Send Orders for Reprints to reprints@benthamscience.ae

Research of Massive Spatiotemporal Data Mining Technology Based on Cloud Computing

Shuxia Wang* and Zenggang Xiong

School of Computer and Information Science; Hubei Engineering University, Hubei, Xiaogan, 432000, China

Abstract: This paper mainly discusses developing a new mass data analysis system in the face of the challenges, based on cloud computing technology, cloud service provided by Amazon and HBase. Based on the research of the EC2, S3 and EMR service provided by Amazon and on the detailed analysis of the client reception server and data handling server of the data analysis system, this paper tries to create a preliminary model of the system. Then, each model and whole system is finalized according to actual demand. Finally, system test and performance test prove that the system can solve the problems faced by traditional data analysis system. The research runs through the whole process of data analytical system design and implementation. Firstly, we present the problem domain and requirements. By investigating the traditional data analysis system and by analyzing the problems existing in this system, we raise the developing requirements of the massive data analysis system based on cloud computing, and point out that the key points of the new system implementation are receiving server and data processing design. Secondly, we design and realize each module of the massive data analysis system. Finally, results of system test and performance test prove that the new data analysis system is stable and that the requirements are satisfied.

Keywords: Big data, cloud computing, data mining, massive spatiotemporal data.

1. INTRODUCTION

With the development of information technology, data analysis is playing an increasing role in business decision and product design. Based on analysis of user behavior data, product designers can design functions with an enhanced user experience and can improve functions of other parts. Decision makers of a company can use the result of data analysis to decide product orientation and the company's developing direction. However, the development of the company, the increase of products and users and other factors lead to a huge increase of the total quantity of data. Facing the challenge of data saving and handling capacity, traditional data analysis system cannot be able to accomplish its mission [1].

With the dramatic increase in the amount of data storage, massive data processing and massive data computing have become important issues in the field of data mining. The traditional serial data mining algorithms are often only able to handle small-scale data, when faced with massive data, their execution speed will be reduced or even not run, so it poses severe challenges and Ordeal for the current data mining. And classification algorithms, as one of the most important part of the data mining, which plays an important role in information retrieval, Web search and CRM. Currently, the vast majority of classification algorithms are serial, feasibly poor and inefficient in dealing with large sets of data, the problem of low classification accuracy have become increasingly prominent, leading to immeasurable computing resources and unlimited extension of implementation time.

2. RELATED THEORY OF DATA MINING AND CLOUD COMPUTING

Today's society has been dealing with massive data before the advent of cloud computing. In the past, when making data mining always looked forward to using high-performance machine or even a large-scale computing device to process. And also, in the context that massive data mining process needs to have a good development and application environment, [2] the use of cloud-based computing approach to data mining is more appropriate. Due to the lack of parallel sorting algorithms, currently large-scale datasets increasingly getting larger day by day; therefore, traditional data mining system can no longer be effective on these massive data and their usage. How to improve the efficiency and parallelism of the algorithm is a problem which needs to be solved urgently.

This thesis is based on an integrated system of comprehensive information security program of laboratory, analysis and research of massive data mining-related technologies, which were applied in this program. As the data, which the public opinion analysis system (POAS) processes, are from the Internet, the amount of data to deal with is becoming very large day by day. In order to train and classify this massive dataset, ensure that POAS is maintained at a stable and efficient rate. How to improve the efficiency and performance of classification of POAS is the main problem of this thesis, as shown in Fig. (**1**).

*Address correspondence to these authors at the School of Computer and Information Science; Hubei Engineering University, Hubei, Xiaogan, 432000, China; Tel: +86- 13407874545; E-mail: 10340012@qq.com

1874-4443/15

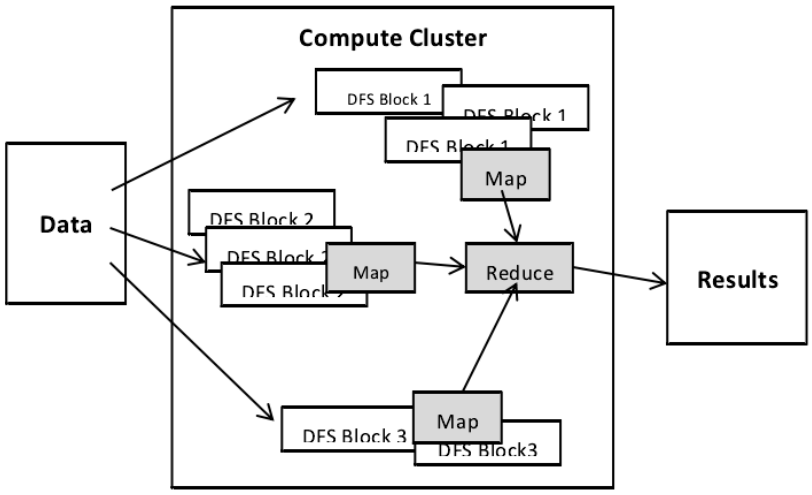

Fig. (1). Architecture of Hadoop.

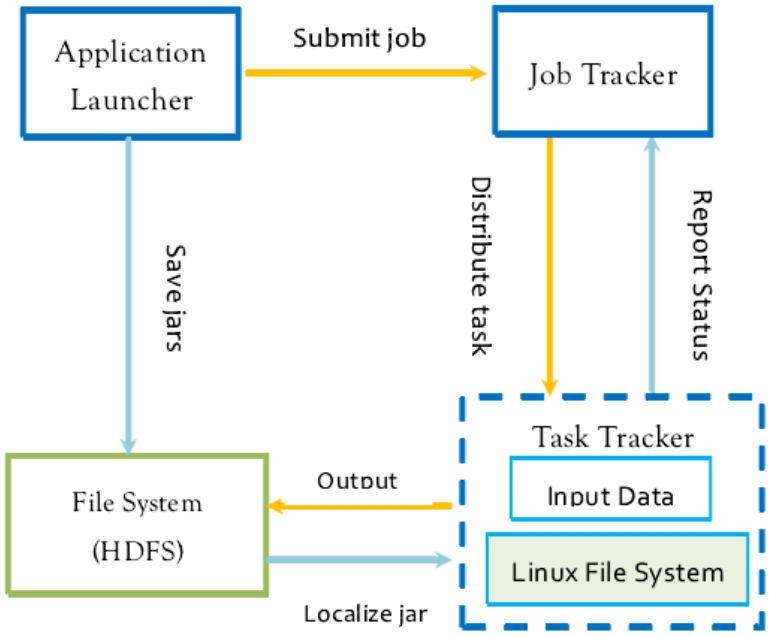

Fig. (2). Hadoop module design.

This thesis puts forward advanced theory that the classification algorithm plays an important role in POAS. According to POAS requirements and system design, a classification algorithm module is designed based on the Strategy pattern for POAS. And designing different parallel classification algorithms, through packed classification algorithms under MapReduce framework, [3] greatly improve the efficiency of classification algorithms, making the classification algorithms speedup close to linear speedup. Through this module, POAS can dynamically call different classification algorithms to classify public opinion data, improving the performance and efficiency of classification algorithms of the system, thereby greatly improving the stability and reliability of POAS, as shown in Fig. (**2**).

Cloud computing is a business computing model, which assigns the computing tasks to a large number of computers in the resource pool. It can provide users with computing power, storage capacity and application service capabilities according to their needs. Cloud computing provides cheap and efficient solutions for storing and analyzing mass data. [4] Data mining is the process of discovering information or patterns that are interesting, non-trivial, implicit, previously unknown and potentially useful in large databases. Data mining plays a guiding role on scientific research, business decisions and other fields, with far-reaching social and economic significance. Data mining needs to use huge computing and storage resource, so integrating cloud computing and data mining can effectively control computing cost and enhance the efficiency of data mining, breaking the bottleneck of the traditional limitations of data mining. It is very important to research the data mining strategies based on cloud computing from the theoretical view and practical view.

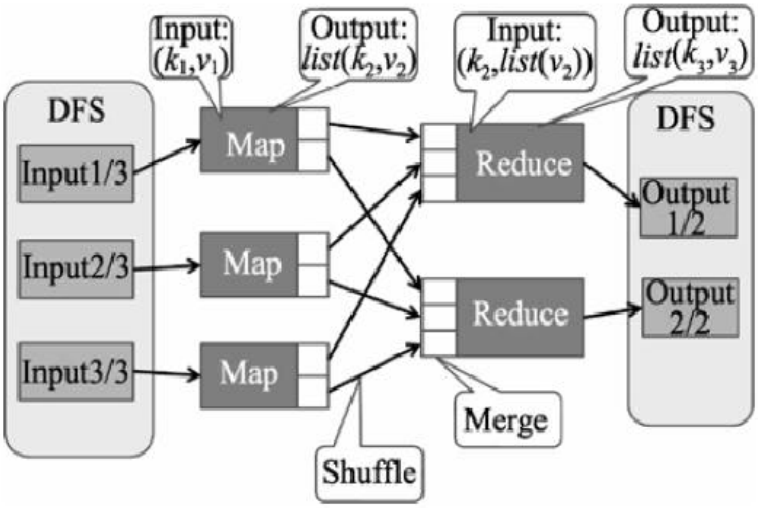

Fig. (3). MapReduce data stream.

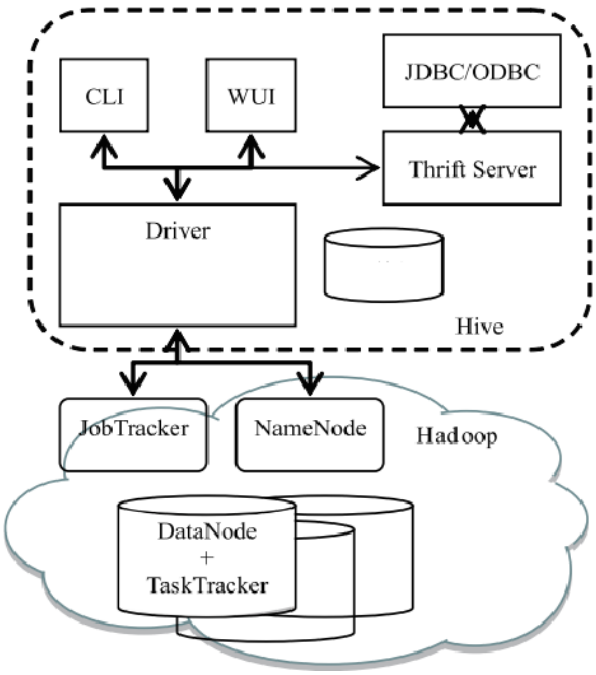

Fig. (4). Hive system architecture.

Hadoop is the most famous open source distributed computing framework, [5] through the use of MapReduce parallel model to effective integration of the computing storage capacity in order to provide a powerful distributed computing capabilities. This paper mainly focuses on the Fig. (**3**).

The development of network techniques brings people in a great deal of information. It also greatly increases the difficulty to find useful knowledge from mass data. The efforts to solve this problem promote the emergence and rapid development of data mining techniques. At present, the data mining technologies and tools have been used in the financial, medical, military, and many other areas of commercial decision-making analysis, as shown in Fig. (**4**).

Cloud computing is a style of computing in which dynamically scalable and often virtualized resources are provided as a service over the Internet. Cloud computing platform can be used to develop high capability programmers. However, it does not provide data reduction service which is one of the bases of data-mining. So, the method to implement data-mining system on cloud computing platform has not been worked out yet.

To solve these problems, in this paper, a data reduction module for accessing isomeric and new type data is designed and implemented to expand Google App Engine cloud platform. Furthermore, a new data-mining system built on top of cloud computing services is designed and implemented to

verify the validity of data reduction module and the efficiency of cloud based data-mining process.

Data reduction module supports uniform definition of datasets by using meta-data definition, dataset definition and dataset instance definition to abstract data type, data structure, and location information and so on. Layered thinking model is used and a new motile layered plug in architecture is induced to enhance the scalability of the system.

Meanwhile, many important thoughts and methods to design and implement the platform are induced in this paper and a prototype of data-mining platform based on this architecture is introduced in the last section of this paper. Experimental results show that the proposed method is very promising.

3. MASSIVE SPATIONTEMPORAL DATA MINING TECHNOLOGY

Lots of implicit knowledge and information exist in GML spatiotemporal data, including spatiotemporal model and its characteristics, the general relationship between spatiotemporal data and common data, and the general data characteristics existing in GML spatiotemporal data and so on. Humans can learn about the natural world by means of getting into relationships, discipline and interaction existing in nature, which are used to make decision for our production and life. However, due to the GML temporal-spatial and half structural characteristics, an accurate mode could not be identified to define GML spatiotemporal data. So, it is much more complicated to extract information from GML spatiotemporal data than traditional data. At the same time, because of characteristics of quantity and the various computing intensive features, [6, 7] the information processing progress is limited to some degree. To resolve all these problems, this paper puts forward a two-parallel clustering mining algorithm in parallel computing Hadoop environment. We designed a parallel clustering GML mining prototype system, followed by showing the clustering visuals in the form of map.

This paper proposes two kinds of parallel clustering algorithm for mining GML spatiotemporal series data. The first kind is put forward with K-means clustering mining algorithm based on time sequence of GML spatiotemporal similarity, considering the spatial attribute and temporal series together to measure the similarity between the two spatial adjacent objects. Then, through the K-means clustering algorithm, the mining is conducted. The second is based on the definition of spatial neighborhood to get the GML spatial neighborhood in GML spatiotemporal objects; and then in this neighborhood, mature the two objects of temporal series of the attribute based on temporal sequence to getting the similarity, combining parallel DBSCAN (STN-DBSCAN) clustering algorithm for data mining. See Fig. (**5**) for detail.

Through constructing Hadoop distributed parallel computing platform and using MapReduce programming model, two kinds of clustering algorithm means and DBSCAN parallel process are realized. By designing and implementing a parallel GML spatiotemporal clustering prototype mining system, a two-parallel clustering algorithms using GML format attributes of the meteorological data is realized. Through the experimental results we verified the effectiveness and quality of high performance of the two-parallel algorithm. Good performance of the algorithm can be expanded.

Finally, we show the results of income cluster in the form of map.

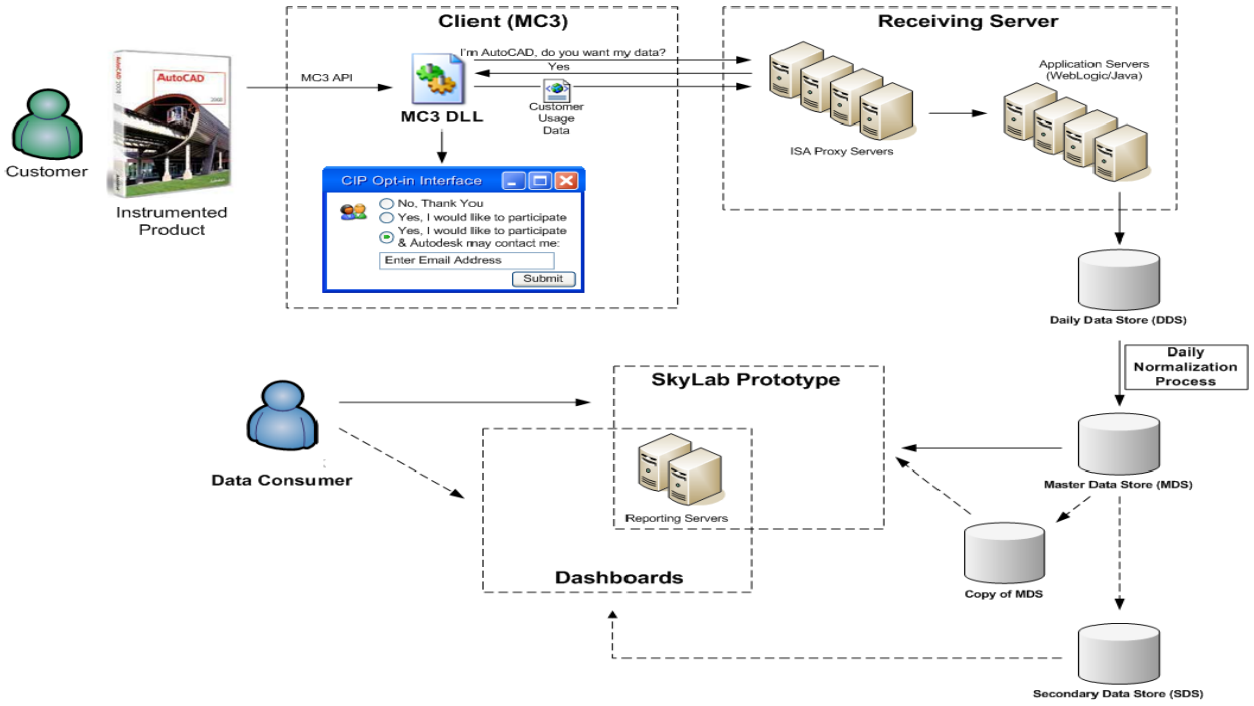

Fig. (5). Architecture of data analysis system.

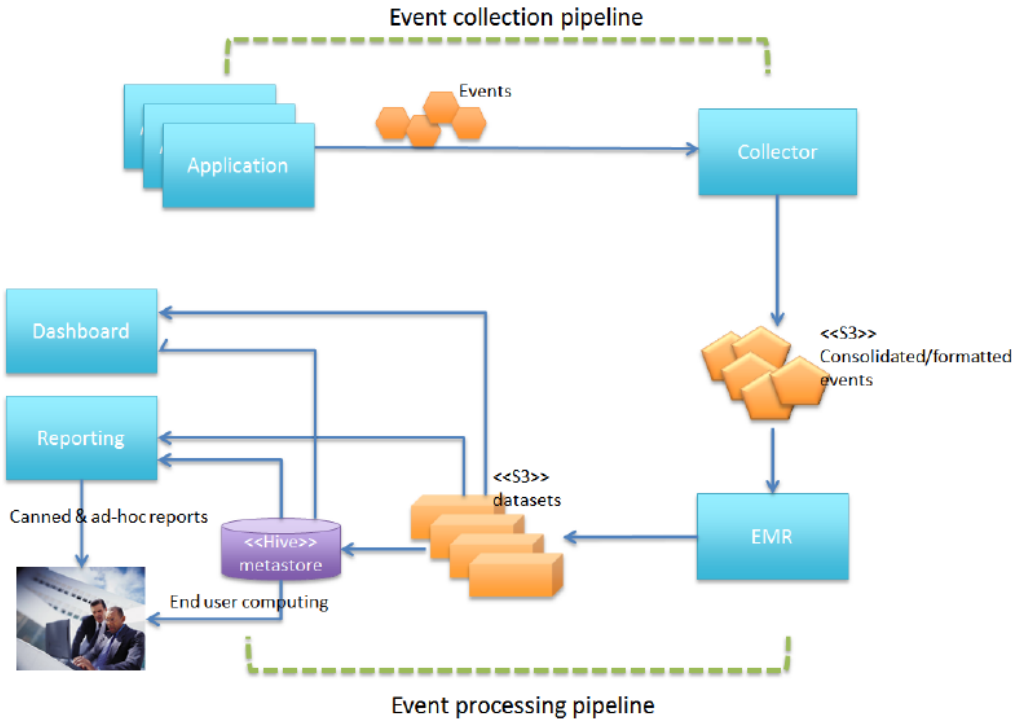

Fig. (6). Architecture of AP2.0 system.

3. DATA ANALYSIS

With the use of computer and Internet technology and expansion of many aspects of human society, data shows explosive growth. Now, the process and storage of large datasets has become a new challenge for enterprises. How to mine valuable and understandable knowledge from the massive data in a more rapid, efficient, low-cost way, to help company make decisions, is a challenge of data mining. The emergence of cloud computing brings new opportunities for data mining technology. Cloud computing distributes the ability of storage and computing among multiple nodes in cloud cluster. So it enables huge dataset storage and computing power. Because you can use a lot of cheap computers for clustering instead of a high priced server, cloud computing can greatly reduce costs. Together with cloud computing technology which can provide the large storage capacity and computing power, data mining technology enters the cloud-based data mining era.

HADOOP is an open source project of Apache for building cloud platform. HADOOP framework will help us implementing clusters easily, fast and more effectively. HADOOP uses HDFS (distributed file system) to achieve large file storage and fault tolerance, and the MapReduce programming model for computing. To make HADOOP apply for data mining, a key question is how to parallel the traditional data mining algorithms. Some specified traditional data mining algorithms can be paralleled easily because of their own characteristics. But some other algorithms are hard to be paralleled. For the algorithms that can be paralleled, combined with MapReduce programming model, we can transplant them to HADOOP platform. Then, the data mining tasks will be completed more efficiently in a parallel way. See the Fig. (**6**).

Cloud computing is a hot topic at home and abroad; it is the development of current high performance computational model, besides being a numerical model of virtualization through the network to provide services through dynamically scalable resources. Through the cloud computing, people can get dynamically extensible computing and storage capacity on the network. Cloud computing can improve the data processing efficiency, while reducing the terminal equipment requirements; it can effectively solve the problems which mass data processing faced. Therefore, cloud computing based on the distributed data mining platform, is a hot research topic.

This paper is based on the practice of the key breakthrough project, by presenting analysis and study of the technology of cloud computing and data mining with a focus on the DBSCAN clustering algorithm based on density. Aiming at the shortcomings of DBSCAN clustering algorithm and combined with the project of charging station data characteristics, this paper proposed a new algorithm. This algorithm is the DBSCAN clustering algorithm based on grid control factor, which is used in the fixed grid size DBSCAN algorithm in the project as the foundation. In order to find a better clustering accuracy grid size, a grid control factor value is required to adjust the size of the grid. The paper has proved that the new algorithm has the improved clustering accuracy by the test of charging station data, and it can effectively reduce the time complexity, as shown in Figs. (**7**-**10**).

Second important problem to be solved in this paper is to perform the parallel processing of the improved algorithm, and then realize it on the cloud computing platform. To carry out the clustering analysis of massive datasets, we must ensure that the system is maintained at a stable, efficient environment. The paper designs a parallel algorithm based on

```
{"event":[
   {"eventType":"image_size"},
   {"version":"v1.0"},
   {"email":"wei.feng@autodesk.com"},
   {"namespace":"pixlr2012"},
   {"language":"java"}],
 "argument":[
   {"name":"image_width","type":"float","required":"yes"},
   {"name":"image_height","type":"float","required":"yes"}]}
```

Fig. (7). JSON String of image_size event definition.

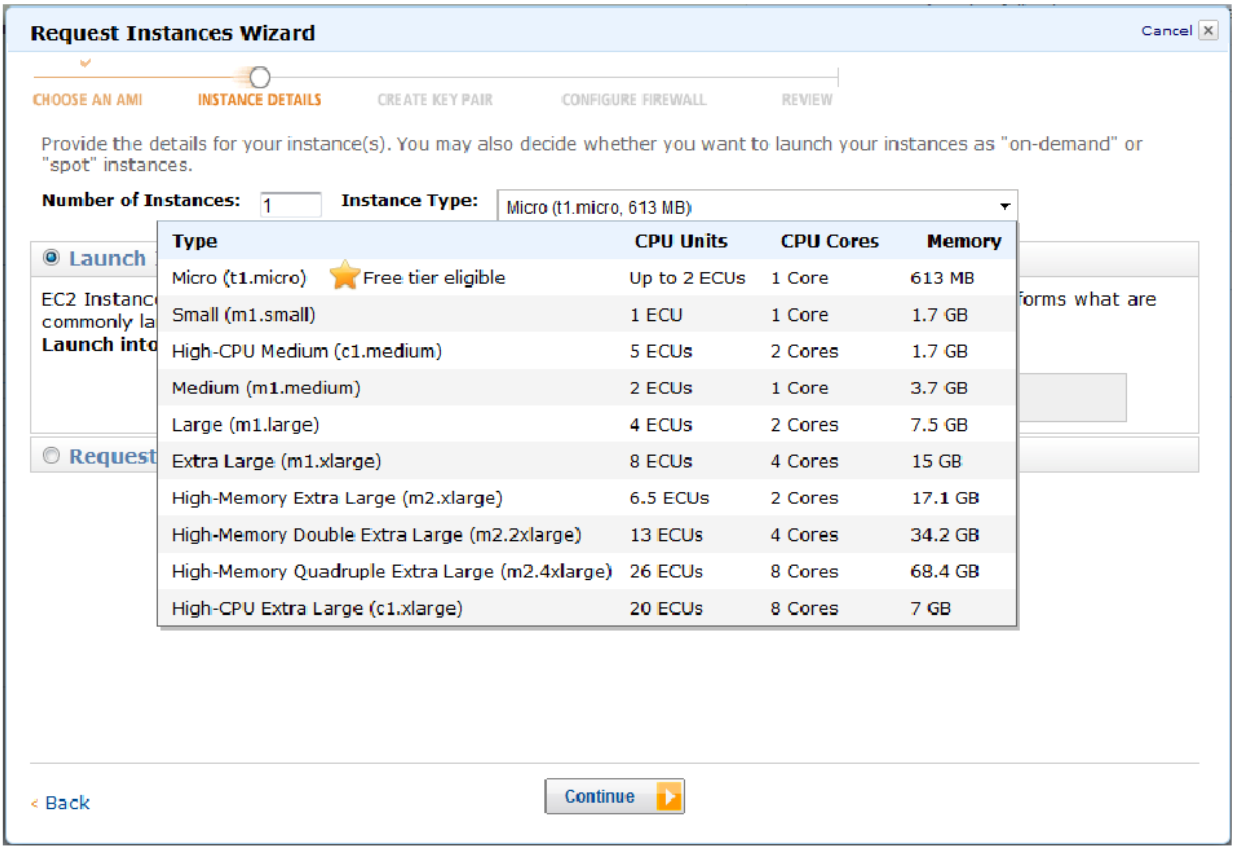

Fig. (8). Types of amazon EC2 instance.

Hadoop, built a simple Hadoop environment, through the encapsulation of the DBSCAN clustering algorithm in the framework of MapReduce, which greatly improved the efficiency of this algorithm. Finally, the improved algorithm is verified based on cloud computing by using replication large-scale charging station data. The experimental results show that the DBSCAN algorithm based on cloud computing greatly improved the processing efficiency of large datasets, on the condition of not reducing the DBSCAN clustering quality.

To help cloud computing take root, it is necessary to adopt various mature technologies to the cloud computing paradigm. Some of them are listed below.

As cloud computing software platform is the heart of a cloud computing system, it will require considerable further research. Hadoop is an open source cloud computing software platform, as an alternative to the platforms developed by Google and others. It appears to be a good vehicle as a launching point for research. Yahoo is a major sponsor of the Hadoop project. IBM has adopted Hadoop for its Blue Cloud solution. Facebook uses Hadoop in its data analysis. Google,

```
Rem ============= No edit allowed below! ==============
Rem call %ap20BackendRootPath%\AggregationClient\lib\getYesterday.bat 2012/01/02
%This line is for test%
call %ap20BackendRootPath%/AggregationClient/lib/getYesterday.bat %SYSDATE%
set DATA_DAY=%YESTERDAY:~0,4%-%YESTERDAY:~5,2%-%YESTERDAY:~8,2%

cd %EMR_CLIENT_ROOT%

ruby elastic-mapreduce --create --name "AP20_New_User_Exposure_Aggregation_Daily"
--num-instances 3 --master-instance-type m1.large --slave-instance-type m1.large
--hadoop-version 0.20 --log-uri %LOG_ROOT% --hive-script --arg %SCRIPT_ROOT%
/tableCreation/ap20_spt_upi.q --hive-versions 0.7 --args -d,DATA_ROOT=%DATA_ROOT%
--step-name Create_Table_ap20_spt_upi --hive-script --arg %SCRIPT_ROOT%
/tableCreation/ap20_agg_startup.q --hive-versions 0.7 --args -d,DATA_ROOT=%DATA_ROOT%
--step-name Create_Table_ap20_agg_startup --hive-script --arg %SCRIPT_ROOT%
/tableCreation/ap20_agg_new_user_history.q --hive-versions 0.7 --args -d,DATA_ROOT=
%DATA_ROOT% --step-name Create_Table_ap20_agg_new_user_history --hive-script --arg
%SCRIPT_ROOT%/tableCreation/ap20_agg_new_user_daily.q --hive-versions 0.7 --args
-d,DATA_ROOT=%DATA_ROOT% --step-name Create_Table_ap20_agg_new_user_daily --hive-script
--arg %SCRIPT_ROOT%
/aggregation/newUserExposureAggregation/ap20_new_user_exposure_aggregation.q
--hive-versions 0.7 --args -d,DATA_DAY=%DATA_DAY% --step-name Join_New_User_Exposure_Data

exit
```

Fig. (9). Configuration of aggregation script.

```
create external table if not exists ap20_agg_new_user_daily
(
 data_day                string,
 region_name             string,
 country_name            string,
 marketing_release_name  string,
 unique_new_user_count   bigint
)
partitioned by (day string)
row format delimited fields terminated by '|' lines terminated by '\n'
stored as textfile
location '${DATA_ROOT}/AggData/AggNewUserExposureDaily';
```

Fig. (10). Script of table creation.

IBM, and Yahoo have donated cloud computing platforms to six US universities. All the computing platforms include Hadoop. Hadoop may become the Linux of cloud computing.

Such means of collaboration as chat, instant messaging, Internet phone calling, *etc*. will be added to various popular applications. Google Docs spreadsheets already make it possible for multiple users to chat while editing a spreadsheet together.

The research on these subjects can leverage the available EAI, EII, and ESB technologies.

Transmitting the bulky multimedia data across the network will continue to be a challenge, and it needs further research to speed up cloud computing. Further, as more data gets pushed to the clouds, including user-created data, the need to analyze (mine) such data to derive business-useful knowledge will increase. The data mining and machine learning communities will need to address this need.

As the clouds proliferate and the users start plugging into multiple clouds, the problems of discovering and composing services that have been subject of research in the service-oriented architecture context, will need to be revisited in the cloud computing context.

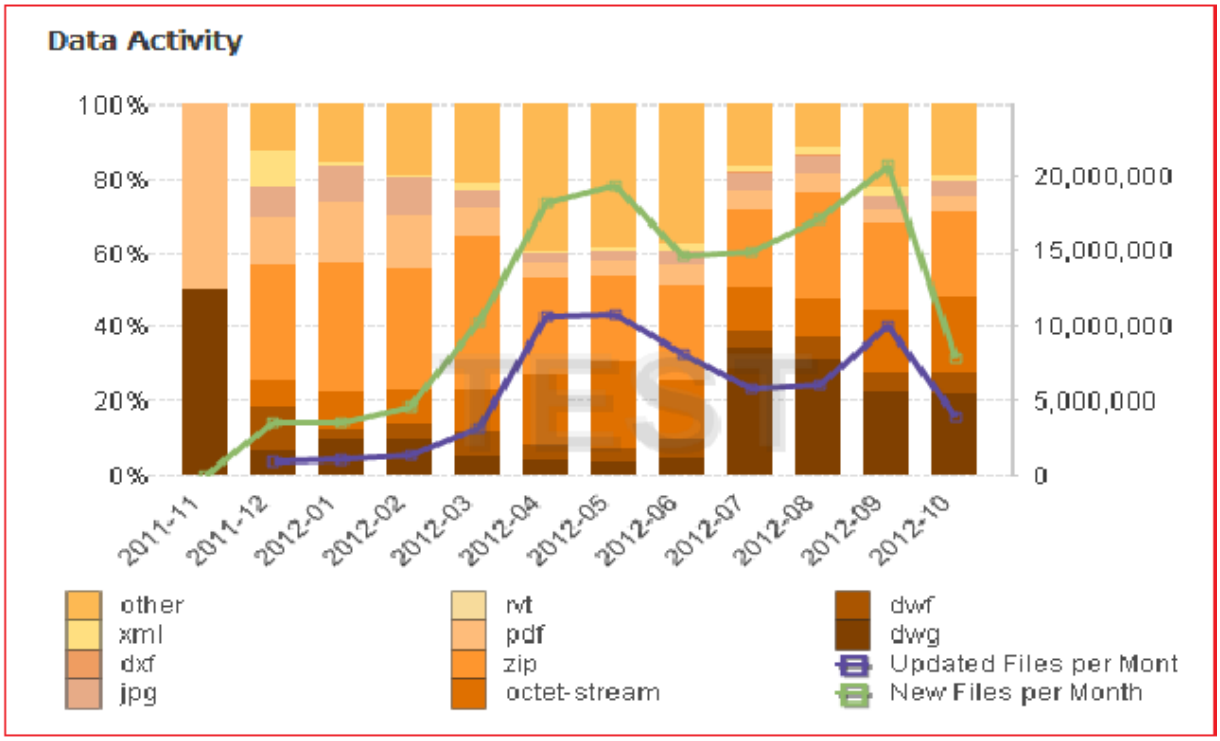

Fig. (11). Data activity analytics.

Data activity analytics are shown in Fig. (**11**). While in many cases more research is needed to make these cryptographic tools sufficiently practical for the cloud, we believe they present the best opportunity for a clear differentiator for cloud computing since these protocols can enable cloud users to benefit from one another's data in a controlled manner. In particular, even encrypted data can enable anomaly detection that is valuable from a business intelligence standpoint. For example, a cloud payroll service might provide, with the agreement of participants, aggregate data about payroll execution time that allows users to identify inefficiencies in their own processes. Taking the vision even further, if the cloud service provider is empowered with some ability to search the encrypted data, the proliferation of cloud data can potentially enable better insider threat detection (*e.g.* by detecting user activities outside of the norm) and better data loss prevention (DLP) (*e.g.* through detecting anomalous content).

CONCLUSION

Currently, the cloud computing and data analytics system are in the booming development period. This research runs through the cloud computing and Amazon cloud Web services, and used them in enterprise data analysis system implementation. The data analysis system has been modified based on could computing, so that the problems which existed in traditional data analysis system could be resolved, in order to arouse data analyzing and handling capacity of the system, save the cost, and cut down development cycle.

After using cloud computing technology, data analysis system is found to be more flexible in data processing and analysis; it can increase or reduce the computing resources to meet the actual requirement. Also, the spending of equipment maintenance and data backup will be saved. Both from the viewpoint of data processing system performance and cost of the project, using cloud computing technology will be the best solution. In this study, the actual project successfully confirms and embodies the advantages of cloud computing technology. In the research and application of cloud computing technology, this paper has some certain reference value.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This paper is founded by science Research Project of Hubei Provincial Department of Education (NO. B2016181).

REFERENCES

[1] J. Dean, and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," In: *Proceeding 6th Symposium on Operating Systems Design and Implementation,* Berkeley: USENIX Association, 2014, pp. 137-150.

[2] F. Chang, J. Dean, S. Ghemawat, W.C. Hsieh, D.A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R.E. Gruber, "Bigtable: A distributed storage system for structured data," In: *Proceeding 7th USENIX Symposium on Operating Systems Design and Implementation,* USENIX A ssociation. Berkeley: 2006, pp. 205-218.

[3] *Welcome to Hadoop MapReduce [EB/OL]*. Available from: http://hadoop.apache.org/mapreduce/

[4] *Google Docs Glitch Exposes Private Files*. Available from:

http://www.pcworld.com/article/160927/google_docs_glitch_exposes_private_files.html.

[5] *IT Cloud Services User Survey*, *pt.2: Top Benefits & Challenges.* Available from: http://blogs.idc.com/ie/?p=210.

[6] D. Boneh, and B. Waters, "Conjunctive, subset, and range queries on encrypted data," In: *The 4th Theory of Cryptography Conference* (TCC 2007), 2007.

[7] *Lithuania Weathers Cyber Attack, Braces for Round 2*. Available from:http://blog.washingtonpost.com/securityfix/2008/07/lithuania_weathers_cyber_attac_1.html.

Received: June 16, 2015 Revised: August 10, 2015 Accepted: September 19, 2015