Test Data Management

Purnima Khurana¹, Purnima Bindal²

¹Assistant Professor, Department Of Computer Science, PGDAV College, University Of Delhi, Delhi, India

Abstract — Management is required in each and every phase of life and so does in any software project. A software project goes through many stages: requirement analysis, designing, development, testing and maintenance. In order to successfully complete each stage, a certain form of management is required. Testing stage also has a lot of tasks to be performed, so management is also required here. Test Data plays a very important role in Testing. It is the information used while executing test cases, or resulting from the execution of test cases. So, its management that is Test Data Management is a big concern and to make any project a success Test Data Management is necessary. Test Data Management is the process of identifying, acquiring, conditioning, populating, and maintaining test data during test planning, test preparation, and test execution and much of the work is automated in order to reduce cost, time required, manual work and errors.

Keywords — Testing, Management, Data, TDM.

I. INTRODUCTION

Data management is the development, execution and supervision of plans, policies, programs and practices that control, protect, deliver and enhance the value of data and information assets[1]. When these management activities are performed for test data, it is termed as Test Data Management. Test data is information used while executing test cases, or resulting from the execution of test cases. Test case data is an asset. Quality of the test data is directly related to the quality of the end product. So, Test Data Management becomes a crucial part to make any software product a success. It is not same as Test Case Management which is defined as “How to create, manage and dispose off test cases”. But it implies creating or obtaining, storing, tracking, or otherwise guiding or handling test data throughout its life cycle[3].

More formally Test Data Management can be defined as:

1) The governance process of identifying, acquiring, conditioning, populating, and maintaining test data during test planning, test preparation, and test execution.

2) The process that defines standards, tasks, ownership, roles, and responsibilities to ensure the accuracy, completeness, and integrity of test data.

Test Data Management activities are required for all phases and types of testing: Unit, System, Integration/UAT, Performance Testing, Production Verification, Regression Testing and Test Automation.

II. WHY MANAGE TEST DATA?

If we are implementing or upgrading large systems or packaged applications, we have undoubtedly encountered testing challenges. To deliver business-critical applications on time and with high quality, we need a reliable and repeatable testing process to accelerate time-to-market and reduce costs. These systems and applications also tend to be highly data-driven and require large amounts of data to test effectively. The testing challenges cross multiple aspects from test planning to data preparation to data protection. Data-driven testing challenges can cause spiralling costs and pose a significant risk to the quality of the resulting application. It is simply not cost-effective to clone large production databases for testing environments—the storage footprint quickly gets out of control. So, what we need is Test Data Management.

Following are the reasons, why we need to manage test data:

1) Testers waste more time preparing test data rather than testing the application: Preparing data for testing is extremely labor intensive and can consume more than 50 percent of our overall testing time and resources.

2) Testers depend a lot on Business Analysts to provide the required test data.

3) Testing deadlines slipped more than once due to delay in test data refresh.

4) Lot of false defects due to data related issues: Lot of defects get raised, then they get rejected mentioning invalid data as the reason. A clear alarm that the data related false defects are growing and needs to be eliminated.

5) There is no reuse of test data: Old test data need to be retired as new test conditions necessitate the need for new data. Also periodic refreshes need to be made.

6) There is a lot of dependency on the upstream systems for the test data to be created: A lot of delay can happen waiting for another system to provide the test data.

7) Need of centrally controlled master test data: Multiple applications can refer the same test data simultaneously, so simultaneous access needs to be controlled in order to maintain data consistency.

8) As the project size increases, it is even more difficult to manage the test data [2], [3].

III. WHY TEST DATA MANAGEMENT IS CRITICAL TO A PROJECT’S SUCCESS?

1) Test data determines the quality of testing: No matter how good testing processes are, if the test data used is not
right or of adequate quality, then the entire product's quality will be affected.

2) Test data should be highly secure: It is absolutely mandatory that test data doesn't contain data from production without being masked. If the data is not secure enough, then there is every chance that a data breach might happen, which can cause a big loss to the organization.

3) Test data needs to be as close to real time as possible: Not only that test data needs to be of quality, it should be as close to real time data / production data as possible to determine that whether the system is working fine or nor when being run in real execution environment.

4) Lowers test data creation time: TDM results in reduce overall test execution time by lowering test data creation times and thus help in timely delivery of product.

5) Testers can focus on testing rather than test data creation: The main focus of trying to automate the test data management process is to allow the testers to focus on the actual testing than worrying about how the data is created and the technicalities surrounding it. This allows the team to remain focused on the job at hand so that it can be done more effectively.

6) Increases efficiency of the process by reducing data related defects: Due to the accuracy of the test data, data related defects will reduce enormously, thereby increasing the efficiency of the process [2].

IV. COMMON CHALLENGES OF PREPARING AND MANAGING TEST DATA

1) Realistic data is hard to collect and sort: With today’s business applications, data is typically spread across multiple databases, including both internal and external ones. Customer records might be stored in one place and purchase orders in another. This can make data extraction a time-consuming process. Many test teams know that accessing and extracting test data can consume a large portion of the time spent in testing efforts. In addition, testing organisations typically have limited access to the production systems or production backup systems where real customer information resides, and they have limited skills for dealing with the range of databases and schemas. This means they have to depend on other stakeholders – such as subject matter experts and database administrators – to provide the data they need. It all adds up to a lot of lost time in the testing process. Accessibility constraints can make test data synchronization an almost impossible task, prompting testers to skip or reduce test scenarios.

2) Data Volatility: Systems under test often run against systems that are in production or shared by other teams. This causes the data to be volatile (changing randomly and often), making validation of results against these volatile systems difficult or impossible.

3) Data Sensitivity: Use of production data is limited by the need to obscure sensitive information such as account, social security numbers, credit card numbers etc. and other personal and business information are an attractive target to hackers, data thieves and others. When production data is used for QA tests, sensitive data can become a soft target for unscrupulous people. They can present a significant compliance challenge, particularly when offshore or outsourcing service providers are used.

4) Impact on Production Systems: Some test scenarios create transactions that impact data in production systems. For instance, if a test places an order, the order then needs to be traced and deleted. This incurs time and costs on both testing and operational staff and can disrupt business operations if not carefully managed.

5) Storage maintenance costs too much: As the number of business applications rises and the amount of data they handle explodes, storage maintenance costs are becoming a significant drain on IT budgets. Given the high cost of storage maintenance, QA team needs to reduce the amount of data it stores and manages. It is not cost effective to clone and maintain an entire production database when we actually need just a relevant subset of the data for testing. It is hard to maintain the referential integrity of data when we pull it out of a production environment. For example, if a customer name changes in one place, we need to make sure it gets changed in other places where it appears.

6) Data is not readily available for reuse: After all the work and expense of extracting data from a production database, the data used by QA specialists is often put to work in just one testing task or phase. QA organisation can get more out of the time and effort it puts into collecting and preparing data if that data is usable by all the testing tools. And since data extraction can be required with every testing iteration, the use of time-consuming manual extraction methods can drive up costs and jeopardise schedules. In addition, human error is common when IT staff attempt to handle large quantities of data manually without a structured automation solution.

7) Volume of Data needed: Deciding amount of data to be needed for testing is also a challenge. For some test scenarios, only a small amount of data can solve the purpose but for other a huge amount of data may be required.

8) Handling of data available in different formats: If there are multiple sources of data available and that too in various formats, then handling such data is also a big challenge. Relevant data for testing has to be collected from these sources and need to be mapped in one format so that it can be used for testing purpose.

9) Dynamically creating data: Sometimes data to be used in testing has to be created dynamically at the time of execution. For ex: data resulting from transactions and from some calculated results. So, creating and managing such data offers a big challenge.

10) Storing Test results in addition to test case data: Due to relationship between test results and test case data, it becomes necessary to store old data. As it also consumes lot of storage space, it’s a challenging task to maintain historical data and relationships among data as well [3],[4],[7].

V. TEST DATA MANAGEMENT LIFE CYCLE

Life Cycle is the various stages that a product/service/artifact goes through before attaining its end of life. So a Test Data Cycle explains the various stages through which the test data goes through in order to reach its
end of life or alternatively start a recurring life cycle. So similar to a test life cycle or a software development life cycle, Test Data goes through the following phases. They are:

A. Requirement Gathering and Analysis
   In this phase, the test data requirements pertaining to the test requirements are gathered. They are categorized into various heads:
   1) Pain Areas: We have to decide the crucial modules of a project in which rigorous testing is need to be done.
   2) Data Sources: There is a need for deciding what data sources are to be considered for preparing Test Data.
   3) Data Security/Masking: Data Masking is the process of masking the sensitive fields from the complete data set. The whole objective of data masking is to ensure that no sensitive data is leaked into non-production regions like testing regions.
   4) Data Volume requirements: In the very step, it should be decided that how much data is required for testing purpose.
   5) Data Archival requirements: Data Archival typically comprises of:
      - Size Management: There must be an efficient mechanism for the database size management. Over time a database size grows and we need to actively manage it.
      - Archival of older data: Older data can be archived to some low disk space occupying area and can be later retrieved whenever needed
   6) Test Data Refresh considerations: Test Data Refresh is the process of loading/refreshing the Test Database with the latest data from the Production database or any other data source.
   7) Gold Copy considerations: Gold Copy is the baseline version of the data that can be used for future releases. For example, if we are trying to load our test database from the production database for the first time. In this case, we can save the copy as a baseline from which future test data refreshes can be made.

This phase is typically carried out in the form of a TDM (Test Data Management) assessment or Test Data Assessment.

B. Planning and Design
   As the name indicates, based on the requirement analysis an appropriate solution is designed to solve the various pain areas in the Test Data. After looking at the problem scale and the feasible solution, a suitable test data process is suggested and we would need to choose between an In-House solution or a Commercial Product or a combination of both. Also in this phase, an effort estimate is done for the entire project. And a test data plan/strategy is also developed that will propose a direction that the project will take and what approaches will be followed to solve the test data problems.

C. Test Data Creation
   In this phase, based on the Test Data Strategy, the solution is developed and test data is created through various techniques depending on the project test data requirements. It can be a combination of manual and automated techniques. Automation techniques might include In-house tools or commercial products. It can be either a refresh from production or generation from scratch or a hybrid approach. The output at the end of this phase will be the actual test data required for the project.

D. Test Data Validation
   In this phase, the created test data is validated against the business requirements. This can be done by Business Analysts or using automated tools if the volumes are very high.

E. Test Data Maintenance
   This is similar to a test maintenance phase, where there might be requests for changes in the test data according to the changes in the tests. Hence again the entire life cycle is followed for maintenance of the test data. This might include creation of Gold Copy for future use, Archives for size management, updating of Gold Copy, Restoration of older data for testing, etc. [2].

Life Cycle Of Test Data Management is pictorially shown below in Fig. 1:

![Test Data Management Lifecycle]

**Fig. 1. Lifecycle of Test Data Management**

### VI. TEST DATA MANAGEMENT TOOLS

Many Test Data Management tools have been developed by different software companies. Some companies have developed a standalone tool for Test Data Management and some have developed a module of it, which is part of a larger Testing tool. A testing tool does a lot of work in addition to Test Data Management.
Some of the Test Data Management tools are listed below in Table 1.

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Tool Name</th>
<th>Vendor</th>
<th>Special Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Oracle Test Data Management Pack</td>
<td>ORACLE</td>
<td>Automatically discover data relationships</td>
</tr>
<tr>
<td>2</td>
<td>SOLIX Enterprise Data Management Suite (EDMS)</td>
<td>SOLIX</td>
<td>Test databases built using automated parallel copy processes to reduce creation time from a day long procedure to just hours</td>
</tr>
<tr>
<td>3</td>
<td>LISA</td>
<td>ITKO</td>
<td>Capturing and manipulating virtualized datasets from messaging streams and data sources</td>
</tr>
<tr>
<td>4</td>
<td>IBM InfoSphere</td>
<td>IBM</td>
<td>Supports testing in cloud and big data systems</td>
</tr>
<tr>
<td>5</td>
<td>HP Test Data Management Software</td>
<td>HP</td>
<td>Supports Custom Data Masking and Shared Data Extraction Rules</td>
</tr>
<tr>
<td>6</td>
<td>TestBench</td>
<td>Original Software</td>
<td>User managed Data Roll Back Capability</td>
</tr>
</tbody>
</table>

These tools with their features are described as follows:

**A. ORACLE Test Data Management Tool**

Oracle Test Data Management Pack helps enterprises shrink storage costs by creating reduced size copies of production data for application development, training and testing while maintaining the referential integrity of the data set. Through data discovery and application modeling, Oracle Test Data Management Pack automatically enforces complex business rules of enterprise applications resulting in accurate subsets of production data.

*Key Features*
1) Automatically discover data relationships.
2) Reduce size of database while securing sensitive, regulated data.
3) Pre-execution analysis and estimation.
4) Low storage overhead execution using Oracle Data Pump.
5) Oracle Database Gateways to support heterogeneous data subsetting.

*Key Benefits*
1) Reduce application storage costs dramatically through data subsetting.
2) Increase developer productivity through right-sized production copies.
3) Sensitive data never leaves production with in-line Masking and Sub setting.
4) Eliminate labor-intensive error-prone manual process to create test systems by automating application discovery, data relationship modeling, dynamic subset rules and parameter-driven execution [5].

**B. SOLIX EDMS (Enterprise Test Data Management Suite)**

The Solix Enterprise Data Management Suite (EDMS) brings together all the tools for an effective Information Lifecycle (ILM) program. From a single web-based console, the platform offers industry-leading database archiving, test data management, data masking and application retirement solutions.

*Key Features of TDM module:*

1) Solix EDMS Test Data Management automates the creation and management of database subsets and clones based on organization-defined business rules.
2) Solix EDMS Test Data Management creates syntactically correct subsets of production data for the most realistic testing possible. Subsets provide accurate application testing results while saving time and reducing infrastructure costs based on storage allocation up to 70%.
3) Solix EDMS Test Data Management eliminates unnecessary security risk and significantly reduces infrastructure costs. Solix EDMS Test Data Management lets you create secure subsets of your production instances for test, development, patching, training and outsourcing. The result is improved security, reduced risk, and a reduction in storage requirements [6].

**C. LISA Solutions for Test Data Management**

ITKO's LISA can automate the creation and management of test data by capturing and manipulating virtualized datasets from messaging streams and data sources. These unique capabilities help eliminate test data conflicts between teams and drives lower costs and manual effort. LISA provides an automated approach to capturing, modelling, desensitizing and maintaining virtual test data across multiple services, integration layers and systems of record.

*Key Features of LISA*

1) **Virtual Dataset Creation**: LISA listens and captures data as it flows across all application tiers and technologies to generate a robust virtual data set with a high level of functional accuracy. Test data can also be imported from databases (JDBC, SQL), Excel spreadsheets, flat files, log files, or XML sources.
2) **Test Data Manipulation**: LISA makes it easy for testers to modify datasets and chain them together into scenarios. Data models can be created and edited within LISA in a point-and-click environment with rules applied for randomization,
sequencing or appending of additional data sources. Or, manipulate data in familiar tools of choice such as spreadsheets.

3) Test Data Masking: Sensitive live data can be obfuscated and automatically processed by LISA into “safe test data” that is valid for testing activities without violating regulatory or security policies.

4) Dynamic Data Stabilization: LISA can stabilize test data so that test scenarios are validated according to expected business rules. This includes maintaining the relevance of variables including changing dates, currency conversions, customer session information and sequential data entry steps.

5) Data Model Self-Healing: LISA can “self-heal” virtual test data sets by validating and recapturing relevant data from live sources or transaction streams on a scheduled or exception basis, increasing the long-term viability of virtual test data and minimizing the need to correct or update data manually.

LISA Benefits for Managing Test Data across Modern Applications

1) Reduced time and cost: Synchronization of test data across multiple dependent systems can be automated with LISA. Testers can now capture and build robust virtual data sets that automatically self-update for longer term use.

2) Eliminates constraints from the software lifecycle: Robust virtual datasets can be leveraged by multiple teams, enabling parallel development and testing activities without dependencies on live systems or conflicts over the use of shared test data.

3) Security and compliance: Automated masking of sensitive data allows testing to proceed without exposing data to unauthorized users or processes.

4) Reduced impact on live systems: LISA’s virtual datasets allow you to avoid accessing production systems to insert, update or delete test data. LISA can route test data requests to virtual data sets or live systems in conjunction with exiting Application Lifecycle Management (ALM), Test Management (TM), and SOA Registry/Repositories as appropriate[7].

D. IBM InfoSphere Optim

It offers comprehensive Test Data Management capabilities which support an iterative repeatable process of creating, maintaining, and protecting data in non production environments to accurately reflect end-to-end business process.

Key Features

1) InfoSphere Optim scales across applications, databases, operating systems and hardware platforms.

2) Secure your test environments, improve quality, accelerate release cycles and reduce costs with InfoSphere Optim.

3) InfoSphere Optim Test Data Management leverages Hadoop as a test data landing zone enabling clients to manage their test data with even more speed at a lower cost.

The InfoSphere Optim Test Data Management solutions can:

1) Reduce Cost: Automate creation of realistic "right sized" test data to reduce the size of test environments and automate comparison functionality to catch errors earlier

2) Reduce Risk: Mask sensitive information for compliance and protection

3) Speed Delivery: Refresh test data speeding testing and application delivery

InfoSphere Optim Test Data Management solutions also helps organizations embrace a new era of computing by supporting testing in the cloud and big data systems[8].

E. HP Test Data Management Software

HP Test Data Management software is a flexible, powerful solution designed to reduce delays and costs associated with data-driven testing. This is accomplished by accelerating test data preparation using automated data extraction and masking.

Key features and benefits of Test Data Management Software:

1) Reduces time required for extracting large amounts of distributed data needed for testing: Implements rapid, policy-guided extraction of relevant subsets of test data derived from multiple, large data sources and delivers them to a single destination while maintaining referential integrity.

2) Accelerates application testing: Streamlines the data preparation phase of the application lifecycle and provides a subset of realistic, but de-identified data for testing purposes.

3) Reduces storage requirements: Reduces the footprint and cost of data storage for databases used in testing environments.

4) Reduces costs through shared data extraction rules: Parameterizes the data extraction rules for ongoing reuse within different extraction activities.

5) Supports data privacy and security: Masks sensitive data with built-in functions for commonly used data types (for example, social security numbers, phone numbers etc.).

6) Flexibly meets unique data security needs: Supports custom data-masking functions for any data type, enabling compliance with corporate data security policies.

7) Efficiently handles different data output types: Generates the output data as either a flat file or a database, while maintaining the same structure and schema as the source database.
8) **Complements the HP suite of testing tools:** Accelerates application testing for both manual and automated tests [9].

**F. TestBench**

TestBench is a tool developed by Original Software for Test Data Management that uniquely addresses the validation of all database effects and the creation of cut down, representative test data to reduce test times and data footprints. Data confidentiality is addressed and the unique user managed data roll-back capability reduces environment downtime, improving testing productivity and accuracy.

TestBench for iSeries is designed for business and IT professionals who recognise the critical importance of thoroughly tested iSeries systems - but lack the automated assistance.

**Key features of TestBench iSeries:**

- Complete systems testing - screens, database and reports.
- Covers batch, interactive and service programs.
- Secure, centralized script and results storage - for easy access and knowledge sharing.
- Automated test data creation and management.
- Stress testing capabilities.
- Automated summary and detail reporting [10], [11].

**VII. CONCLUSIONS**

In this paper we have discussed about Test Data Management which is very crucial for a project’s success. We have discussed why there is a need to manage test data, why TDM is critical to a project’s success, challenges of TDM, lifecycle of TDM and various tools developed for TDM. The major reason behind managing our test data is to save our resources: time, money and labour. It also reduces the chances of error as much of the work is automated rather than doing manual. Test Data Management lifecycle follows the same steps as that of any software cycle. Various tools have also been developed in the last few years by various companies for managing test data. Some companies have developed a standalone tool for TDM while some have integrated a module for TDM in a larger testing module. A testing module does a lot of work in addition to TDM like test case management, data archiving etc. Whatever be the case, the main goals of Test Data Management module are to save resources and provide security by automating various tasks and running them in a controlled manner.

**REFERENCES**