IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-3, March 2019 E-ISSN: 2347-2693

Big data Processing Comparison using Pig and Hive

J. Santosh Kumar[1*], B. K. Raghavendra[2], S. Raghavendra[3]

[1,2]Department of Computer Science and Engineering, KSSEM, Karnataka, India
[3]Dept. of Computer Science and Engineering, Christ Deemed To Be University, Karnataka, India

[*]*Corresponding Author: santosh.kumar.j@kssem.edu.in, Tel.:+91 9035636616*

DOI: https://doi.org/10.26438/ijcse/v7i3.173178 | Available online at: www.ijcseonline.org

Accepted: 12/Mar/2019, Published: 31/Mar/2019

***Abstract*—**Big data is not only about mammoth volume of data along with volume velocity i.e. data generating speed like more than a speed of cheetah and also verity of data like a verity of vegetables in market, which we cannot process using our traditional system, processing is nothing but storing and analyzing the generated huge amount of verity of streaming and non-streaming data. Around us each and every device generates huge amount of structured and unstructured data. From many years many devices and organizations generates the data, generated data is not used by organizations for many years, now a day's organizations thinking of using the generated data for analysis and enhance the performance of organizations. Different data generation sources generate variety of data, i.e. Not of same in nature variety of data like structured whose features (fields) and features types are known, semi structured whose features types are unknown but features are known and unstructured whose features types and features are not known. To process big data Hadoop is developed by Benn cutting of yahoo later enhanced by google and amazon. Now amazon is number one company in the world because of analyzing the generated data. To process big data many tools and software frame work have been developed by many companies like Amazon, Google and Yahoo. Hadoop basically had two components like HDFS and Map Reduce one for storing and other one for processing, later stages YARN is added as recourse manager, before Yarn HDFS takes care of Recourse management which leads poor performance so YARN additional frame work added on top of Hadoop to manage recourse, along with Yarn later stages many other components like H-base-Hive, Sqoop are added to process only structured data and to process unstructured data. Pig and Flume are added to process unstructured data. Main work of Sqoop is to import and export structured data from database to Hadoop and vice versa. whereas flume is to import unstructured data generated from web server, twitter and face-book to Hadoop for analysis. The ecosystem of recent Hadoop are H-base, PIG, hive, Zoo-keeper, Oozie, flume, mahout machine learning tool and many more to make user friendly and to improve the performance of data analysis. Similar spark and flink are also competitors of hadoop spark which overcome limitations of Hadoop and flink which overcome the limitations of spark. In this we wanted to highlight the map- reduce applications for word-count bench mark examples, in our research we executed the bench mark word count program using pig and hive and achieved hive is much faster than PIG.

***Keywords*—**Hadoop;Map-Reduce;Hive;Pig;wordcount;cloudxlab;flink;spark.

I. INTRODUCTION

Big data is the data which we cannot process using traditional system. Humans and devices generates huge amount of data every seconds, we have millions of Petabytes of data around us, many devices like face book twitter sensors devices of IoT systems generates huge amount of data every seconds. Generated data if we analyze and make a knowledgeable out of it will definitely benefits the society, mankind and organizations. Big data is not only the huge data but not able to store and process by our traditional system. There are many tools available to store and process big data. Each and every IoT system device generates the lot of data, data generation is growing like anything past many years lot of data generated but not analyzed and not used properly, and now companies like Amazon used the data for analysis and became number one company in the world. So data analysis is very much important to enhance the organization business, big data is not only about huge data, along with that variety and velocity of data, data generation speed is one of the challenges to store and process. The streaming data is one more challenge. Data is not of uniform in nature variety of data like Un-structured, semi-structured, and structured whose data types and fields are not known. To process variety of data Hadoop have very important components basically it has main two components like HDFS and Map Reduce to store. The eco-system of Hadoop are cloudex-lab, amazon Web services elastic map-reduce components, also amazon cloud added many components like mahout Oozie

zookeeper pig hive and many more as shown in fig 2. In amazon user can invoke or create cluster to process the computation, web services are charged according to usage pay as-use. Machine learning is subject of Artificial intelligence which trains the machine and machine will train itself for making more intelligence for specific application, which allows improving their computational output based on previous input or learns with past errors. With thread parallelism big data computing performance may be improved.

Big data is not only about huge mammoth of volume along with volume velocity of fast generating data each and every living and non-living things generating data each and every fraction of seconds. and verity of data like audio, video, text signs, verity of vegetables in market, which we cannot process using our traditional system. To process big data many tools and software frame work have been developed by many companies like Amazon, Google and Yahoo. Around us each and every device generates huge amount of data that to mix of structured and unstructured data. From many years many devices and organizations generates the data which is not used by organizations now a day's organizations thinking of using the generated data for analysis and enhance the performance of organizations. And data is not of same in nature variety of data like structured whose attributes and attribute types are known, semi structured whose features types are unknown but attributes are known and unstructured whose features types and features are not known. To process big data Hadoop is developed by Benn cutting of yahoo later developed by google and amazon now amazon is number one company because of analyzing the generated data Hadoop basically had two components like HDFS and Map Reduce.

The single point of failure in Hadoop is Name node for that Hadoop maintains Secondary Name nodes with same daemons like main name node. Whenever we come across storage part HDFS Name node and Data Node will come in picture whereas processing part Job Tracker and Task tracker. Name nodes can talk to job tracker and data nodes can talk to task tracker finally work will be done by data nodes. First client will request for read and write of big data to name node, then name node will responds from its metadata to client then client will write data to respective data nodes. Meta data like what all data nodes are free and occupied by which sources, client will write data to data node with replications factor , same block of data will be stored with other data nodes to overcome fault.

Zookeeper will coordinates with all eco-system components for smooth running of jobs.
Oozie will take care of flow of work what all instances to run and which order they should run for efficient utilization and improve the performance of Hadoop.

Three V's of Big Data

1. Variety –The variety of sources which generates the data with their own formats which is not of same type, humans and machines generates data, credit card data health care data twitter data face book data which itself have text images and videos signs etc.
2. Volume – The huge amount petty bytes of data like mammoth each and every seconds every device round us generates huge volume of data, IBM estimates that 2.8 quintillion bytes of data is created each day
3. Velocity – The Speed of data generation is very high, every fraction of seconds each and every user generates the data that's the speed, using traditional system very difficult to store.

Three V's of big data analogy is as shown in fig 1.

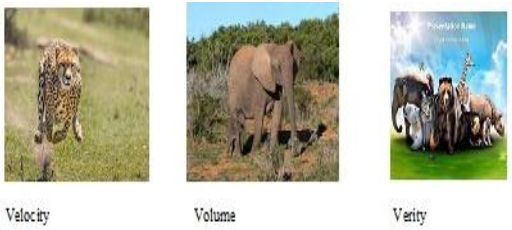

Figure1: Three V's of big data

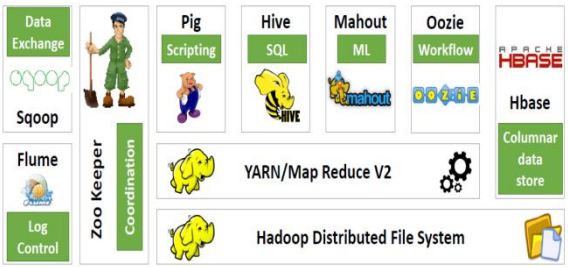

Figure 2: Eco system components of big data.

Fig. 2 above shows the Hadoop eco system components basically had HDFS and Map-Reduce later YARN Zookeeper H-Base Sqoop flume Hive Pig Mahout Oozie frame works are added for user friendly and improve the performance of system on top of this Spark and Flink are also added as competitors Hadoop .

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 and 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 MapReduce and hdfs in storage and processing. It is also a frame work like any other java framework which built on top of OS to utilize memory efficiently and the other devices of CPU efficiently. It's a particularly designed framework for big data processing. Spark has many advantages and disadvantages like 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, flink latency is very less to process big data compared with Spark, flink has many advantages, it processes the data without latency like speed of light, and Memory exception problem is also solved by flink. Flink also 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, terra gen terra sort, Pi, and many more Bench mark applications are along with Hadoop we just need to Run a Jar file of Bench marks to measure the Performance.

Performance Metrics of big data are application metrics and system metrics, Application Metrics is how quickly one can reduce and how many events can be reduced on following factors like rreduction factors, size of the events and time taken for reduction. System Metrics are Memory Usage and caching strategy Input output metrics CPU time for the executors. Time in serialization and spent on Garbage in garbage collection, Data read from HDFS, Network traffic measure.

II. RELATED WORK

Many Author of the paper said about apache Hadoop that it is a framework for processing large distributed data set across cluster of computers and said about scaling the cluster. Due to use of sensors across all devices and network tools of the organizations generating big data, all wanted to store and analyze without investing much cost on managing and service issue of the storage and processing want to deploy everything on cloud so that cloud management organizations will take care of it, these companies can utilize the data for analysis and extract useful knowledge out of it. Map Reduce is the framework which allows large data to be stored across all devices and processed by devices map functions will distribute the data and store across the devices where a reduce will process the query of the client it works on bases of the key value pair. Each line will be treated as key and value that is first word is the key and rest all will be value whenever client request to process the large data first client will approach the name node name node will respond to client with available free nodes after that mapper functions by client will write data to respective data nodes, and whenever client want to process the data it request to name node job tracker then job tracker will communicate to name node to get data information storage then it will assign jobs to task tracker to process the job by name nodes will process the task by their available data then one of the node will aggregate the result and give the result to client.

The author said that increasing the usage of Big Data has led to efficient technologies to manage and process big data. Map Reduce frameworks like Hadoop is replaced by emerging techniques like Spark and Flink, which improves the performance. The author done comparative evaluation of Hadoop, Spark and Flink with Big Data workloads and considered factors as performance and scalability. And the behavior of above frameworks has been characterized by changing some of the parameters of the workloads such as HDFS block size, interconnect network, input data size and thread configuration. The analysis said that replacing Hadoop with Spark or Flink leads to a reduction in execution times by 77% and 70% on average, respectively, for non-sort benchmarks [1].

The Apache Hadoop framework is a Map Reduce for storing and processing big data. However, to achieve good execution performance from this is a huge challenge because of large number configuration parameters. The author said the critical issues of Hadoop system, big data and machine learning analysis, for improving the Hadoop performance. Then a deep learning ML algorithm is proposed for Hadoop system performance improvement [2].

Apache Hadoop is the most widely used frameworks for MapReduce-based applications development. But Hadoop have number of challenges, like management of the resources in the Map Reduce cluster, which will optimize the performance of Map Reduce. The author stated about Dynamic approach for managing speed up of the resources available. It has two operations, one is slot utilization efficiency optimization and utilization optimization. The author Dynamic technique has 3 slot allocation techniques

one is Speculative Execution Performance Balancing, Dynamic Hadoop Slot Allocation and Slot Pre-scheduling. It achieves a increased performance compared with cost-based optimization. Also increases the performance with input data set size[3].

Today's world parallelism is the way of enhancing the performance due to fast data generation from many sources, which needs unprecedented demands on the networking infrastructures and computing. Map Reduce, is generally performed on reserved dedicated servers clusters for parallelism. For non-computing professional and grass-root user, large-scale specific dedicated server clusters deploying, maintaining cost is very high. Whereas the public clouds gives users to rent virtual machines (VMs) and run applications with pay-as-you-go manner [4].

The authors Sid about Distributed processing of big data across clusters of computers using distributed and parallel computing architecture and also the data center deployment, maintenance cost using commodity hardware for high performance Computing. And did compared the performance of traditional single computing system and distributed parallel computing system [5].

The authors discussed about big data and Cloud data management mechanisms and the processing issues of big data, with reference to cloud computing, cloud database, cloud architecture, data storage, Map Reduce optimization techniques. The author also said about the future on big data processing in cloud computing environments [6].

The authors discussed the Resource management for Map Reduce-based applications processing to deploy and resizing Map Reduce clusters Bench marking applications and tool are used for the Map Reduce processing to measure the Map Reduce performance using workloads with big data and to optimize the Map Reduce to process terabytes of data efficiently [7].

The authors discussed about software to improve the scalability of data analytics, Challenges Availability, partitioning, virtualization and scalability, distribution, and elasticity and performance bottlenecks for managing big data [8].

The authors said about Benchmarking a several of high-performance computing (HPC) architectures for data, name node and data node architectures with large memory and bandwidth are better suited for big data analytics on HPC h/w [9].

Map Reduce provides a parallel and scalable programming model for data-intensive business and scientific applications. To obtain the actual performance of big data applications, such as response time, maximum online user data capacity size, and a certain maximum processing capacity [10].

III. RESULTS AND DISCUSSION

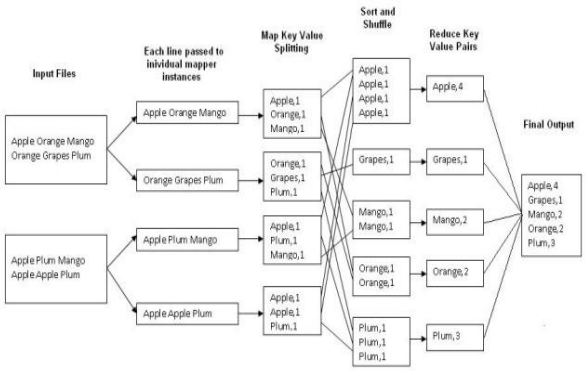

Figure 3: Map Reduce framework for word count

Fig. 3 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.

Fig. 4,5,6,7 shows the execution time of word count program of Pig script and Hive Query for the input file which I have loaded in cloudxlab big data frame work.

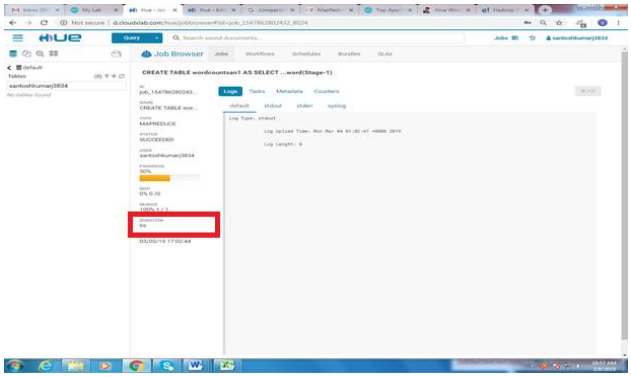

Figure 4: word count Hive Query execution time 6 sec for input file.

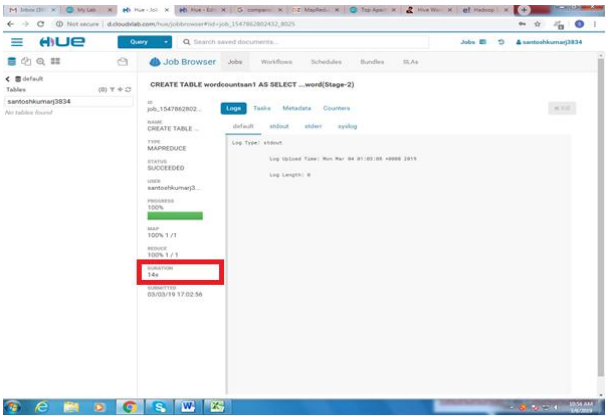

Figure 5: word count Hive Query execution time 14 Sec for input file.

For hive query First we create a table called doc then will load a input file after that we executed the word count hive query program, which showing a time of 14sec + 6sec = 20 Sec to process word count for the given input file on cloudxlab.

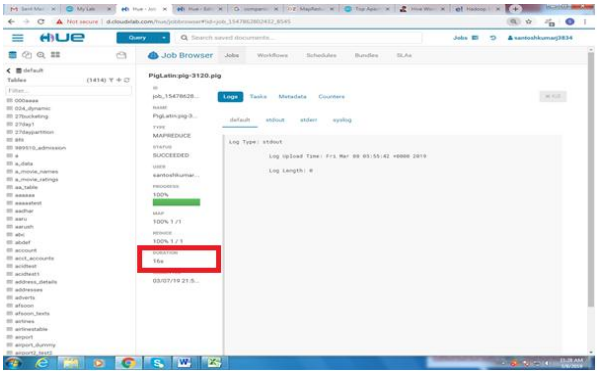

Figure 6: word count program execution time 16 Sec for input file.

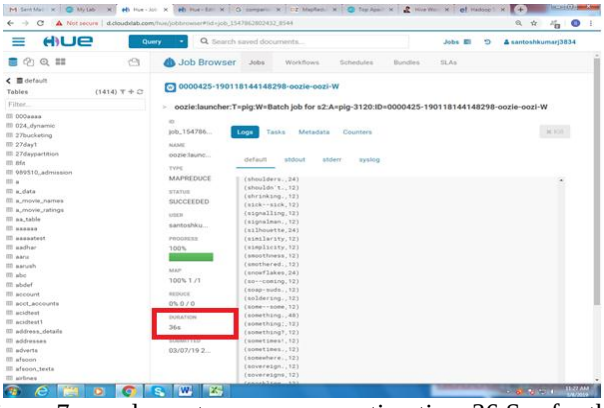

Figure 7: word count program execution time 36 Sec for the input file.

Using Pig script word count bench mark program is executed and got same results as hive query with different execution time i.e. 16sec + 36sec = 52 Sec to process word count for the given input file on cloudxlab.

IV. Conclusion and Future Scope

Hadoop is software framework for processing variety, volume and velocity of big data, companies like google yahoo and Amazon have their own big data framework for processing the big data also they provide cloud based big data eco-system infrastructure to store (HDFS) and process (map-Reduce) big data, from above figures results says that Hive query execution time faster than pig script to process big data input file.

When we executed the bench mark of big data performance word count program with PIG script and Hive query above results are achieved. And figure 4 and figure 5 show the execution time of 20 sec. with Hive query, whereas figure 6 and figure 7 shows the execution time of 52 sec with PIG Script. For the input file which we have loaded on cloudxlab. Further we can compare the execution time with spark and flink, also we can compare big data processing performance using machine learning algorithms.

Acknowledgment

I acknowledge every one for supporting me for doing research.

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

[7] Bogdan Ghiţet. al. "Towards an Optimized Big Data Processing System" 13th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing, 2013

[8] Kyong-Ha Lee et. al. "Parallel Data Processing with Map Reduce: A Survey" SIGMOD Record, December 2011 (Vol. 40, No. 4)

[9] JaliyaEkanayake and Geoffrey Fox "High Performance Parallel Computing with Clouds and Cloud Technologies" International Conference on Cloud Computing, 2009 – Springer
[10] Ashlesha S. Nagdive et al, "Overview on Performance Testing Approach in Big Data" International Journal of Advanced Research in Computer Science, 5 (8), Nov–Dec, 2014, pp165-169.

Authors Profile

Mr. Santosh Kumar Jankatti Pursed Bachelor of Engineering and Master of Technology from VTU, Belagavi, Karnataka. He is pursuing Ph.D from VTU Belagavi, Karnataka and currently working as Associate Professor in Department of Computer Science and Engineering, KSSEM, Bengaluru, Karnataka. He has published more than 5 research papers in reputed international journals. His main research work focuses on Data mining and Big Data Analytics, IoT. He has 10 years of teaching experience and 3 years of Research Experience.

Mr. Raghavendra B. K. Pursed Bachelor of Engineeriing from Bangalore University,Bengaluru,and Master of Technology from VTU, Belagavi, Karnataka. He pursued Ph.D from VTU,Belagavi, Karnataka and currently working as Professor in the Department of Computer Sciences and Engineering, KSSEM, Bengaluru. He has published more than 10 research papers in reputed international journals. His main research work focuses on Data mining and Big Data Analytics. He has 15 years of teaching experience and 10 years of Research Experience.

Mr.Raghavendra S. Pursed Bachelor of Engineeriing and Master of Technology and Ph.D from VTU, Belagavi, Karnataka and currently working as Associate Professor in the Department of Computer Science and Engineering, Christ Deemed To Be University,Bengaluru, Karnataka. He has published more than 10 research papers in reputed international journals. His main research work focuses on Data mining and Big Data Analytics. He has 14 years of teaching experience and 5 years of Research Experience.

