# Master Thesis Proposal

Ville Vainio

March 7, 2019

## 1 Context

The modern economy revolves around stock market. Stock market is way for companies to obtain capital which they can invest into their own business. In exchange, person who invests into the companies stocks technically owns a piece of the company. The investor can make profit by selling these stock in a higher price or by receiving dividends from the company itself.

The price of the stock is simply determined by the law of supply and demand. If somebody is willing to pay a higher price for the stock then the price of the stock can grow. Because of this the stock market is in continuous fluctuation where people are selling and buying the stocks with the price they think the stock is worth using stockbrokers as the middleman. [1]

There are many strategies on how to invest into these stocks which depend on multiple factors such as; how much do you expect to profit with your investment, how much are you willing to take risk, do you want to make money by selling the stocks or by receiving dividends and so on. The underlying principle with every strategy is to minimize the risk you need to take in order to gain as much as profit as possible. Some of the strategies are based on subjective evaluation of the companies, but more technical strategies use metrics that are calculated from the financial statistics or the real-time market values. Strategies using the former data are called fundamental analysis and the strategies using latter data technical analysis. Neither of these approaches can predict the future of the market, but can statistically decrease the probability of larger losses in the market for the investor altough the probability of large losses is still not zero with these methods. [2]

Fundamental analysis is based on the idea that each stock has a intrinsic value that can be larger than the actual price of the stock in the market and buying these will eventually lead to profits.[3] The fundamental analysis focuses on the financial metrics that consist of companys overall statistics. These are for example how much the company has made profit, how much the company has paid dividends and what is companys cash flow. These tell a lot about the growth of the company and how the future of the company looks like. These metrics are usually published quarterly four times a year and present more long-term statistics about the company. Because of this, the amount of data these values present is quite small in terms of space.

The technical analysis that focuses on the real-time market values, on the other hand, needs new data almost daily. Stock exchanges are usually open from morning, opening around 8 to 10am, until evening, closing around 5 to 7pm on weekdays. Before and after this there are more limited pre- and after-hours trading which lasts usually around 1 to 2 hours depending on the exchange in which more limited stock trades can be made. During these hours multiple values are recorded on the prices of the stock from which the most important ones being: the highest price the stock was sold, the highest price the stock was sold and the number of stocks traded during the time interval.

The technical analysis focuses on finding recognizable patterns through this data. [4] Where the data used by the fundamental analysis was relatively small, these values can generate gigabytes of raw data in a week.

Company **A** has created a software application which is mostly focused on automatically perfoming fundamental analysis on companies in Finnish stock markets (Nasdaq First North and Nasdaq Helsinki) and serve these results to clients. The architecture of this application can be found in figure 1. The architecture consist of data ingestion (1), data storage (2) and data analysis (3) components which are hosted in Google Cloud Platform. The components (1) and (3) are implemented as NodeJS applications which are run on Google Cloud Functions (1) and Google App Engine (2). This system works great on Finnish stock market which has around 100 companies listed but does not scale well. The system shows clear delay in values when the number of companies is raised to over 1500 on the part of data ingestion.

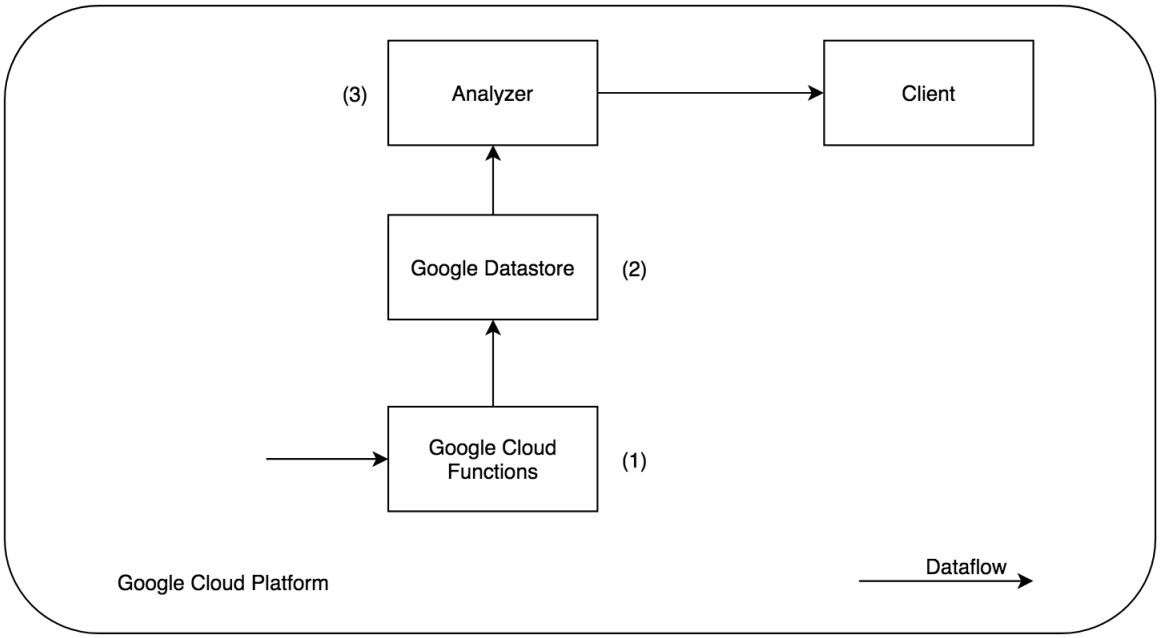


Figure 1: The architecture of company **A**

## 2 Research Questions

The company **A** would now like to include U.S stock market, which consists of around 5000 listed companies, into the system but with this architecture and technologies this is not possible without introducing massive delays in the calculated values. Company **A** would also like to add technical analysis into the analysis step to get more accurate results and more potential clients, but the current ingestion system prevents also this larger flow of data. The objective of this thesis would be to solve these problems and implement a prototype system that the company **A** could use to scale their business.

Data ingestion in the context of stock data is also a problem that has not been researched that much. When searching with terms "big data" and "stock" the resulting papers are mostly focused

on the aspect of analysing the stock data and these papers mostly ignore the steps of ingesting and storing this data. Examples of this kind of papers are [5], [6] and [7]. The main questions this thesis would try to answer would be: What are the requirements and needs for this kinds of system in the context of stock data, based on this what are the state of art techologies to implement this kind of system efficiently in the context of the current infrastructure of company **A** and which solution would be the best cost/perfomance-wise.

# 3 Expected Outcome

For the first research question "What are the requirements and needs for this kinds of system in the context of stock data?" this thesis plans to provide an analysis of the necessary stock market data and its usages. From this analysis, the thesis would derive the main requirements for the system to fulfill in order to satisfy the needs of the possible subsequent analysis stage. This result could then be used in the future if one would want to build their own ingestion system from the ground up as a base.

For the second question "What are the state of art techologies to implement this kind of system efficiently?" the thesis would perform an analysis on the current trends in data ingestion solutions. The thesis would provide information on the latest open-source technologies that could be used to implement this kind of system, how these would fulfill the requirements introduced by the first research question and the needs of company **A** and conclude this with a comparison of these technologies on what are the advantages of using one over another. The result of this part could be used to decide what seems to be the most suitable technology to use to solve the company **A**s technical problem.

For the last question "Which solution would be the best cost/perfomance-wise?" the thesis would implement open-source prototype solution based on the results of the second research question. This prototype could be used by anybody (company or individual) as it is or as a base to build a more complex system on top of it. The system would be targeted for company **A**, which could use the system to scale their business into U.S market and possibly allow them to write more complex analysis algorithms based on the larger amount of data.

# 4 Approach

### Step 1

The thesis would start by going through scientific papers about stock markets and stock analysis. There would first text generally about the characterics of the stock markets, what are they based on, what affects them, can they actually be predicted (random walk hypothesis) and so on. Then the thesis would go through the main directions of analysing the stock markets (fundamental analysis, technical analysis) and explain briefly some of the methods (about three methods per direction) that characterize these directions (Gordon model, Magic formula, LSTM etc.) focusing on the data that these methods need in order to calculate their predictions. This part would be concluded by deriving the requirements for the system based partly on these analysis methods and their data needs, and partly on the definitions that make up a scalable, secure and stable cloud system.

### Step 2

The next step would be to perform literary research on what is currently used to perform big-data ingestion, selecting from the list of technologies mostly those that seem to fulfill the requirements derived in the step 1 and would fit in the infrastructure of company **A** introduced in figure 1. For data ingestion, these technologies would probably be Apache NIFI, Apache Flume, Fluentd etc. This step would consist of first introducing all of the selected technologies and going through how do they work, what are they supposed to solve and what are the advantages and disadvantages of using one. After this, the section would do a comparison of these technologies in the context of stock data and conclude with analysis on which of the technologies would be the most prominent ones to solve this problem.

### Step 3

After this would start the experimental part of the thesis. Based on the results of the step 2, I would implement couple of the most prominent solutions as a prototype systems. This step would include the actual implementations and reporting of these implementations. The report would consist of technical details; what parts does each of the system consists of, what versions were used, where the system was run etc. And subjective remarks; was it easy to implement, was there parts that did not fit together etc. The reporting part would also explain the metrics that would be measured for the subsequent analysis. For these metrics, the dataset used to test the system would be the open-source REST API from IEX [8]. These metrics could be, for example, the time it takes to batch process, processor usage, memory usage, database fetching times etc. These prototypes would be developed inside docker containers so the tools used to measure these metrics would most probably be programming language specific functions and tools that docker provides to measure runtime statistics.

### Step 4

In the final step, the thesis would first inspect the results from the step 3 and based on these make remarks on what could be the best potentially the best implementation in this context. After this there would be an wrap-up on each of the previous sections concluding in retrospective what could've been possibly done better and what could be done in the future, concluding in recommendation what could be based on this thesis the best posssible solution for the company **A** to use in their application.

## References


[1] John L. Person. *Mastering the Stock Market: High Probability Market Timing and Stock Selection Tools*. John Wiley & Sons, Incorporated, 2013.

[2] Justin Fox. *Myth of the Rational Market*. Harper Business, 2009.

[3] Söhnke M. Bartram and Mark Grinblatt. "Agnostic Fundamental Analysis Works". In: *Journal of Financial Economics (JRE)* (June 2017).

[4] John J. Murphy. *Technical analysis financial markets: A comprehensive guide to trading methods and applications*. New York Institute of Finance, 1999.

[5] Liu Lin Wu Yanbin Guo Yiqiang, Huang Ni, and Wang Li. "Trend analysis of variations in carbon stock using stock big data". In: *Cluster Computing* 20 (2017).

[6] Aghakhani Kiarash and Karimi Abbas. "A New Approach to Predict Stock Big Data by combination of Neural Networks and Harmony Search Algorithm". In: *International Journal of Computer Science and Information Security* 14 (2016).

[7] Shyu Jonchi Kao Yu-Cheng and Huang Jim-Yuh. "eWOM for Stock Market by Big Data Methods". In: *Journal of Accounting, Finance & Management Strategy* 10 (2015).

[8] `https://iextrading.com/`.