Efficient big data processing strategy based on Hadoop for electronic commerce logistics

Jiaojin Ci

Institute of Economy and Management, Nanyang Normal University，Nanyang, 473061, China

Corresponding author's e-mail: cijiaojin@163.com

Received 01 October 2013, www.cmnt.lv

Abstract

With the rapid development of cloud computing, more and more electronic commerce applications are confronted with the problems of processing big data, such as big data from the social media posted by the customers of electronic commerce logistics. In order to improve the big data processing efficiency in electronic commerce logistics, an efficient big data processing strategy based on Hadoop is designed, which is named ECLHadoop. In ECLHadoop, those closely related data blocks are placed at the same nodes, which can help to reduce the MapReduce I / O cost, especially the I / O cost at the shuffling stage. The simulation experiment results show that, based on Hadoop, the ECLHadoop can improve the big data computing efficiency for data-intensive analysis in the electronic commerce logistics service.

Keywords: big data, data placement, big data analysis, big data computing strategy, electronic commerce logistics

1 Introduction

With the development of Internet technology and e-commerce, especially in the development of cloud computing and physical networking technology, humans, sensors, PC, servers, mobile phones and other equipment to produce more and more data in daily production and life, how to deal with these large data is facing a huge challenge. Currently, there have been numerous studies abroad designed some large data processing framework, mainly large data processing framework including Google's Google MapReduce [1] and the Apache Hadoop Hadoop MapReduce [2], these two frameworks have their own distributed File System, Respectively GFS [3] and HDFS [4].

Hadoop data placement major consideration in order to balance the distribution of data, but it does not take into account data placement policy relationship between the various data sets. All data stored in HDFS according to workload requirements Hadoop cluster to place, so when executing MapReduce computing, large amounts of data will be migrated, and thus increase the number of I / O cost, especially in Shuffling phase I / O cost.

Since then, the Hadoop big data placement strategies on the basis of many academics, research institutions and universities to design a number of large data Processing frameworks, such as Hadoop ++ [5], CoHadoop [6] and HadoopDB [7]. However, these large data processing frameworks need large data itself larger changes. Hadoop ++ needs Static data indexing and join Trojan Trojan connection, when the new need for these large data again by some static data odifications. CoHadoop is based on IBM's proposed Hadoop new optimization method, which divided according to the application requirements with data blocks. Large data before being submitted to the HDFS, Co-Hadoop application requirements need to attribute them into big data dividing line; due to the large data before being stored in HDFS must enter Line changes, thus increasing the number of treatment costs. Hadoop ++ and CoHadoop carried out to some extent on the basis of Hadoop The degree of modification, and HadoopDB conducted on the basis of Hadoop Greater changes.

While the existing framework of the large data processing to some extent, raise High efficiency of large data processing, but these processing framework more or Less need for big data can be modified. This changes some special Useful for a given application, but will result in a variety of applications of the process frame loss Versatility. Therefore, in order to improve the processing efficiency of large data, RESEARCH A new study, efficient large data processing strategy has important meaning Justice, this strategy should be based on Hadoop for data-intensive Applications, and does not require any changes Hadoop itself. Given these requirements, we design a kind of Hadoop-based Foundation for e-commerce logistics services such data-intensive applications Effective large data processing strategy ECLHadoop (Efficient Electronic Commerce Logistics Big Data Processing Strategy in Hadoop for Data-intensive Analysis).

2. Research status overview

The large data is divided into n data blocks, each corresponding to the Map-tasking data block. This phase belongs Shuffle treatment, will increase the number of I / O cost, especially large number of related data blocks are placed in the data node.

Big Data in Hadoop ++ processing strategy as shown in Figure 1 In large data sources from a data block in the submission to the Map Reduce Before calculating, must be pre-processed. Hadoop ++ will Adding block indexing and Trojan Trojan connection information. When Large data is divided into data blocks, each data block is added to the previous A Trojan index. With the addition of an index and Trojan Trojan Connection, so Hadoop ++ stage or in the Map Reduce Order Segment, all calculations especially Join

query evaluation will be well Improvement. However, adding the index information and Trojan Trojan Union Access the information you need to spend some time, and when the data set changes, Need to re-add the index information and Trojan Trojan joins letter Interest, while the associated data set that will not be placed together, So Hadoop ++ in Shuffle is still a big stage Bottleneck.

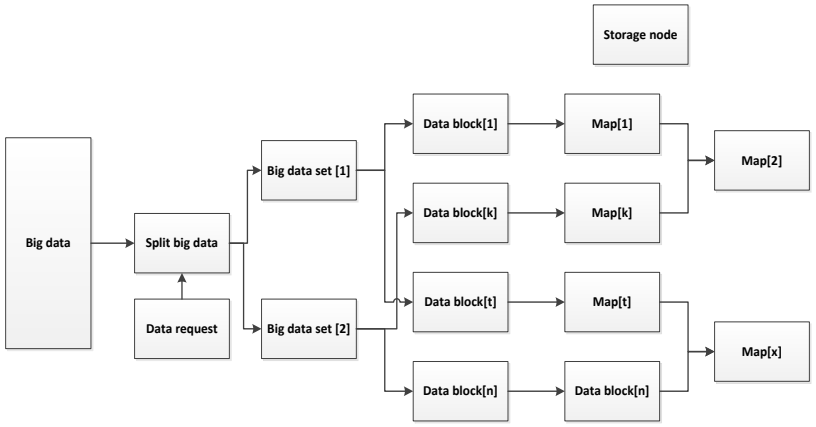

FIGURE 1 CoHadoop big data processing strategy

Firstly, according to Co-Hadoop Big Data application requests will be divided, then the associated data block is assigned to the same data storage node. Thus, the query is calculated especially related queries will be addressed in the same compute node. Map Reduce computation only needs Map Reduce computing without the need to calculate, so you can cancel Shuffle stage, costs are greatly reduced.

In order to improve processing efficiency, and many other optimization methods or strategies have also emerged. For example, the Twister and HaLoop [10] have proposed an iterative calculation for large data problem, UC Berkeley also made Spark [11] to handle large data. In addition, there are some other big data computing strategies such as Hadoop-ML [12] and so on.

The standard GA generates initial population randomly. This method tends to produce a narrow search space, which is disadvantageous to the acquisition of the global optimal solution. To improve the convergence characteristics of GA, the initial population is generated using uniform partition based generation method. This method evenly divides the range of optimized parameters into regions with their number equal to population scale Gs. In every small region, an individual will be generated randomly. This method can quicken the convergence speed and increase the possibility of converging to global optimal solution. Based on the above analysis, the crossover probability P_c^t at *t*-th generation will be defined as:

$$P_c^t = \begin{cases} P_{\max} - \dfrac{(P_{\max} - P_{\text{temp}})(f_b^t - f_{\text{avg}}^t)}{f_{\max}^t - f_{\text{avg}}^t} & f_b^t \geq f_{\text{avg}}^t \\ P_{\max} & f_b^t < f_{\text{avg}}^t \end{cases} \quad (1)$$

where:

$$P_{\text{temp}} = \begin{cases} P_{\min} & P_{\max} e^{-\frac{t}{T_l}} \leq P_{\min} \\ P_{\max} e^{-\frac{t}{T_l}} & P_{\max} e^{-\frac{t}{T_l}} > P_{\min} \end{cases}, \quad (2)$$

where $P_{\max}$ and $P_{\min}$ are the maximum crossover probability and the minimum one, T_l be the maximum iteration times, f_b^t be the bigger fitness of two individuals chosen for crossover operation at t-th generation, $f_{\max}^t$ and f_{avg}^t denote the maximum fitness and average fitness of the population at t-th generation. Similarly, the mutation probability P_m^t at t-th generation will be defined as:

$$P_m^t = \begin{cases} P'_{\max} - \dfrac{(P'_{\max} - P'_{\text{temp}})(f_k^t - f_{\text{avg}}^t)}{f_{\max}^t - f_{\text{avg}}^t} & f_k^t \geq f_{\text{avg}}^t \\ P'_{\max} & f_k^t < f_{\text{avg}}^t \end{cases} \quad (3)$$

where:

$$P'_{\text{temp}} = \begin{cases} P'_{\min} & P'_{\max}(1 - e^{-\frac{t}{T_l}}) \leq P'_{\min} \\ P'_{\max}(1 - e^{-\frac{t}{T_l}}) & P'_{\max}(1 - e^{-\frac{t}{T_l}}) > P'_{\min} \end{cases} \quad (4)$$

where $P'_{\max}$ and $P'_{\min}$ are the maximum mutation probability and the minimum one, f_k^t be fitness of the individual chosen for mutation operation at t-th generation.

3 E-commerce logistics of large data processing strategy –ECLHadoop

Through the above analysis we can see that the existing large data processing strategies are some problems, especially for specific applications, are present bottleneck restricting its computer efficiency. Based on this, we propose a large e-commerce logistics data processing strategy-ECL Hadoop, which is shown in Figure 2.

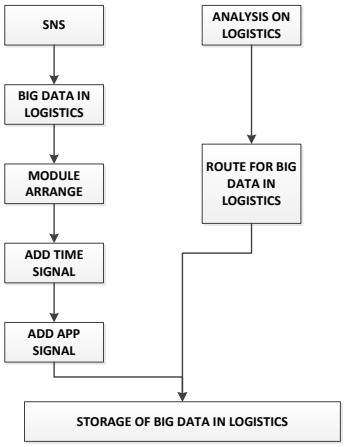

FIGURE 2 Electronic commerce logistics big data processing strategy-ECLHadoop

Data from a large e-commerce logistics commerce logistics service Customer Service released on various social networking sites relevant information. These will be based on Hadoop big data mechanism is divided into 64MB Data blocks will add a time stamp to all data blocks. In addition, we will also add applications associated tag for each block of data; these should be Markers associated with the demand from the early, they and e-commerce associated logistics services. That is, if the analysis needs to tell me they need data sets A and B to link data sets, we can recognize some of the data block with the data sets A and B for the data set of some number.

According to the association of the block, thus, adding a clearance for these data blocks Union flags. After completion of the calculation for all the semantic data blocks, can be Get big e-commerce logistics data placement routing table. Through which complete e-commerce logistics of large data placement, those closely related Data blocks will be stored in the same node. Because in Map Reduce Shuffle avoid processing stage, it can improve the count of efficiency, especially related inquiries calculations.

To better illustrate the e-commerce logistics of large data processing strategy, e-commerce logistics and big data correlation definition and e-commerce logistics of large data correlation calculation is particularly important to do a brief analysis of the following:

(1) E-commerce logistics big data correlation.

In the calculation of e-commerce logistics in big data, Those collections with computing dependent data sets is called the related data sets. As in E-commerce logistics applications, assume there is a very frequent meter Operators relationship - calculated by "Air China" in logistics and transport of fast Delivery traffic. As can be seen from the demand, "Order Form" and "Logistics Business Transportation Company table "is the relationship between two computing Join Table. Therefore, it can be seen that the "Order Form" and "public transportation logistics business Division table "two tables with large data dependencies are two tables, which E-commerce logistics is big data correlation.

(2) E-commerce logistics big data correlation calculation

Big data correlation is calculated primarily through e-commerce logistics for business needs analysis and calculation to achieve. Assuming ecommerce matter The Company's big data flow analysis needs, and they need the distribution table as follows:

1. Business needs 1: {Table (A), Table (C)}
2. Business needs 2: {Table (A), Table (C)}
3. Business needs 3: {Table (A), Table (C)}

Which can be seen, the data set A data set C and data Set D-related, and the data sets and data sets E B related.

Map Reduce mainly includes three aspects of computing, namely, Map phase calculation, Reduce phase calculation and Shuffle stage Calculations. If the data sets associated with A and C in accordance with Hadoop Their random distribution of the distribution mechanism, the data set A And C will be dispersed in the data center a number of data nodes. Whole A Map Reduce computation process includes the following important aspects: First, Map stages needed by hundreds of thousands of data node count Calculation to obtain intermediate results, and these results need to be migrated to the middle Reduce data nodes specified final calculations. Data (intermediate stages thousands meter Map data nodes in the process calculation results) require data migration (Shuffle process), so big Large increases the I / O cost of Map Reduce Shuffle stage Makes the I / O cost of the entire phase of extremely large Map Reduce, Overall computational efficiency is very low.

Different from the above process, if the data in an associated relationship A set of data collection and placed in the same node C, the data sets A and Data set C may be calculated in the same data node completes, without To Shuffle stage of Map Reduce, thus avoiding the Map Reduce The Shuffle bottleneck. So it will greatly enhance the Map Re duce Computational efficiency.

At the same node, there is no relationship between the numbers of calculations associated data storage Data block as Hadoop as completely dispersed into the data center of many thousands of computers. However, we have adopted and Hadoop A copy of the same strategy, the same set of three copies, so by three copies can effectively protect data reliability. In protection can on reliability, ECLHadoop strategy and Hadoop data storage Strategy is the same, there is no conflict.

In the previous analysis, we mainly consider the relevant data sets Can be stored in the case of the same node, but with the amount of data To the larger, there may be some associated data sets is difficult to store In the case of the same data node. To solve the above problem, ECLHadoop treatment strategy is as follows: The data set is stored in the same A network bandwidth performance of the best chassis, the nearest two computing On the machine, if the two computers is not enough, it is stored in the nearest three On the computer, and so on.

In this section, the proposed e-commerce logistics of large data processing strategy ECLHadoop on the basis for such a treatment strategy Detailed case studies. First, assume that e-commerce logistics service Service-related social networking site has four data sets, respectively: Dataset A, dataset B, C data sets and data sets D.

4 Simulation

In order to prove the correctness and rationality of the above

methods were two related simulation. Experiment 1 compared with the time stamp added Join ECLHadoop mind and no computational experiments join query Hadoop computational efficiency under the circumstances; and Experiment 2 will compare Tim Plus the associated application and have labeled ECLHadoop Join joins Charles Inquiry case computational experiments computational efficiency of Hadoop.

Simulation experiments using five computers, one of which is the metadata node, and the other A node outside of the shadow, and the remaining three for e-logistics data Storage nodes. Each computer will be installed the Linux operating system and the Hadoop Distributed File System HDFS.

Simulation results from a use of the major social networking sites the e-commerce logistics service related data. Time simulation data is ranging from January 1, 2013 to May 8, 2013. In Simulation Experiment 1, simulation experiments were carried out in five of the added Time stamp and non-join query ECLHadoop calculated Hadoop comparing the calculation time of both comparative results are shown in Figure 3.

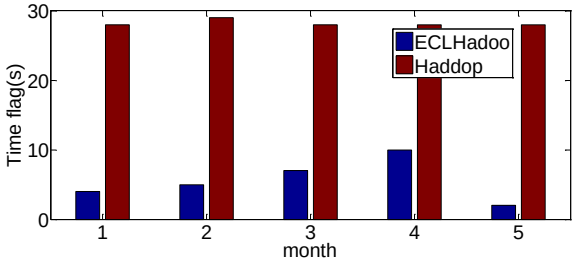

FIGURE 3 Computing time comparison between the ECLHadoop with time flag and Hadoop without Join query computing

From the experimental results, in Hadoop, the pair Any requests, all the data blocks have to be calculated; however, in ECLHadoop, since the effect of the time stamp calculation range,, greatly reduced. For example, when only need to get the time range from 2013 For some years between April 1 to April 30, 2013 data The results of its inquiry, you can discard this time outside the scope of It is the data to reduce the amount of computing time (i.e., count MapReduce Count the time). However, all data in Hadoop because even Form a block, must have been calculated, and therefore the computation time spent significantly More.

Simulation 2 is still used from the social networking website E-commerce logistics services related simulation data. This set of simulation data there are two data sets, ie, data sets and data sets A B. And Simulation A similar experiment, this experiment still contains five sub-group simulation. The simulation has five groups, including: (1) datasets A connection Dataset B. A size of the data set for the 1TB, the dataset B, Size is 1TB. (2) A data set joins dataset B. Number According to the size of the set A is 1TB, while the size of the data set B is 2TB, (3) A data set joins dataset B. A large data sets in Small as 2TB, while the size of the data set B is 1TB. (4) Data sets A coupling dataset B. A size of the data set for 2TB, data the size of the set B is 2TB. (5) A data set joins dataset B. A size of the data set for 2TB, while the size of the dataset B Is 3TB.

5 Conclusions

This paper is designed based on Hadoop for E-commerce logistics data-intensive analysis of large data processing strategy-ECLHadoop. In order to improve e-commerce logistics for large data Computational efficiency, ECLHadoop strategy designed e-commerce Data Service Logistics large data placement strategy will have to calculate the relevance The data is placed together in the same data node. Accordingly, the use of ECLHadoop policy can significantly reduce I / O cost. Without Studies, we will be in big data processing based on open source Implement and improve the basic strategy ECLHadoop on Hadoop Strategy. Moreover, in a further study, some more analysis based on the large data processing strategy ECLHadoop e matter Streaming service analysis applications, the count-based ECLHadoop Count the outcome desired by the user, thereby improving customer e-commerce the new logistics service satisfaction.

References

[1] Pal S K, King R A 1983 On edge detection of X-ray images using fuzzy sets *IEEE Trans Pattern Analysis and Machine Intelligence* **5**(1) 69-77
[2] Nakagawa Y, Rosenfeld Y 1978 A note on the use of local min and max operations in digital picture processing *IEEE Trans Systems, Man & Cybernetics* **8**(8) 632-5
[3] Dean J, Ghemawat S 2004 MapReduce: Simplified data processing on large clusters *Proc of the the 6th Symposium on Operating System Design and Implementation* 1-13
[4] Dittrich J, Quian'e-Ruiz J A, Jindal A, et al 2010 Hadoop ++:Making ayellow elephant run like a cheetah (without it even noticing) *Proceeding of the VLDB Endowment* **3**(1-2) 518-29
[5] Ekanayake J, Li Hui, Zhang Bing-jing, et al 2010 Twister: A runtime for iterative MapReduce *Proc of the 1st International Workshop on MapReduce and Its Applications* 124-41
[6] Bu Y Y, Howe B, Balazinska M, et al 2010 HaLoop: Efficient iterative data processing on large clusters *Proceeding of VLDB Endowment* **3**(1-2) 285-96
[7] Ghoting A, Pednault E. 2009 Hadoop-ML: An infrastructure for the rapid implementation of parallel reusable analytics *Proc of the Large-Scale Machine Learning: Parallelism and Massive Datasets Workshop* 38-48
[8] Brown J 1994 Some motivational issues in computer-based instruction *Educational Technology* **26**(4) 27-9
[9] Kapur J N, Sahoo P K, Wong A K C 1985 A new method for gray level picture thresholding using the entropy of the histogram *Computer Vision, Graphics and Image Processing* 29 273-85
[10] Fu X N, Yin S M, Liu S Q 2003 A improved adaptive fuzzy entropy entropy thresholding method on image segmentation *Acta Photonica Sinica* 32 605-7
[11] Gattoufi S, Oral M, Reisman A 2004 Data envelopment analysis literature: a bibliography update (1951-2001) *Socio-Economic Planning Sciences* 38 159-229
[12] Ridler T W, Calvard S 1978 Picture thresholding using an iterative selection method *IEEE Trans Systems, Man and Cybernetics* 9 630-2

Author

Jiaojin Ci, 1982.11.20, Anqing, Anhui, China.

Current position, grades: lecturerat Institute of Economy and Management in Nanyang Normal University.
University studies: economics.
Scientific interest: logistic management and Electronic commerce.