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 sources below, you need to create a script which will take the data from the government site and load it into your reference tables within our data warehouse.

Once in the warehouse, you need to clean the data, where necessary.

You can use the approach of creating a staging table which is the target of the IMPORT from the .csv files below – then – clean the data and move over to the final reference tables.

Here are the three reference tables we will use.

1. Names (male and female) in Ontario
    a. http://www.ontario.ca/government/open-data-ontario
    b. There are two separate files to down load. One for males and one for females
    c. Just search on "male baby names" and "female baby names" to find the source of data
    d. File name is ontariopbabynames_male_1917-2010_english.csv
    e. File name is ontariopbabynames_female_1917-2010_english.csv
    f. You can merge the two into a single reference table – or – keep them separate.
2. Population Table
    a. http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/demo05a-eng.htm
    b. Look for file: demo05a-eng.csv
    c. It is under the summary tables area
3. Average expected life span (male and female)
    a. http://www5.statcan.gc.ca/access_acces/alternative_alternatif.action?l=eng&keng=5.3&kfra=5.3&teng=Download%20file%20from%20CANSIM&tfra=Fichier%20extrait%20de%20CANSIM&loc=http://www20.statcan.gc.ca/tables-tableaux/cansim/csv/01020512-eng.zip&dispext=CSV
    b. File name is 01020512-eng.zip
    c. You can also extrapolate back in time from the data at this web site
    d. http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/health26-eng.htm

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 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. 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, but, I also want to see the output of a COUNT(*) so I know you got all the data into the reference files.

This assignment is worth 6% of your final mark.