ATLANTIS PRESS

Advances in Computer Science Research (ACSR), volume 73

7th International Conference on Education, Management, Information and Computer Science (ICEMC 2017)

Research on Industry Data Analysis Model Based on Hadoop Big Data Platform

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: Big data; Hadoop; Industry data analysis model; MapReduce; Potential valuable information

Abstract. Big data analysis refers to the huge size of data analysis, from a large amount of data to extract potential valuable information. Hadoop, the core of the Hadoop distributed file system and MapReduce, provides users with the underlying, detailed, transparent distributed infrastructure. In this paper, a business model statistic system based on big data is analyzed. This paper describes the traditional industry data analysis model based on big data technology, and puts forward the existing problems. The paper presents research on industry data analysis model based on Hadoop big data platform.

Introduction

With the continuous popularization of network and information technology, the amount of data produced by human beings is increasing exponentially. The emergence of a large number of new data sources has led to explosive growth of unstructured and semi-structured data. These data have gone far beyond what human beings can handle. How to manage and use these data has gradually become a new field, so the concept of big data comes into being.

Big data era is based on the Internet, Internet of things and other modern network channels, a large number of data resources based on the collection of data storage, value refining, intelligent processing and display of the information age. In this era, people can almost derive valuable knowledge from any data that can be translated into people's lifestyle changes [1]. The arrival of "big data era" has aroused extensive attention in the industry and academia, and a large number of research achievements have been emerging.

With the development of society and computer Internet, data is well preserved, and with the massive accumulation of data, we have entered the era of big data. Large data can be summarized as large amount of data, fast, many types, and low value density. Big data is not simply a fact of big data. Big data contains potentially valuable information that we don't know. The magnitude of the value of the data does not lie in the size of the data, but rather how much valuable information we can dig out from the data. For big data, the most important thing is to analyze it, and obtain a lot of intelligent, in-depth and valuable information through analysis. The method of large data analysis is particularly important in the field of big data. It can be said to be the decisive factor in determining whether the final information is valuable.

Telecom operators, which have decades of business data accumulation, have structured data such as business, sales and marketing, and also involve unstructured data such as text, audio, pictures, and video. From the data source, data from telecom operators to fixed telephone, fixed network access, wireless Internet and mobile voice and all other business, will also involve the family and public customers and enterprise customers, but also to collect contact information of direct sales channels, electronic channels, physical channels etc all kinds of channels.

In the actual work environment, many people will encounter massive data for this complex and difficult problem, its main difficulties are as follows: the large amounts of data, data in what circumstances may exist; the hardware and software requirements, excessive system resources;

processing methods and techniques require very high. Mass data mining is gradually rising based on the face of super massive data, the general algorithm is often used in data mining software or sampling method for processing, the error is not so high, greatly improving the processing efficiency and success rate.

In recent years, research on the desktop log, emerge in an endless stream, are based on the desktop web service log mining of page access behavior accounted for, users access the page order on the user through the analysis of user modeling. The study of behavior analysis for mobile user is in many ways from the desktop research, at the same time using geographic location records mobile terminal equipment, mining user mobile trajectory model, find out the trajectory in an important position and the combination of Internet communication data, log data and mobile application data as the basis of the study, analysis and mining the needs of users, behavior, interests, or even by predicting the user's destination, the next step is to infer the user location in order to provide recommendation for service. The paper presents research on industry data analysis model based on Hadoop big data platform.

Big Data Analysis Model Based on Hadoop Platform

Hadoop is an open source distributed computing platform under the Apache software fund's banner. Hadoop, the core of the Hadoop distributed file system and MapReduce, provides users with the underlying, detailed, transparent distributed infrastructure. High fault tolerance, high scalability and other advantages of the HDFS allows users to deploy Hadoop in cheap hardware, a distributed system, MapReduce distributed programming model allows users to develop parallel applications in distributed systems do not understand the underlying details of the case, as is shown by figure(1).

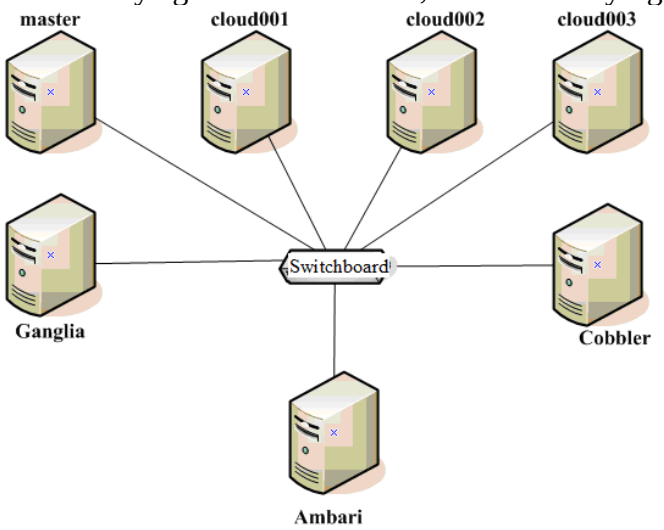

Figure 1. Hadoop cluster deployment diagram

Big data refers to the amount of data involved is huge to not pass the current mainstream software tools, achieve the collection, management, processing and finishing helping business decision-making becomes the purpose of consultation within a reasonable time [2]. Big data not only refers to the number of large amount of data volume (volumes), the initial measurement unit of big data is at least P (1000 T), E (1 million T) or Z (1 billion T), compared with the well-known G, the volume is not big. Secondly, is the data category (variety), data from multiple sources of data, data types and formats is rich, have broken through the structured data category previously defined, including semi-structured and unstructured data.

Chukwa is a large cluster monitoring system based on Hadoop, and it is an open source data collection system. Store data through HDFS and rely on MapReduce to process data.

Index analysis and comparative analysis can be divided into static comparison and dynamic comparative analysis [3]. Static comparison is the same time under different conditions such as general index, comparison of different departments, different countries and regions, also called transverse

comparison; dynamic comparison of numerical indicators in different periods of the same overall condition, also called longitudinal comparison.

How should we understand "big data" from a statistical point of view? It believes that big data is not based on artificial design, by means of traditional methods and the finite, fixed, discontinuous, extensible data structure model, but are based on modern information technology and the tool can automatically record, storage and continuous expansion, far beyond the traditional statistical record and storage capacity of all types of data, as is shown by equation (1) [4].

$$y_i = (1.8e^{\frac{(t-10)}{1.5}} - 1)\hat{y}_i \quad , t \geq 5 \tag{1}$$

In the era of big data, data mining is the most critical work. Data mining from massive, incomplete, noisy, and it is fuzzy and stochastic large databases in the discovery process in which valuable implicit and potentially useful information and knowledge, but also a decision support process. It is mainly based on artificial intelligence, machine learning, model learning, and statistics and so on. According to the data of highly automated analysis, make inductive reasoning, dig out potential models, can help enterprises, businesses and users to adjust the market policy, risk reduction, rational face of the market, and make the right decision [5].

Therefore, users can easily use Hadoop to organize computer resources, so as to build their own distributed computing platform, and can make full use of the computing and storage capabilities of the cluster to complete the processing of mass data.

(1) Extraction: because data may have different structures and types, data extraction process can help us put these complex data into a single or a convenient disposal configuration, in order to achieve the purpose of rapid processing.

(2) cleaning: for big data, not all the value, some data is not our concern, while other data is interference completely wrong, so it is necessary to filter the data by "denoising" to extract the effective data.

The mathematics problem brought about big data, in the mathematical sense, there are larger data set in the computer, there is no absolute big data, all the data in the computer set is a finite set of big data, big data sampling - the minimum sample becomes smaller, find and adapt algorithm set, sampling effects on Algorithm error.

Telecom operators have many years of data accumulation, the breadth and depth of their data resources is difficult to compare to the mobile Internet companies [6]. The advantages of telecom operators in large data applications mainly include the richness, integrity and continuity of data resources. Richness: refers to the data owned by telecom operators, involving a wide range, rich dimensions and huge amount of information. As mentioned earlier, these data relate to the behavior of all kinds of information of hundreds of millions of users, data from the magnitude of TB (1012GB) to PB and ZB, not only relates to the financial income, the amount of business development such as structured data, but also involves pictures, text, audio, video and other unstructured data.

A data mining system can automatically find predictive information in large databases, and a lot of manual analysis is needed in the past, and it can be concluded quickly and directly from the data itself. A typical example is the application of data mining in traffic accident, traffic accident data mining application effect analysis are: analysis of factors affecting the degree of traffic safety and effect, forecast the development trend of traffic accident, as is shown by equation(2), where it is accident identification area, intersection and road traffic accidents can be analyzed; the causes, characteristics, rules and traffic safety work in the weak link, clear of traffic safety management focus and countermeasures [7].

$$P(\beta;\lambda_1,\lambda_2) = \lambda_1 \sum_{j=1}^{p} P_1(\|\beta_j\|) + \lambda_2 \sum_{j=1}^{p} \sum_{m=1}^{M} P_2(|\beta_j^{(m)}|) \tag{2}$$

The principal component analysis, the geographical problems often involve a large number of interrelated natural and social factors, many factors bring great difficulties to constructing a model for

often, users are easy to understand and solve the existing problem of insufficient storage capacity, it is necessary to reduce some of the data and keep the necessary information. Principal component analysis is through statistical analysis, to obtain meaningful expressions essence between the various elements of linear relationship between the compressions of many elements of information expression synthesis variables for some representative; this eliminates redundant variables selection and then selects a few factors, the most abundant information for a variety of clustering analysis structure, application model.

Hive is a tool of data warehouse based on Hadoop, can be structured data file mapping for a database table, and provides a simple SQL query function, the SQL statement can be converted to MapReduce task operation. The advantage is that the learning cost is low, and the simple MapReduce statistics can be implemented quickly through the class SQL statement. Without the development of specialized MapReduce applications, it is very suitable for the statistical analysis of data warehouses.

Research on Industry Data Analysis Model Based on Hadoop Big Data Platform

Index refers to the relative number of changes in social economic phenomena. There are broad and narrow points. According to the index, the range of research can be divided into individual index, class index and total index [8]. Exponential function: one is the direction and extent of the overall number of changes can reflect the social economic phenomenon of the complex; two is the total change of some kind of social economic phenomenon can be analyzed by various factors change degree, this is a kind of factor analysis method. The method of operation is to observe the influence of a change of a factor on the total change through the quantity relation in the exponential system and the assumption that other factors remain unchanged.

Big data has a wide range of applications. Some institutions predict that the development of "big data" will increase the retail profit by more than 60%, and the cost of product development and assembly of the manufacturing sector will be reduced by more than 50%. In the manufacturing industry, the enterprise through the online data analysis to understand customer needs and grasp the market trends, and analysis of large data, it can effectively realize the reasonable procurement and inventory management, greatly reduce the loss of sales due to blind purchase [9]. In business, some foreign supermarkets use of mobile phone positioning and shopping cart obtain residence time inside the mall customers everywhere, analyze customer shopping behavior using video surveillance image software, and optimize the layout and arrangement of the mall shelves.

The biggest advantage of large data compared to sample data is that it has huge data selection space, and can carry out multidimensional and multi angle data analysis. More importantly, due to the large data volume and diversity, not enough to sample some rules of big data can reflect to some small sample; information capture, data can be covered; in the sample considered abnormal value, big data to be recognized. This will greatly enhance our ability to recognize phenomena, avoid losing a lot of important information, and avoid losing a lot of opportunities for decision-making. So, as is shown by equation (3), where it is in the era of big data, big data is both a sample, but also the overall.

$$w_{i+1}^{1}(t+1)=\left(1-wd_{i}^{1}(t)\right)x_{i}^{1}(t)-rs_{i}\alpha N^{1}(t) \tag{3}$$

From the beginning of the 0.20 version of Hadoop, the Hadoop Core project was renamed Common., which is a module of the bottom of the Hadoop system provides a variety of tools for Hadoop sub project, mainly including FileSystem, PRC and serial library [10].

The system should be established by using the mixed data storage technology, including data technology of relational database, MPP database, construction of large data storage structure, can analyze data according to the data model and data storage requirements structured, application support prediction and decision model.

The development of the era of big data open source technology and commercial software can meet as equals the homogeneity trend of operating system, and database platform, intermediate level software of the traditional has become apparent. End users focus on how to solve business problems rather than who to buy a database or operating system.

Summary

The paper presents research on industry data analysis model based on Hadoop big data platform. Big data era has high correlation analysis is required, according to the shortcomings of traditional surveying methods in statistical analysis, correlation analysis of the era of big data first to meet the "two general criteria" and "equality", the results of correlation analysis between variables and the close degree should only about. The effect, should not be in the form of related variables. Hadoop cluster advantages: high reliability, can maintain multiple copies of work data, to ensure that the nodes failed to redistribute processing. Highly scalable, allocating data between clusters of computers and performing calculations that can be easily extended to thousands of nodes.

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] Zhao G, Lai W, Xu C, et al. Analysis of User Behavior in Mobile Internet Using Bipartite Network. Mobile Ad-hoc and Sensor Networks (MSN), 2012 Eighth International Conference on. IEEE, 2012: 38-44.

[2] Hang Yang, Simon Fong, Guangmin Sun et al.. A Very Fast Decision Tree Algorithm for Real-Time Data Mining of Imperfect Data Streams in a Distributed Wireless Sensor Network. International Journal of Distributed Sensor Networks, 2012, 2012.

[3] Hung C C, Peng W C. A regression-based approach for mining user movement patterns from random sample data. Data & Knowledge Engineering, 2011, 70(1): 1-20.

[4] Patricia L. Mabry. Making Sense of the Data Explosion. American Journal of Preventive Medicine, 2011, 40(5).

[5] Sirivimol Thanchalatudom, Namfon Assawamekin, Using Big Data Technology for Information Management in Hybrid Learning System, RNIS, Vol. 12, pp. 179 -182, 2013.

[6] S. G. Wesnousky. Possibility of Biases in the Estimation of Earthquake Recurrence and Seismic Hazard from Geologic Data. Bulletin of the Seismological Society of America, 2010, 10-20.

[7] Rodrigo Rocha Silva, Celso Massaki Hirata, Joubert de Castro Lima, Computing BIG data cubes with hybrid memory, JCIT, Vol. 11, No. 1, pp. 13 -30, 2016.

[8] Radosław Bandomir, Mariusz Krawczyk, Jacek Namieśnik. A New Analyzer Based on Pellistor Sensor with Neural Network Data Postprocessing for Measurement of Hydrocarbons in Lower Explosive Limit Range. Journal of Automated Methods & Management in Chemistry,2005(2).

[9] Hongsheng Xu, Lan Wang and Wenli Gan, Application of Improved Decision Tree Method based on Rough Set in Building Smart Medical Analysis CRM System, International Journal of Smart Home, Vol. 10, No. 1, pp. 251-266, 2016.

[10] Sakr, Sherif & Gaber, Mohamed, Large Scale and Big Data, Auerbach Publications, 2014,pp.20-30.