A Novel Scheduling Algorithm based on Clustering Analysis and Data Partitioning For Big Data

Weiqi Cui, Nan Liu, Yihuan Dong, Jiaqi Li, Qingchen Zhang*

School of Software Technology, Dalian University of Technology , Liao Ning Province, China
854350345@qq.com, nanliudlut@gmail.com, yihuandong@gmail.com, lijiaqi@gmail.com, 623759909@qq.com

Abstract – With the development of the computer technology and network technology, the size of data collected is increasing rapidly. It is difficult to process and analyze so big data in real-time. Cloud computing is an effective tool for real-time processing for big data. How to make full use of cloud computing to analyze big data is an hot issue in recent year. Because of the limit of current network transmission speed, there is much communication between the selected cloud computing nodes during big data processing, which will cause a heavy transmission delay and lead to the reduce of real-time. The current methods select cloud nodes for data processing depend on the quality of services of nodes, which cannot reflect the relations between different cloud nodes. To address this problem, the paper proposes a novel scheduling algorithm based on clustering analysis for big data analysis in cloud environment. The presented method divides the cloud nodes into clusters according to communication cost between different nodes, and then selects a cluster for the big data analysis services. Experimental results show the effectiveness of our scheduling algorithm.

Index Terms - big data, clustering analysis, cloud computing.

1. INTRODUCTION

In recent years, with the rapid development of the Internet of things and electronic commerce, the amount of the collected data is increasing with unprecedented speed[1-5]. We have entered the era of big data. The characteristics of big data include volume, variety and velocity. Many factors contribute to the increase in data volume – transaction-based data stored through the years, text data constantly streaming in from social media, increasing amounts of sensor data being collected, etc. Data today comes in all types of formats – from traditional databases to hierarchical data stores created by end users and OLAP systems, to text documents, email, meter-collected data, video, audio, stock ticker data and financial transactions. By some estimates, 80 percent of an organization's data is not numeric! According to Gartner, velocity "means both how fast data is being produced and how fast the data must be processed to meet demand." RFID tags and smart metering are driving an increasing need to deal with torrents of data in near-real time. How to make full use of cloud computing to analyze big data is an hot issue in recent year [6].

The advent of Cloud Computing[7] is an effective tool for big data analyticst. Cloud computing is Internet-based computing, where shared resources are provided to users on-demand, like a public utility. Cloud applications are usually large-scale and very complex, involving a number of distributed cloud nodes. When deploying a cloud application in a cloud, the application user need to select a number of cloud nodes including servers and virtual machines to run the cloud applications. How to make optimal deployment of cloud applications is a challenging and urgent required research problem.

To select the optimal cloud nodes for deployment purpose, the current methods usually rank the available cloud nodes based on their QoS values and select the best performing ones. The major approaches can be divided into three types: (1) Random appraoches (use random methods to select components). Random strategies have been employed in BOINC [4]. (2) Ranking or rating approaches (cloud nodes is ranked by the order of QoS performance). Ranking strategies have been employed in RIDGE [16] and GridEigenTrust [17]. (3) Matching approaches (matching algorithms are employed to compare the users' requirements and the QoS values of cloud nodes). Matching strategies have been employed in Condor [18]. These previous methods just consider the order of the node performance, and not consider the relationship between nodes, which is important to the communication-intensive cloud applications. In the cloud environment, there are usually a lot of available cloud nodes. When selecting optimal cloud nodes from a set of available cloud nodes for deployment purpose, ranking-based method [6] ranks the available nodes based on their QoS values and selects the best performing ones. A drawback of the ranking methods is that these methods cannot reflect the relations between different cloud nodes. Therefore, such kind of methods cannot be applied to the communication-intensive cloud applications, whose performance is greatly influenced by the communications between the nodes in the application.

To address the above problem, the paper proposes a novel scheduling algorithm based on clustering analysis. The presented method divides the cloud nodes into clusters according to communication cost between different nodes, and then selects a cluster for the big data analysis services. Experimental results show the effectiveness of our scheduling algorithm.

2. THE SCHEDULING ALGORITHM BASED ON CLUSTERING ANALYSIS

There are a number of available distributed nodes in the cloud. Cloud user need to deploy their cloud applications on a number of optimal cloud nodes and use it. Since cloud nodes

* Corresponding author

are usually distributed in different geographic locations, deploying the application on different set of nodes may obtain different level of quality. How to select a subset of optimal cloud nodes to satisfy the requirement of cloud user is an important research problem.

To attack this critical challenge, we divide the cloud nodes into clusters based on clustering method, making the communication between nodes in the same cluster smallest. Further, in order to meet the requirements of load balancing, the computing capacity of the nodes in the same cluster is similar. we use the response time between two nodes to represent the distance between them. If a node has lower response times to other nodes, it means that the node has shorter distances to the other nodes.

Assume there are n cloud nodes distributed in a cloud, the response times between nodes can be represented as an n by n matrix, where p_{ij} is the response time between node i and node j. Apparently, this matrix is a symmetric matrix, in other words, $p_{ij} = p_{ji}$.

$$P = \begin{pmatrix} 0 & p_{12} & \cdots & p_{1n} \\ p_{21} & 0 & \cdots & p_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ p_{n1} & p_{n2} & \cdots & 0 \end{pmatrix}$$

We use p_i to represent the vector of response times from node i to other nodes. i.e., $p_i = (p_{i1}, p_{i2}, ..., p_{i,i-1}, p_{i,i+1}, ..., p_{i,n})$.

A cluster analysis algorithm is designed to divide the cloud nodes into K clusters, $C = \{C_1, C_2, ..., C_K\}$ .They satisfy the following condition.

$$\begin{cases} C_i \neq \varnothing \quad i = 1, 2, ..., K \\ C_i \cap C_j = \varnothing \quad i, j = 1, 2, ..., K \text{ and } i \neq j \\ \bigcup_{i=1}^{K} C_i = D \end{cases}$$

The following formula is used to calculate the distance D between a node to the centroid of the k th clustering.

$$D = \frac{1}{d} \sum_{j:j \in C_k} p_{ij}$$

Where, cal_{ck} denotes the computing capacity of the k th clustering.

After the completion of the clustering, the nodes in the cloud platform are divided into several clusters. And then we select one of the clusters to perform computing tasks. There are many nodes in a cluster, so the algorithm intends to parallelize the big analysis task by dividing input data for multiple computing nodes in the same cluster. The paper partitions the data based on the load balancing so that each node has the same processing delay, including communication delay and calculate the delay, to avoid the reduction of the real-time because of the delay of single node.

3. EXPERIMENT

The data used in our experiment collect from the IoT lab, including three sets of data: temperature, humidity, and carbon dioxide concentration in the digital home laboratory and the digital greenhouse. In the laboratory, we build a cloud computing simulation platform, including a scheduling node and 10 compute nodes. To simulate a heterogeneous cloud environment, there are 7 virtual nodes and 3 computing nodes in our computing platform.

We compare our presented method with the other scheduling approaches including the random-base scheduling algorithm and QoS-based algorithm. The experimental result is shown as the following figure.

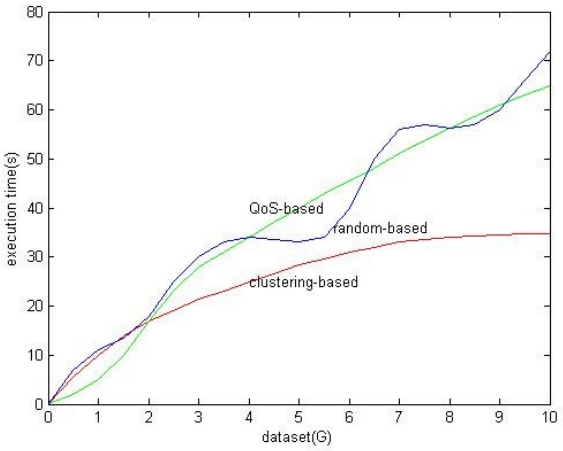

Fig. 1 algorithm comparison result

It can be seen that the running time of the three methods increase with the increasing amount of the data. In the case of a small amount of data, the time spending on computing is more than on the communication, so in this case, the performance of the QoS-based scheduling is better than other algorithm. However, with the increasing amount of data, the cost of communication between nodes increases, the performance of the proposed scheduling method is better than the QoS-based scheduling method, because the proposed scheduling method takes into account computing cost and communication cost in the same time, which show the effectiveness of the proposed algorithm in the real-time of big data analysis.

4. INCLUSION

With the rapid development of the Internet of Things and electronic commerce, the era of big data has come. Real-time processing is a key problem for big data analysis and processing. The cloud computing is an effective tool to process big data in real time. In this paper, a novel scheduling algorithm for big data in cloud environment is proposed, which is based on clustering analysis. The presented method divides the cloud nodes into clusters according to communication cost between different nodes, and then selects a cluster for the big data analysis services. Experimental results show the effectiveness of our scheduling algorithm.

REFERENCES

[1] Atzori L, Iera A, Morabito G. The internet of things: A survey[J]. Computer Networks, 2010, 54(15): 2787-2805.

[2] Turban E, Lee J K, King D, et al. Electronic commerce 2010[M]. Prentice Hall Press, 2009.

[3] Kortuem G, Kawsar F, Fitton D, et al. Smart objects as building blocks for the internet of things[J]. Internet Computing, IEEE, 2010, 14(1): 44-51.

[4] Welbourne E, Battle L, Cole G, et al. Building the internet of things using RFID: the RFID ecosystem experience[J]. Internet Computing, IEEE, 2009, 13(3): 48-55.

[5] QIN Xiong-Pai, WANG Hui-Ju, DU Xiao-Yong, WANG Shan. Big Data Analysis—Competition and Symbiosis of RDBMS and MapReduce. 2012,23(1): 32-45.

[6] http://www.sas.com/big-data/.

[7] Iosup A, Ostermann S, Yigitbasi M N, et al. Performance analysis of cloud computing services for many-tasks scientific computing[J]. Parallel and Distributed Systems, IEEE Transactions on, 2011, 22(6): 931-945.