IMPROVEMENT OF SCHDULING IN HADOOP USING MACHINE LEARNING

[1]Yashverdhan P Singh, [2]Shyam Deshmukh
[1]ME Student, [2]ASST Proffessor
Computer Engineering
GTU PG School, Gandhinagar,India , PICT, Pune

Abstract— The volume of data these services work with interest in parallel processing of commodity cluster. The example of Map reduce is goggle uses the framework to process large amount of data. Other Internet service such as E-commerce website and social networking uses Map Reduce framework. Basically Map Reduce is a framework for preparing large amount of sets gives utilize adhoc solution, issues like web indexing, Data sorting, data searching. Map Reduce framework use for better scheduling so those less straggler occur in the machine. At the time when the straggler occur in the machine the performance of the machine is degrade the machine become slow. There are various technique used to improve the scheduling such as LATE,spectacular execution , smart spectacular execution.
In my research I proposed new technique used in hadoop by using Machine learning. The machine learning technique is proactive techinique. In the proposed model the straggler will be find as soon as possible and improved the scheduling. In this model the straggler found with these past experience task so that tranning period of slave machine will be reduce.

IndexTerms— MapReduce, Straggler, Scheduling, Machine Learning

I. INTRODUCTION

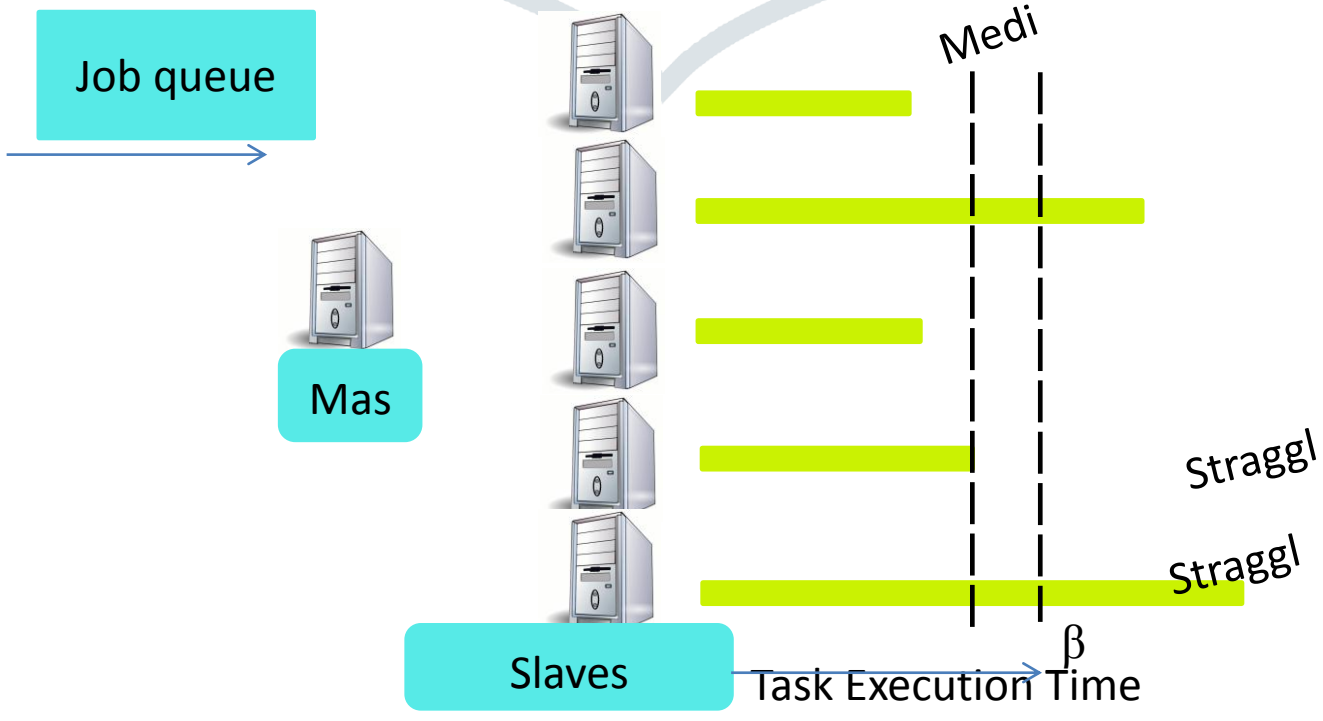
Today's most popular computer application are Internet services millions of users. The volume of data these services work with interest in parallel processing of commodity cluster. The example of Map reduce is goggle uses the framework to process large amount of data. Other Internet service such as E-commerce website and social networking uses Map Reduce framework. Basically Map Reduce is a framework for preparing large amount of sets gives utilize adhoc solution, issues like web indexing, Data sorting, data searching.

Now a days people are highly trained the distribution computing because of demand in paralleling and transparency executing the task the most challenging task to reduce the straggler in distributed system of big data processing network. Map Reduce, and many others at companies implemented big cluster of data to compute the large amounts of raw data, such as documents, Web request logs, etc., to compute various kinds of derived data, such as inverted indices, various representations of the graph structure of Web documents, summaries of the number of pages crawled per host, and the set of most frequent queries in a given day[1].

A Map Reduce job generally splits the input into multiple input which are processed by the map tasks in a completely parallel manner. The model sorting the output where the outputs of the maps, which are then input to the reduce tasks. Equally importantly, if a node is available but is take to more time to perform particular task than its called straggler.the straggler occur in the way the When the straggler occur in the machine it will delay the job execution time and reduce the cluster throughput the number of jobs completed per second in the cluster. The Map Reduce operation is perform the on hadoop.

II. STRAGGLER PROBLEM

The Straggle in the Distributed system when the master allocate the task to slave machine to complete it if the slave machine takes too much time to complete the particular task beyond the threshold limit so we can say the straggler occur in the particular slave machine.



In the above figure the job is given to master machine and the master machine distribute the job into task to slave machine the slave machine has to complete the task in median limit if the slave machine fails to complete the in the time limit then the slave have straggler. So in the figure shown slave machine four and five having straggler because its takes too much time to complete the allocated task.

The Straggler defined as:

Let normalized durations,

$$nd(t_i) = \frac{\text{Task execution time}}{\text{Amount of work done by task } t_i}$$

A straggler is defined if for task t_i of a job J

$$nd(t_i) > (\beta \times median\{nd(t_i)\})$$

where,

β is threshold coefficient ($\beta \sim 1.3$) or signifies the extent to which a task is allowed to slow down before it is called a straggler.

A current study shows that the straggler existing mitigation techniques,the impact of straggler in straggler tasks can be 6- 8× slower than the median task in job on a production cluster. The production is the cluster where the actually job is the job is execute in the system execute. The straggler occur 22% to 28% in the system so the eliminate or mitigate the straggler is our main objective. The hadoop use commodity hardware and the task failure become part of occurring the straggler.

III. STRAGGLER MITIGATION TECHNIQUE

There are two way to mitigate the Straggler they are Reactive Straggler mitigation and proactive straggler mitigation. In reactive straggler mitigation when the straggler occur in the slave machine than the master run the duplicate copy to another slave machine to removing the straggler because of this throughput is degrade and job completion time is increase. There are some reactive technique use to mitigate the straggler.

1. Speculative execution
   The new strategy create named Maximum cost performance. The MCP include the various method such as Use both the progress rate and the process bandwidth within a phase to select slow tasks. The execution of the straggler Use exponentially weighted mo to predict process speed and calculate a task's remaining time Determine which task to backup based on the load of a cluster using a cost-benefit model. The MCP in a cluster of virtual machines running a variety of applications on 30 physical servers.
2. LATE allows the slow nodes in the cluster to be utilized as long as this does not hurt response time. In contrast, a progress rate based scheduler would always re-execute tasks from slow nodes, wasting time spent by the backup task if the original finishes faster.
3. MANTRI
   Mantri where a system that monitors tasks and outliers using cause- and resource-aware techniques. This strategies include restarting outliers, network-aware placement of tasks and protecting outputs of valuable tasks. Using real-time progress reports, Mantri detects and acts on outliers early in their lifetime. Early action frees up resources that can be used by subsequent tasks and expedites the job overall.

Proactive technique

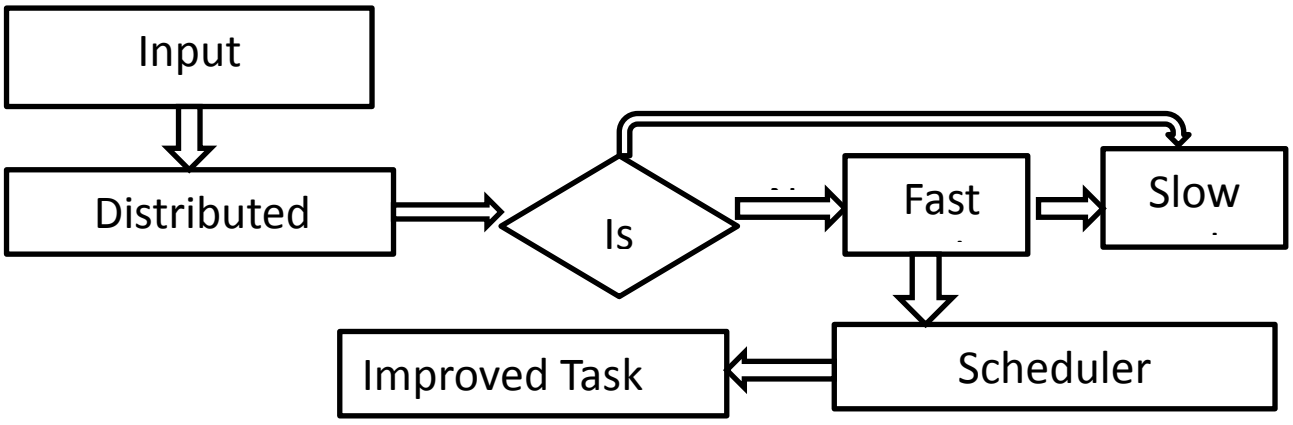
In Proactive Straggler mitigation technique the machine learning technique is used to mitigate the straggler. In proactive straggler mitigation technique where the straggler occurs in slave machine predict earlier before the scheduling. There are various way proactive technique is existing they are

Dolly

Clone of small jobs only marginally increases utilization the execution of the because workloads show that while the majority of jobs are small, they only consume a small fraction of the resources. The main challenge of cloning is, however, that extra clones can cause contention for intermediate data. We use a technique, delay assignment, which efficiently avoids such contention where Evaluation of our system, Dolly, using production workloads.

IV. PROPOSED MODEL

Due to problem occur in the existing system Hadoop scheduling and LATE, Speculative execution, Smart speculative execution the straggler occur in the existing system to improve the scheduling technique the proposed model is improving scheduling using machine leaning.The proposed model is proactive mitigated the straggler in which the model the machine learning is used to predict the straggler. Each slave node having own straggler prediction model from where the slave node predict the straggler and tells to master node.



In the Proposed model architecture the two input parameter will be take from workload. The node parameters and job parameter features will be extract and store in the tabular form. Then that data will be give to distributed SVM (support Vector Machine) the SVM is used to predict the straggler with the help of input feature of node parameter and job parameters and identify the which node is slow or fast.

Then the condition will be check Is node slow if the node will be slow than node will pas through fast node queue otherwise the node will be go through slow node to fast. After passing through to the fast node the scheduler will scheduler the task to slave machine. Workload is done to word count the word count program takes as input.

V. IMPLEMENTATION

There are one master node and four slave node all the machine install the hadoop and ganglia tool. when the jobs come to the master node the jobs are divided into task and allocate to the respective slave node than the node parameter of all the slave node are collected by RRD tool. The RRD tool is the graphical representive of the parameter. The RRD tool are in built with ganglia. The ganglia are monitor the resource are use by the slave machine. When the RRD tool collected the node parameter the unused data is cleaning and labeled is done. When the node parameter done the labeled the python script is run in the SVM. The Support vector machine the machine learning technique used in the machine learning than the MPI lib are used to create the model of each slave node. The node parameter are extract by overload each node and defining the straggler than after the another jobs came the SVM ask the ganglia to identify the straggler than the data is split and training are done and predicted the slave node is straggler or not and the data is test whether the prediction model is right or not and accordingly the labeled is done. By this model building time is reduce and master node the task with fast assignment is improved.

VI. EXPERIMENTAL RESULT

Output of ganglia bytes _in bytes out of the node parameter of all the slave machine.

VII. CONCLUSION

The usage amount of data is process day by day to so the performance of the system is degrade so the various technique are used such as reactive and proactive technique to improved the scheduling. The machine learning SVM technique is used to identify the straggler the system performance is degrade because to large amount of straggler occur during the job execution. So the mitigate the straggler is most important issue. In the proposed model the model is created to individual slave machine in distributed manner. Each slave machine create its own model to identify the straggler and inform the master machine weather the particular node is straggler or not for particular task accordinlay the master machine scheduling the job. The master machine only schedule in fast slave machine. When another task come to the master machine the master machine ask ganglia wheather the slave machine have straggler or not. In the proposed model label based scheduling is done.

REFERENCES

[1] Jeffrey Dean and Sanjay Ghemawat**,"**Map Reduce: Simplified Data Processing on Large Clusters**",** commutation of ACM , January 2008.

[2] Matei Zaharia, Andy Konwinski, Anthony D. Joseph, Randy Katz, Ion Stoica, "Improving MapReduce Performancein Heterogeneous Environments", OSDI,2008.

[3] Qi Chen, Cheng Liu, and Zhen Xiao, "Improving MapReduce Performance Using Smart Speculative Execution Strategy",IEEE, APRIL2014.

[4] Ganesh Ananthanarayanan, Ganesh Ananthanarayanan, Bikas Saha, "Reining in the Outliers in Map-Reduce Clusters using Mantri",2010

[5] Ganesh Ananthanarayanan, " Effective Straggler Mitigation: Attack of the Clones", 10th USENIX Symposium on Networked Systems Design and Implementation (NSDI '13) USENIX Association.2013

[6] Neeraja J. Yadwadkar, Ganesh Ananthanarayanan", *Wrangler:* Predictable and Faster Jobs using FewerResources"IEEE**.**