

DAT218x

Cleansing Data with Data Quality Services

Lab 2-3 | De-duplicating Data with a Matching Policy

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

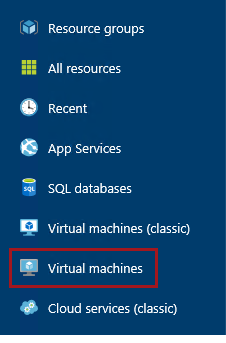
*Go to the next exercise if you are already connected to the lab VM.*

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 **Azure Portal** by using your subscription.
2. In the left pane, select **Virtual Machines**.

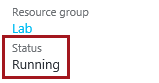


1. In the **Virtual Machines** blade, select the VM you provisioned in **Lab 0-1**.
2. In the VM blade, click **Start**.

C:\Users\PETERM~1\AppData\Local\Temp\SNAGHTML11734f49.PNG

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

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



1. 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.

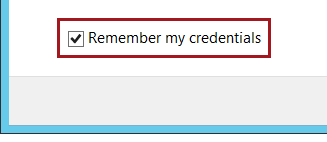
C:\Users\PETERM~1\AppData\Local\Temp\SNAGHTML1174d906.PNG

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.

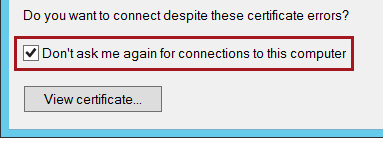
1. When prompted by the web browser to open the Remote Desktop File, click **Open**.

C:\Users\PETERM~1\AppData\Local\Temp\SNAGHTML1147abe6.PNG

1. If prompted to connect to the unknown publisher, click **Connect**.
2. In the **Windows Security** window, enter the credentials you created for your VM.
3. Check the **Remember My Credentials** checkbox.



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



1. Click **Yes**.
2. 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.

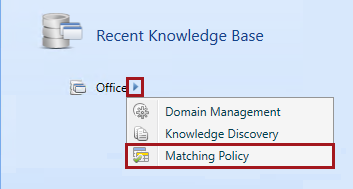
1. Open Data Quality Client.

C:\Users\PETERM~1\AppData\Local\Temp\SNAGHTML117ee0b6.PNG

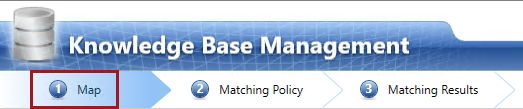
1. In the **Connect to Server** window, click **Connect**.



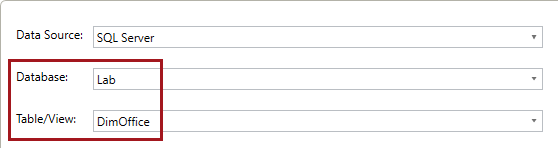
1. To create a matching policy, in the **Knowledge Base Management** panel, click the **Office** knowledge base, and then select the **Matching Policy** activity.



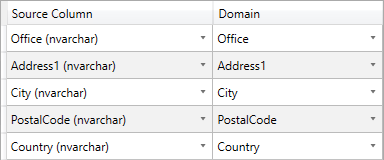
1. Notice that step 1 of the activity is to connect to sample data to create matching policy rules.



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



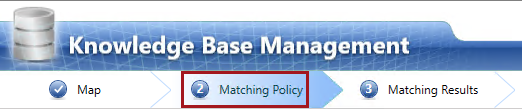
1. Create the following five mappings from source column to domain.



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

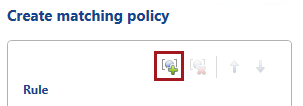


1. Notice that step 2 of the activity is to create a matching policy.

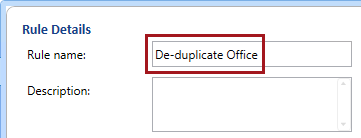


Only one matching policy can be created per knowledge base.

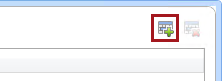
1. Click **Create a Matching Rule**.



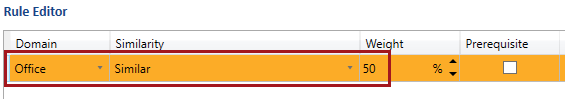
1. In the **Rule Name** box, replace the text with **De-duplicate Office**.



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



1. Configure the domain element based on the following image.

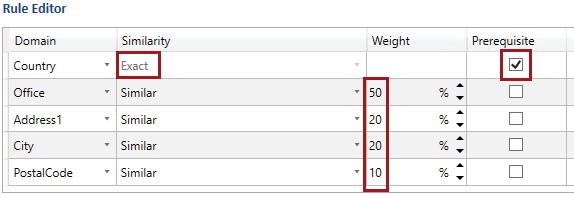


The domain element ensures that the **Office** value can be similar, and its similarity score will contribute 50% to the matching score.

1. 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.

1. Configure additional domain elements based on the following image (order is not important).



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.

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

C:\Users\PETERM~1\AppData\Local\Temp\SNAGHTML396299f.PNG

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

1. 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.

1. 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.



1. Select **Non Overlapping Clusters**.

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



1. Click **Restart**.



1. 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.

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

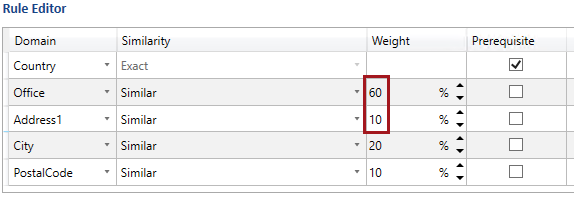


1. In the grid, to sort the data, click the **Office** column header.
2. Notice the two **Tampa, FL** records, and that there **Address1** values are not quite similar.
3. In the **Filter** dropdown list, select **Matched**.



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

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



1. Click **Restart**.

C:\Users\PETERM~1\AppData\Local\Temp\SNAGHTML3a0ae4f.PNG

1. Notice that four clusters were now detected.

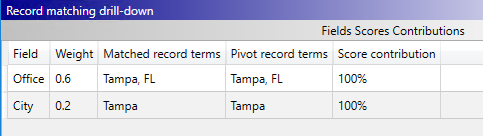
Creating a matching rule is a process of trial-and-error, ultimately arriving at an optimal set of rules to detect duplicate records.

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

C:\Users\PETERM~1\AppData\Local\Temp\SNAGHTML3a71f4b.PNG

1. Review the score details, noting the following:

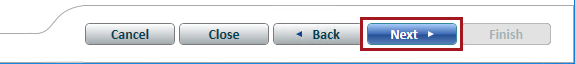
* The matching score is 80%
* The fields that contributed to the score were **Office** and **City**, with exact matches encountered, and so they contributed 60% (0.6 x 100%) and 20% (0.2 x 100%) respectively
* The **Address1** and **PostalCode** fields did not contribute any value to the matching score



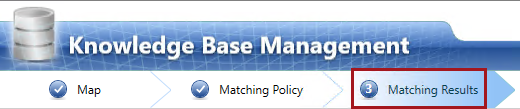
1. Click **Close**.



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



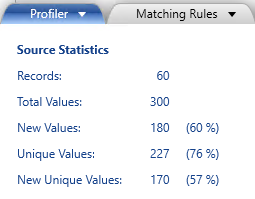
1. Notice that step 3 of the activity is to review matching results.



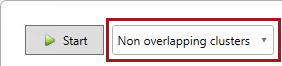
A matching policy can consist of multiple matching rules, and so at this step of the activity the matching would be performed over all rules.

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

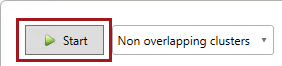
1. In the **Profiler** tab, review the source statistics.
2. These are the result of cleansing the data with the knowledge base.



1. Select **Non Overlapping Clusters**.



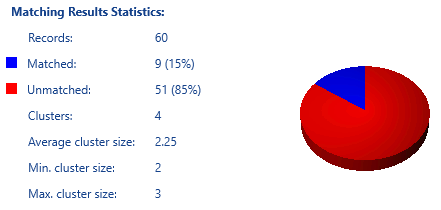
1. To start the matching process, click **Start**.



1. Review the score details for each of the non-pivot records.
2. Open the **Matching Results** pane.



1. Review the matching results statistics.



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





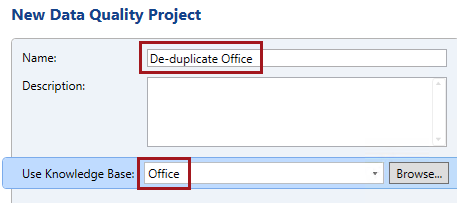
# Exercise 3: De-duplicating Data

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

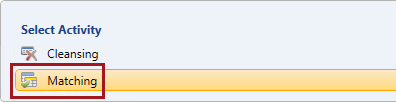
1. To create a new Data Quality Project, in the **Data Quality Projects** panel, click **New Data Quality Project**.



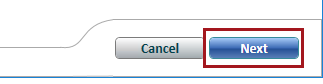
1. In the **Name** box, enter **De-duplicate Office**.
2. In the **Use Knowledge Base** dropdown list, select **Office**.



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



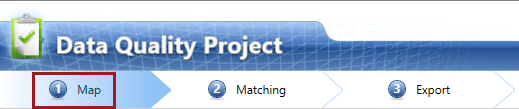
1. Click **Next**.



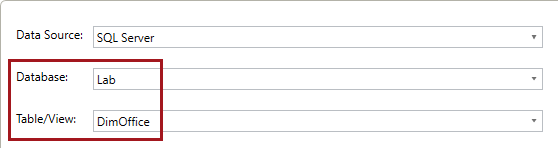
## Mapping the Data to De-duplicate

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.



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



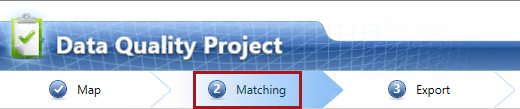
1. Notice that all domains used to define the matching policy are automatically mapped.
2. To proceed to the next step, click **Next**.



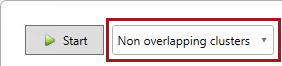
## De-duplicating the Data

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

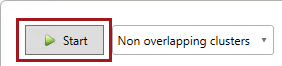
1. Notice that step 2 of the activity is to match (de-duplicate) the source data.



1. Select **Non Overlapping Clusters**.

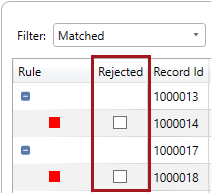


1. To start the matching process, click **Start**.



1. Review the matched records, noting that the grid includes a **Rejected** column.

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



1. 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.

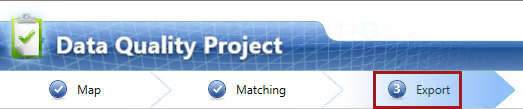
1. To proceed to the next step, click **Next**.



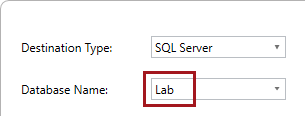
## Exporting the Cleansing Results

In this task, you will export the cleansing results.

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



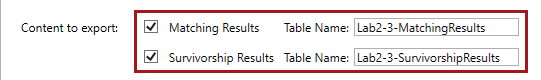
1. To export the results, in the **Database Name** dropdown list, select **Lab**.



1. 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).

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



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



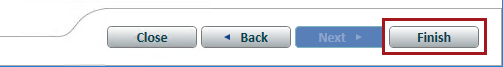
1. Click **Export**.



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



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



1. 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.

C:\Users\PETERM~1\AppData\Local\Temp\SNAGHTML119fc52f.PNG

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



1. To open a script file, on the **File** menu, select **Open | File**.
2. In the **Open File** window, navigate to the **F:\Labs\Lab2-3\Assets** folder.
3. Select the **Script-01-ReviewMatchingOutputs.sql** file, and then click **Open**.
4. In the script file, take note of the first line.

C:\Users\PETERM~1\AppData\Local\Temp\SNAGHTML11b626bc.PNG

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**).

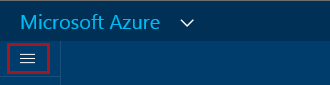
1. Read the comments in the first batch (lines 3-5).
2. Select and execute the only query in the batch (lines 6-7).
3. Read the commented text, and then execute the query for each of the remaining batches in the script.

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.

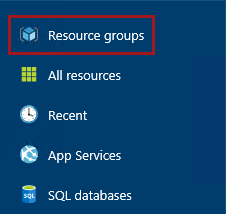
# Finishing Up

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

1. Close the remote desktop window.
2. In the **Azure Portal** Web browser page, open the left pane.



1. Select **Resource Groups**.



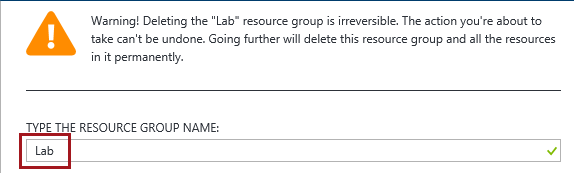
1. In the **Resource Groups** blade, select the **Lab** resource group.



1. In the **Lab** blade, click **Delete**.

C:\Users\PETERM~1\AppData\Local\Temp\SNAGHTML3f497c1.PNG

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



1. Click **Delete**.



1. Sign out of the **Azure Portal**.