ETL (Extract, Transform, Load) Process

From the source system the data is pulled and placed into the data warehouse. ETL process is responsible for this task. In this project SSIS and its toolbox is used for creating the ETL process.

Extraction: Pulling and converting data from the different sources into the format of data warehouse which will be given for transforming. In this project, data from four sources is taken out of which 2 are structured (Kaggle and simplemaps), 1 semi-structured (Numbeo) and 1 is unstructured data(Twitter). Both structured datasets is loaded in the data warehouse as the dataset are in csv format using SSIS import flat file tool. From numbeo, pulled the table into the csv file using R programming in R studio. For unstructured data, extracted the tweets for cities and sentiment analysis is performed using Twitter API and R programming.

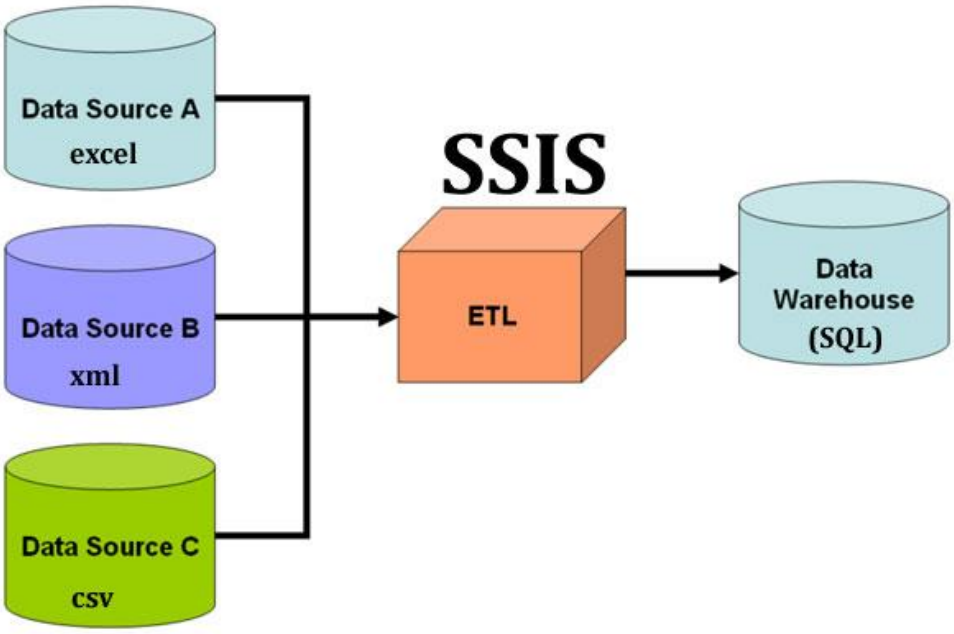


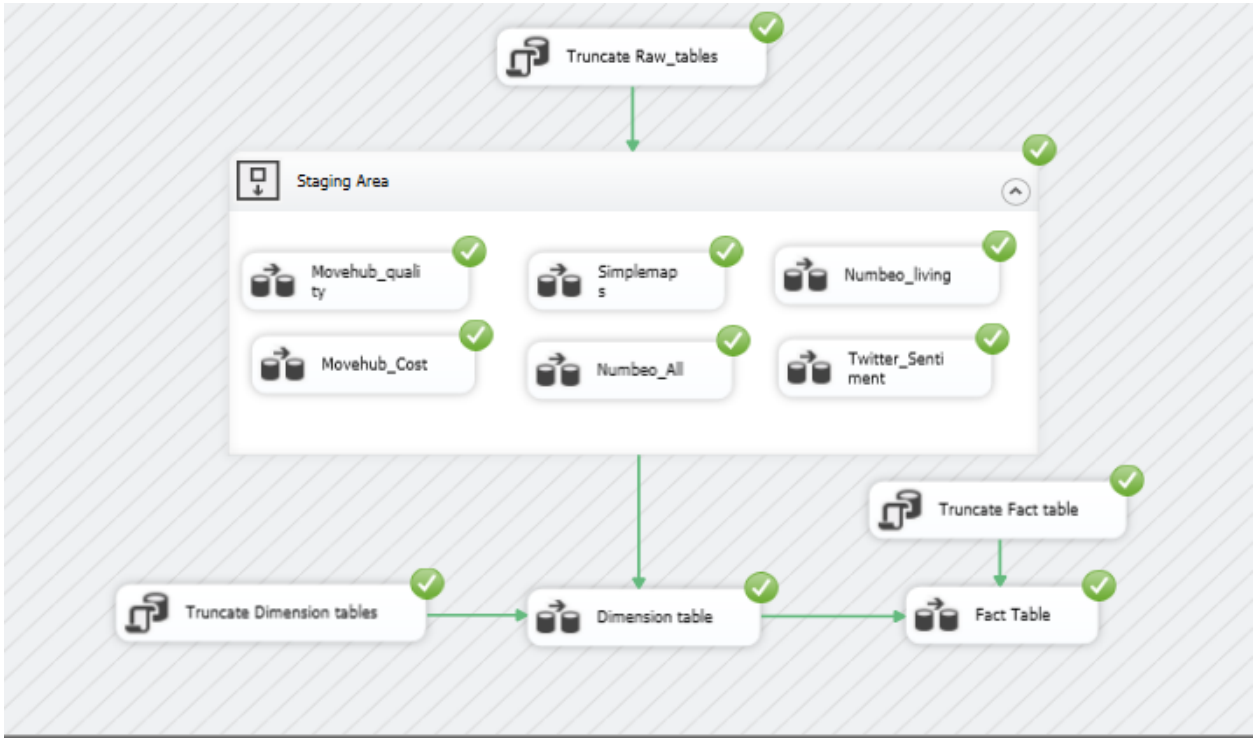
Figure 4 ETL Process in SSIS

Transforming: Following task are performed:

- Cleaning the data
- Filtering the column which are certainly used
- Merging data from different sources

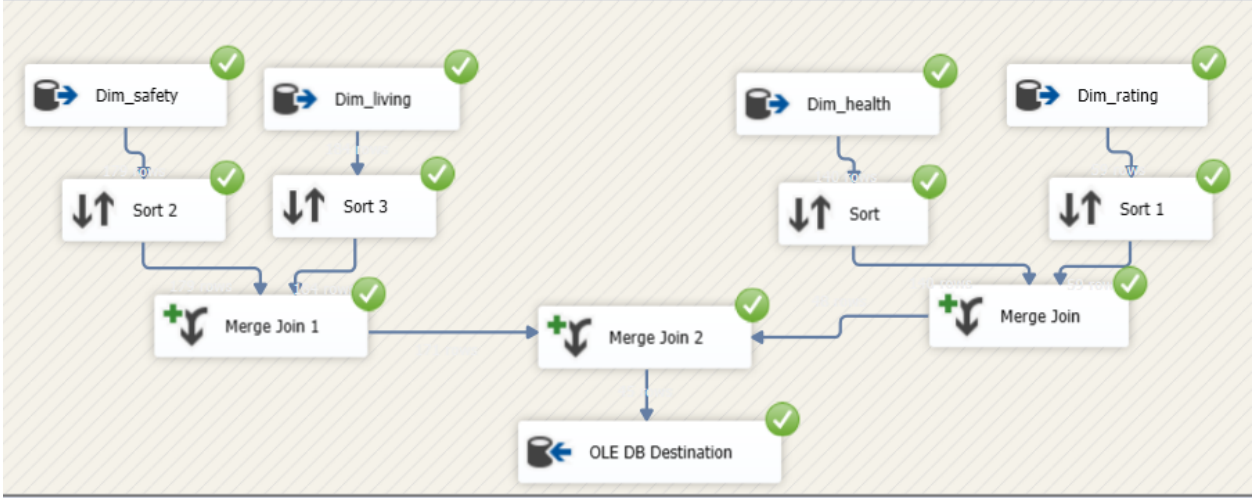
In this project, datasets from kaggle were already cleaned. For other dataset which are from simplemaps and Numbeo are cleaned as they were having alphanumeric values, duplicate records and unwanted columns (after pulling the data through R). For cleaning and transforming the data, used R programs.

Loading: After extracting and transforming the data in the chosen way it is loaded in the database. As data from the different sources is loaded in different tables, it will be used for creating the dimension and fact tables. [3]
In this project, staging tables have the cleaned and transformed data. As the source data is loaded the next task is to load the dimensions followed by fact table creation. Important task is to map the source data to the right table and then map the columns with the respective matching columns in the table.



Control Flow

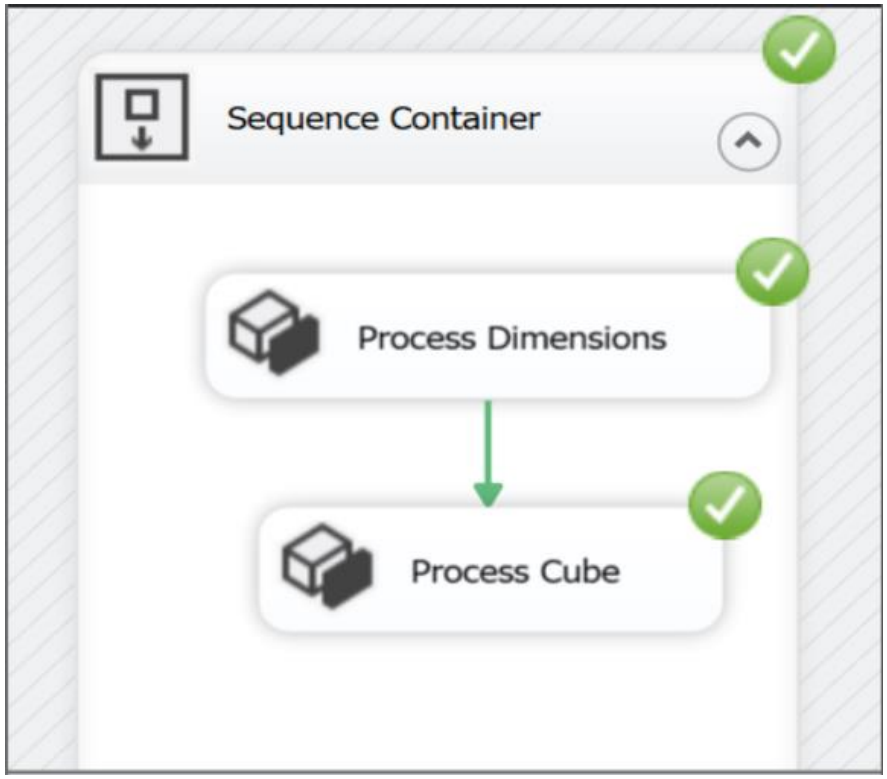
Truncating the table is necessary task because if running the data load multiple times duplicate values will be inserted into the tables. That is why all the raw_tables, dimension tables and fact table are truncated. As per the different city measures dimensions are created. All the dimension tables should have the primary key as the fact table has these primary keys from different dimensions and some dimension measures. For populating the fact table sort, merge-join is used from the SSIS toolbox.



Fact table Data Flow

Deploying the Cube:

From the selected measures one can analyze data based on the Cube formed using SSIS tool. Cube is the multidimensional demonstration of the selected data. There are two ways to deploy the cube- Automated and Manual. In this project cube is deployed with automated method. For the automation SSAS is used. As best practice sequentially, dimensions are processed and then the cube is processed.



Automation of Cube Deployment