Big data Performance Evalution of Map-Reduce Pig and Hive

Santosh Kumar J, Raghavendra S, Raghavendra B.K., Meenakshi

Abstract: Big data is nothing but unstructured and structured data which is not possible to process by our traditional system its not only have the volume of data also velocity and verity of data, Processing means (store and analyze for knowledge information to take decision), Every living, non living and each and every device generates tremendous amount of data every fraction of seconds, Hadoop is a software frame work to process big data to get knowledge out of stored data and enhance the business and solve the societal problems, Hadoop basically have two important components HDFS and Map Reduce HDFS for store and mapreduce to process. HDFS includes name node and data nodes for storage, Map-Reduce includes frame works of Job tracker and Task tracker. Whenever client request Hadoop to store name node responds with available free memory data nodes then client will write data to respective data nodes then replication factor of hadoop copies the blocks of data with other data nodes to overcome fault tolerance Name node stores the meta of data nodes. Replication is for back-up as hadoop HDFS uses commodity hardware for storage, also name node have back-up secondary name node as only point of failure the hadoop. Whenever clients want to process the data, client request the name node Job tracker then Name node communicate to Task tracker for task done. All the above components of hadoop are frame works on-top of OS for efficient utilization and manage the system recourses for big data processing. Big data processing performance is measured with bench marks programs in our research work we compared the processing i.e. execution time of bench mark program word count with Hadoop Map-Reduce python Jar code, PIG script and Hive query with same input file big.txt. and we can say that Hive is much faster than PIG and Map-reduce Python jar code Map-reduce execution time is 1m, 29sec Pig Execution time is 57 sec Hive execution time is 31 sec.

Keywords – HDFS;Hadoop JAR;Pig; Hive;CloudxLab.

I. INTRODUCTION

Big data is not only about the huge volume of data along with volume variety like text, audio, video, signs unstructured and structured data and also the generation speed, which we not able to process with our traditional system. To process we have to use big data frame work like Hadoop developed by many companies like Amazon Microsofty and google. Each and every living and non living things also Humans generates data each and every fraction of seconds data generation is growing like anything every year its doubling the available quantity, Social medias like face book, twitter and many others generates the data very fast millions of users use social media every seconds share data among users, many retails companies gather data of every customer for future recommendation and suggest the customer to enhance the benefits. Sensors of IoT systems generates huge amount of data every fraction of seconds. Generated data need to analyze and take out knowledge out of it to enhance performance of organizations. Also definitely benefits the societal, mankind to come predict future medical and social problems also overcome of problems. Big data is not only about the huge data but it self is not able to store and process by our traditional system. Now we have many frame works, tools to store and process big data. Amazon uses the its retails customer data social media data for analysis and recommend the customer for future to enhance its business by analyzing Amazon doing millions transaction every day and the richest company in the world many other following amazon to enhance their business. So analyzing data is very important to improve the organization, big data huge volume of data along variety and velocity of data, the data generation speed is one of the biggest challenge to process. Big Data is unstructured i.e. not of uniform same type in nature variety of like video ,audio, text, signs, and so on, semi-structured data which have structured and unstructured combination data and structured data whose attribute data types and fields are not known. To process big data many software frameworks like Hadoop are available in market. Hadoop components are HDFS and Map Reduce to store and process. The eco-system of Hadoop have many components like Hdfs, Map-Reduce, Yarn, Pig, Hive, Hbase, Mahout, Oozie, Zookeeper.

Examples of big data frameworks are Amazon web services of elastic map-reduce, cloudex-lab, Microsoft web services, google big data services Oozie zookeeper pig hive and many more as shown in fig 2. Amazon web services user can create cluster to process the big data and other computation. Web services are charged on usage base pay as-use.

Each and every devices and organizations generates the data, simply they use to store and delete after few days, not analyzed it in today’s world organizations thinking of using the stored data for analysis and enhance the profit and enhance the performance.

Revised Manuscript Received on October 30, 2019.

* Correspondence Author

Santosh Kumar J*, Associate Professor in the Department of Computer Science and Engineering at K.S.School of Engineering and Management, Bangalore.

Dr. Raghavendra S., Associate Professor in the Department of Computer Science and Engineering at CHRIST DEEMED TO BE UNIVERSITY, Bangalore.

Dr.Raghavendra B.K. Department of Research Institute is an institution deemed to be university located in Chennai and Masters from Bengaluru

Meenakshi. assistant professor at Jain University Bengaluru India,

Retrieval Number F9002088619/2019©BEIESP
DOI: 10.35940/ijeat.F9002.088619
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Data is of different types like structured whose attributes or features (fields) and attributes or features types are known, semi-structured data whose attributes or features types are not known but attributes or features are known and unstructured whose attributes or features as well as its types not known. To process big data we have frame works like Hadoop which is developed by Benn cutting of yahoo. then enhanced by google and amazon.

The one point of failure is name node of Hadoop for that Hadoop have secondary name node with same daemons as main name node. HDFS means Name node and Data Node whereas processing means Job Tracker and Task tracker. Name nodes can communicate to job tracker and data nodes can communicate to task tracker for work done , finally task will be executed by data nodes. Client will request a task like read and write data from hadoop and to hadoop then name node will respond with free space data nodes client will write the data while processing data nodes will process their tasks and added by reducer for the final output.

Zookeeper will coordinates with all eco-system components for execution of jobs, Oozie will manage flow of work flow which all instances to execute and which order to execute for efficient utilization to improve the performance of Hadoop.

Three V's of Big Data

Variety –Different types sources generates different types of data which we cannot store in a table for processing examples on face book itself audio video text and signs variety of data by single sources etc.

Volume – The large volume of data like elephant as analogy mammoth of data each and every living and non living things generates data every fraction of seconds. IBM estimates that 3 quintillion bytes of data is created per day.

Velocity – The data generation speed is like faster than the speed of cheetah, every fraction of seconds every user generates the data. Three V's of big data analogy is as shown in fig 1.

Figure1. Three V's of big data

Fig. 2 shows the Hadoop eco-system, the components of frame work, early stage we had HDFS and Map-Reduce later stage YARN, PIG, Hive, Sqoop, Zookeeper H-Base, flume Mahout and Oozie frame works are added on top of OS for user friendly to improve the performance of system. later stage Spark is built to overcome the drawbacks of mapreduce then Flink is built to overcome the drawbacks of spark.

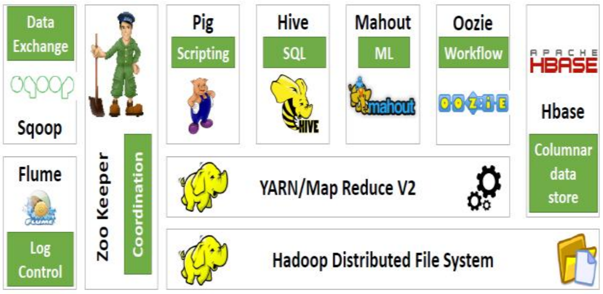


Figure 2. Eco system components of big data.

FLUME AND SQOOP

Sqoop is the frame work for data transfer from data base to Hadoop and Hadoop to database to import data from database to Hadoop no need to create a table where as to export data from Hadoop to database first we have to create a table in data base then we can use export commands, whereas Flume is also data transfer framework for un-structures data generated from web server, face-book and twitter for streaming unstructured data transfer we use Flume which have source, channel and sink components to import data for further analysis the data.

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.

FLINK OVERVIEW.

Flink is also a frame work for all components of Hadoop eco-system. Flink is the frame work for Streaming data, flinklatency is very less to process big data compared with Spark flinkhas many advantages, it processes the data without latency like speed of light, and Memory exception problem is also solved by flink.Flinkalso interact with many devices of which have different storage system to process the data, and it also optimizes the program before execution.

Big data has many performance measurement benchmarks programs as soon we install Hadoop we have test the performance of Hadoop with bench mark programs like Terra Gen, Terra sort, Terra validate, Pi, Word count, and many more Bench mark applications are along with Hadoop we just need to Run a Jar file of Bench marks to measure the .

II. LITERATURE SURVEY

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].

The author of the paper discussed about the big data processing issues, cloud management, Map-reduce optimizations techniques, also discussed the future of big data processing with cloud [6].

III. Results and Discussion

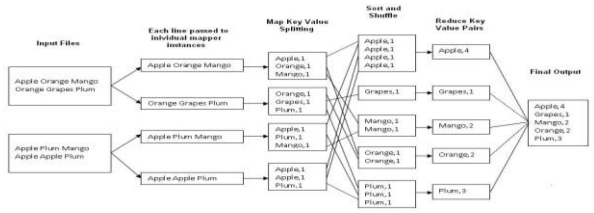
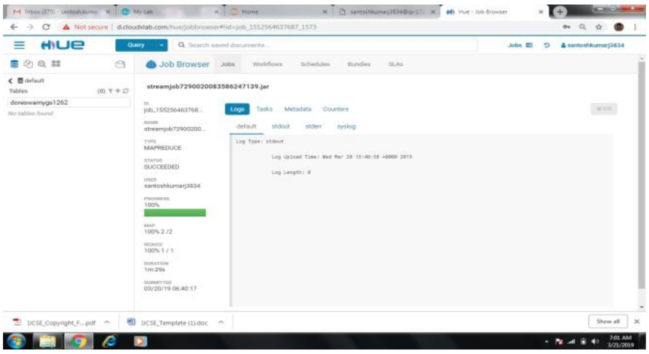


Figure 3. Map Reduce framework for word count

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 fig.3 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

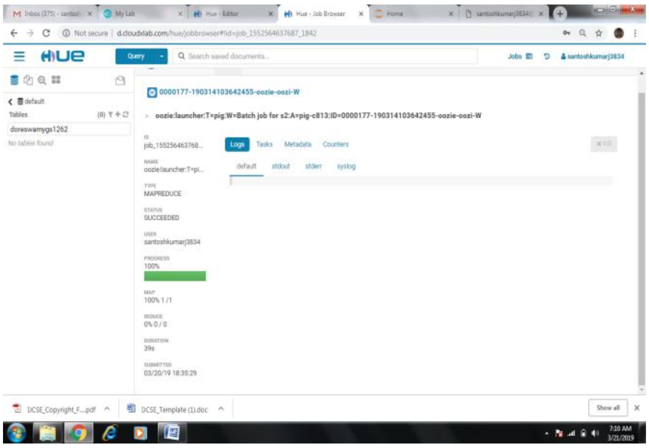
```
Log Type: stdout

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

        Log Length: 0
```

Figure 4. 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%

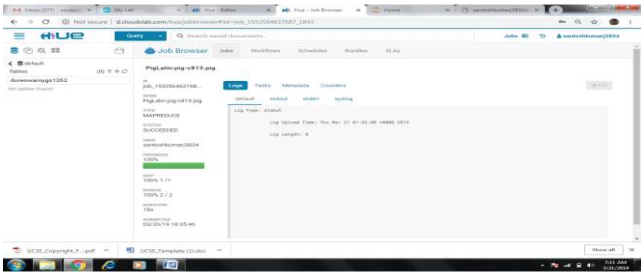
MAP
100% 1 /1

REDUCE
0% 0 / 0

DURATION
39s

SUBMITTED
03/20/19 18:35:29

Figure 5. Word count program execution time 39 Sec for input file.



Hue · Job Browser

PigLatin:pig-e813.pig

job_1552564637768...

PigLatin:pig-e813.pig

MAPREDUCE

SUCCEEDED

100%

100% 1 /1

100% 2 / 2

18s

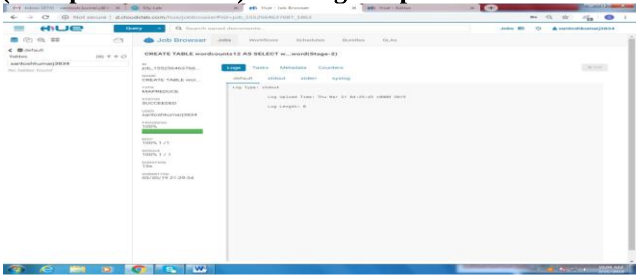
03/20/19 18:25:46

Log Type: stdout

Log Length: 0

Figure 6. Word count program execution time 18 Sec for input file.

Total execution time is 39 Sec + 18 Sec = 57sec (1 Map and 2 reduce) with Pig Script.



Hue · Job Browser

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

MAPREDUCE

SUCCEEDED

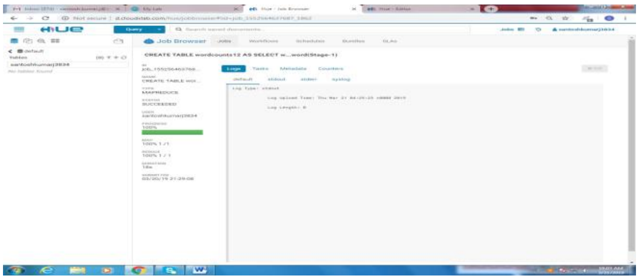
100%

100% 1 /1

100% 1 / 1

13s

Figure 7. Word count program execution time 13 Sec for input file.



Hue · Job Browser

CREATE TABLE wordcounts12 AS SELECT w...word(Stage-1)

MAPREDUCE

SUCCEEDED

100%

100% 1 /1

100% 1 / 1

18s

Figure 8. Word count program execution time 18 Sec for input file.

Total execution time is 13 sec + 18 Sec = 31Sec to execute the input big.txt file with Hive query.

IV. CONCLUSION

Hadoop is a framework for processing variety, volume and velocity of data from above results we can say that Hive is better than Pig and Hadoop map-reduce of python jar program while processing, Hive enhances the execution time. PIG script is faster than Map-Reduce Hadoop Python jar code and , Hive is faster than PIG and Map-Reduce Hadoop Python jar code. **Map-reduce execution time is 1m, 29sec Pig Execution time is 57 sec Hive execution time is 31 sec. as shown in fig.9,** Big.txt input file. With cloudxlab Hadoop big data frame work.

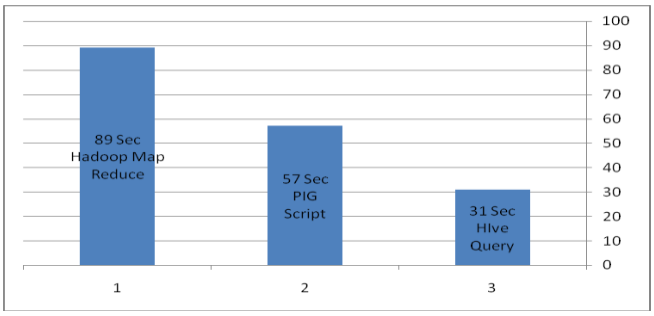


Figure 9. Word count execution time comparison.

ACKNOWLEDGEMENTS

I would like express my deep gratitude to the Principal, HoD and Staff of Computer Science and Engineering department of KSSEM, Bangalore for supporting me in doing this research work.

REFERENCES

1. Jorge Veiga, Roberto R. Expósito et al. "Performance Evaluation of Big Data Frameworks for Large-Scale Data Analytics" 2016 IEEE International Conference on Big Data (Big Data)
2. 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
3. AmanLodha , "Hadoop's Optimization Framework for Map Reduce Clusters " Imperial Journal of Interdisciplinary Research (IJIR) Vol-3, Issue-4, 2017
4. Dan Wang, JiangchuanLiu , "Optimizing Big Data Processing Performance in the Public Cloud: Opportunities and Approaches" IEEE Network • September/October 2015
5. 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
6. ChangqingJi,Yu Li, WenmingQiu et.al. "Big Data Processing in Cloud Computing environments "International Symposium on Pervasive Systems, Algorithms and Networks. 2012

AUTHORS PROFILE

Santosh Kumar J.is currently working as Associate Professor in the Department of Computer Science and Engineering at K.S.School of Engineering and Management, Bangalore. He is pursuing Ph.D. in VTU, Belgaum, Karnataka, India. He has 10 years of teaching and 3 years of industry experience. His interested area are Big data analysis. His reserach topics includesBig data with machine learning.

Dr. Raghavendra S. is currently working as Associate Professor in the Department of Computer Science and Engineering at CHRIST DEEMED TO BE UNIVERSITY, Bangalore. He completed his Ph.D. degree in Computer Science and Engineering from VTU, Belgaum, India in 2017 and has 14 years of teaching experience. His interests include Data Mining and Big data.

Dr.Raghavendra B.K. Pursued P.hd From Dr. M.G.R. Educational and Research Institute is an institution deemed to be university located in Chennai and Masters from Bengaluru University and Bachelors from Bengaluru University Karnataka He published nearly 20 reputed journals and His Area of interest is Data mining and Big data He currently working as Professor and Head CSE department ACU Bengaluru.

Meenakshi. Working as assistant professor at Jain University Bengaluru India, She completed Masters from VTU Belagavi and Bachelors from VTU Belagavi Karnataka Her Area of interest is Data mining and Big data.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication