ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

A Study of Big Data Mining Techniques Used to Improve the Performance of Large Data Processing

G.Suganya MCA.,

Guest Lecturer, Department of Computer Applications, Sri Akilandeswari Womens College, Wandiwash, India

ABSRTACT: Data is the collection of values and variables that are stored in a particular place for further usage using some data base management systems. In traditional small amount of data only stored and retrieved by the users. Nowadays the vast development of networking, data collection and data storage may lead to increase in size of data called as big data. Big data is the large volume of data that comes from various sources. It is a collection of larger data that cannot be processed using traditional database management systems computing techniques. Big data volume measured in terms of terabytes or petabytes. To increase the performance of data storage, retrieval and processing some of the techniques those are used to improve the performance in large collection of structured and unstructured data sets.

KEYWORDS- Big Data Mining-performance-technologies-Hadoop

I .INTRODUCTION

In order to analyze the large volume of data sets arranged in a distributed manner that can be securely store, manage and share large amount of complex data. Big data is a heterogeneous collection of both structured and unstructured data. Businesses are mainly concerned with managing unstructured data. Big data mining is the capability of extracting useful information from these large datasets or streams of data which were not possible before due to its volume, variety, and velocity[21]. Enormous amount of data are generated every minute. Increase of storage capacities, Increase of processing power and availability of data are the main reason for the appearance and growth of big data. Big data refers to the use of large data sets to handle the collection or reporting of data that serves businesses or other recipients in decision making[21].

The rapid development of Internet and mobile technologies has an important role in the growth of data creation and storage. Since the amount of data is growing exponentially, improved analysis of large data sets is required to extract information that best matches user interests. New technologies are required to store unstructured large data sets and processing methods such as Hadoop and Map Reduce have greater importance in big data analysis[21].

With the development of information technology, amount of data are collected from various sensors and devices in multiple formats. Such data processed by independent or connected applications will routinely cross the peta-scale threshold, which would in turn increase the computational requirements. With the fast increase and update of big data in real-life applications, it brings a new challenge to quickly acquire the useful information with big data mining techniques. Huge amount of data is generated from social networking sites, e-commerce, on-line banking, weather stations, market transactions etc.

A. Big Data

Big Data refers to collection of huge data sets with a great diversity in types so that it becomes difficult to process by state-of-the-art data processing approaches or platforms . More generally, we can say that it is formidable to

DOI: 10.15680/IJIRCCE.2017. 0507029

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

perform capture, preparation, analysis and visualization on big data by current technologies. Therefore, big data introduces new challenges for security, processing and analytics such as quick and up-to-date responses of a search query and in-time availability of data . For instance, data obtained from sensor[23] networks like urban management, environment and industrial installation introduces storing, cleansing, query execution and other challenges like security, visualization and analytics .

Similarly in the field of body sensor networks, increasing costs of healthcare and ageing of population are major subjects for re-searchers which have critical information retrieval requirements. For these time sensitive applications, efficiency in query execution and analyzing data is very important for faster decision making . Need of fast data processing and timely responses derive to evaluate the performance of search process so that challenges revealed by the emergence of big data can be highlighted. Indexing is a significant activity even for distributed highly available big data sets to efficiently per-form data retrieval operations . [23]It is impractical to apply full scan on millions of records to accomplish search of a specific result .

Therefore, efficient techniques are required to improve task execution for accessing big data. To improve the efficiency of search and data retrieval process for voluminous data records many solutions have been proposed by re-searchers[23]. For example, vertical partitioning, clustered attribute based indexing for distributed parallel processing systems and clustered adaptive indexing for changing query workload.

Likewise in medical research, large distributed image data sets face the problem of multi-query optimization and a batch processing based image retrieval system contributes in scheduling multiple query requests and minimized response time is achieved. Consequently for distributed and replicated big data storage systems, an efficient indexing technique is needed to serve more number of queries for improved search performance[23].

II.BIG DATA MINING TECHNIQUES

In recent trends there are tremendous amount of data can be processed. But the processing performance may slow their searching process because of large unstructured data seta are available .so, it is not easy to retrieve the needed information or data within a fraction of second. It requires some more time to retrieve the relevant data in the stored data sets. In this paper, we proposed some of the technologies that are used to increase the performance of data processing in big data mining.

A. HADOOP

HADOOP is an open source java framework technology that is used to store, manage and distribute big data across several server nodes. It is a highly archive distributed object oriented programming. This technology can be created by Goug Cutting and Mike Caferella in 2005 for supporting a distributed search engine project. It helps to store access and gain large resources from big data in a distributed fashion at less cost, high degree of fault tolerance, high scalability[6].

Apache Hadoop is an open source software framework written in java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures (of individual machines or racks of machines) are commonplace and thus should be automatically handled in software by the framework. The core of Apache Hadoop consists of a storage part (Hadoop Distributed File System (HDFS)) and a processing part(Map Reduce)[5]. Hadoop splits files into large blocks and distributes them amongst the nodes in the cluster. To process the data, Hadoop Map Reduce transfers packaged code for nodes to process in parallel, based on the data each node needs to process. This approach takes advantage of data locality nodes manipulating the data that they have on hand to allow the data to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture that relies on a parallel file system where computation and data are connected via high-speed networking.

B. HADOOP DISTRIBUTED FILE SYSTEM (HDFS)

The Hadoop File System (HDFS) offers the efficiency to store huge amounts of unstructured data in a reliable way on commodity hardware. In spite of there are file systems with preferable performance, HDFS is an integral part of

DOI: 10.15680/IJIRCCE.2017. 0507029

the Hadoop framework and has already reached the level of a beneficial standard. It has been designed for huge data files and is well suited for hastily ingesting data and agglomeration processing[5,6].

The Hadoop distributed file system (HDFS) is a distributed, scalable, and portable file-system written in Java for the Hadoop framework. Each node in a Hadoop instance typically has a single namenode, and a cluster of data nodes form the HDFS cluster[5]. The situation is typical because each node does not require a data node to be present. Each data node serves up blocks of data over the network using a block protocol specific to HDFS.

C. Map Reduce

MapReduce is a an emerging programming paradigm which is designed for processing extremely large volumes of data in parallel mode by splitting the job into various independent tasks.Map() function and a Reduce() function. The job of Map() is to perform filtering and sorting operations as such, sorting customers by first name into queues[12], by generating one queue for each name and the Reduce() performs a summary/aggregate operations like counting the number of customers in each queue, thereby yielding the name counts fault tolerance.

A MapReduce program in general is a combination of a The "MapReduce System" well known as MapReduce "framework" or “architecture” demonstrates the processing with the distributed servers, running the various tasks in parallel, managing all communications and data transfers between the various parts of the system[12], and providing for redundant data and MapReduce is a framework for processing voluminous data splitted and distributed across huge datasets using a large number of computers (nodes).

The group of nodes collectively treated as a cluster, if all nodes are with similar hardware configurations working on the same local network or else the nodes are treated as a grid, if they are geographically shared and distributed with varying hardware specifications. Processing may occur on the data that is stored either in system log files (unstructured) or in a database (structured). MapReduce takes advantage of locality of data, to minimize the data transfer distance[12].

III. LITERATURE REVIEW

Table 1: Summary of Big data mining technologies that are used to increase the performance of data processing

Techniques	Explanation
Hadoop	**It is an open source technology used to store, access large amount of data in distributed fashion at high degree of scalability and fault tolerance. Connecting single server to thousands of machines with data stored in local (local storage) and computation.[1,5]**
MapReduce	**Designed to process large volume of data in parallel mode. In that we have two functions in separate map() and Reduce().Map()-perform filter and sort operations. Reduce()-Perform aggregation operation. This technology may join with hadoop to increase their performance in next level. It is used to boost the performance.[12,1,5]**
Hadoop Distributed File System(HDFS)	**It is a java based file system. That can be used to provide high throughput access to application data. This may contain Master/Slave Structure. If the data are arranged in an distributed manner then it is easy to access the relevant/needed data without delay.[2,11]**
XHAMI(Extended HDFS and Mapreduce Interface)	**It is used for large scale image processing in mining. XHAMI developed with the combination of HDFS and**

	Mapreduce. XHAMI offers extended library of HDFS and MapReduce to process the single large-scale images with high-level of abstraction over writing and reading the images. XHAMI makes an interface big data image processing in cloud environment.[20]
ST-Hadoop	It is a MapReduce framework that can be used to access big spatio-temporal data. It is a new idea of indexing, where the data are temporarily loaded and divided across computational nodes.[19]
Hybrid Cloud Bursting	Reuse large amount of input data at each iteration. It enables temporary boosting of on-premise resources with additional off-premise resources from public cloud provider for duration of peak usage.
Polyglot persistence	It may handle large number of databases at multiple platforms on same time. Working with multiple nodes at the same time may happen by creating hadoop Master and slave structure. Its main advantage is higher response time and reduced complexity.[17]
Frequent Item set Mining	It is an approach that can be used for optimizing the performance of large scale frequent item set mining. FP growth is an efficient and scalable algorithm to find frequent item sets use novel data structure.[15]
Mobile Distributed File System(MDFS)	It works in mobile cloud. It is capable for big data analytics of unstructured data like media files, text and sensor data. The implementation of MDFs is used to process large amount of data in mobile clouds. MDFS performance result look very promising for real world.[14]
Spark	It is used on top of HDFS, and assurance speeds up to 100 times faster than the two step mapreduce function .That allows data to load in memory and queried frequently, making it especially good for Machine Learning Algorithm.
Apache spark	It is a open source cluster Computing framework. It is emerged as next generation big data processing engine. It maintain Mapreduce linear scalability and fault tolerance.
G-Hadoop	Extension of Hadoop MapReduce framework. It allows MapReduce task to run on multiple clusters. It is simply reuse the user authentication and job submission mechanism of hadoop.[7]
GPHadoop	The semi automatic construction of a crowd-sourced gazetteer can be facilitated by using high performance computing resources .Because it involves the process of mining large volume of geospatial data.
HIPI	HIPI is used for storing and processing big biometric image data .Every day terabytes of image is used in social media. HIPI image bundle may form a hadoop cluster for easier retrieval of big image data.[10]

Hive using Sqoop	**It is used to increase the performance and optimization in hadoop. Very fast it loading the huge volume of relational data into big data warehouse. Fast query processing by increasing the throughput.[10]**
DUSR(Distributed Ultrafast Shape Recognization)	**DUSR is used to identify similar shaped ligand molecules. It showed better performance as compared to the previously used method .It will show structural similarities.[1]**
Stocator	**High performance object store connector for spark. Stocator take advantage of object store semantics to achieve both high performance and fault tolerance. It eliminates rename paradigm by writing each output object to the object's file name. Where it is as much as 18 times faster compared to other data processing techniques.[9]**

IV. CONCLUSION

Big data contain large set of data that are from business, retail industry, marketing and some other social networks. So, every day the volume of data may increase from terabytes to petabytes. Increased amount of storage may lead to reduce data access performance. To avoid this type of problem in future some of the techniques used to increase the performance of data storage and retrieval in a fast and efficient way. In this paper we proposed some of the techniques that may increase the performance in any time and in any situation. This may process not only text but also the other types of files format in an advanced manner. In future a new technique can be generated to solve all those problems in big data mining in an efficient manner.

REFERENCES

1. Vandana Kumari, Rashmi Tripathi, Sunil Patel, Utkarsh Raj, Pritish Kumar Varadwaj:"DUSR(Distribute Ultrafast Shape Recognition): a Hadoop Based Tool to Identify Similar Shaped Ligand Molecules"Indian Journal of Pharmaceutical Education and Research | Vol 51 | Issue 1 |Jan-Mar, 2017.

2. Mrs. Bareen Shaikh, Mrs. Kavita Shinde, Mrs. Sangeeta Borde:" Challenges of Big Data Processing and Scheduling of Processes Using Various Hadoop Schedulers: A Survey" international journal of multifaceted and multilingual studies ,volume-iii, issue-xii issn (online): 2350-0476

3. Song Gao , Linna Li , Wenwen Li , Krzysztof Janowicz , Yue Zhang :" Constructing Gazetteers from Volunteered Big Geo-Data based on Hadoop", Feb 2014

4. Alberto Fernández1 · Sara del Río1 · Nitesh V. Chawla2,3 · Francisco Herrera1:" An insight into imbalanced Big Data classification: outcomes and challenges" DOI 10.1007/s40747-017-0037-9

5. Ch. Shobha Rani , Dr. B. Rama :" MapReduce with Hadoop for Simplified Analysis of Big Data" International Journal of Advanced Research in Computer Science Volume 8, No. 5, May-June 2017 ISSN No. 0976-5697

6. Sudhanshu Shekhar Bisoyi, Pragnyaban Mishra and S. N. Mishra:" Relational Query Optimization Technique using Space Efficient File Formats of Hadoop for the Big Data Warehouse System" Indian Journal of Science and Technology, Vol 10(19), DOI: 10.17485/ijst/2017/v10i19/108088, May 2017 ISSN (Online) : 0974-5645

7.JiaqiZhaoa,LizheWangb,JieTaoc,JinjunChend,WeiyeSunc,RajivRanjane,JoannaKołodziejf, Achim Streitc, Dimitrios Georga kopoulose:" A security frame working Hadoop for Bigdata computing across distributed Cloud data centers" Journal of Computer and System Sciences Journal of Computer and System Sciences 80 (2014) 994–1007

8. Pranati Paidipati1, Dimple Menda2, Radhika Gotmare3: "A Smart and Unified Model for Online Giving: A Complete Study" International Journal of Advance Research and Innovation (ISSN 2347 – 3258) 46th ISTE Annual National Convention & National Conference 2017

9. Gil Vernik, Michael Factor, Elliot K., Pietro Michiardi, Francesco Pace:" Stocator: A High Performance Object Store Connector for Spark" ACM ISBN 978-1-4503-5035-8/17/05.

10. Mohd Ahmed Abdul Mannan, Gulabchand K. Gupta :" HIPI Based Biometric probe image retrieval using Big Image data" International Journal of Advanced Research in Computer Science Volume 8, No. 3, March – April 2017 ISSN No. 0976-5697

11. Jian-bin Hu and Peng Wang: "The Research on Optimization Method of Hadoop MapReduce" 2017 Asia-Pacific Engineering and Technology Conference (APETC 2017), ISBN: 978-1-60595-443-1

12. Nikhat Akhtar1, Firoj Parwej, Yusuf Perwej :" A Perusal of Big Data Classification and Hadoop Technology" International Transaction of Electrical and Computer Engineers System, 2017, Vol. 4, No. 1, 26-38

13. JiaqiZhaoa,LizheWangb,JieTaoc,JinjunChend,WeiyeSunc,RajivRanjane,Joanna Kołodziejf, Achim Streitc,Dimitrios Georga kopoulose :" A security frameworking Hadoop for bigdata computing across distributed Cloud data centres"

14. Johnu George, Chien-An Chen, Radu Stoleru," Hadoop MapReduce for Mobile Clouds" ieee Transactions on cloud computing, vol. 3, no. 1, january 2014

15. Guru Prasad M S, Nagesh H R and Swathi Prabhu:" High Performance Computation of Big Data: Performance Optimization Approach towards a Parallel Frequent Item Set Mining Algorithm for Transaction Data based on Hadoop MapReduce Framework" I.J. Intelligent Systems and Applications, 2017

16. Varsha B.Bobade:" Survey Paper on Big Data and Hadoop" International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 03 Issue: 01 | Jan-2016 www.irjet.net p-ISSN: 2395-0072

17. Ms. Namrata Rawal1, Ms. Vatika Sharma2:" Polyglot Persistence on Oracle Cloud using Hadoop Map Reduce" International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056 Volume: 04 Issue: 04 | Apr -2017 www.irjet.net p-ISSN: 2395-0072

18. Jens Dittrich Jorge Arnulfo Quian ´eRuiz:" Efficient Big Data Processing in Hadoop MapReduce"

19. Louai Alarabi:" ST-Hadoop: A MapReduce Framework for Big Spatio-temporal Data" SIGMOD'17 Student Research Competition May 14-19 2017, Chicago, IL,USA

20. Raghavendra Kune1, Pramod Kumar Konugurthi1, Arun Agarwal,Raghavendra Rao Chillarige and Rajkumar Buyya:" XHAMI – extended HDFS and MapReduce interface for Big Data image processing applications in cloud computing environments" Published online 15 July 2016 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/spe.2525

21.Durgadevi.D and Yamini.G:"A New Approach For Rectifying Erroneous data in Web Usage Mining Using Preprocessing",IJSART Volume 1,Issue 8, August 2015,ISSN[online]:2395-1052

22. Jaseena k.u.1 and julie m. David2: "issues, challenges, and solutions:"big data mining" computer science & information technology (cs & it).

23. T. Jhansi rani :" intrusion detection using big data analysis" e-issn no : 2554-9916 | volume : 3 | issue : 4 | apr 2017

24. Aisha Siddiqaa,, Ahmad Karimb , Tanzila Sabac and Victor Changd:"On the analysis of big data indexing execution strategies".

25. Francisco J. Clemente-Castell´o, Bogdan Nicolaey, M. Mustafa Rafiquey,Rafael Mayo, Juan Carlos Fern´andez_"Evaluation of Data Locality Strategies for Hybrid Cloud Bursting of Iterative MapReduce" HAL Id: hal-01469991

BIOGRAPHY

Suganya Gajendiran is a Guest Lecturer in the Department of Computer Applications, Sri Akilandeswari Women's College, Wandiwash, Thiruvalluvar University. She received Master of Computer Application (MCA) degree in 2015 from Adhiparasakthi Engineering College, Melmaruvathur, Anna University . Her research interests are Computer Networks (wireless Networks), BIG DATA, Algorithms, etc.