Journal of Information Hiding and Privacy Protection
DOI:10.32604/jihpp.2020.010223
Article

Tech Science Press

Design and Implementation of Log Data Analysis Management System Based on Hadoop

Dunhong Yao[1,2,3,*] and Yu Chen[4]

[1]School of Computer Science and Engineering, Huaihua University, Huaihua, 418000, China

[2]Key Laboratory of Wuling-Mountain Health Big Data Intelligent Processing and Application in Hunan Province Universities, Huaihua, 418000, China

[3]Key Laboratory of Intelligent Control Technology for Wuling-Mountain Ecological Agriculture in Hunan Province, Huaihua, 418000, China

[4]School of Computer Science and Engineering, Yulin Normal University, Yulin, 537000, China

[*]Corresponding Author: Dunhong Yao. Email: dh_yao@hhtc.edu.cn

Received: 12 July 2020; Accepted: 25 July 2020

Abstract: With the rapid development of the Internet, many enterprises have launched their network platforms. When users browse, search, and click the products of these platforms, most platforms will keep records of these network behaviors, these records are often heterogeneous, and it is called log data. To effectively to analyze and manage these heterogeneous log data, so that enterprises can grasp the behavior characteristics of their platform users in time, to realize targeted recommendation of users, increase the sales volume of enterprises' products, and accelerate the development of enterprises. Firstly, we follow the process of big data collection, storage, analysis, and visualization to design the system, then, we adopt HDFS storage technology, Yarn resource management technology, and gink load balancing technology to build a Hadoop cluster to process the log data, and adopt MapReduce processing technology and data warehouse hive technology analyze the log data to obtain the results. Finally, the obtained results are displayed visually, and a log data analysis system is successfully constructed. It has been proved by practice that the system effectively realizes the collection, analysis and visualization of log data, and can accurately realize the recommendation of products by enterprises. The system is stable and effective.

Keywords: Log data; Hadoop; data analysis; data visualization

1 Introduction

In recent years, with the continuous increase in the number of Internet users, the data on the Internet is accumulating in geometric multiples, and human society has entered the era of big data. These large amounts of data play an important role in the development of enterprises. If an enterprise can effectively analyze the rich data resources, then effectively promote business cooperation and effectively implement personalized services according to the analysis results, it can make the development of enterprises faster.

Log data is one of the most important information on Internet data [1]. Most of these log data are generated by users browsing, searching, and clicking on network lines, which have no fixed structure and are complicated. Therefore, before processing these data, we need to design its architecture, and we need to use clustering to store the architecture data so that it can be stored for a long time. By analyzing these large quantities of log data and realizing the visualization of data, we can obtain users' usage habits, which can be used for enterprises to implement targeted product recommendations.

2 System Design

The process design of big data processing has four links: Data collection, data storage, data analysis, and data application [2]. Each link is closely related. They play different roles and play different roles in the process of big data processing.

2.1 Data Collection

In the process of data collection, data sources affect the integrity, accuracy, authenticity, consistency, and security of data. Data collection can be divided into two situations, one is the offline collection, and the other is the real-time collection. Data collection techniques include the log collection system flume and the publish-subscribe messaging system Kafka [3]. We store the collected data on the Hadoop cluster [4].

2.2 Data Storage

We collected the original log data is stored on a Hadoop cluster [5], the cluster is storage space will be more than one storage aggregate into comprehensive server storage, it broke the traditional storage method, the log data can be random storage to the cluster nodes, in this way, it can fully guarantee the performance of the storage space and disk utilization, when access to the data, the visitor can be read directly from the cluster, and do not need to care about the data on which node, it also improves the performance of concurrent access.

2.3 Data Analysis

Data analysis is an important part of the process of big data processing. The log data we collect generally has no data structure. Data cleaning can extract, convert, and load the data without structure, so that it can be transformed into a structure known by programmers, and then stored in HDFS [6] or database. The analysis includes simple query analysis, flow analysis, and deeper analysis. The deep analysis of big data is mainly based on large-scale machine learning technology [7].

2.4 Visualization

Data visualization is an effective display of the data. Enterprise users can intuitively see the fluctuation of the data, which is conducive to monitoring the market demand of enterprises and preparing for the further application of the data. Therefore, visualization is a very necessary step.

3 Hadoop Deployments

We need to make some simple configuration before deploying the Hadoop cluster, including the network IP address, the installation of JDK and Hadoop, the setting of SSH non-secret login, and so on.

3.1 Preparation

(1) VMware installation

We install the virtual machine VMware [8] into the Linux system through image file loading. The common image files are CentOS, Ubuntu, etc. In this paper, the Ubuntu image is used. Its installation sequence is to create a new virtual machine first, then load the image file, and finally complete the creation of the virtual machine. After these tasks are completed, we can double-click to open the newly created virtual machine and make the necessary network configuration.

(2) Network configuration

We first check the VMnet [9] IP of this machine, and then open the virtual network editor on the virtual machine, so that we can set the DHCP and NAT of VMnet.

(3) Configure the IP process

In the process of configuring the network IP, we first turn off the firewall and then proceed with the configuration shown in Tab. 1.

Table 1: Configure IP flow

Step	Setting
Start the created virtual machine, enter the startup interface to enter the password, log in to the Hadoop user, then open the terminal, switch to the root user, add the Hadoop user exemption permission, and enter the command line nano /etc/sudoers	Add hadoop ALL = (ALL:ALL) NOPASSWD:ALL
Modify hostname	s100
Modify DNS parsing in the file /etc/hosts, configure network mapping	127.0.0.1 localhost 192.168.137.100 s100 192.168.137.101 s101 192.168.137.102 s102
Modify the hosts in the window system	127.0.0.1 localhost 192.168.137.100 s100 192.168.137.101 s101 192.168.137.102 s102
Set a static IP address in the /etc/network/ interfaces file	auto eth0 iface eth0 inet static address 192.168.137.100 netmask 255.255.255.0 gateway 192.168.137.2 dns-nameservers 192.168.137.2
Restart the network	Command Line
View configuration	ifconfig
Check if the IPod can be connected by ping	ping www.baidu.com

(4) SSH secret login

When using SSH to log in to the server, we need to input IP address, port, user name, password, and other information, which is troublesome and easy to enter wrong. We can achieve password-free quick login through the configuration parameters of the client and server. We save the public key of the server that the client USES to verify the identity of the client, eliminating the need to enter a password, and similarly, the client can simplify the login command by configuring the server parameters.

(5) JDK and Hadoop

When configuring Java's JDK, we first unzip the installation package, use the PWD command to view the installation path, and copy the path, then set the Java home path to the copied path of the directory file /etc/environment, and then type the source /etc/environment command to make the environment effective.

Finally, we execute the command to check the JDK version to verify whether the configuration is successful.

Next, we configure the Hadoop environment, and when Hadoop is installed, we configure its installation path into the environment, which is similar to the path that configured the JDK. The file path is represented by the HADOOP_HOME variable, and the path is added to the path variable.

3.2 Local Mode

Local mode opens only one virtual machine with four core files. When we tested the local mode, we first formatted the name node with the command HDFS name node–format and then started the process of start-all. Sh for viewing. In addition to Jps, there are five other processes in the process, namely Namenode, SecondaryNamenode, DataNode, NodeManager, and ResourceManager, whose structure is shown in Fig. 1. If all five nodes are started successfully, the local pattern is tested successfully.

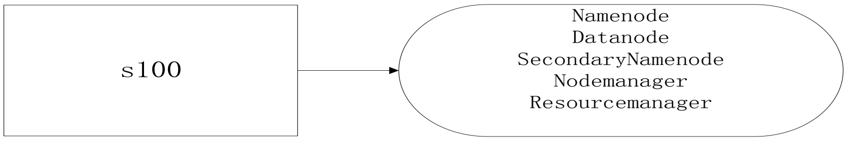

Figure 1: Local mode

3.3 Distributed Mode

There are two types of distributed clusters, one is pseudo-distributed and the other is fully distributed [10]. The pseudo-distributed is similar to the local mode, fully distributed also to require configuration of four core configurations, if the local mode is changed into the cluster mode, then the configuration will automatically obtain the directory, and its structure is shown in Fig. 2.

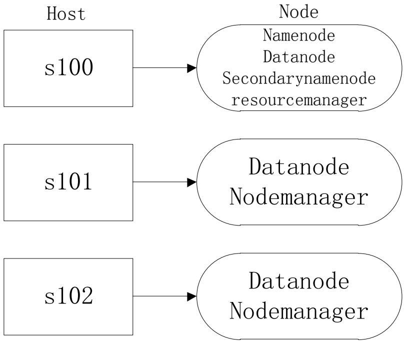

Figure 2: Fully distributed

4 Data Warehouse Hive

Figures and tables should be inserted in the text of the manuscript.

4.1 Hive Modes

Hive has three configuration methods, including embedded Derby, Local, and Remote connection, each configuration has its characteristics but generally used the remote configuration MySQL method.

4.2 Remote Connection MySQL

The specific operation method is to first copy the jar driver, then initialize hive's metadata, and finally test whether it can be started normally. If the startup is successful, it proves that the remote configuration is successful, and then it can interact with MySQL. The configuration is shown in Tab. 2.

Table 2: Hive-site.xml configuration

Name	Value
javax.jdo.option.ConnectionDriverName	com.mysql.jdbc.Driver
javax.jdo.option.ConnectionURL	jdbc:mysql://192.168.137.1:3306/myhive
javax.jdo.option.ConnectionUserName	root
javax.jdo.option.ConnectionPassword	123456

5 Load Balancing Nginx

Nginx loads balancing can distribute all HTTP requests to each machine [11], give full play to the performance of all machines, and improve the quality of service and user experience. In processing millions of log messages, after the asynchronous framework of Nginx processes concurrent requests, it first distributes them to the background server for complex calculation, processing and response, so that it can easily expand the background server when the business volume increases, to maintain stability of the bearing server performance. Fig. 3 is the schematic diagram of configuring the load balancing server.

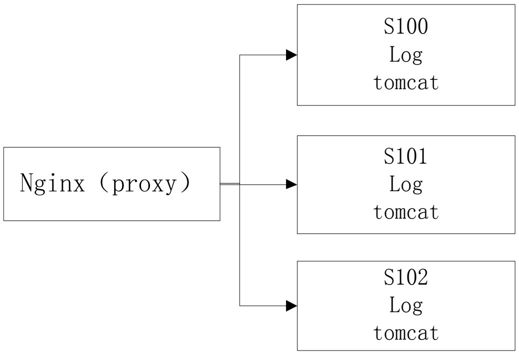

Figure 3: Flowchart for configuring the load balancing server

6 Using Hadoop to Analyze Log Data

To realize log data analysis, we set up a stable Hadoop cluster according to the above configuration requirements, and then used flume to collect log data of the website, because it was configured into a cluster mode, to distribute log data in various directories. If the data volume was relatively large, Kafka needed to be configured separately. When these configurations are complete, we can start the cluster and analyze the log data.

6.1 Data Analysis

After successfully collecting a large amount of log data, we stored it in the items folder on the distributed file system HDFS, named access.log.10. There are nearly 1 million log data in this file, each log data represents a user's access behavior, and the canonical log data structure is shown in Tab. 3.

Table 3: Log data structure

Field	Description
Address	Used to record the user's IP address
User	Used to record the user name of the visiting user
Time	Record the time of user access
Request	Records the URL requested by the user
Status	Record whether the request was successful
products	Record every product that has been accessed
Referrer	It is of great reference value to record the information that is browsed by jumping from other web pages
User agent	Servers that record user visits, such as amazon, are usually the last place in the log data

Then we use MapReduce technology, a computing framework, to cut and analyze these logs, and view the data according to different requirements.

6.2 Data Visualization

In the data visualization part [12], we can use spring to integrate the hibernate framework and combine Echart for data presentation. Echart is a visual chart library that provides simple, beautiful pie charts, line charts, bar charts, and more. On the front end, we usually use this to visualize the page, so we can see the fluctuation of log data. By visualizing some results of log data analysis, pie chart and line chart as shown in Fig. 4 and Fig. 5 can be obtained.

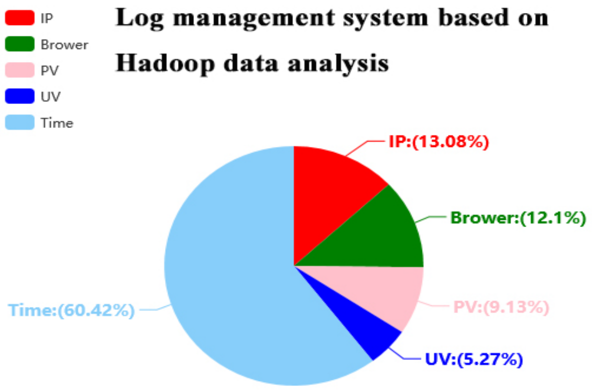

Figure 4: Pie chart

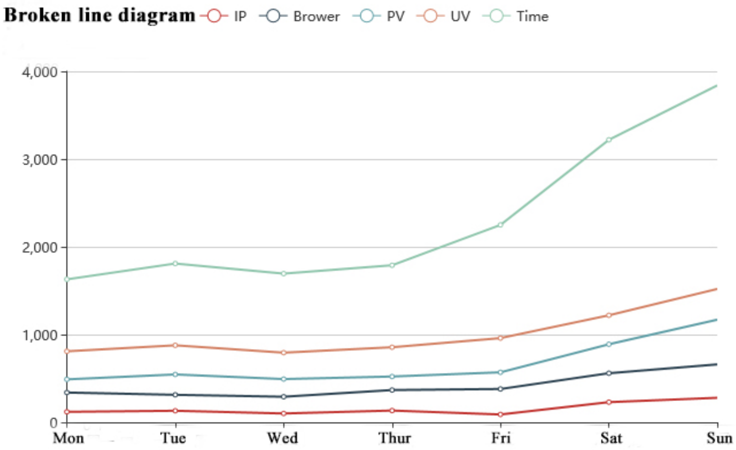

Figure 5: Broken line diagram

From Fig. 4 and Fig. 5, we can intuitively see the time when the number of people that surf the Internet every day is concentrated, and it is also obvious that weekends are the time when most people browse information, especially Sunday, and the other time is relatively less. Therefore, we conclude that these log data may be the data of office workers, so it can be targeted to recommend products of them at this time.

7 Conclusions

Every log on the Internet has a large amount of information. Log big data is data without structure. Before processing, it needs to construct a certain structure, and then put the data onto a well-constructed data warehouse. For analysis and visualization of log data, we build a Hadoop cluster of log data processing, the hive graphs processing technology, data warehouse technology analysis results, such as to visual results show again, realized the log data analysis system, the system can effectively realize data collection, data analysis, and data visualization, and stable running, the effect is good.

Acknowledgement: We would like to thank the Huaihua University Double First-Class initiative Applied Characteristic Discipline of Control Science and Engineering. It provides experimental equipment and environmental support for this work. I would also like to thank the reviewers and editors. We successfully completed the format requirements of the manuscript with their helpful advice and comments.

Funding Statement: This work was supported by the Huaihua University Science Foundation under Grant HHUY2019-24.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] R. Iváncsy and I. Vajk, "Frequent pattern mining in web log data," *Acta Polytechnica Hungarica*, vol. 3, no. 1, pp. 77–90, 2006.

[2] J. Dittrich and J. A. Quiané-Ruiz, "Efficient big data processing in Hadoop MapReduce," in *Proc. of the VLDB Endowment*, vol. 5, no. 12, pp. 2014–2015, 2012.

[3] J. Kreps, N. Narkhede and J. Rao, "Kafka: A distributed messaging system for log processing," in *Proc. of the NetDB*, vol. 11, pp. 1–7, 2011.

[4] S. Singh, R. Garg, P. K. Mishra, "Performance optimization of MapReduce-based Apriori algorithm on Hadoop cluster," *Computers & Electrical Engineering*, no. 67, pp. 348–364, 2018.

[5] Z. Ren, X. Xu, J. Wan, W. Shi and M. Zhou, "Workload characterization on a production Hadoop cluster: A case study on Taobao," in *2012 IEEE Int. Sym. on Workload Characterization*, pp. 3–13, 2012.

[6] D. Borthakur, "HDFS architecture guide," *Hadoop Apache Project*, no. 53, pp. 1–13, 2008.

[7] J. Lin, and A. Kolcz, "Large-scale machine learning at twitter," in *Proc. of the 2012 ACM SIGMOD Int. Conf. on Management of Data*, pp. 793–804, 2012.

[8] Liang, Sheng and G. Bracha, "Dynamic class loading in the Java Virtual Machine," *ACM Sigplan Notices*, vol. 33, no. 10, pp. 36–44, 1998.

[9] H. Wu, Q. Luo, P. Zheng and L. M. Ni, "VMNet: Realistic emulation of wireless sensor networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 18, no. 2, pp. 277–288, 2007.

[10] Y. Kim, T. Araragi, J. Nakamura and T. Masuzawa, "A distributed and cooperative NameNode cluster for a highly-available Hadoop distributed file system," *IEICE Transactions on Information and Systems*, vol. 98, no. 4, pp. 835–851, 2015.

[11] G. Cybenko, "Dynamic load balancing for distributed memory multiprocessors," *Journal of Parallel and Distributed Computing*, vol. 7, no. 2, pp. 279–301, 1989.

[12] D. Keim, H. Qu and K. L. Ma, "Big-data visualization," *IEEE Computer Graphics and Applications*, vol. 33, no. 4, pp. 20–21, 2013.