Abstract

Data analysis is an important functionality in cloud computing which allows a huge amount of data to be processed over very large clusters. Hadoop is a software framework for large data analysis. It provide a Hadoop distributed file system for the analysis and transformation of very large data sets is performed using the MapReduce paradigm. MapReduce is known as a popular way to hold data in the cloud environment due to its excellent scalability and good fault tolerance. Map Reduce is a programming model widely used for processing large data sets. Hadoop Distributed File System is designed to stream those data sets. The Hadoop MapReduce system was often unfair in its allocation and a dramatic improvement is achieved through the Mapper Reducer System. The proposed Mapper Reducer function using the mean shift clustering based algorithm allows us to analyze the data set and achieve better performance in executing the job by using optimal configuration of mappers and reducers based on the size of the data sets and also helps the users to view the status of the job and to find the error localization of scheduled jobs. This will efficiently utilize the performance tuning properties of optimized scheduled jobs. So, the efficiency of the system will result in substantially lowered system cost, energy usage, management complexity and increases the performance of the system.
Performance Tuning and Scheduling of Large Data Set Analysis in Map Reduce Paradigm by Optimal Configuration using Hadoop

References

- Apache, "Hadoop, http://hadoop.apache.org/docs/r0.20.2/hdfs_design.html"

**Index Terms**

Computer Science  
Information Sciences

**Keywords**

Cloud Computing  
Hadoop Distributed file System  
Performance Tuning  
Mean shift  
Clustering

Amazon web services