International Journal of Advance Research in Science and Engineering
Vol. No.5, Issue No. 06 , June 2016
www.ijarse.com
IJARSE
ISSN 2319 - 8354

HIVE, PIG & HBASE PERFORMANCE EVALUATION FOR DATA PROCESSING APPLICATIONS

Vaishali Chauhan[1], Meenakshi Sharma[2]

[1,2]Student of Master of Technology

Department of Computer Science and Engineering, AP Goyal Shimla University, (India)

ABSTRACT

Information extraction has received significant attention due to the rapid growth of unstructured data. Researcher needs a low-cost, scalable, easy-to-use and fault tolerance platform for large volume data processing eagerly. It is very important to evaluate the MapReduce based frameworks for data processing applications. This paper leverages the comparative study of HBase, Hive and Pig.The processing time of HBase, Hive and Pig is implemented on a data set with simple queries and we will observed the performance of the HBase, Hive, Pig and evaluate the result according to it.

Keywords: Big Data ,Hadoop, HIVE , HBASE, PIG, Performance evaluation

I. INTRODUCTION

Big data describes the propinquity between data size and data processing speed in a system. A comprehensible definition of the concept is "data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time" [1]. Data has been a backbone of any enterprise and will do so moving forward. Storing, extracting and utilizing data has been key to many company's operations. In the past when there were no interconnected systems, data would stay and be consumed at one place. With the onset of Internet technology, ability and requirement to share and transform data has been a need. The explosion of data is being experienced by every sector of the computing industry today. Internet giants such as Google, Amazon, Yahoo!, Facebook and the like have to deal with huge amounts of user generated data in the form of blog posts, photographs, status messages, and audio/video files. In the past, the types of information available were limited. Approach to technology has a well-defined set Information management. But in today's world, our world has been the explosion of data volumes. It is Terabytes and petabytes. The large existing data from different databases relational data stored in Big Data. All of this data would be useless if we couldn't store it, and that's where Moore's Law [2] comes in. The law which states that the number of transistors on integrated circuits doubles every two years, since the early '80s processor speed has increased from 10 MHz to 3.6 GHz - an increase of 360 (not counting increases in "word length" and number of cores)? But we've seen much bigger increases in storage capacity, on every level. RAM as moved from $1,000/MB to roughly $25/GB—a price reduction of about 40,000, and all this together with the reduction in size and increase in speed. The first gigabyte disk drives appeared in 1982, weighing more than 100 kilograms; now terabyte drives are consumer equipment, and a 32

GB micro SD card weighs about half a gram. Whether you look at bits per gram, bits per dollar, or raw capacity, storage availability has grown faster than CPU speed.

In the following table we give the units used to measure the data, starting with those most familiar to those that are essential to measure the Big Data.

Name	Symbol	Value in bytes
Gigabyte	GB	10^9(=1000^3)
Terabyte	TB	10^{12}(=1000^4)
Petabyte	PB	10^{15}(=1000^5)
Exabyte	EB	10^{18}(=1000^6)
Zettabyte	ZB	10^{21}(=1000^7)
Yottabyte	YB	10^{24}(=1000^8)

Table 1: Units of data

The term itself is being more formally defined by IBM as the combination of 3 V's is velocity, variety and volume. These are the generic big data properties. However, the acquired properties depicted after entering the system includes value, veracity, variability and visualization. Thus, the 7 V's correctly describes the big data [3][4].

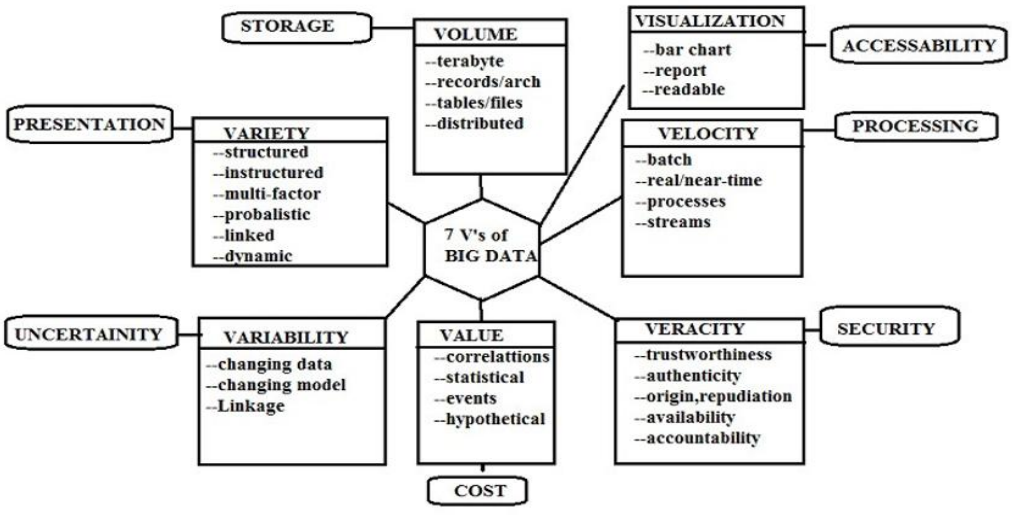

Fig: 1. The 7 V's of big data [3] [4].

- **Volume:** Millions of data is uploaded everyday on Facebook, twitter and other online platforms. The system is generating terabytes, Petabyte and zeta bytes of data. This huge chunk of data is handled through technique of Data Mining expanding its scope to cover big data analytics.
- **Velocity**: The system generates streams of data and multiple sources that require that data. There is an exponential growth in data every hour. Every minute the data is flooded with thousands of online uploads. The widely accepted databases have increased to millions requiring features selection as a vital requirement. Various CI techniques are used for time domain astronomy (TDA) [5].

- **Variety:** Big data is part of structures and unstructured data which include blogs, images, audio, and videos. These data may be analyzed for sentiment and content. Earlier may be the days when companies dealt with only a single data format but today big data provides a platform for all data formats.
- **Variability:** Big data allows handling uncertainty in data with changing data helping in prediction of future behaviour of various customers, entrepreneurs, etc. Basically the meaning of data is constantly changing and the data relies mainly on language processing.
- **Veracity:** In order to ensure the accuracy of big data various security tools are provided for ensuring potential value of the data. This involves automated decision making or feeding data into an unsupervised machine learning algorithm. This ensures the authenticity, availability and accountability of the data.
- **Visualization**: The techniques involved in making the data readable and easily accessible contribute to the 5th V of the Big data. The data needs to be easily understood and the techniques such as the various optimization algorithms provide an advantage of providing an optimal review of the data analyzed.
- **Value:** The value of big data is huge. It enables sentiment analysis, prediction and recommendation. It is massive and rapidly expanding, but it loses its worth when dealt without analysis and visualization that encounters noisy, messy and rapidly changing data.

II. HADOOP

Formal definition of Hadoop by Apache: "The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly available service on top of a cluster of computers, each of which may be prone to failures" [6]. Hadoop was initially inspired by papers published by Google, outlining its approach to handle an avalanche of data, and has since become the standard for storing, processing and analyzing hundreds of terabytes, and even Petabyte of data. Hadoop framework development was started by Doug Cutting and the framework got its name from his son's elephant toy [7]. Hadoop is supposed to have limitless scale up ability and theoretically no data is too big to handle with distributed architecture [8].

The two most important components that are the foundation to Hadoop framework are:

- **Hadoop Distributed File System – HDFS**

  HDFS is a distributed file system designed to run on commodity hardware. HDFS has a master/slave architecture. See Figure 3.1. It's a write-once and read multiple times approach. Hadoop Distributed File System (HDFS) is a file system which is used for storing large datasets in a default block of size 64 MB in distributed manner on Hadoop cluster [9]. An HDFS cluster consists of a single NameNode (latest version 2.3.0 has redundant NameNode to avoid single point of failure), a master server machine that manages the file system and regulates access to the filesystem by the clients. There are multiple data nodes per cluster. The data is split into blocks and stored on these data nodes. NameNode maintains the map of data distribution. Data Nodes are responsible for data read and write operations during execution of data analysis. Hadoop also follows concept of Rack Awareness. What this means is a Hadoop Administrator

user can define which data chunks to save on which racks. This is to prevent loss of all the data if an entire rack fails and also for better network performance by avoiding having to move big chunks of bulky data across the racks. This can be achieved by spreading replicated data blocks on the machines on different racks.

Refer Figure 3 [10], the NameNode and DataNode are commodity servers, typically Linux machines. Hadoop runs different software on these machines to make it a NameNode or a DataNode. HDFS is built using the Java language. Any machine that Java can be run on can be converted to act as the NameNode or the DataNode. A typical cluster has a dedicated machine that runs only the NameNode software. Each of the other machines in the cluster runs one instance of the DataNode software. The NameNode manages all HDFS metadata [11].

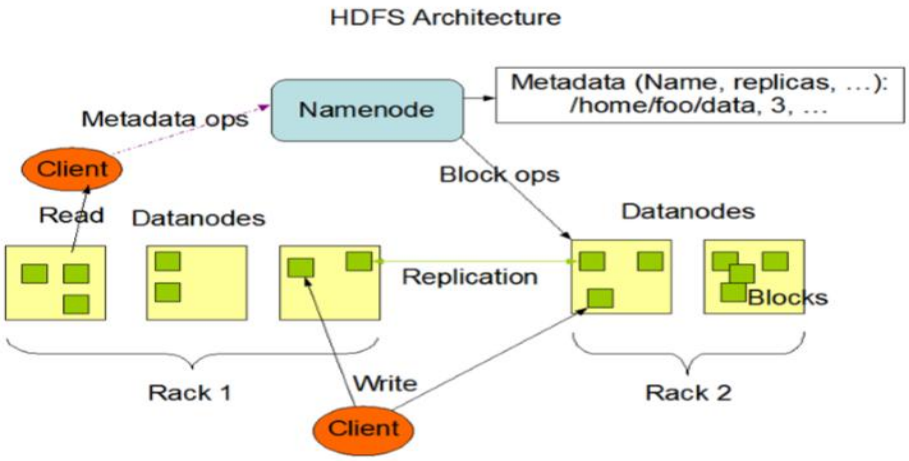

Figure3: Hadoop Architecture [10].

- **Map Reduce Architecture**

It is a programming framework for distributed computing, which was created by Google using divide and conquer method to crack complicated Big Data problems into small units of work and process them in parallel [12]. The basic meaning of Map Reduce is dividing the large task into smaller chunks and then deal with them accordingly, thus, this speed up the computation and increase the performance of the system. Map Reduce can be divided into two steps:

- **Map Stage**

Map is a function that splits up the input text, so map function is written in such a way that multiple map jobs can be executed at once, map is the part of the program that divide up the tasks [13]. This function takes key/value pairs as input and generates an intermediate set of key/value pairs.

- **Reduce Stage**

Reduce is a function that receives the mapped work and produces the final result [14]. The working of Reduce function depends upon merging of all intermediate values associated with the same intermediate key for producing the final result.

International Journal of Advance Research in Science and Engineering
Vol. No.5, Issue No. 06 , June 2016
www.ijarse.com
IJARSE
ISSN 2319 - 8354

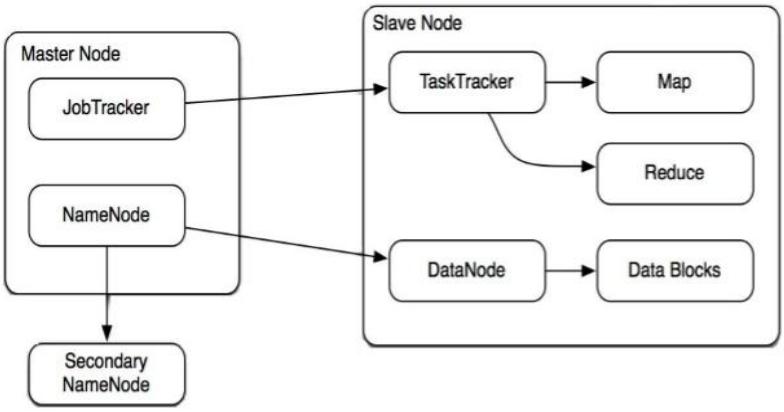

Figure 4: Hadoop Map Reduce Architecture [15].

1. HIVE

Hive is one of the data warehouse infrastructure tools which are used to process structured data in Hadoop. It has been resides on the top of Hadoop to summarize Big Data, making query and analysing them easily [15].Firstly, Hive was developed by Facebook, then Apache Software Foundation developed it further as an open source under the Apache Hive.

Hive is not a Relational Database tool nor it has been designed for Online Transaction Processing (OLTP) nor is it a language for real time queries. Rather it has been designed for OLAP. It stores Schema in a database and process the data into HDFS [16]. Hive is recommended for the developers who are familiar to SQL, Initially Hive was developed by Facebook, later it was taken up by Apache Software Foundation and further as an open source under the name Apache Hive. Hive is designed for OLAP and is fast, scalable and extensible query language [19].

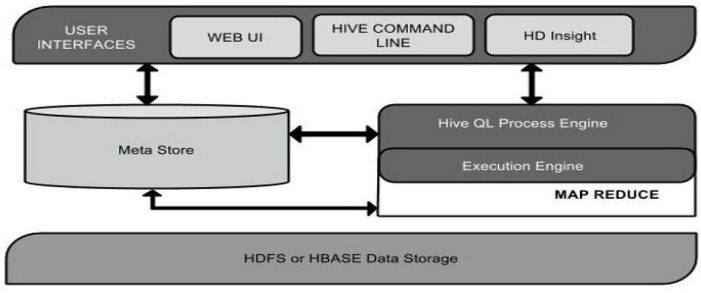

Figure 5: Apache Hive Architecture [18].

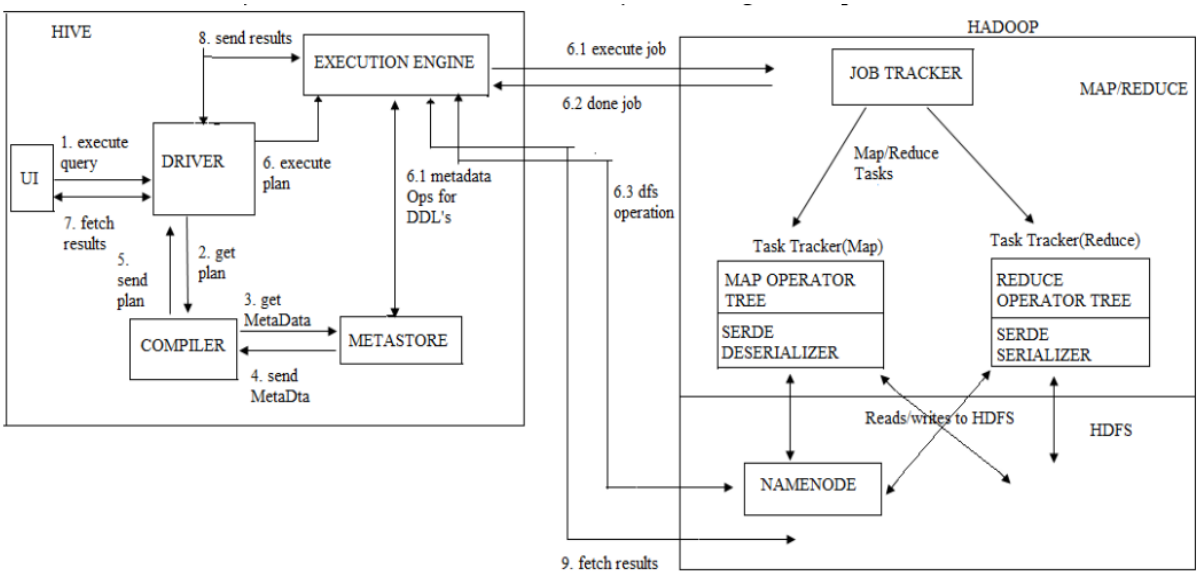

Figure 6: Working Of HIVE [20].

2. Apache HBase:

Hadoop has the limitation that when huge amount of dataset is processed, even for simplest jobs one has to search the entire dataset because data will be accessed only in a sequential manner, HBase are the databases that are used to store huge amount of data and access the data in a random manner [21].

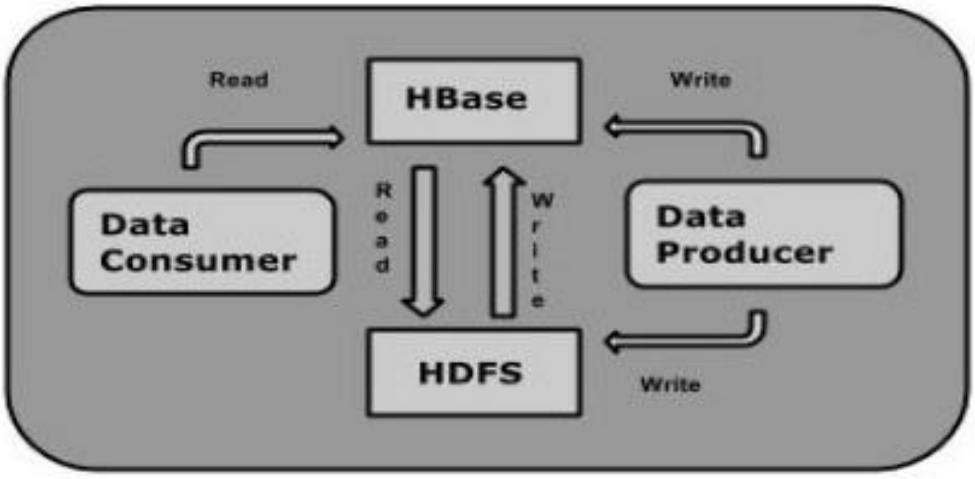

Figure 6: HBase Architecture [21].

3. Apache Pig:

Pig is a scripting language which was initially developed at Yahoo, for creating programs for Hadoop by using procedural language known as Pig Latin, that help in the processing of large data set present in Hadoop cluster. Pig is an alternative to Java for creating MapReduce programs which helps the developers to spend less time in writing mapper &

reducer programs and focuses more on analyzing their data sets. Like actual pigs, who eat almost anything, we can handle any kind of data through the Pig programming language- hence the name Pig!. The rule of thumb is that writing Pig scripts takes 5% of the time compared to writing MapReduce program. The benefit is that you only need to write much fewer lines of code, thus reducing overall development and testing time [18].

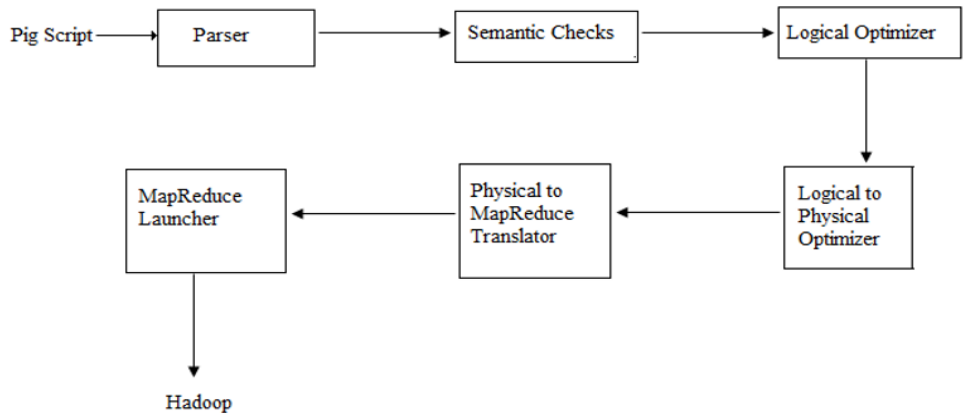

Figure 7: Working Pig [18].

EXPERIMENT SETUP AND SIMULATION RESULTS

We have developed the system and performed the experiment in a virtual machine. Our experiment on a 1 node running Linux Ubuntu 12.04.5 operating system,4GB RAM and 500 GB Hard Drive.

In particular, we installed HADOOP version 2.5.0, Pig version 0.11.1, HBsae version 1.1.1 and Hive version 0.14.0. The CSV format data set is been used for analyzing the version Big Data technologies.

I. Query - Apache HIVE

The query is written to create and load the database, before that we will a database

Hive > use agriculture;

****

create table agr7(commodity_code int, commodity_desc string, country_code string, country_name string, market_year int, calendar_year int, month int, attribute_id int, attribute_desc string, unit_id int, unit_desc string, values int)

row format delimited fields terminated by ',' ;

****

Load data from the path and overwrite it

International Journal of Advance Research in Science and Engineering
Vol. No.5, Issue No. 06 , June 2016
www.ijarse.com
IJARSE
ISSN 2319 - 8354

Hive> load data local inpath ‘agri_alldata1.txt’ overwrite into table agri7;

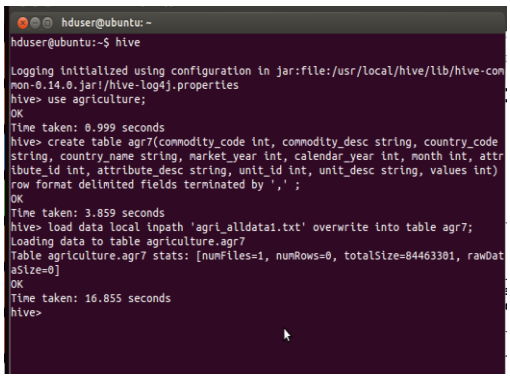

Figure 8: Show loading of data

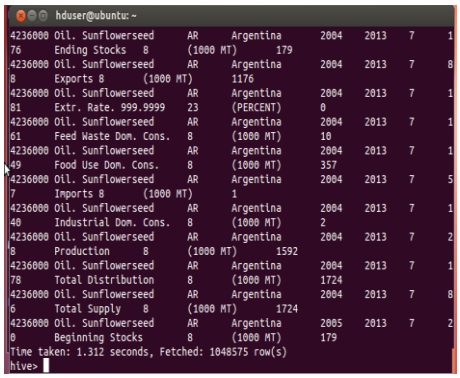

Figure 9: Output of HIVE

II. Query – Apache HBase

The queries written in HBase is the databases that are used to store huge amount of data and access the data in a random manner

****

Hbase> Create ‘agri’,’personal data’,’proffessional data,

****

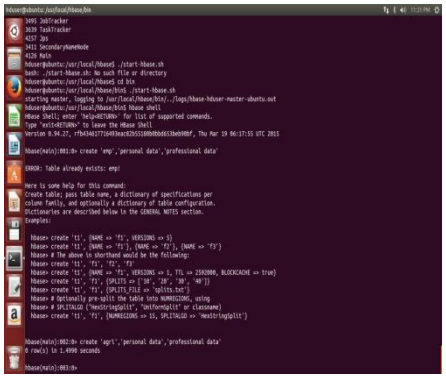

Figure 10: Load the data in HBase

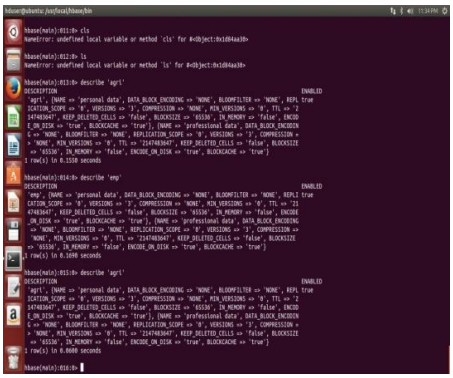

Figure 11: Output of data in HBase

III. Query – Apache Pig

The query written in the pig is in scripted form; we will load the database, then DUMP it and store.

Database been LOAD

****

agri0 = LOAD '/user/hduser/agri1.txt/' USING PigStorage(',') as (commodity_code:int,commodity_desc:chararray,country_code:chararray, country_name:chararray,market_year:int,calendar_year:int,month:int, attribute_id:int,attribute_desc:chararray,unit_id:int,unit_desc:chararray, values:int);

****

Grunt> DUMP agri0;

Grunt> STORE agri0 INTO agriculture;

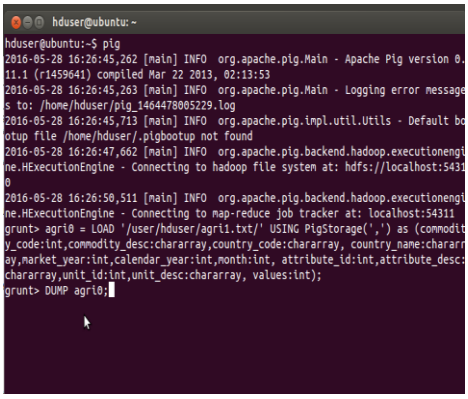

Figure 12: LOAD data

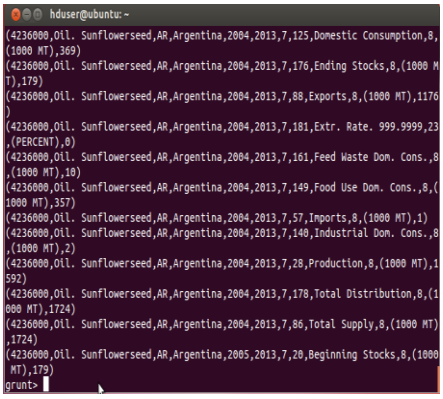

Figure 13: Output in Pig

PERFORMANCE ANALYSIS

We analyzed the various optimization and performance enhancers available in HBase, Hive and Pig. Different analytics led to different results and gave insights about when one should work at what stage of Hadoop framework and which optimization and performance enhancer to leverage at that particular stage.The result is measured on the bases of the execution time been take by the query. As the graph show that development time of Hbase of more as it involves large number of coding, where pig and hive development time is less comparatively to Hbase.

Hive took less time to execute the query in very less time 1.55 sec.

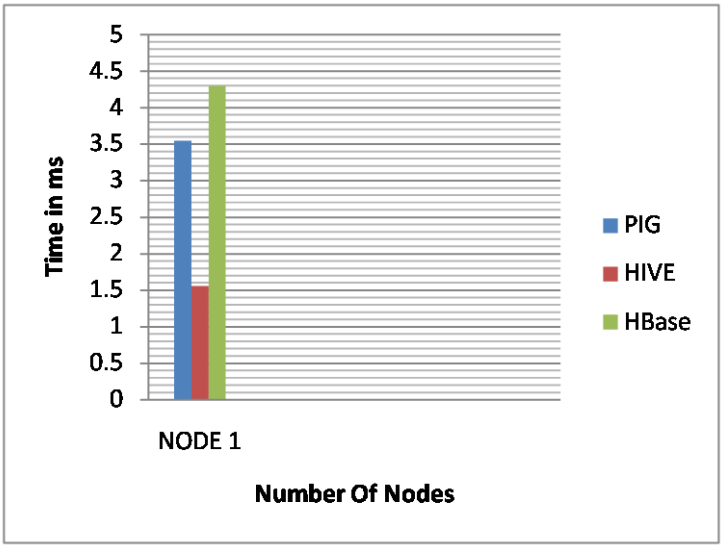

Figure 14: Result on execution time

TABLE 2: Comparative study on BIG DATA Technologies

Features	Apache HBase	Apache Hive	Apache Pig
Available	Open-Source	Open-Source	Open-Source
Developed by	Apache Software Foundation	Facebook	Yahoo
Companies using	EBay, Yahoo, TrendMicro, and Facebook etc.	Facebook, Netflix	Yahoo
Language	Java	HiveQL.	PigLatin
Required Software	JDK version 1.7	Hive version 1.2 above require Java 1.7, Hadoop version 2.x preferred	Java1.6 or above is supported
External file support	Yes	Yes	Yes
When to Use	read/write access to our Big Data	For analytical purposes.	For data processing on Hadoop clusters.

Data Structure it operates on	NOSQL	Apache Derby database	Complex, nested.
Event Driven	No	No	No
Used	To provide quick random access to huge amount of structured data.	Use for effective data aggregation method, ad hoc querying and analysis of huge volumes of data	For processing of large data set present in Hadoop cluster

CONCLUSIONS

There are various ways for Hadoop to run the job. The three programming approaches that are HBase, Hive and Pig are used. But after performing practically we conclude that Hive takes less time as compare to both Pig and HBase. But all approaches have pros and cons so we have to choose according to our need and data**.** For future work, it would be fundamental to compare Hive and Pig on bigger data collections and giving out the best results out of the two.

REFERENCE

[1] Jacobs, The pathologies of big data, Commun. ACM Vol. 52 (8) (2009) pp. 36–44.

[2] R. Schaller, Moore's law: past, present and future in Spectrum, IEEE Vol. 34 (1997): 52-59 (availablehttp://mprc.pku.edu.cn/courses/organization/autumn2013/paper/Moore%27s%20Law/M132oore%27s%20law%20past,%20present%20and%20future.pdfaccessed December 2013).

[3] Rasmus Wegener and Velu Sinha, "The Value of Big data: How analytics differentiates winners", Bain and Company.

[4] Yaochu Jin,, Barbara Hammer, " Computational Intelligence in Big Data", IEEE Computational intelligence magazine, August 2014.

[5] Huijse et al.," Computational intelligence challenges and Applications on Largee Scale Astronomical Time Series Databases", IEEE, 2013.

[6] Apache Hadoop. What Is Apache Hadoop?, 2014. http://hadoop.apache.org/, accessed April 2014.

[7] Wikipedia. Apache Hadoop, 2014. http://en.wikipedia.org/wiki/Apache_Hadoop, accessed April 2014.

[8] T. White. Hadoop – the definitive guide. O'Reilly Media, Inc., Sebastopol, California, 1 edition, 2009.

[9] Qureshi, S. R., & Gupta, A, "Towards efficient Big Data and data analytics: A review", IEEE International Conference on IT in Business, Industry and Government (CSIBIG),March 2014 pp-1-6.

[10] Apache Hadoop. MapReduce Tutorial, 2013. https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html, accessed April 2014.

[11] Apache Hadoop. HDFS Architecture Guide, 2013. http://hadoop.apache.org/docs/r1.2.1/hdfs_design.html, accessed April 2014.

[12] Aravinth, M. S., Shanmugapriyaa, M. S., Sowmya, M. S., & Arun, "An Efficient HADOOP Frameworks SQOOP and Ambari for Big Data Processing," International Journal for Innovative Research in Science and Technology, 2015, pp. 252-255.

[13] http://www.zetta.net/blog/cloud-storage-explained-yahoo

[14] Borthakur, D., Gray, J., Sarma, J. S, et.al, "Apache Hadoop goes realtime at Facebook,"ACM International Conference on Management of data, 2011,pp 1071-1080.

[15] Pandit, A., et al., "Log Mining Based on Hadoop's Map and Reduce Technique.", International Journal on Computer Science & Engineering , 2013pp. 270-274.

[16] http://www.bigdatauniversity.com/web/media/player.php?file=BD030V212EN/Videos/HiveLesson3_HiveDML_Video1_2_1_2.mp4&caption=files.db2university.com/BD030V212EN/Videos/EN/HiveLesson3_Video 1_captions.srt, 25 September,2015

[17] Anshu Choudhary and C.S. Satsangi "Query Execution Performance Analysis of Big Data Using Hiveand Pig of Hadoop",vol. Su-9 no.3,pp.91-93,Sep 2015.

[18] Olston, C., Reed, B., Srivastava, U., Kumar, R., & Tomkins, A, "Pig latin: a not-so-foreign language for data processing," ACM International Conference on Management of data, pp.1099-1110.

[19] Sarkar, D, "Understanding Windows Azure HDInsight Service. In Pro Microsoft HDInsight," Springer, 2013, pp.13-22.

[20] http://www.dezyre.com/Hadoop-Administration/28, 20 September,

[21] Vora ,M. N, "Hadoop-HBase for large-scale data." IEEE International Conference on In Computer Science and Network Technology (ICCSNT), 2011, pp.601-605.