DAT218x

Cleansing Data with Data Quality Services

Lab 2-2 | Cleansing Data with Integration Services

Estimated time to complete this lab is 45 minutes

Overview

In this lab, you will cleanse the Office dataset with Integration Services (SSIS) by using the knowledge base enhanced in **Lab 2-1**.

The labs in this course are accumulative. You cannot complete the following labs if this lab has not been successfully completed.

This document is provided "as-is". Information and views expressed in this document, including URL and other Internet Web site references, may change without notice. This document does not provide you with any legal rights to any intellectual property in any Microsoft product. You may copy and use this document for your internal, reference purposes. © 2016 Microsoft. All rights reserved.

Exercise 1: Connecting to the VM

Go to the next exercise if you are already connected to the lab VM.

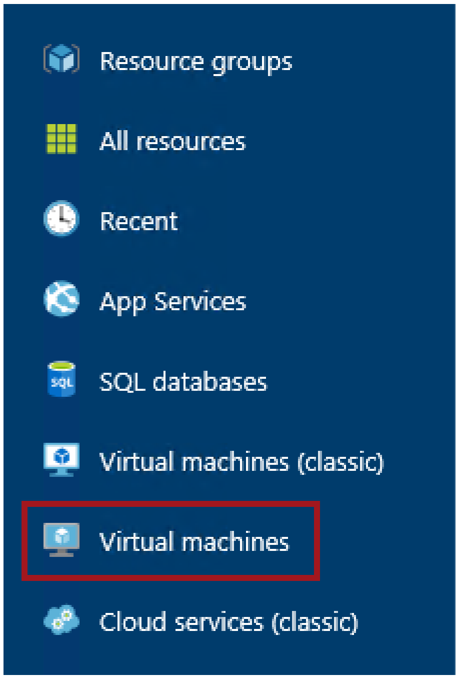
In this exercise, having signed in to the Azure Portal by using your Azure subscription, you will connect to the lab VM which you provisioned in **Lab 0-1**.

Connecting to the VM

In this task, you will sign in to the Azure Portal, and then connect to your lab VM.

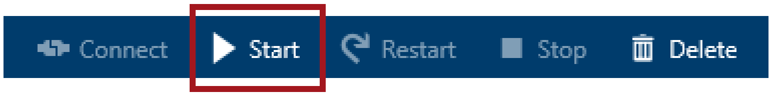
1. Sign in to the **Azure Portal** by using your subscription.

2. In the left pane, select **Virtual Machines**.



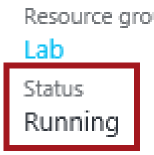
3. In the **Virtual Machines** blade, select the VM you provisioned in **Lab 0-1**.

4. In the VM blade, click **Start**.



5. Wait for the VM status to update to **Running**.

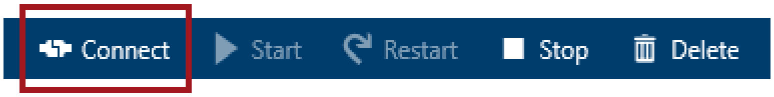
   It usually takes 1-2 minutes for the VM to start.



Resource group
Lab
Status
Running

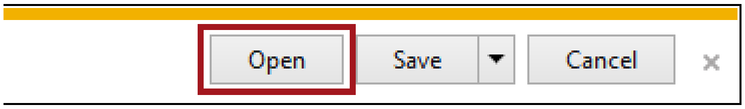
6. To connect to the VM, click **Connect**.

   Take care not to use the RDP file downloaded the previous time. It is likely that a different IP address has be assigned.



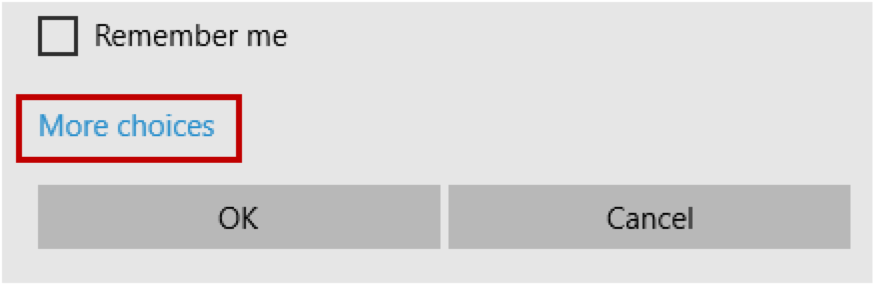
   This file can be used to reconnect to the remote desktop session, but note that when you deallocate the VM and later re-start the VM, it will be likely that a different IP address will be assigned.

7. When prompted by the web browser to open the Remote Desktop File, click **Open**.



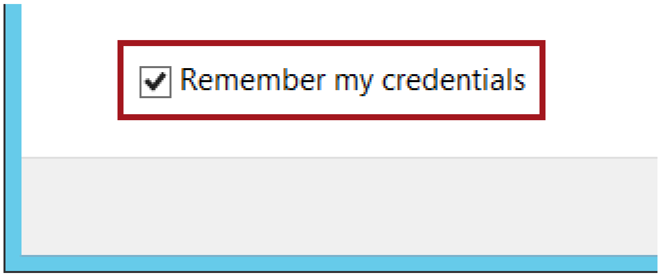
8. If prompted to connect to the unknown publisher, click **Connect**.

   *To enter your credentials, you may need to select **More Choices**, and then select **Use a Different Account**.*



9. In the **Windows Security** window, enter the credentials you created for your VM.

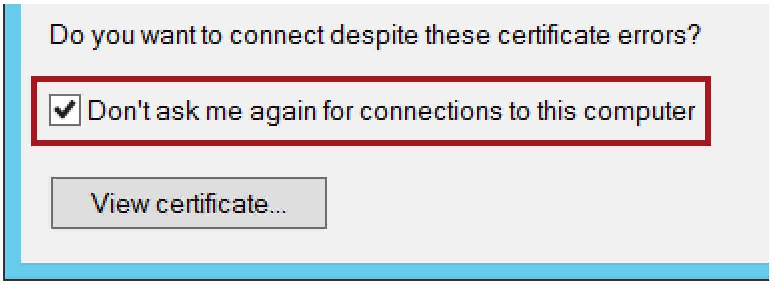
10. Check the **Remember My Credentials** checkbox.



☑ Remember my credentials

11. Click **OK**.

12. In the **Remote Desktop Connection** window, check the **Don’t Ask Me Again for Connections to This Computer** checkbox.



Do you want to connect despite these certificate errors?

☑ Don't ask me again for connections to this computer

13. Click **Yes**.

14. If you have a second monitor, maximize the Remote Desktop window inside a single monitor.

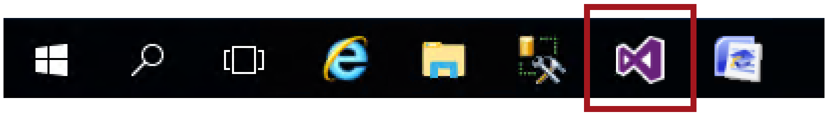
Exercise 2: Cleansing Data with Integration Services

In this exercise, you will cleanse the Office dataset with Integration Services by using the knowledge base enhanced in **Lab 2-1**.

Creating the DQS Connection Manager

In this task, you will open the SSIS project, and then create a DQS connection manager.

1. Open SQL Server Data Tools.



2. To open an existing project, on the **File** menu, select **Open | Project/Solution**.

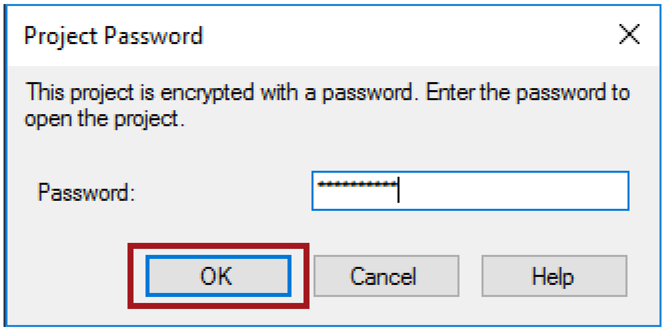
3. In the **Open Project** window, navigate to the **F:\Labs\Lab2-2\Assets\Project** folder.

4. Select **Lab.sln**, and then click **Open**.

   *This is the same project developed in **Lab 1-2**.*

5. In the **Project Password** window, in the **Password** box, enter **Pass@word1**. (Do not enter the period.)

6. Click **OK**.

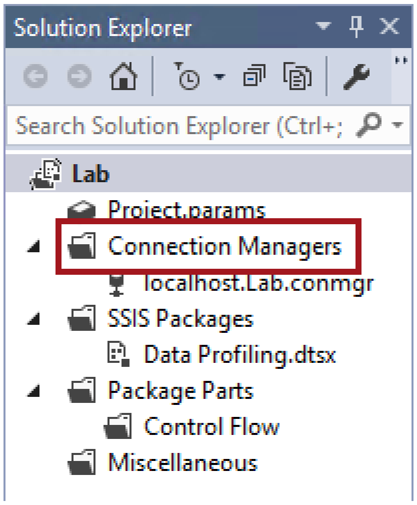


Project Password

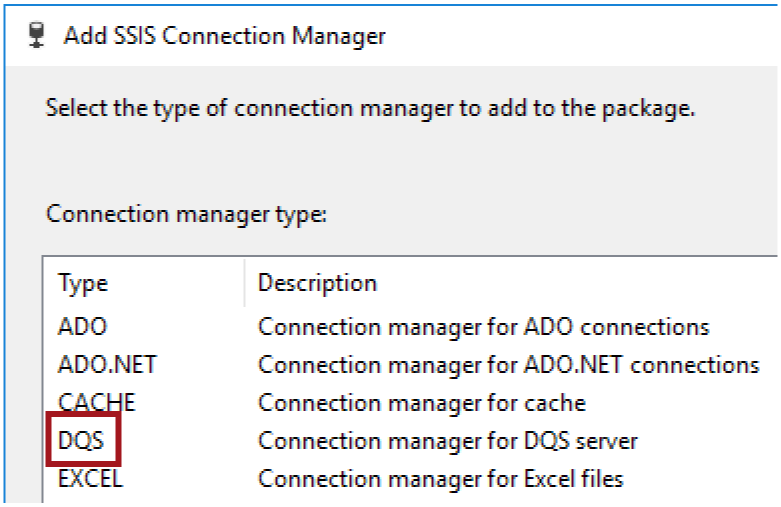
This project is encrypted with a password. Enter the password to open the project.

Password: **********

7. To create an additional connection manager, in **Solution Explorer**, right-click the **Connection Managers** folder, and then select **New Connection Manager**.



8. In the **Add SSIS Connection Manager** window, select the **DQS** connection manager type.

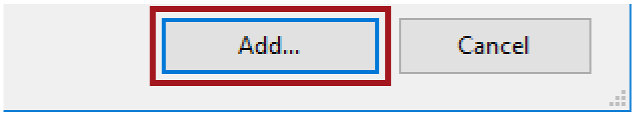


Select the type of connection manager to add to the package.

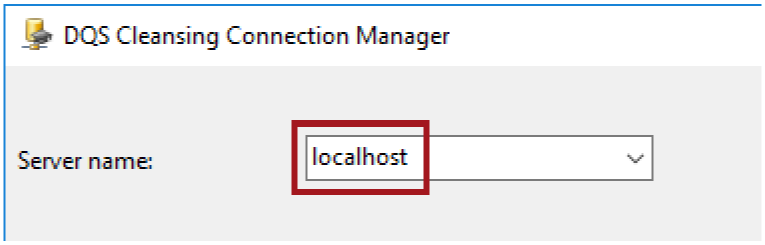
Connection manager type:

Type	Description
ADO	Connection manager for ADO connections
ADO.NET	Connection manager for ADO.NET connections
CACHE	Connection manager for cache
DQS	Connection manager for DQS server
EXCEL	Connection manager for Excel files

9. Click **Add**.

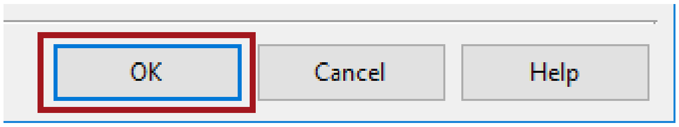


10. In the **Add DQS Cleansing Connection Manager** window, in the **Server Name** dropdown list—do not click the dropdown arrow—enter **localhost**.

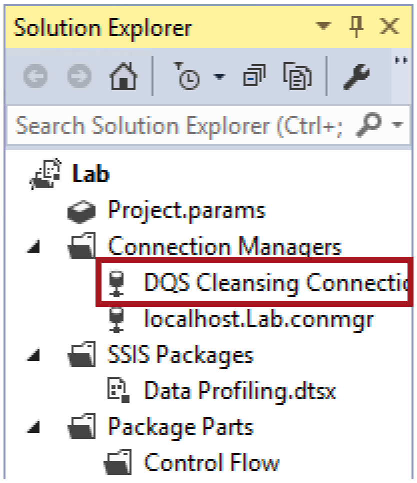


Server name: localhost

11. Click **OK**.



12. In **Solution Explorer**, notice the addition of the connection manager.



Creating the Package

In this task, you will create a package designed to load cleansed office records into a table which represents a data warehouse dimension table. Records that cannot be cleansed will be loaded to an alternate table that can be analyzed by a data steward.

1. In **Solution Explorer**, right-click the **SSIS Packages** folder, and then select **New SSIS Package**.

2. Notice that the package designer opens automatically.

3. To rename the package, in **Solution Explorer**, right-click the **Package1.dtsx** file, and then select **Rename**.

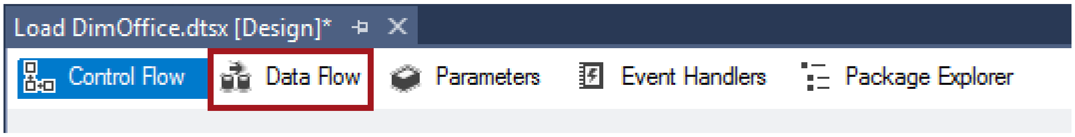
4. Rename the package to **Load DimOffice.dtsx**, and then press **Enter**.

Developing the Data Flow

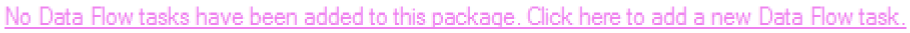
In this task, you will develop a data flow to extract, transform and load (ETL) the office dataset. The transform process will cleanse the data by using the Office knowledge base.

The output of the cleansing will be split into correct and invalid outputs. Correct data will be loaded into the **DimOffice** table, and invalid data will be loaded into the **DimOffice_Error** table.

1. Select the **Data Flow** tab.

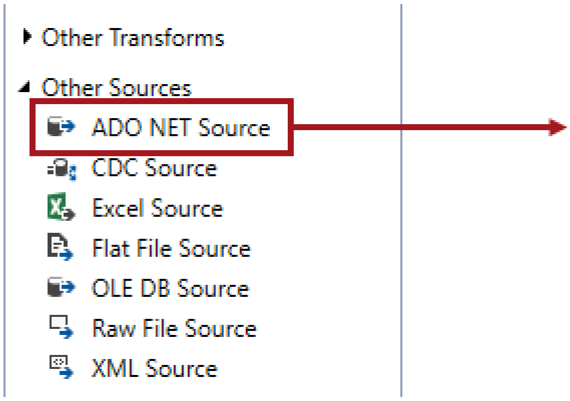


2. To add a data flow task, click the link located at the center of the designer.

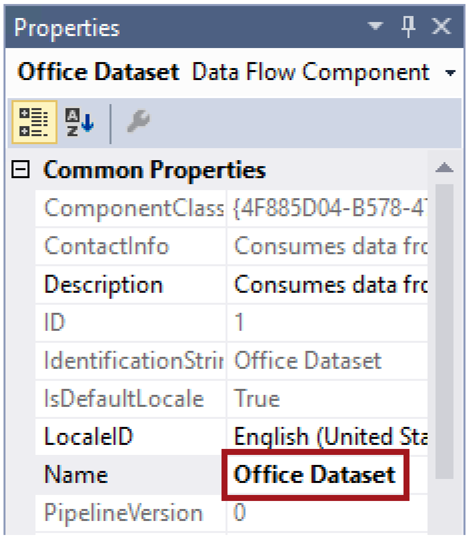


No Data Flow tasks have been added to this package. Click here to add a new Data Flow task.

3. To design the data flow, from the **SSIS Toolbox** (located at the left), expand **Other Sources**, and then drag the **ADO NET Source** to the data flow designer.



4. In the **Properties** pane, set the **Name** property to **Office Dataset**.



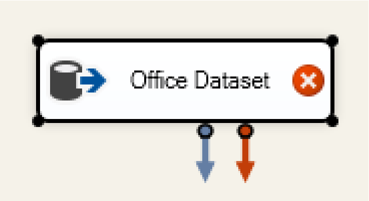
Properties

Office Dataset Data Flow Component

Common Properties

ComponentClass	{4F885D04-B578-4
ContactInfo	Consumes data frc
Description	Consumes data frc
ID	1
IdentificationStrir	Office Dataset
IsDefaultLocale	True
LocaleID	English (United Sta
Name	**Office Dataset**
PipelineVersion	0

5. Verify that the data flow component looks like the following.



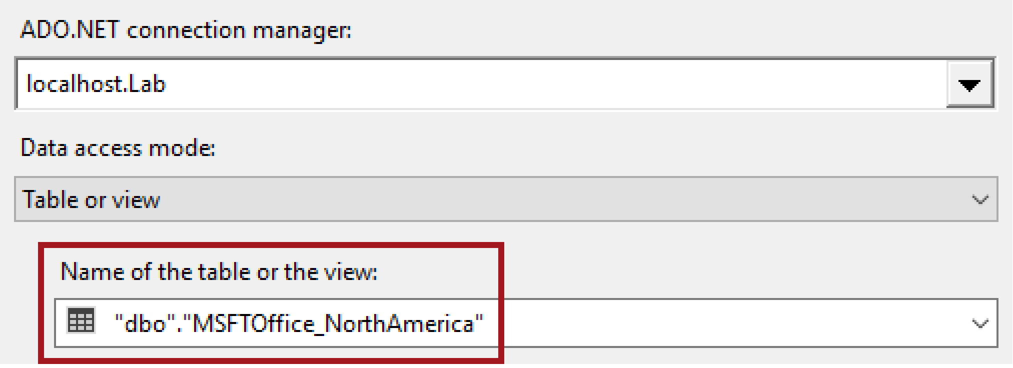
Do not be concerned about the error icon, which will disappear when you complete the next steps.

6. To edit the source component, right-click the component, and then select **Edit**.

7. In the **ADO.NET Source Editor** window, in the **ADO.NET Connection Manager** dropdown list, notice that the **localhost.Lab** connection manager is selected.

8. In the **Name of the Table or the View** dropdown list, select **"dbo"."MSFTOffice_NorthAmerica"**.

   *This is the same dataset you used to data profile in **Lab 1-1**, perform knowledge discovery with in **Lab 1-2**, and source data in the Data Quality Project in **Lab 2-1**.*

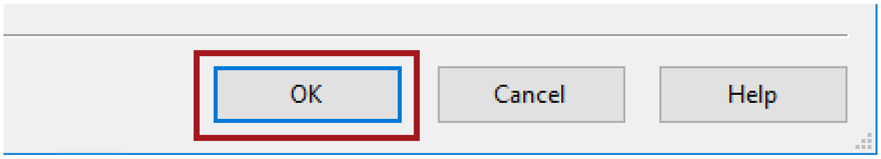


ADO.NET connection manager:
localhost.Lab

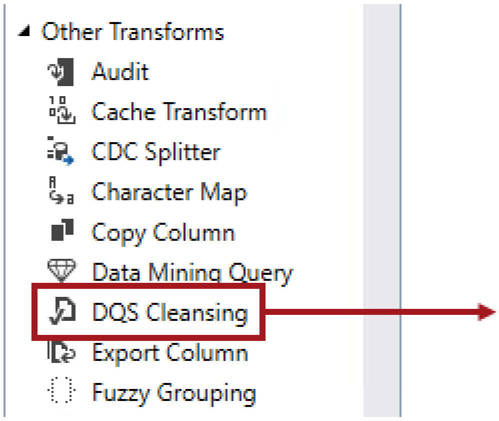
Data access mode:
Table or view

Name of the table or the view:
"dbo"."MSFTOffice_NorthAmerica"

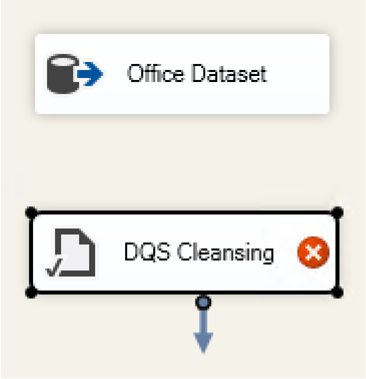
9. Click **OK**.



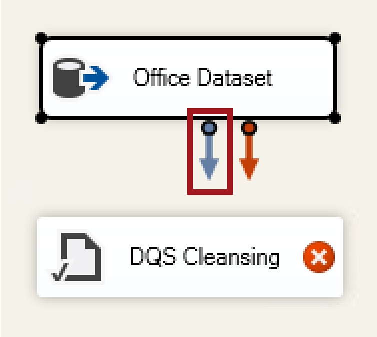
10. From the **SSIS Toolbox**, expand **Other Transforms**, and then drag the **DQS Cleansing** to the data flow designer, and drop it directly beneath the source component.



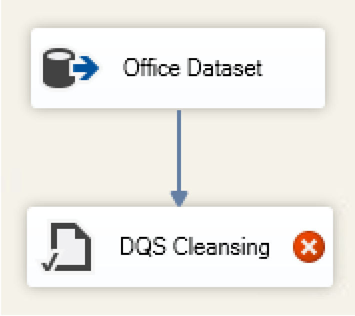
11. Verify that the data flow design looks like the following.



12. To connect the components, first select the **Office Dataset** source component, and then drag the standard output (the left, blue arrow) on top of the cleansing component.



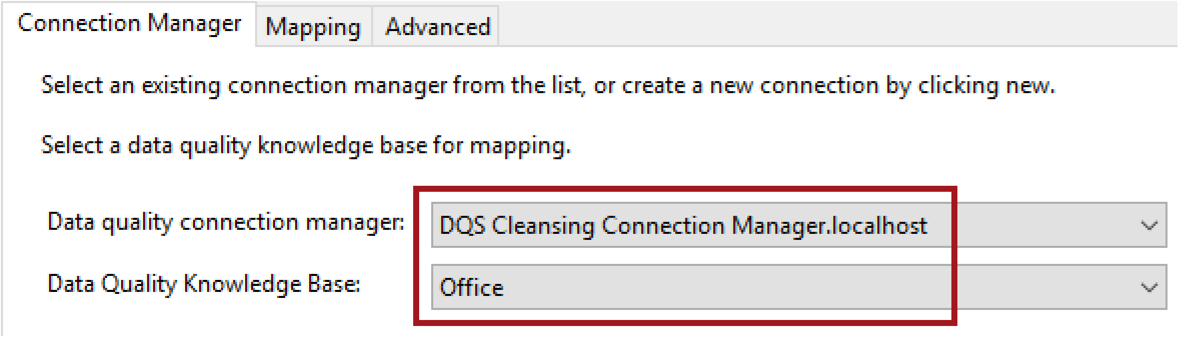
13. Verify that the data flow design looks like the following.



14. To edit the cleansing component, right-click the component, and then select **Edit**.

15. In the **DQS Cleansing Transform Editor** window, in the **Data Quality Connection Manager** dropdown list, select the **DQS Cleansing Connection Manager.localhost** connection manager.

16. In the **Data Quality Knowledge Base** dropdown list, select **Office**.



Select an existing connection manager from the list, or create a new connection by clicking new.

Select a data quality knowledge base for mapping.

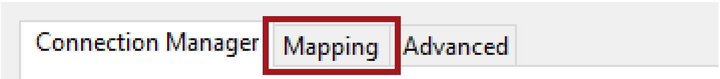
Data quality connection manager: DQS Cleansing Connection Manager.localhost

Data Quality Knowledge Base: Office

17. In the **Available Domains** list, review the knowledge base domains, noticing that the first listed in the composite domain.

You will not use the composite domain to cleanse that data in this package design.

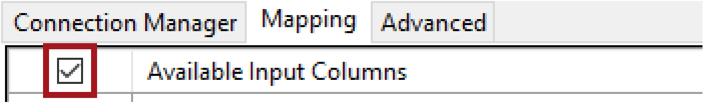
18. Select the **Mapping** tab.



19. Notice the **Available Input Columns** grid.

This grid lists of input columns received from the source component.

20. To select all input columns, check the top-right checkbox.

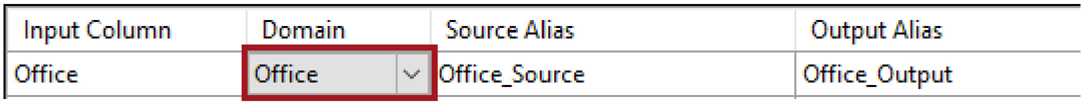


☑	Available Input Columns

21. Notice the second grid that defines the mapping between input columns and the knowledge base domains.

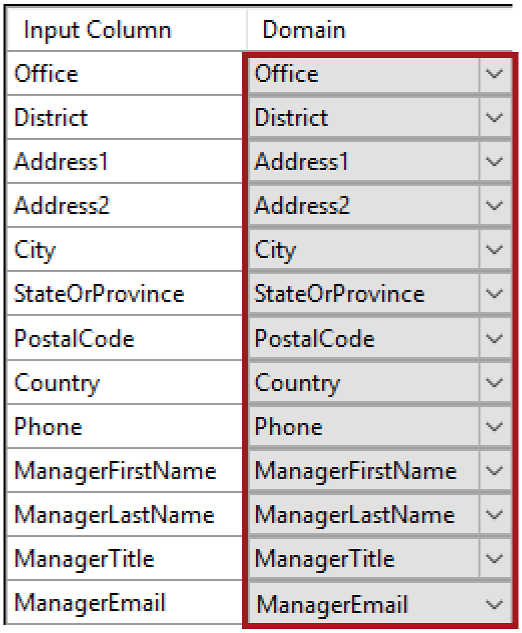
It also defines alias output columns for the source, output and status columns.

22. Set the **Office** input column to map to the **Office** domain.



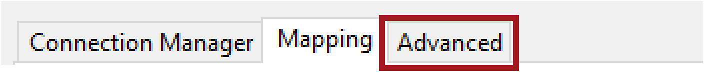
Input Column	Domain	Source Alias	Output Alias
Office	Office	Office_Source	Office_Output

23. Map each input column to its respective domain—do not map the **Address** composite domain.



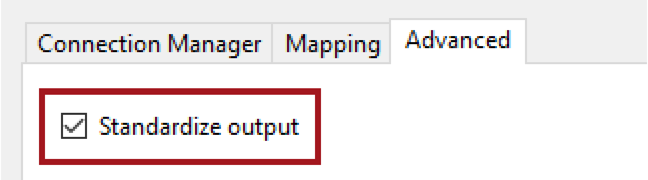
Input Column	Domain
Office	Office
District	District
Address1	Address1
Address2	Address2
City	City
StateOrProvince	StateOrProvince
PostalCode	PostalCode
Country	Country
Phone	Phone
ManagerFirstName	ManagerFirstName
ManagerLastName	ManagerLastName
ManagerTitle	ManagerTitle
ManagerEmail	ManagerEmail

24. Select the **Advanced** tab.



25. Review the available options.

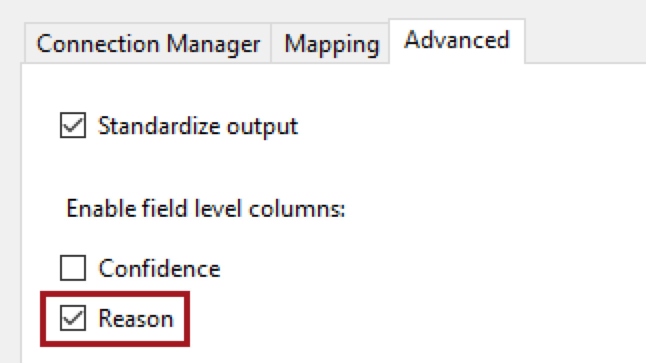
26. Notice that the **Standardize Output** checkbox is selected.



☑ Standardize output

*For your knowledge base, this will mean that **StateOrProvince** values will be set to upper case, and **ManagerEmail** values will be set to lower case.*

27. Check the **Reason** checkbox.



☑ Standardize output

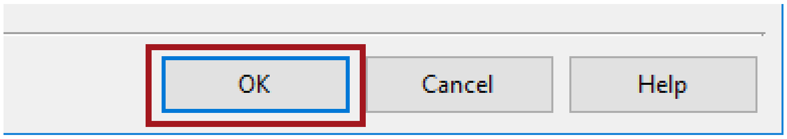
Enable field level columns:

☐ Confidence

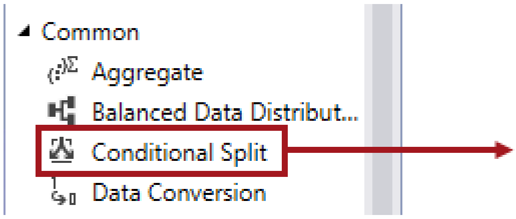
☑ Reason

The reason needs to be output to help explain why values are invalid.

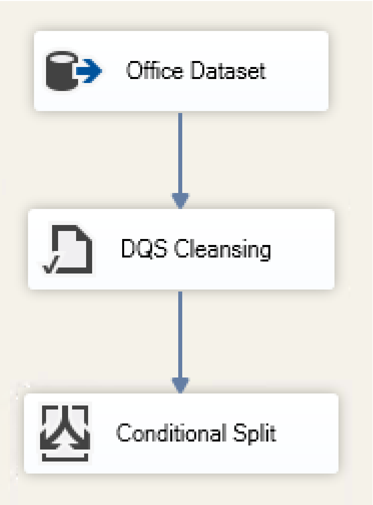
28. To complete the component configuration, click **OK**.



29. From the **SSIS Toolbox**, from inside the **Common** group, drag the **Conditional Split** to the data flow designer, and drop it directly beneath the cleansing component.

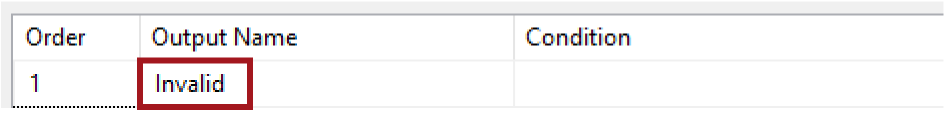


30. Configure the standard output of the cleansing component to connect to the new component, as follows.



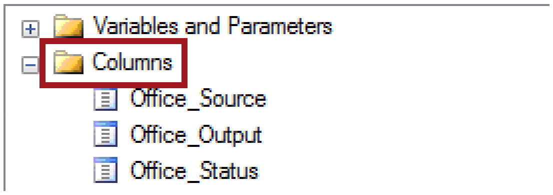
31. Edit the conditional split component.

32. In the grid, in the **Output Name** box, enter **Invalid**.

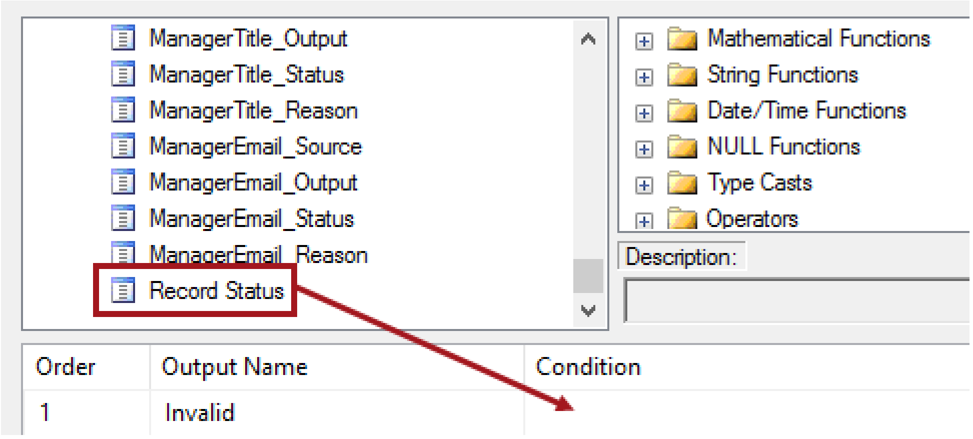


Order	Output Name	Condition
1	Invalid	

33. In the top-right pane, expand the **Columns** folder.



34. Scroll to the bottom of the columns list, and then drag the **Record Status** column into the **Condition** box.

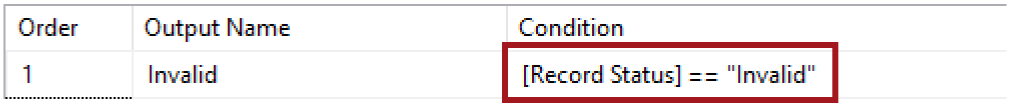


Order	Output Name	Condition
1	Invalid	

35. In the **Condition** box, complete the expression as follows (note that the operator is two equals (=) signs, which tests for equality).

```
[Record Status] == "Invalid"
```

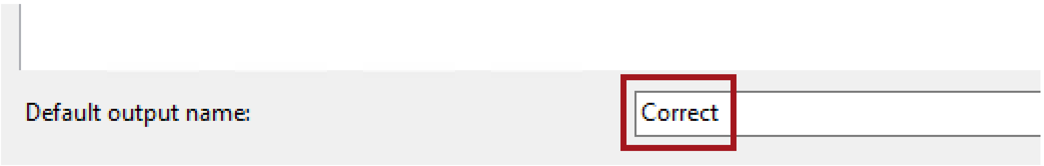
36. Verify that the expression looks like the following.



Order	Output Name	Condition
1	Invalid	[Record Status] == "Invalid"

Any record with an invalid record status will be output to the ***Invalid*** *output.*

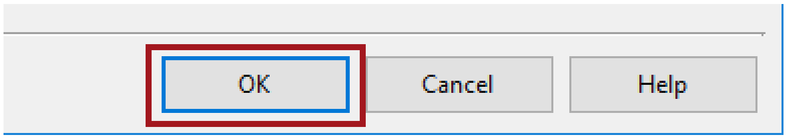
37. In the **Default Output Name** box, replace the text with **Correct**.



Default output name: Correct

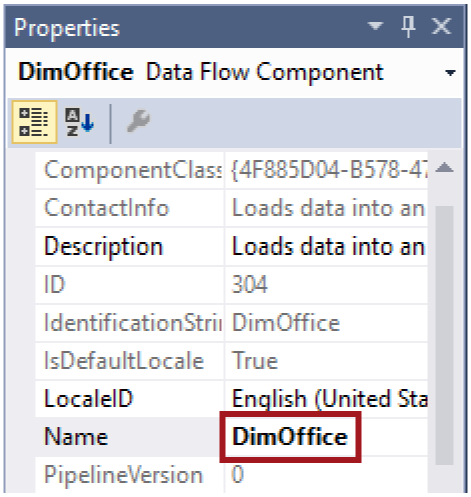
All remaining records will be output to the ***Correct*** *output.*

38. To complete the component configuration, click **OK**.



39. From the **SSIS Toolbox**, expand **Other Destinations** (the last group), and then drag the **ADO NET Destination** to the data flow designer, and drop it beneath, and to the left of, the conditional split component

40. In the **Properties** pane, set the **Name** property to **DimOffice**.



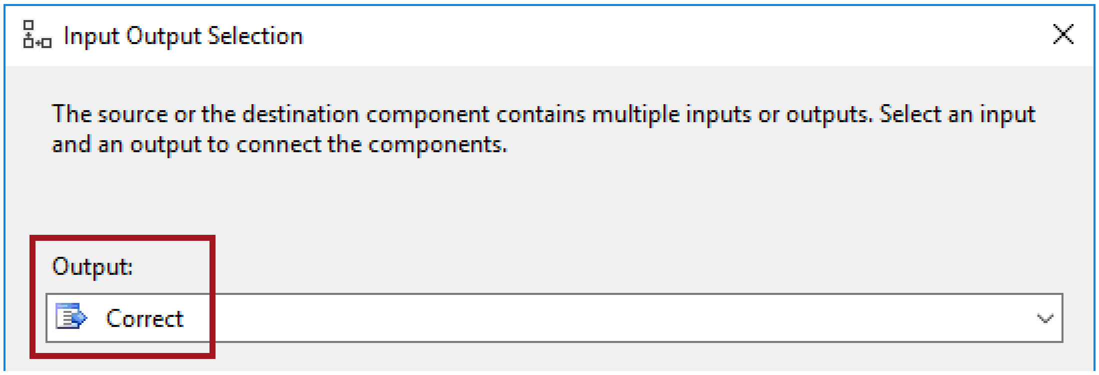
Properties

DimOffice Data Flow Component

ComponentClass	{4F885D04-B578-47
ContactInfo	Loads data into an
Description	Loads data into an
ID	304
IdentificationStri	DimOffice
IsDefaultLocale	True
LocaleID	English (United Sta
Name	DimOffice
PipelineVersion	0

41. Configure the standard output of the conditional split component to connect to the new component.

42. In the **Input Output Selection** window, in the **Output** dropdown list, select **Correct**.



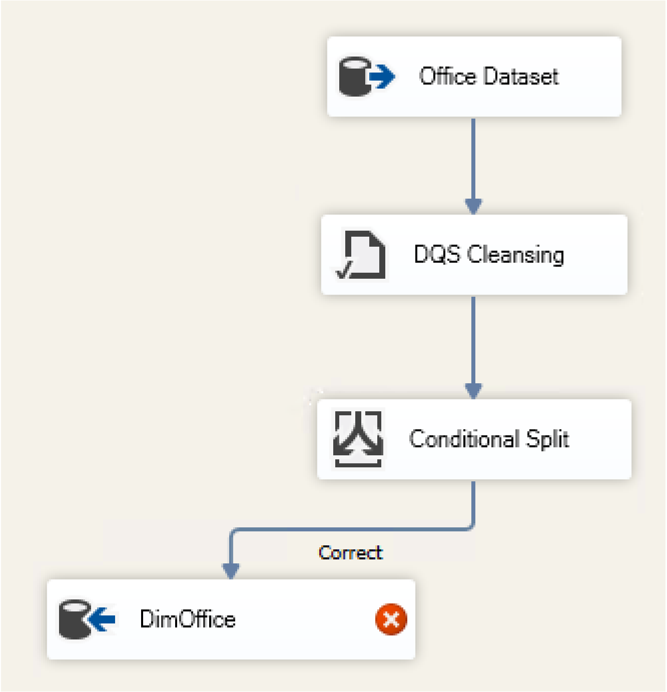
Input Output Selection

The source or the destination component contains multiple inputs or outputs. Select an input and an output to connect the components.

Output: Correct

43. Click **OK**.

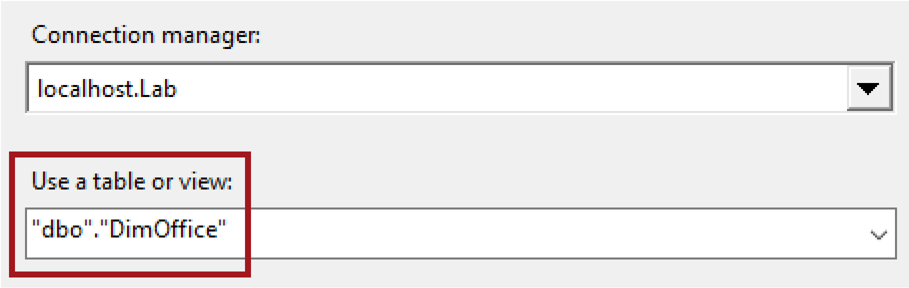
44. Verify that the data flow design looks like the following.



45. Edit the **DimOffice** destination component.

46. In the **ADO.NET Destination Editor** window, in the **Connection Manager** dropdown list, notice that the **localhost.Lab** connection manager is selected.

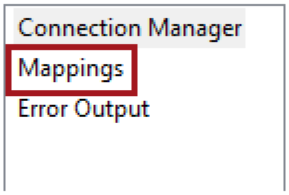
47. In the **Use a Table or View** dropdown list, select **"dbo"."DimOffice"**.



Connection manager:
localhost.Lab

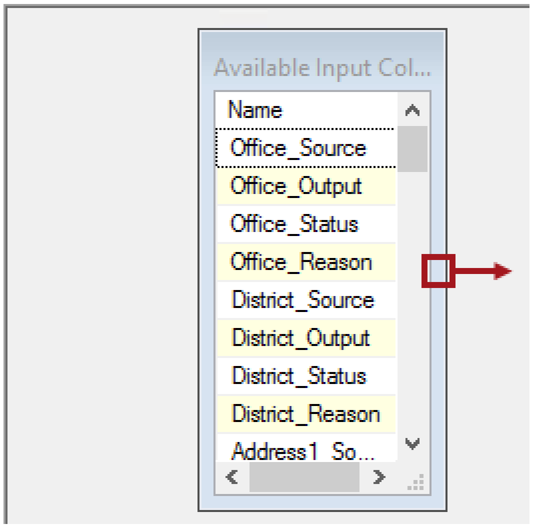
Use a table or view:
"dbo"."DimOffice"

48. In the left pane, select the **Mappings** page.



This page of the editor is used to configure the mappings between the input columns, and the columns of the ***DimOffice*** *table.*

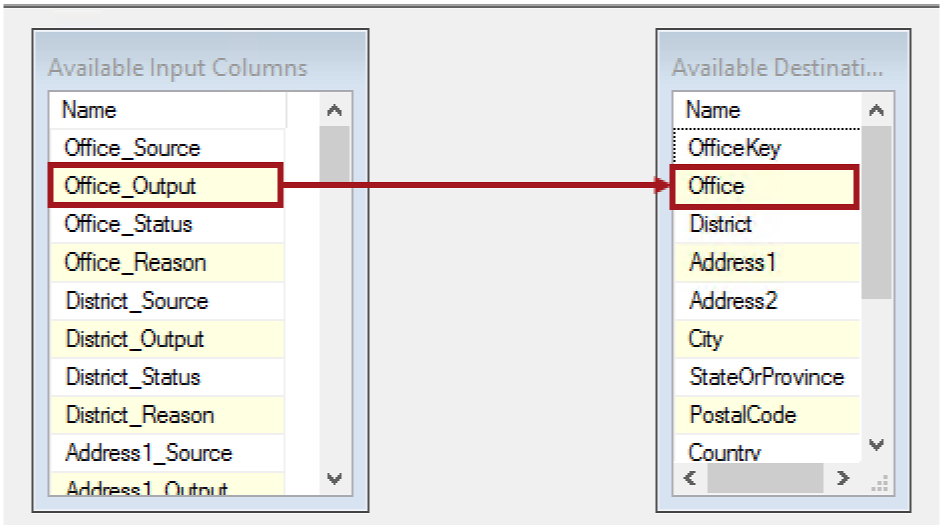
49. To widen the list, drag the right edge of the **Available Input Columns** list, and drag open the **Name** column to reveal the full column names.



Available Input Col...

Name
Office_Source
Office_Output
Office_Status
Office_Reason
District_Source
District_Output
District_Status
District_Reason
Address1_So...

50. From the **Available Input Columns** list, drag the **Office_Output** column to the **Office** columns of the **Available Destination Columns** list.

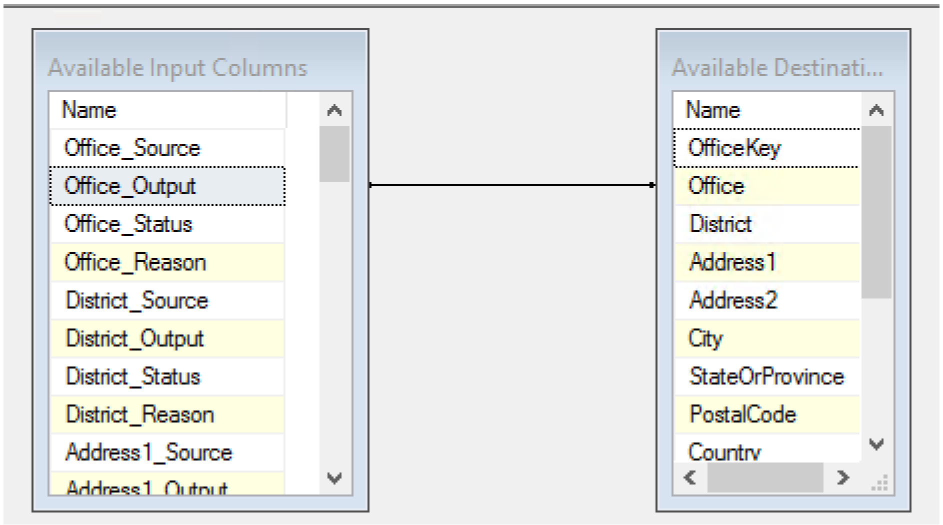


There is no need to map to the ***OfficeKey*** *column, as this is an identity column that will automatically populate a sequence of values when rows are inserted into the table.*

The source columns will contain original values, while the output columns will contain standardized column (i.e. lower case email addresses), so you will map only the output columns.

There is no need to store other column types as the rows passed to this destination are only correct records. Status columns will only ever be ***Correct*** *or* ***Corrected****.*

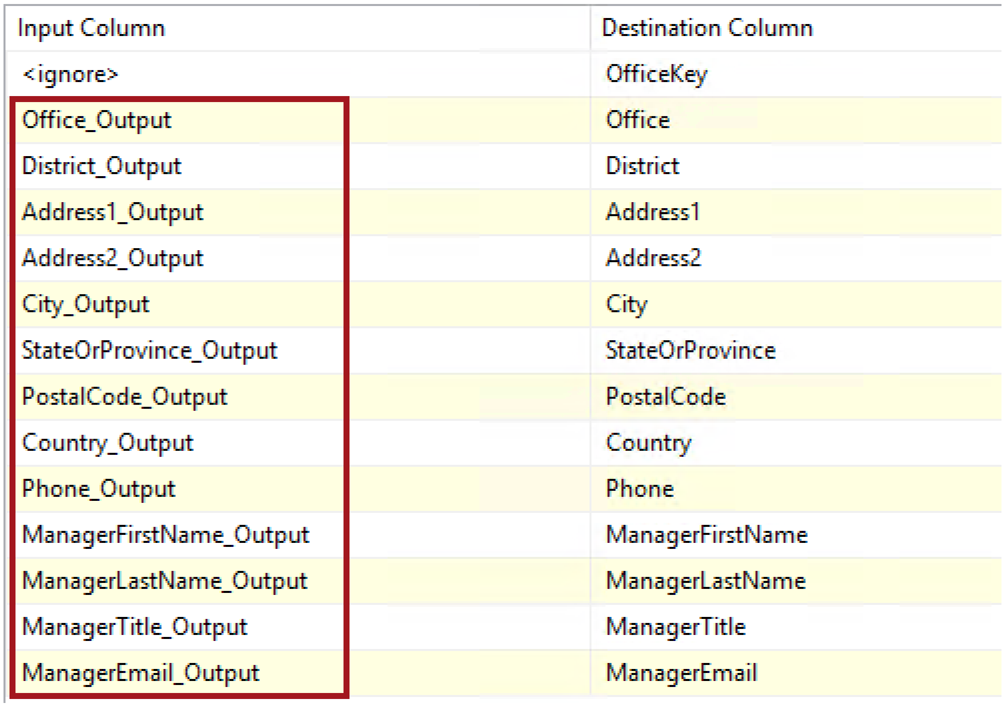
51. Verify that the mapping was created.



52. Map all **"_Output"** columns to the destination columns—except **OfficeKey**.

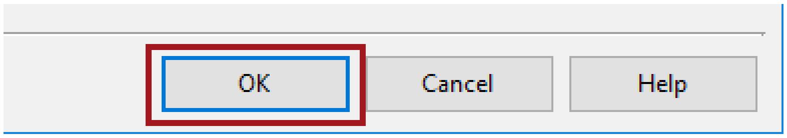
Tip: You can also configure the mappings by selecting the input columns in the lower grid.

53. Verify that all **"_Output"** columns are correctly mapped.



Input Column	Destination Column
<ignore>	OfficeKey
Office_Output	Office
District_Output	District
Address1_Output	Address1
Address2_Output	Address2
City_Output	City
StateOrProvince_Output	StateOrProvince
PostalCode_Output	PostalCode
Country_Output	Country
Phone_Output	Phone
ManagerFirstName_Output	ManagerFirstName
ManagerLastName_Output	ManagerLastName
ManagerTitle_Output	ManagerTitle
ManagerEmail_Output	ManagerEmail

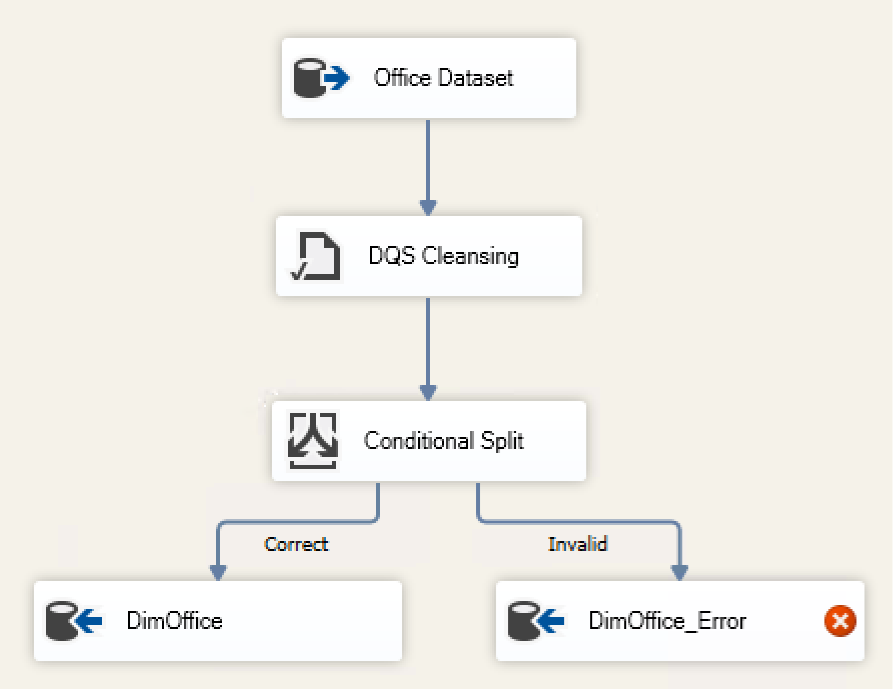
54. To complete the component configuration, click **OK**.



55. Add a second ADO.NET destination component, and then rename it **DimOffice_Error**.

56. Connect the conditional split component to the new destination component.

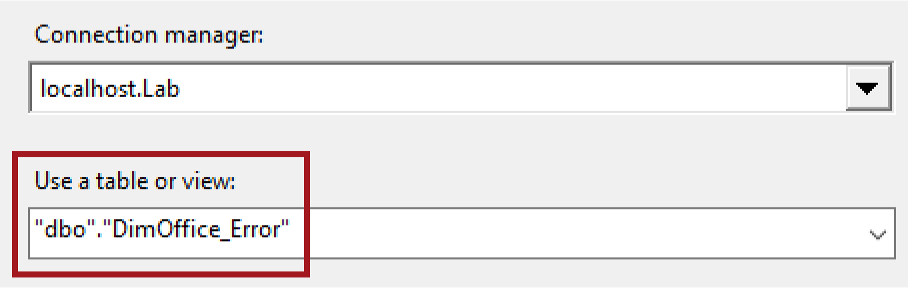
57. Verify that the data flow design looks like the following.



58. Edit the **DimOffice_Error** destination component.

59. In the **ADO.NET Destination Editor** window, in the **Connection Manager** dropdown list, notice that the **localhost.Lab** connection manager is selected.

60. In the **Use a Table or View** dropdown list, select **"dbo"."DimOffice_Error"**.



Connection manager:

localhost.Lab

Use a table or view:

"dbo"."DimOffice_Error"

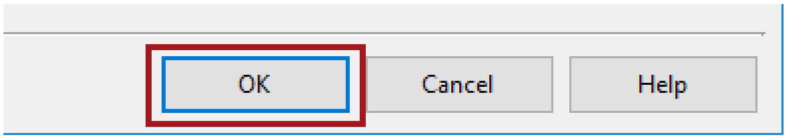
61. Select the **Mappings** page.

62. Notice that the mappings to this table are automatically created.

Mappings are created automatically when there are matching column names and data types between the two tables.

As this table will be used to analyze data quality issues, all output columns will be stored.

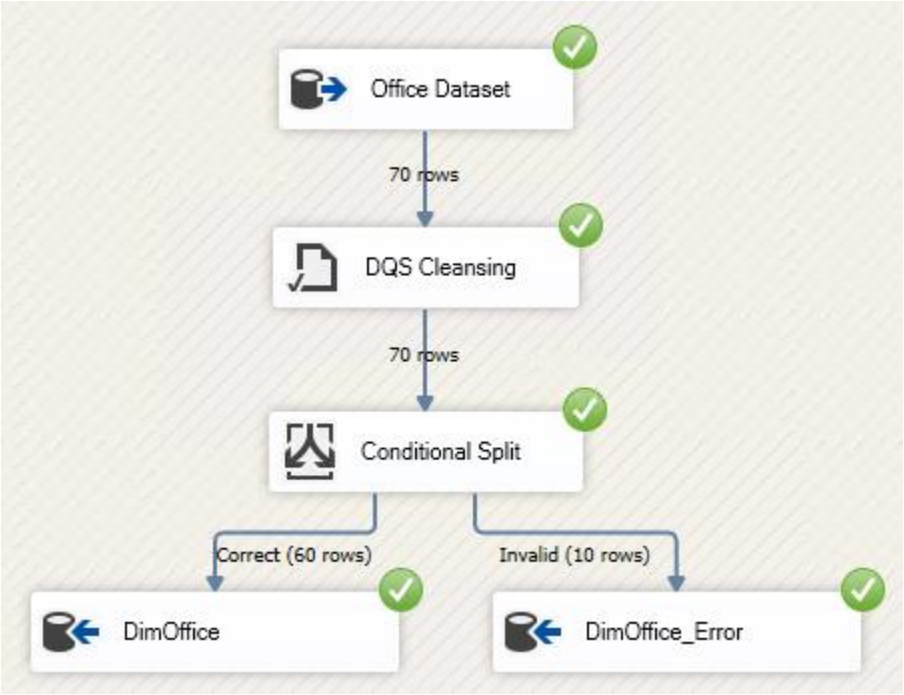
63. To complete the component configuration, click **OK**.



Executing the Package

In this task, you will execute the package and observe the data flow execution.

1. To execute the package, in **Solution Explorer**, right-click the **Load DimOffice.dtsx** package, and then select **Execute Package**.



2. Review the row count statistics for each component output.

3. Note the following:

   - 60 correct rows were loaded into the **DimOffice** table
   - 10 invalid rows were loaded to the **DimOffice_Error** table

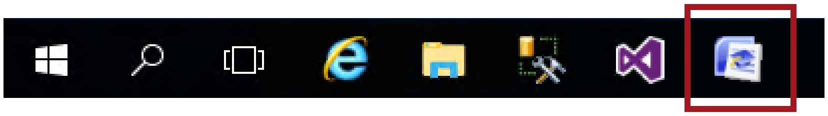
4. To stop the package debugging, on the **Debug** menu, select **Stop Debugging**.

5. To close SQL Server Data Tools, on the **File** menu, select **Exit**.

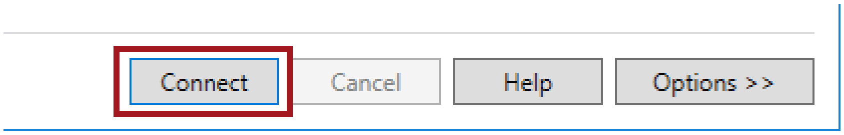
Reviewing Activity Monitoring

In this task, you will review the activity monitoring.

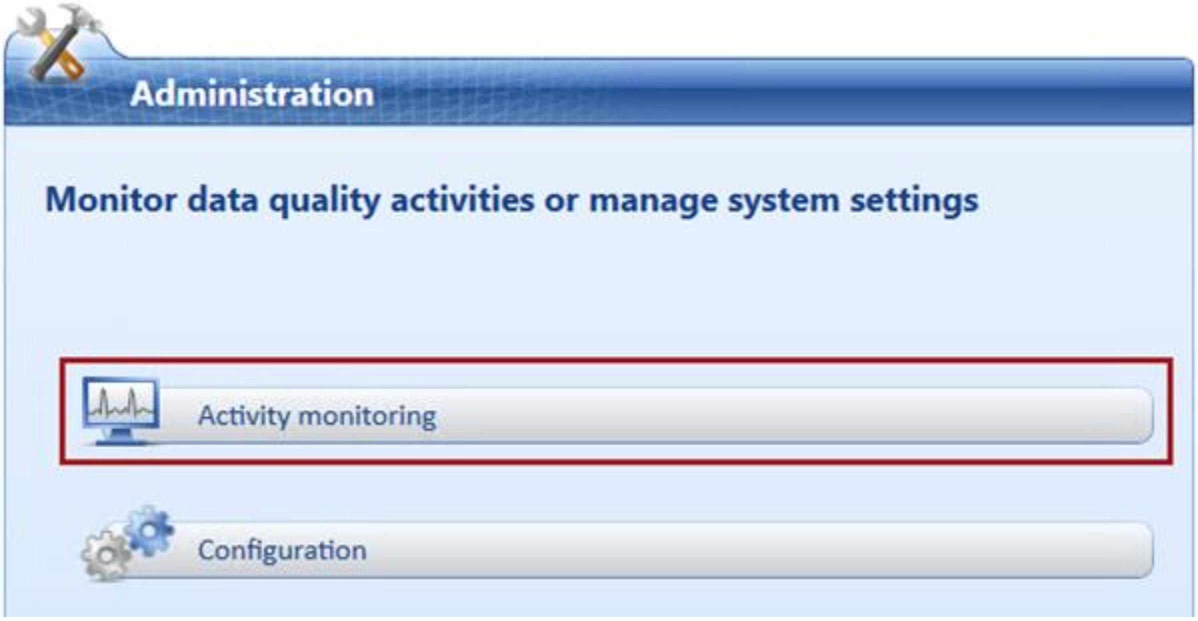
1. Open Data Quality Client.



2. In the **Connect to Server** window, click **Connect**.

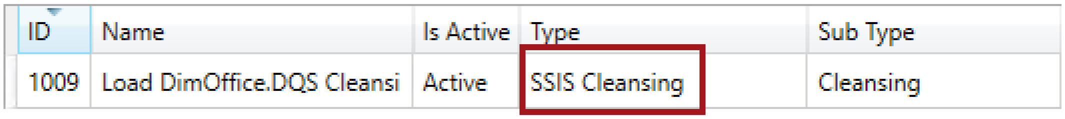


3. To monitor activity, in the **Administration** panel, click **Activity Monitoring**.



4. To sort the activities by descending order, in the activity grid, click the **ID** column header twice.

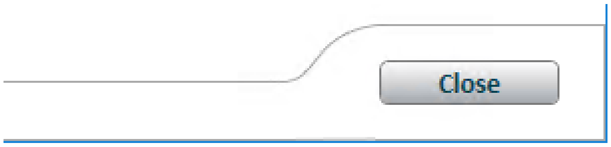
5. Notice the first listed activity is a **SSIS Cleansing** type.



ID	Name	Is Active	Type	Sub Type
1009	Load DimOffice.DQS Cleansi	Active	SSIS Cleansing	Cleansing

Every activity undertaken with the Data Quality Server—even when invoked by SSIS—is logged and remains available for review and audit.

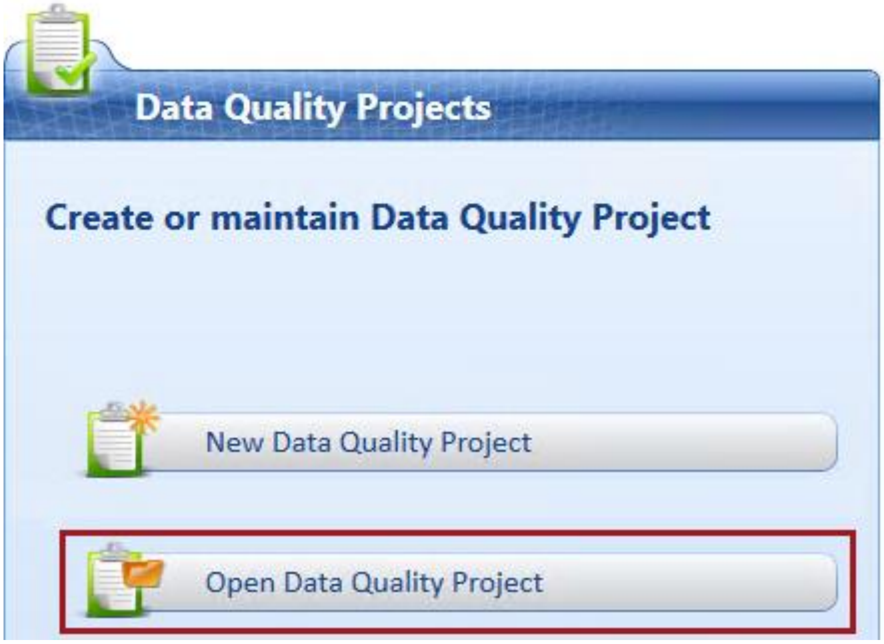
6. Click **Close**.



Reviewing the Cleansing Results

In this task, you will open the Data Quality Project and review the SSIS cleansing results.

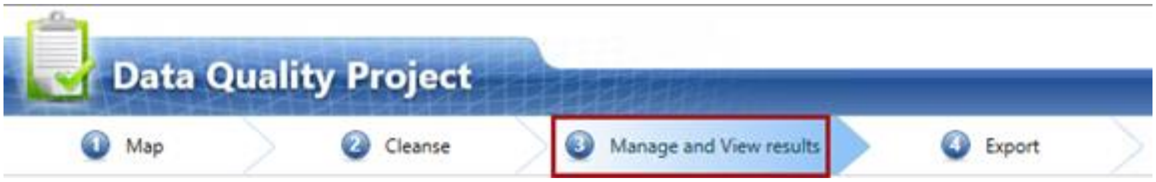
1. To open a Data Quality Project, in the **Data Quality Projects** panel, click **Open Data Quality Project**.


Data Quality Projects
Create or maintain Data Quality Project
New Data Quality Project
Open Data Quality Project

2. In the project grid, right-click the SSIS cleansing project, and then select **Open**.

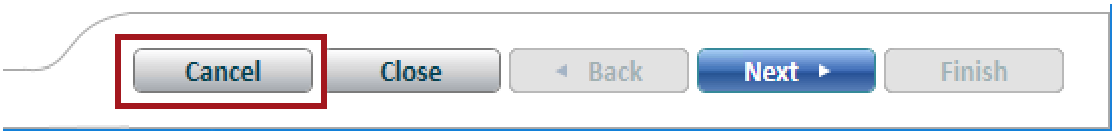
   The SSIS cleansing project is highlighted in red, and is locked.

3. Notice that the project opens at the **Manage and View Results** step.

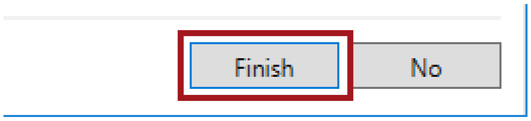


*It is not possible to go back to earlier steps, however it is possible to interactively correct the data, and then export it as you did in **Lab 2-1**. You will not do this in this lab.*

4. Click **Cancel**.



5. When prompted to continue, click **Finish**.

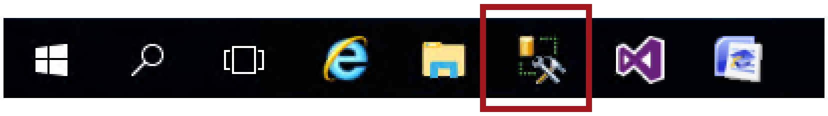


6. Close Data Quality Client.

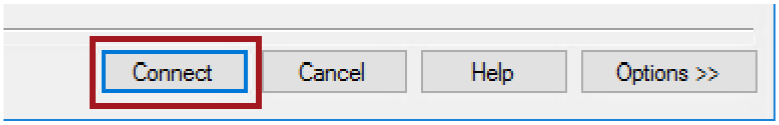
Analyzing the Cleansing Results

In this task, you will execute various queries to analyze the cleansing results output by the SSIS package execution.

1. Open SQL Server Management Studio.



2. In the **Connect to Server** window, click **Connect**.

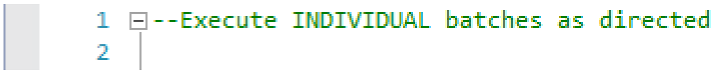


3. To open a script file, on the **File** menu, select **Open | File**.

4. In the **Open File** window, navigate to the **F:\Labs\Lab2-2\Assets** folder.

5. Select the **Script-01-ReviewSsisOutputs.sql** file, and then click **Open**.

6. In the script file, take note of the first line.



```
--Execute INDIVIDUAL batches as directed

```

It is very important that you execute the script in the manner intended. Many script files include multiple batches of statements (completed with the GO keyword), and so you should select the statements together with the GO keyword, and then execute only that selection.

*To execute a subset of a script, select the text you intend to execute, and then click **Execute** (or press **F5**).*

7. Read the comments in the first batch (line 3).

8. Select and execute the only query in the batch (lines 4-5).

9. Read the commented text, and then execute the query for each of the remaining batches in the script.

10. To exit SQL Server Management Studio, on the **File** menu, select **Exit**.

11. If prompted to save changes, click **No**.

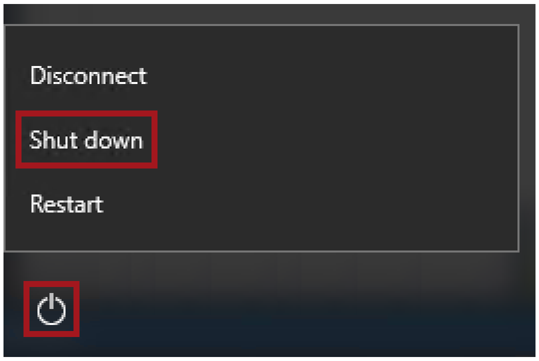
    *You have now completed the lab. If you are not commencing the next lab, you should complete the **Finishing Up** exercise to shut down and stop the VM.*

Finishing Up

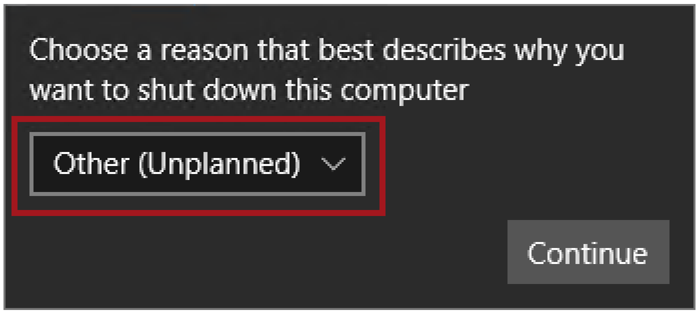
In this exercise, you will shut down and stop the VM.

1. Close all open applications.

2. Press the **Windows** key, and then in the **Start** page, located at the bottom-left, click the **Power** button, and then select **Shut Down**.



3. When prompted to choose a reason, to accept the default.

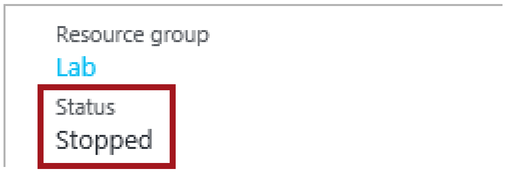


Choose a reason that best describes why you want to shut down this computer

Other (Unplanned)

4. Click **Continue**.

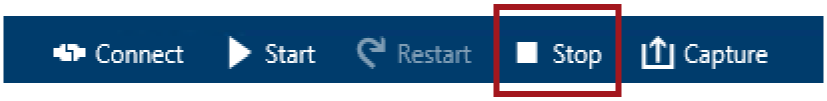
5. In the **Azure Portal** Web browser page, wait until the status of the VM updates to **Stopped**.



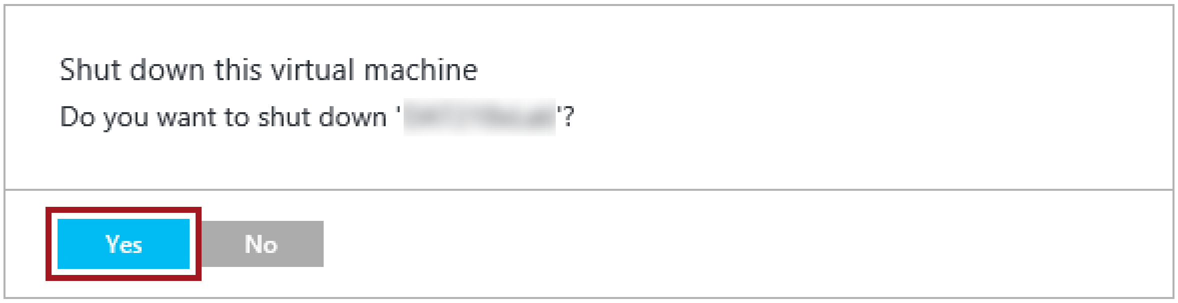
Resource group
Lab
Status
Stopped

In this state, however, the VM is still billable.

6. To deallocate the VM, click **Stop**.



7. When prompted to stop the VM, click **Yes**.

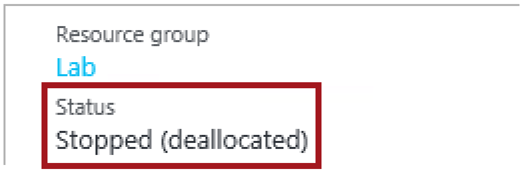


Shut down this virtual machine

Do you want to shut down '[redacted]'?

Deallocation will take some minutes to complete, and also extends the time required to restart the VM. Consider deallocating the VM if you want to reduce costs, or if you choose to complete the next lab after an extended period of time.

8. Verify that the VM status updates to **Stopped (Deallocated)**.



Resource group
Lab

Status
Stopped (deallocated)

In this state, the VM is now not billable—except for a relatively smaller storage cost.

Note that a deallocated VM will likely acquire a different IP address the next time it is started.

9. Sign out of the **Azure Portal**.