Stocks Analysis and Prediction Using Big Data Analytics

Zhihao PENG
Department of Computer Science, Dalian Neusoft Institute of Information, Dalian 116626, China
pengzhihao@neusoft.edu.cn

***Abstract*—Big data analytics are used primarily in various sectors for accurate prediction and analysis of the large data sets. They allow the discovery of significant information from large data sets, otherwise, it is hidden. In this paper, an approach of robust Cloudera-Hadoop based data pipeline is proposed to perform analyses for any scale and type of data, in which selected US stocks are analysed to predict daily gains based on real time data from Yahoo Finance. The Apache Hadoop big-data framework is provided to handle large data sets through distributed storage and processing, stocks from the US stock market are picked and their daily gain data are divided into training and test data set to predict the stocks with high daily gains using Machine Learning module of Spark.**

Keywords- Big data analytics, big data, stock market, machine learning

I. INTRODUCTION

Big data has been attached great importance for the proliferation of a lot of different sectors. It has been extensively employed by business organizations to formalize important business insights and intelligence. Furthermore, it has been utilized by healthcare sector to discover important patterns and knowledge so as to improve the modern healthcare systems. Besides, big data holds significant importance for the information, technology and cloud computing sector.

Recently, the finance and banking sectors utilized big data to track the financial market activity. Big data analytics and network analytics were used to catch illegal trading in the financial markets. Similarly, traders, big banks, financial institutions and companies utilized big data for generating trade analytics utilized in high frequency trading. Besides, big data analytics also helped in the detection of illegal activities such as: money laundering and financial frauds.

In this paper, we hope to build a system which analyses US oil stocks to predict daily gains in US stocks based on the real time data from Yahoo Finance. About all 13 stocks in US oil fund are picked up and their daily gain data are divided into training and test data set to predict the stocks with high daily gains using Machine Learning module of Spark. Based on our analysis we propose a robust Cloudera-Hadoop based data pipeline to perform this analyses for any type and scale of data.

By means of studying a live stream data of US oil stock prices so that it can help us better understand how does US Oil index affects the stock price of other stocks in US oil funds exchange. Besides it can help us predict the profitable stocks for stock traders and provide profits to US Oil stocks trader community.

II. PRELIMINARY KNOWLEDGE

In this paper, we wished to develop a machine learning model to predict the future crude oil prices using the USO(The United State Oil Fund,USO) data and understand which set of features are better in prediction using Hadoop framework. Some basic knowlege of the USO and Hadoop framework are reviewed in this section.

A. USO

USO, the United State Oil Fund, an exchange-traded fund (ETF) security, was designed to track daily price fluctuations of West Texas Intermediate light, sweet crude oil that is being delivered to Cushing, Oklahoma [1]. The objective of the USO is to monitor the daily changes in the percentage terms of its shares. Since USO tracks the near-month contracts listed on NYMEX, the data can be used to understand short-term fluctuations in the crude oil market. The daily closing prices of USO are available on the Yahoo Finance website and can be retrieved to generate various meaningful insights.

B. Hadoop framework.

Apache Hadoop is an open-source big-data framework providing a platform for handing large data sets through distributed storage and processing. The framework is based on the assumption that hardware failures are common and hence is designed such that it automatically takes care of all the possible system failures.

The ecosystem has Hadoop compatible File System(HDFS) and MapReduce at its core hence its very often called as Hadoop MapReduce framework [2]. This framework is a perfect example to showcase power of distributed and parallel computing. Applications to process vast amount of data can be easily written on Hadoop MapReduce framework. The data is processed on multiple machines (distributed architecture) parallelly (parallel computing) on Hadoop ecosystem. In a nutshell, the objective is to distribute the tasks across multiple clusters, which have a compatible file system.

The core of Hadoop Ecosystem is composed of Hadoop Distributed File System(HDFS) and MapReduce(As shown in Fig1.)

978-1-7281-1307-4/19/$31.00 ©2019 IEEE
DOI 10.1109/ICITBS.2019.00081

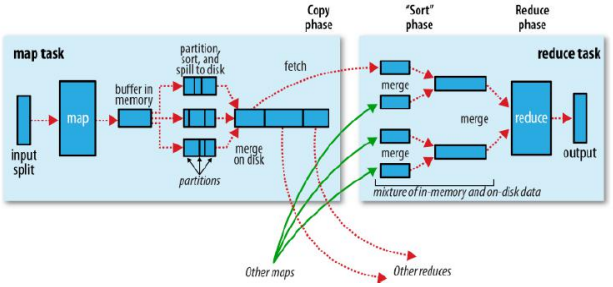


Fig 1.MapReduce Framework[2]

C. *Machine Learning in Spark*

Spark is a leading tool in the Hadoop Ecosystem. MapReduce with Hadoop can only be used for batch processing and cannot work on real-time data. Spark can work stand-alone or over the Hadoop framework to leverage big data and perform real-time data analytics in a distributed computing environment. It can support all sort of complex analysis including Machine Learning, Business Intelligence, Streaming and Batch processing. Spark is 100 times faster than Hadoop MapReduce framework for large scale data processing as it performs in-memory computations thus providing increased speed over MapReduce.

The big-data era has not only forced us to think of fast capable data-storage and processing frameworks but also platforms for implementing machine learning (ML) algorithms that has applications in many domains. With lot of ML tools available, deciding the tool that can perform analysis and implement ML algorithms efficiently has been a daunting task. Fortunately, Spark provides a flexible platform for implementing a number of Machine Learning tasks, including classification, regression, optimization, clustering, dimensionality reduction etc.

III. A CLOUDERA-HADOOP BASED DATA PIPELINE APPROACH

We hope to build a system which analyses stocks to predict daily gains in stocks market based on real time data from Yahoo Finance or other online resources. To illustrate the processes, we select the stocks from USO as example. Our problem is about selecting up all 13 stocks in US oil fund and divide their daily gain data into training and test data set to predict the stocks with high daily gains using Machine Learning module of Spark. Based on our analysis we propose a robust Cloudera-Hadoop based data pipeline to perform this analyses for any type and scale of data.

Our approach is to integrate multiple opensource modalities of Apache Hadoop ecosystems to build a cloud architecture which takes in real time data and process it to produce valuable information to support decision making. Through this approach we propose a cloud based Data Pipeline capable of scraping massive real-time data from Yahoo finance server and generating Machine Learning based insights into global oil stock market. The data are divided into training and test sets, we make our linear regression based learning model learn from the training data and then predict the correlation between stock prices based of coefficients in the regression model. We also calculate the R squared value and Mean Average Error to support our studies, the proposed pipeline is shown in Fig 2.

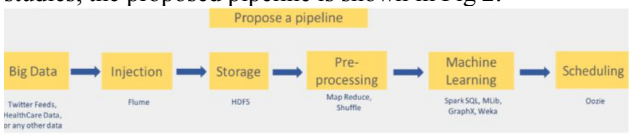


Fig 2. The proposed pipeline

To implement the proposed approach, there are generally five steps: Data Acquisition and Characterization, Data Injection, Storage, Pre-processing and Machine Learning(as shown in Fig 3).

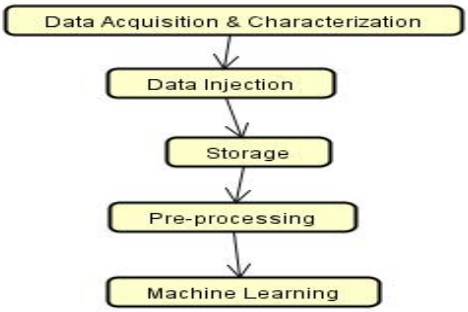
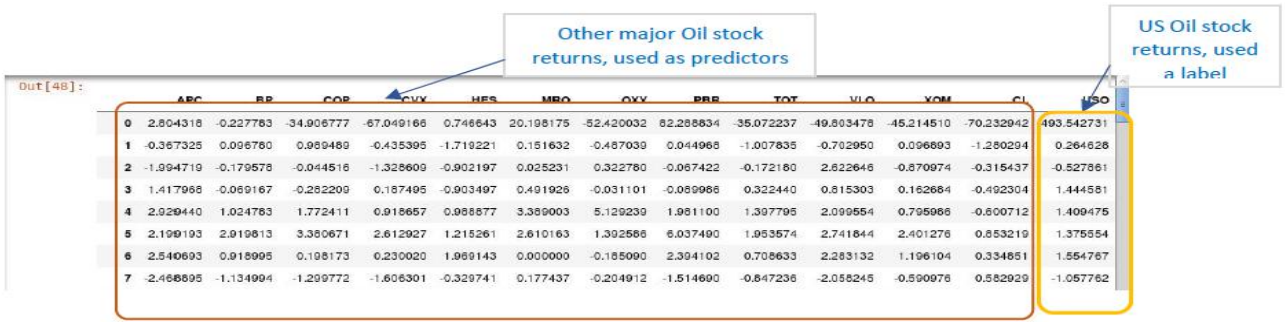


Fig 3. Workflow of the proposed approach

A. *Data Characterization*

Our data set included 13 oil stocks from SP500 stocks available on Yahoo Finance. We have studied prices of these stocks from April, 2006 till now. Starting date so selected was based on the USO (US Oil Fund) being made available for trading oil stocks in the US. We have 13 different numerical variables of float data type. The data we processed contained 2905 rows and 13 columns. There was no null value present in the dataset. The data set is also available in CSV format for local analytics. The returns observed were actually obtained after multiplying with 100.



	APC	BP	COP	CVX	HES	MRO	OXY	PBR	TOT	VLO	XOM	CL	USO
0	2.804318	-0.227763	-34.906777	-67.049166	0.746643	20.198175	-52.420032	62.286834	-35.072237	-49.803478	-45.214510	-70.232942	493.542731
1	-0.367325	0.096760	0.989489	-0.435395	-1.719221	0.151632	-0.487039	0.044968	-1.007835	-0.702950	0.096893	-1.260294	0.264628
2	-1.994719	-0.179578	-0.044516	-1.328609	-0.902197	0.025231	0.322760	-0.067422	-0.172180	2.622646	-0.870974	-0.315437	-0.527861
3	1.417968	-0.069167	-0.282209	0.187495	-0.903497	0.491926	-0.031101	-0.089986	0.322440	0.815303	0.162684	-0.492304	1.444581
4	2.929440	1.024783	1.772411	0.918657	0.988877	3.389003	5.129239	1.981100	1.397795	2.099554	0.795986	-0.600712	1.409475
5	2.199193	2.919813	3.380671	2.612927	1.215261	2.610163	1.392586	6.037490	1.953574	2.741844	2.401276	0.853219	1.375554
6	2.540693	0.918995	0.198173	0.230020	1.969143	0.000000	-0.185090	2.394102	0.708633	2.283132	1.196104	0.334851	1.554767
7	-2.468895	-1.134994	-1.299772	-1.606301	-0.329741	0.177437	-0.204912	-1.514690	-0.847236	-2.058245	-0.590976	0.582929	-1.057762

Fig 4. USO Stock Data

B. *Data Injection*

The Data was injected to HDFS using Flume. Three components worked in concert to push the data into flume, these three were Source, Channel and Sink. The source is accessed by the execution of tail command on shell, local memory acted like the channel and DFS was sink was the sink. The data injection needed configuring local database to HDFS. The following config files were used(as shown in Fig 5).

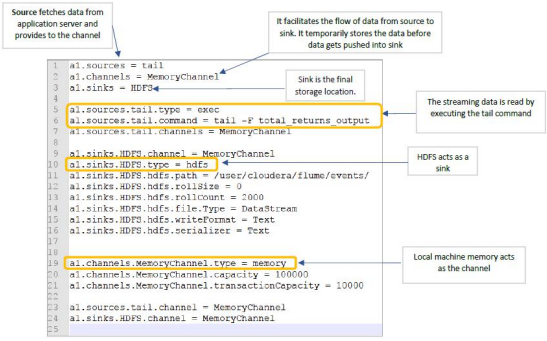


Fig 5. Configuration File

C. *Storage*

For storing the data HDFS on Cloudera was used. The directory address HDFS /Cloudera/flume/events/. The dataset wasn't replicated as we worked in single node. Also the copy of same data could be availed from Yahoo Finance if needed. Data always gets stored in multiple sequential file each entry represented only once(no duplication).

D. *Pre-processing*

PySpark which is the Python API for Spark was used to pre-process data. The sequence file is read into the spark using spark context. The sequential file is converted into a RDD (Resilient Distributed Datasets) . Converted the rdd into a data frame providing it with the suitable schema and specifying the data type of each input.

The input features are converted in to a single feature called features using the Dense Vector function. The dataframe which is ready to be fed into the machine learning has two columns label and the features.

E. *Machine Learning*

Machine was performed using Mlib function in Spark. The data was split into training and testing data sets. Linear regression function is fit to the training dataset. The returns of the USO are predicted for the training dataset. Computed the mean squared error(as shown in Fig 6).

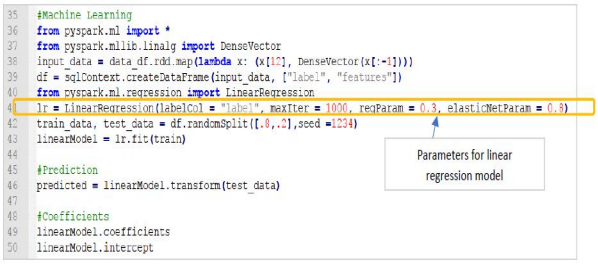


Fig 6.Machine Learning in Spark

IV. RESULTS AND DISCUSSION

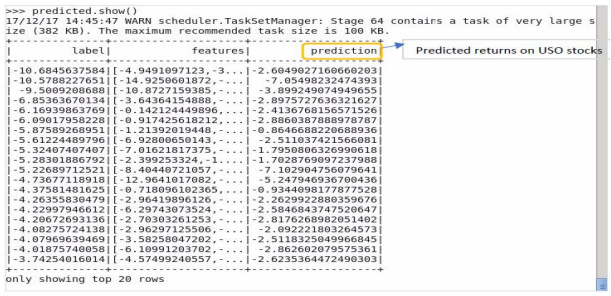
The data is divided into training and test sets and make our linear regression based learning model learn from the training data and then the correlation between stock prices is predicted based on the coefficients in the regression model. Also the R squared value and Mean Average Error are calculated to support the studies.

(1) Coefficients Regression Model

Based historical data get from Yahoo Finance, we can get the Coefficients in the problem are [0.0817,0.0,0.086,0.0,0.1282,0.0675,0.11,0.0438,0.0046,0.0, 0.0,0.0] and the intercept value of the regression model is -0.09793839625089426l, as show in Fig 7.

Fig 7.Coefficients of Regression Model

(2) Predictions from the test data(as shown in Fig 8.)



(3) Evaluation Metrics: R-squared value and MAE(as shown in Fig 8).

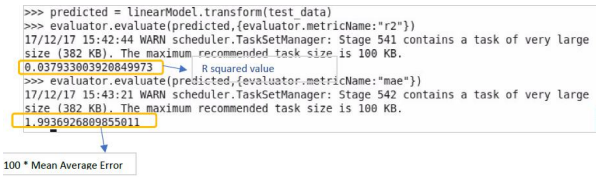


Fig 8.R-squared value and MAE

From the regression model we see not all coefficient values are positive which means that US oil stock prices are not positively correlated with the other US oil stocks that used as predictors in the study. The Model was built using regularization parameter equal to 0.3. The R squared value was calculated using the evaluation function from regression evaluator package from machine learning module of Spark. It explains 3% of USO stock price variations. The Mean Average Error was found to be 1.95% which suggests that linear model of regression is not suitable for predicting stocks return margins from the data with high dimensionality.

V. CONCLUSION

In this paper, the big data analytics are used for efficient stock market analysis and prediction. Generally, stock market is a domain that uncertainty and inability to accurately predict the stock values may result in huge financial losses.

Through our work we were able to propose a approach to help us identify stocks with positive everyday return margins, which can be suggested to be the potential stocks for enhanced trading. Such appoach will act as a Hadoop based pipeline to learn from past data and make decisions based on streaming updates which the US stocks are profitable to trade in. We also try to find scope of improvements to our study in future directions.

We intend to further our study by automating the analysis processes using scheduling module, then obtain periodic recommendations for trading the US stocks. We also plan to test some Neural Network model based learning rather than linear regression aims to accurately predict the US stock prices.

REFERENCES

[1] USO - USCF Investments. (n.d.). Retrieved from http://www.uscfinvestments.com/uso

[2] White, T. (2011). Hadoop: the definitive guide. Sebastopol, CA: OReilly.

[3]Evangelos Triantaphyllou, C.-T.L., Development and evaluation of five fuzzy multiattribute decision-making methods. International Journal of Approximate Reasoning 1996. 14(4): p. 281-310.

[4] Angadi, M. C., & Kulkarni, A. P. (2015). Time Series Data Analysis for Stock Market Prediction Using Data Mining Techniques with R. International Journal of Advanced Research in Computer Science, 6(6).

[5] Attigeri, G. V., MM, M. P., Pai, R. M., & Nayak, A. (2015). Stock market prediction: A big data approach. In TENCON 2015-2015 IEEE Region 10 Conference (pp. 1-5).