Abstract

Hadoop MapReduce is the community accepted platform that deals with the gigantic data in an efficient and cost-effective manner. To cope up with ever growing datasets and shrinking time to analyze them, Hadoop MapReduce leveraged parallelize computations on large distributed clusters consisting of many machines. Careful consideration of the factors affecting the Hadoop MapReduce can enhance its performance. Many researches has been done for improving the total job execution time of MapReduce by optimizing different parameters. The replication factor is still unexplored for its effect on the MapReduce job completion time. This paper focuses on the evaluation of data replication factor on MapReduce job completion time using regression analysis. The performance of the Hadoop MapReduce job in terms of total job completion time is monitored experimentally by changing different values of replication. The evaluation results evidently shows the dependence of the job completion time on the replication factor. The dependence of total job completion time on the replication has been verified both analytically and experimentally.
References

associative classification of big data. Information Sciences, 332, 33-55.


**Index Terms**

Computer Science  
Information Sciences

**Keywords**

Hadoop MapReduce, Big Data, Regression Analysis, Data Replication, Job optimization