MSBD5003 Project Proposal
Real-time Stock Clustering and Prediction

ZHANG, Xichen 20527341
ZHENG, Dongjia 20546139
LI, Haoyang 20533364
WANG, Yuxian 20549997

1 Background

In recent years, increasing numbers of projects in stock market are conducted by big data technology. With the complex features, large volume of data as well as the real-time fluctuating price, and many other influential factors, in order to handle the task efficiently, big data technologies should be applied. Such platform can help organizations and individuals to having a better understanding of the stock market and making correct decisions in different situations.

2 Goal

We plan to do a wide project which aims to build a real-time stock clustering and prediction platform. In this platform, stock data (with indicators including code, name, changing ratio, trade, open, high, low, volume and so on) will be uploaded every second, so the system need to process the data in real time. In terms of clustering, the system will divide the stocks into groups based on their similarities. This will help users have a better understanding of the inner relationship between different stocks.

3 Datasource

Tushare is an open-sourced python financial data interface package. It is stable, free, fast data API which covers all the stocks data in China A-share market. Most importantly, it provides real-time stock data which satisfies the need (streaming data processing) of our project. Users can also obtain data in different granularities such as tick, minute, hour and day, etc.

Most of the data obtained from Tushare, include stock code, stock name, open time, close time, highest price, lowest price, bid price, volume and amount are all in data-frame format. In our project, we are going to feed real-time stock data into spark streaming modules to do real-time processing and analysis. These data can be obtained from Tushare by the corresponding interface easily. Additionally, it can group the stocks data by different sectors, industries, concepts and regions, so that we may find some interesting insights from the clusters.

4 Functions

This project plans to have the following basic functions: data preprocessing, modelling and visualization.

Firstly, stock data was collected using third-party APIs. At the same time, we need to do data cleansing, data storage, data integration and so on. Then, we cluster the stock into different groups in real-time. These grouping result may indicate that these stocks are influenced by similar factors including some industry information, cash flow and profits of some certain countries, etc. As a result, the clustered groups may change from time to time. Instead of predicting a specific stock price, we plan to predict the whole trend of stocks in a cluster. According to the financial information and patterns, we plan to do more feature engineering to predict stock price. Our models are incrementally trained, such that the newest prediction can always combine the latest news.

In order to show our results, we also plan to do visualization at the end, and make the result more visually attractive and intuitive. This visualization part will show the result in real-time.

We also plan to have some advanced functions to do more meticulous data processing. Such as more detailed OLAP and OLTP processing. In the visualization part, we can provide options for users to choose models they want, and to compare accuracy of different models. This may help the users to judge the stock classification, price prediction according to different models and have more insightful understanding of the stock market.

5 Techniques & Architecture

In this project, we are going to use several techniques in different layers to achieve the functions. As the flow graph shows, there are mainly three layers in our design: Data API Layer, Data Process Platform and Visualization Interaction Layer. The data API layer would be a wrapper of third party APIs to make our system compatible with multiple data sources. In this part, we will design a uniformed data structure as the input for the whole system. Some basic python packages such as pandas, numpy will be used here.

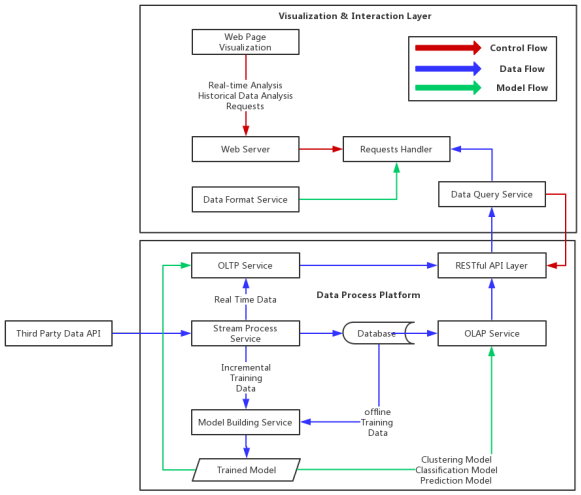

Figure 1: The total existing and proposed open space

The second layer which is the most important part of the system is Data Process Platform. It consists of multiple services with different functions. It supports the fundamental functions of the system. From the users perspective, OLAP and OLTP meet two data analysis demands. We can handle these two different tasks by using Spark Streaming and Spark SQL, which can provide a satisfying performance on the clustering process. Since this is a distributed system, we may use Apache Kafka to do distribute the streaming messages. Furthermore, we use MongoDB as the database to

support the OLAP business, as it is open-sourced and well designed for big data processing. In the bottom of this layer, the clustering, classification and regression algorithms will be driven by Spark MLlib which support distributed training. Moreover, we may use streaming linear regression and streaming k-means algorithm to support the OLTP business. The whole layer will encapsulated by Flask so as to provide REST APIs to the other modules.

As for the visualization and interaction layer, it would be a simple B/S architecture App which support the top business level. It will use tools such as E-Chart as visualization component, since E-Chart provide a good animation effects as well as simplest graph data format. We may use Django or Flask as the back-end of the app which will manage the business service such as graph format and query request. The front-end may use Vue.js as it is lightweight with powerful template and suitable for quick development. AngularJS and React JS are also under consideration.

6 EVALUATION

Our system will be evaluated from two perspective: system performance and model accuracy. System performance will consider from the real-time data processing efficiency and model building speed comparing to the traditional single server platform. And the accuracy will be evaluated by whether those stocks are in the same stock sector in the real life and whether this classifications groups are reasonably explained by some financial factors. At the same time, the prediction will be checked by the real time stock trend.

7 WORKLOAD & COLLABORATION

	ZHANG Xichen	LI Haoyang	ZHENG Dongjia	WANG Yuxian
Data Source		✓		
Database			✓	
Data Streaming			✓	
Model	✓			✓
OLTP & OLAP	✓			
Back-end Services	✓	✓		
Front-end Services			✓	✓
Document	✓	✓	✓	✓

REFERENCES

[1] Microsoft Azure Big Data Achitecture, *https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data/*.

[2] Spark Streaming Programming Guide, *https://spark.apache.org/docs/latest/streaming-programming-guide.html/*.

[3] Machine Learning Library (MLlib) Guide *https://spark.apache.org/docs/latest/ml-guide.html*.

[4] *TuShare 0.4.3 documentation*: http://tushare.org/trading.html

[5] *Flask User's Guide*: http://flask.pocoo.org/docs/1.0/

[6] *E-Charts Documentation*: https://ecomfe.github.io/echarts-doc/public/en/option.html

[7] *Django Documentation*: https://docs.djangoproject.com/en/2.1/

[8] *MongoDB Connector for Apache Spark*: https://www.mongodb.com/products/spark-connector