http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (24): 8398-8404, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.8398.8404

A Parallel Algorithm of Public Bus Accounting Based on Mapreduce

Zhike Han, Shuoping Wang, Jiawei Lu and Hanjian Zhang
School of computer and computing science, Zhejiang University City College
Hangzhou, China

Abstract: Public bus accounting system employs basic accounting methods and financial accounting principles of double-entry bookkeeping. It uses real-time data provided by passenger daily system, real-time billing system, supplies system and human resources system. And it takes the single commercial vehicles and line as the accounting unit to collect revenue, costs. Ultimately, it implements the reflection of bus, lines, fleet and company's operation financial condition accurately. Nevertheless, the number of the data of the accounting process is huge. The problem that accounting process is relatively long and real-time is relatively bad exists as well. The arise of MapReduce framework solves the problem, it provides simple programming interface, hides bottom details, sets programmers free from traditional parallel programming mode. At the same time, its lack of simplicity exists as well, for example, internal expression capability is weak, as for some complex algorithms, it must be separated by programmers into units that can be individually run in MapReduce framework. This study analyses emphatically the problem of long accounting process and bad real-time that using MapReduce programming framework to solve the problem.

Key words: Public bus accounting, parallel algorithm, mapreduce, cost sharing

INTRODUCTION

With the development of society, the relationship between each individual is more and more complex, therefore it brings challenges to traditional method of social network analysis. Because it is impossible that storing the network to internal storage, even external storage, although paid dearly for it. A spontaneous idea is to introduce distributed processing.

In the distributed system, it has many general physical and logical resources and can dynamically allocate tasks, allocated physical and logical resources as well as implement information interchange by computer network. The open-source distributed framework Hadoop (Wang *et al*., 2009) which is developed according to Google cloud computing has been able to finish the task that users submit by so-called MapReduce Job, each Job individually run in a computer node, theoretically as long as increase the amount of machine to increase the amount of Job, it can finish more client task (Dean and Ghemawat, 2010).

MAPREDUCE FRAMEWORK

MapReduce framework description: This framework mainly implements automatic parallelization and big-scale distributed computing by simple and powerful interface, combined with implementation of this interface to implement high-performance computing in a large number of common PCs.

When computing, it uses input Key/Value matching set to produce output Key/Value matching set. Users of MapReduce framework use two functions to express this computation: Map and Reduce. Customized mapper receives input matching set then produces an intermediate Key/Value matching set. MapReduce framework aggregates all the intermediate Value that includes the same intermediate Key to transfer them to reducer, to form a small Value set. Generally, every time it calls Reduce it only produces 0 or 1 output Value (Cohen, 2009).

It can be expressed by two simple formulas:

- Map (k1,v1) ->list (k2,v2)
- Reduce (k2,list(v2)) ->list (v2)

Calling Map automatically splits input data into M pieces that are distributed to several machines, each piece can be parallel processed in different machines (Dean and Ghemawat, 2004). MapReduce framework splits intermediate Key by calling split function that is customized by users to form R pieces, they are distributed to several machines(Venner, 2005).

Corresponding Author: Shuoping Wang, School of Computer and Computing Science, Zhejiang University City College, Hangzhou, China Tel: 086-571-88018551 Fax: 086-571-88015567

PUBLIC BUS ACCOUNTING SYSTEM AND ITS SERIAL ALGORITHM

The current state of the research of Public bus accounting system: Before Spread of information technology in the enterprise, the concept of Bus accounting has been produced. At that time the Bus accounting is mainly based on the most traditional accounting method (Song, 2010), which are carrying on in the study. During that the project involved is extremely various, the amount of data is giant as well. Hangzhou Public Transport Group Co., Ltd., for example, includes 7 subsidiaries, and every subsidiary divided into teams, and route lines. The scale of the company now is very large, involving the wider lines, which includes more than 10,000 buses involved in daily operations among the whole company. In order to monitor the Group's income and expenditure, the company is bound to count the income and expenditure.

The accounting subject involved a large scale, includes main business revenue, main business cost and etc., There are ten or so more top items. Furthermore, top items large items is divided into many subsidiary projects. The number of items is more two hundred. Combining these large number vehicles and so many accounting subject together, the financial accounting becomes a very hard work. Moreover, the requirements for accuracy, but also increased the burden on staff.

Hangzhou Public Transport Group Co., Ltd Given have implemented MIS in relevant departments. These systems that has been running online includes passenger transport daily system, real-time billing system, supplies system and human resources system. Public bus accounting system uses real-time data provided by passenger daily system, real-time billing system, supplies system and human resources system. It can reduce the burden of accounting staffs, and provide group, company, team, circuit, cost of bus and profit statistics more timely and accurately. It offers the leadership making decision faster and more accurate.

Description of serial algorithm for public bus accounting
Accounting elements:

The needs of public bus accounting system of the business logic is very clear, witch is with the aid of basic double entry accounting methods and financial accounting principles, using real-time data of passenger daily, real-time billing, supplies, human resources systems. It takes the single commercial vehicle as the accounting unit to collect revenue, costs and expenses. Then complete comprehensive statistics of the single vehicle, the line, the fleet, the operating companies, the company's cost accounting and profit. Ultimately, it provides accurate reflection of Bus lines, the fleet of the company's operation

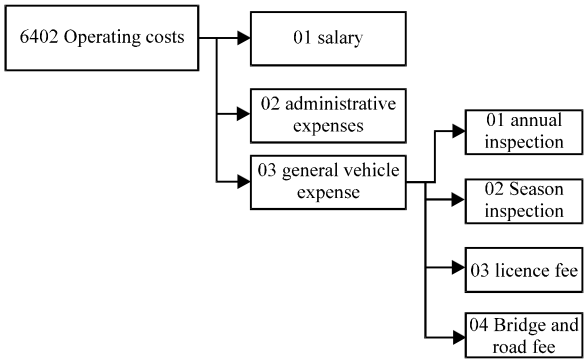

Fig. 1: Sub-project level accounting design

Accounting elements includes: (1) Information elements: vehicle number, line number, worker number, mile,; (2) accounting elements includes: income, costs, expenses, profits, (3) the accounting equation is: income - costs - expenses = profit.

Accounting Subject in the accounting system is the basic unit of accounting. Accounting Subject involved in the system includes: Main business revenue, main business cost, operating overhead, support operating expenses, taxes and surcharges, management fees, finance charges, profit, etc. These projects will maintain by the tree structure, each accounting project has a set of properties to define its distribution and formulas. The accounting subjects and basic financial subjects are same, but according to the company's management requirements will be more refined. Accounting Subject has been classified management, and it has been designed for the non-class limits. It can be infinitely set lower subjects. Numerical code at all levels of the Accounting Subject is recommended to set the format as shown in Fig. 1:

A subject with 4-digit (with the same accounts), subjects with lower 2-digit, level between the use of "." Separated, such as: license fee item number is 6402.03.03

Accounting method: In this study, it uses the cost position of accounting methods (Shen, 2007). In this method, the cost accounting of the accounting methods are usually divided into three categories to answer three questions: (1) What costs happen-cost category accounting, (2) Where these costs take place-the location of cost accounting, (3) For whom these costs occur-costs undertaker accounting. Bus accounting is based on financial accounting, for the revenue cost that cannot be

directly included in buses, and they are shared according to standard operating miles law. Statutory costs is accounted according to required calculation basis and proportion. A brief description about major subjects of the cost accounting method as follows:

- The main business revenue accounting is the actual operating income during a certain accounting period of the single-bus. It is subdivided according to income category on specifics; (2) Tax of main business and surcharges is actual tax accounting during a certain accounting period of the single-bus; (3) The main business cost accounting is the actual cost during a certain accounting period, as well as the cost allocation occurred by this. It is subdivided according to cost category on specifics, (4) Accounting for the indirect costs is the expenses that occurred during the operation of production of teams, including expenditures that station management incurred. The fee which can be charged directly.It should be accounted according to bus, (5) The statutory costs are shared according to required basis for the calculation and proportion; (taxes, three funds, five insurance payments a fee)

Cost-sharing method: As for the cost that need to share, the main business costs and operating overhead costs in the previous section, which algorithm to use has been controversial. The original share point is used in the standard model (Chen, 2008), that is, according to the time scale for bus cost-sharing. But in this way it is more difficult for data statistics and obtainment. This study introduces a kind of bus cost-sharing method by the scale of standard operating miles . The following is an example of personnel salary costs per kilometer to illustrate the cost allocation method. The bus companies are generally divided into management levels, they are generally divided into companies and fleet-level management personnel.

The company management salaries, for example, the name of accounting project: Indirect operating costs (4105)-management staff salaries (01); item code is 4105.01.

Suppose there are n units in a company operating the bus (n = 1, 2, 3 …), i-th vehicle to share the cost of 4105.01 is $cost_i$, Operation of the car mileage for the month is $rang_i$. The company has m management employees (m = 1, 2, 3 …), j-th of management salaries is $salags_j$. The cost 4105.01 that I-th bus allocated is $cost_i$, that is:

$$cost_i = \frac{\sum_{j=1}^{m} salags_j}{\sum_{i=1}^{n} rang_i} * rang_i$$

Similarly, staff sharing algorithm for the team management:

$$cost_i = \frac{\sum_{j=1}^{m} salacd_j}{\sum_{i=1}^{n} rangcd_i} * rang_i$$

Among this:

$$\sum_{i=1}^{n} rangcd_i$$

is the total mileage of fleet of all the operation for the vehicle:

$$\sum_{j=1}^{m} salacd_j$$

is the wages of the general managers of this fleet. Other methods of cost allocation is similar and will not repeat them.

Line cost-sharing method: In actual situation, there is no fixed management for line and vehicle management. For example, the temporary transfer of the vehicle and the changes in vehicle affiliation will cause the adjustment of vehicle line relations. However, due to the introduction of standard operating miles for the accounting, cost accounting methods may rely on the data of the vehicle line's daily report, which is attached to the vehicle mileage of the line of this day to solve the line costing.

The company management salaries, for example, the data which is read from the system daily passenger can draw the bus operation rang (i, j) that the i-th (i = 1, 2, 3 …) bus in the j-th (j = 1, 2, 3 …) line in the single month, the line of total operating mileage of place $rang_j$, then the i-th vehicle to 4105.01 of the subject company management pay costs to the j-line following the $costxl_{(i, j)}$, so:

$$costxl_{(i,j)} = \frac{rang_{(i,j)}}{rang_j} * cost_i$$

J-th line of 4105.01 of the company management wage costs $costxl_j$ is:

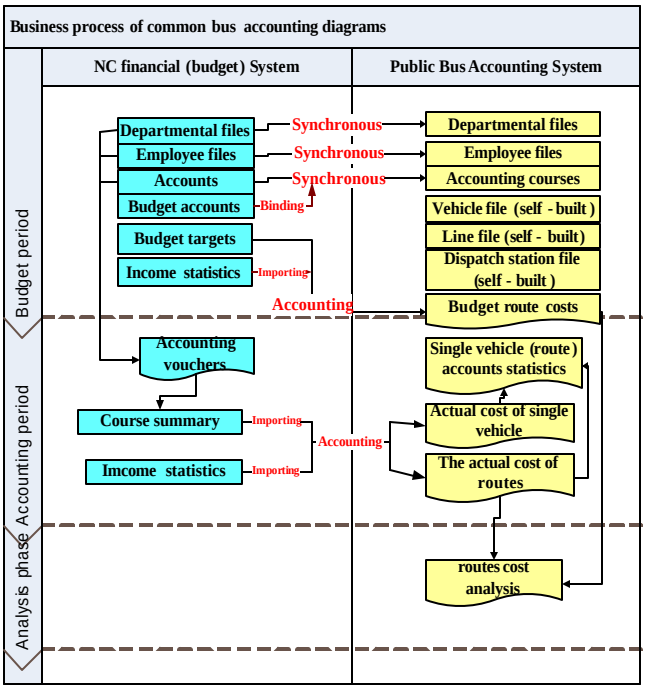

Fig. 2: Accounting business processes

$$\cos txl_j = \sum_{i=1}^{n} \cos txl_{(i,j)}$$

Other subjects are similar.

System construction: Here's an example of Hangzhou Public Transport Group's Public Bus accounting system to describe the system's business processes, where the budget of the UF system using the NC system software presented in Fig. 2. System consists of three components: (1) The first part is the bus accounting of the budget stage: All kinds of basic file which are synchronized with NC budget system, mainly includes: departmental files, personnel files, budget items; NC budget items will be introducing the budget targets and import tables to a revenue budget line accounting system to produce budget bus stage of the line budget costs; (2) The second part is the bus accounting of the accounting stage: at first, the branch does financial accounting in the NC system to form the subjects aggregate results as a Bus cost accounting basis; at the same time import or read data that the passenger Daily, real-time billing, supplies, human resources systems offer; then after bus accounting system and the line cost accounting; finally, account resulting from bus (line) accounting tables; (3) The third part is the analysis stage of the bus accounting: Do budget analysis of the line costs in the budget stage and the cost accounting stage to offer decision support data; the figure shows a business flow chart.

Bus cost accounting is an approach of integrated program, that is, completing the audit, accounting, billing in the same user interface. Specific process is: Firstly, do the accounting of each vehicle revenue, cost and profit according to the raw data, and then to allocate vehicles revenues and costs through the relationship between the vehicle and the line. After forming revenues, costs and profits of line, the team's accounting information is summarized according to the lines. At last, it summarizes the accounting information of branches and Group according to fleets (Fig. 3).

IMPLEMENTATION OF PARRELL ALGORITHM

Pseudo code description of MapReduce algorithm: This study uses scoop tool that apache offered to migrate the data from oracle database to Hadoop, and then implement the parallel algorithms of bus accounting by rewriting the Mapper and Reducer (Fig. 4).

When the program calls each MapReduce program, it will lead to the following operations:

Data partitioning: In the MapReduce model, the data will be divided into a plurality of data corresponding to the

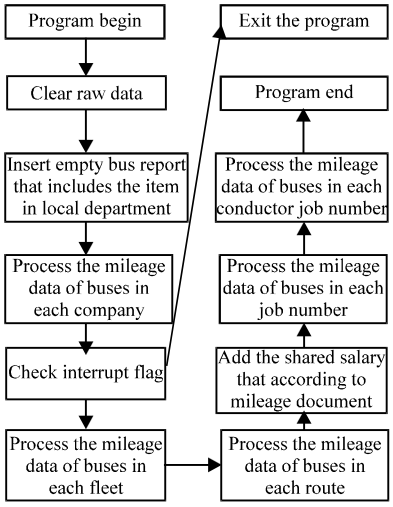

Fig. 3: Process flow diagram of bus accounting

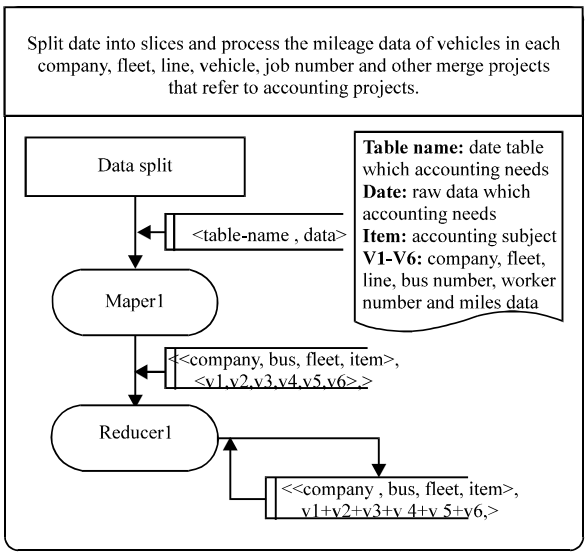

Fig. 4: Process flow diagram of MapReduce

size of the slice (slice data MapReduce model is the basic unit of data parallelism) according to actual needs. Then start on different machines in the cluster to copy slices in the copy program. For example, the data in the figure is a small slice which is divided out. Although other steps don't mark out data partitioning process, the data divide program will be called to do data partitioning before the mapping operation.

Tasks assign: All the programs in the MapReduce model have a master program (master), and others are workstation program (worker), which is referred workstation. Task performing of workstation is assigned by the master program. Master program master divides the task into M Map tasks and R Reduce tasks and select an idle workstation program of the cluster to perform various tasks. In the process of MapReduce parallel algorithms costing, there will be a master program to assign tasks to the workstation.

Workstations perform the mapper (Map): The workstation assigned to perform the mapper reads the corresponding data block, after the data block is processed, the corresponding key / value pairs are resolved from the raw data. Workstations perform the mapping function, then intermediate results of key / value pairs stored into the memory buffer. From the figure we can see the process of MapReduce input and output mapping function key / value pairs. For example, mapper of the input and output of key value pairs are <table-name,data> and << company, bus, fleet, item>, <V1,V2,V3,V4,V5,V6>>.

Processing intermediate results: The intermediate results stored into the buffer are regularly written to the local disk, partition function divides these data into R zones. These intermediate key / value pair results of the program send the location information on the local disk to the master program master, master program master is responsible for sending the location information of the intermediate results to the workstations that execute Reduce function.

The workstation that executes statute functions read intermediate data: The workstation assigned to execute the statute function knows the location that stores intermediate results from the master node and reads intermediate data from local disk on the workstation that executes the mapping function through remotely calling. After the workstation that executes the statute function finishing reading the intermediate data which is to be processed, the intermediate results of the key (key) are used as the key to sort the intermediate results. The key with the same key/value pairs are put into a class. Sorting operation is necessary, because in general, there will be many different key/value pairs sent to the same protocol (Reduce) task, if the same key/value pairs together you can simplify the operation of the Statute.

Workstations perform protocol (Reduce) function: The workstation assigned to execute the statute function traverse all the intermediate results according to each one unique key (key)after sorting, and transmit the intermediate result set to the user-defined statute function for processing. The output of statute function will add to

Table 1: Comparing between the serial and the parallel algorithm

Configuration (Master)	Configuration (Slave)	Algorithm	Time (sec)
1 Master entity		Serial algorithm	6673.67
1 Master entity	2 Slaves nodes, 1 entity machine	Parallel algorithm	6873.88
1 Master entity	4 Slaves nodes, 2 entity machine	Parallel algorithm	5353.65
1 Master entity	6 Slaves nodes, 3 entity machine	Parallel algorithm	2943.61
1 Master entity	8 Slaves nodes, 4 entity machine	Parallel algorithm	1387.76

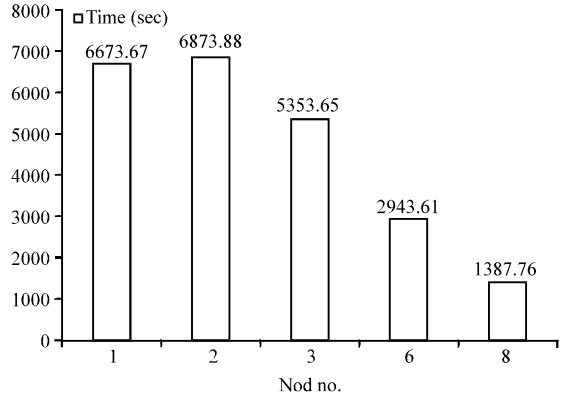

Fig. 5: Performance under different nodes number

the final output file. From the figure we can see the key/value pairs in the Statute of MapReduce process input and output functions, such as: Reduce input and output function key, value pairs were:

- <<company, bus, fleet, item>, <V1,V2,V3,V4,V5,V6>>
- <<company, bus, fleet, item>, V1 + V2 + V3 + V4 + V5 + V6>

Activate user program: When all of the mapping tasks and statute tasks are completed, the master program master activates the user program. When they successfully completed, MapReduce result data stored in R output files. Users can combine these R output files into one file and then migrated back with scoop tool to oracle database as a mean of persistence.

Experimental results: Compare to Hadoop system that spend the dozens or even hundreds of machines, virtual technology is employed to establish a flexible and powerful mini cluster to do real-time Public bus accounting analysis in Hangzhou Public Transport Group Co., Ltd. 10 g data is tested, the results are shown in Table 1, Fig. 5 and 6.

Every slave node configuration is virtual 4 cores, 4g memory, and one master machine Configuration is 8 cores, 8 g memory.

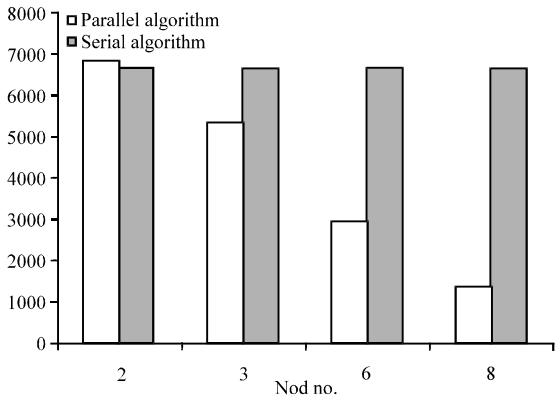

Fig. 6: Performance comparing

According to test results, in the environment of 10g data analysis, parallel algorithm based on MapReduce framework has no brilliant performance than the serial algorithm when the slave nodes is less than four. However, when the slave nodes is more than four, parallel algorithm based on MapReduce framework has showed better performance than the serial algorithm sharply.Yet due to limit of experimental environment, it can't obtain larger data to do experiment. However, when the level of test of data analysis increased, from 10G to 100G even T level, it can be predicted that parallel algorithm based on MapReduce framework will also show better performance than the serial algorithm.

ACKNOWLEGMENT

This study is sponsored by the nonprofit scientific and technological project of Zhejiang Province.China. And the project number is 2013C33077.

CONCLUSION

This study introduces MapReduce to the field of public bus accounting and gives detailed design and implementation of the algorithm, as well as introduces the framework of MapReduce to the implementation of public bus accounting. It is proved by experiment that this system can do huge calculating analysis efficiently. When data storage and calculating meets bottle-neck, the distributed technique has great advantages for flexibility and costs relative to traditional upward extension technique. The arise of open-source programming framework MapReduce can help us use distributed technique more easily. In fact, application space of MapReduce is highly large, these are to be proven by

further experiments. As the system is put into use, the larger data cluster will be tested in the future work. Optimization of the algorithm will be in progress as well to expend the function.

REFERENCES

Chen, H.Y., 2008. Discuss on Implementation of Public bus accounting in Urban public transport enterprises. Urban Public Transport, 9: 13-14.

Cohen, J., 2009. Graph twiddling in a mapreduce world. Comput. Sci. Eng., 11: 29-41.

Dean, J. and S. Ghemawat, 2004. MapReduce: Simplified data processing on large clusters. Proceedings of the 6th Symposium on Operating Systems Design and Implementation, December 26-28, 2004, San Francisco, CA., USA., pp: 137-150.

Dean, J. and S. Ghemawat, 2010. MapReduce: A flexible data processing tool. Commun. ACM, 53: 72-77.

Shen, L.B., 2007. Comparative study of cost accounting method. Friends Account., 7: 22-23.

Song, C.Y., 2010. Research on implementation of public bus accounting in Urban public transport enterprises. Traffic Accounting, No. 2.

Venner, J., 2005. Pro Hadoop. Apress, New York, pp: 6.

Wang, F., J. Qiu, J. Yang, B. Dong, X. Li and Y. Li, 2009. Hadoop high availability through metadata replication. Proceedings of the 1st International Workshop on Cloud Data Management, November 5-8, 2009, Hong Kong, China, pp: 37-44.