

Data Validation Option Integration with Cloudera

Abstract

You can use Cloudera Impala to process queries directly on the Apache Hadoop data stored in HDFS or HBase. Data Validation Option contains a graphical user interface to create and run tests on data including Hadoop data. This article describes how you can integrate Data Validation Option with Cloudera Impala.

Supported Versions

- PowerCenter 10.0 - 10.1.1

Table of Contents

Overview.	2
Prerequisites.	3
Configure Cloudera Impala for Data Validation Option.	3
Step 1. Download the Cloudera ODBC Driver for Impala.	3
Step 2. Install the Cloudera ODBC Driver for Impala.	4
Step 3. Configure a Cloudera ODBC Driver for Impala Data Source on Windows.	5
Step 4. Import Metadata in PowerCenter.	7
Step 5. Create a Connection Object.	8
Step 6. Refresh the Data Validation Option Repository.	10
Step 7. Add a Table Pair in the Data Validation Client.	11
Step 8. Generate Tests for a Table Pair.	12
Step 9. Run Table-Pair Tests.	14

Overview

You can use Data Validation Option to verify the accuracy of the data after data migration, replication, integration or other similar data movement or transformation exercise. You can use Cloudera Impala to enable Business Intelligence (BI), analytics and reporting on Hadoop or Impala-based data.

Data Validation Option automates data validation and makes tests repeatable. You create tests once and run the tests each time PowerCenter loads a batch of data to the target.

Use Cloudera ODBC driver for Impala for direct SQL and Impala SQL access to Apache Hadoop or Impala distributions. The driver transforms the SQL query of an application into the equivalent form in Impala SQL. Impala SQL is a subset of SQL-92. If an application is Impala-aware, you can configure the driver to process the query. The driver gets schema information from Impala to present to a SQL-based application. The driver transforms the queries and joins from SQL to Impala SQL.

Cloudera ODBC Driver for Impala is available for Microsoft Windows and Linux. The driver complies with the ODBC 3.52 data standard and includes Unicode and 32-bit and 64-bit support for high-performance computing environments on all platforms.

To import data from Impala, you can install the Cloudera ODBC driver for Impala drivers and configure an ODBC 32-bit connection in PowerCenter. You can configure a 64-bit connection in PowerCenter to read run-time data from the source and write it to the target. You can use Data Validation Option to run tests to validate data accuracy.

When you integrate Data Validation Option with Cloudera, the driver gets schema information from Impala. You can use Cloudera Impala as a source and target in Data Validation Option. You can then compare data in the source and target

when you setup table pair and rules in the Data Validation Option. Data Validation Option communicates with PowerCenter to run the data. You can view test results in the Data Validation client.

Prerequisites

Before you integrate Data Validation Option with Cloudera Impala, verify that you meet the following prerequisites:

1. On a cluster VM, verify that you have the Impala services installed. Work with your IT organization to enable sudo access on a cluster VM with your local user.
2. To avoid a continuous pop-up dialog box appearing for the Impala drivers, perform the following tasks:
 - a. Navigate to the following PowerCenter client installation to the powmart.ini file path:

```
<Informatica installation directory>\bin\powermart.ini
```
 - b. Add the following entry in the powmart.ini file:

```
text=extodbc.dll
```
 - c. Save the file and restart Designer.
3. Verify that you meet the following system installation requirements:
 - Windows XP with SP3, Windows Vista, Windows 7 Professional, or Windows Server 2008 R2 operating system.
 - Minimum disk space of 25 MB.
 - Driver tested with Impala 1.0.1 and Apache Thrift 0.9.0.
 - Administrator privileges. Ensure that you have administrator privileges to install the driver.

Configure Cloudera Impala for Data Validation Option

You can use Cloudera Impala as a source or target with Data Validation Option.

To configure Cloudera Impala with Data Validation Option, perform the following steps:

1. Download the Cloudera ODBC driver for Impala.
2. Install the Cloudera ODBC driver for Impala.
3. Configure a Cloudera ODBC driver for Impala data source on Windows.
4. Import metadata in PowerCenter.
5. Create a connection object.
6. Refresh the Data Validation Option repository.
7. Add a table pair test in the Data Validation client.
8. Generate tests for a table pair.
9. Run table-pair tests.

Step 1. Download the Cloudera ODBC Driver for Impala

You can download and install Cloudera ODBC driver for Impala to access data in a Hadoop cluster.

Download the Cloudera ODBC driver for Impala from the following Cloudera connector website:

<https://www.cloudera.com/downloads/connectors/impala/odbc/2-5-20.html>

Download the following ODBC drivers from the Cloudera website:

- ClouderaImpalaODBC32.msi for 32-bit applications
- ClouderaImpalaODBC64.msi for 64-bit applications

On Windows, download the 32-bit and 64-bit versions of the Cloudera Impala ODBC driver on the client machine.

If you have Informatica services installed on a 64-bit machine, you need to install a 64-bit Cloudera Impala ODBC driver on the client machine.

Step 2. Install the Cloudera ODBC Driver for Impala

You must install 32-bit and 64-bit Cloudera ODBC drivers for Impala. Use the 32-bit driver to import metadata from the Impala server to PowerCenter. Use the 64-bit driver to read data from the Impala server in PowerCenter.

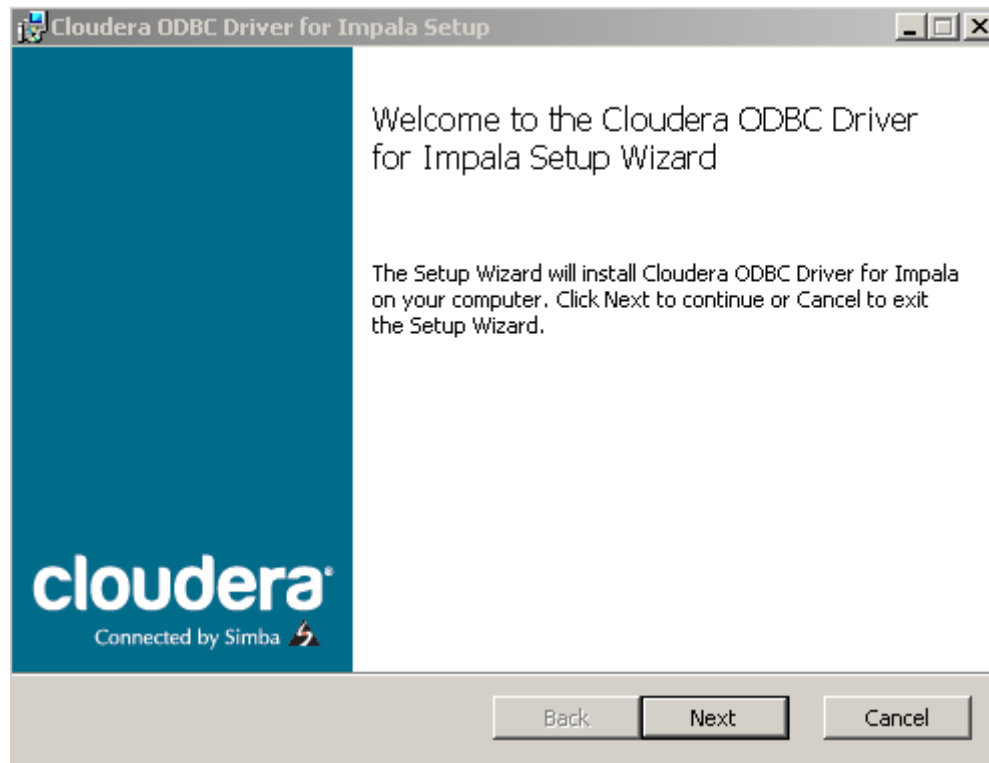
1. Navigate to the location where you download the Cloudera Impala ODBC drivers.
2. Double-click the ODBC driver installer.

Open the 32-bit ODBC driver before you open the 64-bit driver.

Based on your operating system settings, you might get a security warning when you try to open the file.

3. Click **Run**.

The **Cloudera ODBC Driver for Impala Setup Wizard** appears.



4. Click **Next**.
5. To accept the terms of the License Agreement, select the check box and click **Next**.
6. To change the installation directory, click the **Change** button.
7. To set the installation directory, click **OK**.
8. Click **Next**.

9. Click **Install**.
10. After the installation process completes, click **Finish**.
11. Repeat steps [1](#) through [10](#) for the 64-bit driver.

Step 3. Configure a Cloudera ODBC Driver for Impala Data Source on Windows

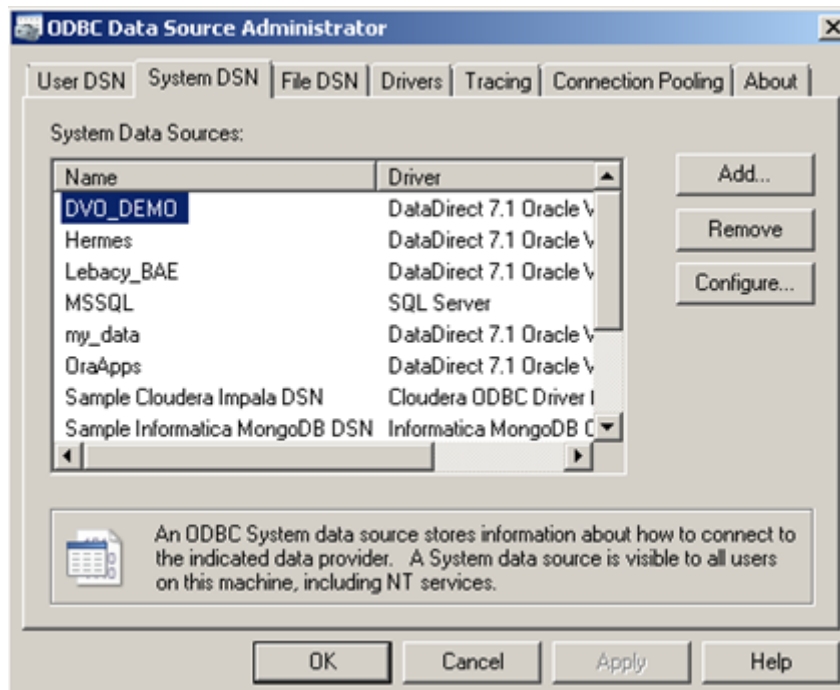
After you install the driver, configure the Data Source Name (DSN). You can use the Windows ODBC Data Source Administrator to create the ODBC data source. When you create the data source, the Windows ODBC Data Source Administrator stores the ODBC driver parameters in the Windows registry.

1. Select the correct version of **ODBC Data Source Administrator**.
 - On a Windows 32-bit system, open the **Control Panel** and click **Administrative Tools**. Then, double-click **Data Sources (ODBC)**.
 - On a Windows 64-bit system, enter the following command at the command prompt:

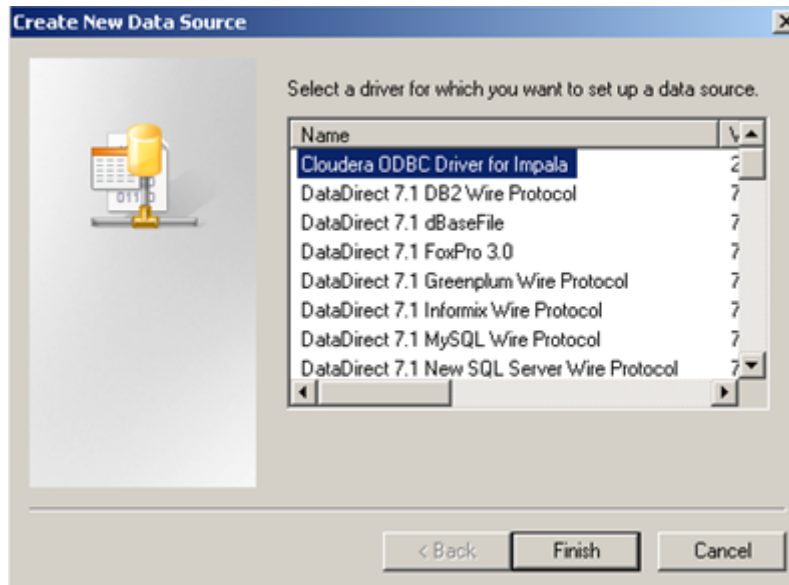
```
%windir%\SysWOW64\odbcad32.exe
```

The **ODBC Data Source Administrator** dialog box appears.

2. In the ODBC Data Source Administrator for the System DSN, click **Add**.

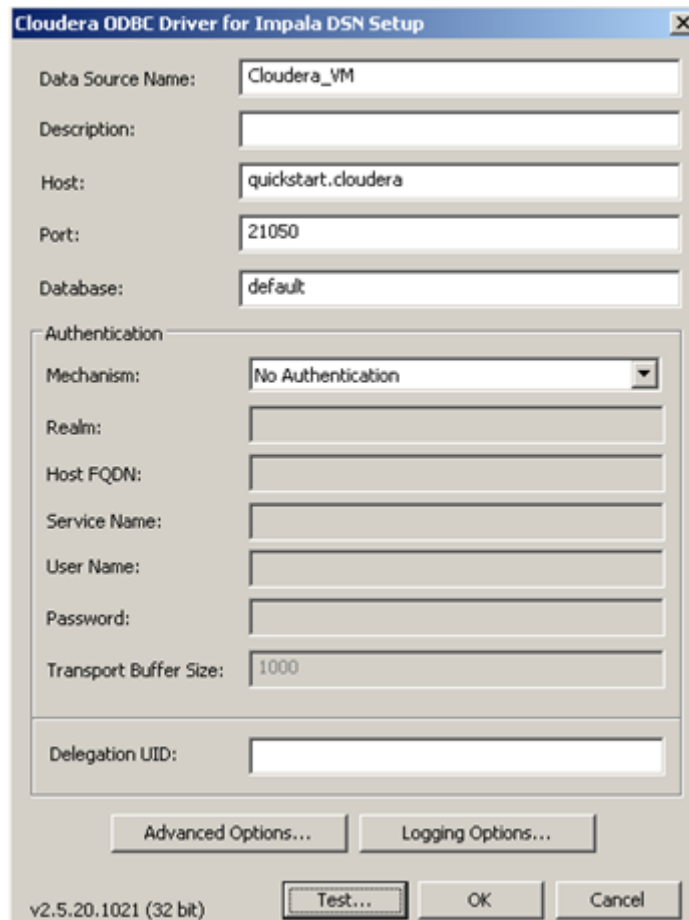


3. Select **Cloudera ODBC Driver for Impala**.



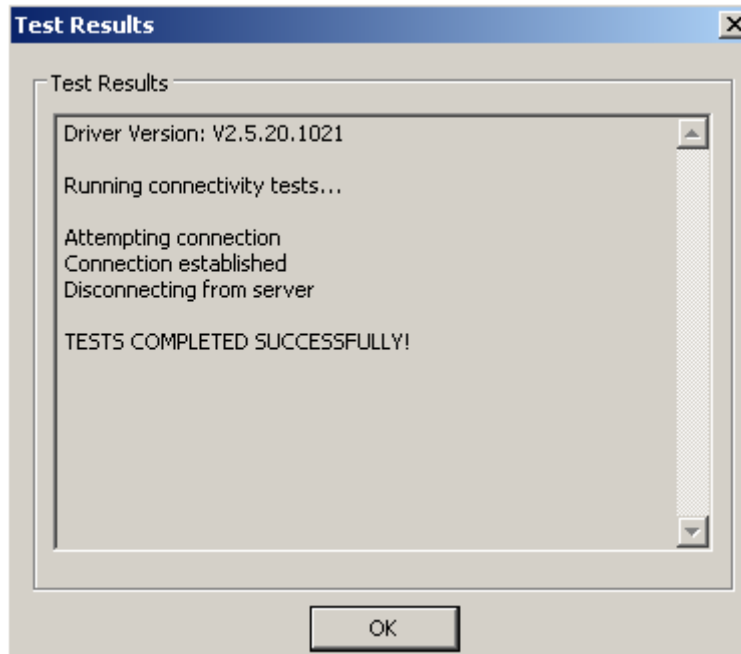
4. Click **Finish**.

The **Cloudera ODBC driver for Impala DSN Setup** window appears.



5. For the Data Source Name, enter a name.

6. For the Host, enter the IP address or host name of the Impala server.
7. For the Port, enter the listening port for the service.
8. To connect to the database, enter **default** for the database.
9. If you require authentication, configure the authentication settings based on the Cloudera documentation.
10. Click **Test** to verify that the connection is valid, and click **OK**.



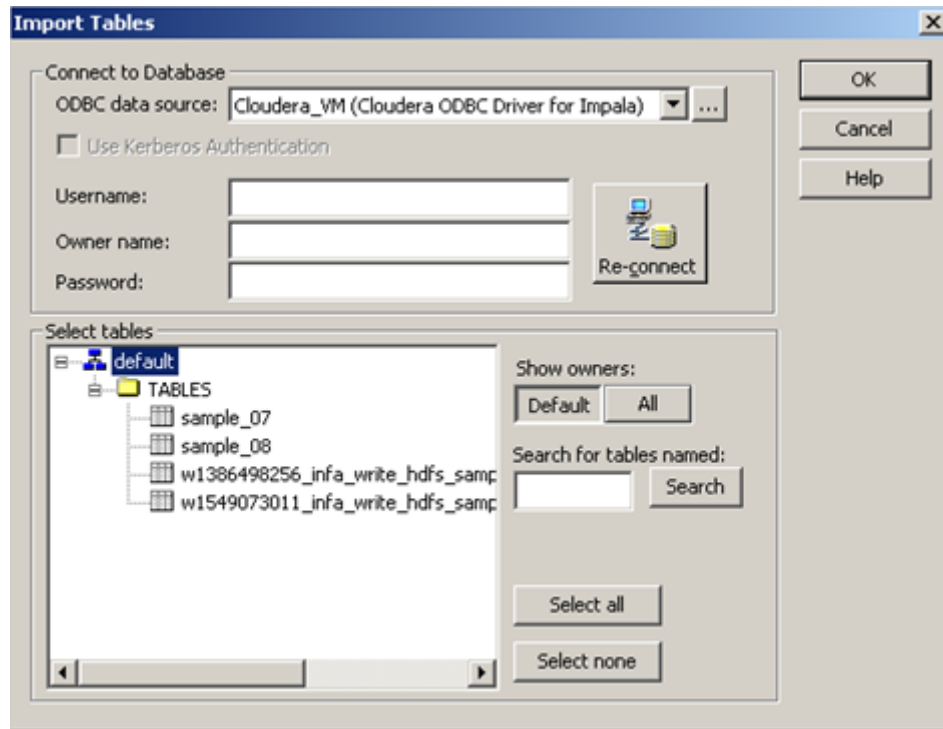
Step 4. Import Metadata in PowerCenter

Import and save the Hive metadata in PowerCenter.

1. Open the Designer and connect to the repository.
2. Open the appropriate folder.
3. In the Source Analyzer, select **Sources > Import from Database** and select the data source name that you created to connect to the Hive instance.

To use HDFS with Data Validation Option, create a Hive view on top of HDFS data or stage the data to the local file system on the machine where Informatica runs.

4. For Show Owners, select **Default**.



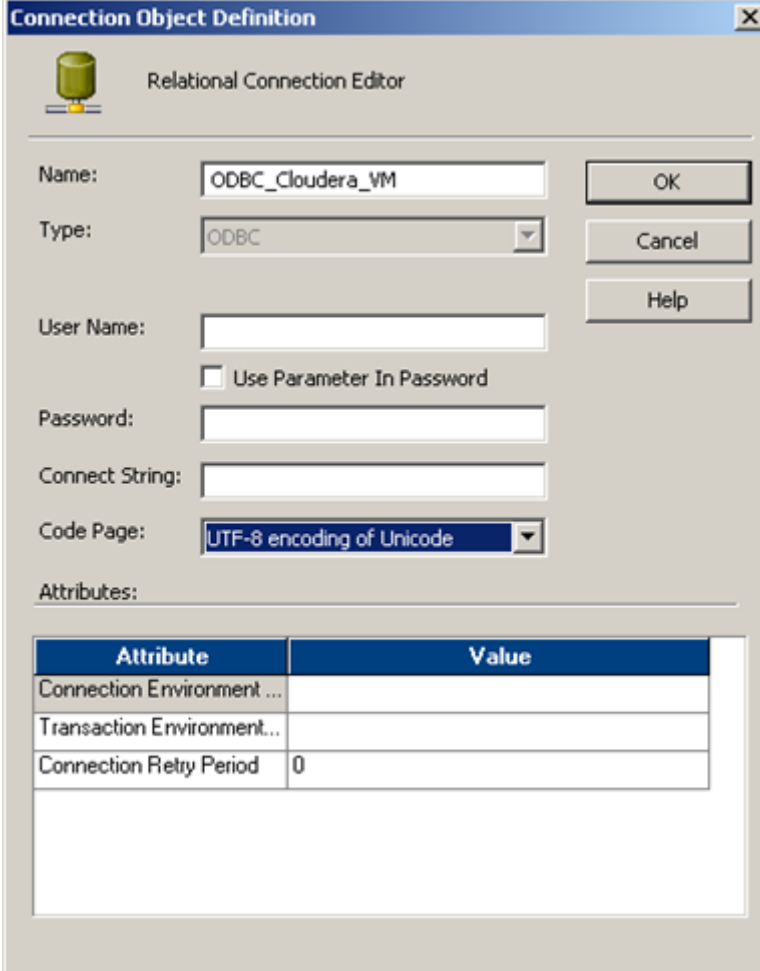
5. To connect to the database, select **Connect**.
6. Select the Hive tables that you want to validate.
7. Click **OK**.
PowerCenter imports the metadata.
8. Save the Hive tables.

Step 5. Create a Connection Object

Before you import and schedule the workflow, you must configure connections in the Workflow Manager. Configure the connection to the repository in the PowerCenter Workflow Manager.

1. In the Workflow Manager, click **Connections** and select **Relational** connection type.
The **Relational Connection Browser** dialog box appears.
2. Click **New**.
The **Select Subtype** dialog box appears.
3. Select **ODBC** from the **Select Type** list.
4. Click **New** to create the connection.
5. Select the database and click **OK**.

The **Connection Object Definition** dialog box appears.



Connection Object Definition

Relational Connection Editor

Name: ODBC_Cloudera_VM

Type: ODBC

User Name:

Use Parameter In Password

Password:

Connect String:

Code Page: UTF-8 encoding of Unicode

Attributes:

Attribute	Value
Connection Environment ...	
Transaction Environment...	
Connection Retry Period	0

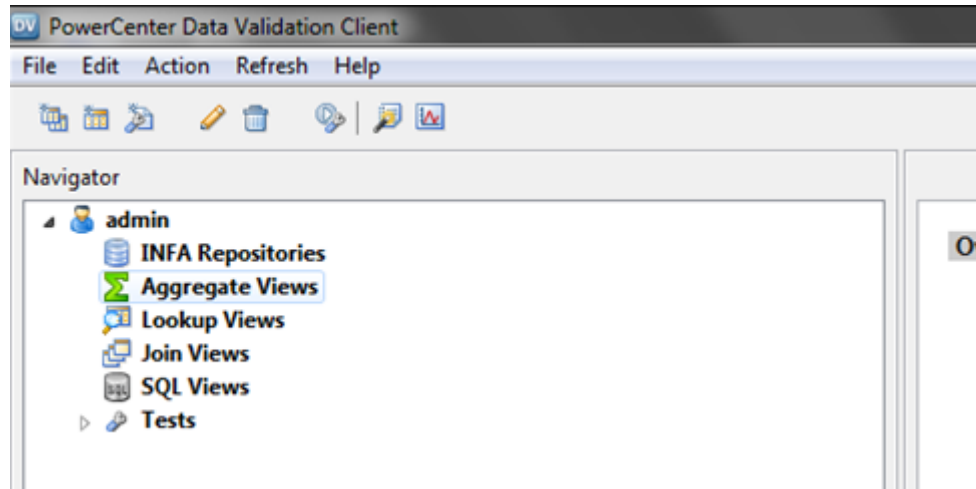
- Specify the relational connection properties. For connection string, enter the 64-bit system data source name that you provided to connect to the Hive instance installation.
- Click **OK**.

The database connection appears in the **Connection Browser** list.

Step 6. Refresh the Data Validation Option Repository

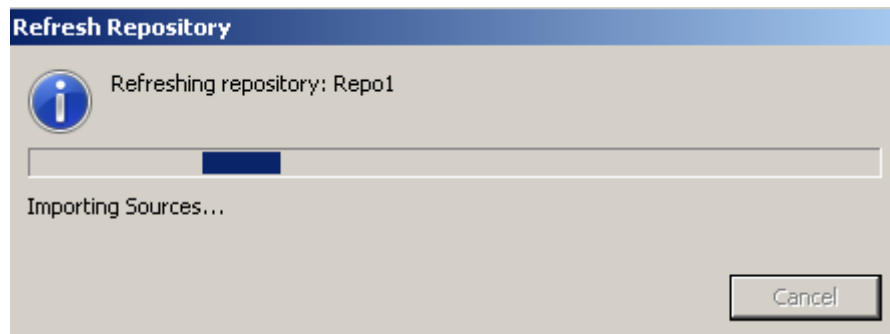
When you refresh a repository, Data Validation Option imports the source, target, folder, and connection metadata from the PowerCenter repository.

1. Open the Data Validation Client.



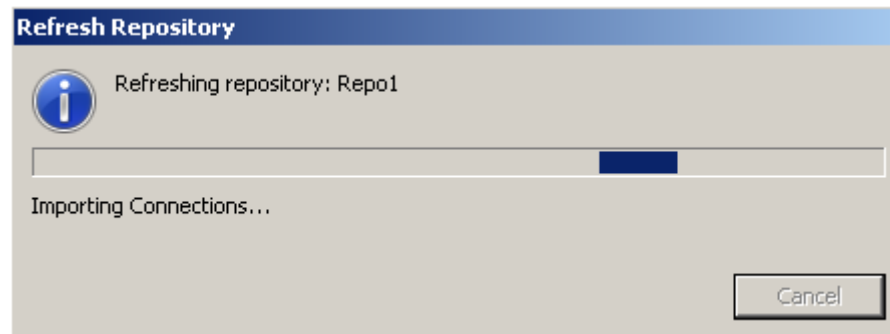
2. Right-click the appropriate repository and select **Refresh Repository**.
3. When you refresh a repository, you can select the objects to refresh. For Data Validation Option to import source and target metadata again, select **Refresh Folder (Sources and Target)**.

The refresh process starts for the contents of source and target folders in each folder in the repository.



4. For Data Validation Option to import connection metadata again, select **Refresh Connections**.

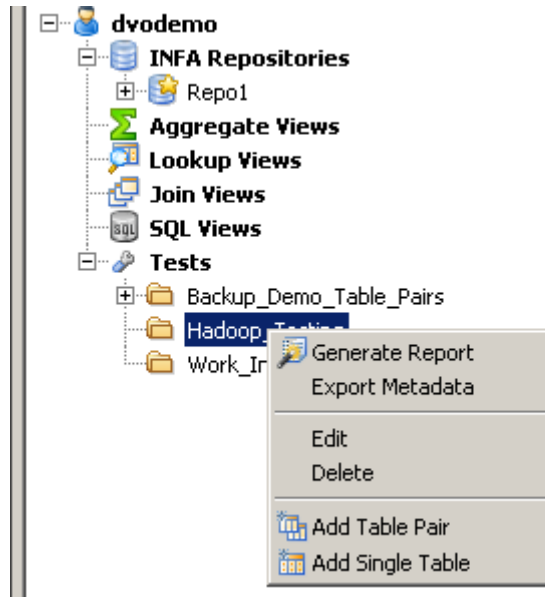
The refresh process starts for the connection metadata.



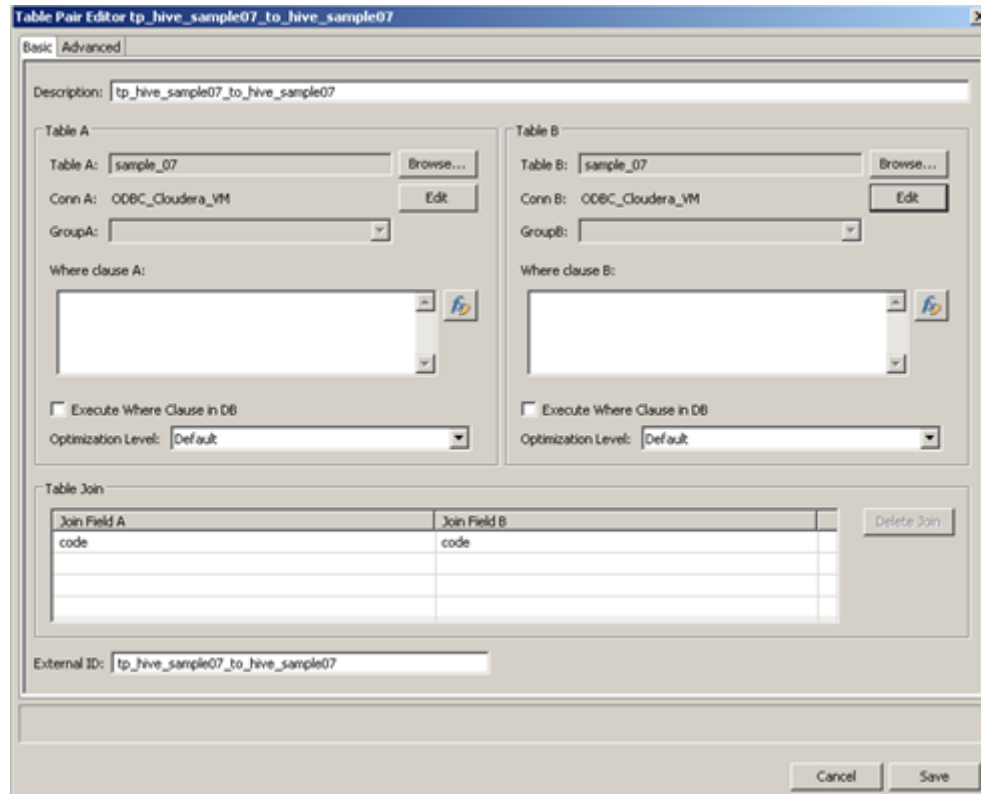
Step 7. Add a Table Pair in the Data Validation Client

You can add a table pair from the file menu or from the shortcut in the menu bar of the Data Validation client.

1. Select the folder to which you want to add the table pair.
2. Right-click the folder and select **Add Table Pair**.



The **Table Pair Editor** window appears.



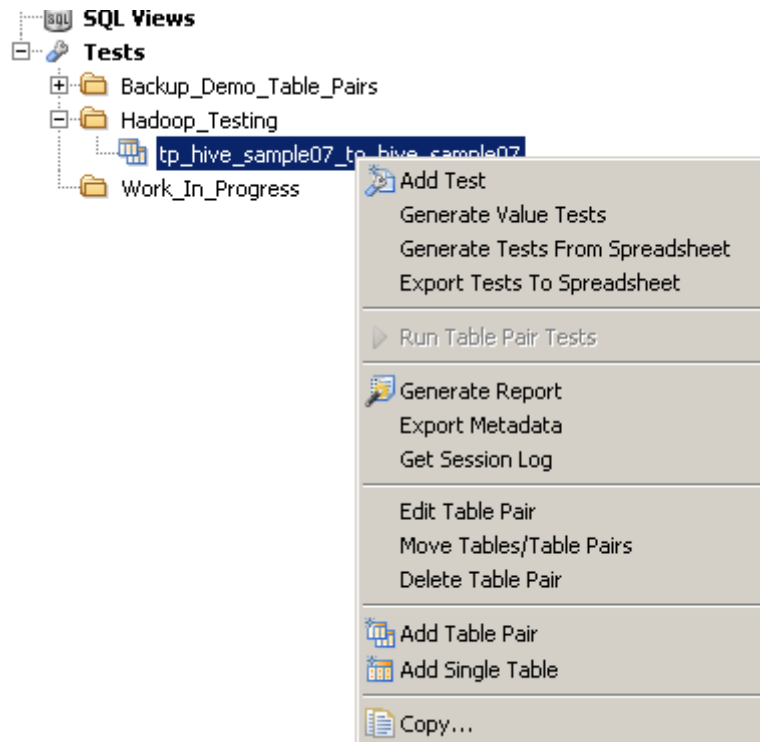
3. On the **Basic** tab, enter the table pair description and the table names.

4. To select the table names, click **Browse**.
5. To configure the connection properties for each table, click **Edit**.
Specify the Cloudera 64-bit data source name for the connection name. For example, ODBC_Cloudera_VM.
6. Select the optimization level as **Default**.
7. For the table join, enter the join values for the fields.
8. To run the Data Validation Option test at the command line, enter the external ID.
9. To save the table pair, click **Save**.

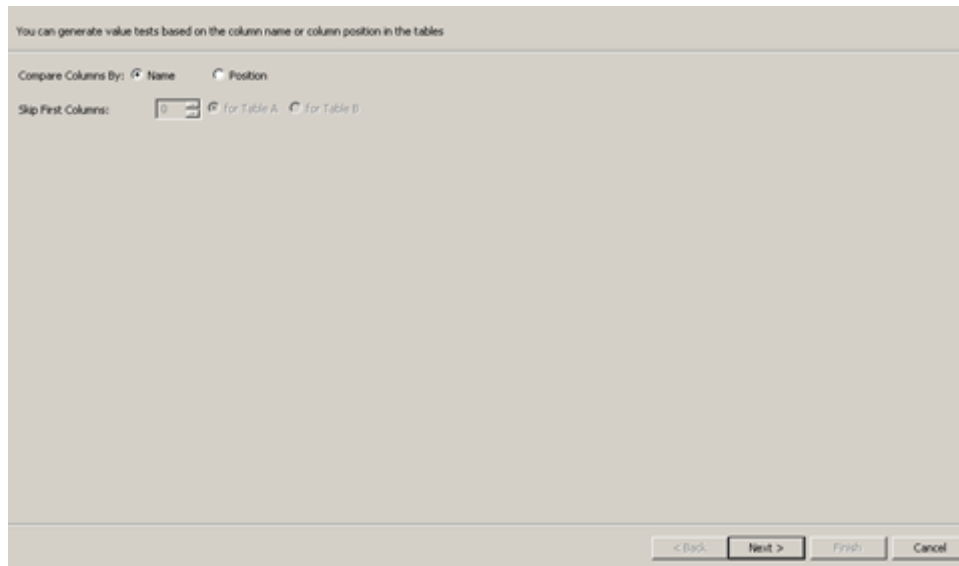
Step 8. Generate Tests for a Table Pair

You can generate count and value tests for an existing table pair.

1. In the Navigator, right-click the table pair that you created and select **Generate Value Tests**.

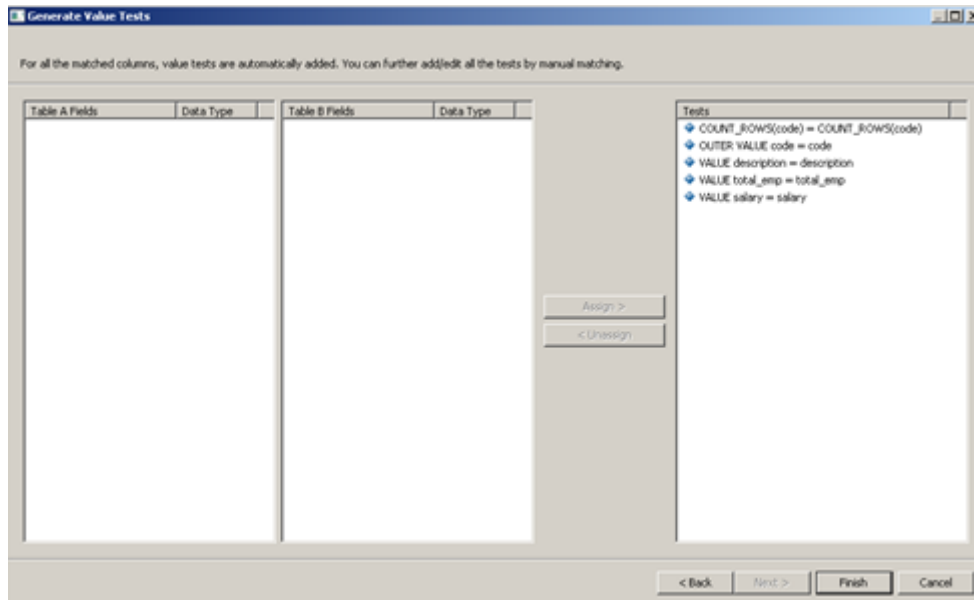


2. Select to compare columns by name.



3. Click **Next**.

The **Generate Value Tests** dialog box shows the tests that are going to automatically be generated based on the matched columns. A blue diamond appears next to each automatically generated test.



4. To manually match additional columns, drag a column in Table A to the matching column in Table B.

5. Click **Finish**.

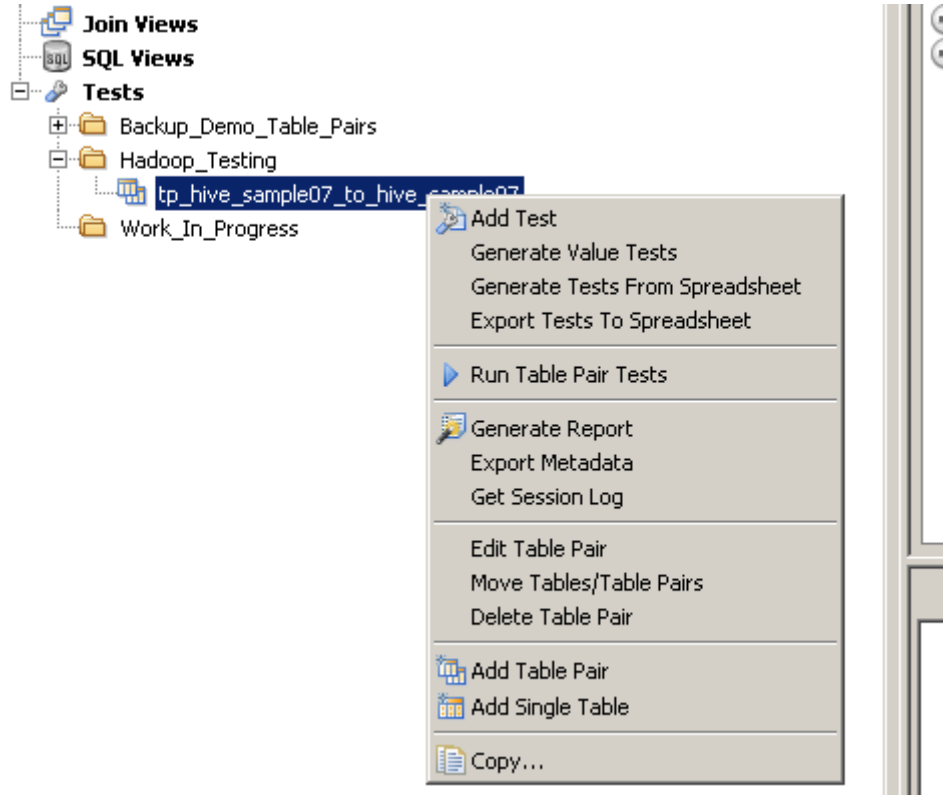
Data Validation Option generates the count and value tests for the matched columns for the table pair.

Step 9. Run Table-Pair Tests

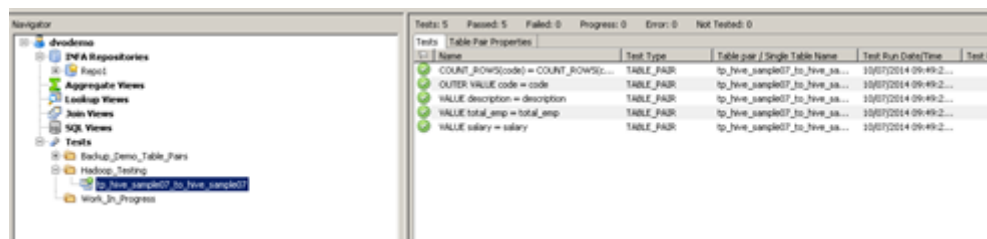
You can run tests from the Data Validation Client or the DVOCmd command line.

Before you run tests, verify that the Data Validation Option target folder is closed in the PowerCenter Designer and the PowerCenter Workflow Manager. If the target folder is open and you run tests, Data Validation Option cannot write to the target folder and the tests fail.

1. To run all tests defined for a table pair, right-click the table pair object in the Navigator.
2. Click **Run Table Pair Tests**.



After you run a test, you can view the results on the Results view in the Properties area.



Authors

Harish Alagarsamy

Sujitha Alexander