Parallel algorithm in Cloud computing

Guo Weiwei and Liu Feng

Heilongjiang University of Technology, Jixi 158100, china,
gwwguoweiwei@163.com, Liufeng8038@163.com

Abstract. With the rapid development of the 3G network, traditional calculation methods are unable to adapt to the data scene that telecom users' Network access behavior's data scale increase rapidly dozens of TB. The cloud techniques such as Hadoop platform are introduced to solve the data storage problem.

Keywords: 3G network, parallel computing, cloud computing

1 Introduction

In the field of telecommunications, technical difficulties sprang up regarding the storage, processing and discovery of enormous data due to rapid increase of communication services and network visiting traffic [1-2]. Right at the moment, lots of new hi-tech computing service models led by cloud computing were successively proposed. They were practically applied by many telecom enterprises home and abroad for business analysis and management activities, forming the distributed mass data architecture based on computer clustering systems [3-4]. In order to get from thousands of TB-scale massive data the business knowledge which is demanded for enterprise users, it generally requires tremendous and complicated mining operations [5-6]. To achieve that goal, the cloud computing platforms were designed with hyperscale computational capability and super strong storage capacity. The typical one is Hadoop platform, which is particularly adopted by China Unicom in our country [7-8].

2 Design and Implementation of Hadoop Platform

The design ideas are shown in Fig 1. To be specific, with the use of clustering computing characteristics of Hadoop, huge data mining tasks are evenly distributed to every single computing node in the clustering system to enhance efficiency and availability of explored knowledge through parallel computation; meanwhile it gives full play of Hadoop’s strong data storing and processing abilities. In the lower levels, we make use of Hadoop’s strengths for data storage and analysis. In the top levels, we invoke transparently relative modules with the help of functional interfaces.
3 Experiment Design and Discussion

3.1. Analysis of rationality

In the design process of the proposed parallel algorithm, by depending on Hadoop platform, the workflow of Sprint algorithm was designed completely and that the concurrent Sprint algorithm was successfully migrated to Hadoop platform. There are following steps:

In different phases, it’s important to keep the parallelized design in the Map process as it affects directly the file processing speed and size of data blocks. Moreover, most time of the algorithm was consumed in the process. But between Reduce and separation process, there is certain inheritance relationship. The paralleling degree is not high so that it’s no need to implement specific parallelization design.

In the parallelized design process, a few Maps are concurrently read in. Then, data are passed to Reduce. For the attribute of continuous values, which are all sequenced for processing, it’s no need for Sprint algorithm to make global ordering at big costs. For giant data volume, the method’s paralleling advantage will more observable.

The design of parallel algorithm takes into full account the own features and superiority of enormous computation of Hadoop platform. Together with the
computational power of the framework, the scalability and efficiency of the proposed method are both ensured to the maximum limit.

3.2. Analysis of effectiveness

From the experimentation above, we can transfer all data sets to attribute lists according to the expression pattern of trees. New attribute tables can be created by binding such tables with corresponding nodes. Then according to the hierarchical structure of trees, mutually separate folders are formed for storing attribute tables and node information in each level.

This experiment shares some nodes, as shown in figure 2.

Fig. 2. Diagram of some nodes storage

4 Conclusion

This paper firstly analyzes and introduces the framework and designs model of Hadoop platform, expounds the characteristics of cloud data mining platform, on the basis of this, the traditional SPRINT parallelize algorithms have been successfully designed.
References

1. Yanan, Z.: Research on parallel clustering algorithm Hadoop based on cloud computing platform. Inner Mongolia University of Science and Technology, 2013
2. Guang, G. J.: Research on key technology of data mining for telecom field. Harbin Engineering University, 2012
5. Shilei, L.: Realization of Hadoop platform and the random forest algorithm research based on image classification system. Xiamen University, 2014