Sogeti

Addressing Regulatory Compliance Around Data Protection

GDPR – The Effects on Test Data Management

Table of Contents

GDPR – The Effects on Test Data Management

Introduction

The EU “General Data Protection Regulation” (GDPR) will take effect in May 2018. The regulation was adopted in April 2016 and will replace the data protection directive 95/46/EC after a two-year transition period. It primarily regulates the handling and protection of personalised data of EU citizens.

What consequences does this new regulation have on tests, test results and organisations? And what should compliant Test Data Management (TDM) look like?

Could Test Data Management Systems (TDMS) help and what criteria will play a decisive role?

Because of the complexity of GDPR, this document will review the main aspects that concern test data management and how it will affect your organisitation.

Main changes to the new EU GDPR

The following list summarises some of the most important changes (see reference documents listed [1] at the end of this paper):

1. Companies need to keep track of all personalised data across the entire company.
2. Data use is subject to a strict purpose. It may only be used for clear and legitimate purposes. If the data is to be used for a different purpose than the one for which it was originally collected, for example tests, consent must be obtained for data use from the persons concerned.
3. Companies need to be able to provide the supervisory authorities with all relevant documentation of the mechanisms the company used in order to track and adequately control personal data across all systems and platforms.
4. EU citizens have the right to be “forgotten“. Companies must demonstrate that they can remove any instance of personal data from all systems and platforms at the request of the person concerned.
5. In the case of a data breach, companies have the duty to notify all persons concerned and the Data Protection Authority without any delay and within 72 hours. A data breach occurs if data is accidentally or unlawfully destroyed or changed, if it is lost or when it is accessed or disclosed by unauthorised persons. The incident does not have to be reported if there is no risk of violating the rights and freedom of the person affected by the data breach.
6. Individual rights have been strengthened. For example:

    a. Anyone concerned must be informed about the nature and use of his data on request. The following information should be part of the information:

    - For what purposes is the data used?
    - What is the exact data?
    - How long will the data be stored?
    - Where does the data come from?
    - Is the data used for automated decision-making?
    - Is the data transferred to third world countries and how is data protection ensured in these countries?

b. The affected person has the restricted right to rectify, delete and limit the processing of his / her data.

c. With GDPR, a right of data transfer is introduced. This refers to the export of data provided by a person in a machine-readable format. In this context, the person concerned may also request the transmission of their data to another service provider.

7. It is also important that the regulation re-regulates transnational data transfer and extends data protection for EU citizens to non-EU countries. If someone is performing offshored test activities and transferring data for test purposes to third world countries, they must now ensure that the EU data protection requirements are at least equivalent.

In article 25 of the regulation, the principles of “Privacy by Design” and “Privacy by Default” are explained. Measures for the protection of personal data should, among other things, be based on state of the art. What is technically feasible should also be done, unless there are weighty reasons against it. This balancing is part of the required documentation.

Failure to comply with the new rules and regulations could mean facing significant fines. The fines are separated into two different types:

- Breaches related to the controller and processor obligations, certification body obligations or monitoring body obligations: Up to €10M or 2% of total worldwide turnover, whichever is the greater (GDPR Article 83(4))
- Breaches related to the basic principles of processing, content, data subject rights, transfer of data, non-compliance: The highest fine states up to €20M or 4% of total worldwide turnover, whichever is the greater (GDPR Article 83(5))

Not all of these changes affect test data management and test operations, but the “right to be forgotten“, for example, can affect the management of test data.

Discovery and documentation

A first step is to analyse and document where, how and for what purpose production and personal data is used in your organisation. It is important that all areas, where personal data is used, are included in this analysis. This also includes development and testing. On the basis of these results, a strategy can be developed and appropriate measures can be derived. It is advisable to consult the Data Protection Supervisor at a very early stage, because the strategy or measures must ultimately be coordinated with him or her.

Change of purpose

One way of dealing with EU requirements is to seek permission from the person concerned to use their personal data in tests. The probability is quite high that your request will not be granted. Perhaps it is possible to extract and manage data in such a way that only the personal data is used for test purposes when consent has been given. The problem becomes far more complex due to the fact that, in most cases, the person concerned is not only the customer, but also the supplier. The third party involved can also be the shareholder or the employee. This is because personal data is stored in so many places.

Data privacy

The use of production data for testing requires changes in a company’s test data management and poses challenges in terms of documentation and data management.

Organisational and technical security measures must be taken and risk management for test data must be introduced so that data breaches such as data theft or unintentional publishing can be made more difficult and the risks can be mitigated. In addition, appropriate reporting channels must be established in order to meet the new 72-hour reporting deadline.

Right to be “forgotten“

If a person exercises his right to be “forgotten”, all corresponding personal data must also be removed from test data in all test environments. This may lead to a daily update of test data. This is not an unlikely scenario for companies with millions of customers.

Test environments

If production data is used for testing, then the same requirements apply in the test environment as for the production environment, with regards to security and access protection. These requirements increase the cost and administrative expense of a corresponding test environment. Specific challenges in this context are the use of near-shore and offshore Test Centres in non-EU countries.

Because of the points listed above, the use of production data for testing is associated with a high level of organisational effort in terms of workload, as well as many other problems. Therefore, you should check whether alternative routes for the creation of test data can be explored.

Alternatives for Test Data Generation

Production data is only one source for test data. Several procedures can be used to generate test data:

- Synthetic test data
- Anonymisation and pseudonymisation of production data
- Masking production data
- Encrypting production data

Each method has its advantages and disadvantages. However, all test data should be generated in such a way that it is immediately recognisable as test data and not production data. In this way you can avoid unnecessary data issues or reports of data loss.

Generation of synthetic test data

A good alternative is the exclusive use of synthetic test data. Synthetic test data generation is the only alternative in which no production data is used for test purposes. In everyday testing, however, synthetic data is often controversial. A common argument is that the data is not realistic enough. Moreover, it is not trivial to create consistent test data in complex IT infrastructures. This sometimes takes up a lot of time, and staff effort.

Empirically, it is also shown more often that artificial data does not reliably detect various fault conditions, which can lead to an increased defect rate in production systems. As a result, there are developers and testers who prefer production data in the test environment. The use of synthetic data testing can therefore have an acceptance problem.

However, there are also clear advantages. If synthetic test data is created carefully, it can be better adapted to the necessary test cases than production data. In addition, there are tools such as test data generators that can facilitate this.

Anonymisation and pseudonymisiation of test data

Another option is the anonymisation and pseudonymisation of production data. The changed data can be used in tests, without being exposed to any risk of theft of production data.

Anonymisation means that personal data is changed in such a way that a person can no longer be identified by the data.

In the case of pseudonymisation, usually multi-character, letter or number combinations, are used in order to exclude or substantially hinder the identification of an individual, by replacing the identifiers.

In contrast to anonymisation, pseudonymisation is used to keep references to different data sets together. Another way is to provide an available key or an applicable rule set that allows the data to be assigned to a person. Without a key or applying the rule, it becomes impossible or extremely difficult to identify individuals.

Often, in the process of data processing, anonymisation is achieved only via pseudonymisation. For example, lookup tables provide pseudonyms that replace names with other names or strings with certain rules. If the lookup table is deleted after this process, you get anonymous data. However, if the lookup table is kept, the data is only pseudonymised. Traceability using the table remains possible.

However, you are not always on the safe side when it comes to anonymising personal data.
De-anonymisation is still possible with feature combinations.

Masking production data

Suppliers of Test Data Management Systems (TDMS) often offer a wide range of masking techniques. The term “data masking” is frequently used out of context and often involves pseudonymisation and anonymisation. Indeed the result of applied masking techniques could be pseudonymisation, anonymisation, or encryption.

However, masking can also work independently. In this way, it serves to (partially) conceal information. This kind of data masking is well-known to anyone who does online shopping, where for example, just the last four digits of the credit card number are displayed. The previous digits cannot be seen or are replaced with an “x”. Full masking is used when entering passwords. In this case, only a placeholder is displayed.

A list of all masking techniques would be too complex. However, it should be mentioned that there is a basic division into static and dynamic masks.

Dynamic means that the tester or developer will only see masked or partially masked data in the affected output field. The data is masked by the database query “on-the-fly”. However, the underlying database contains the data in unmasked form. In static data masking, the data in the database is masked accordingly. A good overview of the various masking techniques can be found in the references section [2].

Encryption of test data

With this method, the stored data is either encrypted in the database or only encrypted when retrieved from the database. The tester only sees the encrypted data in the front end.

If, however, all fields, first names, surnames, addresses, or unique numbers, are stored or provided only in encrypted form, this contradicts the intention of most test requirements and test cases since the correctness of the output must be checked frequently.

In addition, encryption is still very resource-intensive, that is why large amounts of data are usually not encrypted. Normally, test processes are complicated and slowed down by encryption. Therefore, encryption is often not the test method of choice, but in individual cases it can help to protect data.

Again, keys and encryption algorithms can be compromised. Therefore encryption does not provide 100% protection.

Processing Test Data with Test Data Management Systems (TDMS)

If someone in your organisation is dealing with test data, this person will quickly find that it's rare to use homogeneously generated test data. Depending on the test level, test method or requirement, you might use synthetic data, anonymised data, otherwise masked data and/or production data. It is either used serially (e.g. synthetic data in the unit test, production data in the acceptance test) or it's mixed together as the initial data set for all test environments.

Since the process for the administration of heterogeneous test data as well as the generation of test data is very complex, no matter what method is applied, the use of a Test Data Management System (TDMS) is recommended. Such systems offer integrated procedures for masking, anonymising, pseudonymising, and data encryption, and make them available for testing. Some systems also include test data generators.

The possibilities offered by these TDMS are also available via other means such as database specific tools, SQL, scripts, self-created programs as well as data import and export tools.

The main advantage of a TDMS is that it can create and provide heterogeneous test data more quickly. In addition, they offer advantages in the administration of test data and create more transparency. Therefore, they help to fulfil the requirements of the new EU legislation and can support your company in implementing your requirements. Since the systems are quite technical, you should allow sufficient time for the implementation of such a system.

The TDMS market is large, and there is a rich selection of complex TDMS to smaller applications that only meet specific tasks. Nearly every major software vendor offers these solutions. Some vendors are specialised by industry and others are innovative, emerging manufacturers. Which is the right tool, as always depends on your own company's requirements and circumstances. Articles [3] and [4], in the references, provide a good overview of such tools.

TDMS – selection criteria

When compiling the selection criteria for the selection of a Test Data Management System, many things have to be considered. What features need to be included, on which platform does it need to operate and which systems should be supported? The main features of TDMS currently available are:

- Data masking
- Sub-setting
- Data profiling
- Sensitive data analysis
- Data discovery
- Data decommissioning
- Data archiving
- Synthetic data generation
- Automation
- Packages for compliance with legal requirements

Not every suite has all of the above features. It's also important to ask the question: which database management system (DBMS) needs to support the TDMS. Again, not every vendor supports every interface, even if the range is quite large. The same applies to operating system support. In addition to the broad-based providers, there are also those who only specialise in one operating system.

Many TDMS work with enterprise suites or are even certified for them (e.g. SAP certification). However: Not every TDMS supports or collaborates with any enterprise suite. This also applies to integration with test tools from other manufacturers. HPE's ALM is usually supported. Some TDM solutions also have built-in test tools.

Data masking

The masking of data can be relatively complex. Especially in the case of complex data structures and extensive business processes, and where technical knowledge (programming, database) and skills in the area of business analysis and testing are important. By using a good tool the work can be facilitated and the quality can ensured. The tool should be able to replace larger data structures. Data pools, such as addresses or bank connections, should be used for this purpose. It is also important to be able to integrate a company's own data sets. The methods used must be state of the art and make reverse engineering impossible. They must also allow predetermined frequency distributions to be taken into account (e.g. distribution, geographic distribution). With the tool, the process of data masking should ultimately be automated. This makes it possible to create new data without a specialist being involved.

Sub-setting

A very important feature is "sub-setting". Since data sets from production data can be very large (potentially several terabytes), it is useful to reduce data for testing. Selection criteria must therefore be found, which makes it possible, to extract only a smaller corresponding data selection. This is often a special strength of TDMS, enabling you to use test resources more economically.

Data profiling

Data profiling describes the process for the analysis of data (e.g. in a database) by different analysis techniques. This process is largely automated by a TDMS and generally supports various methods. Common data profiling procedures are attribute, record and table analysis. This allows data quality problems to be detected and the causative data to be identified. At the same time, the quality of the analysed data can also be measured.

Sensitive data analysis, data discovery

These two terms describe the analysis of data on data-privacy compliant data and fields. The systems offer options to analyse, classify and categorise this data on the basis of compliance with legal requirements. The results can then be used as a basis for masking. Data discovery is used less consistently as a term. Depending on the manufacturer, it also describes the analysis of sensitive data, but sometimes this only affects the registration of data sources in the system. In some cases, it also describes the analysis on primary and secondary keys, which enables the systems to extract consistent and personal data (when sub-setting).

Data archiving

An important feature can be archiving the test data. Most TDMS are not only able to distribute centralised data, but also archive it. This makes it possible, e.g. to save and archive versions of the test data, adapted to the respective test stage, which can then be used again and again if the test requires it. In this context, there are also systems with test data self-service, which enable test managers or testers to independently import defined test data into test environments.

Data decommissioning

The phasing-out of data may be an important attribute of your TDMS, especially with regards to the right to be “forgotten” of the new EU directive and legislation. But the removal of obsolete archived test data can also help to reduce resources while testing. However: Not all TDMS solutions offer this.

Synthetic data generation

With regards to the new EU regulations, the creation of synthetic data is an important attribute, which should not be omitted from your choice of TDMS, if you decide on such a solution. For systems that do not have this feature, you must either purchase an additional tool or use native database tools and scripts.

Automation

A significant feature of a TDMS is the provision of test data generation functions. However, in the context of progressive test automation, it becomes increasingly important to provide test data “on-demand”. Therefore, it is important that a TDMS can demonstrate the process of test data generation and provisioning as a workflow. The system should be able to generate test data on request and then distribute the data to other systems, enter it into databases or provide it for test automation tools.

Packages for Compliance with Legal Requirements

Most manufacturers of TDMS solutions already provide pre-built masking packages to meet specific legal requirements. Often you will find the following is available:

- **PII -** Personally Identifiable Information, follows US laws and EU directives
- **PCI -** Payment Card Industry Data Security Standard
- **PHI -** Protected Health Information, including medical data from patients and purchase / payment histories

Only the three most important acronyms are listed above. There are other legal requirements which need to be taken into account (NPI, HIPAA, FERPA, etc.) especially for the North American arena. Meanwhile some vendors also advertise “EU GDPR readiness” for 2018.

PII – Personally Identifiable Information

PII is a legal concept that was initially used in the context of US legislation. In the meantime, however, it has been used to summarise all existing international legal regulations. In this sense, old and new EU directives are part of the development and the overall PII concept. Within PII, it is assumed that certain personal data enables you to identify someone directly (full name, address, date of birth, etc.). However, there is also data, which indirectly enables the identification (e.g. surname, web cookies, etc.). In addition, you must always take into account that a combination of several characteristics and attributes - even if they are not directly personal - always has the potential to clearly identify a person.

PCI - Payment Card Industry Data Security Standard

The Payment Card Industry Data Security Standard is a set of rules for processing credit card details in payment transactions. It is supported by all major credit card companies. All companies and service providers that store, transmit, or process credit card transactions must adhere to these regulations. If these regulations are not met, criminal fees and restrictions may be imposed or the acceptance of credit cards may be prohibited. There are 12 general requirements for information processing in connection with credit card data, which are regularly checked by auditors.

PHI - Protected Health Information

Protected Health Information is used as a term in the US legislature, particularly in connection with the Health Insurance Portability and Accountability Act (HIPAA). For US legislation, all information like health status, health care and billing of health care services, which can be attributed to an individual, is covered and collected by a so-called “Covered Entity” (IVF clinics, hospitals, health insurers and billing agencies).

In the EU there is no such special regulation, but this is also regulated by the EU Data Protection legislation and national legislation.

Conclusion

The new EU Data Protection Reform has been in effect since May 2016 and will become the legal regulation no later than 25.05.2018. It is therefore time to deal with the changes and also to consider the effects on test operations and test data management. Test Data Management Systems (TDMS) can effectively help you to comply with the rules and regulations. However, great care and consideration needs to be taken when it comes to the selection of a system, to ensure it meets the requirements of your company, and the new EU legal regulations.

References

1. Documents to be found under http://ec.europa.eu/justice/data-protection/reform/index_en.htm
2. Data Masking: What You Need to Know – Net 2000 Ltd. Whitepaper
3. Sergey Vinogradov, Alexander Pastsyak - Evaluation of Data Anonymization Tools - DBKDA 2012: The Fourth International Conference on Advances in Databases, Knowledge, and Data Applications (2012)
4. Purnima Khurana, Purnima Bindal - Test Data Management - International Journal of Computer Trends and Technology (IJCTT) – Volume 15 Number 4 – Sep 2014
5. Wikipedia: PII
6. Wikipedia: PCI
7. Wikipedia: Protected Health Information
8. Data Masking Best Practice – ORACLE Whitepaper (2013)
9. Judy Fainor - Test Data Extraction Methods for IBM InfoSphere Optim Test Data Management - Finding the right data subsetting strategy (2012)
10. Judy Fainor, Peter Hagelund - Provisioning Test Data with IBM InfoSphere Optim Test Data Management: Part 1 Choosing the right strategy for using production data in testing (2012)
11. Judy Fainor, Peter Hagelund - Provisioning Test Data with IBM InfoSphere Optim Test Data Management: Part 2 Privatizing sensitive data (2012)
12. Lang, Andreas - Anonymisierung/Pseudonymisierung von Daten für den Test - DACH Security 2012 · syssec (2012) pp-pp

About the Authors

Frank Pankalla is Senior Consultant at Sogeti Deutschland GmbH.

With his many years of experience in subjects such as test infrastructure management and test data management, he is incredibly successful at optimising test processes.

Rainer Popella is a Senior Consultant at Sogeti Deutschland GmbH.

He has many years of experience in software development, business analysis and test data management.

About Sogeti

Sogeti is a leading provider of technology and engineering services. Sogeti delivers solutions that enable digital transformation and offers cutting-edge expertise in Cloud, Cybersecurity, Digital Manufacturing, Digital Assurance & Testing, and emerging technologies. Sogeti combines agility and speed of implementation with strong technology supplier partnerships, world class methodologies and its global delivery model, Rightshore®. Sogeti brings together more than 25,000 professionals in 15 countries, based in over 100 locations in Europe, USA and India. Sogeti is a wholly-owned subsidiary of Capgemini SE, listed on the Paris Stock Exchange.

Rightshore® is a trademark belonging to Capgemini.

Learn more about us at www.sogeti.com

Contact Sogeti

24, rue du Gouverneur
Général Eboué,
92136 Issy-les-Moulineaux
FRANCE
Tel: +33 (0)1 58 44 55 66

Get Social: