## IDC3931 Data Quality Lab Exercise

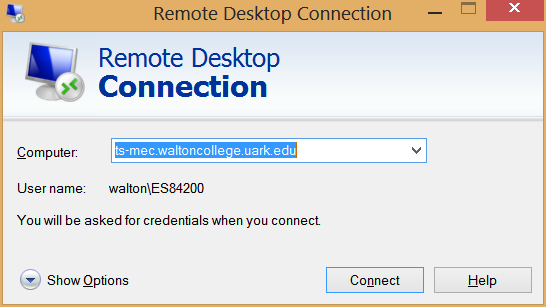
The purpose of this lab is to explore some of the functionality of one of the most powerful data cleansing tools on the market today. It is a component of SQL Server 2012 and is called Data Quality Services.

Personal Laptop: If you have a Windows machine or a Virtual Machine running Windows:

In order to gain access to this SQL Server component, you can either install the BI, or Enterprise, or Develop editions of SQL Server 2012. After installing SQL Server 2012 you will then need to run the Data Quality Server Installer to enable it or we can utilize our relationship with Microsoft in order to do this lab and the subsequent homework. This will ask you for a password that you can set to anything with >= 8 characters and it does not appear that you will ever need to remember that password (but I could be wrong on that).

Classroom Access: In this lab we will utilize the resources available to us because of our role in the Microsoft Academic Alliance. Similar to how we performed our first homework by connecting to Teradata servers at the University of Arkansas, we will be connecting to their Microsoft database servers.

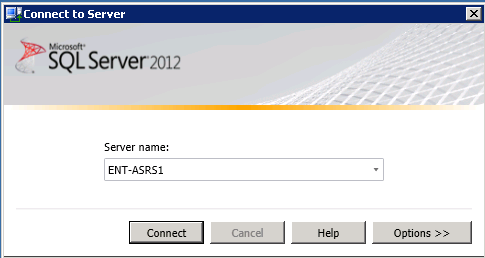
In order to connect to our correctly configured SQL Server 2012 database server, you can use your previously configured Teradata remote desktop connection and just change the TS-TUN to TS-MEC in your existing connection and then **Connect**. See the following screenshot. If you need the document (Microsoft Remote Connection Instructions.docx, showing you how to set this up, you can download it from the in our Blackboard course (Bb>Weekly Tasks/Course Mat'l> Remote Connections Tutorial). The setup for how we did the Teradata server is identical to how we connect to the Microsoft servers.



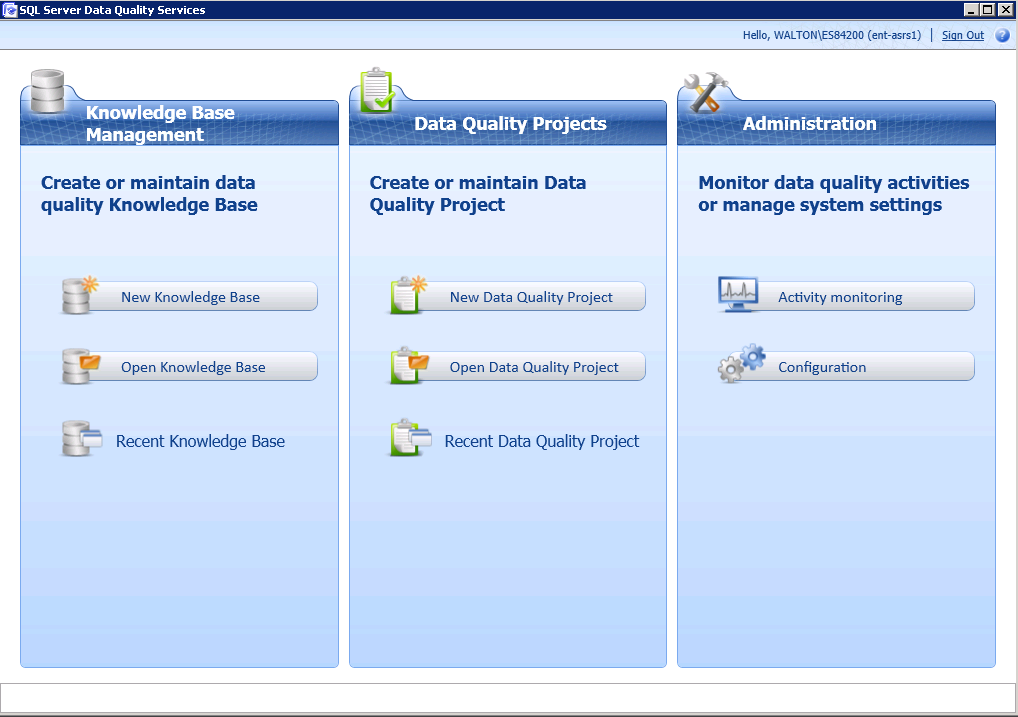
Download the Excel file for this lab and copy it onto your University of Arkansas desktop.

After connecting to your remote machine, Click on START/All Programs/Microsoft SQL Server 2012/Data Quality Services/Data Quality Client.

After the program starts, it will ask you for the name of the server that you will be working on. Enter the following per the following screenshot.



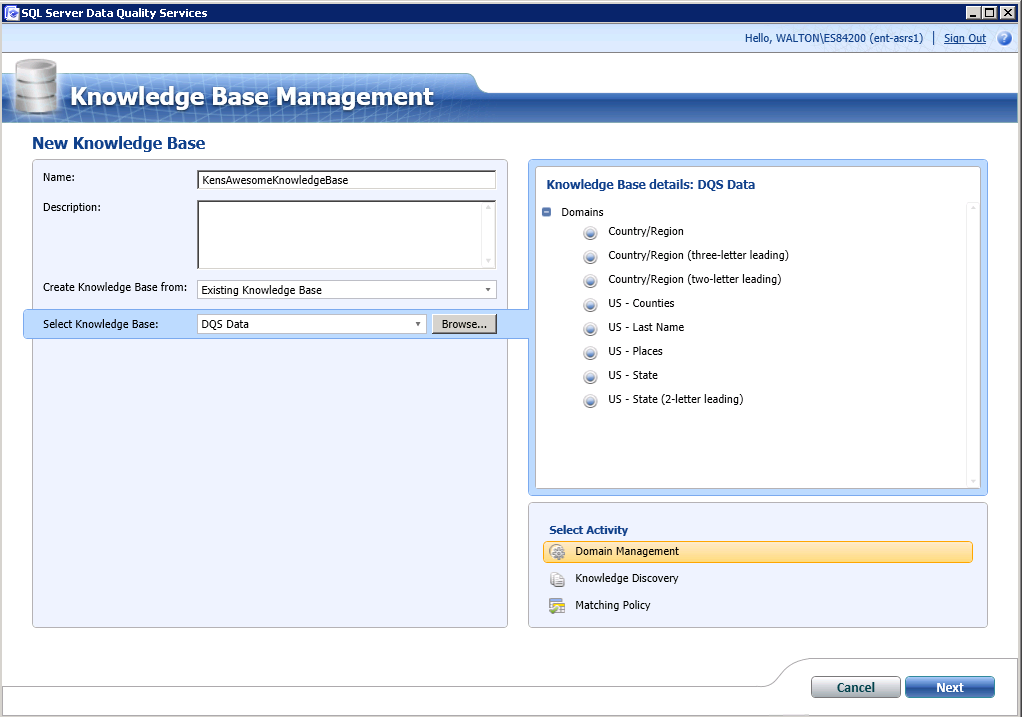
You should be taken to the following screen:



## Creating a Knowledge Base

Knowledge bases are used to house all of the rules about what is good data, what is bad data and the rules that govern how to convert the bad data into good data. When you create a new knowledge base, you can do so from scratch, or you can use an existing knowledge base as a starting point. SQL Server 2012 ships with a pre-existing knowledge base for US-based demographic data named **DQS Data**, and in this example, we will use it as the basis for my own KensAwesomeKnowledgeBase as shown below. Click on “New Knowledge Base”



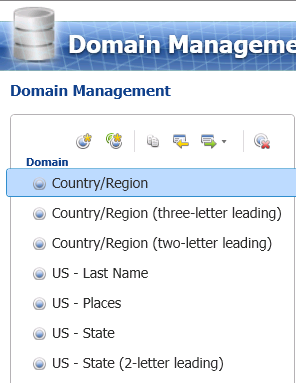
You will then give your knowledge base a name. Please use your last name and then your first name. Please then select **Existing Knowledge Base** and **DQS Data**. 

Before continuing, download/copy the lab’s Excel data file onto the desktop of your remote SQL Server machine. You can do this from the browser on your remote server or your local machine and then copy/paste it from your remote desktop onto the remote server’s desktop… either way works fine. At the end of this lab, you will need a file to output all of your newly-cleansed data into so make a new spreadsheet on your desktop and name it “Cleansed Data”

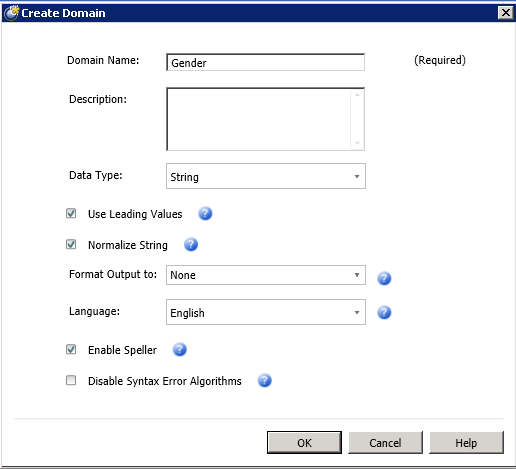
The **DQS Data** knowledge base supplied with SQL Server includes a number of pre-defined domains, as shown in the image above. You can think of these domains as columns that contain the values of good data and bad data with the rules that map the bad into good. You will only need some of these domains for this lab, and you will need to add some additional ones that are specific to your own data for your next HW Assignment.

Under the **Select Activity** (lower right area of the page) you will initially select the **Domain Management** activity as I create the new knowledge base and then “Next”. Your data file does not contain any County information so you can delete that value so delete the “US – Counties” domain. The domains that you are retaining contain official values for country and US state names, and common last names (surnames) based on US demographic data such as the 2000 US census.

You know that there is wrong/inconsistent data in a couple of columns that we need to cleanse. For each column we want to clean up we will need to add a domains to your knowledge base in order to cleanse the two “dirty” columns in our source data file: Gender and Total Children. In order to add a new domain to your knowledge base (KB) click on the “**Create A Domain”** button



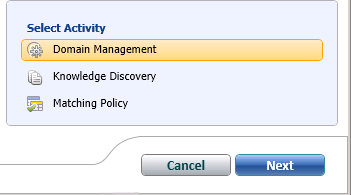
and then filling out the resulting popup window:



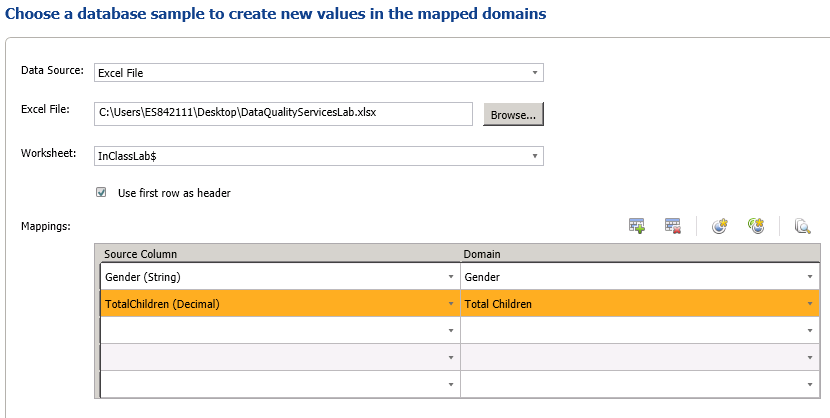
Note that if you are adding Domains that contains Text data like Gender you should choose the **Data Type** of **String**. If you are adding numeric values use decimal and for dates, you should choose the Date data type. Please note that you need to get this data type right on the first try for a domain. If you create the domain with the string data type and later learn that it should be a number, you will have to delete it and start ALL over. i.e. there is no opportunity to edit it after the fact!

For Number of Children, please name the new domain “Total Children” with **Decimal Data Type.** You may be tempted to create this one as an integer since the only numbers in your spreadsheet are 0,1,2 etc but please use the Decimal Data Type for all of your numbers in this exercise/lab

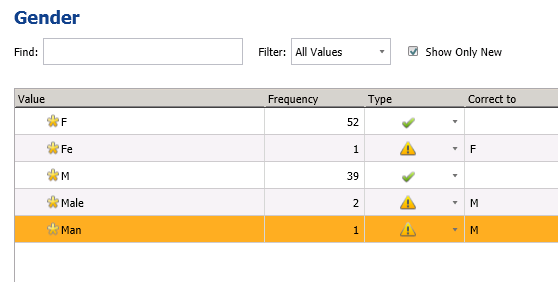
After adding the two new domains, click **Finish** and then **Publish** your updated knowledge base. You will now need to populate the two new domains with values from the data from your data file. From the main menu, click on **Open Knowledge Base** and then select your KB and then **Knowledge Discovery** under the Select Activity header:



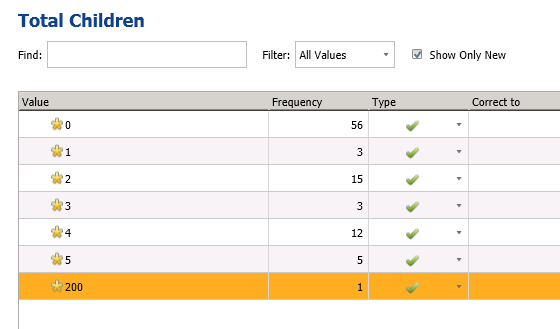
The first thing you want to do is find out all of the values that exist in your data for the Gender and Number of Children columns. We do this by letting SQL Server go through your data and find all of the different values. In order to do this, change the Data Source from SQL Server to Excel file and then populate the fields with the location of your data file (that you hopefully copied onto your desktop a couple of paragraphs ago) as well as which worksheet in that data file. Then map the Gender and Total Children columns from your data file to the new created but sadly empty Gender and Total Children domains.



Click on **Next** and then on the next page click **Start** to discover all values. After the knowledge discovery process is complete, click on one of the domains and you will then be presented with a table that lists all of the unique values from your data. The next step is to determine the right values and the wrong ones that need to be corrected to the right ones. You do this by marking all of the incorrect ones as Invalid (with the yellow warning sign) and then filling in the **Correct to** field with the correctly mapped value.



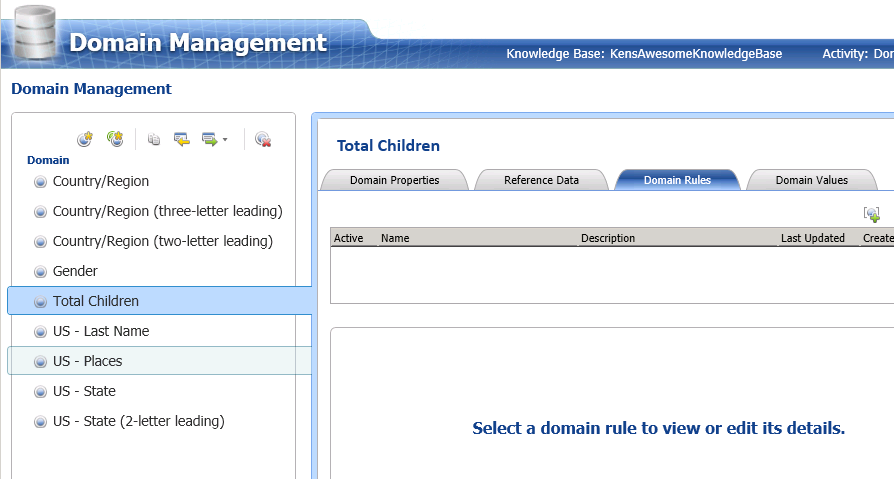
After you have done this for the first Domain (Gender in the above example) then go back and do a knowledge discovery process to see what values are in your dirty Excel file for the other new domain, Total Children. You should get something that looks like this:



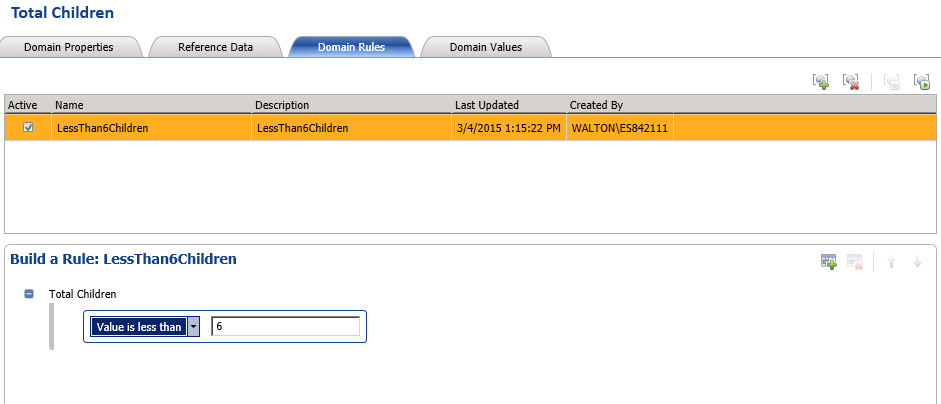
It looks like you have one customer with 200 total children. That is probably a BAD data value. You can either set a rule here to correct it to some value like 5 or something but there is a better way. If you set this 200 to 5 that will take care of the number 200 if it is ever encountered again but what if a 199 or 198 comes in. Wouldn’t it be a better idea to come up with a more universal rule like “if ANY value is over 5, just set it to 5?

Let’s do that! BUT to do that we need to go back to our domain management and set up a more universal rule in our knowledge base. Finish and Publish your KB and then go back and open your KB in Domain Management mode.

Highlight “Total Children” then click on the “Domain Rules” tab and then the plus sign to add a rule.

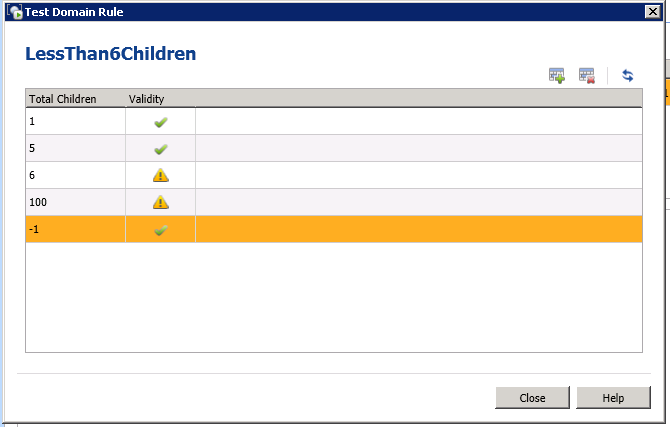


Create a rule that only allows numbers < 6 by filling in the Name and Description in the top section and then selecting “value is less than” and setting the value to 6.

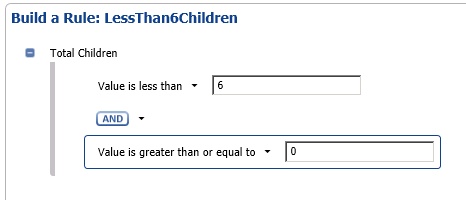


Now test out this rule with dummy test data by clicking on this icon in the upper right area of your domain rules area. Put in test values of 1, 5, 6, 100 and -1 and then test these values by click on the 2 circling arrows in the upper right corner of your test data screen.

You should get the following output:

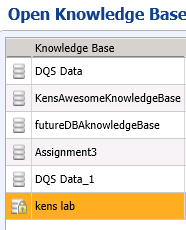


It looks like the positive values are testing out correctly but the current rule is also allowing negative values to be valid. What does a negative child mean? I OWE them a child? Yikes! It sounds like we need to add a condition to make all values < 0 to be invalid as well. Let’s go back to the Build a Rule section and add the additional portion of the rule per the following:



Re-run your test cases to verify that your new rule acts as you hope! If all is well then Finish and Publish your KB.

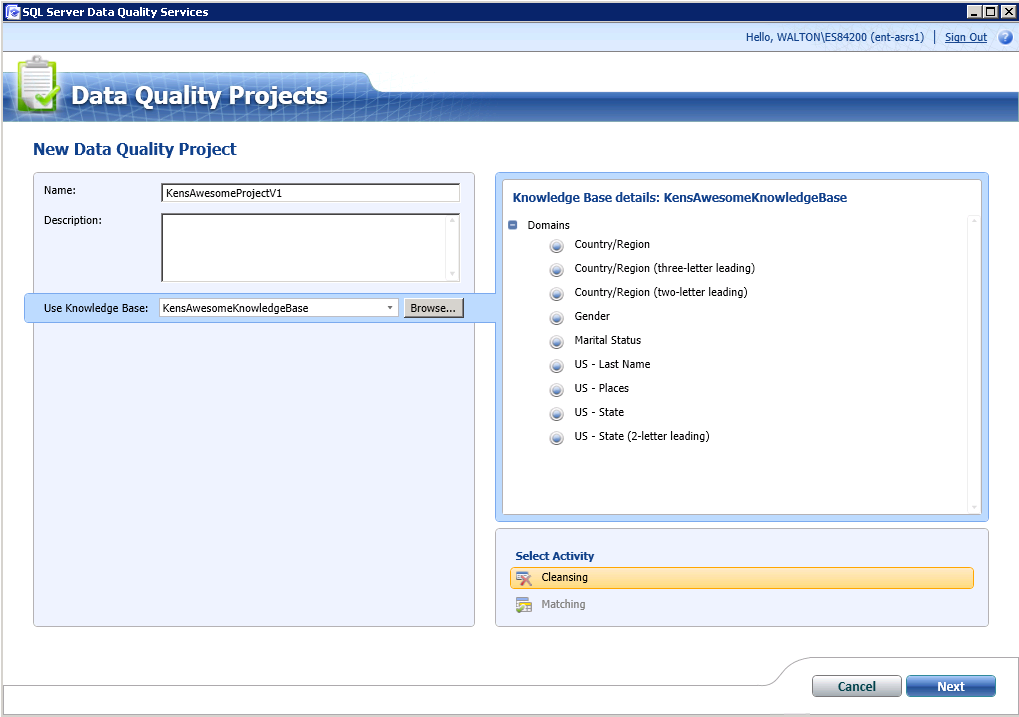
Note that prior to using your newly created and customized KB, it must be in a **published** state. Sometimes you will look out on the “**Open Knowledge Base**” screen and the icon in front of your KB looks different from everyone else’s and may have a small lock on it.



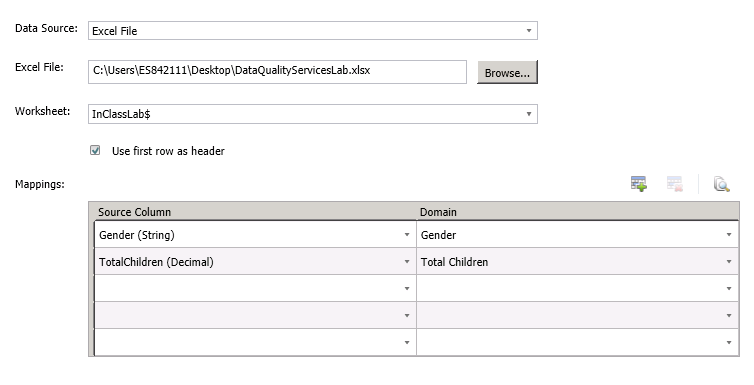
If you see this, then your KB is in an invalid state but you can right click on it and select **Unlock** to put it back in a valid and usable state.

## Scrubbing/Cleansing Your Data

After you have published YOUR new knowledge base, you can then use it to cleanse data from any data source containing columns that can be mapped to the domains defined in it. The simplest way to do this is to create a new data quality project based on the knowledge base (Make sure that you set the “Use Knowledge Base” dropdown to YOUR KB!), specify the **Cleansing** activity, as shown here and then click on the **Next** button.

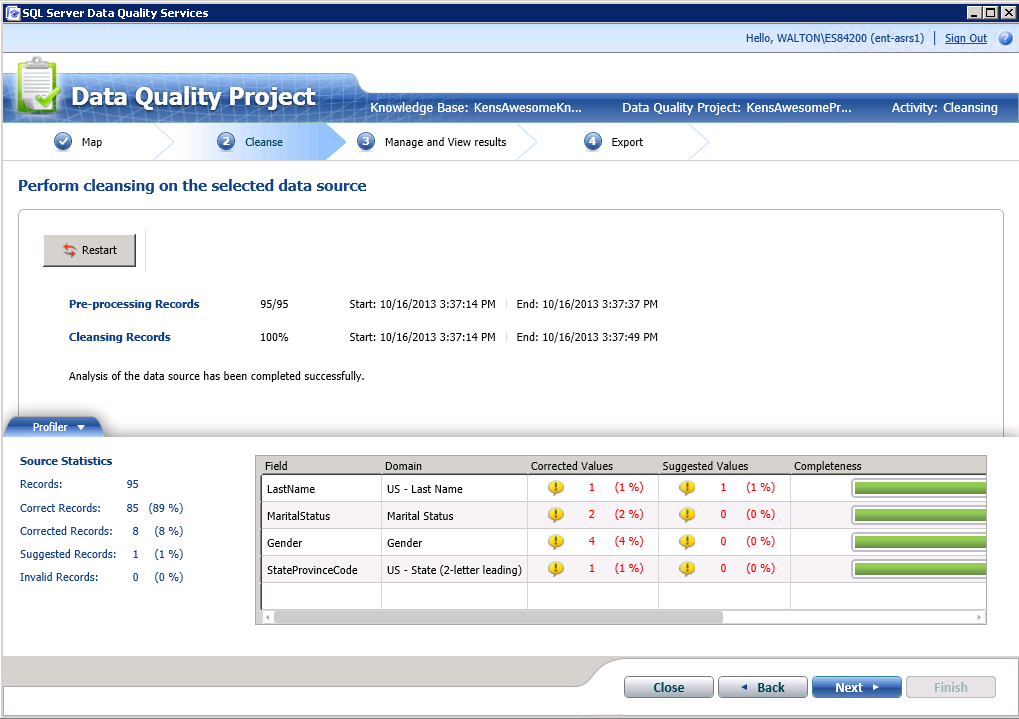


You will next choose which Excel file to cleanse and which columns in that file to cleanse. Change the Data Source to Excel, browse to your dirty data file and then map the columns to be cleansing to their appropriate domains that contain the cleansing rules.



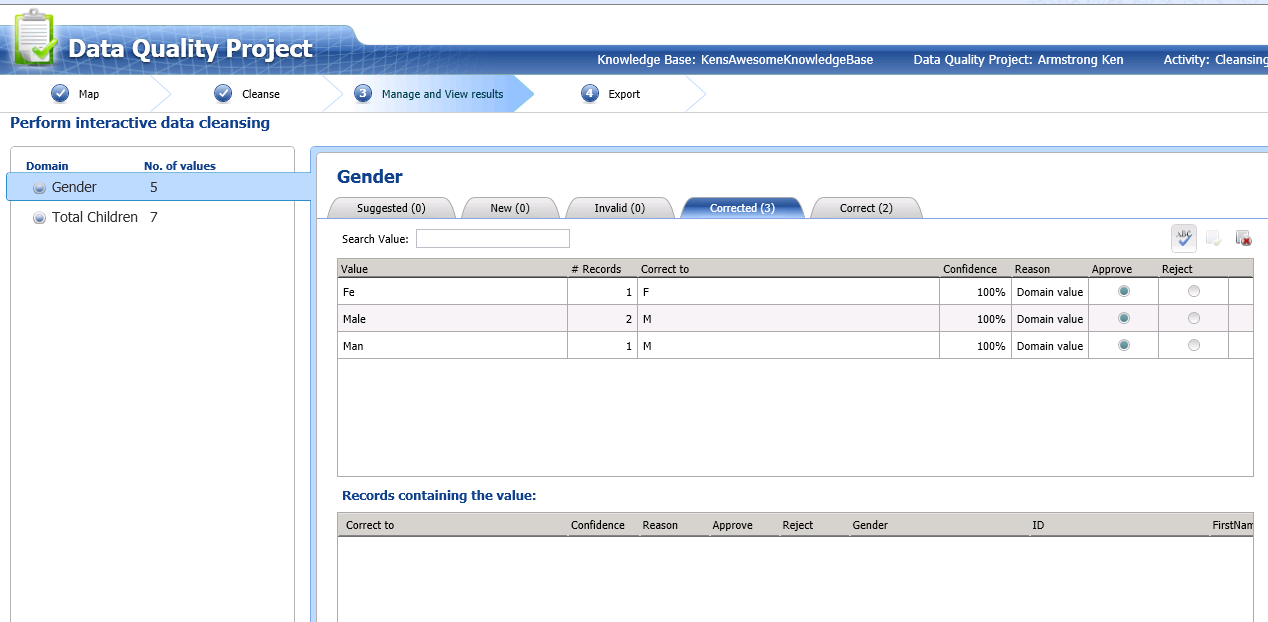
Click **Next** and then on the next page, **Start** the project.

You should get something like this:

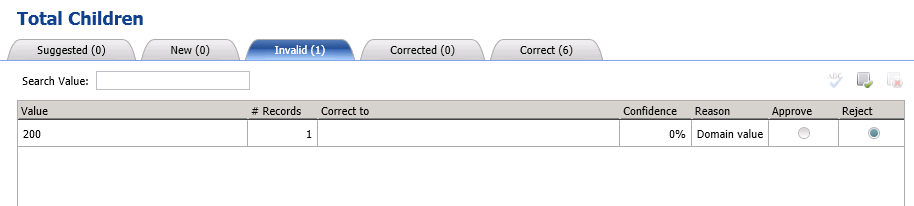


Click **Next** to get to the next page (duhhh).

On the **Manage and View Results** page, look through both of your domains to see if all of the bad data was corrected. In my dataset there were the Genders that were flagged as invalid by your domain rule and that rule corrected all of them so that domain looks good.

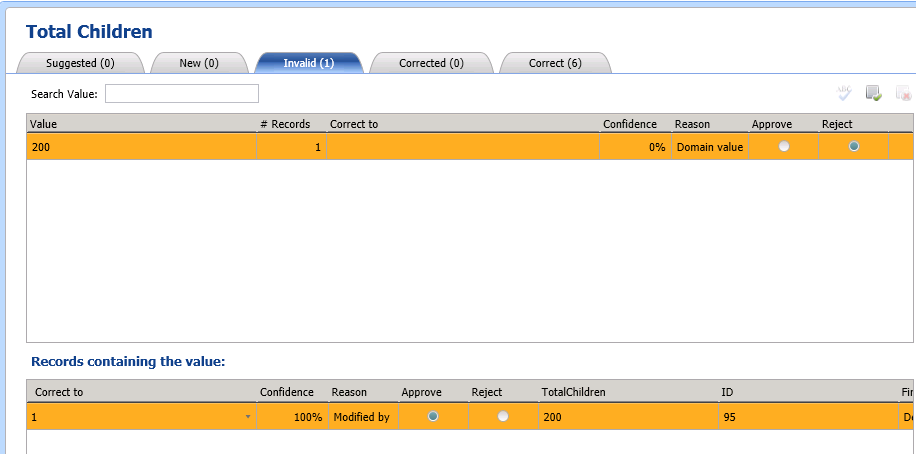


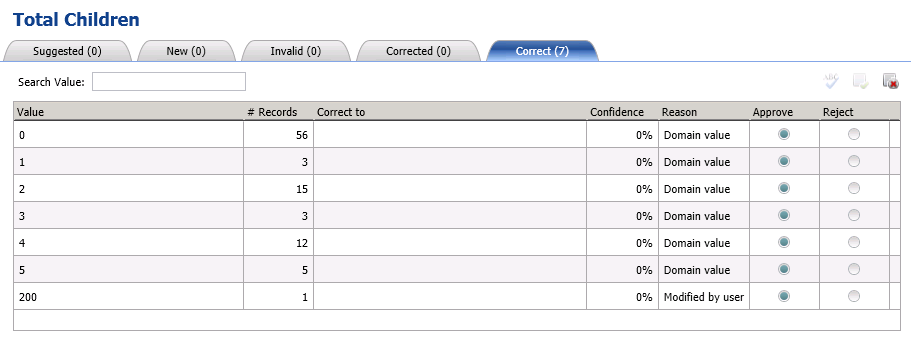
However, if you look into the Total Children domain there were 6 Correct values and 1 Invalid:



What do we want to do with this value? We can just let the KB flag that data as wrong and reject it completely or we can have DQS correct it to whatever value we want. If we look at the underlying data, we see that the average value for that column is 1.2 children so let’s cleanse this value automatically to 1 by doing the following:

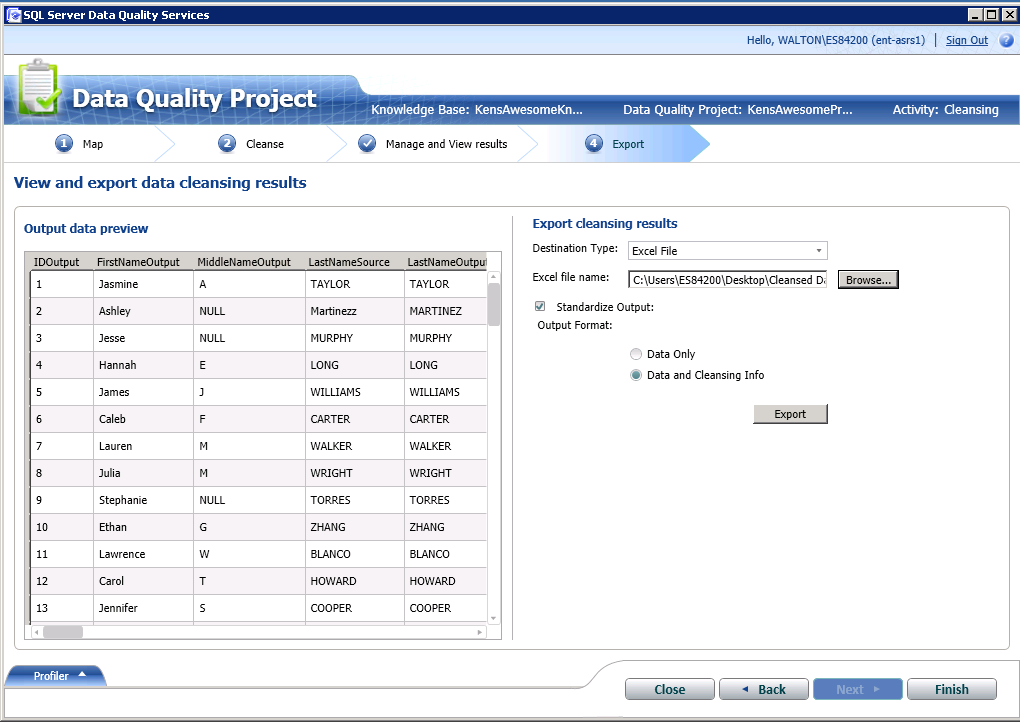
1. In the top pane, click on the bad value (200)
2. In the lower Records containing the value pane, type in the **Correct to** value of 1
3. Set this lower row to Approve



1. In the top pane, click Approve radio button. Immediately after you click this button, that value of 200 will move itself into the “Correct” bucket on the right tab! Here is what my Correct tab now looks like:

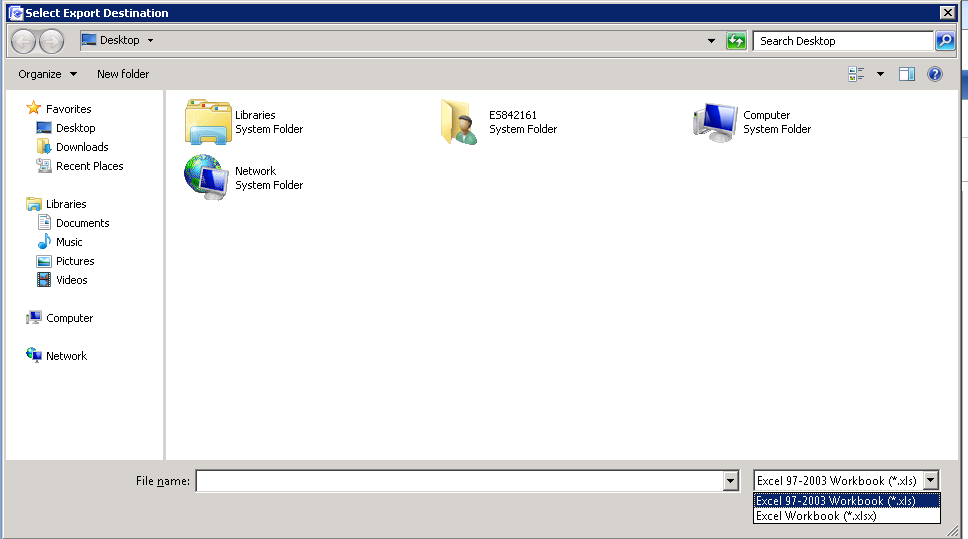
Click **Next** once you are happy with your cleansing results.

You are now ready to export your results to a different file. We will use the file that we created on your desktop back at the beginning of this document called “**Cleansed Data”**.

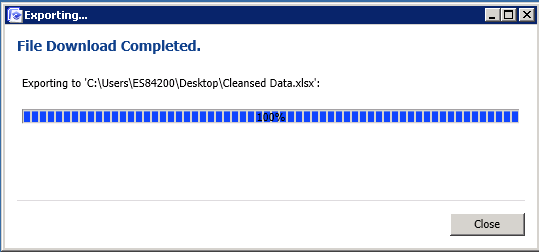


Change the Destination Type to Excel and browse to your newly created **Cleansed Data** file.

You may have to change the file type to xlsx in the dropdown below in order to see that file:



You should select the **Data and Cleansing Info** radio button and then **Export** the data.



Take a look at your output file to see what happened any why! You can easily do this by clicking on Data/Filter and then going to the column on the far right and selecting just the values of “Corrected”.

If you are happy with all of the cleansing that occurred, copy and paste the columns of updated data into your original data file. As a rule of thumb or to implement “best practices”, you should then save this file with a different version name such as **DQS Lab Data – Gold** or something like that signifying that it is done and will not change again.