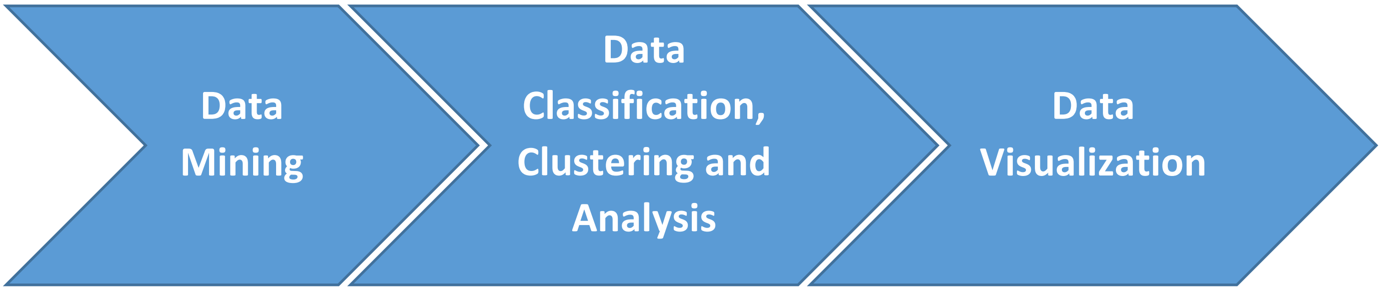
Proposal for Final Project

Group 11

Analyzing and Predicting the Impact of Market Events on Stock Prices

[redacted]

The main aim of this project is to analyse and predict the impact of market events on stock prices. We spilt this project into 3 phases:

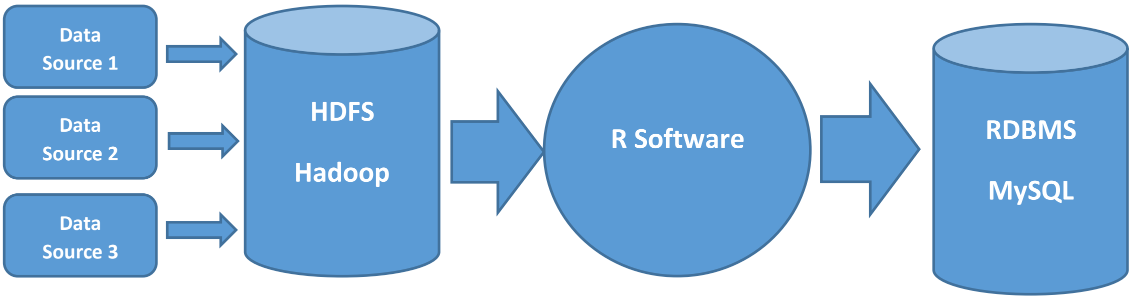


Data Mining:

We are planning to pull the real time market events data using latest concepts Natural Language Processing (NLP). For this, we plan to use text mining package of R – “tm”. In this project, we are planning to explore this package in detail and exhaust its different functions and utilities from different sources viz. Google Finance, Yahoo news coming in different formats like RSS feeds, web news, etc.

We firmly believe that this will give us an opportunity to have a deep dive into “tm.plugin.webmining” package. In this phase of the project, we will explore and demonstrate how we can mine the unstructured data coming from different sources, understand the source format and experiment other third party API viz. OpenNLP etc. We also plan to use other concepts viz tokenization and vector etc. During this phase, we will also look into other R packages for web scraping viz. “rvest” etc.

During Data mining exercise, we may use HDFS to store unstructured information coming from different sites. After we clean the data, we will store it in RDBS database viz. MySQL or others.



Data Mining Architecture

Data Classification, Clustering and Analysis:

After successfully importing the market events data of specific companies, we have plans to classify the data and analyse it further clustering it using machine leaning libraries in R.

The analysis will be divided into two main parts:

1) **Pattern Identification:** In this section, we will run different logistic regression analysis and derive the pattern and correlation between market events and movements in stock prices.
2) **Near Real Time Prediction:** Once a pattern is identified, we will train the model and make prediction for stocks using machine learning algorithms like classification, SVM, random forest. R has several machine learning algorithms available to do prediction. We can use Logistic Regression, k-Means clustering, k-Nearest Classification, Naïve Bayes, Decision tree and Support Vector Machine to analyze the data. We will try to achieve maximum accuracy and precision with different algorithms and compare their results.

Data Visualization:

We will do an extensive exploratory data analysis to draw insights about the data. We will plot histograms, boxplots, scatterplots, and heat maps.

We have plans to future study and expand the scope of this project in following different dimensions:

1. **Real time prediction:** We can enhance the HDFS architecture and implement Apache Spark and Strom framework to capture the instant market events and do real time predictions of stock prices.
2. **Finding the coefficient of impact of market events on Stock Price:** It will be most challenging to derive the coefficients of impact of micro and macro market events on the stock price. We believe that Deep Learning will be the most effective algorithm to find these coefficients. We will include this as a part of future scope of this project.

Note that depending on the time available, we will limit the scope of above phases. We may limit the scope of this project depending on the progress and analysis outcome.

Sample Data

For Data Mining

For Micro Events:

- Data source will be fetched from different website - Yahoo Finance, Google Finance, Google News and Market news and press releases and web corpora from Dow Jones Factiva, LexisNexis etc.

Other Macro Economic Events:

- GDP (Federal Reserve Bank)
- 10 Year Treasury Rates (Federal Reserve Bank)
- Unemployment Rate (US Bureau of Labor Statistics)
- Consumer Price Index (CPI) (US Bureau of Labor Statistics)
- Producer Price Index (PPI) (US Bureau of Labor Statistics)
- Retail Sales (US Census Bureau)
- Foreign Exchange Rates – EUR, JPY, GBP (various currencies)

Tools

- R Studio
- MySQL Database
- Hadoop HDFS Database (may use)
- AWS Cloud (may use)
- Third party APIs viz. OpenNLP and others etc. (may use)

References

- Class Notes, Lectures, Slides by Prof. Dragos Bozdog
- Class Notes, Lectures, Slides by Xingjia Zhang
- Charu C. Aggarwal, Data Classification: Algorithms and Applications. CRC Press, 2015. (ISBN: 978-1-4665-8674-1)
- Charu C. Aggarwal, Data Mining. Springer, 2015. (ISBN: 978-3-319-14141-8)
- Deborah Nolan and Duncan T. Lang, Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving, CRC Press, 2015. (ISBN: 978-1-4822-3481-7)
- Norman Matloff, The Art of R Programming, No Starch Press, 2011. (ISBN: 978-1-59327-384-2)
- Cathy O'Neil and Rachel Schutt, Data Science, O'Reilly, 2014. (ISBN: 978-1-449-35865-5)