DAT218x

Cleansing Data with Data Quality Services

Lab 2-1 | Cleansing Data with a Data Quality Project

Estimated time to complete this lab is 45 minutes

Overview

In this lab, you will cleanse the Office dataset with a Data Quality Project by using the knowledge base created in **Lab 1-2**.

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

Exercise 1: Connecting to the VM

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

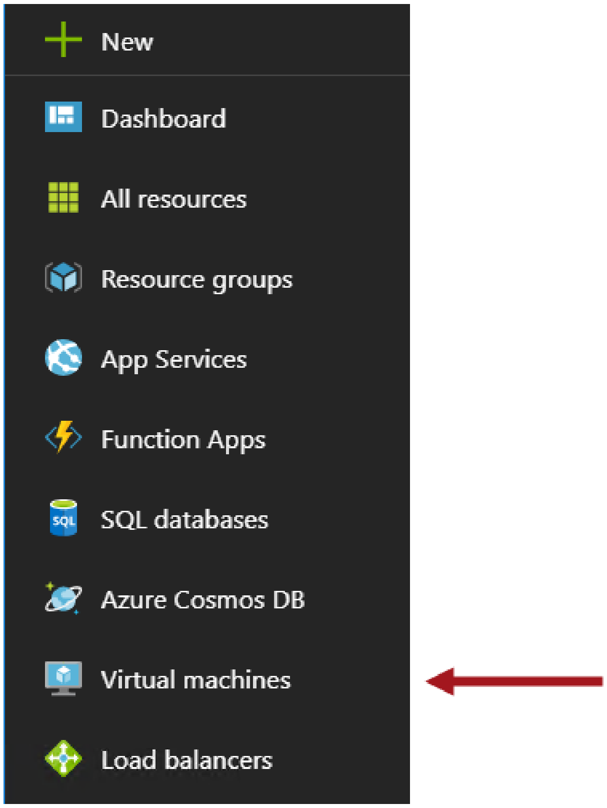
In this exercise, having signed in to the Azure Portal by using your Azure subscription, you will connect to the lab VM which you provisioned in **Lab 0-1**.

Connecting to the VM

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

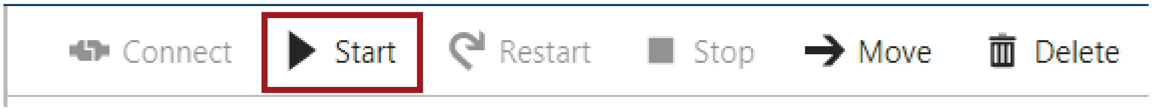
1. Sign in to the **Microsoft Azure Portal** by using your subscription.

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



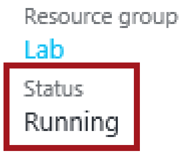
3. In the **Virtual Machines** blade, select the VM you provisioned in **Lab 0-1**.

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



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

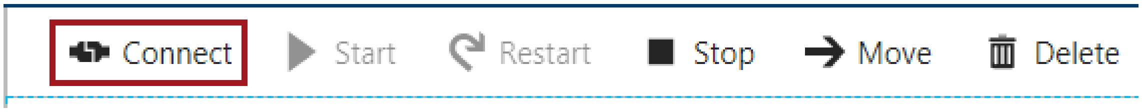
   It usually takes 1-2 minutes for the VM to start.



Resource group
Lab

Status
Running

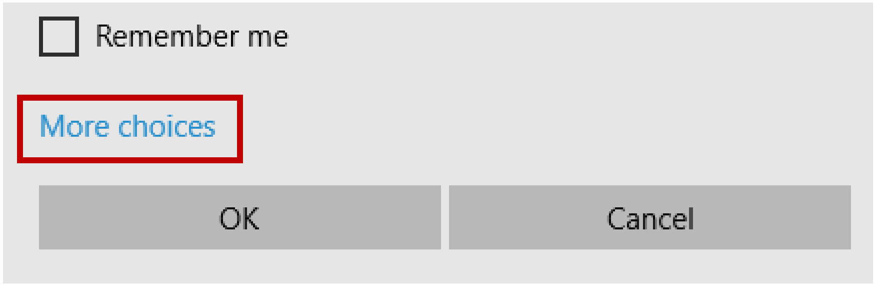
6. To connect to the VM, click **Connect**.



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

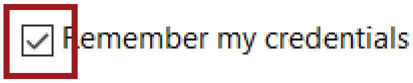
8. If prompted to connect to the unknown publisher, click **Connect**.

   *You need to enter the VM administrator credentials. If the authentication window defaults to an existing account, you will need to select **More Choices**, and then select **Use a Different Account**.*



9. In the **Windows Security** window, enter the VM admin credentials used when provisioning the VM.

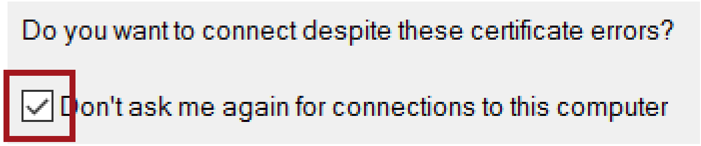
10. Check the **Remember My Credentials** checkbox.



☑ Remember my credentials

11. Click **OK**.

12. In the **Remote Desktop Connection** dialog 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.

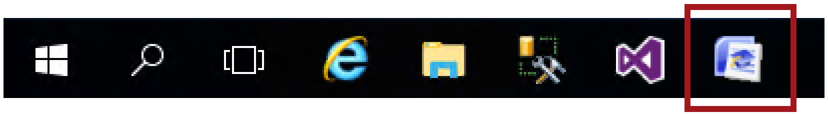
Exercise 2: Cleansing Data with a Data Quality Project

In this exercise, you will cleanse the Office dataset with a Data Quality Project by using the knowledge base created in **Lab 1-2**. This will provide the opportunity to interact with the cleansing results, and also enable domain values to be added back to the knowledge base.

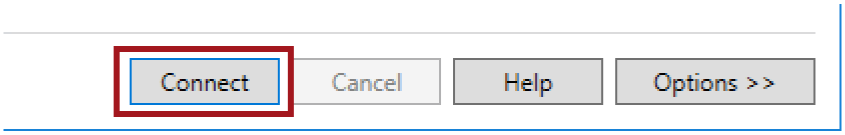
Creating the Data Quality Project

In this task, you will create a Data Quality Project

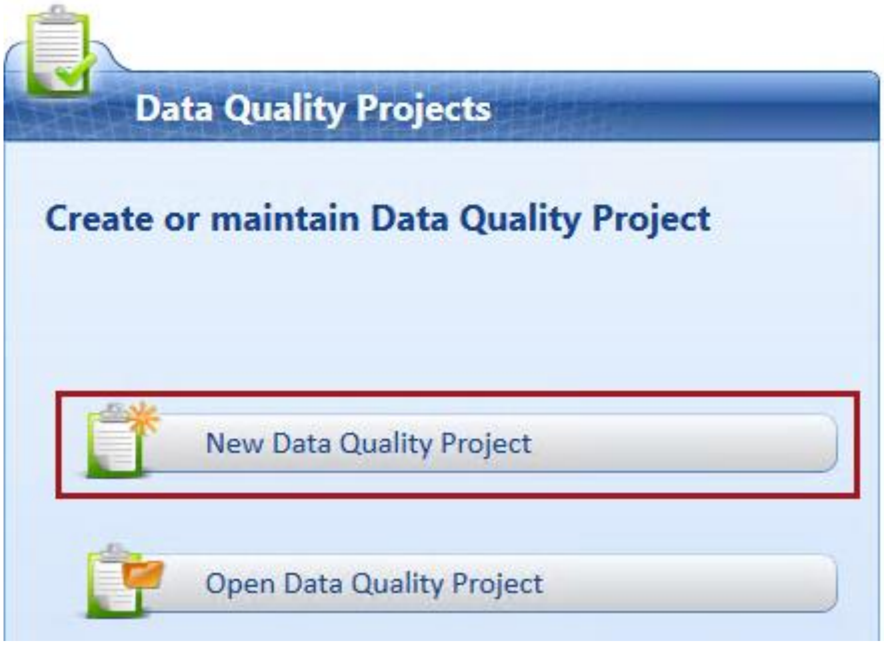
1. Open Data Quality Client.



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

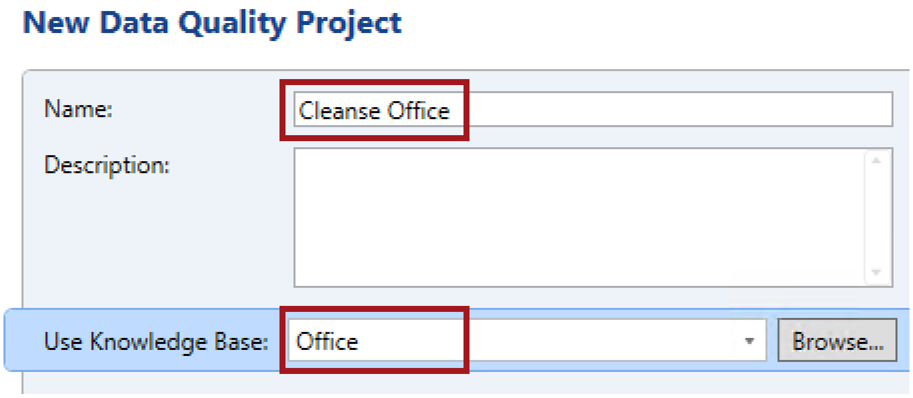


3. To create a new Data Quality Project, in the **Data Quality Projects** panel, click **New Data Quality Project**.


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

4. In the **Name** box, enter **Cleanse Office**.

5. In the **Use Knowledge Base** dropdown list, select **Office**.



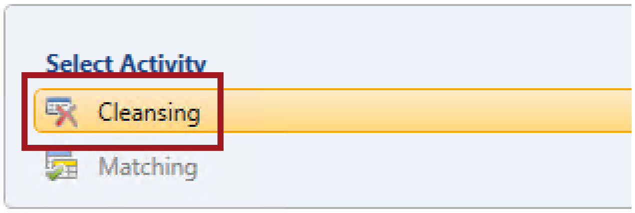
New Data Quality Project

Name: Cleanse Office

Description:

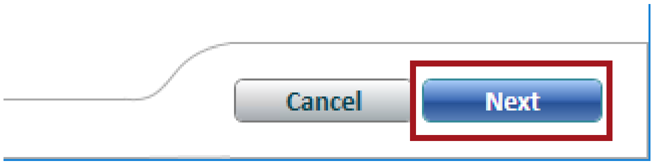
Use Knowledge Base: Office

6. In the lower pane, notice that the **Cleansing** activity is selected.



*The **Matching** activity is disabled, as the selected knowledge base does not yet contain a matching policy. You will create a matching policy in **Lab 2-3**.*

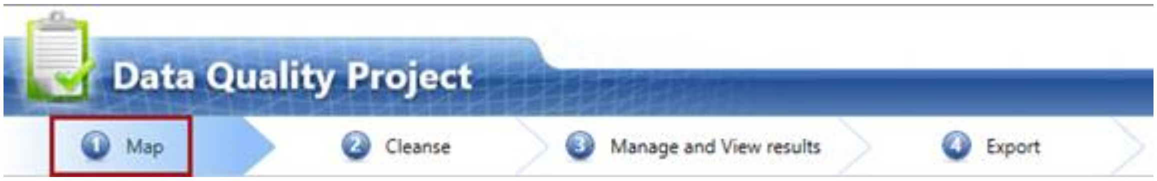
7. Click **Next**.



Mapping the Data to Cleanse

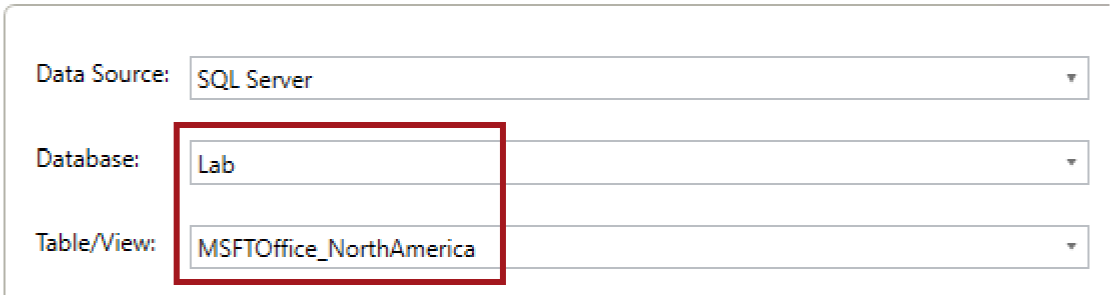
In this task, you will configure the data to cleanse, and also map it to the knowledge base domains.

1. Notice that step 1 of the activity is to map to external data to be cleansed.



2. In the **Database** dropdown list, select **Lab**.

3. In the **Table/View** dropdown list, select **MSFTOffice_NorthAmerica**.



Data Source: SQL Server

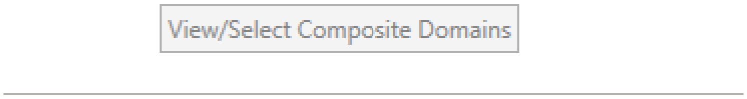
Database: Lab

Table/View: MSFTOffice_NorthAmerica

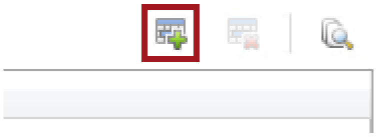
4. In the right pane, review the knowledge base details, including the **Address** composite domain and the domains it comprises.

5. Notice that the domain mappings used during knowledge discovery in the previous lab are automatically configured.

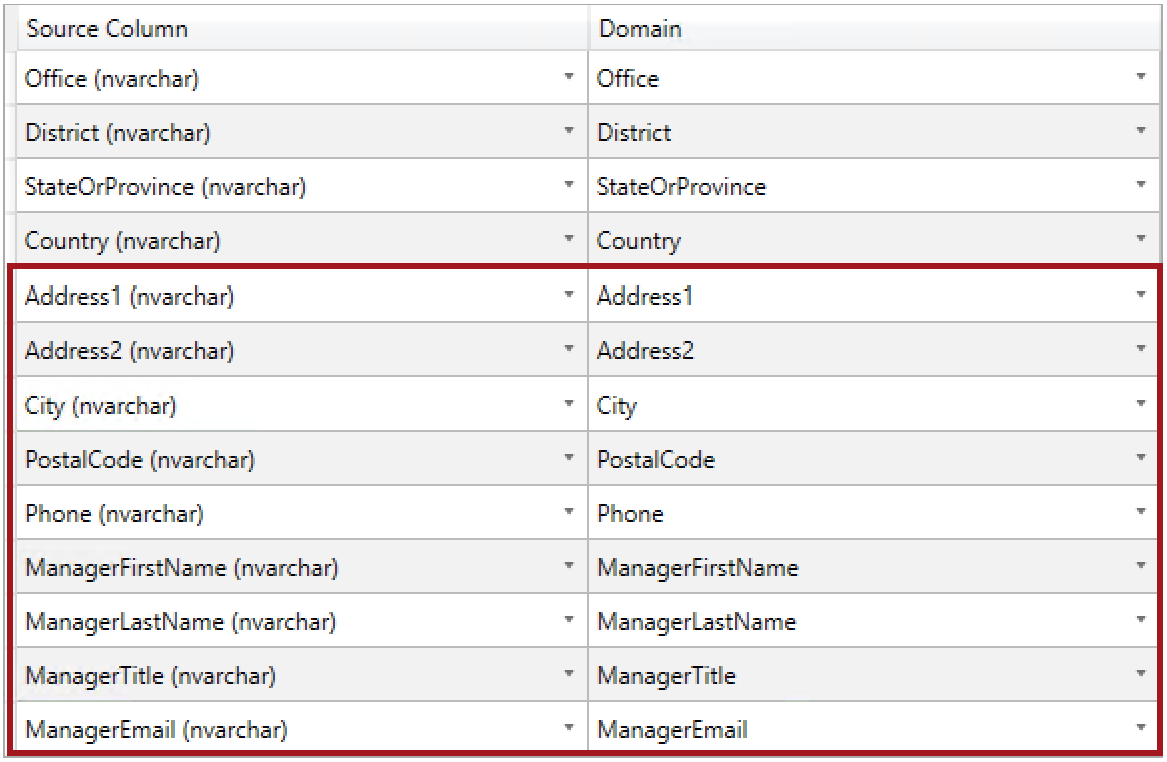
6. Located at the bottom, notice the **View/Select Composite Domains** button that is presently disabled.



7. To add additional mappings, click **Add a Column Mapping**.



8. Map the additional nine source columns (**Address1**, **Address2**, **City**, **PostalCode**, **Phone**, **ManagerFirstName**, **ManagerLastName**, **ManagerTitle** and **ManagerEmail**) to their respective domains.



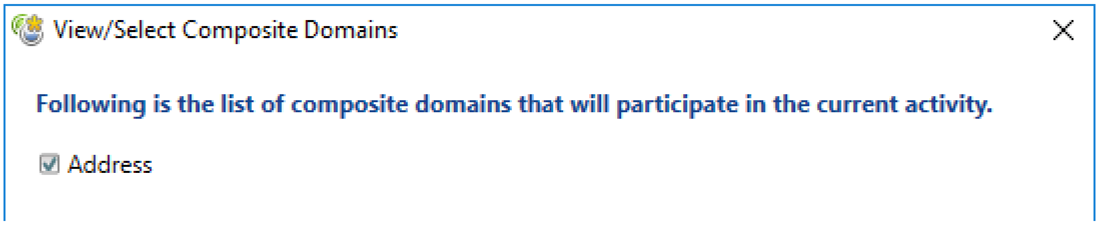
Source Column	Domain
Office (nvarchar)	Office
District (nvarchar)	District
StateOrProvince (nvarchar)	StateOrProvince
Country (nvarchar)	Country
Address1 (nvarchar)	Address1
Address2 (nvarchar)	Address2
City (nvarchar)	City
PostalCode (nvarchar)	PostalCode
Phone (nvarchar)	Phone
ManagerFirstName (nvarchar)	ManagerFirstName
ManagerLastName (nvarchar)	ManagerLastName
ManagerTitle (nvarchar)	ManagerTitle
ManagerEmail (nvarchar)	ManagerEmail

9. Notice the **View/Select Composite Domains** button that is now enabled.

The button will only become enabled when all domains for a composite domain are mapped.

10. Click the **View/Select Composite Domains** button.

11. Notice that the **Address** composite domain is checked, and so it will participate in the cleansing activity.

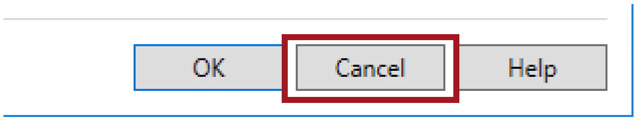


View/Select Composite Domains

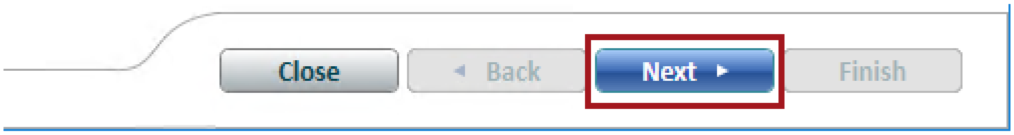
Following is the list of composite domains that will participate in the current activity.

☑ Address

12. Click **Cancel**.



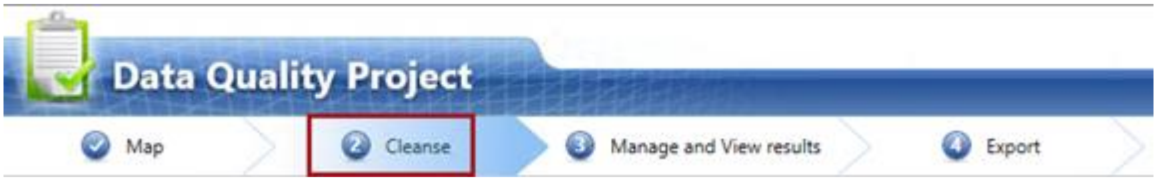
13. To proceed to the next step, click **Next**.



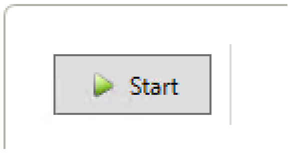
Cleansing the Data

In this task, you will cleanse the data, and then review the profiler results.

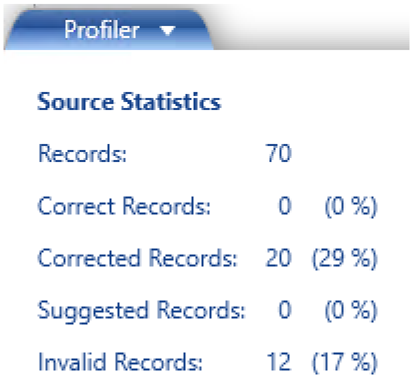
1. Notice that step 2 of the activity is to cleanse the source data.



2. To run the cleansing process, click **Start**.



3. When the cleansing has completed, review the source statistics in the **Profiler** pane.



Source Statistics

Records:	70	
Correct Records:	0	(0 %)
Corrected Records:	20	(29 %)
Suggested Records:	0	(0 %)
Invalid Records:	12	(17 %)

4. Note that all 70 records were cleansed, with 12 invalid records which were unable to be fully cleansed.

5. In the grid, review each domain, noticing that the first domain is in fact the **Address** composite domain comprising multiple fields.

Knowledge Base Check

Lab 2-1 ► Address Composite Domain and Source Statistics

You may need data from this step to answer a lab-based Knowledge Check associated with this module.

It is recommended that you open your Knowledge Check portion of the course in EdX at this time to answer questions as you complete the lab, or expand the section to take a screenshot of **Source Statistics** and **Address** composite domain, to refer to later.

6. Note that the cleansing of values results in outcomes of either:

   - Correct or Corrected
   - Suggested or New, or
   - Invalid

   Correct *means it was a correct value and/or passed all domain rules.*

   Corrected *means a domain value was applied, a term-based relation was applied, or a cross-domain rule resulted in a correction.*

   Suggested *means there is a possible correction subject to a confidence level (and would only be available if a reference data service was used).*

   New *means a value passed all domain rules, but was not a domain value (in this lab, most address values, names, phone numbers and titles are not actual values, and so would be output as* ***New****).*

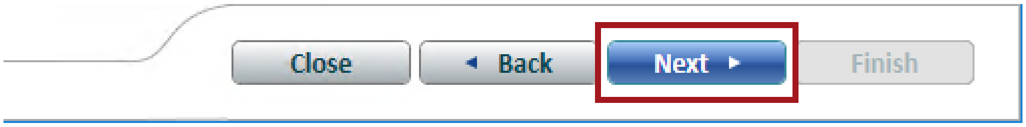
   Invalid *means a value was an error or an invalid domain value without a corresponding correction value, and/or it failed at least one domain rule.*

7. For each domain, review the number of corrected values, the completeness ratio, and also the accuracy ratio.

   You can safely ignore domains which have a high proportion of ***Suggested*** *or* ***New****, as these domains do not store domain values (i.e.* ***Phone*** *and* ***ManagerEmail****).*

   The immediate concern are ***Invalid*** *values that could not be cleansed, and so must be addressed interactively in the next step.*

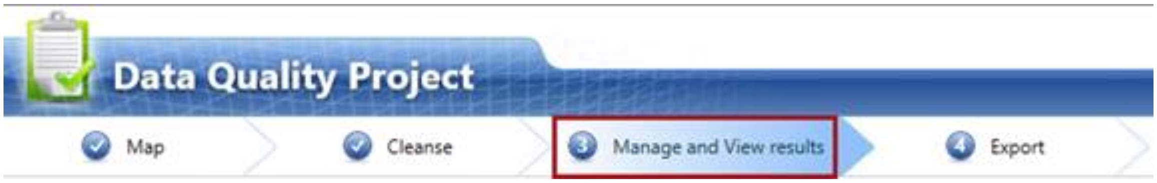
8. To proceed to the next step, click **Next**.



Managing and Viewing the Cleansing Results

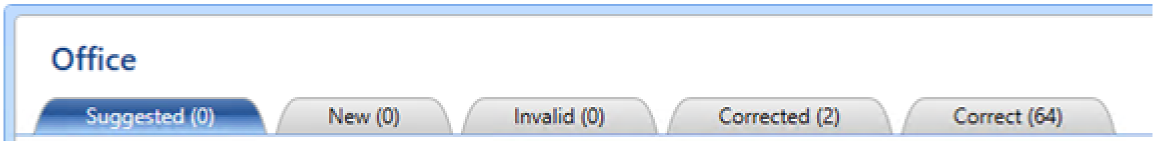
In this task, you will view the details of the cleansing results, and interactively apply corrections.

1. Notice that step 3 of the activity is to manage and view the cleansing results.



2. In the left pane, select the **Office** domain.

3. Review the domain cleansing results and their frequencies.

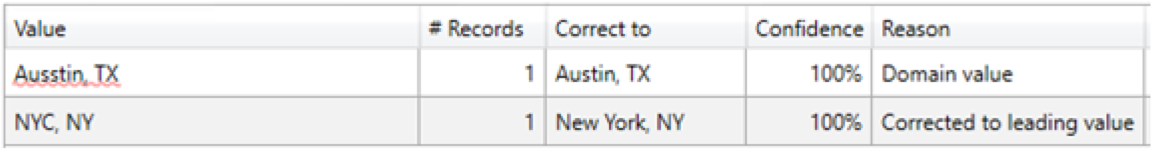


4. Notice the following:

- 2 values were corrected
- 66 values were already correct
- There were no values encountered that were not stored in the domain values
- There were no invalid results

5. Select the **Corrected** tab.

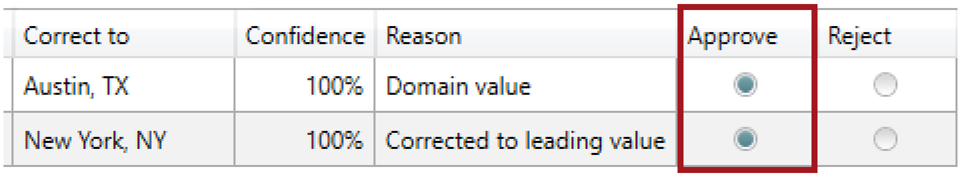
6. Review the two records, noting the source value, corrected value, and the reason.



Value	# Records	Correct to	Confidence	Reason
Ausstin, TX	1	Austin, TX	100%	Domain value
NYC, NY	1	New York, NY	100%	Corrected to leading value

7. Notice also that corrected values are automatically approved.

   Corrected values are always approved, and it is possible for you to reject these.



Correct to	Confidence	Reason	Approve	Reject
Austin, TX	100%	Domain value	●	○
New York, NY	100%	Corrected to leading value	●	○

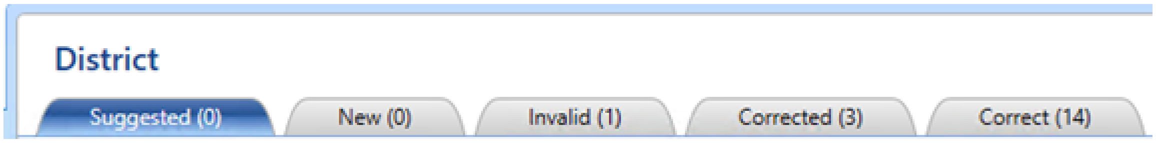
8. Select the **Correct** tab.

9. Review all records, noting that all were correct due to existing correct domain values.

   Correct values, like corrected values, are always approved.

10. Select the **District** domain.

11. Review the domain cleansing results and their frequencies.



12. Select the **Corrected** tab.

13. Notice that domain values resulted in these corrections.

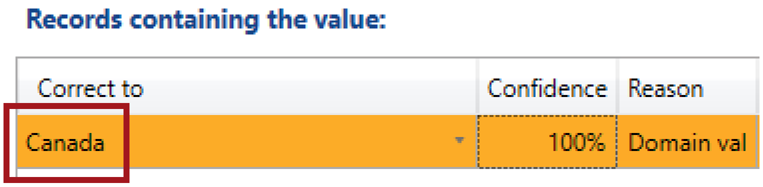
14. Select the **Invalid** tab.

15. Notice that **DQS_NULL** (a missing value) occurred once, and that it has been rejected.

    Invalid values are always rejected.

16. Select the **DQS_NULL** value.

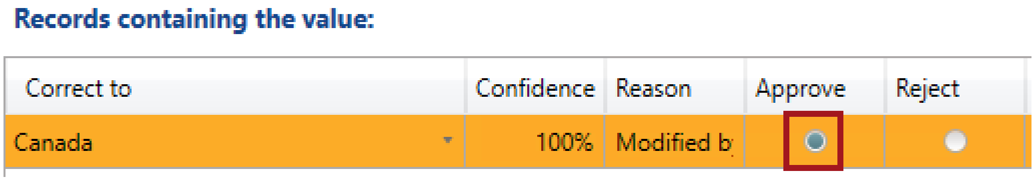
17. In the lower grid, in the **Correct to** box, enter **Canada**.



Records containing the value:

Correct to	Confidence	Reason
Canada	100%	Domain val

18. Approve the record.



Records containing the value:

Correct to	Confidence	Reason	Approve	Reject
Canada	100%	Modified b	●	○

Knowledge Base Check

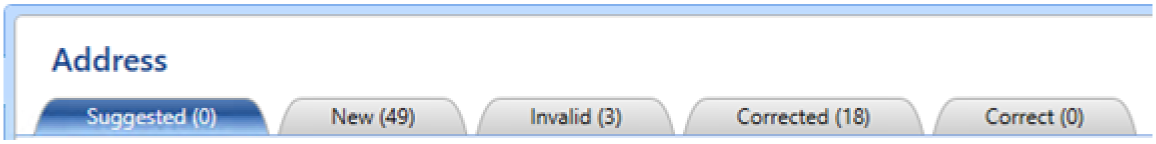
Lab 2-1 ► Invalid ManagerEmail

You may need data from this step to answer a lab-based Knowledge Check associated with this module.

It is recommended that you open your Knowledge Check portion of the course in EdX at this time to answer questions as you complete the lab, or select the **ManagerEmail** domain **Invalid** tab, and take a screenshot of the invalid values, to refer to later.

19. Select the **Address** composite domain.

20. Review the domain cleansing results and their frequencies.



21. Notice the following:

- 49 addresses are correct according to domain rules, but were not found in domain values
- 3 addresses could not be cleansed
- 18 addresses were corrected

22. Select the **Corrected** tab.

23. Notice the address corrections which are highlighted, and refer to the corresponding reason(s) for each.

24. Select the **Invalid** tab.

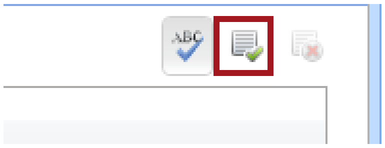
25. Review the reasons associated with the three values.

26. Enter correction values for each domain within the composite domain, as follows.

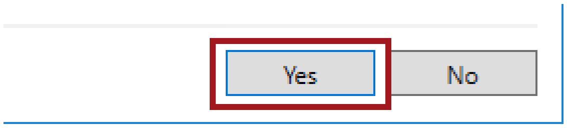
    You must enter the values for all domains.

Address1	Address2	City	StateOrProvince	PostalCode	Country
101 Wood Avenue South		Metro Park	NJ	08830	United States
1950 Meadow Blvd.		Mississauga	ON	L5N 8L9	Canada
7595 Technology Way	Suite 400	Denver	CO	80237	United States

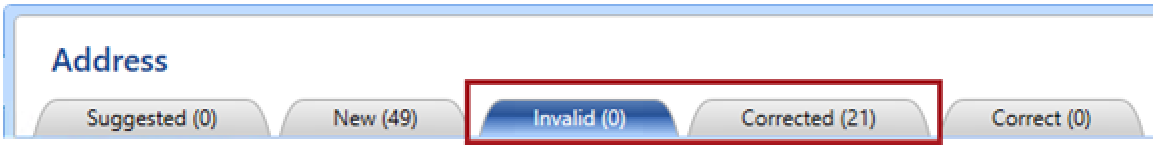
27. To approve all corrections, at the top right of the grid, click **Approve All Items**.



28. When prompted to confirm the approval of all items, click **Yes**.

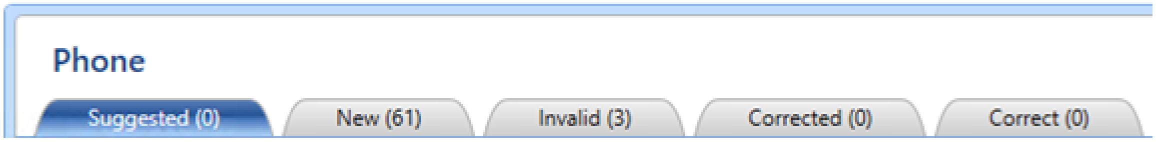


29. Notice that there are no longer any invalid values, and the corrected values have increased from 18 to 21.



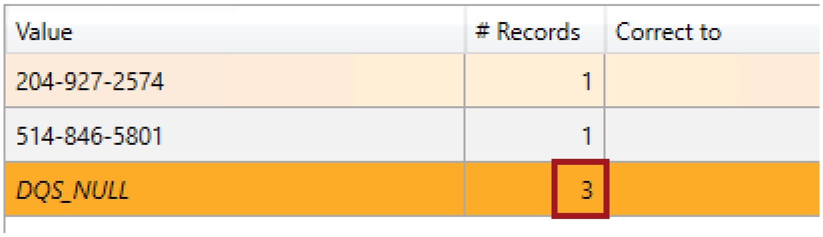
30. Select the **Phone** domain.

31. Review the domain cleansing results and their frequencies.



32. Select the **Invalid** tab.

33. Note the three records have missing telephone numbers.



Value	# Records	Correct to
204-927-2574	1	
514-846-5801	1	
DQS_NULL	3	

34. Enter the corrected to values, as follows.

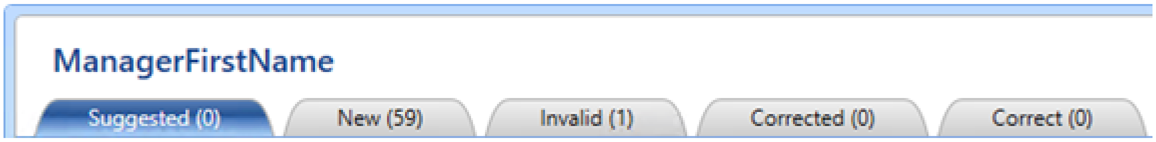
Value	Correct To
204-927-2574	(204) 927-2574
514-846-5801	(514) 846-5801
DQS_NULL	(800) 123-4567

35. Approve all corrections.

You will import these corrections into the knowledge base in the next exercise.

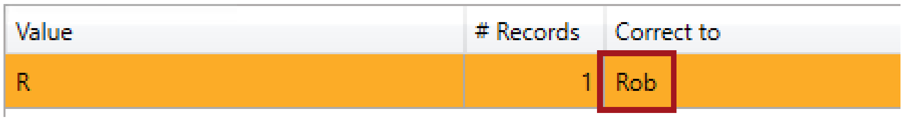
36. Select the **ManagerFirstName** domain.

37. Review the domain cleansing results and their frequencies.



38. Select the **Invalid** tab.

39. Enter the corrected to value, as follows.

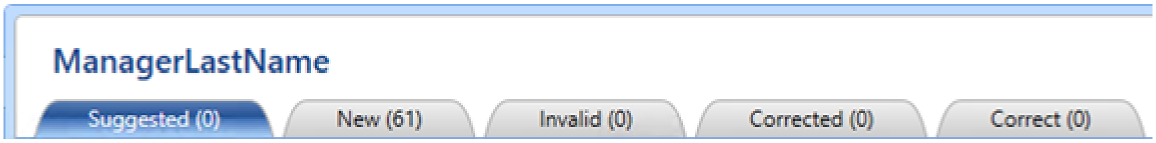


Value	# Records	Correct to
R	1	Rob

40. Approve the correction.

41. Select the **ManagerLastName** domain.

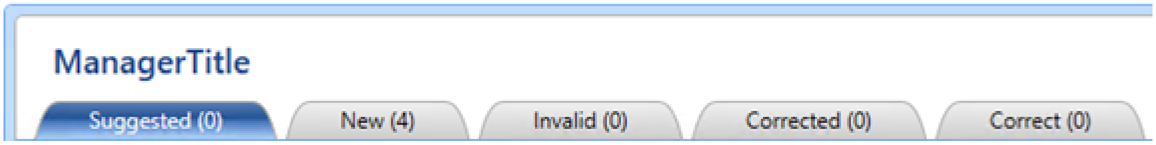
42. Review the domain cleansing results and their frequencies.



There are no invalid values.

43. Select the **ManagerTitle** domain.

44. Review the domain cleansing results and their frequencies.



There are no invalid values. You will, however, import the four new values into the knowledge base in the next exercise.

Knowledge Base Check

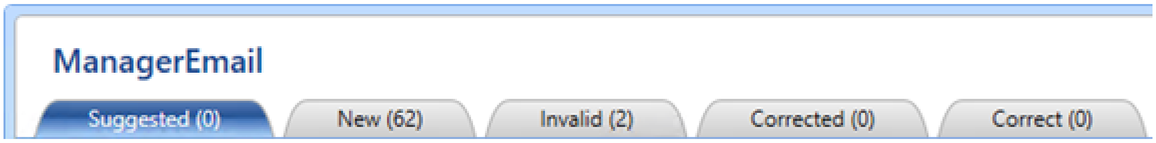
Lab 2-1 ► ManagerTitle Domain

You may need data from this step to answer a lab-based Knowledge Check associated with this module.

It is recommended that you open your Knowledge Check portion of the course in EdX at this time to answer questions as you complete the lab, or select the **ManagerTitle** domain **New** tab, and take a screenshot of the **ManagerTitle** value and **# Records**, to refer to later.

45. Select the **ManagerEmail** domain.

46. Review the domain cleansing results and their frequencies.



47. Select the **Invalid** tab.

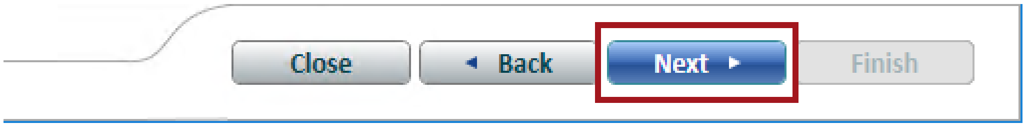
48. Enter the corrections as follows.

Value	Correct To
doris.hartwig@vendor.microsoft.com	doris.hartwig@lab.microsoft.com
MICHAEL.HINES@@lab.microsoft.com	michael.hines@lab.microsoft.com

49. Approve all corrections.

All invalid values have now been interactively cleansed, and approved.

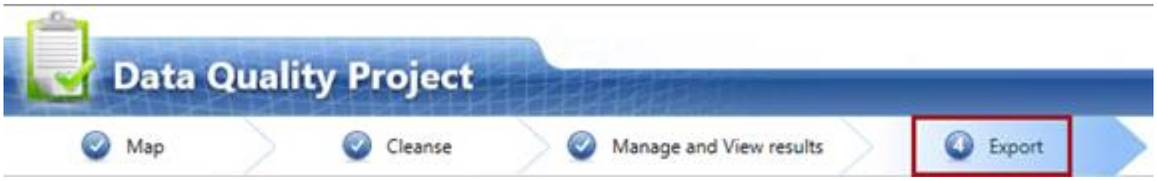
50. To proceed to the next step, click **Next**.



Exporting the Cleansing Results

In this task, you will export the cleansing results.

1. Notice that step 4 of the activity is to export the cleansing results.



2. In the left pane, review the output data preview.

Knowledge Base Check
Lab 2-1 ► District Domain

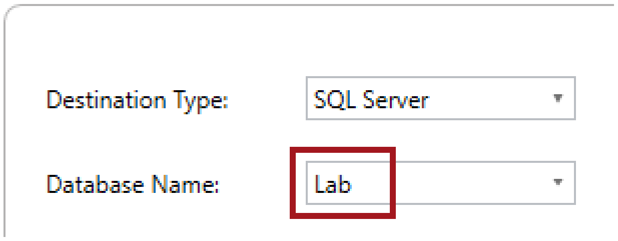
You may need data from this step to answer a lab-based Knowledge Check associated with this module.

It is recommended that you open your Knowledge Check portion of the course in EdX at this time to answer questions as you complete the lab, or take a screenshot of the **DistrictSource** column value for **Denver**, to refer to later.

3. Notice that each domain includes a column to describe:

   - Source
   - Output
   - Reason
   - Confidence
   - Status

4. To export the results, in the right pane, in the **Database Name** dropdown list, select **Lab**.



Destination Type: SQL Server

Database Name: Lab

5. In the **Table Name** box, enter **Lab2-1-DataOnly**.

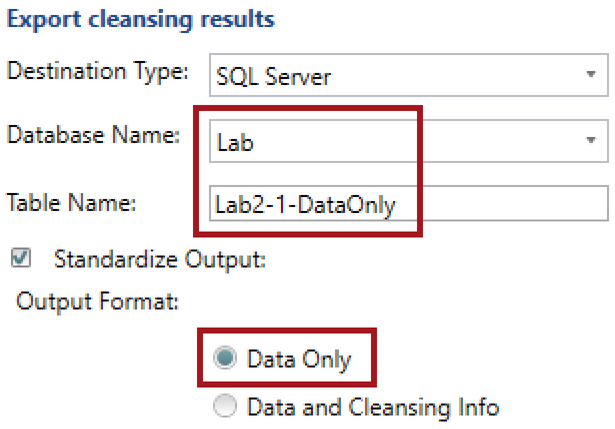
   For your convenience and accuracy, you can copy the table name from the ***F:\Labs\Lab2-1\Assets\Snippets.txt*** *file (open with Notepad).*

6. Notice that the **Standardized Output** checkbox is checked.

☑ Standardize Output:

   Standardized output will ensure that ***StateOrProvince*** *values are formatted in upper case, and that* ***ManagerEmail*** *values are formatted in lower case, as you defined in the domain properties in* ***Lab 1-2****.*

7. For **Output Format**, select the **Data Only** option.



Export cleansing results

Destination Type: SQL Server

Database Name: Lab

Table Name: Lab2-1-DataOnly

☑ Standardize Output:

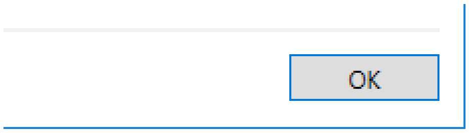
Output Format:

● Data Only

○ Data and Cleansing Info

8. Click **Export**.

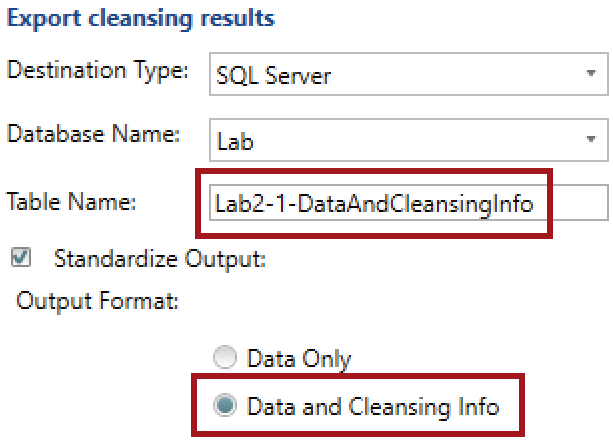
9. When notified that the export to database has completed, click **OK**.



10. In the **Table Name** box, replace the text with **Lab2-1-DataAndCleansingInfo**.

    *For your convenience and accuracy, you can copy the table name from the **F:\Labs\Lab2-1\Assets\Snippets.txt** file (open with Notepad).*

11. For **Output Format**, select the **Data and Cleansing Info** option.



Export cleansing results

Destination Type: SQL Server

Database Name: Lab

Table Name: Lab2-1-DataAndCleansingInfo

☑ Standardize Output:

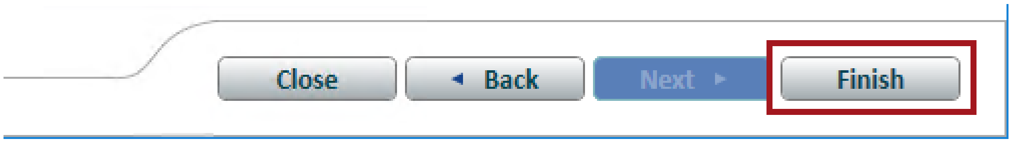
Output Format:

○ Data Only

● Data and Cleansing Info

12. Export the results.

13. To complete the cleansing project, click **Finish**.

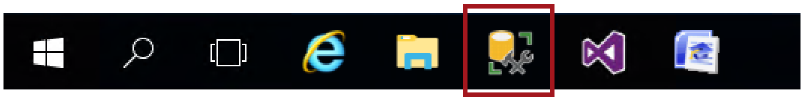


14. Review the activity monitoring, and notice the cleansing activity you have just completed.

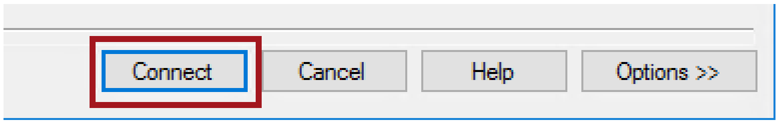
Analyzing the Cleansing Results

In this task, you will execute various queries to analyze the cleansing results.

1. Open SQL Server Management Studio.



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

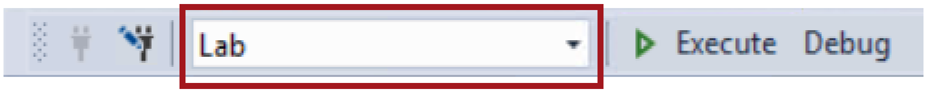


3. To open a script file, on the **File** menu, select **Open | File**.

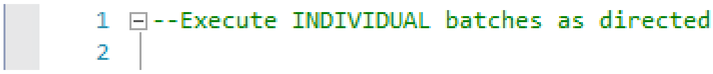
4. In the **Open File** window, navigate to the **F:\Labs\Lab2-1\Assets** folder.

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

6. On the toolbar, select the **Lab** database.



7. 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****).*

8. Read the comments in the first batch (lines 3-5).

9. Select and execute the only query in the batch (lines 6-7).

Knowledge Base Check

Lab 2-1 ► Query Results

You may need data from this step to answer a lab-based Knowledge Check associated with this module.

It is recommended that you open your Knowledge Check portion of the course in EdX at this time to answer questions as you complete the lab, or expand the query results to line 17 to take a screenshot of the results pane to refer to later.

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

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

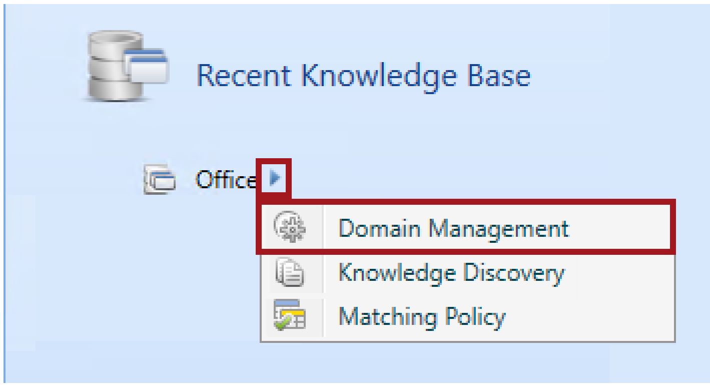
12. If prompted to save changes, click **No**.

Importing Project Values

In this task, you will perform domain management, and import the new values discovered in the cleansing activity for the **ManagerTitle** and **Phone** domains.

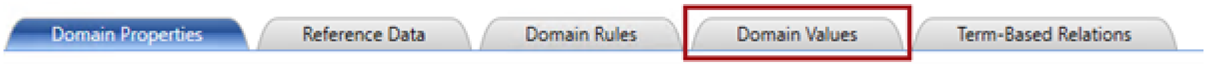
1. Switch to Data Quality Client.

2. To perform knowledge discovery, in the **Knowledge Base Management** panel, click the **Office** knowledge base, and then select the **Domain Management** activity.

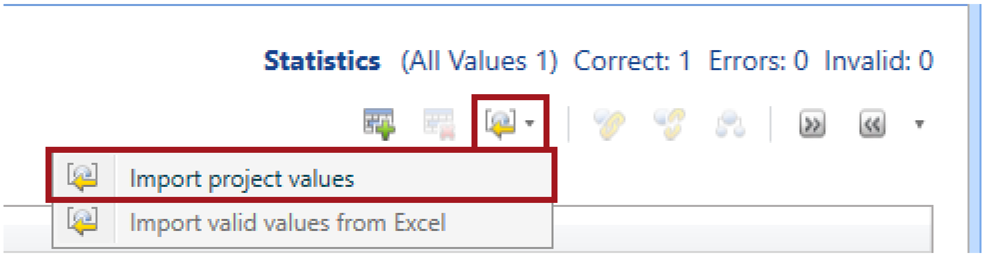


3. Select the **ManagerTitle** domain.

4. Select the **Domain Values** tab.



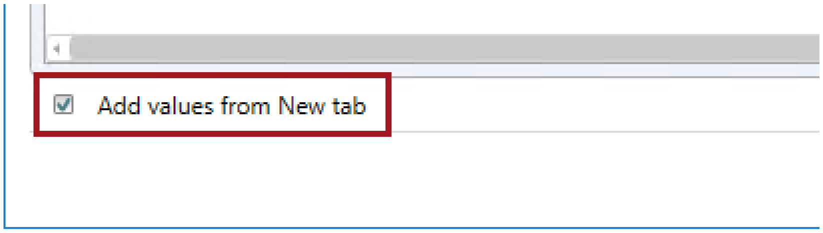
5. Click **Import Values**, and then select **Import Project Values**.


Statistics (All Values 1) Correct: 1 Errors: 0 Invalid: 0
Import project values
Import valid values from Excel

The option to import valid values from Excel is disabled as Microsoft Office is not installed on the VM.

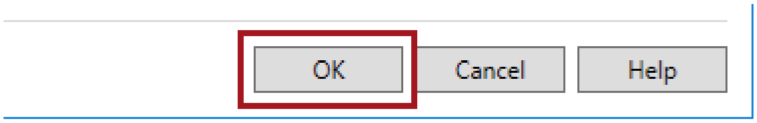
6. In the **Import Project Values** window, notice that the **Cleanse Office** project is selected.

7. Notice also that the **Add Values from New Tab** checkbox is checked.

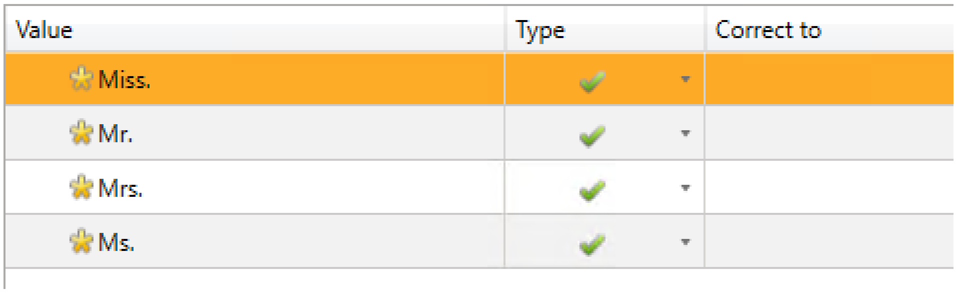


☑ Add values from New tab

8. Click **OK**.



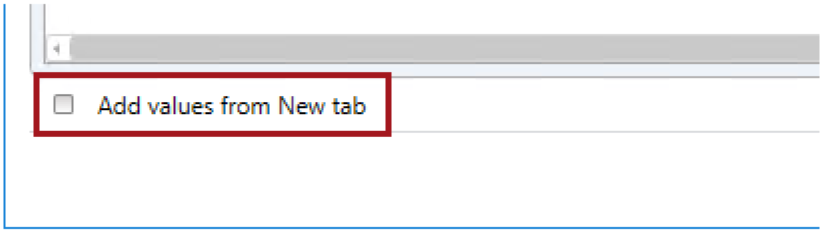
9. Verify that the four titles are added to the domain values.



Value	Type	Correct to
Miss.	✓	
Mr.	✓	
Mrs.	✓	
Ms.	✓	

10. Select the **Phone** domain.

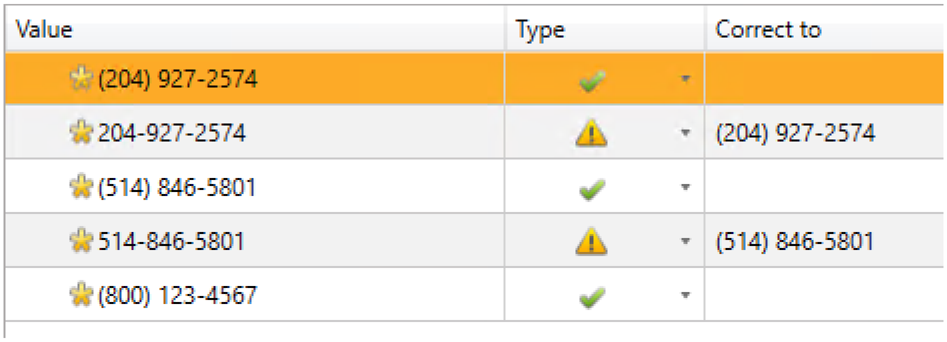
11. Import values from the project, but this time do not add **New** values.



□ Add values from New tab

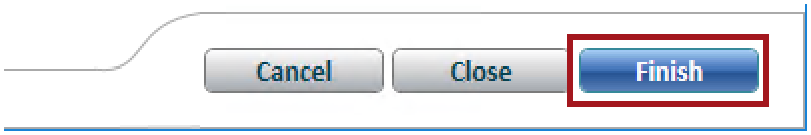
Only manual corrections are added as domain values.

12. Verify that the five values—three correct and two invalid, were added to the domain values.

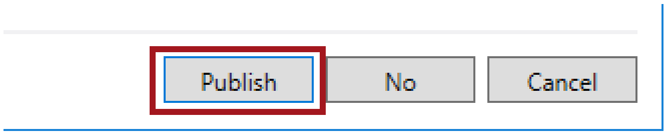


Value	Type	Correct to
(204) 927-2574	✓	
204-927-2574	⚠	(204) 927-2574
(514) 846-5801	✓	
514-846-5801	⚠	(514) 846-5801
(800) 123-4567	✓	

13. To complete the domain management activity, click **Finish**.



14. Publish the knowledge base.



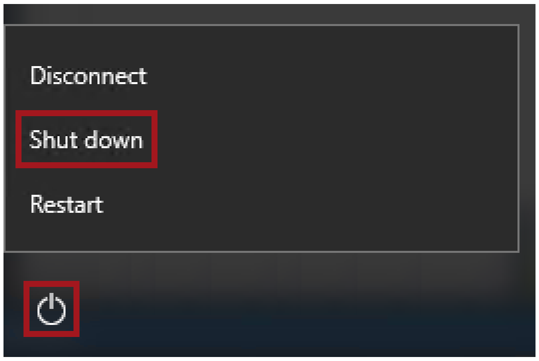
*You have now completed the lab. If you are not commencing the next lab, you should complete the **Finishing Up** exercise to shut down and stop the VM.*

Finishing Up

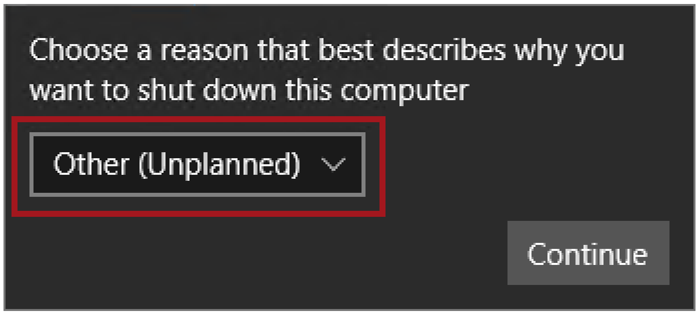
In this exercise, you will shut down and stop the VM.

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**.



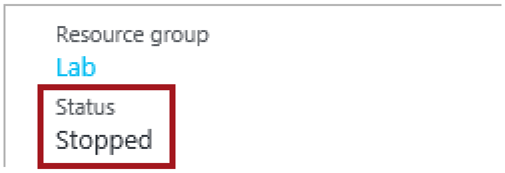
3. When prompted to choose a reason, select **Other (Unplanned)**.


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**.

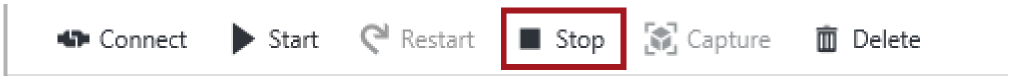

Resource group
Lab

Status
Stopped

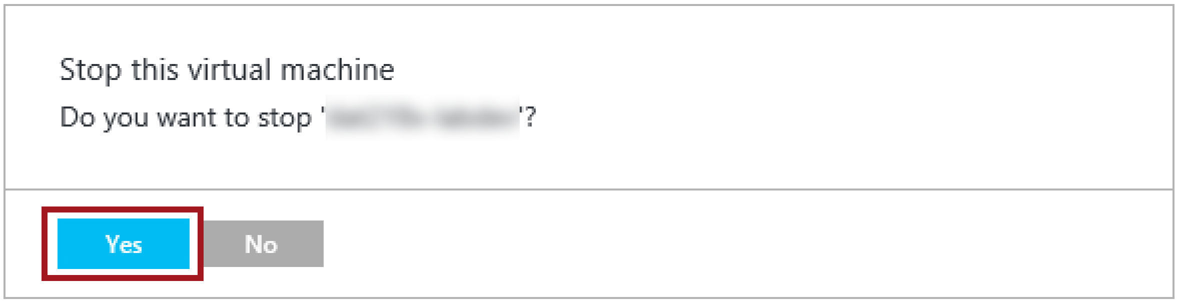
In this state, however, the VM is still billable.

6. Optionally, to deallocate the VM, click **Stop**.

   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.



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

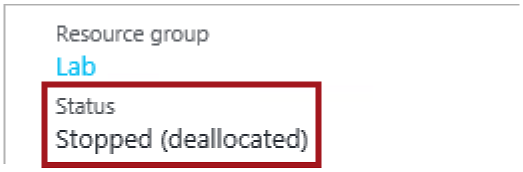


Stop this virtual machine

Do you want to stop '[redacted]'?

   The deallocation can take several minutes to complete.

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**.