2017 Asia-Pacific Engineering and Technology Conference (APETC 2017)
ISBN: 978-1-60595-443-1

Key Problems of Data Management in Data Center in Big Data Context

Teng Lv and Ping Yan

ABSTRACT

This paper analysed the challenges of data management in data center, such as big data volume, data heterogeneous, high time requirement of data processing, and widely separated data sources. We discussed the disadvantages of traditional data management technologies to deal with these problems. We also highlighted the key problems of data management in data center including data integration, data analysis, and representation of data analysis results.

INTRODUCTION

The current information society has entered the era of big data. March 2012, the U.S. government issued a big data research and development initiative[1], which investments 200 million U.S. dollars to start the big data development plan. The U. S. government's plan is another major move following the information highway plan in the field of information of United States.

In the era of big data background, the construction of data engineering is faced with many opportunities and challenges. With the development of various equipments and information technologies, the data center has collected a large variety of data, such as constructed data from relational databases, semi-constructed data from XML documents or native XML databases, and unconstructed data from web pages, audios, and videos, etc. And with the elapse of time, the type of data will be more diverse and the amount of data will be greater, so the data is growing at a rapid pace. It can be seen that in the era of big data, the data is asset. The Data Asset provides guidance for not only building the business case for data quality and data governance, but also for developing methodologies and processes that will enable your organization to better treat its data as a strategic asset[2]. If you can't effectively manage these existing data and new data, you can't play the value of these data. At present, we put forward the following challenges to the data management of data center:

First, the size of data is large. With the extensive application of new equipments and extensive applications of information technologies, data center usually have a

Teng Lv[1] and Ping Yan[2,*]
[1]School of Information Engineering, Anhui Xinhua University, Hefei 230088, China
[2]School of Science, Anhui Agricultural University, Hefei 230036, China
Corresponding author: want2fly2002@163.com

great amount of data to be managed. For such large amount of data, it is beyond the ability of the traditional data management systems to manage and process these data[3].

Second, the data type is various. The data center usually contains a variety of data types[4], including traditional structured data such as data stored in a relational database, multimedia data such as text, images, sounds, videos, and even a variety of cyber sources. These data have differences in structures, and some of the data have incomplete schemas or no schemas at all. All these situations bring a lot of inconveniences to the management of data.

Third, the effectiveness of data process is strong. In data center, some applications have to be finished or processed immediately such as stream data process[5]. If the data cannot be processed within the prescribed time limit, it will lose the value of data.

Lastly, the distribution of data sources is wide. The data center usually collects a huge data from sources of broad geographical distribution[6]. Here are some typical situations of such applications: When data is collected from different departments of a company, or from different sensors of a widespread wireless sensor networks, or from different devices from mobile phones of social networks, or from deferent devices form mobile objects such as automobiles. All of them can cause widespread distribution of data sources in a data center. In particular, maneuvering and deployment of movement in the process of data may change from time to time, which can also cause the delay of collection and management of data in some cases.

Therefore, how to effectively manage the data, how to analyze the valuable information from the given data, and how to show the results of the analysis are urgent problems to be solved for the data engineering in a data center. In fact, the traditional data management technology, especially the relational database technologies faced the difficulties to meet the demand of data management in modern data center, which are showed in the following aspects:

(1) It cannot meet the time requirements of data management in data management, especially when data is distributed across multiple nodes. With the increase of nodes, a relational database for consistency and integrity constraints will greatly reduce the performance of the database, which cause slow response of data query and cannot meet the actual real time needs of data.

(2) It is unable to meet the needs of heterogeneous data management in data engineering. Although the traditional relational database technology is very good at dealing with structured data, but it is cannot efficiently manage a large number of semi-structured, unstructured data in a data center.

(3) It is not effective in dealing with the data with missing schema or without schema at all. The traditional relational database system requires the data to be in the first normal from before it can be stored in relational databases. But for the data in a data center, this limit is often too strong because many data have only missing schemas or even has no schemas at all.

In summary, we believe that the construction of the data center have to learn from the latest experiences and technologies of the current big data research. This paper focuses on the research status of data management in data center and discusses several key scientific problems in common based on big data perspective: (1) data integration; (2) data analysis; and (3) representation of results of data analysis.

DATA INTEGRATION

Data integration is the basis of data management in data engineering. The goal of data integration [7] is to query a variety of different data sources by providing a unified object schema interface to meet the needs of the user's query. In data engineering, data sources are often distributed in different locations, and may use different data schemas. It is difficult to use and manage the heterogeneous data as heterogeneity between the distributed data. This requires data integration of a large number of heterogeneous distributed data. Data integration usually involves the following steps:

First, data preprocessing. Data preprocessing[8] usually includes removing noise data, reducing redundant data, and cleaning the dirty data. Data preprocessing can ensure the quality and reliability of the following data integration stages.

Then, schema mapping[9]. Schema mapping is a mapping between source and target schemas. It is usually required to give a set of mapping rules to define how the related elements in the source schema corresponding to the related elements in the target schema. In the process of schema mapping, special attention should be paid to the processing of the data in the absence of schemas or no schemas. It is also important to deal with semantic consistency and semantic conflict before and after schema conversion.

Lastly, query processing[10]. Data integration system is responsible for converting a query of the global schema to multiple source schemas, i.e., query rewriting. For lossless source schema mapping to global schema, the equivalence of query rewriting can be easily maintained. But in real data engineering, this kind of ideal mapping is often not exist, and most mappings are lossy, i.e., the mappings are only incomplete relations from partial source schemas to a partial global schema. In this case, the data source and its mapping only provide an incomplete view. Therefore, it is necessary to deal with the problem of query problem in lossy mapping efficiently.

In addition, considering the problem of multi-source heterogeneous data in data engineering, it is also need to deal with the uncertainty in data integration. According to the hierarchical level of data integration, the main source of uncertainty in data integration in data engineering is as followings [11]:

(1) The uncertainty of the source data themselves. For example, sensor networks or RF generated data are often uncertain due to the influence of equipments or the environments; some data are inaccurate due to technical or policy reasons such as the lack of data in the database, real-time data, sensitive data and data policy-sensitive data etc.

(2) The uncertainty of schema mapping. In many cases, it is difficult to determine the mapping between a given source schema and target schema; in addition, with the evolution of the data integration system, it becomes more difficult to maintain a certain schema mapping relationship as the original source schemas change in the data center, new source schemas add in the data center, and the application change in the data center.

(3) The uncertainty of query conversion. The query of a user submitted to a data integration system is needed to be converted into a query of related source schema. It will cause a miss match or an inexact match between the original query and the converted query in many cases, which results in the uncertainty of the query conversion.

In summary, the data integration of the data engineering becomes more difficult and challenging in the process of data integration due to the introduction of uncertainty factors in many aspects. How to deal with the data itself, the uncertainty source schemas and target schemas mapping, and the uncertainty of query conversion in the data integration system is very challenging.

DATA ANALYSIS

The purpose of a data engineering construction is to better use of existing data and extract valuable information from the data to serve various operations. Because the data in data center has the characteristics mentioned above, the traditional data analysis [12], such as data mining, machine learning, and statistical methods are very difficult to adapt to the new features of data analysis, so they cannot be directly applied to the data in a data center. We must solve some common problems of data analysis in a data center in the followings:

(1) The traditional data analysis algorithms must solve the problem of performance in large data environment. For the traditional data analysis algorithms, their performances are often no problem when the data size is not very large. But with the increase in the size of the data processing, their performances tend to decline [13], especially in the case of large data processing data in the data center. In addition, the start and end time of the task are mandatory in the real applications of the data engineering. Therefore, it is a new challenge that how to improve the traditional data analysis algorithms in order to meet the requirements of mass data processing in the data engineering. In this case, we can learn from the latest achievements of big data research, such as C-AMAT mathematical model of the new research [14], which can greatly reduce the data I/O delay through the parallel memory read mode to process the cache and storage in parallel.

(2) The traditional data analysis algorithms must solve the data noise problem in big data environment. In the construction of the data center, it is often necessary to integrate data from multiple heterogeneous data sources[15-17]. These data are not the same even divergence greatly in many aspects, such as schemas, formats, forms, and so on. In addition, it often can cause data inconsistency or even conflict due to the transmission, conversion, processing and other stages in the process of data collection. The traditional data analysis algorithms are often helpless in the face of such heterogeneous and large amount of noise data. Therefore, it is also a new challenge how to improve the traditional data analysis algorithms in order to adapt to the large number of noise in the data environment.

PRESENTATION OF THE RESULTS OF THE DATA ANALYSIS

Data engineering is ultimately to service decision-making of various applications and activities. For the results of data analysis, it will reduce the use efficiency of the results or even misleading users if the analysis results are not presented in a way of reasonable, easy to understand and use to the presented to various decision-making users. The main presentation of the results of the data analysis has the following ways:

(1) In the form of text. This form of representation is mainly used to display some simple query results, such as parameters, attributes, and simple information, etc.

(2) In the form of graphics[18]. For some of the more complex analysis results, such as the complex information of statistical data, using text form to present the results of the analysis is not only lengthy, but also difficult to understand. If the graphical form is used to describe these data, it is very intuitive and easy to understand even at a glance. But how to use the graphical form to represent analysis results, many technical problems must resolved such as graph construction, rendering, update and how to determine the complexity of graphics reasonably. These problems are often closely related to space and time and needed to pay special attention to the time and space requirements.

(3) In interact way. It will provide an interactive way so that the decision-making personnel and other types of users can interact with data analysis results. Various users can operate the data and the related analysis of the results of the data can vary accordingly. These operations include the replacement of a part of data, increase a part of data, and change a part of data, etc. in order to visually observe the results of the modified analysis. This kind of way of representing analysis results is particularly suitable for gradually detailed an expectations of personnel.

CONCLUSIONS

The construction of the data center is a major event and critical to the development of an organization, so it is a systematic project. In the process of construction, we must have a detailed macro plan. Only through the use of scientific research methods of the latest contemporary research results, can we construct and use a data center well. In this paper, we refer to the latest developments of the current information technology, especially the latest research trends and results of big data, and consider the challenges facing the construction of data center, such as the of large scale of data, varieties of data types, strong time-bound of data processing, and widely distribution of data source. According to the above analysis, we discusses some key problems of data management based on the perspective of big data, including data integration, data analysis, and presentation of data analysis results. We believe that the research of this paper can provide some ideas of construction and management for the data center, especially in the aspect of data management.

ACKNOWLEDGMENT

The work is supported by Colleges Nature Science Research Key Project of Anhui Province (No.KJ2015A325), Anhui Province Quality Engineering (No.2015zy073), Academic Leader Foundation (No.2014XXK06), and Introduction of Talents Foundation and of Anhui Xinhua University.

REFERENCES

[1] Big Data Initiative - The White House, www.whitehouse.gov/sites/ (2012).
[2] T. Fisher, The Data Asset (2009).
[3] L. Wu, L. Yuan, J. You, Journal of Computer Science and Technology, **30**, 1, (2015).
[4] X. Liu, P.V. Singh, and K. Srinivasan, Marketing Science, **35**, 3, (2016).
[5] S. Kamburugamuve，S. Ekanayake, and M. Pathirage, G. Fox, Ipdps, Hpbdc (2016).
[6] L. Gu, D. Zeng, P. Li, et al, Cloud Networking for Big Data, **2**, 3 (2014).
[7] M. Lenzerini, Proc. of the 21st ACM PODS (2002).

[8] S. García, S. Ramírez-Gallego, J. Luengo et al, Big Data Anal, **1**, 9 (2016).
[9] X.L. Dong, D. Srivastava, Proceedings of the Vldb Endowment, **6**, 11 (2015).
[10] H. Wang, X. Qin, X. Zhou. et al, Front. Comput. Sci., **9** (2015).
[11] X. Dong, A. Halvey and C. Yu, The VLDB Journal, **18**, 2 (2009).
[12] X. Meng and X. Ci, Journal of Computer Research and Development, **50**, 1 (2013).
[13] A. Rajaraman, J. Ullman, Mining of massive datasets, http://infolab.stanford.edu/ ~ullman/mmds.html (2014).
[14] Chinese experts in the United States proposed a new model to significantly improve the computational speed. http://news.xinhuanet.com/tech/2014-03/31/ c_1110023038.htm (2014)
[15] J. Fan, F. Han, H. Liu, arXiv:1308.1479v2 (2016).
[16] H. Hu, Y. Wen, T.S. Chua et al, Access IEEE, **2** (2014).
[17] B. Ramesh, Big Data (2015).
[18] C.K. Chui, F. Filbir, H.N. Mhaskar, Applied & Computational Harmonic Analysis, **38**, 3 (2015).