Send Orders for Reprints to reprints@benthamscience.ae

Open Access

The Cooperative Study Between the Hadoop Big Data Platform and the Traditional Data Warehouse

Ping Hu*

Tongren University Information Engineering College, Tongren, Guizhou, 554300, China

Abstract: In this paper, based on the application conditions of the existing traditional data warehouse and the future forecast of the Hadoop big data platform, this paper proposes the new framework of the cooperation of Hadoop and traditional data warehouse which focus on the cooperation between the traditional data warehouse and the Hadoop technique to solve the problem that the traditional data warehouse can hardly meet customers' demands. The new framework originated from the thoughts of the designers of Cloudera and Teradata, and in this paper, the new architecture is divided into three modules: data acquisition, data storage and data applications, this paper mainly discusses the consideration of structured and unstructured data collection, storage and application problem, and researches the Hadoop and traditional data warehouse in collaboration of data storage and data application. According to data collection and transmission problem, this paper uses the Apache Sqoop technology as the solution; and relies on Hadoop HDFS file system and the Hive data warehouse to store the data. At the same time, this paper also introduces the data application in the Hive. Finally, the prototype system proves the feasibility of the designed structure.

Keywords: Big data, hadoop, data warehouse, traditional data warehouse.

1. INTRODUCTION

With the development of Hadoop technique, it has adopted by more and more companies as the tool in dealing with the big data, it was originally used by Goggle and Facebook as the tool for the storage of large amount of data; the existing traditional data warehouse of the enterprises are being challenged. This paper has put stress upon the study of the coordination, divisions, data collections, transportations, storage and processing between the traditional data warehouse (without specific instruction, the traditional data warehouse mentioned in this paper refers to the single point relational data warehouse) and the Hadoop technique. The support of Hadoop technique is constructed on the base of the original data warehouse, in this way, the traditional data warehouse's deficiency in the processing and storage of the big data can be fixed; the bottle neck in the storage and calculation abilities of the traditional single point data warehouse can be solved through the lateral spreading ability of Hadoop [1].

Nowadays, Big data has become an important direction of development of modern information technology, and sharing and analysis of big data would not only bring immeasurable economic value, but also play a significant role in promoting the development of society. Big Data-as-a-Service (BDaaS) is a new data resource usage pattern and a new form of service economy, by encapsulating heterogeneous data, it can provide ubiquitous service consumers, standardization, on-demand services, including search, analysis or visualization [2].

2. DESIGN AND IMPLEMENTATION OF DATA MINING BASED ON HADOOP

Enterprise information systems usually contain multiple business system, and each business system contains its own set of business system, backup systems and archiving system. The disadvantages of this system are complex management, easily waster of storage space and poor system scalability. In consideration of these shortcomings, this thesis designs and implements a tiered storage system, using a large platform to manage the multiple business systems and making each business system backup and archiving system into one. The tiered storage system provides a data mining and data migration solution based on Hadoop framework.

Research Hadoop key technologies, including MapReduce distributed architecture, database, HDFS distributed file system (Fig. **1**).

Designed and implemented a tiered storage system based on Hadoop architecture, and detailed stated the system and data platform architecture design.

Designed and implemented the data mining module based on MapReduce. Applied the traditional relational database analysis method to the Hadoop database, efficient classifying the data of Hadoop [3].

Designed and implemented a data migration module, making the structured and unstructured data of online business platform migrate to the large data platform. Structure data migration process uses MapReduce data migration, which use thesis designed IO scheduling algorithm, considering the resources use and avoiding assigning tasks to the nodes with heavy IO load. Designed and implemented the unstructured data migration tool, using the FTP to multiple

1874-4443/15 2015 Bentham Open

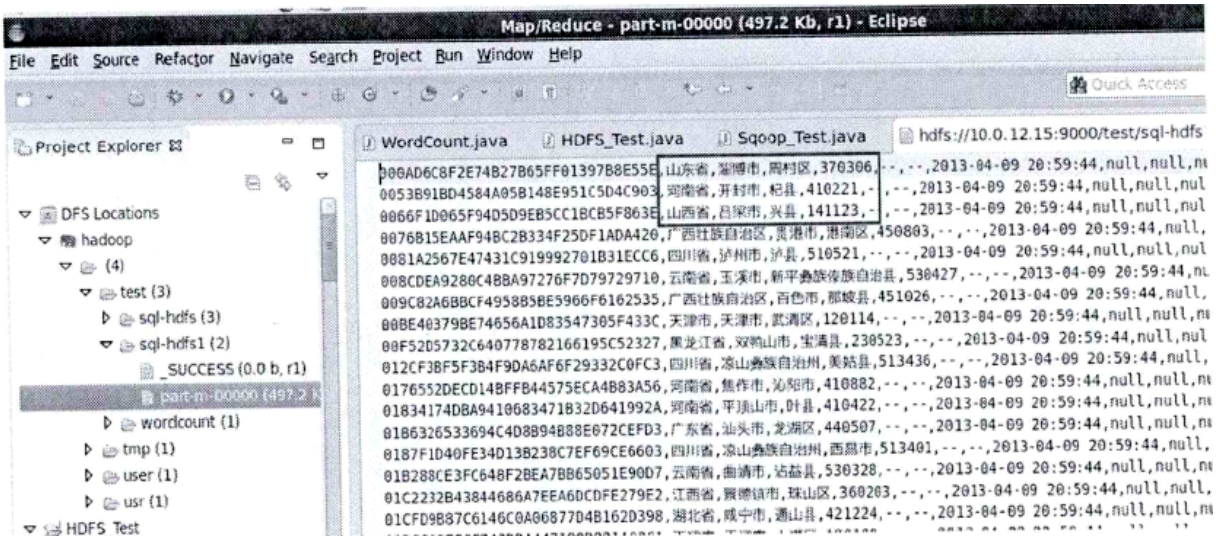

Fig. (1). Data acquisition.

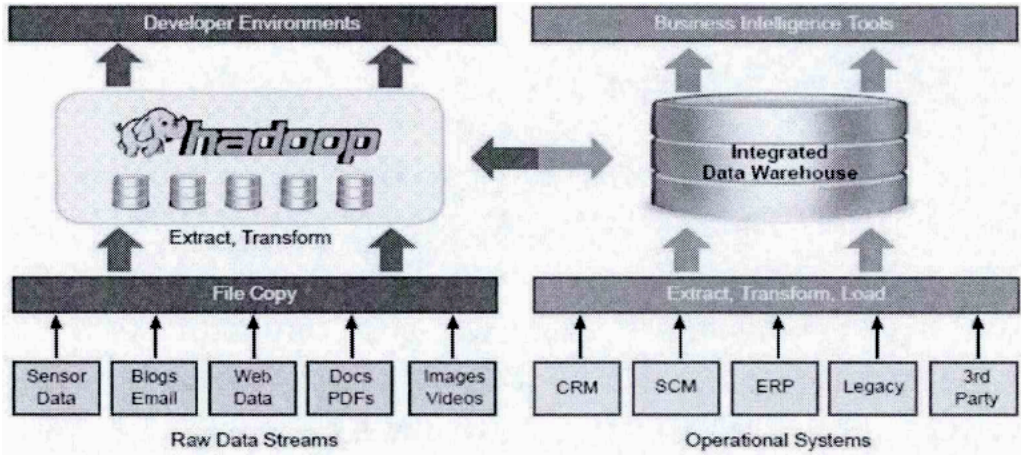

Fig. (2). Hadoop and traditional data.

concurrent migrate the business platform log file to the HDFS specified directory (Fig. **2**).

Complete test of functionality and performance of the system. The test results show that all the functional modules of the system satisfy the design requirement, and the IO scheduler has better performance than the default scheduler.

The thesis design data mining and data migration system to meet the specific needs. The system has better performance of concurrent data migration and better data analysis about consumers and business.

Big data are methods and a revolution in thinking. With hitherto unknown speed big data caused profound changes in society, economy, academic, scientific research, and national defense, military and other fields. Big data in addition to better solve the marketing problems, issues in the philosophy of science and technology; (Fig. **3**) there are all kinds of social problems, the formation of large data people-oriented strategy. Big data related to human being, more and more problems can be solved by large data. Not only in data science and technology level, but also in the commercial pattern, industrial pattern, ecological value and educational level, big data can bring new ideas and new thinking, including government departments, different industry and academia, even individual consumer. Big data and Internet, is not only the field of the information technology revolution, accelerate innovation, leading the social change and start the development of transparent government in the world [4].

This new concept of big data not only refers to the data object size, but also includes the processing and application of data, is the unity of technology and application of data object, the three. Data can be such as the government or enterprise master database this limited data set, can also be as micro-blog, Micro message, social network unlimited data virtual collection. Big data technology includes data acquisition, storage, management, analysis, [5] visualization

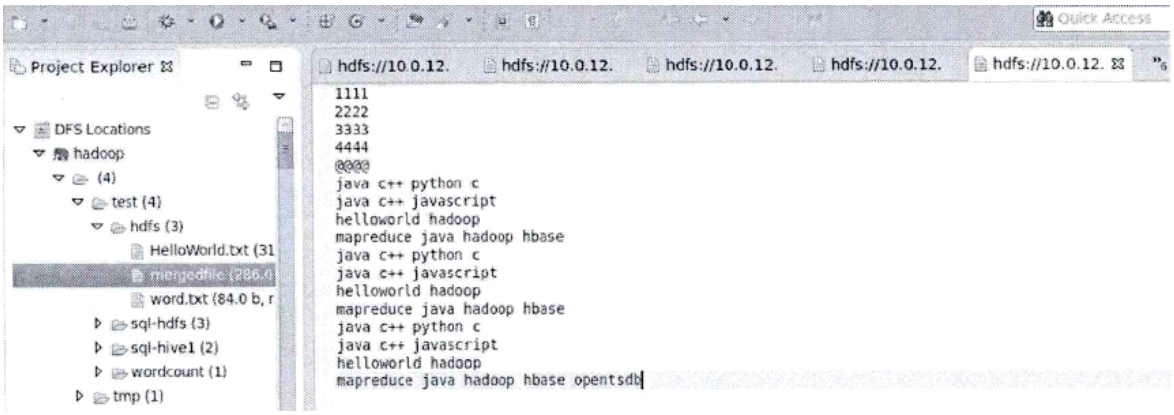

Fig. (3). Text storage.

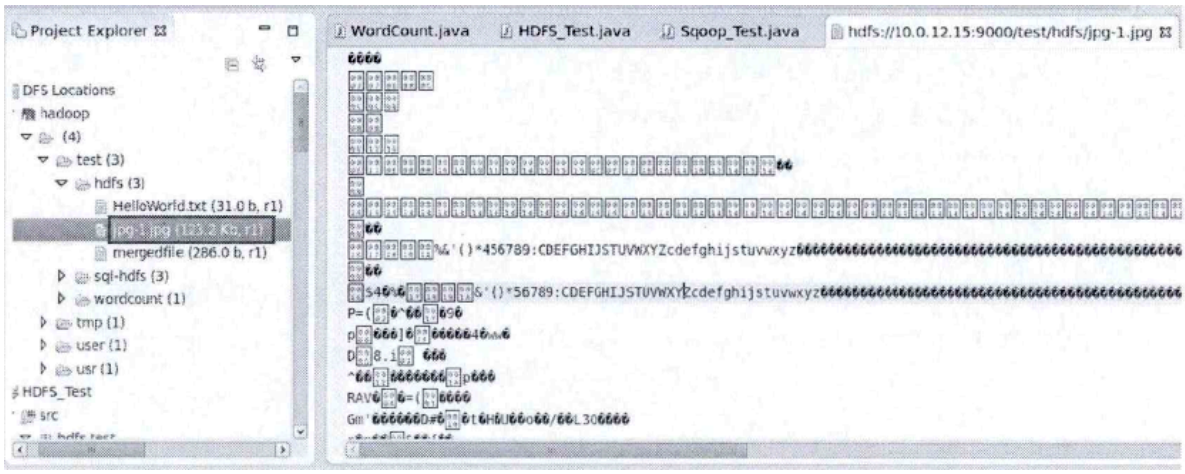

Fig. (4). Figures storage.

technology and its integration. Big data is of application data technology to obtain valuable information on various types of big data sets. The full realizable to the application object, technology, development of three-in-one synchronization. Big data is the area of information technology and the close integration of industry field, with strong demand and broad prospects. The need to seize the opportunities continue to track the research data and continuously enhance the awareness and understanding of big data, insist on technological innovation and the application of collaborative innovation to accelerate at the same time data in all areas of economic and social development and utilization, promote the country, industry, enterprise for the data application requirements and the level of development in the new stage [6].

Big data is of application data technology to obtain valuable information on various types of big data sets. The full realization of the value of big data only adheres to the application object, technology, development of three-in-one synchronization. Big data is a typical area of information technology and the close integration of industry field, with strong demand and broad prospects (Fig. **4**). The need to seize the opportunities continue to track the research data and continuously enhance the awareness and understanding of big data, insist on technological innovation and the application of collaborative innovation to accelerate at the same time data in all areas of economic and social development and utilization, promote the country, industry, enterprise for the data application requirements and the level of development in the new stage as shown in Table **1**.

Table 1. Development tools.

Tools	Edition
Hadoop	hadoop-1 .1.2
Hive	hive-0.11.0
Sqoop	sqoop-1.4.4.bin_ hadoop-1.0.0
SQL Server	SQL Server 2008 R2
MySQL	MySQL-5.5.31-1
Eclipse	eclipse-SDK-4.2.2-linux-gtk-x86-64

```
hive> select province_name,state_name,city_name,city_code from hive_st_def_city
limit 5;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_201312180941_0026, Tracking URL = http://hmaster:50030/jobdet
ails.jsp?jobid=job_201312180941_0026
Kill Command = /usr/hadoop/libexec/../bin/hadoop job  -kill job_201312180941_002
6
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2013-12-20 10:07:14,323 Stage-1 map = 0%,  reduce = 0%
2013-12-20 10:07:18,375 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.66 se
c
2013-12-20 10:07:19,389 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.66 se
c
2013-12-20 10:07:20,432 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.66 se
c
2013-12-20 10:07:21,443 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 1.66
sec
MapReduce Total cumulative CPU time: 1 seconds 660 msec
```

Fig. (5). Command line under the HQL query graph.

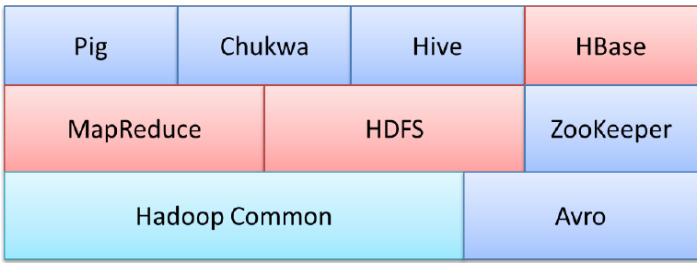

Fig. (6). Hadoop component.

Big data has three main effects on the human enormous economic and social development, to sum up: the first is to promote the realization of the enormous economic benefits, the second is to promote and enhance social management level, and the third is to promote safety and security (Fig. **5**). Big data applications in government and public services can effectively promote government work, government departments to improve service efficiency, decision-making and social management level, has great social value. To make a long story short, big data will provide a powerful new tool for people, so that people can more easily grasp the law of things, more accurate understanding of the world, to predict the future and change the world (Figs. **6-8**).

Big data problem involves a wide range, in the study focus on the analysis of concept analysis and literature. Do the problem oriented, have a definite object in view at the same time, to establish the theory of professional knowledge, the deep research relates to the topic, pay more attention to comprehensive and unified field theory knowledge.

Data share is made to operate the data that stored in other systems by different computers and users, and these computers or users could compute and analysis those dataset. It reported common problem of many software systems and current situation of development of Hadoop in this paper. Also the theory of distributed database and parallel computing and Hadoop technology ware studied. Then key technologies ware analysis.

We researched the architecture and theory and defects of traditional data share models-middleware data share model, data warehouse data share model and P2P data share model (Fig. **9**). In order to resolve defects of traditional data share model, Hadoop data share model is put forward, based on data warehouse data share model. We also analyzed the advantages of Hadoop data share model by number of study method, compared to traditional data share models. At the end, so as to prove the feasibility of Hadoop data share model, the Hadoop data share model is used in material dispatcher system. During the process of realizing material dispatcher system, functional requirements ware presented by function analysis method, then carried on the detailed design and produced main function modules. During the process of realizing material dispatcher system, the problems of material coding disaccord and data join with Hadoop database are solved.

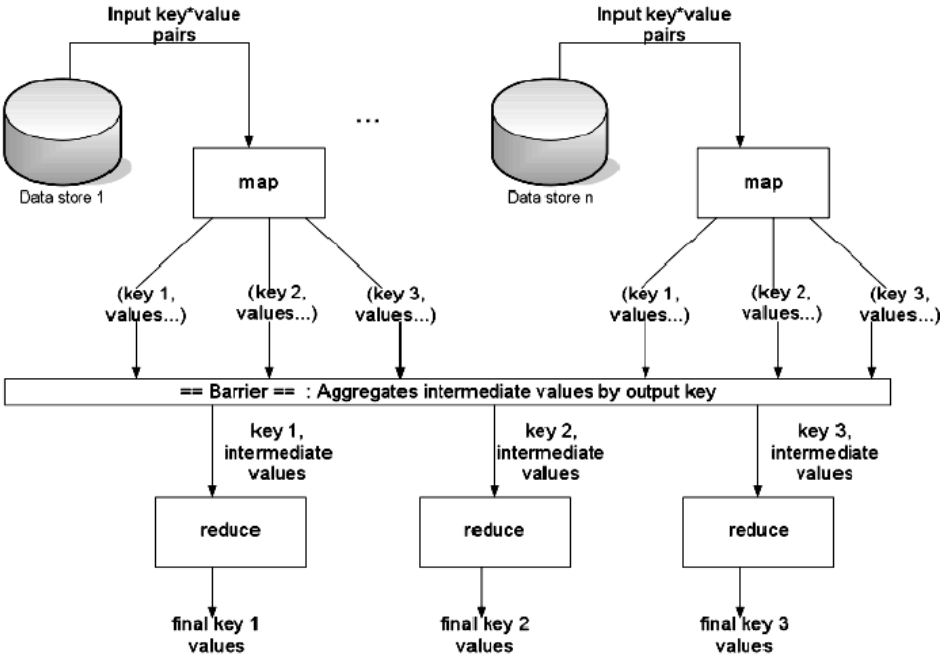

Fig. (7). Workflow of mapReduce data processing.

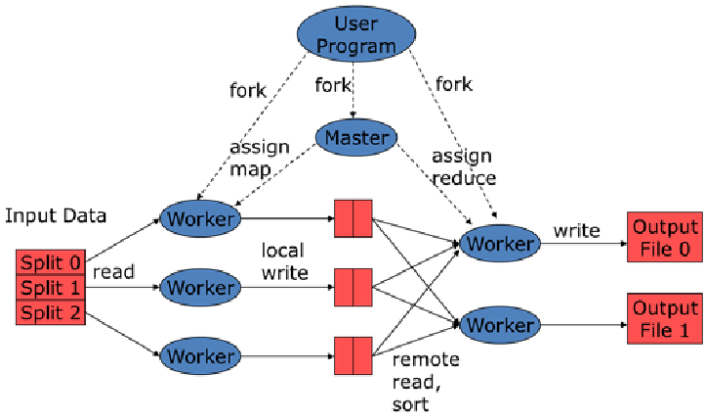

Fig. (8). Hadoop data flow.

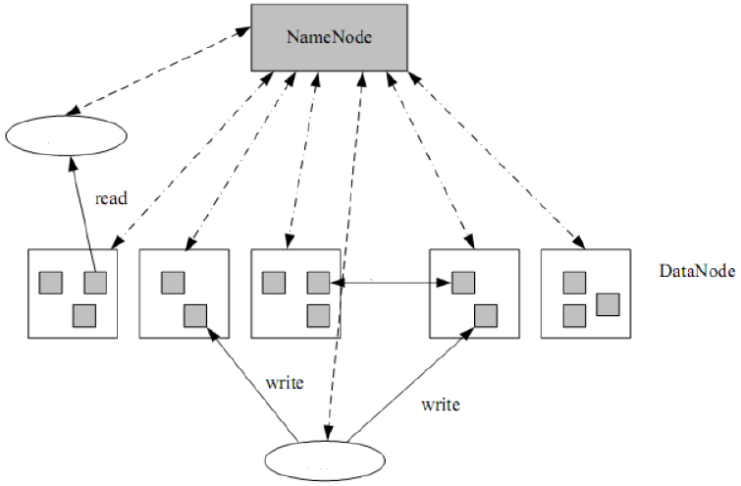

Fig. (9). Hadoop architecture.

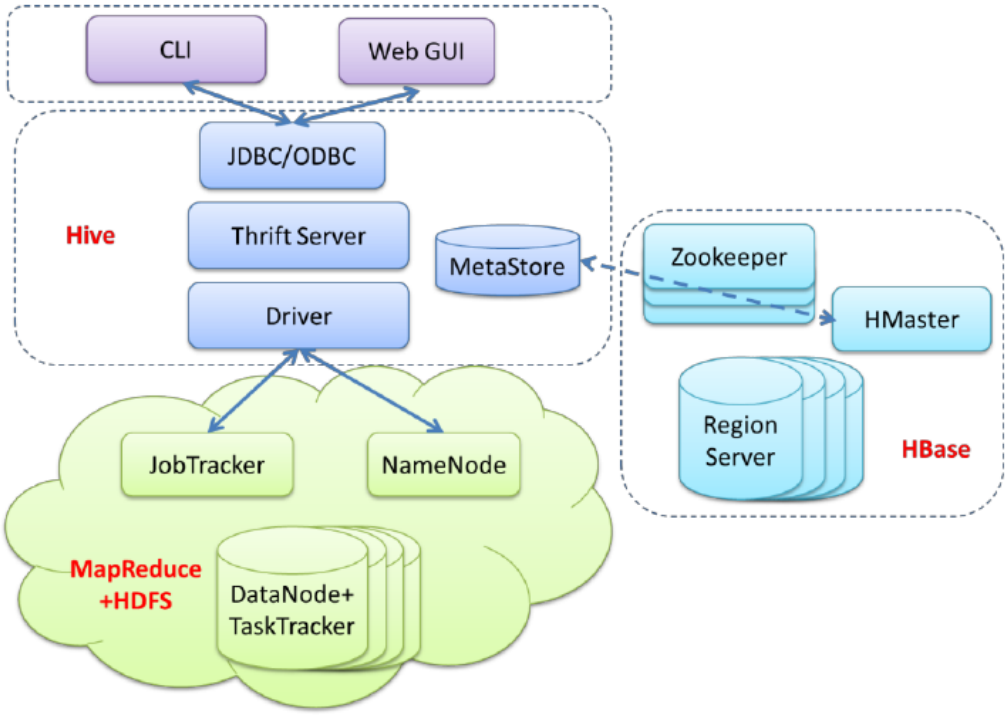

Fig. (10). Data mining module architecture.

Take an example: Taobao.com (referred to as Taobao), attaches great importance to strategic decision support data of operational services and data, and enhance unitized planning strategic level to create a competitive advantage, this company project development and integration for the data warehouse system integrated into the enterprise within the framework of the system of the data platform.

In this background, discusses the design and development of data warehouse systems Taobao data platform. Firstly, the characteristics of the current complex Internet enterprise data and summarized based on the actual needs of the enterprise needs analysis, focusing on analysis of the data warehouse storage engine selection, data extraction\transformation\loading module an important part of the data model and data presentation of the e-commerce industry portal four data warehouse, a detailed description of its function points and establish its use case diagram. The next analysis for the needs of the various modules and systems design, analysis and design to integrate both the technical points of each main module's internal functions and external access, including e-commerce and physical data model design, and finally in system design based on the implementation of the data warehouse system is given.

Since the trial run of the system to good effect on Taobao data analysis and processing, and has made a large contribution to company data operation services and data strategy support (Fig. **10**).

The creation and the application of Data Warehouses is the only way for the enterprise to realize the advanced informationalization. In recent decade, lots of different scales Data Warehouse Systems appear to solve history data management and decision support problem. At the same time, the conflict between the nature anti-lifecycle character of DW and currently popular software development technology appears obvious day by day. And now, the DW technology is faced with such serious problem: at the beginning of its creation, the DW system cannot help users to accurately define business meta-data (for determining the Dimensions-Measures space) and integrated meta-data (for determining the integration space); while running, the software also cannot automatically capture the change users demand and give responses, and it cannot distribute the heavy load of making DW entity and query load of using DW in larger scale. With the application background of the development of PBC Dalian branch "disposal information service system", the decision supporting subsystem of Dalian international cooperating group information system and China life insurance company Dalian branch “Estimate and planning of Market the decision supporting system” paper applies itself to resolve the above conflict and problem, in order to enhance the Self-adaptive of DW system, to implement integration optimization within DW and measure the users compatibility, and to get a self-adaptive DW system .During the process, it uses DW technology, DW quality, CORBA structure and Agent technology (Fig. **11**).

3. HADOOP DATA PLATFORM

With the rapid development of information technology, human society has entered the digital information age. The ability to obtain and master the information has become a symbol of national strength. All information with the needs of different determines the effectiveness of different, but all

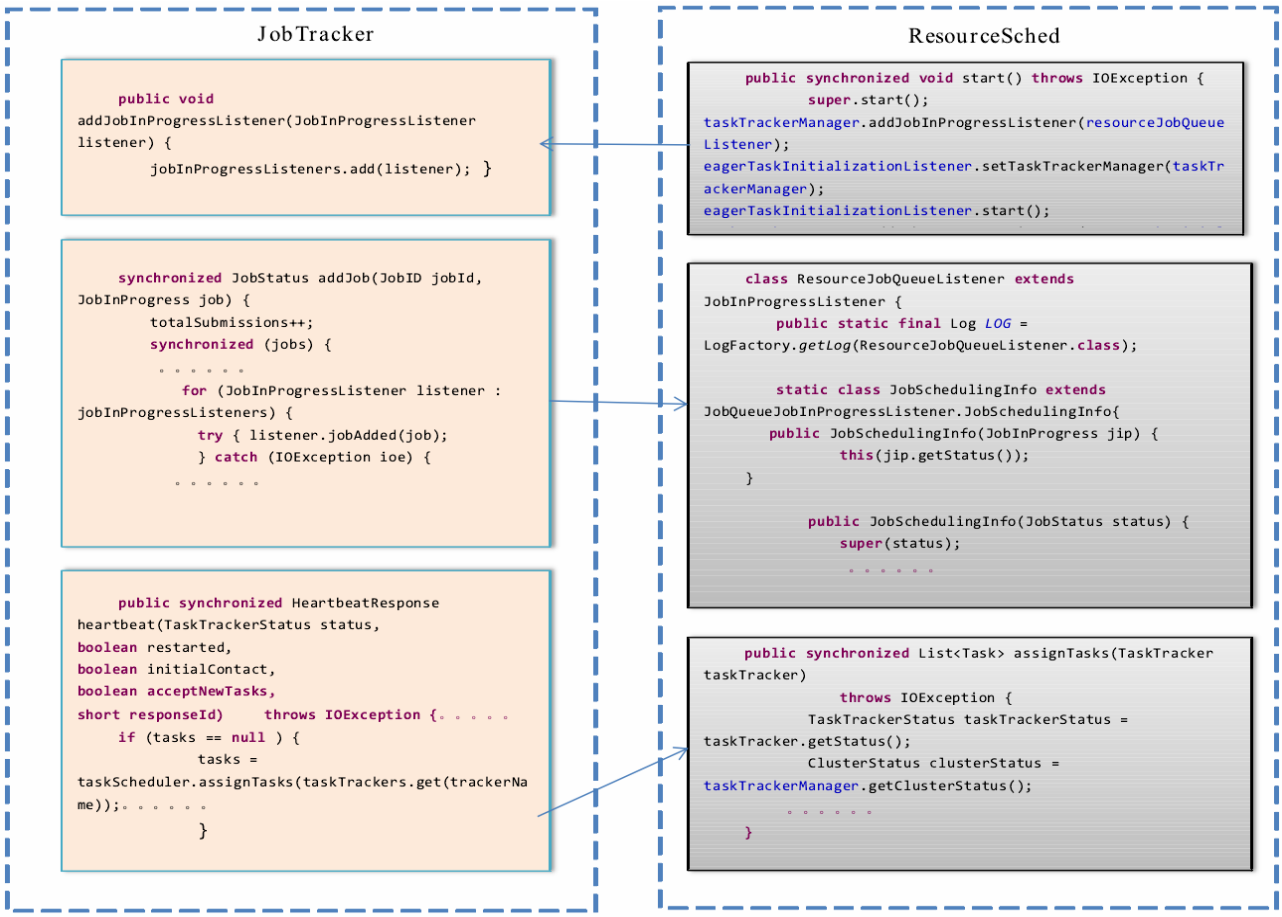

Fig. (11). Relationship between resource scheduler and JobTracher.

the useful information is extracted from large amounts of data. Large amounts of data over time, and continued to produce continuous flow, the formation of big data and diffusion. Big data is not only used to describe the data amount is very large, but also highlighted the speed of processing data. Therefore, big data has become a frontier technology in the field of data analysis. Data become an important factor in every industry and commercial field. Mining and use of a large number of data, not only marks a substantial surplus industrial productivity growth and consumer, but also clearly suggest that the era of big data has come.

Data are becoming the basic factors of production are equally important and material assets and human capital, the use of big data has become the key factor to improve the competitiveness of enterprises. Data become assets, industry unity, the Internet is the three development trends in the age of big data. The scale and the ability of use of a country's data will be an important part of comprehensive national strength. The right of possession and control data will become national core of power as important as a land power, sea power, air power. Data become key elements of production and equally important like land, capital, manpower (Fig. **12**).

In order to decrease the development cost, to improve the multiplexing of components and to realize load balance and distributional calculation in larger range, paper studies and realizes the self-adaptive architecture of DW with the CORBA technology, Agent technology and Java programming language. As a result, the DW Architecture get the ability of initiative, permanence and intelligence. It is a kind of architecture, which can accept many effective components and the "Plug and Play" components, and support many kinds of OS platforms, database platforms and network platforms, can conveniently communicate between components and software architecture, components and components, components and user. And sustained by agent, it realizes the distributional calculation, load balance and self-adaptive according to definite criterion in definite range. This architecture is made of lots of Agent and server software, including the component series, database series, quality series, service series and maintenance series. It fully assimilates and makes use of software development technology and component multiplexing technology. Especially during the calling component process, it breaks through the limits that have to use the object sign to connect and the simple name service, and realizes the load balance and distributional calculation.

3.1. Key Codes

```
publir boolcan sqoop_import sql_ hdfs(){
 tryl
 l-n}pertTee} impTool=new ImportTool();
```

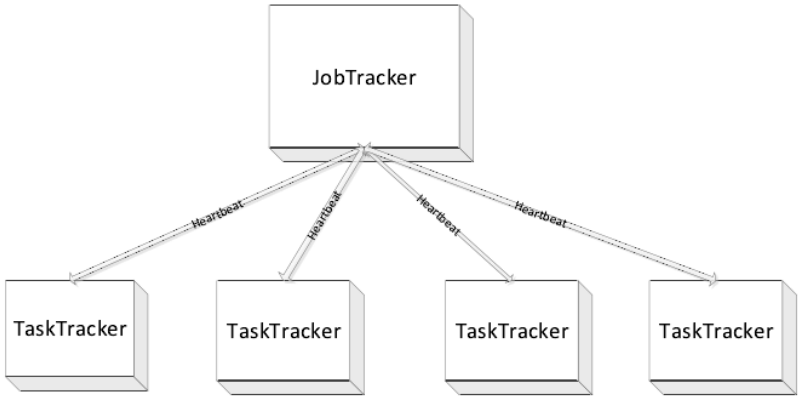

Fig. (12). JobTracker and TaskTracker interactive.

```
}qeeptieps options=new SqoopOption};
c}ptic}n}.sctDriv-
::rC'lass\;}r}L("can}.wi}rcS,iUt.sellsrrw,s    :IcIhc.St})L.S};
werDriv}
op-
tions.setHadoopMapRedHome(Iladoop:llapRedHome);
options.setConnectString(lserwerC:oramclSlring);
options.setUsername(SyhserrerUsernanJe);
options.setPassword(lserverPas.su}arr/);
options. setTargetD ir( I/dfs7 argelDir);
options. setTableName( lldfsTablelVanre );
options.setSpIitByCol(fldfs/mSplilByC ol);
options.setlnputFieIdsTerminatedBy('\001');
options.setNumMappers(1);
impTool.run(options);
}catch (Exception e){
return false;
}
return true;
}
```

3.2. Study and Implementation

To adapt the requirement of DW individuation and respond to the need of DW change, based on the optimization function of the current DW production, paper brings forward DW integration optimize-path searching algorithm and the DW integration scheme self-optimize algorithm, and studies and realizes the DW integration optimization. A task of top priority to the DW system which is used frequently, is to update the DW meta-data constantly and periodically, through the machine study, with the special software system, to make DW respond to users demand with best time-space complexity, and in turn, to enhance the implement efficiency and lengthen the lifetime of the DW system. After giving the entire introduce of DW optimization, paper discusses the solving scheme and realization method of integration optimization in detail. During the realization of integration optimization, based on former integration scheme relation drawings, paper brings forward the integration scheme optimize-path searching algorithm and self-optimize algorithm and gets the integration scheme objects ensemble which is low-cost and satisfies the current users demand.

To quantize the users demand to DW, paper introduces into the concept of "DW Quality", establishes a more complete DW quality architecture through applying it further, and studies and realizes the DW data using quality. The DW data using quality is the best user compatibility of DW. The DW data using quality is subjective, and has to quantize quality to give play to its effect and make it have more targets and decisive factors, which is easy to judge it comprehensively. So, paper studies and analyses the DW data using quality further, and brings out the solving scheme of DW integration users compatibility.

CONCLUSION

Based on the needs for the big data platform of existing enterprises, this paper also gives concerns about the reuse of the existing traditional big data warehouse and studies the cooperative relations between the two sides. Finally, the Demo system realized in this paper provides guidance for the enterprises to realize.

CONFLICT OF INTEREST

The author confirms that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] P. Dong, J. Dong, and T. Huang, "Application of data "warehouse technique in educational decision support system," In: *IEEE International Conference on Service Operations and Logistics and Informatics,* SOLI'06, Shanghai, 2006, pp. 818-822.

[2] F. S. C. Tseng, and C. Chen, "Integrating heterogeneous data warehouses using XML technologies," *Journal of Information Science*, vol. 31, no. 3, pp. 209-229, 2005.

[3] M. N. Vora, "Hadoop-H Base for large-scale data," In: *International Conference on Computer Science and Network Technology,* ICCSNT, Harbin, 2011, pp. 601-605.

[4] S. K. Rahimi, and F. S. Haug, *Distributed Database Management Systems: a practical approach*, John Wily, New York, 2010.

[5] Goldfarb, and F. Charles, "*XML Handbook*," Prentice Hall PTR, London, 2014.

[6] Z. Tari, and O. Bukhres, *Fundamentals of Distributed Object Systems: the CORBA Perspective*, Wiley PTR, London, 2011

Received on: May 26, 2015 Revised on: July 14, 2015 Accepted on: August 10, 2015

© Ping Hu; Licensee *Bentham Open.*

This is an open access article licensed under the terms of the (https://creativecommons.org/licenses/by/4.0/legalcode), which permits unrestricted, non-commercial use, distribution and reproduction in any medium, provided the work is properly cited.