IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.115192

Research on Techniques for Resolving Big Data Issues

Dhanashri D. Shukla[1], Vijay M. Rakhade[2], Lowlesh N. Yadav[3]

Student, Computer Science & Engineering, Shri Sai College of Engineering & Technology Bhadrawati, Chandrapur, India[1]

Assistant Professor, Computer Science & Engineering, Shri Sai College of Engineering & Technology Bhadrawati, Chandrapur, India[2]

Head of Department, Computer Science & Engineering, Shri Sai College of Engineering & Technology Bhadrawati, Chandrapur, India[3]

Abstract: Big data and its analysis are within the focus of current era of massive data. The most production sources of huge data are social media like Facebook, twitter, emails, mobile applications and also the migration of manual to automatic of virtually every entity. Currently, there's a requirement to research and process complex and big sets of information-rich data in all told fields. This paper provides a survey of massive data issues and also the effectual and efficient platforms and technologies which are needed to deal and process the remarkable amount of knowledge. It turns around two major areas namely: clustering and scheduling.

Keywords: Include Big Data Issues, Clustering, Scheduling, Analysing Data.

I. INTRODUCTION

Big data is larger amount of information required new technologies and architectures so it'll become possible to increase values from its capturing and analysis process. Due to such large size of information, it becomes very difficult to control effective analysis using traditional techniques. Big data due to its numerous properties like volume, velocity, variety, variability, complexity and value place forward many provocations. Therefore, recent upcoming technology in market is big data which brings massive profit to the business corporations and it becomes important that several challenges and issues associated in bringing and adapting to the current technology are brought into light. This paper discloses the large data technology together with its importance within the modern times and existing projects which are effective and important in changing the speculation of science into scientific research.

To mark this issues, big data analytics is there to untapped statistics from large volume and sort of data. There are some suggested platforms to manage the problems of huge data like Dryad, Spark, Dremel and Pregel, storm and Hadoop MapReduce. MapReduce is one in all the foremost successful frameworks. Initially it had been suggested by Google. Especially it had been designed for processing big data by exploiting the parallelism among a cluster of machines.

The main goal of this paper is to supply a extensive survey about the massive data problems with big data clusters. The processing techniques are discussed thoroughly, moreover, this paper provides the comparison for the algorithms of the processing technique according per the available resources.

The organization of the paper is as follows: Various sorts of big data issues is discussed first. Then platforms and techniques for giant processing are enlightened. The platform and therefore the big processing techniques are given in a descriptive manner with its further types. Additionally, main results and comparison also are given in styles of tables.

Data Storage Issues: The quantity of information has break out each time and thus need for invention of new storage method. For controlling large volume storage, big data storage companies such as IBM, EMC Amazon utilizing the tools like Apache Drill, SAMOA, NoSQL, IKANOW, Hadoop and Horton Works.

Data Management Issues: The data generation sources are different and thus the data also both by means of format and in items of collection. People contribute digital data in a way which are affordable for them like archives, illustrations, images, audio and video messages. Thus, the collected data is accessible for inquiry and examination. Moreover, information and its provenance will become a serious matter. As Suggested by Gartner Big Data challenges includes more than just handling volumes of data mentioned in this article.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.115192

Security Issues: It is challenging to manage a huge data set in the secure means. More, public and private database and inefficient tools comprise many threats. The security matters occur for distributed systems when huge measures of private information put away in database which is not properly encoded and encrypted.

Techniques to Deal with Big Data: Analysis of big data becomes the important point of current modern time. The quantity of data to be investigated increases on one side. The demand for acceptable time to yield outputs is shrinking on the other hand. There is need of efficient techniques and platforms to store, process and analyse the complex and gigantic sets of information-rich data in all fields now a days.

Big Data Clustering: Big Data clustering is technique for analysis and facilitating big scale data manage, exploration and processing of huge of big data. The clustering procedure consist of dividing up the un-label data entries in different sets. Big data clustering has some techniques:

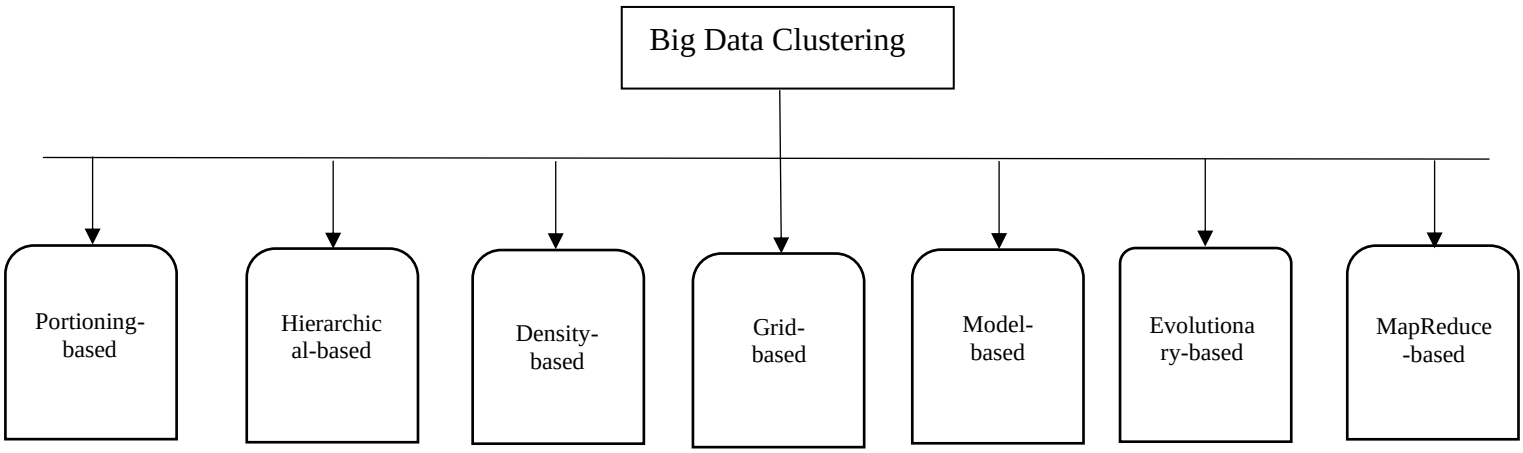

Figure 1: Big Data Clustering Types

Scheduling: Scheduling is one of the processing schemes which deals with the ordering, prioritization and assignment of works to the suitable machines for results. Scheduling is done through ordering the works to the relevant machines is a particular order. To get an optimal schedule for homogenous or heterogenous environments is define as a NP-complete problem.

Scheduling Algorithms for homogenous systems: Many experimenters present scheduling algorithms by incorporating various issues as first in first out scheduler was suggested to take a shot at the possibility of initial things out serve premise. In this scheduler, all employments are given to a solitary queue when a task tracker's pulse just started to the scheduler just gets a job from the head of queue and helps to appoint the undertakings of the activity. The upside of the FIFO scheduler is that the algorithm is quite simple the heap of the job tracker is generally low so it can easily bring about starvation when the little jobs come after the extensive jobs.

Fair scheduler was suggested to make certain the fair distribution of cluster resources such that all jobs to every pool in a way to maintain fairness of the scheduler are the different types of job will have differed sources assigned by the cluster hence, it is going to intensify the high-quality of services. The negative end is that it does not keep in mind the actual loading state of every node so the load stability in every node may not that precise.

After, capacity scheduler stepped forward the feature of HOD (Hadoop on Demand) and encountered the disadvantages of HOD. Capacity scheduling algorithms make the uses of several queues each queue get its sources in keeping with the computing potential, and the rest sources will be to the ones queues that have not met its usage limits, and allocate them base on the load of its calculation which is greater friendly. In case that all the assets of queue are covered through one job, the scheduler applies restrictions to assets of each user within the queue. The benefit of scheduler provisions job priority and can run in parallel to allocate sources dynamically and hence improves the efficiency. Its shortcomings raised by the requirement of extra records of all the job thus making capacity scheduling algorithm more complex and costly.

After, the scheme named as Delay algorithm was proposed. It was proposed to overcome the disadvantages of FIFO scheduler. The mechanism of the scheduler relies on the idea of fair distribution of resources. Thus, in that way the suggested idea makes a long delay for certain types of jobs. Another problem of the scheduler was that if it does non-local in multiple slots.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.115192

This scheduler examines the job and recognize its great. Plan classifier organizes the job employments. According to the asset use, steady jobs will consider for additionally taking care of a good does not make any over weight to the task tracker. Awful jobs will be disallowed. The scheduler considers CPU utilization, one great job finds in the activity queue, job will be picked by the new expected utility limit.

Furthermore, in a scheduling scheme is proposed via proposed via providing some flexibility to the clients. The algorithms allow the customers that have greater than one requests to adjust the priority. Thus, the allocation of the assets would be greater than one requests to adjust the priority thus the allocation of the assets would be granted in accordance with the requirements and the priority demands. In addition, it gives facility to the clients to cut back the task when the demand is greater. But the suggested mechanism became expansive and works kindly for the homogenous surroundings. The comparison of the discussed algorithms of the discussed algorithms are shown in Table 1.

Table 1- Comparison of Big Data Scheduling Algorithms

Author	Algorithm Name	Advantages	Mode
Apache Hadoop, (2009)	FIFO	Easy to understand and Easy to program fair distribution of resources	Non-Pre-emptive Pre-emptive
Apache Hadoop, (2009)	Fair Capacity	Effective utilization of the reasons by utilizing the idle assets of cluster	Non- pre-emptive
(Nita et al, 2015)	MOMTH	Flexibility to clients for priority setting	Pre-emptive

Scheduling Algorithms Handling Straggling Issues: It's been seen generally, that the performance of huge data computing cluster is degraded sometimes due to the incompletion of 1 or variety of tasks. These slow takes are termed as straggles and therefore the phenomenon of this delay is termed as straggling. The initial Google MapReduce framework just starts to dispatch standby tasks when employment is with reference to finishing point. It's been demonstrated that speculative execution can diminish the activity service time by approximately 44%. After that, longest approximate time to finish (LATE) scheme was proposed within which measures the progress rate for the tasks. The late scheduler was intended to handle this unusual terminal of the task. The proposed algorithm measures the progress rate by completion time of task and providing the backups of a number of the tasks consistent with fractional ratio of running phase. Major flaw of this set of rules became that it really works for the slow tasks and have become unable to interrupt the one reasonably of phases of MapReduce at some stage in its progression.

II. CONCLUSION

Big data problems have bought many changes in the way data is processed and managed over time. Today, data is not just posing challenge in terms of volume but also in terms of its highspeed generation. The data quality and validity vary from source to source and thus are difficult to process. The issue has led to the development of several stream processing engines/platforms by different companies such as Yahoo, Linkedin, etc. besides better performance in terms of latency, stream processing overcomes another shortcoming of batch data processing system that is scaling with high velocity data. Availability of several platforms also resulted in another challenge for user organization in terms of selecting the most appropriate stream processing platforms for their needs. In this article we proposed a taxonomy that facilitated the comparison of different features offers by the stream processing platforms.

ACKNOWLEDGMENT

I am Thankful to the college Shri Sai College of Engineering and Technology, Bhadrawati for providing research environment, tools and technical Support to accomplish this work.

REFERENCES

1. 1.Lowlesh Nandkishor Yadav, "predictive Acknowledgement using TRE system to reduce cost and Bandwidth". IJRECE VOL. 7 ISSUE 1 (JANUARY-MARCH 2019).
2. Mr. Vijay M. Rakhade. "Reducing Routing Distraction in IP Network Using Cross Layer Methodology" ICRTEST VOLUME 5 Issue 1 (21-22 January 2017).
3. Hirali Devendra Wadaskar, "Research on Association Rule Mining Algorithms", IJARCCE Vol. 11, Issue 5, May 2022.

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.115192

4. Chandrakant A. Zade. "Research On Data Mining", IJARCCE Vol. 11, Issue 5, May 2022.
5. Agneeswaran, V.S. (2014). Big data analytics beyond Hadoop: real-time applications with storm, spark and more Hadoop alternatives: FT Press.
6. Ananthanarayan, G., Kandula, S., Greenberg, A.G., Stoica, I., Lu, Y., Saha, B., & Harries, E. (2010). Reining in the Outliers in Map-Reduce clusters using Mantri. Paper presented at the Osdi.
7. Ananthanarayan, Ganesh, Michael Chaien-Chun Hung, Xiaoqi Ren, Ion Stoica, Adam Wierman, and Minlan Yu. (2014) "Grass: Trimming Stragglers in Approximation Analytics."
8. Ananthanarayanan, G., Ghodsi, A., Shenker, S., & Stoica, I. (2013). Effective Straggler Mitigation: Attack of the clones. Paper presented at the NSDI.
9. Chen, S., Sun, Y., Kozat, U.C., Huang, L., Sinha, P., Liang, G., Shroff, N. B. (2014) When queueing meets coding: Optimal-latency data retrieving scheme in storage clouds. arXiv preprint arXiv:1404.6687.
10. Chen, Q., Zhang, D., Guo, M., Deng. Q., & Guo, S. (2010). Samr: A self-adaptive mapReduce scheduling algorithm in heterogenous environment. Paper presented at the Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on.
11. Fu, H., Chen, H., Zhu, Y., & Yu, W. (2017). FARMS: Efficient mapreduce Speculation for failure recovery in short jobs. Parallel computing, 61, 68-82.
12. Gartner. "Gartner Survey Reveals That 73 percent of Organizations Have Invested or Plan to invest in big data in the next two years". (2013) Stamford, Conn Hadoop, A. (2009). Hadoop.In./Fifo. Hadoop, A. (2009).