Reasearch of the Real-time Scheduling Algorithm Based on MapReduce

Y.Wu
School of Mathematical Sciences, Baotou Teachers College, Baotou, China

ABSTRACT: MapReduce is now the most widely used parallel scheduling model, the existing scheduling algorithm based on the MapReduce is not make full use of the system of free resources, this paper studies the real-time job scheduling strategy to make up for the defect, the scheduling process is divided into job scheduling and task scheduling. During the job scheduling stage, the prediction of deadline time would be completed, in the phase of task scheduling for the allocation of resource, free slots assign to the map tasks and reduce tasks with great extents, the resources running jobs take possession of can be preempted by a new job, fully meet the user to the real time requirement of the response time.
KEYWORD: MapReduce; real-time job scheduling; deadline prediction; resource allocation

1 INTRODUCTION

MapReduce is a distributed, parallel data processing framework, which has been widely used in cloud computing [1]. The main factors affecting the performance of MapReduce job scheduling is the scheduling algorithm, the existing scheduling algorithms include FIFO, fair scheduling, capacity scheduling, LATE scheduling and real-time scheduling, the scheduling algorithms have different advantages and disadvantages, since we require the cluster response time, in this paper, to study and improve the real-time scheduling strategy that based on MapReduce scheduling process.

2 SCHEDULING PROCESS OF MAPREDUCE

The MapReduce framework consists of clientnode, jobtrackernode, tasktrackernode and HDFS, client node is responsible for submitting jobs to the MapReduce, the jobtracker node is responsible for coordinating the operation, the tasktracker node is responsible for the operation, HDFS (distributed file system) for sharing files. Based on MapReduce job scheduling process of Hadoop [2] as shown in Figure 1.

Figure 1. Caption.' Leave about two lines of space between the figure caption and the text of the paper. Scheduling process of MapReduce consists of two stages: the first part is the scheduling process, jobtracker coordinates process jobs, namely, it decides which job to obtain the system resources; the second part is the task scheduling, after jobtracker received the job, each job can be divided into a number of tasks, and then enter the map phase, the task scheduling is between the scheduling of jobtrackernode and tasktrackernode. Task scheduling of MapReduce in this paper refers to the scheduling which is job scheduling and task scheduling.

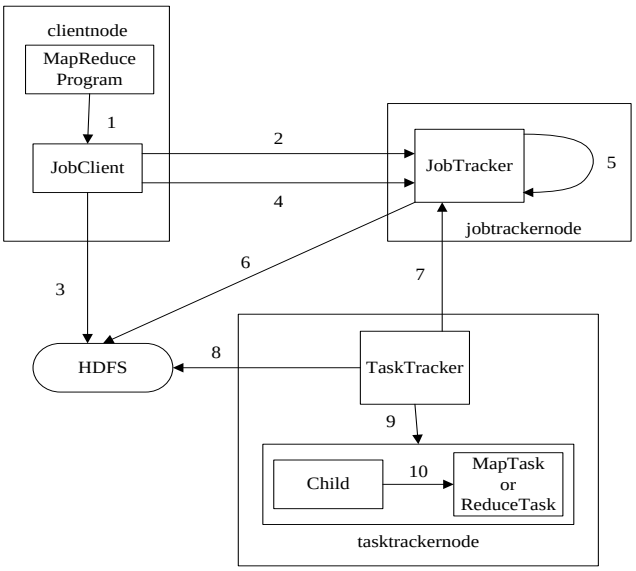

Figure1. Scheduling process of MapReduce

3 SCHEDULING ALGORITHM OF MAPREDUCE

3.1 *FIFO*

FIFO is the default scheduling algorithm, jobs submitted by the users are arranged in the order

presented in a waiting queue, which waits TaskTracker scheduling. This algorithm has the advantages of simple algorithm, but wasting system resources in the process of waiting, the average response time of the job is long[3].

3.2 *Fair scheduling*

This algorithm supports multi-user concurrent scheduling, which assigns multiple queues for multiple users, and it allows multiple users to share the resources of the cluster, it avoids long time operation of waiting, but it does not take into account the real-time job scheduling.

3.3 *Scheduling of computing power*

This algorithm is a multi-user and multi-queue scheduling algorithm which is based on job priority, allocation of resources are distributed according to the proportion of [4], each of the jobs in the queue is scheduling according to FIFO algorithm, the performance for real-time scheduling should be improved.

3.4 *Scheduling of heterogeneous cluster*

The existing scheduling algorithms are based on the homogeneous cluster, LATE algorithm is applicable to heterogeneous cluster, the cluster, the execution time of the same work in the different nodes of the same type is different, the LATE algorithm is scheduling algorithm for the slow work.

4 REAL-TIME JOB SCHEDULING

The core of scheduling of real-time job algorithm for user operation is timely feedback, to meet user requirements in time, so that the operation is completed in the deadline time. Scheduling priority is determined by the deadline time, deadline is smaller the higher the priority; scheduling priority is decided by the resource demand, demand is greater, the priority is higher, therefore, the efficiency of the algorithm is real-time scheduling of deadline time and resource demand.

Deadline time is provided by users, but users often cannot predict the parameters, which requires scheduling algorithm to achieve the prediction of [5] deadline. Each job is divided into a plurality of tasks, task types include map and reduce tasks, but these two kinds of task execution time is different, the deadline prediction strategy can not be biased towards a task, we need to put two kinds of task execution time that are taken into account, and can accurately predict the maximum of two kinds of task execution time, the prediction of the deadline for the map and reduce tasks are more accurate.

The work is divided into a minimum number of tasks, the amount of resources needed for each task is known, because of the number of required resources map tasks and reduce tasks are different, if the free slots (resource slot) assigned to the smallest map task, the system resources are not fully utilized, and it cannot guarantee that all operations are completed in the deadline time. In order to solve this problem, the scheduling strategy and the greedy algorithm are combined, and the idle slots are assigned to each possible jobs of currently running, in order to make full use of idle resources system, if a new job submission, new job deadline before the operation, the priority of the job will be changed, to ensure that all operations are run at the end of the deadline before.

Since the priority tasks in the job scheduling process is based on dynamically changing deadline, so real-time scheduling in this paper is to study the dynamic preemptive scheduling strategy, and it is to solve the problem of dynamic allocation of resources prediction and deadline time.

4.1 *Deadline prediction*

When the jobtracker receives a new user that is submitted work, first to predict deadline time [6] operation. The process of forecast is: because of the job is composed of multiple tasks, assuming the number of map tasks for this job is m, the number of reduce task is r, before the job starts to run is first sampling, and it selects p map tasks and q reduce tasks, map task and reduce task average running time of Tm and Tr, 4.1 and 4.2 respectively in accordance with the formula calculation.

$$T_m = \left\lceil \frac{\sum_{i=1}^{p} (e_{mi} - s_{mi})}{p} \right\rceil \tag{4.1}$$

$$T_r = \left\lceil \frac{\sum_{i=1}^{q} (e_{ri} - s_{ri})}{q} \right\rceil \tag{4.2}$$

Among them, e_{mi} and s_{mi} indicate that the i start time of the map task running and end time of the map task running respectively, e_{ri} and s_{ri} represent the i start time of reduce task running and end time of reduce task running.

In the course of the actual operation, the operation of map task closes to the actual situation of sampling, so Tm is used as the running time of map task. But, the operation of reduce task is not necessarily consistent with the actual situation, So Tr only in the initial stage of reduce task to run, which there is reference value, the actual operation of the process, with a completed reduce task time instead of predictive value, the prediction value of execution time of the reduce task is more closed to the real value, which also reflects the dynamic of deadline prediction.

In the end of deadline, to complete the job scheduling and task scheduling, so, in addition to considering the execution time of map task and reduce task, but also consider the segmentation time of the job, scheduling time and task merging time, setting a time constant Tc for the process, assuming the deadline time is TD, the TD calculation method is as shown in equation 4.3.

$$T_D = \frac{(1+m)*m}{2} T_m + \frac{(1+r)*r}{2} T_r + T_c \tag{4.3}$$

Only through the average of execution times andthe formula 3.1 and formula 3.2. Each map taskand reduce task to measure the average of execution time of the product. m map task execution times is1 times, 2 times,....., m times, so it can only take the average, r reduce task executiontimes is 1 times, 2 times,.....,r times, the averagecoefficient formula 3.3 Tr. According to the formula 3.3 predicts job deadline which considers the execution time of each task, and the task execution times, and time operation during the process of other consumption, so the prediction result is reliable and consistent with the actual situation.

4.2 *Resource allocation*

Work is divided into quantity of the task which can not be too big or too small, because of the task is stored in the HDFS, while the storage unit of HDFS is a task, if the quantity is too small, each task is relatively large, and it is more than the size of HDFS storage unit , which does not facilitate the task of management; if the task is too large, the storage space for each task that is relatively small, which will cause the waste of memory space of HDFS and waste of system resources, therefore, the operation is divided into the size of HDFS block that is the best choice. The size of HDFS block is usually 64MB, it will adjust the size according to the specific situation of Hadoop cluster, if the cluster server configuration is relatively high, which can be adjusted to 128MB or more.

The number of slots idle of cluster resource is assigned to the map task and reduce task, The purpose is to coordinate the number of allocation, which is assigned to the map task and reduce task, not only makes the operation that can be done in deadline time, but also to make use of the system resources in the shortest time to complete the job execution.

If the system is assigned to the map m slots, all the map tasks can be done in parallel, then the map task completion time is Tm, the reduce task in TD-Tm time can be completed; if it only allocates a slots, the map task is only serial scheduling, and complete the time of all map task that is m*Tm , and complete the time of reduce task is only TD- m*Tm. Principles of resource allocation is to make full use of idle resources, and assigns to the current maximum running jobs, and embodies the greedy thought, before the deadline at the end to ensure that the job is running ,which makes running time the shortest .

The number of slots hypothesis free is Es, map is assigned to the number of task slots is Em, reduce is assigned to the number of task slots is Ep, the relationship is satisfied as in formula 4.4.

$$\begin{cases} E_m + E_r \le E_s \\ \frac{E_m}{m} * T_m + \frac{E_r}{r} * T_r \le T_D \end{cases} \tag{4.4}$$

In formula 4.4, the number of free slots Es is not fixed, in the process of operation, due to other operation completed will release the slots, so that the amount of free slots increases, the distribution rules of resources will change, but it always satisfies formula 4.4. In the process of operation, if the MapReduce model receives high priority user job request, i.e. deadline is less than the current operation of deadline, which will happen resource preemption and will be free to assign resources submitted jobs, job priority of currently running is smaller and suspend operation, so the resource allocation is a dynamic process of preemption.

5 STRATEGY OF REAL-TIME JOB SCHEDULING

Two key problems described above on deadline prediction of real-time scheduling and the solution of resource allocation, process with the scheduling combined.

The process of real-time job scheduling includes job scheduling and task scheduling. Job scheduling, namely MapReduce receives the prediction of deadline, and assigns the job situation sorted according to deadline priority of the waiting job; task scheduling, task assignment for free slots, the distribution of the order in accordance with the priority of a task, the task priority order decision by task scheduling, task priority must first complete a high priority, no effect on the other task running low priority scheduling, in accordance with the resource allocation strategy that is based on priority.

5.1 *Job scheduling*

After MapReduce model received the operation, the operation is divided into a plurality of task [7], segmentation principle is at least map task and the least reduce task, the task is less, the average execution times is less, by the formula 3.1 and formula 3.2 to calculate average time of execute of map task and reduce task , according to 3.3 formula calculates the minimal deadline time,

compared with the calculated deadline and the deadline of waiting queue operations , giving priority to all the operations, deadline smaller job priority is higher, then according to the operating priority will put this job in the corresponding position of the waiting queue. Job scheduling is the order of the waiting queue of the first team work. If this process has a new job that is submitted, which will repeat the above process, the priority of the job is the main factors to determine the allocation of resources operation.

Job scheduling algorithm is described as follows:

Input: Operation

Output: wait for the job queue

Begin

a) Work is divided into several map tasks and reduce tasks;
b) Calculate the average running time of map task;
c) Calculate the average running time of reduce task;
d) Predict the deadline time of the job;
e) Determine job priority;
f) According to the priority into the waiting job queue;
g) Select the first team work as a job scheduling.

End

5.2 *Task scheduling*

In the process of scheduling assigns resources to tasks, the allocation of resources in accordance with the formula of 3.4. Task scheduling according to priority, the priority of task is determined by the demand for urgent task of resources, if a task does not run, which will lead to other task that cannot run, urgent degree of this task is high, the priority is the largest, if not the first operation of these tasks, which will lead to other successor task cannot run, work can not complete before the deadline time, which ignores the real-time scheduling. The next step is to free slots statistics node, because there are multiple tasktrackernode in clusters, idle slots of each node is different, statistics of all nodes, and according to the order of nodes will be free slots number sequence. When resources are allocated, firstly it chooses free slots number of the nearest node operation required total number of resources, so as to avoid idle slots the small number of each node, the node slots is not fully utilized.

The process of resource allocation in task scheduling, if a new job is submitted, and in accordance with the process of job scheduling found a higher priority than running jobs, this time the job of the operation should be stopped immediately, it schedules the new work, and allocates resources to the new operation, it reflects the strategy of real-time scheduling jobs and strategy of preemptive resource allocation fully.

The algorithm of task scheduling is described as follows:

Input: Map tasks and reduce tasks

Output: the scheduling results

Begin

a) Map tasks and reduce tasks are sorted;
b) 2. Free slots statistics for each tasktrackernode;
c) 3. Idle slots node puts into the queue;
d) 4. According to the resource allocation principle to assign resources to tasks;
e) 5. A new work of high priority is submitted;
f) 6. Stop the current job scheduling;
g) 7. Scheduling a new job.

End

The process of job scheduling and task scheduling make full use of free slots of the system, and which will assign idle slots to as many map tasks and reduce tasks based on the greedy algorithm, new job of running operation plays resource preemption, and allows multiple concurrent scheduling work, which fully satisfies and embodies the real time of job scheduling.

6 SUMMARY

The algorithm of real time scheduling in this study Conducts deadline prediction and resource allocation in the job scheduling and task scheduling, real-time scheduling multiple concurrent users, compared to FIFO, fair scheduling, capacity scheduling and heterogeneous cluster scheduling calculation and improve the efficiency of job scheduling, improving the utilization rate of resources, and it meets the real-time scheduling, which is a good performance of the strategy which is based on MapReduce scheduling.

REFERENCES

[1] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. ACM, USA. 2008: 107-113.

[2] White T, Min-qi Zhou, Xiao-ling Wang, Che-qing Jin, Wei-ning Qian. Hadoop: The Definitive Guide. 2011, Tsinghua University press.

[3] Yuan Zhou.Hadoop scheduling research and application [Master's thesis]. Lanzhou University, 2012.

[4] Jun-qing Zhou. Research on distributed task scheduling algorithm based on Hadoop platform [Master's thesis], Hunan University, 2012.

[5] Ji Liu, Lanxiang Chen, Dong Dai, Ming-ming Sun, Xuehai Zhou. Design and implementation of a real time MapReduce scheduling algorithm, Application of computer system, 2013, 22(8): 113-119.

[6] Kc K, Anyanwu K. Scheduling Hadoop Jobs to Meet Deadlines. Cloud Computing Technology and Science, 2010 IEEE 2nd International Conference. 2010:388-392.

[7] Zaharia M, Borthakur D, SenSarma J etc al. Job scheduling for multi-user MapReduce clusters. EECS Department, University of California, Berkeley, Tech. Rep.UCB/EECS-2009-55, Apr2009.