Big Data Analysis of Historical Stock Data Using HIVE

Jay Mehta, Jongwook Woo
1Graduate Student, 2Prof, Department of Computer Information Systems California State University Los Angeles, USA
E-mail: 1jmehta3@calstatela.edu, 2jwoo5@calstatela.edu

ABSTRACT
The purpose of this study is to apply Hadoop Big Data to financial analysis and to identify top companies whose volume are traded highest in past years. For this research, historical data of NYSE of each company from January 2000 to December 2014 has been taken. Cloud computing, Azure, is used for storing the historical data and for analyzing the data in Hive. The result contains top 10 companies by its highest volume traded for each industry (Basic, Consumer Durables, Technology, Energy, Transportation, Public Utilities, Consumer Non-Durables, Consumer Services, Capital Goods, Finance, Healthcare, and Miscellaneous). Additionally it has been found that the highest volume traded is for the company Mega Capitalization. It is shown that financial data analysis can be done efficiently and easily using big data technologies like Hadoop and its ecosystem Hive by using cloud services like Microsoft Azure.

Keywords: Hadoop, big data, hive, Data analysis, NYSE, financial data

1. INTRODUCTION
Machine-generated data is growing exponentially from last several years. This data is generated in social networking sites via posts from many users, sensor data: to get climate information, purchase transaction records in large industry and many more. With the help of normal legacy systems, it becomes very difficult and expensive to store and analyze large scale data for data analyst. It is also time consuming process. This kind of large scale data with structured and unstructured format is called Big Data. However, Hadoop framework is growing now a days to store and analyze data and it is convenient for its functions. Hive is one of the ecosystems in Hadoop framework which is built by Facebook to analyze the data on Hadoop cluster. Hive syntax is based on SQL, so a person with the knowledge of SQL can easily work in Hive environment. The syntax used in Hive is called Hive QL (Hive Query Language).

Many companies have been using big data framework to analyze the data and find some patterns and relationship among the data to target customer and market competition. In this study we have used NYSE historical data of 2,480 companies from January 2000 to December 2014 (15 years). We collect the data to keep in HDFS (Hadoop Distributed File Systems) and to analyze it using Hive on Azure blob storage to find top 10 companies by its highest traded volume for each industry.

The remainder of this paper is organized as follows. Section 2 summaries the NYSE data set. Section 3 describes the data analysis view. In Section 4, we describe the components of our data analysis systems and Hive codes. Section 5 discusses the results of our experimental evaluation. We conclude the research in Section 6.

2. HIVE AND NYSE Historical Data
This section briefly describes Hive and NYSE data set. Hive is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis. NYSE data set is an open data set downloaded from the Yahoo Finance [1, 2].

2.1 Hive [3, 4]
MapReduce codes are used to compute and analyze the Big data, But it requires high level expertise of Java, Ruby, Python and Perl. So, to reduce the cost and time Yahoo designed a new data flow language called Pig, which makes easy to analyze the data in 5-10 lines in Pig instead of 100 lines of Java code in MapReduce.

Hive and HiveQL platforms are designed by Facebook which has SQL like syntax. Hive runs on client machine and its queries are submitted to the Hadoop clusters on any local servers or cloud, which is transformed to MapReduce job. In this paper, we have used Hive and Microsoft azure for Hadoop clusters.

2.2 NYSE Historical Data [1, 2]
Daily stock data of each company is available live on yahoo finance for each stock exchange worldwide. We have taken the NYSE stock exchange data for this study. The data set is composed of: company symbol, date, open of the day, high of the day, low of the day, close of the day and volume. This paper is to present main 2 objectives: (1) Top 10 companies which have been traded highest by its volume by each Industry. (2) Top 5 highest volume traded of a specific company by its date.

There are 2,480 csv (comma separated values) files containing following fields:


3. IMPLICATIONS OF DATA ANALYSIS
In this paper, we make the following contributions:

- We download and combine all 2,480 csv files in a single text file using windows shell command.
- We create one storage account and one Hadoop cluster having 4 nodes on Microsoft Azure.
4. Experimental Results

We use ‘Cloud Berry Explorer Azure blob’ to copy the data set in cloud storage of Microsoft Azure. This software is open source available on CloudBerry Lab [5].

Hence we have one file containing data for all companies we create a single table using Hive QL on Hadoop cluster.

This research finds the companies whose volume of stock traded maximum from all the industries of NYSE stock exchange.

Additionally we can have highest volumes of a single company till December 2014.

4. METHODOLOGY

This section describes the Hive queries to create the table and store the data and to visualize the result.

4.1 Characteristics of Hive data set

The size of csv file in local machine is 314 MB. To execute the query, the data should be stored in a table. We first create a table to store the data in a meaningful manner. The query is as follows:

```
CREATE EXTERNAL TABLE nyse
(symbol STRING, date BIGINT, open_of_the_day FLOAT,
high_of_the_day FLOAT, low_of_the_day FLOAT, close_of_the_day FLOAT, volume BIGINT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',
STORED AS TEXTFILE LOCATION 'wasb://gradstudynyse@gradstudynyse.blob.core.windows.net/Folder_Hive/';
```

4.2 Experimental Results

We have taken the list of companies by Industry category from the NASDAQ. We include this list in our query to find the companies whose volume is traded highest in that particular industry. There are total 12 industries in NYSE stock exchange: Basic, Consumer Durables, Technology, Energy, Transportation, Public Utilities, Consumer Non-Durables, Consumer Services, Capital Goods, Finance, Healthcare, and Miscellaneous. Some example queries are presented as follows for industries:

**a. Healthcare Industry**

```
SELECT symbol, MAX(volume) as Max_Volume
FROM nyse
GROUP BY symbol
ORDER BY Max_Volume DESC
LIMIT 10;
```

**Table 1**: Top 10 companies by maximum volume

<table>
<thead>
<tr>
<th>COMPANY_SYMBOL</th>
<th>MAX_VOLUME</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFE</td>
<td>289,735,900</td>
</tr>
<tr>
<td>BSX</td>
<td>243,874,000</td>
</tr>
<tr>
<td>CVS</td>
<td>185,123,500</td>
</tr>
<tr>
<td>RAD</td>
<td>161,471,000</td>
</tr>
<tr>
<td>MRK</td>
<td>145,015,500</td>
</tr>
<tr>
<td>MDT</td>
<td>133,028,500</td>
</tr>
<tr>
<td>BMY</td>
<td>123,619,400</td>
</tr>
<tr>
<td>CI</td>
<td>110,733,900</td>
</tr>
<tr>
<td>BAX</td>
<td>98,811,400</td>
</tr>
<tr>
<td>JNJ</td>
<td>98,440,200</td>
</tr>
</tbody>
</table>

Table 1 and Figure 2 illustrate the top 10 companies in healthcare industry who trade maximum volumes in 2000 and 2014.

**b. Transportation Industry**

```
SELECT symbol, MAX(volume) as Max_Volume
FROM nyse
```

Figure 1: Data analysis system
GROUP BY symbol
ORDER BY Max_Volume DESC
LIMIT 10;

c. Other Industries
We do the same thing similarly for the rest of the industries. The symbols get change in the Hive QL by different industry. Here are Finance Industry in Figures 4 and Energy Industry in Figures 5.

![Figure 2: Top 10 companies by max volume in healthcare industry](image)

**Table 2: Top 10 Companies by Maximum Volume**

<table>
<thead>
<tr>
<th>COMPANY</th>
<th>SYMBOL</th>
<th>MAX VOLUME</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAL</td>
<td>206,372,300</td>
<td></td>
</tr>
<tr>
<td>CSX</td>
<td>110,571,300</td>
<td></td>
</tr>
<tr>
<td>UPS</td>
<td>94,752,600</td>
<td></td>
</tr>
<tr>
<td>UAL</td>
<td>70,265,700</td>
<td></td>
</tr>
<tr>
<td>SWFT</td>
<td>39,961,900</td>
<td></td>
</tr>
<tr>
<td>UNP</td>
<td>39,411,400</td>
<td></td>
</tr>
<tr>
<td>LUV</td>
<td>36,283,400</td>
<td></td>
</tr>
<tr>
<td>NSC</td>
<td>22,727,600</td>
<td></td>
</tr>
<tr>
<td>STNG</td>
<td>21,227,000</td>
<td></td>
</tr>
<tr>
<td>ALK</td>
<td>19,654,000</td>
<td></td>
</tr>
</tbody>
</table>

![Figure 4: Top 10 companies by max volume in finance industry](image)

**Table 3: Highest 10 volumes traded of 'WFC'**

<table>
<thead>
<tr>
<th>Month/Year</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUG-09</td>
<td>478,173,600</td>
</tr>
<tr>
<td>AUG-09</td>
<td>376,167,700</td>
</tr>
<tr>
<td>FEB-09</td>
<td>372,159,100</td>
</tr>
<tr>
<td>MAR-09</td>
<td>293,808,700</td>
</tr>
<tr>
<td>NOV-08</td>
<td>291,565,500</td>
</tr>
<tr>
<td>MAR-09</td>
<td>274,337,900</td>
</tr>
</tbody>
</table>

![Figure 3: Top 10 companies by max volume in transportation industry](image)

d. 10 highest volume at WFC
Our second objective is to find 10 highest volume of a company. Here in the example of query we take the company ‘Wells Fargo & Co’ (WFC). As a result of this query we get the 10 highest volumes of ‘WFC’ with their month and year. Below is the query:

SELECT date, volume FROM nyse
WHERE symbol = 'WFC'
ORDER BY volume DESC LIMIT 10;

![Figure 5: Top 10 companies by max volume in energy industry](image)
5. CONCLUSION

From the above analysis we find the companies who have made profits from each industries. Most of the Investors prefer to invest in such company, which is performing well in the equity market. The data should be useful for financial analyst and users, who works in the stock market and analyze all the past records of a company to advice their clients for investments.

Our second result for WFC shows that company moved fast in 2008 and 2009 among 2000 through 2014. Its highest volumes are traded in these 2 years.

In this paper, we showed the possibility that Big Data Hadoop and Hive can be adopted for financial industry. This approach should be very useful in the field of Business Intelligence where the company act upon the past performance and data. As a future work, we will find more financial data set to find out more useful relationship or pattern.

REFERENCES


AUTHOR PROFILES

Jay Mehta is currently pursuing Masters in Information systems at California State University Los Angeles. He completed his Bachelors in Information Technology Engineering from Gujarat Technological University, India in 2012 followed by Masters in Business Administration in Finance from Gujarat Technological University, India in 2014. His interests include data analysis in Big Data using Hadoop Map Reduce and Hive.

Jongwook Woo is currently a Full Professor at Computer Information Systems at California State University Los Angeles. He is a director of the HiPIC (High-Performance Information Computing Center) at the university. He received the BS and the MS degree, both in Electronic Engineering from Yonsei University in 1989 and 1991, respectively. He obtained his second MS degree in Computer Science and received the PhD degree in Computer Engineering, both from University of Southern California in 1998 and 2001, respectively. His research interests are Information Retrieval /Integration /Sharing on Big Data, Map/Reduce, In-Memory Processing, and functional algorithm on Hadoop Parallel/Distributed/Cloud Computing, and n-Tier Architecture application in e-Business, smartphone, social networking and bioinformatics applications. He has published more than 40 peer reviewed conference and journal papers. He also has consulted many entertainment companies in Hollywood.