International Journal of Research in Advent Technology, Vol.7, No.3, March 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

Big data Processing Comparison with Map-Reduce and PIG

Santosh Kumar.J[1], B. K. Raghavendra[2], Raghavendra.S[3], Meenakshi[4]
Department of Computer Science and Engineering, KSSEM, Karnataka, India[1]
Department of Computer Science and Engineering, KSSEM, Karnataka, India[2]
Department of Computer Science and Engineering, Christ Deemed To Be University, Karnataka, India[3]
Department of computer science and Engineering, YDIT, Bengaluru, India4
Santosh.kumar.j@kssem.edu.in[1], hod.cse@kssem.edu.in[2],
raghav.trg@gmail.com[3],Meenakshib437@gmail.com[4]

Abstract- Big data is the data which is not able to store and analyze from available traditional system. Its Huge volume of data with Variety of types like audio, video, text, sign symbols and many more. With traditional system its difficult to analyze the variety of huge volume of data and also difficult to predict anything out of it, so the big data processing came in picture to store and analyze the big data, in market many big data software frame works like Amazon, Google Big data, Microsoft big data processing tool are available for processing the big data. Every device around us generates huge variety of data every fraction of seconds, that’s the big data which we need to process for enhance the organization, The frame work consists of many sub components like HDFS (Hadoop distributed file system) for storage, Map-Reduce frame work for processing, Yarn for resource management, PIG for processing unstructured data, Hive for processing structured big data, Hbase for storage of structured big data, Zookeper for co-ordination of jobs which jobs need to run first and next so on and also communication between all nodes for processing big data,Sqoop for import and export the structured static data from database to Hadoop and vice versa, flume for import unstructured big data from streaming devices to Hadoop and vice versa, here we are more keen in finding and comparing the execution time of Map-Reduce , PIG Script and Hive query for a bench mark program Word count. Here we have carried research for Map-Reduce program execution, and found that Hive is much faster compared with Map-Reduce and Pig.

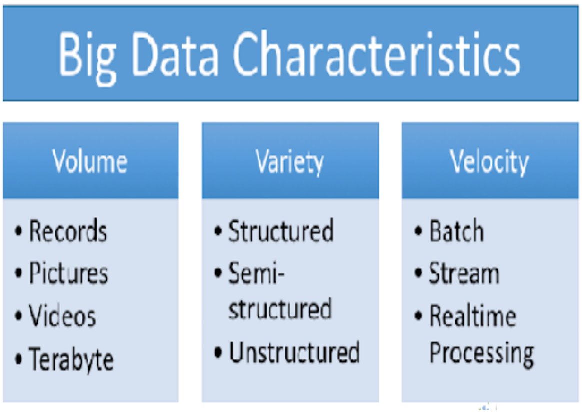
IndexTerms-Hadoop1,Map-Reduce2,Hive3,Pig4,wordcount5,cloudxlab6,Spark7,flink8.

1. INTRODUCTION

World is all about the data everything in the world connected with data each and every device generates and act on data, so data analysis playing very important role in improving the organizations, as every living and non living things generates huge volume, variety, velocity and veracity of data, it’s not only about mammoth of volume of data along with volume it might have velocity and variety parameters, data mining is everywhere almost all fields use data mining to dig knowledge out of generated, stored data, with respective to medical, many hospitals use the big data mining to analyze the patients health, Banking sector use the big data mining for analysis of fraudulent transactions and forecast good and bad customers for the bank, with reference to stock market from many years stock data available but used for analysis now started analyzing the stock data to predict and grow the business, even in education student data can be analyzed for betterment of students and organization, with reference to government sector a lot government sectors are collecting the data from many years but not utilized it properly, if they start analyze and utilize the data many problems around us can be solved, also historical data which say about our culture also can be treated as big data analysis, one more thing in even in research field we can use big data analysis and identify the good research and apply new patents, with reference to social media, twitter face-book and many social media generates huge volume of data very speed means every fraction of second one or the other person generates data with face book like sharing text, audio, video, commenting on particular topic, even his own login and log out data so huge big data is getting generated around us some of the generated data is important and some is not important to identify the important data we should go for analysis after identifying the important data we can save only save important and delete not important data, To process big data software frame works are developed by companies like Amazon, Google.

SPARK OVERVIEW.

Spark is the 100 times faster framework than Map Reduce and hdfs in storage and processing it is also frame work like any other java framework which built on top of OS to utilize memory efficiently and the other devices of CPU efficiently particularly designed framework for big data processing. Spark has many advantages and disadvantages efficient utilizations of Memory management is one of the disadvantage of spark whereas processing big data is advantages compared with map reduce framework and HDFS of Hadoop.



Big Data Characteristics

Volume	Variety	Velocity
• Records	• Structured	• Batch
• Pictures	• Semi-structured	• Stream
• Videos	• Unstructured	• Realtime Processing
• Terabyte		

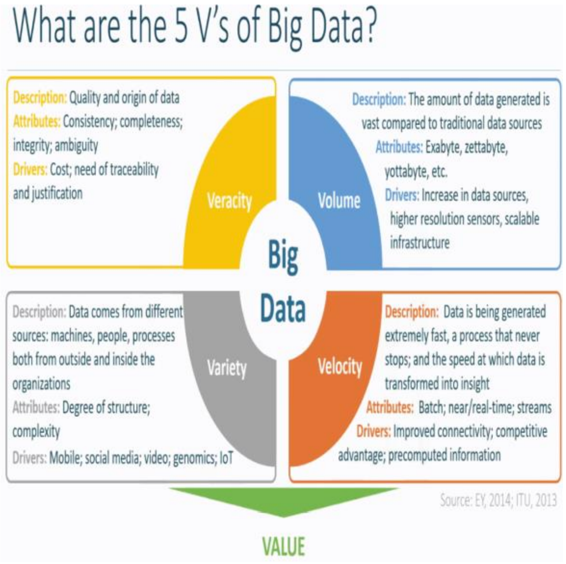
Figure 1: Big data Characteristics.

Big data basically means huge volume variety and velocity of data as shown in figure 1,

Big Data V's are

1. Variety –data generation sources are not same and they won't follow same format while generating data so we have verity of data like text images and videos signals.
2. Volume – as each and every fraction of second living and non-living things are connected with internet and IoT devices and human being generates huge volume of data.
3. Velocity – The data generation speed we cannot imagine.

Five V's of big data analogy is as shown in figure 2.



What are the 5 V's of Big Data?

Description: Quality and origin of data
Attributes: Consistency; completeness; integrity; ambiguity
Drivers: Cost; need of traceability and justification

Description: The amount of data generated is vast compared to traditional data sources
Attributes: Exabyte, zettabyte, yottabyte, etc.
Drivers: Increase in data sources, higher resolution sensors, scalable infrastructure

Description: Data comes from different sources: machines, people, processes both from outside and inside the organizations
Attributes: Degree of structure; complexity
Drivers: Mobile; social media; video; genomics; IoT

Description: Data is being generated extremely fast, a process that never stops; and the speed at which data is transformed into insight
Attributes: Batch; near/real-time; streams
Drivers: Improved connectivity; competitive advantage; precomputed information

Source: EY, 2014; ITU, 2013

Figure 2.Five V's of big data

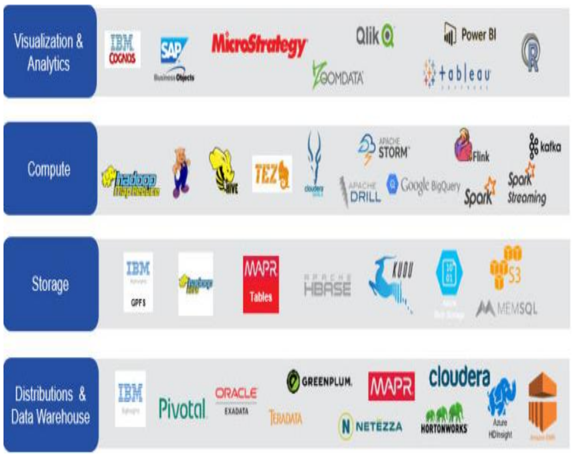


Figure 3.Eco-system components of big data.

Figure 3 above shows the Hadoop eco-system components initially Hadoop had HDFS and Map-Reduce later developed with YARN, H-Base, Zookeeper, Sqoop, Hive Flume, Pig, oozie and Mahout frame works, above all are added for user friendly usability purpose and improve performance of big data processing system.After development of Map-Reduce many started writing code using map-reduce, later Spark developed its very interesting features is efficiency, very fast compared with map-reduce, many industries dropped using of map-reduce and start using spark, it will be like replacing everything all applications of map-reduce to spark, on the other hand imagine having to manage big-data stack. To keep a moderate sized big data functioning, simultaneously developed around 10 different technologies, for storage, computing, and higher-level analytics, also to mention data discovery, data preparation, data quality, data security and data-governance, and data-visualization. New big data and analytics research has started to appear with the cloud. Like Google Deep-Learning offered on-demand on-premise any time. There are some advantages of using the cloud. But a hybrid system-environment is challenge in designing and managing a hybrid data management and architecture that connects the data with clouds.

2. RELATED WORK

The author said about big data techniques like hadoop, spark, flink for processing big data, above all are efficient technologies process big data. Hadoop MapReduce frameworks is replaced by emerging techniques like Spark and spark may replaced by Flink, which enhances the performance. The author compared evaluation of Hadoop, Spark and Flink using Big Data and considered performance and scalability parameters. And processing behavior of the above frameworks has characterized by varying some of the configuration parameters of the hadoop for the given work load, configuration parameters such as block size of HDFS, interconnect network, Size of input data and thread configuration. The said that

Spark or Flink leads to a reduction in execution times by 77% and 70% on average, respectively, for benchmarks [1].

Hadoop framework is Map-Reduce for storing and processing big data. However, to achieve good execution performance is the huge challenge due to large number configuration parameters. The author discussed about configuration parameters changing may enhance the performance, Also stated about machine learning techniques for improving the Hadoop performance. Then a deep learning algorithm is proposed for enhancement of Hadoop system performance [2].

Hadoop is widely used frameworks for MapReduce-based applications. But Hadoop basically have two main component HDFS and Mapreduce which itself have number of challenges, like resource management in Map Reduce cluster, to do that Yarn one more frame work added which which optimize the performance of Map Reduce. The author said Dynamic approach of resource management to enhance the system. The system has two operations one is slot utilization for efficiency optimization and utilization optimization. Also stated about dynamic technique which had 3 slot allocation techniques Out of Speculative Execution Performance Balancing, Dynamic Hadoop Slot Allocation Slot Pre-scheduling. Slot Pre-scheduling achieves a increased performance compared with cost-based optimization. Also enhances the performance with size variable input data [3].

The author of the paper discussed about the parallelism for enhancement of processing performance, huge amount of data is getting generating in today's world processing huge data with traditional system is very difficult must need parallelism, which require Virtual machines concepts, Map-reduce, dedicated clusters of servers large scale servers, to deploy and maintain these all require very high cost to overcome cloud infrastructures of Amazon, Microsoft, Google and many more companies providing Rent VM as pay as use[4].

The author discussed about commodity hardware of big data and distributed concepts i.e. distributed processing of big data, also the architecture of parallel computing, and data center deployment maintenance of system high performance Computing. Also compared the processing performance of single computing system and distributed computing system [5].

3. RESULTS AND DISCUSSION

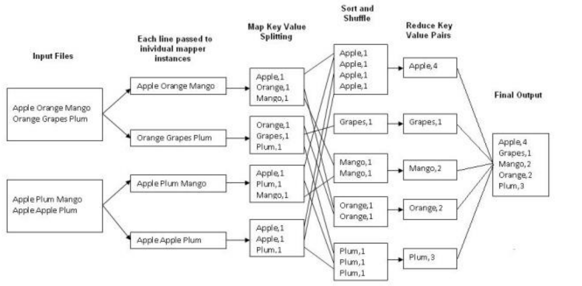


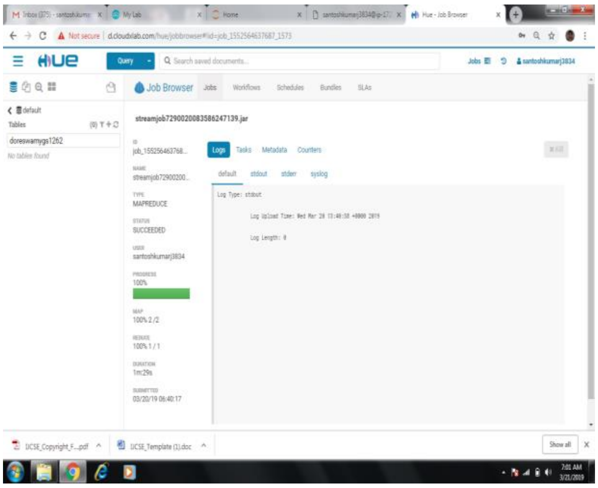
Figure 4: Map Reduce framework for word count

Figure 4 above is the Map-Reduce architectural framework for word count program where huge input file is split as blocks of pages and each pages split as lines and each lines spit as words by spaces to get number of words then all words are shuffled with all the data nodes mappers to count occurrence of each words in each data nodes finally using reduces combines the results achieved by each data node. Cloudxlab is the big data processing frame work provided by Cloudxlab organization for research. Using Cloudxlab following results are achieved for word count Pig script and hive query.

Figure 5,6,7,8 shows the execution time of word count program of Pig script and Hive Query.

Map-Reduce architectural framework is for word count program to count the occurrence of each word in a big data input file as shown in figure.4 where input file is split as of pages and pages split as lines and lines spit as words separated by spaces to get occurrence of words after that results are shuffled with all the mapper nodes to count occurrence of words in data nodes finally using reduces data node combines the results to achieved result.

Fig. 4, 5,6,7,8 shows the word count execution time with Map-Reduce python jar. Pig script and Hive Query.



streamjob7290020083586247139.jar

ID
job_155256463768...

NAME
streamjob72900200...

TYPE
MAPREDUCE

STATUS
SUCCEEDED

USER
santoshkumarj3834

PROGRESS
100%

MAP
100% 2 /2

REDUCE
100% 1 / 1

DURATION
1m:29s

SUBMITTED
03/20/19 06:40:17

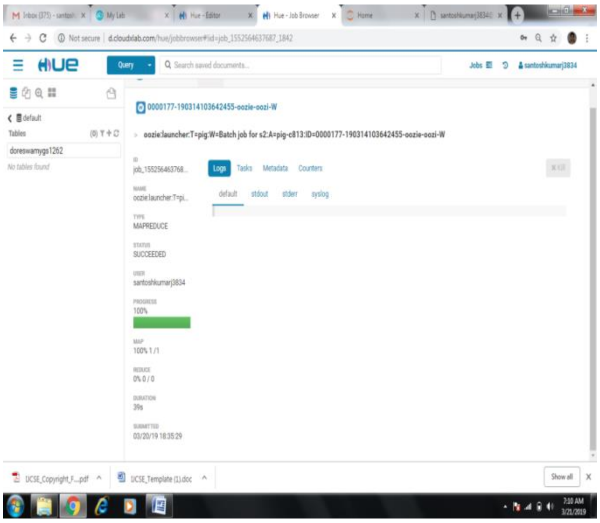
```
Log Type: stdout

        Log Upload Time: Wed Mar 20 13:40:58 +0000 2019

        Log Length: 0
```

Figure 5. Word count program execution time 89 Sec for input file with hadoop python jar program

Jar file of Python program run exec time 1m 29 sec = 89 sec for input file big.txt (2 map and 1 reduce)



0000177-190314103642455-oozie-oozi-W

oozie:launcher:T=pig:W=Batch job for s2:A=pig-c813:ID=0000177-190314103642455-oozie-oozi-W

ID
job_155256463768...

NAME
oozie:launcher:T=pi...

TYPE
MAPREDUCE

STATUS
SUCCEEDED

USER
santoshkumarj3834

PROGRESS
100%

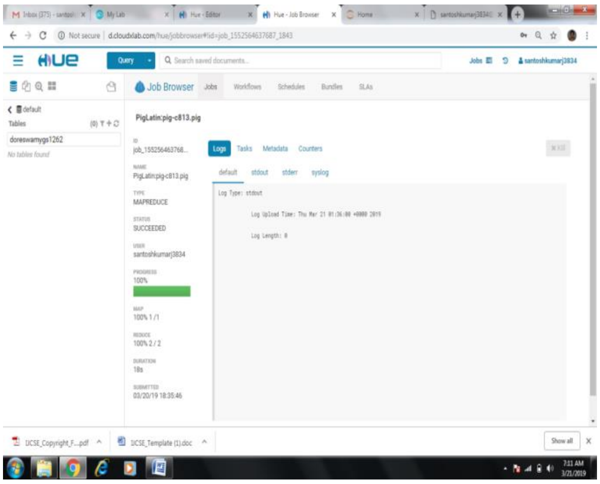
MAP
100% 1 /1

REDUCE
0% 0 / 0

DURATION
39s

SUBMITTED
03/20/19 18:35:29

Figure 6. Word count program execution time 39 Sec for input file.



PigLatin:pig-c813.pig

ID
job_155256463768...

NAME
PigLatin:pig-c813.pig

TYPE
MAPREDUCE

STATUS
SUCCEEDED

USER
santoshkumarj3834

PROGRESS
100%

MAP
100% 1 /1

REDUCE
100% 2 / 2

DURATION
18s

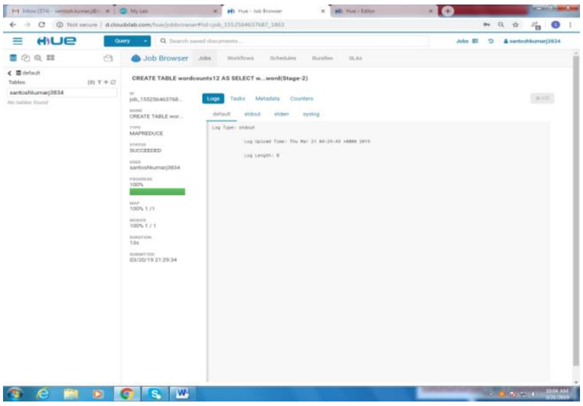
SUBMITTED
03/20/19 18:35:46

```
Log Type: stdout

        Log Upload Time: Thu Mar 21 01:36:00 +0000 2019

        Log Length: 0
```

Figure 7. Word count program execution time 18 Sec for input file.



CREATE TABLE wordcounts12 AS SELECT w...word(Stage-2)

ID
job_155256463768...

NAME
CREATE TABLE wor...

TYPE
MAPREDUCE

STATUS
SUCCEEDED

USER
santoshkumarj3834

PROGRESS
100%

MAP
100% 1 /1

REDUCE
100% 1 / 1

DURATION
13s

SUBMITTED
03/20/19 21:29:34

```
Log Type: stdout

        Log Upload Time: Thu Mar 21 04:29:42 +0000 2019

        Log Length: 0
```

Figure 8. Word count program execution time 13 Sec for input file.

Total of 36 sec + 16 sec = 52 sec of time to execute the word count program for input file.With PIG.

4. CONCLUSION AND FUTURE SCOPE

Hadoop is software framework for processing variety, volume and velocity of data, companies like google, Amazon provide framework for processing the big data, from above results we say that PIG Script is execution time is 52Sec whereas Map-reduce Jar program execution time is 89 sec for the same input file. And we may say that word count program for given input file PIG is better than Map-Reduce jar Program. Further we can compare the execution time with spark and flink with machine learning algorithms.

Acknowledgments

I acknowledge all the Authors and every one for supporting me.

REFERENCES

[1] Md. Armanur Rahman 1 , J. Hossen "A Survey of Machine Learning Techniques for Self-tuning Hadoop Performance" International Journal of Electrical and Computer Engineering (IJECE) Vol. 8, No. 3, June 2018, pp. 1854-1862

[2] AmanLodha , "Hadoop's Optimization Framework for Map Reduce Clusters " Imperial Journal of Interdisciplinary Research (IJIR) Vol-3, Issue-4, 2017

[3] Dan Wang, JiangchuanLiu , "Optimizing Big Data Processing Performance in the Public Cloud: Opportunities and Approaches" IEEE Network • September/October 2015

[4] A. K. M. MahbubulHossen1, A. B. M. Moniruzzaman et. al. "Performance Evaluation of Hadoop and Oracle Platform for Distributed Parallel Processing in Big Data Environments" International Journal of Database Theory and Application Vol.8, No.5 (2015), pp.15-26

[5] ChangqingJi,Yu Li, WenmingQiu et.al. "Big Data Processing in Cloud Computing environments "International Symposium on Pervasive Systems, Algorithms and Networks. 2012.