Big Data Applications [Stock Trade Analysis]

Table of Contents

Team 17

Bharat Nagaraju – A0178258N

Vigneshram Selvaraj – A0178215A

Note* - Keeping the template simple, since it was told to avoid headers and other formats, because it's easy to consolidate.

Introduction and Literature Survey

One of the most challenging, yet important tasks in the financial sector is financial stock analytics. Traditional methods focused on quantitative analytics. However, data such as tweets and other unstructured big data sources is important, which requires the organization to build scalable well architected and engineered big data system for analytics.

The project highlights the objectives of our big data system for publicly traded US equity analytics. The process is challenging for the team, but under the guidance of our lecturers, we managed to build the big data analytics system successfully. This report highlights our learning journey in building the financial big data analytics system.

Financial assets worldwide are estimated to be worth over US$300 trillion, with worldwide publicly traded companies' market capitalization at nearly US$100 trillion [1]. With so much at stake, it is not surprising that one of the most important research in the financial sector is to obtain accurate updated or real-time forecast of financial instrument prices.

Traditionally, the forecast methodology for financial instruments, such as publicly traded equities, are quantitative and statistical methods such Time Series forecast and Stochastic techniques [2] [3] [4], analyzing on just pure quantitative and structured data inputs such as historical prices data and financial report figures.

In recent years, researchers and traders have started to incorporate real-time unstructured information such as tweets, real-time news [5] [6] into their forecast systems and use other more advance machine learning techniques for the prices forecast. This meant that the forecasting system not only must be capable to be scalable pipelines with real time processing and it must also have a scalable well-designed Data Lake for all the data. In addition, the entire big data architecture must also integrate analytics processing capabilities.

Below list of literatures have been reviewed to design the system currently in question

Books	[1] J. Zhang, R . Shan and W . Su, “Applying Time Series Analysis Builds Stock Price Forecast Model,” Mathematical Models and Methods in Applied Sciences, vol. 3, no. 5, p. 152, 2009. [2] R . Weron and A . Misiorek, “FORECASTING SPOT ELECTRICITY PRICES WITH TIME SERIES MODELS,” Econometrics, vol., no , p. , 2005. [3] C. N. Babu and B. E. Reddy, “A moving-average filter based hybrid ARIMA-ANN model for forecasting time series data,” Applied Soft Computing, vol. 23, no. , pp. 27-38, 2014. [4] M . Arias, A . Arratia and R . Xuriguera, “Forecasting with twitter data,” ACM Transactions on Intelligent Systems and Technology, vol. 5, no. 1, p. 8, 2013.
Journal	[1] “Global market cap is about to hit $100 trillion,” Business Insider, 2017. [Online]. Available:

https://www.businessinsider.sg/global-market-cap-is-about-to-hit-100-trillion-2017-

[2] R . Feldman, B . Rosenfeld, R. Bar-Haim and M. Fresko, “The Stock Sonar — Sentiment Analysis of
Stocks Based on a Hybrid Approach,” , 2011.
https://aaai.org/ocs/index.php/iaai/iaai-11/paper/viewpaper/3506

[3] “IEX Trading Executive Team,”
http://www.iextrading.com/about/

[4] . Staff, “Stock Tweets App Integrations,”
http://stocktwits.com/developers/docs/integrations

[5] Big data applications
https://apifriends.com/api-streaming/open-source-apache-big-data-projects/

[6] Impact of Social Media on Box Office: Analysis of twitter activities on Best Picture Nominees | Oscars 2018 with Python
https://towardsdatascience.com/impact-of-social-media-on-box-office-analysis-of-twitter-activities-on-best-picture-nominees-7961c5c8ba40

[7] Simplify Streaming Stock Data Analysis Using Databricks Delta

https://databricks.com/blog/2018/07/19/simplify-streaming-stock-data-analysis-using-databricks-delta.html

[8] Simplifying Streaming Stock Analysis using Delta Lake and Apache Spark
https://databricks.com/blog/2019/06/18/simplifying-streaming-stock-analysis-using-delta-lake-and-apache-spark-on-demand-webinar-and-faq-now-available.html

[9]Streaming financial data
https://medium.com/@yy2799/realtime-financial-market-data-visualization-and-analysis-using-kafka-cassandra-and-bokeh-eac22139e5

[10] Low Latency Streaming
https://www.xenonstack.com/blog/real-time-streaming/

[11] K. S. Umadevi, A. Gaonka, R. Kulkarni and R. J. Kannan, "Analysis of Stock Market using Streaming data Framework," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, 2018, pp. 1388-1390.

[12] C. Lee and I. Paik, "Stock market analysis from Twitter and news based on streaming big data infrastructure," 2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST), Taichung, 2017, pp. 312-317.

	[13] Min Song, Meen Chul Kim, "RT2M: Real-time Twitter Trend Mining System", *Proceedings of IEEE International Conference on Social Intelligence and Technology*, May, 2013 [14] Hana Alostad, Hasan Davulcu, "Directional Prediction of Stock Prices Using Breaking News on Twitter", *Journal of Web Intelligence*, vol. 15, no. 1, 2017.

Problem study

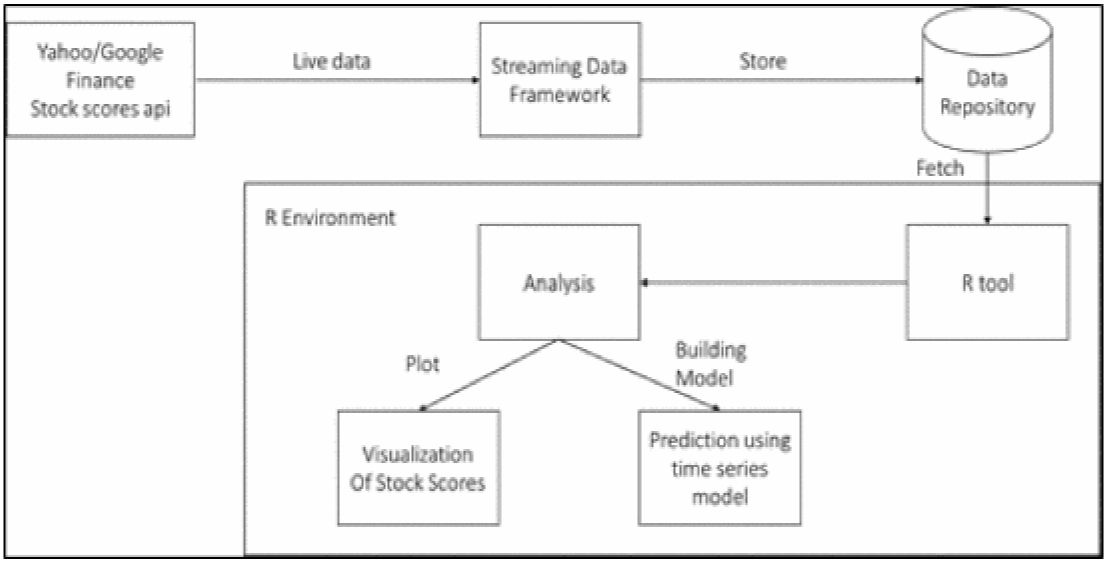
With the rapid development of web, mobile apps and Internet of Things (IoT) devices, a huge volume of data is created at every moment. More than millions of people using social media to contact friends or share their life. These data may contain important event or people's feeling. In fact, these data are more important than you think. They can be used to increase the business revenue, like the personalized recommender system or chat bots. For better customer experience and business revenue, many companies faced the problem of how to process the real-time data. The most important problem is traditional big data frameworks aren't designed for the real-time data. These frameworks usually use batch processing which processes tasks one by one. Like the Apache Hadoop, it is designed for large data processing. Researchers and engineers started to develop new frameworks for processing real-time data. More recent frameworks such as Apache Storm, Apache Spark, Apache Fink were introduced to process the real-time data. These frameworks can process large streaming raw data.

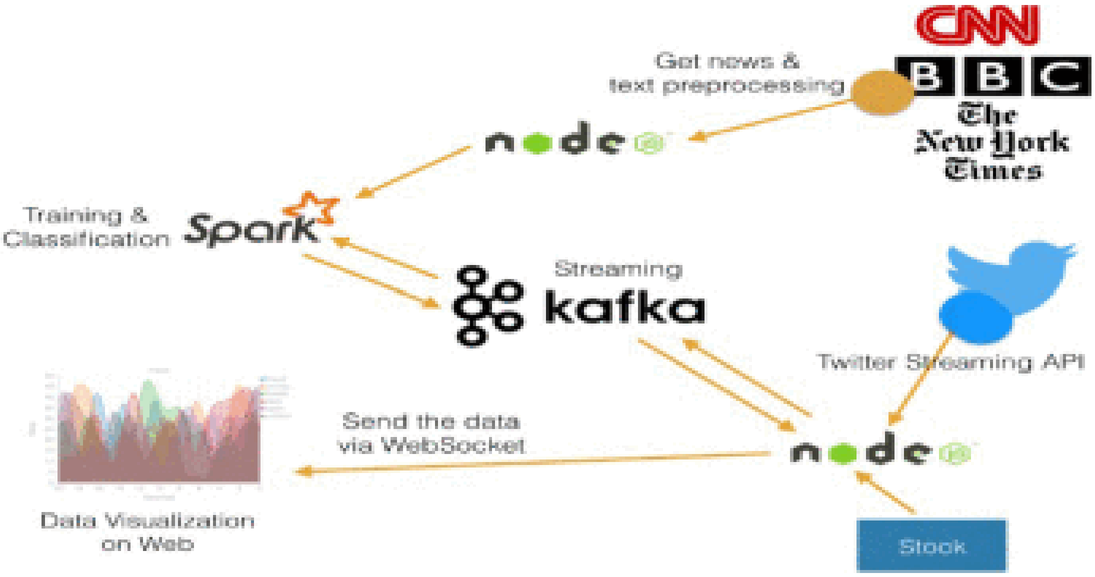
On the other hand, the stock market is an interesting element to analyze. Many people invest money on the stock market. Take an example of stock market. When a company announce its intention to buy another company, news media posts this information in website or social media. The stock value will be affected by Investment. Because they may buy or sell the stock, the company's stock value will rise or fall. The stock market and social media are changing quickly every moment. The trends of the Twitter [13] can affect the stock market. Taking advantage of real-time processing with big data framework, we can find the correlation between stock and Twitter [14].

In this research, we will collect Twitter tweets, stock value and news. For real-time usage, the tool will get the tweets and the stock value in every second. The news tool can get the news from the popular website and do text pre-processing. Next, we use the news data to train the classification model to classify the real-time tweets. With the classified tweets, we put these data and the stock value on the dashboard for visualization. In this dashboard, we can see the real-time trend and the stock market.

Observations and analytics model formulation

High level intended design for the system:




Get news & text preprocessing
CNN
BBC
The New York Times
node
Training & Classification
Spark
Streaming
kafka
Twitter Streaming API
Send the data via WebSocket
node
Data Visualization on Web
Stock

The main objective of this project is to:

a. Design and implement a scalable big data system with well-designed data lake that can ingest and store both real-time traditional structure price data and real-time unstructured data

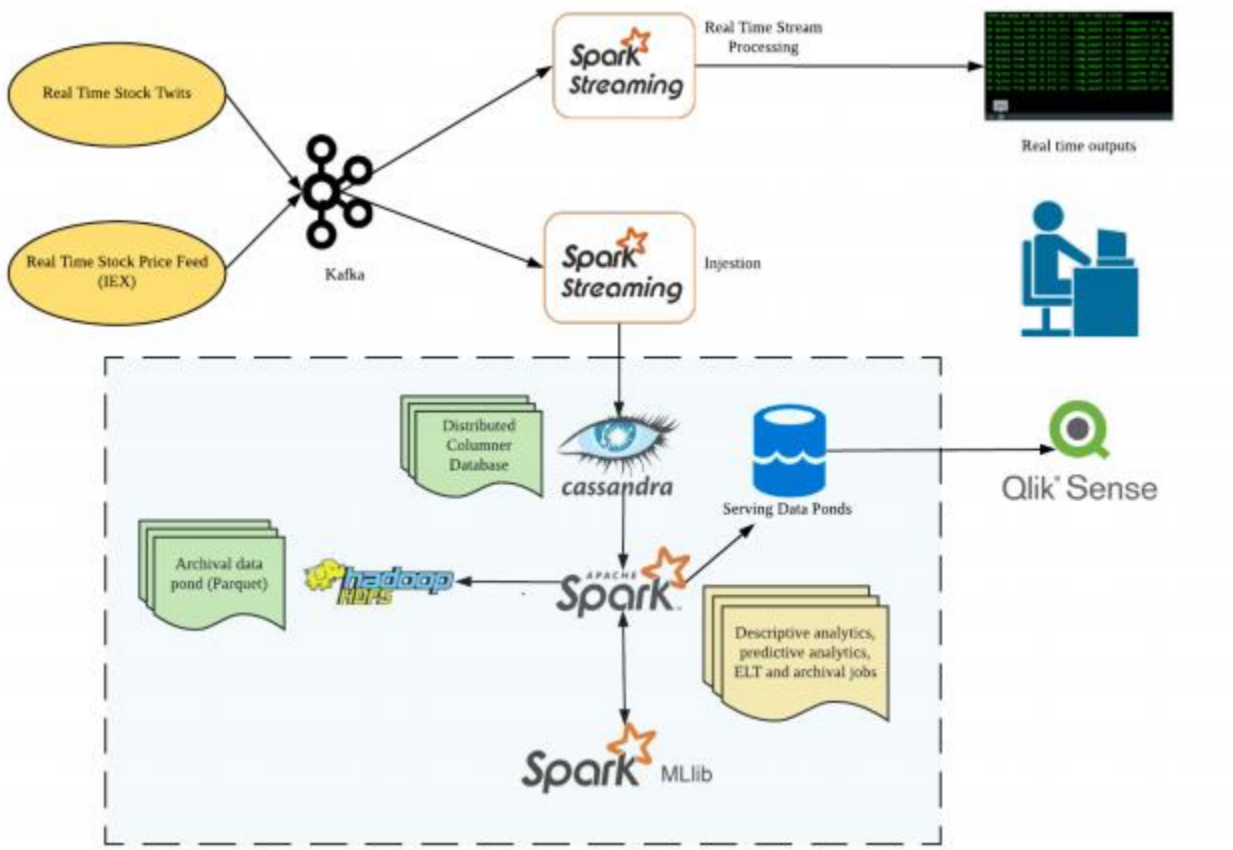
b. The big data system must be capable of real-time stream processing to extract insights from Realtime financial data

c. Integrate machine learning techniques to forecast financial instrument prices the expected output of the system must be able to give users a variety of insights on the financial instrument. o Real Time Streaming insights/analytics

d. Dashboard for near real-time or batch processing analytics and price forecast of the financial instruments

Solution Proposal

Overall Architecture:


Real Time Stock Twits
Real Time Stock Price Feed (IEX)
Kafka
Spark Streaming
Real Time Stream Processing
Real time outputs
Spark Streaming
Injestion
Distributed Columner Database
cassandra
Serving Data Ponds
Qlik Sense
Archival data pond (Parquet)
hadoop HDFS
Apache Spark
Descriptive analytics, predictive analytics, ELT and archival jobs
Spark MLlib

1. There are two real-time data producers which fetch data from 2 different REST APIs. The stock quote API provides updates on the real-time price of the stocks from IEX and the stock twits API provides tweets related to the stock.

2. The producers continuously fetch responses from these APIs and push the retrieved JSON to Kafka topics.
3. The data from the Kafka topics are then processed by Spark streaming jobs in Realtime. There are two category of jobs, one that performs real-time aggregations and visualize in a console and other that pushes to Cassandra.
4. The data at rest in Cassandra are then processed by 2 categories of Spark batch jobs. The first category performs aggregations on the static data and the other category performs batch machine learning.
5. The results from these batch jobs are saved in separate tables in Cassandra.
6. A separate batch job performs archival by routinely converting the data stored in Cassandra to Parquet files. Qlik Sense is used to visualize the results of the batch processing into a dashboard using the Cassandra connector.

Platform Options:

We have currently considered Databricks Cassandra stack along with Kafka. We may use alternatively CDH or Cloud [GCP] based on the computational needs and ease of implementation

Storage:

Currently, Cassandra is planned to be persistence storage for SPARK and Machine Learning jobs and incoming real time data flow from Kafka. However, we may switch it to HBASE or Hive depending on the implementation.

Batch processing:

For Data ingestion, it is performed using Kafka. Batch processing is handled in Spark. MLLIB will be used to generate machine learning models.

Data predictions from Machine learning job and the aggregation results for summarization are stored in Columnar NOSQL Store to be used by any Visualization Tool

Visualization:

Qlik Sense is built on the same core 'in-memory' Data Indexing technology as QlikView and offers dynamic graphs and powerful visualization. Based on the stack used, we may alternatively use POWERBI or Tableau as well to generate visualizations.

Conclusion

In the current system, we use the steaming processing big data architecture to classify Twitter tweets in the real-time. Our system can handle the large data with scalability and fault- tolerance. It's easy to change the current implementation for another real-time analytics usage. For example: The real-time log analysis which is widely used in data centers.

In the future, we want to apply natural language processing or Deep learning. Also, optimized streaming workflow allocation can be studied to get more efficient processing larger streaming data from SNS or IoT services.