
COMPUTING
edge
Data Storage
Data Visualization
Education
5G/6G
AUGUST 2022
www.computer.org
IEEE COMPUTER SOCIETY
IEEE

Get Published in the New *IEEE Open Journal of the Computer Society*

Submit a paper today to the premier new open access journal in computing and information technology.

Your research will benefit from the IEEE marketing launch and 5 million unique monthly users of the IEEE *Xplore*® Digital Library. Plus, this journal is fully open and compliant with funder mandates, including Plan S.

Submit your paper today!
Visit www.computer.org/oj to learn more.

IEEE COMPUTER SOCIETY | IEEE

COMPUTING edge

IEEE COMPUTER SOCIETY computer.org

STAFF

Circulation: *ComputingEdge* (ISSN 2469-7087) is published monthly by the IEEE Computer Society. IEEE Headquarters, Three Park Avenue, 17th Floor, New York, NY 10016-5997; IEEE Computer Society Publications Office, 10662 Los Vaqueros Circle, Los Alamitos, CA 90720; voice +1 714 821 8380; fax +1 714 821 4010; IEEE Computer Society Headquarters, 2001 L Street NW, Suite 700, Washington, DC 20036.

Postmaster: Send address changes to *ComputingEdge*-IEEE Membership Processing Dept., 445 Hoes Lane, Piscataway, NJ 08855. Periodicals Postage Paid at New York, New York, and at additional mailing offices. Printed in USA.

IEEE Computer Society Magazine Editors in Chief

Computer
Jeff Voas, *NIST*

Computing in Science & Engineering
Lorena A. Barba, *George Washington University*

IEEE Annals of the History of Computing
Gerardo Con Diaz, *University of California, Davis*

IEEE Computer Graphics and Applications
Torsten Möller, *Universität Wien*

IEEE Intelligent Systems
Longbing Cao, *University of Technology Sydney*

IEEE Internet Computing
George Pallis, *University of Cyprus*

IEEE Micro
Lizy Kurian John, *University of Texas at Austin*

IEEE MultiMedia
Shu-Ching Chen, *Florida International University*

IEEE Pervasive Computing
Marc Langheinrich, *Università della Svizzera italiana*

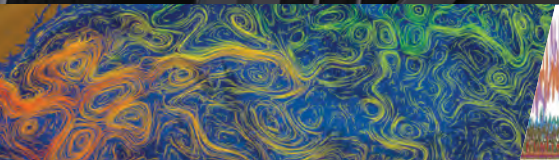
IEEE Security & Privacy
Sean Peisert, *Lawrence Berkeley National Laboratory and University of California, Davis*

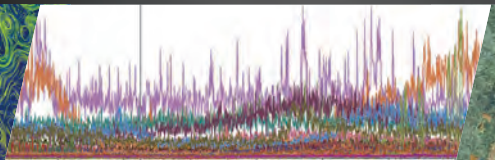
IEEE Software
Ipek Ozkaya, *Software Engineering Institute*

IT Professional
Irena Bojanova, *NIST*

AUGUST 2022 · VOLUME 8 · NUMBER 8

COMPUTING edge





Data Storage

Data Visualization

Education

5G/6G Systems

Departments

Magazine Roundup

The IEEE Computer Society's lineup of 12 peer-reviewed technical magazines covers cutting-edge topics ranging from software design and computer graphics to Internet computing and security, from scientific applications and machine intelligence to visualization and microchip design. Here are highlights from recent issues.

Computer

Changing the Paradigm of Control System Cybersecurity

Current cybersecurity protection relies on network monitoring. Changing the paradigm to monitor process sensors makes it practical to develop workable control system cybersecurity engineering solutions while simultaneously addressing reliability, safety, resilience, and productivity concerns. Read more in this article from the March 2022 issue of *Computer*.

Computing in Science & Engineering

Massively Parallel Particle Hydrodynamics at Exascale

In this article from the January/February 2022 issue of *Computing in Science & Engineering*, the authors introduce the work of the Massively Parallel Particle Hydrodynamics working group, part of the UK's ExCALIBUR software initiative. The aim of the group is to develop extensible software suitable for simulating complex hydrodynamics problems on exascale computing facilities using a Lagrangian particle-based approach. These methods complement mesh-based approaches and are particularly suited to problems with a large and fluid dynamic range or ones that involve free surfaces. The code uses fine-grained task parallelism to achieve a good load balance, when the workload varies greatly from fluid element to element.

IEEE Annals of the History of Computing

Russian Logics and the Culture of Impossible: Part I—Recovering Intelligentsia Logics

This article from the October–December 2021 issue of *IEEE Annals of the History of Computing* reinterprets algorithmic rationality by looking at the interaction between mathematical logic, mechanized reasoning, and computing in the Russian Imperial and Soviet contexts to offer a history of the algorithm as a mathematical object bridging the inner and outer worlds. The authors examine continuities between the turn-of-the-twentieth-century discussions of "poznaniye"—an epistemic orientation towards the process of knowledge acquisition—and the postwar rise of the Soviet school of mathematical logic.

IEEE Computer Graphics and Applications

Segmentation and Recognition of Offline Sketch Scenes Using Dynamic Programming

Sketch recognition aims to segment and identify objects in a collection of hand-drawn strokes. In general, segmentation is a computationally demanding process, since it requires searching through many possible recognition hypotheses. It has been shown that, if the drawing order of the strokes is known, as in the case of online drawing, a class of efficient recognition algorithms becomes applicable. In this article from the January/February 2022 issue of *IEEE Computer Graphics and Applications*, the authors introduce a method that achieves efficient segmentation and recognition in offline drawings by combining dynamic programming with a novel stroke ordering method. They demonstrate that the combined system is efficient and

 2469-7087/22

either beats or matches the state of the art in well-established databases and benchmarks.

IEEE Intelligent Systems

Incremental Computation in Dynamic Argumentation Frameworks

Dealing with controversial information is a challenging and important task for intelligent systems. Formal argumentation enables reasoning on arguments for and against a claim to decide on an outcome. An argumentation framework often models a dynamic situation where arguments as well as the way they interact frequently change over time. Consequently, the sets of accepted arguments (i.e., extensions under a given semantics) often need to be computed again after performing an update. In this article from the November/December 2021 issue of *IEEE Intelligent Systems*, the authors address the problem of efficiently recomputing extensions of dynamic argumentation frameworks. They present an incremental algorithmic solution whose main idea is that of using an initial extension and the update to identify a (potentially small) portion of the argumentation framework, which is sufficient to compute an extension of the whole updated framework.

IEEE Internet Computing

Quantum Information Science

As classical computational infrastructure becomes more limited, quantum platforms offer expandability in terms of scale, energy consumption, and native 3D problem modeling. Quantum information science is a multidisciplinary field drawing from physics, mathematics, computer science, and photonics. Quantum systems are expressed with the properties of superposition and entanglement, evolved indirectly with operators (ladder operators, master equations, neural operators, and quantum walks), and transmitted (via quantum teleportation) with entanglement generation, operator size manipulation, and error correction protocols. This article from the January/February 2022 issue of *IEEE Internet Computing* discusses emerging applications in quantum cryptography, quantum machine learning, quantum finance, quantum neuroscience, quantum networks, and quantum error correction.

IEEE micro

Artificial Intelligence Best Practices in Smart Agriculture

Smart agriculture, with the aid of artificial intelligence (AI), is playing a pivotal role in ensuring agriculture sustainability. AI techniques are employed in soil and irrigation management, weather forecasting, plant growth, disease prediction, and livestock management, which are considered significant domains of agriculture. The authors of this article from the January/February 2022 issue of *IEEE Micro* review recent AI techniques that have been deployed in these domains. They focus on the various AI algorithms used as well as their performance impact. This review not only highlights the effective use of AI at different layers of a smart agriculture architecture, but also identifies future research directions in this field.

IEEE MultiMedia

Are Remote Play Streaming Systems Doomed to Fail? A Network Perspective

Digital games represent one of the most compelling fields in computer science, embodying a wide variety of technical challenges. Thanks to the evolution of streaming and broadband technology, new service provisioning schemes have emerged. Remote play streaming services represent an interesting case study deserving a thorough investigation. To this end, the authors of this article

from the October–December 2021 issue of *IEEE MultiMedia* present a network measurement study that can be useful to create traffic models and help researchers identify issues, guiding architecture, and protocol design. Moving beyond latency and jitter issues, the purpose is to understand whether remote play streaming services can operate through regular connectivity or, on the contrary, are doomed to fail as happened to some pioneer providers.

IEEE Pervasive Computing

Obtaining Labels for In-the-Wild Studies: Using Visual Cues and Recall

The observer effect found in laboratory studies has long posed a problem for researchers. In-the-wild studies reduce the observer effect but have problems with gathering accurately labeled data that is usable for training algorithms. Manual labeling is time-consuming, obtrusive, and unfeasible, and, if done by the researchers, it potentially violates the privacy of the participants. In this article from the January–March 2022 issue of *IEEE Pervasive Computing*, the authors present a labeling workflow based on an in-the-wild study that investigated cognitive state changes through eye-gaze in naturalistic settings. They contribute a setup that enables participants to label their data unobtrusively and quickly. They use J!NS MEME electrooculography glasses, Narrative Clip 2 wearable cameras, and a proprietary data-tagging software package. The setup is reproducible for field studies, preserves data integrity, and maintains participant privacy.

IEEE Security & Privacy

Automated Privacy Preferences for Smart Home Data Sharing Using Personal Data Stores

Personal data stores (PDSs) are decentralized, user-centric data storage and processing environments for implementing privacy-aware smart home data storage. In this article from the January/February 2022 issue of *IEEE Security & Privacy*, the authors' privacy preference recommender system works with PDSs to assist users in making data-sharing decisions to avoid unintended privacy mishaps.

IEEE Software

Hybrid Digital Twins: A Primer on Combining Physics-Based and Data Analytics Approaches

Two popular approaches to building digital twins are pure data-based and physics/simulation-based methods. In this article from the March/April 2022 issue of *IEEE Software*, the authors present a framework for hybrid digital twins that combines the strengths of the two approaches, sharing results and demonstrating applicability to a flow network.

IT Professional

Makers' Studio: Enabling Education and Skill Development Through ICT

Education has been evolving with advancements in information and communication technologies (ICT). The pace of this modernization is determined by the requirements of the society and the technological developments. Although the online mode of education has existed for four decades, an abrupt shift to completely online mode during COVID-19 exposed the lack of proper infrastructure and technological solutions. This article from the January/February 2022 issue of *IT Professional* contributes to filling in this gap and proposes a community-based approach of learning via the establishment of Makers' Studio, where all the stakeholders of the academic community will have contributions and takeaways.

Editor's Note

Big Data Storage

Big data has the power to unlock insights and discoveries in many domains, but it can be challenging to efficiently handle vast amounts of heterogeneous structured and unstructured data. Organizations are increasingly seeking cost-effective, scalable, and flexible data storage solutions. This *ComputingEdge* issue explores the trends and innovations in data storage technology that are allowing organizations to leverage big data.

"It's Time to Talk About HPC Storage: Perspectives on the Past and Future," from *Computing in Science & Engineering*, covers developments in high-performance computing storage system architectural designs that support concurrent and low-latency access to massive volumes of scientific data. "Big Data: Present and Future," from *Computer,* discusses new strategies for storing big data, such as data lakes and NoSQL databases.

Visualization is helpful and sometimes essential for analyzing data and making evidence-based decisions. In *IEEE Computer Graphics and Applications*' "BubbleUp: Supporting DevOps With Data Visualization," the author reports on a tool for rapidly analyzing complex data from distributed systems. In *IT Professional*'s "Rigorous Data Validation for Accurate Dashboards: Experience From a Higher Education Institution," the authors describe a suite of visualization dashboards used at California State University, Fullerton to promote a data-driven operational culture.

Information technology education is critical to preparing today's young people for the workforce and for life beyond school. The authors of *IEEE Security & Privacy*'s "Designing a K-16 Cybersecurity Collaborative: CIPHER" propose a framework for addressing the challenges of creating a pipeline of qualified and diverse cybersecurity professionals. The author of *IEEE Pervasive* Computing's "Beyond Bots and Buttons—New Directions in Information Literacy for Students" suggests interactive methods for teaching misinformation awareness and media literacy.

This *ComputingEdge* issue closes with one article about 5G and 6G cellular networks and systems. Computer's "An Attack Vector Taxonomy for Mobile Telephony Security Vulnerabilities" presents a novel scheme for categorizing threats to 5G networks.

EDITORS: Kathryn Mohror, mohror1@llnl.gov
John M. Shalf, JShalf@lbl.gov

This article originally appeared in
Computing in Science & Engineering
vol. 23, no. 6, 2021

DEPARTMENT: LEADERSHIP COMPUTING

It's Time to Talk About HPC Storage: Perspectives on the Past and Future

Bradley Settlemyer, *Los Alamos National Laboratory, Los Alamos, NM, 87545, USA*
George Amvrosiadis, *Carnegie Mellon University, Pittsburgh, PA, 15213, USA*
Philip Carns and Robert Ross, *Argonne National Laboratory, Lemont, IL, 60439, USA*

High-performance computing (HPC) storage systems area key component of the success of HPC to date. Recently, we have seen major developments in storage-related technologies, as well as changes to how HPC platforms are used, especially in relation to artificial intelligence and experimental data analysis workloads. These developments merit a revisit of HPC storage system architectural designs. In this article, we discuss the drivers, identify key challenges to status quo posed by these developments, and discuss directions future research might take to unlock the potential of new technologies for the breadth of HPC applications.

High-performance computing (HPC) storage systems have become trusted repositories for hundreds of petabytes of data with aggregate throughput rates in the terabytes per second. Numerous research advances have contributed to this success. Object storage technologies helped eliminate bottlenecks related to the management of space on storage devices. The development of separate data and metadata planes facilitated scale-out in the data plane to enable high throughput. The adoption of network portability layers eased porting to new HPC networking technologies. Disaggregation was adopted early, bringing powerful cost and administrative savings and providing flexibility to serve the diverse batch workloads typical of HPC. Together with input/output (I/O) middleware technologies, HPC storage systems have largely addressed the throughput challenges of checkpoint and restart for traditional message passing interface simulation codes, which was their primary driver for many years.

Meanwhile, HPC applications have evolved from numerical simulations to workloads that include artificial intelligence (AI) and analytics. For example, scientists at the Oak Ridge National Laboratory (ORNL) Health Data Sciences Institute are developing AI-based natural language processing tools to extract information from textual pathology reports using Summit, the USA's most powerful supercomputer, due to the vast amounts of memory it provides to its compute cores. Similarly, the High Luminosity Large Hadron Collider (HL-LHC) will further extend the capabilities of the LHC, allowing further investigation of phenomena fundamental to the nature of the universe. To be installed in 2025, these enhancements will lead to annual data generation rates of tens of petabytes, with reduced datasets in the petabyte range being used for analysis. These applications are often read-intensive, and may rely on latency-sensitive transfers, each consisting of small amounts of data. This marks a dramatic shift in how HPC storage systems are used. While some emerging read-intensive workloads may be able to rely on structuring within the data to construct efficient data retrieval plans based on caching or prefetching techniques, AI workloads and many data analytics routines are inherently required to access the data without any predictable ordering. According to the Department of Energy's 2020 AI for Science report:[1]

> *"AI training workloads, in contrast, must read large datasets (i.e., petabytes) repeatedly and perhaps noncontiguously for training. AI models will need to be stored and dispatched to inference engines, which may appear as small, frequent, random operations."*

1521-9615 © 2021 IEEE
Digital Object Identifier 10.1109/MCSE.2021.3117353
Date of current version 16 December 2021.

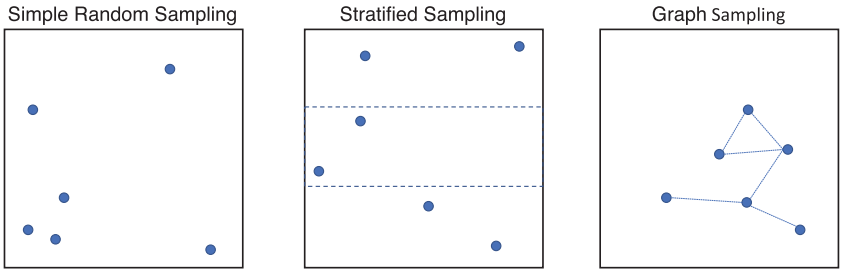


FIGURE 1. Examples of prevalent analysis access patterns: Data-intensive analysis algorithms (propelled by breakthroughs in AI and statistical methods) must extract samples from immense data sets, thereby triggering storage access patterns that *are unpredictable to outside observers.* These workloads put pressure on the storage system's random read input/output operations per second (IOPS) rate and response time in ways that cannot be solved with general purpose caching and prefetching.

Figure 1 shows examples of prevalent access patterns for analytics, which are characterized by this lack of ordering.

Two technology trends have emerged as crucial to data-driven scientific discovery. First, the high-speed networks used within scientific computing platforms provide extremely low-latency access to remote systems, including billions of message injections per second and direct access to remote system memory via remote direct memory access (RDMA) operations. Second, solid-state disks (SSDs) accessed through the non-volatile memory express (NVMe) interface provide more than 1,000 times the performance of traditional hard disk drives for the small random reads used within data-intensive workloads. Interestingly, while HPC storage systems broadly leverage both high-speed networks and SSDs, this adoption was not driven by the need to provide low-latency access to remote storage, but by simulation requirements for fast point-to-point communication between processes and high-throughput requirements for access to HPC storage systems. With the advent of new read-heavy analysis workloads, however, low-latency remote storage access is now also a key enabling technology for new data-driven approaches to computational science.

The evolution of HPC workloads has highlighted previously hidden shortcomings of modern storage systems. This is due to storage architectures that emerged in the early 2000s and remained static using storage servers with designated metadata and data roles, tightly attached to the HPC network, and focused on delivering throughput while relying on software layers that hid latency issues. Traditionally, system architects have relied on the increase of CPU frequencies and scaling out to prevent latency from affecting application performance. In recent years, however, CPUs have increased their computational power through the addition of CPU cores with decreasing frequencies that complicate real-time event processing. Scaling out to meet the IOPS requirements that modern HPC workloads place on the storage system is problematic as well. Balancing work to keep storage and network capability fully utilized is difficult at scale, resulting in underutilized resources and a higher total cost of ownership.

EMERGING CHALLENGES IN HPC STORAGE

Modern HPC storage architectures were shaped by the performance characteristics of conventional hard drives. Conventional hard drives exhibit minimal (if any) onboard processing capability, low random access performance, and high latency. These characteristics placed a ceiling on overall storage system performance, and the remainder of the storage infrastructure was designed around mitigating their limitations as much as possible. Specifically, storage servers were designed to mediate all access to hard drives. By doing so, they could shape traffic (e.g., by serializing and batching), buffer data (e.g., through caching based on locality), and process I/O requests on more powerful host CPUs (e.g., by handling interrupts, packing and unpacking remote procedure call (RPC) requests, and enforcing authorization) to make the most of hard drive capabilities. Hard drive access latency also had subtle implications for other elements of the storage system; there was no incentive to avoid latencies in the client-side operating system or the storage fabric as long as hard drives gated overall performance (see Figure 2).

Low-Latency Access to Storage

The architectural approach shown in Figure 2 was successful: it allowed HPC storage systems to extract maximum aggregate throughput from vast arrays of

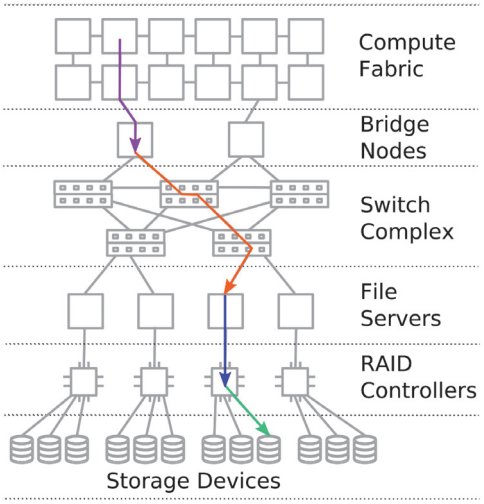


FIGURE 2. Exemplar disaggregated HPC storage architecture. Traditional HPC storage systems have been propelled by simulation workloads to optimize for aggregate bulk synchronous throughput. This is a key disconnect for data-driven analysis: systems designed to maximize aggregate throughput are poorly suited to individual random reads. Each access must traverse multiple distinct protocol hops, where each protocol hop has its own interrupt processing, buffering, handshaking, serialization, and access control conventions. These protocol translations were designed in an era when high-latency storage devices gated overall performance, an assumption that no longer holds today.

commodity hard drives. Limitations are evident, however, now that we attempt to match emerging IOPS and response-time-sensitive workloads to more capable low-latency storage devices. User-space APIs such as libaio or liburing can issue millions of operations per second from a single core, network interface cards can inject hundreds of millions of messages into a network per second, and these rates can be matched by just a few hundred NVMe storage devices. Despite these capabilities, modern storage servers are only able to process 100,000 RPCs per second from a single core. Even an incredibly high-end storage server with 100 high-frequency cores could service only 10 million read or write RPCs per second. Such performance strands over 90% of the network interface capability and saturates fewer than 10 fast NVMe devices.

In other words, the host-based RPC processing that in the past served to optimize access to storage devices has now become a hindrance. The server's ability to deserialize and process an RPC request and then serialize and send an RPC response is now the gating factor in the IOPS rate. The fastest RPC libraries, co-designed with high-performance interconnects and performing no server-side processing, have been unable to achieve even 500,000 RPCs per second per core. The traditional HPC solution of scaling out to achieve higher IOPS is inefficient; expanding the number of server CPU cores will increase complexity, footprint, and power demands, offer diminishing returns on aggregate IOPS rate, and effect no improvement in response time for individual accesses. The classic HPC storage architecture must now be revisited in the context of mixed workloads and the widespread availability of low-latency hardware components.

THE CLASSIC HPC STORAGE ARCHITECTURE MUST NOW BE REVISITED IN THE CONTEXT OF MIXED WORKLOADS AND THE WIDESPREAD AVAILABILITY OF LOW-LATENCY HARDWARE COMPONENTS.

Scaling and Maintaining Low Latency

Science teams driving these data-intensive activities are pushing the scalability of their computations just as teams with simulation codes have before them, and it is paramount that storage systems support that scalability. Traditional caching and prefetching are not generally effective for these algorithms, eliminating a common option for accelerating access. On the other hand, the HPC networking community has learned much that can be applied to next-generation storage systems. Limiting the state associated with connections is an important enabler for scale-out, especially when there is no obvious structure in the communication as there is in many scientific codes.

Devices supporting protocols that require connection establishment are incredibly challenging to employ at the HPC scale, but unfortunately, that is the current direction of network-accessible device protocols such as NVMe-oF. Connectionless models of communication have been demonstrated in HPC[2] and supported in production hardware:[3] it is up to HPC to invent the fast, direct access to remote storage devices that will be a key enabling technology for scalable storage systems. HPC platforms have similarly been at the leading edge of requirements for high-concurrency, low-latency access to remote memory, and extending proven techniques to enable similarly

parallel and low-latency access to storage is a natural research direction. Alterations and alternatives to existing data-transport methods for storage—perhaps built using compute-enabled devices—should be investigated and their potential demonstrated. User-land access to resources has also been shown as critical for maintaining low latencies, which will be critical in the data plane if not also in at least some aspects of the metadata plane. Approaches along these lines have begun to be explored in the larger storage community[4] but must be adapted to the scales and networks of HPC.

Securing Access to Storage Devices

In addition to providing efficient access to storage devices, storage system software is also tasked with providing access control to the data stored within high-performance storage systems. In the current server-mediated access to the storage model, the system software is tasked with enforcing all data access controls. As we move to a storage access paradigm that supports faster, low-latency access to storage devices, a server-mediated access control scheme becomes a bottleneck that paralyzes emerging workloads rather than acting as a useful enforcement mechanism. At the same time, storage devices have gained richer interfaces and capabilities, including zoned namespaces (ZNS) and embedded functions in the form of computational storage, and thus, it is clear that security models that treat storage devices as only a repository for stored data are obsolete.

More direct access to storage devices from large numbers of client processes, which may include user-space access to remote storage devices, must provide new models of security not currently provided by either the network protocols or storage devices. While the NVMe standards body has defined multiple methods for securely accessing storage, none of these mechanisms are currently a good match for data-intensive scientific discovery. The two most common NVMe security methods, in-band authentication and per-request security, are focused on ensuring that clients are authenticated with servers but cannot differentiate between data plane operations that read data or write data and control plane operations that create or destroy on-device namespaces. And while key-per-IO is a novel model that enables every disk access to be secured separately, the overheads of checking an encryption key for every operation is antithetical to low-latency access to storage devices. Instead, new security models that expose the performance advantages of ZNS[5] and leverage scalable approaches to embedded compute, such as computational storage and SmartNICs,[6] require additional research.

ENABLING A FUTURE FOR DATA-DRIVEN SCIENCE

A great deal of effort was required to stabilize HPC storage and make it trustworthy, but it did happen. Multiple production file system options exist for data centers to choose from, and checkpoint and restart for HPC codes has largely been addressed. But storage system designers cannot rest on their laurels, and storage is not a solved problem. Even more than for simulation codes, the potential benefits of HPC for AI and analysis applications hinge on high-performance storage. We need not just innovation, but innovation that goes hand in hand with these scientific objectives.

ARCHITECTURALLY, THE COMMUNITY MUST REVISIT THE DATA PATH BETWEEN ANALYSIS APPLICATIONS AND STORAGE DEVICES.

Architecturally, the community must revisit the data path between analysis applications and storage devices. In much the same way that user-space RDMA access has revolutionized HPC networking (removing handshaking, buffering, and host processing from the interprocess communication path) and allowed networks to keep pace with memory throughput, we must adopt new HPC storage access paradigms that minimize obstructions in the storage data path and allow storage systems to keep pace with NVMe capabilities. The need for RPC processing can be minimized (by thoughtful partitioning of work to control planes), any remaining RPC processing or asymmetric transfer can be offloaded to smart devices, and the complete data path can be holistically evaluated to eliminate duplicate and superfluous protocol translations that collectively leach latency from the system.

From a device interface perspective, storage systems traditionally divide responsibility between the storage device and host rigidly: the device is responsible for handling data block updates, and the host is responsible for data processing. But as SSDs continue to replace hard disks at the front-line storage

tier, block interface support requires complex firmware that affects device performance and cost, and as storage becomes disaggregated from computation, reducing data movement between the device and host becomes crucial. Novel interfaces, like zoned storage, have emerged to reduce firmware complexity by delegating responsibilities to the host, and computational storage allows data to be processed on the device in accordance to application needs, blurring the divide between device and host. Future work will need to adapt popular application types to fully leverage the capabilities of these devices and explore the right balance of near-storage computation for different tasks.

User abstractions are another key piece of the puzzle. Building fast and productive storage systems will require not only addressing these technology challenges but also understanding emerging science needs. The HPC storage community has contributed interface advances in the past, including concepts eventually adopted in the mainstream, but recently most storage abstraction innovation has occurred elsewhere, with cloud service providers offering options such as column stores, document stores, key-value stores, streaming data infrastructure, and object stores. HPC storage researchers must work together with technology providers and domain scientists to find abstractions that match science needs and then to develop scalable storage services embodying those abstractions.

The HPC storage research community also a needs to be reinvigorated. A misconception persists that HPC storage is a solved problem: new storage systems are iteratively designed and deployed by solving formulas based on commodity market forces and logistical constraints. In reality, however, many unsolved problems remain in high-performance storage, especially as high-performance storage comes to the forefront as the key to enabling both simulation and data-driven analytics use cases. The high-performance storage community must innovate within this space and then translate those innovations into solutions for our data-driven science partners. HPC storage and its workloads must become first-class citizens within computer science curricula, coordinated research thrusts, and partnerships between industry, academia, and governments.

CONCLUSIONS

The push to achieve the largest and most complex scientific discoveries using HPC requires heroic efforts from computational scientists, computing system designers, and software developers. But critically, these tremendous efforts have proven to successfully flow downstream and make equally important, but less computationally demanding, scientific discoveries tractable. By design, a calculation that was entirely heroic a decade ago can now be achieved by a handful of highly motivated graduate students. To usher in this same downstream effect for data-driven science, a set of sustained and heroic efforts are needed for building and operating storage systems that can support highly concurrent and low-latency access to massive volumes of scientific data. With this key underpinning under development and then in use, we enable the additional efforts needed to extract new insight and invent new methods for accelerating data-driven scientific discovery. And in several years, as the benefits of new methods for analyzing data are realized and made commonplace, small teams of highly motivated graduate students will perform data-driven searches for discovery that could not be dreamed of as possible within contemporary HPC data centers. The road ahead, and its inevitable roadblocks and detours, will be difficult and surprising, but the rewards at the end of this journey are too great to resist.

ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research, under Contract DE-AC02-06CH11357.

REFERENCES

1. R. Stevens, V. Taylor, J. Nichols, A. B. Maccabe, K. Yelick, and D. Brown, "AI for Science," Office of Scientific and Technical Information (OSTI), Oak Ridge, TN, USA, Tech. Rep. ANL-20/17 158802, 2020, doi: 10.2172/1604756.
2. B. Barrett *et al.*, "The Portals 4.0 network programming interface," *Sandia Nat. Lab.*, Albuquerque, NM, USA, Tech. Rep. SAND2012-10087 392902, 2012.
3. S. Derradji, T. Palfer-Sollier, J. Panziera, A. Poudes, and F. W. Atos, "The BXI interconnect architecture," *Proc. IEEE 23rd Annu. Symp. High-Perform. Interconnects*, 2015, pp. 18–25.
4. Y. Chen *et al.*, "Scalable persistent memory file system with kernel-userspace collaboration," *Proc. 19th {USENIX} Conf. File Storage Technol.*, 2021, pp. 81–95.
5. M. Bjørling *et al.*, "ZNS: Avoiding the block interface tax for flash-based SSDs," in *Proc. USENIX Annu. Tech. Conf.*, 2021, pp. 689–703.
6. H. Li *et al.*, "Leapio: Efficient and portable virtual NVME storage on arm SOCS," in *Proc. 25th Int. Conf. Architectural Support Program. Lang. Operating Syst.*, 2020, pp. 591–605.

BRADLEY SETTLEMYER is currently a Senior Scientist in HPC Design Group, Los Alamos National Laboratory, Los Alamos, NM, USA. He currently leads the storage systems research efforts within Los Alamos' Ultrascale Research Center and his team is responsible for designing and deploying state-of-the-art storage systems for enabling scientific discovery. He is the primary investigator on projects ranging from ephemeral file system design to archival storage systems using molecular information technology. He has authored or coauthored papers on emerging storage systems, long-distance data movement, system modeling, and storage system algorithms. Dr. Settlemyer received the Ph.D. degree in computer engineering from Clemson University, Clemson, SC, USA, in 2009, with a research focus on the design of parallel file systems. Contact him at bws@lanl.gov.

GEORGE AMVROSIADIS is currently an Assistant Research Professor of electrical and computer engineering and, by courtesy, computer science at Carnegie Mellon University, Pittsburgh, PSA, USA, and a member of the Parallel Data Laboratory. He co-teaches courses on cloud computing and storage systems. His current research interests include distributed and cloud storage, new storage technologies, high-performance computing, and storage for machine learning. Dr. Amvrosiadis received the Ph.D. degree in computer science from the University of Toronto, Ontario, Canada, in 2016. Contact him at gamvrosi@cmu.edu.

PHILIP CARNS is currently a Principal Software Development Specialist in the Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, IL, USA. He is also an adjunct Associate Professor of electrical and computer engineering at Clemson University, Clemson, SC, USA, and a Fellow of the Northwestern-Argonne Institute for Science and Engineering. His research interests include characterization, modeling, and development of storage systems for data-intensive scientific computing. Dr. Carns received the Ph.D. degree in computer engineering from Clemson University in 2005. Contact him at carns@mcs.anl.gov.

ROBERT ROSS is currently a Senior Computer Scientist at Argonne National Laboratory, Lemont, IL, USA, and a senior fellow at the Northwestern–Argonne Institute for Science and Engineering, Northwestern University, Evanston, IL, USA. His research interests include system software and architectures for high-performance computing and data analysis systems, in particular storage systems and software for I/O and message passing. He was the recipient of the 2004 Presidential Early Career Award for Scientists and Engineers. Dr. Ross received the Ph.D. degree in computer engineering from Clemson University, Clemson, SC, USA, in 2000. He is the corresponding author of this article. Contact him at rross@mcs.anl.gov.

EDITORS: Norita Ahmad, American University of Sharjah, nahmad@aus.edu
Preeti Chauhan, VP Technical Activities, IEEE Reliability Society, preeti.chauhan@ieee.org

This article originally appeared in
Computer
vol. 54, no. 4, 2021

DEPARTMENT: DATA

Big Data: Present and Future

Preeti Chauhan, *IEEE Senior Member*
Mohit Sood, *University of California, Berkeley*

Big data in the 21st century will impact every individual, organization, and government. Organizations must invest in big data tools for business growth and efficiency while protecting data privacy as we continue toward digitization and datafication.

The term *big data* was first referenced in 1997 in an article by Michael Cox and David Ellsworth in the ACM digital library. The article discusses the challenges of visualization due to large data sets requiring high memory capacity. The authors referred to this as the problem of big data. Later in 2001, Doug Laney, an analyst with the Meta Group, published an article on data management with the "3Vs": volume, velocity, and variety; these terms went on to become the most commonly accepted definitions of big data. Over the years, variability and value were added as other key attributes of big data. In general terms, big data is a large complex set of data that requires additional computation to extract, analyze, and process to drive decision making.

Big data is broadly categorized into three data types: structured, unstructured, and semistructured. Structured data sets are made of clearly defined data types, which makes them easy to search and organize in relational databases. Some examples of structured data are phone numbers, street addresses, and Social Security numbers. Transactional data are another type of structured data and consist of, for example, sales orders, payments, returns, refunds, invoices, purchase orders, inventory-level changes, shipping documents, passport applications, credit card payments, and insurance claims.[1]

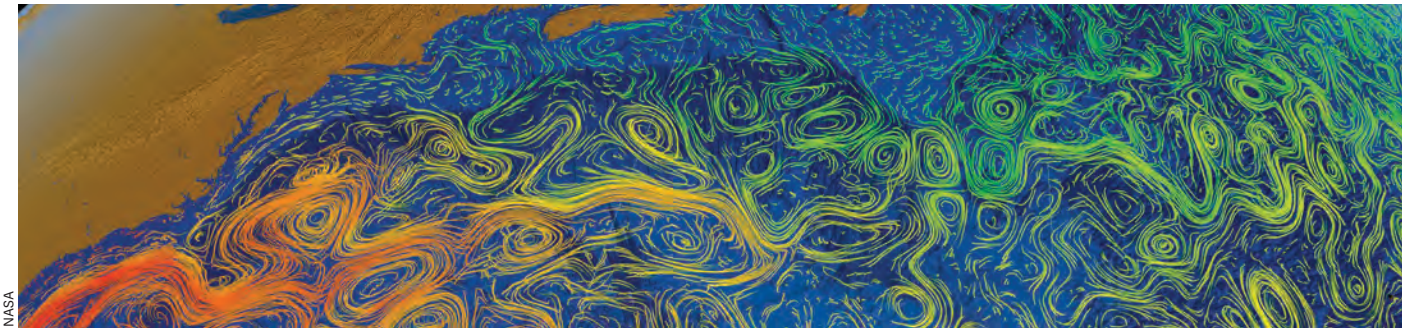
Unstructured data sets are those that are not easily searchable nor stored in structured database format. Such data could be textual, audio, or video and both human and machine generated. Human-generated unstructured data include audio and video data shared on YouTube, Instagram, Facebook, Twitter, and so on. Machine-generated unstructured data include satellite imagery, sensor, and digital surveillance data. Today, only 20% of the existing data are classified as unstructured, but at a 62% growth rate per year, by 2020, unstructured data will form 93% of the data sets.[2]

Lastly, semistructured data are a type that cannot be organized in relational databases, and they do not have a strict structural framework. However, they still have some structural properties. An example would be emails, which can be categorized based on, for instance, sender, subject, recipient, and send date and hence are structured. However, the content in emails falls under the unstructured data type, making emails a semistructured data set overall.

Big data is being created at an astounding pace to say the least. The total volume of data created, captured, copied, and consumed worldwide is expected to be 149 ZB in 2024, up almost two orders of magnitude from 2 ZB in 2010[3] (Figure 1). To put the size in perspective, it would take 181 million years if all of the existing data were to be downloaded. This explosion in data is a result of two macrotrends: an increase in Internet users and Internet-of-Things (IoT)-connected devices and a decrease in data storage and analysis costs, which came about due to a reduction in semiconductor computing (CPUs, GPUs, accelerators, and so on) and storage (memory) costs and advancements in network connectivity in the last two decades. Today,

Digital Object Identifier 10.1109/MC.2021.3057442
Date of current version: 9 April 2021

 Published by the IEEE Computer Society



NASA

there are 4.7 billion Internet users[4] in the world, who are increasingly consuming and creating content on social media, search engines, online entertainment, and news. The Internet users generate about 6,123 TB of traffic every minute, which includes 185 million sent emails, 5.2 million Google searches, 305k Skype calls, and 84,000 photos uploaded on Instagram.[5] Apart from Internet usage, IoT-connected devices (sensors, smart cars, and so on) are expected to increase from 26.7 billion devices in 2020 to 75 billion devices in 2025.[6]

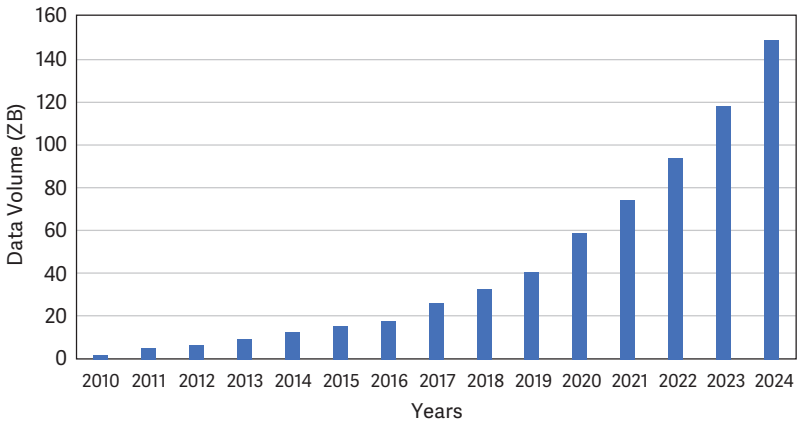


FIGURE 1. Big data volume growth from 2010 to 2024.[3] (Data source: Statistica.)

The cloud-computing industry has grown in lockstep with big data in the last two decades, with significant advances in data extraction, storage, and predictive and prescriptive analysis tools, including artificial intelligence (AI) and machine learning (ML). The global cloud-computing market size is expected to grow from US$371.4 billion in 2020 to US$832.1 billion by 2025, at a compound annual growth rate of 17.5%.[7] This growth rate is further expected to increase with accelerated cloud technology adoption by enterprises in sectors where the work-from-home initiative is helping to sustain enterprise business functions.

APPLICATIONS

Big data in today's world has tremendous potential to provide insights into almost all aspects of our lives, enabling smart decision making, cost reductions, future predictions, production-throughput improvements, and new product offerings. It ranges from providing personalized recommendations for best places to shop or eat based on user history, to playing a pivotal role for health agencies in managing the COVID-19 pandemic through contact tracing and hospital availability analysis. It is also widely accepted that companies focusing on big data analytics to create business values will succeed. This requires both strategic design and a well-thought-out architecture that can utilize the available data streams to meet specific business objectives, determine customer behavioral and usage patterns, and predict market trends.

THE GLOBAL CLOUD-COMPUTING MARKET SIZE IS EXPECTED TO GROW FROM US$371.4 BILLION IN 2020 TO US$832.1 BILLION BY 2025, AT A COMPOUND ANNUAL GROWTH RATE OF 17.5%.

Big data has enabled faster data processing and cost reduction by switching to cloud-based analytics, thus reducing the hardware and associated infrastructure needed for data storage and processing. Faster processing and internalization of complex data has enabled businesses to assess their competitors more quickly against their own offerings to make decisions in a continuously evolving business environment. The

predictive capabilities offered by AI and ML have led companies to make future projections further out with higher accuracy and consistency.

The AI-powered user-pattern analysis from polls, surveys, Internet shopping, search history, location data, and so on has enabled the prediction of future human behavior and the provision of personalized recommendations for shopping, travel, restaurants, politics, weather, and even health. It is now possible to assess customer wants along with resulting satisfaction levels to deliver the right product. This has led many big companies to continuously innovate and launch customer-centric products on a consistent basis. Some of the industries benefiting from big data are health care, banking, media, retail, and energy. Other industries, such as medicine, construction, and transportation, are moving fast toward adopting and integrating big data analytics in their day-to-day operations and decision making.

ON THE MEDICAL FRONTIER, BIG DATA CAN BE FOUND AT THE LEADING EDGE OF THERAPEUTIC AND DIAGNOSTICS RESEARCH.

The health-care industry benefits from big data by enabling the removal of redundant diagnoses from medical records, the early detection of diseases, and the ability to prevent virus outbreaks. The data have structured elements, like a patient's personal information and vitals, and unstructured elements, like X-ray and ultrasound records. Data are typically obtained from patient records or user generated from devices such as health apps,[8] Apple Watches, and Fitbits. The data are analyzed to help hospitals assess the effectiveness of therapies and drug administrations to improve future treatment plans. The most recent example is how health-care agencies and governments were able to do contact tracing for COVID-19 to follow the spread of the pandemic and help regulate social-distancing and shelter-in-place orders.

On the medical frontier, big data can be found at the leading edge of therapeutic and diagnostics research. For example, DeepMind, Google's deep-learning program, made a huge leap through its AlphaFold program to successfully determine the 3D shapes of proteins from their amino-acid sequence and solve a huge, decades-old challenge in biology. AlphaFold solved this problem after getting training on big data comprising approximately 170,000 protein structures.[9] Similarly, AI and big data platforms now provide the capability to sift through years of data to identify possible drugs that are already approved for treating certain diseases and help identify new molecules using this database to accelerate vaccine development.[10]

Although the banking sector has stringent data security regulations and has been relatively slow in adopting innovations, big data has started to play an important role in the banking business. The applications span fraud detection, customer behavior-pattern prediction, market trend detection, improved trade execution, and enhanced customer experience. Banks make use of both structured data (such as demographics, credit scores, and transaction types) and unstructured data (macrodevelopments, geopolitical news, and so on) to grow business and enhance client experience.

With the advent of online streaming, big data analytics has played a key role in driving the growth of the entertainment and media industry worldwide. Netflix, a popular streaming service, has experienced astronomical growth, with subscribers increasing from 21 million to 197 million globally in the last decade.[11] More recently, Disney's video-on-demand channel gained 74 million subscribers within a year of launch.[12] This growth has been due to both good content and big data. These platforms collect an incredible amount of data while their services are being used, to provide personalized recommendations for genres of shows and movies and improve customer engagement. Big data is also used to make decisions on which scripts or shows to produce or license by predicting viewership based on the content and performance of similar shows or movies in the past.[13]

Big data has also revolutionized the retail business forever. Online retail giants such as Amazon and Alibaba use big data to sift through millions of seller options in every product category to provide their users with enhanced experiences. Big data is also used in customer relationship management by almost 90% of businesses to enhance customer experience

and increase sales.[14] Another example is Starbucks, which had harnessed big data from 30 million mobile app users and 20 million loyalty program members at the end of 2020.[15,16] The introduction of in-app payments, which provide valuable data about customer preferences, enables Starbucks to provide targeted offerings and rewards to its customers and increase overall sales.

TOOLS AND KEY PLAYERS

Organizations in general have to ingest both structured and unstructured data generated from disparate sources. Given this heterogeneity of big data, organizations need to make architectural choices about data storage and analytics solutions that provide both agility and flexibility.

The first architectural option is a data lake, which allows cost-effective storage of massive amounts of raw data that have an undefined or unclear purpose but are possibly needed for future use.[17] Data lakes offer fully redundant data storage infrastructure for storage and retrieval with accessibility across geographies, web spaces, or time horizons.[18] Data lakes have now become a prominent offering by cloud providers such as Microsoft (Azure Blob Storage), Amazon Web Services (AWS) (S3), Google (Cloud Storage), IBM (Cloud Object Storage), and Oracle (Cloud Infrastructure Object Storage) among others.

An alternative to a data lake is a not Structured Query Language (NoSQL)-based database that can handle nonstructural data with high availability and durability. A key benefit of NoSQL databases is horizontal scalability,[19] which allows seamless scaling of a single table over hundreds of servers and lower administrative overhead for operating and scaling distributed clusters.[18] While data lakes and databases have different advantages and are more suited to disparate business needs, companies such as Netflix use both to serve various requirements.[18] AWS DynamoDB, MongoDB, Google Cloud Bigtable, and Microsoft Azure Cosmos DB are among many NoSQL-based products that are currently available for database needs.

Once a data source is known, a data warehouse can be built using Extract, Transform, Load or Extract, Load, Transform operations.[20] A data warehouse consists of restructured data that are organized, easy to query, integrated from multiple sources, and of higher quality to ensure that robust reporting and data analysis can be performed for business intelligence (BI) or business analysis purposes. Data warehouse products are offered by all major cloud providers (AWS Redshift, Google Cloud's BigQuery, and Microsoft Azure SQL Data Warehouse), stand-alone providers such as Cloudera, and unique providers such as Snowflake that have the capability to integrate data from Amazon S3, Microsoft Azure, and Google Cloud platforms.[21]

WHILE IT IS ARGUABLE THAT SUCH ACTIONS ARE USER ORIENTED WITHIN BOUNDS OF INTENDED USAGE, THERE ARE GRAVE RISKS AROUND SECURITY BREACH AND ETHICAL USAGE OF BIG DATA TO INFLUENCE PEOPLE ON DEEPLY PERSONAL DECISIONS.

The wide-ranging structured and unstructured stored big data is ultimately processed for predictive and prescriptive analysis to gain better understanding and insights in the application area.[22] The analysis spans from SQL-based BI reports, dashboards, and analysis to help business operations, to unstructured data processing using Apache Hadoop to solve data-intensive and computationally intensive problems[23] and AI and ML tools for building smart applications, such as predictive analytics, deep learning, image identification and classification, and natural language processing.

The rapid growth in the big data industry in the last decade has resulted in the availability of a multitude of tools to execute these analyses. While Tableau, AWS QuickSight, MicroStrategy Analytics, Microsoft Power BI, and Google Data Studio are among the most commonly used BI tools, Amazon EMR, Microsoft Azure HDInsight, and Cloudera Manager are some of the common unstructured data processing platforms that support big data frameworks, such as Apache Hadoop and Apache Spark. A wide-ranging suite of tools (Amazon's Lex, Polly, and Rekognition; Google's AutoML, AI Infrastructure, and Healthcare Natural Language,[24] and so on) are further available to conduct predictive analysis using AI and ML.

TRENDS

Big data and data analytics will continue to grow in the coming years and are expected to be valued at US$230 billion by 2025.[25] An ever-growing number of organizations will continue to adopt and operationalize AI, resulting in a fivefold increase in streaming data and analytics infrastructure by 2024. The 2020 trends of growth in AI/ML, augmented analytics, edge-computing growth, in-memory computation, and continuous intelligence as well as growth or Spark and Databricks tools will continue in the next few years as well.

Augmented analytics utilizes AI and ML techniques to automate data preparation, sharing, and analytics, resulting in the transformation of complex, seemingly unusable data into smaller, usable data sets. It is estimated that augmented analytics markets will become a dominant driver for BI in 2021. Cloud-to-edge transition is also picking up to move away from centralized computing systems requiring high bandwidths. Edge computing will result in faster data analysis and reduced costs since it's better to extract and process data at the edge and then distribute them to relevant users/customers as needed. According to estimates,[26] by 2025, 75% of the enterprise data will be processed by edge computing. Due to the increasing need for real-time data analytics and the decreasing cost of memory, in-memory computing is expected to continue to grow in the coming years. This will be particularly helpful for business clients (banks, retailers, and utilities) for rapid identification of patterns.

Continuous intelligence is also expected to support automatic real-time data analysis and decision making via ML and continuous data analysis. It uses optimization, business rule management, event stream processing, and augmented analytics. It is predicted that more than 50% of new business systems will incorporate continuous intelligence by 2022.[27] Lastly, with the migration to cloud for data ingestion, analytics, and storage, traditional tools working on data center infrastructures such as Hadoop may no longer be the best option. Newer tools such as Spark, which can work with both data centers and cloud-based systems, will start to become more mainstream.

CHALLENGES

As big data is on a continuous-growth path, there are a few areas that need focus so that organizations and societies can continue to benefit their businesses, improve user experiences, and at the same time prevent privacy breaches, erroneous analyses, and disadvantages to small organizations trying to integrate big data into their day-to-day operations.

Ethical aspects of data collection, management, and application are continuously evolving and will influence industry practices in the next decade. Data privacy arising from big data is undoubtedly the biggest challenge affecting the current 4.7 billion Internet users. Big data is essentially the "big boss" in the online world, wherein our every action gets logged somewhere, often permanently, is analyzed, and influences our day-to-day decisions, often without any of us realizing it. The data include, but are not limited to, information about our favorite restaurants and cuisines, travel, shopping history, and search history.

Besides the personal information that users can choose to share, an individual's trails of disparate online data can be used not only to extract additional personal information but also draw inferences on how a person thinks by creating psychographic profiles. Using big data, the "big five" personality traits—openness, conscientiousness, extroversion, agreeableness, and neuroticism—can be determined with high accuracy. Researchers Kosinski and Wu created a model that predicts an average person's personality, sometimes even more accurately than their family and friends can. As an example, skin color (with 95% accuracy) and political affiliation (with 85% accuracy) could be predicted based on an average of 68 Facebook "likes" by a user.[28]

The combination of shared and deduced personal information on users is employed for highly personalized marketing to influence decisions on where to shop, eat, travel, and so forth. While it is arguable that such actions are user oriented within bounds of intended usage, there are grave risks around security breach and ethical usage of big data to influence people on deeply personal decisions. The biggest such situation in recent memory was the leak of the Facebook data of 87 million users—the largest known leak in Facebook history—by Cambridge Analytica to determine and sell psychological profiles of American voters for political campaigns.[29] These profiles, together with personal data, such as land registries, shopping data,

and club memberships, purchased from data brokers like Acxiom and Experian,[28] helped to perform "behavioral microtargeting with psychographic messaging"[29] and influence voters. While the true effectiveness of such messaging has been a matter of public debate, the instance highlights how big data usage can significantly influence our lives on both a personal and a societal level.

Although data privacy affects 59% of the global population who are connected to the Internet,[30] only a very few, piecemeal policy responses to big data regulation, most prominently the European Union's General Data Protection Regulation (GDPR) and the California Consumer Privacy Act, are in place to tackle this challenge. Data-collection practices and usage must be transparent, and companies must abide by them to ensure that user privacy and data breaches do not occur. The increased data mining from social platforms poses an increased risk of data misuse and loss or theft of sensitive personal information.

As companies venture into mainstream big data analytics, they will need to appropriately invest in the development of cybersecurity tools as well. GDPR was implemented in 2018 for data protection to regulate big data by empowering users to have the choice to decide with which businesses to share their data. The intent is to drive trustworthy data sharing with businesses, which in turn is expected to generate more reliable data and their associated analytics. At the same time, big data regulation should not come at the cost of efficiency and technological advancement and should be balanced to allow healthy and transparent sharing of data. A more universal policy framework for regulation of big data that balances the needs of organizations and the privacy of their customers should be deliberated and implemented by both national governments and international bodies.

Organizations currently working with big data or intending to delve into big data space encounter a myriad of challenges spanning adoption and operational issues due to the massive scale and analytics requirements, especially for unstructured data. While there have been continuous investments and advancements in utilizing structured and unstructured data, we are really at the tip of the iceberg for unstructured data. More unstructured data types are being added to the scope and are expected to dominate big data in the years to come, as we traverse the digital age dominated by social media and online platforms. There is a need for corporations to invest in the analytics for unstructured data to drive better BI. This requires big data management tools to have the right people, policies, and technologies to ensure the accuracy, security, and quality of data.

Organizations considering adopting big data often suffer from insufficient understanding and acceptance of it due to their legacy practices. It is therefore important to have a clear business use case and an expected value to be derived from big data analytics. This is especially relevant as the cost of big data analysis—whether done in house or outsourced—can be quite high as the data grow and expand. Organizations, especially nontechnological firms, also face considerable challenges making the right technology decisions themselves or even through external counsel from specialized firms.[31]

A MORE UNIVERSAL POLICY FRAMEWORK FOR REGULATION OF BIG DATA THAT BALANCES THE NEEDS OF ORGANIZATIONS AND THE PRIVACY OF THEIR CUSTOMERS SHOULD BE DELIBERATED AND IMPLEMENTED BY BOTH NATIONAL GOVERNMENTS AND INTERNATIONAL BODIES.

The other set of challenges can be attributed to the inherent massive scale of big data and its associated infrastructure needs. As organizations continuously generate astronomical amounts of data, the scalability of storage and analytic processes becomes increasingly difficult. It is anticipated that once data volumes surge to exabytes, data storage technologies and network bandwidths will become increasingly constrained and require continuous upgrades and technology advancements. Furthermore, the multifold aspects of big data management, including data distribution across geographies, access management, compliance, ownership, and governance, are becoming important from security, legal, and operational perspectives. Lastly, as organizations increasingly adopt big data, the heterogeneous expansion

of data in both richness and variety will require the evolution of data architecture and analytics to process new data types. This increase in scale often compounds the noise, opacity, and relational nature of data, which increases the risk of data corruption and makes data validation, movement, and analysis very challenging.[32]

Big data analytics is another major problem often experienced when using unstructured data to arrive at desired results and conclusions. High data heterogeneity presents issues in determining a relevant data set for the analysis, preparing, or cleaning data and making sense of the unanticipated effects of outliers. Moreover, ascertaining the analytic method that is most appropriate for a problem or data set is not always clear. The decision-making process used to select from the wide range of analyses, such as predictive, prescriptive, and decision modeling; the different forms of analysis, such as quantitative, classification, visual; and the types of analysis, such as graph theory, social network analysis, behavioral analytics, econometric modeling, and control theory, among many others, can be both overwhelming and prone to error.[32]

Big data is undoubtedly one of the most defining trends of the 21st century, and it will impact every individual, organization, and government globally. The astronomical growth in the data generated from the Internet and IoT devices has provided huge opportunities to make improvements in decision making and efficiency in business operations and drive innovations in the research industry through analytic tools and AI/ML advancements. Big data is expected to be the next growth driver in all industries, from health and medicine to retail, banking, entertainment and media, and more. Along with the opportunities and growth associated with big data, organizations also face multiple challenges due to increased data size and complexity. This requires a continuous search for improved tools for data gathering, extraction, storage, and analytics and also necessitates organizations to be on the lookout for concerns of data breach and leaking of sensitive user information. Companies and organizations must address these challenges by continuing to invest in the development and adoption of big data tools and security and privacy practices to drive business improvements, protect user privacy, and ensure that they do not fall behind as the world marches onto digitization and datafication.

REFERENCES

1. D. McGilvray, *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information*. San Francisco, CA: Morgan Kaufmann, 2008.
2. "Structured vs. unstructured data—Best thing you need to know." ProWebScraper. https://prowebscraper.com/blog/structured-vs-unstructured-data-best-thing-you-need-to-know/ (accessed Jan. 30, 2021).
3. A. Holst "Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2024." Statista. Dec. 3, 2020. https://www.statista.com/statistics/871513/worldwide-data-created/ (accessed Jan. 30, 2021).
4. "Digital around the world." Datareportal. https://datareportal.com/global-digital-overview (accessed Jan. 30, 2021).
5. Internet live stats. https://www.internetlivestats.com (accessed Jan. 30, 2021).
6. "Top 10 big data trends of 2020." FinTech News. Nov. 2020. https://www.fintechnews.org/top-10-big-data-trends-of-2020/ (accessed Jan. 30, 2021).
7. "Cloud computing industry to grow from $371.4 billion in 2020 to $832.1 billion by 2025, at a CAGR of 17.5%." Globenewswire. Aug. 2020. https://www.globenewswire.com/news-release/2020/08/21/2081841/0/en/Cloud-Computing-Industry-to-Grow-from-371-4-Billion-in-2020-to-832-1-Billion-by-2025-at-a-CAGR-of-17-5.html (accessed Jan. 30, 2021).
8. "HealthKit." Apple. https://developer.apple.com/healthkit/ (accessed Jan. 30, 2021).
9. "AlphaFold: A solution to a 50-year-old grand challenge in biology." Deepmind. Nov. 2020. https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology (accessed Jan. 30, 2021.)
10. "Big data as a double-edged sword in the fight against COVID-19." ReadWrite. Apr. 2020, https://readwrite.com/2020/04/30/big-data-as-a-double-edged-sword-in-the-fight-against-covid-19/ (accessed Jan. 30, 2021).
11. J. Stoll, "Number of Netflix paid subscribers worldwide from 3rd quarter 2011 to 3rd quarter 2020." Statista. Jan. 2021. https://www.statista.com/statistics/250934/quarterly-number-of-netflix-streaming-subscribers

-worldwide/ (accessed Jan. 30, 2021).
12. J. Bursztynsky. "Disney+ emerges as an early winner of streaming wars, expects up to 260 million subscribers by 2024." CNBC. Dec. 2020. https://www.cnbc.com/2020/12/11/after-showing-massive-growth-disney-hikes-5-year-subscriber-goal-.html (accessed Jan. 30, 2021).
13. E. Dans. "Netflix: Big data and playing a long game is proving a winning strategy." Forbes. Jan. 2020. https://www.forbes.com/sites/enriquedans/ (accessed Jan. 30, 2021).
14. D. Karr. "2020 CRM statistics: The uses, benefits & challenges of customer relationship management platforms." Aug. 2020. https://martech.zone/crm-statistics/ (accessed Jan. 30, 2021).
15. "Starbucks ramps up expansion efforts with focus on digital offerings." Business Insider. Dec. 2020. https://www.businessinsider.com/starbucks-drives-expansion-efforts-with-digital-offerings-2020-12 (accessed Jan. 30, 2021).
16. B. Pearson. "12 ways Starbucks' loyalty program has impacted the retail industry." Forbes. Dec. 2020. https://www.forbes.com/sites/bryanpearson/2020/12/16/12-holiday-gifts-from-the-starbucks-card/?sh=16f603df4534 (accessed Jan. 30, 2021).
17. "Data lake versus data w*arehouse*." Talend. https://www.talend.com/resources/data-lake-vs-data-warehouse/#:~:text=Data%20lakes%20and%20data%20warehouses,processed%20for%20a%20specific%20purpose (accessed Jan. 30, 2021).
18. "Amazon DynamoDB vs Amazon S3." Stackshare. https://stackshare.io/stackups/amazon-dynamodb-vs-amazon-s3 (accessed Jan. 30, 2021).
19. "NoSQL vs relational databases." Mongodb. https://www.mongodb.com/scale/nosql-vs-relational-databases (accessed Jan. 30, 2021).
20. "Extract, transform, and load (ETL)." Microsoft. https://docs.microsoft.com/en-us/azure/architecture/data-guide/relational-data/etl (accessed Jan. 30, 2021).
21. "Snowflake: The data cloud." Snowflake. https://www.snowflake.com (accessed Jan. 30, 2020).
22. "Building a data lake on amazon web services (AWS)." AWS Cloud. https://pages.awscloud.com/rs/112-TZM-766/images/Building-a-data-lake-on-Amazon-Web-Services.pdf (accessed Jan. 30, 2021).
23. "What is Hadoop?" Amazon. https://aws.amazon.com/emr/details/hadoop/what-is-hadoop/ (accessed Jan. 30, 2021).
24. "AI and machine learning products." Google. https://cloud.google.com/products/ai#tab1 (accessed Jan. 30, 2021).
25. "Big data market worth $229.4 billion by 2025." Marketsand markets. https://www.marketsandmarkets.com/PressReleases/big-data.asp (accessed Jan. 30, 2021).
26. "Edge computing by the numbers: 9 compelling stats." The enterprisers project. Apr. 2020, https://enterprisersproject.com/article/2020/4/edge-computing-9-compelling-stats (accessed Jan. 30, 2021).
27. "Gartner identifies Top 10 data and analytics technology trends for 2019." Gartner. https://www.gartner.com/en/newsroom/press-releases/2019-02-18-gartner-identifies-top-10-data-and-analytics-technolo (accessed Jan. 30, 2021).
28. H. Grassegger and M. Krogerus. "The data that turned the world upside down." Vice. Jan. 2017, https://www.vice.com/en/article/mg9vvn/how-our-likes-helped-trump-win (accessed Jan. 30, 2021).
29. N. Confessore. "Cambridge analytica and Facebook: The scandal and the fallout so far." NY Times. Apr. 2018. https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html (accessed Jan. 30, 2021).
30. "Global digital population as of October 2020." Statista. https://www.statista.com/statistics/617136/digital-population-worldwide/#:~:text=Almost%204.66%20billion%20people%20were,percent%20of%20the%20global%20population (accessed Jan. 30, 2021).
31. A. Bekker. "The 'scary' seven: Big data challenges and ways to solve them." Mar. 2018, https://www.scnsoft.com/blog/big-data-challenges-and-their-solutions (accessed Jan. 30, 2021).
32. J. Alberto Espinosa, S Kaisler, F Armour, and W H. Money, "Big data redux: New issues and challenges moving forward," in *Proc. 52nd Hawaii Int. Conf. Syst. Sci.*, 2019 doi: 10.24251/HICSS.2019.131.

PREETI CHAUHAN is a Senior Member of IEEE, Santa Clara, California, USA. Contact her at preeti.chauhan@ieee.org.

MOHIT SOOD is at the University of California, Berkeley, Berkeley, California, 94720, USA. Contact him at mohit_sood@berkeley.edu.

EDITORS: Daniel F. Keefe, dfk@umn.edu
Melanie Tory, mtory@tableau.com

This article originally appeared in
IEEE Computer Graphics and Applications
vol. 41, no. 1, 2021

DEPARTMENT: PEOPLE IN PRACTICE

BubbleUp: Supporting DevOps With Data Visualization

Danyel A. Fisher, *Honeycomb, San Francisco, CA USA*

BubbleUp is a tool that lets DevOps teams—data analysts who specialize in building and maintaining online systems—rapidly figure out why anomalous data have gone wrong. We developed BubbleUp with an iterative, human-centered design approach. Through multiple rounds of feedback, we were able to build a tool that presents a paired-histogram view to help make high-dimensional data make sense.

The alarm pierces 3 A.M. sleep like a lightning bolt: somebody somewhere is having trouble using your service as they expect. As the human on-call, you need to evaluate the following.

- Who is affected (and how many)?
- How bad is the failure?
- What is *actually* wrong?

And you'd better do it fast. Do you wake up the rest of the troubleshooting team in the middle of the night, or can you go back to bed?

No pressure or anything.

The need to evaluate well and fast is the core premise of a group of tools generally called application performance management (APM). APM tools try to help DevOps teams—teams of developer-operators—understand the reliability of their online systems. At Honeycomb, our software product (also called Honeycomb) is one such APM tool: it supports DevOps teams in exploring complex instrumentation data from their distributed systems.

From the perspective of data visualization, DevOps work in a fascinating data analytics domain. They have deep domain knowledge of highly complex systems; they are responsible for both creating and analyzing a data stream dedicated to the task of monitoring and debugging distributed systems that are run on remote servers. The analytics challenges that they solve have impacts that can be measured in both dollars and hours of lost sleep. Most interestingly, because DevOps teams tend to repair bugs after finding them, each investigation is likely to be unique.

The data analytics tasks that DevOps carry out are familiar to the visualization research community, and the lessons that we learn from their work generalize well to other applications of data analytics. They are asking loosely structured questions of high-dimensional data and need to pursue analyses to solve complex problems.

This article discusses the design and development of BubbleUp. A core component of Honeycomb, BubbleUp exists to support DevOps. Its design is the result of working closely with our target users to understand their needs, iterating on the design, and then tracking the use of the tool over time. BubbleUp illustrates a way to help analysts navigate highly complex data; the process of working intensively with our target users helped us narrow down on a solution that would directly address their challenges.

EXAMPLE: A SLOW API

Figures 1–3 show a sample usage of BubbleUp. An operations team is responsible for handling an API that is exposed on the web; client applications call into it. This team is responsible for making sure that performance continues to run at satisfactory levels. They have been alerted that their system is handling some requests intolerably slowly. Fortunately, their system is well-instrumented, and so they can try to dig into their data to figure out what is wrong.

As shown in Figure 1, they issue a query in their dataset to get a heatmap of how long it takes to process requests. Each point in the heatmap represents the performance of a single request being processed. They note the unusual spike, where some requests are

Digital Object Identifier 10.1109/MCG.2020.3024039
Date of current version 12 January 2021.

 Published by the IEEE Computer Society

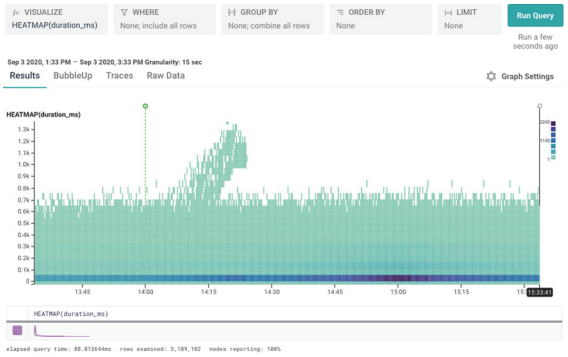


FIGURE 1. Heatmap of the latency for a request in Honeycomb. The darkness of a cell shows the number of requests that were served at that time and latency. The dark line across the bottom of the heatmap shows that most events were served very quickly, but an unusual spike across the top shows some that were much slower.

taking much longer than others, and want to know why they are different.

Using BubbleUp, they select those events (see Figure 2)—the selection is shown as an orange box. BubbleUp responds by showing them a series of histograms comparing the data within the selection to that outside of it (see Figure 3). Each histogram represents one dimension of the data. We compare the events that make up the selection to the events that make up the baseline—all the remaining events. The team can rapidly see that only one `endpoint` and one `app.user_id` were affected: there is only one selection bar on the histogram. In contrast, they can also see that `app.platform` and `app.build_id`, in the second row, do not seem to be important factors: the selection and baseline bars are very similar.

DEBUGGING DISTRIBUTED SYSTEMS

Let's step back to discuss how BubbleUp fits into a broader domain.

Most of the web now runs on distributed systems. An online service might consist of dozens of different

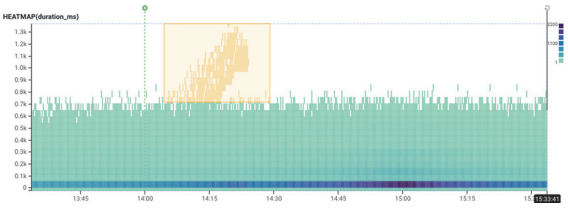


FIGURE 2. User selects a region of the chart.

FIGURE 3. Bubbleup's histograms, one per dimension, comparing the baseline to the selection.

microservices: front-end servers, back-end storage, authentication services, transaction processors, advertising management, and others. This complexity makes it difficult to figure out what has gone wrong when there is a failure. Which service caused a particular slowdown or error? DevOps teams try to instrument their code to describe what their systems are doing, and then try to diagnose and figure out what is going wrong when there is a failure.

The state of the art is to store important metrics—system-level metrics, like memory and CPU usage, and application-level metrics, like the duration of successful API requests—to provide a useful overview of how a service is doing. Each metric can be kept as a single time series. It can be useful to split these metrics out across multiple dimensions: for example, there might be a time series for every distinct API call, split further by whether the requests succeeded or failed.

This makes it extremely fast and effective to offer useful visualizations: a tally of erroneous requests, or the 95th percentile of request duration, for each API endpoint. A talented DevOps team grows experienced with the ways their system can fail and can recognize patterns in the metrics.

Visualization research has looked at this perspective on managing distributed systems. LiveRAC[1] and MeDiCi[2] visualize metrics for many systems simultaneously, for example.

High Cardinality and High Dimensionality

Unfortunately, this still yields a very shallow view of the underlying system and hides a lot of detail about what is actually going on. Accurate diagnosis requires richer information. For example, handling a user request may require a call to authentication, databases, and a webserver to be properly processed. To figure out what's

FIGURE 4. Count of events on a service, broken down by user ID. (Top) all of the customers, and (bottom) one customer, highlighted, showing a diurnal cycle. *Y* axes are obscured to hide actual traffic levels.

wrong with the system, it's helpful to know what user had made the request, which server processed it, what call was made to the database, and how long each of these took and what status code they returned.

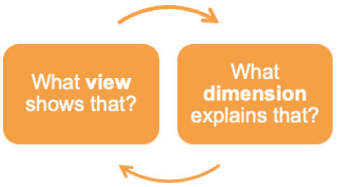
Each of those attributes—the call to the database, the user id—are different dimensions of the dataset; some of those dimensions, like the user id, are extremely high cardinality. Clearly, keeping a combinatoric collection of time series becomes prohibitive. A new generation of systems support those many dimensions, by using column stores. These can provide powerful tools to explore this data. Honeycomb is one of them; the general architecture for such systems follows the example of Facebook's SCUBA.[3] Now it is possible to provide that same count as before, but now split across many different dimensions.

To give an example of how it can look to start using high-cardinality data, Figure 4 shows 50 overlaid time series, representing different users of the system. The top chart shows all the different users overlaid; the bottom chart calls out the one with an unusual diurnal pattern.

High Cardinality and the Core Analysis Loop

This leads to a new analysis dilemma. How do people who develop and operate systems figure out which fields to look at? When there can be hundreds of fields, with millions of values, where do you look to figure out what caused a problem?

I was tasked with helping our users understand the complexity of their data. I went out and interviewed a dozen users, both internal and external to the company. The interviews focused on how they went about going from an alert to a response; in many of them, we chose a recent investigation that they had carried out and reconstructed their process of discovery—including looking at their dead ends. A single strategy recurred in debugging incidents, the *core analysis loop.*



Users would often start with an anomaly in the system that interested them. The core analysis loop started when the user visualized a basic metric that illustrates that anomaly—for example, they might have noticed that some requests were getting slow, so they visualized the median duration of events in the system to see whether it had increased. They would then iteratively try to group that metric by various variables. Their goal was to find a variable that had good explanatory power: that one particular value of one variable could show how the anomaly was different.

For example, let us say that we had encountered the graph at the left side of Figure 5. We wanted to better understand why the 95th percentile of data

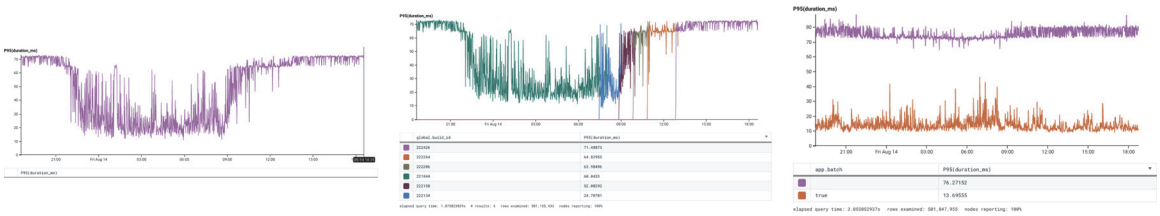


FIGURE 5. Left, the 95th percentile of the duration of requests to the service over time, which dropped from 22:00 to 10:00. Middle and right, grouped by `build_id` and by whether the `app.batch` flag was set. Build does not seem to be a factor, but batch does.

processing time for this query was high at some times and had such a strong dip from 22:00 to 10:00. I suspect that some dimension in my data might explain why the number had dropped. To find out, I might test a series of hypotheses: could this be caused by a specific build of the software? Might it be that a certain set of requests—perhaps those with the `batched` flag—are acting up? For each hypothesis, I would group the data: how does the line look when I aggregate only within builds? What about within the batch flag?

The process of choosing a good variable to check relies on the analyst's experience: if they saw error types that are often associated with database issues, they would turn first to fields that were related to databases. Others would use trial and error, or test hypotheses about their senses of different classes of bugs.

Because Honeycomb is designed as a high-performance query engine, with most queries returning in a few seconds, users could quickly try many different dimensions, looking for an answer. The process could be taxing, as users had to try multiple fields. Honeycomb offers a very loose schema—users can send in whatever sorts of events they want. Many dimensions had no meaningful values, or had too many distinct values, or simply were not relevant to the investigation.

This is a novel problem to the domain we were working in. Many of our competitors simply restricted the number of dimensions that users could send in, often limiting them to under 10. In contrast, it was not unusual for our customers to create datasets with hundreds or thousands of dimensions.

This was a competitive advantage—but also a pain point for our customers. In a low-dimensionality system, it was never hard to find the interesting dimensions. For our customers, there was a risk they would get lost in the noise.

Designing BubbleUp

Honeycomb decided to take on the problem of shortening the core analysis loop. If we could make it simpler to iterate through choices, they would more rapidly converge on their final result. As a secondary advantage, many of our users were unaccustomed to working with high dimensionality data; an experience that helps them understand how powerful their data were would also help them differentiate Honeycomb from our competitors.

We drew inspiration from Scorpion[4] and Macrobase,[5] which highlight the value of explaining anomalies by comparing them to other data.

The heart of the concept is comparing two high-dimensional datasets. Since we knew that our users had already identified anomalous data—such as the dip

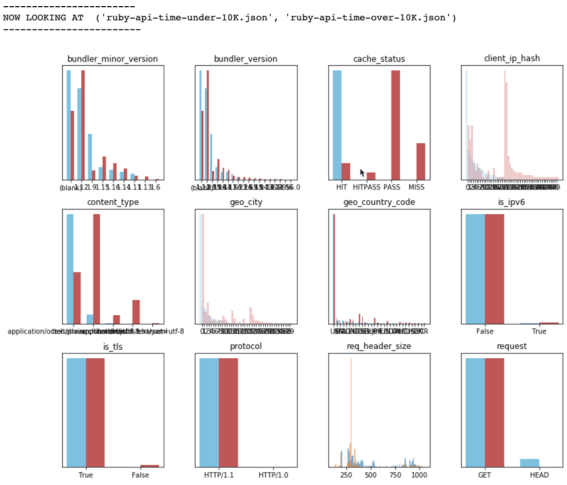


FIGURE 6. First prototype of BubbleUp, built in a Python notebook and visualized using `matplotlib`. Iterating in the notebook allowed us to rapidly explore the design space and validate our decisions.

from 22:00 to 10:00—the question was whether there was a way to separate these groups of points. It would be possible to use high-dimensional analysis techniques that would extract sophisticated, multidimensional explanations. However, we suspected that most explanations we were interested in could actually be much simpler: in many cases, a *single dimension* could distinguish between the anomalous and normal data.

We collected operational data from our servers and started experimenting with prototypes. The first prototypes ran in a Python notebook and generated sheets of side-by-side histograms: each dimension's distribution, shown for the data labeled as anomalous against the baseline.

In these exemplar datasets, it became quickly apparent that there were some dimensions that carried important signals and others that did not. (In Figure 6, which shows our first prototpe, mentioned above, for example, `cache_status` seems relevant, while `bundler_minor_version` probably is not). It also became quickly visible that many dimensions were boring: they had only one value or no values. Perhaps more interestingly, some dimensions had a meaningful value only in the outliers, or in the exceptions.

Admittedly, these were only samples of a single dataset. Every Honeycomb customer has distinctive data, with custom fields that relate to their own business cases. We needed to validate our beliefs about whether this was more broadly true for our customers, too.

One of our users granted us permission to run the code on their data; we sent them a PDF of the python

output. We were delighted at their positive feedback: they instantly understood what was interesting about some of the dimensions, and were able to diagnose a previously unintelligible problem.

This was validation enough to get started on building an implementation. Our design team worked to try to figure out how to incorporate the experience into Honeycomb—a challenge, as the UI did not have a simple way to select a region of the heatmap.

We deployed early (and unstable) betas, first to internal users, then to external users who opted into the experiment. Many of our external users participate in a customer-facing set of Slack channels; we were able to reach out to those users via Slack to build a group of interested users who could exchange feedback.

The feedback we got was a fascinating mix: users would send us long bug-lists and complaints about UI issues and pieces that were difficult to use—and then casually comment that they had used the tool to resolve an incident and that their time to detect issues had dropped from hours to seconds. (One of our insights was that if a user has spent enough time in a tool to complain about small details, then that implies they are finding enough value to dig that deep.)

While beta testing, one user wrote (in a Slack conversation), *[I] was seeing some big latency spikes … look at that it's mostly from one IP address …oh look, it's one IP in Australia.*

Another said, *it "automates" my previous workflow of breaking down and hovering over the table. Some of the use cases we had so far are pinpointing that a specific web process is being slow or seeing that the slowness is being caused by DB queries on a specific endpoint or job.*

The sales team also had a strong reaction to BubbleUp. They had been accustomed to showing how Honeycomb could handle a wide range of data by trying a series of wrong guesses before finding the right answer. BubbleUp allowed them to create a shortened demo (*there's an anomaly, we found it*)—or to walk through the slower process, and then show how BubbleUp short-cut it.

Arguably, the hardest part about releasing BubbleUp was the name: it went from *Smart Drilldown* to *Anomaly Detector* to *Copilot* before we settled on *BubbleUp*; different names were meant to both explain what it did, but also have a personality.

Decisions in Design Iteration

BubbleUp went through numerous rounds of design iteration. It is particularly interesting that we made a number of substantial changes after we released BubbleUp, in response to internal and external feedback on the tool, and our own experience with it.

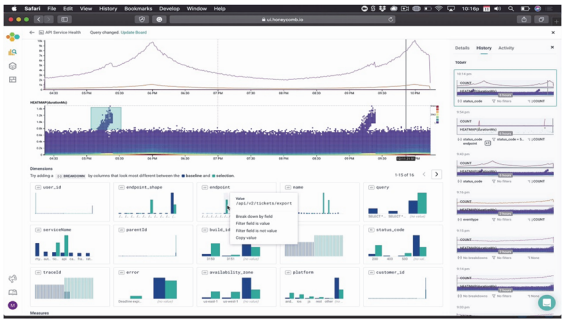


FIGURE 7. BubbleUp in the blue-green color palette. It was difficult to determine which were the selection compared to the baseline, as both colors were used in the heatmap.

Color Coordination

In the first iteration of BubbleUp, we had users compare blue to green histograms (see Figure 7). We got very strong feedback that this was confusing: both colors were well within the color palette of the heatmap, and so users needed to look hard to figure out which was the selection and which was the baseline. By changing to the yellow-and-blue color palette, the questions went away instantly: users understood the yellow mapped to the yellow highlighted area.

Histogram Ranking

We wanted to ensure that the histograms were ordered usefully, to help ensure that users could identify important dimensions. We played with several different metrics and even ran A/B tests comparing ranking algorithms. In the end, we picked a *relative risk* metric (adapted from Macrobase[5]), asymmetrically weighted to highlight fields that had low cardinality values in the selection.

Null Pies

Honeycomb data can be nonrectangular: not all rows of the dataset have all the same fields. For example, a dataset might contain some events that use the Internet—and, therefore, have fields like `http.status` and `http.request`—while other events might use a database, and so have fields like `b.request` and `db.response`. We rapidly found that for many of the most interesting datasets, a single bar for "no meaningful value" turned out to dominate much of the UI. In Figure 8, for example, `status`, `batch`, and `batch_num_datasets` were often empty but dominated the display.

We worked with a designer to create donut charts that could represent *how many nonempty* values there were, instead. In these three images, we see that every event has a defined `trace.trace_id`. Interestingly, most events in the ■ selection have a defined `trace.parent` –

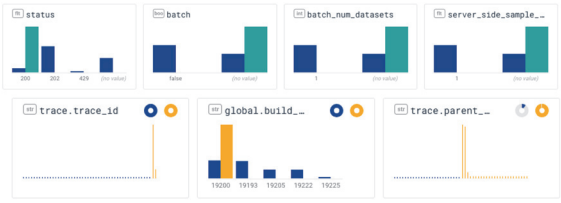


FIGURE 8. Bubbleup with *no value* columns, top, and null pies, bottom. The pie chart shows the percentage of rows that have a valid, nonnull value; the histogram only shows the valid values.

but very few events in the baseline have a defined `trace.parent`. This can help a user rapidly understand how the two groups differ.

Removing Background Events

When we first designed BubbleUp, we contrasted the selection against *everything*. That meant that we counted points inside the rectangle twice: once in the baseline then a second time in the selection. While this had a certain mathematical elegance, it made it actively harder to recognize signals in the data, because data in the selection would also appear in the baseline. Removing points in the selection from the baseline made it far easier to see what was different.

Interacting With BubbleUp

We rapidly realized that one of the most interesting next steps from a BubbleUp was issuing a second query: allowing users to ask *when I eliminate this factor, what's different?*, or *when I focus on this factor, how does it look?*. Following user feedback, we added click through interactions that allow users to create filters and groupings from BubbleUp bars.

BubbleUp for Lines

BubbleUp was efficient to build because it can compare two well-known sets of data points. Still, the most common request to Honeycomb is not comparing heatmap regions—it is understanding why a count, or 95th percentile request, failed. Unfortunately, it is much harder to compute that difference. In those aggregated graphs, we longer know for sure which points sit in the baseline and the selection. Crude techniques—like picking only the slowest points, or all the points in the time region—proved to be insufficiently accurate to provide useful signals. The techniques in Scorpion[4] can help with that computation, and were considering how to incorporate them. Still, BubbleUp-for-Count has been one of the dominant feature requests from our users.

Continuing Life of BubbleUp

BubbleUp continues to be an integral part of the Honeycomb experience. Interestingly, it has had a secondary effect: it is so different from features offered by competitors that it has caused people to see Honeycomb as more substantially differentiated. It emphasizes the value of high dimensional data, and how value can be derived from something that had been seen as out of reach.

The core value of BubbleUp is that it makes it easy to ask novel questions, e.g., *What is special about that particular point?*. This is a key question in the DevOps world—most likely, in many other fields, too—and it drives action. Knowing why a data point is special can help figure out what parameter needs to be tuned, what server needs to be rebooted, or what line of code broke.

We have also begun to build BubbleUp into our product's workflows because it answers the fundamental need to know what happened. For example, Honeycomb recently released features to support service level objectives (SLOs). An SLO computes the ratio of *good* and *bad* events. A key panel of the SLO view shows a BubbleUp of good compared to bad events; we have found that this comparison can rapidly highlight important changes that have caused the SLO to degrade.

CONCLUSION

Comparative Histograms

The heart of BubbleUp is comparing two pools of data to each other as paired histograms. While the applications we have discussed here are for specialized domain, the broader questions about comparing two sets of data should be broadly applicable. I would encourage data analysts in other domains, too, to pick up this sort of a cross-dimensional comparison tool, and to consider paired histograms as a powerful starting step.

Iterative, User-Centered Design

Much of the strength of BubbleUp came, in part, from iterating on feedback from our users throughout the process. Honeycomb users—internal and external—were involved in every development stage, from the first prototypes in Python, through the first release, and then through the incremental improvements. Because of their feedback, we were able to verify that we were going in the right direction, could decide when we were good enough to release, and could prioritize improvements.

This gave us the confidence to massively simplify the fundamental concepts. Comparing sets of histograms on points is computationally simple, rapid to implement, and easy to understand. While techniques like Scorpion[4] and Macrobase[5] are powerful, the cost of building that infrastructure and the complexity of maintaining it was intimidating to a product team in an early stage startup. Building out BubbleUp allowed us to accomplish that the level of value at a fraction of the price.

Fundamentally, BubbleUp has helped our users discover the value of high-dimensional data and to deeply understand their challenges. When that alarms wakes them at 3 A.M., they can find precisely where to look—and faster remediation means better responses to crises.

ACKNOWLEDGMENTS

All of Honeycomb had a hand in BubbleUp, but particular thanks go to E. Freeman, C. Sun, C. Toshok, and I. Wilkes, who implemented the front and back ends for BubbleUp, with visual design by C. H. Chua. BubbleUp was named by D. Mahon. The inspiration of the tool came out of invaluable conversations with E. Wu at Dagstuhl Workshop 17461, *Connecting Visualization and Data Management*. The author is particularly grateful for the feedback from the Honeycomb Pollinators Slack Community.

REFERENCES

1. P. McLachlan, T. Munzner, E. Koutsofios, and S. North, "Liverac: Interactive visual exploration of system management time-series data," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2008, pp. 1483–1492.
2. D. M. Best, S. Bohn, D. Love, A. Wynne, and W. A. Pike, "Real-time visualization of network behaviors for situational awareness," in *Proc. 7th Int. Symp. Vis. Cyber Secur.*, 2010, pp. 79–90.
3. L. Abraham *et al.*, "Scuba: Diving into data at facebook," *Proc. VLDB Endowment*, vol. 6, no. 11, pp. 1057–1067, Aug. 2013.
4. E. Wu and S. Madden, "Scorpion: Explaining away outliers in aggregate queries," in *Proc. VLDB Endowment*, vol. 6, no. 8, pp. 553–564, Jun. 2013.
5. P. Bailis, E. Gan, S. Madden, D. Narayanan, K. Rong, and S. Suri, "Macrobase: Prioritizing attention in fast data," in *Proc. ACM Int. Conf. Manage. Data*, 2017, pp. 541–556.

DANYEL A. FISHER is currently a Principal Design Researcher with Honeycomb.io, San Francisco, CA, USA. He was with Microsoft Research before joining Honeycomb. His work focuses on bringing powerful data visualization tools to end-users. He received the Ph.D. degree from the University of California, Irvine, Irvine, CA, USA. Contact him at danyel@honeycomb.io.

Contact department editor Daniel F. Keefe at dfk@umn.edu or department editor Melanie Tory at mtory@tableau.com.

DEPARTMENT: IT TRENDS

Rigorous Data Validation for Accurate Dashboards: Experience From a Higher Education Institution

This article originally appeared in
IT Professional
vol. 23, no. 3, 2021

Noha Abdou, Afshin Karimi, Rohit Murarka, and Su Swarat, *California State University, Fullerton, CA, 92831, USA*

Data have become an indispensable aspect of our daily lives. The demand for data visualization tools such as dashboards is driven by the desire to make data—and more importantly, the power of using data to inform decision making—accessible to all. During the current pandemic, the availability of real-time data via various dashboards at the global, national, and local levels empowered many to accurately assess the situation and take appropriate actions, a testament to the value of data visualization. For the same reasons, data dashboards are increasingly popular in higher education to promote data consumption and data-driven decision making. California State University, Fullerton (CSUF) is no exception. CSUF has developed a suite of dashboards in the past few years to promote an operational culture that is rooted in evidence.

Discussions about data visualization tools often gravitate towards dashboard design, accessibility, and usability, while neglecting a fundamental (and arguably more critical) issue—the importance of having appropriate and accurate underlying data. The accuracy and adaptability of a dashboard are determined by the accuracy and adaptability of the data behind it, and to ensure such requires a meticulous, streamlined development process. This article is intended to do a "deep dive" into this process.

Institutional Context

In collaboration with the Division of Information Technology (IT), the Office of Assessment & Institutional Effectiveness (OAIE) at CSUF leads systematic and integrated efforts on campus to monitor and demonstrate the impact of university programs, curricula, services, and operations. Being the official data "steward" on campus, OAIE prides itself on providing meaningful data to support strategic planning and decision-making at all levels of the university. To foster a more accessible and easy-to-understand campus culture of data-informed decision-making, OAIE utilizes dashboards to provide important data trends to various campus stakeholders. OAIE's dashboards demonstrate progress across the entire span of students' academic careers from application and admission, major and course enrollment, academic performance, retention, and graduation, to degree completion. Furthermore, OAIE develops dashboards to track key institutional performance indicators including postgraduation alumni outcomes, and faculty and staff diversity.

OAIE regularly maintains two types of dashboards. One type displays static, official, census data, and the other is connected to the Data Warehouse. The connection to the Data Warehouse as opposed to the live operational system is necessary for several reasons.

- Live operational systems [such as CSUF's student information system (SIS) database] are not easily accessible to all users and are not designed for end-user analysis. The live operational databases are typically highly normalized, and querying them involves large numbers of database join operations that could potentially cause system performance issues. Data Warehouses, on the other hand, are designed for quick data retrieval and provide very fast query response times.
- Data Warehouses separate analysis/decision-support from the operational systems. Operational systems are used to manage dynamic data in real-time and support large numbers of transaction processing. They are optimized for "add record"/ "update record" types of operations on well-normalized data. Data Warehouses are designed to handle high-volume analytical processing and are optimized (for the "read" operation) to handle complex queries.

1520-9202 © 2021 IEEE
Digital Object Identifier 10.1109/MITP.2021.3073799
Date of current version 18 June 2021.

- Needed data may reside in different databases on different servers in different formats. Data Warehouses bring consistency in the data definition and format across all collected data, making it easier for the users and the decision-makers to understand, analyze, and share the information. Data Warehouses provide quick access to critical data in one centralized location.
- Data Warehouses support the *ad hoc*, unplanned exploration of data. They also provide an ideal data framework for dashboards and other analytical tools that have drill-down capability.

The Data Warehouse, as well as the extract, transform, and load (ETL) (explained in detail later), process development efforts usually follow a typical software development life cycle model. The stages of such a life cycle model include requirements definition, design, implementation, testing and validation, and release and maintenance.

At CSUF, OAIE serves as the domain data expert who works with a multitude of campus constituents to fulfill their data needs. Furthermore, OAIE is the functional owner of the dashboards, and is therefore responsible for their quality. OAIE's focus, therefore, is on the requirements definition, and testing, and validation aspects of the life cycle model, in addition to the final dashboard development, while IT is mostly charged with the design and implementation aspects of the process. Once the Data Warehouse is developed and validated, OAIE then builds business intelligence (BI) tools, such as dashboards, for campus consumption.

Details of the dashboard warehouse development process that OAIE is responsible for—namely requirements gathering, testing and validation, and dashboard development—will be discussed in the remainder of this article.

REQUIREMENTS DEFINITION

According to the IEEE software engineering glossary, Data Warehouse requirements are defined as follows.

1. A condition or capability needed by a user to solve a problem or achieve an objective.
2. A condition or capability that must be met or possessed by a system or system component to satisfy a contract, standard, specification, or other formally imposed document. The set of all requirements forms the basis for subsequent development of the system or system component.

Abbott defines requirements as: any function, constraint, or other property that must be provided, met, or satisfied to fill the needs of the system's intended user(s). The process of determining requirements for a system is referred to as *requirements definition*, the foundation upon which subsequent stages in system development are built.

Requirements definition is a careful assessment of the needs that a system is to fulfill. It must answer why the system is needed, what features will serve and satisfy the needs, and how the system is to be built. As such, the first crucial step in defining requirements is identifying the subject matter experts (SMEs) who understand the problem definition and can authoritatively address any arising questions.

The requirements analysts work as catalysts with the SMEs in identifying requirements from the various information collected, structure the information by modeling it, and communicate draft requirements to the various audiences. They also draw on their expertise to surface the underlying business needs that the SMEs may not be able to articulate well. This is an exploratory and iterative process, since, oftentimes, SMEs are not able to envision their needs until they see a deliverable, at which point they further refine their requirements. As the requirements definition process typically involves a variety of participants, the requirements must be presented in various forms that are easily understood by the different audiences. Requirements can be divided into functional and nonfunctional requirements.

Functional Requirements

Functional requirements describe what the system should do, e.g., what inputs the system should accept, what outputs the system should produce, and what computations the system should perform.

Nonfunctional Requirements

Nonfunctional requirements describe the general quality characteristics that the system must have to ensure ease of use, optimal system performance, system reliability, and good user experience. Examples include how fast the website loads, or how many concurrent users the system can handle efficiently.

To illustrate the requirements definition process in the higher education setting, let us imagine a university leadership that requests a dashboard to track the daily enrollment registration process for a specific term, in order to gauge the colleges' progress towards meeting their headcount (HC) and full-time equivalent (FTE) enrollment targets. Additionally, they would like

to compare the daily enrollment with the same registration cycle in the prior year to see change year over year. The university's institutional research analysts, who serve as both the requirements analysts and the SMEs, draft the requirements and work with the IT team to develop this dashboard.

Functionally, the leadership wants to view enrollment broken down by the students' major, by the college that offers the course, by undergraduate and graduate/postbaccalaureate student standing, and by new and continuing student status. Due to the different fee structures, they would also like to view enrollment by the students' residence status. Additionally, since the university is encouraging its undergraduates to enroll in 15 units or more per semester in order to graduate in a timely manner, the campus leadership is interested in knowing the students' unit-taking patterns. Finally, they would like drill capability from college-level, department-level, course-level, and class-level to quickly determine enrollment status (including enrollment limit, HC enrolled, available seats, etc.).

In this case, the functional requirements such as computational fields (e.g., FTEs calculations) are clearly defined, and the fields that capture daily enrollment are identified in the campus Data Warehouse. Nonfunctional requirements, on the other hand, include the user experience. Plausible requirements that need to be defined include questions such as when a selection is made from the filter, how quickly the data refreshes on the dashboard? Also, does a selection in one filter concurrently requery the data in another filter? Or, when a user selects a specific college, do the departments change accordingly in the Department filter?

TESTING AND VALIDATION

Once the requirements definition, the design, and implementation stages are completed, it is time to test the quality of the Data Warehouse. To conduct proper testing, the tester needs to have the following.

- Data Warehouse & ETL business rules (mapping documents, transformation rules).
- Environment other than Production (test/development environment).
- Read/Write access to test instances of the source database (data sandboxes).
- Ability to launch the ETL process and have visibility into the Data Warehouse.

ETL processes are the centerpieces in an institution's data management strategy. An ETL process

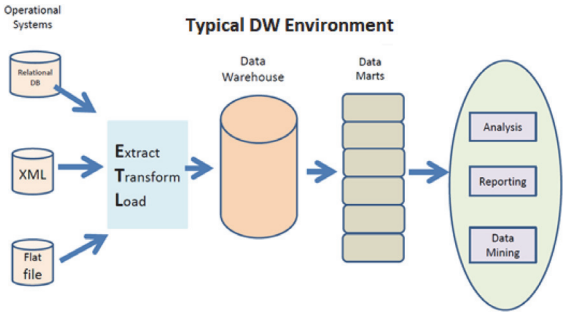


FIGURE 1. Typical Data Warehouse environment.

copies data from one or more source databases, transforms that data to a format that is suitable for reporting and analysis, and finally loads that data to a destination repository (Data Warehouse).

To test the correct execution of this process, a Data Warehouse tester should be knowledgeable about the data source table/field location and information, the destination table/field information and location, as well as the data transformation rules and logic. This information is typically captured in a mapping document. A Data Warehouse tester can use this information to develop and execute test cases.

In addition to having access to an ETL data mapping document, a Data Warehouse tester needs access to a nonproduction Data Warehouse environment, as depicted in Figure 1 below. In a nonproduction test environment, the ETL process loads the data from test instances of different source databases and then loads the transformed data to a test instance of the Data Warehouse.

In such an environment, one can set up test cases by manipulating and setting up data in a source database (e.g., SIS). This implies that the tester has full read/write access to those test database instances. After setting up the test data, the tester launches the ETL process and examines the loaded data in the test instance of the Data Warehouse to see whether the test case passes.

Data Warehouse Testing Example

As a test scenario, let us consider CSUF's Student Success Dashboard. The dashboard, among other indicators, has a student-level flag called "Enrolled" that indicates whether a given student is enrolled at the university.

The business rules indicate that this flag should be set for students enrolled in an undergraduate state-support academic program, who are degree-seekers and are currently enrolled in one or more credit units

TABLE 1. Example test cases to test the data warehouse's "ENROLLED" flag.

Test Setup	Test Execution	Expected Outcome
Locate an undergraduate degree-seeking state-support student enrolled in at last 1 credit unit in the test SIS. . . we call this student A	Launch the ETL	Enrolled flag in test DW should be set to 1
Change the grades for all the courses that student A is enrolled in to 'W' in the SIS	Launch the ETL	Enrolled flag in test DW should be set to 0
Update student A's academic career, program, and plan to a graduate-level program in the SIS	Launch the ETL	Enrolled flag in test DW should be set to 0
Drop all the classes that student A is enrolled in SIS	Launch the ETL	Enrolled flag in test DW should be set to 0
Change student A's academic plan from a degree plan to a credential-only plan in the SIS	Launch the ETL	Enrolled flag in test DW should be set to 0
Add a certificate plan as student A's second academic plan in the SIS	Launch the ETL	Enrolled flag in test DW should be set to 1
Change student A's academic program and plan to reflect a self-support matriculated student in the SIS	Launch the ETL	Enrolled flag in test DW should be set to 0

(excluding course enrollments with "W"grade). Self-support programs, other extended education programs, postbaccalaureate, graduate-level programs, as well as certificate and credential only enrollments, are excluded.

Furthermore, the dashboard displays data stored in the Data Warehouse. To test this flag, a tester should know the names of the tables and fields in the source database that are used to create this flag. The tester should also know the rules that explain the transformation of the source fields into a single binary "Enrolled" flag. The tester can then start writing test cases to test this flag in the Data Warehouse test environment. Table 1 lists some example test cases written to test the "Enrolled" flag.

Different Types of Data Warehouse Software Testing

There are many types of ETL/Data Warehouse testing. Some examples are as follows.

Data transformation testing
Data completeness testing
Data accuracy testing
Metadata testing
Regression testing

- *Data transformation testing:* The source data is transformed ('T' in ETL) before being loaded into the target database. Data can be aggregated, recoded, joined with data from other tables, or transformed based on a certain logic. The transformation rules and logic are in the Data Warehouse mapping document (the tester needs to collect this information if no such document exists). This type of testing may involve writing multiple queries that should run to verify each transformation rule.
- *Data completeness testing:* This testing is to verify that all the data is loaded from the source database into the target database. Simple count comparisons between source data and destination data for variables that are not transformed or are minimally transformed can be performed to satisfy this type of testing. Also, aggregated functions (e.g., sum, max, etc.) can be utilized for this purpose.

One possible reason for having incomplete data in the destination database could be due to an erroneous database join operation. Performing an inner join, instead of an outer join, for example, can cause a certain number of rows to drop from the resulting dataset. The figure below (see Figure 2) shows how the resulting dataset (in red) can vary depending on how the two tables A and B are joined.

1) *Data Accuracy testing:* This test is done to ensure the accuracy of the loaded data. It can include checks for duplicate record loading, as well as tests for effective dates.
2) *Metadata testing:* Metadata refers to data that describe data. In a Data Warehouse, metadata testing includes checking the data types for the target fields and tables, validating the necessary data lengths for different fields, and checking the table indexing.

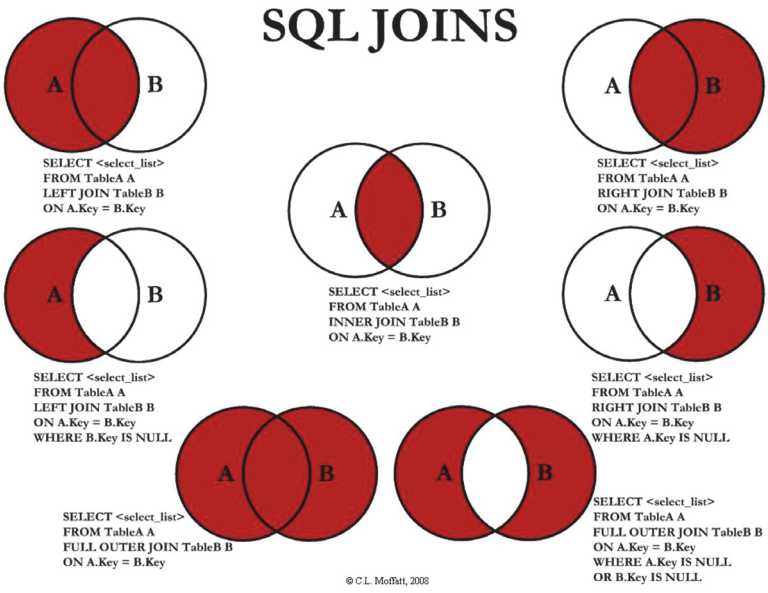


FIGURE 2. Moffatt's visual representation of database join operations 2009.

3) *Regression testing:* Regression testing verifies that any software previously developed still functions correctly after changes are made to the product. The goal is to catch unintended defects introduced when the source code was updated. The Data Warehouse tester should execute regression test cases before every release of the product. Regression test cases typically cover the basic functionality of the product. As defects are fixed or enhancements are made, the corresponding test cases can be added to the regression test suite.

DASHBOARD DEVELOPMENT

After the Data Warehouse is tested and validated, the next step is to develop different BI reports for information dissemination. There are different BI tools that an institution may leverage for the development of such reports. At CSUF, two such tools are utilized: Oracle Business Intelligence Enterprise Edition (OBIEE) suite and Tableau.

Oracle Business Intelligence Edition

CSUF's Data Warehouses are built using Oracle databases. The OBIEE suite provides a seamless way of creating dashboards on top of the Data Warehouses and is the preferred choice when users need to drill down to individual record level data for more information. One such use case is the aforementioned Student Success Dashboard. The dashboard demonstrates the students' academic progression in the university depending on their categorization. This is one of the popular dashboards at CSUF, as it gives users the ability to look at student progression both at the individual and the cohort level, thereby allowing the flexibility of implementing different interventions at different scales.

Registration Snapshot is another dashboard that is built-in OBIEE environment. This dashboard displays the university's enrollment registration numbers and is used to monitor daily registrations when class registration is taking place. This dashboard is instrumental for real-time enrollment monitoring, year-to-year comparison, and class planning (see Figure 3).

Tableau

Tableau is another data visualization tool that is widely used to develop visualizations to disseminate information to the campus community. It can connect to various back-end data sources to extract data. At CSUF, the Tableau dashboards built by OAIE are heavily used to connect to various data sources.

The Tableau connection to the Data Warehouse is done either by using a web connector or connecting directly to the Data Warehouse by signing in to the server from within Tableau. OAIE has used both connection methods. Initially, OAIE used a web connector to connect to the Data Warehouse. Over time, as the size of the Data Warehouse grew, the web connector method could not adequately address the needs as there was a limit to the amount of data that can be accessed by Tableau through this means. Additionally, with new edition releases of Tableau, the web connector also needs to be constantly updated to ensure

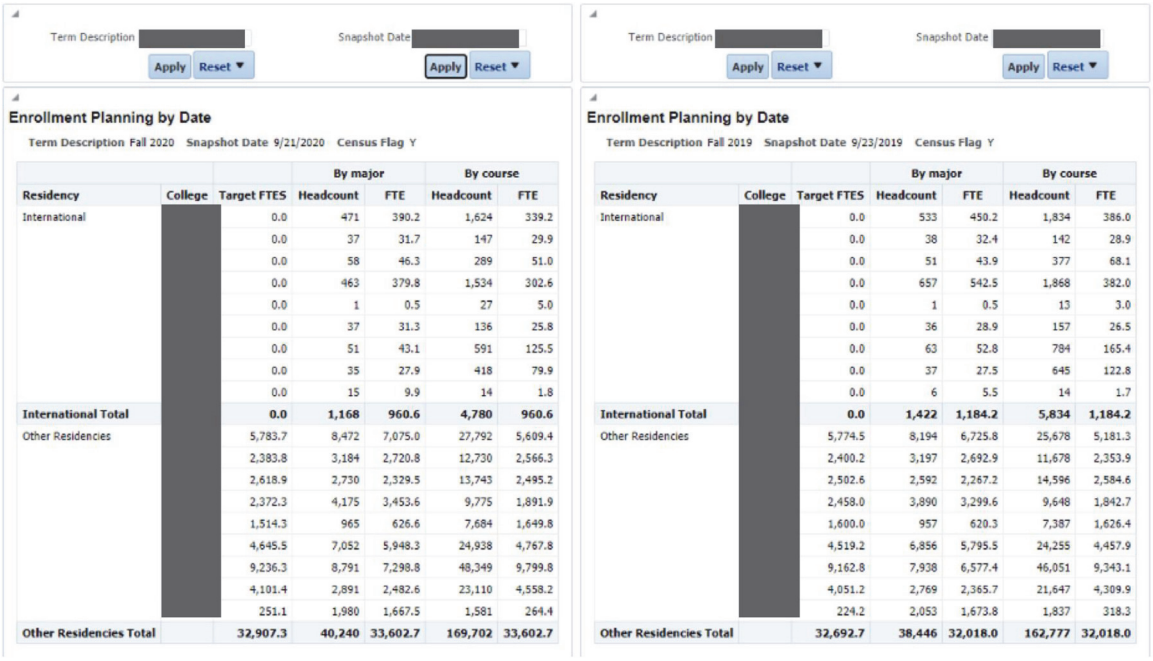


FIGURE 3. Registration snapshot comparison.

smooth communication between Tableau and the Data Warehouse. Upon observing these challenges, OAIE worked with IT to transition to connecting to the Oracle Data Warehouse directly from Tableau. Thus far, this connection method has been able to successfully house dashboards that require daily updates.

CONCLUSION

Dashboards have become a regular part of higher education operations. The ability to visualize data helps empower a broad range of stakeholders with the ability to make data-driven decisions beyond a small group of individuals (e.g., institutional researchers). An important precursor to visualization is having access to accurate data, which are typically housed in Data Warehouses. Data Warehouses are built using an extensive process of requirements gathering, design, implementation, and testing and validation. Only then can dashboard development begin, as described earlier in this article. We have been able to do so at CSUF through collaboration between IT and OAIE in the software development life cycle, a relationship that maximizes the use of each team's expertise. Although we agree with W. Edwards Deming that "In God we trust, all others must bring data," we would add that all must bring *accurate* data through a rigorous data validation process.

REFERENCES

1. "Big data visualization: Turning big data into big insights," Intel IT Center, pp. 1–14, Mar. 2013, doi: 10.1089/big.2013.1507.
2. "Data visualization: Making big data approachable and valuable," SAS, pp. 1–4, Jan.2013.
3. "ScienceMag.org," Accessed 04 2020. [Online]. Available: https://www.sciencemag.org/news/2020/04/every-day-new-surprise-inside-effort-produce-world-s-most-popular-coronavirus-tracker, doi: 10.1126/science.abc1085.
4. "Office of assessment and institutional effectiveness," Accessed 22 February 2021. [Online]. Available: http://www.fullerton.edu/data/
5. *IEEE Standard Glossary of Software Engineering Terminology*, ANSI/IEEE Std., pp. 1–40, 1983, doi: 10.1109/ieeestd.1983.7435207.
6. R. J. Abbott, *An Integrated Approach to Software Development*. New York, NY, USA: Wiley, 1986.
7. A. Karimi, "Methodical data warehouse testing," in *CSU IR Directors Conf.*, 2016.
8. A. Karimi and E. Sullivan, "Student success dashboard at california state university, fullerton," in *Proc. 9th Annu. Symp.*, 2013, pp. 444–454.
9. "Moffatt's visual representation of database join operations," 2009. [Online]. Available: https://www.codeproject.com/Articles/33052/Visual-Representation-of-SQL-Joins

NOHA ABDOU is the Associate Director of Institutional Research, California State University, Fullerton, CA, USA. She is a passionate advocate for data-driven decision making, and for diversity, equity, and inclusion in higher education. Her research interests include all areas of student success, in addition to faculty and staff analytics. She received the Ed.D. degree in educational leadership from the California State University, Fullerton.

AFSHIN KARIMI is a Sr. Research Associate with California State University, Fullerton, CA, USA, specializing in the field of data analytics. Prior to that, he worked for 18 years in the high-tech industry as a Software Developer and Analyst. His expertise is in the areas of software engineering, databases, machine learning and business intelligence. He received the M.S. degree in electrical & computer engineering from The Ohio State University, Columbus, OH, USA.

ROHIT MURARKA is a Senior Research Analyst with California State University, Fullerton, CA, USA. His expertise is in the areas of data mining and statistical analysis as well as data visualization and presentation to support campus wide initiatives. He received the M.S. degree in information systems, concentration in business analytics from California State University, Fullerton, and the B.Tech. degree in information technology from NMIMS University, Mumbai, India.

SU SWARAT is the Associate Vice President for institutional effectiveness with the California State University, Fullerton, CA, USA. Working with campus partners, she and her team significantly transformed the campus culture to embrace data, practice evidence-based decision-making, and foster data literacy. She received the Ph.D. degree in learning sciences from Northwestern University, Evanston, IL, USA, the Master's degree in biology from Purdue University, West Lafayette, IN. USA, and the Bachelor's degree in biology from Peking University, Beijing, China.

EDITORS: Jens Grossklags, jens.grossklags@in.tum.de
Jelena Mirkovic, mirkovic@isi.edu
Bill Newhouse, newhouse@nist.gov

This article originally appeared in
IEEE SECURITY & PRIVACY
vol. 19, no. 2, 2021

DEPARTMENT: EDUCATION

Designing a K–16 Cybersecurity Collaborative: CIPHER

Karen L. Sanzo, Jay Paredes Scribner, and Hongyi Wu, *Old Dominion University*

Cyberattacks have become more common, sophisticated, and harmful, while, at the same time, there is a critical shortage of cybersecurity professionals. For example, from October 2019 to September 2020, there were more than 40,000 unfilled positions for information security analysts.[2] While educational organizations have responded to this burgeoning demand, cybersecurity education and training institutions in the United States have found it difficult to keep pace with the growing call for cybertalent.

Three significant challenges—untapped pools of talent, a lack of diversity in the field of cybersecurity, and inadequate standardization within and across K–16 institutions—impede the identification and cultivation of quality cybersecurity professionals. Although many universities have established cybersecurity degrees, concentrations, and certificate programs, a significant gap exists between K–12 and college education in cybersecurity.[6] It also remains deeply challenging to achieve diversity and inclusion in the cybersecurity field. Additionally, there are no standardized articulations regarding cybersecurity between elementary schools, middle schools, high schools, community colleges, and four-year universities, and there is no guarantee that students at the same grade level are introduced to identical academic content and skills.[3,5]

In this article, we share initial findings from the testing of a proposed framework to address the aforementioned challenges in establishing a K–16 pipeline to prepare cybersecurity professionals. The goal of the initiative is to create a researcher–practitioner partnership (RPP) that paves the way for a national alliance for the development of fundamental, theoretically grounded, and systematic approaches to inclusive K–16 cybersecurity education, especially for students who have a low socioeconomic status (SES). The Cybersecurity Inclusive Pathways Toward Higher Education and Research (CIPHER) model brings together scholars from multiple disciplines and practitioners from various fields to collaborate and fully understand the problems explicated here and to coconstruct a K–16 partnership model to address those challenges.

The planning phase is designed to substantially shape the development of CIPHER. This process, using a design-based implementation research (DBIR) approach, enables us to iteratively test ideas with stakeholders, engage with partners to co-design and test evolutions of CIPHER, and develop a model that can be implemented and scaled up with fidelity.[4] Ultimately, this will lead to a clearly articulated vision and mission for the CIPHER alliance, a well-planned structure and guidelines, and clear road maps for research, education, outreach, and diversity.

DBIR

We use a DBIR approach, which is an extension of design-based research (DBR).[4] Hallmarks of DBR include tenets highlighted by Anderson and Shattuck:[1] being situated in a real educational context, focusing on the design and testing of a significant intervention, using mixed methods, involving

Digital Object Identifier 10.1109/MSEC.2021.3050246
Date of current version: 15 March 2021

 Published by the IEEE Computer Society

multiple iterations, building a collaborative partnership between researchers and practitioners, evolving design principles, and making a practical impact. Further, DBIR extends DBR through a focus "on building organizational or system capacity for implementing, scaling, and sustaining educational innovations."[4] Through this research lens, we are able to present our initial conception of CIPHER to partners and co-design the initiative with educators; make real-time, progressive changes to the model; have tangible and immediate impacts on practice; and study the efficacy of the work. We present the findings from our initial two DBR iterations of CIPHER, including results from our launch meeting and how the model has evolved based on partner collaboration. We conclude with anticipated next steps for the initiative. Readers are encouraged to contact us, in the spirit of DBIR, to provide feedback and if they would like to be part of CIPHER or begin a similar enterprise.

ITERATION 1

The first iteration of CIPHER was developed based on stakeholder survey feedback and a four-hour meeting with more than 35 participants, drawing from K–12 school districts, state-level educational agencies, institutions of higher learning, and industry partners. The survey provided an overview of the purpose of CIPHER and was used to understand if the model resonated with potential partners, ask about who should be included in the planning, and discover initial impressions. These data were used as grounding for the CIPHER launch meeting with stakeholders. The purpose of that meeting was to further explore the model and the emerging vision, present the preliminary concept of having five "task forces" to aid in development, and solicit focus group input through four facilitated breakout groups. The task forces and their purposes are described in the following:

- *Administration and Articulation Task Force*: engage the CIPHER community, identify the leadership team, and develop a plan for coordinating K–16 schools and colleges to establish articulations across the years of schooling to define the pathways for cybersecurity education
- *Diversity Task Force*: develop plans for inspiring and improving the participation of underrepresented groups and low-SES students
- *Human Resource Development Task Force*: identify effective ways to support teachers, counselors, and administrators to integrate cybersecurity into content areas in K–12 curriculum
- *Infrastructure Task Force*: understand the structure support in different schools and classify schools into three tiers: those with high-SES students and substantial computer and Internet infrastructure (tier 1), those with computers for every student but limited Internet access (tier 2), and those with many low-SES students and no computer and Internet infrastructure (tier 3); make recommendations on curricula and infrastructure support to ensure the equity of the proposed cybersecurity education
- *Research and Assessment Task Force*: develop an assessment plan; identify a pilot program to test hypothetical approaches, collect data, and understand what administrators/teachers/counselors/parents need to know and be able to do to support cybersecurity education; create a plan to disseminate research outcomes and solidify CIPHER partnerships.

CYBERATTACKS HAVE BECOME MORE COMMON, SOPHISTICATED, AND HARMFUL, WHILE, AT THE SAME TIME, THERE IS A CRITICAL SHORTAGE OF CYBERSECURITY PROFESSIONALS.

We asked numerous questions: What are your hopes and expectations for a cybersecurity collaborative? What are considerations we should be cognizant of moving forward? Looking ahead three to five years, what will success look like for CIPHER? What should our immediate next steps be? The following focus group themes aided in the development of the second iteration of CIPHER:

- *Creating a central hub*: Attendees expressed a desire to streamline myriad initiatives in

the cybersecurity education space. The collaborative emphasis of CIPHER was seen as beneficial, as there was a lack of collaboration among school districts that had cyberinitiatives (generally due to a lack of capacity and funding resources); a central coalescing mechanism such as CIPHER could address that gap. Essentially, the attendees believed CIPHER was necessary and felt that the initiative's timing was serendipitous.

- *Collaborating, not "bombarding"*: Another interesting but not surprising finding was that K–12 faculty sometimes felt that higher education institutions were "bombarding" them with initiatives, although colleges and universities could also be seen as "partners" and "collaborators." The attendees viewed CIPHER as highly favorable due to the central role K–12 educators would have in the initiative.
- *Building an authentic pipeline*: Much of the conversation around this question revolved around developing a collaborative "pipeline" from K–12 schools to higher education to the workforce. The need to focus on a comprehensive cybersecurity curriculum, access for all students, and training for teachers would be embedded in the concept. While there appeared to be pockets of success in crafting cybersecurity curricula and providing training to teachers, none of the attendees cited a cohesive framework/approach. Some school representatives said there were different departments within the same district that oversaw cybersecurity efforts but rarely interacted. For example, in one district, the responsibility for developing and teaching cybersecurity could be spread across the career and technical education, science, and mathematics departments, with little collaboration.
- *Partnering with employers*: Businesses expressed concern about the challenges to finding cybersecurity talent and the need to establish internships and apprenticeships for high-schoolers and university students. A centralizing hub, such as CIPHER, was seen as a viable mechanism to broker partnerships across educational and industry boundaries while engaging in substantive research to learn iteratively about how to better establish collaborations.
- *Taking action, not just meeting*: The attendees spoke to the need for immediate action and said they did not want to be involved in a years-long "planning-only" initiative. Attendees saw CIPHER's initial focus on planning and development as a drawback and expressed the need to demonstrate immediate, actionable results to prove the concept and build trust with partners. They advocated for small work groups with specific time frames and deliverables. As part of the action orientation, attendees affirmed the work group concept in the original CIPHER design
- *Fostering inclusivity*: Attendees described the need to include representatives from all stakeholder groups, including teachers, administrators, school counselors, business partners, higher-education personnel, and students.

ITERATION 2

Changes were made to the CIPHER planning and to the initiative's mission and vision based on the survey data and focus group findings from iteration 1. We combined the Diversity and Infrastructure task forces into one group to eliminate redundancy. The task forces were populated with volunteers from the initial launch meeting and with additional educators recruited through monthly meetings where the CIPHER concepts continued to be explored and the framework was further developed. Additionally, task force leads were selected from the educator partners to coordinate the teams, including gathering data for future meetings.

Another outcome of iteration 2 was the decision to focus on K–12 partners for five months (September–January) before including additional business and industry collaborators in the design process. The following elements were identified in iteration 2 through focus group (task force) meetings as essential components to the next CIPHER iteration, both in the planning phase and in the initiative's design:

- *Building understanding among internal and external stakeholders*: At times, there was a marked dissonance about the meaning of cyber,

cybersecurity, and computer science. The task force members overwhelmingly expressed the need for K–12 schools to have a common understanding of cybersecurity, career trajectories for students interested in cyberfields, and the preparation teachers need to lead courses in this area. CIPHER was seen as a mechanism to develop this common understanding across districts.

- *Unpacking cybersecurity and computer science*: In relation to developing a common understanding of cybersecurity among stakeholders, task force members described the need to better understand how school districts envision and teach cybersecurity and computer science.
- *Communicating and marketing*: Task force members discussed the necessity for focused, clear, and ongoing communication and marketing campaigns for CIPHER, including during the planning phase, to build interest in and support for the consortium.
- *Fostering hands-on professional development*: As noted, our educator stakeholders emphasize taking action rather than just "discussing" the possibilities for CIPHER. This dovetails with our DBIR approach and, as such, multiple professional development opportunities have been explored and are in the planning phases to be implemented in 2021. They will be used as a vehicle to build interest in CIPHER, provide needed skills to educators, identify areas of additional improvement for teaches and administrators, and gather critical information to develop the K–16 cyberpipeline.
- *Leading cyber in schools*: One area that was underdeveloped in the initial CIPHER design related to exploring the role of K–12 educational leaders. While our original goal was to support school-level leaders in their work leading effective and high-quality cybersecurity training programs, we have also come to better understand how different school districts are developing and implementing cybersecurity curricula in classrooms. As a result, the CIPHER group needs to broaden its understanding of cybersecurity education leaders. In turn, as we work to identify those responsible for developing and implementing cybersecurity programs, we can enhance our support of their work and design development programs tailored to them.
- *Creating an advisory pipeline:* A final area that the task forces have emphasized is to help to create an advisory pipeline that will help students outline their pathways to cybersecurity careers.

Our findings from the first iterations of CIPHER show there is a strong desire among K–12 educators to partner with one another and with institutions of higher education to develop a cohesive approach to preparing students in the area of cybersecurity. In the subsequent months of the CIPHER planning phase, we intend to use the themes to further develop the model, which we anticipate will include incorporating high-school and college students into committees; host professional development sessions focused on cybersecurity (norming around a common understanding, curriculum design and instruction, and the development of academic pathways for students interested in cybersecurity) for teachers, professional school counselors, and administrators; and codesign curricular K–16 cybersecurity pathways for students. Because of our DBIR approach, we are able to design and implement the model while studying the efficacy of the effort and making iterative changes based on ongoing research.

Our hope is that this work will lead to fundamental findings about what administrators, educators, counselors, and parents need to know and be able to do to support cybersecurity education. The anticipated outcomes include the establishment of a CIPHER RPP that engages all stakeholders, a deep understanding of the current knowledge and infrastructure gap, and the creation of an inclusive model for K–16 cybersecurity education, which can be replicated nationwide to bring cybersecurity education to all students.

ACKNOWLEDGMENTS

Correspondence concerning this article should be addressed to Karen L. Sanzo, at ksanzo@odu.edu. This work was supported, in part, by the National Science Foundation, under grant CNS-2012941. The authors

would like to acknowledge the collaboration of Brian K. Payne, Chunsheng Xin, and Danella Zhao in securing the grant.

REFERENCES

1. T. Anderson and J. Shattuck, "Design-based research: A decade of progress in education research?" *Educ. Res.*, vol. 41, no. 1, pp. 16–25, 2012. doi: 10.3102/0013189X11428813.
2. "Hack the gap," CyberSeek, Burning Glass, Boston, MA. Jan. 4, 2021. [Online]. Available: https://www.cyberseek.org/index.html
3. K. Evans and F. Reeder, "A human capital crisis in cybersecurity: A report of the CSIS commission on cybersecurity for the 44th presidency," Center for Strategic & International Studies, Washington, D.C., White Paper, 2010.
4. C. Ford, D. McNally, and K. Ford, "Using design-based research in higher education innovation," *Online Learn.*, vol. 21, no. 3, pp. 50–67, 2017. doi: 10.24059/olj.v21i3.1232.
5. M. Matheny, "CloudPassage study Finds U.S. Universities failing in cybersecurity education," CloudPassage, San Francisco, 2016. [Online] Available: https://www.cloudpassage.com/company/press-releases/cloudpassagestudy-finds-us-universities-failing-cybersecurity-education/
6. W. Zamora. "What K–12 schools need to shore up cybersecurity." Malwarebytes Labs Blog. https://blog.malwarebytes.com/101/2019/02/k-12-schools-need-shore-cybersecurity/ (accessed Dec. 1, 2019).

KAREN L. SANZO is a professor and graduate program director for the PK–12 Educational Leadership Program, Old Dominion University, Norfolk, Virginia, 23529, USA. Her research interests include school leadership development, creating and sustaining partnerships between universities and school districts, and the development and use of formative assessment practices in schools by leaders and teachers. She has served as principal investigator for several national and state-level grants in the areas of school leadership, formative assessment, and science, technology, engineering, and mathematics education. Contact her at ksanzo@odu.edu.

JAY PAREDES SCRIBNER is a professor of educational leadership at Old Dominion University, Norfolk, Virginia, 23529, USA. His research interests include professional and organizational learning and leadership. He has received awards for his research and contributions to the field of educational leadership from the National Staff Development Council, the University Council for Educational Administration, and the University of Wisconsin–Madison School of Education. In 2000, he was awarded a National Academy of Education postdoctoral fellowship. Contact him at jscribne@odu.edu.

HONGYI WU is the Batten Chair of Cybersecurity and the director of the School of Cybersecurity, Old Dominion University, Norfolk, Virginia, 23529, USA, where he is also a professor in the Department of Electrical and Computer Engineering and holds a joint appointment in Department of Computer Science. His research interests include networked and intelligent cyberphysical systems for security, safety, and emergency management applications. He has served on the editorial board of several publications, including *IEEE Transactions on Mobile Computing, IEEE Transactions on Parallel and Distributed Systems*, and *IEEE Internet of Things Journal*. He received the National Science Foundation CAREER Award in 2004, the University of Louisiana at Lafayette Distinguished Professor Award in 2011, and the IEEE Mark Weiser Best Paper Award in 2018. He is a Fellow of IEEE. Contact him at h1wu@odu.edu.

EDITOR: Andrew Kun, andrew.kun@unh.edu

This article originally appeared in IEEE Pervasive Computing vol. 20, no. 4, 2021

DEPARTMENT: EDUCATION AND TRAINING

Beyond Bots and Buttons—New Directions in Information Literacy for Students

Cliff Lampe, *University of Michigan, Ann Arbor, MI, 48109, USA*

The computing profession has recently been deeply involved in combating misinformation. From measuring how misinformation is propagated in computer networks to creating new algorithms and interface options, computer scientists have been at the forefront of fighting misinformation. However, for those of us responsible for training undergraduates, only relying on our core expertise may be a disservice. Unless we expose students to a breadth of perspectives and domains, their ability to identify and combat misinformation will suffer. Information literacy education is necessary, but needs to be treated as a sophisticated and contextual exercise.

In 2020, people in the United States reported wide distrust in traditional media sources, shaped largely by their political inclinations.[1] Critically, science information has also been misused and misreported in various ways at exactly the time that a global pandemic and increasing climate events make trustworthy science important.[2] In complex, diverse societies information quality is of paramount importance. How we deal with shared environmental, medical, and political challenges is based on how well people understand information, and even more fundamentally on a shared understanding about what is true or not. It is challenging enough to agree what to do about something like a global pandemic, but it is even more difficult when people cannot even agree if something is real or not. In this context, how do we train students to be aware of misinformation and to provide them with the tools with which to fight it?

Traditionally, the way we have trained students to combat misinformation is through training them in information literacy frameworks. Some of these frameworks are relatively simple, and some are more complex. For example, the CRAAP test[3] highlights some simple questions around source criticism, asking students to examine currency, relevance, authority, accuracy, and purpose. However, over the past decade there has been criticism of "checklist" versions of information literacy. In an era of such pervasive misinformation, some of which is pushed by malicious actors trying to undermine the concept of objective truth, the checklists can even exacerbate the problem by giving bad actors a roadmap to gaming authenticity. Over the past decade, more complex frameworks have tried to address this problem. For example, the Association of College and Research Libraries (ACRL), which is part of the American Library Association, has been advocating a framework to get people to examine what is true or not.[4] This framework focuses more on the dynamic nature of information, the constructed and contextual nature of authority, and information literacy as an ongoing, strategic process in which people engage. This avoids the problems of less complex methods by teaching people to think more deeply and critically about the information they receive, and by rejecting simple ideas of what is true or false.

In addition to information literacy, there has been a call for different types of media literacy as well.[5] Similar to information literacy principles, media literacy campaigns try to get consumers of information to critically engage with how the channels or media that transmit information shape its quality and character. The main difference between these two types of literacy campaigns—media and information—is a matter more of focus on the information transmission process. In media literacy the design of the channel, the economic models of the industry, and the social processes around the technology become as important as the content of the information itself.

Information literacy campaigns have been around for decades in the United States, yet it seems like misinformation and conspiracy theories are as relevant as ever. What happened? One issue could be that media and information literacy campaigns backfired. Literacy campaigns told people to be arbiters of what was true

1536-1268 © 2021 IEEE
Digital Object Identifier 10.1109/MPRV.2021.3109579
Date of current version 3 December 2021.

or not, downplayed the role of experts, and substituted individual literacy for institutional expertise. This may have worked, if not for the advent of self-publication via the Internet, which now created a context in which people felt enabled to determine their own truth, and then to find plenty of information to be able to back up that viewpoint.[6] It is also possible that human literacy became swamped by the sheer amount of algorithmically curated content that suddenly became available. A goal of misinformation is not necessarily to convince someone that a lie is true, but it is to create so much doubt about the nature of truth that people give up on analysis and rely on heuristics instead. Sometimes this is referred to as "flooding," in which so much contradictory information is made available that people cannot process effectively what is true or not. Depending on heuristics to define information then enables identity-based decision making. For example, in the United States right now there is a deeply partisan divide on some issues on whether something is true or not. Party identity is an effective heuristic when people feel like they cannot trust information anymore.

Discerning Truth—a course to teach information and media literacy to undergraduates

In this context where information literacy seems to be struggling to keep up with the amount of misinformation available, I was charged with the task of developing an undergraduate course to help students determine what was true or not. The course was a survey of students from all over the university and part of a theme semester related to civics on campus. Taking lessons from recent literature on information literacy, the goal was to help students treat misinformation not as a checklist, but as a deeply contextual concern.

Information literacy in contextual domains. To some extent, an entire university is defined by the question "what is true?" In order to cover the wide range of options for epistemology available in higher education, we focused on a variety of speakers and interactions for students. Broadly, we had three categories of speakers, all from information disciplines.

- Computer scientists: these speakers were experts in using computational methods to study online misinformation. From these interactions students learned about the challenges of operationalizing the concept of misinformation, observing information consumption at scale, and converting findings into possible solutions.
- Legal experts: these speakers spoke to the nature of evidence. From a first amendment lawyer, the students learned about "close reading" of texts and how to understand policy. From a federal prosecutor, the students learned about using evidence to build a case and working with multiple other disciplines to come to a joint sense of truth.
- National and local journalists: these speakers taught students about how the financial structure of the new industry shapes information processes, how journalism ethics works in a modern media environment, and the interplay between media production and social aspects like commenting.

Overall, the speakers highlighted the contextual, complex nature of information and misinformation. While some speakers were able to address issues that include more global perspectives as well as those of people who have been traditionally excluded, those issues should be more prevalent in future iterations of this effort. In addition, while these three areas are important aspects of information, one could imagine including more philosophy, public health, public policy, social work, and a range of other disciplines. In modern U.S. universities, colleges and departments are run administratively separately it will take concerted efforts to train students in information literacy in a way that honors all of the different disciplines with something to contribute.

Information and media literacy workshops. An essential aspect of modern information literacy training is learning through practice. To accomplish this goal, we had students engage in multiple exercises that would highlight the complex nature of misinformation. Exercises included:

- A trolling workshop where students were tasked with creating misinformation campaigns in a sandbox social site. Students used memes, videos, and image macros in order to create their campaigns, and determined success through peer evaluation.
- A moderation exercise where students took turns moderating content in an active social media site, collaborating in groups to build cases for what content should be deleted or not.
- Misinformation treasure hunts, where students found pieces of misinformation online and then did fact-checking exercises to refute the information.

These exercises helped drive home many lessons of misinformation that would be hard to absorb from readings and speakers. For example, misinformation is

as much an affective experience as it is cognitive. In the trolling workshop students reported strong emotional reactions, and really connected with affective arguments. For information literacy, it is important that we go beyond readings and reflections so that students can really experience the multiple dimensions of misinformation.

In computing, we tend to focus on algorithmic and interface solutions to misinformation. However, there is still a strong role for information literacy training to bridge the sociotechnical gap that remains where computing solutions fail.

REFERENCES

1. A. Mitchell, M. Jurkowitz, J. B. Oliphant, and E. Shearer, "How Americans navigated the news in 2020: A tumultuous year in review." pewresearch.org. Accessed: Feb. 2021. [Online]. Available: https://www.pewresearch.org/journalism/2021/02/22/how-americans-navigated-the-news-in-2020-a-tumultuous-year-in-review/
2. J. D. West and C. T. Bergstrom, "Misinformation in and about science," in *Proc. Nat. Acad. Sci*, vol. 118, no. 15, Art. no. e1912444117, Apr. 2021, doi: 10.1073/pnas.1912444117.
3. S. Blakeslee, "The CRAAP test," *LOEX Quart.*, vol. 31, no. 3, 2004, Art. no. 4.
4. Accessed: Sep. 1, 2021. [Online]. Available: https://acrl.libguides.com/framework/toolkit
5. D. Gillmor, "Towards a new model for journalism education," *J. Pract.*, vol. 10, no. 7, pp. 815–819, 2016.
6. D. Boyd, "Did media literacy backfire?," *J. Appl. Youth Stud.*, vol. 1, no. 4, pp. 83–89, 2017.

CLIFF LAMPE is a Professor of information with the School of Information, University of Michigan, Ann Arbor, MI, USA. He is the corresponding author of this article. Contact him at cacl@umich.edu.

EDITOR: James Bret Michael, IEEE Senior Member, bmichael@nps.edu

DEPARTMENT: CYBERTRUST

This article originally appeared in
Computer
vol. 54, no. 4, 2021

An Attack Vector Taxonomy for Mobile Telephony Security Vulnerabilities

Matthew Lanoue, Chad A. Bollmann, James Bret Michael, and John Roth, *Naval Postgraduate School*
Duminda Wijesekera , *George Mason University*

A simplified cybersecurity threat matrix may provide a unifying way to define the security risk posed by current and future generations of mobile telephony.

A security framework proposed by Adam Shostack simplifies threat modeling by asking four questions:[1]

1. What are we working on?
2. What can go wrong?
3. What are we going to do about it?
4. Did we do a good job?

While the first and last questions may be easier to answer, the second and third questions require substantial effort to address. To answer the second question, Shostack proposes using the STRIDE method to identify potential threats. This method invites scientists and engineers to imagine how common attack methods such as spoofing, tampering, repudiation, information disclosure, denial of service (DoS), and escalation of privilege may be used to target a system. For simple systems, this may be effective, but it is not an efficient method for brainstorming threats for a complex system such as mobile telephony.

Addressing the security risks for mobile telephony is a multidisciplinary endeavor. The prevailing practice of examining methodologies and how attacks might be applied does not result in an intuitive representation of attack vectors usable by most experts in areas other than security. We believe that an improved approach to identifying potential threats to mobile telephony, consistent with the four-question framework, is to categorize attacks in terms of attack vectors and their relationship to the user plane (UP), control plane (CP), and their interactions throughout the radio front-end and the core network. Our proposed matrix can aggregate more detailed taxonomies in the literature to enable unified views of what can go wrong and facilitate decisions on what to do about it.

5G MOBILE TELEPHONY

The fifth generation of mobile telephony, also known as new radio (5G NR), promises increased bandwidth, reduced latency, customizability, and greater cellular coverage. The apps envisioned for 5G can be broadly categorized into three use cases: enhanced mobile broadband, massive machine-type communication, and ultrareliable low-latency communication. Examples of planned apps include vehicle-to-everything (V2X) communication, smart cities with networked sensors, and mobile streaming of 3D video with ultra-high resolution and ultra-reliable low-latency communication. 5G virtualized many services of the previous generations and created app-specific, virtual network slices that specify performance and cybersecurity requirements—a significant advance. But 5G systems have also become complex in the quest to enable the flexibility to meet anticipated functional and performance requirements over the next 10 years. 6G may introduce additional complexity with its move from centralized service provisioning to a decentralized network and service architecture.

Digital Object Identifier 10.1109/MC.2021.3057059
Date of current version: 9 April 2021

 Published by the IEEE Computer Society

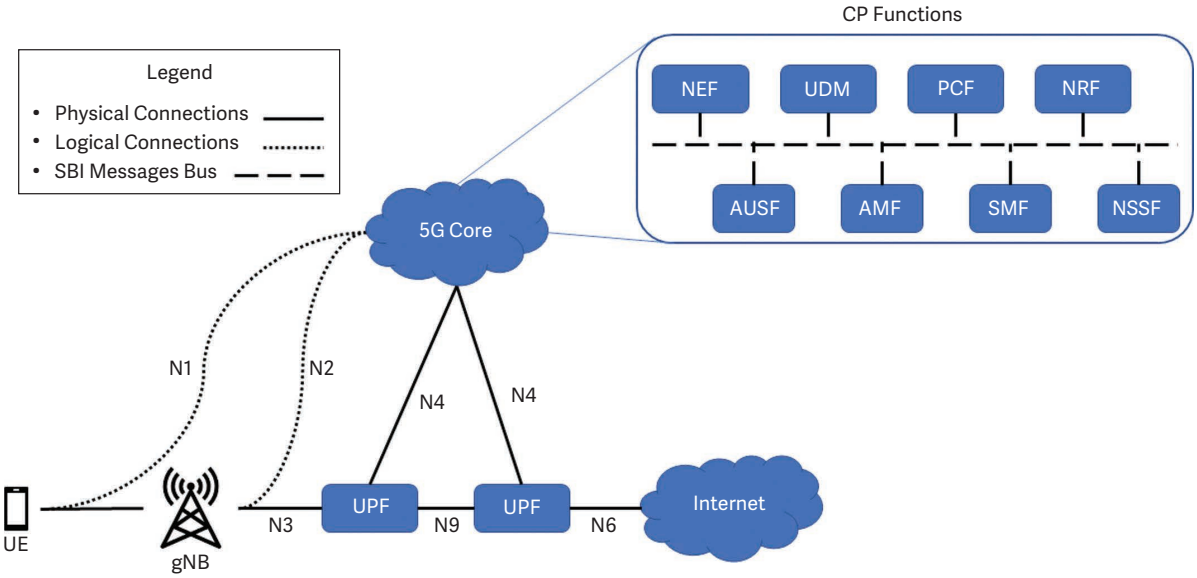


FIGURE 1. An overview of the 5G network architecture. SBI: service-based infrastructure; UE: user equipment; UPF: UP function; UDM: unified data management; AMF: access and mobility management function; SMF: session management function; NEF: network exposure function; AUSF: authentication server function; PCF: policy control function; NRF: network repository function; NSSF: network slice selection function.

With the rapid adoption of 5G NR, one might assume that stakeholders have already thoroughly scrubbed 5G NR artifacts (for example, standards, protocols, architecture, and designs) for security vulnerabilities, although recent studies indicate otherwise.[2,3] Others have enumerated security vulnerabilities in 5G,[4–6] but the current taxonomies and threat models suffer from drawbacks, including complex methodologies that can be difficult for nonsecurity experts to rapidly apply. Some of these other taxonomies also afford limited flexibility to adapt to changes in technology, architecture, design, and usage.

Each engineering and science discipline involved in mobile telephony contributes in some way to understanding and assessing the security risks within the overall engineering tradeoff space. A common, easily graspable taxonomy to categorize the threats, which parts of the network they affect, and where they originate is necessary to facilitate the effective engagement and investment of resources spread across the research, development, sustainment, and defense of the mobile telephony infrastructure.

BUT 5G SYSTEMS HAVE ALSO BECOME COMPLEX IN THE QUEST TO ENABLE THE FLEXIBILITY TO MEET ANTICIPATED FUNCTIONAL AND PERFORMANCE REQUIREMENTS OVER THE NEXT 10 YEARS.

Accordingly, we developed a threat-matrix-based approach to the categorization of attacks based on attack vectors instead of methods. The approach is straightforward to use and adaptable. Let's take a look at how the approach can be applied to 5G NR.

ADDITIONAL DETAILS ON 5G ARCHITECTURE

RADIO ACCESS NETWORK (RAN)

Per our approach, the access plane (consisting primarily of the RAN) is rolled into the UP. The Open RAN Alliance (O-RAN ALLIANCE) splits the RAN into three parts by separating the functional modules that compose a single logical radio node (gNB): the radio unit (RU), distributed unit (DU), and centralized unit (CU). In 5G literature, fronthaul refers to the link between the RU and DU, midhaul to the link between the DU and CU, and backhaul to the link between the CU and the 5G core network. This split, illustrated in Figure 2, supports integrated access and backhaul, promotes interoperability among vendors' implementations of these modules within the RAN, and can (or will) be used to provide complementary services such as optimized traffic flow based on artificial intelligence techniques.[S1] The additional computing power available in the RAN enables the creation of new services further away from the core network—these are called mobile edge services.

UP

The N1 and N2 connections are used to pass information to and from the CPFs by the UE and gNB, respectively. As depicted in Figure 2, the UPF can be further classified by its connection to other UP features. The packet data session anchor UPF connects to a gNB via the N3 link, and the intermediate UPF (IUPF) connects to another UPF via the N9 link. The IUPF connected to an external data network such as the Internet via the N6 link may also support inter public land mobile network UP security to protect the network from incoming malicious traffic.[S2]

CP

Important CPFs include the AMF, SMF, and UDM. The AMF controls the process for new UE and gNBs to connect to the 5G network and UE handoffs between gNBs. Within the CP, the destination for information carried over the N1 and N2 connections is the AMF. Upon the UE's request, the SMF creates, updates, and terminates sessions as permitted by the AMF and manages the session context with the UPF over the N4 connection. The UDM replaces the home subscriber server in the 4G standard; it manages user data and authentication credentials. Further information on CPFs is found in 3GPP Technical Specification 23.501; network functions and entities are listed in clause 4.2.2, and further details about each network function are in clause 6.2.[S2]

NETWORK FUNCTION VIRTUALIZATION AND NETWORK SLICING

Instead of each network function residing on separate machines, the network functions share common hardware and become virtual network functions (VNFs). While this concept has been applied to some 4G networks, it will be fully adopted in 5G, including within the RAN.[4] This is evident in the decoupling of gNB functionality, as previously discussed. A network slice can thus be thought of as all of the related VNFs servicing a certain network app. For instance, a server rack may contain a network slice for a massive machine-type communication network for the factory floor and another slice for an ultrareliable low-latency communication network for connecting supervisors to multiple geographically dispersed factory floors.

REFERENCES

S1. "Open RAN explained." Nokia, Espoo, Finland, Oct. 16, 2020. [Online]. Available: www.nokia.com/about-us/newsroom/articles/open-ran-explained

S2. "System architecture for the 5G system (5GS)." 3GPP, Sophia Antipolis, Tech. Spec. 23.501. V16.6.0," June 24, 2020.

5G NR ARCHITECTURE

Figure 1 illustrates the key components of a simplified 5G architecture. The 5G architecture organizes telephony capabilities as being in the CP or UP in addition to specifying the relationships between the two planes. The specific implementation of any 5G network may differ between service providers and granularity can be added (such as an access plane).

The CP in Figure 1 consists of the 5G core and CP functions (CPFs), the logical connections N1 and N2 to the UE and logical radio nodes known as gNBs, and the N4 connections to the UPF. CPFs in the 5G core communicate with one another via HTTP/2 queries and responses over the SBI bus. For the interested reader, more details are contained in "Additional Details on 5G Architecture."

The UP of Figure 1 consists of the UE, radio access network (RAN), UPF, and the connections N3, N6, and N9. Within the RAN, the actual mechanism by which the base station (gNB) interacts with the core network depends on the mode of operation and equipment involved. It could consist of a simple gNB such as in

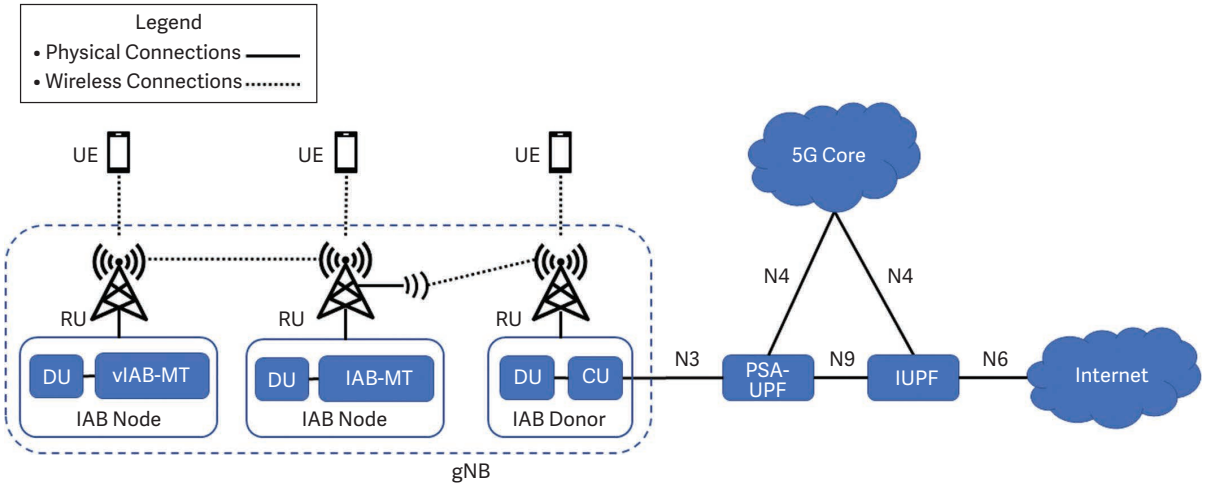


FIGURE 2. The disaggregation of RAN functions for IAB. DU: distributed unit; CU: centralized unit; PSA-UPF: packet data session anchor UPF; IUPF: intermediate UPF; IAB-MT: integrated access and backhaul mobile termination; vIAB-MT: virtual IAB-MT.

Figure 1 or multiple radio units (RUs) working together as a single, virtual gNB to provide integrated access and backhaul (IAB), as depicted in Figure 2.

In a 5G deployment, the UP may consist of a single gNB or multiple gNBs as well as a single UPF or multiple UPFs. IAB provisioning itself is a potential UP attack (UPA) target.

In addition to the changes in network architecture, 5G introduces or improves various security mechanisms. These measures include encrypting the International Mobile Subscriber Identity (IMSI) whenever it is required to be transmitted over the air, limiting which network functions have access to the IMSI (as part of the Subscription Permanent Identifier), and the improved 5G Authentication and Key Agreement protocol.[7,8] A deeper investigation of the security architecture for 5G is beyond the scope of this article. Instead, we focus on how our methodology can be applied to mobile telephony, using a brief overview of the 5G network architecture as a backdrop.

MOBILE TELEPHONY ATTACK VECTOR CATEGORIZATION SCHEME

Much of the literature on the security of mobile telephony centers on attack methods instead of the origins or targets of attacks. While approaches from MITRE and the European Union Agency for Cybersecurity (ENISA) may help to consolidate attack vectors based upon objectives or methods such as outages[5] or credential access,[6] these approaches do not adequately address complex systems such as 5G in which functionality is partitioned among multiple planes. For example, a DoS attack (that is, an outage) could be utilized against a gNB to prevent access within a network cell or against a service provider's UDM to prevent access to an entire network.

TABLE 1. The proposed 5G attack vector matrix.

		Type of Attack (Targeted Function)	
		UPA	CPA
Attack source	UP		
	CP		

Although these attacks have the same objective (outage) and method (target gNB), the attacks differ in their implementation and effect. Similarly, credential access could potentially be accomplished both in the UP as well as the CP. Thus, while binning attack vectors on the method is convenient, it does not provide sufficient clarity to the network engineers, who are constantly refining the standards, and the vendors and service providers, who are constantly refining the UE and infrastructure. Binning on methodology also requires a more complex taxonomy that is difficult to visualize.

Considering that the CP and UP split is fundamental to 5G, we instead propose the approach described in Table 1 that bins attack vectors based on their source and targeted function [that is, a UPA or CP attack

USED IN CONJUNCTION WITH SHOSTACK'S FRAMEWORK, ONE CAN ASSESS THE POLICIES AND MECHANISMS FOR MITIGATING THE SECURITY RISKS OF 5G IN ADDITION TO ASSESSING THE GOODNESS OF THE UNDERLYING TRUST ASSUMPTIONS FOR THE SECURITY ARCHITECTURE.

(CPA)] within these two planes. This simplified top-level model provides for aggregating many of the more detailed taxonomies in the literature, such as ENISA[5] and ATT&CK Mobile,[6] and can assist stakeholders in forming their own mental models of security risks.

To account for the complexity of the multiplane system, the user of the matrix categorizes the attack vectors based on the plane of their origin (UP or CP) and the type of attack (UPA or CPA), as shown in Table 1. In a CPA, the targeted function or feature lies in the CP, whereas in a UPA, the function or feature that is targeted lies in the UP.

The following sections analyze selected attack methods (recently proposed as well as legacy) and examine how they might be applied to mobile telephony, specifically against a 5G network. These attack methods are then fleshed out into more specific attack vectors and, in turn, categorized within our proposed taxonomy. While some of these legacy attack vectors presented may not be viable against a 5G network, we use them to illustrate our taxonomy. Used in conjunction with Shostack's framework, one can assess the policies and mechanisms for mitigating the security risks of 5G in addition to assessing the goodness of the underlying trust assumptions for the security architecture. The power of categorizing attacks based on attack vectors is in presenting a simplified visualization of weaknesses in the mobile telephony architecture to facilitate threat discovery. We believe that some of the first attack vectors employed against 5G networks will be derivatives of legacy attack vectors.

CPAs

The proprietary nature by which individual organizations implement CP features may deter unsophisticated or poorly funded would-be attackers. History, however, teaches us that security through obscurity is inadequate. In addition, the network functions and large repositories of information are likely to be tempting targets for attackers. The 5G CP will also be the most complex, and history also shows that each added feature typically introduces bugs (that is, vulnerabilities or weaknesses) and dependencies; some of the weaknesses may be exploitable for cyberattacks. Finally, the diversity of services and dynamic reconfigurability envisioned in 5G will require extensive virtualization in control nodes and many more edge control nodes. The net result of these changes significantly increases both the attack surface and access vectors to the 5G CP.

Network function spoofing

The use of software-defined networking and network function virtualization to create multiple app-specific network slices will rely on configurations and updates that traverse the Internet using well-documented and recognized network protocols (for example, HTTP/2, TCP, and IP). A capable adversary could insert itself into the routing chain and determine the location and functions contained in a particular network slice. Instead of probing and exploiting vulnerabilities within the network slice application programming interface and functions, they could redirect 5G traffic from a particular RAN to a spoofed network core, thus enabling the collection of information for all connected UE and the manipulation of any network function within the slice.

Database record access and manipulation

Using an initial methodology similar to network function spoofing, an adversary could exploit a vulnerability within the network slice instead of simulating the entire slice. Such attacks are potentially more damaging because the effects could be more difficult to detect and longer lasting. A spoofed slice or database is fixed by reestablishing the proper connections, but one that is manipulated requires additional effort to determine the scope of corruption of the database and restore its integrity.

Signaling interception, manipulation, and jamming

The radio frequencies associated with 5G are

tremendously diverse. Frequency bands are deployed differently by country, but low- and midbands with longer propagation paths start at about 600 MHz (for low bands) and 2.5 GHz for the midrange. The highest bands, millimeter-wave (mmWave), span the 20–60 GHz range. The highest frequencies typically have much shorter and less robust propagation paths due to higher attenuation. However, the increased directionality can support ultradense deployments in urban areas and permit high-data-rate and low-latency apps that were simply not feasible with the available 4G spectrum. 5G design calls for significantly greater base station density and pushing many core services of the network to the edge to handle short propagation paths and novel apps.

The CP is subject to jamming attacks from the middle and both ends. End devices, legitimate or rogue, can target the CP for DoS and certain other types of cyberattacks. At the end of the day, the edge node is just another piece of commodity silicon running an operating system hosting many virtualized apps.[9] Similar attacks, targeting data in transit of user information such as location and other privacy data, can originate from the Internet (used for global backhaul). Regarding "the middle," dense deployments in urban areas require radio frequency (RF) backhaul that will enable classes of attacks not available in 4G.

While the encryption and directionality of backhaul links will mitigate many types of interception and manipulation attacks, jamming is possible through both RF and physical means. The high attenuation and directionality of mmWave signals permit low-tech jamming solutions such as antenna or path blockage. In safety- or time-critical apps, such as the Industrial Internet of Things (IIoT) or connected vehicle communications, DoS jamming attacks on the CP present a real threat to public safety and have been proven feasible in the 4G and, by extension, the 5G CPs.[9]

In the context of applying the proposed classification taxonomy, jamming a link in the chain of nodes used to facilitate IAB (see Figure 2) could be used to manipulate artificial intelligence (AI) logic that routes user data to the CU connected via N3 to the PSA-UPF. If a compromised gNB was inserted in the new path determined by AI and used to harvest user information, then the targeted jamming would be considered a CPA launched from the UP.

UPAs

While the transition from 4G to 5G may alter the technical details of UPAs, UPAs tend to be the most heavily researched and publicized subset of 5G attack vectors. Examples of UPAs that originate in the UP include DoS via cell jamming, key stealing, and hardware vulnerability exploitation. Just as a DoS attack can be accomplished via multiple methods, a spoofed base station (gNB) can be used to accomplish a variety of attacks. Using a compromised gNB to deny service to a group of UE within the cell would represent a UPA launched from the UP. However, using the compromised gNB to create a database of user credentials and security keys by simulating access requests to the core network represents a CPA launched from the UP.

JUST AS A DOS ATTACK CAN BE ACCOMPLISHED VIA MULTIPLE METHODS, A SPOOFED BASE STATION CAN BE USED TO ACCOMPLISH A VARIETY OF ATTACKS.

UE spoofing

Part of the 5G NR appeal lies in its billing as the "one standard to rule them all" by enabling the convergence of the previously isolated silos composed of V2X, industrial control systems (ICSs), the IoT/IIoT, and mobile broadband (MBB) communications (among others). Although service providers try to offer robust authentication methods to thwart the use of spoofed (or rogue) UE, the added complexity of 5G NR creates opportunities for developing new classes of UE spoofing attacks. For instance, one might leverage the integration of ultralow-latency (ULL) communications provisioned for safety and industrial apps by exploiting known vulnerabilities in legacy manufacturing stations. As noted previously, industrial or safety use cases present new consequences for DoS-type attacks against devices in the UP.

ICSs typically have stringent timing, availability, and reliability constraints that require the ULL promised by 5G. These use cases require the addition of quality-of-service (QoS) provisions to the 5G NR standards or slices thereof. These new specifications can, in turn, be leveraged to induce new classes of DoS

attacks.[10] For instance, the 5G NR standards permit ULL UE to preempt transmissions from MBB UE. In one recent study, authors simulated both throughput degradation via MBB preemption and breaking typical ULL QoS guarantees using as few as five pieces of rogue UE.[10] Ultimately, these attacks warp 5G's flexibility to preempt normal device-to-base-station communications with spoofed higher priority requests; the resulting DoS does not have to be highly effective to break QoS guarantees for safety applications.

Cell jamming

The novelty of 5G ensures that many of the proposed 5G jamming attacks are still theoretical. But, in principle at least, individual frequencies or mobile devices (UE) are subject to many of the same attacks that have already been proven against 4G and are well described by Lichtman et al.[9]

The frequency diversity of 5G NR makes it much more difficult for an adversary to jam all possible downlinks from an entire cell. In turn, this means broad-based noise jamming will remain effective but require high transmission powers and be readily observable to counterjamming sensors. Selective jamming strategies will be stealthier and likely more effective at targeting smaller sets of UE. Selective jamming strategies target specific cell-to-UE connection information carried by cell downlinks such as the primary synchronization signal (PSS) and physical broadcast channel (PBCH). These attacks would typically leverage the cell's own broadcast information to identify and select the specific time and frequency blocks to target. Note that there are many more potential targets than the PSS and PBCH alone; Lichtman et al. provide a thorough discussion of these attacks.[9]

Base station spoofing

Base station spoofing (both Wi-Fi and 3G/4G) is a proven attack used by both law enforcement and criminals to target individuals and devices via the UP. Rogue base stations can be configured to enable different types of attacks. If configured to forward connection information from UE to the CP and connect to a legitimate cellular provider, the base station acts as if it was the target's device. By reading or recording the intercepted traffic, the base station can impose nearly any attack on the connected UE, including confidentiality and integrity compromise, geolocation, and DoS.[11,12] Less-complex attacks have also been demonstrated that do not require the breaking of encryption, such as collecting UE identifying data, service downgrade, and battery-draining attacks.[5]

Hardware vulnerability exploitation

The popularity of location-based apps has resulted in many providers of telecommunications chips bundling global navigation satellite system (GNSS) functionality into their mobile telephone system as a system-on-a-chip solution. GNSS, being a set of well-documented standards designed to work with a low signal-to-noise ratio, enables a myriad of terrestrial spoofing attacks. For instance, CVE-2019–2254 represents a vulnerability applicable to numerous Qualcomm chip sets that uses the spoofed commands sent to the GNSS chip to conduct a buffer overflow and execute arbitrary code on a system.[13,14] The prevalence of nonmobile telephone features on modern smartphones represents a significant attack vector for hardware vulnerabilities by increasing the device's attack surface, as discussed previously.

Key catching and stealing

Encryption between UE and RAN utilizes a permanently stored key in the UE's subscriber identification module (SIM) card. The IMSI and International Mobile Equipment Identity of the UE are protected by temporary IDs once the device is connected to the RAN. However, an attacker can force a situation where the temporary identifiers (IDs) stored in the UE and network fall out of synchronization. In this case, the network may request the UE to use its permanent ID to reestablish communications. In 4G networks, this mechanism could be exploited by the attacker to force all UEs served by a particular gNB to broadcast their IMSI, which was done in the clear, referred to as IMSI catching. Since the attack uses a CP authorization mechanism to steal user credentials, it would be a UPA launched from the CP.

Additionally, the use of permanent symmetric keys has fallen under scrutiny after The Great SIM Heist revealed a large-scale attack against the SIM provisioning process.[4,15,16] Other methods for storing permanent authorization credentials include embedded universal integrated circuit cards or some other form of "soft

SIM" that can be remotely provisioned. Adversarial manipulation of any remote provisioning process could be used in a DoS attack or to enable the connection of unauthorized devices to a network. Such key stealing operations are UPA launched from the UP because the credentials stored in a SIM card are part of the UE.

Timing attacks

Precise time synchronization is essential for mobile apps, such as the synchronization of base stations to enable call handoff.[17] While clock synchronization can be accomplished via cellular signals, it can also be accomplished via network-based protocols such as the Network Time Protocol (NTP) and Precision Time Protocol (PTP) or by wireless signals, such as those from GNSS constellations supporting real-time kinematic positioning. Each method has an associated accuracy and interface—which introduces its own associated set of vulnerabilities. Clock synchronization protocols such as GNSS are one way and are thus susceptible to man-in-the-middle (MITM) attacks. PTP is a two-way clock synchronization protocol, where the round-trip time delay can be measured to detect MITM attacks. However, this alone is insufficient to guarantee protocol security.[17]

Timing attacks could be utilized to increase network delays, degrade communication links, or create opportunities for UE to connect to a spoofed network infrastructure. Some levels of security against this type of attack are offered by the variety of timing synchronization options available to the gNB. An attack that manipulates the timing inputs to the gNB to increase delays or degrade service is a UPA launched from the CP since timing is a service provided through the CP.

MITIGATIONS AND THE WAY AHEAD FOR IMPROVED SECURITY IN 5G AND BEYOND

Agreement on a common scheme for categorizing security vulnerabilities is a required first step to enabling us to be efficient and effective at improving the security of mobile telephony. For example, the scheme can help us identify areas of overlap where investing our scarce time and people resources will maximize our return on investment in addressing potential threats. Vulnerabilities can then be identified and investigated by using industrial-strength formal methods, simulations, and experimentation, which, in turn, feed into threat modeling and security risk reduction. The goal here is to systematically improve 5G security and inform the research and development of follow-on work regarding architectures, standards, protocols, and implementations. Let's not repeat security missteps that occurred in prior generations of mobile telephony.

Table 2 illustrates our vector matrix, which is partially populated using the attack examples described previously. This scheme is invariant to the evolution of equipment, standards, and implementations, making it possible to apply it to 6G and beyond—which may introduce new functional planes. It also helps consolidate multiple attack methods, making common threat vectors within the 5G architecture easier to discern and assess.

Each threat warrants a discussion of mitigations that are beyond the scope of this article. But we can identify a couple of overarching mitigation considerations by considering two specific attacks discussed previously: cell jamming and base station spoofing.

Regarding selective jamming attacks, because most of this "free" targeting information is codified in the 5G NR standards, mitigating potential attacks in scalable, compatible ways is a challenging and open

TIMING ATTACKS COULD BE UTILIZED TO INCREASE NETWORK DELAYS, DEGRADE COMMUNICATION LINKS, OR CREATE OPPORTUNITIES FOR UE TO CONNECT TO A SPOOFED NETWORK INFRASTRUCTURE.

TABLE 2. A partially populated 5G attack vector matrix.

		Type of Attack	
		UPA	CPA
Attack source	UP	• Cell jamming used to conduct DoS attack • Spoofed base station conducts DoS to a subset of attached UE • Spoofed UE used to attach an unauthorized device to the network • Hardware vulnerability exploitation • Key stealing	• Spoofed gNB obtains user credentials by simulating multiple access requests • Jam select links in IAB path to reroute user data through compromised gNB • Modified gNB conducts DoS against multiple gNBs attached to 5G channel by falsely advertising favorable channel characteristics
	CP	• IMSI catching • Timing attacks • Use privileged access to core network to authorize attachment of modified gNB that captures user credentials	• Use privileged access to core network to manipulate billing • Use privileged access to core network to facilitate DoS attack by modifying routing policies, QoS, and so on

area of research. Hardware changes will have to be avoided to ensure backward compatibility. Software changes for security may unacceptably reduce the data rate or battery life. Because of the difficulty of changing a mature technology, the most effective mitigation—restricting the amount of timing and frequency detail that is provided to UE prior to authentication with a base station and encryption of the link—will have to be reserved for 6G and beyond.

Some of these same mitigations would help to address the security risks related to base station spoofing. Additional service provider mitigations could include automated cellular network sensing and anomaly detection based on unusual gNB control messages (for example, frequent RRC_REJECT and RRC_IDLE), unusually high signal strength reports, or inconsistent blacklisted cell data.[11] UE solutions could involve additional checks on certain control messages from the base station as well as reports to the CP.[11] Both UE and edge reports would effectively increase the number of sensors available to a provider to detect malicious and accidental network issues.

Many of the rogue base station mitigations could also be implemented in software in the CP. While imposing computational and data overhead, providers can implement these fixes relatively quickly. But to ensure smooth roaming and protection across cellular network operators, some fixes would require changes to the standards. Standards changes are the source of urgency for fully identifying 5G threats and mitigations; 6G standards are currently being refined, and the window for accepting changes may close before all of the useful changes are identified.[18]

We've presented a novel scheme for categorizing attack vectors based on a dual-plane architecture, such as that specified for 5G. Given that mobile telephony is part of the Internet, applying our scheme requires thinking about how other parts of the Internet interact with mobile telephony. The Internet can be thought of as an extension of the RAN, carrying CP and UE messages. With this perspective, how do we classify the cyberattacks emanating from outside of the mobile telephony environment in the proposed scheme? For example, in Table 2 we listed timing attacks as UPA from the CP. To us, this makes sense if NTP or PTP is the feature we are looking at. If, instead, we looked at a border router peered with a 5G network, one could argue that the router is part of the UP since it is part of the data-routing path.

Regardless of whether an attacker attempts to leverage weaknesses of the UP, CP, or both, improving society's trust in the dependability of deployed 5G NR networks and the apps they enable will be challenging. By embracing the concept of the functional-plane split, our taxonomy is usable for 6G and beyond while remaining simple enough to facilitate engagement and understanding by experts across technical domains. By focusing on vectors rather than specific methods, our scheme can aggregate more detailed methodologies and accommodate the evolution of mobile telephony technology and cyberattacks. The matrix can also be scaled as desired, while improved threat visualization enables allocating limited resources to obtaining the most fruitful security enhancements.

REFERENCES

1. J. Beyer, "Adam Shostack on threat modeling," *IEEE Softw.*, vol. 37, no. 6, pp. 110–112, 2020. doi: 10.1109/MS.2020.3017406.
2. F. Rashid. "5G networks will inherit their predecessors' security issues," IEEE Spectrum. June 23, 2020. https://spectrum.ieee.org/tech-talk/telecom/security/5g-networks-will-juggle-legacy-security-issues-for-years
3. A. Shaik, R. Borgaonkar, S. Park, and J.-P. Seifert, "New vulnerabilities in 4G and 5G cellular access network protocols: Exposing device capabilities," in *Proc. 12th Conf. Security Privacy Wireless Mobile Netw.*, 2019, pp. 221–231. doi: 10.1145/3317549.3319728.
4. P. Schneider and G. Horn, "Towards 5G security," in *Proc. IEEE Trustcom/BigDataSE/ISPA*, 2015, pp. 1165–1170. doi: 10.1109/Trustcom.2015.499.
5. "ENISA threat landscape for 5G networks," European Union Agency for Cybersecurity, Attiki, Greece, Dec. 2020. doi: 10.2824/802229.
6. "Mobile matrices," MITRE Corporation. https://attack.mitre.org/versions/v8/matrices/mobile/ (accessed Jan. 4, 2021).
7. "Security architecture and procedures for 5G system," 3GPP, Sophia Antipolis, Tech. Spec. 35.501. V16.5.0, Dec. 16, 2020.
8. A. Koutsos, "The 5G-AKA authentication protocol privacy," in *Proc. 2019 IEEE Euro. Symp. Security Privacy*, pp. 464–479. doi: 10.1109/EuroSP.2019.00041.
9. M. Lichtman, R. Rao, V. Marojevic, J. Reed, and R. Jover, "5G NR jamming, spoofing, and sniffing: Threat assessment and mitigation," in *Proc. IEEE Int. Conf. Commun. Workshops*, 2018, pp. 1–6. doi: 10.1109/ICCW.2018.8403769.
10. C. Chen, G. Hung, and H. Hsieh, "A study on a new type of DDoS attack against 5G ultra-reliable and low-latency communications," in *Proc. Euro. Conf. Netw. Commun.*, 2020, pp. 188–193. doi: 10.1109/EuCNC48522.2020.9200956.
11. S. Jha et al., *Private communication*, Oct. 2020.
12. S. Hussain, M. Echeverria, O. Chowdhury, N. Li, and E. Bertino, "Privacy attacks to the 4G and 5G cellular paging protocols using side channel information," in *Proc. Netw. Distrib. Syst. Security Symp.*, 2019, pp. 1–15. doi: 10.14722/ndss.2019.23442.
13. "CVE-2019-2254 detail," NIST, Gaithersburg, MD, NVD-CVE-2019-2254, 2019. [Online]. Available: https://nvd.nist.gov/vuln/detail/CVE-2019-2254
14. "Multiple vulnerabilities in Google Android OS could allow for arbitrary code execution," Center for Internet Security, East Greenbush, NY, 2019. [Online]. Available: https://www.cisecurity.org/advisory/multiple-vulnerabilities-in-google-android-os-could-allow-for-arbitrary-code-execution_2019-069/
15. J. Scahill and J. Begley. "How spies stole the keys to the encryption castle," The Intercept, Feb. 19, 2015. https://theintercept.com/2015/02/19/great-sim-heist/
16. "Rationale and track of security decisions in Long Term Evolved (LTE) RAN/3GPP System Architecture Evolution (SAE)," 3GPP, Sophia Antipolis, Tech. Spec. 33.821. V9.0.0," June 12, 2009.
17. L. Narula and T. E. Humphreys, "Requirements for secure clock synchronization," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 4, pp. 749–762, 2018. doi: 10.1109/JSTSP.2018.2835772.
18. X. Giao, Y. Huang, S. Dustdar, and J. Chen, "6G vision: An AI-driven decentralized network and service architecture," *IEEE Internet Comput.*, vol. 24, no. 4, pp. 34–40, 2020. doi: 10.1109/MIC.2020.2987738.

MATTHEW LANOUE is a lieutenant in the U.S. Navy and a graduate student in the Naval Postgraduate School's Department of Electrical & Computer Engineering, Monterey, California, 93943, USA. Contact him at matthew.lanoue@nps.edu.

CHAD A. BOLLMANN is a permanent military professor in the Naval Postgraduate School's Department of Electrical & Computer Engineering, Monterey, California, 93943, USA. Contact him at cabollma@nps.edu.

JAMES BRET MICHAEL is a professor in the Naval Postgraduate School's Department of Computer Science and Department of Electrical & Computer Engineering, Monterey, California, 93943, USA. Contact him at bmichael@nps.edu.

JOHN ROTH is an assistant professor in the Naval Postgraduate School's Department of Electrical & Computer Engineering, Monterey, California, 93943, USA. Contact him at jdroth@nps.edu.

DUMINDA WIJESEKERA is a professor in the Department of Computer Science at George Mason University, Fairfax, Virginia, 22030, USA. Contact him at dwijesek@gmu.edu.

IEEE COMPUTER SOCIETY ELECTION

Make Your Voice Heard

Vote for the Leaders Driving Tomorrow's Technology

Vote by Monday, 12 September at 12PM EDT

www.computer.org/election2022

IEEE COMPUTER SOCIETY

IEEE

IEEE Computer Society Has You Covered!

WORLD-CLASS CONFERENCES — Stay ahead of the curve by attending one of our 210 globally recognized conferences.

DIGITAL LIBRARY — Easily access over 800k articles covering world-class peer-reviewed content in the IEEE Computer Society Digital Library.

CALLS FOR PAPERS — Discover opportunities to write and present your ground-breaking accomplishments.

EDUCATION — Strengthen your resume with the IEEE Computer Society Course Catalog and its range of offerings.

ADVANCE YOUR CAREER — Search the new positions posted in the IEEE Computer Society Jobs Board.

NETWORK — Make connections that count by participating in local Region, Section, and Chapter activities.

Explore all of the member benefits at www.computer.org today!

IEEE COMPUTER SOCIETY

IEEE

Conference Calendar

IEEE Computer Society conferences are valuable forums for learning on broad and dynamically shifting topics from within the computing profession. With over 200 conferences featuring leading experts and thought leaders, we have an event that is right for you. Questions? Contact conferences@computer.org.

SEPTEMBER

6 September

- CLUSTER (IEEE Int'l Conf. on Cluster Computing), Heidelberg, Germany

12 September

- ARITH (IEEE Symposium on Computer Arithmetic), virtual

18 September

- QCE (IEEE Quantum Week), Broomfield, Colorado, USA

19 September

- AI4I (Int'l Conf. on Artificial Intelligence for Industries), Laguna Hills, USA
- AIKE (IEEE Int'l Conf. on Artificial Intelligence and Knowledge Eng.), Laguna Hills, USA
- ESEM (ACM/IEEE Int'l Symposium on Empirical Software Eng. and Measurement), Helsinki, Finland
- TransAI (Int'l Conf. on Transdisciplinary AI), Laguna Hills, USA

26 September

- ASE (IEEE/ACM Int'l Conf. on Automated Software Eng.), Ann Arbor, USA
- LCN (IEEE Conf. on Local Computer Networks), Edmonton, Canada

OCTOBER

3 October

- ICSME (IEEE Int'l Conf. on Software Maintenance and Evolution), Limassol, Cyprus
- NAS (IEEE Int'l Conf. on Networking, Architecture and Storage), Philadelphia, USA

4 October

- IMET (Int'l Conf. on Interactive Media, Smart Systems and Emerging Technologies), Limassol, Cyprus

16 October

- MODELS (ACM/IEEE Int'l Conf. on Model Driven Eng. Languages and Systems), Montreal, Canada
- VIS (IEEE Visualization and Visual Analytics), Oklahoma City, USA

NOVEMBER

2 November

- SBAC-PAD (IEEE Int'l Symposium on Computer Architecture and High-Performance Computing), Bordeaux, France

6 November

- IISWC (IEEE Int'l Symposium on Workload Characterization), Austin, USA

7 November

- BIBE (IEEE Int'l Conf. on Bioinformatics and Bioengineering), Taichung, Taiwan

10 November

- ASONAM (IEEE/ACM Int'l Conf. on Advances in Social Networks Analysis and Mining), virtual

11 November

- IPCCC (IEEE Int'l Performance, Computing, and Communications Conf.), Austin, USA

13 November

- SC22 (Int'l Conf. for High-Performance Computing, Networking, Storage and Analysis), Dallas, USA

14 November

- SuperCheck (IEEE/ACM Int'l Symposium on Checkpointing for Supercomputing), Dallas, USA

17 November

- CHASE (IEEE/ACM Conf. on Connected Health: Applications, Systems and Engineering Technologies), Washington, DC, USA

21 November

- ATS (IEEE Asian Test Symposium), Taichung, Taiwan

23 November

- ITNAC (Int'l Telecommunication Networks and Applications Conf.), Wellington, New Zealand

28 November

- ICA (IEEE Int'l Conf. on Agents), Adelaide, Australia
- PRDC (IEEE Pacific Rim Int'l

Symposium on Dependable Computing), virtual

29 November

- AVSS (IEEE Int'l Conf. on Advanced Video and Signal-Based Surveillance), Madrid, Spain

30 November

- ICDM (IEEE Int'l Conf. on Data Mining), Orlando, USA
- ICKG (IEEE Int'l Conf. on Knowledge Graph), virtual

DECEMBER

5 December

- BigMM (IEEE Int'l Conf. on Multimedia Big Data), Naples, Italy
- IRC (IEEE Int'l Conf. on Robotic Computing), Italy
- ISM (IEEE Int'l Symposium on Multimedia), Naples, Italy
- RTSS (IEEE Real-Time Systems Symposium), Houston, USA
- SEC (IEEE/ACM Symposium on Edge Computing), Seattle, Washington
- SMDS (IEEE Int'l Conf. on Smart Data Services), Barcelona, Spain

6 December

- BDCAT (IEEE/ACM Int'l Conf. on Big Data Computing, Applications and Technologies), Portland, Oregon
- BIBM (IEEE Int'l Conf. on Bioinformatics and Biomedicine), Las Vegas, USA
- UCC (IEEE/ACM Int'l Conf. on Utility and Cloud Computing), Portland, Oregon, USA

7 December

- SNPD (IEEE/ACIS Int'l Winter Conf. on Software Eng., Artificial Intelligence, Networking and Parallel/Distributed Computing), Taichung, Taiwan

12 December

- AIVR (IEEE Int'l Conf. on Artificial Intelligence and Virtual Reality), Virtual Conference

13 December

- CloudCom (IEEE Int'l Conf. on Cloud Computing Technology and Science), Bangkok, Thailand

14 December

- CIC (IEEE Int'l Conf. on Collaboration and Internet Computing), Virtual Conference
- CogMI (IEEE Int'l Conf. on Cognitive Machine Intelligence), Virtual Conference
- CSDE (IEEE Asia-Pacific Conference on Computer Science and Data Engineering), Gold Coast, Australia
- ICPADS (IEEE Int'l Conf. on Parallel and Distributed Systems), Nanjing, China
- TPS-ISA (IEEE Int'l Conf. on Trust, Privacy and Security in Intelligent Systems, and Applications), Virtual Conference

17 December

- Big Data (IEEE Int'l Conf. on Big Data), Osaka, Japan

18 December

- HiPC (IEEE Int'l Conf. on High Performance Computing, Data, and Analytics), Bengaluru, India
- iSES (IEEE Int'l Symposium on Smart Electronic Systems), Warangal, India

2023

2 January

- WACV (IEEE/CVF Winter Conference on Applications of Computer Vision), Waikoloa, Hawaii

1 February

- ICSC (IEEE Int'l Conf. on Semantic Computing), Laguna Hills, USA

8 February

- SaTML (IEEE Conference on Secure and Trustworthy Machine Learning), San Francisco, USA

25 February

- HPCA (IEEE Int'l Symposium on High-Performance Computer Architecture), Montreal, Canada

IEEE COMPUTER SOCIETY JOBS BOARD

Evolving Career Opportunities Need Your Skills

Explore new options—upload your resume today

www.computer.org/jobs

Changes in the marketplace shift demands for vital skills and talent. The **IEEE Computer Society Jobs Board** is a valuable resource tool to keep job seekers up to date on the dynamic career opportunities offered by employers.

Take advantage of these special resources for job seekers:

- JOB ALERTS
- TEMPLATES
- WEBINARS
- CAREER ADVICE
- RESUMES VIEWED BY TOP EMPLOYERS

No matter what your career level, the IEEE Computer Society Jobs Board keeps you connected to workplace trends and exciting career prospects.

IEEE COMPUTER SOCIETY

IEEE