Good Practices – Data Generation

V1.0
November 17, 2017

Contents

1 Data Generation Methods

There are two methods of synthetic data creation:

1. File Based – Creation of data in files such as delimited, fixed-width etc.
2. DB Based – Data creation directly in relation databases like Oracle, SQL Server etc.

The following subsections outline some best practices for data generation.

1.1 File Based Data Generation

1.1.1 Validation of Sample Data and Layouts

- Ensure that sample files received from team has been pre-tested to be working by running through application. This helps avoid data content related issues.
- Compare sample data with layout to ensure that both are in sync. This can be achieved manually or by registering the layout and importing the sample file in TDM.

1.1.2 TDM Generated Output Comparison

- Publish the sample file imported into TDM into relevant format and then compare with original sample file to ensure that TDM can generate the file correctly.
- Comparison can be done using tools like Beyond Compare, Ultra Edit etc. for quicker turnaround

1.1.3 Validate TDM Repeat Logic

- Create the basic modelling rules to publish multiple records.
- Publish the sample file by increasing the repeat count to 5 or so to validate if the basic modelling for repeat functionality is working.

1.1.4 Generate Sanity Files

- Once basic modelling is complete then output files generated from TDM for both single repeat and multiple repeat to be provided to testing team for validation.
- Repeat as needed so as to ensure that basic modelling is complete and validated before starting advanced modeling operations.

1.1.5 Separate Data Pools for Positive and Negative Scenarios

- Create separate data pools for positive and negative scenarios so that data conditions and modelling can be separated.
- Files should be generated separately for both these scenarios as well so that issue identification is easier.

1.2 DB Based Data Generation

1.2.1 Creation of ER Diagrams

- Generation of ER diagrams for databases with relationships defined or based on custom relationships are recommended to ensure that the DB relationships are documented.
- This ER diagram can be reviewed with DBA/App team to ensure that relationships have been correctly defined especially in the case of relationships not available with database.

1.2.2 Run Schema Comparison

- It is recommended to run a schema comparison between currently registered objects and target schema to ensure that both are in sync before triggering data publishes.
- This helps to validate any schema related issues before publishing so that inconsistent data is not pushed into db especially when foreign keys are not already defined.

1.3 Generic Recommendations

1.3.1 Categorization of Data Entities

- Segregate data entities into static, reference and dynamic values and maintain as data dictionary for quick reference.
- Modeling can then be accomplished for the reference and dynamic values based on business rules provided by application team.