Efficient Analysis on Big data Using Map Reduce in Cloud Environment

Rajvi N Vora[1], Darshil J Shah[2]

[1] *Rajvi N Vora Masters in, Computer Silver Oak College of engineering & Technology, Gujarat, India*
[2] *Darshil J Shah Masters in, Computer Silver Oak College of engineering & Technology, Gujarat, India*

ABSTRACT

Abstract Nowadays clouds are overloaded by dozen of peta bytes & zeta bytes sized data. Data Inputs are large like social media, industry data, logs of every field and to handle these data we have very few technology like Map reduce, Hadoop, Hive, PIG, WibiData, Platfora, etc. Big Data has many challenges like Handling data volume, Analysis of Big Data, Privacy of data, Storage of huge amount of data, Data visualization, Job scheduling , F ault tolerance. In this paper we are showing improved Fair Scheduling algorithm based on node health degree for job allocation to nodes. Jobs are assigned to node equally and Fair4s algorithm can ensure to reduce job fail rate and improve cluster throughput..

Keyword *Hadoop, Map Reduce, HDFS, Job Scheduling, Big data, Cloud Computing*

1. INTRODUCTION

Number of people large amounts of data has become available on hand to decision makers. Big data refers to data sets that are not only big, , as well as high in assortment and speed; they are difficult to handle using traditional tools and techniques. Rapid growth of this data, solutions need to be studied and provided in order to handle and extract value and knowledge from these data. sets. Furthermore, decision makers need to be able to gain valuable insights from such fluctuated and quickly evolving information, ranging from daily transactions to customer interactions and social network data. Such value can be provided using big data analytic, which is the application of advanced analytic techniques on big data. Researcher gives more interested to build a job scheduling algorithms that are well-suited and appropriate in cloud computing of the critical situation. Job scheduling is one task in cloud computin g because the user have to pay for services based on usage time.

1.1 HADOOP

Hadoop is an open source Java-based, programming structure that backings the preparing of extensive information sets in a circulated figuring environment. Hadoop was motivated by Google's Map Reduce (GFS), a product structure in which an application is separated into various little parts. Any of these squares or pieces can be keep running on any hub in the group. Doug Cutting, Hadoop's maker, named the structure after his kid's full toy elephant. The present Apache Hadoop biological community comprises of the Hadoop piece, Map-Reduce, the Hadoop disseminated record framework (HDFS) and various related undertakings, for example, Apache Hive, HBase and Zookeeper.

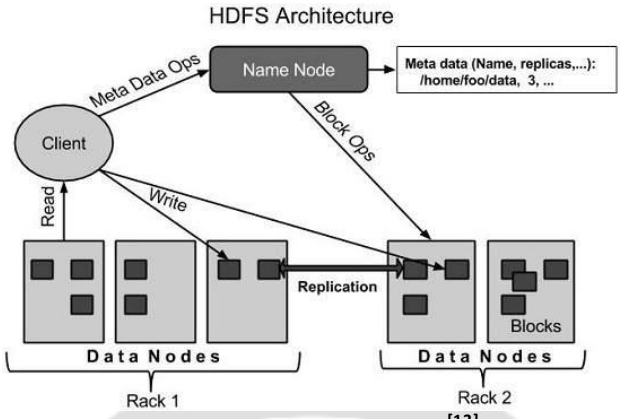

Figure 1 HDFS Architecture [13]

1.2 MAPREDUCE

MapReduce is a handling system and a project model for circulated registering in view of java. The MapReduce calculation contains two imperative assignments, to be specific Map and Reduce. Map takes an arrangement of data and proselytes it into another arrangement of information, where singular components are separated into tuples (key/value sets). Furthermore, diminish errand, which takes the yield from a guide as an info and consolidates those information tuples into a littler arrangement of tuples. As the grouping of the name MapReduce suggests, the lessen errand is constantly performed after the guide work.

Map function takes input pairs and produces a set of intermediate key and value pairs and passes them to the Reduce function in order to combine all the values associated with the same key. Reduce function accepts an intermediate key as a set of values for that key; it merges together these values to prepare a proper smaller set of values to produce the output file. For doing proper map reduce function we need to do scheduling with big data hence from lots of scheduling algorithm we are selecting best algorithm **Fair4s** algorithm to have a proper output.

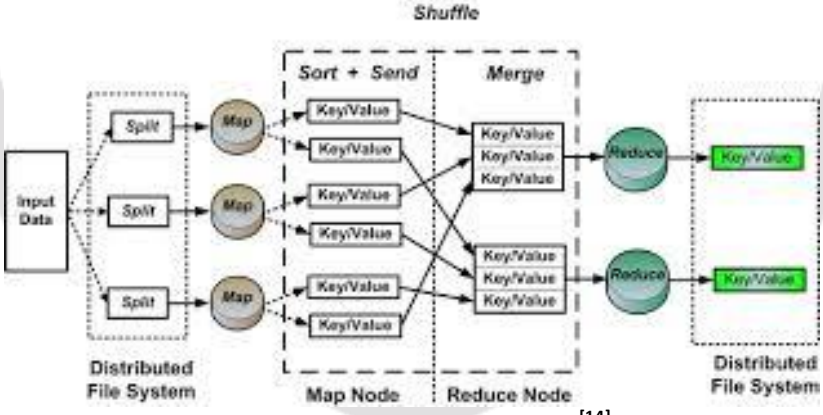

Figure 2 MapReduce Framework [14]

2. PROPOSED SYSTEM USING FAIR SCHEDULER

Research uptill now is saying that we can improve efficiency of big data files on cloud with going through process of fair scheduler and improving their output with checking node health degree. Jobs are assigned to idle node first and then after checking node health degree plus characteristic variables.

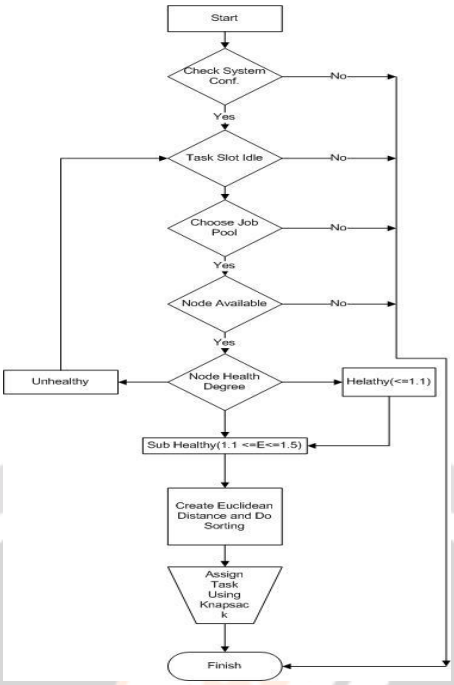

Figure 3 Proposed System Flow Chart

2.1 ALGORITHM OF PROPOSED SYSTEM

1) When an idle task slot can be used, a pool is chosen according to the Fair scheduler;

2) If the node on which the idle task is located is healthy, then go to 5;
3) If the node on which the idle task is located is sub healthy, characteristic variables of the node are separately multiplied by the reciprocal of Health Degree of node, then go to 5;
4) if the node on which the idle task is located is unhealthy, then go to 8,and set timer on node ,its Health Degree is multiplied by 0.9 every x minutes until it reaches the state of sub healthy;
5) Computing Euclidean distances between characteristic variables of job task and characteristic variables of node, sorting them from small to large;
6) Choosing an appropriate job according to Delay Scheduler in the sorted set;
7) Assigning the task of a chosen job to a task slot to run;
8) Finish.
In step 4, generally the timer interval of unhealthy node is set to 5, when the node Health Degree reaches sub health, the node can be schedule again. In step 5,

In step 6, Delay Scheduler is used so that the algorithm achieves higher locality.

3.1 RESULTS AND EXPECTED OUTCOMES

To get results for proposed system we need to check system configuration first if it's configured than we need to have Linux operating system for execution of Hadoop files. After successful installation of Hadoop and it's jar files you need to make java programs for scheduling, mapping, reducing time allocation after getting java files go to Hadoop environment in select your instance to execute and get output of uploaded file.

Overcoming disadvantage of MapReduce and getting Fault tolerance High scalability with the use of fair scheduler and you can use knapsack algorithm for better output.

Table -1: Expected Outcome

Name	Map Total Time(ms)	Map Time(ms)	Map Combine time(ms)	Quick Sort Time(ms)	Execution Time(ms)
Map/Reduce job	84824.61	56535.23	2096.37	1122.93	969.53
Map/Reduce Job	717209.12	598814.12	141745.31	102390.23	6803.11

4. CONCLUSIONS

In this examination going to execute enhanced Fair4s calculation in which change in its adaptation to non-critical failure, Based on wellbeing degree undertakings are relegated. Additionally the examination of Fair4s on various sorts of burden, Data of at whatever time comes in database and checking proficiency and diminishing fall flat rate of occupation. Our hubs are chosen by ascertaining weight of burden and assignments are given by measuring wellbeing degree and until hubs are coming in sub sound degree errands are not allocated to that hub. Insatiable backpack is tasking for quicker and better designation of employments to hubs. Along these lines we are expanding effectiveness of booking and diminishing adaptation to internal failure

6. REFERENCES

[1]Sosinsky, Barrie. Cloud computing bible. Vol. 762. John Wiley & Sons, 2010Manyika, James, et al. "Big data: The next frontier for innovation, competition, and productivity." (2011)
[2]Miller, Michael. Cloud computing: Web-based applications that change the way you work and collaborate online. Que publishing, 2008.
[3]Manyika, James, et al. "Big data: The next frontier for innovation, competition, and productivity." (2011).
[4]Guo, Yingjie, et al. "The Improved Job Scheduling Algorithm of Hadoop Platform." arXiv preprint arXiv:1506.03004 (2015).
[5]Harshitha, R., and G. S. Rekha. "A Survey on Scheduling Techniques in Hadoop." International Journal of Engineering Development and Research. Vol. 3. No. 1 (Jan 2015). IJEDR, 2015.
[6]Daneshyar, Samira, and Majid Razmjoo. "Large Scale data Processing Using Mapreduce in Cloud Computing Environment." International Journal on Web Service Computing (IJWSC) 3.4 (2012): 1-13.
[7]Gautam, Jyoti V., et al. "A survey on job scheduling algorithms in Big data processing." Electrical, Computer and Communication Technologies (ICECCT), 2015 IEEE International Conference on. IEEE, 2015.Brown, L. D., Hua, H., and Gao, C. 2003.
[8]Guo, Yingjie, et al. "The Improved Job Scheduling Algorithm of Hadoop Platform." arXiv preprint arXiv:1506.03004 (2015)..
[9]http://www.sas.com/en_us/insights/big-data/what-is-big-data.html
[10]http://searchcloudcomputing.techtarget.com/definition/big-data-Big-Data
[11]http://searchcloudcomputing.techtarget.com/definition/Hadoop
[12]http://computer.howstuffworks.com/internet/basics/google-file-system.htm
[13] http://www.tutorialspoint.com/hadoop/hadoop_hdfs_overview.htm
[14] http://gppd-wiki.inf.ufrgs.br/index.php/MapReduce