SQL Server 2014 BI

Lab 02

Working with Data Quality Services in SQL Server 2014

Jump to the Lab Overview
Terms of Use

© 2014 Microsoft Corporation. All rights reserved.

Information in this document, including URL and other Internet Web site references, is subject to change without notice. Unless otherwise noted, the companies, organizations, products, domain names, e-mail addresses, logos, people, places, and events depicted herein are fictitious, and no association with any real company, organization, product, domain name, e-mail address, logo, person, place, or event is intended or should be inferred. Complying with all applicable copyright laws is the responsibility of the user. Without limiting the rights under copyright, no part of this document may be reproduced, stored in or introduced into a retrieval system, or transmitted in any form or by any means (electronic, mechanical, photocopying, recording, or otherwise), or for any purpose, without the express written permission of Microsoft Corporation.

For more information, see Microsoft Copyright Permissions at http://www.microsoft.com/permission

Microsoft may have patents, patent applications, trademarks, copyrights, or other intellectual property rights covering subject matter in this document. Except as expressly provided in any written license agreement from Microsoft, the furnishing of this document does not give you any license to these patents, trademarks, copyrights, or other intellectual property.

The Microsoft company name and Microsoft products mentioned herein may be either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries. The names of actual companies and products mentioned herein may be the trademarks of their respective owners.

This document reflects current views and assumptions as of the date of development and is subject to change. Actual and future results and trends may differ materially from any forward-looking statements. Microsoft assumes no responsibility for errors or omissions in the materials.

THIS DOCUMENT IS FOR INFORMATIONAL AND TRAINING PURPOSES ONLY AND IS PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND, WHETHER EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, AND NON-INFRINGEMENT.
## Contents

**TERMS OF USE** .................................................................................................................. 2

**CONTENTS** .......................................................................................................................... 3

**ABOUT THE AUTHOR** ......................................................................................................... 4

**DOCUMENT REVISIONS** ....................................................................................................... 4

**LAB OVERVIEW** .................................................................................................................. 5

**EXERCISE 1: CREATING A KNOWLEDGE BASE** ................................................................. 6
  Task 1 – Exploring the New Customer Data ........................................................................... 6
  Task 2 – Creating the Knowledge Base ................................................................................. 7
  Task 3 – Performing Knowledge Discovery ........................................................................ 8
  Task 4 – Managing the StateCode Domain ........................................................................ 11
  Task 5 – Creating the EmailAddress Domain ..................................................................... 13

**EXERCISE 2: CREATING A DATA QUALITY PROJECT** .......................................................... 16
  Task 1 – Creating the Data Quality Project ....................................................................... 16

**EXERCISE 3: USING INTEGRATION SERVICES TO CLEANSE DATA** ............................. 20
  Task 1 – Reviewing the LoadDimCustomer Package ......................................................... 20
  Task 2 – Updating the Package Data Flow ......................................................................... 21
  Task 3 – Executing the Package ....................................................................................... 22
  Task 4 – Reviewing the Logged Data .............................................................................. 23
  Task 5 – Finishing Up ........................................................................................................ 23

**SUMMARY** ............................................................................................................................ 24
About the Author

This lab was designed and written by Peter Myers.

Peter Myers has worked with Microsoft database and development products since 1997. Today, he specializes in all Microsoft BI products and provides mentoring, technical training, and education content authoring for SQL Server, Office, and SharePoint. Peter has a broad business background supported by a bachelor’s degree in applied economics and accounting, and he extends this with solid experience backed by current MCSE and MCT certifications. He has been a SQL Server MVP since 2007.

Document Revisions

<table>
<thead>
<tr>
<th>#</th>
<th>Date</th>
<th>Author</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>19-OCT-2014</td>
<td>Peter Myers</td>
<td>Initial release</td>
</tr>
</tbody>
</table>
Lab Overview

Introduction
In this lab, you will create a Data Quality Services knowledge base to cleanse US state codes and email addresses. You will then create a Data Quality Project to cleanse data sourced from an Excel workbook. Finally, you will enhance an existing Integration Services package design to use the DQS Cleansing component to cleanse loaded data as part of an ETL process.

Objectives
The objectives of this exercise are to:

- Create a Data Quality Knowledge Base
- Use knowledge discovery to create a domain
- Manage domains
- Create a Data Quality Project
- Use the DQS Cleansing component in an Integration Services package

Exercises
This hands-on lab comprises the following exercise:

1. Creating a Knowledge Base
2. Creating a Data Quality Project
3. Using Integration Services to Cleanse Data

Estimated time to complete this lab: 30 minutes
Exercise 1: Creating a Knowledge Base

In this exercise, you will commence by exploring a small data extract of new customers that needs to be loaded into the data warehouse. To support the cleansing of the data, you will then create the *AdventureWorksBI* knowledge base consisting of two domains to cleanse US state code and email address data.

**Task 1 – Exploring the New Customer Data**

In this task, you will open and explore the new US customer data that must be cleansed and loaded in the *AdventureWorksDW2014* database’s *DimCustomer* table.

1. To open Excel, on the taskbar, click the Excel shortcut.

   ![Figure 1](image1.png)

   *Figure 1*

   *Selecting the Excel Shortcut*

2. To open an existing workbook, in the **Recent** pane (located at the left), click **Open Other Workbooks**.

   ![Figure 2](image2.png)

   *Figure 2*

   *Locating the Open Link*

3. In the **Open** pane, click **Computer**, and then click **Browse**.

4. In the **Open** window, navigate to the **D:\SQLServerBI\Lab02\Assets** folder, select the **NewCustomers.xlsx** file, and then click **Open**.
5. Notice the table of data representing 20 new customers to be loaded into the data warehouse **DimCustomer** table. In particular, notice the cells highlighted in red that represent data quality problems. The two email addresses are invalid, and the state values are required to be the two character US state postal code, not the state abbreviation.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>Q</td>
<td>Black</td>
<td>F</td>
<td><a href="mailto:jodi22@hotmail.com">jodi22@hotmail.com</a></td>
<td>7102 Serrano Way</td>
<td>CA</td>
<td>Chandler</td>
<td>85225</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>Q</td>
<td>Emily</td>
<td>F</td>
<td><a href="mailto:emily34@live.com">emily34@live.com</a></td>
<td>9344 Wishing Circle</td>
<td>PA</td>
<td>Bensalem</td>
<td>19010</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>Q</td>
<td>Edward</td>
<td>M</td>
<td><a href="mailto:edward25@hotmail.com">edward25@hotmail.com</a></td>
<td>9875 Grant St</td>
<td>WA</td>
<td>Bremerton</td>
<td>98322</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>Q</td>
<td>Gabriel</td>
<td>M</td>
<td><a href="mailto:gabriel42@hotmail.com">gabriel42@hotmail.com</a></td>
<td>6510 Northridge Drive</td>
<td>TN</td>
<td>Memphis</td>
<td>38115</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>Q</td>
<td>Kaylee</td>
<td>F</td>
<td><a href="mailto:kaylee60@live.com">kaylee60@live.com</a></td>
<td>377 Trees Drive</td>
<td>CA</td>
<td>El Cajon</td>
<td>92020</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>Q</td>
<td>Summer</td>
<td>F</td>
<td><a href="mailto:summer3@live.com">summer3@live.com</a></td>
<td>5705 Tweed Lane</td>
<td>CA</td>
<td>San Francisco</td>
<td>94119</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>Q</td>
<td>Franklin</td>
<td>M</td>
<td><a href="mailto:Franklin23@hotmail.com">Franklin23@hotmail.com</a></td>
<td>38 First Ave</td>
<td>IL</td>
<td>Chicago</td>
<td>60610</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>Q</td>
<td>Tara</td>
<td>F</td>
<td><a href="mailto:tara11@hotmail.com">tara11@hotmail.com</a></td>
<td>4148 Valley Ave</td>
<td>CA</td>
<td>Clearwater</td>
<td>31755</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>Q</td>
<td>Gary</td>
<td>M</td>
<td><a href="mailto:groper@live.com">groper@live.com</a></td>
<td>2950 Minell Place</td>
<td>WY</td>
<td>Casper</td>
<td>82001</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>7</td>
<td>Q</td>
<td>Johnathan</td>
<td>M</td>
<td><a href="mailto:johnathan13@hotmail.com">johnathan13@hotmail.com</a></td>
<td>2250 Pablo Nueva</td>
<td>WA</td>
<td>Woodinville</td>
<td>98072</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>8</td>
<td>Q</td>
<td>Keith</td>
<td>M</td>
<td><a href="mailto:keith04@hotmail.com">keith04@hotmail.com</a></td>
<td>4077 Chilouquap Ct</td>
<td>UT</td>
<td>Riverton</td>
<td>84065</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>9</td>
<td>Q</td>
<td>Jenny</td>
<td>F</td>
<td><a href="mailto:jenny25@live.com">jenny25@live.com</a></td>
<td>3829 Baltic Sea Ctr</td>
<td>OR</td>
<td>Milwaukee</td>
<td>53222</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>10</td>
<td>Q</td>
<td>Randy</td>
<td>M</td>
<td><a href="mailto:randy22@hotmail.com">randy22@hotmail.com</a></td>
<td>P.O. Box 2382</td>
<td>AZ</td>
<td>Phoenix</td>
<td>85004</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>11</td>
<td>Q</td>
<td>Dominique</td>
<td>M</td>
<td><a href="mailto:dominique16@live.com">dominique16@live.com</a></td>
<td>6685 Norms Court</td>
<td>IN</td>
<td>Daleville</td>
<td>47334</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>12</td>
<td>Q</td>
<td>Carla</td>
<td>F</td>
<td><a href="mailto:carla0@live.com">carla0@live.com</a></td>
<td>2782 San Simon</td>
<td>KP</td>
<td>Newport</td>
<td>40071</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>13</td>
<td>Q</td>
<td>Meghan</td>
<td>F</td>
<td><a href="mailto:maghan18@live.com">maghan18@live.com</a></td>
<td>9321 Spring Court</td>
<td>GA</td>
<td>Columbus</td>
<td>31901</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>14</td>
<td>Q</td>
<td>Frederick</td>
<td>M</td>
<td><a href="mailto:frederick11@hotmail.com">frederick11@hotmail.com</a></td>
<td>6176 Cow Street</td>
<td>AL</td>
<td>Birmingham</td>
<td>35203</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>15</td>
<td>Q</td>
<td>Toni</td>
<td>M</td>
<td><a href="mailto:toni4@hotmail.com">toni4@hotmail.com</a></td>
<td>7367 Wests Way</td>
<td>WA</td>
<td>Chantilly</td>
<td>20151</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>16</td>
<td>Q</td>
<td>Marshall</td>
<td>M</td>
<td><a href="mailto:marshall8@hotmail.com">marshall8@hotmail.com</a></td>
<td>3849 Santa Fe Court</td>
<td>CA</td>
<td>Houston</td>
<td>77003</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>17</td>
<td>Q</td>
<td>Jorge</td>
<td>M</td>
<td><a href="mailto:Jorge08@live.com">Jorge08@live.com</a></td>
<td>9764 Mt Emna Drive</td>
<td>WY</td>
<td>Cheyenne</td>
<td>82001</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 3**

*Reviewing the New Customer Data*

6. To close Excel, click the **X** button located at the top right corner.

**Task 2 – Creating the Knowledge Base**

In this task, you will create a knowledge base and commence its creation with the knowledge discovery activity.

1. To open the Data Quality Client, on the taskbar, click the Data Quality Client shortcut.

**Figure 4**

*Selecting the Data Quality Client Shortcut*

2. In the **Connect to Server** window, in the **Server Name** dropdown list, enter **localhost**, and then click **Connect**.
3. To create a new knowledge base, in the Data Quality Client, in the **Knowledge Base Management** section, click **New Knowledge Base**.

![New Knowledge Base](image)

**Figure 5**
Creating a New Knowledge Base

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

5. In the **Select Activity** section, select **Knowledge Discovery**.

![Select Activity](image)

**Figure 6**
Selecting the Knowledge Discovery Activity

6. Click **Next** (located at the bottom right).

**Task 3 – Performing Knowledge Discovery**

In this task, you will perform knowledge discovery for the **StateCode** domain by retrieving the existing US state codes used in the data warehouse.

1. Switch to the Data Quality Client.

2. Notice the progression of steps that you will work through to discover knowledge.

![Knowledge Discovery Steps](image)

**Figure 7**
Knowledge Discovery Steps
3. In the Map step, in the Database dropdown list, select the AdventureWorksDW2014 database.

4. In the Table/View dropdown list, select the vUSStateCode view (located at the bottom of the list).

   **Note:** This view retrieves all US state codes stored in the DimGeography table.

5. In the Mappings table, in the Source Column column, select StateCode (nvarchar).

6. Click Create a Domain.

   **Note:** You can hover over the buttons to reveal a tooltip that describes their functionality.

---

**Figure 8**

*Creating a Domain*
7. In the **Create Domain** window, in the **Domain Name** box, enter **StateCode**.

![figure9](image)

**Figure 9**
*Creating the StateCode Domain*

8. Click **OK**.

9. Ensure the **Mappings** table looks like the following.

<table>
<thead>
<tr>
<th>Source Column</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>StateCode (nvarchar)</td>
<td>StateCode</td>
</tr>
</tbody>
</table>

![figure10](image)

**Figure 10**
*Verifying the Mapping*

10. Click **Next**.

11. In the **Discover** step, click **Start**.

12. When the discovery process has completed, in the **Profiler** pane, notice that 36 unique values (US states) have been retrieved.

13. Click **Next**.
14. In the Manage Domain Values step, review the 36 state codes retrieved from the vUSStateCode view.

**Note:** The star next to each state code indicates that it is a new value, and its type is defaulted to Correct (the green check). It is possible to add additional values, remove values, and to configure different types (Error and Invalid). You will add Error values in the next task.

15. To complete the knowledge discovery process, click Finish.

16. When prompted to publish the knowledge base, click No.

**Note:** Once published, the knowledge base is available for use. In the next two tasks, the StateCode domain will be enhanced with new members, and an additional domain will be added. The knowledge base will remain exclusively locked by until the domains are fully defined, and then it will be published.

**Task 4 – Managing the StateCode Domain**

In this task, you will manage the StateCode domain by adding two corrections. Note that the Tex. value (as has been used in the NewCustomers.xls workbook) will not be corrected. In the final exercise of this lab, you see how to manage invalid data in an ETL process.

1. In the Data Quality Client, in the Knowledge Base Management section, in the Recent Knowledge Base list, click the arrow to the right of the AdventureWorksBI domain, and then select Domain Management.

![Performing Domain Management](image)

**Figure 11**

*Performing Domain Management*

2. Notice that the StateCode domain is selected.
3. Select the **Domain Values** tab, and notice the values loaded during the knowledge discovery process.  

**Note:** In addition, the **DQS_NULL** value is included to support the correction of NULL (missing) values.

4. To add a new value, click **Add New Domain Value**.

![Figure 12](image12.png)

*Figure 12*

*Adding a New Domain Value*

5. In the new row, in the **Value** column, enter **Calif.** (include the period).

6. In the corresponding **Type** dropdown list, select the red cross (error).

7. In the corresponding **Correct To** box, enter the two character state code **CA** (do not include a period).

8. Verify that the new value looks like the following.

![Figure 13](image13.png)

*Figure 13*

*Verifying the New Value*

9. Press **Enter** to commit the new value.

10. Repeat the steps in this task to add a second new value for **Fla.** (include the period) that will correct to two character state code **FL** (no period).

11. Scroll to the top of the domain values list, and notice the association of the new domain values to their corrected domain values.

![Figure 14](image14.png)

*Figure 14*

*Reviewing the CA Value*
Task 5 – Creating the EmailAddress Domain

In this task, you will create a new domain to test the validity of email addresses. The domain will be based on a rule that uses a regular expression.

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

   ![Create a Domain](image)

   **Figure 15**
   Creating a New Domain

2. In the Create Domain window, in the Domain Name box, enter EmailAddress, and then click OK.

3. Notice that the new domain is selected.

4. Select the Domain Rules tab.

5. Click Add a New Domain Rule.

   ![Add a New Domain Rule](image)

   **Figure 16**
   Creating a New Domain Rule

6. In the Name box, enter Valid EmailAddress.

7. In the Build a Rule section, in the dropdown list, select Value Matches Regular Expression.

8. In the rule box, enter the following regular expression.

   **Note:** For convenience, the regular expression can be copied from the D:\SQLServerBI\Lab02\Assets\Snippets.txt file.

   **Regular Expression**
   \b[A-Z0-9._%+-]+@[A-Z0-9.-]+\.[A-Z]{2,4}\b
9. Verify that the rule looks like the following.

![Figure 17]

Verifying the Domain Rule

10. To test the rule, click **Run the Selected Domain Rule On Test Data**.

![Figure 18]

Testing the Domain Rule

11. In the **Test Domain Rule** window, click **Add a New Testing Term For the Domain Rule**.

![Figure 19]

Adding a New Testing Term

12. In the **EmailAddress** box, enter **john@hotmail.com** (a valid email address).

13. Repeat the last two steps to add the following two invalid email addresses:

<table>
<thead>
<tr>
<th>EmailAddress</th>
</tr>
</thead>
<tbody>
<tr>
<td>john**@hotmail.com</td>
</tr>
<tr>
<td>john@hotmaiil</td>
</tr>
</tbody>
</table>
14. Click **Test the Domain Rule On All the Terms.**

![Figure 20](image)

*Figure 20*

*Testing the Domain Rule*

15. Review the icon in the **Validity** column for each test **EmailAddress** value.

16. Click **Close**.

17. Click **Finish**.

18. When prompted to publish the knowledge base, click **Publish**.

19. When prompted to acknowledge that the knowledge base was published, click **OK**.

© Copyright 2014 Microsoft Corporation. All rights reserved.
Exercise 2: Creating a Data Quality Project

In this exercise, you will create a Data Quality Project to cleanse and review the data in the NewCustomers.xls file.

Task 1 – Creating the Data Quality Project

In this task, you will create a Data Quality Project to cleanse and review the data in the NewCustomers.xls file.

1. In the Data Quality Client, in the Data Quality Projects section, select New Data Quality Project.

![New Data Quality Project](NewDataQualityProject.png)

*Figure 21*
Creating a New Data Quality Project

2. In the Name box, enter NewCustomers.

3. In the Use Knowledge Base dropdown list, ensure the AdventureWorksBI knowledge base is selected.

4. In the Select Activity section, notice that Cleansing is selected.

5. Click Next.

6. Notice the progression of steps that you will work through to create the data quality project.

![Data Quality Project Steps](DataQualityProjectSteps.png)

*Figure 22*
Data Quality Project Steps

7. In the Map step, in the Data Source dropdown list, select Excel File.

8. To the right of the Excel File box, click Browse.

9. In the Select an Excel File window, navigate to the D:\SQLServerBI\Lab02\Assets folder, select the NewCustomers.xls file, and then click Open.
10. In the Worksheet dropdown list, select **NewCustomers**.

11. In the Mappings table, in the Source Column column, select **EmailAddress (String)**.

12. In the corresponding Domain column, select the **EmailAddress** domain.

13. Repeat the last two steps to map the **State (String)** source column to the **StateCode** domain.

14. Verify that the mappings look like the following.

<table>
<thead>
<tr>
<th>Source Column</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>EmailAddress (String)</td>
<td>EmailAddress</td>
</tr>
<tr>
<td>State (String)</td>
<td>StateCode</td>
</tr>
</tbody>
</table>

*Figure 23*

Verifying the Mappings

15. Click **Next**.

16. In the Cleanse step, click **Start**.

17. When the discovery process has completed, in the Profiler pane, to the left, notice the source statistics.

*Source Statistics*

<table>
<thead>
<tr>
<th>Records:</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>0</td>
</tr>
<tr>
<td>(0 %)</td>
<td></td>
</tr>
<tr>
<td>Corrected Records:</td>
<td>4 (20 %)</td>
</tr>
<tr>
<td>Suggested Records:</td>
<td>0 (0 %)</td>
</tr>
<tr>
<td>Invalid Records:</td>
<td>2 (10 %)</td>
</tr>
</tbody>
</table>

*Figure 24*

Reviewing the Source Statistics

18. In the grid, notice the statistics associated with each mapped domain.

The **EmailAddress** field contained no corrected or suggested values, which is expected because the domain includes only a rule and no domain values with corrections.

The **State** field included four corrections for the customers in California (CA) and Florida (FL).
19. Scroll to the very right of the grid to reveal the **Accuracy** column. Hover over each accuracy bar to reveal statistics about the correct, corrected, suggested, new or invalid results.

![Figure 25](image)

*Figure 25*
*Reviewing the Accuracy Statistics for the State Field*

20. Click **Next**.

21. In the **Manage and View Results** step, notice that the **EmailAddress** domain is selected.

22. Notice the tabs and the number of records associated with each classification.

![Figure 26](image)

*Figure 26*
*Reviewing the EmailAddress Tabs*

23. Select the **New** tab, and then scroll to review the valid email addresses.

   **Note:** These are considered new values because there are no domain values defined.

24. Select the **Invalid** tab, and then review the invalid email addresses. Scroll the grid to the right, and then notice that the **Reject** radio buttons for both values are selected. It is possible to approve them if it made sense to do so.

25. In the left pane, select the **StateCode** domain.
26. Select the **New** tab, and notice the one value for *Tex.*. The discovery of this new value would prompt the data steward to manage the domain to translate this value to the state code **TX**. You will not update the domain in this instance.

27. Select the **Corrected** tab, and notice the two values and their corresponding corrected values.

28. To review the individual Californian records, select the **Calif.** row. In the grid below, notice the three records for this state, and that they have a confidence of 100% and that they have been approved.

29. Click **Next**.

30. In the **Export** step, notice the left grid that displays the cleansed records.

31. Scroll to the right of the grid and notice the additional statistical columns that describe reason, confidence and status for the columns mapped to domains.

32. On the right, notice that the cleansed results can be output to a SQL Server table or a CSV file.

33. Click **Finish**.
Exercise 3: Using Integration Services to Cleanse Data

In this exercise, you will review the **LoadDimCustomer** Integration Services package, and then configure the DQS Cleansing component to use the **AdventureWorksBI** knowledge base. You will then execute the package and review the logged invalid values.

**Task 1 – Reviewing the LoadDimCustomer Package**

In this task, you will review the **LoadDimCustomer** package.

1. To open Visual Studio, on the taskbar, click the Visual Studio shortcut.

![Selecting the Visual Studio Shortcut](Figure 27)

*Selecting the Visual Studio Shortcut*

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

3. In the **Open Project** window, navigate to the
   - `D:\SQLServerBI\Lab02\Assets\AdventureWorksBI` folder, select the
   - **AdventureWorksBI.sln** file, and then click **Open**.

4. If the **Error List** opens, close the list to increase space in the designer window.

![Closing the Error List](Figure 28)

*Closing the Error List*

5. In **Solution Explorer**, right-click the **LoadDimCustomer.dtsx** package, and then select **Open**.

6. Read the comments in the control flow design.
7. In the **Load DimCustomer** Data Flow task, disregard the error notification. The error will be addressed when you complete the configuration in the next task.

**Task 2 – Updating the Package Data Flow**

In this task, you will update the package data flow to use the **AdventureWorksBI** knowledge base and map the input columns to the two domains.

1. Right-click the **Load DimCustomer** Data Flow task, and then select **Edit**.

2. Read the comments related to each data flow component. You can also open the components to fully appreciate the package design, but do not modify any properties.

3. Right-click the **DQS Cleansing** component, and then select **Edit**.

4. In the **DQS Cleansing Transformation Editor** window, in the **Data Quality Connection Manager** dropdown list, select **DQS**.

5. In the **Data Quality Knowledge Base** dropdown list, select the **AdventureWorksBI** knowledge base.

6. Select the **Mapping** tab.

7. In the **Available Input Columns** grid, scroll to the bottom of the grid, and then check **EmailAddress**.

8. In the mapping grid, in the **Domain** dropdown list, select **EmailAddress**.

9. Repeat the last two steps to map the **StateProvince** input column to the **StateCode** domain.

10. Verify that your mappings look like the following.

    ![Figure 29 Reviewing the Mappings](image)

    **Figure 29**

    **Reviewing the Mappings**
11. Click **OK**.

12. To review the Conditional Split components, right-click the **Test Customers** component, and then select **Edit**.

![Figure 30: Locating the Conditional Split Components](image)

13. Notice the condition used to isolate valid customers, and then click **Cancel**.

14. Repeat the last two steps to review the **Test EmailAddress** and **Test StateProvince** components to understand the logic used to isolate the invalid records.

**Task 3 – Executing the Package**

In this task, you will execute the **LoadDimCustomer** package and review the execution statistics.

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

2. When the package execution has completed, note the row counts passed to each of the three conditional split components.

![Figure 31: Reviewing the Row Counts](image)

3. To return to design mode, on the **Debug** menu, select **Stop Debugging**.
Task 4 – Reviewing theLogged Data
In this task, you will review the logged invalid data. Appreciate that a data steward could be notified of records inserted into these tables, and that would then prompt them to either fix the data, or update the knowledge base to learn from these rows.

1. To open SQL Server Management Studio, on the taskbar, click the SQL Server Management Studio shortcut.

   ![Selecting the SQL Server Management Studio Shortcut](image)

   Figure 32
   Selecting the SQL Server Management Studio Shortcut

2. In the Connect to Server window, ensure the Server Type dropdown list is set to Database Engine, and that the Server Name text is set to localhost.

3. Click Connect.

4. In Object Explorer (located at the left), expand Databases | AdventureWorksDW2014 | Tables.

5. Right-click the etl.Log_DimCustomer_InvalidEmailAddress table (located second last in the list), and then select Select Top 1000 Rows.

6. Review the rows returned.

7. Repeat the last two steps for the etl.Log_DimCustomer_InvalidStateCode table.

Task 5 – Finishing Up
In this task, you will close all open applications.

1. To close SQL Server Management Studio, on the File menu, select Exit.

2. If prompted to save changes, click No.

3. To close Visual Studio, on the File menu, select Exit.

4. If prompted to save changes, click No.

5. To close the Data Quality Client, in the top right corner, click the X button.
Summary

In this lab, you created a Data Quality Services knowledge base to cleanse US state codes and email addresses. You then created a Data Quality Project to cleanse data in an Excel workbook. Finally, you updated an Integration Services package to use the DQS Cleansing component to cleanse the data as part of an ETL process.