Cracking the Stock Market with Spark

Methods of collecting insights for large and distributed datasets 100x faster

Redwan Ibne Seraj Khan, Computer Engineering’19, UB Advanced Honors College

Introduction

The stock market is one of the most vital parts of the economy of a country. It plays an important role in the growth of the economy. So, a stock market prediction is considered an attracting point for financial investors [3]. As the stock market is full of uncertainty, its prediction is one of the most important exertions in business and finance [2]. Traders need to take decisions regarding which stocks will generate the most profits. Hence there lies great importance in the analysis of stocks. This poster will show the limitations of the existing models and propose a new model basing on older models to address the concerns of the modern age.

Types of Analysis

Fundamental Analysis

Fundamental analysis is the analysis of stocks through social media data by performing sentimental analysis. As stocks get influenced with people’s motives and selling patterns, this analysis is of vital importance to traders [1].

Technical Analysis

Technical analysis is the analysis of stocks by using historical data prices of stocks [2].

Related Work

Keeping in mind the importance of the stock market, efforts have been made for its analysis. *Kanade et al.* have proposed an architectural model using Hadoop and MongoDB [2]. Hadoop is a way to distribute very large files across multiple machines. It uses MapReduce, which is a way for splitting a computation task. It consists of a Job Tracker which sends code to multiple task trackers for computation. The task trackers then allocate CPU and memory needed for the tasks and monitor them on the worker nodes.

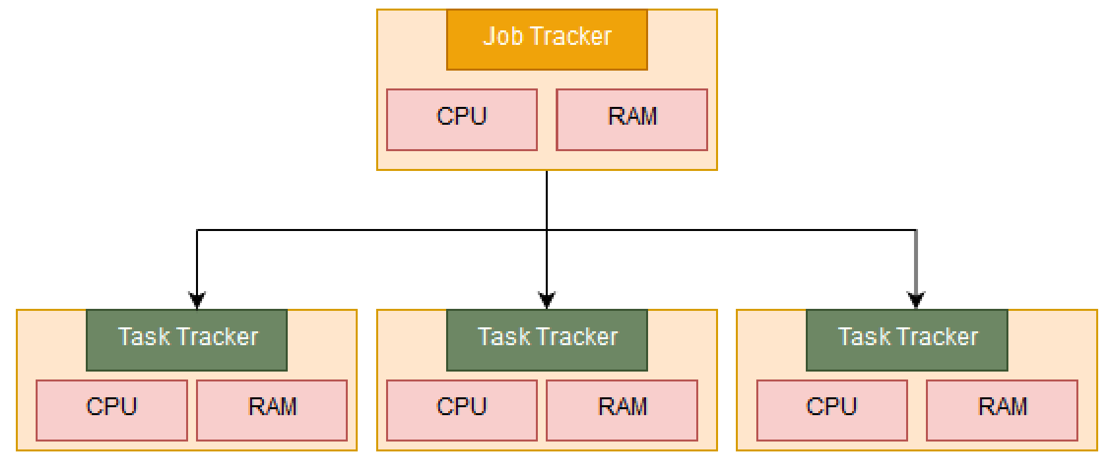

Figure: MapReduce Architecture

Although this model works, it would not be able to address the needs of fast computation. *M. M. Seif et al.* has come forward with another model [3] using spark which is much faster. Both of the models does not take into account people’s transactions and security issues while evaluating the market. Moreover, the model in [3] limits itself to using only HDFS as data storage and the model in [2] to MongoDB.

Proposed Architectural Model

The model that I am proposing would use Spark for computations. For the storage of data I would use HDFS, Cassandra or AWS S3 as either one of these will be able to work with spark. For the sake of clarity, I would just explain the architecture of HDFS and why it would be beneficial in this regard.

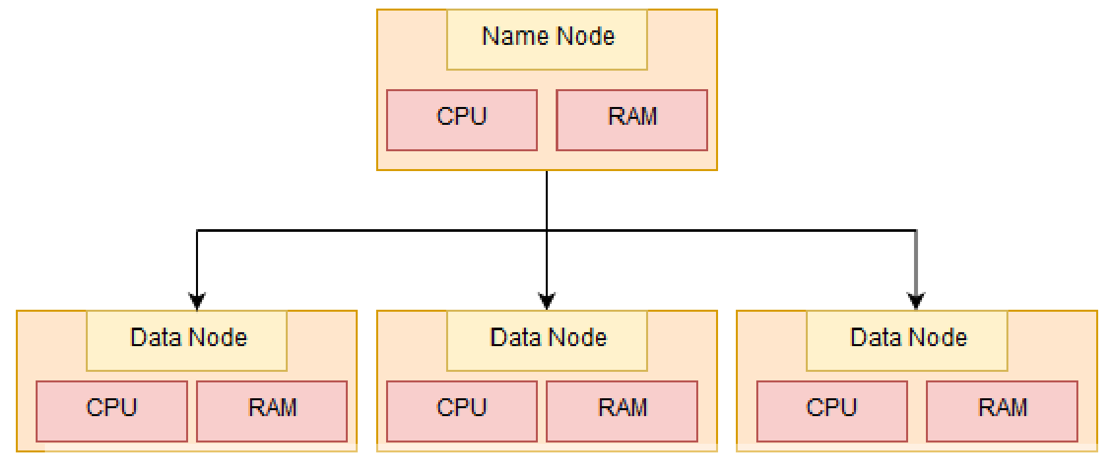

Figure: HDFS Architecture

HDFS uses blocks of data with a size of 128 MB by default. Each of these blocks is replicated three times and are distributed in such a way that supports fault tolerance. As the blocks are small, more parallelization can be performed during computations. As each block is replicated, so failure of a node does not cause loss of data. Fast computational power of spark coupled with HDFS will truly make a remarkable model.

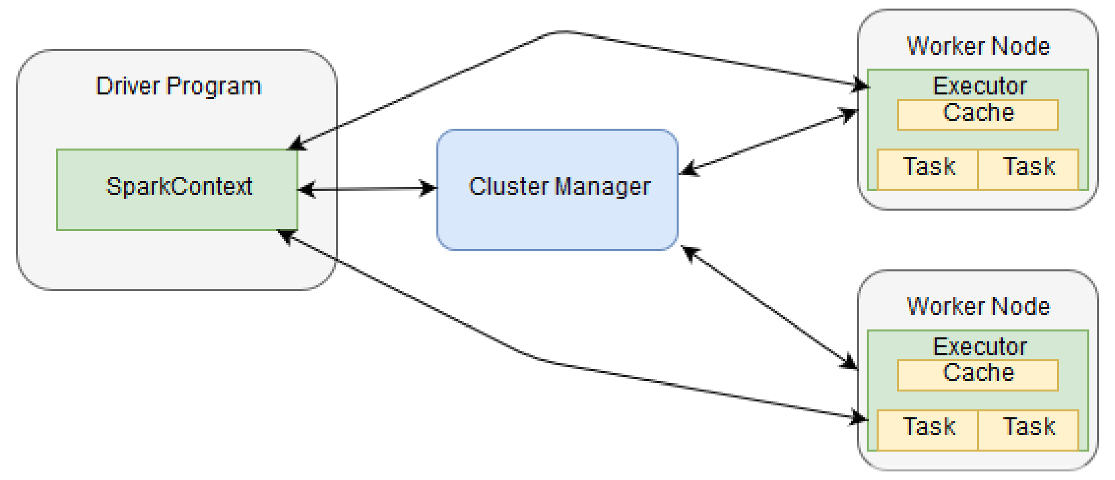

Figure: Spark Architecture

The main difference between Spark and MapReduce is that the latter writes most data to disk after each map and reduce operation whereas the first keeps most of the data in memory after each transformation and can spill over to disk if memory is filled.

Spark uses RDDs (Resilient Distributed Datasets) which are fault tolerant, parallely operated and has the ability to use many data sources. Moreover, RDDs are immutable, lazily evaluated and cacheable. Through a series of transformations and actions Spark can basically process everything that exists in data analysis.

Given the higher speed of Spark, it would give results in a much faster time than Hadoop. People’s transactions can have a lot of effect on the prices of stocks. Price of a stock can increase significantly if it gets popular among the general people. For added security and to establish trust among traders, all of the real time data is passed through a layer of blockchain.

All of the collected data would be pre-processed and features would be extracted which would then be used for classification. Spark Dataframes serve as the standard way of exploiting Spark’s Machine Learning library which provides great tools for manipulating data and predicting.

Towards the end a binary classifier would suggest the user if buying/selling a stock would be the best option to consider. The users will also have a web interface which would allow them to keep track of the various decisions that they have taken from the system. The analyzed data is further passed on to the storage which would later be used for future predictions.

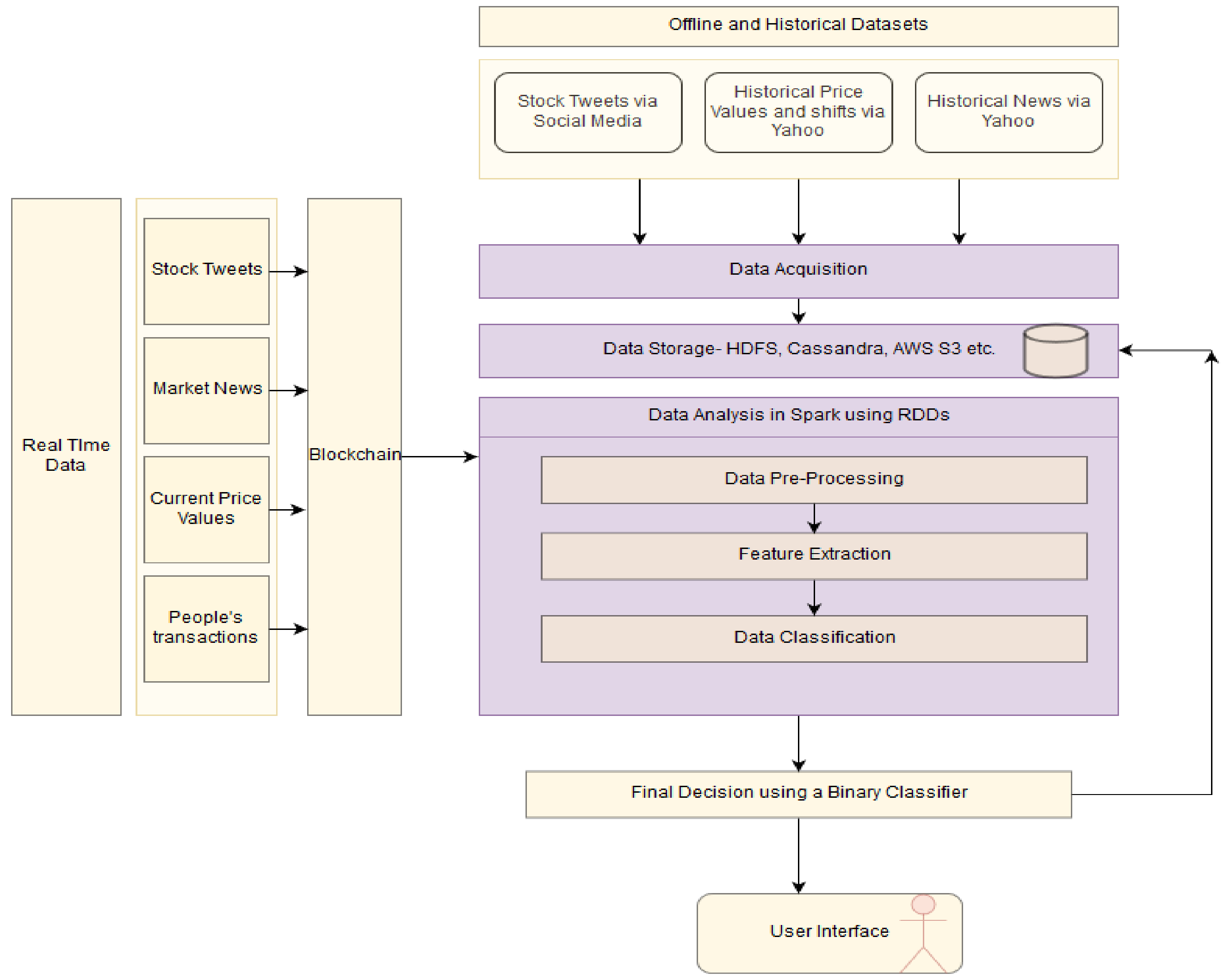
Offline and Historical Datasets
Stock Tweets via Social Media
Historical Price Values and shifts via Yahoo
Historical News via Yahoo
Real TIme Data
Stock Tweets
Market News
Current Price Values
People's transactions
Blockchain
Data Acquisition
Data Storage- HDFS, Cassandra, AWS S3 etc.
Data Analysis in Spark using RDDs
Data Pre-Processing
Feature Extraction
Data Classification
Final Decision using a Binary Classifier
User Interface

Figure: Proposed Architectural Model

Future Work and Conclusion

Spark is becoming increasingly popular in today’s age. If its power can be harnessed in the right way, computations and analysis in different fields can be done in a much faster and efficient way. Moreover, with the advent of newer machine learning models it remains to be seen just how accurate we can get in analysis of stock market data with Spark.

References

1. Drakopoulou, Veliota. "A Review of Fundamental and Technical Stock Analysis Techniques." (2016).
2. Kanade, Vivek, et al. "Stock market prediction: using historical data analysis." *International Journal* 7.1 (2017).
3. Seif, Mostafa Mohamed, Essam M. Ramzy Hamed, and Abd El Fatah Abdel Ghfar Hegazy. "Stock Market Real Time Recommender Model Using Apache Spark Framework." *International Conference on Advanced Machine Learning Technologies and Applications*. Springer, Cham, 2018.

University at Buffalo The State University of New York

Department of Computer Science and Engineering
School of Engineering and Applied Sciences