Creating a Subset of Production Data
Abstract
You can create a small, referentially intact copy of production data for a nonproduction environment. A nonproduction environment might include development, test, or training environments. This article describes how to create a data subset operation in Persistent Data Masking and Data Subset (TDM).

Supported Versions
- Persistent Data Masking and Data Subset 9.5.2

Table of Contents
Overview .....................................................................  2
Step 1. Create a Project ...........................................................  3
Step 2. Import Source Metadata to the Project .............................................  3
Step 3. Create Logical Constraints ....................................................  4
Step 4. Create an Entity ...........................................................  5
Step 5. Create a Plan ................................................................  6
Step 6. Generate a Workflow ........................................................  9
Step 7. Run the Workflow ..........................................................  9

Overview
Data Subset creates a small, targeted, and referentially intact copy of production data to use in a non-production environment. You can create a subset of production data that includes tables or files from multiple sources.

Data Subset is a component of Persistent Data Masking and Data Subset (TDM). Create a data subset operation with the TDM Workbench.

A developer needs to create a subset of company department data to test a new application. The developer needs to create a Department table that is about 10% of the size of the production data. The developer also needs to include the DepartmentDetails table in the subset. Each table must include department numbers that are greater than 250. However, the Department table and the DepartmentDetails tables do not have a primary key-foreign key relationship. All the DepartmentDetails rows must have corresponding Department rows in the subset.

Complete the following steps to create the data subset in the TDM Workbench:
1. Create a project.
2. Import metadata that describes the source data you want to use.
3. Create constraints that define a parent-child relationship between the tables.
4. Create an entity to define the tables to include in the data subset.
5. Create a plan that defines the components of the subset project to include in a workflow.
6. Generate the workflow.
7. Run the workflow and monitor it.
**Step 1. Create a Project**

A project is the top-level container that you use to organize the components for the subset operation.

1. In the TDM Workbench, click **Projects**.
   A list of the current projects appears on the **Overview** view.
2. Click **Actions > New**.
3. Enter Department_Subset as the project name. Enter a description.

The default folder name is the project name. The owner is the user name of the project creator.

**Step 2. Import Source Metadata to the Project**

Import source metadata that describes the tables or files in the source data.

You can import source metadata from the PowerCenter repository or you can import metadata by selecting a connection to a database. For this example, import metadata from a folder in the PowerCenter repository.

1. In the **Overview > Data Sources** view, click **Import Data Sources**.
2. Choose to import metadata from the PowerCenter repository and click **Next**.
3. Select the folder in the repository that contains the metadata to import. Click **Next**.
4. Select the tables or files in the repository folder that describe the source data.
   In this example, select the Department and DepartmentDetails tables.

5. Choose to import the metadata immediately instead of scheduling the import.
6. Click **Finish**.

The TDM Workbench notifies you when the import completes successfully.
Step 3. Create Logical Constraints

Constraints define parent-child relationships between source tables. Use constraints to create relationships between tables that you want to include in the data subset.

The Department and DepartmentDetails tables are not related by key. To create a relationship, define a constraint using the Dept_ID column in the DepartmentDetails table.

1. In the Department_Subset project, click the Discover view.
   The TDM Workbench shows a list of the tables and files in the project.
2. Select the DepartmentDetails table in the Discover | Tables view.
3. Click Constraints.
4. Click the Create New Constraint icon.
   The Constraints wizard appears.

5. Enter a name to identify the constraint.
6. Accept the default severity level.
7. Click Select to choose the parent table in the relationship.
   The TDM Workbench shows the DepartmentDetails columns in the left panel and the Department columns in the right panel.
8. To establish the relationship between the tables, select the Dept_ID in the left panel. Select the Dept_ID in the right panel. Click the Link icon to create a link between them.

9. Click Finish.
   The relationship appears in the Relationship panel of the Discover view.
Step 4. Create an Entity

When you create a data entity, you select one table for the entity. The table that you select is the driving table. The TDM Workbench adds tables to the entity based on the constraints that other tables have with the driving table. The TDM Workbench adds all children and parents of the driving table to an entity.

For this example, the driving table is Department. Department is the parent of DepartmentDetails.

1. In the project, click Define | Data Subset.
2. Click New > Entities.
   The Entity Wizard appears.
3. Enter the entity name and description.
4. Click Select Table to select the driving table.
   The Select Driving Table dialog box appears.
5. Select the Department table as the driving table.
   The TDM Workbench adds the DepartmentDetails table to the entity because it is a child of the Department table. The Entity Map view shows the relationships between the tables.
6. Drag the right side of the Entity Map view toward the center of the view to decrease the size.
   The Table Properties view appears on the right.
7. Click the Department table in the Entity Map view.

8. Click Edit to configure subset criteria.

9. Select the Dept_ID column. Select the Greater Than operator, and enter 250 as the value. The subset data must contain department numbers greater than 250.

10. Click Save.


**Step 5. Create a Plan**

A plan defines a data subset operation. A plan includes the components that you need to generate a workflow. The plan can also include additional criteria to limit the data in the target database.

1. Click Actions > New to create a plan.
2. Enter a name and description of the plan and click Next.
3. Click Next to skip adding data masking components to the plan.
4. Click Add Subset Components.
5. Select the Department_Entity entity to add it to the plan.

6. Click Next.

   Define additional criteria to limit the subset results for the workflow.

7. Select the Department_Entity subset component in the Plan Components column.

   In the next step, create a limit that limits the size of the subset target.

8. Click Limit.

9. Click the Add icon to add a limit.

   The Tables view shows a list of tables in the entity.
10. Select the Department table and click **OK**.
11. Configure the **Valuetype** as percent and change the **Value** to 10.

   The data subset operation creates a subset that contains 10% of the driving table rows. The subset also contains the corresponding DepartmentDetails rows.

12. Click **Save** and return to the **Plan | Properties** view.
13. In the **Plan Settings**, edit the source and target connections.

14. In the **Source and Target Properties**, choose to truncate the target table.
15. Click **Save**.
Step 6. Generate a Workflow

When you generate a workflow, the PowerCenter Repository Service generates mappings for the workflow.

2. Choose Schedule Now to generate the workflow immediately.
3. Click Generate Workflow. The PowerCenter Repository Service generates the workflow.
4. Navigate to the project tab. Click Monitor to view the status of the workflow generation task.
5. Click Job ID to view a log of the Generate Workflow task.

Step 7. Run the Workflow

Run a workflow in the Workflow view. View the run status and the workflow logs from the Workflow Executions view.

1. In the project, click the Workflow view.
2. Select the workflow you want to run. The workflow details appear. You can view each session name in a workflow and the number of tables that the session processes.
3. Click Actions > Execute Workflow.
4. Choose the PowerCenter Integration Service to run the workflow.
5. Click the plan Workflow Executions view to view the workflow status. The Workflow Executions view shows a list of workflows that you ran for the plan.
6. Select a workflow to view the session statistics. You can click the Workflow ID to view the workflow log. Click the Session ID to view the session log.

Authors

Ellen Chandler
Principal Technical Writer