ETL Project Report
Comparative Presidential Executive Orders (From President Truman to President Obama)

Adote A
Mohammad Shams
Emmanuel Olawuyi

The dataset for this project is derived from data.world website. The data are US presidential executive orders from 1945 to 2015. Datasets included ".csv" files for documents of executive orders, topics and subtopics and their descriptions. In order to start the analysis, we had to extract the data using Panda into Jupyter Notebook. After providing the notebook's dependencies, the datasets were read and displayed.

After the extraction, we have to transform the data by extracting the columns needed for the project from the four dataset derived from data.world website. Also, the columns were renamed and some data types were converted to float format. We also merged the data frames obtained from CSV files and were able to create two tables "description" and "executive".

Table "executive" provides us information on Executive Order number, the year signed, the name of the President and the their political parties and the number categorizing the topic of the executives orders.

Table "description" provides us information of the executive orders, the topics and the subtopics codes and their title with a description of the executive orders.

All the transformed data obtained from the ".csv' files and combined into two tables are stored in SQL database named "ETL_db". The database and tables were created on SQL using SQL scripts and from Jupyter notebook, we loaded the data into the tables into the SQL workbench.

The obstacle we faced in this project included data type mismatch, renaming columns in order to create a effective merge of the dataframes and also in order for us to insert the values for the table we had to use "replace" instead of "append" in the code below:

executive_df.to_sql(name='executive', con=engine, if_exists='replace', index=True).

Lastly, the time for the project was insufficient, but we finally managed to submit on time.