RESEARCHERID
THOMSON REUTERS

*[Taram,*5(5): May 2018] ISSN 2348 – 8034
DOI- 10.5281/zenodo.1255371 Impact Factor- 5.070

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

SMART INVESTMENT DECISIONS THROUGH SHARE MARKET ANALYSIS USING BIG DATA SYNTHESIS ANDHIVE BASED OPTIMIZATION TECHNIQUES IN HADOOP ECOSYSTEM

Manish Taram[*1] &Dr Vikash Kumar Singh [2]
[*1]PhD Scholar, Dept. of Comp. Sci, IGNTU, India
[2]Associate Professor, Dept. of Comp. Sci, IGNTU

ABSTRACT

Big Data is an advance technology which is used to reduce complexity of huge amount of data (data can be in a structured form, unstructured form or semi-structured form). Big data technology has potential to take over all the fields that are capable of producing huge amount of data. Big data is a group of multiple techniques and tools. It has multiple characteristics and techniques. Big data can be understood as huge amount of data that is still growing as a result of continuous data generation by industries, social media, sensory data and even from space related data as satellite images. It can be any combination of audio, video text and images as well. It can be applied in many areas like it is useful in a business, healthcare, law firm, education as we can identify various patterns in data, perform future prediction etc. In this paper our aim is to describe an experiment pertaining to analysis of business file by displaying various traits so as to how to get data records from the data sets and conclude various filtration and data synthesis techniques.

Keywords:*Hadoop, Bigdata, Pig, Hive, Sqoop, Zookeeper*

I. INTRODUCTION

From last 30-35 years, data management tasks are based on physical devices that are used only by the employees. Nowadays the Internet represents a big space where great amounts of information are added every day [1]. Many industries used to waste millions of $(dollars) per year for data management at that time data is limited but now days each & every minute 20-30 million people get connected via communication mediums like (social media, email, video call etc.). Big data analysis has now become an integral part of many computational and statistical departments. Analysis of petabyte scale of data is having an enhanced importance in the present day scenario [2]. So actually we are generating brontobytes of data every year which very hard to store only analytics is far-fetched dream those days. Digital data is rapidly increasing resulting in Big data generation which can consists of text files, image files, audio files and videos files. It can play an important role by analyzing data year by year of every type. Analytics solutions that mine structured and unstructured data are important as they can help organizations gain insights not only from their privately acquired data, but also from large amounts of data publicly available on the Web [3]. In a business field, data has an important role because industrial areas are producing high amount of structured data, unstructured data and semi-structured data. Big data consist of 5 Vs concept- 1st V is called volume (vast amount of data generated every second). 2nd V is called variety (different types of data which contributes to the problems such as text, images, audio, video file). 3rd v is called velocity (speed at which new data is generated and move around). 4th V is called value (that refers to the trustworthiness of the data). 5th V is called veracity (having an access to big data is not good unless we turn in to valuable content.

II. BACKGROUND DETAILS & RELATED WORK

Data-driven decision making, popularized in the 1980s and 1990s, is evolving into a vastly more sophisticated concept known as big data that relies on software approaches generally referred to as analytics [9]. Garter analyst Doug Laney introduced the 3 Vs concept. These 3 Vs are volume, variety and velocity. These three Vs concept were

defining big data but later more Vs get added to meet the demands of time. Earlier, every work or task was done by employees. Data is generated but in a limited amount. Now mobiles, laptops and all electrical things are helping people in getting connected through social networks and all other types of communication networks. In the recent years, datasets generated by machines have been large in terms of volume and have been globally distributed [4]. At basis, the key necessities of big data storage are that it can handle very huge amounts of data and continuous balancing to keep upwith expansion and that it can provide the input/output operations per second (IOPS) necessary to deliver data to analytics tools [7]. So data now generates in huge amount every day. Hence at the present scenario it is more practical to accept big data technology. Companies that figure out how to combine domain expertise with data science will pull away from their rivals [8].

Despite being Herculean in nature, Big Data applications are almost ubiquitous- from marketing to scientific research to customer interests and so on [5]. This enormous amount of data is like raw gold of information and information is power which needs to be disseminate and analyzed, this present scenario of data accumulation and analysis leads to the invention of Apache hadoop ecosystem which is an operating system and is a combination of following technologies-

Grid computing- Grid computing is a type of shared architecture of computer network. Large amount of data can't be processed by one computer so grid computing concept come into place. In a grid computing, multiple computer nodes get connected to a one centralized computer node via grid processing system. In this computing we use resources from multiple computers.

Distributed computing- Distributed computer system is Client-Server based architecture. In a distributed computing, multiple computer nodes are situated in a remote location. All get connected to each other by one dedicated server.

Hadoop Operating System- Hadoop is an operating system which runs upon another operating system that is Linux. It consists of group of tools which are used together by various companies in various domains for various tasks. Hadoop is a type of cluster or open source framework which is written in java language. It is not similar to another framework; it has its own family for processing different types of data. It is a batch processing framework for processing large sets of data. A Hadoop framework provides distributed storage through HDFS (Hadoop distributed file system) and computation (processing) through Map Reduce.

HDFS- (Hadoop Distributed File System)- It is main component of Hadoop. It stores data in a distributed manner. Data get saved into blocks of 64MB or 128 MB. It is highly fault-tolerant and developed on a low cost hardware. HDFS have two core components-java libraries, utilities and YARN which is used for management of various resources.

Map Reduce- Map reduce is a parallel programming model of Hadoop framework. We can write a map reduce program using any language like java python, C++, Perl, Ruby, etc.

SQOOP- Sqoop is a data extraction tool which is used to extract data from various other databases.

HBase- It is non-relational data base that runs on top of HDFS. HBASE is created for large datasets mostly for unstructured data types. It contains millions of rows and columns.

HIVE-Hive is created by Facebook and later acquired by Apache foundation. Hive deals with structured data. Its query language is known as HQL (hive query language). HQL is used in place SQL to perform all the queries in a way same as SQL language.

Pig- Apache pig is a tool, developed by Yahoo. It is a data processing tool that runs over top of the Hadoop. Pig has its own language called pig Latin. It is a data flow language. Pig firstly loads the data and then performs various functions like filtering, grouping, sorting etc.

*[Taram,*5(5): May 2018] ISSN 2348 – 8034
DOI- 10.5281/zenodo.1255371 Impact Factor- 5.070

Mahout- Mahout is an open source machine learning library of Apache written in java.

OOZIE- OOZIE is a work flow scheduler system that manages Hadoop's jobs. It is a server based workflow engine.

Zookeeper-Zookeeper is a coordinator of a Hadoop Ecosystem. Its main purpose is to ensure that each and every tool is able to communicate with each other without any interruption or not. It performs synchronization, grouping, naming and maintenance. In this paper we are proposing an efficient way to synthesize data for an industrial data file which is in .csv format. File taken to perform the experiment is a sample dataset of dividends given by various companies which are registered in Newyork Stock Exchange we will first try to abstract related data required for our experiment then we will load this data to our hadoop ecosystem and perform various filtration operations on it is. Our experiment will show that how we can systematically find a pattern or squeeze the given data to get the desired outcome using various available techniques available in Hadoop ecosystem. Lastly we will try to show an outcome which will display a result that can be achieved by various filtration methods and which gives a vital output which very useful for stock market analysis.

III. EXPERIMENTAL SETUP AND RESULTS

Apache's Pig is an important component of Hadoop system which reduces the coding and analyzing time for Big Data. Big data is collection of complex and large data sets, which include information, may be produced by multiple services [6]. To analysis the we will use a powerful tool that is Apache Pig. It is generally used by data scientist for analysis or performing different operations on the huge amount of data which are as follows-

To process huge data sources such as web logs
- To perform data processing for search platforms.
- To process time sensitive data loads.

To perform the analysis structural data, we are using hadoop Apache Pig software and its native language Pig-latin that comes intrigated with Hadoop's Apache Pig software.

How to Run Pig
To run the Pig environment, we have to enter in grunt interactive shell. This shell can be initiated in two different modes using the following commands -
- Local mode only in one system **(pig –x local)**
- Hadoop, distributed mode access in distributed mode **(pig –x mapred)**

Experiment details
- ✓ First of all, install host operating system Ubuntu and load the all component of Hadoop ecosystem. I have one Ubuntu OS that has all components or necessary files in it.
- ✓ Hadoop is the distributed file system. it can reduce or analysis any data or dataset even they have max size (million's bytes of data) of it. Hadoop can analyze and give us appropriate outcome.
- ✓ I have data file of **New-York** Stock Exchange namely **NYSE_dividends.** It has numbers of dataset like exchange, dates, profits, symbols etc.
- ✓ We have done all the possible operation on that particular dataset and saw a variety of outputs on each operation.
- ✓ In this experiment we can perform the different operation on that particular dataset **NYSE_dividends** and see the appropriate outcomes.
- ✓ All experimental operation can be done only in command prompt of Ubuntu system that is LX-Terminal.
- ✓ First of all, start all the function of the Hadoop system by using – start-all.sh
- ✓ Start the all three machine of Hadoop architecture – jps
- ✓ To open or start the pig grunt shell- pig-x local
- ✓ To quit the grunt shell – quit

*[Taram,*5(5): May 2018] ISSN 2348 – 8034
DOI- 10.5281/zenodo.1255371 Impact Factor- 5.070

- ✓ To exit the grunt shell - exit

Performing various filtration operation in our dataset

We are using a sample dataset of Newyork Stock exchange namely ***NYSE_dividends***which is .csv file (a file with comma separated values) containing data in rows and columns pertaining to dividends given by various companies that are registered with a certain stock exchange.

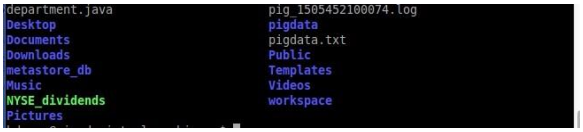

Fig. 1. Locating dividends file in the list of all datasets.

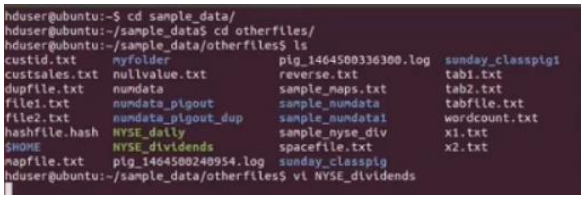

Fig. 2. Displaying the contents of data set

In the above figure 2 we can see that *vi* command has been issued to display the contents of the data file.

Fig. 3. Creation of a relation to separately process the dataset

In the above figure 3 we are using *Load* command to store our file in a separate relation namely *nyse.*

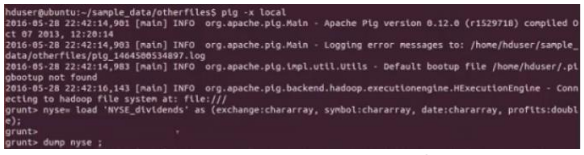

Fig. 4. Displaying the contents of nyse relation

In the above figure 4 we can see that by issuing *dump* we can see the contents of relation.

Fig. 5. Filtration of dataset using merged queries with averaging our dataset

In the figure 5. we can see that an *AVG* command is used to finally find the average dividends.

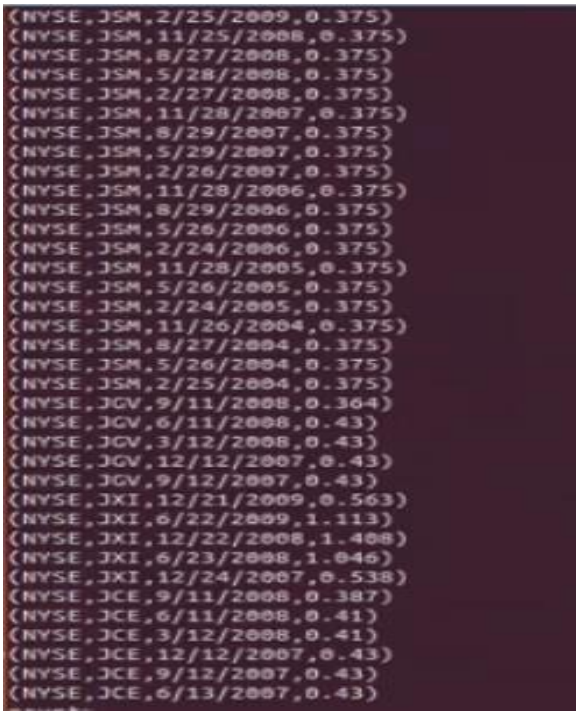

Fig. 6. Displaying output of filter data with constraint as profit column greater than 0.3

Above figure 6 shows the list of companies which had given dividends of greater than 0.3 on their shares on stamped dates. In the above experiment as a result we have a list of companies that provides dividend greater than 0.3percent on their shares. Like this we canxperform many operations on a particular dataset which may include filtration of data, ordering the arranging, Output of lower case first column, Output of Order Command in Decreasing Order, putting constraints like Profit column should be greater than value 0.3 filtrations etc., This knowledge can then be used as input for the selection and formulation of mathematically tractable models of tie-formation [10].

IV. CONCLUSION

From the above experiment we can conclude that all the tasks performed in the hadoop operating system are done using the power of various distributed resources, as we have first selected a dataset that is to be loaded in the hadoop system which is done by issuing simple load command then we have created a relation to store our dataset for further filtration operations with plenty of various available commands in the hadoop ecosystem and finally we get the result by grouping the filtered dataset and constraining the last profit column to show only dividends greater than 0.3 percent. This experiment shows how we can use distributed resources in a clustered hadoop system with great processing power and storage capacity to perform the analysis on any kind of data either it is structured or unstructured which in turn opens a gateway to perform the analysis and storage of datasets from various fields such Healthcare, Education, Space and Business etc

REFERENCES

1. *Elena Geanina ULARU, Florina Camelia PUICAN, Anca APOSTU, Manole VELICANU Perspectives on Big Data and Big Data Analytics, Database Systems Journal vol. III, no. 4/2012.*
2. *Anjali P P and Binu A, A Comparative Survey Based on Processing Network Traffic Data Using Hadoop Pig and Typical Map-reduce, International Journal of Computer Science & Engineering Survey (IJCSES) Vol.5, No.1, February 2014.*

3. *Marcos D. Assunção a,⁎, Rodrigo N. Calheiros b, Silvia Bianchi c, Marco A.S. Netto c,Rajkumar Buyyab, Big Data computing and clouds: Trends and future directions, M.D. Assunção et al. / J. Parallel Distrib. Comput. 79–80 (2015) 3–15, Elsevier.*
4. *Chowdam Sreedhar, Nagulapally Kasiviswanath,Pakanti Chenna Reddy, Clustering large datasets using K-means modified inter and intra clustering (KMI2C)in Hadoop, Journal of Big Data,December 2017, Springer.*
5. *Samiddha Mukherjee, Ravi Shaw, Big Data – Concepts, Applications, Challenges and Future Scope, International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 2, February 2016.*
6. *Ms. Sarika Rathi, A brief Study of Big Data Analytics using Apache Pig and Hadoop Distributed File System, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)Volume 6, Issue 1, January 2017.*
7. *K.R.Kundhavai, S.Sridevi, IoT and Big Data- The Current and Future Technologies: A Review, International Journal of Computer Science and Mobile Computing, Vol.5 Issue.1, January- 2016, pg. 10-14across.*
8. *Andrew McAfee and Erik Brynjolfsson, Big Data: The Management Revolution, Harvard business review, October 2012.*
9. *Anthony G. Picciano, The Evolution Of Big Data And Learning Analytics In American Higher Education, Journal of Asynchronous Learning Networks, Volume 16: Issue 3.*
10. *Chris Snijders, Uwe Matzat, Ulf-Dietrich Reips, "Big Data": Big Gaps of Knowledge in the Field of Internet Science, International Journal of Internet Science ; 7 (2012), 1. - S. 1-5*