International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 4, Number 2, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v4n2.36
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Application Status of Hadoop in Data Cloud Computing

Pinpeng Jiang *

Teensen Academic High School, Nanchang, 330007, China
*Corresponding Author: jpp070307@163.com

ABSTRACT

This paper summarizes the Hadoop platform and its application in network public opinion analysis, and emphasizes the advantages of Hadoop in big data storage and processing, as well as the development of cloud computing technology and its importance in data collection, processing and analysis. This paper also discusses the core technologies based on Hadoop and Spark, including keyword retrieval, web crawler and text mining, which together constitute a comprehensive toolkit for analyzing and managing online public opinion.

KEYWORDS

Hadoop; Spark; Cloud computing; Network public opinion

1. OVERVIEW OF HADOOP PLATFORM

Hadoop platform can effectively distribute massive data through programming model. At present, the main sub-project of Hadoop platform is HDFS. With HDFS, large-scale data can be effectively stored, and it has the characteristics of high fault tolerance and can be accessed on a large scale through high throughput.It is especially popular in the internet industry, supporting such as Yahoo! Advertising and search, data analysis of Facebook, search log mining of Baidu, e-commerce data processing of Taobao, and data analysis service of China Mobile. Despite the competition from technologies such as Apache Spark, Hadoop and its ecosystem are expected to coexist with Spark for a long time due to their unique advantages in data storage, data warehouse and real-time processing. The market predicts that the Hadoop market will grow at an annual growth rate of 16.10% from 2019 to 2029, with the fastest growth in the Asia-Pacific region and the largest market in North America. In 2023, the global market scale has reached 64.485 billion yuan, and it is expected to increase to 152.193 billion yuan in 2029. Many well-known enterprises, including IBM, Adobe and A9.com, as well as Baidu, Alibaba, Tencent and Huawei in China, have widely applied Hadoop technology in their big data strategies. Hadoop plays an important role in key fields such as data mining, log analysis, data warehouse and recommendation engine. Looking forward to the future, Hadoop will remain the key technology of big data processing and continue to play its value in many industries [1].

2. THE OVERALL APPLICATION OF COMPUTING IN NETWORK PUBLIC OPINION

2.1. Application Status of Cloud Computer

Cloud computing technology has been applied in China's civil technology and other fields, and will become the future storage trend. The storage structure will replace the traditional storage mode, and the log data of online learning behavior will be stored in the corresponding database in different data

formats. As the key direction of information technology, cloud computing is rapidly growing in the global market by 2024, and it is expected to reach the trillion-dollar scale by 2027. The cloud computing market in China is also growing rapidly, and the market scale has expanded significantly in 2022. Technically, Serverless architecture and cloud native technology promote the simple deployment of applications and the digital transformation of enterprises. Cloud security has become particularly important because of the cloudization of enterprise IT, and cloud native security technology has gradually occupied the market center. At the same time, the integration of AI and machine learning enhances the data processing and intelligence of cloud services. Facing the increase of network attacks, enterprises pay more and more attention to cloud security, and cloud native security has become a new field of industry development. Looking forward to the future, cloud computing will tend to be safer, smarter and more environmentally friendly.

2.2. Collecting Data

As the initial link of the network public opinion keyword monitoring process, data collection plays a foundation role. In the process of data collection, it is necessary to flexibly adopt corresponding collection methods according to different data sources [2]. At present, the main data information of online public opinion mostly comes from major online platforms, such as Tik Tok, Baidu, Sina Weibo and so on. When data information comes from news websites, we mainly use Nutch tools to collect data. Because the distributed system shows high efficiency and stable performance in actual operation, we usually choose to use distributed Nutch in this case. As for the data collection of Weibo website, it mainly relies on API interface, and ensures that the client can access related applications through the real authorization of Weibo platform.

2.3. Processing Data

Data processing is a link between the preceding and the following, and it is very important to realize the monitoring of online public opinion keywords. The current technology can not directly process the collected data, so it is necessary to take digital processing measures. In the process of processing, because the domestic network public opinion data is mainly Chinese data, which is different from English data processing, it is necessary to strengthen the preprocessing of Chinese word segmentation. In addition, the construction of text vector space model is also an important part of data processing, which contains basic elements such as word frequency, meaning, part of speech and title, and sets corresponding weight ratios for different types of characteristic words. After data preprocessing, effective data clustering is needed, corresponding clustering modules are set up, and hierarchical clustering algorithm is used for orderly processing. When using hierarchical clustering algorithm, the factors such as stability in dealing with high-dimensional data, dependence on parameters and anti-interference should be taken as practical considerations to ensure that the algorithm can play a practical role [2].

2.4. Analysis Data

The third step is data analysis, which is the decisive link to realize network public opinion keyword monitoring. In the whole network public opinion monitoring system, public opinion analysis module is the core part. Only when the public opinion analysis module can run stably can we monitor the network public opinion keywords effectively [2]. There are three main forms of online public opinion keyword monitoring. The first is keyword monitoring of sensitive topics. In order to avoid the influence of sensitive keywords on society, the monitoring system will make full and effective matching with the help of sensitive word thesaurus. When it is found that the keywords spread by the network obviously match the words in the sensitive word thesaurus, the system can monitor them in time. The second is the keyword monitoring of hot topics. In view of the current hot topics in society, the hot topics, hot articles and hot comments spread on the Internet are effectively analyzed by data

clustering technology, and their attention is counted respectively, and they are arranged in order according to the numerical value, so as to accurately and effectively identify the hot topics in society in a certain period. The third is content tendency monitoring, which is based on the information publisher's own subjective feelings to get the information publisher's personal position and attitude about information content. At the same time, with the help of data clustering technology, emotional words are effectively matched and accurately calculated according to the corresponding weights. This monitoring method can deeply explore the emotional tendency of information publishers and provide more abundant information for public opinion analysis.

3. THE CORE TECHNOLOGY OF INTERNET PUBLIC OPINION BASED ON HADOOP AND SPARK

3.1. Hadoop System

Hadoop ecosystem consists of many components, such as HDFS, MapReduce, HBase, and other intelligent big data tools, as well as data migration tool Sqoop, data collection tool Flume and coordination service Zookeeper [3]. This system can operate efficiently on Linux operating system and support multiple programming languages. Hadoop is famous for its high reliability, fault tolerance and scalability, and is specially designed to store and process large-scale data sets for in-depth analysis. Therefore, Hadoop has become a popular big data processing solution. As a conventional method to solve the shortage of Hadoop cluster resources, it is necessary to expand hardware resources, such as adding servers or storage units. Hadoop environment, with its advantages in distributed data computing, provides great convenience for developers, enabling them to easily develop applications without having to grasp the underlying details of storage. Modern stream computing frameworks, such as Spark, are designed to handle large-scale data streams, and have high-efficiency real-time data processing capabilities, which is in sharp contrast with those old stream computing systems that rely on event-driven models and have limited processing capabilities. The advantage of streaming computing systems is that they can quickly analyze real-time data streams in memory, and can immediately deliver the analysis results to users or store them for emergencies.

3.2. Keyword Retrieval

There is a huge amount of data on the Internet, so comprehensive retrieval is not only time-consuming, but also difficult to ensure the accuracy of the results. Keyword retrieval technology has become the dominant method of information retrieval because of its high efficiency and timeliness. In the network public opinion monitoring, this technology plays a core role, and its implementation methods are diverse: extraction based on preset semantic rules, statistical analysis using big data, and application of machine learning algorithms. In order to improve the practicability of keyword retrieval, we can simplify the operation process by the following steps: first, deeply analyze the information content, select keywords with more close meaning, and build a standardized thesaurus; Then, these keywords are processed to extract keywords from the text; Finally, screening is carried out according to the weight of keywords to ensure the relevance and accuracy of the final results[4].This process optimizes the application of keyword retrieval technology and improves its efficiency and effect in network public opinion monitoring.

3.3. Web Crawler Technology

Web crawler is an automatic software that collects information on the Internet according to preset rules. It is very important for keyword search engines, which can identify and extract keywords from web pages efficiently, and it is an important tool for network public opinion monitoring. The starting point of a crawler is a series of seed URLs, which usually point to the homepage of a portal or forum. When performing a task, the web crawler first adds these seed URLs to the queue to be processed,

and then accesses the URLs in the queue one by one to grab the web page content and store it locally. After that, the crawler will parse these pages and identify and collect links to other pages. Web crawler technology is mainly divided into two types: universal and focused. Universal web crawler is suitable for a wide range of application scenarios and is common in large search engines. Focused web crawlers focus on specific topics or fields and are suitable for special situations [4].

3.4. Text Mining Technology

Text mining is a key technology in the field of data mining, which consists of three core links: text preprocessing, text classification and text clustering. Text preprocessing is the cornerstone of the whole process, which plays a decisive role in the efficiency and accuracy of mining. It includes Chinese word segmentation, that is, dividing continuous Chinese characters into independent words; And text feature representation, that is, converting text information into formats that can be recognized by computers, such as probability model and space vector model. Text classification relies on supervised learning algorithms, which can identify the categories of text. Classification algorithm classifies texts accurately by analyzing their contents and meanings. Text clustering adopts unsupervised learning algorithm, which can automatically group texts and make similar texts gather together [4]. Generally speaking, through these three links, text mining can effectively extract information from a large number of texts, classify and cluster them to support deeper data analysis and knowledge discovery.

4. CONCLUSION

Hadoop platform and its components can be used to distribute and store massive data sets with high fault tolerance and accessibility. The core of these technologies is modern frameworks such as Hadoop system and Spark, aiming at processing large-scale data with high reliability and real-time processing ability. Keyword retrieval plays a core role in efficiently extracting relevant information from massive data sets, and it adopts technologies such as big data analysis and machine learning algorithm. Web crawler automatically collects web page information, and general and special crawlers serve different needs, while text mining includes preprocessing, classification and clustering to mine deeper insights from text data. Generally speaking, these technologies form a comprehensive toolkit for analyzing and managing Internet public opinion.

REFERENCES

[1] Jharikar, P.B., Raut, S.N., Patil, A.R. et al. (2017) Public Opinion Analysis Using Hadoop, International Journal on Recent and Innovation Trends in Computing and Communication, 5: 91-96.

[2] Hu, Z.Y., (2020) Analysis on the Monitoring System of Internet Public Opinion Keywords Based on Hadoop, Public Relations Rorld, 80-82.

[3] Akshi, K. (2022) News and Public Opinion Multioutput IoT Intelligent Modeling and Popularity Big Data Analysis and Prediction, 3567697.

[4] Bai, R. (2019) Design of Network Public Opinion Monitoring System Based on Cloud Computing and Hadoop. Jingdezhen College, Electronic Design Engineering, 27:141-150.