**Data Profiling through IBM Quality Stage**

Many of the assumptions we have about our source data are probably not accurate. Today most of the Organisations are facing a Data Quality problem. In this blog we will look how we can achieve Data Quality using IBM tools.

We have two step processes to achieve Data Quality:

1) Data Profiling

2) Data Quality

We can do Data Profiling Using IBM Information Analyser and IBM Infosphere Data stage and Quality Stage. We apply the Data Quality Rules on source data using Quality Stage Rules. Mainly, these kind of profiling and standardization of source data will be handled in MDM implementation projects to cleanse and standardize the name and address information.

**Setting up and linking an Investigate job**

Create each job by adding Data Quality stages and IBM Info Sphere Data Stage Sequential files and stages to the IBM Info Sphere Data Stage and Quality Stage Designer canvas. Each icon on the canvas is linked together to allow the data to flow from the source file to each stage.

open the Designer client.

1. From the left pane of the Designer, Click **Palette** > **Data Quality** to select the Investigate stage. If you do not see the palette, click **View** > **Palette**.

2. Drag the Investigate stage onto the Designer canvas and drop it in the middle of the canvas.

3. Drag a second Investigate stage and drop it beneath the first **Investigate** stage. You must use two investigate stages to create the data for the reports.

4. Click **Palette** > **File** and select **Sequential File**.

5. Drag the **Sequential File** onto the Designer canvas and drop it to the left of the first Investigate stage. This sequential file is the source file.

6. Click **Palette** > **Processing** and select the Copy stage. This stage duplicates the data from the source file and copies it to the two Investigate stages.

7. Drag the Copy stage onto the Designer canvas and drop it between the **Sequential File** and the first Investigate stage.

8. Click **Palette** > **File**, and drag a second **Sequential File** onto the Designercanvas and drop it to the right of the first Investigate stage. The data from the Investigate stage is sent to the second **Sequential File** which is the target file.

9. Drag a third **Sequential File** onto the Designer canvas and drop it to the right of the Investigate stage and beneath the second **Sequential File**. You now have a source file, a Copy stage, two Investigate stages, and two target files.

11. Drag a fourth **Sequential File** onto the Designer canvas and drop it beneath the third **Sequential File** as the final target file. In the next step, you link all the stages together.

12. Click **Palette** > **General** > **Link**.

a. Right-click and drag a link from the source file to the Copy stage. If the link is red, click to activate the link and drag it until it meets the stage. It should turn black. When all the icons on the canvas are linked, you can click on a stage and drag it to change its position.

b. Continue linking the other stages. The following figure shows the completed Investigate job with the names



**Configuring the source file**

The source data and metadata are attached to the Source File as the source data for the job. The goal of this lesson is to attach the input data of customer names and addresses and load the metadata. To add data and metadata to the Investigate job, configure the source file to locate the input data file **MOCK\_DATA.csv** stored on your computer and load the metadata columns.

To configure the source file:

1. Double-click the **Source File** icon to open the **Properties** tab on the Source File - Sequential File window.

a. Click **Source** > **File** to activate the **File** field.

b. Click in the **File** field and select **Browse for File**.

c. Locate the directory on the server where you copied the input.csv file (for example, C:\IBM\InformationServer\Server\Projects\).

d. Click **MOCK\_DATA.csv** to select the file, then click **OK**.

2. Click **Options** > **First Line is Column Names**, and then select **True** from the **First Line is Column Names** list.

3. Click the **Columns** tab.

4. Click **Load**.

5. From the Table Definitions window, Browse the path (Ex : /home/dsadm/users/Venkat)

6. Select the MOCK\_DATA.csv file and click on IMPORT.

7. Click **Input**, in the Table Definitions folder, and click **OK**.

8. Click **OK** in the Select Columns window to load the sample metadata.

9. Click **View Data** to display the quality of the input data.

10. In the first Data Browser window for the output link, select the number of

rows to display and click **OK**. You can leave the number of rows as 100.

11. In the second Data Browser for the output link, you see bank customer names and addresses. The addresses are shown in a disorganized way making it difficult for the bank to analyze the data.

12. Click **Close** to close the Data Browser window.

13. Click **OK** to update the Sequential File stage with the changes that you made.

**Configuring the Copy stage**

The Copy stage duplicates the source data and sends it to the two Investigate stages. This lesson explains how to configure a Processing stage, the Copy stage, to duplicate the source and send the output metadata to the two Investigate stages.

To configure a Copy stage:

1. Double-click the Copy stage icon to open the **Properties** tab on the Copy - Copy window.

2. Click the **Input** > **Columns** tab. The metadata you loaded in the Source File has propagated to the Copy stage.

3. Click the **Output** > **Mapping** tab to map the columns in the left **Columns** pane to the right **To Name** pane.

4. In the **Output name** field above the **Columns** pane of the screen, select **To Name** if it is not already selected.

Selecting the correct output name ensures that the data goes to the correct Investigate stage, Investigate Name, or Investigate City State Zip stage.

5. Copy the data from the **Columns** pane to the **To Name** pane:

a. Place your cursor in the **Columns** pane; right-click and select **Select All** from the shortcut menu.

b. Right-click and select **Copy** from the shortcut menu.

c. Place your cursor in the **To Name** pane, right-click and select **Paste Column** from the shortcut menu. The column metadata is copied into the **To Name** pane and lines are displayed to show the linking from the **Columns** pane tothe **To Name** pane.

6. In the **Output name** field above the **Columns** pane, select **To City State Zip** from the drop-down menu.

7. Repeat step 5 to map the **Columns** pane to the **To City State Zip** pane.

8. Click **OK** to save the updated Copy stage.

**Investigate stage, configuring to review names**

The Word Investigate option of the Investigate stage parses name and address data into recognizable patterns by using rule sets that classify personal names and addresses. The Investigate stage analyzes each record from the source file. In this lesson, you select the NAME rule set to apply USPS standards.

**To configure the Investigate stage:**

1. Double-click the **Investigate Name** icon.

2. Click the **Word Investigate** tab to open the Word Investigate window.

3. Select **Name** from the **Available Data Columns** section and click > to move the **Name** column into the **Standard Columns** pane. The Investigate Name stage analyzes the **Name** column by using the rule set that you select in step 4.

4. In the **Rule Set:** field, click to select a rule set for the Investigate Name stage.

a. In the Rule Sets window, double-click the **Standardization Rules** folder to open the Standardization Rules tree.

b. Double-click the **USA** folder, double-click the **USNAME** folder, and then select **USNAME**. The USNAME rule set parses the **Name** column according to United States Post Office standards for names.

c. Click **OK** to exit the Rule Sets window.

5. Click the **Token Report** check box in the **Output Dataset** section of the window.

6. Click the **Stage Properties** > **Output** > **Mapping** tab.

7. Map the output columns:

a. Click the **Columns** pane.

b. Right-click and select **Select All** from the shortcut menu.

c. Right-click and select **Copy** from the shortcut menu.

d. Click in the **Name Investigation** pane.

e. Right-click and select **Paste Column** from the shortcut menu. The columns on the left side map to the columns on the right side. Your **Name Investigation** map should look like the following figure:



8. Click the **Columns** tab. Notice that the Output columns are populated when you map the columns in the **Mapping** tab.

9. Extend the data type for the qsInvWord and qsInvClassCode columns.

a. In the row of the **Columns** grid for the qsInvWord column, select **Unicode** from the **Extended** list.

b. Repeat step 9a for the qsInvClassCode column.

c. Click **OK**.

10. Click **OK**, then click **File** > **Save** to save the updated investigation job.

**Investigate stage, configuring to review geographic regions**

The Word Investigate option of the Investigate stage parses name and address data into recognizable patterns by using rule sets that classify personal names and addresses. The Investigate stage analyzes each record from the source file. In this lesson, you apply the USAREA rule set to apply USPS standards.

**To configure the InvestigateCityStateZip icon:**

1. Double-click the **InvestigateCityStateZip** icon.

2. Click the **Word Investigate** tab to open the Word Investigate window.

3. Select the following columns in the **Available Data Columns** pane to move to the **Standard Columns** pane. The second Investigate stage analyzes the address columns by using the rule set that you select in step 5.

V City

V State

V Zip5

V Zip4

4. Click to move each selected column to the **Standard Columns** pane.

5. In the **Rule Set:** field, click to locate a rule set for InvestigateCityStateZip.

a. In the Rule Sets window, double-click the **Standardization Rules** folder to open the Standardization Rules tree.

b. Double-click the **USA** folder and double-click on the **USAREA** folder and select the **USAREA** file. The USAREA rule set parses the City, State, Zip5 and Zip4 columns according to the United States Post Office standards.

c. Click **OK** to exit the Rule Sets window. USAREA.SET is shown in the **Rule Set** field.

6. Click the **Token Report** and **Pattern Report** check boxes in the **Output Dataset** section of the window. When you assign data to 2 outputs, you mustverify that the link ordering is correct. Link ordering assures that the data issent to the correct reports through the assigned links that you named in

7. If you need to change the display order of the links, click the **Stage Properties** > **Link Ordering** tab and select the output link that you want to move.

8. Move the links up or down as described next:

The following figure shows the correct order for the links.



9. Click the **Output** > **Mapping** tab. Since there are two output links from the second Investigate stage, you must map the columns to each link:

a. From the **Output name** list above the **Columns** pane, select **PatternReport**.

b. Select the **Columns** pane.

c. Right-click and select **Select All** from the shortcut menu.

d. Right-click and select **Copy** from the shortcut menu.

e. Select the **Pattern Report** pane, right-click and select **Paste Column** from the shortcut menu. The columns are mapped to the **Pattern Report** output link.

f. From the **Output name** list above the **Columns** pane, select **TokenReport**.

g. Repeat steps b through e, except select the **Token Report** pane in step e.

10. Extend the data type for the columns for both output links.

a. Click the **Output** > **Columns** tab.

b. From the **Output name** list above the **Columns** page, select **Pattern Report**.

c. In the row of the **Columns** grid for each column that has an SQL type of Var Char or Char, select **Unicode** from the **extended** list.

d. From the **Output name** list above the **Columns** page, select **Token Report**.

e. Repeat step 10c for the columns for this output link.

f. Click **OK**.

11. Click **OK** to close the InvestigateCityStateZip window.

**Configuring target reports**

The source data information and column metadata are propagated to the target data files for later use in creating Investigation reports. The Investigate job modifies the unformed source data into readable data which is later configured into Investigation reports.

**To configure the data files:**

1. Double-click the **Name Token Report** icon on the Designer client canvas.

2. In **Input** > **Properties**, click **Target** > **File**.

3. In the **File** field, click and browse to the path name of the folder on the server computer where the input data file resides. In the following steps, you are going to specify target file names on stage input tabs.

4. In the **File name** field, type tokrpt.csv to display the path and file name in the **File** field, (for example, C:\IBM\InformationServer\Server\Projects\venkat\Name\_tokrpt.csv).

5. Specify how the stage collects data before the stage writes the data to the sequential file. You specify collection details because the Investigate stage runs in parallel mode and the Sequential File stage runs in sequential mode.

a. Click the **Partitioning** tab.

b. From the **Collector type** list, select **Ordered**. This method reads all of the rows from the first partition, then all of the rows from the second partition, and so on.

c. In the **Sorting** section, select **Perform sort**.

d. From the available columns, click **qsInvCount**.

e. Click **OK**.

6. Double-click the **CityStateZipPatternReport** icon.

7. Repeat steps 2 to 5 except type areapatrpt.csv for the file name.

8. Double-click the **CityStateZipTokenReport** icon.

9. Repeat steps 2 to 5 except type areatokrpt.csv for the file name.

**Compiling and running jobs**

Test the Investigate job by running the compiler followed by running the job to process the data for the reports.

Compile the Investigate job in the Designer client. After the job compiles successfully, open the Director client and run the job.

**To compile and run the job:**

1. Click **File** > **Save** to save the Investigate job on the Designer canvas.

2. Click to compile the job. The Compile Job window opens and the job begins to compile. When the compiler finishes, the following message is shown

Job successfully compiled with no errors.

3. Click **Tools** > **Run Director**. The Director application opens with the job shown in the status view.

4. Click to open the Job Run Options window.

5. Click **Run**.

After the job runs, **Finished** is shown in the **Status** column.

**Name Token Report**



Token report output columns:

QsInvCount indicates the number of times this token was encountered across the entire input data.

QsInvWord identifies the individual token or word value that is found inside the selected input columns.

QsInvClassCode identifies the classification of the token that is based on the selected rule set classification table. Unclassified tokens, if selected, get a question mark “?” for alpha or a carat “^” for numeric.

**City State Zip Pattern Report**



**Pattern Report output Columns:**

QsInvColumnName indicates the names of the column that is investigated.

QsInvPattern displays the character and includes the character in the frequency count and pattern analysis.

QsInvSample shows one or more samples of the content of this column. The number to be displayed is configurable.

QsInvCount shows the actual number of occurrences of the value in the qsInvPattern column.

QsInvPercent shows the percentage occurrences of the value in the qsInvPattern column to the total number of records on this file.

**City State Zip Token Report**



QsInvCount indicates the number of times this token was encountered across the entire input data.

QsInvWord identifies the individual token or word value that is found inside the selected input columns.

QsInvClassCode identifies the classification of the token that is based on the selected rule set classification table. Unclassified tokens, if selected, get a question mark “?” for alpha or a carat “^” for numeric.