THINK INDIA JOURNAL

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

Improving Hadoop Job Scheduling Using Ant Genetic Algorithm

Amritpal Singh
Assistant Professor
Lovely Professional University
Amritpal.17673@lpu.co.in

Varun Singla
Assistant Professor
Lovely Professional University
Varun.17705@lpu.co.in

Ravinder Singh
Assistant Professor
Lovely Professional University
Ravinder.17750@lpu.co.in

***Abstract*— Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, creation, manage, and process data within a tolerable elapsed time. Big data has the now great attention in the information industry and in society due to the existence of large amounts of data and the crucial need for changing such data into useful information and knowledge. Generally data is managed by the distributed file system (DFS) and Hadoop DFS called HDFS. But when the data contain some time-series, then it is more efficient to apply some more level of data processing at the data source before storing the data. Using Hadoop the data is stored as well as process simultaneously. Scheduling in Hadoop refers to the distribution of jobs. Hadoop allows the user to configure a job, requesting the job, execution of the job, and control over the job and query the state of the system. Massive job scheduling problem is an important research area in big data research era. The experimental results showed that the nature inspired algorithms can improve efficiency of job scheduling on Hadoop clusters effectively.**

***Index Terms*—: Big data; Hadoop job scheduling; Map reduce; Ant-genetic algorithm; Optimization**

I. INTRODUCTION

Today, there is a large amount of data generated in every sector of the business, science, and our personal life. The speed of growth of data seems to out speed the advanced computer structure[1]. This challenge is known as Big Data. Still people have different definitions of the big data. Big data is a technology that is used for the gathering of the large and complex data sets. Big data defines the both structure and non-structured data. Big data becomes difficult to process using traditional data processing application. Big data arise when we have a problem with the size of data. Any instance of big data might be of any size like petabytes (1024 terabytes) or Exabyte (1024 petabyte) of data consisting of billions of record of people. Today big data "size" is a moving target as its ranging from dozen of terabytes to petabytes. When the organization deals with the large data sets, the organizations faces some problems to generate, operate and control big data. Big data is basically difficulties in computer science as well as business analytics because standard tools or traditional tools and procedures are not compatible to search and analyze the large sets of data. Data mining technique helps to extract the data from the large database for analyze the data to make it clean for the future purpose. The methodology of data mining describes how to handle the large data sets while processing it. Big data also have some issues like the data integrity

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

when the data preservation is more important. Security of the data is the main issue when we talked about the data. So here we got a solution for this by splitting the data into several clusters. Big data helps in better understanding of the customer, revenue and in many other fields like insurance, healthcare, governmental and manufacturing. There is an approach defined for learning the forming of the large data sets from a training data [2]. Job scheduling is a very important area related to the big data research. In this paper authors discussed about the searching of best node for the current jobs and improve the efficiency of the scheduling. For this we can use the FIFO (first in first out) method or fair scheduling (used by the google). Clusters in Hadoop may contain thousands of server and thousands of CPU. So it is a big challenge to process the large amount of jobs over the large cluster. To optimize the job scheduling a nature inspired algorithm is applied to the MapReduce function to improve the matching speed of the current jobs to the best suitable node [3]. The challenges in big data include the analysis, represent, maintain, search, sharing, storage and shifting of the data. The challenge of the volume is related to the size of the data that is facing by the every organization of any fields like science, computer etc. The next aspect or characteristic of the big data is the velocity. This is related to the speed of upcoming new data and the updating the existing data. The third characteristic variety is related to the source of the data or the category of the data [4]. This may be scientific data, web data and transactional data. The main characteristic of the big data is its complexity. The data management becomes a complex process, when it deals with the large data sets from multiple data sources. The use of big data becomes a very urgent way for leading organizations to outperform their peers. Big data is not only directed to the size of the data. It is a moment to find a new perception in new and appearing types of data and quantity to make your business more active. Big data need many technologies to well-organized process large data sets within enough time. These techniques may be data fusion and integration, genetic algorithm, machine learning, natural language processing, and visualization, which further have different methods to manage the big data [5]. The big data storage related to the repository and management of large-scale datasets while attaining reliability and availability of the data accessing. Various storage system arise to meet the demands of massive data. We have two storage technologies that are Direct Attached Storage (DAS) and Network Attached Storage (NAS). Actually, we may have big data but should be actionable. Big data is a technique that is used to maintain context that is missing in the structured database.

Nature Inspired Algorithms

Nature has been always a source of inspiration for many researches. Nature inspired problem solving techniques are widely used to solve complex problems. These techniques are mainly used due to their decentralized and self-organized behavior. Such behavior is observed in social systems such as Evolutionary algorithm, particle swarm optimization (PSO), ant colony optimization etc. Nature-inspired techniques are being used for data clustering in big data, hybridization with traditional clustering techniques and their effectiveness and efficiency [6].

Ant Colony Algorithm

Ant colony algorithm was proposed by Marco Dorigo et al. in 1996. First version of the algorithm was Ant system that was widely used in the travelling salesman problem. ACO algorithm is inspired by the behavior of real ant which is able to find out the shortest path between its nest and the food source. Ants of the colony have the different properties:

i. An ant searches for the minimum cost solution.
ii. An ant k has a memory of $2^{k.}$
iii. An ant starts from the start level and move to the feasible neighbor state.
iv. When ant colony move from one node to another then an ant can also update its pheromone.
v. Once an ant colony finds their target then they can go back by the same path and also update the pheromone.

II. BIG DATA APPLICATIONS

Big data in healthcare**:** The practices and policies of healthcare are different everywhere throughout the world. There is the main purpose related to the healthcare system. The very first goal is to enhance the patient experience (experience contains quality and satisfaction). Another goal is to improve worldwide population health and to make the healthcare cost to its low value and the last one is a traditional method to manage healthcare as well as create modern technology to analyze large amount of information about the patients. Since, it will take long time for every medical staff to collect the large amounts of data or

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

information in healthcare. High-performance analysis of data by new technologies makes large datasets easier to process or change the large quantity of data into appropriate one that is used to provide better care [7].

Big data in network security: When we talk about the network security of any organization then it requires the whole database to use the security mechanism over it. Nowadays Big data is emerging as the empowerment of the landscape of various technologies of the network security. Big data may be used in network monitoring, forensics and SIEM. Big data can also be used to construct a world where we can maintain the control over the confession of our customized information is being challenged statically. Various mobiles and credit card organization have conducted large-scale fraud detection for decades and analysis and processing the large scale data [8]. Big data in market and business: Big Data is the largest Ace-card (game-changer) opportunity in the business of marketing and sales, 20 years ago the Internet work as main stream, because of the unequalled array of perception into customer needs and behavior makes it feasible. But many decision-maker who agree that this is true, they are not sure that how to make the most of it and later they also find themselves faced with huge and inordinate amounts of data and it leads to quickly changing in behavior of the customer's , And it also increase the complexity and the competitive pressure of the organization. If we use real-world examples like additional resources which are downloadable, non-technical language and a good quality of humor will help you to search the treatments offered by data driven marketing. Big data disclose the behavior of the customer's and provide a way to evaluate the feedback and experience of the customer. These perceptions can help to make sure the success of your business [9].

Big data in education and sports: We can use the big data technologies in the education as well as in sports; if we use the big data analytics in the big data then we can have the extraordinary results. Online data of the student's behavior can deliver a good perception to the educators, like if any of the students need some supplementary attention, the topic covered in the class is not clear, or if the course has to be modified. Then Students are commanded to answer guide questions, so before the class they can go through the set of online content related to the topic to be covered in the class. Hence by counting the number of students those have completed their online module, and the time taken by the students to solving those module and the accuracy of the students while solving the problem of the module, a lecturer can have the better information about his student's performance and can improve the lesson plan accordingly [10].

Now when we come to the optimization of big data, as we all know the optimization is a process to make anything fully effective, efficient and fully perfect. So, big data optimization states that we are going to deals with the volume, velocity and variety of big data. These are the main characteristics of the big data which we have to optimize and make our data more efficient and good. The size of data makes it more ubiquitous, to deal with this problem we need the optimization of big data. As we all know big data is not only about the data size, there are some different factors like heterogeneity of data source, levels of noise which increase the complexity of the big data problem. Generally data is managed by the distributed file system (DFS) and Hadoop DFS called HDFS. But when the data contains some time series, then it is more efficient to apply some more level of data processing at the data source before storing the data. Using Hadoop the data are stored as well as the process simultaneously. Scheduling in Hadoop refers to the distribution of jobs. Hadoop allows the user to configure a job, requesting the job, execution of the job and control over the job and query the state of the system. Every jobs have independent tasks and time slot for the execution. Hadoop file system is a type of master-slave architecture where name node acts as the master and data node act as the slaves. Job tracker receive the requests from the job client and then send to the task tracker. Task tracker shuffle the jobs by partitioning them to reduce to the best pair.

III. SCOPE OF STUDY

As the big data processing is a challenging task for every organization and companies. As per the survey, there is a need to develop a method which makes big data fully perfect and efficient. So it is needed to optimize the big data and data clusters to make the data processing easy and make the data secure and it will reduce the cost of maintenance of data and the cost of the network over the databases. Hadoop is a distribution file system and data processing used to handle the large volume of data in any

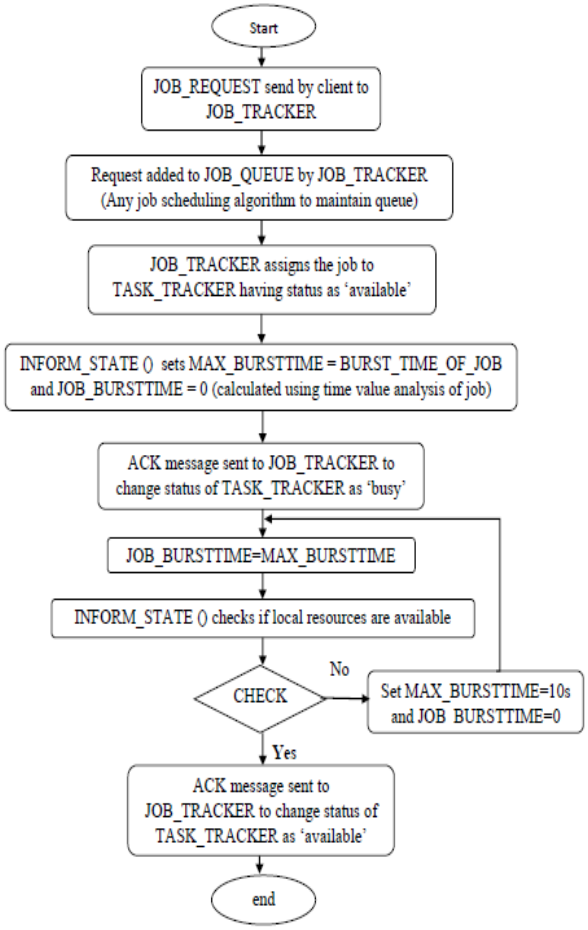
ISSN: 0971-1260
Vol-22-Issue-16-August-2019

structure. Hadoop has two components HDFS (Hadoop distributed file system) and MapReduce. HDFS supports the data in structured relational form or in any unstructured form. While MapReduce is used to process the parallel data processing in distributed servers. Using hybrid Ant-Genetic algorithm Optimization in the job scheduling of the Hadoop makes it more efficient and decrease the overhead of the data nodes.

IV. RESEARCH METHODOLOGY

For the robustness in the Hadoop job scheduling in case the master node crash then snapshot will be taken every 30 minutes. So if the master node crashes then the recent data node or task tracker will be selected as the master node and the rollback is made using previous snapshot [11]. Following are key terms that we have used in proposed work:

- JOB_CLIENT- User or Job Provider.
- JOB_TRACKER- Name node act as a master node in the MapReduce framework. Also responsible for maintenance of the status (idle or busy) of all the task tracker.
- TASK_TRACKER- Data node act as a slave node in the MapReduce framework.
- JOB_QUEUE- Sequences of incoming job requests sent by JOB_CLIENT (User).
- INFORM_STATE () - A function used in tast_tracker to check whether the local server resources are idle or not and inform back to the job_tracker.
- JOB BURST_TIME- Average execution time for different types of jobs in fair scheduling calculated by time analysis.



V. EXPECTED OUTCOMES

Hadoop is a framework for storing data and running applications on clusters of commodity hardware. While storing the data into the large data sets over the distributed clusters. There is need to maintain the complexity (time and space efficiency) in the job scheduling in Hadoop. In this research work we will explore all the factors related to the working of job scheduling in Hadoop to handle the big data. And will also try to enhance the performance of the Hadoop jobs scheduling by using nature inspired algorithm.

VI. SUMMARY AND CONCLUSION

Big data now has great attention in many field related to the social media, Business perspectives, Industrial organization etc. So it is the main issue to make our data and data processing fully perfect so that it is easy to store and process the large data sets simultaneously. To solve this problem

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

Hadoop engine is used for handling the large data sets. Hadoop consists of lots of jobs like production jobs, Ad-hoc jobs and large extension jobs. So by optimizing the jobs scheduling in Hadoop it will be easy to dealing with the data processing over the distributed clusters.

REFERENCES

[1] Ms. VibhavariChavan, Prof. Rajesh. N. Phursule, "Survey Paper on Big Data", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 2014, 7932-7939.

[2] Min Chen, Shiwen Mao, Yunhao Liu, " Big Data: A Survey", © Springer Science and Business Media New York 2014 Mobile Netw Appl (2014) 19:171–209.

[3] Sandeep U Mane and Pankaj G. Gaikwad, "Nature Inspired Techniques for Data Clustering", Circuits, Systems, IEEE Communication and Information Technology Applications (CSCITA), 2014 International Conference on, pp 419-424.

[4] Zewu Peng, Jingliang Liao and YiqiaoCai, "Differential evolution with distributed direction information based mutation operators: an optimization technique for big data", Springer-Verlag Berlin Heidelberg 2015.

[5] Sabia and Sheetalkalra, "Applications of big Data: Current Status and Future Scope", International Journal on Advanced Computer Theory and Engineering (IJACTE)ISSN (Print): 2319-2526, Volume -3, Issue -5, 2014.

[6] Divyakant Agrawal, Sudipto Das and Amr El Abbadi, "Big Data and Cloud computing: Current State and Future Opportunities", Proceedings of the 14th International Conference on Extending Database Technology, pp 530-533.

[7] Xiaoli Cui, Pingfei Zhu, Xin Yang, Keqiu Li and Changqing Ji, "Optimized big data Kmeans clustering using MapReduce" ,Published online: 19 June 2014 © Springer Science+Business Media New York 2014.

[8] Chanchal Yadav,Shuliang Wang and Manoj Kumar, "Algorithm and approaches to handle large Data-A Survey", IJCSN International Journal of Computer Science and Network, Vol 2, Issue 2, 2013.

[9] Prafulkoturwa,SheetalGirase, DebajyotiMukhopadhyay, "A survey of claasification technique in era of big data", IJAFRC, Vol.1, Issue 11, November 2014, ISSN: 2348-4853.

[10] Tapan P Gondaliya , Hiren D joshi, Hardik J joshi, " New Big Things in Era of Digital Data: Big data and big data challenges and its solution using different tools", 10th International CALIBER-2015 HP University and IIAS, Shimla, Himachal Pradesh, India March 12-14,

[11] Xiaofei Huang, Hui Zhou and Wei Wu, "Hadoop Job Scheduling Based on Hybrid Ant-Genetic Algorithm", 2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, 978-1-4673-9200-6/15 © 2015 IEEE.