ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 156

7th International Conference on Management, Education and Information (MEICI 2017)

Innovation Model Analysis of Data Flow Calculation and Storage Technology based on Cloud Computing

Hongsheng Xu[1,2 a *], Ke Li[1,2] and Ganglong Fan[1,2]

[1]Luoyang Normal University, Luoyang, 471934, China

[2]Henan key Laboratory for Big Data Processing & Analytics of Electronic Commerce, Luoyang, 471934, China

[a]85660190@qq.com

Keywords: Cloud computing; Data flow calculation; Big data storage; Cloud storage; Data mining

Abstract. Common forms of cloud storage can be divided into distributed file systems and distributed database. Among them, the distributed file system uses large-scale distributed storage nodes to meet the demand of large files stored. The paper describes big data acquisition and processing key technology based on data mining. Big data is divided into offline batch calculation, real-time interaction and flow calculation. The paper presents innovation model analysis of data flow calculation and storage technology based on cloud computing.

Introduction

Based on the total volume, the accuracy of big data analysis will surpass the traditional mathematical statistics, and the causality will be replaced by the correlation, which is not so optimistic. On the one hand, traditional mathematical statistics are not outdated. Its four hundred years of development still play an important role in all aspects of economic society. For example, sampling is an ancient and mature statistical method. If the target is clear and the method is scientific, the conclusion is correct in most cases, and it is not inferior to the full data.

A common view is that Internet searches are the main cause of data diversity, and that is partly true. Data, however, the diversity of the increase is mainly due to the new data structure, and includes the web logs, social media, the Internet search, mobile phone call records and sensor network data types. Some of these sensors are installed on trains, cars and planes, each increasing the diversity of data.

The three concepts of data processing time shift: you all don't sampling, efficiency do not absolute precision and related do not cause and effect [1]. Specific large data processing method in fact there are many, but according to the long time practice, the author summarizes a basic big data processing flow, and the process should be able to help you straighten out the big data processing.

The whole process can be summarized as four steps, are gathering, import, and pretreatment, statistics and analysis, and mining.

Big data due to the volume is huge, is growing again at the same time, so the unit in lowering the value of the data density. But at the same time, large data in improving the overall value, big data by analogy for oil and gold, so it can discover the huge commercial value to find hidden patterns from huge amounts of data, to the depth of data mining and analysis [2]. Data mining and traditional data mining model also is put in bigger difference: general data volume small, the traditional data mining algorithm is relatively complex, slow convergence speed. Large data of the great amount of data, however, the data storage, cleaning, ETL (extract, transform, load) need to be able to cope with the demand of the large amount of data and the challenge, to a great extent, need to adopt the distributed parallel processing.

Transition from large-scale production to mass customization of the enterprise, must grasp the needs of users. In the Internet age, the demand characteristic is often in the act of user inadvertently revealed. Through the information association, reference, and it is methods such as clustering, classification, analysis, to get the answer. "Big data" between Internet and traditional enterprise establish an intersection. It promotes the Internet enterprise integration into the traditional supply chain, and in the traditional enterprise gene under Internet. The combination of the traditional enterprises and the Internet,

Internet users and consumers, is bound to lead to consumption patterns, manufacturing mode and management mode of great change.

Around the data and the end user, we observed the development of the computer industry has three directions: first application software will be suffused with the Internet. Second and it is industry vertical integration. Company, as near to the end user will have more say in the industrial chain. Third, the data will be assets. Suffused with the Internet is a important way of collecting data, without extensive application of Internet, the company is difficult to get the user's behavior data; Industry vertical integration trend in the data is applied, through collecting a large number of user data, closer to the user, a better understanding of users, to provide more appropriate services; Data become assets more emphasis on strategic significance. Put forward three trends; expand the research of the large data theme, opened up a new perspective and logic to observe software company growth path and investment value. The paper presents innovation model analysis of data flow calculation and storage technology based on cloud computing.

Big Data Acquisition and Processing Key Technology Based on Data Mining

We know that big data analysis technology initially originated from the Internet industry. Web archiving, the user clicks, the relationship between commodity information, the user data to form the sustained growth of huge amounts of data sets [3]. These large data contains a lot of can be used to enhance the user experience, improve service quality and develop new applications of knowledge, and how to efficiently and accurately find the knowledge is the basic determines the position of each big Internet companies in the fierce competition environment.

First of all, the technology, led by Google, the Internet company put forward the technical framework of graphs, using cheap PC server cluster, large-scale concurrent processing bulk transaction..

Big data is the data analysis of the cutting edge of technology. In short, from various types of data, the rapid ability to obtain valuable information is the big data technology. Big data can be divided into big data technologies, data engineering, big data science and data applications and other fields.

People talk about at present most is big data technology and data applications. Engineering and scientific problems has not been taken into account. Big data project refers to the operational management of planning and construction of the large data of system engineering; Big data science focused on discovery and validation data network development and the process of operation law of big data and the relationship between the natural and social activities, as is shown by equation(1).

$$q_{ii} = \lim_{h \to 0^+} \frac{p_{ii}(h) - 1}{h} = -(\lambda_i + u_i) \tag{1}$$

The overall situation and development trend of large data mainly reflected in several aspects: big data and academic, big data and human activities, large data security and privacy, the key application, the influence of system processing, and the whole industry [4].

Big data on the overall situation, the data will become bigger, the size of the data resource, highlighted the value of the data, data Shared privatization and alliance.

The efficiency problem is more complicated. Because the level of database model may be a similar most SOA to complete the process of information service bus, an important step is to ensure that related to the arrangement of overhead lines in a minimum degree. This can help to reduce the data access overhead associated with SOA, but it can't overcome the problem of storage system itself.

Because this storage system has been through out of SOA component level model, it is easy to neglect the problem related to the amount of delay and data transmission, in particular, if the database is the distribution of the cloud, so use them can produce variable network latency.

Analysis and the market segment, the be fond of according to individual or group, or offers rich personalized products, consumer behavior, for example, the Marketing Department can collect some valuable information, to find a shopper's interest, and then targeted to organize some marketing activities, thereby increasing the enterprise in the competition advantage.

Big data is no longer simply is the fact that big data, and the reality of the most important is to analyze big data, only through analysis to earn a lot of smart, in-depth, valuable information.

So more and more applications involve large data, and these big data attributes, including the number, speed, diversity and so on are all present the growing complexity of the large data, so the big data analysis methods in the field of big data is particularly important, is to determine whether the final information valuable decisive factor, as is shown by equation (2) [5].

$$\frac{dp_0(t)}{dt} = -(\lambda + up_0(t)) + u \quad , p_0(0) = 1, p_1(0) = 0 \tag{2}$$

The three concepts of data processing time shift: you all don't sampling, efficiency do not absolute precision and related do not cause and effect. Specific large data processing method in fact there are many, but according to the long time practice, the author summarizes a basic big data processing flow, and the process should be able to help you straighten out the big data processing. The whole process can be summarized as four steps, are gathering, import, and pretreatment, statistics and analysis, and mining.

Big data collection techniques: data is to point to by RFID data, the sensor data, social network interactive data and mobile Internet data way so as to obtain the various types of structured, semi-structured or called weakly structured and unstructured massive amounts of data is the basis of large data knowledge service model [6]. Key to break through the distributed high-speed high reliable crawl or data acquisition, large data collection technology such as high-speed data whole image;

Break through the high-speed data parsing, transformation and loading etc. Large data integration technology and it is design quality evaluation model, the development of quality and technical data.

The actual efficiency of data mining is needed before we value in big data mining to evaluate the questions very carefully. Do not necessarily all data mining plan can get ideal result. The first thing you need to guarantee the authenticity and completeness of the data itself, if the information collected by the noise caused by itself or some key data is not included, then dug up by the law of value is discounted. The second is to consider the costs and benefits of value to mining, if the mining project investment of human resources, hardware and software platform costly, project cycle is longer, and dig out the information for enterprise decision-making, cost benefit contribution is not big, so one-sided believe and rely on the power of data mining is unrealistic and not worth the cost.

Innovation Model Analysis of Data Flow Calculation and Storage Technology based on Cloud Computing

SPC is a kind of distributed stream processing middleware, to support the application of extracting information from the massive data stream. SPC contains for the realization of the distributed, dynamic and extensible application and provide programming mode and development environment, including its programming model is used to declare and create processing units (PE) API, as well as the assembly, test, debug, and deploy application toolset. Unlike other stream processing middleware, SPC, in addition to support relational operators, also supports the relational operators and user-defined functions.

Complex form of big data has led many to "rough knowledge" of measurement and assessment of the related research questions [7]. Known optimization, data envelopment analysis, expectancy theory, management science, utility theory can be applied to the study of how the subjective knowledge into rough knowledge of data mining in the process of "secondary mining". Here the human-computer interaction will play a crucial role.

Hadoop platform mainly off-line batch application oriented, typically through a static task scheduling batch operation data, the calculation process is relatively slow, some inquiries may take several hours or even longer to produce a result, the higher request for real-time applications and services is the co-action. Graphs are a kind of very good cluster parallel programming model, to meet the needs of most applications. Although graphs is distributed/parallel computing a good abstract, but it is not necessarily suited to solve any problem in the field of computing.

Strict traditional relational database design pattern, in order to ensure the consistency and give up poor performance and scalability problems are gradually exposed in big data analysis. Became popular,

no data storage model. No, others understand as Not Only SQL, is Not a specific data storage model, it is a kind of a relational database. Its characteristic is: there is no fixed data table model, can be distributed and horizontal extension. No is not a simple object to relational database, but against a supplement and extension of its shortcomings, as is shown by equation (3) [8].

$$L = \sum_i \sum_j np_{ij} = 0p_{00} + 1p_{01} + 1p_{10} + 2p_{11} + 2p_{b1} = \frac{4\rho + 5\rho^2}{H} \tag{3}$$

With the development of cloud computing technology to a wide range of applications, Hadoop distributed storage system based on open source and graphs data processing mode analysis of the system also has been widely used. Hadoop through data block and the recovery mechanism can support the petabytes of distributed data storage, and graphs based distributed processing model to analyze the data and processing [9]. Graphs programming model can be easily to be more general data processing tasks and on large-scale cluster parallel operation, and has an automatic failover. Graphs in open source software such as Hadoop programming model driven is widely applied, is applied to the Web search, fraud detection and so on all kinds of practical application.

Big idea, "4 v" characteristics of data processing and handling all determine the traditional way of data processing hardware and software implementation, innovating the mode of the big data applications. From the perspective of technology research, under the precondition of no loss of value, in order to improve the data quality, reduce the data scale as the goal of data mining technology, to extract value as the goal of data correlation analysis and the depth of mining technology and aiming at fast and efficient new big data calculation.

Experiments and Analysis

Big data analysis is dependent on the data quality and data management, the high quality of the data and effective data management, in both academic research and in the field of commercial applications, will be able to ensure that the results of the analysis of real and valuable. Big data analysis is the basis of the above five aspects, further large data analysis, of course, there are many, many more features, more in-depth, more professional large data analysis method.

In large-scale distributed database, HBase and Cassandra mainstream no database is mainly provide support and high extensibility will accordingly in terms of consistency and availability of sacrifice, in the traditional RDBMS ACID semantics, transaction support, etc. There are some drawbacks in Google's Megastore is trying to do a no merge with traditional relational database, and provides a strong consistency and high availability guarantee.

No database is a new data processing model based on cloud platform, no (in many cases, also known as cloud database. Because of the process data model is completely distributed in a variety of low-cost servers and storage disk, so it can help web pages and the huge amounts of data in the process of rapid processing various interactive applications [10]. It for the company, AOL, Cisco and other enterprises to provide web application support, as is shown by equation (4).

$$M(x,y)dx + N(x,y)dy = 0, \exists \mu(x,y) \neq 0 \tag{4}$$

Modeling, machine learning and statistical analysis and data are often linked to, is used to predict the events and actions. There are some things that are easy to be predicted, such as bad weather can influence the turnout of voters, but some are difficult to predict. Among voters, for example, change the vote on the decisive factor.

To cope with the challenge of huge amounts of data, some commercial database system attempts to combine the traditional RDBMS technology and distributed and parallel computing technology, to deal with large data requirements. Many systems or from the aspect of hardware is to accelerate. Data processing is a typical system has the IBM Netezza, Oracle's Exadata does, EMC Greenplum, HP Vertica, as well as the Teradata. Will tell from the function of these systems, can continue to support the operation of the traditional database and data warehouse and semantic analysis model, and on the

extensibility, also can use large-scale cluster resources data parallel processing, greatly accelerate the data loading, index and query processing time.

Summary

The paper presents innovation model analysis of data flow calculation and storage technology based on cloud computing. For large data processing of the data query, statistics, analysis, mining and other requirements, led to large data calculation of different calculation model, on the whole we put the big data is divided into offline batch calculation, real-time interaction and flow calculation.

Acknowledgements

This paper is supported by Henan key Laboratory for Big Data Processing & Analytics of Electronic Commerce, and also supported by the science and technology research major project of Henan province Education Department (13B520155, 17B520026).

References

[1] Olston C, Reed B, Srivastava U, Kumar R, Tomkins A. Pig latin: a not-so-foreign language for data processing, Proceedings of the 2013 ACM SIGMOD international conference on Management of data. ACM, 2013: 110-123.

[2] Gupta R, Gupta H, Mohania M. Cloud Computing and Big Data Analytics: What Is New from Databases Perspective, Big Data Analytics. Springer Berlin Heidelberg, 2012: 42-61.

[3] Schroedl S, Kesari A, Neumeyer L. Personalized ad placement in web search, Proceedings of the 4th Annual International Workshop on Data Mining and Audience Intelligence for Online Advertising (AdKDD), Washington USA. 2014.

[4] Rabkin A, Katz R. Chukwa: A system for reliable large-scale log collection, Proceedings of the 26th international conference on Large installation system administration. USENIX Association, 2012: 80-100.

[5] H.-s. XU, R.-l. ZHANG, "Semantic Annotation of Ontology by Using Rough Concept Lattice Isomorphic Model", International Journal of Hybrid Information Technology, Vol.8, No.2, 2015, pp.93-108.

[6] Neumeyer L, Robbins B, Nair A, Kesari A. S4: Distributed stream computing platform, Data Mining Workshops (ICDMW), 2015 IEEE International Conference on. IEEE, 2015: 260-277.

[7] Zhao Y, Hategan M, Clifford B, Foster I, von Laszewski G, Nefedova V, Raicu I, Stef-Praun T, Wilde M. Swift: Fast, reliable, loosely coupled parallel computation, Services, 2011 IEEE Congress on. IEEE, 2011: 460-481.

[8] Khetrapal A, Ganesh V. HBase and Hypertable for large scale distributed storage systems, Dept. of Computer Science, Purdue University, 2013.

[9] Yan XF, Zhang DX. Big Data Research. Computer Technology and Development, 2013, 23(4): 168-172.

[10] Wang S, Wang HJ, Qin XP, Zhou X. Architecting Big Data: Challenges, Studies and Forecasts. Chinese Journal of Computers, 2011, 34(10): 1741-1752.