OPTIMIZATION OF WORKLOAD PREDICTION BASED ON MAP REDUCE FRAME WORK IN A CLOUD SYSTEM

V.Sivaranjani[1], R.Jayamala[2]

[1] Student, Pervasive Computing Technology, Bharathidasan Institute of Technology,Tamil Nadu, India
[2] Assistant Professor, Computer Science and Engineering, Bharathidasan Institute of Technology, Tamil Nadu, India

Abstract

Nowadays cloud computing is emerging Technology. It is used to access anytime and anywhere through the internet. Hadoop *is an open-source* Cloud computing *environment that implements the Googletm MapReduce* framework*. Hadoop is a framework for distributed processing of large datasets across large clusters of computers. This paper proposes the workload of jobs in clusters mode using Hadoop. MapReduce is a programming model in hadoop used for maintaining the workload of the jobs. Depend on the job analysis statistics the future workload of the cluster is predicted for potential performance optimization by using genetic algorithm.*

***Key Words:** Cloud computing, Hadoop Framework, MapReduce Analysis, Workload*

---***---

1. INTRODUCTION

The large scale data processing is very important aspects of the multimode cluster setup. It is very challenging problem. The MapReduce framework [1] is proposed by Google provides an efficient and scalable solution for working large-scale data. The basic concept of MapReduce framework is used to distribute the data among many nodes and process them in parallel manner. Hadoop is a open-source implementation of MapReduce framework. Hadoop use the Yahoo, Facebook, Twitter etc.

The MapReduce consists of the two Phases. 1) Map and 2) Reduce. The Map is used to split the job into several independent chunks and each chunks assigned to different computing data node. In the reduce phase, the data is aggregated, summarized, filtered or combining the given data. The result is stored in a Distributed File System.

Hadoop[2] is an open-source implementation of a MapReduce framework. The components of the MapReduce framework are 1) Job Tracker, 2) Task Tracker, 3) Name Node 4) Data Node.

The Name Node stores the file system metadata. Which file are maps to what block locations and which blocks are stored on which data node. The data node is where the actual data resides. All data nodes send the heartbeat messages to name node every 3 seconds to say data nodes are alive. If name node does not receive the heartbeat message from data node for 10 minutes, that data node is dead. All data node talks each other to rebalance the data, move and copy. The Job Tracker is used to managing the Task tracker and resource management that is tracking resource availability and time management of each job. The Task tracker is pre-configured a number of tasks and accept of each task. The Job Tracker consists of Job History. Get the required information from Job History to predict the future workload.

This paper describe about work load prediction on map reduce framework. The chapter 2 describes about System Architecture Design. Chapter 3 describes about Load prediction. Chapter 4 describes optimization process. Chapter 5 describes about Implementation and analysis. Chapter 6 describes Conclusion and Future work.

2. SYSTEM ARCHITECTURE DESIGN

The Job executes in cluster setup to get the job history information from the job tracker. The architecture design of the optimization of workload prediction based on the map reduce framework in a cloud system.

Fig- 1. Represents the MapReduce framework consists of different components are Name Node, Job Tracker and Task Tracker. The Name node stores the file in a distribute file system. The Job Tracker monitoring the resource availability and resource management of MapReduce framework.

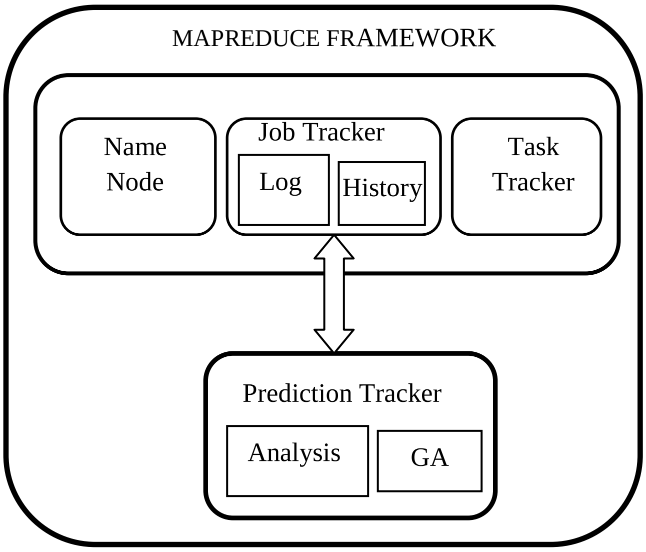

Fig -1: System Architecture Design

The Job Tracker consists of two phases. 1) Logs and Job History. Job History maintaining the past job description and provides different parameters like number of nodes in a cluster, number of the jobs, job Id, execution time and memory usage of each job etc. The Task Tracker is where the data is store resides and maintains data node information.

This paper proposes the prediction tracker component in MapReduce framework. The prediction tracker consists of two components 1) Analysis 2) GA (Genetic Algorithm).The analysis component get the job history related information from the Job Tracker. The GA is used to predict the future workload in optimized manner.

3. LOAD PREDICTION PROCESS

The load prediction mainly focuses the prediction tracker. The Analysis components of prediction tracker acquire the require job history information from the Job Tracker. The genetic algorithm is used to get the optimized solution for workload prediction based on the historical data.

The description of paper is listed as follows.

- Collect the workload of each job from the Hadoop cluster.
- Analysis the workload of each job
- Based the results, optimization performance is evaluated.

The trace file [3] of the job tracker data are JobID (a unique job identifier), job status (successful, failed or killed), job submission time, job launch time, job finish time, the number of map tasks, the number of reduce tasks, total duration of map tasks, total duration of reduce tasks, read/write bytes on HDFS (Hadoop Distributed File System), read/write bytes on local disks.

4. OPTIMIZATION PROCESS

Hadoop framework gives the trace file of the job tracker to get the job submission time. Prediction process [3][4] is based on the job submission time, duration of job completion time.

Forecast (Prediction) is an essential aspect of managing any organization is planning for the future. It is used to determine future inventory, costs, capacities and interest rate changes. There the two basic approaches of forecasting: qualitative approach, quantitative approach [6]. Qualitative approach is subjective, they are appropriate when past data are not available. Quantitative approach is used to forecast future data when past data are available.

This paper focuses on quantitative approach, based on an analysis of historical data which consider time series. A time series is set of observations measured at successive points in time. Time series is used to predict future values based on previously observed value [7].

Genetic algorithm is used to find the predicted value using historical data[8]. First step of the algorithm, select the population depends upon the original data element. Each element converted to the binary number to make a binary string or chromosome. The crossover point is selected and performs the crossover process and mutation process. Binary strings are converted to the real value. All actual value is converted to the binary strings or chromosomes. Operators of the genetic algorithm are three type’s selection, crossover and mutation.

The genetic algorithm [9] is used to
1. Initialize the population with random individuals.
2. Evaluate the fitness value of the individuals.
3. Select good solutions by using s-wise tournament selection without replacement
4. Create new individuals by recombining the selected population using single point crossover
5. Evaluate the fitness value of all offspring.
6. Repeat steps 3–5 until some convergence criteria are met.

Calculate the error rate using mean absolute percentage error. The mean absolute percentage error (MAPE) is also known as mean absolute percentage deviation (MAPD). It is a measure the accurate method for constructing acceptable time series values in statistics. The formula of MAPE

$$M = \frac{1}{n}\sum_{t=0}^{n} \left| \frac{\mathrm{A_t} - \mathrm{F_t}}{\mathrm{A_t}} \right|$$

$\mathrm{A_t}$ - Actual value

$\mathrm{F_t}$- Forecast value

n – Number of absolute value.

M – Mean Absolute percentage Error

5. IMPLEMENTATION AND ANALYSIS

In this paper, hadoop framework is installed in ubuntu operating system. Job history detail inferred from the job tracker with time series based. Table -1 represents the error rate of workload prediction.

Table -1: Example of Error value calculation

SI.NO	Predicted Value	Actual Value	Error Rate
1	12	15	0.2
2	15	14	0.07142
3	4	5	0.2
MAPE error rate(%)			9.04733

6. CONCLUSION AND FUTURE WORK

In this paper, we have presented the analysis of Hadoop trace derived from a single-node production Hadoop cluster. The trace covers the jobs execution files. In the future, we plan to work on the implications derived from this work and integrate them into the multi node cluster in real time.

REFERENCES

[1]. J. Dean and S. Ghemawat, "*Mapreduce: Simplified data processing on large clusters,*" in *OSDI*, 2004, pp. 137–150.

[2]. T. White, *Hadoop - The Definitive Guide*. O'Reilly, 2009.

[3]. Zujie Ren, Xianghua Xu, Jian Wan et.al "*Workload Characterization on a Production Hadoop Cluster: A Case Study on Taobao*" Proceedings of the 2012 IEEE International Symposium on Workload Characterization**,** 2012.

[4]. Sheng Di, Cho-Li Wang, "*Error-Tolerant Resource Allocation and Payment Minimization for Cloud System*" Proc. IEEE Transactions on parallel and distributed systems, VOL. 24, NO. 6, 2013, pp-1097-1106.

[5]. Zhen Xiao, Weijia Song, and Qi Chen "*Dynamic Resource Allocation Using VirtualMachines for Cloud Computing Environment*" proc. IEEE Transactions on parallel and distributed systems, VOL. 24, NO. 6, JUNE 2013, pp. 1107-1117.

[6]. http://www.wikipwedia.com/wiki/Time_series.

[7]. Sam Mahfound and Ganesh Mani "*Financial Forecasting Using Genetic Algorithms*" http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.86.9698&rep=rep1&type=pdf.

[8]. Satyendra, ArghyaGhosh, Subhojit Roy, J. Pal Choudhury, S. R. Bhadra Chaudhuri "*A Novel Approach of Genetic Algorithm in Prediction of Time Series Data*" in Proc of Special issues of international journal of computer application (ACCTHPCA), June 2012.

[9]. Abhishek Verma, Xavier Llora, David E. Goldberg and Roy H. Campbell,*"Scaling GeneticAlgorithms using MapReduce*" Proceedings of journal of cluster computing, special issue, 2011.

BIOGRAPHIES

V.Sivaranjani is a student,of M.E in Pervasive Computing Technology at Bharathidasan Institute of Technology. Her current research focuses on the cloud computing and parallel computing.

Mrs.R.Jayamala, Asst. Professor under the Department of Computer Science and Engineering at Bharathidasan Institute of Technology. Her research focuses on the cloud computing and Networks.