УДК 339.138:004.9

BASIC ELEMENTS OF THE TECHNICAL CONTENT OF BIG DATA ON THE INTERNET

Wang Xu, postgraduate student, midixiaozi@163.com
Kievich A.V., Doctor of Economics, Professor, a.v.kievich@yandex.ru
Polessky State University
Ван Сюй, аспирант
Киевич Александр Владимирович, д.э.н., профессор
Полесский государственный университет

Annotation. The article states that focusing on the whole life cycle process of information resources, the basic technology of big data on the Internet includes big data collection technology, big data pre-processing technology, big data storage and computing technology, big data analysis and mining technology, etc.

Keywords: big data, information architecture, volumetric storage, new discoveries, business challenges.

The development of Internet big data technology system has been constantly enriched and perfected, and the integration and convergence with other information and communication technologies, such as Internet of Things and artificial intelligence, is now more mature. Focusing on the whole life cycle process of data resources, the basic technology of Internet big data includes big data collection technology, big data preprocessing technology, big data storage and computing technology, big data analysis and mining technology, etc. [1].

Internet big data collection technology.

Internet big data collection is the first part of the big data life cycle. With the development of various types of technology and applications, there are various sources of data, including numerous unstructured databases, the Internet of Things and so on, in addition to traditional relational databases [2]. Data types are also richer and richer, including the original structured data, more or semi-structured data and unstructured data. According to the different sources of data, big data collection techniques and methods also differ greatly, here we are in accordance with the database data collection, network data collection, Internet of Things data collection classification description.

Database data collection.

Database data collection varies depending on the type of database and whether the type of data stored in it is structured or unstructured.

For traditional relational databases, ETL (data extraction, transformation and loading) tools, SQL coding, ETL tools and SQL coding combination of three ways. ETL tools after years of development, has formed a relatively mature product system, especially for traditional relational databases, typical representatives include: Oracle's OWB, IBM's Datastage, Microsoft's DTS, Informatica and so on. With the help of ETL tools, database data can be quickly collected and pre-processed, shielding the complex coding tasks, which can improve the speed and reduce the difficulty, but lack of flexibility. Database data acquisition through SQL coding is more flexible than using ETL tools, which can improve the efficiency of data acquisition and preprocessing, but the coding is complex and requires high technical requirements [3]. The combination of ETL tools and SQL coding can combine the advantages of the first two methods, which can greatly improve the speed and efficiency of data acquisition and preprocessing.

For unstructured database collection and data transfer between different types of databases, the current use of more open source projects to provide ETL tools, typical representatives include: Sqoop, Kettle and Talend, etc., designed for big data, can take into account the offline and real-time data collection, you can achieve mainstream unstructured databases (eg, HDFS, HBase and other) and traditional relational databases (eg, MySQL, Oracle, PostgreSQL, etc.) between two-way data transfer. HDFS, HBase and other mainstream NoSQL databases) and traditional relational databases (e.g., MySQL, Oracle, PostgreSQL, etc.) to deliver data in both directions.

Comparatively speaking, database data has high value density and is mainly collected through log files, system interface functions, etc. The collection technology is standardised and there are more available tools, and the unified collection technology for different types of databases will become an important development trend in the future.

Network Data Acquisition.

Network data collection can be subdivided into two categories based on the type of data collected: Internet content data collection and web log collection.

Internet content data collection is mainly the process of acquiring content data from websites by using web crawler technology and the open application programming interface (API, Application Programming Interface) of websites and other means, supplemented by the comprehensive use of the word splitting system, task and indexing system to realise content data acquisition from websites. This way can extract semi-structured data and unstructured data from web pages on the Internet and store them as unified local data files in a structured way, supporting the collection of pictures, audio, video and other files or attachments as well as automatic association. Web crawler is a kind of programme or script that automatically crawls Internet content according to certain rules. Web crawler technology was first mainly used in search engines, Internet search engines and web page holders through the Robots agreement between what information can be crawled, what information should not be crawled.

Web log collection is currently used more open source log collection system, typical representatives include: Flume, Scribe, Logstash, Fluentd, etc. Flume is an open source log collection system project contributed by Cloudera to Apache, with high availability, high reliability and distributed features, which can achieve real-time and dynamic collection and transmission of massive logs. Scribe is Facebook's open source log collection system project, with scalability and high fault-tolerance features, can achieve the

distributed collection of logs and unified processing [4]. Logstash deployment is relatively simple to use, pay more attention to the pre-processing of log data, you can do a good job for the subsequent log parsing pad.Fluentd deployment and Flume Flumed deployment and Flume is relatively similar, scalability is very good, the application is also quite extensive.

IoT Data Acquisition.

Whether it is consumer IoT, industrial IoT, or smart city IoT, it may involve RFID tags, positioning devices, infrared sensing devices, LIDAR, and a variety of sensors and other devices, and it can be said that the role of the IoT terminal equipment is to collect IoT data, which may involve the collection of various types of data, such as sound, light, heat, electric current, pressure, location, and biometrics. IoT data involves a wide range of data, relatively dispersed data, huge differences in data types, and large differences in data collection methods and collection means [5].

Let's separately highlight the big data preprocessing methods themselves.

The data needed for big data analysis and mining is often collected through multiple channels with multiple types of data, and the data collected through the above big data collection techniques often have data quality problems such as data redundancy, data missing values, data conflicts, etc. It is necessary to improve the data quality through big data preprocessing techniques to make the data more in line with the needs of analysis and mining in order to ensure the correctness and validity of the big data analysis and to obtain high-quality The data quality needs to be improved through big data preprocessing technology to make the data more in line with the needs of analysis and mining, so as to ensure the correctness and effectiveness of big data analysis and obtain high-quality results. Big data preprocessing technology can clean, fill, smooth, merge, specification and consistency check operations on the collected raw data, transforming the messy raw data into a relatively single and easy to deal with structural types, laying the foundation for the later big data analysis and mining. Big data preprocessing mainly includes: data cleaning, data integration, data conversion and data protocol.

Data cleansing.

Data cleaning is mainly through the detection of redundancy, errors, inconsistencies and other problems in the data, using a variety of cleaning techniques to de-noise the data to form a consistent data set, including the removal of duplicate data, fill in the missing data, eliminating the noise data and so on. Removing duplicate data generally uses statistical analysis methods such as similarity calculation. There are two ways to deal with missing data, one is to ignore incomplete data, i.e., clearing missing data, and the other is to fill in missing data through statistical methods, classification or clustering methods to ensure data availability. In practice, the data collection process will also generate a large amount of noise data (outside the reasonable data domain) for a variety of reasons, which, if left unprocessed, will result in inaccurate and unreliable results of subsequent analysis and mining [6]. Commonly used methods to eliminate noisy data include statistical and mathematical methods such as binning, clustering, and regression.

The main data cleaning tools include the aforementioned open source ETL tools such as Sqoop, Kettle and Talend, as well as open source ETL tools such as Apache Camel, Apache Kafka, Apatar, Heka and Scriptella. In addition, Potter's Wheel, a data cleansing tool that is highly interactive and integrates bias detection and data transformation, is also used.

Data integration.

Data integration refers to the merging of heterogeneous data from multiple data sources into a consistent database. This process mainly involves schema matching, data redundancy, detection and processing of data value conflicts, and the main tools are still the open source ETL tools mentioned above. Schema matching is mainly used to discover and map the attribute correspondence between two or more heterogeneous data sources, and machine learning algorithms such as plain Bayes and stacking are more widely used in schema matching. Data redundancy may come from the inconsistency of data attribute naming, and Pearson product-moment correlation coefficient can be used to measure the consistency of data attribute naming, and the larger the absolute value indicates that the correlation between the two is stronger. Data value conflict is mainly manifested as the same entity with different data values from different sources, which sometimes needs to be supplemented by manual determination rules to deal with the data value conflict problem.

Data transformation.

Data transformation is the process of dealing with inconsistencies in the collected data, including the transformation of data names, granularity, rules, data formats, units of measurement, etc., as well as the combination of new data fields, segmentation and other transformations. Data transformation actually also includes data clarity related work, which requires the cleaning of inconsistent data according to business rules to ensure the accuracy of subsequent analysis results. The main purpose of data transformation is to transform the data into a form suitable for analysis and mining, and the choice of data transformation methods depends on the big data analysis and mining algorithms. Commonly used transformation methods include: function transformation, the use of mathematical functions for each attribute value mapping; data normalisation, proportional scaling of the data attribute values, as far as possible to fall into a smaller specific interval. Normalisation helps in the implementation of various classification and clustering algorithms, but also avoids over-reliance on units of measure, and circumvents the problem of weight imbalance.

Data protocol.

Data reduction refers to the premise of maintaining the original appearance of the data as much as possible, looking for the most useful features to reduce the size of the data, the maximum streamlining of the data, involving techniques and methods mainly include high-dimensional data dimensionality reduction processing methods (dimensional reduction), instance statute, discretisation techniques, and machine learning algorithms such as unbalanced learning. Data statute technology can be used to get a statute representation of the data set, making the data set smaller, but at the same time still close to maintaining the integrity of the original data, which can improve the efficiency of analysis and mining under the premise of ensuring the accuracy of analysis and mining. At present, the data generalisation technique based on massive data has become one of the important issues in big data preprocessing.

The technologies of big data storage and computation themselves should be emphasized separately.

Big data storage and computing technology is the foundation of the whole big data system. There are two main types of current big data system architectures: one is the MPP database architecture, and the other is the layered architecture of the Hadoop system. These two architectures have their own advantages and corresponding applicable scenarios. In addition, with the development of fibre-optic network communication

technology, the big data system architecture is developing towards the architecture of storage and computing separation and cloud architecture [6].

MPP.

MPP (Massively Parallel Processing) architecture. MPP architecture is the connection of multiple processing nodes through the network, each node is an independent machine, the processing unit within the node exclusive use of its own resources, including memory, hard disk, IO, etc., that is, each node within the CPU can not access the memory of another node. The MPP architecture servers need to implement complex scheduling mechanisms as well as parallel processing processes through software. Focusing on industry big data, it adopts the Shared Nothing architecture, through column storage, coarse-grained indexing and other big data processing technologies, combined with the highly efficient distributed computing model of the MPP architecture, it completes the support for analytical applications, and the operating environment is mostly low-cost PC Server with high performance and high scalability, which is widely used in the field of enterprise analytics applications. Application.

These MPP products can effectively support PB-level structured data analysis, which is beyond the capabilities of traditional database technology. For the new generation of enterprise data warehouse and structured data analysis, the better choice is MPP database.

Hadoop.

Hadoop is a distributed systems infrastructure developed by the Apache Foundation. Users can develop distributed programs without knowing the underlying details of the distribution. It makes full use of the power of clusters for high-speed computing and storage. It is reliable, efficient and scalable.The core of Hadoop is HDFS and MapReduce.

HDFS (Hadoop Distributed File System), is the foundation for data storage management in the Hadoop architecture. It is a highly fault-tolerant system that detects and responds to hardware failures and is used to run on low-cost general-purpose hardware. HDFS simplifies the file consistency model and provides high-throughput application data access capabilities through streaming data access, suitable for applications with large datasets. It provides the mechanism of write once read many times, data in the form of chunks, simultaneously distributed on different physical machines of the cluster.

MapReduce (Distributed Computing Framework) is a distributed computing model for performing computations on large data volumes. It shields the details of the distributed computing framework and abstracts the computation into two parts: map and reduce, where map performs specified operations on independent elements of a data set to generate intermediate results in the form of key-value pairs, and reduce performs statutes on all the values of the same key in the intermediate results to obtain the final result. "MapReduce is very suitable for data processing in a distributed parallel environment consisting of a large number of computers.

Around Hadoop derived from related big data technology, to deal with traditional relational database is more difficult to deal with the data and scenes, such as for the storage and calculation of unstructured data, etc., make full use of the advantages of Hadoop open source, along with the continuous progress of related technology, its application scenarios will be gradually expanded, and the most typical application scenario is to achieve the most typical application scenario is to achieve the extension and encapsulation of Hadoop for The most typical application scenario is to extend and encapsulate

Hadoop to achieve support for Internet big data storage and analysis. There are dozens of NoSQL technologies here, and they are also being further subdivided. For unstructured and semi-structured data processing, complex ETL processes, complex data mining and computational models, the Hadoop platform is better at it. At present the mainstream choice is the distributed architecture, and in the distributed architecture system, Hadoop can be said to be a more mature and stable big data platform system has been tested, so many enterprise big data platforms, are based on Hadoop to build.

Thus, it can already be argued that with the emergence of Big data technology, technology will become increasingly integrated with artificial intelligence technology to empower computing systems with the ability to understand, reason, discover and make decisions about data so that they can derive more accurate and deeper insights from data and extract the value behind it [7].

With the development of artificial intelligence, it becomes possible to mine useful information and form knowledge in massive data sets, and machine systems gradually acquire cognitive ability, which stimulates the development of cognitive computing. Cognitive computing is the product of the continuous development of artificial intelligence, including natural language processing, speech recognition, computer vision, machine learning, deep learning, robotics and so on. As long as people realize the close relationship between big data and analytics, they will realize that cognitive computing is as indispensable to big data analysis as analytics, and the importance of cognitive computing will be recognized more and more.

Список использованных источников

1. ULR: [Электронный ресурс] – Режим доступа: https://www.oracle.com /cis/big-data/what-is-big-data/#how – Дата обращения: 25.09.2024 г.

2. Wang Xu., Kievich A.V. The main trends in the digital economy and finance that shape the current landscape and vector of development of industries / Wang Xu., A.V. Kievich // Economy and Banks. 2024. № 1. С. 42-51.

3. Ван С., Киевич А.В. Анализ основных социальных сетей и их возможностей для SMM-продвижения в Республике Беларусь / Ван Сюй, А.В. Киевич // Устойчивое развитие экономики: состояние, проблемы, перспективы : сборник трудов XVII международной научно–практической конференции, Пинск, 28 апреля 2023 г. : в 2 ч. / Министерство образования Республики Беларусь [и др.] ; редкол.: В.И. Дунай [и др.]. – Пинск : ПолесГУ, 2023. – Ч. 1. – С. 13–17.

4. ULR: [Электронный ресурс] – Режим доступа: https://eecs.csuohio.edu /~sschung/cis612/CIS612_Lecture1_IntroBigDataAnalyricsCloud.pdf. - Дата обра-щения: 28.09.2024 г.

5. Ван С., Киевич А.В. Цифровой рубль в России – третья форма денег в виде цифрового кода / Ван Сюй, А.В. Киевич // В сборнике: Банковская система: устойчивость и перспективы развития. Сборник научных статей четырнадцатой международной научно-практической конференции по вопросам финансовой и банковской экономики. Пинск, 2023. С. 29-34.

6. ULR: [Электронный ресурс] – Режим доступа: https://www.mokosmart. Com/ru/iot-in-industry-4-0/. – Дата обращения: 21.09.2024 г.

7. Zhang, L., & Liu, W. (2021). International cooperation in AI and digital finance under the Belt and Road Initiative. Global Journal of Economics and Business, 9(4), 45-58.