

DAT218x

Cleansing Data with Data Quality Services

Lab 2-3 | De-duplicating Data with a Data Quality Project

Estimated time to complete this lab is 45 minutes

Overview

In this lab, you will further enhance the knowledge base by creating a matching policy to identify duplicate offices. You will then use a Data Quality Project matching activity to identify duplicate records stored in the **DimOffice** table.

This is the final lab in the course. Once you have completed the lab, you will be guided to delete the VM.

Exercise 1: Connecting to the VM

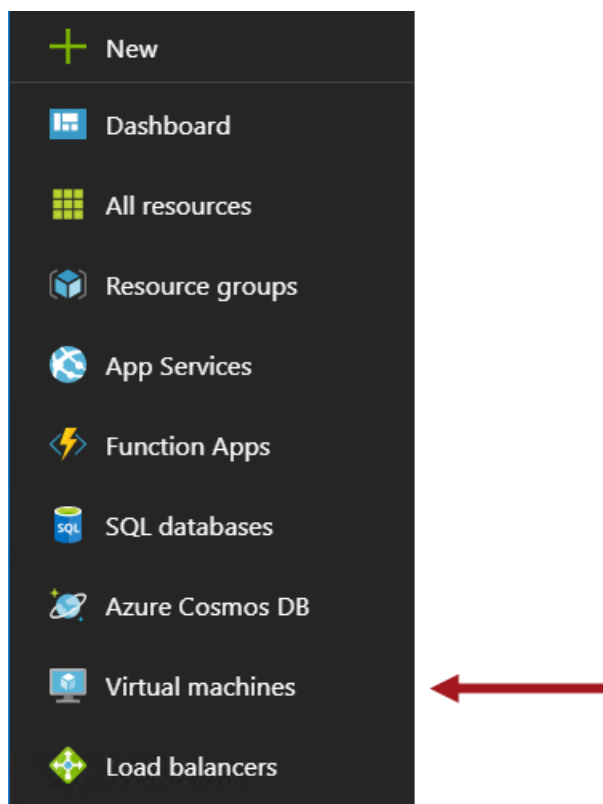
Go to the next exercise if you are already connected to the lab VM.

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**.

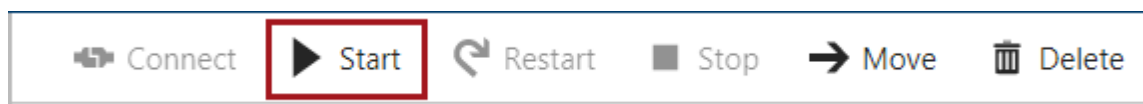
Connecting to the VM

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**.

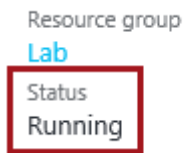


3. In the **Virtual Machines** blade, select the VM you provisioned in **Lab 0-1**.
4. In the VM blade, click **Start**.



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

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

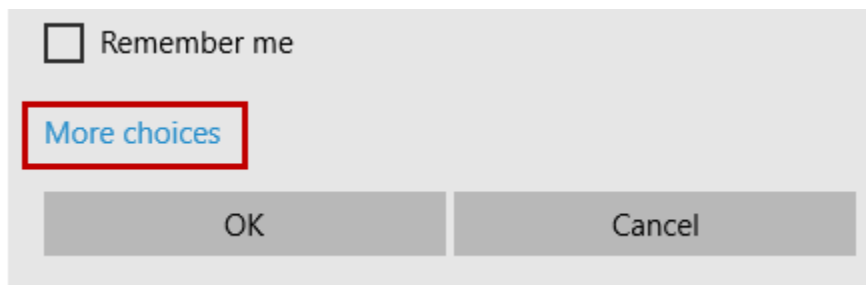


- To connect to the VM, click **Connect**.

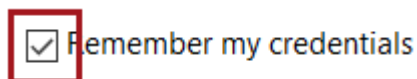


- When prompted to open the Remote Desktop File, click **Open**.
- If prompted to connect to the unknown publisher, click **Connect**.

*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**.*

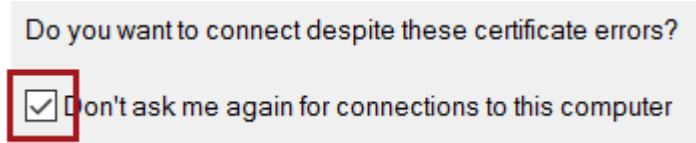


- In the **Windows Security** window, enter the VM admin credentials used when provisioning the VM.
- Check the **Remember My Credentials** checkbox.



- Click **OK**.

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



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

Exercise 2: Creating a Matching Policy

In this exercise, you will further enhance the knowledge base by creating a matching policy to identify duplicate offices.

Creating the DQS Connection Manager

In this task, you will create a DQS connection manager.

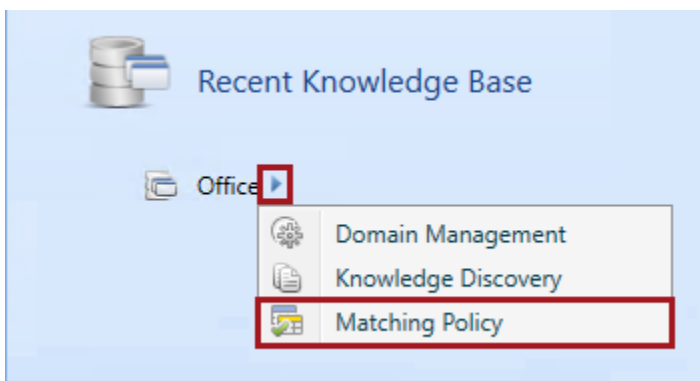
1. Open Data Quality Client.



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



3. To create a matching policy, in the **Knowledge Base Management** panel, click the **Office** knowledge base, and then select the **Matching Policy** activity.



4. Notice that step 1 of the activity is to connect to sample data to create matching policy rules.



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

6. In the **Table/View** dropdown list, select **DimOffice**.



Data Source: SQL Server

Database: Lab

Table/View: DimOffice

7. Create the following five mappings from source column to domain.

Source Column	Domain
Office (nvarchar)	Office
Address1 (nvarchar)	Address1
City (nvarchar)	City
PostalCode (nvarchar)	PostalCode
Country (nvarchar)	Country

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



Cancel Close Back Next Finish

9. Notice that step 2 of the activity is to create a matching policy.



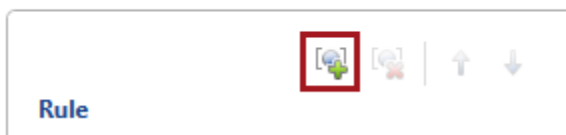
Knowledge Base Management

1 Map 2 Matching Policy 3 Matching Results

Only one matching policy can be created per knowledge base.

10. Click **Create a Matching Rule**.

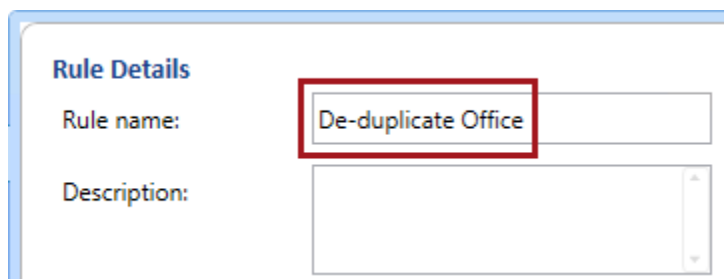
Create matching policy



Rule

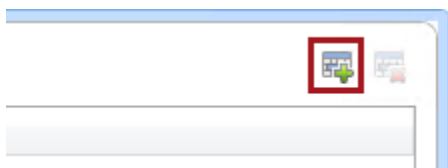
Add Remove Up Down

11. In the **Rule Name** box, replace the text with **De-duplicate Office**.



The screenshot shows a 'Rule Details' form. The 'Rule name:' field contains the text 'De-duplicate Office', which is highlighted with a red rectangular box. The 'Description:' field is empty.

12. To create a domain element for the rule, click **Add a New Domain Element**.



13. Configure the domain element based on the following table.

Domain	Similarity	Weight
Office	Similar	50

Rule Editor

Domain	Similarity	Weight	Prerequisite
Office	Similar	50	<input type="checkbox"/>

The domain element ensures that the **Office** value can be similar, and its similarity score will contribute 50% to the matching score.

14. Add four additional domain elements, and notice that they are configured for the remaining mapped domains.

The rule editor grid cannot be resized, and so it requires some patience to achieve the desired configuration.

15. Configure additional domain elements based on the following table (order is not important).

Domain	Similarity	Weight	Prerequisite
Country	Exact		Checked
Address1	Similar	20	Unchecked
City	Similar	20	Unchecked
PostalCode	Similar	10	Unchecked

Rule Editor

Domain	Similarity	Weight	Prerequisite
Country	Exact		<input checked="" type="checkbox"/>
Office	Similar	50	<input type="checkbox"/>
Address1	Similar	20	<input type="checkbox"/>
City	Similar	20	<input type="checkbox"/>
PostalCode	Similar	10	<input type="checkbox"/>

Like **Office** values, the **Address1**, **City** and **PostalCode** values can be similar, and together their weight values add to 100%.

The **Country** values must be an exact match, and also a prerequisite meaning that if the country values do not match, then the two records cannot be considered duplicates.

16. To test the rule, click **Start**.



DQS will index the source data, and then perform matching.

Knowledge Base Check

Lab 2-3 ► Matched Record ID, PostalCode and Office Key

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 of the data from the **Record ID** and associated **PostalCode** and **Office** key columns, to refer to later.

17. Notice that four clustered were detected.

Record Id values in bold indicate a pivot record, which is the record that will be retained (survivor). The records within the cluster are duplicates that will be discarded. You will have some influence over which record is assigned as the pivot record when performing a Data Quality Project matching project.

18. Notice that the last two clusters are for the same New York office.

Within these two clusters, notice that the **OfficeKey** values repeat, meaning that records have been assigned to more than one cluster.

41	New York Metro District
42	New York Metro District
41	New York Metro District
42	New York Metro District

19. Select **Non Overlapping Clusters**.

Non overlapping clusters will ensure that records relate to only one cluster.



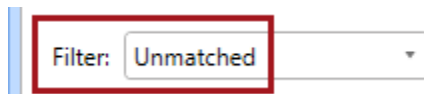
20. Click **Restart**.



21. Notice this time, that only three clusters were detected.

*The two **Tampa, FL** offices were not detected, due to a matching score less than the minimum matching score (80%) configured.*

22. In the **Filter** dropdown list, select **Unmatched**.



23. In the grid, to sort the data, click the **Office** column header.

24. Notice the two **Tampa, FL** records, and that their **Address1** values are not quite similar.

25. In the **Filter** dropdown list, select **Matched**.



26. In the rule editor, modify the domain elements by:

- Increasing the **Office** weight to **60%**, and
- Decreasing the **Address1** weight to **10%**.

Rule Editor

Domain	Similarity	Weight	Prerequisite
Country ▾	Exact ▾		<input checked="" type="checkbox"/>
Office ▾	Similar ▾	60 % ▴▾	<input type="checkbox"/>
Address1 ▾	Similar ▾	10 % ▴▾	<input type="checkbox"/>
City ▾	Similar ▾	20 % ▴▾	<input type="checkbox"/>
PostalCode ▾	Similar ▾	10 % ▴▾	<input type="checkbox"/>


27. Click **Restart**.



28. Notice that four clusters were now detected.

Creating a matching rule is a process of trial-and-error, ultimately arriving at an optimal set of rules to detect duplicate records.

29. Right-click the non-pivot **Tampa, FL** record, and then select **View Details**.

	1000017	1000017		Tampa, FL
	1000018	1000017	80%	Tampa, FL

30. Review the score details, noting the following:

- The matching score is 80%
- The fields that contributed to the score were **Office** and **City**, with exact matches encountered, and so they contributed 60% ($0.6 \times 100\%$) and 20% ($0.2 \times 100\%$) respectively
- The **Address1** and **PostalCode** fields did not contribute any value to the matching score

Record matching drill-down					
Fields Scores Contributions					
Field	Weight	Matched record terms	Pivot record terms	Score contribution	
Office	0.6	Tampa, FL	Tampa, FL	100%	
City	0.2	Tampa	Tampa	100%	

31. Click **Close**.



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



33. Notice that step 3 of the activity is to review matching results.



A matching policy can consist of multiple matching rules, and so at this step of the activity the matching would be performed over all rules.

In this lab, your matching policy consists of only the one rule, and so the results will not differ from the previous step.

34. In the **Profiler** tab, review the source statistics.

Knowledge Base Check

Lab 2-3 ► Source Statistics – New and Unique Columns

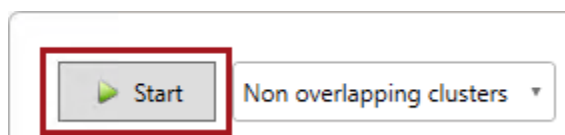
You may need data from this step to answer a lab-based Knowledge Check associated with this module.

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 the data from the **Source Statistics** results, including the **New** and **Unique** values, to refer to later.

35. Select **Non Overlapping Clusters**.



36. To start the matching process, click **Start**.



37. Review the score details for each of the non-pivot records.

Knowledge Base Check

Lab 2-3 ► Non Overlapping Clusters Score Details

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 the data from the **Score Details** results, including the **Record ID** and **Score** columns, to refer to later.

38. Open the **Matching Results** pane.



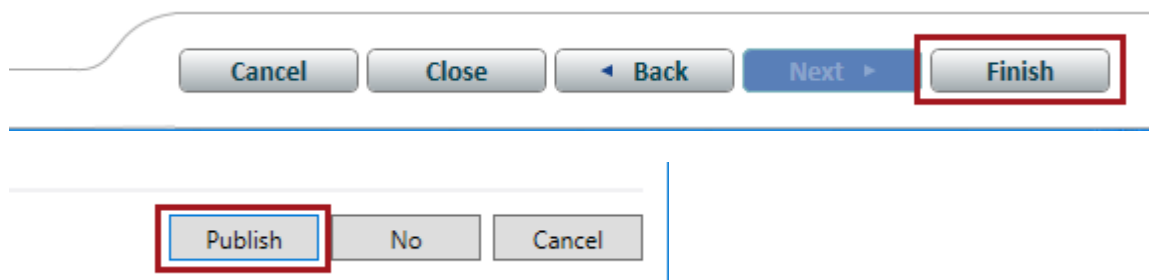
39. Review the matching results statistics.

Matching Results Statistics:

Records:	60
■ Matched:	9 (15%)
■ Unmatched:	51 (85%)
Clusters:	4
Average cluster size:	2.25
Min. cluster size:	2
Max. cluster size:	3



40. Finish the matching policy activity, and publish the knowledge base.



Exercise 3: De-duplicating Data with a Data Quality Project

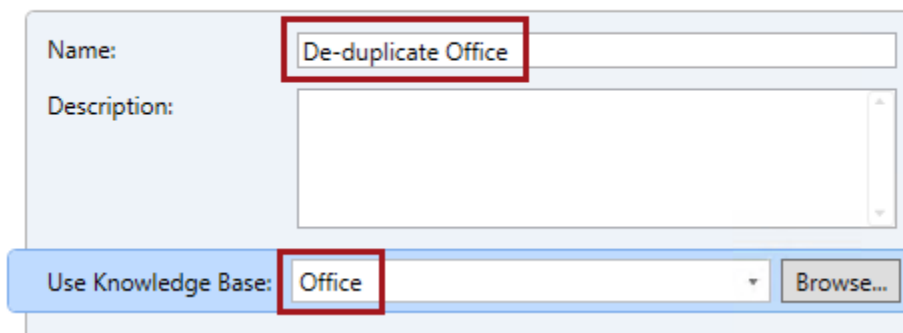
In this exercise, you will use a Data Quality Project matching activity to identify duplicate records stored in the **DimOffice** table.

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

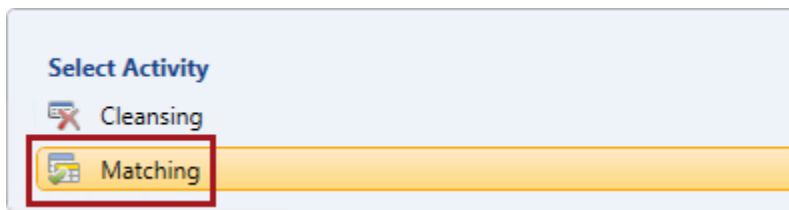


2. In the **Name** box, enter **De-duplicate Office**.
3. In the **Use Knowledge Base** dropdown list, select **Office**.

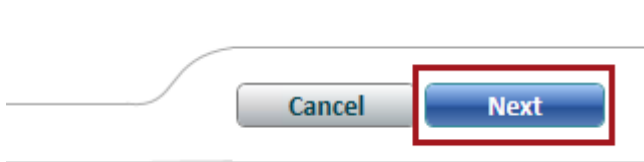
New Data Quality Project

The screenshot shows a form titled 'New Data Quality Project'. It has two main sections: 'Name:' and 'Description:'. The 'Name:' section has a text box containing 'De-duplicate Office', which is highlighted with a red rectangular box. The 'Description:' section has a large empty text area. At the bottom, there is a 'Use Knowledge Base:' section with a dropdown menu showing 'Office', which is also highlighted with a red rectangular box. To the right of the dropdown is a 'Browse...' button.

4. In the lower pane, select the **Matching** activity.



5. Click **Next**.



Mapping the Data to De-duplicate

In this task, you will configure the data to de-duplicate, 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 de-duplicated.

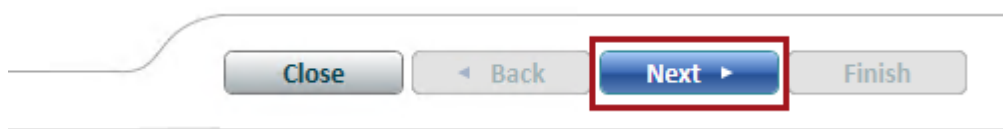


2. In the **Database** dropdown list, select **Lab**.
3. In the **Table/View** dropdown list, select **DimOffice**.

A screenshot of a configuration form for the 'Map' step. It has three dropdown menus: 'Data Source' set to 'SQL Server', 'Database' set to 'Lab', and 'Table/View' set to 'DimOffice'. The 'Database' and 'Table/View' fields are highlighted with a red rectangle.

4. Notice that all domains used to define the matching policy are automatically mapped.

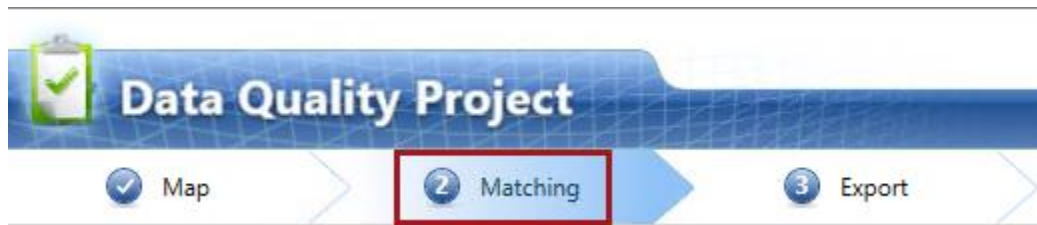
5. To proceed to the next step, click **Next**.



De-duplicating the Data

In this task, you will de-duplicate the data, and then review the profiler results.

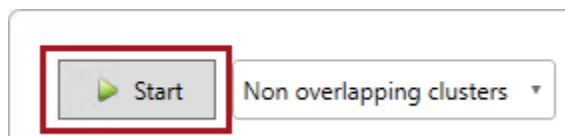
1. Notice that step 2 of the activity is to match (de-duplicate) the source data.



2. Select **Non Overlapping Clusters**.

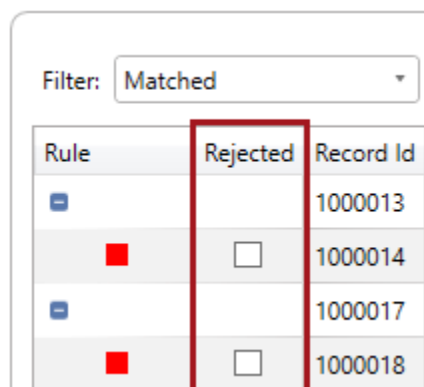


3. To start the matching process, click **Start**.



4. Review the matched records, noticing that the grid includes a **Rejected** column.

There is no need to reject any matches in this lab.

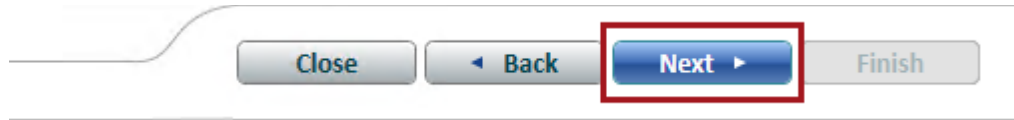


Rule	Rejected	Record Id
		1000013
	<input type="checkbox"/>	1000014
		1000017
	<input type="checkbox"/>	1000018

- Review the output in both the **Profiler** and **Matching Results** panes.

They produce the same outputs as the matching policy activity in the previous exercise.

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



Exporting the Cleansing Results

In this task, you will export the cleansing results.

- Notice that step 3 of the activity is to export the cleansing results.



- To export the results, in the **Database Name** dropdown list, select **Lab**.

A screenshot of the export configuration form. It has two dropdown menus. The first is labeled 'Destination Type:' and has 'SQL Server' selected. The second is labeled 'Database Name:' and has 'Lab' selected. The 'Database Name' dropdown is highlighted with a red rectangular box.

- In the **Content to Export** group, check the **Matching Results** checkbox, and in the corresponding **Table Name** box, enter **Lab2-3-MatchingResults**.

*For your convenience and accuracy, you can copy the table names from the **F:\Labs\Lab2-3\Assets\Snippets.txt** file (open with Notepad).*

- Check the **Survivorship Results** checkbox, and in the corresponding **Table Name** box, enter **Lab2-3-SurvivorshipResults**.

A screenshot of the 'Content to export' section. It shows two rows. The first row has a checked checkbox for 'Matching Results' and a text box containing 'Lab2-3-MatchingResults'. The second row has a checked checkbox for 'Survivorship Results' and a text box containing 'Lab2-3-SurvivorshipResults'. The entire section is highlighted with a red rectangular box.

- Notice—but do not change—the selected **Survivorship Rule** option.

Survivorship Rule

- ☒ Pivot record
- ☐ Most complete and longest record
- ☐ Most complete record
- ☐ Longest record

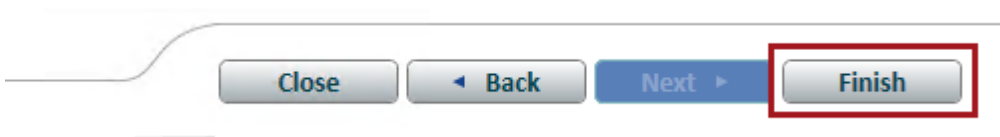
- Click **Export**.



- When notified that the export to database has completed, click **Close**.



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

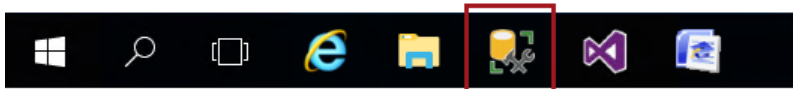


- Review the activity monitoring, and notice the cleansing activity you have just completed.

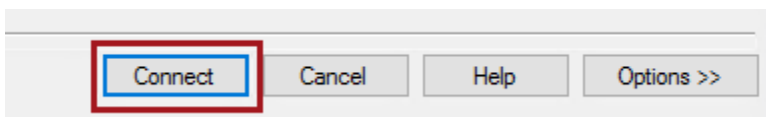
Analyzing the Cleansing Results

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

- Open SQL Server Management Studio.



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



- To open a script file, on the **File** menu, select **Open | File**.
- In the **Open File** window, navigate to the **F:\Labs\Lab2-3\Assets** folder.

5. Select the **Script-01-ReviewMatchingOutputs.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.

```
1 --Execute INDIVIDUAL batches as directed
2
```

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

*You have now completed the lab. This is the final lab in the course, and so you should now complete the **Finishing Up** exercise to delete the VM.*

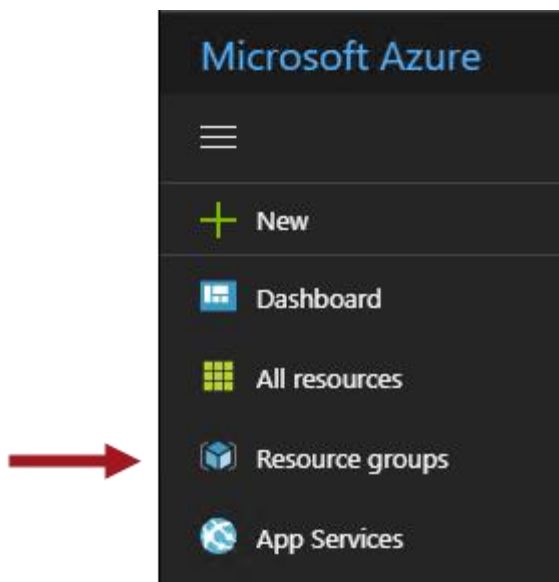
Finishing Up

In this exercise, you will delete the **Lab** resource group, which will delete the VM.

Knowledge Base Check Before You Move On

Before deleting the resource group, it is recommended that you open your Knowledge Check portion of the course within EdX and answer any outstanding end-of-module questions for modules 1-2.

1. Close the remote desktop window.
2. In the **Azure Portal** browser page, select **Resource Groups**.



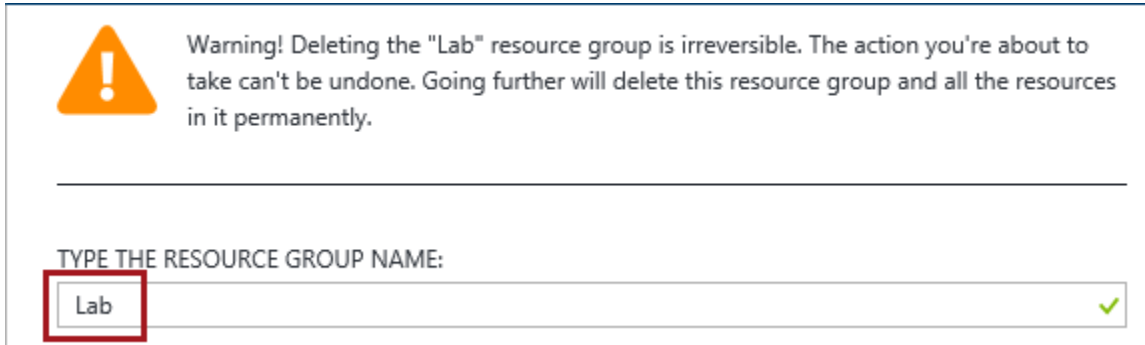
3. In the **Resource Groups** blade, select the **Lab** resource group.



4. In the **Lab** blade, click **Delete Resource Group**.



- When prompted to delete the resource group, in the **Type the Resource Group Name** box, enter **Lab**.



A warning dialog box with a yellow triangle icon containing an exclamation mark. The text reads: "Warning! Deleting the 'Lab' resource group is irreversible. The action you're about to take can't be undone. Going further will delete this resource group and all the resources in it permanently." Below the text is a horizontal line. Underneath the line is the label "TYPE THE RESOURCE GROUP NAME:" followed by a text input field. The input field contains the text "Lab" and has a green checkmark at the end. The "Lab" text and the input field are highlighted with a red rectangular border.

- Click **Delete**.



Two buttons are shown side-by-side. The first button is blue with the text "Delete" in white. The second button is light blue with the text "Cancel" in blue. The "Delete" button is highlighted with a red rectangular border.

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