Big Data Emerging Technologies: A CaseStudy with Analyzing Twitter Data using Apache Hive

Aditya Bhardwaj[1], Vanraj[#], Ankit Kumar[3], *Yogendra Narayan, Pawan Kumar[5]
[1, 3, 5] Computer Science & Engineering Department, *Electrical Engineering Department,
#Mechanical Engineering Department,
National Institute of Technical Teachers Training and Research, Chandigarh, India
Email: [1]adityaform@gmail.com, [#]vanraj1010@hotmail.com, [3]ankitmerwal@gmail.com, [*]narayan.yogendra1986@gmail.com

***Abstract*—These are the days of Growth and Innovation for a better future. Now-a-days companies are bound to realize need of Big Data to make decision over complex problem. Big Data is a term that refers to collection of large datasets containing massive amount of data whose size is in the range of Petabytes, Zettabytes, or with high rate of growth, and complexity that make them difficult to process and analyze using conventional database technologies. Big Data is generated from various sources such as social networking sites like Facebook, Twitter etc, and the data that is generated can be in various formats like structured, semi-structured or unstructured format. For extracting valuable information from this huge amount of Data, new tools and techniques is a need of time for the organizations to derive business benefits and to gain competitive advantage over the market. In this paper a comprehensive study of major Big Data emerging technologies by highlighting their important features and how they work, with a comparative study between them is presented. This paper also represents performance analysis of Apache Hive query for executing Twitter tweets in order to calculate Map Reduce CPU time spent and total time taken to finish the job.**

Keywords—Big Data Analytics; Hadoop; Apache Pig; Hive; Sqoop; Hbase; Zookeeper; Flume.

I. INTRODUCTION

Digital universe is flooded with large amount of data generated by number of users worldwide. These data are of diverse in nature, come from various sources and in many forms. Every time we use Internet, send an email, make a phone call, or pay a bill, we create data. All this data needs to be stored in huge data chunks. These data chunks are stored in thousands of disks or hard drives. Around 2.72 zettabytes of data were created until 2012 and it is expected to double every two years reaching about 8 zettabytes at the end of 2015 [2]. Multimedia industries and increase use of social networking sites are the major source of Big Data generation. Every minute Facebook users shares nearly 3.3 million pieces of content, Twitter user sent 347,22 tweets, 100 hours of videos uploaded on YouTube, and 4.1 million search queries are executed on Google every minute. So, in order to get business value from this large amount of data generated Hadoop and its Ecosystems are the popular solution which can help out for better Big Data Analytics solution. This paper is organized as follows. In section II overview of BigData and Hadoop is presented. In section III and IV, Hadoop Ecosystem with analyzing Twitter data by using Apache Hive configured on Microsoft HDInsight Hadoop cluster is discussed. Finally, this paper is concluded in section V.

II. OVERVIEW OF BIG DATA AND HADOOP

In the last two decades, there has been an abundant amount of data generated due to the continuous increase of Internet usage. In today's competitive environment, companies are eager to know about, whether customers are purchasing their products, are they finding their products and applications interesting and easy to use, or in the field of banking they need to find out how customers are doing their transactions. In a similar way, they also need to know about how to improve advertising strategies to attract the customers, and how to predict the behavior of customers to formulate a policy or make a decision for maximizing the profits. To answer all these entire sets of questions, Big Data Analytics is a solution. So, Big Data Analytics is the process of examining large amount of data in an effort to uncover hidden patterns or unknown correlations [3].

1. Hadoop For Big Data Processing

The main challenge in front of IT world is to store and analyze huge quantities of data. Every single day data is generated in huge amount from various fields like Geography, Engineering, and Economics & Science etc. To analyze such huge amounts of data for better understanding of users there is a need to develop data intensive applications which are highly available, highly scalable and based on the reliable storage system. To cope up with these requirements in 2003, Google developed Distributed File System (DFS) called Google File System GFS [6] and introduced Map-Reduce [7] programming model to achieve high performance by moving tasks to the nodes where the data is stored and by executing them in parallel. GFS was a great discovery in order to handle massive data for storing, retrieving, processing and analyzing. But, the major issue with GFS was that this file system was proprietary [10], so the researcher team of Yahoo developed an open source implementation of GFS and Map-Reduce and later this open-source project was named as Apache Hadoop [14] [15]. Hadoop was created by Doug Cutting, an employee at Yahoo for the Nutch search engine project [17]. By seeing his son's toy elephant Doug named it as Hadoop with yellow elephant like symbol. Hadoop architecture mainly comprise of two main components: HDFS for storing Big Data and MapReduce for Big Data analytics.

1.1 HDFS Architecture:

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 [5].

978-1-4673-8253-3/15/$31.00 ©2015 IEEE

Hadoop cluster means running a set of daemons on different servers of the network.

The following table highlights various Hadoop daemons as described below:

Table I. Function of Hadoop Daemons

Hadoop	Description
Name Node	NameNode is the master node which controls DataNodes and it stores metadata for directories.
Data Node	Data Nodes are the slave nodes present in the Hadoop cluster on which TaskTracker node runs. It is responsible for serving read and write request for the client.
Secondary NameNode	Secondary NameNode is the backup of NameNode.
JobTracker	For each MapReduce there is one JobTracker node which is responsible for distributing of mapper and reducer functions to available TaksTrackers and monitoring the results.
Task Tracker	Actual job is run by the TaskTracker and then result is communicated back to the JobTracker.

1.2 MapReduce 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 [28]. 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:

a) 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 [30]. This function takes key/value pairs as input and generates an intermediate set of key/value pairs.

b) Reduce Stage

Reduce is a function that receives the mapped work and produces the final result [34]. The working of Reduce function depends upon merging of all intermediate values associated with the same intermediate key for producing the final result.

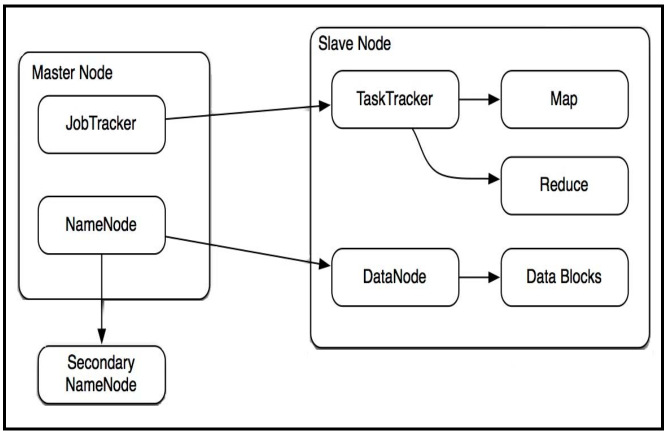


Fig. 1: Hadoop MapReduce Architecture [6].

III. HADOOP ECOSYSTEM

For increasing the efficiency and performance of Hadoop there are many technologies which are built on the top of the Hadoop. Hadoop along with these sets of technologies is known as Hadoop Eco System. Basic Hadoop Eco system consists of following emerging technologies [9].

***a) 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 focus 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 scirpts 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[35].

***b) Apache Hive**:* Pig is scripting language like PigLatin for Hadoop, and Hive is similar to standard with SQL like queries for Hadoop that allows the developers for writing Hive Query Language(HQL). Hive is recommended for the developer 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 [31].

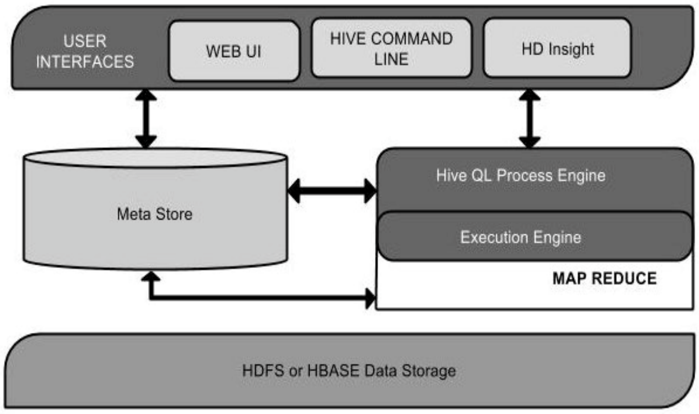


Fig. 2: Apache Hive Architecture [35].

The following components explain the working of Hive:

User Interface: Hive supports three user interfaces such as Web UI, Command line, and Hive HD Insight.

Meta Store: For storing the schema or metadata of tables, column of table, their data types, and HDFS mapping database server is used in Hive.

HiveQL Process Engine: HiveQL is similar to SQL which is a replacement of traditional approach for MapReduce program.

Execution Engine: Execution engine is used to process the query and generates results which are similar to MapReduce results.

HBASE or HDFS: HBASE or Hadoop Distributed File System are data storage techniques for storing data into the file system.

978-1-4673-8253-3/15/$31.00 ©2015 IEEE

c) Apache Sqoop:- Sqoop is a tool that is used to transfer data between Hadoop and Relational Database Servers like Oracle, MySQL. In short, sqoop is used to import data from relational databases to Hadoop HDFS, and export data from Hadoop File System to relational databases [29].

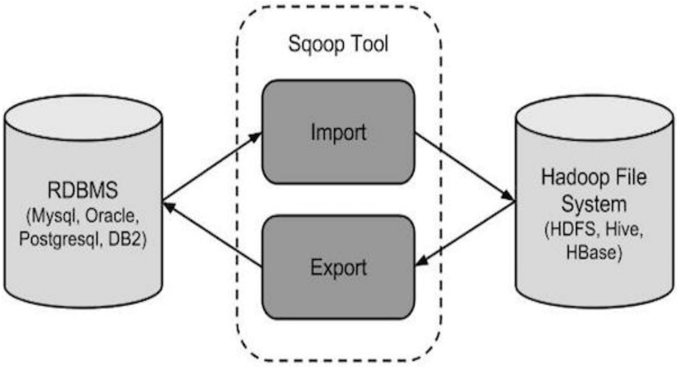


Fig. 3: Sqoop Architecture[28]

d)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 [32].

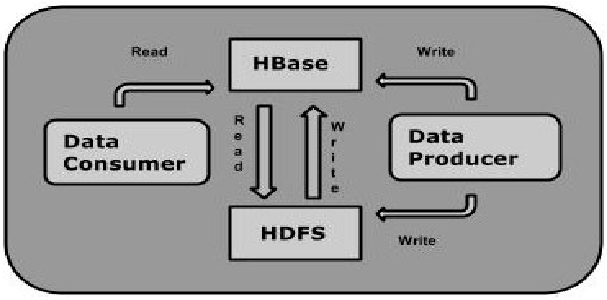


Fig. 4: HBase Architecture[32]

e) Apache Zookeeper: Zookeeper is an open source project by Apache that provides centralized infrastructure which helps in synchronization across the Hadoop cluster [39]. Zookeeper has hierarchical name space architecture, in which each node in the namespace is called Znode. Znodes are similar to the files in a traditional UNIX like system, Znode can be updated by any node in the cluster, and any node in the cluster informed its changes to Znode [42].

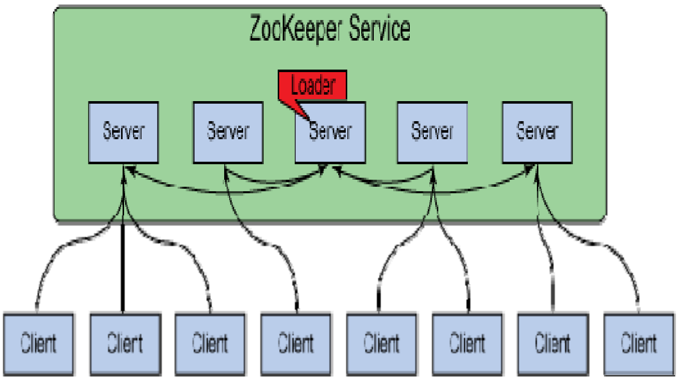


Fig. 5: Apache Zookeeper Architecture [41].

f) Apache Flume: Flume is application built on top of Hadoop which is used for moving of huge amounts of streaming datasets into Hadoop Distributed File System (HDFS) [43]. Sources of stream data are sensor, machine data, logs and social media. Various components of Apache Flume are as follows: ***Source***: Entity through which data enters into Flume

- **Sink:** For delivering the data to the destination this entity is used.
- **Channel:** It is the medium between source and sink
- **Agent**: It is the physical Java virtual machine running flume.

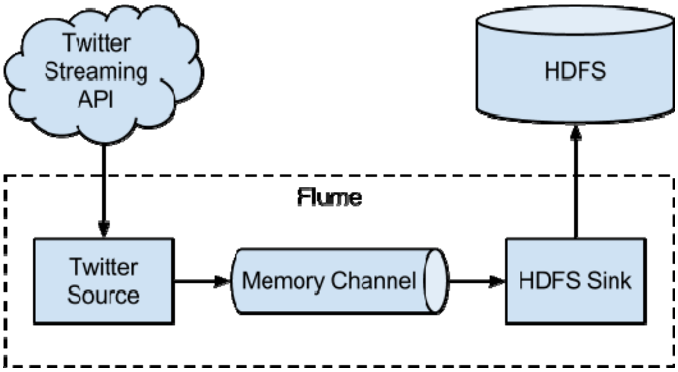
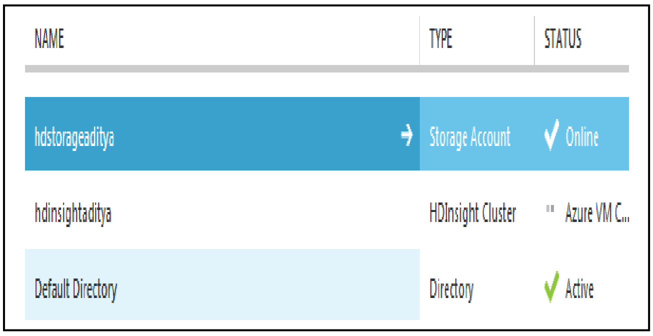


Fig. 6: Apache Flume Architecture[45].

IV. EXPERIMENTAL SETUP

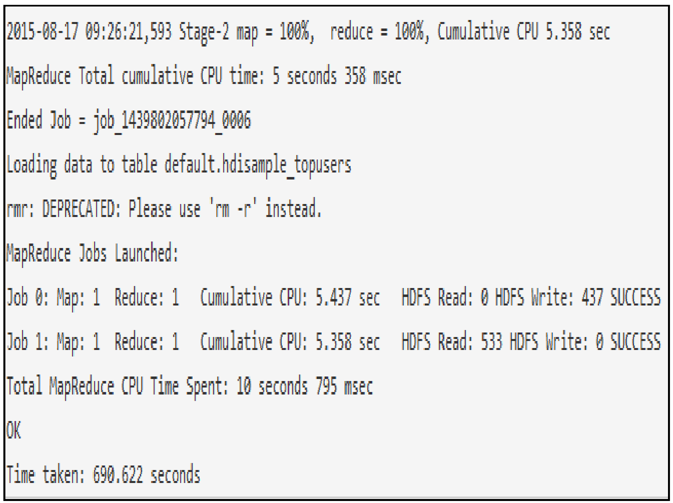
Social websites like Twitter are a useful source of Big-Data for analyzing and understanding users trends. The main objective of this work is to fetch and analyze Twitter tweets which are stored as JavaScript Object Notation (JSON) format on cloud based Apache Hive solution. For the implementing of this work we have used Microsoft Azure cloud services. Two Infrastructure- as a Service (IaaS) services were used: one is HDInsight Hadoop solution for creating Apache Hive setup and second one is Microsoft Azure Blob Storage for data storage. Blob Storage is a general-purpose Hadoop compatible Azure storage solution that enable the HDInsight computation to store and delete data safely without losing user data. First, Twitter live data were fetched by using Twitter Streaming API and then fetched raw data was stored into Blob Storage after that it is transferred into Hive Table. Now HDInsight cluster of various nodes size was created on which Apache Hive queries were executed to analyze the Twitter tweets. Fig.7, shows the snapshot for creating cluster named ‘hdinsight adtiya’ on Microsoft Azure cloud data center location at EastUS and Blob Storage name ‘hdstorage’ for storing the input and output data. Cluster was created by using Window7 as hostmachine and Window Server 2012 as guest machine. Now, after successful creation of Azure cluster, execute Hive query and calculate the required results.



NAME	TYPE	STATUS
hdstorageaditya	Storage Account	✓ Online
hdinsightaditya	HDInsight Cluster	Azure VM C...
Default Directory	Directory	✓ Active

Fig. 7: Snapshot Microsoft Azure Cloud Service.

Hive query was executed on the data stored in Hive table and the results of a number of tweets count were calculated. The result of HDInsight cluster for running Hive query were analyzed based on two parameters: first one is Total Map Reduce CPU Time Spent for running Hive query and second is Total Time taken for running this job. HDInsight Hadoop cluster of size 1 node, 2 nodes, 4 nodes, and 6 nodes are used. 1 node cluster size means MasterNode and DataNode are on the same machine. 2 nodes cluster size means 1 MasterNode, and 1 DataNode are running on two machines. 4 nodes cluster size means 1 MasterNode, 1 SecondaryNode and 2 DataNodes are running on separate machines. 6 nodes cluster size means 1 MasterNode, 1 SecondaryNode, and 4 DataNodes are running on separate machines Fig.8 below represents the snapshot for Hive query execution on 2 node Hadoop cluster size.



```
2015-08-17 09:26:21,593 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 5.358 sec
MapReduce Total cumulative CPU time: 5 seconds 358 msec
Ended Job = job_1439802057794_0006
Loading data to table default.hdisample_topusers
rmr: DEPRECATED: Please use 'rm -r' instead.
MapReduce Jobs Launched:
Job 0: Map: 1  Reduce: 1   Cumulative CPU: 5.437 sec   HDFS Read: 0 HDFS Write: 437 SUCCESS
Job 1: Map: 1  Reduce: 1   Cumulative CPU: 5.358 sec   HDFS Read: 533 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 10 seconds 795 msec
OK
Time taken: 690.622 seconds
```

Fig. 8: SnapShot for Executing Hive Query

Table II. shows the results for mapreduce CPU time spent on Hadoop cluster measured in seconds when the hive query was executed for feteching and predicting tweets count.From the table II, Fig.9 is drawn which illustrates MapReduce CPU time spent for executing hive query on HDInsight Hadoop cluster. In Fig.9, X-axis represents the increase in number of nodes in HDInsight Hadoop cluster and Y-axis represents total time taken for executing Hive query. From the Fig.9, it is observed that as the number of nodes in HDInsight cluster increase the mapreduce slot time for executing hive query increase because more number of nodes in cluster means more switching of mapper and reducer function on the cluster nodes.

TABLE II. MapReduce CPU Time Spent for Hive Query.

No. of Nodes in Hadoop Cluster	MapReduce CPU Time Spent (sec)
1	7.263
2	10.795
4	13.892
6	16.382

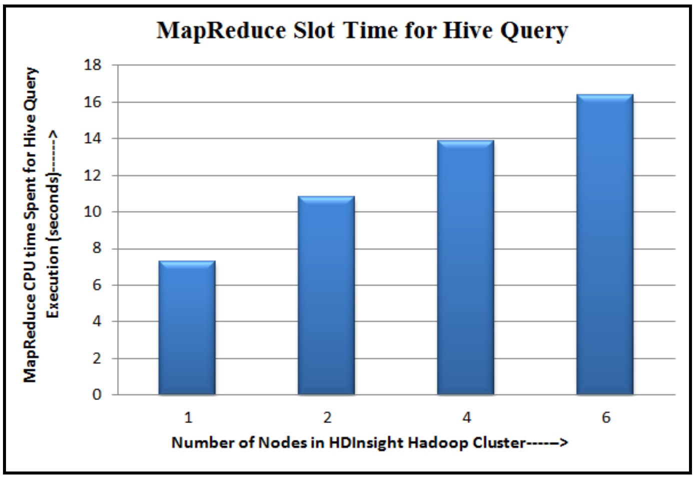


Fig. 9: MapReduce Slot Time for Hive Query Execution.

Table III. shows the results for total time taken for executing Hive query. Total time taken means time including Map-Reduce phase, time to fetch Twitter tweets and time to execute Hive query. From the Table III, Fig.10 is drawn which illustrate total time taken to execute hive query on HDInsight Hadoop cluster for twitter data analysis. In Fig.10, X-axis represents the increase in number of nodes in HDInsight Hadoop cluster and Y-axis representst total time taken for executing hive query on Hadoop cluster of varying size. From the Fig.10, it is observed that as the number of nodes in HDInsight cluster increase total time taken to execute Hive query decrease because if we increase number of nodes in HDInsight cluster then processing of Hive query can take place parallely and which will decrease the query execution time.

TABLE III. Observation of Total Time Taken to Execute Hive Query

No. of Nodes in Hadoop Cluster	Total Time Taken (sec)
1	981.461
2	690.622
4	240.35
6	78.718

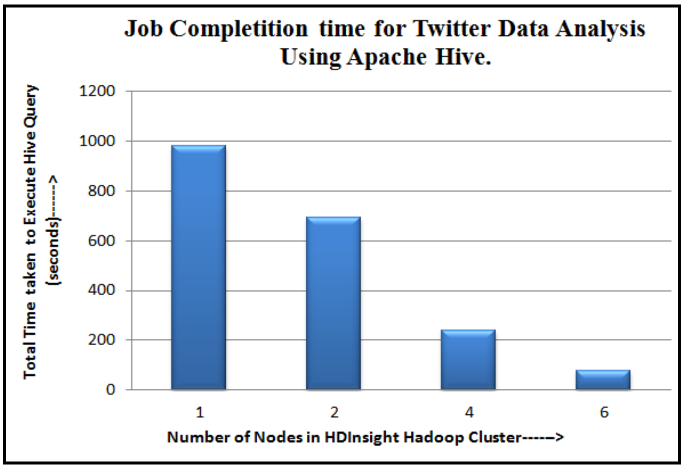


Fig. 10: Total Time Taken to Execute Hive Query for Twitter Analysis

978-1-4673-8253-3/15/$31.00 ©2015 IEEE

TABLE IV. Comparative Study of BigData Emerging Technologies

Features	Apache Pig	Apache Hive	Apache Sqoop	Apache HBase	Apache Zookeeper	Apache Flume
Developed By	Yahoo	Facebook	Cloudera	Apache Software Foundation	Yahoo	Cloudera
Available	Open-Source	Open-Source	Open-Source	Open-Source	Open-Source	Open-Source
Language supported	PigLatin	SQL-like language called HiveQL, or HQL.	MYSQL, Microsoft SQL server, PostgreSQL, IBM DB2.	Java	Java and C	Java
When to Use	For data processing on Hadoop clusters.	For analytical purposes.	When there is need to import and export data from RDBMS to Hadoop	When we need random, read/write access to our Big Data	For Distributed Applications.	For moving large amount of data to a centralized data store.
Data Structure it operates on	Complex, nested	Apache Derby database	Simple	NOSQL	Kafka data structures	Simple
Schema	Optional	Required	Optional	Required	Required	Required
External file support	Yes	Yes	Yes	No	Yes	Yes
Required Software	Java1.6 or above is supported.	Hive version 1.2 above require Java 1.7, Hadoop version 2.x preferred.	No such requirements.	JDK version 1.7 Recommended.	JDK 6 or greater, 2 GB of RAM, Three ZooKeeper servers is the minimum recommended size	Java 1.7 Recommended, sufficient memory and disk space for source, channel, and sink.
Event Driven	No	No	No	No	No	Yes
Companies using	Yahoo	Facebook, Netflix	Yahoo, Amazon	EBay, Yahoo, TrendMicro, and Facebook etc.	Rackspace, Yahoo etc.	Yahoo, Google etc.
Used for	For processing of large data set present in Hadoop cluster	Use for effective data aggregation method, adhoc querying and analysis of huge volumes of data.	To transfer data between Hadoop and Relational databases.	To provide quick random access to huge amount of structured data.	To Provide centralized control for synchronization across the Hadoop cluster.	For moving streaming web log data into HBase.

V. CONCLUSION

Big Data analysis is the latest area of interest for the research communities around the globe. Big Data refers to the volume of data beyond the traditional database technology capacity to store, access, manage and compute efficiently. By analyzing this large amount of data companies can predict the customer behavior, improved marketing strategy, and get competitive advantages in the market. Hadoop is a flexible and open source implementation for analyzing large datasets using MapReduce. There are various emerging technologies such as Apache Pig, Hive, Sqoop, HBase, Zookeeper, and Flume that can be used to improve the performance of basic Hadoop MapReduce framework. Apache Pig which is a scripting language that can be used to reduce development time of MapReduce program because it requires less number of lines of code and provides nested data types that are missing from MapReduce. Hive provides easy to use platform for the developers who are comfortable in SQL language for Map Reduce programming. HDFS has the inability of random read/write to BigData that can be provided by HBase. If we want to transfer data between Hadoop and RDBS system Sqoop can be used. Zookeeper can be used for synchronization of Hadoop cluster and finally Flume can be used for moving streaming web log data to HDFS. This paper also discussed fetching and executing Twitter tweets by using Hive query on HDInsight cluster and results shows that as we increase number of nodes in the cluster, then MapReduce slot time increase but overall total time taken for executing Hive query decease. Future work will consist of implementation of these Big Data emerging technologies to improve the performance of basic Hadoop MapReduce framework.

REFERENCES

[1] Chen, M., Mao, S., & Liu, Y, "Big data: A survey", Mobile Networks and Applications Springer, volume 19, issue 2, April 2014, pp. 171-209.

[2] Sagiroglu, S., & Sinanc, D, "Big data: A review", IEEE International Conference on Collaboration Technologies and Systems (CTS), 2013, pp 42-47.

[3] Pal, A., & Agrawal, S "An experimental approach towards big data for analyzing memory utilization on a Hadoop cluster using HDFS and MapReduce", IEEE, First International Conference on Networks & Soft Computing (ICNSC), August 2014, pp.442-447.

[4] Zhang, J., & Huang, M. L., "5Ws model for bigdata analysis and visualization," IEEE 16th International Conference on Computational Science and Engineering, 2013, pp.1021-1028.

[5] 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.

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

[7] Ibrahim, S., Jin, H., Lu, L., Qi, L., Wu, S., & Shi, X "Evaluating mapreduce on virtual machines: The Hadoop case" Springer , 2009, pp. 519-528.

[8] Patnaik, L. M, "Big Data Analytics: An Approach using Hadoop Distributed File System.", International Journal of Engineering and Innovative Technology (IJEIT), vol 3, May 2014, pp. 239-243.

[9] Bedi,P.,Jindal,V., & Gautam, A,"Beginning with Big Data Simplified", IEEE International Conference on Data Mining and Intelligent Computing (ICDMIC), 2014, pp. 1-7.

[10] Li, J., Wang, Q., Jayasinghe, D., Park, J., Zhu, T., & Pu, C., "Performance overhead among three hypervisors: An experimental study using Hadoop benchmarks", IEEE International conference on Big Data, 2013, pp. 9-16.

[11] Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. "The rise of big data" on cloud computing: Review and open research issues" ELSEVIER, 2015.

[12] Assuncao, M. D., Calheiros, et al. " Big Data computing and clouds: Trends and future directions " Journal on Information System, ELSEVIER, 2014, pp. 98-115.

[13] Chandarana, P., & Vijayalakshmi, M, "Big Data analytics frameworks" IEEE International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA), 2014, pp.430-434.

[14] Singh, J., & Singla, V, "Big Data: Tools and Technologies in Big Data" International Journal of Computer Applications, 2015.

[15] Lakavath, S. "A Big Data Hadoop Architecture for Online Analysis", 2014.

[16] Stephen K, Frank A, J.. Alberto E, William M, "Big data:Issues and challenges moving forward," IEEE International conference on System Scinece, 2013.

[17] Sachchidanand S, Nirmala S, "Big data analytics," IEEE International conference on Communication,Information & Computing technology, Oct 2012, pp.19-20.

[18] Shidaganti, G., & Prakash, S, "Feedback analysis of unstructured data from collabrative networking a BigData analytics approach," IEEE International Conference on Circuits, Communication, Control and Computing, 2014, pp.343-347.

[19] Han, D., & Stroulia, E, "Federating Web-Based Applications on a Hierarchical Cloud.," IEEE 7th International Conference on Cloud Computing (CLOUD), 2014, pp. 946-947.

[20] Han, D., & Stroulia, E, "A three-dimensional data model in hbase for large time-series dataset analysis," IEEE 6th International Workshop on Maintenance and Evolution of Service-Oriented and Cloud-Based Systems, 2012 , pp.47-56.

[21] Bhandarkar, M, "MapReduce programming with apache Hadoop," IEEE International Symposium on Parallel & Distributed Processing (IPDPS), 2010, pp.1-2.

[22] Dubey, A. K., Jain, V., & Mittal, A. P," Stock Market Prediction using Hadoop Map-Reduce Ecosystem" IEEE 2nd International Conference on Computing for Sustainable Global Development, 2015, pp.616-621.

[23] Batool, R., Khattak, A. M., Maqbool, J., & Lee, S, "Precise tweet classification and sentiment analysis," IEEE 12th International Conference on Computer and Information Science, 2013,pp.461-466.

[24] Abuín, J. M., Pichel, J. C., Pena, T. F., Gamallo, P., & Garcia, M, "Efficient execution of Perl scripts on Hadoop clusters," IEEE International Conference on Big Data, 2014, pp.766-771.

[25] Gupta, C., Bansal, M., Chuang, T. C., Sinha, R., & Ben-romdhane, S., "A predictive model for anomaly detection and feedback-based scheduling on Hadoop.," IEEE International Conference on Big Data, 2014, pp.854-862.

[26] Wang, L., Tao, J., Marten, H," MapReduce across distributed clusters for data-intensive applications" IEEE International Symposium on *Parallel & Distributed Processing (IPDPS)*, 2012, pp.2004-2011.

[27] Kumar, R., Parashar, B. B., Gupta, S., Sharma, Y., & Gupta, N, "Apache Hadoop, NoSQL and NewSQL Solutions of Big Data," International Journal of Advance Foundation and Research in Science & Engineering (IJAFRSE), pp. 28-36.

[28] 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.

[29] Cloudera-http://www.cloudera.com

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

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

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

[33] Dagli, M. K., & Mehta, B. B, "Big Data and Hadoop: A Review" International Journal of Applied Research in Engineering and Science, 2014.

[34] 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.

[35] 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.

[36] Padhy, R. P,"Big Data Processing with Hadoop-MapReduce in Cloud Systems," International Journal of Cloud Computing and Services Science (IJ-CLOSER), 2012, pp 16-27.

[37] Xie, J., Yin, S., Ruan, X., Ding, Z., Tian., "Improving mapreduce performance through data placement in heterogeneous Hadoop clusters," IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW),2010,pp:1-9.

[38] Krishna, P. V., Misra, S., Joshi, D., & Obaidat, M. S, "Learning automata based sentiment analysis for recommender system on cloud," IEEE International Conference on Computer, Information and Telecommunication Systems, 2013, pp. 1-5.

[39] Hadoop-http://Hadoop.apache.org/

[40] Wu, X., Zhu, X., Wu, G. Q., & Ding, W, "Data mining with big data." IEEE Transactions on Knowledge and Data Engineering, 2014, pp.97-107.

[41] Zhang, F., Cao, J., Khan, S. U., Li, K., & Hwang, K, "A task-level adaptive MapReduce framework for real-time streaming data in healthcare applications " Future Generation Computer Systems Elsevier,2015, pp.149-160.

[42] Jin, S., Yang, S., & Jia, Y, "Optimization of task assignment strategy for map-reduce," 2nd IEEE International Conference on Computer Science and Network Technology (ICCSNT) , 2012,pp. 57-61.

[43] Palanisamy, B., Singh, A., & Liu, L, "ost-effective resource provisioning for mapreduce in a cloud," IEEE Transactions on Parallel and Distributed Systems, 2015, pp:1265-1279.

[44] Tang, Z., Jiang, L., Zhou, J., Li, K., & Li, K, "A self-adaptive scheduling algorithm for reduce start time" Future Generation Computer Systems, Elsevier, 2015, pp:51-60.

[45] Zheng, Z., Zhu, J., & Lyu, M. R, "ervice-generated big data and big data-as-a-service: an overview," IEEE International Congress on Big Data (BigData Congress), 2013, pp: 403-410.