# Research on Large Data System and Its Analytical Technology

Rongjian Yuan

GuiZhou Vocational Technology College of Electronics & Information, Kaili Guizhou, 556000, China

**Keywords:** Big data system, Data analysis, Application technology.

**Abstract.** Over the years, the rapid development of large data has aroused great concern in various fields of experts and scholars, many authoritative magazines will also be large data development and application as the world's development opportunities and challenges. On this basis, this paper analyzes the application of large data system and technology according to the characteristics and applications of different forms of large data, and discusses various countermeasures for reference.

## Introduction

Compared with the traditional data, the large data system is mainly characterized by large volume, high speed, modal number, difficult to identify and low density, and the difficulty of large data system application is not the large amount of data, The processing of structured data, the integration of unstructured data, and the authenticity of data are faced with great challenges. The large data system has a lot of unpredictable application factors. On this basis, the author conducts analysis for large data systems and application technology.

## Research on Large Data Processing System

### Batch data processing system

This is a use of batch data for deep-seated mining application model, tap the deep value, easy to develop the right decision-making, which is the important work of large data batch processing, after the first storage calculation of some real-time requirements are not high The task. Batch data processing system features include the following aspects: (1) a huge amount of data, especially from the TB level to the PB level, a lot of data stored in the static form of hard disk, update slow, long time, repeated use (2) the accuracy of the data is higher, especially in the application of precipitation out of the various data in the business has become a very valuable asset; (3) the data is more valuable, it is not easy to move or backup a large number of data; ) Data density is low, such as video batch data, many applications in the continuous monitoring of the available data is extremely small, there must be a reasonable algorithm to extract useful value, and these data processing time is usually more, to provide User system interaction means less, so the results of the treatment and the expected value is often a large contrast, more suitable for some large-scale mature operations [1].

Internet of things, the Internet, cloud computing and so are important sources of large data, batch data in which the processing can have a new insight. Such as the Internet field, people commonly used Sina microblogging, Facebook and other social networking exchanges have produced a lot of pictures, video, text, the bulk of these data processing is the analysis of social networks, used to find hidden between people Relationship or co-access to the community.

### Streaming data processing systems and interactive data processing systems

The streaming data processing system is a real-time processing system for data information. It has two forms: interactive data processing and streaming data processing. The streaming data processing in the large data background is the evolution of the server log real-time acquisition, interactive data Processing is to shorten the processing time. The characteristics of streaming data are the following: (1) the source of the elements of the sequence of diverse, different formats, from the database point of view, different elements is different Yuan Zu [2]. (2) In different occasions, different characteristics of streaming data, such as the characteristics of the elements Su Liang, flow size, format, etc., but

most have a common feature, in order to common in the flow of data systems for processing. Application of common data acquisition applications, need to take the initiative to obtain massive real-time data, mining value information, and through the sensor acquisition, log collection and other forms to complete. For example, the financial data generated in the financial banking industry is relatively short, this system can help financial banks to develop real-time decision-making, than the traditional business intelligence BI have more dynamic, timelier response to demand.

The characteristics of interactive data processing are flexible, intuitive and easy to control, can be the first time to receive the operator's request, through the data dialog input prompt information, and guide the operator to complete the operation to get the results. While the data files stored in the system can be modified in a timely manner and the results can be used immediately. Such as e-mail, search engines, social networks, etc., users as long as the platform to share information, such as Yahoo, Baidu, etc., forcing the interactive platform applications more and more frequent, diversity and diversification, more data to meet the real-time Claim.

**Graphic data processing system**

Because this system has a very good structure, it can clearly show that the link between things, including the map data query, storage, the shortest path consulting, keyword consulting and so on. But with the increase in the number of nodes, the complexity of the data processing complexity defects will gradually be reflected. The characteristics of the graph data are expressed as follows: (1) the association between nodes. (2) There are many types of graph data, such as biology, chemistry, information retrieval and so on. (2) There are many kinds of graphs, such as labels, semantic graphs and so on. (2) There are many kinds of graphs, (3) The graph data calculation has a strong coupling between the graph data, and the data of the graph data is between the graph data and the data in the graph. Is connected, so there is a certain relationship between the calculation of the data, the existence of coupling makes large data calculation has a higher challenge, because it can not be handled according to a single machine, and the map of each vertex is more difficult to divide For a number of independent subgraph processing, so this has become a map data processing system problems and challenges [3]. In the application, the social network is to establish a large number of online social network relationship, with the structure of the map to express the relationship between different individuals, such as E-mail that the relationship between people and the relationship between the expression of social groups. In the field of transportation, the plan is the shortest path to dynamically query the traffic network. It is undeniable that the processing of graph data has a very good application prospect.

From the above finishing, we can see that the large data processing system type is very large, and distinctive, summarized its future development trends are the following: (1) data processing engine specialization. This is a way to reduce the cost, improve efficiency, get rid of the traditional model of the development of the traditional, from the current development point of view, domestic and foreign Internet companies are in the pursuit of the development of large-scale, low cost and high throughput expansion system; 2) Data processing platform diversification. This trend is inevitable, the world in 2008 after the cloning of Google's GFS, this phenomenon is very common, but the main goal is to focus on how different data processing platform to significantly improve the system performance and ensure that System ecosystem and the integrity of the environment; (3) real-time data calculation. This is a supplement to the batch calculation, especially for the PB level data, as short as possible to deal with time to ensure real-time [4].

**Big data analysis**

**Deep learning**

The most important problem with large data analysis is to ensure that data is effectively expressed, interpreted, and learned in depth, including image, sound, and text data. Because the traditional model application has limitations, such as too dependent on data expression, so the general effect of learning,

and large data appears after the model is more complex, to characterize and explain the data in detail, so the depth of learning through the hierarchical Architecture to learn the different levels of object expression, can solve the complex abstract problem. In-depth learning usually uses artificial neural networks, the more common is the multi-layer perceptron MLP depth architecture.

The most important problem with large data analysis is to ensure that data is effectively expressed, interpreted, and learned in depth, including image, sound, and text data. Because the traditional model application has limitations, such as too dependent on data expression, so the general effect of learning, and large data appears after the model is more complex, to characterize and explain the data in detail, so the depth of learning through the hierarchical Architecture to learn the different levels of object expression, can solve the complex abstract problem. In-depth learning usually uses artificial neural networks, the more common is the multi-layer perceptron MLP depth architecture.

Over the years, deep learning has been extended to the image, voice, natural language, in 2009, Microsoft Research Institute Dahl has been using deep neural network to deal with voice, to a large extent reduce the voice error rate, making it the first depth Learning to extend the field. By 2012, Hinton's use of deep-level convolution neural network evaluation is to reduce the error rate from 26% to 15%, which applied to the training data and the value of the parameters. In 2013, this data dropped to 11.2%, and new performance was found by neural network face recognition, and the correctness of face recognition was close to human level. In addition, it should be noted that in the field of images there are some can not monitor the depth of learning, such as trying to completely mark the image training to achieve the face feature detection test has achieved results [5]. At present, China's Baidu has also set up a depth study and research institute, the depth of learning algorithms to study, have also thought about the depth of learning technology has become Baidu on-line products, the development of Baidu far-reaching, such as face recognition, image search and many more.

**Knowledge calculation**

The basis of large data analysis is the knowledge calculation, high-end analysis of the data, it is necessary in the large data which has the knowledge of the value of knowledge, to be able to support the query, analysis and application of the knowledge base. At present, there are more than 50 large data base types in various countries around the world, and the application system is more than 100 kinds, such as KnowItAll, Probase, Satori and so on. There are also some of the more well-known commercial websites or companies also use this platform, such as the US official website Website Data.gov, Wolfram knowledge computing platform wolframalpha. China Academy of Sciences Institute of Mathematical Sciences proposed knowware, and Shanghai Jiaotong University, the earliest use of the building knowledge platform is zhishi.me, these are the application of knowledge computing [6]. Analysis of knowledge base, including three parts, namely, the construction of knowledge base, the integration of multi-source knowledge and knowledge base updates. Construction includes the concept of knowledge, examples, attributes, relationships, etc., to build a manual construction and automatic construction, such as China's application of more knowledge network, the concept of hierarchical network. Multi-source knowledge fusion is to solve the problem for some reuse, we must be aware of any knowledge base construction costs are very large, then in order to prevent the process of building a variety of repetitive knowledge of the conflict, we must introduce knowledge The concept of reuse and sharing, the integration of different sources of knowledge, the different data to clean up. According to the integration method can be divided into manual integration and automatic integration, if the knowledge base is small, then you can take the way of manual integration, but its time-consuming and prone to error must also be improved. For example, Wikipedia is the integration of knowledge classification system structure, the internal classification of the integration of the system, there are entities and concepts of the digestion problem, and for different areas to establish a number of corresponding areas, according to different needs of different knowledge fusion algorithm , In order to achieve rapid integration of multiple knowledge, which is the future need to do the work.

## Social computing

*Research on Online Social Network Structure*

From the micro level, the online social network has the characteristics of random disorder, and the macro level can be included in the scope of regularization. Clarifying the seemingly contradictory structure of the network is an important part of the online social network. In general, the community structure contains a number of internal links in the connection relation of the network nodes, and the community analysis work includes the definition, measurement and structure evolution. In particular, attention should be paid to evolution, which is the root cause of the large-scale trend of information network content and structure. Over the years, people have begun to study through the evolution of time, such as Palla et al is in the complete sub-map seepage community found the nature of community evolution, and thus get a more interesting conclusion, that is, the stability of the smaller community is Its existence in the premise, if the scope of the larger, then the dynamic has become the basis. In addition, in the introduction of the MMSB model is also actively introduced the concept of time, for example, some scholars believe that the two time films in the role of choice and its relationship with the first-order Markov nature, which the application of social computing problems Have good tips and help.

*Online social network information dissemination*

Throughout the knowledge of information dissemination research, the most profound and extensive study of infectious disease model, in addition to random walk model. Over the years, researchers have found that information dissemination and the spread of infectious diseases exist in different characteristics, such as the proposed Persistence and Stickiness concept, is applied in the field of Twitter communication theory, the establishment of Twitter news for the degree of measurement, Establish a system for presenting real-time information to customers.

## Visualization

The visualization of large data is fundamentally different from the visualization of traditional information. The first challenge is the scale, and the integrated information has large-scale, high-dimensional, multi-source and dynamic evolution. In view of the challenges of large data visualization in today's application, this paper presents four main research directions: (1) Compressing information flow or deleting data to deal with and simplify redundant information, such as simplification of information node grid, The algorithm is extended to tetrahedral calculations. (2) the design of multi-scale and multi-level approach to ensure that users can autonomously control the resolution, such as the way the fixed grid system will be built on the hierarchy, then you can use the quadtree texture layer or triangular one-sided binary tree to display data The ability to adapt. (3) the external data of innovative data, allowing users to interact with each other more convenient and quick access to information, and design innovative geometric algorithms to solve visual problems. (4) to propose a new visual metaphor method to fully display the data, such as the tree browser is to dynamically adjust the tree size of the deployment area, the expression of information.

## Conclusions

To sum up, in the application of large data systems, people face the data complexity, computational complexity and system complexity of the triple challenge. This paper analyzes the different data processing forms such as batch processing data and interactive data, summarizes the development trend of the future system such as engine specialization and platform diversification, and analyzes the depth of learning, knowledge calculation, social computing and visualization Data technology, that the relevant personnel of the large data system technology to provide a reference.

## References

[1] Zhang Lei, Du Dongmei. Discussion on Large Data System and Analysis Technology, *Modern economic information*,2014,12(11):336.

[2] Wang Chao, Ma Yanchao. Research on Log Statistics and Analysis System Based on Large Data Technology, *Database and Information Management*, 2017,2 (15): 9-11.

[3] Wu Huinan. Summary of large data systems and analytical techniques, *Information recording material*,2016,17(3):2-5.

[4] Cheng Xueqi, Wang Yuanzuo. Large data system and analysis technology, *Journal of Software* ,2014,9(18):1907-1908.

[5] Cao Juncheng, Ming Yangyang. Summary of Large Data Analysis Technology for Energy Internet, *South China Power Grid Technology*, 2015,11 (1): 1-12.

[6] Cheng Xueqi. Large data research: the future of science and technology and economic and social development of the major strategic areas, *Chinese Academy of Sciences*, 2012,2 (27): 647-657.