A Survey of Data Management System for Cloud Computing: Models and Searching Methods

Linhua Zhou, Kejia He, Xiaotao Sheng and Bobo Wang
School of Electronic and Information, Ningbo University of Technology, Ningbo 315016, China

Abstract: At present, the research of data storage and management in cloud computing mainly focuses on dealing with data expression and search. This study gives a comprehensive survey of numerous models and approaches of data-intensive applications in cloud computing in both academic and industrial communities. We review various approaches and their ideas of design. And then, we attempt to summarize and appraise the open issues.

Keywords: Computing pattern, data modeling, data querying and searching

INTRODUCTION

Cloud computing, as a new computing pattern having the capacity of sharing computing and storage, has been developed steadily and rapidly during recent years. It is selected by the famous scientific journal Technology Review published by MIT as one of the ten newly developed technologies which have changed the world. At present cloud computing has been the key focus of many enterprises and research groups.

Recently, with the development of technology of wireless sensor, communication, computer and storage, great amount of data are produced. The increasing speed of data scale has excelled that of Moore's law. The exponential increase will not stop in the long term. However, how to organize and manage those huge amounts of data so as to get the useful information for the users has been a new theme in the research area of cloud computing. With the rapid development of cloud computing technology and great investment that many companies and enterprises have provided the cloud computing service related to it, there is no need for many medium and small sized companies to buy expensive computers and storage equipment. They can rent equipments from companies with cheap prices without caring for the upgrading and maintaining of these equipments or the expansion of building the system based on the equipment. This is a great attraction to medium and small sized companies which care deeply about the cost budget. However, there still exist some problems: On the one hand, because of the different architecture of platforms or systems that each company offer, it results in the isolated information system which is hard to integrate and finally create an isolated information island. So there is an important problem in the cloud computing on how to present the data in the cloud environment effectively to help the customers to find the useful data among the existing cloud data center. Under such circumstance, the concern of users is the presentation of data and their searching, not the server where the data are located or the cloud computing center which provides resources. For the cloud computing users, their concern is to obtain the useful data that users need from the relevant data resources, or whether one’s data can be accessed by relevant companies conveniently. Therefore, how to get the data timely, accurately and reliably plays an important role in the success of the company’s decision making.

Apart from that, many companies are not willing to pull down the former information system to rebuild a new one. They would like to transfer to the cloud computing platform based on the existing system so as not to cause bad influence on the company’s present decision making and at the same time they can make full use of the advantages of cloud computing in the dynamic of dealing with data. Therefore, how to satisfy the coming dynamic searching need, compatibility of searching process of present data management system and the research of algorithm have been the key focus of cloud computing research. Especially in the 21st century, along with the development of data collecting technology with the large scale, great flux and universality and wide employ of giant storage
equipment, cloud computing platform has been the indispensable basic equipment and computing environment for the construction and processing of all kinds of distributing application system. Constructing such data presentation model and searching methods has been the key problem influencing the promotion of cloud computing application. Researching and exploiting the cloud data querying and searching methods among the data center of different architecture and establishing the corresponding model and method can not only help the users to describe their data accurately and conveniently, but also help them to adjust and optimize the present data searching strategies to gain more accurate data information so as to provide more reliable decision making support.

In this study, we review and summarize different innovative approaches to organize and manage giant data in the circumstance of cloud computing. Cloud data modeling is the foundation of cloud computing application and the searching algorithm upon it is the key issue of cloud computing application. They are gaining significant attention in the cloud computing community. Researchers have started to explore the modeling design and corresponding searching algorithm of the cloud computing application system, especially directing at relational database model, semi-structured data model and Unstructured data model.

**DATA MODEL AND SYSTEM ARCHITECTURE**

At present, one of the crucial research topics on data organization and management of the circumstance of cloud computing is how to describe the cloud data. In this section, some typical data models and system architectures are introduced. And we divide them into two kinds: the data models and system architectures of the academic community and the data models and system architectures of the industrial community.

**The academic community:**

**Semantic web models:** The semantic web technology which possesses the inference ability has become more and more popular in internet application domain. And it also arouses the enthusiasm of researchers in cloud computing community. Urbani et al. (2009) introduces the RDF data description model which aim at Map Reduce architecture. They utilize the property of RDF (such as RDF: sub Class of) to describe the cloud data. Neumann and Weikum (2010) also use RDF describe data, but different from Urbani et al. (2009) they put their focus on the query optimization. Sun and Jin (2010) state how to store the RDF data into HBase system and process them through Map Reduce approach. Ranabahu and Sheth (2010) employ DSL (Domain Specific Language) represent the cloud data, DSL is a programming language which utilize annotation to describe the specific domain data.

**Grid models:** In addition to semantic web technology, some researchers have introduced the successful results of the grid community to the cloud computing community. Foster et al. (2008) compare the cloud computing with grid computing and state these two computing patterns are common on target, architecture and technology, but differ from the programming model, computing model, business model and security. And they also suggest the community of cloud computing can borrow some successful ideas from the grid community. Voicu and Schuld (2009) shift their research finding of data grid to clouding computing domain. They use the data management protocol of the data grid to describe and manage the cloud data. Giunta et al. (2008) go a step further on the data management protocol. They introduce the WSRF (Web Service Resource Framework) to the cloud computing community and verify the application of WSRF on the e-science domain.

**Other models:** There are some researchers who consider introducing other computing models for resolving the problem of the data description of cloud environment. For example, Oliveira and Ogasawara (2010) apply the approaches of middleware community to the cloud computing environment. They propose the cloud middleware (SciCumulus) to describe and manage the workflow data of scientific activity. Zeng et al. (2009) adopt web service architecture to construct their cloud computing application and then based on this architecture they explore how to use web service to describe and manage their application data. Ludwig et al. (2009) however utilize the restful-based methods to describe and access the configuration information of the server side of cloud computing platform. Cui et al. (2010) use metadata to describe the personal information and integrate user data from different web sites. They emphasize the role of metadata and tag. Miyuki (2009) lay stress on the hierarchy processing. They propose the SLA (Service Level Agreement)-based method to integrate the data information coming from the different social network application system. Yiu et al. (2010) argue the owner of data should transform the original data when they provide the data and propose adopting transformation approach to describe the customer’s private data information.

**The industrial community:** Differing from the academic community, the industrial community is focusing very strongly on the simplicity, flexibility and scalability of the cloud data description model. Google Company, as the precursors of cloud computing, leads the world in the cloud computing research and application. Its core business, such as searching engine, Google Docs and Google Earth, all built on the distributed File System (GFS), large-scale distributed database (Big Table) and corresponding running...
framework (Map Reduce). Now hadoop project http://hadoop.apache.org that has been a popular platform of cloud computing is an open implementation of the Google Company’s cloud computing framework. The HDFS which is used for data storage is an open implementation of the GFS and the HBase which provide big table storage based on the HDFS is an open implementation of the Big Table and the Hadoop Map Reduce which is used for process the data of the clusters is an open implementation of Map Reduce programming framework. Both Google Company’s cloud computing platform and Hadoop project are designed for searching the contents of web pages, so their data description model cater to the big grain and file format data. Amazon company develops the Dynamo system and S3 (Simple Storage Service) for their customers, while the Dynamo uses the key/value to present customer’s data and S3 service permit customer to define themselves data description model. Brattner et al. (2008) construct a relational data model based on S3.

Microsoft Company is also launching their cloud computing system, Dryad system (for internal use) and SQL Azure database system (for external use). Facebook Company introduces the Cassandra system which adopts the form of key/value to present structural data.

QUERYING AND SEARCHING METHODS

Data searching is the most important function in data management. Gounaris (2009) argue, nowadays, the querying tools and technology designed for web services does not fit the querying requirement of cloud computing environment. And they suggest researching the next generation searching engine and tools that against the cloud computing application. At present, there are two kinds of research direction to the data searching methods of the circumstance of cloud computing: Map Reduce-oriented searching methods and SQL-like querying methods (Husain et al., 2011).

The map reduce-oriented searching methods: The Map Reduce-oriented searching methods provide the outstanding performance in large-scale data searching and analysis. It has become the key issue of the searching research of the cloud data. The research result of Dean and Ghemawat (2010) prove that the Map Reduce framework works well in processing data among heterogeneous system. With comparing to other framework, Pavlo et al. (2009) concludes the Map Reduce framework is appropriate for applications of searching large-scale data sets. Jiang et al. (2011) select Map Reduce framework as their searching platform and then analyze their data. The Pregl system developed by Malewicz et al. (2010) uses Map Reduce framework store and process large-scale distributed gallery. Combining with semantic web technology, Olston et al. (2008) use the Map Reduce framework of Hadoop to realize its RDF database searching which are stored in the HDFS file system. Husain et al. (2010) also use the Map Reduce framework to deal with the searching in the RDF map. Ma et al. (2009) applies the Map Reduce framework to the large scale mobile data query.

SQL-like data query methods: Compared with Map Reduce, SQL query has stronger expression ability such as join (Stonebraker et al., 2010). Therefore, some researchers and developers have developed corresponding SQL-Like data query according to their own system. The goal of Hive project is to build a data warehouse platform and use Hive QL language of SQL-Like to search the data stored in HDFS. Yahoo uses Pig Latin (Gates et al., 2009) which is similar to SQL query to search the data by constructing a query plan to operate the corresponding query of the data stored in the cloud computing environment. Microsoft Company proposes SCOPE (Structured Computations Optimized for Parallel Execution) by assimilating the characteristics of SQL such as organizing the data in the dataset by row and column method, supporting join operating (Chaiken et al., 2008).

Some researchers have shifted some traditional searching methods to the cloud computing environment in some specific areas. Teregowda et al. (2010) use such method to analyze its availability and flexibility in the data information management of digital library. Kraska et al. (2009) transfer the search and management of dataset log information to S3 cloud computing platform of Amazon and make a comparative analysis of its performance. Li et al. (2009) apply the traditional text query technology to cloud computing to find the searching performance of cloud computing.

CONCLUSION AND OPEN ISSUES

The aim of cloud computing is to make people compute and store the resources as easily as using water and electricity. Under such kind of computing model, more and more enterprises will transfer the storage data to cloud computing center. And the scale becomes more and more large. For now, the research of data storage and management in cloud computing mainly focuses on dealing with data expressing and searching. In this study, we provided a survey on numerous models and approaches of tackling these problems. We review the various approaches and its idea of design. There are still many open issues, such as, mobile data management in cloud environment, backend-support for data-intensive applications of cloud computing, integration of different cloud computing platform, data mining in cloud computing, etc., which can be considered as scopes for future studies.

ACKNOWLEDGMENT

This study is supported in part by Zhejiang Mobile Network Application Key Lab program (No. 246


