MASSIVE DATA BALANCE SCHEDULING IN CLOUD COMPUTING ENVIRONMENT

Xiuran Wei[1], Feng Wang[2]
[1]College of Information and Management Science
Henan Agricultural University, Zhengzhou, 450046, China
[2]College of Software
North China University of Water Resources and Electric Power
Zhengzhou, 450045 China
3398849494@qq.com

Abstract - At present, in the cloud computing environment, the massive data has different attribute characteristics, which cannot be made balance scheduling. There are often long scheduling time, and the balance of CPU, memory and bandwidth is poor. Based on the improved clonal selection algorithm, the balance scheduling method for massive data in cloud computing environment is proposed. Firstly, the massive data balance loading and scheduling, as well as the minimized task execution time are taken as the target, to construct the data scheduling model in the cloud computing environment. According to the principle of clonal selection, the attributes of the massive data in the cloud computing environment are defined as antigens, and the antibody encoding mode of the massive data balance scheduling is designed. Meanwhile, the individual data with the higher affinity to the antigens are selected from the antibody for mutation treatment, and the balance of massive data scheduling is quantified in cloud computing environment. After obtaining the quantization function, the control parameters of massive data scheduling are analyzed and studied, so as to build a balance scheduling model for massive data in cloud computing environment. The simulation results show that the proposed algorithm can make effectively balance scheduling for the massive data in the cloud computing environment, its scheduling time is short, the balance of CPU, memory and bandwidth is higher and the reliability is strong.

Keywords: Massive Data; Cloud Computing; Clonal Selection Algorithm; Balance Scheduling.

1. Introduction

At present, with the continuous improvement of the level of computer science and technology, cloud computing technology occupies a central position in the field of IT [1-3].

The advantages of cloud computing are generally reflected by the degree of their tasks operation. Therefore, how to efficiently schedule large amounts of data in cloud computing environment is the main problem to be solved in this field [4, 5]. As the balance scheduling method for the massive data in cloud computing environment has far-reaching significance, it has become the focus of research in the industry, and has been widespread concern, but also a lot of good methods are appeared [6, 7]. In article [8], a classification and optimization scheduling method for massive data in cloud computing environment based on Bayesian theory is proposed. This method realizes the classification and optimization scheduling of the massive data in the cloud computing environment by using the statistical probability of the cloud computing task scheduling. The method is simple, but there is a problem that the load balancing cannot be considered in a short time when the classification and optimization scheduling is carried out. In article [9], the classification and optimization scheduling method for massive data in cloud computing environment based on ant colony algorithm is studied emphatically. According to the principle of ant foraging, the massive data classification and optimization scheduling in cloud computing environment is finished. The method has the characteristics of global optimal solution, but there is a drawback that the search time is long. In article [10], the classification and optimization scheduling method for the massive data in the cloud computing environment based on the time-first algorithm is adopted. This method can effectively perform the classification and optimization scheduling of the massive data in the cloud computing environment by calculating the execution time of each scheduling task. The scheduling efficiency of this method is high, but there are disadvantages of greater cost. In article [11], a balance scheduling and modeling method of massive

data in cloud computing environment based on particle swarm optimization (PSO) is proposed.

In this method, the clustering algorithm is used to cluster the massive data loading, and the massive data which are overloaded are migrated to realize the loading balance of the massive data in the cloud computing environment. The method is stable, but there is a problem that the calculation is too cumbersome and time consuming.

Aiming at the above problems, this paper proposes a balance scheduling method for massive data in cloud computing environment. Firstly, the massive data balance loading and scheduling, as well as the minimized task execution time are taken as the target, to construct the data scheduling model in the cloud computing environment. According to the principle of clonal selection, the attributes of the massive data in the cloud computing environment are defined as antigens, and the antibody encoding mode of the massive data balance scheduling is designed. Meanwhile, the individual data with the higher affinity to the antigens are selected from the antibody for mutation treatment, and the balance of massive data scheduling is quantified in cloud computing environment by combining multi-objective optimization algorithm. According to the different attribute characteristics of massive data, the corresponding distribution strategy is put forward, so as to construct the massive data balance scheduling model in the cloud computing environment.

2. Modeling Principle of Massive Data Balance Scheduling in Cloud Computing Environment

In the process of massive data balance scheduling in cloud computing environment, the data resources in the cloud environment should be classified according to their attribute information, and the load balance of the cloud computing system should be guaranteed before scheduling the resources after classification [12, 13]. It is also possible to summarize the scheduling steps of the massive data in the cloud computing environment as: the cloud resources are classified firstly according to the level before scheduling, and the massive data balance loading and scheduling, as well as the minimized task execution time are taken as the target, which are combined with the different types of cloud resources and task scheduling of the implementation time, and thus to effectively complete the massive data balance scheduling. The specific steps are as follows:

In the process of massive data balance scheduling and modeling in the cloud computing environment, if the data resource represented by r_i can satisfy the task scheduling requirement represented by ts_j, that is, the condition of $\left(r_i, ts_j\right) \geq 0$, then ts_j can be allocated to the resource r_i and carry out. There will be a number of data resources r_i to meet the task ts_j in the conventional state. In order to achieve the load balancing of cloud computing system, the task ts_j will send the p_{ij} class attribute data to the resources r_i. The use of the following formula can express the attribute matrix of system task scheduling in the cloud environment:

$$p_{n\times m} = \frac{\left(p_{ij}\right)_{n\times m}}{\left(r_i, ts_j\right) \geq 0} \tag{1}$$

In the process of massive balance scheduling and modeling under the cloud computing environment, according to the following formula, the balance scheduling model of massive data can be established:

$$S_{n\times m} = p_{n\times m} \times \left(x_{ij}\right)_{n\times m} \tag{2}$$

In the above equation, $\left(x_{ij}\right)_{n\times m}$ represents the task's production rate vector.

In summary, it can be described that the modeling principle of massive data balance scheduling in cloud computing environment is effective, and the massive balance scheduling model in cloud computing environment is established effectively.

3. Optimization Principle of Massive Data Balance Scheduling in Cloud Computing Environment

Aiming at the problem of that load balancing is not fully taken into account and scheduling error is large when using the current algorithm to establish the massive data balance scheduling model in the cloud computing environment [14, 15], a balance scheduling method for massive data in cloud computing environment based on the improved clonal selection algorithm is proposed. According to the principle of clonal selection, the attributes of the massive data in the cloud computing environment are defined as antigens, and the antibody encoding mode of the massive data balance scheduling is designed. Meanwhile, the individual data with the higher affinity to the antigens are selected from the antibody for mutation treatment, and the balance of massive data scheduling is quantified in cloud computing environment [16]. According to the different attribute characteristics of massive data, the corresponding distribution strategy is put

forward, so as to construct the massive data balance scheduling model in the cloud computing environment.

- ***Variation process analysis of individual data***

In the process of optimizing the modeling of massive data balance scheduling in the cloud computing environment, the model of the data resource balance scheduling in the cloud computing environment is established with the aim of minimized task execution time and the load balancing factor. According to the principle of clonal selection, the attributes of the massive data in the cloud computing environment are defined as antigens, and the antibody encoding mode of the massive data balance scheduling is designed. Meanwhile, the individual data with the higher affinity to the antigens are selected from the antibody for mutation treatment. The specific steps are as follows: in the process of optimization modeling the massive data equalization scheduling in the cloud computing environment, the massive data load parameters represented by l_j are divided into the CPU utilization, memory utilization, memory utilization, network bandwidth, IO access rate and the total number of processes of the system represented by C_j, M_j, N_j, IO_j and P_j respectively, then the use of the following formula can express the massive data load in the cloud environment:

$$l_j = \frac{C_j w_1 + M_j w_2 + N_j w_3 + IO_j w_4 + p_j w_5}{G = \{R, T, E\}} \tag{3}$$

In the above formula, w_1, w_2, w_3, w_4, w_5 represents the weight of the massive data load parameter respectively, and which needs to meet the following conditions:

$$w_1 + w_2 + w_3 + w_4 + w_5 = \frac{1}{l_j} \tag{4}$$

In the optimization modeling process of the massive data balance scheduling under cloud computing environment, the implementation time of the task set T in the massive data set R is defined as $makespan$ which represents the task execution time span. t_s and t_e respectively represent the time of task t to start the task and end the task on the massive data r, the process of task implementation can be represented by $Sch = \{t, r, t_s, t_e\}$. Then the following formula can be used to calculate $makespan$:

$$makespan = \{t, r, t_s, t_e\} \times \frac{\sqrt{\max\{t_{ei}\} - \min(t_{sj})} \times Sch}{\left[w_1 + w_2 + w_3 + w_4 + w_5 = \frac{1}{l_j} \right]} \tag{5}$$

In the above equation, $\max\{t_{ei}\}$ represents the latest end time of all tasks in the cloud computing environment, and $\min(t_{sj})$ represents the earliest start time of all tasks.

In the process of optimizing the modeling of massive data balance scheduling in the cloud computing environment, according to the principle of clonal selection, the massive data balance scheduling in the cloud computing environment are defined as antigens, and E representing the matrix element in the massive data distribution set is defined as an antibody gene, then the following formula can be obtained:

$$P = \frac{p_n}{E \oplus [makespan]} \tag{6}$$

In the above equation, n represents the total number of tasks, and P_n represents the massive data assigned to the task t_i.

In the process of optimizing the modeling of massive data balance scheduling in the cloud computing environment, the optimal solution of the massive data balance scheduling is needed to measure the time span and the load balance of the massive data. Therefore, the affinity function of antibody and antigen can be calculated by the following formula:

$$Fit(p) = \frac{1 \times p}{makespan} \tag{7}$$

In the above formula, $Fit(p)$ represents the affinity function of antibody and antigen.

In the process of optimizing the modeling of massive data balance scheduling in the cloud computing environment, assuming that the scale population of antibody is K represented by $A = \{P_1, P_2 \ldots P_K\}$, the affinity of any two antibodies can be obtained by the following formula:

$$t(p_k) = Fit(p) \times \frac{1}{A \times k} \tag{8}$$

In the above formula, $t(p_k)$ represents the affinity of any two antibodies.

In the process of optimizing the modeling of massive data balance scheduling in the cloud computing

environment, the individuals with high affinity for antigen are selected from the antibodies:

$$H_j(K)=\frac{-\sum_{i=1}^{K} p_i(j)\lg p_i(j)}{t(p_k)} \tag{9}$$

In the above formula, $p_i(j)$ represents the probability of the one which is same as the j-th gene of antibody p_i, and $H_j(K)$ represents the individual with higher affinity for the antibody after mutation.

In summary, we can show that in the process of optimizing the modeling of massive data balance scheduling in the cloud computing environment, the minimized task execution time and the load balancing factor are taken as the target firstly, to construct the data scheduling model in the cloud computing environment. According to the principle of clonal selection, the massive data balance scheduling in the cloud computing environment are defined as antigens, and the antibody encoding mode of the massive data balance scheduling is designed. Meanwhile, the individual data with the higher affinity to the antigens are selected from the antibody for mutation treatment, and the balance of massive data scheduling is quantified in cloud computing environment, to lay the foundation for establishing the optimization model of massive data balance scheduling under the cloud computing environment.

- ***Realization of massive data balance scheduling in cloud computing environment***

In the cloud computing environment, when the data traffic demand exceeds the network bandwidth, the network signal will collide and conflict, the data will accumulate in the node to cause congestion [17, 18]. Therefore, in the process of classification optimization scheduling of massive data in the cloud computing environment, the execution value of each task on all resources is calculated on the basis of individual data variation, that is, the product of the corresponding resource level of each scheduling task and the minimum execution time. The optimal scheduling of the massive data in the cloud computing environment is achieved using the task with the smallest product [19-24].

In the process of classification optimization scheduling of massive data in the cloud computing environment, the balance of massive data scheduling in cloud computing environment is quantified on the basis of individual data variation. Assuming the data set is X_n, then:

$$X_n=\{x_1,x_2,x_3.....x_n\} \tag{10}$$

The quantification model of the obtained data is:

$$x_i(n+1)=\frac{1}{2}\left[x_i(n)H_j(K)-\left|x_i(k)-T(k)\right|\right] \tag{11}$$

Where x is the time series of load sampling point, T is the communication signal delay, and then the corresponding distribution strategy for the different attribute characteristics of massive data is put forward, so as to construct the model of massive data balance scheduling under cloud computing environment.

Then the time span of data operation can be expressed as:

$$t=\sum_{i=1,j=1}^{n} T(x)E(i,j) \tag{12}$$

Where $E(i,j)$ is the efficiency function of data signal, $e(i,j)$ is the specific form of scheduling model function:

$$E(i,j)=\begin{cases}\dfrac{e_{ij}-e(i,j)}{e_{\max}-e(i,j)}\\ \dfrac{e_{ij}-e(i,j)}{e(i,j)e_{\min}}\end{cases} \tag{13}$$

The appropriate objective function of massive data scheduling is selected. Let ω be the weight of the data classification attribute, A be the weighted constraint equalization ratio, M be the sampling period, and then the balance parameter of data scheduling τ is solved as follows:

$$\tau=\omega_1 A_i t+\omega_2 T_i+\omega_3 M_i \tag{14}$$

According to the parameter of data balance scheduling, the corresponding delay response parameter ς and the data flow control parameter ψ can be selected and determined. The parameter solving process is:

$$\varsigma=\frac{\tau R_1 R_2 \chi}{4\pi\chi h^2}\sum_{i=1}^{n}\kappa(1-\sqrt{\frac{1}{2}\sin\beta\cos\alpha\sin\varphi}) \tag{15}$$

$$\psi=\frac{\tau\sqrt{R_1 R_2\phi}}{2\pi}\sum_{i=1}^{n}\kappa(1-\sqrt{\frac{1}{2}\sin\beta\cos\alpha}) \tag{16}$$

Thus it obtains the model of massive data balance scheduling in the cloud computing environment:

$$f(x)=\sum_{i=1}^{n}\min(x_i^2,\tau^2)-\sum_{i=1}^{n}(x_i)\frac{\varsigma^2}{2t}-t\sum_{i=1}^{n}(x_i)\psi^2 \tag{17}$$

4. Experimental Simulation

In order to prove the validity of the modeling method for massive data balance scheduling in the cloud computing environment based on the improved clonal selection algorithm, an experiment is carried out. Cloud computing simulation tool *CloudSim* is used to establish a experimental simulation platform of the massive data balance scheduling under cloud computing environment. In the course of the experiment, there are 38 massive data to set up cloud computing resource pool, on this basis, the scheduling set composed by the independent task is made scheduling, and the number of tasks is between 29 to 301. The algorithm in literature [8] and [9], and the improved algorithm is used to simulate the massive data balance scheduling in the cloud computing environment, the parameters of the experiment are set as follows: 29-301 is the length of the antibody, $K = 99$ is the size of the antibody population, $m = 39$ is the total number of massive data, $p_m = 0.5$ represents the mutation probability, $t = 1$ represents the number of current iteration, and $t_{\max} = 199$ represents the maximum number of iterations.

- ***Comparison of task execution time span of different algorithms***

In this paper, we use the improved algorithm, the algorithm in literature [8] and the literature [9] to carry on the massive data balance scheduling experiment under the cloud computing environment, and compare the task execution time span and the resource balance factor obtained by the three algorithms.

The comparison results are shown in Fig. 1 and Fig. 2.

As can be seen from Fig. 1, the task execution time span of the improved algorithm is the smallest, and although the number of tasks is increasing, the average value has good stability.

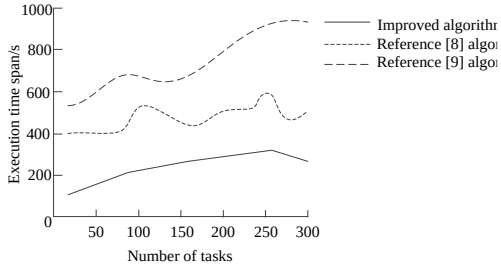

Figure 3: Comparison of CPU balance by Different Algorithms

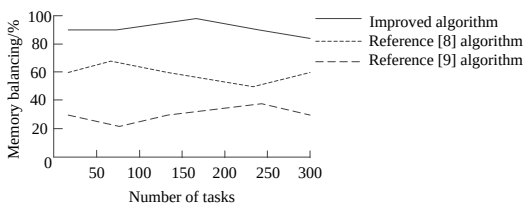

Figure 4: Comparison of memory balance by different algorithms

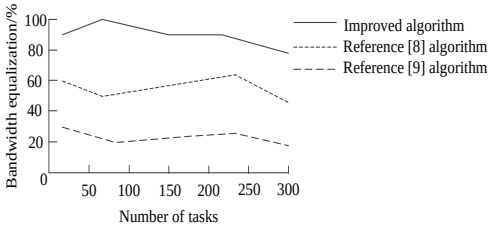

Figure 5: Comparison of bandwidth balance by different algorithms

It can be shown in Figure 3, Figure 4 and Figure 5 that the improved algorithm has more obvious load balancing advantage than other algorithms because the improved algorithm quantifies the balance of the massive data in the cloud computing environment. The control parameters of the massive data of various attributes in the environment are analyzed, and the load balancing of the improved algorithm is ensured. Experimental results show that the method of massive data balance scheduling based on improved clonal selection algorithm is highly efficient and reliable in cloud computing environment.

5. Conclusions

Aiming at the problem of load balancing of resource scheduling is not fully considered and scheduling error is large when using the current algorithm to establish the model of massive data balance scheduling in the cloud computing environment. In order to solve the problem, in this paper, a modeling method for massive data balance scheduling in cloud computing environment based on improved clonal selection algorithm. Firstly, the minimized task execution time and load balancing factor are taken as the target, to construct the data scheduling model in the cloud computing environment.

According to the principle of clonal selection, the attributes of the massive data in the cloud computing environment are defined as antigens, and the antibody encoding mode of the massive data balance scheduling is designed. Meanwhile, the individual data with the higher affinity to the antigens are selected from the antibody for mutation treatment, and on this basis, the balance of massive data scheduling in cloud computing environment is quantified. After obtaining the quantization function, the control parameters of massive data scheduling are analyzed and studied, so as to build a balance scheduling model for massive data in cloud computing environment. The simulation results show that the proposed algorithm can make effectively balance scheduling for the massive data in the cloud computing environment, its scheduling time is short, the balance of CPU, memory and bandwidth is higher and the reliability is strong.

Acknowledgements

Key scientific and technological research projects of Henan Province - A meteorological data reconstruction system based on compressed sensing theory (No. 152102210112);
Key project of science and technology research of Henan Provincial Department of Education - Research and application of meteorological data set processing based on compressed sensing (No. 13A520713).

References

[1] Ratten V. International Consumer Attitudes toward Cloud Computing: A Social Cognitive Theory and Technology Acceptance Model Perspective. Thunderbird International Business Review, 2015, 57(3): 217-228.
[2] Mulfari D, Celesti A, Villari M, et al. Providing Assistive Technology Applications as a Service Through Cloud Computing. Assistive Technology the Official Journal of Resna, 2015, 27(1): 44-48.
[3] Ge L, Wang S, Ge X. Framework design of cloud computing technology application in power system transient simulation // Power and Energy Engineering Conference. IEEE, 2015, 1-6.
[4] Zhang Y. Web Data Mining Technology on Cloud Computing. Applied Mechanics and Materials, 2014, 543(47): 3490-3493.
[5] Zi Y, Gao J, Sun Y, et al. Visualized Data Processing Technology under Cloud Computing // Eighth International Conference on Measuring Technology and Mechatronics Automation. IEEE, 2016, 905-909.
[6] Visnjevic V, Herman F, Licul A. Insight into glacier climate interaction: reconstruction of the mass balance field using ice extent data // EGU General Assembly Conference. EGU General Assembly Conference Abstracts, 2016.
[7] Gordon M, Li S M, Staebler R, et al. Determining air pollutant emission rates based on mass balance using airborne measurement data over the Alberta oil sands operations. Atmospheric Measurement Techniques, 2015, 8(5): 4769-4816.
[8] Chen H, Zhu X, Guo H, et al. Towards energy-efficient scheduling for real-time tasks under uncertain cloud computing environment. Journal of Systems and Software, 2015, 99(2): 20-35.
[9] Zuo L, Shu L, Dong S, et al. A Multi-Objective Optimization Scheduling Method Based on the Ant Colony Algorithm in Cloud Computing. IEEE Access, 2015, (19)3: 2687-2699.
[10] Wang Q. An Analysis on Effective Classification Method for Massive Data in Cloud Computing Environment. Applied Mechanics and Materials, 2014, 51(17): 2315-2319.
[11] Zhang Y, Gong D, Hu Y, et al. Feature selection algorithm based on bare bones particle swarm optimization. Neurocomputing, 2015, 148(1): 150-157.
[12] Nouaouria N, Boukadoum M. Improved global-best particle swarm optimization algorithm with mixed-attribute data classification capability. Applied Soft Computing, 2014, 21(8): 554-567.
[13] Bhutani K, Kumar M, Aggarwal S. Multi-attribute data classification using Neutrosophic probability // India Conference. IEEE, 2016, 1-5.
[14] Nezarat A, Dastghaibifard G H. Efficient Nash Equilibrium Resource Allocation Based on Game Theory Mechanism in Cloud Computing by Using Auction.. Plos One, 2015, 10(1): 138-144.
[15] Mazalov V, Lukyanenko A, Luukkainen S. Equilibrium in cloud computing market. Performance Evaluation, 2015, 92(3): 40-50.
[16] Gąsior J, Seredyński F. Decentralized Job Scheduling in the Cloud Based on a Spatially Generalized Prisoner's Dilemma Game. International Journal of Applied Mathematics and Computer Science, 2015, 25(4): 737-751.
[17] Botta A, De Donato W, Persico V, et al. Integration of Cloud computing and Internet of Things. Future Generation Computer Systems, 2016, 56(3): 684-700.
[18] Díaz M, Martín C, Rubio B. State-of-the-art, challenges, and open issues in the integration of Internet of things and cloud computing. Journal of Network and Computer Applications, 2016, 67(3): 99-117.
[19] Chen D F. Research on application of cloud computing in mobile Internet. Electronic Design Engineering, 2015, 13(1): 50-58.
[20] Zhao T, Zhou S, Guo X, et al. A Cooperative Scheduling Scheme of Local Cloud and Internet Cloud for Delay-Aware Mobile Cloud Computing // IEEE GLOBECOM Workshops. IEEE, 2015, 1-6.
[21] Liu F, Peng Z Y, Gao X. Simulation of port resource equilibrium scheduling model in cloud computing environment. Computer Simulation, 2016, 33(8): 297-300.
[22] Verzi, D, Alcala, J, Fletes, E, Gonzalez, C, Hernandez, D, Palomares, E, and Vega, P. A Computational Study of Pancreatic Beta Cell Dynamics in the Progression to Diabetes. Journal of Interdisciplinary Mathematics, 2018, 21(3): 695-716.
[23] Susanth, C, and Kalayathankal, S J. Operations on Independence Numbers of Certain Graph Classes. Journal of Discrete Mathematical Sciences and Cryptography, 2018, 21(1): 75-82.
[24] Mallik, A, and Arefin, M A. Clean Water: Design of an Efficient and Feasible Water Treatment Plant for Rural South-Bengal. Journal of Mechanical Engineering Research and Developments, 2018, 41(1): 156-167.