Technical White Paper

On

Use of OEDQ in Conversion Process

Velsy Thomas

Fujitsu Consulting India Pvt. Ltd.

Contents

Introduction:

This document explains use of OEDQ in conversion program through Data Cleansing, Validation and Loading process.

What is OEDQ:

OEDQ provides a comprehensive data quality management environment that is used to understand, improve, protect and govern data quality. OEDQ facilitates best practice master data management, data integration, business intelligence, and data migration initiatives. OEDQ provides integrated data quality in customer relationship management and other applications.

Why to prefer OEDQ:

a) By delivering the required fit data in the appropriate format for their business critical application hence ensuring maximum efficiency.
b) Enables quick identification and resolution of problems in underlying data
c) The OEDQ provides features that convert unsorted data to get a meaningful set of data hence enabling Customers to identify new opportunities, improve operational efficiency.

Features of OEDQ:

Following are the key features of OEDQ:

- Integrated data profiling, auditing, cleansing and matching
- Browser-based client access
- Ability to handle all types of data (for example, customer, product, asset, financial, and operational)
- Connection to any Java Database Connectivity (JDBC) compliant data sources and targets
- Multi-user project support (role-based access, issue tracking, process annotation, and version control)
- Services Oriented Architecture (SOA) support for designing processes that may be exposed to external applications as a service
- Designed to process large data volumes
- A single repository to hold data along with gathered statistics and project tracking information, with shared access
- Intuitive graphical user interface designed to help you solve real world information quality issues quickly
- Easy, data-led creation and extension of validation and transformation rules
- Fully extensible architecture allowing the insertion of any required custom processing

How OEDQ can be used in Conversion Program

Conversion is a Process where existing data from the client's old system is extracted, cleansed, formatted, and installed into a new system. These are One-time only process that requires extensive testing and preparation. They must be executed and performed before a system goes into production. This section will display how OEDQ can be used in the conversion program.

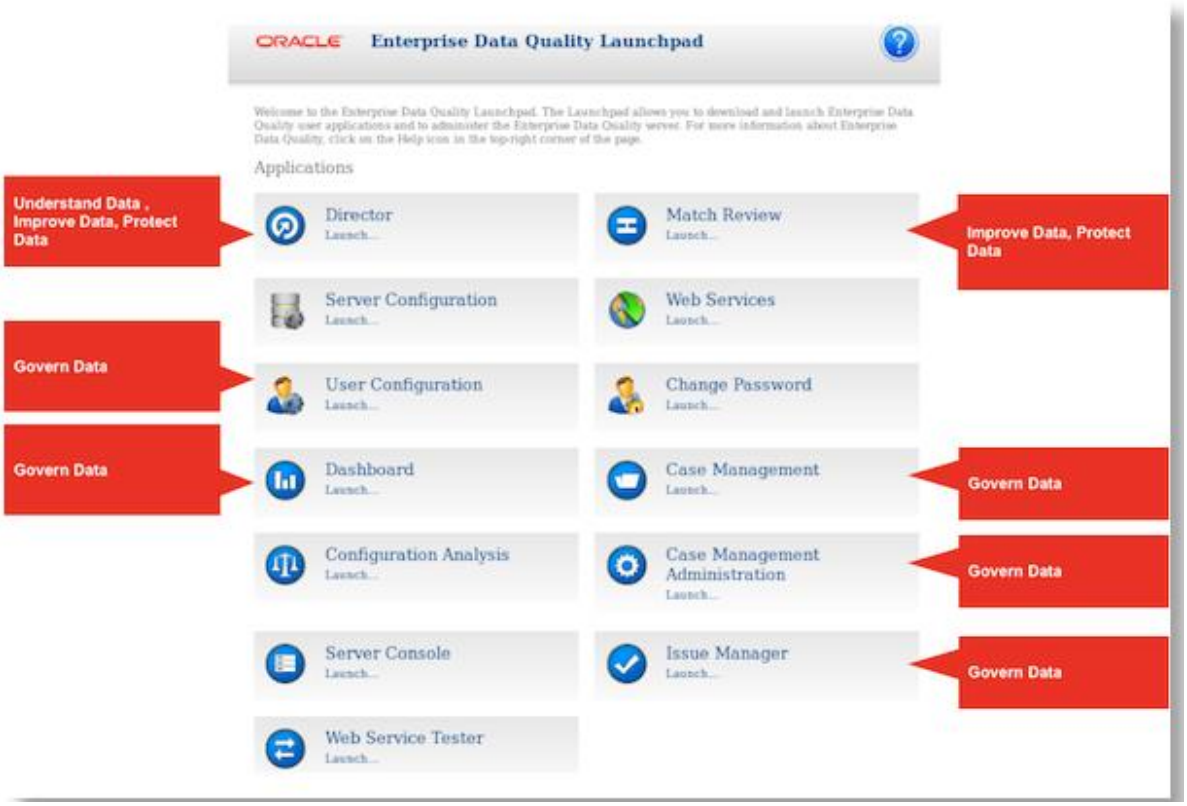
Role of OEDQ

Data plays an important role in conversion. Text data is rarely available in a completely neat, ordered fashion. Typical problems include:

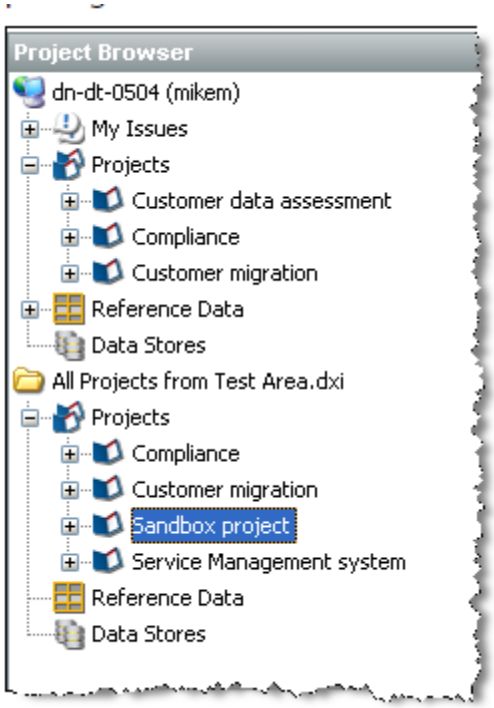
- Misfielded data, such as names, comments, or telephone numbers in address blocks
- Poorly structured data such as addresses, where data can flow from one field to the next Field
- Notes fields that store information the data structure doesn't support but contain useful semi structured data that normally cannot be analyzed or extracted, such as names, comments, or telephone numbers in address blocks

The file from the legacy system needs to analyzed, validated and then finally converted to the required format to be acceptable in the new environment. In Oracle, the task of cleansing, validation and extraction is performed at the program level using PLSQL. Using OEDQ, we can perform the cleansing check like removing duplicates, Merging, Validations and then finally loading the data to the staging table. Also these data can be extracted to the required format in case of Manual conversion.

Understanding the Tool



Director: Used for profiling, analyzing and cleaning your data. The Director will comprise of a Project, Snapshot, and tools to perform the profiling, analyzing and cleaning and Export the data to the required Table.

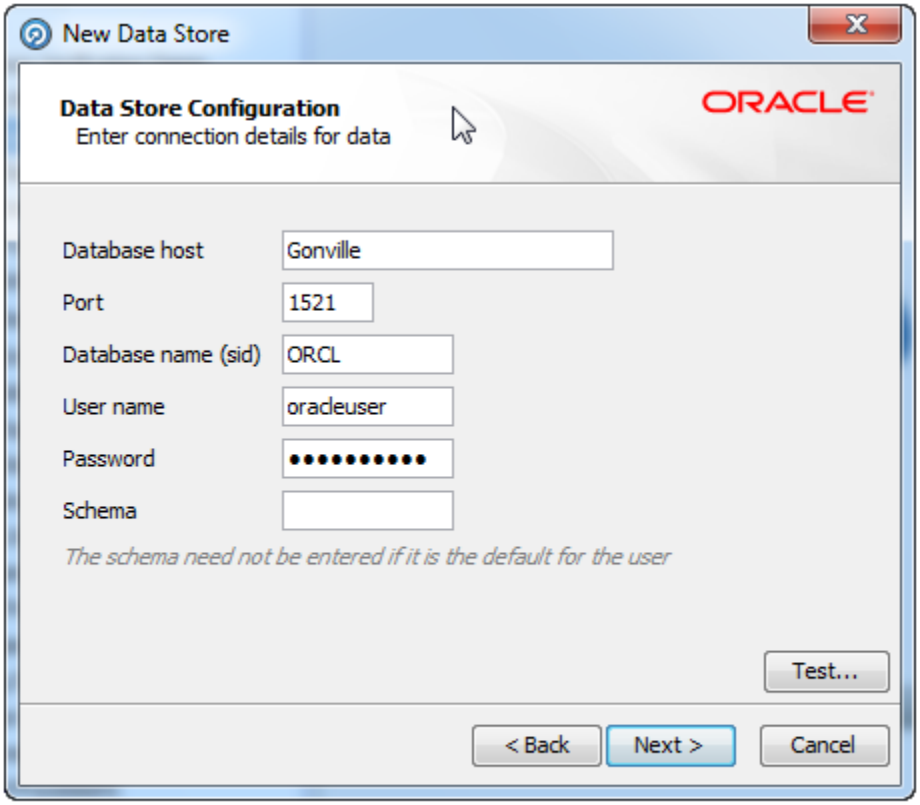


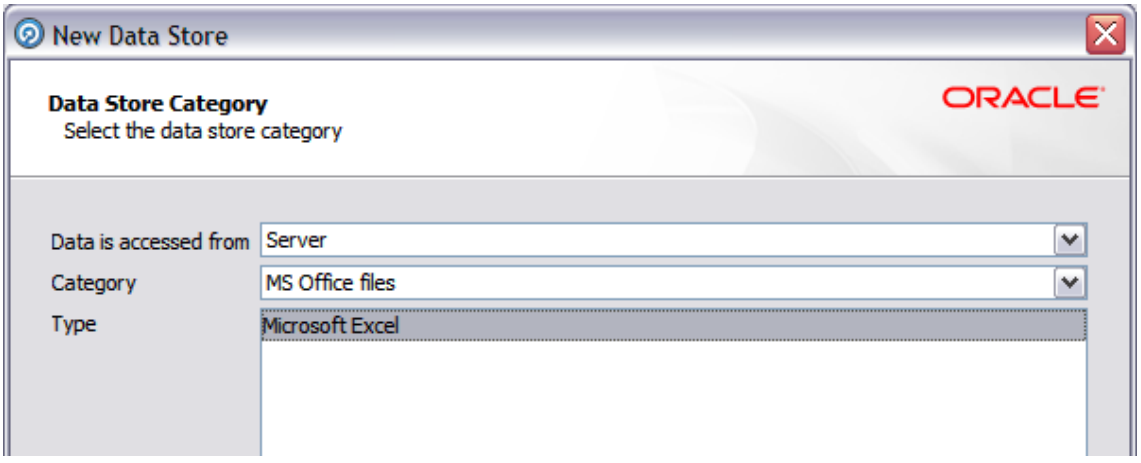
Projects Tab: The tab lists the number of projects. Like for example if we are performing a Customer Conversion, then the same can be considered as one Project. So inside this project we will perform the required cleansing, transformation, validation and finally exporting data to the required staging table.

Data Store: 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 the written Staged Data results of a process to a data store, or both. OEDQ supports native connections to the following types of data store:

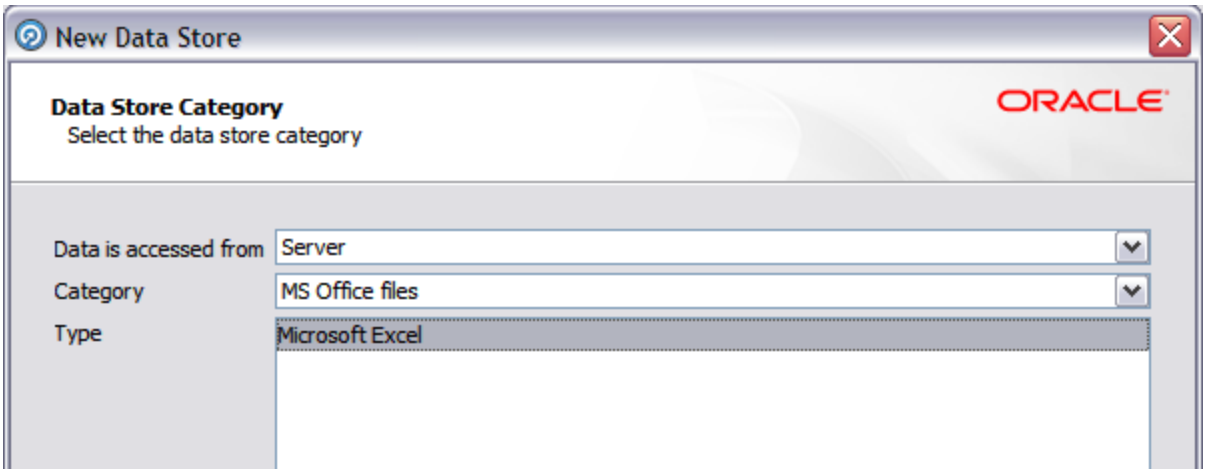
- Oracle
- PostgreSQL
- DB2
- DB2 for i5/OS (see below)
- MySQL
- Microsoft SQL Server
- Sybase
- Microsoft Access
- JNDI

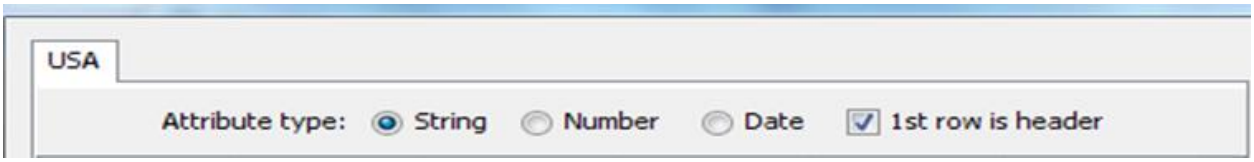
They are of 2 types: Client Side and Sever Side. Client Side is used when the file is stored in the local Machine. Server Side is used when the file is stored in the Server landing area. Usually it is recommended to connect to the data store via the Server. The **file (Access, Text, Excel or XML)**, enter the name of the file as it exists (or will exist) in the server landing area, including the file suffix. For **fixed-width text files**, the fields in the file must be defined appropriately. For a **Database**, specify the **Database host**, **Port number** (if not using the default port number), **Database Name**, **User Name, Password,** and **Schema** (if different from the default Schema for the User).



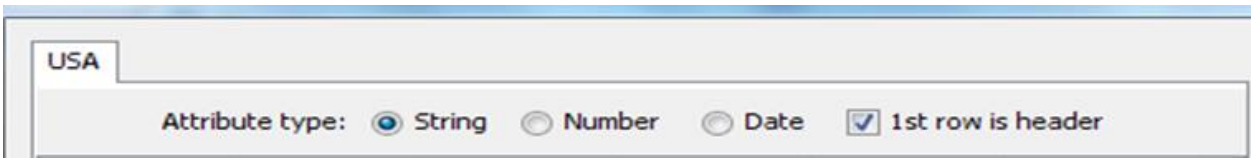


The connection to the new data store can be checked using the **Test** button.

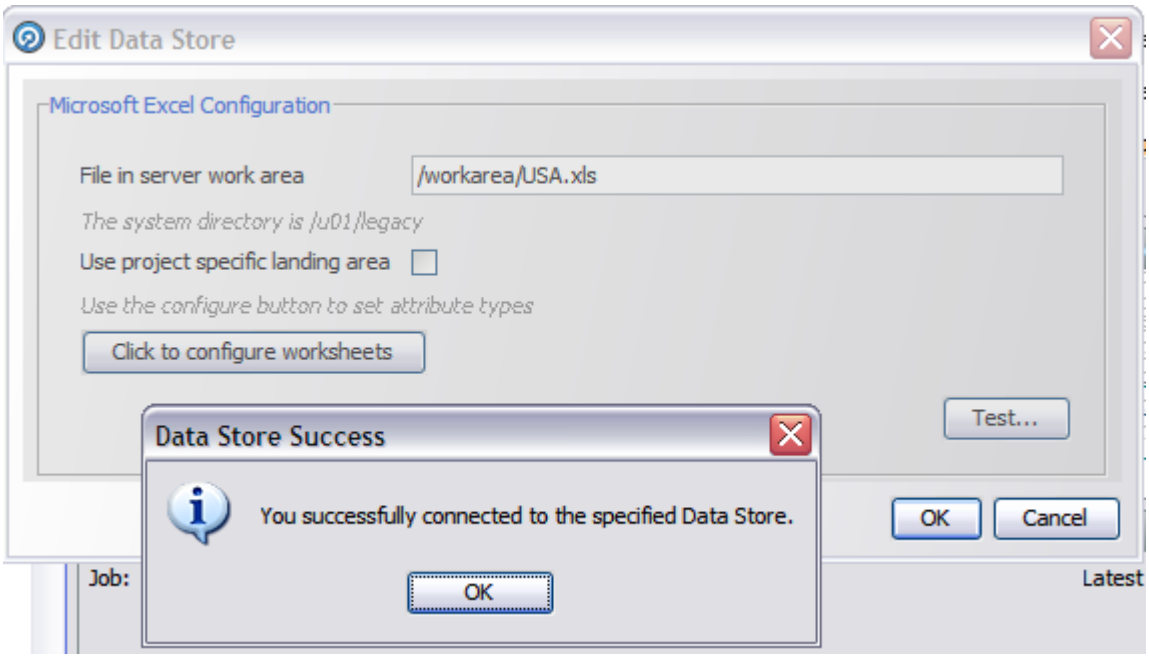




Click on 'Click to configure worksheets'
Enable first row as header option
Classify fields which needs to String, Date as well as Number



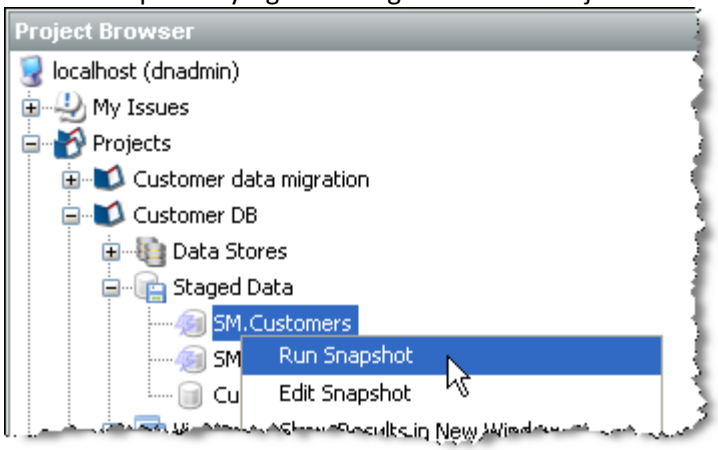
Test the data connection



By this step, Data file is being accessed from the required server location.

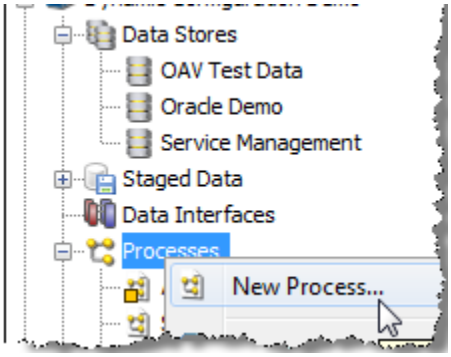
Snapshot

A snapshot is a staged copy of data in a Data Store that is used in one or more processes. Create a snapshot based on the above required data store. This step will create a copy of the data which is being referred as Snapshot. We can also create a Snapshot by selecting the table or view and also specifying the SQL query to get the data from required database tables. Sorting and Filtering of the required columns are also available at the Snapshot level. The snapshot is the one in which further cleansing and validation will be performed. Once a snapshot configuration has been added, you can run the snapshot by right-clicking on it in the Project Browser, and selecting Run Snapshot:



Process

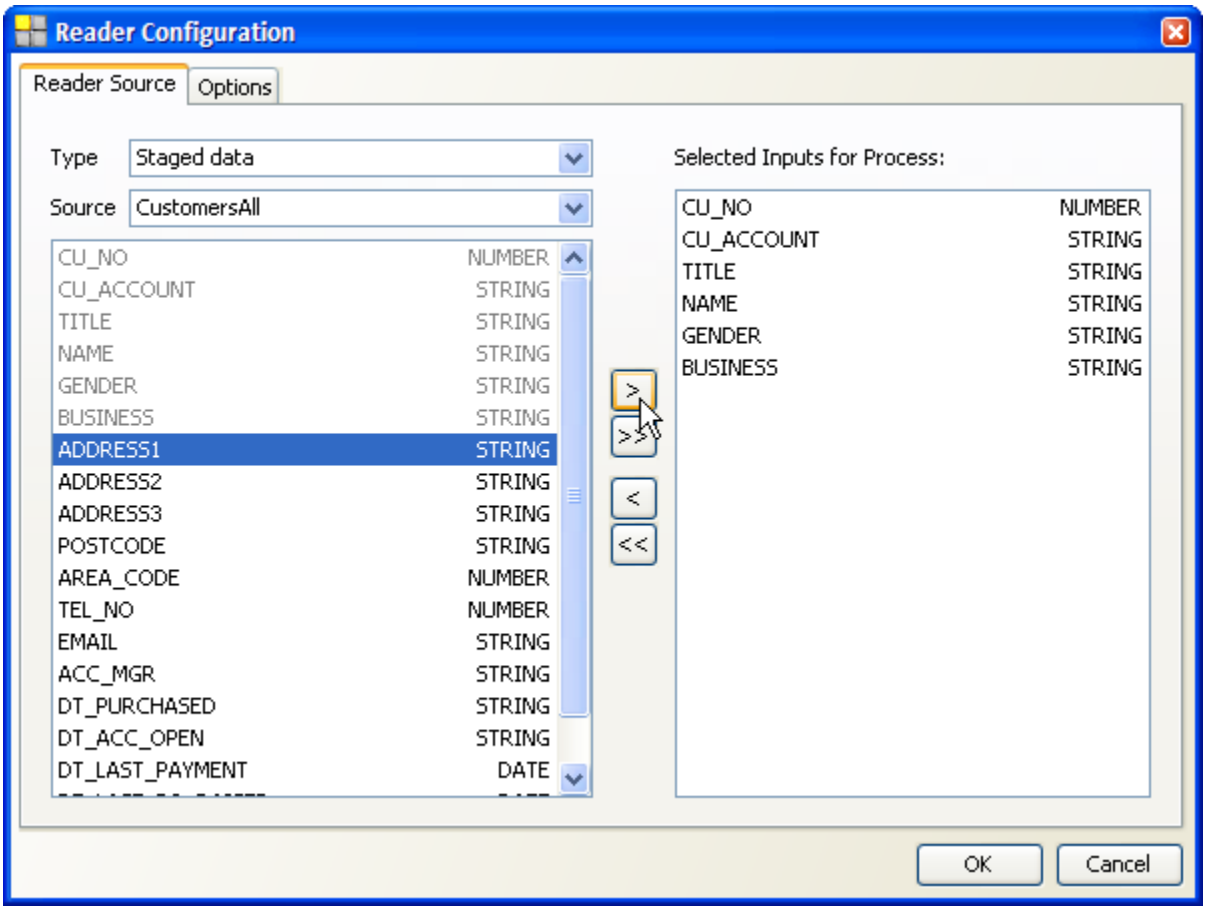
Process is used for analyzing the data from the snapshot. Process needs to be created by selecting the appropriate snapshot. In the process, processors will be added to perform further cleansing and validation. The cleansed and validated records will be mapped to the required staging table. The Records which fails the validation can be exported to an excel sheet for further reference and correction.



Reader

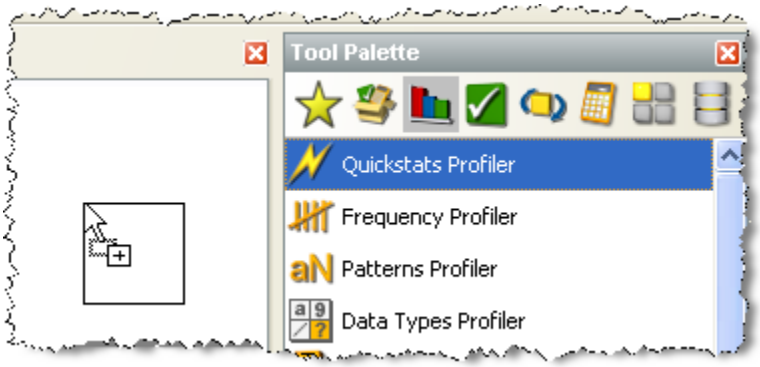
Readers are used at the beginning of processes in order to select the sources of data and any selection and reordering of data attributes from that data source For example, for the purposes of a specific process, option is available to select only required columns from the data file.

All the available attributes in the data appear in the left pane. Select only those that is required with in the process by using the arrow buttons to select, and de-select attributes:



Tool Palette

The Tool Palette provides a list of all the available processors. To use a processor in the definition of a process, drag it from the Tool Palette and drop it onto an open process:



Tools are categorized by Profilers, Audit Checks, Transformers, Advanced Tool like adding script. In the Cleansing, we can use the below listed tools to identify the required records which will undergo validation. Listing tools from a Conversion point of activity.

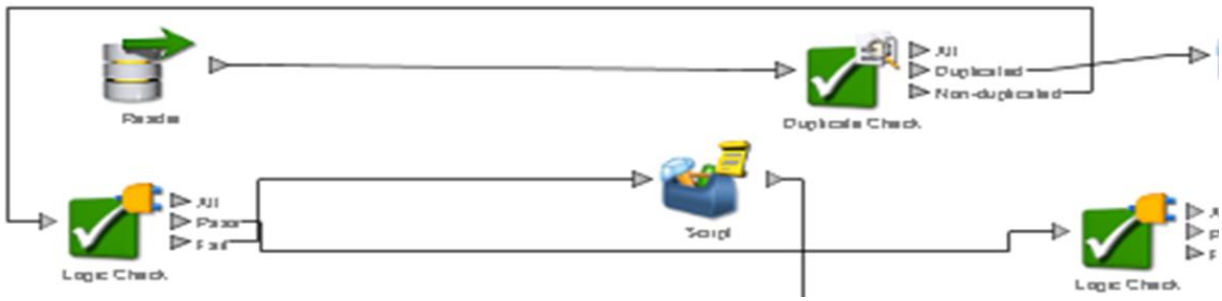
Audit Checks:

Type of Rule	Example business rule	Audit processor
Whether or not the attribute is allowed to contain null values	The CU_NO attribute must not be null	No Data Check
The allowed or expected length of the data in the attribute	The CU_ACCOUNT attribute must be between 10-11 characters in length, and must not contain spaces	Length Check
The data type consistency in an attribute	There must be no numeric values in the NAME attribute	Data Type Check
The validity of values in an attribute	Values in the TITLE attribute must match a list of valid titles	List Check
The validity of specific characters in an attribute	The values in a NAME attribute must not contain characters such as #~@;:/?.>,<%$£!^*	Invalid Character Check
Duplication of values in an attribute	There must be no duplicate CU_NO values	Duplicate Check
Check one attribute's value against another	The DATE_OF_BIRTH attribute must be before the DATE_OF_DEATH attribute	Cross-attribute Check
Check for related data in a reference table	There must be at least one active Contact record for a Customer	Lookup Check

Check for data which passes a Logic expression	There is a valid DATE_OF_BIRTH attribute and a valid Postcode and a valid email address	Logic Check
Check that data has a specific value, or value range	All male Customers must have a Gender value of 'M'	Value Check

Duplicate Check

The Records from reader are mapped to the duplicate check tool which checks for duplicate data for all fields



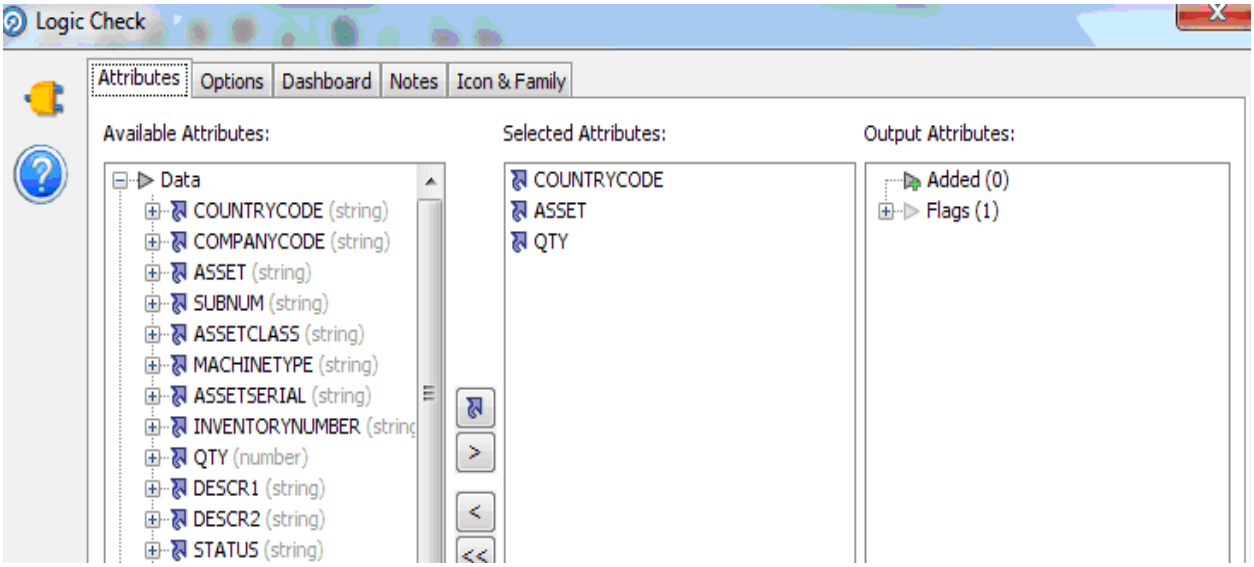
Non Duplicated: Non Duplicated Records can be used for further processing.
Duplicated: Duplicated records can assigned to a script from advanced tool where a message can be written to display the existing record is duplicated with the corresponding column.

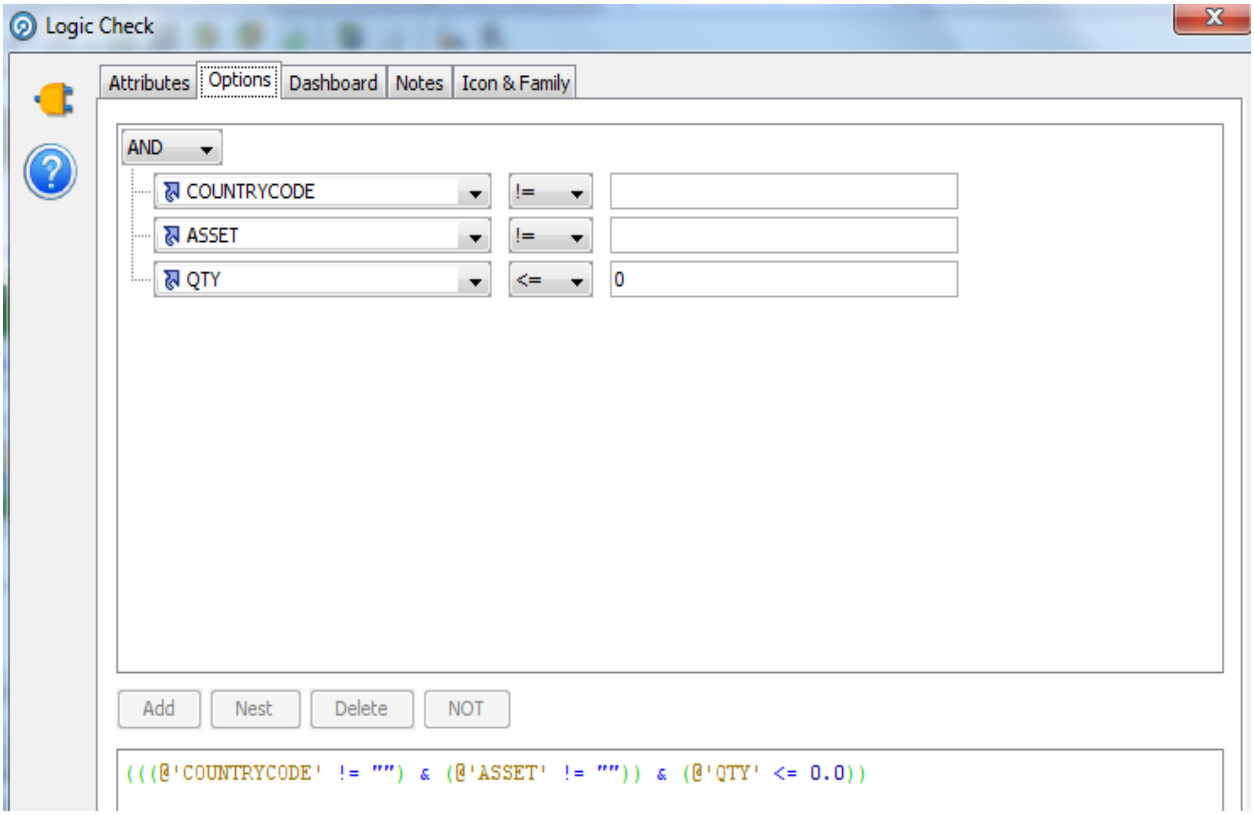
Logic Check

In the Logic check, specific validation condition can be written and verified. Those Records which pass the condition will move to the next round of validation and those which fail can be assigned with a script with required error message which can be written to excel sheet that can be used for future reference.

Example:

Validation Rule: Verify records as Passed where Country code is NOT NULL, Asset Number is NOT NULL and Quantity (QTY) <=0. The result will be provided as Pass and Fail section,





Pass: The records passed can be moved for further validation

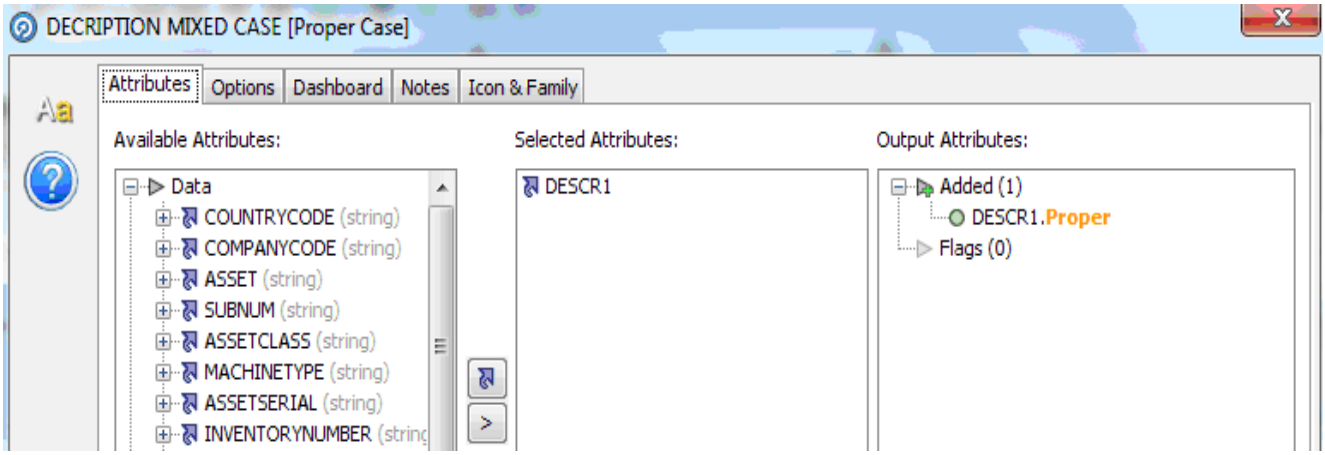
Fail: The records which fail for the above conditions will display the below error message (**Advanced Tool: Script)** which will be written to a result book.

var res = ' '

res = input1[0]||'The record fails as Country code, Asset are null and Quantity is less than 0'

output1 = res

input1[0] → Pass the unique column to script which identifies the record so while correction it is easier to correct the same and provide the appropriate file again during next testing phase.



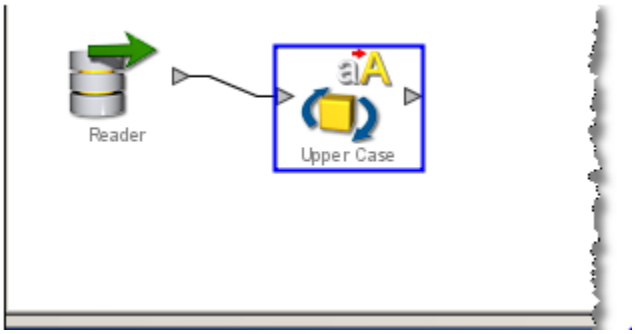
Transformation

Transformation processors take one or more input attributes, transform them, and output the transformed values in new attributes. Transformer Processor include Character Replace, Cross Attribute Check, Replace, Proper Case, Trim Whitespace, Convert Number to String, Convert Date to String, Convert String to Date etc.

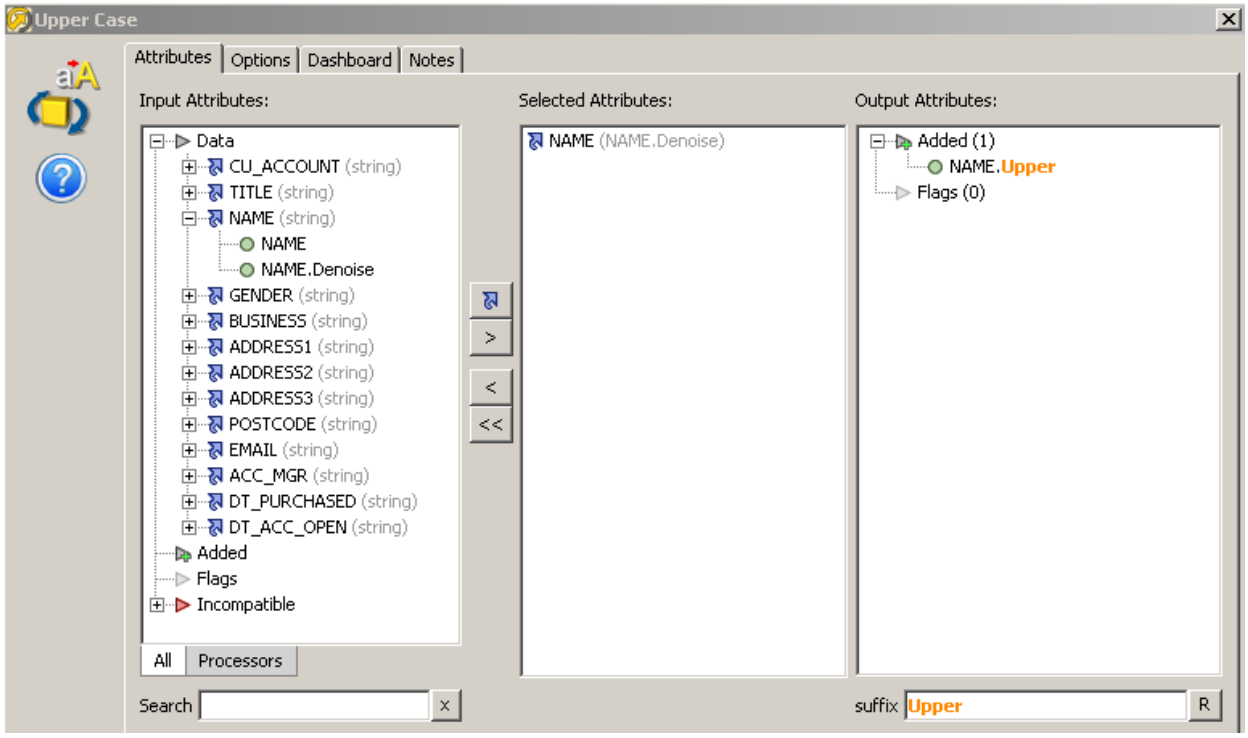
Example:

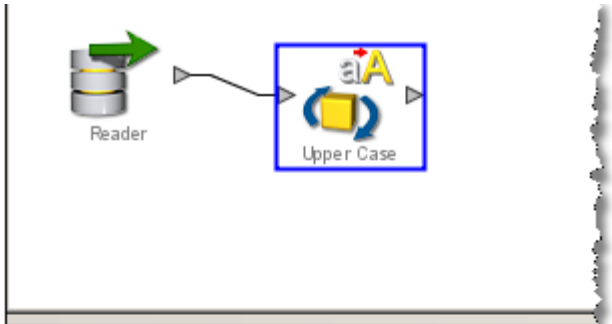
Uppercase Processor

Drag and drop the Upper Case processor from the tool Pallete



Select the File column which needs to be converted in Upper Case





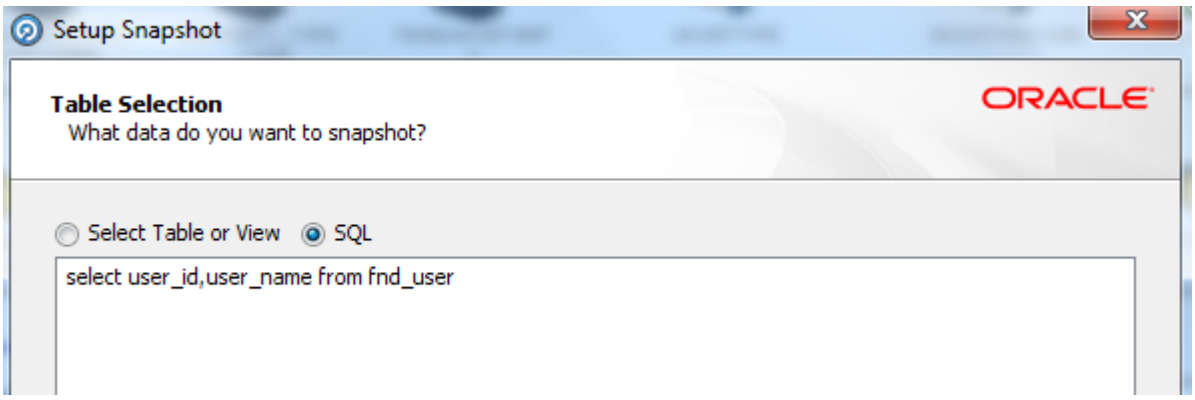
Results Browser - Upper Case

NAME (asc)	NAME.Upper
# MCAULAY	# MCAULAY
# RAE	# RAE
# SWAN	# SWAN
# WILLIAM	# WILLIAM
A Test	A TEST
Abigail ANDERSON	ABIGAIL ANDERSON
Adam BALL	ADAM BALL
Aidan HODGSON	AIDAN HODGSON
Akosua GODWIN	AKOSUA GODWIN
Alan BLACKWELL	ALAN BLACKWELL
Alan BLATCH	ALAN BLATCH
Alan GILLS	ALAN GILLS
Alan HENTHORN	ALAN HENTHORN

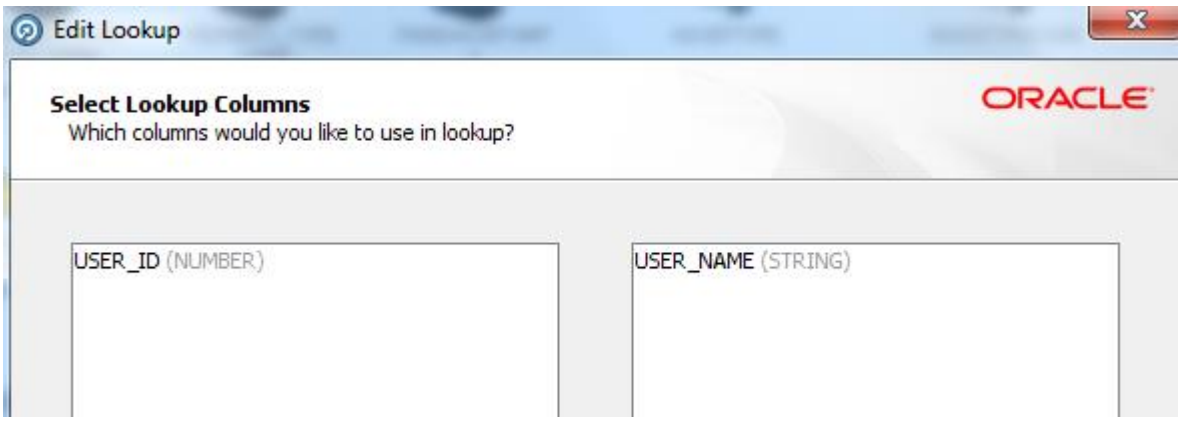
Once the process has been designed by adding the required validation and cleansing rules, we can write the staged data.

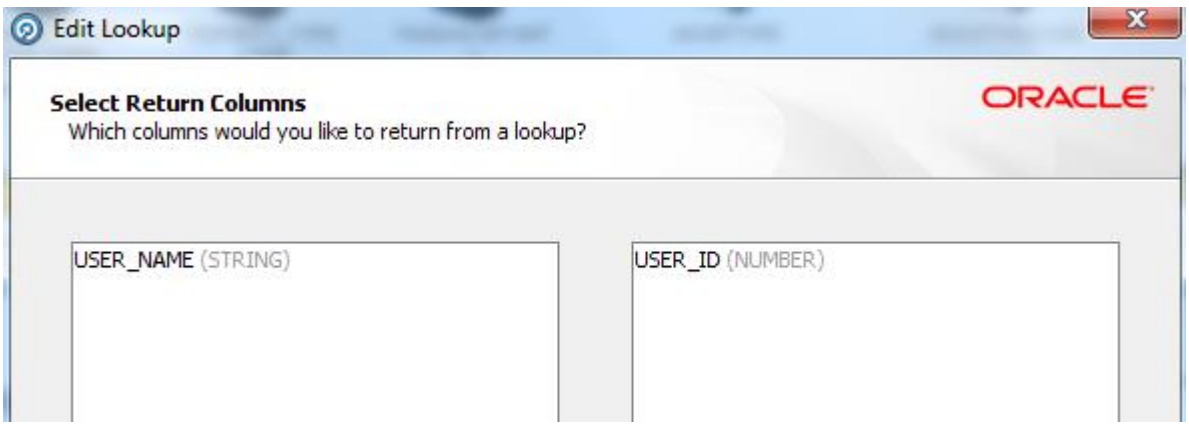
Use of Look up and Reference

The Lookup and Return processor allows you to look up related data in a Reference Data source, and return the data for use in downstream processing. Reference data can be set of static values against which data in the source file needs to be validated. Reference also can be generated by connecting the required server and getting data by writing SQL statement using required tables.

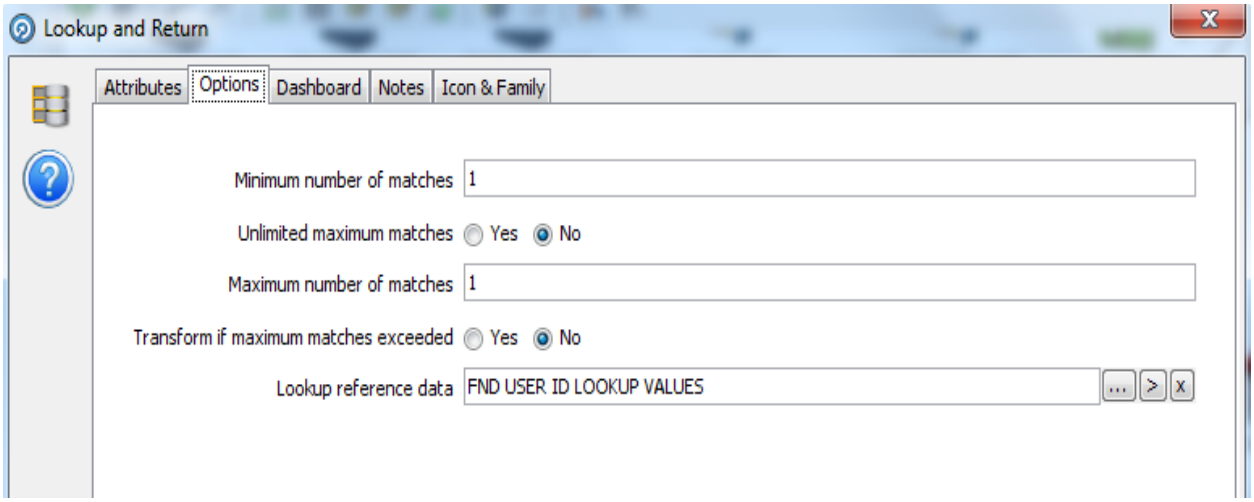


Create a look up based on the above SQL.





The above lookup and return takes the input value as user name and return user id.



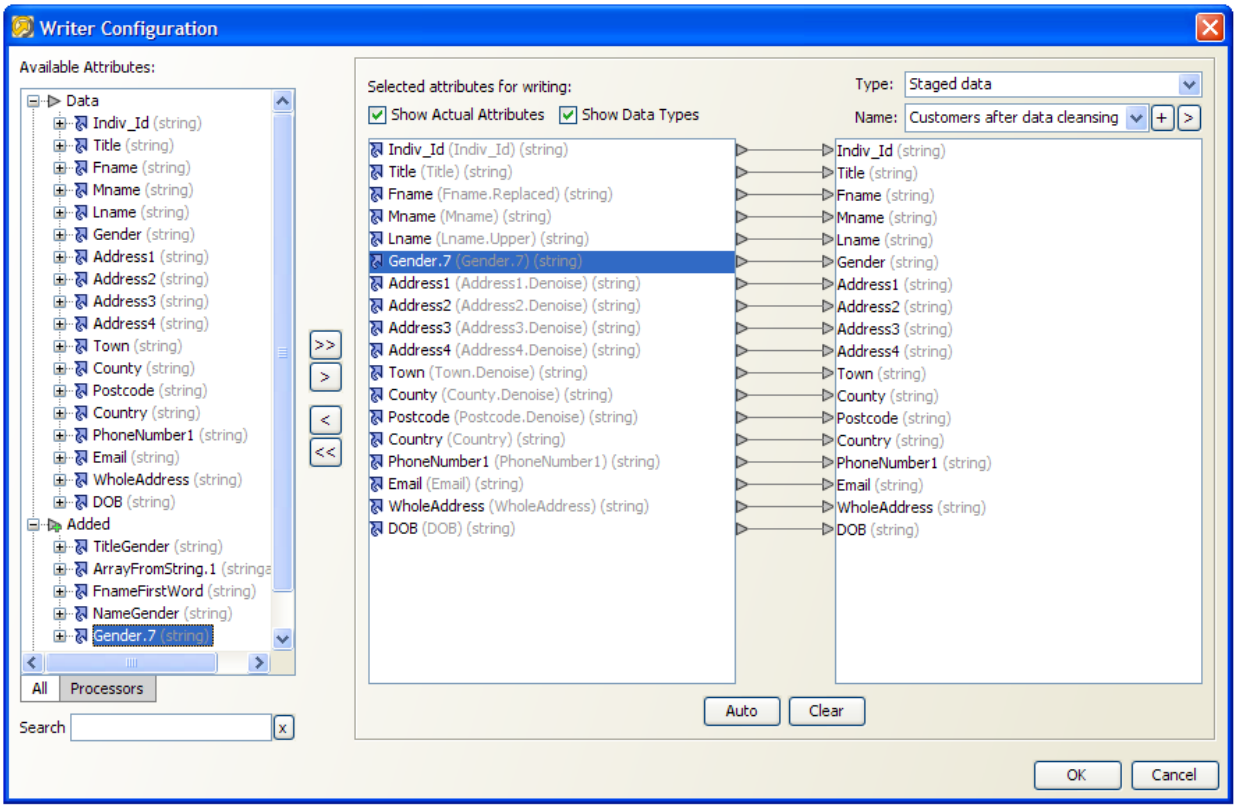
Transformed: These records will go to the next level of validation.
Untransformed: These records will error out and will be capture in the result book to be referred later.

Staged Data

Staged Data will be created referring to the required column structure of the required target table. Different Staged data can be created as per the requirement. The preferred option is to create Valid Staged Data referring all the validated Records. In valid Staged data referring to invalid records when the process is run. Before using the Writer, the staged data needs to be defined.

Writer

A Writer is a special type of processor used to write out records to a Staged Data table, Data Interface or Real time consumer (for real time responses). Select the attributes which needs to be written and map the same with corresponding columns in the staged data name which is referring the target table.



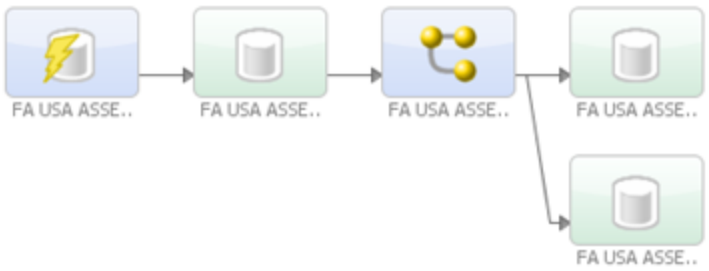
Export

Export can map Staged Data attributes to the attributes in an external database table or file, or it can auto-create the target database table or file when it runs. The appropriate data connection detail needs to be mentioned. Need to mention additional details like data should be appended to the table, whether the data should be overwritten, or whether to replace only target records with a matching primary key in the staged data set. Exports are defined separately from Writers in processes to allow more flexibility. Running and Scheduling Processes

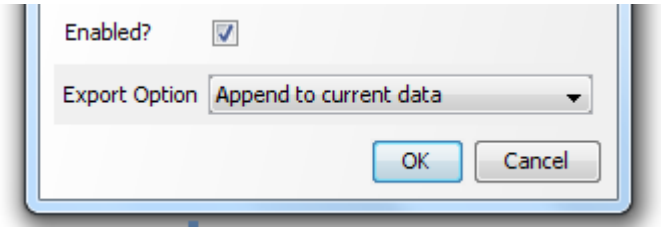
Job

A Job is an organized configuration of one or more ordered tasks. A task is the execution of a Snapshot, a Process, an Export, a Results Book Export or an External Task

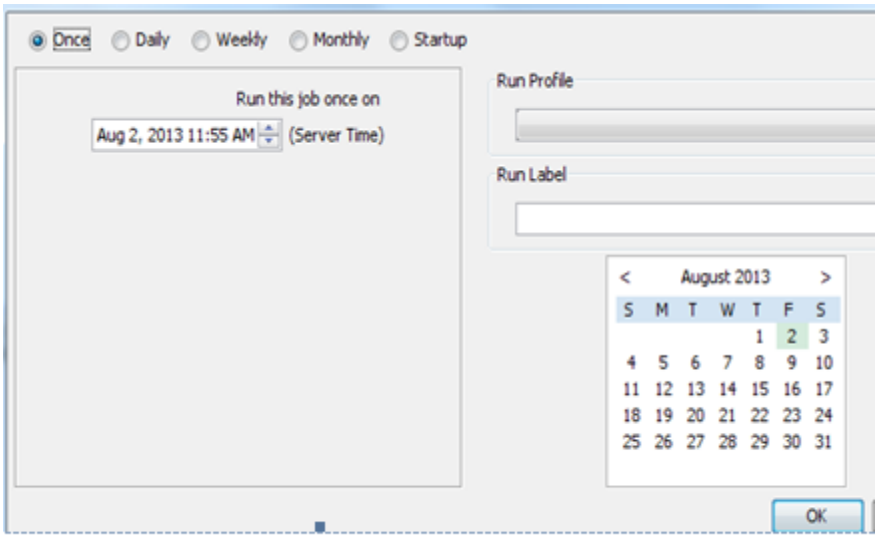
In the job creation, drag and drop the required the process to be included. It will be added in the job as shown below in first section. The second section displays the export where the validated data is exported to the required target table.



While adding the Export to the Job, it will ask for the option of appending or overwriting as shown below:



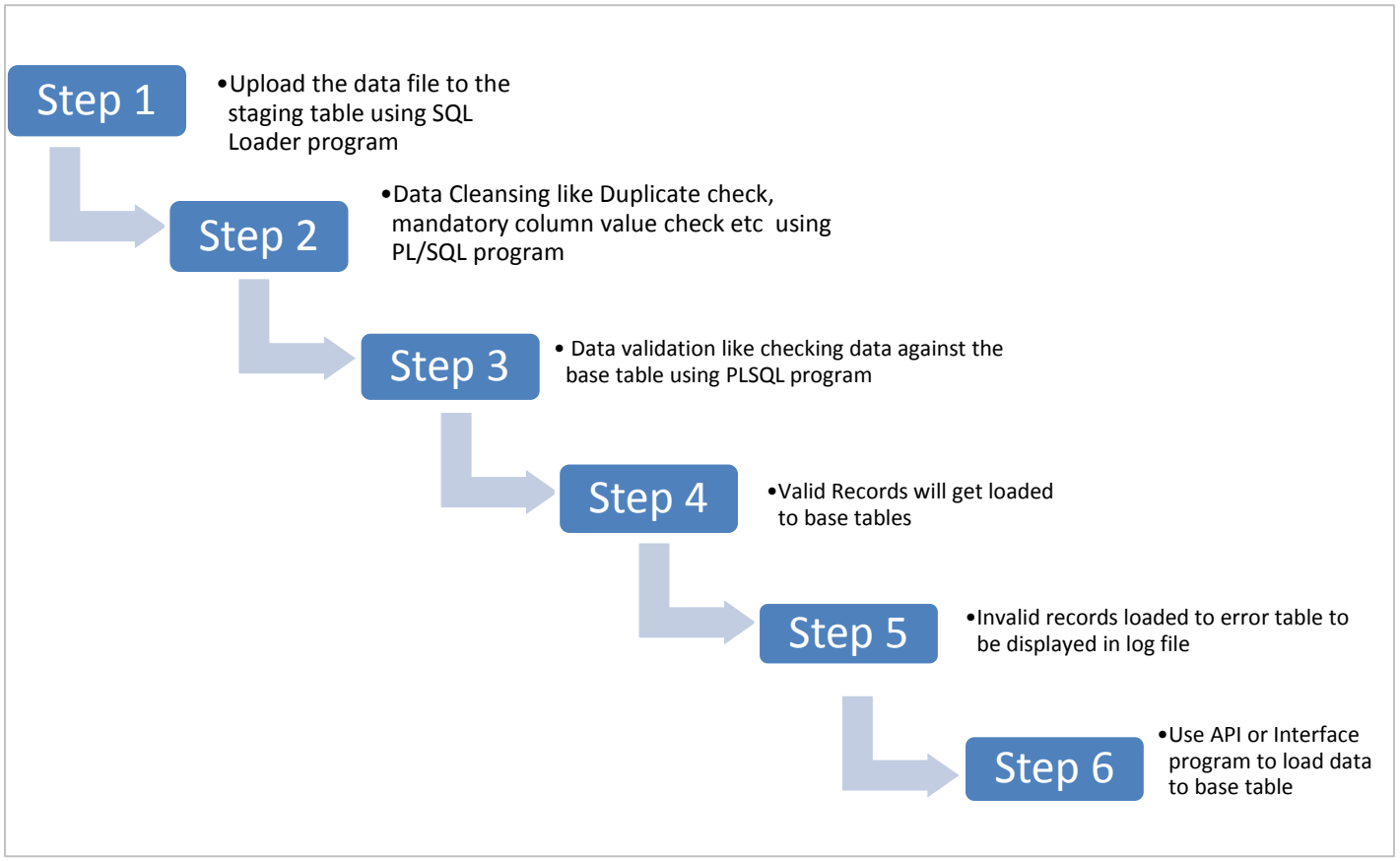
Job execution can also be scheduled as given below:



So Once we have defined the above we need to run the job, automatically the validated data will be inserted the required staging table at the oracle ERP side. Then after at the program level minimum complex validation as well As API or Interface table can be called to transfer the data to the required destination base table.

Comparison of the programmatic conversion process and using OEDQ

Normal Conversion Process



Using the OEDQ

The above described Step1 to Step5 can be performed in OEDQ with more ease and better performance. Hence only task of complex validation (complex SQL queries) and loading data to base table using API or Interface table needs to be performed at the Oracle ERP side.

Benefits

- Reduces the programming level at the Oracle ERP level
- Loading data to the staging table at the Oracle ERP side can be done much faster and hence increasing the performance.
- OEDQ provides a better output for the error records in excel. This will help the legacy team to correct the invalid records by seeing appropriate error message.
- Time Consumed for the validation and cleansing activity is less

- Pictorial representation help to identify the issues or errors quickly hence enables to resolves the issue fast.