SAP Data Anonymization – FAQ

PUBLIC

(1) WHEN SHOULD I USE ANONYMIZATION?
Numerous laws in different countries restrict the processing of personal data, especially of sensitive personal data. However, if the data is anonymized, these restrictions don’t apply anymore, meaning that the data can be processed and analyzed as required.

So the main usage scenario for anonymization is to create data sets that can be analyzed without identifying individual persons.

The second usage scenario is to protect confidential company data. If a company contributes information to a data set, but does not want anyone to know their true data. In this case, the company itself is the individual and must be protected. A typical use case for this scenario is benchmarking within industries.

(2) WHAT CAN I DO WITH ANONYMIZATION THAT I COULDN’T DO BEFORE?
In many situations, users have limited access to data sets due to either legal or company governance restrictions. Anonymization is a way to fulfill privacy (or governance) requirements by ensuring a released data set no longer contains any sensitive information that can be linked to a person.

With anonymization you can safely give others access to a data set containing sensitive data without compromising the privacy of individuals. If you are an application developer, you can access data that was previously unavailable due to privacy concerns. If you are an analyst, you can get additional information and use it for your tasks without having to worry about privacy issues.

(3) CAN I ANONYMIZE MY DATA SET BY MASKING SENSITIVE COLUMNS?
Masking and anonymization are different concepts and are used in different situations. Masking takes a single value in the data set and applies a pre-defined filter. This means that only a certain part of the information may be changed or made invisible. A popular example is removing certain digits of credit card numbers. Someone with access to that specific line item can only see the remaining digits.

Anonymization, on the other hand, takes the complete data set under consideration and tries to modify it in a way that no sensitive information can be linked to any individual. If the credit card number is considered sensitive, then I do not know to whom it belongs at all.

Masking and anonymization cover different use cases. Consider a scenario where support agents need to be able to see a fraction of the credit card number for verification. Masking is the feature to choose here since the agents need the link between the individual and the number. Now imagine a scenario where you want to analyze typical consumer behavior. In this case, you do not need a link between a specific individual and their credit card data.

Generally speaking, masking should not be used for anonymization. Either it destroys too much information, or there are privacy issues. For more information and a detailed explanation please refer to question “Why is anonymization difficult?”

(4) WHY IS ANONYMIZATION DIFFICULT?
Anonymization is difficult for two reasons. First, simple measures, like removing identifiers, are not enough to anonymize a data set. Second, anonymization should not lead to a data set that is of no use. Both can be explained with the following example.

Consider a patient data set containing illnesses and additional information like the weight.
Name | Birth | City     | Weight | Illness
---|---|---------|--------|--------
Paul | 07-1975 | Walldorf | 82 kg | AIDS
Marin | 10-1975 | Hamburg | 110 kg | Lung Cancer
Nils | 01-1975 | Munich | 70 kg | Flu
Annika | 09-1097 | Berlin | 58 kg | Multiple Sclerosis

Of course, such data must be kept private; no one should know which illness a certain person has. However, analysis of such data is important to gain new medical insights. An analyst (doctor) might ask “How many people who weigh more than 95 kg have cancer?” to deduce certain patterns.

An intuitive step to anonymize the data would be to simply remove the name or replace it with a pseudonym. The pseudonymized table would look like this.

Name | Birth | City     | Weight | Illness
---|---|---------|--------|--------
0c4a67 | 07-1975 | Walldorf | 82 kg | AIDS
df89aa | 10-1975 | Hamburg | 110 kg | Lung Cancer
305be2 | 01-1975 | Munich | 70 kg | Flu
7422c2 | 09-1097 | Berlin | 58 kg | Multiple Sclerosis

The doctor would still get the correct answer to his question since counting cancer patients weighing over 95 kg does not involve the “Name” column. However, this is not proper anonymization since someone who knows that Martin is overweight can still re-identify him as the second row in the data set because of the weight column. One might think that removing the weight column as well would prevent such an attack. Of course, this particular attack will not work anymore, but the analyst is now unable to determine the number of overweight cancer patients. Additionally, attacks following the same pattern can also happen with the “City” and the “Birth” column.

Consequently, anonymization requires structured methods to prevent such attacks but keep the utility of the data as high as possible. Please refer to the documentation to learn more about the anonymization methods available.

(5) WHAT IS THE DIFFERENCE BETWEEN DYNAMIC DATA MASKING AND DIFFERENTIAL PRIVACY? I CAN ALSO CALCULATE AVERAGES USING DYNAMIC DATA MASKING WITHOUT EXPOSING ANY SENSITIVE DATA.

Yes, masking will keep the utility of the data but usually does not lead to an anonymized data set. Please refer to the question “Why is anonymization difficult?” for more information.

(6) HOW DOES THIS TECHNOLOGY WORK?

We apply research-based methods to the data to anonymize it. All methods are structured approaches with certain guarantees regarding privacy. For more information on the methods and detailed explanations on how they work, please refer to the documentation.

(7) WHICH ANONYMIZATION METHOD DO I CHOOSE?

Choosing the method usually depends on the privacy requirements and your data set. The method differential privacy has stronger and provable guarantees regarding privacy. However, applying differential
privacy tends to have a larger utility loss. k-anonymity, on the other hand, does not have such a strong statistical privacy guarantee but keeps more utility and is also more intuitive to understand.

In the current implementation, the differential privacy method is limited to numerical columns, whereas k-anonymity can process many different data types (both numerical and text).

The final decision on the method and the parameters usually has to be approved by data privacy officers within your organization.

(8) WHAT ARE THE PREREQUISITES FOR USING THE DIFFERENT ANONYMIZATION METHODS (E.G. SIZE/COMPLEXITY OF DATA SET)?

Making a data set k-anonymous is a computationally complex task. Thus, the service might take longer and require more memory to k-anonymize a data set compared to applying differential privacy. Due to the nature of a stateless service, there are limitations on the file size you can upload and on the processing time.

(9) WHAT HAPPENS IF I SHARE AN ANONYMIZED DATA SET WITH EXTERNAL RESEARCHERS, BUT THEY ONLY USE HALF OF THE DATA? DOES ANONYMIZATION STILL WORK?

Anonymization always covers the complete data set provided regardless of how much used. Even though only single line items might be queried and the remaining part is ignored, it does not change anything with respect to the privacy guarantees.

(10) WHICH GUARANTEES CAN YOU GIVE THAT DATA IS TRULY ANONYMIZED? WHAT CAN I TELL MY DATA PROTECTION OFFICER?

In a nutshell: k-anonymity makes an individual indistinguishable within a group of at least k members. Applying the differential privacy method makes sure that an individual contribution does not change the probability of a query result. You'll find more detailed explanations on the guarantees in the documentation. Additionally, both methods are included in the EU Opinion 05/2014 (http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf).

(11) DIFFERENTIAL PRIVACY: HOW CLOSE ARE RESULTS CALCULATED ON ANONYMIZED DATA TO THE ACTUAL DATA?

This depends on the parameters set. Rule of a thumb: The lower the sensitivity and the larger the provided epsilon (which reflects the impact of an individual contribution on the probability of any outcome), the fewer number of aggregated records are required for a precise result close to the original one.

(12) WHAT ARE THE ATTACK SCENARIOS THAT ANONYMIZATION PROTECTS AGAINST?

See the documentation.
I’VE HEARD THAT DIFFERENTIAL PRIVACY/K-ANONYMITY IS BROKEN. IS THIS TRUE?

“Broken” is a word usually used in the context of encryption systems. It means that it is possible to decipher an encrypted message without requiring the appropriate key, or in other words, without having the legitimation to read the clear text message. Anonymization however works differently: The whole purpose is to provide clear text data for analytics. Of course, deciphering and seeing the whole original message does not make sense.

Anonymization methods give certain guarantees. If applied correctly, differential privacy gives strong statistical guarantees and k-anonymity gives intuitive guarantees of one individual being indistinguishable from another within a certain group. The k-anonymity guarantee does not cover any detailed estimation on the information within a certain group: Assuming a group contains potential terminal diseases, then an attacker might gain some information without knowing the exact disease. However, this does not mean that k-anonymity is broken since the intuitive guarantee still holds. It rather means that k-anonymity might not be the method of choice or should be configured differently.

To be on the safe side, anonymization is usually not used standalone, but together with standard data protection mechanisms like access control/authorization. For instance, even though the data set is anonymized only a limited number of persons have access to it.

ARE YOUR ANONYMIZATION METHODS CERTIFIED?

At this point in time, there are no certifications available for anonymization methods since only complete end-to-end processes can be compliant in terms of privacy legislation and an anonymization method is only one piece of this. However, the methods are mentioned in the EU Opinion 05/2014 (http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf) and are commonly known as being able to protect privacy.

APART FROM DIFFERENTIAL PRIVACY AND K-ANONYMITY, WHAT OTHER ANONYMIZATION METHODS ARE THERE?

Differential privacy is a criterion rather than a specific method. The implementation we provide in the service is an implementation based on the La Place noise “method”. Of course, there are different instances available covering other data types such as GPS coordinates and others.

k-anonymity also has some successors providing additional guarantees for the information in a certain group, that is, how many different entries of the sensitive attribute are in a k-group. However, in contrast to differential privacy, k-anonymity and derivates do not provide strong statistical guarantees.

MORE INFORMATION

http://www.sap.com/data-anonymization