On the Discussion of Computer Mass Data Access and Its Characteristics in Time Limit

Xingping Wu¹, Xusheng Liu¹, Ziqian Li¹, Zhidong Deng¹, Qing Zhu¹, Li Zhang²,*

¹State Grid Customer Service Center.
²Nanjing Data Foundation Technology Co., Ltd.
*Corresponding author e-mail: zhangli_66666@126.com.

Abstract. With the rapid development of the Internet era, the Internet of Things, big data, mobile payments, and the digital age of AI, information technology is gradually covering people's lives and learning. People come into contact with tens of thousands of information every day. Getting useful data can not only shorten the time effectively, but also meet specific needs. For example, the State Grid will conduct big data analysis on our daily electricity consumption according to our daily electricity consumption, analyzing the peak and trough of electricity consumption, and the abnormal electricity consumption in January. In all the consumption data analysis, the results of the analysis are used in the power transmission and transformation business planning of the power sector, carry out targeted preparation plan in the specific transmission of power, and further improve the quality of power consumption. Based on this, the paper discusses and analyzes the massive data access, massive data analysis and data effectiveness.

1. Introduction
In one minute, the number of new data on Twitter exceeded 100,000. Facebook, a social network, has seen more than 6 million views. The huge amount of logs and communication records brought about by the popularization of broadband, the personal information constantly updated by social networks every day, video records such as video communication, medical images, geographic information, surveillance videos, data information generated by non-traditional IT equipment such as sensors and navigation equipment, and pictures and information generated by various intelligent terminals that are continuously increasing, these explosive growth data are flooding the entire network. At the same time, the data structure of these massive data is more diversified, and images, videos and documents account for a large proportion. According to the data, the growth rate of unstructured data such as email, video, microblog, post, mobile phone call and web page click is 80% every year. And these data contain a lot of valuable information. If we can effectively dig out their value, it will undoubtedly bring huge economic benefits to enterprises. Therefore, for enterprises, the explosive growth of massive data driven by the Internet is both a huge opportunity and a huge challenge. Because on the premise of massive data, IT solutions must not only be able to store and manage more efficiently and at low cost, but also be able to retrieve and analyze more quickly, flexibly, and stably. This article discusses the computer's massive data access and its characteristics of timeliness.

2. Computer Mass Data Access and Its Research on Aging
Faced with such a huge amount of data, the first problem is that the size of this data has exceeded the load capacity of a single machine. Using machine clustering and parallelization technology is an effective way to solve the problem. The following methods are popular in recent years.

2.1. High Performance Computing

Broadly speaking, HPC (High Performance Computing) is to assign jobs to a cluster of machines that access a shared file system and are managed by a Storage Area Network (SAN). It is very suitable for jobs that are mainly compute intensive. But when the data flow that a node needs to access is large (for example, hundreds of gigabytes of data), a problem will arise: because the network bandwidth becomes the "bottleneck", the computing node is idle [1].

At present, the most widely used parallel programming environment in high-performance computer systems at home and abroad is MPI (Message Passing Interface), which has become an international standard for parallel programs. MPI has many advantages such as good portability, powerful functions, high efficiency and so on, and it has many different free, efficient and used implementation versions. Almost all parallel computing machine manufacturers provide support for it, which is incomparable to all other parallel programming environments [2].

2.2. Grid computing

By utilizing the unused resources (CPU cycles and disk storage) of a large number of heterogeneous computers (usually desktops), it is used as a virtual computer cluster embedded in a distributed telecommunications infrastructure, providing a model for solving large-scale computing problems. Grid computing focuses on the ability to support cross-management domain computing, making it different from traditional computer clusters or traditional distributed computing.

No matter in a narrow sense or a broad sense, the goal of the grid is nothing more than to use the Internet to form a "virtual supercomputer" of computers scattered in different geographical locations, so as to realize the overall sharing of computing resources, storage resources, data resources, information resources, software resources, communication resources, knowledge resources, expert resources, etc. Each participating computer is a node, just like the chess pieces placed on the weiqi board, and the crisscross lines on the board correspond to the real world network, so the whole system is called "grid". Computing on the grid is like playing go. It's not done by a single piece of chess, but by all pieces working together to form a joint force. Traditional Internet realizes the connection of computer hardware, web realizes the connection of web pages, and grid tries to realize the comprehensive connection of all resources on the Internet.

Grid system is a resource sharing model, and resource providers can also become resource consumers. Grid focuses on how to combine distributed resources into dynamic virtual organizations. It takes computing as the center, and computing resources and storage resources are distributed in every corner of the Internet. It does not emphasize that computing and storage resources required by tasks are located in the same place. Due to the limitation of network bandwidth, the data transmission time in grid computing accounts for a large part of the total running time [3].

2.3. Cloud computing

In data-intensive computing, the size of data is like a long wooden stick in a performance tent. Big data, once the exclusive domain of scientific and engineering HPC environments, now runs through the data range of large data centers. Whether it is in the detection of e-commerce fraud, improving security through better intelligence collection, or 3D high definition TV or movies on the Web, the era of data-intensive computing has come [4].
As shown in figure [4], data-intensive computing can be thought of as a filtering process. It starts with a large amount of data and ends with the final decision (yes or no, buy or not, live or die). It needs to take into account a large number of structured or unstructured data, and transform them into useful information through a series of processes to improve intelligence, knowledge, experience and other insights, and ultimately make better decisions.

The effective way to solve the problem of data intensive computing is to move the computing to the data, not the traditional way. Cloud computing is a better way to solve this problem. For example, Hadoop Map Reduce framework uses distributed computing nodes to form a distributed storage environment, and the data is divided into a plurality of blocks for storage. Each computing node only needs to use its own local data and does not need to transfer a large amount of data [1,5].

Generally speaking, cloud computing has the following forms.

SAAS (Software as a Service): This type of cloud computing transmits programs to thousands of users through browsers. In the user's view, this will save the expenditure on server and software authorization. From the supplier's point of view, only one procedure needs to be maintained, which can reduce costs. SAAS is commonly used in human resource management programs and ERP. Google Apps and Zoho Office are similar services.

Utility Computing: Practical computing was only recently reborn in Sun, IBM and other companies that provide storage services and virtual servers. This cloud computing is to create a virtual data center for IT industry, enabling it to centralize memory, I/O devices, storage and computing capabilities into a virtual resource pool to provide services for the entire network.

Web Service: Closely related to SAAS, network service providers can provide API for developers to develop more Internet-based applications instead of providing stand-alone programs. For example, cloud computing has related applications in online games. Drunk Happy has set a precedent for the innovative application of cloud computing in the field of online games. It adopts a diversified computing mode of cloud computing and integrates the development of traditional computer technology and network technology, making game data more efficient and safe to operate calculations, thus optimizing the speed of game computation and enabling people to experience the game world more smoothly.

Platform as a Service (PaaS): In this form, cloud computing provides a development environment as a service. Users can use middlemen's equipment to develop their own programs and transmit them to users through the Internet and its servers.

Management service provider (MSP): one of the oldest cloud computing applications. This kind of application is more oriented to the IT industry rather than the end-user, and is often used in e-mail virus scanning, program monitoring, etc.

Business service platform: SAAS and MSP mixed application. This type of cloud computing provides a platform for interaction between users and providers. For example, the user's personal...
expense management system can manage its expenses according to the user's settings, and coordinate the various services it orders.

(7) Internet integration: to integrate companies providing similar services on the internet so that users can compare and select their own service providers more conveniently.

3. Mass Data Access and Its Timeliness
There are five main characteristics of massive data: Volume (large capacity), Variety (various types), Velocity (fast speed), veracity (difficult to identify) and most important Value (low value density).

Volume (large capacity) refers to the huge data volume and data integrity of the computer's massive data access. It can refer to the large amount of data contained in the computer massive data access collection, or to the large number of sub-data contained in the network constituting the computer massive data access.

Variety (various types) means to find the inherent correlation between massive and diverse data. There are many types of data included in computer massive data access, which can include various structured data types, as well as various unstructured data types, and even other data types.

Velocity can be understood as meeting the real-time requirements faster. The structure and content of the computer's massive data access can be dynamically changed, and the frequency of change is high, the speed is fast, and the range is wide. The data form is extremely dynamic, and the processing needs extremely fast real-time.

Veracity (difficult to identify) can be reflected in many aspects such as the content, structure, processing, and association of sub-data contained in the data. Computer massive data access can include many random numbers with different probability distributions and many fuzzy numbers with different domains. The associations between data are ambiguous and may change randomly at any time.

Value (low value density) refers to the low value density of computer massive data access. In massive data, useful data is often mixed with a large amount of useless data. Therefore, a necessary job for processing massive computer data is to "continuously remove garbage, remove noise". The value of massive computer data access is like gold rushing in the sand. The larger the amount of data, the less valuable it is.

4. Computer Mass Data Access and Its Technical Methods in Time Effect
In order to apply technical barriers such as data diversity, format diversity, cross-system and cross-language and improve the access standard of computer communication data, a scientific and well-structured and practiced technical organization mode should be adopted, and a data sharing mode of separating real-time data from offline data should be adopted to access and share data. Firstly, the computer network communication system should be optimized, the application of data encryption technology, data authentication and other technologies should protect the data itself, build multi-level fault tolerance, and ensure data consistency. The technology of data access and data effectiveness will be described below.

4.1. Aggregation Application under Large Data Volume
The value of big data aggregation technology will be analyzed in specific application scenarios. The general technical framework is as follows:
Figure 2. The general technical framework.

The collection and aggregation layer usually uses open source technology components such as:

File beat is an indicator such as collecting logs; file beat itself is technically open source and universal. Meet current mainstream acquisition technology architecture requirements;

metric beat collects the operating indicators of the host, the technology is open source, and the collection is real-time and accurate.

Apm agent collects business operation indicators; the technology is open source, collecting real-time and accurate data;

Flume: Provides a highly available, highly reliable, distributed massive log collection, aggregation, and transmission system. Flume supports customizing various types of data senders in the log system for data collection. At the same time, Flume provides the ability to easily process data and write to a variety of data recipients (customizable). It is more like a pipe, taking on the responsibility of data transmission, which is characterized by real-time and high availability.

Real-time data is stored in Elastic search; Elastic Search is a Lucene-based search server. It provides a distributed multi-user full-text search engine based on RESTful web interface. Elastic search is developed in Java language and released as an open source under Apache license terms. It is a popular enterprise search engine. Elastic Search is used in cloud computing and can achieve real-time search. It is stable, reliable, fast and easy to install and use. Official clients are available in Java, .NET (C #), PHP, Python, Apache Groovy, Ruby, and many other languages. According to the ranking of DB-Engines, Elastic search is the most popular enterprise search engine, followed by Apache Solr, which is also based on Lucene. Ensure that data is provided to third-party audiences in real time and accurately.
Techniques related to offline data storage shown in Figure 3:

HADOOP: Hadoop is a distributed system infrastructure developed by Apache Foundation. Users can develop distributed programs without knowing the details of the distributed bottom layer. Make full use of the power of clusters for high-speed computing and storage. Hadoop implements a Hadoop Distributed File System (HDFS). HDFS has high fault tolerance and is designed to be deployed on low-cost hardware. Moreover, it provides high throughput to access application data, which is suitable for applications with large data sets. HDFS relaxes the requirements of POSIX and can access data in the file system in streaming form. The core design of the Hadoop framework is: HDFS and Map Reduce. HDFS provides storage for large amounts of data, while Map Reduce provides calculations for large amounts of data [6].

HDFS: Hadoop Distributed File System (HDFS) refers to a distributed file system (Distributed File System) that is designed to run on commodity hardware. It has a lot in common with existing distributed file systems. But at the same time, it is also very different from other distributed file systems. HDFS is a highly fault-tolerant system suitable for deployment on inexpensive machines. HDFS can provide high throughput data access and is very suitable for applications on large-scale data sets. HDFS relaxes some POSIX constraints to realize streaming reading of file system data. HDFS was initially developed as the infrastructure for Apache Nutch search engine projects. HDFS is part of the Apache Hadoop Core project.

HIVE: hive is a Hadoop-based data warehouse tool used for data extraction, conversion, and loading. This is a mechanism that can store, query, and analyze large-scale data stored in Hadoop. The hive data warehouse tool can map structured data files to a database table, and provides SQL query functions, which can transform SQL statements into Map Reduce tasks for execution. Hive has the advantage of low learning cost. It can realize fast Map Reduce statistics through SQL-like statements, making Map Reduce simpler without developing special Map Reduce applications. Hive is very suitable for statistical analysis of data warehouse and Windows registry file [7]. Similar technologies include Spark.

Make full use of the current big data processing technology in the processing of computer massive data, deal with the continuous optimization of the current convergence and processing technology architecture, and reasonably adopt the processing technology under different scenes to deal with different data processing requirements. On the one hand, it can meet the needs of real-time data processing by computers, and on the other hand, it can provide guarantee for the storage of massive data.

4.2. Data Processing Technique
After studying the current mainstream data formats such as log, text, database, http interface, socket interface, etc., how to deploy and collect data in plug-in mode for different data formats is also the
current research direction. At present, there is no common technology or component that can interface the repeated data formats and needs to be customized. At present, log processing currently has filebeat component to deal with mainstream log processing, if the text processing is structured data, only light-weight modification is required, which fully ensures the accuracy of data transmission. Text and database have corresponding mature processing technologies. Ali's database connection pool druid technology can adapt to various databases. Socket and http also have mature technologies for docking.

4.3. Application of Data Encryption Technology
Ensuring data security is not an easy task. With the continuous development of computer technology, big data processing technology is increasingly coping with different application scenarios. The technologies adopted by people are also constantly changing, which directly affects the risk of data leakage, while the processing technologies to deal with data security are more professional and not easy to crack.

The application of data encryption technology to computer network communication can prevent network data from being stolen and copied, thus attracting the attention of many enterprises. At the same time, the enterprise department should also regularly check and kill viruses, carry out backup processing on important data, and adopt encryption technologies of different security levels to provide security for enterprise data. On the research of data encryption technology, the United States and Britain developed relatively early. The application of asymmetric key and symmetric key to network systems can further improve the security of system operation.

5. Summary
The above-mentioned author has described the relevant factors that affect the computer massive data access, massive data analysis, data timeliness and so on. At present, China's computer communication has certain capabilities in data access, data analysis and data timeliness. However, the consistency between data security and data privacy is the foundation of the operation of the country and even enterprises. The problems of data leakage and data embezzlement are gradually prominent, which to some extent affects the effective development of real-time sharing of computer data, data opening and privacy protection. In order to improve the computer's access to massive data, massive data analysis and data effectiveness should have and give full play to its characteristics, and fully improve the computer data security system, data architecture mode, data transmission effectiveness, data consistency and data multi-level fault tolerance system, laying a foundation for the application of computer data value.

References
[1] White T. Hadoop Definitive Guide (Chinese version) [M]. Zeng Dazhen, Zhou Aoying. Trans. Beijing: Tsinghua University Press, 2010.
[2] Du Zhihui. Parallel programming technology of high performance computing——MPI parallel programming [DB / OL]. (2010-08-25) [2011-01-03].
[3] Chen Quan, Deng Qianni. Cloud computing and its key technologies [J]. Development and Application of High Performance Computing, 2009, 29 (9): 2562-2567.
[4] Tanasescu C, Reed T. Data intensive computing [EB / OL]. [2010 23]
[5] Anand R, Jeffrey D. Mining of massive datasets [EB / OL]. [2011-0103]
[6] Xu Jiye, Zhu Jiehua, Wang Haibin, Meteorological Big Data, Shanghai Science and Technology Press, 2018.09, p. 46.
[7] Wan Chuanmei, edited by Xie Zhenglan. Practical solution to HADOOP application development (revised edition): China Railway Press, 2014.08: page 24.