Building Effective Test Data Management In Distributed Environment

By: Saksham S. Sarode

Session No: TH 6
Agenda

1. Test Data Management – Concept And Constraints
2. TDM In Distributed Environment – Strategy and Approach
3. Real Life Scenario - Challenges And Solutions
4. Industry Standards – Tools And Frameworks
Why to manage test data, it's not productive?

My requirements are complex, interdependent and overlapping.

Where can I get the data best suited for my testing needs?

Do we really need tools for test data management?

Managing data for testing is consuming too much of my time & effort.

Is there any reusability for test data at all?

Why to manage test data, it's not productive?
Typical Challenges

Where to look for
- Multiple test data locations
- Distributed environment
- Poor quality and test data views

What to trust
- Lack of trusted Information
- Inaccurate, out-of-data, incomplete data
- Difficult to track information through organization

What is really important
- Huge amount of test data
- Not using customer analysis to tailor requirements
- Not leveraging metadata

What are the stakes
- Escalating cost
- Changing needs
- Impact on quality
Test Data Management

WHAT???

Refers to the need to manage test data used in non-production environments along with its constraints and challenges by using scientific methodology and systematic approach

WHY???

• Quality
• Cost
• Time
• Security
• ......
**Requirement Gathering**

### Activities

- Conduct survey in the organization
- Analyze status quo
- Define process and build standard templates
- Identify SPOC for different teams
- Set up a team
- High level scoping
- Create Data Landscape
- Establish SLAs

### Entry Criteria

- Mangmnet Initiative
- Project Kick Off Meeting

### Exit Criteria

- Signed-off Templates
- Signed-off Test Data
- Landscape document

### Work Items

- Questionnaire
- Templates
Planning

**Activities**

- Data profiling exercise
- Version control
- Data catalog
- Requirement consolidation
- Analyze requirements
- Gap analysis

**Entry Criteria**

- Templates are available
- Signed-off Test Data
- Landscape document
- Team is available

**Exit Criteria**

- Reviewed Data Profile
- Reviewed Data Catalog
- Reviewed Analysis document

**Work Items**

- Data Request form(s)
- Requirements Document
- Data Distribution Log
- Data Security
- Policy document
Design

### Activities

- Identify test region(s)
- Identify data sources and providers
- Identify tools for data extracting, masking, creating, loading etc
- Create Data distribution plan
- Create coordination and communication plan
- Create Data Strategy document or Data Plan

### Entry Criteria

- Signed-off test data requirements are available.
- Signed-off Analysis document is available.

### Exit Criteria

- Signed-off Data Strategy Document or Data Plan

### Work Items

- Data Requirements document
- Data Distribution Log
- Analysis document
- Data Strategy document or Data Plan
Execution

**Activities**

- Execute test data plan
- Take back-up of data from source region
- Carry out masking (optional)
- Load data dump to target region
- Create Data
- Communicate data readiness and request

**Entry Criteria**

- Signed-off test data plan is available
- Environment is available
- Identified Tools are available

**Exit Criteria**

- Reviewed updated data distribution log
- Reviewed updated Data catalog
- Reviewed updated Data Profile

**Work Items**

- Data Plan
- Data catalog
- Data Distribution Log
- Data Profile document
## Maintenance

### Activities

- Support CR / unplanned data needs/incidents
- Scheduled perfective maintenance:
  - Update Data requirements
  - Assign priority to the request
  - Communicate data readiness and request validation
  - Take back-up
- Review status of ongoing projects
- Schedule maintenance
- Communicate schedule
- Data profiling exercise
- Gap analysis

### Entry Criteria

- Simple data requests
- Unplanned data requirements
- Scheduled perfective maintenance

### Exit Criteria

- Updated data distribution Log
- Reviewed updated Data Profile Document

### Work Items

- Data Request form
- Data Requirements document
- Data Distribution Log
- Data Profile document
Real Life Challenges And Solutions
Challenges

Production data contains sensitive information which cannot be exposed to other users.

Best Practices

- Intelligent Masking
- Dynamic generation
- Credit Card number generation based on Luhn’s algorithm
- SSN Numbering scheme
- DSN logic for Email IDs
- Based on valid pre-defined seed values
- Standard scrubbing techniques
Challenges

Production database size is huge, cloning the production environments to development and test environments requires huge disk space and complicates the data management.

Best Practices

• Not to clone but populate only data that is required
• Sampling / Subset involves the extraction of data based on specified filter criteria
• Combination of sub-setting techniques
Challenges

To create test data for testing special conditions governed by timelines

Best Practices

• Shift the dates as per the aging needs defined
• Shift the dates according to the business days
• Define customized calendars based on the business rules
• Use the customized calendar to shift the dates appropriately
• Intelligent aging based on the business day calculations of the application
Time to set up test database/test environment is very high

**Challenges**

**Best Practices**

- Identify and analyze the different data sources
- Extract / Subset the required data from production
- Condition extracted data based on data security requirements
- Enhance the data as per the needs
- Validate the data
- Store the processed test data as a Base-line so that it could be reused for further releases/iterations
Challenges

Test data gets exhausted after usage for every release/iteration/cycle. Hence there is a periodic need to recreate/refresh the test data.

Best Practices

- Identify and analyze the different data sources
- Identify the test data refresh cycle timelines, refresh procedures
- Define timelines for test data refresh
- Enhance the data as per the needs

Validate the data
- Refresh the generated data into the test database based on the refresh cycle
- Repeat the process for each Test Data Refresh
Challenges

Test data that is available for testing is not exhaustive enough to simulate all real time conditions

Best Practices

- Sampling / Subset extraction based on specified filter criteria
- Combination of different samples to increase the test data coverage
- Combination of representative sets of data spanning entire range of data and with adequate test data coverage
Test data gets exhausted after usage for every release/iteration/cycle. Hence there is a periodic need to recreate/refresh the test data.

- Identifying data that is not currently required but may be required for future use
- Select the data based on the filter criteria
- Archive both referential and/or non-referential tables
- Archive the data based on the selection criteria and store the data as flat files (CSV) which occupy low disk space
Challenges

Test data typically comes from numerous data sources and components, Accessing them and Managing them becomes time consuming

Best Practices

• A common test data repository to store all test data from various data sources
• Centralized repository helps different application owners
• Centralized repository helps administrators
• Centralized test data repository helps in automatic re-use of the data
Test Data Management Industry Standards
Process Framework

Enterprise Data Strategy

Test Data Principles
- Data Isolation
- Data Reuse
- Initial state reusability
- Data Safety
- Data Profiling
- Data cataloguing
- Version Control
- Data De-personalization

Test Data Strategy
- Sampling
  - Data profiling
  - Business entity profile definition
  - Data quality issue resolution
  - Data cardinality & distribution
  - Sampling methods
- Creation
  - Analyze existing test data
  - Level of test integration
  - Data Source priority
  - Data volume
  - Data variation
  - Functional Scope
- Maintenance
  - Stability / Segregation
  - Initial State
    - Data refresh
    - Data reinitialize
    - Data reset
    - Data roll forward

Delivery Management
- Integration
- Scope
- Time/Cost
- Quality
- Resource
- Communication
- Risk
- Procurement
Enables organizations to efficiently and effectively respond to emergent, data-intensive business opportunities

Accelerate test data management through rapid data preparation, sub-setting and masking

Reduces level of complexity often associated with managing distributed data and preparing data for testing

Allows companies to securely manage test data and development processes
CTDM Architecture

Source Files
- Excel, CSV

Extract Engine
- File Extract Engine
- Database Extract Engine

Test Data Generation Engine
- Base data
- Conditioned data

Test Data Repository
- Report data

Publish Engine
- Database Publish Engine
- File Publish Engine

Secured Test Data
- Test DB
  - SQL Server
  - Oracle
  - DB2

Test data file
- Excel, CSV

Archived Files (CSV)

Archive Engine

Obfuscation Engine
- Aging Engine

Data Conditioning Engine

Process Reports
Q & A