Team Analytics Project Proposal – First Draft

Part 1. General Information

Team Name (optional):

Team Members:
Yun-shao Sung
Hongzhi Ren
Jiajie Tang

Project Title:
Stock Analysis Intelligent System

Project Description (short description of your project):

The project provides the US stock investors an intelligent stock analytical tool to help them make wise investment decisions. The basic idea underneath is to effectively apply advanced analytical technologies (including statistical analysis, machine learning, data mining and natural language processing, etc.) under the framework of "Big Data Technology". New information related to stock price and company financial performance is being produced from various sources on a daily basis. In order to take advantage of the overwhelming amount of realtime information and to make timely investment decisions, investors need an intelligent system to help them get useful insights, not only historical facts, but realtime analysis results as well, in just a second. Our project is built on both historical and realtime data/information, and delivers investment analysis to users. Specifically, the first potential topic is to analyze the correlation-behavior between different stocks' price. We will apply MapReduce to figure out the pattern of price movement between any two stocks based on daily/weekly updated data sources. Stocks that have similar behavior or totally opposite behavior will be ranked and presented to users, who will use the information to dig out potential investment opportunities. The second topic would be financial report analysis for each company. Several important key-ratios that reflect the performance of the company will be calculated and compared across the industry/market. The complicated computation is also based on MapReduce method, and the data is updated quarterly/yearly. The third goal is to use NLP technology to parse text information/opinions from equity research websites. Sentimental analysis will be made and the result will serve as an indicator for investors. The last topic is to conduct machine learning algorithms for stock price or financial statement to dig out hidden information that might be useful for decision making.

Team Analytics Project Proposal – First Draft

Part 2. General Data Source Information

Data Sources - E.g. tweets	**Data Source Description (brief)**	**Data Size** Estimate size, e.g. MB? GB? TB?
1. MorningStar Inc.	Morningstar provides data on approximately 500,000 investment offerings, including stocks, mutual funds, and similar vehicles, along with real-time global market data on more than 14 million equities, indexes, futures, options, commodities, and precious metals, in addition to foreign exchange and Treasury markets. Morningstar also offers investment management services through its registered investment advisor subsidiaries and has approximately $170 billion in assets under advisement and management as of Dec. 31, 2014.	The size of the data source should be TB unit. But we will only use part of it (GB unit).
2. Bloomberg L.P.	Bloomberg L.P. is a global, multimedia-based distributor of information services, combining news, data and analysis for financial markets and businesses. Bloomberg provides real-time pricing, data, history, analytics and electronic communications 24 hours a day, 365 days a year and is used by over 250,000 financial professionals in 90 countries worldwide. The Bloomberg terminal is an essential tool for investors and other professionals planning to take advantage of the opportunities in Emerging Markets.	The size of the whole data source should be TB unit. But we will only use part of it (GB unit)
3. Google Inc.	Google Finance is a website launched on March 21, 2006 by Google. The service features business and enterprise headlines for many corporations including their financial decisions and major news events. Stock information is available, as areAdobe Flash-based stock price charts which contain marks for major news events and corporate actions. The site also aggregates Google News and Google Blog Search articles about each corporation, though links are not screened and often deemed untrustworthy. It contains up to 40 years of data for U.S. stocks, and richer	The size of the whole data source should be TB unit. But we will only use part of it (GB unit)

	portfolio options.	
4. GuruFocus	This site tracks the stock buys, sells and commentaries of guru investors such as Warren Buffett, Peter Lynch, and the best mutual fund managers.	The size of the whole data source should be TB unit. But we will only use part of it (MB unit)

Team Analytics Project Proposal – First Draft

Part 3. Detailed Data Source Information

Data Sources - From Part 2 above	**Data Characteristics** - Is data source a realtime source? - Is it realtime and stored (e.g. a log)? - Is it statically loaded data (e.g. historic)?	**Data Frequency** - If realtime data, what is the frequency?
1. Morningstar.com	The project will use stock price data from this source. It is a realtime data source. The financial report historical information is stored in its database.	The stock price is updated per second. The financial report data is updated per quarter/year. The other news/data are all realtime.
2. Bloomberg Terminal	Both statically loaded and realtime data.	The stock price, market data, portfolio data and so on are all updated per second.
3. Google Finance	Mostly realtime data/news.	Prices and news are updated per second.
4. GuruFinance	Mostly realtime data—stock research reports and opinions.	Reports and opinions are updated per second if there is new information coming out.
5.		

Team Analytics Project Proposal – First Draft

Part 4. Technologies

List technologies. Will your project make use of MapReduce? Pig? Flume? HBase? Hive? Impala? Spark? Mahout? Other?

Our team will try to implement following technologies: MapReduce, Pig, Hive, Mahou/Mllib.

Part 5. References

References – Please add references to all papers/articles read by the team (should be at least two references per team member).

1. Reference 1 : Machine Learning with Mahout, by Nibeesh K, MCS 10 206(P) Seminar II, Seminar Report

2. Reference 2 : Case Study Evaluation of Mahout as a Recommender Platform, by Carlos E. Seminario, David C. Wilson

3. Reference 3 : Dache: A Data Aware Caching for Big-Data Applications Using the MapReduce Framework

4. Reference 4 : Extending MapReduce across Clouds with BStream

5. Reference 5 : Stock Exchange Forecasting Using Hadoop Map-Reduce Technique, by KUSHAGRA SAHU, REVATI PAWAR, SONALI TILEKAR, RESHMA SATPUTE

6. Reference 6 : Processing of Kuala Lumpur Stock Exchange Resident on Hadoop MapReduce, by H. Law, S. Aghabozorgi, S. Lim, Y. Teh and T. Herawan

...