Data Analytics for Microsoft Stock

Mini Project

Surbhi Sonkiya

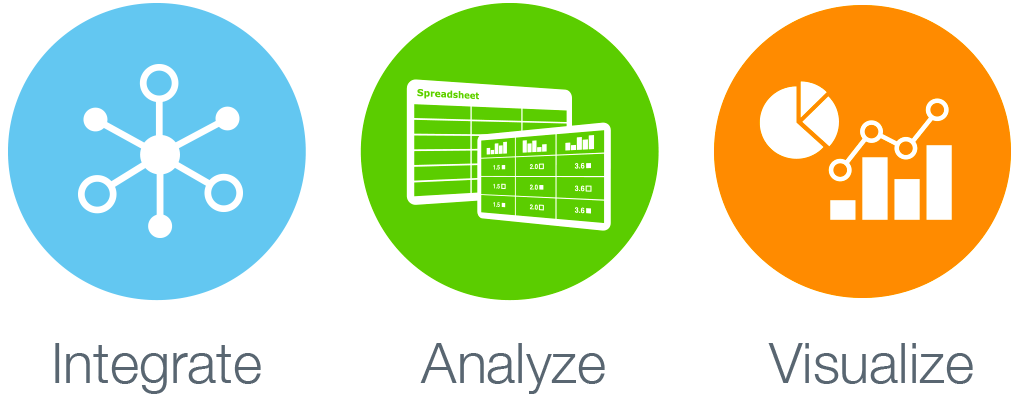
(EIT Digital Master School)

## Introduction

To perform data analytics on the historical stock market dataset and to plot the graph of price variation over several years.

## Aim of the project

To understand basic process flow of bigdata i.e. Integrate, Analyze and Visualize.



#### Integrate

Though this means to pull in datasets from various data sources and integrate them into one output dataset to be used for next process. In this mini project, I have pulled the dataset from internet. I have transformed the acquired dataset into csv format and supplied this static dataset to next process where I read this csv file and generated spark dataframe.

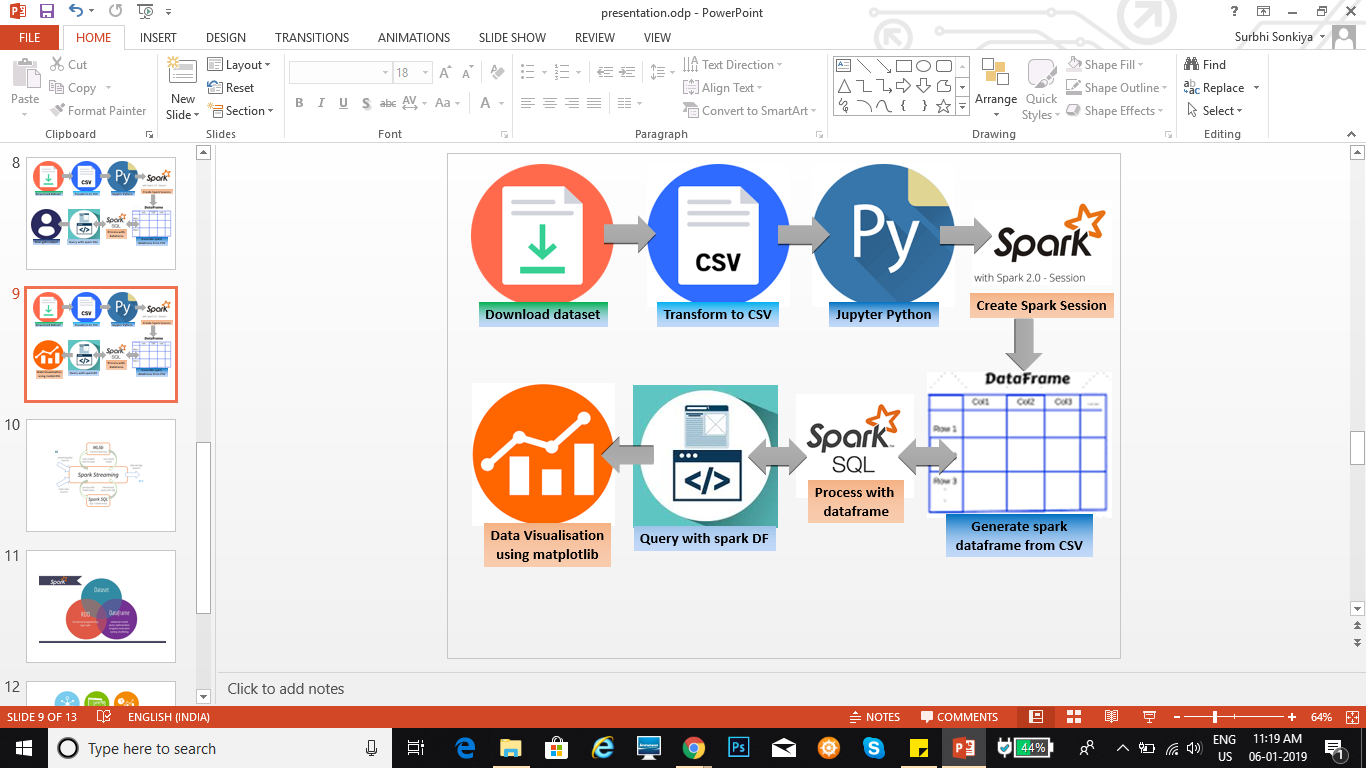
#### Analyze

I have used spark dataframe to perform data analytics and derive some insights (explained in detail below) about the data. I also explored another aspect of Dataframes. It is called Feature Engineering, which means, adding of new features into the existing dataset to derive further insights from the data. For this mini project, I added one feature which is explained below.

#### Visualize

It is not everyone’s cup of tea to gather insights from big datasets. Also, not everyone knows how to extract valuable information from big datasets. However, depiction of some data in visual format is easily understandable even by a layman. Also, I believe people in general prefer visual depiction of data rather than tabular format. For this reason, I decided to plot the visualization of the variation in prices over the past 18 years for Microsoft stock.

## Project workflow

****

#### Explanation

**Step 1**: I choose the dataset of Microsoft shares. So the first step was to acquire the dataset from Google or Yahoo finance. On research, it was found that these finance API by Google and Yahoo are shut down. Another alternative was to download the dataset via executing a VBscript. I collected dataset from 01-Jan-2000 to 31-Oct-2018 for the purpose of this mini project.

**Step 2**: I generated spark dataframe from the dataset and performed data analytics on that. For this I have used spark dataframe and python to extract below mentioned results from the dataset.

On research it was found that for the purpose of this project, querying on spark dataframes had advantages over using Spark SQL. Therefore, spark dataframes has been used more widely in this project. However, I have also used Spark SQL at some places to understand the concept and its implementation in general.

* What are the column names?
* What does the Schema look like?
* Print out the first 5 columns.
* Use describe function to learn about the DataFrame.
* Formatting the column values to limit the values to show up to two decimal places.
* What was the peak ‘High’ and which day had the peak ‘High’ in price?
* What is the mean of the ‘Close’ column?
* What is the max and min of the ‘Volume’ column?
* How many days was the closing lower than avg. value of ‘Close’?
* What percentage of the time was the ‘High’ greater than threshold?
* What is the max ‘High’ per year?
* What is the average ‘Close’ for each Calendar Month? I.e. across all the years, what is the average Close price for Jan, Feb, Mar, etc.

##### **Feature Engineering**

* Created a new dataframe by adding a new feature called HV Ratio i.e. the ratio of the High Price versus volume of stock traded for a day.

**Step 3**: Results and Data Visualization: for the purpose of this mini project, I have used matplotlib to plot the graph of price vs year to visualize the variation in price over several years. I have also explored D3 for visualisation of the same data. With D3, I plotted the visualisation in Bubble and Sunburst format.

## Required Dependencies

* Python 3
* Spark 2.2.0
* Linux (Ubuntu 16.04)
* Jupyter Notebook

## Future scope

A machine learning model can be trained using the past dataset and several machine learning algorithms can be applied to predict the future price of the stock. It will help investors know the predicted price of the stock based on the data analytics performed on the past dataset. This will give some insight to the investors whether to invest in a particular stock or not. How profitable the stock can be in future and how much can they expect to benefit from it.

## References

1. https://spark.apache.org/docs/2.2.0/streaming-programming-guide.html#dataframe-and-sql-operations
2. https://spark.apache.org/docs/2.2.0/sql-programming-guide.html
3. http://moonshinenews.com/tips-leverage-cloud-big-data/
4. https://www.bloomberg.com/professional/blog/3-ways-big-data-changing-financial-trading/
5. https://www.analyticsvidhya.com/blog/2014/08/big-data-profit-stock-market/
6. https://databricks.com/blog/2015/02/17/introducing-dataframes-in-spark-for-large-scale-data-science.html
7. http://www.signalsolver.com/download-historical-stock-price-data-excel/
8. https://d3js.org/