Data Quality Management using DecisionSpace® Data Quality

Release 5000.10.3.0

Volume I

HALLIBURTON | Landmark

Part Number 230070 Revision B

October 2015

Copyright and Trademarks

Halliburton | Landmark
10200 Bellaire Blvd., Houston, Texas 77072-5206, USA
P.O. Box 42810, Houston, Texas 77242-2810, USA
Phone: 281.575.3000, Fax: 713.839.2015
Internet: https://www.landmarksoftware.com

Trademarks

Note

Third Party Licenses and Attributions

Halliburton acknowledges that certain third party code has been bundled with, or embedded in, its software. The licensors of this third party code, and the terms and conditions of their respective licenses, may be found at the following location:

PathNameInInstallationDir`/Third_Party.pdf`

Disclaimer

The programs and documentation may provide links to external web sites and access to content, products, and services from third parties. Halliburton is not responsible for the availability of, or any content provided on, third party web sites. You bear all risks associated with the use of such content. If you choose to purchase any products or services from a third party, the relationship is directly between you and the third party. Halliburton is not responsible for: (a) the quality of third party products or services; or (b) fulfilling any of the terms of the agreement with the third party, including delivery of products or services and warranty obligations related to purchased products or services. Halliburton is not responsible for any loss or damage of any sort that you may incur from dealing with any third party.

Data Quality Management using DecisionSpace® Data Quality Release 5000.10.3.0

Volume I

Chapter 1
Introduction

Welcome to the Data Quality Management using DecisionSpace® Data Quality class. This course introduces you to the DecisionSpace Data Quality software that helps you evaluate, correlate, correct, monitor, and synchronize data across the enterprise.

It covers the functionality and use of the modules within the Data Quality application so that you can manage projects, users, data source connections, and the rules repository to:

- Identify the full spectrum of data issues including ambiguous, incomplete, inaccurate, inconsistent, or missing data
- Clean the data before it impacts end-user workflows or results in poor decision making

Course Objectives

This manual is designed to be used with the DecisionSpace Data Quality application. Refer to the exercises during class and use the manual for reference when you return to your workplace.

Briefly, this manual covers the following topics:

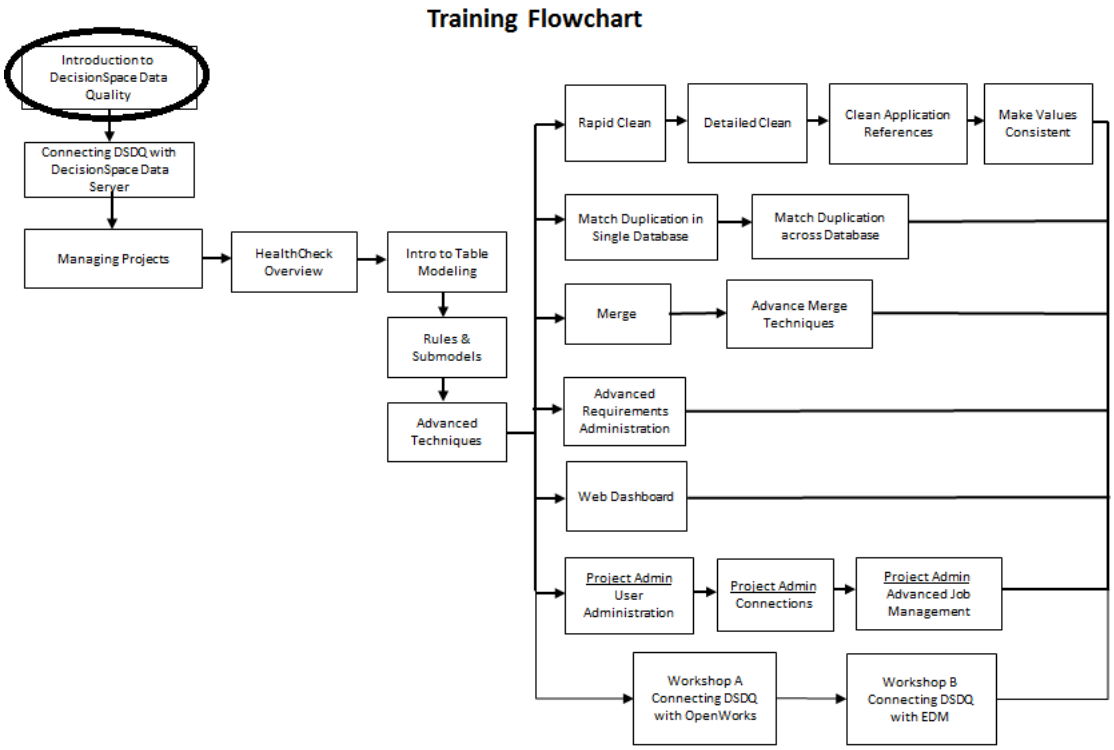
- An overview of the DecisionSpace Data Quality (DSDQ) application
- Connecting the Data Quality application with DecisionSpace Data Server
- Managing projects in the Data Quality application
- Using the Data Quality Phases i.e. HealthCheck, Clean, Match & Merge
- Managing the Data Quality Rules Repository
- Using the Data Quality Web Dashboard for analyzing the data quality assurance process
- Managing users and projects

Chapter Overview

In this chapter, you will learn about:

- The DecisionSpace Data Quality software including the Data Quality Project window and components
- Installing the Data Quality application
- Starting the DecisionSpace Data Quality software

Topics covered in each chapter are outlined in the following illustration. Those specific to the current chapter will be circled in black for your reference.



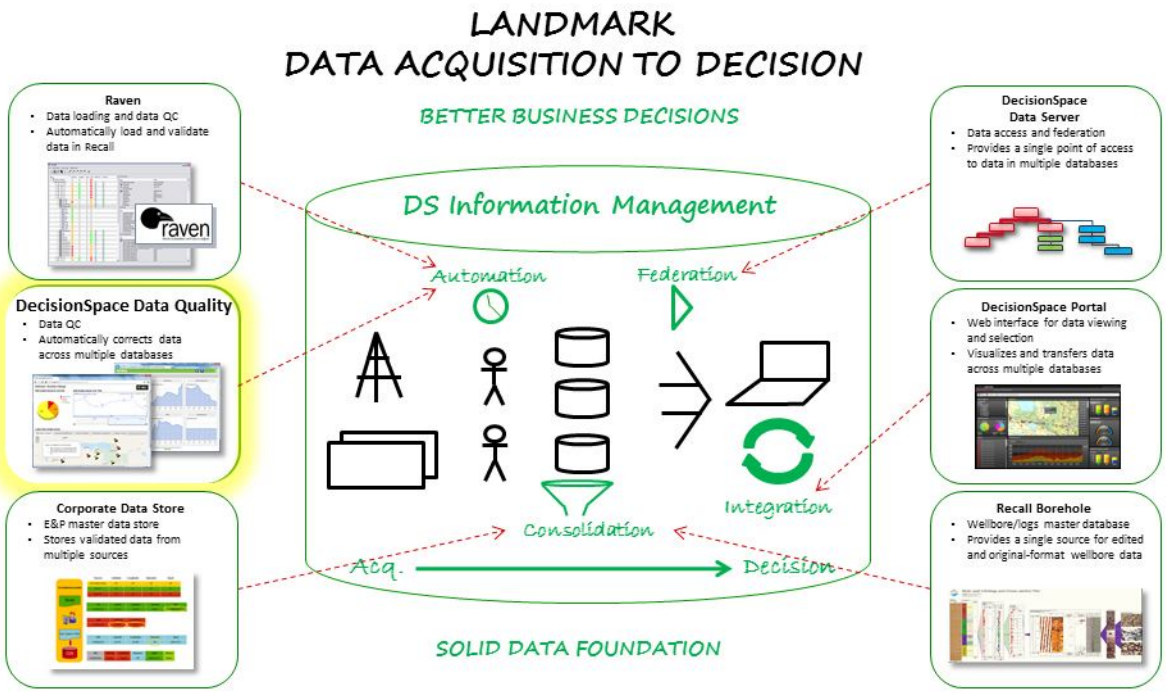
DecisionSpace Data Quality Software: Introduction

The growing number of new data types from complex resources plays coupled with a lack of skilled data management experts to ensure availability of quality data is consistently driving E&P companies to automate their data quality management efforts. To create long term, sustainable value, the upstream E&P industry is looking for a solution that:

- Addresses data quality issues across disparate data stores
- Promptly identifies and corrects data quality issues helping in decreasing the risk associated with bad quality data
- Resolves data quality issues with reusable, repeatable and automated quality assurance processes

The DecisionSpace Data Quality software helps you deal with this avalanche of data by offering data quality tools designed to evaluate, correlate, correct, monitor and synchronize data across the enterprise. Thus, enabling you to:

- Quickly assess the health of your data
- Automate and schedule data quality jobs for perpetual monitoring and continuous data improvement
- Removing data bottlenecks that hinder a project's progress
- Communicate data quality improvements over time to management and end users



Users of the Data Quality Application

The Data Quality application is intended for users who need help with:

- Data profiling prior to consolidations and clean-ups. Includes data types such as seismic, well, drilling, land and production
- Ongoing data quality management
- Data assessments conducted post-merger and acquisition activities
- Creating a business case to justify data quality and ROI to senior management
- Confirming operated/non-operated interest wells against large public repositories
- Developing and maintaining a cross-master well list across data stores and unique well identifier (UWI) changes
- Identifying and correcting duplicate records during the merger of two companies

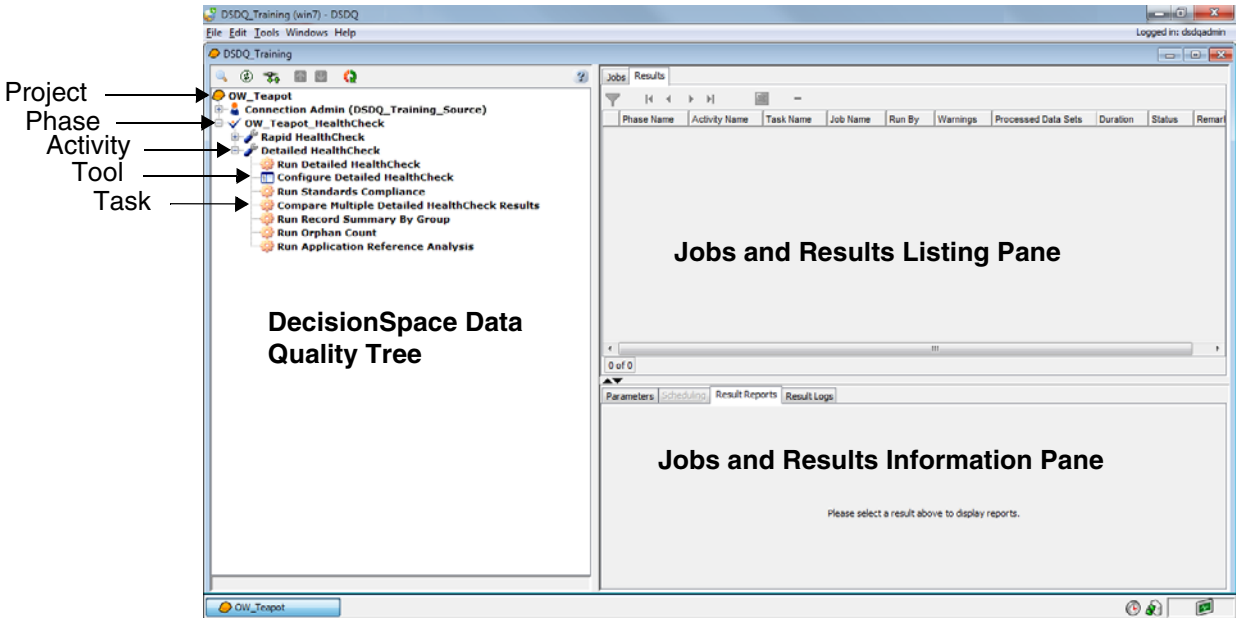
- Matching well data, land records, facilities information and production data in one “view”
- Maintaining corporate standards for reference purposes

DecisionSpace Data Quality Project Window

DecisionSpace Data Quality is a straightforward, yet powerful tool for managing data quality. Part of the ease of use of this application is that all of the work to be performed is controlled from one window, the DecisionSpace Data Quality Project Window. This window allows you to maneuver through the numerous functions of the application. It is divided into 3 panes:

- The DecisionSpace Data Quality Tree
- The Jobs and Results Listings Pane
- The Jobs and Results Information Pane

The Project Window can be maximized, minimized, or resized according to you preference.



DecisionSpace Data Quality Components

Before using the Data Quality application, it is important to understand the following components. A brief description of each component is given in the table below:

Component	Description
Project	A project contains all the Phases, Activities, Tools, Tasks, Jobs, and Results for a defined area of work. This component works as a Master folder that holds all the files created by you and the application. A project is identified in the application by the hard hat icon.
Phase	A Phase is composed of a set of Activities that are designed to achieve a particular set of objectives, which in turn contains Tools and/or Tasks. Phases are added to the project by you as the Project is created. An icon is specified in the application for each type of Phase (i.e. HealthCheck Phase, Clean Phase, Match Phase and Merge Phase).
Activity	An Activity is a group of Tools and/or Tasks that are steps toward reaching a specific goal. When you add a Phase to a Project, the Activities are saved to the Project. Activities do not require any action on your part, but instead hold all the Tools and Tasks that you may perform. An activity is identified in the application by a wrench icon.
Tool	A Tool can be thought of as a mini-application. When a Tool is activated, an application window opens, allowing for data manipulation and/or configuration. The information supplied by the user is then used within the application. A Tool is identified in the application by the window icon.
Task	The purpose of a Task is to create jobs, provide reports on the data to the user, and/or to set-up the data for a specific processing operation. A Job in a Task may be run after it has been configured via the relevant Tool. The user creates a job with specific parameters within the Task before running it. A Task is identified in the application by the gear icon.
Job	Jobs are tasks with specific Parameters. These specifications can make jobs unique and generate different results. An existing job can be run by selecting it and clicking on the **Run Selected Job** button in the Jobs toolbar. A new job can be created, and saved and run at the same time by clicking on the **Run Job** button in the Parameters toolbar.
View	View is a window that is an interactive display of results generated by a task. Views are accessed in the **Result Reports** tab in the Results Information pane and are launched by clicking the **Basic View Frame** icon once a result is selected.

Component	Description
Submodel	A Submodel is a core data structure within the Data Quality application that is used by all Activities. Created in **Perform Table Modeling**, a Submodel is a grouping of data tables within a database, identified by the three circle icon.
Service Level	A Service Level is a grouping of Data Quality rules that can be configured for specific tasks, projects, or even intended for a single submodel. Identified by the book icon, Service Levels are used directly with Submodels to set up and run rules against the Application Model.

At the project window level, you can perform an action either by right-clicking an item or by using the toolbar. At the DecisionSpace Data Quality Tree level, there are some actions that can be performed by right-clicking the component to display the pop-up menu. Double-clicking a tool or task opens the tool or initiates the process of job creation for a task.

DecisionSpace Data Quality Installation

This section outlines the requirements and procedures for installing DecisionSpace Data Quality on a server system and deploying it to workstations.

Hardware Requirements

The table below describes the minimum and recomended hardware configurations for running the DecisionSpace Data Quality application:

Workstation	Recommendation
Processor and Memory	• 64-bit CPU (AMD64 or Intel 64) with 12 GB of RAM. (Recommended) • 64-bit CPU (AMD64 or Intel 64) with 6 GB of RAM. (Minimum)
Hard Drive Space	• Free space on the hard drive for application and TEMP directory files: 10 GB
Windows operating systems	• Windows 7 (64-bit)

Workstation	Recommendation
Server	
Processor and Memory	• 64-bit CPU (AMD64 or Intel 64) with 12 GB of RAM. (Recommended) • 64-bit CPU (AMD64 or Intel 64) with 6 GB of RAM. (Minimum)
Hard Drive Space	• Free space on the hard drive for application and TEMP directory files: 20 GB
Windows operating systems	• Server 2008 (Certified) (64-bit)
Database	
Application data storage of projects and results can be held in either:	• PostgreSQL: Bundled with DecisionSpace® Data Quality installer, requires no database administrator configuration • Oracle 10/11: Storage and schemas configured by your database administrator • Microsoft SQL Server 2008/2012: Storage and schemas configured by your database administrator
Storage space required depends on size and amount of data sources connected to by DecisionSpace Data Quality	• A good guideline is to allocate 100 GB of tablespace for production systems initially, with room to grow as more projects and connections are added.

DecisionSpace Data Quality Server offers a web dashboard, web deployment of DecisionSpace Data Quality desktop client, and runs scheduled jobs in unattended mode.

Application Database Storage

A new feature in DecisionSpace 5000.10.3.0 is automatic data storage using the Postgres database. During the DecisionSpace Data Quality installation, you only need to select Postgres Database Server from the list of components to install.

Postgres automatically installs and is configured within the main DecisionSpace Data Quality installation's 'database' folder. Optionally, during installation, one can specify a different folder to hold Postgres data. For example, this is useful if the installation of the DecisonSpace Data Quality application occurs on the C:\ drive, while the much larger

D: drive should be used for DecisionSpace Data Quality application data storage.

> **Note**
>
> **Automatic Storage Using Postgres**:
>
> If automatic data storage using Postgres is selected, then there is no need to manually configure application storage in Oracle or MS SQL Server as described in the next two sections.

Alternatively, for those who prefer to manage DecisionSpace Data Quality Application data storage manually, database administrators will need to create a minimum of three empty data schemas in either Oracle or MS SQL Server:

- **DSDQ_MASTER** is the DecisionSpace Data Quality application configuration repository and requires at least **10 GB** of table space.

- **DSDQ_REFERENCE** holds the standards tables that Landmark provides with DecisionSpace Data Quality that can be used as a reference when cleaning the user database. These references can also be modified by the user. It is recommended that you make available at least **1 GB** of table space.

- Workspace schemas, often named **DSDQ_TEST_WKSP**, are used to capture the project results. DecisionSpace Data Quality software creates the necessary tables in this workspace schema to capture results. The user has the option to keep the results from each of the tasks, so this database can grow quite large. It is recommended to make available at least **25 GB** of table space for each workspace; however, **additional space may be required** depending on the data retention policies set by DecisionSpace Data Quality users. A separate workspace should be created for every new DecisionSpace Data Quality project.

The names above are a guideline only. You may use any database user names to match your local IT policies. You will be prompted for the names you use in the DecisionSpace Data Quality Setup Wizard and Desktop applications.

Application Licensing

DecisionSpace Data Quality requires an existing Landmark License Application Manager (LAM) to be configured. For more information about setting up an LAM, please refer to the files and documentation found on the Landmark Software Manager site.

Once the LAM is configured, all systems running DecisionSpace Data Quality require a **LM_LICENSE_FILE** environment variable to be configured. This environment variable will direct DecisionSpace Data Quality to the location of the LAM license file. The **LM_LICENSE_FILE** environment variable should point to the LAM server, using the configured port, which is 2013 by default.

To configure a new environment variable:

1. Select **Start > Control Panel** from the Windows taskbar and then Systems and Security from the Control Panel window.

2. Select System.

3. Click the **Advanced system settings** option under **Control Panel Home**.
   The **System Properties** dialog box appears with the **Advanced** tab selected by default.

4. Click the **Environment Variables…** button.

5. Click the **New…** button on the **System Variables** group box.

6. For **Variable name**, enter LM_LICENSE_FILE.

   For **Variable value**, enter a value according to the configured LAM Server.

   For example, **2013@LocalServer**, where **2013** is the LAM configured port, and **LocalServer** is the server name that the LAM is configured on.

Installation Steps

In order to complete the installation of DecisionSpace Data Quality, you will need to:

- Set up Database Storage on either Oracle or MS SQL Server (For this exercise, we will use Oracle 11 database)

- Install DSDQ on a server system, alongside an existing DecisionSpace Data Server installation
- Deploy DSDQ to workstations, either automatically or manually

Setting up Database Storage for Oracle 10/11

To setup application database storage for Oracle, you will need to create new **DSDQ_MASTER**, **DSDQ_<PROJ>_WKSP** and **DSDQ_REFERENCE** users and grant access. A sample script is provided in the *docs/sql/directory* for Oracle databases.

1. Log in to the Oracle database as an administrator to ensure that you have the permission required to create new users (using a tool like SQLPlus).

2. From the command prompt, execute the following command in order to connect to Oracle:

```
sqlplus /nolog
```

3. From sqlplus, execute the following command to connect as sysdba:

```
SQL>connect / as sysdba;
```

4. Execute the following command to create the table space:

```
CREATE TABLESPACE DSDQ DATAFILE
'c:\app\student\oradata\DSDQ\dsdq.dbf' SIZE 1024M
AUTOEXTEND ON MAXSIZE unlimited EXTENT MANAGEMENT
LOCAL UNIFORM SIZE 4M ONLINE;
```

5. Execute the following commands to create the DSDQ_MASTER , DSDQ_REFERENCE and DSDQ_ WORKSPACE users and grant access:

```
SQL> create user DSDQ_MASTER identified by landmark
default tablespace DSDQ temporary tablespace temp
account unlock;

SQL> create user DSDQ_REFERENCE identified by landmark
default tablespace DSDQ temporary tablespace temp
account unlock;

SQL> create user DSDQ_TEST_WKSP identified by landmark
default tablespace DSDQ temporary tablespace temp
account unlock;
```

```
SQL> grant connect, resource to DSDQ_MASTER;

SQL> grant connect, resource to DSDQ_REFERENCE;

SQL> grant connect, resource to DSDQ_TEST_WKSP;

SQL> exit
```

Note

It is necessary to set all options by default in each step for this exercise.

Setting up Database Storage for MS SQL 2008/2012

Note

This procedure is for your reference. For the purpose of this training, please skip this section.

To setup application database storage for MS SQL, you will need to create new DSDQ_MASTER, DSDQ_TEST_WKSP and DSDQ_REFERENCE users and grant access.

1. Log in to MS SQL as an administrator to ensure that you have the permission required to create new users (using a tool like SQL Server Management Studio).

2. Right-click on the Databases folder and select New Database… from the drop-down menu.
   The New Databases window appears.

3. Specify a Database Name and Owner for each user that you create.

4. Repeat steps 2 and 3 to create all three users.

Note

The Owner will be associated with the User parameter in the Create Connections window within DecisionSpace Data Quality.

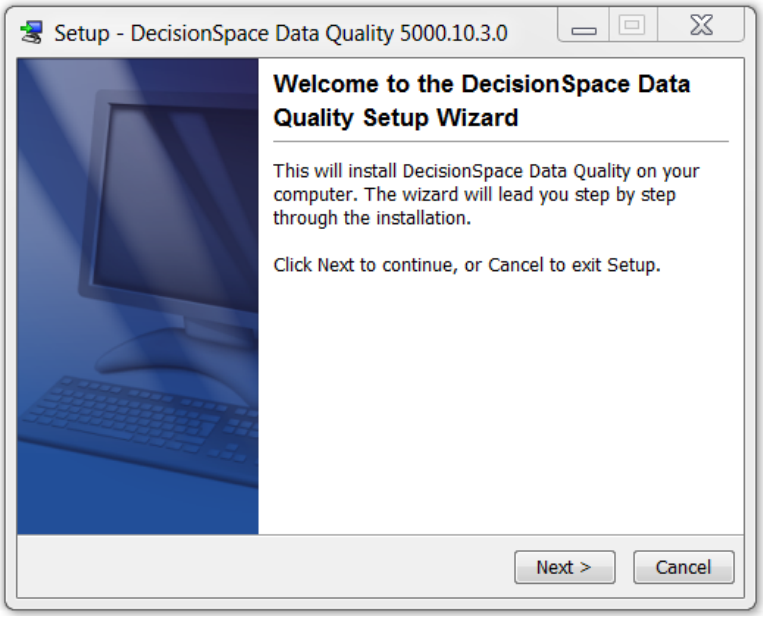
Installing DecisionSpace Data Quality on a Server System

There are two types of DecisionSpace Data Quality installers (Windows example provided):

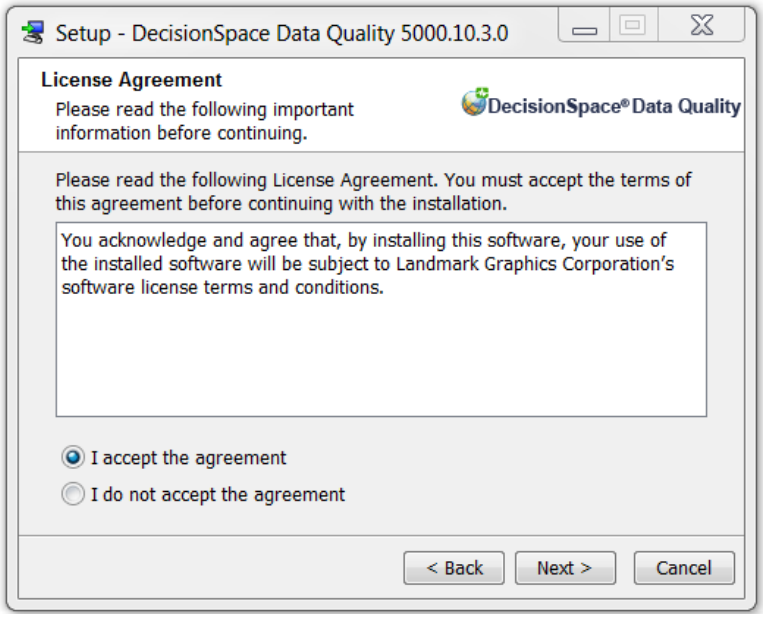
- DSDQ_Server_Win64_5000_10_3_0.exe: 64-bit installer for 64-bit Windows operating systems (for server installation).

- DSDQ_Client_Win64_5000_10_3_0.exe: 64-bit installer for 64-bit Windows operating systems (for client installation).

1. Double-click the provided installer to launch the **DecisionSpace Data Quality Setup Wizard**.
   The DecisionSpace Data Quality Setup Wizard screen appears.

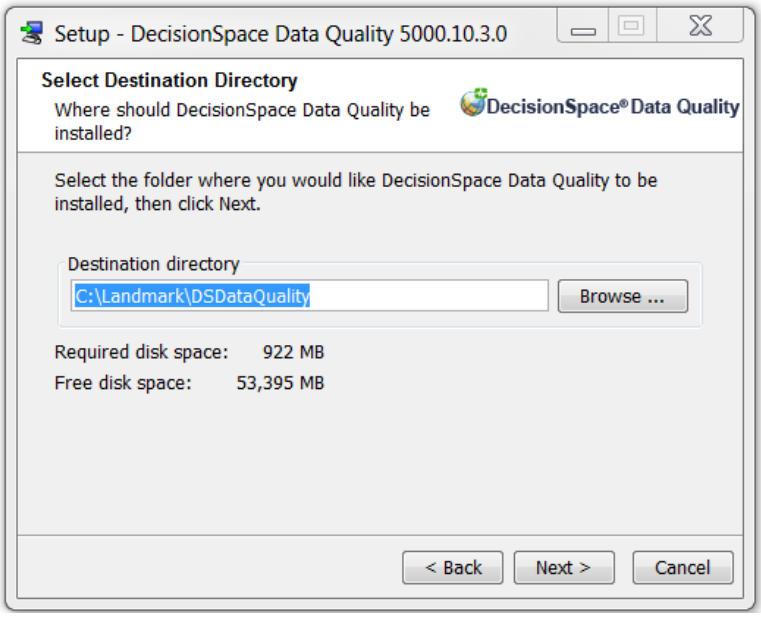


2. Click **Next** to continue.



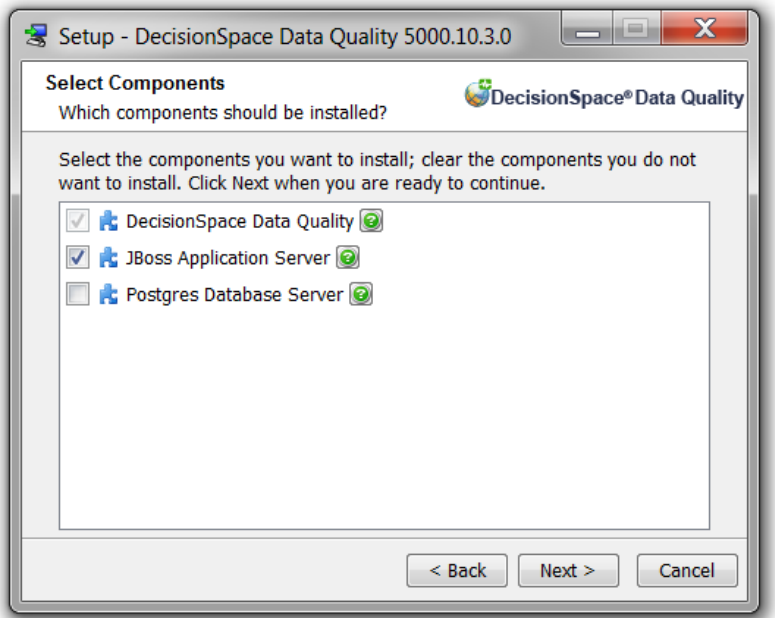
3. Review the software license agreement information given in the **License Agreement** screen and select the **I accept the agreement** option.

4. Click **Next** to continue.



5. Select the destination directory where you would like DSDQ to be installed by clicking the **Browse…** button. If the specified directory does not exist, it will be created.

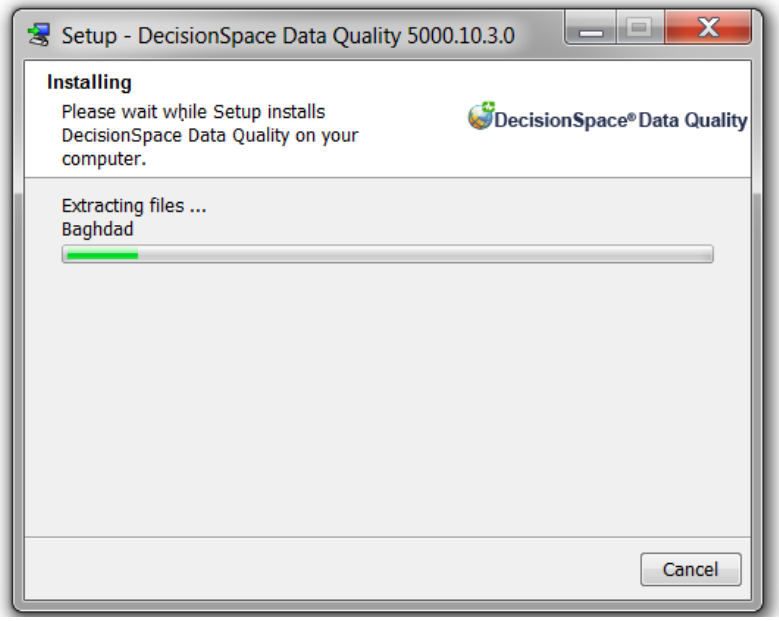
6. Click **Next** to continue.



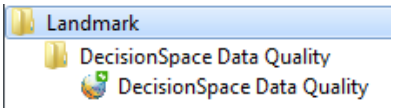
7. Select the following component:

   - **DecisionSpace Data Quality**
   - **JBoss Application Server**

8. Click **Next** to install DecisionSpace Data Quality on your computer.



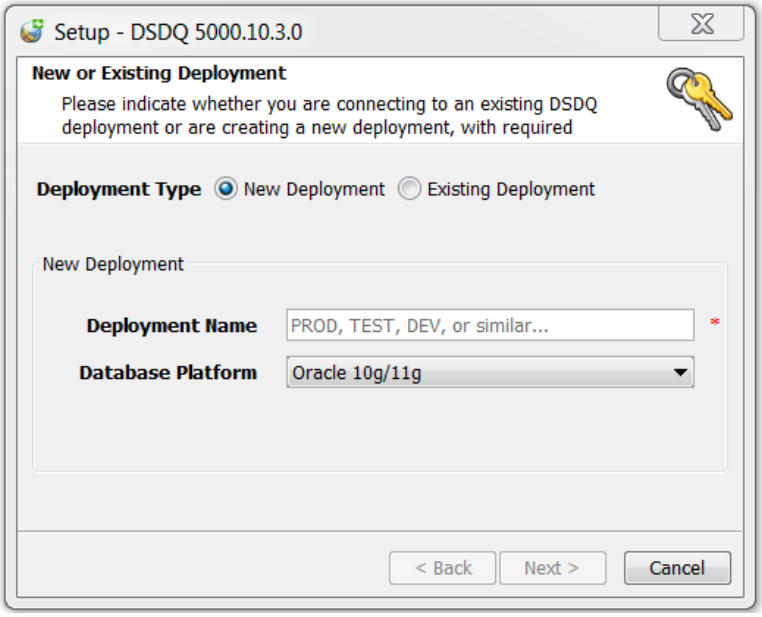
9. Once the installation process is complete, the setup process will begin.
   An application shortcut will be automatically created on your desktop along with a *DecisionSpace Data Quality* program group under *Landmark* in the Windows Start Menu.



10. Continue with the application setup process by entering the following information in the **New or Existing Deployment** screen:

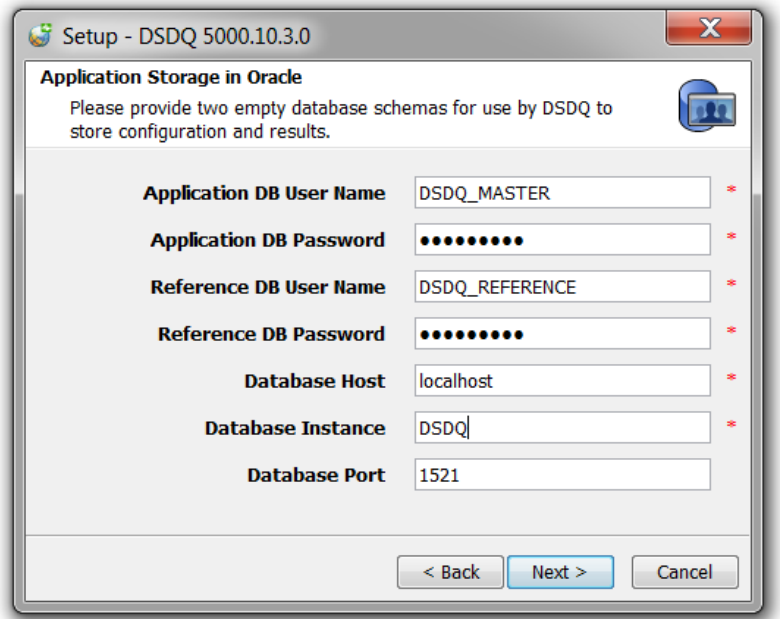
    - Select **New Deployment** for **Deployment Type**.
    - Enter a name in the **Deployment Name** field.

- Select **Oracle 10g/11g** from the **Database Platform** drop-down list.



11. Click **Next** to continue.

12. Enter the **User Name** and **Password** to connect to the DSDQ_MASTER and DSDQ_REFERENCE databases and then the **Data Host** and **Instance** information (modify the **Database Port** if required).

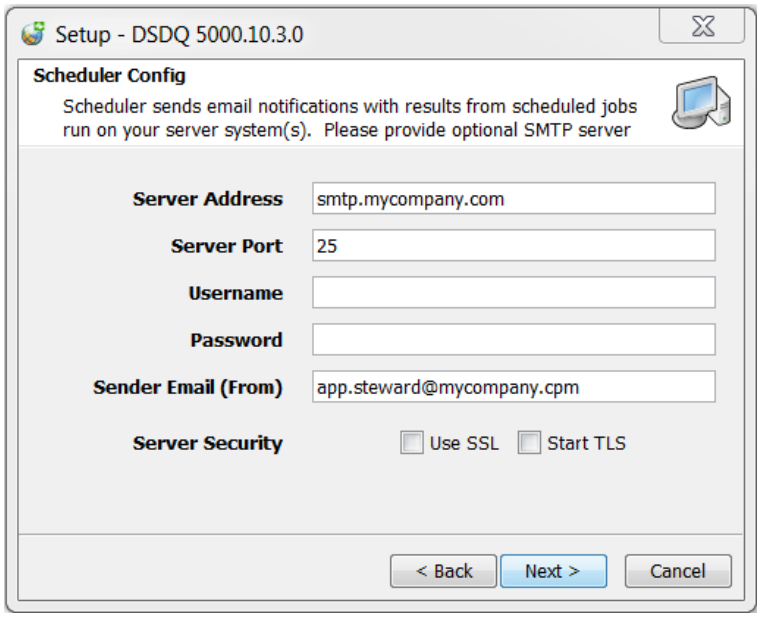


13. Click **Next** to continue.
    The application will attempt to connect to the databases. If this attempt is successful, the specified databases will be installed and the setup process will continue.

> **Note**
>
> There must be empty database schemas created solely for the use of DecisionSpace Data Quality.

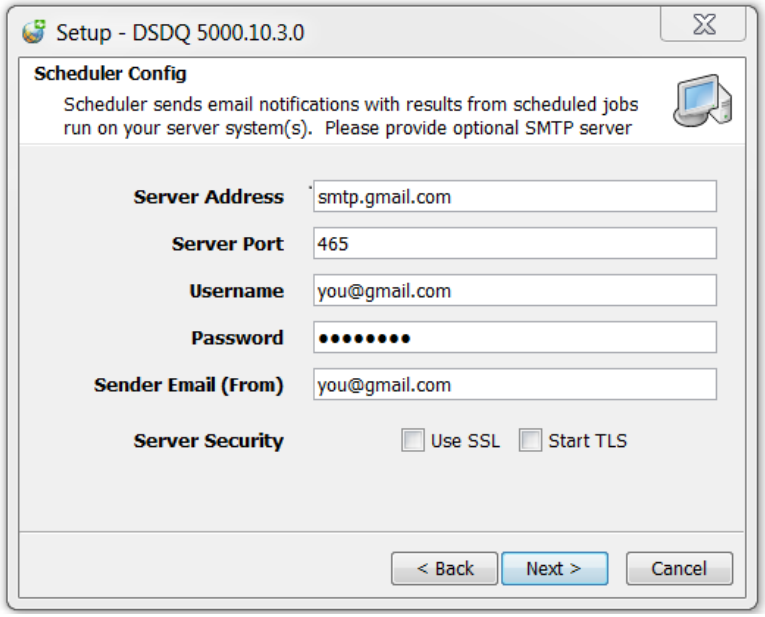
14. Enter the **Server Address** and **Server Port** if you have setup the DSDQ Server package and you are planning to schedule jobs to run and wish to receive email notifications regarding the scheduled jobs.

15. Specify the email address or email alias of your local DecisionSpace Data Quality application steward in the **Sender Email (From)** field.
    This address will be used in the From: field of email notifications sent to users in order to control who they reply to with their questions.



16. Optionally, if your mail server requires authentication in order to send email, please enter the SMTP **Username** and **Password**.

17. If your mail server requires additional connection security, select the **Use SSL** and **Start TLS** options as appropriate.
    Shown below are some example settings for sending to a secure SMTP server.

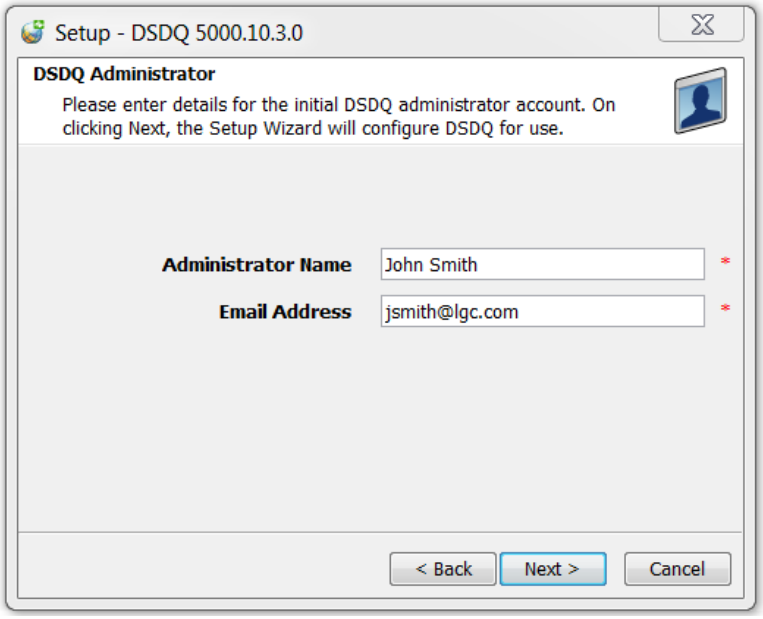


18. Leave all options as default and click **Next** to continue.
    The Setup Wizard attempts to connect to the Mail Server. The **DSDQ Administrator** window appears.

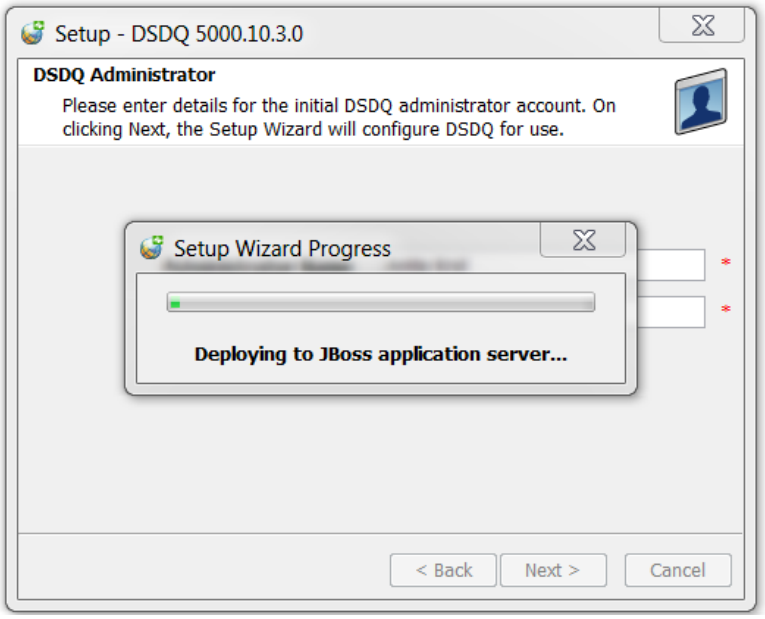
19. Create a DecisionSpace Data Quality login account by specifying an **Administrator Name** and **Email Address** in the **DSDQ Administrator** window.

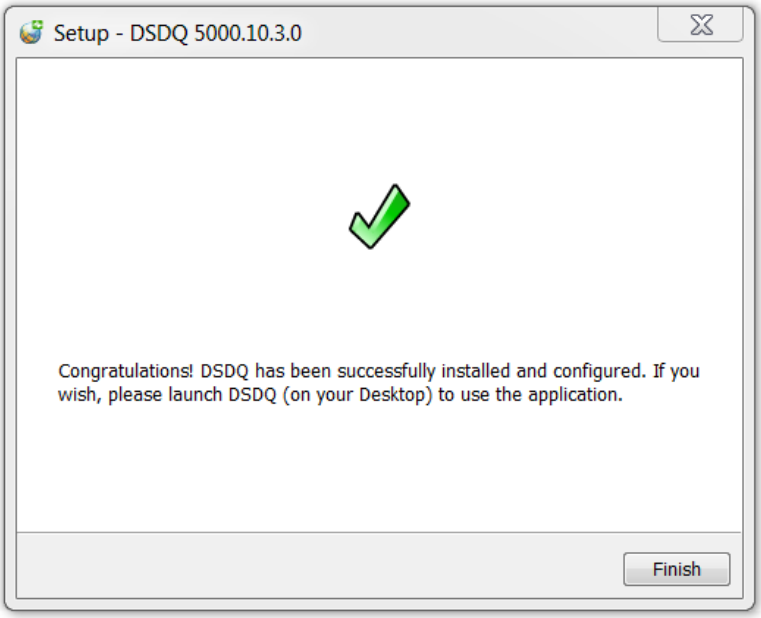
> **Note**
>
> This initial setup will allow the application to email information about scheduled jobs to specified users.

The system administrator account will be created and used for logging into DSDQ.



20. Click **Next** to continue.
    The System Account will be validated, DSDQ_MASTER and DSDQ_REFERENCE data will be loaded and the application will be registered. Please be patient as this step may take a few minutes.





21. Click **Finish** to complete the DecisionSpace Data Quality setup process.

Deploying DecisionSpace Data Quality to Workstations - Automatic Deployment

DecisionSpace Data Quality automatic deployment allows you to launch the DecisionSpace Data Quality Client application from a web page hosted on your intranet, your desktop or from the Start menu. As a pre-requisite to automatic deployment, your computer must have Java version 7 or higher installed on it. For information on installing Java, please visit www.java.com.

The primary benefit of automatic deployment is that updates applied to the DecisionSpace Data Quality Server system are automatically provided to workstation users the next time they run DecisionSpace Data Quality Client.

To automatically deploy DecisionSpace Data Quality:

1. Ensure that the Server installation has been completed successfully.

2. Enter the web server address and port in the address bar of your web browser. E.g. http://ServerMachine:8091

3. Click on the icon button to launch the **DecisionSpace Data Quality Desktop Client** on the desktop.
   The **Opening dsdq.jnlp** window appears.

4. Select **Open with Java™** Web Start Launcher (default) and click **OK**.
   The DSDQ application will be launched and a shortcut will be created in the **All Programs** list on the Start menu.

Deploying DecisionSpace Data Quality to Workstations - Manual Deployment

This option is for IT departments who wish to script their own deployment of DecisionSpace Data Quality Client for their users. In this scenario, subsequent DecisionSpace Data Quality Server updates are not automatically pushed to workstations. Therefore, system administrators are responsible for keeping all deployed clients up-to-date. Manual deployment of DecisionSpace Data Quality includes the correct version of Java, negating the need to install Java 7 or later.

To manually deploy DecisionSpace Data Quality:

1. Ensure that the Server installation has been completed successfully.

2. Install the DecisionSpace Data Quality application on the client machine.

3. Cancel the DecisionSpace Data Quality Setup: License Key window that will automatically open up.

4. Copy and paste the dvconnections file (dvconnections.xml) from the server to the "..*local DSDQ install\conf\DataVera*" directory.

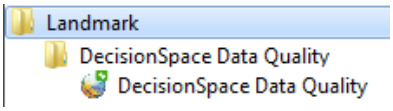
5. Run the DSDQ executable to start the application.

Starting the Data Quality Application

When the Data Quality application is installed on your computer, a shortcut to the application may be placed on your desktop. The relevant program group, Data Quality, also gets set up under the Windows Start menu giving you access to the Data Quality application.

To start the Data Quality application:

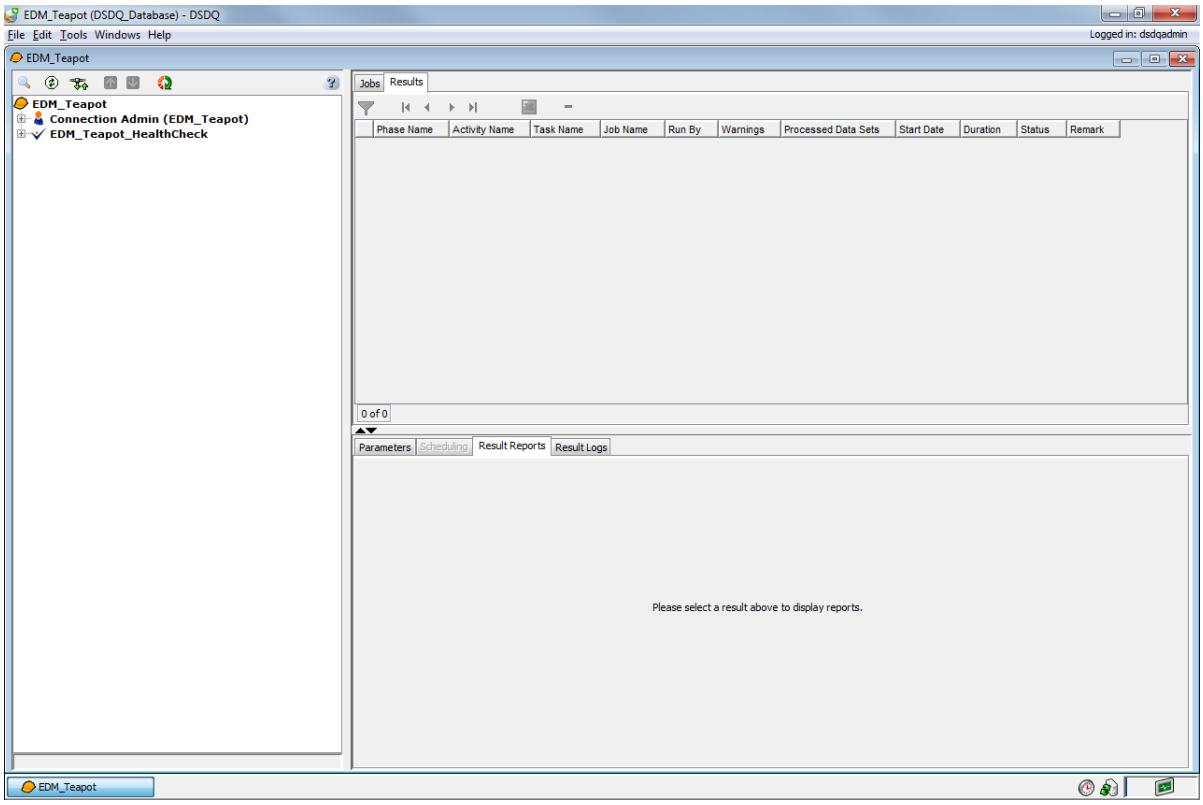
1. Double-click the Data Quality icon on the desktop or select **Start > All Programs > Landmark > DecisionSpace Data Quality > DecisionSpace Data Quality** from the Windows taskbar.



Note

If a shortcut is not created on your desktop already, you can create one by right-clicking the DSDQ.exe file and selecting Send to > Desktop (create shortcut) from the pop-up menu. The DSDQ.exe file is located in the 'local installed folder' directory.

The Data Quality **Project Window** displays.



To Start DSDQ from a Shared Network Drive

If the Data Quality application is set up on a shared network drive, you can access it from your machine (i.e., the client machine) in the following way:

1. Browse to the network drive that the Data Quality application is installed on. (Consult your system administrator if you do not know the location.)

2. Select the network path where '..\Installed DecisionSpace Data Quality' directory.

3. Right-click on the DSDQ.exe file and select **Create Shortcut** to create a shortcut and dragging the DSDQ shortcut to local desktop.

For Assistance

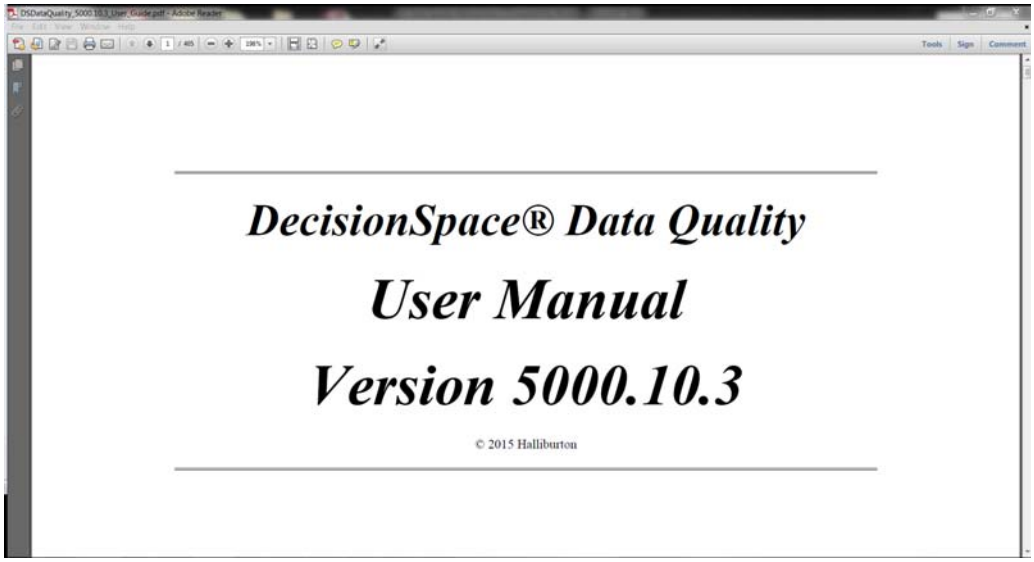
The Data Quality software comes with a user guide that provides an overview of the Data Quality software as well as step-by-step procedures for using the application. For additional explanation from one of Landmark's support specialists, call or e-mail the Landmark help desk.

Using the User Guide

There are multiple ways to access the Data Quality User Guide:

Click the **Help** button wherever available for help. This brings up more information for the relevant window via the user guide.

- Access the User Guide by selecting **View Help** from the **Help** menu.



Contacting Landmark Customer Support

Landmark software operates Technical Assistance Centers (TACs) in Australia, the United Kingdom, and the United States. Additional support is also provided through regional support offices around the world.

- **Support via Web Portal**
- **Technical Assistance Centers**
- **Regional Offices**

Support via Web Portal

Support information is always available on the Landmark Customer Support internet page. You can also submit a support request directly to Landmark Customer Support though the Landmark Customer Support Portal:

http://www.landmarksoftware.com/Pages/ContactSupport.aspx

To request support in the Landmark Customer Support Portal:

1. In the **PIN** and **Password** text boxes in the Please Sign In area, enter your registered personal identification number and password.
2. Click the **Sign In** button.
3. In the Case & Defect Information area, click the **Create a New Case** link.
4. In the **Create Case** area, fill in the necessary information. Provide details about your technical concern, including any error messages, the workflow steps where the problem occurred, and attachments of screen shots that display the problem. To help understand the concern, you can also attach other files too, such as example data files.
5. Click the **Submit** button. A support analyst in the nearest Technical Assistance Center will respond to your request.

Technical Assistance Centers

Asia, Pacific	**61-8-9481-4488 (Perth, Australia)**
8:00 am - 5:00 pm Local Time Monday-Friday, excluding holidays	Toll Free 1-800-448-488 Fax: 61-8-9481-1580 Email: **apsupport@lgc.com**
Europe, Africa, Middle East	**44-1372-868686 (Leatherhead, UK)**
9:00 am - 5:30 pm Local Time Monday - Friday, excluding holidays	Fax: 44-1224-723260 (Aberdeen, UK) Fax: 44-1372-868601 (Leatherhead, UK) Email: **support@lgc.com**
Latin America	**713-839-3405 (Houston, TX, USA)**
(Spanish, Portuguese, English) 7:00 am - 5:00 pm Local Time	Fax: 713-839-3646 Email: **soporte@lgc.com**
North America	**713-839-2200 (Houston, TX, USA)**
7:30 am - 5:30 pm Central Standard Time Monday - Friday, excluding holidays	**Toll Free 1-877-435-7542** **(1-877-HELP-LGC)** Fax: 713-839-2168 Email: **support@lgc.com**

Regional Offices

For contact information for regional offices, see the Contact Support page located at:

http://css.lgc.com/InfoCenter/index?page=contact&section=contact

If problems cannot be resolved at the regional level, an escalation team is called to resolve your incidents quickly.

Chapter **2**

Connecting DecisionSpace Data Quality (DSDQ) with DecisionSpace Data Server (DSDS)

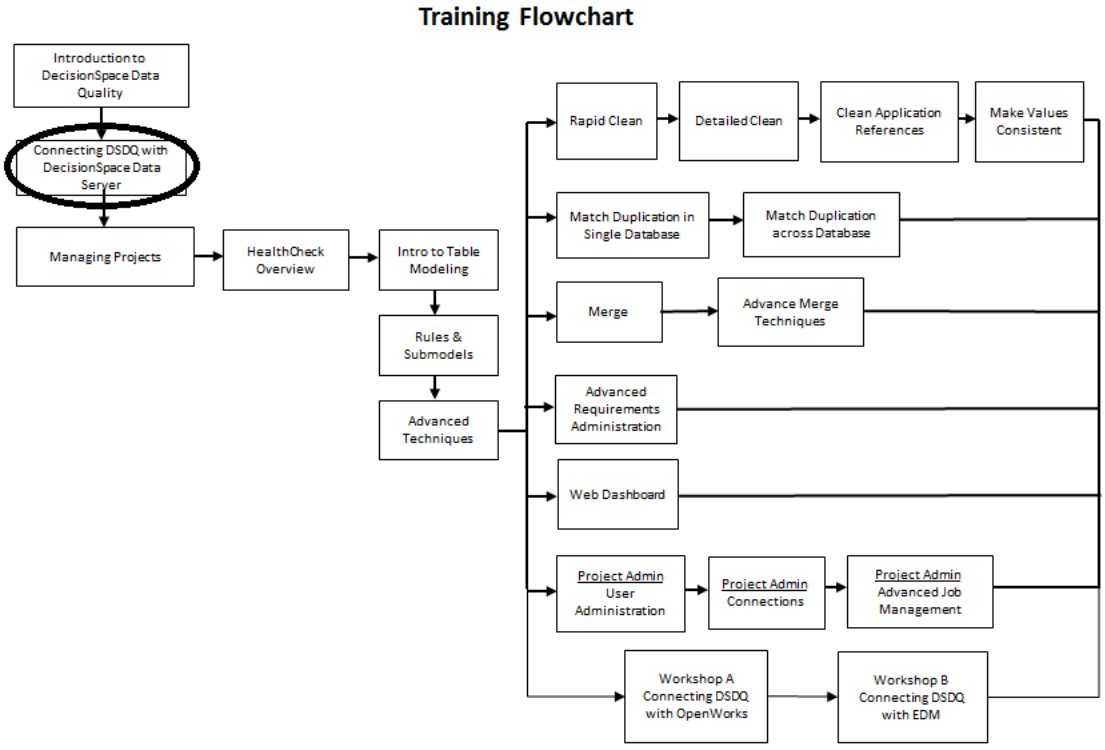
DecisionSpace Data Server allows the Data Quality application to connect to Landmark application databases, such as OpenWorks® and EDM™ and non-Landmark data sources. This enables the Data Quality application to automatically load preconfigured models and rules against these databases.

Chapter Overview

In this chapter, you will learn about:

- The DecisionSpace Data Server software including its advantages and key components
- Installing DecisionSpace Data Server
- Connecting DecisionSpace Data Quality (DSDQ) with DecisionSpace Data Server (DSDS)

Topics covered in each chapter are outlined in the following illustration. Those specific to the current chapter will be circled in black for your reference:

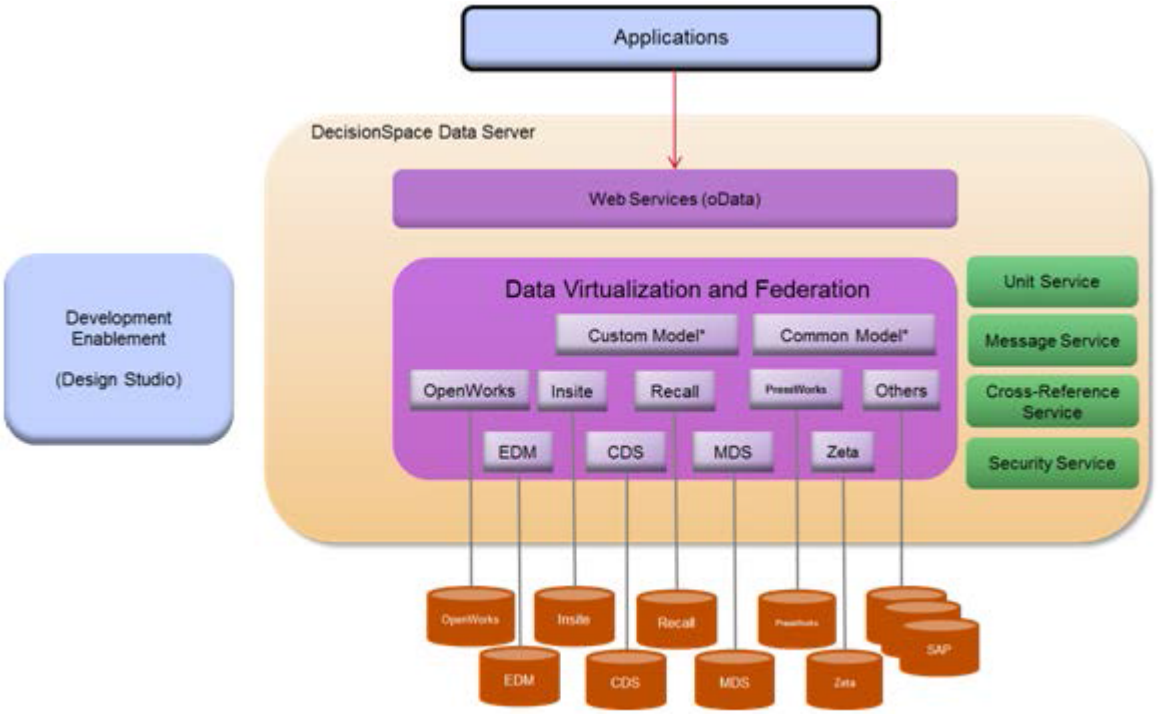


DecisionSpace Data Server: An Introduction

Decision Space Data Server provides applications with a common data access to Landmark and non-Landmark data sources. It enables access to data via services instead of individual development kits for different databases. DecisionSpace Data Server also provides tools for developers and consultants to create connections to additional data sources and expose the data as a service.

It is a simple, yet efficient tool that acts as a bridge between DecisionSpace Data Quality application and various data sources. The following features set DecisionSpace Data Server apart from other applications with similar functionality:

- Enterprise services for Exploration and Production data
- Data integration/virtualization from heterogeneous models and sources
- Data interoperability
- Open standards based web accessible data
- Integration with enterprise security for identity, authentication and authorization



Key components of DecisionSpace Data Server are:

- **JBoss Application Server**: JBoss is a widely used application server. It is cross-platform and open source software written in Java that implements standards from J2EE. JBoss also supports the LDAP login module. To configure single sign on (SSO) using LDAP authentication in DecisionSpace Data Server, refer to the DecisionSpace Data Server System Administration Guide.

- **OData4j**: The Open Data Protocol (OData) is a web protocol for querying and updating data that provides a way to unlock your data by applying and building upon web technologies such as HTTP, Atom Publishing Protocol (AtomPub) and JSON to provide access to information from a variety of applications, services, and stores. OData4j is a new open-source toolkit for building first-class OData producers and first-class OData consumers in Java.

- **Teiid**: Teiid is a data virtualization system that allows applications to use data from multiple, heterogeneous data stores. It is a query engine for joining and unioning data from multiple sources in an optimal manner.

- **OData Consumers**: OData consumers are the clients written in .Net/Java and other scripting languages to remotely access the data

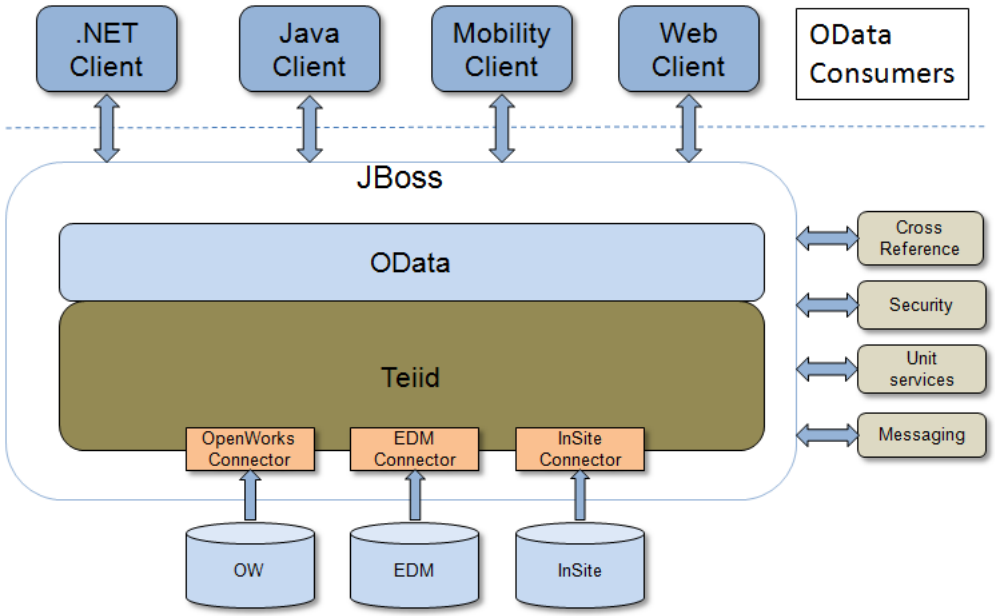
produced by the DecisionSpace data service. The data is published in two formats (ATOMPub and JSON) by service. Consumers can request data in either of the supported formats.

- **Security**: DecisionSpace Data Server can be configured to use several different authentication providers, including file-based, database-based, LDAP-based, or a combination of one or more of these. It can also be configured for single sign on with Active Directory-based authentication service. By default, a file-based security module is used.

- **Messaging**: DecisionSpace Data Server can publish transaction messages to any JMS based messaging broker.

- **Connectors**: Connectors are Teiid plugins that enable access to relational and non-relational data sources. Currently, DecisionSpace Data Server ships with connectors for the following data sources:

  - OpenWorks
  - EDM
  - Insite



DecisionSpace Data Server Layout

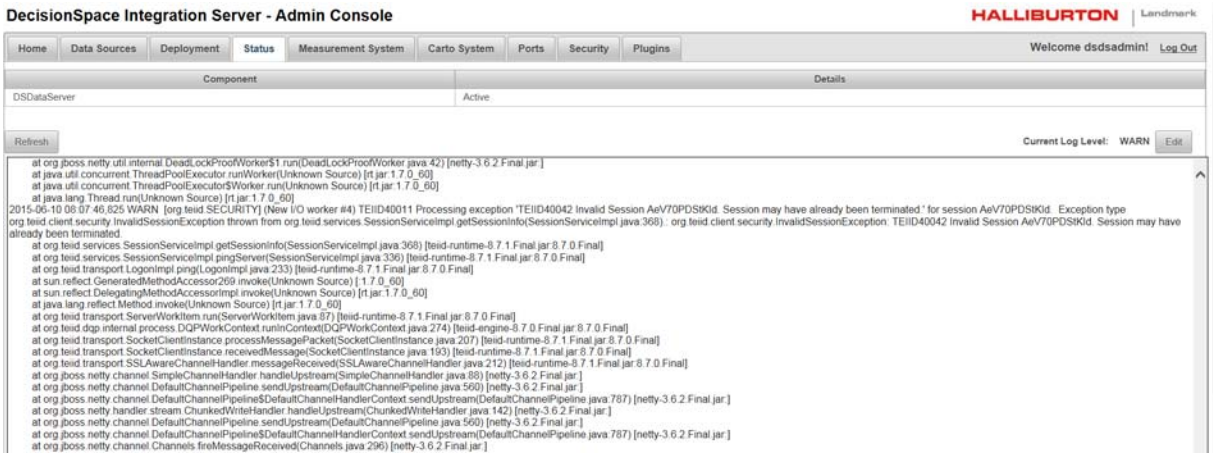
DecisionSpace Data Server Console has nine tabs, located on the top-left corner of the console window:

Home: This is a default tab. This tab displays summary information about the contents of the other tabs.

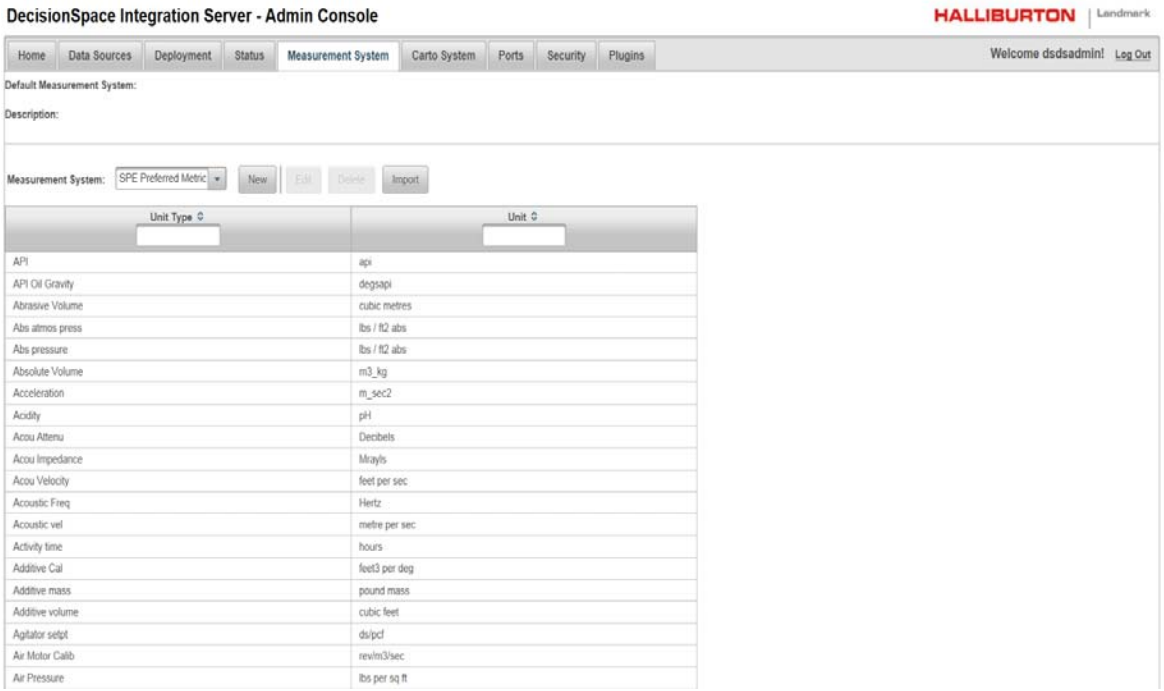
Data Sources: This tab allows you to create groups in order to manage data sources and then generate VDBs (virtual database connections). To the left you see a tree of default Data Sources and any new additions are listed here. We recommend that you organize the data sources by group.

Deployment: This tab displays a list of deployed VDBs and their related details and status. This tab also allows the deletion of previously created VDBs or the addition or removal of data sources to an existing VDB.

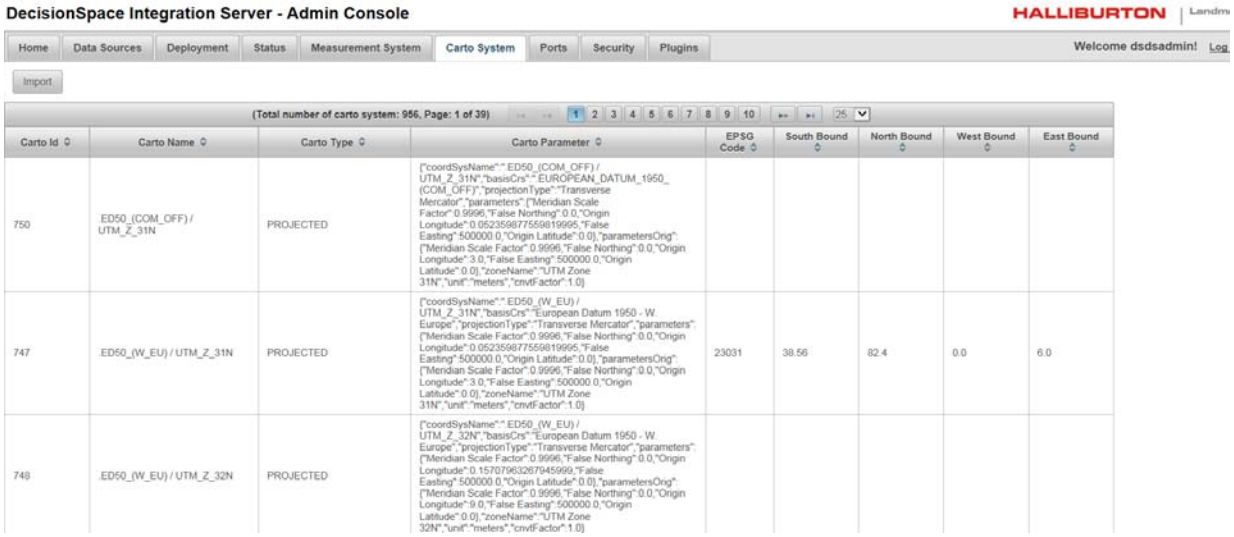
Status: This tab allows you to check the server.log file, helpful for troubleshooting issues.



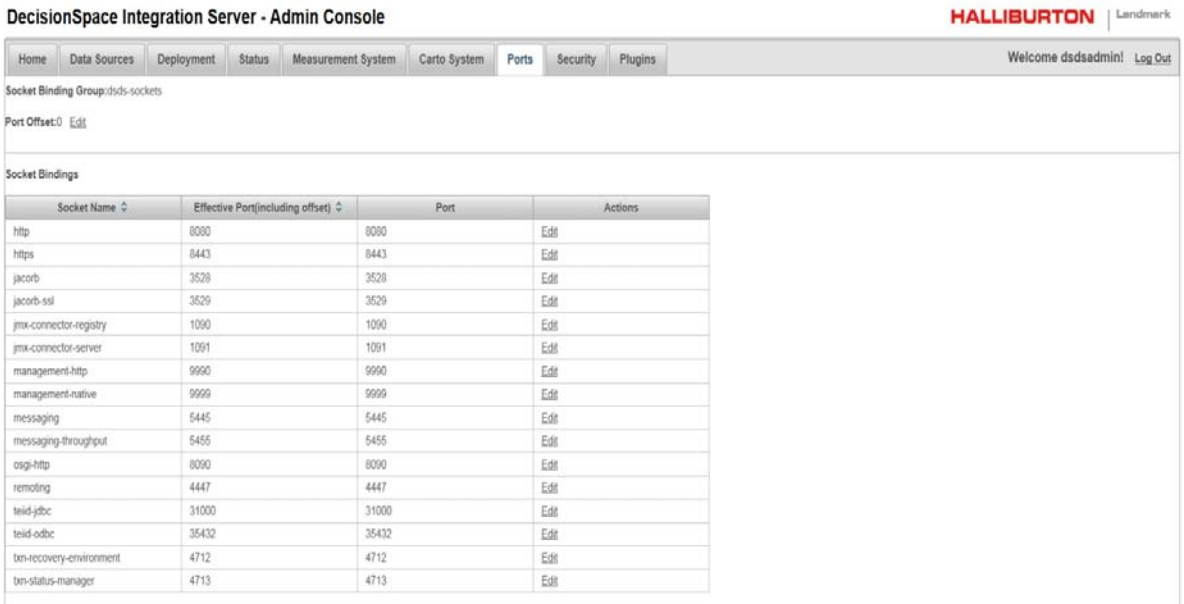
Measurement System: This tab shows the list of measurement systems with the associated unit types and units. This window allows a user to create custom measurement systems.



Carto System: This tab shows the list of cartographic systems. Allows you to import CRS definitions that do not exist in the Data Server's CRS service.

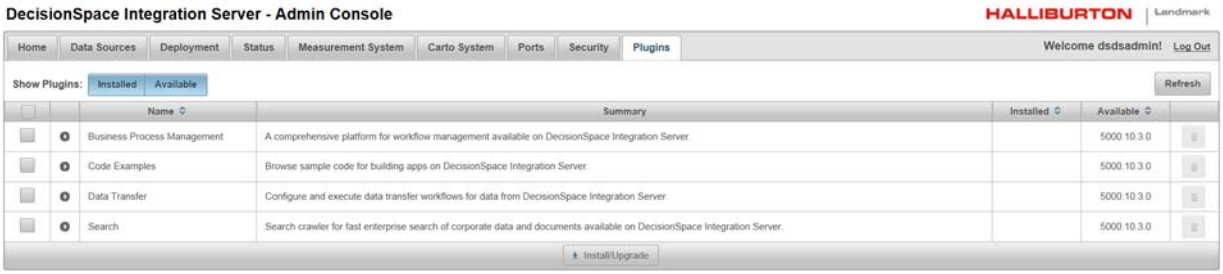


Ports: This tab shows the port settings used during the initial DecisionSpace Data Server installation. This window allows customizing the port settings, if necessary.



Security: This tab allows you to configure SSO and LDAP settings, manage users, assign roles, change user passwords.

Plugins: This tab shows the list of available and currently installed plugins. The Console also has a session time out, which causes the session to expire after 30 minutes of inactivity and directs the user back to the initial login page.



DecisionSpace Data Server Installation

This section outlines the requirements and procedures for installing DecisionSpace Data Server on a workstation.

System Requirements

Before you start installing DecisionSpace Data Server, please ensure that your workstation meets the following requirements:

Data Server	
Resource	**System Requirements**
Operating System	• Windows 2008 R2 Server x64 • Windows 7 x64 • RHEL AS 6 (Primary) • RHEL AS 5.x (Secondary)
Third-Party Software (packaged with Data Server)	• JBoss Enterprise Application Platform (EAP) ver. 6.0.1 Alpha 1 • Teiid 8.4.0 Final • OData 3.0 • JRE 1.6.0 64-bit • DecisionSpace Messaging (ActiveMQ 5.7.0) • DecisionSpace Data Server Design Studio (Teiid Designer 8.2) available using the Custom install option)
Landmark Software Dependencies	• LAM 5000.0.3 (Flex 11.7) • Licensed using FlexLM DSDATASERVER ver. 5000.8 which is read from the system environment

Data Source-specific	• OpenWorks Connector — Oracle Client 11.2.0.2 Administrative version 64-bit — OpenWorks 5000.8.3 Client • EDM Connector — EDM 5000.1.0 installed with 5000.1.10 Update • Insite Connector — InSite Server — JNBridge Pro 6.1.x (prerequisite: VC++ 2010 redist x86 and .Net 4.0)
Client[a]	
JAVA Client	• JAVA Client Binding (JRE 1.6 64-bit)
.NET Client	• .NET Client Binding (4.0)
OData Client	• OData v3.0
Browsers	
Browsers	• Internet Explorer 9 or higher
	• Firefox 20.x or higher
	• Google Chrome
General	
Minimum Java Heap Size	• 1024
Minimum Memory Size	• 2048

a. DecisionSpace Data Server users will write their own OData consumers, which are clients written in .Net/Java and other scripting languages to remotely access the data produced by the DecisionSpace data service.

Installing DecisionSpace Data Server

In this exercise, you will install DecisionSpace Data Server 5000.10.3.0 locally, on a Windows 7 x64 Platform. During the installation procedure, the following third-Party software will be installed:

- JBoss Enterprise Application Platform (EAP) ver. 6.0.1 Alpha 1
- Teiid 8.4.0 Final
- OData 3.0
- JRE 1.6.0 64-bit
- DecisionSpace Messaging (ActiveMQ 5.7.0)
- DecisionSpace Data Server Design Studio (Teiid Designer 8.2) (available using the Custom install option)

The following Landmark Software Dependencies must be verified in order to run DecisionSpace Data Server after installation:

- LAM 5000.0.3 (Flex 11.7)
- Licensed using FlexLM DSDATASERVER ver. 5000.8 (the instructor will tell you where the license file is located)

This procedure must be performed by a domain user with administrator rights on the computer where the application is to be installed. Prior to initiating installation, you must ensure that there is at least 1 GB of space available in the installation directory.

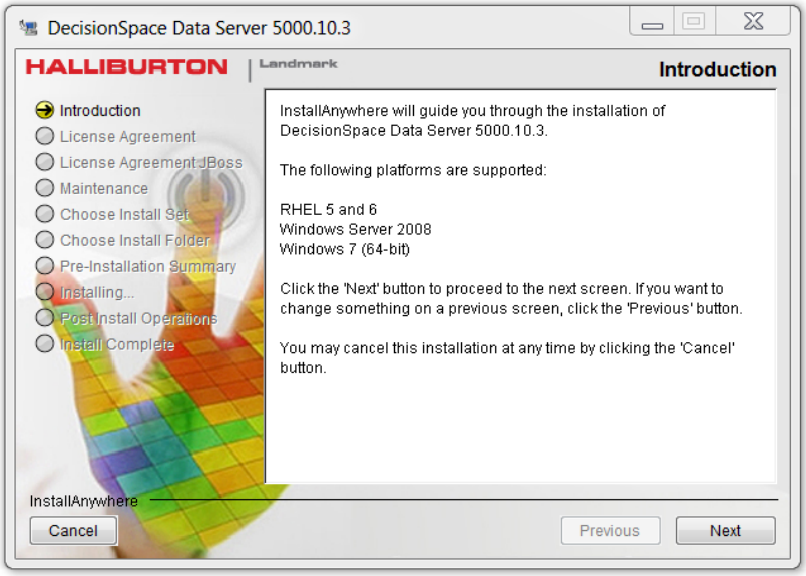
To install the DecisionSpace Data Server software:

1. Open a terminal session and navigate to the directory where the Data Server installer resides (For example: ***\Landmark\DSDataServer5000.10.3.0***).
2. Depending on the operating system on which you have to install the application, enter one of the following commands in the terminal window:
    - **On Windows:** DSDS_5000_10.3_Win.exe

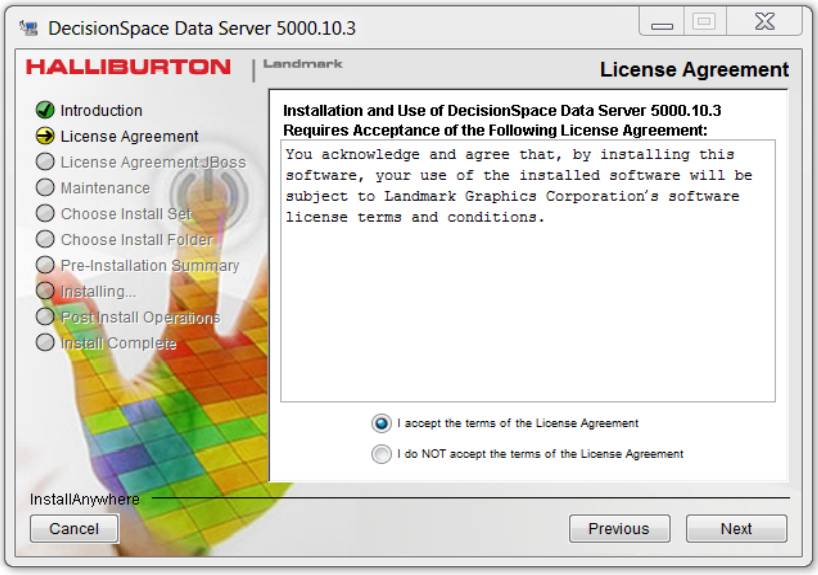
- **On Linux:** ./DSDS_5000_10.3_RHL.bin

> **Note**
>
> If you are installing DSDS on Linux, please ensure that Java 1.6.0 is loaded and that the $JAVA_HOME environment variable is set in your environment.

The installer launches and the **Introduction** screen appears.

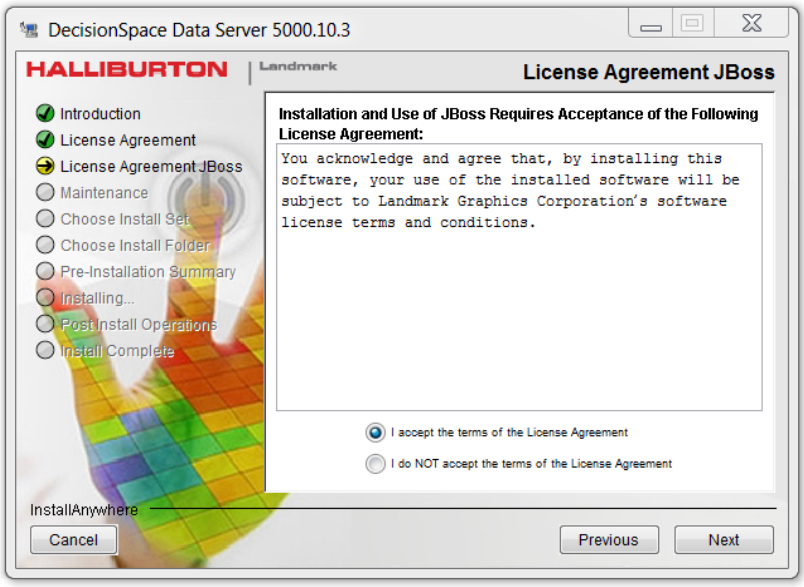


3. Review the information in the **Introduction** screen and click **Next** to continue.
   The **License Agreement** screen appears.



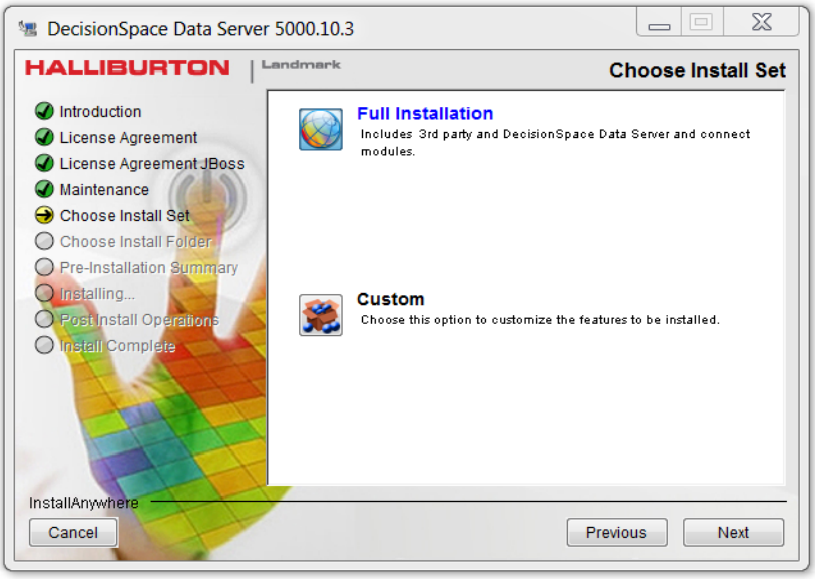
4. Review the software license agreement information given in the **License Agreement** screen and select **I accept the terms of the License Agreement** option.

5. Click **Next** to continue.
   The **License Agreement JBoss** screen appears.



6. Review the JBoss software license agreement information given in the **License Agreement JBoss** screen and select **I accept the terms of the License Agreement** option.

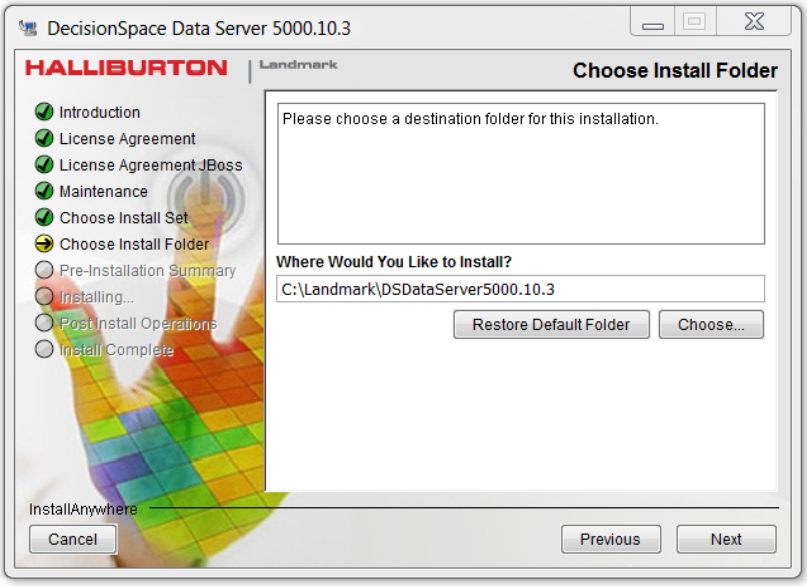
7. Click **Next** to continue.
   The **Choose Install Set** screen appears.



8. Select **Full Installation** to install all software modules required for running DecisionSpace Data Server. These include the Data Server, OpenWorks Connector, EDM Connector, InSite Connector, DecisionSpace Messaging, PowerHub Connector, and OPC-DA Connector.

> **Note**
>
> Custom installation is selected to install a subset of the components. Currently, the options include OpenWorks Connector, EDM Connector, InSite Connector, DecisionSpace Messaging, PowerHub/CDS Connector, Petrel Connector, Recall Connector, and DecisionSpace Data Server Design Studio. For more information, refer to the DSDS_5000.10.3_Release_InstallGuide.

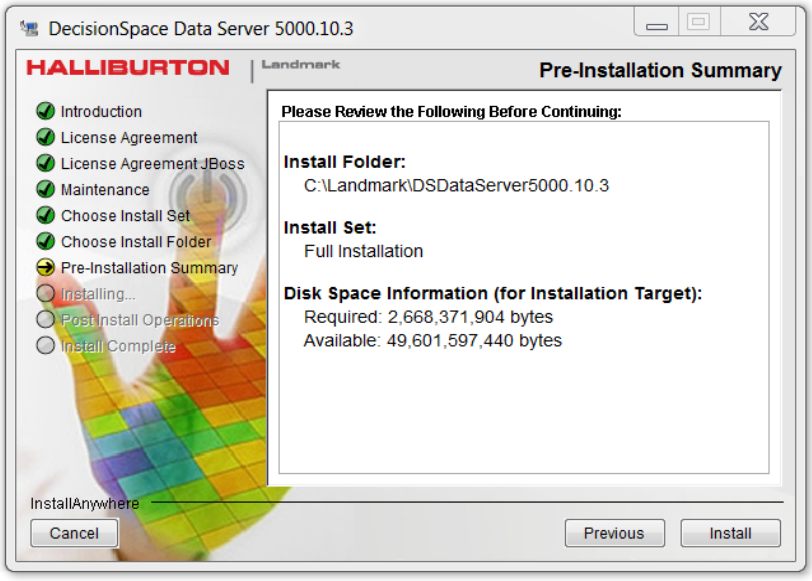
9. Click **Next** to continue.
   The **Choose Install Folder** screen appears.



10. Install DSDS at the default location by clicking **Next**.

> **Note**
>
> If you are installing on Linux and you want to install DSDS on a different location than the default: Select "Choose...", navigate to the desired location, select "OK". The "Where Would You Like To Install?" line now shows your selection. Make sure to add "/DSDataServer5000.10.3" (or similar) to the path (for example: ***/apps/lgc/R5000/DSDataServer5000.10.3)***. The "DSDataServer5000.10.3" folder will be created by the installer.

The **Pre-Installation Summary** screen appears.



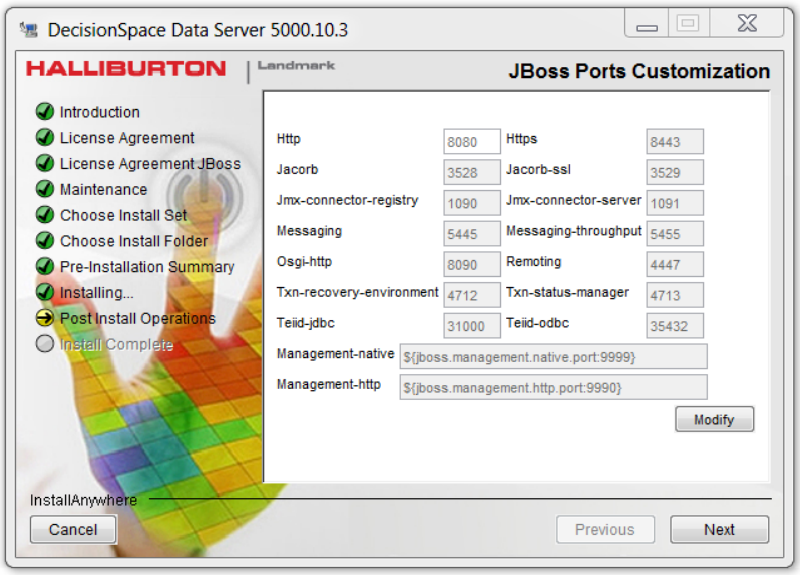
11. As appropriate, perform one of the following procedures:

    - To use the license server and port number settings in the Landmark license file (LM_LICENSE_FILE) on your system, leave these fields blank and click **Next**.

    - To override the license server and port number settings in the license file on your system, enter the **License Server** and **Port** number, and then click **Next**.

> **Note**
>
> Your instructor will tell you where the license file is located.

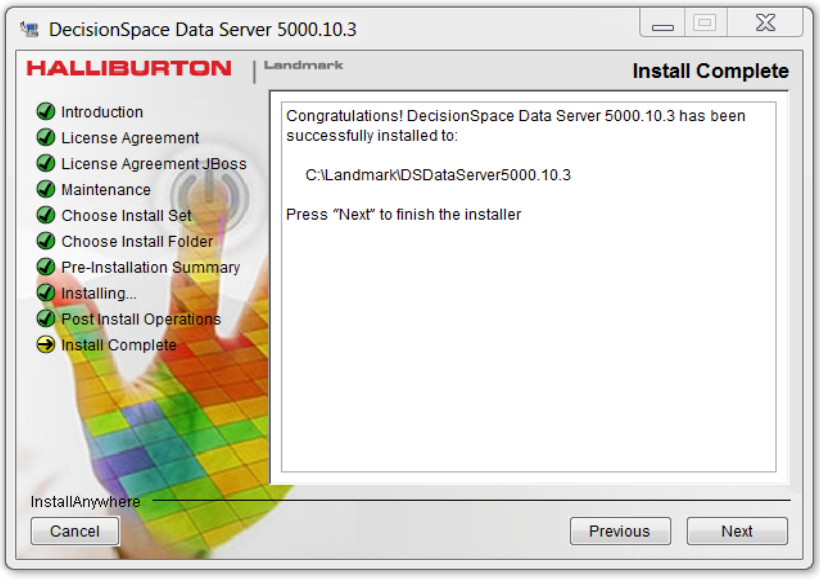
The **JBoss Ports Customization** screen appears.



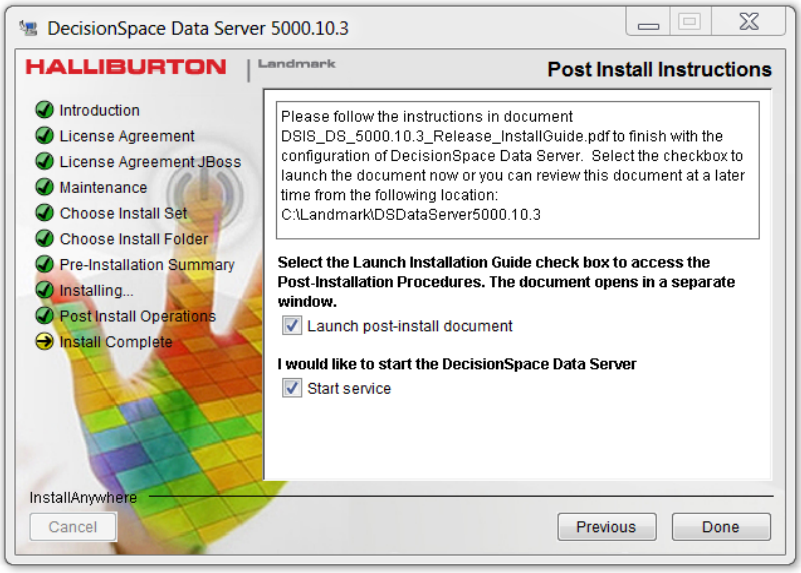
12. As appropriate, perform one of the following procedures:

    - Click **Next** to accept the default ports.
    - Click **Modify** to change the ports followed by **Validate Ports** to check the validity of the ports.

13. Click **Next** to continue.
    A progress screen displays during installation. Once the installation is complete, the **Install Complete** screen appears.



14. Click **Next** to continue.
    The **Post Install Instructions** screen appears.



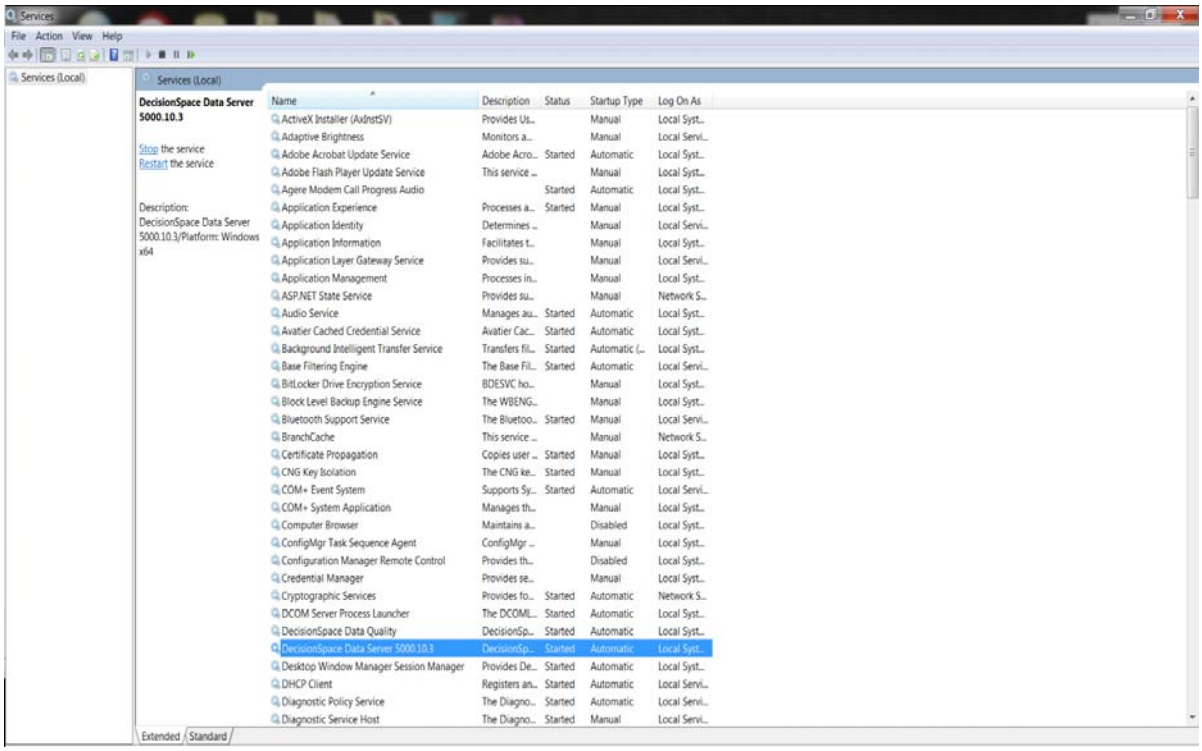
15. Select the **Start service** check box to start the **JBoss** server.

16. Click **Done** to exit the **installer**.

Starting the DecisionSpace Data Server Service

On Windows:

1. Enter **services** in the Search box on the **Start** menu and then press **<Enter>** on your keyboard.
   The **Services** screen appears.



2. Select **DecisionSpace Data Server 5000.10.3**.

3. Click **Start the service** to start the DecisionSpace Data Server service.

4. Double-click the service to change the **Startup Type** to **Automatic**.

> **Note**
>
> The DSDS Service in Windows can also be launched from a command window, by executing the following script:
>
> `<DSDS_INSTALL_HOME >\ApplicationServer\bin\runDSDS.bat`

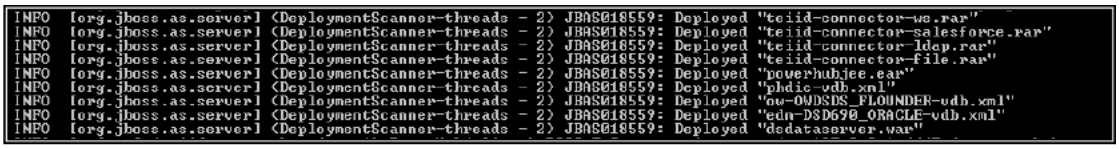
On Linux:

From the terminal, execute the following script:
`<DSDS_INSTALL_HOME>/bin/runDSDS.sh`

> **Note**
>
> **DO NOT** close the window if using this method. If you wish the program to run in the background use the '&' sign at the end of the command which will allow you to continue to use the same window for other functions.

On Both Platforms:

If you are running the **runDSDS** script, wait for a few moments and verify that all JBoss modules are deployed (as shown in the illustration below). Ensure that Teiid VDBs are set to **ACTIVE**.

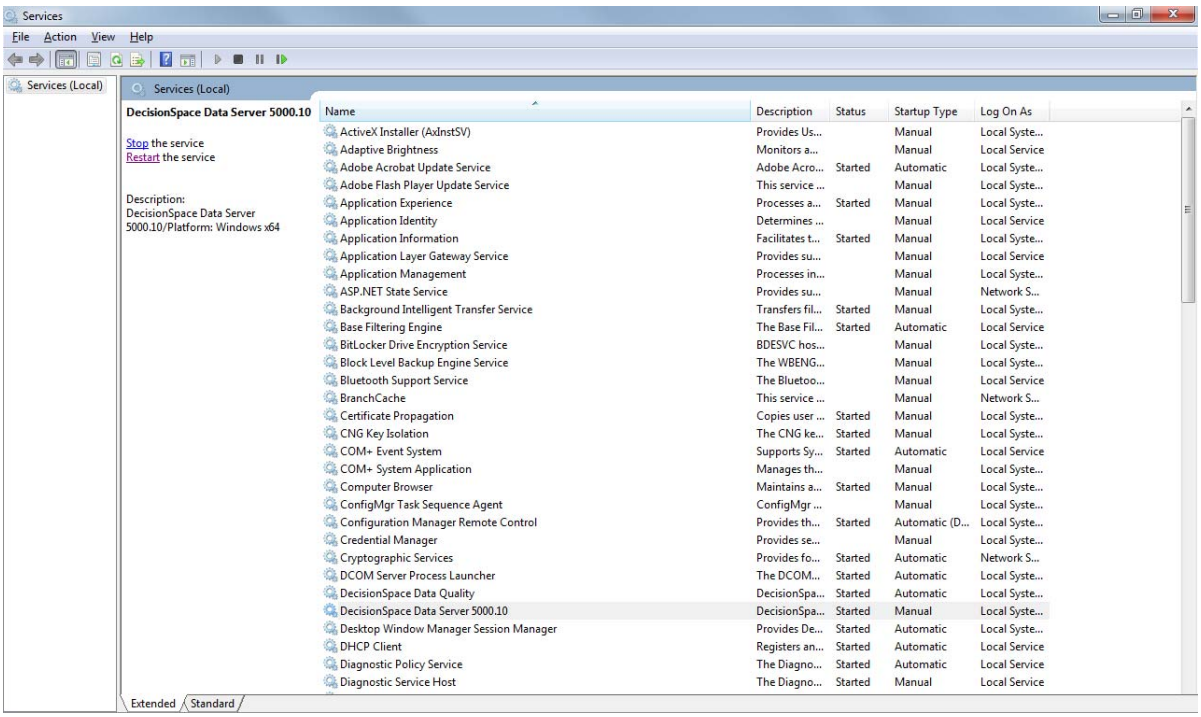


Exercise: Stopping the DecisionSpace Data Server Service

To stop the DecisionSpace Data Server service:

On Windows:

1. Enter **services** in the Search box on the **Start** menu and then press **<Enter>** on your keyboard.
   The **Services** screen appears.



2. Select **DecisionSpace Data Server 5000.10.3**.

3. Click **Stop the service** to stop the DecisionSpace Data Server service.

On Linux:

Terminate the process by entering <**Ctrl+C**> in the terminal running the service.

Post-Installation Procedures

Once the DecisionSpace Data Server is installed and configured, it can connect to a few internal development data sources (currently shipped as samples). In order to use the services against your own data, the following steps are necessary.

- Add a Data Source
- Test the Connection to the Data Source
- Generating VDBs (Virtual Databases)

Once the DSDS service is running, the DSDS Admin Console can be accessed using one of these methods:

- Start Menu -> All Programs ->Landmark -> DecisionSpace Data Server 5000.10.3 -> Start DecisionSpace Data Server Console
- Open the Web browser and enter the following URL:
  - ***http://<server_name>:8080/dsdataserver-console,*** *or*
  - ***http://localhost:8080/dsdataserver-console*** (if it is installed locally)

When prompted for credentials, the default user name and password for the console is **dsdsadmin**.

Exercise: Adding a Data Source

Prior to adding a data source and generating your VDB (virtual database), collect the necessary information for the type of data source you want to add. In order to determine what is necessary, highlight the data source type you wish to create and click **Create**. The list of expected parameters is displayed.

For example, if you are adding a **JDBC EDM-SQLServer** data source, you will be asked for the following parameters:

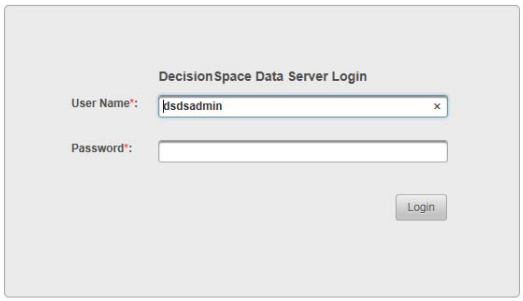
```
jdbc:sqlserver://
[serverName[instanceName][:portNumber]][;property=value[;
property=value]]
```

Where:

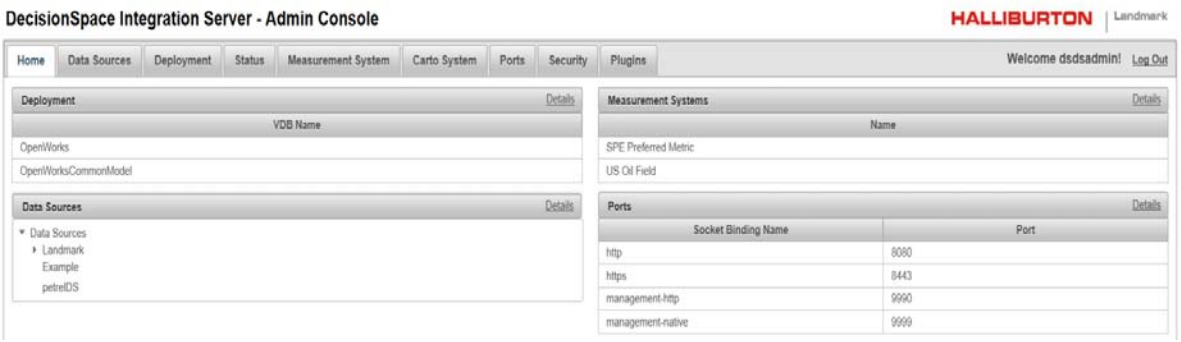
```
jdbc:sqlserver://localhost\EDM5000:0;DatabaseName=EDMDB
```

To add a data source:

1. Open your internet browser and enter the following URL in the address bar: ***http://<Server-IP Address: Port>/dsdataserver-console (e.g., http://localhost:8080/dsdataserver-console)***.
   The **Authentication Required** window appears.



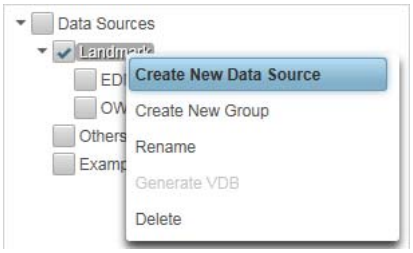
2. Enter **dsdsadmin** in the **User Name** and **Password** fields and click **Login**.
   The DecisionSpace Data Server Console window appears with the **Home** tab selected by default.



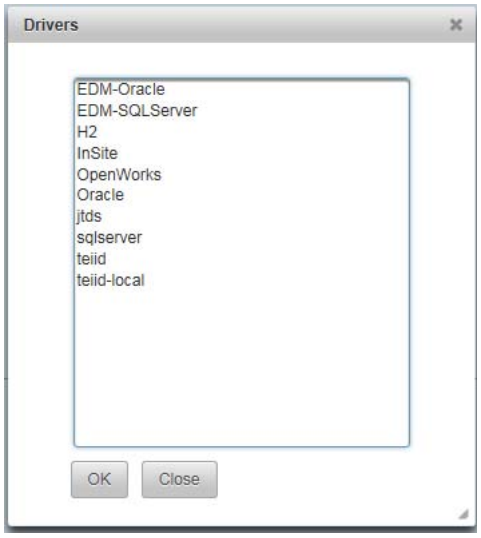
3. Select the **Data Sources** tab.
   A list of existing data source types appears on the left side of the DecisionSpace Data Server Console window.

4. Select the **Landmark** data source check box.

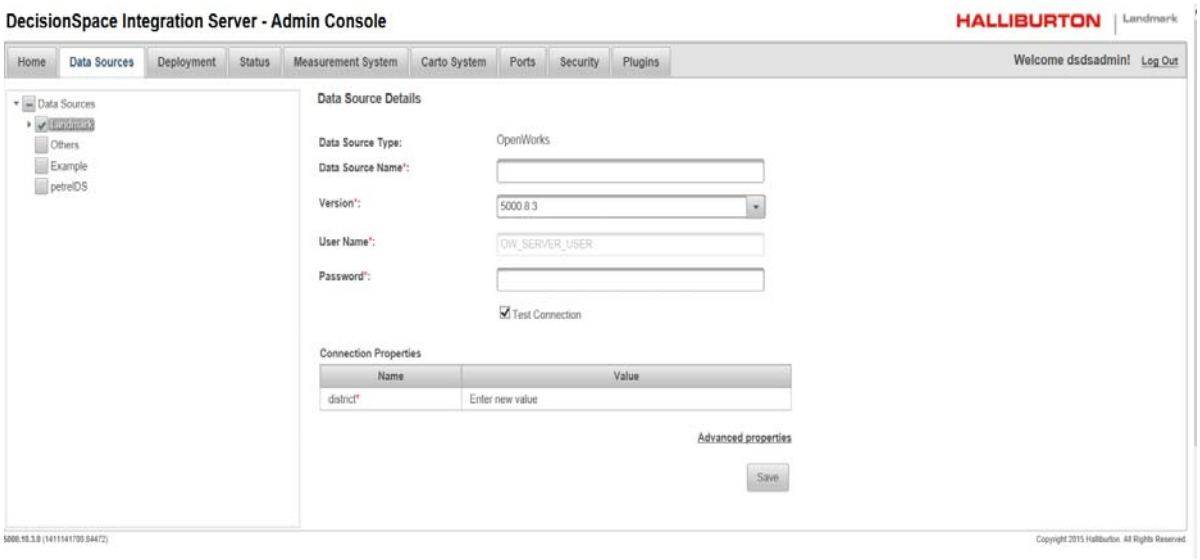
5. Right-click the **Landmark** data source and select **Create New Data Source** from the context menu.



The **Drivers** window appears.



6. Select the **OpenWorks** data source and click **OK**.
   The **Data Source Details** window appears.

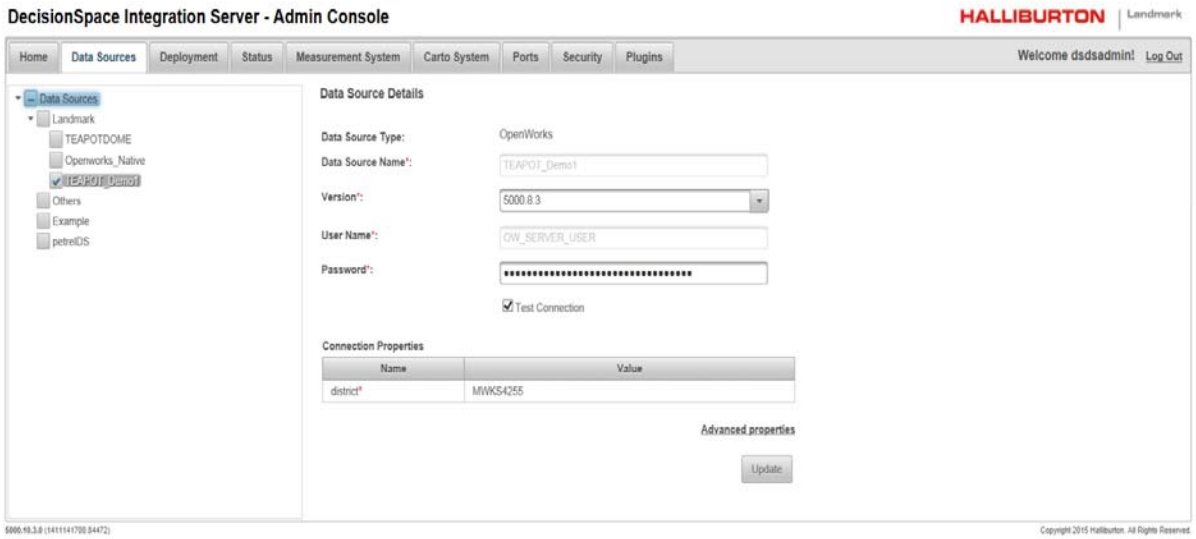


7. Enter **TEAPOT_Demo1** in the **Data Source Name** field.

8. Enter password for **OW_SERVER_USER**.

9. In the **Connection Properties** group box, enter a value for the district.

10. Click **Save**.
    The new data source displays in the tree list on the left side of the DecisonSpace Data Server Console window.



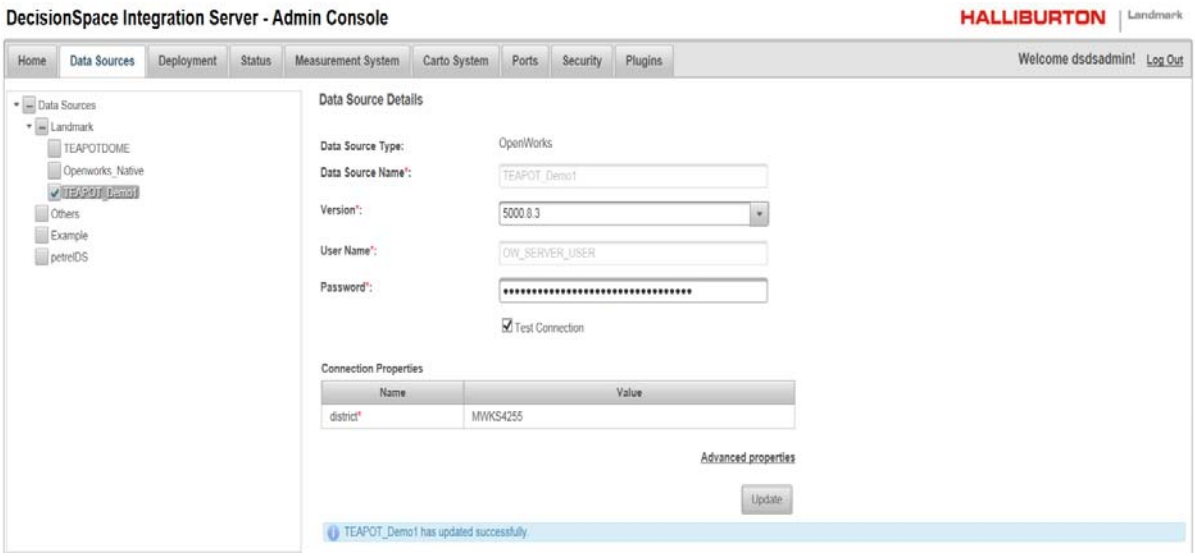
Testing the Connection to the Data Source

To test a connection of the data source that you have just added:

1. Select the **Data Sources** tab.

2. Select the **TEAPOT_Demo1** check box from the list on the left side of the DecisionSpace Data Server Console.

3. Right-click the **TEAPOT_Demo1** option and select **Test** from the context menu.
   A message appears on the DecisionSpace Data Server Console window displaying the status of the connection.



Generating Virtual Databases (VDBs)

You can create single and multi-source VDBs. Multi-source VDBs must use the same data source type as you cannot mix data source types. To select two or more data sources, press and hold the **<CTRL >** key on your keyboard and click on each data source in the **Data Sources** tree to select it.

> **Note**
>
> When generating a VDB for data sources listed under the "Others" section, it is necessary to gather more information about model, group and schema, prior to generating the actual VDB. For more details on the required information, see the "Generate VDB – Others" section below.

To create a single-source VDB for the data source that you have just added i.e., **TEAPOT_DEMO1**.

1. Select the **Data Sources** tab.

2. Select the **TEAPOT_Demo1** check-box from the list on the left side of the DecisionSpace Data Server Console.

3. Right-click the **TEAPOT_Demo1** option and select **Generate VDB** from the context menu.
   A message appears on the DecisionSpace Data Server Console displaying the status of the connection

Errors (if any) generated during this process are stored in the **server.log** file. Select the **Status** tab or browse to the **JBOSS_HOME/standalone/log** directory to view this file.

> **Note**
>
> After deployment of a VDB on the server, the data sever will try to connect to the data source(s) selected for VDB generation and load the metadata of those data sources (tables, relations etc). It may take a few moments before the data becomes available. If anything goes wrong during this process, the VDB will not get deployed. To view the full details of a deployment, select the **Deployment** tab.

4. To check if the deployment was successful, click on the **Deployment** tab.
   The OpenWorks connection displays in the **Deployment** tab.

5. Click on **Details** to see more details about the deployment of the **VDB**.

> **Note**
>
> If you want to delete a **VDB**, click **Delete**. The **VDB** is deleted along with the connection information.

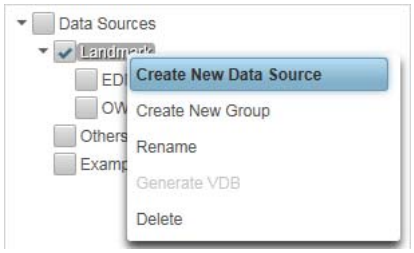
Exercise: Creating Database Connections

The purpose of this exercise is to create, test, generate & deploy theVDB and it also shows deletion of database connections in DSDS. The databases used in this exercise are OpenWorks and EDM:

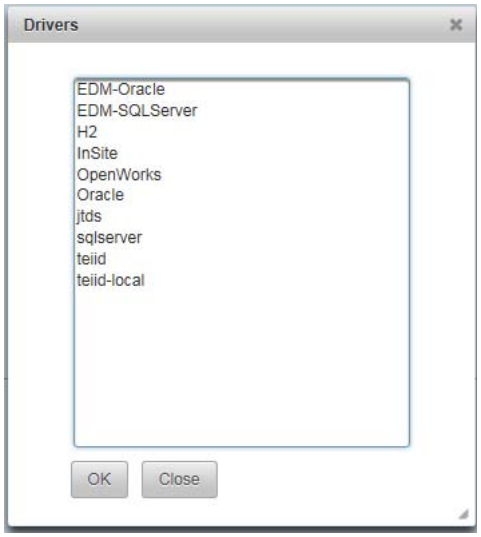
1. Select the **Data Sources** tab of the DecisionSpace Data Server Console.
   A list of existing data source types appears.

2. Select the **Landmark** data source check box.

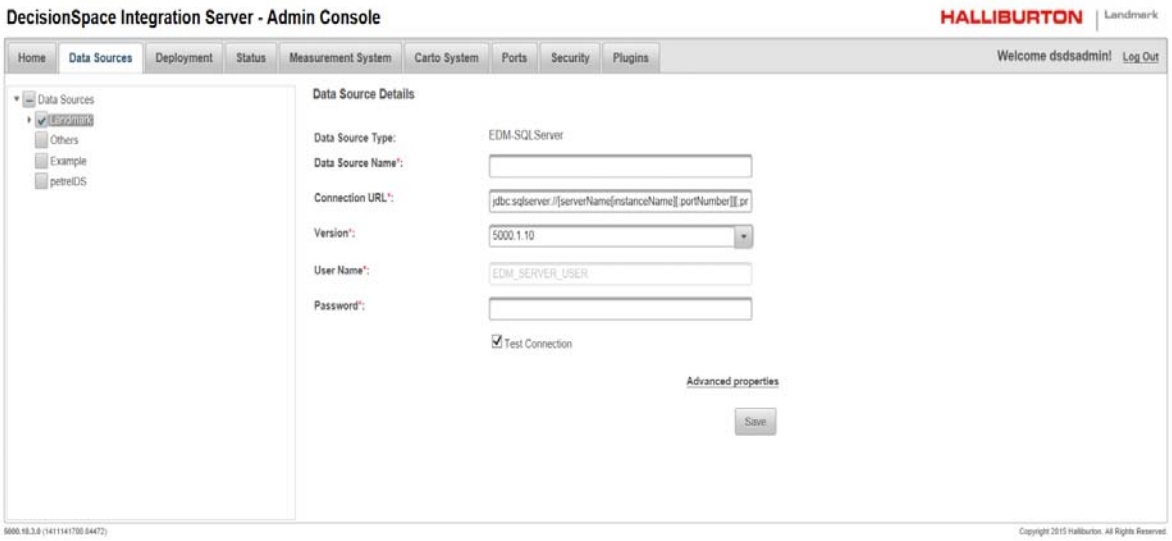
3. Right-click the **Landmark** data source and select **Create New Data Source** from the context menu.



The **Drivers** window appears.



4. Select the **EDM-SQL Server** data source and click **OK**.
The **Data Source Details** window appears.



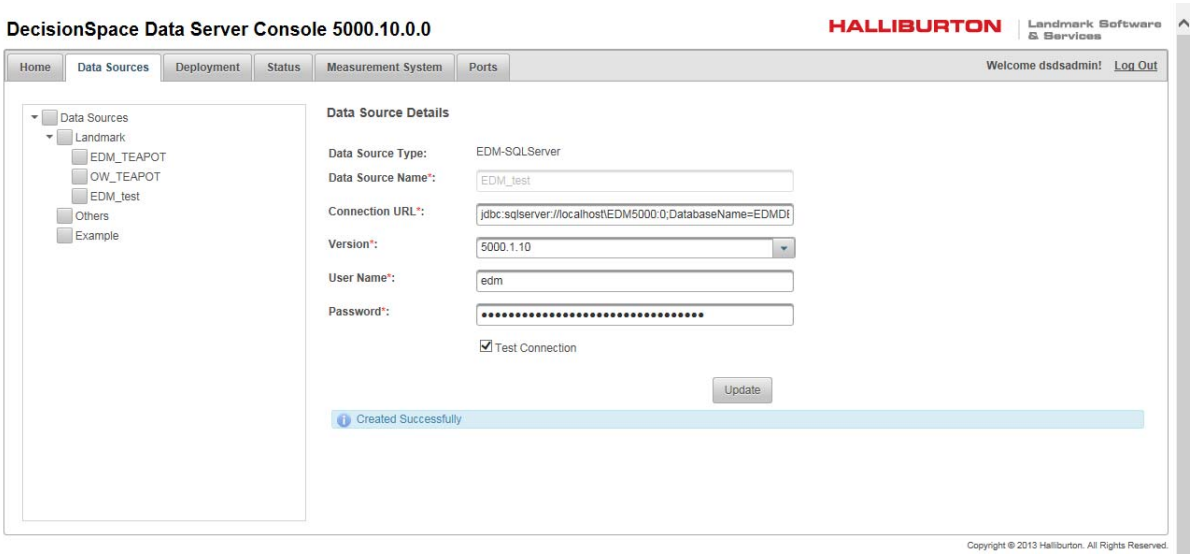
5. Enter **EDM_test** in the **Data Source Name** field.

6. Enter the following syntax in the **Connection URL** field:

```
jdbc:sqlserver://localhost\EDM
5000:0;DatabaseName=EDMDB
```

7. Enter **edm** and **Landmark1** in the **Username** and **Password** fields respectively.

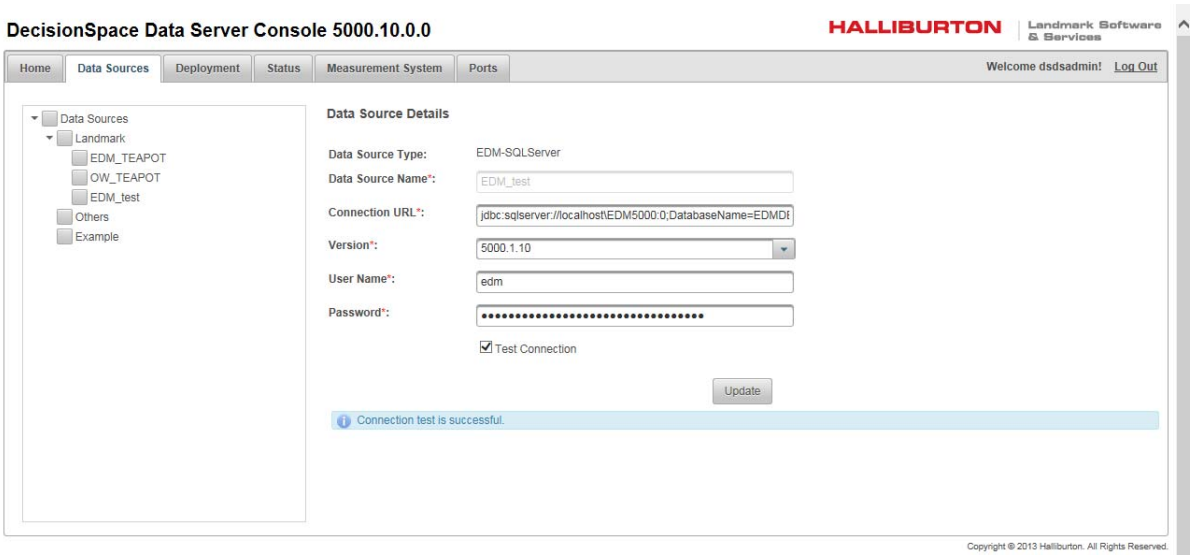
8. Click **Save**.
   The new data source displays in the tree list on the left side of the DecisonSpace Data Server Console window.



9. Select the **EDM_test** check-box from the list on the left side of the DecisionSpace Data Server Console.

10. Right-click the **EDM_test** data source and select **Test** from the context menu.

A message appears on the DecisionSpace Data Server Console displaying the status of the connection.



11. Select the **EDM_test** check-box from the list on the left side of the DecisionSpace Data Server Console.

12. Right-click the **EDM_test** data source and select **Generate VDB** from the context menu.
    A message appears on the DecisionSpace Data Server Console displaying the status of the connection

Errors (if any) generated during this process are stored in the **server.log** file. Select the **Status** tab or browse to the **JBOSS_HOME/standalone/log** directory to view this file.

Note

After deployment of a VDB on the server, the data sever will try to connect to the data source(s) selected for VDB generation and load the metadata of those data sources (tables, relations etc). It may take a few moments before the data becomes available. If anything goes wrong during this process, the VDB will not get deployed. To view the full details of a deployment, select the **Deployment** tab.

13. To check if the deployment was successful, click on the **Deployment** tab.

14. Click on **Details** to see more details about the deployment of the VDB.

> **Note**
>
> If you want to delete a **VDB,** click **Delete**. The **VDB** is deleted along with the connection information.

Exercise: Connecting to DSDS using Web Browser, Excel (PowerPivot)

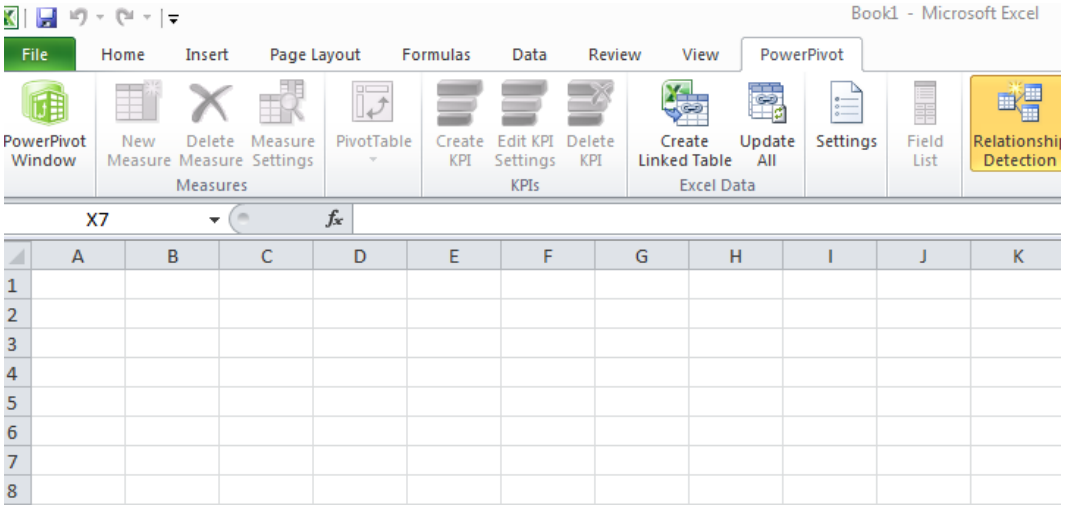
In this exercise you will be using PowerPivot to analyze of the data from OpenWorks and EDM.

Once all the tables are imported you can run the analysis on the data. We will be using the data for the table MD_PK_PDEN_TYPE and the analysis we want to do is to find out the count of PDEN_TYPE.

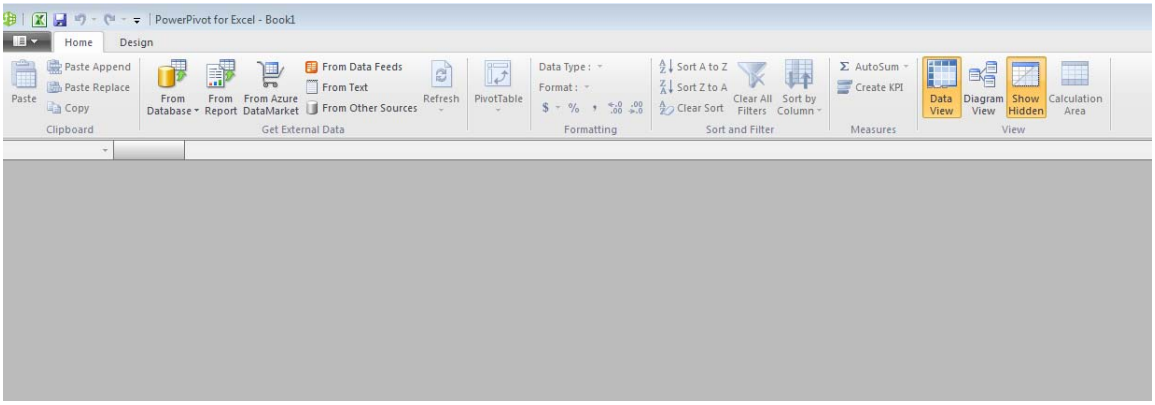
After this exercise you will be able to perform different analyses based on requirements.

1. Launch **Excel**.

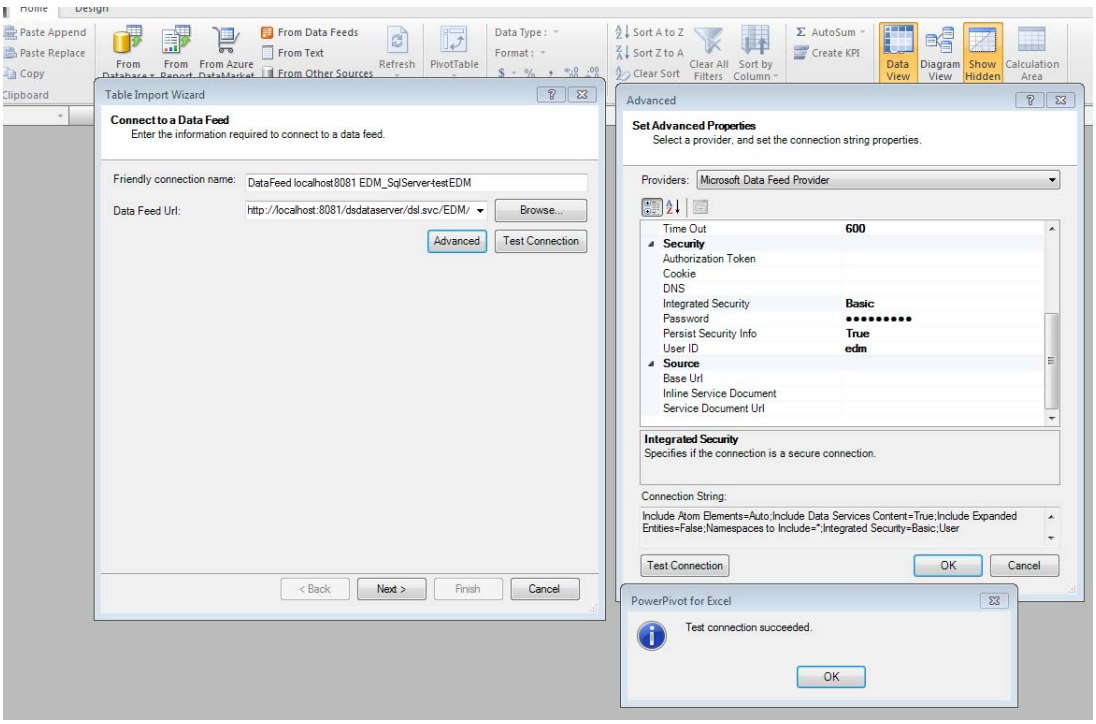
2. Launch **PowerPivot**.



3. Launch the **PowerPivot** Window.



4. Create the connection to **EDM** data feeds.



5. Click **OK** on the **PowerPivot for Excel** dialog box.

6. Click **Next**.

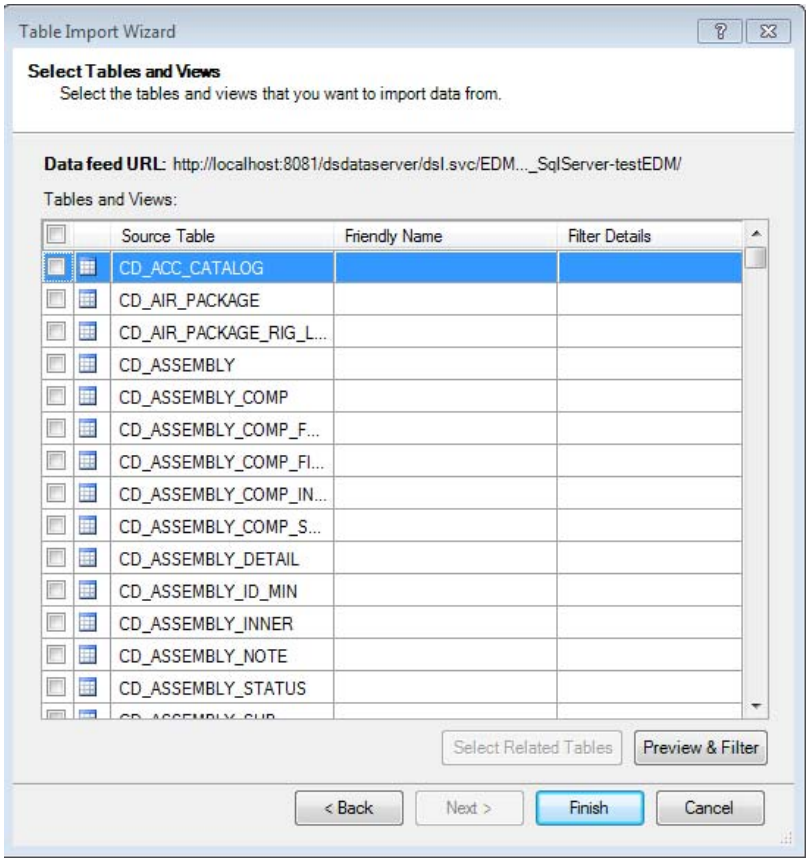
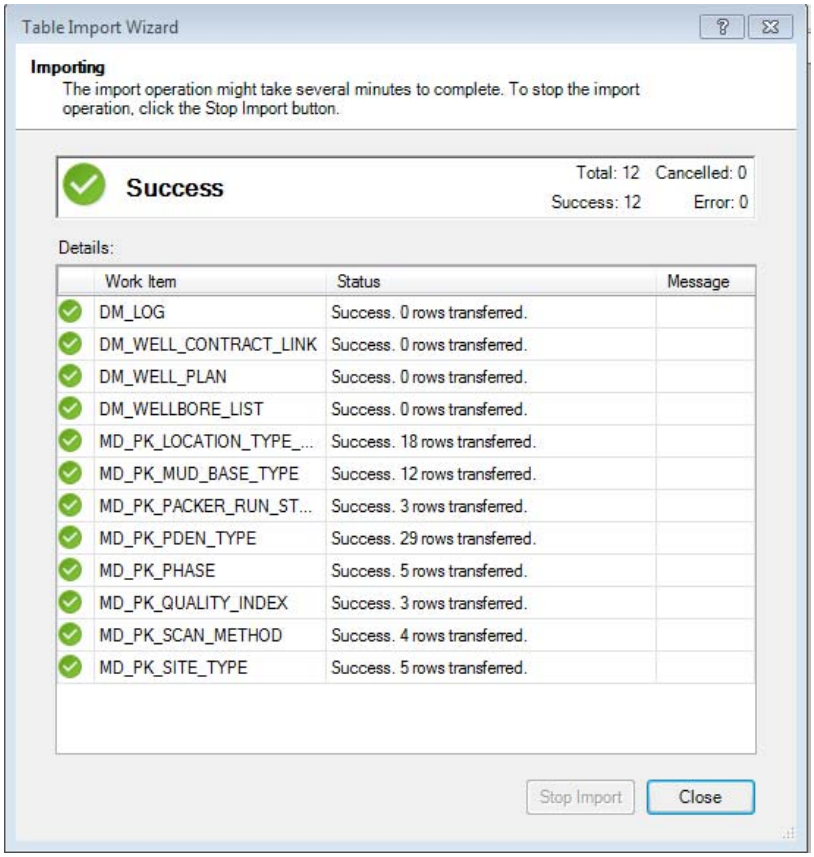
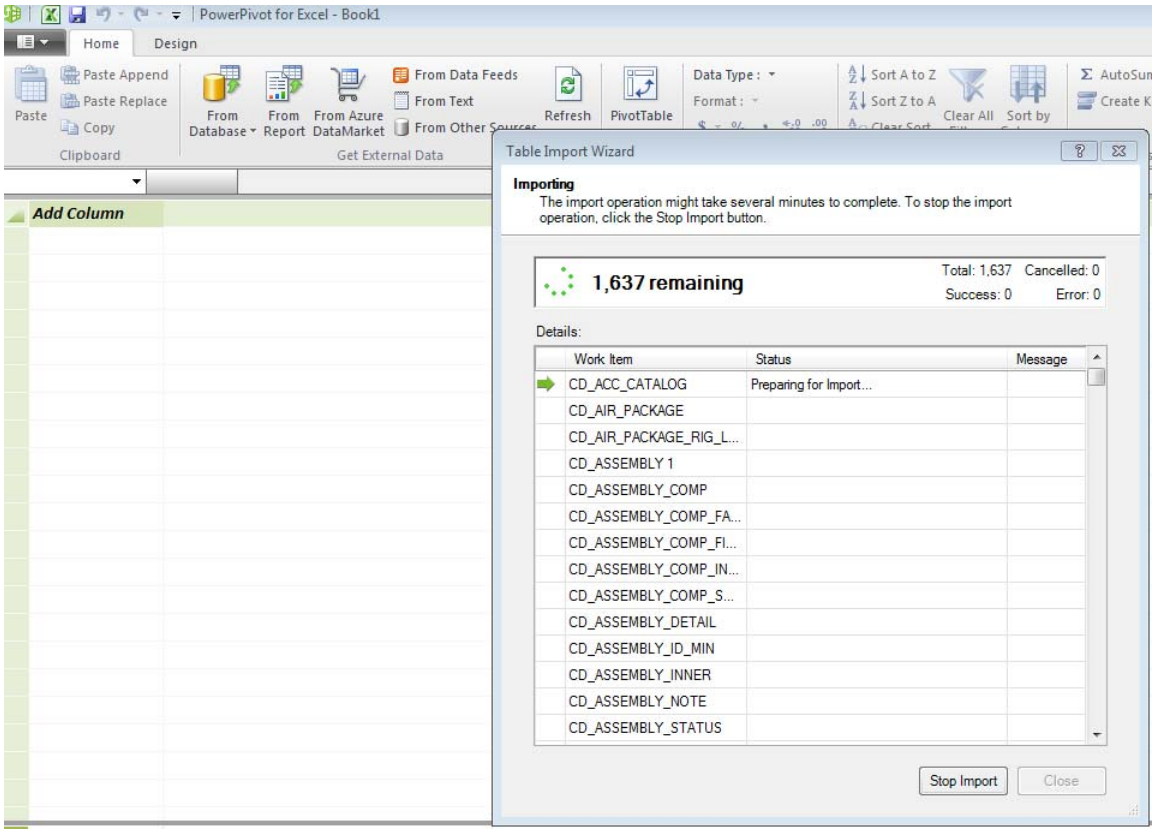


Table Import Wizard
Select Tables and Views
Select the tables and views that you want to import data from.
Data feed URL: http://localhost:8081/dsdataserver/dsl.svc/EDM..._SqlServer-testEDM/
Tables and Views:
Source Table
Friendly Name
Filter Details
CD_ACC_CATALOG
CD_AIR_PACKAGE
CD_AIR_PACKAGE_RIG_L...
CD_ASSEMBLY
CD_ASSEMBLY_COMP
CD_ASSEMBLY_COMP_F...
CD_ASSEMBLY_COMP_FI...
CD_ASSEMBLY_COMP_IN...
CD_ASSEMBLY_COMP_S...
CD_ASSEMBLY_DETAIL
CD_ASSEMBLY_ID_MIN
CD_ASSEMBLY_INNER
CD_ASSEMBLY_NOTE
CD_ASSEMBLY_STATUS
Select Related Tables
Preview & Filter
< Back
Next >
Finish
Cancel

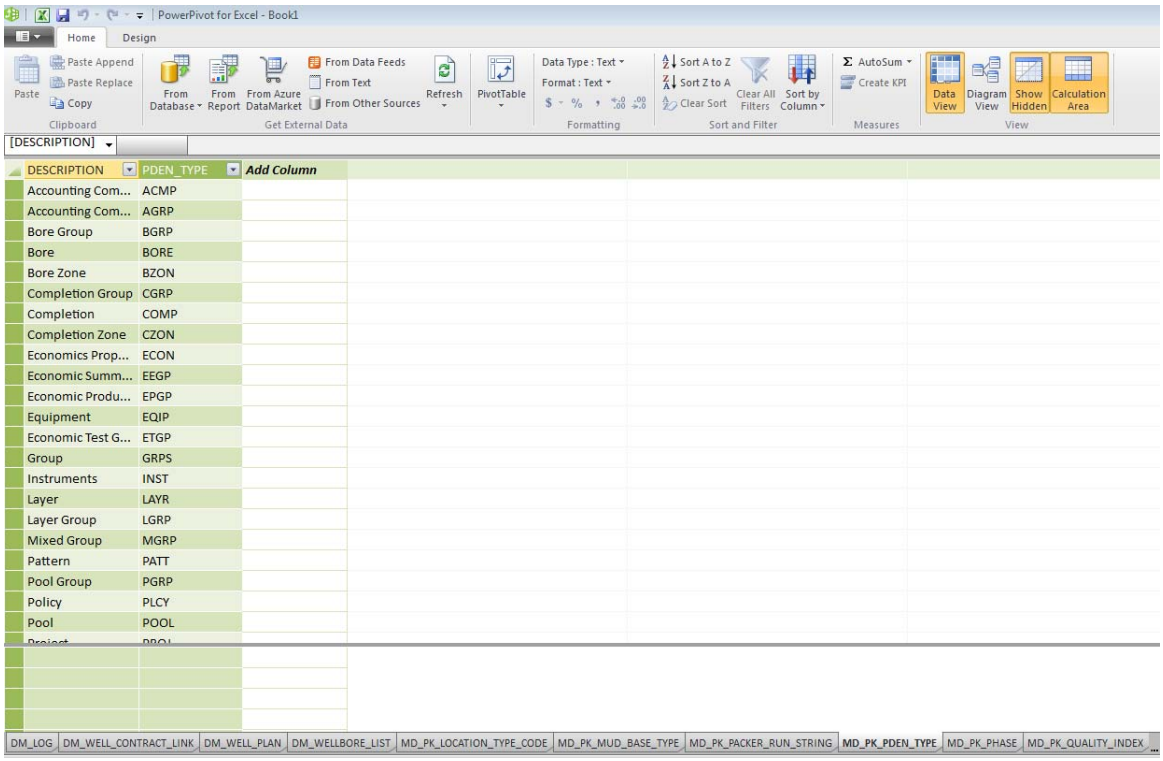
7. Select all the source tables that you want to do analysis on and then click **Finish**.

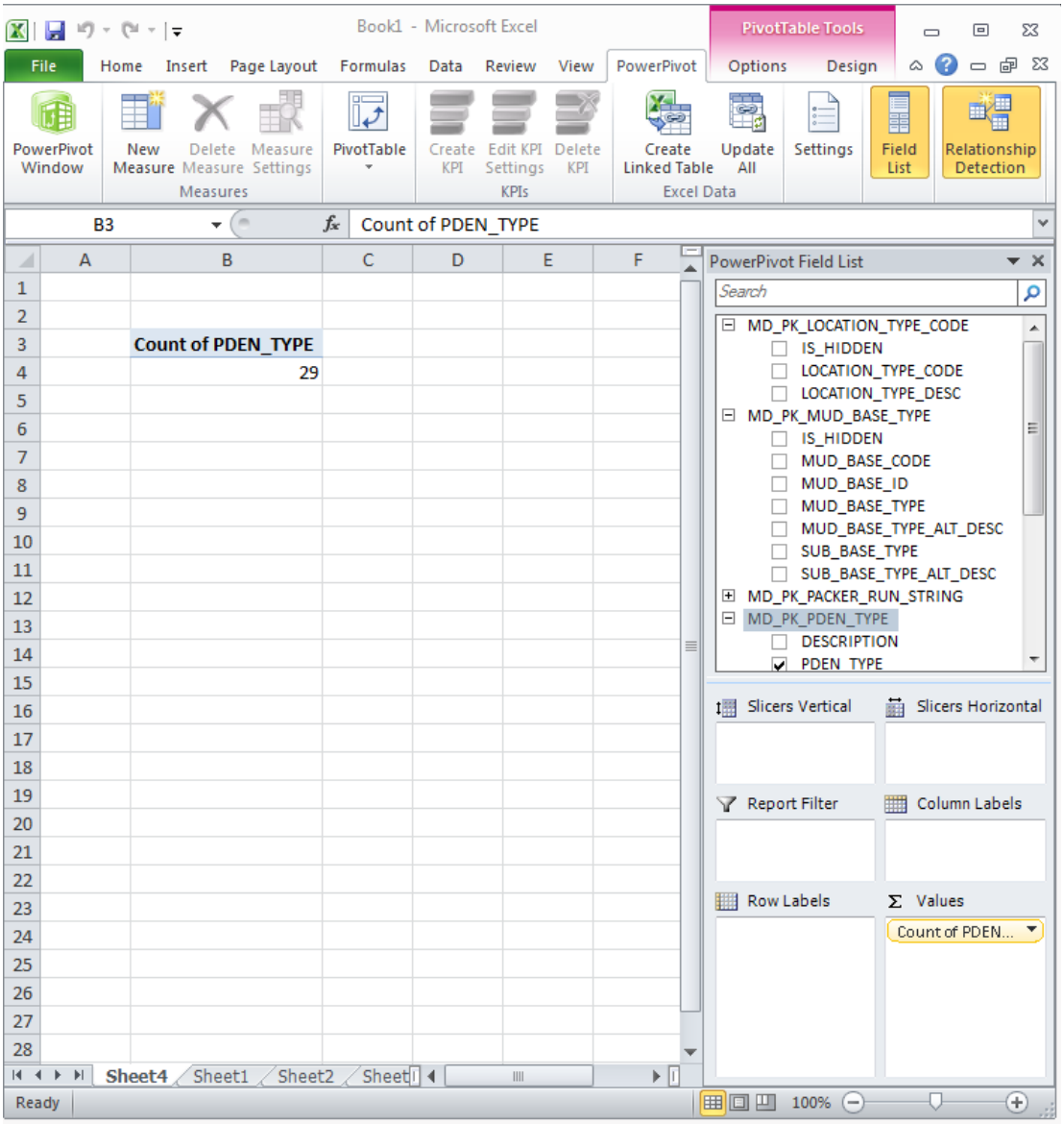


8. Select all tables.
   MS Excel imports all the tables.



9. Run the analysis on the imported data.
   Data for the table displays.




Book1 - Microsoft Excel
PivotTable Tools
File
Home
Insert
Page Layout
Formulas
Data
Review
View
PowerPivot
Options
Design
PowerPivot Window
New Measure
Delete Measure
Measure Settings
Measures
PivotTable
Create KPI
Edit KPI Settings
Delete KPI
KPIs
Create Linked Table
Update All
Excel Data
Settings
Field List
Relationship Detection
B3
Count of PDEN_TYPE
Count of PDEN_TYPE
29
PowerPivot Field List
Search
MD_PK_LOCATION_TYPE_CODE
IS_HIDDEN
LOCATION_TYPE_CODE
LOCATION_TYPE_DESC
MD_PK_MUD_BASE_TYPE
IS_HIDDEN
MUD_BASE_CODE
MUD_BASE_ID
MUD_BASE_TYPE
MUD_BASE_TYPE_ALT_DESC
SUB_BASE_TYPE
SUB_BASE_TYPE_ALT_DESC
MD_PK_PACKER_RUN_STRING
MD_PK_PDEN_TYPE
DESCRIPTION
PDEN_TYPE
Slicers Vertical
Slicers Horizontal
Report Filter
Column Labels
Row Labels
Values
Count of PDEN...
Sheet4
Sheet1
Sheet2
Ready
100%

Connecting DecisionSpace Data Quality with DecisionSpace Data Server

Connections in the DecisionSpace Data Quality consist of the following three Source Dataset elements:

- Data Source (Data Owner) - where data is read from
- Workspace - schema where results are written
- Data Model - location where detailed information about data source tables and columns, table relationships, and element assignments are stored

These three elements help connect each phase of the Data Quality software to its own unique configuration. The 'Data Owner' connections can only be deleted from the application, when all the connected phases are deleted. Workspace connections and models are editable and can be updated as and when required.

Note

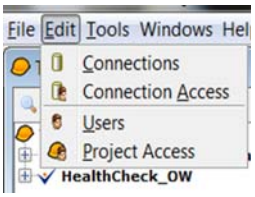
A warning message appears to contact Support if the maximum number of purchased connections is exceeded.

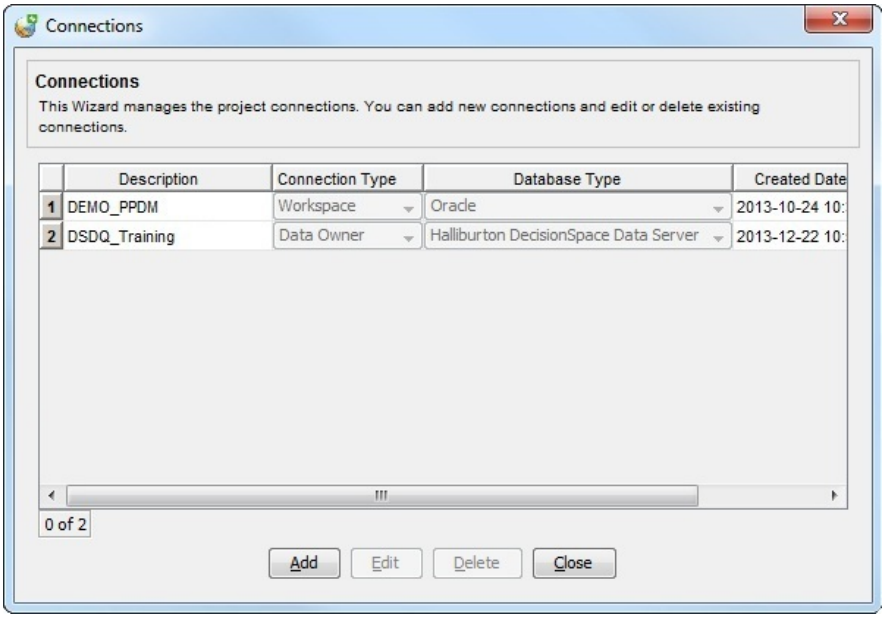
Exercise: Creating Connections

A Data Owner (the data source that the application reads from) must exist prior to reaching this step. This section outlines the process to create a connection between the Data Quality application and the Workspace schema (where results are written) created prior to installing DecisionSpace Data Quality.

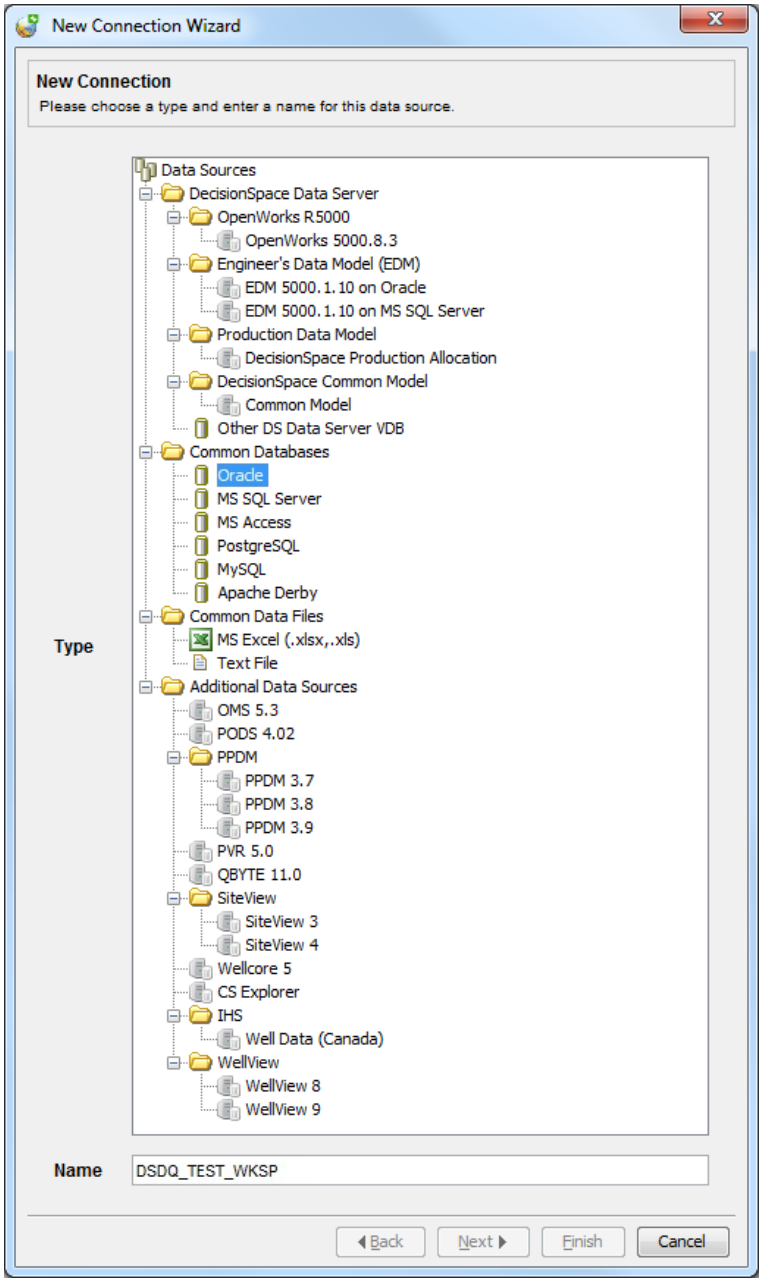
To add a new Data Owner connection:

1. Select **Edit > Connections** from the menu bar on the **DSDQ Project Window.**





2. Click **Add** to display the **New Connection** window.

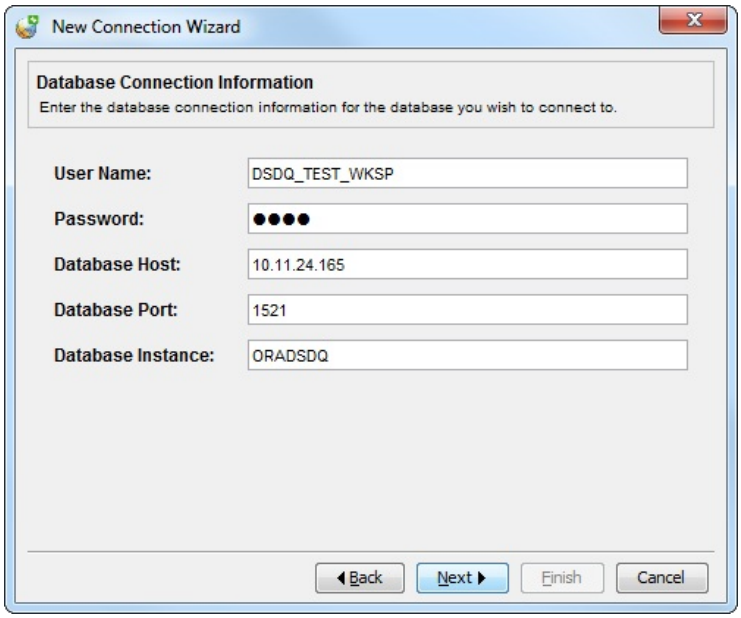


3. Select **Oracle** as the connection type from the **Type** tree list.

4. Enter **DSDQ_TEST_WKSP** in the **Name** field. This could be a real database name or an alias. This name is used for user reference only and appears in the drop-down list after the connection setup is complete.

> **Note**
>
> Supported Databases for Connection Type connections include: DecisionSpace Data Server, Oracle, MS SQL Server, Apache Derby, MySQL, PostgreSQL, PPDM 3.8, MS Access, .DBF files, .CSV files, and Excel spreadsheets.

5. Click **Next** to display the **Database Connection Information** window.



6. Enter **DSDQ_TEST_WKSP** as the user name for the database scheme that you are trying to connect to in the **User Name** field.

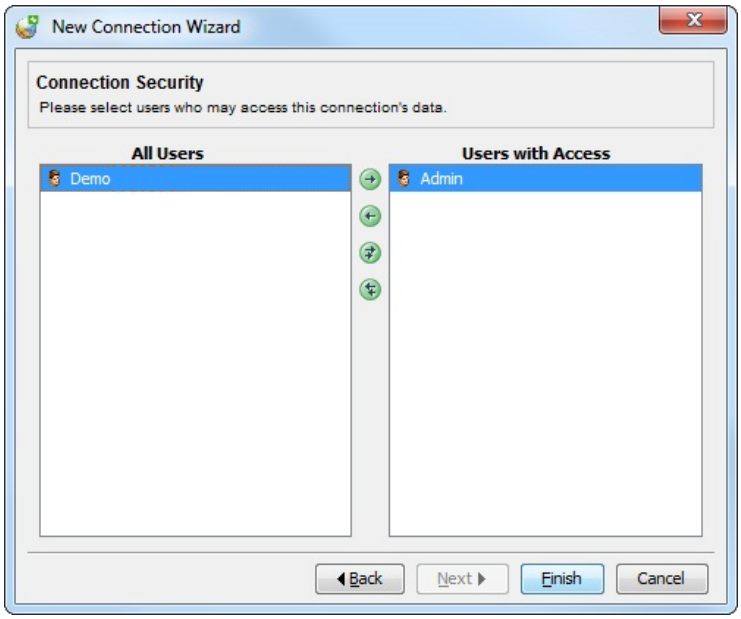
7. Enter the password used to connect to the database schema in the **Password** field.

8. Enter the host information for the server where the database schema resides in the **Database Host** field.

9. Enter the port number for connection to your data source in the **Database Port** field. The default value is 1521.

10. Enter a name for the database instance in the **Database Instance** field.

11. Click **Next** to display the **Connection Security** window.



12. Select user(s) who should have access to the data source from the **All Users** list and then click to move selected users to the **Users with Access** list.

13. Click **Finish** to complete the process of creating a connection between the Data Quality application and the Workspace schema (where results are written).

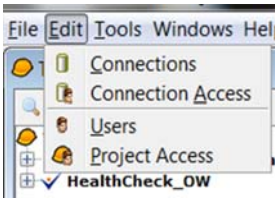
DecisionSpace Data Server Quick Start Connections

DecisionSpace Data Server allows users to connect to Landmark application databases, such as OpenWorks® and EDM™. By connecting to these application databases through the DecisionSpace Data Server, the Data Quality software can automatically load preconfigured models and rules against these databases.

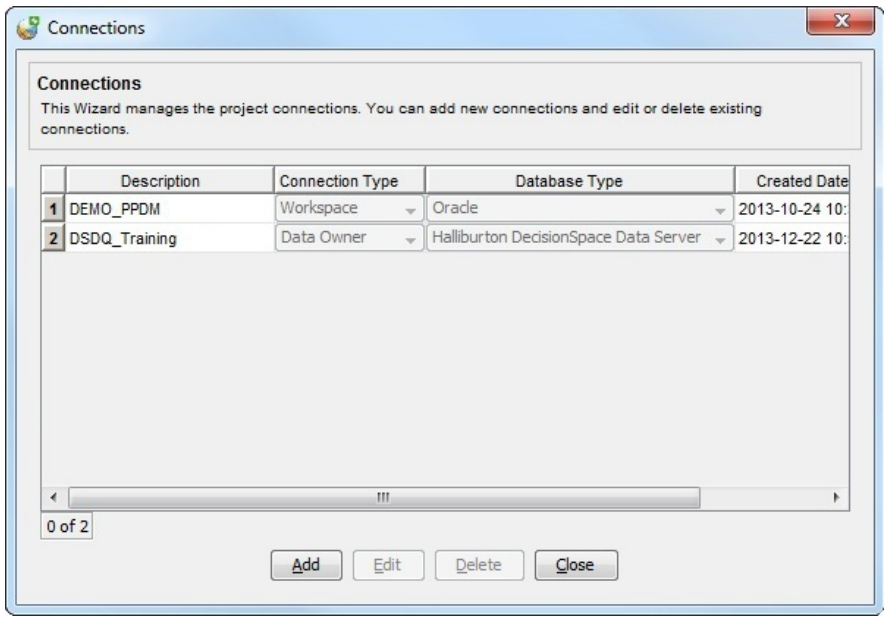
Exercise: Creating New DecisionSpace Data Server Quick Start Connections

To create a DecisionSpace Data Server Quick Start Connection:

1. Select **Edit > Connections** from the menu bar on the **DSDQ Project Window.**

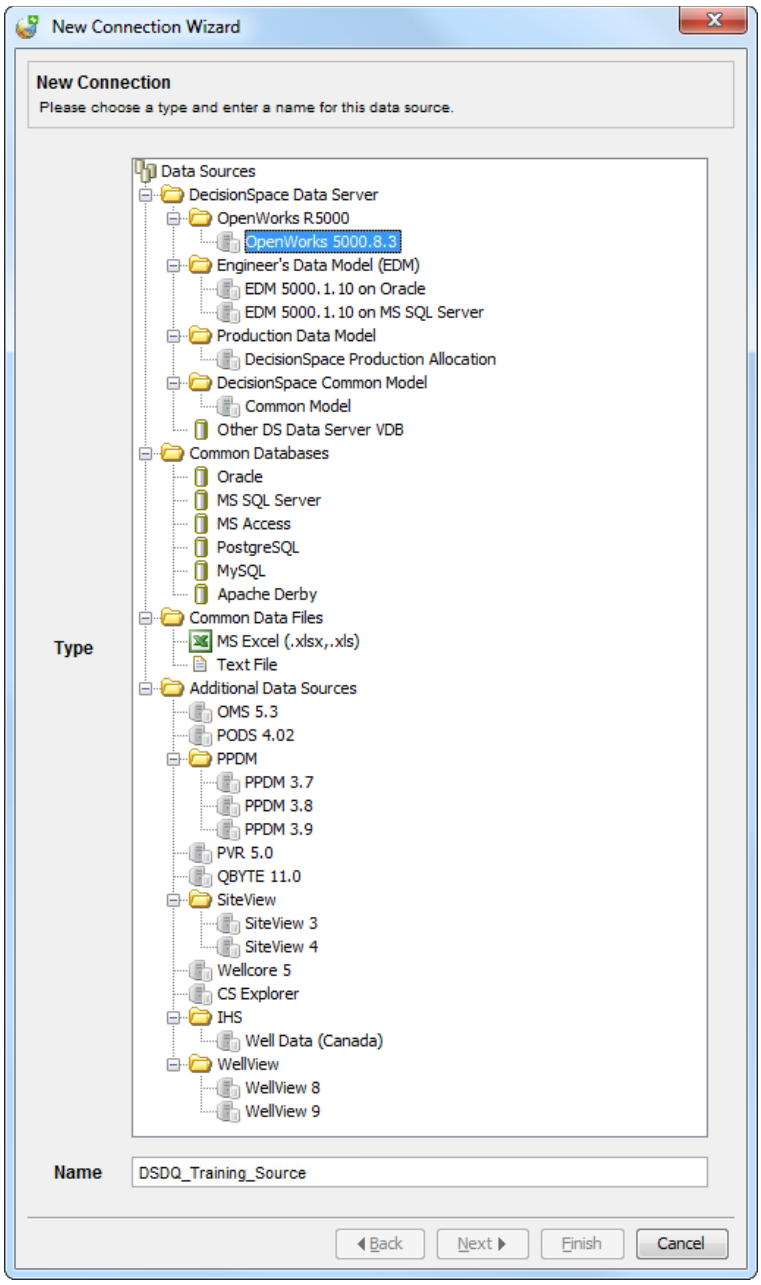


The **Connections** window displays.



2. Click **Add** to display the **New Connection** window.

The **New Connection Wizard** window displays.

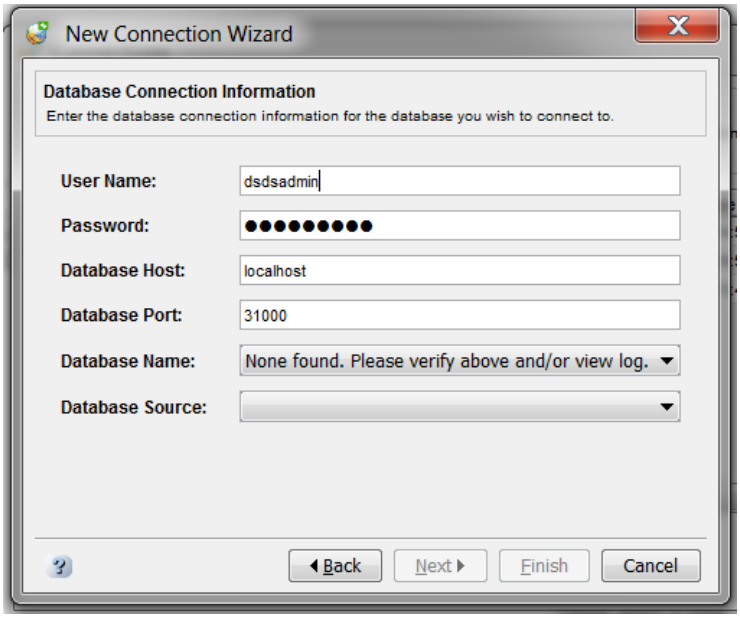


3. Select **OpenWorks 5000.8.3** as the connection type from the **Type** tree list.

> **Note**
>
> Supported DecisionSpace Data Server Quick Starts are OpenWorks R5000 and OpenWells.

4. Enter **DSDQ_Training_Source** as the name for the connection in the **Name** field. This could be a real database name or an alias. This name is used for user reference only and appears in the drop-down list after the connection setup is complete.

5. Click **Next** to display the **Source Connection Wizard** window.



6. Enter **dsdsadmin** as the user name for the database schema that you are trying to connect to in the **User Name** field.

7. Enter the password used to connect to the database schema in the **Password** field.

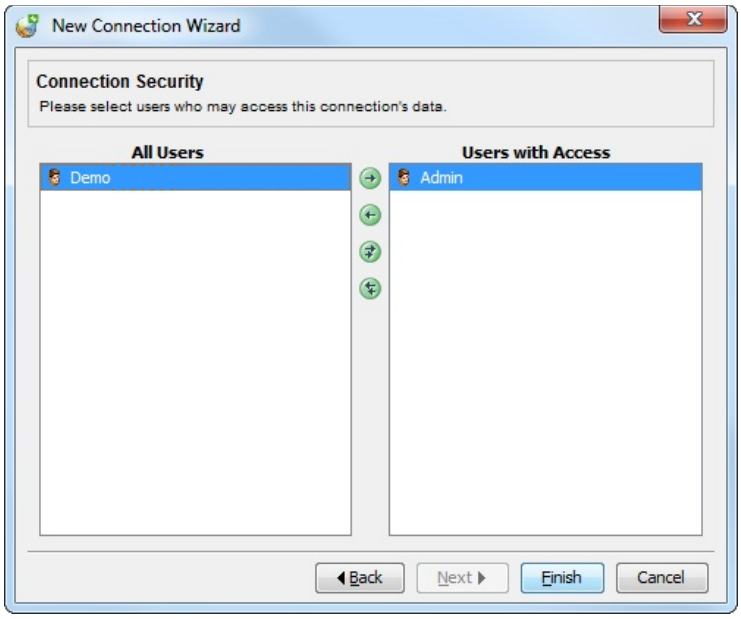
8. Enter the host information for the server where the database schema resides in the **Database Host** field.

9. Enter the port number for connection to your data source in the **Database Port** field. The default value is **31000**.

10. Select the database you want to connect to from the **Database Name** drop-down list.

11. Enter the name of the database source.

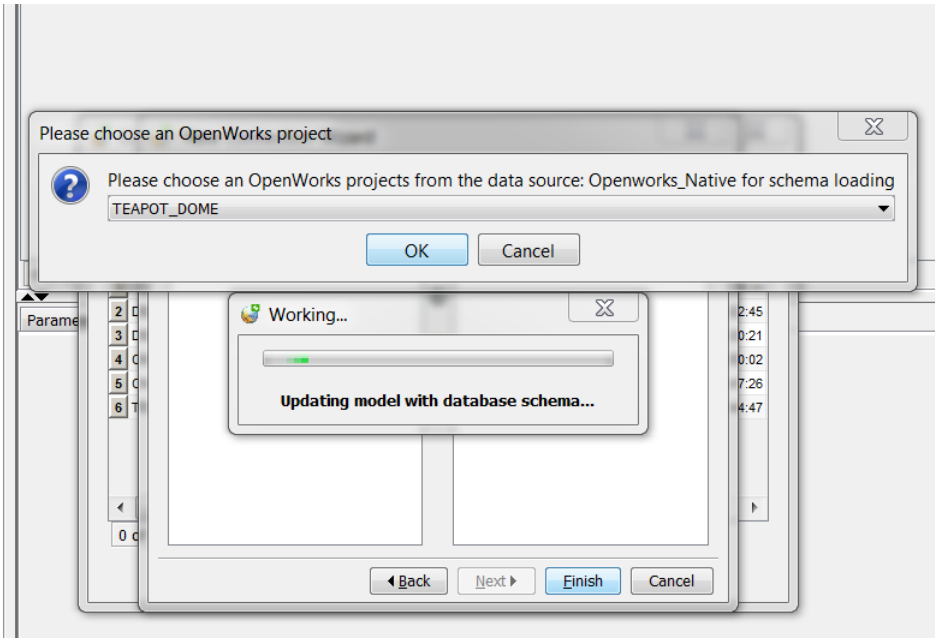
12. Click **Next** to display the **Connection Security** window.



13. Select user(s) who should have access to the data source from the **All Users list** and then click to move selected users to the **Users with Access** list.

14. Click **Finish** to connect the **OpenWorks 5000.8.3** application database with the Data Quality application.

15. Choose an OpenWorks project from the data source.



Chapter 3
Managing Projects

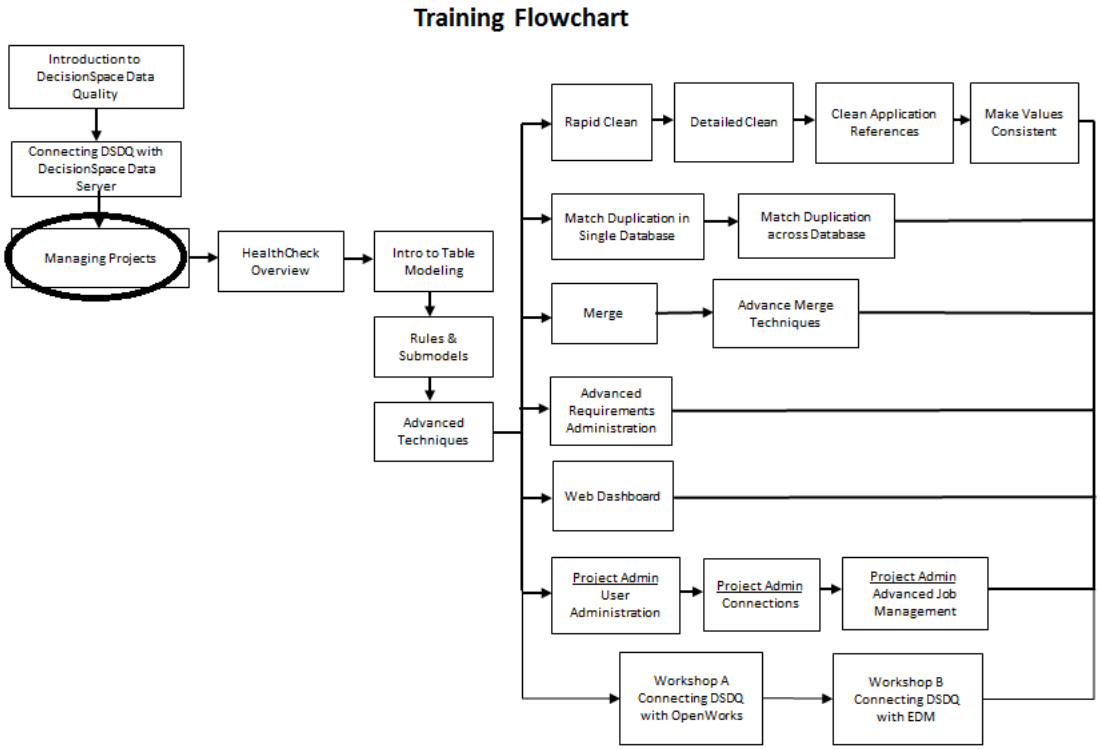
Before you load data that you want processed in DecisionSpace® Data Quality (DSDQ), you must create a project to contain all the Phases, Activities, Tools, Tasks, Jobs and Results associated with the source data. In DSDQ, a Project is a master folder that holds all the files created by end-users and the application while processing data. Rules to profile, audit and validate data quality (i.e., Phases) can only be defined once a project has been created. A Project in DSDQ is identified by a hard hat icon.

Chapter Overview

In this chapter, you will learn about:

- Creating a Project
- Managing Projects (i.e. opening and deleting projects)
- Importing/Exporting Projects

Topics covered in each chapter are outlined in the following illustration. Those specific to the current chapter will be circled in black for your reference:

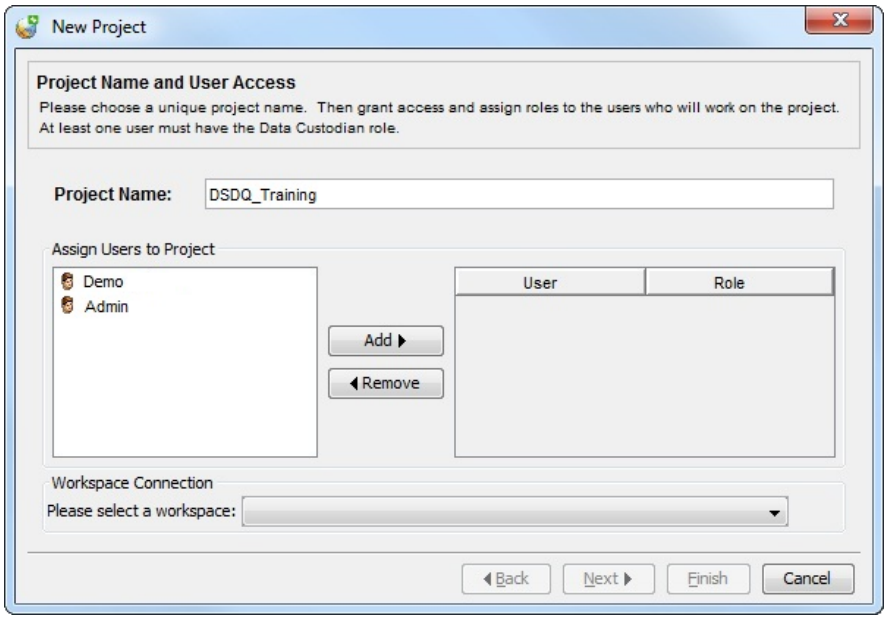


Creating a Project

A project can only be created by a user with **Administrator** rights in DSDQ. It is recommended that you specify a unique name for each project that you create. During this process, you will:

- Assign users to the project and their roles while they work on the project
- Select a Workspace Connection (the database where results will be written)
- Select the desired Phase and source connection (data source that the application reads from)

In all instances, the **New Project** wizard opens displaying all available users.



There are two roles within the Data Quality software-- Data Custodian and Data Steward. These roles carry with them an implicit description based on their roles in an organization:

- **Data Custodian:** Data custodians are responsible for facilitating the safe transport and storage of data. Often working directly in information technology and security, they aim to maintain data infrastructure and business rules. In the Data Quality application, the **Data Custodian** role relates to responsibilities such as setup

and configuration of rules, creating associations between columns and tables, and creating connections to data stores for **Data Stewards**

- **Data Steward:** Data Stewards have the primary responsibility of managing the content in data stores, as well as controlling any modifications to them. They deal with the daily governance of a company's information, and are often be subject matter experts on their governed data. Within the Data Quality application, **Data Stewards** are responsible for tasks directly related to the data within data stores, such as running jobs, confirming data matches, and consolidating information. **Data Stewards**, however, do not have access to create new jobs, or modify their configurations

Each role has a set of attached permissions. Permissions are organized by Activities/Tools as they appear in the Data Quality Tree or Tools Menu. Examples of Activities are Project Administration and Project Tools. Examples of Menu Tools are Manage Users and Manage Project Access. Every Activity/Tool can have either **read** or **execute** permissions, or be **locked** to the user. **Execute** defines the ability to create, configure, and run an activity, tool, or menu. **Read** access allows the running of a created job, and extends to allow some work within the results of jobs. Additionally, every **Data Custodian** or **Data Steward** can be granted the **Administrator** role, granting access to Data Quality meta-tasks. The following table summarizes Roles and their Permissions.

Roles and Permissions	Data Custodian	Data Steward
Project Administration	execute	read
Project Tools	execute	locked
Unit of Measure Aliases	execute	read
Test Data	execute	read
Rapid HealthCheck	execute	read
Detailed HealthCheck	execute	read
Rapid Clean	execute	read
Detailed Clean	execute	read
Clean Application References	execute	read
Make Values Consistent	execute	read

Roles and Permissions	Data Custodian	Data Steward
Detailed Match	execute	read
Manage Duplication	execute	read
Setup and Manage Registry	execute	read
Setup and Manage Alias Set	execute	read
MasterSet HealthCheck	execute	read
Manage Master Records	execute	read
Merge Setup	execute	locked
Merge	execute	read
Advanced Scheduling	execute	read
Job Administrator	execute	read
Requirements Administrator	execute	locked
Reference Data Administrator	execute	locked
Unit of Measure Administrator	execute	locked
Regular Expression Helper	execute	execute
Manage Users	Admin task	Admin task
Manage Project Access	Admin task	Admin task
Manage Connection Access	Admin task	Admin task

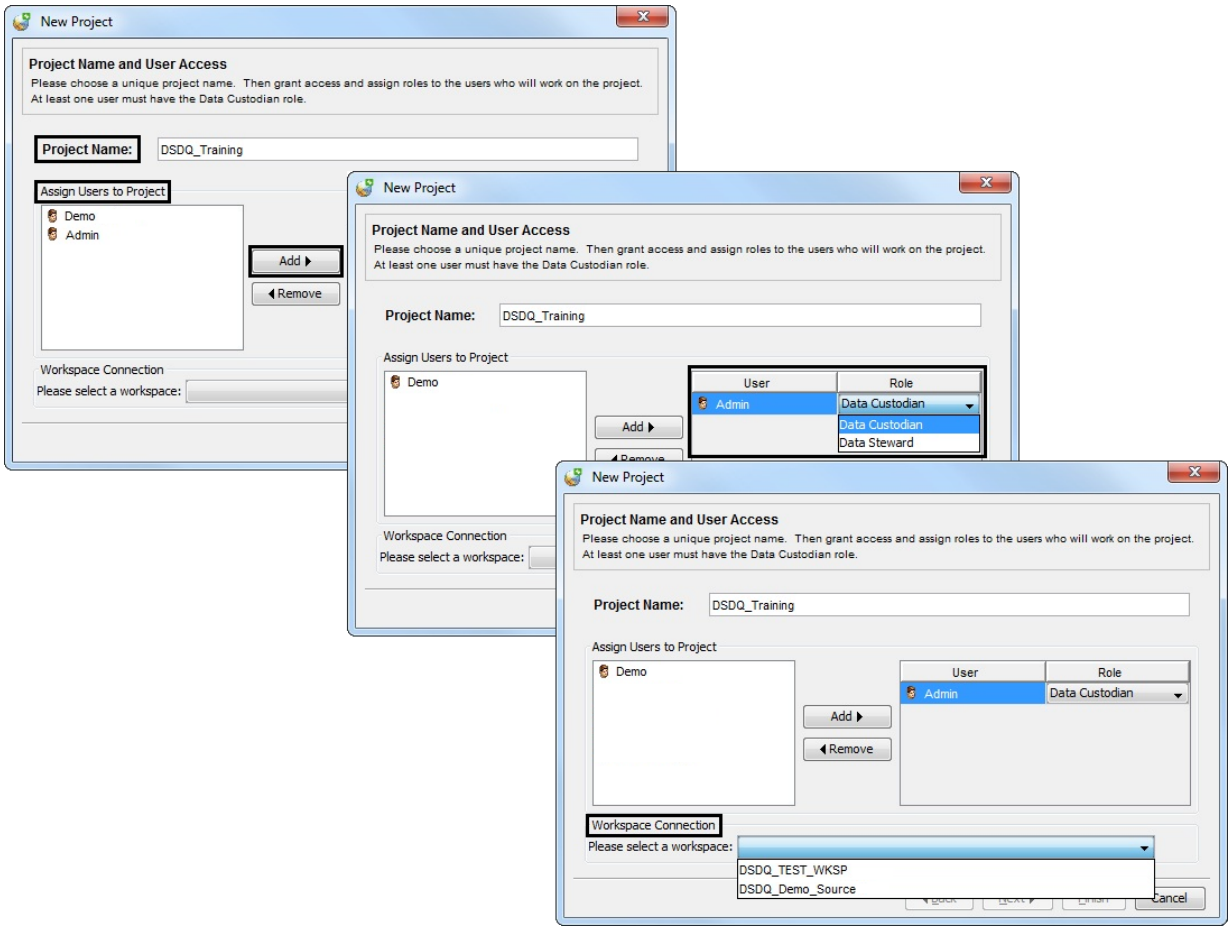
Exercise: Creating a Project

In this exercise, you will be creating a new project in DSDQ, assign user and role under a workspace previously created and add a HealthCheck Phase to the project.

A new project can be created in any of the following three ways:

- When the software is initially installed and no projects exist
- By selecting **New Project** from the **File** menu
- By clicking the **New** button in the **Open an Existing Project** window

1. Select **File > New** from the menu bar on the **DSDQ Project Window**.
   The **New Project** window appears.



2. Enter **DSDQ_Training** in the **Project Name** field.

> **Note**
>
> A project name must be less than or equal to 50 characters.

3. Select **Admin** from the **Assign Users to Project** group box. Access to the project will be given to selected users only.

4. Click the + Add button to assign project access to the selected user.
   You can only select one user at a time. Repeat steps **3** and **4** for each user that you create. The selected users i.e. **Admin** is added to the box on the right side of the **Assign Users to Project** group box.

5. Assign the **Admin** user a role in the project by selecting it from the Role drop-down list. Options available for selection include **Data Custodian** and **Data Steward**. Select the **Data Custodian** role for the **Admin** user.

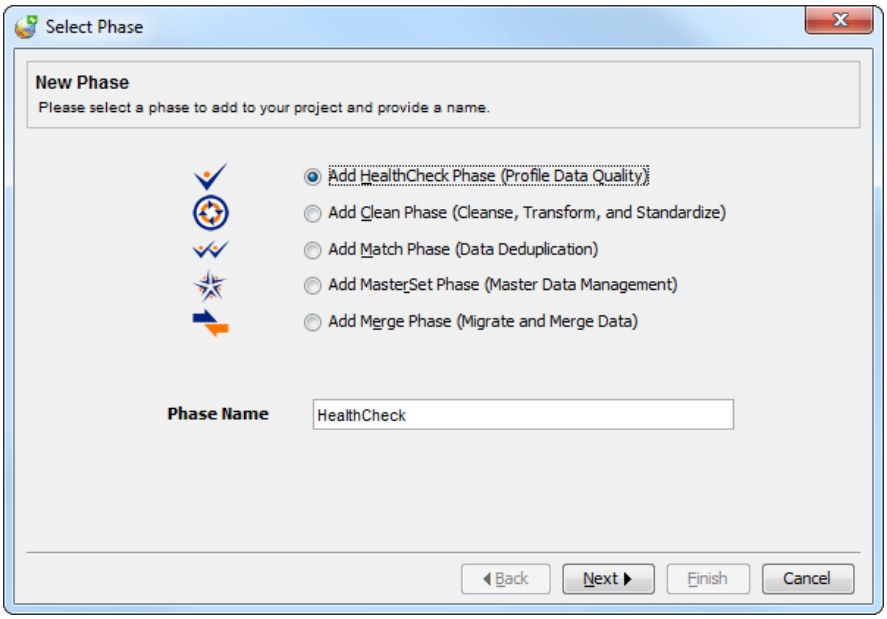
6. Select **DSDQ_TEST_WKSP** from the **Please select a workspace** drop-down list.

> **Note**
>
> A workspace database is a repository where results will be written. Ensure that the workspace database that you connect to this project (i.e., **DSDQ_TEST_WKSP**) is not used by other Data Quality projects and phases.

Supported Databases for **Workspace** connection include:

- Oracle
- MS SQL Server
- Apache Derby

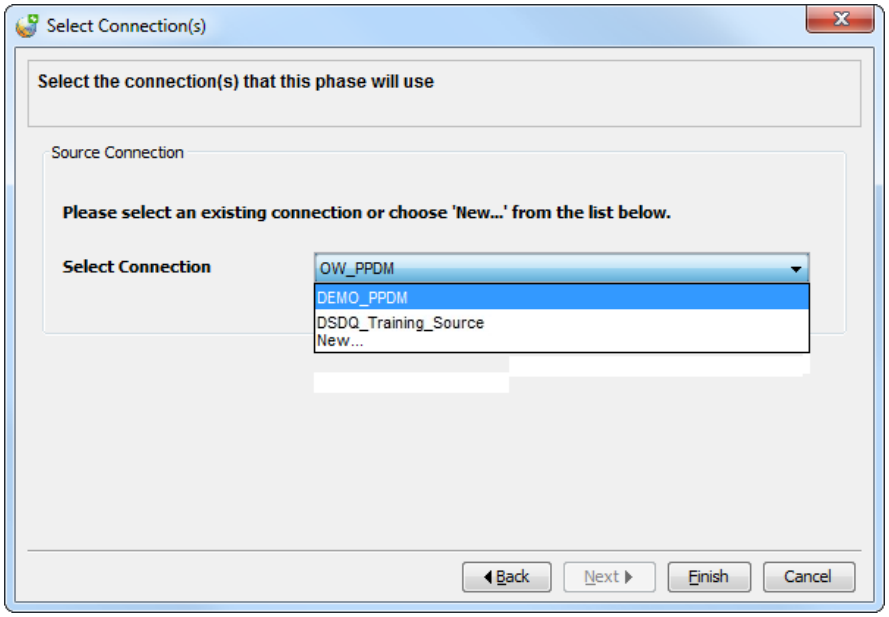
7. Click **Next** on the **New Project** window to continue.
The **Select Phase** window appears with the **Add HealthCheck Phase (Profile Data Quality)** option selected by default.



8. Select the desired phase to add it to your project, if different from the one selected by default.

9. Optionally, change the description of the selected phase by entering a brief narrative for it in the **Phase Name** field. The phase name must be unique, and can be renamed after being created.

10. Click **Next** to continue.
    The **Select Connection(s)** window appears.



11. Select **DSDQ_Training_Source** as the data owner connection (the data source that the application reads from) from the **Select Connection** drop-down list.

> **Note**
>
> For more information on data owner connections, refer to Adding a New Data Owner Connection section in Chapter 2. Connecting DecisionSpace Data Quality with DecisionSpace Data Server.

12. Click **Finish** to complete the process of creating a new project in DSDQ. Once created, this project will be added to the Data Quality Tree on the **DSDQ Project Window**.

> **Note**
>
> The Data Quality data model that holds the data source objects for a particular connection is automatically created and given the same name as that of the connection.

Working with Projects

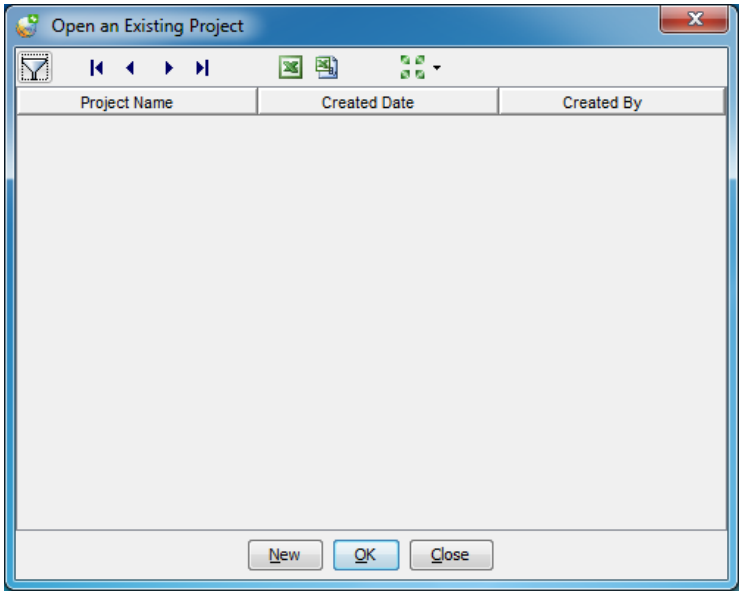
This section provides you information about opening existing projects, deleting unwanted projects and transferring projects to and from the application.

The functionality to import and export projects in DSDQ enables you to:

- Transfer full project configurations between the Data Quality installation
- Ensure a backup mechanism is in place for a project's connection, model and service level configurations

Exercise: Opening an Existing Project

1. Select **File > Open Project** from the menu bar on the **DSDQ Project Window**.
   The **Open an Existing Project** window appears with the most recent project i.e. **DSDQ_Training** selected by default.



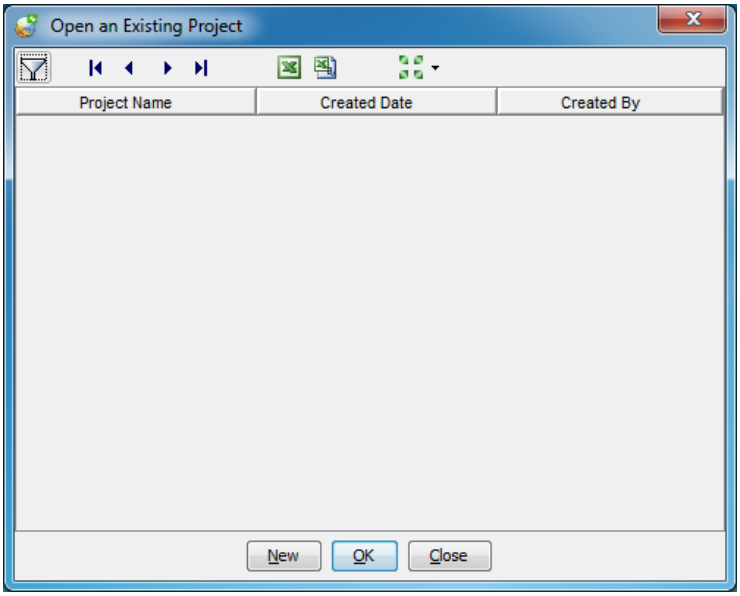
2. Click **OK** to open it.

> **Note**
>
> The most recent project will open automatically when you launch the application. In case the project that you want to open is not the one created recently, you will select it from this window and click OK to open it.

Exercise: Exporting a Project

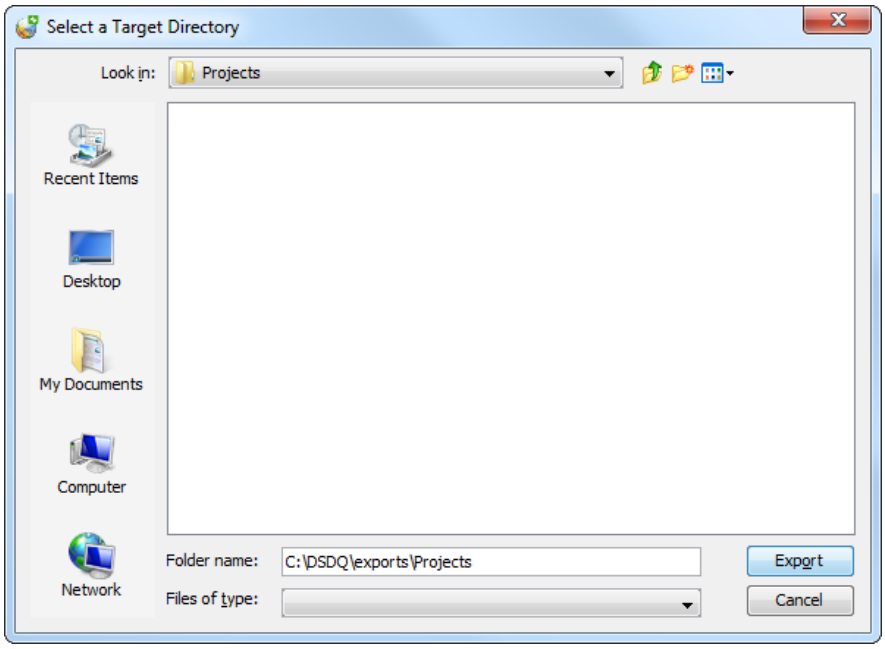
Project Export allows you to transfer full project configuration between new Data Quality installations, as well as providing you with a backup mechanism for your connection, model, and service level configurations.

1. Select **File > Open Project** from the menu bar on the **DSDQ Project Window**.
   The **Open an Existing Project** window appears with the most recent project i.e. **DSDQ_Training** selected by default.



2. Select the **DSDQ_Training** project from the list.

3. Click the arrow on the **Import/Export Project** icon and select the **Export Selected Project(s)** option from the drop-down menu.
   The **Select a Target Directory** window appears.

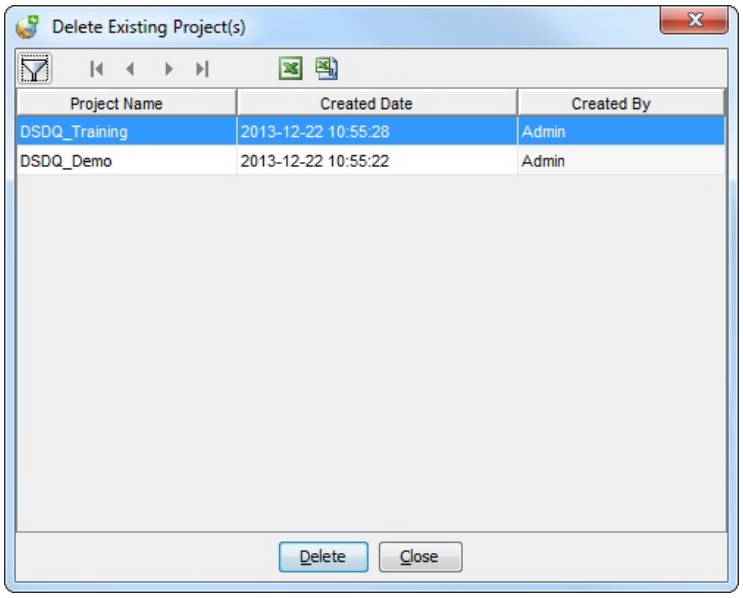


4. Select an export project path and click **Export**.
   The selected project is exported at the desired location.

Exercise: Deleting a Project

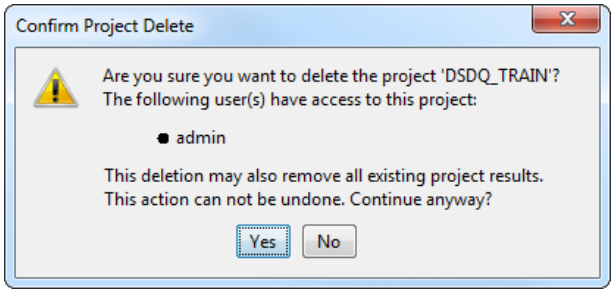
Deleting a project removes all the data associated with it as well. Before you delete a project, ensure you have the required access privilege in DSDQ to do so i.e. **Administrator** rights.

1. Select **File > Delete Project** from the menu bar on the **DSDQ Project Window**.
   The **Delete Existing Project(s)** window appears with the most recent project i.e. **DSDQ_Training** selected by default.



> **Note**
>
> In case the project that you want to delete is not the one created recently, you will select it from this window and then click **Delete** to remove it.

2. Click **Delete**.
   The **Confirm Project Delete** message box appears listing all users associated with the project.

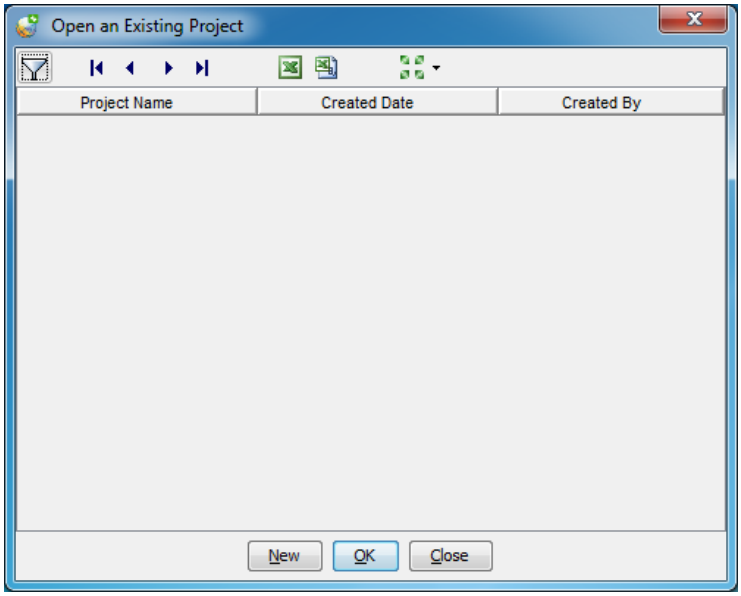


3. Click **Yes** to continue.

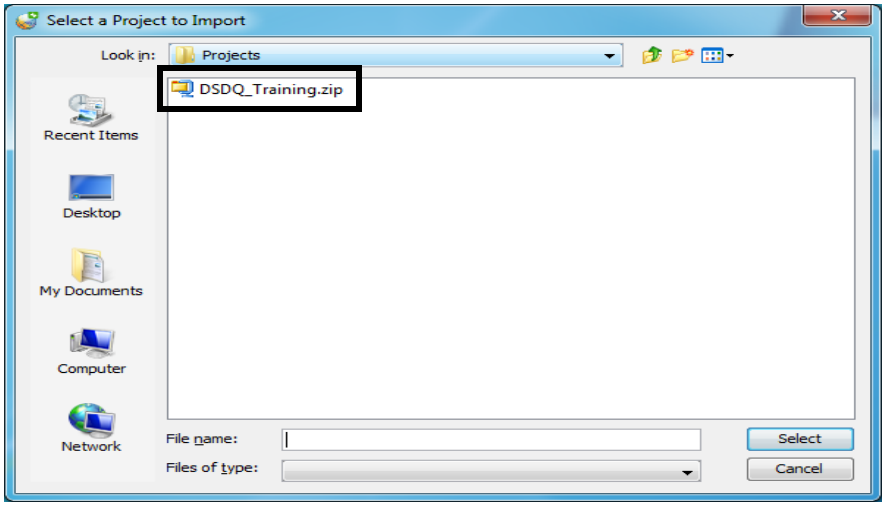
Exercise: Importing a Project

Project Import allows you to load full project configuration on a new Data Quality installation. Before you import a project, ensure you have the required access privilege in DSDQ to do so i.e., **Administrator** rights.

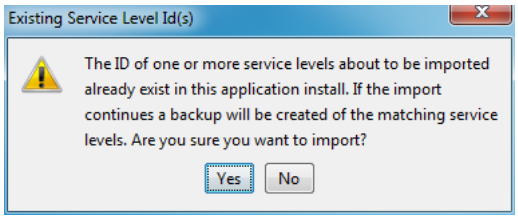
1. Select **File > Open Project** from the menu bar on the **DSDQ Project Window**.
   The **Open an Existing Project** window appears.



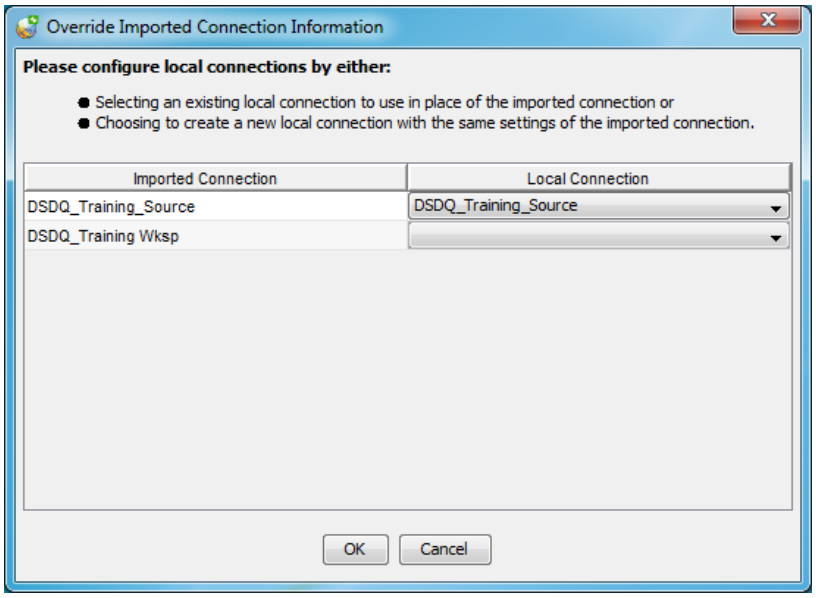
2. Click the arrow on the Import/Export Project icon and select the **Import Project** option from the drop-down menu.
   The **Select a Project to Import** window appears.



3. The Existing Service Level Id(s) message box displays, prompting you to confirm whether you want to import this project even though there are already matching service levels. Click **Yes**.



4. Select the **DSDQ_Training** project file and click **Select** to display the **Override Imported Connection Information** window.



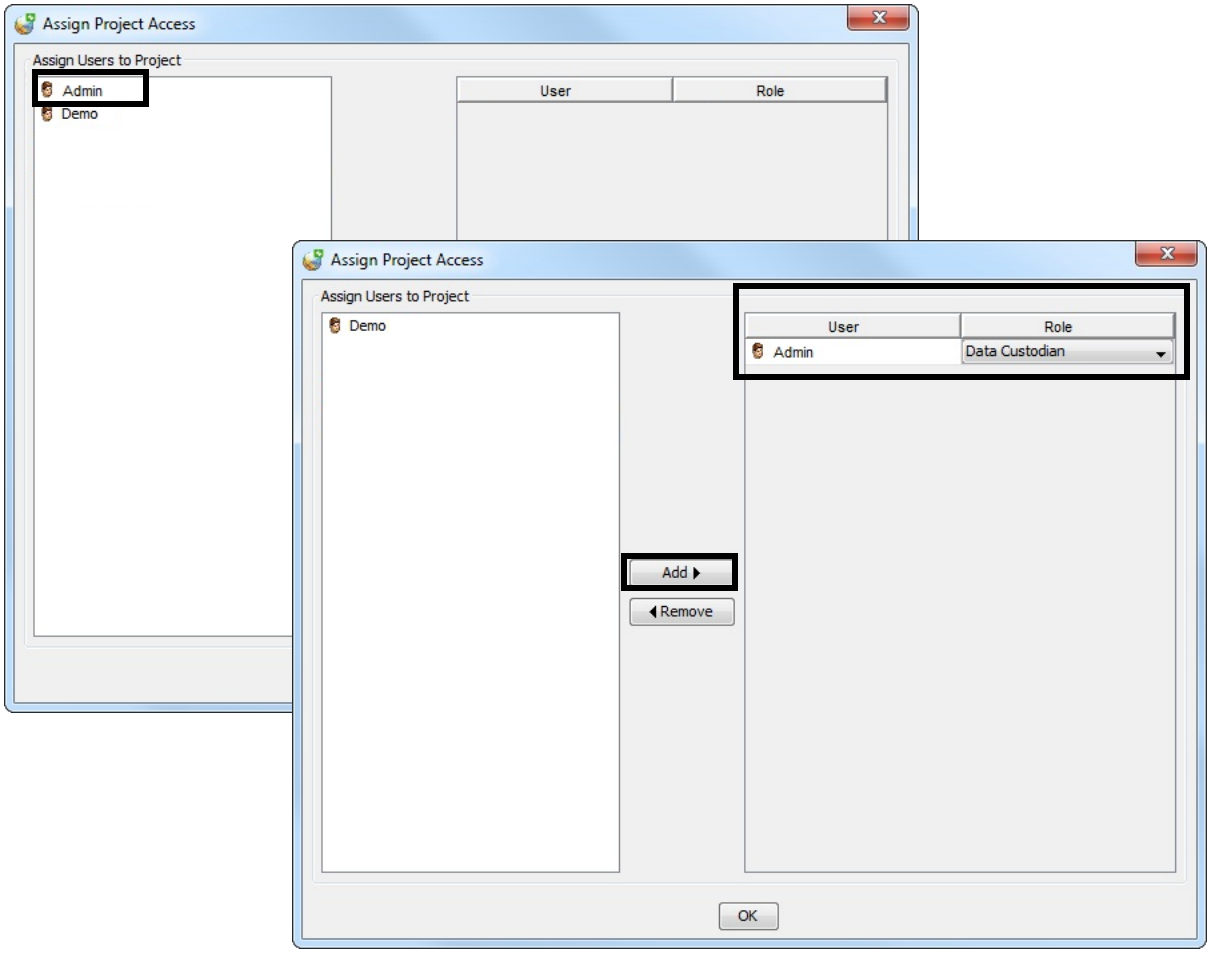
Note

If the selected project export files match a project that already exists in the current Data Quality install, you are prompted if you want to backup that project in order to proceed with the new import process. Clicking Yes causes an export file backup of the existing project to be automatically created in the %DSDQ_HOME%/exports/Projects folder with the same <project name> 1.zip. That project is then deleted from the current Data Quality install and the new one is imported. Clicking No aborts the import process.

Connection information for the imported project is selected from here. Connections under the **Imported Connection** column represent connections coming from the imported project. The **Local Connection** column lists connections available in the current Data Quality install. These can be used by the imported project. By

selecting a connection from the **Local Connection** column, once the import process is complete, the imported connection shares connection details with the selected local connection. Select **New…** from the **Local Connection** drop-down list to create a new connection for the imported project. For more information on new connections, refer to Creating Connections section in Chapter 2. Connecting DecisionSpace Data Quality with DecisionSpace Data Server.

5. Click **OK** to continue the import process.
   The **Assign Project Access** window appears. This window allows you to assign the desired Data Quality user access to the imported project. Users given access to an imported project will gain access to all of its connections.



6. Select **Admin** from the **Assign Users to Project** group box. Access to the project will be given to selected users only.

7. Click the Add button to assign project access to the selected user.
   The selected user i.e. **Admin** is added to the box on the right side of the **Assign Users to Project** group box.

8. Assign the **Admin** user a role in the project by selecting it from the Role drop-down list. Options available for selection include **Data Custodian** and **Data Steward**. Select the **Data Custodian** role for the **Admin** user.

9. Click **OK** to complete the project import process.
   Once the project has been successfully imported, an Import Validation Warning may appear. This warning prompts you of possible, noncritical issues encountered during the import process. These issues are stored in the client.log file available at ***Landmark/Decisionspace Data Quality/logs***.

Chapter 4
Data Evaluation

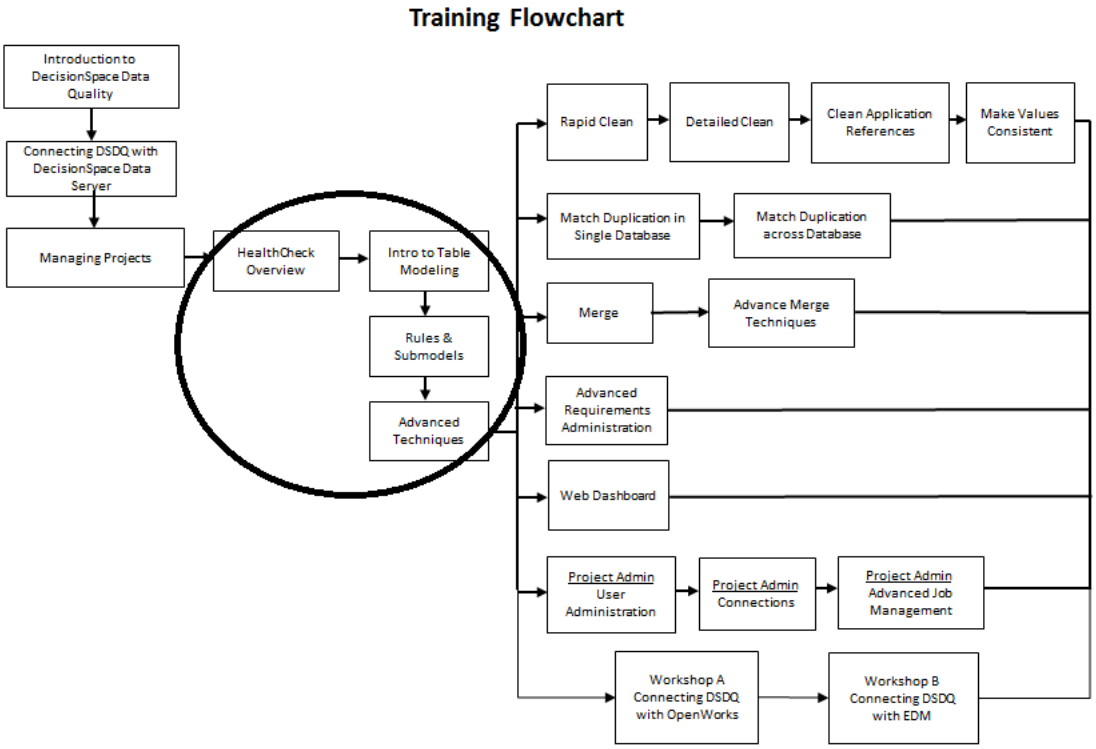
The first step in a successful data quality improvement project begins with data profiling, which uncovers the “where”, “what”, and “why” problems within your valuable data assets. Data profiling identifies the full spectrum of data issues, including incomplete, inaccurate, inconsistent, missing or ambiguous information. HealthCheck provides a comprehensive look at the quality of the source dataset. The **HealthCheck** Phase is composed of the **Rapid HealthCheck** and **Detailed HealthCheck** Activities. The series of reports that are produced by these activities assist with the completion of data quality analysis.

Chapter Overview

In this chapter, you will learn about:

- Performing table modeling
- Evaluating data volume and quality
- Identifying data issues

Topics covered in each chapter are outlined in the following illustration. Those specific to the current chapter will be circled in black for your reference:



Perform Table Modeling

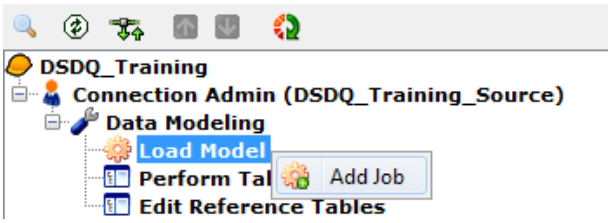
Table modeling defines the table hierarchy for the source database. All tasks within the Data Quality application are dependent on the information provided within table modeling.

Exercise: Loading a Data Model

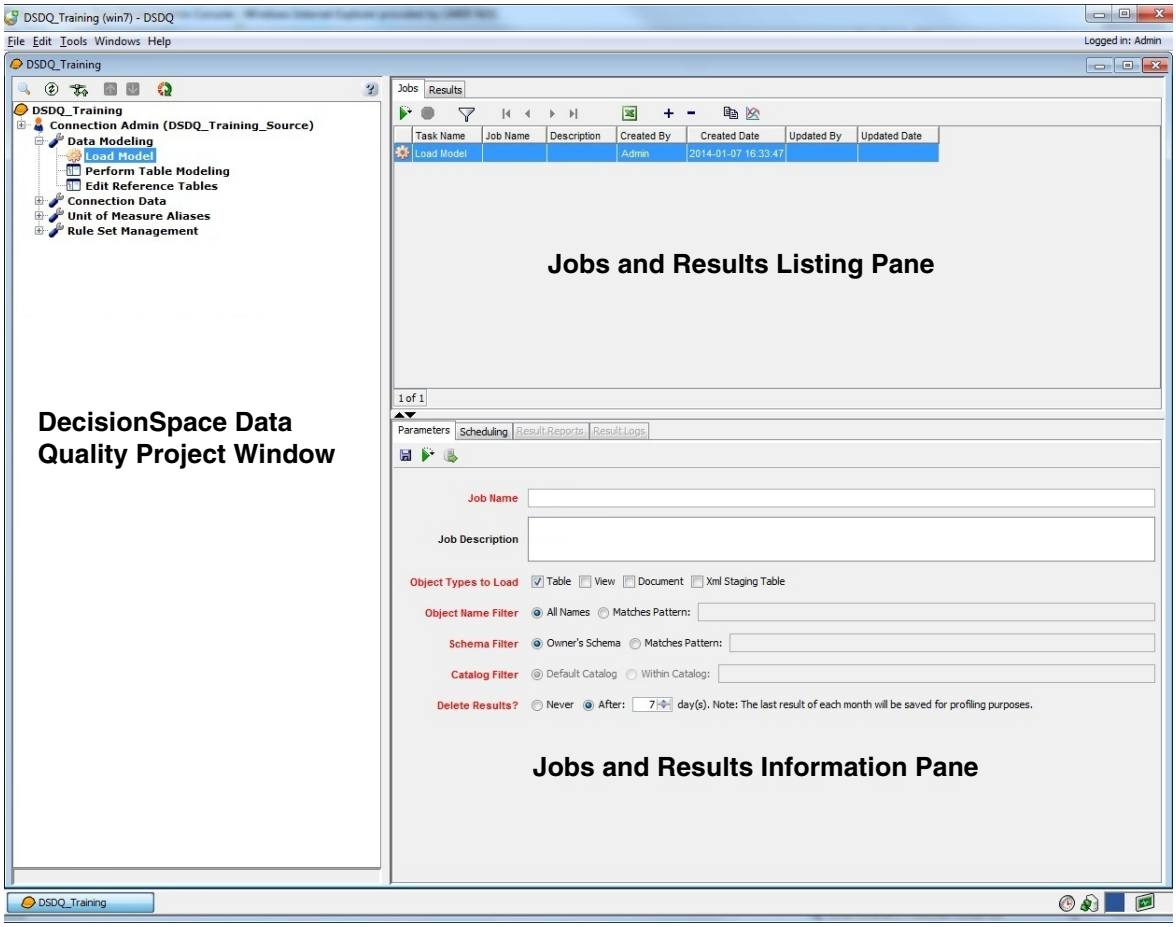
To Load a Data Model:

1. Click ⊞ on the DecisionSpace Data Quality Tree to expand the **Data Modeling** Activity.

2. Double-click the **Load Model** Task or right-click the **Load Model** Task and select **Add Job** from the pop-up menu.



A new job is initiated and displays on the **Jobs and Results Listing Pane**.



3. Enter **Job-01** in the **Job Name** field.

4. Enter **Load Model** in the **Job Description** field.

5. Select the **Table** option for **Object Types to Load**.

6. Select the **All Names** option for **Object Name Filter**.

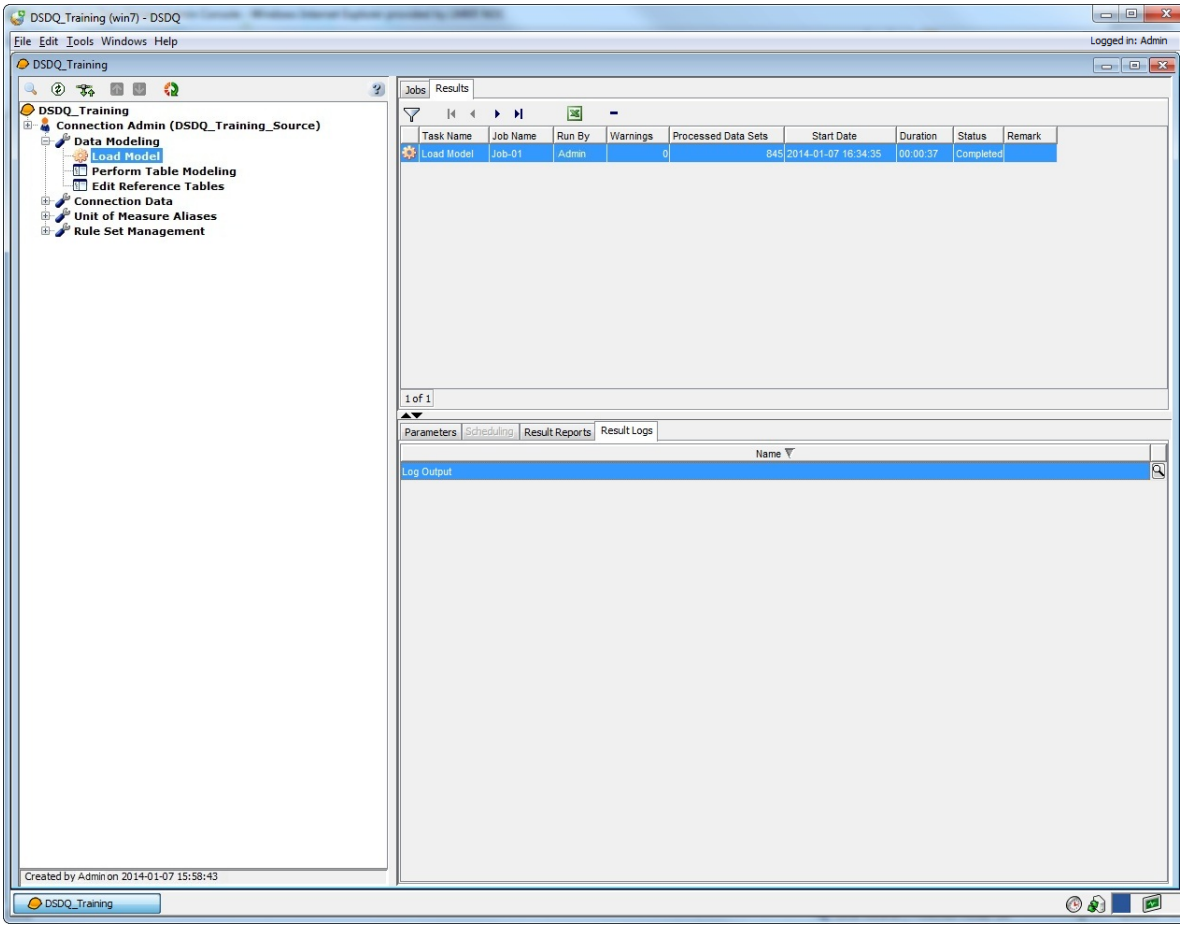
7. Select the **Owner's Schema** option for **Schema Filter**.
   The **Catalog Filter** field will remain unchanged.

8. Select the **After** option for **Delete Results?** Leave the number of days to **7**.

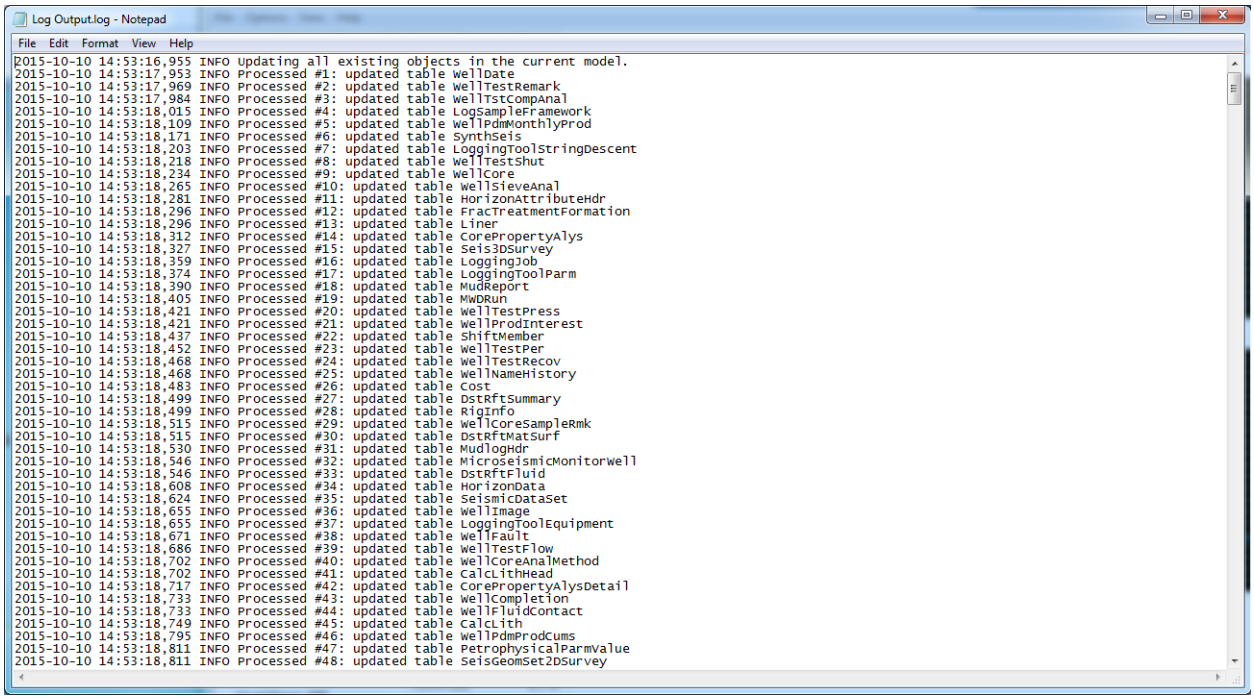
9. Click to save changes in the **Parameters** tab.

10. Click .
    The **Load Model** task runs and displays results in the **Result** Logs tab of the **Job and Results Information Pane.**

11. Select the **Results** tab on the **Job and Results Listing Pane** to view the values in the **Result Logs** tab on the **Job and Results Information Pane**.



12. Double-click on **Log Output** on the **Jobs and Results Information Pane** to display the **Log Output** in TXT format.

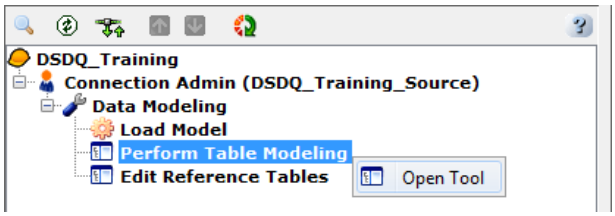


13. Click **File > Exit** to close the log file.

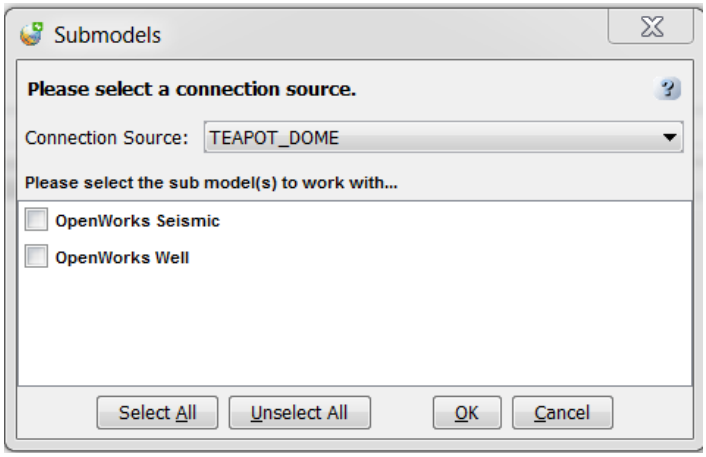
Exercise: Perform Table Modeling

To perform Table Modeling:

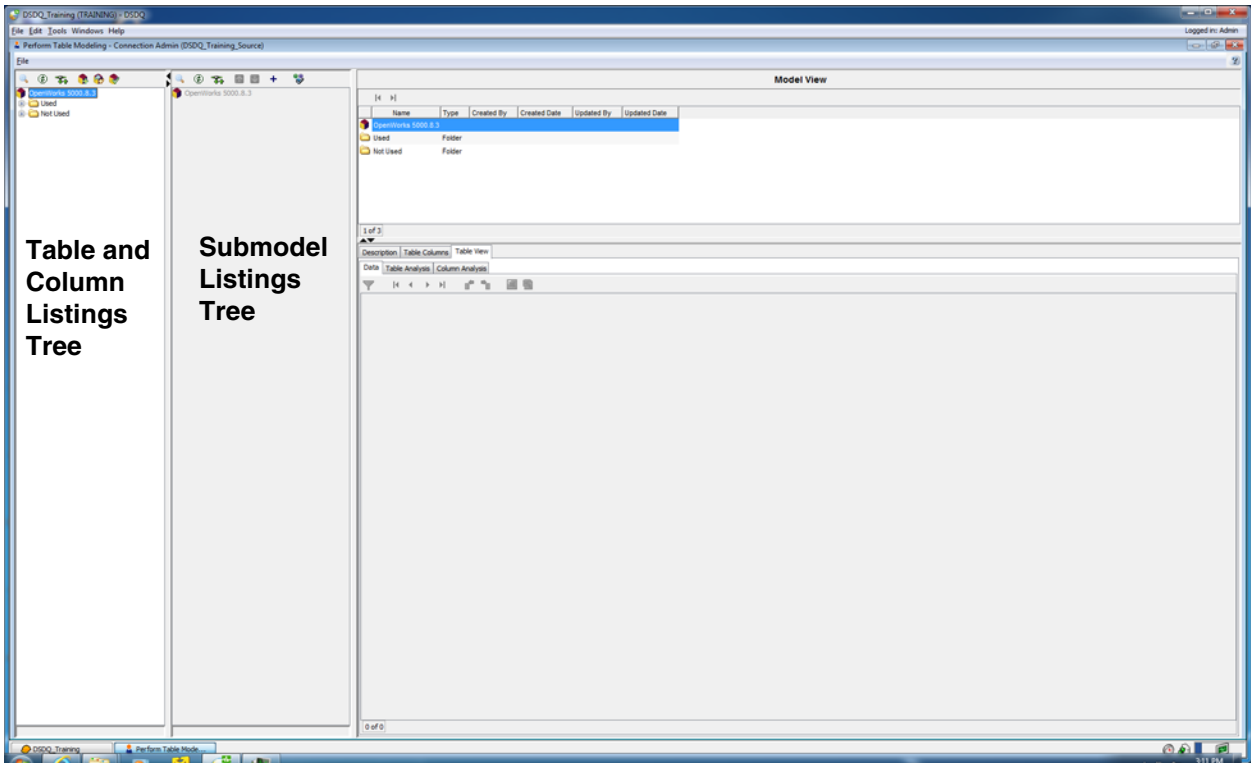
1. Double-click the **Perform Table Modeling** tool or right-click the **Perform Table Modeling** tool and select **Open Tool** from the pop-up menu.



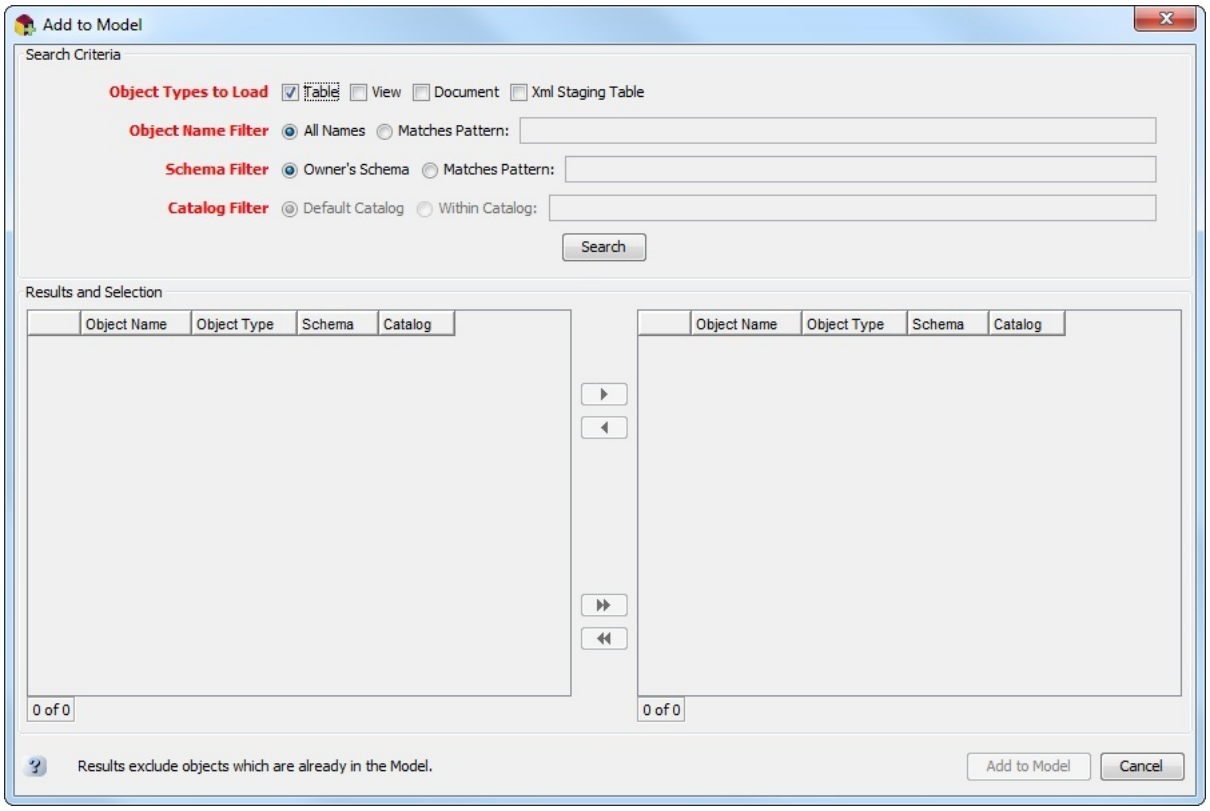
The **Submodels** dialog box displays. Select the **TEAPOT_DOME** project from the **Connection Source**. Clear the selection of the quickstart OpenWorks submodels (**OpenWorks Seismic** and **OpenWorks Well**) and click **OK**.



The **Perform Table Modeling** screen appears.



2. Click to add database objects to the model.
   The **Add to Model** window appears.



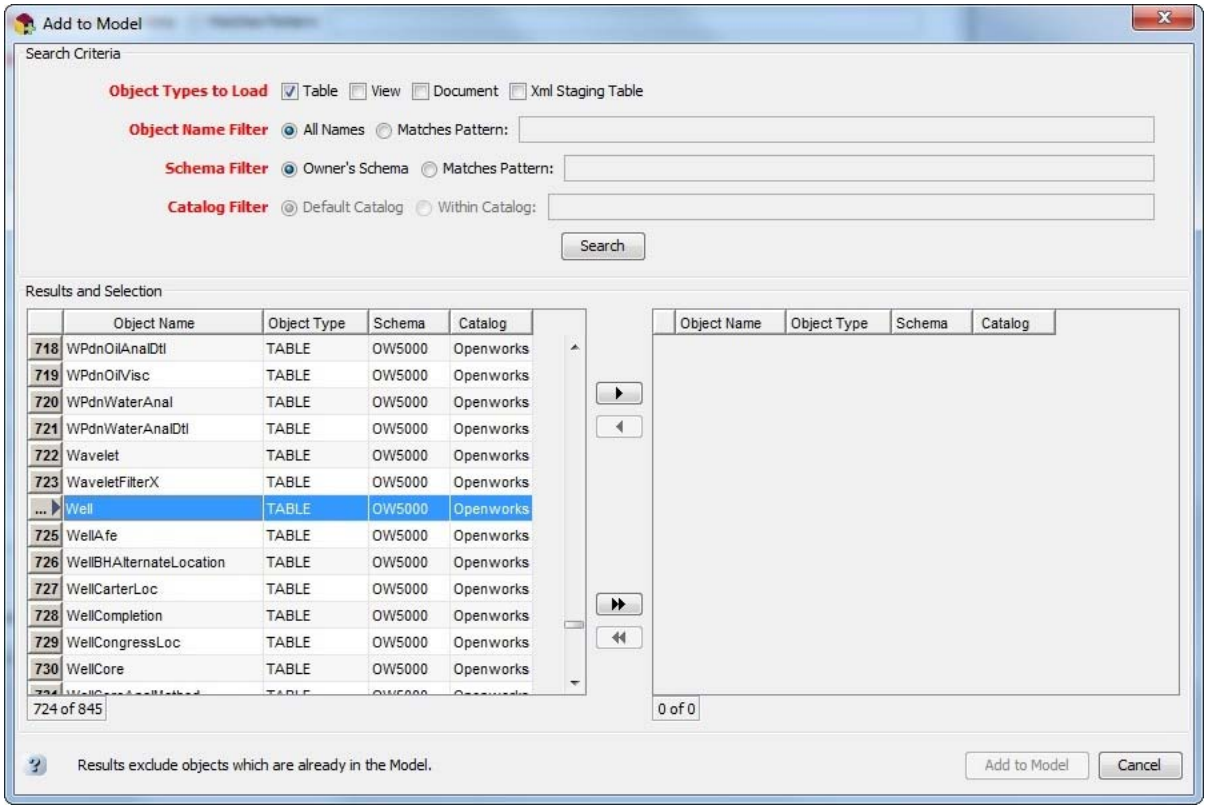
3. Select the **Table** option for **Object Types to Load**.

4. Select the **All Names** option for **Object Name Filter**.

5. Select the **Owner's Schema** option for **Schema Filter**.

6. The **Catalog Filter** options will be disabled for the current selection.

7. Click **Search**.
   The search results appear based on your search criteria.

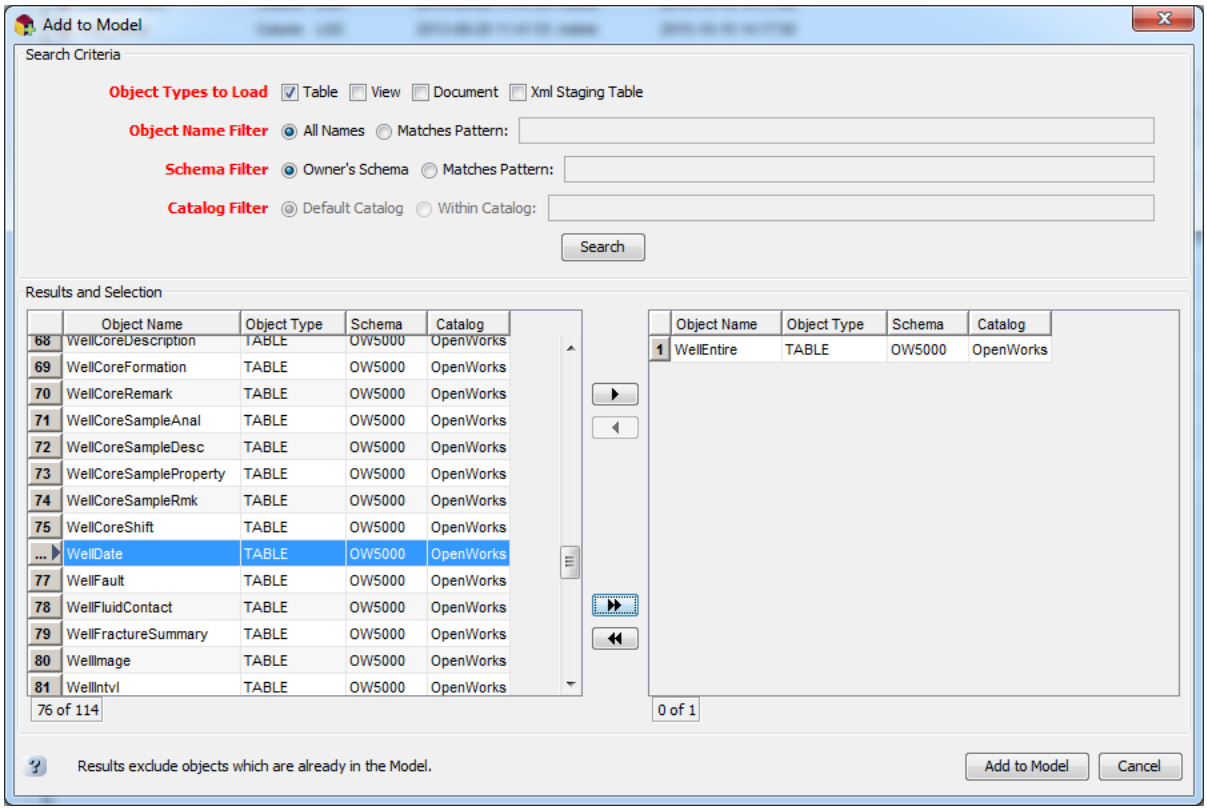


8. Browse through the database objects and select **Well**.

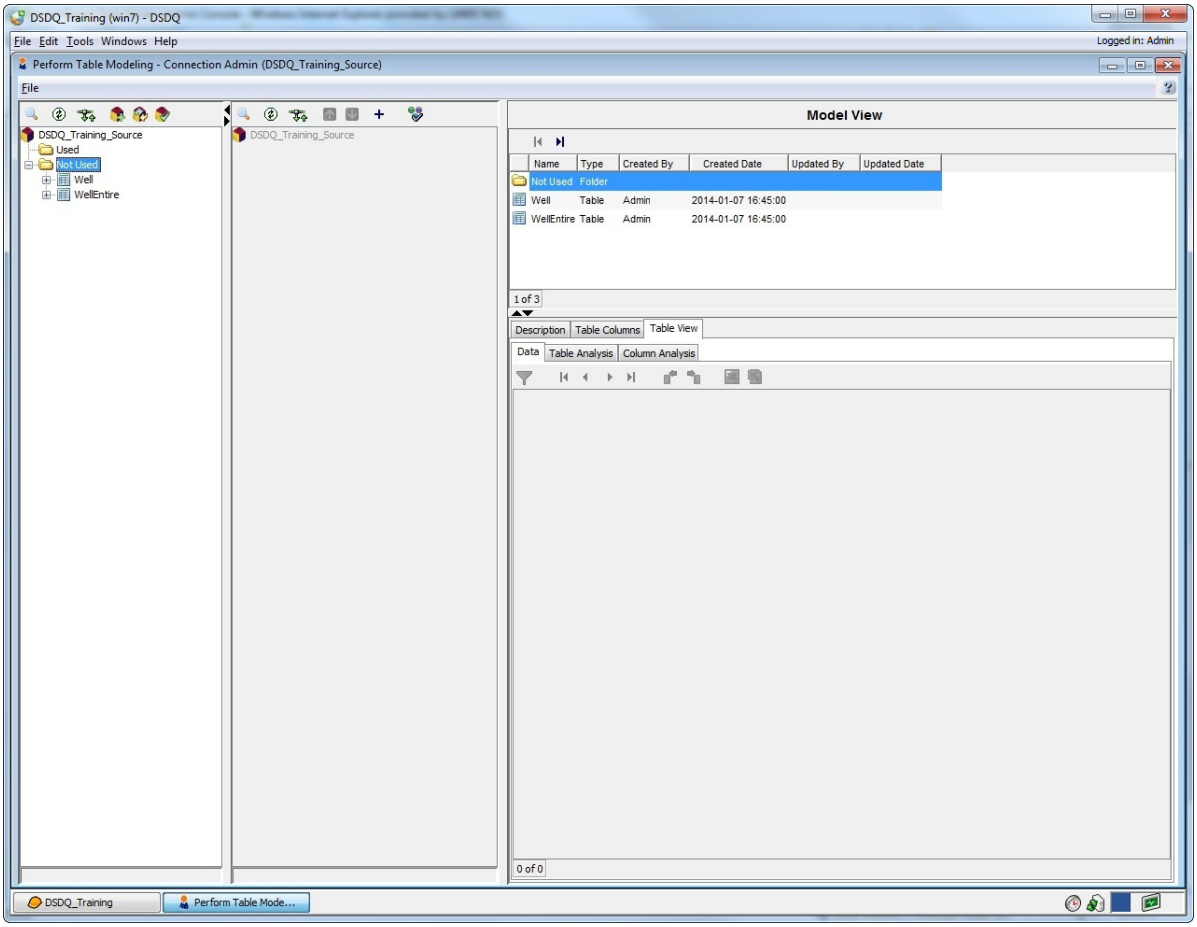
> **Note**
>
> Press <Ctrl> key, and click on the required objects to select them.

9. Click ▸ to send selected objects to the table on right side.

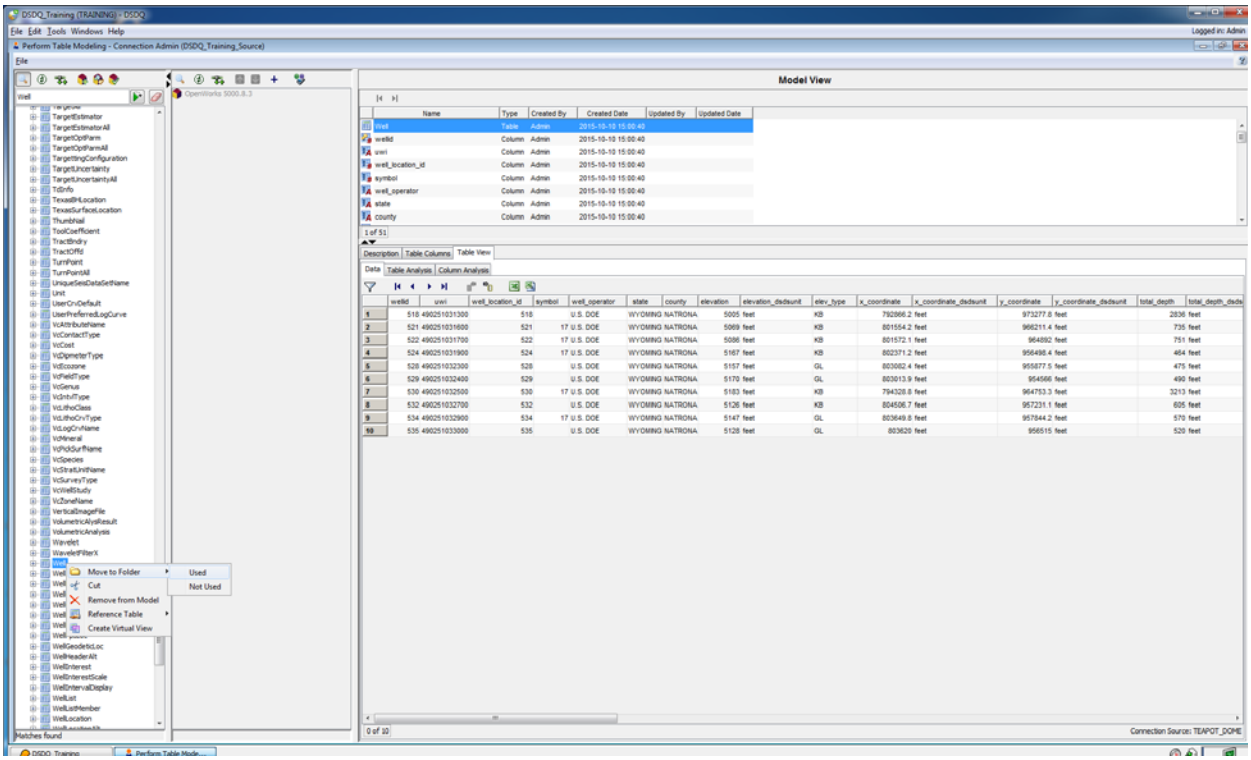
10. Click **Add to Model** to add selected database objects to the model.



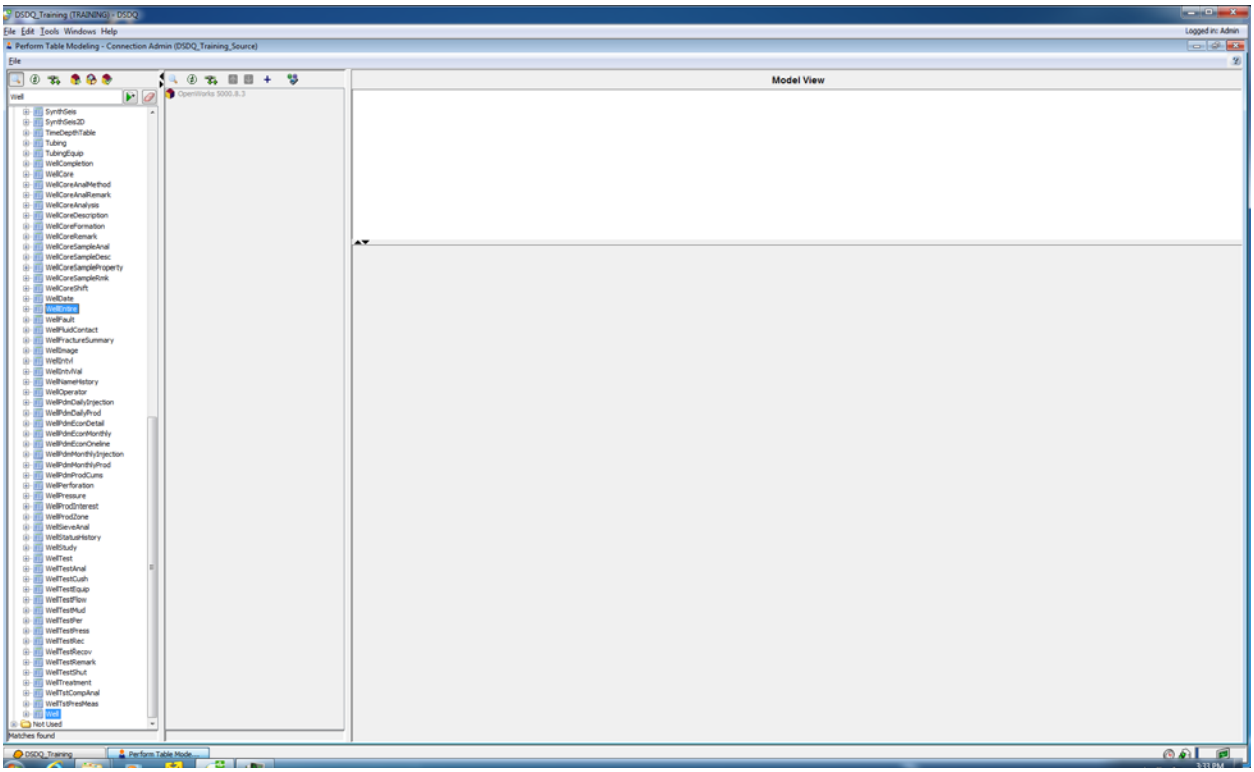
Search for **Well** and **WellEntire** in the **Table and Columns Listings Tree**. If the database objects appear under the **Not Used** tree, right-click and select **Move to Folder > Used**.



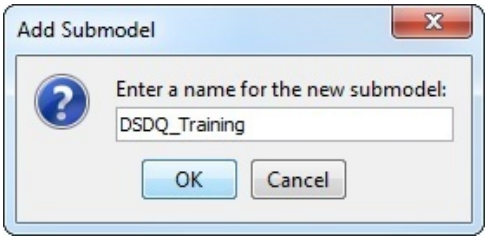
11. Select the database objects and right-click them. Select the **Move to Folder > Used** option from the pop-up menu.



The selected objects now appear under the **Used** tree.

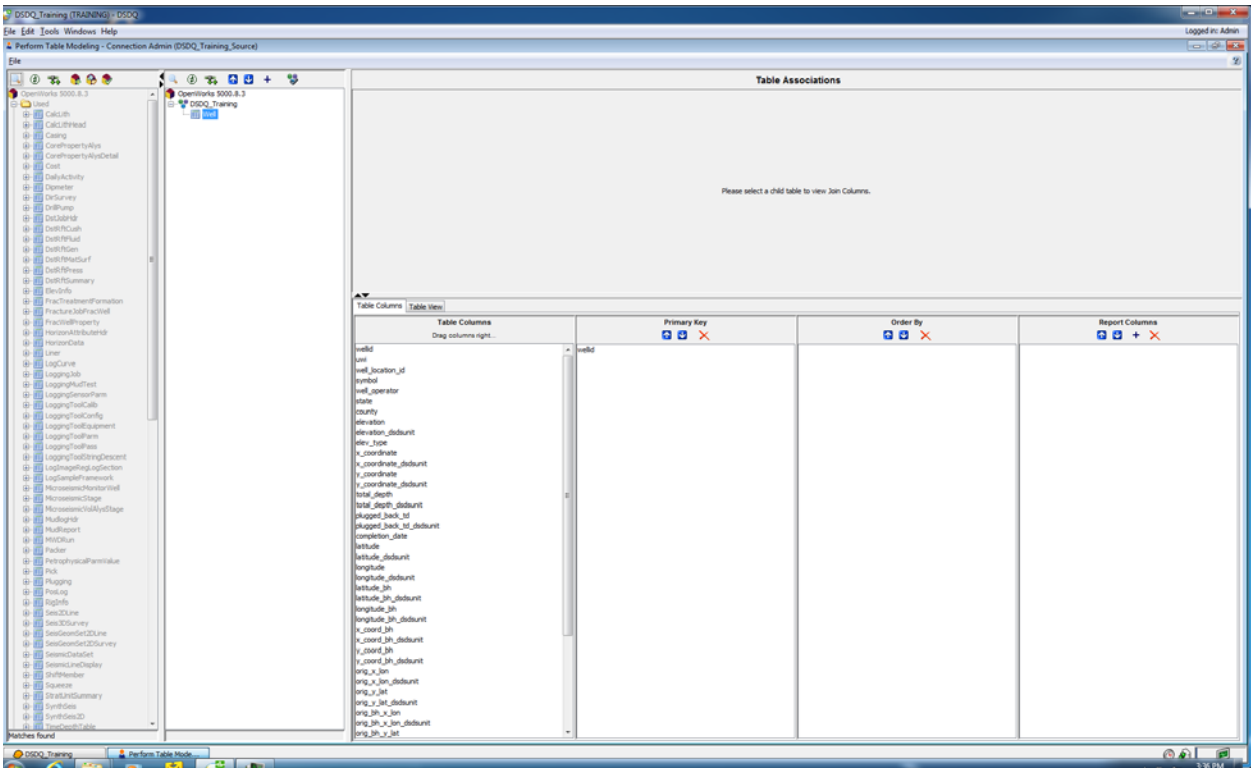


12. Click ⊞ adjacent to **Well** on the **Submodel Listings Pane**. The **Add Submodel** dialog box appears.



13. Enter **DSDQ_Training** in the **Enter a name for the new submodel** field.

14. Click **OK**.



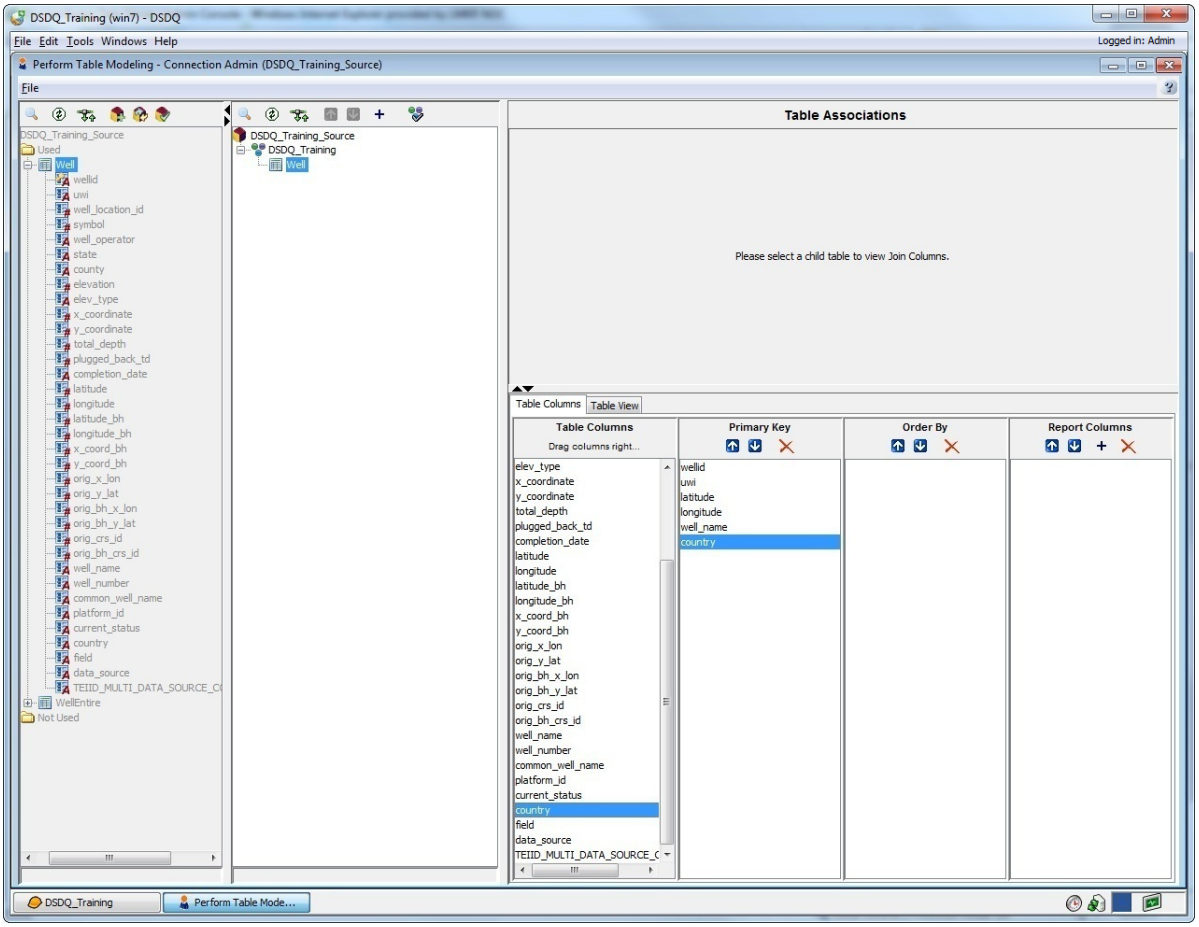
15. Drag the **Well** table from **Table and Column Listings Tree** to **Submodel Listings Tree,** under the newly created submodel **DSDQ_Training.**

16. From **Table Columns** pane, select the following objects:

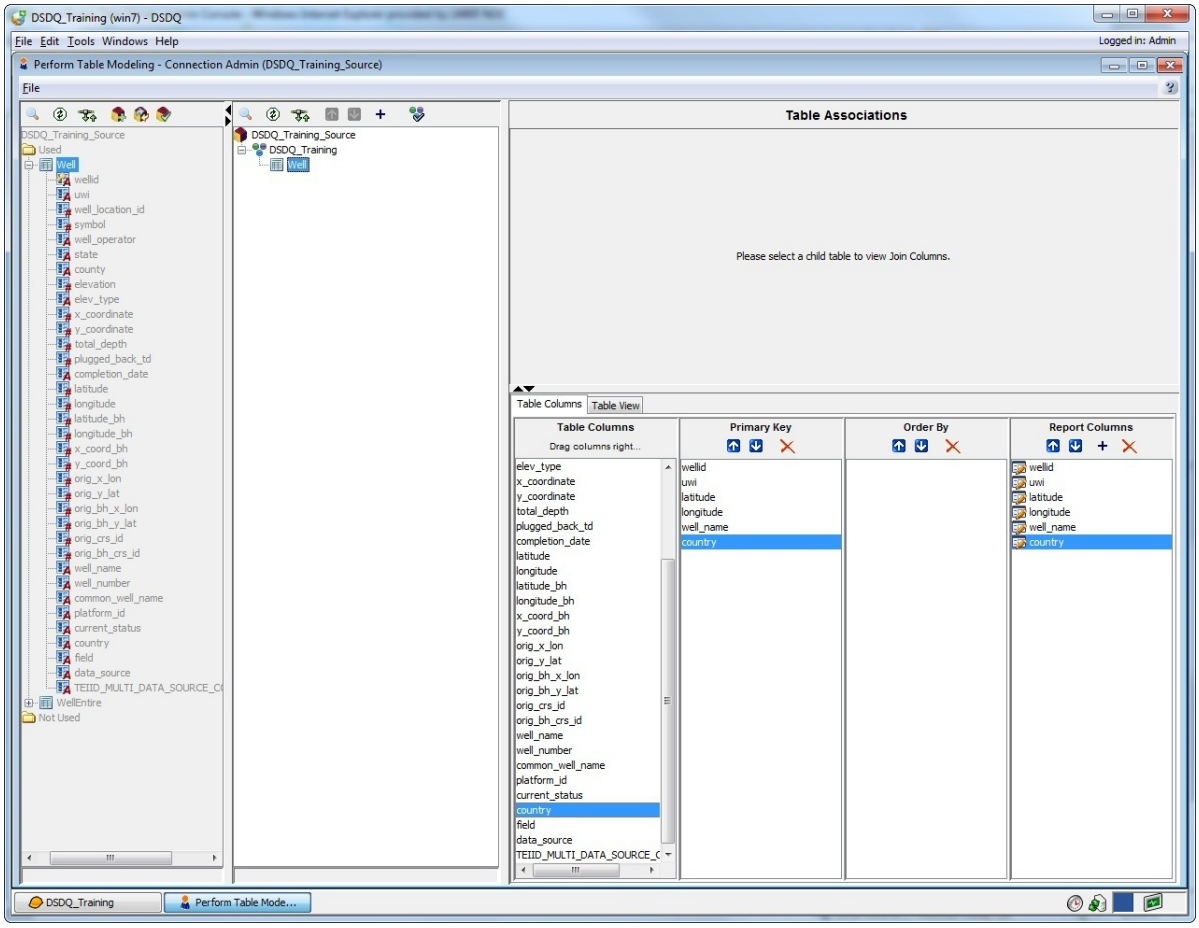
    - wellid
    - uwi

- latitude
- longitude

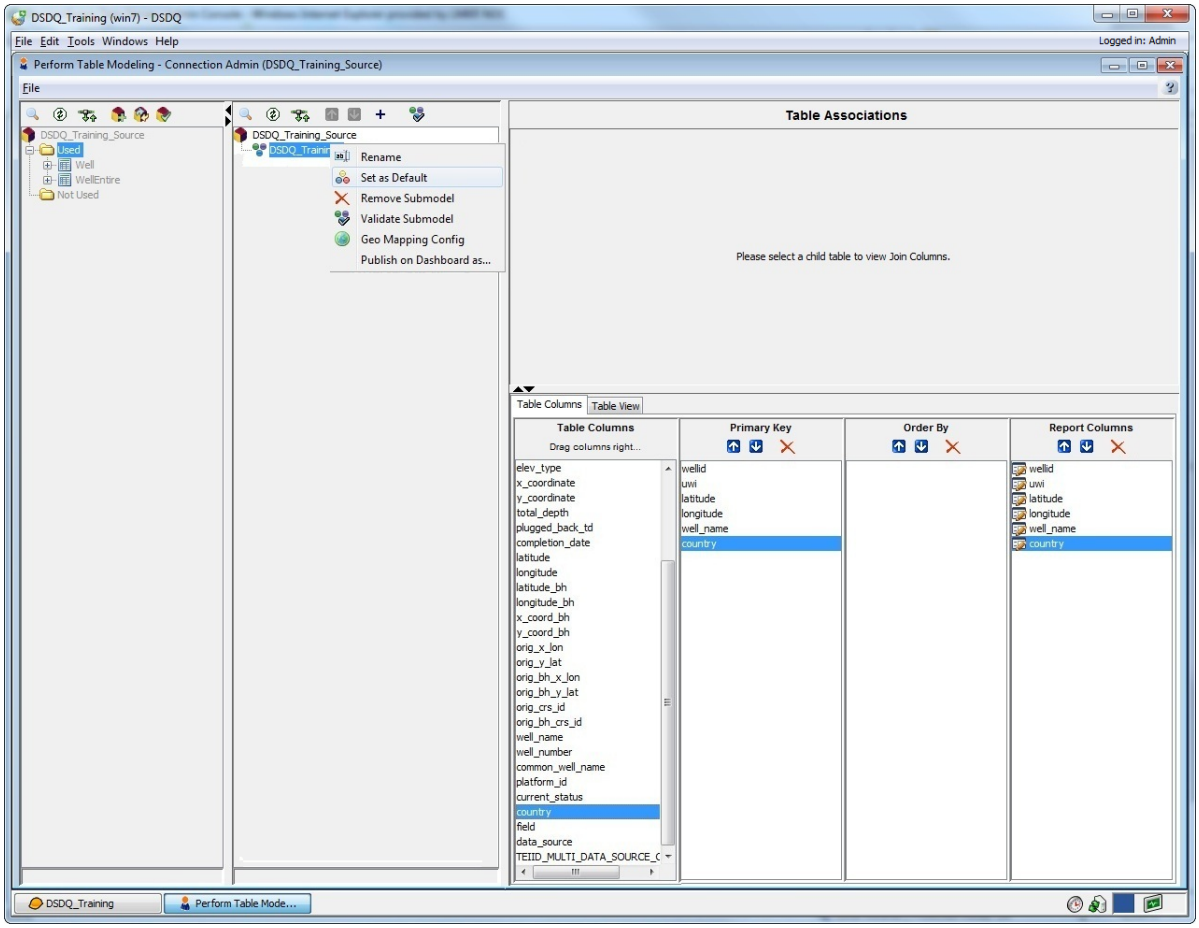
17. Drag all the selected objects to the **Primary Key** pane.



18. Similarly, move all these objects to the **Report Columns** pane as well.

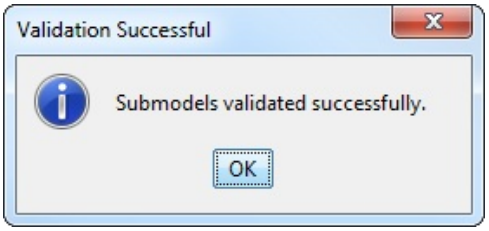


19. Right-click on the **DSDQ_Training** submodel and select **Set as Default** from the pop-up menu.



The submodel's icon will change from to .

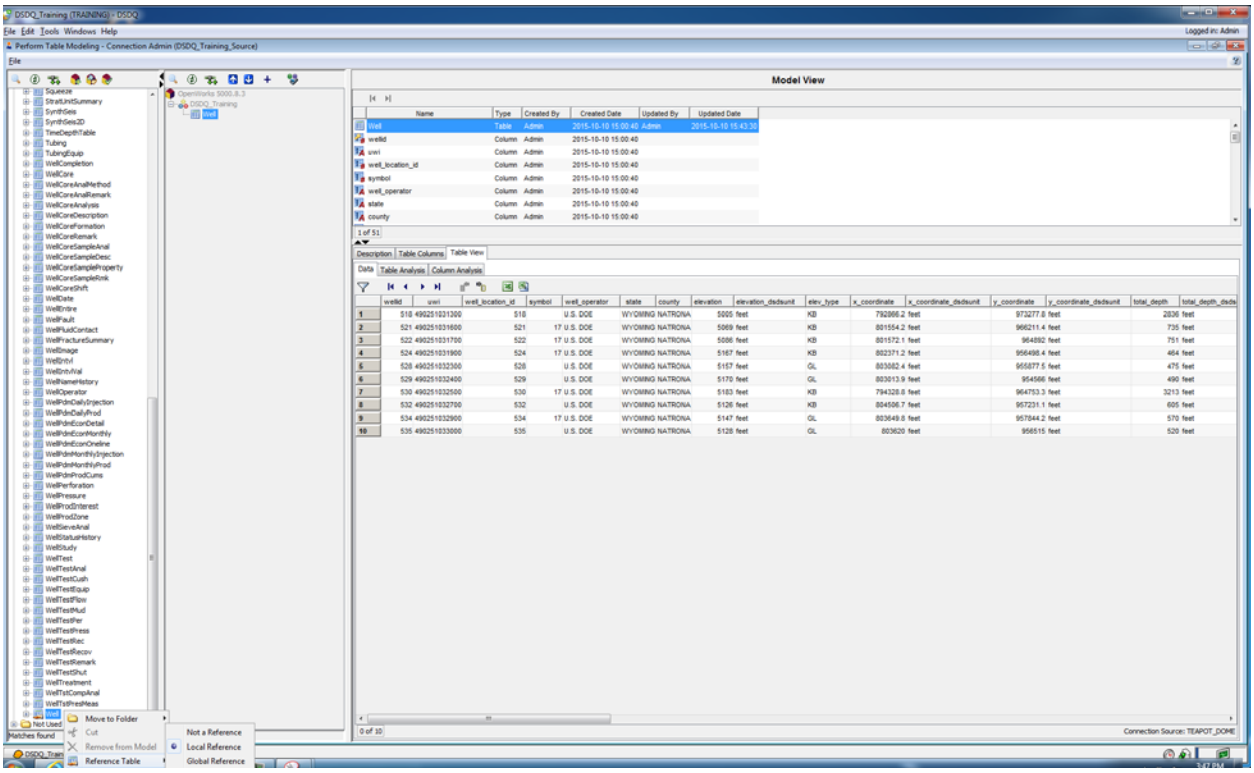
20. Click to validate the submodel.



21. Click **OK**.

22. Right-click on the **Well** table in the **Table and Column Listings Tree**, and select **Reference Table > Local Reference** from the pop-up menu.

    The **Well** database object's icon will change from  to  .



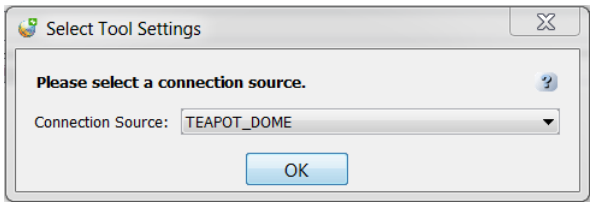
23. Select **File > Exit** to close the **Perform Table Modeling** window.

Exercise: Editing a Reference Table

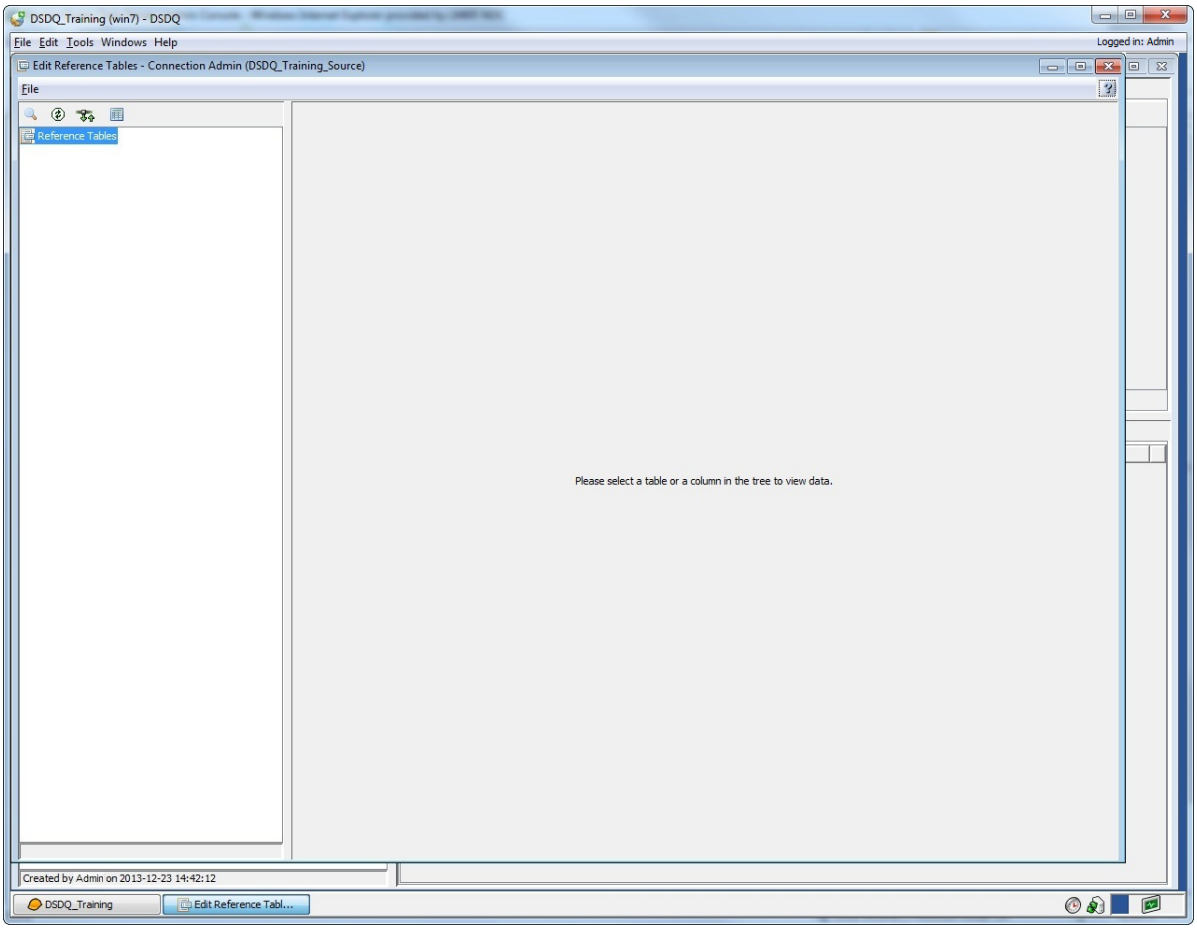
The **Edit Reference Tables** Tool allows the modification of data in application reference tables.

To edit a reference table:

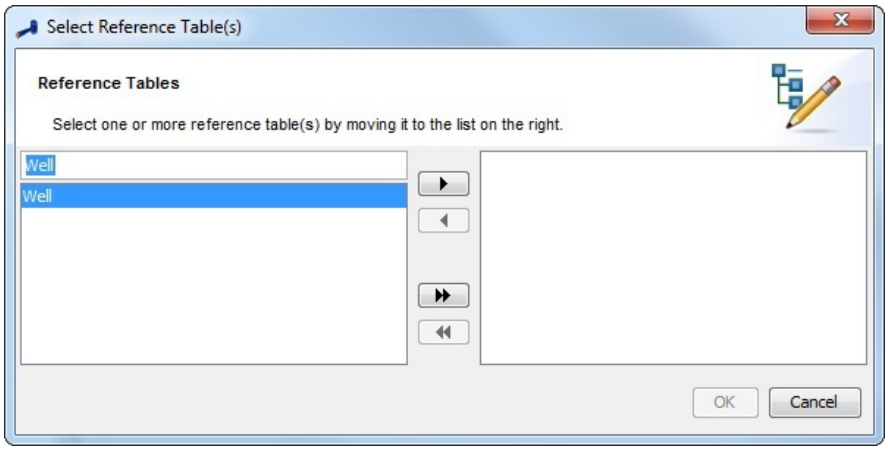
1. Double-click the **Edit Reference Tables** tool or right-click the **Edit Reference Tables** tool, and select **Open Tool** from the pop-up menu. The Select Tool Settings dialog box displays. Select the **Connection Source** and click **OK**.



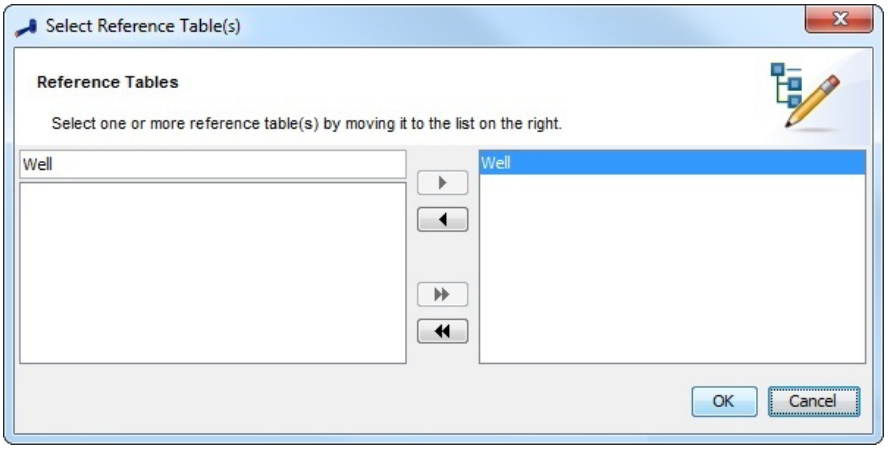
   The **Edit Reference Tables** window appears.



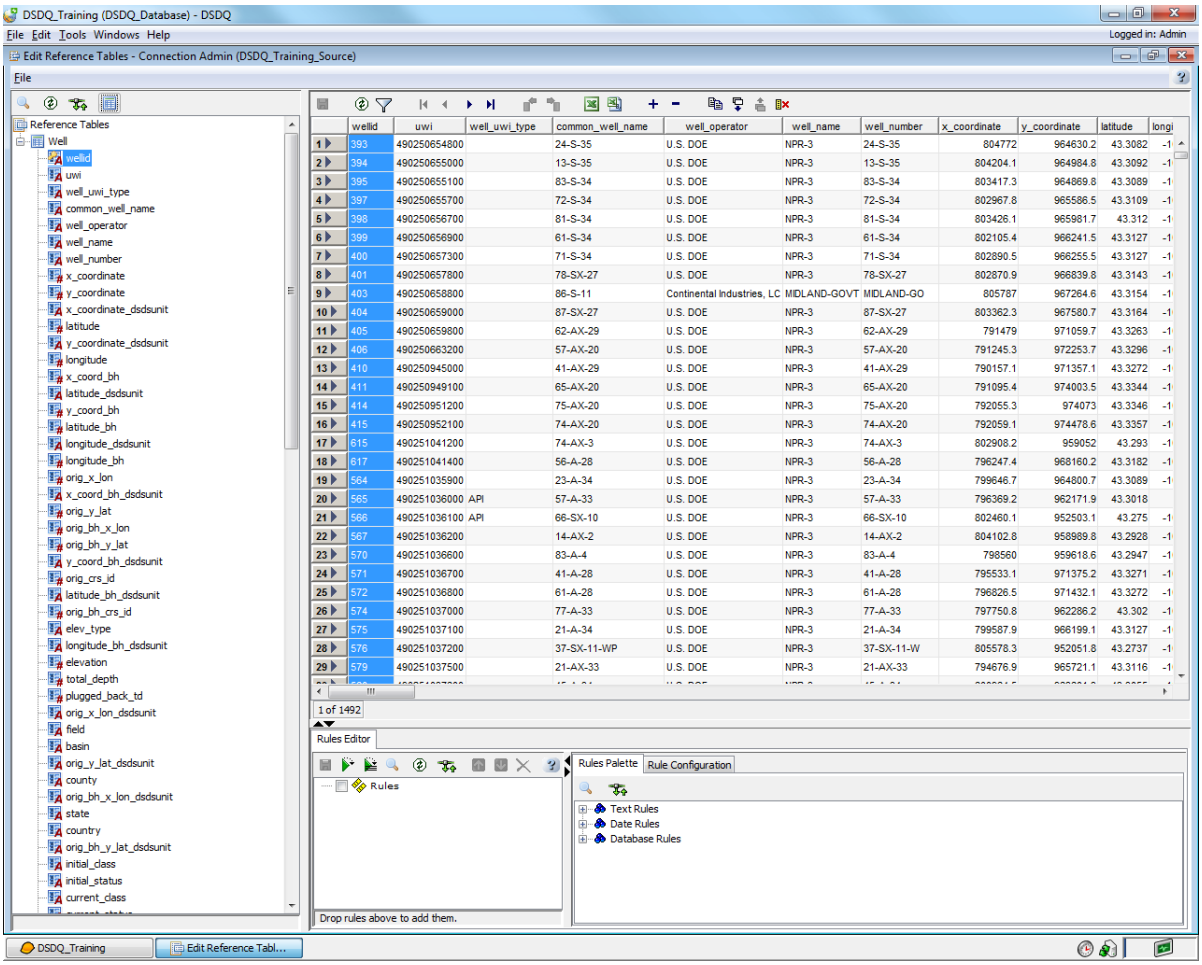
2. Click to select reference tables.
   The **Select Reference Tables** dialog box appears.



3. Select the **Well** table and click to move it to the selected tables list.



4. Click **OK**.
   The **Edit Reference Tables** screen appears.



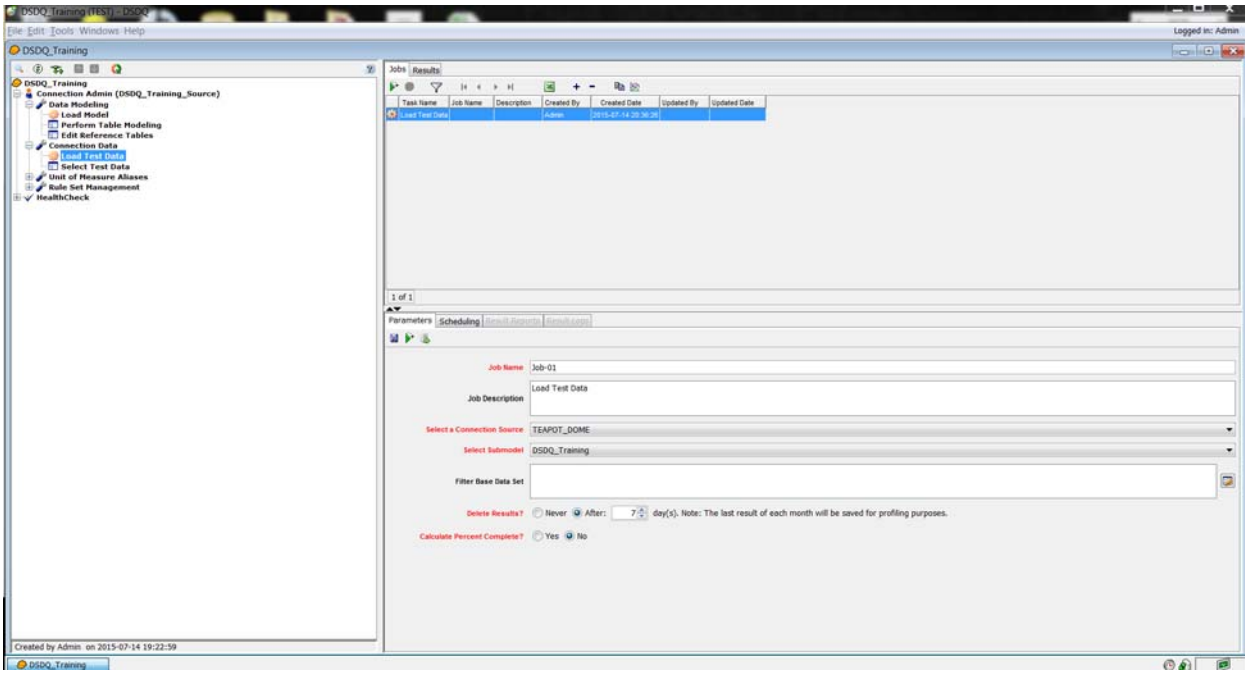
5. Select **File > Exit** to close the **Edit Reference Tables** window.

Exercise: Loading Test Data

The **Load Test Data** Task loads the source submodel data into the test data tables.

To load test data:

1. Double-click the **Load Test Data** Task or right-click the **Load Test Data** Task and select **Add Job** from the pop-up menu.



2. Enter **Job-01** in the **Job Name** field.

3. Enter **Load Test Data** in the **Job Description** field.

4. Select **TEAPOT_DOME** from **Select a Connection Source**.

5. Select **DSDQ_Training** from the **Select Submodel** drop-down list.

6. Optionally, set a filter on the data subset.

7. Select the **After** option for **Delete Results?** Change the number of days to **7**.

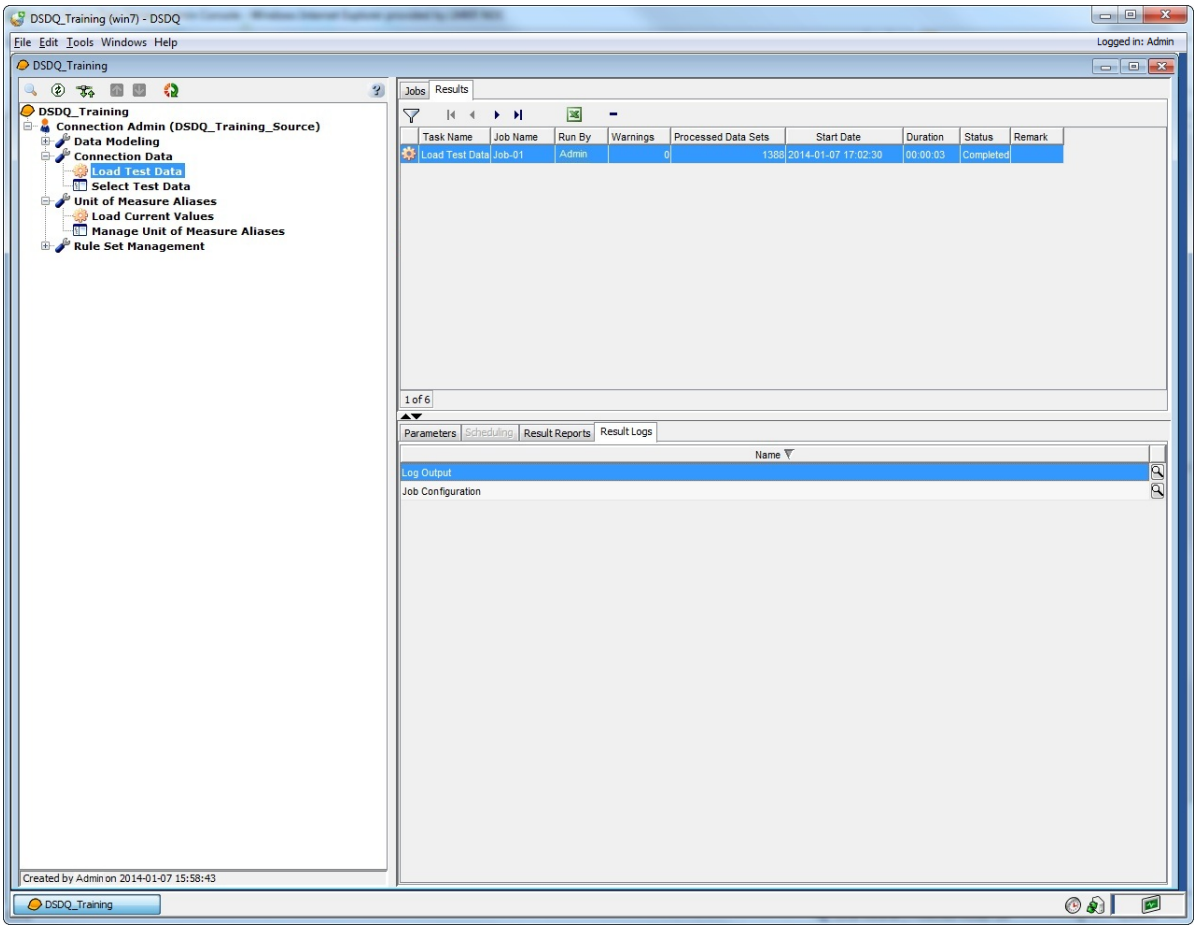
8. Select the **No** option for **Calculate Percent Complete?**

9. Select the **No** option for **Enable Data Read Ahead?**

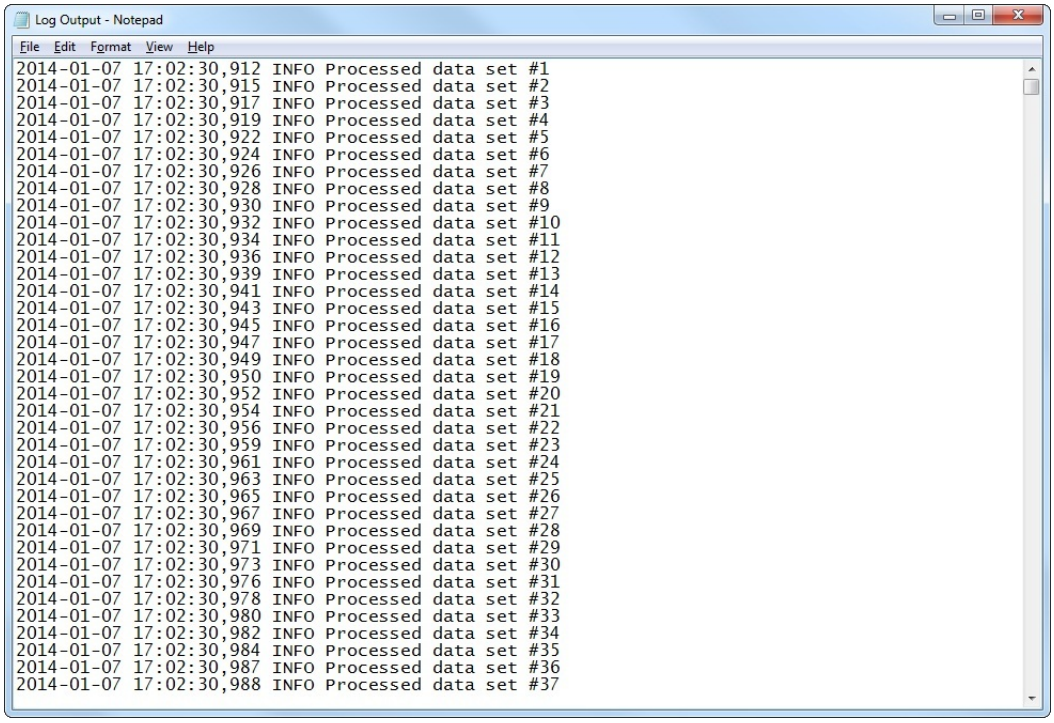
10. Click to save changes in the **Parameters** tab.

11. Click .

12. Select the **Results** tab.
    The **Load Test Data** task runs and displays results in the **Result Logs** tab on the **Jobs and Results Information Pane**.



13. Double-click the **Log Output** file on the **Jobs and Results Information Pane** to display the **Log Output** in TXT format.



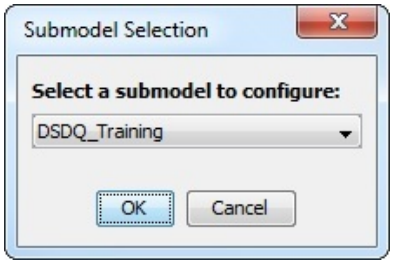
14. Select **File > Exit** to close the log file.

Exercise: Selecting Test Data

The **Select Test Data** Tool helps in selecting data subset for performing the **HealthCheck** and **Clean** Phases. The **Load Test Data** Task must be run on the submodel prior to running this Tool.

To select test data:

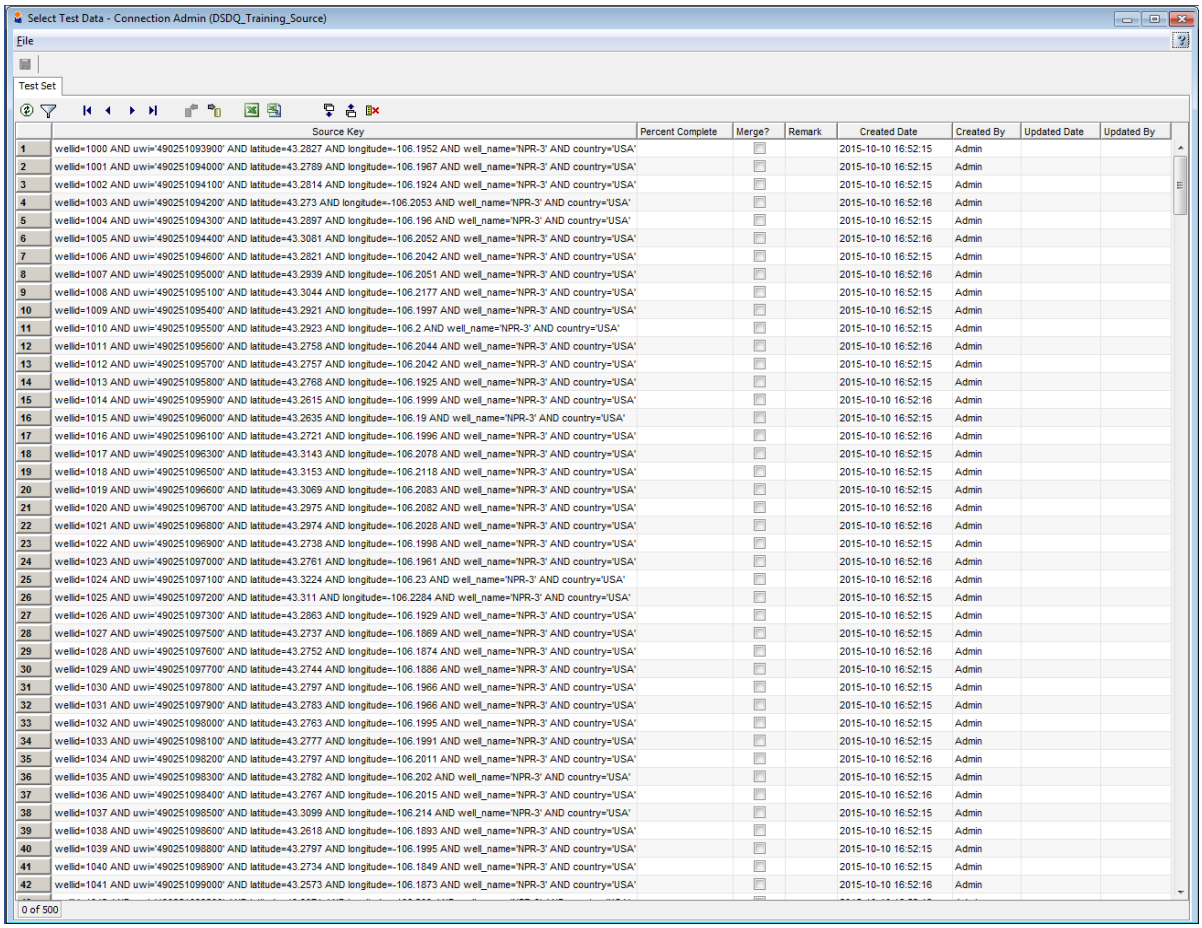
1. Double-click the **Select Test Data** Tool or right-click the **Select Test Data** Tool and select **Open Tool** from the pop-up menu.
   The **Submodel Selection** dialog box appears.



> **Note**
>
> Only Submodels that have been loaded will be displayed in the drop-down list. Refer to **Loading Test Data** for more details.

2. Select **DSDQ_Training** from the **Select a submodel to configure** drop-down list.

3. Click **OK**.



4. Select the check box for the records that will be used for testing purposes in the **Merge?** column. You can select check boxes for multiple phases that will be used during the testing process.

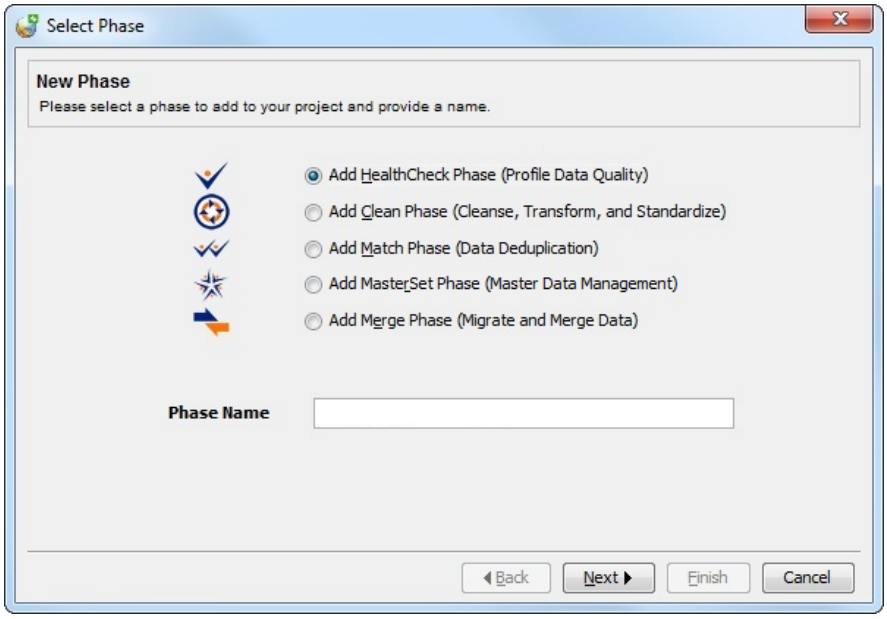
5. Click to save changes.

6. Click **File > Exit** to close the window.

Data Evaluation in DecisionSpace Data Quality

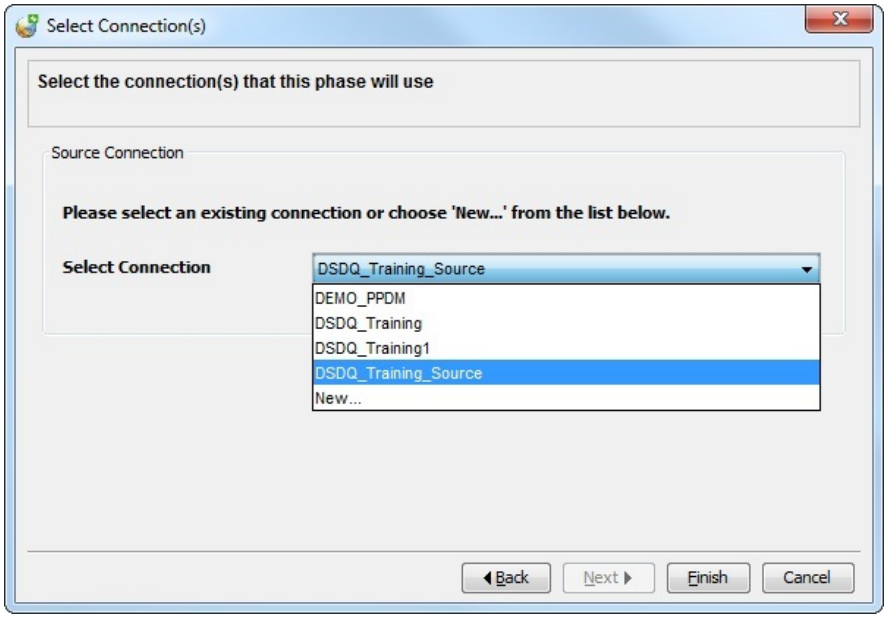
To add a HealthCheck Phase:

1. Click the **Add New Phase** button on the Project toolbar. The Select Phase window appears with the **Add HealthCheck Phase (Profile Data Quality)** option selected by default.



2. Enter **Training_HealthCheck** in the **Phase Name** field.

3. Click **Next** to continue.
   The **Select Connection(s)** window appears.



4. Select **DSDQ_Training_Source** from the **Select Connection** drop-down list.

5. Click **Finish**.
   The **HealthCheck** Phase is created and displayed in the **DecisionSpace Data Quality Project** Window.

Evaluating Data Volume and Quality

The **Rapid HealthCheck** Activity provides a quick look at the volume and quality of the data. It is fast to run and does not require a great deal of configuration. The **Run Table Analysis on All Tables** task does a simple row count, which is useful for identifying tables for modeling. The **Run Column Analysis on All Columns** task offers basic data profiling by checking the following parameters within a column:

- "Rows": Number of rows
- "# Not Null": Number of not null values
- "% Populated": Percentage of rows populated
- "# Unique": Number of unique values
- "Minimum Value": Minimum value in the column
- "Maximum Value": Maximum value in the column
- "# Mixed Case": Number of values with mixed cases
- "# NPC": Number of Non-Printable Characters
- "# PWS": Number of Preceding White Spaces
- "# TWS": Number of Trailing White Spaces
- "# DWS": Number of Double White Spaces (between words)

Exercise: Profiling Data Using SQL Query

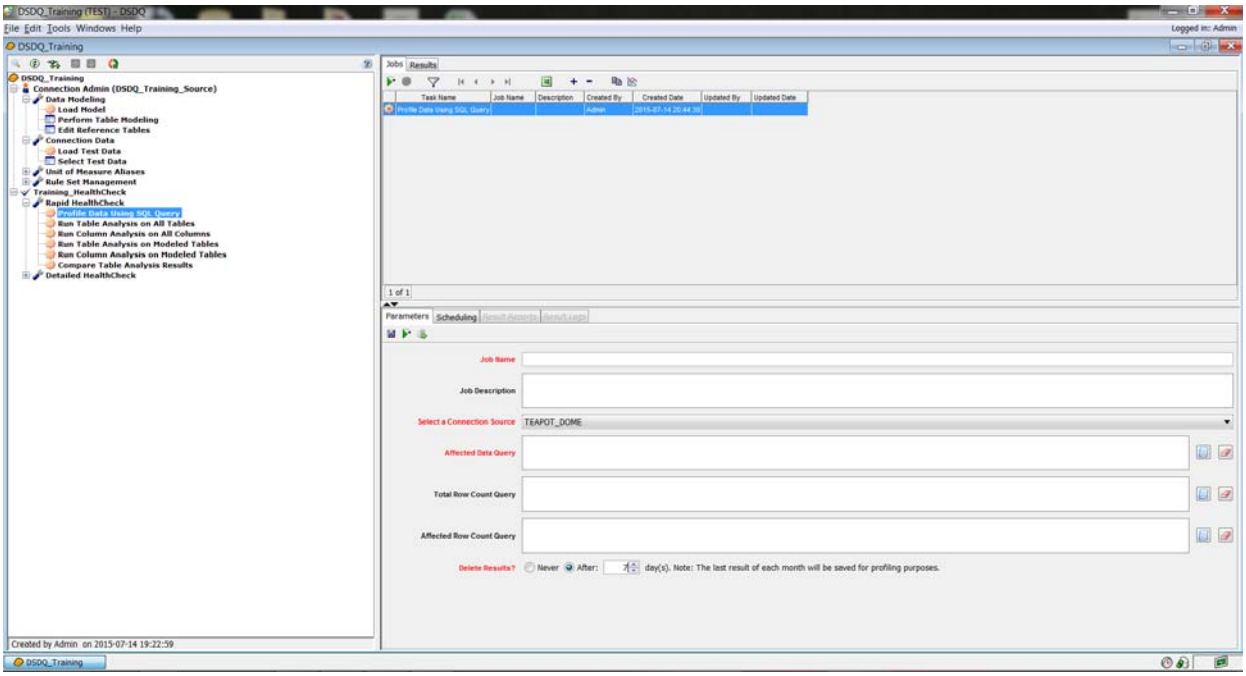
SQL queries can be customized (as per your requirement) and applied to the data source.

To profile data using SQL query:

1. Click ⊞ on the DecisionSpace Data Quality Tree to expand the **Rapid HealthCheck** Activity.

2. Double-click the **Profile Data Using SQL Query** task or right-click the **Profile Data Using SQL Query** task, and select **Add Job** from the pop-up menu.

   A new job is initiated and displays on the **Job and Results Information Pane** on the right side of the **DecisionSpace Data Quality Project Window**.



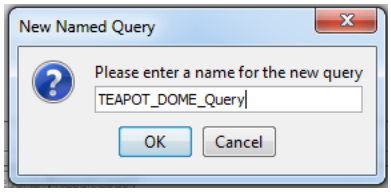
3. Enter **Job-01** in the **Job Name** field.

4. Enter **SQL Profiling** in the **Job Description** field.

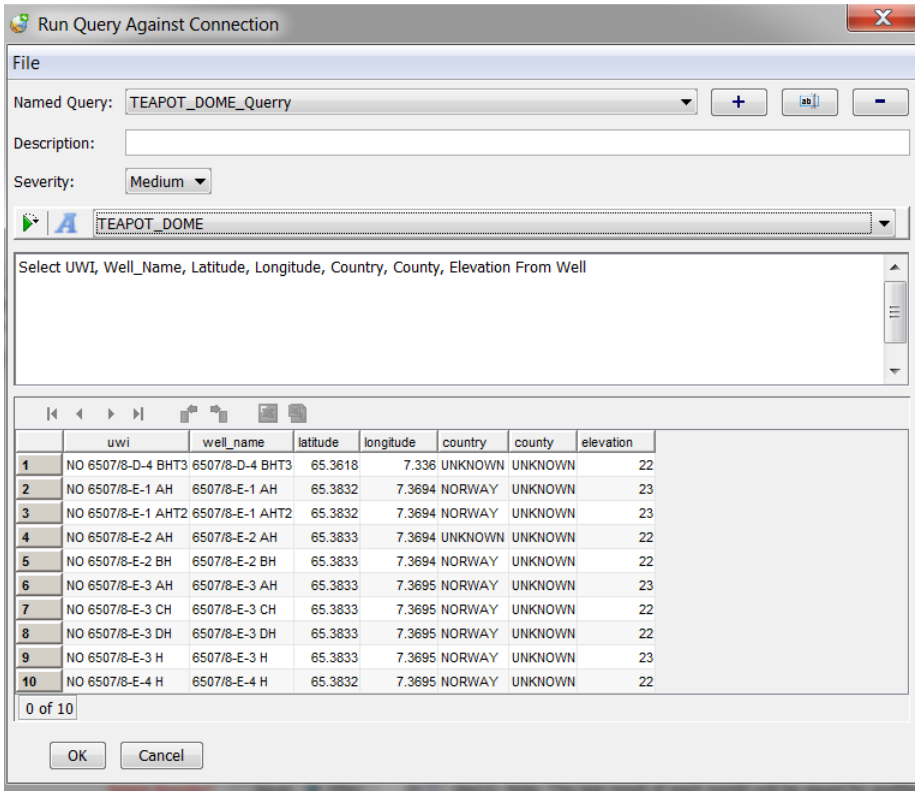
5. Select **TEAPOT_DOME** from **Select a Connection Source**.

6. Click in the **Affected Data Query** field.

7. Enter the name **TEAPOT_DOME_Query** in the New Named Query message box.



8. The Run Query Against Connection dialog box displays.



9. Enter **Select UWI, Well_Name, Latitude, Longitude, Country, County, Elevation From Well** in the SQL Query field.

10. Click and then click **OK** on the Run Query Against Connection dialog box.

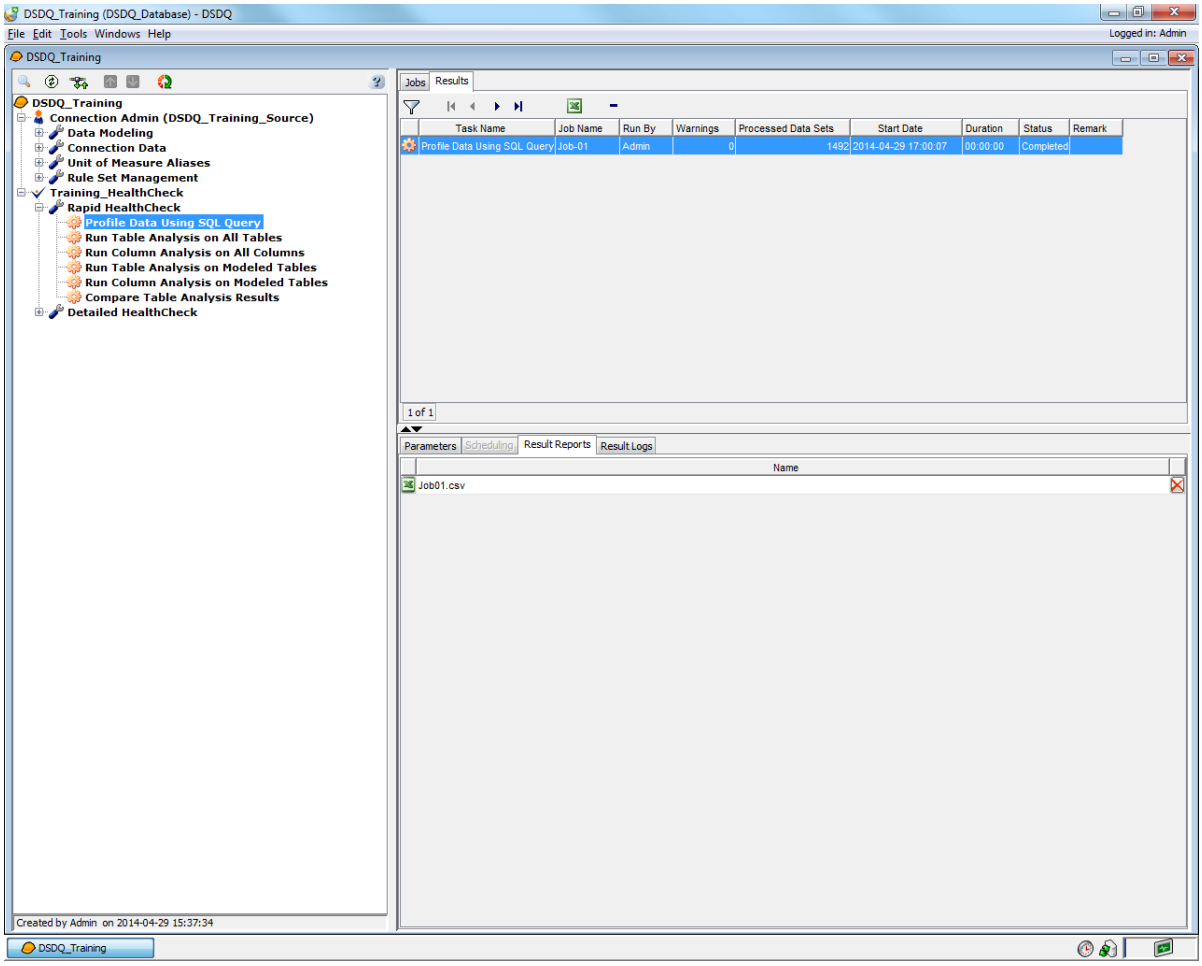
11. Select the **After** option for **Delete Results?** Leave the number of days as **7**.

12. You can edit the query in the text area provided.

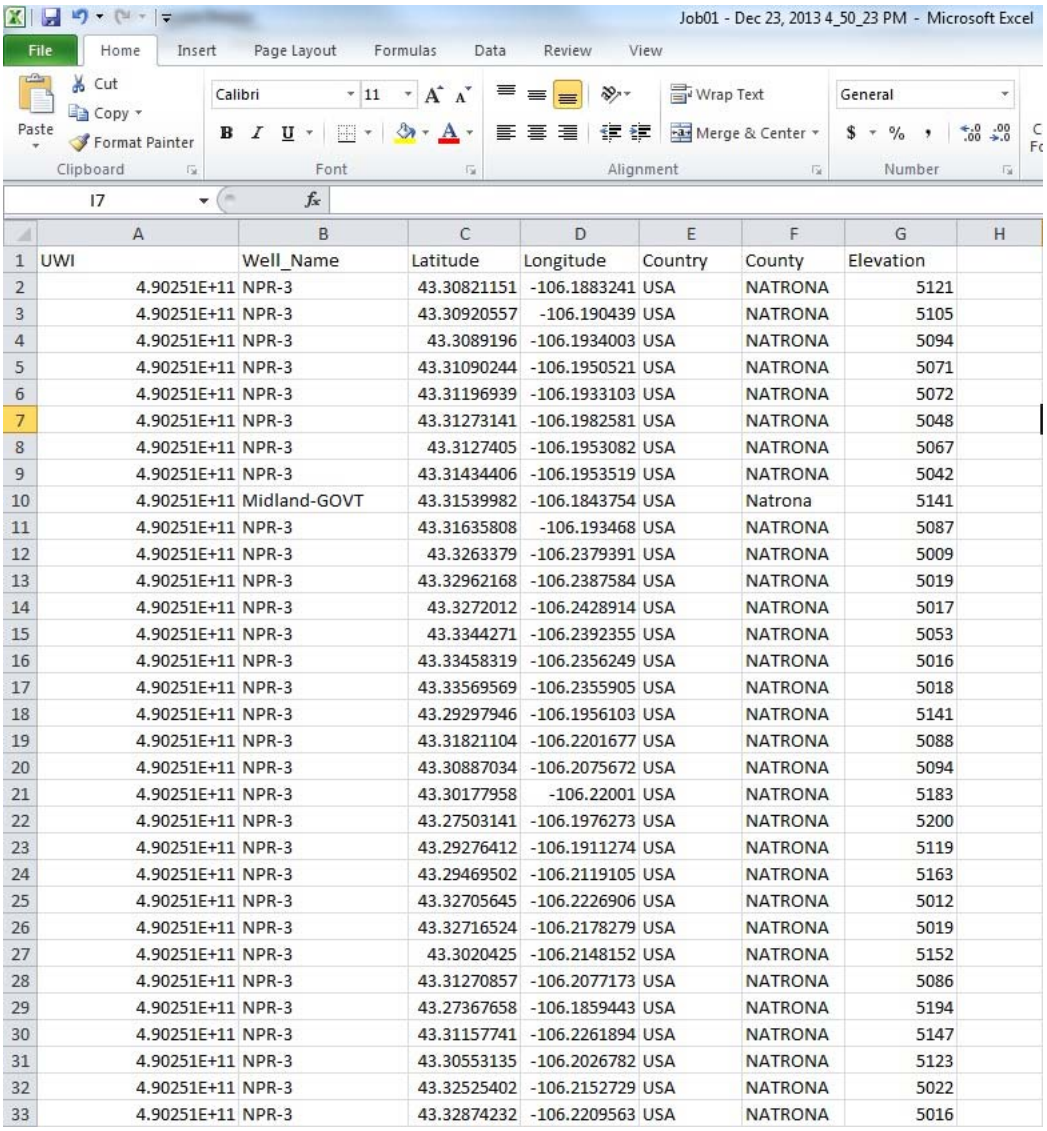
13. Click to save changes in the **Parameters** tab.

14. Click .
    The **Profile Data Using SQL Query** task runs and displays results in the **Result Reports** tab of the **Job and Results Information Pane**.

15. Select the **Results** tab on the **Job and Results Listing Pane** to view the values in the **Result Reports** tab on the **Job and Results Information Pane**.



16. Click on the **Result Reports** tab on the **Job and Results Information Pane** to display the **Profile Data using SQL Query** results in XLSX format.



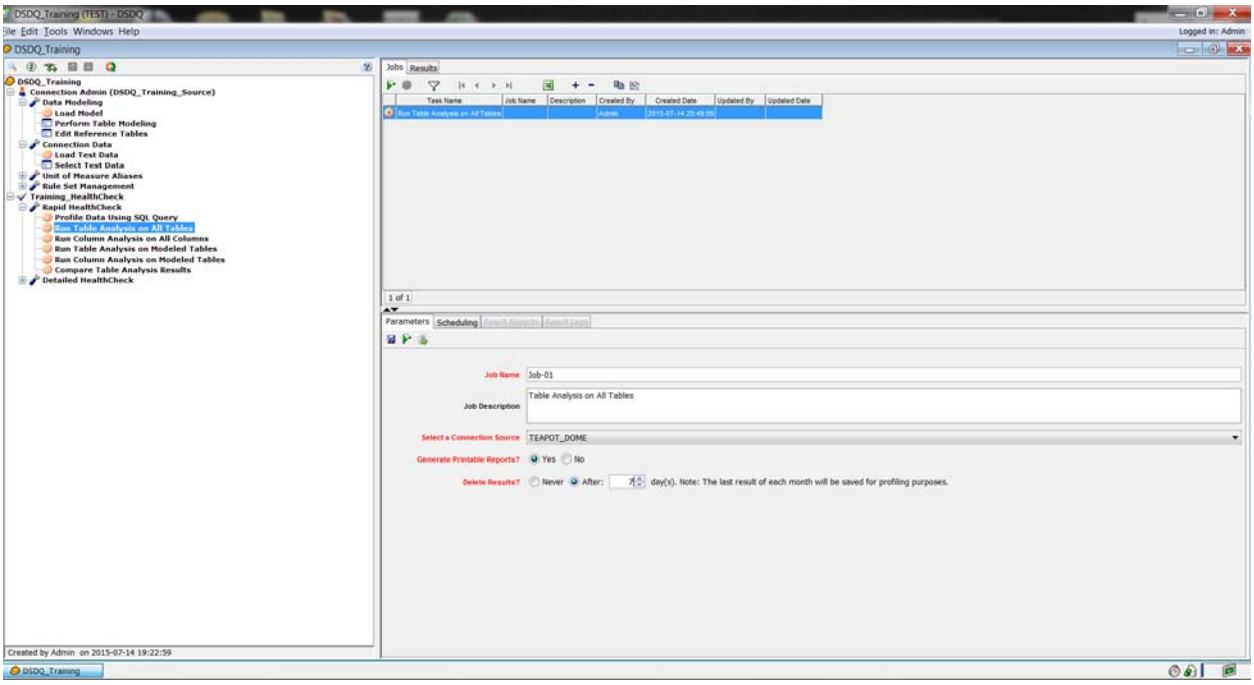
	A	B	C	D	E	F	G	H
1	UWI	Well_Name	Latitude	Longitude	Country	County	Elevation	
2	4.90251E+11	NPR-3	43.30821151	-106.1883241	USA	NATRONA	5121	
3	4.90251E+11	NPR-3	43.30920557	-106.190439	USA	NATRONA	5105	
4	4.90251E+11	NPR-3	43.3089196	-106.1934003	USA	NATRONA	5094	
5	4.90251E+11	NPR-3	43.31090244	-106.1950521	USA	NATRONA	5071	
6	4.90251E+11	NPR-3	43.31196939	-106.1933103	USA	NATRONA	5072	
7	4.90251E+11	NPR-3	43.31273141	-106.1982581	USA	NATRONA	5048	
8	4.90251E+11	NPR-3	43.3127405	-106.1953082	USA	NATRONA	5067	
9	4.90251E+11	NPR-3	43.31434406	-106.1953519	USA	NATRONA	5042	
10	4.90251E+11	Midland-GOVT	43.31539982	-106.1843754	USA	Natrona	5141	
11	4.90251E+11	NPR-3	43.31635808	-106.193468	USA	NATRONA	5087	
12	4.90251E+11	NPR-3	43.3263379	-106.2379391	USA	NATRONA	5009	
13	4.90251E+11	NPR-3	43.32962168	-106.2387584	USA	NATRONA	5019	
14	4.90251E+11	NPR-3	43.3272012	-106.2428914	USA	NATRONA	5017	
15	4.90251E+11	NPR-3	43.3344271	-106.2392355	USA	NATRONA	5053	
16	4.90251E+11	NPR-3	43.33458319	-106.2356249	USA	NATRONA	5016	
17	4.90251E+11	NPR-3	43.33569569	-106.2355905	USA	NATRONA	5018	
18	4.90251E+11	NPR-3	43.29297946	-106.1956103	USA	NATRONA	5141	
19	4.90251E+11	NPR-3	43.31821104	-106.2201677	USA	NATRONA	5088	
20	4.90251E+11	NPR-3	43.30887034	-106.2075672	USA	NATRONA	5094	
21	4.90251E+11	NPR-3	43.30177958	-106.22001	USA	NATRONA	5183	
22	4.90251E+11	NPR-3	43.27503141	-106.1976273	USA	NATRONA	5200	
23	4.90251E+11	NPR-3	43.29276412	-106.1911274	USA	NATRONA	5119	
24	4.90251E+11	NPR-3	43.29469502	-106.2119105	USA	NATRONA	5163	
25	4.90251E+11	NPR-3	43.32705645	-106.2226906	USA	NATRONA	5012	
26	4.90251E+11	NPR-3	43.32716524	-106.2178279	USA	NATRONA	5019	
27	4.90251E+11	NPR-3	43.3020425	-106.2148152	USA	NATRONA	5152	
28	4.90251E+11	NPR-3	43.31270857	-106.2077173	USA	NATRONA	5086	
29	4.90251E+11	NPR-3	43.27367658	-106.1859443	USA	NATRONA	5194	
30	4.90251E+11	NPR-3	43.31157741	-106.2261894	USA	NATRONA	5147	
31	4.90251E+11	NPR-3	43.30553135	-106.2026782	USA	NATRONA	5123	
32	4.90251E+11	NPR-3	43.32525402	-106.2152729	USA	NATRONA	5022	
33	4.90251E+11	NPR-3	43.32874232	-106.2209563	USA	NATRONA	5016	

Exercise: Running Table Analysis on All Tables

This task is used to analyze all the tables for issues and inconsistencies. In this particular exercise, we are analyzing the tables and count the number of rows in them. Rows are counted when values are entered in them.

To run table analysis on all tables:

1. Double-click the **Run Table Analysis on All Tables** Task or right-click the **Run Table Analysis on All Tables** Task and select **Add Job** from the pop-up menu.
   A new job is initiated and displays on the **Job and Results Information Pane** on the right side of the **DecisionSpace Data Quality Project Window**.



2. Enter **Job-01** in the **Job Name** field.

3. Enter **Table Analysis on All Tables** in the **Job Description** field.

4. Select a project from the **Select a Connection Source** list.

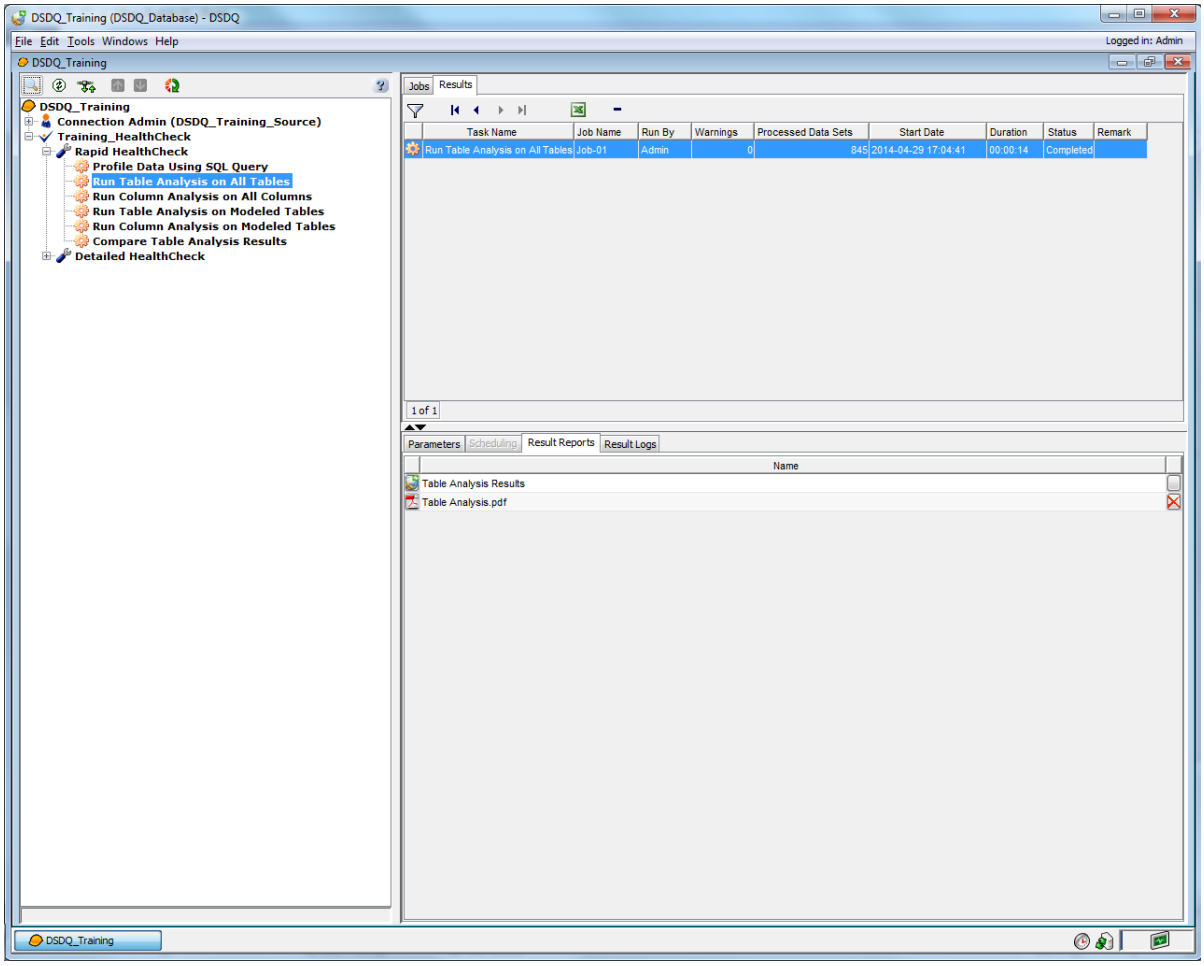
5. Select the **Yes** option for **Generate Printable Reports?**

6. Select the **After** option for **Delete Results?** Leave the number of days as **7**.

7. Click to save changes in the **Parameters** tab.

8. Click .
   The **Run Table Analysis on All Tables** task runs and displays results in the **Result Reports** tab on the **Jobs and Results Information Pane**.

9. Select the **Results** tab on the **Jobs and Results Listing Pane** to view the values in the **Result Reports** tab on the **Job and Results Information Pane**.



10. Click on the **Result Reports** tab to display results for **Table Analysis on All Tables** in PDF format.

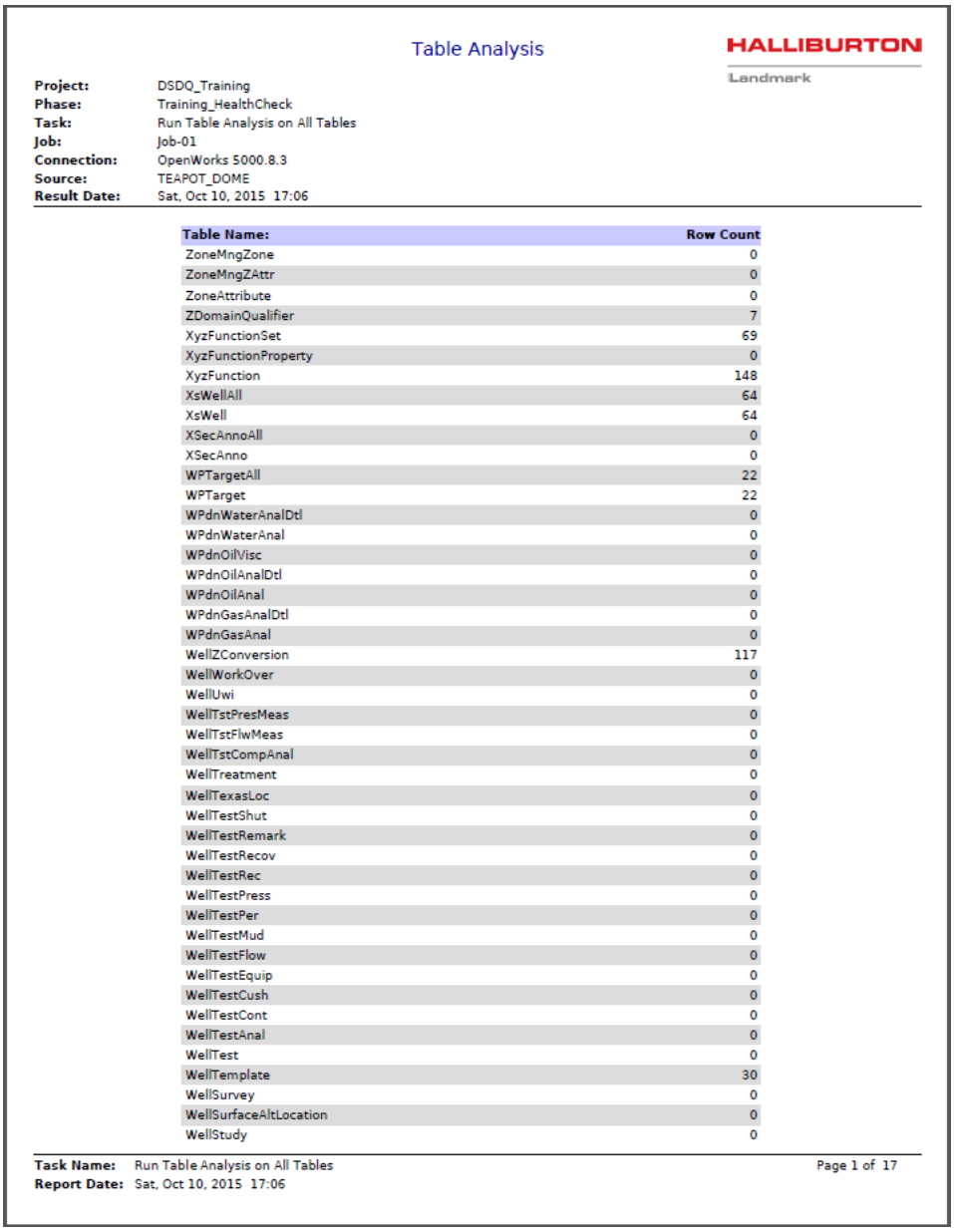


Table Analysis

HALLIBURTON
Landmark

Project: DSDQ_Training
Phase: Training_HealthCheck
Task: Run Table Analysis on All Tables
Job: Job-01
Connection: OpenWorks 5000.8.3
Source: TEAPOT_DOME
Result Date: Sat, Oct 10, 2015 17:06

Table Name:	Row Count
ZoneMngZone	0
ZoneMngZAttr	0
ZoneAttribute	0
ZDomainQualifier	7
XyzFunctionSet	69
XyzFunctionProperty	0
XyzFunction	148
XsWellAll	64
XsWell	64
XSecAnnoAll	0
XSecAnno	0
WPTargetAll	22
WPTarget	22
WPdnWaterAnalDtl	0
WPdnWaterAnal	0
WPdnOilVisc	0
WPdnOilAnalDtl	0
WPdnOilAnal	0
WPdnGasAnalDtl	0
WPdnGasAnal	0
WellZConversion	117
WellWorkOver	0
WellUwi	0
WellTstPresMeas	0
WellTstFlwMeas	0
WellTstCompAnal	0
WellTreatment	0
WellTexasLoc	0
WellTestShut	0
WellTestRemark	0
WellTestRecov	0
WellTestRec	0
WellTestPress	0
WellTestPer	0
WellTestMud	0
WellTestFlow	0
WellTestEquip	0
WellTestCush	0
WellTestCont	0
WellTestAnal	0
WellTest	0
WellTemplate	30
WellSurvey	0
WellSurfaceAltLocation	0
WellStudy	0

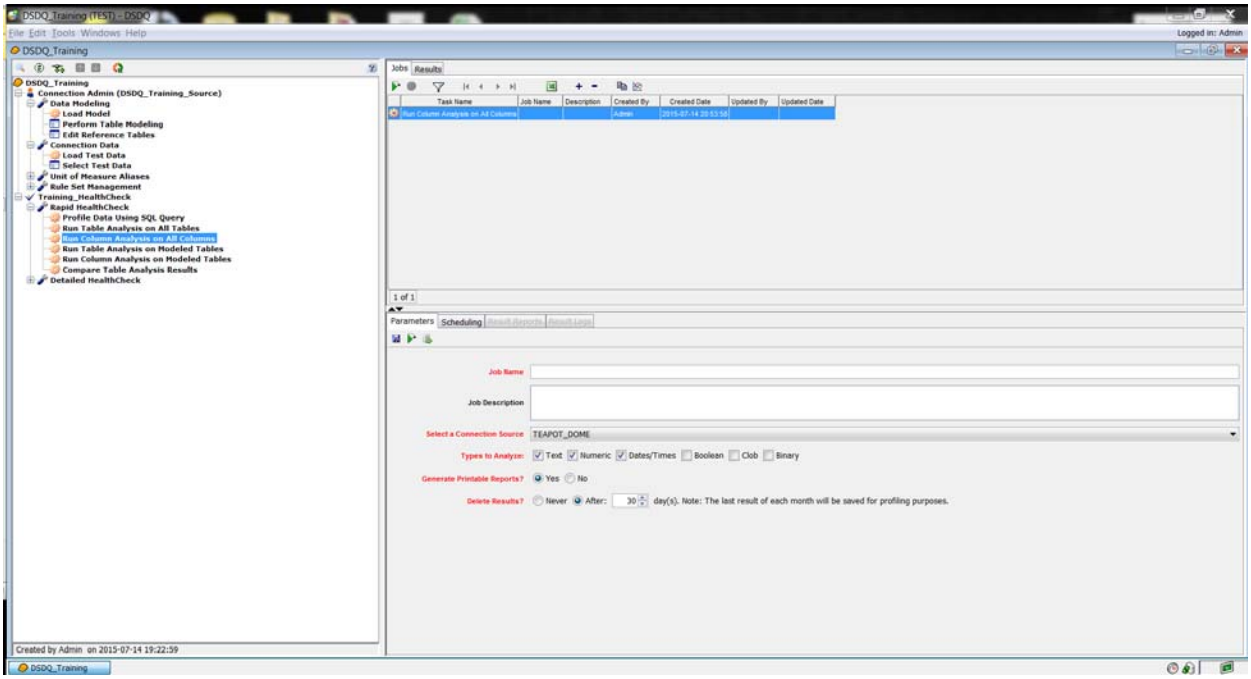
Task Name: Run Table Analysis on All Tables
Report Date: Sat, Oct 10, 2015 17:06

Page 1 of 17

Exercise: Running Column Analysis on All Columns

To run column analysis on all the columns:

1. Double-click the **Run Column Analysis on All Columns** task or right-click the **Run Column Analysis on All Columns** task, and select **Add Job** from the pop-up menu.
   A new job is initiated and displays on the **Job and Results Information Pane** on the right side of the **DecisionSpace Data Quality Project Window**.



2. Enter **Job-01** in the **Job Name** field.

3. Enter **Column Analysis on All Columns** in the **Job Description** field.

4. Select **TEAPOT_DOME** from the **Select a Connection Source** drop-down list.

5. Select all options for **Types to Analyze**.

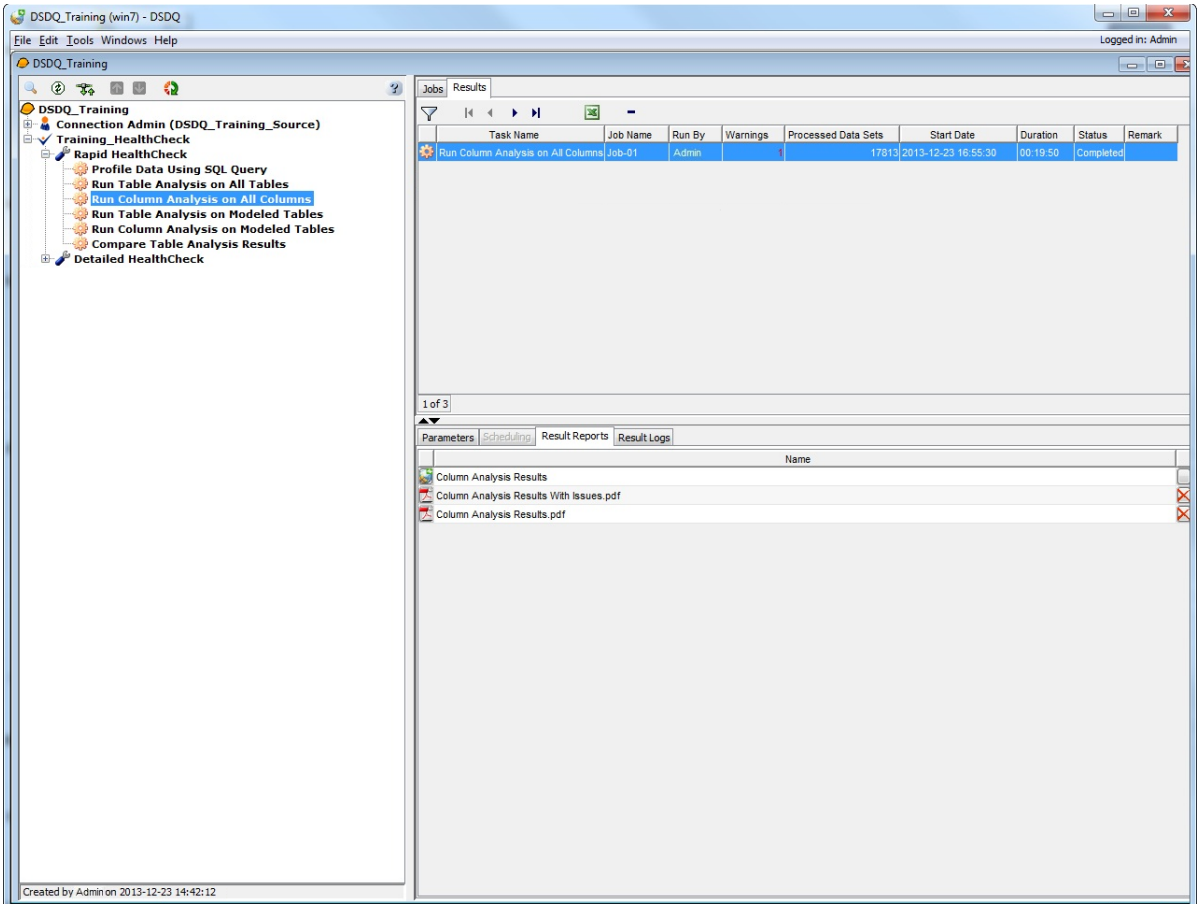
6. Select the **Yes** option for **Generate Printable Reports?**

7. Select the **After** option for **Delete Results?** Leave the number of days as **7**.

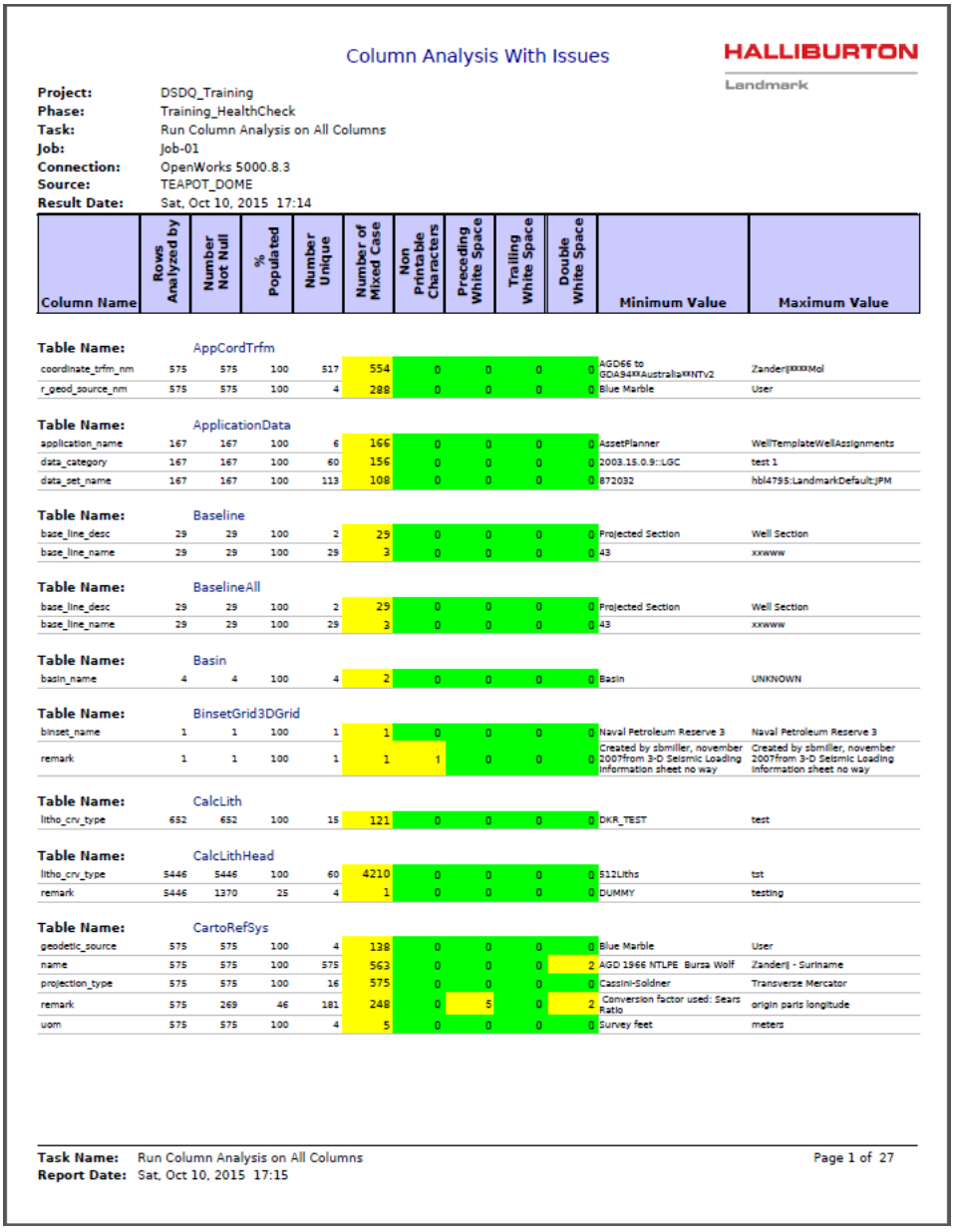
8. Click to save changes in the **Parameters** tab.

9. Click .
   The **Run Column Analysis on All Columns** task runs and displays results in the **Result Reports** tab on the **Jobs and Results Information Pane**.

10. Select the **Results** tab on the **Jobs and Results Listing Pane** to view the values in the **Result Reports** tab on the **Job and Results Information Pane**.



11. Click on the **Result Reports** tab to display **Column Analysis Results with Issues** in PDF format.



Column Analysis With Issues

HALLIBURTON
Landmark

Project: DSDQ_Training
Phase: Training_HealthCheck
Task: Run Column Analysis on All Columns
Job: Job-01
Connection: OpenWorks 5000.8.3
Source: TEAPOT_DOME
Result Date: Sat, Oct 10, 2015 17:14

Column Name	Rows Analyzed by	Number Not Null	% Populated	Number Unique	Number of Mixed Case	Non Printable Characters	Preceding White Space	Trailing White Space	Double White Space	Minimum Value	Maximum Value
Table Name: AppCordTrfm											
coordinate_trfm_nm	575	575	100	517	554	0	0	0	0	AGD66 to GDA94[XX]Australia[XX]NTv2	Zander[XXX]Mol
r_geod_source_nm	575	575	100	4	288	0	0	0	0	Blue Marble	User
Table Name: ApplicationData											
application_name	167	167	100	6	166	0	0	0	0	AssetPlanner	WellTemplateWellAssignments
data_category	167	167	100	60	156	0	0	0	0	2003.15.0.9::LGC	test 1
data_set_name	167	167	100	113	108	0	0	0	0	872032	hbl4795:LandmarkDefault:JPM
Table Name: Baseline											
base_line_desc	29	29	100	2	29	0	0	0	0	Projected Section	Well Section
base_line_name	29	29	100	29	3	0	0	0	0	43	xxwww
Table Name: BaselineAll											
base_line_desc	29	29	100	2	29	0	0	0	0	Projected Section	Well Section
base_line_name	29	29	100	29	3	0	0	0	0	43	xxwww
Table Name: Basin											
basin_name	4	4	100	4	2	0	0	0	0	Basin	UNKNOWN
Table Name: BinsetGrid3DGrid											
binset_name	1	1	100	1	1	0	0	0	0	Naval Petroleum Reserve 3	Naval Petroleum Reserve 3
remark	1	1	100	1	1	1	0	0	0	Created by sbmiller, november 2007from 3-D Seismic Loading Information sheet no way	Created by sbmiller, november 2007from 3-D Seismic Loading Information sheet no way
Table Name: CalcLith											
litho_crv_type	652	652	100	15	121	0	0	0	0	DKR_TEST	test
Table Name: CalcLithHead											
litho_crv_type	5446	5446	100	60	4210	0	0	0	0	512Liths	tst
remark	5446	1370	25	4	1	0	0	0	0	DUMMY	testing
Table Name: CartoRefSys											
geodetic_source	575	575	100	4	138	0	0	0	0	Blue Marble	User
name	575	575	100	575	563	0	0	0	2	AGD 1966 NTLPE Bursa Wolf	Zanderij - Suriname
projection_type	575	575	100	16	575	0	0	0	0	Cassini-Soldner	Transverse Mercator
remark	575	269	46	181	248	0	5	0	2	Conversion factor used: Sears Ratio	origin parts longitude
uom	575	575	100	4	5	0	0	0	0	Survey feet	meters

Task Name: Run Column Analysis on All Columns
Report Date: Sat, Oct 10, 2015 17:15

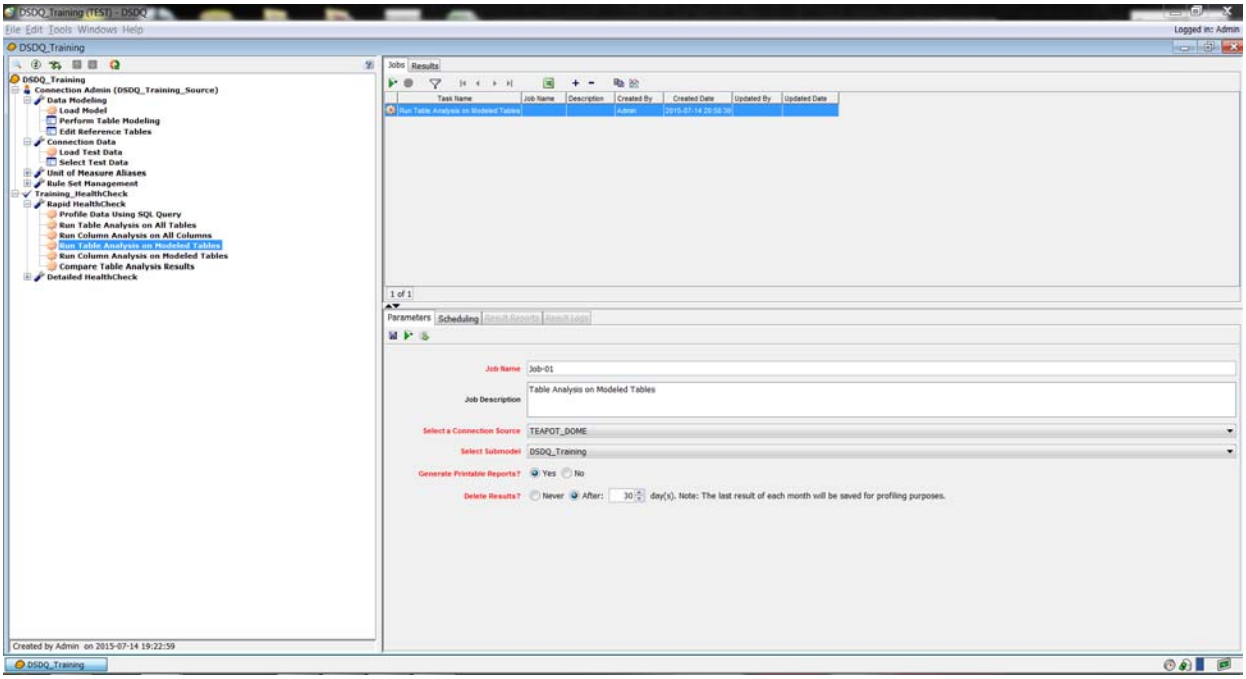
Page 1 of 27

Exercise: Running Table Analysis on Modeled Tables

The **Run Table Analysis on Modeled Tables** task runs only on tables that have been modeled in the **Perform Table Modeling** tool.

To run table analysis on all the modeled tables:

1. Double-click the **Run Table Analysis on Modeled Tables** task or right-click the **Run Table Analysis on Modeled Tables** task, and select **Add Job** from the pop-up menu.
   A new job is initiated and displays on the **Job and Results Information Pane** on the right side of the **DecisionSpace Data Quality Project Window**.



2. Enter **Job-01** in the **Job Name** field.

3. Enter **Table Analysis on Modeled Tables** in the **Job Description** field.

4. Select **TEAPOT_DOME** from the **Select a Connection Source** drop-down list.

> **Note**
>
> For more information on data owner connections, refer to Adding a New Data Owner Connection section in Chapter 2, Connecting DecisionSpace Data Quality with DecisionSpace Data Server.

5. Select **DSDQ_Training** from the **Select Submodel** drop-down list.

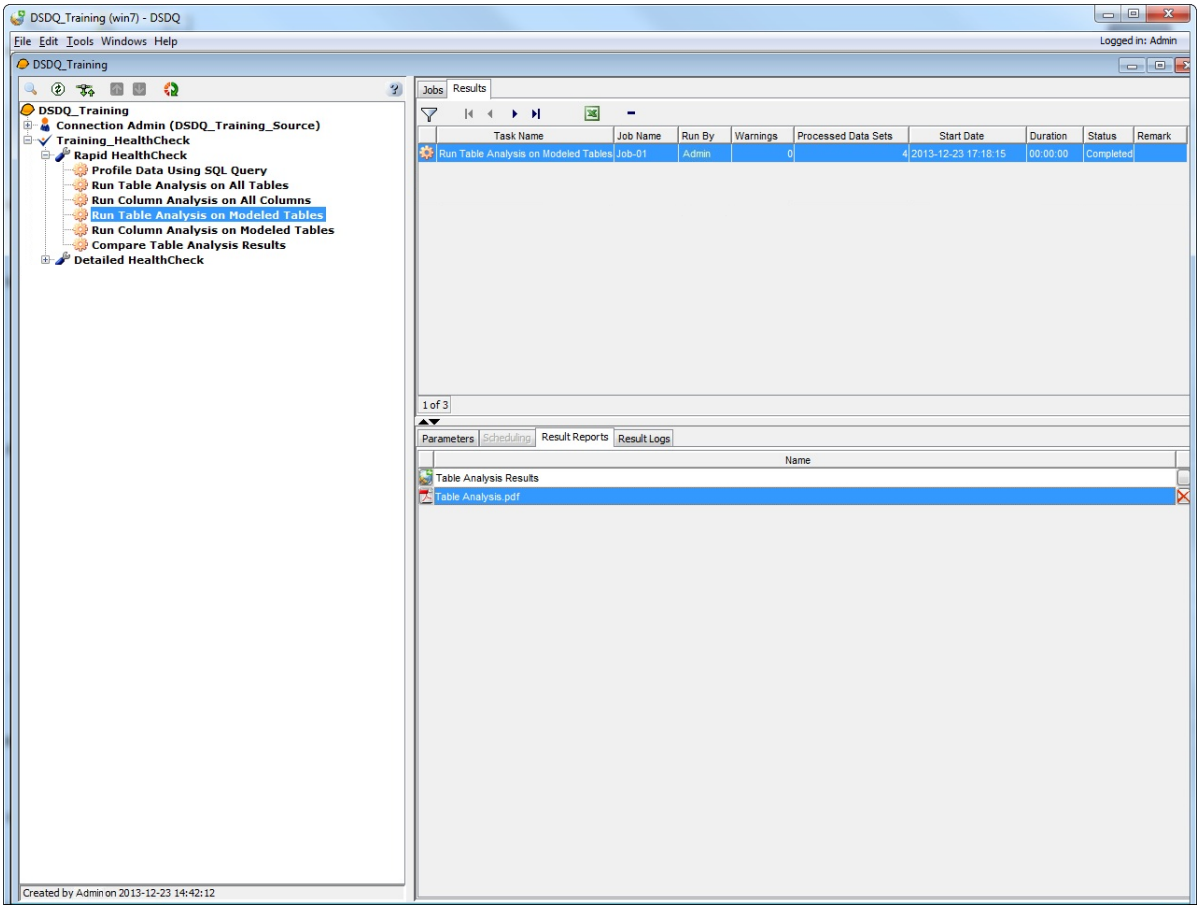
6. Select the **Yes** option for **Generate Printable Reports?**

7. Select the **After** option for **Delete Results?** Leave the number of days as **7**.

8. Click to save changes in the **Parameters** tab.

9. Click .
   The **Run Table Analysis on Modeled Tables** Task runs and displays results in the **Result Reports** tab on the **Jobs and Results Information Pane**.

10. Select the **Results** tab on the **Jobs and Results Listing Pane** to view the values in the **Result Reports** tab on the **Job and Results Information Pane**.



11. Click on the **Result Reports** tab to display **Table Analysis on Modeled Tables** results in PDF format.

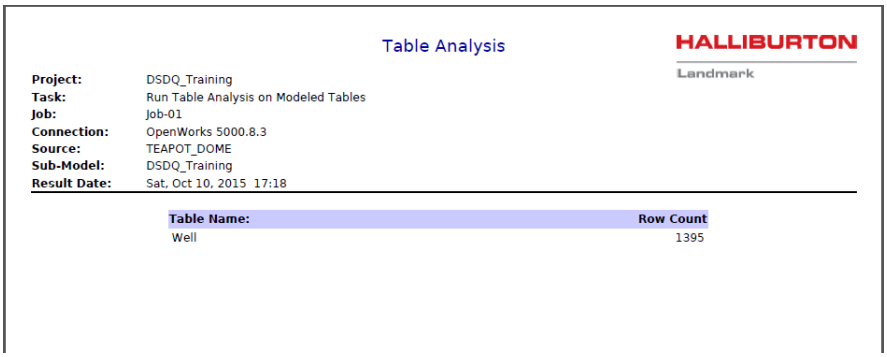


Table Analysis

HALLIBURTON
Landmark

Project:	DSDQ_Training
Task:	Run Table Analysis on Modeled Tables
Job:	Job-01
Connection:	OpenWorks 5000.8.3
Source:	TEAPOT_DOME
Sub-Model:	DSDQ_Training
Result Date:	Sat, Oct 10, 2015 17:18

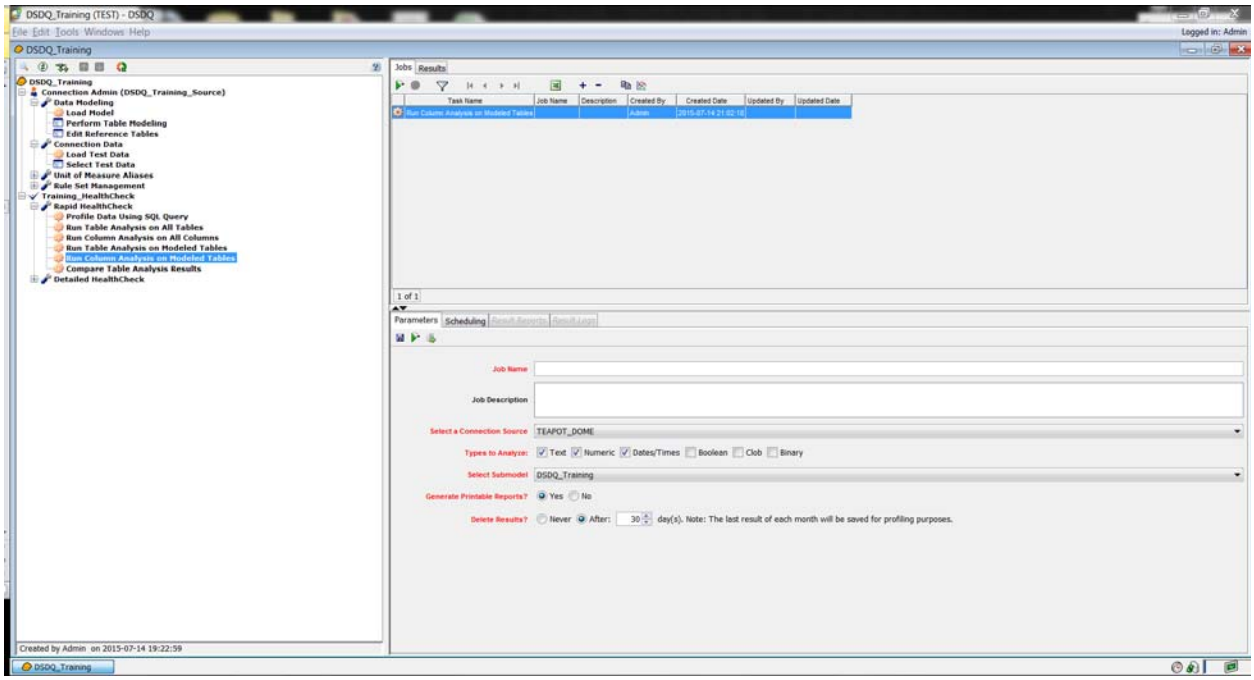
Table Name:	Row Count
Well	1395

Exercise: Running Column Analysis on Modeled Tables

The **Run Column Analysis on Modeled Tables** task runs only on tables that have been modeled in the **Perform Table Modeling** tool.

To run column analysis on all the modeled tables:

1. Double-click the **Run Column Analysis on Modeled Tables** task or right-click the **Run Column Analysis on Modeled Tables** task, and select **Add Job** from the pop-up menu.
   A new job is initiated and displays on the **Job and Results Information Pane** on the right side of the **DecisionSpace Data Quality Project Window**.



2. Enter **Job-01** in the **Job Name** field.

3. Enter **Column Analysis on Modeled Tables** in the **Job Description** field.

4. Select **TEAPOT_DOME** from the **Select a Connection Source** drop-down list.

> **Note**
>
> For more information on data owner connections, refer to Adding a New Data Owner Connection section in Chapter 2, Connecting DecisionSpace Data Quality with DecisionSpace Data Server.

5. Select all the options for **Types to Analyze**.

6. Select **DSDQ_Training** from the **Select Submodel** drop-down list.

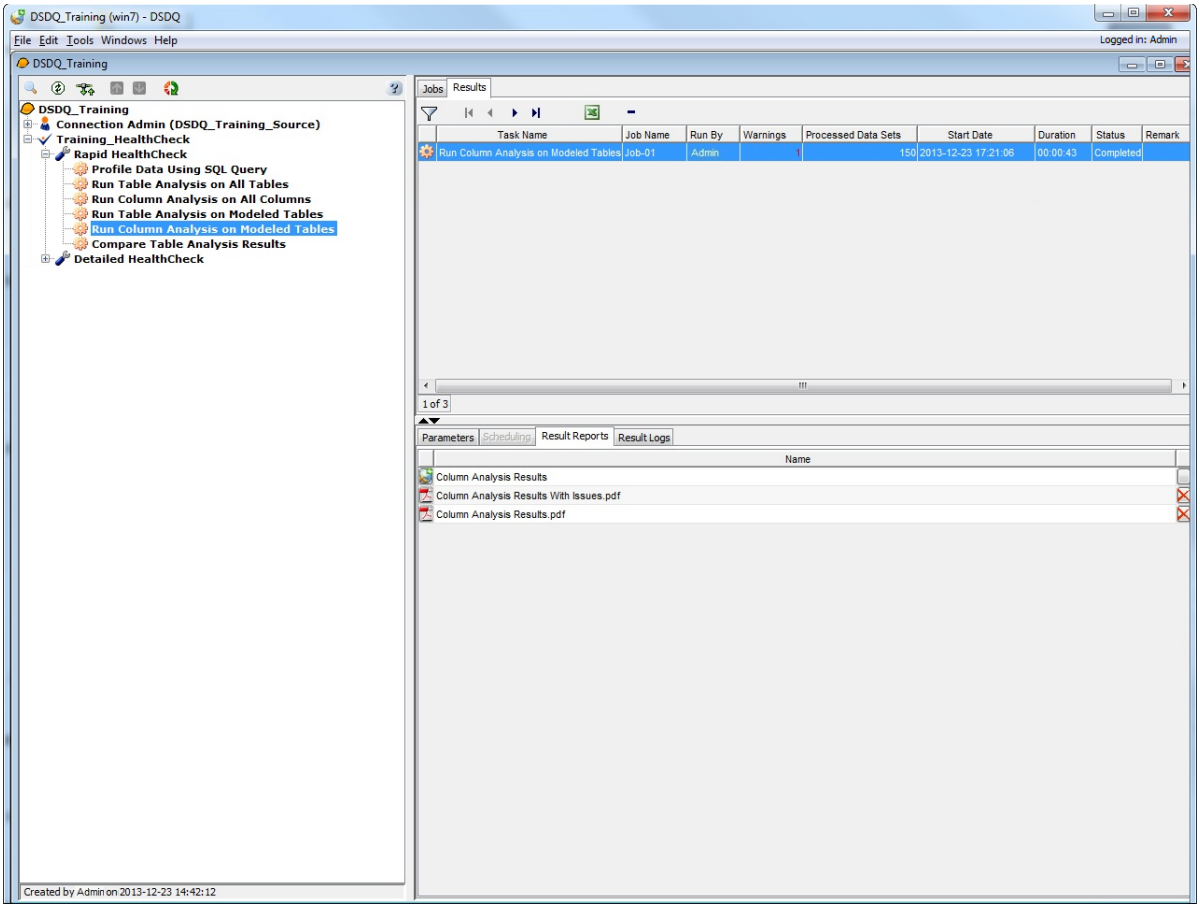
7. Select the **Yes** option for **Generate Printable Reports?**

8. Select the **After** option for **Delete Results?** Leave the number of days as **7.**

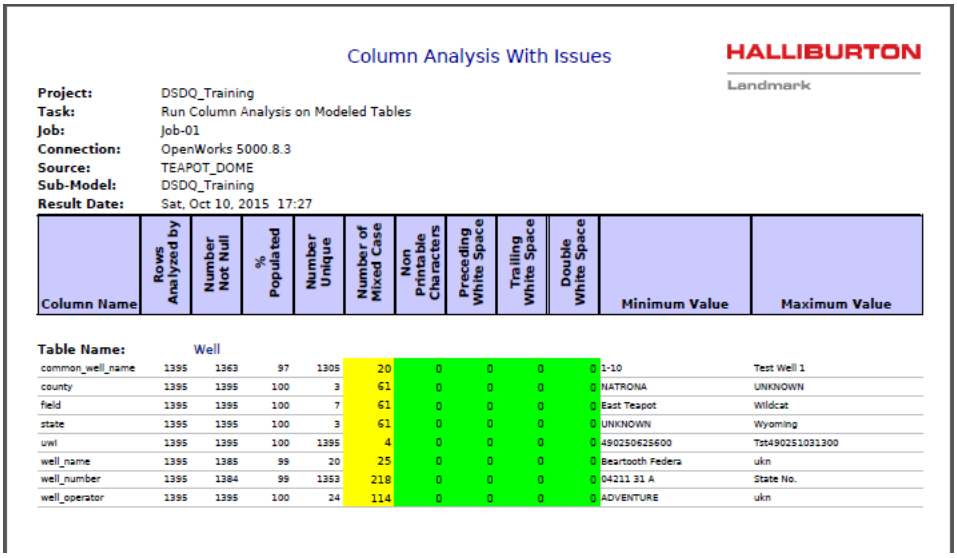
9. Click to save changes in the **Parameters** tab.

10. Click .
    The **Run Column Analysis on Modeled Tables** Task runs and displays results in the **Result Reports** tab on the **Jobs and Results Information Pane**.

11. Select the **Results** tab on the **Jobs and Results Listing Pane** to view the values in the **Result Reports** tab on the **Job and Results Information Pane**.



12. Click on the **Result Reports** tab to display **Column Analysis Results with Issues** in PDF format.



Column Analysis With Issues

HALLIBURTON
Landmark

Project: DSDQ_Training
Task: Run Column Analysis on Modeled Tables
Job: Job-01
Connection: OpenWorks 5000.8.3
Source: TEAPOT_DOME
Sub-Model: DSDQ_Training
Result Date: Sat, Oct 10, 2015 17:27

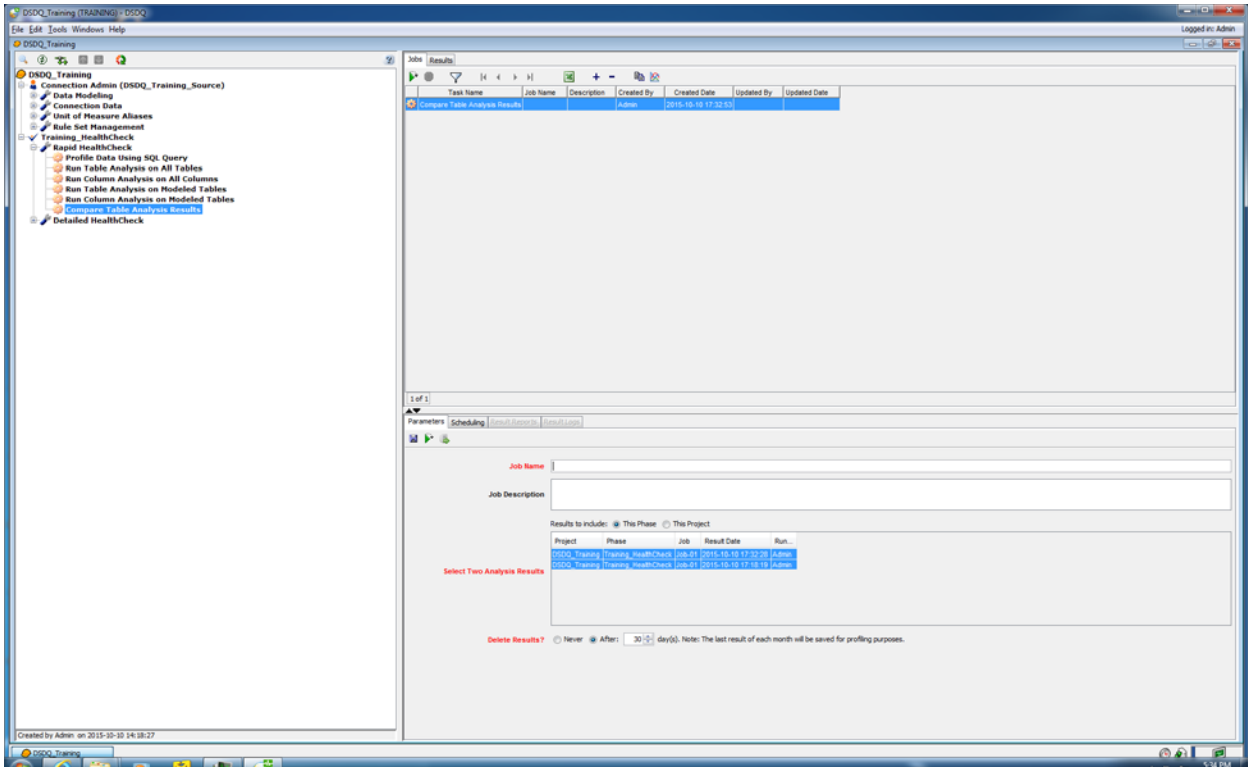
Table Name: Well

Column Name	Rows Analyzed by	Number Not Null	% Populated	Number Unique	Number of Mixed Case	Non Printable Characters	Preceding White Space	Trailing White Space	Double White Space	Minimum Value	Maximum Value
common_well_name	1395	1363	97	1305	20	0	0	0	0	1-10	Test Well 1
county	1395	1395	100	3	61	0	0	0	0	NATRONA	UNKNOWN
field	1395	1395	100	7	61	0	0	0	0	East Teapot	Wildcat
state	1395	1395	100	3	61	0	0	0	0	UNKNOWN	Wyoming
uwi	1395	1395	100	1395	4	0	0	0	0	490250625600	Tst490251031300
well_name	1395	1385	99	20	25	0	0	0	0	Beartooth Federa	ukn
well_number	1395	1384	99	1353	218	0	0	0	0	04211 31 A	State No.
well_operator	1395	1395	100	24	114	0	0	0	0	ADVENTURE	ukn

Exercise: Comparing Table Analysis Results

To compare the table analysis results:

1. Double-click the **Comparing Table Analysis Results** task or right-click the **Comparing Table Analysis Results** task, and select **Add Job** from the pop-up menu.
   A new job is initiated and displays on the **Job and Results Information Pane** on the right side of the **DecisionSpace Data Quality Project Window**.



2. Enter **Job-01** in the **Job Name** field.

3. Enter **Compare Table Analysis** in the **Job Description** field.

4. Select the **This Phase** option for **Results to include**:

   - **This Phase:** The results can be compared within the same phase.

   - **This Project:** The results can be compared within multiple phases in the project.

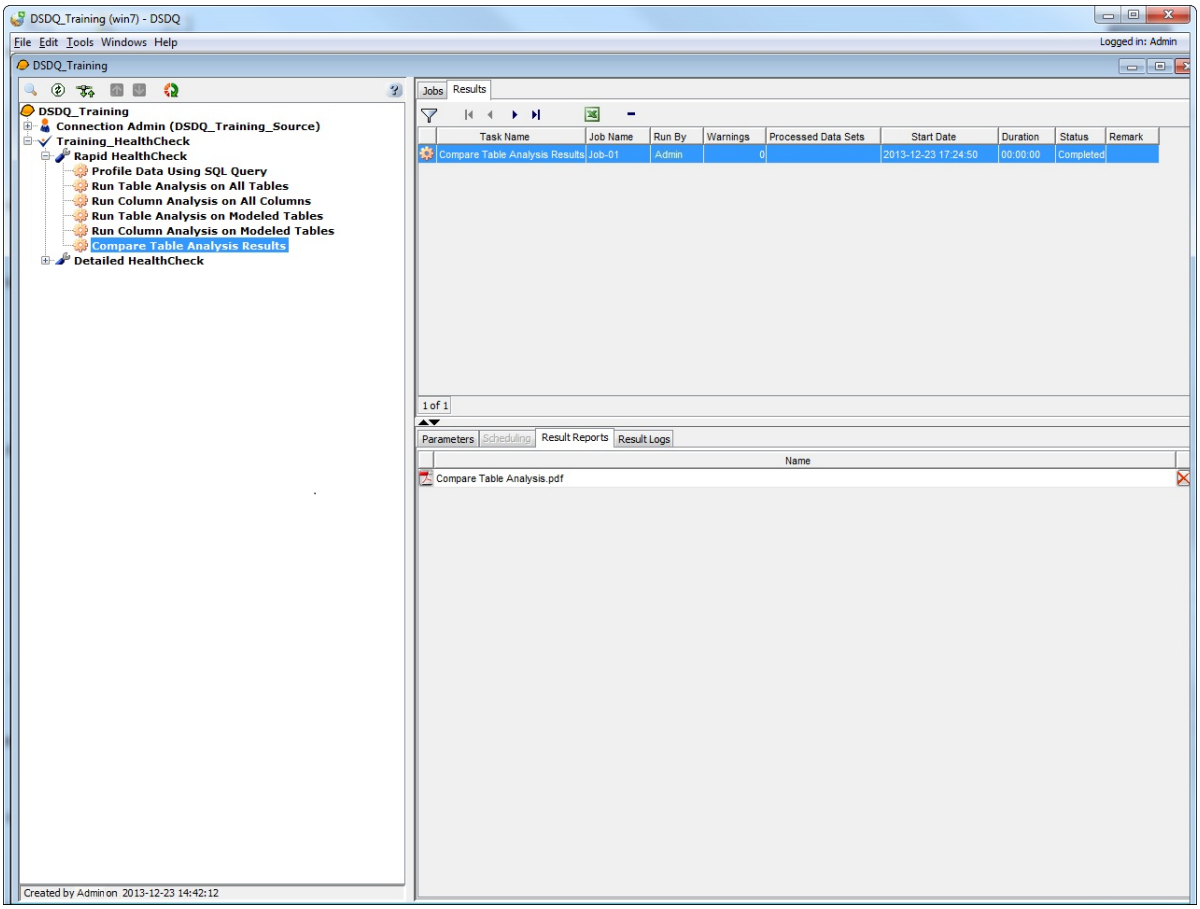
5. Select two results that you want to compare from the **Select Two Analysis Results** list.

6. Select the **After** option for **Delete Results?** Leave the number of days as **7**.

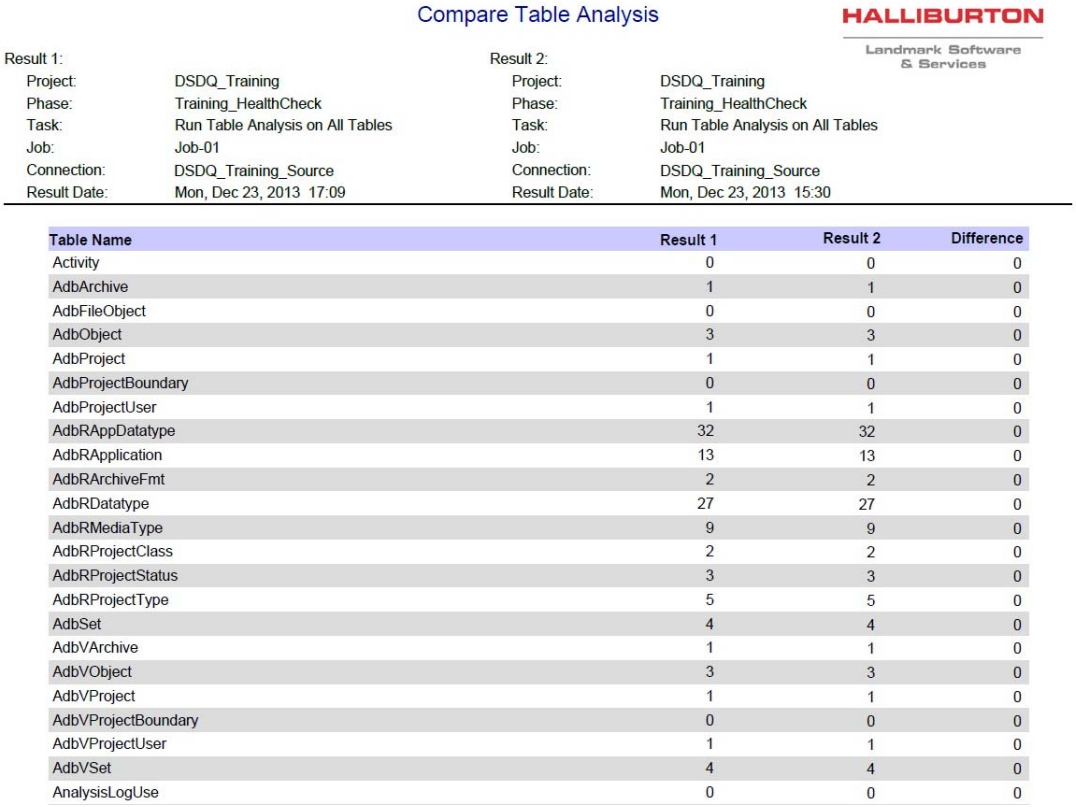
7. Click to save changes in the **Parameters** tab.

8. Click .
   The **Compare Table Analysis Results** task runs and displays results in the **Result Reports** tab on the **Jobs and Results Information Pane**.

9. Select the **Results** tab on the **Jobs and Results Listing Pane** to view the values in the **Result Reports** tab on the **Job and Results Information Pane**.



10. Click on the **Result Reports** tab to display **Compare Table Analysis** results in PDF format.



Compare Table Analysis

HALLIBURTON
Landmark Software & Services

Result 1:		Result 2:	
Project:	DSDQ_Training	Project:	DSDQ_Training
Phase:	Training_HealthCheck	Phase:	Training_HealthCheck
Task:	Run Table Analysis on All Tables	Task:	Run Table Analysis on All Tables
Job:	Job-01	Job:	Job-01
Connection:	DSDQ_Training_Source	Connection:	DSDQ_Training_Source
Result Date:	Mon, Dec 23, 2013 17:09	Result Date:	Mon, Dec 23, 2013 15:30

Table Name	Result 1	Result 2	Difference
Activity	0	0	0
AdbArchive	1	1	0
AdbFileObject	0	0	0
AdbObject	3	3	0
AdbProject	1	1	0
AdbProjectBoundary	0	0	0
AdbProjectUser	1	1	0
AdbRAppDatatype	32	32	0
AdbRApplication	13	13	0
AdbRArchiveFmt	2	2	0
AdbRDatatype	27	27	0
AdbRMediaType	9	9	0
AdbRProjectClass	2	2	0
AdbRProjectStatus	3	3	0
AdbRProjectType	5	5	0
AdbSet	4	4	0
AdbVArchive	1	1	0
AdbVObject	3	3	0
AdbVProject	1	1	0
AdbVProjectBoundary	0	0	0
AdbVProjectUser	1	1	0
AdbVSet	4	4	0
AnalysisLogUse	0	0	0

Identifying Data Issues

The **Detailed HealthCheck** Activity allows you to run business rules against the dataset to identify data problems. It produces a series of comprehensive reports that assist in rating the quality of the data. The tables must be modeled and the elements assigned prior to running the **Detailed HealthCheck** Activity. It also allows you to assign columns from selected submodels to HealthCheck requirements and test the service level to view results in the **Configure Detailed HealthCheck** Tool. The user can select which requirements are to be enabled/disabled in the service level. The user can also select a subset of the total data to be used when testing a service level.

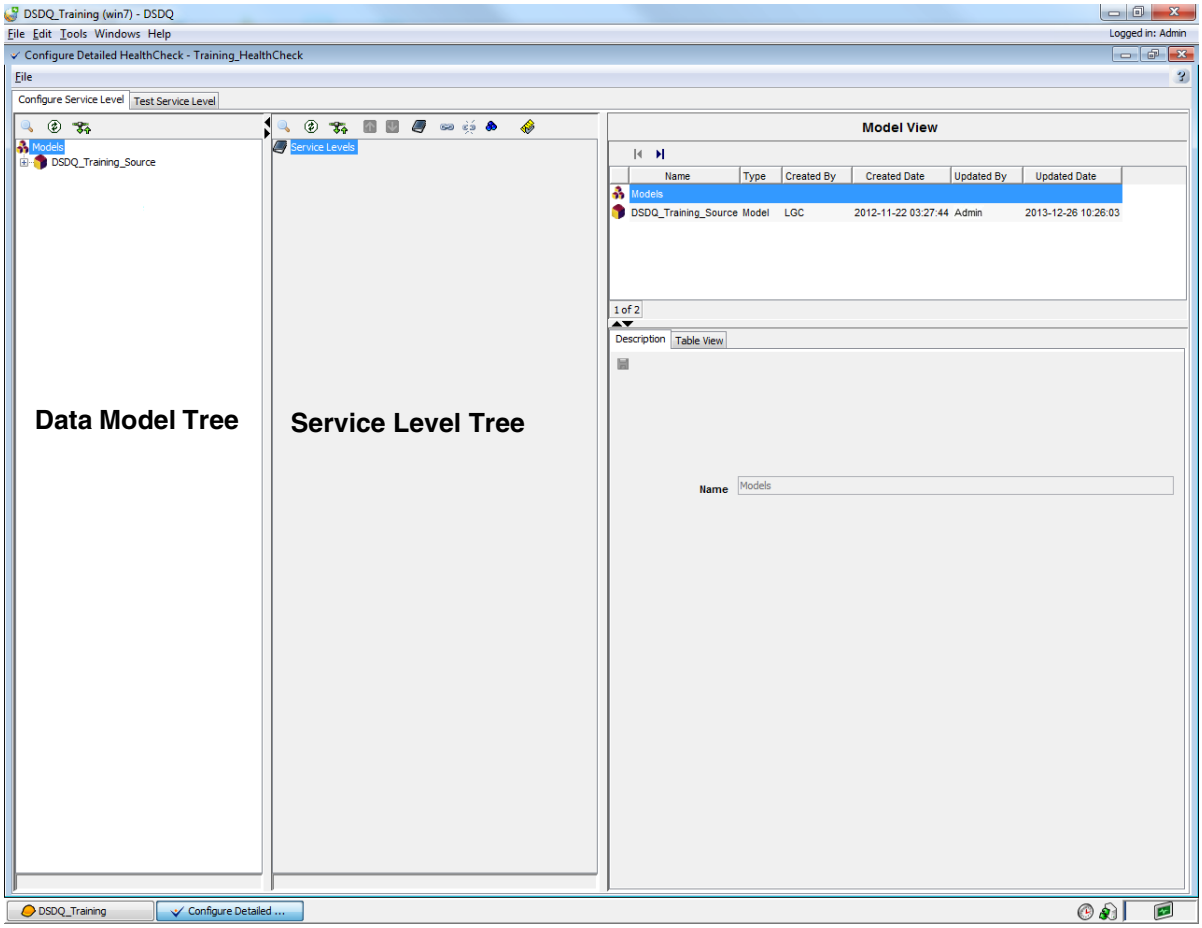
Before running Detailed HealthCheck, you will need to create a new Service Level which is going to be your collection of rules. Refer to Chapter 8, page 7. Refer back to this page after step 4. Enter the following name for your service level: **DSDQ_Training_Rules**.

Exercise: Configuring the Detailed HealthCheck Tool

To configure the detailed HealthCheck Tool:

1. Click ⊞ to expand the **Detailed HealthCheck** Activity.

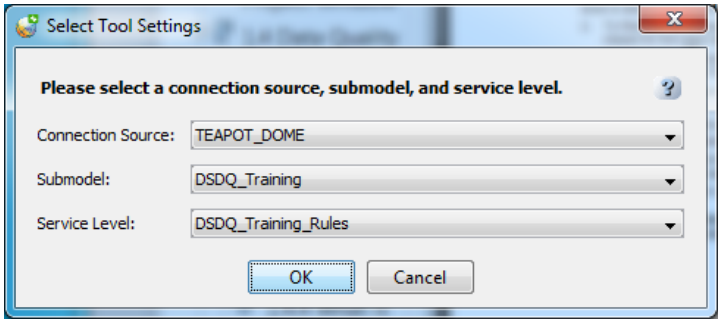
2. Double-click the **Configure Detailed HealthCheck** tool or right-click the **Configure Detailed HealthCheck** tool, and select **Open Tool** from the pop-up menu.
   The **Configure Detailed HealthCheck** window appears.



Note

The **Configure Detailed HealthCheck** Tool has two tabs: **Configure Service Level** and **Test Service Level** tabs located at the top left. With the **Configure Service Level** tab selected, the Data Model Tree (left tree) displays tables used in the currently selected submodel. The Service Level Tree (right tree) shows the currently selected Service Level. The Table View Pane populates with table names or requirements, depending on what tree is selected. The **Description** tab displays the model name.

3. Select **File > Settings** from the File menu bar.
   The **Select Tool Settings** dialog box appears.

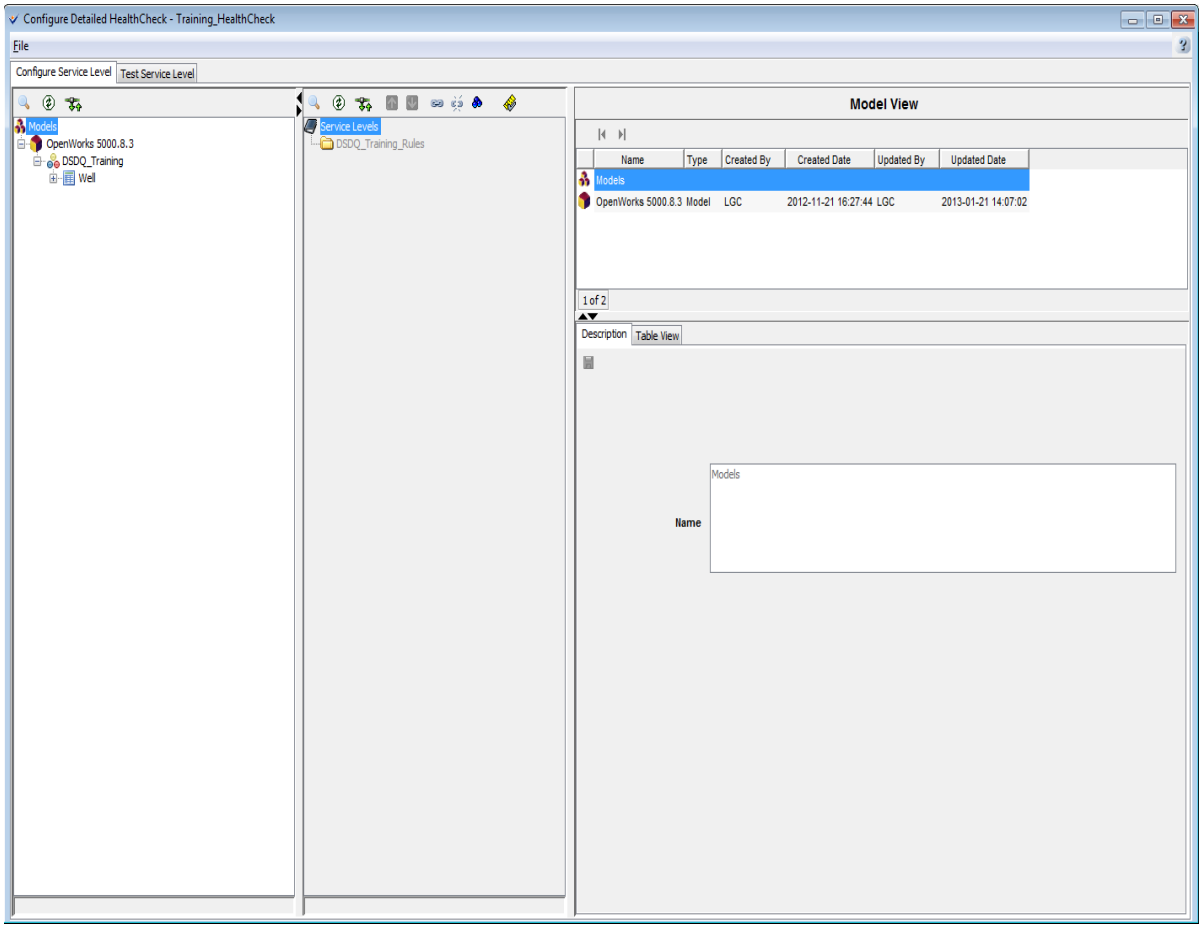


4. Select **TEAPOT_DOME** from the **Connection Source** drop-down list.

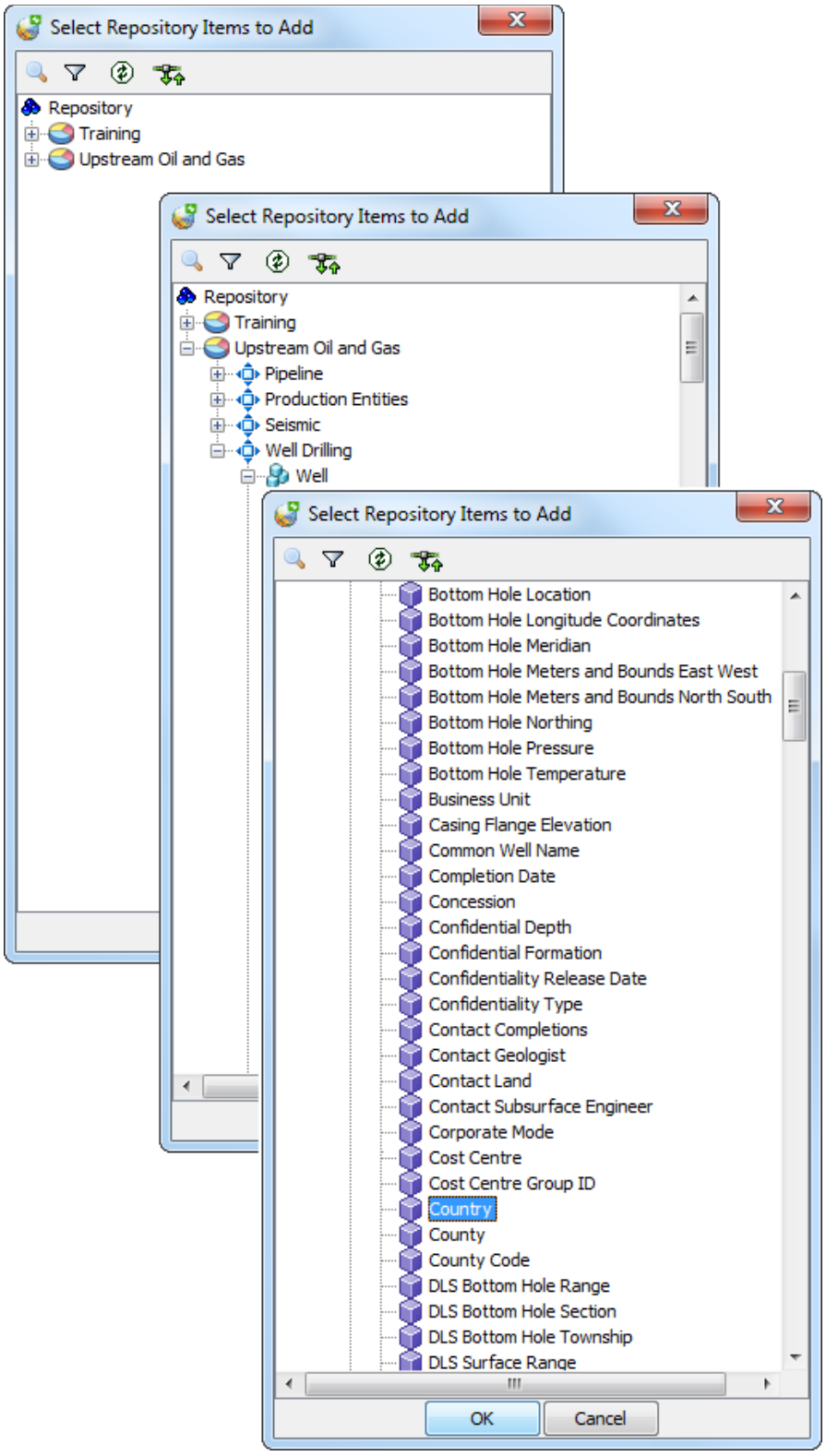
5. Select **DSDQ_Training** from the **Submodel** drop-down list.

6. Select **DSDQ_Training_Rules** from the **Service Level** drop-down list.

7. Click **OK**.
   The **Configure Detailed HealthCheck** window appears with **DSDQ_Training** displaying in the Data Model Tree.



8. Click on the Service Level Tree toolbar.
   The **Select Repository Items to Add** dialog box opens.

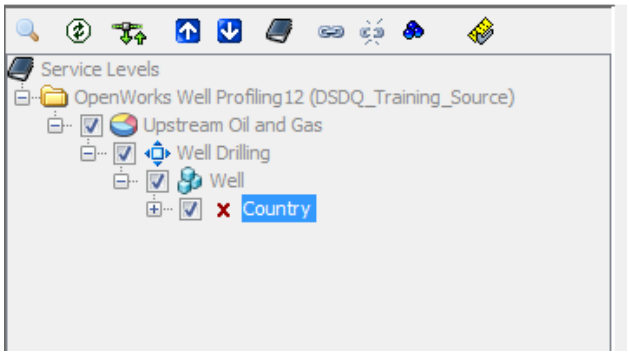


9. Click ⊞ to expand the **Upstream Oil & Gas** sector.

10. Expand the **Well Drilling** area.

11. Expand the **Well** element group.

12. Select the **Country** element and click **OK.**
    The **Country** element is added to the Service Level Tree. A red cross will appear adjacent to the element in the Service Level Tree indicating that the element is not linked to any Column.

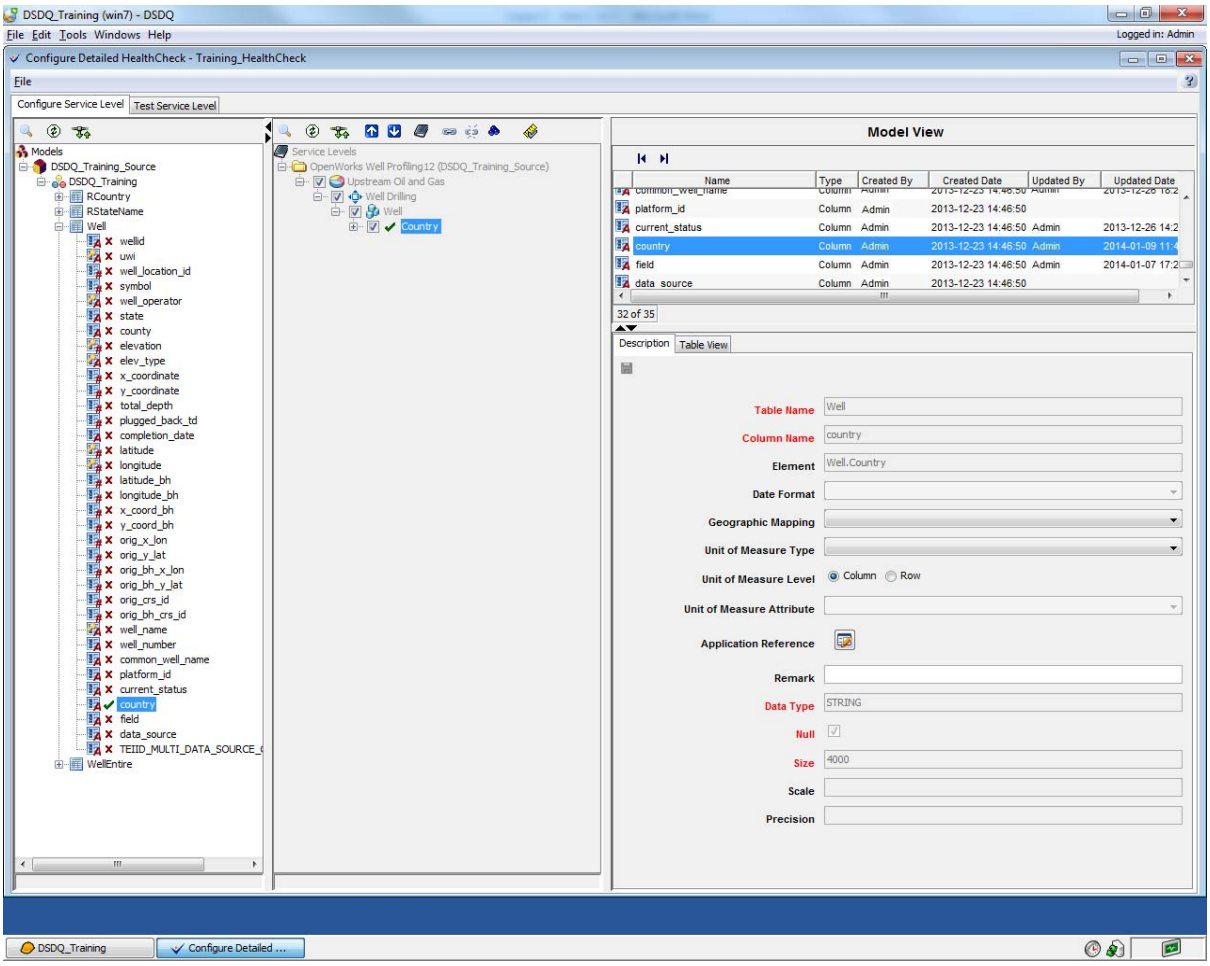


13. Click ⊞ on the Data Model Tree to expand **DSDQ_Training** table.

14. Expand the **Well** table and select the **Country** column.

15. Drag and drop the **Country** column onto the **Country** element in the Service Level Tree.

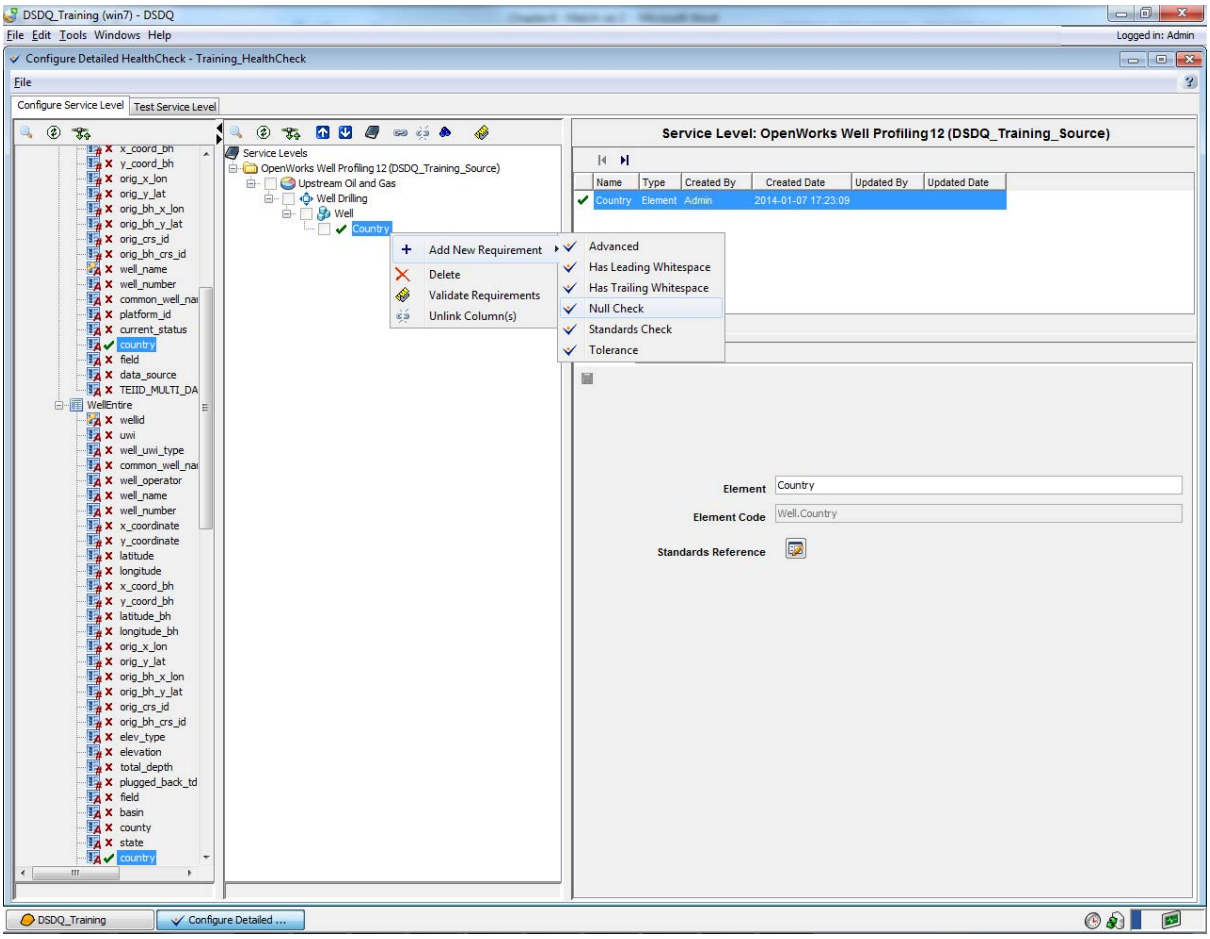
    A green check mark will appear adjacent to the column and element that have just been associated. Only one column from the same table can be linked to the same element. However, it is possible to link many columns to the same element if the columns come from different tables. Alternatively you can select the column & element to link and click the **Link Column to Element** button on the Service Level Tree toolbar.



Note

You can unlink elements or columns. Select the element or column to unlink and click the **Unlink Columns from an Element** button on the Service Level Tree toolbar.

16. Right-click the **Country** element in the Service Level Tree and select **Add New Requirement > Null Check** from the pop-up menu.



Note

The Null Check Requirement determines the number of rows in the column are not populated.

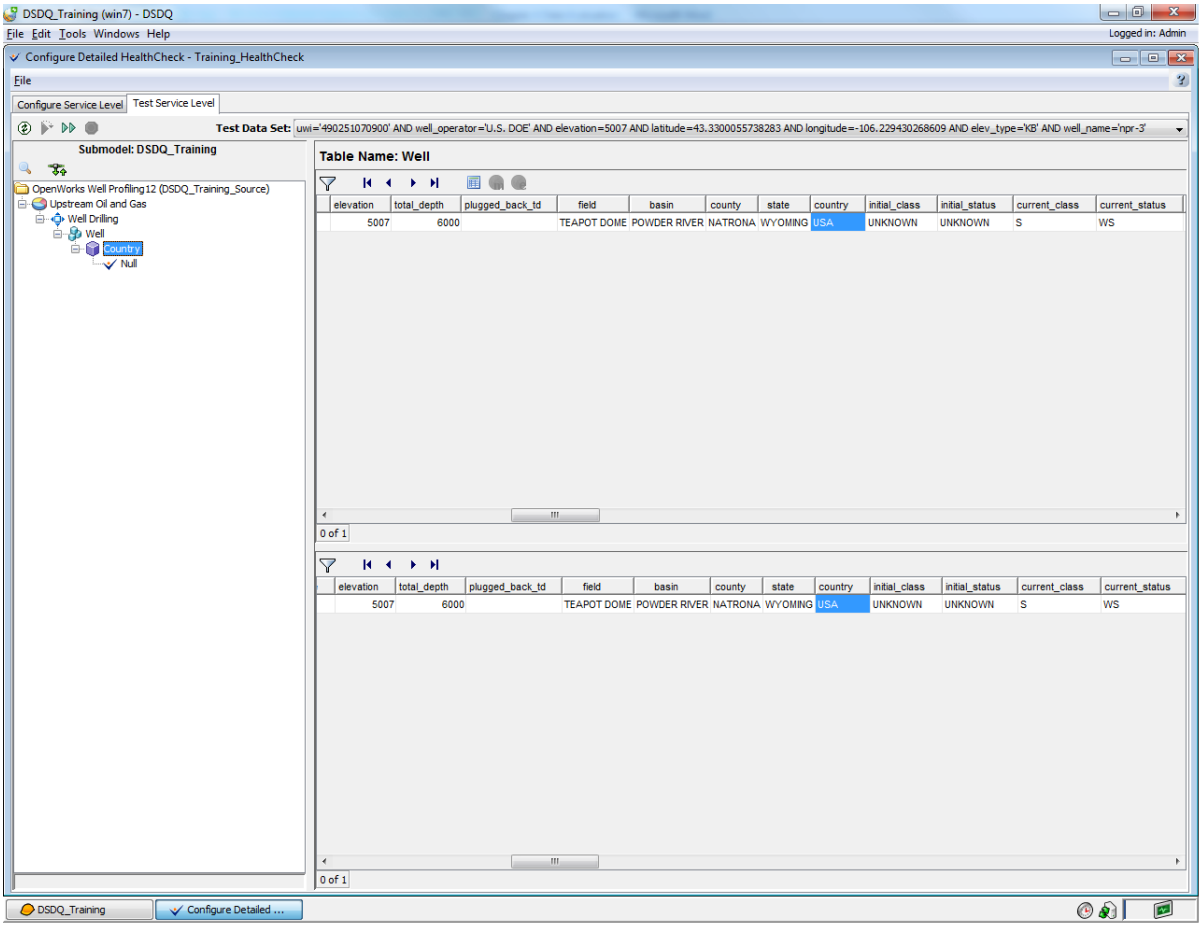
The **Enter Name** dialog box appears.

17. Optionally, specify a user-defined name for the requirement.

18. Click **OK** to add the requirement to the selected element.

19. Optionally, repeat steps **8**to **18** to add all elements for HealthCheck.

20. Click the **Test Service Level** tab.
    The test is automatically executed for the first record of the test data subset.



The corresponding column that the element has been assigned to is highlighted in both panes on the right. The original row data for the record is displayed in the top right pane. The bottom right pane shows the row of data after the service level rules and requirements have been run. Temporary and check columns defined in the rules and requirements are also shown (to see them you may have to scroll all the way to the right). By looking at the columns that have been changed and temporary columns, the user can verify that the behavior of the service level is correct prior to running the **Run Detailed HealthCheck** task. The service level is tested one record at a time.

Note

To view only the affected columns for the selected element in the table view, click the **View Affected Columns** button in the table view toolbar.

21. Click the **Next Data Set** button to test the next record.

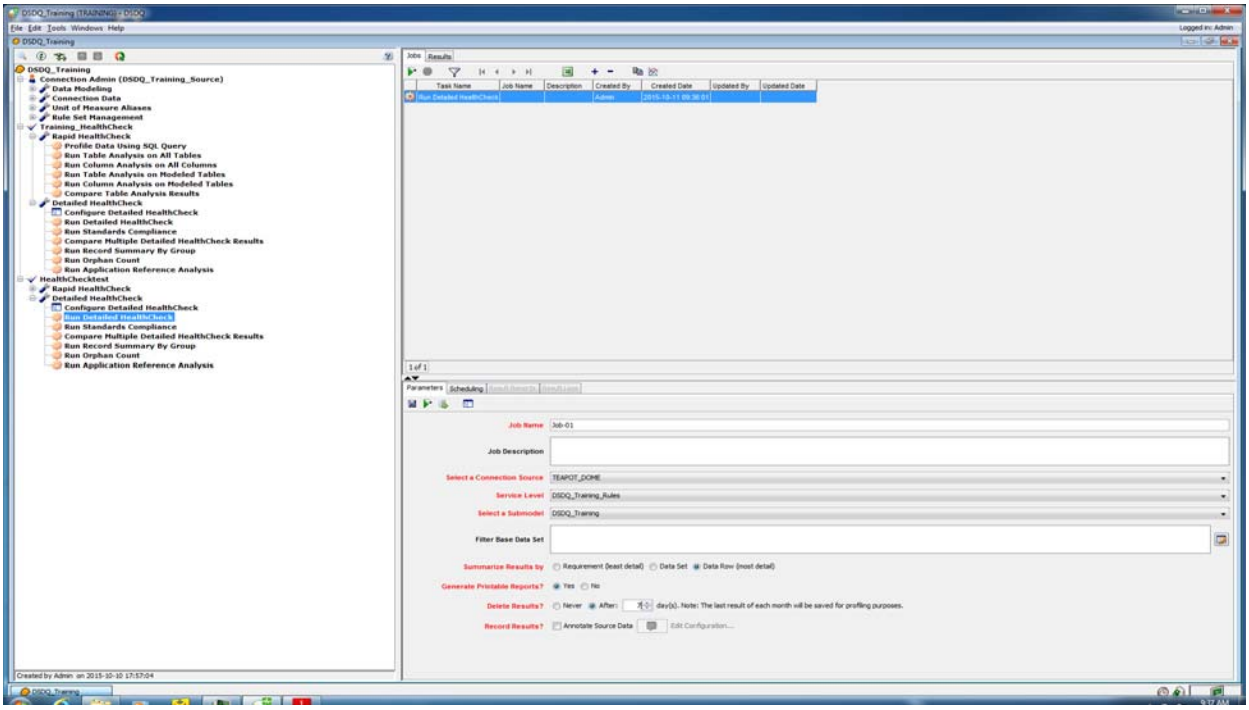
22. Repeat step **21** to test all records.

23. Select **File > Exit** from the menu bar on the **Configure Detailed HealthCheck** window.

Exercise: Running Detailed HealthCheck

To run detailed HealthCheck:

1. Double-click the **Run Detailed HealthCheck** task or right-click the **Run Detailed HealthCheck** task, and select **Add Job** from the pop-up menu.
   A new job is initiated and displays on the **Job and Results Information Pane** on the right side of the **DecisionSpace Data Quality Project Window**.



2. Enter **Job-01** in the **Job Name** field.

3. Enter **Detailed HealthCheck** in the **Job Description** field.

4. Select **TEAPOT_DOME** from the **Select a Connection Source** drop-down list.

> **Note**
>
> For more information on data owner connections, refer to **Adding a New Data Owner Connection** section in Chapter 2, **Connecting DecisionSpace Data Quality with DecisionSpace Data Server**.

5. Select **DSDQ_Training_Rules** from the **Service Level** drop-down list.

6. Select **DSDQ_Training** from the **Select a Submodel** drop-down list.

7. Optionally, you can select a filter for the dataset.

8. Select **Data Row (most detail)** option for **Summarize Results by**.

9. Select the **Yes** option for **Generate Printable Reports?**

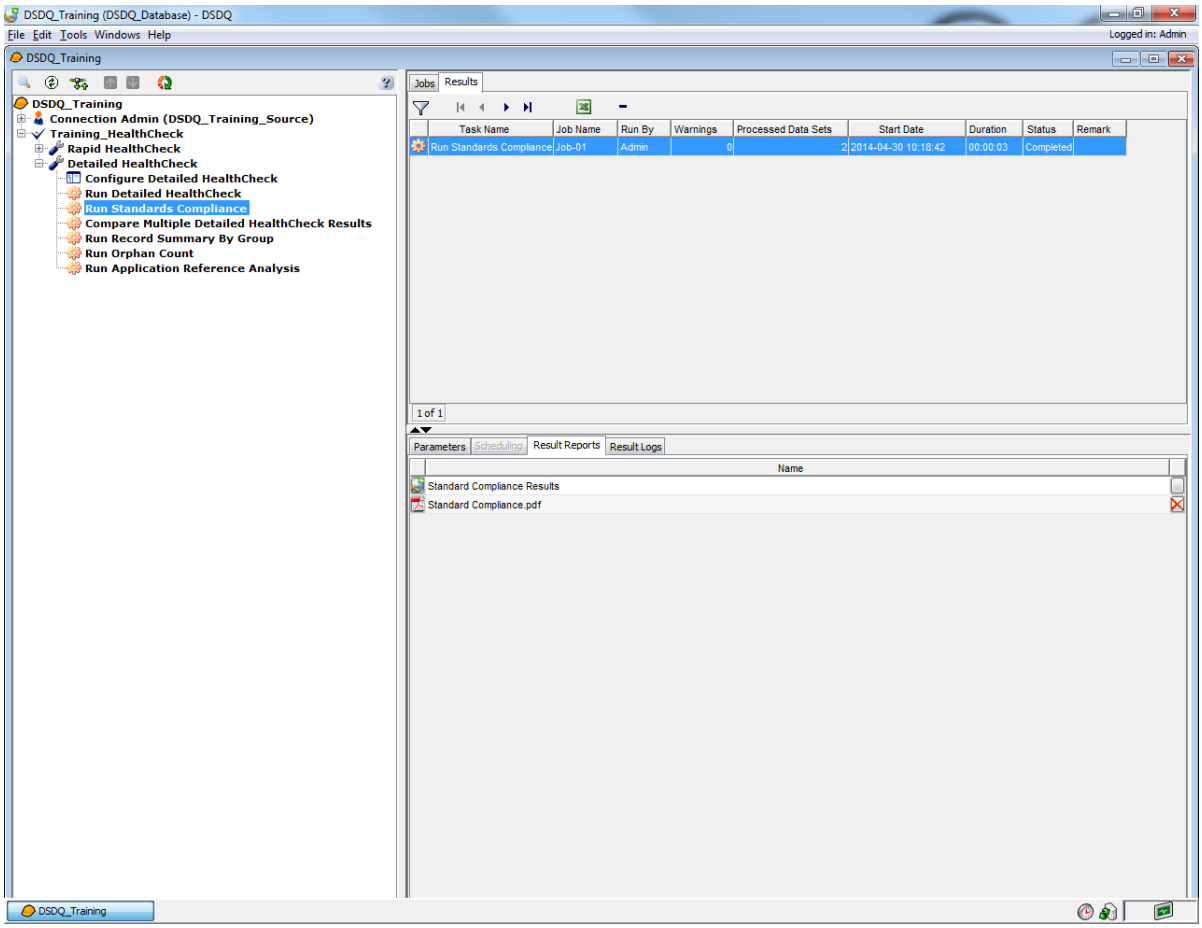
10. Select the **After** option for **Delete Results?** Leave the number of days as **7**.

11. Do not select the check box for **Record Results?**

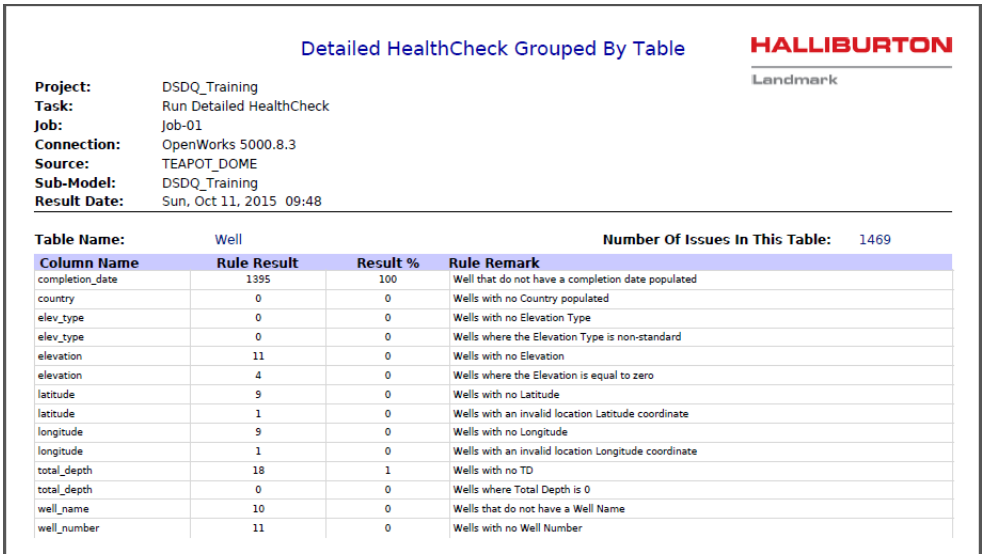
12. Click  to save changes in the **Parameters** tab.

13. Click  .
    The **Detailed HealthCheck** task runs and displays results in the **Result Reports** tab on the **Jobs and Results Information Pane**.

14. Select the **Results** tab on the **Jobs and Results Listing Pane** to view the values in the **Result Reports** tab on the **Job and Results Information Pane**.



15. Click on the **Results Reports** tab to display results for **Detailed HealthCheck Grouped by Table** in PDF format.



Detailed HealthCheck Grouped By Table

HALLIBURTON
Landmark

Project: DSDQ_Training
Task: Run Detailed HealthCheck
Job: Job-01
Connection: OpenWorks 5000.8.3
Source: TEAPOT_DOME
Sub-Model: DSDQ_Training
Result Date: Sun, Oct 11, 2015 09:48

Table Name: Well **Number Of Issues In This Table:** 1469

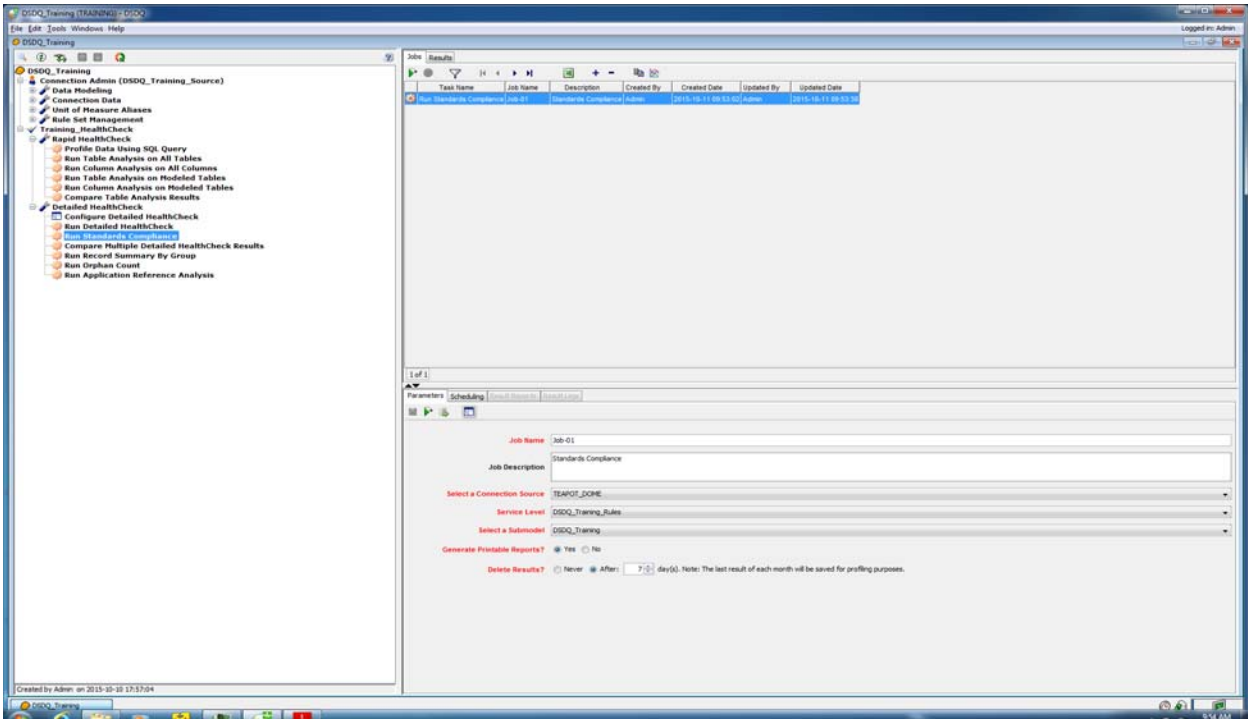
Column Name	Rule Result	Result %	Rule Remark
completion_date	1395	100	Well that do not have a completion date populated
country	0	0	Wells with no Country populated
elev_type	0	0	Wells with no Elevation Type
elev_type	0	0	Wells where the Elevation Type is non-standard
elevation	11	0	Wells with no Elevation
elevation	4	0	Wells where the Elevation is equal to zero
latitude	9	0	Wells with no Latitude
latitude	1	0	Wells with an invalid location Latitude coordinate
longitude	9	0	Wells with no Longitude
longitude	1	0	Wells with an invalid location Longitude coordinate
total_depth	18	1	Wells with no TD
total_depth	0	0	Wells where Total Depth is 0
well_name	10	0	Wells that do not have a Well Name
well_number	11	0	Wells with no Well Number

Exercise: Running Standards Compliance

The **Run Standards Compliance** task checks the number of values within a column. It can also be used to compare the values against the pre-configured values in either the Applications or Standards Reference tables.

To run standards compliance:

1. Double-click the **Run Standards Compliance** task or right-click the **Run Standards Compliance** task and select **Add Job** from the pop-up menu.
   A new job is initiated and displays on the **Job and Results Information Pane** on the right side of the **DecisionSpace Data Quality Project Window**.



2. Enter **Job-01** in the **Job Name** field.

3. Enter **Standards Compliance** in the **Job Description** field.

4. Select **TEAPOT_DOME** from the **Select a Connection Source** drop-down list.

   Note

   For more information on data owner connections, refer to Adding a New Data Owner Connection section in Chapter 2, Connecting DecisionSpace Data Quality with DecisionSpace Data Server.

5. Select **DSDQ_Training_Rules** from the **Service Level** drop-down list.

6. Select **DSDQ_Training** from the **Select a Submodel** drop-down list.

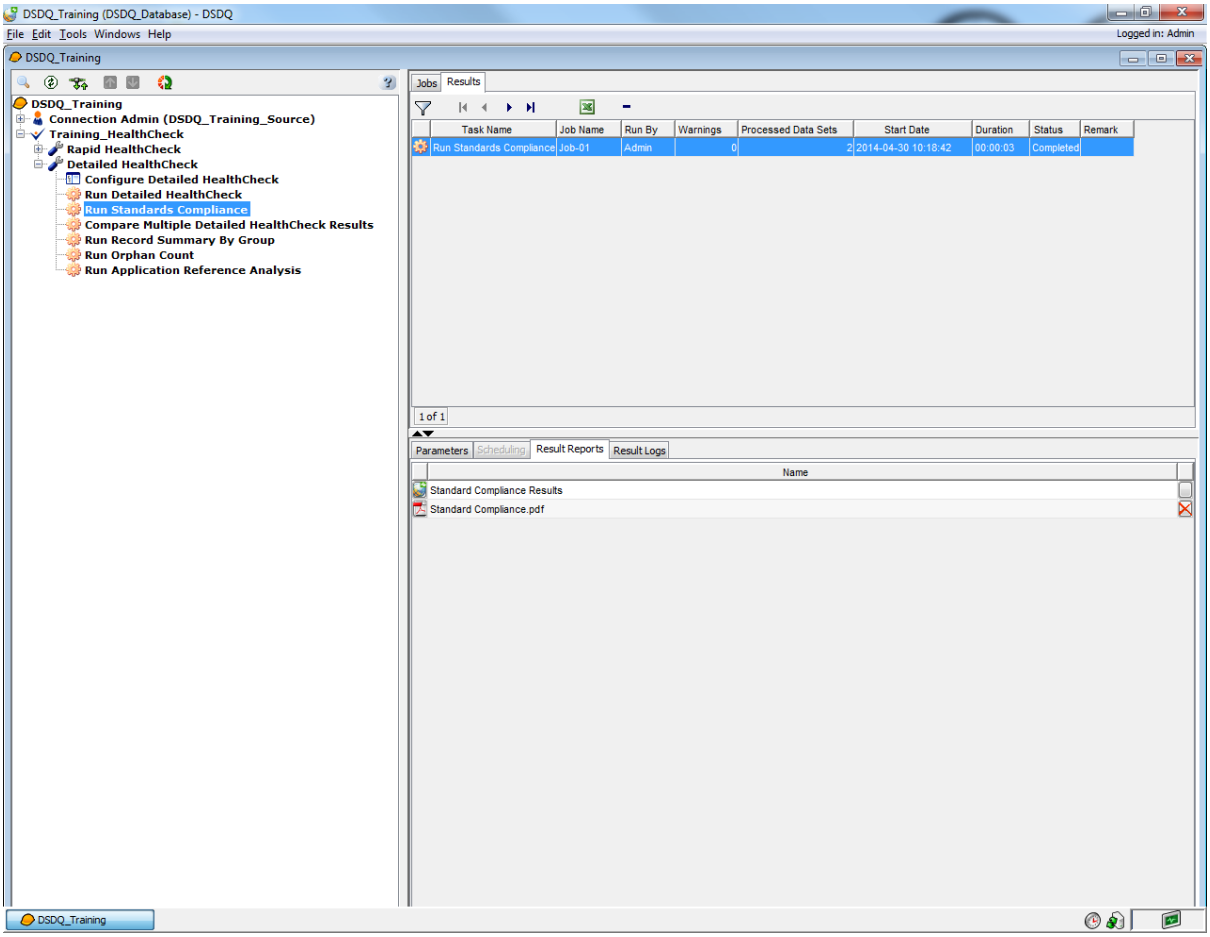
7. Select the **Yes** option for **Generate Printable Reports?**

8. Select the **After** option for **Delete Results?** Leave the number of days as **7**.

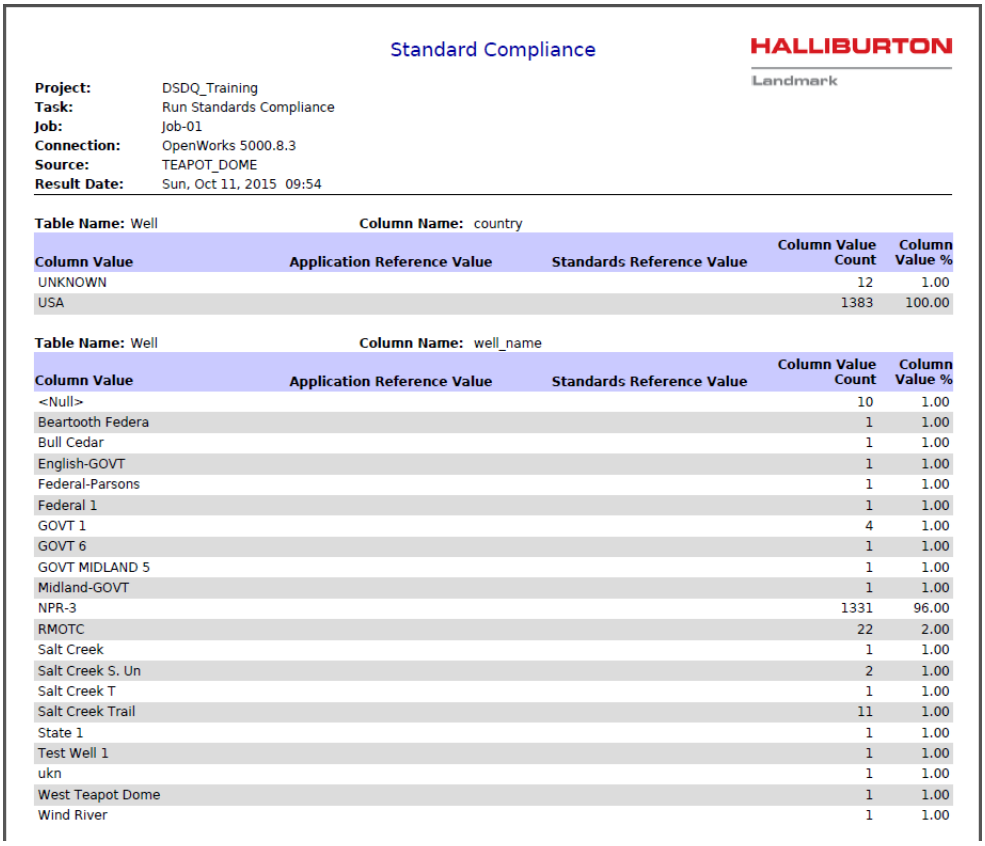
9. Click to save changes in the **Parameters** tab.

10. Click .
    The **Run Standards Compliance** task runs and displays results in the **Result Reports** tab on the **Jobs and Results Information Pane**.

11. Select the **Results** tab on the **Jobs and Results Listing Pane** to view the values in the **Result Reports** tab on the **Job and Results Information Pane**.



12. Click on the **Result Reports** tab to display results for **Standard Compliance** in PDF format.



Standard Compliance

HALLIBURTON
Landmark

Project: DSDQ_Training
Task: Run Standards Compliance
Job: Job-01
Connection: OpenWorks 5000.8.3
Source: TEAPOT_DOME
Result Date: Sun, Oct 11, 2015 09:54

Table Name: Well **Column Name:** country

Column Value	Application Reference Value	Standards Reference Value	Column Value Count	Column Value %
UNKNOWN			12	1.00
USA			1383	100.00

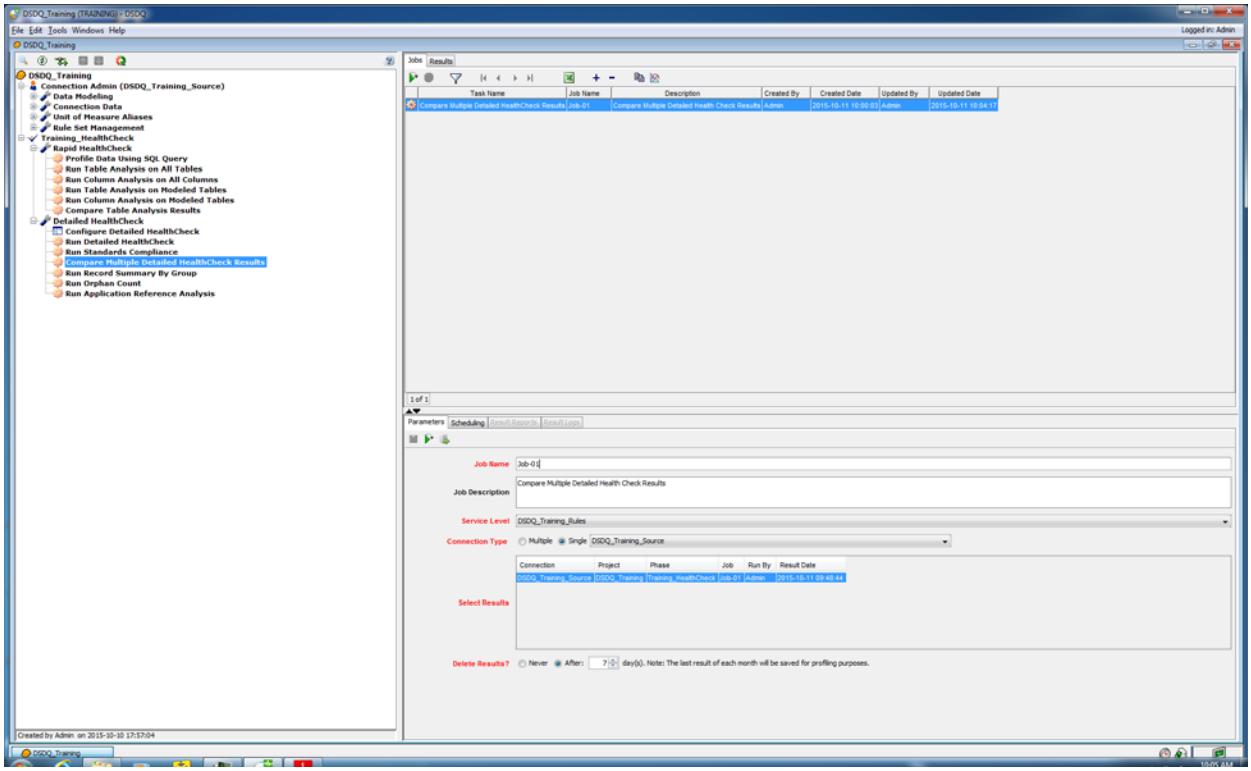
Table Name: Well **Column Name:** well_name

Column Value	Application Reference Value	Standards Reference Value	Column Value Count	Column Value %
<Null>			10	1.00
Beartooth Federa			1	1.00
Bull Cedar			1	1.00
English-GOVT			1	1.00
Federal-Parsons			1	1.00
Federal 1			1	1.00
GOVT 1			4	1.00
GOVT 6			1	1.00
GOVT MIDLAND 5			1	1.00
Midland-GOVT			1	1.00
NPR-3			1331	96.00
RMOTC			22	2.00
Salt Creek			1	1.00
Salt Creek S. Un			2	1.00
Salt Creek T			1	1.00
Salt Creek Trail			11	1.00
State 1			1	1.00
Test Well 1			1	1.00
ukn			1	1.00
West Teapot Dome			1	1.00
Wind River			1	1.00

Exercise: Comparing Multiple Detailed HealthCheck Results

To compare multiple Detailed HealthCheck results:

1. Double-click the **Compare Multiple Detailed HealthCheck Results** task or right-click the **Compare Multiple Detailed HealthCheck Results** task, and select **Add Job** from the pop-up menu.
   A new job is initiated and displays on the **Job and Results Information Pane** on the right side of the **DecisionSpace Data Quality Project Window**.



2. Enter **Job-01 i**n the **Job Name** field.

3. Enter **Compare Multiple Detailed HealthCheck Results** in the **Job Description** field.

4. Select **DSDQ_Training_Rules** from the **Service Level** drop-down list.

5. Select the **Single** option for **Connection Type**.

    - **Multiple** connections automatically list the latest Detailed HealthCheck job run for a connection (i.e., PPDM, WellView, etc.) against the selected Service Level.

    - **Single** connections list only the jobs run under the specified connection for the selected Service Level.

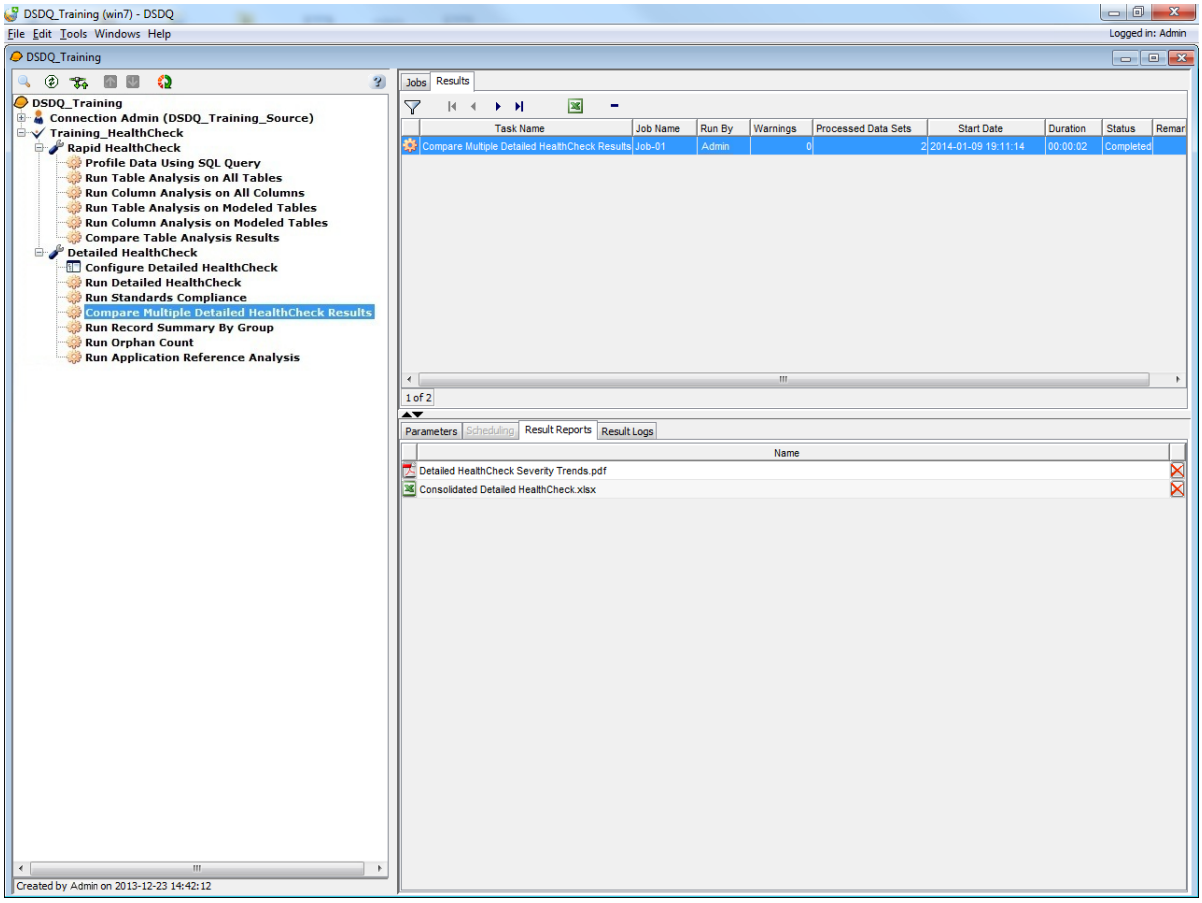
6. Select **DSDQ_Training_Source** from the **Connection Type** drop-down list.

7. Select **DSDQ_Training_Source** from the from **Select Results** list.

8. Select the **After** option for **Delete Results?** Leave the number of days as **7**.

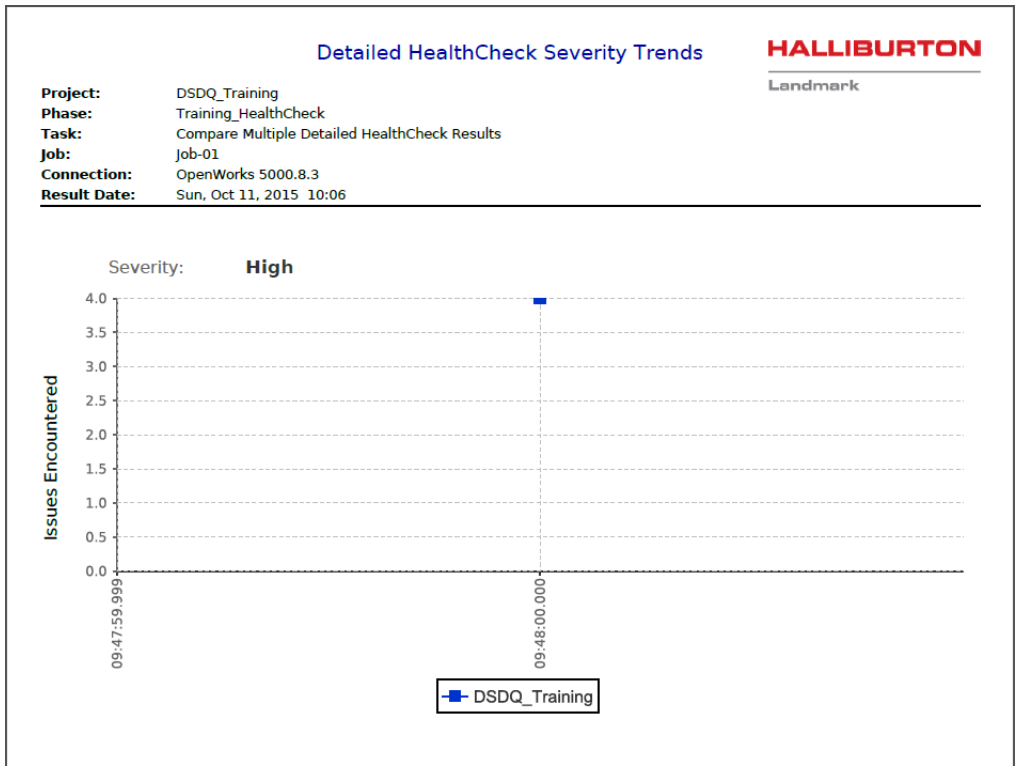
9. Click  to save changes in the Parameters tab.

10. Click .
    The **Compare Multiple Detailed HealthCheck Results** task runs and displays results in the **Result Reports** tab on the **Jobs and Results Information Pane**.



11. Select the **Results** tab on the **Jobs and Results Listing Pane** to view the values in the **Result Reports** tab on the **Job and Results Information Pane**.

12. Click on the **Result Reports** tab to display **Detailed HealthCheck Severity Trends** in PDF format.

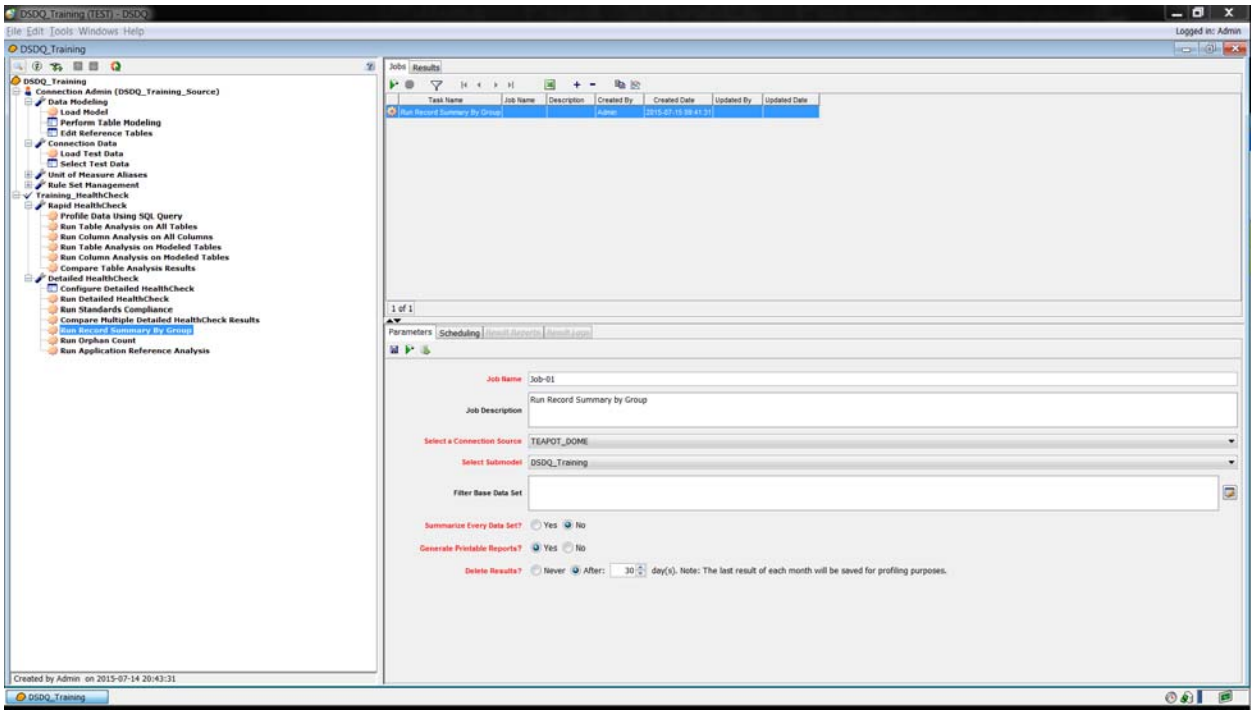


Exercise: Running Record Summary by Group

The **Run Record Summary by Group** task provides a basic record count based on data groups. Depending on your choice, the task checks the basic row count by table or data group. For example, if the Well Header data group has 10 records and the Drilling Event data group has 50 records, the task counts the number of events that belong to each well.

To run record summary by group(s):

1. Double-click the **Run Record Summary By Group** task or right-click the **Run Record Summary By Group** task, and select **Add Job** from the pop-up menu.
   A new job is initiated and displays on the **Job and Results Information Pane** on the right side of the **DecisionSpace Data Quality Project Window**.



2. Enter **Job-01** in the **Job Name** field.

3. Enter **Summary by Group** in the **Job Description** field.

4. Select **TEAPOT_DOME** from the **Select a Connection Source** drop-down list.

> **Note**
>
> For more information on data owner connections, refer to Adding a New Data Owner Connection section in Chapter 2, Connecting DecisionSpace Data Quality with DecisionSpace Data Server.

5. Select **DSDQ_Training** from the **Select a Submodel** drop-down list.

6. Optionally, set a filter on the data subset.

7. Select **No** from **Summarize Every Data Set**.

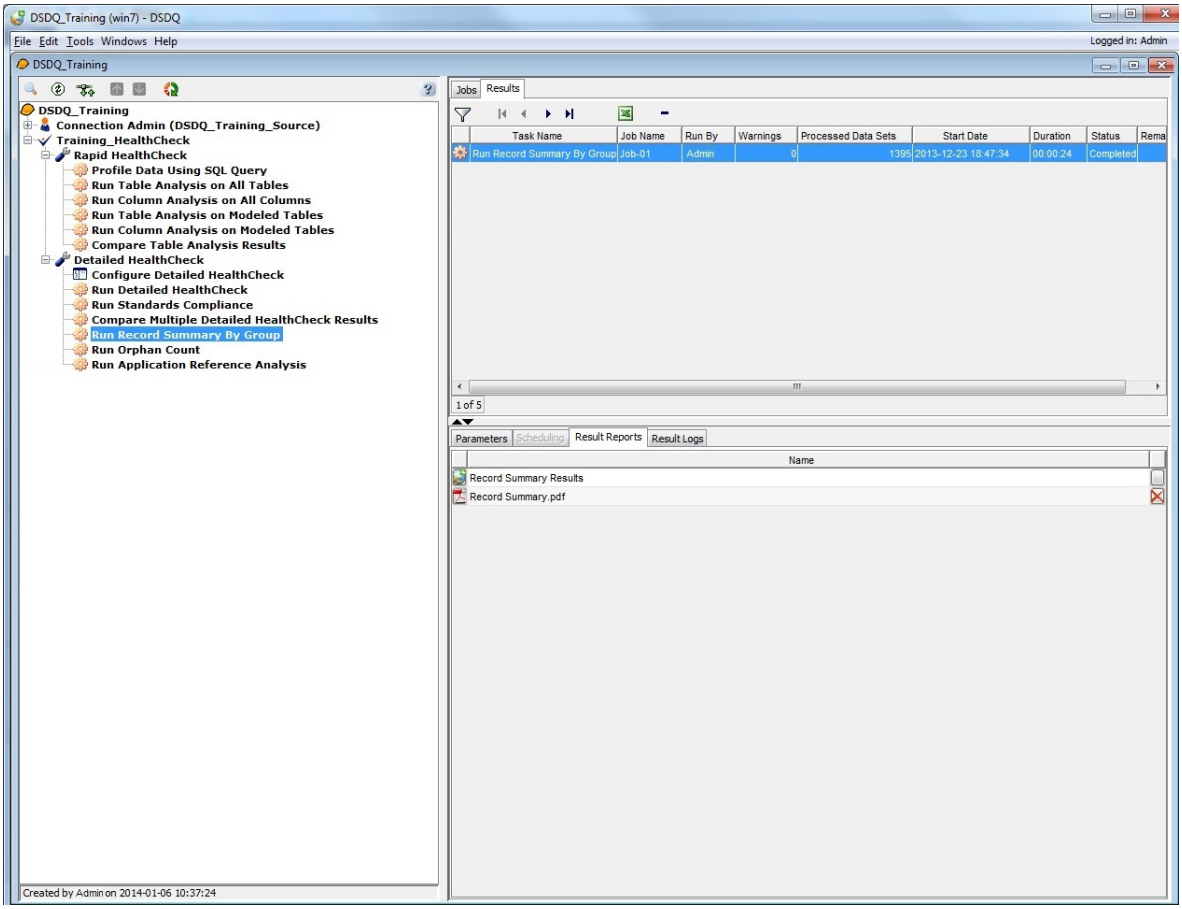
8. Select **Yes** from **Generate Printable Reports?**

9. Select the **After** option for **Delete Results?** Leave the number of days as **7**.

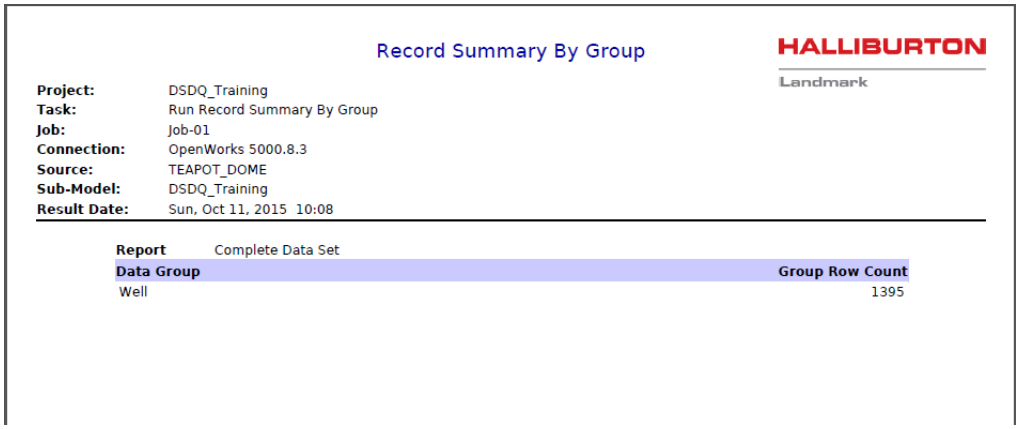
10. Click to save changes in the **Parameters** tab.

11. Click .
    The **Run Record Summary by Group** task runs and displays results in the **Result Reports** tab on the **Jobs and Results Information Pane**.

12. Select the **Results** tab on the **Jobs and Results Listing Pane** to view the values in the **Result Reports** tab on the **Job and Results Information Pane**.



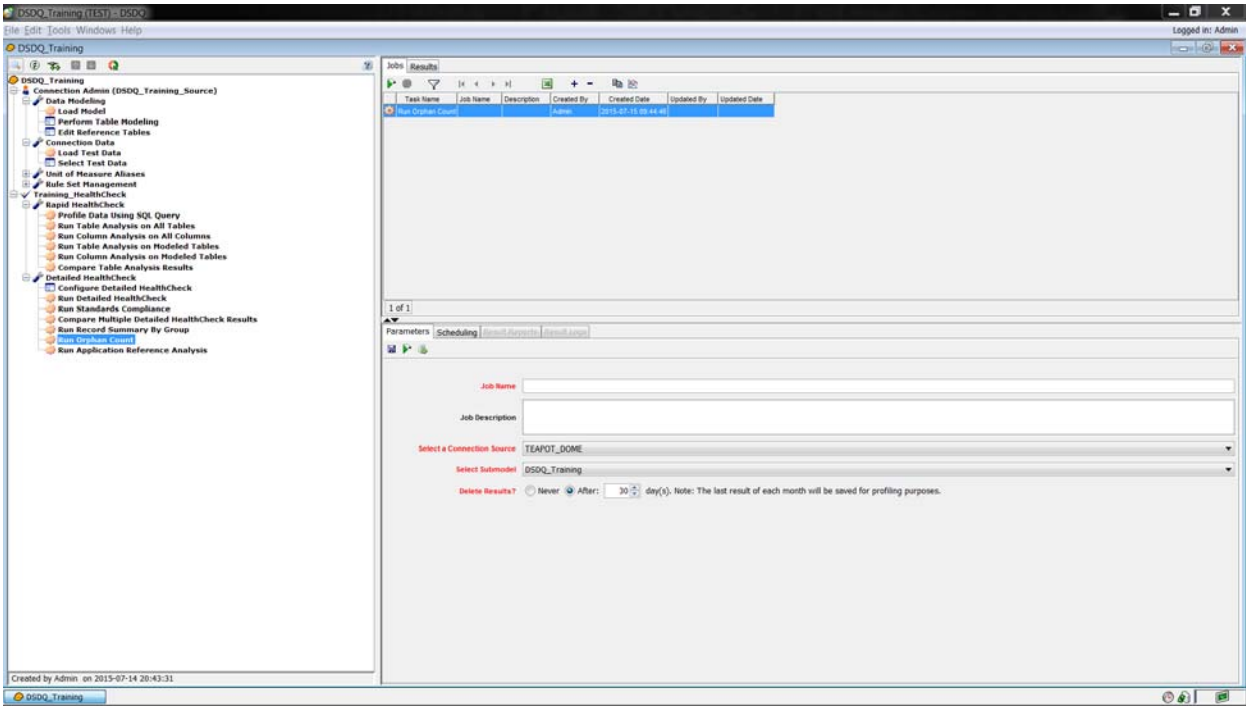
13. Click on the **Result Reports** tab to display **Record Summary By Group** in PDF format.



Exercise: Running Orphan Count

To run orphan count:

1. Double-click the **Run Orphan Count** task or right-click the **Run Orphan Count** task, and select **Add Job** from the pop-up menu. A new job is initiated and displays on the **Job and Results Information Pane** on the right side of the **DecisionSpace Data Quality Project Window**.



2. Enter **Job-01** in the **Job Name** field.

3. Enter **Orphan Count** in the **Job Description** field.

4. Select **TEAPOT_DOME** from the **Select a Connection Source** drop-down list.

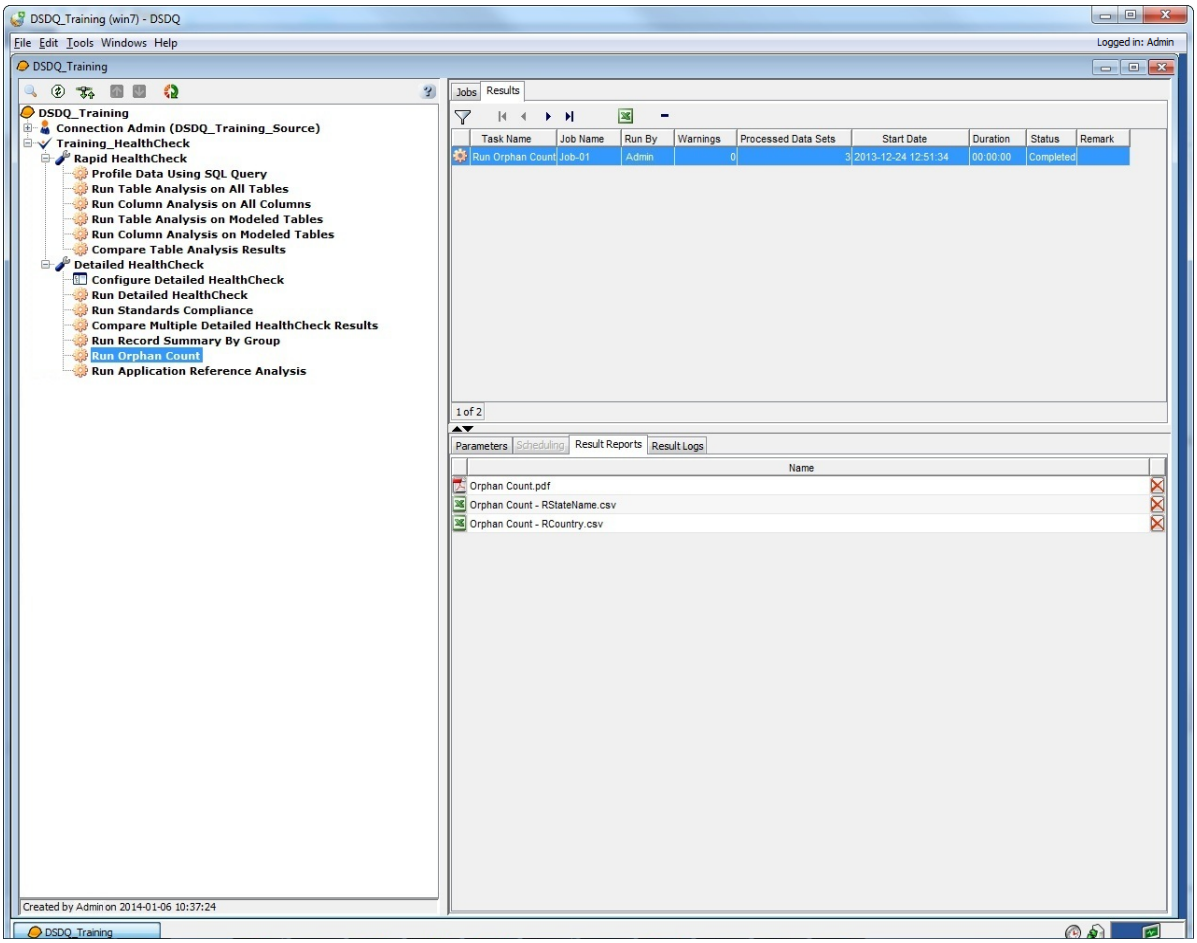
5. Select **DSDQ_Training** from the **Select a Submodel** drop-down list.

6. Select the **After** option for **Delete Results?** Leave the number of days as **7**.

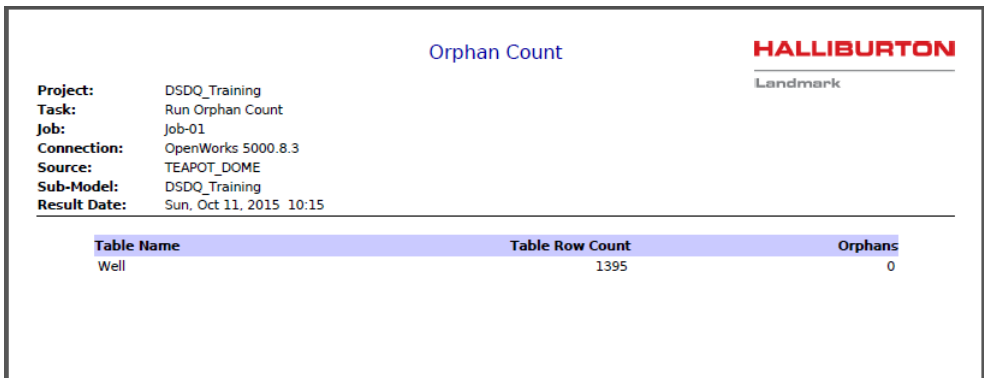
7. Click to save changes in the **Parameters** tab.

8. Click .
   The **Run Orphan Count** task runs and displays results in the **Result Reports** tab on the **Jobs and Results Information Pane**.

9. Select the **Results** tab on the **Jobs and Results Listing Pane** to view the values in the **Result Reports** tab on the **Job and Results Information Pane**.



10. Click on the **Result Reports** tab to display the **Orphan Count** in PDF format.



Exercise: Running Application Reference Analysis

The **Run Application Reference Analysis** task is used to check which columns are associated with a particular Application Reference table, and how many values in each column match a specific Reference table value. In order to perform this task, you will need to set an Application Reference Table, and assign it to a Column. (Application References and Standard References are lookup Columns.)

1. Double-click **Perform Table Modeling**.

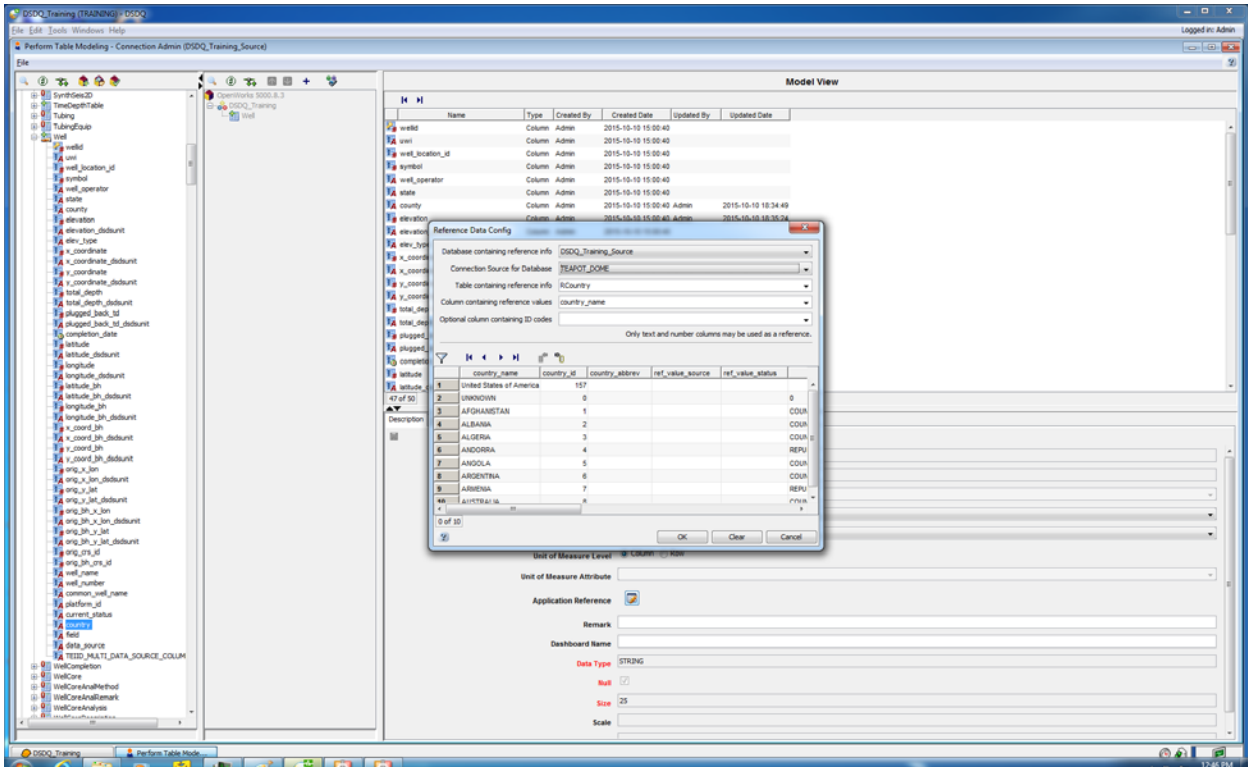
2. Right-click **RCountry** from the Table and Column Listings Tree and select the **Move to Folder > Used** option from the menu.

3. Right-click the **RCountry** table again and select **Reference Table > Local Reference** from the menu.

4. Once RCountry has been added as a local reference, it needs to be set to a column.

5. Expand **Well** in the left tree and click the country column.

6. Click the description tab and then click the **Application Reference** icon.
   The Reference Data Config dialog box displays.



7. Select **DSDQ_Training_Source** for **Database containing reference info**.

8. Select **TEAPOT_DOME** for **Connection Source for Database**.

9. Select **RCountry** for **Table containing reference info**.

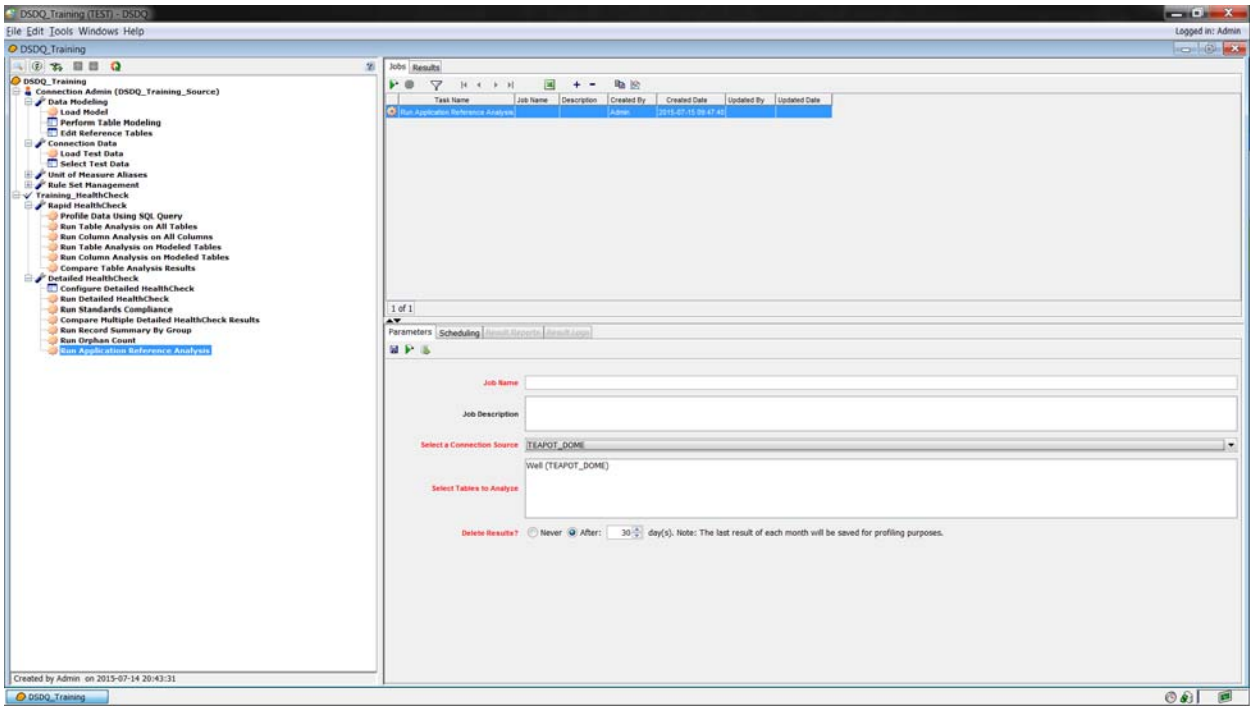
10. Select **country_name** for **Column containing reference values**.

11. Leave the Optional column containing ID Codes blank.

12. Click **OK**. Now the Application Reference is set.

To run application reference analysis:

1. Double-click the **Run Application Reference Analysis** task or right-click the **Run Application Reference Analysis** task, and select **Add Job** from the pop-up menu.
   A new job is initiated and displays on the **Job and Results Information Pane** on the right side of the **DecisionSpace Data Quality Project Window**.



2. Enter **Job-01** in the **Job Name** field.

3. Enter **Application Reference Analysis** in the **Job Description** field.

4. Select **TEAPOT_DOME** from the **Select a Connection Source** drop-down list.

> **Note**
>
> For more information on data owner connections, refer to Adding a New Data Owner Connection section in Chapter 2, **Connecting DecisionSpace Data Quality with DecisionSpace Data Server**.

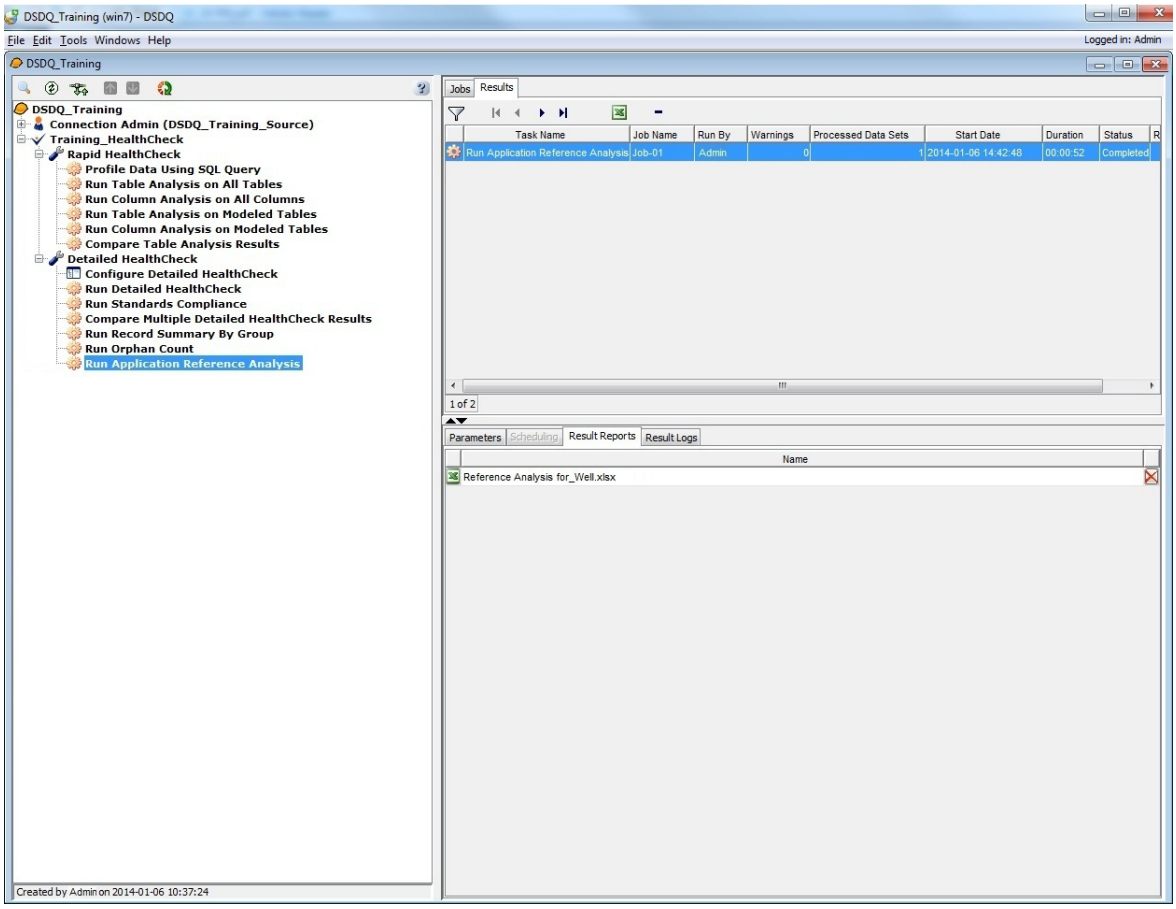
5. Select **Well** from the **Select Tables to Analyze** list.

6. Select the **After** option for **Delete Results?** Leave the number of days as **7**.

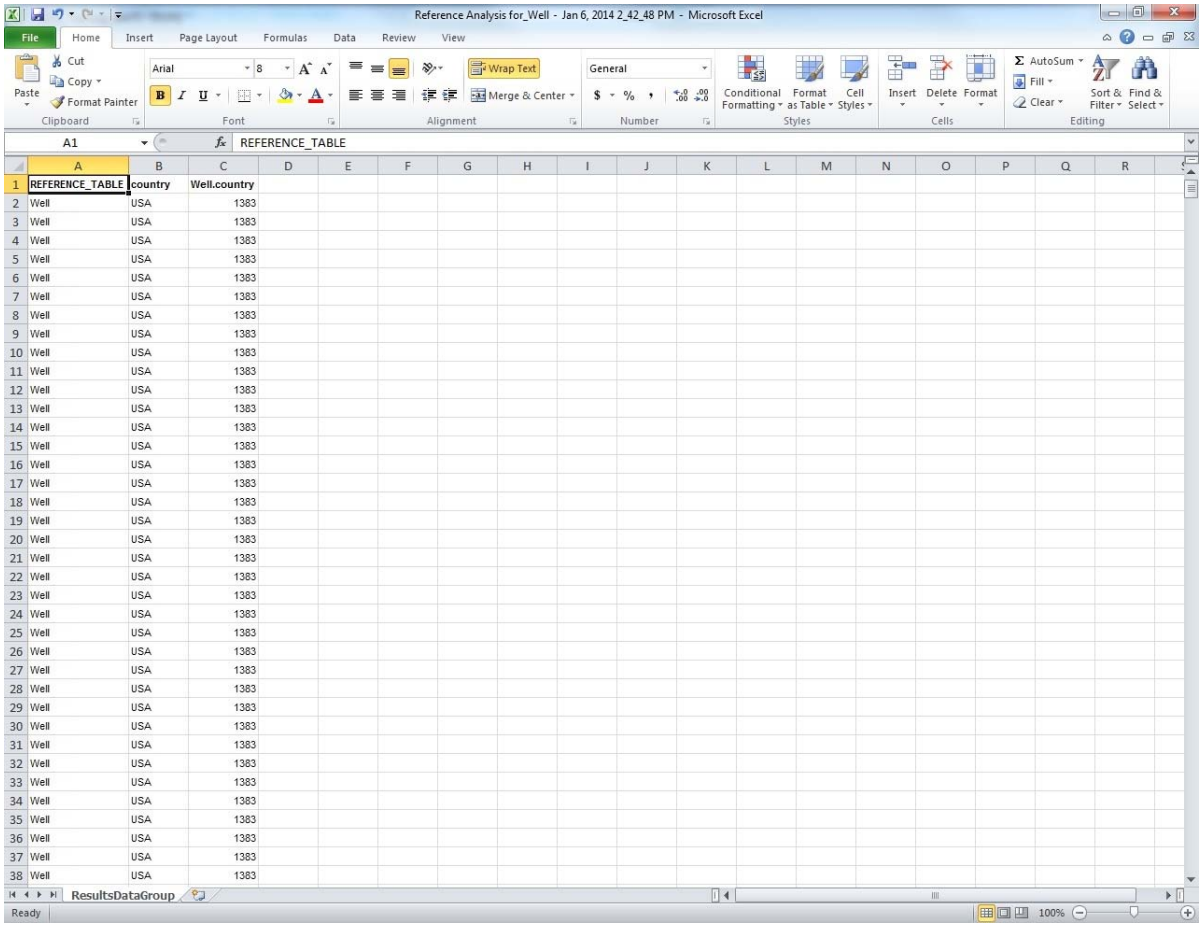
7. Click to save changes in the **Parameter** tab.

8. Click .
   The **Run Application Reference Analysis** task is executed and displays results in the **Result Reports** tab.

9. Select the **Results** tab on the **Jobs and Results Listing Pane** to view the values in the **Result Reports** tab on the **Job and Results Information Pane**.



10. Double-click on the **Result Reports** tab to display the **Reference Analysis for_Well spreadsheet** in Microsoft Excel.
    The **Reference Analysis for_Well s**preadsheet appears.



	A	B	C
1	REFERENCE_TABLE	country	Well.country
2	Well	USA	1383
3	Well	USA	1383
4	Well	USA	1383
5	Well	USA	1383
6	Well	USA	1383
7	Well	USA	1383
8	Well	USA	1383
9	Well	USA	1383
10	Well	USA	1383
11	Well	USA	1383
12	Well	USA	1383
13	Well	USA	1383
14	Well	USA	1383
15	Well	USA	1383
16	Well	USA	1383
17	Well	USA	1383
18	Well	USA	1383
19	Well	USA	1383
20	Well	USA	1383
21	Well	USA	1383
22	Well	USA	1383
23	Well	USA	1383
24	Well	USA	1383
25	Well	USA	1383
26	Well	USA	1383
27	Well	USA	1383
28	Well	USA	1383
29	Well	USA	1383
30	Well	USA	1383
31	Well	USA	1383
32	Well	USA	1383
33	Well	USA	1383
34	Well	USA	1383
35	Well	USA	1383
36	Well	USA	1383
37	Well	USA	1383
38	Well	USA	1383

Chapter 5
Data Cleansing and Standardization

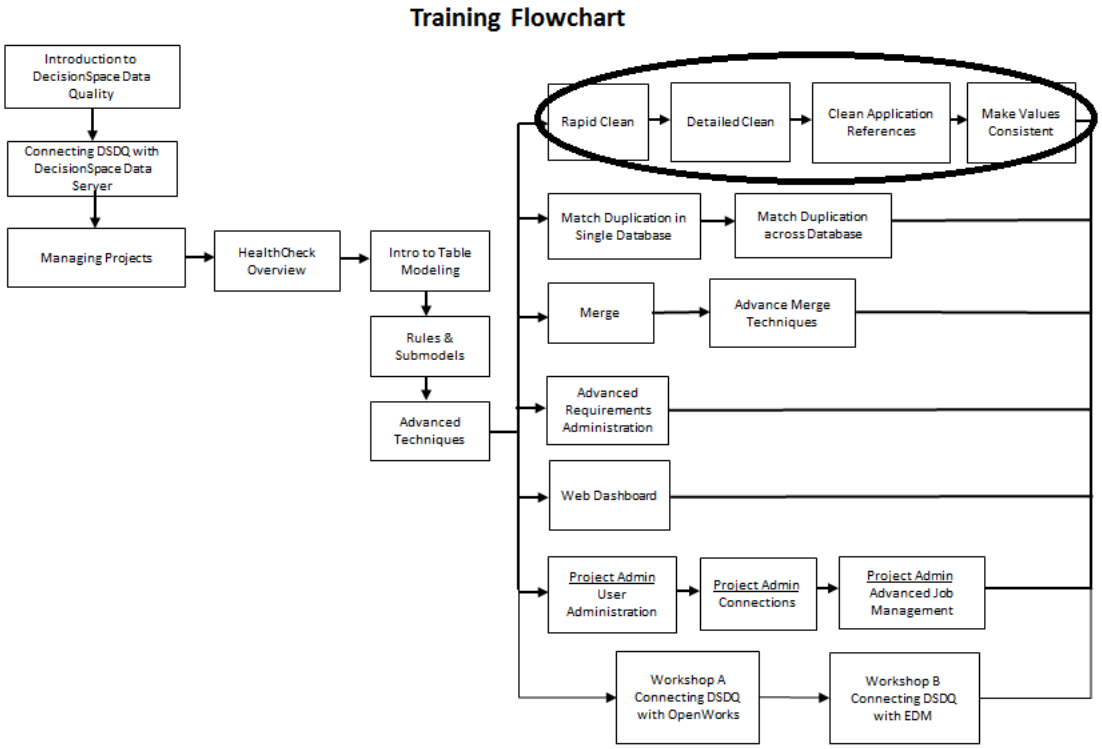
The Clean Phase enables you to address data quality issues across disparate data stores by means of quality control queries and a repeatable cleansing methodology. This Phase includes two Activities, **Rapid Clean** and **Detailed Clean**, and it is designed to work through data issues found during the HealthCheck phase.

Chapter Overview

In this chapter, you will learn about:

- Using the Rapid Clean Activity
- Using the Detailed Clean Activity
- Cleaning Application References
- Using the Making Values Consistent Tool

Topics covered in each chapter are outlined in the following illustration. Those specific to the current chapter will be circled in black for your reference:



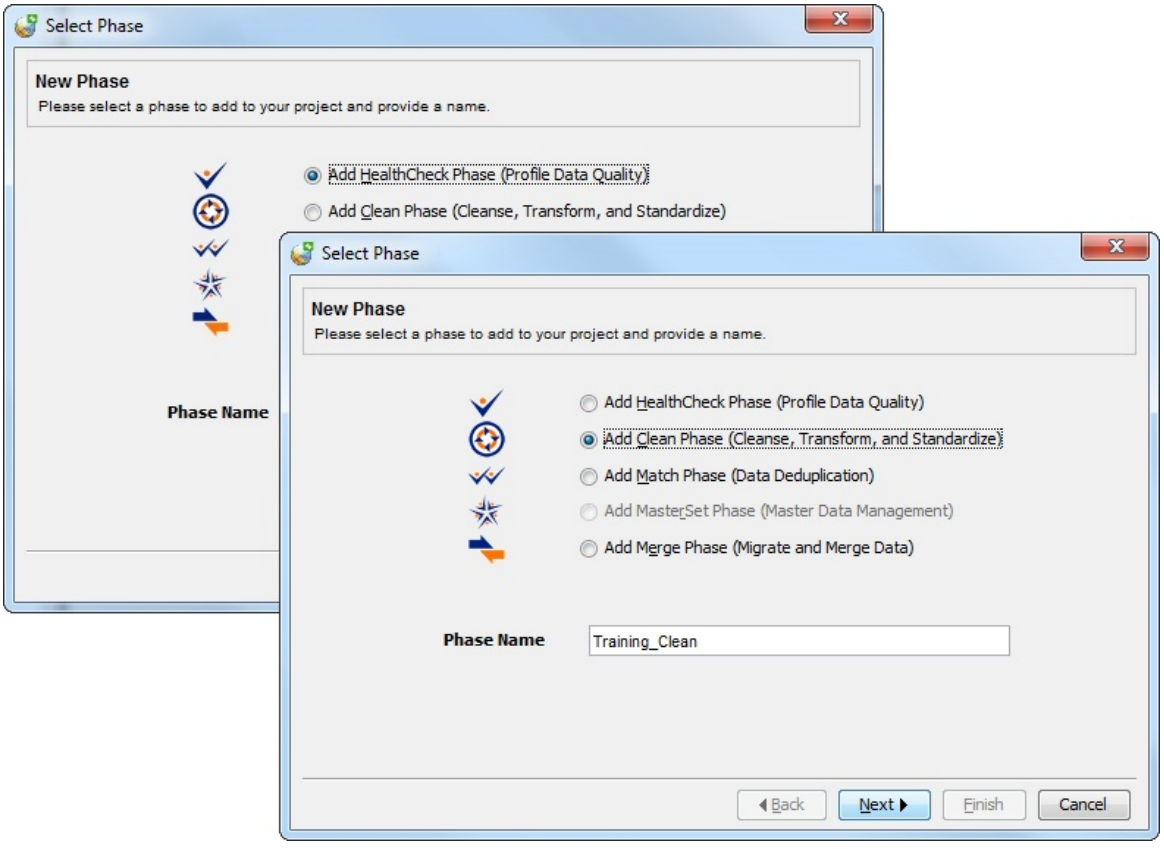
Resolving Data Issues

Issues found in the data during the HealthCheck Phase are cleaned by using the Clean Phase. The Clean Phase comprises of Rapid Clean and Detailed Clean Activities. By using this Data Quality feature, you can apply individual policies to transform your data while ensuring that the original database is updated with the correct policies.

Exercise: Adding a Clean Phase

To add a Clean Phase:

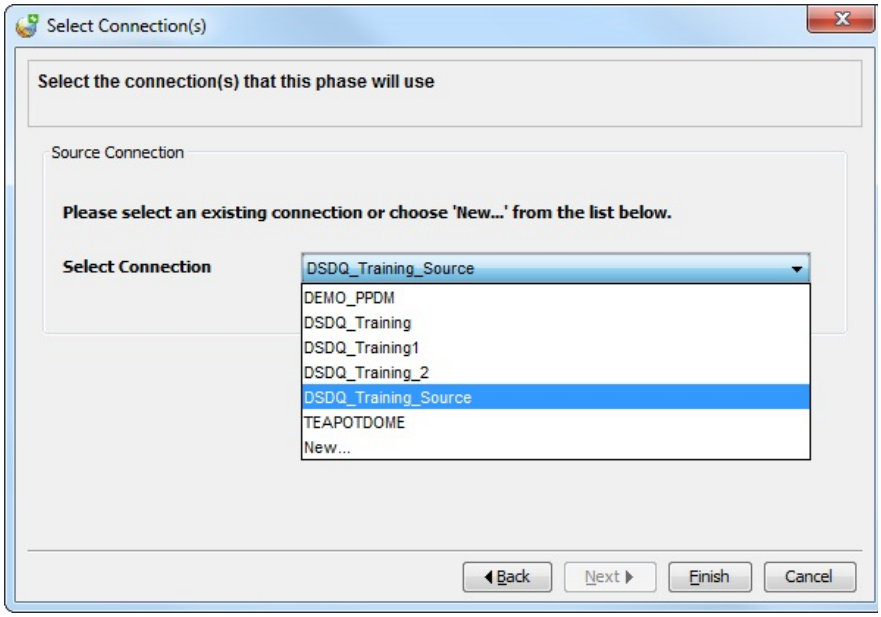
1. Click the **Add New Phase** button on the Project toolbar. The **Select Phase** window appears with the **Add HealthCheck Phase (Profile Data Quality)** option selected by default.



2. Select the **Add Clean Phase (Cleanse, Transform, and Standardize)** option.

3. Enter **Training_Clean** in the **Phase Name** field.

4. Click **Next** to continue.
   The **Select Connection(s)** window appears.



5. Select **DSDQ_Training_Source** from the **Select Connection** drop-down list

6. Click **Finish**.
   The **Clean** Phase is created and displays in the **DecisionSpace Data Quality Project Window**.

Rapid Clean Activity

The **Rapid Clean** Activity cleans out issues. Issues that can be cleaned out or corrected with this tool are:

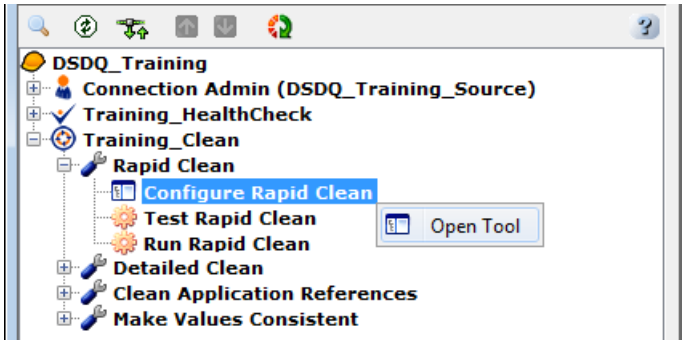
- Mixed Case
- Non Printable Characters
- Preceding White Space
- Trailing White Space
- Double While Space

> **Note**
>
> You must have already run a **Column Analysis** on Modeled Tables to run Rapid Clean. The Analysis information is required to display the issues found.

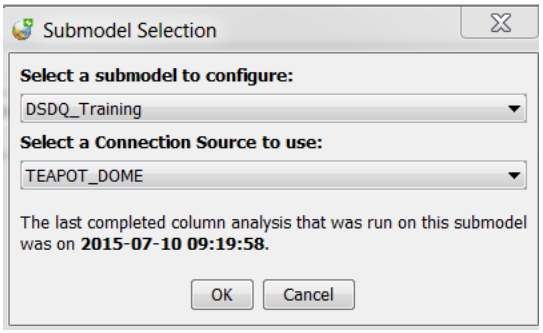
Exercise: Configuring the Rapid Clean Tool

To configure the Rapid Clean Tool:

1. Click ⊞ on the DecisionSpace Data Quality Tree to expand the **Training_Clean** Phase.
2. Click ⊞ to expand the **Rapid Clean** Activity.
3. Double-click the **Configure Rapid Clean** Tool or right-click the **Configure Rapid Clean** Tool and select **Open Tool** from the pop-up menu.



The **Submodel Selection** dialog box appears.



Note

The purpose of the date and time display is to let the user know that any clean changes made after this stated date and time will not be shown when the tool is opened. To see the updated changes after the date and time displayed here requires re-running the **Run Column Analysis on Modeled Tables** Task.

4. Select **DSDQ_Training** from the **Select a submodel to configure** drop-down list

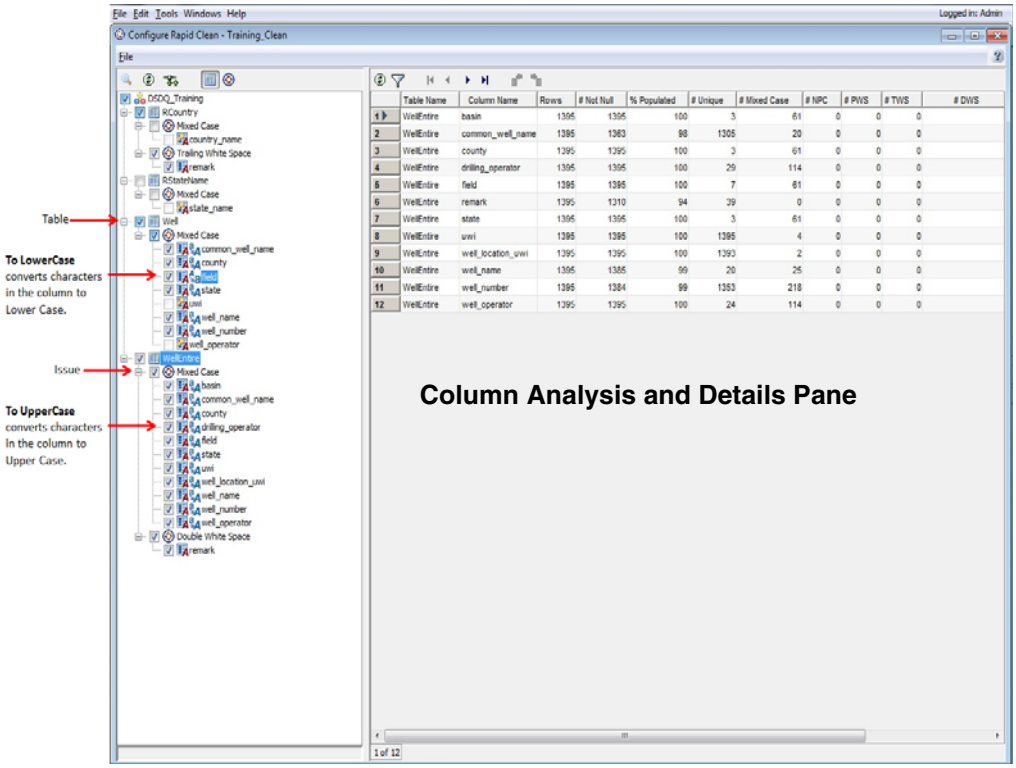
5. Click **OK**.

Note

The **Submodel Selection** dialog box only displays the submodels on which **Run Column Analysis on Modeled Tables** has been run.

The **Configure Rapid Clean** window appears. On the left side of the window is the Column Issues Tree, a tree that displays issues found during column analysis, and the columns and tables in which they occur. On the right side is the **Column Analysis Details** Pane, which displays column analysis information for the currently

selected node in the tree. You can expand a table or an issue by clicking the plus ⊞ sign next to it. When a table is expanded, all issues for that table are displayed. When an issue is expanded, all column(s) in the table with the issue are displayed.



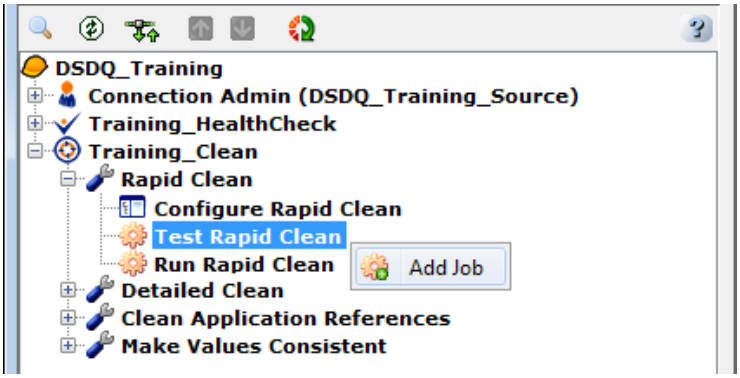
6. Select the check box(s) next to table(s) with issues.
   Issue(s) in the column for the selected table will be highlighted in the **Column Analysis Details** Pane. The selected issues will be corrected when the Rapid Clean task is run. For columns containing Mixed Case characters, an extra icon appears to the left of the column name. Clicking this icon will toggle between converting mixed case characters to uppercase or lowercase. The default is set to uppercase. Selecting will convert all characters in the column to uppercase and selecting will convert all characters in the column to lowercase when the Rapid Clean task is run.

7. Select **File > Exit** from the menu bar on the **Configure Rapid Clean** window to close it.

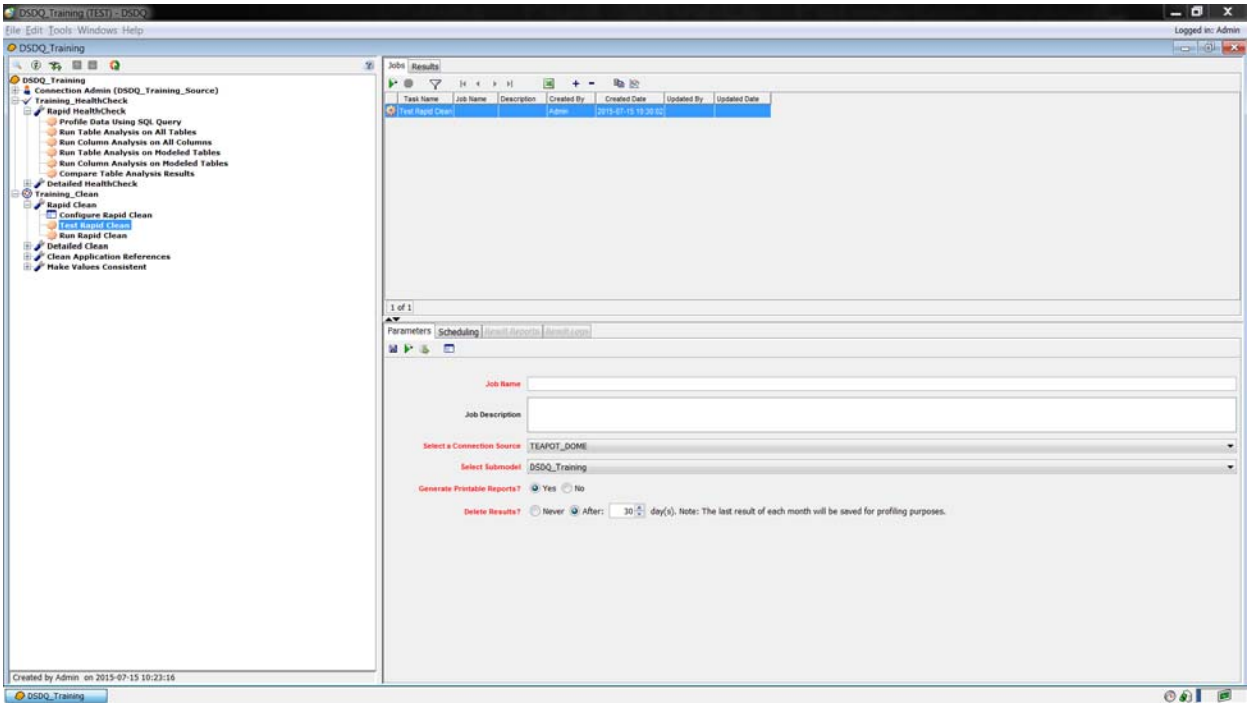
Exercise: Running the Test Rapid Clean Task

After cleaning the issues using the **Configure Rapid Clean** tool, the **Test Rapid Clean** task is run to make sure that the expected results are seen before running the **Run Rapid Clean** task to fix the entire dataset in the submodel. To run the **Test Rapid Clean** task:

1. Double-click the **Test Rapid Clean** task or right-click the **Test Rapid Clean** task, and select **Add Job** from the pop-up menu.



A new job is initiated and it displays on the **Job and Results Information** Pane on the right side of the **DecisionSpace Data Quality Project** window.



2. Enter **Job-01** in the **Job Name** field.

3. Enter **Rapid Clean Test** in the **Job Description** field.

4. Select **TEAPOT_DOME** from **Select a Connection Source**.

5. Select **DSDQ_Training** from the **Select Submodel** drop-down list.

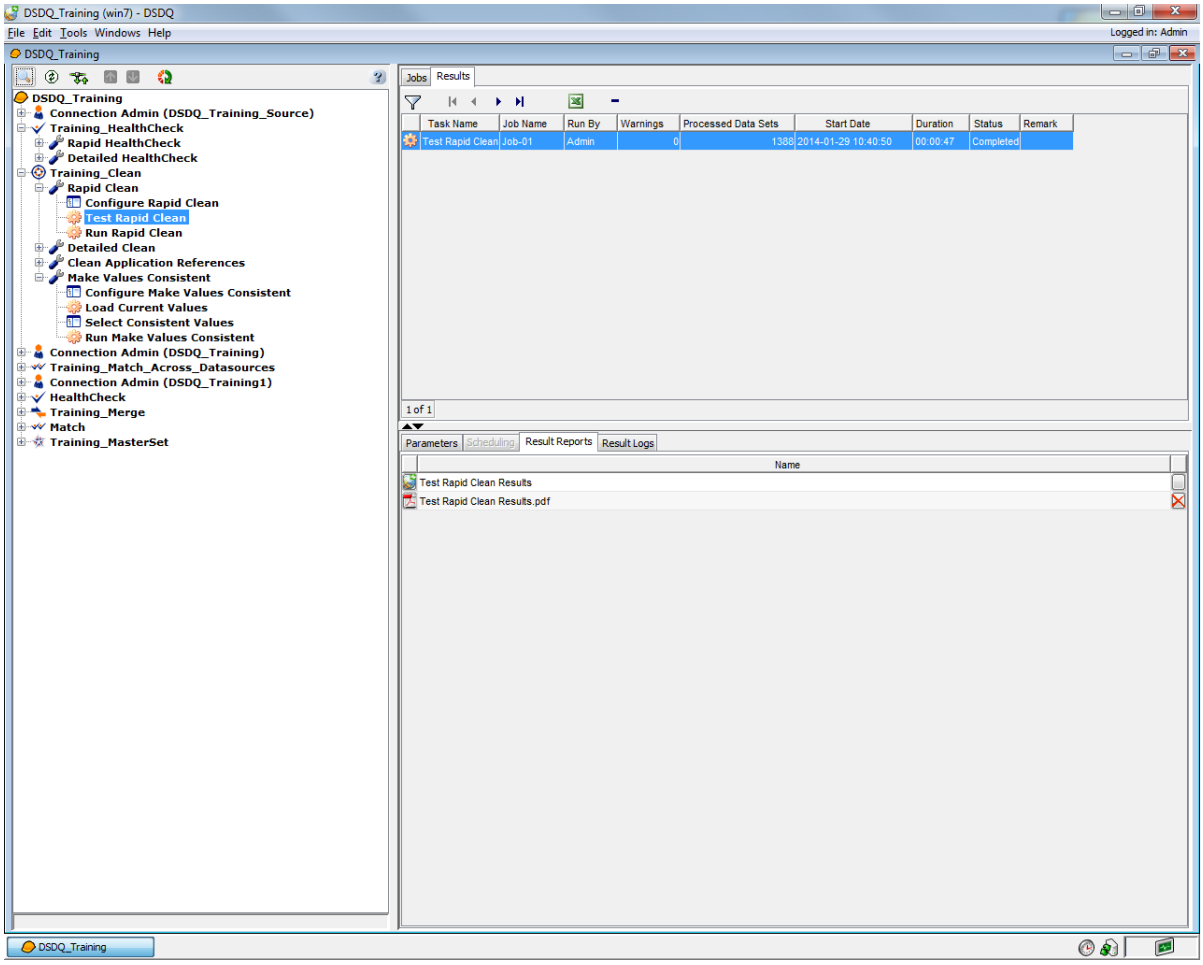
6. Select the **Yes** option for **Generate Printable Reports?**

7. Select the **After option for Delete Results?** Leave the number of days as **7**.

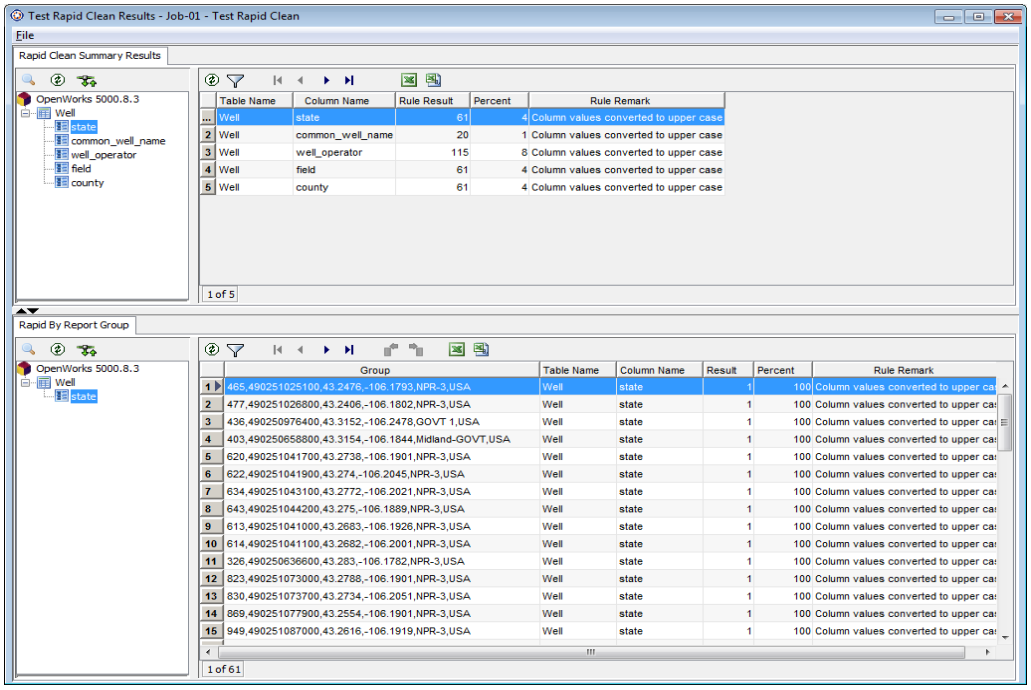
8. Click to save changes in the **Parameters** tab.

9. Click .
   The **Test Rapid Clean** task runs and displays results in the **Result Reports** tab of the **Job and Results Information** Pane.

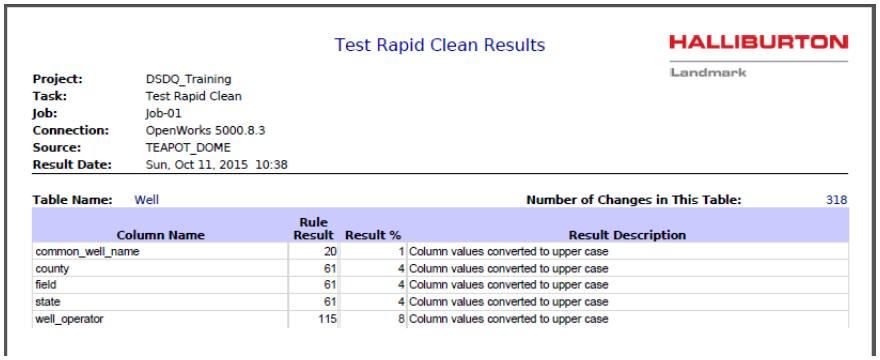
10. Select the **Results** tab on the **Job and Results Listing** Pane to view the values in the **Result Reports** tab on the **Job and Results Information Pane**.



11. Click on the **Result Reports** tab of the J**ob and Results Information** Pane to display the **Test Rapid Clean Results**.



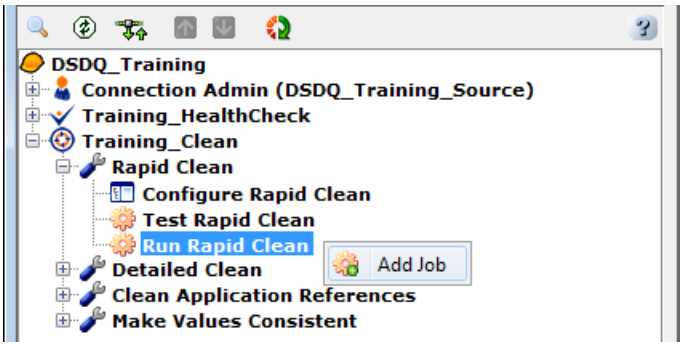
12. Click on the **Result Reports** tab on the J**ob and Results Information** Pane to display the **Test Rapid Clean Results** in PDF format.



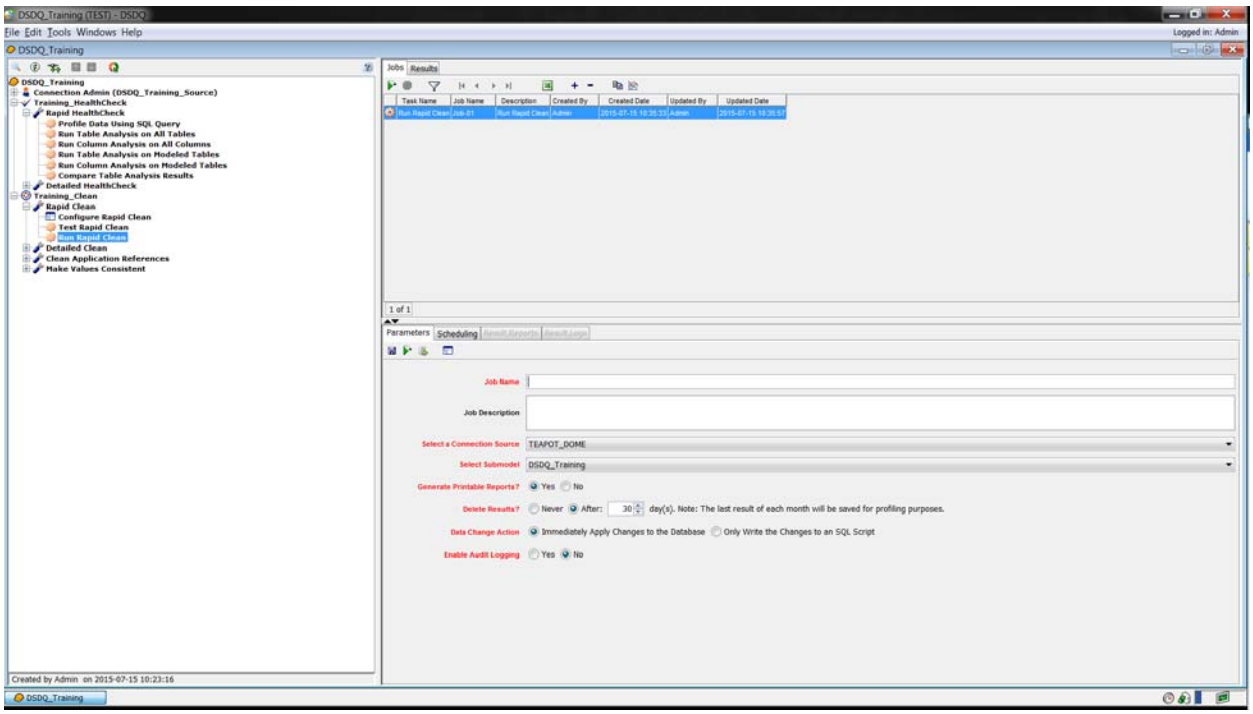
Exercise: Running the Rapid Clean Task

The **Run the Rapid Clean** task fixes the issues that were selected in the **Configure Rapid Clean** tool for the specific submodel.

1. Double-click the **Run Rapid Clean** task or right-click the **Run Rapid Clean** task and select **Add Job** from the pop-up menu.



A new job is initiated and it displays on the **Job and Results Information Pane**.



2. Enter **Job-01** in the **Job Name** field.

3. Enter **Rapid Clean** in the **Job Description** field.

4. Select **TEAPOT_DOME** from **Select a Connection Source**.

5. Select **DSDQ_Training** from the **Select Submodel** drop-down list.

6. Select the **Yes** option for **Generate Printable Reports?**

7. Select the **After** option for **Delete Results?** Leave the number of days as **7**.

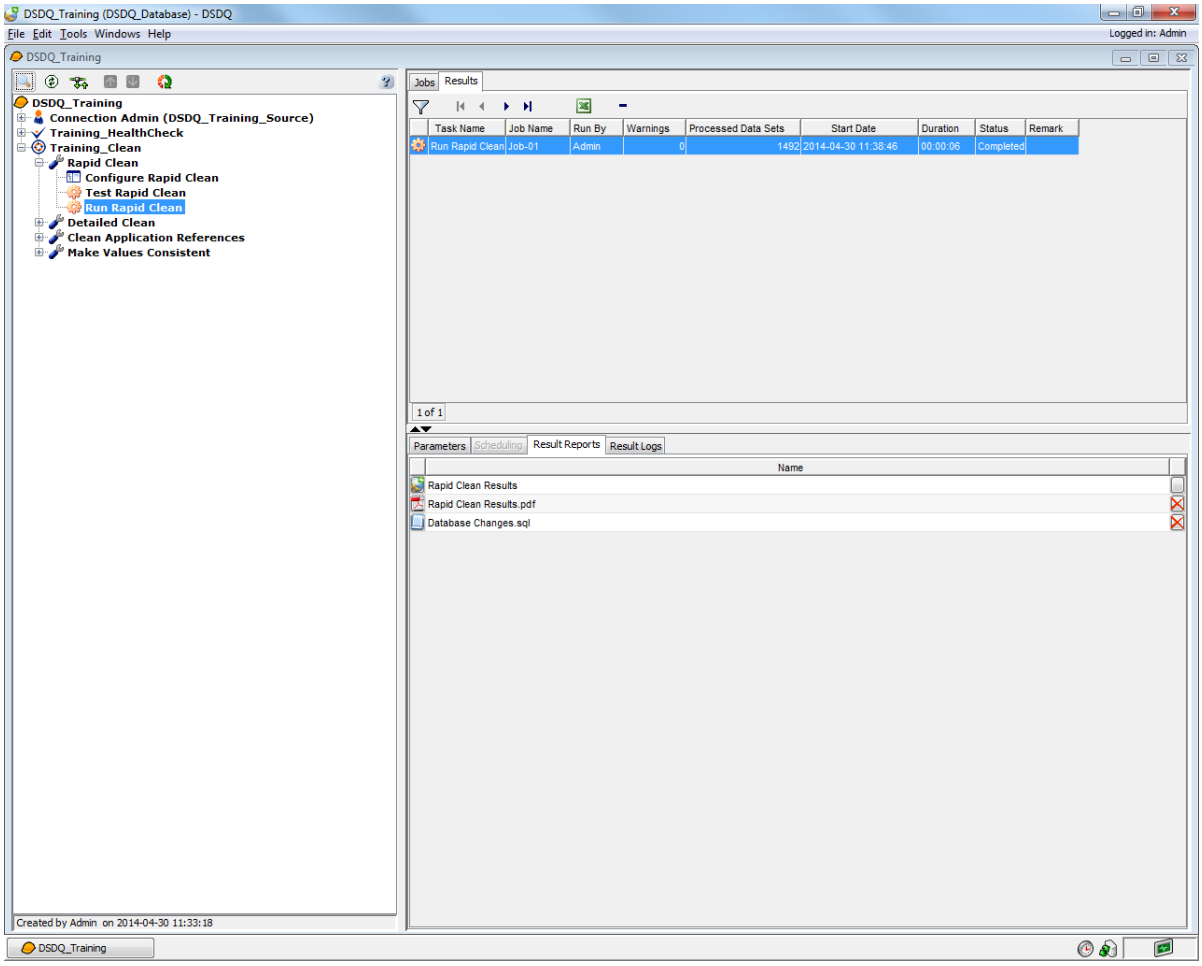
8. Select the **Immediately Apply Changes to the Database** option for **Data Change Action**.

9. Select the **No** option for **Enable Audit Logging**.

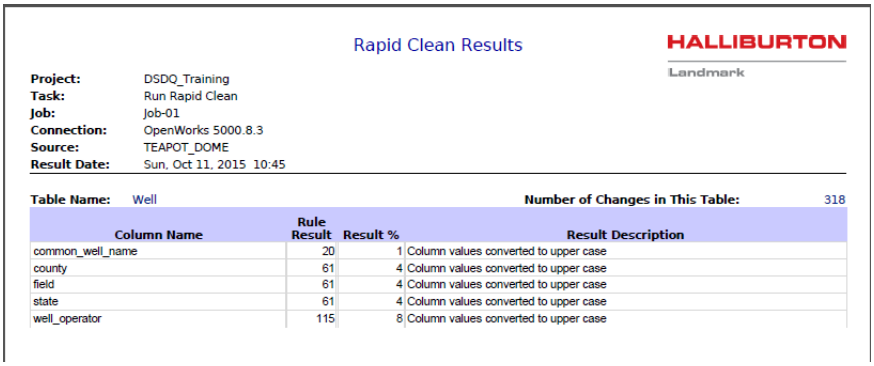
10. Click to save changes in the **Parameter** tab.

11. Click .
    The **Rapid Clean** task runs and displays results in the **Result Reports** tab of the **Job and Results Information** Pane.

12. Select the **Results** tab on the **Job and Results Listing** Pane to view the values in the **Result Reports** tab on the **Job and Results Information** Pane.



13. Click on the **Result Reports** tab on the **Job and Results Information** Pane to display **Rapid Clean Results** in PDF format.



Rapid Clean Results

HALLIBURTON
Landmark

Project: DSDQ_Training
Task: Run Rapid Clean
Job: Job-01
Connection: OpenWorks 5000.8.3
Source: TEAPOT_DOME
Result Date: Sun, Oct 11, 2015 10:45

Table Name: Well **Number of Changes in This Table:** 318

Column Name	Rule Result	Result %	Result Description
common_well_name	20	1	Column values converted to upper case
county	61	4	Column values converted to upper case
field	61	4	Column values converted to upper case
state	61	4	Column values converted to upper case
well_operator	115	8	Column values converted to upper case

Detailed Clean Activity

The **Detailed Clean** Activity gives you the ability to assign columns from selected submodels to clean requirements, test a service level with selected test data and view test results in the **Configure Detailed Clean** tool.

Exercise: Configuring the Detailed Clean Tool

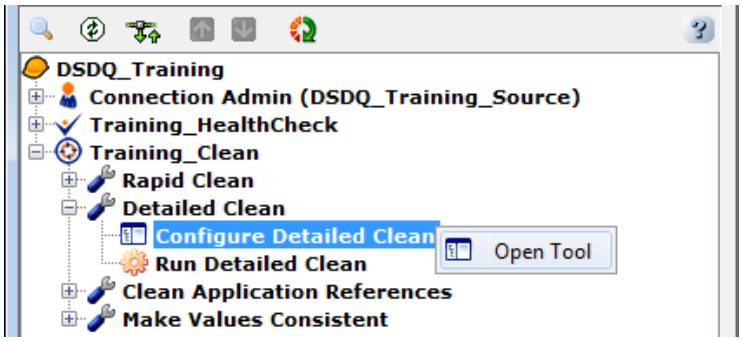
The **Configure Detailed Clean** tool is used to configure service levels for testing prior to running the **Run Detailed Clean** task. During this process, you can select which requirements in the service level to enable/disable and when testing a service level, what subset of the total data to use. A service level containing clean requirements must exist prior to opening the **Configure Detailed Clean** tool. This process requires you to execute the following steps:

- Select a submodel
- Select a service level
- Assign elements to columns
- Add a new service level requirement
- Validate the service level requirement
- Test a service level

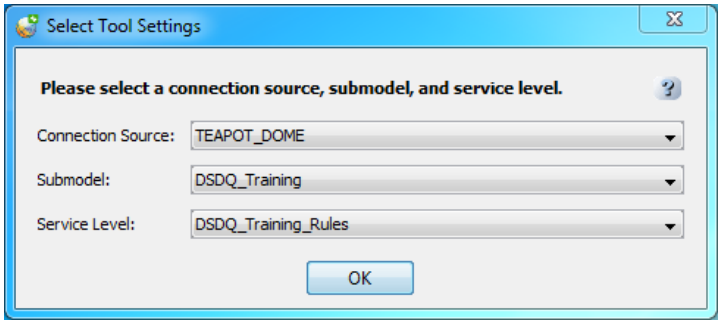
Information about each step is provided in the following section:

1. Click ⊞ on the DecisionSpace Data Quality Tree to expand the **Detailed Clean** Activity.

2. Double-click the **Configure Detailed Clean** tool or right-click the **Configure Detailed Clean** tool, and select **Open Tool** from the pop-up menu.



The **Select Tool Settings** dialog box displays:

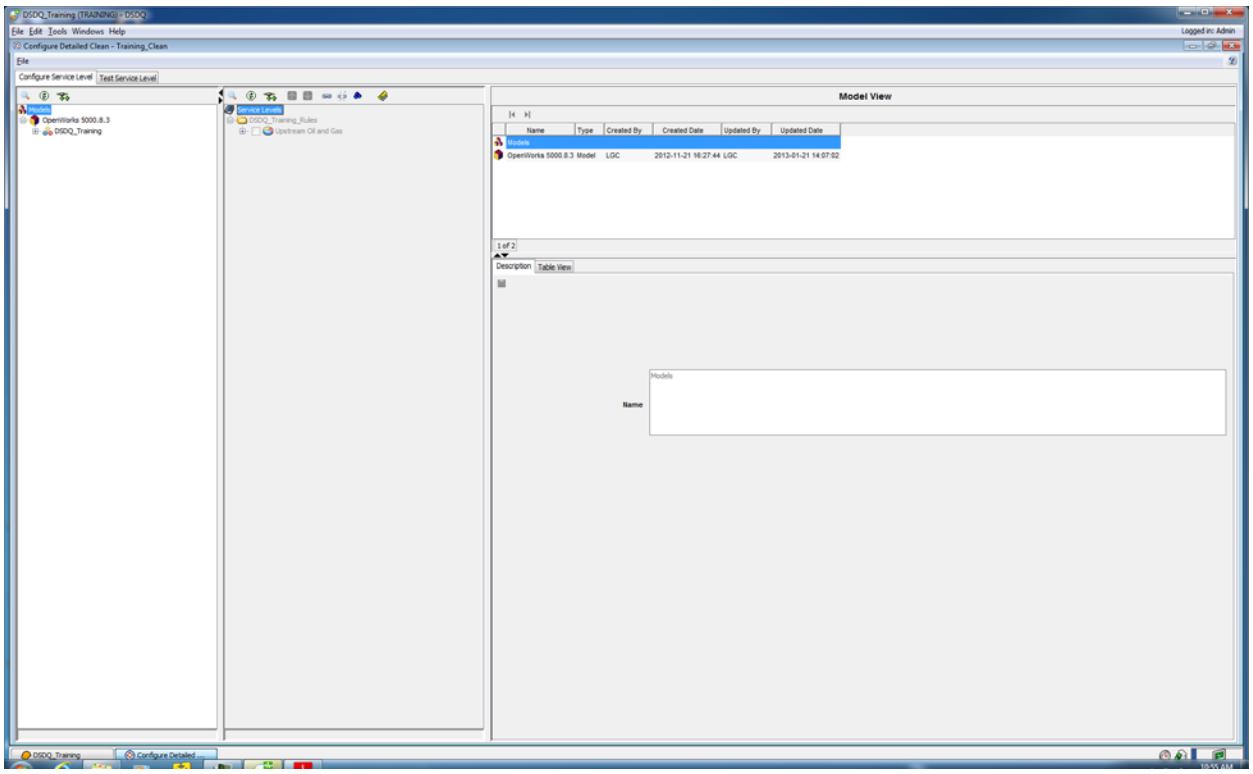


3. Select **TEAPOT_DOME** from the **Connection Source** drop-down list.

4. Select **DSDQ_Training** from the **Submodel** drop-down list.

5. Select **DSDQ_Training_Rules** from the **Service Level** drop-down list.

6. Click **OK**.
   The **Configure Detailed Clean** window appears with the selected service level displaying in the Service Level Tree.



Note

The selected service level is remembered when the **Configure Detailed Clean** tool is closed and is automatically displayed when the tool is re-opened. Only one service level can be configured at any given time.

7. Click ⊞ on the Data Model Tree to expand the **DSDQ_Training** submodel.

8. Expand the **Well** Table.

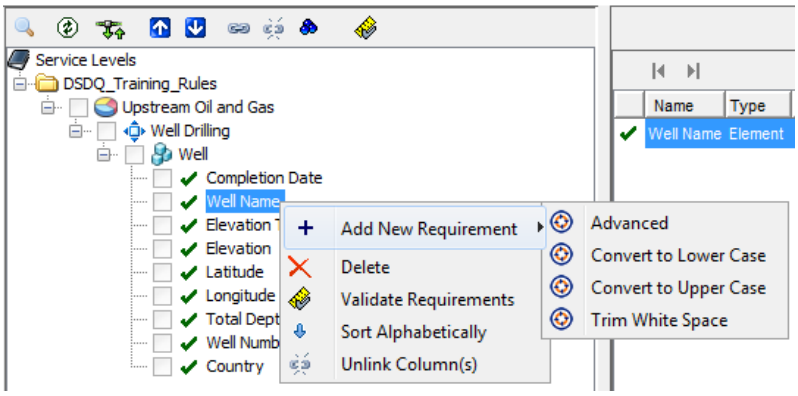
9. Click ⊞ on the Service Level Tree to expand the **DSDQ_Training_Rules** service level.

10. Expand the **Upstream Oil & Gas** sector.

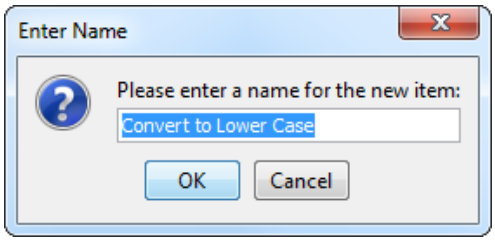
11. Expand the **Well Drilling** area.

12. Expand the **Well** element group.

13. Right-click the **Well Name** element in the Service Level Tree and select **Add New Requirement > Convert to Lower Case** from the pop-up menu.

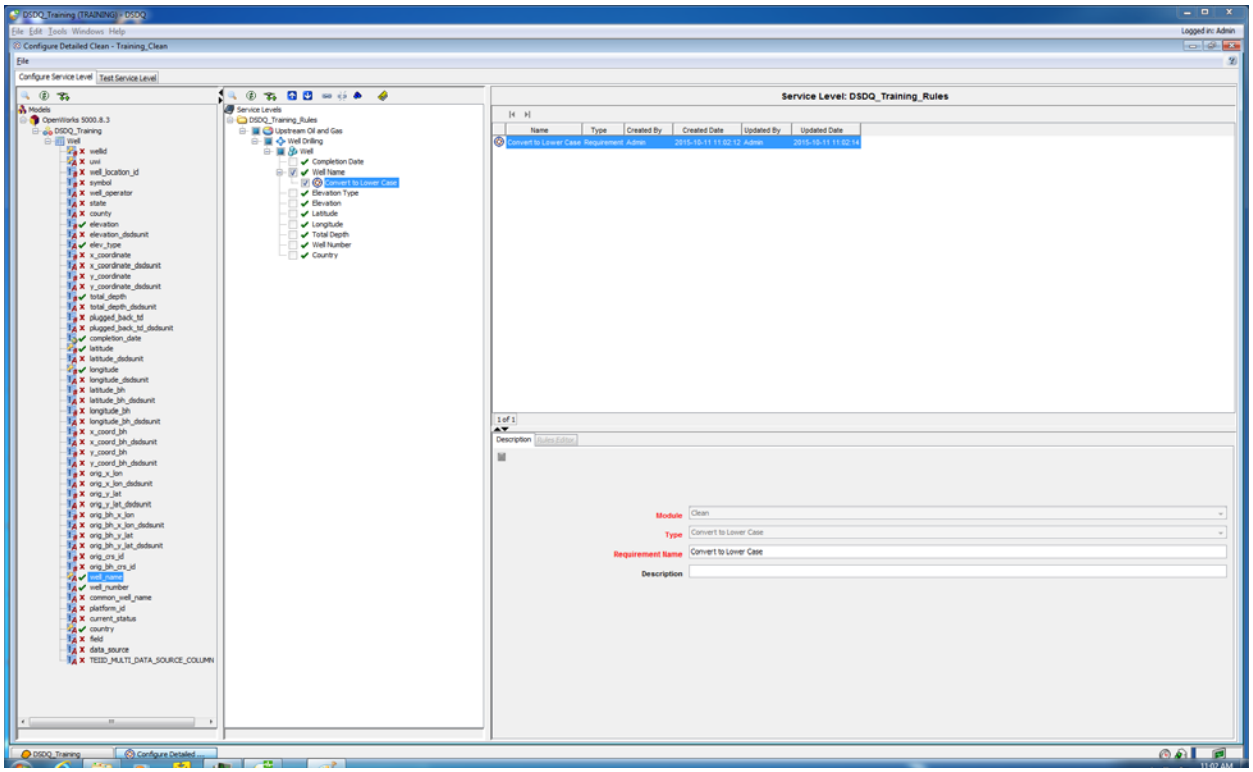


The **Enter Name** dialog box appears.

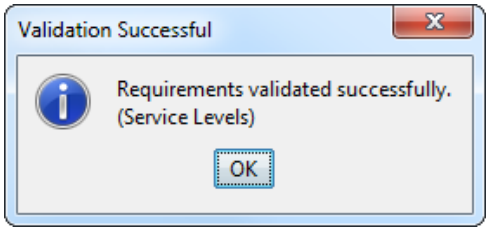


14. Optionally, specify a user-defined name for the requirement.

15. Click **OK** to add the requirement to the selected element.



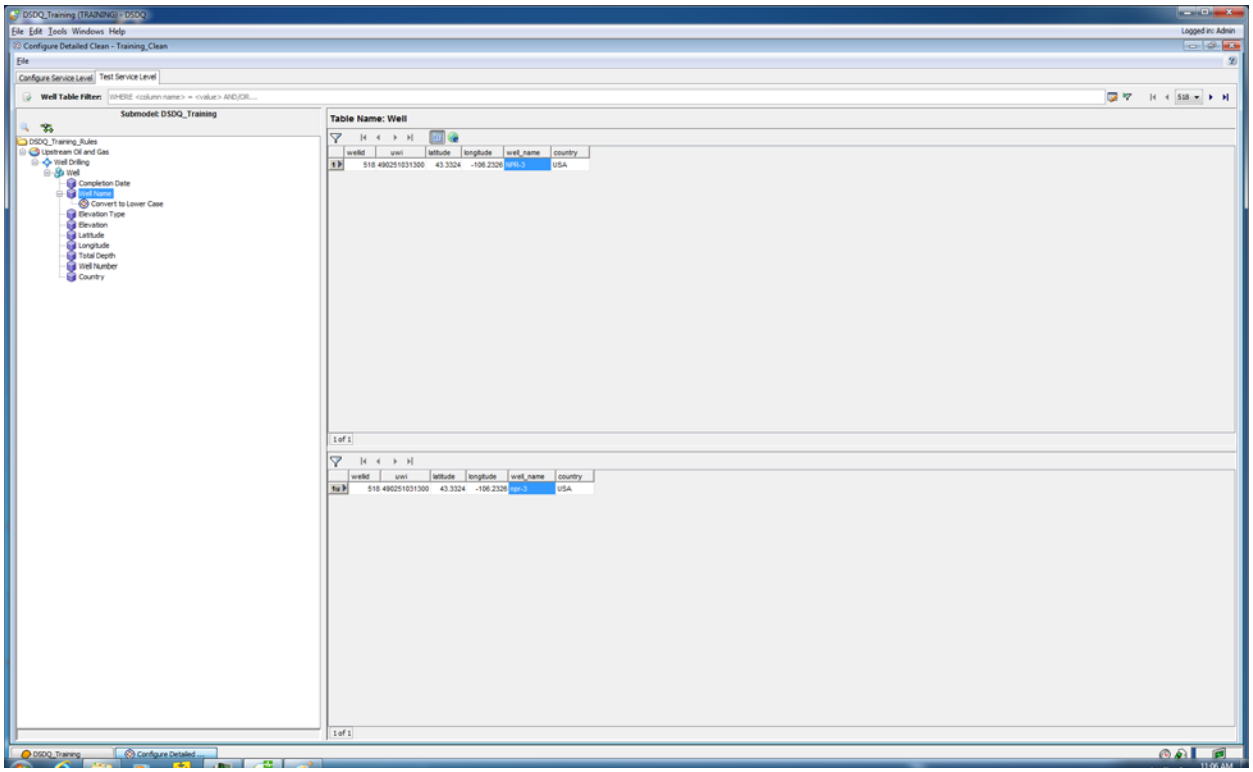
16. Select the **Convert to Lower Case** requirement and click on the Service Level Tree toolbar to validate the requirement.
A **Validation Successful** message is displayed once the selected requirement has been validated.



Note

You will need to remove **Well Name** as a Primary Key in Perform Table Modeling in order to perform this task.

17. You can test a service level by selecting the **Test Service Level** tab. The service level test is automatically executed. Results are displayed on the top and bottom panes on the right side of the window.

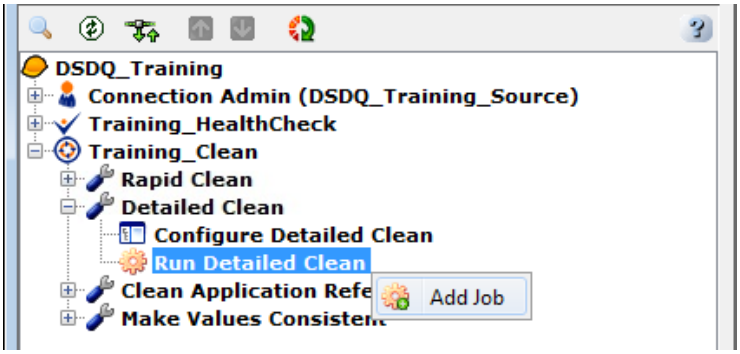


18. Select **File > Exit** from the menu bar on the **Configure Detailed Clean** window.

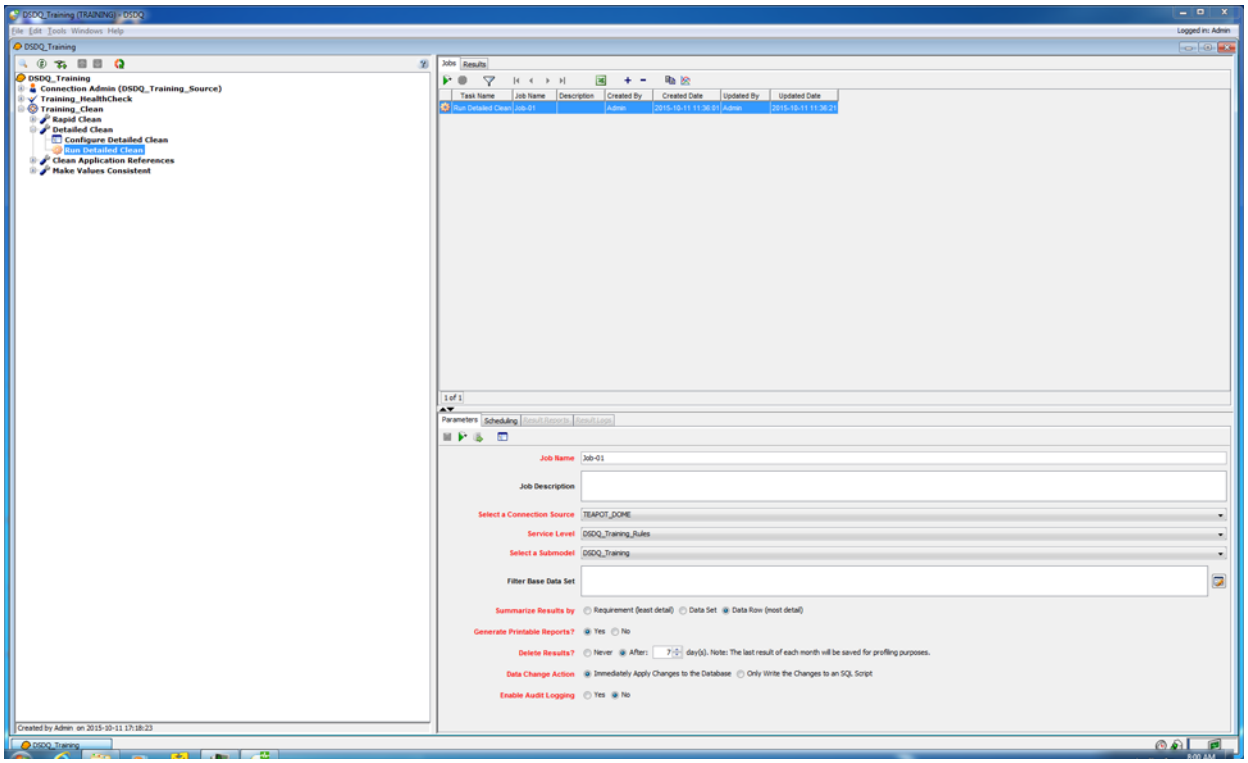
Exercise: Running the Detailed Clean Task

To run the Detailed Clean Task:

1. Double-click the **Run Detailed Clean** task or right-click the **Run Detailed Clean** task and select **Add Job** from the pop-up menu.



A new job is initiated and displays on the **Jobs and Results Listing Pane**.



2. Enter **Job-01** in the **Job Name** field.

3. Enter **Detailed Clean** in the **Job Description** field.

4. Select **TEAPOT_DOME** from **Select a Connection Source**.

5. Select **DSDQ_Training_Rules** from the **Service Level** drop-down list.

6. Select **DSDQ_Training** from the **Select Submodel** drop-down list.

7. Optionally, set a filter on the data subset.

8. Select the **Data Row (most detail)** option for **Summarize Results by**.

9. Select the **Yes** option for **Generate Printable Reports?**

10. Select the **After option for Delete Results?** Leave the number of days as **7**.

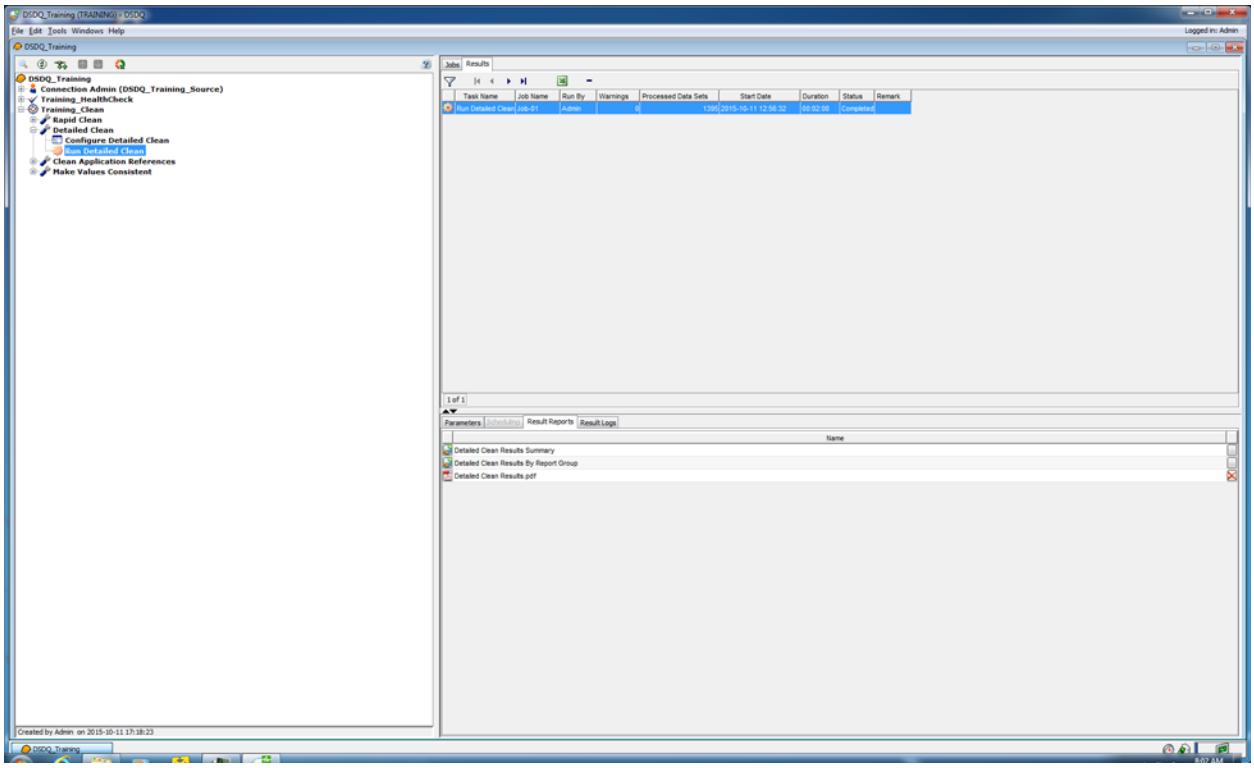
11. Select the **Immediately Apply Changes to the Database** option for **Data Change Action**.

12. Select the **No** option for **Enable Audit Logging**.

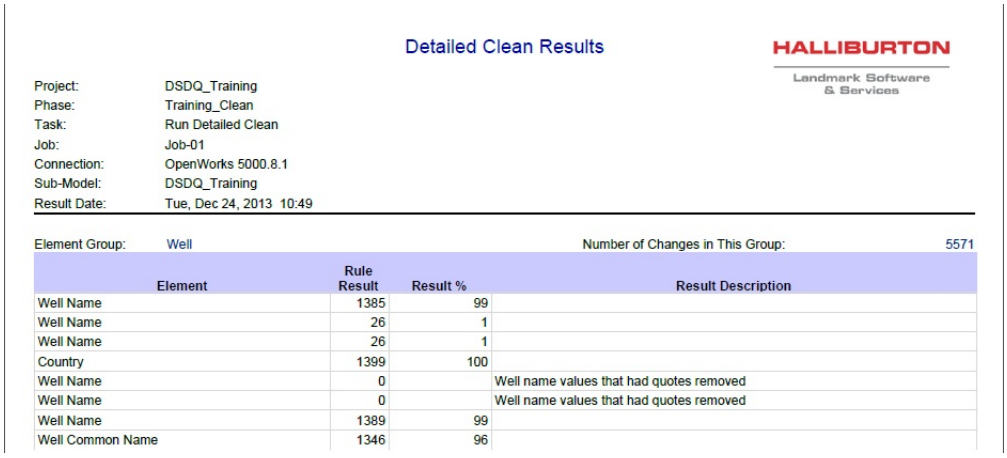
13. Click to save changes in the **Parameters** tab.

14. Click .
    The **Detailed Clean** task runs and displays results in the **Result Reports** tab on the **Jobs and Results Information** Pane.

15. Select the **Results** tab on the **Jobs and Results Listing** Pane to view the values in the **Results Report** tab on the **Job and Results Information** Pane.



16. Click on the **Results Reports** tab to display **Detailed Clean Results** in PDF format.



Detailed Clean Results

HALLIBURTON
Landmark Software & Services

Project:	DSDQ_Training
Phase:	Training_Clean
Task:	Run Detailed Clean
Job:	Job-01
Connection:	OpenWorks 5000.8.1
Sub-Model:	DSDQ_Training
Result Date:	Tue, Dec 24, 2013 10:49

Element Group: Well — Number of Changes in This Group: 5571

Element	Rule Result	Result %	Result Description
Well Name	1385	99	
Well Name	26	1	
Well Name	26	1	
Country	1399	100	
Well Name	0		Well name values that had quotes removed
Well Name	0		Well name values that had quotes removed
Well Name	1389	99	
Well Common Name	1346	96	

Cleaning Application References

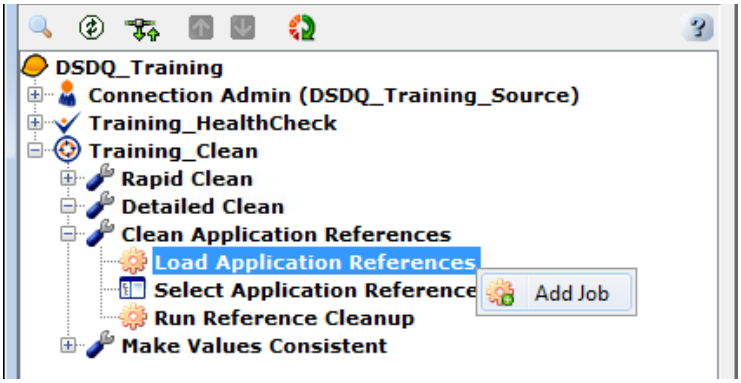
The Clean Application References Activity enables you to clean up data references to the application reference tables. In order to do so, you will load the current values in the application reference tables followed by configuring values that are correct for your data and application reference tables.

Exercise: Loading Application References

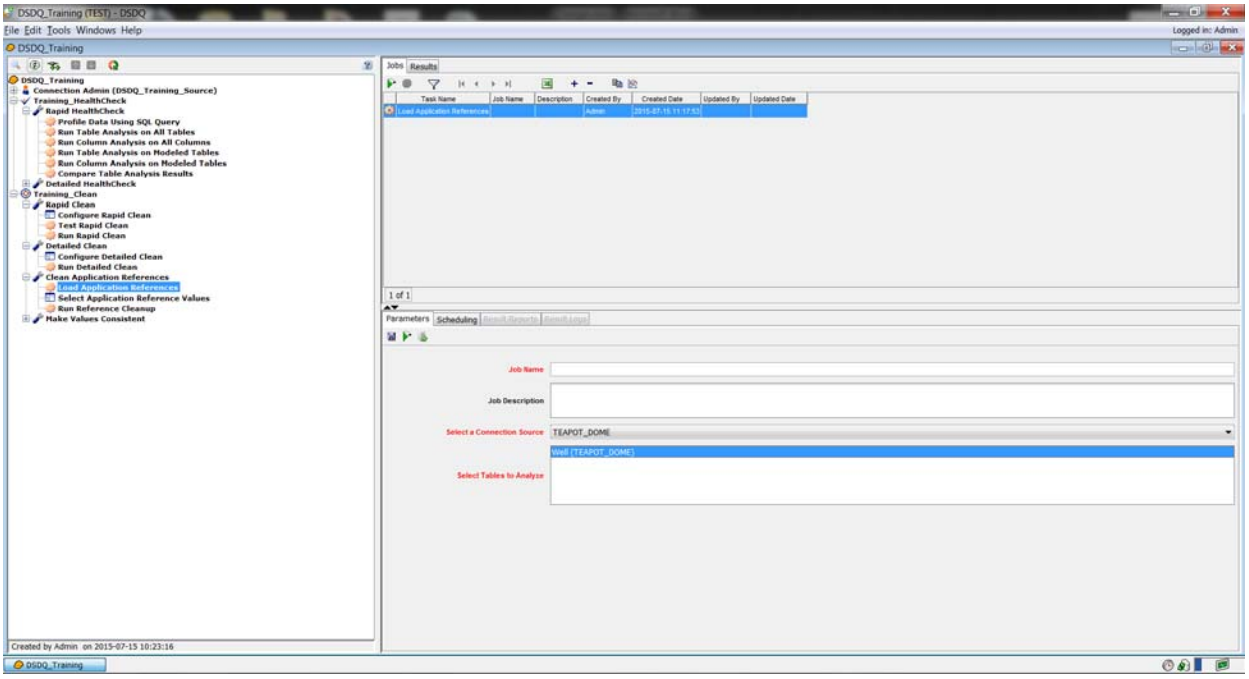
The **Load Application Reference** task loads the current values in the application reference tables that have been configured. To Load Application References:

1. Click ⊞ on the DecisionSpace Data Quality Tree to expand the **Clean Application References** Activity.

2. Double-click the **Load Application References** task or right-click the **Load Application References** task, and select **Add Job** from the pop-up menu.



A new job is initiated and displays on the **Jobs and Results Listing Pane**.



3. Enter **Job-01** in the **Job Name** field.

4. Enter **Application References** in the **Job Description** field.

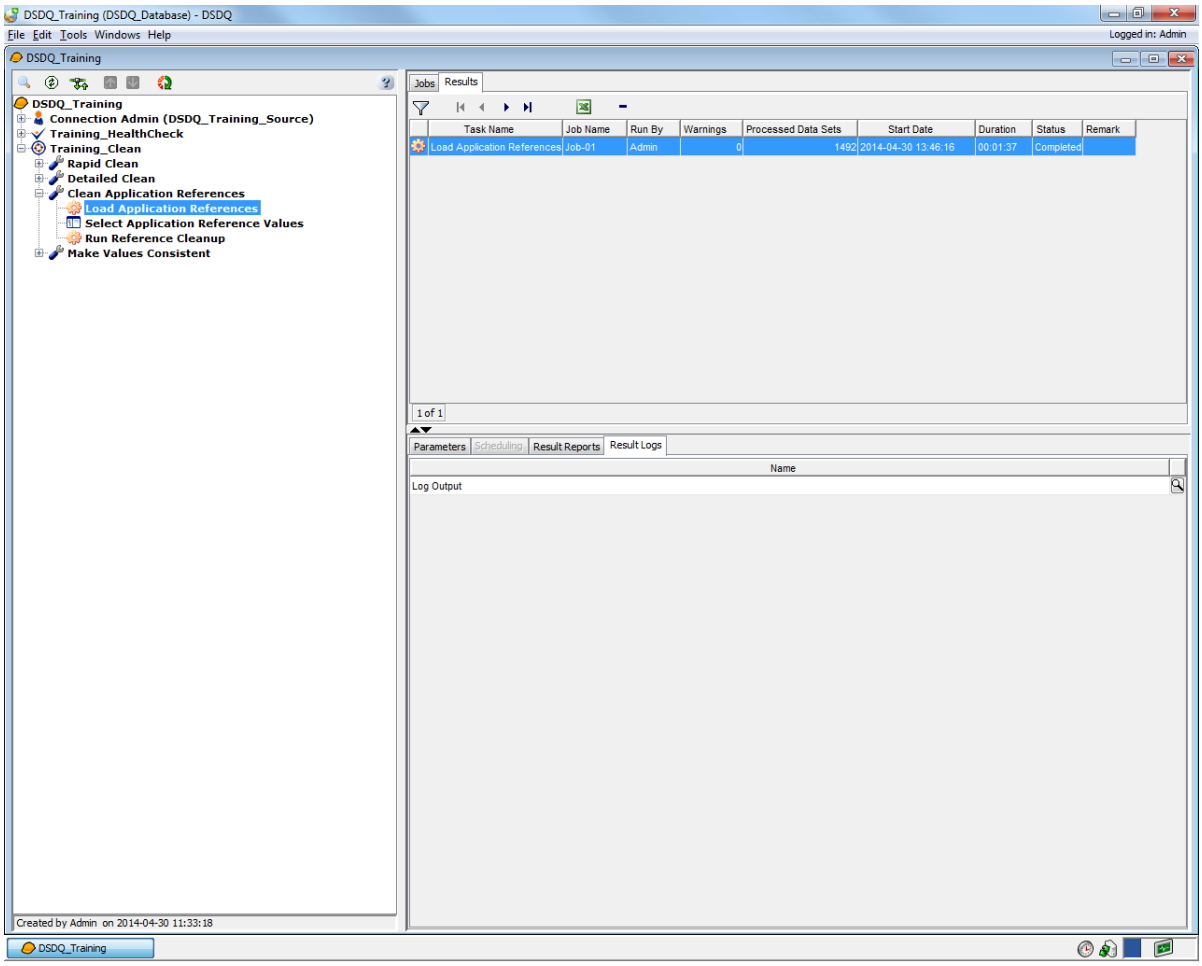
5. Select **TEAPOT_DOME** from **Select a Connection Source**.

6. Select **RCountry** from the **Select Tables to analyze** list box.

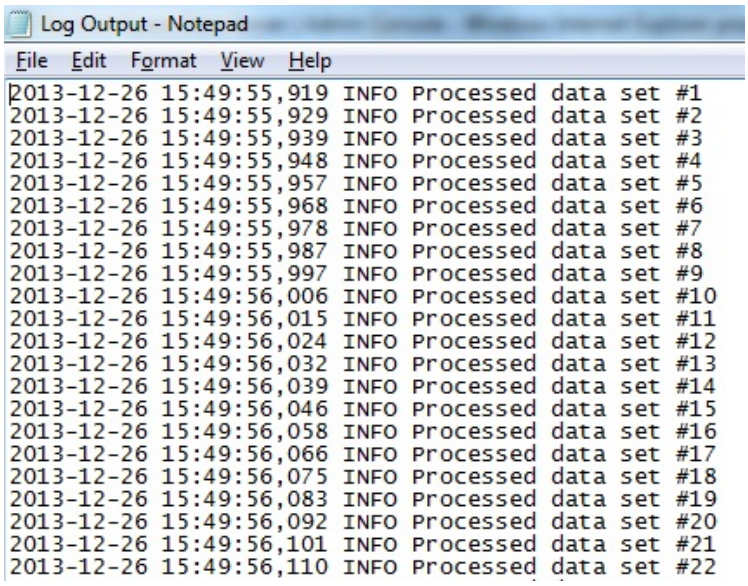
7. Click to save changes in the **Parameter** tab.

8. Click .
   The **Load Application References** task is executed and displays in the **Result Reports** tab on the **Jobs and Results Information** Pane.

9. Select the **Results** tab on the **Job and Results Listing** Pane and then the **Result Logs** tab on the **Jobs and Results Information Pane**.



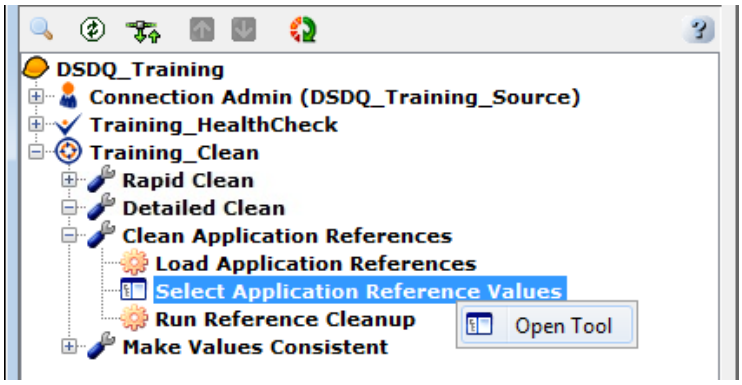
10. Double-click **Log Output** on the **Result Logs** tab to view the loaded references.



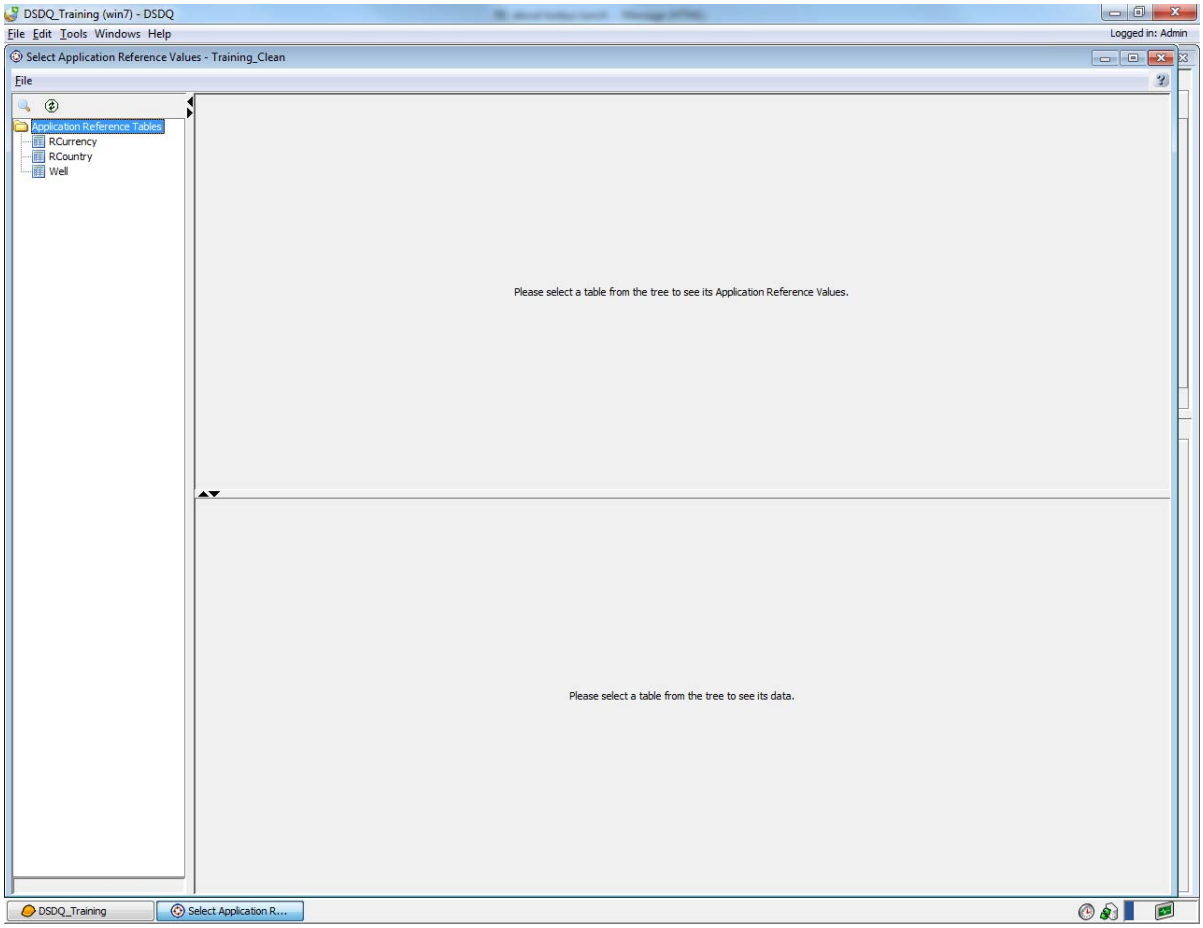
Exercise: Selecting Application Reference Values

The **Select Application Reference Values** tool allows you to select and configure the values that are correct for your data and application reference tables. To select application reference values:

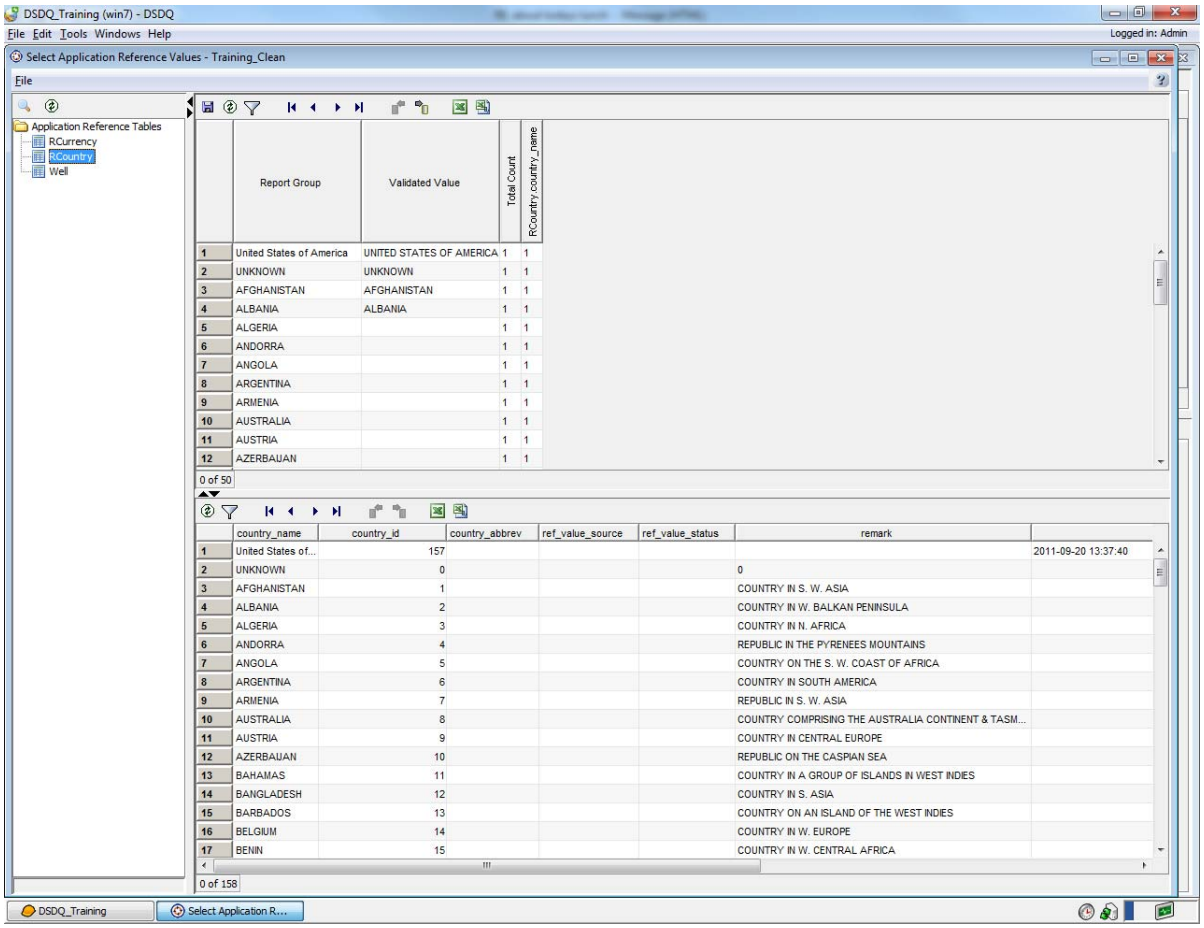
1. Double-click the **Select Application Reference Values** tool or right-click the **Select Application Reference Values** tool and select **Open Tool** from the pop-up menu.



The **Select Application Reference Values** window appears.



2. Select the **RCountry** table from the **Application Reference Table** to view the results of the table in the top right table.



> **Note**
>
> Table and column headers specify the application reference that was configured and the Total Count of matching values found. Selecting a result in the top right table displays the application reference in the bottom table. The bottom table displays the application reference values for the selected table.

3. To change a value, enter the correct value for application references in the **Validated Value** column. Only those values that exist in the application reference table can be entered here.

4. Click to save changes made to the reference table.

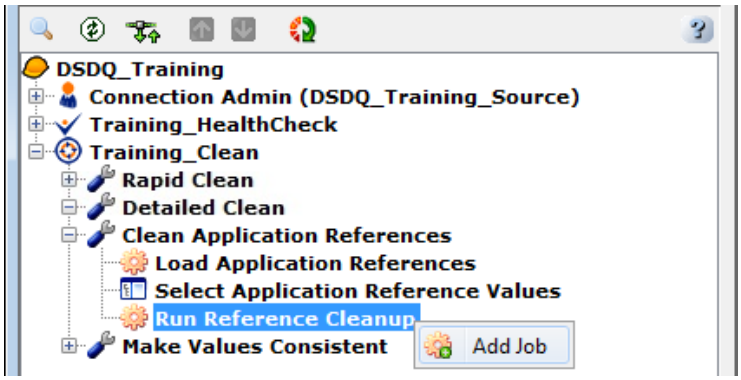
5. Select **File > Exit** from the menu bar on the **Select Application Reference Values** window.

> **Note**
>
> The application reference tables have to be configured with at least one **Primary Key**. To view data from a specific column, configure it as a **Report Column** in the **Perform Table Modeling** tool.

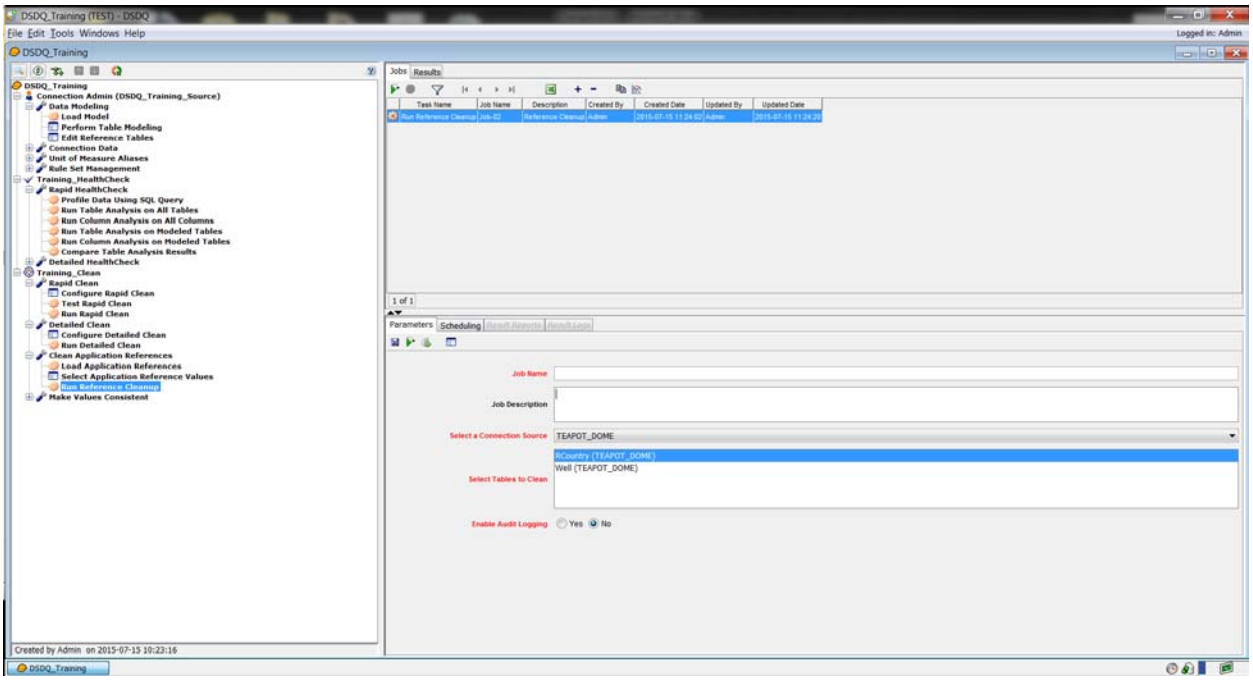
Exercise: Running Reference Cleanup

The **Run Reference Cleanup** task modifies the values as per the configuration in the **Select Application Reference Values** tool.

1. Double-click the **Run Reference Cleanup** task or right-click the **Run Reference Cleanup** task, and select **Add Job** from the pop-up menu.



A new job is initiated and displays on the **Jobs and Results Listing Pane**.

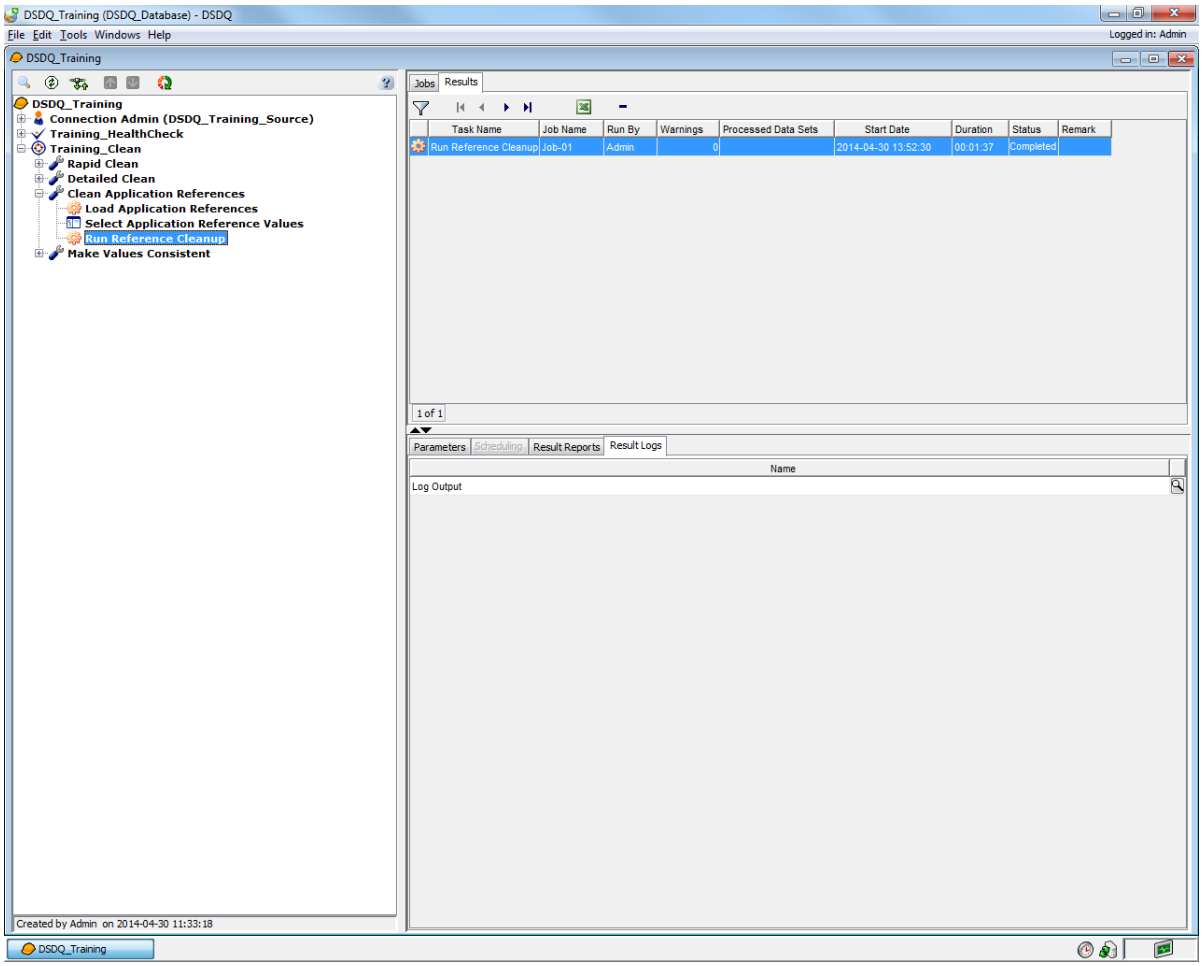


Note

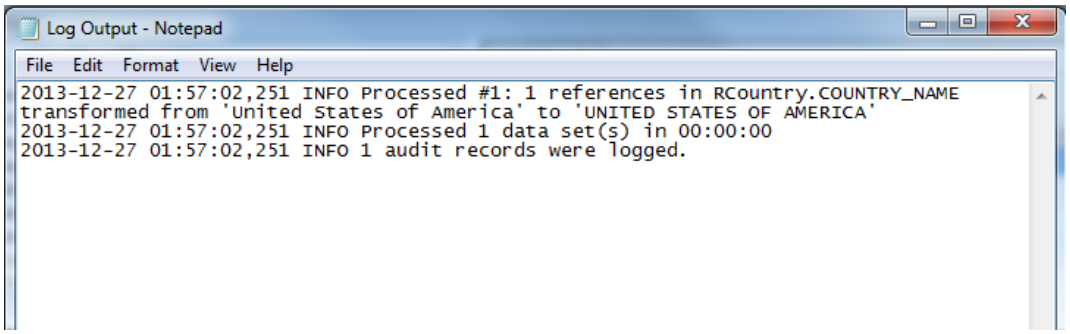
The job creation is initiated only the first time the task is run. To add an additional job, right-click the **Run Reference Cleanup** task and select the **Add Job** option from the pop-up menu, or click the **Add New Job** button on the Jobs toolbar.

2. Enter **Job-01** in the **Job Name** field.
3. Enter **Reference Cleanup** in the **Job Description** field.
4. Select **TEAPOT_DOME** from **Select a Connection Source**.
5. Select **RCountry** in the **Select Tables to Clean** field.
6. Select the **Yes** option for **Enable Audit Logging**.
7. Click to save changes in the **Parameter** tab.
8. Click .
   The **Run Reference Cleanup** task is executed and displays results in the **Result Reports** tab on the **Jobs and Results Information Pane**.

9. Select the **Results** tab on the **Job and Results Listing** Pane and then the **Result Logs** tab on the **Jobs and Results Information Pane**.



10. Double-click **Log Output** on the **Result Logs** tab to view results of the Run Reference Cleanup task.



Making Values Consistent

The Make Values Consistent tool is used to assign elements that have Make Values Consistent requirements to associated columns. In order to do so, you will load current values for this requirement, select consistent values (manually, or by selecting these from reference tables), and finally update the data with the new values.

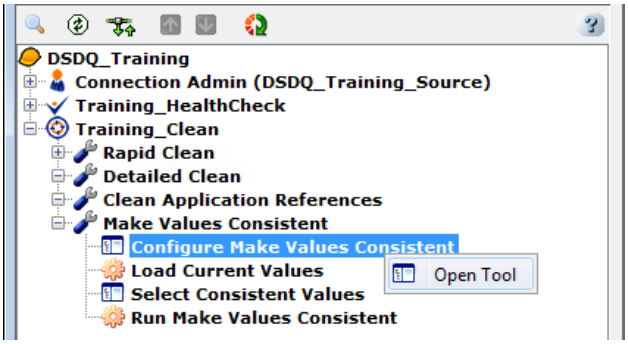
Exercise: Configuring Make Values Consistent

The **Configure Make Values Consistent** tool is used to configure service levels with make values consistent requirements for testing prior to running the actual **Run Make Values Consistent** task. The user can select which requirements in the service level to enable/disable when testing. A service level containing make values consistent requirements must exist prior to opening the **Configure Make Values Consistent** tool. This process requires that you execute the following steps:

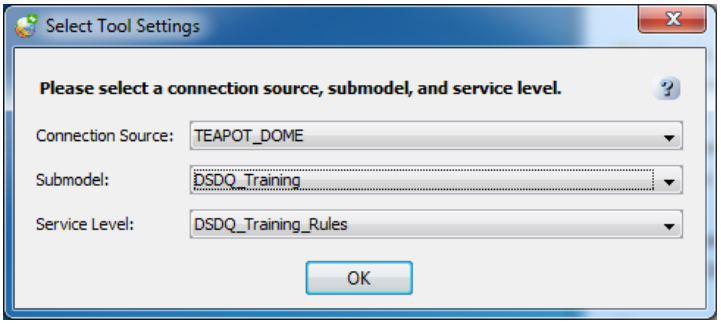
- Select a submodel
- Select a service level
- Assign elements to columns
- Add a new service level requirement

1. Click ⊞ on the DecisionSpace Data Quality Tree to expand the **Make Values Consistent** Activity.

2. Double-click the **Configure Make Values Consistent** task or right-click the **Configure Make Values Consistent** task, and select **Open Tool** from the pop-up menu.



The **Select Tool Settings** dialog box appears:

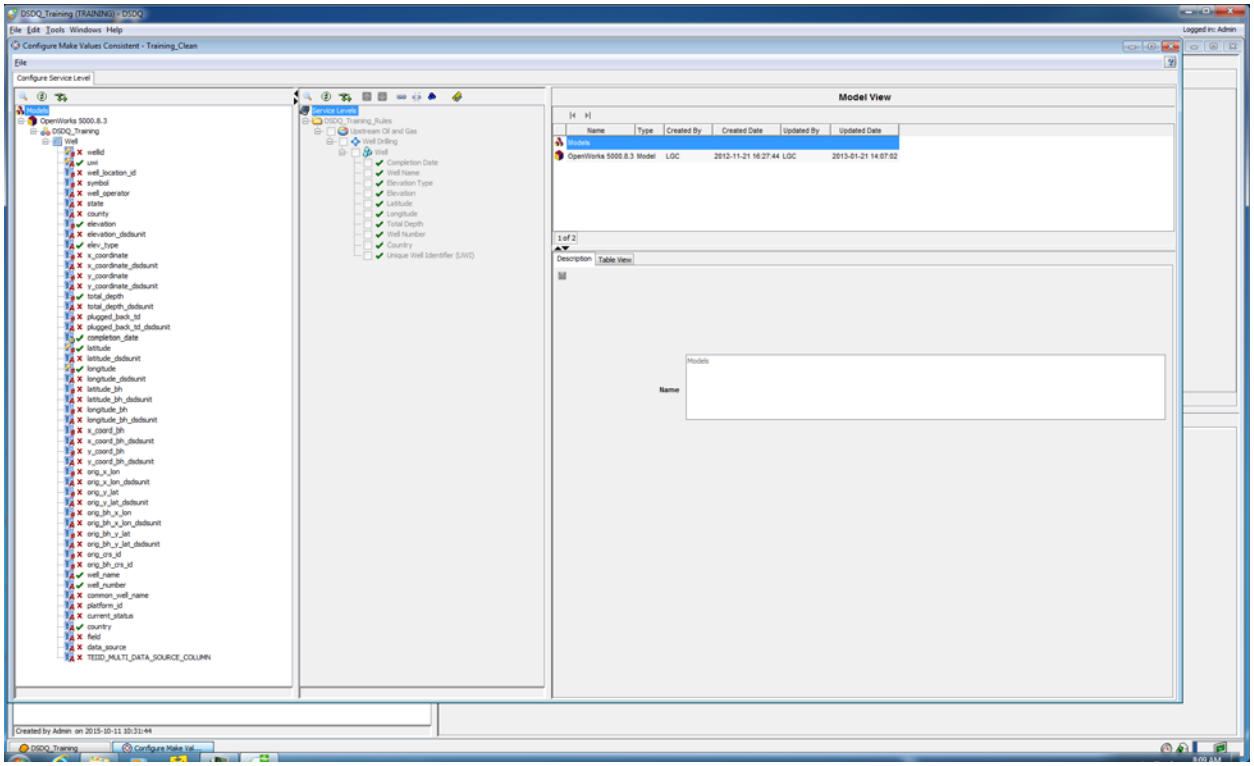


3. Select **TEAPOT_DOME** from the **Connection Source** drop-down list.

4. Select **DSDQ_Training** from the **Submodel** drop-down list.

5. Select **DSDQ_Training_Rules** from the **Service Level** drop-down list.

6. Click **OK**.
   The **Configure Make Values Consistent** window appears with the selected service level displaying in the Service Level Tree.



7. Click ⊞ on the Data Model Tree to expand the **DSDQ_Training** Submodel.

8. Expand the **RCountry** table.

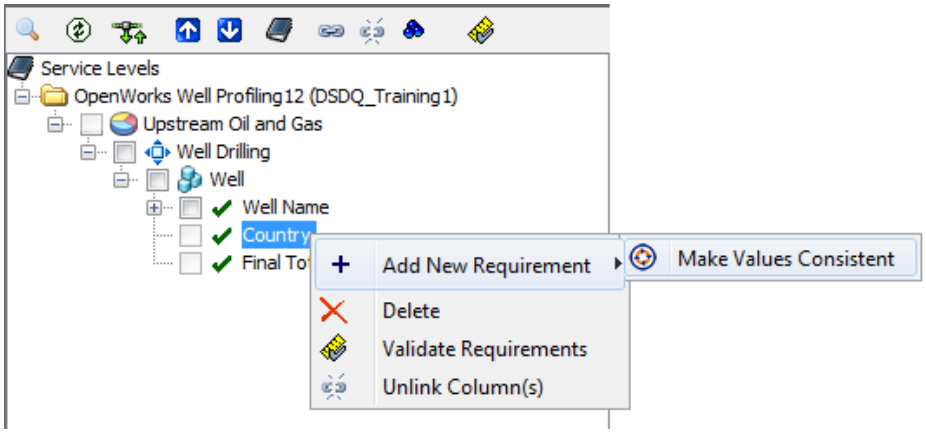
9. Click ⊞ on the Service Level Tree to expand the **DSDQ_Training_Rules** service level.

10. Expand the **Upstream Oil & Gas** sector.

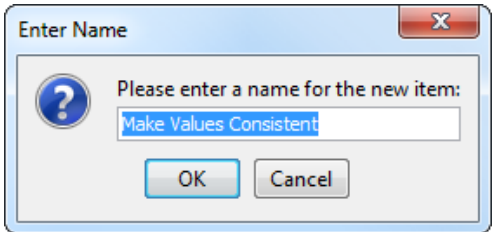
11. Expand the **Well Drilling** area.

12. Expand the **Well** element group.

13. Right-click the **Country** element on the Service Level Tree and select **Add New Requirement > Make Values Consistent** from the pop-up menu.

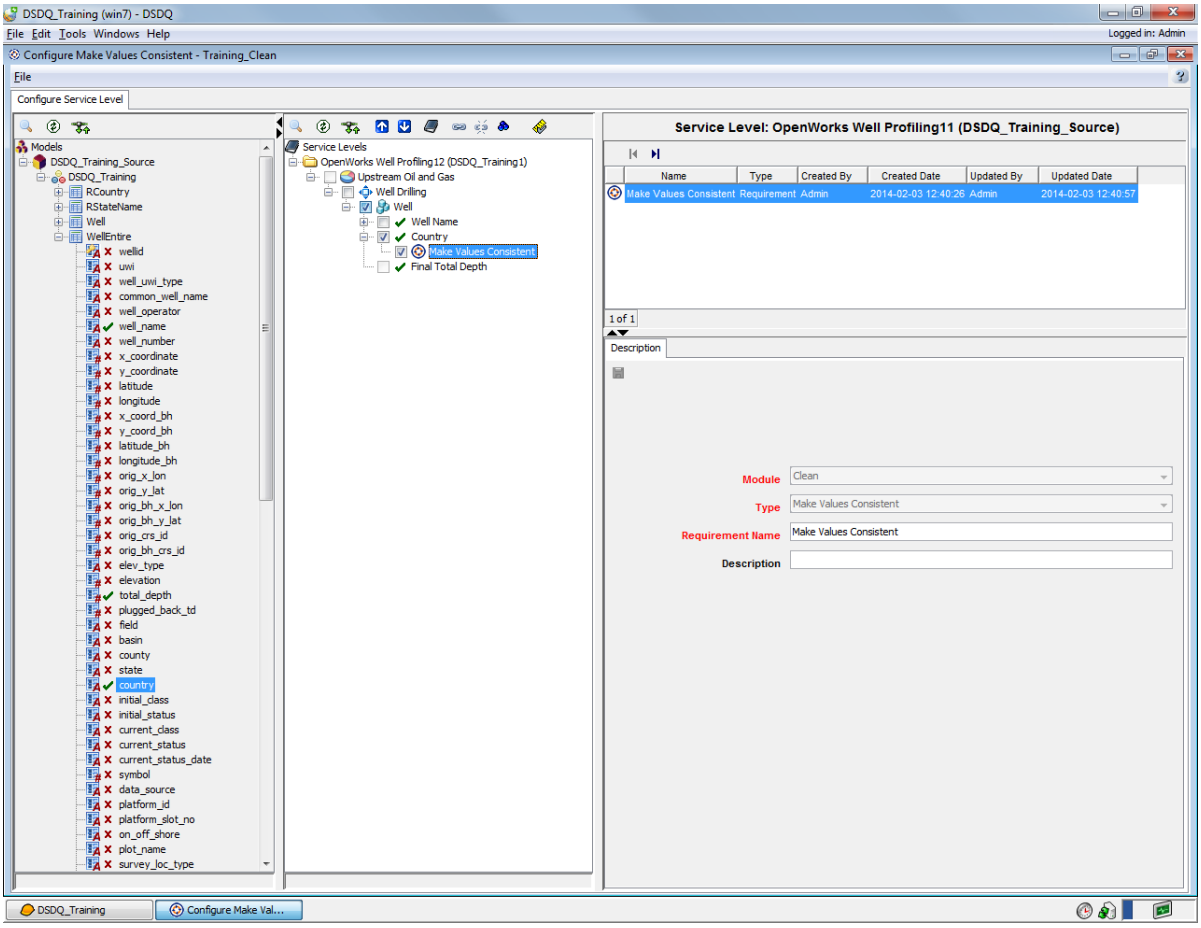


The **Enter Name** dialog box appears.



14. Optionally, specify a user-defined name for the requirement.

15. Click **OK**.
    The requirement is added and displays in the Service Level Tree.

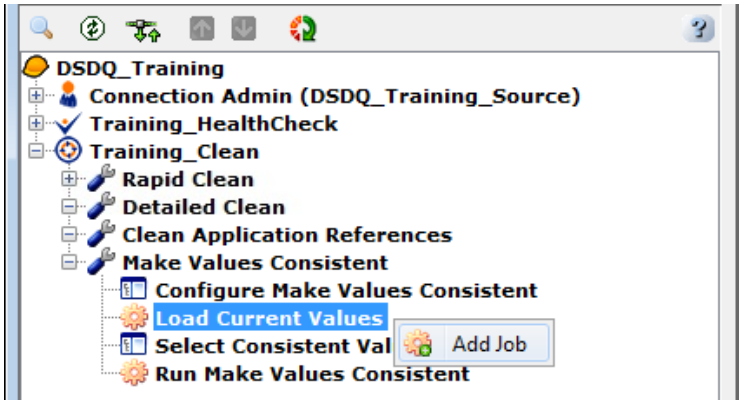


16. Select **File > Exit** from the menu bar on the **Configure Make Values Consistent** window.

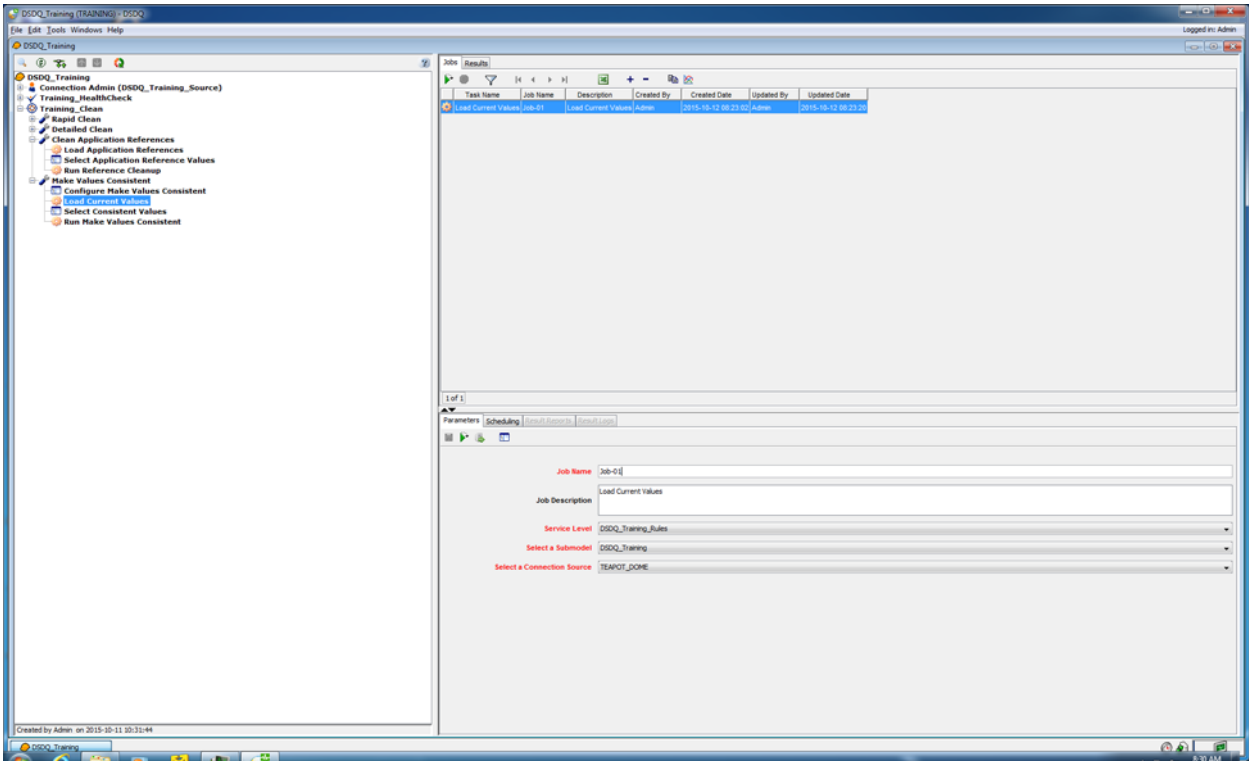
Exercise: Loading Current Values

To load current values in columns assigned to make values consistent:

1. Double-click the **Load Current Values** task or right-click the **Load Current Values** task, and select **Add Job** from the pop-up menu.



A new job is initiated and displays on the **Jobs and Results Listing Pane**.



2. Enter **Job-01** in the **Job Name** field.

3. Enter **Load Current Values** in the **Job Description** field.

4. Select **DSDQ_Training_Rules** from the **Service Level** drop-down list.

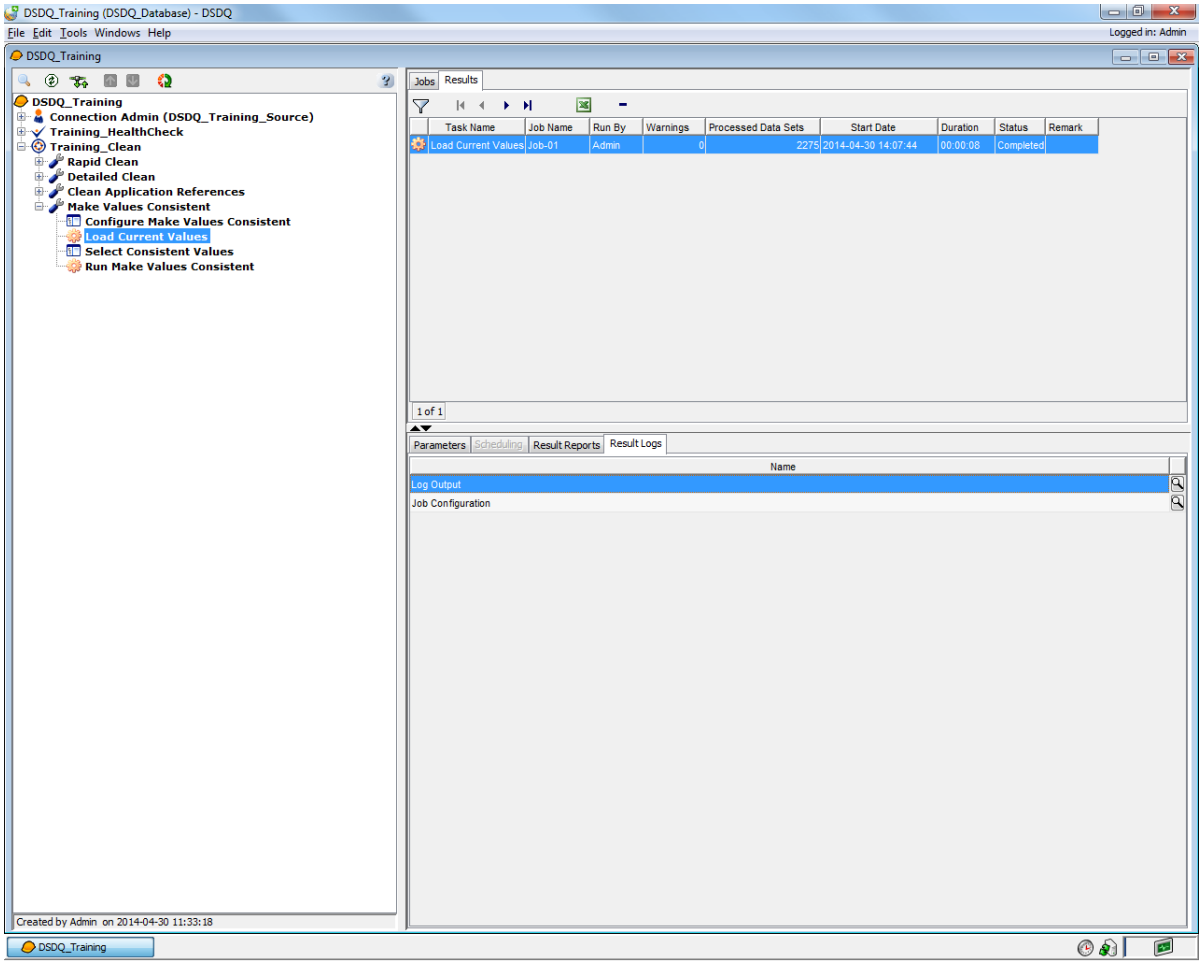
5. Select **DSDQ_Training** from the **Select a Submodel** drop-down list.

6. Select **TEAPOT_DOME** from **Select a Connection Source**.

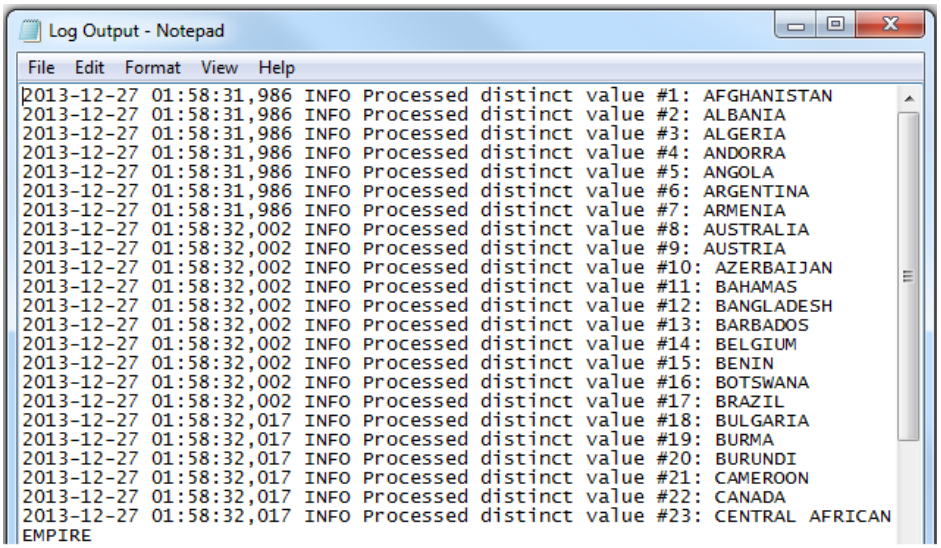
7. Click to save changes in the **Parameter** tab.

8. Click .
   The **Load Current Values** task runs and displays results in the **Result Reports** tab on the **Jobs and Results Information** Pane.

9. Select the **Results** tab on the **Jobs and Results Listing** Pane to view the values in the **Result Reports** tab on the **Job and Results Information** Pane.



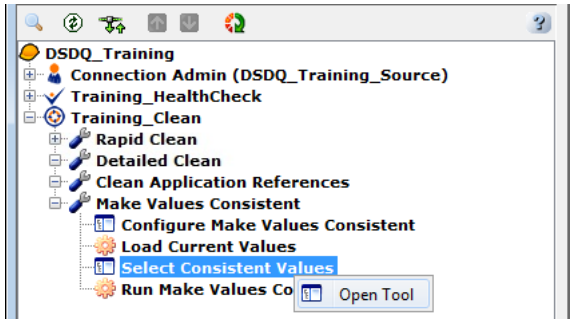
10. Double-click **Log Output** on the **Result Logs** tab to view results of the Run Reference Cleanup task.



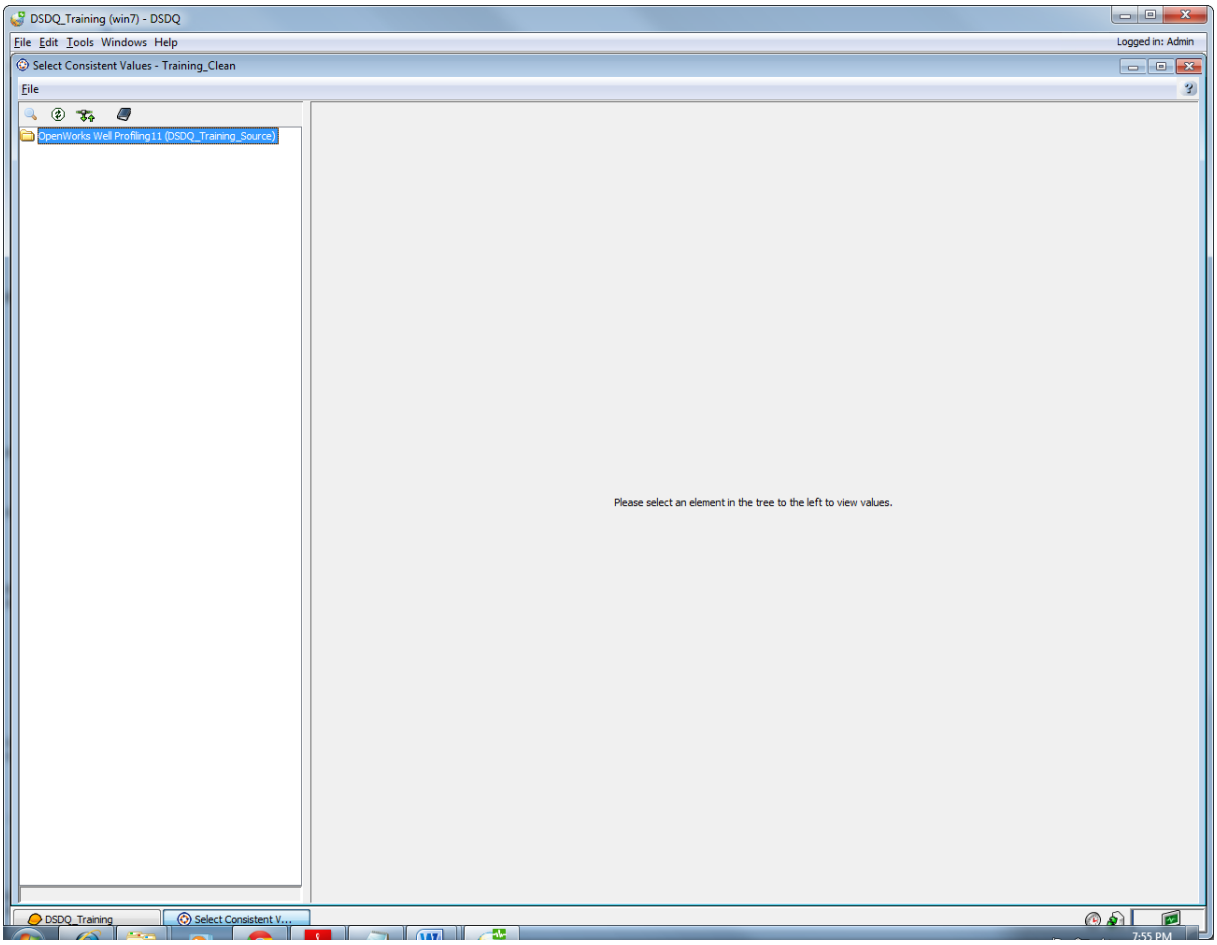
Exercise: Selecting Consistent Values

The **Select Consistent Values** tool allows you to enter consistent values for columns that have been assigned elements with make values consistent requirements. To select consistent values:

1. Double-click the **Select Consistent Values** tool or right-click the **Select Consistent Values** tool, and select **Open Tool** from the pop-up menu.



The **Select Consistent Values** window appears.

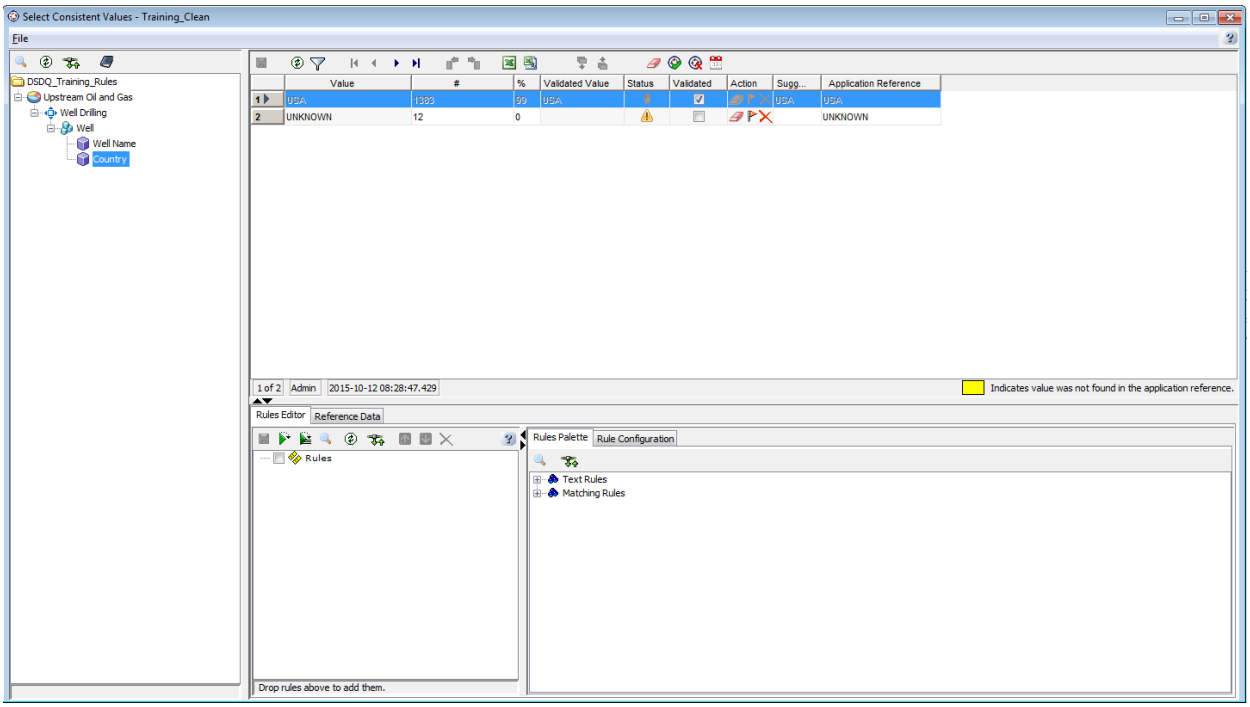


2. Click on the Service Level Tree toolbar.

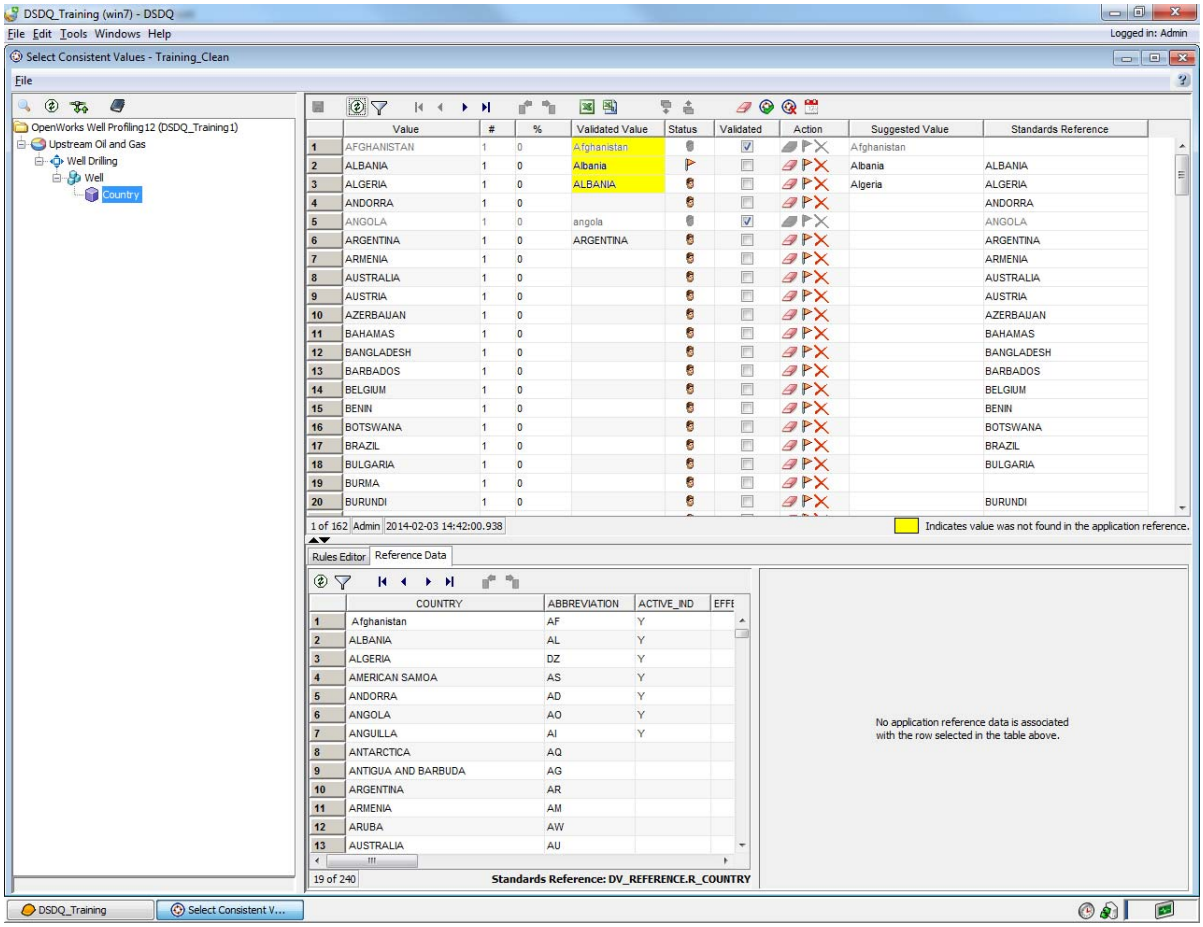
The **Select Service level** window appears.

3. Select **DSDQ_Training_Rules**.

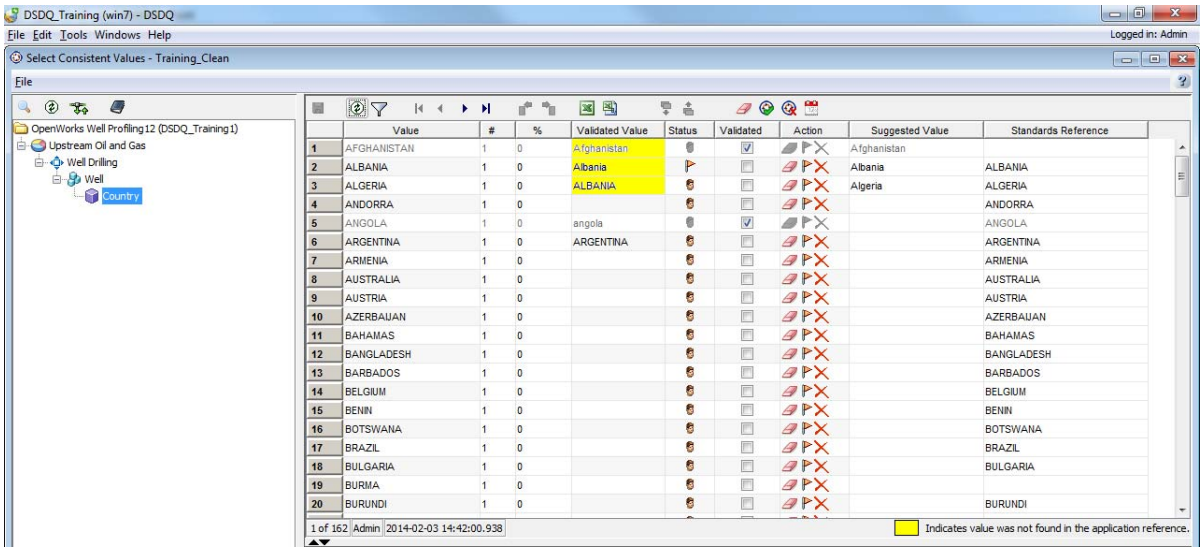
4. Click **OK**.
   All service levels that have make values consistent requirements are displayed.



5. Select an element in the Service Level Tree for which you want to select consistent values.The top right table displays the values for the column that is associated to the element.

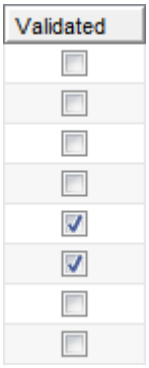


6. Select a value in the top table. If the value does not exist in the application reference table, its **Validated Value** will appear in yellow.



7. Enter the final value of the element in the **Validated Value** column. This value will be used when the **Run Make Values Consistent** job is processed.

8. Select the check box in the **Validated** column for the element for which you want to **Select Consistent Values**.



Note

The **Validated** column can only be checked if a Validated Value has been entered. Once the **Validated** column has been checked, all values in that row will no longer be editable. Once you unchecks the **Validated** column, the row becomes editable.

The **Action** column has the following three buttons:

- **Clear Validated Value** - clear the Validated Value for that row.
- **Flag Value** - set the value to be looked at in future.
- **Set Value to Null** - remove the current value and set it to Null.

On clicking the **Validate All Rows** button in the toolbar, the Status column will be set to **Updated by User** for all records and the Validated column will be checked, as long as a Validated Value exists for that row.

On clicking the **Clear All Validated Rows** button in the toolbar, the Status and Validated columns values will be unchecked.

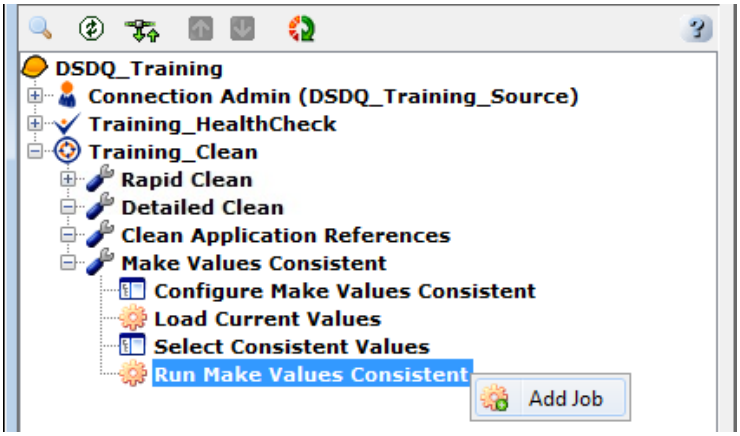
Once the elements and validated values have been correctly configured, the values can be modified by running the **Run Make Values Consistent** task. If the **Load Current Values** task is run again, then the processed records are stored and can be viewable by clicking on the **Previous Validated Values** button. This is a toggle button and clicking it again will allows you to view only current data.

9. Click to save changes.
10. Select **File > Exit** from the menu bar on the **Select Consistent Value** window.

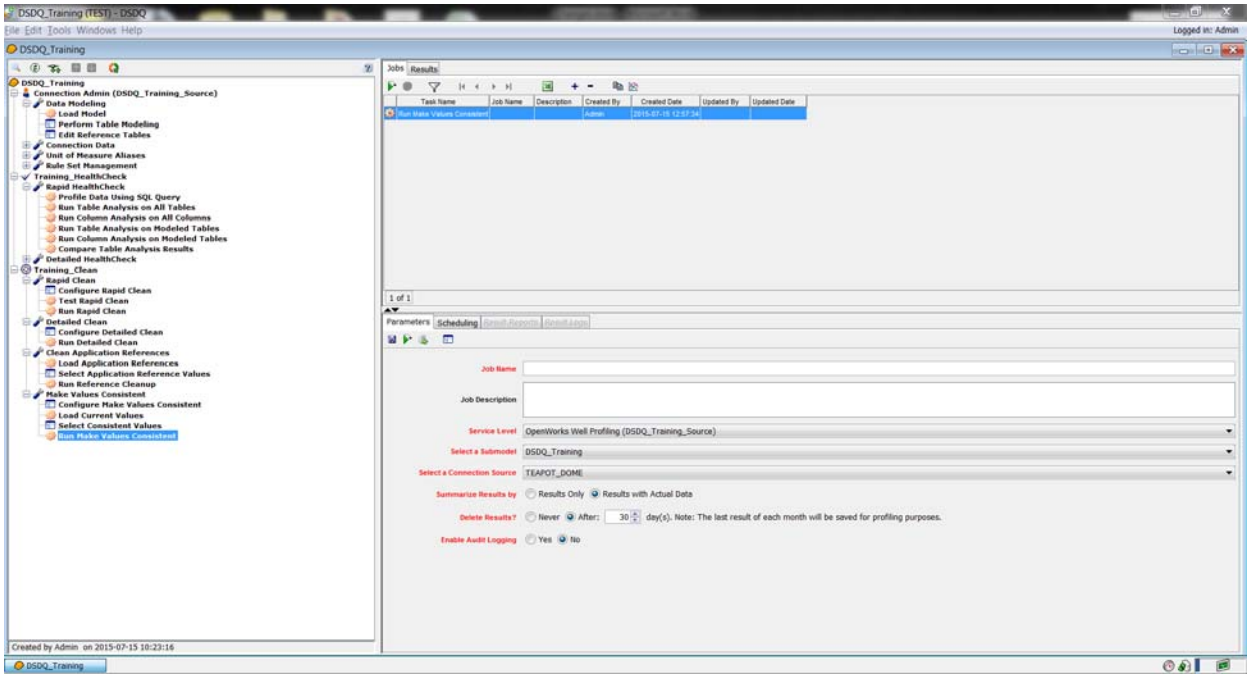
Exercise: Running Make Values Consistent

The Run Make Values Consistent task modifies values with Validated Values for the data and updates the actual data. To run the Make Values Consistent task:

1. Double-click the **Run Make Values Consistent** task or right-click the **Run Make Values Consistent** task, and select **Add Job** from the pop-up menu.



A new job is initiated and displays in the **Job and Results Listing Pane**.



2. Enter **Job-01** in the **Job Name** field.

3. Enter **Make Values Consistent** in the **Job Description** field.

4. Select **DSDQ_Training_Rules** from the **Service Level** drop-down list.

5. Select **DSDQ_Training** from the S**elect Submodel** drop-down list.

6. Select **TEAPOT_DOME** from the **Select a Connection Source**.

7. Select the **Results with Actual Data** option for **Summarize Results by**.

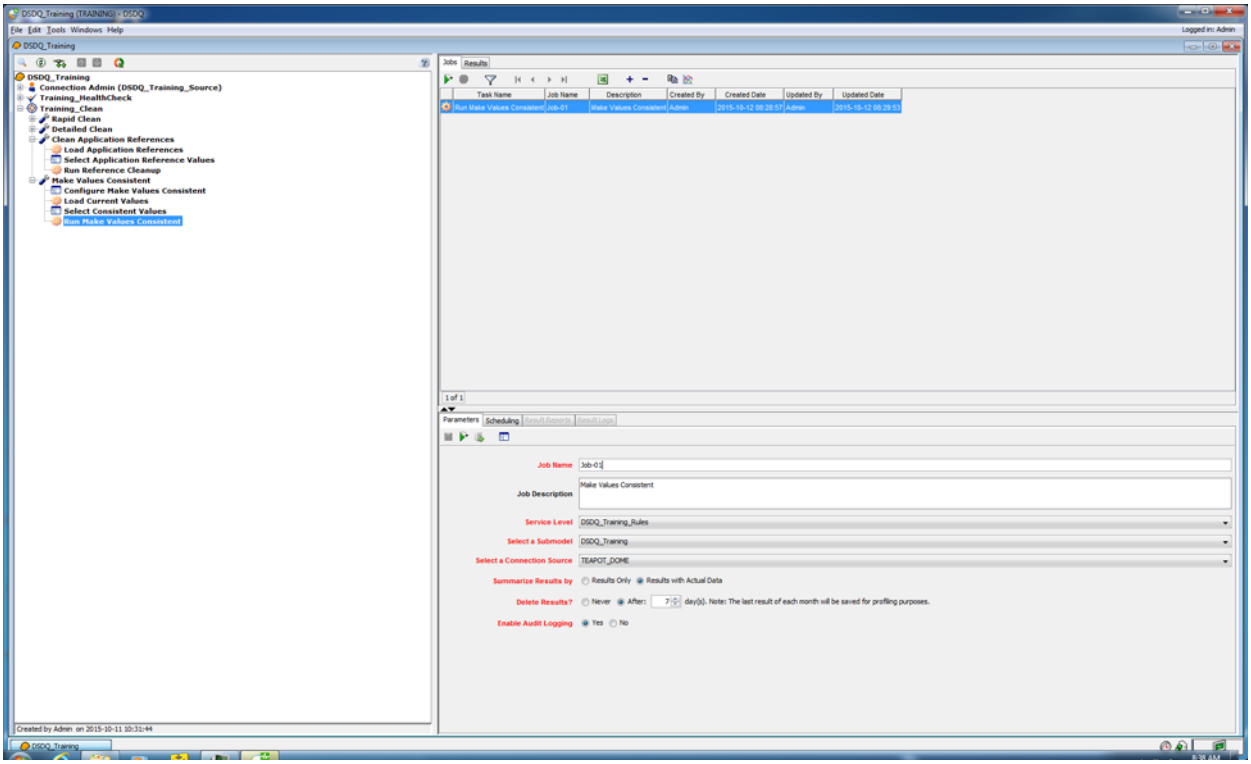
8. Select the **After** option for **Delete Results?** Leave the number of days as **7**.

9. Select the **Yes** option for **Enable Audit Logging**.

10. Click  to save changes.

11. Click  .
    The **Run Make Values Consistent** task is executed and displays results in the **Result Reports** tab.

12. Select the **Results** tab on the **Job and Results Listing** Pane to view the values in **Result Reports** tab on the **Job and Results Information Pane**.



13. Click on the **Result Reports** tab to display **Make Values Consistent Results** in PDF format.

Make Values Consistent Results

HALLIBURTON
Landmark Software & Services

Project: DSDQ_Training
Phase: Training_Clean
Task: Run Make Values Consistent
Job: job-01
Connection: DSDQ_Training_Source
Result Date: Sat, Dec 28, 2013 02:08

Table Name: RCountry **Column Name:** country_name

Original	Accepted	Rows Affected	Result %	Remark
AFGHANISTAN	Afghanistan	1	0	
ALGERIA	Algeria	1	0	
ALBANIA	Albania	1	0	

Chapter 6
Managing Data Duplication in DecisionSpace Data Quality

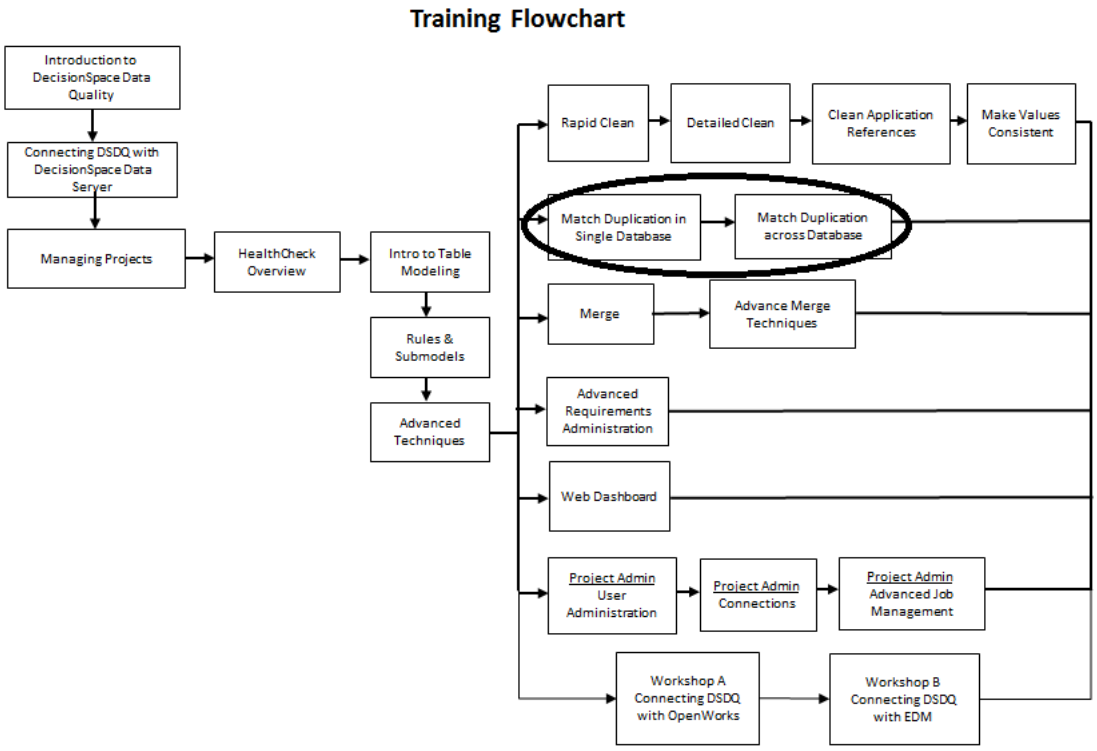
The Match Phase enables you to address data duplication issues found in data sources. Duplication can occur within a data source as well as across data sources. Once data duplication has been identified, you have the ability to clean or merge the duplicated data.

Chapter Overview

In this chapter, you will learn about:

- Data duplication
- Using the **Detailed Match** Activity for a single data source
- Managing duplication for a single data source
- Using the **Detailed Match** Activity across data sources
- Managing duplication across data sources

Topics covered in each chapter are outlined in the following illustration. Those specific to the current chapter will be circled in black for your reference:



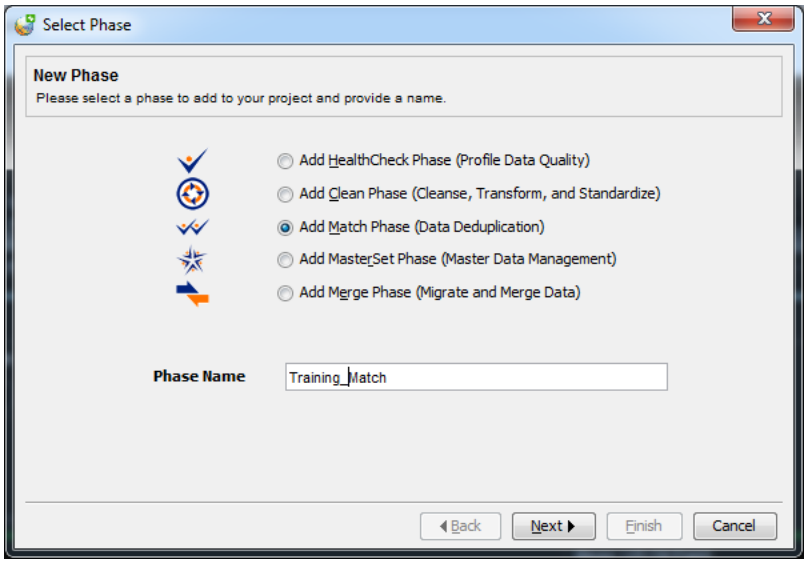
Data Duplication

Data Duplication occurs when multiple copies of the same data are created. This process is performed to create backups of specific data and/or can occur due to automated duplication of records. In situations where unwanted duplicate data copies are created, a procedure is implemented to remove or merge the duplicated copies into one clear source of accurate and updated information. The data quality application matches common records and removes or merges duplication. This not only helps in saving databases space but also assists in data organization.

Exercise: Adding a Match Phase for a Single Data Source

To add a Match Phase for a single data source:

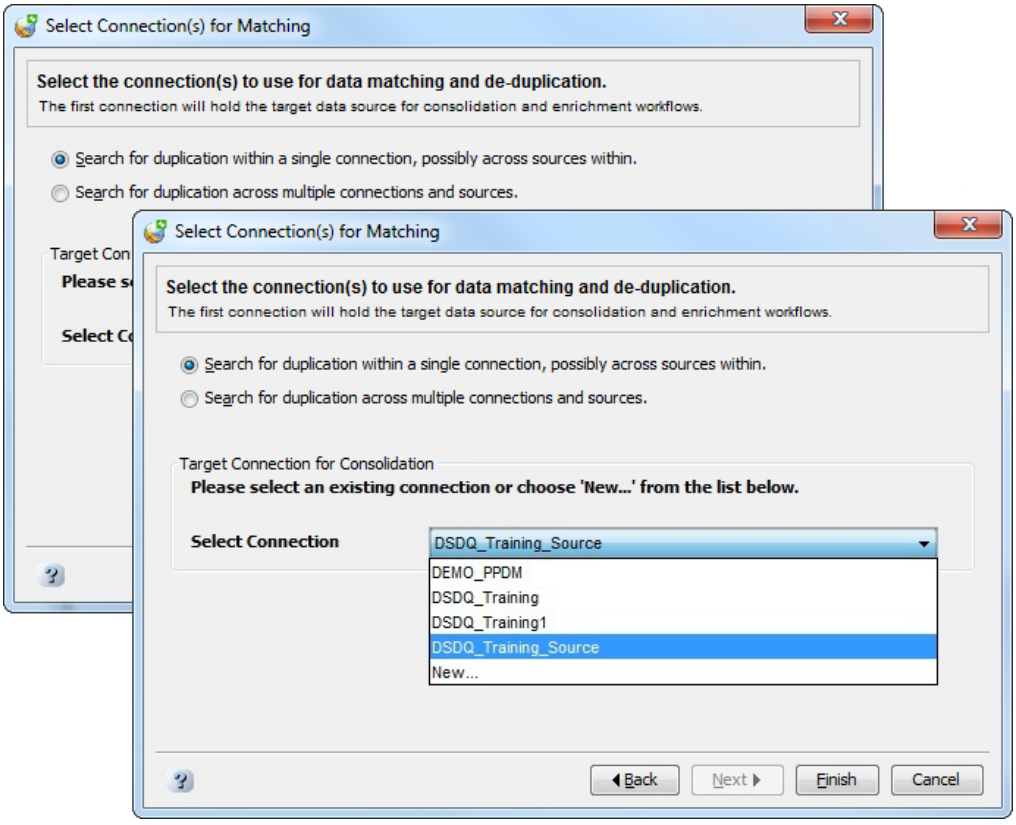
1. Click the **Add New Phase** button on the project toolbar. The **Select Phase** window appears with the **Add HealthCheck Phase (Profile Data Quality)** option selected by default.



2. Select the **Add Match Phase (Data Deduplication)** option.

3. Enter **Training_Match** in the **Phase Name** field.

4. Click **Next** to continue.
   The **Select Connection(s) for Matching** window appears with the **Search for duplication within a single connection, possibly across sources within** option selected by default.



5. Select **DSDQ_Training_Source** from the **Select Connection** drop-down list and click **Finish**.
   The Match Phase is created and displayed in the DecisionSpace Data Quality Project window.

Detailed Match for a Single Data Source

The **Detailed Match** Activity helps you in creating your own source data sets within match groups. Duplicated matches are found within the source data and you can setup specific match requirements for finding matches. Once the matches are found, they can be fixed. During the **Detailed Match for a Single Data Source** activity, only a specific data source is configured for duplication removal.

Exercise: Configuring the Detailed Match for a Single Data Source

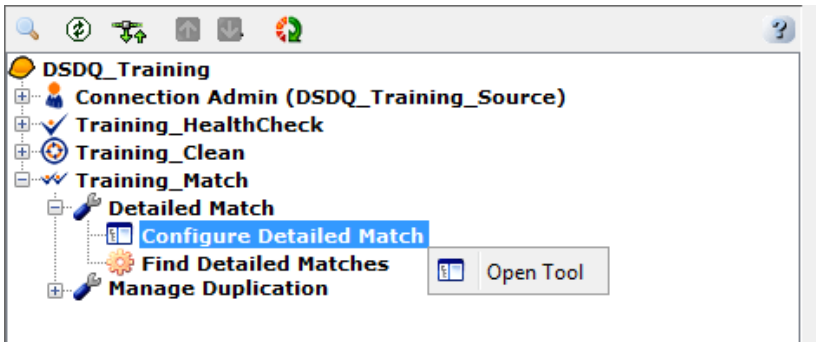
The **Configure Detailed Match** tool is used to create match groups and configure service levels for testing prior to running the **Find Detailed Matches** task. The match group consists of a single submodel. The user can select which requirements to enable/disable in the service level. You will also need to select a subset of data for testing.

To configure detailed match for a single data source:

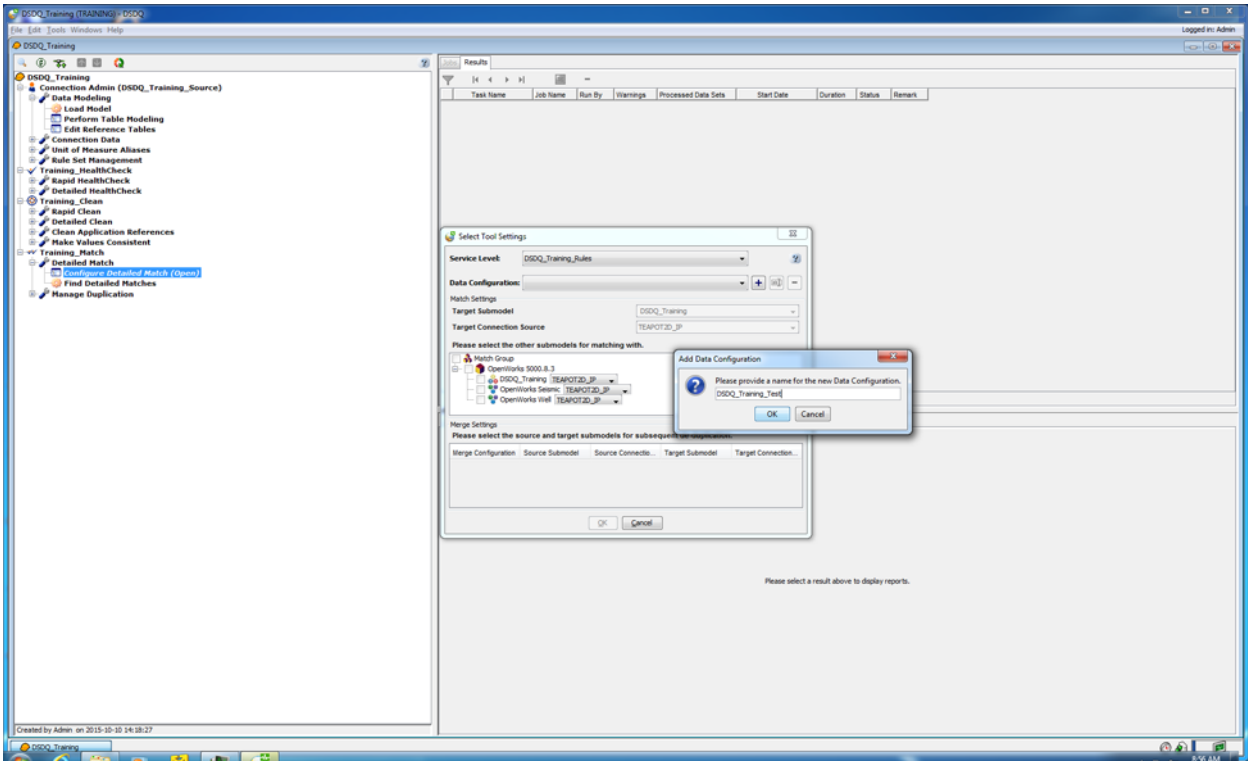
1. Click ⊞ on the DecisionSpace Data Quality Tree to expand the **Training_Match** phase.

2. Click ⊞ to expand the **Detailed Match** activity.

3. Double-click the **Configure Detailed Match** tool or right-click the **Configure Detailed Match** tool, and select **Open Tool** from the pop-up menu.

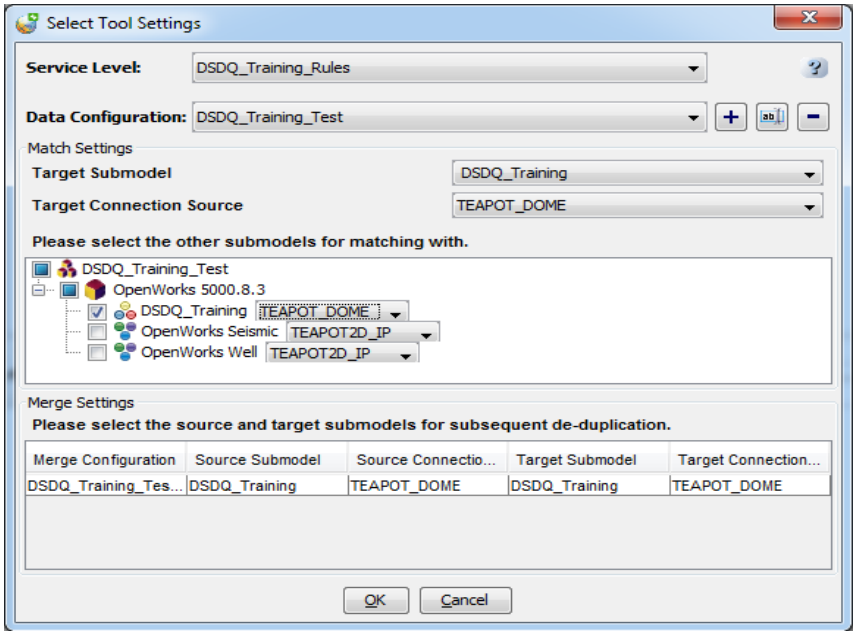


The **Add Data Configuration** dialog box appears by default the first time you run the **Configure Detailed Match** Tool.



4. Enter **DSDQ_Training_Test** in the **Add Data Configuration** dialog box.

5. Click **OK**.
   The **Select Tool Settings** window appears.



6. Select **DSDQ_Training_Rules** from the **Service Level** drop-down list.

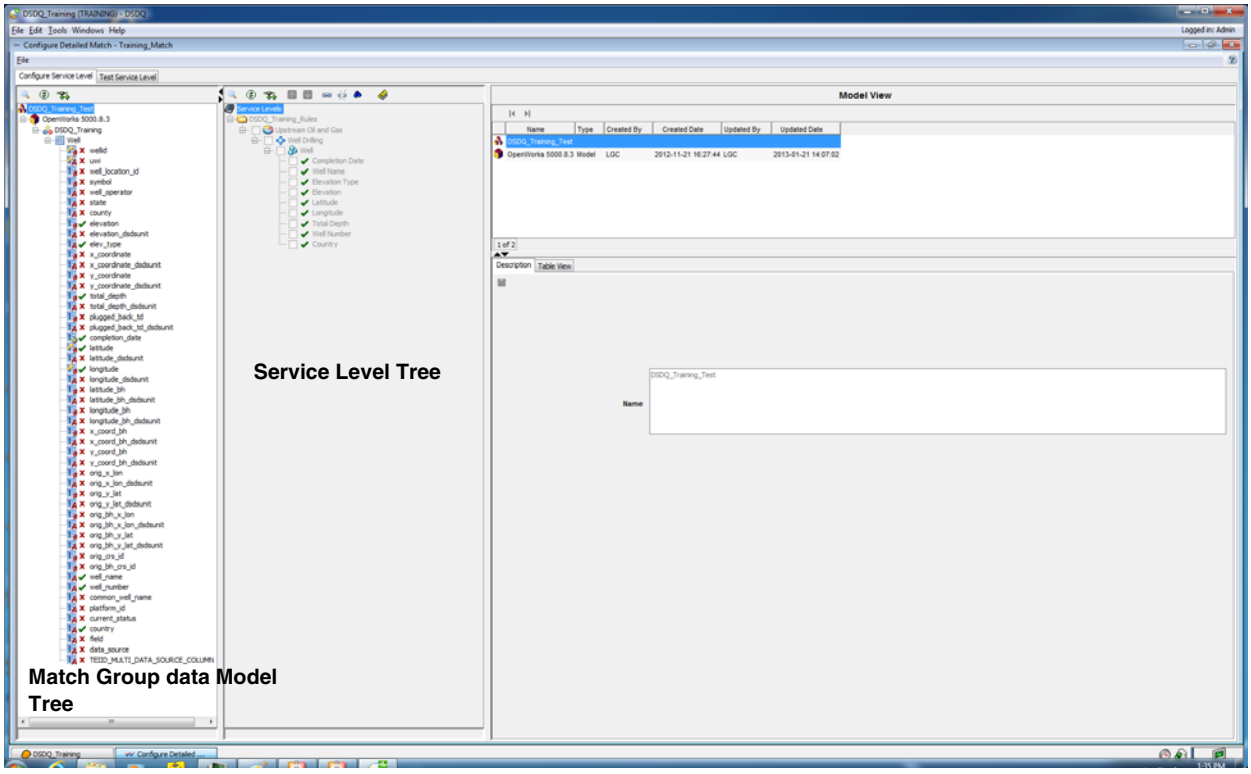
7. Select **DSDQ_Training_Test** from the **Data Configuration** drop-down list.

8. Select **DSDQ_Training** from the **Target Submodel** drop-down list.

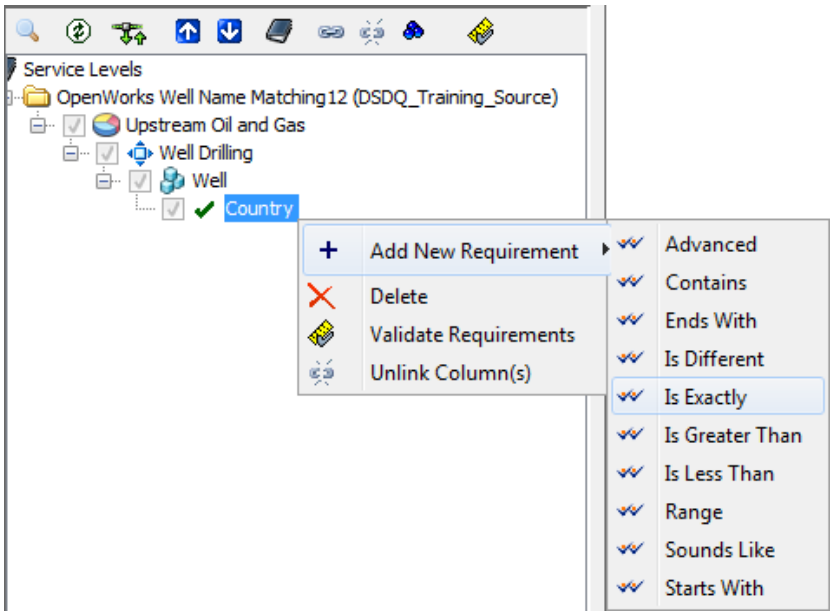
9. Select **TEAPOT_DOME** from the **Target Connection Source**.

10. Select the **DSDQ_Training_Test** check box and then select **TEAPOT_DOME** from the drop-down list.

11. Click **OK**.
    The **Configure Detailed Match** window appears with **DSDQ_Training_Test** and **DSDQ_Training_Rules** displaying in the Match Group Data Model Tree and Service Level Tree, respectively.



12. Right-click the **Country** element in the Service Level Tree and select **Add New Requirement > Is Exactly** from the pop-up menu.



The **Enter Name** dialog box appears.

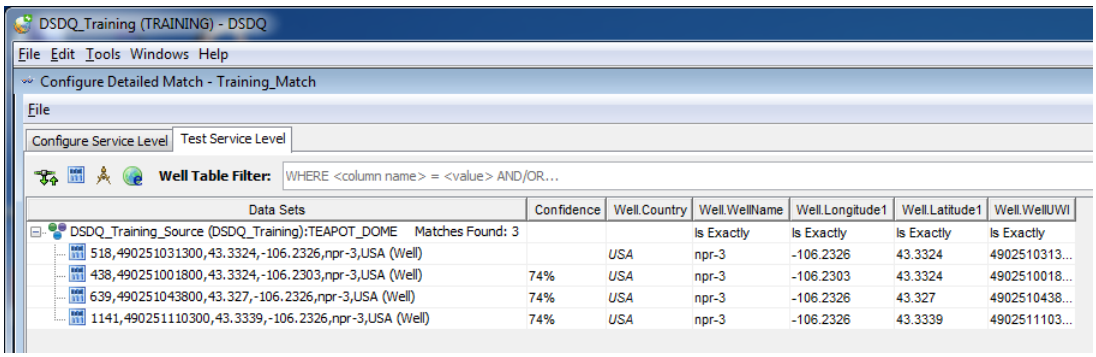
13. Optionally, specify a user-defined name for the requirement.

14. Click **OK** to add the requirement to the selected element.

15. Optionally, repeat steps 11 through 14 to add all elements for matching. It is recommended that you use at least the following:

    - Well Name
    - UWI
    - Latitude
    - Longitude

16. Click the **Test Service Level** tab.
    The test is automatically executed for the first record of the test data subset.



17. Verify all entries have correct matches.

18. Click the **Next Data Set** button to test the next record.

19. Repeat step 25 to test all records.

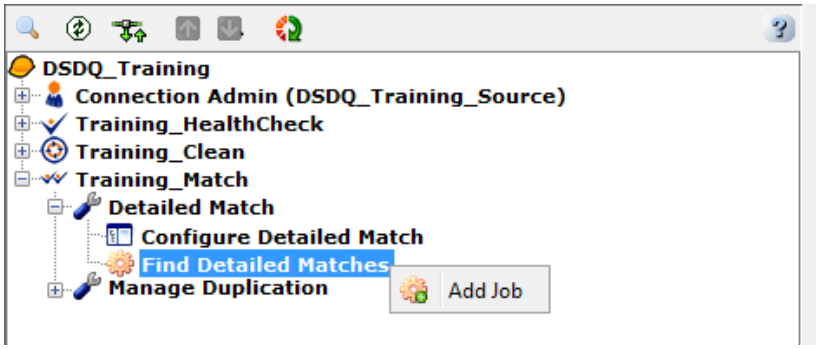
20. Select **File > Exit** from the menu bar on the **Configure Detailed Match** window.

Exercise: Finding Detailed Matches within a Single Data Source

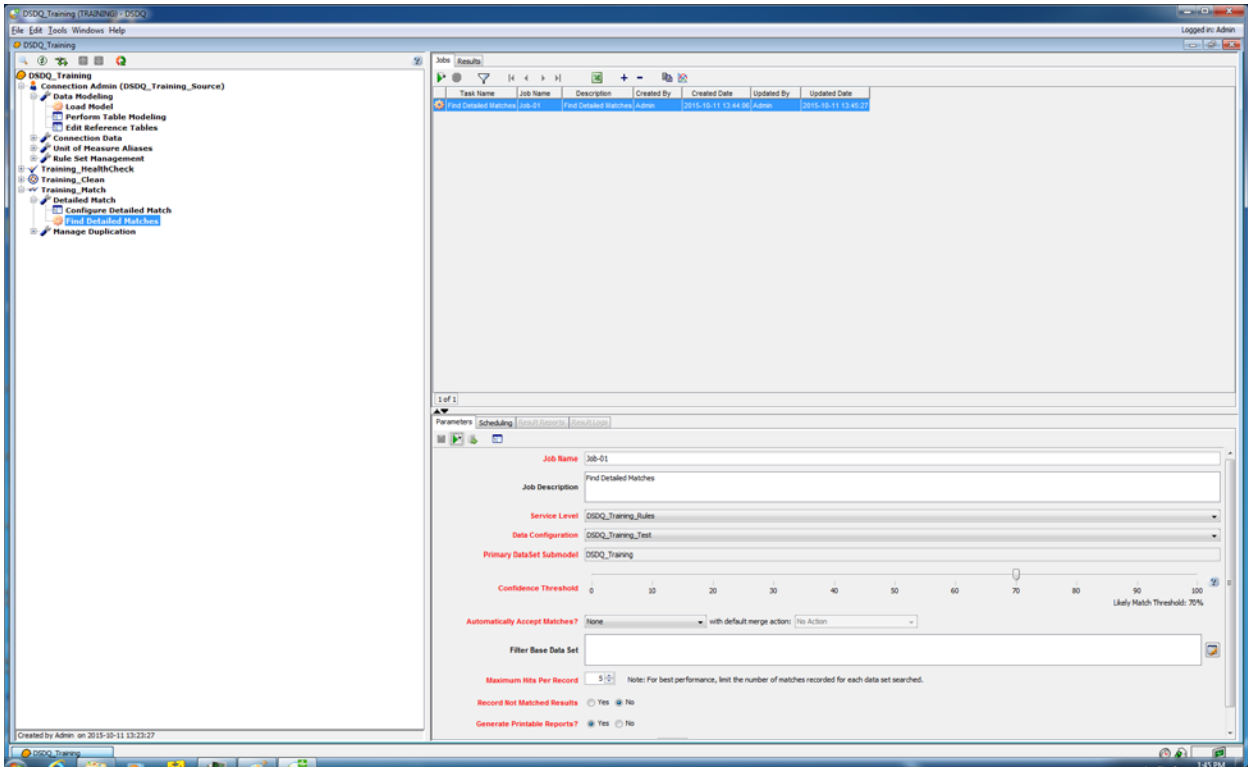
The **Find Detailed Match** task allows you to check values that match, or are redundant within the database.

To find Detailed Matches within a Single Data Source:

1. Double-click the **Find Detailed Matches** task or right-click the **Find Detailed Matches** task, and select **Add Job** from the pop-up menu.



A new job is initiated and displays on the **Jobs and Results Listing Pane**.



2. Enter **Job-01** in the **Job Name** field.

3. Enter **Find Detailed Matches** in the **Job Description** field.

4. Select **DSDQ_Training_Rules** from the **Service Level** drop-down list.

5. Select **DSDQ_Training_Test** from the **Data Configuration** drop-down list.
   The **Primary DataSet Submodel** field populates automatically.

6. Set the **Confidence Threshold** option as **70**.

> **Note**
>
> **Confidence Thresholds** allows you to set the minimum required match confidence for a selected match hit to be recorded. Setting a higher value here will be more selective in which possible matches appear in the Match results viewer. Setting a lower value will be more permissive, allowing for more matches to be presented, though they may possibly be incorrect upon expert review.

7. Select **None** from the **Automatically Accept Matches** drop-down list.

> **Note**
>
> **Automatically Accept Matches** allows the **Find Detailed Matches** task to automatically match depending on the criteria selected: either **All** or **Having only one match per source**. It will then mark the merge action on each match according to the default merge action selected.

8. Optionally, set a filter on the data subset.

9. Set the **Maximum Hits Per Record** option as **5**.

10. Select the **No** option for **Record Not Matched Results**.

11. Select the **Yes** option for **Generate Printable Reports?**

12. Select the **No** option for **Enable Data Read Ahead?**

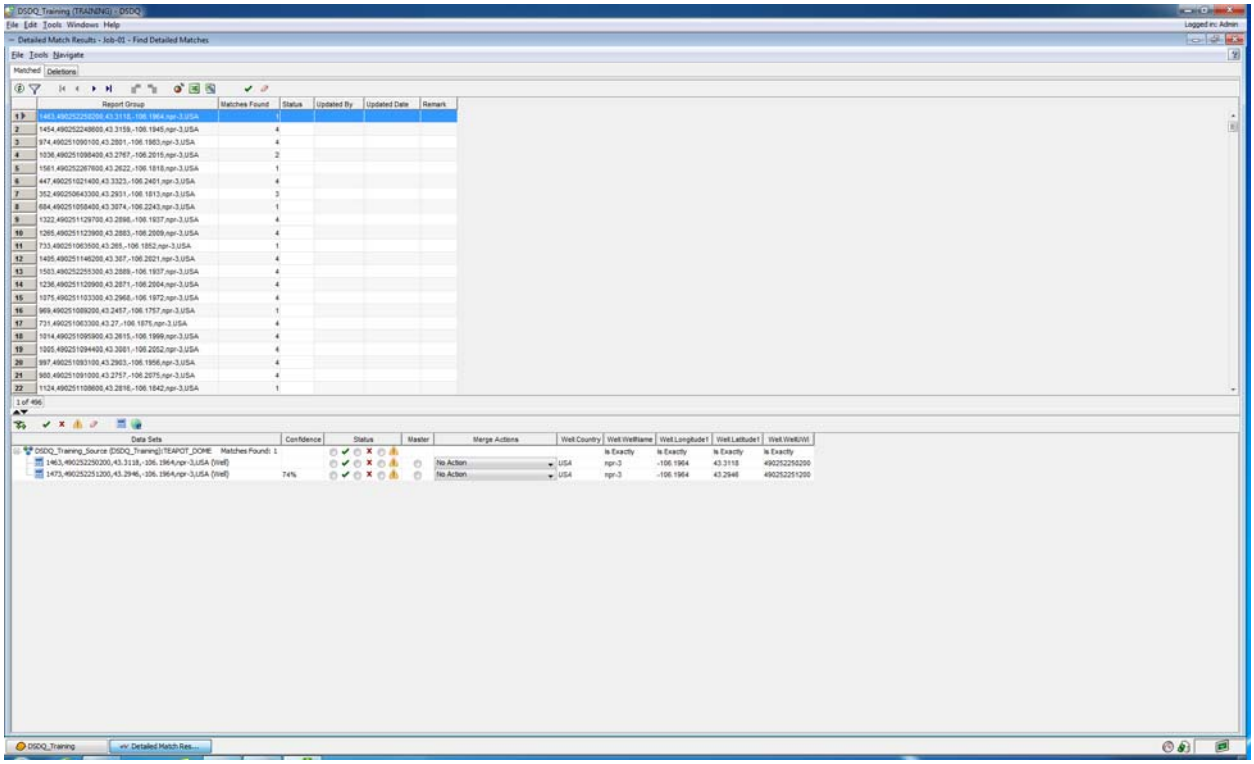
13. Select the **After** option for **Delete Results?** Leave the number of days as **30**.

14. Click to save changes in the **Parameters** tab.

15. Click .
    The **Find Detailed Matches** task runs and displays results in the **Result Reports** tab on the **Jobs and Results Information Pane**.

16. Select the **Results** tab on the **Jobs and Results Listing Pane** to view the values in the **Result Reports** tab on the **Job and Results Information Pane**.

17. Double-click **Detailed Match Results** on the **Result Reports** tab.
    The **Detailed Match Results** window appears.



18. Select all the rows in the **Matches Details** Pane and click ✔ on the Matches Found toolbar.

> **Note**
>
> Selecting the **Yes Status** ✔ radio button implies that you want the result to match the source.
>
> Selecting the **No Status** ✖ radio button implies that you do not want to match it to the source.
>
> Selecting the **Review Status** ⚠ radio button implies that the record still needs to be reviewed. You can set these statuses for all by clicking the respective buttons in the toolbar.
>
> To clear the status for this result set, click the **Clear Status** button.

19. Optionally, select the data and click the **Show on Browser Map** button on the toolbar to display the location of the data set on a map in the **DecisionSpace Data Quality** Dashboard. You can also display the data in a Pie chart by clicking the **Generate Status Chart** button.

20. Select **File > Exit** from the menu bar on the **Detailed Match Results** window.

Managing Duplication for a Single Data Source

The **Manage Duplication** activity manages the process of duplication removal with the right set of policies. This ensures that duplication removal standards are met. A submodel is configured for adding matches, which are then removed or merged as desired.

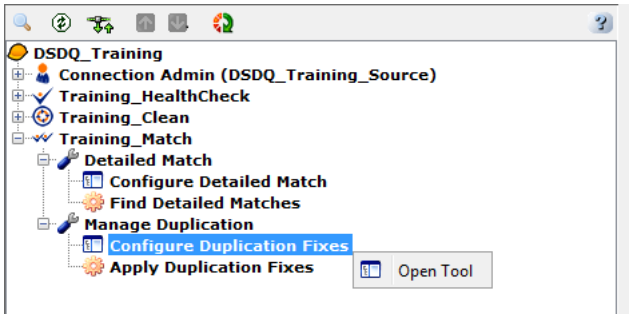
Exercise: Configuring Duplication Fixes for a Single Data Source

The **Configure Duplication Fixes** tool enables you to setup table and column mappings for the automatically created merge group, generated from the Match Group that you created in the **Detailed Match** activity.

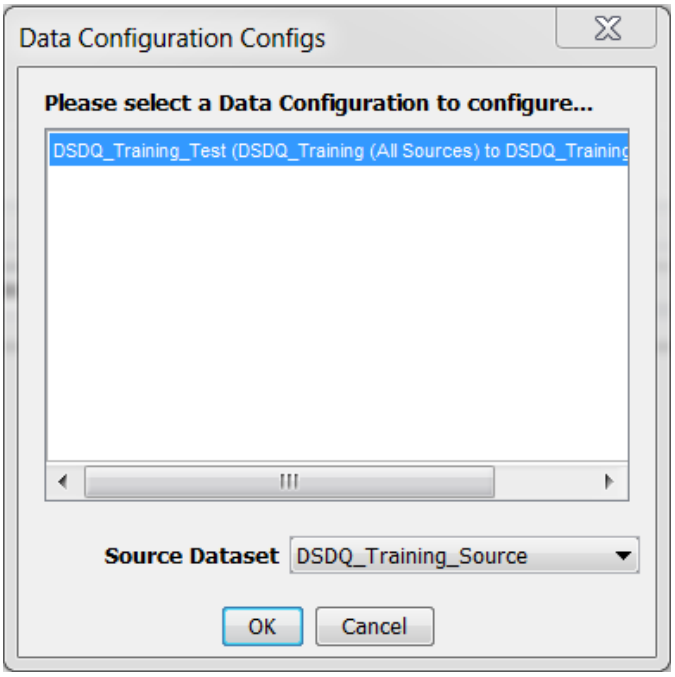
To Configure Duplication Fixes for a Single Data Source:

1. Click ⊞ to expand the **Manage Duplication** Activity on the DecisionSpace Data Quality Tree.

2. Double-click the **Configure Duplication Fixes** tool or right-click the **Configure Duplication Fixes** tool, and select **Open Too**l from the pop-up menu.



The **Data Configuration Configs** window appears.



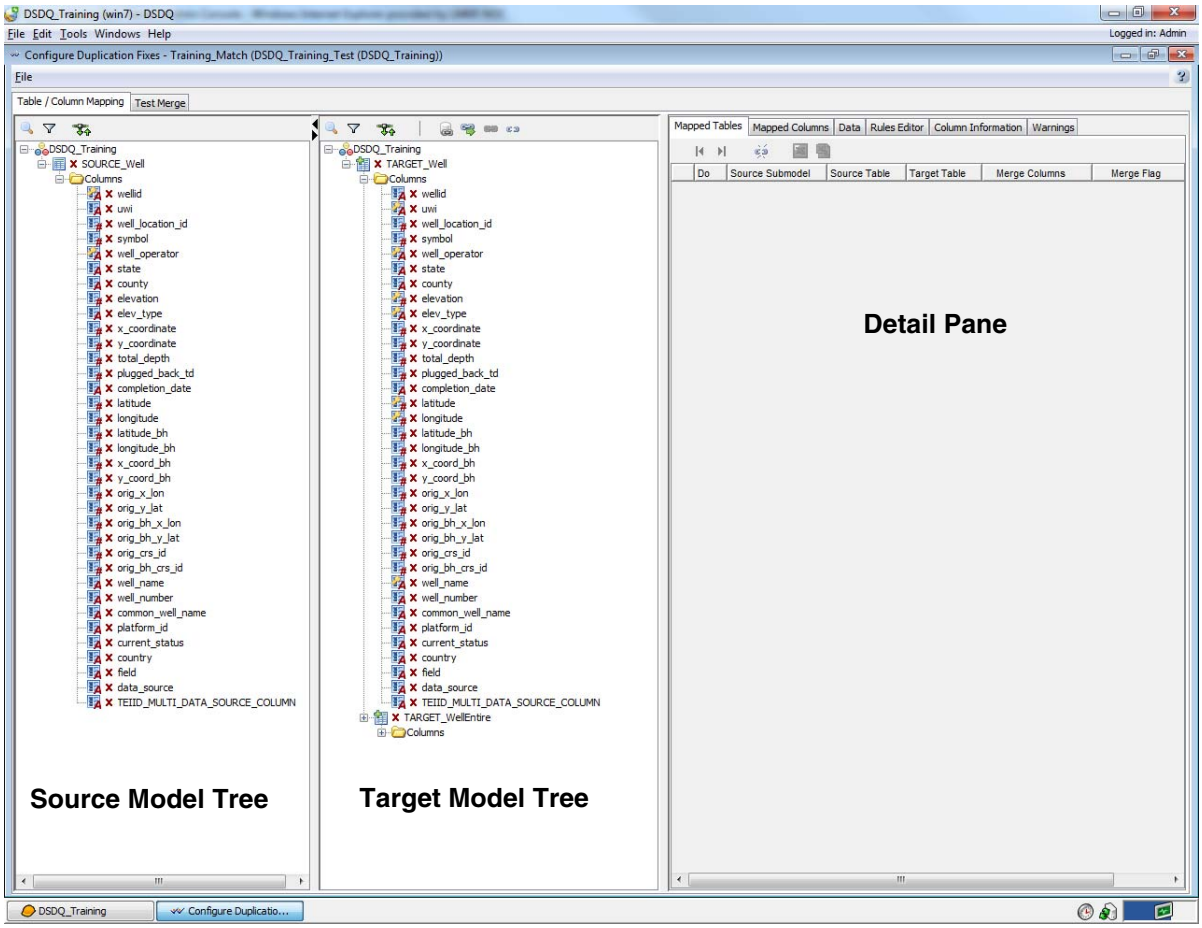
3. Select the **DSDQ_Training_Test (DSDQ_Training)** option.

4. Click **OK**.

   The **Configure Duplication Fixes** window appears, displaying tables and columns for the Source Model Tree, Target Model Tree and the Detail pane. The Detail Pane has six tabs:

Mapped Tables	This area displays information about the tables that have been mapped in the source and target trees.
Mapped Columns	This area displays the mapping between the target column and all its mapped sources.

Data	This area displays the source and target tables' data. Selecting a column in either the Source or Target tree highlights the corresponding column in the data view if available. If the selected column has already been mapped, the mapped column data is highlighted in each of the corresponding trees. The **Source Data** View and **Target Data** View toolbars can be used to **Filter Data**, to move to the **First Record in Current View** or **Last Record in Current View**, or to move to the **Next Data Set** or **Previous Data Set**. The data can also be saved in Excel format by clicking on the corresponding button or exported to a CSV file by selecting the **Create a CSV Export File** button on the toolbar. Data in the Data Detail Pane can be sorted by clicking a column header. Column information on any column can be viewed by right-clicking a column header and selecting Column Info from the pop-up menu. Other columns information can be viewed by right-clicking a column header and selecting Columns Filter from the pop-menu.
Rules Editor	Used to apply rules to specific columns. Simply drag the rule to the target column that the rule has to be applied to. The **Methods** tab is automatically populated with the relevant information. Make changes to the fields as needed.
Column Information	Displays basic information about the selected column: "Data Type", "Column Size", etc. The tab is divided into two vertical panes: the left one holds the source column information, and the right one holds target column information.
Warnings	This area will display any inconsistencies between the mapped columns, e.g. source column length is greater than the target column length. Initially this tab is blank. When the first warning is logged, the tab name turns red and a warning icon appears next to its name.



5. Select the **Source_Well** table from the Source Model Tree.

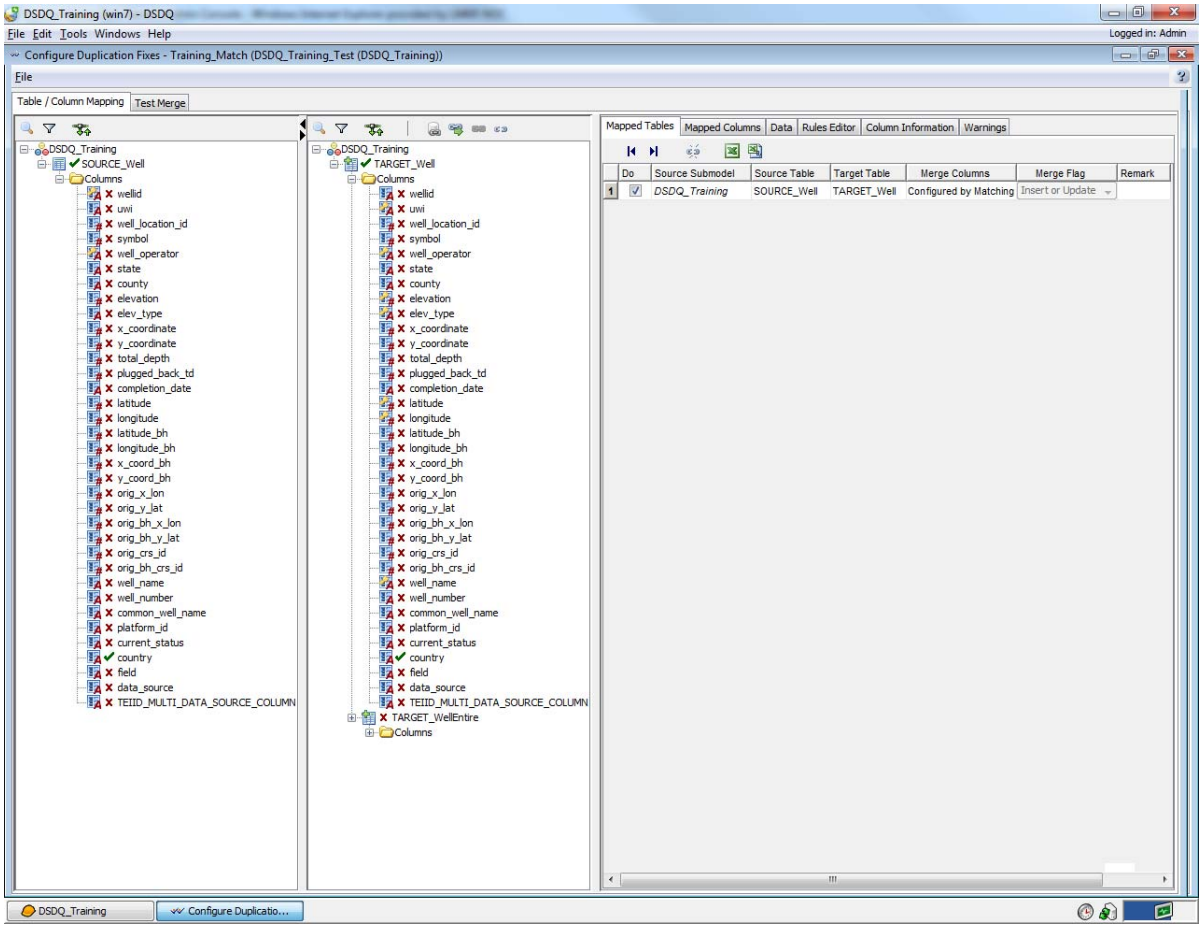
6. Select the **Target_Well** table from the Target Model Tree.

7. Click the **Auto Map selected Table/Columns** button on the toolbar.
   A green check mark appears adjacent to the selected tables.

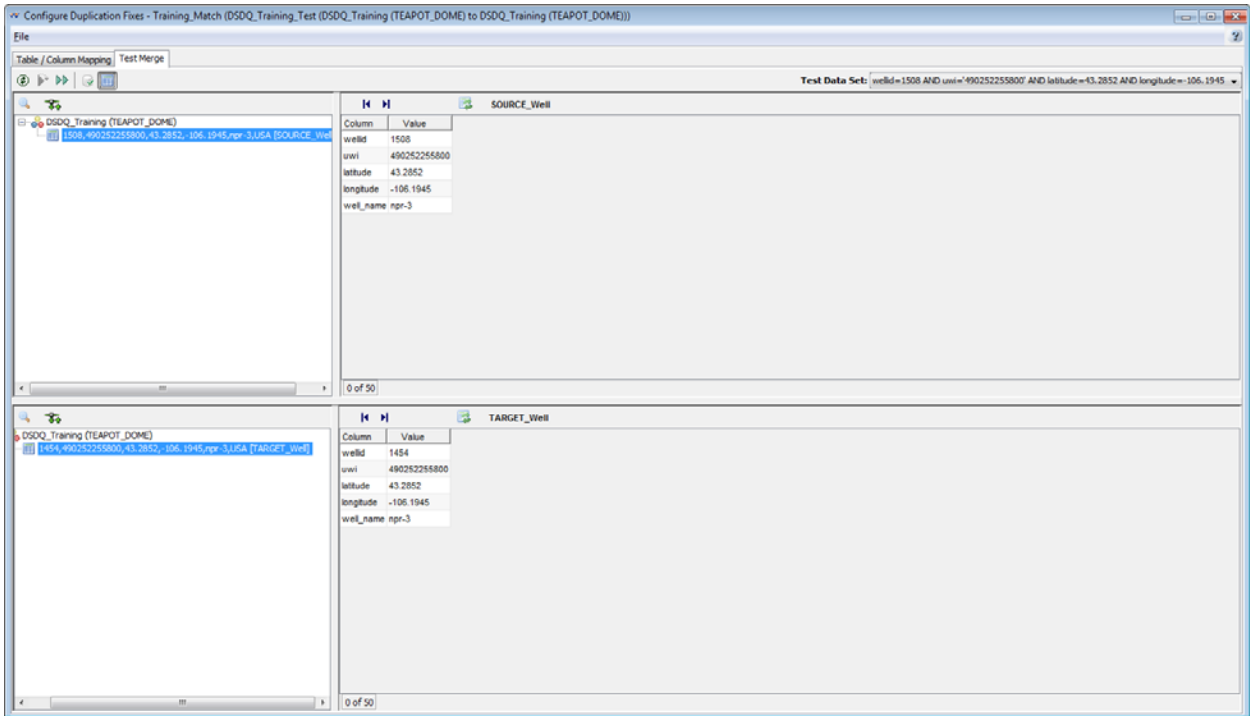
8. Select the following columns from the **Source_Well** and **Target_Well** table:

   - UWI
   - Latitude
   - Longitude
   - Well Name

9. Click the **Auto Map selected Table/Columns** icon on the toolbar.
   A green check mark appears adjacent to the **Country** column in both the Source Model Tree and Target Model Tree.



10. Select the **Test Merge** tab to test all match result configurations.



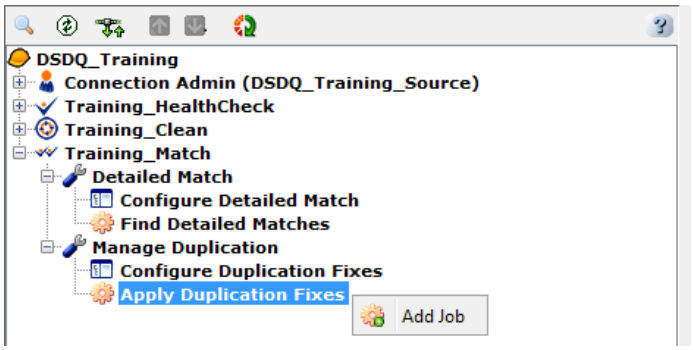
11. Select **File > Exit** from the menu bar on the **Configure Duplication Fixes** window.

Exercise: Applying Duplication Fixes for a Single Data Source

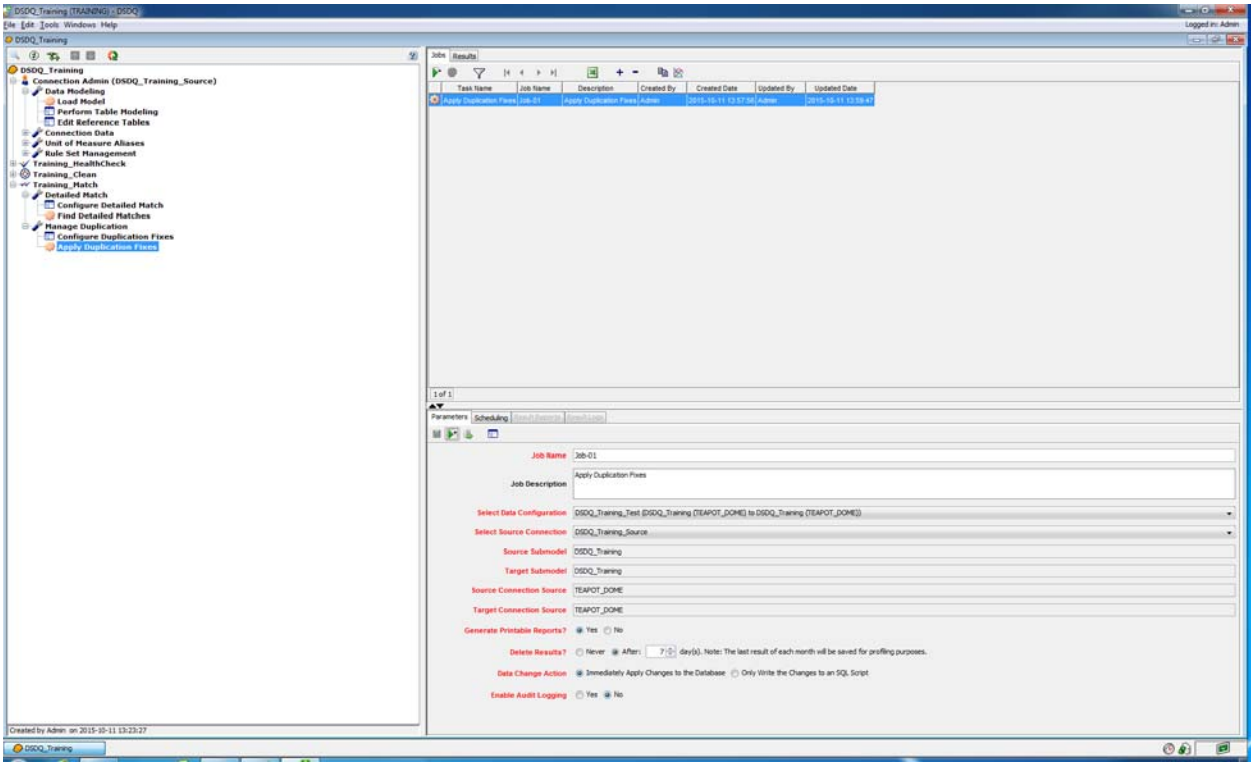
After the duplications have been tested in the **Configure Duplication Fixes** Tool, the **Apply Duplication Fixes** Task can be used to move the complete data set over to the target database and remove the duplication.

To Apply Duplication Fixes for a Single Data Source:

1. Double click the **Apply Duplication Fixes** task or right-click the **Apply Duplication Fixes** task and select **Add Job** from the pop-up menu.



A new job is initiated and displays on the **Jobs and Results Listing Pane**.



2. Enter **Job-01** in the **Job Name** field.

3. Enter **Apply Duplication Fixes** in the **Job Description** field.

4. Select **DSDQ_Training_Test (DSDQ_Training)** from the **Select Data Configuration** drop-down list.

5. Select **DSDQ_Training_Source** from the **Select Source Connection** drop-down list.
   The **Source Submodel** and **Target Submodel** fields populates automatically.

6. Select the **Yes** option for **Generate Printable Reports?**

7. Select the **After** option for **Delete Results?** Set the number of days as **7**.

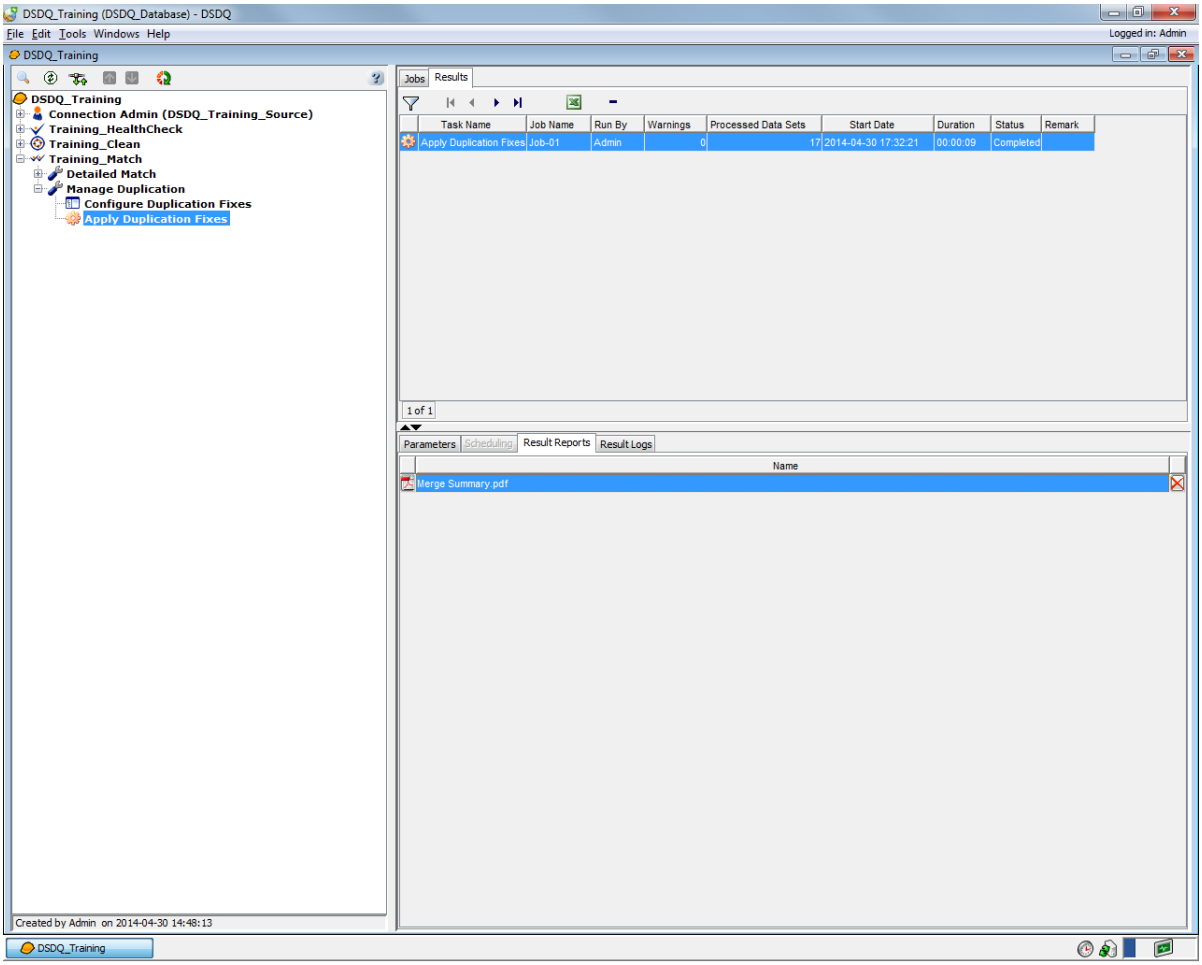
8. Select the **Immediately Apply Changes to the Database** option for **Data Change Action**.

9. Select the **No** option for **Enable Audit Logging**.

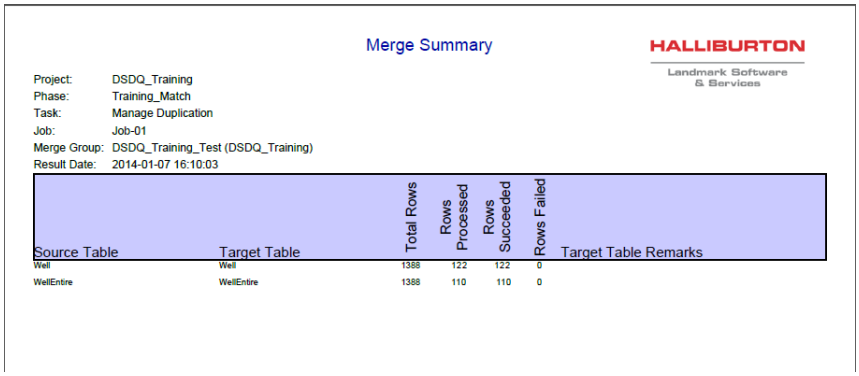
10. Click to save changes in the **Parameter** tab.

11. Click .
    The **Apply Duplication Fixes** Task is executed and displays results in the **Result Reports** tab.

12. Select the **Results** tab on the **Jobs and Results Listing Pane** to view the values in the **Result Reports** tab on the **Job and Results Information Pan**e.



13. Click on the **Result Reports** tab to display **Duplication Fixes** results in PDF format.



Merge Summary

HALLIBURTON
Landmark Software & Services

Project: DSDQ_Training
Phase: Training_Match
Task: Manage Duplication
Job: Job-01
Merge Group: DSDQ_Training_Test (DSDQ_Training)
Result Date: 2014-01-07 16:10:03

Source Table	Target Table	Total Rows	Rows Processed	Rows Succeeded	Rows Failed	Target Table Remarks
Well	Well	1388	122	122	0	
WellEntire	WellEntire	1388	110	110	0	

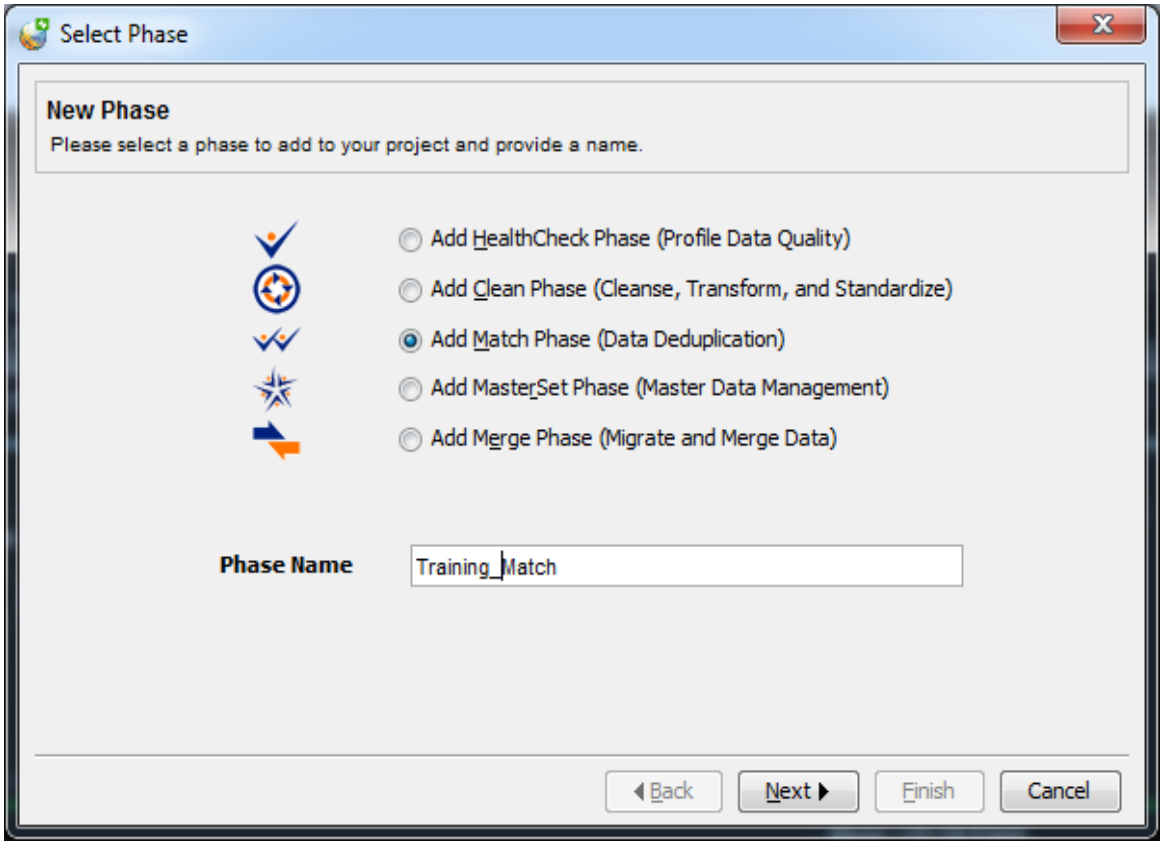
Detailed Match across Data Sources

The Detailed Match across Data Sources Activity helps in searching verifying and removing duplication across several data sources. The procedure is almost identical to the one followed for clearing out a single data source with the exception of creating a primary submodel and searchable submodels. The primary submodel contains the source data set that matches will be found for. The searchable submodels contain the data sets that matches will be found from.

Exercise: Adding a Match Phase across Data Sources

To add a Match Phase across Data Sources:

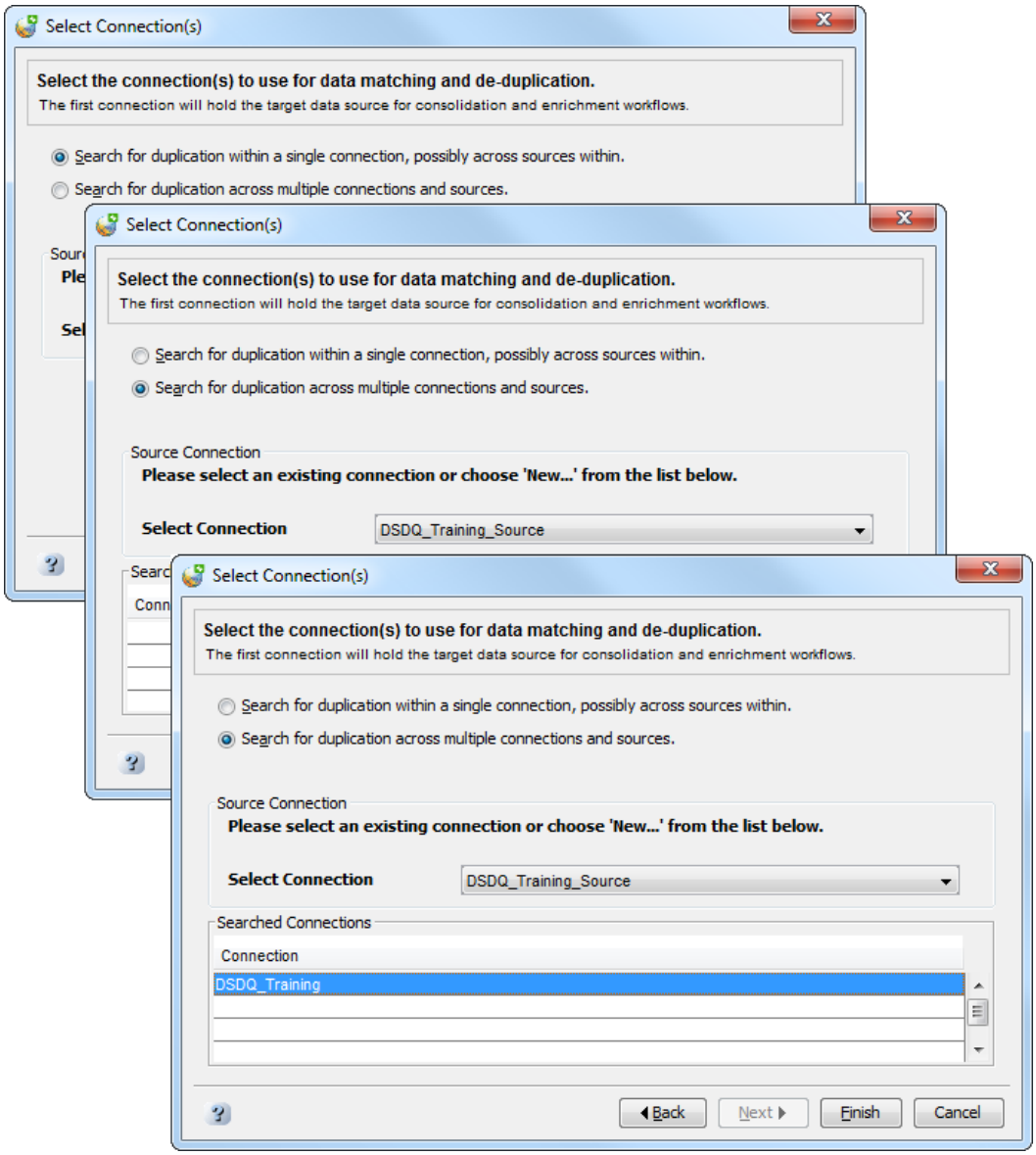
1. Click the **Add New Phase** button on the Project toolbar. The **Select Phase** window appears with the **Add HealthCheck Phase (Report on Data Quality Profiling)** option selected by default.



2. Select the **Add Match Phase (Data Deduplication)** option.

3. Enter **Training_Match_Across_DataSources** in the **Phase Name** field.

4. Click **Next** to continue.
   The **Select Connection(s)** dialog box appears with the **Search for duplication within a single connection, possibly across sources within** option selected by default.



5. Select the **Search for duplication across multiple connections and sources** option.

6. Select **DSDQ_Training_Source** from the **Select Connection** drop-down list.

> **Note**
>
> If it is not there, you will need to create this connection (DSDQ_Training) and add the Well table to your OpenWorks Well submodel as seen in Chapter 4: Perform Table Modeling, pg.4-8. Repeat steps 1-11 and 22. (In Step 1 select OpenWorks Well, you will be adding the Well column to it. Delete WellEntire from the submodel before Validating. After exiting, you will need to link new columns to their respective elements in Configure Detailed Match.)

7. Click on the first empty row under the **Searched Connections** text box and select **DSDQ_Training** from the **Connections** drop-down list.

8. Click **Finish.**
   The Match Phase is created and displayed in the DecisionSpace Data Quality Project Window.

Exercise: Configuring Detailed Match across Data Sources

The **Configure Detailed Match** Tool is used to create Match Groups (a selection of primary/searchable submodels and their respective Merge Groups) and configure Service Levels for testing prior to running the **Find Detailed Matches** Task. The Match Groups consist of a primary submodel, searchable submodel(s), and their respective Merge Groups. The user can select which requirements in the service level to enable/disable, and when testing a service level, what subset of the total data to use. A service level containing Match requirements must exist prior to opening the **Configure Detailed Match** Tool.

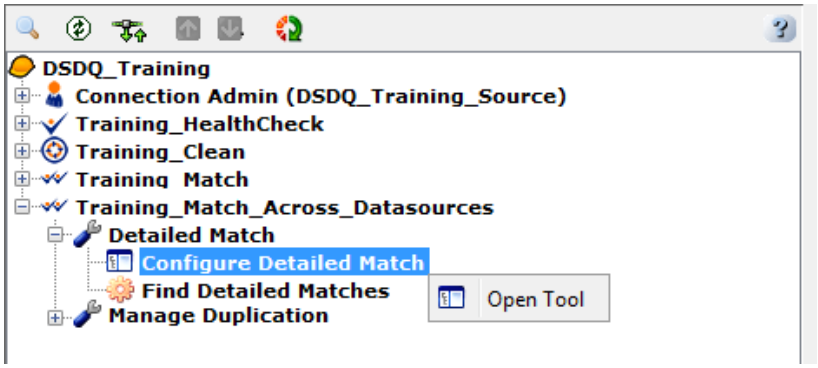
> **Note**
>
> Only one Service Level can be configured at any given time.

To Configure Detailed Match across Data Sources

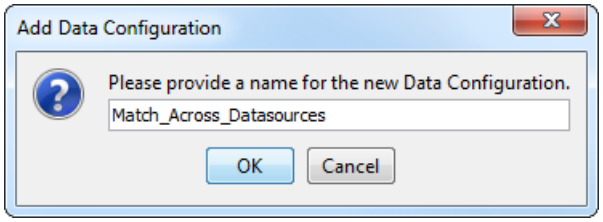
1. Click ⊞ on the DecisionSpace Data Quality Tree to expand the **Training_Match_Across_DataSources** Phase.

2. Click ⊞ to expand the Detailed Match Activity.

3. Double-click the **Configure Detailed Match** Tool or right-click the **Configure Detailed Match** Tool and select **Open Tool** from the pop-up menu.

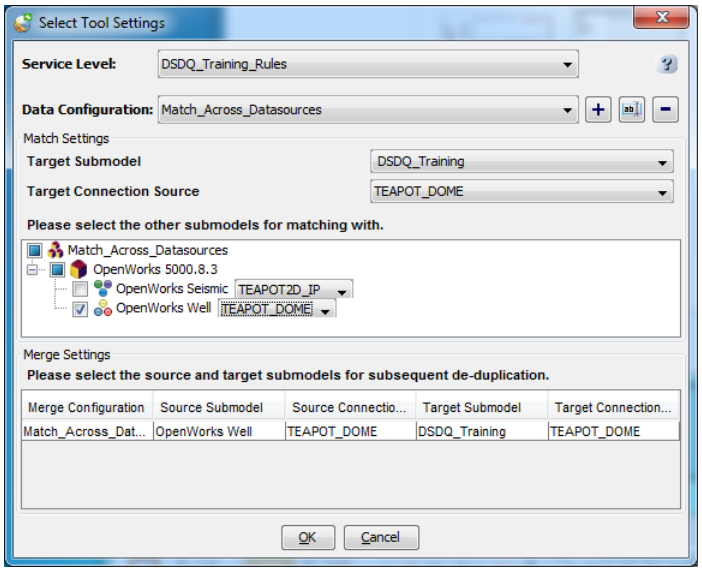


The **Configure Detailed Match** window appears. The **Add Data Configuration** dialog box appears by default the first time you run the **Configure Detailed Match** tool.



4. Enter **Match_Across_Datasources** in the **Please provide a name for the new Data Configuration** field.

5. Click **OK**.
   The **Select Tool Settings** window appears with **DSDQ_Training_Test** added to the **Searchable Submodel** Tree.



6. Select **DSDQ_Training_Rules** from the **Service Level** drop-down list.

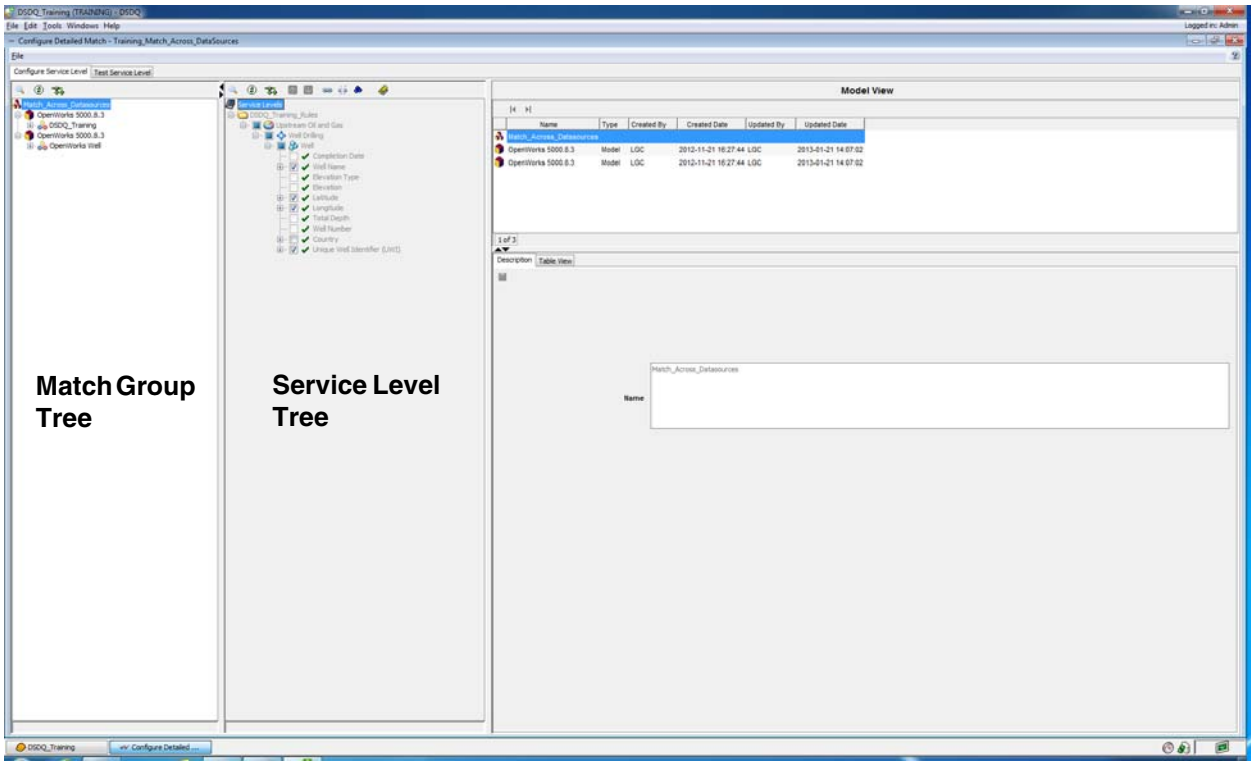
7. Select **Match_Across_Datasources** from the **Data Configuration** drop-down list.

8. Select **DSDQ_Training** from the **Target Submodel** drop-down list.

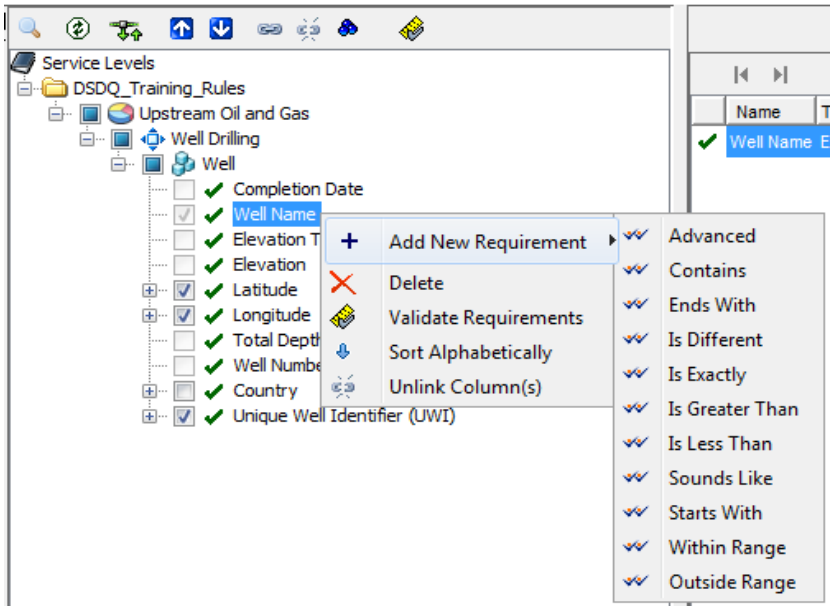
9. Select **TEAPOT_DOME** from the **Target Connection Source** drop-down list.

10. Select the **OpenWorks Well** check box from the **Searchable Submodel** Tree and select **TEAPOT_DOME2** from the drop-down list.

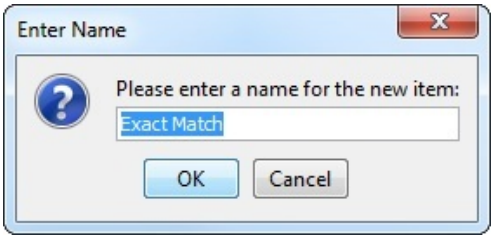
11. Click **OK**.
    The **Configure Detailed Match** window appears with **DSDQ_Training** added to the **Match Group** Tree.



12. Right-click the **Well Name** element in the Service Level Tree and select **Add New Requirement > Is Exactly** from the pop-up menu.



The **Enter Name** dialog box appears.



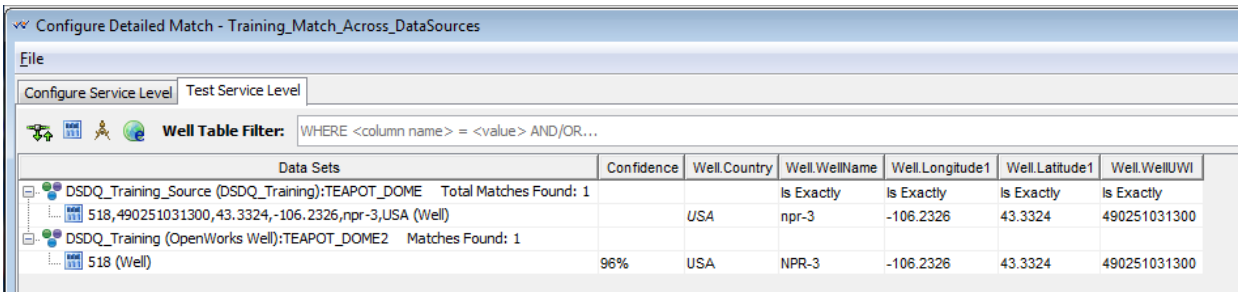
13. Optionally, specify a user-defined name for the requirement.

14. Click **OK** to add the requirement to the selected element.

15. Optionally, repeat Steps 11 through 14 to add all elements for matching. For better results, it is recommended that you use the following:

    - Well Name
    - Latitude
    - Longitude
    - UWI

16. Click the **Test Service Level** tab.
    The test is automatically executed for the first record of the test data subset.



17. Verify all entries have correct matches.

18. Click the **Next Data Set** button to test the next record.

19. Repeat step 19 to test all records.

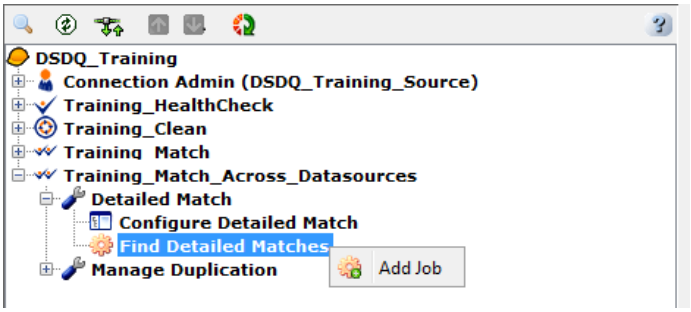
20. Select **File > Exit** from the menu bar on the **Configure Detailed** window.

Exercise: Finding Detailed Matches across Data Sources

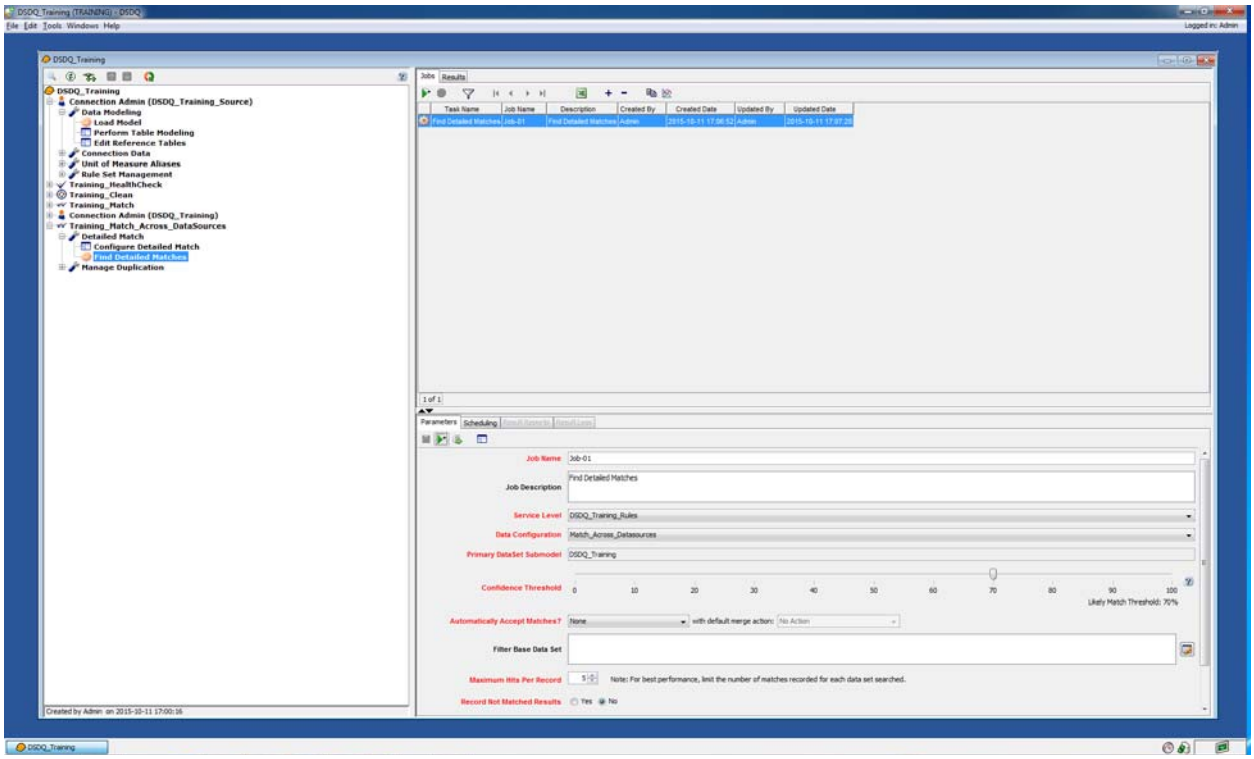
The **Find Detailed Matches** task looks for matching values within the specified columns using the associated match requirements.

To find Detailed Matches across Data Sources:

1. Double-click the **Find Detailed Matches** Task or right-click the **Find Detailed Matches** task and select **Add Job** from the pop-up menu.



A new job is initiated and displays on the **Jobs and Results Listing Pane**.



2. Enter **Job-01** in the **Job Name** field.

3. Enter **Find Detailed Matches** in the **Job Description** field.

4. Select **DSDQ_Training_Rules** from the **Service Level** drop-down list.

5. Select **Match_Across_Datasources** from the **Data Configuration** drop-down list.
   The **Primary DataSet Submodel** field populates automatically.

6. Set the **Confidence Threshold** option as **70**.

7. Select **None** from the **Automatically Accept Matches** drop-down list.

8. Optionally, set a filter on the data subset.

9. Set the **Maximum Hits Per Record** option as **20**.

10. Select the **No** option for **Record Not Matched Results**.

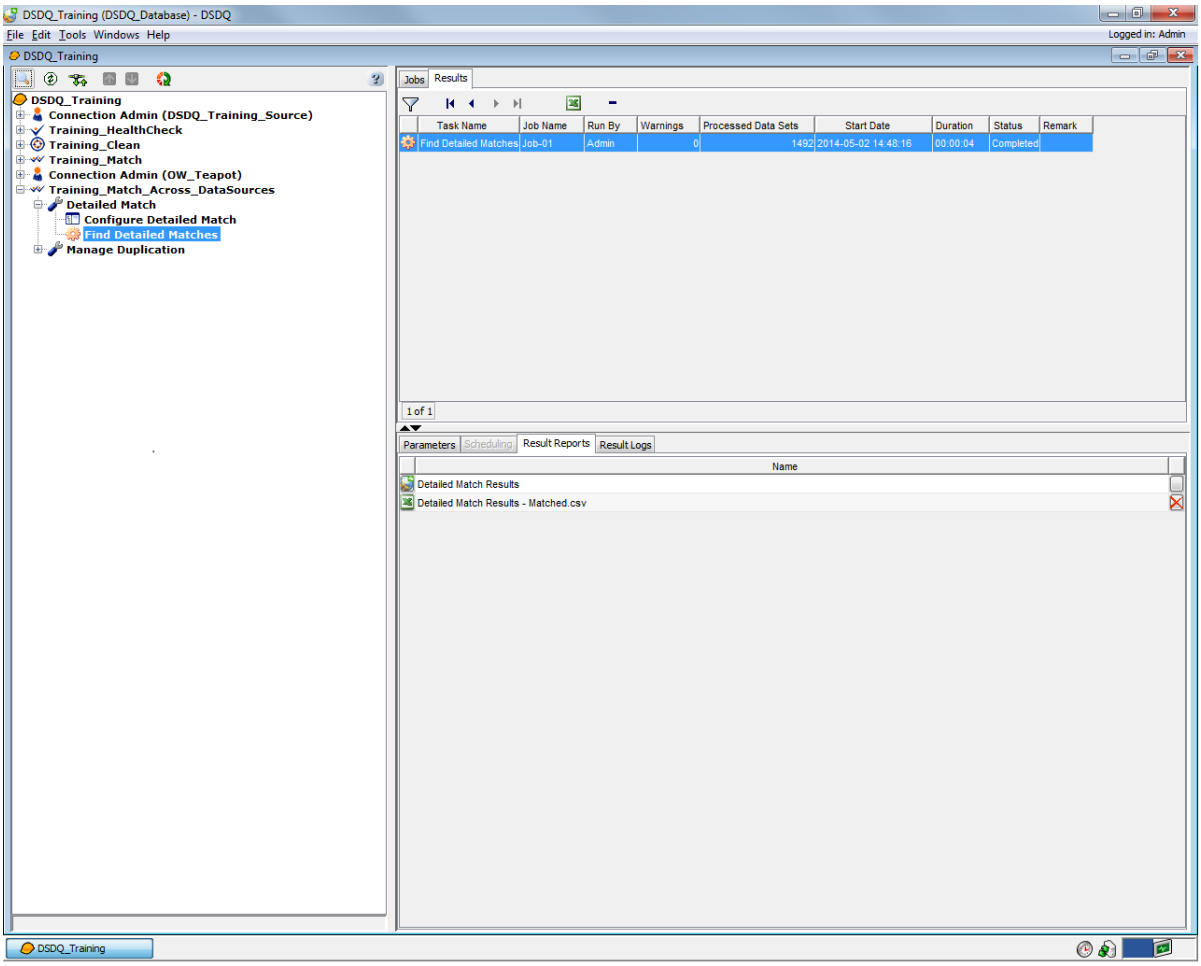
11. Select the **Yes** option for **Generate Printable Reports?**

12. Select the **After** option for **Delete Results?** Set the number of days as **7**.

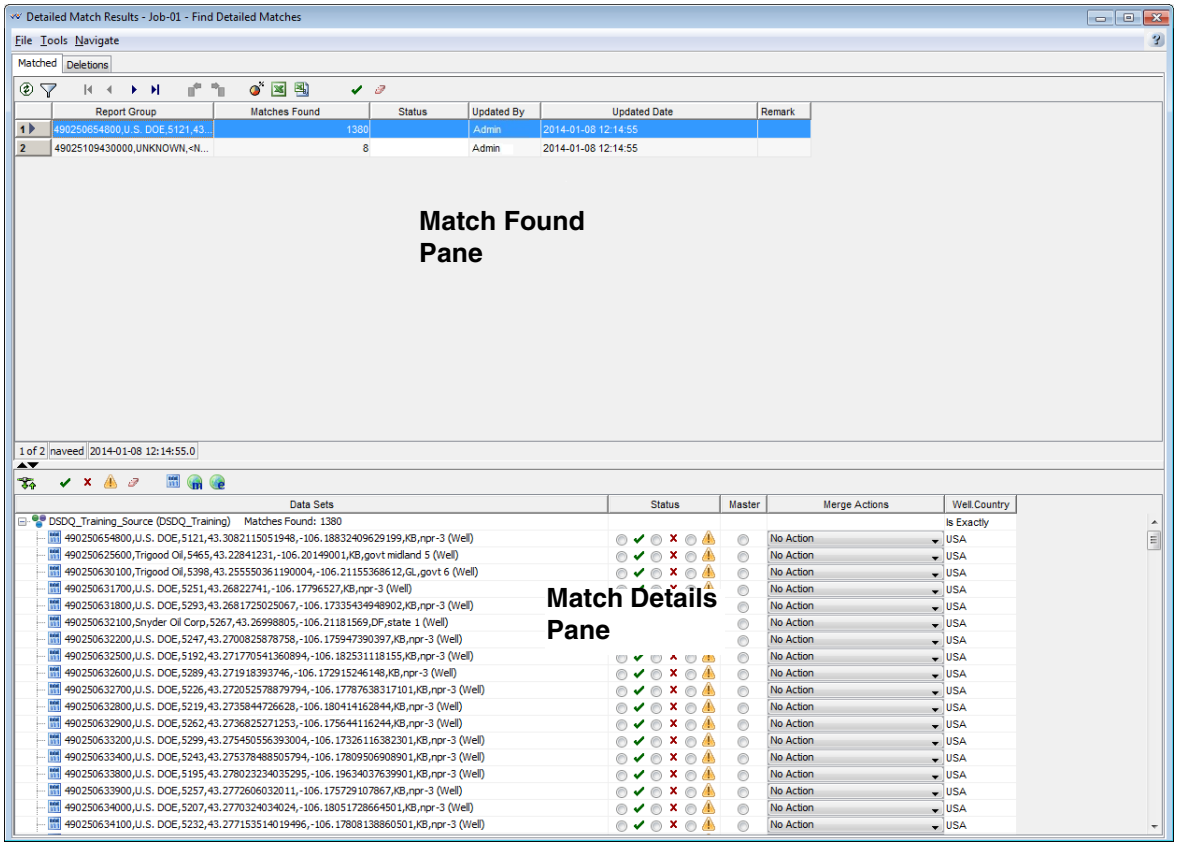
13. Click to save changes in the **Parameters** tab.

14. Click .
    The **Find Detailed Matches** task runs and displays results in the **Result Reports** tab on the **Jobs and Results Information Pane**.

15. Select the **Results** tab on the **Jobs and Results Listing Pane** to view the values in the **Result Reports** tab on the **Job and Results Information Pane**.



16. Double-click **Detailed Match Results** on the **Result Reports** tab. The **Detailed Match Results** window appears.



17. Select all the rows in the **Matches Details** Pane and click ✔ on the **Matches Found** toolbar.

18. Optionally, select the data and click the **Show on Browser Map** button on the toolbar to display the location of the data set on a map in the **DecisionSpace Data Quality Dashboard**. You can also display the data in a Pie chart by click the **Generate Status Chart** button.

19. Select **File > Exit** from the menu bar on the **Detailed Match Results** window.

Manage Duplication across Data Sources

The Manage Duplication across Data Sources Activity helps in removing duplication from multiple data sources. Submodels are configured to make the process straightforward while meeting standards of duplication removal.

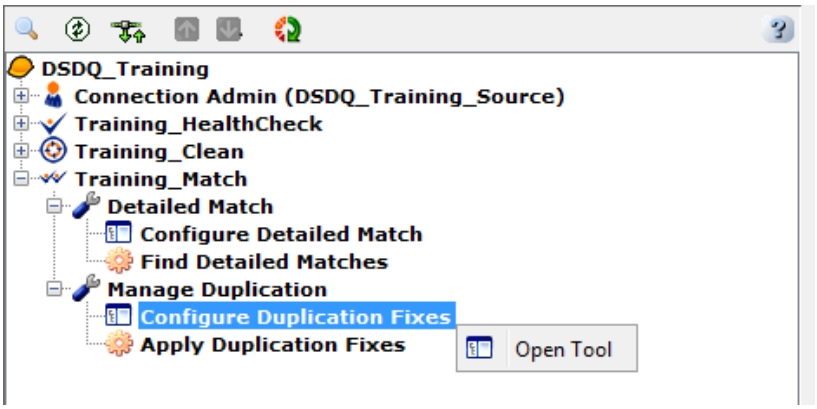
Exercise: Configuring Duplication Fixes across Data Sources

The **Configure Duplication Fixes** Tool enables you to set up table and column mappings for the selected **Merge Group**. The subset of data can then be selected for testing before running it on the actual data set. The interface also enables users to apply rules in order to modify data before moving it to the target database.

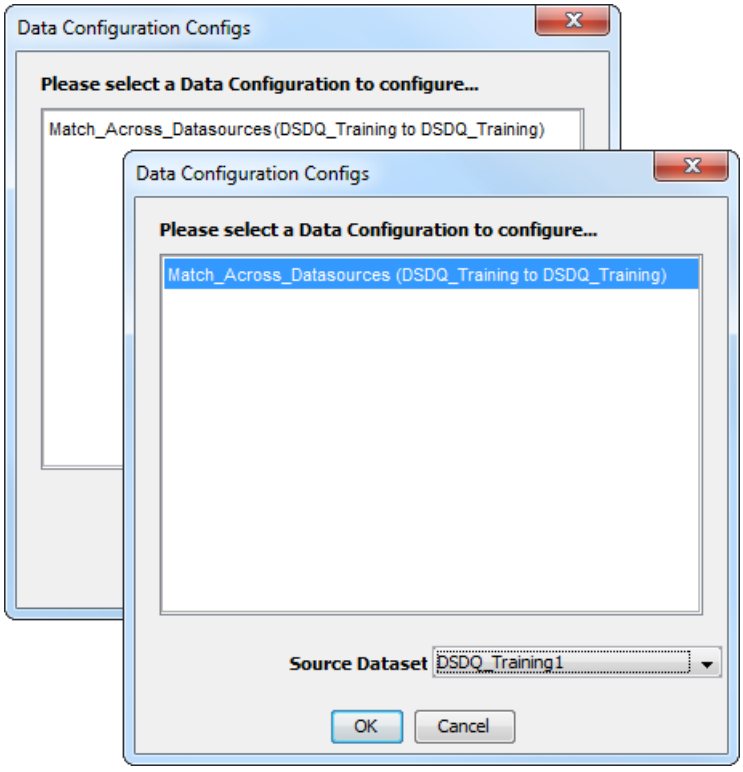
To Configure Duplication Fixes across Data Sources:

1. Click ⊞ to expand the **Manage Duplication** Activity.

2. Double-click the **Configure Duplication Fixes** Tool or right-click the **Configure Duplication Fixes Tool** and select **Open Tool** from the pop-up menu.



The **Data Configuration Configs** window appears.

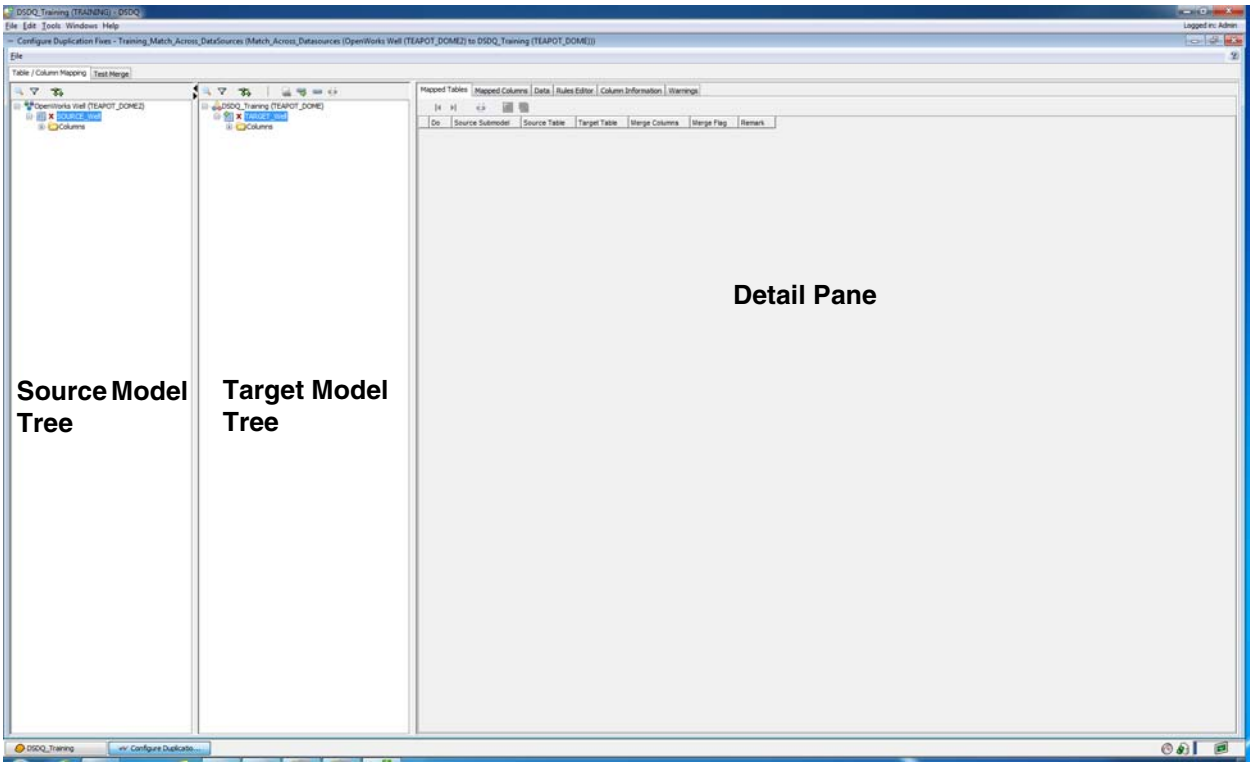


3. Select the **Match_Across_Datasources (DSDQ_Training to DSDQ_Training)** option.

4. Click **OK**.
   The **Configure Duplication Fixes** window displays, displaying tables and columns for the Source Model Tree, Target Model Tree and the Detail pane. The Detail Pane has six tabs:

Mapped Tables	This area display information about the tables that have been mapped in the source and target trees.
Mapped Columns	This area displays the mapping between the target column and all its mapped sources.
Data	This area displays the source and target tables' data. Selecting a column in either the Source or Target tree highlights the corresponding column in the data view if available. If the selected column has already been mapped, the mapped column data is highlighted in each of the corresponding trees. The **Source Data** View and **Target Data** View toolbars can be used to **Filter Data**, to move to the **First Record in Current View** or **Last Record in Current View**, or to move to the **Next Data Set** or **Previous Data Set**. The data can also be saved in Excel format by clicking on the corresponding button or exported to a CSV file by selecting the **Create a CSV Export File** button on the toolbar. Data in the Data Detail Pane can be sorted by clicking a column header. Column information on any column can be viewed by right-clicking a column header and selecting Column Info from the pop-up menu. Other columns information can be viewed by right-clicking a column header and selecting Columns Filter from the pop-menu.
Rules Editor	Used to apply rules to specific columns. Simply drag the rule to the target column that the rule has to be applied to. The **Methods** tab is automatically populated with the relevant information. Make changes to the fields as needed.
Column Information	Displays basic information about the selected column: "Data Type", "Column Size", etc. The tab is divided into two vertical panes: the left one holds the source column information, and the right one holds target column information.

Warnings	This area will display any inconsistencies between the mapped columns, e.g. source column length is greater than the target column length. Initially this tab is blank. When the first warning is logged, the tab name turns red and a warning icon appears next to its name.



5. Select the **Source_Well** table from the Source Model Tree.

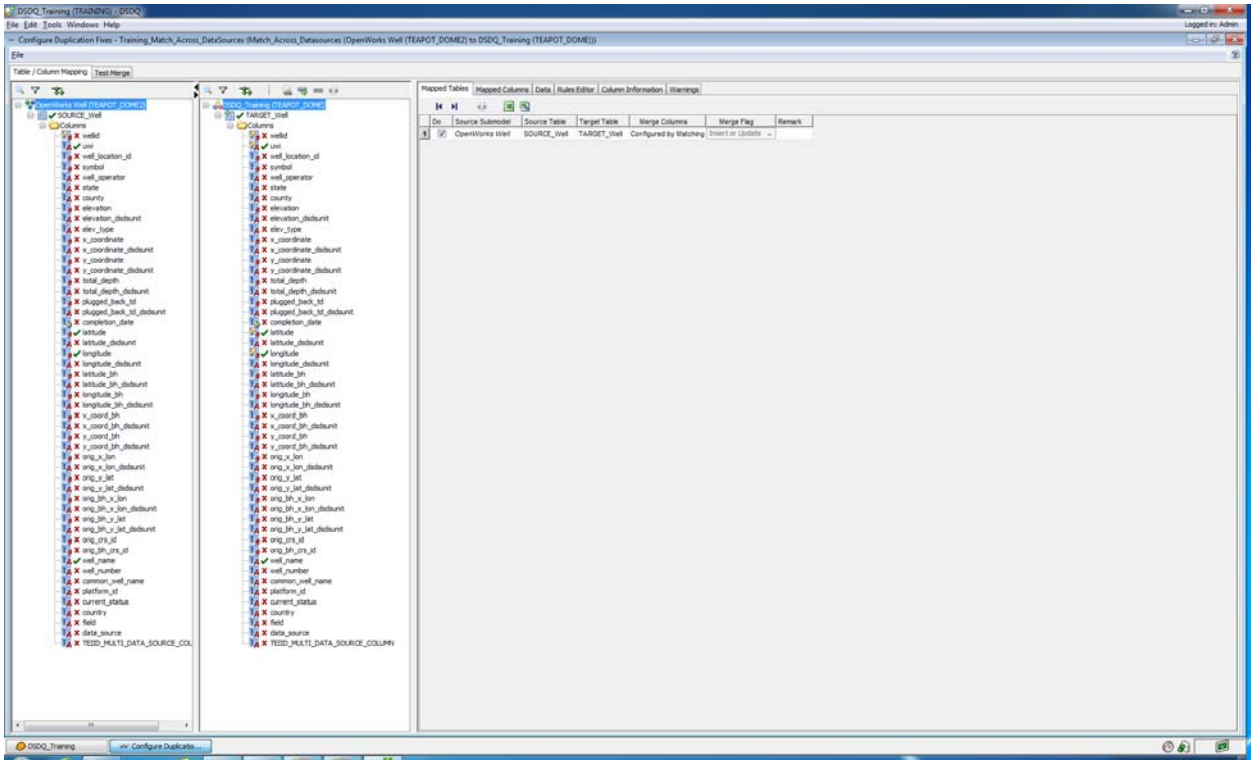
6. Select the **Target_Well** table from the Target Model Tree.

7. Click the **Auto Map selected Table/Columns** button on the toolbar.
   A green check mark appears adjacent to the selected tables.

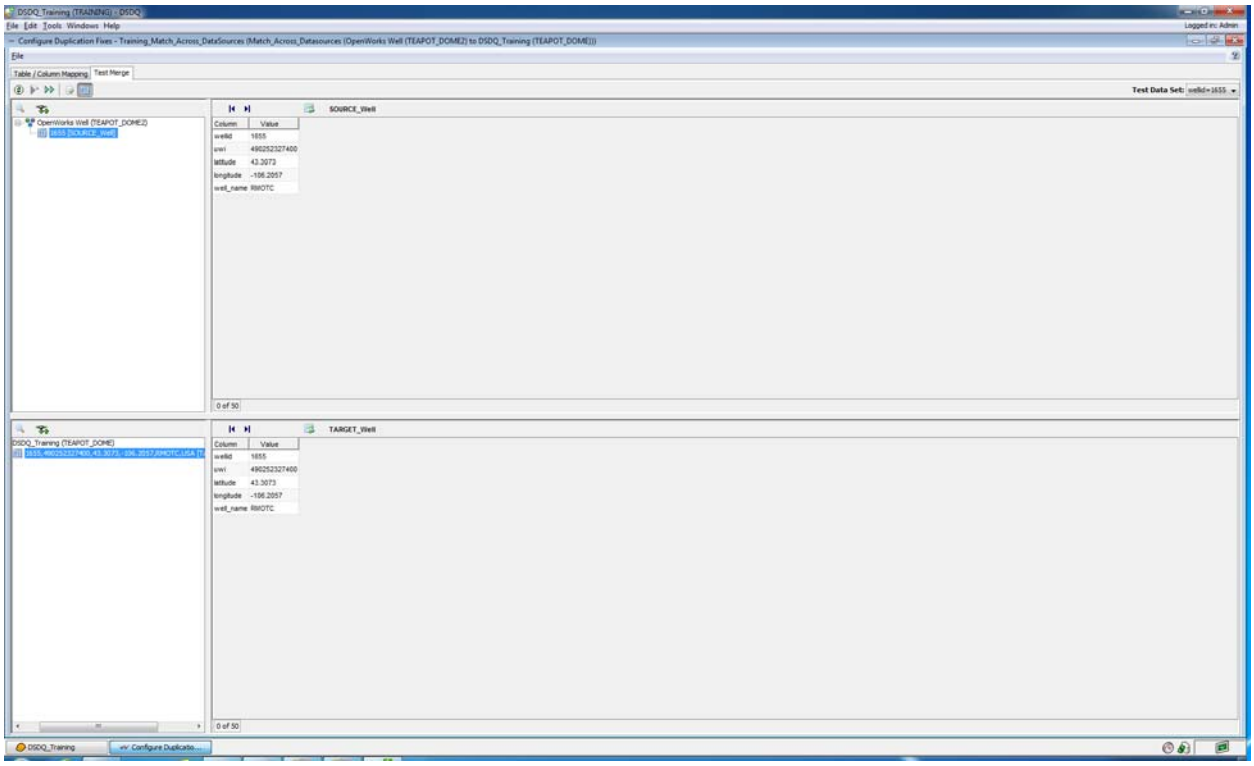
8. Select the following columns from the Source_Well & Target_Well table:

   - UWI
   - Latitude
   - Longitude
   - Well Name

9. Click the **Auto Map selected Table/Columns** button on the toolbar.
   A green check mark appears adjacent to the columns in both the Source Model and Target Model Tree.



10. Select the **Test Merge** tab to test all match result configurations.



11. Select **File > Exit** from the menu bar on the **Configure Duplication Fixes** window.

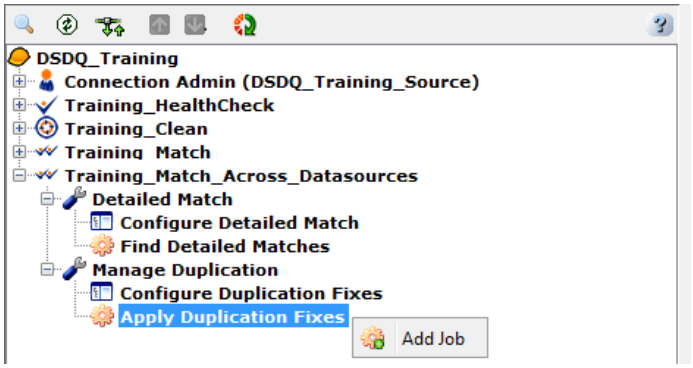
Exercise: Applying Duplication Fixes across Data Sources

After the duplications have been tested in the **Configure Duplication Fixes** Tool, the **Manage Duplication** Activity and **Apply Duplication Fixes** Task is used to move the complete data set over to the target database.

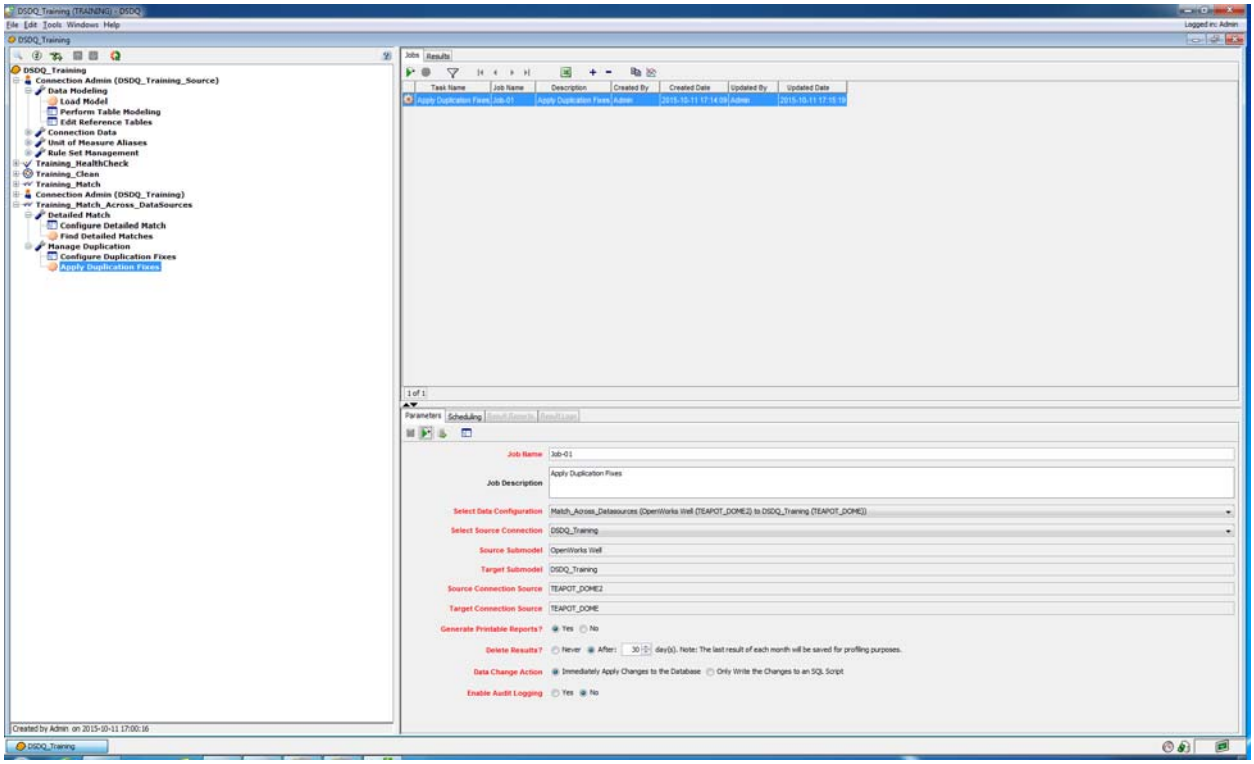
To Apply Duplication Fixes across Data Sources:

1. Double click the **Apply Duplication Fixes** Task or right-click the **Apply Duplication Fixes** task and select **Add Job** from the pop-up

menu.



A new job is initiated and displays on the **Jobs and Results Listing Pane**.



2. Enter **Job-01** in the **Job Name** field.

3. Enter **Apply Duplication Fixes** in the **Job Description** field.

4. Select **Match_Across_Datasources (OpenWorks Well (TEAPOT_DOME2) to DSDQ_Training (TEAPOT_DOME))** from the **Select Data Configuration** drop-down list.

5. Select **DSDQ_Training** from the **Select Source Connection** drop-down list.

The **Source Submodel** and **Target Submodel** fields populates automatically.

6. Select the **Yes** option for **Generate Printable Reports?**

7. Select the **After** option for **Delete Results?** Set the number of days as **7**.

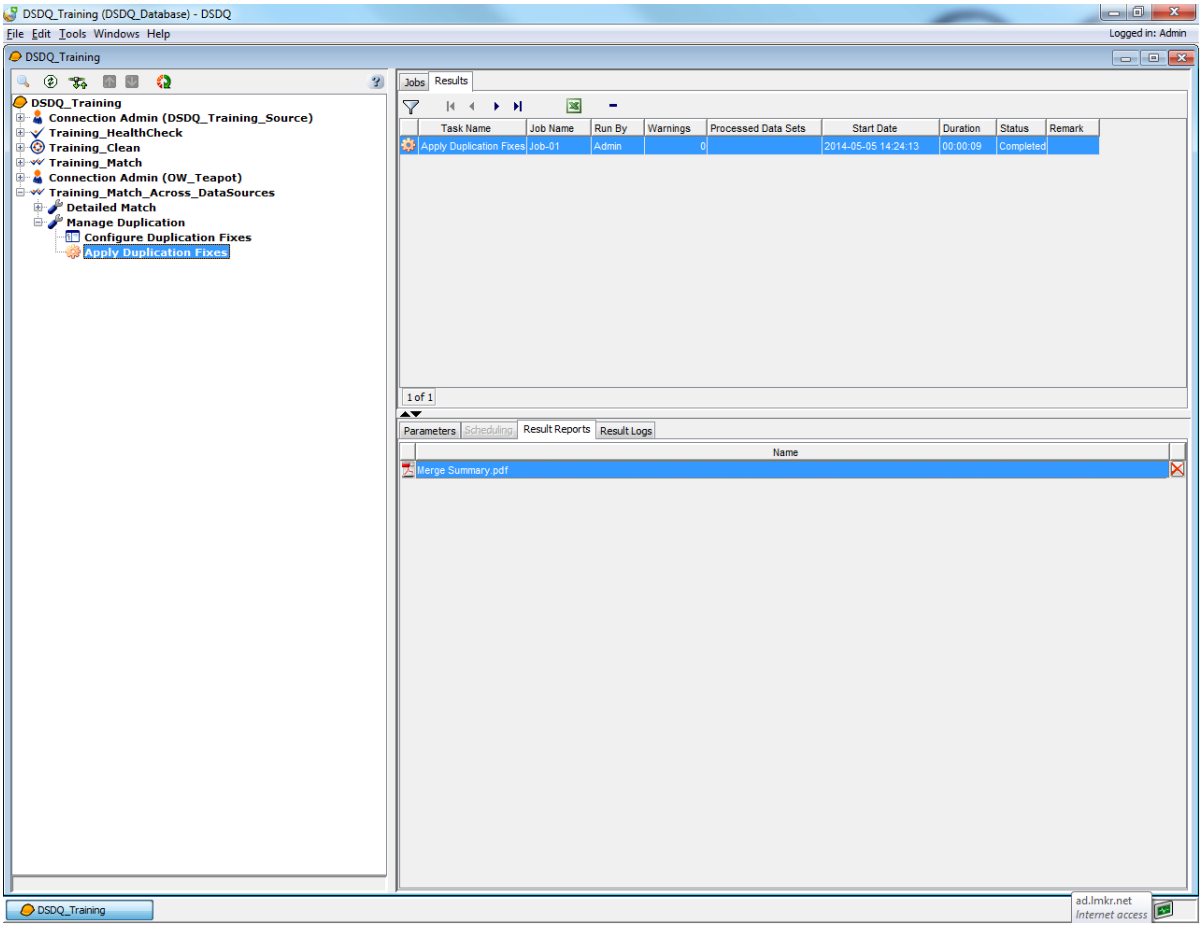
8. Select the **Immediately Apply Changes to the Database** option for **Data Change Action**.

9. Select the **No** option for **Enable Audit Logging**.

10. Click to save changes in the **Parameter** tab.

11. Click .
    The **Apply Duplication Fixes** Task is executed and displays results in the **Result Reports** tab.

12. Select the **Results** tab on the **Jobs and Results Listing Pane** to view the values in the **Result Reports** tab on the **Job and Results Information Pane**.



13. Click on the **Results Reports** tab to display Duplication Fixes results in PDF format.