

DAT218x

Cleansing Data with Data Quality Services

Lab 1-2 | Creating a Knowledge Base

Estimated time to complete this lab is 100 minutes

Overview

In this lab, you will be introduced to the lab scenario that involves the cleansing of a dataset of North American Microsoft office locations.

You will gain a complete understanding of the dataset by using Integration Services (SSIS) data profiling. You will then create a knowledge base to address many of the data quality requirements.

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

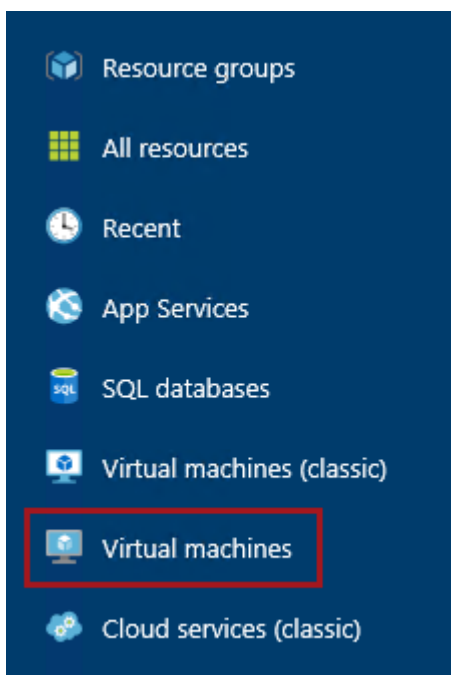
Go to the next exercise if you are already connected to the lab VM.

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

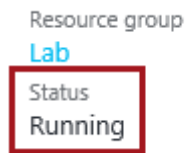


3. In the **Virtual Machines** blade, select the VM you provisioned in **Lab 0-1**.
4. In the VM blade, click **Start**.



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

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



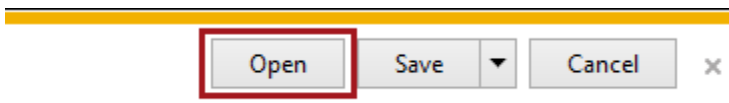
- To connect to the VM, click **Connect**.

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

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



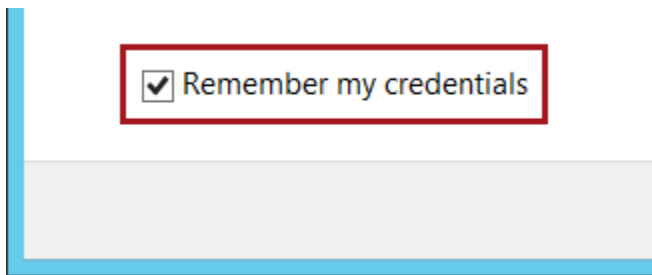
- If prompted to connect to the unknown publisher, click **Connect**.

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

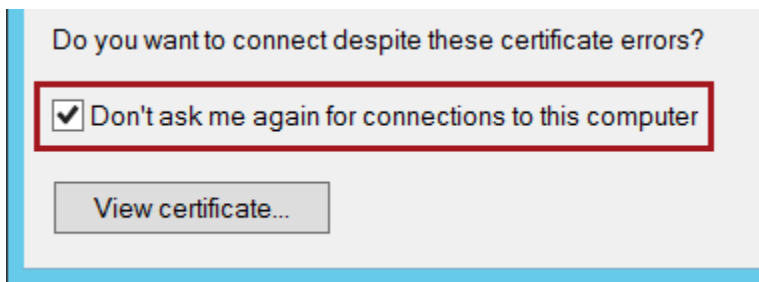


- In the **Windows Security** window, enter the credentials you created for your VM.

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



11. Click **OK**.
12. In the **Remote Desktop Connection** 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: Exploring the Office Dataset

In this exercise, you will be introduced to the office dataset that includes records for 70 Microsoft offices located in the United States and Canada.¹

Specifically, each office is described in terms of its name, district, address and phone number, and also the office manager who is described in terms of their first and last names, title and email address.

The labs in this course are designed address data quality issues in this dataset. In this lab you will create a knowledge base to address many of the data quality issues. In later labs of this course, you will apply the knowledge base to cleanse and de-duplicate the dataset.

The following is a list of the dataset columns, and also a description of each column together with a list of the data quality requirements.

Column	Purpose	Data Quality Requirements
Office	The complete name of the office, often named as the city location (e.g. Pittsburgh, PA)	<ul style="list-style-type: none">• Cannot be missing
District	A geographic grouping of offices (e.g. Mid Atlantic District)	<ul style="list-style-type: none">• Cannot be missing• The word "District" must not be abbreviated
Address1	The office first address line.	<ul style="list-style-type: none">• Cannot be missing
Address2	The office second address line, typically containing building, floor or suite details—if required	
City	The office city name	<ul style="list-style-type: none">• Cannot be missing

Table continued on the next page...

¹ The office dataset was sourced from:

- <https://www.microsoft.com/en-us/about/companyinformation/usaoffices/default.aspx>
- <https://www.microsoft.com/en-ca/about/locations.aspx>

In some instances, the office details have been modified to intentionally create data quality issues. Also, the dataset has been extended with office manager details which have been sourced from the Adventure Works sample databases, and so this data is fictitious. Microsoft do not support the use of this dataset outside of the labs in this course.

Column	Purpose	Data Quality Requirements
StateOrProvince	The office state or province. For United States offices, this is a state code. For Canadian offices, this is a province or territory code.	<ul style="list-style-type: none"> • Cannot be missing • Must conform to the two-letter state codes as used by the United States Postal Service, or Canada Post • Must be output in upper case format
PostalCode	The office postal code. For United States offices, this is the ZIP Code. For Canadian offices, this is the postal code.	<ul style="list-style-type: none"> • Cannot be missing • For the United States, this must either be the five-digit ZIP Code (e.g. 15212), or the extended ZIP+4 Code (e.g. 15212-1234) • For Canada, this must be a six-alphanumeric string conforming to the sort rules, with a mandatory space separating two groups of three characters (e.g. T2P 0T1)
Country	The office country	<ul style="list-style-type: none"> • Cannot be missing • The country name must be spelled in full (i.e. either United States, or Canada)
Phone	The office telephone number	<ul style="list-style-type: none"> • If missing, the telephone number should be set to the toll-free 800 number • Must be expressed with the area code enclosed within parentheses, followed by a space, followed by three digits, followed by a hyphen, followed by the remaining four digits (e.g. (412) 323-6700)
ManagerFirstName	The office manager first name	<ul style="list-style-type: none"> • Cannot be missing • Cannot be expressed as only an initial

Table continued on the next page...

Column	Purpose	Data Quality Requirements
ManagerLastName	The office manager last name	<ul style="list-style-type: none"> Cannot be missing Cannot be expressed as only an initial
ManagerTitle	The office manager title	<ul style="list-style-type: none"> Cannot be missing Valid values are either "Mr.", "Miss.", "Ms.", or "Mrs."
ManagerEmail	The office manager email	<ul style="list-style-type: none"> Cannot be missing Must belong to the email domain "lab.microsoft.com" Must be output in lower case format

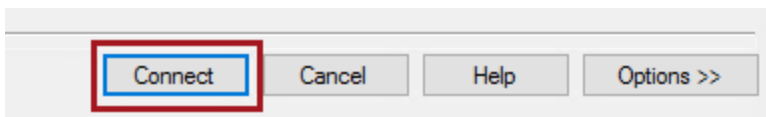
Exploring the Office Dataset

In this task, you will query and explore the data in the office dataset.

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\Lab1-2\Assets** folder.
5. Select the **Script-01-ExploreOfficeDataset.sql** file, and then click **Open**.

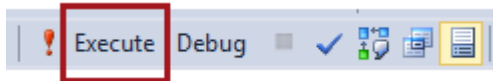
Within a lab, all script file names follow a numeric sequence, and contain a brief description of their purpose.

6. 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**.



7. Select and execute the only batch in the script (lines 7-8).

Do not spend a lot of time trying to understand this dataset, as you will apply data profiling techniques in the next exercise to easily and quickly understand the data.

8. To close SQL Server Management Studio, on the **File** menu, select **Exit**.
9. If prompted to save changes, click **No**.

Exercise 3: Data Profiling the Office Dataset

In this task, you will open an Integration Services (SSIS) project, and then design a package to data profile the Office dataset.

Opening the SSIS Project

In this task, you will open an Integration Services (SSIS) project.

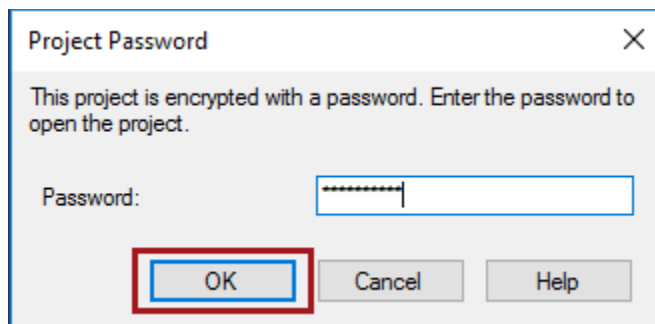
1. Open SQL Server Data Tools.



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\Lab1-2\Assets\Project** folder.
4. Select **Lab.sln**, and then click **Open**.
5. In the **Project Password** window, in the **Password** box, enter **Pass@word1**. (Do not enter the period.)

The sensitive content of the SSIS project and packages is protected by this password.

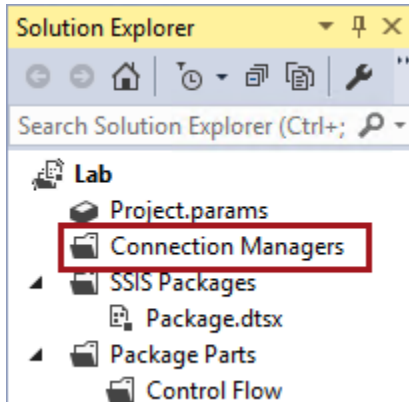
6. Click **OK**.



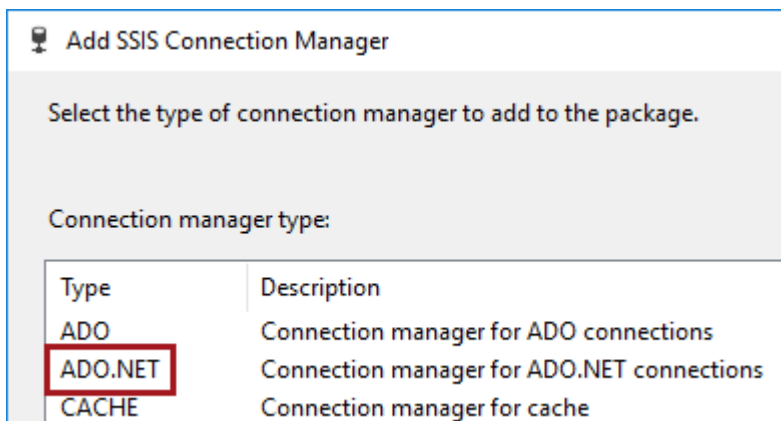
Creating a Connection Manager

In this task, you will create a connection manager to the **Lab** database.

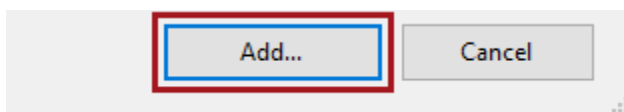
1. To create a connection manager, in **Solution Explorer** (located at the right), right-click the **Connection Managers** folder, and then select **New Connection Manager**.



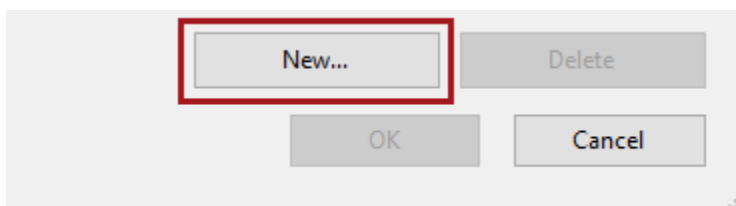
2. In the **Add SSIS Connection Manager** window, select the **ADO.NET** connection manager type.



3. Click **Add**.



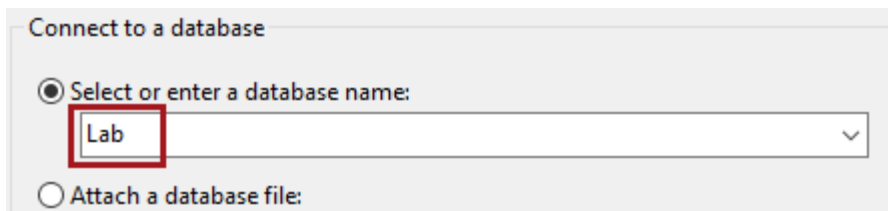
4. In the **Configure ADO.NET Connection Manager** window, click **New**.



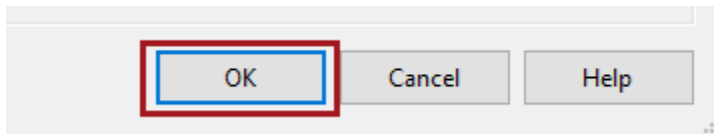
5. In the **Connection Manager** window, in the **Server Name** dropdown list—do not click the dropdown arrow—enter **localhost**.



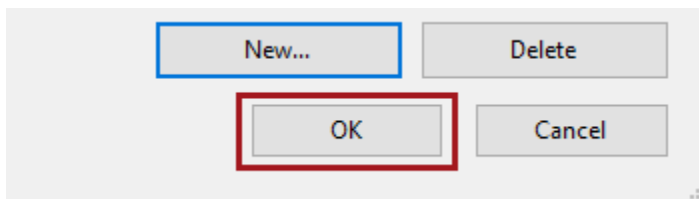
6. In the **Select or Enter a Database Name** dropdown list, select **Lab**.



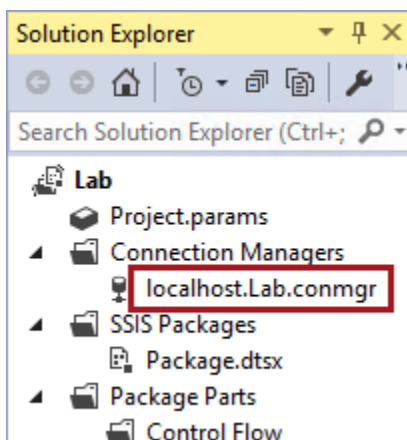
7. Click **OK**.



8. In the **Add ADO.NET Connection Manager** window, click **OK**.



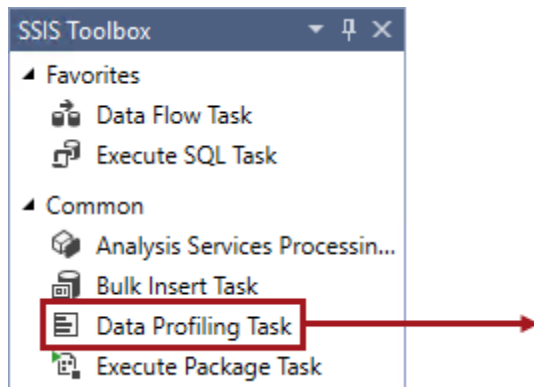
9. In **Solution Explorer**, notice the addition of the connection manager.



Design the Data Profiling Package

In this task, you will design the **Data Profiling** package.

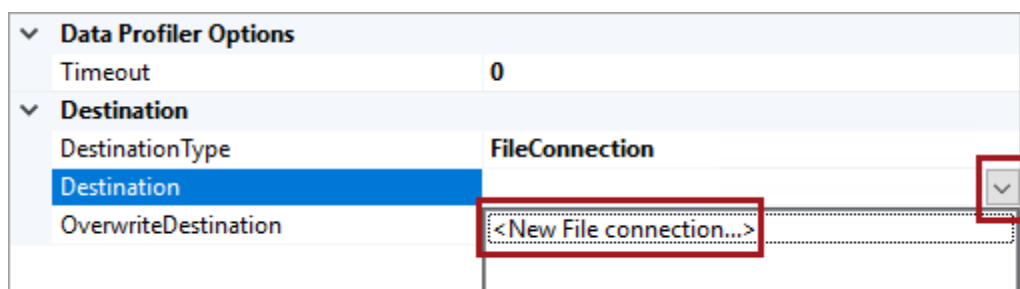
1. To rename the existing package, in **Solution Explorer**, right-click the **Package.dtsx** file, and then select **Rename**.
2. Rename the package to **Data Profiling.dtsx**, and then press **Enter**.
3. Notice that the package opens in the package designer.
4. To open the **SSIS Toolbox**, click inside the package designed, and then on the **SSIS** menu, select **SSIS Toolbox**.
5. To design the package, from the **SSIS Toolbox** (located at the left), from inside the **Common** group, drag the **Data Profiling Task** to the package designer.



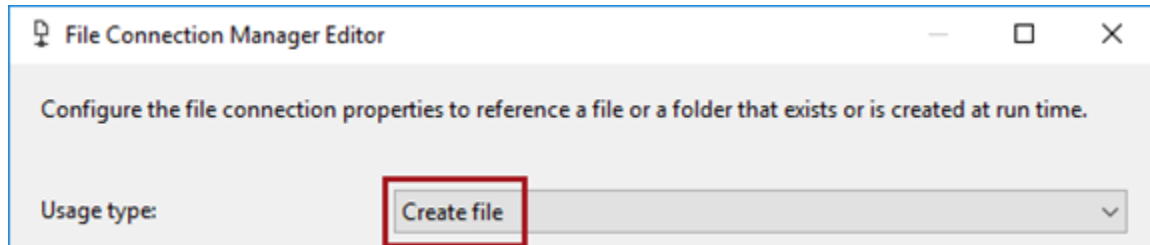
6. To edit the task, in the package designer, right-click the **Data Profiling Task**, and select **Edit**.

You will configure the task to perform data profiling of the Office dataset. It will achieve this by outputting an XML file to the file system, and you will then explore this output by using the Data Profile Viewer application.

7. In the **Data Profiling Task Editor** window, select the **Destination** property, then click the down-arrow, and then select **<New File Connection>**.

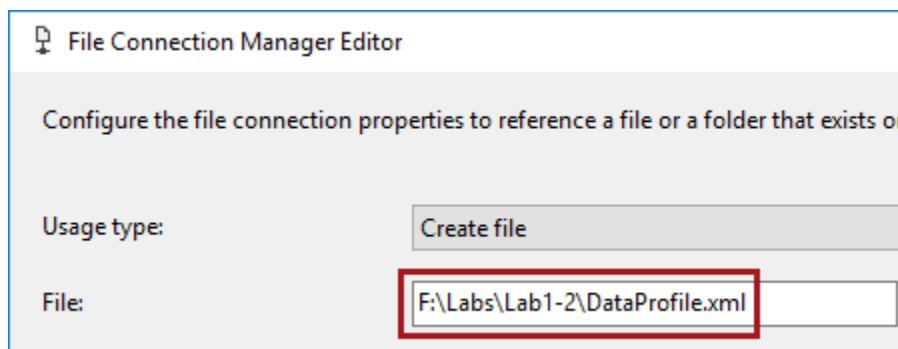


8. In the **File Connection Manager Editor** window, in the **Usage Type** dropdown list, select **Create File**.

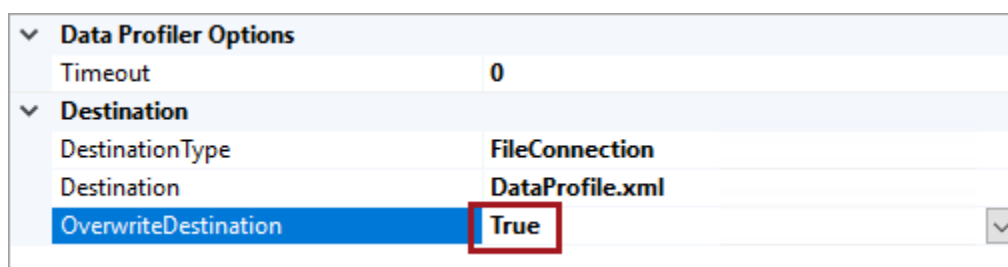


9. In the **File** box, enter the **F:\Labs\Lab1-2\DataProfile.xml** file path.

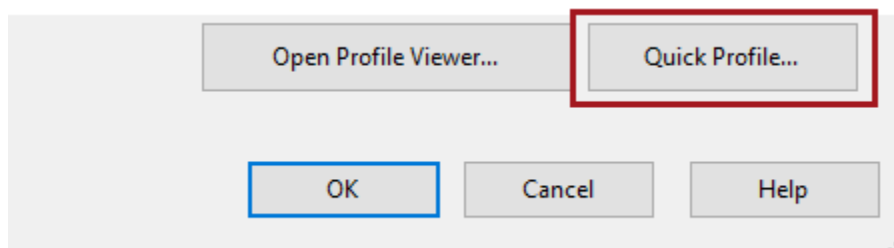
*You can type the file path, or click **Browse** to browse to the file, or for your convenience, you can copy the file path from the **F:\Labs\Lab1-2\Assets\Snippets.txt** file (open with Notepad).*



10. Click **OK**.
11. Set the **OverwriteDestination** property to **True**.



12. To configure the profile requests, click **Quick Profile**.



13. In the **Single Table Quick Profile Form** window, configure the following.

Single Table Quick Profile Form

You can profile a table or view on all applicable columns using default settings. Choose the table and the profiles you want.

ADO.NET Connection: localhost.Lab New...

Table or View: [dbo].[MSFTOffice_NorthAmerica]

Compute:

- ☒ Column Null Ratio Profile
- ☒ Column Statistics Profile
- ☒ Column Value Distribution Profile
- ☒ Column Length Distribution Profile
- ☒ Column Pattern Profile
- ☒ Candidate Key Profile

for up to 1 Column keys

☐ Functional Dependency Profile

for up to 1 Columns as Determinant Columns

14. Click **OK**.

OK Cancel Help

15. In the **Data Profiling Task Editor** window, click **OK**.

OK Cancel Help

16. To execute the package, in **Solution Explorer**, right-click the **Data Profiling.dtsx** package, and then select **Execute Package**.

17. Verify that the task executed successfully.

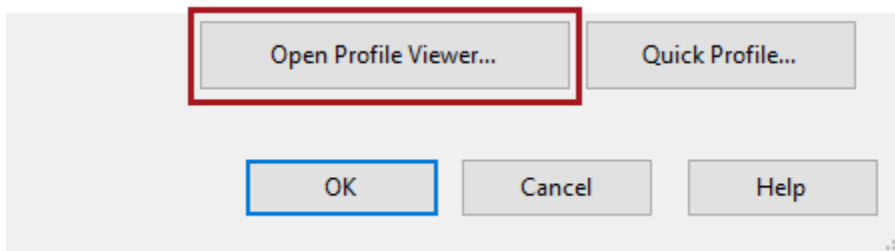


18. To stop the package debugging, on the **Debug** menu, select **Stop Debugging**.

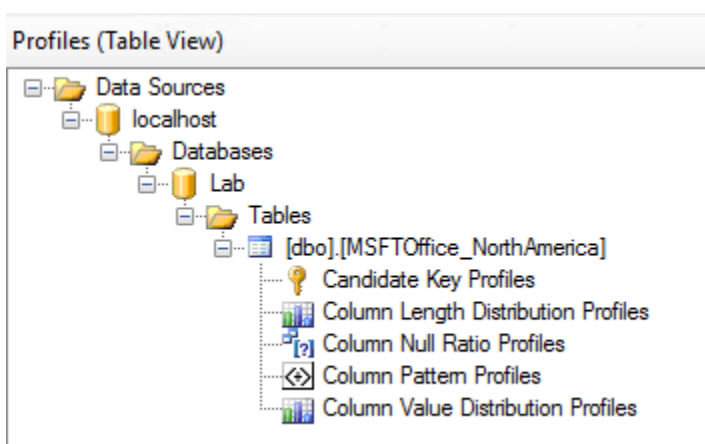
Data Profiling the Office Dataset

In this task, you will use Data Profile Viewer to gain a good understanding of the office dataset. You will commence by reviewing columns with missing values.

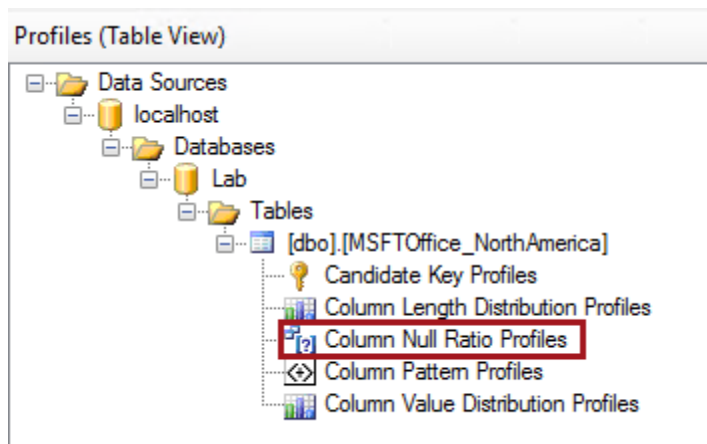
1. In the package designer, right-click the **Data Profiling Task**, and select **Edit**.
2. Click **Open Profile Viewer**.



3. Maximize the **Data Profile Viewer** window.
4. In the left pane, notice the hierarchy from data source, to database, to table.



5. In the left pane, select **Column Null Ratio Profiles**.



6. In the right pane, click the **Null Count** column header twice to sort the values in descending order.

Knowledge Base Check

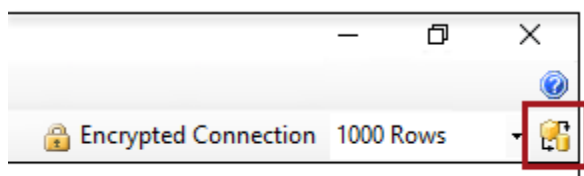
Lab 1-2 ► Null Percentage

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 data from the **Column**, **Null Count** and **Null Percentage** columns to refer to later.

7. Review the column null counts and percentages.
8. Note the following:
- Most **Address2** values are missing, and this is acceptable (i.e. not all addresses describe a suite, floor or building number)
 - There are three missing **Phone** values, and this is not acceptable
 - There is one missing **Country** value, and this is not acceptable
 - There is one missing **District** value, and this is not acceptable
9. To view the records with missing **Phone** values, select the row representing the **Phone** column.

10. Located at the top-right corner, click **Drill Down**.



11. In the lower pane, review the three records.

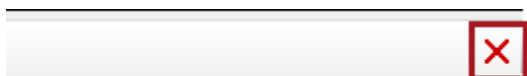
Knowledge Base Check

Lab 1-2 ► Column Null Ratio Profiles

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 data from the **Address1** column, and then related **City** column to refer to later.

12. To close the drill down result, inside the lower pane, click the red **X**.

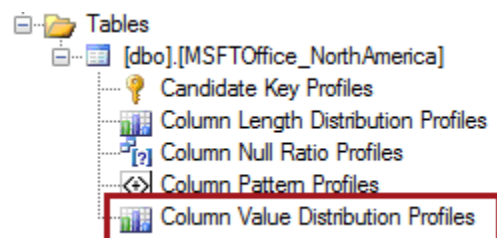


13. Optionally, drill down to the other records with missing values.

Data Profiling for Column Values

In this task, you will use the Data Profile Viewer application to review column values.

1. In the left pane, select **Column Value Distribution Profiles**.



- To sort the columns, in the right pane, click the **Number of Distinct Values** column header twice.

Column	Number Of Distinct Values
Address1	69
PostalCode	68
Office	66
ManagerEmail	64
City	63

Knowledge Base Check

Lab 1-2 ► Number of Distinct Values

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 **Number of Distinct Values** column and associated data, to refer to later.

- At a glance, it is possible to quickly understand which columns contain unique, or near-unique, values.

Recall that there are 70 office records.

- Select the **Office** column.
- To sort the columns in the frequency list, click the **Count** column header twice.
- Notice that there are four duplicate office names, and that this is not acceptable.

Value	Count	Percentage
Tampa, FL	2	2.8571 %
Reston, VA	2	2.8571 %
New York, NY	2	2.8571 %
Charlotte, NC	2	2.8571 %
Bentonville, AR	1	1.4286 %

*You will identify and remove duplicate records in **Lab 2-3**.*

7. Drill down to the two **Tampa, FL** records, and notice that they have different addresses.

Knowledge Base Check

Lab 1-2 ► Tampa FL Address1

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 data for the **Address1** column, along with the **Number of Distinct Values** column, to refer to later.

8. In the **Column Values Distribution Profiles** pane (top pane), select the **District** column.
9. Note the following:
 - Several district names include the abbreviation **Dist.**, and this is not acceptable
 - **Greater South East District**, and **Greater Southeast District** are the same district (the first is misspelled)
 - **Mid Atlantic District**, and **Mid-Atlantic District** are variant spellings of the same district
 - **Midwestt Dist.** is misspelled and abbreviated
10. In the top pane, select the **StateOrProvince** column.
11. Note the following:
 - Not all values are in the two-character format, with some abbreviations, and also a full spelling for **California**, and this is not acceptable
 - Some values are in lower case, and this is not acceptable
12. In the top pane, select the **Country** column, and note that there should only be two values: **United States** and **Canada**.
13. In the top pane, select the **Phone** column.
14. Scroll to the end of the values list, and note that two telephone numbers are not in the required format which includes parentheses.

15. Note also, when you sort the value list by **Count** descending, that there are duplicate phone numbers.

When duplicates of identifiers—like account numbers, Social Security Numbers, telephone numbers and email addresses—are encountered, this is usually strong evidence that duplicate records exist.

Knowledge Base Check

Lab 1-2 ► Duplicate Phone Numbers

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

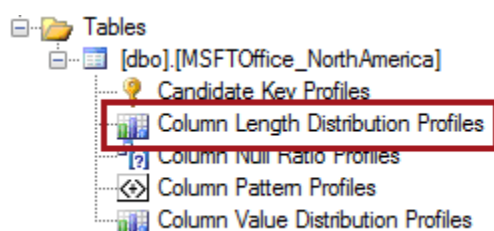
It is recommended that you open your Knowledge Check portion of the course in EdX at this time to answer questions as you complete the lab, take a screenshot of the data for the descending sorted **Value** column, along with the **Count** and **Percentage** columns, to refer to later.

16. In the top pane, select the **ManagerTitle** column.
17. Note that all four values are acceptable terms, and so no cleansing will be required for this column.

Data Profiling for Column Value Lengths

In this task, you will use Data Profile Viewer to review column value lengths.

1. In the left pane, select **Column Length Distribution Profiles**.



2. In the top pane, select the **PostalCode** column.
3. Review the length distributions, and note that most records have a value length of either five or seven characters, and these are likely to be valid US (five-digit) or Canadian (seven-character) values.

4. Drill down to review the records that have values lengths other than five or seven, and note the following:
 - The four-digit value 8830 needs to be corrected to 08830
 - The six-character value L5N8L9 requires that a space be inserted in the middle
 - The 10-character value is an acceptable ZIP+4 Code
5. In the top pane, select the **ManagerFirstName** column.
6. Note that the one record with a value length of one character is an initial, and not a full name, and that this is not acceptable.

Knowledge Base Check

Lab 1-2 ► PostalCode Length Distribution

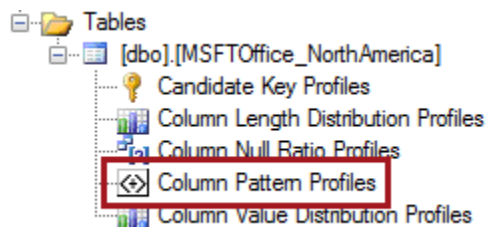
You may need data from this step to answer a lab-based Knowledge Check associated with this module.

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 data for the **Length**, **Count** and **Percentage** columns, to refer to later.

Data Profiling for Column Patterns

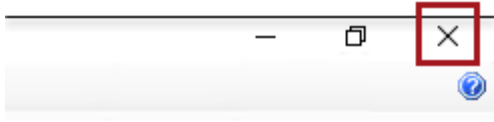
In this task, you will use Data Profile Viewer application to review column patterns.

1. In the left pane, select **Column Pattern Profiles**.

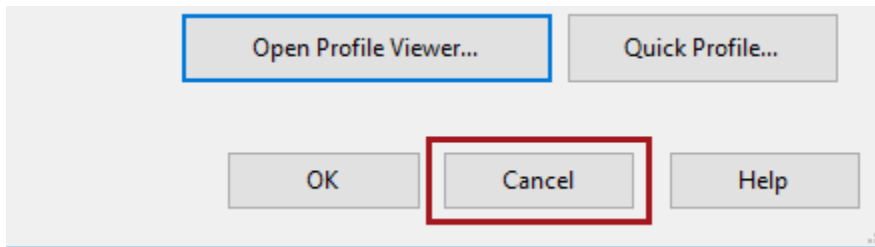


2. In the top pane, select the **PostalCode** column.
3. Note that two patterns were detected, representing 95% of all column values.
4. Note the following:
 - The **\d\d\d\d\d** pattern represents five digits (numbers), and are valid ZIP Codes
 - The **\w\w\w \w\w\w** (there is an embedded space between the two groups of three letters) pattern represents valid Canadian postal codes
5. In the top pane, select the **Phone** column.
6. Note the pattern found for 97% of telephone numbers.

7. In the top pane, select the **ManagerEmail** column.
8. Note the pattern found for 90% of email addresses, which describes the correct email domain (**lab.microsoft.com**).
9. To close Data Profile Viewer, at the top-right corner, click **X**.



10. In SQL Server Data Tools, to close the task editor, click **Cancel**.



11. To close SQL Server Data Tools, on the **File** menu, select **Exit**.
12. When prompted to save changes, click **Yes**.

Exercise 4: Creating a Knowledge Base

In this exercise, you create a knowledge base to address many of the data quality requirements for the office dataset.

Creating the Knowledge Base

In this task, you will create the knowledge base by using the domain management activity.

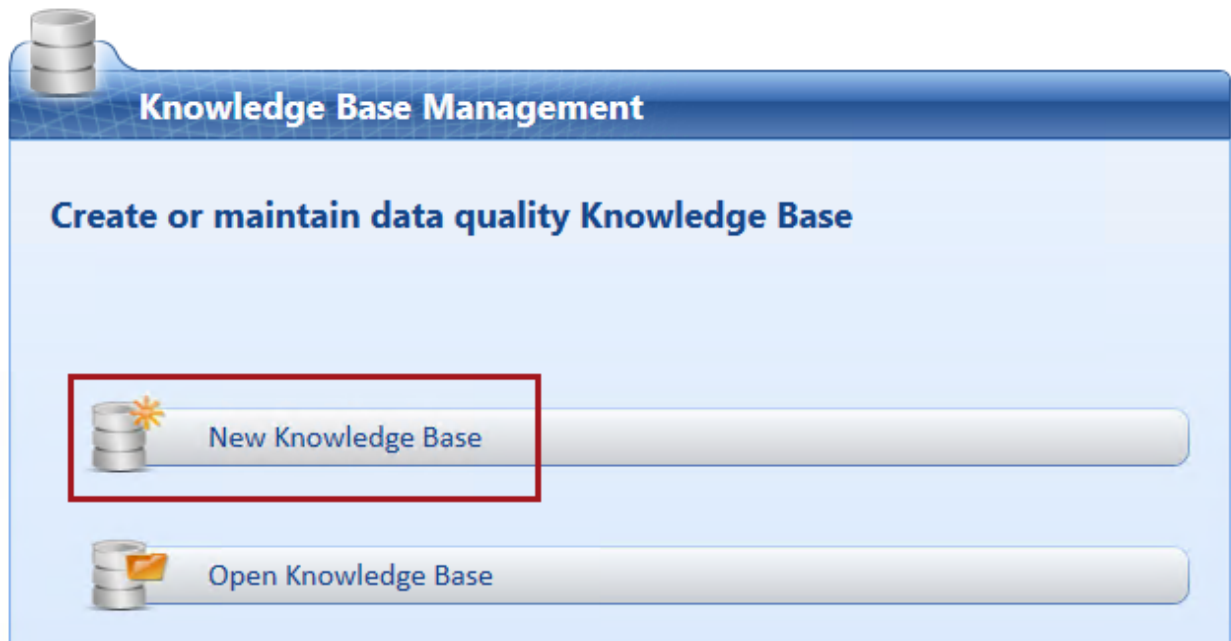
1. Open Data Quality Client.



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



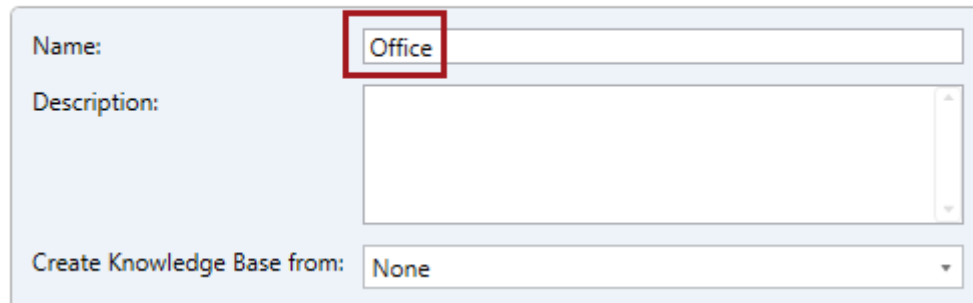
3. To create a new knowledge base, in the **Knowledge Base Management** panel, click **New Knowledge Base**.



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

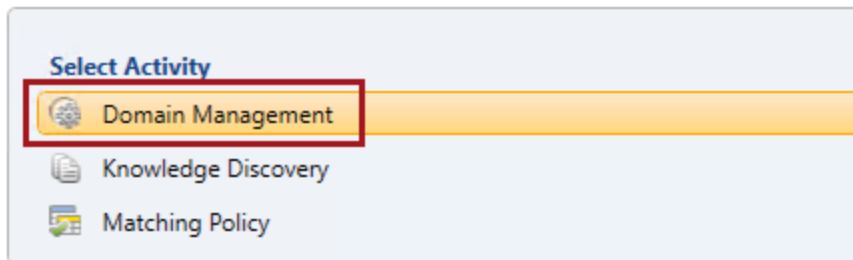
Important: When naming objects in this lab, be sure to enter the names exactly as the lab describes. Incorrect name values may result in errors later in the lab.

New Knowledge Base

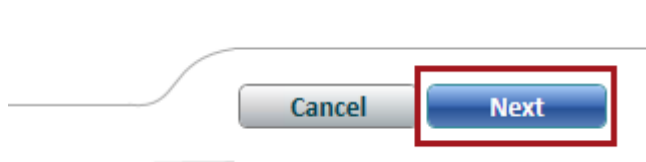


The 'New Knowledge Base' dialog box has a light blue header. Below the header, there are three input fields. The first field is labeled 'Name:' and contains the text 'Office'. The second field is labeled 'Description:' and is empty. The third field is labeled 'Create Knowledge Base from:' and contains the text 'None'. A red rectangle highlights the 'Name' field.

5. In the lower pane, notice that the **Domain Management** activity is selected.



6. Click **Next**.



Creating the Domains

In this task, you will create the knowledge base domains.

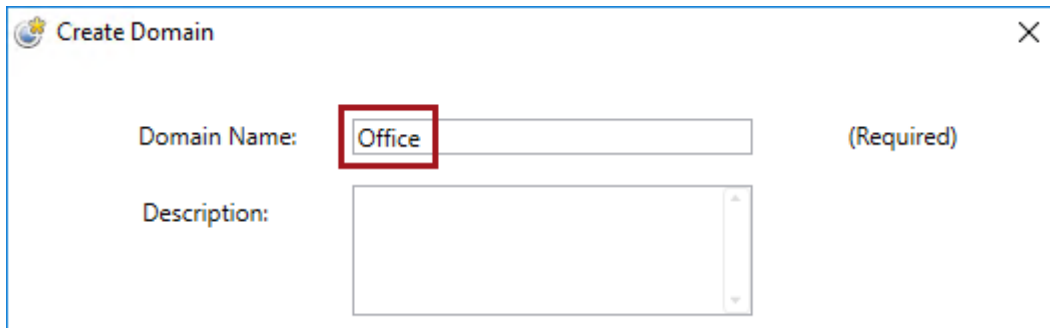
1. To create a domain, click **Create a Domain**.

Tip: In Data Quality Client, commands are available either as icons, or right-click context menus. To determine what an icon does, hover the cursor over it to reveal a tooltip.

Domain Management

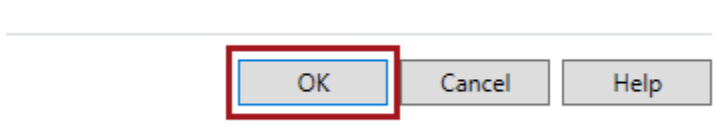


2. In the **Create Domain** window, in the **Domain Name** box, enter **Office**.



There are many domain properties that can also be set when creating the domain, and these can be modified at any time during domain management.

3. Click **OK**.



4. Create the following additional 12 domains.

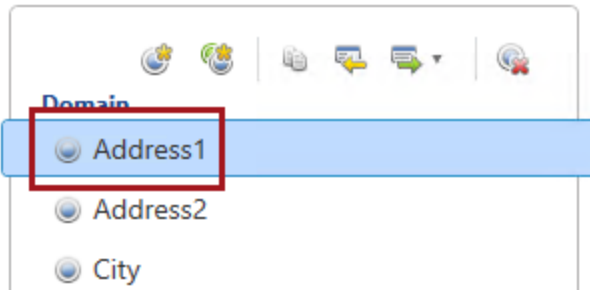
Reminder: Take care to name the domains exactly as the lab describes.

- District
- Address1
- Address2
- City
- StateOrProvince
- PostalCode
- Country
- Phone
- ManagerFirstName
- ManagerLastName
- ManagerTitle
- ManagerEmail

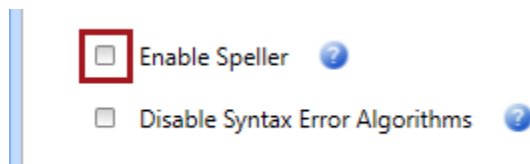
5. Verify that you have 13 domains.

6. Select the **Address1** domain.

Important: It is a common mistake to configure the wrong domain, which later involves determining which domain to undo (there is no Ctrl-Z to undo). Always take care to select the correct domain before configuring it.



7. In the **Domain Properties** tab, uncheck **Enable Speller**.



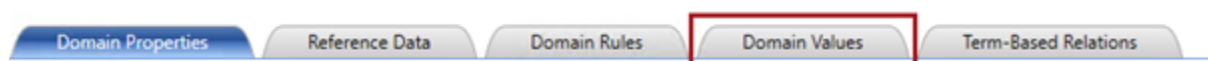
8. Configure the following additional domain properties.

Domain	Action
Address2	Enable Speller: Uncheck
StateOrProvince	Format Output to: Upper Case Enable Speller: Uncheck
ManagerEmail	Format Output to: Lower Case Enable Speller: Uncheck

Configuring Domain Values

In this task, you will configure domain values and define a synonym.

1. Select the **Office** domain.
2. Select the **Domain Values** tab.



3. Notice that the domain values already includes the **DQS_NULL** value.

All domains include this value, and it cannot be deleted.

4. Set the **DQS_NULL** value to **Invalid**.

Value	Type	Correct to
DQS_NULL		

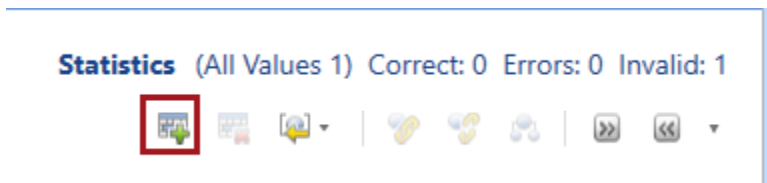
*This configuring ensures that missing **Office** values will result in an invalid record.*

5. Repeat the last step to set the **DQS_NULL** value to **Invalid** for the following additional domains:

- District
- Address1
- City
- PostalCode
- StateOrProvince
- Country
- Phone

6. Select the **Country** domain.

7. To add a domain value, click **Add New Domain Value**.



8. In the new row added to the domain value grid, enter **Canada**.

Value	Type	Correct to
DQS_NULL		
Canada		

9. Press **Enter**.
10. Notice that the domain value is set to type **Correct** (green check mark).

11. Add two additional domain values:

- United States
- US

12. In the grid, notice that domain values sort alphabetically, and that new domain values added during the activity are adorned with a yellow star.

Value	Type	Correct to
★ Canada	✓	
DQS_NULL	⚠	
★ United States	✓	
★ US	✓	

13. To define synonyms, first select the **United States** domain value, and then while pressing the **Control** key, select the **US** domain value.

14. Right-click the selection, and then select **Set as Synonyms**.

15. Notice the arrangement of domain values, with the **United States** domain value as the leading value.

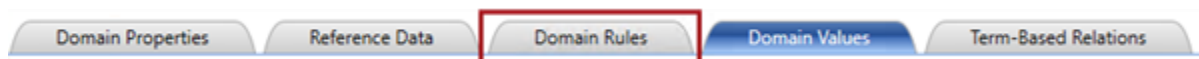
*While **US** is regarded as correct domain value, it will be corrected to the leading value.*

Value	Type	Correct to
★ Canada	✓	
DQS_NULL	⚠	
★ United States	✓	
★ US	✓	United States

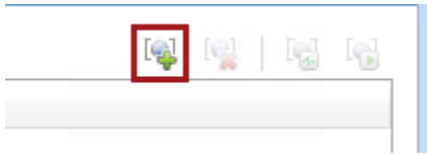
Configuring Domain Rules

In this task, you will configure domain rules.

1. Select the **ManagerFirstName** domain.
2. Select the **Domain Rules** tab.



- To add a domain rule, click **Add a New Domain Rule**.



- In the domain rule grid, in the **Name** box, enter **Not an initial**.

Active	Name	Description
<input checked="" type="checkbox"/>	Not an initial	

It is important to be clear—yet concise—when defining a rule name, as it will be provided as the reason when data cannot conform to the rule.

- In the **Build a Rule** section, modify the operator to **Length is Greater Than or Equal to**, and then in the corresponding box, enter **2**.

Build a Rule: Not an initial

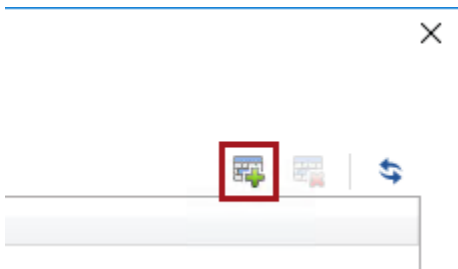
ManagerFirstName

Length is greater than or equal to 2

- To test the domain rule, click **Run the Selected Domain Rule on Test Data**.

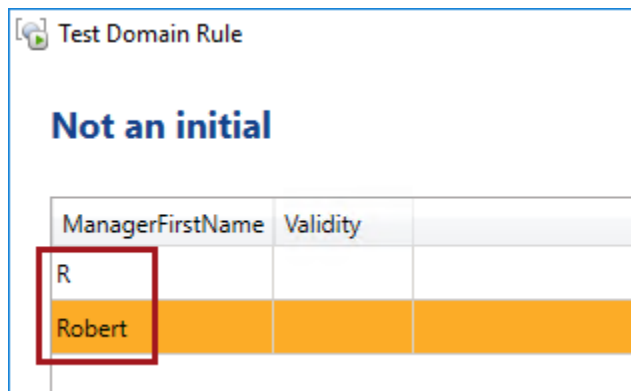


- In the **Test Domain Rule** window, click **Adds a New Testing Term for the Domain Rule**.

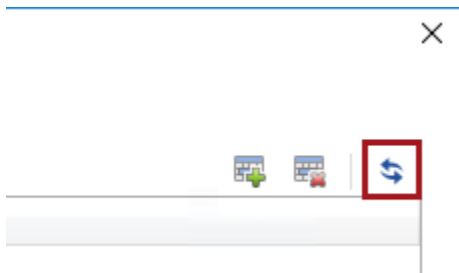


- In the **ManagerFirstName** box, enter **R**.

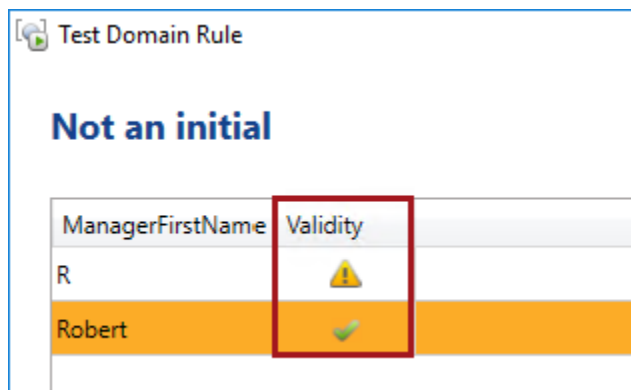
9. Add a second testing term with the value **Robert**.



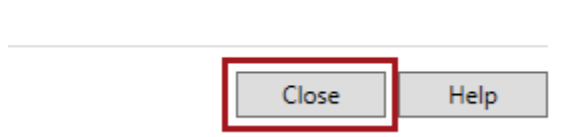
10. Click **Test the Domain Rule On All the Terms**.



11. Verify that the value **R** is invalid, while the value **Rob** is correct.



12. Click **Close**.



13. Select the **ManagerLastName** domain.
14. Repeat the steps in this task to create the **Not an initial** domain rule.
15. Select the **Phone** domain.

16. Create a domain rule named **Valid phone format**, and configure the following rule logic.

*For your convenience and accuracy, you can copy the pattern from the **F:\Labs\Lab1-2\Assets\Snippets.txt** file (open with Notepad).*

Build a Rule: Valid phone format

Phone

Value matches pattern (000) 000-0000

17. Test the domain rule with the following terms:

- 800 123 4567
- 800 123-4567
- (800) 123-4567

18. Verify that the only the final term is correct.

Test Domain Rule

Valid phone format

Phone	Validity	
800 123 4567	⚠	
800 123-4567	⚠	
(800) 123-4567	✅	

19. Select the **ManagerEmail** domain.
20. Create a domain rule named **Valid email address**, and configure the following rule logic.

*For your convenience and accuracy, you can copy the regular expression from the **F:\Labs\Lab1-2\Assets\Snippets.txt** file (open with Notepad).*

Build a Rule: Valid email address

ManagerEmail

Value matches regular expression \b[A-Z0-9._%+-]+@[A-Z0-9.-]+\.[A-Z]{2,4}\b

21. Note that this domain rule only tests valid email addresses, and not the additional requirement that the email address must belong to a particular domain.
22. To add a new condition, click **Add a New Condition to the Selected Clause**.



23. Complete the configuration of the rule logic on the following.

*For your convenience and accuracy, you can copy the string from the **F:\Labs\Lab1-2\Assets\Snippets.txt** file (open with Notepad).*

Build a Rule: Valid email address

ManagerEmail

Value matches regular expression ▾ \b[A-Z0-9._%+-]+@[A-Z0-9.-]+\.[A-Z]{2,4}\b

AND ▾

Value ends with ▾ @lab.microsoft.com

24. Test the domain rule with the following terms:

- rob@hotmail.com
- rob@@lab.microsoft.com
- rob@lab.microsoft.com

25. Verify that the only the final term is correct.

Test Domain Rule

Valid email address

ManagerEmail	Validity	
rob@hotmail.com	⚠	
rob@@lab.microsoft.com	⚠	
rob@lab.microsoft.com	✅	

26. Select the **PostalCode** domain.

27. Create a domain rule named **Valid postal code format**, and configure the following rule logic.

*For your convenience and accuracy, you can copy the two regular expressions from the **F:\Labs\Lab1-2\Assets\Snippets.txt** file (open with Notepad).*

The first regular expression validates a US postal code (ZIP Code) allowing also for the Zip+4 Code format. The second regular expression validates a Canadian postal code, requiring a space at the fourth character.

Build a Rule: Valid postal code format

PostalCode

Value matches regular expression ▾

AND ▾

Value matches regular expression ▾

28. To modify the operator, to the right of the **AND** operator, click the down-arrow, and then select **OR**.



29. Verify that the domain rule looks like the following.

Build a Rule: Valid postal code format

PostalCode

Value matches regular expression ▾

OR ▾

Value matches regular expression ▾

30. Test the rule with the following terms:

- 1234
- 12345
- 12345-123
- 12345-1234
- A1A1A1
- A1A 1A1 (the fourth character is a space)

31. Verify that the only the second, fourth and last terms are correct.

Test Domain Rule

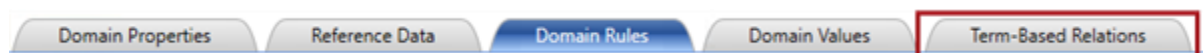
Valid postal code format

PostalCode	Validity	
1234	⚠	
12345	✓	
12345-123	⚠	
12345-1234	✓	
A1A1A1	⚠	
A1A 1A1	✓	

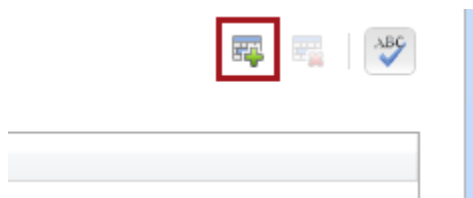
Configuring a Term-Based Relation

In this task, you will configure a term-based relation.

1. Select the **District** domain.
2. Select the **Term-Based Relations** tab.



3. To add a term-based relation, click **Add New Relation**.



4. In the term-based relation grid, in the **Value** box, enter **Distr.** (include the period).

5. In the **Correct To** box, enter **District** (do not include the period).

Value	Correct to
Distr.	District

This relation will ensure all abbreviated instances will be corrected to the full name.

Configuring a Composite Domain

In this task, you will configure a composite domain to be composed of the address-related domains.

1. In the left pane, click **Create a Composite Domain**.

Domain Management

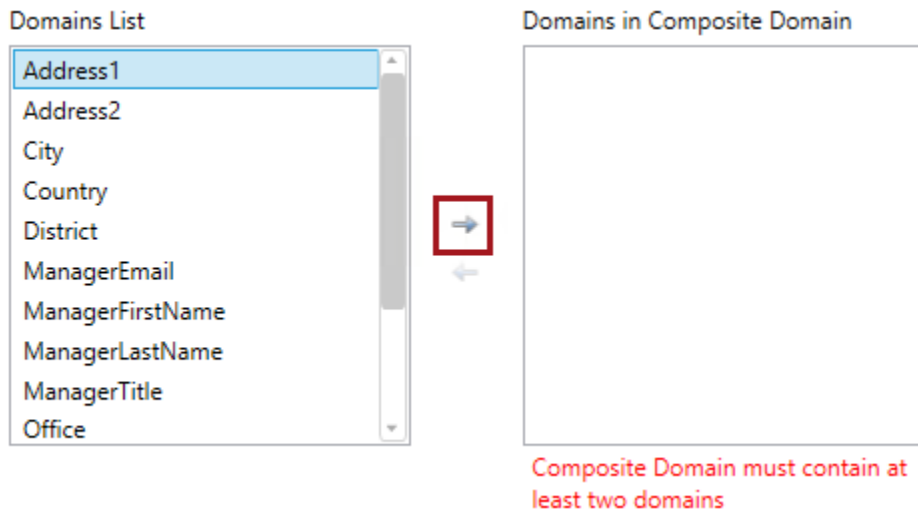


2. In the **Create a Composite Domain** window, in the **Composite Domain Name** box, enter **Address**.

A screenshot of the 'Create a Composite Domain' window. The title bar says 'Create a Composite Domain'. Inside the window, there is a label 'Composite Domain Name:' followed by a text box containing the word 'Address'. To the right of the text box is the text '(Required)'. Below the text box is a label 'Description:' followed by a large empty text area.

3. In the **Domains List**, select **Address1**.

4. To add the domain to the composite domain, click the right-arrow.

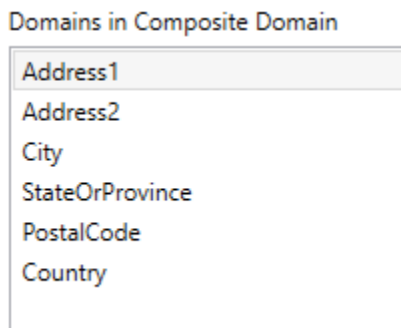


5. Add the following domains also to the composite domain, ensuring that they are added in the order listed.

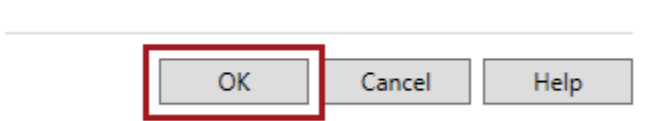
*Tip: You can double-click each domain to add it to the list, and you can also multi-select items in order by pressing the **Control** key, and then add them by clicking the right-arrow.*

- Address2
- City
- StateOrProvince
- PostalCode
- Country

6. Verify that the **Domains in Composite Domain** list includes the following domains, in the order presented.

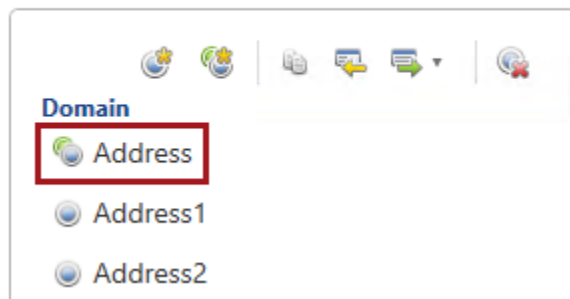


7. Click **OK**.



8. In the left pane, notice the addition of the composite domain, and that it is adorned with a different icon.

Domain Management



Configuring a Composite Domain Rule

In this task, you will configure a composite domain rule to correct **StateOrProvince** values for the city of **Vancouver**.

Vancouver is a city in both the US state of Washington, and the Canadian province of British Columbia.

1. Ensure that the **Address** composite domain is selected.
2. Select the **CD Rules** tab.



3. To add a composite domain rule, click **Add a New Domain Rule**.



4. In the cross-domain rules grid, in the **Name** box, enter **Vancouver CA**.

Active	Name
<input checked="" type="checkbox"/>	Vancouver CA

5. In the **Build a Rule** section, configure the following rule logic.

Build a Rule: Vancouver CA

Value is equal to

6. To add a clause, right-click inside a blank area of the **Build a Rule** section, and then select **Add Clause**.

Build a Rule: Vancouver CA

Value is equal to

Add clause

Delete clause

Move up

Move down

7. Complete the configuration of the rule logic based on the following.

Build a Rule: Vancouver CA

Value is equal to

Value is equal to

8. In the **Then** section, configure the following.

Then

StateOrProvince

Value is equal to

BC

This configuration is referred to as a *definitive cross-domain rule*. A definitive cross-domain rule is one that uses the **Values is Equal to** operator in the **Then** logic, and it is able to correct values, rather than just validate values.

The value **BC** will be added to the **StateOrProvince** domain values as a result of configuring this rule.

9. Test the cross-domain rule with the following terms.

Test Composite Domain Rule

Vancouver CA

City	Country	StateOrProvince	Validity	Correct To
Vancouver	Canada			
Vancouver	United States			

10. Verify that the first term would be corrected to **BC**.

Test Composite Domain Rule

Vancouver CA

City	Country	StateOrProvince	Validity	Correct To
Vancouver	Canada		⚠	BC
Vancouver	United States		?	

11. Create a second cross-domain rule named **Vancouver US**, and configure the following rule logic.

Build a Rule: Vancouver US

City

Value is equal to

Vancouver

AND

Country

Value is equal to

United States

Then

StateOrProvince

Value is equal to

WA

The value **WA** will be added to the **StateOrProvince** domain values as a result of configuring this rule.

12. Test the cross-domain rule with the following terms.

Test Composite Domain Rule

Vancouver US

City	Country	StateOrProvince	Validity	Correct To	
Vancouver	Canada				
Vancouver	United States				

13. Verify that the second term would be corrected to **WA**.

Test Composite Domain Rule

Vancouver US

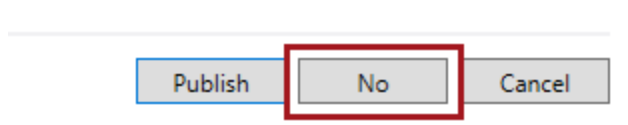
City	Country	StateOrProvince	Validity	Correct To	
Vancouver	Canada		?		
Vancouver	United States		!	WA	

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



15. When prompted to publish the knowledge base, click **No**.

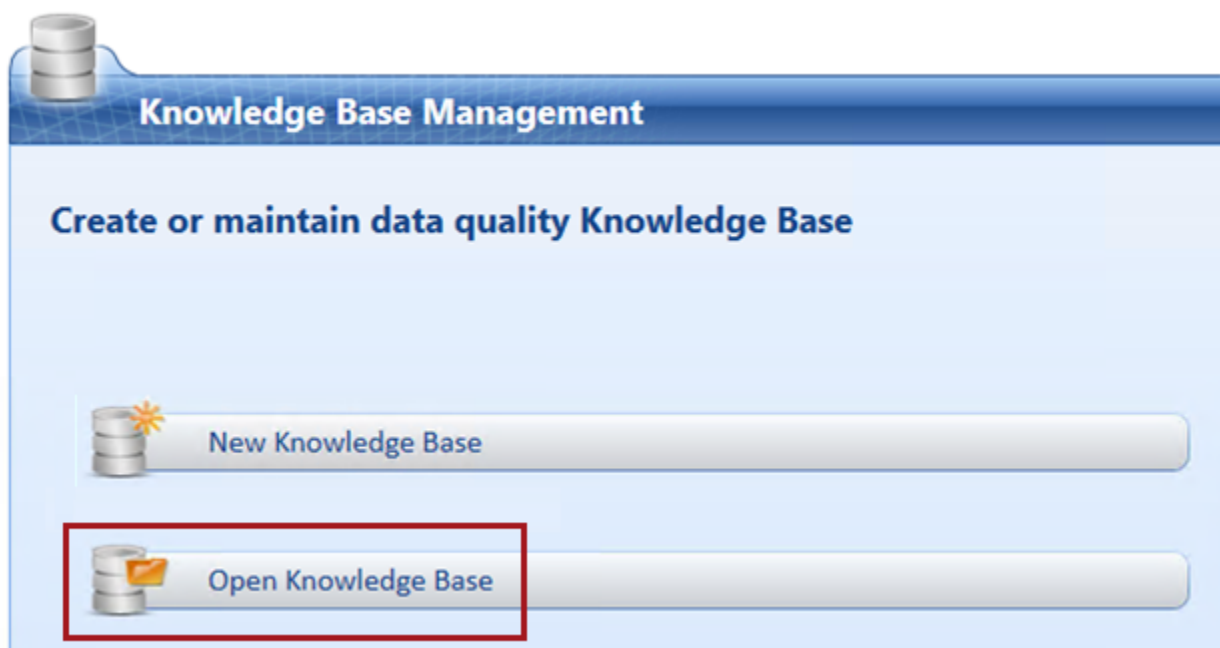
The knowledge base is not yet ready to applied to a cleansing activity. You will continue to enhance the knowledge base with knowledge discovery activities in the next exercise.



Reviewing Knowledge Base Status

In this task, you will review the status of the knowledge base.

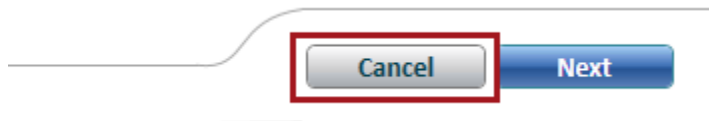
1. To create a new knowledge base, in the **Knowledge Base Management** panel, click **Open Knowledge Base**.



2. In the grid, notice that the **Office** knowledge base is locked, and has the state **In Work**.

The knowledge base cannot be used until it is unlocked. You will unlock the knowledge base when you publish it in the next exercise.

3. Click **Cancel**.



Monitoring DQS Activity

In this task, you will monitor the DQS activity.

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



2. To sort the activities by descending order, in the activity grid, click the **ID** column header twice.
3. Notice the first listed activity is the one you just completed.

Every DQS activity undertaken is logged and remains available for review and audit.

Filter By: Value:

ID	Name	Is Active	Type	Sub Type
1002	Office	Ended	Knowledge Management	Domain Management
1001	DQS Data	Ended	Knowledge Management	Domain Management
1000	DQS Data	Ended	Knowledge Management	Domain Management

4. Click **Close**.



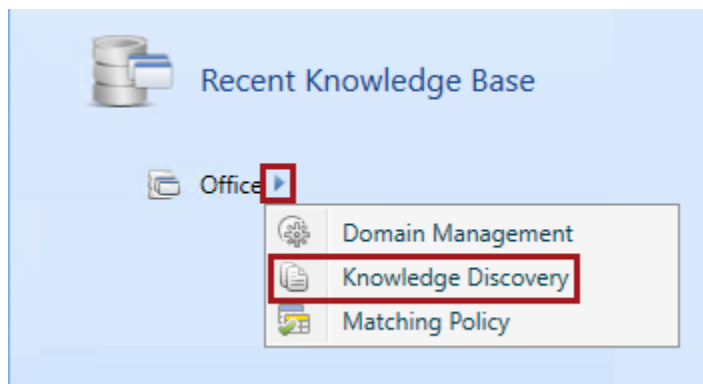
Exercise 4: Performing Knowledge Discovery

In this exercise, you will perform knowledge discovery to add domain values to the knowledge base.

Adding Trusted Knowledge

In this task, you will add trusted state and province codes to the **StateOrProvince** domain values. This trusted knowledge was acquired from the US and Canadian postal authorities.

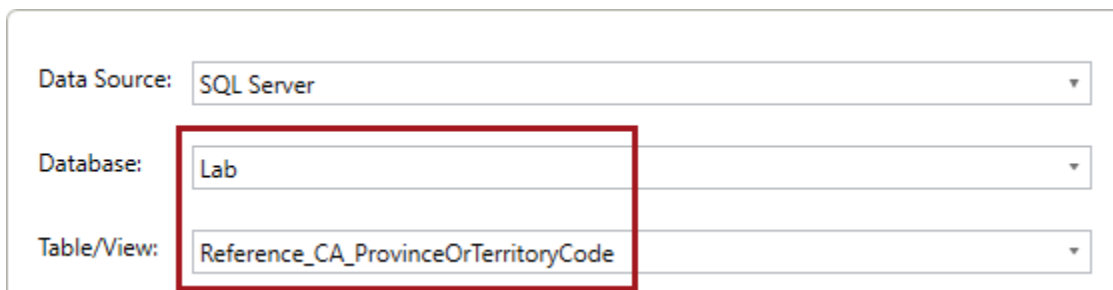
1. To perform knowledge discovery, in the **Knowledge Base Management** panel, click the **Office** knowledge base, and then select the **Knowledge Discovery** activity.



2. Notice that step 1 of the activity is to map to external data containing knowledge.



3. In the **Database** dropdown list, select **Lab**.
4. In the **Table/View** dropdown list, select **Reference_CA_ProvinceOrTerritoryCode**.

A screenshot of a configuration form for a 'Map' step. It contains three dropdown menus: 'Data Source' set to 'SQL Server', 'Database' set to 'Lab', and 'Table/View' set to 'Reference_CA_ProvinceOrTerritoryCode'. The 'Database' and 'Table/View' dropdowns are highlighted with a red box.

5. At the top-right corner of the **Mappings** grid, click **Preview Data Source**.

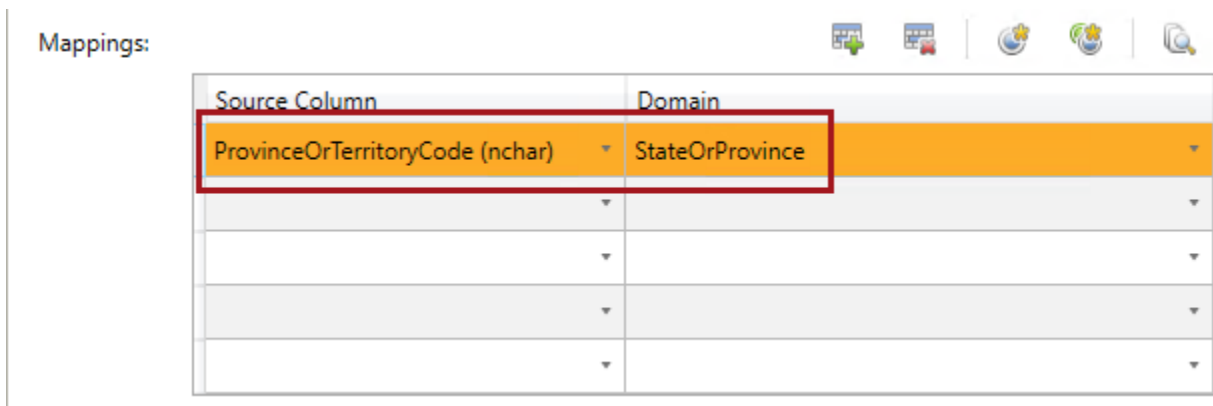


6. Review the source data, and then click **Close**.



7. In the **Mappings** grid, in the first row, in the **Source Column** column, select the **ProvinceOrTerritoryCode** column.

8. In the corresponding **Domain** column, select the **StateOrProvince** domain.



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



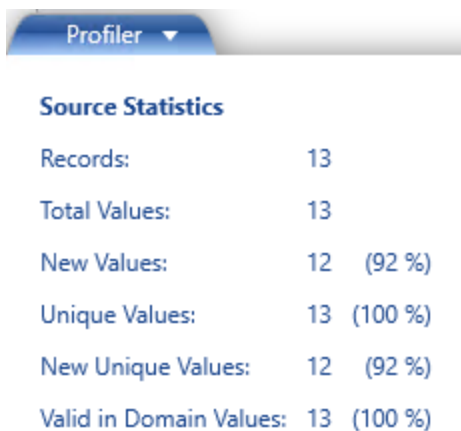
10. Notice that step 2 of the activity is to discover knowledge from the source.



11. Click **Start**.



12. When the discovery analysis has completed, review the source statistics in the **Profiler** pane.

The Profiler pane shows source statistics for a dataset. It includes a tab labeled "Profiler" and a section titled "Source Statistics".

Source Statistics	
Records:	13
Total Values:	13
New Values:	12 (92 %)
Unique Values:	13 (100 %)
New Unique Values:	12 (92 %)
Valid in Domain Values:	13 (100 %)

13. Note that 13 unique values were detected, of which 12 are new values for the domain.

*In the previous exercise, when you added the cross-domain rules, both **BC** and **WA** were added to the domain values. **BC** (British Colombia) was included in the source data, but not added to the domain values as it already exists.*

14. To proceed to the next step, click **Next**.
15. Notice that step 3 of the activity is to manage domain values.

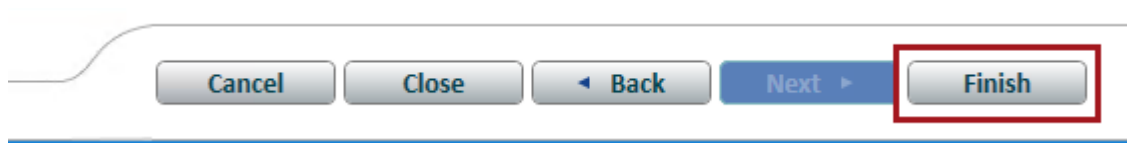


16. Review the list of domain values, and notice that this is a list of what has been added in this activity.
17. To reveal all domain values, uncheck the **Show Only New** checkbox.

StateOrProvince

Find: Filter: All Values ☐ Show Only New

18. To complete the knowledge discovery process, click **Finish**.



19. Do not publish the knowledge base.
20. Repeat the steps in this task to perform another knowledge discovery activity, this time sourcing data from the **Reference_US_StateCode** table.

A configuration form with three dropdown menus. The first is 'Data Source' with 'SQL Server' selected. The second is 'Database' with 'Lab' selected. The third is 'Table/View' with 'Reference_US_StateCode' selected. A red rectangular box highlights the 'Database' and 'Table/View' fields.

Knowledge Base Check

Lab 1-2 ► Profiler – Source Statistics

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

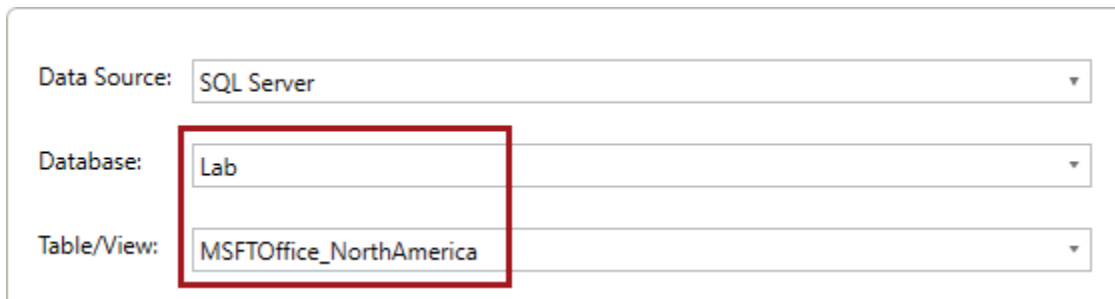
It is recommended that you open your Knowledge Check portion of the course in EdX at this time to answer questions as you complete the lab, or take a screenshot of the **Profiler – Source Statistics** pane, to refer to later.

21. Do not publish the knowledge base.

Adding Additional Knowledge

In this task, you will add knowledge sourced from the Office dataset. There are known issues with this data, and so judgement will need to be applied to ensure domain values are appropriately added.

1. Perform knowledge discovery a third time on the **Office** knowledge base.
2. Source data from **Lab** database, and the **MSFTOffice_NorthAmerica** table.



The screenshot shows a configuration interface with three dropdown menus. The first dropdown is labeled 'Data Source:' and has 'SQL Server' selected. The second dropdown is labeled 'Database:' and has 'Lab' selected. The third dropdown is labeled 'Table/View:' and has 'MSFTOffice_NorthAmerica' selected. A red rectangular box highlights the 'Database:' and 'Table/View:' sections.

3. Map only the following four source columns to their respective domains.

Source Column	Domain
Office (nvarchar)	Office
District (nvarchar)	District
StateOrProvince (nvarchar)	StateOrProvince
Country (nvarchar)	Country

*The rationale for performing knowledge discovery for the **StateOrProvince** domain is to detect and appropriately configure anomalies.*

4. Notice that domains that can be cleansed by domain rules (i.e. **Phone** and **ManagerEmail**) are not included in this knowledge discovery activity. Some domains do not need to have possible values stored as domain values.
5. Proceed to the discovery step, and start the discovery process.

- Review the profiler statistics.

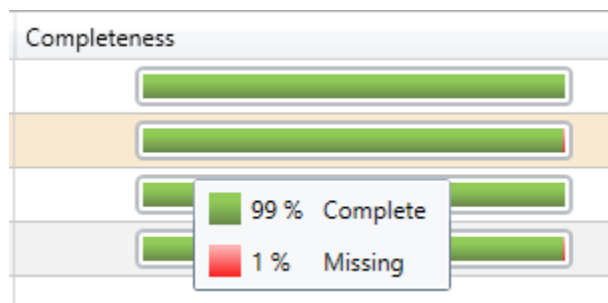
Knowledge Base Check

Lab 1-2 ► MSFTOffice_NorthAmerica 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 take a screenshot of the **Profiler – Source Statistics** pane, to refer to later.

- In the profiler grid, review the statistics at domain level, and also hover the cursor **Completeness** bar.
- Notice that the **District** and **Country** domains have a small proportion of missing values (these were detected during the SSIS data profiling exercise earlier in this lab).



- For the **Country** domain, notice the notification icon in the **New** column.

Field	Domain	New
Office	Office	70 (100 %)
District	District	69 (99 %)
StateOrProvince	StateOrProvince	12 (17 %)
Country	Country	 3 (4 %)

- Hover the cursor over the notification icon to reveal a tooltip describing a possible issue.

You can ignore the issue in this lab.

- Proceed to the domain management step.
- In the left pane, select the **Office** domain.

13. Notice the domain value **Ausstin, TX** has a red squiggly.

As this domain has spelling enabled, DQS used its dictionary to check spelling.

Value	Frequency
✳ Albany, NY	1
✳ Alpharetta, GA	1
✳ Ausstin, TX	1
✳ Bellevue, WA	1

14. Right-click the **Ausstin, TX** text, and then select the correct spelling suggestion: **Austin, TX**.
15. Notice that the correction has assigned the domain value as an error, and corrected it to a new domain value.

The knowledge base now understands how to correct any instance of this misspelled office.

Value	Frequency	Type	Correct to
✳ Albany, NY	1	✓	▼
✳ Alpharetta, GA	1	✓	▼
✳ Austin, TX	0	✓	▼
✳ Ausstin, TX	1	✗	▼ Austin, TX

16. Scroll down the list to locate the **Lehi, UT** office domain value (which is, in fact, correctly spelled).

Value	Frequency
✳ Kansas City, KS	1
✳ Lehi, UT	1
✳ Los Angeles, CA	1

17. Right-click the **Lehi, UT** domain value, and then select **Add to Dictionary**.
18. Notice that the red squiggly has been removed.

19. Show all domain values.

It is useful to show all values when managing synonyms that may involve existing members.

Office

Find: Filter: All Values ☐ Show Only New

20. Locate the two adjacent domain values for New York.

★ New York, NY	2	✓	▼
★ NYC, NY	1	✓	▼

21. Multi-select the two domain values, right-click the selection, and then select **Set as Synonyms**.

22. Ensure that **New York, NY** is the leading value.

☑ ★ New York, NY	2	✓	▼	
★ NYC, NY	1	✓	▼	New York, NY

23. Select the **District** domain.

*There is no need to be concerned about the **Distr.** abbreviation used in these domain values, as in the previous exercise you configured a term-based relation to replace any instances of the abbreviation.*

24. Use the dictionary to correct the **Midwesst Distr.** Domain value to **Midwest Distr.**
25. Show all domain values, and notice how the misspelled Midwest domain value corrects to an existing domain value.

☑ ★ Midwest Dist.	4	✓	▼	
★ Midwestt Dist.	1	✗	▼	Midwest Dist.

26. Set the **Greater South East District** member to **Error**.

Value	Frequency	Type
🌟 Canada	10	✓
🌟 Desert Mountain District	3	✓
DQS_NULL	1	⚠
🌟 Greater South East District	1	✓
🌟 Greater Southeast District	7	✓
🌟 Gulf Coast Dist.	3	✗

27. In the adjacent **Correct to** box, enter the correct domain value, **Greater Southeast District**, and then press **Enter**.
28. Notice how the error domain value relates to the correct domain value.

Value	Frequency	Type	Correct to
🌟 Canada	10	✓	
🌟 Desert Mountain District	3	✓	
DQS_NULL	1	⚠	
🌟 Greater Southeast District	7	✓	
🌟 Greater South East District	1	✗	Greater Southeast District

29. Correct also the **Mid-Atlantic Dist.** value to the **Mid Atlantic District** value.

🌟 Mid Atlantic District	5	✓	
🌟 Mid-Atlantic Dist.	1	✗	Mid Atlantic District

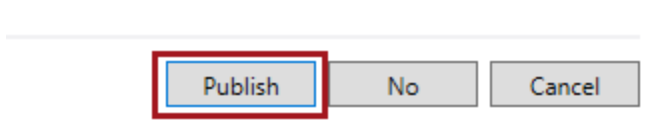
30. Select the **StateOrProvince** domain.
31. Correct each of the five new domain values, based on the following.

Value	Frequency	Type	Correct to
🌟 Calif.	2	✗	CA
🌟 California	1	✗	CA
🌟 Fla.	3	✗	FL
🌟 Ohio	2	✗	OH
🌟 Tex.	1	✗	TX

32. Show all domain values, and notice how the corrections relate to an existing domain values.
33. Select the **Country** domain.
34. Correct the **CA** domain value to the **Canada** domain value.

Value	Frequency	Type	Correct to
🌟 CA	2	✖	Canada
🌟 United State	1	✖	United States

35. Notice that DQS used its dictionary to correct the misspelled **United States** domain value.
36. Show all domain values, and notice how the corrected domain values relate to the existing domain values.
37. Finish the knowledge discovery activity, and publish the knowledge base.



38. When notified that the knowledge base has been published, click **OK**.



*You will use the knowledge base in **Lab 2-1** to cleanse the Office dataset.*

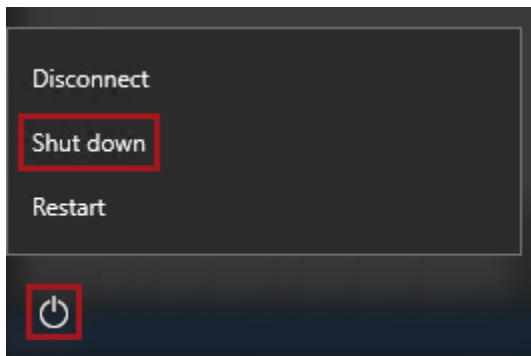
39. Review the knowledge base status, and notice that it is no longer locked, and has not state (i.e. it is open).
40. Review the activity monitoring, and notice the three knowledge discovery activities you have just completed.

*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.*

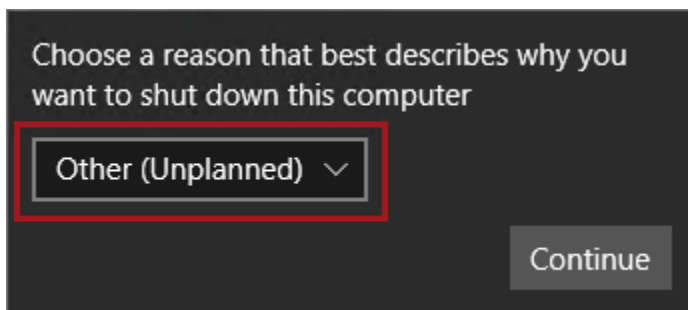
Finishing Up

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

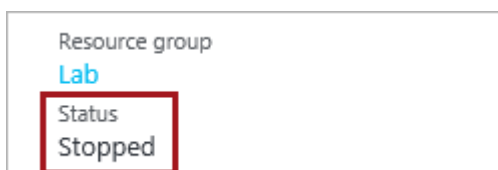
1. Close all open applications.
2. Press the **Windows** key, and then in the **Start** page, located at the bottom-left, click the **Power** button, and then select **Shut Down**.



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



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

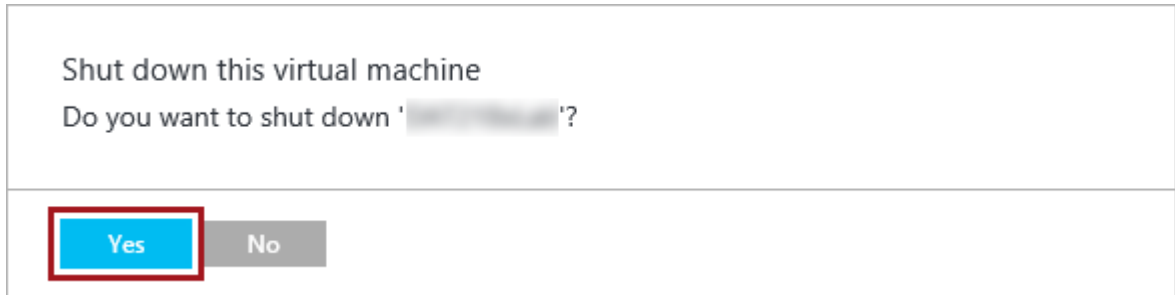


In this state, however, the VM is still billable.

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

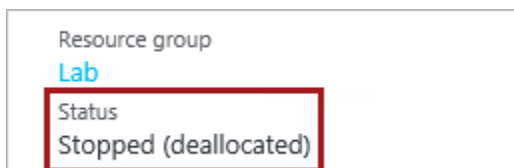


- When prompted to stop the VM, click **Yes**.



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.

- Verify that the VM status updates to **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.

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