IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-8, Special Issue-1, Feb 2020 E-ISSN: 2347-2693

A New Technique for New York Stock Exchange (NYSE) Data Analysis Using Apache Pig

Sayantan Halder[1*], Dristi Dugar[2], Ira Nath[3], Pranati Rakshit[4], Dharmpal Singh[5]

[1,2,3,4,5]Dept. of Computer Science & Engineering, JIS College of Engineering, Kalyani, India

**Corresponding Author: samsayan181@gmail.com*

Available online at: www.ijcseonline.org

***Abstract*—** In the given project entitled NYSE Data Analysis using Big Data (Apache Pig), a database of New York Stock Exchange collected from the open source of New York Stock Exchange daily report where we will analyse the data and produce the required output. Here, the data is referred to the daily stocks of all the companies or industries which is enlisted in the NYSE daily report.

We will use the Apache Pig which is used to analyse large data sets representing them as data flows. It is designed to provide an abstraction over MapReduce, reducing the complexities of writing a MapReduce program. Basically, the MapReduce uses java program to execute and to analyse the dataset and it has to use certain logics to provide the desired output. Apache Pig gives us a better platform to do the work more efficiently and quickly using certain logical lines. Apache Pig is one of the best platforms for the analyzation of Big Data.

***Keywords*—** analyse, NYSE, Big Data, MapReduce, Apache Pig

I. INTRODUCTION

In the given project entitled NYSE Data Analysis using Big Data (Apache Pig), a database of NYSE collected from the open-source of NYSE daily report where we will analyse the data and produce the required output. We will use the Apache Pig which is used to analyse large data sets representing them as data flows. It is designed to provide an abstraction over MapReduce, reducing the complexities of writing a MapReduce program. We as Data Analyst divided the data according to our need to solve a few queries. The goal is to process the data for transformation, inspection, cleansing and modelling to discover useful information, conclusion and to support decision-making. After analysing the data, it is a major goal to solve the queries and we use Apache Pig as an interface to solve it.

A. DATA ANALYSIS

The process of evaluating data using analytical and logical reasoning to examine each component of the data provided is the main goal of Data Analysis. Data Analysis is one of the many steps that must be completed while conducting a research experiment. Data from various sources is gathered, reviewed, and then analyzed in some sort to find a particular conclusion. The goal is to process the data for transformation, inspection, cleansing and modelling to discover useful information, conclusion and to support decision-making.

B. BIG DATA

About Big Data:

a) Storing and accessing large amounts of (unstructured) data.
b) Process of high-volume data streams.
c) Makes sense of the data.
d) Predictive technologies.

Characteristics of BIG DATA

1. Volume:

The quantity of data generated and stored. By the size of the data we can determine the value and the potential insight of the data and whether it can be considered as big data or not.

2. Variety:

The data type and nature of data. This is helpful for the people who analyse it to effective result insight. Big data draws from text, images, audio, video and it also completes missing pieces through data fusion.

3. Velocity:

It defines, the speed the speed of the generated data to meet the demand and challenge that lie in the path of growth and development. Big data can also be available in real-time.

4. Veracity:

The data quality of the captured data can vary greatly, also affecting the accurate analysis.

Fig. 1: 8 V's of Big Data which always need to be followed.

C. APACHE PIG

Apache Pig is an abstraction over MapReduce. It is a tool/platform which is used to analyse larger sets of data representing them as data flows. Pig is used with Hadoop, so we can perform, the data manipulation operations in Hadoop with the help of Apache Pig.

To write programs for data analysis, Pig uses a high-level programming language known as Pig Latin. The language will provide various types of operators, which programmers will use to develop their own functions for reading, writing and to process data. Analysation of data using Apache Pig, programmers need to write scripts using the language Pig Latin. All the Pig Latin scripts are converted internally to Map and Reduce tasks. Apache Pig has a component known as Pig Engine that will accept the Pig Latin script as an input and converts the script into MapReduce jobs.

People who are not a good java programmer, normally used to struggle to work with Hadoop, especially while performing any MapReduce tasks. Apache Pig is a substitute for all such programmers. By the use of Pig Latin, programmers will be able to perform MapReduce tasks easily without having to type complex codes in Java. Apache Pig uses a multi-query approach, so it reduces lengthy codes. Let's take an example, an operation that requires 200 lines of code (LoC) in Java can be easily done by just 10 LoC in Apache Pig. So, Apache Pig reduces the development time by almost 16 times.

Pig Latin has a similarity like SQL language and is easy to learn when you are familiar with SQL. Apache Pig provides many in-built operators to support data operations like joins, filters, ordering, etc. In addition to this, it also has nested data types like tuples, bags, and maps that are missing from MapReduce.

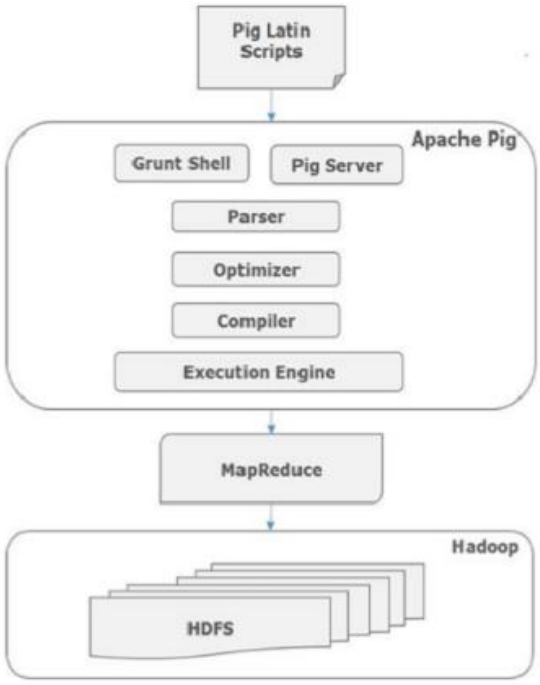

Fig. 2: Architecture of Apache Pig

II. RELATED WORK

During the past few years, several changes are happening in the IT field especially in the area of Cloud computing, Big Data, mobility and Internet of things. It creates a new platform for the enterprises to penetrate into new business. Due to internet and social media penetration vast amount of data produced significantly in the past two years. Every day's data generation exceeds 2.5 quintillion bytes of data. Today's 90% of the data is created within past two years. The growth of data is in enormous speed because of the service and the user, production of vast amount of data is increasing day-by-day. Internet of Things will be an important trigger for database growth for the near future. Estimated number internet connected devices in the year 2020 will be 16 to 50 billion. By the collaboration of Machine and users, the generation and production of data and management in present days would be the Big Data challenge. This is going to be the big challenge for large data streams that we receive from everyday devices and finding useful and meaningful hidden information from the large stream and it is very hard to detect the behavioural patterns out of it. Higher level of decision making and prediction capabilities of the applications and services are very important to get the full benefits from the context aware data intensive applications and services and make the valuable or important information transparent and available at a much higher frequency.

BIGDATA TECHNOLOGY

Big Data is not actually a new concept. Enterprises are having high volume of databases and data warehouses for

many years. The difference in its size, and the complications and increasing growth. It requires new tools to handle the challenges. Traditional RDBMS is not sufficient to handle Big Data. It requires efficient and effective technology to process huge volume of data in an efficient manner. Modern technologies and latest cloud-based applications required to overcome the limitations of traditional RDBMS. Facebook, Twitter, Google, Amazon, Linked In required latest database management technologies to handle dynamic and complicated datasets. NoSQL was initiated by these companies. NoSQL are essential for the Enterprises to handle huge dataset generated through Cloud computing, IoT, Big Data and Big Users. NoSQL does not use SQL as a querying language but it is a type of database management system in a distributed architecture. This is not another RDBMS. NoSQL has following key properties. Higher scalability Ability of partitioning and distribution of data Simplified protocols and interfaces Query capabilities are low Eventual consistency rather ACID property Efficient storage management through distributed indexing Dynamic addition of new attributes to the records. Many Big Data tools which are open source are available in the market.

Big Data Analysis Platforms and Tools.

1) Hadoop And MapReduce: This is one of the popularly used Big Data tool. Hadoop MapReduce is a Big Data programming model used for writing applications to process very huge amount of data in parallel on various clusters of commodity hardware in a reliable and fault tolerant manner. The scheduling, monitoring and re-execution of the failed tasks taken care by the master and the slave execute the tasks as per the direction of the master.

2) Gridgain: This is an alternative of MapReduce and this also supports HDFS. This is used for fast analysis of real time data using in –memory processing.

3) Hpcc: Its expansion is High performance computing cluster. Both paid version and open source is available. 4) Storm: It works in many programming languages and owned by Twitter. It works under Linux operating system. B. Data Bases / Warhouses.

4) Neo4j: It is a graph database model. The speed is thousand times high than the traditional DBMS. It works under REST interface or Java API.

5) Apache CouchDB: It performs MapReduce queries➢ thourgh JavaScript. It provides synchronization even in Smart Objects. ➢

6) Terrastore: This works in all the operating system. It is highly scalable and consistent.

7) FlockDB: It is a graph-oriented database and works in all➢ Operting system

8) RIAK: is another open source distributed key-value data➢ store. It works with map/reduce, HTTP, REST and JSON. 9) Hypertable: This is designed after Bigtable. It runs on the top of HDFS, GlusterFS, or the Kosmos File System (KFS). Its own querying language is HQL (Hypertable querying language)

10) Hive: Like Hypertable it uses its own querying language called HiveQL. This runs in all operating system. Hive is the Hadoop based data warehouse. It provides insights from various data collected from various sources. Talend, Jaspersoft, Jedox, Pentaho, SpagoBI, Knime, BIRT are some of the popular BI tools used for Big Data. Data Mining The primary aim of the data mining it to derive the required information in an understandable format from the available data set. RapidMiner/RapidAnalytics, Mahout, Orange, Weka, jHepWork, KEEL, SPMF, Rattle are popularly available Data Mining tools.

11) Apache Pig: Pig is a high-level scripting language used generating MapReduce programs using Hadoop. Pig Latin is the textual language available in the language layer of Pig.

III. METHODOLOGY

New York Stock Exchange (NYSE) data analysis using Apache Pig

This paperwork will help you in gaining some insights on the NYSE data using Apache Pig. The New York Stock Exchange is formerly known as American stock exchange. It is the world's largest stock exchange by market capitalization of its listed companies at US$21.3 trillion as of June 2017. The daily's average trading value was approximately US$169 billion in 2013. NYSE Daily stock data for each company is available live on yahoo finance for each stock exchange worldwide. We have taken the data of NYSE stock exchange for research and study. This 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 the main 2 objectives: (1) Top 10 companies which have been traded highest by its volume by each Industry. (2) Top 5 highest volume trade of a specific company followed by its date. This dataset contains 2,480 CSV (comma separated values) files containing following fields: • NYSE: Company Symbol, Date, Open of the Day, High of the Day, Low of the Day, Close of the Day, Volume.

PROBLEM STATEMENTS

Now, using Apache pig, we will try to gain more insights from these datasets.

- **Problem Statement 1:**
  - Find the top 100 data from the datasheet.
- **Problem Statement 2:**
  - Generate unique symbols from the data.
- **Problem Statement 3:**
  - Find maximum of high for each symbol
- **Problem Statement 4:**
  - Find minimum of close for each symbol.

Find the top 100 data from the datasheet.

```
grunt> nyse_subset100 = LIMIT nyse 100;
grunt> dump nyse_subset;
```


Output

```
(NYSE,CLI,2009-08-13,31.9,32.33,31.01,31.82,1197000,30.97)
(NYSE,CLI,2009-08-14,31.7,32.13,31.23,31.99,1449100,31.13)
(NYSE,CLI,2009-08-17,31.06,31.27,30.27,30.39,1861400,29.58)
(NYSE,CLI,2009-08-18,30.54,30.87,30.09,30.44,960100,29.63)
(NYSE,CLI,2009-08-19,30.0,30.78,29.65,30.63,940400,29.81)
(NYSE,CLI,2009-08-20,30.76,31.74,30.46,31.66,1236000,30.81)
(NYSE,CLI,2009-08-21,32.02,33.38,32.02,32.49,1140500,31.62)
(NYSE,CLI,2009-08-24,32.58,33.03,32.12,32.27,1129900,31.41)
(NYSE,CLI,2009-08-25,32.58,33.16,32.29,32.68,1044500,31.81)
(NYSE,CLI,2009-08-26,32.62,32.78,32.0,32.37,901300,31.5)
(NYSE,CLI,2009-08-27,32.23,32.79,31.76,32.74,914500,31.86)
(NYSE,CLI,2009-08-28,33.15,33.37,32.3,32.68,1268300,31.81)
(NYSE,CLI,2009-08-31,32.2,32.51,31.73,32.03,1303300,31.17)
(NYSE,CLI,2009-09-01,31.7,32.14,30.25,30.29,1745600,29.48)
(NYSE,CLI,2009-09-02,30.09,30.64,29.82,30.13,1157100,29.32)
(NYSE,CLI,2009-09-03,30.54,31.09,29.85,31.01,955300,30.18)
(NYSE,CLI,2009-09-04,30.93,31.81,30.55,31.71,1250300,30.86)
(NYSE,CLI,2009-09-08,32.23,32.93,32.01,32.92,1345200,32.04)
(NYSE,CLI,2009-09-09,32.99,33.54,32.39,33.48,1341100,32.58)
(NYSE,CLI,2009-09-10,33.34,33.82,32.78,33.76,996800,32.86)
(NYSE,CLI,2009-09-11,33.79,34.08,33.08,33.4,783400,32.51)
(NYSE,CLI,2009-09-14,33.11,35.19,33.05,35.06,982100,34.12)
(NYSE,CLI,2009-09-15,34.92,35.82,34.49,35.12,1404400,34.18)
(NYSE,CLI,2009-09-16,35.47,36.58,35.3,36.4,1222200,35.43)
(NYSE,CLI,2009-09-17,36.37,37.63,35.63,36.26,1131700,35.29)
(NYSE,CLI,2009-09-18,36.52,37.26,35.43,36.44,1238900,35.47)
(NYSE,CLI,2009-09-21,35.68,36.31,35.05,35.12,1580500,34.18)
(NYSE,CLI,2009-09-22,35.54,35.88,35.02,35.72,1820700,34.76)
(NYSE,CLI,2009-09-23,35.84,35.84,33.45,33.57,1879200,32.67)
(NYSE,CLI,2009-09-24,33.89,33.98,31.71,32.09,2144100,31.23)
(NYSE,CLI,2009-09-25,31.9,32.44,31.63,31.83,1376400,30.98)
(NYSE,CLI,2009-09-28,32.18,33.34,31.88,33.25,1143500,32.36)
(NYSE,CLI,2009-09-29,33.42,33.6,32.5,32.88,1031600,32.0)
(NYSE,CLI,2009-09-30,33.04,33.24,32.02,32.33,1783800,31.47)
```

Generate unique symbols from the data

```
grunt> nyse_subset = FOREACH nyse GENERATE symbol;
grunt> unique_sym = DISTINCT nyse_subset;
grunt> dump unique_sym;
```

Find maximum of high for each symbol

grunt> gd = GROUP nyse BY symbol;
grunt> res = DISTINCT(FOREACH gd GENERATE FLATTEN(nyse.symbol), FLATTEN(MAX(nyse.high)));
grunt> dump res;

Find minimum of close for each symbol.

grunt> gd = GROUP nyse BY symbol;
grunt> min1 = DISTINCT(FOREACH gd GENERATE FLATTEN(nyse.symbol), FLATTEN(MIN(nyse.close)));
grunt> dump min1;

Methods of Research Used

For this study, the descriptive method was utilized. We have utilized the method so that the study would be cheap and quick. It could also suggest unanticipated hypotheses. Nevertheless, it could be very hard to rule out any alternative explanations and especially infer causations. Thus, this study used to use a descriptive approach. This kind of descriptive type research utilizes observations to study. We have also taken the help of research papers for the deep study of the technology.

We have implemented this method is to describe the nature of a situation, as it is an existing method, we study and explore the cause of particular phenomena. The researcher has opted this kind of research method, considering the desire of the researcher to obtain first-hand data from the respondents so as to formulate rational and sound conclusions and recommendations for the study.

To come up with pertinent findings and provide credible recommendations, this study utilized two sources of research: primary and secondary. We have obtained primary research data through this new research study. On the other hand, the secondary research data were obtained from the research papers and online youtube videos previous studies on the same.

IV. RESULTS AND DISCUSSION

We have taken the data set of NYSE (New York Stock Exchange) to analyze and provide the output of certain queries. This method is very much efficient to analyze the big data sets. It gives the output in minimum time comparison to other technologies. We know that stock exchange has very large sets of data and to continue the program and to give the right knowledge of stocks one need to analyze that large set of data. It takes a very large time and that old technology is affecting the market. So big data plays an important role in analyze the data.

Even though the Big Data boom started few years ago, the opportunities are growing as the speed of data keeps growing. A global survey conducted by McKinsey on Big Data to understand the innovation, competition, and productivity. The survey covered Healthcare, Public sector, Retail, Manufacturing, Telecommunications. The deep study has been carried out with the help of existing literature reviews and various interview with industry personals. The research conducted in Economics and management. The research focused on Productivity, Competitiveness and growth. The evolution of global financial market and the economic impact of technology. Following key domains will have the great opportunities. A. Marketing Big Data automatically will not lead to better marketing. The deeper and richer insights derived from Big Data drives the success and to read the pulse of the customers. Proper analytics leads to the prediction of tomorrow's requirements by today's purchase. B. In healthcare, a large amount of data carried out for better analysis. Critical insights derived from clinical data will provide good care to the patients. Clinics can play a better role due to the availability of transparent and largely available information. Best practices to be deployed to meet the challenges and complete rethinking and change in IT structure required at the time of deployment. In the genetic field, a big revolution is occurring in the present world. New research direction arrived by Genome project. Big Data plays significant role in data storage, retrieval, sequence analysis

and visualization. C. Social Media Many companies are interested to understand the e-commerce transactions and social media postings to understand the public interest. Get valuable insights from the flooded data is today's challenge. D. Automation Current emerging trend is collection of sensor data in the IoT environment. There is an urgent need for storage, manage and analyse the increasing data which is collected from IoT. E. Manufacturing Industries Manufacturing and IoT are interrelated because many companies are having automated machines in the production environment which generates more data. Big Data tools will be helpful to the manufacturing industries to Store, Retrieve and analyse the data. F. Defence Information is an important treasure in arms race. Data received from satellites, aircraft and messages from various devices are important in the military tech. G. Smart City Smart city is going to change our living environment and infrastructure. This will bring the IoT into reality by embedding advanced technology and data driven methods. Big companies like IBM and Cisco are working seriously to make it real.

V. CONCLUSION AND FUTURE SCOPE

Since our main aim was to know and come across the objectives of Data Analysis of Live Streaming of data using Big Data, we came across various important points related to the given topics.

We learnt how a Data Analyst works to take out the real data required from a huge amount of data. Then we used different Hadoop features mainly APACHE PIG to solve the different queries as required for our project outcome.
Later in future we will be using CLOUD as a platform to fulfil various objectives.
After going through the given project, we came to know about the following:

- Data Analysis
- Big Data
- Hadoop
- Components of Hadoop
- Hadoop Core Components
    i) MapReduce
    ii) Hadoop Distributed File System (HDFS)

SCOPE OF THE PROPOSED SYSTEM
Apache Pig is an abstraction over MapReduce. It is a platform, where larger sets of data are used to analyze and represent them as data flows. Pig is commonly used with Hadoop. We can perform all the data manipulation tasks in Hadoop using Pig.

We can analyze very large data in a minimum time through big data. Big data can be analyzed using certain tools. We are using Piglatin tool to analyze the New York Stock Exchange data within a couple of minutes.

A. BUSINESS CONTEXT

Existing analytical tools takes a long time to execute very large set of data. But using big data tools we can execute it in a couple of minutes, depends on the system configuration.

B. USER OBJECTIVES

Data Set
First access the data set which is to be analyzed.
Query
As per the requirement of the client, the queries are to be given input for the desired output.
Result
As the query is given as input, then the desired output is produced.

REFERENCES

[1] http://www.thesojo.net/key-domains-with-opportunities-in-big-data/
[2] http://www.datamation.com/data-center/50-top-open-source-tools-for-big-data-1.htm
[3] Apache: Couchdb (Online; Oct 2015)
[4] MongoDB: Mongodb (Online; Oct 2015)
[5] Neo Technology, I.: Neo4j, the world's leading graph database.
[6] Pig.apachi.org (online Oct 2015)
[7] https://www.ijraset.com/fileserve.php?FID=3679
[8] https://www.tutorialspoint.com/apache_pig/index.htm

Authors Profile

Sayantan Halder perused Bachelor of Technology in Computer Science & Engineering from JIS College of Engineering, Kalyani in 2019. Currently perusing Masters of Technology in Computer Science & Engineering from JIS College of Engineering, Kalyani. His main research work focuses on Data Mining, Data Analytics, Big Data. He is member of CSI.

Dristi Dugar is currently perusing Bachelor of Technology in Computer Science & Engineering from JIS College of Engineering, Kalyani. She is a member of CSI.

Ira Nath has received her M.Tech degree in Software Engineering from the Maulana Abul Kalam Azad University of Technology, West Bengal, India (MAKAUT) in 2008 formerly West Bengal University of Technology (WBUT), India. She is currently pursuing her Ph.D in Computer Science & Technology at Indian Institute of Engineering Science and Technology (IIEST), Shibpur, India formerly Bengal Engineering and Science University (BESU), Shibpur, India.. She is currently an assistant

professor in the department of Computer Science & Engineering, JIS College of Engineering, Kalyani, Nadia. Her research interests include WDM optical Networks, Mobile Adhoc Network and network security. She is a life time member of CSI.

Dr. Pranati Rakshit is an Assistant Professor in the Department of Computer Science and Engineering of JIS College of Engineering, India. She has more than 19 years of work experience associated with teaching and research. She has completed her Ph.D. degree from Jadavpur University, Kolkata, West Bengal, in the field of Pattern Recognition and Medical image analysis. She has completed her Master degree and B.E. Degree in Computer Science & Engineering. She has worked in the field of Data Mining, image Processing, IOT also. She has supervised more than 20 M.Tech Projects. She has a good number of research publications. She is a life member of Indian Society of Technical Education.

Mr Dharmpal Singh received his Bachelor of Computer Science and Engineering and Master of Computer Science and Engineering from West Bengal University of Technology. He has done his Ph.D in year 2015. He has about 12 years of experience in teaching and research. At present, he is with JIS College of Engineering, Kalyani, and West Bengal, India as an Associate Professor and Head of the department. He has published 32 papers in referred journal and conferences index by Scopus, DBLP and Google Scholar and editorial team and senior member of many reputed journal index by SCI, Scopus, DBLP and Google Scholar. He has organized seven national levels Seminar/Workshop, published two patents and has applied for the AICTE Research Project (MRP) in year of 2019.

