# Appendix for C_HANAIMP_16 Exam: Data Security

from the book

## SAP HANA® 2.0 Certification Guide

by Rudi de Louw


"Your all-in-one SAP HANA study guide!"

SAP PRESS

# Appendix for C_HANAIMP_16 Exam: Data Security

Readers of the *SAP HANA 2.0 Certification Guide*, third edition, for the C_HANAIMP_15 exam will know that the book also includes all the relevant updates for SAP HANA 2.0 SPS 04. This enables you to take the next certification exam as well. The topic that wasn't on the C_HANAIMP_15 exam but is on the C_HANAIMP_16 exam is *data anonymity*, which is addressed in this appendix in its wider context of data security.

## Techniques You'll Master:

- Know what data security is
- Understand data masking and data anonymity
- Learn how to implement data security in your SAP HANA models
- Explain analytic privileges

In Chapter 13, we discussed how to implement security on the information models you've built. Data security complements your knowledge of SAP HANA modeling techniques and provides you with more tools to address legal requirements and regulations about who may see what data.

**Real-World Scenario**

You're asked to verify that your information models comply with certain international regulations to ensure data security and privacy. You use data masking and data anonymization techniques with your calculation views to make sure that users' personal data is protected at all times, even when data scientists and analytics require the use of personal data.

## Objectives of this Portion of the Test

The purpose of this portion of the SAP HANA certification exam is to test your knowledge of SAP HANA data security capabilities.

The certification exam expects you to have a good understanding of the following topics:

- SAP HANA data security concepts
- Masking and anonymization for data security and privacy

**Note**
This portion contributes 1% to 2% of the total certification exam score.

## Key Concepts Refresher

The topic of *data security* has always been important with SAP HANA because it's used as a database and a development platform. Management of how and by whom the data is consumed are inherent and expected features.

One of the basic principles of system security is *segregation of duties*, which means that you're not allowed to have only one person responsible for the administration of security. Different people have different responsibilities, and the responsibilities are divided in such a way that one single person can't both create and assign rights to themselves or others that give them special privileges that can be used for corruption or fraud.

As a modeler, you might be creating analytic privileges. Security administrators then add these as part of a role, and the system administrators can assign this role to people in the organization. The security auditors will check the system regularly to ensure that the segregation of duties principle isn't violated.

The system administrators set up the integration to the Lightweight Directory Access Protocol (LDAP) server, such as Microsoft Active Directory, to enable Single Sign-On (SSO) in the SAP HANA system. They might also

configure the Security Assertion Markup Language (SAML) servers for enabling SSO in web browsers. This is required for SAP HANA extended application services, advanced model (XSA), to work securely in organizations.

Roles are only set up in a development server; they are assigned to users in the production server. This segregates privileges across the landscape. Similarly, developers and modelers can only write code and create information models in the development server, but these models are run by end users with the correct roles in the production server.

SAP HANA has had data security features such as encryption of data files (data volumes, log volumes, and backups) and audit logs for many years now. These are configured by the system administrators and checked by the security auditors. You can also encrypt individual table columns with SAP HANA client-controlled keys.

## Data Masking

Data masking is used when you want to reduce the exposure of sensitive information. An example of data masking that you already know is when enter your password. Your password isn't shown as text, but "masked" by displaying dots or blanks.

Another typical use case is when working with credit card numbers. You have to hide this information from database administrators (DBAs) and users with broad access. Even users that work directly with the user data, such as call center agents, shouldn't see credit card details.

Data masking can completely hide such information or just hide parts of it. A call center agent might see just the last three digits of the credit card number to verify with the customer that the correct card is used for a transaction.

Masking data in a column affects the display of the data. The data in the database stays unchanged and can be still used in other information models and programs.

You can find the data masking feature in the **Semantics** node of a calculation view. In the **Columns** tab, you'll see a **Data Masking** column, as shown

on the right side of Figure 1.1. Here we created data masks for both the **PHONENUMBER** and **ACCOUNTNUMBER** fields.

When you click in the **Data Masking** column field, a dialog screen opens where you can add an **Expression**. You can create customizable mask expressions and even use the built-in functions of the expression editor. In Figure 1.1, the account number—that could be a credit card number—is configured so that it will be masked out but still show the last three digits.

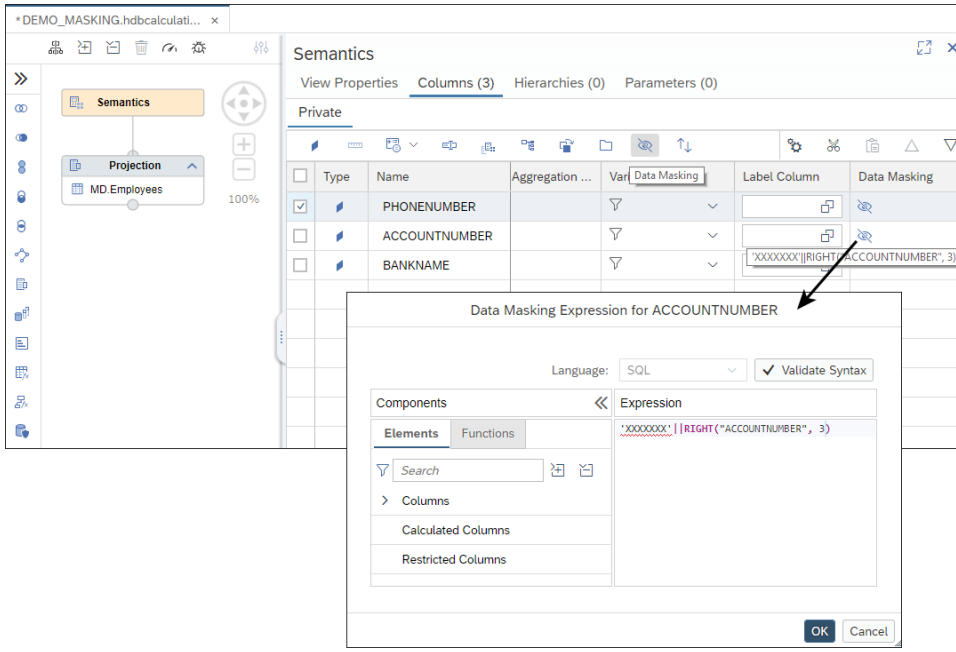


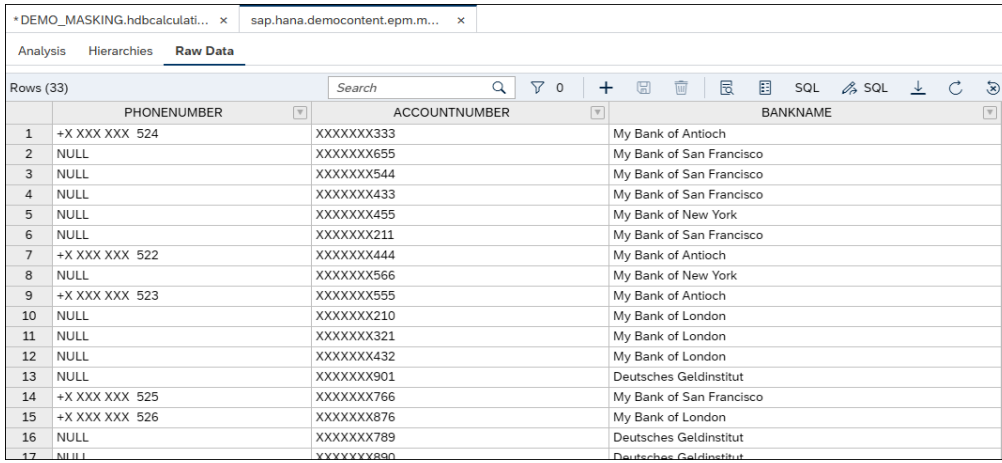

**Figure 1.1** Data Masking in the Semantics Node with a Data Masking Expression

**Note**
Data masking only works for columns of type VARCHAR and NVARCHAR.

Figure 1.2 shows what the masked data will look like to a user. Sensitive data in both the **PHONENUMBER** and **ACCOUNTNUMBER** fields is masked.

Let's say that a few special end users, such as the call center managers, need to see the actual cleartext data. You could create another calculation view

for these users, but that would require double maintenance each time the view needs to be updated. In this case, you grant the **UNMASKED** object privilege to the role assigned to these users, as shown in Figure 1.3.




| | PHONENUMBER | ACCOUNTNUMBER | BANKNAME |
|---|---|---|---|
| 1 | +X XXX XXX 524 | XXXXXXX333 | My Bank of Antioch |
| 2 | NULL | XXXXXXX655 | My Bank of San Francisco |
| 3 | NULL | XXXXXXX544 | My Bank of San Francisco |
| 4 | NULL | XXXXXXX433 | My Bank of San Francisco |
| 5 | NULL | XXXXXXX455 | My Bank of New York |
| 6 | NULL | XXXXXXX211 | My Bank of San Francisco |
| 7 | +X XXX XXX 522 | XXXXXXX444 | My Bank of Antioch |
| 8 | NULL | XXXXXXX566 | My Bank of New York |
| 9 | +X XXX XXX 523 | XXXXXXX555 | My Bank of Antioch |
| 10 | NULL | XXXXXXX210 | My Bank of London |
| 11 | NULL | XXXXXXX321 | My Bank of London |
| 12 | NULL | XXXXXXX432 | My Bank of London |
| 13 | NULL | XXXXXXX901 | Deutsches Geldinstitut |
| 14 | +X XXX XXX 525 | XXXXXXX766 | My Bank of San Francisco |
| 15 | +X XXX XXX 526 | XXXXXXX876 | My Bank of London |
| 16 | NULL | XXXXXXX789 | Deutsches Geldinstitut |
| 17 | NULL | XXXXXXX890 | Deutsches Geldinstitut |

**Figure 1.2** Data Preview with Data Masking

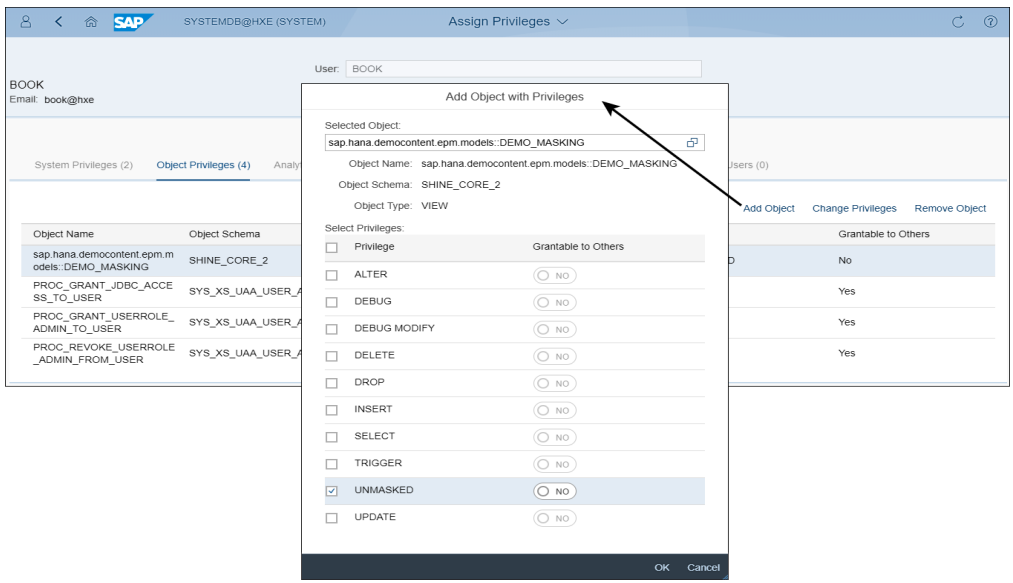



**Figure 1.3** Granting the Unmasked Object Privilege

If you need to use SQLScript instead of graphical calculation views, you can use the `WITH MASK` clause in your SQLScript statements, for example, when creating tables or SQL views.

## Data Anonymization

*Data anonymization* is another feature that came in with SAP HANA 2.0 SPS 03 and was greatly enhanced with SAP HANA 2.0 SPS 04. Data anonymization looks at how to enable analytics on complex data sets without exposing the private information of individuals.

The European Union (EU), for example, has strict regulations (e.g., General Data Protection Regulation [GDPR]) that limit the usage and storage of personal data from any person with an EU passport or identity. How do you comply with such regulations while still gleaning value from data in your systems? Data anonymization enables you to get insights from data that could otherwise not be leveraged due to regulations.

You can use data anonymization in SAP real-time access to anonymized data while the original data stays unchanged. In Chapter 6, we quickly discussed the anonymization node. Let's now look at this node in more detail.

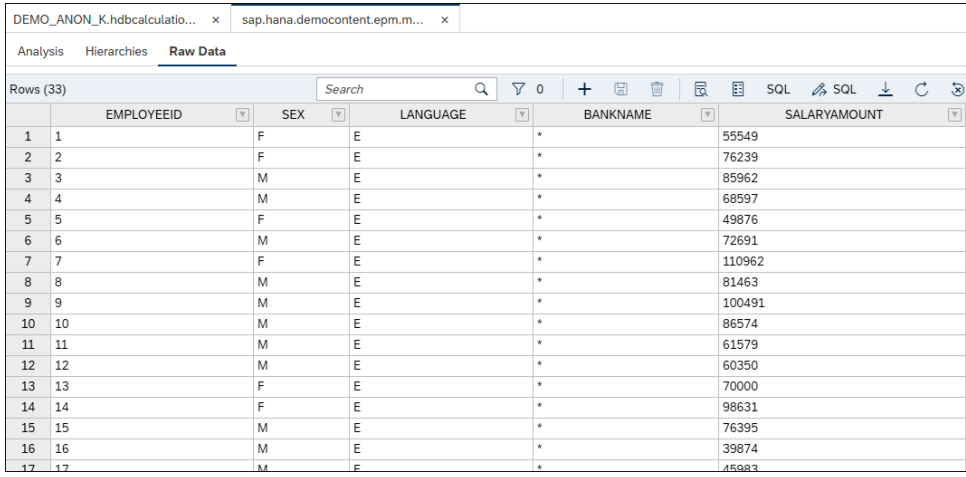
### Data Anonymization Using k-Anonymity

The *k-anonymity* method is intuitive and widely used for modifying data for privacy protection. It hides the individual record in a group of similar records ("crowd"), thus significantly reducing the possibility that the individual can be identified.

Let's say that you create a report of the sick leave taken by employees in your company to determine the reasons for absenteeism. You decide not to include the names of the employees. However, if only two employees had cancer, you might be able to deduce who they are even if their names aren't mentioned.

The “k” in k-anonymity refers to the size of the “crowd” or group. If you set “k” to 20, the data in an individual record will only show if there are 20 people in the group. Obviously, a k-value of 1 is useless as it will show all the data in individual records. The larger the k-value, the more anonymized the data will be.

Figure 1.4 shows the salary and banking details of employees. If you know the gender, language, and name of the bank of the employees, you can deduce the salaries of some employees. In this case, the data anonymization algorithm doesn’t show the name of the bank to prevent users from deducing the salary amount for some users.


DEMO_ANON_K.hdbcalculatio... × sap.hana.democontent.epm.m... ×

Analysis Hierarchies Raw Data

Rows (33) Search 0 SQL SQL

| | EMPLOYEEID | SEX | LANGUAGE | BANKNAME | SALARYAMOUNT |
|---|---|---|---|---|---|
| 1 | 1 | F | E | * | 55549 |
| 2 | 2 | F | E | * | 76239 |
| 3 | 3 | M | E | * | 85962 |
| 4 | 4 | M | E | * | 68597 |
| 5 | 5 | F | E | * | 49876 |
| 6 | 6 | M | E | * | 72691 |
| 7 | 7 | F | E | * | 110962 |
| 8 | 8 | M | E | * | 81463 |
| 9 | 9 | M | E | * | 100491 |
| 10 | 10 | M | E | * | 86574 |
| 11 | 11 | M | E | * | 61579 |
| 12 | 12 | M | E | * | 60350 |
| 13 | 13 | F | E | * | 70000 |
| 14 | 14 | F | E | * | 98631 |
| 15 | 15 | M | E | * | 76395 |
| 16 | 16 | M | E | * | 39874 |
| 17 | 17 | M | E | * | 45983 |

**Figure 1.4** Data Preview with Data Anonymization

Figure 1.5 shows how you can use the anonymization node with k-anonymity. You add an anonymization node by selecting the last icon on the left toolbar. You can add one data source to this node. In the **Details** tab, you choose the **k-Anonymity** method from the dropdown. We’ll look at the **Dif-ferential Privacy** method in the next section. You can also specify the value of **k** (the minimum size of the group), and the columns that we might need to anonymize. You click on each column and then select **Import Values**.

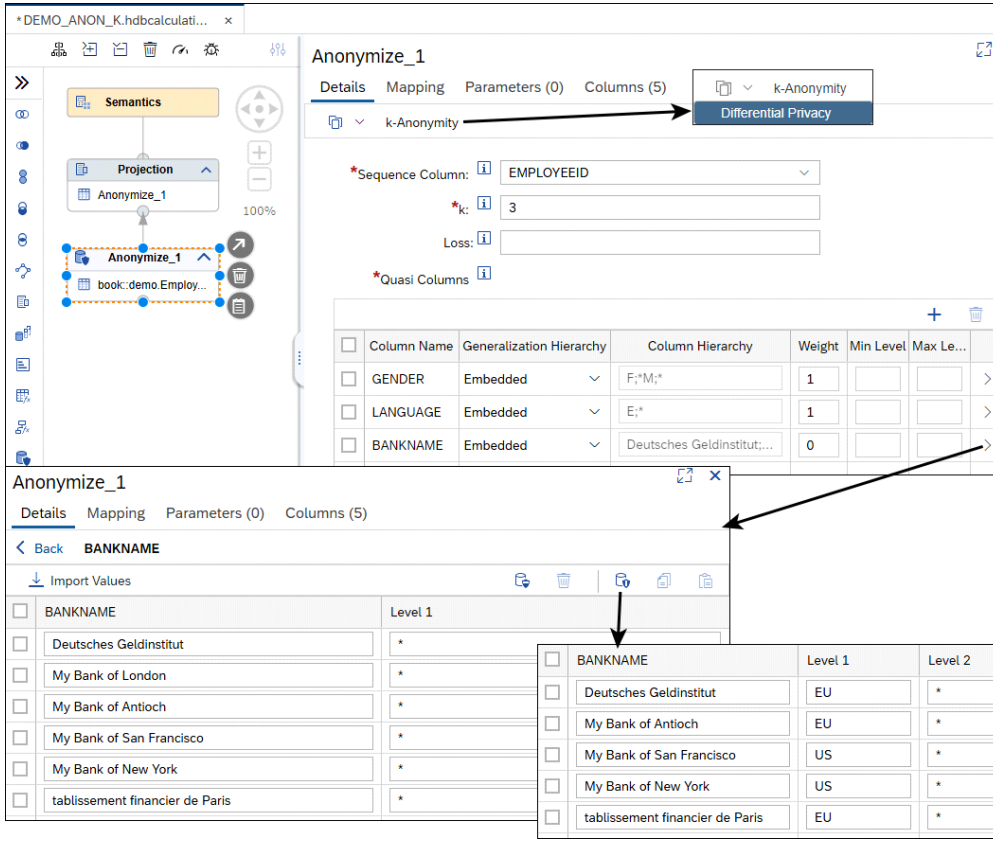



**Figure 1.5** Data Anonymization with k-Anonymity

In Figure 1.4 we saw that the values of the **BANKNAME** are not showing in the report because we have too few records. We might as well remove the field. But, in this case, we can still get some value from the field by grouping banks together. By adding another level, we can group the banks per region to get a minimum of three people (our chosen value for K) per group. Although we don't get the actual name of the bank in our report when there are too few records, we can still see in which region of the world the bank is. This is much better than just getting blank values.

Using k-anonymity doesn't guarantee privacy. If you include the names of the employees, your efforts will be in vain. Be careful, think, and double-check with other people (e.g., auditors) to ensure that you've applied it correctly.

In the preceding example, the system chose to show you the gender (**SEX**) of the people, but not their bank. The system could also have chosen to show the **BANKNAME** instead of **SEX**. What if you wanted to always see the **SEX** field? In SAP HANA 2.0 SPS 04, you have control over which fields are prioritized. The **Quasi Columns** have extra columns called **Weight**, **Min Level**, and **Max Level**. By specifying the **Weight** of the **SEX** field as **1** and the **Weight** of the **BANKNAME** field as **0**, you ensure that if the system has to choose between these two fields, it will always choose to show the **SEX** field.

The **Min Level** and **Max Level** parameters refer to the level as illustrated in Figure 1.5 with the **BANKNAME**. At **Level 1**, we use the name of the region (in this case, **EU**) rather than specify the name of the bank.

You also have control over whether you want to use the entire result set or just a specified percentage of the available records. If you use all the records, some fields may appear blank to protect the privacy of individuals. With the **Loss** parameter, you can specify that the system must drop the records for those individuals and give you a result set with no blank fields where possible. The **Loss** parameter requires a value between **0** and **1**, indicating the percentage of records that you're prepared to lose, for example, 0.5 indicates 50%.

The additional controls available in SAP HANA 2.0 SPS 04 enable what is called *l-diversity*. This is applied in addition to k-anonymity. Strictly speaking, k-anonymity is just a special case of l-diversity where *l=1*.

The Generalization Hierarchy column allows you to also use external hierarchies. If you've created a SQL hierarchy, you can reuse it here, instead of having to define duplicate hierarchies. This simplifies maintenance.

### Data Anonymization Using Differential Privacy

The second method you can choose from the dropdown in the **Details** tab is **Differential Privacy**. Differential privacy allows you to gain statistically valid insights from your data while still protecting the privacy of individuals. It works by modifying individual records without changing the aggregated statistics. For example, you can "scramble" salary data so that no one can find out an individual's salary, but the aggregates (sum, average, etc.)

still stay almost the same. Here you apply noise to the data in a column to hide sensitive information.

Differential privacy is a formal statistical method that guarantees privacy within the sensitivity levels that you specify. It can anonymize salary data but still allow users to gain valid statistical insights from the data.

Figure 1.6 shows how the **Details** tab changes when you select the **Differential Privacy** method. You set the sensitivity levels with the **Epsilon** and **Sensitivity** values, and the data in the **Noised Column** gets anonymized.

Figure 1.7 shows the results in the **Data Preview**. The **SALARYAMOUNT** column has been anonymized. You can see that the statistical method isn't that intelligent because employee **12** now has a negative salary. That poor employee now seemingly has to pay to work for the company.

While the data for each record might be rubbish, you can still gain valid statistical insights. Analytics on the data create valid statistical results. In Figure 1.8, the average salary is calculated for both the real and the anonymized data set.

The results for the anonymized data set are extremely close to the real values—certainly within a known error margin that depends on the number of records. The more records you use, the closer these two results will be.

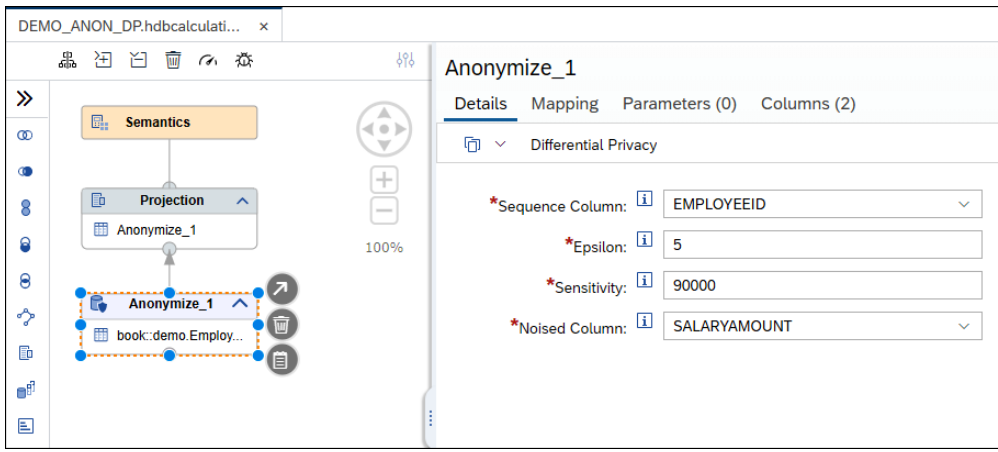



**Figure 1.6** Data Anonymization with Differential Privacy

DEMO_ANON_DP.hdbcalculati... × sap.hana.democontent.epm.m... ×

Analysis Hierarchies **Raw Data**

Rows (33) Search 0 SQL

| | EMPLOYEEID | SALARYAMOUNT |
|---|---|---|
| 4 | 4 | 64132.950605186255 |
| 5 | 5 | 42922.01321477542 |
| 6 | 6 | 69677.40909565346 |
| 7 | 7 | 111227.95433511464 |
| 8 | 8 | 127341.92112300106 |
| 9 | 9 | 123837.36288429565 |
| 10 | 10 | 92865.56385951245 |
| 11 | 11 | 51258.728320671864 |
| 12 | 12 | -9460.128537202312 |
| 13 | 13 | 77473.05978663219 |
| 14 | 14 | 125108.17768731393 |
| 15 | 15 | 24078.52528268342 |
| 16 | 16 | 33648.530018259306 |
| 17 | 17 | 128519.29118557046 |

**Figure 1.7** Data Preview with Data Anonymization Using Differential Privacy

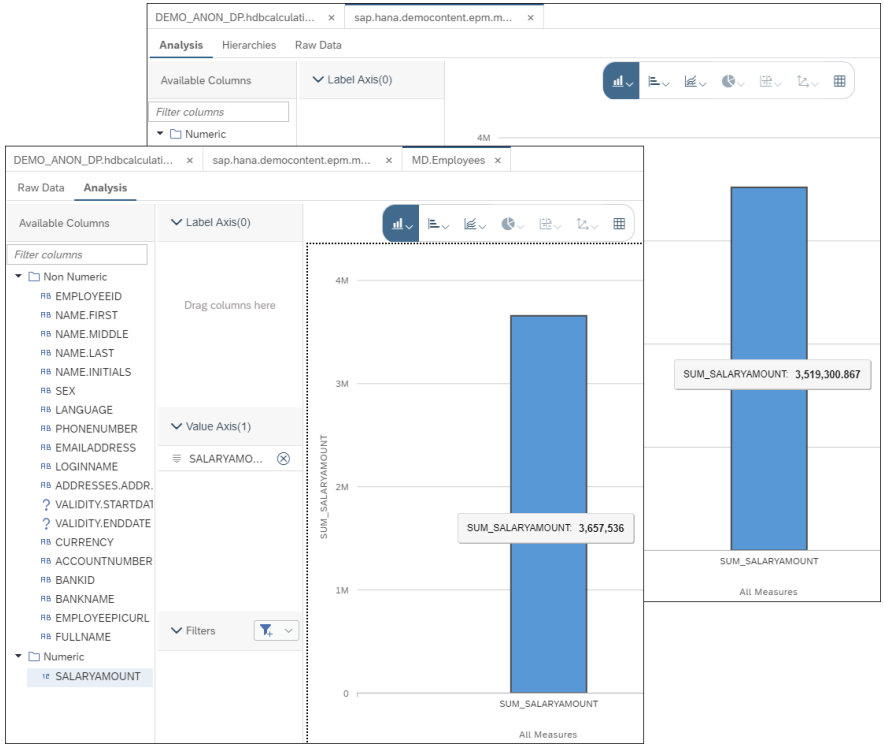



**Figure 1.8** Comparing the Aggregate Results of Data Anonymization with Differential Privacy to the Aggregate Results of the Source Data

The SAP HANA cockpit provides an anonymization report that you can use for technical teams and data protection officers. Figure 1.9 shows the **Anonymization Report** area where you can view all calculation views with anonymization scenarios, methods, and parameters.

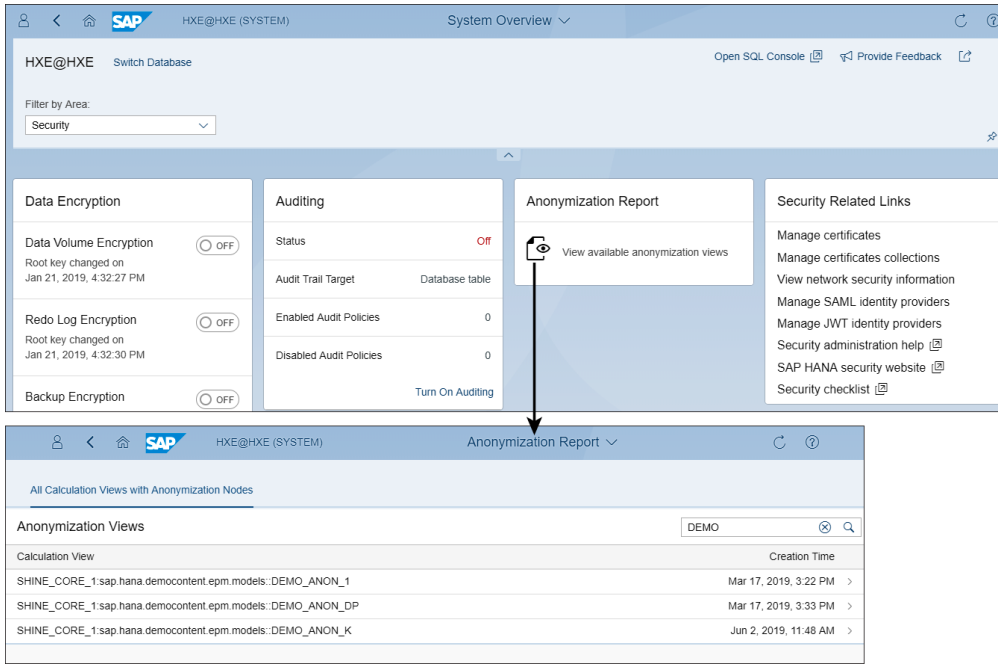



**Figure 1.9** Anonymization Report in the SAP HANA Cockpit

## Choosing the Right Tool

Medical information is very personal and may be embarrassing to people should it become public. Let's look at some of the people and processes involved when handling such data, as well as which SAP HANA feature you should use in each case:

- When patients go to a medical specialist, their personal data needs to be captured. This will include medical insurance information and sometimes banking details. The receptionist needs to type in this data if this is the patient's first visit to this specialist. If the patient has been there before, you can use data masking to show only partial informa-

tion that the receptionist can use to check the information. The receptionist shouldn't be able to see the patient's medical history. You can block access to this using standard authorization concepts.

- Some patients might not like the fact that the receptionist can see their personal information when capturing the information. These patients might use an app such as LogBox (www.logbox.co.za) to manage their own data. Such apps use blockchain technology to prove that the patient gave permission to share some of their personal information with the medical specialist. If you work with such an app, you can use analytic privileges to allow patients to change only their own data.
- All the medical insurance companies in a country need data about what patients were treated for to calculate the medical insurance rates and risk profiles. You can use data anonymization to calculate aggregates correctly without allowing them to see the data of each patient's visits to specialists.

This completes our tour of the data security aspects you need for working with information models in SAP HANA XSA and SAP HANA Deployment Infrastructure (HDI) containers.

## Practice Questions

These practice questions will help you evaluate your understanding of the topic. The questions shown are similar in nature to those found on the certification examination. Although none of these questions will be found on the exam itself, they allow you to review your knowledge of the subject. Select the correct answers, and then check the completeness of your answers in the "Practice Question Answers and Explanation" section. Remember, on the exam, you must select all correct answers and only correct answers to receive credit for the question.

1. What do you use to hide the identifying data of any person by ensuring that the data columns are only displayed when they are guaranteed to be part of a larger group?

☐ **A.** Data masking

☐ **B.** Data anonymization using k-anonymity

☐ **C.** Data anonymization using differential privacy

2. What do you use to ensure that each account holder of a bank account can only see his own identity document details in a calculation view?

- ☐ **A.** Data anonymization
- ☐ **B.** Analytic privileges
- ☐ **C.** Data masking
- ☐ **D.** Audit logs

## Practice Question Answers and Explanation

1. Correct answer: **B**
   You use data anonymization using k-anonymity to hide the identifying data of any person by ensuring that the data columns are only displayed when they are guaranteed to be part of a larger group.

2. Correct answer: **B**
   You use analytic privileges to ensure that each account holder of a bank account can only see his own identity document details in a calculation view.

## Takeaway

You should now have a good overview of data security in SAP HANA, especially as it relates to information modeling. You should understand when to use data masking and data anonymity, as well as how they affect users' privacy.

## Summary

In this appendix, to help you prepare for the C_HANAIMP_16 exam, we looked at data security and privacy using data masking and data anonymization.

Best wishes for your exam!