Complying with the EU General Data Protection Regulation

The Implications for Test Data Management

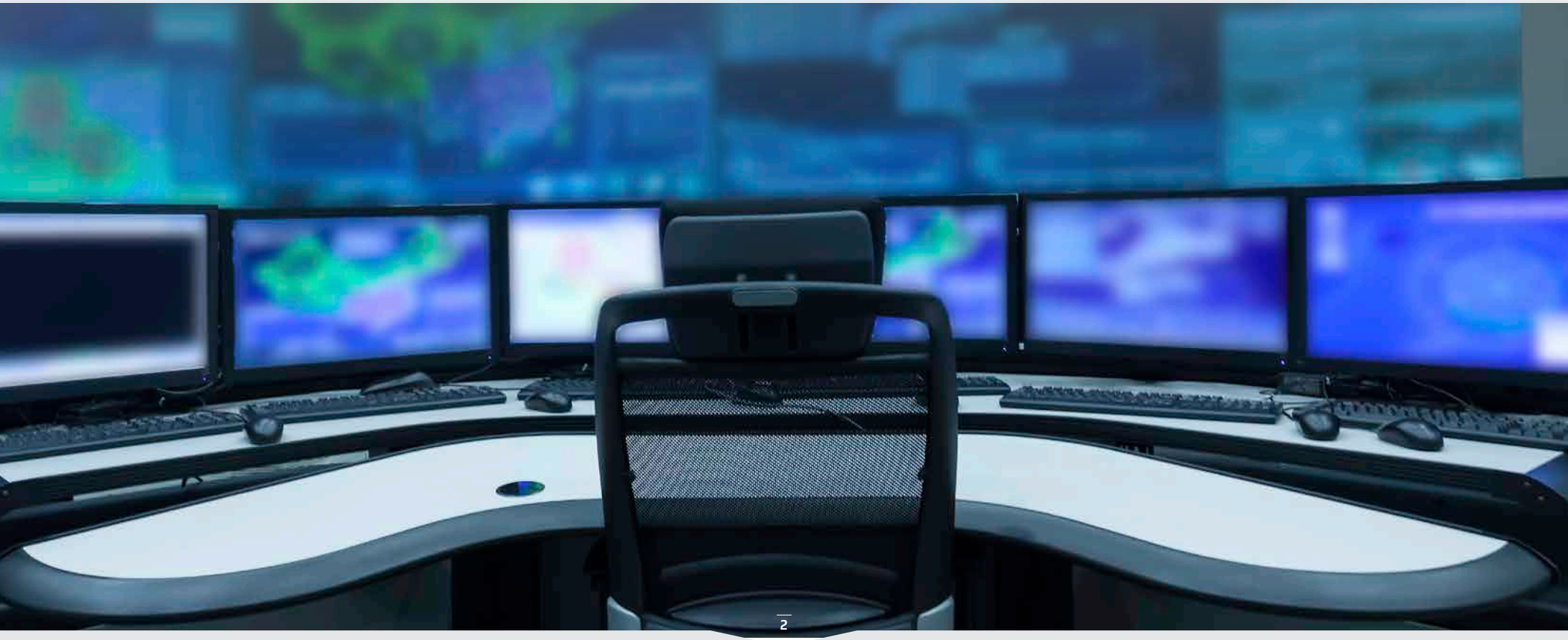
ca technologies

Section 1:

Know Everything About Your Data and Protect It

The GDPR is set to have wide-ranging implications for the type of data which can be used in non-production environments. Organizations will need to understand exactly what data they have and who's using it, and must be able to restrict its use to tasks for which consent has been given.

One way to avoid exposing personal data to test environments is to not provision it in the first place, even in a masked form. Synthetic data generation offers a technique which could enable organizations to transition to fully virtualized, fictitious test environments.



Section 2:

Key Implications of the EU GDPR

First proposed in 2012, the EU General Data Protection Regulation (GDPR) aims to more clearly define and reform existing regulation to address the technological issues and advances that have emerged since the 1995 Data Protection Directive. The GDPR seeks to make legislation more homogenous across the EU, and the extended scope of the legislation means that organizations worldwide will have no excuses for being unprepared when it comes into effect.

The GDPR was finalized on April 14, 2016. Though there is a two-year implementation period, organizations should not be complacent. Considering the time it will take to work towards compliance and the harsh penalties proposed, organizations need to address their data management processes and technology now.

Any organization that processes EU citizen's personal data must comply with the GDPR, which will introduce wide ranging reform for testing and development teams—especially for the still-prominent role played by production data in testing. In this paper, we draw from some of the key GDPR articles, highlighting their potential and practical implications for the technology and processes in place at numerous organizations.

Regulation scope: who it applies to and the consequences of non-compliance

Scope:

Any organization that collects, processes or controls EU citizen's data will be subject to the regulation, even if they are outside of the EU. Responsibility is further placed on the data controllers, who will be held jointly liable with the data processors. Organizations in both categories must be able to demonstrate compliance and show that they have technical and organizational measures in place to ensure it is enforced.

Harmonisation:

One goal of the GDPR is to make legislation more homogenous, stringent and enforceable across the EU. Mechanisms will be in place to ensure the coordinated enforcement of the regulation by the European Data Protection Authorities (DPA), and the GDPR will further introduce the concept of a lead DPA to provide businesses with a single point of contact. This so-called one-stop-shop mechanism is intended for use in cross-border or transnational cases, and aims to ensure the fast and consistent application of the regulation.

Process innovations:

The GDPR introduces a move toward privacy by design, meaning that organizations will have to build safeguards into processes, such as testing and development, from beginning to end. Privacy Impact Assessments (PIAs) will further need to be performed in specific cases that are judged to be high risk, and organizations will have to consult the supervisory authority if mitigating steps are not taken.

Enforcement:

Maximum fines are set depending on the nature of the infringement, and range from two percent of annual revenues, or €10 million, to four percent, or €20 million. When you consider the reputational damage and the increased weight that consumers place on their privacy, non-compliance becomes a very costly risk.

The international dimension:

The GDPR permits the transfer of data outside the EU, but under strict rules and regulation. When transferring data to a third country, that country must be judged by the EU to provide adequate data protection. If this is not the case, organizations must use other legal mechanisms to provide the required level of protection. This includes, for instance, Binding Corporate Rules (BCRs) or standard contractual clauses.

The where, who and what of sensitive data

Consent has long been one of the cornerstones of data protection directives and the GDPR is not new in this respect. However, in the context of the regulation, some organizations might look back on their policy and the relationship between consent and data usage.

The need for consent:

The GDPR effectively rules out the possibility of opt-out consent, something organizations have previously deployed. Instead, the level of consent required for the use of data strikes a middle ground between unambiguous consent and the stronger legal definition of explicit consent. This means that consent must be constituted by an affirmative action, while the terms of consent must be set out in an intelligible and accessible form and must be clearly distinguishable from other matters. The definition of consent must include all relevant information, such as the nature of the data that will be processed, the purposes of the processing, the identity of the controller and the identity of any other recipients of the data. Silence or inactivity will not constitute consent and individuals must also be informed of their right to withdraw consent (see below).

Data minimization and purpose limitation:

In addition to expiring or withdrawn consent, the GDPR affirms previous legislation regarding data minimization and purpose limitation. Consent has to be specific to the processing operations and the controller cannot request open-ended or blanket consent to cover future processing. This means organizations can only collect as much data as is required to fulfill the reasons for which consent was given, and must ensure that it is kept only for as long as needed and is used only by those who need it. Responsibility is placed on data controllers to specify the legitimate interests for which they are using data, whether statutory or contractual. Importantly, a service cannot be made conditional on consent unless fulfilling the service requires processing data.

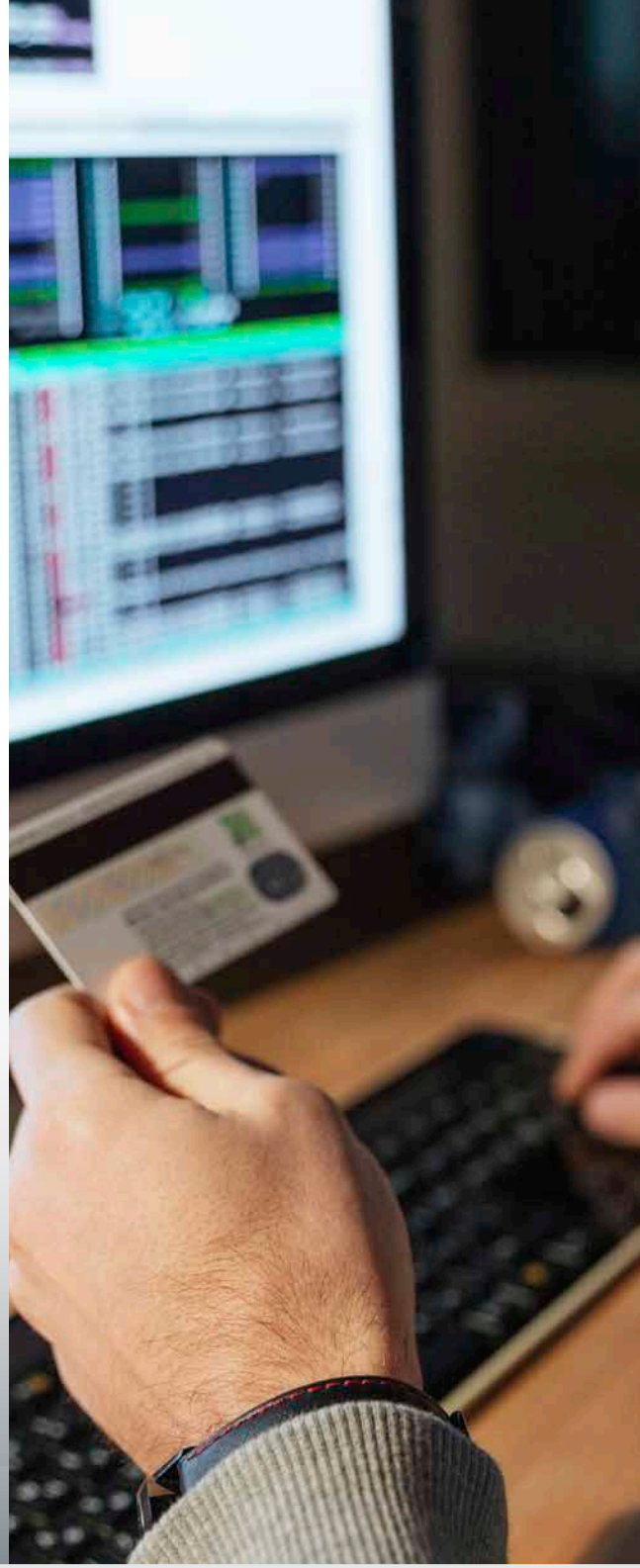
In other words, data cannot be collected and provisioned indefinitely, and organizations will need to know exactly where personal data is, when the data was collected, who's using it and for what purpose. For organizations that store data inconsistently, for example in uncontrolled spreadsheets and across environments, it will be extremely difficult to guarantee that there are no instances where data is being used for illegal purposes, or that it has been kept too long.

Data portability:

For organizations that do not understand their complex data that's spread across legacy and test environments, the much-debated data portability requirement exacerbates the issue of knowing where sensitive data resides. EU citizens will be entitled to request a copy of their personal data, in a form usable by them and transmissible to another system. If organizations do not know where this data is or cannot make sense of that data, they are in danger of non-compliance.

The right to erasure:

Also known as the Right to be Forgotten, this provision further demands that organizations know exactly where an individual's data is across their systems, so that it can be deleted upon request. The regulation also stipulates that it must be as easy to withdraw consent as it is to give it and, once withdrawn, the data should no longer be used for processing. Given how 46 percent of 500 global IT professionals said that they have received customer requests to remove data in the last 12 months, and yet 41 percent admitted that they do not have definite processes, technology or documentation to remove the data,[1] the rights to erasure and portability are likely to present a headache unless organizations take measures to address this.



Section 3:

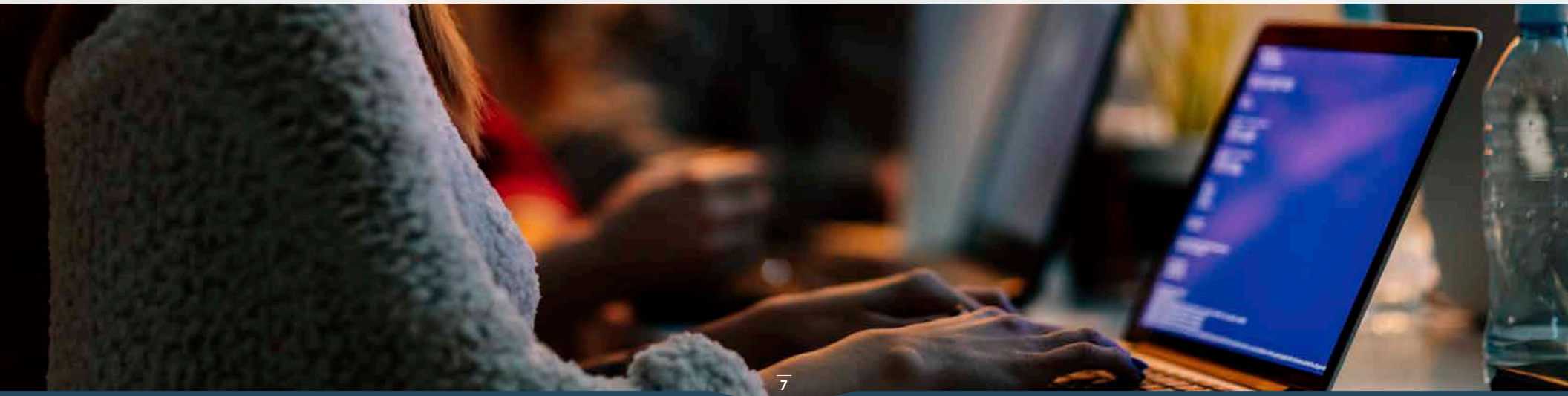
Maintain Compliance and Improve Testing Quality with CA Test Data Manager

Understand Your Data and Where Sensitive Data Exists

When testing and developing software, data can end up spread across test and development as well as complex environments. Testers might copy data to their environment for a given use, but organizations must know how long the data is used for, and that it's used with consent and for a legitimate purpose.

Data profiling from CA Test Data Manager (TDM) can help with this key point of compliance by identifying exactly where sensitive data is stored enterprise-wide, and by using statistical analysis to find personal data stored across multiple file formats and applications. Using a cubed view to create an accurate picture of data, TDM identifies sensitive information reflected in related systems, components or applications.

Custom, mathematically based filters mean that data can be sieved through on a granular level to identify every instance of information relating to an individual. This data can include credit card numbers, email addresses, home addresses and the like, helping organizations fulfill the right to data portability. The data discovery offered by TDM is fully auditable, so that organizations can demonstrate compliance.



Data masking and its drawbacks

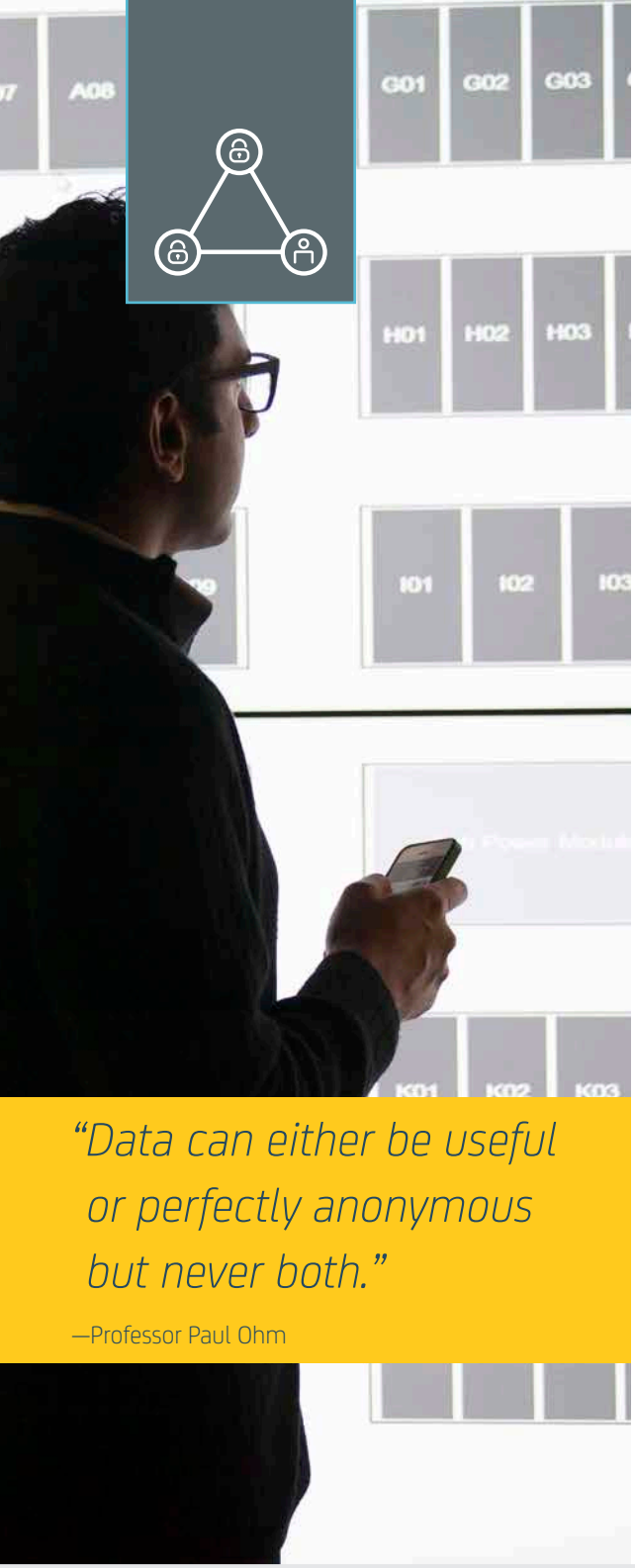
Once personal information has been found across production databases, it can be masked to make it compliant for use in non-production environments. TDM offers multiple, native masking engines that are capable of masking millions of rows of sensitive information in minutes. Personal data is replaced with realistic but fictitious values, while maintaining the referential integrity needed for testing across each system. This means that testers and developers don't need to use sensitive content.

However, masking data cannot completely mitigate the risk of non-compliance in non-production environments. One reason for this is that masking data to the extent that it can be considered "pseudoanonymized" under the GDPR is highly complex, and historically, organizations have failed to implement it successfully. For data to be considered pseudoanonymized, it must not be possible to identify a data subject without the use of additional information, which is stored separately from the masked data.[3]

In practice, this will mean that a lot of data needs to be masked. Case in point: 87 percent of Americans can be identified by three unique identifiers (gender, date of birth and zip code),[2] and the GDPR itself extends the definition of personal information to include factors such as genetic, mental, economic, cultural or social identity. All of these identifiers will need to be scrubbed and these may extend across multiple databases, to prevent individuals from being identified by cross-referencing masked test databases.

All of this information will have to be removed, while retaining the referential integrity needed for testing. This is highly complicated and organizations often compromise, leaving some original content in the masked data. In this instance, however, they are open to the risk of non-compliance and masking no longer appears to be the quick and easy way to become compliant.

87 percent of Americans can be identified by three unique identifiers (gender, date of birth and zip code)[2]



"Data can either be useful or perfectly anonymous but never both."

—Professor Paul Ohm

Even if organizations can successfully remove this content, the masked data will bear at least some resemblance to the original. This is because the complex relationships within and across systems must be retained for testing. As Professor Paul Ohm notes, "Data can either be useful or perfectly anonymous but never both."[4] If data is suitable for testing, then it might be possible to reverse engineer the masked data using either information external to the data or indirect identifiers. For example, this information might be who a bank's largest customer is, or knowing when someone made a transaction of a certain amount.

Synthetic data generation

If data has been properly profiled and modeled, it is just as easy with TDM to synthetically generate new data from scratch, as it is to mask existing data. Fortunately, this is also the way to provision test data which contains absolutely no resemblance to the original personal information.

TDM works directly with relational database management systems (RDBMs), mainframe platforms or API layers to create realistic personal data as quickly as processing power will allow. Data is generated based on a multidimensional model of existing data, or based on a model of all possible tests taken from CA Agile Requirements Designer, creating the data needed for 100-percent test coverage. A comprehensive list of combinable SQL functions, system and default variables and seed tables are provided, so that referentially intact data can be tailored to specific test cases and fed into multiple systems at once. In other words, no environment is too complex; if data can be masked, it can also be generated. If more data is needed for testing, organizations can use bulking scripts to quickly produce millions of rows of rich data and avoid the need to recourse to production data.

A hybrid approach

Of course, it's not going to be possible to replace all the data in non-production environments in one go. Even with high-performance tools and processes, masking a database or generating data from scratch requires time. This presents a challenge for testing and development: How can one database be masked if the same data is found in a different, interdependent part of the system? In some instances, there might be hundreds or thousands of interdependent systems with data stored in different formats across them.

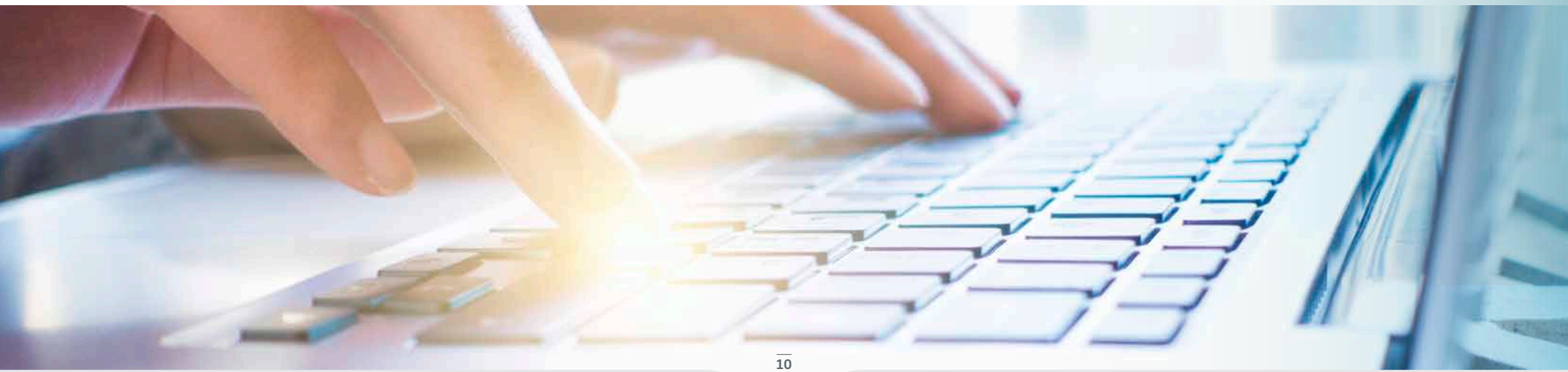
CA advocates a hybrid approach where the implementation period is used to move towards a combination of masked and synthetic data, and ultimately, entirely synthetic data. Development systems should be treated like production systems, while user activity should be simulated and pushed in at various points. This way, fictitious, synthetic data that is like production data can be pushed in.

Synthetic data generation could be used in a number of ways in this transitionary context. For example, database records could be copied and a separate version of them created, while automation frameworks could be used to pump fictitious data through the front end-under test. Another approach might be to simulate message queues for the majority of systems, which today use files or messages to communicate.

Limit data access and use to authorized individuals

As already discussed, the GPDR will require that organizations only use data for the explicit reasons that it was given, can only keep it for as long as it's needed and can only use it for a legitimate purpose. Enterprises can't keep or use data indefinitely, nor can it be used by an indefinite number of individuals.

TDM centralizes data requests under the remit of a central security team and stores data as reusable assets in a central test data warehouse



where testers can request it using a self-service, web-based portal. Access to data is granted on a project-by-project and user-by-user basis, while each project in TDM is set up with a set of data pools and tasks, and individuals are given permission to perform certain tasks. Only authorized individuals can access personal data and permission is set up on a highly granular basis, not just via a role-based approach.

TDM further supports the ability to restrict data usage to named individuals. Access to the tool requires user, group or role security authorization, which can be defined internally to the tool or using Lightweight Directory Access Protocol (LDAP). The ability to apply masking operations, for example at run time, depends on the credentials and permissions used when setting up a connection profile, which is required to access a target data source. Permissions must be given for a task to be performed and the same rules can be applied to multiple databases using separate connection profiles.

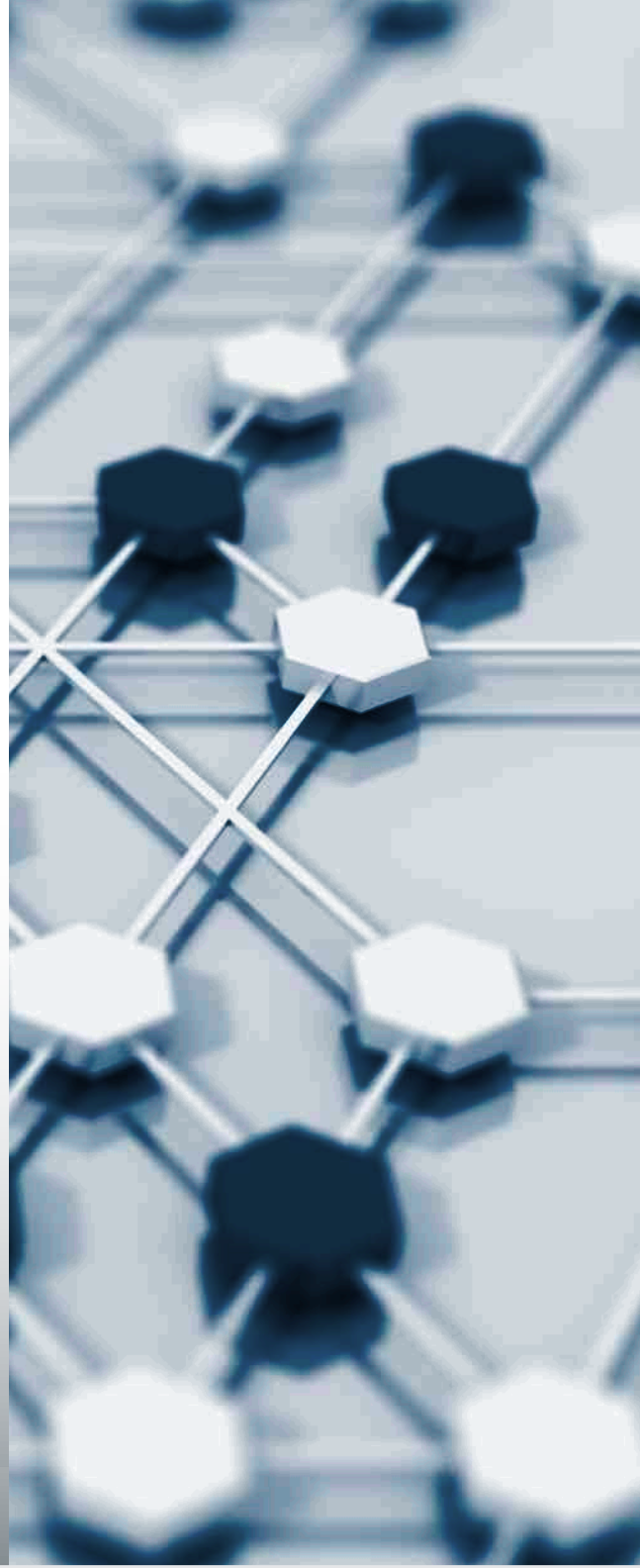
Previous masking and data-generation routines can be stored in the warehouse, along with data models and collected metadata. This maximizes rework for increased efficiency and limits access to authorized individuals only.

About the Author

Huw joined CA in 2015 as Vice President of Application Delivery, when specialist testing vendor Grid-Tools was acquired into the CA DevOps portfolio. During his 30-year career, Huw has gained a deep understanding of the challenges faced by modern organizations, and, with a thorough understanding of testing, how to solve them.

Huw has helped launch numerous innovative products that have recast the testing model. In 1988, he set up data archiving specialists BitbyBit, and was soon joined by long-term partner, Paul Blundell. After BitbyBit was acquired, Huw and Paul co-founded data migration and application conversion firm Move2Open.

In 2004, they set up Grid-Tools Ltd, and Huw quickly began redefining how large organizations approach testing. He helped oversee the development Datamaker (now CA Test Data Manager), pioneering a data-centric approach to testing, and later played a visionary role in the design and development of Agile Designer (now CA Agile Requirements Designer).



See how synthetic test data creation works.
Contact CA Sales to see a demo of CA TDM:
https://www.ca.com/us/contact/sales.register.html

CA Technologies (NASDAQ: CA) creates software that fuels transformation for companies and enables them to seize the opportunities of the application economy. Software is at the heart of every business, in every industry. From planning to development to management and security, CA is working with companies worldwide to change the way we live, transact and communicate – across mobile, private and public cloud, distributed and mainframe environments. Learn more at **ca.com**.

1 Phil Muncaster, "Firms Already Swamped by Right to be Forgotten Requests," January 2016

2 Latanya Sweeney, "Simple Demographics Often Identify People Uniquely," Carnegie Mellon University, 2000

3 Gabe Maldoff, Top 10 operational impacts of the GDPR: Part 8 - Pseudonymization, 2016. Cited from https://iapp.org/news/a/top-10-operational-impacts-of-the-gdpr-part-8-pseudonymization/ on 05/27/2016

4 Paul Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization," 2009

CS200-198117_0516

ca technologies