**ETL Project Report (Dow, NasDaq..)**

**Introduction**:

The purpose of the project is to extract data from the available websites, analyze the data, cleanup the data, transform the data to required formats, prepare the SQL table to store the data and upload the data into the DB tables.

**Datasource**:

The data was taken from <https://data.world> and <https://kaggle.com> and from <https://finance.yahoo.com> , initially some of the CSV files are downloaded and analyzed, the required CSV files are categorized and used for analysis purposes.

The data from finance.yahoo.com is web scrapping project, the real time data are web scrapped and analyzed.

**Data files:**

The “**sp.csv**” file has the data for “Standard and Poolrs 500” indices, the data that is available

Date,   
Price (closing price),   
Open (opening price),   
High ( the day high price reached),   
Low( the day low price reached),   
Volume ( the data volume) and   
Percentage of change

The column values are

|  |
| --- |
| ﻿Date,  Price,  Open,  High,  Low,  Vol.,  Change % |

This CSV file is parsed, the required data is transformed to numeric after removing char from the data like “,”.. and processed, it is successfully transformed and converted to json object and uploaded into the DB table.

The column names are rearranged and successfully uploaded into the DB table.

|  |
| --- |
| "Date": "trans\_date",  "Price": "price",  "Open": "openvalue",  "High": "dayhighvalue",  “Low": "daylowvalue",  "Change %": "percent\_change" |

The associated table

|  |
| --- |
| create table sandp (  trans\_date date,  price float,  openvalue float,  dayhighvalue float,  daylowvalue float,  percent\_change text  ) |

The other file chosen is “**bitcoin\_data.csv**” this file has columns

|  |
| --- |
| ﻿Timestamp,  Open,  High,  Low,  Close,Volume (BTC),  Volume (Currency),  Weighted Price |

The columns are rearranged and successfully loaded into the DB table.

|  |
| --- |
| "Timestamp":"trans\_date",  "Close":"price",  "Open":"openvalue",  "High":"dayhighvalue",  "Low":"daylowvalue",  "Volume (BTC)":"volume\_btc",  "Volume (Currency)":"volume\_currency",  "Weighted Price":"weighted\_price" |

The data is cleansed, cleaned, transformed to resolve all the numeric, and other data types issues. The table associated to the csv

|  |
| --- |
| create table btc\_usd (  trans\_date date,  price float,  openvalue float,  dayhighvalue float,  daylowvalue float,  volume\_btc integer,  volume\_currency integer,  weighted\_price float  ) |

**WebScrapping:**

The yahoo website finance section is taken for analysis, since yahoo has numerous data on their page, scrapping the data is much challenge and lots of show stoppers on getting the required data. The most active link on finance website is used to get the top 100 active symbols of the data.

<https://finance.yahoo.com/most-active> , the notebook is developed to web scrap the finance website and get the required data for the symbol, the table associated to the screen scrapping has the below columns

|  |
| --- |
| create table yahoo\_most\_active (  trans\_date date,  symbol text,  company\_name text,  price float,  day\_change text,  percent\_change text,  volume float,  avg\_volume float,  market\_cap float  ) |

The data transformation rule

1. Remove the “,” from the data that has to be stored as numeric
2. Apply the M rule for the data that needs to be stored as numeric and as M shows as Million, convert to float and multiply by 1000000
3. Apply the B rule for the data that needs to be stored as numeric and as B shows as Billion, convert the data to float and multiply by 1000000000
4. The same for Trillion T
5. If there are no data available, update the numeric value to 0
6. If the data stored as string, update as NA

Prepare the json object after applying the rule and update, load the json into the DB table. The sample data after screen scrapping, transforming, cleaning

|  |
| --- |
| [{'symbol': 'VLO',  'company\_name': 'Valero Energy Corporation',  'price': '41.97',  'day\_change': '+0.69',  'percent\_change': '+1.67%',  'volume': 5801000.0,  'avg\_volume': 4588000.0,  'market\_cap': 17114000000.0,  'trans\_date': '2020-10-23'},  {'symbol': 'KR',  'company\_name': 'The Kroger Co.',  'price': '32.76',  'day\_change': '+0.12',  'percent\_change': '+0.37%',  'volume': 6026000.0,  'avg\_volume': 7645000.0,  'market\_cap': 25367000000.0,  'trans\_date': '2020-10-23'},  …. |

The same process for Trending yahoo website and Bitcoin Yahoo website

Trending: <https://finance.yahoo.com/trending-tickers>

BitCoin: <https://finance.yahoo.com/cryptocurrencies>

**The other DB SQLs:** that are used for Trending and Bitcoin

|  |
| --- |
| create table yahoo\_most\_trending (  trans\_date date,  symbol text,  company\_name text,  price float,  day\_change text,  percent\_change text,  volume float,  avg\_volume float,  market\_cap float  )  create table yahoo\_btc\_coin (  trans\_date date,  symbol text,  company\_name text,  price float,  day\_change text,  percent\_change text,  total\_volume float,  currency\_supply float,  market\_cap float,  currency\_value float,  currency\_value\_old float  ) |

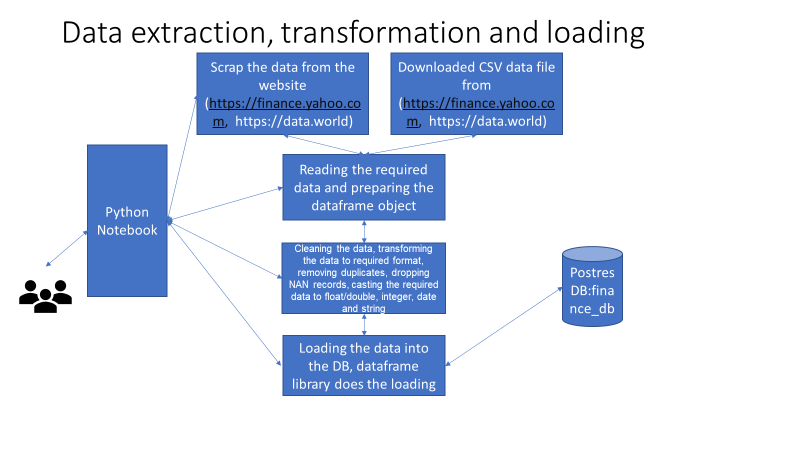
**Jupyter Notebook:**

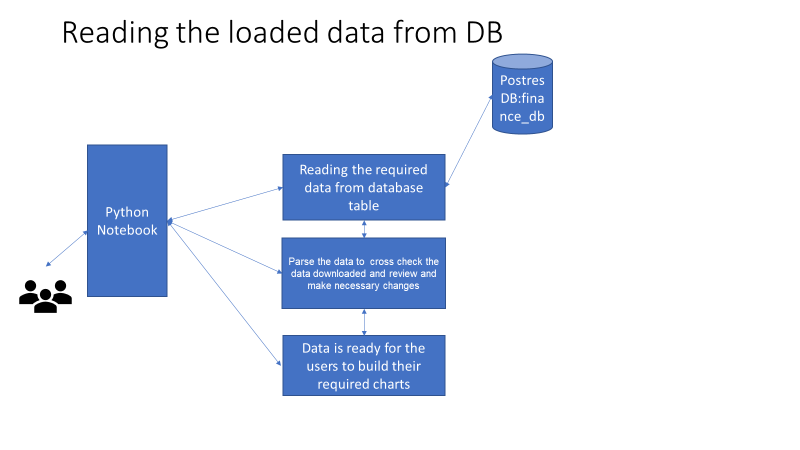
1. Data\_etl.ipynb ( used to process S and P data from CSV, source <https://data.world> )
2. BitCoin\_Data.ipynb ( used to process BitCoin data from CSV, source <https://data.world>)
3. Web\_etl.ipynb ( does web scrapping from <https://finance.yahoo.com/most-active> website and processes and stores in the DB table)
4. Web\_etl\_trending.ipynb ( does web scrapping from <https://finance.yahoo.com/trending-tickers> website and processes and stores in the DB table)
5. Web\_etl\_bit\_coin.ipynb ( does web scrapping from <https://finance.yahoo.com/cryptocurrencies> website and processes and stores in the DB table)

The data are stored in database table and completed review and validations.

ETL Overview:

* The main purpose of this project is to extract the data from the website data.world and yahoo.com to get the data for dow, Nasdaq, SandP indices, we also included some of the analysis for BitCoins.
* The financial data includes daily volume, open value, close value, day low, day high, these datas are segregated for each trading day.
* Most of the industries, manufacturing comes and technology based companies are traded in Dow and Nasdaq.
* The SAndP 500 provides the standard value of the economy, this depends on the market value of the indices traded in Dow and Nadaq.
* The BitCoins are traded through BlockChain, where user needs to create their own account and buy the Bitcoin, there are BitCoins for each Currency and most standard Bitcon comes in USD, users could buy Bitcoin in fractions and it is tradable all the time, there is no specific time range limits on it.
* In this project we planned to take top 100 indices on each segment and do necessary transformation and upload the data in the postgres DB tables.
* In this analysis we planned to take indices, price open, price change, total volume, avg 3 months from the data website.  
  In the transaction the datas are transformed to double, integer, string objects in standard format to store in the table.
* The symbolic representation of M ( millions) and B(Billions..) will be converted to required double/integer values.
* The data that are available to download are in viewable and formatted data, these data needs to be updated to make it readable for ourpurposes.
* Some of the example datas are, the currency values comes as “M” ( stands for million), similarly for B (Billion) and T(Trillion), this needs to be converted to required numerical value.
* The currency value has currency Symbol and “,” in it, this needs to be replaced with empty char and so easily uploadable as integer/float/double value to the DB table.
* The date comes as non standard format and this needs to be converted to required date format to import into the DB table., also this format changes across different websites, some common rule/module needs to be in place to process those data.
* Most of the data downloaded has “empty space” at the beginning or at the end, this needs to be trimmed.
* Some of the duplicate records needs to be dropped
* The NAN record needs to be dropped
* Before making a column as index, there are issues some time data doesn’t meets unique value constraint.
* Some of the records has special characters that could not be uploaded into the DB table, this needs to be dropped.
* Some time if required value comes as “-” value, this needs to be removed in case if this comes in required integer/double value.
* So each data set needs to go through the process to make required corrections, modifications, transformations, before uploading into the DB table.
* ETL plays more critical role in extracting the data from website, transforming the data to required formats and loading the data.





The above diagram shows the flow chart, interactions on modules in the notebook