Real Time Analysis System of User Behavior from the Perspective of Big Data

Zhiqiang Cai¹ and Jiaai Zhang¹,*
¹Jilin University of agricultural science and technology, 132109, JiLin, China
*Corresponding author e-mail: 572619424@qq.com

Abstract. Nowadays, the speed of data generation is growing rapidly. The analysis and mining of data on enterprise websites can bring business value to enterprises. There are many ways for users to access a website. What the real-time analysis system of user behavior from the perspective of big data needs to do is to collect all the access ways of users, and finally get a real-time dynamic effect map through technical storage, statistics, analysis. Through the analysis of the effect map, we can know which way users visit the website and which way is the main one, so as to serve as a reference for enterprise product promotion. In particular, because the resulting renderings are dynamic and real-time, with the passage of time, the amount and percentage of each access mode will change, from which we also can know the change trend of access mode, so as to adjust the promotion strategy in time.

Keywords: Big Data, Popularize, Real-time Analysis, Spark

1 Introduction
Big data computing technology perfectly solves the problems of collection, storage, calculation and analysis of massive data. The era of big data opens another era for human society to use data value. Through the analysis and mining of massive data information, we can extract the user behavior information which has important value to the enterprise and bring business value to the enterprise. User behavior analysis refers to the statistics and analysis of the data we want to know under the condition of obtaining the historical data or real-time data of the website, finding out the relevance information we want to know, and unifying these relevance information with our promotion and marketing, etc., which provides a strong basis for further changing the promotion mode.¹,²,³,¹⁰

User behavior analysis is a new method of user research in recent years. Many people have no deep understanding of it, and there are many loopholes in the theory. However, the academic department has already focused on it as the main means to study consumer psychology.

This system is based on big data, and real-time user behavior analysis. Through collecting how users enter the website, then statistics, storage, analysis, and finally display in the form of charts, enterprises can understand their own promotion situation through chart comparison, know which promotion is necessary and which promotion is cost-effective, so as to carry out Effective promotion decision. In particular, because the resulting renderings are dynamic and real-time, with the passage of
time, the amount and percentage of each access mode will change, from which we can know the change trend of access mode, so as to adjust the promotion strategy in time. Based on spark architecture, the system collects a large number of data generated by users visiting the website, processes the data in depth, mines the relevant user behavior attributes, constructs the user behavior analysis model, and effectively explores the specific user behavior application and displays the final results. The completed work includes data collection, caching, storage, retrieval and effect display. Common user behavior analysis scenarios are as follows: (1) user access times in different time periods, (2) what chains do users access to the website more frequently, (3) what are the differences between the forms of registered users and visitors reading the website, (4) which regions are more willing to visit the website.[10]

2 Requirement Analysis

2.1 Project Overview
Develop a real-time analysis system of user behavior from the perspective of big data. Through real-time processing and analysis of the data generated by how the user enters the target website, you can know the quality of the external chain on the day in real time and make targeted promotion changes in time.

2.2 Project Requirements
The system is mainly implemented from the following requirements: (1) design and development based on big data. (2) data operation must be real-time. (3) the results of data analysis should be more humanized and comparative.

2.3 Introduction to Development Software

2.3.1 Flume Flume is a distributed log collection system, which collects data from each server and sends it to a specified location, for example, to HDFS as shown in Figure 2-1. Simply speaking, flume collects logs.

![Figure 2-1. Flow Chart of Flume Operation Mechanism](image)

Agent is composed of source, channel and sink.
Source: obtain data and transfer the received data to one or more channels.
Channel: channel is a temporary storage container, which caches data through the event format received from the source until they are consumed by the receiver, acting as a bridge between the source and the sink. This ensures consistency in getting and sending data. It can link to any number of sources and sinks. There are many types supported.
Sink: stores data in centralized storage, which consumes data from the channel and passes it to the destination. The target may be another receiver or HDFS, Hbase, etc.

Flume is mainly an agent. There are two interactions outside the agent, one is to accept the input of data, the other is to output data. Sink’s task is to send the data to other specified libraries. After receiving the data, the source sends the data to the channel as a data buffer, temporarily stores the data, and then sends the data from the sink to the specified location, such as HDFS, etc. note: the temporary
data can only be deleted after the receiver successfully sends the data from the channel. This also ensures the reliability of data transmission.[8]

2.3.2 Kafka Kafka is a distributed flow processing platform. In Kafka version 0.8x, Kafka is mainly a distributed, divisible log service system with replica number.

Kafka belongs to the Apache organization and is a high-performance distributed publish subscribe message queuing system. Its main features are: ① high throughput, a single server can achieve a throughput of several hundred thousand/s. ② support message division between servers, support distributed consumption, and ensure message order in each partition. ③ light weight, support real-time data processing and offline data processing.

2.3.3 Zookeeper Zookeeper is a distributed and open-source distributed application coordination service, which provides simple and original functions. Based on the distributed functions, the distributed application can realize more advanced services, such as distributed synchronization, configuration management, and queue management.[7]

Zookeeper is the manager of the cluster. It monitors the status of each node in the cluster and performs the next reasonable operation according to the feedback submitted by the node. Finally, a simple and easy-to-use interface and a stable, high-performance system are provided for users.

There are two main functions of zookeeper: ① data storage system: znode system is a complete data view on each node. ② monitoring mechanism: monitoring the change of stored data. Enable services to respond quickly.

3 Conclusions

3.1 Overview of Big Data and User Behavior Analysis
In the actual production environment, because the data in the project is generated by the user’s click in behavior, which requires the click of many users. The processing method to obtain these click data is to bury points on the corresponding external chain, and then carry out ETL operation on the obtained data, and then analyze and process these ETL data.

After data collection, it is necessary to collect data through flume, and then flow to Kafka for data caching. By writing code of sparkstreaming, the data in Kafka can be consumed in real time, and then stored in HBase database. Because HBase needs coordinated selection of zookeeper and storage address entry, here we also need to install zookeeper. HBase is distributed and stored in itOf course, the data of should also be stored in the distributed system. Here we choose Hadoop HDFS. Finally, we use spring boot to write the corresponding code to take out the data in HBase and use echarts to show the effect. As shown in Figure 3-1, it is the architecture flow chart of our project.[2,4,5,6,7,9]

![Figure 3-1. Architecture Flow Chart of the Project](image)

4 System Detailed Design
The system is mainly divided into five parts: data collection, log collection, message cache, data
real-time storage and retrieval, data display. Next, the five parts are introduced respectively.

4.1 Data Acquisition
Because the data in the project is generated by the user’s click in behavior, which requires the click of many users. The processing method to obtain the click data is to bury the points on the corresponding external chain, and then carry out ETL operation on the obtained data, and then analyze and process the data passing through ETL.

4.2 Data Acquisition
Log collection is to collect the data generated by users. Here, flume is used for implementation. Three parameters need to be set: source, channels, and sink.Exec is selected for source.type, because the previous collected data is stored in one log.log every 0.3s. The file can be monitored through exec’s tail-f. a piece of data is entered in log.log, and the system collects one piece of data. You may have a question: the data has been stored in the log.log. Why do you need to collect it again? In fact, it’s wrong. Because it’s based on big data, there can’t be only one server in the real environment, and it must be distributed. In the case of multiple servers, the data in these log.log files need to be collected again, and the data needs to be cached, flume’s sink can do this. Channels. Type we choose memory, while sink. Type is Kafka.

4.3 Message Cache
Caching the collected data can ensure that the data will not be lost. When consuming, it can also determine which consumption and which consumption are not. Kafka has done a good job in these aspects. The method of this system is to create a topic of category in Kafka, and store all data flowing from flume in the same topic. When starting java code that stores data in HBase, Kafka will consume from the topic of category. Of course, which topic to consume from is specified in Java code.[10]

4.4 Real Time Storage and Retrieval of Data
When it comes to real-time data, we choose spark streaming, which is widely used at present. Data storage code is mainly composed of two parts. One is to set the relevant parameters of Kafka and write the consumption data code. The other is to store the consumption data in HBase. The difficulty here is to convert a whole piece of data into the format we want to store. For example, such a piece of data:

ver = 1 & en = e-PV & PL = website & SDK = JS & B-RST = 1920 * 1080 & UD=12gh4079-223e-4a57-ac60-1a04d8f7a2f&L=zhCN&USD=8e9559b3-da35-44e1-AC98-85EB37 D1F263&c_time=1525104000&p_url=http://list.mytest.com/www/6/-..html.

Suppose 6 corresponds to Baidu, and the format is rowkey 2018-05-01 and cell (Baidu, 1).

Data extraction is still complex. To create a scanner, a prefix filter is installed on the scanner, and then the table in HBase is sent to the scanner for scanning, and then the rowkey is segmented. After a series of operations, the results suitable for display can be obtained.

4.5 Data Display
Data display is the ultimate goal of this project. A series of operations of the system is to display the final effect of data. Through the visual comparison of charts, we can understand the quality of the external chain, so as to make better promotion. Here, spring boot + echarts are used to display the renderings.

5 Test

5.1 Data Display
After a series of work, the data display diagram as shown in Figure 5-1 can be obtained.
6 Epilogue
In this Internet era, enterprises must learn to use big data to analyze and mine the data information on the website, extract valuable information, and find new profit sources and growth points. In this paper, the real-time analysis system of user behavior from the perspective of big data is to use big data computing technology to analyze the user behavior of the website in real time. Real time statistics of the proportion of users choosing the form of entrance and chart display, to understand which ways users use to visit the website, to know which access ways are the main and which are becoming more and more popular, so as to adjust the promotion strategy in time.

Acknowledgements
National Undergraduate Science and technology innovation and entrepreneurship training program(201911439025).

References
[1] Luo X, Wang J, Shen Q, et al. User behavior analysis based on user interest by web log mining. 2017 27th International Telecommunication Networks and Applications Conference (ITNAC). IEEE Computer Society, 2017.
[2] Kadam S D, Motwani D, Vaidya S A. Big data analytics-recommendation system with Hadoop Framework. 2016 International Conference on Inventive Computation Technologies (ICICT). IEEE, 2017.
[3] Liu, Zixia, H. Zhang, and L. Wang. "Hierarchical Spark: A Multi-Cluster Big Data Computing Framework." 2017 IEEE 10th International Conference on Cloud Computing (CLOUD) IEEE, 2017.
[4] Amini, Sasan, I. Gerostathopoulos, and C. Prehofer. "Big data analytics architecture for real-time traffic control." 2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS) IEEE, 2017.
[5] Yadav, Vinit. Working with HBase. Processing Big Data with Azure HDInsight. Apress, 2017.
[6] Kreps, Jay. "I ♥ Logs: Apache Kafka and Real-time Data Integration - O'Reilly Media Free, Live Events." (2018).
[7] Artho, Cyrille, et al. "Model-Based API Testing of Apache ZooKeeper." 2017 IEEE International Conference on Software Testing, Verification and Validation (ICST) IEEE, 2017.
[8] Birjali, Marouane, A. Beni-Hssane, and M. Erritali. "Analyzing Social Media through Big Data using InfoSphere BigInsights and Apache Flume." Procedia Computer Science 113(2017):280-285.
[9] BarbaGonzález, Cristóbal, GarcíaNieto, José, AldanaMontes, José F, et al. Multi-Objective Big Data Optimization with jMetal and Spark. 2017.
[10] Ashayer A, Yasrobi S, Thomas S, et al. Performance Analysis of Hadoop Cluster for User Behavior Analysis. 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS). IEEE, 2018.