IBM Software

Back to basics: Fundamentals of test data management

How to build and deliver high-quality applications fast using realistic test data

IBM

Business moves fast—which means that software development teams need to move even faster. The emergence of new software development models, such as the agile development process, has given organizations powerful tools for responding to events quickly and has helped organizations evolve through collaboration between self-organizing, cross-functional teams.

To make the most of agile processes, organizations need effective and efficient testing strategies—complete with processes for governing test data. However, many development, testing and quality assurance (QA) teams struggle to create and maintain the required test data. IT departments often lack confidence about test data preparation and data usage within the testing discipline, and it may not be clear how to use and administer data efficiently.

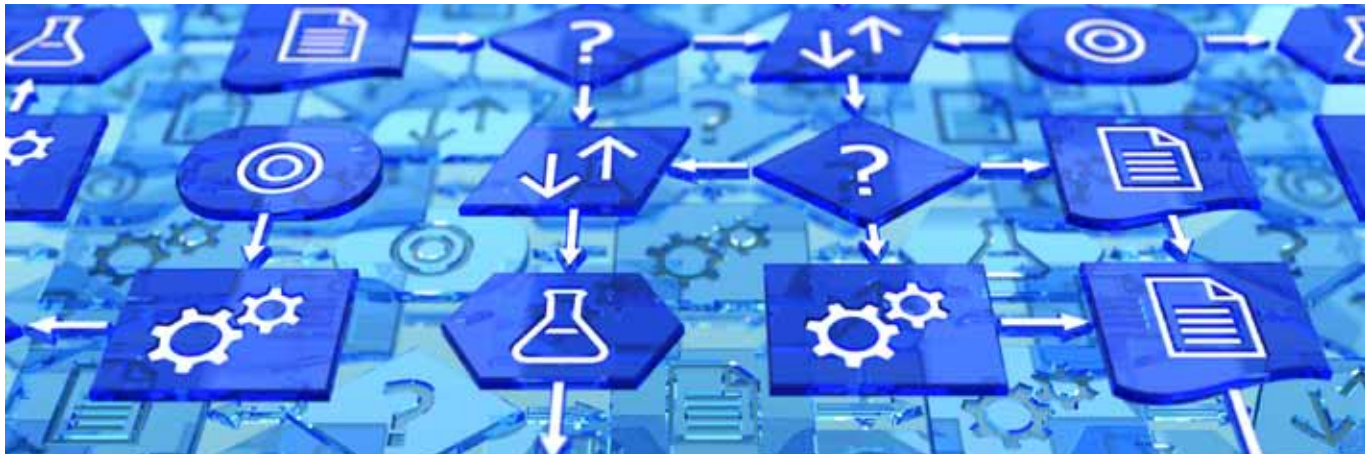
That's where test data management comes in.

What is test data management?

Simply stated, test data management is the process of creating realistic test data for non-production purposes such as development, testing, training or QA.

Research shows that projects cancelled due to poor data quality are 15 percent more costly than successful projects of the same size and type.[1] A better test data management strategy not only ensures greater development and testing efficiencies, but helps organizations identify and correct defects early in the development process, when they are cheapest and easiest to fix.

Typically, test data management involves two major activities: test data preparation and test data usage.



Test data preparation involves manufacturing data by copying or subsetting data from production or by developing test data generation scripts and provisioning them for multiple testing environments.

Referential integrity, data quality and data relationships must be retained during the preparation stage. The skills required to complete these tasks typically lie with DBAs, since they are the ones with knowledge of the underlying data model.

Typical approaches to test data preparation can include cloning production databases, subsetting data from production databases or writing scripts to synthetically create test data. Subsetting is the recommended method, but each has advantages and drawbacks.

METHODS	PROS	CONS
Cloning production databases	Relatively simple to implement	• Expensive in terms of hardware, license and support costs • Time-consuming: Increases the time required to run test cases due to large data volumes • Not agile: Developers, testers and QA staff can't refresh the test data • Inefficient: Developers and testers can't create targeted test data sets for specific test cases or validate data after test runs • Not collaborative between DBA and testing teams • Not scalable across multiple data sources or applications • Laborious: Production systems are typically large • Risky: Nonproduction environments might be compromised or misused (developers, testers and QA staff need realistic data to do their jobs—but they do not have a valid business reason to access sensitive data such as corporate secrets, revenue projections or customer information)
Generating synthetic test data	Safe	• Resource-intensive: Requires a huge commitment from highly skilled DBAs with deep knowledge of the underlying database schema, as well as knowledge of implicit relationships that might not be formally detailed in the schema • Tedious: DBAs must intentionally include errors and set boundary conditions within the synthetic data set to ensure a robust testing process, which adds time to the test data creation process • Challenging: Despite the time and effort put forth by the DBA to generate synthetic test data, testers find it challenging to work with because synthetic test data doesn't always reflect the integrity of the original data set or retain the proper context • Time-consuming: Process is slower and can be error-prone
Subsetting production databases	Less expensive compared to cloning or generating synthetic test data	• Skill-intensive: Without an automated solution, requires highly skilled resources to ensure referential integrity and protect sensitive data

Test data usage shifts focus to the tester or developer, who may not be database-savvy. This may create inefficiencies because the tester or developer absolutely requires proper test data—and if this test data is not available, the tester must go back to a DBA for help. The tester understands "test conditions" and tries to map those to accurate, physically available data in the test environment. The tester's mission is to ensure safe passage of the required tests, not to create high-quality, referentially intact test data.

Because DBAs and the application delivery team (developers, testers and QA personnel) have different skill sets and job roles, it is critical that everyone in the testing process works closely together. A strategic test data management strategy can help.

Most applications rely on relational database technology, which can create challenges for testing teams. The application data model may contain dozens, hundreds or even thousands of tables—and just as many interrelationships. What's more, data model complexity is not limited to large-scale systems: even a database of less than a dozen tables may contain relationships that make navigating the data model difficult.

Many organizations store data in a variety of relational databases. In addition, data may be stored in hierarchical or non-relational formats, such as IBM® Virtual Storage Access Method (VSAM) files and IBM IMS™ databases. All database management systems have different methods for handling data, which further complicates test data preparation.

From a test data usage perspective, it is not uncommon to require test data from multiple related databases—including both relational and non-relational data sources. In addition, each phase of the testing process, from unit testing through system integration and acceptance testing, has unique requirements and varying levels of complexity. Any problems that are discovered must be resolved, and the test data must be refreshed before testing can continue. And after a test is executed, IT organizations need a way to verify the results.

How can you improve both test data preparation and usage? First, you'll need to develop a strong test data management strategy.

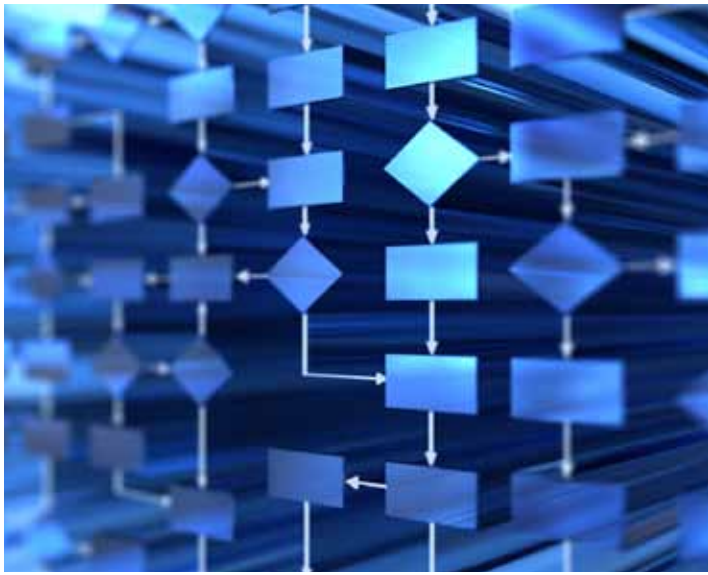
Five tenets of a good test data management strategy

When implementing a test data management approach, five best practices help streamline test data preparation and usage:

1. **Start by discovering and understanding test data.** Data is scattered across systems and resides in different formats. In addition, different rules may be applied to data depending on its type and location. Organizations should identify their test data requirements based on the test cases—which means they must capture the end-to-end business process and the associated data for testing. This could involve a single application or multiple applications. For example, a business may have a CRM system, an inventory management application and a financial application that are all related and require test data.
2. **Subset production data from multiple data sources.** Subsetting is designed to ensure realistic, referentially intact test data from across a distributed data landscape without added costs or administrative burden. In addition, the best subsetting approaches include metadata in the subset to accommodate data model changes quickly and accurately. In this manner, subsetting creates realistic test databases small enough to support rapid test runs but large enough to accurately reflect the variety of production data. Part of an automated subsetting process involves creating test data to force error and boundary conditions. This includes inserting rows and editing database tables, along with multi-level undo capabilities.
3. **Mask or de-identify sensitive test data.** Masking helps secure sensitive corporate, client and employee information and supports compliance with government and industry regulations. Capabilities for de-identifying confidential data must ensure a realistic look and feel and should consistently mask complete business objects, such as customer orders, across test systems.

4. **Refresh test data.** During the testing process, test data often diverges from the baseline, resulting in a less-than-optimal test environment—but refreshing test data can improve testing efficiencies. Refreshing test data helps to streamline the testing process and maintain a consistent, manageable test environment, which improves predictability and repeatability of testing efforts.

5. **Automate test data result comparisons.** The ability to identify data anomalies and inconsistencies during testing is essential to the overall quality of the application. The only way to truly achieve this goal is to deploy an automated capability for comparing the baseline test data against results from successive test runs—and speed and accuracy are essential. Automating these comparisons saves time and helps identify problems that might otherwise go undetected.



The bottom line: Managing test data nets real business value

Production data doesn't stand still—and neither should test data. Organizations need test data management solutions that are designed to accommodate changing test requirements.

Support a complete test data management strategy with IBM InfoSphere Optim solutions

The IBM InfoSphere® Optim™ Test Data Management solution offers comprehensive test data management capabilities for creating rightsized, fictionalized test databases that accurately reflect end-to-end business processes. InfoSphere Optim software scales to meet development and testing requirements across multiple applications, databases, operating systems and hardware platforms. It also helps facilitate modern software delivery models—including agile development—by making test data continuously accessible to testers and developers so they can quickly meet test requirements (see Figure 1).

The InfoSphere Optim Test Data Management solution helps improve application quality and delivery efficiency by:

- Reducing costs by intelligently creating and subsetting realistic test data from complete business objects
- Reducing risk by masking sensitive information
- Speeding delivery of test data through refresh capabilities

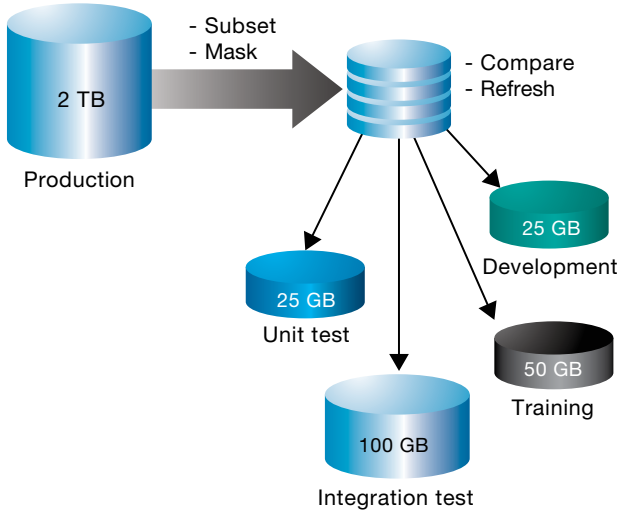


Figure 1: InfoSphere Optim software enables users to create referentially intact, rightsized and secure test databases.

IBM takes a unique test data management approach by automatically discovering referential integrity for undocumented or poorly documented applications and defining the data required for test cases. The InfoSphere Optim Test Data Management solution can capture database relationships and discover the complex associations that are typically hidden from view.

InfoSphere Optim Test Data Management also enables IT departments to subset referentially intact data accurately so they can create realistic test databases—no matter how many tables or relationships are involved. It can de-identify sensitive data by examining data values from multiple sources to determine the complex rules and transformations that can hide sensitive content, as well as transform test data to meet specific test case requirements. The solution also helps ensure that masked data is contextually appropriate for the data it replaced so as not to impede testing.

"Since implementing Optim, we have reduced the time allocated for creating and managing our testing environments, and the content of the test databases is more realistic and reliable. Now, we can generalize a consistent testing methodology across our organization."

– **Michèle Davain**
DBA Manager, Technical Department
Cetelem

Using InfoSphere Optim, companies can easily refresh and maintain test environments for developers and testers. Users can browse and edit test data to force error conditions and resolve problems. By enabling developers to compare the test data before and after testing, the software helps validate expected test results and identify hidden errors (see Figure 2).

InfoSphere Optim software supports custom and packaged ERP applications in heterogeneous environments. It can also automate the creation of rightsized masked test data in private cloud and virtual infrastructures.

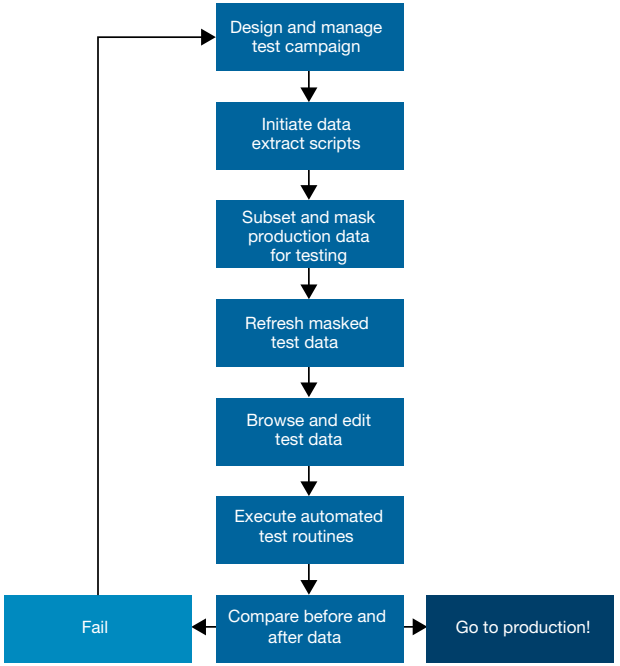


Figure 2: InfoSphere Optim Test Data Management software enhances testing discipline by automating several key processes.

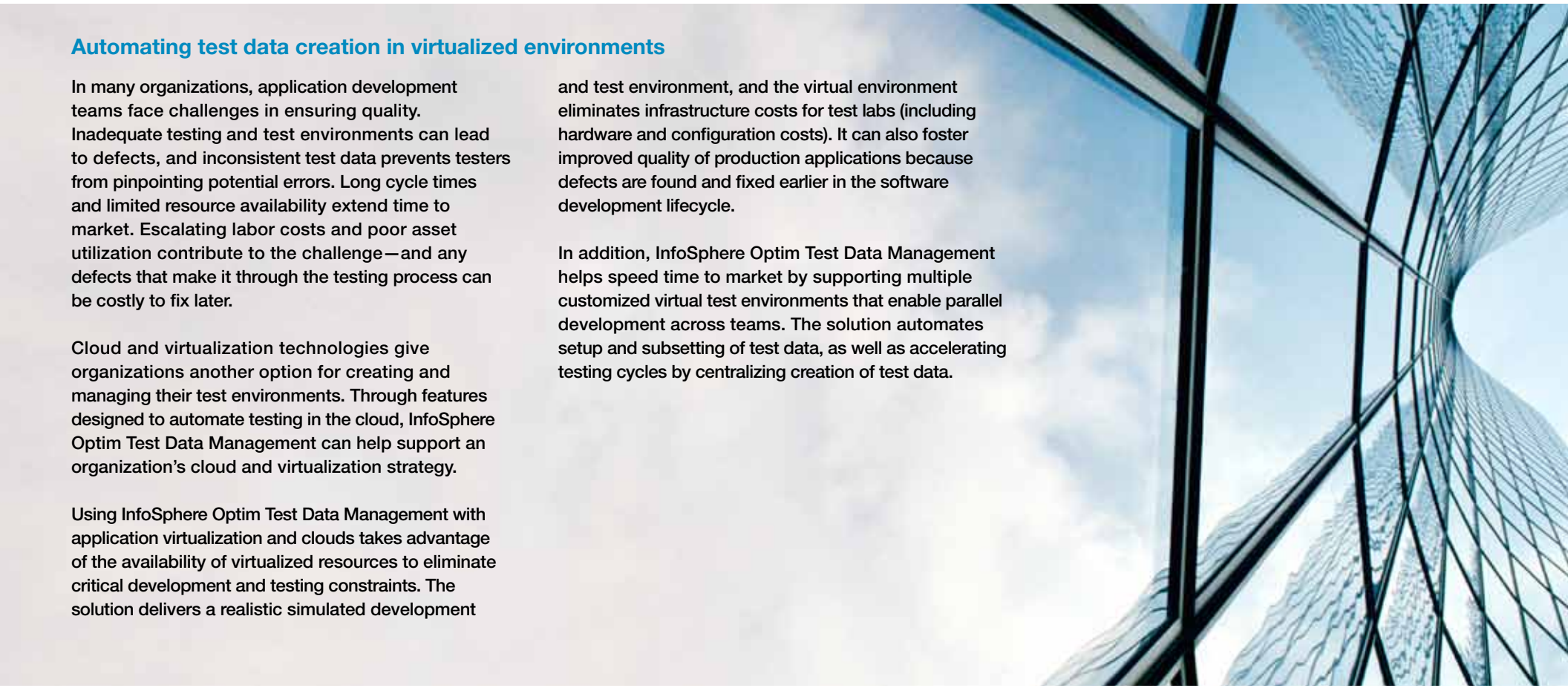
Automating test data creation in virtualized environments

In many organizations, application development teams face challenges in ensuring quality. Inadequate testing and test environments can lead to defects, and inconsistent test data prevents testers from pinpointing potential errors. Long cycle times and limited resource availability extend time to market. Escalating labor costs and poor asset utilization contribute to the challenge—and any defects that make it through the testing process can be costly to fix later.

Cloud and virtualization technologies give organizations another option for creating and managing their test environments. Through features designed to automate testing in the cloud, InfoSphere Optim Test Data Management can help support an organization's cloud and virtualization strategy.

Using InfoSphere Optim Test Data Management with application virtualization and clouds takes advantage of the availability of virtualized resources to eliminate critical development and testing constraints. The solution delivers a realistic simulated development and test environment, and the virtual environment eliminates infrastructure costs for test labs (including hardware and configuration costs). It can also foster improved quality of production applications because defects are found and fixed earlier in the software development lifecycle.

In addition, InfoSphere Optim Test Data Management helps speed time to market by supporting multiple customized virtual test environments that enable parallel development across teams. The solution automates setup and subsetting of test data, as well as accelerating testing cycles by centralizing creation of test data.



Closing the gap between DBAs and testers

The right test data management solution accelerates time to value for business-critical applications and builds relationships and efficiencies across the organization. IBM InfoSphere Optim Test Data Management closes the gap between DBAs and application developers by providing all teams with accurate, appropriately masked and protected data for their work. Developers can confirm that new application functionalities perform as expected. QA staff can validate that the application performs as intended based on the test cases, and that integrations work properly. And business leaders can be more confident that competitive functionality will be delivered on time with less risk.

Resources

For more information on the IBM InfoSphere Optim Test Data Management solution and other data management offerings, please explore these resources:

- Visit **ibm.com**/optim
- Download the white paper: "Enterprise Strategies to Improve Application Testing"
- Check out the IBM InfoSphere Optim Smarter Testing webcast

IBM Corporation
Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
June 2012

[1] "Software quality in 2010: A survey of the state of the art," Software Productivity Research LLC, November 2, 2010. http://www.sqgne.org/presentations/2010-11/Jones-Nov-2010.pdf

Please Recycle

IMM14092-USEN-00