Send Orders for Reprints to reprints@benthamscience.ae

The Open Automation and Control Systems Journal, 2015, *7*, 1621-1628

Open Access

Research of Social Network Log Analysis System Based on MongoDB

Yuzhen Shi*

College of Software, Pingdingshan University, Pingdingshan, Henan, 467002, China

Abstract: This paper mainly studies the design of an efficient social network log analysis program based on the distributed database, MongoDB. The so-called social network log analysis is to gather and store the log information generated when users access the social network pages, and then transform, clean and excavate. This article compares MongoDB database with traditional relational database, analyzes its advantages and application scenarios. Its anti-paradigm design due to the nest avoids the association, making queries and storage of the large data efficiently by: storing social network logs in the MongoDB; directly analyzing the logs with its built-in MapReduce programming model, and saving the results of the analysis as files for business people to use. Our study aims to discover the hidden users' access rules and patterns in the log data by effective data mining of the social network log data, thereby providing helpful information for optimizing website structure and business model.

Keywords: Log analysis, MapReduce, MongoDB.

1. INTRODUCTION

With the growth and expansion of the Internet enterprise scale, social network log information also grew step by step. In order to provide better services, it is necessary to understand the features and needs of the user's access, by analyzing and studying the user's behavior, so analysis of the Social network logs was suggested. It combined the traditional data mining and social network logs, by getting the useful information from a large number of social network logs data, counting and analyzing the users' behavior and page view, to finally infer the user's access modes. [1] It can play a role in many situations, such as network security, build the social network site and perform the market analysis of the e-commerce. It is a new research direction for data mining.

NoSQL is the general name of the non-relational databases, which is a new data storage technology to meet the needs of the rapid growth of the Internet applications. Since it is easy to extend, has high write and read performance even for the large amount of data, and has flexible data models, it have been well developed in some application scenarios. MongoDB is a representative of NoSQL databases, the document-oriented data model it used, can automatically split the data and store them on different machines. This automatic slicing mechanism achieves a distributed extension, the collections and documents in the database can be stored in many database nodes. The applications of MongoDB are very wide because of its good horizontal scalability: it is suitable for storing low-value and large-sized files; offers a data management technology which satisfies the high concurrency and the magnanimous data processing from the Internet development to cloud computing. Internet provides to meet the high concurrent development of cloud computing. Such characteristics make it possilbe to have a good development in the field of social network log analysis.

*Address correspondence to this author at the College of Software, Pingdingshan University, Pingdingshan, Henan, 467002, China;
Tel: +86 1896765445; E-mail: yuzhen@163.com

2. RELATED TECHNOLOGIES

With the rapid development of the Internet, logging a large amount of data generated by the network every day is a huge task. How to solve the problem of massive log data processing has been a very important research topic in the field of log analysis. With the rapid development of network technology, [2] data on the Web is exponentially growing, and has the forms such as a massive, diverse, heterogeneous, dynamic change, which makes centralized log analysis based on a single node at the platform impossible to meet the massive data network analysis requirements. To design a common scalable platform to effectively deal with the massive log data, and perform analysis of Web page visits, has been the inevitable choice of the Internet enterprise development.

For solving the problem, an analysis of the key technical features of the existing distributed storage and computing systems was made, combined with the analysis and research of the Hadoop platform which was designed and implemented based on the mass data of the distributed computing platform log analysis system; this system was used for finding Web visits statistics. [3] This paper described in detail the various functional modules of the system presented in distributed platforms that warrant the efficiency of experimental analysis. Experiments show that the proposed analysis system, works through multiple resources to complete the original work being undertaken by a node; whether it is the implementation of data processing or regular tasks, its efficiency is higher than stand-alone centralized environment-based Web log analysis as shown in Fig. (**1**).

1874-4443/15

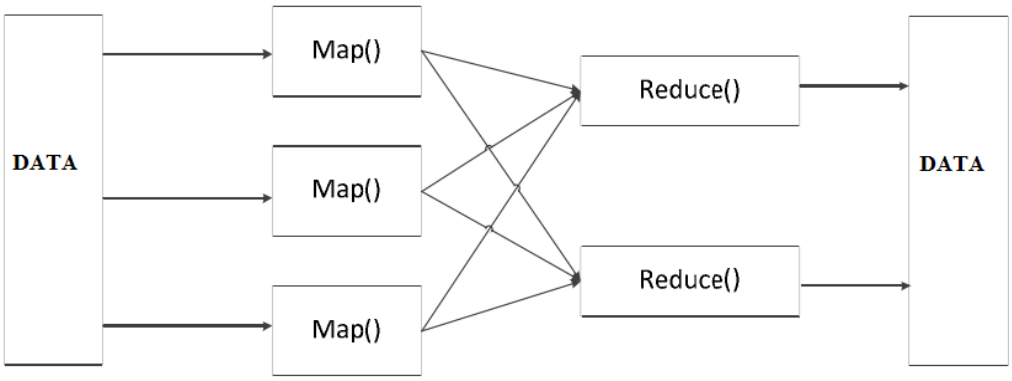

Fig. (1). Map reduce flows.

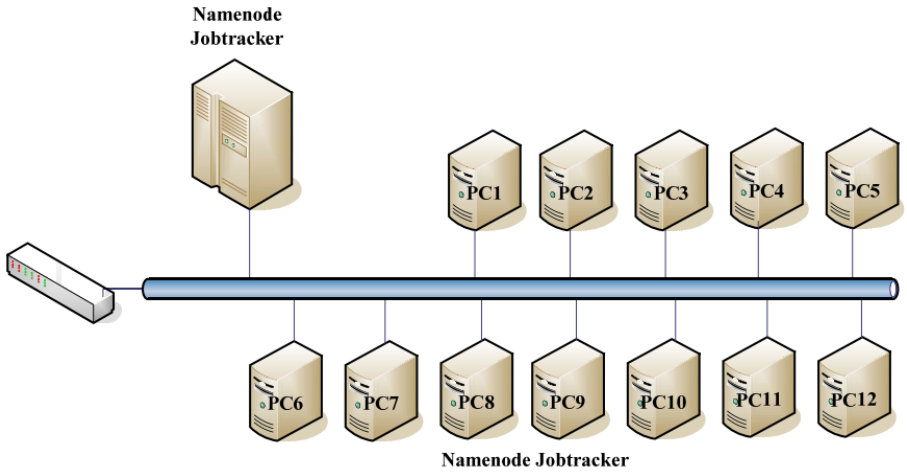

Fig. (2). Platform archtecture.

Recent social networking websites such as Twitter, Facebook, LinkedIn, YouTube, and Wikipedia have not only connected large user populations but have also captured Exabyte's of information associated with their daily interactions. Since beginning, social networking has been a subject of interest for social scientists in the context of human social networks; for mathematicians and physicists, in the context of complex network theory; and, most recently, for computer scientists, in the examination of information or Internet-enabled social networks. We can thus separate major research challenges into these areas as shown in Fig. (**2**) [4].

The first step of Data integration is to effectively extract information from data sources, taking efficiency and data transmission quantity into consideration. For the purpose, the best method involves timely detection of incremental data. The classification of information sources mainly contains databases and flat files. Based on the features of information sources and the needs at business level, one can choose most suitable one from a variety of incremental acquisition methods.

Understanding social networks evolves into a big data problem when business, management, or information systems specialists hope to predict behavior to ultimately enhance marketing, sales, and online commerce. Many social networking sites have 10 to 200 million users, so data sampling is central to most studies. Although it is significantly time consuming, gaining insight from the entire dataset might provide the most optimal solutions. Big data is usually characterized by the "three Vs" - that is, volume, velocity, and variety. In terms of volume, at the end of 2011, Facebook had 721 million individuals and 68.7 billion friendship edges (see http://arxiv.org/abs/1111.4503) [5], (Fig. **3**). In terms of velocity, Twitter and Facebook respectively generate 7 Tbytes and 10 Bytes of data daily. These data also need to be processed at the speed of thought. For example, on 11 November 2012, a sales event at TaoBao, the largest online shopping marketplace in China, generated 100 million transactions and reached a peak transaction rate of 205,000 per minute [6]. In terms of variety, data today come from various sources, ranging from surveillance videos, to satellite images, to mobile tweets, to sensors and meters in the power grid.

Comparison between the original data and store data is shown in Table **1**.

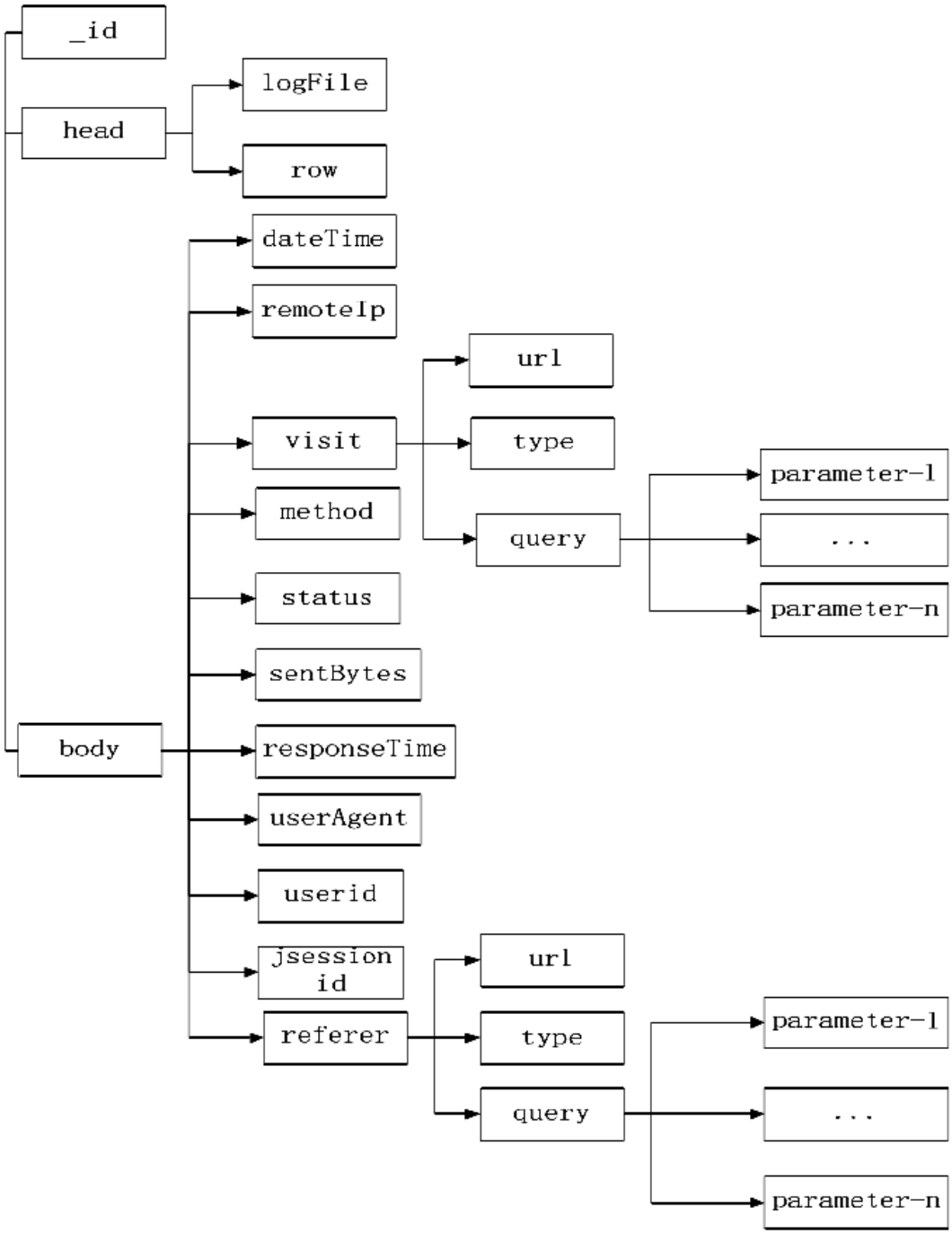

Fig. (3). Document storage structure diagram.

This paper focuses on the log analysis method, which is dependent on the database system with a log component. I have implemented a log analysis method on the oracle database system. Also, this study summarizes the theoretical basis and uses notes of log miner tool from existing data, which are based on my own practice. First, two major problems are described: the effectiveness of the log extraction and net effect of incremental data, especially for this piece of log extraction is a key to the method. On the basis of two different analytical modes, and three steps into log analysis: *i.e.* the initialization operation, normal operation, and interruption restart, this paper illustrates the likely problems and how to resolve them. The main purpose is to reliably and effectively extract specific IDU operation in accordance with the specific object. The second problem describes the theoretical models of net effect processed according to the transaction record's interchangeability and simplified rules, and on this basis, deals with different situations and resolves difficult problems. Secondly, this study makes a brief introduction to multi-source analysis, composite example, transaction concepts, and describes a number of existing difficulties, for the control of composite transactions, multi-source operations, a large transaction processing, *etc*. (Fig. **4**).

Table 1. The original data and store data.

Original Data	Store Data
27/Sep/2013:01:09:09 +0800	201309270109098000
219.142.53.236	219.142.53.236
GET /wap/order.do?op=confirm&shopCartIDs=940&buyNow Flag=1&buyGoodNum=1 HTTP/1.1	GET
	/wap/order.do
	op=confirm shopCartIDs=940 owFlag=1 buyGoodNum=1 buyN
	HTTP/1.1
500	500
510	510
1.431	1.431
Mozilla/5.0 (compatible; MSIE 10.0; Windows NT 6.1 Trident/6.0)	Mozilla/5.0 (compatible; MSIE 10.0; Windows NT 6.1;Trident/6.0)
C69166DB C16E17F20B3671FB7C24E1F6	C69166DBC16E17F20B3671FB7C24E1F6
C69166DB C16E17F20B3671FB7C24E1F6	C69166DBC16E17F20B3671FB7C24E1F6

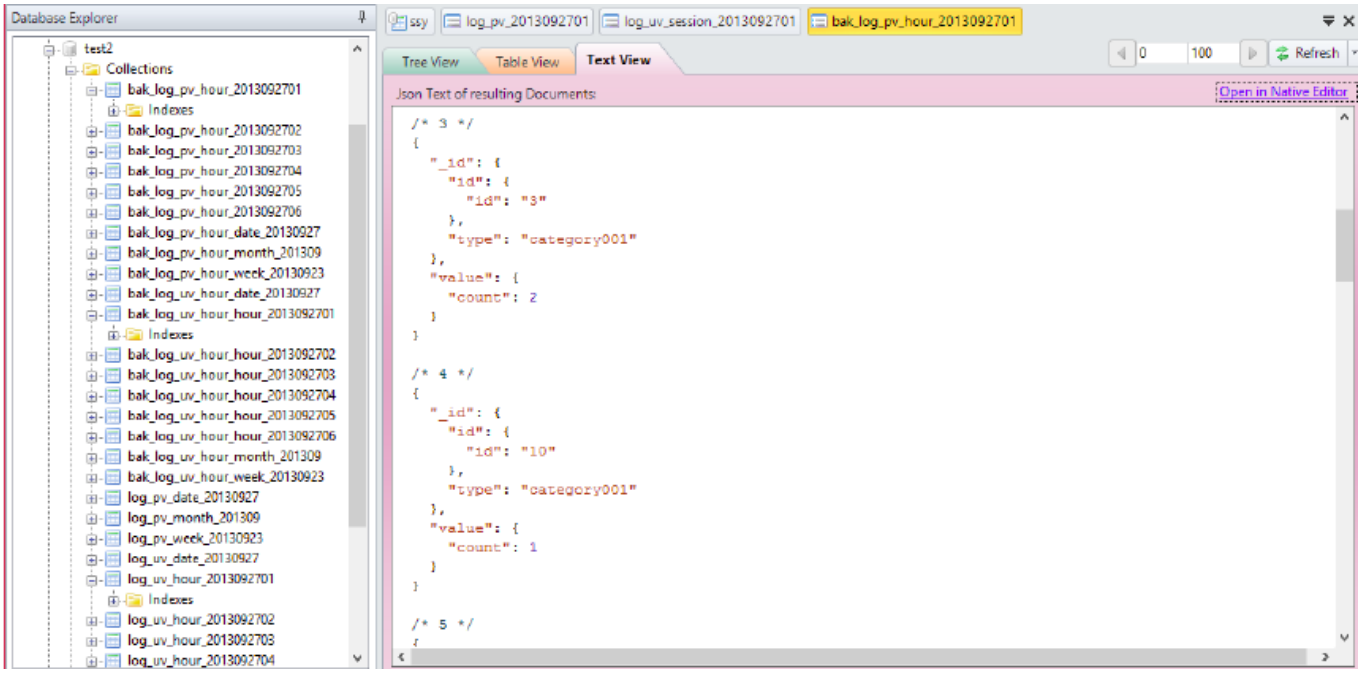

Fig. (4). The results of the statistics of document.

Storage format is:

{ [27/Sep/2013 :01:10:04+0800]}{124.65.196.2} }GET/ wap/order.do?op=confirm&shopCartIDs=940&buyNowFlag =1&buyGoodNum=1WindowsNT6.1;Trident/6.0)}{C69166 DBC16E17F20B3671FB7C24E1F6}{C69166DBC16E17F2 0B3671FB7C24E1F6}HTTP/1 .1}}500}}510}{1.431}{- }}Mozilla/5.0 (compatible; MSIE 10.0.

```
reduce=function(key, emits)
{
total=0;
for(var i in emits)
{
```

```
total+=emits [i] . count;
return }"count": total};
db.runCommand(
 {mapreduce:
 map:，reduce:[, query : ][, sort : ] [, limit : ][，out : ][, keeptemp: ][, finalize : ][，scope:
 [, verbose:true] } );
db.collection.mapReduce(<map>,<reduce>,
{<out>,<query>,<sort>,<limit>,<keytemp>,<finalize>,<scope>,
sMode>,<verbose>})
{
 "map": "function Map(){emit({'urfahis.body.referer, 'jsessionid'ahis.body.jsessionid,'mobile'ahis.bod
y. mobile,'type':'referer'} ,
 {'total': 1,'size': this.body.sentBytes});}”，
 "reduce": "function Reduce(key,values){ var size=O;var total=O;for(var i in values){size+=values
[i].size;total+=values[ i].total;}return{'total': total,'size': size};}”，
 "query": {`body.visit.type':{'$nin': ['do','htmf] },`
body.referer.type': {'$in':['do','htmf] }},
 "out":”$!{pvSize.pvSizeCollName}”
```

Grid software logs all events and activities for the storage grid, grid administrators can use analysis of log information so as to equip the health grid in order to make more optimal decisions. During the analysis and the use of the grid log, however, it faced with a wide range of problems such as a huge amount of hard to read data, which caused great difficulties for grid administrators and ordinary users. Therefore, we urgently needed a log analysis system for grid software. The system should be able to extract useful information from the grid which could do log analysis and mining to helps users make better use of the grid to meet the production needs.

Based on the guiding ideology of Software Engineering, from the system needs analysis, several aspects of the design of the system architecture, system design, system coding and system testing, this dissertation elaborated on the whole process of the research and development of grid software log analysis system, its final design and completed a grid software log analysis, based on log analysis and decision support systems. The business process shows that the system first obtained grid log files from the grid nodes for pretreatment, followed by log analysis and mining output analysis results, final reports and charts. In order to reduce the reporting engine load and improve the efficiency of report generation, the pretreatment process will log the initial merge log after merging datasets on the basis of the analysis and the log of the association rule mining. The analysis mainly includes: a summary of statistical information grid analysis, refinement multidimensional statistical analysis, statistical analysis of resource access, access to user statistics analysis, intrusion detection analysis, and log analysis to describe the operation of the grid system from the full multi-angle range.

Look at Fig. (**5**), each network is positioned based on its relative level of generality and its ability to execute. (Some online services included in this figure, such as Amazon EC2, Globus Online, Galaxy, and caGrid, are arguably social networks by themselves. However, we list them here because they all provide an open collaborative environment that's very close to a social network and can rapidly evolve toward that direction.)

3. DATA MODEL AND HIGH-LEVEL ABSTRACTION

Relational models and SQL provide an abstraction layer between the database's physical layer and the application layer. This feature lets users specify a query in a language dependent and declarative manner, while a query engine schedules and optimizes its execution. No similar solution exists for big data analysis. Instead, NoSQL data stores offer various forms of data structures - such as document, graph, row-column, and key-value pair — that are directly exposed to users. So, users must understand data's physical organization and employ vendor-specific APIs to manipulate these data. Current state of the art attempts to devise a SQL layer on top of NoSQL, but without an abstract data model, this effort is ad hoc and limited to the underlying technology.

3.1. Incremental Processing and Approximate Result

In some big data applications, such as financial fraud detection and market promotion, long delays aren't tolerable. A newly emerging paradigm called stream computing enables continuous queries over streaming data such as social media feeds and call data records. Stream computing opens a gateway to real-time analytics, but a few challenges remain. One is the interplay between building the batch mode model and sensing the real-time streams. On one hand, the accumulated historical data in the data warehouse can help information specialists build a statistical model to guide stream processing - for example, deciding which features to observe and helping set the reacting threshold. On the other hand, the new data from the stream system should be leveraged to tune the model to reflect the recent trends. An incremental data processing and model tuning mechanism is vital to this interplay

3.2. Implementation

System implementations include: in order to achieve the system easy to upgrade and easy to expand the target system uses J2EE Spring Struts framework; for the log format conversion and processing efficiency, system uses log Description XML format; for log analysis in order to be more efficient in order to improve optimization, and implementation of the association rules analysis, the system uses the improved association rules Apriori algorithm; for improving operating efficiency and performance of the system in order to achieve efficient use of memory space, the system introduces design patterns; and to optimize system interactive performance, system uses a visual display for log analysis results using reports and charts. Finally, a comprehensive test is performed based on the function and performance of

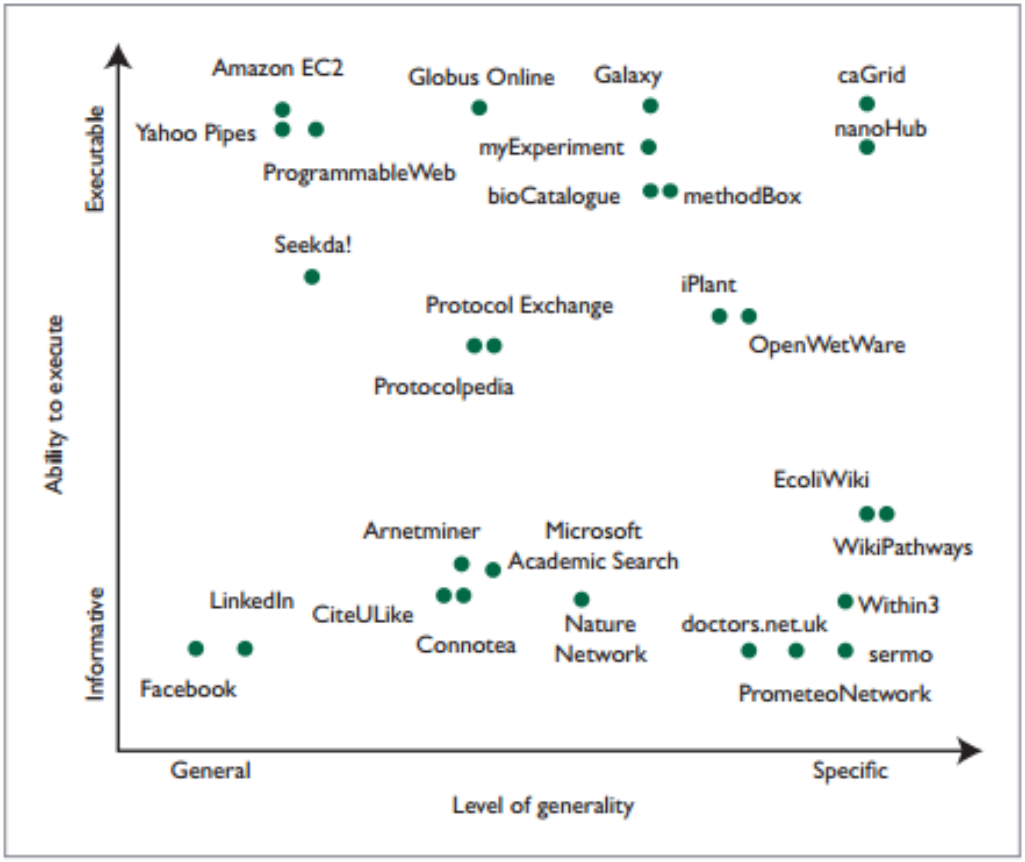

Fig. (5). Social networks for scientist.

the system, which can be seen from the test results. The grid software log analysis system is designed and completed on the basis of features and performance meeting the expected design goals, the software can be used as grid software technical support auxiliary analysis tool for promoting the gird use.

4. CONNECTED DATA: NEW CHALLENGES FOR CLOUDS AND SOCIAL NETWORKS

The research has shown that users primarily employ social networking sites to articulate and make visible their existing social networks. In other words, users on these sites aren't usually trying to connect with strangers but are primarily communicating with people who are already part of their direct or extended social network. This observation implies that a level of trust already exists between social network users, and that these users share at least one aspect of their lives: career, hobbies, political views, and so on. We envision that these characteristics are vital to enabling interesting opportunities, including establishing security policies that leverage existing trust relationships, promoting data and resource sharing within networks of people with similar interests, and optimizing data analytics by leveraging the fact that people in the same network potentially share the same interests and will thus submit similar queries. Finally, we propose leveraging the wisdom of socially connected individuals to build and maintain service reputation systems. Clouds comprising social network connections open numerous research opportunities.

As shown in Fig. (**6**), based on their social link, new users can be automatically classified into group as they join the network.

4.1. An Example

Baidu FengChao is a newly promoted advertisement auction system, exploiting the daily billions of web searches, which brings huge income for both business customers and Baidu. Till 2010, the income from FengChao occupies more than 20% of Baidu's total income. However, according to the online running and customer feedbacks, FengChao still faces many problems in advertisement quality measuring, presence and optimization. These problems will cause economic loss for customers and brings bad effects for FengChao. To address these problems, this paper designed and implemented a massive advertising log analysis system based on Hadoop, aiming to mine abnormal data from massive advertisement log, and further provide visual statistics on the abnormal data from different views to help FengChao find potential problems, after a thorough analysis of the reasons for the abnormal data, finally propose effective solutions.

With the rapid progress of computing techniques and communication techniques, the information service has been widely used in regular life. It persistently provides trusted service for 7x24 hours a week. The infrastructure of such service is usually based on distributed systems that are constructed with a large number of computing resources, handle large amount of user requests and store large amount of user

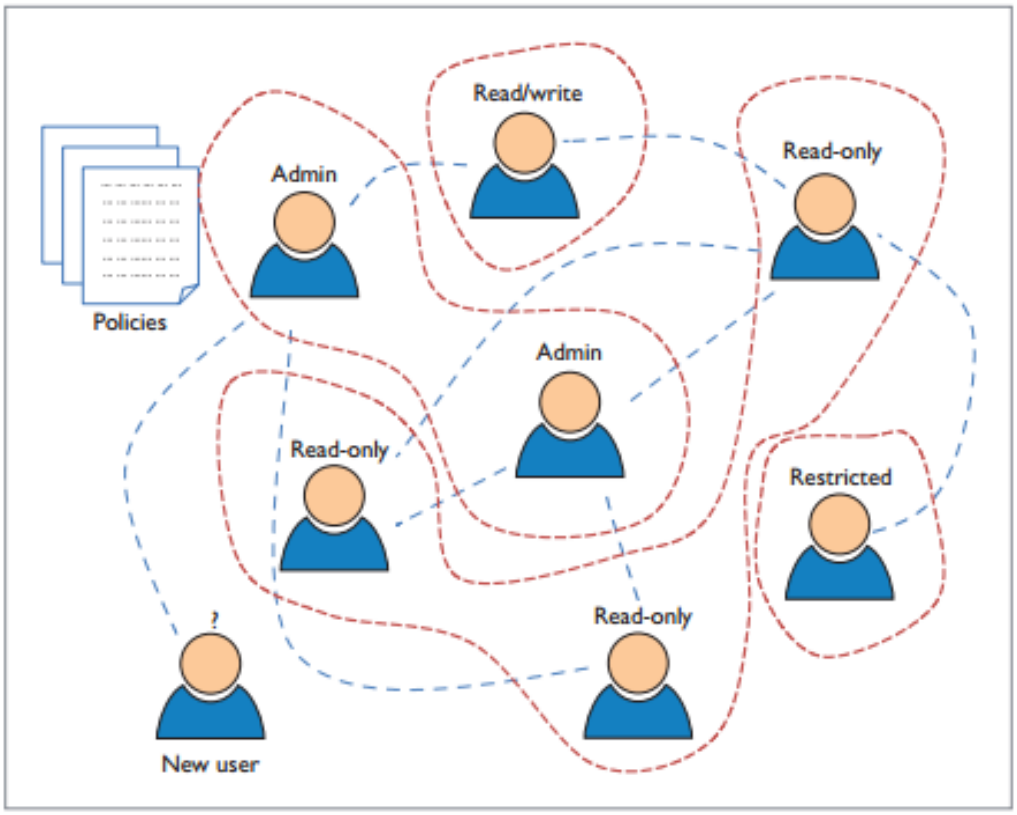

Fig. (6). Overlaying the social graph with security group, roles, and policies.

data. In order to build a large-scale trusted distributed system, it requires a dramatic increase in the complexity of system behavior and system running environment, for the following two reasons: software bugs are difficult to get rid of, and it is difficult to anticipate all the situations that could encounter at runtime. A direct result of such complexity is that the system should be persistently providing trusted service while the system composing component persistently fails, which seriously affects the 7x24 high trusted services required.

Monitoring-Action is a widely used runtime trust grantee mechanism. It uses the fault localization techniques to accurately diagnose the root cause of system failure, and then takes the corresponding actions to grantee the runtime trust property of the system. Event log has been widely used in fault localization techniques as an effective abstraction of system behavior. There are mainly two ways to localize a fault: using fault model to identify a known fault; using normal behavior model to identify the abnormal behavior. However, due to the complexity system behavior and system environment, there still exist great challenges to localize fault by event logs: there exist a large mass of noisy logs through the fault modeling process, which could lead to false positives and false negatives in the fault localization process; it is difficult to extract the normal behavior model under distributed environment, and the low detection efficiency makes it difficult to detect runtime abnormal system behavior.

First, our paper determines the requirement of a log analysis system based on Fengchao's business functionalities, and then designs the functional structure of this log analysis system, which can be divided into three modules: log parsing module, log analysis and mining module, and web presentation module. Log parsing module completes the preprocessing of the original log data. Log analysis and mining module is the key part of this system. It builds computation model for different business monitoring and mine abnormal data pertaining to a business, then does a multi-view statistics of the abnormal data. The log analysis and mining module mainly consists of three business themes: advertisement quality, advertisement census and advertisement optimization. The web presentation module provides statistics result on a web page with dynamic trend graph and tables.

In the implementation, log parsing and log mining modules fully utilized the advantages of Hadoop in processing the big data. The massive original log data and analysis result are both stored in HDFS(Hadoop Distributed File System), establishing a different set of MapReduce computing program to realize the data processing based on Hadoop MapReduce algorithm. The web module adopts LAMP (Linux+Apache+MySQL+PHP) and a popular web application open source framework CakePHP. Finally, the log analysis system's function and performance were tested and verified for commercial effect, and it has been established that the log analysis system can help Fenchao solve potential problems, by effectively reducing the Fengchao's online error rate, and providing effective basis for decision making.

CONCLUSION

This study designs and implements a system of social network data acquisition and structure analysis; the main work is as follows:

This dissertation describes some concepts and techniques used in the process of designing and realizing a new system.

The system has been designed and implemented with the capability to implement Baidu Fengchao users network data acquisition and structure analysis, and that the system can be used to get: the real relationship data between users from Baidu Fengchao; the data de-noising processing, and the graph of relationship network structure. Finally, the complex network analysis method is used to analyze characteristics of the network topology. Moreover, the system is capable to implement co-author network data acquisition and structure analysis, and the system can be used to: get to the co-author network data from the papers recorded by the DBLP database up to four levels of conferences, the theme of which is "data mining"; process data to generate network structure graph; detect the top 100 structure holes and opinion leaders.

CONFLICT OF INTEREST

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

ACKNOWLEDGEMENTS

This work is supported by the Annual Subsidy Scheme Young Backbone Teachers of Colleges and Universities in Henan province Foundation of China under Grant No.183 and The Henan Province Science and Technology Projects Foundation of China under Grant No. KJT142102210226.

REFERENCES

[1] O. Kai, "*Analysis and Classification of NoSQL Databases and Evaluation of their Ability to Replace an Object-Relational Persistence Layer*," Technische University, China, 2010.

[2] C. Fay, D. Jeffery, G. Sanjay, H. Wilson, W. Deborah, B. Michael, C. Tushar, F. Andrew, and G. Robert, "BigTable: a distributed storage system for structured data," *ACM Transactions on Computing System*, vol. 26, no. 2, pp. 1-26, 2008.

[3] K. Chodorow, "*Scaling MongoDB. Sebastopol*", CA: O'Reilly Media, 2011, pp. 1-3.

[4] R. Lammel, *Data Programmability Team, Google's Map Reduce Programming Model-Revisited*, Microsoft Corp: Redmond, WA, USA, 2007.

[5] F.J. Benjamin, W. Stefan, and U. Brian, "Multi-university research teams: shifting impact, geography, and stratification science," *Science*, vol. 322, no. 21, pp. 1259-1262, 2008.

[6] Y. Liao, M. Moshtaghi, B. Han, and M.B. Mining, "Opportunitiesand challenges," In*: Proceedings of Social Networks: Computational Aspects and Mining*, Springer, vol. 3, 2011, pp. 1-28.

Received: June 02, 2015 Revised: August 04, 2015 Accepted: September 10, 2015