An Integration Method of Massive Heterogeneous Data based on Big Data

Bo He[1,a], Lili Zhu[1,b], Huanli Zhang[1,c]

[1]School of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054, China.

[a]29659807@qq.com, [b]52980551@qq.com, [c]heboswnu@sohu.com

Abstract

Massive heterogeneous data has the characteristics of huge amount of data, high distribution, heterogeneous data and incremental data. The existing data integration methods can not solve the bottleneck problem of massive heterogeneous data integration for small-scale data. To solve this problem, this paper puts forward an integration method of massive heterogeneous data based on big data by using the advantages of MapReduce in processing massive data and the advantages of data view and large database HBase in integrating massive heterogeneous data.

Keywords

Big Data; Massive Heterogeneous Data; Data Integration.

1. Introduction

Massive heterogeneous data has the characteristics of huge amount of data, high distribution, heterogeneous data and incremental data.

In March 2012, the Obama administration announced the launch of the "big data research and development plan" [1], which raised "big data" from commercial behavior to national strategy.

Big data [2,3] refers to a data set that takes more time than can be tolerated to acquire, manage, mine and process data by using common software tools.

Data integration is the logical or physical integration of data from different sources, formats and characteristics, so as to provide comprehensive data sharing.

The existing data integration methods are aimed at small-scale data and are not suitable for the integration of massive heterogeneous data.

MapReduce [4], a big data computing model, decomposes the problems to be executed into map and reduce operations, which is very suitable for the integration of massive heterogeneous data.

2. Current situation of massive heterogeneous data integration methods

Data integration centralizes and preprocesses heterogeneous data from multiple sources, which is the basis of further data mining. Some scholars have made preliminary research on data integration methods. Representative research achievements include "research and implementation of a text data integration method" published by Chen Feiyan et al. [5], research on parallel data integration method of cloud key value data warehouse published by Liu Junqiang et al. [6], research on data integration method of distributed information system in grid environment published by Qiu Shuwei et al. [7], and published by Huang pan et al "Research and application analysis of data integration method in information system" [8].

The above research results are aimed at the integration of small-scale data and can not solve the bottleneck problem of massive heterogeneous data integration. Therefore, according to the characteristics of massive heterogeneous data, how to integrate massive heterogeneous data is an urgent research. The development trend is to use the advantages of MapReduce in processing massive data and the advantages of data view and large database HBase in integrating massive heterogeneous data, This paper studies the integration method of massive heterogeneous data based on big data.

Association rules describe the rules of frequent itemsets in a given transaction set. The key of association rule knowledge discovery is to obtain frequent itemsets. Common association rule knowledge discovery methods include apriori, FP growth, etc.

Apriori is a typical association rule method. It adopts the iteration of layer by layer search and uses k-itemset to generate K + 1 itemset. This method is simple, but it has the problems of many times of scanning data, many times of synchronization and low execution efficiency.

The frequent pattern tree is a tree structure that meets the following three conditions: (1) it consists of a tree marked "null". (2) each node in the item prefix subtree contains three fields: item name, count and node link, where item name records the item name, count records the number of transactions represented by the path to the node, and node link points to the same item name value in the frequent mode tree Node link is null when the next node does not exist. (3) each table item in the frequent item header table contains two fields: item name and head of node link, where head of node link is a pointer to the first node with the same item name value in the frequent mode tree.

The FP growth method is based on the frequent pattern tree and only needs to scan the data twice, which greatly reduces the scanning times and calculation time of the data.

The existing data integration methods can not solve the bottleneck problem of massive heterogeneous data integration for small-scale data. To solve this problem, this paper puts forward an integration method of massive heterogeneous data based on big data.

3. Integration method of massive heterogeneous data based on big data

According to the characteristics of massive heterogeneous data, taking advantage of MapReduce's advantages in processing massive data and the advantages of integrating massive heterogeneous data with data view and large database HBase, an integration method of massive heterogeneous data based on big data is proposed by establishing MapReduce, unified data view and large database HBase.

Firstly, build a unified data view for massive heterogeneous data and establish the mapping relationship between the unified data view and massive heterogeneous data. Secondly, according to the mining theme, use map decomposition task to extract data from massive heterogeneous data of different networks. Then, use reduce to merge and integrate the extracted data into the large database HBase. Finally, analyze the data in the large database HBase Preprocess the data to obtain massive isomorphic data, as shown in Figure 1.

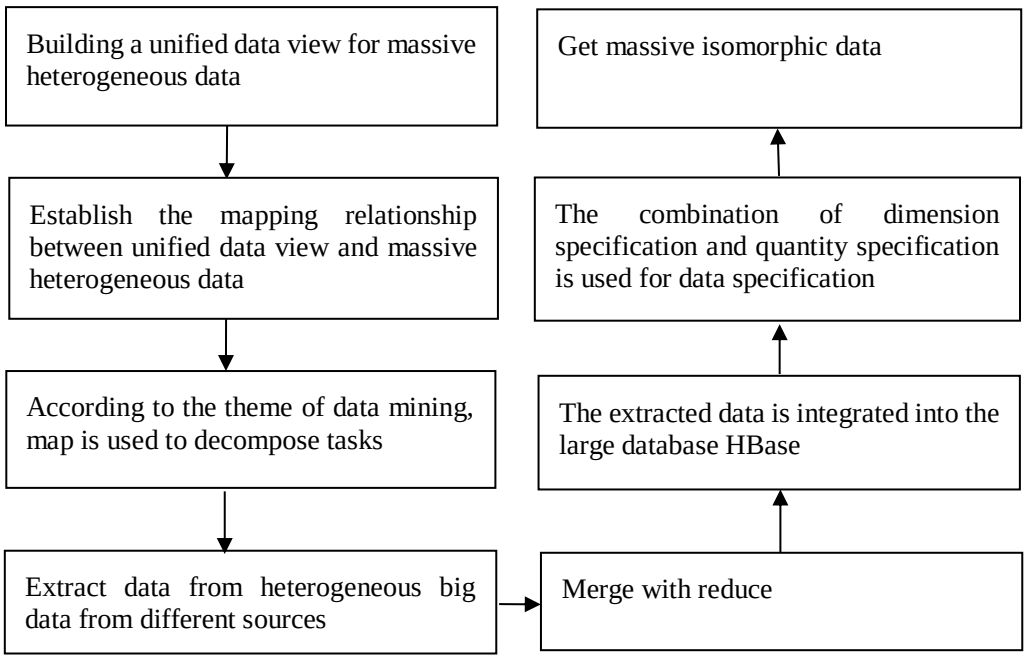

Figure 1. An integration method of massive heterogeneous data based on big data

4. Conclusion

Using the advantages of MapReduce processing massive data and the advantages of data view and large database HBase integrating massive heterogeneous data, this paper puts forward an efficient massive heterogeneous data integration method based on big data. The next step is to experiment and apply the proposed method.

Acknowledgments

This research is supported by the research fund for humanities and social sciences of the ministry of education under grant No.19XJA910001.

References

[1] Big Data Across the Federal Government [EB]. http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final_1.pdf, 2012.

[2] Science. Special Online Collection: Dealing with Data [EB]. http://www.sciencemag.org/site/special/data/ , 2011.

[3] Meng Xiaofeng, CI Xiang. Big data management: concepts, technologies and challenges [J]. Computer research and development, 2013,50 (1): 146-149

[4] Zhang Jifu, Li Yonghong, Qin Xiao, Xun Yaling. Local outlier data mining algorithm based on MapReduce and correlation subspace [J]. Journal of software, 2015,26 (5): 1079-1095

[5] Chen Feiyan, Hu Liang. Research and implementation of a text data integration method [J]. Journal of Northeast Normal University (NATURAL SCIENCE), 2016 (1): 78-83

[6] Liu Junqiang, Zuo Hongfu, Peng Zhiyong. Research on parallel data integration method of cloud key value data warehouse [J]. Computer application research, 2015,32 (8): 2458-2460

[7] Qiu Shuwei, Zheng Lin, Huang Jianxin. Research on data integration method of distributed information system in grid environment [J]. Journal of Guangzhou University (NATURAL SCIENCE EDITION), 2012,11 (2): 70-75

[8] Huang pan, Wang Dongdong, Wang lulu. Research and application analysis of data integration method in information system [J]. Shandong industrial technology, 2015 (7): 179-179.