1. Getting to know data collection

1.2 Data collection process

Hello, students. Today, let's learn the process of data collection.

**(1) The process of data collection**

Data collection can extract unstructured data from web pages or business processing systems, store it as a unified local data file, and store it in a structured way. It supports the collection of pictures, audio, video and other files or attachments, and attachments and text can be automatically associated.

(1) Collection process of web crawling

In the process of web page data collection, it generally needs to go through three steps of collection, cleaning, and storage. The details are as follows.

Step 1: Collect data.

The first step for a web crawler is to locally send a request to the starting URL to get the response it returns to extract the data contained within it. This step is generally implemented through Python.

Extracting data is essentially parsing a web page. Two things need to be done. One is to extract the links on the web page, and the other is to extract the resources on the web page.

1) Obtaining links: In essence, it refers to obtaining links to other webpages that exist on the webpage to be parsed. The web crawler needs to send requests to these links, and so on, until all the specific websites are crawled.

2) Obtaining data: Obtaining data is the purpose of web crawlers. Common data types are listed as follows:

Web page text: HTML, JSON, etc.;

Image: JPG, GIF, PNG, etc.;

Video: MPEG-1, MPEG-2 and MPEG-4, AVI, etc. [/et ˈset(ə)rə]

Step 2: Clean the data.

Data cleaning is a very important step after data collection. Data cleaning can unify the data format, reduce many problems in data analysis, accurately analyze data, and improve data analysis efficiency. In web pages, you can remove some tags that are not related to content, such as styles, scripts, etc.:

Step 3: Store the data.

Storing the data is the last step of the web crawler, and the acquired data can be saved and used for further analysis after proper processing.

Using the relevant knowledge of web crawlers can not only crawl website pages, but also crawl page-related information in apps. Due to the popularity of mobile phones, relevant information in apps is also an indispensable part of big data analysis.

(2) Data collection of log data types

In the data collection process of log data type, ETL plays a very important role. ETL is the process of extracting, cleaning and transforming the data of the business system and then loading it into the data warehouse. In order to achieve better analysis results, provide analysis basis for enterprise decision-making.

Specific steps are as follows:

Step 1: Collect data. The collection is completed by configuring the collection tool (Filebeat, etc.).

Step 2: Clean the data. Receive raw log data through open source tool (Logstash, etc.) configuration, split and verify.

Step 3: Store the data. The collection results are stored in a database or file through open source tools (Logstash, etc.).

**( 2 ) Knowledge and ability system for data collection**

**1) Data analysis ability**

The types of data vary widely. The data directly generated by the software main body include database storage data, database behavior data, Internet application data, business system behavior/status data, operating system behavior data, middleware container behavior/status data, etc.; data, including Internet device data, industrial control device data, multimedia device data, etc. With the rapid development of software and hardware technology in the computer field, the types and scales of various software data and hardware data are rapidly increasing, and the differences are becoming larger and larger.

Different types of data have their own specific formats, standards, meanings, protocol specifications, etc., requiring professionals to have the ability to analyze various data scenarios, and use scientific methods and means to interpret, analyze, collect, clean, storage. Provide sufficient and reliable data raw materials for subsequent data processing and mining.

**2 ) Data collection capability**

Acquisition tools vary by data. According to the differences in sources and characteristics of different data, it is necessary to use software tools and software systems to complete the entire process of data acquisition. For some specific hardware devices, it is also necessary to use corresponding hardware-type acquisition devices to collect hardware data. Data acquisition engineers must have the ability to customize acquisition tools such as software and hardware.

The huge scale of data volume puts forward higher requirements on the performance, efficiency, stability and storage of data collection. In terms of technical depth, data acquisition engineers are required to master higher-end technologies such as high concurrency, clustering, and virtualization.

**3 ) Planning ability of data collection scheme**

Professionals who meet the requirements of the certificate standards will have comprehensive capabilities, and can formulate collection plans, develop and customize collection tools according to the actual on-site data collection environment, and comprehensively utilize collection systems, platforms, and frameworks to complete data collection, preprocessing, and storage. completion process.

Students, in this lesson, we learned the process of data collection, and we will continue to learn the method of data collection in the next lesson . See you next class!