Alain Magazin - Gabriel Zamy

Project Proposal

Cloud & Big Data Final Project

The U.S stock market is one of the oldest and biggest financial market in the world. The New York Stock Exchange (NYSE) traces its routes all the way to 1792 and therefore stores (with its newer counterpart the NASDAQ) a immense quantity of data about the the history of the financial sector in the USA, about the industrial progress and the general economy.
The history of sock markets is therefore a very large data set and to analyse it efficiently we must use modern technology like Big Data processing. With Big Data we can extract valuable information out of raw data and use to analyse historical events, stock market performance throughout the years and connect all of this to general modern world history.

The data set we have comes from the website www.kaggle.com where a user posted a ready-to-use data set with all records of the NYSE and NASDAQ since their creation. Originally, this user used the yahoo finance api to extract this data. This data set is comprised of four folders (two of which we will not use because not completely relevant to our project, S&P 500 and Forbes 2000). We will instead use the NASDAQ and NYSE data sets which record the stock price details of all traded assets in the history of these two markets. The format is the following: each traded asset has its file in both CSV and JSON (we will use CSV). Each line is a date (a trading day), and comprises of columns detailing the stock for that day (opening price, close price, volume, etc…). In total, the two folders count for 6.52 GB of data.

To achieve the goals of our project we will use various tools seen in class of Cloud & Big Data. The most important being pyspark, as it is the best solution for Big Data processing in python in our opinion. We will use it to process our data and give ut results that we will then print and plot using various python libraries such as Matplotlib or Numpy. This project will be located in its own GitHub repository and will have a webpage hosted on GitHub describing it, showing and explaining the results and detailing our work process.