Test Data Management (TDM): an overview of process, challenges, and solutions.

By Steve Anderson

Efficient management of data used for testing is essential to maximizing return on investment and supplementing the testing efforts for the highest levels of success and coverage. If the data used in testing does not promote ease of use and adaptation, poorly represents the sampled source, or consumes excessive resources for preparation and maintenance, a negative impact on the desired outcome quickly manifests and continues to degrade the quality of results. To balance in favor of positive results and improved returns, consider the process, potential challenges, and possible solutions involved in TDM.

The TDM Process
A Tester cannot simply claim “there are probably defects” in a system and never attempt to identify and report the defects. They must interact with the system and replicate potential defects that have been found. Similarly, a tester can’t provide adequate results if they do not have access to relevant systems and an appropriate sample of data the system utilizes. For data to return the most value, it must be managed using quality processes. The key phases involved in a TDM process are:

- Planning
- Analysis
- Design
- Build
- Maintenance

Table of the test data management process:

<table>
<thead>
<tr>
<th>Phase</th>
<th>Steps Involved</th>
</tr>
</thead>
</table>
| Planning    | 1. Assign Test Data Manager (TDM)  
2. Define data requirements and templates for data management  
3. Prepare documentation including list of tests and data landscape reference  
4. Establish a service level agreement  
5. Set up the test data management team  
6. Appropriate plans and papers signed off |
| Analysis    | 1. Initial set up and synch exercises involve data profiling for each individual data store assignment/recording of version numbers for existing data in all environments  
2. Collection/consolidation of data requirements  
3. Update project lists |
4. Analyze data requirements and latest distribution log
5. Assess for gaps and impact of data modification
6. Define data security, back up, storage, and access policy
7. Prepare reports

**Design**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Decide strategy for data preparation</td>
</tr>
<tr>
<td>2.</td>
<td>Identify regions needing data to be loaded/refreshed</td>
</tr>
<tr>
<td>3.</td>
<td>Identify appropriate methods</td>
</tr>
<tr>
<td>4.</td>
<td>Identify data sources and providers</td>
</tr>
<tr>
<td>5.</td>
<td>Identify tools</td>
</tr>
<tr>
<td>6.</td>
<td>Data Distribution plans</td>
</tr>
<tr>
<td>7.</td>
<td>Coordination/communication plan</td>
</tr>
<tr>
<td>8.</td>
<td>Test activities plan</td>
</tr>
<tr>
<td>9.</td>
<td>Document for data plan</td>
</tr>
</tbody>
</table>

**Build**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Execute plans</td>
</tr>
<tr>
<td>2.</td>
<td>Execute masking/de-identification where applicable</td>
</tr>
<tr>
<td>3.</td>
<td>Back up data</td>
</tr>
<tr>
<td>4.</td>
<td>Update logs</td>
</tr>
</tbody>
</table>

**Maintenance**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Support change requests, unplanned data needs, problems/incidents</td>
</tr>
<tr>
<td>2.</td>
<td>Prioritize requests where applicable</td>
</tr>
<tr>
<td>3.</td>
<td>Analyze requirements and consider if they can be met from existing/modified current data including data assigned to other projects</td>
</tr>
<tr>
<td>4.</td>
<td>Required data modification</td>
</tr>
<tr>
<td>5.</td>
<td>Back up new data</td>
</tr>
<tr>
<td>6.</td>
<td>Assign version markers and log with appropriate description</td>
</tr>
<tr>
<td>7.</td>
<td>Review status of ongoing projects</td>
</tr>
<tr>
<td>8.</td>
<td>Data profile exercises</td>
</tr>
<tr>
<td>9.</td>
<td>Assess/address gaps</td>
</tr>
<tr>
<td>10.</td>
<td>Refresh data where needed</td>
</tr>
<tr>
<td>11.</td>
<td>Schedule and communicate maintenance</td>
</tr>
<tr>
<td>12.</td>
<td>If necessary, redirect requests</td>
</tr>
<tr>
<td>13.</td>
<td>Documentation and reports</td>
</tr>
</tbody>
</table>

**Tools**

The use of quality tools promotes quality results in any line of work, and it is no different when it comes to TDM. Links with useful tools are provided below.

Challenges

There are many challenges that can complicate the TDM process such as sensitive data masking and resource consumption. An overlooked challenge can cause major setbacks. Several common topics for consideration have been listed below.

**Challenges of Test Data Management include:**

- Additional time for data set up/management instead of actual testing
- Additional administrative efforts in test data management
- Additional expense including personnel and hardware
- Inaccurate/difficult to access data negatively impacts testing
- Sensitivity of private information (credit cards, medical records, etc.)
- Storage required for test data
- Potential for data loss
- Use of real data versus fake data generated from scratch
- Data requests poorly communicated result in inadequate data returns
- Identification of data anomalies
- Test priority confliction
- Timely data reversions

**Data masking and de-identification**

Data masking and de-identification is essential to comply with privacy laws and standards. There are several approaches that may be taken to use realistic data without betraying the confidentiality of sensitive data:

- You could go through and remove all sensitive information, such as credit cards or social security numbers, but this may not always be the correct method to accurately cover test requirements.
- One method is to generate fake data from scratch that fits the appropriate format. This can be time consuming for personnel; however, an automated script can be used to quickly generate required data.
- If you need to return the data to its original format, in some circumstances, a reversible algorithm can be used to alter the data. However, if the algorithm is known or discovered this could potentially allow for the private data to be compromised.
- A numeric variance, such as +/- 10%, can be used to change information (finance, demographics, etc.) just enough to make it untrue but still valid enough for appropriate use.
- Data Encryption is a very extensive approach that may not be as effective as it appears if access rights are carelessly given out.
- Masking out with viewed values being changed, such as with XX or **, can allow systems to still use the data without making the data available for easy access.
Solutions

Once challenges are reviewed we need to consider solutions to help mitigate the impact of these challenges. Considerations for TDM improvement have been listed below:

Solutions to reduce challenge impact include:

- Ensure connectivity of relevant parties before data set up
- Testing environments and data requirements are well-defined
- Smaller data sets that accurately sample full data coverage
- Involved parties meet and confirm requirements are fully addressed
- Back up data and assign versions
- Log the versions with relevant details for quick reference and conversions
- Data partitions are assigned to entire teams/projects, not to individual members
- Maintain records of data distribution
- Unused data/partitions made available for other relevant projects
- Masking and de-identification of sensitive information
- Scope of project defines masking tools for complete and consistent masking with realistic representation
- Masking tools jointly decided by relevant parties
- Standard request and documentation templates
- Refresh test data as needed, including periodic updates with new extracts, to accurately cover customer data
- Subset of metadata to accommodate changes
- Regular scheduled maintenance
- Insert row and database editing changes with multilevel undo capabilities
- Cloud storage (may violate privacy protection)
- Outsourcing of processes to expert companies
- Networking with other professionals
- Automation can be used to expedite processes and lower resource cost, including:
  - Masking/De-identification of sensitive information
  - Comparisons between baseline and successive test runs

Summary

Efficient Test Data Management (TDM) improves quality of testing results. Improved results lead to an improved product and higher return on investment. A process with good understanding and meeting of requirements, coupled with quality solutions to relevant challenges, will help provide the efficiency desired in TDM. Once TDM is optimized, increases in productivity, results, and profitability should quickly manifest, allowing more resources and focus can be utilized on continuing quality products and services.