DBW624 – Assignment 3
ETL Application

Now we are going to focus on the reference tables and get a feel for what is involved with moving and cleansing data.

For each of the REFERENCE TABLE SOURCES IN ASSIGNMENT 1, you need to create a script which will take the data from the source file (usually an xls or .csv format) and load it into your reference tables within our data warehouse.

Once in the warehouse, you need to clean the data, where necessary, to ensure it maps to the logical model defined in Assignment 1.

My recommendation is that you use the approach of creating a staging table which is the target of the IMPORT/LOAD from the .xls or .csv files – then – clean the data and move over to the final reference tables.

The process would look something like:

1. **Download your data from the government web site – this is now in a .csv or .xls file**
2. **Create a table which has the same schema as your downloaded file – This is the staging table**
3. **Ingest all the data from your downloaded files into the staging table**
4. **Perform cleansing on the staging table**
5. **Create your final table as outlined in your logical model**
6. **Move the cleansed data from your staging table into your final table**

Here are the three reference tables we will use (see Assignment 1 for details)

1. **Baby Names**
2. **Population**
3. **Life Expectancy**

All of this assignment can be done with SQL, however, you are free to use any programming language you like.

What you need to hand in is:
1 – Your ETL/ELT/ELTL script: A copy of all the steps you are taking from extracting, cleansing and loading the data into the warehouse. Basically your ETL or ELT script. This will cover all the steps defined above.
2 – Sample Data from Final Reference Tables. You also need to provide a sample of the data from within your reference tables, showing the clean data. Showing 20 rows of data from each table is fine.
3 – The output of a SELECT COUNT(*) for each of the three final reference tables.

This assignment is worth 6% of your final mark.