Zain Publications
Company Registration No: UAN TS02E0015729
(Govt. of India, Ministry of Micro, Small & Medium Enterprises)
Contact Person: Dr. Mohd Zainuddin
Address: #12-1-898, North Lalaguda, Secunderabad, Telangana, India

M.V.S. Prasad *et al,* International Journal of Research in Information Technology, Volume 2, Issue 2, Feb 2018, Pg: 32-39
ISSN (Online): 2001-5569
UGC Approved Journal, Impact Factor: 1.021

PERFORMANCE ANALYSIS ON BIG DATA FRAMEWORKS

M.V.S. Prasad[1], Dr. O. Naga Raju[2]
[1]Research Scholar, Department of Computer Science and Engineering,
Acharya Nagarjuna University Andhra Pradesh, India -522510
[2]Lecturer in Computer Science, Ph.D & HoD of Computer Science,
S. K. B. R Govt. Degree College, Macherla, Guntur, Andhra Pradesh, India -522426
[1] mvs.prasad2009@gmail.com, [2] dr.o.nagaraju@gmail.com

Abstract: ***In the present digital life every day many people using different services and resources. Many applications such as Social Networks, Cloud Storages and E-commerce are generated big volumes of data. Every day the size of data will increase and manage these data is very essential. In all data mining applications maintains of data, reuse data and analytical support for future using big data frameworks. In this paper we discuss about the features of big data frameworks and also discuss the performance of big data frameworks. All these frameworks works as big data fashion on large volumes of data set. Here mainly discuss about the Map Reduce , Apache Spark and Apache Mahout. In this paper we compare the performance of Map Reduce , Apache Spark with the Apache Mahout.***
Keywords: ***HDFS, Map Reduce, Apache Spark, Apache Mahout***

1. Introduction:

In modern computer networks everyday users of internet increase tremendously. To manage the users information in the field of computer networks is very high complicated. For maintains and manage the information computer networks industry using distributed file system. Many applications like business operations, social networks, web applications, e-commerce applications day to day release large volumes of data sets. In Computer Systems have high revolutions from the few years. To process and scale out the large volumes data sets distributed file system plays a vital role. The distributed file systems not only use in web industry, it is also use in traditional enterprises[7]. The computation cost is more for the process of big data sets. Mostly the large computation is processor dependent and it handle small amount of data. When the data is copied to memory then the processor start the computation of data. The traditional

Zain Publications
Company Registration No: UAN TS02E0015729
(Govt. of India, Ministry of Micro, Small & Medium Enterprises)
Contact Person: Dr. Mohd Zainuddin
Address: #12-1-898, North Lalaguda, Secunderabad, Telangana, India

M.V.S. Prasad *et al,* International Journal of Research in Information Technology, Volume 2, Issue 2, Feb 2018, Pg: 32-39
ISSN (Online): 2001-5569
UGC Approved Journal, Impact Factor: 1.021

data processing methodology is very expensive in terms of computation, time and functionality of hardware. So the intention of distributed file system is reduce the computation cost and scale up the hardware functionality[3].

The data processing techniques are required to improve the speed and performance of the system. In many areas these techniques are useful like machine learning algorithms, queries and result analysis. Even though in streaming of real time data sources and batch processing this techniques are make useful. To make future platform easy and scale up the system and support new applications we use Distributed File System for large data sets.

To handle the large volumes of datasets and increase the processing speed in distributed file system have many frameworks. Here, we discuss about the procedure of frameworks and its efficiency in processing large volumes of data sets. In this paper mainly focus on Hadoop Mapreduce, Apache Spark and Apache Mahout.

2. Related Work:

Big data term using in distributed file system, it has large amount of data and unstructured various kinds of data. The traditional database management system unable to process and analyze such data because it large in size, unstructured in nature. The aim of the many applications is reduce the computation cost and scale up the performance of the system. In industry many business applications and e-commerce applications of intention is compute the unused large dataset and find out related data to make right decision. To process the big data efficiently have many open source technologies like Hadoop Mapreduce, Apache Spark and Apache Mahout[6].

For process the big data and maintenance more used best open source framework is Hadoop Mapreduce. Hadoop was designed based on Google File System and MapReduce framework. So in Hadoop framework , HDFS and MapReduce are the primary components. According to the programming model of MapReduce first the job divide the input data into chunks and allocate map tasks for processing simultaneously. The HDFS offered a efficient, scalable and replica based storage of data at different nodes. HDFS is works on master and salve

Zain Publications
Company Registration No: UAN TS02E0015729
(Govt. of India, Ministry of Micro, Small & Medium Enterprises)
Contact Person: Dr. Mohd Zainuddin
Address: #12-1-898, North Lalaguda, Secunderabad, Telangana, India

M.V.S. Prasad *et al,* International Journal of Research in Information Technology, Volume 2, Issue 2, Feb 2018, Pg: 32-39
ISSN (Online): 2001-5569
UGC Approved Journal, Impact Factor: 1.021

model. In this name node we cal master node and data node we cal slave node it had original data[3]. The replication factor is main concern, it configure accordingly user and usage type. The second main concern is MapReduce. It give permission to process replicated data and make successful in programming language techniques map and reduce. To achieve the big data computation Map phase is implemented with various mappers in distributed portions of datasets. The output of this phase are sorting and shuffling and goes to next phase. In this phase resultant data is aggregated and gives final output.

In processing of big data the Apache Spark is another way to work in distributed fashion. The Resilient Distributed Dataset (RDD) is introduced by Apache Spark. It gives application programming interface centered on data structure . RDD operates simultaneously and it had distributed collection of object. Map Reduce follow the cluster computing model, each map and reduce phases disk read and write operations are performed, it leads execution delay. The Apache Spark's RDD provides wide variety of technologies , its reduces the execution cost and time. The Apache Spark's RDD makes availability of all the data set data in local file system in the process of shuffle. Spark's also have other libraries for implementation of machine learning, steaming , SQL and graph programming.

In data mining for classification, clustering and frequent queries mostly preferred framework is Apache Mahout machine learning framework. Apache Mahout work in distributed paradigm , open source, scalable and had rich machine learning libraries. The main features of Apache Mahout is:

1. Mahout had different type of algorithms when we compared with other frameworks.
2. Mahout is useful to build scalable algorithms and its programming environment is simple extensible.
3. For implementation classification and clustering and etc ,. it had set of built-in libraries[2].

Zain Publications
Company Registration No: UAN TS02E0015729
(Govt. of India, Ministry of Micro, Small & Medium Enterprises)
Contact Person: Dr. Mohd Zainuddin
Address: #12-1-898, North Lalaguda, Secunderabad, Telangana, India

M.V.S. Prasad *et al,* International Journal of Research in Information Technology, Volume 2, Issue 2, Feb 2018, Pg: 32-39
ISSN (Online): 2001-5569
UGC Approved Journal, Impact Factor: 1.021

3. Processing Model:

3.1 Map Reduce:

In Big Data to process large volumes of data sets parallel mostly useful framework is MapReduce. It is mainly focus on data distribution and fault-tolerance specially designed by Google. High availability , parallel processing and batch processing are the features of MapReduce. MapReduce is built on Java API, so it's easy for programmers to develop the applications and deploy across cluster nodes. In MapReduce we have two types of phases, first one map phase and second one reduce phase.

3.1.2 Map Phase:

In Map Phase the input data is splitting to small task by task tracker. The split input data is actual input data format for processing of data inside the task tracker. The programmer have idea when developing the application, because the input data format will be different from one application to another. In map phase we have two internal process those are Splitting and Shuffling[5] .

3.1.2 Reduce Phase:

The output data of map phase will be input data for the reduce phase. The reducer take shuffle and sorted data. In this phase the incoming data is combine and write into hadoop distributed file system. The execution flow of MapReduce is First In First Out(FIFO)[5].

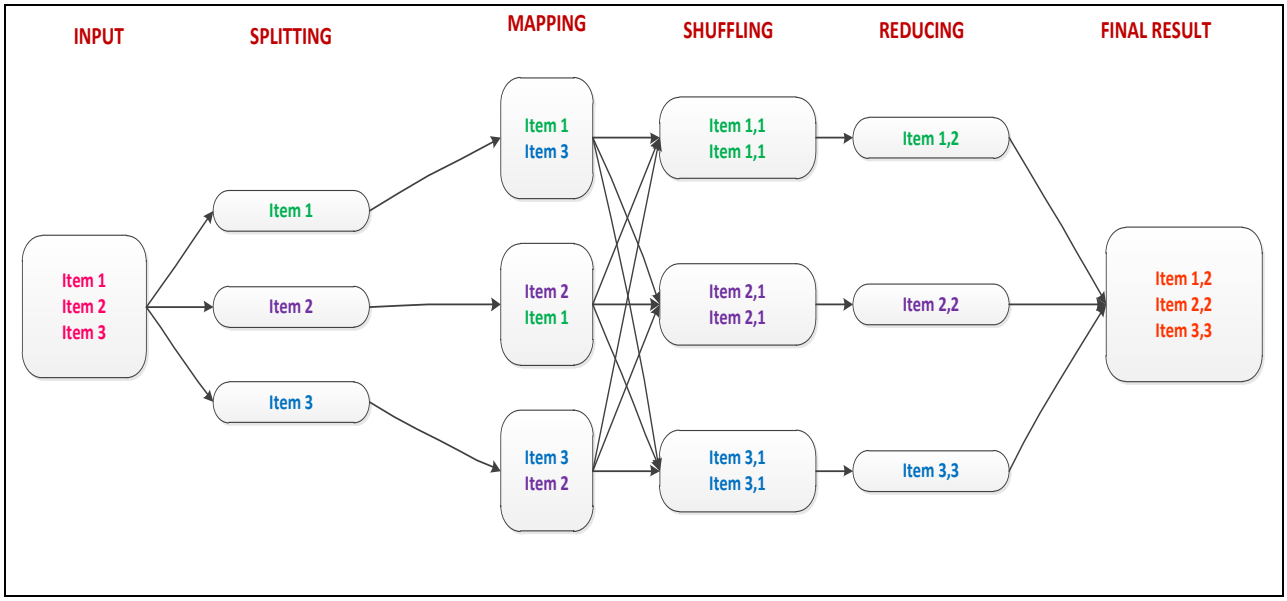

Fig : 1 MapReduce work flow

Zain Publications
Company Registration No: UAN TS02E0015729
(Govt. of India, Ministry of Micro, Small & Medium Enterprises)
Contact Person: Dr. Mohd Zainuddin
Address: #12-1-898, North Lalaguda, Secunderabad, Telangana, India

M.V.S. Prasad *et al,* International Journal of Research in Information Technology, Volume 2, Issue 2, Feb 2018, Pg: 32-39
ISSN (Online): 2001-5569
UGC Approved Journal, Impact Factor: 1.021

3.2 Apache Spark:

Apache Spark is open source framework to process large amount of data sets. It have high speed, user friendly and good for analysis purpose. It had more advantages when compare with other Big Data frameworks. Apache Spark is high flexible for any type of input data and it convert into varies required output(text, graph and etc)s.

Apache Spark is designed to work on multiple times on same data set. Spark execution engine works both in-memory and on-disk. Spark also perform external operations. First Spark store data into memory and then split on disk.

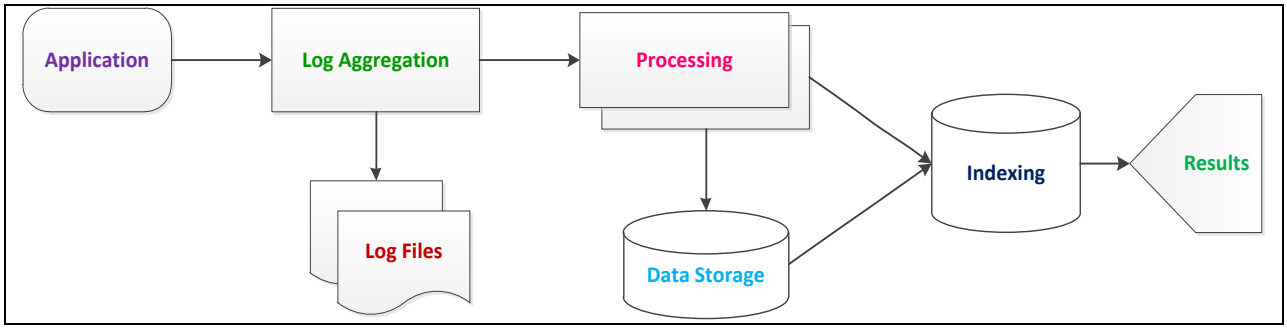

Fig 2: Apache Spark work flow

3.3 Apache Mahout:

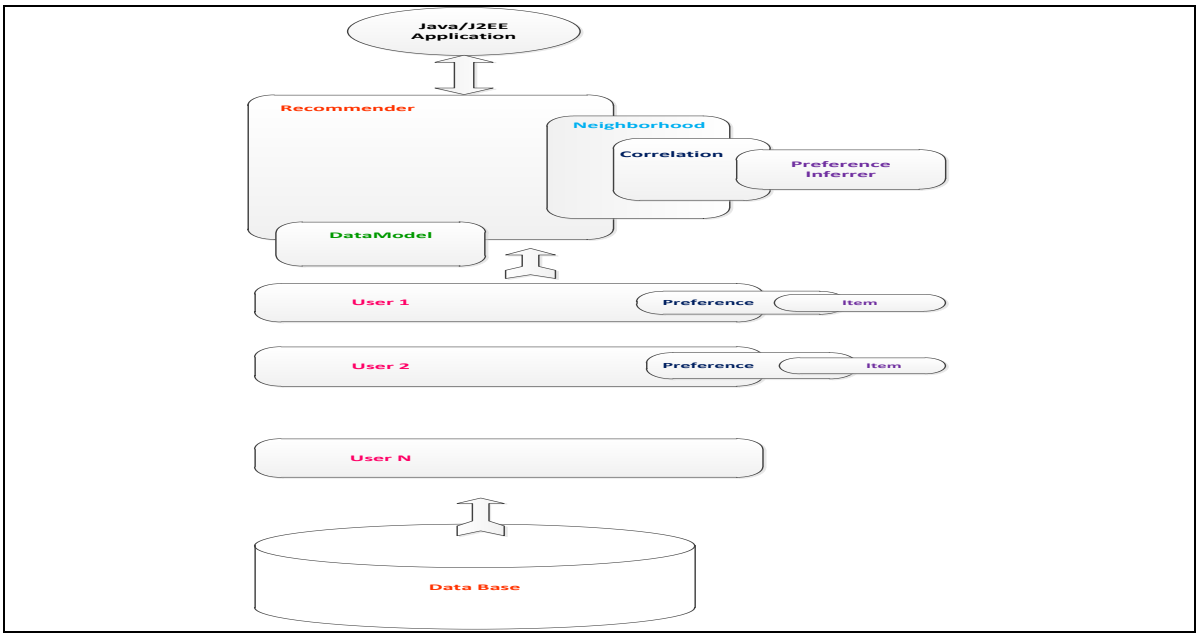

Fig 3: Apache Mahout work flow

Zain Publications
Company Registration No: UAN TS02E0015729
(Govt. of India, Ministry of Micro, Small & Medium Enterprises)
Contact Person: Dr. Mohd Zainuddin
Address: #12-1-898, North Lalaguda, Secunderabad, Telangana, India

M.V.S. Prasad *et al,* International Journal of Research in Information Technology, Volume 2, Issue 2, Feb 2018, Pg: 32-39
ISSN (Online): 2001-5569
UGC Approved Journal, Impact Factor: 1.021

Apache Mahout is most preferable and scalable framework to process massive large data sets. It had more built-in algorithms and rich machine learning API, when we compare with other frameworks. Apache Mahout is more preferable framework for researchers even in work standalone application. Apache Mahout is built on top of HDFS and MapReduce[6].

4. Result Analysis:

In experimental results we compare the performance of different big data frameworks. For analyzing those frameworks we taken different size of large data sets and measure the execution time. In the figure 4, we consider the three big data frameworks like MapReduce, Apache Spark, Apache Mahout. We take two different input size of files and calculate the execution time for each framework. In this experiments we observe that the MapReduce took high execution time than Apache Spark, Apache Mahout. So we can conclude Apache Mahout good performance when compare with Apache Spark and MapReduce. In bar chart red color represent the 1240 MB file size, other one is 62 Mb file size[4].

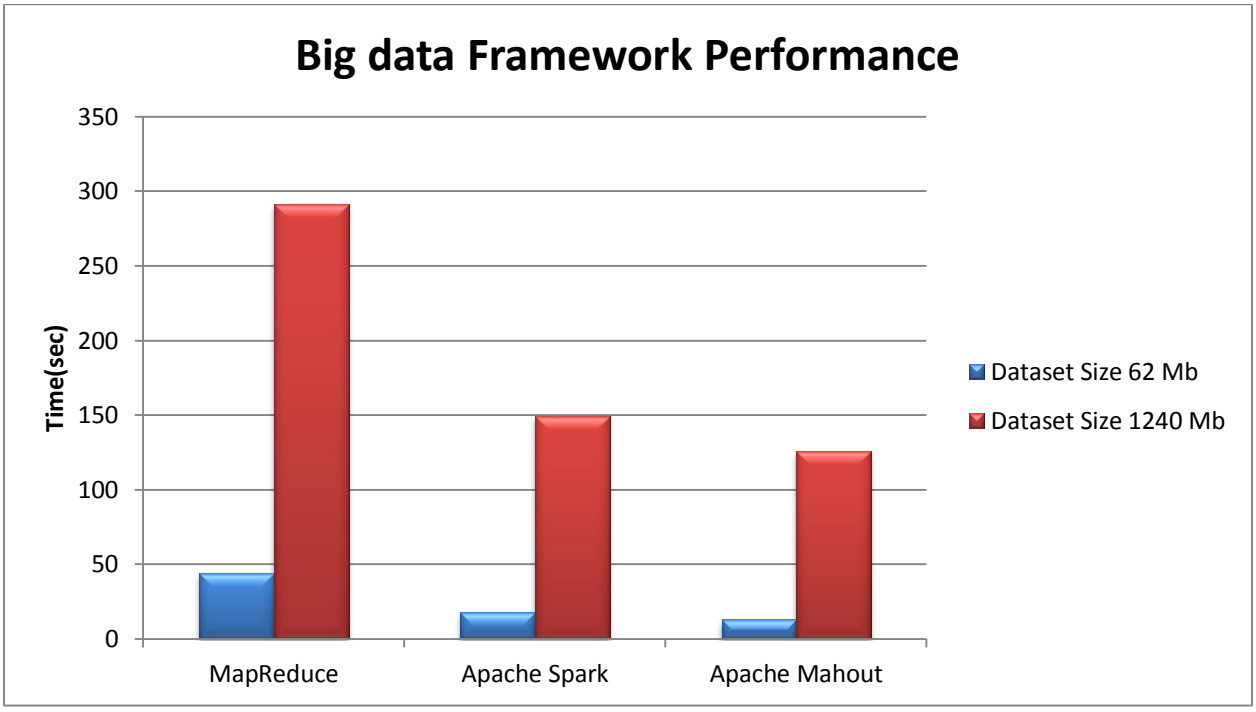

Fig 4 : Performance of Big data frameworks

Zain Publications
Company Registration No: UAN TS02E0015729
(Govt. of India, Ministry of Micro, Small & Medium Enterprises)
Contact Person: Dr. Mohd Zainuddin
Address: #12-1-898, North Lalaguda, Secunderabad, Telangana, India

M.V.S. Prasad *et al,* International Journal of Research in Information Technology, Volume 2, Issue 2, Feb 2018, Pg: 32-39
ISSN (Online): 2001-5569
UGC Approved Journal, Impact Factor: 1.021

In the figure 5, we compare three big data frameworks MapReduce, Apache Spark, Apache Mahout. We take different input size of files and calculate execution time for each framework. In this scenario we observe that the MapReduce took high execution time than Apache Spark, Apache Mahout. So Apache Mahout had better performance when compare with MapReduce and Apache Spark[2].

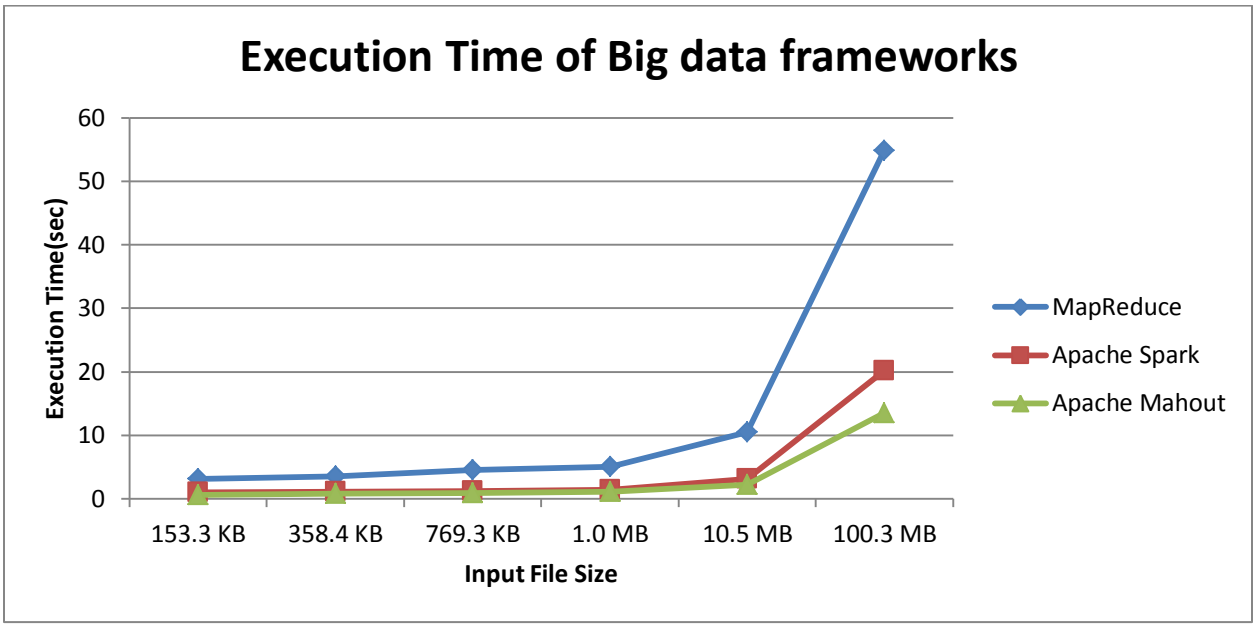

Fig 5: Execution time of big data frameworks

5. Conclusion:

In this paper done experiments on different type of files by using three big data frameworks. This paper gives the view for comparison results of big data frameworks. The experimental results shows that Apache Mahout framework given high performance when compare with MapReduce and Apache Spark frameworks. By finding the observing of all frameworks , the Apache Mahout had high capabilities for real time industry as well as enterprises applications. Apache Mahout fulfill all the needs of big data in processing. MapReduce, Apache Spark and Apache Frameworks are work in distributed fashion but Apache Mahout proven handling large amount of data sets efficiently and scalable .

Zain Publications
Company Registration No: UAN TS02E0015729
(Govt. of India, Ministry of Micro, Small & Medium Enterprises)
Contact Person: Dr. Mohd Zainuddin
Address: #12-1-898, North Lalaguda, Secunderabad, Telangana, India

M.V.S. Prasad *et al,* International Journal of Research in Information Technology, Volume 2, Issue 2, Feb 2018, Pg: 32-39
ISSN (Online): 2001-5569
UGC Approved Journal, Impact Factor: 1.021

References:

1. Dr.Venkateswara Reddy Eluri,Dr.M.RAMESH et al. (2016). A Comparative Study of Various Clustering Techniques on Big Data Sets using Apache Mahout. IEEE. 2 (7), 1-4.
2. Jai Prakash Verma, Atul Patel. (2016). Comparison of MapReduce and Spark Programming Fra-meworks for Big Data Analytics on HDFS. IJCSC. 7 (2), 80-84.
3. Prakasam Kannan. (2015). Beyond Hadoop MapReduce Apache Tez and Apache Spark. Computer Science Department San Jose State University. 2 (3), 1-6.
4. Satish Gopalani, Rohan Arora et al. (2015). Comparing Apache Spark and Map Reduce with Performance Analysis using K-Means. International Journal of Computer Applications. 113 (1), 1-4.
5. George John. (2016). Hadoop Map Reduce Architecture. Available: http://www.a4academics.com. Last accessed 16th Jan 2017.
6. Apache Software Foundation. (2016). Apache Mahout. Available: https://mahout.apache.org. Last accessed 6th Mar 2017.
7. Matei Zaharia. (2014). An Architecture for Fast and General Data Processing on Large Clusters. Electrical Engineering and Computer Sciences University of California at Berkeley. 12 (2), 1-128.