ISBN 978-981-11-0008-6
Proceedings of 2016 6th International Workshop on Computer Science and Engineering (WCSE 2016)
Tokyo, Japan, 17-19 June, 2016, pp. 386-389
doi: 10.18178/wcse.2016.06.062

Design of a Big Data Accessing and Processing Architecture Using Cloud Computing Technologies

Chao-Tung Yang [1], Hsin-Wen Lu [1] and Wen-Chung Shih [2 +]

[1] Department of Computer Science, Tunghai University, Taichung, Taiwan, R.O.C.
[2] Department of Computer Science and Information Engineering, Asia University, Taichung, Taiwan, R.O.C.

Abstract. Big Data analytics has emerged as a promising topic and gained more and more attention. In order to have scalable load capacity for data platforms, we must build them in good architecture. Some issues must be considered in order to use the cloud computing to quickly integrate big data into database for analyzing, searching, and filtering big data to obtain valuable information. In this paper we propose a design for cloud storage platform with HBase for storing and analyzing big data and improve the performance of importing data into database. It also can operates through Hadoop MapReduce for HBase database to do distributed data processing, and to provide functions, including keyword search, data filtering, and basic statistics.

Keywords: big data, cloud computing, data access, data processing

1. Introduction

According to analysis by IBM, by 2015 the global fast-growing big data [1], [2] is expected to exceed 8000 Exabyte. For example, the amount of medical information will increase about 50 times. Analysis of big data will be of great value in the next 10 years, covering medical, manufacturing, telecommunications, retail, energy, transportation, security, and other application markets [3]. Data analysis is performed on information and data generated in the past; after analysis, it forms the basis for future actions and decisions making. In fact, it is quite common practice in life. However, the data processing and analysis have become more complex as data volume gets bigger and bigger, with more various types. Hence, how to use cloud computing to integrate big data into database [4], [5] and how to analyze, search and filter big data to obtain valuable information and extract useful information for analysis have become important issues.

Big data refers to data of a huge amount generated within a reasonable period of time which cannot be artificially captured, managed, processed, interpreted and organized as information that can be read by human beings. With the assumption of the same total amount of data, compared to individual analysis of small independent data sets, analysis of assembled individual mini data sets can draw a lot of additional information and data relationships that can be used to detect business trends, determine the quality of research, avoid spread of diseases, fight crimes, detect real-time traffic states and so on. Those prospects of usages are the reasons for the prevalence of big data analytics. However, big data almost cannot be processed by most of the traditional database management system, but must be processed by software running in parallel on tens, hundreds or even thousands of servers. Definition of big data depends on the capacity of the institution holding data sets, as well as its software ability to analyze and process data.

2. System Design and Implementation

2.1. Platform overview

+ Corresponding author. Tel.: +886-4-2332-3456; fax: +886-4-2332-0718.
E-mail address: wjshih1@gmail.com.

Due to the fact that big data is usually scattered in different institutions and its format is varied, the integration of big data with increasing amount is challenging. In order to have sufficient load capacity of these data platforms, well designed architecture is called for and it is the main topic of this work. Therefore, this work proposes to use several servers to compose a cloud cluster, namely a master server and several nodes. It will be carried out in three stages when importing data records through the user interface, as shown in Figure 1:

- Step 1: From the end user to upload data to the cloud platform, and the data stored to the dispersed nodes.
- Step 2: After the data is stored, the cluster will provide services to allow users to process, search, and filter data via the cloud.
- Step 3: Through processing, searching, and filtering data, the system will be able to provide useful information to users.

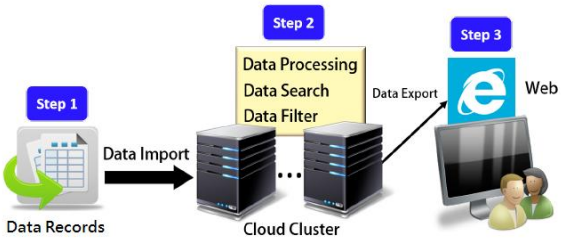

Fig. 1: The overview of the proposed platform.

2.2. System architecture

This work proposes to develop a platform for big data processing on cloud. Most records are in dispersed data formats, so they are assumed to be saved in comma separated values (CSV) format. Data can be easily stored using CSV format in platforms. CSV is a common format that can be easily accessed by most programs. When data are imported into the cloud-based platform, the data collection service will start putting data into a distributed database system. At the same time, the data will also be stored in cloud storage. Data Search Service and Data Processing Service do computing and processing through the cloud computing platform after depositing data into the cloud storage, and Web Application Service operates via the front-end user interface to do data exchange. Finally, jQuery is adopted to implement the user interface, and the web page presents visual interface using HTML5, JavaScript and CSS 3 technology, as shown in Figure 2.

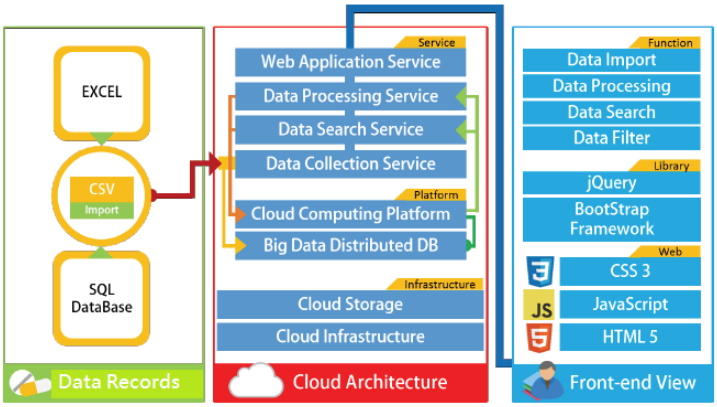

Fig. 2: System architecture.

2.3. Cloud-based architecture

The cloud-based implementation, as shown in Figure 3, has *n* hosts, and each host has its own CPU, memory and storage space to form a distributed computing system as Apache Hadoop cloud computing clusters. Data storage in the cluster uses Apache HBase NoSQL database storage, data access to HDFS on

HBase, and Hadoop MapReduce uses ZooKeeper to access the data. Cloudera CDH Cluster is responsible for deploying and monitoring.

- Data Collection Service: After building the cloud platform, collection of big data becomes the next important issue, i.e., how to quickly and accurately import the big data into database. In this paper three methods are used for data import. The first method is to use the single-threaded; the second, MapReduce; and the third, Completetbulkload method to import.
- Data Collection Service with Single-Threaded: The single-threaded method is to import data by individually scanning each row and column. When data is imported into the distributed database, it should communicate to HBase through ZooKeeper. The data will be put into HBase one by one. HBase will store data in HDFS. This method is suitable for use for a small amount of data import. When the amount of data becomes larger, much time will be required and this data import method is inefficient.
- Data Collection Service with Completebulkload: To import tab-separated values (TSV) data using MapReduce, first the data is converted into HBase files called HFile, then the data is added by the Completebulkload method and becomes HBasetables in HBase. This method is suitable for large-scale data import. Because the file is turned into HFile and then sent to HBase, omitting the process that causes waste of resources and time.
- Data Processing Service: Data Processing Service is through Hadoop MapReduce to filter data and perform basic data statistics. First, the user needs to send his requests to Data Processing Service. Second, Data Processing Service will organize the user's requests, which will be converted as the MapReduce program for execution. Third, the MapReduce program will be divided into Map() and Reduce() in two parts; Map() will get stored data through HBase and do basic filtering, and Reduce() will integrate these data and return the result to Data Processing Service. Fourth, Data Processing Service will send the results back to the user.
- Data Search Service: Data Search Service provides users with the keyword search service. When a user starts information searching, first of all, the keyword and search rules will be sent to Data Search Service. Second, Data Search Service will convert the search request into the Hadoop MapReduce program and send it to Hadoop for execution. Third, when executed by MapReduce, Map() divides the data from HBase into many blocks, each block is independently processed by individual Map(). Information found by each Map() will be sent to Reduce() for arrangement, and then the result of Reduce() will be sent back to Data Search Service. Fourth, Data Search Service sends the data back to the user.
- Web Application Service: The Web Application Service, responsible for communication with the front end, receives service requests from the front-end and coordinates other system services such as Data Search Service and Data Processing Service. Web Application Service sends results back to the Front-end View.

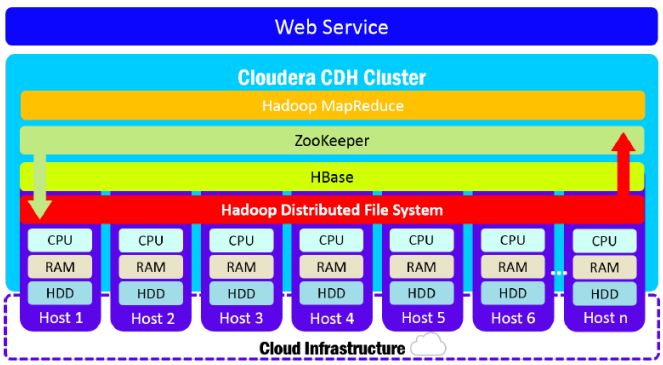

Fig. 3: The cloud-based implementation.

2.4. System implementation: A medical case

Six hosts were used to build a cloud cluster platform by using Cloudera Manager to set up ZooKeeper, Hadoop HDFS, Hadoop MapReduce, and HBase. On deployment of the platform environment, Cloudera Manager is used to monitor system states, such as Hosts, HBase, HDFS, MapReduce and ZooKeeper, and to monitor resources, such as CPU usage, Disk I/O, network and HDFS I/O.

Cloudera Manager can also monitor the status of each node, confirming normal connections of each host. Cloudera Manager checks at regular intervals, and it will warn if connections are abnormal or the connection quality is poor. Cloudera Manager can remove nodes at any time, adding or removing nodes into or out of the cluster. Through the six hosts, i.e., the one NameNode and five DataNodes, the Hadoop NameNode shows that the cluster provides 5.3 TB of big data storage space. This information also shows the number of live nodes.

3. Conclusion

In this paper we propose a cloud platform architecture to process big data. Through distributed cloud architecture a database is constructed to build HBase database that can load and store big data. Hadoop MapReduce distributed computing architecture is used for data processing and analysis. One of the important functions of the platform is to quickly store a large number of data records into the cloud platform. In future work we will compare the cost of time via different ways to import data to HBase and identify the best performance of the experimental design for data import to accelerate the speed of the cloud platform. In the architecture, users can use the web interface on the platform to efficiently search data by specifying ”RowKey”, ”RowKey Range” and ”Keyword.” The platform also provides a filter function for users to filter data stored on the platform by choosing filtering rules and users can add rules to further filter for data in each column. In summary, in this paper a big data processing platform on cloud is proposed. In addition to store big data on the cloud platform, data processing functions such as search, filter, and data analysis of the stored data are also designed.

4. Acknowledgements

This research was supported by Ministry of Science and Technology of Republic of China under the number of MOST 104-2511-S-468-004 and MOST 104-2511-S-468 -001 -MY2.

5. References

[1] R. Li, Z. Xu, W. Kang, K. Yow, and C. Xu. Efficient multi-keyword ranked query over encrypted data in cloud computing. Special Issue on Extreme Scale Parallel Architectures and Systems, Cryptography in Cloud Computing and Recent Advances in Parallel and Distributed Systems, fICPADSg 2012 Selected Papers, vol. 30, pages 179-190, 2014..

[2] S. Rio, V. Lopez, J. Benitez, and F. Herrera. On the use of mapreduce for imbalanced big data using random forest. *Information Sciences*, 2014.

[3] Z. Chen, S. Yang, H. Zhao, and H. Yin. An objective function for dividing class family in nosql database. *Proceedings of 2012 International Conference on Computer Science Service System (CSSS)*, pages 2091-2094, Aug 2012.

[4] W. Hsu, H. Lu, C. Yang, J. Liu and W. Chu. Implementation of data transform method into nosql database for healthcare data. *Proceedings of 2013 Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT'13)*, 2013.

[5] S. Lombardo, E. Di Nitto, and D. Ardagna. Issues in handling complex data structures with nosql databases. *Proceedings of 14th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pages 443-448, Sep. 2012.