Getting Started with EDQ

Release 12.2.1

Hands-On Lab

Version 1.0

Authors:

- Yash Patel (yash.patel@oracle.com)
- David Hecksel (david.hecksel@oracle.com)

Table of Contents

Introduction

Enterprise Data Quality is Oracle's premiere solution for delivering 'Data Fit for Use' and 'Accurate' Analytics that satisfy an organization's current and emerging Data Governance requirements for your Data Warehouse / Data Mart / Data Lake environments/initiatives. Oracle Enterprise Data Quality provides organizations with an integrated suite of data quality tools that provide an end-to-end solution to measure, improve and manage the quality of data from any domain, including Customer / Citizen / Student / Employee / Patient and others. Oracle Enterprise Data Quality also combines powerful data profiling, cleansing, matching and monitoring capabilities while offering unparalleled ease of use. Features of Oracle Enterprise Data Quality include:

- Advanced data profiling to identify and measure poor quality data and identify rule requirements to resolve your Project and/or Enterprise data quality issues
- Semantic and pattern-based recognition to accurately parse and standardize data that is poorly structured
- An innovative Open Reference Data Architecture enabling easy creation, customization and maintenance of business rules that adapt and learn from your data to enable and expedite automated solutions to jumpstart, continually improve and socialize your data quality over time
- "Easy Button" integration with Oracle Data Integrator Enterprise Edition

Key Benefits of adopting Enterprise Data Quality

- Data Quality Firewall: Provide automated processes for detecting, fixing and/or establishing a "return to sender" process on inbound data from other systems or partners prior to ingestion into your critical Data Warehouse and/or DataMart assets
- Cost Efficiencies: Eliminate costly manual Data "break-fix" and other one off workarounds by DBA / Business Owner staff
- Reporting and Compliance: Provides a consistent and collaborative platform for data quality governance (including DQ Firewall metrics and DQ over time) by enforcing data standards and enabling shared understanding of DQ rules by all project participants across systems and processes

Oracle Enterprise Data Quality Profiling for Data Integration

Oracle Enterprise Data Quality Profiling for Data Integration provides a basis for understanding data quality issues and a foundation for building data quality rules for defect remediation and prevention. EDQ enables users to understand their data by discovering, highlighting and communicating data anomalies within the data being profiled. Create, communicate, investigate, collaborate, and close data quality Incidents. EDQ Data Profiling provides a Data Quality Firewall for your Data Warehouse, Mart or other key data assets. Why put just 'any data' in your critical to business Data Warehouse and Marts? Examples include: University students ranging in age from -47 to 114 years old; students currently enrolled in the University system but at campus' that were closed the prior year; 27% of patients within the past 10 years mysteriously born on Jan 1, 1970. All of that is valid data from a database standpoint – it inserted and updated just fine – but is it really 'Data Fit for Use'? Would Analytics done on that

data be 'Accurate Analytics'? Profiling users can easily collaborate with others in the organization by opening and assigning Incidents to Business Owners, DBAs, Data Scientists and others within the EDQ user interface while exploring and discovering anomalies.

Oracle Enterprise Data Quality Audit and Dashboard for Data Integration

Oracle Enterprise Data Quality Audit and Dashboard for Data Integration provides:

- A variety of flexible and extensible Audit processors for validating the format, content and adherence of the data at hand to business rules. Records with non-conforming column content can easily be separated out for separate error path processing / correction and/or notification. Audit processors are merely added to Data Quality workflows and visually configured – no coding required.

- Pre-built dashboards and dashboard templates for reporting on DQ metrics in easy to understand and configurable executive dashboards viewed in your favorite web browser. DQ teams, executives, source data providers, ETL teams, Data Governance Centers of Excellence team members and consumers of the resulting 'Data Fit for Use' can view both point in time and rolling period comparison dashboards. Finally, dashboards complement Incident workflow team communication as the extended DQ team, sponsors and interested parties can review the latest published Dashboard results as projects progress through the Development, QA and Pre-Production SDLC phases.

Oracle Enterprise Data Quality Batch Processing for Data Integration

Oracle Enterprise Data Quality Batch Processing for Data Integration provides:

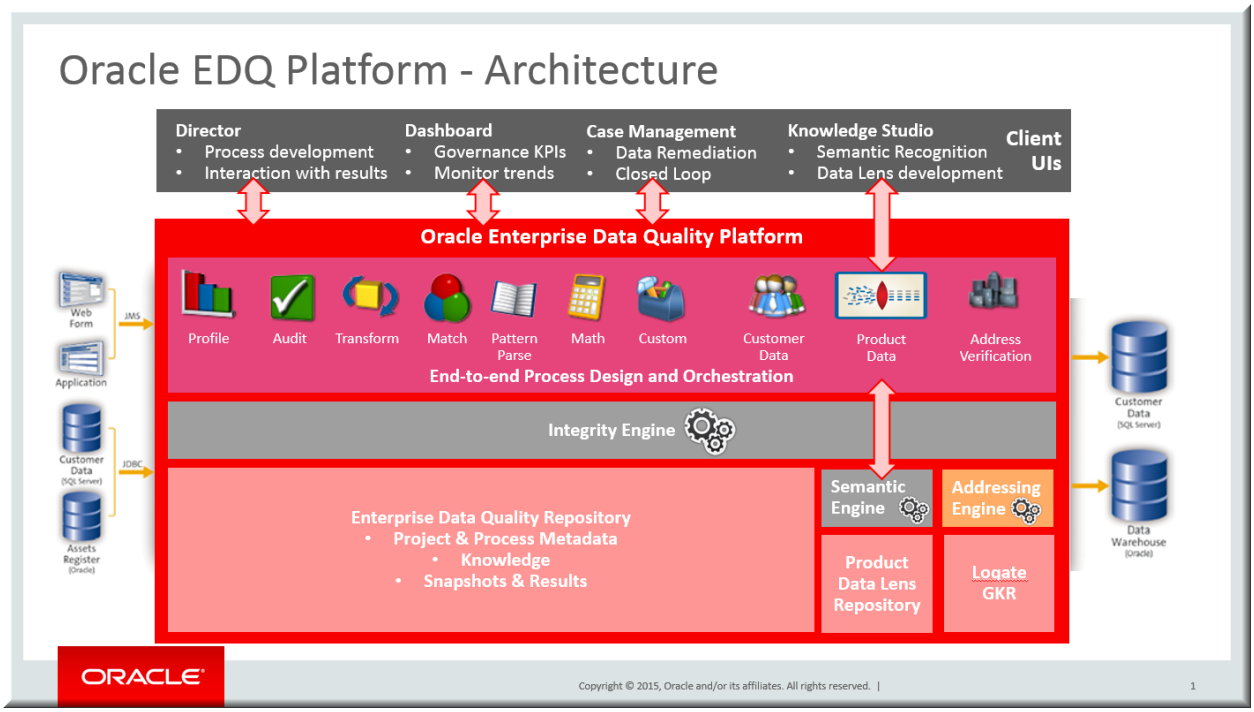
- A variety of flexible and extensible Transformation processors for standardizing and enriching source data into data that conforms to business rules and project requirements. For example, one set of Customer data had 18 different representations / spelling of the city 'Toronto'. Toronto, toronto, TORONTO, TOR, tor, torontto, … Leveraging EDQ's extensive set of out of the box Reference Data and EDQ's 'secret sauce' Open Reference Data Architecture, the Customer used a single Lookup and Return processor and pointed it to the Toronto reference data file and all invalid representations were transformed to 'Toronto' – simplifying the Business Intelligence developer's job and adding to the integrity of the data.

- A rich and configurable palette of visual Processors for adding match, merge and de-duplication record processing into DQ workflows

Now that data has been validated, transformed, enriched and standardized, optional match / merge / de-duplication processing can be done. Lab 4 focuses on adding match, merge and de-duplication functionality to your Hands-on Lab process workflows.

Oracle Enterprise Data Quality Component Architecture

Oracle Enterprise Data Quality (OEDQ) is a Java Web Application with the benefit of having a rich client for creating DQ Process workflows, creating/viewing Dashboards, data profiling, audit, advanced parsing, standardizing, performing Match Review approval processing and product Administration (among others).



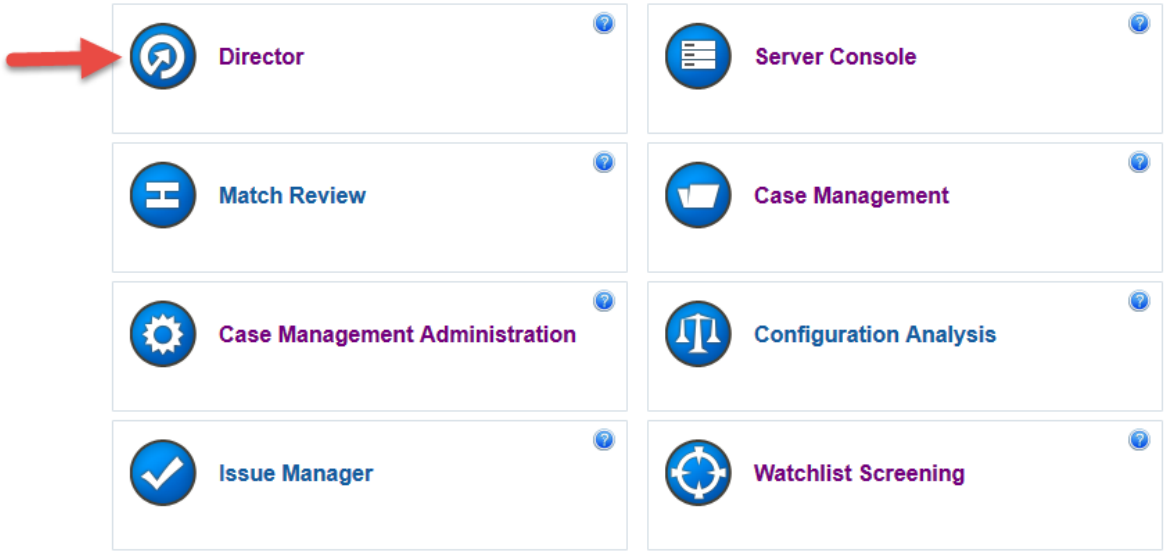
Lab 1a: Enterprise Data Quality Director

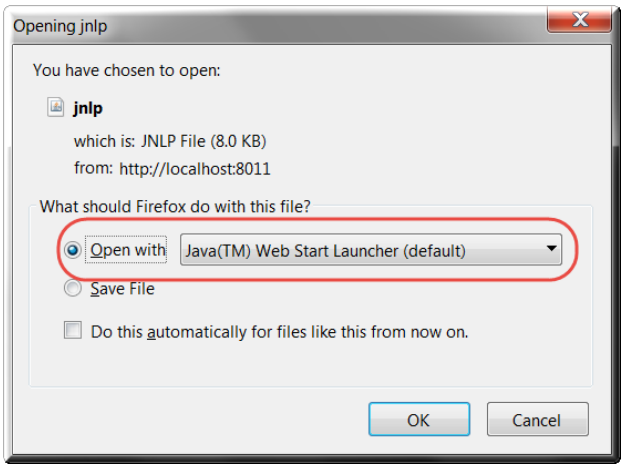
Navigate to the Enterprise Data Quality Launchpad and start Director

1. Launch an internet browser (Mozilla Firefox, Internet Explorer, Google Chrome) using your host operating system and navigate to URL http://localhost:8011/edq

> *Note that your host operating system will require an up to date Java 7 runtime environment (JRE 1.7.0_55+) installed. Alternatively, you can complete this workshop by opening the Mozilla Firefox browser within the virtual machine and navigating to a slightly different URL: http://localhost:8001/edq*

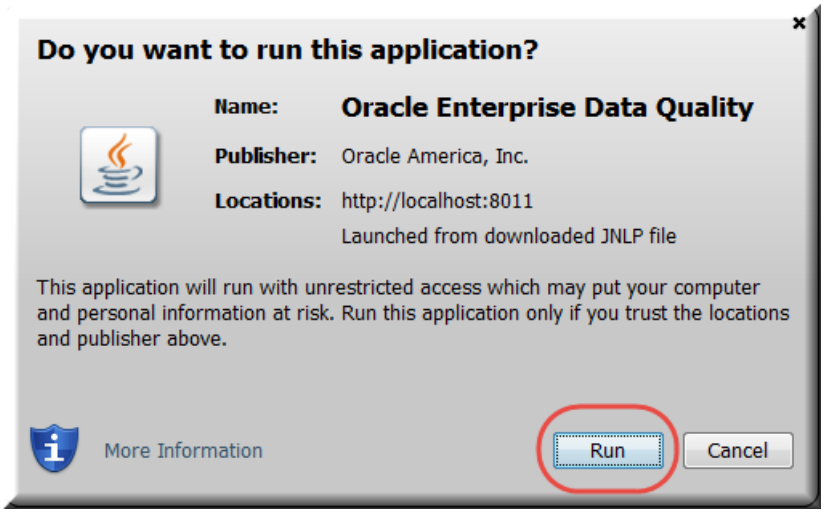
2. Click on the Director icon to launch the Director user interface





3. You may be prompted to open a file. Choose `open with` and select the `Java Web Start` application and click **OK**

*This may take a few moments to download the web application. You may be prompted to allow the application to run, click **Run***

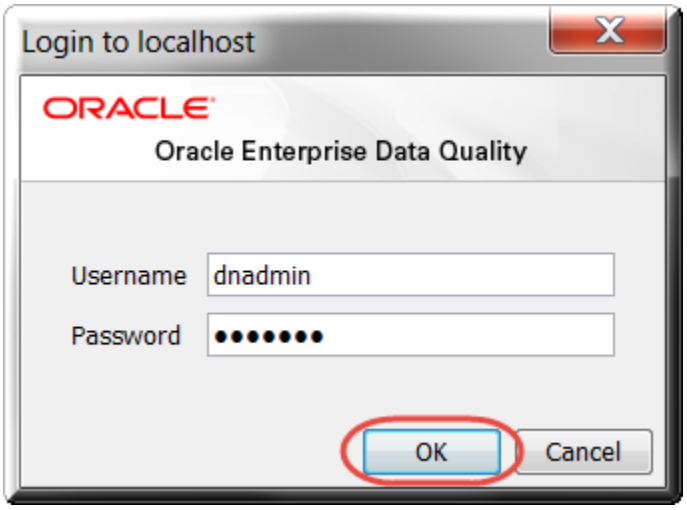


4. You will be prompted to login to Enterprise Data Quality. Use the following credentials:

   User Name: *dnadmin*

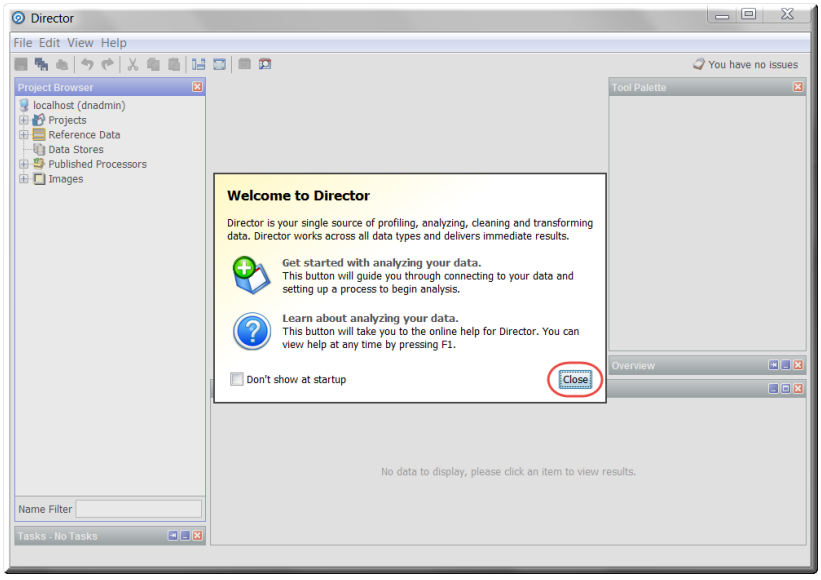
   Password: *dnadmin*

   Then click **OK**

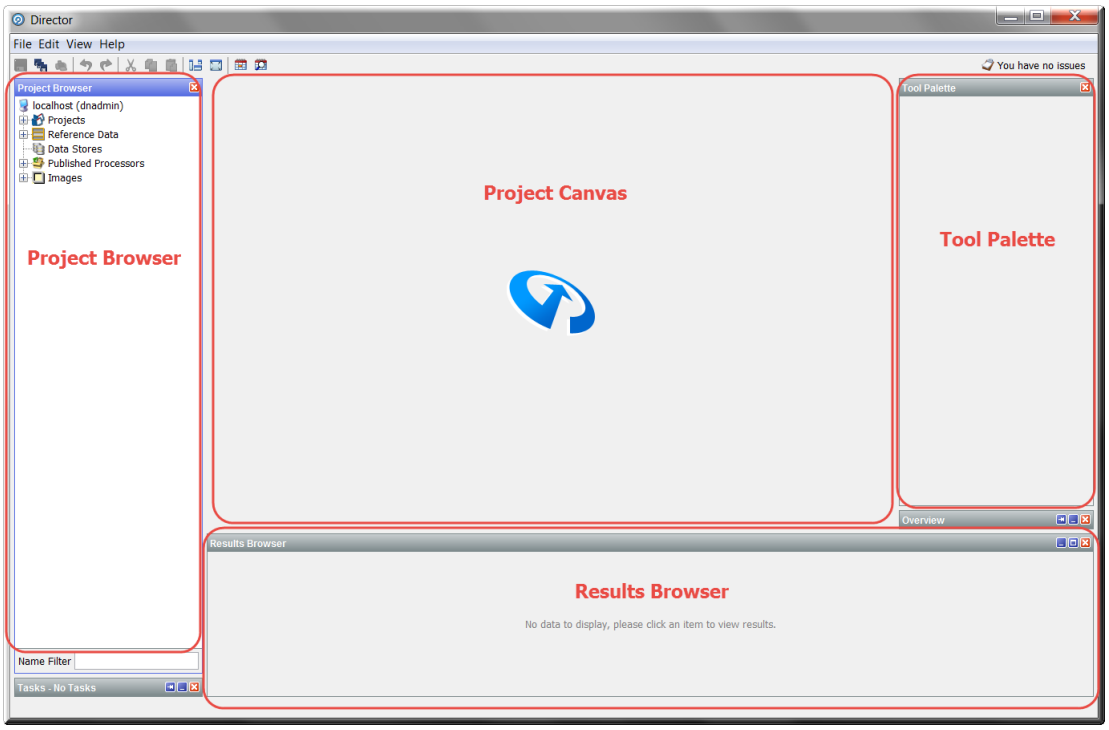


Explore the Director User Interface

5. After Director launches, click on **close** in the **Welcome to Director** message:



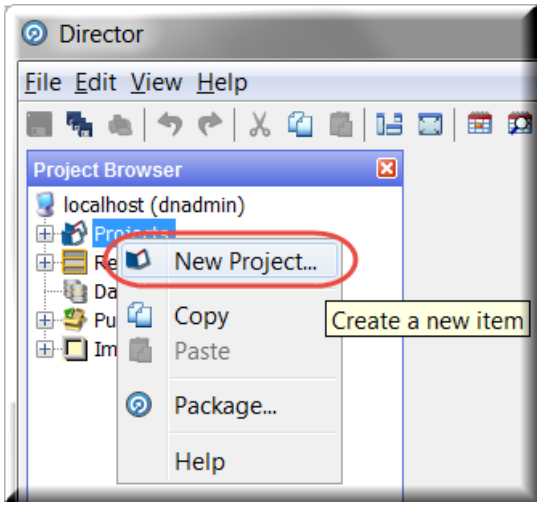
6. Take a moment to familiarize yourself with the Terminology of each of the four different areas of the **Director** application



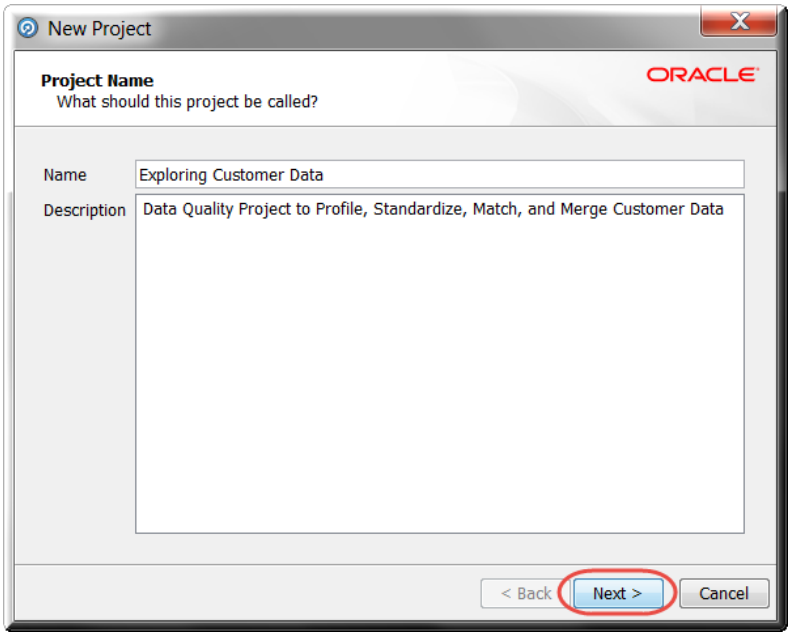
Create a New Project

Projects are created in the Project Brower and are generally utilized to hold data and processes related to a Data Quality initiative. You can set permissions and access levels at a project level. We will now begin a data quality project utilizing sample customer data from a US based order management system throughout the rest of the lab.

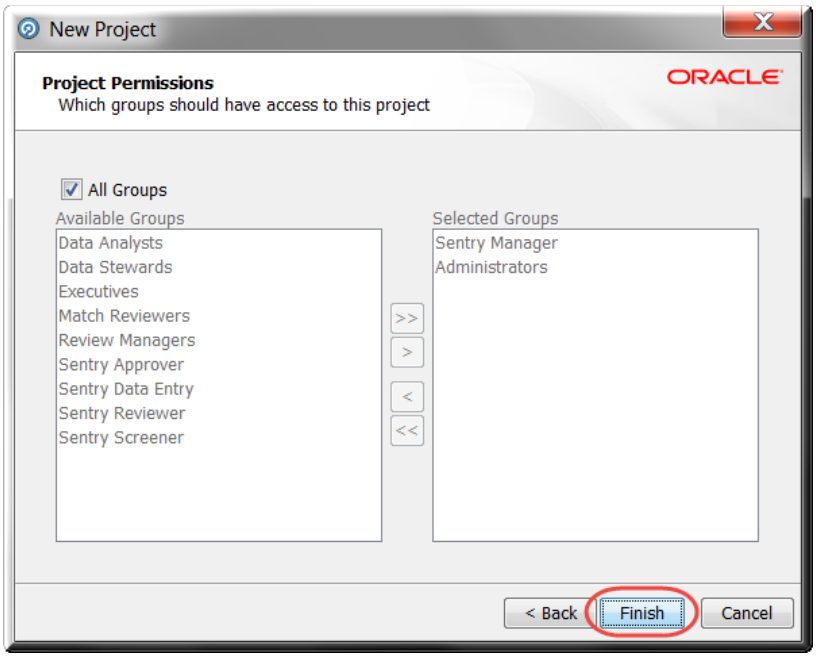
1. In the **Project Browser**, right-click **Projects** and select **New Project...** to start the Wizard



2. For **Name**, enter *`Exploring Customer Data`* and optionally add a **Description**. Click **Next >** to continue



3. Ensure the **All Groups** checkbox is selected in **Project Permissions**. This will ensure any user can view and use the project. Click **Finish** to create the new project



The **Exploring Customer Data** Project now appears in the **Projects** list!

Add a Data Store

Now that we have created a Project for the Labs, the next step is enabling access to your Data that needs Profiling / Enrichment / Standardizing and optional Match / Merge / De-duplication. Turning your Data into:

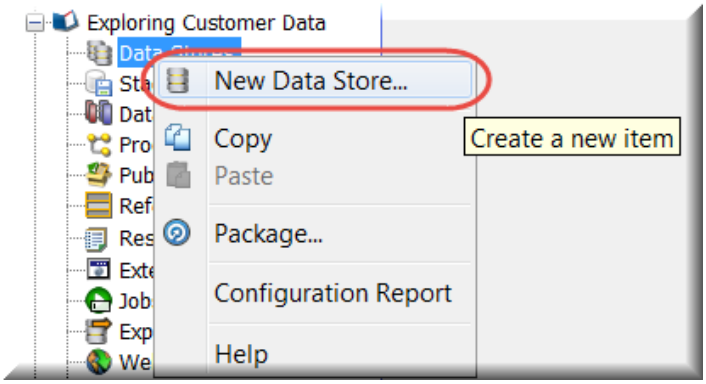
- 'not just Data', but 'Data Fit for Use'
- 'not just Analytics', but 'Accurate Analytics'

A Data Store is a connection to a store of data, whether the data is stored in a database or in one or more files. The data store may be used as the source of data for a process, or you may export written Staged Data results of a process to a data store, or both.

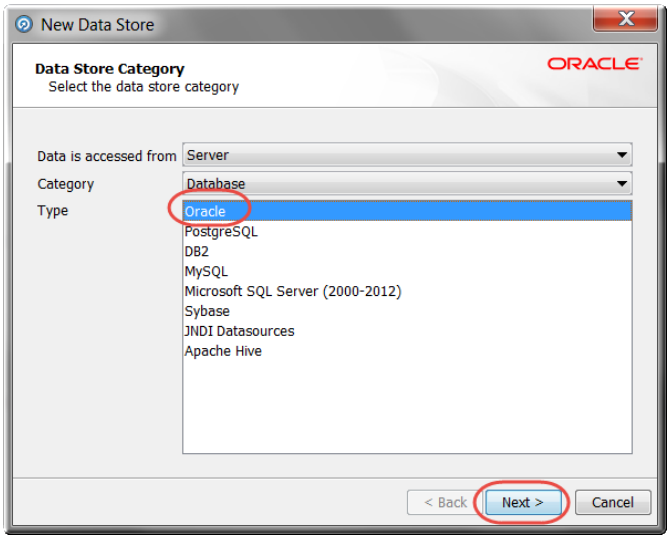
> *It is normally recommended to connect to the data store via the server. When connecting to files, this means that the files must exist in the server landing area to ensure that the server will be able to access them. However, it is also possible to pull the data onto the server using a client connection.*

EDQ supports native connections to many types of type of data stores. Let's begin by adding a data store to connect to an Oracle Database:

1. Expand the newly created project **Exploring Customer Data**, right click **Data Stores**, and select **New Data Store** to launch the **New Data Store** Wizard

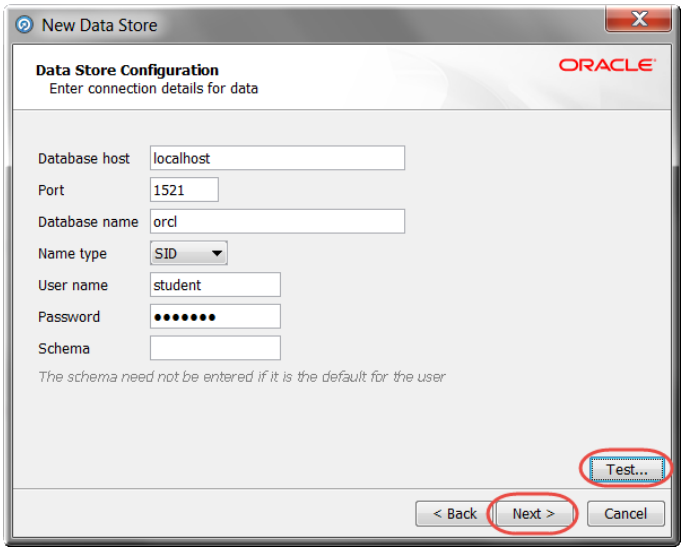


2. Select **Server** and **Database** in the two dropdown boxes, then select **Oracle**. Click **Next >** to continue

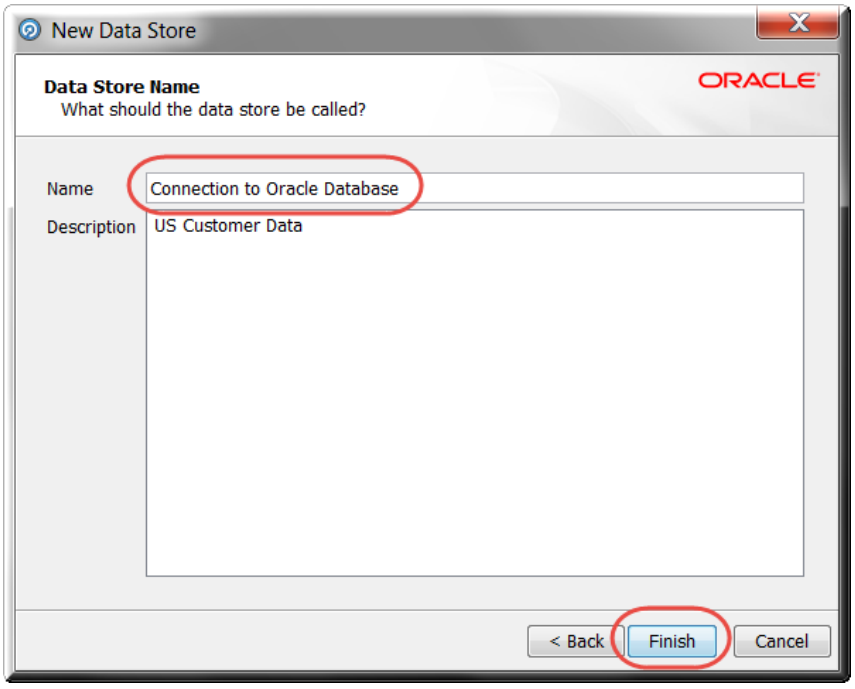


3. Enter the following connection details:
   - **Database Host**: *localhost*
   - **Port**: *1521*
   - **Database name** (SID): *orcl*
   - **User Name**: *student*
   - **Password**: *student*

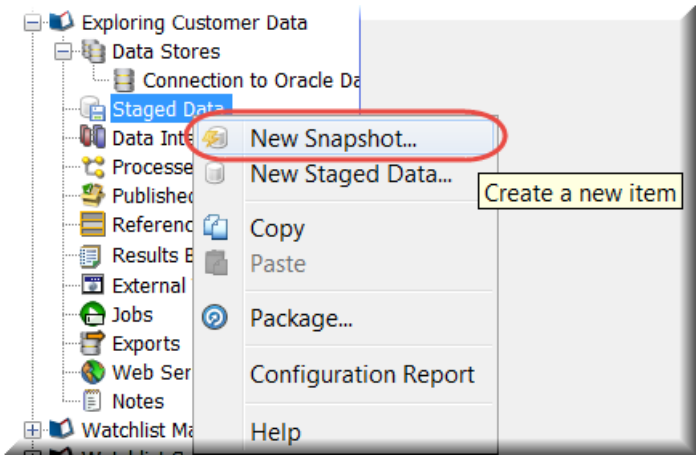
   Click **Test** to check whether Director can access the Server Database, then click **Next >** to continue



4. Now it is time to name our Data Store. For the **Name** text entry field, enter `Connection to Oracle Database`. You may optionally enter text in the Description text area to assist with communication and collaboration on the intended data source in the Description, click **Finish**

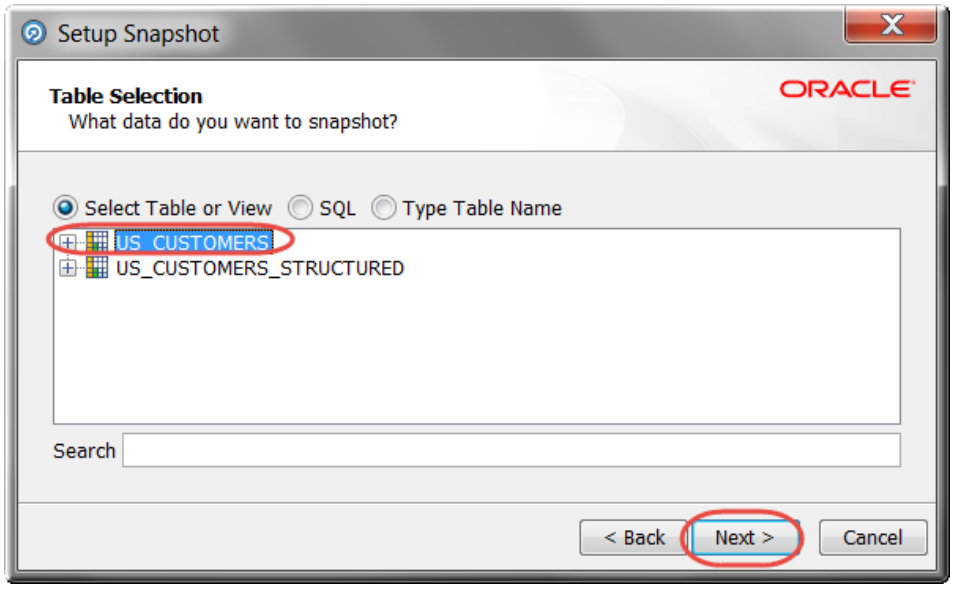


5. Navigate back to the **Project Browser** and right click **Staged Data** under your **Exploring Customer Data** project and select **New Snapshot...**

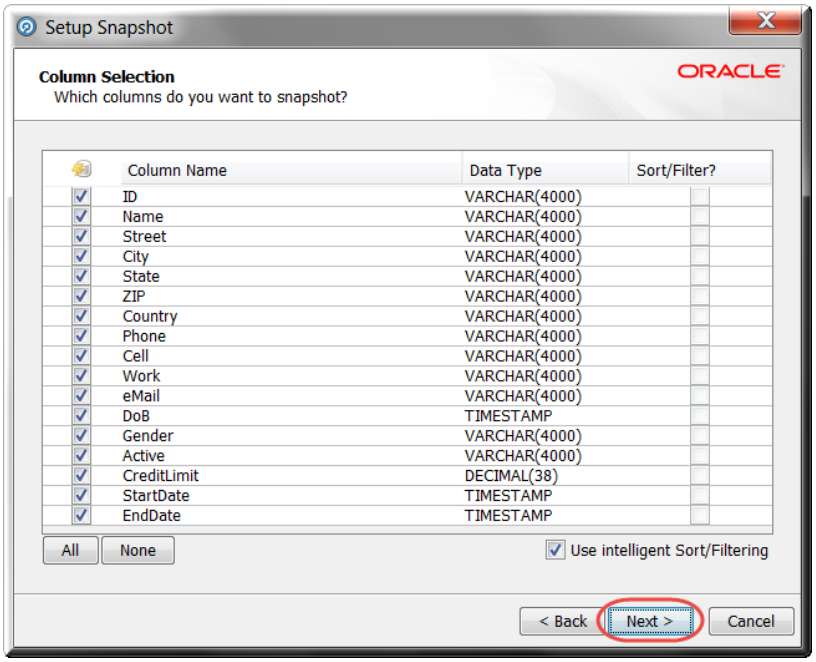


6. Select the newly created database connection – **Connection to Oracle Database**, then click **Next >** to continue. This is where the data for the snapshot will come from. For **Table Selection**, only one table is presented in the list and is auto selected. The

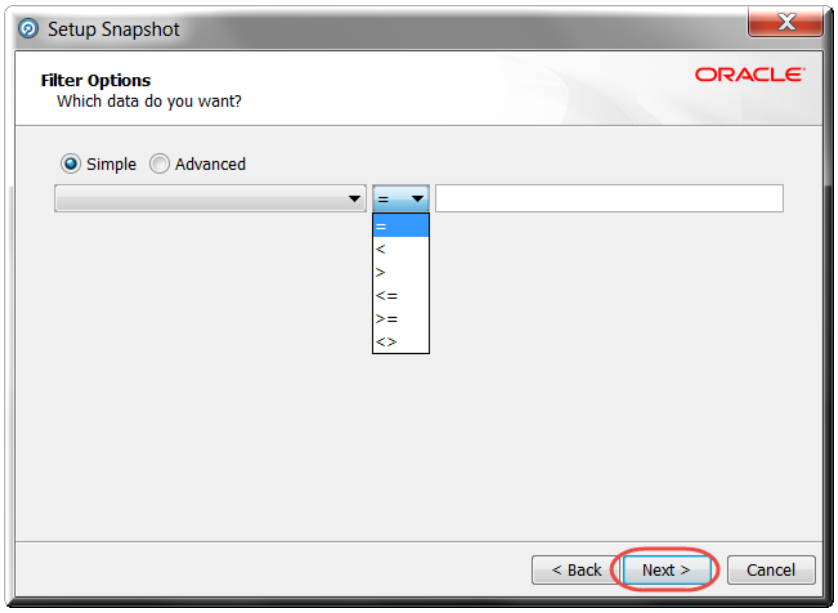
source of this particular EDQ Staged Data is the table **US_CUSTOMER_DATA**. Keep the **Select Table or View** radio button selected. Click **Next >** to continue



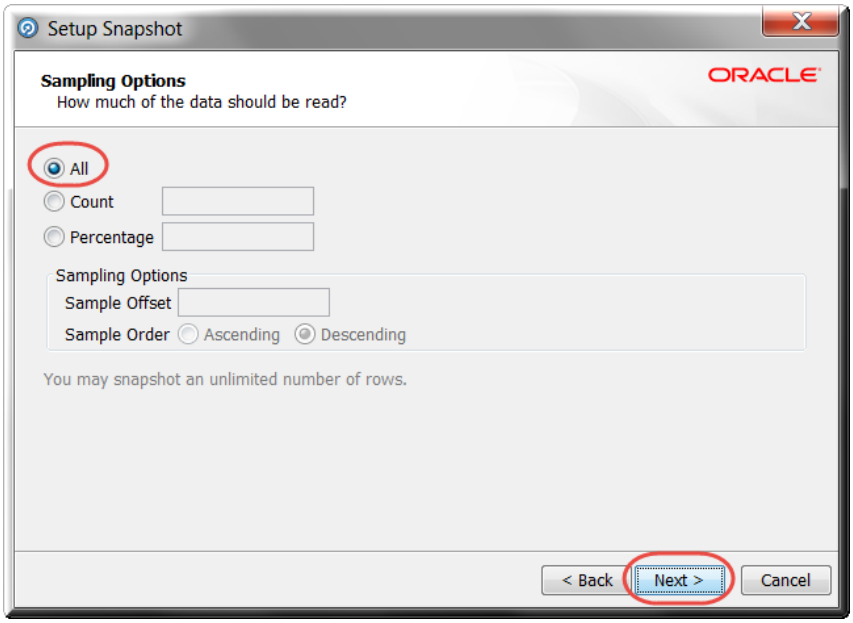
7. For **Column Selection**, ensure all columns are selected for setting up this data snapshot, then click **Next >** to continue



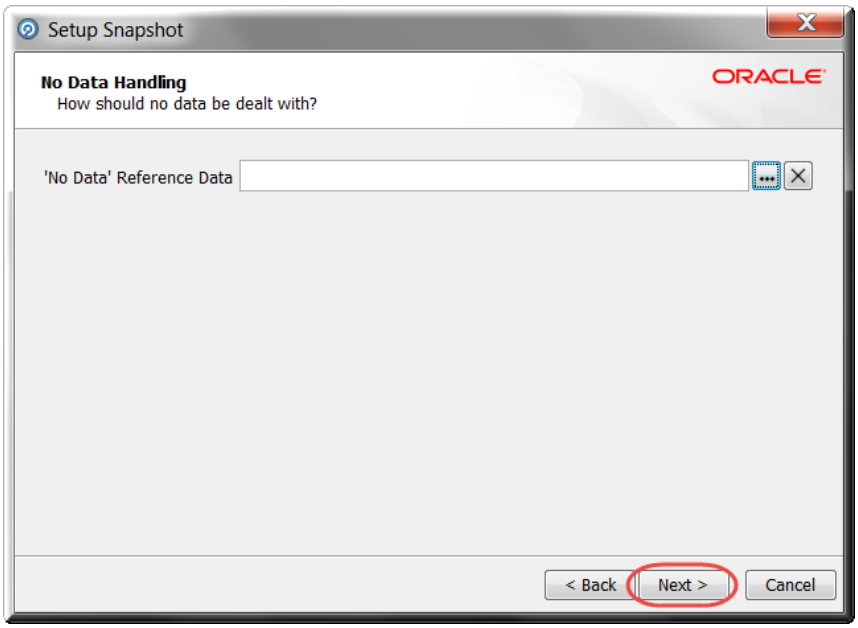
8. For **Filter Options**, notice the **Simple** and **Advanced** radio buttons to filter data by column and operators. The **Advanced** option allows for adding your own SQL query. Leave the default value of **Simple** and click **Next >** to continue



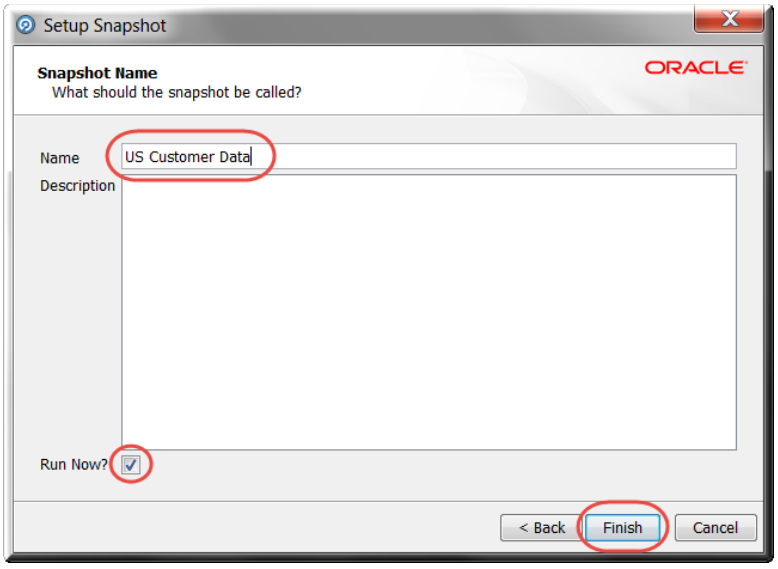
9. **Sampling Options** allows the behavior of the amount of data read into the snap shot to vary. By default, the **All** radio button selection is selected. If needed, you can specify a certain **Count** or **Percentage** of data to be read for the snapshot. In this example, select **All** for the sampling options, then click **Next >** to continue



10. Leave the default empty value for the **'No Data' Reference Data** Textfield. We will work with Reference Data later in another lab

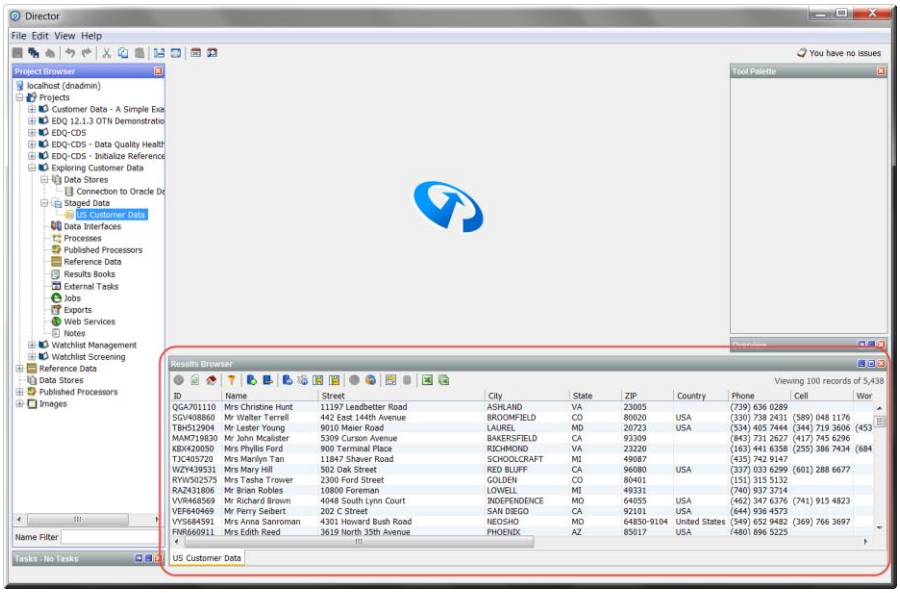


11. Give your snapshot the following name – *US Customer Data*. Ensure the **Run Now?** Checkbox is checked, then click **Finish** to complete and close the **New Snapshot** wizard



Notice that after a short delay, the **Results Browser** is populated with data originating from the Oracle Database and sourced from an EDQ Snapshot. Taking the Snapshot causes

Enterprise Data Quality to stage the data from the database into the EDQ data repository – meaning that a copy of the data from the database is placed in the Enterprise Data Quality repository. From now on we will be working with the data residing in the US Customer Data Snapshot



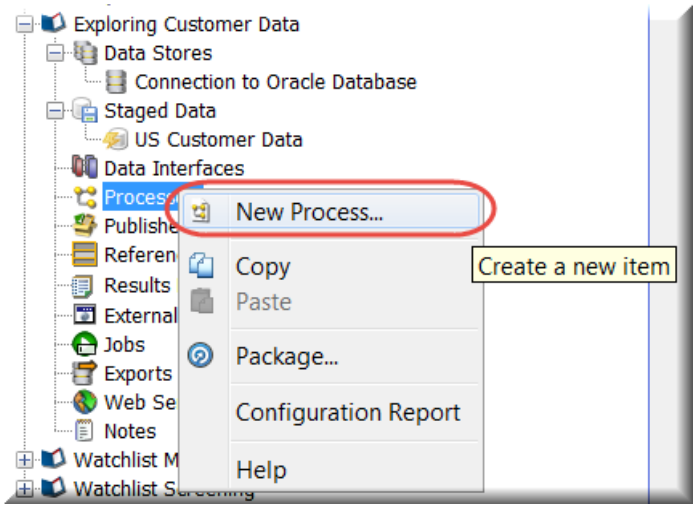
Lab 1b: Profiling your Data

The first step in improving the quality of your data is to understand it. You know you have 'Data' – but is it 'Data Fit for Use'? Enterprise Data Quality allows the user to quickly assess, find, investigate and understand anomalies regarding data content, standardization, relationships and duplication among others. EDQ enables users to understand their data by discovering, highlighting and communicating data anomalies within the data being profiled.

As you will learn, Profiling can lead to many different insights on your data sources and targets including outliers, minimum and maximum values, invalid dates and record completeness. It can also show the frequency with which particular values occur. For instance, how many unique values do you have in a field, and how many times is a particular value duplicated? These profiling topics and more are the focus of our following lab: Create and Run a Profiling Process.

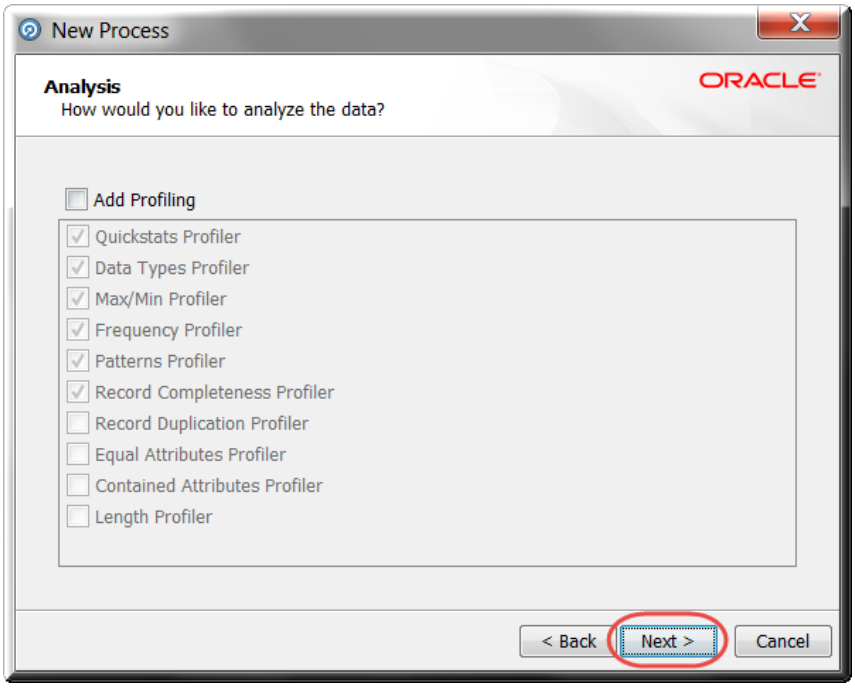
Create and Run a Profiling Process

1. Navigate to the Project Browser and right-click on **Processes** under your **Exploring Customer Data** Project, then click **New Process...**



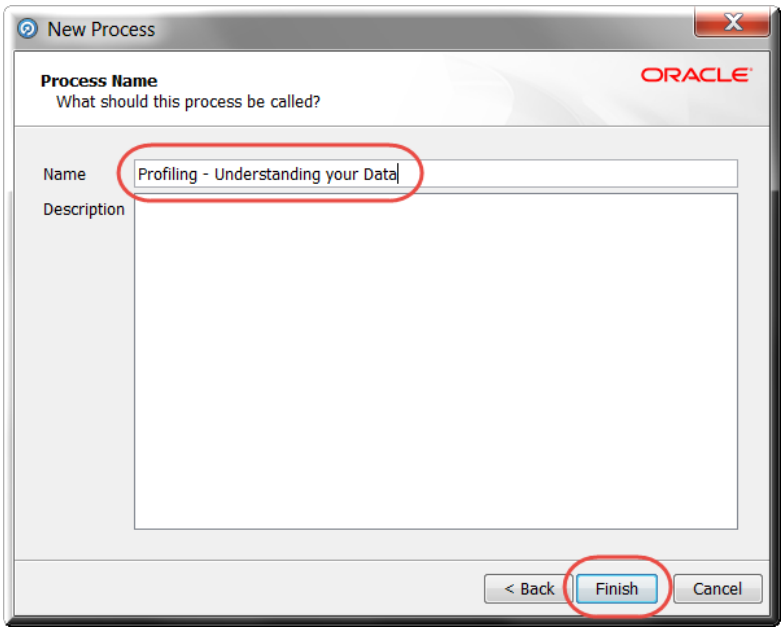
2. Select the previously Staged Data, **US Customer Data**, then click **Next >** to continue

3. Notice that you can optionally select **Add Profiling** while creating this New Process. We will add our own Profiling processors in the next few steps. Leave the checkbox unchecked and click **Next >** to continue



*When selecting **Add Profiling** when creating a new process, the data will be profiled for all Profilers selected. When you are creating a process for a large or wide dataset, it is recommended to add Profilers separately after creating the Process as it may take a long time to execute all the profilers as they are particularly intermediate data generation and compute intensive*

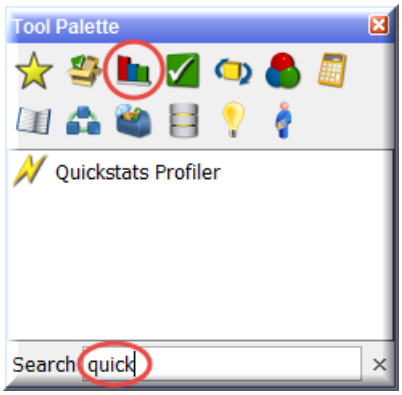
4. Give this process a name: ***`Profiling - Understanding your Data`*** – then click **Finish**



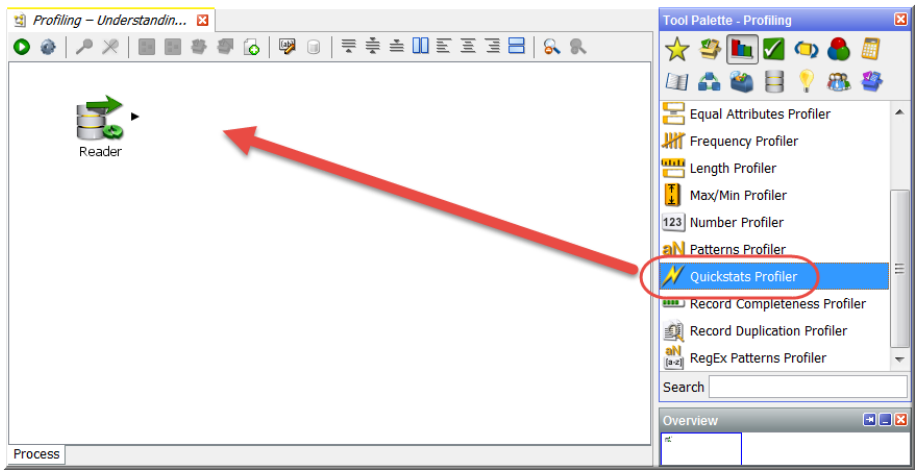
Congratulations! A tabbed Project Canvas is now presented with your newly created Process. You will note a **Reader** processor is automatically added to the Project Canvas. The term processor will be used to refer to the different pre-built objects that are dragged and dropped from the **Tool Palette**. In short, each processor can be configured in a process to perform some kind of operation on your data. In this case, since the Staged Data, **US Customer Data**, was selected while creating the process, a **Reader** processor that ingests **US Customer Data** has already been added to the canvas.

5. Navigate to the **Tool Palette** and find the **Profiling** icon. Next, find and select the **Quickstats Profiler** among the Profiling processor family

*Using the **Search** box underneath the Tool Palette allows you to quickly find Processors also.*

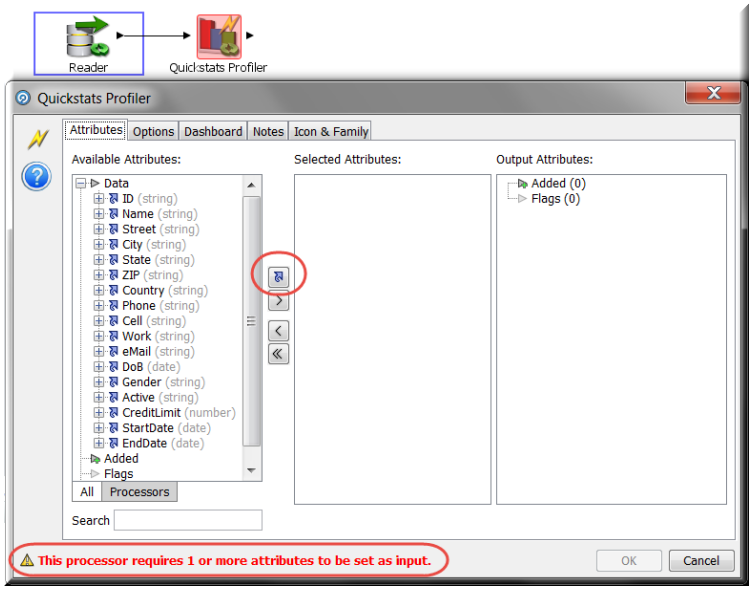


6. Drag and drop the **Quickstats Profiler** onto the Project Canvas



*Notice that the Process Canvas becomes active, the Process is created and a Reader and **Quickstats Profiler** has been added. The **green circular arrows** on any processor means it has yet to be executed.*

7. Hover over the output triangle of the **Reader** processor. An information tool-tip appears with the name and brief description of the processor. Click and drag from the output triangle of the **Reader** processor to the input triangle of the **Quickstats Profiler**. Upon successful connection and release of the mouse, the **Quickstats Profiler** configuration dialog will appear:

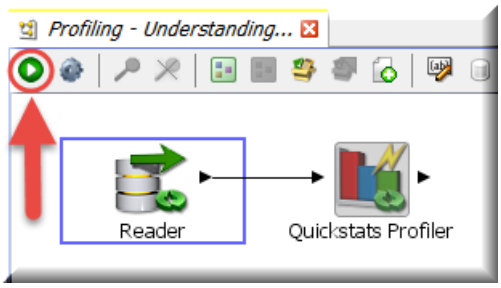


8. Note the message in red informing you that **This processor requires at least one attribute to be set as an input**. Click the **Select All** icon, as shown in the screenshot above. This will select all available attributes from our **US Customer Data** staged data. Click **OK** to save

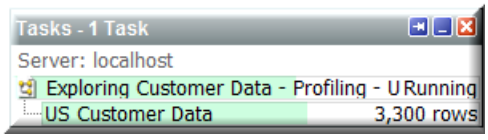
The new processor also shows the 'not yet run' green icon.

Once a process is run, results are stored in the Enterprise Data Quality repository and the green 'not yet run' icons will disappear. As subsequent processors are added, only those processors need to be run as long as the other processors are left unchanged. As you recall, when we created our **New Process**, we left the **Add Profilers** checkbox unchecked. As a result, adding the processors separately will save time and disk space because only the Profiler processors we need will be added and run rather than the default set of multiple profiler processors (some of which we may not need for the Profiling the staged data representing the **Reader)**.

9. The process now has a **Reader** and a **Quickstats Profiler**. Click the **Run** icon in the toolbar to run the process

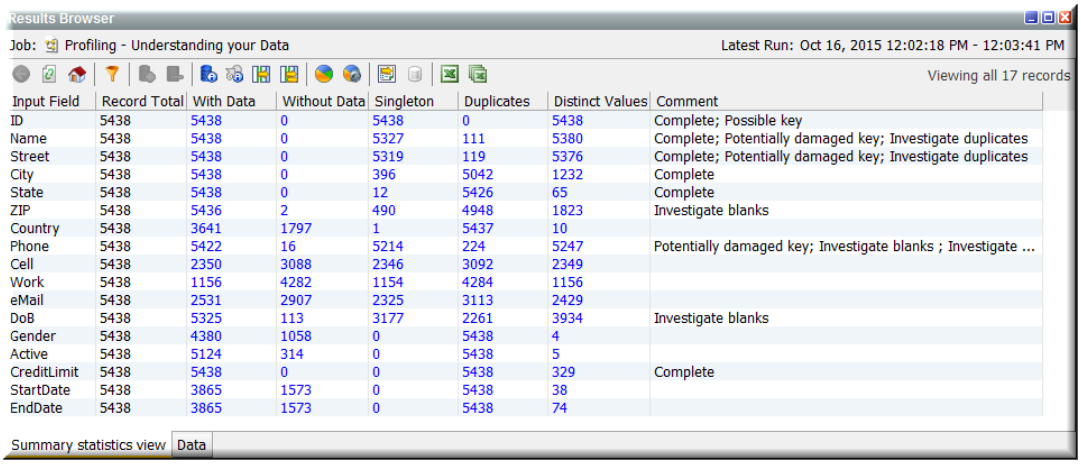


*The progress can be observed in the **Task Bar** in the bottom-left of the Director as the process runs. When the process has finished, the 'not yet run' icons will disappear from the canvas to show that the processors have data associated with them.*



10. Click the **Reader** processor to see the raw input data stored in the staged data snapshot. This will be displayed in the **Results Browser**

11. Next click the **Quickstats Profiler** to see the output of the processor



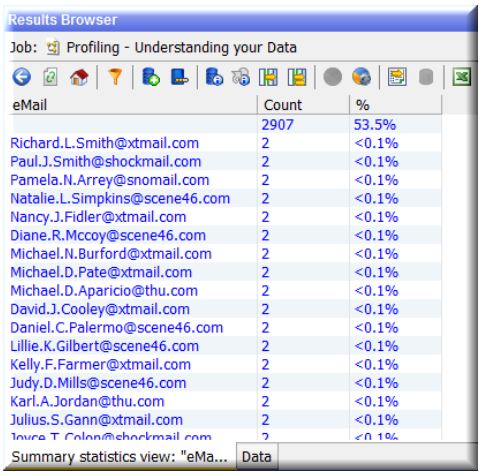
Input Field	Record Total	With Data	Without Data	Singleton	Duplicates	Distinct Values	Comment
ID	5438	5438	0	5438	0	5438	Complete; Possible key
Name	5438	5438	0	5327	111	5380	Complete; Potentially damaged key; Investigate duplicates
Street	5438	5438	0	5319	119	5376	Complete; Potentially damaged key; Investigate duplicates
City	5438	5438	0	396	5042	1232	Complete
State	5438	5438	0	12	5426	65	Complete
ZIP	5438	5436	2	490	4948	1823	Investigate blanks
Country	5438	3641	1797	1	5437	10	
Phone	5438	5422	16	5214	224	5247	Potentially damaged key; Investigate blanks ; Investigate ...
Cell	5438	2350	3088	2346	3092	2349	
Work	5438	1156	4282	1154	4284	1156	
eMail	5438	2531	2907	2325	3113	2429	
DoB	5438	5325	113	3177	2261	3934	Investigate blanks
Gender	5438	4380	1058	0	5438	4	
Active	5438	5124	314	0	5438	5	
CreditLimit	5438	5438	0	0	5438	329	Complete
StartDate	5438	3865	1573	0	5438	38	
EndDate	5438	3865	1573	0	5438	74	

The **Quickstats Profiler** provides fundamental quality metrics for a number of records or transactions, highlighting:

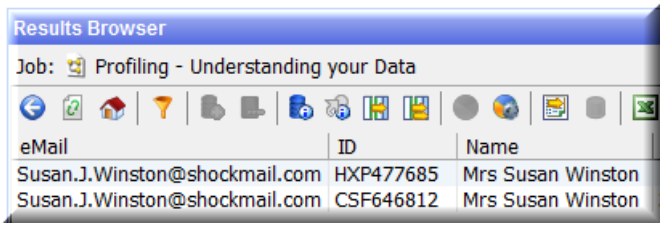
- Candidate key columns
- Completeness and missing data
- Duplication
- Uniqueness and diversity of values

For each **Input Field**, the number of records (**Record Total**), **With Data**, **Without Data**, **Singleton**, **Duplicates**, and **Distinct Values** are shown. These results can be observed and investigated to quickly find data anomalies. For instance, there are **4 Distinct Values** for the **Gender** attribute, when there should really only be two: Male and Female. You can also drill down on any blue text to see the data underneath.

12. Click the number **3113** listed for **eMail** under the **Duplicates** column in the **Results Browser**. Click the **Count** hyperlink for the **eMail** address containing no content / blank

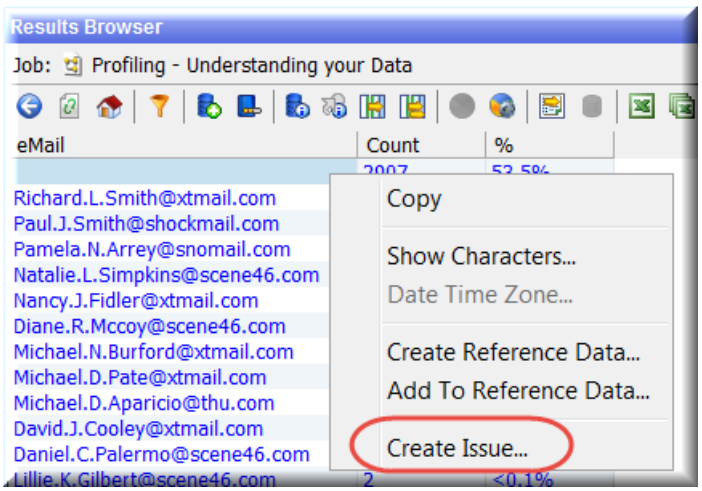


13. Click in the **Results Browser** to return to the previous view of Drill-down on one of the non-null values. We observe that there are a number of duplicate **eMail** values (**Count** of **2**) in the system that may require further investigation from a duplicate record standpoint

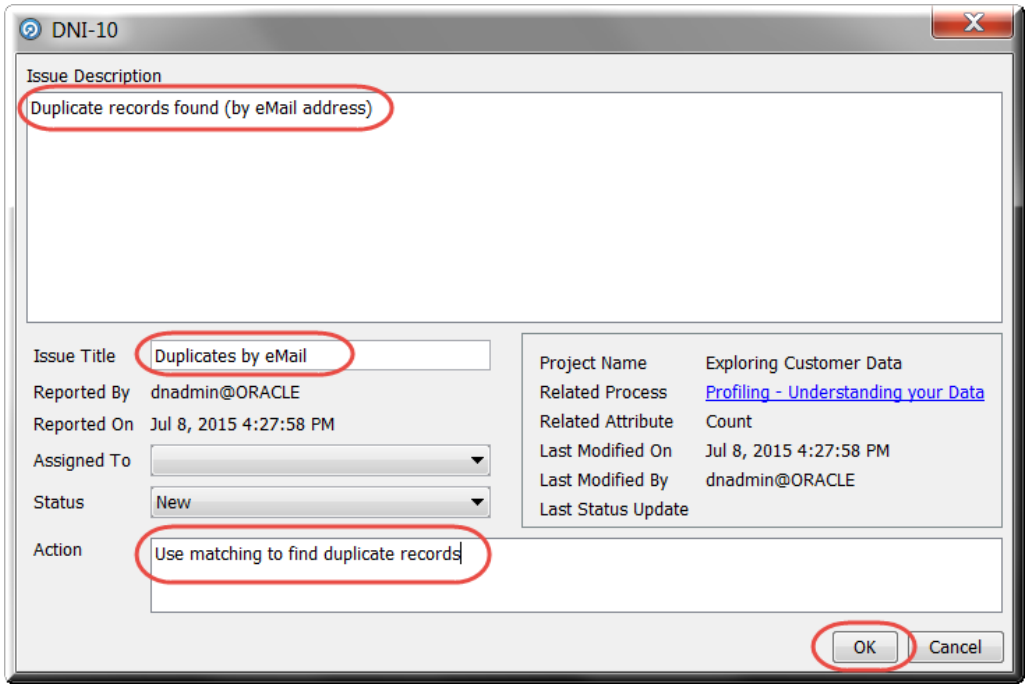


The large number of empty **email** values (**53.5**%) may also represent a 'Data Fit for Use' issue depending on the requirements / data SLA of the Customer data

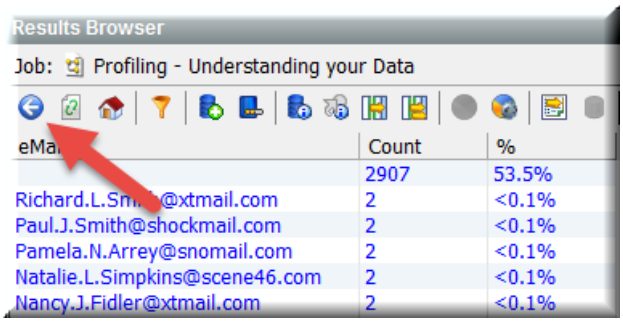
14. Enterprise Data Quality has many features to create a collaborative environment where users can raise issues to DBA's, Application Developers, ETL Developers, BI Dashboard writers and Business Owners as they are found. We will explore the **Issue Manager** in greater detail in an upcoming lab: Issue Management. To add an issue for duplicate (**Count** > 1) eMail records, right-click on a hyperlink field where the **Count** value **is 2** in the **Results Browser** and choose **Create Issue...**



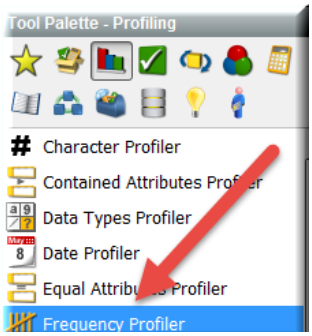
15. It is possible to assign the issue to yourself (**dnadmin@ORACLE**) or another user. You can also type in a follow-up **Action**: ***Use matching to find duplicate records***. The issue also includes a link to the process and results view where the issue was created. Under **Issue Description**, type ***Duplicate records found (by eMail address)***. Click OK and the issue is created



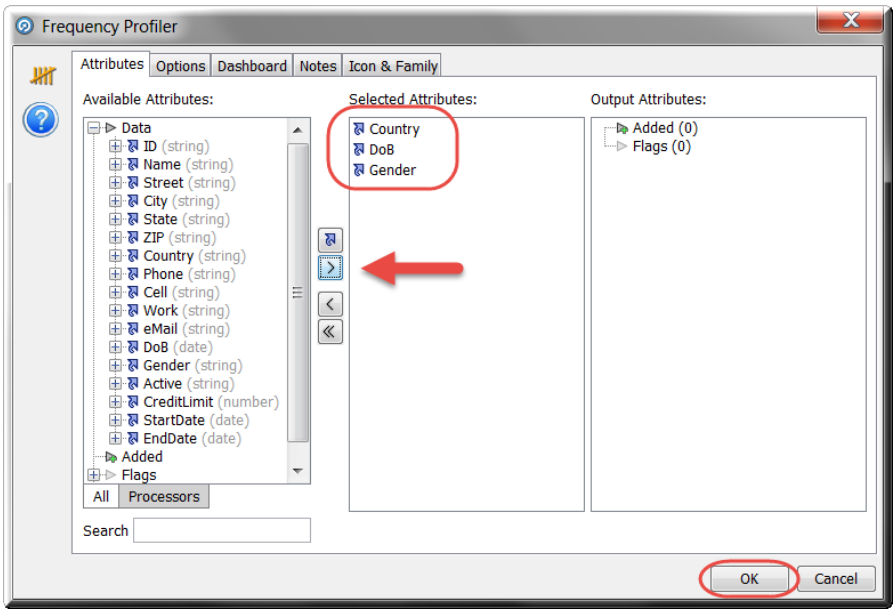
16. Click the back icon in the **Results Browser** to return to the results of the **Quickstats Profiler**



17. Return to the **Tool Palette - Profiling** and find the **Frequency Profiler**. Drag and drop the processor onto the Project Canvas and link the output triangle of the **Quickstats Profiler** to the input of the **Frequency Profiler**



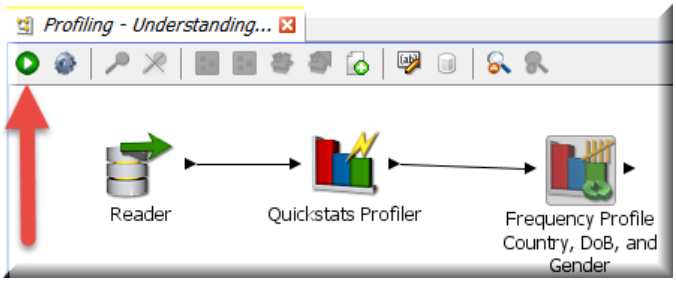
18. The **Frequency Profiler** dialog appears. Multi-select the **Country**, **DoB**, and **Gender** **Available Attributes** and select the  icon to add the attributes to your **Selected Attributes**, then click **OK**



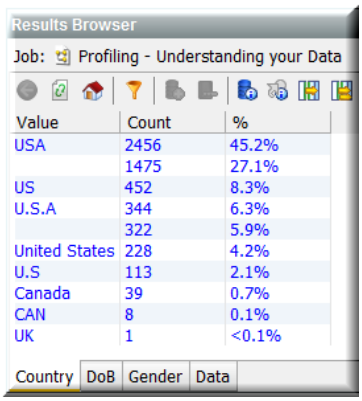
19. Processors can be renamed by double-clicking on the name of the processor within the canvas. Double click on the existing label of the **Frequency Profiler** and enter ***`Frequency Profile Country, DoB, and Gender`*** to rename the processor

Right clicking on the processor icon and selecting **Rename** also allows renaming of the processor

20. Click the **Run** icon  to start the process as the Frequency profiler has yet to run and we want to view the results. Wait for execution to complete



21. Click the **Frequency Profile Country, DoB, and Gender** processor to view the results in the **Results Browser**. Notice the 4 distinct tabs at the bottom left corner: **Country, DoB, Gender** and **Data**. Let's take a moment to analyze what these different tabs tell us about our data set



Having separate tabs for each **Selected Attribute** is one of the many nice user interface features of EDQ – usability being one of the top 3 differentiators for EDQ

This specific processor happens to tell us a lot about our data set just by observing the different values. Notice how many different representations of United States of America there are: **USA**, **US**, **U.S.A**, **United States**, and **U.S**. This discrepancy can cause major issues with Business Intelligence (BI) dashboard authors and analytics dashboards results. If the BI team is asked to report on United States sales, the Business Intelligence dashboard author must do one of the following (neither of which are appealing):

- Undercount US Sales because he/she picks "USA" to look for and does not know there are 5 (and perhaps rising over time) different (but similar) valid representations of United States in the Country column. Such an approach results in the undercounting of US Sales (a 55% undercounting to be exact).

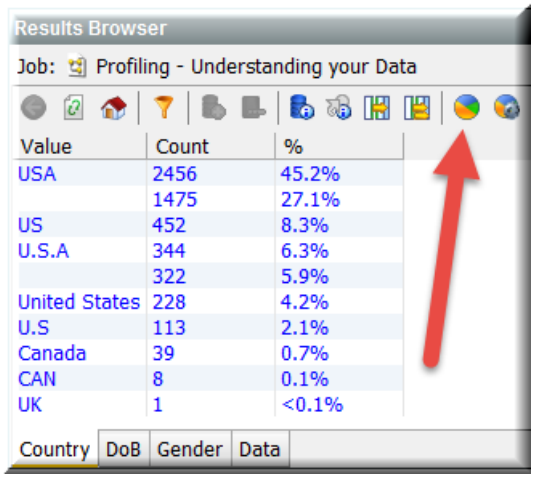
- Have to ask "Which United States?"

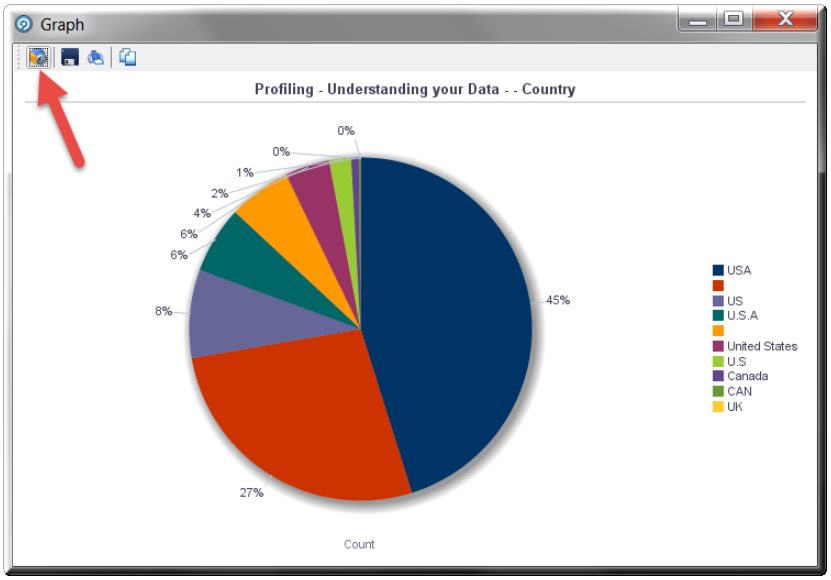
- Be put under the burden to identify all the variants of acceptable representations of "United States" in the Country column. And not just Country column, all Columns can have standardization content issues – putting the onus on the Business Intelligence report and dashboard authors to ferret out and identify all possible acceptable variants and incorporate that logic in the Business Intelligence report/dashboard implementations. Why put 'dirty data' in the Data Lake – why not clean the water before it is put in the Lake?

Finally there are over 27% of the rows with no data in the Country column. While we will not do so in this Workshop, one could easily create States Reference Data and if Country is blank and the State column is one of the valid 50 states – set the Country to the standardized value of "**United States**" – furthering the efforts in transforming the Customer dataset from

something more than "just data", but 'Data Fit for Use". Not just "Analytics", but "Accurate Analytics".

22. Click the **Graph Results** graphical button  in the **Results Browser** to see a chart of the different values presented in the **Graph** dialog



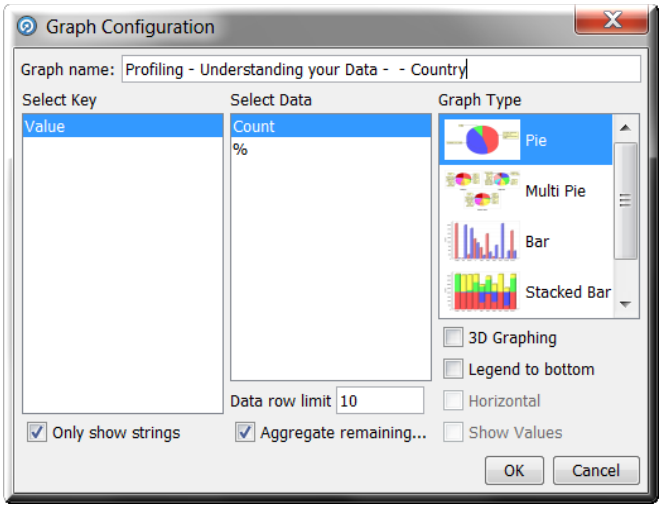


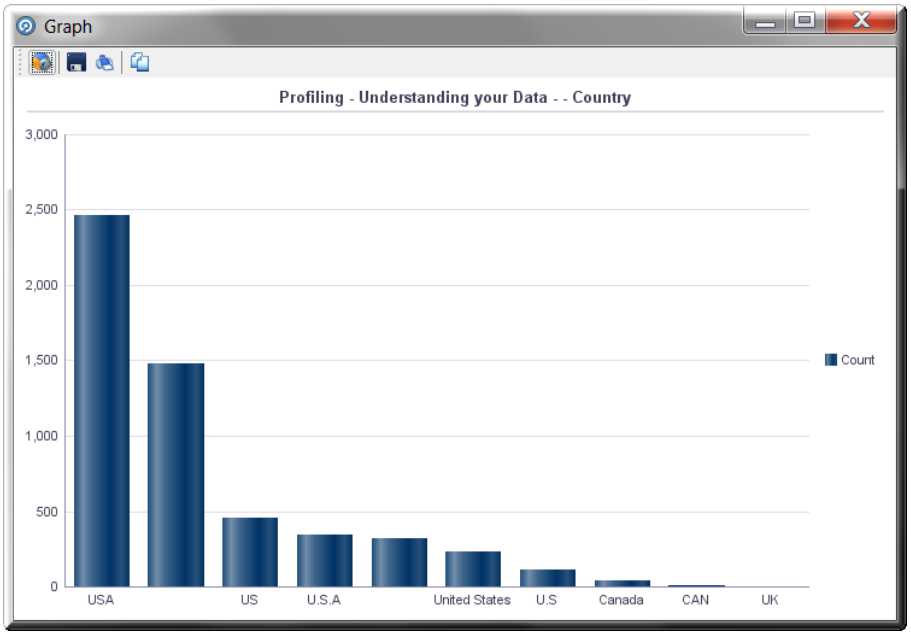
As we can see – despite the fact that 1% of the orders are outside the US, over half of the data contains a value other than the most common value of **USA**

23. The first button in the toolbar as shown in the screenshot above, will allow you to change the title of the Chart, configure the Chart to a different type of visualization or modify the type of data displayed. Click this button to open the **Graph Configuration** dialog. While the graphing capabilities in EDQ are not an OBI-EE, Cognos or Tableau,

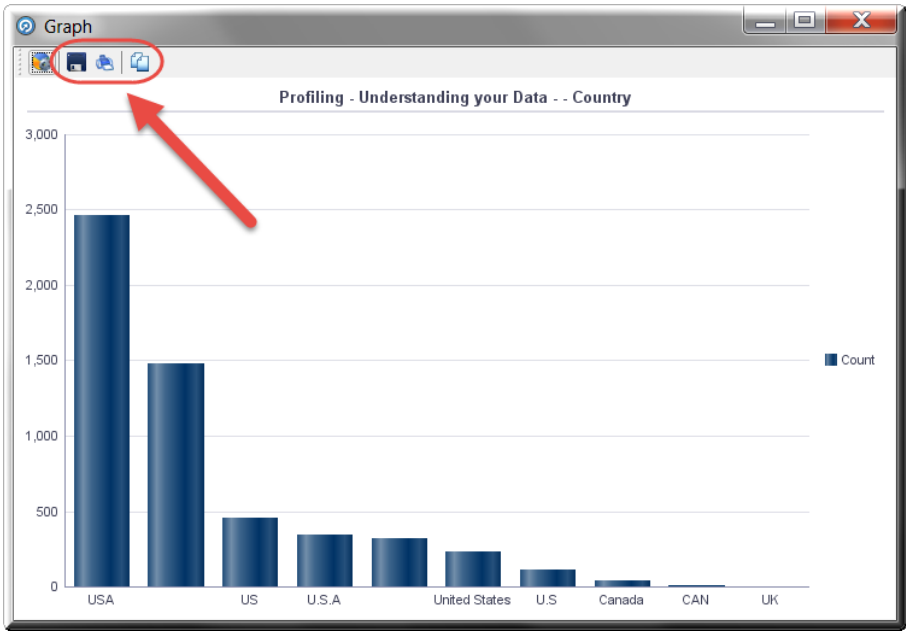
being one click away from visualizing the **Results Browser** … you know the old adage – 'a picture is worth 1,000 words' and worth its weight in gold when collaborating with others on the DataMart / Data Warehouse / Data Governance / MDM / Data Scientist and/or Chief Data Officer teams(s)

24. Select the **Bar Graph Type**, then click **OK**

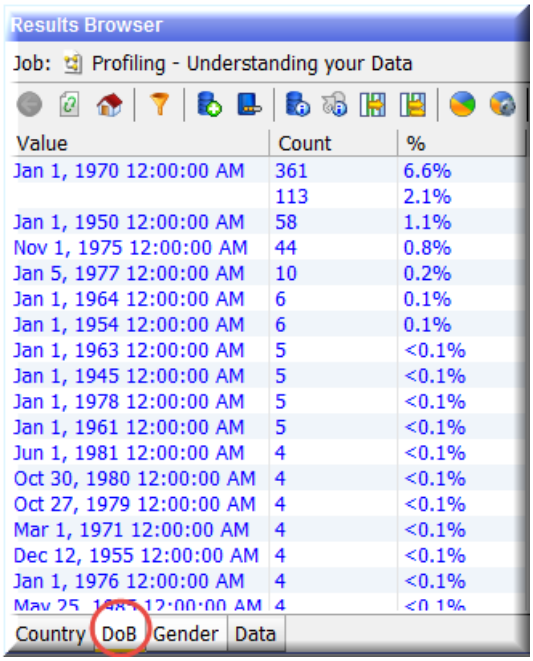




25. The next three buttons in the chart allow you to save it to an image file, print, or copy to the clipboard to enable sharing with others



26. Close the **Graph** dialog window and click the **DoB** tab in the bottom of the **Results Browser** to view the results of the **Frequency Profiler**



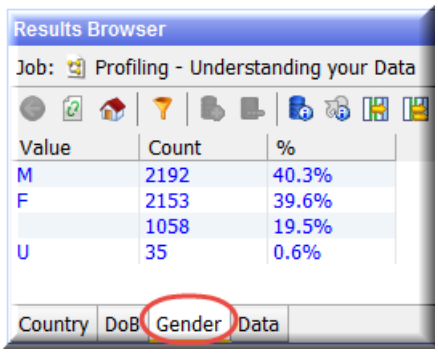
Note that the columns can be sorted by clicking on the various column headers (click on **Count** if not already sorted by count). Column sorting in the **Results Browser** can be another good

exploratory technique to quickly identify additional issues with your data. For instance, you will notice many individuals with a birthday on **Jan 1**. This may indicate that there was some sort of default value used with Jan 1 and that the quality of the **DoB** column may be low.

As another example, if you saw one DoB value entry showing an unusually high percentage of the total row counts (such as **6.6%** having **Jan 1, 1970**) – this would represent another hidden issue with the Data (likely a spurious default value used by the source data system if the value were whitespace). Every single row with Jan 1, 1970 of course inserted into the source database just fine … but the data is not 'Fit for Use' for age banding or other age analysis of the data unless attempts are made to fix / enrich the DoB field.

Lack of 'Fit for Use' of data (default, blank and null values for DoB) limits 'Fit for Use' – limiting the data's 'Fit for Use' as Report / Dashboard authors unknowingly include erroneous data in the results of their Dashboards. Additionally, Dashboard consumers losing confidence in their Reports/Dashboards is the **number 1 cause of Data Warehouse / Data Mart / Data Governance project failures (Gartner)**

27. Click the **Gender** tab in the bottom of the **Results Browser** to view the results of the **Frequency Profiler**

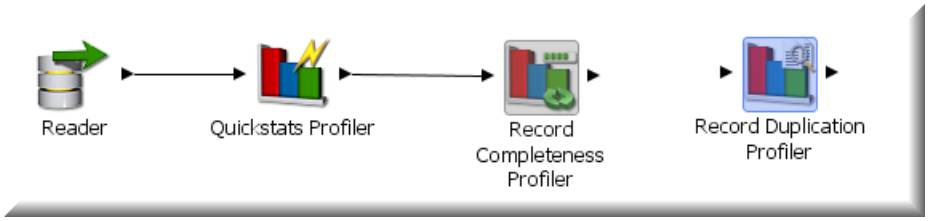


By now, you will surely see that this dataset needs some fine tuning to make it usable for accurate analytics. For instance, 19.5% of the gender values above are blank. In a later lab we will be able to use a processor to enrich the blank gender values so that more than 98% of gender values are 'Fit for Use' (enabling analysis by Gender as needed)

28. Let's add a few more processors to the **Profiling – Understanding your Data** process. Return to the **Tool Palette - Profiling** and find the **Record Completeness Profiler** and **Record Duplication Profiler** by using the Search box. Type ***Record*** in the ***Search*** textfield to find the processors and drag and drop these processors to the Project Canvas

29. Click and drag from the output triangle of the **Quickstats Profiler** processor so that the connector line reaches the input triangle of the **Record Completeness Profiler**

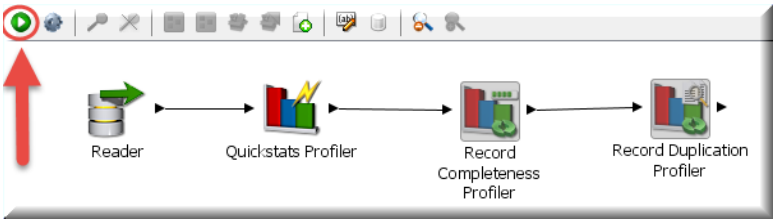
30. The **Record Completeness Profiler** configuration dialog appears. Click the select all icon to have all the data columns participate, then click **OK**



31. Click and drag the output triangle of the **Record Completeness Profiler** to the input triangle of the **Record Duplication Profiler**

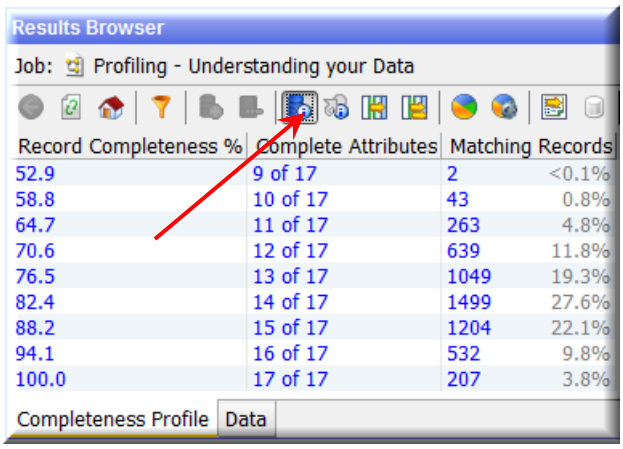
32. The **Record Duplication Profiler** configuration dialog appears. Click and select the **Name** attribute from **Available Attributes** and click the icon to move the attribute to the **Selected Attributes**. Similarly, click and select the **Zip** attribute, then click **OK**

33. Click the **Run** icon in the toolbar to run the process

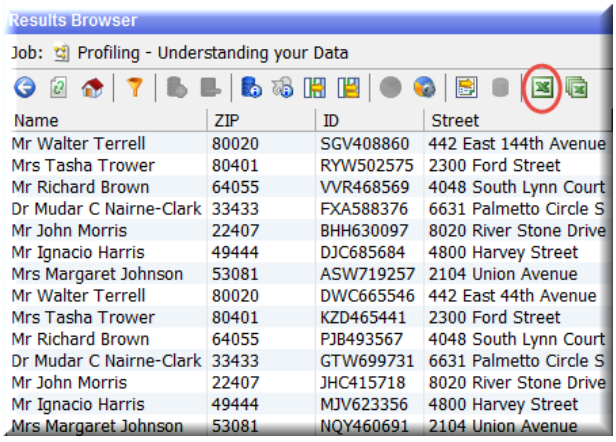


Just as the **Quickstats Profiler** provided many details about the dataset, the **Record Completeness Profiler** will analyze records with all of the selected attributes to display completeness. The **Record Duplication Profiler** will analyze records for duplicates across the selected Name and Zip attributes.

34. Click on the **Record Completeness Profiler** processor to view the results in the **Results Browser**. You can see that only **207** of the customers in the **US Customer Data** have all **17 of 17** attributes filled. Click the **Show Additional Information** icon . Notice that those **207** complete records only make up **3.8%** of the entire dataset



35. Click on the **Record Duplication Profiler** to view the results. Drill down on the **14** representing Duplicated records. We learned how to raise issues earlier in the lab, you can also export the results to an Excel file to send to an individual in the organization for further investigation. Click the **Export to Excel** icon in the **Results Browser** toolbar to save the file. (You do not need to save the file)



 In the case where there are multiple tabs within the Results Browser, you can click the button next to Export to Excel to **Export all tabs to Excel**.

While we can continue to add Profilers to further investigate the US Customer data, perhaps it is best to move on to explore the next family of EDQ Processors – **Audit** – which will help us check and standardize the data

Lab 2: Auditing your Data to confirm variances. Creating and using Reference Data.

In this Lab, we will focus on EDQ Audit processors. So what is an Audit processor and why would we want to use them? Audit processors, or **checks**, check input data using business rules in order to assess whether or not it is fit for its business purpose. Which data attributes to check are often determined by:

- Data Profiling activities
- Business rules comprising a set of data acceptance checks forming a data acceptance SLA (Service Level Agreement) between two different parties – the sender and the recipient. Why would you and your fellow co-workers perform data break fix tasks if a large percentage of the input data from the sender is deemed not 'fit for use'? A policy of 'return to sender' for fixing the data may be a more appropriate policy for failing to pass data acceptance SLA checks

What data would you check (Audit) for 'fit for use' compliance as a result of Lab number one? Country? DoB?

Audit processors categorize each input record as to whether it is **valid** or **invalid** according to the **check**. Audit processors provide separate and easily accessible data output streams for valid and invalid records enabling separate workflows for handling valid and invalid records within an EDQ Process. Audit processors implicitly use the business rules that you apply to a given data attribute when profiling. For each type of business rule that you can apply, there is an Audit processor.

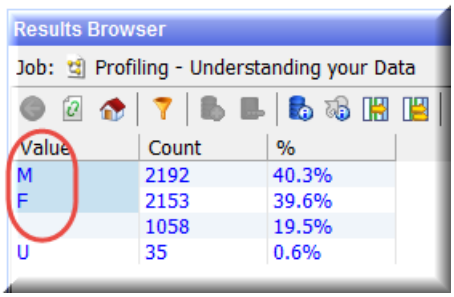
In Lab 2, we expand upon our Data Profiling activities performed in Lab 1 and carry out Audit data checks using EDQ Audit processors (these checks can be thought of as your "**Data Quality Firewall**" for your Data Warehouse / Data Mart) leveraging:

- the data anomalies we discovered in Lab 1
- the rich set of open and extensible EDQ Reference Data (EDQ's 'secret sauce')

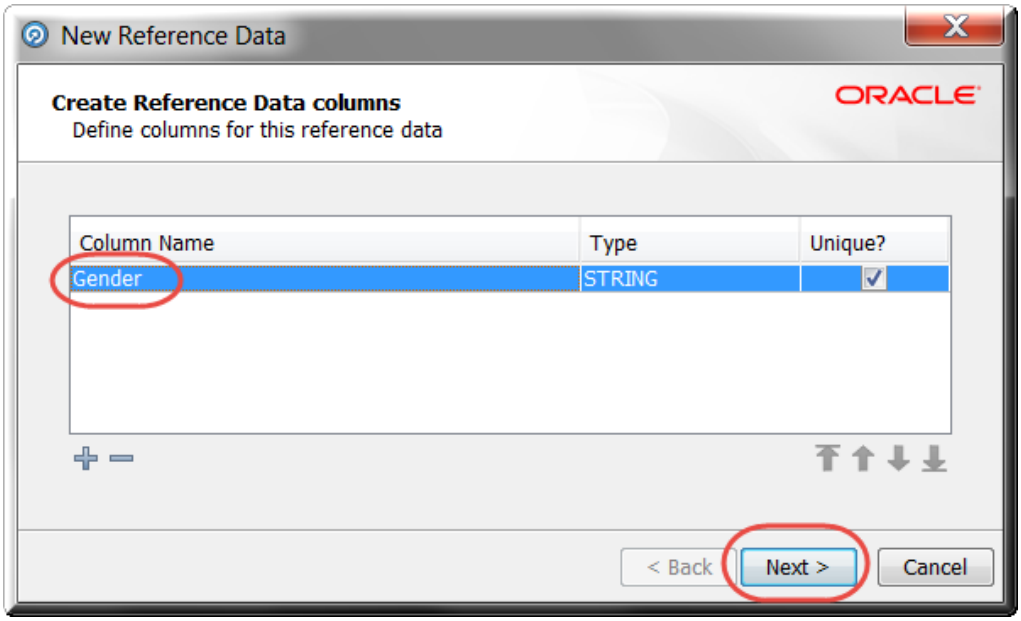
Create Reference Data

1. First, return to your **Profiling – Understanding your Data** process in the left-hand side of **Director**. You should be here already, but if not, find your project (**Exploring Customer Data**) under the Project Browser and expand/double-click on your Profiling process, **Understanding your Data**. Click on the **Frequency Profile Country, DoB, and Gender**, then click on the **Gender** tab in the bottom left corner of the **Results Browser**

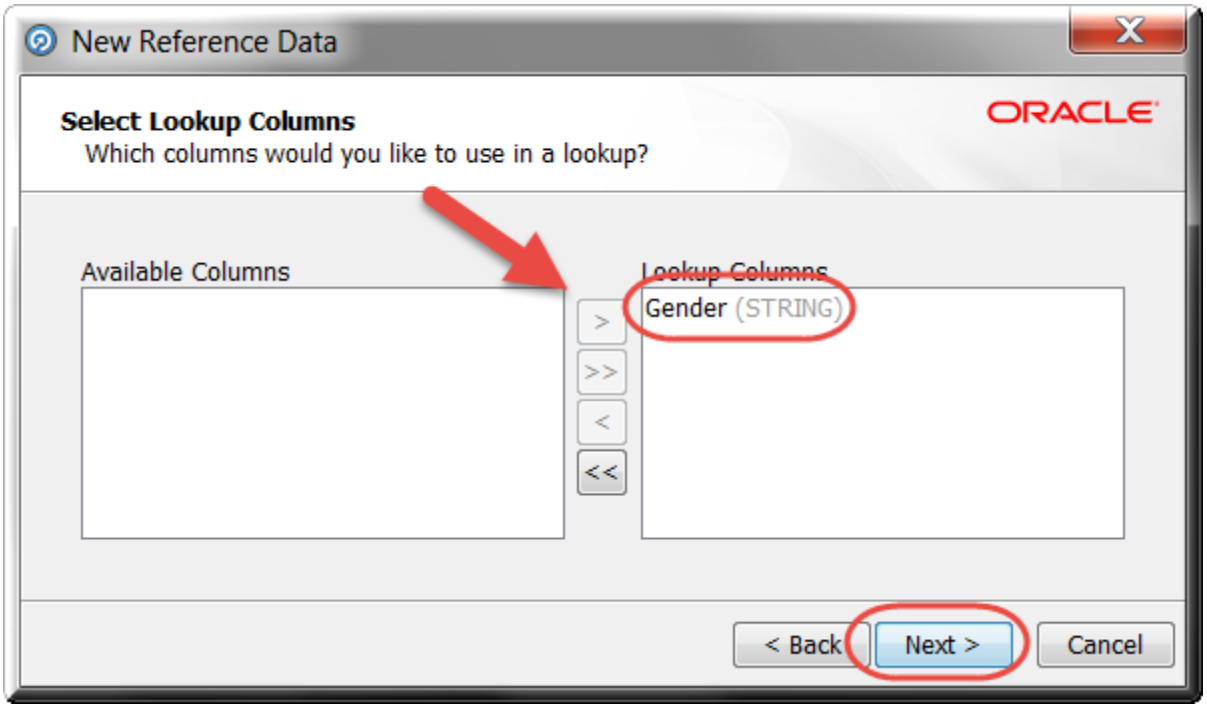
2. Hold down CTRL key and click on the **M** and **F** values

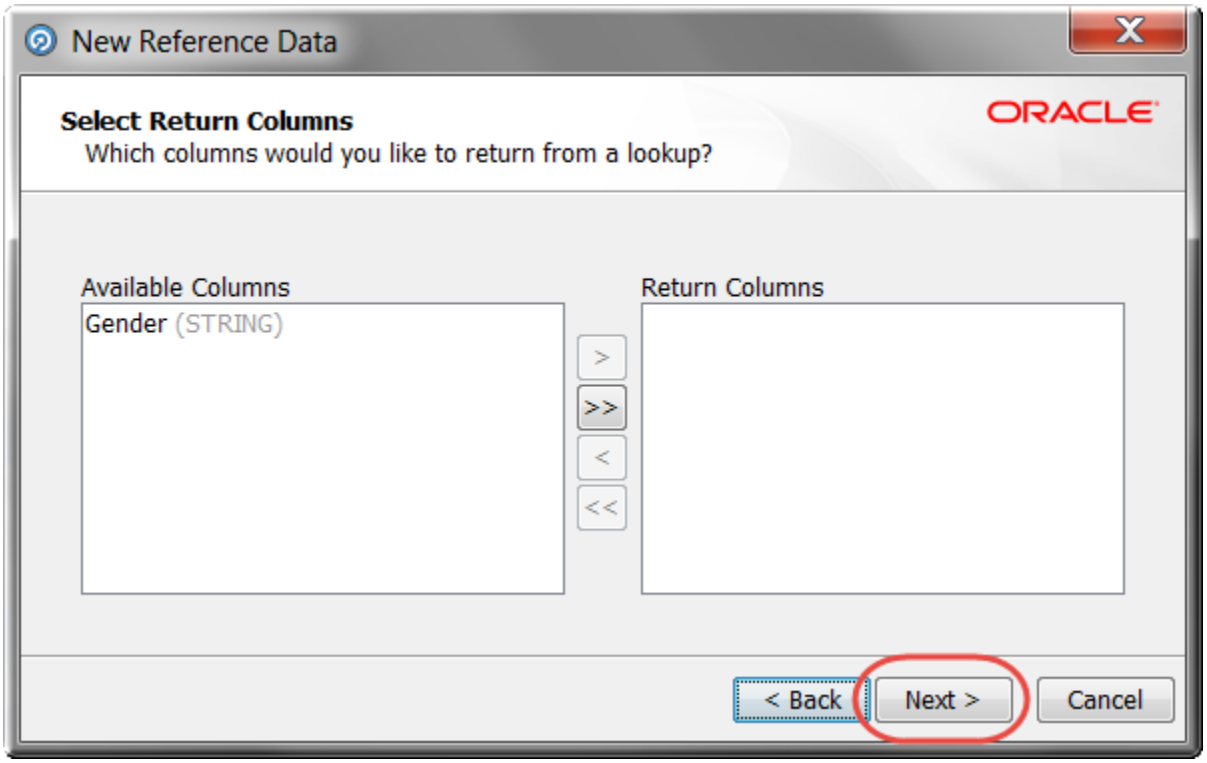


3. Right-click and select **Create Reference Data**. The **New Reference Data** dialog appears. Rename the attribute name to ***Gender***, click **Next** to continue

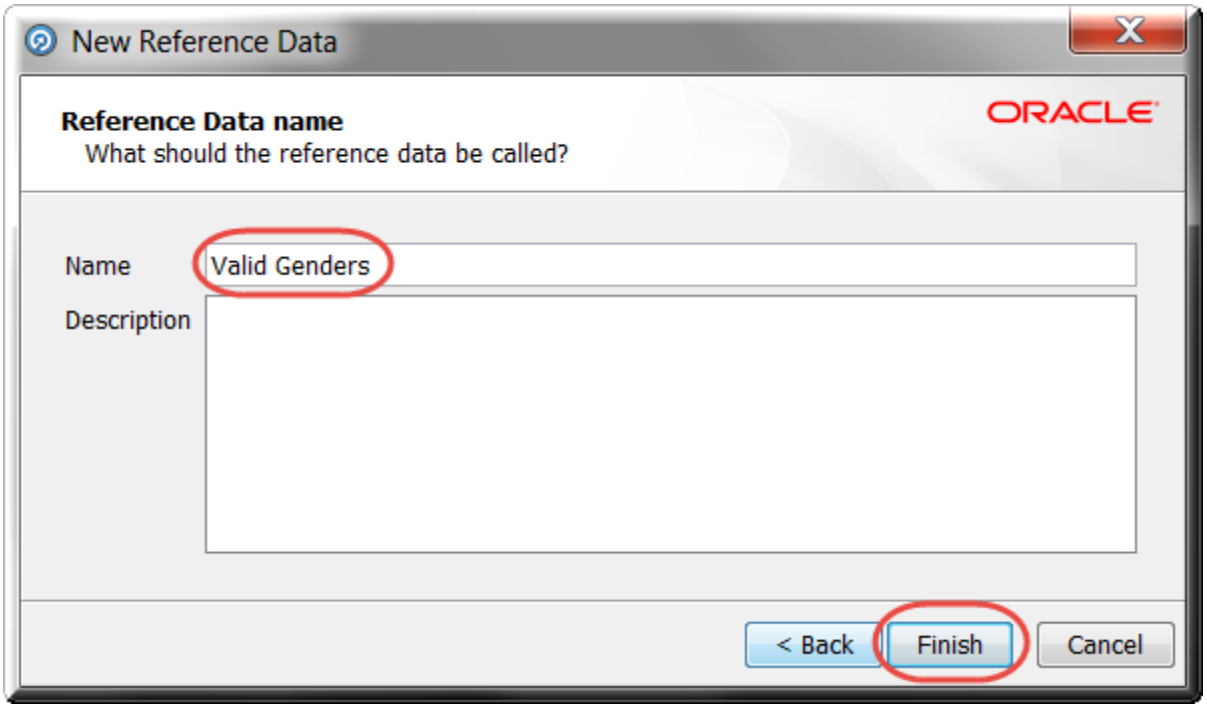


4. Add **Gender** to the **Lookup Column** using the **>** button, then click **Next >**. Click **Next >** on the next two screens to keep the default settings to continue (we will not add a return column or associate (classify) this reference data with any category)





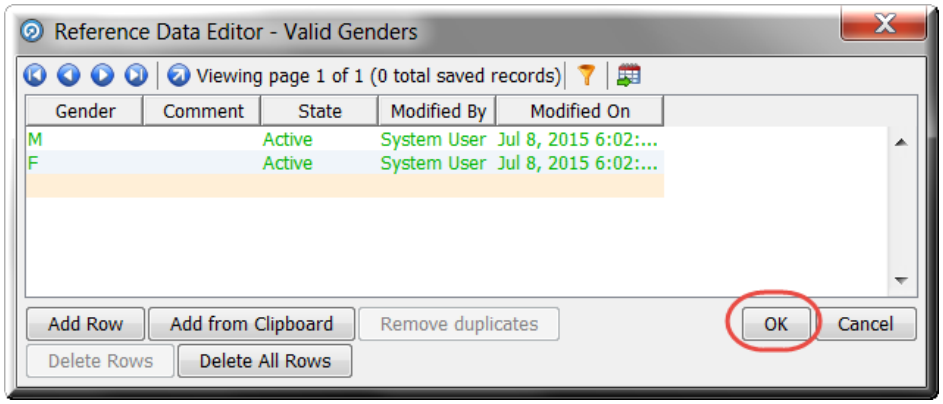
5. Finally, provide a name for this Reference Data: ***`Valid Genders`***, then click **Finish**



6. The **Reference Data Editor** appears next

Here, you can modify the Reference Data to Add Rows or Delete Rows. EDQ comes with many different types of Reference Data out of the box which can dramatically speed up the time it takes to create data check processes. We will see some examples of the out of the box Reference Data in subsequent labs.

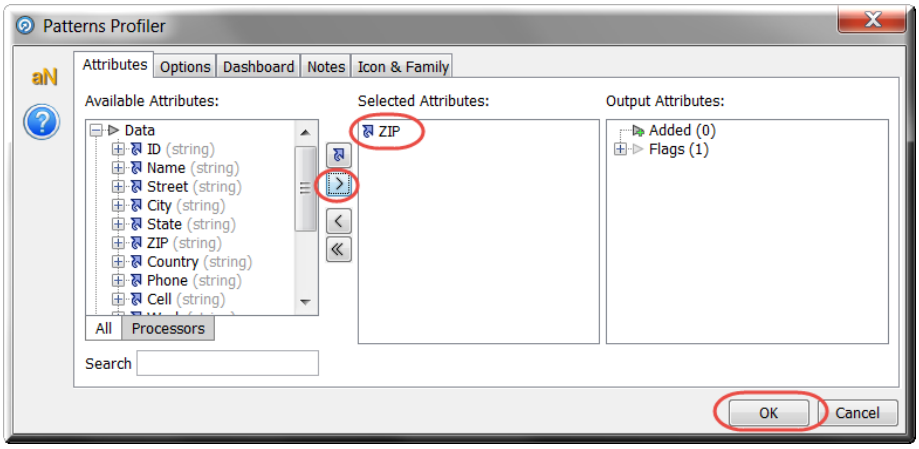
7. Here, you can see that EDQ has harvested the M and F values from your Profiling results to create your own custom Reference Data. We have just created the **single source of truth** definition of a **valid gender**. Click **OK** to return to the Project Canvas



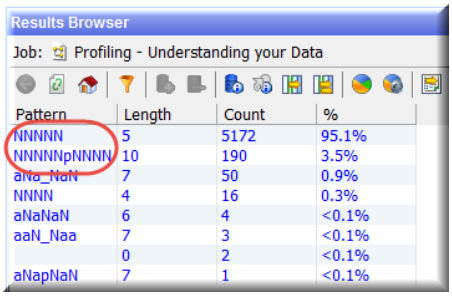
Next, we will need to create Reference Data for the valid types of ZIP Codes (our single source of truth for what a valid zipcode is). In this case we want 5-digit number strings or a 5-digit number followed by a hyphen and a 4-digit number. To easily create this Reference Data, we need to profile the ZIP code attribute using a Pattern Profiler.

8. Navigate to the **Tool Palette** and find the **Pattern Profiler**. Drag and drop this into the **Project Canvas** and drag and drop the end triangle from the **Record Duplication Profiler** to the **Pattern Profiler**

9. The **Pattern Profiler** configuration dialog appears. Select **ZIP** from the **Available Attributes** and press the **>** button to add it to **Selected Attributes**, click **OK** to continue



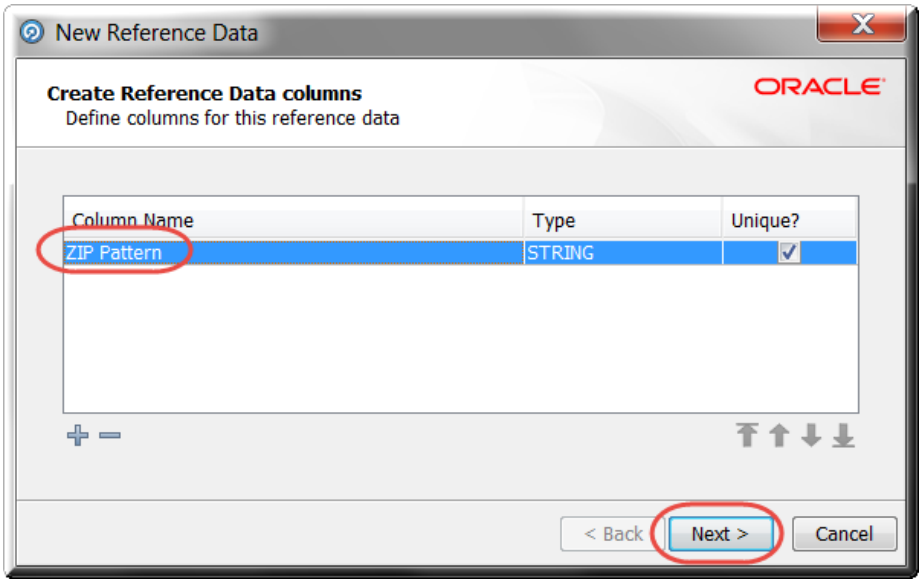
10. Click the **Run** button in the **Toolbar** in the top left corner above the **Process Tab**. After the Process completes, click the **Pattern Profiler** and view the results in the **Results Browser**



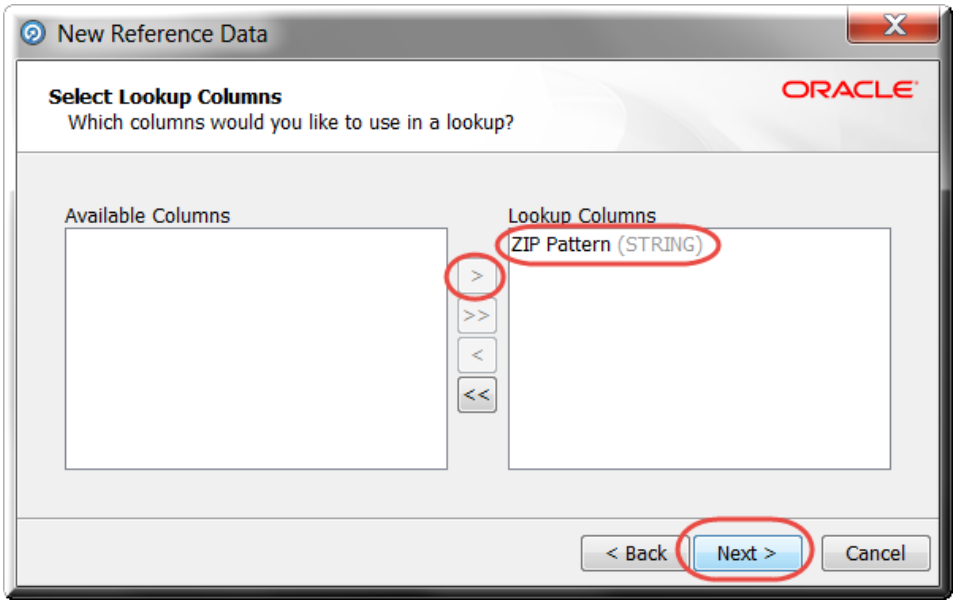
***N** signifies a number, **p** signifies punctuation, **a** signifies an alpha character, and _ signifies a space.*

11. Now that we have Patterns to create Reference Data from, repeat the steps taken when creating the Valid Genders Reference Data. Since we want 5 digit or 5 digit followed by 4 digits, CTRL click on **NNNNN** and **NNNNNpNNNN**

12. Right-click on the **NNNNN** and select **Create Reference Data**, and the **New Reference Data** Dialog appears. Rename the **Column Name** to ***Zip Patterns*** by double clicking on **Pattern** under Column Name, click **Next >** to continue



13. Add **ZIP Pattern** to the **Lookup Columns** using the **>** button, then click **Next >** on the next two screens (we will not add return columns or select a category) to continue



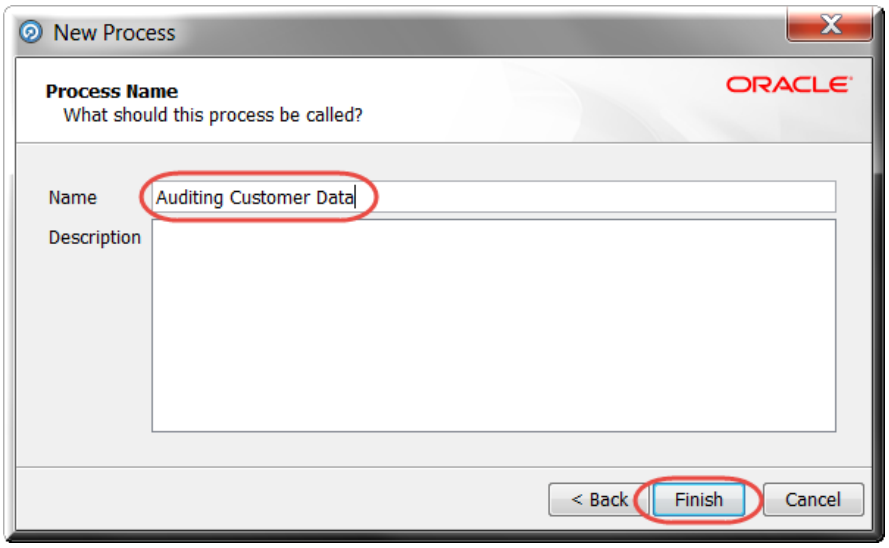
14. Last, give this New Reference Data a name ***`Valid ZIP Patterns`***, click **Finish**. The **Reference Data Editor** appears. If any values were selected by accident, remove those rows. Otherwise, click **OK** to continue. Voila! We have just created a **single source of truth** definition for the pattern of a valid zipcode

We will now begin to create a new Process for Auditing our US Customer Data. The Reference Data we just created in the past few steps will be utilized by some of the out of the box Audit Processors within our Audit (data checking) Process.

Create Audit Process

15. Return to the **Project Browser** in the left side of your Director window, and underneath your **Project (Exploring Customer Data)**, right-click on **Processes** and click **New Process...**

16. Select the **US Customer Data**, then click **Next >**. Click **Next** on the next screen (we will not add any more profiling here). Name this new process ***`Auditing Customer Data`***, then click **Finish** to continue

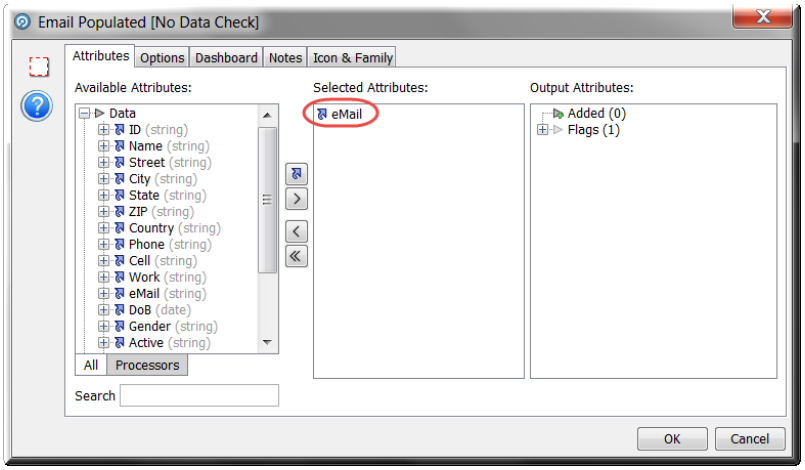


17. As with the first process we created, **a Reader Processor** is automatically added to the Project Canvas. Navigate to the **Tool Palette** to find the Audit  icon

Take a moment to review the different Audit Processors. Hover your mouse-tip over the different entries to get a brief description.

18. First, drag and drop a **No Data Check** processor onto the Process canvas. Right click on the **No Data Check** processor and select **Rename** to re-name it to ***Email Populated*** and press the enter key. Drag and Drop the end triangle from the **Reader** to your newly named **Email Populated** Audit process

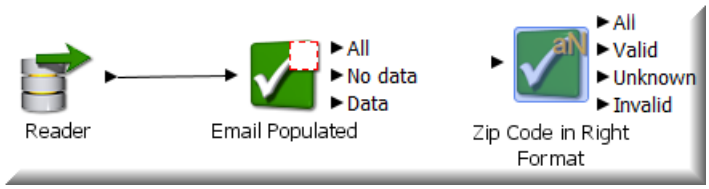
19. The **Email Populated** configuration dialog appears. Select **eMail** from **Available Attributes** and click the button to add it to **Selected Attributes**, click **OK** to continue



20. Click **Run** in the **Toolbar** above the **Auditing Customer Data** process tab and select the **Email Populated** audit processor to view the results

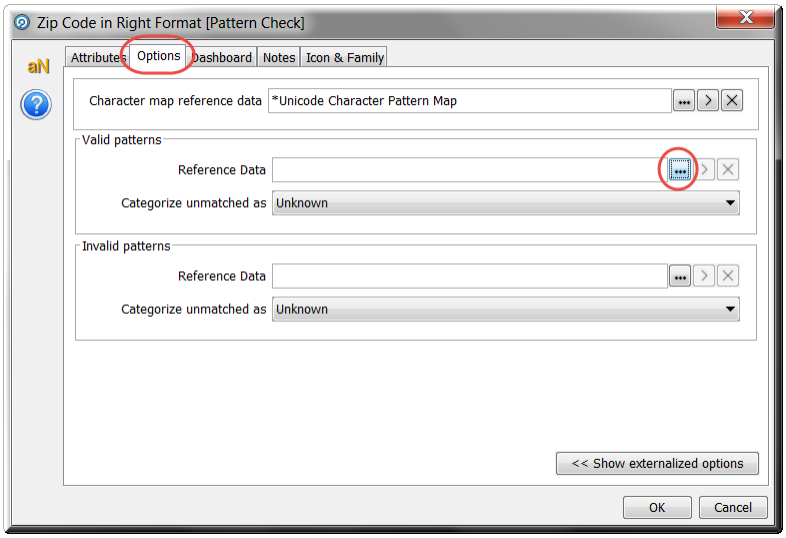
> Note the **Without Data** and **With Data** values in the **Results Browser**. If desired, we can continue to develop this process using one or more of the end point output data stream triangles from the Processor by choosing **Data**, **No Data** or **All**.

21. Next, find the **Pattern Check** Processor in the Tool Palette. Drag and drop it into the canvas and rename it to ***Zip Code in Right Format*** by right clicking on the Pattern Check Processor and pressing the enter key

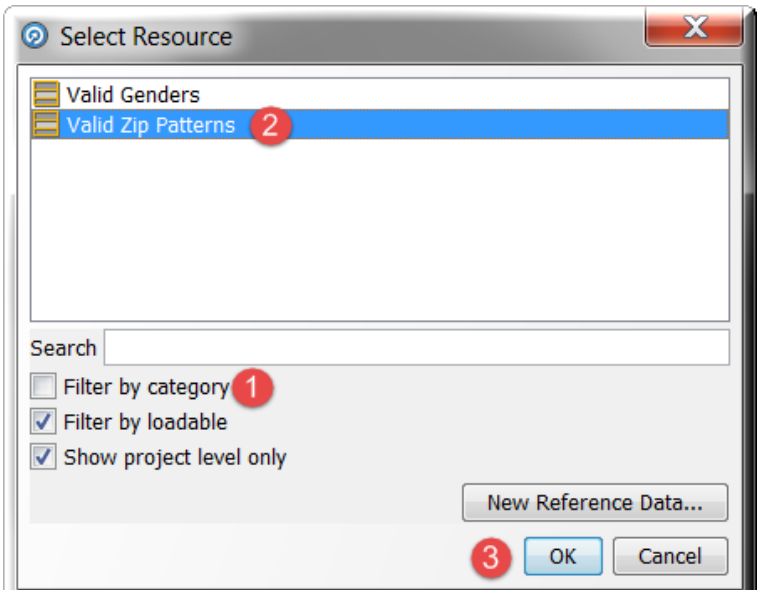


22. Connect **the All** end triangle from **Email Populated** to the **Zip Code in Right Format** processor. The configuration dialog for the **Pattern Check (Zip Code in Right Format)** processor appears. Select the **ZIP** from **Available Attributes** as the Field for validation using the  button

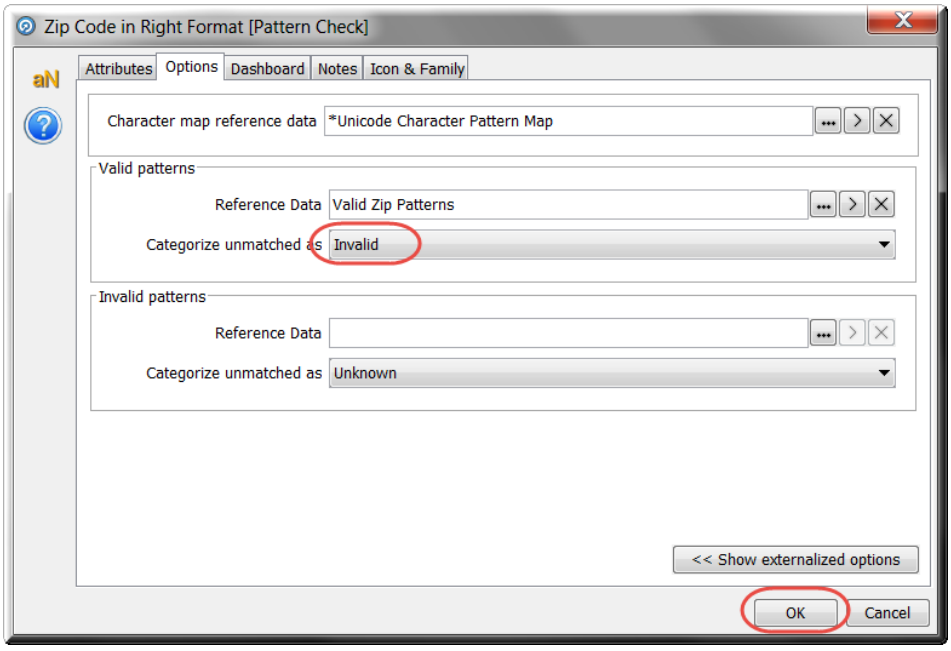
23. Click the **Options** tab at the top of the dialog box, and then click the  button in the **Valid Patterns** section in the middle of the window



24. Uncheck **Filter by Category** in the **Select Resource** Window. This is where you will select the Reference Data we created for the different types of valid Zip Codes (single source of truth for a valid Zipcode format) at the beginning of the lab. Click on **Valid Zip Patterns**, then click **OK**



25. In the section under **Valid Patterns**, click the drop-down box to change **Categorize unmatched as** to **Invalid**, then click **OK** to continue

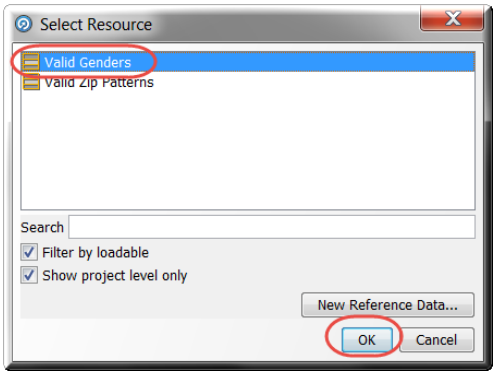


26. Click the **Run** icon in the toolbar to start the process. Click the **Zip Code in Right Format** to view the results

Notice that there are **5362 Valid Records** and **76 Invalid Records**. That is, there are 76 records that fail the 'fit for use' rule that do not match NNNNN or NNNNNpNNNN.

27. Return to the **Tool Palette** and find the **List Check** processor. Drag and drop it onto the Project Canvas and link the **All** triangle from **Zip Code in Right Format** to the **List Check** Processor

28. Select the **Gender** attribute and add to the **Field for validation**. Then click the **Options** tab in the top of the dialog box to add the Reference Data for Genders (single source of truth definition for a valid gender) created at the beginning of this lab. Click the ... icon to the right of the **Reference Data** text field in the section for **Valid Values** to browse for Reference Data and select the **Valid Genders** Reference Data, then click **OK** to continue

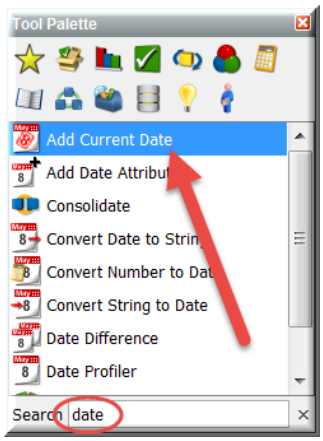


29. Click **OK** to close the **List Check** dialog box. Double-click the **List Check** processor to rename it to ***`Check for Valid Gender`***. Finally, click the **Run** button in the toolbar to start the process

> Some of this information was presented in the Profiling lab; however, these processors allow you to branch into using other types of processors to work with *All, Valid, Unknown* and/or *Invalid* results

In this next auditing example, we will check our US Customer Dataset for individuals under the age of 18 and write those records out to Staged Data. Since Age is not included in the dataset, we will calculate it using the DoB column and the current date.

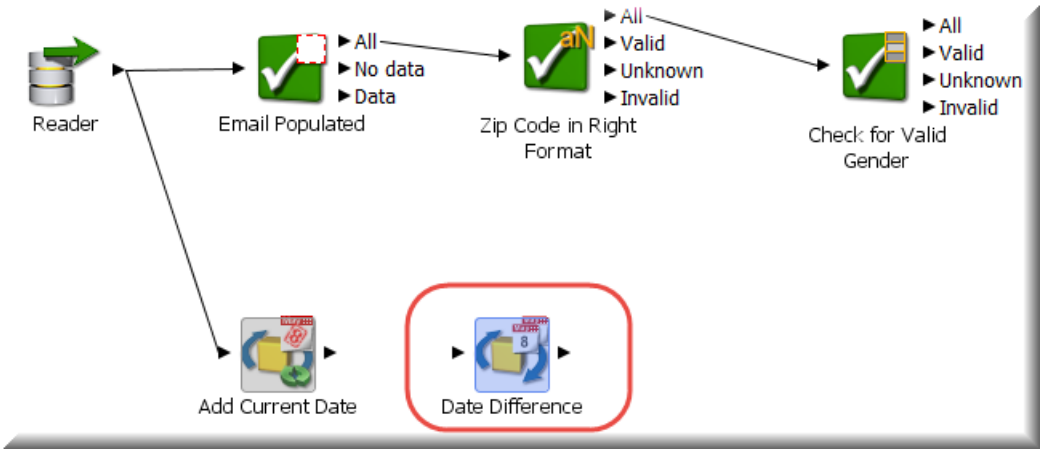
30. To begin, return to the **Tool Palette**, and search for the **Add Current Date** processor. Drag and drop it onto the Project Canvas (you can find the processor by entering ***`date`*** into the Tool Palette search text field)



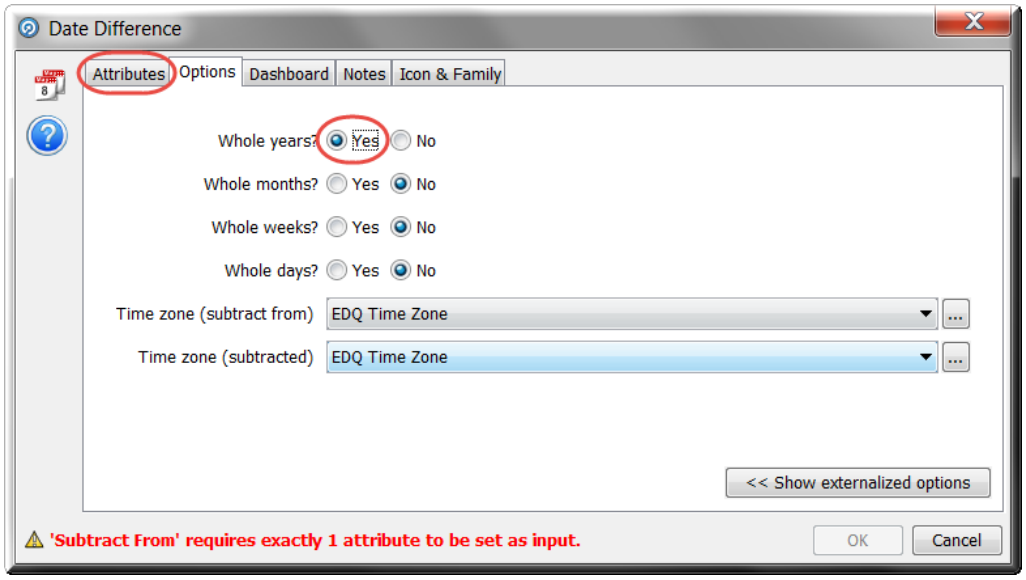
31. Connect the end triangle from the **Reader** to the input triangle of the **Add Current Date** Processor. No other configuration is needed for this processor, click **OK**

32. Return to the **Tool Palette** to find the **Date Difference** Processor and drag and drop it onto the Project Canvas. Connect the end triangle from **Add Current Date** to the input triangle of the **Date Difference** processor

The **Date Difference** and **Add Current Data** processors will be used to calculate the Age of each Customer. This is also an example of how EDQ can (and often does) add additional metadata columns for use by various downstream processors within an EDQ Process



33. Click the **Options** tab at the top of the Date Difference dialog box. Select the **Yes** radio button selection option for **Whole years?**



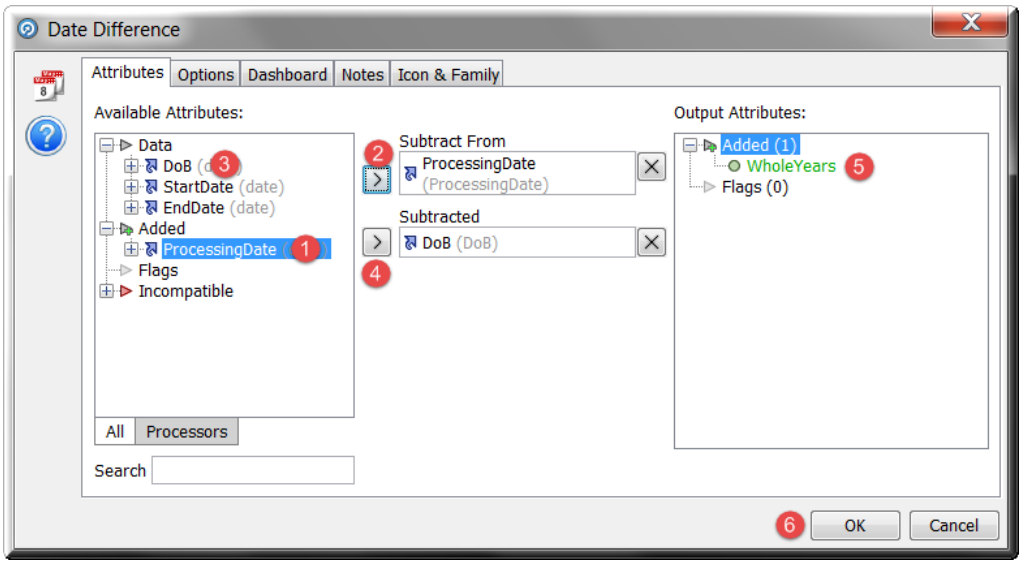
If you wanted to calculate the difference in Age (or Date) in Months, Weeks, or Days it can be done by selecting those options. Additionally, since both **DoB** and the date added with the **Add Current Date** processor happen to be in the same Time Zone, we will leave the last two drop-downs as-is.

34. Return to the **Attributes** tab by clicking it on the top-left corner of the Date Difference Dialog Box. Add the **ProcessingDate** to the **Subtract From** option using the button

35. Add the **DoB** to the **Subtracted** option using the button, then click the **Options** tab at the top of the Date Difference dialog box

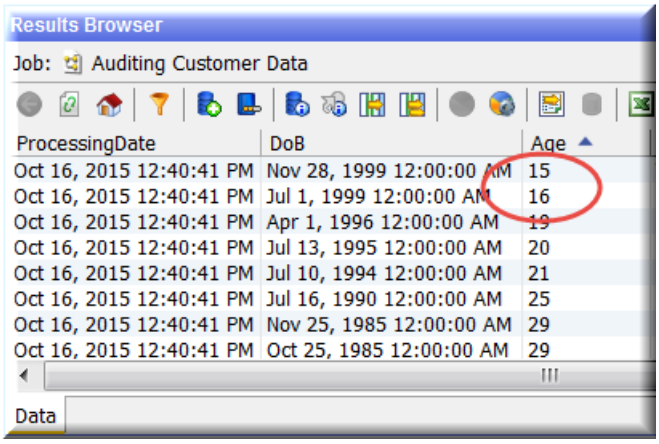
An Output Attribute has now been added to the right-side of the dialog box. An example of EDQ providing additional metadata that can aid in turning the data into 'fit for use'

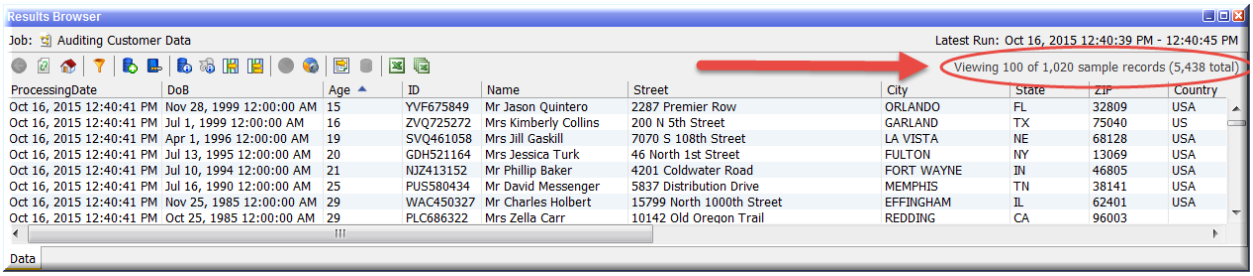
36. Double-click **WholeYears** within the **Output Attributes** on the right side of the **Date Difference** dialog box and rename it to ***Age***, click **OK** to continue



37. Click the **Run** button in the top left corner of the Project Canvas to run the process. Afterwards, sort the records in the Results Browser by clicking on the **Age** column

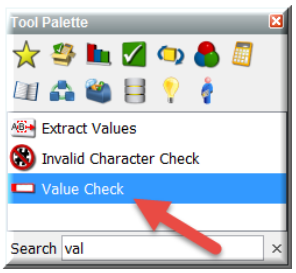
Notice that there are indeed Customers under the age of 18. At this time only **2** are visible, but if you bring your attention to the right side of the Results Browser, you will notice only 100 of 1,099 sample records are displayed of 5,438 total.



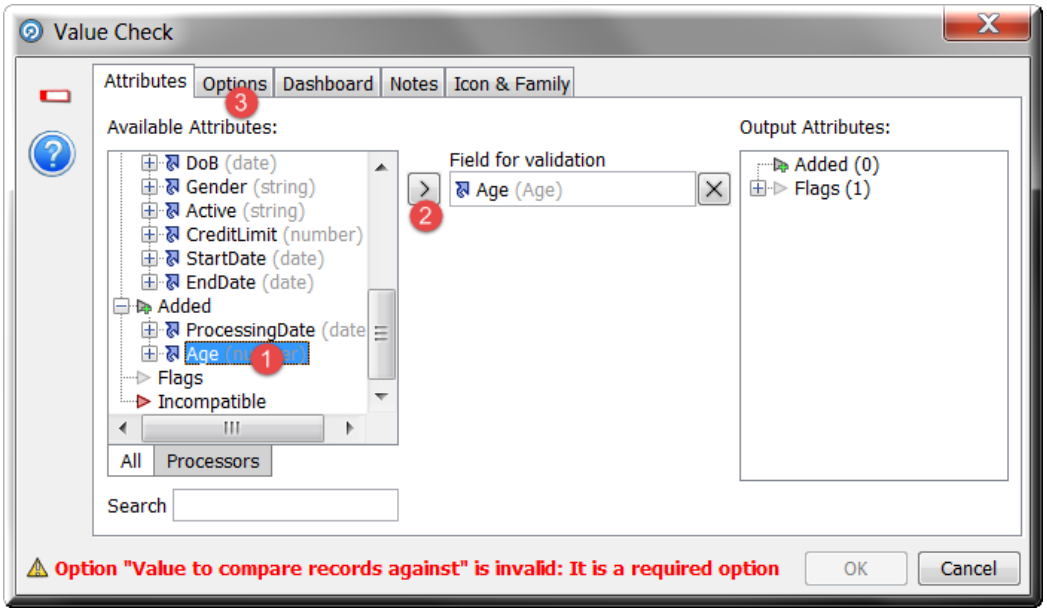


38. Return to the **Tool Palette** to find the **Value Check** processor. Drag and drop the **Value Check** processor to the Project Canvas

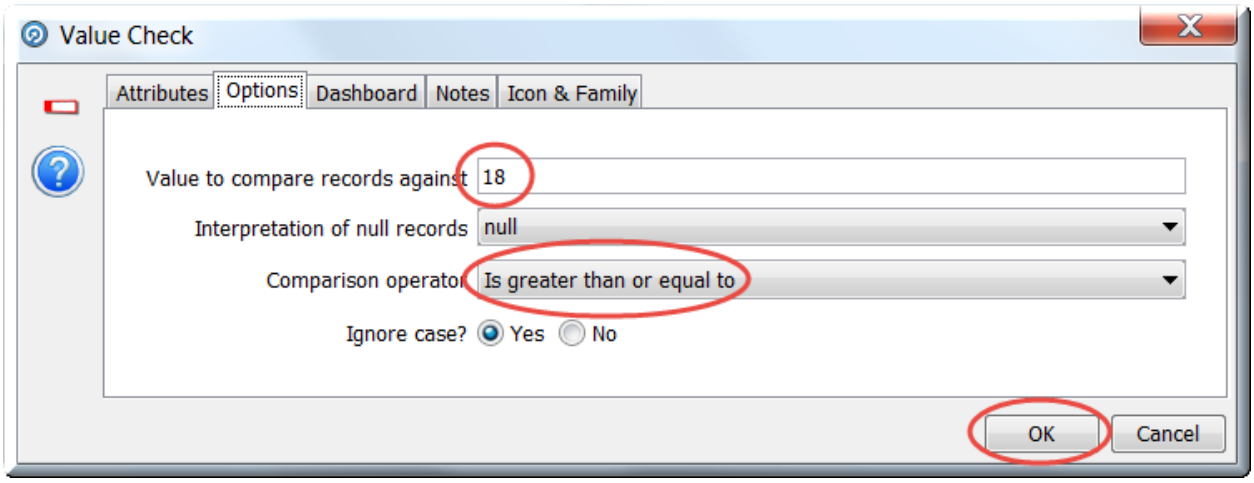
This process will be used to find values greater than or equal to 18. Afterwards, we will link the failed records from this Processor to a Writer.



39. Drag the end triangle from the **Date Difference** processor to the input triangle of the **Value Check** processor. Select the **Age** attribute as the **Field for Validation**, then click the **Options** tab on the top of the **Value Check Dialog Box** to continue

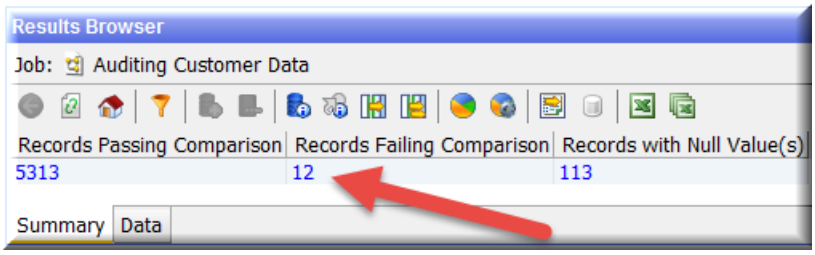


40. Enter ***18*** as the **Value to Compare Records Against**. Change the Comparison operator to **Is greater than or equal to**, then click **OK** to continue



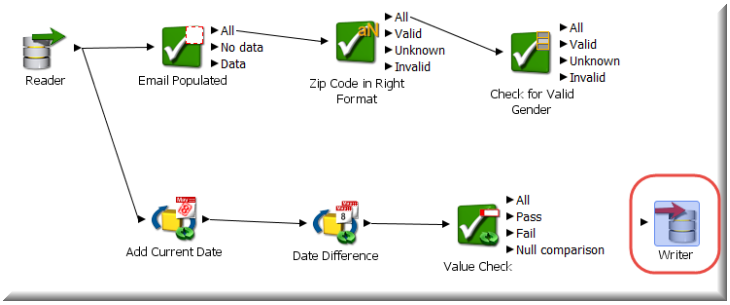
41. Click **Run** in the top left corner of the Project Canvas to run the process

Notice that there are **12** customers that are under the age of 18, whereas **5313** fit the criteria of being 18 or older.



Earlier in the lab we saw how we can export records from the Results Browser to an Excel file. We can also drag one or more **Writer** Processors to the Project Canvas to write results back to a data store.

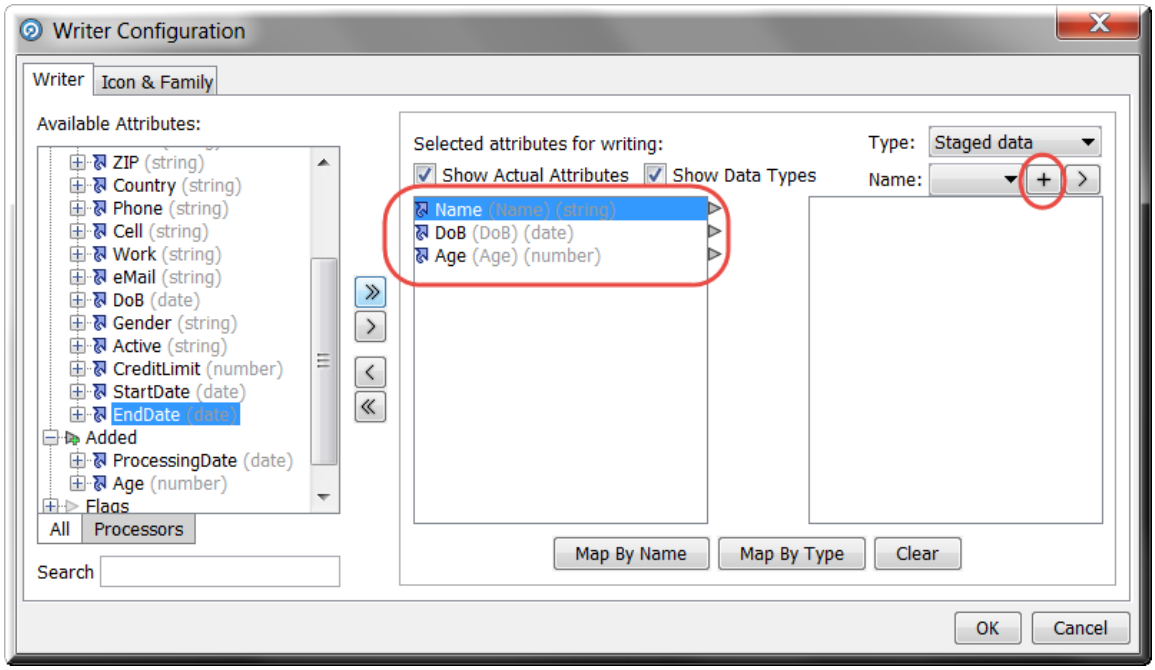
42. Return to the **Tool Palette** to search for the **Writer** and drag and drop it to the Project Canvas



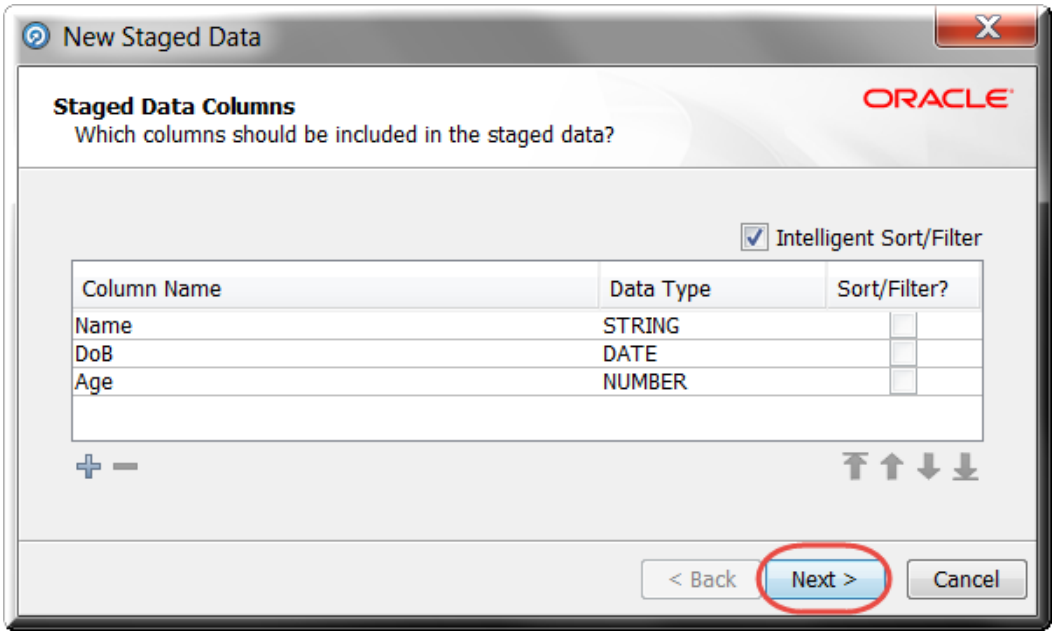
43. Connect the **Fail** end triangle of the **Value Check** Processor to the input triangle of the **Writer**

This will allow us to export those Customers that are under 18 for use in other scenarios.

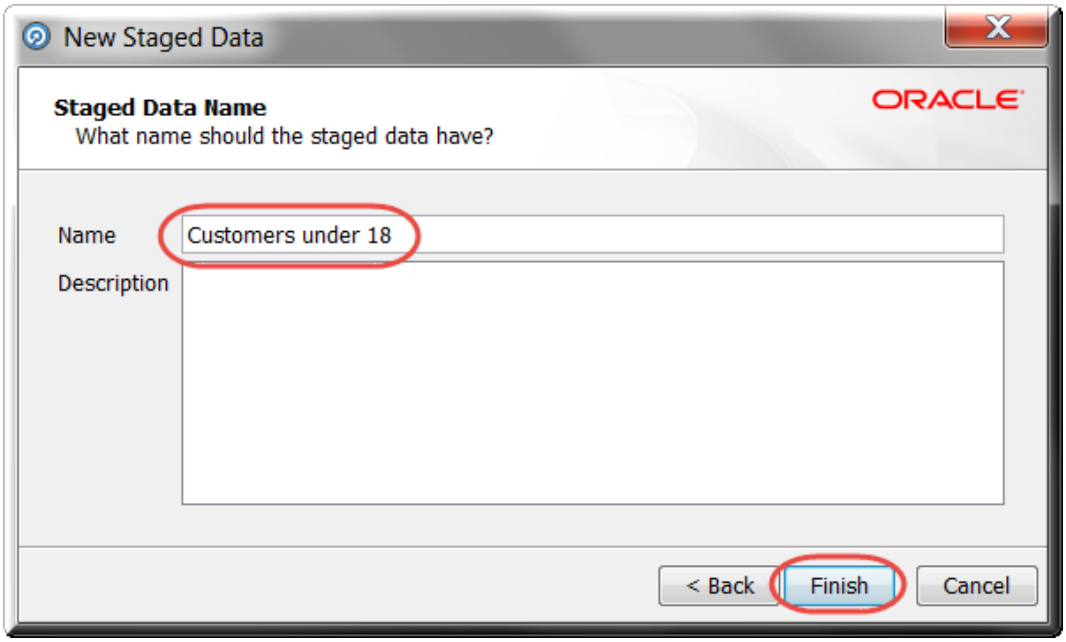
44. Multi-select the **Name**, **DoB,** and **Age** Attributes from the **Available Attributes** and move them to the **Selected attributes for writing** list area. Then click the **+** underneath the **Type: Staged data** dropdown to create a new **Staged data** object since one does not exist yet



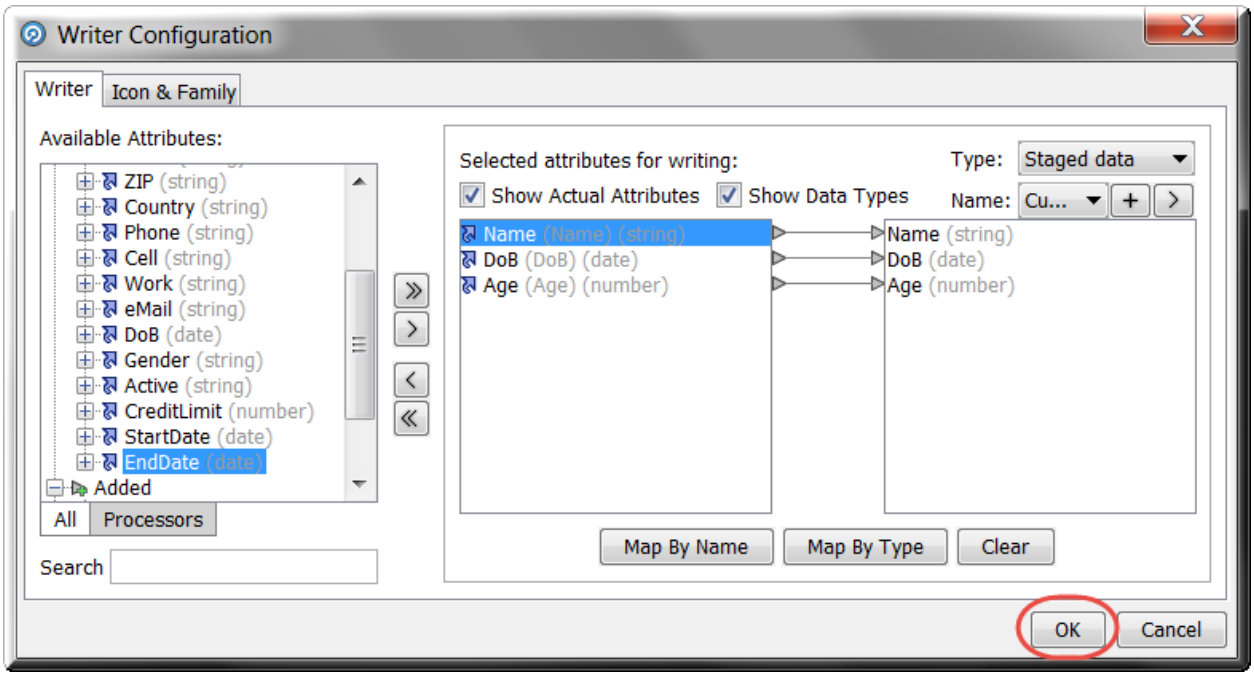
45. The first dialog box when creating **New Staged Data** will prompt you to select the columns to include. In this case we will keep the defaults, click **Next >** to continue



46. Give this Staged Data set a name: ***`Customers Under 18`***, then click **Finish**

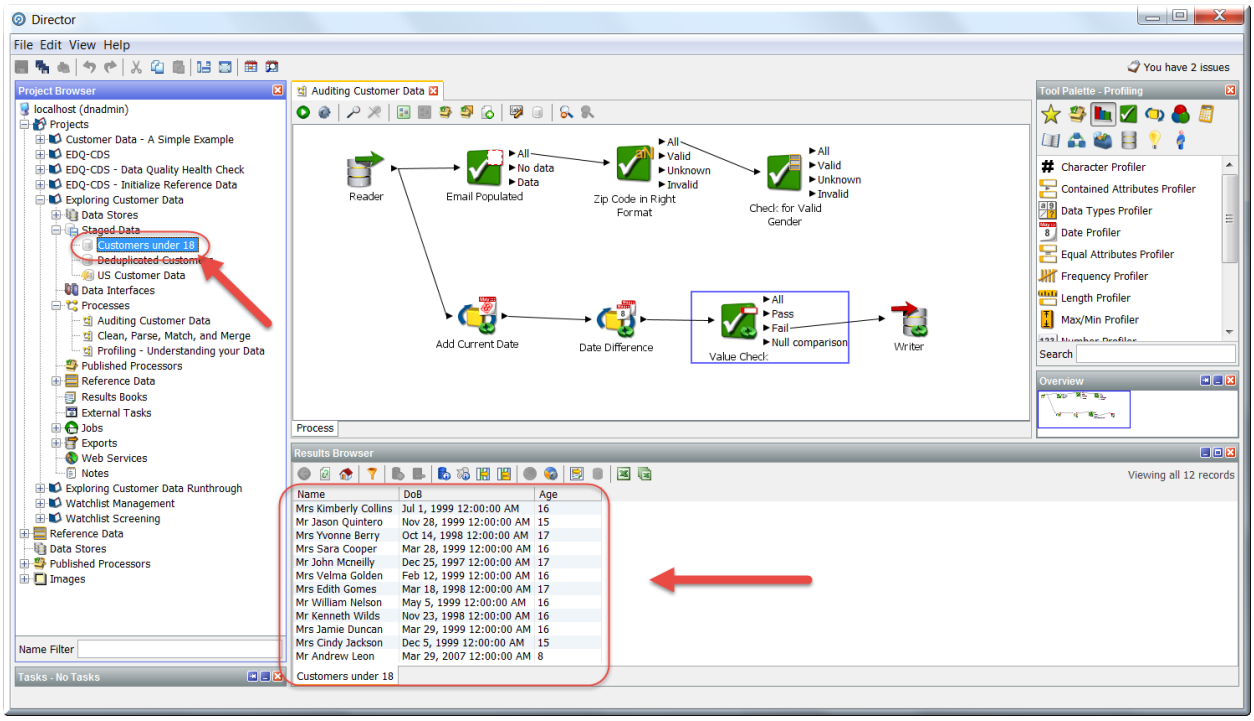


47. Afterwards, you will notice the **Name**, **DoB**, and **Age** are attributes mapped to the **Staged Data** set you just created. Click **OK** to finish configuring the Writer



48. Click **Run** in the top left corner of the Project Canvas to run the process

You will notice the focus of the Director shifts back to the **Project Browser** and your new Staged Data, **Customers under 18** is selected. The **Results Browser** also has changed to display the **6 Customers** who failed the Value Check for Greater than or is equal to 18.

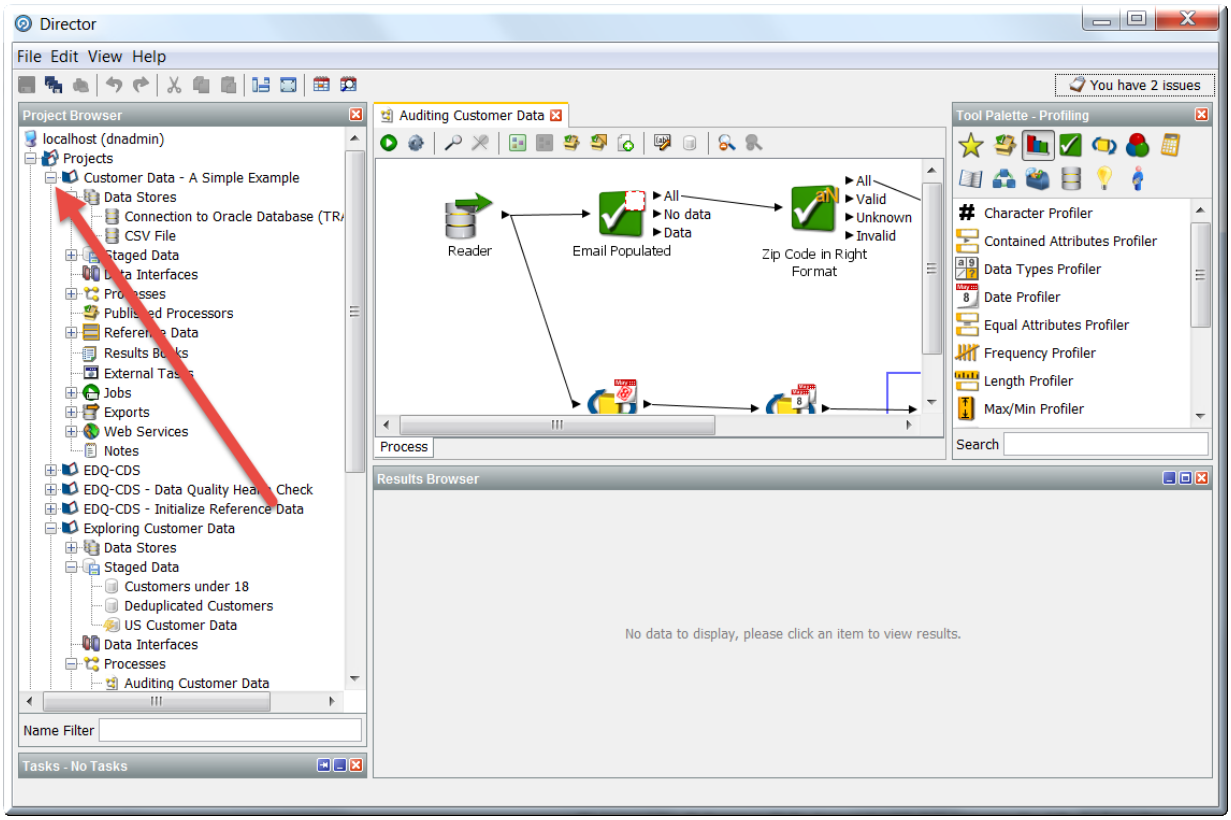


In the next Lab, we will be able to use what we learned about our data during the Profile and Auditing stages to Clean, Enrich, Parse, Transform, Match and Merge information. Before that, we need to prepare two additional sets of Reference Data to be able to standardize the problem found with the Country column.

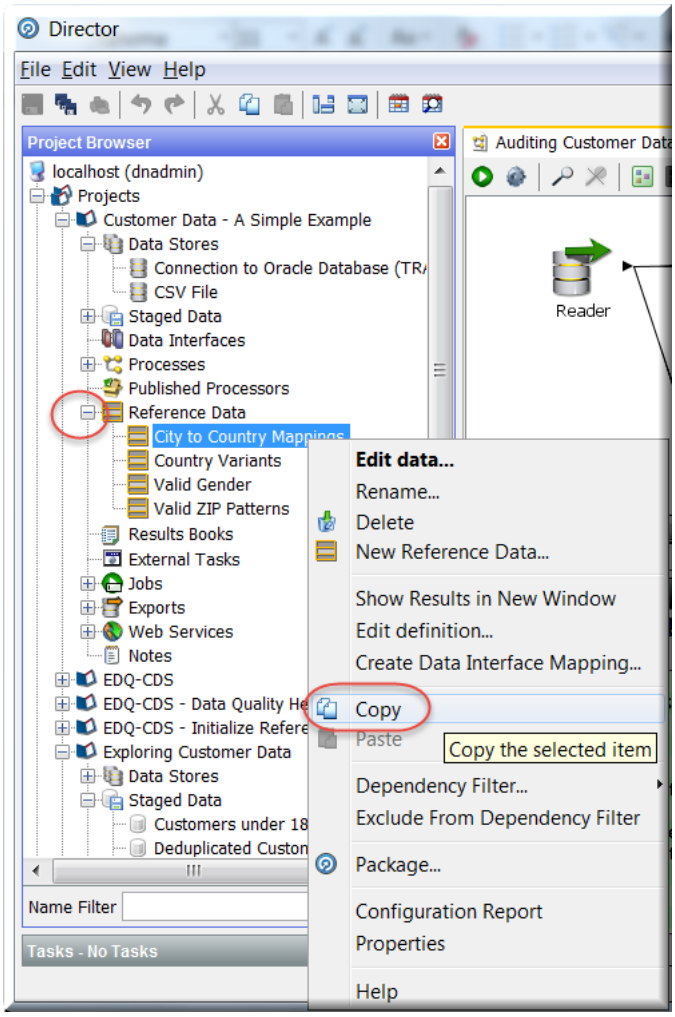
We have already looked at creating issues to create a collaborative environment, and we will take a deeper dive into that in Lab 5. There also are other ways of making EDQ a collaborative tool. For instance, Reference Data and other items within the Project Browser can easily be shared by simply copying and pasting it from one project to another.

In order to save time, we will re-use Reference Data created during another project.

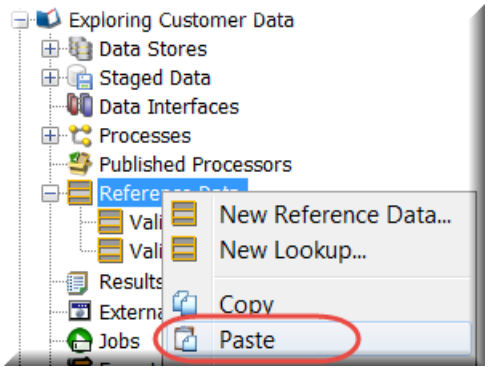
49. Navigate to the Project Browser and find the Project titled **Customer Data – A Simple Example**



50. Expand the **Reference Data** item and right-click to **Copy** the **City to Country Mappings**



51. Return to your **Project - Exploring Customer Data** and find the **Reference Data** item, right-click and **Paste** the contents of the clipboard



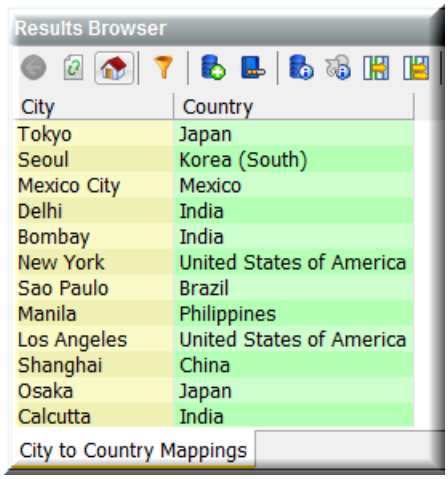
52. Navigate back to the **Project Browser** and find the Project titled **Customer Data – A Simple Example**

53. Expand the **Reference Data** item if it is not already and right-click and **Copy** the **Country Variants** Reference Data

54. Return to your **Project - Exploring Customer Data** and right-click on **Reference Data** and **Paste** the contents of the clipboard

55. In the **Reference Data** section of your Project, **Exploring Customer Data**, click on **City to Country Mappings**

Notice in the **Results Browser** that this is a Two Column Reference Data set. The **City** (yellow) column signifies the Lookup Column, similar to the two sets of reference data we created in previous steps for Gender and Zip Code. The **Country** (green) column signifies the Return Column, which we have not yet explored. In this case, wherever the Lookup Column value contains data, the **City to Country Mappings** Reference Data set can be used to return the Country that City is in. This will help us standardize the values found in the Country column in subsequent steps.



56. Click the **Country Variants** Reference Data within the Project, **Exploring Customer Data** in the Project Browser

Notice that this is also a Two Column Reference Data set with **Country Variants (USA, US, U.S.A, etc.)** as the Lookup Column and a **Standardized value (United States of America)** as the Return Column. If you flip back to the page where we created the Pie Chart and Bar Graphs for the Country Profiling exercise, you will remember that we have several representations of the same Logical country within the data set. This Reference Data will allow us to standardize the multiple representations of a single Country into a single standardized representation of Country – dramatically increasing the 'fitness of use' for the data when done for all relevant columns. The primary beneficiary: Analytics dashboard authors and consumers of dashboard Analytics. 'Not just Analytics, but Accurate Analytics'.

 This is a unique copy of Reference Data. If the Reference Data is modified in your project, it will stay intact for the other project. If you happen to find an additional variant of a country within the Results Browser, right-click the Reference Data to edit it and add the necessary rows. This allows you to modify the process and/or processors that utilize the Reference Data without making drastic changes that could be time consuming.

Lab 3: Creating and Automating Data Standardization and Improvement Solutions

In this lab we will perform the following using EDQ Processors:

- Standardize
- Transform
- Parse
- Enrich

All of these steps are 'critical to success' for a data centric initiative by creating "Data fit for use" and "Accurate Analytics".

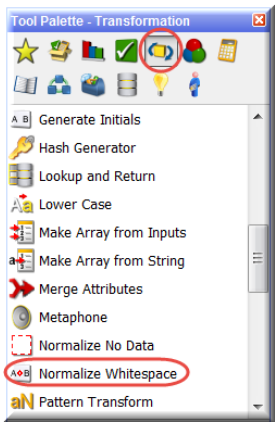
Create Standardization Process

1. Create a New Process under your **Project (Exploring Customer Data)** in the **Project Browser** by right-clicking on **Processes** and clicking **New Process...**

2. Select the **US Customer Data** and click **Next >**. Click **Next >** without adding any Profiling

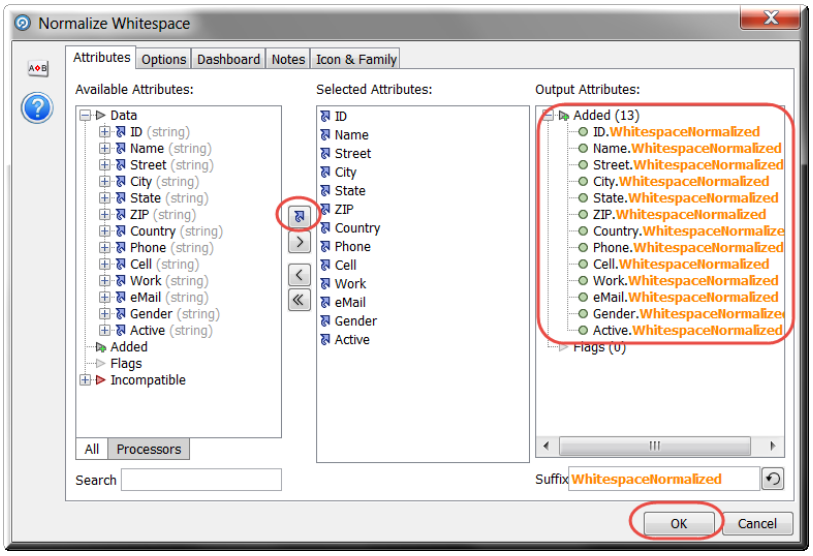
3. Name your process ***`Clean, Parse, Match, and Merge`*** then click **Finish**

4. Return to the **Tool Palette** on the right side of the **Director** and find the **Normalize Whitespace** processor by clicking on the **Transformation** icon



5. Connect the **Reader** to the **Normalize Whitespace** Processor

6. The **Normalize Whitespace Dialog Box** appears. Click the **Select All** icon and click **OK**. Next click the **Run** icon in the toolbar to start the Process

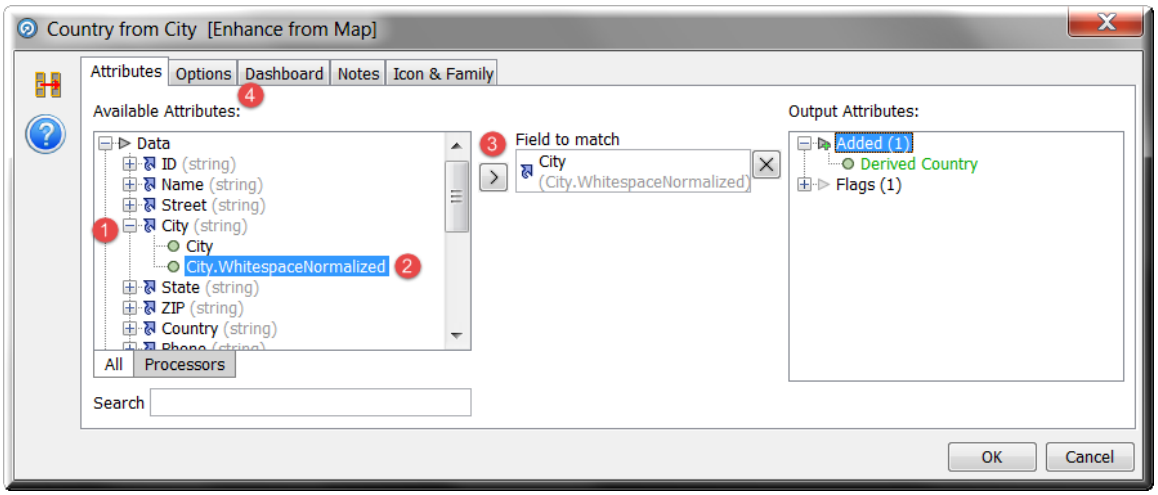


Next we will begin the process to Standardize the Country column.

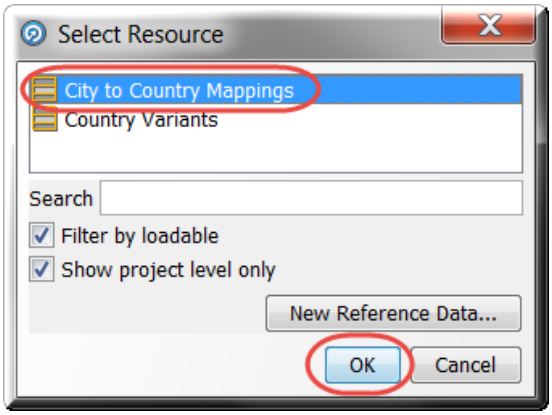
7. Find the **Enhance from Map** Processor from the **Tool Palette** by using the **Search** textfield: type ***Enhance from Map***. Drag and drop the Processor to the Process canvas. Right click the **Enhance from Map** and select Rename to rename it to ***Country from City***

8. Connect the end triangle from the **Normalize Whitespace** Processor to the **Country from City** Processor

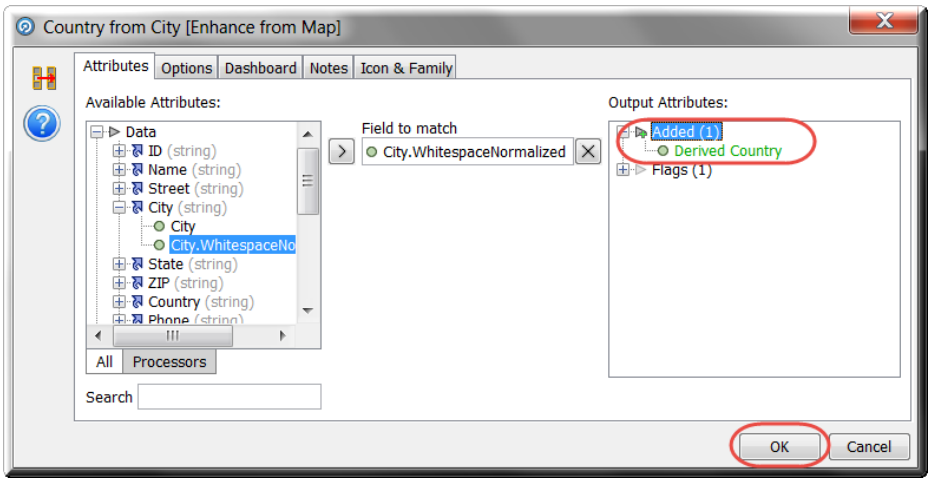
9. In the **Country from City Dialog Box**, expand the **City** field from the **Available Attributes** to observe the extra metadata value(s) for selection created by the Normalize Whitespace Processor. Double-click on **City** to add **City** to the **Field to Match**. Next, click the **Options** tab at the top of the dialog box



10. Click the Browse button within the **Value Map** area to connect the Reference Data we copied earlier. Select **City to Country Mappings**, click **OK** to continue



11. Click **Attributes** on the top left corner of the dialog box to adjust the name of the **Output Attribute: EnhancedResult** by double-clicking to ***Derived Country***, then click **OK**

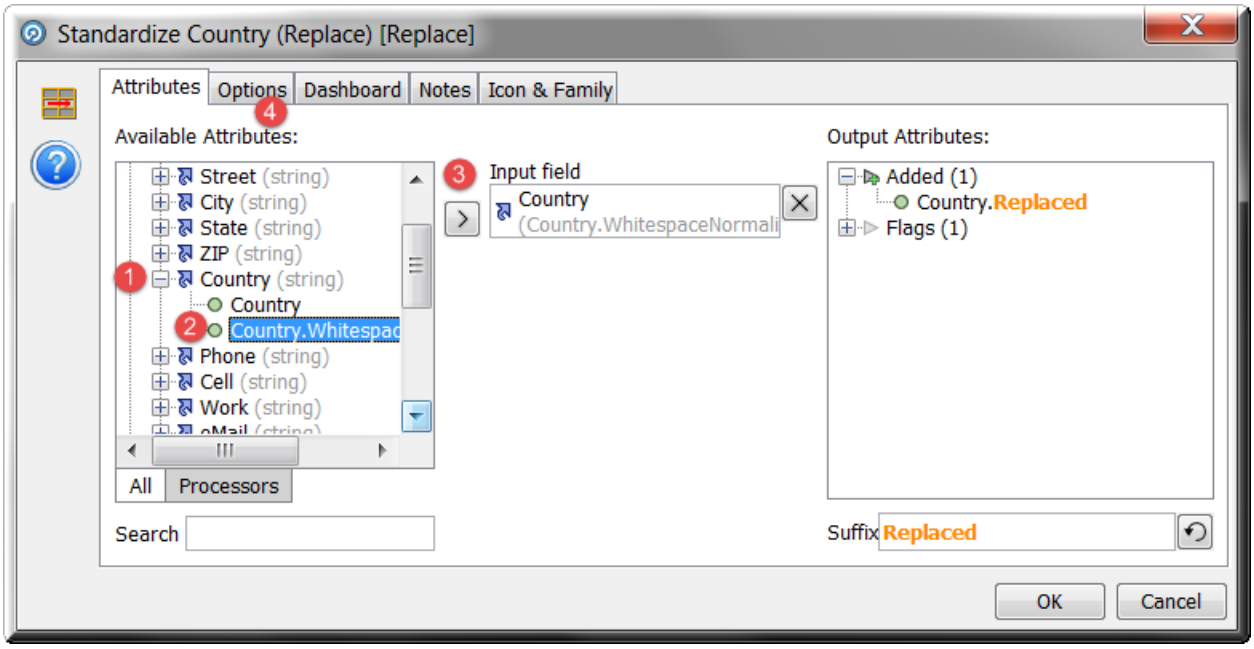


12. Click **Run** to start the **Country from City** process

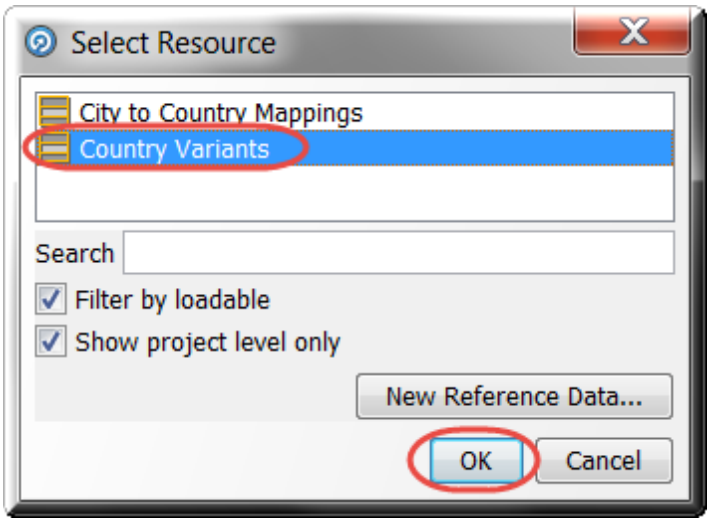
You will notice several unenhanced results, but we are not done yet! Feel free to drill down on the Enhanced and Unenhanced values to glance at what this processor did.

13. Return to the **Tool Palette** to search for the **Replace** Processor. Drag and Drop the **Replace** Processor to the Project Canvas. Right click the **Replace** Processor and select **Rename** to rename it ***`Standardize Country`***

14. The **Standardize Country** configuration dialog appears. Connect the **All** end triangle from the **Country from City** Process. Double-click the **Country** field and add **Country.WhitespaceNormalized** to the Input Field. Then click the **Options** tab at the top of the dialog box to setup the Replacements



15. Click the Browse button next to **Replacements** to add the **Country Variants** Reference Data, click **OK**

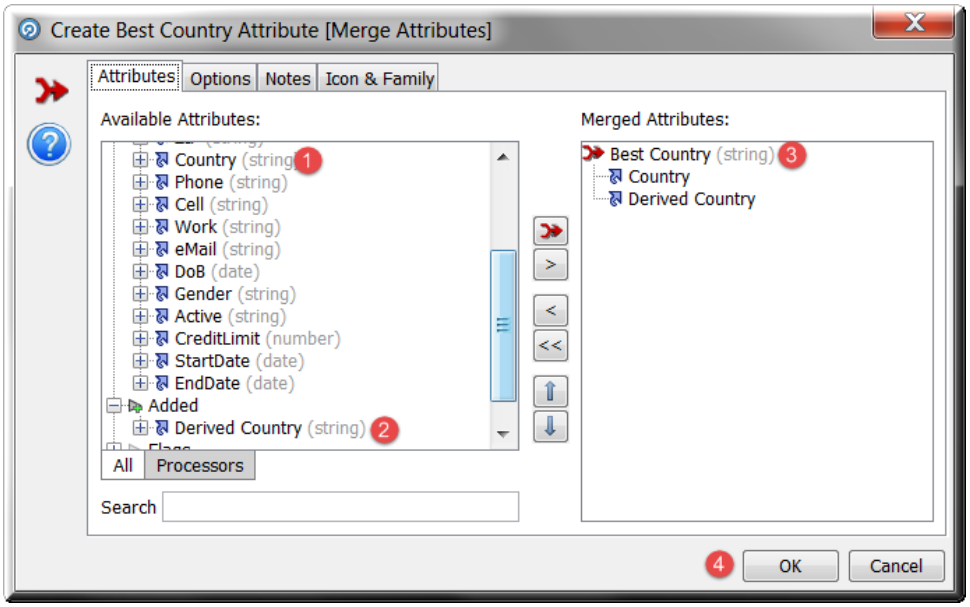


16. Click the **Run** button in the toolbar to start the process

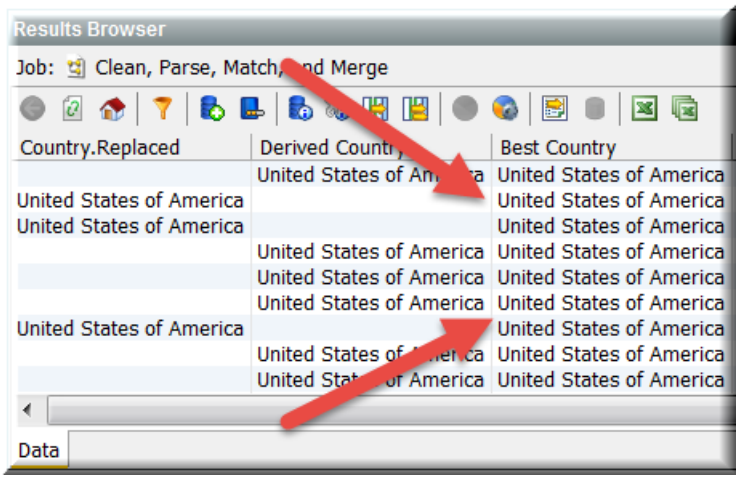
17. Return to the **Tool Palette** and search for **Merge**. Drag and drop the **Merge Attributes** Processor onto the Project Canvas and rename it to ***`Create Best Country Attribute`***

We will use this to combine the Country and Derived Country columns to a single attribute that can be used further down the Process, hence, getting closer to the 'single source of truth'.

18. Connect the **All** end triangle from **Standardize Country** to the **Create Best Country Attribute** Processor. Double-click on **Country** and **Derived Country** from the Available Attributes to add to the **Merged Attributes**. Lastly, rename the Merged Attribute to **`Best Country`**, then click **OK** to continue

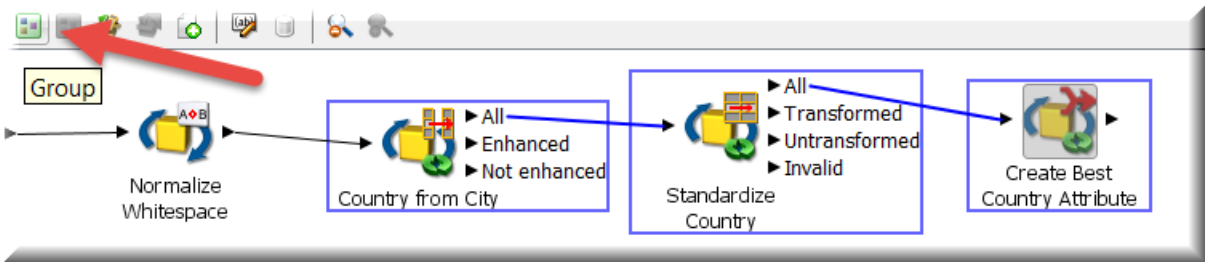


19. Click the **Run** button in the top left corner of the Project Canvas and then click the **Create Best Country Attribute** Processor to view the results in the **Results Browser**



Notice how the two columns, **Country.Replaced** and **Derived Country** have now been merged to a single column, **Best Country**.

20. It is also possible to group several processors together by highlighting the desired processors to group and clicking the **Group Button** in the Toolbar, click-and-drag a box around the **Country from City**, **Standardize Country**, and **Create Best Country Attribute** processors. Then click the **Group Button** in the toolbar



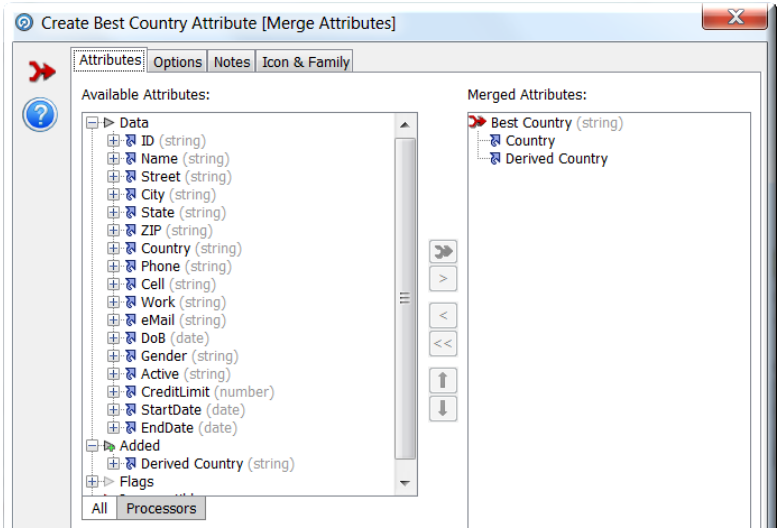
21. Afterwards, you can click on **Group** in the green box it forms to re-name the grouping. In this case, rename it ***`Standardize/Add Country`***

Grouped Processors can be published as Processors so that other EDQ users can use the Processor in their projects. We will not cover this capability in this lab guide.

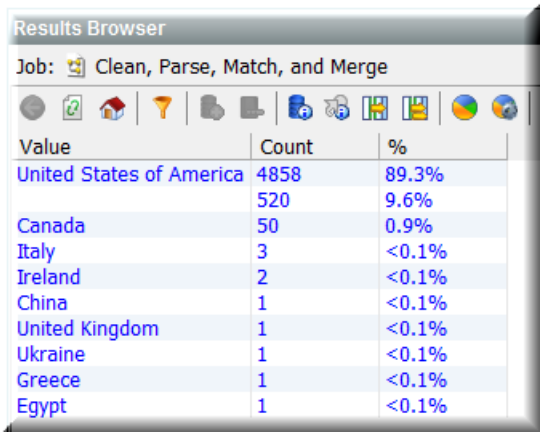
Before we continue, let's add a Frequency Profiler to the end of the **Standardize/Add Country** grouping to check on the progress of making this data set fit for use (by standardizing the data)

22. Return to the **Tool Palette** and enter ***`Frequency Profiler`*** in the **Search** textfield. Drag and drop the **Frequency Profiler** to the Project canvas and connect the end triangle from **Create Best Country Attribute** to the **Frequency Profiler** and the configuration dialog box appears

23. Double click the **Best Country** attribute to add it to the **Selected Attributes** column, and click **OK**.



24. Re-name the Frequency Profiler by right clicking your mouse, select **Rename** and enter ***`Frequency Profile Countries`***. Click the **Run** button in the toolbar to start the process



Observe the uniform values for the Countries. We have standardized five (5) different representations of United States (USA, US, U.S.A, United States and U.S) into one and only one Standardized representation: **United States**. Without standardization, one of two things is guaranteed to happen:

- Undercounting the sales in the United States when reported in Dashboard Analytics. The Dashboard author will pick one of the Country values. But what about the U.S and U.S.A values? What Country will those sales get reported on?

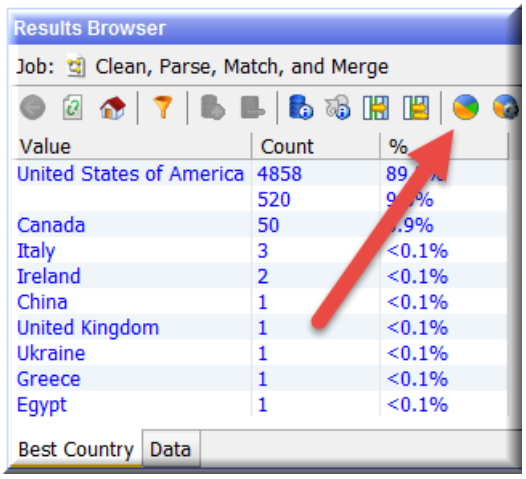
- The Dashboard author must know in advance all non standardized values of Countries exist that exist and handle all the anomalies themselves. How likely is that going to happen for all anomalies in each and every Attribute in the data? Even if it did – how much longer would it take to develop and test these dashboards versus Dashboards written against Standardized data ? Think productivity ! Think business value of inaccurate Dashboard results?

Also note that there are still 520 or 9.6% of the Country values that still need attention.

25. Right-click on **520** and **Create Issue...**

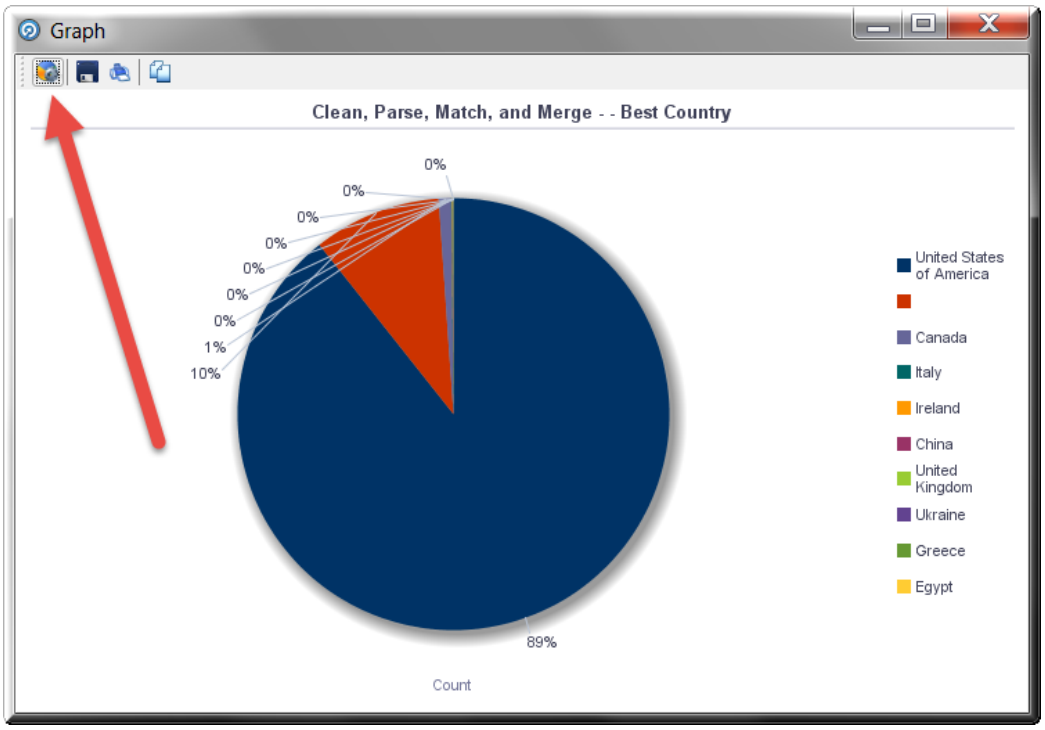
26. Type a description in the **Issue Description** Text Area for the issue so that we can come back to it during the next Lab exercise – ***520 entries still exist after standardization/add country process***, select a user for **Assigned To** and click **OK** to close the issue dialog

27. Click the **Graph Results** icon in the toolbar of the **Results Browser**

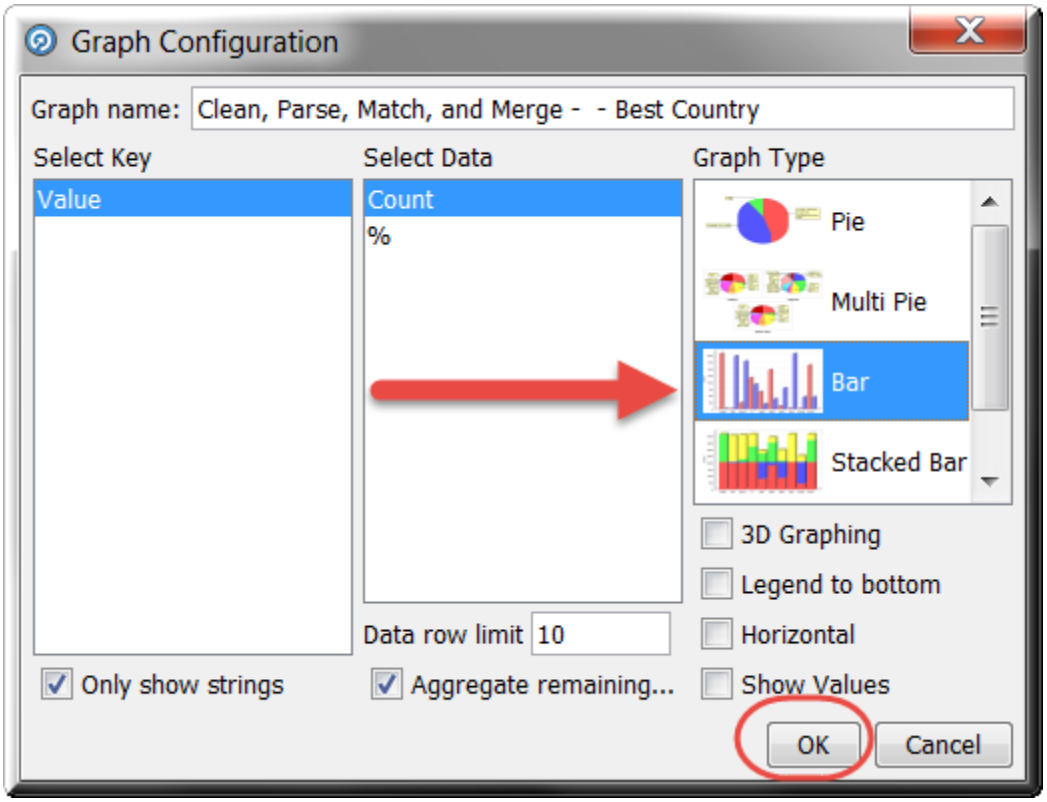


Notice how this chart looks drastically different from the one we created in Lab 1b: Profiling. The various representations of USA have now all been standardized to United States of America

28. Click the **Graph Configuration Button** in the toolbar of the chart

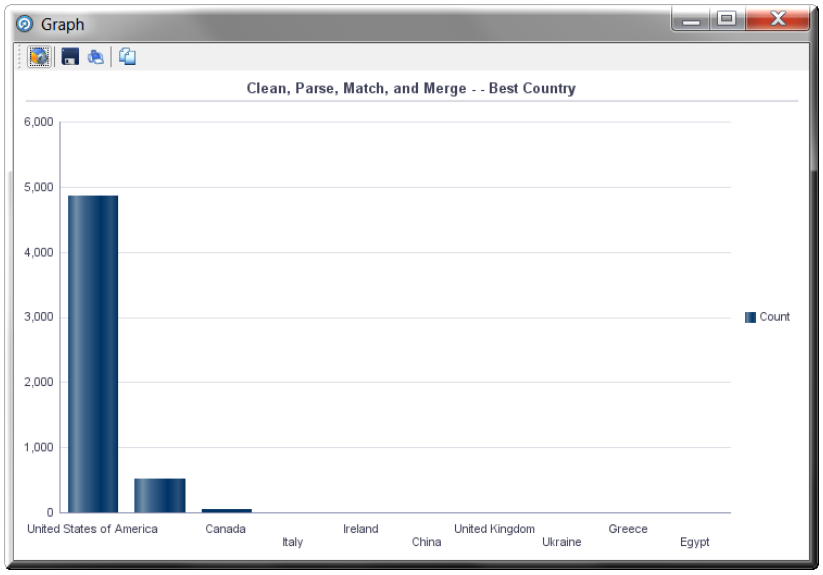


29. Change the **Graph Type** to **Bar**, click **OK** to continue

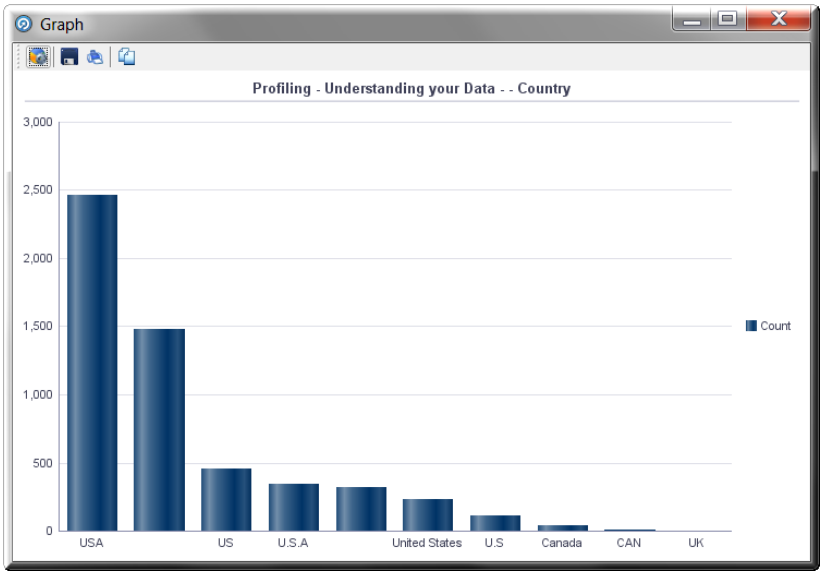


Observe the difference again of the following chart compared to the previously created chart prior to Standardization. We can absolutely say the Country column has much higher fitness for use for downstream data initiatives consuming and/or reporting on the data. Additionally, other countries that we did not see before have been added.

After Standardization



Before Standardization

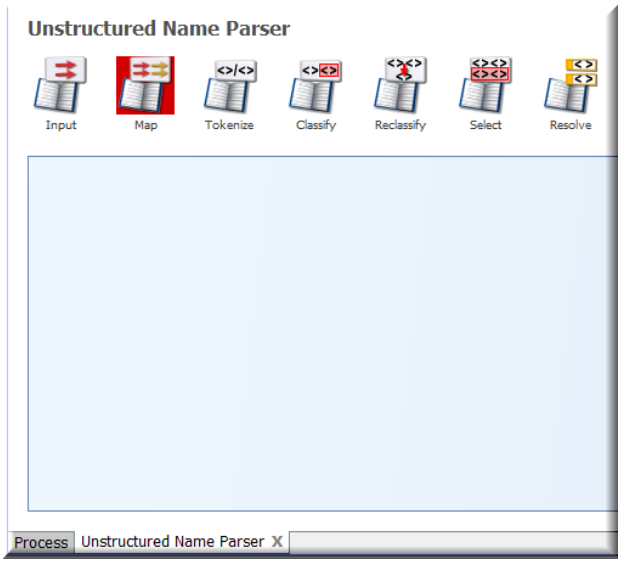


Next we will begin to examine and apply structure to the Name field.

30. Navigate to the **Tool Palette** and search by typing ***`Unstructured Name Parser`***. Drag and Drop the **Unstructured Name Parser** to the Project Canvas and connect the **Normalize Whitespace** endpoint to the **Unstructured Name Parser**

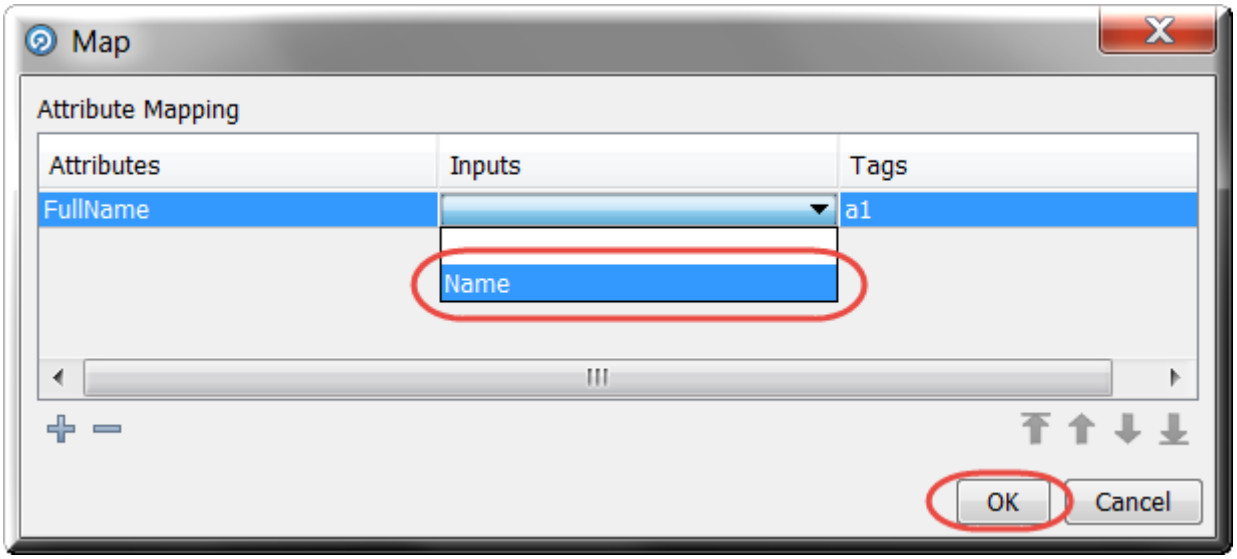
31. The **Unstructured Name Parser Dialog Box** appears, select **Name** from the **Available Attributes** and add it to the **Selected Attributes**, click **OK** to continue

    The Parse processor consists of several configuration screens. This processor is preconfigured to work with Names, so we will not have to worry about configuring every sub-processor. Also, you will notice the Project Canvas has multiple tabs on the bottom left corner (above the **Results Browser**). You can switch back and forth as needed between the Project Canvas and the **Unstructured Name Parser** configuration.

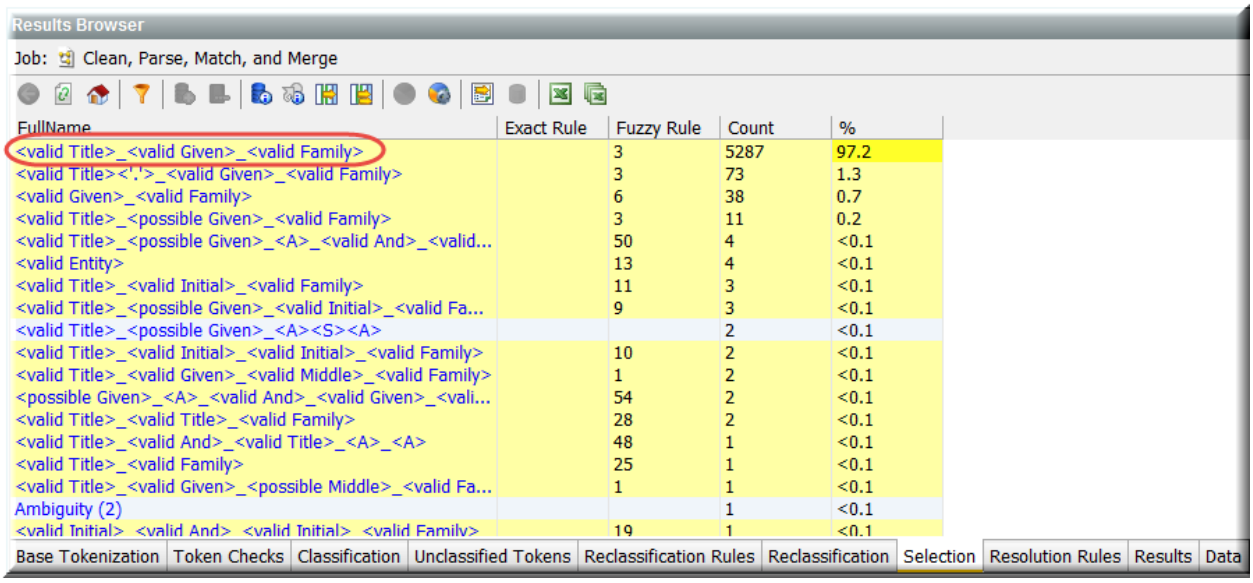


At the top of the Project Canvas you will notice several sub-processors**: Input, Map, Tokenize, Classify, Reclassify, Select, and Resolve**. Any sub-processor that is **shaded in Red** needs to be configured.

32. Double-click on **Map** to configure it. Since we selected one attribute for this processor, it only has a single input – **FullName**. Select the **Name** attribute under the **Input** dropdown, and click **OK** to continue

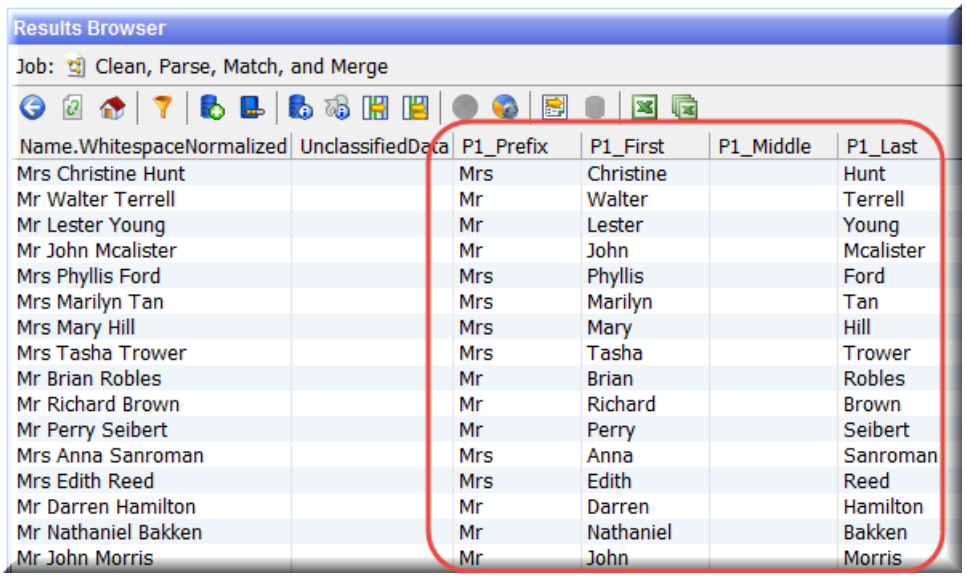


33. Click the **Run** button in the toolbar to run the process. After a few moments, the processor will show the selected token patterns found in the Name field within the **Selection** tab within the Results Browser



Notice that over 97% of the records have the pattern <valid Title>_<valid Given>_<valid Family>.

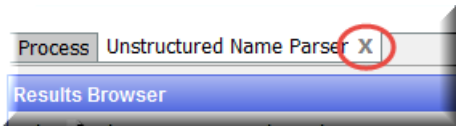
34. Drill down on the pattern by clicking **<valid Title>_<valid Given>_<valid Family>** to see how this parser split up the **Name** attribute



Name.WhitespaceNormalized	UnclassifiedData	P1_Prefix	P1_First	P1_Middle	P1_Last
Mrs Christine Hunt		Mrs	Christine		Hunt
Mr Walter Terrell		Mr	Walter		Terrell
Mr Lester Young		Mr	Lester		Young
Mr John Mcalister		Mr	John		Mcalister
Mrs Phyllis Ford		Mrs	Phyllis		Ford
Mrs Marilyn Tan		Mrs	Marilyn		Tan
Mrs Mary Hill		Mrs	Mary		Hill
Mrs Tasha Trower		Mrs	Tasha		Trower
Mr Brian Robles		Mr	Brian		Robles
Mr Richard Brown		Mr	Richard		Brown
Mr Perry Seibert		Mr	Perry		Seibert
Mrs Anna Sanroman		Mrs	Anna		Sanroman
Mrs Edith Reed		Mrs	Edith		Reed
Mr Darren Hamilton		Mr	Darren		Hamilton
Mr Nathaniel Bakken		Mr	Nathaniel		Bakken
Mr John Morris		Mr	John		Morris

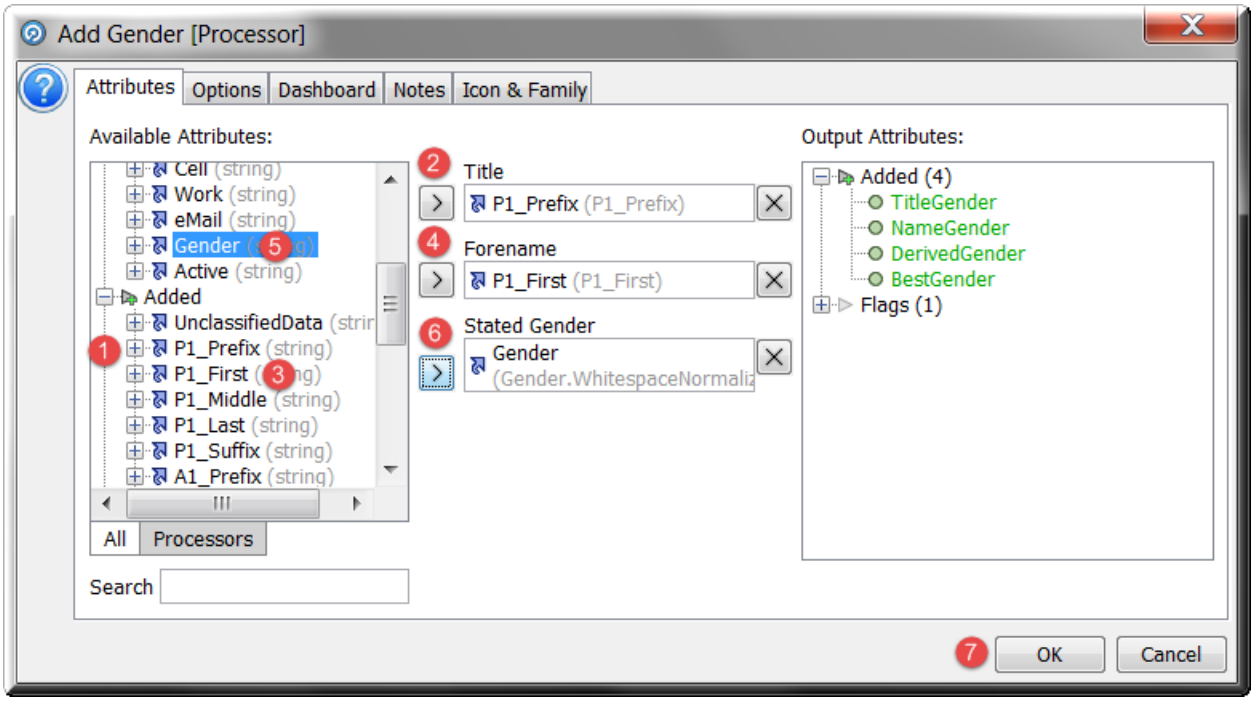
We can now use these additional attributes (P1_Prefix, P1_First, etc.) further down in this process. For instance, we can derive the Gender based on the P1_Prefix column to further improve the quality of our Gender data as noticed in Lab 1.

35. Click the **X** on the **Unstructured Name Parser** in the lower left corner of the Project Canvas to return to the **Process Canvas**

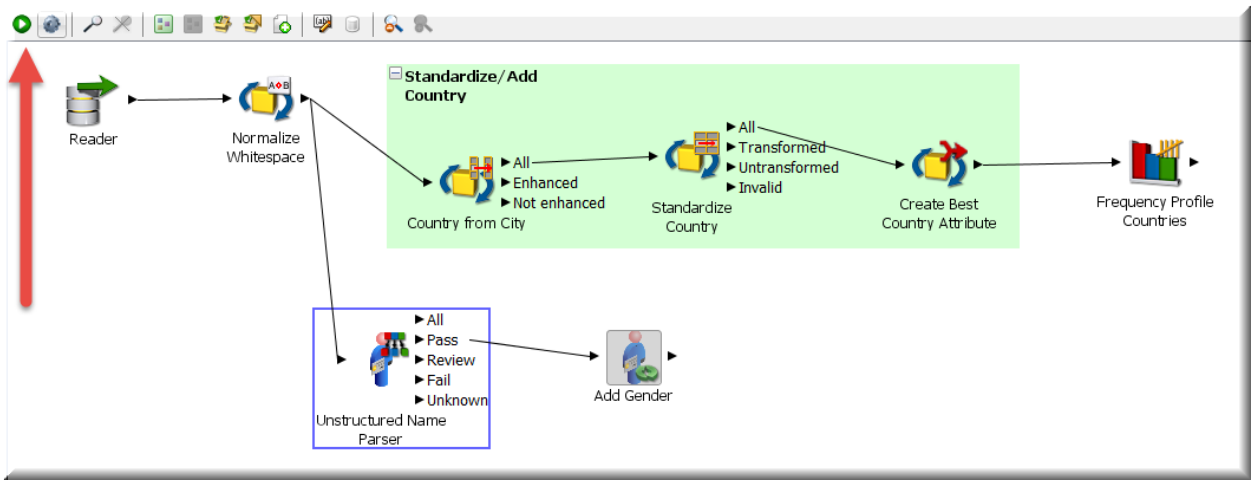


36. Return to the **Tool Palette** and search by typing ***Add Gender***. Drag and Drop the **Add Gender** processor to the Project Canvas and connect the **Pass** output triangle from the **Unstructured Name Parser** to the input of the **Add Gender** processor

37. Select **P1_Prefix** and press the **>** icon to move it to **Title.** Select **P1_First** and move to **Forename**. Select **Gender** and move it to **Stated Gender**, then click **OK** to continue

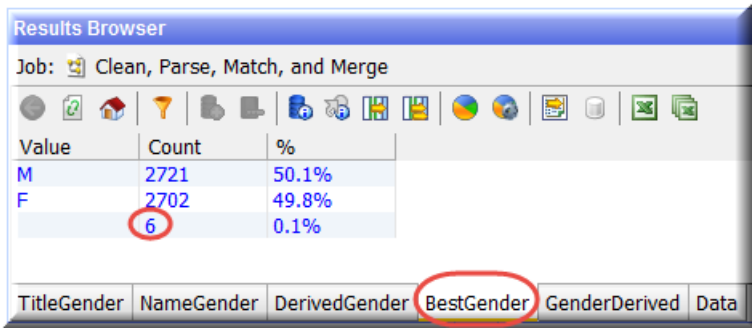


38. Click the **Run** icon in the toolbar to start the process

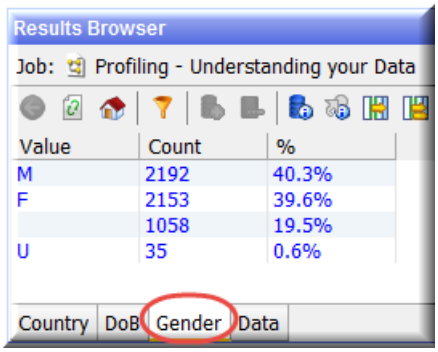


39. In the **Results Browser**, click the **BestGender** tab. This shows the successful and unsuccessful count and percentage of records derived from the Title and name. Click **6** below the **Count** column to drill-down

After Adding Gender (Enriched / Standardized Gender):

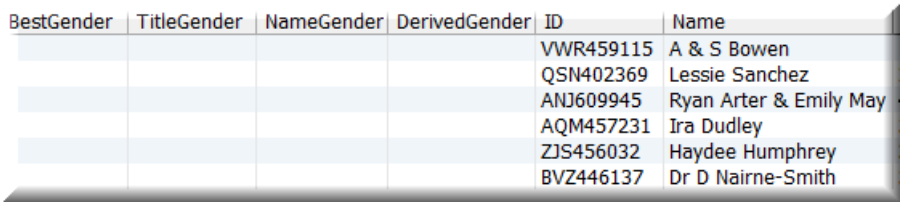


Before Adding Gender (Screenshot from Lab 1b)…Data un-fit for use



We have automated enrichment of the Gender attribute (and subsequent 'Fit for Use') dramatically reducing the percentage of missing **Gender** from **19.5%** to less than **0.1%.** Imagine the before/after increase in effectiveness of Marketing initiatives (or Analytics trending analysis)

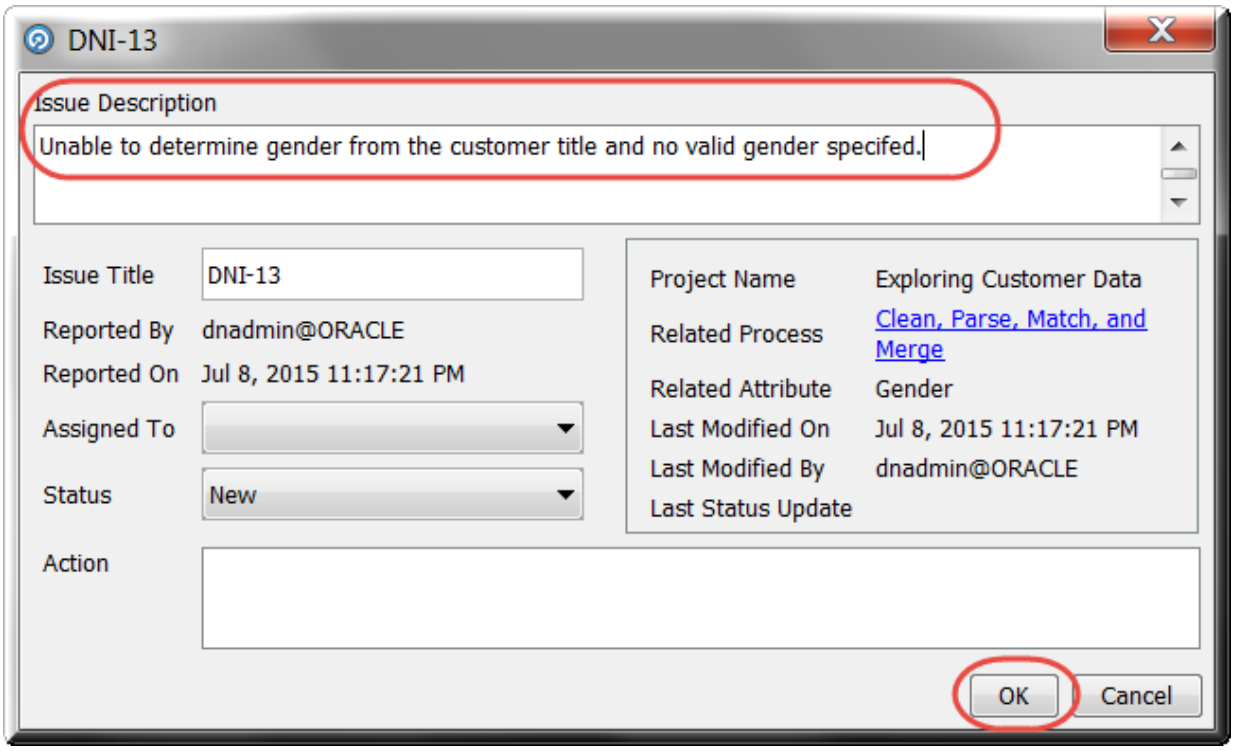
40. By observing the following **Name** column, you can see why the **Gender** was not derived



41. Scroll further to the right until you see the **Gender** column.

Notice there are invalid characters. Let's create an issue and come back to it in Lab 4. Remember, Issues can be assigned to users or groups. We will work with Issues more in a few moments.

42. CTRL-click on the 4 **U** values below the **Gender** column. Right-click and click **Create Issue...** Type ***Unable to determine gender from the customer title and no valid gender specified*** for the **Issue Description**.

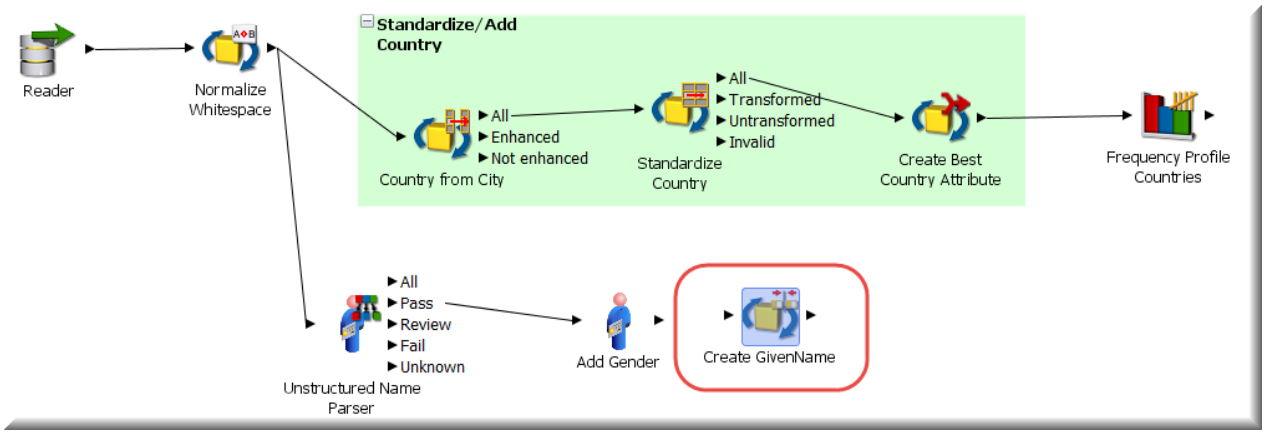


Lab 4: Match, Merge, and De-duplicate data to create Single Source of Truth for your key data entities

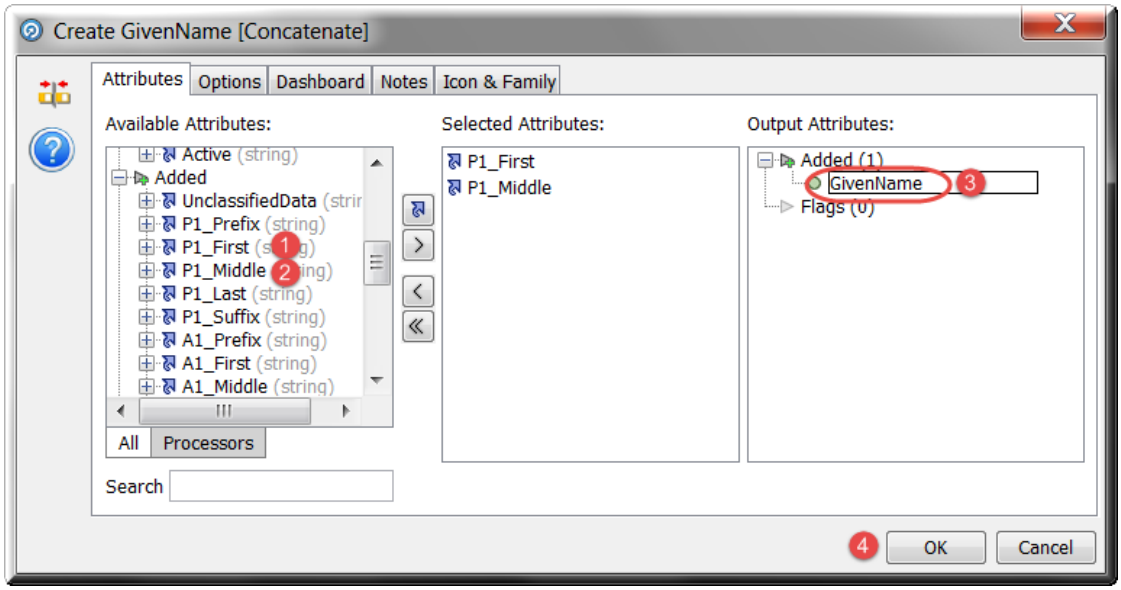
Now that we have Enriched and Standardized the Country and Gender columns (there are other data issues too – those are the ones we focused on in Lab 3), we will begin to prepare data for matching and then run a pre-configured matching processor to match, merge and de-duplicate the data.

1. Return to the **Tool Palette** and search for the **Concatenate** processor by typing ***`Concatenate`***

2. Drag and drop the **Concatenate** processor on the Project Canvas. Double-click on the label **Concatenate** to rename the processor to ***`Create GivenName`***



3. Connect the end triangle from **Add Gender** to **Create GivenName**. Select the attributes **P1_First** and **P1_Middle**. Be sure to select these attributes in this order. Double-click on **Concat** in the **Output Attributes** and rename it to ***`GivenName`***, click **OK** to continue

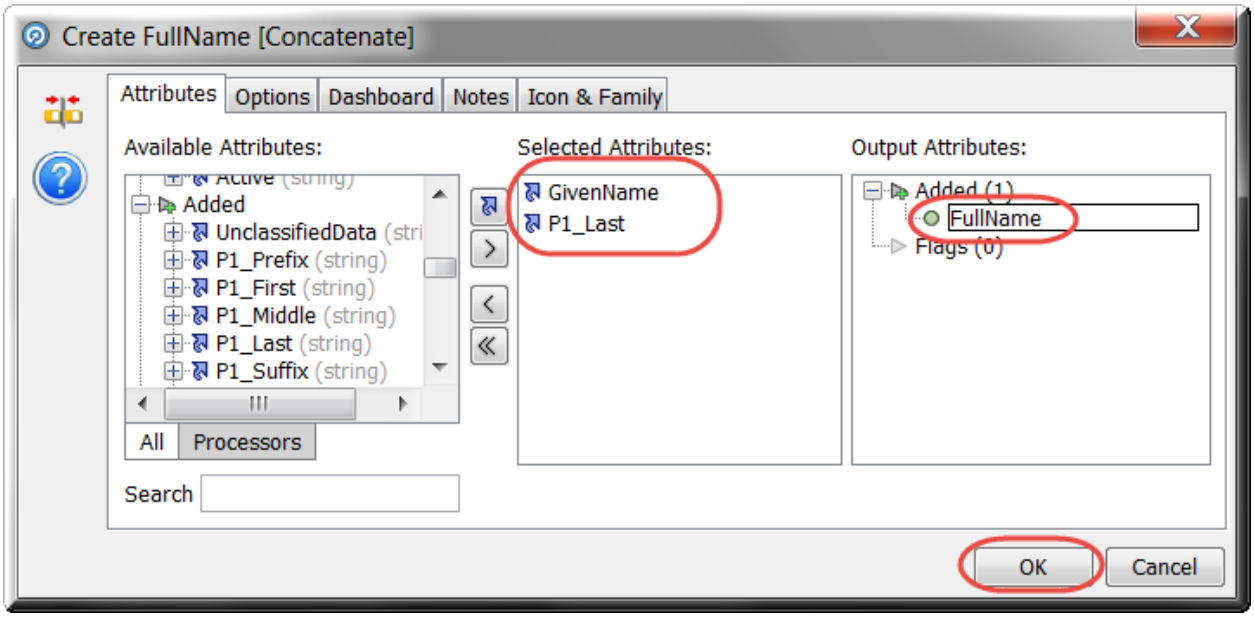


4. Right-click on the **Create GivenName** processor and click **Copy**

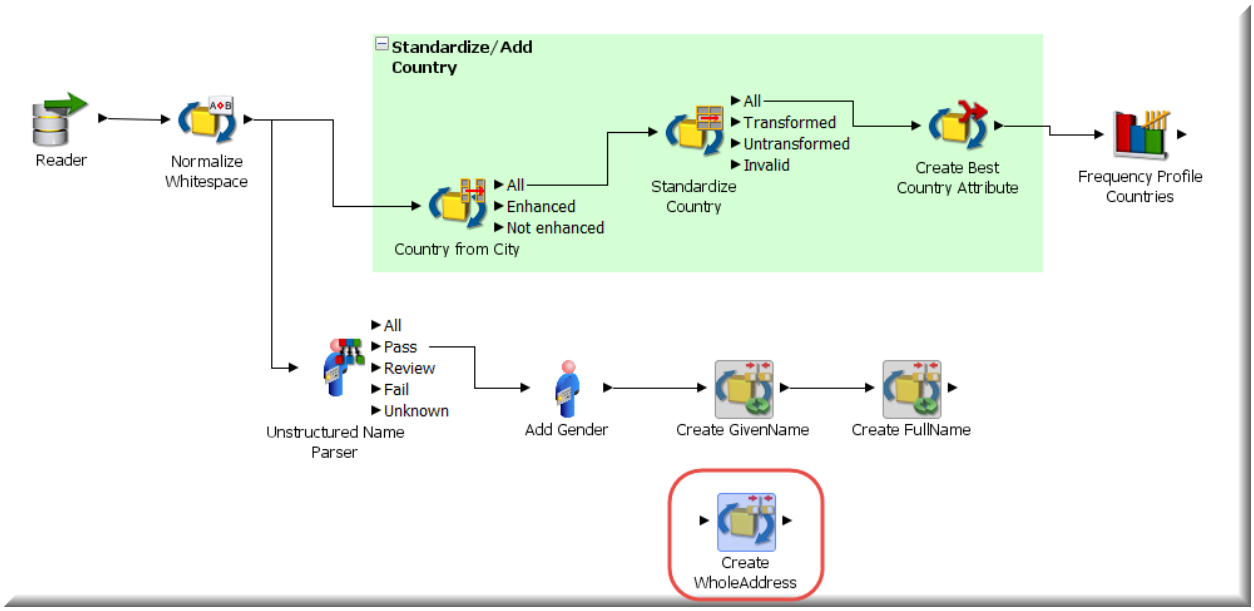
5. Right-click on the Project Canvas and click **Paste** to add another copy of the **CreateGivenName** processor. Double-click on the name of the copied processor to rename it to ***Create FullName***. Next, connect the Processor **Create GivenName** to **CreateFullName**

6. Connect the end triangle of **Create GivenName** to the input triangle of **Create FullName**. First, remove the existing **Selected Attributes** by double-clicking on **GivenName** and **P1_Last**

7. Select **GivenName** and **P1_Last** in this order. Rename the **Output Attribute** on the right-hand side of the dialog box to ***FullName*** by double-clicking on the **existing Output Attribute**



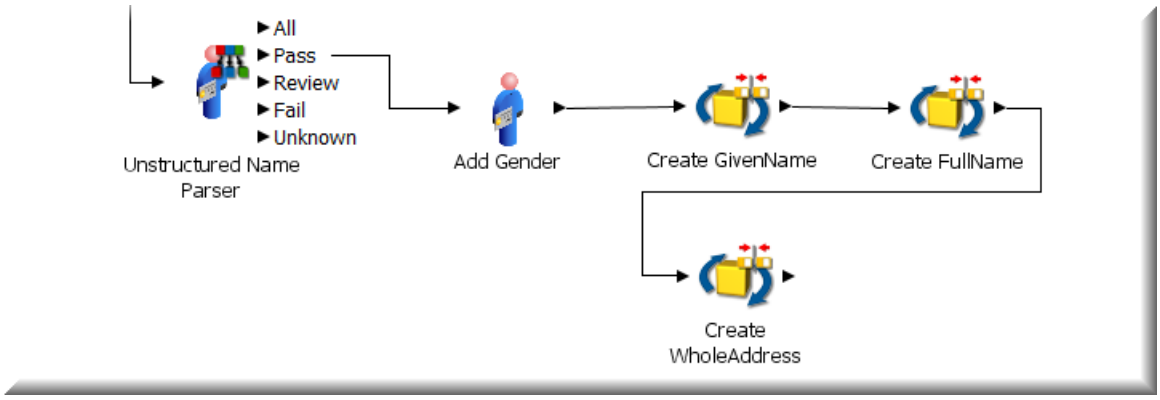
8. Navigate back to the **Tool Palette** and add another **Concatenate** processor to the Project Canvas below the existing **Create GivenName** and **Create FullName** Processors. Double-click on the name of the processor to rename it to ***Create WholeAddress***



You can double-click on the line (pipe) connecting the different processors to change it to an elbow connection.

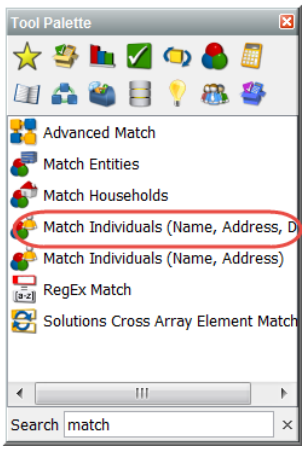
9. Connect the endpoint of **Create FullName** to the input of the **Create WholeAddress** processor. Select the **Street, City, State,** and **ZIP** attributes as the **Selected Attributes**. Rename the **Output Attribute** by double-clicking on the existing value to ***WholeAddress***, click **OK** to continue

10. Double-click the **line** connecting the **Create FullName** and **Create WholeAddress** processors to change it to an elbow connection

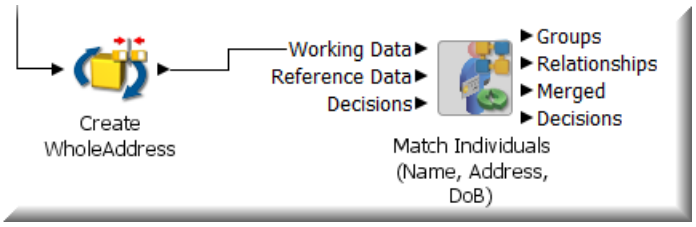


Next, we will find and add a Match Processor to Match Individuals.

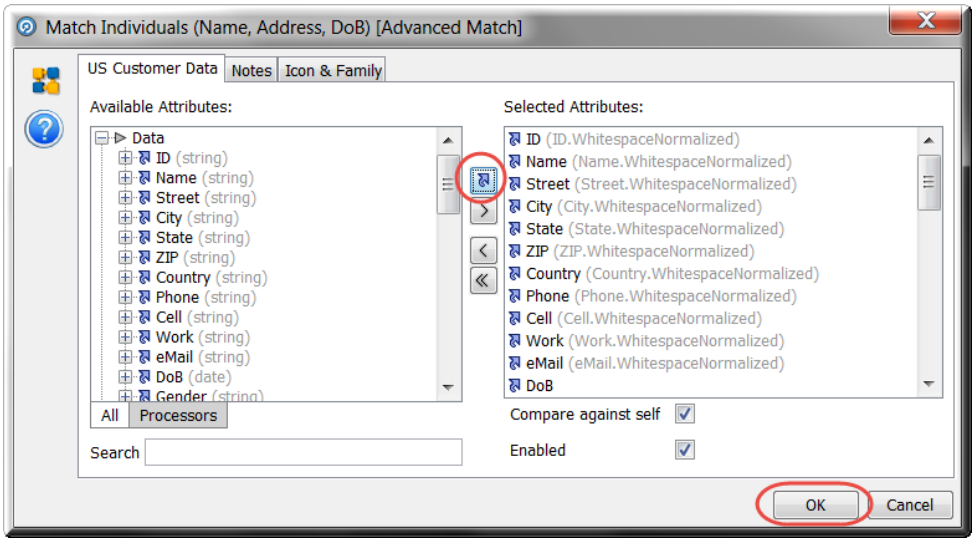
11. Navigate to the **Tool Palette** and search by typing ***Match***. You may need to expand the **Tool Palette** to find **Match Individuals (Name, Address, DoB)**. Click and drag the left border of the **Tool Palette** to resize it



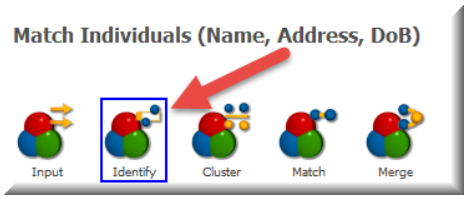
12. Drag the **Match Individuals (Name, Address, DoB)** onto the Project Canvas and connect the end triangle from **Create WholeAddress** to the **Working Data** input port of the **Match Individuals (Name, Address, DoB)** Processor



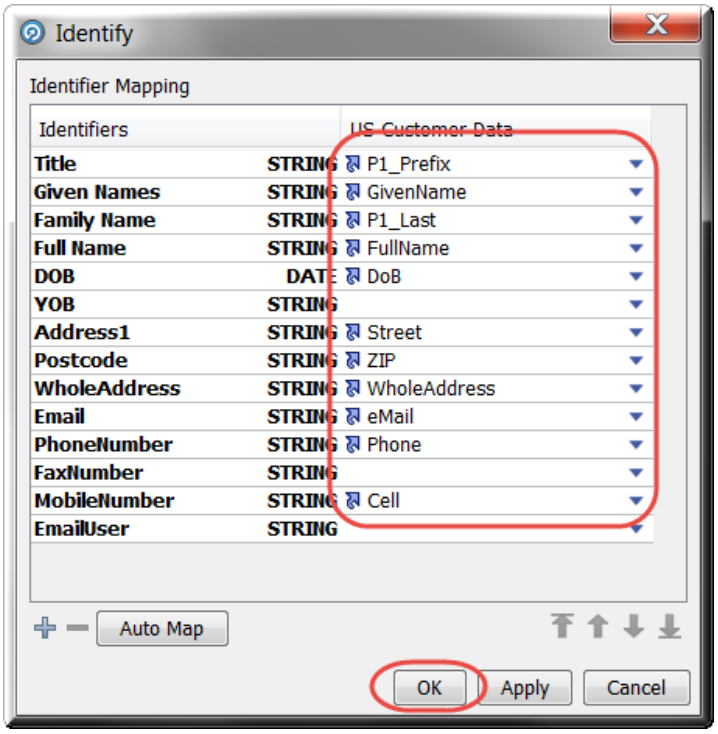
13. Use the **Select All** button to Select All attributes, click **OK** to continue



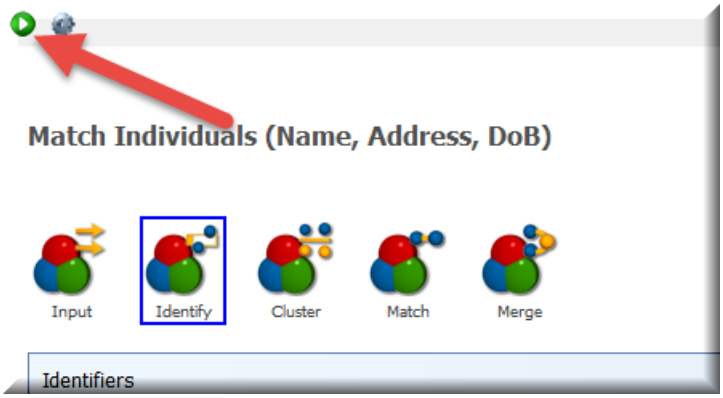
14. Double-click on the **Identify** sub-processor



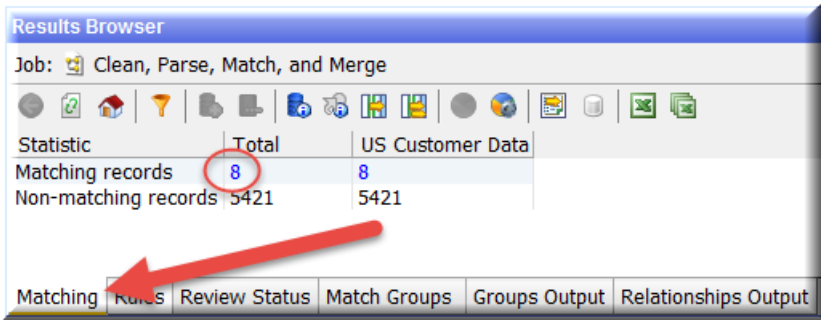
15. Connect the input attributes to matching identifiers as depicted in the graphic on the next page



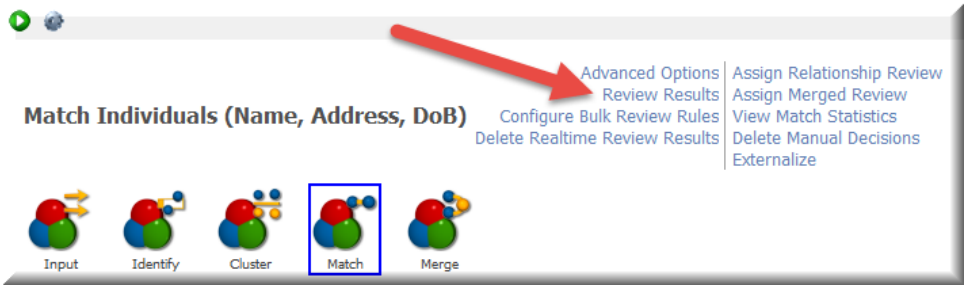
16. Click the **Run** icon in the toolbar to run the process



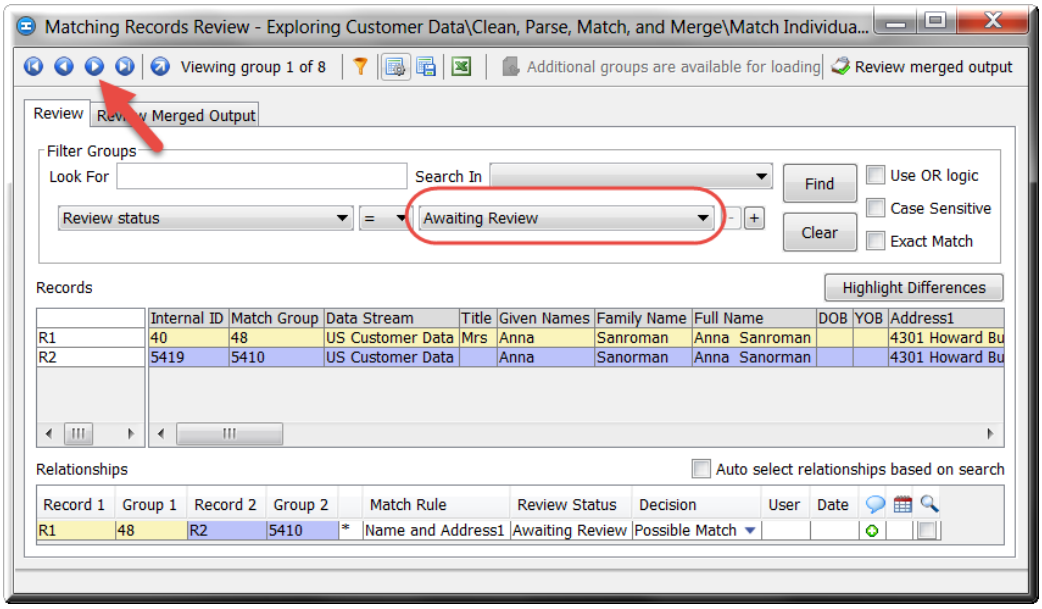
17. When the Process has finished, single-click the **Match** sub-processor to see the results in the **Results Browser**. Select the **Matching** tab on the bottom edge of the Results Browser and you will see that the Match Processor has found **8 records** that match automatically



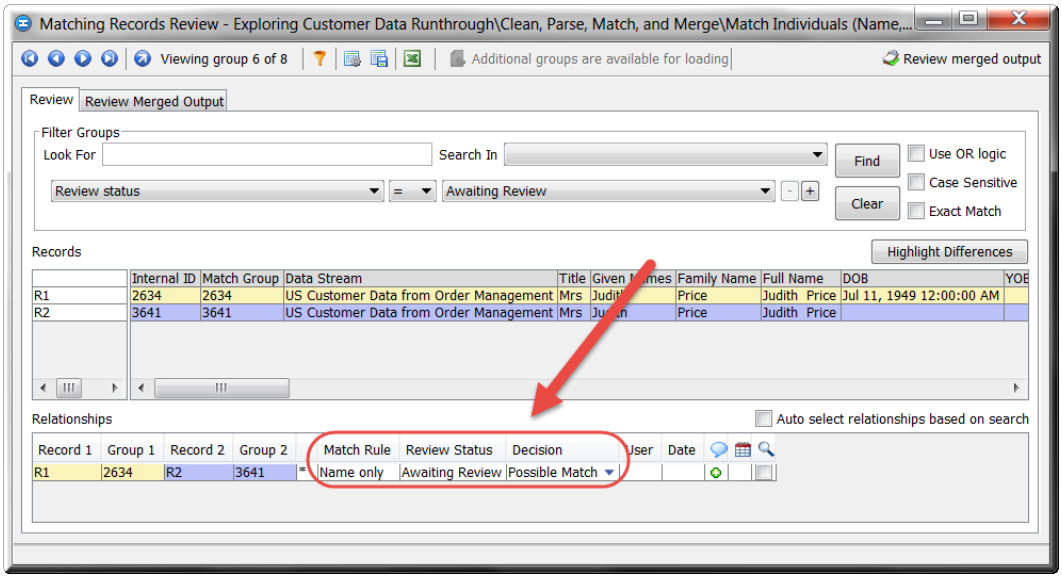
18. Click the **Review Results** link on the top right corner of the **Project Canvas** to see the matching results in more detail



19. Use the buttons in the top left to navigate through the groups that are awaiting review

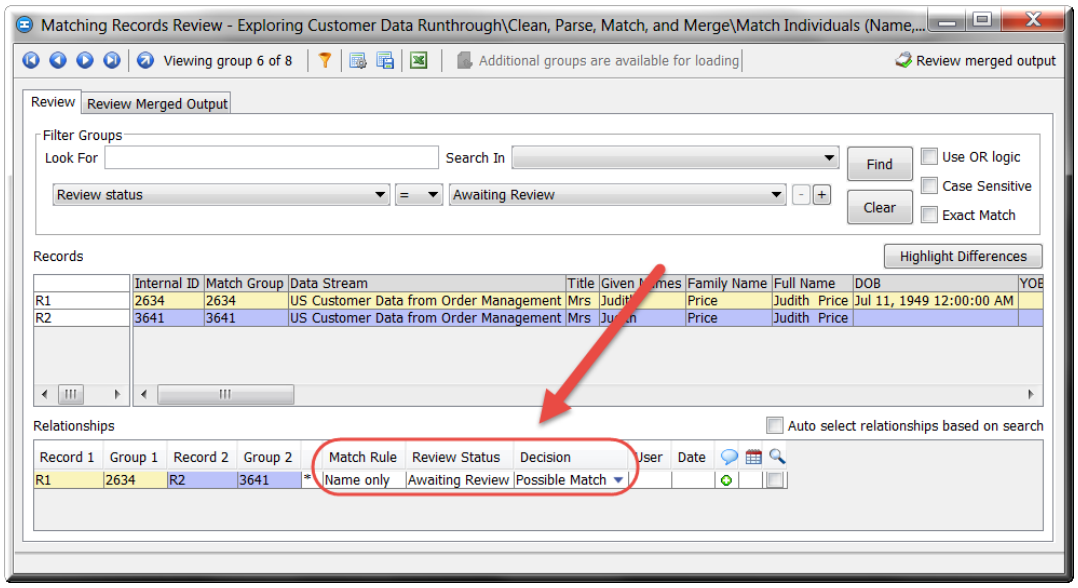


Notice at the bottom of the **Matching Records Review** window, there are additional columns Displaying the **Match Rule**, **Review Status**, and **Decision**.

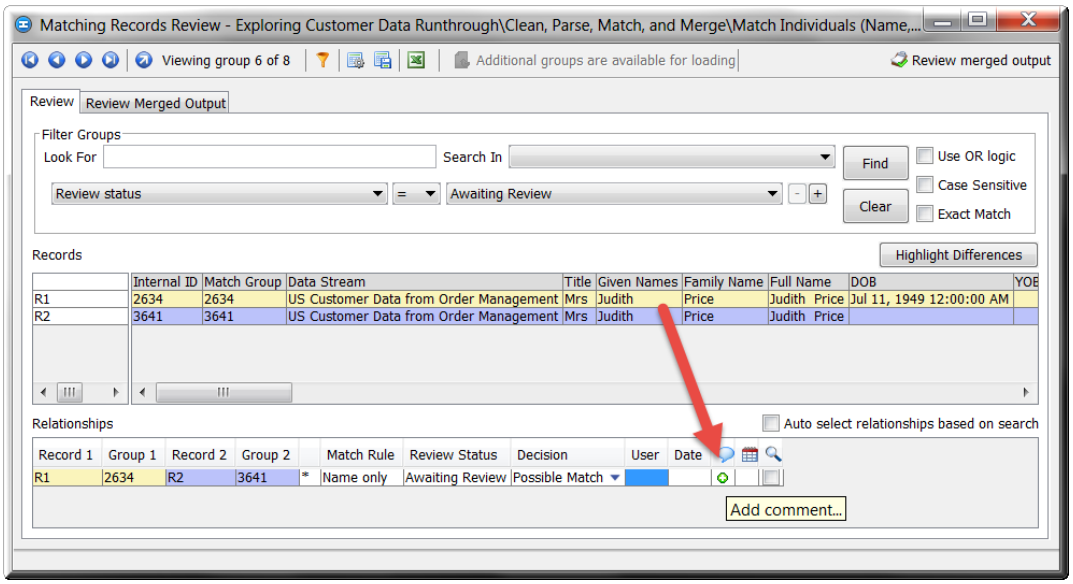


20. Click the **Decision** drop down to view the available options

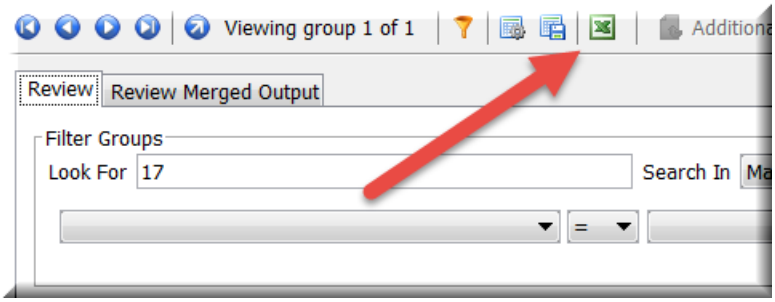
If the issue can be verified by a data quality expert, then one of the following decisions can be selected to match the record.



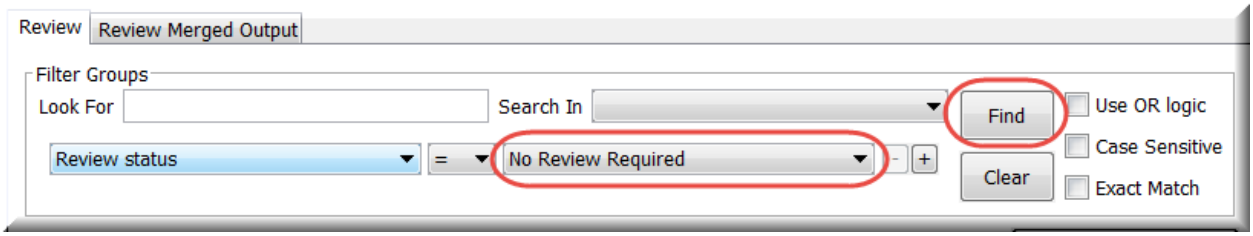
21. Additionally, if further detail is required, or a data quality user would like to place a comment, the **Green + Sign** ⊕ under the **Speech Bubble** can be clicked to add a comment



22. As in previous Results Browser screens, an **Export to Excel** button is available here to send the Matching Records Review results to another user for verification



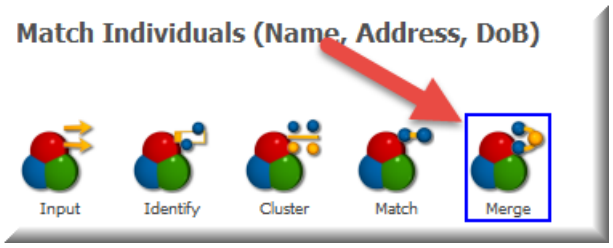
23. To see what was matched automatically, change the on-screen filter to display groups with a **Review Status = No Review Required**



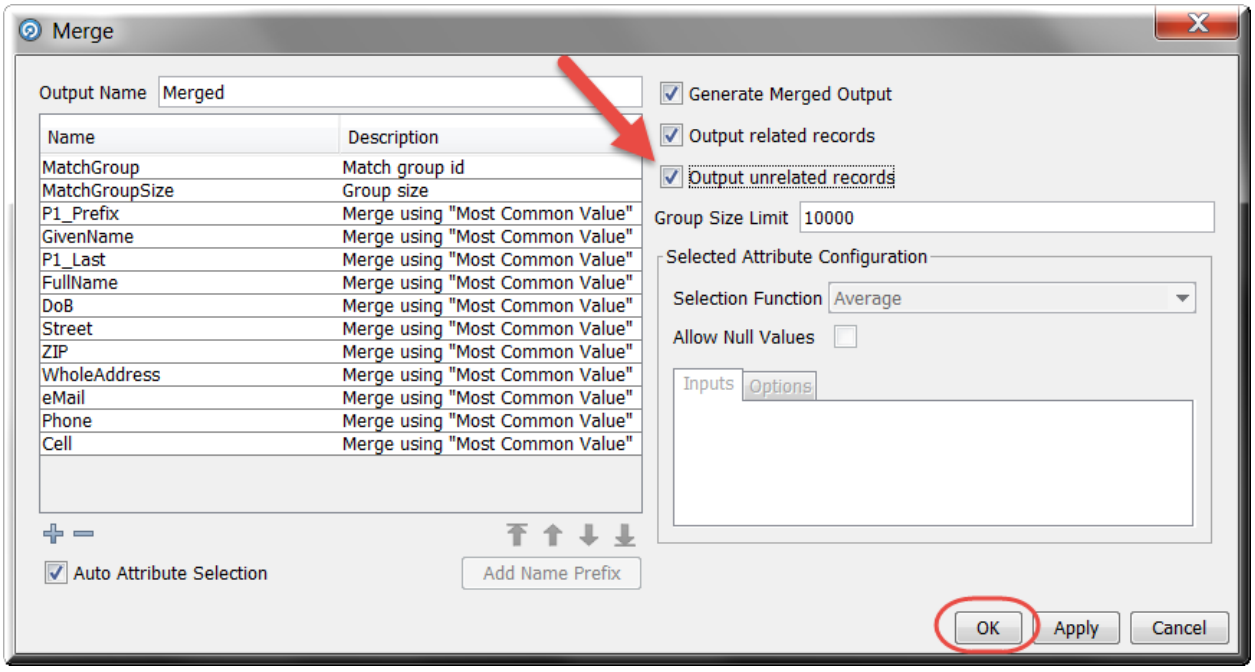
24. Navigate through the groups to review the matching results using the **Left** and **Right** arrows at the top left corner of the **Matching Records Review** dialog box. When finished, close the **Matching Records Review** window

So far we have Cleaned, Parsed, and Matched data from this customer dataset. At this point, we will use the Merge sub-processor to write out de-duplicated records from the Match Process using default merging rules.

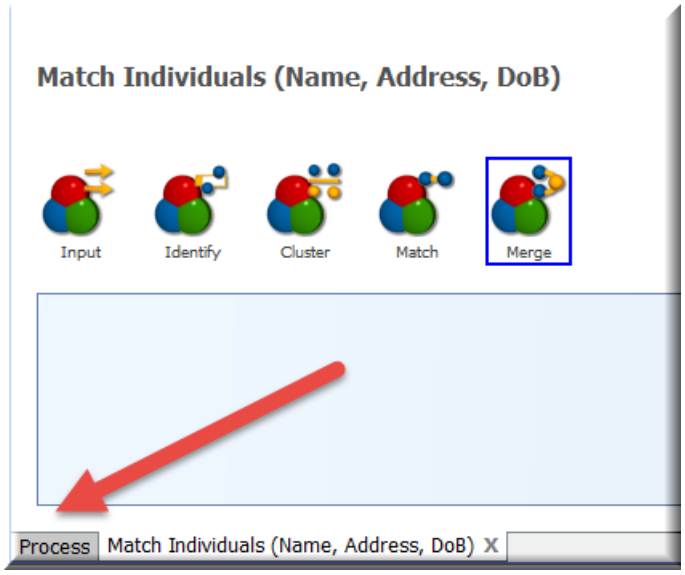
25. Double-click the **Merge** sub-processor to begin and open the **Merge** dialog box



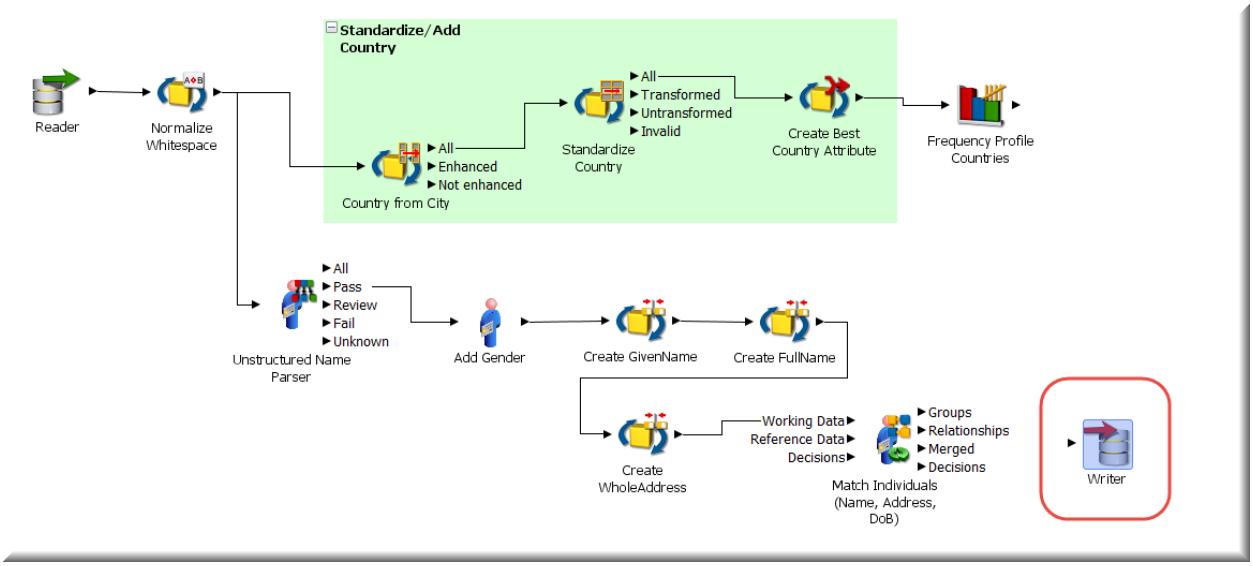
26. Tick the box next to **Output unrelated records** in the **Merge** dialog, click **OK** to continue



27. Click the **Process tab** in the middle-left of the canvas to return to the Project Canvas



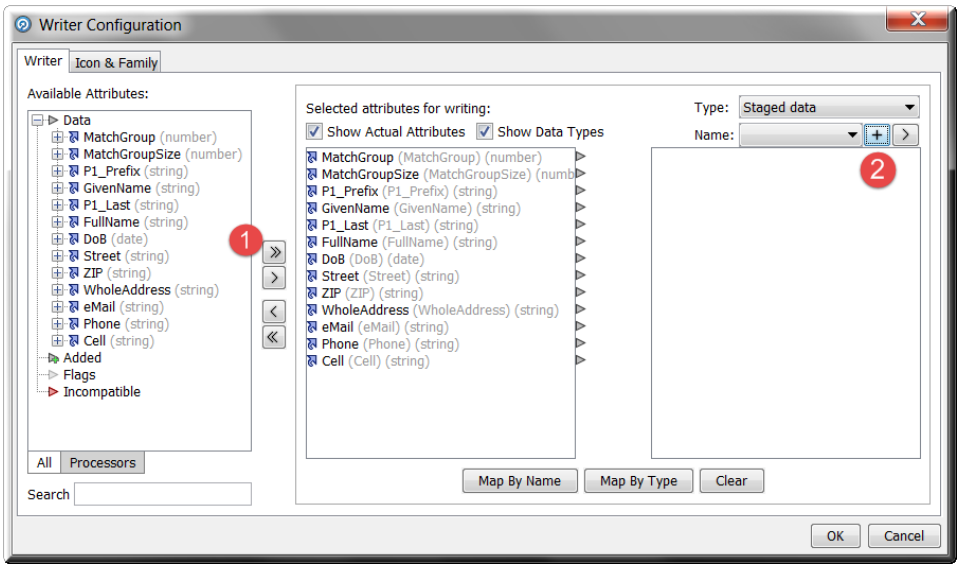
28. In the **Tool Palette**, search by typing ***Writer***. This processor enables an EDQ process to write data to different types of data stores, for example, Staged Data. Drag and drop the **Writer** to the right of the **Match Processor** on the Project Canvas



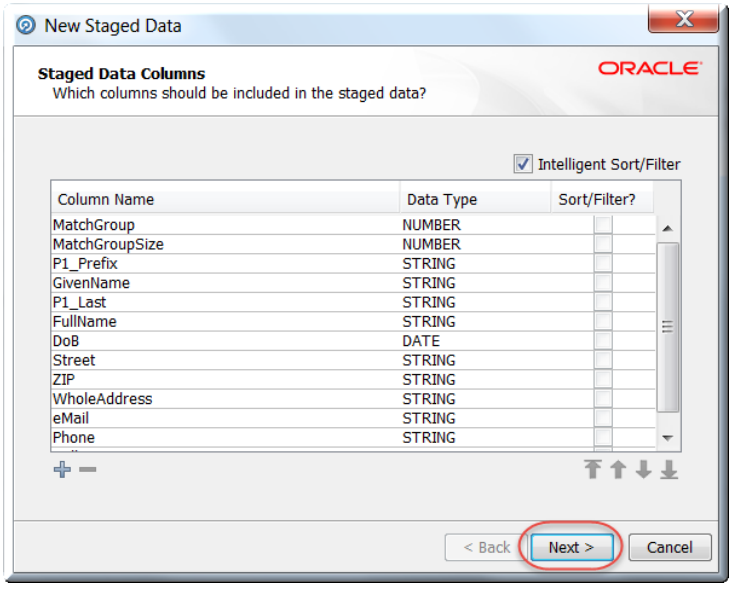
29. Connect the end triangle from the **Merged** output from the **Match Individuals** Processor to the **Writer**

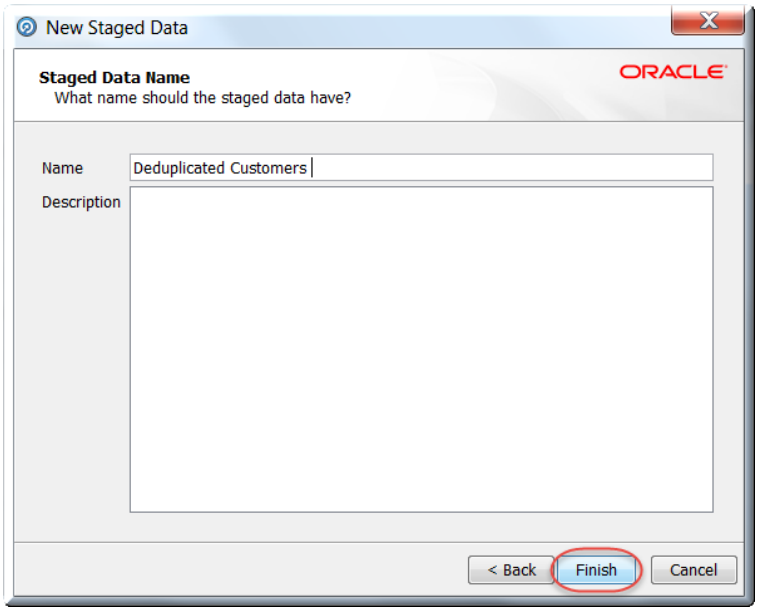
There are other forms of outputs for the Match Individuals processor. For example, Groups, which are the input records organized into match groups. Relationships, which are all links between matching records.

30. Click the » button to Select All **Available Attributes**. Then press the + button on the right side of the **Writer Configuration** dialog box to add a new **Staged Data** set

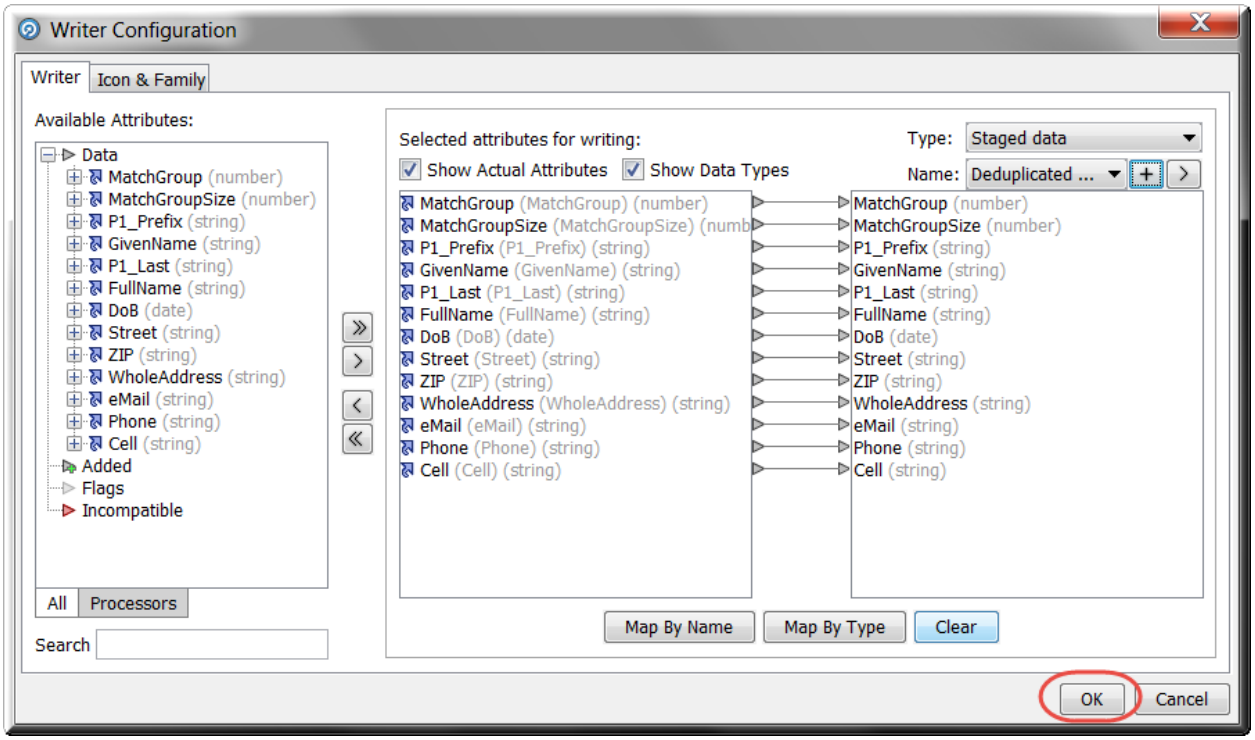


31. Click **Next >** to accept the default configuration of creating the Staged Data Set. Give the Staged Data set a name by typing ***Deduplicated Customers***, click **Finish**

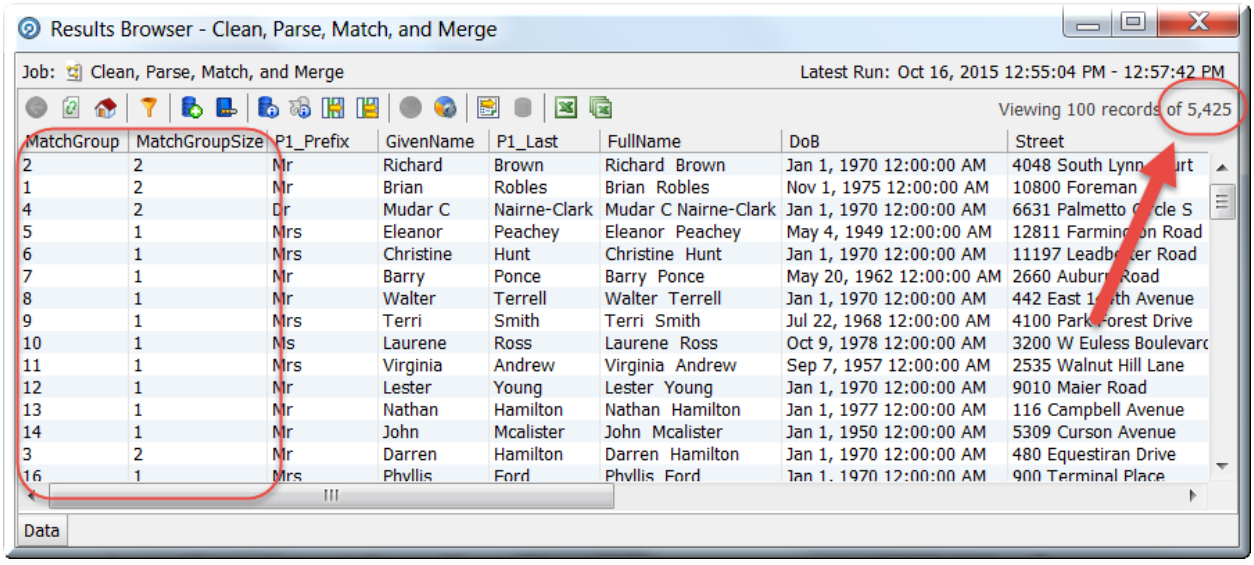




32. Click **OK** to finish setting up the **Writer Configuration**



33. Click the **Run** icon in the toolbar to run the process. When this process has completed, right-click on the **Writer** and click **Show results in new window**

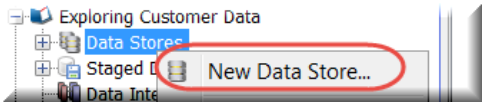


There are various match groups with a MatchGroupSize of 2 - each single record was written out for the 2 matching input records. Also, the total volume of written records is 5425, rather than the 5438 in the **Reader** processor.

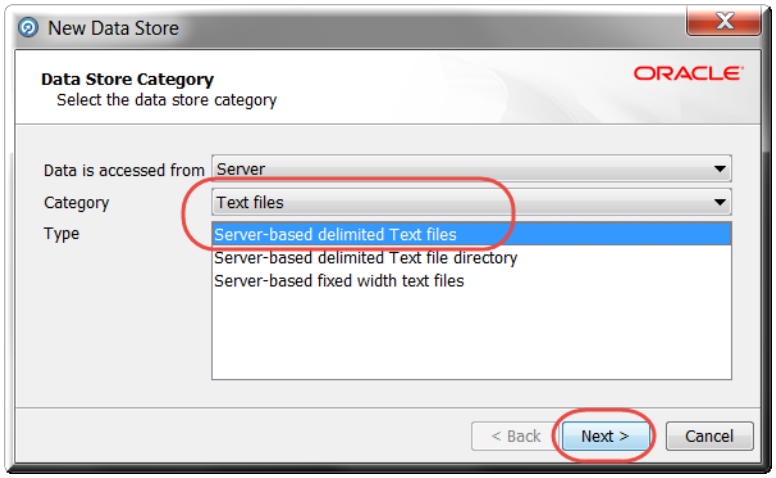
Lab 5: Create a Job to Automate Data Quality with ODI and EDQ Integration

Before we move on to Lab 6, we will create a job that runs data through the process we just created. Next, we will configure the Job to run within an ODI package to automate the Data Quality project.

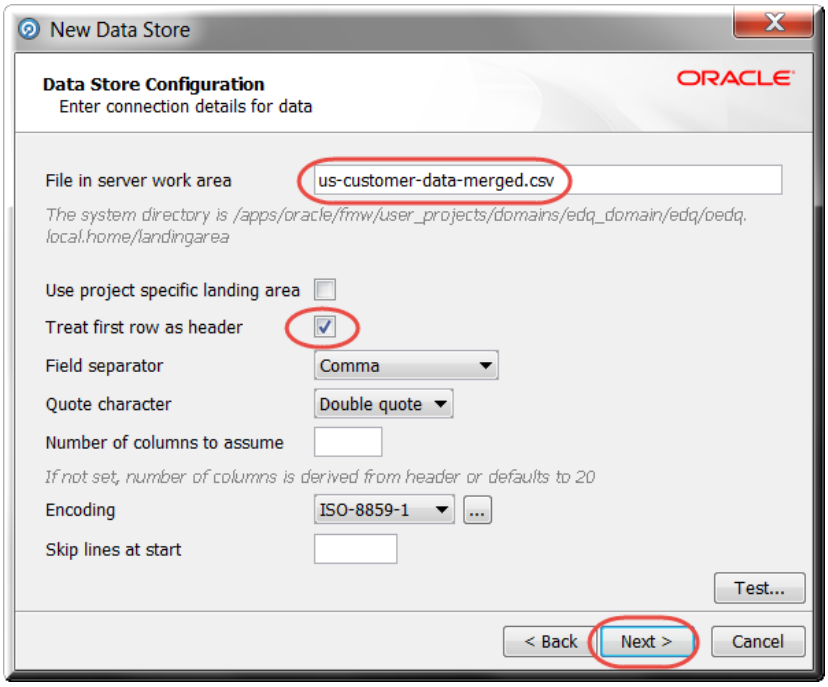
1. In the Project Browser on the left-side of the screen, right-click **Data Stores** and click **New Data Store** under your **Exploring Customer Data** project



2. Select **Server** and **Text files** in the two drop down boxes, then click **Server-based delimited Text files**, and click **Next >** to continue

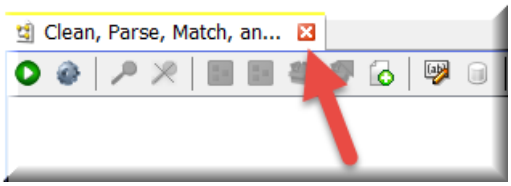


3. In the File in server work area, enter a name for the export file by typing ***`us-customer-data-merged.csv`***. Also, check the option to **Treat first row as header**

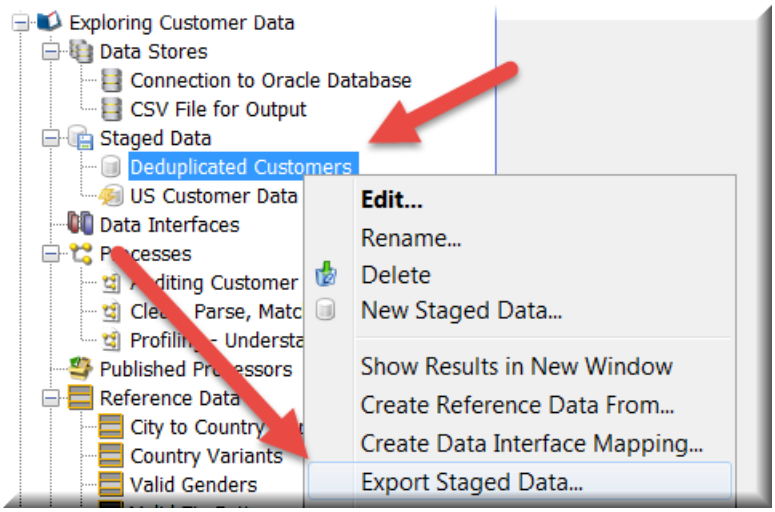


4. Click **Next >**, and name this Data Store – type ***`CSV File for Output`***, click **Finish**

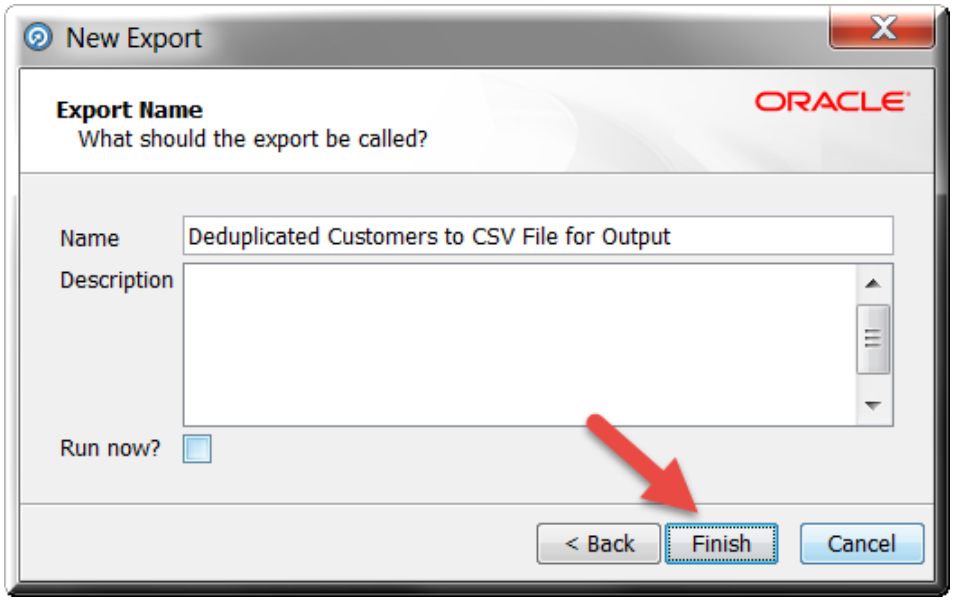
5. Close your **Process tabs** at the top of the **Project Canvas** if you haven't already. If you are prompted to save a given process, click **Yes**



6. Navigate to the Project Browser on the left side of the screen and right-click on **Deduplicated Customers** under the **Staged Data** category. Expand the Staged Data category if you do not see it. Select **Export Staged Data...** Click **Next >** on the first window of the dialog since the Staged Data is already selected



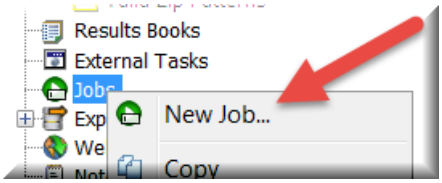
7. Select the **CSV File for Output** data store as the target for the data export. Click **Next >** and **Finish** to keep the default export name and to complete the wizard for configuring the new Export Configuration



Create a Job

Next we will create a job that can be invoked externally, for example, from Oracle Data Integrator.

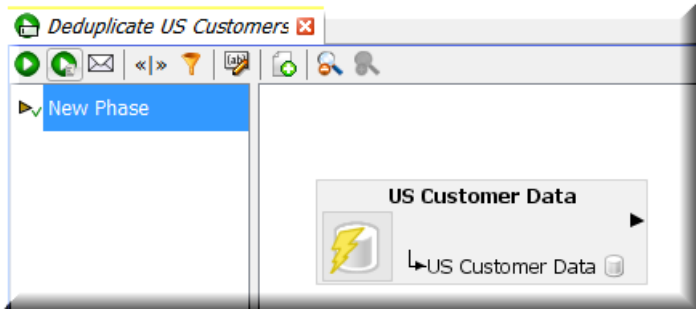
8. Navigate to the Project Browser, and select the **Jobs** category. Right-click it and select **New Job...**



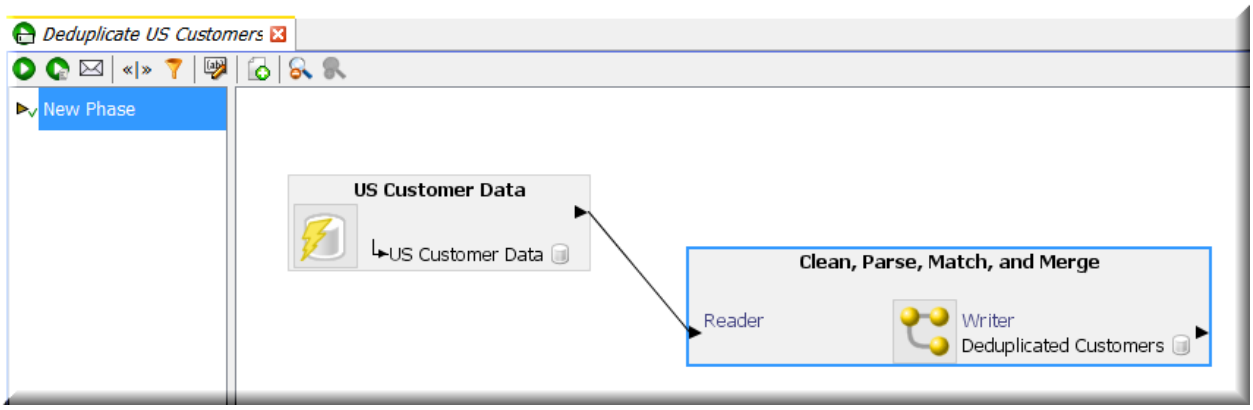
9. Name the job by typing ***`Deduplicate US Customers`***, click **Finish**

The Job Canvas is displayed next with a slightly different Tool Palette. The **Tool Palette** contains all of the runnable configuration tasks in the project including snapshots, processes, exports, etc. Notice the **icons at the top of the Tool Palette** are different – click through them to explore.

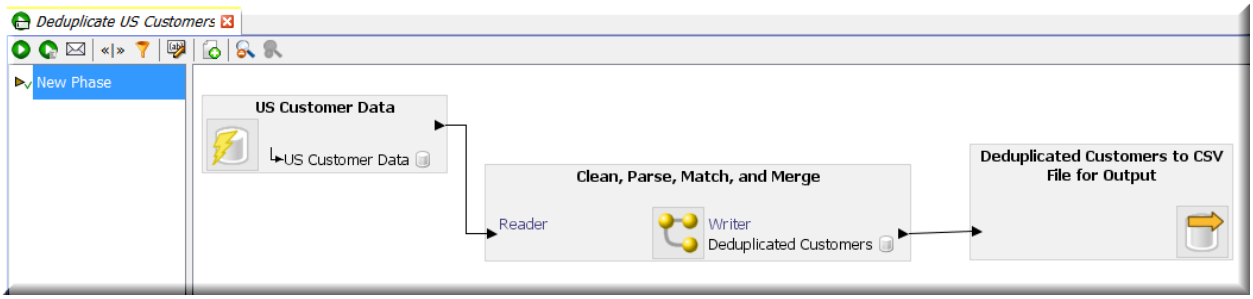
10. Click the  icon in the **Tool Palette** to display the **Snapshots** and drag the **Connection to Oracle Database.US_Customer_DATA** snapshot onto the canvas



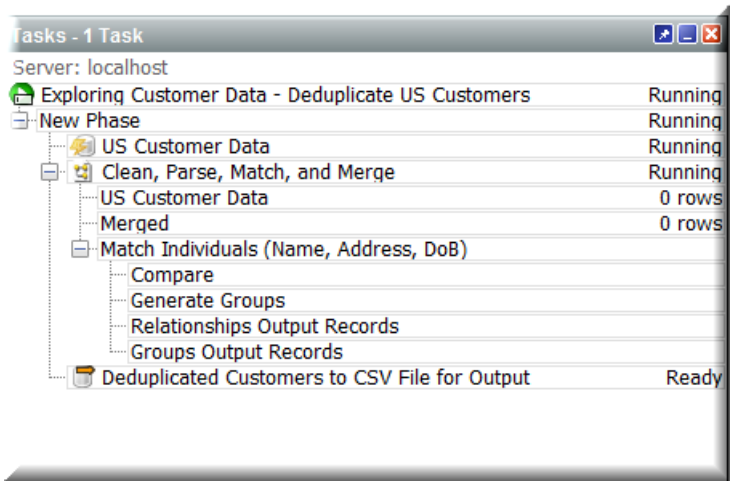
11. Next, click the  icon in the **Tool Palette** to display the **Processes** and drag the **Clean, Parse, Match, and Merge** process onto the canvas



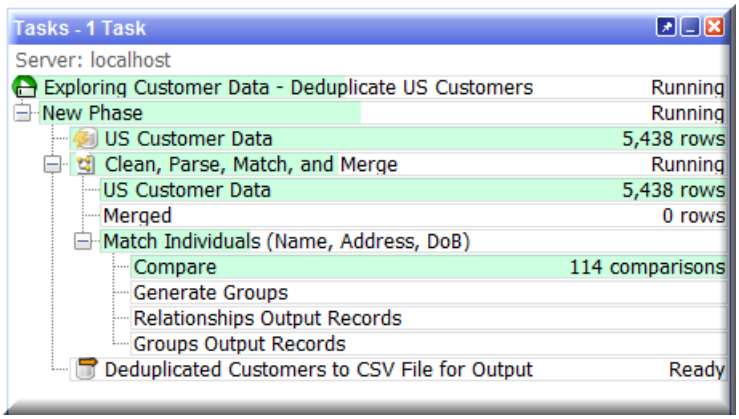
12. Lastly, click the  on the right side of the **Tool Palette** to display the **Exports**. Drag and drop the **Deduplicated Customers to CSV File** for Output export onto the canvas



13. Click the **Run** icon in the toolbar to run the job. Note the **Tasks** Window in the bottom left of the Director



There are three tasks – Snapshot, Process, and Export) all run together to refresh the input data, running it through the process you created, and writing out the file to a new CSV file in the server landing area.

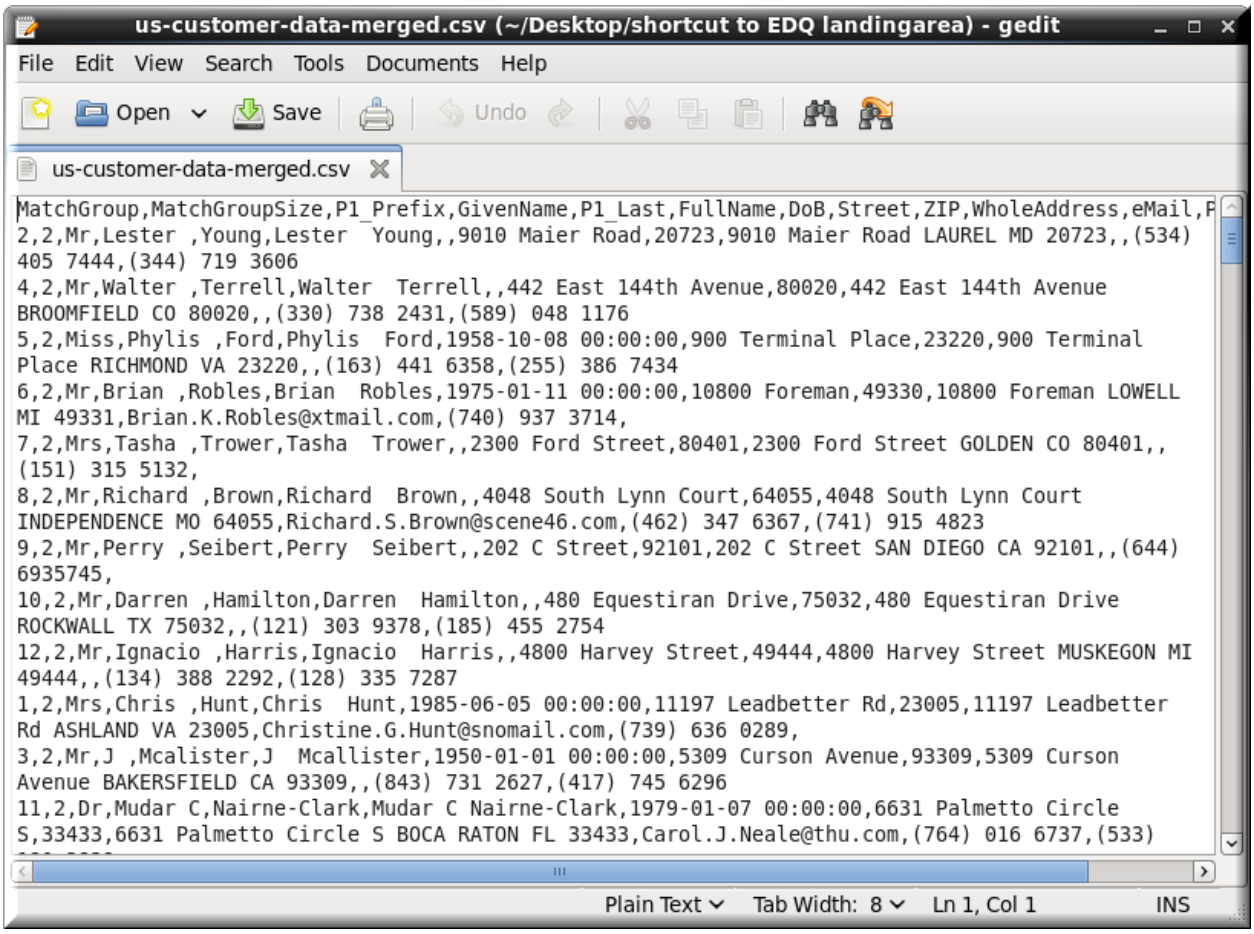


When the job has finished running you can click the **shortcut to EDQ landing area** on the Virtual Machine's desktop and see the **us-customer-data-merged.csv** file that the job has created.

14. Click the  icon in your task bar and switch to the desktop of the EDQ 12.2.1 Virtual Machine

15. Double click the 

shortcut and find the **us-customer-data-merged.csv**



This job can now be used with Oracle Enterprise Data Quality's built-in scheduler, any external schedule, or using the pre-integrated agent with Oracle Data Integrator to create Fit for Use Data Warehouses and Accurate Analytics.

ODI and EDQ Integration – Follow Along (*No Hands-On Exercise*)

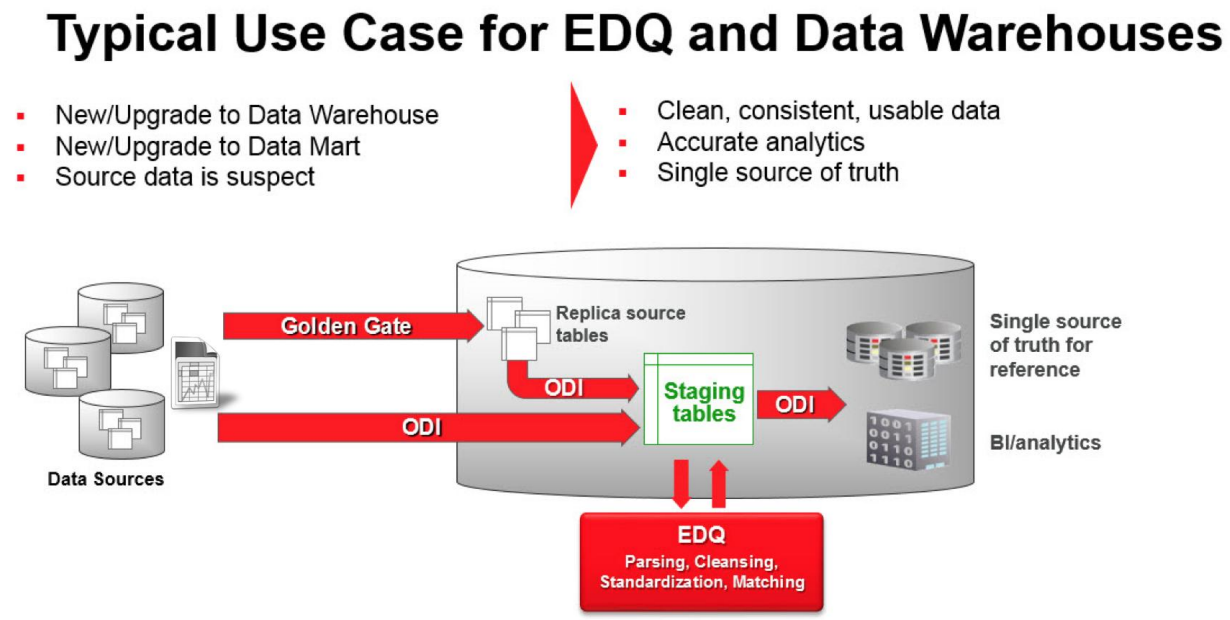
In this short example, you will see how to automate Data Quality projects and how integrate a job with ODI. Since ODI is not installed within the EDQ workshop environment, the following descriptions and screenshots are meant to show you one example of how the integration between the two products can be accomplished.

Since there are several other components and concepts involved with integrating ODI and EDQ, it's important to bring up a few basics. It is possible to create several mappings in ODI to accomplish several different types of data flows and transformations. The mappings created in ODI can be used in packages, which allow the ODI user to create a flow and order in which those mappings should execute. In addition to including mappings in packages, ODI also

includes Open Tools that can be used within packages to introduce additional actions like sending an email, retrieving files using FTP, or playing a Beep on the success or failure of the package.

An additional Open Tool included with ODI is called the EnterpriseDataQuality Open Tool. This particular tool can be used to invoke EDQ jobs from an ODI package using a Java Management Extension (JMX) interface. After dragging this tool into an ODI package, a few parameters will need to be specified.

There can be several different types of use cases for integrating ODI with EDQ.



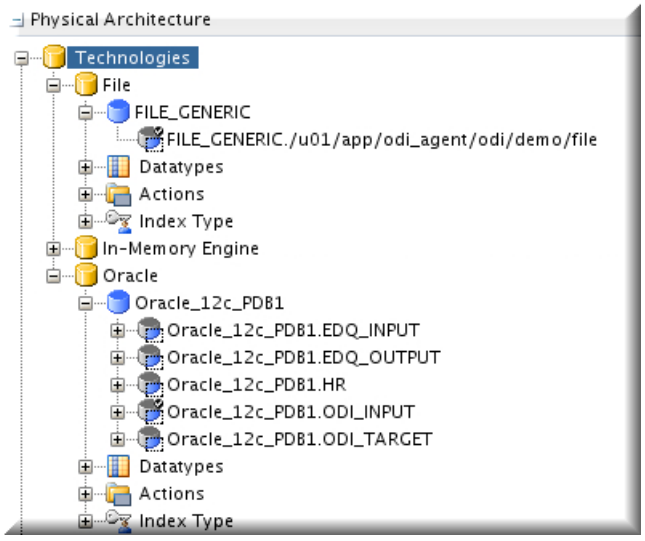
In the following example, we will create an ODI mapping to load a .CSV file to a database table, which is the ODI Input source to be used to load an EDQ Input table. EDQ will read the EDQ Input to process it using an EDQ Job, like the one created earlier. This will profile, audit, parse, cleanse, match and merge and de-duplicate the data as we performed in the previous labs.

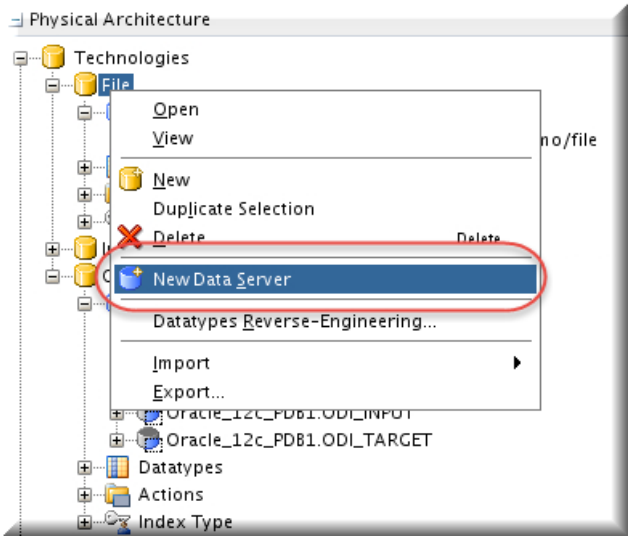
Once the EDQ process is complete, EDQ will write the output of the job back to a database table, which we will refer to as the EDQ Output. Lastly, an ODI Mapping will load the EDQ Output to an ODI Target table. For example, the ODI Target could be a data mart or data warehouse that is used for analytics.

Sample ODI Topology Setup

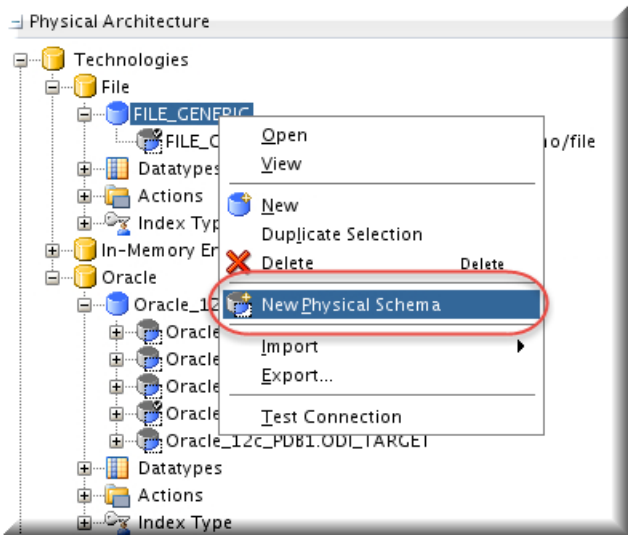
For this example, we will setup a few data stores in the ODI Topology using the File and Oracle Physical Technologies.

1. To accomplish data movement from a file to database staging tables to a database target, the following data servers and data stores need to be setup. Right-clicking on the technologies, for example **File** and **Oracle** listed below reveals the option to create

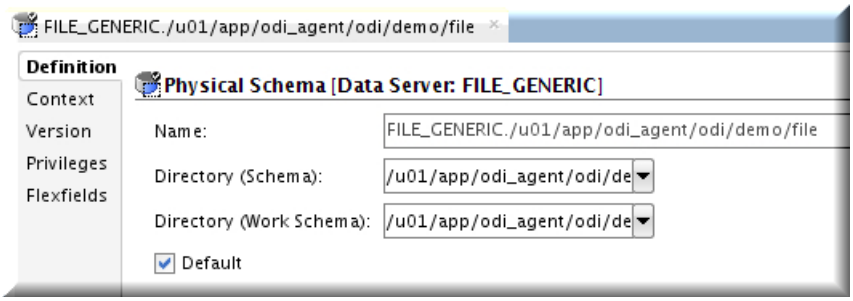




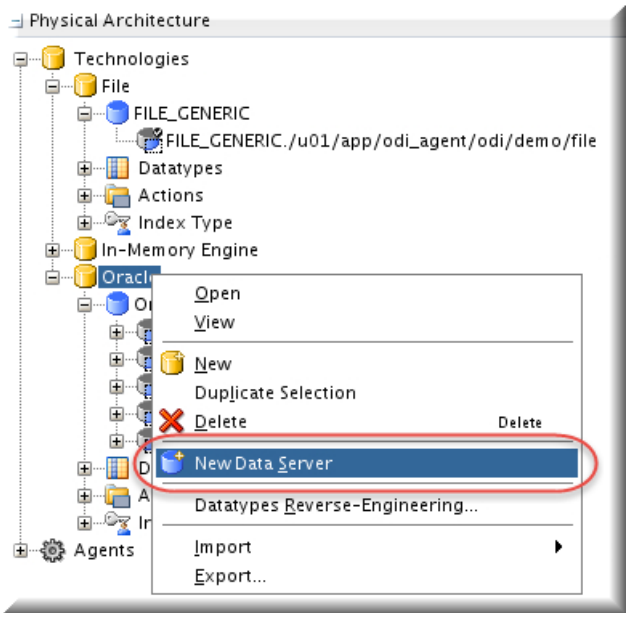
2. In the case of **File**, the default values for the **FILE_GENERIC** data server are used. Right-Clicking on the **FILE_GENERIC** data server reveals the options to create a **New Physical Schema**

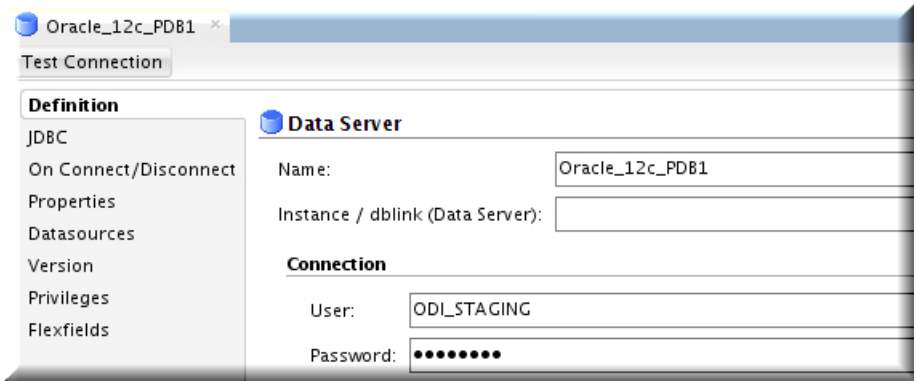


3. The physical schema was created as depicted in the following screenshot. In the case of **Directory (Schema)** and **Directory (Work Schema)**, the location of the .CSV file containing customer data was provided: ***/u01/app/odi_agent/odi/demo/file***. This .CSV file is stored on the same storage as the client running ODI Studio



4. Next, the physical connections to the Oracle Database will need to be created to facilitate the data movement from the .CSV file to the ODI Input and subsequent staging tables. An **Oracle** data server connection was created by right clicking on the **Oracle** technology

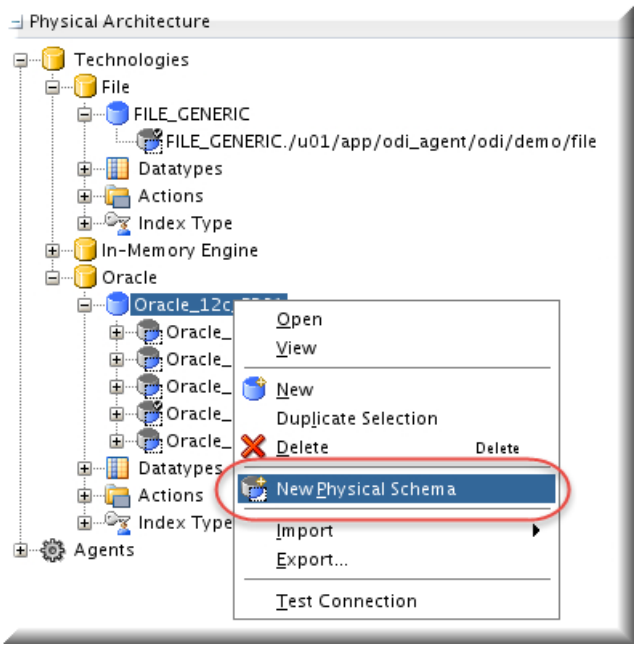




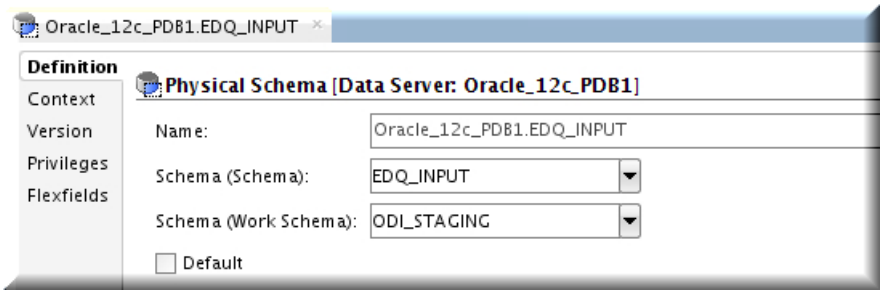
The connection string to this Oracle database was provided on the **JDBC** tab. Afterwards, you can click the **Test Connection** button in the upper-left corner of the data store configuration screen. Once the connection is successful, you can begin to create the configuration for the data stores (schemas) to connect to.

5. This example will require 4 different data store configurations: **ODI Input**, **EDQ Input**, **EDQ Output**, and **ODI Target**. The input and output tables will serve as staging tables for performing data quality work in preparation for the load to a target table.

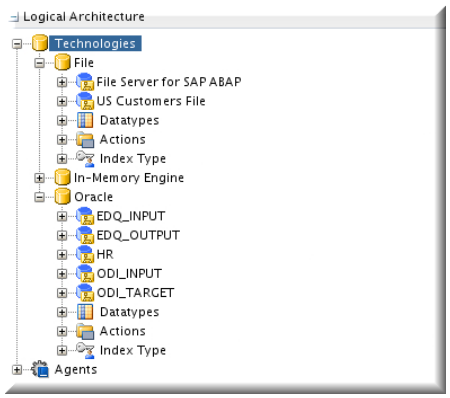
6. Right-clicking on the **Data Store** beneath **Oracle** in the Topology – Physical Architecture tab allows for setting up a **New Physical Schema**



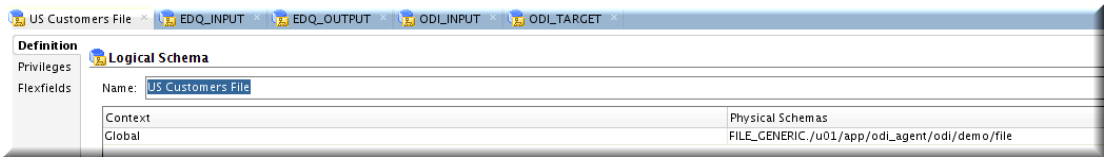
7. In order to reduce repetition, all four of the physical schemas will be created using identical steps. The only difference will be the schema selected in the **Schema (schema)** dropdown. Each selection will refer to previously created schemas called ODI Input, EDQ Input, EDQ Output, and ODI Target

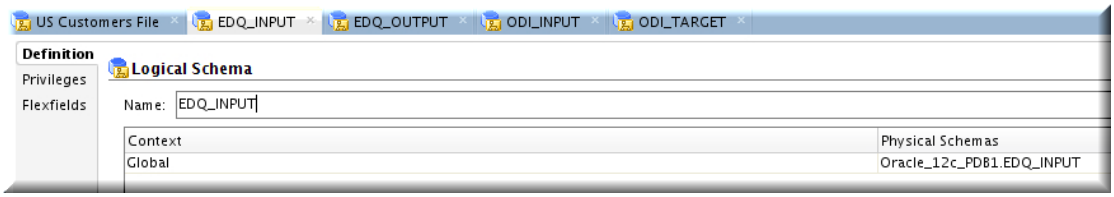


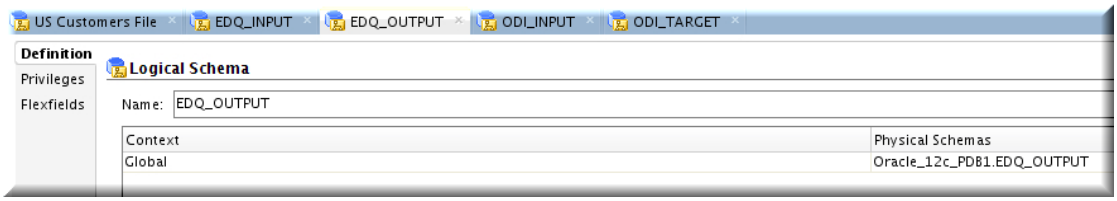
8. As with typical ODI Topology setup, a logical schema must be created for each of the data stores created above. Right-clicking on the technologies **File** and **Oracle** will reveal the option to create **New Logical Schema**. The following logical schemas were created

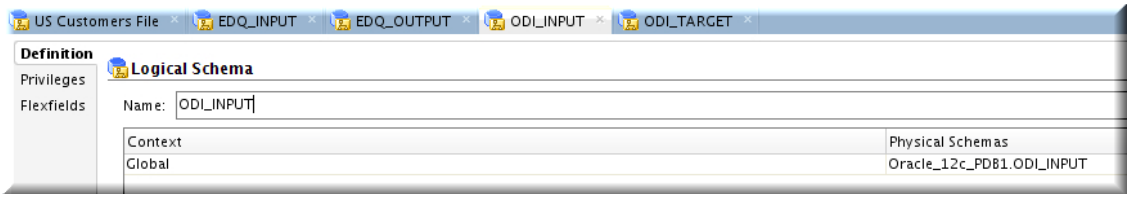


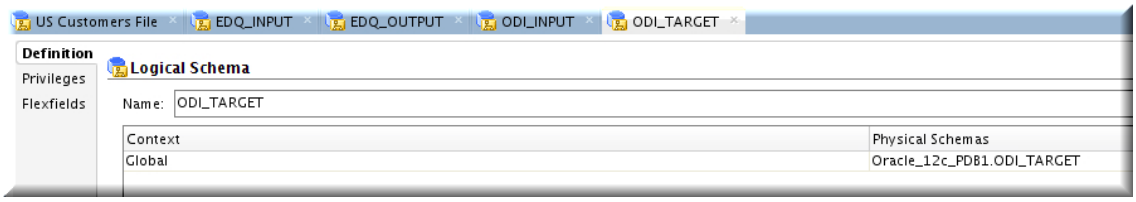
9. For each logical schema created, the appropriate **Physical Schema** was selected on the **Definition** tab



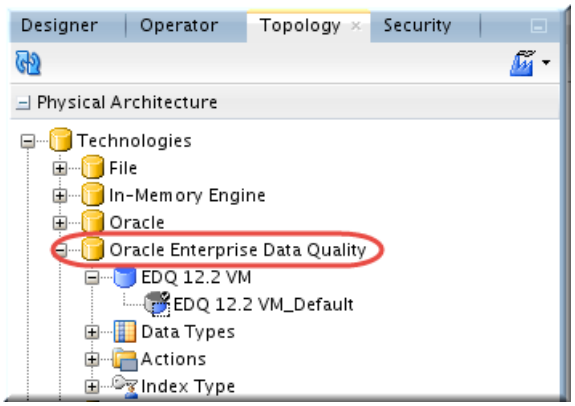




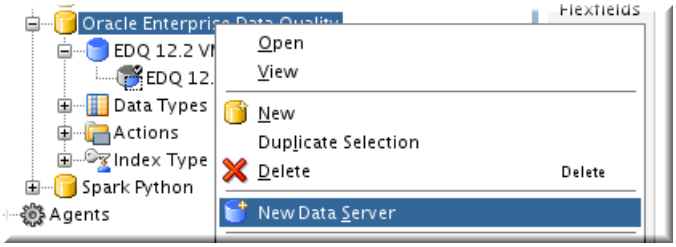




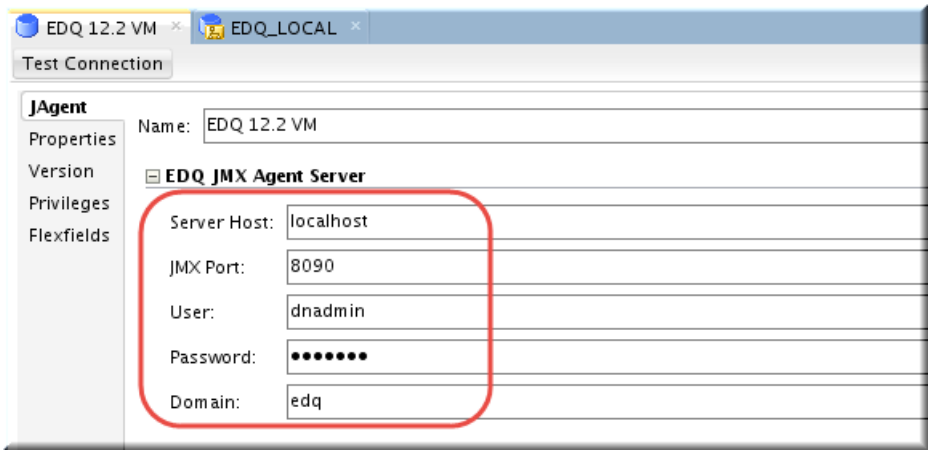
10. In Oracle Data Integrator 12.2.1, Enterprise Data Quality is a technology option within the Topology Navigator. A data server was created to store the details of the EDQ server to be used within an ODI Package.



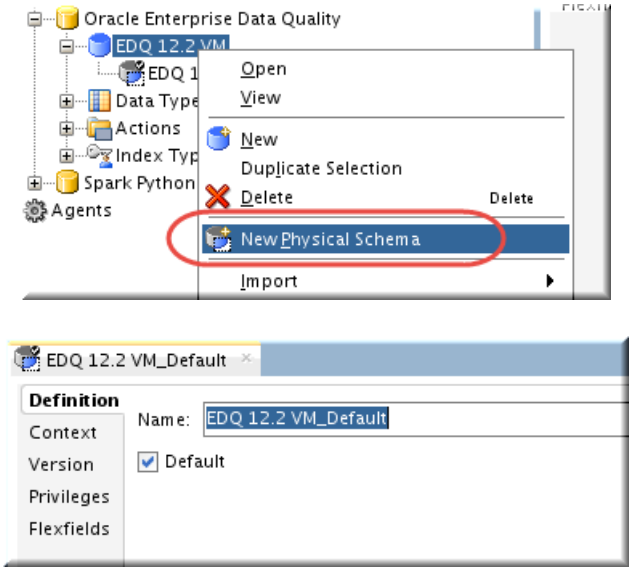
11. Right-clicking on **Oracle Enterprise Data Quality** and selecting **New Data Server** brings up the dialog to setup the EDQ server connection:



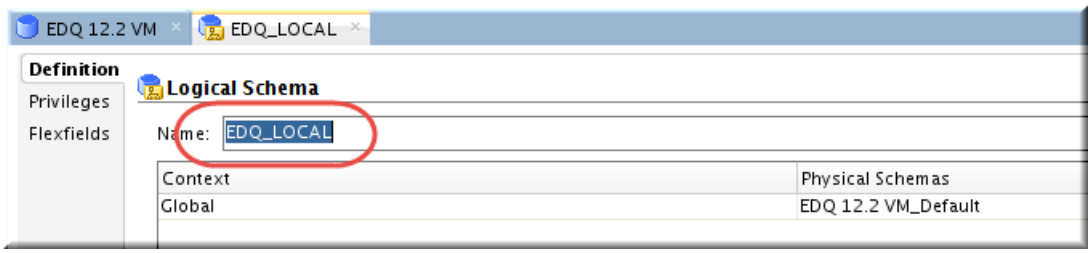
12. A **Name** for the EDQ data server and the **EDQ JMX Agent Server** details were filled in as the following. 8090 is the default JMX port configured on this EDQ VM.



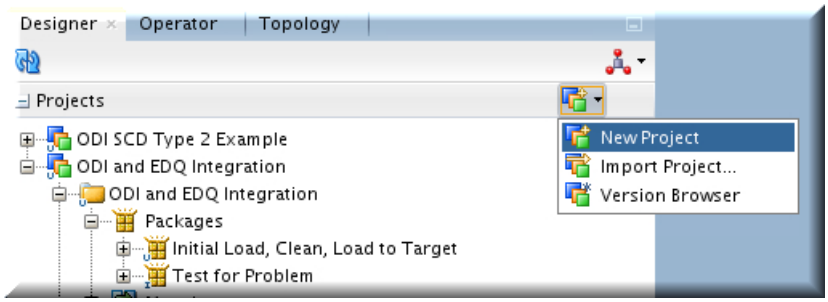
13. Afterwards, a physical schema was created and saved with the default information



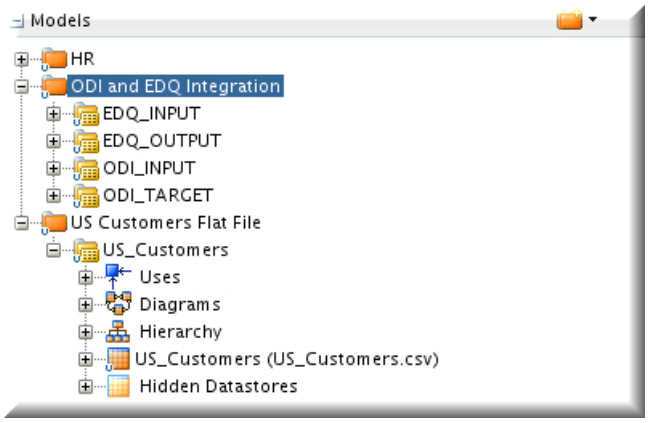
14. Just as with the file and database data servers, a Logical Schema was created for the EDQ Physical Data Server. This detail will be provided later on when using the EDQ Open Tool within an ODI Package



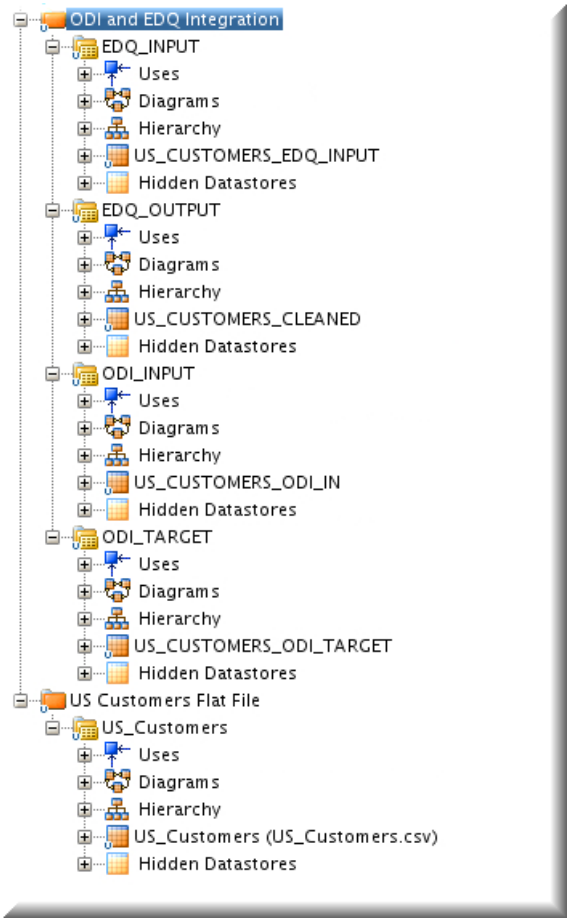
15. A **Project** is created within the **Designer** tab to hold the knowledge modules, mappings, and package for this sample data quality integration



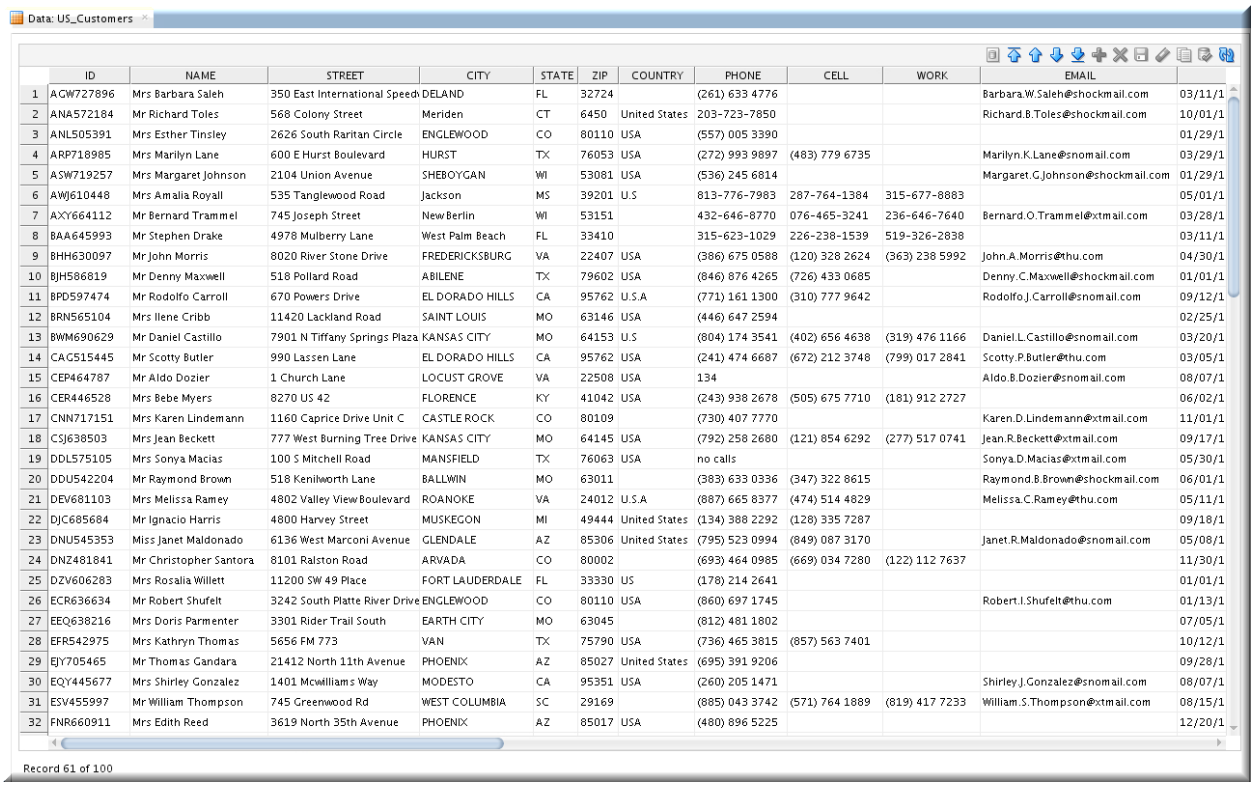
16. The following model folder and models were created within the **Designer** tab and **Models** section of ODI



17. Reverse-engineering was completed to generate the models required to create mappings in ODI



18. Right clicking the **US_Customers (US_Customers.csv)** model allows you to see the raw data that we will cleanse in later steps. Clicking **View Data** reveals the raw data



Data: US_Customers

	ID	NAME	STREET	CITY	STATE	ZIP	COUNTRY	PHONE	CELL	WORK	EMAIL	
1	AGW727896	Mrs Barbara Saleh	350 East International Speedv	DELAND	FL	32724		(261) 633 4776			Barbara.W.Saleh@shockmail.com	03/11/1
2	ANA572184	Mr Richard Toles	568 Colony Street	Meriden	CT	6450	United States	203-723-7850			Richard.B.Toles@shockmail.com	10/01/1
3	ANL505391	Mrs Esther Tinsley	2626 South Raritan Circle	ENGLEWOOD	CO	80110	USA	(557) 005 3390				01/29/1
4	ARP718985	Mrs Marilyn Lane	600 E Hurst Boulevard	HURST	TX	76053	USA	(272) 993 9897	(483) 779 6735		Marilyn.K.Lane@snomail.com	03/29/1
5	ASW719257	Mrs Margaret Johnson	2104 Union Avenue	SHEBOYGAN	WI	53081	USA	(536) 245 6814			Margaret.G.Johnson@shockmail.com	01/29/1
6	AWJ610448	Mrs Amalia Royall	535 Tanglewood Road	Jackson	MS	39201	U.S	813-776-7983	287-764-1384	315-677-8883		05/01/1
7	AXY664112	Mr Bernard Trammel	745 Joseph Street	New Berlin	WI	53151		432-646-8770	076-465-3241	236-646-7640	Bernard.O.Trammel@xtmail.com	03/28/1
8	BAA645993	Mr Stephen Drake	4978 Mulberry Lane	West Palm Beach	FL	33410		315-623-1029	226-238-1539	519-326-2838		03/11/1
9	BHH630097	Mr John Morris	8020 River Stone Drive	FREDERICKSBURG	VA	22407	USA	(386) 675 0588	(120) 328 2624	(363) 238 5992	John.A.Morris@thu.com	04/30/1
10	BJH586819	Mr Denny Maxwell	518 Pollard Road	ABILENE	TX	79602	USA	(846) 876 4265	(726) 433 0685		Denny.C.Maxwell@shockmail.com	01/01/1
11	BPD597474	Mr Rodolfo Carroll	670 Powers Drive	EL DORADO HILLS	CA	95762	U.S.A	(771) 161 1300	(310) 777 9642		Rodolfo.J.Carroll@snomail.com	09/12/1
12	BRN565104	Mrs Ilene Cribb	11420 Lackland Road	SAINT LOUIS	MO	63146	USA	(446) 647 2594				02/25/1
13	BWM690629	Mr Daniel Castillo	7901 N Tiffany Springs Plaza	KANSAS CITY	MO	64153	U.S	(804) 174 3541	(402) 656 4638	(319) 476 1166	Daniel.L.Castillo@snomail.com	03/20/1
14	CAG515445	Mr Scotty Butler	990 Lassen Lane	EL DORADO HILLS	CA	95762	USA	(241) 474 6687	(672) 212 3748	(799) 017 2841	Scotty.P.Butler@thu.com	03/05/1
15	CEP464787	Mr Aldo Dozier	1 Church Lane	LOCUST GROVE	VA	22508	USA	134			Aldo.B.Dozier@snomail.com	08/07/1
16	CER446528	Mrs Bebe Myers	8270 US 42	FLORENCE	KY	41042	USA	(243) 938 2678	(505) 675 7710	(181) 912 2727		06/02/1
17	CNN717151	Mrs Karen Lindemann	1160 Caprice Drive Unit C	CASTLE ROCK	CO	80109		(730) 407 7770			Karen.D.Lindemann@xtmail.com	11/01/1
18	CSJ638503	Mrs Jean Beckett	777 West Burning Tree Drive	KANSAS CITY	MO	64145	USA	(792) 258 2680	(121) 854 6292	(277) 517 0741	Jean.R.Beckett@xtmail.com	09/17/1
19	DDL575105	Mrs Sonya Macias	100 S Mitchell Road	MANSFIELD	TX	76063	USA	no calls			Sonya.D.Macias@xtmail.com	05/30/1
20	DDU542204	Mr Raymond Brown	518 Kenilworth Lane	BALLWIN	MO	63011		(383) 633 0336	(347) 322 8615		Raymond.B.Brown@shockmail.com	06/01/1
21	DEV681103	Mrs Melissa Ramey	4802 Valley View Boulevard	ROANOKE	VA	24012	U.S.A	(887) 665 8377	(474) 514 4829		Melissa.C.Ramey@thu.com	05/11/1
22	DJC685684	Mr Ignacio Harris	4800 Harvey Street	MUSKEGON	MI	49444	United States	(134) 388 2292	(128) 335 7287			09/18/1
23	DNU545353	Miss Janet Maldonado	6136 West Marconi Avenue	GLENDALE	AZ	85306	United States	(795) 523 0994	(849) 087 3170		Janet.R.Maldonado@snomail.com	05/08/1
24	DNZ481841	Mr Christopher Santora	8101 Ralston Road	ARVADA	CO	80002		(693) 464 0985	(669) 034 7280	(122) 112 7637		11/30/1
25	DZV606283	Mrs Rosalia Willett	11200 SW 49 Place	FORT LAUDERDALE	FL	33330	US	(178) 214 2641				01/01/1
26	ECR636634	Mr Robert Shufelt	3242 South Platte River Drive	ENGLEWOOD	CO	80110	USA	(860) 697 1745			Robert.I.Shufelt@thu.com	01/13/1
27	EEQ638216	Mrs Doris Parmenter	3301 Rider Trail South	EARTH CITY	MO	63045		(812) 481 1802				07/05/1
28	EFR542975	Mrs Kathryn Thomas	5656 FM 773	VAN	TX	75790	USA	(736) 465 3815	(857) 563 7401			10/12/1
29	EJY705465	Mr Thomas Gandara	21412 North 11th Avenue	PHOENIX	AZ	85027	United States	(695) 391 9206				09/28/1
30	EQY445677	Mrs Shirley Gonzalez	1401 Mcwilliams Way	MODESTO	CA	95351	USA	(260) 205 1471			Shirley.J.Gonzalez@snomail.com	08/07/1
31	ESV455997	Mr William Thompson	745 Greenwood Rd	WEST COLUMBIA	SC	29169		(885) 043 3742	(571) 764 1889	(819) 417 7233	William.S.Thompson@xtmail.com	08/15/1
32	FNR660911	Mrs Edith Reed	3619 North 35th Avenue	PHOENIX	AZ	85017	USA	(480) 896 5225				12/20/1

Record 61 of 100

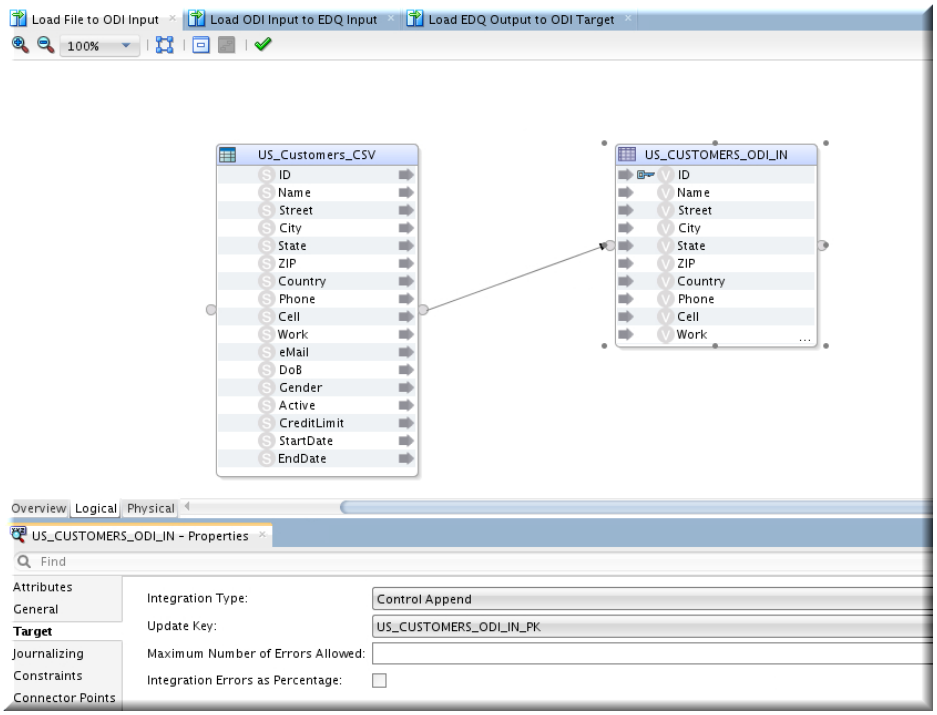
Create Mappings in ODI

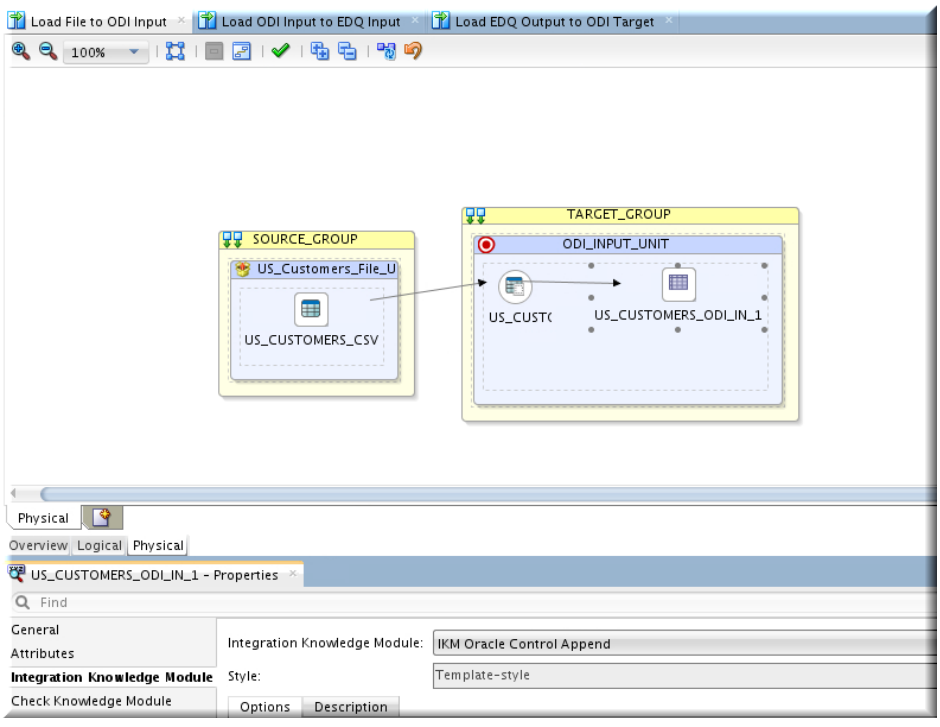
Three mappings will be created in the next section to enable loading the file to database tables.

- The first coordinates movement from the raw data, a comma-separated-values file **US_Customers.csv** to an Oracle table, **US_CUSTOMERS_ODI_IN**.
- The second uses the ODI Input, **US_CUSTOMERS_ODI_IN** as the source to load a the EDQ Input table **US_CUSTOMERS_EDQ_INPUT**.
- The last uses the EDQ Output table **US_CUSTOMERS_CLEANED** to load the ODI Target table **US_CUSTOMERS_ODI_TARGET**.
- Each mapping will use the Control Append integration technique.

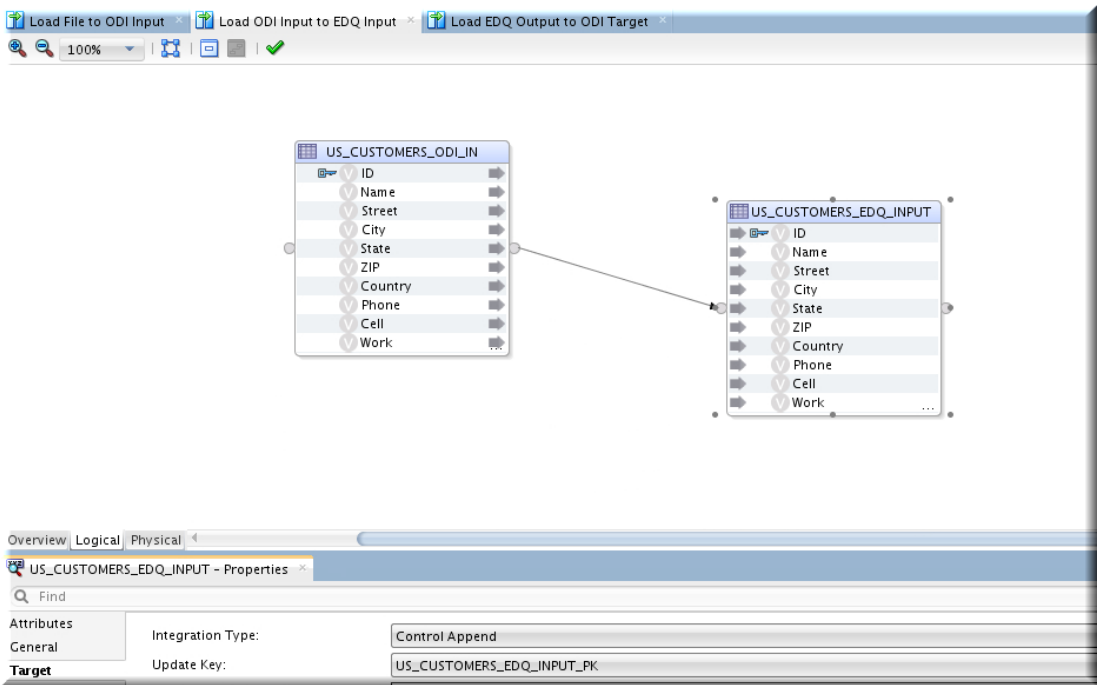
It is worth noting that although there is no direct mapping in ODI to load the EDQ Output table **US_CUSTOMERS_CLEANED**, EDQ will write data to that table after successfully completing a data quality job that ODI will invoke.

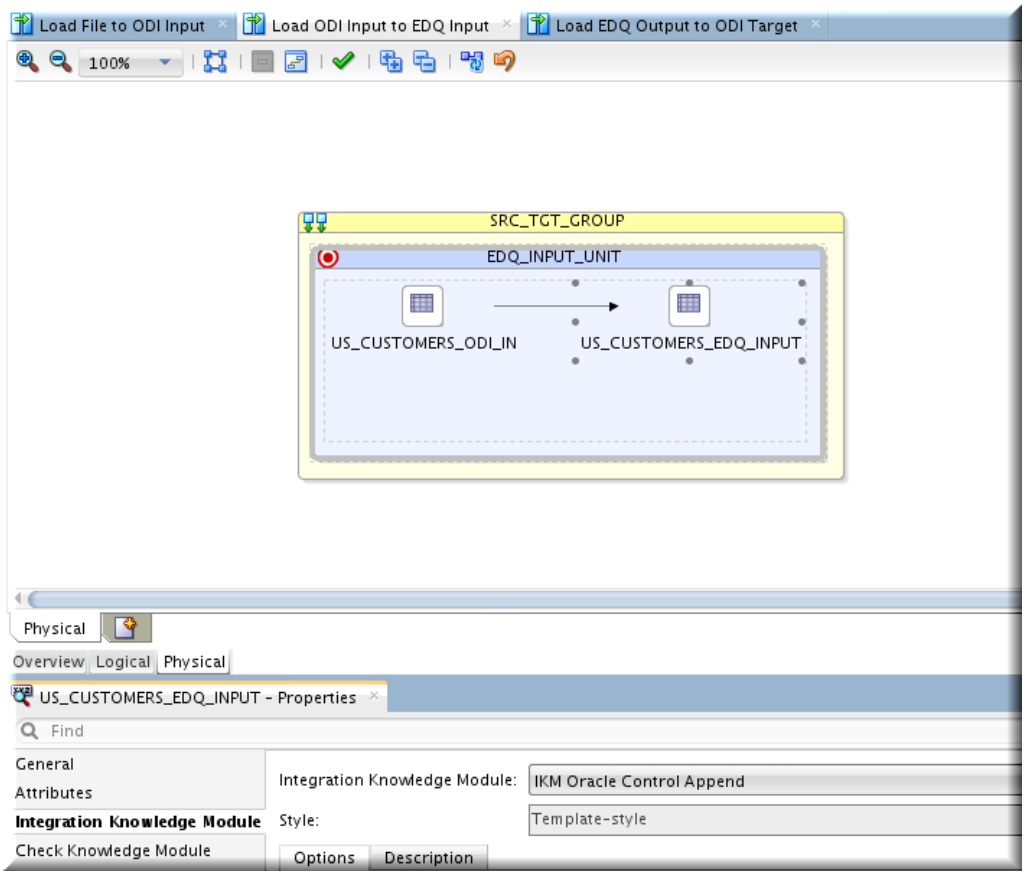
19. The model **US_CUSTOMERS_ODI_IN** serves as the ODI Input for EDQ and is loaded from the **US_Customers** file. A mapping ***Load File to ODI Input*** was created. Since the columns between the file and table are identical in this example, the auto-map option was used.



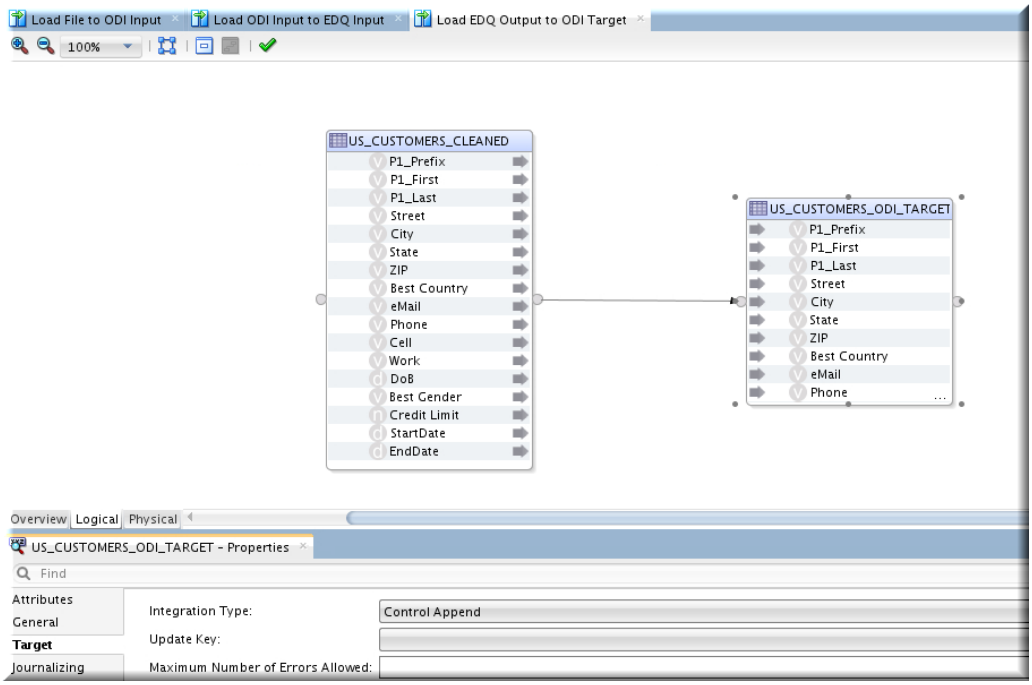


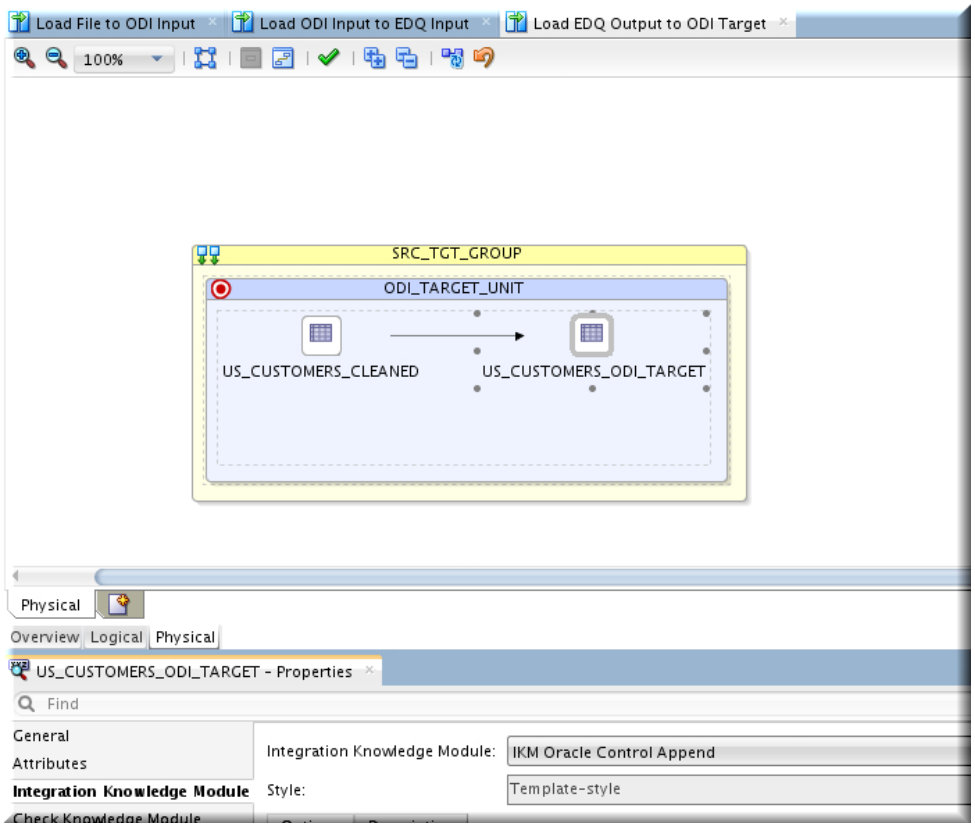
20. Another mapping ***Load ODI Input to EDQ Input*** is created as shown below





21. Lastly, a mapping ***Load EDQ Output to ODI Target*** is created as shown below

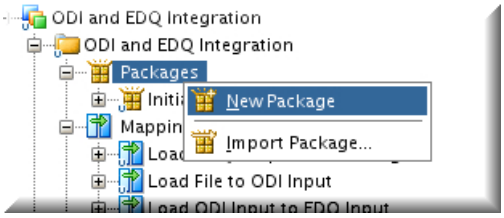




Create an ODI Package

The ODI Package will serve as a means to automating the tasks that take place during the execution of each mapping. Additionally, the Open Tool for Enterprise Data Quality will be utilized in the package to call an existing EDQ job.

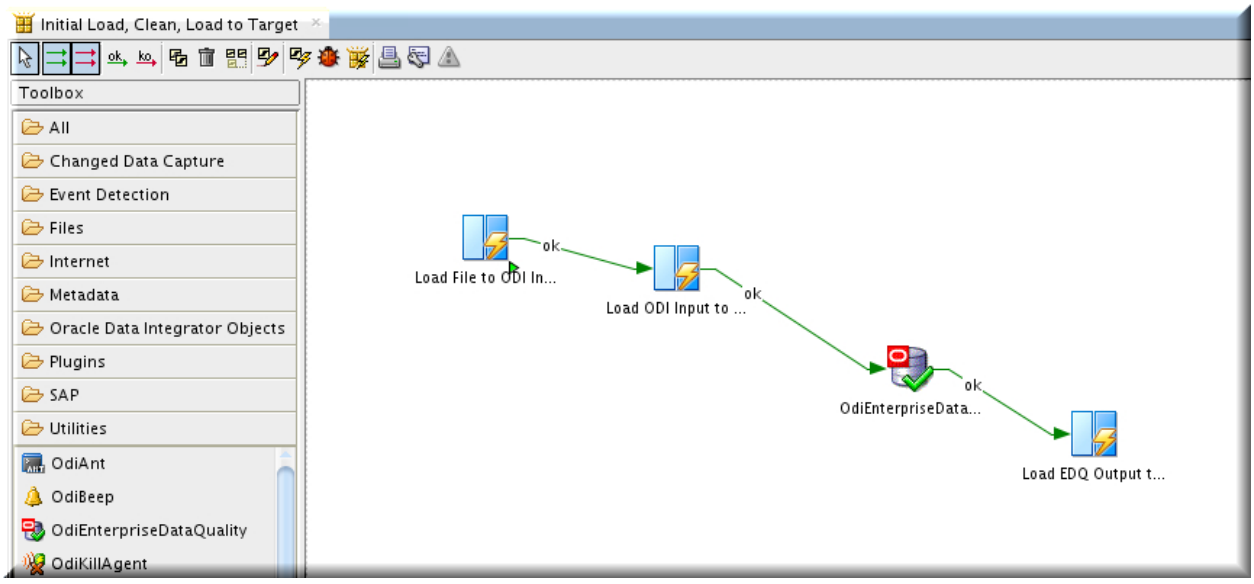
22. Right clicking on **Packages** within the **Designer** tab and **Projects** section reveals the option to create a **New Package**



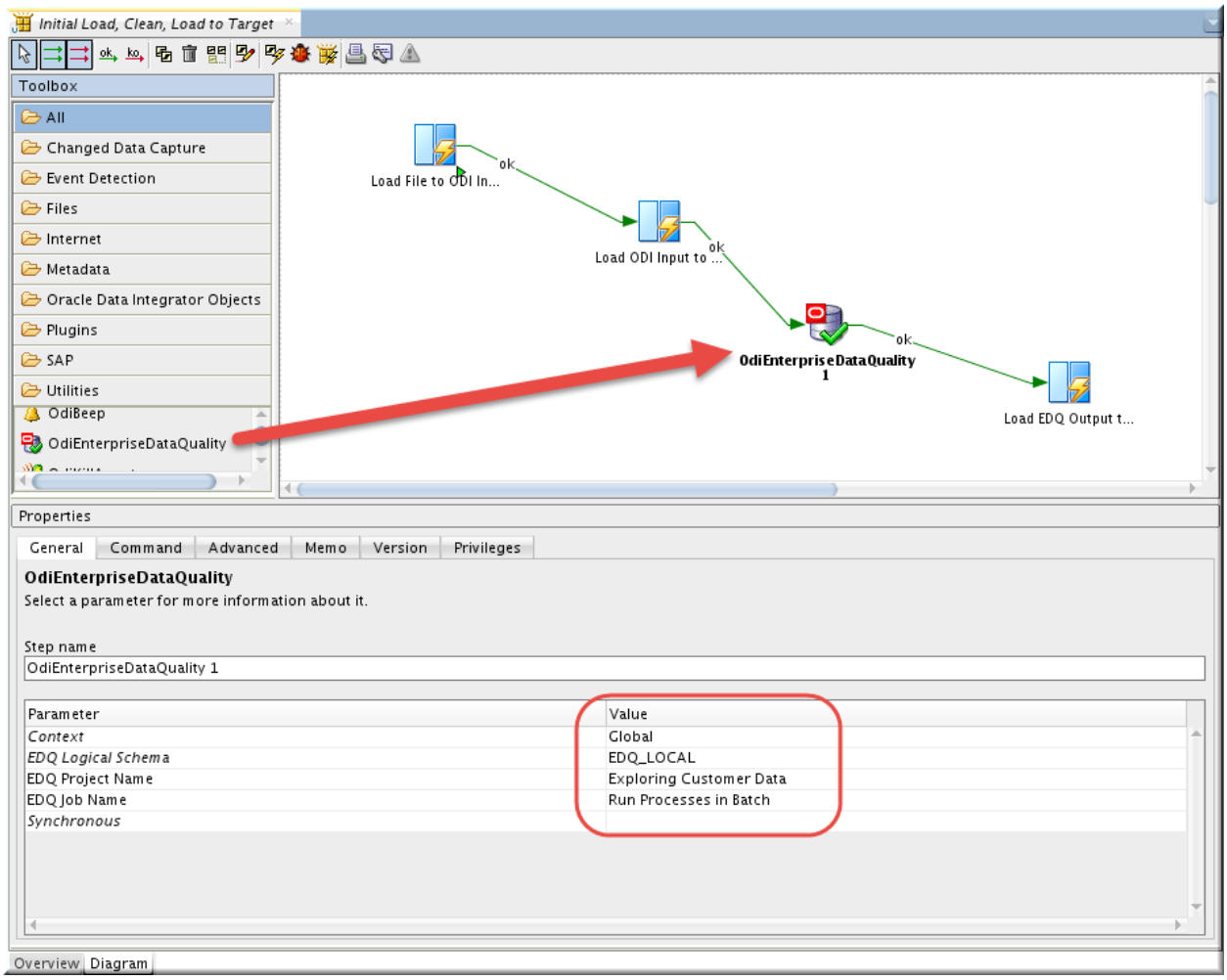
23. The three mappings from the **Designer** tab are dragged and dropped onto the canvas in the **Package** tab. In this case, the package created was named ***Initial Load, Clean, Load to Target***

24. The **OdiEnterpriseDataQuality** Open Tool was dragged and dropped onto the canvas

25. Lastly the **OK→**  option was selected to create a sequence to execute the mappings and invoke the EDQ job on successful completion of each step as shown in the screenshot below



26. Clicking on the **OdiEnterpriseDataQuality** icon in the canvas reveals the properties window towards the bottom of **ODI Studio**

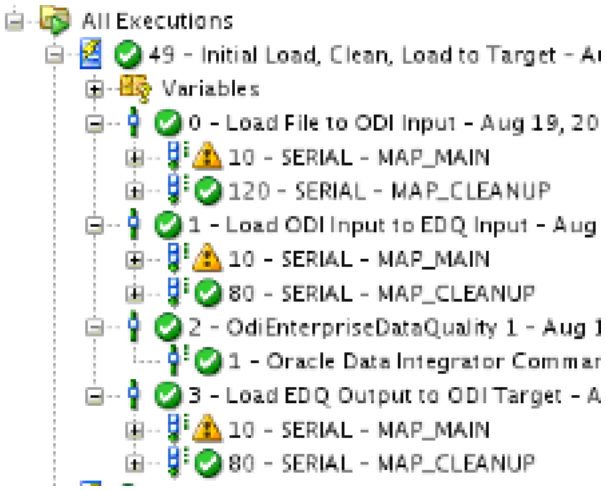


27. There are 6 **Parameters** to fill in to successfully invoke a job from EDQ from an ODI Package. The **Synchronous** option will use the default value (Yes), so it can be left blank

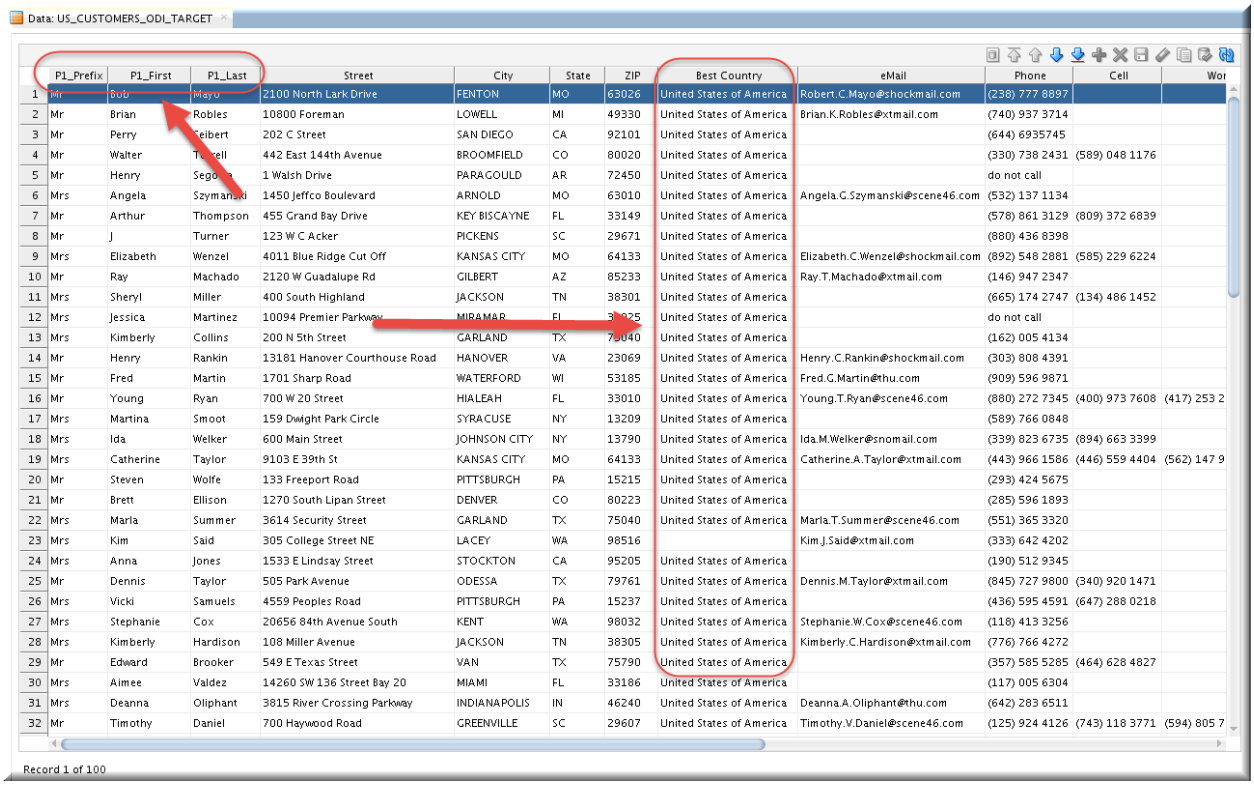
Context	***ODI Execution Context from ODI Topology***
EDQ Logical Schema	***EDQ Logical Schema from ODI Topology***
EDQ Project Name	***EDQ Project Name***
EDQ Job Name	***EDQ Job name from EDQ Project***

28. After the parameters are entered. Right click the **Package** (Initial Load, Clean, Load to Target) to start the process of loading the .CSV file to staging tables. Since the OK→  arrows were used for each mapping and the open tool, as long as each mapping is

successful, each step will continue. The execution can be observed within the **Operator** tab



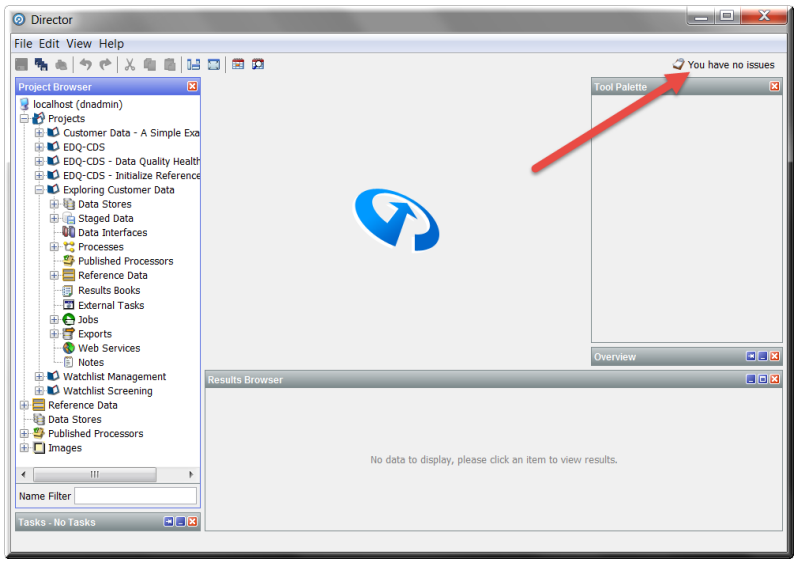
29. Return to the **Designer** tab and the **Models** section. Right clicking the ODI Target **US_CUSTOMERS_ODI_TARGET** allows you to view the cleansed data. Notice the column headers retain the metadata generated by Enterprise Data Quality. Additionally, the column for country is standardized to a uniform set of values



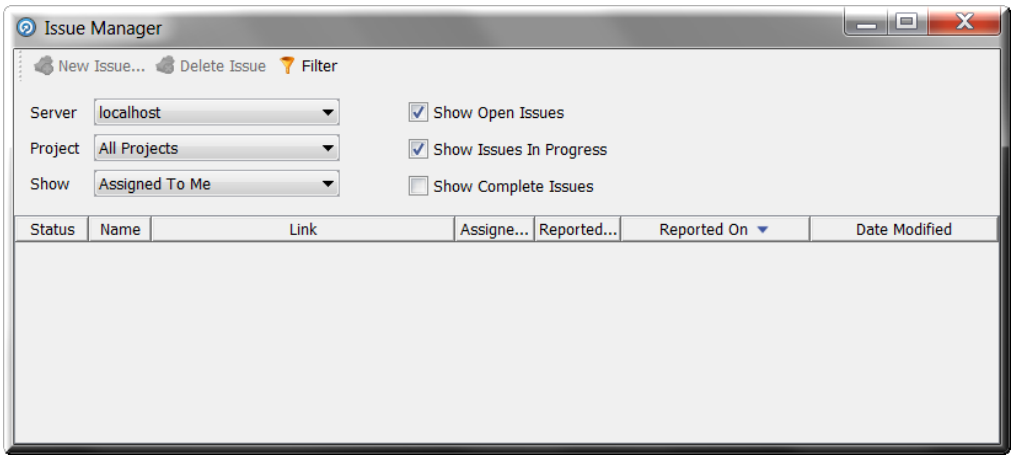
Lab 6: Issue Management

Issue Management in EDQ allows you to have a means to keeping track of data quality problems that need to be addressed. Issues can be assigned to you, another user, or a group for investigation and resolution. The Issue Manager console can be found within the Director application or via the EDQ Launchpad

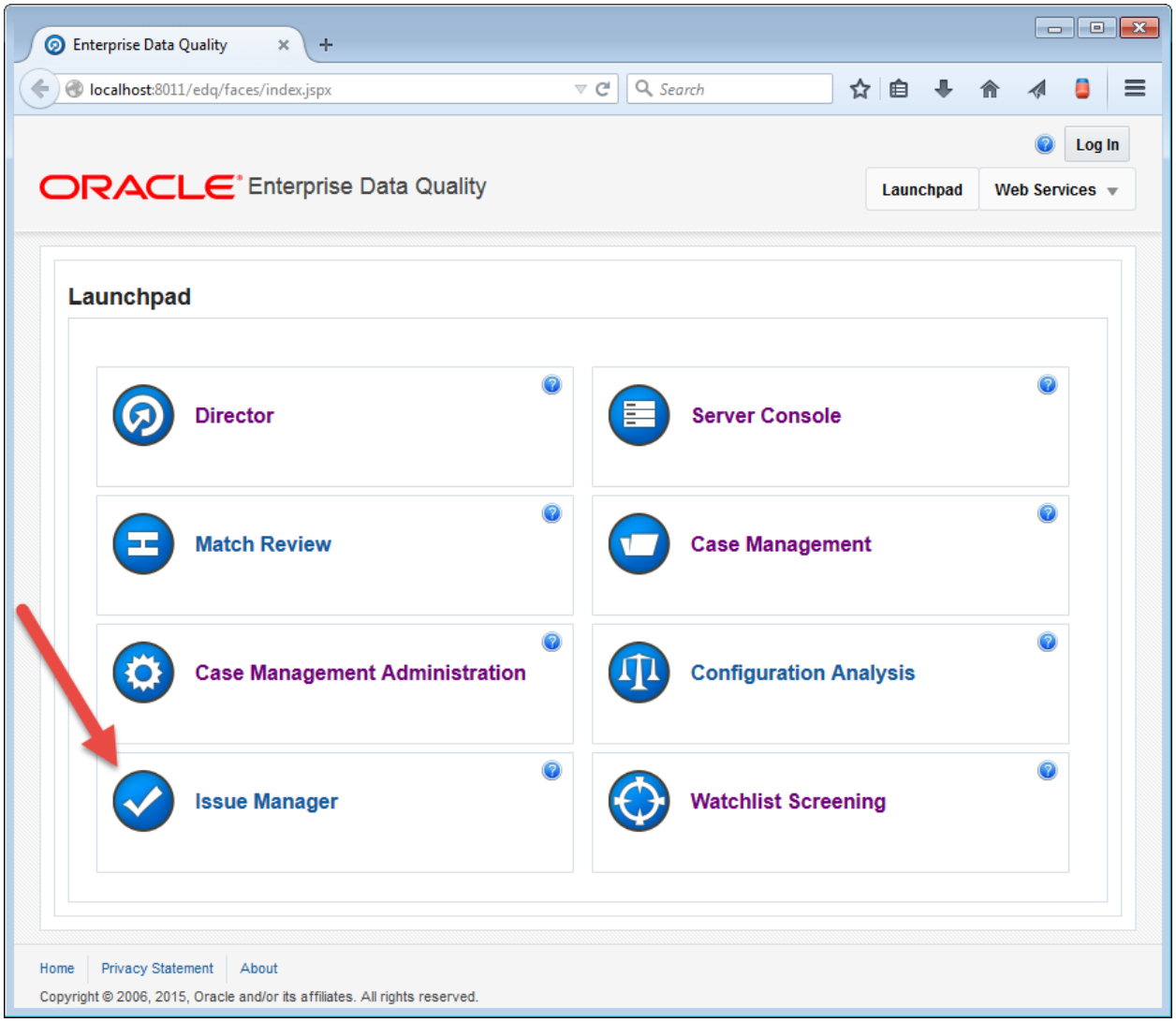
1. Navigate to the **Issue Manager** within the EDQ Director application. Find the  icon on the upper-right corner of Director. Depending on who you assigned the issues created in the previous lab, there may be a **message stating you have some issues waiting**



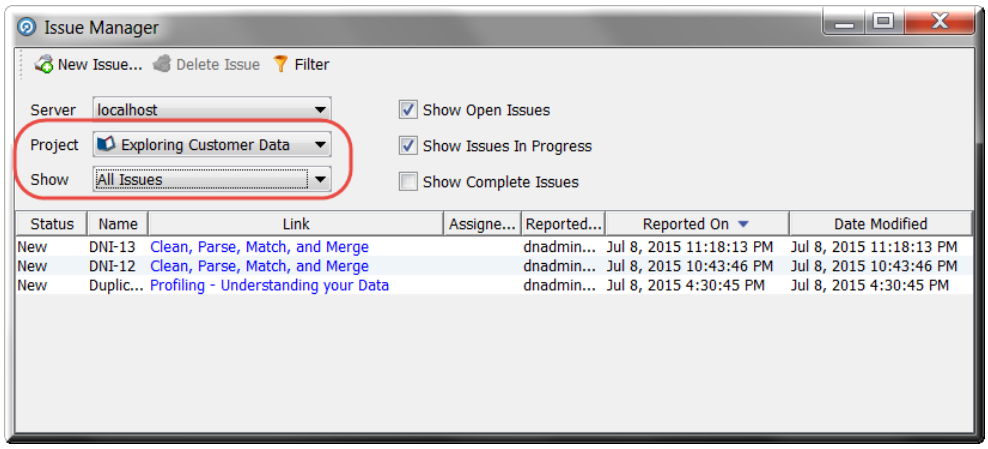
2. Take a moment to explore the interface of the **Issue Manager**



The Issue Manager was just opened using the Director application, but for users solely responsible for issue management, the application can also be opened via the EDQ Launchpad (can be accessed at http://localhost:8011/edq/faces/ in this lab)



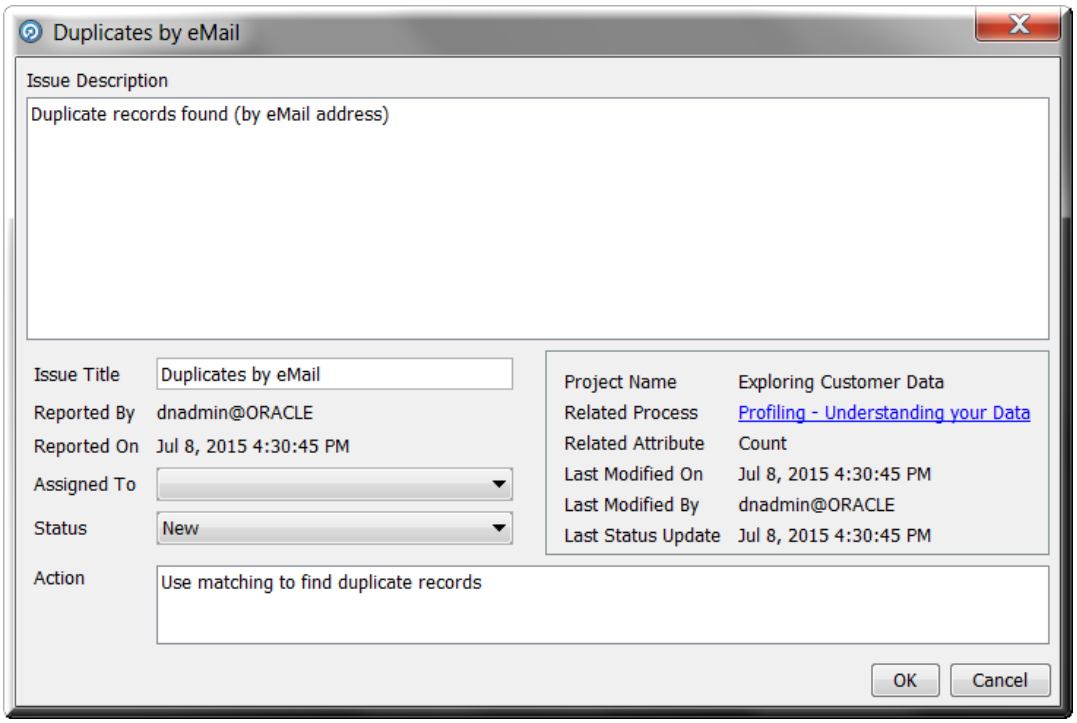
3. In the Issue Manager, begin by changing the drop-downs for **Project** to **Exploring Customer Data** and **Show** to **All Issues**



The Issue Manager will display issues by Status, Name, Link (from where the issue was opened), Assignment, User who opened the issue, Reported Date and the last time the issue was modified.

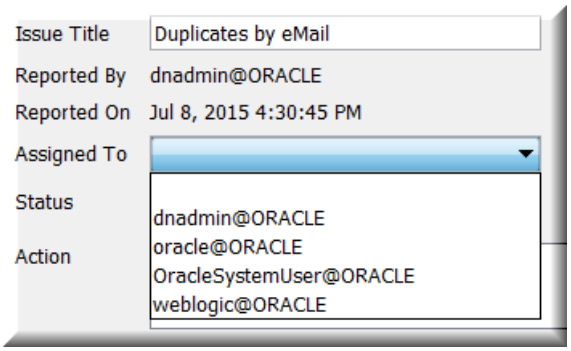
4. Open an issue by double-clicking on a record beneath the **Name** column in the **Issue Manager**

 The **Link** column is an active hyperlink that will return the Director application to the location where the issue was opened. You can also open an issue by right-clicking a record and clicking **Open Issue**.

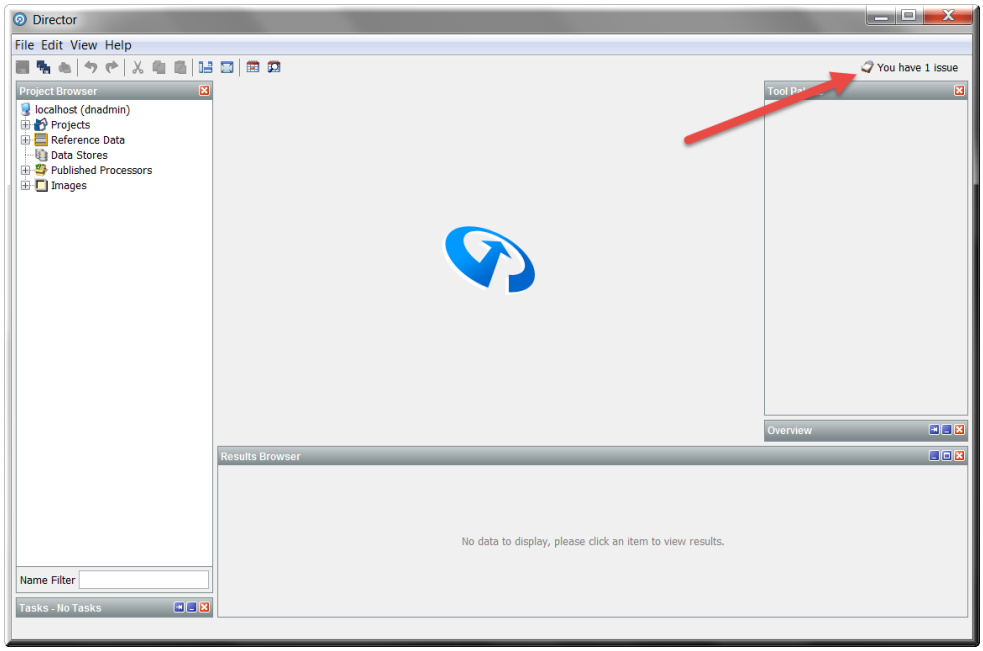


The **Issue Title** reflects the Name column from the Issue Manager. Also, an action item can be noted at the bottom for the individuals assigned to resolve the issue. Additional detail is provided on the bottom right corner of the issue window. Note that the Project Name, Related Process, and Related Attribute is recorded.

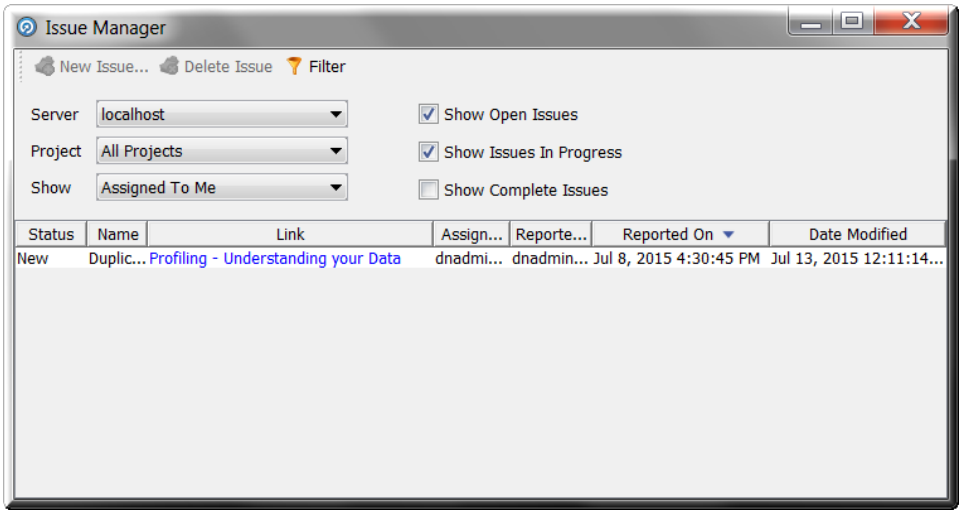
5. Select the drop-down titled **Assigned To** and assign the issue to **dnadmin@ORACLE**, click **OK** to continue



At this point, you will have at least 1 issue assigned to the user you are logged in as. Note the **top right corner of the Director** reflects this.

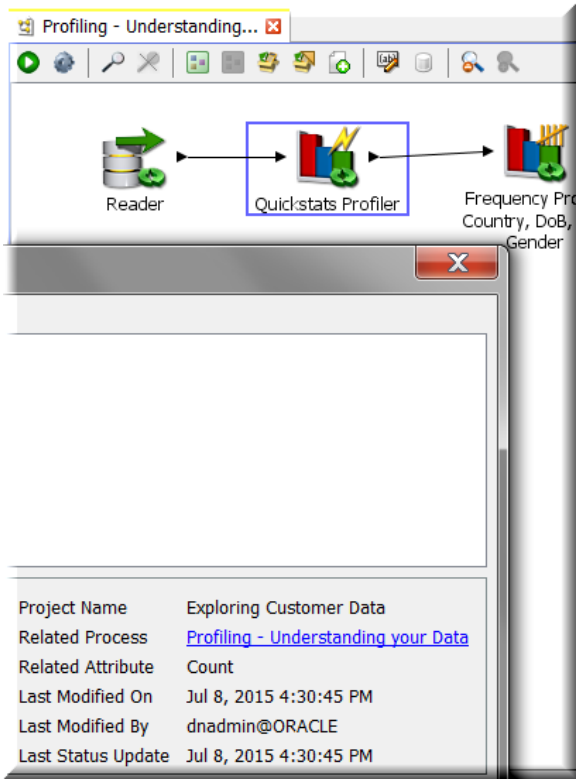


6. Return to the **Issue Manager** by clicking the icon or the message stating you have issues. Notice by default the filters for **Project** and **Show** are defaulted to **All Projects** and **Assigned to Me**. This ensures that when you log in, you are seeing all the important issues you are assigned to resolve



If you click **Filter** at the top of the Issue Manager, you will notice that the drop-down and other options on the right side of the Issue Manager are hidden.

7. Click the hyperlink under the **Link** column, and return to the Director. You will notice that it brings you to your project and the **Profiling – Understanding your Data** process. Furthermore, notice that the **Quickstats Profiler** is already selected because that is where the issue was created



8. The **Results Browser** may state 'This process needs to be run before viewing results. Click the run button in the upper-left corner of the Director window to run the process. Afterwards, the issue can be further investigated by the user by drilling down into the results found in the Results Browser

As users begin to investigate and fix the data quality issues raised using the Issue Manager functionality, it is important to change the **Status** to **In Progress** or **Complete** so that the EDQ users can remain productive resolving issues that are New or unassigned.

The Issue Management feature of EDQ allows your users to be collaborative when working on Data Quality initiatives and projects. In addition to the functionality found within the Director or Issue Manager, EDQ can also be configured to notify the user when new issues are raised.

Tips and Tricks for Deployment and using Oracle Data Integrator and EDQ together

In this section, you will review some best practices related to the usage and deployment of Enterprise Data Quality. In addition to the best practices, a short list of resources can be found below.

- EDQ has a large number of properties that can be modified to fine-tune the environment. The full guide can be found here
  - http://docs.oracle.com/middleware/1213/edq/DQSAG/perf_tuning.htm

Many tuning controls are explained in detail in the documentation linked above, but two of the most common are discussed below:

- One of the common adjustments many environments follow is to modify the Client Heap Size due to issues that may occur when for example you are attempting to export a large amount of data to a client-side Excel file

  - Find the `blueprints.properties` file in the `config/properties` directory of the EDQ server
  - You will see `*.jvm.memory = 512m` within the file already. The memory allocation can be changed to affect all client applications (i.e. Director, Issue Manager, etc.) To modify the heap size for a given client application, replace the * with `director` to change it for Director or `issues` to change it for Issue Manager
  - The complete list of client applications can be found in the link above for fine-tuning the environment

- EDQ has a voracious disk space appetite during runtime. Be sure to set your database table tablespace sizes appropriately to take into account a data size inflation factor of 10x to up to 25x. So, if your source data is 2GB, you will need between 20 and 50GB of tablespace when running your EDQ process (particularly if a given Process is using Profiler processes)

- In the case that the database server is more powerful than the machine hosting the EDQ application server, it may be worth increasing the `workunitexecutor.outputThreads` parameter to modify the number of threads (and database connections) used when writing results and staged data to the database
  - This parameter is found in the `director.properties` file in the `oedq_local_home` configuration directory

Additional Resources

- Oracle Enterprise Data Quality Learning Library
  - http://oracle.com/oll - Click Search and use the drop-down menus to adjust the Search Filter to:
    - Product Family – Fusion Middleware
    - Product – Enterprise Data Quality
- Oracle Docs – Oracle Enterprise Data Quality 12.2.1
  - http://docs.oracle.com/middleware/1221/edq/index.html
- Oracle Enterprise Data Quality Home Page
  - http://www.oracle.com/technetwork/middleware/oedq/overview/index.html
- Oracle Data Integration Blogs
  - https://blogs.oracle.com/dataintegration/