The Research of Massive Data Analysis and Processing Based on Hadoop

Julan YI[1, a]

[1]XINYU UNIVERSITY, Xin Yu 338004, China

[a]julanyi@126.com

Keywords: Hadoop; Massive Data; Data Processing

Abstract. how to quickly extracted from these massive data out of the enterprise value of useful information has become the most vexing problems in the development of application software programmers encounter in the course. Based on the starting point of this issue, this paper analyzes the key technical foundation and other existing distributed storage and computing on the combination of Hadoop cluster technology research as well as their business needs and the actual hardware and software strength, we propose a massive Hadoop-based data processing model and data structure design in several ways, the program process organization and use programming techniques and other methods to introduce the development of the model, and finally applied to model the log data preprocessing large site.

Introduction

With the rapid development of computer technology and Internet technology, vast amounts of data are continuously generated an urgent need for businesses to change their traditional architecture, in the face of massive data, how to analyze these data and how to effectively use the value of the data. At the same time, how to optimize their business has become a modern enterprise transformation process inevitable question. The amount of data is only one aspect of the massive data challenges of massive data, referring to the other two aspects of speed and diversity. Speed indicates the response speed requirements gathering, processing and data query data. The diversity refers to the format and content of the ever-changing data [1-2]. In the massive data processing, how to efficiently and quickly dig out from massive amounts of data and the potential value of conversion capacity for decision-making is based, will become the core competitiveness of enterprises. The importance of data analysis no doubt, but with the speed of data generated faster and faster, the amount of data increases, the challenges encountered in data processing technology is also growing. How to dig out from the mass of data useful value, analyze the deeper meaning, and then transformed into actionable information, it has become a problem all Internet companies have to deal with.

Carried out on the cloud platform of massive data processing model and algorithm parallelization, distributed research, with sufficient research value. In this paper, as the basic platform Hadoop cloud platform research, conducted massive Web log data preprocessing models on top of and research Mining Based Apriori distributed data to performance mass data processing technologies effectively enhance the cloud platform of hope able to promote the development of massive data processing techniques make whatever contribution.

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..

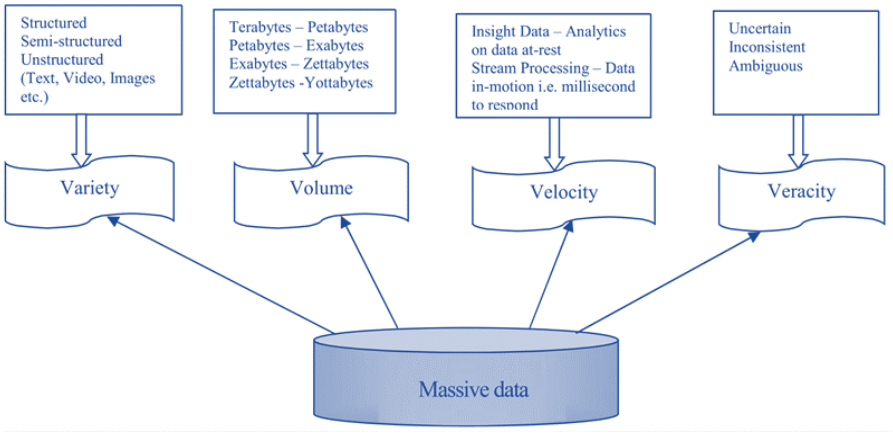

Figure 1. Characteristics of massive data

Data volume is huge. Level jumped from TB to PB and the EB level. So far, the amount of data that human production of all printed materials are 200PB, and the amount of data for all of the words in the history of mankind have said about 5EB.

Data type range. This type of diversity also allows data to be divided into structured data and unstructured data. Compared with the previous text-based for easy storage of structured data, produce more and more unstructured data to all manufacturers have posed a challenge. Thanks to the Internet and the rapid development of communications technology in recent years, thanks to today's data type had not a single text, in addition to web logs, audio, video, pictures, location information, and so many types of data to the data processing capacity of a more high requirements.

Low value density. Value density level and inversely proportional to the amount of data. In the video, for example, a one hour video, in uninterrupted monitoring process, it may be useful data with only twelve seconds. Value by how quickly complete algorithm is more powerful machine data "purification" is currently under turbulent background big data problem solved.

Fast processing speed. This is a big distinction between data in traditional data mining the most significant features. In front of the vast amounts of data, the efficiency of data processing is their life.

Hadoop cloud platform architecture structure

Hadoop is a distributed system infrastructure, users can 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.

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.

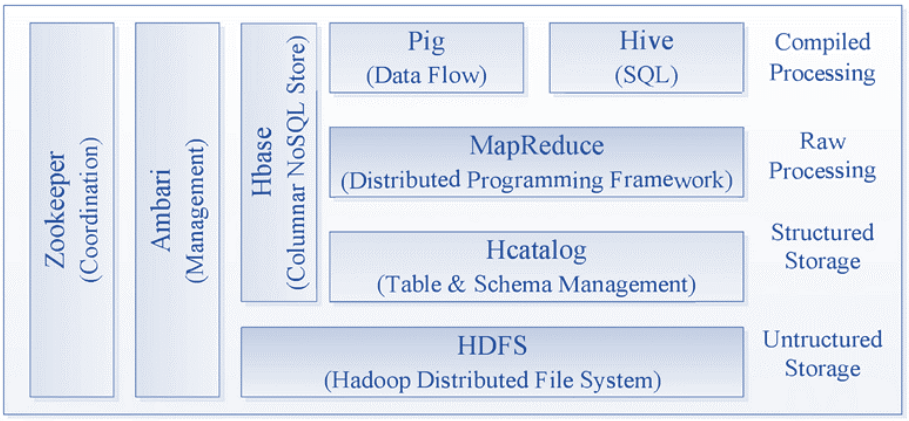

Figure 2. Hadoop cloud platform architecture structure

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.

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.

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.

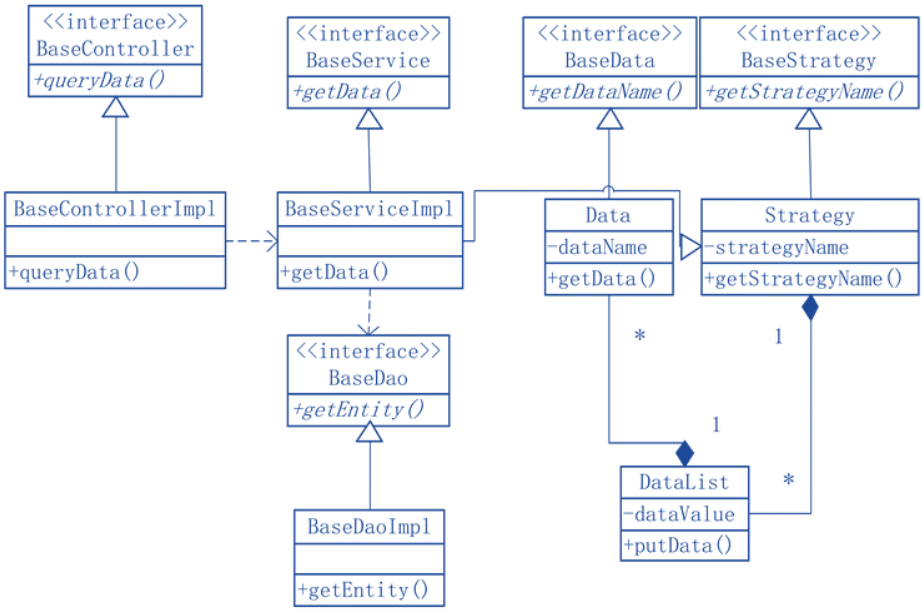

Figure 3. Hadoop-based mass data analysis class diagram

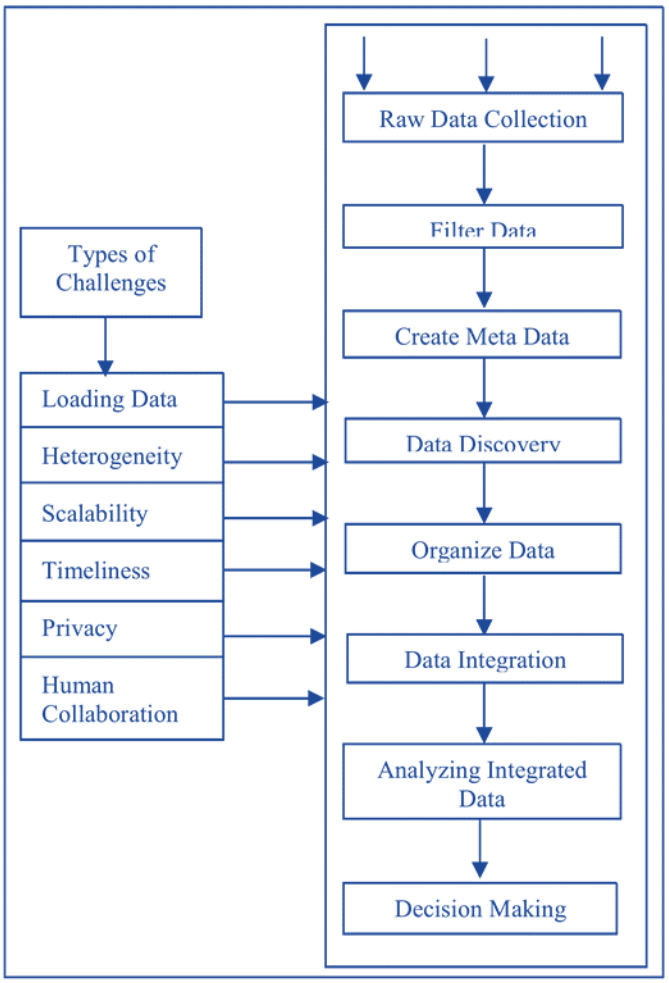

Figure 4. Hadoop-based mass data processing

Conclusion

Development and produce huge amounts of data in this article starting from personalized network, research and analysis of the existing mass data processing solutions, combined with distributed computing theory proposed method and countermeasures to deal with the personalized

massive data-processing network. Firstly, the theory of distributed computing and distributed platform for discussion and analysis, and then select the basic platform Hadoop as a research project. After Hadoop cloud platform for in-depth research and analysis in a laboratory to try to build and deploy confirmatory Hadoop cloud platform. Next, the combined pre-discussed theory of distributed computing, massive data Massive Web log data preprocessing models under study Hadoop cloud platform, and gives the performance improvement program and analyzed. After performed at Hadoop platform improvements and performance analysis of distributed data mining algorithms based Apriori research and discussion, and gives. Based on the final completion of the entire contents of Hadoop massive data processing key technology research.

Reference

[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.