IJCSET – Volume 3, Issue 2 – February 2017. ISSN: 2455-9091 Pages: 22-28

INCREASING COMPUTAIONAL CAPABILITY IN BIG DATA USING AN ENSEMBLE CLASSIFIER

[1] V.N.ANUSHYA, [2] A.X.SUGANYA GLADIES, [3] M. VIGNESH, [4] S. BHUVANESH,
[1, 2, 3, 4] ASSISTANT PROFESSOR,
[1, 2, 3, 4] DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING,
[1, 2, 3, 4] DHAANISH AHMED INSTITUTE OF TECHNOLOGY, COIMBATORE,

ABSTRACT: Big data refers to huge volume of data. Big data is the process of handling large datasets. In today's scenario, data is growing exponentially faster than ever so the concept of Big data has emerged. It can perform data storage, data analysis, and data processing as well as data management techniques in parallel. Big data can process several peta bytes (10^{15}) of data in seconds. It can handle both structured and unstructured data at a time. The aim of this project is to use the classification technique before mapping the tasks into the resources. For mapping the tasks, MapReduce programming model is used which reduces the workload on the resources. The MapReduce will take more time to decide the resource for performing the tasks which is to be allocated. Parallel Database technology is used to increase the performance of Big data because it allocate the tasks in parallel into the resources.

In this model, for classifying the tasks, Ensemble Classifier is used. An Ensemble Classifier is the group of different classifiers which make the classifiers to process in parallel and also shares the knowledge of fastest processing classifier to others. The Support Vector Machine, Decision Tree and K-Nearest Neighbor are the classifiers used to produce an Ensemble Classifier. Therefore, the data's will be processed with minimal scheduling time (the map class will not take time to decide to which resource the task has to be allocated). Along with Ensemble Classifier, Map Reduce model and Parallel Database Technology is used which increases the efficiency and throughput of Big Data by reducing the scheduling time.

Keywords: [MapReduce, Hadoop, Ensemble Classifier, Parallel Database]

1. INTRODUCTION

Big data is capable of handling large datasets at a time. It can perform data storage, data analysis, and data processing and data management techniques in parallel. Big data can process several peta bytes (10^{15}) of data in seconds. It can handle both structured and unstructured data at a time. Big data spends 70% of the time on gathering and retrieving the data and remaining 30% of the time is spend on analyzing the data. Big data can process even several peta bytes of data in seconds. Big data analytics will be most useful for hospital management and government sectors especially in climate condition monitoring. There are three characteristics of big data namely volume, velocity and variety.

The characteristics are explained in detail below.

Volume

Many factors contribute to the increase in data volume. Volume refers to the sense of storage in Big data. For example, in facebook 2.5 peta bytes of data are processed per day. Through the internet 2.5 Quintillion (10^{27}) bytes of data are processed per day. In order to store those large amounts of data we use Bigdata.

Velocity

Data is streaming in at unprecedented speed and must be dealt with in a timely manner. Velocity refers to the speed and performance of Big data. For example, internet can process only 4-5 Mb of data per second but in Big data 10Mb of data can be processed per second.

Variety

Data today comes in all types of formats. Variety refers to the types of data used in Big data. Structured data refers to numeric data in traditional databases. Unstructured data like text documents, email, video, audio, stock ticker data and financial transactions. For example, in Data warehousing and Data Mining either structured or unstructured data can be used. But in Big data both structured and unstructured data can be used at a time.

Hadoop

Hadoop is the most popular open source framework used in Big data to handle large datasets. It is a batch oriented system. Hadoop is used to analyze user interaction data. It is linear scalable on low cost commodity hardware. It is designed to parallelize data processing across computing nodes to speed computations and hide latency.

Hadoop Architecture

The architecture of Hadoop consists of one master node and many slave nodes. In the master node there will be a MapReduce model which is used for computation purpose and a Hadoop Distributed File System (HDFS) which is used to store large amount of data. Also in each slave node there will be a Map Reduce as well as HDFS. In the Map Reduce, the master node will take care of allocating the tasks to the slave nodes for processing. The HDFS in the master node will allocate the storage space to each slave node and also keeps track of where each data is located. Once the slave node finishes the given tasks it will send the results back to the master node. There will be commodity hardware in the Hadoop architecture which is used to improve the system's performance.

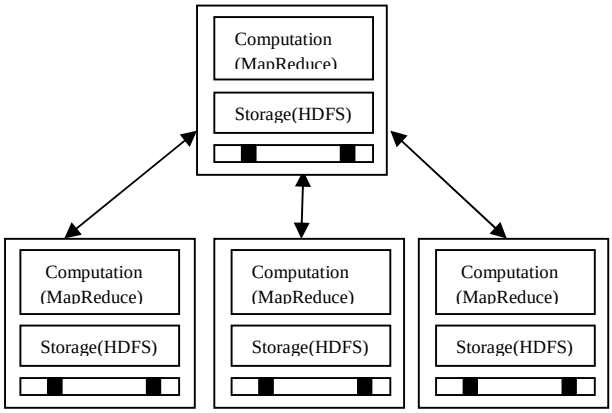

Figure. 1 Architecture of Hadoop

HDFS

HDFS is the storage component of Hadoop and is based on Google's Google File System (GFS). It is optimized for high throughput and works best when reading and writing large files. The blocks are replicated to nodes throughout the cluster based on the replication factor (default is 3). Replication increases reliability and performance. The architecture of HDFS consists of three daemons: They are:

- NameNode (Master)
- Secondary NameNode (Master)
- DataNode (Slave)

Name Node

The NameNode manages the filesystem namespace. It maintains the filesystem and the metadata for all the files. The namenode also knows the datanodes on which all the blocks for a given file are located, however, it does not store block locations persistently, since this information is reconstructed from datanodes when the system starts. Without the namenode, the filesystem cannot be used. For this reason, it is important to make the namenode resilient to failure, and Hadoop provides two mechanisms for this.

Secondary Name Node

Secondary namenode does not act as a namenode. Its main role is to periodically merge the namespace image with the edit log to prevent the edit log from becoming too large. The secondary namenode usually runs on a separate physical machine, since it requires plenty of CPU and as much memory as the namenode to perform the merge.

Data Node

Datanodes are the workhorses of the file system. The actual contents of the file are stored as blocks. They store and retrieve blocks and they report back to the namenode periodically with lists of blocks that they are storing. Datanodes must send block reports to both namenodes since the block mappings are stored in a namenode's memory, and not on disk.

Map Reduce

MapReduce was introduced by Google in 2004 for executing set of functions against large amount of data. It is a software framework for processing large datasets in a distributed fashion over several machines. The core idea behind MapReduce is mapping the dataset into collection of key/value pair and then reducing all pairs with same key.

Map Step

The master node takes the input, divides it into smaller sub-problems and distributes them to worker nodes. A worker node processes the smaller problem and passes the answer back to its master node.

Reduce Step

The master node collects the answer to all the sub-problems and combines them in some way to form the output.

2. LITERATURE SUPPORT

A. EXPLORATION ON BIG DATA ORIENTED DATA ANALYZING AND PROCESSING TECHNOLOGY

Big Data Processing Technology

Processing of big data is done by hybrid structure technology based on Map Reduce technology and parallel database technology.

MapReduce Technology

In 2004, Map Reduce was proposed by Google, it is an object-oriented programming model to deal with the large data, primarily used for processing internet data. The Map Reduce technology includes two basic operation conceptions:

- Map (Mapping) and
- Reduce (Simplication).

The Map technology mainly processes a group of input data record and distributes data to several servers and operation systems. Its means of processing data is a strategy based on the key/value. The Reduce technology mainly occupies itself in summarizing and processing the result after processing the above key/value. Map Reduce is designed for mass composed of low-end computer cluster, its excellent scalability has been fully verified in industry. Map Reduce has low requirement to hardware. Map Reduce can store data in any format, can achieve a variety of complex data processing function. Analysis based on the Map Reduce platform, without the need of complex data preprocessing and writing in the database process.

Parallel Database Technology

In the processing and analysis of the big data, data parallel processing manner is essential, because the processing strategy of "divide and rule" provides unlimited reverie to extend system performance.

Parallel computing includes two aspects:

- Data parallel processing and
- Task parallel processing.

1. In terms of the data parallel processing means, a large-scale task to be solved can be dissembled in to various system sub-tasks with the same scale and then each sub-task will be processed.
2. In terms of the task parallel processing, mode might cause the disposal of tasks and coordination of relationships overly complicated.

A. Constructing the Pattern on Big Data Processing

The hybrid structure mode of associating Map Reduce with Parallel Database is key for constructing the big data processing. SQL, as a universal relationship database system scripting language, therefore, the SQL scripting language can act as a entry point for the combination of the two. Map Reduce defines a self-defined interface function for SQL scripting sentence and provides the same grammatical form as common SQL scripting sentence. Within the self-defined function realizes the data processing done based on parallel computing. The interface function of such Map Reduce can normally operate under the database system circumstance, and it's returned result sets remain a usable data table.

B. ALGORITHM AND APPROACHES TO HANDLE LARGE DATA-A SURVEY

Data mining environment produces a large amount of data that need to be analyzed; patterns have to be extracted from that to gain knowledge. It has become difficult to process, manage and analyze patterns using traditional databases and architectures. This presents a review of various algorithms necessary for handling such large data set. To extract patterns and classify data with high similar traits, Data Mining approaches such as Genetic algorithm, neural networks, support vector Machines, association algorithm, clustering algorithm, cluster analysis, were used. Big Data architecture typically consists of three segments:

- Storage system
- Handling and
- Analysis

Big Data typically differ from data warehouse in architecture; it follows a distributed approach whereas a data warehouse follows a centralized one. The Data Mining termed Knowledge; its architecture was laid describing extracting knowledge from large data. Data was analyzed using software Hive and Hadoop. For the analysis of data with different format cloud structure was laid.

Contribution of Algorithms in Big Data Handling

Many algorithms were defined earlier in the analysis of large data set. In the beginning different Decision Tree Learning was used earlier to analyze the big data. The approach is to have a single decision system generated from a large and independent n subset of data. This paper uses a hybrid approach combining both genetic algorithm and decision tree to create an optimized decision tree thus improving efficiency and performance of computation. The earlier techniques were inconvenient in real time handling of large amount of data so in Streaming Hierarchical Clustering for Concept Mining, defined a novel algorithm for extracting semantic content from large dataset. The algorithm was designed to be implemented in hardware, to handle data at high ingestion rates.

In Hierarchical Artificial Neural Networks for Recognizing High Similar Large Data Sets., described the techniques of SOM (self-organizing feature map) network and learning vector quantization (LVQ) networks. SOM takes input in an unsupervised manner

whereas LVQ was used supervised learning. It categorizes large data set into smaller thus improving the overall computation time. The term maximal information coefficient (MIC) was defined, which is maximal dependence between the pair of variables. It was also suitable for uncovering various nonlinear relationships. It was compared with other approaches was found more efficient in detecting the dependence and association. It had a drawback –it has low power and thus because of it does not satisfy the property of equitability for very large data set. Then in 2012, we generate interaction between among objects and then grouping them into clusters. This algorithm (Combination of Genetic Algorithm and decision Tree) was compared with K-Means, CURE, BIRCH, and CHAMELEON and was found to be much more efficient than them. The advantage of this approach is that efficiency and performance of the computation will get improved when different algorithms are used to handle large data.

3. PROPOSED WORK

The classification technique is used to classify the whole dataset before mapping the tasks into the resources so that it reduce the time span, whereas during later period each and every data of whole dataset were analyzed individually and then mapped into the resources which consumes more time to complete the task. To classify and analyze the data before mapping, an Ensemble Classifier is used along with MapReduce model and Parallel Database technology to increase the efficiency and throughput of Big data.

A. ENSEMBLE CLASSIFIER

An Ensemble Classifier is the group of different classifiers which make the classifiers to process in parallel and also shares the knowledge of fastest processing classifier to others.

B. MAPREDUCE

Map step: The user will load the dataset to the MapReduce model. The map step will map the datasets based on key/value pair in parallel to other resources i.e the slave nodes to perform computation. After the computation work is over all the slave nodes will send the computed results to the reduce step.

Reduce step: In the reduce step shuffle and sort will be performed based on keys and values. After the values get sorted the computed results will be aggregated to produce a final output.

ALGORITHM FOR MAPREDUCE

The algorithm for Map Reduce includes a simple word count example.

The map step processes one key and value pair at a time.

Map (key: uri, value: text)
For word in tokenize (value)
Emit (word, 1) #found one occurance of word
Inverted Index
Map (key: uri, value: text)
For word in tokenize (value)
Emit (word, key) #word and uri pair.

The reduce step processes one key and all values that belong to it.

Reduce (key: word type, value:list of 1s)
Emit (key, sum (value))
Inverted Index
Reduce (key:word type, value:list of URIs)
#perhaps transformation of value
Emit (key, value)

PARALLEL DATABASE TECHNOLOGY

The parallel Database Technology is used to perform the computation of the given tasks in parallel. When all the slave nodes perform in parallel the performance of the system will be greatly improved.

4. THE IMPLEMENTATION TECHNIQUES

In this chapter the implementation details is done using Hortonworks Sandbox 2.0 with the Hadoop version 2.2.0 is used. We

can work with Hadoop in windows environment using Hortonworks Sandbox.

The implementation techniques are as follows:

1. Establishing the distributed resource clusters through Hadoop File System.
2. Incorporating the Ensemble Classifier in Hadoop File System.
3. Construction of Efficient data Analysis framework through hybrid structures
3.1. Incorporating the Parallel database as framework utilized for resource utilization.
3.2. Enabling a MapReduce Programming Model for handling the production workloads.
4. Performance Analysis through Make span and Response time factors.

The description of the above Modules is as follows,

Establishing the distributed resource clusters through Hadoop File System

The framework Hadoop is installed which distributes the resource clusters through Hadoop Distributed File System to perform processing. Hortonworks Sandbox 2.0 is installed with the Hadoop version 2.2.0.

Incorporating the Ensemble Classifier in Hadoop File System

An Ensemble classifier is incorporated in Hadoop File system to classify the tasks based on the classifiers used. An Ensemble classifier is a group of different classifiers which make them to process in parallel with the given data and also shares the knowledge of the fastest processing classifier to others. Therefore, the data will be processed with minimal scheduling time. An Ensemble Classifier such as Support Vector Machine, Decision Tree and K-Nearest Neighbor (KNN) classifiers will be used.

Construction of Efficient Data Analysis Framework through Hybrid structures

An efficient data analysis framework is constructed using MapReduce programming model and Parallel Database Technology. MapReduce programming model is established to map the incoming tasks into the resources and also greatly reduces the workloads of the resources. Parallel Database Technology is used for the processing of the tasks to be done in parallel by utilizing the resources efficiently, thereby increasing the performance of Big data.

Incorporating the Parallel database as framework utilized for resource utilization.

Parallel computing includes two aspects: data parallel processing and task parallel processing. In terms of the data parallel processing means, a large-scale task to be solved can be dissembled into various system sub-tasks with the same scale and then each sub-task will be processed. As such, compared to the whole task, it will be easy to process. Adopting the task paralleling processing mode might cause the disposal of tasks and coordination of relationships overly complicated. Using the parallel database technology is a means for realizing the parallel processing of data information.

Enabling a Map Reduce Programming Model for handling the production workloads

The Map technology mainly processes a group of input data record and distributes data to several servers and operation systems. Its means of processing data is a strategy based on the key/value. The Reduce technology mainly occupies itself in summarizing and processing the result after processing the above key/value. Map Reduce is designed for mass composed of low-end computer cluster, its excellent scalability has been fully verified in industry. Map Reduce has low requirement to hardware. Map Reduce can store data in any format, can achieve a variety of complex data processing function. Analysis based on the Map Reduce platform, without the need of complex data preprocessing and writing in the database process.

Performance Analysis through Make span and Response time factors

To measure the time taken to complete the computation work by classifying the tasks before mapping it into the resources, the performance analysis is done. The response time for each input should be efficient when compared to previous results. During performance analysis, the time taken to complete each task will be measured and also how the computation work is responding for each data will also be measured. The time taken to load the dataset will also be calculated and will be compared with the time taken by the dataset loaded previously. The large dataset will be analyzed and processed faster when compared to previous results as the dataset is loaded in the Hadoop. The comparison of the results will be represented in a graph.

CONCLUSION

In this project, the study of MapReduce programming model is done to reduce the workloads on the resources and also to allocate the tasks into the resources. The Parallel Database Technology is used to perform the computation tasks in parallel which increases the performance of Big data. In order to reduce the scheduling time for allocating the tasks into resources, classification technique is used before MapReduce and Parallel Database technology. For classifying the tasks an Ensemble classifier is used. An Ensemble classifier is a group of different classifiers such as Support Vector Machine (SVM) classifier, Decision Tree classifier, K-Nearest Neighbor (KNN) classifier etc. The study of these various types of classifiers is done to share the knowledge of the fastest processing classifier to others which will greatly reduce the scheduling time. Along with Ensemble Classifier, MapReduce programming model and Parallel Database Technology is used to increase the efficiency and throughput of Big data.

REFERENCES

[1]. Anwar M. A. and Naseer Ahmed,“Knowledge Mining in Supervised and Unsupervised Assessment Data of Students’ Performance”, 2011 2nd International Conference on Networking and Information Technology.

[2]. Chanchal Yadav, 2 Shuliang Wang, 3 Manoj Kumar 1 CSE, Amity University Noida, Uttar Pradesh, India, “Algorithm and Approaches to Handle Large Data - A Survey”, Volume 2, Issue 3, June 2013.

[3]. Dean J, Ghemawat S. “MapReduce: Simplified data processing on large clusters”/ / Proceedings of the 6th Symposium on Operating System Design and Implementation(OSDΓ 04) .San Francisco, California, USA, 2004: 137-150.

[4]. Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall and Werner Vogels, “Dynamo: Amazon’s Highly Available Key-value Store”,SOSP’07, October 14-17, 2007.

[5]. Kiminori Matsuzaki¤, Kento Emoto¤, Hideya Iwasakiy and Zhenjiang Hu, “A Library of Constructive Skeletons for Sequential Style of Parallel Programming”.

[6]. Mehmet Koyutu¨rk, Ananth Grama, and Naren Ramakrishnan “Compression, Clustering, and Pattern Discovery in Very High-Dimensional Discrete-Attribute Data Sets”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 17, NO. 4, APRIL 2005.

[7]. XIAO DAWEI, AO LEI, “Exploration on Big Data Oriented Data Analyzing and Processing Technology”, Vol. 10, 2013.