LUPINE PUBLISHERS

Current Trends in Computer Sciences & Applications

DOI: 10.32474/CTCSA.2023.03.000153

ISSN: 2643-6744

Research Article

Research on Data Mining Solutions Based on Cloud Computing Technology

Yubo Wang*

Innovation Research Institute, CETC Taiji Computer Corp, China

***Corresponding author:** Yubo Wang, CETC Taiji Computer Corp, Innovation Research Institute, China

Received: July 20, 2023 **Published:** August 02, 2023

Abstract

The article based on the traditional data mining with large data mining perspective, probes into the connotation of the large data mining, puts forward the cloud computing service integration of data mining and mining system frame work, and to merge the multi-function large Hadoop data mining platform as an example, the analysis of large data mining internal work processes and analyzes the advantages and challenges, So as to provide a reference for users' cognition and application needs of big data mining.

Keywords: Big data; Big data mining; Cloud computing; System architecture

Introduction

With the continuous development of the Internet, the Internet of Things, cloud computing and the popularity of intelligent terminals, massive complex and diverse data show an explosive growth, prompting the arrival of the era of big data. As an important factor of production, big database become a strategic asset with huge potential value, promoting industrial upgrading and rise, and influencing therefrom of scientific thinking and research methods. However, big data, relying on its rich resource reserve and powerful computing technology, also brings challenges. Massive, dynamic, and uncertain data make traditional data processing systems face storage and computing bottlenecks. At the same time, traditional data mining technology can no longer meet the needs of users because of its limited function on how to quickly and real-time mining valuable information and knowledge from complex big data. Therefore, in the context of big data, an applicable technology, namely big data mining, is needed to deal with the current challenges. In view of the big data environment, the mining system. Developed with reference to the traditional data mining construction ideas cannot provide users with satisfactory services, so the architecture of big data mining is indispensable to meet the construction and application requirements. However, at present, there are few models for reference, and some departments put forward big data mining solutions according to business requirements, because the system is not portable and the integration of internal components is poor, its application has obstacles. This paper discusses big data mining by comparing with traditional data mining, proposes the architecture of big data mining based on cloud computing and builds a specific big data mining system, and objectively evaluates the advantages and disadvantages of big data mining by taking the workflow as the main line, so as to provide reference methods for promoting its application and development.

Big Data Mining

Big data mining is to mine information and knowledge with huge potential value from big data with huge volume, diverse types, dynamic and rapid flow and low value density, and provide it to users in the form of services. Compared with traditional data mining, it also aims at mining valuable information and knowledge. However, in terms of the background of technological development, the data environment and the breadth and depth of mining, they are different.

Development Background

Both have evolved because of advances in technology, the mass production of data, and the need for valuable data. However, in the technology of advanced degree and the volume of data, analysis of variety complexity and processing ability, the traditional data

mining without big data era has enriched environment technical conditions, in the database, data warehouse, and under the background of development of the Internet, has realized from the independence, transverse to the longitudinal development of data mining. Big data mining, on the other hand, benefits from the emergence and development of cloud computing, Internet of things, mobile intelligent terminals and other technologies under the background of big data. Aiming at the characteristics of big data and the problems faced by existing mining systems, it is a system antically integrated and improved with the help of advanced technologies. Compared with the traditional data mining has been quite mature application, algorithm research and system tool development, the research and application of its technology is still in continuous development, for massive data mining is mainly based on cloud computing for the integration of related technologies to achieve.

Processing Objects

Table 1: Comparison of the characteristics of objects processed by traditional data mining and big data mining.

category	segmentation		Represents Bigtable, HBase
Transtiond the database	NOSQL system	The key value system	Dynamo, Cassandra etc.
		Documnt storage system	Mongo, Couchbase,etc
		Figure database	Secondary, OrientDB,etc
	NEWSQL system	General purpose database	Spanner, NuoDB,etc
		Memory based number According to the library	SQLFire, VoltDB,etc
Analytical the database	Database based on MapReduce		Hive, Pig
	Hadoop based database		HAWQ, Impala Hadapt

Due to the different data environment between big data mining and traditional data mining, there are differences in the characteristics of the processing objects. The data source of traditional data mining is mainly the passive data generated in a specific range of management information system, plus a few active data generated by users in Web information system. The type of data is mainly structured data, plus a small amount of semi-structured or unstructured data. Besides management information systems and Web information systems, the data sources of big data mining also include simulation data automatically generated by sensing devices such as sensing information systems. Compared with traditional data mining, big data mining has wider data sources, huge volume and more complex types. Accordingly, the collection mode is no longer limited to passive, the collection range is more comprehensive, the throughput is high, and the processing is real-time and fast. However, due to the low requirements on the accuracy of the data, the redundancy and uncertainty of the data are high, as shown in Table 1.

Big Data Mining Architecture Based on Cloud computing.

The traditional data mining system is usually run on a single machine or client/server, and the architecture is usually a two-layer structure of client/server or three-layer structure of Web browser/server. Its system structure is roughly divided into the data source, data storage, analysis, mining front show 4-layer, processing mainly with the method of data to calculate, the pretreatment of data loaded into the data warehouse, data mart store, focus on analysis of mining on the server, and migration will eventually result in interactive way is presented to the user. However, when dealing with massive distributed and dynamically heterogeneous big data, this centralized batch processing mode of first storage and then processing undoubtedly increases the time, space complexity and transmission cost. In addition, there are also the following problems:

a) There are obstacles in the expansion of traditional mining systems. In heterogeneous environment, the computing power of cluster is poor, and the storage is easily limited by the size and type of data, but the cost of vertical expansion is huge.

b) The mining effect cannot meet the expected requirements. Traditional analysis tools and mining algorithms are not portable and scalable for multi-dimensional and complex big data [1]. which leads to low-quality and inefficient mining results due to insufficient analysis. For example, traditional clustering algorithms process high-dimensional data at the cost of original data loss, low-quality clustering results and high time complexity [2].

c) The effect of user interaction experience is not good. The time-consuming and cumbersome manual sorting in the pre-processing stage and the passive and non-intelligent cognitive process of user needs are consistent with the standards of simplicity, quickness, intelligence and real-time required by users.

Big Data Mining Platform Based on Hadoop

In order to facilitate the specific analysis of big data mining, this paper builds a multi-functional Hadoop big data mining platform to understand each processing link. The multi-functional big data mining based on Hadoop platform is divided into three layers: data source, big data mining platform and user layer. Data source is a complex processing object formed by structured, semi-structured

and unstructured data. The big data mining platform is based on Hadoop to integrate various computing modes, analysis, mining and other functions, and combine the characteristics of real-time data for corresponding processing; In the user layer, system cognition and service acceptance are carried out in an interactive way, as shown in Figure 1.

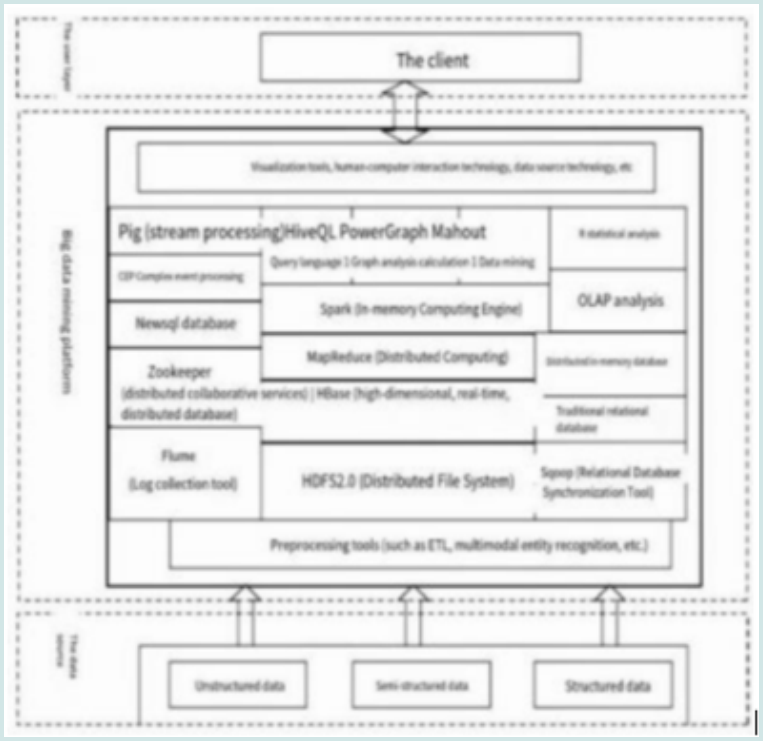

Figure 1: Integrated multi-functional big data mining based on Hadoop platform.

Data Preprocessing

In traditional data mining, there is a pattern before data. Through the established pattern, tools such as ETL and DB Put and driving methods such as query and update are used to preprocess static data [3]. It pays attention to maintaining the integrity and accuracy of data, and the processed data is of high quality. Big data mining comes after data. There is a pattern, which is not to determine a certain pattern in advance, but the uncertainty model changes continuously with the change of data [4]. Big data preprocessing is mainly based on Map Reduce integrated with traditional preprocessing technology, real-time DataStream processing, multi-modal entity recognition, Deep Web integration and remote automatic acquisition and fusion technology, to improve the ability of parallel computing, iterative computing, data merging and sharing in the process of preprocessing. For example, the use of Flume or Sqoop's flow computing technology and embedded middleware multi-level data processing technology for data transmission and migration, to achieve Historical data and data streams are processed synchronously to improve the efficiency of real-time data processing. However, since the processing of big data focuses more on the correlation between data rather than the causal connection and focuses more on the real-time processing of data rather than the integrity and accuracy, and focuses on data rather than model, the quality of processed big data is not good [5], and the accuracy and credibility of mining results are not high. For example, in 2013, the prediction of influenza by Google using big data showed a high error rate [6].

Data Storage

The storage management of traditional data mining is based on data warehouse and operation database. The relational database system, such as unified system and file system, mainly uses row storage to store static and definite structured data in E-R (entity and relation) or multidimensional data model. The storage is passive, and the access mode is random. The specific mode is generally defined internally by the system, and the flexibility and scalability are poor. Storage with high requirements on ACID (Atomicity, Consistency, Isolation, and Durable-Y) and low fault tolerance but big data mining. In addition to traditional data storage, it also includes distributed storage, which can store structured, semi-structured and unstructured data. The storage strategy is mainly column storage or mixed column storage, and the mode is generally implemented externally [7]. In general, it does not support ACID but supports BASE (Basically available, Soft State, Eventually Consistent), and it supports limited functions compared with relational databases [8]. For example, Google build synopsis data structure, realizes the dynamic and uncertain data directly storage processing. The developed big table uses column storage and a new data model Ordered Table to store data. The schema is flexible

and simple, and has strong scalability, but there are problems in data consistency and compatibility with relational data model. The Spanner system supports synchronous cross-center replication and visual sharing and provides SQL user interfaces to effectively achieve high scale ability and ACID integration. In addition, for uncertain data, large data storage has corresponding uncertain data base management systems, uncertain data lineage management technology, etc. [9], the data to model uncertainty relation storage, storage way and strict order directly, and can be based on a memory disk rather than build synopsis data structure, realizes the dynamic and uncertain data directly storage processing Table 2.

Table 2: Distributed database system.

Contrast	Traditional data mining	Big Data Mining
The data source	Less	Numerous,extensive
Colletion scope	Local sampling	Global access to
Acquisiton methods	Passive is given priority to	Active and automatic
The data type	Relatively simple to structure Focus on data	Complex and diverse, with semi-structure and non-structure
Data redundancy	LOW	High
Total data volume	Measured in TB	Mass, measured in EB or TB
Processing unit	The unit is MB	PB is not very
Accuracy of data	Demand is high	demanding
Processing efficiency	A longer time	Real time and fast

Data calculation and analysis

Compared with the centralized batch processing mode in traditional data mining where data is moved to computing, big data mining uses the integration of multiple computing modes to process big data in parallel. For a small number of static data with fewer dimensions, traditional data mining shows high query and analysis performance due to repeated times, accurate query methods, strong flexibility and fast processing and analysis ability of OLAP [10]. However, in the face of massive data with various dimensional attributes and huge data cubes, traditional OLAP cannot automatically and deeply analyze [1], and the query language mainly based on SQL is difficult to express the complex analysis model to be constructed, so the quality and efficiency of its query analysis will be seriously affected. However, aiming at the problems of poor scale ability of traditional analysis tools and weak analysis function of the existing cloud platform, big data mining integrates system functions [4-5] to improve the distributed parallel computing ability of the original analysis mining and the analysis ability of the supporting platform. Figure 1 shows the deep integration of R analysis software and Hadoop, and the integration and improvement of traditional mining algorithms and existing algorithms based on Hadoop. For dynamic graph data, memory distributed data management system can support query processing with low latency. For the data stream, the sliding window model-oriented method is used to process the data stream directly through the probability dimension index.

Conclusion

The emergence of big data not only brings rich and diverse resources with potential value, but also revolutionizes both the traditional data management mode and the scientific way of thinking. In the face of massive, complex and uncertain dynamic data, traditional data processing methods are facing severe challenges in both computing and storage capabilities, and their scalability and flexibility cannot meet the requirements of real-time processing of big data. Cloud computing provides powerful computing and storage power for the processing of big data, and big data mining provides an opportunity for the deep integration of big data and cloud computing. This paper uses the method of comparing big data mining with traditional data mining, discusses the connotation, architecture and the technology and method in the workflow of big data mining, and objectively analyzes the advantages and disadvantages of each. In general, traditional data mining is suitable for centralized batch processing of static and small amount of structured data samples and has high query analysis performance and mining efficiency. Its calculation mode and storage type are single, and it pays more attention to data quality and requires higher data accuracy. However, in the face of large-scale, rapid flow of uncertain data, traditional data mining the smaller storage capacity, throughput, and to improve the efficiency of computing the internal managing strategies cannot meet the requirements of processing, at the same time, system extensibility, flexibility is not strong and complicated and complicated mining algorithms do not have scalability also affects the full play of function analysis of mining. On the other hand, big data mining effectively makes up for the shortage of traditional data mining storage and computing. Based on cloud computing, it integrates a variety of computing and storage modes, which is suitable for real-time and rapid batch or stream processing of massive, big data with mixed multi-structures, and has high scalability, scalability and fault tolerance. Big data mining pays more attention to the processing speed of massive data and does not pay much attention to the accuracy of data. Its algorithm complexity is low, the quality of data is poor, and big data mining is not as flexible and efficient as traditional data mining when

dealing with a small amount of data. In addition, big data mining still has privacy security, sharing and other issues. At the same time, big data mining still faces challenges in realizing the integration of intelligent analysis mining, visualization and automatic mining, and creating human-computer interaction interfaces that are beneficial to user cognition. Therefore, the subsequent research work on the above issues needs to be further promoted and deepened to meet users' needs for intelligent, reliable, efficient and high-quality features of big data mining.

References

1. Han JW, Micheline K (2012) Data Mining concepts and techniques [M]. Fan Ming, Meng Xiaofeng, translated. Beijing, Machine Press. China.
2. Wang Yuan zhuo, JIN Xiao long, Cheng Xue qi, et al. (2013) Network big data: Current situation and prospect [J]. Chinese Journal of Computers 36(6): 1125-1138.
3. LI Jianzhong, Liu Xianmin (2013) An important aspect of big data: data availability [J]. Journal of Computer Research and Development 50(6): 1147-1162.
4. Lazer D, Kenndy R, King G, Vespignani A (2014) The para-ble of google flu: traps in big data analysis [J]. Science 343 (6176): 1203-1205.
5. Enrico B, Marco G, Mauro I (2014) Performance evaluation of NoSQL big-data applications using multi-formalism Model [J]. Future Generation Computer Systems 37(1): 345-353.
6. Shen Derong, Yu Ge, Wang Xite, et al. (2013) A survey of NoSQL system supporting big data management [J]. Journal of Software (8): 1786-1803.
7. Gao M, Jin C Q, Wang X L, et al. (2010) Review of data lineage management technology [J]. Chinese Journal of Computers 33(3): 373-389.
8. Zhao Bo, YE Xiaojun (2011) Research and Implementation of OLAP Performance Test Method [J]. Computer Research and Development 48(10): 1951-1959.
9. Pei Jian (2013) Some new progress in analyzing and mining uncertain and probabilistic data for big data analytics [J]. Lecture Notes in Computer Science 8170(1): 38-45.
10. Ding Y, Yang Q P, Qian Y M, et al. (2013) Architecture and key technologies of data mining platform based on cloud computing [J]. Zte Communications Technology 19(1): 53-60.

DOI: 10.32474/CTCSA.2023.03.000153