International Journal of Advanced Smart Convergence Vol.5 No.1 16-22 (2016)
http://dx.doi.org/10.7236/IJASC.2016.5.1.16

IJASC 16-1-3

Map Reduce-based P2P DBaaS Hub system

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 139-701, Korea*
***Department of Computer Science, Kwangwoon University, 20 Kwangwoon–ro, Nowon-gu, Seoul, 139-701, Korea*
****, +Department of Electronic Engineering, Kwangwoon University, 20 Kwangwoon-ro, Nowon-gu, Seoul 139-701, korea*
E-mail: blueshark2011@kw.ac.kr, gdchung@kw.ac.kr, jyonglee@kw.ac.kr, gdchung@kw.ac.kr

Abstract

The database integration is being emphasized to one way of the companies collaboration. To database integration, companies are use like one database what their own, it can be provided more efficient service to customer. However, there exist some difficulty to database integration. that is the database security and database heterogeneity problems. In this paper, we proposed the MapReduce based p2p DBaaS hub system to solve database heterogeneity problem. The proposed system provides an environment for companies in the P2P cloud to integrate a database of each other. The proposed system uses DBaaS Hub for a collection of data in the P2P cloud, and use MapReduce for integrating the collected data.

Keywords: *Cloud, P2P-Cloud, P2P, DBaaS, Map-Reduce*

1. Introduction

With the development of the IT industry, companies have seeking a various method for providing a convenient service to clients. One of them is collaboration between the companies. Recently, one way of collaboration between companies, and it is focused on the database integration. In this paper, we propose a MapReduce-based P2P DBaaS hub system that provides a database integration environment. The DBaaS is providing a database in the cloud with on-demand format. It is the cloud computing paradigm that enable to use the database from multiple companies as a single database.[1, 2] However, the provision of DBaaS has a several problem. that is the database security and database heterogeneity problems between companies.[1] In this paper, we focused on database heterogeneity problems, and we proposed P2P DBaaS Hub System for solve this problem. The Database heterogeneity is appeared each database's schema structure is different, or

when users using other type of database(MSSql, MySQL and Oracle etc...) each other.[1] because of the

Manuscript Received: Nov. 13, 2015/ Revised: Nov. 26, 2015/ Accepted: Jan 15, 2016
+ Corresponding Author: gdchung@kw.ac.kr
Tel:+82-2-940-5288
Department of Electronic Engineering, Kwangwoon University, Korea

database heterogeneity, there is difficult to normal method to exchange data each other. Proposed system solve this database heterogeneity to using DBaaS hub, and collect data from database inside cloud using DBaaS hub. In the proposed system, database integration is consist of the two sequences, data collection and data integration. Data collection is as follow.

1. Data request.
2. Data extract and convert
3. Data translate and collect

Collected data from database inside cloud is gathering to data requester. For output the data collected as you want, using the MapReduce to integrate the data.[3, 4, 5] Through the data collection and data integration, the user can use the database in the cloud as if as a single database. The paper is organized as follows. In the section 2, describes the related works of this paper. In the section 3, describes the proposed system and architectures. In the section 4, describes the application example of proposed system In the section 5, System comparison of proposed system and P2P-MapReduce: Parallel data processing in dynamic Cloud environment's P2P-MapReduce. Finally in the section 6, describes conclusion and future research.

2. Related Works

In this section describes the related works of this paper. P2P Cloud for the cloud environment used in the proposed system, DBaaS Hub System for proposed method of the data collection and integration, and MapReduce for data integration.

2.1 P2P Cloud

Cloud computing is a technique to a user connected to a network, for providing resources on demand type, As the popularity of cloud computing paradigm, the cloud of have much type of service for provide more friendly service to users, Various types of cloud that appeared to provide more efficient service to the users. P2P cloud is one of those cloud computing.[6, 7, 8] Generally, cloud services are centralized cloud gathered one of computer constituting the cloud is configured as a cluster. However P2P cloud are computer constituting the cloud and not to form a cluster, each computer that is connected to the same structure as the P2P network.[6] The P2P cloud is possible to provide all the services of an existing centralized cloud,. and p2p cloud is more stable, because it has multiple access point that not like centralized cloud.[6] It is mainly used in many fields of online games or media streaming, Unlike centralized need to configure the cluster to build, only resources that already exist, there are advantages to build this possible.[6, 8] In this paper, we use these P2P cloud. By using the P2P cloud, users, it is possible to be provided a more stable service.[6]

2.2 DBaaS Hub System

DBaaS Hub system is one of the important technologies to construct DBaaS. Database used in many different companies has different schema structure or it has been used different types of database. DBaaS Hub system takes care heterogeneity problem by standardizing the data in these databases based on ontology. In DBaaS Hub system when requesting data, and use the global query. As a global query is a query statement with no clauses From, The field name to request a data non-local field name for the database, and makes use of standard field names from the standard metadata.[1]

2.3 MapReduce

MapReduce is software framework developed for process a large amount of data or unstructured data in real

time. MapReduce mainly process the data more than petabytes or unstructured data, difficult to process the traditional relational database. Data mining is the main use areas, such as Web crawling. Hadoop is a frame work is a typical example.[3, 4, 5] MapReduce is to process large amounts of data, divided data distributed parallel processing multiple computers. Because process data come distributed to multiple computers, a single operation is made by a simple operation.[3, 4, 5] MapReduce is characterized separated by Map and Reduce functions, and organize the data into the Map and the Reduce is eliminate duplication of data.
In this paper, we proposed MapReduce based P2P DBaaS Hub system that using these MapReduce's feature. the proposed system integrate data collected from multiple computer to using MapReduce. And it has several different to P2P-MapReduce: Parallel data processing in dynamic Cloud environments P2P-MapReduce[4], it will describes in section 5.

3. MapReduce-based P2P DBaaS Hub system

In this section, describe about MapReduce-based p2P DBaaS Hub System. Proposed system is system that collect data using DBaaS Hub to using cloud member's database like a single database, and output collected data through the MapReduce. Data search from database in the cloud is composed as follows.
1. Input condition.
2. Data collection.
3. Data integration and output
cloud is p2p cloud based on p2p network, Figure 1 is overview of the proposed system.

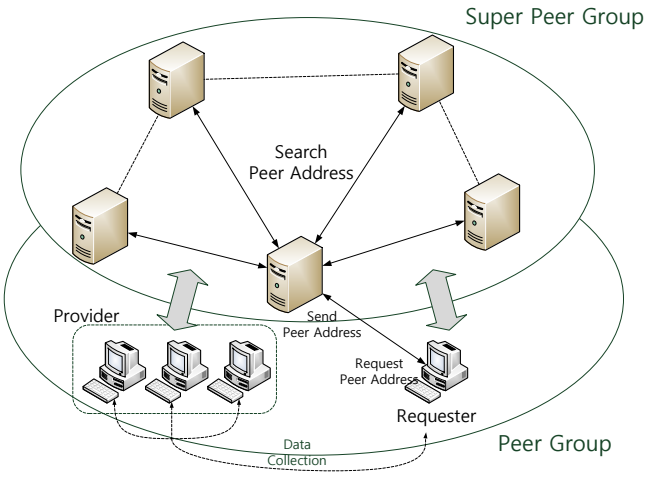

Figure 1. MapReduce based P2P DBaaS Hub System overview

Cloud network of the proposed system, as shown in Figure 1 consists of a two types. That is the Super peer and peer. Peer has a database with DBaaS Hub that receives or provides data to each other, are a kind of computer user layer. However Super-Peer does not even have a DBaaS Hub, nor does it provide the data. In the P2P system, super peer is not connecting peer to peer directly, connect to broker like one of the hybrid p2p method. Super Peer is responsible for processing a search query on behalf of the Peer connected to them.[6, 7] As in Figure 1, if the requester requesting the address of the provider for data collection, Super peer searches the address of the provider from the Super peer in the Super peer group returned to the requestor.[9, 10] if using super peer, can reduce network traffic, search delay time, and it is possible to

prevent an infinite loop when Peer searching.[10] Figure 2. shows the architecture of Super Peer and Peer in the proposed system.

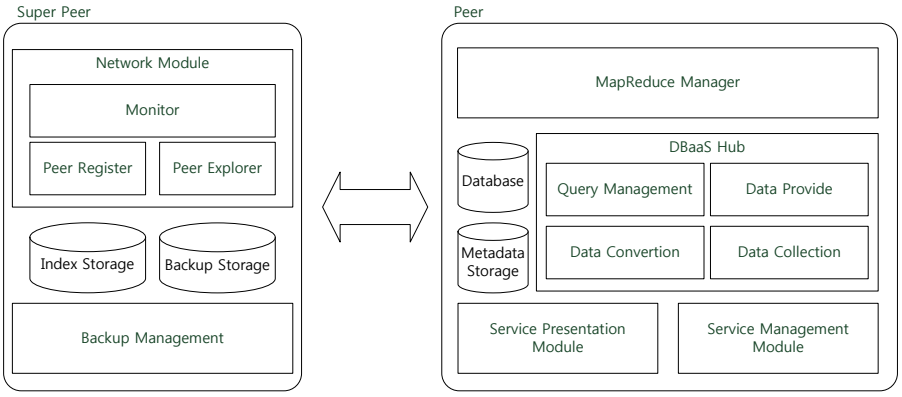

Figure 2. System architecture

Super peer under the management of their Peer, and is composed of modules for performing a function to respond to a search query request from a Peer. Also, in contrast to other Super Peer death has the ability to back up data with one another.

- **Network Module**: The network module performs a function for managing the peer, or in response to a search query. it consists Monitor, Peer Register, and Peer Explorer.
- **Monitor**: Monitor is periodically check the connection status of the peers, and serves to update the index information.
- **Peer Register**: Executes the registration process of the new peer has been connected to the network. Finally, generating index information of the peer and stored in the index storage.
- **Peer Explorer**: The response to the search query received from other peers or Super Peer. Search is through the index information exists in the Index Storage.
- **Index Storage**: Index store stores the index information indicating the data type of peer, peer addresses, database type of peer.
- **Backup Storage**: In contrast to the death of another Super peer, store the backup data of another Super peer.
- **Backup Management**: Periodically backing up the data in case that the Super Peer or other Super peer is be dead, if the other Super peer dies. Super peer performs the function on behalf of that serves.

Peer: Peer has a function of DBaaS Hub for request data or provide data, and MapReduce for integrate data and output. All Peer has a DBaaS Hub and the database by default.

- **MapReduce Manager**: function to integrate collected data from the data provider peer and output integrated data.
- **Metadata Storage**: Ontology-based, there is a IS-A, HAS-A relationship table about the local metadata and standard metadata. When requesting the data, converting data, standard metadata is necessary.
- **DBaaS Hub**: It performs the function of the management of Metadata and collection of data. it consists of Query Management, Data Conversion, Data Provide. And Data Collection.
- **Query Management**: Generate global query to using user inputted condition, IS-A, HAS-A Table in the metadata storage. Or convert the global query into local queries.
- **Data Conversion**: Mapping the data extracted from the database through the IS-A, HAS-A

relationship in metadata store to be converted to a standard data.

- **Data Provide**: Output the data to a standard document-oriented database format file and sends it to the requester.
- **Data Collection**: It stores a document-oriented database format file transmitted from the provider to the temporary storage. File stored in the temporary storage are the MapReduce Manager is used for data integration.
- **Service management Module**: The service management module to manage the process for the service used by the user. That is from the condition input to the output of the data.
- **Service Presentation Module**: A service management module is responsible for periodically updating the standard metadata stored in the user's metadata store.

The proposed system user enable to using the database in the cloud as a single database. For this DBaaS Hub functions to collects data from the Peer in the cloud. Figure 3. is to describe the process of data collection in the proposed system. If the user inputs a search conditions in the interface, the requester for data collection peer requests an provider peer's address search in the Super peer. Requestor peer initiates a connection with the provider peer, and the data collection begin.

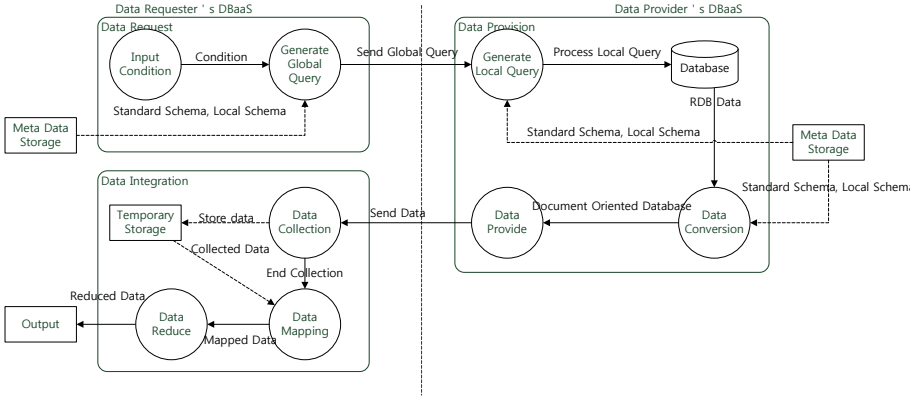

Figure 3. Sequence of data collection

1. The data of the request: the search conditions entered by the user, using the IS-A, HAS-A relationship table stored in the metadata storage to generate a global query. The generated global query is sent to the connected provider.
2. The extraction and conversion of data: provider convert global query to local query as IS-A, HAS-A Relationship mapping . extract data to input converted local query in database, extracted data is convert standard format data to using IS-A, HAS-A relationship mapping. Do this standardized data is output to the document-oriented database format file, the provider transmits the output file to the requester.
3. The collection and transmission of data: the transfer document-oriented database format files are stored in temporary storage by the data collection.

After the data collection is complete, MapReduce Manager, and outputs the integrated data on the MapReduce to Load the document-oriented file data stored on a temporary storage. The next chapter describes the MapReduce for Applications in the proposed system.

4. Application Example

This section describes the MapReduce to the system proposed through the application. In the proposed system, subjected to a process MapReduce before such as condition input, data collection. Figure 4 shows an overview of the application example of the proposed system MapReduce.

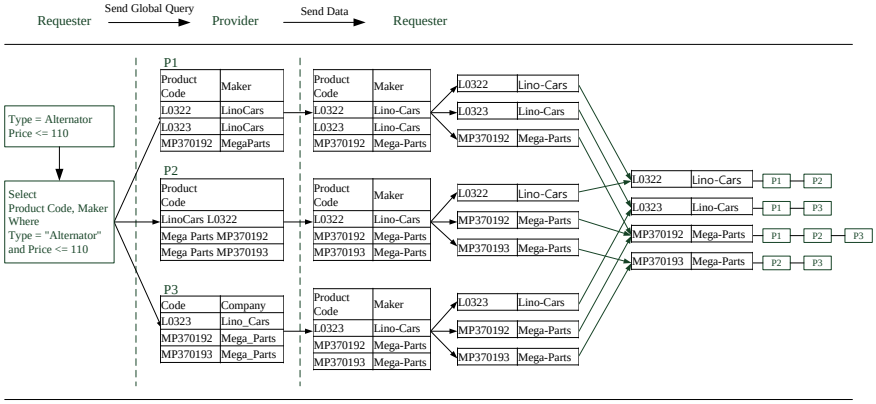

Figure 4. Overview of MapReduce based P2P DBaaS Hub System application example

In Figure 4 it is MapReduce proceeds through several processes. The first to inputting a search condition. The search conditions is Type = Alternator, Price <= 110. Second, generates global query by using the inputted search conditions. Third, the providers extract data using global query. Fourth, the extracted data is transferred to the requester is standardized. Fifth, Map the collected data. Sixth, Reduce the mapped data. data the reduce completed is finally merge and output.

5. System comparison

With the development of the IT industry, companies have seeking a various method for providing a convenient service to clients. One of them is collaboration between the companies. Recently, one way of

Table 1. Main parameters

Compare	P2P-MapReduce[4]	Proposed System
Operating Environment	Cloud Environment	Cloud Environment
Data processing methods	processing in the peer group	processing in the DBaaS hub System
Data Collection	The data collected from the outside	The data collected from peers in the P2P network through the DBaaS hub

The proposed system and P2P-MapReduce are operates in cloud environments. However there is some different, first P2P-MapReduce MapReduce to process the input data collected from the external environment. But the proposed system MapReduce data collected within the cloud environment. Second, P2P-MapReduce the MapReduce is performed in the P2P network, But in proposed system is performed in the requester's system. Both system are based on MapReduce, but shown in Table 1, difference occurs.

6. Conclusion

In this paper, we proposed a MapReduce based P2P DBaaS Hub system. the proposed system is provide integrated database environment to P2P cloud user based on MapReduce. In the proposed system, database integration is consist of three sequence, condition input, data collection, data integration and output. Data collection a function of gather data from a database is existing in P2P cloud. Proposed system's p2p cloud is based on super peer, it can reduce network traffic, prevent infinite loop, and search delay time. When the data collection, data exchange of peer to peer is using standard document-oriented format file through the DBaaS Hub system. it can standardlize data to using ontology based table, and exchange is possible structure of the database schema is different from each other. Because exchange of data in the form of a document-oriented database files even if a heterogeneous database can exchange data with each other. Finally, the data integration is through the MapReduce. As future research will be the study of MapReduce based adaptive P2P cloud system can applicate the different types of works.

References

[1] Kye-Dong Jung, et al. "The Study of DBaaS hub System for Integration of Database In the Cloud Environment." Journal of Digital Convergence 12.9 (2014): 201-207.

[2] Lehner, Wolfgang, and K-U. Sattler. "Database as a service (DBaaS)." Data Engineering (ICDE), 2010 IEEE 26th International Conference on. IEEE, 2010.

[3] Abouzeid, Azza, et al. "HadoopDB: an architectural hybrid of MapReduce and DBMS technologies for analytical workloads." Proceedings of the VLDB Endowment 2.1 (2009): 922-933.

[4] Marozzo, Fabrizio, Domenico Talia, and Paolo Trunfio. "P2P-MapReduce: Parallel data processing in dynamic Cloud environments." Journal of Computer and System Sciences 78.5 (2012): 1382-1402.

[5] Lee, Kyungyong, et al. "Parallel processing framework on a P2P system using map and reduce primitives." Parallel and Distributed Processing Workshops and Phd Forum (IPDPSW), 2011 IEEE International Symposium on. IEEE, 2011.

[6] Ozalp Babaoglu, Moreno Marzolla, Michele Tamburini. Design and Implementation of a P2P Cloud System. SAC '12 Proceedings of the 27th Annual ACM Symposium on Applied Computing. Pages 412 417. ACM New York, NY, USA. 2012

[7] Graffi, Kalman, et al. "Towards a p2p cloud: reliable resource reservations in unreliable p2p systems." Parallel and Distributed Systems (ICPADS), 2010 IEEE 16th International Conference on. IEEE, 2010.

[8] Li, Jin. "Erasure resilient codes in peer-to-peer storage cloud." Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on. Vol. 4. IEEE, 2006.

[9] Beverly Yang, B., and Hector Garcia-Molina. "Designing a super-peer network." Data Engineering, 2003. Proceedings. 19th International Conference on. IEEE, 2003.

[10] Min, Su-Hong, Joanne Holliday, and Dong-Sub Cho. "Optimal super-peer selection for large-scale p2p system." Hybrid Information Technology, 2006. ICHIT'06. International Conference on. Vol. 2. IEEE, 2006.