The Application of Cloud Computing in Large-Scale Statistic

SUN Xiuli
Shandong Women's University
Jinan, Shandong Province, China
197935@gmail.com

LI Ying
Department of Information Technology
Shandong Women's University
Jinan, Shandong Province, China
cherry_jn@126.com

HU Baofang
Department of Information Technology
Shandong Women's University
Jinan, Shandong Province, China
hbf0509@126.com

SUN Hongfeng
Department of Information Technology
Shandong Women's University
Jinan, Shandong Province, China
nameshf@163.com

***Abstract*—The main challenge in current statistical work is the huge pressure of the statistical analysis along with the huge amount and diversity of the statistical data. This paper established a framework model of large-scale data processing by bringing in cloud computing. By studying the resource allocation algorithm of cloud computing, we proposed an accelerating genetic algorithm of double-target fitness function which considered the safety of statistical data and the responding time of work, as well as analyzed the convergence speeds of the algorithm in various weights in order to test each target's effect on iterations.**

Keywords- statistical analysis;cloud computing;double-target genetic algorithm

I. INTRODUCTION

Online submission of statistical data can enhance its timeliness, provide government departments with authentic analysis basis for macro economic trends, as well as promote sound and rapid economic and social development in our country. In 1990, the Office of National Statistics established a cyber direct reporting platform of statistical data from 5000 industrial enterprises and 3000real estate enterprises. Later, according to the requirements of the Office of National Statistics, statistical bureaus at provincial and municipal level have regarded the cyber direct report as the key program to collect the statistics of grassroots construction. On the basis of the target model and overall thought of "a set of meters + computerization", most provincial capital cities have the comprehensive statistic management platform, among which provinces like Guangdong, Jiangsu, Shandong, Zhejiang have basically established a municipal-level direct reporting platform of statistics. At present, based on the previous achievements, the Office of National Statistics ranks the big four projects as the key to the following work， namely the establishment of inventory of basic departments, system of a set of meters of company, integrated software system construction and the cyber direct reporting system, to enhance the capacity of statistics, the equality of statistical data and credibility of statistics, and gradually realize the unity and normalization of the statistical work to ensure the reality and accuracy of statistical data.The main challenge in current statistical work of data processing is the huge pressure of the statistical analysis along with the huge amount and diversity of the statistical data, however, the effective utilization of resources for data processing in statistic units at all levels still has room for improvement. If we can make the best use of the existing computing resources, lower cost and efficiency accomplishment of statistical work are the important issues need seriously studying and solving.

As the representative of the new generation of calculation model, cloud computing technology possess several key technologies which can be applied to solve the above-mentioned problems through comprehensive evolution of a couple of techniques such as grid computing, utility computing, service computing and so on. From the perspective of construction technology of cloud, it is a model which uses a lot of concurrent working processors to deal with high performance computational problems. Compared with other high performance computing technologies, it has three essential features, that's to say, based on large-scale cheap server clusters, the cloud platform maximize the efficiency of hardware resources through the cooperation between basic facility and application program at upper layers, and at the time tolerate errors of several nodes by means of software[1] [2].

By applying cloud computing to the statistical data processing, this paper studied the following four parts: Analyzed the statistics task and the distribution of computing resources.

- Established the model of application framework of cloud computing technology in large-scale statistical data processing.
- Optimized the computing resources in the established framework model by adopting genetic algorithm.
- Analyzed the advantages by applying cloud computing to statistical data processing.

II. DISTRIBUTION OF COMPUTING RESOURCES IN THE STATISTICAL WORK

The statistics task has the characteristics of varied data, heavy work, higher accuracy and timeliness. At present, the analytical calculation of cyber direct reporting system

generally rely on the computing platform of the nodes at national and provincial level. For large-scale cyber direct reporting system, it has limited computing power, poor expandability and high upgrading cost. The current cyber direct report is limited to above-scale enterprises in national economic industries, such as industry, construction industry and real estate industry. With further expansion of the direct report, social statistics such as price investigation and demographic census will gradually adopt cyber direct report which will engender amazing data size. One of the important challenges of cyber direct reporting system in the future is the serious shortage of data storage and analytical ability, and the large-scale processing capacity of statistical data will inevitably become a bottle-neck to the development of cyber direct reporting technology.

Through analysis of the computing resources and characteristics of statistical data of the statistical department, we discovered that the calculating pressure of statistical departments at all levels presents a hierarchical structure which means that calculating pressure will be more serious in higher levels. Computing resources of the Office of National Statistics is usually running at full capacity, while computing resources of statistical departments at lower levels cannot make the most of them because their statistical work are limited to local areas and thus have few and light tasks.

III. APPLICATION FRAMEWORK MODEL OF THE STATISTICAL CLOUD

For the sake of the data features and the distribution of computing resources in the statistical work, we analyzed the effect of virtualization technology in the cloud computing model on the integration of computing resources in the statistical work, and proposed to effectively organize and use the computing resources of statistical departments at lower levels by use of the computing technology in order to establish a unified sharing platform by gathering the computing resources of statistical departments at all levels. For the whole statistical cloud system, we sat up a unified resource pool constituted by all the available resources, and the pool consisted of statistical data, processors, memorizers and network resources. What's more, we allotted tasks and resources for virtual nodes through virtualization technology and the real-time transport technology of virtual machine, thus to improve the utilization rate of computing resources of the whole statistical system.

The establishment of statistical cloud involves the submission of statistical cloud, infrastructures of network storage, servers which provide computing power, systematical management platform, applications for carrying services and so on. Together with principles of the cloud computing, the model of statistical cloud platform in this paper was composed of the resource layer, the virtualization layer, the management layer, the service layer and the display layer (see in figure2).

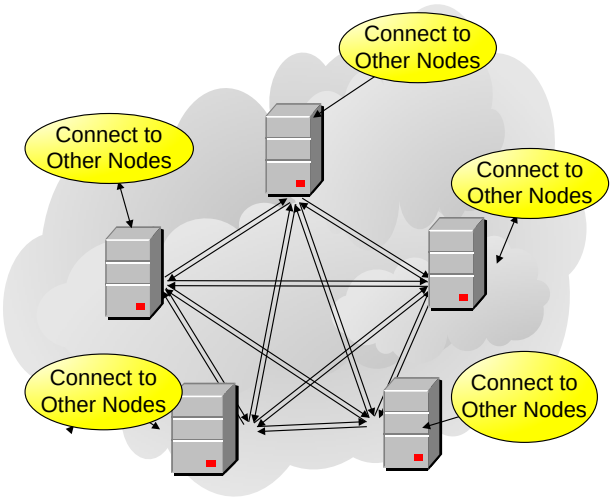

Figure1. Structure of the Statistical Cloud Platform

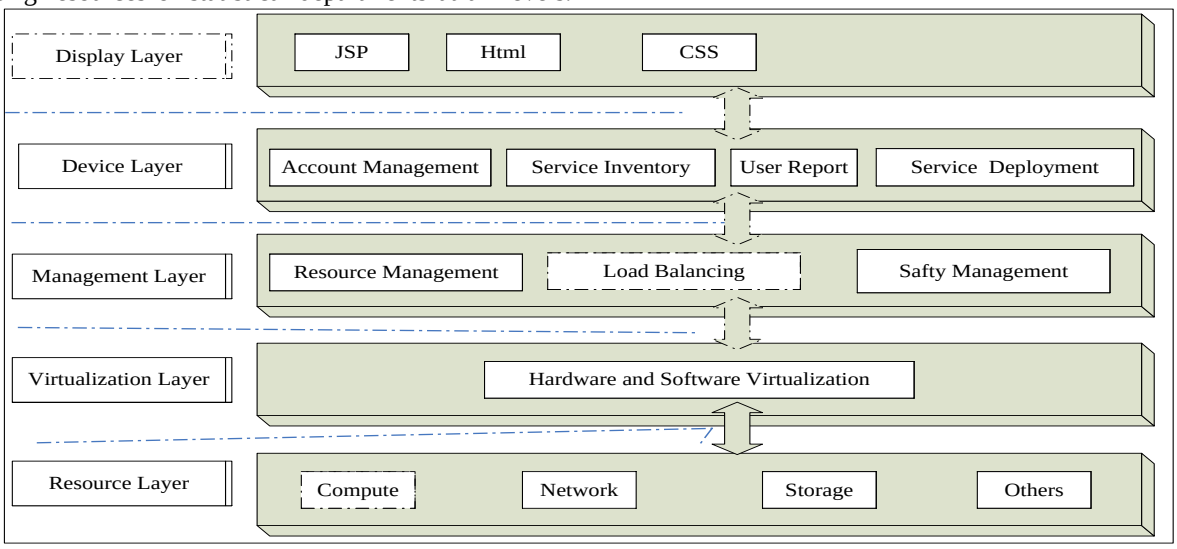

Figure2. The Framework Structure of The Statistical Cloud

IV. DISTRIBUTION OF COMPUTING RESOURCES BASED ON DOUBLE-TARGET GENETIC ALGORITHM

At the present, techniques such as scheduling algorithm, distribution, load balancing and storage redundancy of computing resources in the cloud computing have been frequently studied at home and abroad. Li Jianfeng and other people have proposed a genetic task scheduling algorithm appropriate for the cloud computing which took both the total and average execution time of the task into consideration[3]; Hua Xiayu al de have put forward the load algorithm of computing resources based on ant optimization, that's to say, when distributing computing resources, we should firstly predict the computing equality of potential nodes, then according to features of the environment of the cloud computing to present a resource distribution algorithm based on ant optimization by analyzing the effects of factors such as circuit equality, occupied bandwidth and responding time on loads [4] .

The report and delivery of statistical data should be done at regular time. On the one hand, we should pay attention to the timeliness of task responding; on the other hand, since statistical data is related to secret information of enterprises and individuals, we should attach special importance to the safety of the data. Based on these two aspects, this paper proposed a double-target accelerating genetic algorithm to distribute statistics tasks among different computing nodes. Besides the responding time of tasks, the fitness function has also considered the safety of tasks.

Assuming that the model meets the following requirements:

1) The computing power of each node is known to us, and it could obtain the type and the speed of processors.

2)Whenever submitting a statistical task, the relevant statistical department has to set a safety demand (SD). SD represents safety demand of tasks which includes execution environment, visit control and so on.

3)Setting the safety level (SL) of each node according to their defensive capacity. SL is determined by invading detection, firewall, anti virus/worm, responsiveness for attack and so on.

4)The prerequisite of the calculation of node computing is that the SL of nodes should greater than the SD of the task; otherwise the task will reselect executive nodes. The failure probability of the task is presented as equation1.

$$PF=\begin{cases} 0 & \text{if } SD<=SL \\ 1-e^{-(SD-SL)} & \text{if } SD>SL \end{cases} \quad (1)$$

The flow chart of the algorithm can be seen in figure3.

The genetic algorithm is a method to select the evolution objects of next generations through the fitness function in order to find the optimal solution to the problem. Therefore, the choice of the fitness is quit important because it is related to the convergence rate of the algorithm and the quality of the solution [3] [6]. The fitness function in this paper is presented by equation2.

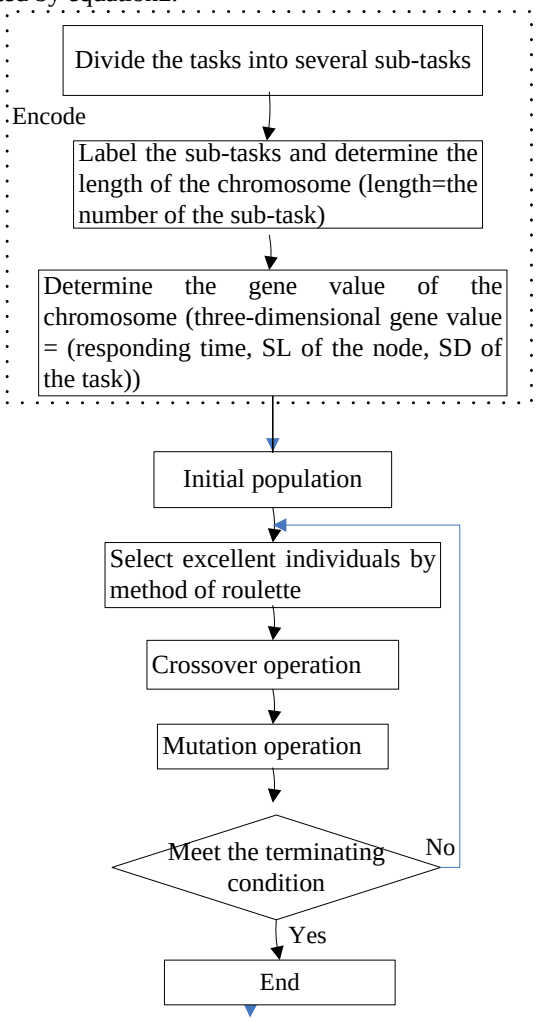

Figure3. The Flow Chart of the Algorithm

$$f(m)=\frac{1}{W1*ExecutionTime+W2*PF} \quad (2)$$

ExecutionTime means the product of the calculated amount of the computing power and the sub-task of the node, PF represents the failure probability of the node when executing a certain task, W1 and W2 are the weight proportions of the safety of execution time and the sub-task in the node fitness function, which means that the individual with less execution time and lower failure rate will has greater fitness value, thus has greater probability to be chosen.

This paper can adjust the proportion of each objective in the node fitness through setting different weight value, and five kinds of weight group can be seen in table1.

The construction of the simulation environment in this paper adopted the Gridsim [5]. In the same environmental conditions, it will compare the convergence speed by use of different weights. And the terminating conditions of the algorithm are all setted:

- The mixmum iteration is 200;
- If the execution time of the successive 50 generations as generally the same, we can say that the algorithm is basically converged, thus end the algorithm [3][6].

The results of the alorithm can be seen in figure4.

TABLE I. DIFFERENT WEIGHT GROUP

Weight group	W1	W2
WC1	0.8	0.2
WC2	0.7	0.3
WC3	0.5	0.5
WC4	0.3	0.7
WC5	0.2	0.8

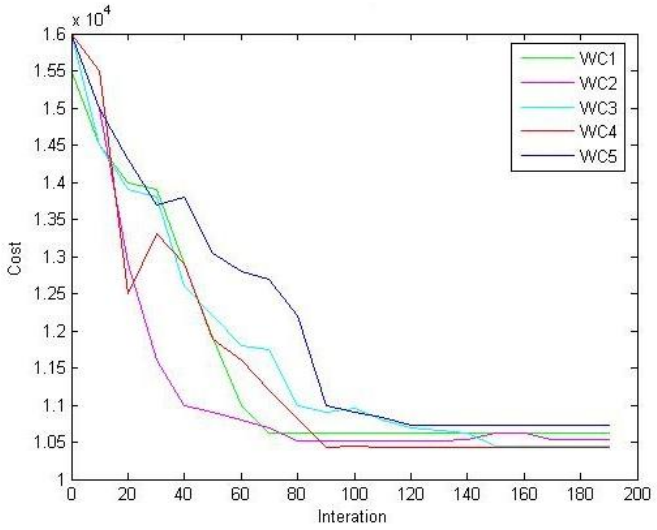

Figure4. The Simulation Results

From the above results, we can know that the responding time has big proportion in the evolution time, and the greater weight of the responding time, the faster of the evolution speed. If W1=0.7, W2=0.3, the iteration is 67. Therefore, for compute tasks with lower SL, we can enlarge the weight of the time to speed up the convergence of the algorithm, decrease the iterations and reduce the cost of the algorithm; for tasks with higher SL, we should choose the weight installation with higher node failure rate.

V. THE ADVANTAGES OF APPLYING THE CLOUD COMPUTING TO STATISTICAL DATA PROCESSING

The statistical cloud platform based on the framework model proposed in this paper has the following advantages:

- It is beneficial to the unified management of the statistical data. In the framework model of the statistical cloud established in this paper, for statistical agencies at all levels, the unified organization of all the statistical data in the cloud has just one requirement that meets the access ritght of the data, which has increasd the usability of the data.
- It can make the best of the various computing respources which has greatly relieved the computing pressure brought by the top statistical agencies after the cyber direct report. If the calculating task of the node was not crowed in a certain statistical agent, and the computing resources such as the server are left unused, it can use the idle resources by the method of virtual machine which will accordingly strenghten the computing power of the top statistical agencies.
- It has lightened the load of the statistical agenties at premary level. The structure of the framework model suggested in this paper can make use of the available resources and facilities for cyber direct report of statistical data, therefore, statistical data at primary level can realize cyber direct submission of statistical data if they have browers and internet, and the existing resources in statistical data at provincial, prefectural and municipal levels are involved in the unified management of the Office of National Statistics, thus save up a large number of excellent statisticians to statistical agencies at primary level.

VI. CONCLUSIONS

This paper has analyzed the distribution characteristics of statistical data and computing resources in statistical work. By introducing the cloud computing into statistical work, it established a framework model for large-scale statistical data processing based on cloud computing, based on which it studied the distribution algorithm of computing resources in cloud computing and proposed an accelerating genetic algorithm of double-target fitness function with different weights according to the traits that the statistical work not only require the timeliness but also the safety of the data. Thai paper has also analyzed the convergence speed of the algorithm of different weight values which provided the setting of the weight value with theoretical basis.

ACKNOWLEDGMENT

This work was supported partially by a project of Shandong province higher educational science and technology program (J11LG56), the national statistics scientific research program of China (2011LZ028).

REFERENCES

[1] Liu Zhen, Liu Feng al de. The Application of Cloud Computing Model in Large Scale Railway Data Processing [J]. Journal of Beijing Jiaotong University. 2010. 10:14-19

[2] Mi Le, Jiang Jinle. Cloud Computing [M]. Beijing: Machinery Industry Press, 2009.

[3] Li Jianfeng, Peng Jian. Task Scheduling Algorithm Based on Improved Genetic Algorithm in Cloud Computing Environment [J]. Computer Applization. 2011.1:184-186

[4] Hua Xiayu, Zheng Jun, Hu Wenxin. Distribution Algorithm of Ant Optimizing Computing Resource Based on Cloud Computing [J]. Journal of Huadong Normal University (Column of Nature and Science). 2010.1

[5] HU Baofang,SUN xiuli,LI Ying, SUN Hongfeng.An Improved Adaptive Genetic Algorithm in Cloud Computing.[J].PDCAT2012

[6] Wang Xiaoping, Cao Liming. Genetic Algorithm [M]. Xi'an: Press of Xi'an Jiaotong University, 2002.