Advances in Engineering, volume 126

5th International Conference on Machinery, Materials and Computing Technology (ICMMCT 2017)

Research on Big Data and Data Acquisition

Ning Huang

Central South University, Changsha, Hunan, 410083, China

Abstract

The big data tide is sweeping the globe and becoming a new kind of productivity. The big data technology not only means access to large amounts of data, more importantly is through the analysis and integration of massive data to get the valuable information hidden in the data. And as a new way of thinking, big data will give a new change to our society.

Keywords: big data, data analysis, data acquisition

1 The concept of big data

Big Data refers to those data beyond the traditional database system processing power. Its data size and transfer speed requirements are high, or its structure is not suitable for the original database system. In order to obtain the value of big data, we must choose another way to deal with it.

2 The development of big data

The first agency which pointing out the big data era has come to the world's leading consulting firm is McKinsey. McKinsey in the study pointed out that the data has penetrated into each industry and business functions, and gradually become an important production factor. And the use of massive data will signal a new wave of productivity growth and the arrival of the wave of consumer surpluses. Big data quickly became the computer industry competing to read the popular concept, also caused a high degree of strategy within the industry.

Although the big data is still in the initial stage in the country, but the commercial value has been revealed. In the future, the overall trend and development trend of Big Data will be reflected in several aspects: big data and academic, big data and human activities, big data security privacy, critical applications, system processing and the whole industry. Big data overall situation, the scale of the data will become larger, data resources, highlighting the value of data, the emergence of data privatization and alliance sharing. With the development of big data, data sharing alliance will gradually grow into a core part of the industry. Big data development will spawn many new professions, will produce data analysts, data scientists, data engineers, have a very rich data experience will become scarce talent. With the growing sharing of big data, privacy issues also ensued, such as the daily call, location, etc., but it brings convenience to the privacy of the problem. Data resources, big data in the national enterprises and social level to become the most important strategic resources, become the new strategic high ground and snapped new focus.

3 The classification of big data

3.1 According to real-time data analysis.

According to real-time data analysis divided into real-time data analysis and offline data analysis of two. Real-time data analysis is generally used for financial, mobile and Internet B2C and other products, often require a few seconds to return to hundreds of millions of rows of data analysis, so as not to affect the user experience purposes, for most of the feedback time requirements are not so demanding applications, such as off-line statistical analysis, machine learning, search engine reverse index calculation, recommendation engine calculation, etc., should be used offline analysis, through the data acquisition tool will log data into a dedicated analysis platform. But in the face of massive data, the traditional ETL tools are often completely ineffective, mainly due to the cost of data format conversion is too large, the performance cannot meet the massive data collection needs.

3.2 According to the amount of data in big data,

According to the amount of data in big data, it divided into memory level, three levels of mass, BI level. The memory level here refers to the amount of data that does not exceed the maximum memory of the cluster.

The level of mass refers to the database and BI products have been completely invalid or costly data volume BI level refers to those who are too large for the amount of memory data, but generally can be placed in the traditional BI products and specially designed of the BI database for analysis.

4 The characteristics of big data

4.1 Massive

Enterprises are facing large-scale growth in the amount of data. For example, IDC's latest report predicts that by 2020, the global data volume will expand 50 times. At present, the size of big data is still a constant indicator of change, a single data set size ranging from tens of TB to several PB range. In short, storing 1PB of data will require 20,000 PCs with a 50GB hard drive. In addition, various unexpected sources can generate data.

4.2 Volatility

Big data has a multi-layer structure, which means that big data will show a variety of forms and types. Compared with the traditional business data, big data is irregular and fuzzy features, resulting in difficult or even unable to use the traditional application software for analysis. Traditional business data has evolved over time to a standard format that can be recognized by standard business intelligence software. At present, the challenge is to deal with and from all forms of complex data mining value.

4.3 Diversity

A common view is that the use of Internet search is the main reason for the formation of data diversity, this view is partly correct. However, the increase in data diversity is mainly due to the new multi-structure data, as well as including network log, social media, Internet search, cell phone call records and sensor networks and other data types. Some of these sensors are installed on trains, cars and airplanes, and each sensor adds to the diversity of data.

4.4 High speed

High-speed describes the speed at which data is created and moved. In the era of high-speed network, through the realization of software performance-based high-speed computer processors and servers, creating real-time data flow has become a popular trend. Enterprises need to not only know how to quickly create data, but also must know how to quickly process, analyze and return to the user to meet their real-time needs. According to IMS Research on data creation speed of the survey, it is predicted that by 2020 the world will have 22 billion Internet-connected devices.

5 The acquisition technology of big data

5.1. Differences between data acquisition and big data acquisition

Computer science relies heavily on models and algorithms before big data appears. If people want to get accurate conclusions, the need to establish a model to describe the problem, at the same time, need to straighten out the logic, understand the cause and effect, the design of sophisticated algorithms to draw close to reality. Therefore, a problem, can get the best solution, depending on the modeling is reasonable, a variety of algorithms to determine the success or failure of the key. However, the emergence of big data completely changed people's dependence on modeling and algorithms. For example, suppose that there is Algorithm A and Algorithm B to solve a problem. When running in a small amount of data, the result of algorithm A is obviously better than that of algorithm B. That is, algorithm A can bring better results in terms of the algorithm itself; however, it has been found that when the amount of data grows, algorithm B runs better on a large amount of data than algorithm A does on small amounts of data In the results of the operation. This discovery has led to a milestone in both the computer science and the computer-derived disciplines: when the data grows larger, the data itself (rather than the algorithms and models used to study the data) ensures the validity of the data analysis results. Even if the lack of precise algorithms, as long as there is enough data, but also get close to the fact that the conclusion. The data is thus known as the new productivity.

When the data is sufficient, it does not need to understand the specific causal relationship to be able to draw conclusions. For example, Google does not set up various grammar and translation rules to help users translate. But rather use the Google database to collect all the user's words used to compare the recommendation. Google checks the writing habits of all users, the most commonly used, the highest frequency of translation recommended to the user. In this process, the computer cannot understand the logic of the problem, but when the user behaviours' records more and more data, the computer cannot understand the logic of the problem under the circumstances, to provide the most reliable results. Visible, massive data and analytical tools to deal with these data, to understand the world provides a complete new way.

With the ability to handle multiple data structures, big data can be analyzed to the fullest extent possible using human behaviour data recorded on the Internet. The big data appears before the computer can handle the data need to be pre-structured, and recorded in the corresponding database. But big data technology requirements for the structure of the data greatly reduced on the Internet people left behind the social information, geographic location information, behavioural habits information, preferences and other dimensions of information can be processed in real time, three-dimensional and complete outline of each one various characteristics of the individual.

5.2 The lack of traditional data collection

Traditional data acquisition sources are single, and the amount of data stored, managed and analyzed is relatively small. Most relational databases and parallel data warehouses can be used. Traditional parallel database technology is highly consistent and fault-tolerant, and it is difficult to guarantee its usability and scalability according to the CAP theory in terms of speeding up data processing by relying on parallel computing.

5.3 The new methods of big data acquisition

Network data collection is through the web crawler or website public API, etc. from the website to obtain data information. This method can be unstructured data extracted from the Web page, store it as a unified local data files, and stored in a structured manner. It supports pictures, audio, video and other documents or attachments to the collection, attachments and text can be automatically associated. In addition to the content contained in the network, the collection of network traffic can be handled using bandwidth management techniques such as DPI or DFI.

For data that requires high confidentiality, such as production and business data or academic research data, data can be collected through cooperation with enterprises or research institutions using specific system interfaces and other relevant methods. Flume is Apache's open source, highly reliable, highly scalable, easy to manage, to support customers to expand the data acquisition system.

Fluentd is another open source data collection framework. Fluentd uses C / Ruby development, using JSON files to unify log data. Its pluggable architecture supports data sources and data outputs in a variety of different formats and formats. Finally, it also provides high reliability and good scalability. Treasure Data, Inc. provides support and maintenance for this product.

Splitter is a distributed machine data platform with three main roles: Search Head is responsible for data search and processing, providing information extraction when searching; Indexer is responsible for data storage and indexing; Forwarder, responsible for data collection, cleaning, and sent it to Indexer.

References

[1] Tu Zipei. Big Data. Guangxi Normal University Press. Nanning. pp. 68-70, 2012

[2] By Victor Meyer Schoenberg. Sheng Yangyan, Zhou Tao translation. Big Data Era. Zhejiang People Publishing House. pp.87- 91, 2013

[3] Xiong Yi. "Big Data" brightens the future. Journal of Harbin Institute of Technology, 12(10), pp. 18- 20, 2012

[4] Jin Zongze, Feng Yali, Ji Bo, Zhang Xi, Gao Kuai. Association Mining in Big Data Analysis China Building Materials Science and Technology, 9(6), pp.58- 61, 2011