριμένες και εγγενής σε μια μόνο τεχνολογία Big Data. Χρησιμοποιούνται συνήθως από τις βάσεις δεδομένων NoSQL, τις βάσεις δεδομένων εντός μνήμης (In Memory Databases) και τις λύσεις αποθηκών δεδομένων (datawarehouse), όπου κάθε τεχνολογία προσπαθεί να δημιουργήσει προστιθέμενη αξία στο προιόν της μέσω της καινοτομίας στη μορφή δεδομένων. Αυτές οι μορφές χρησιμοποιούν επίσης ειδικό μηχανισμό για τη συμπίεση δεδομένων.

Οι εκάστοτε εφαρμογές ενδέχεται να χρειαστεί να χρησιμοποιούν πολλαπλές μορφές για τα διαφορετικά στάδια του χειρισμού των δεδομένων. Απο πλευρά της πηγής, για την κατανάλωση δεδομένων από άλλα εργαλεία / τεχνολογίες σε μια τεχνολογία Big Data, τα δεδομένα αυτά είναι συνήθως σε μορφοποιημένο κείμενο (Delimited Text format). Θα πρέπει δηλαδή να αναλωθούν ως έχουν και στη συνέχεια να μετατραπούν στην προτιμόμενη μορφή για αποτελεσματικότερη επεξεργασία / ανάλυση. Ορισμένα εργαλεία υποστηρίζουν την άμεση μετατροπή στη μορφή δεδομένων που απαιτεί ο στόχος εντός του βήματος λήψης δεδομένων. Κατά την εξαγωγή δεδομένων από μια τεχνολογία Big Data σε μια στοχευμένη τεχνολογία / εργαλείο, η μορφή δεδομένων στόχου είναι επίσης τυπικά σε μορφή Delimited Text format . Για σκοπούς Analytics που αφορούν αθροίσεις/συγκεντρωτικά στοιχεία σε επίπεδο στήλης, οι μορφές στήλης δεδομένων όπως το Parquet ή ORC θεωρούνται οι καλύτερες. Η χρήση του Avro έχει νόημα όταν οι ορισμοί των στηλών είναι πιθανόν να αλλάξουν στο μέλλον για έναν δεδομένο πίνακα και οι περισσότερες στήλες μιας εγγραφής χρησιμοποιούνται συχνά από τις περιπτώσεις χρήσης του στόχου. Μεταξύ όλων αυτών των μορφών, το Parquet γίνεται γρήγορα πρότυπο στις διάφορες Τεχνολογίες Big Data, λόγω του γενικού και ανοικτού πηγαίου κώδικα και της αποτελεσματικής δομής αποθήκευσης που προσφέρει.

2.7.6 Ευρετηρίαση | Indexing

Στις τεχνολογίες Big Data, ένα αρχείο δεδομένων είναι συνήθως χωρισμένο σε πολλαπλά κομμάτια/μπλοκ δεδομένων στους διάφορους κόμβους δεδομένων μια συστοιχίας. Η ευρετηρίαση στα Big Data εξυπηρετεί δύο σκοπούς:

-**Πρώτον**, είναι ο τρόπος για τον εντοπισμό συγκεκριμένης εγγραφής σε ένα αρχείο, προσδιορίζοντας πρώτα σε ποιο κομμάτι/μπλοκ δεδομένων ανήκει η εγγραφή.

-**Δεύτερον,** βοηθά επίσης στο να εντοπιστεί ταχύτερα η ακριβής τοποθεσίας της εγγραφής του αρχείου εντός ενός μπλοκ δεδομένων.

Η ευρετηρίαση είναι ιδιαίτερα σημαντική για τις περιπτώσεις ανάγνωσης/γραφής για τυχαίας αναζήτησης μιας συγκεκριμένης εγγραφής από ένα μεγάλο σύνολο δεδομένων. Ωστόσο, η ευρετηρίαση παίζει σημαντικό ρόλο στην επεξεργασία μαζικών δεδομένων όπως επίσης για την εύρεση των σωστών μπλοκ δεδομένων (ενός αρχείου) προς επεξεργασία αντί για επεξεργασία / σάρωση όλων.

Διάφοροι τύποι μηχανισμών ευρετηρίασης λοιπον χρησιμοποιούνται από διαφορετικές τεχνολογίες Big Data. Χωρίζονται στις εξής:

Traditional / Non-AI-based Indexing

Οι πιο δημοφιλείς είναι το **B Tree** , το **Inverted Index** , το **Radix Tree**, το **Bitmap Index** κ.α Υπάρχει επίσης ένας ειδικός τύπος τεχνικής ευρετηρίασης, ο οποίος είναι διαφορετικός από τον κανονικό τρόπο εντοπισμού μιας εγγραφής στόχου σε ένα DataNode / Data Block. Αυτή η τεχνική δεν παρέχει πληροφορίες σχετικά με το πού υπάρχουν τα δεδομένα και αντίθετα λέει πού να μην ψάξει κάποιος τα επιθυμητά δεδομένα. Τα παραδείγματα τέτοιων τεχνικών είναι το **Filter Bloom** , το **Zone Map** κ.α. Αυτή η τεχνική θεωρείται καλός ενισχυτής απόδοσης σε περίπτωσεις αναζήτησης σε μεγάλο όγκο δεδομένων, καθώς αυτό εξοικονομεί πολλά input / output δίσκου.

Η πιο δημοφιλής τεχνική ευρετηρίασης bulk δεδομένων είναι η χρήση του σκεπτικού διαχωρισμού σε λογικά τμήματα όπου κάθε πίνακας / αρχείο χωρίζεται με βάση λίγα ποιοτικά χαρακτηριστικά (όπως ημερομηνία, χώρα κ.λπ.) που είναι διαθέσιμα σε κάθε εγγραφή. Τα δεδομένα που περιέχουν την ίδια τιμή για ένα χαρακτηριστικό δηλαδή συντηρούνται σε έναν ξεχωριστό φάκελο. Κατά τη επεξεργασία δεδομένων, ο καθορισμούς αυτού του χαρακτηριστικού βοηθάει πολύ καθώς το σύστημα επεξεργάζεται τα διαθέσιμα δεδομένα μόνο σε εκείνους τους φακέλους που ικανοποιούν τα κριτήρια και όχι ολόκληρου του dataset.

AI-based Indexing | Ευρετηρίαση με βάση την Τεχνητή Νοημοσύνη

Αυτή η προσέγγιση είναι σε θέση να ανακαλύψει άγνωστη συμπεριφορά μεγάλων δεδομένων χρησιμοποιώντας μια βάση γνώσης, παρέχοντας αποτελεσματική ευρετηρίαση δεδομένων και επομένως αποτελεσματική αναζήτηση και ανάκτηση δεδομένων. Ωστόσο, γενικά χρειάζεται περισσότερο χρόνο για να απαντήσει στο ερώτημα αναζήτησης σε σύγκριση με μια προσέγγιση ευρετηρίασης χωρίς AI. Οι ήπιες υπολογιστικές τεχνικές ευρετηρίασης που βασίζονται σε AI συνδυάζουν ασαφείς μεθόδους και υπολογιστικές μεθόδους νευρώνων για την ευρετηρίαση δεδομένων, ενώ οι τεχνικές με βάση τη Μηχανική Μάθηση (ML) βελτιώνουν την ευρετηρίαση δεδομένων χρησιμοποιώντας μεθόδους μηχανικής μάθησης όπως πχ η **manifold learning**. Η ευρετηρίαση βάσει γνώσης Αναπαράστασης και Λόγου (Knowledge Representation and Reasoning (KRR)-based) το επιτυγχάνει αυτό με την χρήση σημασιολογικών οντολογιών.

Collaborative AI-based Indexing | Συνεργατική Ευρετηρίαση με βάση την τεχνητή νοημοσύνη

Αυτή η προσέγγιση βελτιώνει την ακρίβεια της ευρετηρίασης δεδομένων και την αποτελεσματικότητα της αναζήτησης, στηριζόμενη στη συνεργατική τεχνητή νοημοσύνη, με στόχο την παροχή μεγαλύτερων συνεργατικών λύσεων ευρετηρίασης δεδομένων. Παρέχονται συνεργατικές μέθοδοι ευρετηρίασης με βάση την μηχανική μάθηση και συνεργατικές μέθοδοι ευρετηρίασης που βασίζονται σε KRR

Αποθήκευση Δεικτών

Εκτός από την επιλογή του σωστού μηχανισμού ευρετηρίασης είναι επίσης σημαντικό για τις τεχνολογίες Big Data να αποθηκεύονται και οι δείκτες των ευρετηρίων στη σωστή θέση. Υπάρχουν λίγες προσεγγίσεις που συνήθως ακολουθούνται όπως:

-**Centralized Indexing**, οι δείκτες δηλαδή για όλα τα δεδομένα διατηρούνται σε κεντρικό σημείο και η αναζήτηση γίνεται σε αυτό για να εντοπίστεί η σωστή τοποθεσία/κόμβο δεδομένων που πρέπει να αναζητηθούν τα δεδομένα.

-**Partitioned Index**, όπου υπάρχει ένας συγκεντρωτικός πίνακας παρατήρησης ο οποίος διατηρεί τον κατάλογο με το εύρος δεδομένων που είναι αποθηκευμένα σε κάθε Data Node ενώ ξεχωριστά ευρετήρια δημιουργούνται στην πραγματικότητα σε κάθε Data Node για τα δεδομένα που διατηρούνται στον συγκεκριμένο κόμβο.

-Στην τρίτη προσέγγιση, τα δεδομένα ταξινομούνται πάντοτε με βάση το πρωτεύον κλειδί και οι πληροφορίες σχετικά με το εύρος τους αποθηκεύονται σε έναν συγκεντρωτικό πίνακα αναζήτησης.

Τελος, πολλαπλά δευτερεύοντα ευρετήρια για τα διάφορα χαρακτηριστικά των δεδομένων υποστηρίζονται επίσης σε μερικές από τις τεχνολογίες Big Data. Δημιουργούνται είτε σε διακριτούς κόμβους (για τα δεδομένα που υπάρχουν σε εκείνο τον κόμβο) είτε εφαρμόζονται ως ξεχωριστοί πίνακες Ευρετηρίου.

2.7.7 Διατήρηση Δεδομένων | Data Persistence

Η συντήρηση είναι μία από τις σημαντικότερες πτυχές των τεχνολογιών Big Data καθώς η είσοδος/έξοδος των δεδομένων στους δίσκους και η μετακίνησή τους μέσω του δικτύου είναι το βασικό έναντι της αποτελεσματική επεξεργασία και πρόσβαση σε μεγάλο όγκο κατανεμημένων δεδομένων. Σε υψηλό επίπεδο, η συντήρηση των δεδομένων αντιμετωπίζεται με δύο τρόπους:

Στην **πρώτη προσέγγιση**,αντιμετωπίζεται αποθηκεύοντας καταλληλα τα δεδομένα στο τοπικό δίσκο του κάθε κόμβου δεδομένων.

Στη **δεύτερη προσέγγιση**, η συντήρηση γίενται μέσα από ένα γενικό σύνολο APIs του Κατανεμημένου Συστήματος Αρχείων τα οποία μάλιστα μπορούν να υλοποιηθούν από άλλο προϊόν/προμηθευτή που θα εξασφαλίζει ότι πληρούνται όλοι οι κανόνες του ίδιου του API αλλά και οι απαιτήσεις ποιότητας της υπηρεσίας. Αυτή η δευτερεύουσα προσέγγιση γνωρίζει την μεγαλυτερη ανάπτυξη στις μέρες μας καθώς δίνει περισσότερες επιλογές στην υλοποίηση μιας λύσης Big Data. Υπάρχουν δε δύο τύποι υλοποιήσεων για API κατανεμημένου συστήματος αρχείων:

-Στην πιο παραδοσιακή επιλογή τα δεδομένα αποθηκεύονται κατά κύριο λόγο σε έναν τοπικό δίσκο των κόμβων δεδομένων με λογική **Shared Nothing** (π.χ. HDFS, Gluster42S. MazumderFS, Spectrum Scale FPO κ.λπ.).

-Η άλλη επιλογή, η οποία ακόμα βρίσκεται στα πρώτα της βήματα είναι περισσότερο προσανατολισμένη στην κεντρική μνήμη(π.χ.Tachyon, Apache Ignite) . Σε αυτήν την προσέγγιση η πλειοψηφία των δεδομένων φορτώνεται στη μνήμη ως κατανεμημένα μπλοκ δεδομένων/κομμάτια ανακαλούμενα από τα διάφορα Data Nodes.Μετά την ολοκλήρωση της επεξεργασίας, τα δεδομένα διατηρούνται σε οποιαδήποτε διασυνδεδεμένο χώρο αποθήκευσης δεδομένων, από SAN, Cloud Store όπως S3, Βάσεις Δεδομένων NoSQL ή ακόμα και HDFS

2.8 Κλιμάκωση | Scaling

[35. Ian Gorton]

Η κλιμάκωση είναι η ικανότητα του συστήματος να προσαρμόζεται σε αυξημένες απαιτήσεις όσον αφορά την επεξεργασία δεδομένων.Όσων αφορά τον σχεδιασμό κλιμακούμενων λύσεων Big Data είναι γενικά απόδεκτές οι εξής εμπειρικές αρχές :

"**Δεν μπορείτε να επεκτείνετε την ανθρωποπροσπάθεια και το κόστος για την κατασκευή ενός συστήματος Big Data με τον ίδιο ρυθμό που θα επεκτείνετε τις ικανότητες/χωρητικότητα του συστήματος**". Εάν υπολογίσετε ότι εντός ενός έτους το σύστημά σας θα είναι 4 φορές μεγαλύτερο, δεν είναι λογικό να αναμένετε ότι μπορείτε/χρειάζετε έχετε 4 φορές μεγαλύτερη ομάδα υποστήριξης

"**Όσο πιο σύνθετη λύση, τόσο λιγότερη κλιμάκωση θα είναι εφικτή**" - επιλέξτε την τεχνολογία με σύνεση. Όσο πιο κινούμενα/προσαρμόσιμα μέρη διαθέτει, τόσο πιο δύσκολο είναι να καταλάβουμε πώς να τα λειτουργήσουμε σε συνθήκες 10 η 100 η 1000 φορές περισσότερων δεδομένων.

"**Όταν αρχίζετε την επέκταση, ο ακριβής προσδιορισμός της υφιστάμενης κατάστασης αποτελεί ένα πραγματικό πρόβλημα καθώς είναι πολύ δύσκολο να εξισορροπηθεί το φορτίο των διαφόρων αντικειμένων σε όλους τους διαθέσιμους εξυπηρετητές**" - είναι δύσκολος ο χειρισμός αποτυχιών, γιατί το να χάνεις την υφιστάμενη κατάσταση και να την αναδημιουργήσεις είναι δύσκολο. Η χρήση αντικειμένων αγνώστου κατάστασης είναι προτιμότερος.

"Η αποτυχία είναι αναπόφευκτη – ο πολλαπλασιασμός και η ανθεκτικότητα στην αποτυχία είναι η λύση". Πρέπει να μπορείτε να χειριστείτε την αποτυχία και να είστε έτοιμοι για προβλήματα με πολλά μέρη του συστήματος, που πολλές φορές μάλιστα εκδηλώνονται ταυτόχρονα.

Για να υποστηριχθεί η επεξεργασία των Big Data, οι διαφορετικές πλατφόρμες ενσωματώνουν την κλιμάκωση σε διαφορετικές μορφές. Από μια ευρύτερη προοπτική, οι μεγάλες πλατφόρμες δεδομένων μπορούν να κατηγοριοποιηθούν στους ακόλουθους δύο τύπους κλιμάκωσης *[8.Singh, Reddy]:*

Οριζόντια Κλιμάκωση: Περιλαμβάνει τη κατανομή του φόρτου εργασίας σε πολλους εξυπηρετητές που μπορεί να είναι ακόμη και μηχανές βασικών προϊόντων. Είναι επίσης γνωστή ως "**scale out** ", όπου προστίθενται πολλαπλές ανεξάρτητες μηχανές για να βελτιώσουν την ικανότητα επεξεργασίας. Συνήθως, υπάρχουν πολλές εγκαταστάσεις του λειτουργικού συστήματος που λειτουργούν σε ξεχωριστές μηχανές.

Κάθετη κλιμάκωση: Περιλαμβάνει την εγκατάσταση περισσότερων επεξεργαστών, περισσότερης μνήμης και ταχύτερου υλικού, συνήθως, μέσα σε ένα μόνο διακομιστή. Είναι επίσης γνωστό ως "**scale up**" και συνήθως περιλαμβάνει μια μοναδική εγκατάσταση ενός λειτουργικού συστήματος.

	Πλεονεκτήματα	**Μειονεκτήματα**
Οριζόντια Κλιμάκωση	-Αυξάνει την απόδοση σε μικρά βήματα όσο χρειάζεται -Το κόστος της επένδυσης που απαιτείται συνήθως είναι λιγότερο -Μπορεί να επεκταθεί σε όσο βαθμό χρειάζεται	-Το λογισμικό πρέπει να χειριστεί όλη την απαραίτητη κατανομή λογισμικού και τον παραληλλισμό -Πολύ μικρός αριθμός υφισταμένων λύσεων λογισμικού μπορεί να εποφεληθεί ικανοποιητικά
Κάθετη Κλιμάκωση	-Μπορεί να οφελήσει πιο εύκολα το μεγαλύτερο μέρος του λογισμικού -Η εγκατάσταση και διαχείριση υλικού εντός μιας μόνο μηχανής είναι πιο εύκολη	-Απαιτεί εκτεταμένες οικονομικές επενδύσεις -Το σύστημα οφείλει να είναι σε θέση να χειριστεί και μελλοντική αύξηση του φόρτου εργασίας η οποία όμως δεν μπορεί να εκτιμηθεί εκ των προτέρων -Πιθανόν η περαιτέρω κλιμάκωση να μήν είναι εφικτή πέρα ενός συγκεκριμένου ορίου

Παρότι η κάθετη κλιμάκωση μπορεί υπο συνθήκες να κάνει την διαχείριση και την εγκατάσταση ευκολότερη, περιορίζει την ικανότητα κλιμάκωσης της πλατφόρμας δεδομένου ότι κάτι τέτοιο θα απαιτήσει εκ νέου σημαντικές οικονομικές επενδύσεις. Για την διαχείριση του μελλοντικού φόρτου εργασίας, πάντα χρειάζεται να προστίθεται υλικό που είναι ισχυρότερο από τις τρέχουσες απαιτήσεις λόγω του περιορισμένου χώρου και δυνατοτήτων επέκτασης που είναι διαθέσιμες σε κάθε μηχάνημα. Αυτό αναγκάζει τον χρήστη να επενδύσει περισσότερο από ότι απαιτείται για τις τρέχουσες ανάγκες επεξεργασίας.

Από την άλλη πλευρά, η οριζόντια κλιμάκωση δίνει στους χρήστες τη δυνατότητα να βελτιώνουν διαρκώς την απόδοση με μικρές αυξήσεις κάθε φορά που απαιτούν κατά βάση μικρότερες χρηματοοικονομικές επένδυσεις. Επίσης, δεν υπάρχει όριο για την ποσότητα κλιμάκωσης που μπορεί να γίνει και μπορεί κανείς να κλιμακώνει οριζόντια το σύστημα όσο χρειάζεται. Παρά τα πλεονεκτήματα αυτά, το κύριο μειονέκτημα είναι η περιορισμένη διαθεσιμότητα εξειδικευμένου λογισμικού που μπορεί να χρησιμοποιηθεί αποτελεσματικά για την ενορχήστρωση της οριζόντιας κλιμάκωσης.

2.8.1 Οριζόντια Κλιμάκωση στην Υποδομή

[4.NIST]

Συστήμα Αρχείων Διαμοιράζομενων Δίσκων | Shared-Disk File Systems

Προσεγγίσεις, όπως τα SAN και το NAS, χρησιμοποιούν μια συγκεκριμένη δεξαμενή αποθήκευσης (storage pool), η οποία είναι προσβάσιμη από πολλούς υπολογιστικούς πόρους. Ενώ οι τεχνολογίες αυτές επιλύουν πολλές πτυχές της πρόσβασης σε πολύ μεγάλα σύνολα δεδομένων από πολλαπλούς κόμβους ταυτόχρονα, υποφέρουν σε ζητήματα που σχετίζονται με το κλείδωμα (data locking) και τις ταυτόχρονες ενημερώσεις (simultaneously update) και, το σημαντικότερο, δημιούργούν ένα στένωμα απόδοσης(bottleneck) (από κάθε λειτουργία εισόδου / εξόδου) που περιόριζει την ικανότητά τους να κλιμακώνονται για να καλύψουν τις ανάγκες πολλών εφαρμογών Big Data. Αυτοί οι περιορισμοί ξεπεράστηκαν με την εφαρμογή πλήρως κατανεμημένων συστημάτων αρχείων.

Κατανεμημένα Συστήματα Αρχείων | Distributed File Systems

Στα κατανεμημένα συστήματα αποθήκευσης αρχείων, τα πολλαπλως δομημένα σύνολα δεδομένων διανέμονται σε όλους τους υπολογιστικούς κόμβους της συστοιχίας διακομιστών (server cluster). Τα δεδομένα μπορούν να διανεμηθούν σε επίπεδο αρχείου/συνόλου (dataset) ή πιο συχνά σε επίπεδο μπλοκ (block), επιτρέποντας ταυτόχρονα σε πολλαπλούς κόμβους της συστοιχίας να αλληλεπιδρούν με διαφορετικά μέρη ενός μεγάλου αρχείου / συνόλου δεδομένων ταυτόχρονα. Τα Big Data frameworks σχεδιάζονται συχνά για να εκμεταλλευτούν την χωροθέτηση των δεδομένων (data locality) σε κάθε κόμβο κατά τη διανομή της επεξεργασίας, γεγονός που μειώνει την ανάγκη μετακίνησης των δεδομένων μεταξύ κόμβων. Επιπλέον, πολλά συστήματα κατανεμημένων αρχείων εφαρμόζουν επίσης αναπαραγωγή σε επιπέδο αρχείου/μπλοκ όπου κάθε αρχείο/μπλοκ αποθηκεύεται πολλές φορές σε διαφορετικές μηχανές τόσο για λόγους αξιοπιστίας/ανάκτησης (τα δεδομένα δεν χάνονται αν

αποτύχει ένας κόμβος της συστοιχίας) όσο και για ενίσχυση της τοπικότητας. Οποιοσδήποτε τύπος δεδομένων και πολλά μεγέθη αρχείων μπορεί να αντιμετωπιστεί χωρίς ETL διαδικασίες, με ορισμένες τεχνολογίες να έχουν αξιοσημείωτα καλύτερη απόδοση για μεγάλα μεγέθη αρχείων.

Διαμοιραζόμενης Μνήμης | Shared Memory

Όταν η ποσότητα των δεδομένων είναι πολύπλοκη, αδόμητη ή απαιτούνται πολλοί αλγόριθμοι στα δεδομένα, θεωρείται προτιμότερη η δημιουργία ενός σύστηματος κοινής μνήμης.Το μεγαλύτερο μέρος των δεδομένων δηλαδή θα μπορούσε να διατηρηθούν στη μνήμη του συστήματος και οι διαφορετικές διαδικασίες να τα χρησιμοποιούν απευθείας. Για παράδειγμα, η παρακολούθηση χιλιάδων τροφοδοσιών βίντεο για τον προσδιορισμό οποιασδήποτε συσχέτισης μεταξύ των εικόνων θα μπορούσε να επωφεληθεί από τη διατήρηση όλων των τροφοδοσιών αυτών στην κύρια μνήμη και την ύπαρξη πολλαπλών εφαρμογών που θα επεξεργάζονται τα δεδομένα. Με την προσέγγιση κοινόχρηστης μνήμης, οι εφαρμογές γίνονται ευκολότερες στην ανάπτυξη καθώς και στην αποσφαλμάτωση. Παρότι η διαχείριση ενός και μοναδικού συστήματος μεγάλης κλίμακας και η αποθήκευση και επεξεργασία όλων των δεδομένων είναι ευκολότερη, αυτού του τύπου τα συστήματα συνήθως είναι αρκετά πιο ακριβά.Η μείωση δηλαδή του κόστους και της πολυπλοκότητας διαχείρισης έχει αντιστάθμισμα το υψηλό κόστος υλικού.

Ομότιμα Δίκτυα | Peer-to-Peer Networks

Τα δίκτυα Peer-to-Peer περιλαμβάνουν εκατομμύρια μηχανές συνδεδεμένες σε ένα δίκτυο. Είναι μια αποκεντρωμένη και κατανεμημένη αρχιτεκτονική δικτύου όπου οι κόμβοι στα δίκτυα (γνωστοί ως ομότιμοι) χρησιμεύουν ως πόροι αλλά και καταναλώνουν και οι ίδιοι πόρους. Είναι μια από τα παλαιότερες τεχνικές κατανεμημένης υπολογιστικής που υπάρχουν. Συνήθως, μια διεπαφή μετάδοσης μηνυμάτων (Message Passing Interface MPI) είναι το σχήματα επικοινωνίας που χρησιμοποιείται σε μια τέτοια εγκατάσταση για την επικοινωνία και την ανταλλαγή των δεδομένων μεταξύ των ομότιμων συμμετεχόντων. Κάθε κόμβος μπορεί να αποθηκεύσει τις παρουσίες δεδομένων και το scale out είναι πρακτικά απεριόριστο (μπορεί να είναι εκατομμύριων κόμβων). Η κύρια αιτία συμφόρησης σε μια τέτοια εγκατάσταση έγκειται στην επικοινωνία μεταξύ διαφορετικών κόμβων δεδομένων. Η μετάδοση μηνυμάτων σε δίκτυο peer-to-peer είναι φθηνότερη αλλά οι λειτουργίες συνδυασμού και συνάθροισης αποτελεσμάτων είναι πολύ πιο ακριβή. Επιπλέον, τα μηνύματα αποστέλλονται μέσω του δικτύου με τη μορφή ενός spanning δέντρου ορίζοντας αυθαίρετα έναν κόμβο ως τη ρίζα όπου γίνεται η εκκίνηση. Το MPI, το οποίο είναι το πρότυπο παράδειγμα επικοινωνίας λογισμικού που χρησιμοποιείται σε αυτό το δίκτυο, έχει χρησιμοποιηθεί αρκετά χρόνια και είναι καλά εδραιωμένο και λεπτομερώς αποσφαλματωμένο. Ενα εκ των κύριων χαρακτηριστικών του MPI είναι η διαδικασία διατήρησης κατάστασης συντήρησης, δηλαδή οι διαδικασίες μπορούν να ζήσουν όσο λειτουργεί το σύστημα και δεν χρειάζεται να διαβάζουμε τα ίδια δεδομένα ξανά και ξανά όπως στην περίπτωση άλλων λύσεων (πχ όπως το MapReduce). Όλες οι παράμετροι δηλαδή μπορούν να διατηρηθούν τοπικά. Ως εκ τούτου, σε αντίθεση με MapReduce, το MPI είναι κατάλληλη επεξεργασία που απαιτεί επαναλήψεις. Ένα άλλο χαρακτηριστικό του MPI είναι η ιεραρχιση προτύπου master/slave. Όταν το MPI δηλαδή αναπτύσσεται ως μοντέλο master-slave, η slave η μηχανή μπορεί να ο μετατραπεί σε master για άλλες διαδικασίες. Αυτό μπορεί να είναι εξαιρετικά χρήσιμο για περιπτώσεις όπου απαιτείται δυναμική κατανομή πόρων λόγω του ότι οι σκλάβοι έχουν τεράστια ποσά δεδομένων για επεξεργασία. Το MPI τέλος διατίθεται για πολλές γλώσσες προγραμματισμού. Περιλαμβάνει μεθόδους για την αποστολή και λήψη μηνυμάτων και δεδομένων. Ορισμένες άλλες διαθέσιμες μέθοδοι με το MPI είναι τύπου “εκπομπης” όπου χρησιμοποιείται για τη μετάδοση των δεδομένων ή των μηνυμάτων σε όλους τους κόμβους καθώς το "Barrier", το οποίο είναι μια άλλη μέθοδος που μπορεί να θέσει ένα εμπόδιο-έλεγχο και επιτρέπει σε όλες τις διαδικασίες να συγχρονίζονται και να απαιτεί να φτάσουν μέχρι ένα συγκεκριμένο σημείο προτού συνεχίσουν.Αν και το MPI φαίνεται να είναι ιδανικό για την ανάπτυξη αλγορίθμων για αναλύσεις σε Big Data, έχει μερικά σημαντικά μειονεκτήματα. Ένα από τα κύρια μειονεκτήματα είναι η δυσανεξία σε σφάλματα δεδομένου ότι η MPI δεν διαθέτει μηχανισμό αντιμετώπισης σφαλμάτων. Όταν χρησιμοποιείται πάνω από τα ομότιμα δίκτυα, που είναι εντελώς αναξιόπιστο ως υλικό, μια αποτυχία ενός κόμβου μπορεί να προκαλέσει την κατάρρευση ολόκληρου το σύστηματος. Οι χρήστες πρέπει να εφαρμόσουν κάποιο είδους μηχανισμό ανοχής σφάλματος μέσα στο πρόγραμμα για να αποφευχθούν τέτοιες ατυχείς καταστάσεις. Με την διάδοση αλλων είδων τεχνολογιών(όπως οι Hadoop που είναι ανθεκτικές στην ανοχή σφάλματος)να καθίστανται ευρέως δημοφιλείς, η χρήση του MPI πλέον είναι περιορισμένη.

2.8.2 Πλατφόρμες Κάθετης Κλιμάκωσης

[8.Singh, Reddy]

Υψηλής Απόδοσης Συστοιχίες Υπολογιστών (HPC)

Τα clusters HPC , που ονομάζονται επίσης λεπίδες ή υπερυπολογιστές, είναι μηχανές με χιλιάδες πυρήνες επεξεργασίας. Μπορούν να έχουν διαφορετική ποικιλία ως προς την οργάνωση των δίσκων, μνήμης cache, μηχανισμους επικοινωνίας κ.λπ. ανάλογα με τις απαιτήσεις του χρήστη. Αυτά τα συστήματα χρησιμοποιούν καλοσχεδιασμένο και ισχυρό υλικό το οποίο είναι βελτιστοποιημένο για ταχύτητα και απόδοση. Λόγω της κορυφαίας ποιότητας υλικού υψηλής , η ανοχή σε σφάλματα σε τέτοια συστήματα δεν είναι προβληματική δεδομένου ότι οι αποτυχίες υλικού είναι εξαιρετικά σπάνιες. Το αρχικό κόστος ανάπτυξης ενός τέτοιου συστήματος μπορεί να είναι πολύ υψηλή λόγω της χρήσης του υλικού υψηλής τεχνολογίας. Δεν είναι τόσο επεκτάσιμες υποδομές όσο πχ Hadoop ή Spark, αλλά είναι ακόμα σε θέση να επεξεργαστούν terabyte δεδομένων.Όμως το κόστος κλιμάκωσης ενός τέτοιου συστήματος είναι πολύ υψηλότερο σε σύγκριση με τις ομάδες Hadoop ή Spark. Το σχήμα επικοινωνίας που χρησιμοποιείται για τέτοιες πλατφόρμες είναι τυπικά MPI.

Multicore CPU

Το Multicore αναφέρεται σε ένα μηχάνημα που διαθέτει δεκάδες πυρήνες επεξεργασίας . Συνήθως έχουν κοινή μνήμη αλλά μόνο ένα δίσκο. Τα τελευταία χρόνια, οι CPU έχουν αποκτήσει εσωτερικό παραλληλισμό. Ακόμα πιο πρόσφατα, ο αριθμός των πυρήνων ανά πλακέτα καθώς και ο αριθμός των διεργασιών που μπορεί να εκτελέσει ένας πυρήνας έχουν αυξηθεί σημαντικά. Νεότερες εκδόσεις μητρικών καρτών επιτρέπουν πολλαπλές CPU σε ένα μόνο μηχάνημα αυξάνοντας έτσι τον παραλληλισμό. Μέχρι τα τελευταία χρόνια, οι επεξεργαστές ήταν κυρίως υπεύθυνοι για την εκτέλεση των αλγορίθμων σε αναλύσεις Big Data.

Ο παραλληλισμός του φόρτου στις επιμέρους CPU επιτυγχάνονται κυρίως μέσω πολυνηματικής επεξεργασίας (multithread). Όλοι οι πυρήνες μοιράζονται την ίδια μνήμη. Η εργασία πρέπει να διαμοιραστεί σε νήματα(threads). Το κάθε νήμα εκτελείται παράλληλα σε διαφορετικούς πυρήνες CPU. Οι περισσότερες από τις γλώσσες προγραμματισμού παρέχουν βιβλιοθήκες για την δημιουργία threads και την εκμετάλλευση του παραλληλισμού της CPU. Η πιο δημοφιλής επιλογή αυτού του τύπου προγραμματισμού είναι η γλώσσα Java. Δεδομένου ότι οι CPU πολλαπλών πυρήνων υπάρχουν εδώ και αρκετά χρόνια, ένας μεγάλος αριθμός των εφαρμογών λογισμικού και προγραμματιστικών περιβαλλοντων είναι καλά αναπτυγμένα για χρήση σε τέτοιες πλατφόρμες. Οι αναπτύξεις στις CPU δεν είναι του ίδιου ρυθμού συγκρινόμενες με τις GPU. Ο αριθμός των πυρήνων ανά CPU παραμένει σε διπλά ψηφία με την ισχύ επεξεργασίας κοντά στα 10Gflops ενώ μία GPU έχει περισσότερους από 2500 πυρήνες επεξεργασίας με 1000Tflops της ισχύος επεξεργασίας. Αυτός ο μαζικός παραλληλισμός στις GPU τις καθιστά μια απο τις πιο ελκυστικες επιλογές για εφαρμογές παράλληλου υπολογισμού.

Το μειονέκτημα των CPU είναι ο περιορισμένος αριθμός πυρήνων επεξεργασίας τους και η εξάρτηση τους από τη μνήμη του συστήματος για πρόσβαση σε δεδομένα. Η μνήμη του συστήματος περιορίζεται σε μερικές εκατοντάδες gigabytes και αυτό περιορίζει το μέγεθος των δεδομένων που μια CPU μπορεί να επεξεργαστεί αποτελεσματικά.

Όταν το μέγεθος των δεδομένων υπερβεί τη μνήμη του συστήματος, η πρόσβαση στο δίσκο προκαλέι τεράστια συμφόρηση. Ακόμα όταν τα δεδομένα πληρώσουν στη μνήμη του συστήματος, η CPU μπορεί να επεξεργάζεται δεδομένα πολύ πιο γρήγορα από την ταχύτητα πρόσβασης στη μνήμη που καθιστά τη πρόσβαση στην μνήμη δυσχερή. Οι GPU αποφεύγει αυτό το πρόβλημα χρησιμοποιώντας τη μνήμη τύπου DDR5 σε σύγκριση με την πιο αργή μνήμη DDR3 που χρησιμοποιείται συνήθως. Επίσης, η GPU έχει κρυφή μνήμη υψηλής ταχύτητας για κάθε πολυεπεξεργαστή το οποίο επίσης που επιταχύνει την πρόσβαση δεδομένων.

Μονάδα Επεξεργασίας Γραφικών | GPU- Graphics Processing Unit

Η Μονάδα Επεξεργασίας Γραφικών (GPUs) είναι ένα εξειδικευμένο υλικό σχεδιασμένο να επεξεργάζεται τη δημιουργία εικόνων σε ένα πλάισιο προσωρινής μνήμης που προορίζεται για την έξοδο οθόνης/προβολής. Μέχρι πρόσφατα, οι GPU χρησιμοποιήθηκαν κυρίως για γραφικές λειτουργίες όπως την επεξεργασία βίντεο και εικόνων, την επιτάχυνση της επεξεργασίας γραφικών κλπ. Ωστόσο, λόγω της αρχιτεκτονικής παράλληλης μαζικής επεξεργασίας, τις πρόσφατες εξελίξεις στο υλικό GPU και τα σχετικά προγραμματικά πλαισία έχουν οδηγήσει στην ανάπτυξη σε GPGPU (general-purpose computing on graphics processing units -Υπολογισμοί Γενικού Σκοπού για Μονάδες Επεξεργασίας Γραφικών). Η GPU διαθέτει μεγάλο αριθμό πυρήνων επεξεργασίας (συνήθως

γύρω στους 3000) σε σύγκριση με μια multicore CPU. Εκτός από τους πυρήνες επεξεργασίας, η GPU συνήθως διαθέτει δικής της μνήμη υψηλής χωρητικότητας DDR5 η οποία είναι πολλές φορές γρηγορότερη μια τυπικης DDR3 μνήμης. Η απόδοση της GPU έχει αυξηθεί σημαντικά τα τελευταία χρόνια σε σύγκριση με εκείνη των CPU. Πρόσφατα για παράδειγμα η Nvidia ξεκίνησε την σειρά Tesla GPU που έχουν σχεδιαστεί ειδικά για υπολογισμούς υψηλής απόδοσης. Η Nvidia κυκλοφόρησε επίσης το πλαίσιο CUDA το οποίο έκανε τον προγραμματισμό GPU προσιτό σε όλους τους προγραμματιστές χωρίς να χρειάζεται από πλευράς τους εμβάθυνση στις λεπτομέρειες του υλικού. Αυτές οι εξελίξεις δείχνουν ότι ο GPGPU κερδίζει πράγματι μεγαλύτερη δημοτικότητα.

Μια GPU έχει συνήθως δύο επίπεδα παραλληλισμού. Στο πρώτο επίπεδο, υπάρχουν αρκετοί πολυεπεξεργαστές (Multiprocessors MPs) και μέσα σε κάθε πολυεπεξεργαστή υπάρχουν πολλοί επεξεργαστές ροής (Stream Proccessors SPs). Για να λειτουργήσει αυτήν τη διάταξη, το πρόγραμμα GPU χωρίζεται σε νήματα τα οποία εκτελούνται στα SP και επιπλεόν είναι ομαδοποιημένα ώστε να σχηματίσουν ένα μπλοκ νημάτων που θα τρέξει σε έναν πολυεπεξεργαστή. Τα νήματα εντός του μπλοκ μπορούν να επικοινωνούν μεταξύ τους και να συγχρονίζονται. Κάθε ένα από αυτά τα νήματα έχει πρόσβαση σε μικρή αλλά εξαιρετικά γρήγορη προσωρινή μνήμη και στην ακόμα μεγαλύτερη καθολική κεντρική μνήμη.Τα νήματα δεν μπορούν να επικοινωνήσουν με νήματα άλλου μπλοκ καθώς αυτά μπορεί να έχουν προγραμματιστεί με διαφορετικούς χρόνισμους. Αυτή η αρχιτεκτονική υποδηλώνει ότι για κάθε εργασία για να μπορεί να τρέξει σε GPU, πρέπει να σπάσει σε μπλοκ υπολογισμού που θα μπορούν να τρέξουν ανεξάρτητα χωρίς μεταξύ τους επικοινωνία. Αυτά τα μπλοκ έπειτα θα πρέπει να σπάσουν περαιτέρω σε μικρότερες εργασίες που θα εκτελούνται σε ένα μεμονωμένο νήμα το οποίο μπορεί να επικοινωνεί με άλλα νήματα στο ίδιο μπλοκ.

Οι GPU χρησιμοποιήθηκαν για την ανάπτυξη γρηγορότερων αλγορίθμων μηχανικής μάθησης. Ορισμένες βιβλιοθήκες όπως ο GPUMiner εφαρμόζουν λίγους αλγόριθμους μηχανικής μάθησης σε GPU χρησιμοποιώντας το πλαίσιο CUDA. Τα πειράματα έχουν δείξει πως σε τέτοια προβλήματα υπάρχει κέρδος πολλών κύκλων ταχύτητας χρησιμοποιώντας μια GPU σε σύγκριση με μια πολυεπίπεδη CPU.

Η GPU απο την άλλη όμως έχει τα δικά της μειονεκτήματα. Το κύριο μειονέκτημα είναι η περιορισμένη αυτόνομη μνήμη που μπορεί να περιέχει. Με μέγιστη μνήμη 12 GB ανά GPU (μέχρι τη σημερινή γενιά), δεν θεωρούνται κατάλληλες για την επεξεργασία δεδομένων κλίμακας terabyte.Επίσης μόλις το μέγεθος των δεδομένων γίνει μεγαλύτερο από το μέγεθος της μνήμης GPU, η απόδοση μειώνεται σημαντικά καθώς απαιτείται πρόσβαση στο δίσκο. Ένα άλλο μειονέκτημα είναι ο περιορισμένος αριθμός λύσεων λογισμικού και οι αλγόριθμοι που είναι διαθέσιμοι για GPU. Λόγω του τρόπου με τον οποίο απαιτεί η κατανομή των εργασιών για εκτέλεση στην GPU, πολλοί από τους υπάρχοντες αναλυτικούς αλγόριθμοι δεν είναι εύκολα μεταφέρσιμοι σε GPU.

Προγραμματιζόμενες Μήτρες Πυλών Πεδίου (FPGA - Field Programmable Gate Arrays)

Τα FPGA είναι εξαιρετικά εξειδικευμένες μονάδες υλικού που είναι ειδικά σχεδιασμένες για συγκεκριμένες εφαρμογές. Μπορούν να βελτιστοποιηθούν πολύ για ταχύτητα και να γίνουν τάξεις μεγέθους πιο γρήγορες σε σύγκριση με άλλες πλατφόρμες .Προγραμματίζονται χρησιμοποιώντας γλώσσα περιγραφής υλικού (Hardware descriptive language -HDL). Λόγω του προσαρμοσμένου υλικού, το κόστος ανάπτυξης τους είναι συνήθως πολύ μεγαλύτερο σε σύγκριση με άλλες πλατφόρμες. Έπειτα από πλευράς λογισμικού, η κωδικοποίηση πρέπει να γίνει σε HDL και απαιτεί γνώση του υλικού σε πολύ λεπτομερές και χαμηλό επιπεδο που αυξάνει πολύ το κόστος ανάπτυξης αλγορίθμων. Ο χρήστης πρέπει να διερευνήσει προσεκτικά την καταλληλότητα μιας συγκεκριμένης εφαρμογής για FPGA δεδομένου ότι είναι αποτελεσματικές μόνο για ένα ορισμένο σύνολο εφαρμογών.

Τα FPGAs χρησιμοποιούνται σε διάφορες εφαρμογές πραγματικού κόσμου. Ένα παράδειγμα όπου αναπτύχθηκαν με επιτυχία είναι στις εφαρμογές ασφάλειας δικτύων. Σε μια μια τέτοια εφαρμογή, το FPGA χρησιμοποιείται ως τείχος προστασίας υλικού και είναι πολύ γρηγορότερο από το λογισμικό firewalls για τη σάρωση μεγάλων ποσοτήτων δεδομένων δικτύου. Τα τελευταία χρόνια βέβαια, η ταχύτητα των multicore επεξεργαστών έχει αρχίσει να πλησιάζει περισσότερο την ταχύτητα των FPGAs.

2.9 Διασύνδεση στα Big Data | Big Data Integration

[37. Hadi Fadlallah]

2.9.1 Integration σε Επίπεδο Δεδομένων

Τα στοιχεία της πλατφόρμας Big Data διαχειρίζονται δεδομένα με νέους τρόπους σε σχέση με την παραδοσιακή βάση δεδομένων(σχεσιακή) και ο λόγος είναι η ανάγκη επεκτασιμότητας και υψηλής απόδοσης για τη διαχείριση τόσο δομημένων όσο και μη δομημένων δεδομένων. Για όλα τα στοιχεία των οικοσυστημάτων Big Data στις εγκαταστάσεις Hadoop και NoSQL ισχύει πως το καθένα από αυτά έχει τη δική του προσέγγιση για την εξαγωγή, μετατροπή και φόρτωση δεδομένων.

Επιπλέον, τα παραδοσιακά εργαλεία ETL εξελίσσονται για να χειριστούν τα νέα χαρακτηριστικά που επιτάσσουν τα Big Data. Ενώ οι παραδοσιακές μορφές διασύνδεσης (integration) έχουν λάβει νέες ερμηνείες στον κόσμο των Big Data, εξακολουθούν να χρειάζονται μια κοινή πλατφόρμα που να υποστηρίζει την ποιότητα και τη προτυποποίηση των δεδομένων.

Η παραδοσιακή διασύνδεση(integration) των δεδομένων πραγματοποιούνταν ανέκαθεν χρησιμοποιώντας επεξεργασία τύπου batch (δεδομένα σε αναμονή), ενώ η διασύνδεση στα Big Data μπορεί να γίνει τόσο σε πραγματικό χρόνο όσο και σε επεξεργασία batch. Αυτό κάνει τις φάσεις ETL να αναδιατάσσονται ώστε να γίνουν ELT σε ορισμένες περιπτώσεις, έτσι τα δεδομένα αφου εξάγονται,πρώτα φορτώνονται σε κατανεμημένα συστήματα αρχείων και στη συνέχεια μετασχηματίζονται πριν χρησιμοποιηθούν.

Προκειμένου να ληφθούν καλές επιχειρηματικές αποφάσεις βάσει ανάλυσης σε μεγάλο όγκο δεδομένων, τα δεδομένα πρέπει να είναι αξιόπιστα και κατανοητά σε όλα τα επίπεδα του οργανισμού. Πρέπει να παραδοθεί προς χρήση με έναν αξιόπιστο, ελεγχόμενο, συνεπή και ευέλικτο τρόπο σε όλες τις μονάδες της επιχείρησης. Για την επίτευξη αυτού του στόχου, χρησιμοποιούνται τρεις βασικές τεχνικές:

- Χαρτογράφηση Σχήματος | Schema Mapping
- Τεκμηρίωση σε επίπεδο Εγγραφής | Record Linkage
- Συγκερασμός Δεδομένων | Data Fusion

Το Schema Mapping και το Record Linkage χρησιμοποιήθηκαν και στην παραδοσιακή διασύνδεση δεδομένων, αλλά περιλαμβάνονταν μόνο στη διαδικασία ETL και ήταν πολύ λιγότερα υποσχόμενα σε σχέση με το πως ενσωματώθηκαν στις τεχνολογίες Big Data. Στις επόμενες ενότητες, θα περιγραφούν καθεμία από αυτές τις τεχνικές και θα παρατεθεί το ποιες είναι οι τελευταίες έρευνες που έγιναν για την υιοθέτηση αυτών των τεχνικών στο Big Data Integration.

Schema Mapping

Γενικά

Στα αρχικά στάδια της ανάλυσης μεγάλων δεδομένων, είναι απιθανό να υπάρχει το ίδιο επίπεδο ελέγχου των ορισμών δεδομένων όπως συμβαίνει στα ιδιόκτητα επιχειρησιακά δεδομένα. Ωστόσο, μόλις εντοπιστούν τα μοτίβα που σχετίζονται περισσότερο με τους σκοπούς της επιχείρησης, θα πρέπει να υπάρχει η δυνατότητα αντιστοίχισης των στοιχείων των δεδομένων σε κοινούς ορισμούς. Αυτός ο κοινός ορισμός μεταφέρεται στη συνέχεια σε καθημερινά δεδομένα λειτουργίας, αποθήκες δεδομένων, αναφορές και επιχειρησιακές διαδικασίες.

Το Schema Mapping μπορεί να θεωρηθεί ως διαδικασία που αποτελείται από δύο φάσεις. Η πρώτη, δημιουργώντας ένα ενδιάμεσο (καθολικό) σχήμα και, στη συνέχεια, εντοπίζοντας τις αντιστοιχίσεις μεταξύ του ενδιάμεσου σχήματος και του τοπικού σχήματος των πηγών δεδομένων να προσδιορίστεί ποια συγκεκριμένα χαρακτηριστικά (ή σύνολα) περιέχουν τις ίδιες πληροφορίες.

Για παράδειγμα, σκεφτείτε ότι συλλέγουμε δεδομένα σχετικά με παίκτες του "Cricket". Έχουμε τρεις πηγές δεδομένων: S1, S2 και S3. Κάθε πηγή δεδομένων περιέχει διαφορετικό χαρακτηριστικό όπως φαίνεται στον παρακάτω πίνακα:

S1	Name, Games, Runs
S2	Name, Team, Score
S3	Name, Club, Matches

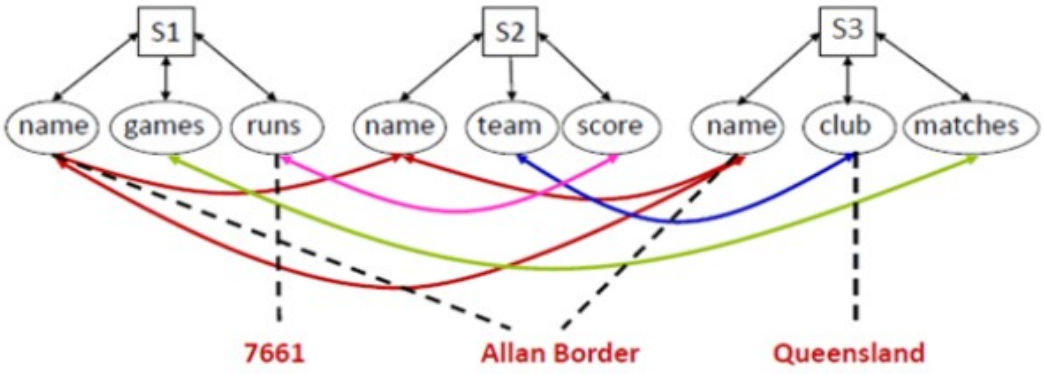
Εικόνα 15.Παράδειγμα Διαφορετικών Καταχωρήσεων

Πρώτον, προσδιορίζουμε τέσσερα χαρακτηριστικά που θέλουμε στο διαμεσολαβητικό σχήμα, που είναι το όνομα, ο αριθμός των παιχνιδιών που παίζονται, το συνολικό σκορ και η ομάδα. Στη συνέχεια, κάνουμε τη χαρτογράφηση μεταξύ του σχήματος διαμεσολάβησης και του σχήματος των πηγών δεδομένων, όπως φαίνεται στον παρακάτω πίνακα:

MS attribute	S1	S2	S3
Name	Name	Name	Name
Number of games	Games		Matches
Total score	Runs	Score	
Team		Team	Club

Εικόνα 16.Χαρτογράφηση βάση Σχήματος Διαμεσολάβησης

Αφού ολοκληρωθεί η αντιστοίχιση, μπορούμε να υποβάλουμε ερώτημα σε όλες τις πηγές δεδομένων, καθώς είναι πλέον μία λογική πηγή δεδομένων. Η παρακάτω εικόνα δείχνει ένα παράδειγμα για το πώς μπορούμε να συλλέγουμε δεδομένα κατά την αναζήτηση ενός παίκτη που ονομάζεται "Allan Border".



Εικόνα 17.Αναζήτηση σε Διασύνδεμένα Δεδομένων

Μέθοδοι Schema Mapping

Πιθανοτική Ευθυγράμμιση Σχήματος | Probabilistic Schema Alignment

Για να αντιμετωπιστεί η ασάφεια των χαρακτηριστικών, γίνεται προσθήκη πιθανοτήτων στην αντιστοίχιση χαρακτηριστικών και στη χαρτογράφηση σχημάτων (προτάθηκε από τον Das Sarma) και αξιολογήθηκε σε πίνακες ιστού που ανιχνεύτηκαν από πέντε διαφορετικά domains, με καθένα από τα domain περιέχει περίπου 50-800 πίνακες ιστού. Επιπλέον, παρατηρήθηκε πως συνολικά δίνει καλύτερα αποτελέσματα από ντετερμινιστικές μέθοδους

Διασύνδεση Δεδομένων Deep Web | Integrating Deep Web Data

Ο όρος «Deep Web» χρησιμοποιείται για να περιγράψει τις ιστοσελίδες που δεν είναι προσβάσιμες από τις μηχανές αναζήτησης. Προκειμένου να προσφέρει πρόσβαση σε deep web, ένας αλγόριθμος προτάθηκε από τον Madhavan και δοκιμάστηκε σε ένα δείγμα 500.000 φορμών HTML και πέτυχε μια καλή κάλυψη της υποκείμενης βάσης δεδομένων.

Διασύνδεση Πινάκων Ιστού | Integrating Web Tables

Οι "Πίνακες Ιστού" είναι ετερογενείς δεδομένων σε μορφή πίνακα που αποθηκεύονται σε HTML. Δεν έχουν ένα σαφές και ακριβές σχήμα. Ο Cafarella πρότεινε μια μέθοδο αναζήτησης λέξεων-κλειδιών σε αυτούς τους πίνακες ιστού βασισμένη σε δύο προσεγγίσεις κατάταξης: το Feature Rank και το Schema Rank.

Ο Das Sarma πρότεινε ένα πλαίσιο για την ανάκτηση πινάκων ιστού που σχετίζονται με έναν δοσμένο πίνακα. Πειραματίστηκαν το πλαίσιο αυτό στους πίνακες της Wikipedia και έλαβαν καλά αποτελέσματα.

Για την εξόρυξη γνώσης από τους Πίνακες Ιστού, ο Limaye πρότεινε μια λύση βασισμένη σε μοντέλο γράφων για την περιγραφή των Πίνακες Ιστού, και τα πειράματά τους έδειξαν ότι το μοντέλο αυτό κερδίζει μεγαλύτερη ακρίβεια από το παραδοσιακό μοντέλο.

Ιχνηλάτιση Εγγραφής | Record Linkage

Το Record Linkage είναι μια εργασία στην οποία εντοπίζουμε εγγραφές που αναφέρονται στην ίδια λογική οντότητα σε διαφορετικές πηγές δεδομένων, ειδικά όταν δεν γίνεται κοινή χρήση κοινού αναγνωριστικού μεταξύ

των πολλαπλών πηγών δεδομένων (όπως πχ ο αριθμός ταυτότητας ή το ΑΦΜ για άτομα). Στην παραδοσιακή ενοποίηση δεδομένων, κάτι τέτοιο λειτουργεί μόνο για τη σύνδεση δομημένων δεδομένων.

Στην διασύνδεση Big Data όμως, οι πηγές δεδομένων είναι ετερογενείς στις δομές τους και συλλέγονται από πολλές πηγές (μέσα κοινωνικής δικτύωσης, αρχεία καταγραφής αισθητήρων κ.α) οι οποίες συνήθως που παρέχουν μη-δομημένα δεδομένα σε μορφή κειμένου και ενώ ταυτόχρονα οι ίδιες οι πηγές δεδομένων είναι δυναμικές και συνεχώς εξελισσόμενες γεγονός που καθιστά αυτή την τεχνική πολύ δύσκολη.

Ως ένα πολύ απλό παράδειγμα του τι σημαίνει Record Linkage, ας συλλέξουμε δεδομένα για ασθενείς από διαφορετικό νοσοκομεία. Υποτεθίστω ότι συλλέγουμε δεδομένα για έναν ασθενή με το όνομα «Mohammad Hassan» και γεννήθηκε το έτος 1953, βρίσκουμε δύο αρχεία από δύο νοσοκομεία Α και Β ,όπως καταχωρήθηκαν στους παρακάτω πίνακες

Patient Name	Mohammad Hassan
Visit ID	837720
Date of Birth	20/03/1953
File No	00001245
Entry Date	01/03/2000

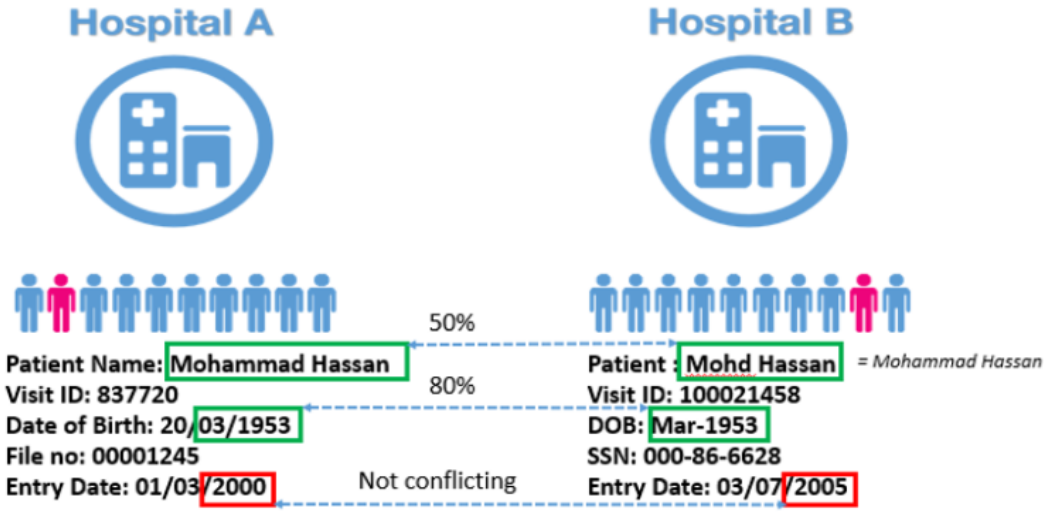
Table 3 - Record from Hospital A

Patient	Mohd Hassan
Visit ID	100021458
DOB	Mar-1953
SSN	000-86-6628
Entry Date	03/07/2005

Table 4 - Record from Hospital B

Εικόνα 18.Παράδειγμα Διαφορετικών Πηγών

Συγκρίνοντας μετά την αντιστοίχιση των χαρακτηριστικών αυτού του σχήματος και με κάποια εμπειρική σύγκριση μπορούμε να διακρίνουμε πως υπάρχει υψηλη πιθανότητα αυτές οι εγγραφές να ανήκουν στον ίδιο ασθενή. Η παρακάτω εικόνα δείχνει μια υπόθεση σύγκρισης:



Εικόνα 19.Πιθανατοτική Υπόθεση Σύγκρισης

Πρώτα η σύγκριση και των δύο ονομάτων δίνει πιθανότητα 50%, και στη συνέχεια η σύγκριση του έτους γέννησης αυξάνει αυτήν την πιθανότητα στο 80%. Επιπλέον, μπορούμε να βρούμε από τις ημερομηνίες εισαγωγής ότι οι εγγραφές δεν έρχονται σε διένεξη μεταξύ. Επομένως, μπορούμε να υποθέσουμε με πιθανότητα 80% ότι αυτές οι καταχωρίσεις ανήκουν στον ίδιο ασθενή.

Γενικα χρησιμοποιούνται αρκετές τεχνικές όπως:

Αντιστοίχιση κατά ζεύγη (Pairwise matching) : Αυτή η τεχνική χρησιμοποιείται για τη σύγκριση ενός ζεύγους εγγραφών για να ελέγξει εάν ανήκουν στην ίδια λογική οντότητα.

Συσταδοποίηση | Clustering: Αυτή η τεχνική χρησιμοποιείται για να ληφθεί μια καθολική συνεπής απόφαση για την αποτελεσματική τμηματοποίηση των εγγραφών μέσω της επιβεβαιώσης ότι κάθε κατάτμηση θα συμπεριλάβει μια ξεχωριστή οντότητα.

Πακετοποίηση | Blocking : Αυτή η τεχνική χρησιμοποιείται για την κατάτμηση των εισερχόμενων εγγραφών σε πολλαπλά μπλοκ, ώστε να επιτρέπεται μόνο αντιστοίχιση ζευγαριών με εγγραφές σε ένα ίδιο μπλοκ.

Μέθοδοι Ιχνηλάτιση Εγγραφής

Μετα-Αποκλεισμός: Κλάδεμα στην Αντιστοίχιση Ζευγαριών | Meta-Blocking: pruning pairwise matching's

Παπαδάκης και άλλοι πρότεινε μετα-αποκλεισμό αντί για προσέγγιση πολλαπλών αποκλεισμών βασισμένων στα κλειδιά, για να αντιμετωπιστεί κάποια ανεπάρκεια που παρουσιάστηκε κατά την εργασία με ετερογενή δεδομένα μεγάλης κλίμακας. Επιπλέον, έκανε πειράματα για να αξιολογήσει τα είδη κλαδέματος, και απέδειξε ότι το μετα-μπλοκάρισμα βελτιώνει την αποτελεσματικότητα του μπλοκαρίσματος.

Αυξητική Ιχνηλάτιση Εγγραφών | Incremental Record linkage

Οι Whang και Garcia-Molina έχουν επικεντρωθεί στην διαχρονική εξέλιξη των κανόνων αντιστοίχισης κατά ζεύγη και εντόπισαν μια γενική κατάσταση που μπορεί αυτό να εφαρμοστεί. Σε αυτή την κατεύθυνση ο Gruenheid πρότεινε κάποιες στοιχειώδεις τεχνικές που δίνουν περισσότερη αποτελεσματικότητα από τη ιχνηλάτιση μέσω διαδικασίας batch και τους παλιούς αλγόριθμους ιχνηλάτισης

Σύνδεση Τμημάτων από Δεδομένα Κειμένου σε Δομημένα Δεδομένα

Οι Cortez και da Silva πρότειναν τη χρήση τεχνικών εξαγωγής πληροφοριών από μη δομημένα δεδομένα για τη κατάρτιση δομημένων δεδομένων πριν από τη χρήση τεχνικών ιχνηλάτισης.

Για την σύνδεση εκατοντάδες χιλιάδες προσφορές προϊόντων με τις δομημένες πληροφορίες των προϊόντων, οι ο Kannan πρότεινε μια καινοτόμα προσέγγιση για τη σύνδεση τμηματων κειμένου σε δομημένα δεδομένα, η οποία βασίζεται κυρίως σε:

- Σημασιολογική ανάλυση αποσπάσματων κειμένου με χρήση ετικετών, ανάλυσης με λογικούς κανόνες και βέλτιστων τεχνικών δοκιμής
- Μια συνάρτηση αντιστοίχισης που επιστρέφει την πιθανότητα ταύτισης μεταξύ των εγγραφών.

Προσωρινή Ιχνηλάτιση Εγγραφών

Οι Li πρότεινε μια τεχνική για τον προσδιορισμό των ληγμένων τιμών χαρακτηριστικών χρησιμοποιώντας ένα μοντέλο ανάπτυξης οντοτήτων στην πάροδο του χρόνου, τα οποία επιτρέπουν την χρονική ιχνηλάτιση της κάθε εγγραφής.

Ο Chiang ανέπτυξε λεπτομερή πιθανολογικά μοντέλα για καλύτερη καταγραφή της εξέλιξης των οντοτήτων και πρότεινε ταχύτερους αλγόριθμους χρονικής ιχνηλάτισης εγγραφών. Επιπλέον, έκαναν πειράματα σε σύνολα δεδομένων πραγματικού κόσμου, συμπεριλαμβανομένου του συνόλου δεδομένων DBLP (σύνολο δεδομένων βιβλιογραφίας επιστήμης υπολογιστών) για την αξιολόγηση της εργασίας τους και πέτυχαν καλά αποτελέσματα σε δύσκολες περιπτώσεις του DBLP.

Ιχνηλάτιση Εγγραφής με Προϋποθέσεις Μοναδικότητας

Guo πρότεινε μια τεχνική σύνδεσης η οποία έδειξε πολλά υποσχόμενα αποτελέσματα με την παρουσία τόσο εσφαλμένων δεδομένων όσο και πολλαπλών αναπαραστάσεων της ίδιας τιμής χαρακτηριστικού.Πρότεινε επίσης έναν συνδυασμό ιχνηλάτισης εγγραφών και συγχώνευσης δεδομένων για τον εντοπισμό ψευδών τιμών χαρακτηριστικών και τη διαφοροποίησή τους από εναλλακτικές αναπαραστάσεις της σωστής τιμής.

Σύντηξη Δεδομένων | Data Fusion

Όταν οι μη δομημένες πηγές δεδομένων ενσωματώνονται σε δομημένα λειτουργικά δεδομένα, πρέπει να υπάρχει βεβαιότητα ότι τα αποτελέσματα θα έχουν νόημα.

Η σύντηξη δεδομένων είναι ένας συνδυασμός τεχνικών που στοχεύουν στην επίλυση συγκρούσεων από μια συλλογή πηγών και στην εύρεση της αλήθειας που πρέπει να αντανακλά τον πραγματικό κόσμο. Πρόκειται για ένα νέο πεδίο που εμφανίστηκε πρόσφατα. Το κίνητρό του είναι ακριβώς η ακρίβεια των δεδομένων: ο Ιστός έχει διευκολύνει τη δημοσίευση και τη διάδοση ψευδών πληροφοριών σε πολλαπλές πηγές, οι οποίες καθιστούν τον διαχωρισμό του "σιταριού από το στάχυ" πολύ κρίσιμο ζήτημα στην παρουσίαση δεδομένων υψηλής ποιότητας.Υπάρχουν τρεις τεχνικές που χρησιμοποιούνται στο Data Fusion:

Ανίχνευση Διπλοτυπων: Εντοπίσμός διπλότυπων τιμών και μειώση του αντίστοιχου συντελεστή βάρος τους

Ψηφοφορία: Ανίχνευση της πιο κοινής τιμής για κάθε χαρακτηριστικό.

Ποιότητα Πηγής: Μετά την ψηφοφορία, δίνουμε περισσότερο βάρος σε ενημερωμένες πηγές (που παρουσιάζουν δηλαδή τον υψηλότερο αριθμό κοινών χαρακτηριστικών)

Για παράδειγμα, θεωρήστε ότι έχουμε πέντε πηγές δεδομένων S1… S5 με τα ίδια πέντε χαρακτηριστικά Att1… Att5 όπως φαίνεται στον παρακάτω πίνακα:

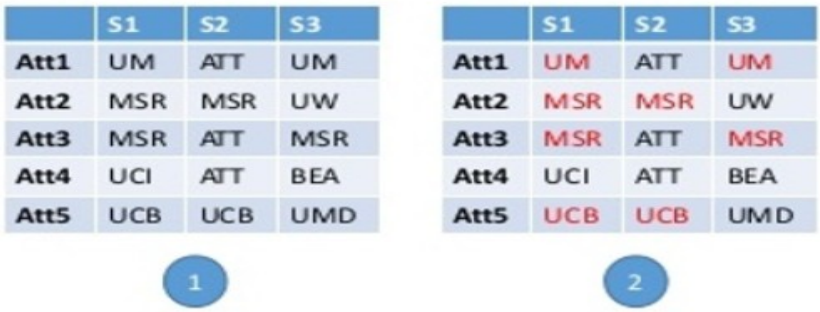
	S1	S2	S3	S4	S5
Att1	UM	ATT	UM	UM	UI
Att2	MSR	MSR	UW	UW	UW
Att3	MSR	ATT	MSR	MSR	MSR
Att4	UCI	ATT	BEA	BEA	BEA
Att5	UCB	UCB	UMD	UMD	UMD

Εικόνα 20.Παράδειγμα Καταχωρήσεων

Όπως φαίνεται στο παρακάτω σχήμα, πρώτα αναζητούμε διπλότυπες πηγές.Στην προκείμενη κοιτάζωντας τα S3, S4 και S5, διαπιστώνουμε ότι τα S3 και S4 είναι πανομοιότυπα. Επιπλέον, υπάρχει μόνο μία διαφορά με το S5. Επομένως, το βάρος των S4 και S5 είναι μειωμένο.

Εικόνα 21.Απομείωση Βάρους

Μετά από αυτό, είναι η φάση ψηφοφορίας στην οποία αναζητούμε τις πιο κοινές τιμές για κάθε χαρακτηριστικό όπως φαίνεται στην Εικόνα. 6 (οι τιμές με κόκκινο χρώμα είναι οι πιο κοινές τιμές)



Εικόνα 22.Εντοπισμός πιο κοινών

Στην τελευταία φάση -την ποιότητα πηγής- διαπιστώνουμε ότι το S1 έχει τον υψηλότερο αριθμό κοινών χαρακτηριστικών, γι 'αυτό το δίνουμε περισσότερο βάρος.

Εικόνα 23.Στάθμιση Βαρών

Τεχνικές

Ανακάλυψη Αλήθειας | Truth Discovery

Έγιναν πολλές προσπάθειες για τη μέτρηση της αξιοπιστίας των πηγών, ο Dong πρότεινε μια διαδικασία συντηξης δεδομένων που αποτελείται από 3 φάσεις: Ανακάλυψη Αλήθειας, Αξιολόγηση Αξιοπιστίας και Ανίχνευση Διπλότυπων. Μετά από αυτό, πολλοί ερευνητές πρότειναν ορισμένες επεκτάσεις για κάθε μια από αυτές τις φάσεις

Online Data Fusion

Σε πολλούς τομείς, τα διαδικτυακά δεδομένα αλλάζουν με την πάροδο του χρόνου και σε πολλές περιπτώσεις, πρέπει να αξιολογούνται σε πραγματικό χρόνο. Για την αντιμετώπιση αυτού του προβλήματος, ο Liu πρότεινε μια διαδικτυακή τεχνική σύντηξης δεδομένων στην οποία η ακρίβεια της πηγής και η συσχέτιση των διπλοτύπων αξιολογούνται σε λειτουργία εκτός σύνδεσης και κατά το χρόνο λειτουργίας στο κάθε ερώτημα αξιολογείται ο βαθμός η ανακάλυψη αλήθειας. Αυτή η τεχνική αξιολογήθηκε με πειραματισμό σε ένα σύνολο δεδομένων βιβλίων και δείχνει υψηλή απόδοση.

Δυναμική Data Fusion

Τα διαδικτυακά σύνολα δεδομένων είναι συχνά δυναμικά, σε πολλές περιπτώσεις αλλάζουν σε υψηλή ταχύτητα, που δημιουργεί προβλήματα στις τεχνικές σύντηξης δεδομένων. Για την αντιμετώπιση αυτού του προβλήματος, ο Dong πρότεινε και αξιολόγησε έναν δυναμικό αλγόριθμο σύντηξης δεδομένων.

2.9.2 Integration σε Επίπεδο Επιχειρησιακής Πλατφόρμας

[14. Deloitte]

Η Deloitte πρότεινε τα εξής πρότυπα διασύνδεσης των νέων τεχνολογιών Big Data στους οργανισμούς

Warm-Cold Data Store

-Νέες τεχνολογίες χρησιμοποιούνται για να αρχειοθετηθούν δεδομένα υψηλού όγκου και χαμηλής αξίας

-Η μορφοποιημένη πληροφορία μετακινείται από την αποθήκη δεδομένων σε νέα ζώνη αποθήκευσης.

-Τα δεδομένα γίνονται διαθέσιμα για χρήση από Αναλυτικά Sand Boxes ή μέσω εργαλείων reporting κατευθείαν από την νέα ζώνη

Enterprise DataWarehouse Augmentation

-Νέα τεχνολογία χρησιμοποείται πρωτίστως για την εκτέλεση διαδικασιών ELT.

-Μεγάλοι όγκοι αδόμητων η ημιδομημένων δεδομένων προτυποποιούνται για μεταγενέστερη κατανάλωση

-Η πληροφορία μπορεί να προσπελαστεί κατευθείαν από τις νέες δομές αποθήκευσης και να συγχωνευθεί με τα υφιστάμενα παραδοσιακά σύνολα δεδομένα χρησιμοποιώντας τεχνικές εικονικοποίησης (data virtualization)

Integrated Analytics Platform

-Σε αυτό το σενάριο οι νέες τεχνολογίες διασυνδέονται σε πολύ μεγαλύτερο βαθμό με τα παραδοσιακά εργαλεία ώστε να χρησιμοποιηθούν για deeper analytics

-Αδόμητα και δομημένα δεδομένα συνδυάζονται

-Τα αποτελεσματικά δεδομένα τότε γίνονται διαθέσιμα για περισσότερη ανάλυση και αναφορές κατευθείαν από νέα εργαλεία και προωθούνται και στα παραδοσιακά εργαλεία.

Real Time Desicion Platform

-Οι νέες τεχνολογίες χρησιμοποιούνται για την ανάλυση δεδομένων σε πραγματικό χρόνο

-Τα analytics είναι ενσωματωμένα εντός των καθημερινών διαδικασιών ώστε να υποστηρίζουν την λήψη αποφάσεων πρώτης γραμμής.

3. Αρχιτεκτονικές Συστημάτων Big Data

3.1 Γενικά

[32. Garrett Alley]

Η αρχιτεκτονική μεγάλων δεδομένων είναι το γενικό υπερσύστημα που χρησιμοποιείται για την απορρόφηση και την επεξεργασία τεράστιων ποσοτήτων δεδομένων (αυτά δηλαδή που αναφέρονται ως "Big Data") έτσι ώστε να μπορούν να αναλυθούν για επιχειρησιακούς σκοπούς. Μπορεί να θεωρηθεί το προσχέδιο για την υιοθέτηση μιας ολοκληρωμένης λύσης Big Data βασισμένης στις επιχειρηματικές ανάγκες ενός οργανισμού. Η αρχιτεκτονικές αυτές συνήθως έχουν σχεδιαστεί για να χειρίζονται τους ακόλουθους τύπους εργασιών:

- Μαζική Επεξεργασία Πηγών Big Data.
- Επεξεργασία Big Data σε πραγματικό χρόνο.
- Προγνωστική Ανάλυση(Predictive Analytics) και Μηχανική Μάθηση.

Μια καλά σχεδιασμένη αρχιτεκτονική συνήθως εξοικονομεί χρήματα στην επιχείρηση και να βοηθάει ουσιαστικά στην πρόβλεψη μελλοντικών τάσεων, ώστε να είναι εφικτή η λήψη καλών επιχειρηματικών αποφάσεων

Οι αρχιτεκτονικες συνήθως ποικίλλουν σημαντικά ανάλογα με τις υποδομές και τις ανάγκες της κάθε επιχείρησης, αλλά σε γενικές γραμμές συνήθως περιέχουν τα ακόλουθα στοιχεία:

Πηγές Δεδομένων. Όλη η αρχιτεκτονική μεγάλων δεδομένων ξεκινά από τις πηγές. Αυτες μπορεί να περιλαμβάνουν δεδομένα από βάσεις δεδομένων, δεδομένα από πηγές σε πραγματικό χρόνο (όπως συσκευές IoT) και στατικά αρχεία που δημιουργούνται από εφαρμογές, όπως πχ αρχεία καταγραφής των Windows.

Κατάνάλωση Μηνυμάτων σε Πραγματικό Χρόνο. Εάν υπάρχουν πηγές σε πραγματικό χρόνο, θα χρειαστεί να δημιουργηθεί ένας αντίστοιχος μηχανισμός στην αρχιτεκτονική για να απορροφηθούν αυτά τα δεδομένα.

Αποθήκευση Δεδομένων. Προφανώς θα χρειαστεί χώρο αποθήκευσης για τα δεδομένα που θα υποβληθούν σε επεξεργασία μέσω αρχιτεκτονικής. Συχνά, τα δεδομένα θα αποθηκεύονται σε μια **Data Lake**(λίμνη δεδομένων), η οποία είναι μια μεγάλη μη δομημένη βάση δεδομένων που κλιμακώνεται εύκολα.

Συνδυασμός Επεξεργασίας batch και Πραγματικού Χρόνου. Συνήθως χρειάζεται ο χειρισμός δεδομένων τόσο σε πραγματικό χρόνο όσο και στατικά δεδομένα, επομένως ένας συνδυασμός των δύο τύπων επεξεργασίας να πρέπει να ενσωματωθεί στη αρχιτεκτονική.

Αποθήκευση Αναλυτικών Δεδομένων. Αφού προετοιμάστούν τα δεδομένα για ανάλυση, πρέπει να συγκεντρώθούν σε ένα μέρος, ώστε να μπορεί να εκτελεστεί ανάλυση σε ολόκληρο το σύνολο δεδομένων. Η σημασία αυτής της προσέγγισης είναι στο ότι όλα τα δεδομένα βρίσκονται σε ένα μέρος ώστε η ανάλυσή να είναι ολοκληρωμένη και βελτιστοποιημένη για συγκεντρωτικές αναφορές και χαμηλού επιπέδου συναλλακτική πληροφορία. Αυτό μπορεί να έχει τη μορφή αποθήκης δεδομένων βασισμένης σε τεχνολογία Cloud ή σχεσιακής βάσης δεδομένων, ανάλογα με τις ανάγκες.

Εργαλεία Ανάλυσης και Reporting. Μετά την συλλογή και την επεξεργασία διαφόρων πηγών δεδομένων, θα πρέπει να συμπεριληφθεί ένα εργαλείο για την ανάλυση των δεδομένων. Συχνά, θα χρησιμοποιηθεί ένα εργαλείο BI (Business Intelligence) για αυτήν την εργασία και ενδέχεται να απαιτείται από έναν επιστήμονα δεδομένων(data scientist) να εξερευνήσει τα δεδομένα.

Αυτοματοποίηση. Η μετακίνηση των δεδομένων μέσω των διαφόρων συστημάτων απαιτεί ωποσδήποτε ενορχήστρωση και κάποια μορφή αυτοματοποίησης. Η απορρόφηση και ο μετασχηματισμός των δεδομένων, η μετακίνηση σε στρώματα batch και stream ροες, η φόρτωσή τους σε μια αναλυτική βάση αποθήκευσης και, τέλος, η δημιουργία Insights πρέπει να εντασσεται σε μια επαναλαμβανόμενη ροή εργασίας, ώστε να είναι εφικτή συνεχώς η λήψη Insights από τα Big Data.

Οφέλη της Αρχιτεκτονικής Big Data

Ο όγκος των δεδομένων που είναι διαθέσιμα για ανάλυση αυξάνεται καθημερινά. Υπάρχουν περισσότερες πηγές συνεχούς ροής από ποτέ, συμπεριλαμβανομένων των διαθέσιμων δεδομένων από αισθητήρες κίνησης, αισθητήρες υγείας, αρχεία καταγραφής συναλλαγών και αρχεία καταγραφής δραστηριοτήτων. Αλλά η κατοχή των δεδομένων είναι μόνο το ένα ζήτημα. Πρέπει επίσης να είστε σε θέση να κατανοήσετε τα δεδομένα και να τα χρησιμοποιήσετε εγκαίρως για να επηρεάσετε κρίσιμες αποφάσεις. Η χρήση μιας αρχιτεκτονικής Big Data μπορεί να βοηθήσει την επιχείρισή σας να εξοικονομήσει χρήματα και να λάβει κρίσιμες αποφάσεις, όπως:

- **Μείωση Κόστου**ς. Τεχνολογίες Big Data όπως το Hadoop και τα Analytics που βασίζονται σε Cloud μπορούν να μειώσουν σημαντικά το κόστος όσον αφορά την αποθήκευση μεγάλων ποσοτήτων δεδομένων.
- **Λήψη ταχύτερων και καλύτερων αποφάσεων**. Χρησιμοποιώντας το στοιχείο ροής πραγματικού χρόνου της αρχιτεκτονικής δεδομένων, μπορείτε να λάβετε αποφάσεις σε πραγματικό χρόνο.
- **Πρόβλεψη μελλοντικών αναγκών και δημιουργία νέων προϊόντων.** Τα μεγάλα δεδομένα μπορούν να σας βοηθήσουν να μετρήσετε τις ανάγκες των πελατών και να προβλέψετε μελλοντικές τάσεις χρησιμοποιώντας τα αναλυτικά στοιχεία.

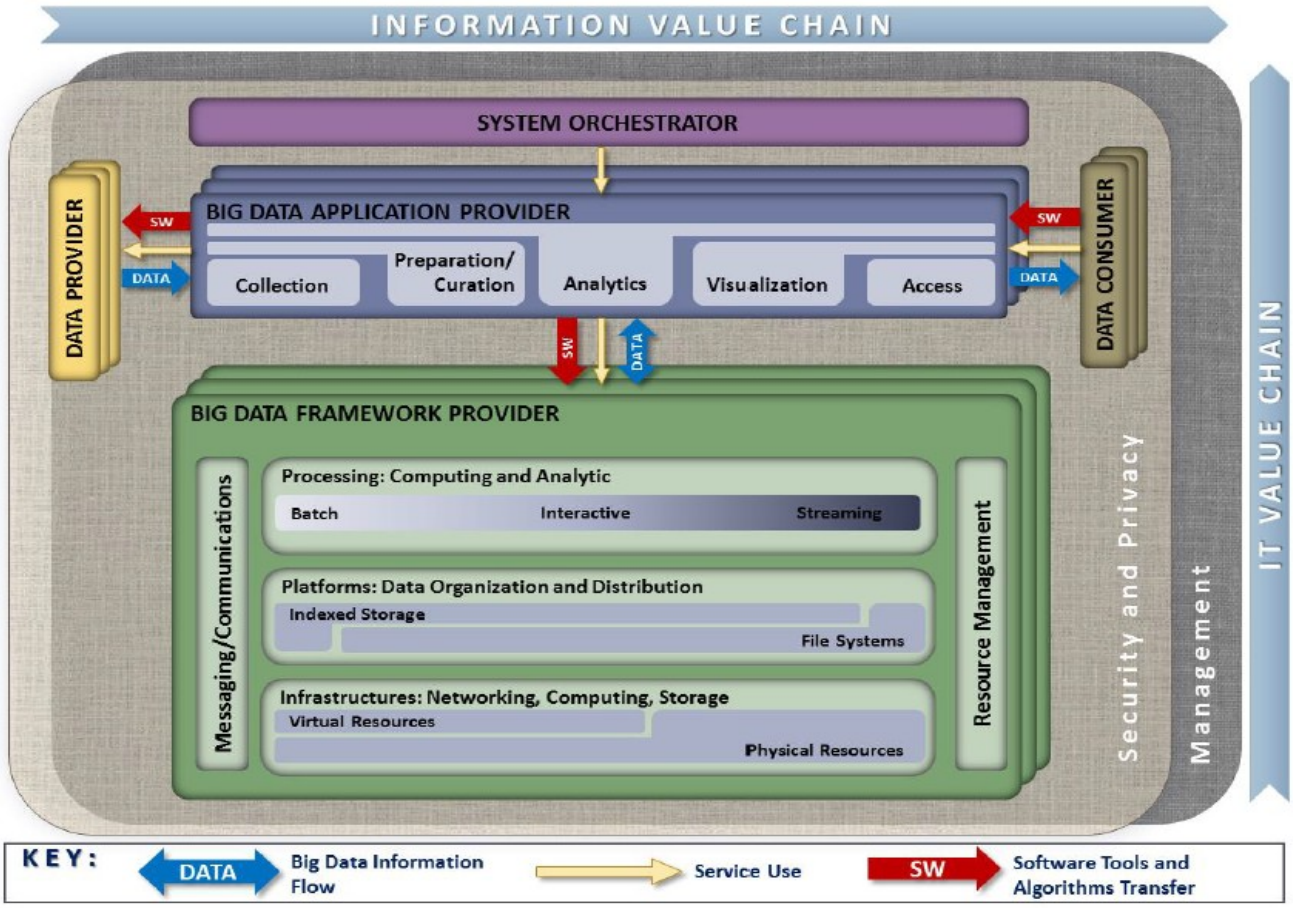
3.2 NIST Big Data Reference Architecture - NBDRA

by NIST [7]

Αποτελεί ένα εννοιολογικό αρχιτεκτονικής Big Data ουδέτερο από προμηθευτή, τεχνολογία και υποδομή. Αυτό το εννοιολογικό μοντέλο, το **NBDRA-NIST Big Data Reference Architecture**, φαίνεται στο παρακάτω σχήμα και αντιπροσωπεύει ένα σύστημα Big Data που αποτελείται από πέντε λογικά λειτουργικά στοιχεία που συνδέονται με διεπαφές διαλειτουργικότητας (services). Δύο υπόβαθρα περιβάλλουν τα συστατικά αυτά, το πρώτο είναι η διαχείριση και το δεύτερο η ασφάλεια και η ιδιωτικότητα, με τρόπο που αντιπροσωπεύει την αλληλένδετη φύση τους με τα πέντε στοιχεία.

Το NBDRA έχει ως στόχο να επιτρέψει στους μηχανικούς συστημάτων, τους επιστήμονες δεδομένων, τους προγραμματιστές λογισμικού, τους αρχιτέκτονες δεδομένων και τους ανώτερους υπεύθυνους λήψης αποφάσεων να αναπτύξουν λύσεις σε ζητήματα που απαιτούν την σύγκλιση διαφορετικές προσεγγίσεων σε ένα ενιαίο και διαλειτουργικό οικοσύστημα Big Data. Περιγράφει ένα πλαίσιο για την υποστήριξη μιας ποικιλίας επιχειρηματικών περιβαλλόντων, συμπεριλαμβανομένων των αυστηρά ορισμένων και ιδιόκτητων συστημάτων πληροφοριακών συστημάτων επιχειρήσεων και των χαλαρά συνδεδεμένων κάθετων τομέων οικονομίας/βιομηχανιών, ενισχύοντας την κατανόηση του τρόπου με τον οποίο τα Big Data συμπληρώνουν αλλά και διαφέρουν από τα υπάρχοντα συστήματα analytics, business intelligence και βάσεων δεδομένων.

Η αρχιτεκτονική που προτείνεται λοιπόν,οργανώνεται γύρω από πέντε κύριους ρόλους συμμετεχόντων (stakeholders(και πολλαπλούς δευτερεύοντες ρόλους ευθυγραμμισμένους κατά μήκος δύο αξόνων που αντιπροσωπεύουν τις δύο αλυσίδες αλυσίδες Big Data: Information Value (οριζόντιος άξονας) και IT Value (κάθετος άξονας). Κατά μήκος του άξονα Information Value , η αξία δημιουργείται από τη συλλογή δεδομένων, την διασύνδεση, την ανάλυση και την εφαρμογή των αποτελεσμάτων. Κατά μήκος του άξονα IT, η αξία δημιουργείται παρέχοντας δικτύωση, υποδομή, πλατφόρμες, εργαλεία εφαρμογών και άλλες υπηρεσίες πληροφορικής για τη φιλοξενία και τη λειτουργία των Big Data στην υποστήριξη απαιτούμενων εφαρμογών δεδομένων. Στη διασταύρωση και των δύο αξόνων βρίσκεται ο ρόλος του Big Data Application Provider, που δείχνει ότι η ανάλυση δεδομένων και η εφαρμογή της προσδίδουν προστιθέμενη αξία στους ενδιαφερόμενους των Big Data και στις δύο αλυσίδες τιμών. Ο όρος “πάροχος” ως μέρος του προσδιορισμού τόσο του Big Data Application Provider και του Big Data Framework Provider υπάρχει για να δείξει ότι αυτοί οι ρόλοι παρέχουν ή εφαρμόζουν συγκεκριμένες δραστηριότητες και λειτουργίες μέσα στο σύστημα. Δεν ορίζει ένα συγκεκριμένο μοντέλο υπηρεσίας ή επιχειρηματικής οντότητας.



Εικόνα 24.NIST Big Data Reference Architecture

Παρακάτω αναλύονται οι ρόλοι που προβλέπει το πλαίσιο

ΠΑΡΟΧΟΣ ΔΕΔΟΜΕΝΩΝ

Ο ρόλος του Παρόχου Δεδομένων είναι εισάγει νέα δεδομένα ή πηγές πληροφοριών στο σύστημα Big Data για σκοπους ανακάλυψης, πρόσβασης και μετασχηματισμού από το σύστημα. Οι νέες ροές δεδομένων είναι διακριτές από τα δεδομένα που υπάρχουν ήδη προς χρήση από το σύστημα και διατηρούνται στις διάφορες αποθήκες. Παρόμοιες τεχνολογίες μπορούν να χρησιμοποιηθούν για πρόσβαση τόσο στις νέες ροές δεδομένων όσο και στα υπάρχοντα δεδομένα. Οι φορείς παροχής δεδομένων μπορεί να είναι οτιδήποτε, από έναν αισθητήρα, μέχρι έναν άνθρωπο που εισάγει δεδομένα χειροκίνητα.

Οι δραστηριότητες του παροχου δεδομένων περιλαμβάνουν τα ακόλουθα, τα οποία είναι κοινά στα περισσότερα συστήματα :

- Συλλογή πρωτογενών δεδομένων
- Η διατήρηση των δεδομένων όπου απαιτείται
- Παροχή λειτουργιών μετασχηματισμού για την επεξεργασία δεδομένων ευαίσθητων πληροφοριών, όπως προσωπικές πληρφορίες αναγνώρισης(PII)
- Δημιουργία μεταδεδομένων που περιγράφουν την πηγή δεδομένων, τις πολιτικές χρήσης / τα δικαιώματα πρόσβασης και άλλα τα σχετικά χαρακτηριστικά
- Εφαρμογή δικαιωμάτων πρόσβασης σε δεδομένα.
- Καθιέρωση επίσημων ή ανεπίσημων συμβάσεων για τις άδειες πρόσβασης σε δεδομένα.
- Την πρόσβαση στα δεδομένα μέσω κατάλληλων προγραμματιζόμενων διεπαφών
- Παροχή μηχανισμών πρόσβασης
- Δημοσίευση της διαθεσιμότητας των πληροφοριών και των μέσων πρόσβασης σε αυτά.

SYSTEM ORCHESTRATOR

Οι δραστηριότητες στο πλαίσιο του ρόλου του System Orchestrator καθορίζουν τη συνολική ιδιοκτησία, τη διακυβέρνηση και την πολιτική λειτουργίας για το σύστημα Big Data ορίζοντας τις αντίστοιχες απαιτήσεις. Οι δραστηριότητες αυτές πραγματοποιούνται κυρίως κατά τη φάση ορισμού/σχεδιασμού του συστήματος, αλλά πρέπει να επανεξετάζονται περιοδικά καθ 'όλη τη διάρκεια του κύκλου ζωής του συστήματος. Η άλλη πρωταρχική πτυχή των δραστηριοτήτων αυτού του ρόλου είναι η παρακολούθηση της συμμόρφωσης με τις σχετικές απαιτήσεις.

Ορισμένες κατηγορίες δραστηριοτήτων που θα μπορούσαν να οριστούν για αυτόν τον ρόλο στην αρχιτεκτονική περιλαμβάνουν τον καθορισμό απαιτήσεων και την παρακολούθηση της συμμόρφωσης για θέματα όπως:

• **Business Ownership:** Αυτή η τάξη δραστηριότητας ορίζει ποιοι ενδιαφερόμενοι κατέχουν και έχουν ευθύνη για τα διάφορα μέρη του Big Data System. Ορίζει την ιδιοκτησία και την ευθύνη για τις δραστηριότητες και τις λειτουργικές συνιστώσες του υπόλοιπου συστήματος και πώς αυτη η ιδιοκτησία θα παρακολουθείται.

• **Διακυβέρνηση:** Καθορίζει τις πολιτικές και τη διαδικασία διακυβέρνησης του συστήματος συνολικά. Αυτές οι απαιτήσεις διακυβέρνησης θα εκτελούνται και θα παρακολουθούνται με τη σειρά τους από φορείς που ορίζονται ως ιδιοκτήτες για τα αντίστοιχα τμήματα του συστήματος.

• **Αρχιτεκτονική Συστήματος:** Περιλαμβάνει τον καθορισμό των γενικών απαιτήσεων που πρέπει να καλύπτεται από την αρχιτεκτονική του συστήματος. Σε γενικές γραμμές, οι δραστηριότητες αυτής της κατηγορίας καθορίζουν τις τεχνικές κατευθυντήριες γραμμές που πρέπει να πληρεί το συνολικό σύστημα και στη συνέχεια να παράσχει τις αντίστοιχες πολιτικές παρακολούθησης της συνολική αρχιτεκτονικής για να βεβαιωθεί ότι εξακολουθεί να συμμορφώνεται με τις απαιτήσεις.

• **Επιστήμη Δεδομένων:** Καθορίζουν πολλές από τις απαιτήσεις που πρέπει να πληρούνται από μεμονωμένους αλγορίθμους ή εφαρμογές εντός του συστήματος. Αυτά θα μπορούσαν να περιλαμβάνουν για παράδειγμα την ακρίβεια του υπολογισμών ή την ακρίβεια/ανάκληση των αλγορίθμων εξόρυξης δεδομένων.

• **Ασφάλεια / Προστασία Προσωπικών Δεδομένων:** Παρόλο που καμία κατηγορία δραστηριοτήτων δεν θεωρείται υποχρεωτική, είναι ασφαλώς πολύ κρίσιμη καθώς οποιαδήποτε αρχιτεκτονική χωρίς σαφώς καθορισμένες απαιτήσεις ασφάλειας και προστασίας της ιδιωτικής ζωής και η σχετική παρακολούθηση της είναι εξαιρετικά επικίνδυνη. Η ασφάλεια ασχολείται με τον έλεγχο πρόσβασης στο σύστημα και τα δεδομένα του και απαιτείται να διασφαλίζει την ιδιωτικότητα των προσωπικών ή εταιρικών πληροφοριών. Το απόρρητο αφορά τόσο την εξασφάλιση προσωπικών πληροφοριών, όσο και τον καθορισμό των πολιτικών και των ελέγχων με την οποία οι πρωτογενείς πληροφορίες ή οι παράγωγες πληροφορίες μπορούν να μοιραστούν ή όχι.

Άλλες κατηγορίες δραστηριοτήτων που μπορούν να αντιμετωπιστούν περιλαμβάνουν τα εξής:

• **Διαχείρισης της Ποιότητας,**

• **Διαχείριση Υπηρεσιών**, και

• **Απαιτήσεις Ελέγχου**.

BIG DATA APPLICATION PROVIDER

Ο ρόλος του Big Data Application Provider Φορέα Παροχής Εφαρμογών Big Data εκτελεί ένα συγκεκριμένο σύνολο λειτουργιών καθ'όλο το κύκλου ζωής των δεδομένων ώστε να ανταποκρίνεται στις απαιτήσεις που έχει θέσει ο System Orchestrator, καθώς και να καλύπτει την ασφάλεια και τις απαιτήσεις του απόρρητο. Είναι το στοιχείο αρχιτεκτονικής που ενσωματώνει την επιχειρηματική λογική και λειτουργικότητα που θα εκτελεστεί από την αρχιτεκτονική. Οι δραστηριότητες του περιλαμβάνουν τα ακόλουθα:

<u>Συλλογή</u>

Σε γενικές γραμμές, η δραστηριότητα συλλογής του Παρόχου Εφαρμογών χειρίζεται τη διασυνδεση με τον Πάροχο Δεδομένων. Αυτή μπορεί να είναι μια γενική υπηρεσία, όπως ένας διακομιστής αρχείων ή ένας διακομιστής ιστού που έχει οριστεί από τον Systems Orchestrator να αποδέχεται ή να εκτελεί συγκεκριμένες συλλογές δεδομένων ή ακόμα μπορεί να είναι μια εξειδικευμένη υπηρεσία στην χρήση συγκεκριμένης εφαρμογής που έχει σχεδιαστεί για να τραβάει δεδομένα ή να λαμβάνει ωθήσεις δεδομένων από τον πάροχο δεδομένων. Δεδομένου ότι κατ'ελάχιστον λαμβάνει δεδομένα, πρέπει να αποθηκεύει(μόνιμα η προσωρινά) τα ληφθέντα δεδομένα μέχρις ότου προωθηθούν για διατήρηση στον Big Data Framework Provider. Αυτή η διατήρηση δεν χρειάζεται να είναι σε φυσικά μέσα, αλλά μπορεί απλά να είναι σε μια ουρα αναμονής ή μνήμης η οποιαδήποτε άλλη υπηρεσία παρέχεται από τα πλαίσια επεξεργασίας του Big Data Application Provider. Η συλλογή συνήθως στοχεύει σε αυτά που πρόκειται να εισαχθούν σε κύκλους Extract-Transform-Load-ETL/Extract Load Transform

ELT. Στο αρχικό στάδιο, σύνολα δεδομένων (π.χ. αρχεία δεδομένων) από παρόμοιες δομές συλλέγονται (και συνδυάζονται), οδηγώντας σε ομοιόμορφο επίπεδο ασφάλειας, πολιτικών και θεωρήσεων. Δημιουργούνται τα αρχικά μεταδεδομένα (π.χ. αναγνωρίζονται οι οντότητες και τα κλειδιά τους) ώστε να διευκολυνθούν τα επόμενα στάδια aggregations ή look-up.

Προετοιμασία

Η δραστηριότητα της προετοιμασίας είναι όπου πιθανώς εκτελούνται οι κύκλοι Extract-Transform-Load-ETL /Extract Load Transform ELT αν και η δραστηριότητα της ανάλυσης πιθανόν να έχει ήδη εκτελεστεί και σε προηγμένα τμήματα του μετασχηματισμού. Οι εργασίες που εκτελούνται από αυτή τη δραστηριότητα θα μπορούσαν να περιλαμβάνουν επικύρωση δεδομένων (π.χ., αθροίσματα ελέγχου / hashes, έλεγχοι μορφοποιήσεων), καθαρισμός (π.χ., εξαλείψη προβληματικών εγγραφών / πεδίων), απομάκρυνση ακραίων τιμών, τυποποίηση, επαναδιαμόρφωση ή ενσωμάτωση. Είναι επίσης η δραστηριότητα όπου τα δεδομένα των πηγών συνήθως διατηρούνται πριν την οριστική αρχειοθέτηση τους στον Framework Provider όπου και θα επαλήθευτούν. Η προσάρτηση μπορεί να περιλαμβάνει βελτιστοποίηση των δεδομένων μέσω ειδικών χειρισμών (π.χ. deduplication) και ευρετηρίαση για την βελτιστοποίηση τη διαδικασία ανάλυσης. Αυτή η δραστηριότητα μπορεί επίσης να συγκεντρώνει δεδομένα από διαφορετικούς παροχείς δεδομένων, αξιοποιώντας τα κλειδιά μεταδεδομένων ώστε να δημιουργηθεί ένα εκτεταμένο και ευρύτερο σύνολο δεδομένων.

Analytics

Η δραστηριότητα της ανάλυσης περιλαμβάνει την χαμηλού επιπέδου κωδικοποίηση της επιχειρηματική λογική στο συστήμα Big Data (ακολουθώντας την λογική επιχειρηματικών διαδικασιών υψηλότερου επιπέδου που έχει κωδικοποιηθεί από τον System Orchestrator). Η δραστηριότητα υλοποιεί τεχνικές εξαγωγής γνώσεων από τα δεδομένα σχετικά με τις απαιτήσεις της κάθε εφαρμογής. Οι απαιτήσεις καθορίζουν τους αλγόριθμους επεξεργασίας δεδομένων για την καθαυτού επεξεργασία των δεδομένων αλλά και για την παραγωγή νέων στοιχείων που θα αντιμετωπίσουν τον όποιο τεχνικό στόχο. Η δραστηριότητα της ανάλυσης επίσης θα αξιοποιήσει τα επεξεργαστικά πλαίσια για την εφαρμογή της σχετικής λογικής. Αυτό συνήθως περιλαμβάνει την παροχή προς εκτέλεση κατάλληλου λογισμικού που υλοποιεί την αναλυτική λογική στα στοιχεία batch ή/και ροής του πλαίσιου επεξεργασίας. Το πλαίσιο ανταλλαγής μηνυμάτων / επικοινωνίας του Big Data Framework Provider μπορεί να χρησιμοποιηθεί για να μεταβιβάσει δεδομένα ή λειτουργίες ελέγχου στη λογική εφαρμογής που εκτελείται. Η αναλυτική λογική μπορεί επίσης να διαιρεθεί σε πολλαπλές ενότητες που θα εκτελεστούν μεν αυτόνομα από το επεξεργαστικό πλαίσιο αλλα θα επικοινωνούν μεταξύ τους μεταδίδοντας μηνύματα μέσω του πλαισίου ανταλλαγής μηνυμάτων/επικοινωνίας αλλά και με άλλες λειτουργίες που δημιουργούνται από τον Big Data Application Provider

Οπτικοποίηση

Προετοιμάζει στοιχεία των επεξεργασμένων δεδομένων και την έξοδο της αναλυτικής δραστηριότητας για παρουσίαση στους καταναλωτές δεδομένων. Ο στόχος αυτής της δραστηριότητας είναι να μορφοποιηθούν και να παρουσιάστούν τα δεδομένα με τέτοιο τρόπο ώστε να μεταδίδεται με τον καλύτερο τρόπο το νόημα τους και η γνώση που φέρουν.Η προετοιμασία οπτικοποίησης μπορεί να περιλαμβάνει την παραγωγή μιας έκθεσης με βάση κείμενο ή την απόδοση των αναλυτικών αποτελεσμάτων σε κάποια μορφή γραφημάτων. Η προκύπτουσα έξοδος μπορεί επίσης να είναι μια στατική απεικόνιση που να αποθηκευτεί μέσω του Big Data Framework Provider για μελλοντική πρόσβαση. Ωστόσο, η δραστηριότητα απεικόνισης συχνά εξαρτάται από τη δραστηριότητα πρόσβασης, τη δραστηριότητα ανάλυσης και τον πάροχο μεγάλων δεδομένων (πλαίσιο επεξεργασίας) για την παροχή διαδραστικής απεικόνισης των δεδομένων προς τον Καταναλωτή Δεδομένων βάσει παραμέτρων που ίδιος ο καταναλωτής παρέχει έχωντας αντίστοιχες δυνατότητες πρόσβασης. Η δραστηριότητα απεικόνισης μπορεί να είναι εντελώς εφαρμόσιμη σε συγκεκριμένη ανάγκη,να χρησιμοποιεί μιας ή περισσότερες βιβλιοθήκες αλλων εφαρμογών ή μπορεί να χρησιμοποιεί πλαισια εξειδικευμένα στην επεξεργασία απεικόνισεων εντός του Big Data Framework Provider

Πρόσβαση

Επικεντρώνεται στην επικοινωνία / αλληλεπίδραση με τον Καταναλωτή Δεδομένων. Παρόμοια με τη δραστηριότητα συλλογής, η δραστηριότητα αυτη μπορεί να είναι μια γενική υπηρεσία όπως ένας διακομιστής ιστού ή διακομιστής εφαρμογών που έχει ρυθμιστεί από τον System Orchestrator για να χειρίζεται συγκεκριμένα αιτήματα του καταναλωτή δεδομένων. Αυτή η δραστηριότητα επίσης διασυνδέεται με την απεικόνιση και την ανάλυση ώστε να απαντήσει στα αιτήματα του Καταναλωτή Δεδομένων (ο οποίος μπορεί να είναι άτομο) και να

χρησιμοποιεί τις πλατφόρμες επεξεργασίας για την ανάκτηση δεδομένων που ανταποκρίνονται στα αιτήματα των Καταναλωτών. Επιπλέον, επιβεβαιώνει ότι καταγράφονται τα περιγραφικά και διοικητικά μεταδεδομένα ή ακόμα και τα μεταδεδομένα που σχετίζονται με την πρόσβαση η το είδος των δεδομένων που παραδόθηκαν στον καταναλωτή.

BIG DATA FRAMEWORK PROVIDER

Ο πάροχος πλαισίου αποτελείται συνήθως από μία ή περισσότερες ιεραρχικά οργανωμένες διαστάσεις των στοιχείων της αλυσίδας τεχνολογικής αξίας της προτεινόμενης αρχιτεκτονικής NBDRA . Δεν υπάρχει προϋπόθεση ότι όλες οι περιπτώσεις ενός επιπέδου της ιεραρχίας πρέπει να είναι της ίδιας τεχνολογίας. Στην πραγματικότητα, οι περισσότερες υλοποιήσεις Big Data είναι υβρίδια που συνδυάζουν πολλαπλές προσεγγίσεις τεχνολογίας προκειμένου να τους παρέχουν ευελιξία ή να καλύπτουν πλήρως μια σειρά απαιτήσεων, οι οποίες καθοδηγούνται από τον Big Data Application Provider

Πολλές από τις πρόσφατες εξελίξεις που σχετίζονται με το Big Data συνέβησαν στον τομέα των πλαισίων που έχουν σχεδιαστεί για την εξυπηρέτηση αναγκών κλιμάκωσης (π.χ. αντιμετώπιση όγκου, ποικιλίας, ταχύτητας και μεταβλητότητας) διατηρώντας όμως ταυτόχρονα γραμμική ή σχεδόν γραμμική απόδοση. Αυτές οι πρόοδοι έχουν προσελκύσει μεγάλο μέρος του τεχνολογικού ενθουσιασμού. Κατά συνέπεια, υπάρχουν πολύ περισσότερες πληροφορίες σχετικά με τον τομέα των πλαισίων σε σύγκριση με τα άλλα στοιχεία.

Ο Big Data Framework Provider επομένως περιλαμβάνει τους ακόλουθους τρεις υπο-ρόλους (από κάτω προς τα πάνω)

• Πλαίσια Υποδομής (βλέπε κεφάλ.2.2)

• Πλαίσια Πλατφόρμας Δεδομένων(βλέπε κεφάλ.2.3), και

• Πλαισία Επεξεργασίας/Υπολογιστικής (βλέπε κεφάλ.2.3)

ΚΑΤΑΝΑΛΩΤΗΣ ΔΕΔΟΜΕΝΩΝ

Παρόμοια με τον πάροχο δεδομένων, ο ρόλος του καταναλωτή δεδομένων στο πλαίσιο του NBDRA μπορεί να είναι ένας πραγματικός τελικός χρήστης ή ένα άλλο σύστημα. Ο ρόλος των καταναλωτών περιλαμβάνει τα εξής:

- Αναζήτηση και Ανάκτηση,
- Τοπική Αποθήκευση (Download)
- Τοπική Ανάλυση
- Reporting
- Οπτικοποίηση
- Δεδομένα που πρέπει να χρησιμοποιηθούν για τις δικές τους διαδικασίες.

Ο Data Consumer χρησιμοποιεί τις διασυνδέσεις ή τις υπηρεσίες που παρέχει ο Big Data Application Provider για να πάρει πρόσβαση στις πληροφορίες που ενδιαφέρουν. Αυτές οι διεπαφές μπορούν να περιλαμβάνουν αναφορές επί των δεδομένων, την ανάκτηση δεδομένων και την αναδρομολόγηση δεδομένων.

Αυτός ο ρόλος αλληλεπιδρά γενικά με τον Big Data Application Provider μέσω της λειτουργίας πρόσβασης ώστε να εκτελεστούν αναλύσεις και τις απεικονίσεις που υλοποιεί ο πάροχος εφαρμογής. Αυτή η αλληλεπίδραση μπορεί να βασίζεται στη ζήτηση, όπου ο καταναλωτής δεδομένων εκκινεί την εντολή / συναλλαγή και ο πάροχος εφαρμογής απαντά. Η αλληλεπίδραση επίσης θα μπορούσε να περιλαμβάνει αλληλεπιδραστικές οπτικοποιήσεις, δημιουργία αναφορών ή ανάλυση δεδομένων μέσω στοιχείων επιχειρηματικής ευφυΐας που παρέχονται από τον πάροχο εφαρμογών. Εναλλακτικά, η αλληλεπίδραση μπορεί να βασίζεται σε διαδικασία ροής ή push, όπου ο καταναλωτής δεδομένων απλώς εγγράφεται ή ακούει μία ή περισσότερες αυτοματοποιημένες εξόδους της εφαρμογής. Σε όλες σχεδόν τις περιπτώσεις, το στρώμα Ασφάλειας και Προστασία Προσωπικών Δεδομένων της αρχιτεκτονικής Big Data θα υποστηρίξει την επαλήθευση ταυτότητας και την εξουσιοδότηση μεταξύ του Καταναλωτή Δεδομένων και της αρχιτεκτονικής, με τις δύο πλευρές να είναι το ίδιο ικανές να εκτελέσουν το ρόλο του ελεγκτή ταυτότητας/ εξουσιοδότησης και η άλλη πλευρά να παράσχει τα διαπιστευτήρια. Σαν τη διασύνδεση μεταξύ της αρχιτεκτονικής Big Data και του Data Provider, της διεπαφής μεταξύ του Καταναλωτή Δεδομένων και του Παρόχου Εφαρμογής θα περάσουν επίσης από τις τρεις ξεχωριστές φάσεις του την έναρξη, τη μεταφορά δεδομένων και του τερματισμού.

ΥΠΗΡΕΣΙΕΣ ΔΙΑΧΕΙΡΙΣΗΣ ΤΗΣ NBDRA

Τα χαρακτηριστικά μεγάλου όγκου, ταχύτητας, ποικιλίας και μεταβλητότητας απαιτούν μια ευέλικτη διαχείριση της πλατφόρμας αποθήκευσης, επεξεργασίας και διαχείρισης πολύπλοκων δεδομένων. Η διαχείριση των συστημάτων Big Data πρέπει να χειρίζονται τόσο πτυχές του συστήματος Big Data όσο και δεδομένα που αφορούν το σύστημα. Περιλαμβάνει δύο γενικές ομάδες δραστηριοτήτων: Την **Διαχείριση Συστήματος**(System Management) και την **Διαχείριση του Κύκλου Ζωής** BDLM Big Data Lifcycle Management)

Το System Management περιλαμβάνει δραστηριότητες όπως πρόβλεψη, παραμετροποίηση, διαχείριση πακέτων, διαχείριση λογισμικού, διαχείριση αντιγράφων ασφαλείας, διαχείριση δυνατοτήτων, διαχείριση πόρων και διαχείριση απόδοσης. Το BDLM περιλαμβάνει δραστηριότητες που περιβάλλουν τον κύκλο ζωής της συλλογής, προετοιμασίας /επεξεργασίας, ανάλυσης, οπτικοποίησης και πρόσβασης.

Όπως αναφέρθηκε παραπάνω, η NBDRA αντιπροσωπεύει ένα ευρύ φάσμα συστημάτων Big Data από αυστηρά αντιστοιχισμένες επιχειρησιακές λύσεις που διασυνδέονται μέσω τυποποιημένων ή ιδιόκτητων διεπαφών σε χαλαρά συζευγμένα κάθετα συστήματα που συντηρούνται από διάφορους εμπλεκόμενους φορείς ή αρχές που δεσμεύονται από συμφωνίες, τυποποιήσεις ή de facto διασυνδέσεις. Ως εκ τούτου, μπορούν να ισχύσουν ταυτόχρονα διαφορετικές εκτιμήσεις και τεχνικές λύσεις για κάθε διαφορετική περιπτώση.

ΣΤΟΙΧΕΙΑ ΑΣΦΑΛΕΙΑΣ ΚΑΙ ΠΡΟΣΤΑΣΙΑΣ ΠΡΟΣΩΠΙΚΟΥ ΑΣΦΑΛΕΙΑΣ ΤΗΣ NBDRA

Τα στοιχεία ασφάλειας και προστασίας της ιδιωτικής ζωής συνιστούν μια βασική πτυχή της NBDRA. Έτσι το Υλικό Ασφάλειας και Προστασίας Προσωπικών Δεδομένων περιβάλλει τα πέντε κύρια εξαρτήματα, υποδεικνύοντας ότι όλα τα εξαρτήματα της αρχιτεκτονικής επηρεάζονται από λόγους ασφάλειας και προστασίας της ιδιωτικής ζωής. Ο Πάροχος Δεδομένων και ο Καταναλωτής Δεδομένων σχεδόν πάντα περιλαμβάνονται στα Στοιχεία Ασφάλειας και Προστασίας Προσωπικών Δεδομένων, και μάλιστα μπορεί συχνά να συμφωνούν ονομαστικά σε πρωτόκολλα και μηχανισμούς ασφαλείας. Οι κύριες κατηγορίες δραστηριοτήτων που σχετίζονται με αυτό το στρώμα είναι:

Ταυτοποίηση: Αυτή η κατηγορία δραστηριοτήτων περιλαμβάνει την επικύρωση ότι ο χρήστης ή η διαδικασία είναι όντως αυτό ισχυριζονται πως ειναι. Οι συγκεκριμένες δραστηριότητες επαλήθευσης ταυτότητας μπορούν να καθορίσουν τον τύπο της επαλήθευσης ταυτότητας, όπως π.χ. δύο-παράγοντων ή με ιδιωτικό κλειδί.

Εξουσιοδότηση: Διασφαλίζει ότι ο χρήστης ή η διαδικασία έχει τα δικαιώματα πρόσβασης σε πόρους ή υπηρεσίες. Τα στοιχεία ελέγχου πρόσβασης μπορούν να ορίζουν τα συγκεκριμένα δικαιώματα πρόσβασης (π.χ. δημιουργίας, ενημέρωσης, διαγραφής) για τα δεδομένα ή τις υπηρεσίες. Οι δραστηριότητες εξουσιοδότησης μπορεί να καθορίζονται με ευρείας βάσης έλεγχους πρόσβασης ή πιο λεπτομερείς έλεγχους πρόσβασης που βασίζονται σε χαρακτηριστικά.

Έλεγχος: Καταγράφουν γεγονότα που συμβαίνουν στο σύστημα για να υποστηρίξουν τους ελεγκτές στην ανάλυση τους σε περίπτωση παραβίασης ή καταστροφής δεδομένων, καθώς και για την ιεραρχία των δεδομένων.

3.3 Reference Architecture for Big Data Systems

by Pekka Pääkkönen & Daniel Pakkala [16]

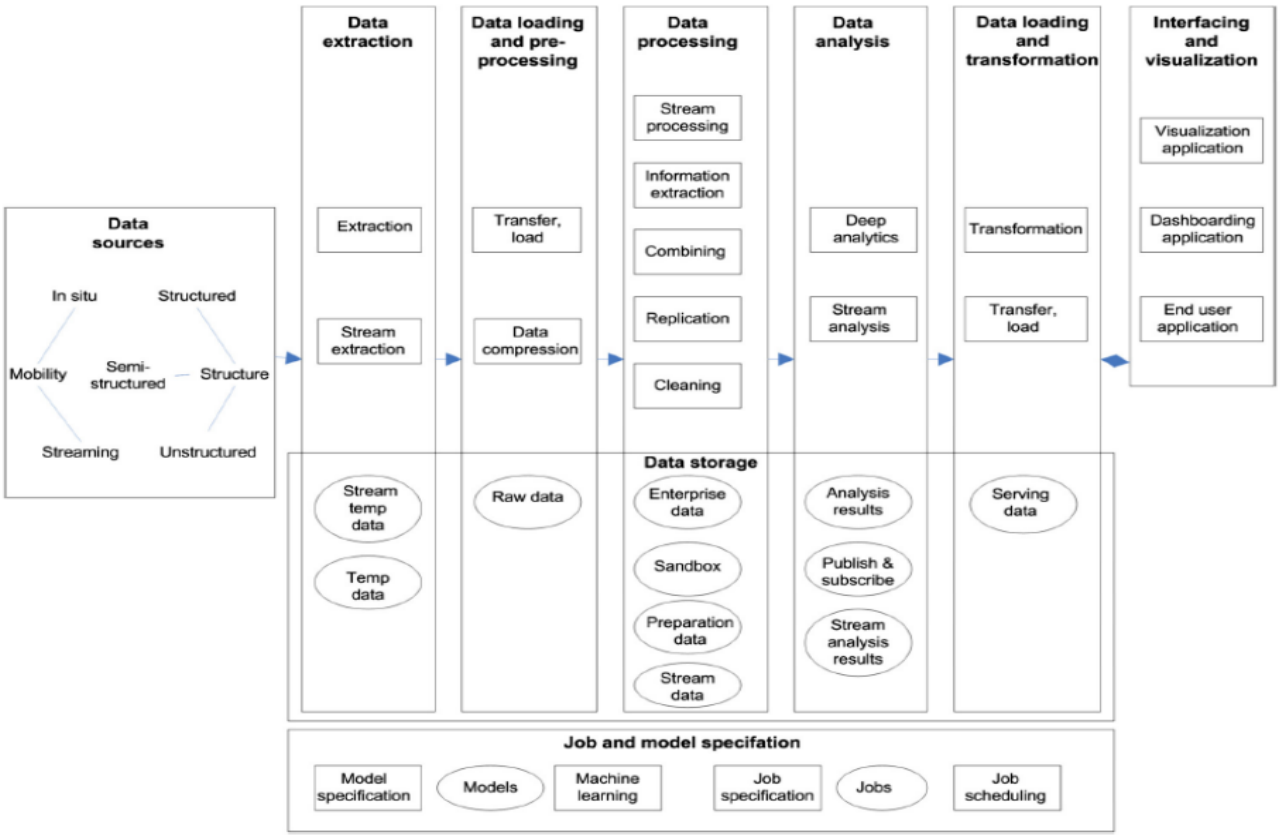
Γενικά

Η αρχιτεκτονική αναφοράς αναπτύχθηκε με το σκεπτικό ότι θα ήταν χρήσιμη με τους ακόλουθους τρόπους: Να διευκολύνει τη δημιουργία συγκεκριμένων αρχιτεκτονικών, και να αυξήσει την κατανόηση των συστημάτων Big Data σαν μια συνολική εικόνα που χρησιμοποιεί τυποποιημένες λειτουργίες και ροές δεδομένων. Σχεδιάστηκε με επαγωγικό συλλογισμό επί δημοσιευμένων περιπτώσεων χρήσης. Συγκεκριμένα, αναλύθηκαν η λειτουργικότητα, οι ροές δεδομένων και η αποθήκευση δεδομένων των αρχιτεκτονικών υλοποίησης σε επτά μεγάλες περιπτώσεις χρήσης δεδομένων. Στη συνέχεια, η αρχιτεκτονική αναφοράς κατασκευάστηκε με βάση αυτήν την ανάλυση.

Δομή της Αρχιτεκτονικής Αναφοράς

Το σχήμα παρουσιάζει το σχεδιασμό της αρχιτεκτονικής αναφοράς. Οι **λειτουργικοτητες** (ορθογώνια), οι **τομείς αποθήκευσης δεδομένων**(ελλείψεις) και **ροές δεδομένων**(βέλη) χρησιμοποιούνται για την αναπαράσταση της αρχιτεκτονικής. Η λειτουργικότητα της επεξεργασίας δεδομένων παρουσιάζεται ως αγωγός, όπου η κίνηση των δεδομένων γίνεται κυρίως από αριστερά προς τα δεξιά. Παρόμοιες λειτουργικότητες έχουν ομαδοποιηθεί σε λειτουργικές περιοχές. Οι τομείς αποθήκευσης δεδομένων παρουσιάζονται μαζί με τις αντίστοιχες λειτουργικές περιοχές. Αντίθετα, ο καθορισμός των εργασιών και των μοντέλων έχουν απεικονιστεί ξεχωριστά από τον αγωγό δεδομένων.

Σχηματική Αναπαράσταση



Εικόνα 25.Reference Architecture for Big Data Systems

Οι δομικές εννοιες της αρχιτεκτονική αναφοράς αναλύονται παρακάτω

Οι **πηγές δεδομένων (data sources)** ορίζονται σε δύο διαστάσεις, την **κινητικότητα(mobility)** και τη **δομή(structure)** των δεδομένων. Πρώτον, ως **in situ** αναφέρεται σε δεδομένα, τα οποία δεν κινούνται. Ένα παράδειγμα in situ δεδομένων είναι ένα αρχείο Hadoop προς επεξεργασία μέσω MapReduce. Τα **streaming** δεδομένα αναφέρονται σε μια ροή δεδομένων προς επεξεργασία σε πραγματικό χρόνο, π.χ. μια ροή Twitter. Δεύτερον, προσδιορίζεται η δομή της πηγής δεδομένων. Τα δομημένα δεδομένα (**structured**) έχουν ένα αυστηρό

μοντέλο δεδομένων. Τέτοιο παράδειγμα είναι περιεχόμενο μιας σχεσιακής βάσης δεδομένων, η οποία είναι δομημένη βάσης ένα σχήμα βάσης δεδομένων. Τα μη δομημένα (**unstructured**) δεδομένα δεν συνδεόνται με κάποιο μοντέλο δεδομένων. Τα περιεχόμενα ιστοσελίδων ή εικόνες γενικά θεωρούνται μη δομημένα δεδομένα. Τα ημι-δομημένα (**semi-structured**) δεδομένα δεν είναι ακατέργαστα ή αυστηρά τυποποιημένα. Άλλες πτυχές των ημιδομημένων δεδομένων περιλαμβάνουν την παρατυπία, την εμπλοκή και τη μερική-δομή , και ένα εξελισσόμενο και ευέλικτο μοντέλο σχήματος / δεδομένων. Παραδείγματα ημι- δομημένων δεδομένα περιλαμβάνουν έγγραφα XML και JSON.

Η εξαγωγή (**extraction**)αναφέρεται στην εισαγωγή δεδομένων **in situ** στο σύστημα. Οταν τα **in situ** δεδομένα εξάγονται **(extracted)**, μπορεί να αποθηκευτούν προσωρινά σε μια αποθήκη δεδομένων δεδομένα (**Temp Data Store**) ή μεταφέρονται και φορτώνονται σε κάποιο **Raw Data Store** . Τα δεδομένα ροής(streaming) μπορούν επίσης να εξάγονται και να αποθηκεύονται προσωρινά,(**Stream Temp Data Store**). Η αποτελεσματικότητα μπορεί να βελτιωθεί από την **συμπίεση -compressing** των δεδομένων που εξάγονται πριν από τη μεταφορά (**transfer**) και τις εργασίες φόρτωσης (**load**). Ο σκοπός του **Raw Data Store** είναι να διατηρεί μη συμπιεσμένα τα δεδομένα. Τα δεδομένα από το **Raw Data Store** μπορούν να καθαριστούν ή να συνδυαστούν και να αποθηκευτούν σε μια νέα αποθήκευση δεδομένων προετοιμασίας, το οποίο προσωρινά βρίσκεται σε επεξεργασία δεδομένων. Ο καθαρισμός(**cleansing**) και ο συνδυασμός(**combining**) αφορούν τη βελτίωση της ποιότητας των πρωτογενών μη επεξεργασμένων δεδομένων. Τα ακατέργαστα και επεξεργασμένα δεδομένα μπορούν να επαναληφθούν(**replicated**) μεταξύ δομών αποθήκευσης. Επίσης, μπορούν να εξαχθούν νέες πληροφορίες (**information**) από το raw data store για σκοπους **Deep Analytics**. Η εξαγωγή πληροφοριών(**information-extraction**) αναφέρεται στην αποθήκευση ακατέργαστων δεδομένων σε δομημένη μορφή. Η επιχειρησιακή αποθήκη δεδομένων (**Enterprise Data Store**) χρησιμοποιείται για την αποθήκευση καθαρισμένων και επεξεργασμένων δεδομένων. Το **Sand-box Store** χρησιμοποιείται στη συγκέντρωση δεδομένων για πειραματικούς σκοπούς της ανάλυσης δεδομένων.

Τα **Deep Analytics** αφορούν την εκτέλεση εργασιών επεξεργασίας batch σε in situ δεδομένα. Τα αποτελέσματα της ανάλυσης μπορούν να αποθηκευτούν ξανά στο αρχικές δομές αποθήκευσης, σε μια δομή αποθήκευσης δεδομένων ανάλυσης (**Analysis Results Store**) ή σε μια δομή δημοσίευσης & συνδρομής (**Publish & Subscribe Store**). Η δομή δημοσίευσης & συνδρομής επιτρέπει η αποθήκευση και η ανάκτηση των αποτελεσμάτων της ανάλυσης γίνεται έμμεσα μεταξύ των συνδρομητών και όσων δημοσιεύουν στο σύστημα. Η επεξεργασία ροής(strem proccessing) αναφέρεται στην επεξεργασία εξερχόμενων δεδομένων συνεχούς ροής, τα οποία πιθανόν να αποθηκεύονται κιόλας προσωρινά πριν την ανάλυση. Η ανάλυση ροής (**Stream Analysis**) αναφέρεται στην ανάλυση **stream data** (τα οποία αποθήκευονται σε αποτελέσματα ανάλυσης ροής-**stream analysis results**). Τα αποτελέσματα της ανάλυσης δεδομένων μπορεί επίσης να μετασχηματιστούν σε μια δομή δεδομένων εξυπηρέτησης **Serving Data Store**, το οποίο εξυπηρετεί εφαρμογές διασύνδεσης και οπτικοποίησης **interfacing and visualization applications**. Μια τυπική εφαρμογή για μετασχηματισμό (**transformation**) και δεδομένων εξυπηρέτησης **Serving Data Store** εξυπηρετεί OLAP ερωτήματα.

Τα αναλυθέντα δεδομένα μπορούν να απεικονιστούν με διάφορους τρόπους. Η εφαρμογή **Dashboard** αναφέρεται σε ένα απλό UI, όπου τυπικά οι βασικές πληροφορίες (όπως π.χ. τα KPI Key Performance Indicator) απεικονίζονται χωρίς έλεγχο απο το χρήστη. Η εφαρμογή οπτικοποιήσης (**Visualization Application**) παρέχει λεπτομερή απεικόνιση και λειτουργίες ελέγχου και συνήθως υλοποιείται μεσω ένός εργαλείο επιχειρηματικής ευφυίας(BI Business Intelligence) στο domain της επιχείρησης. Η εφαρμογή τελικού χρήστη (**End User Application**) έχει ένα περιορισμένο σύνολο λειτουργιών ελέγχου και θα μπορούσε να δωθεί στους τελικούς χρήστες και στην μορφή εφαρμογής για κινητές συσκευές

Οι εργασίες (**jobs**) επεξεργασίας batch μπορεί να καθορίζονται στη διεπαφή χρήστη. Οι εργασίες αυτές μπορούν να αποθηκευτούν και να προγραμματιστούν με εργαλεία προγραμματισμού εργασιών **job scheduling tools**. Μοντέλα και αλγόριθμοι μπορούν επίσης να καθορίζονται στη διεπαφή χρήστη (**Model specification**). Η μηχανική μάθηση (**machine learning**) μπορεί να χρησιμοποιηθεί για την εκπαίδευση μοντέλων που βασίζονται στα νέα εξαγόμενα δεδομένα.

3.4 Big Data Architecture Framework BDAF

by Yuri Demchenko, Cees de Laat, Peter Membrey [20]

Μοντέλα Δεδομένων, Δομές και Τύποι | Data Models, Structures, and Types:

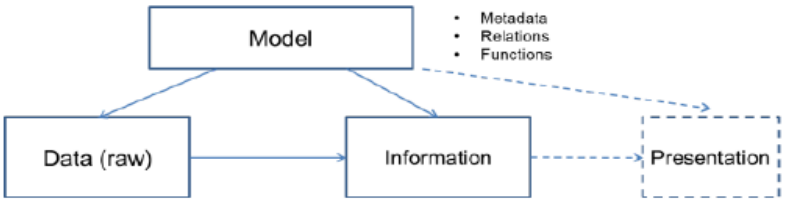
Το BDAF υποστηρίζει μια ποικιλία τύπων δεδομένων που παράγονται από διαφορετικές πηγές. Αυτά τα δεδομένα πρέπει να αποθηκεύονται και να επεξεργάζονται και, σε ορισμένες να καθορίζουν τις τεχνολογίες και τις λύσεις των υποδομών Big Data που θα ακολουθηθούν. Οι τύποι που δέχεται το μοντέλο συμφωνούν με αυτούς που έχει ορίσει το NBDWG-NIST Big Data Working Group δηλαδή:

-Δεδομένα που περιγράφονται μέσω ενός τυπικού μοντέλου δεδομένων (formal data model)

-Δεδομένα που περιγράφονται μέσω μιας επίσημης γραμματικής(formalized grammar)

-Δεδομένα που περιγράφονται μέσω μιας τυποποιημένης μορφοποίησης(standard format)

-Αυθαίρετα δεδομένα σε κείμενο ή δυαδική μορφή



Εικόνα 26.Δομές Δεδομένων Big Data

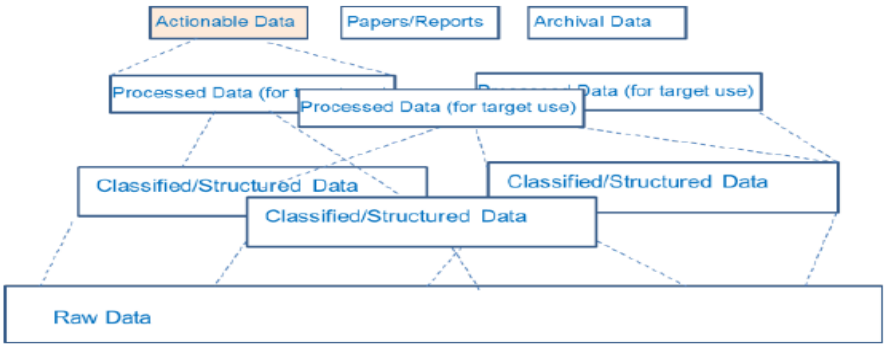
Επίσης ορίζονται τα εξής είδη επιστημονικών δεδομένων:

-Τα ακατέργαστα δεδομένα (**raw data**) που συλλέγονται από την παρατήρηση και από το πείραμα (σύμφωνα με ένα αρχικό ερευνητικό μοντέλο)

-Δομημένα δεδομένα και σύνολα δεδομένων (**structured data**) που διήλθαν από φιλτράρισμα δεδομένων και την επεξεργασία (υποστηρίζοντας ένα συγκεκριμένο επίσημο μοντέλο)

-Δημοσιευμένα δεδομένα (**published**) που υποστηρίζουν μία επιστημονική υπόθεση,ένα αποτέλεσμα έρευνας ή μια δήλωση

-Στοιχεία που συνδέονται με δημοσιεύσεις για την υποστήριξη της ευρείας ενοποίησης στην έρευνα, την ολοκλήρωση και τη διαφάνεια.

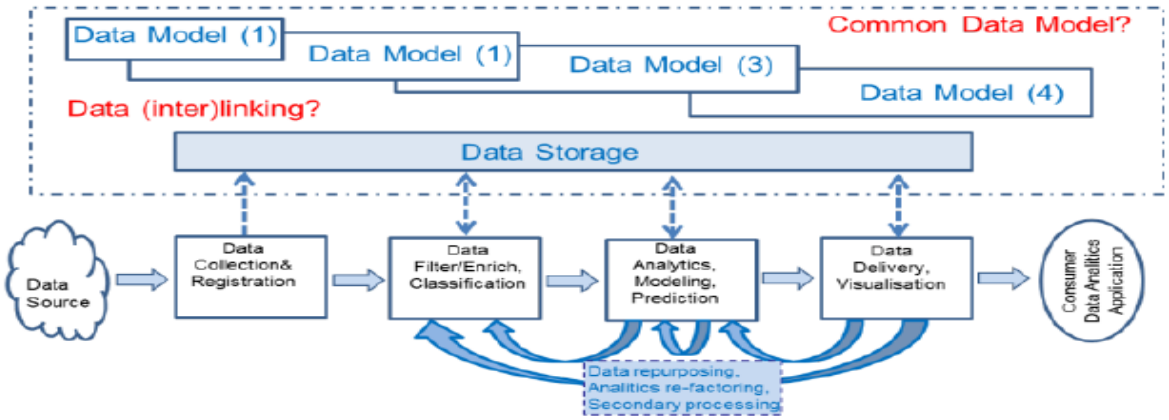


Εικόνα 27.Linkage Δεδομένων Big Data

Διαχείρισης Big Data και Κύκλος Ζωής | BDLM

Το BDAF υποστηρίζει τη Διαχείριση Κύκλων Ζωής Big Data δηλαδή την προέλευση, διατήρηση και αρχειοθέτηση με ένα μοντέλο που ονομάζει BDLM - **Big Data Lifecycle Management**. Το BDLM λοιπόν υποστηρίζει τα σημαντικά στάδια μετασχηματισμού δεδομένων: συλλογή, καταχώριση, φιλτράρισμα, ταξινόμηση, ανάλυση, μοντελοποίηση, πρόβλεψη, παράδοση, παρουσίαση και οπτικοποίηση. Οι δυνατότητες διαχείρισης Big Data μπορούν να αντιμετωπιστούν εν μέρει με τον καθορισμό επιστημονικών ή επιχειρηματικών ροών εργασίας και

χρησιμοποιώντας αντίστοιχα συστήματα διαχείρισης ροών εργασίας. Στο νέο BDLM απαιτείται όλα τα αυτά τα στάδια αποθήκευσης και συντήρησης δεδομένων να επιτρέπουν την επαναχρησιμοποίηση/επανεξέταση (re-use/re-purposing) των δεδομένων και δευτερογενώς την διερεύνηση/αναλύση σχετικά με τα επεξεργασμένα δεδομένα και τα δημοσιευμένα αποτελέσματα. Ωστόσο, αυτό είναι δυνατό μόνο εάν η πλήρης ταυτοποίηση, διασταύρωση και ιχνηλάτηση εφαρμόζονται στο επίπεδο της υποδομής **BDI**. Η ακεραιότητα των δεδομένων και ο έλεγχος πρόσβασης θα πρέπει να υποστηρίζεται καθ 'όλη τη διάρκεια του κύκλου ζωής των δεδομένων. Η επιδιόρθωση/συμπλήρωση δεδομένων είναι επίσης ένα σημαντικό στοιχείο του μοντέλου το οποίο πρέπει επισης να υλοποείται κατά τρόπο ασφαλή και εμπιστευτικό.



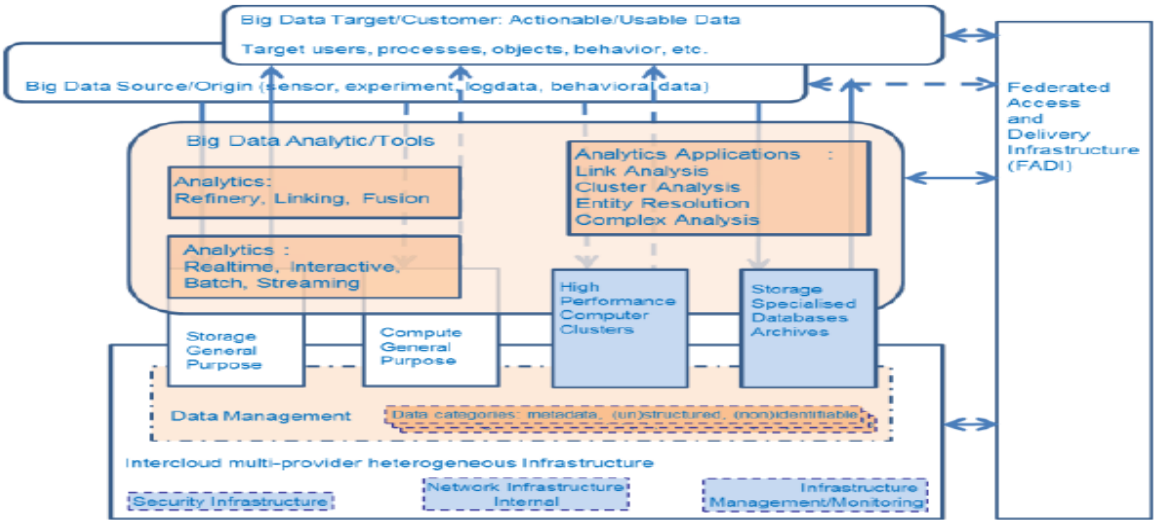
Εικόνα 28.Κύκλος Ζωής Δεδομένων Big Data

Big Data Analytics και Εργαλεία

Αυτά αφορούν τις απαιτούμενες λειτουργίες μετασχηματισμού δεδομένων καθώς και τα σχετικά συστατικά στοιχεία της υποδομής. Εκτός από τις γενικές υπηρεσίες υποδομής βάσης cloud (αποθήκευση, υπολογιστική, διαχείριση υποδομής / Virtual Machines) οι παρακάτω εφαρμογές και υπηρεσίες απαιτούνται για την υποστήριξη Big Data και άλλων data-centric εφαρμογών της οποίες αναφέρονται γενικότερα ως **Big Data Analytics Infastructures (BDAI)**:

- Υπηρεσίες Συστοιχιών (Cluster Services)
- Υπηρεσίες και εργαλεία που σχετίζονται με το Hadoop
- Ειδικά εργαλεία ανάλυσης δεδομένων (αρχεία καταγραφής-logs, γεγονότα-events, εξόρυξη γνώσης κτλπ)
- Βάσεις Δεδομένων / Εξυπηρετητές SQL, NoSQL
- Βάσεις Δεδομένων MPP (Massive Parallel Processing)

Εργαλεία ανάλυσης μεγάλων δεδομένων προσφέρονται επίσης από τους μεγάλους παρόχους υπηρεσιών cloud όπως: Amazon Elastic MapReduce και το Dynamo, το Microsoft Azure HDInsight, το IBM Big Data Analytics. Κλιμακούμενα εργαλεία ανάλυσης δεδομένων και Hadoop επίσης προσφέρονται από λίγες εταιρείες που θεωρούνται εταιρίες αμιγώς Big Data όπως η Cloudera και άλλες μικρότερες περισσότερο εξειδικευμένες σε συγκεκριμένους κλάδους (10Gen,ClearStoryData, Climate FiedView,CDAP,Dataguise,NuoDB,ZoomData)



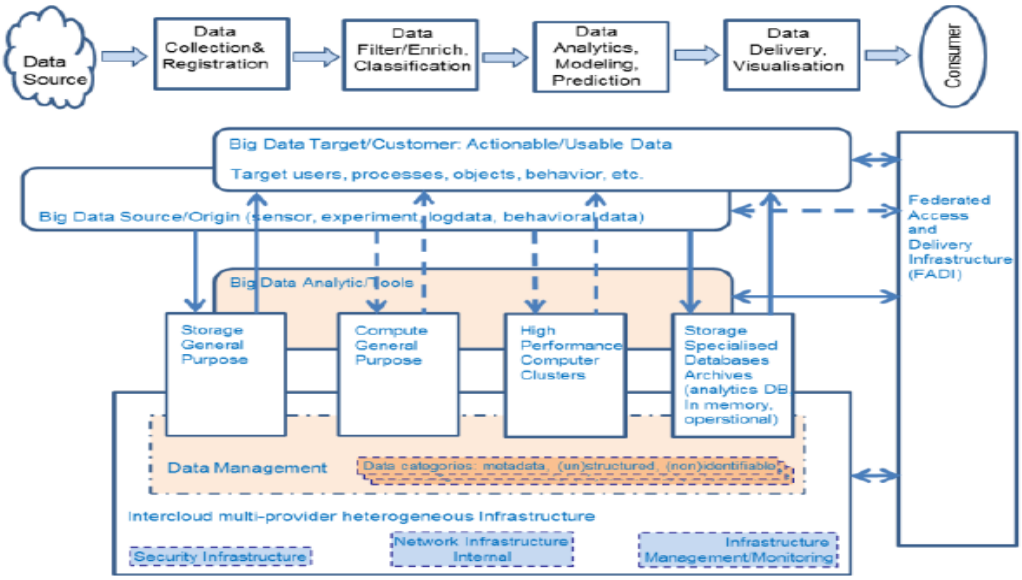
Εικόνα 29.Υποδομή Big Data Analytics

Υποδομή Big Data | Big Data Infastructure - BDI

Αυτό το στοιχείο περιλαμβάνει αποθήκευση, υπολογιστική υποδομή, την υποδομή δικτύου, τα δίκτυα αισθητήρων και τις τελικές ή ενδιάμεσες συσκευές. Μια γενική άποψη περιλαμβάνει τη γενική υποδομή και την γενικού χαρακτήρα διαχείριση δεδομένων, συνήθως βασισμένη σε Cloud τεχνολογία, το Big Data Analytics μέρος που θα απαιτήσει clusters υψηλών υπολογιστικών επιδόσεων, οι οποίες με τη σειρά τους θα απαιτήσουν υψηλής απόδοσης δίκτυο με μικρό περιθώριο λαθών. Οι γενικές υπηρεσίες και τα στοιχεία του BDI περιλαμβάνουν

- Εργαλεία Διαχείρισης Δεδομένων Big Data
- Μητρώα, ευρετηρίαση / αναζήτηση, σημασιολογία, χώροι ονομάτων
- Υποδομή ασφαλείας (έλεγχος πρόσβασης, επιβολή πολιτικών, εμπιστευτικότητα, εμπιστοσύνη, διαθεσιμότητα, ιδιωτικό απόρρητο)
- Συνεργατικό περιβάλλον (διαχείριση ομάδων)

Επίσης ορίζεται η **Federated Access and Delivery Infrastructure (FADI)** ως σημαντική συνιστώσα του γενικού BDI που διασυνδέει διάφορα στοιχεία του cloud / Intercloud υποδομής που συνδυάζει προβλέψης δέσμευσης πόρων συνδεσιμότητας δικτύου και ομοσπονδιακό έλεγχος πρόσβασης



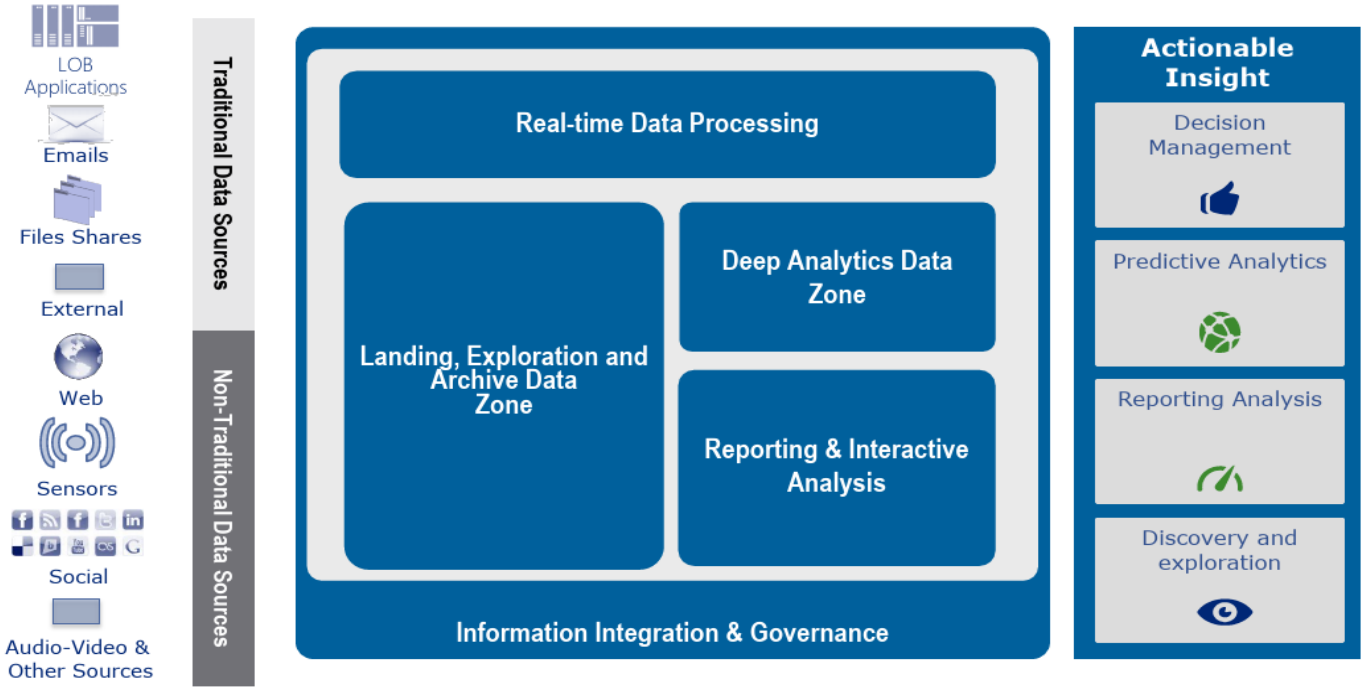
Εικόνα 30.Λειτουργικά Στοιχεία Υποδομής Big Data

Ασφάλεια Big Data

Η ασφάλεια πρέπει να προστατεύει τα δεδομένα σε ηρεμία και κίνηση, εξασφαλίζοντας αξιόπιστη επεξεργασία περιβάλλοντα επεξεργασίας και αξιόπιστη λειτουργία της BDI, παρέχοντας λεπτομερή έλεγχο πρόσβασης και προστασία των χρηστών αλλά και των προσωπικών πληροφοριών

3.5 DataZones Architecture

by Deloitte [14]



Εικόνα 31.DataZones Architecture

Landing, Exploration and Archive DataZone

Ζώνη Προσεδάφωσης | Landing Zone

Η ζώνη προσγείωσης είναι αυτή που ήταν παλαιότερα γνωστή ως **staging**. Το Hadoop διαδραματίζει κεντρικό ρόλο σε αυτή.Είναι μια περιοχή όπου:

• Ο χώρος αποθήκευσης δεν είναι πια πρόβλημα:

-Αντί να λαμβάνουμε υποσύνολα, όλα τα δεδομένα μπορούν να είναι από εν λειτουργία σύστημα

-Επιτρέπεται στους χρήστες να αναλύουν και να ανακαλύπτουν δεδομένα οι ίδιοι

• Τα δεδομένα μπορούν να αποθηκευτούν χωρίς να χρειάζεται να προκαθοριστεί ένα σχήμα,πράγμα χρήσιμο για περιπτώσεις μη-δομημένων δεδομένων.

Η νέα Ζώνη Staging

Με την εξέλιξη των τεχνολογιών, η "παλιά" staging διαδικασία μπορεί τώρα να χωριστεί σε μια ζώνη **Landing** και μια διακριτή ζώνη **Staging**. Ωστόσο, με την υιοθέτηση του Hadoop, ο μετασχηματισμός μπορεί να γίνει ταχύτερος και αποθήκευση φθηνότερη.

Στη ζώνη αυτή, τα δεδομένα μετατρέπονται στο μοντέλο δεδομένων στόχου και τοποθετούνται σε μια "staging" ζώνη, από την οποία μπορούν να φορτωθούν απευθείας στο **Enterprise Data Warehouse**-EDW

Αυτή η ζώνη διατηρείται συνήθως στο ίδιο σύστημα Hadoop με τη ζώνη landing

Ζώνη Αρχειοθέτησης | Archiving Zone

Το Hadoop αποτελεί επίσης κατάλληλη περιοχή αρχειοθέτησης. Χρησιμοποιώντας ευρέως διαδεδομένο υλικό γίνεται αρκετά φθηνότερη. Εάν όλα τα δεδομένα σταδιακά μεταπέσουν σε τεχνολογία Hadoop, τα παλαιά δεδομένα μπορούν να καθαριστούν από το EDW.

Η ζώνη αρχειοθέτισης πλέον μετατρέπεται σε μια ζώνη διαθέσιμη για ερωτήματα,η οποία και χτίζεται καθημερινά ξεκινώντας από την ημέρα μηδέν.

Η ζώνη staging αποθηκεύει τα δεδομένα σε ένα μοντέλο ίδιο ή παραπλήσιο του EDW, δημιουργώντας έτσι ένα αντίγραφο των περισσότερων δεδομένων στο Hadoop, σε μορφή αξιοποιήσιμη μέσω queries

Ζώνη Εξερεύνησης | Exploration Zone

Το Hadoop χρησιμοποιούμενο ως ζώνη προσεδάφωσης είναι ισχυρό καθώς παρέχει τη δυνατότητα να "κάνουν κάτι " με τα δεδομένα πριν αυτά μετατραπούν ή δομηθούν. Ενώ λοιπόν είναι ο τόπος όπου συλλέγονται τα πρωτογενή δεδομένα και αποθηκεύονται, η ζώνη εξερεύνησης παρέχει την επιθυμητή αυτή πρόσβαση στα δεδομένα.

Είναι δηλαδή μια περιοχή όπου:

• Τα μεταδεδομένα μπορούν να οριστούν για να δημιουργήσουν διασύνδεση μεταξύ της πληροφορίας που εξάγεται από μη δομημένα δεδομένα,την προέλευση δεδομένων κλπ

• Οι χρήστες μπορούν να συνδυάσουν τα δεδομένα με διαφορετικούς τρόπους και να τα οργανώσουν σύμφωνα με τις δικές τους ειδικές επιχειρηματικές τους ανάγκες

• Μπορούν να γίνουν δοκιμές διαφορετικών υποθέσεων, ανακάλυψη μοτίβων,συσχετισμοί

• Τα δεδομένα είναι απο μόνα τους αυτά που ορίζουν την επιχειρηματική κατεύθυνση

Δημιουργία Αναφορών και Διαδραστική Ανάλυση | Reporting & Interactive Analysis

Η ζώνη **Reporting & Interactive Analysis** είναι ένας τρόπος παρουσίασης των δεδομένων σε διαφορετικές επιχειρηματικές ομάδες. Ιστορικά, αυτό γίνεται με την κατασκευή διακριτών **datamarts**.

Αυτές οι απαιτήσεις θα παραμείνουν και πιθανότατα θα συνεχίσουν να βασίζονται σε σχεσιακές βάσεις δεδομένων και στο προσεχές μέλλον.

Η κύρια πρόοδος σε αυτόν τον τομέα είναι η υιοθέτηση τεχνολογιών **Massively Parallel Processing** (MPP) ή **in-memory** με σκοπό τη βελτίωση των επιδόσεων των ερωτημάτων.

Ζώνη Ανάλυσης Δεδομένων Εις Βάθος | Deep Data Analytics

Διαφορετικά συστήματα είναι ειδικά κατασκευασμένα για αυτόν τον τύπο φόρτου εργασίας αναλυτικών διεργασιών, ιδιαιτερα αυτά που αποκαλούνται Appliances.

Τα appliances:

- Έχουν εξ'αρχής σχεδιαστεί για φόρτο εργασίας αναλύσεων και όχι για τυπικά ερώτημα
- Σχηματίζουν ένα ολοκληρωμένο σύνολο διακομιστών, αποθήκευσης, λειτουργικών συστημάτων, προκαθορισμένων βάσεων δεδομένων
- Χρησιμοποιούν τεχνικές **MPP** ή **in-memory**
- Περιέχουν προκαθορισμένους αλγόριθμους ανάλυσης

Σημαντικό έδαφος στην ζώνη αυτή έχει κερδίσει τα τελευταία χρόνια το Spark το οποίο λειτουργεί ως ένα ενιαίο στρώμα για τους επιστήμονες δεδομένων.Υποστηρίζει γενικό φόρτο εργασίας καθώς και streaming, διαδραστικά ερωτήματα και μηχανική μάθηση παρέχοντας αρκετά οφέλη από άποψη απόδοσης

Ζώνη Επεξεργασίας Δεδομένων σε Πραγματικό Χρόνο | Real-Time Data Processing

Οι απαιτήσεις πραγματικού χρόνου εμφανίστηκαν για να επεξεργαστούν τον μεγάλο όγκο και την ταχύτητα των δεδομένων (Clickstreams, Internet Of Things, Logs).

Διαφορες μηχανές επεξεργασίας σύνθετων συμβάντων είναι ήδη διαθέσιμες στην αγορά (Π.χ SAS Event Stream Processing, IBM InfoSphere Stream).

Έτσι προσφέρεται η δυνατότητα στις εταιρείες να επεξεργάζονται τα δεδομένα τη στιγμή ακριβώς που παράγονται και να εξαγάγουν συγκεκριμένες πληροφορίες χωρίς να χρειάζεται να αντιγράψει ή να υποδεχτούν όλα τα περιττά δεδομένα, λύνοντας έτσι το ζήτημα της αποθήκευσης και ενδεχομένως την απόρριψη δεδομένων λόγω έλλειψης χώρου.

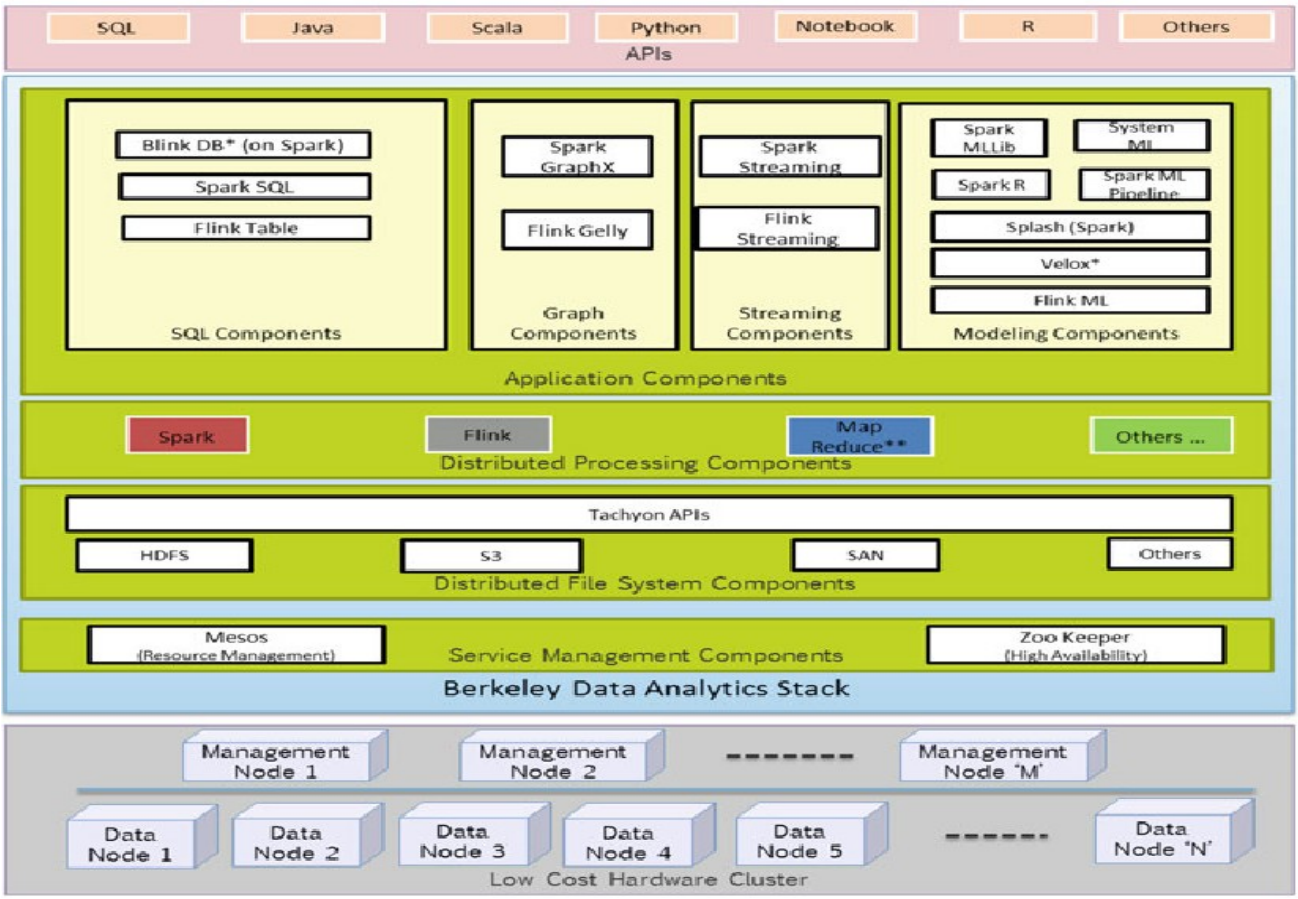
Αυτή η ζώνη δεν αποθηκεύει τα δεδομένα, είναι μια ζώνη επεξεργασίας όπου τα δεδομένα αναλύονται ή μετασχηματίζονται όπως παράγονται από διάφορες πηγές.

3.6 The Berkeley Data Analytics Stack – BDAS

by Berkeley's Amp Lab [19]

Το Berkeley Data Analytics Stack (BDAS) αποτελείται από ένα σύνολο πλαισίων Big Data που είναι σχετικά νέα στον κλάδο. Ωστόσο, έχουν λάβει τεράστια δημοτικότητα και υιοθέτηση τα τελευταία χρόνια που άξιζε να το αναγνωρίσουμε ως μια ξεχωριστή στοίβα αυτόνομων Big Data Tools και Platforms. Πρωτίστως τροφοδοτείται από την έρευνα στο Amp Lab στο Berkeley, όπως και το Hadoop Ecosystem,τα πλαίσια εντός του BDAS είναι επίσης γνωστά ως Cluster Computing Frameworks. Μπορούν να παρέχουν βασικές δυνατότητες που απαιτούνται για τις τυπικές απαιτήσεις Big Data όπως η γρήγορη επεξεργασία δεδομένων σε μια συστοιχία χρησιμοποιώντας μνήμη που είναι διαθέσιμη στους διάφορους κόμβους δεδομένων της συστοιχίας, αποθηκεύοντας τα δεδομένα στη μνήμη με κατανεμημένο τρόπο, διαχείριση υπολογιστικών πόρων σε διάφορους τύπους περιπτώσεων χρήσης κ.λπ. Καθώς τα πλαίσια στο BDAS εφαρμόζουν κυρίως περιπτώσεις χρήσης Big Data παρόμοιες με αυτές αντιμετωπίζονται από το Hadoop Ecosystem, έχουν ως στόχο την επίτευξη το ίδιο με πολύ μικρότερο μέγεθος συστοιχίας μειώνοντας έτσι το κόστος και τα γενικά τα έξοδα διαχείρισης. Επιπλέον, σε αντίθεση με το Hadoop Ecosystem, το BDAS έχει πρόσβαση και αποθήκευση δεδομένων από / προς τοπικές, δικτυακές ή cloud αποθηκεύσεις δεδομένων (S3, NFS κ.λπ.) και άλλες βάσεις δεδομένων (Cassandra, HBase, RDBMS κ.λπ.). Επίσης το BDAS χρησιμοποιεί την "**Analytics First**" προσέγγιση αντί για την "**Storage First**" προσέγγιση του Hadoop Ecosystem και άλλων τεχνολογιών Big Data. Οι τεχνολογίες στη στοίβα BDAS έχουν δημιουργηθεί και αναπτυχθεί για την υποστήριξη της δημιουργίας **insights** και της λειτουργικότητας με τέτοιο τρόπο ώστε τα ληφθέντα ανεπεξέργαστα δεδομένα να ακολουθούν πρώτα όλα τα πολύπλοκα βήματα ανάλυσης που απαιτούνται για τη δημιουργία του insight χωρίς απαίτηση ενδιάμεσης αποθήκευσης. Στη συνέχεια, το τελικό Insight μπορεί να αποθηκευτεί (να γίνει push) σε μια μόνιμη μορφή αποθήκευσης καταλληλη για κατανάλωση.

Τα διάφορα στρώματα BDAS που προτείνονται είναι τα παρακάτω:



Εικόνα 32.Τα επίπεδα του Berkeley Data Analytics Stack

Τα στοιχεία επομένως της αρχιτεκτονικής διακρίνονται στα εξής στρώματα/επίπεδα :

Το **Κατανεμημένου Συστήματος Αρχείων (Distributed File System),** όπου βοηθούν στην αποθήκευση και πρόσβαση στα δεδομένα διάφορων λύσεων αποθήκευσης δεδομένων όπως HDFS,Amazon S3, SAN κ.λπ. Βοηθούν επίσης την πρόσβαση των διάφορων Στοιχείων Κατανεμημένης Επεξεργασίας (όπως Spark, Map Reduce, Flink κλπ.) να έχουν πρόσβαση στα δεδομένα που είναι αποθηκευμένα στη μνήμη.

Της **Κατανεμημένης Επεξεργασίας (Distributed Processing)** βοηθούν στην πραγματική επεξεργασία των δεδομένων σε πολλαπλούς κόμβους του υλικού διαμοιράζοντας παραλληλα την υπολογιστική ισχύ του κάθε κόμβου. Οι πρώτοι υποψήφιοι εδώ είναι οι Spark και Flink. Ωστόσο στο μέλλον άλλα πλαίσια όπως το MapReduce μπορούν επίσης να χρησιμοποιηθεί με τον ίδιο τρόπο όπως το Tachyon υποστηρίζει το πλήρες σύνολο των APIs για HDFS.

Τις **Εφαρμογές (Application)** που παρέχουν τις απαραίτητες αφαιρέσεις επι με της Κατανεμημένης Επεξεργασίας (προηγούμενο επίπεδο) έτσι ώστε το τελευταίο να μπορεί να χρησιμοποιηθεί για διαφορετική περιπτώσεις χρήσης Big Data σχετικα με διαδραστικά ερωτήματα, Ροες, Μοντελοποίηση και Επεξεργασία Γράφων, κ.λπ.

Το **Επίπεδο API** που είναι το επίπεδο που υποστηρίζει κοινές διεπαφές όπως SQL, Java, R,Scala, και Python, έτσι ώστε τα δεδομένα να είναι εύκολα προσβάσιμα και να ενσωματωθούν από άλλα εργαλεία και τεχνολογίες επιχειρήσεων.

Την **Διαχείριση Υπηρεσιών (Service Management)** που βοηθούν στη διαχείριση των υπηρεσιών για υψηλού επιπέδου διαθεσιμότητα, δυναμική προτεραιοποίηση της χρήσης των πόρων και κατανομή.

3.7 Lamda Architecture

[23.Marz, Warren]

Η κύρια ιδέα της αρχιτεκτονικής Lambda είναι η κατασκευή συστημάτων Big Data ως μια σειρά από στρώματα(layers). Κάθε στρώμα ικανοποιεί ένα υποσύνολο των ιδιότητες και βασίζεται στη λειτουργικότητα που παρέχεται από τα παρακάτω επίπεδα.Όλα ξεκινούν από την εξίσωση:

```
query = function (all data)
```

Στην ιδανική περίπτωση, μπορείτε να εκτελέσετε τις συναρτήσεις εν κινήσει για να λάβετε τα αποτελέσματα. Δυστυχώς, ακόμη και αν αυτό ήταν δυνατό, θα χρειαζόταν ένα τεράστιο ύψος πόρων που πρέπει να γίνουν και θα το καταστήσουν υπερβολικά ακριβό. Φανταστείτε για παράδειγμα ότι πρέπει να διαβάσετε ένα σύνολο δεδομένων μεγέθους petabyte κάθε φορά που θέλετε να απαντήσετε στο ερώτημα της τρέχουσας τοποθεσία κάποιου ατόμου. Η πιο προφανής εναλλακτική προσέγγιση είναι να έχει υπολογιστεί εκ των προτέρων η συνάρτηση ερωτήματος. Ας ονομάσουμε την συνάρτηση του προκαθορισμένου ερωτήματος ως **batch view**. Αντί να υπολογίσει το ερώτημα εκείνη την στιγμη, διαβάζετε τα αποτελέσματα από την προυπολογισμένη batch view. Η προ-υπολογισμένη view είναι ευρετηριασμένη έτσι ώστε να είναι προσβάσιμη και μέσω τυχαίων αναγνώσεων . Αυτό το σύστημα μοιάζει με αυτό

```
batch view = function(all data)
query = function(batch view)
```

Σε αυτό το υπόδειγμα συστήματος λοιπόν, εκτελείτε μια συνάρτηση σε όλα τα δεδομένα για να δημιουργήσετε την batch view. Τότε όταν θα θελήσετε να μάθετε την τιμή για ένα ερώτημα, εκτελείτε μια συνάρτηση επ'αυτής της batch view.Έτσι καθιστάται δυνατή η γρήγορη λήψη των τιμών που χρειάζεστε από αυτήν, χωρίς πρέπει να σαρώσετε τα πάντα.

Γενικά λοιπόν ,η αρχιτεκτονική Lambda είναι μια τεχνική επεξεργασίας δεδομένων που μπορεί να χειρίζεται τεράστιο όγκο δεδομένων με αποτελεσματικό τρόπο. Η αποτελεσματικότητα αυτής της αρχιτεκτονικής γίνεται εμφανής με τη μορφή αυξημένης απόδοσης, του μειωμένου λανθάνοντος χρόνου και αμελητέων σφαλμάτων. Ο στόχος επικεντρώνεται συνήθως σε εφαρμογές σχεδόν πραγματικού χρόνου (**near real time**). Αυτό επίσης επιτρέπει στους προγραμματιστές να καθορίσουν δέλτα μηχανισμούς με τη μορφή λογικής κώδικα ή κανόνων επεξεργασίας φυσικής γλώσσας (NLP) σε μοντέλα επεξεργασίας δεδομένων συμβάντων(event-based) για να επιτύχουν ευρωστία, αυτοματοποίηση και αποτελεσματικότητα αλλά και να βελτιώσουν την ποιότητα των δεδομένων. Επιπλέον, οποιαδήποτε αλλαγή στην κατάσταση των δεδομένων αποτελεί για το σύστημα ένα συμβάν και στην πραγματικότητα είναι δυνατόν να γεννήσει μια εντολή, ένα ερώτημα ή να πραγματοποιήσει διαδικασίες δέλτα ως απάντηση στα εν εξελίξει αυτά γεγονότα.

Ο χειρισμός όλων των πηγών με την λογική των συμβάντων αποτελεί την ιδέα της χρήσης των συμβάντων για εξαγωγή προβλέψεων καθώς και για την αποθήκευση των αλλαγών ενός συστήματος σε πραγματικό χρόνο για κάθε αλλαγή κατάστασης, μια ενημέρωση στις βάσεις δεδομένων ή οποιοδηποτε άλλο γεγονός μπορεί να γίνει κατανοητό ως αλλαγή. Για παράδειγμα, εάν κάποιος αλληλεπιδράσει με μια ιστοσελίδα ή ένα προφίλ κοινωνικού δικτύου, συμβάντα όπως η προβολή σελίδας, το "Μου αρέσει" ή το "Προσθέστε ως φίλο" κλπ. ενεργοποιούν

γεγονότα που μπορούν να επεξεργαστούν ή να εμπλουτιστούν και δεδομένα που μπορούν να αποθηκευτούν σε μια βάση δεδομένων.

Η επεξεργασία δεδομένων ασχολείται με τις ροές συμβάντων και το μεγαλύτερο μέρος του επιχειρησιακού λογισμικού που ακολουθεί το **Domain Driven Design** χρησιμοποιεί τη μέθοδο επεξεργασίας ροής(**stream processing**) για να προβλέψει ενημερώσεις στο βασικό μοντέλο και να αποθηκεύσει τα ξεχωριστά συμβάντα που χρησιμεύουν ως πηγή για προβλέψεις σε ένα πραγματικού σύστημα δεδομένων. Για την αντιμετώπιση πολυάριθμων συμβάντων που συμβαίνουν ή τη επεξεργασίας τύπου δέλτα, η αρχιτεκτονική Lambda καθιστά εφικτή την επεξεργασία δεδομένων εισάγοντας τρία διαφορετικά επίπεδα. Περιλαμβάνει δηλαδή το Batch Layer, το Speed Layer (γνωστό επίσης ως Stream layer) και το Serving Layer τα οποία αναλύονται παρακάτω.

Στρώμα Πατρίδας | Batch Layer

Ονομάζεται το τμήμα της αρχιτεκτονικής Lambda που υλοποιεί τη συνάρτηση

```
batch view  = function(all data)
```

Το batch layer αποθηκεύει το κύριο αντίγραφο του σύνολο δεδομένων(**master dataset**) και προυπολογίζει Batch Views επί αυτού του συνόλου.Το master dataset μπορεί να θεωρηθεί ως μια πολύ μεγάλη λίστα εγγραφών. Το Batch Layer είναι σε θέση να κάνει δύο πράγματα: αποθηκεύει ένα αμετάβλητο, συνεχώς αναπτυσσόμενο σύνολο κύριων δεδομένων και να υπολογίζει γενικές συναρτήσεις επι αυτού του συνόλου δεδομένων. Αυτός ο τύπος επεξεργασίας γίνεται αποδοτικότερα με τη χρήση συστημάτων μαζικής επεξεργασίας(πχ Hadoop). Η απλούστερη μορφή του Batch Layer μπορεί να αναπαρασταθεί σε ψευδοκώδικα όπως ο παρακάτω:

```
function runBatchLayer():
    while(true):
        recomputeBatchViews()
```

Δηλαδή τρέχει σε ένα βρόχο υπό την συνθήκη while(true)και υπολογίζει συνεχώς τη Batch Views από την αρχή.

Τα νέα δεδομένα γενικά λαμβάνονται ως συνεχόμενη τροφοδοσία για το σύστημα δεδομένων. Σε κάθε περίπτωση τροφοδοτείται ταυτόχρονα το στρώμα batch και στο στρώμα ταχύτητας. Κάθε νέα ροή δεδομένων που έρχεται στο batch layer του συστήματος δεδομένων υπολογίζεται και επεξεργάζεται σε μία **Data Lake**. Όταν τα δεδομένα αποθηκεύονται στο data lake χρησιμοποιώντας βάσεις δεδομένων τύπου **in-memory** ή μακροχρόνιας συντήρησης τύπου **NoSQL**, το batch layer χρησιμοποιεί για να επεξεργαστεί τα δεδομένα χρησιμοποιώντας υπολογιστικά μοντέλα τύπου **MapReduce** ή **Μηχανικής Μάθησης** (ML) για να κάνει προβλέψεις για τις επερχόμενες Batch Views.

Στρώμα Εξυπηρέτησης | Serving Layer

Το στρώμα παρτίδας λοιπόν εκπέμπει τις batch views ως αποτέλεσμα των λειτουργιών του. Το επόμενο βήμα είναι να φορτωθούν οι προβολές αυτές κάπου έτσι ώστε να μπορούν να τους υποβληθούν ερωτήματα. Εδώ είναι το σημείο όπου λαμβάνει χώρα το στρώμα εξυπηρέτησης. Το επίπεδο εξυπηρέτησης συνήθως είναι μια εξειδικευμένη κατανεμημένη βάση δεδομένων που φορτώνεται με τις batch views και καθιστά δυνατή την τυχαία προσπέλαση ανάγνωσης σε αυτές. Όταν καινούργιες batch views γίνουν διαθέσιμες, το στρώμα εξυπηρέτησης τις αντικαθιστά αυτόματα ώστε να είναι όσο το δυνατόν πιο ενημερωμένα τα διαθέσιμα αποτελέσματα.

Μια βάση δεδομένων στο επιπέδου εξυπηρέτησης υποστηρίζει μαζικές ενημερώσεις(updates) και τυχαίες αναγνώσεις(random reads) Κυρίως, δεν χρειάζεται να υποστηρίζει τυχαίες εγγραφές(random writes). Αυτό είναι ένα πολύ σημαντικό σημείο καθώς τα random writes είναι γνωστό πως προκαλούν το μεγαλύτερο μέρος της πολυπλοκότητας στις βάσεις δεδομένων. Μη υποστηρίζοντας random writes, αυτές οι βάσεις δεδομένων είναι εξαιρετικά απλές γεγονός που τις καθιστά ισχυρές, προβλέψιμες, εύκολες στη παραμετροποίηση και στη χρήση.

Γενικά λοιπόν,οι έξοδοι από το στρώμα παρτίδας με τη μορφή batch views και από στρώμα ταχύτητας με τη μορφή near-real time views προωθούνται στο επίπεδο εξυπηρέτησης το οποίο χρησιμοποιεί αυτά τα δεδομένα για την κάλυψη των εκκρεμών αιτημάτων σε επίπεδο ad-hoc.

Στρώμα Ταχύτητας | Stream Layer

Το επίπεδο εξυπηρέτησης ενημερώνεται κάθε φορά που το στρώμα παρτίδας τελειώνει με τον υπολογισμό μιας Batch View. Αυτό σημαίνει ότι τα μόνα δεδομένα που δεν αντιπροσωπεύονται στην Batch View είναι τα δεδομένα που εισήχθησαν ενώ ο υπολογισμός έτρεχε. Το μόνο επομένως που μένει να κάνουμε για να έχουμε ένα σύστημα δεδομένων πραγματικού χρόνου είναι η ενσωμάτωση και αυτών των τελευταίων ωρών δεδομένων. Αυτός ακριβώς είναι ο σκοπός του στρώματος ταχύτητας. Όπως υποδηλώνει το όνομά του, στόχος του είναι να

διασφαλίσει ότι η αντιπροσώπευση των νέων δεδομένων στο ερώτημα λειτουργεί τόσο γρήγορα όσο απαιτείται για τις ανάγκες της εφαρμογής. Μια μεγάλη διαφορά είναι ότι το στρώμα ταχύτητας βλέπει μόνο πρόσφατα δεδομένα, ενώ το στρώμα παρτίδας εξετάζει όλα τα δεδομένα ταυτόχρονα. Μια άλλη μεγάλη διαφορά είναι ότι για να επιτευχθούν οι κατά το δυνατόν λιγότεροι λανθάνοντες χρόνοι, το επίπεδο ταχύτητας δεν βλέπει όλα τα νέα δεδομένα ταυτόχρονα. Αντ 'αυτού, ενημερώνει τις Realtime Views σε πραγματικό χρόνο καθόσον λαμβάνει νέα δεδομένα αντί για τον εκ νέου υπολογισμό από το μηδέν που κάνει το στρώμα παρτίδας.Γενικά λοιπόν, οι ροές δεδομένων που επεξεργάζονται στο batch layer έχουν ως αποτέλεσμα την ενημέρωση της διαδικασίας Δέλτα ή του MapReduce ή του Μοντέλου Μηχανικής Μάθησης που χρησιμοποιούνται περαιτέρω από το stream layer για την επεξεργασία των νέων δεδομένων που εισάγονται σε αυτό. Το stream layer παρέχει τις εξόδους στην βασική διαδικασία εμπλουτισμού και υποστηρίζει το Serving Layer στην μείωση της καθυστέρησης στην απόκριση των ερωτημάτων.

Εφαρμογές Αρχιτεκτονικής Lamda

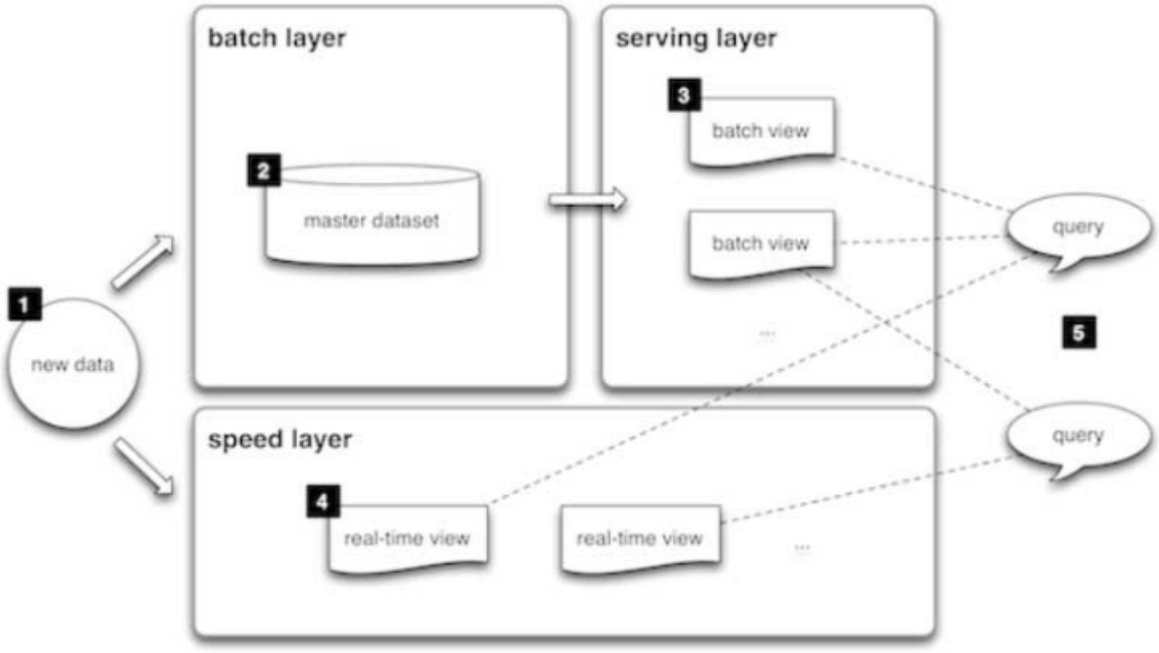
[30. Iman Samizadeh]

Η αρχιτεκτονική Lambda μπορεί να αναπτυχθεί για εκείνα τα επιχειρησιακά μοντέλα επεξεργασίας δεδομένων όπου:

- Τα ερωτήματα των χρηστών πρέπει να εξυπηρετούνται σε ad-hoc επίπεδο χρησιμοποιώντας τον αμετάβλητο χώρο αποθήκευσης δεδομένων.
- Απαιτούνται γρήγορες απαντήσεις και το σύστημα πρέπει να είναι ικανό να χειρίζεται διάφορες ενημερώσεις με τη μορφή νέων ροών δεδομένων.
- Καμία από τις αποθηκευμένες εγγραφές δεν θα διαγραφεί και θα πρέπει να επιτρέπεται την προσθήκη ενημερώσεων και νέων δεδομένων στη βάση

Η αρχιτεκτονική Λάμδα μπορεί να θεωρηθεί ως αρχιτεκτονική επεξεργασίας δεδομένων σχεδόν σε πραγματικό χρόνο. Όπως αναφέρθηκε παραπάνω, μπορεί να αντέξει τα σφάλματα καθώς επίσης επιτρέπει την κλιμάκωση. Χρησιμοποιεί τις λειτουργίες του στρώματος παρτίδας και του στρώματος ροής και συνεχίζει να προσθέτει νέα δεδομένα στον κύριο χώρο αποθήκευσης, διασφαλίζοντας παράλληλα ότι τα υπάρχοντα δεδομένα θα παραμείνουν ανέπαφα. Εταιρείες όπως το Twitter, το Netflix και το Yahoo χρησιμοποιούν αυτήν την αρχιτεκτονική για να ικανοποιήσουν τα πρότυπα ποιότητας των υπηρεσιών που παρέχουν.

Παρακάτω, ένα βασικό διάγραμμα του μοντέλου Lambda Architecture:



Εικόνα 33.Πρώτυπο Αρχιτεκτονικής Λάμδα

Πλεονεκτήματα *[30. Iman Samizadeh]*

- Το στρώμα παρτίδας της αρχιτεκτονικής Lambda διαχειρίζεται τα ιστορικά δεδομένα με την χρήση κατανεμημένη αποθήκευσης ανθεκτικής σε σφάλματα, η οποία εξασφαλίζει χαμηλή πιθανότητα

σφαλμάτων ακόμη και αν το σύστημα καταρρεύσει. Αποτελεί εγγύηση ότι όλες οι αιτήσεις θα λάβουν απάντηση σχετικά με το εάν ήταν επιτυχείς ή όχι.

- Δίνει μια καλή ισορροπία μεταξύ ταχύτητας και αξιοπιστίας.
- Κλιμακούμενη αρχιτεκτονική για την επεξεργασία δεδομένων.Μπορεί να κλιμακωθεί είτε αυτόματα είτε με την ενσωμάτωση επιπλέον χωρητικότητάς.
- Επιχειρηματική Ευελιξία - Αντιδρά σε πραγματικό χρόνο στα μεταβαλλόμενα σενάρια επιχειρήσεων και της αγοράς

Μειονεκτήματα *[30. Iman Samizadeh]*

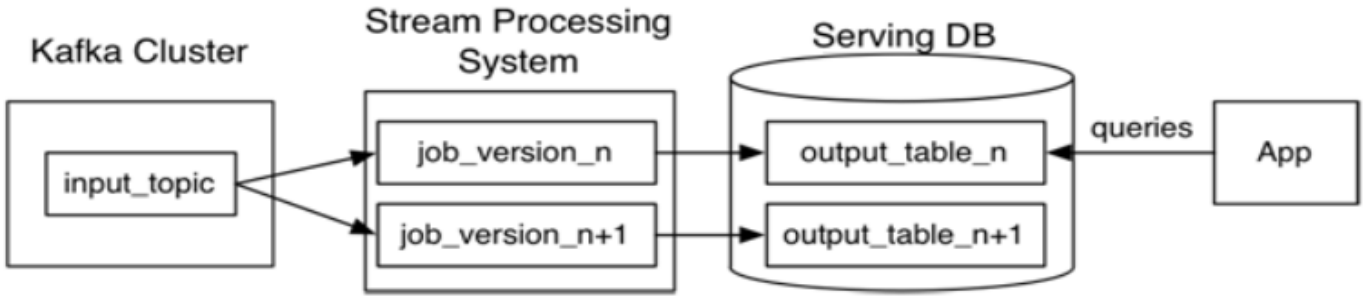
- Μπορεί να οδηγήσει σε υπερπληθώρα κώδικα λόγω της ανάγκης για ολοκληρωμένη επεξεργασία.
- Απαιτεί καθολική επανεπεξεργασία για κάθε κάθε κύκλο του batch γεγονός που δεν είναι και τόσο επωφελές σε ορισμένα σενάρια.
- Τα δεδομένα που έχουν μοντελοποιηθεί με την αρχιτεκτονική Lambda είναι δύσκολο να μετακινηθούν ή να αναδιοργανωθούν σε άλλη αρχιτεκτονική

3.8 Kappa Architecture

[28. Jay Kreps]

Το 2014 ο Jay Kreps ξεκίνησε μια συζήτηση όπου επεσημάνθηκαν κάποιες αδυναμίες σχετικά με την αρχιτεκτονική Lambda και οδήγησε την κοινότητα Big Data σε μια άλλη εναλλακτική αρχιτεκτονική που χρησιμοποιεί λιγότερους πόρους κώδικα και ήταν ικανή να αποδίδει καλά σε μερικά επιχειρηματικά σενάρια όπου η χρήση πολυεπίπεδης αρχιτεκτονικής Lambda έμοιαζε υπερβολή.

Η αρχιτεκτονική Kappa δεν μπορεί να θεωρηθεί ως υποκατάστατο της αρχιτεκτονικής Lambda αντίθετα πρέπει να θεωρηθεί ως μια εναλλακτική λύση που μπορεί να χρησιμοποιηθεί σε εκείνες τις περιπτώσεις όπου η λειτουργία του batch layer δεν είναι απαραίτητη για την επίτευξη της τυπικής ποιότητας υπηρεσιών. Αυτή η αρχιτεκτονική βρίσκει τις εφαρμογές της στην επεξεργασία διακρτιών γεγονότων σε πραγματικό χρόνο. Παρακάτω φαίνεται βασικό διάγραμμα της αρχιτεκτονική Kappa που δείχνει τα δύο επίπεδα λειτουργίας του συστήματος:



Εικόνα 34.Πρώτυπο Αρχιτεκτονικής Καππα

Αν μεταφράστεί η αλληλουχία των λειτουργιών της αρχιτεκτονικής kappa σε μια εξίσωση η οποία καθορίζει οποιοδήποτε ερώτημα στο πεδίο των Big Data.

```
Query = K (New Data) = K (Live streaming data)
```

Η εξίσωση σημαίνει ότι όλα τα ερωτήματα μπορούν να ικανοποιηθούν με την εφαρμογή της λειτουργίας κάππα στις ζωντανές ροές δεδομένων του στρώματος ταχύτητας. Σημαίνει επίσης ότι η επεξεργασία ρεύματος δεδομένων συμβαίνει στο στρώμα ταχύτητας.

Εφαρμογές Αρχιτεκτονικής Kappa

[30. Iman Samizadeh]

Ορισμένες παραλλαγές των εφαρμογών κοινωνικών δικτύων, οι συσκευές που είναι συνδεδεμένες με ένα cloud-based σύστημα παρακολούθησης(monitoring), το Διαδίκτυο των Πραγμάτων (IoT) χρησιμοποιούν παραμετροποιημένες έκδοσεις της αρχιτεκτονικής Lambda, η οποία χρησιμοποιεί κυρίως τις υπηρεσίες του

στρώματος ταχύτητας(speed layer) σε συνδυασμό με το στρώμα συνεχούς ροής(stream layer) για την επεξεργασία των δεδομένων στην Data Lake.

Η αρχιτεκτονική Kappa μπορεί να εφαρμοστεί σε εκείνα τα επιχειρησιακά μοντέλα επεξεργασίας δεδομένων όπου:

- Πολλαπλά γεγονότα στα δεδομένα ή ερωτήματα καταγράφονται σε μια ουρά για να τροφοδοτήσουν ένα κατανεμημένο σύστημα αποθήκευσης αρχείων ή ιστορικό.
- Η σειρά των γεγονότων και των ερωτημάτων δεν είναι προκαθορισμένη. Οι πλατφόρμες επεξεργασίας ροής μπορούν να αλληλεπιδράσουν με τη βάση δεδομένων ανά πάσα στιγμή.
- Παραμένει σταθερό και σε υψηλή διαθεσιμότητα καθώς απαιτείται χειρισμός Terabytes αποθήκευσης από κάθε κόμβο του συστήματος ώστε να υποστηριχθεί η αναπαραγωγή(replication)

Τα παραπάνω σενάρια δεδομένων συνήθως αντιμετωπίζονται με την εντατική χρήση του Apache Kafka, το οποίο είναι εξαιρετικά γρήγορο, ανθεκτικό σε σφάλματα και οριζόντια κλιμακούμενο. Επιτρέπει έναν καλύτερο μηχανισμό για τη διαχείριση των ροών δεδομένων. Ένας ισορροπημένος έλεγχος μεταξύ των επεξεργαστών ροής και των βάσεων δεδομένων καθιστά δυνατή την εκτέλεση των εφαρμογών σύμφωνα με τις προσδοκίες.Το Kafka διατηρεί τα ταξινομημένα δεδομένα για μεγαλύτερο χρονικό διάστημα και φροντίζει ανάλογως τα ερωτήματα συνδέοντάς τα με την κατάλληλη θέση του αρχείου καταγραφής. Το LinkedIn και κάποιες άλλες εφαρμογές χρησιμοποιούν αυτή την θεώρηση στην επεξεργασία των Big Data και αποκομίζουν όφελος από τη διατήρηση μεγάλου όγκου δεδομένων για την κάλυψη των ερωτημάτων που είναι απλώς ισοδύναμα μεταξύ τους.

Πλεονεκτήματα *[30. Iman Samizadeh]*

- Η αρχιτεκτονική Kappa μπορεί να χρησιμοποιηθεί για την ανάπτυξη συστημάτων δεδομένων που απαιτούν απευθείας online ενημέρωση και συνεπώς δεν χρειάζονται το στρώμα παρτίδας.
- Η επανεπεξεργασία απαιτείται μόνο όταν αλλάξει ο κώδικας.
- Μπορεί να αναπτυχθεί διατηρώντας σταθερό εύρως μνήμης.
- Μπορεί να χρησιμοποιηθεί για οριζόντιως κλιμακούμενα συστήματα.
- Απαιτούνται λιγότεροι πόροι αποθήκευσης καθώς η μηχανική μάθηση πραγματοποιείται σε πραγματικό χρόνο.

Μειονεκτήματα *[30. Iman Samizadeh]*

- Η απουσία του στρώματος παρτίδας μπορεί να οδηγήσει σε σφάλματα κατά την επεξεργασία δεδομένων ή κατά την ενημέρωση της βάσης δεδομένων που απαιτούν την ύπαρξη διαχειριστή εξαιρέσεων για την επανεπεξεργασία ή την ευθυγράμμιση.

3.9 Cloud Customer Architecture for Big Data and Analytics v2.0

[22 . Object Management Group]

Η Big Data & Analytics αρχιτεκτονική σε Cloud που δημιουργήθηκε από το **Cloud Standards Customer Council** του **Object Management Group** μπορεί να βοηθήσει τις επιχειρήσεις στην κατανόηση εφαρμοσμένων προτύπων αρχιτεκτονικής που έχουν αναπτυχθεί σε πολλά επιτυχημένα έργα επιχειρησιακής κλίμακας.

Οι εφαρμογές Cloud προσφέρουν μια επιλογή ιδιωτικών, δημόσιων και υβριδικών αρχιτεκτονικών. Το ιδιωτικό cloud περιλαμβάνει εσωτερικά δεδομένα και στοιχεία επεξεργασίας που λειτουργούν πίσω από εταιρικά τείχη προστασίας και δεσμευμένο Cloud. Το δημόσιο cloud προσφέρει υπηρεσίες μέσω Διαδικτύου με δεδομένα και υπολογιστικούς πόρους σε προσβάσιμους διακομιστές. Τα υβριδικά περιβάλλοντα έχουν ένα μείγμα συστατικών που λειτουργούν τόσο ως αποκλειστικές υπηρεσίες cloud όσο και σαν δημόσιο cloud με τα δεδομένα να ρέουν μεταξύ τους.

Υπάρχουν επίσης επιλογές στα επίπεδα των υπηρεσιών που μπορεί να προσφέρει ένας πάροχος cloud για μια λύση Analytics:

Υπηρεσίες Πλατφόρμας Βασικής Υποδομής Δεδομένων - όπως το Hadoop as service - που παρέχει προεγκατεστημένες και διαχειριζόμενες υποδομές. Με αυτό το επίπεδο υπηρεσιών, η επιχείρηση είναι υπευθυνη για τη φόρτωση, διοργάνωση και διαχείριση των δεδομένων και των αναλυτικών στοιχείων.

Μια Ελεγχόμενη Υπηρεσία Διαχείρισης Δεδομένων - όπως μια υπηρεσία Data Lake - που παρέχει διαχείριση δεδομένων, υπηρεσίες καταλόγου, ανάπτυξη analytics, ασφάλεια και υπηρεσίες διαχείρισης πληροφοριών πάνω από μία ή περισσότερες πλατφόρμες δεδομένων. Με αυτό το επίπεδο υπηρεσίας, η επιχείρησης είναι αρμόδια για τον καθορισμός των πολιτικών στον τρόπο διαχείρισης των δεδομένων και για τη σύνδεση πηγών δεδομένων

στην λύση Cloud. Οι κάτοχοι δεδομένων έχουν άμεσο έλεγχο του τρόπου φόρτωσης, ασφάλειας και χρήσης των δεδομένων τους.Οι καταναλωτές δεδομένων μπορούν να χρησιμοποιήσουν τον κατάλογο για να εντοπίσουν τα δεδομένα που θέλουν, να ζητήσουν πρόσβαση και κάνουν χρήση των δεδομένων μέσω διεπαφών αυτονομης εξυπηρέτησης.

Μια Υπηρεσία Παροχής Δεδομένων και Insights- όπως μια υπηρεσία ανάλυσης πελατών(customer analytics service). Με αυτό το επίπεδο υπηρεσίας, η επιχείρηση είναι υπεύθυνη για τη σύνδεση πηγών δεδομένων με την λύση Cloud Analytics. Η λύση παρέχει API για πρόσβαση σε συνδυασμούς των δεδομένων σας και πρόσθετων πηγών δεδομένων, τόσο ιδιόκτητα όσο και ανοιχτά δημόσια δεδομένα, μαζί με αναλυτικές πληροφορίες που προκύπτουν από αυτά τα δεδομένα.

Η χοροθέτηση των δεδομένων και της επεξεργασίας είναι μία από αυτές τις πρώτες αρχιτεκτονικές αποφάσεις σε ένα έργο cloud analytics. Αυτό επιτρέπει την ευελιξία στα μοντέλα λειτουργίας και τη βέλτιστη τοποθέτηση τόσο φόρτου εργασίας δεδομένων όσο και των αναλύσεων στις διαθέσιμες πλατφόρμες επεξεργασίας. Οι νομικές και κανονιστικές απαιτήσεις μπορεί επίσης να επηρεάσουν το που μπορούν να τοποθετηθούν τα δεδομένα καθώς πολλές χώρες έχουν νόμους περί κυριαρχίας δεδομένων που εμποδίζουν δεδομένα σχετικά με φυσικά πρόσωπα, οικονομικές οντότητες και ορισμένους τύπους πνευματικής ιδιοκτησίας να μετακινηθούν πέρα από την επικράτεια της χώρας.

Η επιλογή των αρχιτεκτονικών cloud επιτρέπει στα υπολογιστικά στοιχεία να μετακινούνται κοντά στα δεδομένα με σκοπό τη βελτιστοποίηση της απόδοσης όταν ο όγκος δεδομένων ή / και οι περιορισμοί του εύρους ζώνης οδηγούν σε σημεία συμφόρησης για απομακρυσμένα δεδομένα και την μετακίνηση.

Παρακάτω αναλύονται τα συστατικά της προτεινόμενης αρχιτεκτονικής

Στοιχεία Δημοσίου Δικτύου | Public Network Components

Χρήστες Cloud | Cloud Users

Ένας χρήστης cloud είναι ένα άτομο που συνδέεται με τη λύση cloud analytics μέσω του Διαδικτύου. Αυτό το άτομο μπορεί να ανεβάζει νέα δεδομένα, να αναζητά και να ανακτά δεδομένα, να παρέχει ανατροφοδότηση σχετική με τα δεδομένα, να αιτείται νέες αναλύσεις ή εκτέλεση υφιστάμενων.Ενδεικτικοί ρόλοι τέτοιων χρηστών είναι

- Knowledge Worker and Citizen Analyst
- Data Scientist
- Application Developer
- Data Engineer
- Chief Data Officer (CDO)

Όλα αυτά τα διαφορετικά πρόσωπα έχουν τα ακόλουθα κοινά χαρακτηριστικά:

- Θέλουν Αυτοεξυπηρέτηση. Συχνά ακολουθούν την προσέγγιση «do-it-yourself», με τη δυνατότητα δημιουργίας sandboxes για να δοκιμάζουν νέες υποθέσεις και να μετακινήνούν τα ενεργά πλάνα δράσης στην παραγωγή.
- Θέλουν πρόσβαση στα σωστά δεδομένα για να ολοκληρώσουν τις αναλυτικές εργασίες (αυτό μπορεί να περιλαμβάνει μεγάλο όγκο δεδομένων), ανεξάρτητα από το πού αποθηκεύονται τα δεδομένα, με καλή κατανόηση της ποιότητας και προέλευσης των δεδομένων
- Συχνά χρειάζονται πρόσβαση σε πολλά διαφορετικά εργαλεία και δυνατότητες, πολλά από τα οποία είναι ανοιχτού κώδικα η μπορεί να είναι κατ 'απαίτηση (βάσει φόρτου εργασίας και κλιμάκωσης).
- Τέλος, χρειάζονται συνεργασία αναμεταξύ τους. Ένας τρόπος με τον οποίο οι πάροχοι cloud το επιτρέπουν αυτό είναι με οικοδόμηση μιας γνωσιακής βάσης χρησιμοποιώντας τεχνολογία γράφων και επιτρέποντας στον χρήστη να δει τη συσχέτιση των χρηστών και των δεδομένων, δηλαδή ποιοι και πως χρησιμοποίησαν ποια δεδομένα, προσδιορίζοντας έτσι και τα σωστά άτομα με τα οποία ο χρήστης πρέπει να συνεργαστεί.

Εφαρμογές Software as a Service (SaaS)

Όλο και περισσότερο, οι οργανισμοί κάνουν χρήση των εφαρμογών που προσφέρονται ως υπηρεσία μέσω cloud. Αυτός ο τύπος υπηρεσίας cloud ονομάζεται λογισμικό ως υπηρεσία ή SaaS. Οι τύπου SaaS εφαρμογές στο δημόσιο δίκτυο είναι κυρίως εφαρμογές για κινητά και εφαρμογές ιστού που για αλληλεπίδραση με πελάτες, όπως για παράδειγμα, διαδικτυακές τραπεζικές συναλλαγές, διαδικτυακές αγορές, διαδικτυακές κράτησεις για ταξίδια, εφαρμογές IoT όπως συνδεδεμένα αυτοκίνητα και έξυπνα σπίτια, πρόγνωση καιρού και μέσων κοινωνικής δικτύωσης, κ.λπ.

Πηγές Δεδομένων | Data Sources

Οι δημόσιες πηγές δεδομένων περιέχουν εξωτερικές πηγές δεδομένων για τις λύσεις ανάλυσης δεδομένων που προέρχονται από παρόχους μέσω του Διαδικτύου.Περιλαμβάνουν

- Μηχανές & Αισθητήρες: Δεδομένα που παράγονται από συσκευές, αισθητήρες, δίκτυα και συναφή αυτοματοποιημένα στοιχεία συμπεριλαμβανομένου του Internet of Things (IoT).
- Εικόνα & Βίντεο: Δεδομένα που καταγράφουν οποιαδήποτε μορφή πολυμέσων (εικόνες, βίντεο κ.λπ.) και που μπορούν να σχολιαστούν με ετικέτες, λέξεις-κλειδιά και άλλα μεταδεδομένα.
- Κοινωνικά: Δεδομένα για πληροφορίες, μηνύματα και εικόνες / βίντεο που δημιουργούνται σε εικονικής μορφής κοινότητες και δίκτυα.
- Σύνολα Δεδομένων από το Διαδικτύο: Δεδομένα που αποθηκεύονται σε ιστότοπους, κινητές συσκευές και άλλα συστήματα συνδεδεμένα στο Διαδίκτυο.
- Μετεωρολογικά Δεδομένα
- Δεδομένων Τρίτων (Third Party) Δεδομένα που χρησιμοποιούνται για την αύξηση και την ενίσχυση των υπαρχόντων δεδομένων με νέα χαρακτηριστικά όπως δημογραφικά, γεωχωρικά ή CRM.

Υπηρεσίες Edge | Edge Services

Οι υπηρεσίες Edge περιλαμβάνουν υπηρεσίες που επιτρέπουν στα δεδομένα να ρέουν με ασφάλεια από το Διαδίκτυο προς στο σύστημα επεξεργασίας και ανάλυσης δεδομένων που φιλοξενείται είτε στον πάροχο cloud είτε στην επιχείρηση. Περιλαμβάνουν στοιχεία όπως

- Domain Name System Server (DNS)
- Content Delivery Networks (CDN)
- Firewall
- Load Balancers

Στοιχεία Παρόχου Cloud | Provider Cloud Components

Ο πάροχος Cloud αντιπροσωπεύει τη λύση analytics που φιλοξενείται στο cloud. Διαθέτει στοιχεία για την προετοιμασία δεδομένων για analytics, αποθήκευση δεδομένων, εκτέλεση αναλύσεων και επεξεργασία των αποτελεσμάτων αυτών των συστημάτων.Παρακάτω αναλυόνται τα στοιχεία τα οποία τον συνθέτουν.

Πρόσβαση Δεδομένων και Διαχείριση API | Data Access and API Management

Ο γενικός σκοπός του στοιχείου Πρόσβασης Δεδομένων είναι να εκφράσει τις διάφορες δυνατότητες που απαιτούνται για την αλληλεπίδραση με το στοιχείο **Data Repositories**. Οι δυνατότητες εξυπηρετούν τις ανάγκες πρόσβασης των επιστημόνων δεδομένων, επιχειρησιακών αναλυτών, προγραμματιστών και όλους τους υπόλοιπους που χρειάζονται πρόσβαση σε πολύτιμα δεδομένα.Οι υπηρεσίες που επιτελεί είναι:

- Data Access
- Data Virtualization
- Data Federation
- Open APIs

Υπολογιστική Ρευμάτων Πραγματικού Χρόνου | Streaming Computing

Τα συστήματα επεξεργασίας πραγματικού χρόνου (streams) μπορούν να απορροφήσουν και να επεξεργαστούν μεγάλους όγκους εξαιρετικά δυναμικών, ευαίσθητων στο χρόνο συνεχών ροών δεδομένων από μια ποικιλία εισόδων, όπως συσκευές παρακολούθησης που βασίζονται σε αισθητήρες, συστήματα μηνυμάτων και

τροφοδοσίες χρηματοοικονομικής αγοράς. Το μοντέλο “store-and-pull” των παραδοσιακών περιβάλλοντων επεξεργασίας δεδομένων δεν είναι κατάλληλο για αυτήν την κατηγορία εφαρμογών ροής χαμηλού λανθάνοντος χρόνου ή πραγματικού χρόνου όπου δεδομένα πρέπει να υποβληθούν σε επεξεργασία εν κινήσει. Οι δυνατότητες περιλαμβάνουν:

- Streaming Analytics
- Complex Event Processing (CEP)
- Data Enrichment
- Real Time Ingestion

Γνωσιακά Υποβοηθούμενη Διασύνδεση Δεδομένων | Cognitive Assisted Data Integration

Το στοιχείο Cognitive Assisted Data Integration εστιάζει στις διαδικασίες και τα περιβάλλοντα που ασχολούνται με τη συλλογή, την πιστοποίηση, την επεξεργασία και τη μετακίνηση δεδομένων προκειμένου να προετοιμάστούν για αποθήκευση στα αποθετήρια αναλυτικών Data Lakes, τα οποία στη συνέχεια κοινοποιούνται στα στοιχεία Discovery & Exploration και των Actionable Insights, μέσω του στοιχείου Data Access. Το στοιχείο Data Ingestion & Integration μπορεί να επεξεργαστεί δεδομένα σε προγραμματισμένα διαστήματα παρτίδας ή σε διαστήματα σχεδόν πραγματικού χρόνου "just-in-time", ανάλογα με τη φύση των δεδομένων και τον επιχειρηματικό σκοπό της χρήση τους. Διάφορες γνωστικές τεχνολογίες όπως η μηχανική μάθηση και η επεξεργασία φυσικής γλώσσας μπορούν να αξιοποιηθούν για την ημι-αυτοματοποίηση της διαδικασίας συλλογής και διασύνδεσης των δεδομένων .

- Batch Ingestion
- Change Data Capture
- Document Interpretation & Classification
- Data Quality Analysis

Αποθετήρια Δεδομένων | Data Repositories

Το στοιχείο Data Repositories είναι ένα σύνολο ασφαλών αποθετηρίων δεδομένων που επιτρέπει την αποθήκευση για κατανάλωση από εργαλεία ανάλυσης και χρήστες. Αυτά τα αποθετήρια αποτελούν την καρδιά του περιβάλλοντος εφαρμογής των Analytics. Μπορεί να διαφέρουν από ένα και μοναδικό αποθετήριο Hadoop ή Enterprise Data Warehouse, σε πολλά αποθετήρια που χρησιμοποιούνται για διαφορετικούς σκοπούς από διαφορετικά αναλυτικά εργαλεία. Στο στοιχείο αυτό δεν περιλαμβάνεται η αποθήκευση λειτουργικών και συναλλακτικών δεδομένων(όπως OLTP, ECM,ERP κ.λπ.) καθώς κατά βάση εντάσσονται στο στοιχείο των Data Sources.

Οι τύποι αποθετηρίων δεδομένων περιλαμβάνουν:

- Landing Zone & Data Archive
- History
- Deep & Exploratory Analytics
- Sand Boxes
- Data Warehouses & Data Marts
- Predictive Analytics

Γνωσιακή Αναλυτική Εξευρεύνηση & Ανακάλυψη | Cognitive Analytics Discovery & Exploration

Ο γενικός σκοπός του στοιχείου Discovery & Exploration επικεντρώνεται στην ενεργοποίηση μιας νέας (και παλιάς) κλάσης πελατών δεδομένων. Επιστήμονες δεδομένων,οι επιχειρηματικοί αναλυτές, οι μηχανικοί δεδομένων και οι προγραμματιστές εφαρμογών πρέπει να μπορούν να βρούν την κρυμμένη αξία στα δεδομένα γρήγορα. Έτσι, πρέπει να έχουν την ικανότητα να συνεργάζονται και αλληλεπιδρούν εύκολα ακόμη και με τα πιο περίπλοκα αποθετήρια δεδομένων μέσω νέων και αναδυόμενων τεχνικών της επιστήμης δεδομένων. Χρειάζονται επίσης το δυνατότητα σημασιολογικής αναζήτησης τόσο σε δομημένο όσο και σε μη δομημένο περιεχόμενο για να έχουν την πλήρη εικόνα της οντολογίας δεδομένων.Σε αυτό το στοιχείο εντάσσονται οι εξής λειτουργίες:

- Data Science
- Search and survey/shopping for data

Εφαρμόσιμες Γνωσιακές Προβλέψεις | Cognitive Actionable Insight

Ο συνολικός σκοπός του στοιχείου Actionable Insight είναι η ανάλυση δεδομένων από διάφορες πηγές με συνεπή και δομημένο τρόπο και η αντλήση πληροφοριών που είναι ουσιαστικές και εφαρμόσιμες για τον επιχειρηματικό τομέα. Μια ποικιλία τεχνικών χρησιμοποιούνται για να αντλήσουν αυτήν την πολύτιμη εικόνα: οπτικοποίηση των δεδομένων, ευφυής αναζήτηση σε πολυδιάστατα δεδομένα, χρήση στατιστικών μοντέλων, εξόρυξη δεδομένων, ανάλυση περιεχομένου, βελτιστοποίηση και γνωσιακές λειτουργίες. Οι γνωσιακές τεχνολογίες μπορούν να εφαρμοστούν εδώ για να επιλέξουν αυτόματα το σωστό μοντέλο ανάλυσης, να αλληλεπιδρούν με χρήστες με πιο φιλικό προς τον άνθρωπο τρόπο. Λειτουργίες

- Visualization & Storyboarding
- Reporting, Analysis & Content Analytics
- Decision Management
- Predictive Analytics & Modeling
- Cognitive Analytics
- Insight as a Service

Εφαρμογές Software as a Service |SaaS Applications

Η χρήση στο επίπεδο του δικτύου παρόχου αυξάνεται διαρκώς για τις επιχειρήσεις.Συνηθέστερες λειτουργίες:

- Customer Experience
- New Business Models
- Financial Performance
- Risk
- Fraud & Preparations
- IT Economics

Μετασχηματισμός και Συνδεσιμότητα | Transformation and Connectivity

Το στοιχείο μετασχηματισμού και συνδεσιμότητας επιτρέπει ασφαλείς συνδέσεις με εταιρικά συστήματα με τη δυνατότητα φιλτραρίσματος, συγκέντρωσης, τροποποίησης ή μορφοποίησης δεδομένων ανάλογα με τις ανάγκες. Ο μετασχηματισμός δεδομένων συχνά απαιτείται όταν τα δεδομένα δεν ταιριάζουν σε εταιρικές εφαρμογές. Οι βασικές δυνατότητες περιλαμβάνουν:

- Enterprise Security Connectivity
- Transformations
- Enterprise Data Connectivity

Επιχειρηματικό Δίκτυο | Enterprise Network

Το εταιρικό δίκτυο είναι το σημείο όπου βρίσκονται τα εσωτερικά συστήματα και οι χρήστες της επιχείρησης

Εταιρικοί Χρήστες

Οι εταιρικοί χρήστες είναι άτομα που ενεργούν τόσο ως χρήστες cloud όπως αυτοί συζητήθηκαν παραπάνω, όσο και ως εξειδικευμένοι χρήστες που συνδέονται με τη λύση cloud analytics μέσω του εσωτερικού δικτύου του οργανισμού. Οι εταιρικοί χρήστες μπορούν επίσης να είναι χρήστες με διαχειριστικό ρόλο, ρυθμίζοντας το αναλυτικό σύστημα επεξεργασίας και παρακολουθόντας την απόδοση και της διαθεσιμότητα της λύσης συνολικά

Επιχειρησιακές Εφαρμογές

Οι εταιρικές εφαρμογές είναι βασικές πηγές δεδομένων σε μια λύση Analytics. Μπορούν επίσης να γίνουν ο προορισμός για νέες πληροφορίες, ή μπορεί να λειτουργήσουν ως πλατφόρμα ανάπτυξης για αναλυτικά μοντέλα πραγματικού χρόνου που αναπτύχθηκαν στην Data Lake και μπορούν επίσης να παρέχουν πληροφορίες στις εφαρμογές SaaS.Τα υποστοιχεία ταυτίζονται με αυτά που έχουν αναλυθεί στην παράγραφο Εφαρμογές SaaS

Επιχειρησιακά Δεδομένα

Στα εταιρικά δίκτυα, οι επιχειρήσεις συνήθως φιλοξενούν έναν αριθμό εφαρμογών που παρέχουν κρίσιμες επιχειρηματικές λύσεις μαζί με την ανάλογη υποστήριξη υποδομών όπως αποθήκευση δεδομένων. Τέτοιες εφαρμογές είναι βασικές πηγές για δεδομένα που μπορούν να εξαχθούν και να διασυνδεθούν με υπηρεσίες που παρέχονται από τη λύση Cloud Analytics.

Τα εταιρικά δεδομένα περιλαμβάνουν μεταδεδομένα σχετικά με τα δεδομένα καθώς και κεντρικές καταχωρίσεις για επιχειρησιακές εφαρμογές. Τα εταιρικά δεδομένα μπορούν να ρέουν απευθείας προς ενοποίηση ή στα αποθετήρια δεδομένων παρέχοντας ανατροφοδότηση στο αναλυτικό σύστημα.Περιλαμβάνουν:

- Reference Data
- Master Data
- Transactional Data
- Application Data
- Log Data
- Enterprise Content Data
- Historical Data
- Archived Data

Κατάλογος Επιχειρησιακών Χρηστών

Ο κατάλογος εταιρικών χρηστών περιέχει τα προφίλ χρηστών τόσο για τους χρήστες cloud όσο και για τους επιχειρησιακούς χρήστες. Ένα προφίλ χρήστη παρέχει έναν λογαριασμό σύνδεσης και παραθέτει τους πόρους (σύνολα δεδομένων, API και άλλες υπηρεσίες) που το άτομο έχει δικαίωμα πρόσβασης. Οι υπηρεσίες ασφαλείας και οι edge υπηρεσίες τον χρησιμοποιούν ως οδηγό για πρόσβαση στο εταιρικό δίκτυο, τις εταιρικές υπηρεσίες ή τις εξιδεικευμένες επιχειρησιακές υπηρεσίες cloud.

Καθολικά Στοιχεία Περιβάλλοντος | Cross All Environment Components

Διακυβέρνηση Πληροφορίας

Τα στοιχεία Διαχείρισης και Διακυβέρνησης Πληροφορίας σας βοηθούν να δημιουργήσετε εμπιστοσύνη στα δεδομένα σας διατηρώντας μια αξιόπιστη, ακριβή προβολή κρίσιμων επιχειρηματικών δεδομένων, παρέχοντας μια τυποποιημένη προσέγγιση στο να ανακαλύπτετε τα στοιχεία πληροφορικής σας και να ορίσετε μια κοινή επιχειρηματική γλώσσα. Το αποτέλεσμα είναι καλύτερη και γρηγορότερη λήψη αποφάσεων που οδηγεί σε λειτουργική αποδοτικότητα και ανταγωνιστικό πλεονέκτημα.

- Data Lifecycle Management
- Master & Entity Data
- Reference Data
- Data Catalog
- Data Models
- Data Quality Rules

Ασφάλεια

Το στοιχείο Ασφαλείας είναι κρίσιμο σε όλες τις αρχιτεκτονικές δεδομένων και τα πλάνα. Με έμφαση στην προστασία των δεδομένων υπάρχουν συγκεκριμένα χαρακτηριστικά που πρέπει να δοθεί σημασία. Αυτά είναι οι δυνατότητες κάλυψης/απόκρυψης δεδομένων σε χαμηλό επίπεδο για εκείνους που πρέπει να συνεχίσουν αλληλεπιδράσουν με αυτό, τη δυνατότητα κρυπτογράφησης των δεδομένων από όλους χρήστες, τη δυνατότητα να γνωρίζουν ποιος έχει πρόσβαση σε αυτα και γιατί, και τη δυνατότητα να έχουν μια συνολική εικόνα όλων αυτών των δραστηριοτήτων.Υποστοιχεία:

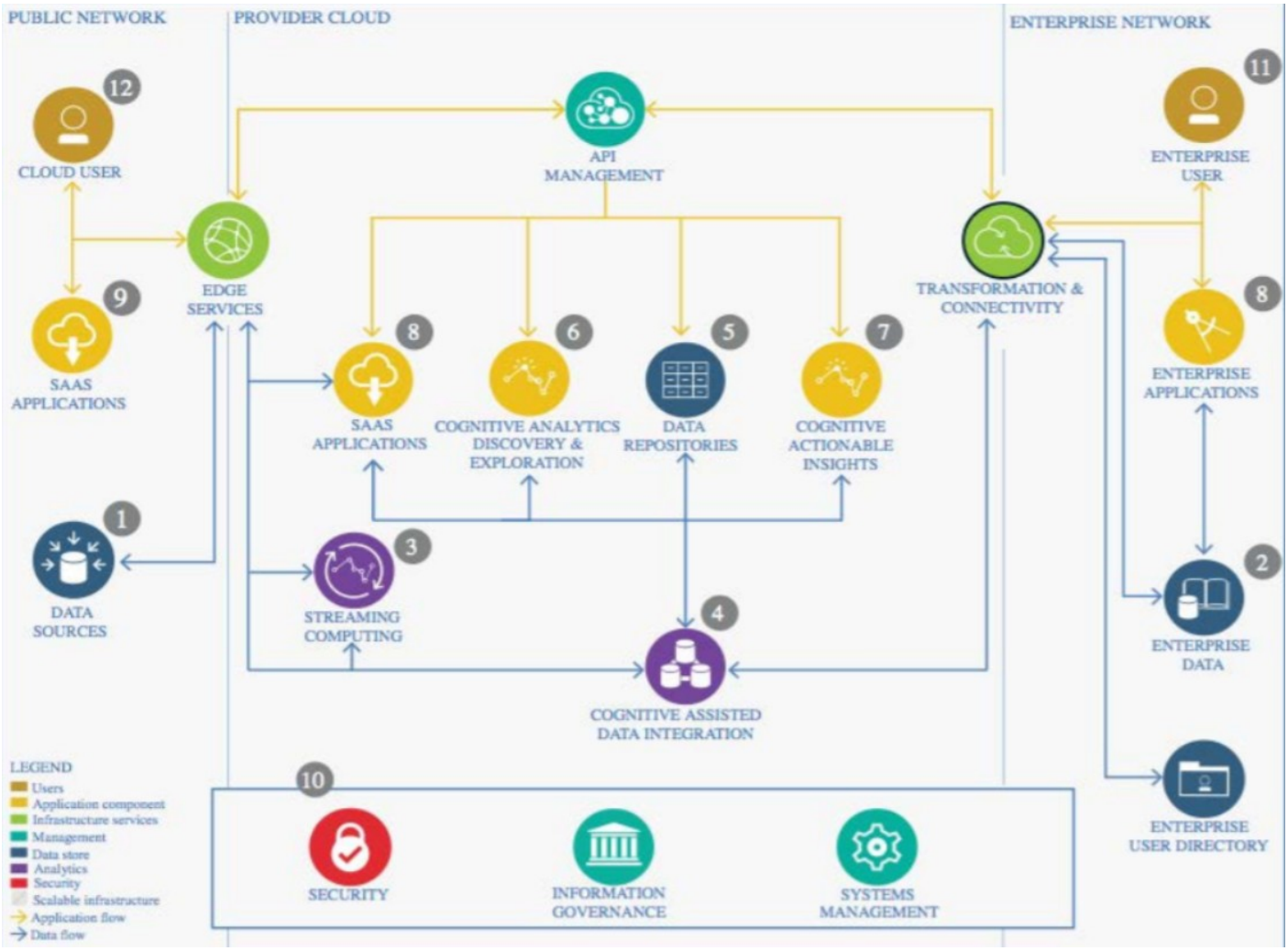
- Data Security
- Identity & Access Management
- Infrastructure Security
- Application Security
- Secure DevOps

- Security Monitoring & Intelligence
- Security Governance

Διαχείριση Συστήματος

Περιλαμβάνει την διαχείριση και οι λειτουργίες της υπηρεσίας Cloud αναφέρονται σε όλες τις δραστηριότητες που εκτελούνται από έναν οργανισμό για τον σχεδιασμό, την παράδοση, τη λειτουργία και τον έλεγχο των υπηρεσιών πληροφορικής και cloud που προσφέρονται στους πελάτες. Οι Συμφωνίες Επιπέδου Υπηρεσιών(SLA-Service Level Agrrements) του παρόχου cloud ενδέχεται να καλύπτουν όλες τις λεπτομέρειες.

Παρακάτω ακολουθεί μια γραφική αναπαράσταση όλων των συστατικών στοιχείων της προτεινόμενης αρχιτεκτονικής και των αντίσοιχων ροών:



Εικόνα 35.Cloud Customer Architecture for Big Data and Analytics

Ροές Εκτέλεσης

1. Δεδομένα από προφίλ πελατών (δημογραφικά στοιχεία, προτιμήσεις, υπηρεσίες παραγωγής, ιστορική συμπεριφορά κτλπ) και τα δεδομένα αλληλεπιδράσεων από εταιρικές βάσεις δεδομένων συλλέγονται σε διαδικασίες batch η/και πραγματικού χρόνου. Δημόσια Δεδομένα,όπως δεδομένα κοινωνικών μέσων (Facebook & Twitter) συλλέγονται επίσης.

2 & 3. Τα επιχειρηματικά και δημόσια δεδομένα διασυνδέονται και μετασχηματίζονται χρησιμοποιώντας υπηρεσίες πληροφορικής (κυρίως Apache Spark)

4. Μετασχηματισμένα δεδομένα, τα οποία είναι ένας συνδυασμός δομημένων δεδομένων πελατών από εταιρικές πηγές και τα δεδομένα αλληλεπίδρασης που μπορεί να ποικίλουν στη δομή για μια συγεκριμένη χρονική περίοδο, αποθηκεύονται σε ένα JSON αποθετήριο δεδομένων στην Data Lake. Τα προσωρινά αποθηκευμένα δεδομένα για αιτήματα API σε πραγματικό χρόνο αποθηκεύονται σε in-memory βάση δεδομένων.

5. Η εκτέλεση analytics βασισμένη σε κανόνες για την ανάλυση της προσωπικής εμπειρίας του πελάτη που βρίσκεται στο υποκατάστημα ή χρησιμοποιεί την εφαρμογή για κινητά. Οι αλλαγές παρουσιάζονται στο προσωπικό του καταστήματος ή στο κινητό του πελάτη με σκοπό τη βελτίωση της εμπειρίας του πελάτη σε πραγματικό χρόνο.

6. Στο περιβάλλον έρευνας και ανάπτυξης, οι επιστήμονες δεδομένων ανακαλύπτουν και διερευνούν πρότυπα. Οι περιπτώσεις χρήσεις που συνεχώς εξέλισσονται περιλαμβάνουν πληροφορίες πελατών, προσωπική εμπειρία πελάτη, διευκόλυνση αξιολόγησης σχετικά με την πίστωση, ιδιωτική τραπεζική εμπειρία και αλληπέδρσαση μέσω λειτουργικών εφαρμογών για κινητά.

7. Τα API διαβάζουν δεδομένα μέσω αιτημάτων πραγματικού χρόνου που συνήθως προέρχονται από κατά τόπους καταστήματα.

8. Μια cloud πλατφόρμα μάρκετινγκ χρησιμοποιείται για όλες τις αλληλεπιδράσεις με τον πελάτη σε όλα των διαφορετικών κανάλια, ειδικά σε ότι αφορά την επόμενη καλύτερη δράση ή την επόμενη καλύτερη προσφορά. Για παράδειγμα, η εφαρμογή για κινητά λαμβάνει υπόψη είσοδο από μέσα κοινωνικής δικτύωσης, εντοπισμού και κινητής τηλεφωνίας για να αποφασίσει την καλύτερη ενέργεια σε πραγματικό χρόνο, χρησιμοποιώντας τη δύναμη των διαθέσιμων πληροφοριών για την κατά το δυνατόν καλύτερη προσέγγιση με τους πελάτες. Αυτό μπορεί να οδηγήσει σε μια καλύτερη εμπειρία πελάτη και βοηθήσει να εντοπιστούν όμαδες μέ όμοια χαρακτηριστικά μεταξύ των προφίλ πελάτων με γνώμωνα το ατομικό τους περιβάλλον, τη συμπεριφορά και τις προτιμήσεις τους.

9. Τα γνωσιακά API χρησιμοποιούνται επί του παρόντος για ερωτήσεις και απαντήσεις(Q&A) για τους πράκτορες μάρκετινγκ πιστωτικών καρτών.

4. Εργαλεία και Ολοκληρωμένες Λύσεις Λογισμικού

Η παρακάτω ταξινόμηση-κατάταξη αποτελεί διασκευή αυτής που αρχικά προτάθηκε από τον *Bas Geerdink [38]*

4.1 Μηχανές Εισαγωγής & Διασύνδεσης Δεδομένων -Importing & Integration Engines-

Εικόνα 36.Εμπορικά Εργαλεία Εισαγωγής και Integration

Apache Chukwa	
Ιστοθέση	http://chukwa.apache.org
Τύπος	Συλλογή Δεδομένων
	**Περιγραφή**
	Το Apache Chukwa είναι ένα σύστημα συλλογής δεδομένων ανοιχτού κώδικα για την παρακολούθηση μεγάλων κατανεμημένων συστημάτων. Είναι χτισμένο πάνω από το Hadoop Distributed File System (HDFS) και το Map / Reduce πλαίσιο και κληρονομεί την επεκτασιμότητα και την ευρωστία του Hadoop.Περιλαμβάνει επίσης μια ευέλικτη και ισχυρή εργαλειοθήκη για την εμφάνιση, παρακολούθηση και ανάλυση αποτελεσμάτων με σκοπό την καλύτερη δυνατή χρήση των συλλεγόμενων δεδομένων.

IBM InfoSphere Data Explorer	
Ιστοθέση	https://www.ibm.com/support/knowledgecenter/beta/en/SS8NLW_9.0.0/dataexplorer_9.0.0.html
Τύπος	Ανακάλυψη/Εξόρυξη Δεδομένων
	Περιγραφή
	Καθολικός τρόπος πλοήγησης και ανακάλυψης σε ένα ευρύ φάσμα εφαρμογών, πηγών δεδομένων και τύπων αρχείων

MuleSoft Anypoint	
Ιστοθέση	https://www.mulesoft.com/platform/enterprise-integration
Τύπος	Διασύνδεση Δεδομένων (Data Integration)
	**Περιγραφή**
	Εργαλείο διασύνδεσης τύπου Cloud SaaS-Software as a Service

Pentaho Data Integration (Kettle)	
Ιστοθέση	https://help.pentaho.com/Documentation/8.2/Products/Data_Integration
Τύπος	Διασύνδεση Δεδομένων
	**Περιγραφή**
	Ολοκληρωμένη Πλατφόρμα Data Integration

Rapid Miner	
Ιστοθέση	https://rapidminer.com
Τύπος	Εξαγωγή Δεδομένων
	**Περιγραφή**
	Εργαλείο για εξόρυξη δεδομένων, απόκτηση δεδομένων, ETL και ανάλυση δεδομένων

Splunk	
Ιστοθέση	https://www.splunk.com
Τύπος	Συλλογή Δεδομένων
	**Περιγραφή**
	Συλλογή και ευρετηρίαση δεδομένων που παράγονται απο μηχανές και υπολογιστικά συστήματα

Talend Open Studio	
Ιστοθέση	https://www.talend.com/products/big-data/big-data-open-studio/
Τύπος	Συλλογή, Φόρτωση, Απόκτηση Δεδομένων
	**Περιγραφή**
	Φόρτωση, εξαγωγή, μετασχηματισμός και επεξεργασία μεγάλων και διαφορετικών συνόλων δεδομένων

CloverDx	
Ιστοθέση	https://www.cloverdx.com/product
Τύπος	Πλατφόρμα Διαχείρισης Δεδομένων
	**Περιγραφή**
	Σχεδιασμός, εντοπισμός σφαλμάτων, εκτέλεση και αντιμετώπιση μετασχηματισμών δεδομένων και ροών εργασιών. Ενορχηστρώση φόρτωση εργασίας δεδομένων. Ανάπτυξη του φόρτου εργασίας δεδομένων σε ένα ισχυρό επιχειρησιακό περιβάλλον. Σε μορφή cloud ή on-premise.Διάθεση των δεδομένων στους ανθρώπους,τις εφαρμογές και αποθηκευτικούς χώρους σε μία ενιαία πλατφόρμα.

StreamSets	
Ιστοθέση	https://streamsets.com
Τύπος	Ολοκληρωμένη Πλατφόρμα DataOps
	**Περιγραφή**
	Συλλογή και μετασχηματισμός δεδομένων

MQTT	
Ιστοθέση	http://mqtt.org
Τύπος	IoT
	**Περιγραφή**
	Το MQTT σημαίνει MQ Telemetry Transport. Πρόκειται για ένα πρωτόκολλο δημοσίευσης / εγγραφής, εξαιρετικά απλό και ελαφρών μηνυμάτων, σχεδιασμένο για περιορισμένες συσκευές και χαμηλό

	εύρος ζώνης, υψηλού λανθάνοντος χρόνου ή για αναξιόπιστα δίκτυα. Οι αρχές σχεδιασμού είναι να ελαχιστοποιήσουν το εύρος ζώνης δικτύου και τις απαιτήσεις πόρων των συσκευών, ενώ προσπαθούν επίσης να διασφαλίσουν την αξιοπιστία και κάποιο βαθμό διασφάλισης της παράδοσης

Eclipse Kura™	
Ιστοθέση	https://www.eclipse.org/kura/
Τύπος	IoT
	**Περιγραφή**
	Εκτάσιμο ανοιχτού κώδικα IoT Edge Framework που βασίζεται σε Java / OSGi. Το Kura προσφέρει πρόσβαση API στις διεπαφές υλικού των IoT Gateways (σειριακές θύρες, GPS, φύλακας, GPIO, I2C κ.λπ.). Διαθέτει έτοιμα προς χρήση πρωτόκολλα πεδίου (συμπεριλαμβανομένων των Modbus, OPC-UA, S7), ένα application contatiner και μία διεπαφή χρήστη μέσω διαδικτύου για προγραμματισμό ροών για την απόκτηση δεδομένων από τα πεδία, την επεξεργασία του στις τερματικές συσκευές (edge) και τη δημοσίευσή τους στις κορυφαίες πλατφόρμες Cloud IoT μέσω συνδεσιμότητας MQTT.

Apache Edgent	
Ιστοθέση	https://edgent.incubator.apache.org
Τύπος	IoT
	**Περιγραφή**
	Το Apache Edgent είναι ένα μοντέλο προγραμματισμού και runtime kernel που μπορεί να ενσωματωθεί σε πύλες και συσκευές μικρού εκτοπίσματος που επιτρέπουν τοπικές, σε πραγματικό χρόνο, αναλύσεις σχετικά με τις συνεχείς ροές δεδομένων που προέρχονται από εξοπλισμό, οχήματα, συστήματα, συσκευές και αισθητήρες όλων των ειδών (για παράδειγμα, Raspberry Pis ή έξυπνα τηλέφωνα). Σε συνεργασία με κεντρικά συστήματα ανάλυσης, το Apache Edgent παρέχει αποτελεσματικές και έγκαιρες αναλύσεις σε ολόκληρο το οικοσύστημα IoT

4.2 Μηχανές Επεξεργασίας | Processing Engines

4.2.1 Προετοιμασία & Καθαρισμός Δεδομένων | Data Preparation & Cleaning

Apache Avro	
Ιστοθέση	https://avro.apache.org
Τύπος	Data Serialization
	**Περιγραφή**
	Πλούσιες δομές δεδομένων. Μια συμπαγής, γρήγορη, δυαδική μορφή δεδομένων. Ένα αρχείο container, για την αποθήκευση μόνιμων δεδομένων. Κλήση απομακρυσμένης διαδικασίας (RPC). Απλή ενσωμάτωση με δυναμικές γλώσσες. Δεν απαιτείται δημιουργία κώδικα για την ανάγνωση ή εγγραφή αρχείων δεδομένων ούτε για τη χρήση ή την εφαρμογή πρωτοκόλλων RPC.

Apache Sqoop	
Ιστοθέση	https://sqoop.apache.org
Τύπος	Data Transportation / Μεταφορά Δεδομένων
	**Περιγραφή**
	Το Apache Sqoop (TM) είναι ένα εργαλείο που έχει σχεδιαστεί για την αποτελεσματική μεταφορά μαζικών δεδομένων μεταξύ Apache Hadoop και δομημένων βάσεων δεδομένων, όπως σχεσιακές βάσεις δεδομένων.

DataCleaner	
Ιστοθέση	https://datacleaner.org
Τύπος	Data Cleaning / Καθαρισμός Δεδομένων
	**Περιγραφή**
	Εφαρμογή Ανάλυσης Ποιότητας Δεδομένων

OpenRefine	
Ιστοθέση	https://openrefine.org
Τύπος	Data Cleaning / Καθαρισμός Δεδομένων
	**Περιγραφή**
	Εργαλείο για τον χειρισμό ακατέργαστων δεδομένων, τον καθαρισμό, τη μετατροπή του από τη μία μορφή σε άλλη, την επέκτασή τους με υπηρεσίες ιστού και τη σύνδεσή τους με βάσεις δεδομένων

4.2.2 Πλοήγηση στα Δεδομένα | Data Exploration Engines

4.2.2.1 Αναζήτηση

Εικόνα 37.Εμπορικά Εργαλεία Αναζήτησης

Apache Lucene	
Ιστοθέση	https://lucene.apache.org
Τύπος	Αναζήτηση
	Περιγραφή
	Εργαλείο με τεχνολογία αναζήτησης και ευρετηρίασης , καθώς και ορθογραφικό έλεγχο, επισήμανση επιτυχίας και προηγμένες δυνατότητες ανάλυσης / διακριτικοποίησης

Apache Nutch	
Ιστοθέση	http://nutch.apache.org
Τύπος	Web Crawler / Ανίχνευση Ιστου
	Περιγραφή
	Επεκτάσιμο και κλιμακούμενο λογισμικό ανίχνευσης ιστού, βασισμένο στο Lucene

Apache Solr	
Ιστοθέση	https://lucene.apache.org/solr/
Τύπος	Search Server/Διακομιστής Αναζήτησης
	Περιγραφή
	Διακομιστής αναζήτησης υψηλής απόδοσης, κατασκευασμένος με χρήση του Lucene

Sphinx	
Ιστοθέση	http://sphinxsearch.com
Τύπος	Search Server/Διακομιστής Αναζήτησης
	**Περιγραφή**
	Διακομιστής Αναζήτησης Γενικής Χρήσης

Xapian	
Ιστοθέση	https://xapian.org
Τύπος	Search Server/Διακομιστής Αναζήτησης
	**Περιγραφή**
	Βιβλιοθήκη ανάκτησης πληροφοριών βασισμένη στις πιθανότητες και πλήρης μηχανή αναζήτησης κειμένου

4.2.2.2 Quering | Υποβολή Ερωτημάτων

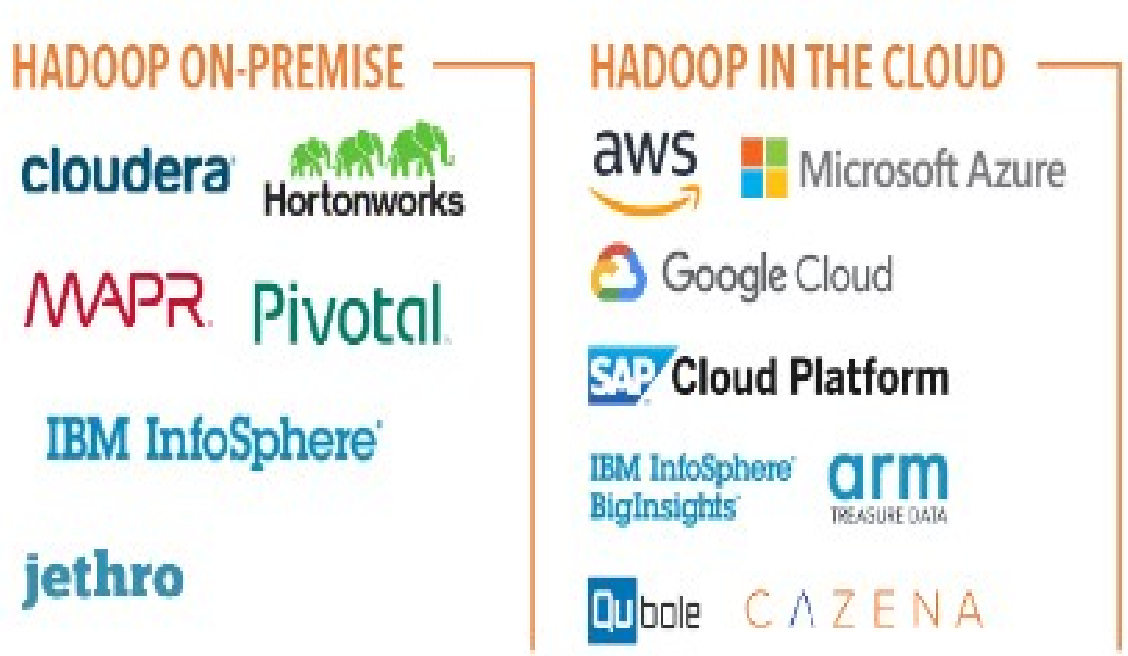
Apache Pig	
Ιστοθέση	https://pig.apache.org
Τύπος	Πλατφόρμα Εξερεύνησης
	**Περιγραφή**
	Πλατφόρμα για την ανάλυση μεγάλων συνόλων δεδομένων που αποτελείται από μια γλώσσα υψηλού επιπέδου για την σύνθεση προγραμμάτων ανάλυσης δεδομένων, σε συνδυασμό με υποδομή για την αξιολόγηση αυτών των προγραμμάτων. Η εμφανής ιδιότητα των προγραμμάτων Pig είναι ότι η δομή τους επιδέχεται ουσιαστικής παραλληλοποίησης, η οποία με τη σειρά της τους επιτρέπει να χειρίζονται πολύ μεγάλα σύνολα δεδομένων.

Apache Hive	
Ιστοθέση	https://hive.apache.org
Τύπος	Μηχανή DataWarehouse
	**Περιγραφή**
	Λογισμικό αποθήκης δεδομένων (data warehouse) που διευκολύνει την ανάγνωση, τη γραφή και τη διαχείριση μεγάλων συνόλων δεδομένων που βρίσκονται σε κατανεμημένο χώρο αποθήκευσης χρησιμοποιώντας SQL. Η δομή μπορεί να προβάλλεται σε δεδομένα που είναι ήδη αποθηκευμένα. Ένα εργαλείο γραμμής εντολών και ένα πρόγραμμα οδήγησης JDBC παρέχονται για τη σύνδεση των χρηστών

Apache Impala	
Ιστοθέση	https://impala.apache.org/index.html
Τύπος	Αναλυτική Βάση Δεδομένων
	**Περιγραφή**
	Το Apache Impala είναι ανοιχτού κώδικα, εγγενής αναλυτική βάση δεδομένων για το Apache Hadoop. Το Impala παρέχει χαμηλό λανθάνοντα χρόνο και υψηλή ταχύτητα συγχρονισμού για BI / Αναλυτικά Queries στο Hadoop (δεν παραδίδεται από πλαίσια επεξεργασίας batch όπως το Apache Hive). Το Impala κλιμακώνεται επίσης γραμμικά, ακόμη και σε πολυπαραγοντικά περιβάλλοντα.

Apache Drill	
Ιστοθέση	https://drill.apache.org
Τύπος	Μηχανή SQL
	**Περιγραφή**
	Το Drill είναι μια ανοιχτού κώδικα μηχανή ερωτημάτων SQL για εξερεύνηση Big Data. Το Drill έχει σχεδιαστεί εξολοκλήρου για να υποστηρίξει ανάλυση υψηλής απόδοσης σε ημι-δομημένα και ταχέως εξελισσόμενα δεδομένα που προέρχονται από σύγχρονες εφαρμογές Big Data, παρέχοντας παράλληλα την εξοικείωση και το οικοσύστημα του ANSI SQL, της πρότυπης γλώσσας ερωτημάτων. Το Drill παρέχει επίσης διασύνδεση με υπάρχουσες εφαρμογές όπως Apache Hive και Apache HBase.

4.2.3 Μηχανές Επεξεργασίας Batch/Hadoop



Εικόνα 38.Εμπορικές Πλατφόρμες Hadoop

IBM InfoSphere BigInsights

Ιστοθέση:https://www.ibm.com/support/knowledgecenter/SSPT3X_3.0.0/com.ibm.swg.im.infosphere.biginsights.product.doc/doc/bi_features_architecture.html

Γενική πλατφόρμα big data, συμπεριλαμβανομένης της Hadoop InfoChimps Cloud Hadoop Solution Proprietary Suite για ισχυρές, επεκτάσιμες υπηρεσίες μεγάλων δεδομένων που βασίζονται σε cloud

MapR

Ιστοθέση:https://mapr.com/products/

Διανομή Hadoop και συναφών έργων Microsoft Windows Azure HDInsight Hadoop Solution Proprietary Service που αναπτύσσει και προβλέπει συστοιχίες Hadoop στο cloud, παρέχοντας ένα πλαίσιο λογισμικού σχεδιασμένο για τη διαχείριση, ανάλυση και το reporting σε big data

Microsoft Windows Azure HDInsight

Ιστοθέση:https://azure.microsoft.com/en-us/services/hdinsight/

Hadoop Solution Proprietary Service that deploys and provisions Hadoop clusters in the cloud, providing a software framework designed to manage, analyze and report on big data

Υπηρεσία Hadoop Solution που αναπτύσσει και παρέχει συστοιχίες Hadoop στο cloud, παρέχοντας ένα πλαίσιο λογισμικού σχεδιασμένο για τη διαχείριση, ανάλυση και αναφορά μεγάλων δεδομένων

Oracle Big Data Appliance

Ιστοθέση:https://www.oracle.com/engineered-systems/big-data-appliance/

Μια ολοκληρωμένη πλατφόρμα μεγάλων δεδομένων σχεδιασμένη να αποκτά, να οργανώνει και να αναλύει μεγάλους φόρτους εργασίας δεδομένων από διάφορες πηγές σε ταχύτητα και με δυνατότητα κλιμάκωσης. Το οικονομικά αποδοτικό σύστημα της Oracle προσφέρει λειτουργικότητα για ανάλυση υψηλής ποιότητας.

Teradata Appliance for Hadoop

Ιστοθέση:https://docs.teradata.com/reader/0EWpaXyr07ATwpcM_b83ew/pgDY9HXi6WnBmMkbGeLphA

Ενσωματωμένη στοίβα υλικού / λογισμικού, βελτιστοποιημένη για αποθήκευση μεγάλων δεδομένων εταιρικής κλάσης και βελτίωση του Qubole Hadoop Solution Proprietary Hadoop-as-a-Service που εκτελείται στο Amazon AWS

Amazon EMR (Elastic MapReduce)

Ιστοθέση:https://aws.amazon.com/emr/

Το Amazon EMR είναι η κορυφαία πλατφόρμα δεδομένων cloud για την επεξεργασία τεράστιων ποσοτήτων δεδομένων χρησιμοποιώντας εργαλεία ανοιχτού κώδικα όπως Apache Spark, Apache Hive, Apache HBase, Apache Flink, Apache Hudi και Presto

Το πλαίσιο Hadoop που εκτελείται σε υποδομή κλίμακας ιστού του Amazon Elastic Compute Cloud (Amazon EC2) και της υπηρεσίας απλής αποθήκευσης Amazon (Amazon S3)

Ανοιχτού Κώδικα

Apache Hadoop MapReduce	
Ιστοθέση	http://hadoop.apache.org
Τύπος	Πλαίσιο Ανάπτυξης Εφαρμογών
	**Περιγραφή**
	Πλαίσιο που επιτρέπει την κατανεμημένη επεξεργασία μεγάλων συνόλων δεδομένων σε ομάδες υπολογιστών χρησιμοποιώντας απλά μοντέλα προγραμματισμού όπως Cloudera Hadoop Distribution Proprietary Distribution of Hadoop και σχετικά έργα

Disco	
Ιστοθέση	http://discoproject.org
Τύπος	Πλαίσιο Ανάπτυξης Εφαρμογών
	**Περιγραφή**
	Ελαφρύ πλαίσιο για κατανεμημένους υπολογισμούς με βάση το πρότυπο MapReduce, που αναπτύχθηκε από τη Nokia. Το Disco είναι ισχυρό και εύκολο στη χρήση, χάρη στην Python. Διανέμει και αναπαράγει τα δεδομένα και προγραμματίζει αποτελεσματικά τις εργασίες. Το Disco περιλαμβάνει ακόμη και τα εργαλεία που χρειάζονται για την ευρετηρίαση δισεκατομμυρίων σημείων δεδομένων και την υποβολή ερωτημάτων σε πραγματικό χρόνο.

4.2.4 Επεξεργασιας Συνεχούς Ρεύματος Δεδομένων | Streaming Processing

Εικόνα 39.Εμπορικά Εργαλεία Streaming

Esper	
Ιστοθέση	http://www.espertech.com
Τύπος	Complex Event Processing
EsperTech	**Περιγραφή**
	Το Esper είναι μια γλώσσα, ένας μεταγλωττιστής και μία runtime εφαρμογή για σύνθετη επεξεργασία συμβάντων (CEP) και αναλύσεις ροής, διαθέσιμο για Java καθώς και για .NET.

TIBCO StreamBase	
Ιστοθέση	https://www.tibco.com/resources/datasheet/tibco-streambase
Τύπος	Complex Event Processing
TIBCO StreamBase	**Περιγραφή**
	Το TIBCO StreamBase είναι πλατφόρμα επεξεργασίας συμβάντων για την εφαρμογή μαθηματικής και σχεσιακής επεξεργασίας σε ροές δεδομένων σε πραγματικό χρόνο. Επιτρέπει στους οργανισμούς να δημιουργούν και να αναπτύσσουν ταχέως εφαρμογές που βασίζονται σε συμβάντα για την αυτοματοποιημένη διαδικασία γρήγορων δεδομένων σε πολύ μικρότερη κλίμακα κόστους και του κινδύνου εναλλακτικών λύσεων.

Apache Storm	
Ιστοθέση	https://storm.apache.org
Τύπος	Complex Event Processing
	**Περιγραφή**
	Το Apache Storm είναι ένα σύστημα υπολογισμού σε πραγματικό χρόνο δωρεάν και ανοιχτού κώδικα.Καθιστά εύκολη την αξιόπιστη επεξεργασία μη περιορισμένων ροών δεδομένων, κάνοντας σε επεξεργασία πραγματικού χρόνου αυτό που έκανε το Hadoop σε επεξεργασία batch. Το Apache Storm είναι απλό, μπορεί να χρησιμοποιηθεί με οποιαδήποτε γλώσσα προγραμματισμού. Εχει πολλές περιπτώσεις χρήσης: αναλυτικά στοιχεία σε πραγματικό χρόνο, διαδικτυακή μηχανική μάθηση, συνεχόμενος υπολογισμός, κατανεμημένο RPC, ETL και άλλα. Είναι γρήγορο: ένα σημείο αναφοράς το έφτασε σε πάνω από ένα εκατομμύριο πλειάδες επεξεργασμένες ανά δευτερόλεπτο σε κάθε κόμβο. Είναι επεκτάσιμο, ανεκτικό σε σφάλματα, εγγυάται την επεξεργασία των δεδομένων και είναι εύκολο να ρυθμιστεί και να λειτουργήσει.

Apache Flink	
Ιστοθέση	https://flink.apache.org
Τύπος	Υπολογιστικό Πλαίσιο
	**Περιγραφή**
	Το Apache Flink είναι ένας μηχανισμός πλαισίου και κατανεμημένης επεξεργασίας για ντετερμενιστικούς υπολογισμούς σε μη περιορισμένες και οριοθετημένες ροές δεδομένων. Το Flink έχει σχεδιαστεί για να λειτουργεί σε όλα τα κοινά περιβάλλοντα συστοιχιών, να εκτελεί υπολογισμούς σε ταχύτητα μνήμης και σε οποιαδήποτε κλίμακα.

Apache Kafka	
Ιστοθέση	https://kafka.apache.org
Τύπος	Streaming & Messaging
	**Περιγραφή**
	Apache Kafka είναι πλατφόρμα λογισμικού για επεξεργασία ροών δεδομένων. Αναπτύχθηκε αρχικά από την εταιρεία πίσω από το κοινωνικό δίκτυο LinkedIn και κατόπιν δόθηκε ως δωρεά στο Ίδρυμα Λογισμικού Apache. Είναι γραμμένη στις γλώσσες προγραμματισμού Scala και Java, ενώ πρόκειται για λογισμικό κώδικα ελεύθερου προς ανάπτυξη από όλους. Το έργο αποσκοπεί να παρέχει μια ενιαία πλατφόρμα για χειρισμό ροών δεδομένων σε πραγματικό χρόνο, με χαρακτηριστικά την υψηλή απόδοση και ελάχιστες περιόδους αδράνειας. Στην αρχιτεκτονική του το επίπεδο αποθήκευσης είναι

	ουσιαστικά μια ουρά δημοσίευσης και κατανάλωσης μηνυμάτων, με τεράστια επιδεκτικότητα διεύρυνσης, σχεδιασμένη σαν ένα κατανεμημένο αρχείο καταγραφής συναλλαγών

4.2.5 Επεξεργασίας και Ελέγχου Αρχείων Καταγραφής

-Log Processing & Monitoring-

Elastic Kibana & Logstash	
Ιστοθέση	https://www.elastic.co/logstash https://www.elastic.co/kibana
Τύπος	Διαχείριση και Οπτικοποίηση Αρχείων Καταγραφής
	**Περιγραφή**
	Το **Kibana** επιτρέπει την οπτικοποιήση των δεδομένων του ElasticSearch και την πλοήγηση στο Elastic Stack, ώστε να μπορείτε να κάνετε οτιδήποτε, από την παρακολούθηση του φορτίου των queries έως την κατανόηση του τρόπου με τον οποίο τα αιτήματα ρέουν μέσω των εφαρμογών σας. Το **Logstash** είναι ένας αγωγός επεξεργασίας δεδομένων ανοιχτού κώδικα για διακομιστές, που απορροφά δεδομένα από πολλές πηγές ταυτόχρονα, τα μετατρέπει και, στη συνέχεια, τα στέλνει στο επιθυμητό "stash".

Graylog2	
Ιστοθέση	https://www.graylog.org
Τύπος	Διαχείριση Log
	**Περιγραφή**
	Εργαλείο για την διαχείριση Log

Loggly	
Ιστοθέση	https://www.loggly.com
Τύπος	SaaS Διαχείριση Log
	**Περιγραφή**
	Το Loggly είναι μια λύση SaaS για τη διαχείριση δεδομένων καταγραφής. Κάνει εφικτή την μεταφορά αρχείων καταγραφής από τα βάθη ολόκληρης της υποδομής σε ένα μέρος όπου μπορεί να παρακολουθείτε η δραστηριότητα και να γίνεται ανάλυση των τάσεων. Επειδή το Loggly είναι μια διαχειριζόμενη υπηρεσία, δεν απαιτείται επιπλέον υλικό ή λογισμικό για να γίνει εφικτή η χρήση του Loggly και επίσης κλιμακώνεται δυναμικά κατά τη λειτουργία σας.

Ανοιχτού Κώδικα

Apache Flume	
Ιστοθέση	https://flume.apache.org
Τύπος	Υπηρεσία Διαχείρισης Log
	**Περιγραφή**
	Το Flume είναι μια κατανεμημένη, αξιόπιστη και διαθέσιμη υπηρεσία για αποτελεσματική συλλογή, συγκέντρωση και μεταφορά μεγάλων ποσοτήτων δεδομένων καταγραφής. Έχει μια απλή και ευέλικτη αρχιτεκτονική που βασίζεται σε ροές δεδομένων ροής. Είναι ανθεκτικό σε σφάλματα με συντονιζόμενους μηχανισμούς αξιοπιστίας και πολλούς μηχανισμούς ανακατεύθυνσης και ανάκτησης. Χρησιμοποιεί ένα απλό επεκτάσιμο μοντέλο δεδομένων που επιτρέπει online αναλυτική εφαρμογή.

Fluentd	
Ιστοθέση	https://www.fluentd.org
Τύπος	Διαχείριση Log
	**Περιγραφή**
	Εργαλείο για τη συλλογή συμβάντων και αρχείων καταγραφής με δυνατότητες προσθηκών

4.3 Πλαίσια Λογισμικού Υποδομής

Εικόνα 40.Πλαίσια Υποδομής

Apache Mesos	
Ιστοθέση	http://mesos.apache.org
Τύπος	Εικονικοποίηση/Ομοιογενοποίηση Πόρων Υλικού
	**Περιγραφή**
	Το Apache Mesos δημιουργεί αφαιρέσεις για την CPU, τη μνήμη, την αποθήκευση και άλλους υπολογιστικούς πόρους ανεξάρτητες από μηχανήματα (φυσικά ή εικονικά), επιτρέποντας την ανοχή σφαλμάτων και τα επεκτάσιμα κατανεμημένα συστήματα να κατασκευάζονται εύκολα και να λειτουργούν αποτελεσματικά. Το Mesos κατασκευάζεται χρησιμοποιώντας τις ίδιες αρχές με τον πυρήνα Linux, αλλά σε διαφορετικό επίπεδο αφαίρεσης. Ο πυρήνας του λειτουργεί σε κάθε μηχανή και παρέχει εφαρμογές (π.χ. Hadoop, Spark, Kafka, Elasticsearch) με API για διαχείριση πόρων και προγραμματισμός σε ολόκληρο το κέντρο δεδομένων και το περιβάλλον cloud.

Docker	
Ιστοθέση	https://www.docker.com
Τύπος	Εικονικοποίηση Πόρων
	**Περιγραφή**
	Το Docker είναι ένα σύνολο προϊόντων ως Υπηρεσία(PaaS) που χρησιμοποιεί εικονικοποίηση σε επίπεδο λειτουργικού συστήματος για την παράδοση λογισμικού σε πακέτα που ονομάζονται containers. Οι containers είναι απομονωμένα το ένα από το άλλο και ομαδοποιούν το δικό τους λογισμικό, βιβλιοθήκες και αρχεία διαμόρφωσης. Μπορούν να επικοινωνούν μεταξύ τους μέσω καλά καθορισμένων καναλιών.Όλοι οι containers εκτελούνται από έναν πυρήνα λειτουργικού συστήματος και επομένως χρησιμοποιούν λιγότερους πόρους από τις εικονικές μηχανές.

Kubernetes	
Ιστοθέση	https://kubernetes.io
Τύπος	Ενορχήστρωση Containers/Εικονικοποίηση
	**Περιγραφή**
	Είναι ένα σύστημα ενορχήστρωσης ανοιχτού κώδικα για αυτοματοποίηση ανάπτυξης, κλιμάκωσης και διαχείρισης εφαρμογών. Αρχικά σχεδιάστηκε από την Google και τώρα συντηρείται από το Cloud Native Computing Foundation. Στόχος του είναι να παρέχει μια «πλατφόρμα για την αυτοματοποίηση της ανάπτυξης, κλιμάκωσης και λειτουργίας των containers εφαρμογών σε συστοιχίες διακομιστών». Λειτουργεί με μια σειρά εργαλείων containers, συμπεριλαμβανομένου του Docker. Πολλές υπηρεσίες cloud προσφέρουν μια πλατφόρμα ή υποδομή που βασίζεται σε Kubernetes ως υπηρεσία (PaaS ή IaaS) στην οποία τα Kubernetes μπορούν να αναπτυχθούν ως υπηρεσία παροχής πλατφόρμας. Πολλοί προμηθευτές παρέχουν επίσης τις δικές τους επώνυμες διανομές Kubernetes.

Apache TEZ	
Ιστοθέση	https://tez.apache.org
Τύπος	Πλαίσιο Ανάπτυξης Εφαρμογών
	**Περιγραφή**
	Το έργο Apache TEZ® στοχεύει στη δημιουργία ενός πλαισίου εφαρμογής που επιτρέπει ένα σύνθετο κατευθυνόμενο-ακυκλικό γράφημα εργασιών για την επεξεργασία δεδομένων. Είναι επί του παρόντος χτισμένο πάνω στο Apache Hadoop YARN.

Apache Thrift	
Ιστοθέση	https://thrift.apache.org
Τύπος	Πλαίσιο Ανάπτυξης Εφαρμογών
Apache Thrift™	**Περιγραφή**
	Το Thrift είναι μια γλώσσα ορισμού διεπαφών και ένα πρωτόκολλο δυαδικής επικοινωνίας που χρησιμοποιείται για τον ορισμό και τη δημιουργία υπηρεσιών για πολλές γλώσσες.Αποτελεί ένα πλαίσιο κλήσης απομακρυσμένης διαδικασίας (RPC) και αναπτύχθηκε στο Facebook για "ανάπτυξη κλιμακούμενων διαγλωσσικών υπηρεσιών". Συνδυάζει μια στοίβα λογισμικού με μια μηχανή δημιουργίας κώδικα για τη δημιουργία υπηρεσιών πολλαπλών πλατφορμών που μπορούν να συνδέσουν εφαρμογές γραμμένες σε διάφορες γλώσσες και πλαίσια, όπως ActionScript, C, C ++, C #, Cappuccino, Cocoa, Delphi, Erlang, Go, Haskell, Java, JavaScript, Objective-C, OCaml, Perl, PHP, Python, Ruby, Elixir , Rust, Smalltalk και Swift.Αν και αναπτύχθηκε στο Facebook, είναι πλέον ένα έργο ανοιχτού κώδικα στο Ίδρυμα Λογισμικού Apache.

Apache Spark	
Ιστοθέση	https://spark.apache.org
Τύπος	In-Memory Ενοποιημένη Μηχανή Analytics
APACHE Spark™	**Περιγραφή**
	Το Apache Spark είναι ένα ανοιχτού κώδικα κατανεμημένο υπολογιστικό πλαίσιο γενικού σκοπού. Το Spark παρέχει μια διεπαφή για τον προγραμματισμό ολόκληρων ομάδων με έμμεσο παραλληλισμό δεδομένων και ανοχή σφαλμάτων. Επιτυγχάνει υψηλή απόδοση τόσο για δεδομένα batch όσο και για ροή δεδομένων, χρησιμοποιώντας έναν υπερσύγχρονο προγραμματιστή DAG, έναν βελτιστοποιητή ερωτημάτων και μια μηχανή φυσικής εκτέλεσης. Αρχικά αναπτύχθηκε στο Πανεπιστήμιο της Καλιφόρνιας, στο AMPLab του Μπέρκλεϋ, η βάση δεδομένων Spark αργότερα δωρίστηκε στο Apache Software Foundation, το οποίο το έχει διατηρήσει έκτοτε.

Akka	
Ιστοθέση	https://akka.io
Τύπος	Πλαίσιο Ανάπτυξης Εφαρμογών
akka	**Περιγραφή**
	Το Akka είναι ένα δωρεάν κιτ εργαλείων ανοιχτού κώδικα που απλοποιεί την κατασκευή ταυτόχρονων και κατανεμημένων εφαρμογών στο Java VM. Η Akka υποστηρίζει πολλαπλά μοντέλα προγραμματισμού για ταυτόχρονη χρήση, αλλά δίνει έμφαση στον συγχρονισμό που βασίζεται σε ρόλους, με έμπνευση από την Erlang

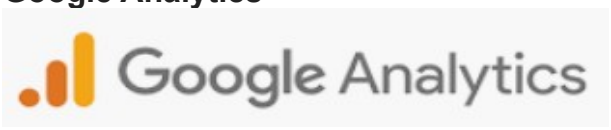
4.4 Μηχανές και Πλατφόρμες για Analytics, Επιστήμη Δεδομένων και Τεχνητή Νοημοσύνη
-Engines and Platforms for Analytics, Data Science & AI-

4.4.1 Πλατφόρμες Data Analytics

Εικόνα 41.Εμπορικές Πλατφόρμες Analytics

4.4.1.1 Analytics Platforms

Google Analytics



Ιστοθέση:https://marketingplatform.google.com/about/analytics/
Το Google Analytics είναι μια διαδικτυακή υπηρεσία analytics που προσφέρεται από την Google, η οποία παρακολουθεί και καταγράφει την κυκλοφορία ιστότοπων, επί του παρόντος ως πλατφόρμα εντός του πακέτου επωνυμία Google Marketing Platform με δυνατότητες οπτικοποίησης και API

SAP Analytics

Ιστοθέση:https://www.sapanalytics.cloud
Εργαλείο Analytics που χρησιμοποιεί ως βάση την SAP HANA

Oracle Business Analytics

Ιστοθέση:https://www.oracle.com/business-analytics/
Μηχανή για εξιδεικευμένα Analytics

SAS Analytics

Ιστοθέση:https://www.sas.com/el_gr/software/all-products.html
Ολοκληρωμένο περιβάλλον για predictive και descriptive modeling, εξόρυξη δεδομένων, ανάλυση κειμένου, προβλέψεις, βελτιστοποίηση, προσομοίωση, πειραματικός σχεδιασμός και άλλα

4.4.2.2 Analytics Εργαλεία

CrossFunctional

Salesforce Marketing Cloud

Το Salesforce Marketing Cloud είναι ένας πάροχος λογισμικού και υπηρεσιών αυτοματοποίησης και ανάλυσης ψηφιακού μάρκετινγκ.

FICO

Ιστοθέση:https://www.fico.com/en/products/fico-analytics-workbench
Το FICO® Analytics Workbench ™ παρέχει ένα ολοκληρωμένο σύνολο προηγμένων δυνατοτήτων ανάλυσης για τη δημιουργία λύσεων επιχειρηματικών αποφάσεων υψηλής επίδρασης. Πρόκειται για μια cloud-based λύση analytics που παρέχει στους επιχειρησιακούς χρήστες και στους επιστήμονες δεδομένων εξελιγμένη αλλά και εύχρηστη εξερεύνηση δεδομένων, διαμόρφωση οπτικών δεδομένων, σχεδιασμό στρατηγικής αποφάσεων, μοντελοποίηση advanced scorecard και μηχανική μάθηση.

Online Web/Mobile/Commerce

Εικόνα 42.Εμπορικα Εργαλεία On-Line Analytics

Clicky	
Ιστοθέση	https://clicky.com
Τύπος	Διαχείριση Log
clicky	**Περιγραφή**
	Παρακολούθησης, ανάλυσης και αντίδρασης στην κίνηση του ιστολογίου ή του ιστοτόπου σε πραγματικό χρόνο.

Crazyegg	
Ιστοθέση	https://www.crazyegg.com
Τύπος	Διαχείριση Log
crazyegg	**Περιγραφή**
	Χρήση των οπτικων αναφορών και τις ατομικών εγγραφών συνεδρίας για αναγνώριση των επισκέπτών του ιστότοπού σας - από πού προέρχονται, από πού πλοηγούνται και από πού κολλάνε - ώστε να μπορούν να γίνουν βέλτιστες αλλαγές στη σχεδίαση.

Social Media

Εικόνα 43.Εμπορικα Εργαλεία Social Media Analytics

BrandWatch	
Ιστοθέση	https://www.brandwatch.com
Τύπος	Social Media Analytics
Brandwatch	**Περιγραφή**
	Η κάλυψη του εργαλείου περιλαμβάνει ιστολόγια, ιστότοπους ειδήσεων, φόρουμ, βίντεο, κριτικές, εικόνες και κοινωνικά δίκτυα, συμπεριλαμβανομένων των Twitter, Facebook, Instagram και Reddit. Οι χρήστες μπορούν να αναζητήσουν δεδομένα χρησιμοποιώντας αναζήτηση κειμένου και εικόνας, και να χρησιμοποιήσουν γραφήματα, κατηγοριοποίηση, ανάλυση συναισθημάτων και άλλες δυνατότητες για να παρέχουν περαιτέρω πληροφορίες και ανάλυση. Η Brandwatch έχει πρόσβαση σε περισσότερες από 80 εκατομμύρια πηγές.

4.4.2 AI / Machine Learning / Deep Learning

4.4.2.1 Οριζόντιες Λύσεις

Εικόνα 44.Ολοκληρωμένες Εμπορικές Πλατφόρμες AI

4.4.2.2 Μηχανικής Μάθησης & Αναγνώρισης Εικόνων & Επεξεργασίας Φυσικής Γλώσσας

Machine Learning	Vision	NLP
Azure Machine Learning Amazon SageMaker Google Cloud AutoML Vision H2O.ai DataRobot gamalon ViSENZE ELEMENT AI deepsense.ai	Microsoft Azure Amazon Rekognition Cloud Vision API clarifai 商汤 sensetime EVER AI deepomatic neurala twentybn UBIQUITY6 AiBEE 依图 YITU trax 云从科技 CLOUDWALK synthesia DataGrid	Google Cloud twilio amazon alexa Amazon Translate Amazon Polly narrative science semanticmachines WolframAlpha Mobvoi EigenTechnologies SoundHound Inc. voicera PRIMER MindMeld cortical.io NUANCE maluuba Gridspace cogito snips Amenity Analytics SMARTLING Unbabel PolyAI RASA

Εικόνα 45.Εξειδικευμένες Εμπορικές Πλατφόρμες AI

4.4.2.3 Ανοιχτού Κώδικα

Εικόνα 46.Πλατφόρμες AI Ανοιχτού Κώδικα

Apache Mahout	
Ιστοθέση	https://mahout.apache.org
Τύπος	Machine Learning Platform
MAHOUT	**Περιγραφή** Το Apache Mahout (TM) είναι ένα κατανεμημένο αλγεβρικό γραμμικό πλαίσιο και μαθηματικά εκφρασμένο Scala DSL σχεδιασμένο για να επιτρέπει στους μαθηματικούς, τους στατιστικούς και τους επιστήμονες δεδομένων να εφαρμόσουν γρήγορα τους δικούς τους αλγόριθμους. Το Apache Spark είναι η συνιστώμενη διανομή back-end-of-the-box, ή μπορεί να επεκταθεί και σε άλλα κατανεμημένα backend.

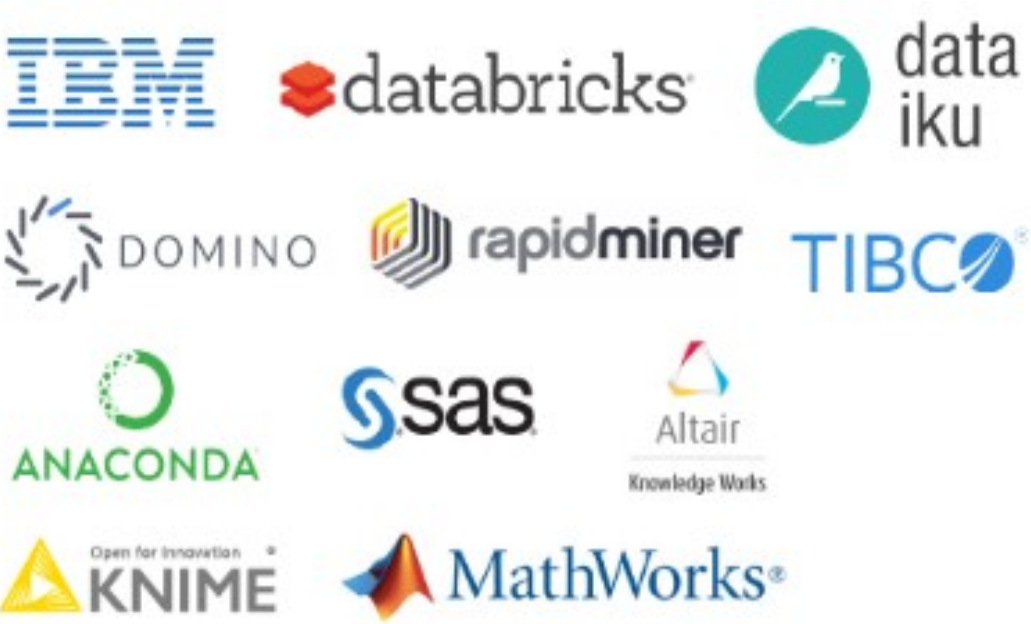
	Υποστήριξη και σε άλλα κατανεμημένα Backends και Modular Native Solvers για επιτάχυνση CPU / GPU / CUDA

Orange	
Ιστοθέση	https://orange.biolab.si
Τύπος	Machine Learning
	**Περιγραφή**
	Εργαλείο ανάλυσης δεδομένων με υποστήριξη για προσθήκες, που αναπτύχθηκε από το Πανεπιστήμιο της Λιουμπλιάνα. Το Orange είναι μια ανοιχτή κώδικα εργαλειοθήκη οπτικοποίησης δεδομένων, μηχανικής μάθησης και εξόρυξης δεδομένων. Διαθέτει οπτικό προγραμματιζόμενο front-end για διερευνητική ανάλυση δεδομένων και διαδραστική οπτικοποίηση δεδομένων.

Salford Systems SPM	
Ιστοθέση	https://www.salford-systems.com/products
Τύπος	Predictive Modeling
	**Περιγραφή**
	Πλατφόρμα ανάλυσης και εξόρυξης δεδομένων για τη δημιουργία προγνωστικών, περιγραφικών και αναλυτικών μοντέλων από βάσεις δεδομένων οποιουδήποτε μεγέθους, πολυπλοκότητας ή οργάνωσης

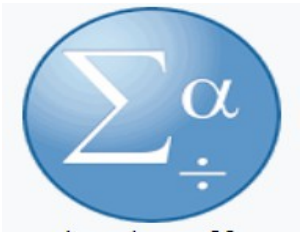
4.4.3 Στατιστικά Εργαλεία,Γλώσσες & Επιστήμη Δεδομένων

-Statistical Tools ,Languages & Data Science Platforms-



Εικόνα 47.Πλατφόρμες Data Science

Mathematica	
Ιστοθέση	https://www.wolfram.com/mathematica/
Τύπος	Mathematical Analytics
	**Περιγραφή**
	Το Wolfram Mathematica (συνήθως ονομάζεται Mathematica) είναι ένα σύγχρονο υπολογιστικό σύστημα που εκτείνεται στους περισσότερους τομείς της υπολογιστικής - συμπεριλαμβανομένων των νευρικών δικτύων, της μηχανικής μάθησης, της επεξεργασίας εικόνας, της γεωμετρίας, της επιστήμης δεδομένων, των οπτικοποιήσεων και άλλων. Χρησιμοποιείται σε πολλούς τεχνικούς, επιστημονικούς, μηχανικούς, μαθηματικούς και υπολογιστικούς τομείς

IBM SPSS	
Ιστοθέση	https://www.wolfram.com/mathematica/
Τύπος	Statistical Analysis
	**Περιγραφή**
	Το SPSS Statistics είναι ένα πακέτο λογισμικού που χρησιμοποιείται για διαδραστική ή παρτίδα, στατιστική ανάλυση.

Ανοιχτού Κώδικα

Εικόνα 48.Γλώσσες Προγραμματισμού Data Science

R	
Ιστοθέση	https://www.r-project.org
Τύπος	Statistical analysis and visualization engine
	Περιγραφή
	Η R είναι μια γλώσσα προγραμματισμού και ένα περιβάλλον ελεύθερου λογισμικού για στατιστική πληροφορική και γραφικά που υποστηρίζονται από το R Foundation for Statistics Computing. Η γλώσσα R χρησιμοποιείται ευρέως μεταξύ των στατιστικολόγων και των αναλυτών δεδομένων για την ανάπτυξη στατιστικού λογισμικού και την ανάλυση δεδομένων.

4.5 Πλατφόρμες Εξαγωγής Αναφορών & Οπτικοποίησης & Επιχειρησιακής Ευφυίας

-Reporting & Visualization & BI Platforms-

Εικόνα 49.Εμπορικές Πλατφόρμες BI

D3.js	
Ιστοθέση	https://d3js.org
Τύπος	Statistical analysis and visualization engine
	Περιγραφή
	Βιβλιοθήκη που χρησιμοποιεί JavaScript, HTML, SVG και CSS για την απόδοση διαγραμμάτων και γραφημάτων

Gephi	
Ιστοθέση	https://gephi.org
Τύπος	Graph Analysis
	Περιγραφή
	Πακέτο λογισμικού ανάλυσης και οπτικοποίησης δικτύων και γράφων

IBM Cognos	
Ιστοθέση	https://www.ibm.com/products/cognos-analytics
Τύπος	Reporting and BI engine
	**Περιγραφή**
	Το IBM Cognos Business Intelligence είναι μια ολοκληρωμένη σουίτα επιχειρηματικής ευφυΐας που από την IBM. Παρέχει ένα σύνολο εργαλείων για αναφορές, αναλυτικά στοιχεία, κάρτες αποτελεσμάτων και παρακολούθηση συμβάντων και μετρήσεων

QlikView	
Ιστοθέση	https://www.qlik.com/us
Τύπος	Reporting and BI engine
	**Περιγραφή**
	End-to-end πλατφόρμα που περιλαμβάνει ενοποίηση δεδομένων, επιχειρηματική ευφυΐα καθοδηγούμενη από τον χρήστη και συγκριτική ανάλυση.

TIBCO Spotfire	
Ιστοθέση	https://www.tibco.com/products/tibco-spotfire
Τύπος	Reporting and BI engine
	**Περιγραφή**
	Το TIBCO Spotfire είναι μια λύση analytics που παρέχει αναζήτηση και συστάσεις που υποστηρίζονται από μια ενσωματωμένη μηχανή τεχνητής νοημοσύνης. Επιτρέπει τη δημιουργία απλών μετρήσεων ταμπλό, προγνωστικών εφαρμογών ή δυναμικών εφαρμογών ανάλυσης σε πραγματικό χρόνο, παρέχει πολλές κλιμακούμενες δυνατότητες, συμπεριλαμβανομένων οπτικοποιήσεων, wrangling δεδομένων , analytics προβλέψεων, analytics τοποθεσίας και ροής.

Tableau	
Ιστοθέση	https://www.tableau.com
Τύπος	Reporting and BI engine
	**Περιγραφή**
	Πλατφόρμα οπτικής ανάλυσης που προσφέρει διαδραστική οπτικοποίηση δεδομένων

Visually	
Ιστοθέση	https://visual.ly
Τύπος	Οπτικοποίηση
	Περιγραφή
	Το Visual.ly είναι μια πλατφόρμα κοινότητας για οπτικοποίηση δεδομένων,γραφήματα, infographics

Yellowfin	
Ιστοθέση	https://www.yellowfinbi.com
Τύπος	Reporting and BI engine
	Περιγραφή
	Πλατφόρμα επιχειρησιακών analytics που επιτρέπει στους οργανισμούς να εξαγάγουν αξία από τα δεδομένα τους, επειδή συνδυάζουν πίνακες ελέγχου με βάση τη δράση, αυτοματοποιημένη ανακάλυψη δεδομένων και αφήγηση δεδομένων σε μια ενιαία, ολοκληρωμένη πλατφόρμα.

Zoomdata	
Ιστοθέση	https://www.zoomdata.com
Τύπος	Analytics
	Περιγραφή
	Το Zoomdata είναι μια εταιρεία λογισμικού επιχειρηματικής ευφυΐας που ειδικεύεται στην οπτικοποίηση μεγάλων δεδομένων σε πραγματικό χρόνο , σε ροές δεδομένων και σε ανάλυση πολλαπλών πόρων. Τα προϊόντα της εταιρείας είναι διαθέσιμα on-prem, στο cloud και ενσωματωμένα σε άλλες εφαρμογές

Minitab	
Ιστοθέση	https://www.minitab.com/en-us/products/minitab/
Τύπος	Reporting and BI engine
	Περιγραφή
	Το Minitab είναι ένα στατιστικό εργαλείο για εκμεταλλευση των δεδομένων, ανακαλύψη τάσεων, πρόβλεψη μοτίβων, αποκάλυψη κρυφών σχέσεων μεταξύ μεταβλητών, οπτικοποίηση των αλληλεπιδράσεων δεδομένων ο εντοπίσμός σημαντικών παραγόντων.

4.6 Διαχείρισης & Ενορχήστρωσης | Orchestration & Management

Εικόνα 50.Πλατφόρμες Διαχείρισης

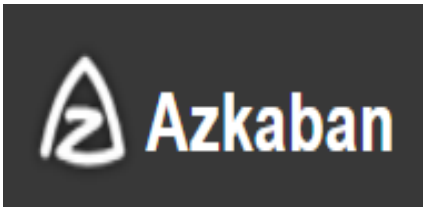
Apache Ambari	
Ιστοθέση	https://ambari.apache.org
Τύπος	Provisioning, Monitoring
Apache Ambari	**Περιγραφή** Το έργο Apache Ambari στοχεύει στην απλοποίηση της διαχείρισης Hadoop αναπτύσσοντας λογισμικό για την παροχή, διαχείριση και παρακολούθηση συστοιχιών Apache Hadoop. Παρέχει ένα διαισθητικό, εύχρηστο διαδικτυακό περιβάλλον εργασίας χρήστη που υποστηρίζεται από τα RESTful API του. Υποστήριζει για Hadoop HDFS, Hadoop MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig and Sqoop

Apache Hadoop YARN	
Ιστοθέση	https://hadoop.apache.org
Τύπος	Resource management
hadoop YARN	**Περιγραφή** Πλαίσιο για προγραμματισμό εργασιών και διαχείριση πόρων συστοιχιών Hadoop Μια πλατφόρμα υπεύθυνη για τη διαχείριση πόρων υπολογιστών και τη χρήση τους από τον προγραμματισμό των εφαρμογών των χρηστών

Apache Airflow	
Ιστοθέση	https://airflow.apache.org
Τύπος	Workflow Management
	**Περιγραφή**
	Το Apache Airflow είναι μια πλατφόρμα διαχείρισης ροών εργασίας ανοιχτού κώδικα

Apache Oozie	
Ιστοθέση	https://oozie.apache.org
Τύπος	Workflow Management
	**Περιγραφή**
	Σύστημα ροής εργασίας / συντονισμού για τη διαχείριση και τον προγραμματισμό εργασιών Apache Hadoop

Apache ZooKeeper	
Ιστοθέση	https://oozie.apache.org
Τύπος	Coordination
	**Περιγραφή**
	Το ZooKeeper είναι μια κεντρικοποιημένη υπηρεσία για τη διατήρηση πληροφοριών διαμόρφωσης, την ονομασία, την παροχή κατανεμημένου συγχρονισμού και την παροχή υπηρεσιών ομάδας. Όλα αυτά τα είδη υπηρεσιών χρησιμοποιούνται σε κάποια μορφή από άλλες κατανεμημένες εφαρμογές.

Azkaban	
Ιστοθέση	https://azkaban.github.io
Τύπος	Workflow
	**Περιγραφή**
	Χρονοπρογραμματιστής ροής εργασιών batch που δημιουργήθηκε στο LinkedIn για την εκτέλεση των εργασιών Hadoop. Το Azkaban επιλύει την χρονική ακολουθία μέσω εξαρτήσεων μεταξύ των εργασιών και παρέχει ένα εύχρηστο διαδικτυακό περιβάλλον εργασίας χρήστη για τη συντήρηση και παρακολούθηση των ροών εργασίας σας.

Chef	
Ιστοθέση	https://www.chef.io
Τύπος	Deployment Management/Provisioning
	**Περιγραφή**
	Ο Chef είναι μια εταιρεία και το όνομα ενός εργαλείου διαχείρισης διαμόρφωσης γραμμένο σε Ruby και Erlang. Χρησιμοποιεί μια καθαρή Ruby, γλώσσα για συγκεκριμένους τομείς (DSL) για τη σύνταξη των παραμετροποιήσεων συστήματος που ονομάζονται "συνταγές". Ο Chef χρησιμοποιείται για τον εξορθολογισμό της διαμόρφωσης και συντήρησης διακομιστών μιας εταιρείας και μπορεί να διασυνδεθεί σε πλατφόρμες που βασίζονται σε cloud όπως το Internap, το Amazon EC2, το Google Cloud Platform, το Oracle Cloud, το OpenStack, το SoftLayer, το Microsoft Azure και το Rackspace με στόχο την αυτόματη παροχή και διαμόρφωση νέων μηχανήματων.

Kong	
Ιστοθέση	**https://konghq.com**
Τύπος	APIs & Microservices Management
	**Περιγραφή**
	Πλατφόρμα ανοιχτού κώδικα και υπηρεσία cloud για διαχείριση, παρακολούθηση και κλιμάκωση των διεπαφών προγραμματισμού εφαρμογών (APIs) και των μικροϋπηρεσιών.

etcd	
Ιστοθέση	https://etcd.io
Τύπος	Store For Distributed Systems
	**Περιγραφή**
	Το Etcd είναι μια πολύ σταθερή, κατανεμημένη βάση δεδομένων κλειδιού-τιμής που παρέχει έναν αξιόπιστο τρόπο αποθήκευσης δεδομένων στα οποία απαιτείται πρόσβαση από ένα κατανεμημένο σύστημα ή μια συστοιχία μηχανών. Αντιμετωπίζει με ευκολία τις επιλογές των κυρίαρχων κόμβων κατά τη διάρκεια της κατάτμησης του δικτύου και μπορεί να ανεχθεί την αποτυχία των μηχανών, ακόμη και αν συμβεί στον κόμβο του ηγέτη.

4.6.1 Πλατφόρμες Data Governance

Εικόνα 51.Εργαλεία Data Governance

4.7 Κατανεμημένα Σύστηματα Αρχείων

Amazon S3	
Ιστοθέση	https://aws.amazon.com/s3/
amazon S3	**Περιγραφή**
	Υπηρεσία αποθήκευσης αρχείων στο διαδίκτυο, που προσφέρεται από το Amazon Web Services

Apache Hadoop HDFS	
Ιστοθέση	https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html
hadoop HDFS	**Περιγραφή**
	Κατανεμημένο σύστημα αρχείων που παρέχει πρόσβαση υψηλής απόδοσης σε δεδομένα εφαρμογών

Ceph	
Ιστοθέση	https://ceph.io/ceph-storage/
	Περιγραφή
	Κατανεμημένη αποθήκευση αντικειμένων και σύστημα αρχείων σχεδιασμένο να παρέχει εξαιρετική απόδοση, αξιοπιστία και επεκτασιμότητα. Το Ceph είναι μια πλατφόρμα αποθήκευσης ανοιχτού κώδικα, που υλοποιεί αποθήκευση αντικειμένων σε μια μόνο κατανεμημένη συστοιχία υπολογιστών και παρέχει διεπαφές 3i n1 για: αποθήκευση επιπέδου αντικειμένου, μπλοκ και αρχεία. Στοχεύει πρωτίστως σε πλήρως κατανεμημένη λειτουργία χωρίς κανένα σημείο βλάβης, κλιμάκωση σε επίπεδο exabyte και ελεύθερη διαθεσιμότητα.

Red Hat Gluster Storage	
Ιστοθέση	https://www.redhat.com/en/technologies/storage/gluster
	Περιγραφή
	Το Red Hat Gluster Storage είναι μια πλατφόρμα αποθήκευσης που καθορίζεται από λογισμικό (SDS Ssoftware-Defined storage). Έχει σχεδιαστεί για να χειρίζεται φόρτους εργασίας γενικού σκοπού, όπως δημιουργία αντιγράφων ασφαλείας και αρχειοθέτησης, καθώς και analytics. Είναι ιδανικό για υπερσυγκέντρωση δεδομένων. Είναι οικονομικά αποδοτικό και, σε αντίθεση με τα παραδοσιακά συστήματα αποθήκευσης, μπορεί να αναπτυχθεί σε περιβάλλοντα κέντρου δεδομένων, εικονικοποίησης, container και cloud.

Alluxio	
Ιστοθέση	https://www.alluxio.io
	Περιγραφή
	Το Alluxio είναι ένα σύστημα εικονικών κατανεμημένων αρχείων ανοιχτού κώδικα (VDFS) Βρίσκεται μεταξύ του επιπέδου υπολογισμών και αποθήκευσης στη οργανωτική στοίβα των Big Data Analytics. Παρέχει ένα επίπεδο αφαίρεσης δεδομένων για υπολογιστικά πλαίσια, επιτρέποντας στις εφαρμογές να συνδέονται σε πολλά συστήματα αποθήκευσης μέσω μιας κοινής διεπαφής. Οι εφαρμογές βάσης δεδομένων, όπως τα Analytics, η μηχανική μάθηση και το AI, χρησιμοποιούν API (όπως Hadoop HDFS API, S3 API, FUSE API) που παρέχονται από το Alluxio για να αλληλεπιδρούν με

	δεδομένα από διάφορα συστήματα αποθήκευσης με γρήγορη ταχύτητα

MooseFS	
Ιστοθέση	https://moosefs.com/about/#about
	**Περιγραφή**
	Το Moose File System (MooseFS) είναι ένα σύστημα διανομής αρχείων ανοιχτού κώδικα, συμβατό με τις αρχές POSIX, το οποίο αναπτύχθηκε από την Core Technology. Στοχεύει να είναι ενα ανεκτικό σε σφάλματα, υψηλής διαθεσιμότητας, υψηλής απόδοσης, κλιμακούμενο σύστημα γενικής χρήσης, κατανεμημένο σύστημα αρχείων για κέντρα δεδομένων.

4.8 Λύσεις Αποθήκευσης & Βάσεων Δεδομένων

Εικόνα 52.Πλατφόρμες Αποθήκευσης

4.8.1 Λύσεις Αποθήκευσης Υπολογιστικού Νέφους | Cloud Storage Solutions

Amazon DynamoDB
Ιστοθέση:https://aws.amazon.com/dynamodb/
Το Amazon DynamoDB είναι μία Cloud-based βάση δεδομένων τύπου εγγράφων και κλειδιού-τιμής που παρέχει μονοψήφια απόδοση χιλιοστών του δευτερολέπτου σε οποιαδήποτε κλίμακα. Είναι μια πλήρως διαχειριζόμενη,

διαμοιρασμένη σε διάφορες περιοχές, ανθεκτική βάση δεδομένων με ενσωματωμένη ασφάλεια, δημιουργία αντιγράφων ασφαλείας και επαναφορά(backup and restore) και μνήμη cache για εφαρμογές κλίμακας Διαδικτύου.

Amazon SimpleDB
Ιστοθέση:https://aws.amazon.com/simpledb/
Cloud βάση δεδομένων τύπου columnar. Αποτελεί μια κατανεμημένη βάση δεδομένων γραμμένη σε Erlang από την Amazon.com. Χρησιμοποιείται ως υπηρεσία διαδικτύου σε συνδυασμό με το Amazon Elastic Compute Cloud (EC2) και το Amazon S3 και είναι μέρος του ευρύτερου πακέτου Amazon Web Services.

Google BigTable
Ιστοθέση:https://cloud.google.com/bigtable
Είναι μια Cloud-based αποθήκευση τύπου key-value .Το Bigtable είναι ένα συμπιεσμένο, υψηλής απόδοσης, ιδιόκτητο σύστημα αποθήκευσης δεδομένων που βασίζεται στα Google File System, Chubby Lock Service, SSTable (log-δομημένος χώρος αποθήκευσης όπως το LevelDB) και μερικές άλλες τεχνολογίες Google

Microsoft Windows Azure Table Storage
Ιστοθέση:https://azure.microsoft.com/en-us/services/storage/tables/
Cloud-based αποθήκευση τύπου key-value.

4.8.2 NoSQL

4.8.2.1 Key-Value

Oracle NoSQL	
Ιστοθέση	https://www.oracle.com/database/technologies/related/nosql.html
	**Περιγραφή**
	Η βάση δεδομένων Oracle NoSQL (ONDB) είναι μια κατανεμημένη βάση δεδομένων κλειδί-τιμής τύπου NoSQL από την Oracle Corporation.Παρέχει συναλλακτική σημασιολογία για χειρισμό δεδομένων, οριζόντια επεκτασιμότητα και απλή διαχείριση και παρακολούθηση.

Redis	
Ιστοθέση	https://redis.io
	**Περιγραφή**
	Το Redis είναι μια ανοιχτού κώδικα (με άδεια BSD) δομή αποθήκευσης δεδομένων στη μνήμη, που χρησιμοποιείται ως βάση δεδομένων, cache και μεσίτης μηνυμάτων. Υποστηρίζει δομές δεδομένων όπως συμβολοσειρές, κατακερματισμούς (hashes), λίστες, σύνολα, ταξινομημένα σύνολα με ερωτήματα εύρους, bitmaps, υπερ-καταλόγους, γεωχωρικά ευρετήρια με διανυσματικά ερωτήματα ακτίνας και ροές δεδομένων. Το Redis διαθέτει ενσωματωμένο πολλαπλασιασμό δεδομένων, scripting τύπου Lua,

	έξαγωγή LRU, συναλλαγές και διαφορετικά επίπεδα διατήρησης στο δίσκο και παρέχει υψηλή διαθεσιμότητα μέσω του Redis Sentinel και αυτόματης κατάτμησης με το Redis Cluster

Voldemort	
Ιστοθέση	https://www.project-voldemort.com/voldemort/
	**Περιγραφή** Το Voldemort είναι ένα κατανεμημένο σύστημα αποθήκευσης κλειδιού/τιμής.Τα κυριότερα χαρακτηριστικά του είναι: ➢ Τα δεδομένα αναπαράγονται αυτόματα σε πολλούς διακομιστές. ➢ Τα δεδομένα κατανέμονται αυτόματα έτσι κάθε διακομιστής να περιέχει μόνο ένα υποσύνολο των συνολικών δεδομένων ➢ Παρέχει συντονισμένη συνέπεια (από αυστηρή συμμετοχή έως ενδεχόμενη συνέπεια) ➢ Η αποτυχία του διακομιστή αντιμετωπίζεται με διαφάνεια ➢ Τα στοιχεία δεδομένων έχουν εκδοθεί για μεγιστοποίηση της ακεραιότητας των δεδομένων σε σενάρια αποτυχίας χωρίς να διακυβεύεται η διαθεσιμότητα του συστήματος ➢ Κάθε κόμβος είναι ανεξάρτητος από άλλους κόμβους χωρίς κεντρικό σημείο αποτυχίας ή συντονισμού ➢ Καλή απόδοση κάθε κόμβου: 10-20k λειτουργίες ανά δευτερόλεπτο ανάλογα με τα μηχανήματα, το δίκτυο, το σύστημα δίσκου και τον παράγοντα αναπαραγωγής δεδομένων

Basho Riak	
Ιστοθέση	https://riak.com/products/#riak
	**Περιγραφή** Το Riak είναι ένας κατανεμημένος χώρος αποθήκευσης δεδομένων NoSQL τύπου κλειδιού-τιμής που προσφέρει υψηλή διαθεσιμότητα, ανοχή σφαλμάτων, λειτουργική απλότητα και επεκτασιμότητα.Εκτός από την έκδοση ανοιχτού κώδικα, διατίθεται σε μια υποστηριζόμενη εταιρική έκδοση και μια έκδοση αποθήκευσης cloud. Ο Riak εφαρμόζει τις αρχές από την Dynamo του Amazon με μεγάλη επιρροή από το θεώρημα CAP. Γραμμένη σε Erlang, η Riak είναι ανεκτική σε σφάλματα μέσω πολλαπλασιασμού δεδομένων και αυτόματη κατανομή δεδομένων σε όλο τη συστοιχία για απόδοση και ανθεκτικότητα.

4.8.2.2 Document-Store

Apache CouchDB	
Ιστοθέση	https://couchdb.apache.org
	**Περιγραφή** Το Apache CouchDB είναι μια βάση δεδομένων NoSQL εγγράφων ανοιχτού κώδικα, υλοποιημένη σε Erlang. Χρησιμοποιεί πολλές μορφές και πρωτόκολλα για αποθήκευση, μεταφορά και επεξεργασία των δεδομένων του, χρησιμοποιεί JSON για αποθήκευση δεδομένων, JavaScript ως γλώσσα ερωτήματων χρησιμοποιώντας το MapReduce και HTTP για το API.

MongoDB	
Ιστοθέση	https://www.mongodb.com
	**Περιγραφή** Το MongoDB είναι μια βάση δεδομένων έγγραφων πολλαπλών πλατφορμών. Καθώς κατατάχθηκε ως βάση δεδομένων NoSQL, χρησιμοποιεί έγγραφα σχήματος τύπου JSON. Αναπτύσσεται από τη MongoDB Inc. και διαθέτει άδεια χρήσης τύπου Δημόσιας Άδειας Διακομιστή (SSPL-Server Side Public License).

4.8.2.3 Wide Column Store

Apache HBase	
Ιστοθέση	https://hbase.apache.org
	**Περιγραφή** Επεκτάσιμη, κατανεμημένη βάση δεδομένων τύπου columnar που υποστηρίζει αποθήκευση δομημένων δεδομένων για μεγάλους πίνακες. Χρησιμοποιείται όταν απαιτείται τυχαία, σε πραγματικό χρόνο πρόσβαση ανάγνωσης / εγγραφής στα Big Data. Στόχος αυτού του έργου είναι η φιλοξενία πολύ μεγάλων πινάκων - δισεκατομμυρίων γραμμών X εκατομμυρίων στηλών - πάνω από συστοιχίες εξιδεικευμένου υλικού. Το Apache HBase είναι μια ανοιχτού κώδικα, κατανεμημένη, αναπτυσσόμενη, μη-σχεσιακή βάση δεδομένων που έχει διαμορφωθεί σύμφωνα με το Bigtable της Google. Ακριβώς όπως το Bigtable αξιοποιεί τον κατανεμημένο χώρο αποθήκευσης δεδομένων που παρέχεται από το σύστημα αρχείων Google, το Apache HBase παρέχει δυνατότητες τύπου Bigtable πάνω από το Hadoop και το HDFS.

Apache Cassandra	
Ιστοθέση	https://cassandra.apache.org
	Περιγραφή
	Το Apache Cassandra είναι ένα δωρεάν, ανοιχτού κώδικα, κατανεμημένο, τύπου wide-column, σύστημα διαχείρισης βάσεων δεδομένων NoSQL που έχει σχεδιαστεί για να χειρίζεται μεγάλες ποσότητες δεδομένων σε πολλούς διακομιστές, παρέχοντας υψηλή διαθεσιμότητα χωρίς κανένα σημείο αποτυχίας. Η Cassandra προσφέρει ισχυρή υποστήριξη για συστοιχίες που εκτείνονται σε πολλαπλά κέντρα δεδομένων με ασύγχρονο πολλαπλασιασμό (replication) χωρίς επιτήρηση που επιτρέπει λειτουργίες χαμηλού λανθάνοντος χρόνου για όλους τους καταναλωτές. Η Cassandra προσφέρει το σχεδιασμό κατανομής του Amazon Dynamo με το μοντέλο δεδομένων του Bigtable της Google.

Apache Accumulo	
Ιστοθέση	https://accumulo.apache.org
	Περιγραφή
	Ταξινομημένη, κατανεμημένη βάση δεδομένων κλειδιού / τιμής που αναπτύχθηκε στο NSA, με ασφάλεια σε επίπεδο κελιού για την εκχώρηση δικαιωμάτων σε μεμονωμένα κελιά πίνακα Είναι εξαιρετικά επεκτάσιμο και έχει ως βάση σχεδιασμού το Bigtable της Google.Δομείται πάνω από τα Apache Hadoop, Apache ZooKeeper και Apache Thrift. Ώντας γραμμένο σε Java, το Accumulo διαθέτει ετικέτες πρόσβασης σε επίπεδο κυψέλης και μηχανισμούς προγραμματισμού από την πλευρά του διακομιστή.

4.8.2.4 Graph Databases

Εικόνα 53.Βάσεις Αποθήκευσης Γράφων

Apache Giraph	
Ιστοθέση	https://giraph.apache.org
	**Περιγραφή** Το Apache Giraph είναι ένα επαναληπτικό σύστημα επεξεργασίας γράφων που έχει δημιουργηθεί για υψηλή επεκτασιμότητα. Για παράδειγμα, αυτή τη στιγμή χρησιμοποιείται στο Facebook για την ανάλυση του κοινωνικού γραφήματος που σχηματίζουν οι χρήστες και οι συνδέσεις τους. Το Giraph δημιουργήθηκε ως αντίστοιχο ανοιχτού κώδικα στο Pregel. Και τα δύο συστήματα είναι εμπνευσμένα από το μοντέλο Massal Synchronous Parallel κατανεμημένου υπολογισμού που εισήγαγε ο Leslie Valiant. Το Giraph προσθέτει πολλά χαρακτηριστικά πέρα από το βασικό μοντέλο Pregel, συμπεριλαμβανομένων των βασικών υπολογισμών, των θρυμματισμένων αθροιστών, της εισόδου προσανατολισμένης στην ακμή, τον εκτός πυρήνα υπολογισμό κ.α

4.8.3 NewSQL Databases

[39.Wikipedia]

Το NewSQL είναι μια κατηγορία σχεσιακών συστημάτων διαχείρισης βάσεων δεδομένων που επιδιώκουν να παρέχουν την επεκτασιμότητα των συστημάτων NoSQL για φόρτους εργασίας διαδικτυακής επεξεργασίας συναλλαγών (OLTP) διατηρώντας παράλληλα τις εγγυήσεις ACID ενός παραδοσιακού συστήματος βάσεων δεδομένων.

Πολλά εταιρικά συστήματα που χειρίζονται δεδομένα υψηλού προφίλ (π.χ. συστήματα οικονομικών και επεξεργασίας παραγγελιών) είναι πολύ μεγάλα για συμβατικές σχεσιακές βάσεις δεδομένων, αλλά έχουν απαιτήσεις συναλλαγών και συνέπειας που δεν είναι πρακτικές για συστήματα NoSQL. Οι μόνες επιλογές που ήταν προηγουμένως διαθέσιμες για αυτούς τους οργανισμούς ήταν είτε η αγορά ισχυρότερων υπολογιστών είτε η ανάπτυξη προσαρμοσμένου ενδιάμεσου λογισμικού που διανέμει αιτήματα μέσω συμβατικών DBMS. Και οι δύο προσεγγίσεις διαθέτουν υψηλό κόστος υποδομής ή / και κόστος ανάπτυξης. Τα συστήματα NewSQL προσπαθούν να συμφιλιώσουν τις συγκρούσεις.

Εικόνα 54.Βάσεις NewSQL

Apache Ignite	
Ιστοθέση	https://giraph.apache.org
	Περιγραφή
	Είναι μια ανοιχτή κώδικα κατανεμημένη βάση δεδομένων καθώς και πλατφόρμα προσωρινής αποθήκευσης και επεξεργασίας που έχει σχεδιαστεί για να αποθηκεύει και να υπολογίζει μεγάλους όγκους δεδομένων σε συστοιχίες κόμβων.Η βάση δεδομένων του χρησιμοποιεί τη μνήμη RAM ως το προεπιλεγμένο επίπεδο αποθήκευσης και επεξεργασίας, επομένως, ανήκει στην κατηγορία των υπολογιστών στη μνήμη.Ανεξάρτητα από το API που χρησιμοποιείται, τα δεδομένα στο Ignite αποθηκεύονται με τη μορφή ζευγών κλειδιών-τιμών. Η βάσης δεδομένων κλιμακώνεται οριζόντια, διανέμοντας ζεύγη τιμών-κλειδιών στην συστοιχία με τέτοιο τρόπο ώστε κάθε κόμβος να κατέχει ένα τμήμα του συνόλου δεδομένων και εξισορροπούνται αυτόματα κάθε φορά που ένας κόμβος προστίθεται ή αφαιρείται από την συστοιχία.Υποστηρίζει μια ποικιλία API, συμπεριλαμβανομένων όσων συμμορφώνονται με JCache, τις συνδέσεις με ANSI-99 SQL, των συναλλαγών ACID, καθώς και τους υπολογισμούς MapReduce όπως υπολογισμούς.

4.9 Αναλυτικές Βάσεις, Λύσεις High Performance Computing & Massive Parallel Processing

Εικόνα 55.Βάσεις για Αναλύσεις Υψηλών Απαιτήσεων

IBM Netezza	
Ιστοθέση	https://www.ibm.com/products/netezza
Τύπος	Data Warehouse
	**Περιγραφή**
	Ολοκληρωμένη λύση υλικού και λογισμικού για αποθήκευση δεδομένων υψηλής απόδοσης και προηγμένες εφαρμογές ανάλυσης.

Kognitio WX2	
Ιστοθέση	https://kognitio.com/products/
Τύπος	In-memory Database
	**Περιγραφή**
	In-memory analytics πλατφόρμα βάσεων δεδομένων

SAP HANA	
Ιστοθέση	https://www.sap.com/products/hana.html
Τύπος	In-memory Database
	**Περιγραφή**
	Το SAP HANA είναι ένα σχεσιακό σύστημα διαχείρισης βάσεων δεδομένων στη μνήμη, οργανωμένο σε στήλες και αναπτύχθηκε και διατίθεται εμπορικά από τη SAP SE.Η κύρια λειτουργία του ως διακομιστή βάσης δεδομένων είναι η αποθήκευση και ανάκτηση δεδομένων όπως ζητείται από τις εφαρμογές. Επιπλέον, εκτελεί προηγμένα analytics(predictive analytics, επεξεργασία χωρικών δεδομένων, ανάλυση κειμένου, αναζήτηση κειμένου, αναλύσεις ροής, επεξεργασία δεδομένων γράφων) και περιλαμβάνει δυνατότητες εξαγωγής, μετασχηματισμού, φόρτωσης (ETL) καθώς και διακομιστή εφαρμογών.

GreenPlum	
Ιστοθέση	https://greenplum.org
Τύπος	NoSQL MPP/Shared-nothing database for massive parallel processing
	**Περιγραφή**
	Το Greenplum είναι μια μεγάλη τεχνολογία δεδομένων που βασίζεται στην αρχιτεκτονική MPP και στην τεχνολογία βάσης δεδομένων Postgres ανοιχτού κώδικα

Exasol	
Ιστοθέση	https://www.exasol.com/en/
Τύπος	In-memory
	**Περιγραφή**
	Exasol, ένα σχεσιακό σύστημα διαχείρισης βάσεων δεδομένων στη μνήμη, προσανατολισμένο σε αποθήκευση στηλών.

Vertica Analytics Platform	
Ιστοθέση	https://www.vertica.com
Τύπος	NoSQL Datawarehouse
	**Περιγραφή**
	Η πλατφόρμα Vertica Analytics είναι προσανοτολισμένη στην αποθήκευση στηλών και σχεδιάστηκε για τη διαχείριση μεγάλων, ταχέως αναπτυσσόμενων όγκων δεδομένων και παρέχει πολύ γρήγορη απόδοση ερωτημάτων όταν χρησιμοποιείται για αποθήκες δεδομένων και άλλες εφαρμογές που απαιτούν ένταση ερωτήματος. Η οργάνωση της αποθήκευσης σε στήλες, αρχιτεκτονική μαζικής παράλληλης επεξεργασίας (MPP), διασύνδεση με Standard SQL και πολλές άλλες ενσωματωμένες δυνατότητες ανάλυσης, machine learning επί βάσης δεδομένων, υψηλή συμπίεση, αρχιτεκτονική shared-nothing, αυτοματοποιημένη διαχείριση φόρτου εργασίας,βελτιστοποίηση ερωτημάτων και αποθήκευσης, υποστήριξη για τυπικές διεπαφές προγραμματισμού.

Βιβλιογραφικές Πηγές

1. **Chaowei Yang, Qunying Huang, Zhenlong Li, Kai Liu, Fei Hu**,*"Big Data and cloud computing: innovation opportunities and challenges"*, International Journal of Digital Earth, 2016

2. **Zira Karakaya** *"Software Engineering Issues In Big Data Application Development"* ,Department of Computer Engineering, Atilim University Ankara/Turkey, 2nd International Conference on Computer Science & Engineering 2017

3. **Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding**, *"Data Mining with Big Data"*, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 1, JANUARY 2014

4 .**NIST** Big Data Public Working Group Definitions and Taxonomies Subgroup, *"Big Data Interoperability Framework: Volume 1, Definitions"*,Version 2, June 2018

5. **NIST** Big Data Public Working Group Definitions and Taxonomies Subgroup, *"Big Data Interoperability Framework: Volume 7, Standards Roadmap"*, Version 3

6. **NIST** Big Data Public Working Group Definitions and Taxonomies Subgroup, *"Big Data Interoperability Framework: Volume 3, Use Cases and General Requirements",* Version 3

7. **NIST** Big Data Public Working Group Definitions and Taxonomies Subgroup, *"Big Data Interoperability Framework: Volume 6, Reference Architecture"*, Version 2, June 2018

8. **Dilpreet Singh, Chandan K Reddy**, *"A survey on platforms for big data analytics"*, Journal of Big Data 2014

9. **Shui Yu , Song Guo**, *"Big Data Concepts, Theories, and Applications"*, 2016, ISBN 978-3-319-27761-5

10. **Ian Gorton, John Klein**, Software Engineering Institute, "*Distribution,Data,Deployment Software Architecture Convergence in Big Data Systems",* IEEE COMPUTER SOCIETY BIG DATA SOFTWARE SYSTEMS 2015

11. **ORACLE**, "Big Data in Financial Services and Banking - Architect's Guide and Reference Architecture Introduction" | ORACLE ENTERPRISE ARCHITECTURE WHITE PAPER | FEBRUARY 2015

12. **Deloitte, Jurriaan Tressel, Guus van de Plasse**,*"Big data in the cloud"*, EMEA AIM Bootcamp 2017 Amsterdam, March 2017

13. **Deloitte**, *"Data Modeling for Big Data",* EMEA AIM Bootcamp, Amsterdam, March 2017

14. **Deloitte** , *"Design Big Data Architecture"*, EMEA AIM Bootcamp Amsterdam, March 2017

15. **Oliver Hummel, Holger Eichelberger, Andreas Giloj, Dominik Werle, Klaus Schmid**, *"A Collection of Software Engineering Challenges for Big Data System Development"*, 2018

16. **Pekka Pääkkönen, Daniel Pakkala,** *"Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems"* , Big Data Research , 2015

17. **Kshitij A Doshi, Tao Zhong, Zhongyan Lu ,Xi Tang ,Ting Lou, Gang Deng , "***Blending SQL and NewSQL Approaches ,Reference Architectures for Enterprise Big Data Challenges"* , International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery , 2013

18. **Trivadis Guido Schmutz**, *"Architecture of Big Data Solutions"* , Frankfurt 2017

19. **UNIVERSITY OF CALIFORNIA** *THE BERKELEY DATA ANALYSIS SYSTEM (BDAS):AN OPEN SOURCE PLATFORM FOR BIG DATA ANALYTICS"*, BERKELEY SEPTEMBER 2017

20.**Yuri Demchenko, Peter Membre** , *"Defining Architecture Components of the Big Data Ecosystem",* 2014

21. **MUTAZ BARIKA , SAURABH GARG,ALBERT Y. ZOMAYA,LIZHE WANG,AAD VAN MOORSEL,RAJIV RANJAN** , *"Orchestrating Big Data Analysis Workflows in the Cloud: Research Challenges, Survey, and Future Directions"* , ACM Computing Surveys, Vol. 52, No. 5, Article 95. Publication date: September 2019

22.**Cloud Standards Customers Council**, *"Cloud Customer Architecture for Big Data and Analytics V2.0"* , 2017

23.**Nathan Marz, James Warren**, *"Big Data: Principles and best practices of scalable real-time data systems"* ,Manning Publications Co.2015

24.**T. Ramalingeswara Rao , Pabitra Mitra, Ravindara Bhatt, A. Goswami**,*"The big data system, components, tools, and technologies: a survey"*, 2018

25. **Hsinchun Chen, Roger H. L. Chiang, Veda C. Storey**, *"BUSINESS INTELLIGENCE AND ANALYTICS: FROM BIG DATA TO BIG IMPACT"*, MIS Quarterly Vol. 36 No. 4, pp. 1165-1188/December 2012

26. **Andrew McAfee, Erik Brynjolfsson**, *"Big Data: The Management Revolution"*, Harvard Business Review October 2012

27. https://towardsdatascience.com/scalable-efficient-big-data-analytics-machine-learning-pipeline-architecture-on-cloud-4d59efc092b5

28. https://www.oreilly.com/radar/questioning-the-lambda-architecture/

29. https://databricks.com/glossary/lambda-architecture

30.https://towardsdatascience.com/a-brief-introduction-to-two-data-processing-architectures-lambda-and-kappa-for-big-data-4f35c28005bb

31. https://dzone.com/articles/what-is-big-data-architecture

32. https://docs.microsoft.com/en-us/azure/architecture/guide/architecture-styles/big-data

33. https://www.educba.com/business-intelligence-vs-big-data/

34. https://www.datanami.com/2012/09/17/big_data__scale_up_or_scale_out_or_both/

35. https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=310358

36. https://www.waitingforcode.com/general-big-data/massively-parallel-processing/read

37. https://towardsdatascience.com/big-data-integration-9a2fb2d78529

38. **Bas Geerdink ,**“*A Reference Architecture for Big Data Solutions: Introducing a model to perform predictive analytics of enterprise data, combined with open data sources, using big data technology*” ,MASTER’S THESIS ,Utrecht University of Applied Science ,August 2013

39. https://en.wikipedia.org/wiki/NewSQL

Πηγές Εικόνων

Εικόνα 0: Deloitte, “Data Modeling for Big Data”, EMEA AIM Bootcamp, Amsterdam, March 2017

Εικόνα 1-4: Deloitte,”Design Big Data Architecture”, EMEA AIM Bootcamp Amsterdam, March 2017

Εικόνα 5: NIST, Big Data Public Working Group Definitions and Taxonomies Subgroup, “Big Data Interoperability Framework: Volume 1, Definitions”,Version 2, June 2018

Εικόνα 6: T. Ramalingeswara Rao , Pabitra Mitra, Ravindara Bhatt, A. Goswami,”The big data system, components, tools, and technologies: a survey”, 2018

Εικόνα 7: NIST Big Data Public Working Group Definitions and Taxonomies Subgroup, “Big Data Interoperability Framework: Volume 6, Reference Architecture”, Version 2, June 2018

Εικόνα 8-11: Shui Yu,Song Guo,“Big Data Concepts, Theories, and Applications”,2016,ISBN 978-3-319-27761-5

Εικόνα 12-13: Deloitte,”Design Big Data Architecture”, EMEA AIM Bootcamp Amsterdam, March 2017

Εικόνα 14: Shui Yu,Song Guo,“Big Data Concepts, Theories, and Applications”,2016,ISBN 978-3-319-27761-5

Εικόνες 15-23: https://towardsdatascience.com/big-data-integration-9a2fb2d78529

Εικόνα 24: NIST Big Data Public Working Group Definitions and Taxonomies Subgroup, “Big Data Interoperability Framework: Volume 6, Reference Architecture”, Version 2, June 2018

Εικόνα 25: Pekka Pääkkönen, Daniel Pakkala, “Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems” , Big Data Research , 2015

Εικόνα 26-30: Yuri Demchenko,Peter Membre,“Defining Architecture Components of the Big Data Ecosystem”, 2014

Εικόνα 31. Deloitte,”Design Big Data Architecture”, EMEA AIM Bootcamp Amsterdam, March 2017

Εικόνα 32. Shui Yu,Song Guo,“Big Data Concepts, Theories, and Applications”,2016,ISBN 978-3-319-27761-5

Εικόνα 33. http://lambda-architecture.net

Εικόνα 34. https://www.oreilly.com/radar/questioning-the-lambda-architecture/

Εικόνα 35. https://www.omg.org/cloud/deliverables/CSCC-Cloud-Customer-Architecture-for-Big-Data-and-Analytics.pdf

Εικόνα 36-55: Matt Turck, Lisa Xu & FirstMark, **“**Data & AI Landscape 2019”,June 2019, https://mattturck.com/data2019/