Query Execution Performance Analysis of Hive and Pig in Cloud Architecture

[1] Mahesh Godara, [2]Shrwan Ram

[1]M.E scholar Department of CSE, MBM Engineering College, Jai Narian Vyas University Jodhpur

[2] Asst. Professor Department of CSE, MBM Engineering College, Jai Narian Vyas University Jodhpur

Abstract**—***Cloud platforms require a skilled computing infrastructure. At this level, a large amount of data is generated in fractions of a second, so traditional computing techniques are not enough. Big data provides answers to such huge calculations and supports measurement storage based on application requirements. Big data is the next generation storage infrastructure. This paper examines the big data environments and compares data retrieval techniques. For comparative research, Pig and Hive techniques are chosen. These technologies provide effective data processing capabilities. Hadoop storage is designed for comparative research and then configures pig and hive with the help of the MapReduce framework. In addition, in order to evaluate the efficiency of query execution in terms of processing time, a list of similar questions is prepared and used for processing each query. Both technologies are used for the resulting queries. The hive queries have been found to be processed inless time as compared to Pig in library dataset.*

Index Terms—Big Data; Hadoop; Hive;Pig; Performance Analysis; Data Processing; Query Execution Time.

I. INTRODUCTION

In order to rapidly develop technologies and inventions in the fields of computer science and technology, it is necessary to effectively use the data generated by these technologies. Many organizations use data warehousing services to analyze their data. These organizations use data analytics to make decisions for their development. In this way, we need to process the data accurately to make accurate decisions. But now the amount of data is increasing and is available in PETA bytes, and the centralized server architecture cannot be used to control the amount of data.

With the introduction of the Internet, it has improved small-scale business development. Large Internet giants like Facebook, Google, etc. are using big data to manage their data. Facebook generates 2-3 terabytes of data per day [1]. The large amount of data and usage of distributed computing presents a new set of challenges for managing and performing computational operations such as mining, machine learning, and artificial Intelligence. Managing and using large amounts of data requires a lot of time and cost so, efficiency is critical in data analysis.

The technology is rapidly increasing, which increases the demand of the users and increases the cost of data processing in an organization. A large number of resources (such as computing power and data transmission capabilities) in data processing have also increased, making traditional technologies and tools obsolete. Exploring new equipment and technologies. Big data [4] and big data analysis provide a good solution for large data processing. Hadoop is a widely used solution for processing and storing big data in open source software platform platforms.

In this article, the cloud environment Microsoft Azure HDInsight is used to process data. Pig and Hive, Hadoop [4] devices have been used to store and process data sets. Large data set [6] of 6 GB is used to analyze data. All necessary installations are completed in the initial phase, and then the dataset is stored on the Azure storage engine. All data sets are then loaded into the hive and pig tables. Finally, a series of four different types of problems were performed on the hive and pig. Execution is done on 8 node cluster and the results are analyzed.

II. Background Study

A. Big Data and Analytics

The growth and availability of large amounts of data with all possible variants is often referred to as big data. This is one of the most popular terms in today's automation world, perhaps in the form of the Internet, where big data is becoming a common value for business and society. Big data and its technologies provides a platform for analyzing datasets and gather valuable insights from these datasets. Big data researchers see the big data as follows:

In terms of volume:

This is one of the most important factors, contributed to the emergence of big data. The amount of data has multiplied by various factors. For decades, organizations and governments have been recording transaction data, and social media has always had unorganized data, automation, sensor data, and steam-to-machine data pumping. In the past, data storage itself was a problem, but setting up storage today is not a big challenge due to advanced technology and economical storage devices. But there are still number of other challenges, such as correlations in large data volumes and value use analysis to gather information from the data [3].

In terms of Velocity:

The amount of data is challenging but what is developing a serious challenge is to deal with time and skills. Internet streaming, RFID tags, automation and sensors, robotics and more technologies actually require processing large amounts of data in real time. Therefore, the speed of data growth is a huge data challenge, standing in front of every large organization [3].

In terms of Variety:

Increased the amount of data is a big problem but the diversity of data is a huge challenge. The diversity of data includes structured, unstructured, relational and non-relational data which is growing in a variety of file formats such as video, image, multimedia, financial data, aeronautical data and scientific data. The challenge now is to find a way to correlate all data diversification from time to time to derive value from these data. Today, many organizations are trying to find better solutions to this challenge [3].

In terms of variability:

As the variety increases, the data grows rapidly which can be a problem but Big data ups and downs with the traffic is a huge challenge. It is necessary that analysis of a lot of social media response data for global events has to completed on time before the trend changes. The impact of global events on financial markets, which increases and when dealing with unstructured data it is affects the results of the markets. So it’s necessary to process the real time data [5].

According to complexity:

All of the above factors do produce big data challenges, huge amounts, increased source diversity, and continuous multiplication with unexpected trends. Despite all these facts, the data should be processed to establish a meaningful relationship between the different hierarchies and links before data becomes out of control. This explains the complexity involved in big data [5].

B. Hadoop

Hadoop is an open-source software framework for data processing and execution of tasks or jobs on many group of nodes [17]. Hadoop has the processing power for any type of data and it has the ability to run and manage unlimited concurrent tasks or jobs. It consists of cluster nodes which are made of commodity hardware.

There are two major parts in Apache Hadoop. The first is Hadoop Distributed File System (HDFS) which is designed for the primary storage of distributed file systems. The second part is MapReduce and it is a parallel processing software framework. This is a sub-project for the Apache Hadoop project. MapReduce is not a programming language, it's a programming model. The entire Hadoop Ecosystem runs on the HDFS and the MapReduce algorithm, which is used by the application that runs the Hadoop. In Hadoop data is processed in parallel nodes of different machines or systems.

The Hadoop Cluster is designed to store and analyze large amounts of unorganized data in distributed computing environments. We can build or use a wide range of complex applications at the top of the Hadoop platform. Hadoop is an advanced platform and is scalable enough to handle large and different variations of unstable data. A small Hadoop Cluster consists of a master or head node and multiple client or slave nodes. The JobTracker, the TaskTracker, the name node,data node and head node are the elements of the Hadoop. A worker works side by side with the master node, slave data node and TaskTracker.

Apache Hive

Hive Data Warehouse Infrastructure runs on the top of Hadoop. It provides a language called Hive QL to organize and execute data. Hive QL is similar to SQL using a declarative programming model [9]. Pig separates language from Latin, which uses a more procedural approach. The final result required for Hive QL in SQL is described in a large query. Instead, using pig Latin, a series of query assignments is created step-by-step. Apache hive developers specifically enable SQL developers to write queries in the Hive query language HQL. HQL is similar to standard query language. HQL can break Hive queries to communicate in the map reduce the jobs executed in the Hybrid cluster.

Apache Pig

Pig is a tool which provides a platform to perform large data analysis. The amount of large data substantial parallelization of functions is a very important feature of the pig program, which enables them to handle large scale data sets. While pig and hive are meant to perform similar tasks [6]. Pig is better suited for data preparation phase of data processing, whereas the hive improves the data warehousing and presentation landscape. The idea is that since the data is increasing, it is first cleared using the tools provided by the pig and then stored. From that point the hive is used to run a query that analyzes the data. During this work, the incremental buildup of the data warehouse is not enabled and is done using both data preparation and pig script. The feasibility of using pig and hive in combination is yet to be tested.

III. System Architecture

To perform the query analysis a Big Data environment is developed first using the Hadoop and MapReduce technology. Hadoop provides a platform for storing data and environment for the analysis of data which can be scaled as per the requirement. MapReduce is a programming model which provides the map and reduce function for data analytics. MapReduce is not a programming language it's just a programming model.

The input datasets are hosted on the Azure storage repository and then using MapReduce jobs data is processed in Hive using HiveQL and Pig Script in Pig.Command Line Interface is used to query the data in both Hive and Pig with dataset of 6GB. When the processing of data and the execution of queries are done, the amount time took by the query for the execution is noted as a performance analysis of the query.

The proposed system architecture in cloud using the Microsoft HDInsight for the study done in query execution performance is shown in the Fig. 1. below:

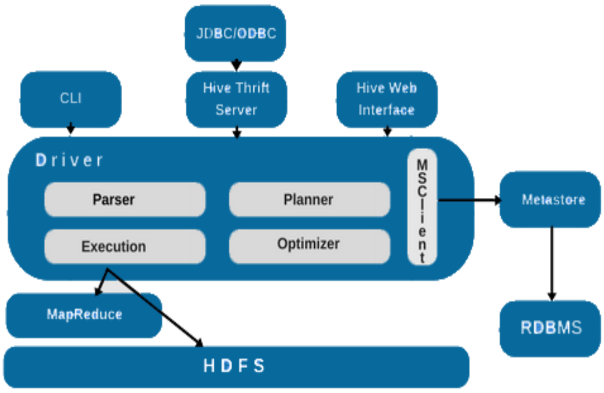

Fig. 2. Hive Architecture.

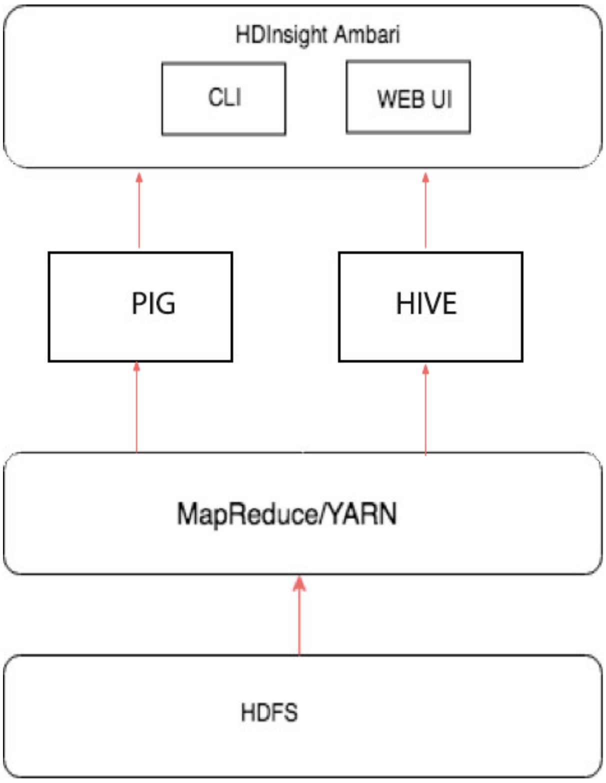

Fig. 1. System Architecture

The architectural diagram of Hive with their components are shown in Fig. 2.

UI: The user interface is made for user interaction with the HDFS. User can submit queries and other operations to the system. The user interfaces that Hive supports are a Web User Interface, command line, and HDInsight.

Meta Store: It stores the schema or metadata of the tables and its partitions in the warehouse. It also stores the column and column data types, the serializer and de-serializer, HDFS Mapping of files where the data is stored.

Driver: This component takes the queries as input for the MapReduce program. It handles session and executes and fetch API's modeled on JDBC/ODBC interfaces.

Compiler: This component parses the Hive queries and semantic analysis on every query block is done. Finally, execution plan with the help of hive table and the metadata from metastore is generated.

Execution Engine: It executes the execution plan created by the compiler. It processes the query and results are generated like MapReduce results.

HDFS or HBASE: Hadoop distributed file system or HBASE are the primary data storage used by Hadoop to store data into the file system.

Figure 3 shows the hierarchical architecture [20] of pig. In this figure, the initial HDFS file system is used to store data and MapReduce is used for further processing. In order to measure the performance of Queries, Pig and MapReduce is attached as an one system. Pig is an application that works on top of MapReduce, Pig is written in Java, and the pig Latin is compiled in MapReduce's work. As shown in the figure, Apache pig structure has various components. The key components are described below.

Parser: Initially the pig script is handled by the parser. It checks the syntax of the script, checks the typing, and performs a variety of other checks. The production of parser will be a Directed Acyclic Graph (DAG), which represents the pig Latin statement and logical operators. In DAG, nodes are the representation of logical operator and edges for the data flow.

Adapter: The resulted logical planning DAG is then passed through the Logical Optimizer, which performs logical optimization tasks such as projection and pushdown.

Compiler: Compiler compiles customized logical plans in a series of jobs.

Execution engine: In the end, the MapReduce jobs are put in a sequence of order to the Hadoop. After all, these MapReduce jobs are executed on Hadoop, generating desired results.

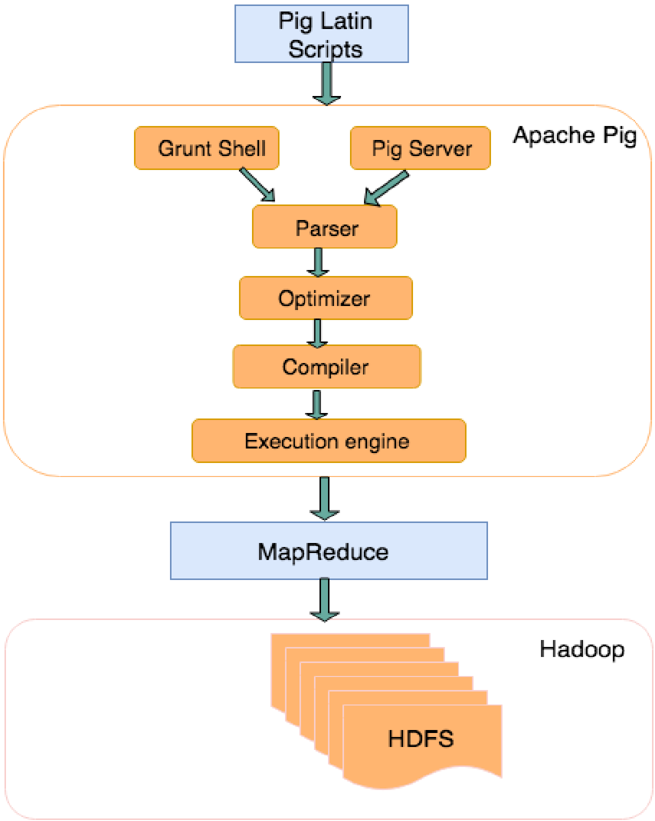

Fig. 3. Pig architecture

IV. Performance Analysis

A. Experimental Setup

Hardware: To measure the performance of the query engine, we used cloud cluster using the Microsoft HDInsight. The cluster consists of 2 Head nodes which were made fixed for

all the setup and 6 worker nodes. Each of the head nodes have 4 Cores (CPU),1 TB disk storage and 27.46GB memory. All the worker nodes have 8 cores (CPU). This cluster has sufficient storage for the real-world and synthetic data, and also has the memory required to allow query engines to benefit from in-memory caching of query inputs or outputs.The Table. 1 shows the Hadoop cluster parameter details.

*Software:*The Hadoop 3.6 version is used for study. Hive uses YARN as resource manager, so we have used Hive 2.0 to carry out our experiments. All query engines under test run on top of a 64-bit Ubuntu 16.0 operating system. Since the queries we run compute results over large amounts of data, the configuration parameters of the distributed file system this data is stored on (HDFS) are crucial. We keep all the parameters fixed throughout the experiment. One of these parameters includes the HDFS block size, which we keep to the default of 64 MB.

Head Node Cores	4
Work Node Cores	8
OS	Linux -ubuntu16 (x86_64)
Hadoop version	3.6
Hive version	2.0
RAM	27 GB
Pig version	0.17.0

Table 1. Hadoop Cluster Configuration

*Datasets:*The dataset used for performing experiment is of size 6GB in CSV file format [7]. It has 91980693 rows.

*Query Types:*There are 4 types of queries are used to carry out the experiment. We have tried to use every type of common queries including the simple SELECT query with WHERE clause and some logical operator like AND, >, < and GROUP BY query and to more complex queries Like INNER JOIN, ORDER BY queries. After the set-up of the Hadoop cluster for the experiment, the queries shown in Table 2 are executed on Hive and PIG with 6GB datasets are loaded into the Hive and Pig tables externally and queries are made to run on the 8 nodes.The execution time as a parameter of performance is used.

S.No	Type of Queries
Q1	Sql SELECT ,WHERE ,AND / > / <
Q2	Sql GROUP BY
Q3	Sql ORDER BY
Q4	Sql INNER JOIN

Table 2. Query Types

B. Experimental Results

The amount of time taken during the query fires and it fetched the record from the datasets is termed here as the query execution time. To measure the query execution time, the queries types listed in the Table 2 are fired on the Hive interface. First with the 6GB dataset queries are fired with number of nodes 8 and all the observation times are observed and then the same process is repeated 4 times and all the execution time are recorded in Table 3. The graphical representation of the experiment is shown in Fig. 4

Query	**Result 1**	**Result 2**	**Result 3**	**Result 4**
Q1	88	84	89	81
Q2	52	52	47	49
Q3	241	239	238	233
Q4	180	1266	1229	1311

Table 3. Execution Time (in seconds) for Hive dataset

After that same type of queries are fired on Pig datasets with 8 nodes and the execution time is recorded and repeated for four times to get accurate results. The execution time for 6GB datasets is Pig shown in table 4 and its graphical visualization is shown in Fig. 5.

After getting the performance using the different repetition of experiments a mean or average performance is calculated. The average result is shown in Table 5.

Query	**Result 1**	**Result 2**	**Result 3**	**Result 4**
Q1	126	125	120	122
Q2	228	265	223	244
Q3	360	377	354	367
Q4	1440	1607	1466	1630

Table 4. Execution (in seconds) for Pig dataset

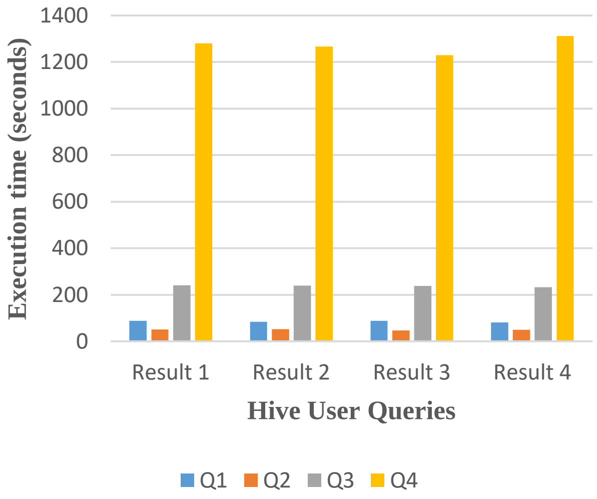

Fig. 4. Hive Graphical representation of the query performance

C. Comparative Performance

The comparative performance of Hive and Pigin terms of query execution time for the given 6 GB dataset is given in Fig. 6. In order to provide the graphical format of both the system, the X axis contains the user queries and the Y axis contains the time consumed in query execution in the given Hadoop configuration. The mean execution time obtained from both Hive and Pig from all the four queries are used to plot the graph. According to the results obtained we can conclude that the performance of the Hive is much more effective as compared to the Pig for the 6 Gb datasets used on 8 node cluster.

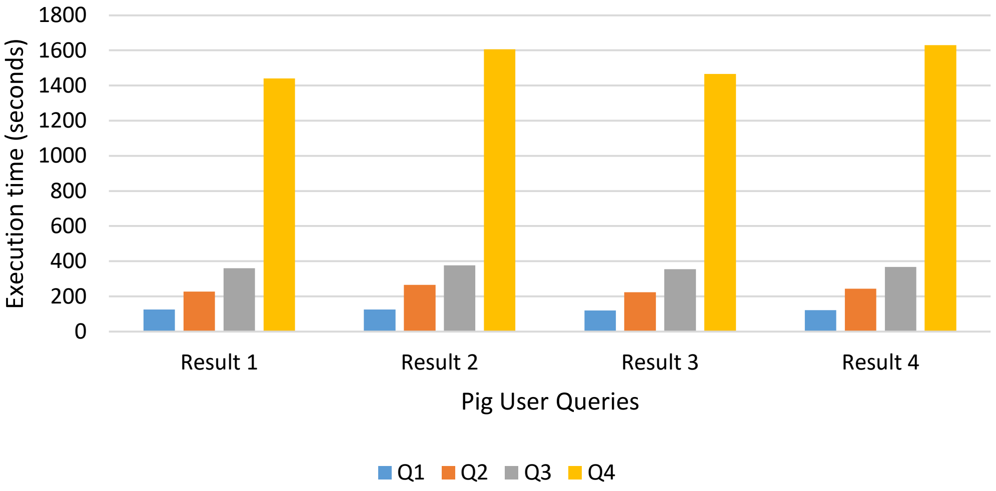

Fig. 5. Pig Graphical representation of the query performance

Query	Hive Mean	Pig Mean
Q1	85.36	123.25
Q2	50.12	240.00
Q3	237.83	364.50
Q4	1271.49	1535.75

Table 5. Average Execution (in seconds) for Pig and Hive

V. CONCLUSION AND FUTURE WORK

The main aim of the proposed study is to have the comparative study of Hive and Pig data processing techniques in the cloud architecture. Therefore, the query execution time is assumed as the key factor for the study. To perform the experimentation on 6 GB dataset which is hosted on Microsoft Azure datastore and using the Hadoop to run the queries performance of both Pig and Hive is evaluated. From the comparative study, the performance of the Hive is found to be more effective and time efficient for data processing as compared to Pig for the same dataset. While working with Hadoop tools Pig and Hive there were couple of areas for improvement identified which can be used for future work. Hadoop was not the easiest tool to work with like other open source tools available in the market. The speed of Hadoop is very less compared to the other alternatives available like Apache Spark is 100 times faster than Hadoop in terms of processing speed.

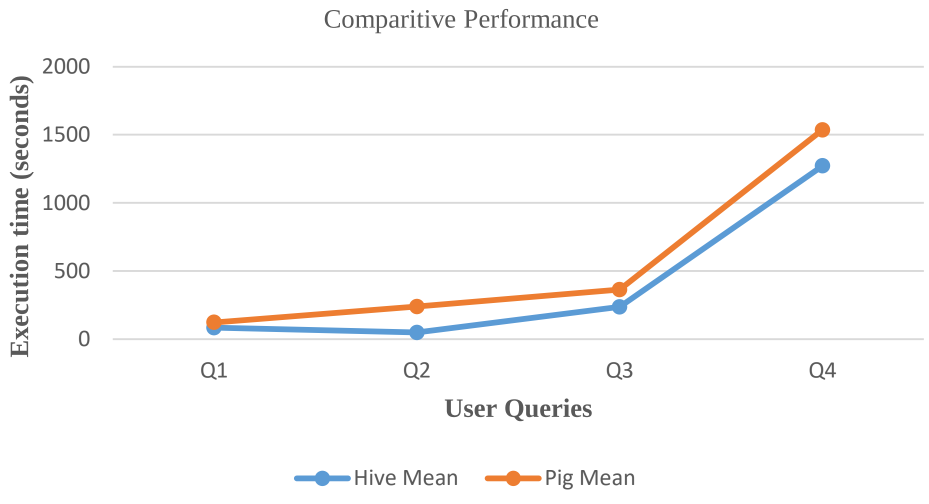

Fig. 6. Pig Comparative performance of Hive and Pig

REFERENCES

[1] Bharath Vissapragada, “Optimizing SQL Query Execution over Map-Reduce,” M.S. thesis, Department of Comp. Sc., Center for Data Engineering IIIT, Hyderabad, India, September 2014.

[2] Ammar Fuad, Alva Erwin, and Heru PurnomoIpung, “Processing Performance on Apache Pig, Apache Hive and MySQL Cluster,” International Conference on Information, Communication Technology and System, IEEE, 2014.

[3] Changqing Ji, Yu Li, Wenming Qiu, Uchechukwu Awada, and Keqiu Li, “Big Data Processing in Cloud Computing Environments,” International Symposium on Pervasive Systems, Algorithms and Networks, IEEE, Dalian, China, 2012.

[4] Apache Hadoop, A vailable: http://wiki.apache.org/hadoop.

[5] Munesh Kataria, Ms.Pooja Mittal, “Big Data and Hadoop with Components like Flume, Pig, Hive and Jaql,” IJCSMC, Vol. 3, July 2014, pp. 759 – 765.

[6] Dataset that is used in this project, A vailable: https://www.kaggle.com/seattle-public-library/seattle-library-checkout-records/data

[7] Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Ning Zhang, Suresh Antony, Hao Liu and Raghotham Murthy, “Hive – A Petabyte Scale Data Warehouse Using Hadoop,” ICDE Conference, IEEE, 2010.

[8] Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Prasad Chakka, Suresh Anthony, Hao Liu, Pete Wyckoff and Raghotham Murthy, “Hive – A Warehousing Solution

[9] Sai Prasad Potharaju, Shanmuk Srinivas, Ravi Kumar Tirandasu, “Case Study of Hive Using Hadoop,” DBLP, V olume-1, Issue-3, 2014.

[10] Gang Zhao, “A Query Processing Framework based on Hadoop,” International Journal of Database Theory and Application, Vol.7, No.4, 2014, pp. 261-272.

[11] Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.;Akram, W. (18-22 Dec. 2012) “Shared disk big data analytics with Apache Hadoop”

[12] Tom White, “Meet Hadoop”, in Hadoop: The Definitive Guide, 4th Edition, O'Reilly, California, 2009, pp. 12-14.

[13] Aditya B. Patel, Manashvi Birla, Ushma Nair (6-8 Dec.2012) “Addressing Big Data Problem Using Hadoop and Map Reduce”

[14] Radhiya A. Arsekar, Ankita V. Chikhale, Vaibhav T. Kamble and Vinayak N. Malavade, “Comparative Study of MapReduce and Pig in Big Data”, International Journal of Current Engineering and Technology, Vol.5, No.2, April 2015

[15] Prof R.Angelin Preethi ,Prof J.Elavarasi ‘big data analytics using hadooptools – apache hive vs apache pig’ International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE) ISSN: 0976-1353 Volume 24 Issue 3