INFORMATICA®

Introduction to ILM Test Data Management

Part Number:

Table of Contents

Introduction to ILM Test Data Management

Product Overview

Test Data Management integrates PowerCenter and Informatica application services to provide a solution that you can use to manage non-production data in your organization. Test Data Management integrates with PowerCenter 8.6.1, 9.0.1, and 9.1.0.

With Test Data Management, you can protect sensitive data to perform test, development, or training tasks. You can also create a subset database to reduce the cost and maintenance of non-production systems.

Use data subset to create a small environment for non-production tasks. You can define the type of data that you want to include in the subset database. You might create a subset database with data based on time, function, or geographic location. For example, a time-based subset database might include recent payment transactions from all invoice data in a production system.

Use data masking to change sensitive production data to appear as realistic data in non-production environments. After you identify the data to mask, you define masking rules. You can configure different masking techniques for a masking rule, such as key, substitution, or special masking techniques. For example, you can use key masking to mask a Customer ID with a random number that is repeatable with each instance of the ID.

Before you create a data subset plan, you can run primary and foreign key profiles to discover potential primary-to-foreign key constraints and add them to source tables. Constraints define relationships between parent and child tables and must be in place to run data subset operations in a workflow.

Before you create a data masking plan, run sensitive field profiles to discover which ports contain sensitive data. Use the profile results to determine which columns to mask, and which data masking rules to create. Then, create and assign rules to the ports that contain sensitive data.

To complete data subset and masking operations, you generate and run workflows from data subset and data masking plans in the ILM Workbench.

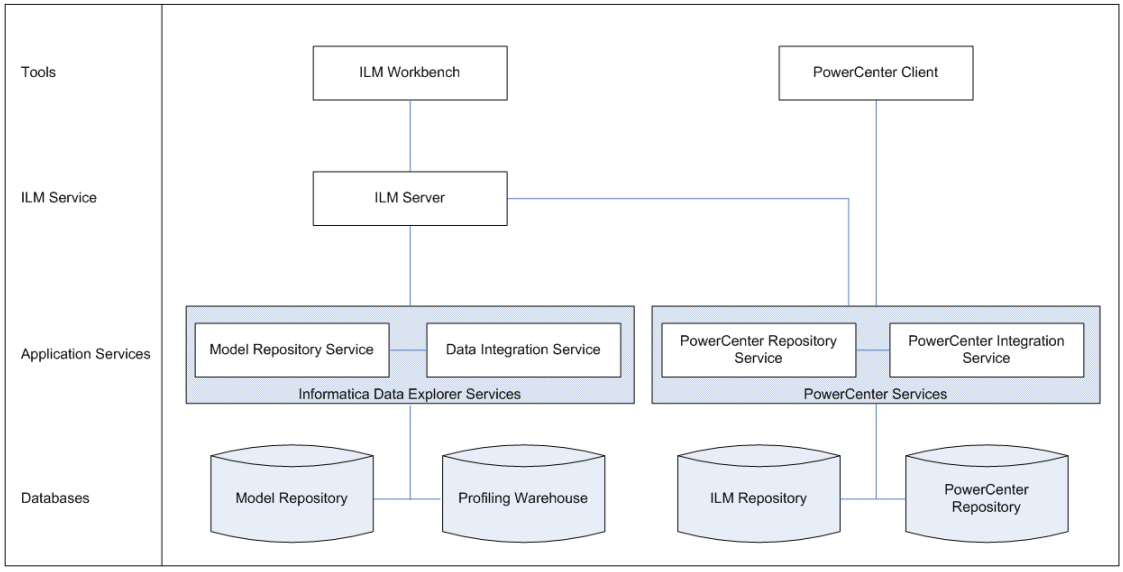
You can also use the ilmcmd command line program to complete a subset of the ILM Workbench tasks from the command line.

ILM Workbench users have roles and privileges that determine which tasks they can perform through the ILM Workbench or the ilmcmd command line program. A compliance officer role creates and manages rules, policies, and patterns. An ILM developer role can view and manage folders, plans, applications, and application components. The ILM administrator role manages roles and privileges for all users.

Test Data Management Architecture

The Test Data Management architecture consists of tools, services, and databases.

The following figure shows the components of Test Data Management:



Tools Component

The tools component of Test Data Management consists of the ILM Workbench and PowerCenter Client applications.

ILM Workbench

A web-based client application that you can use to perform data discovery, such as primary key, foreign key, and sensitive field discovery, and perform data subset and masking operations for data sources.

PowerCenter Client

A client application that you use to define sources and targets, import source metadata into the PowerCenter repository, and define connections to relational sources. The PowerCenter Client connects to the PowerCenter repository through the PowerCenter Repository Service to modify PowerCenter repository metadata. It connects to the PowerCenter Integration Service to start workflows.

Services Component

The services component of Test Data Management consists of the ILM Server, which runs as a service, Informatica Data Explorer services, and PowerCenter services. The Informatica Data Explorer services include the Model Repository Service and Data Integration Service, which perform data discovery operations and manage the metadata of profile objects. The PowerCenter services include the PowerCenter Repository Service and PowerCenter Integration Service, which perform data subset and masking operations.

Data Integration Service

An application service that performs data discovery operations for the ILM Workbench.

ILM Server

An application service that runs the ILM Workbench and manages connections between service components and ILM Workbench users.

Model Repository Service

An application service that manages the Model repository.

PowerCenter Integration Service

An application service that runs workflows that you generate in the ILM Workbench for data subset and masking operations, and extracts data from sources and loads data to targets.

PowerCenter Repository Service

An application service that manages the ILM repository and PowerCenter repository.

The PowerCenter Repository Service accepts requests from the ILM Server to create and modify repository metadata and accepts requests from the PowerCenter Integration Service for metadata when a workflow runs.

Databases Component

The databases component of Test Data Management consists of the ILM repository, Model repository, PowerCenter repository, and profiling warehouse.

ILM repository

A relational database within the PowerCenter repository that stores metadata for sources that you import into the ILM Workbench, constraints that you add to tables in the ILM Workbench, components that you create in the ILM Workbench, and data dictionaries that you use to mask data.

Model repository

A relational database that stores the metadata for tables that you profile for data discovery in the ILM Workbench.

PowerCenter repository

A relational database that stores metadata for the sources and targets that you import into the PowerCenter Client, and the workflows that you generate from plans in the ILM Workbench.

Profiling warehouse

A database that stores data for tables that you profile for data discovery in the ILM Workbench.

Test Data Management Connections

To perform data subset, masking, and discovery operations, you need to connect to databases, repositories, and services. The connections that you need depend on the operations that you want to perform.

To perform data discovery operations, an application requires connections to a database source and a Data Integration Service.

To perform data subset and masking operations, workflows that you generate from plans require connections to services, the ILM repository, external database sources, and PowerCenter sources and targets.

Service Connections

Test Data Management requires connections to services.

Data Integration Service

To perform data discovery operations, an application requires a connection to a Data Integration Service, which is the service in the Informatica domain that performs the data discovery operation.

You can select a Data Integration Service for each application.

PowerCenter Integration Service

To perform data masking or data subset operations, a workflow that you generate from a plan requires a connection to a PowerCenter Integration Service.

The first time that you generate a workflow from a plan, you select a connection to a PowerCenter Integration Service. That connection is available to all workflows in all applications.

PowerCenter Repository Service

When you import metadata into the ILM Workbench from a PowerCenter source, the ILM Server sends a request to the PowerCenter Repository Service to extract source metadata from the PowerCenter repository and load it to the ILM repository.

You install the ILM Server on the machine that hosts the Informatica Server. During the ILM Server installation, you provide connection details to the PowerCenter repository. The ILM Server installation connects to the PowerCenter repository and creates the ILM repository tables.

Repository Connections

Test Data Management requires connections to repositories.

ILM repository

When you use the ILM Workbench to create a data subset or masking component such as an entity, rule, or data dictionary, the PowerCenter Integration Service stores the component in the ILM repository.

To perform data subset or masking operations, a workflow requires a connection to the ILM repository. After you select the ILM repository connection for a workflow, it is available to all workflows in all applications.

Model repository

When you run profiles to discover data, the ILM Server sends a request to the Data Integration Service to extract data for the profiled tables. The Data Integration Service sends a request to its associated Model Repository Service to load the metadata for the profiled tables to the Model repository.

When you create a Model Repository Service in the Administrator tool, you create or use an existing Model repository.

Profiling warehouse

When you run profiles to discover data, the ILM Server sends a request to the Data Integration Service to extract data for the profiled tables. The Data Integration Service loads the data for profiled tables to the profiling warehouse.

When you create a Data Integration Service in the Administrator tool, you configure a profiling warehouse.

PowerCenter repository

When you import metadata into the ILM Workbench from a PowerCenter source, the ILM Server sends a request to the PowerCenter Repository Service to copy source metadata from the PowerCenter repository into the ILM repository.

You install the ILM Server on the machine that hosts the Informatica Server. During the ILM Server installation, you provide connection details to the PowerCenter repository. The ILM Server installation connects to the PowerCenter repository and creates the ILM repository tables.

Database Connections

Test Data Management requires connections to databases.

To connect Test Data Management to databases, you do not need to install a separate driver.

External Database Connection

To perform data discovery operations, an application requires a connection to the database source for which you want to discover data. Configure the connection in the **Design** perspective of an application. For each application, you can configure one database connection. You can change a configured connection.

PowerCenter Relational Database Connections

When you generate workflows from data subset or masking plans, you must select connections for any relational sources and targets. You create relational database connections for sources and targets in the PowerCenter Workflow Manager, and you select these connections in the ILM Workbench when you configure a workflow. These connections are usable by all workflows that you configure in the ILM Workbench.

Test Data Management uses these connections to generate workflows from plans, and the PowerCenter Integration Service uses these connections to run the workflows.

Test Data Management Process

When you begin a data masking or subset project, you import source metadata into the ILM Workbench.

Then you run data discovery profiles to discover primary and foreign keys and sensitive fields in tables. Primary and foreign key discovery enables you to add constraints to parent and child tables, while sensitive key discovery helps you determine which data masking rules to create and which ports to assign them to in the source.

You create and generate a workflow from a data subset plan. You run this workflow to create a subset database.

You create and generate a workflow from a data masking plan. You run this workflow to mask data in a target database.

One plan can generate a workflow that performs both data subset and data masking operations.

Data Import Process

Import metadata for the sources on which you want to perform data subset or masking operations. You can import metadata for a PowerCenter source or an external database source.

Complete the following tasks in the ILM Workbench to implement the data import process:

1. Import metadata from a PowerCenter source or external database.
2. Create applications and associate them with the imported sources. Those sources are then available to data subset and data masking plans.

When you import PowerCenter source metadata, the ILM Server sends a request to the PowerCenter Repository Service to extract source metadata from the PowerCenter repository and load it to the ILM repository. When you import external database metadata, the ILM Server extracts metadata from source tables and loads it into the ILM repository.

Data Discovery Process

Run data discovery profiles to discover primary and foreign key data and sensitive field data in source tables.

Complete the following tasks in the ILM Workbench to implement the data discovery process:

1. Configure and test a connection to a database.
2. Configure a Data Integration Service to discover the data in the configured database.

3. Create and run primary key and foreign key profiles to discover potential constraints.
4. Add constraints to tables.
5. Create and run a sensitive field profile to discover sensitive columns in one or more tables.

When you run data discovery profiles, the ILM Server sends a request to the Data Integration Service to extract data for the profiled tables. The Data Integration Service loads the data for profiled tables to the profiling warehouse. When you add constraints to tables, the ILM Server stores the constraints in the ILM repository.

Data Subset Process

To create a subset database, create an application and add entities, groups, and templates to it. Entities define parent and child tables that you want to include in the subset database, while groups define unrelated tables that you want to include in the subset database. A template contains entities and groups.

When you create a data subset plan, the entities, groups, and templates in the application are available to add to the plan. To create the subset database, generate and start a workflow from the plan.

Complete the following tasks in the ILM Workbench to implement the data subset process:

1. Create an application, and associate it with the sources that you want to extract data from to load into the subset database.
2. Create entities, groups, and templates to define the tables that you want to copy to the subset database. An entity defines a set of tables that are related based on physical or logical constraints. A group defines a set of unrelated tables. A template contains entities and groups.
3. Create a data subset plan, and add entities, groups, and templates to it. For each port in the parent table, define one or more parameters to filter the data in the port.
4. Generate and start a workflow from the plan.
5. Monitor the progress of the workflow.

When you create applications, entities, groups, templates, and plans, the ILM Server stores these components in the ILM repository. When you generate and run workflows from plans, the PowerCenter Integration Service runs the workflows and loads the data into the subset database.

Data Masking Process

To perform data masking operations, you create an application, rules, and policies. You can also import mapplets from PowerCenter as rules. You assign rules to ports in an application source. Then, you assign rules to policies, and the policies to plans.

A rule defines a data masking technique and its masking parameters. A policy defines the data masking rules, the data to mask, and masking parameters for a source.

Complete the following tasks in the ILM Workbench to implement the data masking process:

1. Use the results of sensitive field data discovery to determine which columns to mark as sensitive in application sources.
2. Define patterns. Use a pattern as part of filtering criteria to search for the ports that you want to mask.
3. Create an application, and associate it with the sources from which you want to extract and mask data.
4. Create rules and assign them to ports in the application source. Optionally, select a pattern.
5. Create policies and assign rules to the policies.
6. Create a data masking plan and assign policies to it.
7. Generate and start a workflow from the plan.

8. Monitor the progress of the workflow.

When you create applications, rules, policies, and plans, add constraints to tables, or define data dictionaries for data masking, the ILM Server stores these components in the ILM repository. When you generate and run workflows from plans, the PowerCenter Integration Service runs the workflows and masks and loads the data into the target database.

Test Data Management Example

A company wants to enforce a policy for masking sensitive data in disparate and large application environments.

The test professionals in the organization need to access representative data to test the new policy and ensure that the sensitive data is not compromised.

The company uses Test Data Management to establish and enforce the policy.

To begin the process, the compliance officer gathers data masking requirements for the new policy. Then, the compliance officer creates data masking rules and policies in the ILM Workbench.

To create a subset database to test the new policy with representative data from the various application environments, an IT developer runs primary and foreign key profiles. The profiles discover potential primary-to-foreign key constraints in the source tables, which the IT developer adds to the tables. With these constraints in place, the IT developer can create entities in an application, which make the entities available for use in plans. An entity defines a set of tables that are related based on constraints. The IT developer also creates groups in the application so that he can include unrelated tables in the subset database. A group defines a set of unrelated tables.

Then, the IT developer runs a sensitive field profile to identify sensitive columns in the source tables. The IT developer discovers columns that contain Social Security number data and assigns the SSN masking rule defined by the compliance officer to the sensitive ports in the application source. Similarly, when he discovers columns that contain credit card data, he assigns the credit card masking rule to the sensitive ports in the application source.

The IT developer creates a plan that performs both data masking and subset operations. He adds entities and groups to the plan to define which data to copy to the subset database. To mask the sensitive columns in the source data, he adds the data masking policies defined by the compliance officer to the plan. When he generates and runs a workflow from the plan, the PowerCenter Integration Service runs the workflow and loads the masked data into the subset database.

The test professionals verify that the data masking rules correctly masked the data in the subset database.

Finally, the compliance officer validates the results in the subset database.

INDEX