Survey on Efficient Processing of Job by Enhancing Hadoop MapReduce Framework

[1]Patil Pankaj Shantaram, [2]Prof. R. B. Wagh

[1]PG Student, [2]Assistant Professor, [1,2]Department of Computer Engineering, R. C. Patel Institute of Technology Shirpur, Maharashtra, India.

[1]*patilpankajs1107@gmail.com*, [2]*rajanikantw@gmail.com*

Abstract - From long time users who used to store and analyze data would store data in database and further process through SQL queries. Data might be unstructured and large which is difficult to store and process. Such problem faced by Google organizations, they build MapReduce framework which used for large-scale data processing consist of map and reduce functions. Implementation of MapReduce paradigm referred as Hadoop. In Cloud Computing MapReduce platforms, resource allocation remains a challenge. So, enhanced Hadoop architecture is developed known as H2Hadoop which is used to reduce the computation cost which is consist of BigData. H2Hadoop also provides solutions over issue of resource allocation also find DNA sequence in text data. Distributed file system is provided by four types of nodes such as NameNode, DataNodes, JobTracker, and TaskTracker which are also known as nodes of Hadoop. Hadoop is also consisting of distributed file system to perform given job known as Hadoop Distributed File System; in short it is referred as HDFS. H2Hadoop reduces CPU time by directly assigning job to DataNode which contain required data without assigning to whole cluster. Also reduce number of read operations.

Keywords — BigData, Cloud Computing, Hadoop, H2Hadoop, MapReduce.

I. INTRODUCTION

Genomic data might be structured, semi-structured or unstructured data which is processed using open source cloud computing platforms. Such data can be complex and difficult to understand by users which require efficient algorithms and computational power to translate data into human readable form. Cloud computing provides resource sharing with middleware's, application development platforms, and business applications. Cloud is consisting of usable and accessible resources which are used for resource utilization. For DNA sequence alignment large and complex amounts of data processing is required and also required targeting and scheduling of resources to solve this problem. Also it is necessary to define certain definitions of BigData, Hadoop and MapReduce which are used to increase the performance of enhance Hadoop MapReduce framework.

A. BigData Concepts

BigData is consist of large datasets which is not processed using existing computing techniques in a less processing time, also used in many areas of business and technology. BigData might be relational database which is also called as structured data which consist of stock market data. Also BigData would be non-relational database which is also called as semi structured data which consist of word, PDF, DNA datasets, text or social media data. Thus BigData consist of huge volume up to Zetabyte, high velocity which means the speed of data generation is high, extensible variety of data which means generated data is numeric data, sequence data or binary data, and veracity which means how clear the data is. These are also known as 4V's of BigData.

To process BigData it is important to develop new framework which makes BigData easy to manage, fast to implement, cost reductions, and can protect data privacy and security. NoSQL BigData system works based on real-time data by taking advantage of cloud computing architectures with minimum amount of coding and without any need of additional infrastructure. There are several challenges faced by BigData such as process of extracting the information from database, physical storage, which is used to store

 DOI : 10.18231/2454-9150.2017.0008

BigData, integration, cleaning, aggregation, and representation of data. BigData needs environment which overcomes on these challenges and it is called as Hadoop which is a framework used to process BigData.

B. Hadoop Overview

Hadoop is the framework which is used for distributed processing and storage of large datasets into cluster. As Hadoop works with the MapReduce algorithm and get written in java which also known as Apache open source framework provides solutions for BigData processing and analysis. Figure 1 shows Hadoop architecture which has two major layers namely, Processing or Computation layer consist of MapReduce which provides the system analysis and Storage layer provides data reliability consist of Hadoop Distributed File System. Hadoop consist of core tasks in which divides data into directories and files then distributed such data for further processing. HDFS is located on top of the local file system which supervises the processing in which blocks are replicated used in handling hardware failure and checking that the code was executed successfully. Hadoop framework is designed for detecting and handling failures at application layer and automatically distributes the data and utilizes parallelism.

MapReduce is used to process large amount of data, on clusters which has thousands of nodes which is about multi-terabyte datasets. The processing of data is done in fault-tolerant manner where MapReduce algorithm has been used to access log analysis, document clustering, generating search indexes and data analysis. A MapReduce job splits the input dataset into independent blocks and stores them in HDFS. MapReduce job is also known for distributed computation across clusters and also the job results are stored in HDFS.

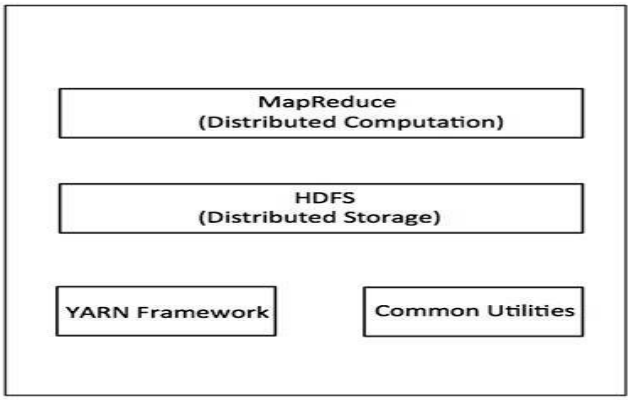

Fig 1: Hadoop Architecture.

Hadoop consist of file system known as Hadoop Distributed File System (HDFS) which is used to store large data files across the cluster. In HDFS, data files are written only once, in which Hadoop divides the data into blocks then duplicated in the HDFS and read multiple times. In HDFS, the cluster consist of Hadoop is get divided into two components, such as the master node called as NameNode and the slaves called as DataNodes. Single NameNode is responsible for saving the data and assigning jobs to appropriate DataNodes that store application data. Also there are two core components of Hadoop framework such as Hadoop Common which are Java libraries and utilities and Hadoop YARN which is a framework for job scheduling and cluster resource management.

C. MapReduce Analysis

MapReduce model is used to write a simple program and run it on thousands of machines. It is also responsible for speeding up the development process of system. Also, it is easy to use without experiencing parallel and distributed system by programmers. MapReduce is one of the component of Hadoop which is responsible for providing system analysis, such as providing data managing system, stream reading access and runs tasks on cluster of nodes. MapReduce reads input data in text format which must be read as text file and output is also in text format. It is also called as simplified data processing on cluster.

MapReduce consist of two functions such as Map and Reduce which are written by the user. Map takes input and provides the set of intermediate key/value pair and gets buffered in memory. Then the groups of all intermediate values which are consist of similar intermediate key further provided to Reduce function. Reduce function accepts that intermediate key and set of values for that key and merges together which finally gives the smaller set of values. MapReduce library in user program first splits input files into number of pieces which consist of 16 MB to 64 MB per piece, from that one program which is special and called as Master. And remaining are Slaves on which work is assigned by Master. Master selects any idle Slave and assigns each one Map task or Reduce task and also pings every Slave periodically.

D. H2Hadoop

In present architectures, the location of data blocks which is present in HDFS file system, known to the NameNode. NameNode then assign jobs to DataNode, knowing which DataNode present in HDFS holds the blocks of required data. It is also able to direct jobs to specific DataNode and get the results without going through entire cluster. In H2Hadoop implementation of some pre-processing is performed in NameNode, which is used for dividing job into various tasks and also assign jobs to client. The pre-processing process consists of building metadata table, which contains location

 DOI : 10.18231/2454-9150.2017.0008

of required data blocks. Also it can only read data from specified blocks without sending job to whole data again. So, H2Hadoop is not able to read hole data blocks from HDFS, which is responsible for reducing CPU time and reduces number of read operation. By adding such pre-processing techniques and features it will process the job efficiently, which gives enhancement to Hadoop MapReduce framework.

E. MapReduce Scheduling Algorithm

Now a day's it is very hard to process large amount of data, such as Facebook daily creates and processes 15 TB of data which is very critical and important to handle. Also Akamai is going to capture and processes nearly about 75 million events each and every day. By considering such requirements in mind, MapReduce technique is introduced which is used to process and handle such large amount of distributed data. Also there is MapReduce Scheduler which is used to acquire some requirements while processing BigData such as user priority, data locality, job performance, network decongestion, resources utilization, energy efficiency and reliability. Here MapReduce Scheduling Algorithm is introduced which is used to meet some quality requirements and also manages the large clusters of hardware.

MapReduce Scheduling Algorithm is used to schedule some set of entities such as tasks, users and jobs. It is also used to acquire runtime environments data, workload characteristics and resources required to schedule in datasets. In this survey MapReduce Scheduling Algorithm is presented for primary quality requirements addressing such as throughput, fairness, response time, availability, energy efficiency, data locality, resource utilization and security aspects. MapReduce Scheduling Algorithm can follow static rules in which tasks, jobs and users are get scheduled at runtime is fixed. Also it can follow dynamic conditions at runtime which are based on some properties like resource, data, workload and job which get decided to allocate task and also decide optimal task allocation.

II. LITERATURE SURVEY

M. Ming *et al.* have discussed the method to process the information in bioinformatics which is called as sequence alignment. Also it has great significance for finding the structure and the function of protein sequences and nucleic acids and the information of evolution. They also describe the issues of sequence alignment and the most common local sequence alignment algorithms called Blast algorithm. Blast algorithm is not meet actual demand for the flood of biological data which is provided by stand-alone. They develop Blast-Parallel algorithm by introducing some additional features as compared to Hadoop-Blast algorithm. Which may results in translation of sequential data into meaningful information and calculates the degree of similarities among multiple sequences [1].

E. E. Schadt *et al.* describes that hundreds of gigabases of data is generated in a week with efficient cost such as DNA and RNA sequencing data. For such semi structured data efficient targeting and scheduling of resources is carried out to solve complex problems. DNA sequence alignment requires large and complex amounts of data processing and computational requirements which provides low-cost, high-throughput rate of data generation. Success in this is depends on ability of providing computational solutions to large-scale data management and analysis large-scale, high-dimensional of datasets. Also discuss that how different kinds of computational environments get mastered and solves BigData problems successfully [2].

M. Chen *et al.* present general background of BigData and technologies, such as could computing, Internet of Things, data centers, and Hadoop. And focus on the four phases of BigData, which are data analysis, data generation, data acquisition, and data storage. Finally represent the several applications of big data, which including enterprise management, Internet of Things, online social networks, media applications, collective intelligence, and smart grid. Also the 4 V's of BigData are introduced as first, Volume of the data, which gives data size in terms of Zetabyte. Second, Velocity of data, which calculate speed of data when it is generated. Third, Varity of the data, which means the form of data which is sequential data, numeric data or binary data. Fourth, Veracity of the data, which determines that how clear the data is [3].

T. White describes the process of development and maintenance of scalable, reliable, distributed systems and defines Hadoop is an Apache open-source software framework which is written in Java which provides solutions for BigData processing and analysis. Also discuss Hadoop Distributed File System (HDFS) which store large datasets also provides an interface between the user's applications and the local file system and also reliable for sharing of the resources for efficient data analysis. Datasets are analyzing with Hive also structured and semi-structured data is provided by HBase. Distributed systems are building by ZooKeeper which is used to design, build, and administer a Hadoop cluster. Also describes an approach called Write-once and read-many that permits data files to be written only once in HDFS and then allows it to be read many times over with respect to the numbers of assigned jobs [4].

 DOI : 10.18231/2454-9150.2017.0008

M. Hammoud *et al.* have discussed that to process and handle large amounts of data, Hadoop gives proper solutions. By considering principle of moving computation towards data gives proper solution than moving large data blocks towards computation and also Hadoop uses HDFS to store large data files in cluster. To process BigData in efficient manner MapReduce provides programming model, in which programmer write functional-style code which gets divided into multiple map and/or reduce tasks. Google create MapReduce framework which is used to process 20 petabytes of data per day and gives efficient solution and hides architectural details which is fault tolerance procedure. Also Amazon develops a new framework, to analyze vast amount of data which is a cost effective process called as Amazon Elastic MapReduce. To reduce network traffic and improve performance, Hadoop schedules map tasks of its input, which gives power to cloud, vendors like Facebook, Google, Microsoft, and Yahoo. By transmitting data repeatedly towards nodes is leverage the bottleneck problem, so the traffic in network is reduced by making Hadoop reduce task aware of partition of network locations which also called as Locality-Aware Reduce Task Scheduler (LARTS) [5].

J. Dean *et al.* designed a MapReduce framework which performs simplified data processing on large clusters but hides the details of parallelization, load balancing, data distribution and fault-tolerance. Such data processing is done by the map and reduces functions and many other functional languages. MapReduce is called as programming model which provides stream reading access, runs tasks on a cluster of nodes, and provides a data managing system for a distributed data storage system. Storing data in HDFS has different forms such as <Key, Value> concept to determine the given parameter (Key) and to retrieve the required result (Value) when the job is completed. Users provides a map function which provides a set of intermediate key/value pairs by processing key/value pair, and a reduce function that insert all intermediate values. Run-time is used for partition of input data, also used to handle machine failure problems, manages inter-machine communication, and scheduling of programs without any experience with parallel and distributed systems. Large number of clusters required for MapReduce which processes many terabytes of data on thousands of machines. By using hundreds of MapReduce programs which also consist of thousand MapReduce jobs which get executed on Google's clusters every day [6].

S. Wu *et al.* describes that MapReduce has become relevant tool used to analyze large data. Some vendors such as Massively Parallel Processing (MPP) which is a data warehouse system now effectively use MapReduce into their system which provides two user-friendly interfaces such as map and reduce. MapReduce face major problem with SQL, which hides some implementation details required for user. Pig and Hive are high-level languages which overcome such problem and resemble SQL which shows details to user while working on it. Also the given query is transformed into a set of MapReduce jobs. HDFS consist of "Write-Once and Read-Many" model to store data in distributed DataNodes which means data is written only one time in DataNode and get read multiple times. NameNode is responsible to present list of DataNodes to assign job which is received from the client and update such list when a DataNode is not present due to failure in hardware or network issues. Hive is a MapReduce-based query processing system used for query optimization and translates the query into a set of MapReduce jobs sentence by sentence when it is an SQL query. By combining MapReduce and data warehouse system achieve better performance is which require less time for query optimization [7].

G. S. Sadasivam *et al.* have discussed about alignment which provides arrangement between two or more sequences of nucleotides or amino acids which leads to maximize the similarities among the sequences. Alignments has two approaches such as global and local in which global method is more reliable than local method, which gives maximum number of matched residues by using entire sequences. Multiple Sequence Alignment (MSA) is consist of three or more sequences whereas Pair wise alignment uses two sequences. Once a sequence has been aligned to MSA it will not modify again but it is a fast and straightforward method which aligns the closest sequences first. They also describes Dynamic programming algorithms like Needleman-Wunsch and Smith-Waterman which not works on small number of sequences but provides accurate alignments and require high computational power. If the size of alignment sequences is get increased then complexity of these algorithms also get increased exponentially. Scalable parallel processing Hadoop framework has been discussed and implemented for the sequence alignment of genomic data which produces quality alignment in time efficient manner. To compute parallelism of data Hadoop data grids is introduced which is used to improve the speed and accuracy of sequence alignment [8].

H. Alshammari *et al.* describes about explosion of biological data is made by improvement in high throughput of data generation tools and large scale of genomic research. The challenges faced by data driven biomedical industries due to analysis of such large non relational, distributed datasets require large and complex amounts of data processing and it can scale up to few terabytes of data. Hadoop based cloud

 DOI : 10.18231/2454-9150.2017.0008

architecture consists of Hadoop Distributed File System (HDFS), MapReduce programming model and Apache Zookeeper coordination service as shown in figure 1. Whereas ZooKeeper provides coordination and messaging across applications and also require some capabilities such as distributed synchronization, naming, and group services. By utilizing several number of servers and Hadoop Apache Map-Reduce frameworks, HDFS distributed computing model provides genomic data to align over the framework. By building a MapReduce program it will find a chromosome and DNA sequence by making pattern sequence as a key. It is found that to search human genome much faster than single node with reduction in search time, which requires DataNodes get increased in the cluster. In current Hadoop architecture, NameNode contains location of data blocks in HDFS and directly assign jobs to the clients without going through all DataNodes [9].

M. N. Vora presents Hadoop-HBase database which is used to store large scale of data. To manage and process digital data some database management tools face difficulty to store and analyze such data. Here the goal is to evaluate the performance of random reads and writes of data storage location information which also changing the type of data found in databases. To overcome this difficulty researchers develop HBase which is called as Apache open source software. HBase is also one kind of NoSQL databases which works on top of Hadoop Distributed File System (HDFS). HBase also used as indexing table and provides solution which works successfully and also by creating a key-value data structure. HBase is a column-oriented database which expanded horizontally. HDFS contains the non-textual data like images and location of such data which is stored in HBase for evaluation of hybrid architecture which enables faster search and retrieval of the data which is need of any organization who are flooded with data [10].

III. CONCLUSION

Enhanced Hadoop framework also known as H2Hadoop, which provides access to NameNode which further identify the blocks which consist of certain information in the cluster. In H2hadoop less data is get read, so some Hadoop factors such as number of read operations, which get reduced by the number of DataNodes. DataNodes are carrying the source data blocks which require reading data from dataset. MapReduce scheduling algorithm is used in managing large clusters of hardware nodes. Also meets multiple quality requirements by managing and distributing users, jobs, and tasks. It is also expected to improve the existing system and its performance in which existing system reduces the data transfer within the network and reduces the cost of execution of the MapReduce job as the number of active DataNodes during the action of a job reduces.

REFRENCES

[1] M. Ming, G. Jing, and C. Jun-jie, "Blast-Parallel: The parallelizing implementation of sequence alignment algorithm based on Hadoop platform,'' in *6th International Conference on Biomedical Engineering and informatics (BMEI)*, pp. 465-470, 2013.

[2] E.E. Schadt, M. D. Linderman, J. Sorenson, L. Lee, and G. P. Nolan, "Computational solutions to large-scale data management and analysis,'' *Nature Reviews Genetics*, vol. 11, pp. 647-657, 2010.

[3] M. Chen, S. Mao, and Y. Liu, "Big Data: A Survey,'' *Mobile Networks and Applications*, vol. 19, pp. 171-209, 2014.

[4] T. White, "Hadoop: The definitive guide,'' *O'Reilly Media, Inc.*, 2012.

[5] M. Hammoud and M. F. Sakr, "Locality-Aware Reduce Task Scheduling for MapReduce,'' in *IEEE Third International Conference on Cloud Computing Technology and Science (CloudCom)*, pp. 570-576, 2011.

[6] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters,'' *Communication of the ACM*, vol. 51, pp. 107-113, 2008.

[7] S. Wu, F. Li, S. Mehrotra, and B. C. Ooi, "Query optimization for massively parallel data processing,'' in *Proceedings of the 2nd ACM Symposium on Cloud Computing*, pp. 12-23, 2011.

[8] G. S. Sadasivam and G. Baktavatchalam, "A noval approach to multiple sequence alignment using Hadoop data grids,'' in *Proceeding of the 2010 Workshop on Massive Data Analytics on the Cloud*, pp. 1-7, 2010.

[9] H. Alshammari, H. Bajwa, and J. Lee, "Hadoop Based Enhanced Cloud Architecture,'' in *ASEE, USA*, 2014.

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

 DOI : 10.18231/2454-9150.2017.0008