2021 International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA) | 978-1-6654-2490-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/CAIBDA53561.2021.00041

Research on Big Data Analysis Data Acquisition and Data Analysis

Hong Li
Guangdong Peizheng College
Guangzhou, China
Email :honglhappy@126.com

***Abstract*--Using big data analysis algorithm, this paper discusses the source of data acquisition, points out the technical characteristics of data source, and explains the data acquisition methods and requirements. Clear the purpose of big data analysis, build data analysis system, describe the process of data analysis. There are four steps in big data analysis: data acquisition, data storage, data analysis and data mining. Data acquisition can be divided into two parts: acquisition and preprocessing, which is just narrow data acquisition. The data analysis of the Internet of things is a process of organizing and purposefully collecting data, processing data and analyzing data to form information. The analysis process is the support and execution process of the algorithm. In the data processing cycle, data analysis should be used properly in every link from data acquisition in the field perception layer to data transmission in the communication layer to data processing in the application layer, so as to improve the effectiveness and rationality of data processing.**

Keywords: data; Collection; analysis; Research

I. Introduction

The purpose of data collection is to solve the problem of data island. Whether it is structured data, unstructured data or discrete data, without data collection, these data from various sources can only be independent and meaningless. Data acquisition is to write these data into the data warehouse, integrate the scattered data together, and then comprehensively analyze these data.

According to the classification of big data sources, data collection can be divided into several categories

Collection of system file log.

Network big data collection.

Software program access.

Provide database and data file.

Discrete data or random data selection.

A. Collection of system a file log

Many Internet enterprises have their own massive data collection tools, which are mostly used for system log collection, such as chukwa of Hadoop, flume of cloudera, scribe of Facebook, etc. these tools adopt distributed architecture, which can meet the needs of log data collection and transmission of hundreds of MB per second.

For example, scribe is an open source log collection system of Facebook, which can collect logs from various log sources and store them in a central storage system for centralized statistical analysis and processing. Scribe provides a scalable and error tolerant solution for distributed collection and unified processing of logs.

For example, chukwa provides a complete set of solutions and frameworks for the collection, storage, analysis and presentation of large amount of log data, which can be used to monitor the overall operation of large-scale Hadoop clusters and analyze their logs.

B. Network big data collection

This is what we often know as web crawler. In theory, Web Data Collection refers to obtaining data information from web sites through web crawler or website open API. This method can extract unstructured data from web pages, 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 can be automatically associated with the body.

In the Internet era, web crawler is mainly to provide the most comprehensive and up-to-date data for search engines. At present, there are hundreds of web crawler tools, which can be divided into three categories.

Distributed web crawler tools.

The Java Web crawler tool.

Non Java Web crawler tools.

C. Software interface mode

It needs the data interface provided by the supplier of each system to realize data collection and aggregation.

Implementation process:

Coordinate with multi software manufacturer engineers to get to know all system business processes and database related table structure design, and determine the feasibility plan through detailed deliberation;

Code testing, commissioning phase delivery

The data reliability and value of interface docking mode are high, and there is no data duplication in general; The data can be transmitted in real time through the interface to meet the requirements of data real-time application. But the cost of interface development is high; Need to coordinate multiple software vendors, heavy workload and easy to fail; The

978-1-6654-2490-5/21/$31.00 ©2021 IEEE
DOI 10.1109/CAIBDA53561.2021.00041

scalability is not high. For example, due to the new business needs each software system to develop new business modules, the data interface between it and the big data platform also needs to be modified and changed, and even to overturn all the previous data interface codes, which is heavy workload and time-consuming.

D. Database and data file supply

Data file data

Database records data

Data stored in a dataset

E. Discrete data or random data selection

Randomly selected data and discrete data

Big data statistics is applied in the process of data analysis. Through the use of probability theory to establish mathematical model, collect the data of the observed system, carry out quantitative analysis, summary, and then infer and predict, provide the basis and reference for relevant decision-making. The basic theories of statistics include: probability limit theory and its application in statistics, tree probability, spatial probability, random, Poisson approximation, random network Markov process and field theory, Markov convergence rate, Brownian motion and partial differential equation, limit of space branch population, large deviation and random median, cross boundary problem in sequential analysis and time series analysis, one-to-one correspondence between Markov process and Dirichlet table, central limit theorem in function estimation, stability problem of limit theorem Causality and statistical inference, prediction inference, network inference, likelihood, estimator and maximum likelihood estimation, precise approximation in parametric model, adaptive method in nonparametric estimation, new content in multivariate analysis, theory and application of time series, nonlinear time series, comparison between deterministic model and stochastic model in time series, extreme value statistics, Bayesian calculation Change point analysis, random estimation, measure value processing, function data statistical analysis.

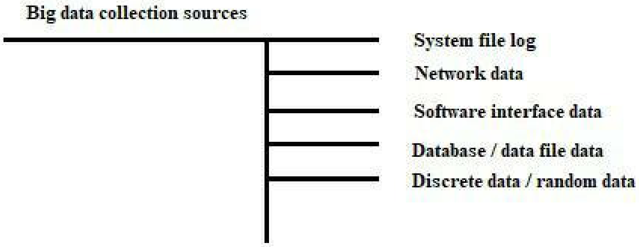
Big data collection sources
System file log
Network data
Software interface data
Database / data file data
Discrete data / random data

Figure 1 Schematic diagram of big data collection sources

II. Data processing and analysis process

Data analysis is the hot spot of IT application. The focus of data analysis lies in the data technology itself, including data storage cost, data storage method and data processing technology. The storage level of data storage is constantly improving. Get new data and store it locally or in the cloud. The improvement of data volume promotes the upgrading of data processing technology, from the previous single machine data reading to the generation of system data architecture, greatly improving the magnitude and speed of data processing. When the system invested a lot of time and resources, established a complete set of data processing system, obtained the data required by the system. The value of data essence is the core reason of data processing.

Correlation analysis of big data is a statistical analysis method to study the correlation between two or more random variables in the same status. For example, between height and weight; The correlation between relative humidity and rainfall is a problem of correlation analysis. The difference between correlation analysis and regression analysis: regression analysis focuses on the dependence between random variables, so that one variable can be used to predict another variable; Correlation analysis focuses on finding various correlation characteristics among random variables. Correlation analysis has applications in industry and agriculture, hydrology, meteorology, social economy and biology.

Regression analysis of big data, in statistics, regression analysis refers to a statistical analysis method to determine the quantitative relationship between two or more variables. Regression analysis can be divided into single regression and multiple regression according to the number of variables involved; According to the number of dependent variables, it can be divided into simple regression analysis and multiple regression analysis; According to the relationship between independent variables and dependent variables, it can be divided into linear regression analysis and nonlinear regression analysis.

Principal component analysis of big data is a statistical method. Through orthogonal transformation, a group of variables that may have correlation are transformed into a group of linearly unrelated variables, which are called principal components.

The description statistics of big data is a method to sort out and analyze the data by charts or mathematical methods, and estimate and describe the distribution state, digital characteristics and the relationship between random variables.

The description statistics are divided into three parts: centralized trend analysis, out of center trend analysis and correlation analysis.

The analysis of concentration trend mainly depends on the statistical indicators of average, medium and mode to show the trend of data concentration. For example, what is the average score of the subjects? Is it positive or negative? The trend analysis of the middle distance mainly depends on the statistical indexes such as the total distance, the quarterdifference, the average difference, the variance (covariance: the statistics used to measure the relationship between two random variables), and the standard deviation to study the data out of the middle trend.

Cluster analysis is a multivariate statistical classification. Regression analysis is a statistical method to determine the quantitative relationship between two or more variables. Principal component analysis (PCA) transforms the possible correlated variables into linearly uncorrelated variables.

Data analysis is the basis of data management system of Internet of things. The data analysis system should evaluate its effectiveness by analyzing the following problems when appropriate. Data analysis is a complex process. It is a process of checking, cleaning, transforming and modeling data. The purpose is to find useful information, draw conclusions and promote decision-making. The general analysis process can be divided into the following steps:

Clear analysis purpose → comb business to form analysis ideas → build analysis index system → collect data → process analysis data → output data information

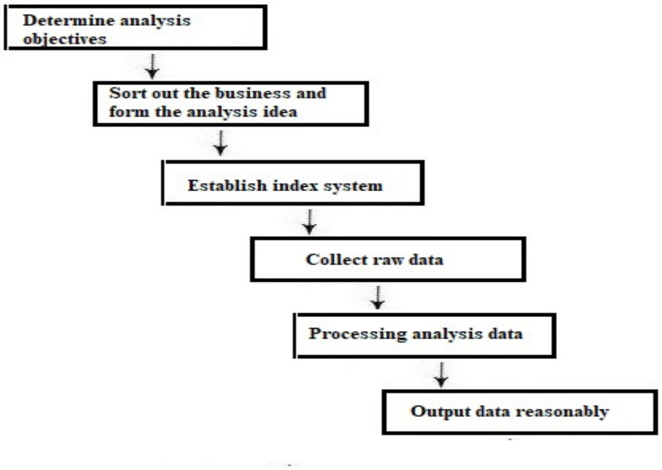

Figure 2 Schematic diagram of big data analysis steps

III. Data analysis

A. Clear analysis purpose and business

Analysis should be purposeful and directional. Is it to analyze a problem now, or to sort out the overall business status, or to forecast and monitor a certain index in the future. In short, it is to solve doubts, monitor and predict, with the purpose of improving efficiency and gain.

After clarifying the purpose, we need to sort out the ideas, how to sort out? If it is to analyze the reasons for the general decline in sales in the past month, it is necessary to restore the progress of the whole thing from the bottom up. The purchase process involves trading volume, customer unit price and discount rate, and then is divided into various products; The browsing link involves the browsing volume, PV / UV; The user dimension includes loss rate, activity, repurchase rate and so on. The analysis purpose is divided into several different analysis points, and then the analysis method and specific analysis index are determined for each analysis point.

B. Build analysis index system

To build an analysis index system is to analyze the whole body, systematize the analysis framework, make clear what index each point is, and any analysis path can correspond to the index (of course, there will not be such a perfect system in reality). Take e-commerce as an example, follow the thinking logic of people and goods yard. Common business analysis scenarios include sales, commodities, channels, competing products, members and so on. Commodities can be further subdivided into inventory, profit and related sales analysis. In the whole business analysis system, ensure the systematization, that is, what to analyze first, then what to analyze, so that each analysis point has a logical connection, so that the analysis results are convincing.

C. To acquire and collect data

SQL is the most basic database language, no matter from what database, data warehouse, big data platform, all need to master. Hive and spark are both based on big data. Hive can map structured data files to a database table, and quickly realize simple MapReduce statistics through SQL like statements.

D. Cleaning and data processing

The original data comes from various business systems. If the indicator caliber is not correct, there will always be inconsistent, repetitive, incomplete (the attribute of interest has no value), error or exception (deviation from the expected value) data. All of this can be done through data cleaning: removing noise and irrelevant data.

Data integration: combine data from multiple data sources and store them in a consistent data store

Data transformation: transform the original data into a form suitable for data mining

Data reduction: data cube aggregation, dimension reduction, data compression, numerical reduction, discretization and concept stratification.

E. Reasonable data information output

analysis results using output template or report tools. Through building a platform to complete the targeted data analysis and display.

Data integrity: a database is a general data processing system for an enterprise or an application field. What it stores is a collection of relevant data belonging to enterprises and institutions, groups and individuals. The data in the database is established from the overall point of view, which is organized, described and stored according to a certain data model. Its structure is based on the natural relationship between data, so it can provide all the necessary access paths, and the data is no longer for a certain application, but for the whole organization, with the overall structural characteristics.

Data sharing: the data in the database is established for many users to share their information, which has got rid of the restrictions and constraints of specific procedures. Different users can use the data in the database according to their own usage; Multiple users can share data resources in the database at the same time, that is, different users can access the same data in the database at the same time. Data sharing not only meets the requirements of users for information content, but also meets the requirements of information communication between users.

IV. Conclusion

In the process of building the Internet of things system model, the most important part is the data processing in the system process. The sensor layer of the Internet of things collects a large amount of data, which needs to be processed. The purpose of data processing is to transform the original data into useful information. Data is the input or raw value of data processing. The output of data processing is information, and the output can be presented in different forms, such as text file, data file, chart, spreadsheet or image.

Reference

[1] J.P.Sun , (2015)Coal mine accident analysis and coal mine big data and Internet of things [J]. Industrial and mining automation.

[2] Wu X.F.Wu ,(2017) New thinking on Internet of things and big data [J]. Communication world.

[3] G.R.Bian ,(2017)uangrong. Application of military Internet of things in ammunition support of ordnance warehouse [J]. Packaging engineering,.

[4] Y.JZhang,(2017)ajuan. Review on information security and privacy protection of Internet of things [J]. Logistics technology.

[5] Y.Huang,(2020) Cloud computing technology of computer network [J]. Journal of Jiamusi vocational college.

[6] J. Yang ,(1990)Application of artificial intelligence technology in separation process synthesis [D]. Beijing University of chemical technology.