Analysis and Processing of Massive Data Based on Hadoop Platform

Chenxiang Zhang

Department of Information Technology, Suzhou Industrial Park Institute of Services Outsourcing, Suzhou 215123, China

zhangchenxiang@126.com

Keywords: Massive Data, Data Processing, Hadoop

Abstract: How to extract useful information from massive data quickly becomes the most difficult problem that application software developers encounter in curriculum development. Based on the analysis of the key technical foundation and other existing distributed storage and calculation researches on Hadoop cluster technology combination, as well as their business needs and the actual hardware and software capabilities, the paper proposes a large-scale Hadoop data processing based on model and data structure design program in several processes of organizing and using the programming methods, introduces the development of the model, model of log data preprocessing and its application to large website.

1. Introduction

With the rapid development of computer technology and Internet technology, a large amount of data is constantly emerging. It is urgent for enterprises to change their traditional architecture, and how to analyze these data and how to make full use of data value in the face of massive data. At the same time, how to optimize enterprise management has become an inevitable problem in the process of modern enterprise transformation. The amount of data is only one aspect of the challenge of massive data mass data, referring to the other two aspects of speed and diversity. Speed represents response speed requirement collection, processing, and data query data. Diversity refers to the format and content of changing data. In massive data processing, how to mine potential value and transformation ability from mass data efficiently and quickly will provide a basis for decision making, and will become the core competitiveness of enterprises. The importance of data analysis is unquestionable. But with the faster and faster data generation and larger data volume, data processing technology faces more and more challenges. How to excavate useful value from massive data, analyze deeper meaning and transform it into operable information has become a problem that Internet Co must deal with.

The parallel and distributed research of mass data processing models and algorithms on the cloud platform has sufficient research value. In this paper, as the research platform of Hadoop cloud platform, for a large number of Web log data preprocessing model, the research of massive data processing performance of the Apriori algorithm based on distributed data mining, effectively improve the cloud platform, make a contribution to promote the development of large data processing technology.

2. Characteristics of massive data

Massive data is generally used to describe a lot of unstructured data and semi-structured data, the data in a relational database for downloading to spend too much time and money when analyzing. Massive data analysis and cloud computing often linked together, because real-time analysis of large data sets requires the same as Map Reduce framework to assign to computer tens, hundreds or even thousands of jobs [3-4].

Massive data requires special techniques to effectively deal with a lot of tolerance through time data. Suitable for mass data technologies, including massively parallel processing (MPP) database, data mining grids, distributed file system, distributed databases, cloud computing platform, the

Internet and scalable storage system [5]. For the characteristics of massive data, you can use Volume, Variety, Value, Velocity to summarize, as shown in figure 1.

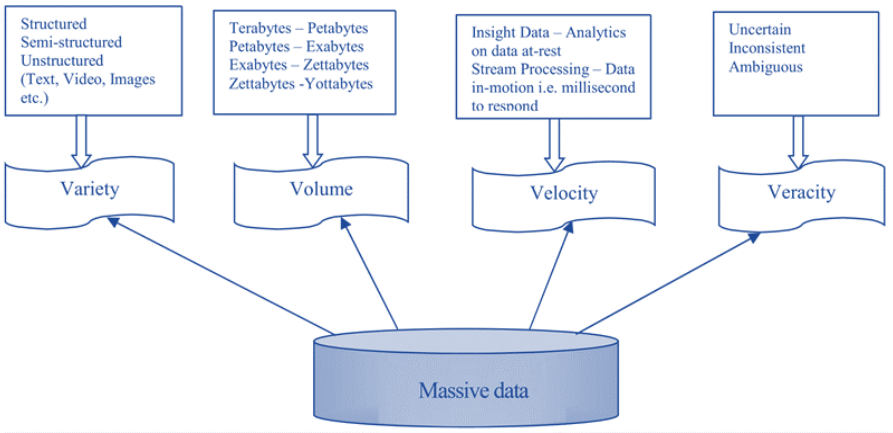

Figure 1. Characteristics of massive data

The amount of data is huge. Jump from TB to PB and EB levels. So far, that's the amount of data produced by 200pb printed material, and the amount of 5eb data in all the history of human history.

Data type range. This diversity also allows data to be divided into structured and unstructured data. Compared with the previous text-based structured data storage, more and more unstructured data have been created to bring challenges to all manufacturers. Due to the rapid development of Internet and communication technology, in recent years, due to the fact that data type is not single text in recent years, besides data types such as web logs, audio, video, pictures and location information, data processing capability is higher.

Low value density. The level of the value density is inversely proportional to the amount of data. In video, for example, an hour of video, in an uninterrupted process of monitoring, it may be useful data, only twelve seconds. How to complete the more powerful machine data "purification" at a faster speed is a solution to the problem of large data in the current turbulence background.

Fast processing speed. This is the most significant feature that is different from the traditional data mining. In the face of large amounts of data, the efficiency of data processing is their life.

3. Hadoop cloud platform architecture structure

Hadoop is a distributed system infrastructure, users can be distributed without knowing the underlying details of the development of distributed applications, take advantage of the cluster of high-speed computing power and storage [6]. Hadoop includes a plurality of sub-projects, but mainly by the Distributed Storage (HDFS), Distributed Computing (Map Reduce) composed of two basic parts, the typical basic deployment architecture shown in Figure 2.

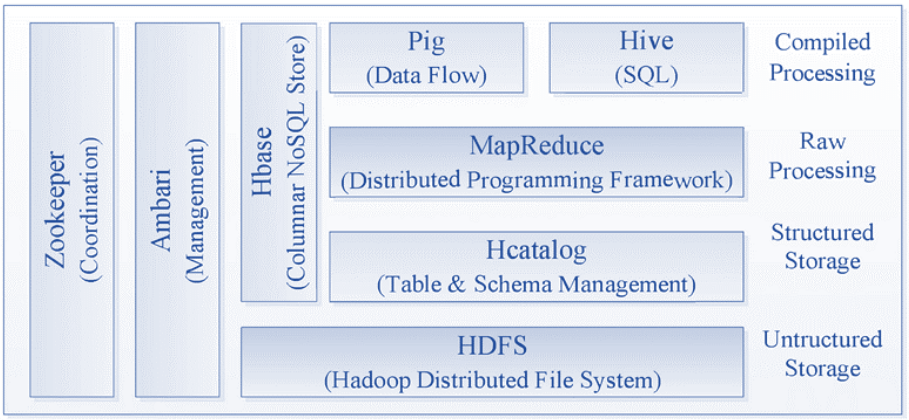

Figure 2. Hadoop cloud platform architecture structure

Distributed File System HDFS.HDFS (Hadoop Distributed File System) is a project

development for the Hadoop distributed file system, which uses master / slave architecture. HDFS by an Narne Node (document indexing server) and numerous Data Node (data nodes). HDFS provides users with the appropriate file name space for user data in the form of file storage. HDFS general will file these documents cut into several pieces, sliced file block will be stored on a data server. Then provided opened by Name Node, close, and rename files and directories basic functions, is also responsible for the file block is mapped to the Data Node. Then by the Data Node responsible for responding to client specific file reads and writes, while handling the creation initiated by the Name Node, delete, and requests the backup data block.

Parallel computing architecture Map Reduce. Map Reduce is a computer designed for multiple parallel processing of large amounts of data parallel computing framework. Input data Map Reduce job usually divided into separate blocks of data, data divided by a plurality of generally parallel processing tasks Map. Mapper removed from the HDFS data in the local hard disk, Reducer further calculations storing process will result in the local hard disk made by Network Mapper output or outputs the result to the HDFS. Map Reduce framework focuses scheduled task, and the task of monitoring the status of implementation, if fails, will re-execute the task. Compute nodes and storage nodes are usually together, which means that the same node and Map Reduce used in Hadoop HDFS used in. This makes Map Reduce framework can be distributed according to the stored data. Situation to schedule tasks. Map Reduce framework includes a separate master server Job Tracker (work distribution server) and a group of mounted together with the Data Node from the server Task Tracker (task execution server). The master server is responsible for scheduling from the server to, and monitoring tasks, re-execute the failed task.

4. Hadoop-based mass data processing

Massive data analysis system of internal business processing logic is basically the same, are the user sends a request, the system processing business logic and returns the results to the client show. The following software engineering UML class diagram and a message sequence chart for massive data analysis system, the core data query, for example, based on the design of the entire Hadoop system will be described, massive data analysis system class diagram shown in Figure 3.

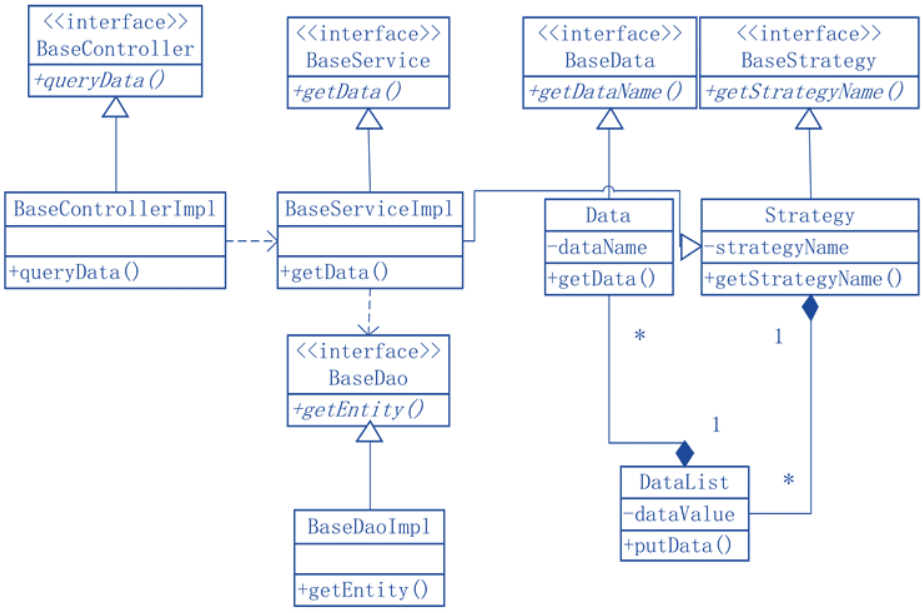

Figure 3. Hadoop-based mass data analysis class diagram

In the massive data analysis system class diagram design, Base Service for all business processing logic unit interface, Base Service Imp1 for all actual business logic of the parent class implements Base Service method, such as query data on Core Business It can be defined as Query Data Service, which inherited the Base Service Imp1 class, and has its own core method of querying the core data were independently of this business expansion, so the entire service layer reusability

greatly enhanced. Base Controller interface to all of the requests mapping service interface, the task is to deal with all of this interface by the user pages http requests sent to the server, the request is forwarded to the business logic layer in accordance with certain rules, until after the business logic layer processing is complete, calls show view, the results are presented to the user. Base Controller Imp1 implements the Base Controller Interface, Base Dao interface to all of the business logic objects interface entity, the task of this interface is that entity objects encapsulate business logic by these entities access to database objects and core data file, this interface is interactive safeguard business logic and real data source. Base Dao Imp1 Base Dao implements an interface, the parent class of business logic objects to interact with the database so that you can put all the interaction terms are instantiated, more convenient object-oriented programming. Hadoop-based mass data processing is shown in Figure 4.

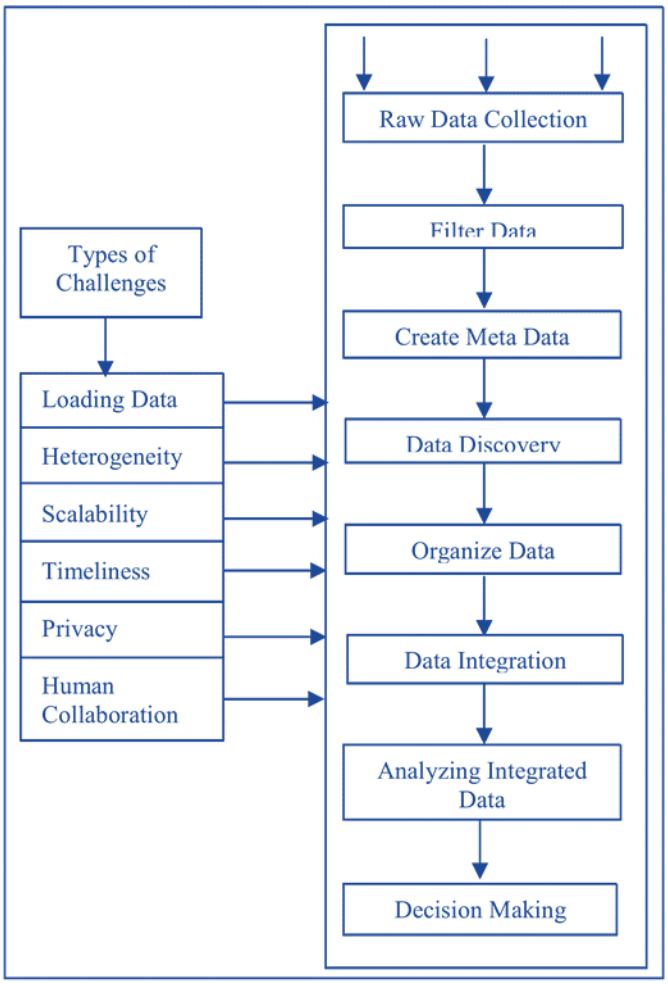

Figure 4. Hadoop-based mass data processing

5. Conclusion

Starting from personalized network development and massive data generation, this paper analyzes existing massive data processing schemes, and puts forward methods and Countermeasures for handling personalized mass data processing network combined with distributed computing theory. First, the theory of distributed computing and distributed platform is discussed and analyzed, and then the basic platform of Hadoop is chosen as the research project. After in-depth research and analysis of the Hadoop cloud platform, we try to build and deploy a verifying Hadoop cloud platform in the laboratory. Next, combined with the theory of distributed computing discussed earlier, we studied the massive mass Web log data preprocessing model under Hadoop cloud platform, and gave performance improvement plan and analysis. After that, the improvement and performance analysis are carried out on the Hadoop platform, and the distributed

data mining algorithm based on Apriori algorithm is studied and discussed, and the results are given. Finally, the whole content of the key technology research of mass data processing based on Hadoop is completed.

References

[1] Dittrich J, Quiané-Ruiz J A. Efficient big data processing in Hadoop MapReduce[J]. Proceedings of the VLDB Endowment, 2012, 5(12): 2014-2015.

[2] Zikopoulos P, Eaton C. Understanding big data: Analytics for enterprise class hadoop and streaming data[M]. McGraw-Hill Osborne Media, 2011.

[3] Lee K H, Lee Y J, Choi H, et al. Parallel data processing with MapReduce: a survey[J]. AcM sIGMoD Record, 2012, 40(4): 11-20.

[4] Tan H, Luo W, Ni L M. Clost: a hadoop-based storage system for big spatio-temporal data analytics[C]//Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, 2012: 2139-2143.

[5] Wang L, Tao J, Ranjan R, et al. G-Hadoop: MapReduce across distributed data centers for data-intensive computing[J]. Future Generation Computer Systems, 2013, 29(3): 739-750.

[6] Shang W, Jiang Z M, Hemmati H, et al. Assisting developers of big data analytics applications when deploying on hadoop clouds[C]//Proceedings of the 2013 International Conference on Software Engineering. IEEE Press, 2013: 402-411.