International Journal of Advanced Smart Convergence Vol.5 No2 59-65 (2016)
http://dx.doi.org/10.7236/IJASC.2016.5.2.59

IJASC 16-1-2

A Design of DBaaS-Based Collaboration System for Big Data Processing

Yean-Woo Jung*, Jong-Yong Lee**, Kye-Dong Jung**

**Department of Information System, KwangWoon University Graduate School of Information Contents, 20 Kwangwoon-ro, Nowon-gu, Seoul 01897, Korea*
***Ingenium college of liberal arts, KwangWoon University, 20 Kwangwoon–ro, Nowon-gu, Seoul, 01897, Korea*
e-mail: {blueshark2011, jyonglee, gdchung}@kw.ac.kr

Abstract

With the recent growth in cloud computing, big data processing and collaboration between businesses are emerging as new paradigms in the IT industry. In an environment where a large amount of data is generated in real time, such as SNS, big data processing techniques are useful in extracting the valid data. MapReduce is a good example of such a programming model used in big data extraction. With the growing collaboration between companies, problems of duplication and heterogeneity among data due to the integration of old and new information storage systems have arisen. These problems arise because of the differences in existing databases across the various companies. However, these problems can be negated by implementing the MapReduce technique. This paper proposes a collaboration system based on Database as a Service, or DBaaS, to solve problems in data integration for collaboration between companies. The proposed system can reduce the overhead in data integration, while being applied to structured and unstructured data.

Keywords: *Cloud, Database, DBaaS, MapReduce, Collaboration*

1. Introduction

As Big Data is emerging as a new paradigm in the IT market, the technology for processing big data on the cloud computing environment are focused. Big data as a means of unstructured and large, while at issue was Internet users is increased to increase the emergence and spread of mobile devices Web 2.0[1]. Big data is closely related to cloud computing. Big data has been emerged by increase of cloud and internet users. In addition, because the emergence of cloud computing can able to easily build a system for big data.

Cloud computing should continue to provide the service to users. But when all the data in a single repository to store and process, there is the risk of a service interruption when a problem occurs in the repository. Therefore, cloud computing uses a distributed file system. As big data processing technologies development, acceptance of data continues to increase problem was occurring in the distributed file storage technologies for cloud computing, and cope with a variety of fault, optimizes the output performance and prevent errors in the data and storage of data and data problems such as duplication was also possible solving [2] [3].

Big data occur in collaboration between global company. Data of generated in entire world is large capacity, generate in real time and has variability.

And there is various types of data exist, because each company in the same field are manage data in different ways. integration of these data has two problems.

Manuscript Received: April 20, 2016 / Revised: May 5, 2016 / Accepted: May 20, 2016
+ Corresponding Author: gdchung@kw.ac.kr

Tel:+82-2-940-5288

Ingenium college of liberal arts, KwangWoon University, Korea

Each company has built a different database. The heterogeneity of data generated from differences in schemas and data storage formats. And it can be a number of duplicate data is generated in the processing of data. duplicate data has the disadvantage that the decreasing data integrity and increasing the overhead of the search.

In this paper, to solve this problem, define the data generated in the between companies collaboration as big data and propose a DBaaS based collaboration system for processing big data. In the proposed system solves the data heterogeneity between companies and redundancy issues through DBaaS to provide data integration, management and mapping capabilities of meta data, and big data processing technology MapReduce. And collaborative systems by applying a document-oriented database technology, it is possible to improve the storage capacity optimization and data exchange rate. Document-oriented database enables the data exchange between different heterogeneous databases because the information, such as relationships or attributes for each data to store the data required[4].

This paper is organized as follows. Section 2 describes DBaaS, Document-Oriented database and MapReduce. Section 3 describes DBaaS based Collaboration System. Section 4 describes application example. Section 5 describes Comparison with other system. Section 6 describes conclusion and future research.

2. Related Works

In this section describes DBaaS to provide integration of data in the proposed system, MapReduce that provides redundant data removed And document-oriented database used for the exchange and processing of data.

2.1 DBaaS

DBaaS (Database as a Service) is a cloud computing services and resources, including operational monitoring from the database to the user installation[5]. The service range of existing Cloud is widened been raised the need to integrate multiple databases. And when providing the database to a cloud, the type of database, the decision on the methodology to provide the schema, the resource, and the predicted response to the load that occurs in a variety of situations is required. Also it need to solve the heterogeneity of data generated at each integrated database.

DBaaS has the advantage that these problems can be solved. The integration takes place in DBaaS has the following types: It is integrated in the integration and database level in the schema level. There are many difficulties in the implementation of the integration scheme level, but there is advantage exists that can be efficiently if it implementation. The proposed system can solve data heterogeneity using scheme level integration proposed by DBaaS Hub System.

DBaaS can approach to each data provider using non-From query and SQL, because it using existing database. In addition, the data exchange among the integrated database uses document-oriented database[4].
In the proposed system, request data to provider has each database using global query and standard data mapping through MDR(Metadata repository) for data request to DBaaS, and solve the heterogeneity of collected data.

2.2 Document Oriented Database

Document-Oriented Database is non-relationship database based on document format such as Json and XML. In Document-Oriented Database, it has no complicated system changes when change stored data, because it include schema in document file[6]. However, it is difficult to express relationship between data, Because data reference does not almost exist.

Document-oriented database Access allows the rapid processing of the data by decreasing the reference count, and it is possible to ensure the scalability of a storage capacity because it is file-based. Can also

manage data more small capacity as compared to the relational database, and is suitable for exchanging data, because significantly independent of the structure[6][7][8][9].

2.3 MapReduce

MapReduce for big data processing has a feature that processes using the multiple computers to split a large amount of data. It consists of a Map and Reduce process. In Map converts the data to a key / Value form and input to the data node, In Reduce removes redundancy of the data assigned to the node. MapReduce is widely used Big data analysis tools, distributed file system and etc[10][11][12][13].

The proposed system removes the redundancy of the data to be gathered from the database and generates the integrated data throughthe MapReduce. Because in this paper based on the resource database, all data is composed of a field name / value format. Therefore collaboration system serves to map a field name corresponding to the key of the data in a standard field name to solve the heterogeneity of data further problem consisting in the integrated database resource.

3. Collaboration System

In this section we describe the DBaaS based collaboration system. The proposed system is to process big data generated through the data integration between companies, and provides services that provide integrated data. Data exists in each companies database has different metadata each other. Therefore, the system proposed map the companys data to standard metadata, and then integrate data. Proposed system use Json to Data exchange with each data provider, processing collected data and integrated data storage. Collaboration system provides a function to provide the integrated database resource to the user. Figure 1 shows the architecture for a collaboration system.

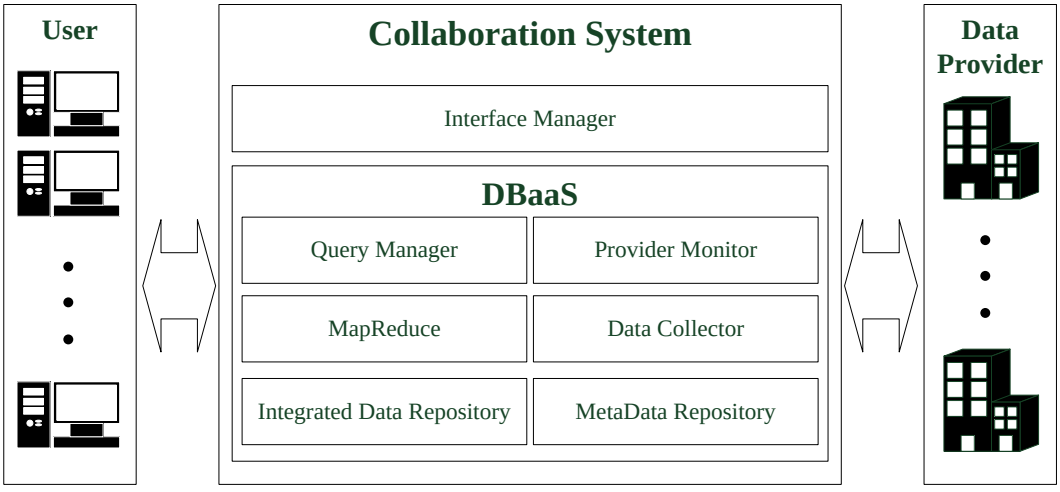

Figure 1. System architecture.

A description of each module of the collaboration system is as follows.

- Interface Manager: Interface Manager provides an interface for the convenience of users accessing the collaborative system.
- DBaaS: DBaaS shall serve to monitor, collect and integrate the resource database of the data provider. DBaaS consists Query Manager, Provider Monitor, MapReduce, Data Collector, Integrated Data Repository and MetaData Repository.
- Query Manager: Query Manager creates a Global Query to request the necessary data to the data provider.

- Provider Monitor: Provider monitor shall serve to monitor the process for the collection and integration of data, monitor and status of the data provider.
- MapReduce: MapReduce, it created an integrated data from JSON-formatted data that is collected from the Data Collector.
- Data Collector: Data Collector serves to collect and integrate the data from the data provider and stored in the Integrated Data Repository.
- Integrted Data Repository (IDR): IDR store collected from the data provider and integrated data. The IDR provides an integrated database resources to the users.
- MetaData Repository (MDR): MDR is Repository that stores Local Metadata, Standard Metadata, and Mapping information between local an standard.

Figure 4 shows the process of data integration and provision of collaborative systems through sequence diagram.

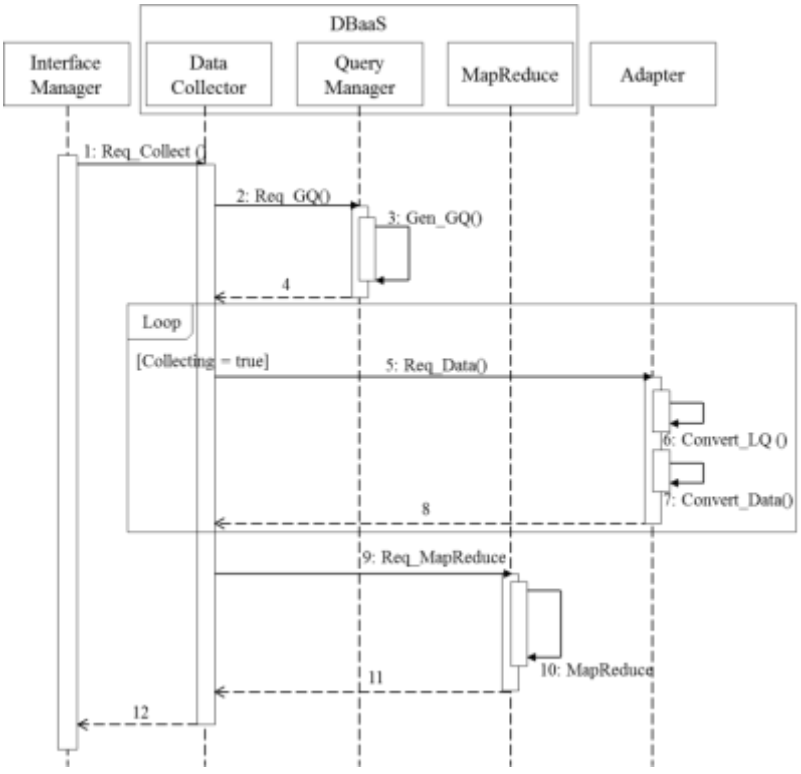

Figure 2. Sequence of data collection.

There is Adapter in data provider. Adapter is convert global query to local query from collaboration system, extract data from database and convert data do Json type document-oriented database file. Operating procedure of the proposed system is follows:

First Data Request(1). Second generate Global Query(2~4). Third send Global Query to each provider's Adapter(5). fourth convert global query to local query and extract data from provider's database(6). fifth convert data to Json type document-oriented database file(7). Sixth send data to collaboration system(8). seventh integrate data through MapReduce(9~11). finally provide integrated data to user.

4. Application Example

In this section to evaluate the performance of proposed system, the Collaboration system, was applied to a

travel system. Figure 3 and Figure 4 shows a interface for application examples.

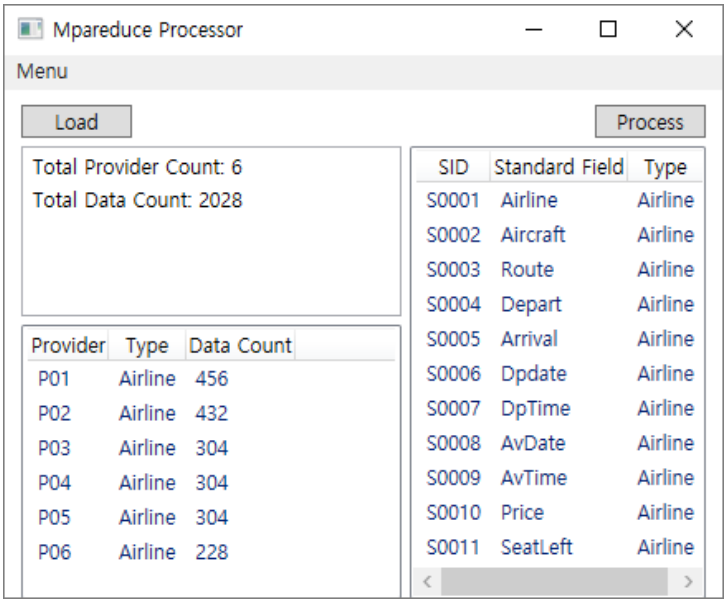

Figure 3. Before data processing.

Data was collected from a total of six provider, the type of the collected data is a data for Airline. In Figure 3, the interface output total provider information, provider-specific information and standard feild name information.

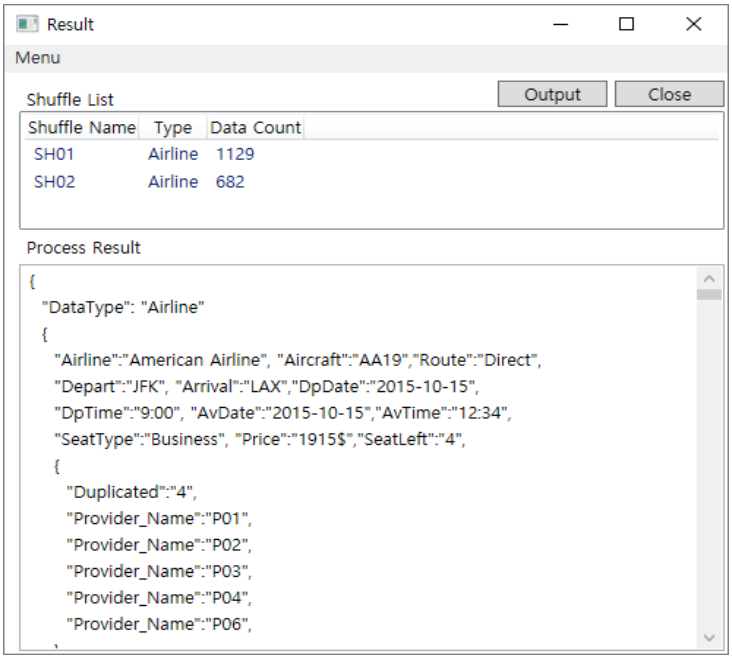

Figure 4. Result of data processing.

In Figure 4 Shuffle List shows the data count after the data integration through MapReduce. Processed data is output to Json format as shown in Figure. The number of the integrated data was reduced as compared

to the total number of data before processing, because redundant data is removed.

5. System comparison

In this section, a comparison of the existing system and the proposed system. Table 1 shows a comparison of six items of this system and MapReduce in Hadoop Distributed-FileSystem(HDFS)[13] and The Realtime Complex Event Detection System(RCEDS)[12].

Table 1. System Comparison.

	Hadoop Distributed-File System	The Real time Complex Event Detection System	Collaboration System
Data Processing	Processing the data divided in a number of nodes	Processing by the complex event	Structured data through multi-threaded processing
Data Type	Analyzing unstructured text data by SNS	Analyzing unstructured text data by SNS	Structured data
Mapping Type	Key-Value Map	Data Type Mapping	Metadata mapping by MDR
Data Collection	Data generated in real time	Filtering through keyword set	Filtering the data through the requirements
Required Information	Information that are not required Decrease in efficiency with unnecessary data cleansing	Information that are not required Increased efficiency in filtering events	The requested information Increased efficiency in filtering data
Data Heterogeneity	Standard data definitions difficulties	Defined by a complex event processing	Defined by the metadata

First, in the data processing item, HDFS is to distribute data to multiple nodes and processes the data and, RCEDS is a complex event to create multiple complex event, proposed system processes the formal data stored in the RDB through the multi-threaded. Second, in the data type item, two other system analyzes text data in the SNS, but proposed system analyzes the structured data. Third, in the mapping type item, HDFS mapping between a Key, RCEDS maps to the type of data, proposed system performs a mapping based on the metadata based on the MDR. Fourth, data collection HDFS is targeted to data generated in real time on the network, RCEDS is reduced the amount of data to the data generated in real time through the keyword set filter, proposed system reduces the amount of data to be subjected to the analysis by the data filtering through the requirements of the query. Fifth, in the Required Information item, HDFS collects unnecessary data, and thus uses the MapReduce for improved efficiency. RCEDS improve efficiency through the filter of the event information that is not required. proposed system is effective for the processing efficiency because it performs the analysis with the required information collected by the query. Sixth, the heterogeneity of collected data, HDFS is hard to define the standard data in the process of analyzing by key-value pair for the text data generated in real time. RCEDS and is solved by information about the heterogeneous data through the definition for complex event processing, proposed system solves problem through define analysis and mapping of metadata.

6. Conclusion

In this paper, we proposed a collaboration system for solving the data heterogeneity, redundancy and compatibility between the database generated from Big Data processing. Solution of the problem is presented as follows. The data heterogeneity resolution solved by metadata mapping using MDR in DBaaS. and using document-oriented database of Json format to solve compatibility, guaranteed throughput and scale of the efficiency of the data. In addition, data redundancy problem is solved through the MapReduce.

Feature of the system is as follows. Data providers because it maintains an existing RDB, using the global

data query in the data acquisition requests. Definition of the standard data is possible through the simple and efficient processing, filtering, because using structured data. Because solves heterogeneity and redundancy can reduce the overhead in the integrated data search.

Future research not only structured data, will be the research and evaluation of methods for solving the heterogeneity of unstructured data.

References

[1] Agrawal, Divyakant, Sudipto Das, and Amr El Abbadi. "Big data and cloud computing: new wine or just new bottles?." Proceedings of the VLDB Endowment 3.1-2 (2010). pp 1647-1648. 2010

[2] Changqing Ji, Yu Li, Wenming Qiu, Uchechukwu Awada and Keqiu Li. "Big data processing in cloud computing environments." Pervasive Systems, Algorithms and Networks (ISPAN), 2012 12th International Symposium on. IEEE. pp17-23. 2012.

[3] Yean-Woo Jung, Seongsoo Cho, Jong-Yong Lee and KyeDong Jeong. "A Design of P2P Cloud System Using The Super P2P". The International Journal of Internet, Broadcasting and Communication(IJIBC) VoL.7 No.1. pp42-48. 2015

[4] Kye-Dong Jung, Chi-Gon Hwang, Jong-Yong Lee, Hyo-Young Shin. "The Study of DBaaS Hub System for Integration of Database In the Cloud Environment ". The Society of Digital Policy and Management. Journal of Digital Convergence Vol.12 No. 9. pp201-207. 2014.

[5] Yean-Woo Jung, Jong-Yong Lee, Seong Ro Lee and Kye-Dong Jung. "Design of the P2P cloud system based MapReduce". The 2nd International Conference on Contents(ICCPND). pp36-37. 2015

[6] Shimura, Takeyuki, Masatoshi Yoshikawa, and Shunsuke Uemura. "Storage and retrieval of XML documents using object-relational databases." Database and Expert Systems Applications Volume 1677 of the series Lecture Notes in Computer Science pp 206-217. 1999.

[7] Padhy, Rabi Prasad, Manas Ranjan Patra, and Suresh Chandra Satapathy. "RDBMS to NoSQL: Reviewing some next-generation non-relational databases." International Journal of Advanced Engineering Science and Technologies Vol.11 No.1. pp15-30. 2011.

[8] Wei-ping, Zhu, Li Ming-Xin, and Chen Huan. "Using MongoDB to implement textbook management system instead of MySQL." Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on. IEEE, pp300-305. 2011.

[9] Nurzhan Nurseitov, Michael Paulson, Randall Reynolds, Clemente Izurieta. "Comparison of JSON and XML Data Interchange Formats: A Case Study." Caine 9. pp157-162. 2009.

[10] Lammel, R. "Google's MapReduce programming model-Revisited". Science of computer programming Vol.70 No.1. pp1-30. 2008.

[11] Hung-chih Yang, Ali Dasdan, Ruey-Lung Hsiao and D. Stott Parker. "Map-reduce-merge: simplified relational data processing on large clusters".Proceedings of the 2007 ACM SIGMOD international conference on Management of data. pp1029-1040. 2007.

[12] Junhui Lee. "The Realtime Complex Event Detection System in Bigdata using Heterogeneous Data Integration". Inha University Graduate School. Master's Thesis. 2013.

[13] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler. "The hadoop distributed file system." Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on. IEEE, pp1-10. 2010.