Construction of Data Acquisition and Statistical Analysis Model based on Big Data Processing

Hongsheng Xu[1,2 a *], Ganglong Fan[1,2] and Ke Li[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: Data acquisition; Statistical analysis; Big data processing; Data mining; Distributed computing

Abstract. Statistics and analysis mainly use distributed database or distributed computing cluster to analyze and classify the massive data stored in it. The main features and challenges of statistics and analysis are that the amount of data involved in the analysis is large, and the system resources, especially I/O, will be greatly occupied. The method of big data analysis is an important factor in determining the value of the final information. The paper presents construction of data acquisition and statistical analysis model based on big data processing. Big data analysis, mining and utilization will bring great business value to enterprises.

Introduction

Big data has a multi tier structure, which means big data will change in a variety of forms and types. Compared to traditional business data, large data has irregular and fuzzy characteristics, which makes it difficult to even use traditional software for analysis. Traditional business data evolves over time and has a standard format that can be identified by standard business intelligence software. The challenge for companies today is to process and mine data from complex data presented in various forms.

The vertical integrated data model to connect to data services to more specific applications, the application data in a manner such that the customer relationship management, enterprise resource planning or dynamic data authentication is largely separated from each other in the service level, this separation is directly related to the data infrastructure [1]. In some cases, these applications may have SOA components that can access the storage / data services directly.

The original data is fragmented and, after filtering and organizing information (Information), related information integration and effective rendering of knowledge, deep understanding of knowledge and sublimation to understanding the nature of things and can extrapolate for wisdom. So data is the source, and it is the cornerstone of decision making and value creation. The so-called big data, on the one hand refers to in a certain period of time can not be conventional information technology and traditional database management software and hardware tools of perception, acquisition and processing of huge data set [2].

Data processing speed is in the case of very large data, but also to achieve real-time data processing. The last characteristic refers to the authenticity of the data is high, with the new data, social data, trading enterprise content and application of data source in traditional data source limitation is broken, companies increasingly need effective information resources to ensure its authenticity and safety.

Big data will bring one after another the IT technology revolution. In order to solve the problem of massive data, the growing diversity of data, data processing timeliness and other issues, will be in memory, data warehouse, system architecture, artificial intelligence, data mining and analysis of information communication and so on emerging breakthrough technology in today's world. Big data will lead to various types of innovation in all walks of life. With the development of big data, the industry gradually integrated, previously considered unrelated industry, through big data technology has interlinked channels.

The strategic significance of big data technology is not to master huge data information, but to deal with these meaningful data professionally [3]. In other words, if the big data compared to an industry, then the key to this industry profitability lies in improving the processing capacity of the data, through the "processing" to achieve value-added data". And China Internet of things school alliance believes that the development of Internet of things can not be separated from big data, relying on large data to provide sufficient resources.

RDMBS for fixed pattern, structured data, and has formed a mature storage, query, statistical processing. With the rapid development of Internet, Internet and mobile communication network, data formats and types in the changing and development. So for these different types of heterogeneous data, we need to adopt different data processing mode and storage that combination of structured and unstructured data storage. In the data management model and the architecture of the whole, also need to use a distributed file system and model. The distributed NoSQL database architecture can adapt to the large amount of data and the changing structure. The paper presents construction of data acquisition and statistical analysis model based on big data processing.

Big Data Application Solution Analysis

Therefore, the value of large data is to obtain the maximum data value through data sharing and cross multiplexing. In his view, the future of big data will be like infrastructure, data providers, managers, regulators, data cross multiplexing, big data into a big industry. With the growing share of big data, privacy issues follow, such as the daily phone calls, location, and so on. But it brings convenience as well as personal privacy.

Big data technology refers to the rapid acquisition of valuable information from various types of large amounts of data. The core of solving big data problems is big data technology. At present, "big data" not only refers to the size of the data itself, but also includes tools, platforms and data analysis systems for data collection [4]. Big data research and development aims to develop large data technology and its application to related fields, through the solution of huge data processing problems to promote its breakthrough development. Therefore, the challenges posed by the era of big data not only reflected in how to deal with large amounts of data, to obtain valuable information, but also reflected in how to strengthen the big data technology research and development, and seize the forefront of the development of the times.

Market analysis, big data association using a more accurate understanding of consumer behavior, explore new business model; logistics optimization, supplier collaborative work using big data, can ease the contradiction between supply and demand, control of expenditure, improve service. In the financial field, the application of large data within the enterprise has been rapid development, as is shown by equation (1) [5].

$$e_i^* = \frac{e_i}{\sigma} = \frac{y_i - \hat{y}_i}{\sigma}, i = 1,2,\cdots,n \tag{1}$$

Big data acquisition is generally divided into big data intelligent perception layer: mainly includes data sensing system, network communication system, sensor adapter system, intelligent identification system and software and hardware resources access system, to achieve massive data of structured, semi-structured and unstructured intelligent identification, positioning, tracking, access, transmission, signal conversion preliminary, monitoring, treatment and management etc.. We must focus on capturing intelligent identification, perception, adaptation, transmission and access to large data sources. The basic support layer: the virtual server that provides the big data service platform, the database of structured, semi-structured and unstructured data, and the basic support environment of network resources [6]. The key to distributed virtual storage technology, data acquisition, storage, organization, analysis and decision making operation visual interface technology, network transmission and data compression technology, big data privacy protection technology etc..

Statistics and analysis of the main use of the distributed database, or distributed computing analysis and classification of common summary of mass data storage within the cluster, in order to meet the demand analysis of the most common, in this regard, some real-time requirements will be used EMC GreenPlum, Oracle Exadata, and MySQL based storage Infobright so, some of the batch, or based on semi-structured data needs can use Hadoop [7]. The main features and challenges of statistics and analysis are that the amount of data involved in the analysis is large, and the system resources, especially I/O, will be greatly occupied, as is shown by equation(2).

$$d_i^1(t) = \frac{x_i^1(t) - x_{i+1}^1(t+1)}{x_i^1(t)} \tag{2}$$

The analysis of large data users with large data analysis expert, as well as ordinary users, but they are the two most basic requirements for large data analysis is visual analysis, because the visual analysis can reveal characteristics of big data directly, and can be easily accepted by the readers, like talk as simple.

The capacity of large data is usually up to the PB scale, and the massive data storage system needs to have corresponding scalability and throughput capability [8]. The growth rate of unstructured data is much faster than that of existing storage technology, which poses a challenge to traditional data storage and processing technologies.

With the arrival of the big data information wave, the library has also ushered in the era of large service for readers. The content of user services extends from data integration, management to data mining, analysis and display. At the same time, the library industry is also facing severe challenges and threats. How to strengthen the infrastructure construction of IT Library of data center, improve data collection, mining, processing, integration, analysis and decision-making ability, the data resources are efficiently converted into library information assets and productivity, has become an important problem facing the era of big data library industry.

Construction of Data Acquisition and Statistical Analysis Model Based on Big Data Processing

The data analysis method for random sampling analysis, maximize the use of all the data, rather than rely on a small portion of the data, the amount of data processing technology in the early and relatively backward "small data era", for the processing of data usually adopt random sampling analysis method to obtain valuable information through the analysis of the sample the data, which is a method based on minimum data to get the most information. And the full data model of full data analysis, its comprehensive data, beyond the sample analysis of the accuracy of the sample requirements.

But the NoSQL data storage has the following problems: one is relative to the strict access control and privacy management technology of SQL, the current NoSQL can not use the SQL mode, but also to adapt to the NoSQL memory model is not mature; the two is that although NoSQL software from the traditional data storage has experience, but NoSQL still exist all kinds of loopholes, the following formula show [9].

$$w_{i+1}^1\left(t+1\right) = \left(1 - wd_i^1\left(t\right)\right)x_i^1\left(t\right) - rs_i\alpha N^1\left(t\right) \tag{3}$$

The core idea of big data is data mining. With the help of computer knowledge discovery and data mining patterns hidden from massive data, is an interdisciplinary integration of computer, statistics and other fields of knowledge, the core of artificial intelligence, machine learning, pattern recognition and the theory of knowledge management implementation in the last century 90 times when there has been a remarkable progress. In essence, the "big data thinking big change and some data driven business intelligence model innovation, is the extension of the theory of data mining, data mining is expressed as relative to the statistical thinking brought by the change may be more accurate.

The development of big data will lead to many new and emerging jobs, which will generate data analysts, data scientists, data engineers, and people with very rich data experience will become scarce talent. With the development of big data, the data sharing alliance will gradually grow and become the core of the industry. With the growing share of big data, privacy issues follow, such as the daily phone calls, location, and so on. But it brings convenience as well as personal privacy. Data resources, big data in the national and enterprise and social level become an important strategic resource, become the new strategic high ground and the new focus of buying, as is shown by equation(4).

$$x^{(1)}(k+1)=\left(x^{(0)}(1)-\frac{b}{a}\right)e^{-ak}+\frac{b}{a} \tag{4}$$

In order to provide more uniform data integrity and management, management server can be used as SOA components to operate all kinds of database system, perform common tasks in specific database, such as weight and integrity check [10]. This approach is more easily adapted to legacy applications and data structures, but it destroys SOA as a service principle in asking data access, and can also lead to consistency issues in data management.

One of the core subjects of Intelligence. Statistical analysis: hypothesis test, significance test, variance analysis, correlation analysis, T test, variance analysis, chi square analysis, partial correlation analysis, distance analysis, regression analysis and simple regression analysis, multivariate regression analysis, stepwise regression, regression prediction and residual analysis, ridge regression, logistic regression analysis, curve estimation, factor analysis, cluster analysis, principal component analysis, factor analysis, clustering method and clustering analysis, discriminant analysis, correspondence analysis and multiple correspondence analysis (optimal scale analysis), bootstrap technology and so on.

The diversity of "big data" determines the complexity of data acquisition sources, ranging from smart sensors to social networking data, from sound pictures to online trading data, and the possibilities are endless. Selecting the right data source and cross analysis can create the most significant benefits for the enterprise. With the explosive growth of data sources, the diversity of data has become an urgent problem to be solved in big data applications.

Experiments and Analysis

Because of the semi-structured and unstructured features of large data, the structured "rough knowledge" (latent pattern) generated by data mining based on large data is also accompanied by some new features. This structured rough knowledge can be processed and transformed by subjective knowledge to generate semi-structured and unstructured intelligent knowledge. The search for "intelligent knowledge" reflects the core value of big data research.

Deep integration with cloud computing. Big data cannot do without the cloud computing technology, cloud computing and big data provides flexible and scalable infrastructure supporting environment and data services, data model, provides a new business value of cloud computing, therefore, from the beginning of 2013, big data technology and cloud computing technology will inevitably enter the period with more perfect the. Overall, cloud computing, Internet of things, mobile Internet and other emerging computing forms, both produce big data place, but also needs large data analysis methods.

In the storage and security of big data, big data format changing, due to the existence of a huge volume of features, it also brings a lot of challenges. For structured data, relational database management system RDBMS after decades of development, has formed a complete set of storage, access, security and backup control system. Due to the huge volume big data, and it is also caused by the impact of the traditional RDBMS, as mentioned earlier, the centralized data storage and processing in turn distributed parallel processing.

Big data can be divided into large data technology, big data engineering, big data science and big data applications and other fields. At present, the most talked about is big data technology and big data applications. Engineering and scientific issues have yet to be taken seriously. Big data refers to the system of engineering project planning and construction management of large data; data discovery and scientific attention to validate the relationship between big data rule and natural and social activities of the big data network development and operation process.

Summary

The paper presents construction of data acquisition and statistical analysis model based on big data processing. The strategic significance of big data technology is not to master huge data information, but to deal with these meaningful data professionally. In other words, if the big data compared to an industry, then the key to this industry profitability lies in improving the processing capacity of the data, through the "processing" to achieve value-added data". And China Internet of things school alliance believes that the development of Internet of things can not be separated from big data, relying on big data can provide sufficient resources.

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] Malewicz G, Austern MH, Bik AJC, Dehnert JC, Horn I, Leiser N, Czajkowski G. Pregel: a system for large-scale graph processing, Proceedings of the 2014 ACM SIGMOD International Conference on Management of data. ACM, 2014: 266-287.

[2] Foster I, Zhao Y, Raicu I, Shiyong L. Cloud computing and grid computing 360-degree compared,Grid Computing Environments Workshop, 2012. GCE'12. Ieee, 2012: 15-20.

[3] Zheng QL, Fang M, Wang S, Wang XQ, Wu XW, Wang H. Scientific Parallel Computing Based on MapReduce Model. Micro Electronics & Computer, 2015, 26(8):13-17.

[4] Keahey K, Freeman T. Contextualization: Providing one-click virtual clusters, eScience, 2015. eScience'15. IEEE Fourth International Conference on. IEEE, 2015: 447-589.

[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] Nurmi D, Wolski R, Grzegorczyk C, Obertelli G, Soman S, Youseff L, Zagorodnov D. The eucalyptus open-source cloud-computing system, Cluster Computing and the Grid, 2011. CCGRID'11, 9th IEEE/ACM International Symposium on, IEEE, 2011: 266-289.

[7] SONAJHARIA M, RAJNI J, “Rough Set Based Decision Tree Model for Classification”, 5th International conference on data warehousing and knowledge discovery, Prague, Czech Republic: DEXA Society, Vol.2737, 2013, pp.172-181.

[8] Ghemawat S, Gobioff H, Leung ST. The Google file system. In: Proc. of the 19th ACM Symp. on Operating Systems Principles. New York: ACM Press, 2016, 50−68.

[9] Baker J, Bond C, Corbett JC, Furman JJ, Khorlin A, Larson J, Leon JM, Li YW, Lloyd A, Yushprakh V. Megastore: Providing Scalable, Highly Available Storage for Interactive Services, CIDR. 2014, 66: 500-514.

[10] Li GJ, Cheng XQ. Research Status and Scientific Thinking of Big Data, Bulletin of Chinese Academy of Sciences, 2012, 27(6): 647-657.