Construction of Big Data Mining Platform Based on Cloud Computing

Mali SUN

Anhui Xinhua University, Hefei Anhui, China

Keywords: Cloud computing, Data mining, Key technologies.

Abstract. Cloud computing is the product of the Internet era, which can realize dynamic resources disposition, customized services for demand, and centered on the Internet, render service measurable and resources transparent. It plays a big role in the in data mining field of various industries in today's society. Building the data mining platform based on cloud computing is helpful to achieve efficient data mining. This paper analyzes the basic architecture of the big data mining platform based on cloud computing and the key technologies for its construction on the basis of relevant theories of cloud computing and data mining.

Introduction

With the advent of the cloud era and the rapid development of mobile Internet, China has entered the information age with too big amount of information. According to a survey, By the end of 2014, the amount of data has exceeded 3 ZB worldwide. Such a large amount of data has brought problems to the implementation of data mining system, making large data processing complicated. The computing power of the system cannot meet the requirements, nor can the computing resources of traditional stand-alone servers. Thus, it is necessary to use distributed computing technology for mass calculation. The emergence of cloud computing makes the big data mining platform have a new development direction and makes its construction possible. However, the big data mining platform based on cloud computing has not yet been built to perfection, which requires continuous scientific and technological research on building a new data mining system.

Cloud Computing and Data Mining

Data Mining. Also known as data or knowledge discovery, data mining refers to searching out potential and valuable data from a large number of fuzzy and random actual data by computing. Data mining is closely related to computer technology, which is realized through statistics, on-line analytical processing, information retrieval, machine learning, expert system and pattern recognition. Data mining is an important technology in the field of knowledge discovery, specific techniques of which mainly include collection, extraction, warehousing, analysis and statistics; it is widely used in areas such as internet, finance, telecommunications and scientific research at present.

Cloud Computing. Cloud computing is a computing method based on the Internet, which shares software and hardware resources and information to computers and other equipment. Cloud refers to network, in particular the Internet. In using cloud computing, the user does not need to know the details of the infrastructure in the "cloud", nor does he/she need to have corresponding professional knowledge and direct control over the whole computing process. Cloud computing mainly has the following characteristics. Firstly, it achieves dynamic resource distribution, different resources division according to the needs of the user and increase of available resources. Secondly, it realizes customized services, that is, provide users with self-support resource service, who do not need to

interact with the suppliers. Thirdly, it is centered on network, through which it provides users with services. Fourthly, services can be measurable and optimized for the user while the use of resources can be controlled. Fifthly, resources provided for the user are transparent so that the user does not have to understand its internal structure.

Architecture of Big Data Mining Platform Based on Cloud Computing

The development of network cloud brings new problems and challenges as well as a new direction of development to data mining. The big data mining platform based on cloud computing is helpful to solve the problems pertinent to traditional data mining technology such as low efficiency, backward function, postponement and lag of information and high cost. Cloud computing belongs to a commercial calculation model with the combination of network computing, parallel computing and distributed computing, the power of which realizes great efficiency of big data mining. With the realization of standardization and normalization of the SaaS function of cloud computing, big data mining based on cloud computing of SaaS is gradually understood and put into application. This article constructs a SaaS platform of big data mining from three perspectives, namely, service of big data mining based on cloud computing, parallelization of data mining algorithm and componentization of data mining algorithm.

The overall structure of the big data mining platform based on cloud computing is shown as in Fig. 1. The bottom level of the structure is supported by cloud computing, adopting cloud computing to provide distributed storage and computing capacity for data mining performance. The design of the data mining platform lies in the middle. Finally the top level data mining capability is accessed through a third party algorithm capability, and then so exposed that it can be called according to the need of business system.

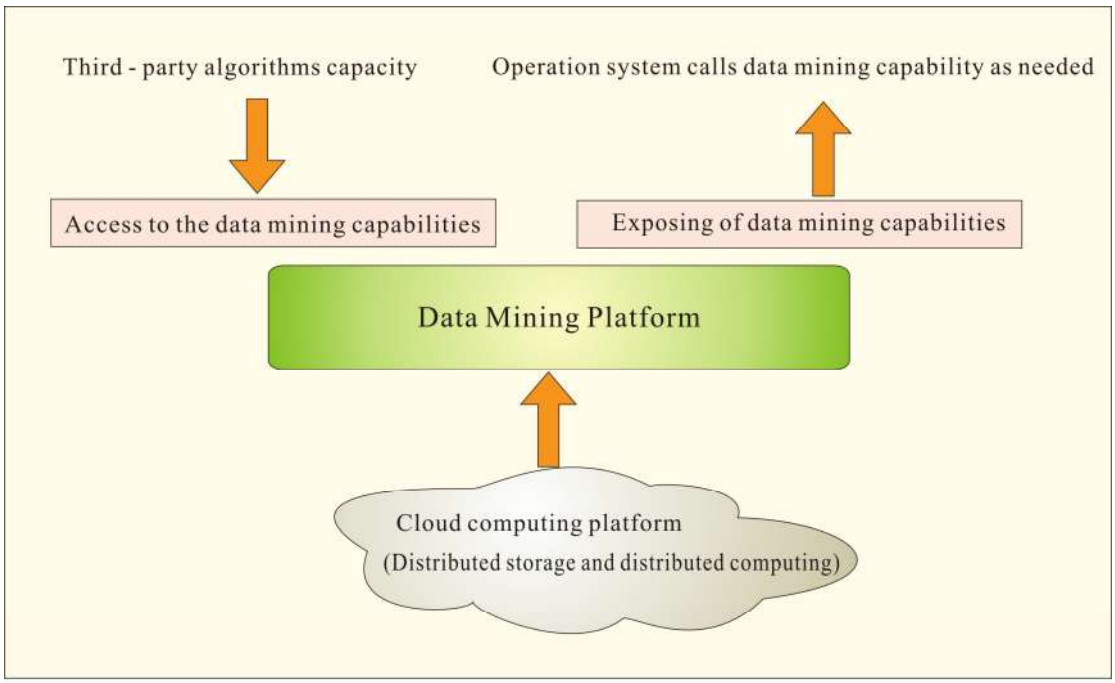

Figure 1 Architecture of the Big Data Mining Platform Based on Cloud Computing

The traditional architecture of data mining technology is built on the basis of relational database, which is unable to satisfy the computing of huge amount of data. Cloud computing using its distributed storage and computing forms constructs a new type of cloud computing data mining platform, as shown in Fig. 2, which mainly includes three layers from the bottom to the top: the third layer of cloud computing support platform, the second data mining ability layer, and the first

layer of cloud service of data mining.

Firstly, the cloud computing support platform can be created in a quick, simple and extensible way to manage large and complex IT infrastructure, which is mainly composed of cloud platform mainly to store data, cloud computing platform given priority to data processing and comprehensive cloud computing platform dealing with data storage and processing at the same time.

Secondly, the layer of data mining capability mainly provides infrastructural power of data mining, the main function of which is parallelizing data processing algorithms, dispatching service management framework, providing internal system data mining processing, recommending algorithmic library and enabling the third party data mining algorithm into the layer. This layer is the basis of mining providing and the core of the whole data mining system.

Thirdly, the main function of the cloud service layer of data mining is to provide cloud service and relevant engines for language and statement access so as to facilitate automatic use of cloud service. The interface forms service capability encapsulation are diversified, mainly including web service under simple object access protocol, scalable markup language (XML), hypertext transfer protocol (HTTP) and local application programming interface (API). According to different situations, each business system of the cloud service layer can restructure and call data mining cloud service.

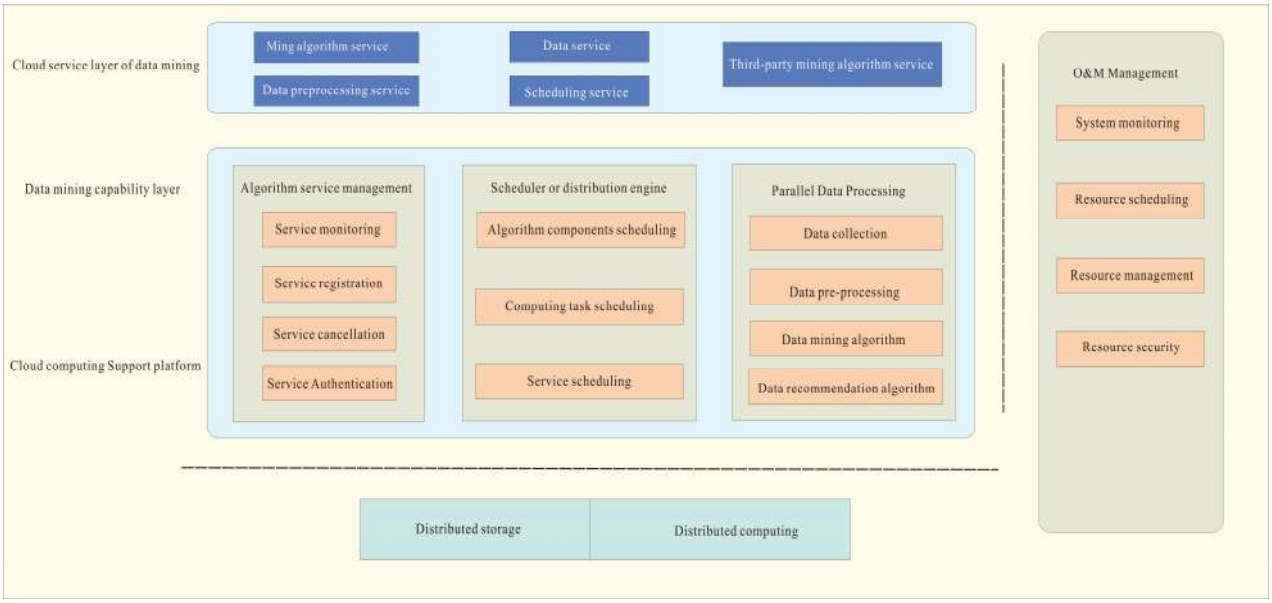

Figure 2 Framework of Data Mining Platform Based on Cloud Computing

Key Technologies to Constructing Big Data Mining Platform Based on Cloud Computing

The construction of big data mining platform based on cloud computing is dependent on the support from advanced science and technology, for which the key technologies needed are listed as follows.

Cloud Computing Technology. The first is distributed storage technology, which takes use of disk space on the computer through the network and constitutes a virtual storage device by scattered resources, thus achieving data storage in a distributed manner. It embodies the reliability and economic efficiency of cloud computing in data, processing and helps to provide the user with an inexpensive and excellent way of data mining.

The second is virtualization technology, which refers to computer components operating on the basis of the virtual environment. It helps to expand the capacity of the hardware and to simplify the software reconfiguration process as well allows running multiple operating systems on a single

platform, wherein the programs are independent of each other and avoid mutual influence, eventually significantly improving the efficiency of the computer.

The third is parallel cloud computing technology, which is conducive to effective implementation of data mining technology and to encapsulate the details of cloud computing such as task parallelism, task scheduling, task fault tolerance, system fault tolerance or data distribution, etc. The user does not need to care about these details so as to improve the development efficiency.

Data Collection Control Center. Its function is collecting different types of data, that is, completing the collection work of all the business data that have been accessed to the cloud computing data mining platform, and to solve the contradictions pertinent to relevant provisions and protocols between different data, making them adapt to various source data formats.

Service Scheduling and Management Technology. In the big data mining platform based on cloud computing, corresponding service scheduling and management technology is so important that it enables different business systems to use the computing platform. Service scheduling helps to ensure safety and reliability of cloud service while management technology unifies the functions of service exposing and registration. The latter makes it feasible for the third party data mining and is conducive to expanding the influence of the service platform.

Mining Algorithm Parallelization Technology. Data mining algorithm parallelization under cloud computing efficient utilize the basic capabilities the cloud computing platform provides, mainly including algorithm parallelism, parallel strategy selecting, etc. Parallelization of data mining algorithms of is conducive to make full use of the resources of each workstation and to implement unified scheduling and coordinate processing, thus finally achieving efficient parallel computing.

Summary

Nowadays, as Internet technology is developing rapidly and information amount is increasing unceasingly, it is of great importance construct the big data platform based on cloud computing. The platform helps to achieving high performance, reliability and economy in data mining. This paper starts to construct the big data mining platform from three angles of servitization of cloud computing of large data mining, the parallelization of data mining algorithms, the componentization data of mining algorithms, and analyzes three core cloud computing technologies pertinent to data collection control center, service scheduling and management and mining algorithm parallelization. It attempts to provide reference for the construction of big data mining platform.

Author in brief

Mali SUN, born in Hefei, Anhui province in 1981, master, lecturer, whose main research direction is data mining.

References

[1] J. Ding, S. L. Yang, H. Luo and S. Ding, Data mining service model in cloud computing environment, Computer Science. 6A, 39 (2012) 217-219+237.

[2] X. Y. Wang, On cloud computing, Journal of Taiyuan University. 3 (2012) 135-137.

[3] Y. Gao, T. Nie and Y. Mao, Research on principles and implementations of cloud computing, Computer CD Software and Applications. 16 (2014) 105-106.

[4] Y. Shen, The research of high efficient data mining algorithms for massive data sets, PhD. Dissertation, Jiangsu University, 2013.