1. Getting to know data collection

1.1 Definition of data collection, application value, division of data type and scope

Hello, classmates. Today, let's learn the definition of data collection, the application value, the type of data, and the division of scope.

1. What is data collection

**(1) Definition of data collection**

Data acquisition, also known as data acquisition, is an interface that uses a device or program to collect data from outside the system and input it into the system. Data collection generally has the following three characteristics:

1) Data collection is mainly automated, and try to get rid of manual input;

2) The collection content is mainly collected in full, and the method of sampling data is eliminated;

3) Diversified collection methods and enriched content, getting rid of the previous method of only collecting basic data.

Data collection is the entrance of big data analysis and a very important link. Therefore, data collection needs to have the following three characteristics:

1) Comprehensiveness: The amount of data has analytical value, and the data surface is sufficient to support the analysis needs.

2) Multidimensionality: The collected data is more important to meet the analysis needs. Flexibility and speed to customize multiple properties and different types of collected data to meet different analysis goals.

3) Efficiency: It includes the efficiency of technical execution, the efficiency of collaboration among team members, and the efficiency of data analysis requirements and goals. That is to say, it is necessary to clarify the purpose of data collection, collect information with problems, and make the collection more efficient and targeted. In addition, the timeliness of the data must also be considered.

**(2) The application value of data**

Data collection is the cornerstone of the big data industry. However, data collection is not easy. The informatization construction of all walks of life, including government departments, is carried out in a closed manner. Massive data is sealed in different software systems. The data sources are diverse, the amount of data is large, and the data is updated quickly.

The focus of data collection is not on the data itself, but on how to truly solve the actual business problems in data operations. However, in order to solve business problems, you must make data collection generate value, and you must do data analysis and data mining. Before data analysis and data mining, we must first ensure the collection of high-quality data. Only by comprehensively and accurately collecting the required data, forming a data flow scale, and then analyzing the data flow, can the analyzed data results have a guiding role in decision-making.

**(3) Type of data**

From the perspective of the type of collected data, the types of data are complex and diverse, including structured data, unstructured data and semi-structured data.

1) Structured data: The most common type of structured data is data with schemas.

2) Unstructured data: refers to irregular or incomplete data structure without a predefined data model, including all formats of office documents, text, pictures, various reports, images, and audio/video information.

13) Semi-structured data: Data between structured and unstructured, such as XML, HTML and JSON are common semi-structured data.

**(4) Division of data collection range**

According to the different subjects of data generation, the scope of data collection mainly includes database collection, system log collection, network data collection, sensing device data collection, etc.

1) Database collection: mainly use relational databases such as MySQL and Oracle, and NoSQL databases such as Redis, MongoDB and HBase for data collection. Enterprises complete the collection of big data by deploying a large number of databases on the collection side and implementing load balancing and fragmentation among these databases.

2) System log collection: It is mainly used to collect a large amount of log data generated by the company's business platform for offline and online big data analysis systems. The log collection system has high availability, high reliability, and scalability. The system log collection tools all adopt a distributed architecture, which can meet the collection and transmission requirements of hundreds of megabytes of log data per second.

3) Network data collection: It is the process of obtaining data information from public websites or APIs through web crawler technology. When using a web crawler, the content on each web page will be obtained from the URL of one or several initial web pages, and the web page will be crawled. In the process of continuously extracting new URLs from the current page and putting them in the queue, until the set stop conditions are met. The crawled data is stored in the local storage system.

4) Data collection of sensing equipment: sensing equipment is a detection device that can sense the measured information, and can convert the sensed information into electrical signals or other required forms of information output according to certain rules to meet the needs of the information. Requirements for transmission, processing, storage, display, recording and control. Various sensors will be installed at the work site, such as pressure, temperature, flow, sound, electrical parameters, etc. The sensors have strong adaptability to the environment and can cope with various harsh working environments.

Students, in this class, we have learned the definition of data collection, the application value, the type of data, and the division of the scope. In the next class, we will continue to learn the data collection process. See you next class!