OPTIMAL PATH SEQUENCES OF MAPREDUCE JOBS USING HADOOP

Juliet A Murali[1], Tiny Molly V[2]
Assistant Professor ,Viswajyothi College of Engineering and Technology,Vazhakulam

Abstract
Big data analysis is one of the major issues in cloud computing. MapReduce has been widely used as a big data processing platform. Global MapReduce is a Hadoop based framework for cloud computing. GMR make use of GEO execution path. This paper deals with creation of optimal executing sequences of MapReduce jobs on geo-distributed data sets by incorporating Global MapReduce.

Key Words: Cloud, MapReduce, GMR, DTG

1. Introduction

Big data represents the large volume of both structured and unstructured data. It is very difficult to process the them by using traditional data processing applications. Other than processing, it also very difficult to capture manage and retrieve data from big data.

Computer cluster consists of a set of loosely or tightly connected computers. It can be viewed as a single system. The clusters can be inter-connected to form cloud. Cloud computing, is a kind of Internet-based computing that provides shared processing resources and data to computers and other devices on demand. A data center is a computer system having some functionality. Cloud computing provide host services in cost effective manner in association with large capacity data centers. Cloud computing includes both computational and storage services. MapReduce and Hadoop is an open source implementation by Apache .It can be used as a programming model for cloud- based data processing.

2. MapReduce

MapReduce has been widely used as a big data processing platform ,proposed by google. It is a programming model and an associated implementation for processing and generation large data sets. From the name itself , the data processes is done in two phases Map and Reduce. [2]

The map function processes a key/value pair to get intermediate key/value pair. [6]

The input to the map function is a split file contains key/value pair. The reduce function make use of an intermediate key. The reduce function merge all intermediate values. Figure 1 shows the MapReduce Model .

Scheduling is one of the most critical aspects of MapReduce. The inbuilt MapReduce scheduling algorithms include FIFO (First Input First Output), Fair Scheduler and Capacity Scheduler. [7] FIFO is the default Hadoop scheduler. The FIFO scheduler schedules jobs based on their priorities in first come first-out of first serve order. The fair scheduler was developed by Facebook and capacity scheduler by yahoo. [1]

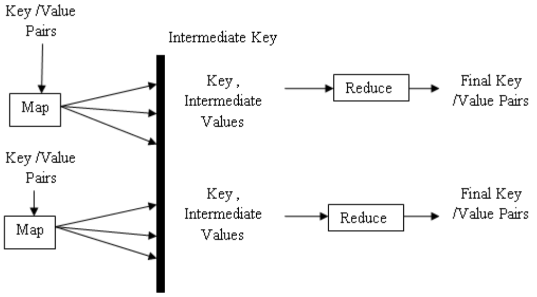

Figure 1: MapReduce Model

The goal of the fair scheduler is to provide fast response times for small jobs . The capacity scheduler was developed by Yahoo. Capacity scheduling algorithm puts jobs into multiple queues in accordance with the conditions, and allocates certain system capacity for each queue. [3] , [4],[8]

3. Hadoop

Hadoop is Java based open source implementation of the MapReduce platform. Hadoop runs over a distributed file system called Hadoop.

Distributed File System (HDFS) which has the same architecture as Google File System [1]. HDFS has master/slave architecture. HDFS consists of one the master server, called NameNode and there are a number of slaves, called DataNodes. NameNode which controls several DataNodes, and the DataNodes store actual data. Namenode supervises metadata such as information of directories, access log from users, detail of data location, and system logs. Datanode keeps data in Blocks. A Block is a basic unit for data storing in HDFS. Figure 2 briefly describes the Hadoop Architecture. [5]

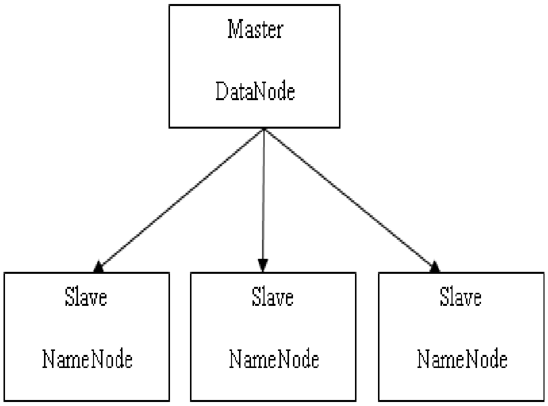

Figure 2 : Hadoop Architecture

The MapReduce framework has master/slave architecture. It has a single master server or JobTracker and several slave servers or TaskTrackers, one per node in the cluster. [5]

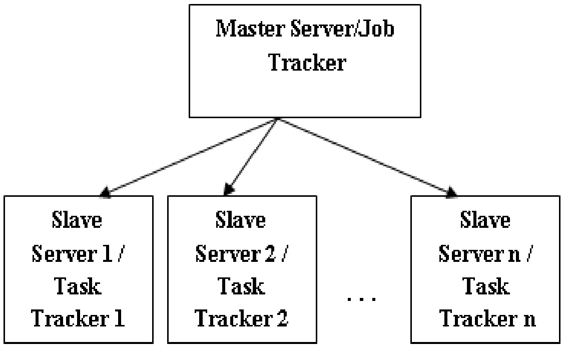
Master Server/Job Tracker
Slave Server 1 / Task Tracker 1
Slave Server 2 / Task Tracker 2
. . .
Slave Server n / Task Tracker n

Figure 3: MapReduce Master/Slave Architecture

The JobTracker is the point of interaction between users and the framework. Users submit map/reduce jobs to the JobTracker, which puts them in a queue of pending jobs and executes them on a firstcome/first-served basis. The JobTracker manages the assignment of map and reduce tasks to the TaskTrackers. The TaskTrackers execute tasks upon instruction from the JobTracker and also handle data motion between the map and reduce phases.Figure 3 MapReduce Master/Slave Architecture. [2], [3]

4. Global MapReduce

GMR is a Hadoop based framework for cloud computing. It operates on multiple datacenters instead of a single datacenter. They collect data from different datacenter and place it in the selected datacenter. A global MapReduce cluster is created from geographically distributed datanodes. [1],[9]

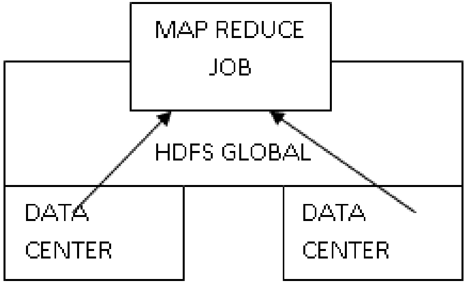

Figure 4 : G-MR Architecture

The collected data is stored into a global HDFS from different locations and run the MapReduce job over these global compute resources. Other than collecting data it can also execute instructions using Hadoop clusters. It provide sequence of map reducing , that means output of one job may become output of another one.GMR make use of data transformation graph algorithm DTG.DTG provides the optimal path for MapReduce sequencing. G-MR uses the DTG algorithm an optimized path to perform the sequence of MapReduce jobs and uses Hadoop MapReduce clusters. Figure 4 shows the G-MR Architecture. [3]

The GMR mainly consists of GroupManager JobManagers .The GroupManager is the main component. Each JobManager uses CopyManager and AggregationManager during the processing time. Figure 5 shows G-MR Structured Architecture

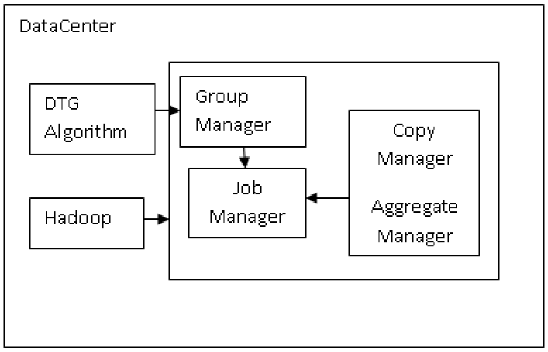

Figure 5: G-MR Structured Architecture

The GroupManager at any one of the Datacenter is initiated and it runs DTG algorithm. The DTG algorithm provides the optimal sequencing of MapReduce jobs. The GroupManager allocate MapReduce job to JobManagers corresponding to the Datacenter and also give information about subset of data to be processed. The Job Manager has the privilege to call CopyManager, when it need to copy data from other Datacenter. It can also call AggregationManager when it needs aggregation.

5. DTG Algorithm

DTG creates a data transformation group,that contains all possible MapReduce path. Each node represents number of MapReduce phases. The edge is a weighted edge , it represents the execution time or cost. DTG uses Dijkstra's shortest path algorithm inorder to determine optimal path. Te optimal path has lowest execution time or cost. [1]

6. Conclusion

The GMR is a MapReduce framework , that make use of Hadoop. It also uses the features of the copy execution path and aggregation path. It uses DTG algorithm for sequencing MapReduce .GMR provide an optimal path having minimum execution time or cost.

Reference

[1] Hadi Yazdanpanah, Amin Shouraki, Abbas Ali Abshirini,"A Comprehensive View of MapReduce Aware Scheduling Algorithms in Cloud Environments " , International Journal of Computer Applications (0975 – 8887) Volume 127 – No.6, October 2015

[2]A Dhineshkumar,M Sakthivel ," Big Data Processing of Data Services in Geo Distributed Data Centers Using Cost Minimization Implementation",International Journal of Innovative Research in Computer and Communications Engineering Vol. 3,Issue 3,March 2015

[3] Kirtimalini N. Kakade,T A Chavan ," Improving Efficiency of GEO-Distributed Data Set using Pact",International Journal of Current Engineering and Technology E-ISSN 2277-4106 P-ISSN 2347-5161June 2014

[4] Chamikara Jayalath, Julian Stephen, and Patrick Eugster," From the Cloud to the Atmosphere: Running MapReduce across Data Centers " , IEEE Transactions on computers, vol. 63, no. 1, january 2014

[5] Jeffrey Dean,Sanjay Ghemawat ," MapReduce:Simplified Data Processing on Large Clusters",OSDI'04: Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, December, 2004.

[6] Hadoop, "Hadoop home page." http://hadoop.apache.org/.

[7] M. Zaharia, A. Konwinski, A. D. Joseph, R. Katz and I. Stoica, "Improving MapReduce performance in heterogeneous environments", In: OSDI 2008: 8th USENIX Symposium on Operating Systems Design and Implementation, 2008.

[8] Hadoop's Fair Scheduler. https://hadoop.apache.org/docs/r1.2.1/fair_scheduler.

[9] J. Chen, D. Wang and W. Zhao, "A Task Scheduling Algorithm for Hadoop Platform", JOURNAL OF COMPUTERS, VOL. 8, NO. 4, APRIL 2013, pp. 929-936.