IBM InfoSphere QualityStage
Version 11 Release 3

Tutorial

SC19-4327-00

IBM InfoSphere QualityStage
Version 11 Release 3

Tutorial

SC19-4327-00

Note

Before using this information and the product that it supports, read the information in “Notices and trademarks” on page 49.

Contents

InfoSphere QualityStage parallel job tutorial

Use the parallel job tutorial to learn the basic skills that you need to develop parallel jobs.

"Tutorial project goals" on page 3 are to use Designer client stages to cleanse customer data to remove all the duplicates of customer addresses and provide a best case for the correct address.

About IBM InfoSphere QualityStage

IBM® InfoSphere® QualityStage® is a data cleansing component that is part of the IBM InfoSphere DataStage® and QualityStage Designer client.

The Designer client provides a common user interface in which you design your data quality jobs. In addition, you have the power of the parallel processing engine to process large stores of source data.

The integrated stages available in the Repository provide the basis for accomplishing the following data cleansing concepts:

- Resolving data conflicts and ambiguities
- Uncovering new or hidden attributes from free-form or loosely controlled source columns
- Conforming data by transforming data types into a standard format
- Creating one unique result

Learning objectives

The key points that you should keep in mind as you complete this tutorial include the following topics:

- How the processes of standardization and matching improve the quality of the data
- The ease of combining both InfoSphere DataStage and QualityStage Designer client stages in the same job
- How the data flows in an iterative process from one job to another
- The surviving data results in the best available record

Projects in IBM InfoSphere QualityStage

The IBM InfoSphere DataStage and QualityStage Designer client provides a view to projects. The projects are a method for organizing your re-engineered data. You define data files and stages and you build jobs in a specific project. IBM InfoSphere QualityStage uses these projects to create and store files on the client and server.

Each InfoSphere QualityStage project contains the following components:

- InfoSphere QualityStage jobs
- Stages that are used to build each job
- Match specifications
- Standardization rules
- Table definitions

In this tutorial, you will create a project and use the data that is provided to create jobs in the project.

About InfoSphere QualityStage jobs

IBM InfoSphere QualityStage uses jobs to process data.

To start a InfoSphere QualityStage job, you open the Designer client and create a new Parallel job. You build the InfoSphere QualityStage job by adding stages, source and target files, and links from the Repository, and placing them onto the Designer canvas. The Designer client compiles the Parallel job and creates an executable file. When the job runs, the stages process the data by using the data properties that you defined. The result is a data set that you can use as input for the next job.

In this tutorial, you build four InfoSphere QualityStage jobs. Each job is built around one of the Data Quality stages and additional IBM InfoSphere DataStage stages.

IBM InfoSphere DataStage and QualityStage stages

A stage in IBM InfoSphere DataStage and QualityStage performs an action on data. The type of action depends on the stage that you use.

The stages in the InfoSphere DataStage and QualityStage Designer client are stored in the Designer tool palette. You can access all the IBM InfoSphere QualityStage stages in the Data Quality group in the palette. You configure each stage to perform the type of actions on the data that obtain the required results. Those results are used as input data to the next stage. The following stages are included in InfoSphere QualityStage:

- Investigate stage
- Standardize stage
- Match Frequency stage
- One-source Match stage
- Two-source Match stage
- Survive stage
- Standardization Quality Assessment (SQA) stage

In this tutorial, you use most of the InfoSphere QualityStage stages.

You can also add any of the IBM InfoSphere DataStage stages to your job. In some of the lessons, you add InfoSphere DataStage stages to enhance the types of tools for processing the data.

Server and client components

You load the client and server components that are used to build jobs to cleanse data.

The following server components are installed on the server:

Repository
: A central store that contains all the information required to build an IBM InfoSphere QualityStage job.

InfoSphere Information Server engine
: Runs the InfoSphere QualityStage jobs.

The following InfoSphere DataStage client components are installed on a personal computer:

- IBM InfoSphere DataStage and QualityStage Designer
- IBM InfoSphere DataStage and QualityStage Director
- IBM InfoSphere DataStage and QualityStage Administrator

In this tutorial, you use all of these components when you build and run your InfoSphere QualityStage project.

Tutorial project goals

The goal of this tutorial is to use IBM InfoSphere DataStage and QualityStage Designer stages to cleanse customer data by removing all the duplicates of customer addresses and providing a best case for the correct address.

In this tutorial, you have the role of a database analyst for a bank that provides many financial services. The bank has a large database of customers; however, there are problems with the customer list because it contains multiple names and address records for a single household. Because the marketing department wants to market additional services to existing customers, you need to find and remove duplicate addresses.

For example, a married couple has four accounts, each in their own names. The accounts include two checking accounts, an IRA, and a mutual fund.

In the bank's existing system, customer information is tracked by account number rather than customer name, number, or address. For this one customer, the bank has four address entries.

To save money on the mailing, the bank wants to consolidate the household information so that each household receives only one mailing. In this tutorial, you are going to use InfoSphere QualityStage to standardize all customer addresses. In addition, you need to locate and consolidate all records of customers who are living at the same address.

Learning objectives

The purpose of this tutorial is to provide a working knowledge of the InfoSphere QualityStage process flow through the jobs. In addition, you learn how to do the following tasks:

- Set up each job in the project
- Configure each stage in the job
- Assess results of each job
- Apply those results to your business practices

After you complete these tasks, you should understand how InfoSphere QualityStage stages restructure and cleanse the data by using applied business rules.

This tutorial will take approximately 2.5 hours to complete.

Skill level

To use this tutorial, you need an intermediate to advanced level of understanding of data analysis.

Audience

This tutorial is intended for business analysts and systems analysts who are interested in understanding InfoSphere QualityStage.

System requirements

- IBM InfoSphere Information Server
- Microsoft Windows XP or Linux operating systems

Prerequisites

To complete this tutorial, you need to know how to use

- IBM InfoSphere DataStage and QualityStage Designer
- Personal computers

Expected results

Upon the completion of this tutorial, you should be able to use the Designer client to create your own InfoSphere QualityStage projects to meet the business requirements and data quality standards of your company.

Setting up the tutorial

The setup process for the tutorial includes creating a folder, creating a project, copying the job and input data file to the project and then starting the project. You must complete the setup tasks before you begin Module 1 of the tutorial.

Creating a folder for the tutorial files

Copy the tutorial files to a folder that you create on your IBM InfoSphere QualityStage client computer.

Procedure

1. Create a folder on your computer (for example, `C:\TutorialData`).
2. Locate the `Parallel_job_tutorial.zip` file, which is on the installation media. In the directory that contains the installation media, the `Parallel_job_tutorial.zip` file is in the *`parent_directory`*`\TutorialData\QualityStage` directory. For example, the `Parallel_job_tutorial.zip` file might be in the `is-client\TutorialData\QualityStage` directory.
3. Extract the files from the `Parallel_job_tutorial.zip` file to the folder that you created in step 1.

Creating the tutorial project

Create a new project for the tutorial to keep your tutorial exercises separate from the other work on InfoSphere QualityStage.

About this task

You must have IBM InfoSphere QualityStage Administrator privileges.

To create the tutorial project:

Procedure

1. Select **Start** > **All Programs** > **IBM InfoSphere Information Server** > **IBM InfoSphere DataStage and QualityStage Administrator**.
2. In the Attach to DataStage window, type your user name and password, and click **OK**.
3. In the Projects tab, click **Add** to open the Add New Project window.
4. In the **Name** field, specify the name of the new project (for example, Tutorial).
5. Click **OK** to create the new project.
6. Click **Close** to close the Administrator client.

Copying tutorial data

Copy the tutorial data files from the tutorial folder you created on the client computer to the project folder or directory on the IBM InfoSphere QualityStage computer where the engine tier is installed.

About this task

When you created the project for the tutorial, you automatically created a folder or directory for the tutorial project on the computer where the engine tier is installed. The InfoSphere Information Server engine tier can be installed on the same Windows computer as the clients, or it can be on a separate Windows, UNIX, or Linux computer. Sometimes the engine tier is referred to as the IBM InfoSphere DataStage and QualityStage server.

Procedure

1. Open the tutorial folder `TutorialData\QualityStage` that you created on the client computer and locate the input.csv file.
2. Open the project folder on the computer where the engine tier is installed for the tutorial project you created. Where *tutorial_project* is the name of the project you created, examples of path names are:
   - For a Windows server: `C:\IBM\InformationServer\Server\Projects\`*`tutorial_project`*
   - For a UNIX or Linux server: `opt/IBM/InformationServer/Server/Projects/`*`tutorial_project`*
3. Copy the .csv file to the project folder on the server.

Starting a project

Use a project in the IBM InfoSphere DataStage and QualityStage Designer client as a container for your IBM InfoSphere QualityStage jobs.

About this task

Open the Designer client to begin the tutorial. The Designer Parallel job provides the executable file that runs your InfoSphere QualityStage jobs.

Procedure

1. Click **Start** > **All Programs** > **IBM InfoSphere Information Server** > **IBM InfoSphere DataStage and QualityStage Designer**. The Attach to Project window opens.
2. In the **Domain** field, type the name of the server that you are connected to.
3. In the **User name** field, type your user name.
4. In the **Password** field, type your password.
5. In the **Project** field, select the project you created (for example, Tutorial).
6. Click **OK**. The New window opens in the Designer client.

Creating a job

The IBM InfoSphere DataStage and QualityStage Designer client provides the interface to the parallel engine that processes the jobs. You are going to save a job to a folder in the metadata repository.

Before you begin

If it is not already open, open the client.

Procedure

1. From the New window, select the **Jobs** folder in the left pane and then select the **Parallel Job** icon in the right pane.
2. Click **OK**. A new empty job design window opens in the job design area.
3. Click **File** > **Save**.
4. In the Save Parallel Job As window, right-click the Jobs folder and select **New** > **Folder** from the shortcut menu.
5. Type in a name for the folder (for example, MyTutorial).
6. Click the new folder (MyTutorial) and in the **Item name** field, type `Investigate1`.
7. Click **Save** to save the job.

Results

You have created a new parallel job named Investigate and saved it in the folder Jobs\MyTutorial in the repository. Using these procedures, create 3 more parallel jobs in this folder and name them Standardize1, Unduplicate1, and Survive1.

What to do next

Import the tutorial data into your project.

Importing tutorial components

Use the IBM InfoSphere DataStage and QualityStage Designer client to import the tutorial components, which include sample jobs and table definitions, into the tutorial project.

About this task

Import the tutorial components to begin the tutorial lessons.

To import tutorial components:

Procedure

1. Select **Start** > **All Programs** > **IBM InfoSphere Information Server** > **IBM InfoSphere DataStage and QualityStage Designer**.
2. In the Attach to DataStage window, type your user name and password.
3. Select the **Tutorial** project from the **Project** list and click **OK**. The Designer client opens and displays the New window.
4. Click **Cancel** to close the New window because you are opening an existing job, not creating a new job or object.
5. Select **Import** > **DataStageComponents**.
6. In the **Import from file** field, go to the directory on the client into which you copied the tutorial data, for example: C:\TutorialData\QualityStage. Select the **QualityStage_Tutorial.dsx** file.
7. Ensure **Import All** is selected. You can also select **Perform impact analysis**.
8. Click **OK** to import the sample jobs and table definitions into a repository folder named QualityStage Tutorial.

Results

The components are displayed in the repository under **QualityStage Tutorial** folder. You can open each job and look at how they are designed on the canvas. Use these jobs as a reference when you create your own jobs.

What to do next

You can begin Module 1.

Module 1: Investigating source data

This module explains how to set up and process an Investigate job to provide data from which you can create reports in the IBM Information Server Web console.

You can use the information in the reports to make basic assumptions about the data and the steps you must take to attain the goal of providing a legitimate address for each customer in the database.

Learning objectives

After completing the lessons in this module, you should know how to do the following tasks:

- Add IBM InfoSphere DataStage and QualityStage stages and links to a job
- Configure stage properties to specify which action they take when the job is run
- Load and process customer data and metadata
- Compile and run a job
- Produce data for reports

This module should take approximately 30 minutes to complete.

Lesson 1.1: Setting up and linking an Investigate job

Create each job by adding Data Quality stages and IBM InfoSphere DataStage sequential files and stages to the IBM InfoSphere DataStage and QualityStage Designer canvas. Each icon on the canvas is linked together to allow the data to flow from the source file to each stage.

If you have not already done so, open the Designer client.

1. From the left pane of the Designer, go to the MyTutorial folder you created for this tutorial and double click on Investigate1 to open the job.
2. Click **Palette** > **Data Quality** to select the Investigate stage.

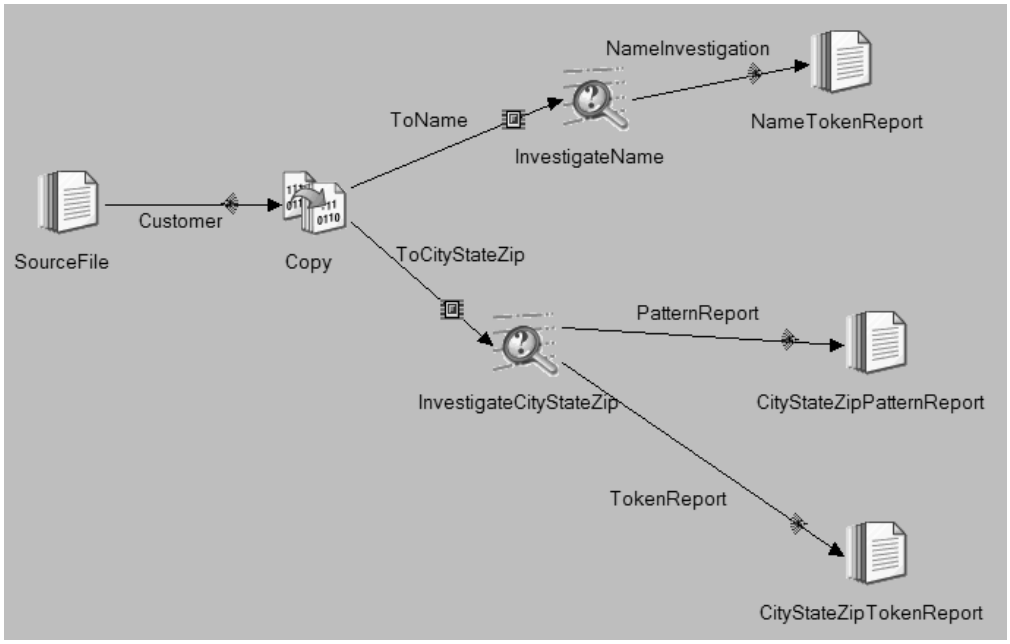
   If you do not see the palette, click **View** > **Palette**.
3. Drag the Investigate stage onto the Designer canvas and drop it in the middle of the canvas.
4. Drag a second Investigate stage and drop it beneath the first **Investigate** stage. You must use two investigate stages to create the data for the reports.
5. Click **Palette** > **File** and select **Sequential File**.
6. 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.
7. 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.
8. Drag the Copy stage onto the Designer canvas and drop it between the **Sequential File** and the first Investigate stage.
9. Click **Palette** > **File**, and drag a second **Sequential File** onto the Designer canvas 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.
10. 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 you will assign to the stages and links in the next lesson.



Lesson checkpoint

When you set up the Investigate job, you are connecting the source file and its source data and metadata to all the stages and linking the stages to the target files.

In completing this lesson, you learned the following about the Designer:

- How to add stages to the Designer canvas
- How to combine Data Quality and Processing stages on the Designer canvas
- How to link all the stages together

Lesson 1.2: Renaming links and stages in an Investigate job

When creating a large job in the IBM InfoSphere DataStage and QualityStage Designer client, it is important to rename each stage, file, and link with meaningful names to avoid confusion when selecting paths during stage configuration.

When you rename the links and stages, do not use spaces. The Designer client resets the name back to the generic value if you enter spaces. The goal of this lesson is to replace the generic names for the icons on the canvas with more appropriate names.

To rename icons on the canvas:

1. To rename a stage, complete the following steps:
   a. Click the name of the source SequentialFile until a highlighted box appears around the name.
   b. Type `SourceFile` in the box.
   c. Click outside the box to deselect the box.
2. To rename a link, complete the following steps:
   a. Right-click the generic link name DSLinkXX that connects SourceFile to the Copy stage and select **Rename** from the shortcut menu. A highlighted box appears around the default name.
   b. Type `Customer` and click outside the box. The default link name changes to Customer.
3. Right-click the generic link name that connects the Copy stage to the first Investigate stage.
4. Repeat step 2, except type `ToName` in the box.
5. Right-click the generic link name that connects the Copy stage to the second Investigate stage.
6. Repeat step 2, except type `ToCityStateZip` in the box.
7. Click on the names of the following stages and type the new stage name in the highlighted box:

Stage	**Change to**
Copy	`Copy`
Investigate (the first one)	`InvestigateName`
Investigate (the second one)	`InvestigateCityStateZip`

8. Rename the three target files from the top in the following order:
   a. NameTokenReport
   b. CityStateZipPatternReport
   c. CityStateZipTokenReport
9. On the names of the following links, select **Rename** and type the new link name in the highlighted box:

Link	**Change to**
From InvestigateName to NameTokenReport	`NameInvestigation`
From InvestigateCityStateZip to CityStateZipPatternReport	`PatternReport`
From InvestigateCityStateZip to CityStateZipTokenReport	`TokenReport`

Renaming the elements on the Designer canvas provides better organization to the Investigate job.

Lesson checkpoint

In this lesson, you changed the generic stages and links to names appropriate for the job.

You learned the following tasks:

- How to select the default name field in order to edit it

- The correct method to use in changing the name

Lesson 1.3: Configuring the source file

The source data and metadata are attached to the SourceFile 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 **input.csv** stored on your computer and load the metadata columns.

To configure the source file:

1. Double-click the **SourceFile** icon to open the **Properties** tab on the SourceFile - Sequential File window.
2. Select the tutorial data file:
   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 from the DVD (for example, C:\IBM\InformationServer\Server\Projects\tutorial).
   d. Click **input.csv** to select the file, then click **OK**.
3. Click **Options** > **First Line is Column Names**, and then select **True** from the **First Line is Column Names** list.
4. Click the **Columns** tab.
5. Click **Load**.
6. From the Table Definitions window, click the **QualityStage Tutorial** > **Table Definitions** folder. This folder was created when you imported the tutorial sample metadata.
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.

Lesson checkpoint

In this lesson, you attached the input data (customer names and addresses) and loaded the metadata.

You learned how to do the following tasks:

- Attaching source data to the source file
- Adding column metadata to the source file

Lesson 1.4: 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 SourceFile has propagated to the Copy stage.
3. Click the **Output** > **Mapping** tab to map the columns in the left **Columns** pane to the right **ToName** pane.
4. In the **Output name** field above the **Columns** pane of the screen, select **ToName** if it is not already selected. Selecting the correct output name ensures that the data goes to the correct Investigate stage, InvestigateName, or InvestigateCityStateZip stage.
5. Copy the data from the **Columns** pane to the **ToName** 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 **ToName** pane, right-click and select **Paste Column** from the shortcut menu. The column metadata is copied into the **ToName** pane and lines are displayed to show the linking from the **Columns** pane to the **ToName** pane.
6. In the **Output name** field above the **Columns** pane, select **ToCityStateZip** from the drop-down menu.
7. Repeat step 5 to map the **Columns** pane to the **ToCityStateZip** pane.
8. Click **OK** to save the updated Copy stage.

This procedure shows you how to map columns to two different outputs.

Lesson checkpoint

In this lesson, you mapped the input metadata to the two output links to continue the propagation of the metadata to the next two stages.

You learned how to do the following tasks:

- Adding a IBM InfoSphere DataStage stage to a IBM InfoSphere QualityStage job
- Propagating metadata to the next stage
- Mapping metadata to two output links

Lesson 1.5: Investigate stage, configuring to review names

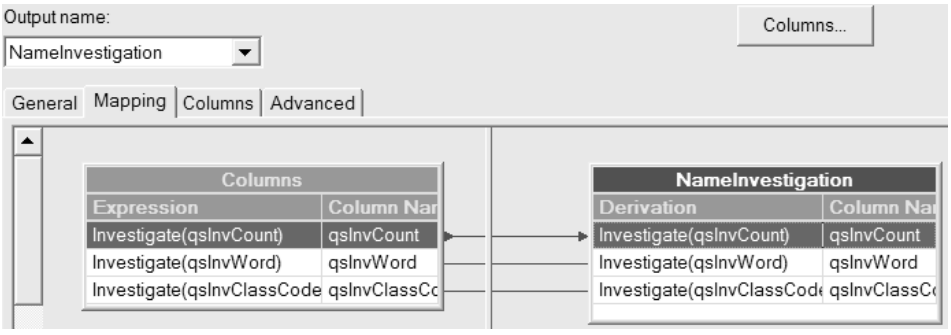
The Word Investigate option of the Investigate stage parses name and address data into recognizable patterns by using rule sets that classify personal names and addresses.

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 **InvestigateName** 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 InvestigateName 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 InvestigateName 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, 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 **NameInvestigation** 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 **NameInvestigation** 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.

Lesson summary

This lesson explained how to configure the Investigate stage by using the USNAME rule set.

You learned how to configure the Investigate stage in the Investigate job by doing the following tasks:

- Selecting the columns to investigate
- Selecting a rule set to apply to the data
- Mapping the output columns

Lesson 1.6: 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.
   - City
   - State
   - Zip5
   - 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 must verify that the link ordering is correct. Link ordering assures that the data is sent to the correct reports through the assigned links that you named in Lesson 1.2. The **Link Ordering** tab is not displayed if there is only one link.
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:

- Click  to move the link name up a level.
- Click  to move the link name down a level.

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 **PatternReport** pane, right-click and select **Paste Column** from the shortcut menu. The columns are mapped to the **PatternReport** output link.
   f. From the **Output name** list above the **Columns** pane, select **TokenReport**.
   g. Repeat steps b through e, except select the **TokenReport** 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 **PatternReport**.
   c. In the row of the **Columns** grid for each column that has an SQL type of VarChar or Char, select **Unicode** from the **Extended** list.
   d. From the **Output name** list above the **Columns** page, select **TokenReport**.
   e. Repeat step 10c for the columns for this output link.
   f. Click **OK**.
11. Click **OK** to close the InvestigateCityStateZip window.

Lesson summary

This lesson explained how to configure the second Investigate stage to the AREA rule set.

You learned how to configure the second Investigate stage in the Investigate job by doing the following topics:

- Selecting the columns to investigate
- Selecting a rule set to apply to the data
- Verifying the link ordering for the output reports
- Mapping the output columns to two output links

Lesson 1.7: 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 **NameTokenReport** 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\tutorial\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.

Lesson checkpoint

This lesson explained how to configure the target files for use as reports.

You configured the three target data files by linking the data to each report file.

Lesson 1.8: 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.

Lesson checkpoint

In this lesson, you learned how to compile and process an Investigate job.

You processed the data into three output files by doing the following tasks:

- Compiling the Investigate job
- Running the Investigate job in the Director

Module 1: Summary

In Module 1, you set up, configured, and processed an Investigate job in IBM InfoSphere DataStage and QualityStage Designer.

An Investigate job looks at each record column-by-column and analyzes the data content of the columns that you select. The Investigate job loads the name and address source data stored in the database of the bank, parses the columns into a form that can be analyzed, and then organizes the data into three data files.

The Investigate job modifies the unformed source data into readable data that you can configure into Investigation reports using the IBM InfoSphere Information Server Web console. You select **QualityStage Reports** to access the reports interface in the Web console.

The next module organizes the unformed data into standardized data that provides usable data for matching and survivorship.

Lessons learned

By completing this module, you learned about the following concepts and tasks:

- How to correctly set up and link stages in a job so that the data propagates from one stage to the next
- How to configure the stage properties to apply the correct rule set to analyze the data
- How to compile and run a job
- How to create data for analysis

Module 2: Standardizing data

This module explains how to set up and process a Standardize job to standardize name and address information derived from the database of the bank.

When you worked on the data in Module 1, some addresses were free form and nonstandard. Removing duplicates of customer addresses and guaranteeing that a single address is the correct address for that customer would be very difficult without standardizing the data.

Standardizing or conditioning ensures that the source data is internally consistent, that is, each type of data has the same type of content and format. When you use consistent data, the system can match address data with greater accuracy by using one of the matching stages.

Learning objectives

After completing the lessons in this module, you should know how to do the following tasks:

1. Add stages and links to a Standardize job
2. Configure the various stage properties to correctly process the data when the job is run
3. Work with handling nulls by using derivations
4. Generate the frequency distribution for standardized data

This module should take approximately 60 minutes to complete.

Lesson 2.1: Setting up a Standardize job

Standardizing data is the first step in data cleansing. In Lesson 2.1, you add a variety of stages to the IBM InfoSphere DataStage and QualityStage Designer canvas. These stages include the Transformer stage which applies derivations to handle nulls and the Match Frequency stage which adds frequency data.

If you have not already done so, open the Designer client.

As you learned in Lesson 1.1, you must add stages and links to the Designer canvas to create a standardize job. The Investigate job that you completed helped you determine how to formulate a business strategy by using Investigation reports. The Standardize job applies rule sets to the source data to condition it for matching.

To set up a Standardize job:

1. From the left pane of the Designer, go to the MyTutorial folder you created for this tutorial and double click on Standardize1 to open the job.
2. Drag the following icons onto the Designer canvas from the palette.
   - **Data Quality** > **Standardize** icon to the middle of the canvas
   - **File** > **Sequential File** icon to the left of the Standardize stage
   - **File** > **Data Set** icon to the right of the Standardize stage
   - **Processing** > **Transformer** icon between the Standardize stage and the **Data Set** file
   - **Processing** > **Copy** icon between the Transformer stage and the **Data Set** file
   - **Data Quality** > **Match Frequency** icon below the Copy stage

- Second **File** > **Data Set** icon to the right of the Match Frequency stage

After linking the stages and files, you can adjust their location on the canvas.

3. Right-click the **Sequential File** icon and drag to create a link from the **Sequential File** icon to the Standardize stage icon.
4. Drag links to the remaining stages like you did in step 3.

   If the link is red, click to activate the link and drag it until it meets the stage. It should turn black.

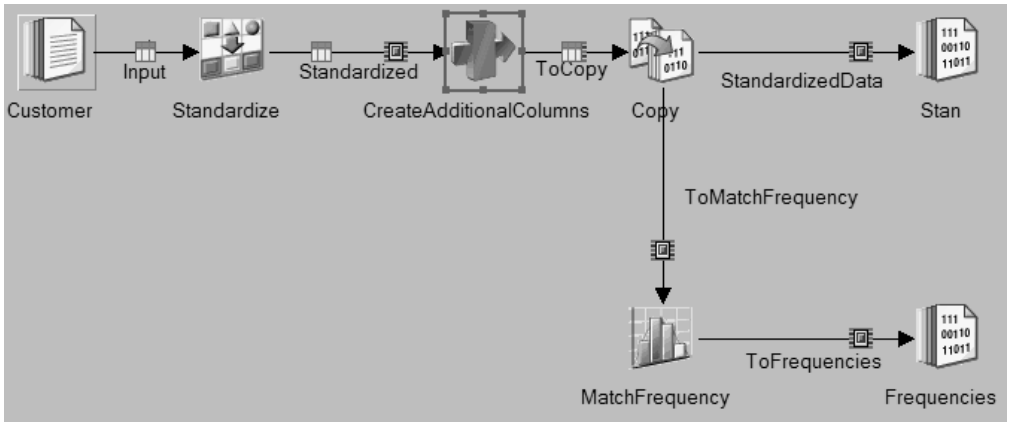
   When all the icons on the canvas are linked, you can click on the stages and drag them to change their positions.
5. Click on the names of the following stages and type the new stage name in the highlighted box:

Stage	**Change to**
SequentialFile	`Customer`
Standardize stage	`Standardize`
Transformer stage	`CreateAdditionalColumns`
Copy stage	`Copy`
Data_Set file (the output of the Copy stage)	`Stan`
Match Frequency stage	`MatchFrequency`
Data_Set file (the output of the Match Frequency stage)	`Frequencies`

6. Right-click on the names of the following links, select **Rename** and type the new link name in the highlighted box:

Link	**Change to**
From Customer to Standardize	Input
From Standardize to CreateAdditionalColumns	Standardized
From CreateAdditionalColumns to Copy	ToCopy
From Copy to Stan	StandardizedData
From Copy to MatchFrequency	ToMatchFrequency
From MatchFrequency to Frequencies	ToFrequencies

The following figure shows the Standardized job stages and links.



Lesson checkpoint

In this lesson you learned how to set up a Standardize job. The importance of the Standardize stage is to generate the type of data that can then be used in a match job.

You set up and linked a Standardize job by doing the following tasks:

- Adding Data Quality and Processing stages to the Designer canvas
- Linking all the stages
- Renaming the links and stages

Lesson 2.2: Configuring the Standardize job stage properties

The properties for each of the stages in the Standardize job must be configured on the IBM InfoSphere DataStage and QualityStage Designer canvas.

Complete the following tasks to configure the Standardize job:

- Load the source data and metadata
- Add compliant rule sets for United States names and addresses
- Apply derivations to null sets
- Copy data to the two output links
- Create frequency data

Configuring the Customer file properties

To configure the Customer (source file) stage properties:

1. Double-click the **Customer** source file icon to open the **Properties** tab on the Customer - Sequential File window.
2. Click **Source** > **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.
4. Select `input.csv` and then click **OK**. This is the source file the Standardize stage reads when the job runs.
5. Click **Options** > **First Line is Column Names**, and then select **True** from the **First Line is Column Names** list.
6. Click the **Columns** tab and click **Load**.
7. From the Table Definitions window, click the **QualityStage Tutorial** folder. This folder was created when you imported the tutorial sample metadata.
8. Click **Table Definitions** > **Input**. The table definitions load into the **Columns** tab of the **Customer** source file.
9. Click **OK** to close the Table Definitions window.
10. Click **OK** again in the Select Columns window to load the sample metadata.
11. Click **OK** to close the **Customer** source file.

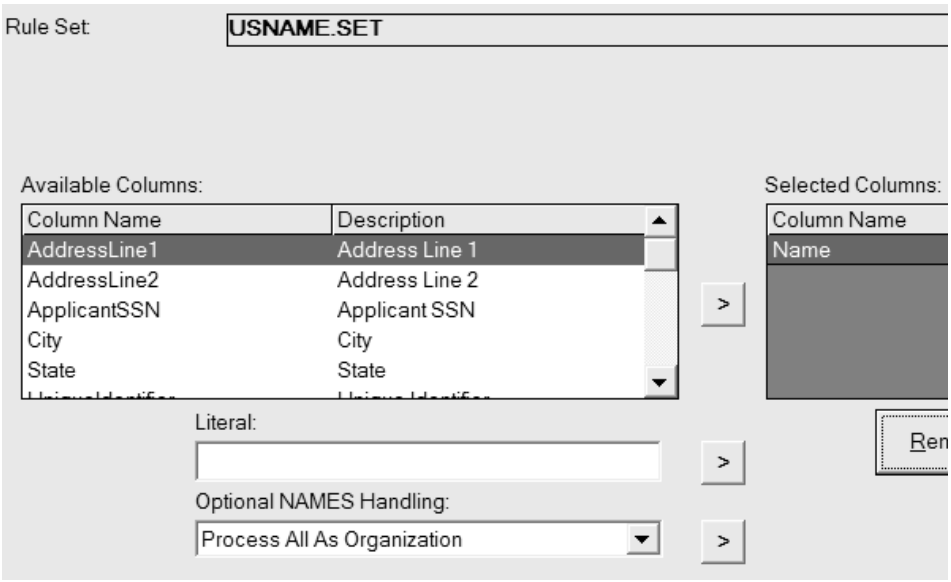
The source data is attached to the **Customer** source file and table definitions are loaded to organize the data into standard address columns.

Configuring the Standardize stage

The Standardize stage applies rules to name and address data to parse the data into a standard column format.

To configure the Standardize stage:

1. Double-click the Standardize stage icon to open the Standardize Stage window.
2. Click the **New Process** tab to open the Standardize Rule Process window.
3. In the **Rule Set** field, click **Standardization Rules** > **USA**. The rule sets in the **Standardization Rules** folder are domain specific for standardization jobs. You select rule sets from this folder to create consistent, industry-standard data structures and matching structures.
4. Open the **USNAME** folder.
   a. Select the **USNAME** rule set and click **OK**. USNAME.SET displays in the **Rule Set** field. You select this rule set because the name and address data is from the United States.
   b. In the **Available Columns** pane, select **Name**.
   c. Click  to move the **Name** column into the **Selected Columns** pane. The **Optional NAMES Handling** field is activated.
   d. Click **OK**.



5. Click the **New Process** tab to open the Standardize Rule Process window.
6. In the **Rule Set** field, click **Standardization Rules** > **USA** and select the **USADDR** rule set.
7. Select the following column names in the **Available Columns** pane and move them to the **Selected Columns** pane:
   - AddressLine1
   - AddressLine2
8. Click **OK**.
9. Click the **New Process** tab to open the Standardize Rule Process window.
10. In the **Rule Set** field, click **Standardization Rules** > **USA** and select the **USAREA** rule set.
11. Select the following column names in the **Available Columns** pane and move them to the **Selected Columns** pane:
    - City
    - State

- Zip5
- Zip4

Note: Maintain the order of the columns. Zip5 should precede Zip4.

12. Click **OK**.
13. Click the **New Process** tab to open the Standardize Rule Process window.
14. In the **Rule Set** field, click **Standardization Rules** > **USA** and select the **USTAXID** rule set.
15. Select the following column name in the **Available Columns** pane and move it to the **Selected Columns** pane:
    - ApplicantSSN
16. Click **OK**.
17. Map the Standardize stage output columns.
    a. Click the **Stage Properties** tab.
    b. Click the **Output** > **Mapping** tab.
    c. In the **Columns** pane, right-click and select **Select All** from the shortcut menu.
    d. Right-click and select **Copy** from the shortcut menu.
    e. Move to the **Standardized** pane, right-click and select **Paste Column** from the shortcut menu.
18. Save the table definitions to the **Table Definitions** folder.
    a. Click the **Columns** tab.
    b. Click **Save**. The Save Table Definition window opens.
    c. In the **Data source type** field, type `Table Definitions`.
    d. In the **Data source name** field, type `Standardized`.
    e. In the **Table/file name** field, type `Standardized`.
    f. Click **OK** to open the Save Table Definition As window.
    g. Save the Standardization table definitions in the Table Definition folder that is one level down from the project folder, for example, **QualityStage Tutorial** > **Table Definitions**.
    h. Confirm your changes and exit the windows.

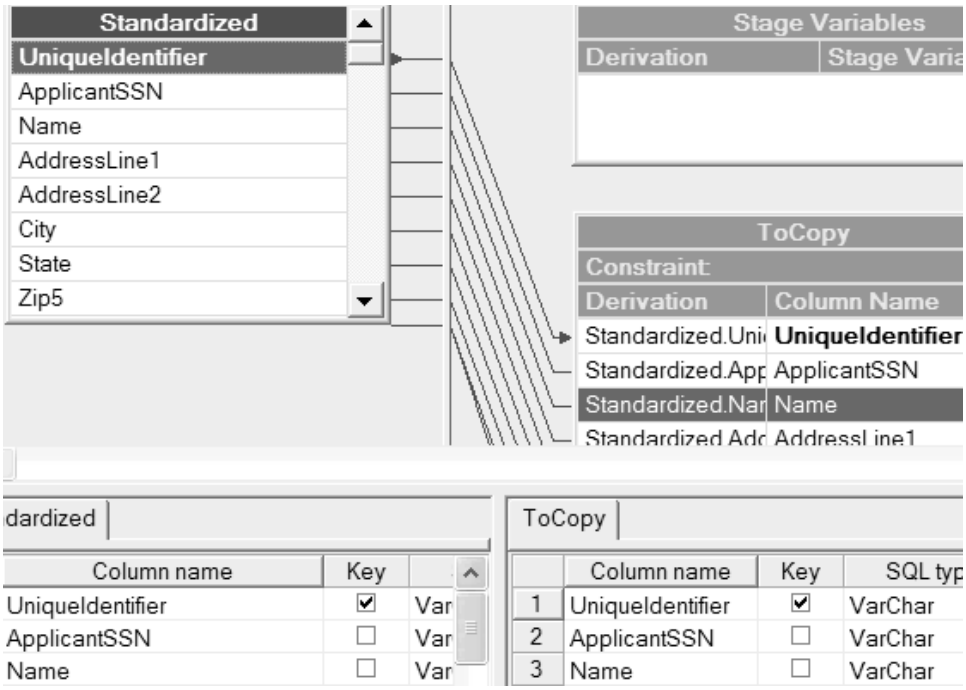
You configured the Standardize stage to apply the USNAME, USADDR, USAREA, and USTAXID rule sets to the customer data and saved the table definitions.

Configuring the Transformer stage

The Transformer stage increases the number of columns that the matching stage uses to select matches. The Transformer stage also applies derivations to handle null values.

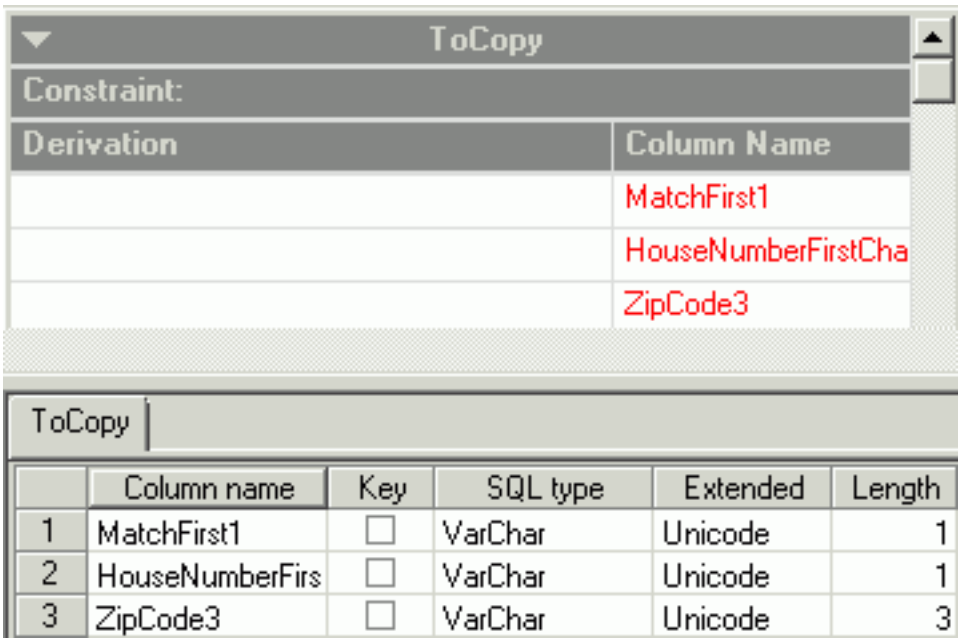
To configure the transformer properties:

1. Double-click the CreateAdditionalColumns stage icon to open the Transformer Stage window.
2. In the upper section of the window, right-click any column in the **Standardized** link box and select **Select All** from the shortcut menu. All of the columns in the **Standardized** link box are highlighted.
3. Right-click and select **Copy** from the shortcut menu.
4. Move to the **ToCopy** pane in the upper section of the window, right-click and select **Paste Column** from the shortcut menu. The mapping of input columns to specified derivations should look like the following figure:

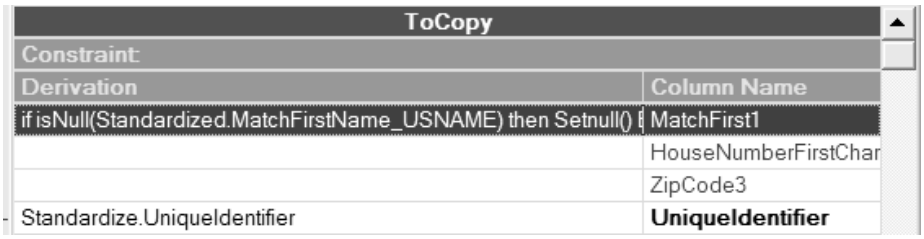


5. In the lower right section of the window, select the top row, row 1, of the **ToCopy** pane and add three derivations and columns to the CreateAdditionalColumns stage:
   a. Right-click the row and select **Insert row** from the shortcut menu.
   b. Add two more rows using the procedure explained in step 5a.
   c. Right-click the top inserted row and select **Edit row** from the shortcut menu to open the Edit Column Meta Data window.
   d. In the **Column name** field, type `MatchFirst1`.
   e. From the **SQL type** list, select **VarChar**.
   f. In the **Length** field, enter**1**.
   g. From the **Nullable** list, select **Yes**.
   h. Click **Apply**, then click **Close** to close the window.
   i. On the Parallel page in the lower section of the window, select **Extended (Unicode)**.
   j. Right-click the next row and select **Edit row** from the shortcut menu.
   k. In the **Column name** field, type `HouseNumberFirstChar`.
   l. Repeat substeps 5e to 5h.
   m. Right-click the last new row and select **Edit row** from the shortcut menu.
   n. In the **Column name** field, type `ZipCode3`.
   o. Repeat substeps 5e to 5h, except in the **Length** field, select **3**.

   The mapping of input columns to specified derivations should look like the following figure:



6. Add derivations to the columns:
   a. Double-click the cell that is in the **Derivation** column and in the same row as the **MatchFirst1** column, in the ToCopy window. Type the derivation: `if IsNull(Standardized.MatchFirstName_USNAME) then Setnull() Else Standardized.MatchFirstName_USNAME[1,1]`. This expression detects whether the MatchFirstName column contains a null. If the column contains null, it handles it. If the column contains a string, it extracts the first character and writes it to the MatchFirst1 column.



   b. Repeat substep a for the HouseNumberFirstChar column and type the derivative: `if IsNull(Standardized.HouseNumber_USADDR) then Setnull() Else Standardized.HouseNumber_USADDR[1,1]`.
   c. Repeat substep a for the ZipCode3 column and type the derivative: `if IsNull(Standardized.ZipCode_USAREA) then Setnull() Else Standardized.ZipCode_USAREA[1,3]`.
7. Map the three derivations and columns to the input columns.
   a. Move to the upper left pane and scroll the **Standardized** pane until you locate MatchFirstName_USNAME.
   b. Click and drag the cell into the **ToCopy** pane and into the cell that contains Standardized.MatchFirstName_USNAME.
   c. When prompted to override existing data, click **Yes**.
   d. Repeat substeps a through c for HouseNumber_USADDR and ZipCode_USAREA, matching the column names in the **Standardized** pane to the similarly named columns in the **ToCopy** pane.
   e. Click **OK** to close the Transformer Stage window.

Configuring the Copy stage

The Copy stage duplicates data and writes it to more than one output link. In this lesson, the Copy stage duplicates the metadata from the Transformer stage and writes it to the Match Frequency stage and the target file.

The metadata from the Standardize and Transformer stages is duplicated and written to two output links.

To configure the Copy stage:

1. Double-click the Copy stage icon to open the Copy Stage window.
2. Click the **Output** > **Mapping** tab.
3. Copy the data to the **StandardizedData** output link:
   a. In the **Output name** field above the **Columns** pane, select **StandardizedData**.
   b. Right-click in the **Columns** pane and select **Select All** from the shortcut menu.
   c. Right-click and select **Copy** from the shortcut menu.
   d. Move to the **StandardizedData** pane, right-click and select **Paste Column** from the shortcut menu.
4. To copy the data to the **ToMatchFrequency** output link, select **ToMatchFrequency** in the **Output name** field above the **Columns** pane and repeat steps b through d, pasting the data to the **ToMatchFrequency** pane.
5. Click **OK** to copy the data and close the Copy stage.

Configuring the Match Frequency stage

The Match Frequency stage generates frequency distribution information by analyzing data that is used to perform matching.

The Match Frequency stage processes frequency data independently from executing a match. The output link of this stage carries four columns:

- qsFreqVal
- qsFreqCounts
- qsFreqColumnID
- qsFreqHeaderFlag

To configure the Match Frequency stage:

1. Double-click the Match Frequency stage icon to open the Match Frequency Stage window.
2. Select the **Do not use a Match Specification** check box. At this point you do not know which columns are used in the match specification.
3. Click the **Stage Properties** tab.
4. Click the **Output** > **Mapping** tab.
   a. In the **Output name** field, select **ToFrequencies**.
   b. Right-click in the **Columns** pane and select **Select All** from the shortcut menu.
   c. Right-click and select **Copy** from the shortcut menu.
   d. Move to the **ToFrequencies** pane, right-click and select **Paste Column** from the shortcut menu.
5. Save the table definitions to the **Table Definitions** folder.
   a. Click the **Columns** tab.

b. Click **Save**. The Save Table Definition window opens.
c. In the **Data source type** field, type `Table Definitions`.
d. In the **Data source name** field, type `ToFrequencies`.
e. In the **Table/file name** field, type `ToFrequencies`.
f. Click **OK** to open the Save Table Definition As window.
g. Save the Standardization table definitions in the Table Definition folder that is one level down from the project folder, for example, **QualityStage Tutorial** > **Table Definitions**.
h. Click **Save**.
i. Confirm your changes and exit the windows.

6. Click **OK** to close the **Output** > **Column** tab and the Match Frequency stage.
7. Click **OK** to close the stage.

Lesson checkpoint

This lesson explained how to configure the source file and all the stages for the Standardize job.

You have now applied settings to each stage and mapped the output files to the next stage for the Standardize job. You learned how to do the following tasks:

- Configure the source file to load the customer data and metadata
- Apply United States postal service compliant rule sets to the customer name and address data
- Add additional columns for matching and create derivations to handle nulls
- Write data to two output links and associate the data to the correct links
- Create frequency data

Lesson 2.3: Configuring the target data sets

The two target data sets in the Standardize job store the standardized and frequency data that you can use as source data in the One-source Match job.

Complete the following tasks to configure the target data sets:

- Attach the file to the Stan target data set
- Attach the file to the Frequencies data set

To configure the target data sets:

1. Double-click the **Stan** target data set icon to open the Data Set window.
2. Click **Input** > **Properties** and select **Target** > **File**.
3. Click  and browse to the folder on the server computer where the input data file (for example, input.csv) resides.
4. In the **File name** field, type `Stan` and then click **OK** to display the path and file name in the **File** field, (for example, C:\IBM\InformationServer\Server\Projects\tutorial\Stan).
5. Save the table definitions to the **Table Definitions** folder.
   a. Click the **Columns** tab.
   b. Click **Save**. The Save Table Definition window opens.
   c. In the **Data source type** field, type `Table Definitions`.
   d. In the **Data source name** field, type `StandardizedData1`.
   e. In the **Table/file name** field, type `StandardizedData1`.

f. Click **OK** to open the Save Table Definition As window.
g. Save the table definitions in the Table Definition folder that is one level down from the project folder, for example, **QualityStage Tutorial** > **Table Definitions**.
h. Confirm your changes and exit the windows.

6. Double-click the **Frequencies** target data set icon.
7. Repeat steps 2 through 5 for the Frequencies file except replace `Stan` with `Frequencies` and replace `StandardizedData1` with `ToFrequencies1` in the appropriate fields. The Stan file and the Frequencies file are the source data sets for the One-source Match job.
8. Click **File** > **Save** to save the Standardize job.
9. Click  to compile the job in the Designer client.
10. Click  to run the job.

The job standardizes the data according to applied rules and adds additional matching columns to the metadata. The data is written to two target data sets as the source files for a later job.

Lesson checkpoint

This lesson explained how to attach files to the target data sets to store the processed standardized customer name and address data and frequency data.

You have configured the Stan and Frequencies target data set files to accept the data when it is processed.

Module 2: Summary

In Module 2, you set up and configured a Standardize job.

Running a Standardize job conditions the data to ensure that all the customer name and address data has the same content and format. The Standardize job loads the name and address source data stored in the database of the bank and adds table definitions to organize the data into a format that can be analyzed by the rule sets. Further processing by the Transformer stage increases the number of columns and frequency data is generated for input into the match job.

Lessons learned

By completing this module, you learned about the following concepts and tasks:

- How to create standardized data to match records effectively
- How to run IBM InfoSphere DataStage and Data Quality stages together in one job
- How to apply country or region-specific rule sets to analyze the address data
- How to use derivations to handle nulls
- How to create the data that can be used as source data in a later job

Module 3: Grouping records with common attributes

This module explains how to set up and run a One-source Match job using standardized data and generated frequency data to match records and remove duplicate records.

The One-source Match stage is one of two stages that matches records while removing duplicates and nonmatched records. The other matching stage is the Two-source Match stage.

The One-source Match stage groups records that share common attributes. The match specification that you apply was configured to separate all records with weights above a certain match cutoff as duplicates. The master record is then identified by selecting the record within the set that matches to itself with the highest weight.

Any records that are not part of a set of duplicates are nonmatched records. These records along with the master records are used for the next pass. Do not include duplicates because you want them to belong to only one set.

Using a matching stage ensures data integrity because you are applying probabilistic matching technology. This technology is applied to any relevant attribute for evaluating the columns, parts of columns, or individual characters that you define. In addition, you can apply agreement or disagreement weights to key data elements.

Learning objectives

After completing the lessons in this module, you should know how to do the following tasks:

- Add IBM InfoSphere DataStage links and stages to a job
- Use standardized data and frequency data as the source files
- Configure stage properties to specify which action the stages take when the job is run
- Remove duplicate addresses after the first pass
- Apply a match specification to determine how matches are selected
- Funnel the common attribute data to a separate target file

This module should take approximately 30 minutes to complete.

Lesson 3.1: Setting up a One-source Match job

Collecting records into groups with related attributes is the next step in data cleansing. In this lesson, you add the Data Quality One-source Match stage and a Funnel stage to match records and remove duplicates.

If you have not already done so, open the IBM InfoSphere DataStage and QualityStage Designer client.

As you learned in the previous module, you must add stages and links to the Designer canvas to create a One-source Match job. The Standardize job you just completed created a **Stan** data set and a **Frequencies** data set. The information from these data sets is used as the input data when you design the One-source Match job.

To set up a One-source Match job:

1. From the left pane of the Designer, go to the MyTutorial folder you created for this tutorial and double click on OneSource1 to open the job.
2. Drag the following icons to the Designer canvas from the palette.
   - **Data Quality** > **One-source Match** icon to the middle of the canvas.
   - **File** > **Data Set** icon to the top left of the **One-source Match** icon.
   - A second **File** > **Data Set** icon to the lower left of the **One-source Match** icon.
   - **Processing** > **Funnel** icon to the upper right of the **One-source Match** icon.
   - Three **File** > **Sequential File** icons, one to the right of the Funnel stage and the other two to the right of the One-source Match stage.
3. Right-click the top **Data Set** icon and drag to create a link from this data set to the One-source Match stage.

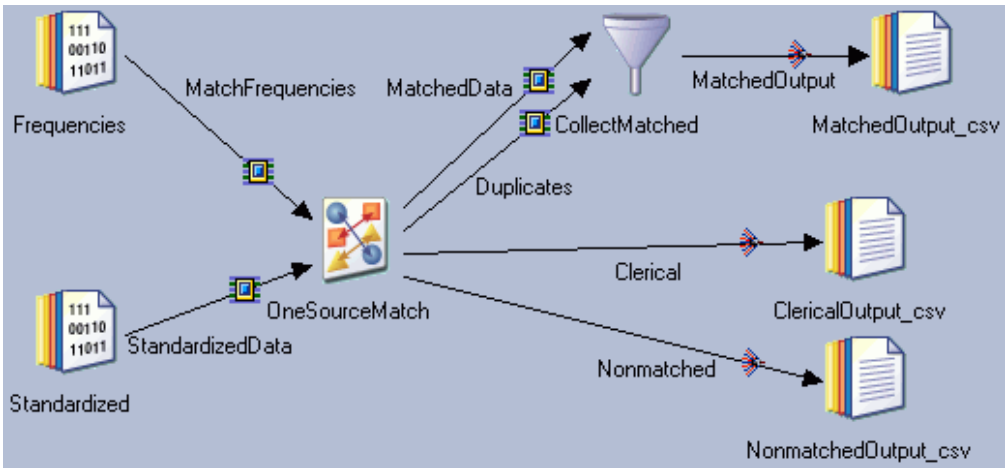
   Note: The order in which you create links affects the successful run of your job. Later in this tutorial, you will modify stage properties in order to change the order of some of the links.
4. Drag links to the remaining stages. Drag two links from the One-source Match stage to the Funnel stage.
5. Click on the names of the following stages and type the new stage name in the highlighted box:

Stage	Change to
top left Data Set	`Frequencies`
lower left Data Set	`Standardized`
One-source Match	`OneSourceMatch`
Funnel	`CollectMatched`
top right Sequential File	`MatchedOutput_csv`
middle right Sequential File	`ClericalOutput_csv`
lower right Sequential File	`NonMatchedOutput_csv`

6. Right-click on the names of the following links, select **Rename** from the shortcut menu and type the new link name in the highlighted box:

Links	Change to
From Frequencies to OneSourceMatch	`MatchFrequencies`
From Standardized to OneSourceMatch	`StandardizedData`
OneSourceMatch to CollectMatched	`MatchedData`
OneSourceMatch to CollectMatched	`Duplicates`
CollectMatched to MatchOutput_csv	`MatchedOutput`
OneSourceMatch to ClericalOutput_csv	`Clerical`
OneSourceMatch to NonMatchedOutput_csv	`NonMatched`

7. Click **File** > **Save** to save the job.



Lesson checkpoint for the One-source Match job

In this lesson, you learned how to set up an One-source Match job. During the processing of this job, the records are matched using the match specification created for this tutorial. The records are then sorted according to their attributes and written to a variety of output links.

You set up and linked an One-source Match job by doing the following tasks:

- Adding Data Quality and Processing stages to the Designer canvas
- Linking all the stages
- Renaming the links and stages with appropriate names

Lesson 3.2: Configuring the One-source Match job stage properties

Configure the properties for each stage of the One-source Match job on the Designer canvas.

Complete the following tasks to configure the One-source Match job:

- Load data and metadata for two source files
- Apply a match specification to the One-source Match job and select output links
- Combine unsorted records

To configure the Frequencies and Standardized data sets:

1. Double-click the **Frequencies** data set icon to open the **Properties** tab on the Frequencies - Data Set window.
2. Click **File** > **Source**.
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.
4. In the **File name** field, type `Frequencies`. (For example, C:\IBM\InformationServer\Server\Projects\tutorial\Frequencies).
5. Click **OK** to close the window.
6. Click the **Columns** tab and click **Load**. The Table Definitions window opens.
7. Select the *Project_folder* > **Table Definitions** > **ToFrequencies1** file and click **OK**.

8. Confirm your changes and exit the windows. The table definitions load into the **Columns** tab of the source file.
9. Double click the **Standardized** data set icon.
10. Repeat steps 2 to 9 except type `Stan` in step 4 and select the **StandardizedData1** file in step 7.

The data from the Standardize job is loaded into the source files for the One-source Match job.

Configuring the One-source Match stage

The One-source Match stage groups records with common attributes.

To configure the One-source Match stage:

1. Double-click the One-source Match stage icon.
2. Click the Match Specification  button.
3. From the Repository window, expand folders until you locate the **NameandAddress** folder.
4. Right-click on the **NameAndAddress** match specification and select **Provision All** from the shortcut menu.
5. Click **OK** to attach the **NameAndAddress** one-source match specification for the tutorial.
6. Click the check boxes for the following **Match Output** options:
   - Match - Sends matched records as output data.
   - Clerical - Separates those records that require clerical review.
   - Duplicate - Includes duplicate records that are above the match cutoff.
   - Nonmatched- Separates records that are not matched, duplicate, or clerical records as nonmatched records.
7. Keep the default **Dependent** setting in the **Match Type** panel. After the first pass is run, duplicates are removed with every additional pass.
8. Click the **Stage Properties** > **Link Ordering** tab. Make sure the input and output links are displayed in the following order. If necessary, use the up and down arrow buttons to move the links into the correct order.

Input Link label	**Link name**
Data	StandardizedData
DataFreq	MatchFrequencies

Output Link label	**Link name**
Match	MatchedData
Clerical	Clerical
Duplicate	Duplicates
Nonmatched	NonMatched

9. Click the **Output** > **Mapping** tab and map the following columns to the correct links:
   a. From the **Output name** list above the **Columns** pane, select **MatchedData** .
   b. Right-click in the **Columns** pane and select **Select All** from the shortcut menu.

c. Right-click and select **Copy** from the shortcut menu.
d. Move to the **MatchedData** pane, right-click and select **Paste Column** from the shortcut menu.
e. Select **Duplicates** from the **Output name** list above the **Columns** pane.
f. Repeat steps 9b on page 31 -9d for the Duplicates data.
g. Select **Clerical** from the **Output name** list above the **Columns** pane.
h. Repeat steps 9b on page 31 -9d for the Clerical data.
i. Select **NonMatched** from the **Output name** list above the **Columns** pane.
j. Repeat steps 9b on page 31 -9d for the Nonmatched data.

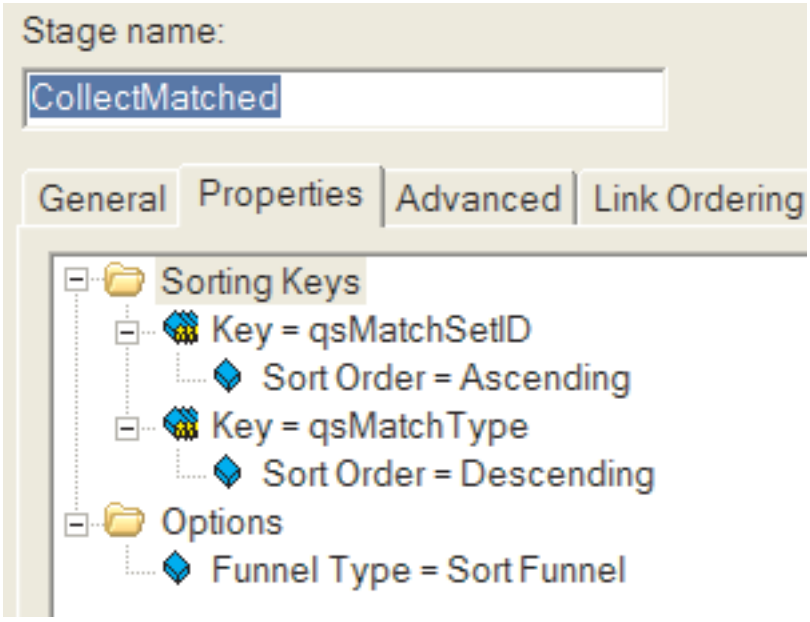
10. Click **OK** to close the Stage Properties window.
11. Click **OK** to close the stage.

Configuring the Funnel stage

The Funnel stage combines records when they are received in an unordered format.

To configure a funnel:

1. Double-click the CollectMatched stage icon and click the **Stage** > **Properties** tab.
2. In the **Options** tree, select **Funnel Type**.
3. From the **Funnel Type** list, select **Sort Funnel**.
4. Click **Sorting Keys** > **Key**, and then select **qsMatchSetID** from the **Key** list. The default sort order is **Ascending**.
5. Click **Sorting keys** again.
6. In the **Available properties to add** field, click **Key**.
7. From the **Key** list, select **qsMatchType**.
8. Click **Sort Order**, and then select **Descending** from the **Sort Order** list.



9. Click the **Output** > **Mapping** tab.
10. Right-click in the **Columns** pane and select **Select All** from the shortcut menu.
11. Right-click and select **Copy** from the shortcut menu.

12. Move to the **MatchedOutput** column, right-click and select **Paste Column** from the shortcut menu.
13. Click **OK** to close the stage window.

Lesson 3.2 checkpoint

In Lesson 3.2, you configured the source files and stages of the One-source Match job.

You learned how to do the following tasks:

- Load data and metadata generated in a previous job
- Apply a match specification to process the data and identify matches and duplicates
- Combine records into a single file

Lesson 3.3: Configuring One-source Match job target files

To configure the target files for the One-source Match job, you must attach files to the four output records. The records in the `MatchedOutput` file become the source records for the next job.

To configure the target files:

1. Double-click the **MatchedOutput_csv** icon to open the **Properties** tab in the MatchedOutput_csv - Sequential File window. You are attaching a file name to the matched records.
2. Click **Target** > **File**.
3. Next to the **File** field, click  and browse to the folder on the server computer where the input data file resides.
4. In the **File name** field, type `MatchedOutput.csv` to display the path and file name in the **File** field, (for example, C:\IBM\InformationServer\Server\Projects\tutorial\MatchedOutput.csv).
5. Click **Options** > **First Line is Column Names** and change the value to **True**.
6. Click the **Format** tab.
7. Right-click **Field Defaults**, and then click **Add sub-property** > **Null field value**.
8. Type `""`in the **Null field value** field. The null field value is a set of two double quotation marks with no space between them.
9. Specify how the stage collects data before the stage writes the data to the sequential file. You specify collection details because the previous 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 **qsMatchDataID**.
   e. Click **OK**.
10. Save the table definitions to the **Table Definitions** folder.
    a. Click the **Columns** tab.
    b. Click **Save**. The Save Table Definition window opens.
    c. In the **Data source type** field, type `Table Definitions`.

d. In the **Data source name** field, type `MatchedOutput1`.
e. In the **Table/file name** field, type `MatchedOutput1`.
f. Click **OK** to open the Save Table Definition As window.
g. Save the Standardization table definitions in the Table Definition folder that is one level down from the project folder, for example, **QualityStage Tutorial** > **Table Definitions**.
h. Confirm your changes and exit the windows.

11. Repeat steps 1 on page 33 - 10 on page 33 for each of the following stages:
    - For the ClericalOutput_csv stage, type `ClericalOutput.csv` and `Clerical1` in the appropriate fields.
    - For the NonMatchedOutput_csv stage, type `NonMatchedOutput.csv` and `NonMatched1` in the appropriate fields.
12. Click **File** > **Save** to save the job.
13. Click  to compile the job in the IBM InfoSphere DataStage and QualityStage Designer client.
14. Click **Tools** > **Run Director** to open the IBM InfoSphere DataStage and QualityStage Director Director. The Director opens with the One-source Match job visible in the Director window with the Compiled status.
15. Click  .

You configured the target files, and then you compiled and ran the job.

Lesson checkpoint

In this lesson, you combined the matched and duplicate address records into one file. The nonmatched and clerical output records were separated into individual files. The clerical output records can be reviewed manually for matching records. The nonmatched records are used in the next pass. The matched and duplicate address records are used in the Survive job.

You learned how to separate the output records from the One-source Match stage to the various target files.

Module 3: Summary

In Module 3, you set up and configured a job using the One-source Match stage to consolidate matched and duplicate name and address data into one file.

In creating a One-source Match stage job, you added a match specification to apply the blocking and matching criteria to the standardized and frequency data created in the Standardize job. After applying the match specification, the resulting records were sent out through four output links, one for each type of record. The matches and duplicates were sent to a Funnel stage that combined the records into one output, which was written to a file. The nonmatched records were sent to a file, as were the clerical output records.

Lessons learned

By completing Module 3, you learned about the following concepts and tasks:

- How to apply a match specification to the One-source Match stage
- How the One-source Match stage groups records with similar attributes

- How to ensure data integrity by applying probability matching technology

Module 4: Creating a single record

This module designs a Survive job to isolate the best record for the name and address of each customer.

The One-source Match job identifies a group of records with similar attributes. In the Survive job, you specify which columns and column values from each group creates the output record for the group. The output record can include the following information:

- An entire input record
- Selected columns from the record
- Selected columns from different records in the group

Select column values based on rules for testing the columns. A rule contains a set of conditions and a list of targets. If a column tests true against the conditions, the column value for that record becomes the best candidate for the target. After testing each record in the group, the columns declared best candidates combine to become the output record for the group. Column survival is determined by the target. Column value survival is determined by the rules.

Learning objectives

After completing the lessons in this module, you should know how to do the following tasks:

- Add stages and links to a Survive job
- Choose the selected column
- Add the rules
- Map the output columns

This module should take approximately 20 minutes to complete.

Lesson 4.1: Setting up a Survive job

Creating the best results record in the Survive stage is the last job in the data cleansing process. The best results record is the name and address with the highest probability of being correct for every bank customer.

In this lesson, add the Data Quality Survive stage, the source file of combined data from the One-source Match job, and the target file for the best records.

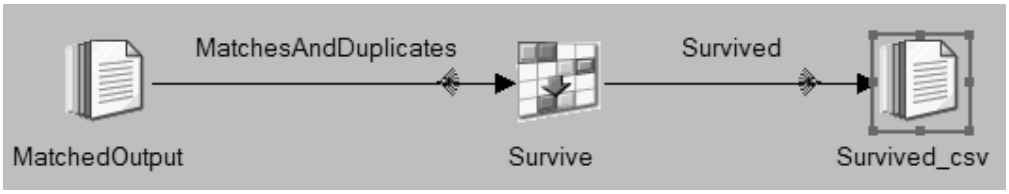
To set up a Survive job:

1. From the left pane of the IBM InfoSphere DataStage and QualityStage Designer client, go to the MyTutorial folder you created for this tutorial and double click on Survive1 to open the job.
2. Drag the following icons to the Designer canvas from the palette:
   - **Data Quality** > **Survive** icon to the middle of the canvas
   - **File** > **Sequential File** icon to the left of the Survive stage
   - Second **File** > **Sequential File** icon to the right of the Survive stage
3. Right-click the left **Sequential File** icon and drag a link to the Survive stage.
4. Drag a second link from the Survive stage to the output **Sequential File** icon.
5. Click on the names of the following stages and type the new stage name in the highlighted box:

Stage	Change to
left Sequential file	MatchedOutput
Survive stage	Survive
right Sequential file	Survived_csv

6. Right-click on the names of the following links, select **Rename** from the shortcut menu and type the new link name in the highlighted box:

Links	Change to
From MatchedOutput to Survive	MatchesAndDuplicates
From Survive to Survived_csv	Survived



Lesson checkpoint

In this lesson, you learned how to set up a Survive job by adding as source data the results of the One-source Match job, the Survive stage, and the target file as the output record for the group.

You have learned that the Survive stage takes one input link and one output link.

Lesson 4.2: Configuring Survive job stage properties

To configure Survive job stage properties, load matched and duplicates data from the One-source Match job, configure the Survive stage with rules that test columns to a set of conditions, and configure the target file.

In the Survive job, you are testing column values to determine which columns are the best candidates for that record. These columns are combined to become the output record for the group. In selecting a best candidate, you can specify that these column values be tested:

- Record creation data
- Data source
- Length of data in a column
- Frequency of data in a group

To configure the source file:

1. Double-click the **MatchedOutput file** icon to access the **Properties** page.
2. Click **File** > **Source**.
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.
4. In the **File name** field, type `MatchedOutput.csv` to display the path and file name in the **File** field, (for example, C:\IBM\InformationServer\Server\Projects\tutorial\MatchedOutput.csv).

5. Click **Options** > **First Line is Column Names** and change the value to **True**.
6. Click the **Format** tab.
7. Right-click **Field Defaults**, and then click **Add sub-property** > **Null field value**.
8. Type ""in the **Null field value** field. The null field value is a set of two double quotation marks with no space between them.
9. Click the **Columns** tab and click **Load**. The Table Definitions window opens.
10. Click the *Project_folder* > **Table Definitions** > **MatchedOutput1** file. The table definitions load into the **Columns** tab of the source file.
11. Confirm your changes and exit the windows.

You attached file the MatchedOutput.csv file and loaded the Table Definitions into the MatchedOutput file.

Configuring the Survive stage

Configure the Survive stage with rules to compare the columns against a best case.

To configure the Survive stage:

1. Double click the Survive stage icon.
2. Click **New Rule** to open the Survive Rules Definition window. The Survive stage requires a rule that contains one or more targets and a TRUE condition expression.

   You define rules by specifying the following elements:

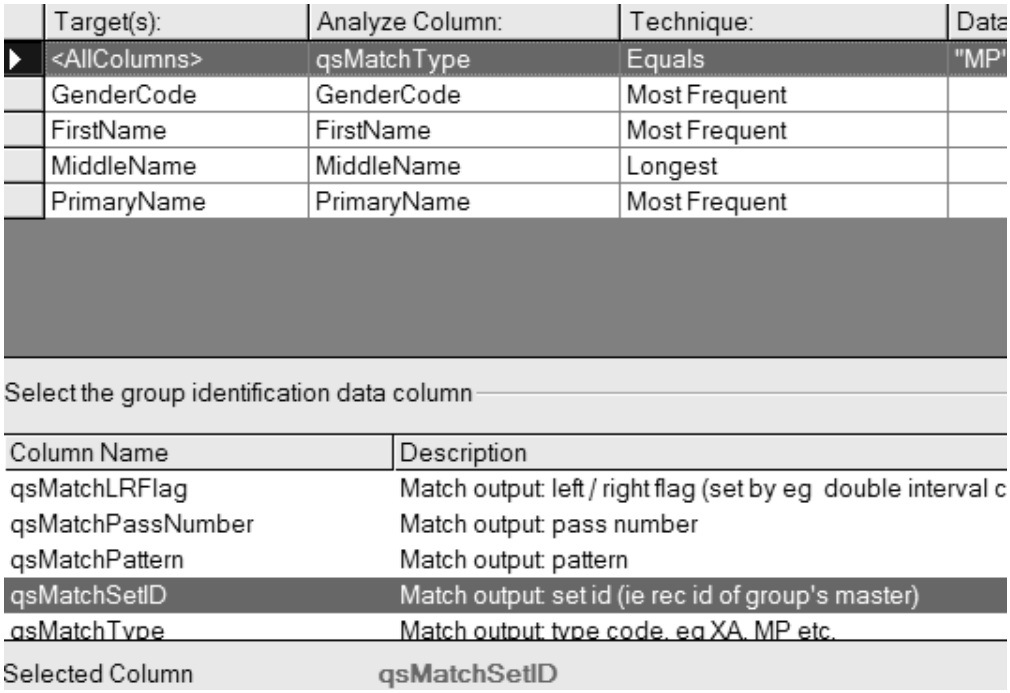
   - Target column or columns
   - Column to analyze
   - Technique to apply to the column being analyzed

3. In the **Available Columns** pane, select **AllColumns** and click  to move **AllColumns** to the **Target** pane. When you select AllColumns, you are assigning the first record in the group as the best record.
4. In the **Survive Rule** section of the window, select **qsMatchType** from the **Analyze Column** list. You are selecting qsMatchType as the target to which to compare other columns.
5. From the **Technique** list, select **Equals**.
6. In the **Data** field, type `MP`. MP signifies Match Pair for the One-source Match stage.
7. Click **OK** to close the Survive Rule Definition window.
8. Repeat steps 2 - 5 to add the following columns and rules. Do not enter values in the **Data** field.

Specify Output Columns > Target(s)	Analyze Column	Technique
GenderCode_USNAME	GenderCode_USNAME	Most Frequent (Non-blank)
FirstName_USNAME	FirstName_USNAME	Most Frequent (Non-blank)
MiddleName_USNAME	MiddleName_USNAME	Longest
PrimaryName_USNAME	PrimaryName_USNAME	Most Frequent (Non-blank)

   You can view the rules you added in the Survive grid.
9. In the **Select the group identification data column** section, select the qsMatchSetID column.

10. Click **Stage Properties**, and then click the**Output** > **Mapping** tab.
11. Right-click in the **Columns** pane and select **Select All** from the shortcut menu.
12. Select **Copy** from the shortcut menu.
13. Move to the **Survived** pane, right-click and select **Paste Column** from the shortcut menu.
14. Confirm your changes and exit the windows.



Configuring the target file

You are configuring the target file for the Survive stage.

1. Double click the **Survived_csv** target file icon and click **Target** > **File** to activate the **File** field.
2. In the **File** field, click  and browse to the folder on the server computer where the input data file resides.
3. In the **File name** field, type `record.csv` to display the path and file name in the **File** field, (for example, C:\IBM\InformationServer\Server\Projects\tutorial\record.csv).
4. Click **Options** > **First Line is Column Names** and change the value to **True**.
5. Click the **Format** tab.
6. Right-click **Field Defaults**, and then click **Add sub-property** > **Null field value**.
7. Type ""in the **Null field value** field. The null field value is a set of two double quotation marks with no space between them.
8. Confirm your changes and exit the windows.
9. Click **File** > **Save** to save the job.
10. Click  to compile the job in the Designer client.

11. Click **Tools** > **Run Director** to open the DataStage Director. The Director opens with the Standardize job visible in the Director window with the Compiled status.
12. Click **Run**.

Lesson checkpoint

You have set up the survive job, renamed the links and stages, and configured the source and target files, and the Survive stage.

With Lesson 4.2, you learned how to specify simple rules which are then applied to a selected column. This combination is then compared against all columns to find the best record.

Module 4: Summary

In Module 4, you completed the last job in the IBM InfoSphere QualityStage work flow. In this module, you set up and configured the Survive job to select the best record from the matched and duplicates name and address data that you created in the One-source Match stage.

In configuring the Survive stage, you selected a rule, included columns from the source file, added a rule to each column and applied the data. After the Survive stage processed the records to select the best record, the information was sent to the output file.

Lessons learned

In completing Module 4, you learned about the following concepts and topics:

- How to use the Survive stage to create the best candidate in a record
- How to apply simple rules to the column values

IBM InfoSphere QualityStage Tutorial: summary

From the lessons in this tutorial, you learn how InfoSphere QualityStage can be used to help an organization manage and maintain its data quality. It is imperative for companies that their customer data be high quality; thus it needs to be up-to-date, complete, accurate, and easy to use.

The tutorial presented a common business problem which was to verify customer names and addresses, and showed the steps to take by using InfoSphere QualityStage jobs to reconcile the various names that belonged to one household. The tutorial presented four modules that covered the four jobs in the InfoSphere QualityStage work flow. These jobs provide customers with the following assurances:

- Investigating data to identify errors and validate the contents of fields in a data file
- Conditioning data to ensure that the source data is internally consistent
- Matching data to identify all records in one file that correspond to similar records in another file
- Identifying which records from the match data survive to create a best candidate record

Lessons learned

By completing this tutorial, you learned about the following concepts and tasks:

- About the InfoSphere QualityStage work flow
- How to set up a InfoSphere QualityStage job
- How data created in one job is the source for the next job
- How to create quality data by using InfoSphere QualityStage

Appendix A. Product accessibility

You can get information about the accessibility status of IBM products.

The IBM InfoSphere Information Server product modules and user interfaces are not fully accessible.

For information about the accessibility status of IBM products, see the IBM product accessibility information at http://www.ibm.com/able/product_accessibility/index.html.

Accessible documentation

Accessible documentation for InfoSphere Information Server products is provided in an information center. The information center presents the documentation in XHTML 1.0 format, which is viewable in most web browsers. Because the information center uses XHTML, you can set display preferences in your browser. This also allows you to use screen readers and other assistive technologies to access the documentation.

The documentation that is in the information center is also provided in PDF files, which are not fully accessible.

IBM and accessibility

See the IBM Human Ability and Accessibility Center for more information about the commitment that IBM has to accessibility.

Appendix B. Contacting IBM

You can contact IBM for customer support, software services, product information, and general information. You also can provide feedback to IBM about products and documentation.

The following table lists resources for customer support, software services, training, and product and solutions information.

Table 1. IBM resources

Resource	Description and location
IBM Support Portal	You can customize support information by choosing the products and the topics that interest you at www.ibm.com/support/entry/portal/Software/Information_Management/InfoSphere_Information_Server
Software services	You can find information about software, IT, and business consulting services, on the solutions site at www.ibm.com/businesssolutions/
My IBM	You can manage links to IBM Web sites and information that meet your specific technical support needs by creating an account on the My IBM site at www.ibm.com/account/
Training and certification	You can learn about technical training and education services designed for individuals, companies, and public organizations to acquire, maintain, and optimize their IT skills at http://www.ibm.com/training
IBM representatives	You can contact an IBM representative to learn about solutions at www.ibm.com/connect/ibm/us/en/

Appendix C. Accessing the product documentation

Documentation is provided in a variety of formats: in the online IBM Knowledge Center, in an optional locally installed information center, and as PDF books. You can access the online or locally installed help directly from the product client interfaces.

IBM Knowledge Center is the best place to find the most up-to-date information for InfoSphere Information Server. IBM Knowledge Center contains help for most of the product interfaces, as well as complete documentation for all the product modules in the suite. You can open IBM Knowledge Center from the installed product or from a web browser.

Accessing IBM Knowledge Center

There are various ways to access the online documentation:

- Click the **Help** link in the upper right of the client interface.
- Press the F1 key. The F1 key typically opens the topic that describes the current context of the client interface.

  Note: The F1 key does not work in web clients.
- Type the address in a web browser, for example, when you are not logged in to the product.

  Enter the following address to access all versions of InfoSphere Information Server documentation:

  `http://www.ibm.com/support/knowledgecenter/SSZJPZ/`

  If you want to access a particular topic, specify the version number with the product identifier, the documentation plug-in name, and the topic path in the URL. For example, the URL for the 11.3 version of this topic is as follows. (The ⇒ symbol indicates a line continuation):

  ```
  http://www.ibm.com/support/knowledgecenter/SSZJPZ_11.3.0/⇒
  com.ibm.swg.im.iis.common.doc/common/accessingiidoc.html
  ```

  Tip:

  The knowledge center has a short URL as well:

  `http://ibm.biz/knowctr`

  To specify a short URL to a specific product page, version, or topic, use a hash character (#) between the short URL and the product identifier. For example, the short URL to all the InfoSphere Information Server documentation is the following URL:

  `http://ibm.biz/knowctr#SSZJPZ/`

  And, the short URL to the topic above to create a slightly shorter URL is the following URL (The ⇒ symbol indicates a line continuation):

  ```
  http://ibm.biz/knowctr#SSZJPZ_11.3.0/com.ibm.swg.im.iis.common.doc/⇒
  common/accessingiidoc.html
  ```

Changing help links to refer to locally installed documentation

IBM Knowledge Center contains the most up-to-date version of the documentation. However, you can install a local version of the documentation as an information center and configure your help links to point to it. A local information center is useful if your enterprise does not provide access to the internet.

Use the installation instructions that come with the information center installation package to install it on the computer of your choice. After you install and start the information center, you can use the **iisAdmin** command on the services tier computer to change the documentation location that the product F1 and help links refer to. (The ⇒ symbol indicates a line continuation):

Windows

```
IS_install_path\ASBServer\bin\iisAdmin.bat -set -key ⇒
com.ibm.iis.infocenter.url -value http://<host>:<port>/help/topic/
```

AIX® Linux

```
IS_install_path/ASBServer/bin/iisAdmin.sh -set -key ⇒
com.ibm.iis.infocenter.url -value http://<host>:<port>/help/topic/
```

Where <host> is the name of the computer where the information center is installed and <port> is the port number for the information center. The default port number is 8888. For example, on a computer named server1.example.com that uses the default port, the URL value would be http://server1.example.com:8888/help/topic/.

Obtaining PDF and hardcopy documentation

- The PDF file books are available online and can be accessed from this support document: https://www.ibm.com/support/docview.wss?uid=swg27008803&wv=1.
- You can also order IBM publications in hardcopy format online or through your local IBM representative. To order publications online, go to the IBM Publications Center at http://www.ibm.com/e-business/linkweb/publications/servlet/pbi.wss.

Notices and trademarks

This information was developed for products and services offered in the U.S.A. This material may be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

Notices

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785 U.S.A.

For license inquiries regarding double-byte character set (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law:
INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Corporation
J46A/G4
555 Bailey Avenue
San Jose, CA 95141-1003 U.S.A.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

All statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information is for planning purposes only. The information herein is subject to change before the products described become available.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Each copy or any portion of these sample programs or any derivative work, must include a copyright notice as follows:

© (your company name) (year). Portions of this code are derived from IBM Corp. Sample Programs. © Copyright IBM Corp. _enter the year or years_. All rights reserved.

If you are viewing this information softcopy, the photographs and color illustrations may not appear.

Privacy policy considerations

IBM Software products, including software as a service solutions, (“Software Offerings”) may use cookies or other technologies to collect product usage information, to help improve the end user experience, to tailor interactions with the end user or for other purposes. In many cases no personally identifiable information is collected by the Software Offerings. Some of our Software Offerings can help enable you to collect personally identifiable information. If this Software Offering uses cookies to collect personally identifiable information, specific information about this offering’s use of cookies is set forth below.

Depending upon the configurations deployed, this Software Offering may use session or persistent cookies. If a product or component is not listed, that product or component does not use cookies.

Table 2. Use of cookies by InfoSphere Information Server products and components

Product module	Component or feature	Type of cookie that is used	Collect this data	Purpose of data	Disabling the cookies
Any (part of InfoSphere Information Server installation)	InfoSphere Information Server web console	• Session • Persistent	User name	• Session management • Authentication	Cannot be disabled
Any (part of InfoSphere Information Server installation)	InfoSphere Metadata Asset Manager	• Session • Persistent	No personally identifiable information	• Session management • Authentication • Enhanced user usability • Single sign-on configuration	Cannot be disabled

Table 2. Use of cookies by InfoSphere Information Server products and components (continued)

Product module	Component or feature	Type of cookie that is used	Collect this data	Purpose of data	Disabling the cookies
InfoSphere DataStage	Big Data File stage	• Session • Persistent	• User name • Digital signature • Session ID	• Session management • Authentication • Single sign-on configuration	Cannot be disabled
InfoSphere DataStage	XML stage	Session	Internal identifiers	• Session management • Authentication	Cannot be disabled
InfoSphere DataStage	IBM InfoSphere DataStage and QualityStage Operations Console	Session	No personally identifiable information	• Session management • Authentication	Cannot be disabled
InfoSphere Data Click	InfoSphere Information Server web console	• Session • Persistent	User name	• Session management • Authentication	Cannot be disabled
InfoSphere Data Quality Console		Session	No personally identifiable information	• Session management • Authentication • Single sign-on configuration	Cannot be disabled
InfoSphere QualityStage Standardization Rules Designer	InfoSphere Information Server web console	• Session • Persistent	User name	• Session management • Authentication	Cannot be disabled
InfoSphere Information Governance Catalog		• Session • Persistent	• User name • Internal identifiers • State of the tree	• Session management • Authentication • Single sign-on configuration	Cannot be disabled
InfoSphere Information Analyzer	Data Rules stage in the InfoSphere DataStage and QualityStage Designer client	Session	Session ID	Session management	Cannot be disabled

If the configurations deployed for this Software Offering provide you as customer the ability to collect personally identifiable information from end users via cookies and other technologies, you should seek your own legal advice about any laws applicable to such data collection, including any requirements for notice and consent.

For more information about the use of various technologies, including cookies, for these purposes, see IBM's Privacy Policy at http://www.ibm.com/privacy and IBM's Online Privacy Statement at http://www.ibm.com/privacy/details the section entitled "Cookies, Web Beacons and Other Technologies" and the "IBM Software Products and Software-as-a-Service Privacy Statement" at http://www.ibm.com/software/info/product-privacy.

Trademarks

IBM, the IBM logo, and ibm.com® are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at www.ibm.com/legal/copytrade.shtml.

The following terms are trademarks or registered trademarks of other companies:

Adobe is a registered trademark of Adobe Systems Incorporated in the United States, and/or other countries.

Intel and Itanium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows and Windows NT are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java™ and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

The United States Postal Service owns the following trademarks: CASS, CASS Certified, DPV, LACS^{Link}, ZIP, ZIP + 4, ZIP Code, Post Office, Postal Service, USPS and United States Postal Service. IBM Corporation is a non-exclusive DPV and LACS^{Link} licensee of the United States Postal Service.

Other company, product or service names may be trademarks or service marks of others.

Index

Printed in USA

SC19-4327-00