Send Orders for Reprints to reprints@benthamscience.ae

Research of the Big Data Platform and the Traditional Data Acquisition and Transmission Based on Sqoop Technology

Changai Chen* and Shan Jiang

School of Information Technology, Henan University of Traditional Chinese Medicine, Zhengzhou, Henan 450000, China

Abstract: In this paper, 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.

Keywords: Sqoop, big data platform, data acquisition and transmission, transitional data.

1. INTRODUCTION

With the development of Sqoop 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 Face book 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 Sqoop technique [1].

The support of Sqoop 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 Sqoop [2].

Data acquisition and transmission system is closely related to people's lives and in most cases it is used in production environments where real-time data monitoring is in demand. It helps relieve people from trivial work, increases productivity and reduces cost. A common data acquisition system only supports one transmission method, which has greatly limited its range of application, so a data acquisition and transmission platform based on C8051F MCU is proposed in this thesis.

2. BASIC KNOWLEDGES AND RELATED TECHNOLOGY

Based on the application conditions of the existing traditional data warehouse and the future forecast of the Sqoop big data platform, this paper proposes the new framework of the cooperation of Sqoop and traditional data warehouse which focus on the cooperation between the traditional data warehouse and the Sqoop 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 Cloud era and Terawatt, 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, [3] and researches the Sqoop 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 Sqoop 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 (Fig. **1**). Finally, the prototype system proves the feasibility of the designed structure [4].

A general-purpose data acquisition and transmission platform is introduced in this thesis. The main platform hardware includes C8051 MCU and CP2200 Ethernet controller, and the platform software provides data acquisition, wireless data transmission, cable data transmission and some other functionalities. The platform makes a feature of providing two methods to transmit data. In locations where cabling is difficult, it provides remote data transmission through 3G wireless network by connecting it to the mobile communication module; [5] while in locations where wireless signal is weak, the platform uses the Ethernet controller to access the Internet and transmit data. Tests show that the platform works well, and it not only meets the requirements of data acquisition and transmission but also is general purpose and stable. (Fig. **2**)

Data acquisition and transmission technology are usually applied to many situations, such as scientific research and industry. Due to the speed and difficulty of installment, the further development and application of the data acquisition

1874-4443/15

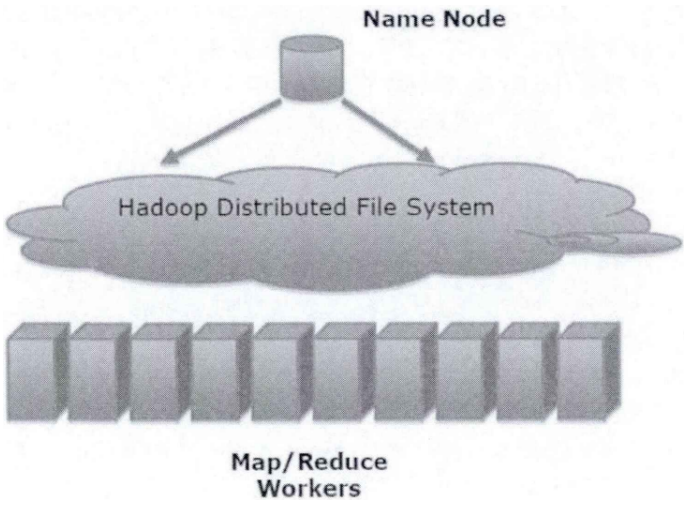

Fig. (1). HDFS and map/reduce.

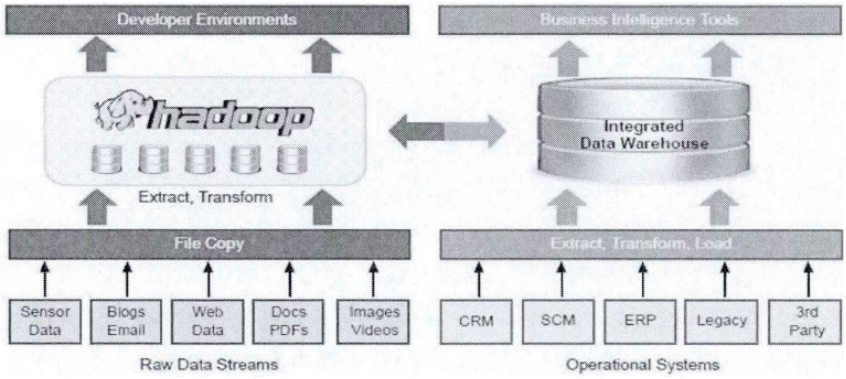

Fig. (2). Sqoop and traditional data.

Fig. (3). A Modular data center was built at Nanyang Technological University (NTU) for System/Tested Research. The tested hosts 270 servers organized into 10 racks.

instruments were confined to the traditional communication ways. As a good solution to the problems, the USB (Universal Serial Bus) technology is widely used because of the attributes of hot-plug-in, plus-and-play, easy to expand, engrossing less system recourse. The Sqoop has been paid much attention depending on the theory transfer rate of 480Mbps (Fig. **3**).

The data acquisition and transmission system was developed with Sqoop technology in the hardware this paper, based on deeply analyzing of Sqoop protocol, which built up and technology. As for software platform for data acquisition and transmission with Sqoop application, [6] the system was use in multimember dynamic ECG recorder and contaminative insulators on-line detecting system, and the bulk and interrupt transmission were carried out in these system respectively (Fig. **4**).

① The hardware circuit of system was designed, which based on ISP1581 Sqoop interface chip of Philips corporation. The bus work mode and data accessing mode were mainly considered in the process of design.

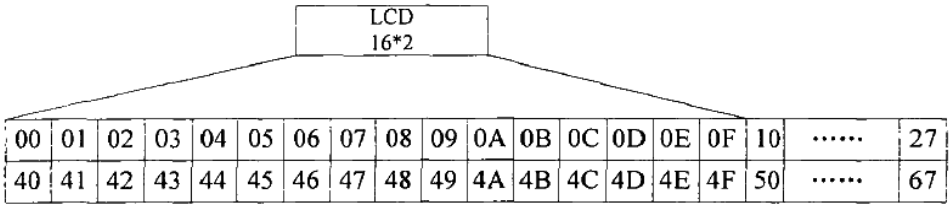

Fig. (4). Screen position and RAM address mapping diagram.

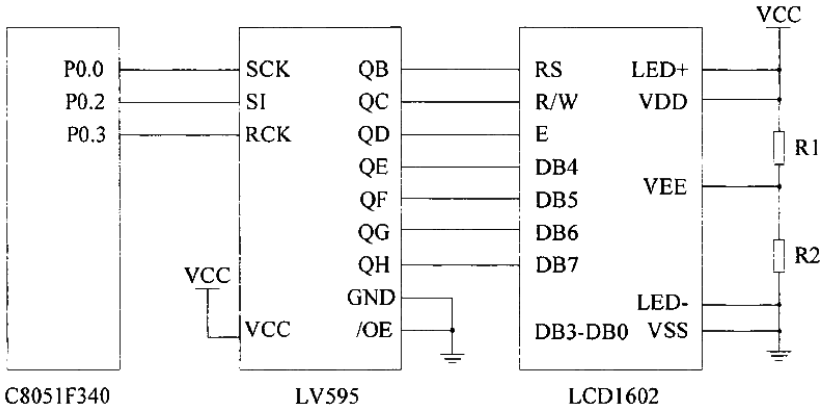

Fig. (5). Circuit of LCD module transmission in this system.

② The software platform of system was built up, which made up of firmware, device driver and application program. The firmware was complied with language C, whose primary functions were USB bus enumeration, A/D conversion and information communication between device and host computer. The device driver was developed with the saddlebag driver works of Numeral Company. The application program was developed to control device, read-write data, display waveform and read-back record [7].

③ The multimember dynamic ECG recorder based on Sqoop technology was designed, which mainly completed acquisition of ECG analog signals, switch of power supply mode, data management and data transmission with USB bulk transfer mode.

④ The contaminative insulators on-line detecting system based on Sqoop and SMS technology was designed, which was a practical application of Sqoop data acquisition and transmission system with its interrupt transfer type. The develop of LP-1 insulators on-line detector and the establishing of data transfer channel were the most important part in the whole system. The format of detecting data was set down for SMS (Fig. **5**).

The paper combine Sqoop stud with its practical application, in order to solve interface problem between device and host computer, and complete data acquisition and information communication. Based on the need of subject, Sqoop technology is applied to practical data acquisition system, which made convenient connect between acquisition device and host computer, and provides Sqoop advantage to acquisition device. The successful application of Sqoop technology provides a good foundation for further study and application of Sqoop (Fig. **6**).

3. ANALYSIS

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.

3.1. Key Codes

```
publir boolcan sqoop_import sql_ hdfs(){
  tryl
    l-n}pertTeel impTool=new ImportTool();
    }qeeptieps options=new SqoopOption};
    c}ptic}n}.setDriv-
;:r('lass\;}r}L("can}.wi}rcS,iUt.setlsrrw,s      :IcIhc.St})L.S};
werDriv}
    options.setHadoopMapRedHome(Iladoop:llapRedHome);
    options.setConnectString(lserwerC: oramclSlring);
    options.setUsername(SyhserrerUsernanJe);
     options.setPassword(lserverPa.s.smarcl);
    options. setTargetD ir( I/dfs7 argelDir);
    options. setTableName( lldfs加TablelVanre );
```

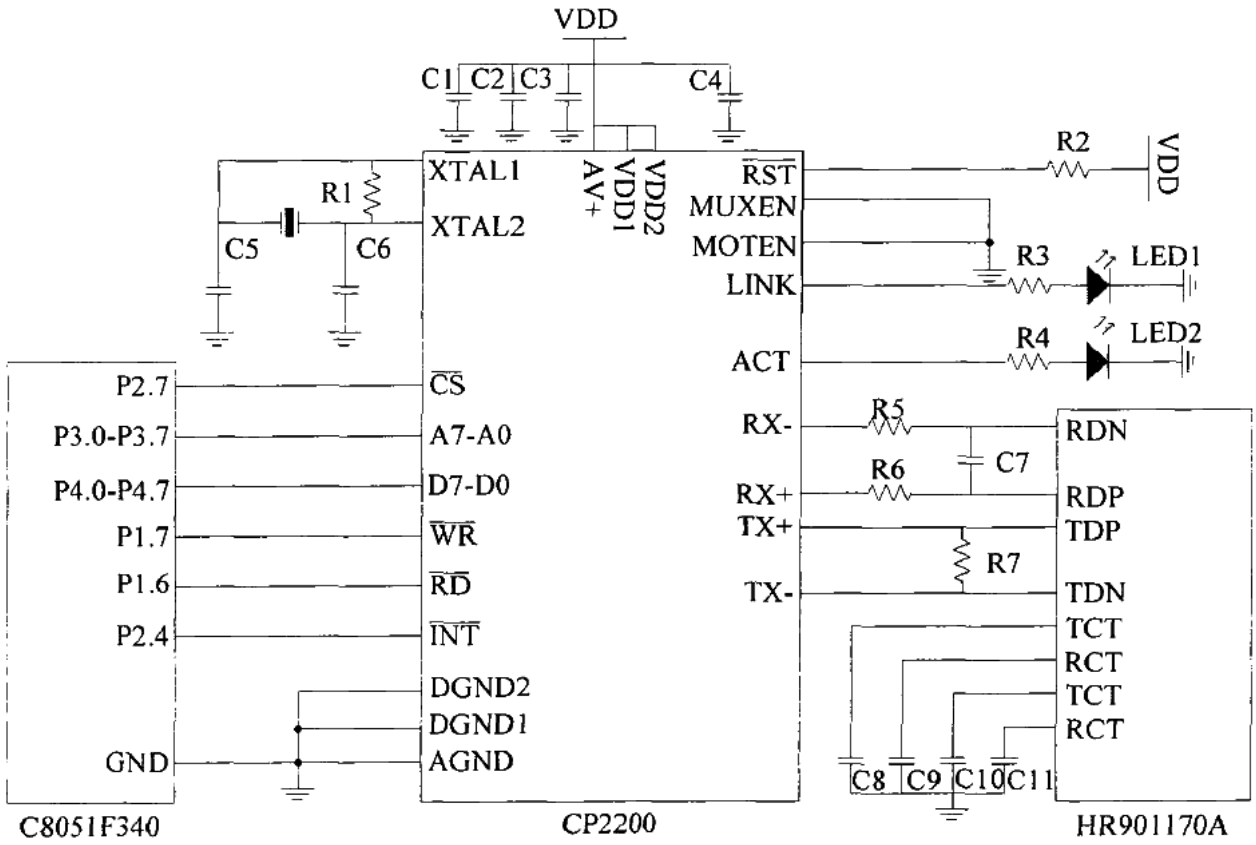

Fig. (6). Circuit of cable transmission module.

Table 1. Comparison between big data and traditional data.

	Traditional Data	Big Data
Volume	GB	constantly updated (TB or PB currently)
Generated Rate	per hour, day,…	more rapid
Structure	structured	semi structured or unstructured
Data Source	centralized	fully distributed
Data Integration	easy	difficult
Data Store	RDBMS	HDFS,NoSQL
Access	interactive	batch or near real-time

```
  options.setSpIitByCol(fldfs/mSplilByC ol);
  options.setlnputFieIdsTerminatedBy('\001');
   options.setNumMappers(1);
  impTool.run(options);
 }catch (Exception e){
   return false;
}
return true;
}
```

3.2. Design and Implementation

Due to the research of BDaaS is in the conceptual discussion stage, it still faces four challenges: 1)There is no standardized, user experience based BDaaS architecture which can shield the complexity of data sources and operations; 2)The lack of generic unstructured data model which reflects user behavior characteristic, made BDaaS for unstructured

data diff cult to build; 3)Existing data model follows the Web services model, however, so far, holistic BDaaS service model with the characteristics of big data has not yet appeared; 4)There is no appropriate solution in providing data retrieval, analysis and visualization services and optimizing service capacity.

In order to solve the above problems, four key technologies of BDaaS architecture, data model, BDaaS service model, as well as BDaaS applications will be in-depth study. Firstly, this paper designed a User Experience-oriented BDaaS Architecture, so as to provide a high level of standardization guidance for building a platform. Secondly, in terms of the data model, in order to unify description unstructured data, the user behavior-based unstructured data model has been designed. Thirdly, in terms of the service model, algebraic model has been established by using process algebra, and then extended OWL-S ontology-based BDaaS model and the service composition approach were designed. Finally, service processes of retrieval, analysis and visualization have been described in detail, and the two measures of improving the retrieval services accuracy and service efficiency have been used to optimize the BDaaS capacity.

4. PLATFORM TEST

As existing unstructured data models were difficult to meet the demand for BDaaS, the Galaxy Data Model (GDM) has been proposed, which is a user behavior based unstructured data model. By monitoring the behavior of data generator people, a generic model with fully attributes like user behavior, semantic background have been created, which is the basis for the realization of the BDaaS for unstructured data. The case study shows GDM not only has good versatility and comprehensiveness, but also has a lightweight, easy-to-use description language and operating language. In addition to the traditional file system, GDM also supports modeling and retrieval of unstructured data in HDFS. In addition, GDM has application in the National Pre-pregnancy Check Information Management System (NPCIMS) to verify its feasibility and practicality.

Due to the holistic BDaaS service model with the characteristics of big data has not yet appeared, Extended OWL-S based Big Data-as-a-Service model (EO-BDaaS) has been proposed. By add properties of the data sources, data types, service operation in the OWL-S in order to build many types of BDaaS such as search, analysis, visualization, and to compose service dynamically. Case study shows, compared with the existing data services, EO-BDaaS with a more comprehensive description on attributes and operations. Besides, it has capabilities such as strong semantic comprehension and automatic service composition, and integrated the unique combination operations of BDaaS into the implementation of data services seamlessly. To solve the problem of low accuracy of retrieval services, this paper presents the heat sensitive unstructured data retrieval ranking algorithm HotRank. First heat score was calculated, which is the match degree between the tasks attributes of search results and task attributes of services consumers, after that assigned the scores to each of the search results, and then sorted search results based on heat score. By using such means to make search results more in line with the preference of the user. The simulation results show that, the Precision-Recall of HotRank is better than Windows Search ranking algorithm. Therefore as the improving of retrieve accuracy, HotRank is able to optimize not only the user experience, but also the service capacity.

A data heat recognition-based Hybrid Prefect Algorithm (HPA) has been proposed to meets the quickly respond requirements of the BDaaS. First, by analyzing the log of user data operation and develop data heat determine rules, then according the dynamic and static prefect rules to get candidate data, finally prefect data would be take into the cache. The simulation results show that average hit rate of HPA is 55%, the average accuracy rate of HPA is 43%, which indicates that the algorithm not only has good ability to predict user operation of data, but also to optimize the BDaaS capacity. In addition, HPA-based Hybrid Prefect based Persistent Caching architecture has been applied in the National Prepregnancy Check Management Information System (NPCMIS) in order to verify its effectiveness.

As shown in Fig. (**7**), it pivots on two axes, i.e., data value chain and timeline. The data value chain divides the data lifecycle into four stages, including data generation, data acquisition, data storage, and data analytics. In each stage, we highlight exemplary technologies over the past 10 years. Data acquisitionrefers to the process of obtaining information and is subdivided into data collection, data transmission, and data pre-processing. First, because data may come from a diverse set of sources, websites that host formatted text, images and/or videos - data collection refers to dedicated data collection technology that acquires raw data from a specic data production environment. Second, after collecting raw data, we need a high-speed transmission mechanism to transmit the data into the proper storage sustaining system for various types of analytical applications. Finally, collected datasets might contain many meaningless data, which unnecessarily increases the amount of storage space and affects the consequent data analysis. For instance, redundancy is common in most datasets collected from sensors deployed to monitor the environment, and we can use data compression technology to address this issue. Thus, we must perform data pre-processing operations for efcient storage and mining.

5. WIRELESS TRANSMISSION

Wireless sensor networks bridge the gap between physical world information and digital world, enhance the data collection ability of computer, and open up a broad application prospect for the fields of military, medicine, environment omni-Coring, intelligent home system and so on, and cause wide concern. The sensors in the sensor networks are connected wirelessly with ad hoc mode. Efficient communecanon supports are needed to build network. The researches on traditional ad hoc networks focus on better service quality and higher throughput. The key issue of wireless sensor

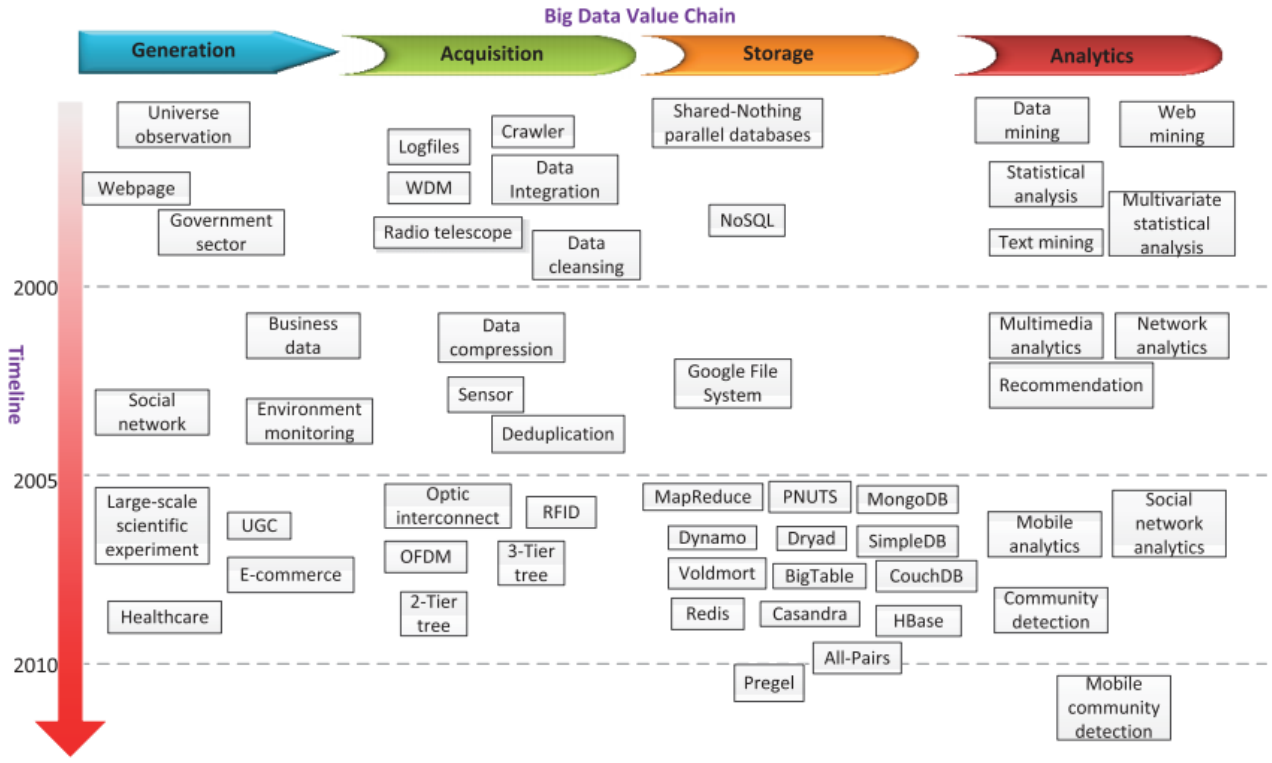

Fig. (7). Big datatechnology map.

networks is energy efficiency. So, efficient communication strategies with low energy consumption are required. We design transmission scheme for MAC layer, cross-layer between MAC and route layer, and special scenario in structural health monitoring.

We have designed an asynchronous duty cycle based MAC layer broadcast protocol for wireless sensor networks. The sender repeats the broadcast packets with dynamic interval. The repeated transmissions would make the period sleeping neighbors receive the packet when they wake up. The dynamic interval would reduce the packets collision caused by hidden nodes problem. We implement our scheme in MICAz motes to validate the applicability. We also have performed further expert- mints with NS-2 simulation environment. The results show that our protocol could deal with packet collision well. The extra power consumption caused by collision is reduced and the data reception ratio is enhanced.

We have proposed a cross layer designed communication protocol for data collection in wireless sensor networks. In the MAC layer, we inherit the low power listening approach to save the power consumption. Instead of single target adopted in traditional MAC protocol, the sender could select mufti-nodes to be the candidate targets and transmit the data to the node that wakes up first. In the route layer, we evaluate the latency from each node to sink and select the nodes with low latency to be the next hop targets. The sleeping based energy efficient protocol trades high latency for low energy consumption. Our work could achieve lower energy consumption with much less latency introduced. The experiments shows the energy consumption and latency are lower than state of art sleeping based energy efficient protocols. Our protocol is much efficient in large scale network.

We have introduced a novel scheme of using the elevators to assist data collection for wireless sensor networks used in structural health monitoring application. A base station is attached to an elevator. A representative node on each floor collects and transmits the data to the base station using short range communication when the elevator stops at or passes by this floor. We formulate the problem as an optimization problem where the data traffic should be transmitted on time and the lifetime of the sensors should be maximized. We show that if we know the movement pattern of the elevator in advance, this problem can be solved optimally. We then study the online version of the problem and show that no online algorithm has a constant competitive ratio against the offline algorithm. We show that knowledge of the future elevator movement will intrinsically improve the data collection per-romance. We discuss how the information could be collected and develop online algorithms based on different level of knowledge of the elevator movement patterns.

A comprehensive set of simulations and MICAz tested experiments have demon-started that our algorithm substantially outperforms conventional multi-hop routing and naive waiting for elevator scheme.

CONCLUSION

Firstly, the development background of the platform and relevant technologies are introduced in this thesis. Then, the overall platform architecture is given in system design, and the system hardware and software are divided according to function into modules respectively. The hardware implementation and software implementation are given next. In hardware implementation, the schematic diagram of each hardware module is presented, and the driver for some specific hardware module is also given. In software implementation, detailed implementation and key technologies are described, including AT command programming and the simplification to the TCP/IP protocol stack. Then, the platform test process is conducted. Finally, the conclusion of this thesis and the outlook are given for further study.

CONFLICT OF INTEREST

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

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] E. Dumbill, *Apache Hadoop What You Need to Know About this Important Big Data Tool*. [http://www.forbes.com/sites/oreillymedia/2012/02/07/apache-hadoop-what-you-need-to-know-about-this-important-big-data-tool/]

[2] J . Ekanayake, H. Li, B. Zhang, T. Gunarathne, S. Bae, J. Qiu, and G. Fox, "Twister: A runtime for iterative mapreduce," In: *Proc. 19th Assoc. Comput. Mach. (ACM) Int. Symp. High Perform. Distrib. Comput.*, 2010, pp. 810-818.

[3] D. Logothetis, and K. Yocum, "Ad-hoc data processing in the cloud," In: *Proc. VLDB Endowment*, vol. 1, no. 2, pp. 1472-1475, 2008.

[4] T. Condie, N. Conway, P. Alvaro, J. M. Hellerstein, K. Elmeleegy, and R. Sears, "Mapreduce online," In: *Proc. 7th USENIX Conf. Netw. Syst. Des. Implement.*, 2010, p. 21.

[5] B. Li, E. Mazur, Y. Diao, A. McGregor, and P. Shenoy, "A platform for scalable one-pass analytics using mapreduce," In: *Proc. Assoc. Comput. Mach. (ACM) SIGMOD Int. Conf. Manag. Data*, 2011, pp. 985-996.

[6] D. Jiang, B. C. Ooi, L. Shi, and S. Wu, "The performance of mapreduce: An in-depth study," *Proc. VLDB Endowment*, vol. 3, no. 1-2, pp. 472-483, 2010.

[7] A. Ghazal, T. Rabl, M. Hu, F. Raab, M. Poess, A. Crolotte, and H. Jacobsen, "Bigbench: Towards an industry standard benchmark for big data analytics," In: *Proc. Assoc. Comput. Mach. (ACM) SIGMOD Int. Conf. Manag. Data*, 2013, pp. 1197-1208.

Received: September 16, 2014 Revised: December 23, 2014 Accepted: December 31, 2014