1. Getting to know data collection

1.3 Data collection method

Hello, classmates. Today, let's learn how to collect data together.

**1. method of data collection**

three main methods of data collection :

(1) Manual programming

Python: is a free and open source language that is often compared to R for its ease of use. Unlike R, Python is easy to pick up and easy to use. Simple data crawling can be achieved using the resource library provided by Python.

(2) Open source tools

Filebeat: It is an open source local file log data collector that can monitor the log files in the log directory. During use, the collection process of the general log format can be realized through simple command configuration. Two components are involved in the use process - the finder Prospector and the collector Harvester to read the file and send the event data to the specified output.

Logstash: It is a device that collects logs in a log relationship system. Simply put, Logstash is a pipeline with real-time data transmission capability, which is responsible for transmitting data information from the input end of the pipeline to the output end of the pipeline. The root pipe can also be added with a filter according to your own needs.

Flume: is a distributed, highly reliable, and highly available log collector, mainly used to collect, aggregate and move log data from large batches of different data sources to the data center - distributed file system (Hadoop Distributed File System, HDFS) for storage.

(3) Business tools

Inspur data collection platform: provide a basic scientific research environment for multi-scenario data calculation and analysis and mining, fully combine relevant data of industry topics, and use big data technology to dig and analyze deeply to meet the scientific research needs of industry big data and further improve the university's large-scale data. The level of data scientific research, with the help of a perfect "production-university-research" system, realizes the transformation of scientific research results into business value.

Log Easy: It is a professional log analysis tool. The platform provides powerful and easy-to-use search methods, including Fan Guo query, field filtering, regular expressions, NOT/AND/OR Boolean values, fuzzy matching, etc. And can highlight the query field, locate the context of the log, TB - level massive data can quickly return search results

Octopus Collector: is a free, simple and intuitive web crawling tool that can scrape data from many websites without coding. In order to reduce the difficulty of use, Octopus has prepared a "Simple Website Template" for beginners, covering most mainstream websites on the market. Users can collect data without task configuration when using the simple template. The simple template is established for collecting Xiaobai. Confidence. In addition, you can also set up timed cloud collection, obtain dynamic data in real time, and export data to the database or any third-party platform regularly.

Video demonstration of Octopus collector ]

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