Transforming Test Data Management for Increased Business Value

By centralizing, automating and securing test data, IT organizations can create the highest quality software at the lowest possible cost.

Executive Summary

Whatever your industry, speed and agility are key to success. Companies that are quick to roll out new applications or services for employees, customers or partners are in a prime position to reap competitive advantage.

Ensuring application quality at a competitive price requires quality test data, which all too often is a costly process bottleneck. TDMaxim is a customizable test data solution framework that combines our test data management (TDM) service and Informatica’s application information lifecycle management software to deliver faster, higher-quality test data across database platforms, test environments and lines of business.

TDMaxim reduces costs and helps meet regulatory requirements by creating smaller test subsets of live data, using a variety of methods to mask sensitive fields. For one retail client, TDMaxim reduced test data provisioning time by more than 95% and test cycle time by 33%. For another client, the use of TDMaxim, along with a global delivery model, cut overall test data management costs by 40%. TDMaxim has also been shown to reduce test data storage needs by more than 50%.

This white paper describes the challenges of creating and managing test data, six best practices for meeting those challenges and how TDMaxim implements those best practices to deliver the highest quality software at the lowest possible cost.

Test Data Challenges

Proper application software testing requires the identification of data that the application must access, process and display. Such data includes customer names, credit card or account numbers and transaction details, such as SKU numbers or transaction numbers. This data ensures the proper functioning of not only the core application code but also related functions, such as database calls, the business rules that govern the sharing and processing of information and the user interface.

All too often, generating, tracking and refreshing test data becomes a bottleneck. Research indicates that testers spend over 30% of their time resolving data-related defects. These include setting up the data for multiple environments, downtime due to corrupt data and filling in missing data. In addition, a lack of coordinated processes in the extraction, creation and storage of data adds unnecessary delays and costs, which include the expense of storing unneeded test data.
The absence of proper test data causes nearly one-third of the incidents we see in QA/non-production environments and is a major reason why nearly two-thirds of business applications reach production without being properly tested. The resulting application failures cause significant amounts of downtime, with an average price tag of $100,000 per hour for mission-critical applications, according to industry estimates. Such software failures also force nearly half of deployed applications, in our experience, to be rolled back into development, at significant cost to user productivity and added effort.

One reason test data is such a challenge is that different applications require very different and specific test data. For example:

- Testing retail applications may require ID numbers for customers in loyalty programs and transaction identifiers to automatically validate returns. This is necessary for test cases, such as ensuring that when a customer returns a product that earned her a discount, the amount of the discount is deducted from the refund.
- For insurers, testing billing or claims processing systems may require older data for policies that have been in force for decades or more.
- Banking applications, such as those governing transactions at ATMs, must be tested to ensure they can accurately access a customer’s current balance to ensure withdrawals do not exceed current balances.

In addition, large organizations must coordinate the creation and refreshing of test data with application release schedules across multiple business units, as well as for applications that rely on distributed systems, multiple downstream and upstream services and various platforms. They must also avoid overly long test cycles that can result in the use of old and inaccurate test data. Such poor test data can lower the quality of testing and the applications themselves, while increasing time to market.

While it’s essential that all test teams have enough data to mimic production environments, IT organizations must avoid overly large test data archives that are expensive to store and manage and (even worse) could lead to the use of the wrong data for a given test. Extremely large test data sets, and a lack of referential integrity across applications (ensuring consistent relationships between tables), also make it harder to retrieve proper data for testing.

Finally, and critically, organizations must scrub or depersonalize different types of data, depending on the legal, regulatory and business requirements of each application. These include standards within specific industries, such as PHI (Protected Health Information,) PII (Personally Identifiable Information) and PCI (Payment Card Industry). In addition, various countries and regions, such as the European Union, have their own strict standards for protecting personal information.

For all these reasons, software testing and test data management is not only a back-office function but also a business-critical enabler of competitiveness.

To deliver business value through improved test data management, we developed TDMaxim in collaboration with Informatica. TDMaxim advances the state-of-the-art of TDM solutions, combining Informatica’s industry-leading TDM products with our best practices to deliver superior data management capabilities and solutions for global organizations across industries.

TDMaxim: Delivering Quality Through Testing

TDMaxim leverages both our test data management practice and Informatica’s TDM software. A leader in Gartner’s Magic Quadrants for static data masking and data integration, TDM provides more than 350 clients worldwide with best practices in test data management, including defining, discovering, subsetting and masking sensitive test data.

By centralizing TDM efforts and enabling on-demand reusability of test data, TDMaxim can reduce test data provisioning times by as much as 95% and overall test cycle time by 33%. This enables the fastest possible deployment of high-quality applications at the lowest possible cost, while ensuring that sensitive data is masked to meet regulatory, legal and business needs.

Best Practice 1: Centralize TDM Services

A centralized TDM office should serve as a center of excellence (CoE), recommending tools, technologies and best practices that lines of business and geographies can then customize to meet their needs.
The CoE can provide, for example, staging servers for data extraction and scrubbing. It can train and make available domain experts, DBAs, system architects, tool experts, data leads and others to help local test units improve their work. It can share expertise and knowledge across business units and increase utilization of test data team members by cross-training them in different applications and data stores.

While the CoE provides a single framework for test data management, each business unit must be free to adapt it to its own environment, business processes and testing needs. To facilitate this, the CoE team should have a strong understanding of the data landscape, including upstream and downstream applications, as well as the current project/release lifecycle so it can help tailor solutions for local needs. Defining an end-to-end process is very important to ensure that the correct data, as well as the correct volume of data, is provisioned through the use of real data, synthetic data creation or a hybrid approach using both (see Figure 1).

To ensure coordination and communication, the CoE should also provide a process owner (a test data manager) who is responsible for coordinating test data release across projects and environments. This manager should report to a TDM manager and ensure that test data requests are collected early enough to ensure their proper coordination. This position should also ensure maximum reuse of test data during the upfront creation of the test environment and throughout the full testing process. The CoE should also define the tools strategy for organizations or business units across the enterprise to ensure they enhance ease of adoption, improve process efficiency and deliver the appropriate data. Among the recommended tools is a centralized ticketing system to track requests and measure the quality of data provided.

Such centralized TDM expertise will provide a single, end-to-end view of test data management requirements across the organization, eliminating confusion and wasteful, redundant efforts. The workflows and best practices included within TDMaxim incorporate our experience implementing such TDM services.

Best Practice 2: Identify Automation Opportunities
Just as with any process that involves repetitive tasks and large amounts of data, automating the creation of test data can significantly reduce costs while reducing manual error. One best practice we have learned is to embed the notion of automation within the entire test data process, continually evaluating whether, and how, each step might be automated.
In our experience, the areas that benefit most from automation include:

- **Data mining to identify common test data elements**, usable across multiple applications or test scenarios to minimize redundant test data.
- **Aging, masking and archiving of test data** through tools-based automation.
- **Prioritization and allocation** of test data requests.
- **Generating and allocation of test data** through metrics, such as the time required to generate test data.
- **Utilizing APIs** to generate test data.
- **Generating, distributing and reporting the results** of user satisfaction surveys.
- **Creating and implementing business rules** for the use of test data across environments, while ensuring proper masking for privacy and security.
- **Building an automation suite for master data preparation** through a GUI-based automation tool or through the execution of SQL queries.
- **Automating the masking, profiling, versioning and aging of data** with a commercial off-the-shelf tool, applying data creation rules specific to each business process.

**Best Practice 3: Create a Gold Copy and Manage Versions of the Test Data**

Studies show that more than 20% of test efforts go toward preparing test data, with testers spending hours mining data, discussing requirements with subject matter experts and tweaking new data for every code release. All this time could be better spent on other areas, such as test design.

Organizations can greatly reduce the cost and effort of creating test data by creating and maintaining a gold copy: a centralized, version-controlled repository of reusable data. This single, enterprise-wide repository helps ensure that consistent, accurate test data is always available by creating subsets of the data to meet the needs of specific applications, test scenarios, platforms or lines of business. Functional testing, for example, might require a great variety (but not a massive amount) of data to test the widest variety of business rules or usage scenarios. Performance testing, on the other hand, might require a far higher volume of data (but not such a wide variety) in order to test the application under the highest possible load.

For each new or modified test scenario, the test staff identifies its unique data requirements and determines whether that data is available in the gold copy. If it is not, the staff extracts the needed data from production or other test environments, or creates or simulates the test data using TDM. This new data is added first to the test environment for testing, and then to the gold copy to ensure it is available, as needed, for other test scenarios.

By replacing multiple, and possibly conflicting, local test databases, this gold copy also reduces the cost of storing excess and unneeded data. It boosts agility by reducing the cost of, and delays in, providing data for each project from the production data. In our experience, it can also reduce, by as much as 60%, the analysis required to sift millions of records to identify and extract the subset required for a specific test effort.

After each test, result accuracy should be assessed to determine data quality, with the gold copy being updated or revised, as needed. As many as five or six versions of the gold copy should be archived for use in testing different versions of an application, or for regression testing to assess whether a new version of the code introduced new errors.

**TDMaxim provides a centralized framework for automating the identification of test data, the extraction of test data from other sources, the creation of subsets for various business needs and their distribution to local test environments.**

TDMaxim provides a centralized framework for automating the identification of test data, the extraction of test data from other sources, the creation of subsets for various business needs and their distribution to local test environments. To further enhance automation, TDMaxim helps implement this centralized yet flexible model with a variety of techniques to mask sensitive data and remove personal details. It also speeds testing with accelerators for PeopleSoft, SAP NetWeaver, Siebel eBusiness and the Oracle e-Business suite.

TDMaxim also is equipped with a framework that uses queries to mine project-specific data from cross-application test environments to meet complex demands for test data. The environment includes techniques to synthetically create data, such as for testing in service-oriented architectures or graphical user interfaces.

TDMaxim also provides a centralized framework for automating the identification of test data, the extraction of test data from other sources, the creation of subsets for various business needs and their distribution to local test environments.
Quick Take

TDMaxim: Proven Results

As brick-and-mortar retailers face increasing online competition, having a reliable, fast and user-friendly Web site is often a strategic necessity. However, the data needed to test such a site often comes from multiple legacy databases, organized around complex hierarchies and containing stale data, such as discontinued products and old promotional campaigns.

We recently completed a proof-of-concept creating test data using production data from one of our customers, a major global high-tech retailer. The test database consisted of records for about 80,000 SKUs (stock-keeping units) drawn from Microsoft SQL Server, native AS/400 and IBM DB2 databases on the AS/400. The product database consisted of multiple hierarchies (such as cables, printer cables and network cables) and multiple dimensions for each product, such as price, promotional campaigns and coupons.

This complexity made it difficult to create proper subsets of test data for various test scenarios. In addition, the production database contained out-of-date information about products and special offers, such as coupons that had to be made current to ensure proper testing.

Using Informatica ILM TDM, we defined a policy and generated, subsetted and updated more than four million records of test data in less than seven days. This eliminated work previously performed by the retailer, allowing testers to refresh their test data in minutes, with the click of a mouse.

TDMaxim supports the management of a gold copy, or copies, of test data, with comprehensive support for enterprise data sources, including databases, mainframe and legacy file types, as well as data from packaged applications, such as SAP, Oracle, Siebel or other cloud sources.

It also features vertical-market-specific accelerators that help organizations more quickly determine which test data to produce, mask and manage, including for applications that must meet standards designed to protect, for example, medical or credit card information.

Best Practice 4: Mask Business-Sensitive Data

In today’s regulatory and legal environment, data security in non-production areas such as test environments is an important but often overlooked requirement.

The Ponemon Institute, for example, estimates that nearly three-quarters of all database administrators can view any data in the enterprise, increasing the risk of a breach. It also says half of the enterprises it surveyed report that data has been compromised or stolen by a malicious insider, such as a privileged user.

Such risks require the masking or obfuscation of the large amounts of data that may be needed to properly test applications. Some of this masking is required to meet regulatory requirements such as HIPAA (the Health Insurance Portability and Accountability Act) or PCI.

Knowing what data to mask and how to mask it requires guidance from information security experts, as well as risk analysis to identify the cost/benefit of the masking effort, and to ensure that the referential integrity of the data across applications is maintained after masking.

Masking can be accomplished using several methods, including fixed rules, random processing and customized rules based on business needs. Fixed rules include masking specific fields, such as credit card or Social Security numbers, for example. Custom rules apply to business processes unique to one application and its data. In either case, masking can be performed with various techniques, including shuffling characters, selective masking of only some data and making random changes to characters.

TDMaxim provides domain experts whose experience in various industries helps prioritize data masking efforts, and who can suggest which alternatives (such as scrubbing or obfuscation) are most cost-effective. It also provides industry-specific frameworks, defining which data elements must be masked to help organizations prioritize their efforts.
Best Practice 5: Implement Proper Test Data Governance and Metrics

Like any other costly resource essential to meeting business challenges, the use of test data should be carefully managed to comply with internal audit requirements, minimize costs and maximize results. This includes measuring the cost and effort required to create and manage it, user satisfaction levels and the success (or lack thereof) of data masking for audit, security or risk teams.

Another critical governance area is ensuring that test data is properly generated and subsetted to minimize storage needs, while ensuring that proper test data is available when needed for various geographies or business units.

As with any other governance program, use of the proper metrics is essential. Test data governance metrics we have found useful include:

- **Size of the TDM team** to evaluate resource usage.
- **The volume of test data provisioning requests**, which helps identify data-intensive applications and proactively collect and efficiently schedule the fulfillment of those requests.
- **The number of repeated data requests**, to identify opportunities for automation.
- **The turnaround time for data requests**, to help assess user service and satisfaction levels and to track improvements over time.
- **The number of test data refreshes per year**, to identify areas for automation.
- **The number of bad data requests**, in order to track and reduce unproductive TDM activities.

In tracking these metrics, we recommend the following benchmarks:

- **Team size**: A centralized CoE should have between one and 10 members, with two to three dedicated staff for every 100 team members in local lines of business.
- **Data provisioning**: Test data should be refreshed two to three times in each monthly or quarterly release cycle.
- **Data requests**: The number of data requests can range from 100 to 1,000, depending on the size of the project. Such requests should be tracked in a centralized ticketing system.
- **Size of test data**: Storage required for test data can range up to 500G bytes per application, minimized through archiving.

- **Request response time**: The ideal is 24 hours, although a full data refresh may range from two days to two weeks based on complexity.
- **Number of test data refreshes**: This depends on releases, but frequency should not exceed two to three refreshes per build.
- **Number of invalid requests**: Fewer than 5% of requests should be for data that already exists, or for undefined business processes. Levels over 10% result in test data that does not match business requirements and reduces the quality of testing.
- **Mix of on-site/globally sourced staff**: We recommend a 40-60 split in on-site vs. off-site staff.
- **Automation**: Our best practices suggest 60% to 70% of all TDM activities should be automated. Through its best-practice-based workflow for communication and coordination, TDMaxim enables the real-time provisioning of optimal test data, as needed, to meet ever-changing business needs.

Best Practice 6: Ensure Use of Appropriate Tools

Application testing can be very complex, and each test project may have unique requirements. The proper tools are essential to cost-effectively perform each step, from capturing data requirements from end users, to tracking the fulfillment of each service request, to subsetting the data, masking sensitive data and storing the gold copy or copies.

Such tools should provide a single integrated platform to ensure test data processes are well documented and automated with centralized workflow tools. The platform should be accountable for ensuring that the handling of data requests, the prioritization of data delivery across projects/releases, the tracking of efforts and status, trend analysis and continuous improvement are communicated across the organization and performed efficiently (see Figure 2, next page).

TDMaxim helps achieve this with a best-practices-based toolset that reduces the cost and increases the quality of software testing. It provides tools to capture the data requirements, record the servicing of data requests through the use of existing data, create synthetic data and capture metrics, documents and project information.

Because of the unique nature of each test process, we recommend clients follow a framework to ensure they have chosen the most appropriate
tools for each application. Among the issues to consider are cost and ease of use, as well as how well the tool can meet project-specific requirements, such as:

- The amount and types of test data needed.
- Support for legacy systems and databases.
- Version control.
- Support for federated databases.
- Reporting.
- Dependencies on other tools.

Better Test Data for Higher Software Quality

As organizations share more digital information with employees, customers and business partners, the quality of the applications that deliver that data becomes more important. With its unique combination of best-practice workflows and disciplines, along with the unparalleled data handling, masking and workflow capabilities of Informatica TDM, TDMaxim delivers higher application quality, in less time, with reduced costs for everything from testing coordination to data storage.

Footnotes

About Cognizant

Cognizant (NASDAQ: CTSH) is a leading provider of information technology, consulting, and business process outsourcing services, dedicated to helping the world's leading companies build stronger businesses. Headquartered in Teaneck, New Jersey (U.S.), Cognizant combines a passion for client satisfaction, technology innovation, deep industry and business process expertise, and a global, collaborative workforce that embodies the future of work. With over 50 delivery centers worldwide and approximately 156,700 employees as of December 31, 2012, Cognizant is a member of the NASDAQ-100, the S&P 500, the Forbes Global 2000, and the Fortune 500 and is ranked among the top performing and fastest growing companies in the world. Visit us online at www.cognizant.com or follow us on Twitter: Cognizant.

© Copyright 2013, Cognizant. All rights reserved. No part of this document may be reproduced, stored in a retrieval system, transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the express written permission from Cognizant. The information contained herein is subject to change without notice. All other trademarks mentioned herein are the property of their respective owners.