Suresh Gyan Vihar University,
Jaipur International Journal of Converging Technologies and Management (IJCTM)
Volume 1, Issue 3, 2015
ISSN : 2455-7528

Study of Distributed Computing and hadoop in big Data implementing Map Reduce Programming Model

Shweta[1], Piyusha Tiwari[2], Swati Kumari [3], Vaibhav Kumar[4]

1. *Scholar, B.tech ,CS department, Poornima College of Engineering, Jaipur(Rajasthan),India*
2. *Faculty IT department, Poornima College of Engineering, Jaipur(Rajasthan),India*
3. *Scholar, B.tech ,CS department, Poornima College of Engineering, Jaipur(Rajasthan),India*
4. *Scholar, B.tech ,IT department, Poornima College of Engineering, Jaipur(Rajasthan),India*

Abstract:

Relational Database failures give rise to the distributed computing using hadoop architecture with the help of mapreduce programming model and HDFS (Hadoop Distributed File System) as the storage. The hadoop provides big data environment for the programming model called as map reduces. The hadoop provides two important features of split execution and scalability and hence it reduces the fault tolerance. Map Reduce is a programming model based on the functional strategy which uses two works to be carried out i.e. Map () and Reduce (). The map function process a key/value pair and reduce function merges all intermediate key/value pairs. Program in the functional style are parallelized and executed on the large cluster of the machine. The hadoop architecture takes care of scheduling the program execution across the machine, handling the machine failure, manage the inter-machine communication. So user without any knowledge of parallel and distributed computing can easily utilize the resource of a large distributed system. The mapreduce being a distributed model carry out its execution in the parallel mode. The jobs in the mapreduce are scheduled using various schedulers. Our aim is to study about various adaptive schedulers in the mapreduce and their features.

Keyword: Distributed Computing, Hadoop, HDFS, Map Reduce, Relational database, Scheduler

Introduction:

Data is creating at the unprecented rate that 90% of the data has been generated in last 2 years due to enormously increased networking device. The relational databases are the traditional database to execute the queries on the large database. But the relational Database being a structured schema system which work on the dataset can't process unstructured data, data in the form of islands neither it can scale the data. It uses datawarehouse and datamart for carrying out the processing with two approaches i.e. Normalization and

Dimensional. The queries that could be fired limits in thousands and we need to process queries in lakhs so system can fail. Here the concept of Big Data comes into existence. The big data is all about data that is too big, too hard and too fast. So it is characterized by 4 V's i.e. Velocity, Volume, variety and Veracity. It defines the relationship of unstructured data, tapping of the devices and to carry out process in faster and accurate manner. Here schema less

Suresh Gyan Vihar University,
Jaipur International Journal of Converging Technologies and Management (IJCTM)
Volume 1, Issue 3, 2015
ISSN : 2455-7528

database is used e.g. NoSQL which follow ACID property. Hence to avoid the failure and achieve efficient use of data we use distributed computing using hadoop architecture with the help of mapreduce programming model and HDFS (Hadoop Distributed File System) as the storage.

Distributed Computing:

The system is designed to work in such manner that work is divided on the number of system to carry out its execution but it give resemblance like it's happening on the single system. Here there are two types of node. First is the master node and second is the slave node. The master system decides the job scheduling, division and number of replication that need to be done. It also can be replicated to a number of nodes which can be initially defined. After the execution of the work done by slave, they submit the final result to the master node and hence master node updates and each master node updates simultaneously.

Hadoop:

It is a software framework basically an architecture which allows hardware device to be compatibility with the distributed computing on the large datasets clusters for e.g. map reduce programming model and to carry out big data operations. It provides failure handling. The hadoop architecture consist of hadoop common, Map reduce programming model, Yarn scheduler and Hadoop Distributed File System.

[1]Hadoop Common: It is the common utilities for module support.

[2]Map Reduce Programming Model: It is the Parallel Processing of Hadoop [3]Yarn Scheduler: It is scheduler and resource manager.

[4]Hadoop Distributed File System: It is the file system of hadoop in which master and slave file system is used. Master node controls the slave node.

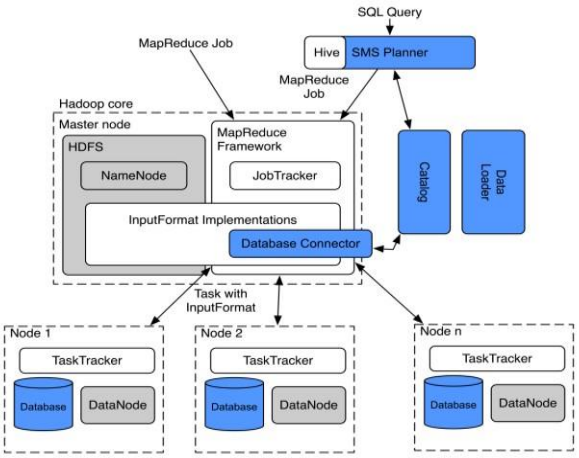

Fig.1. Hadoop Database Architecture

Whereas the Hadoop Database includes database connector, data loader, Catalog and query interface. [1]Database Connector: It is used to access multiple database system by executing SQL
Queries

[2]Data Loader: It splits the data into smaller chunks and provide coordinate their parallel load into the database system.

[3]Catalog: It contains the information of the data and the metadata and the information of chunks of data that is stored.

[4]Query Interface: It provides either API or SQL interface to execute the query.

MapReduce Programming model:

The mapreduce programming model is the parallel processing engine of the hadoop. It is based on the batch processing programming model. It is inspired by the functional programming model. It allows the distributed computing on the massive amount of data. It is

Suresh Gyan Vihar University,
Jaipur International Journal of Converging Technologies and Management (IJCTM)
Volume 1, Issue 3, 2015
ISSN : 2455-7528

based on two function Map () and Reduce (). It is fault tolerable and it is Scalable. It is based on divide and conquer algorithm. It includes two phases

[1]Map Phase: In this phase the map function takes the argument f (it takes single element) and this argument is applied to all elements in the list.

[2]Fold/Reduce Phases: It takes the function g (which takes 2 arguments) and it takes one initial value. In this the function g is applied to the initial value and the first element in the list. Then the result is stored in the intermediate storage that acts as the input to the next function of the g. The process is repeated till all the elements in the list ends

We can see the map function as the transformation over the dataset. In which the transformation is specified by function f and each and every functions used to happen in the isolation. The each function can be parallelized on the different machine. Similarly we can see the fold/reduce function as the aggregation function which is defined by the function g. the data need to brought together (local) to proceed further. The data sets must be grouped so that they can process in parallel.

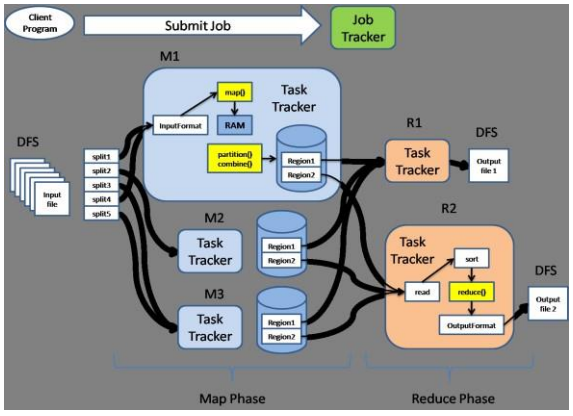

Fig.2. Working of the Map Reduce Programming Model

Here the key/value pair is used for processing the cluster of data. The key /value pair can be basic data structures like integer, float, string ,any arbitrary dataset where the arbitrary dataset include the URL of the website, URL may be considered as the key and the html content of the page may be treated as the value. The two functions are defined as below

[1] Map: (K1, v1) → [(K2, v2)]
[2]Reduce: (K2, [v2]) → [(K3, v3)]

The mapreduce consist of mapper that is applied to generate the intermediate key/value pair and the reducer is applied to the all intermediate key/value pair to generate the output key/value in the dataset of the distributed file system which is split across.

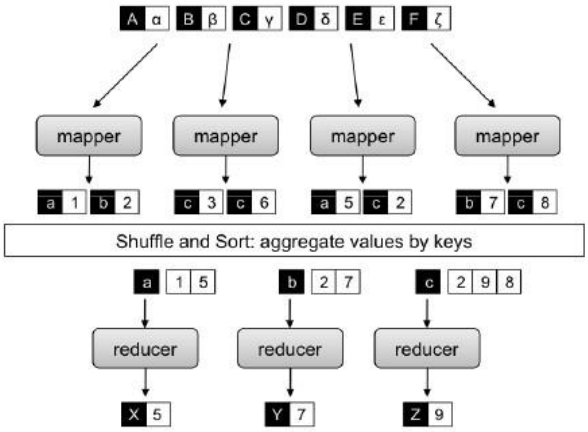

Fig.3. a simplified View of MapReduce

A simple program to print the hello world in the mapreduce consists of following pseudo code.

```
1: class MAPPER
2:     method MAP(docid a, doc d)
3:         for all term t ∈ doc d do
4:             EMIT(term t, count 1)

1: class REDUCER
2:     method REDUCE(term t, counts [c1, c2, ...])
3:         sum ← 0
4:         for all count c ∈ counts [c1, c2, ...] do
5:             sum ← sum + c
6:         EMIT(term t, count sum)
```

It is used for the word count of the program.
It takes following functions:

[1]Input: It takes a document file and a key attribute associated with it.
[2]Mapper: It takes a key-value pair, key is a word and value is a integer.
[3]Framework: It checks and guarantees that all values associated with same keys are bought to same reducer.
[4]Reducer: It returns all values to same keys, key is the word and the value is the count.

Scheduling in the mapreduce:

Each job in the distributed computing is divided into number of task. The number of task created may or may not exceed the number machines available so the jobs need to be assigned on each machine properly and that is where scheduling counts. So for better efficiency of the machine scheduler need to be customize. Nowadays adaptive scheduling has been taken into consideration. There are four types of scheduling: [1]Data locality ameliorator schedulers: When data is near then it is used.
[2]Adaptive schedulers based on speculative execution: This scheduler identifies the slow process and then launches several backups. [3]Performance manager Scheduler: It is capable to manage user goal.
[4]Resource Contention Reducer scheduler: It takes into consideration of the metrics of the task tracker.

Conclusions:

The mapreduce programming model develop on the platform of the java is one of the way to implement the Distributed computing on the large dataset. There is various ways to schedule the program in order to increase the efficiency and adaptive scheduling is the best way out of it.

Acknowledgement:

I want to give my sincere gratitude to the **Jawahar Nehru University** for providing the platform for the paper presentation. I would like to thank to my guide and mentor **Ms. Piyusha Tiwari** for her proper guidance and support. I really appreciate **Mr. Ramchandra** (Business Analyst, Toyata Company) for his proper guidance and motivation for competition of the paper work. I really thank to my parents for his proper and continuous support.

References:

[1] Sangeeta Bansal, Dr. Ajay Bansal, Transition Relational database to the big data, Department of computer Science, Amity University, Noida(UP) India, january 2014.
[2]Pietro Michiardi, tutorial on the mapreduce pdf (Eurecom), 2013
[3] Kamil Bajda, Daniel J. Abadi, avi Silberschatz, Erik Paulson, Efficient Processing of the Dataware House queries in the Split Execution Environment,
Hadapt Inc., 2Yale University, 3University of Wisconsin-Madison ,2011
[4]Jeffery dean and Sanjay Ghemawat, MapReduce: Simplified Processing on large dataset.Google.Inc.jeff@google.com;sanjay@g oog le.com, 2004
[5]Maedeh Mozakka, Faramarz Safi Esfahani ,Mohammad H. Nadimi, A survey on the adaptive Scheduler of the mapreduce, Faculty of computer engineering, Najafabad Branch, Islamic Azad
University, Najafabad, Iran, 2014