National Conference on Engineering Innovations and Solutions (NCEIS – 2018)
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 4 | Issue 6 | ISSN : 2456-3307

Big Data Analytics With Business Intelligence: A Survey

Revati. K*[1], Padmashree. V[1], Veeresh Hiremath[1], Vinutha. D. C[2], Chandini. S. B[2]
[1]Student, Department of ISE, VVCE, Mysuru, Karnataka, India
[2]Associate professor, Department of ISE, VVCE, Mysuru, Karnataka, India

ABSTRACT

Everything in today's world stands on data. Usage of many applications is resulting in the generation of several petabytes of data every day. This generated data is very important in order to take business decisions. Thus to analyse this big data business intelligence systems are built. There are many platforms to perform this analysis such as hadoop, spark, orange etc. and also there are many algorithms to perform scheduling such as FCFS, capacity scheduling, priority scheduling, shortest job scheduling etc. The main aim of this paper is to build an efficient business intelligence system which uses a scheduling algorithm called MSHEFT-"Memory sensitive heterogeneous earliest finish time".

I. INTRODUCTION

Big data is huge collection of data. It is the mixture of structured, semi-structured and unstructured data. There will be a variety of data like text files, images, videos, xml files etc. The multi-V (volume, velocity, variety, veracity, and value) model is frequently used to characterize big data processing needs. Volume defines the amount of data, velocity means the rate of data production and processing, variety refers to data types, veracity describes how data can be a trusted function of its source, and value refers to the importance of data relative to a particular context. This data is very important in order to make business decision. Business intelligence systems are built in order to make this possible. Business intelligence and analytics (BI&A) and the related field of big data analytics have become increasingly important in both the academic and the business communities over the past two decades. Business systems will have several constraints on building such as cost, efficiency, maintenance etc. These systems should be very accurate in analyzing the data so as to reduce the risk involved in decision making. We have several platforms which is used for data analysis. The idea of using multiple platforms in BI will definitely increases its efficiency. Scheduling plays and important role in big data optimization, especially in reducing the time for processing. The main goal of scheduling in big data platforms is to plan the processing and completion of as many tasks as possible by handling and changing data in an efficient way with a minimum number of migrations. This paper aims to explain how multi-platform business intelligence system can be built using MSHEFT scheduling algorithm with high availability, high scalability and high capacity.

II. BUSINESS INTELLIGENCE

Business intelligence comprises of strategies and technologies used by the enterprises for the data analysis of business information. BI systems provide historical, current and predictive views of business operations. Common functions of BI include reporting, online analytical processing, analytics, data mining, process mining, complex event processing, business performance management, bench marking, text mining, predictive analytics and prescriptive analytics. BI can handle large amount of structured and unstructured data to help, identify, develop and otherwise create new strategies business

opportunities. All these helps in taking decisions to improve business. BI system should be built in low cost and high efficiency. Thus combining multiple platforms for analysis will help in achieving this goal.

III. BIGDATA ANALYSIS PLATFORMS.

To perform analysis there are many platforms available. Tools such as Hadoop, Spark, Pentaho, Karmasphere studio etc. Most used platforms are Hadoop and Spark. They are efficient, fast and consistent platforms for analysis.

A. Hadoop

Hadoop is an open source framework that is used to process large amount of data in an inexpensive and efficient way and job scheduling is key factor for achieving high performance in big data processing. There are two components of hadoop, HDFS (hadoop distributed file system) and MapReduce. HDFS is used for data storage while MapReduce is used for data processing. MapReduce has two functions: Map and Reduce. The functions are both written by the user and the function take vales as input key value pairs and output the result as a set of key value pairs.

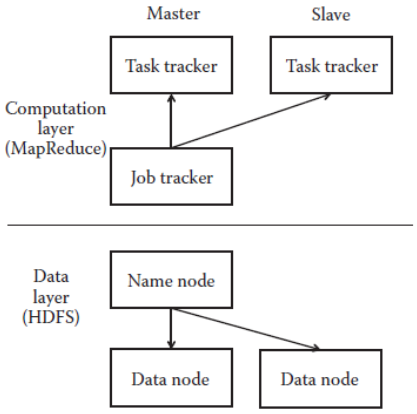

Figure 3.1.1 Hadoop general architecture

B. Apache Spark

Apache spark is an open source cluster computing framework. Spark provides an interface for programming entire clusters with implicit data parallelism and fault tolerance. Spark has Resilient distributed dataset (RDD), a read-only multiset of data items distributed over a cluster of machines. Spark is developed in response to the drawbacks of MapReduce. Spark facilitates implementation of both iterative analysis and interactive data analysis.

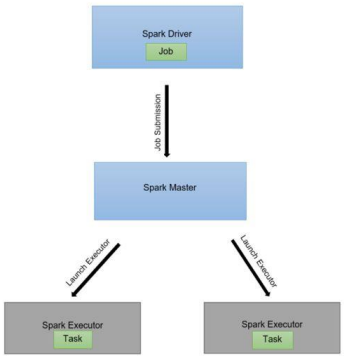

Figure 3.2.1 Spark framework

C. Comparison between hadoop and spark

Table 3.3.1. comparison between hadoop and spark

HADOOP	SPARK
• Slow in computing results. • Difficult to program • Difficult to manage	• 100 xs faster in memory and 10 xs faster in disk. • Spark is easy to program. • Easy to manage.
• Fails to process real-time data.	• Work well with real-time data.
• High latency.	• Low latency.
• More secure	• Less secure.

As it is seen from the table Spark is faster than Hadoop. But when we consider the cost, to build Spark system it costs more when compared to Hadoop. Thus to build BI systems we can use the combination of multiple platforms in order to make it cost effective. An algorithm to schedule the tasks to optimize the speed in such environment should be used.

IV. SCHEDULING ALGORITHMS

General scheduling algorithms are discussed below.

A. **First In First Out (FIFO) Scheduling** FIFO scheduling is based on queue mechanism. So in this first the job is divided into many tasks and then it will be given to those slots which are free and available TaskTracker nodes . Jobs will have to wait for the execution due to acquisition of clusters. Hence the jobs will have to wait till there turn come. All jobs need to complete in a time manner and provide better response time to every job.

B. **Fair Scheduling**

The main objective of this Fair Scheduling is to provide a fair share of cluster capacity over a time. So for every user group jobs in to job pools, there will be a guaranteed minimum number of Map and Reduce slots. It supports preemption i.e. to give the slots to the pool running under capacity, the scheduler forcefully kill tasks in job pools running over capacity. Priority is also assigned to various pools. Facebook develops the Fair Scheduler to manage the Hadoop cluster.

C. **Capacity Scheduling**

Capacity scheduler shares fair percent of cluster. It supports FIFO scheduling within every queue with the pre-emption. When a TaskTracker slot becomes free, the job with the lowest load and lowest arrival time is chosen. A task is then scheduled from that job. The Yahoo developed the capacity scheduler . The intention of Yahoo was to concentrate on the conventional situation wherein there are large number of users and the goal was to ensure a fair number of resources among the users

V. MSHEFT ALGORITHM

For multiple platform business intelligence system. Heterogeneous Earliest Finish Time (HEFT) is used for scheduling the communication time of previous set of dependent task of heterogeneous network, HEFT tries to search for local optimization and eventually makes the whole optimal. HEFT algorithm is modified to Memory-Sensitive Heterogeneous Earliest Finish Time (MSHEFT) where the priority is considered first, then the size of data file is considered as the second condition, and finally an extra factor is considered, which is "Remaining Amount of Memory". The pseudo code of MSHEFT algorithm is shown below.

- ✓ Compute rank for all the nodes by traversing graph upward starting from exit node.
- ✓ Sort the nodes in the list in increasing order of rank values.
- ✓ When there are unscheduled nodes in the list the compare priority.
- ✓ Select the first job from list and remove it.
- ✓ If the memory size is > 0.6 GB then assign the task to the processor that minimizes the(EFT) value of node else wait for the remaining memory size and again repeat.

Here priority and memory size both are considered in order to schedule the task. If the memory size is below 75% and priority is less then task will go to Hadoop platform and high priority and high volume of data which takes more time for computation will be taken care by Spark. Thus the algorithm gives optimum results in multiplatform BI systems when compared to other algorithms.

VI. CONCLUSION

Today's business is data driven. In order to make proper decisions, there is a need of business intelligence system. This system should be cost effective, efficient and fast. Using multiple platforms and suitable scheduling algorithms it is possible to build cost effective sophisticated BI systems. MSHEFT algorithm works well in such environment and provides solutions in faster way.

VII. REFERENCES

1. Bao Rong Chang, Yo-Ai Wang, Yun-Da Lee, and Chien-Feng Huang "Development of Multiple Big Data Analysis Platforms for Business Intelligence". Proceedings of the 2017 IEEE International Conference on Applied System Innovation IEEE-ICASI 2017 - Meen, Prior & Lam (Eds).

2. Nagina, Dr. Sunita Dhingra – "Scheduling Algorithms in Big Data: A Survey ". International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 5 Issue 8 August 2016 Page No. 17737-17743 .
3. Art of scheduling for big data sciences.
4. "Business intelligence and analytics:from big data to big impact" Hsinchun Chen Eller College of Management, University of Arizona, Tucson, AZ 85721 U.S.A. {hchen@eller.arizona.edu} Roger H. L. Chiang Carl H. Lindner College of Business, University of Cincinnati, Cincinnati, OH 45221-0211 U.S.A. {chianghl@ucmail.uc.edu} Veda C. Storey J. Mack Robinson College of Business, Georgia State University.
5. Optimizing Load Balancing and Data-Locality with Data-aware Scheduling Ke Wang*, Xiaobing Zhou§, Tonglin Li*, Dongfang Zhao*, Michael Lang†, Ioan Raicu*‡ *Illinois Institute of Technology, §Hortonworks Inc., †Los Alamos National Laboratory, ‡Argonne National Laboratory kwang22@hawk.iit.edu, xzhou@hortonworks.com,{tli13,dzhao8}@hawk.iit.edu,mlang@lanl.gov, iraicu@cs.iit.edu.