ISSN: 2688-8653

Design and Implementation of E-commerce User Behavior Analysis System Based on Spark

Akang Zhang*, Zeyu Zang and Ling Wu

School of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu, Anhui, China.

* Corresponding Author: 1803622124@qq.com

Abstract

E-commerce user behavior data contains a lot of valuable information, and mining and analyzing massive data to obtain user behavior laws is particularly important for e-commerce enterprises. Using Spark technology, the e-commerce user behavior analysis system is designed, and the user behavior data is processed from four aspects: Top10 analysis of popular categories, Top3 analysis of commodities in various regions, website conversion statistics, real-time analysis for real-time advertising click flow statistics, and introduces the design ideas and processes of the system in detail. The system can help e-commerce platforms analyze user behavior data, obtain interesting business indicators and enhance risk control, which has good practical application value.

Keywords

E-commerce; User behavior; Spark; Offline analysis; Real-time statistics.

1. Introduction

With the rapid development of the Internet, online shopping has gradually become the choice of most people. According to the Statistical Report on the Development of Internet in China [1] published by China Internet Information Center on September 29, 2020, as of June 2020, the scale of online purchasing users in China has reached 749 million. When users visit e-commerce websites, a series of behavior records will be generated, including browsing, clicking, purchasing, joining shopping carts, etc. E-commerce websites will generally persist these data. E-commerce user behavior data contains a large amount of valuable information. Mining and analyzing the large amount of data to obtain user behavior rules and make decisions is particularly important for commercial enterprises [2]. Document [3] uses Spark, a distributed framework based on memory iterative computing, combined with distributed file storage technology in Hadoop framework, and uses K-Means++ clustering algorithm to classify e-commerce user session behavior data. Document [4] Designs a user behavior analysis system fusion framework based on Spark platform and using multiple MLlib mining algorithms. Document [5] makes use of a large amount of data storage, processing and retrieval capabilities to analyze and query user data. Literature [6] shows that XGBoost prediction model based on Spark platform has better prediction accuracy and stability than traditional machine learning algorithm model. Document [7] uses Spark and Hadoop Distributed File Systems to analyze e-commerce user behavior. On the basis of the above research, this project uses the latest Spark technology to design an e-commerce user behavior analysis system, which conducts offline and real-time analysis of various user behaviors (access behavior, web browsing behavior and advertising click behavior). Offline analysis includes Top10 analysis of popular categories, Top3 analysis of commodities in various regions, website conversion statistics, real-time analysis for real-time advertising click flow statistics, totaling 4 topics.

ISSN: 2688-8653

2. Related Technologies

In the development of big data projects, clusters composed of many big data ecosystem components will be used, such as Hadoop cluster, ZooKeeper cluster, Spark cluster, etc. These clusters are interconnected and interdependent to form a complete big data system. The four characteristics of big data: scale, diversity, high-speed and value, namely the so-called "4V", are also realized through the collaborative operation of these clusters. The project also uses the components of the big data ecosystem described below.

2.1. Linux operating system

Linux is a free and open source UNIX-like operating system. It is also a multi-user, multi-task, multi-thread and multi-cPU operating system based on POSIX and Unix. Linux system has become the first choice of servers due to its advantages of stable performance, efficient performance of firewall components, and simple configuration.

2.2. Hadoop cluster

Hadoop is a distributed computing platform, which mainly solves the storage, analysis and calculation of massive data. It is composed of Hadoop Distributed File System (HDFS), MapReduce and Yarn. Hadoop provides users with a distributed infrastructure with transparent details at the bottom of the system.

2.3. ZooKeeper cluster

ZooKeeper is a distributed, open source distributed application coordination service and an important component of Hadoop and HBase. It is a software that provides consistency services for distributed applications. Its functions include configuration maintenance, domain name service, distributed synchronization, group service, etc.

2.4. Spark cluster

Spark is a cluster computing framework for real-time data processing. Its main feature is memory computing. The Spark ecosystem is mainly composed of Spark Core, Spark SQL, Spark Streaming and other components.

2.5. HBase cluster

HBase is a column-oriented distributed storage database. The operation of HBase depends on Hadoop and ZooKeeper. HBase uses HDFS as its file storage system and ZooKeeper as its distributed application coordination service. At the same time, storing the metadata information of HBase cluster can provide the automatic failover function for HBase cluster to ensure the high availability of HBase cluster.

2.6. Kafka cluster

Kafka is a distributed publish and subscribe information system based on ZooKeeper with high throughput, which can process all the action flow data of consumers in the website. Generally, Kafka is used to build a data pipeline between systems or applications to convert or respond to real-time data, so that the data can be calculated in a timely manner and the corresponding results can be obtained.

2.7. Phoenix query engine

MyBatis, a data persistence layer framework that supports SQL queries, is used in the data visualization part of the project. However, the HBase database used in the project does not support JDBC access and SQL statement queries, which leads to that the data visualization system we built cannot directly use the MyBatis framework to access the HBase database.

ISSN: 2688-8653

Therefore, we need to use the Apache Phoenix query engine to make HBase support access through JDBC, And convert SQL queries into HBase related operations.

3. Data Set Analysis

The same data set is used in Top10 analysis of popular categories and Top3 analysis of popular categories in various regions, including seven fields: unique value used to identify user behavior (user_session), user behavior type (event_type), product category ID (category_id), user ID (user_id), product ID (product_id), area (address_name), and specific time (event_time).The data set used in the website conversion rate statistics contains four fields: the user's time to visit the page (actionTime), the unique value used to identify the user's behavior (sessionid), the user's page ID (pageid), and the user ID (userid).The data set used in the real-time statistics of ad click stream contains four fields: time stamp, user ID, ad ID and city.

4. Design and Implementation of the System

4.1. Top10 analysis of popular categories

When users visit e-commerce websites, they usually generate many behavioral events for products, such as viewing, adding to shopping cart and purchasing. The implementation idea of Top10 analysis of popular categories is shown in Figure 1: ① First read the data. In order to facilitate the subsequent aggregation processing, accumulate the value values of the same key, calculate the total number of different behaviors in each category, and then count the number of times that the products in each category are added to the shopping cart, purchased and viewed. Here, the behavior type and category ID are used as the key, and the value 1 is used as the value; ② Count the times of viewing, adding to shopping cart and purchasing each category; ③ Consolidate the values with the same Key value in order to combine the viewing times, shopping cart adding times and purchase times of the same category into one line; ④ Create a custom sorting class. The sorting rules are sorted in descending order according to the number of items viewed, added to the shopping cart and purchased in each category. Use Spark's secondary sorting; ⑤ Obtain the top 10 categories and implement persistence processing; ⑥ Encapsulate the program as a jar package, and upload the program to the big data cluster environment to run the program in Spark on YARN mode.

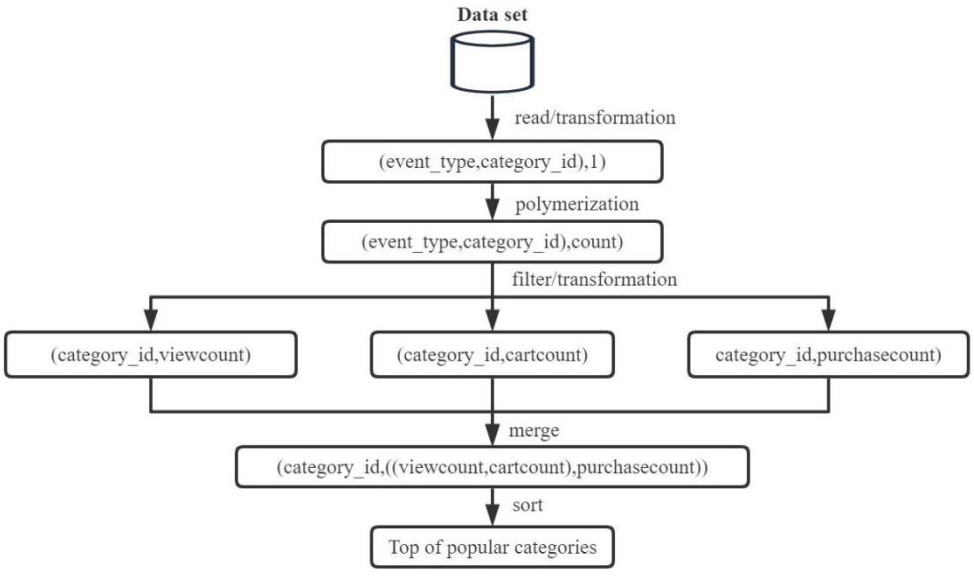


Figure 1. Top10 analysis process of popular categories

ISSN: 2688-8653

4.2. Top3 analysis of popular categories by region

When a user visits an e-commerce website, the website will store the data of the region where the user triggered the action through the IP address or location information. Get the most popular products in each region by counting the number of times different products are viewed in each region. This part will analyze the user behavior data stored on e-commerce websites, so as to count the top 3 popular products in each region. The implementation idea is as shown in Figure 2: ① Read the area name, behavior type and commodity ID data in the data set, and filter out the data with the behavior type of view, so that the value values of the same key can be accumulated during subsequent aggregation processing. Here, we need to convert the data format, and use the area name and commodity ID as the key, and the value 1 as the value. Get the data whose user behavior type is viewing goods; ② Use the area name as the Key, the product ID and the number of times the product is viewed as the Value, and then group the converted data according to the Key to count the number of times each product is viewed; ③ The statistical results are grouped according to the region, and the data in each group are sorted in descending order; ④ Obtain the first three data of each group and implement persistence processing; ⑤ Encapsulate the program as a jar package, and upload the program to the big data cluster environment to run the program in Spark on YARN mode.

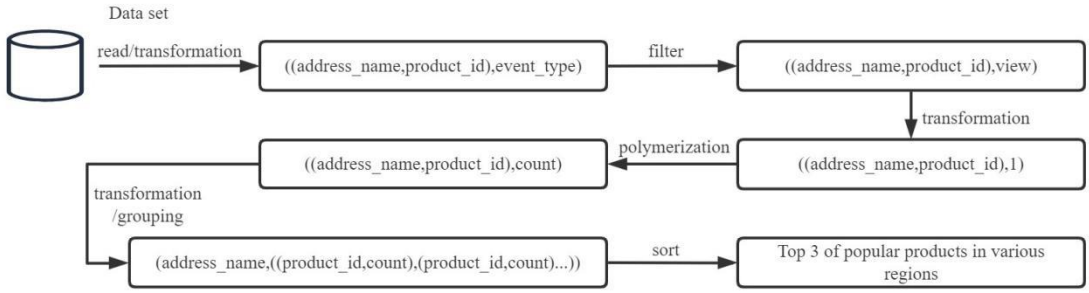


Figure 2. Top3 analysis process of popular products in different regions

4.3. Website conversion rate statistics

The website conversion rate refers to the ratio of the number of visits to the total number of visits that users have made corresponding target actions. The corresponding target actions mentioned here can be a series of user actions such as user login, user registration, user browsing, user purchase, etc. Page single-hop conversion rate is a statistical form of website conversion rate. It optimizes the page layout and marketing strategy by counting the page single-hop conversion rate, so that users who visit the website can browse the website at a deeper level. The implementation idea is shown in Figure 3: ① Read the time of user accessing the data set, the unique value of user behavior, the ID of user browsing the web page and user ID; ② Then aggregate the current browsing interface and another page to which you jump; ③ Sort the data set according to the user ID and access time, and obtain the order of each user browsing the web page; ④ Group the sorted data according to the user ID and merge the pages browsed by the same user; ⑤ Then, the grouped data is converted into single-hop form according to the browsing order of the same user's web pages; ⑥ Aggregate the converted data and count the total number of each single hop. Finally, the single-hop conversion rate of A → B page is calculated by the formula A → B page single-hop conversion rate=A → B page single-hop total number/A total number of visits.

ISSN: 2688-8653

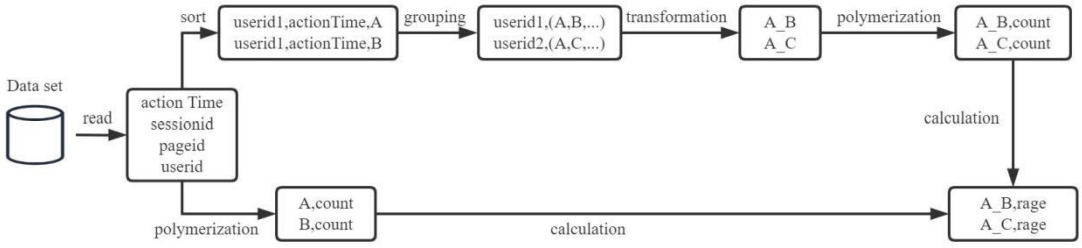


Figure 3. Statistical analysis process of website conversion rate

4.4. Real-time statistics of advertising click stream

commerce websites usually have some advertising positions. When users browse the website, the advertising content will be displayed in the corresponding advertising positions. The user may click the advertisement to jump to the corresponding interface to view the details, so as to improve the user's browsing depth and purchase probability on the website, and conduct real-time calculation and statistics for the real-time data of the user's advertising click behavior. The implementation idea is shown in Figure 4: ① Generate and read Kafka real-time production user ad click stream data. Convert the data format into a data format with userid as key, and adid and city as a whole as value; ② SparkStreaming, as a consumer, reads the data produced by Kafka in real time, merges the data in the blacklist user table in HBase database, and filters the data containing blacklist users; ③ Perform two aggregation operations on the filtered data. The first aggregation converts the data format to userid as the key and value 1 as the value, and counts the number of clicks of each advertisement in different cities. In the second aggregation, the data format is converted into a key with adid and city as a whole, and a value of 1 is used as a value to count the number of times users appear, which is used to add users with more than 100 clicks of advertisements to the blacklist users.

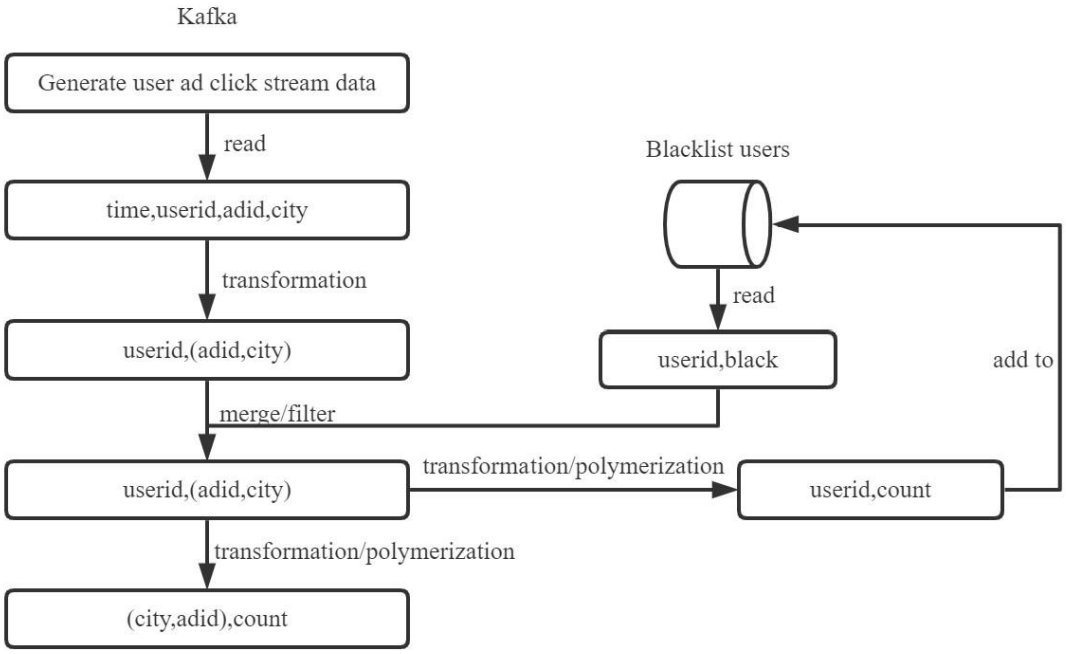


Figure 4. Real-time statistical analysis process of advertising click stream

5. Data Visualization

Data visualization can express data or information as visual objects in graphics to convey data or information. The goal is to clearly and effectively convey information to users, so that users can easily understand the complex relationship in data or information. The system realizes the visualization function of offline analysis and real-time analysis.

ISSN: 2688-8653

5.1. Offline data visualization

The visualization of Top10 analysis of popular categories, Top3 analysis of regional products, and website conversion rate statistics can be divided into the following steps: ① First, create tables in Phoenix to map existing data tables in HBase database; ② Define the entity class Entity, which is used to store the data obtained from the database. Define the data access interface Dao, which is used to access data in the database. Define the controller class Controller, which is used to implement the interface to obtain data in the database and transfer data to HTML through the Model object; ④ Define the HTML page, read the data in the Model object and fill it into the ECharts template to realize data visualization.

5.2. Real-time data visualization

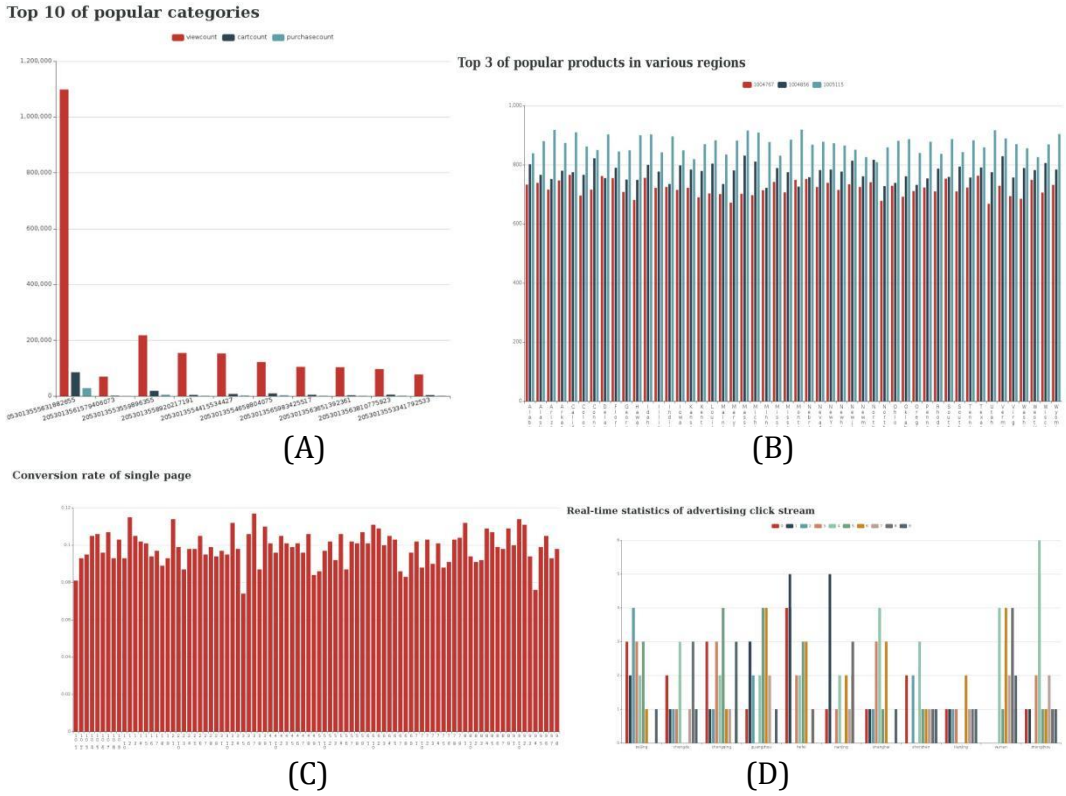
The visualization of real-time statistics of advertising click stream is divided into the following steps: ① First, create a data table in the HBase database to store real-time analysis results; ② Create tables in Phoenix to map existing data tables in HBase database; ③ Start Kafka producers to write data to Kafka; Spark Streaming consumes the data in Kafka, processes the data according to the business logic, and stores the processing results in the corresponding table of HBase database. ④ Define the entity class Entity to store the data obtained from the database. Define the data access interface Dao to access the data in the database. Define the controller class Controller, implement the interface to obtain data in the database, respond to Ajax requests sent by HTML pages and return data. ⑤ Define the HTML page, obtain the data returned by the Controller through Ajax request and fill it into the ECharts template to realize data visualization.

5.3. Visualization

Offline data visualization: In Figure 5, sub-figure (A) shows the display effect of Top10, a popular category. The abscissa is the product number, and the three bar charts represent the number of views, the number of added shopping carts, and the number of purchases, respectively. The ordinate is a numerical value; Subfigure (B) shows the display effect of Top3, a popular product in each region. The abscissa is the name of the region, the three bar charts represent three types of products, and the ordinate is the value of the product; Subfigure (C) shows the display effect of website conversion rate statistics, in which the abscissa is the page label of the user's browsing, and the ordinate is the value of the conversion rate of a single page.

Real-time data visualization: The sub-graph (D) of Figure 7 shows the display effect of real-time statistics of advertising click stream, in which the abscissa is the city name, the ten bar charts represent ten advertisements respectively, and the ordinate is the statistics of the number of advertisements clicked. In the process of real-time statistics, the statistical value will be refreshed every five seconds, and the sub-graph (D) is the display effect of one refresh.

ISSN: 2688-8653



(A) (B)

(C) (D)

Figure 5. Visual display effect

6. Conclusion

With the development of Internet big data, e-commerce platforms enrich people's daily life and also produce massive user behavior data every day. How to analyze and mine the common points of daily, weekly and monthly behavior data from the massive user behavior data has become one of the research topics of many e-commerce platforms. This system is composed of two functional modules: offline data analysis and real-time data analysis, which can better help e-commerce platforms understand users' preferences within a certain period of time, so as to adjust product sales quantity and sales method. Through the analysis of user behavior data, the characteristics of each user and valuable users in the user group are mined, and effective expansion marketing is carried out for different users, so as to improve the company's profits.

Acknowledgments

This work is supported by Undergraduate scientific research and innovation fund project of Anhui University of Finance and Economics, *Spark-based e-commerce user behavior analysis system under big data environment* (Grant No: XSKY22147), and General teaching and research project of Anhui University of Finance and Economics, *Research and practice of case-driven big data application system development capability training* (Grant No: acjyyb2022021).

References

[1] China Internet Information Center. Statistical Report on the Development of Internet in China [EB/OL].[2019-09-29].

[2] Chen Wei. Design and implementation of e-commerce user behavior analysis system based on Hadoop [J]. Journal of Suzhou Institute of Education, 2021,24 (03): 120-125.

[3] Sun Kang. Research on Spark-based e-commerce user behavior analysis system [D]. Tutor: He Xiaojun. Shenyang University of Technology, 2022.

ISSN: 2688-8653

[4] Yin Le, Yao Yuan, Liu Chen. Research on the framework of user behavior analysis system based on Spark [J]. Network Security Technology and Application, 2018 (02): 56-57.

[5] Chen Meng Design and implementation of user behavior analysis system based on big data platform [D]. Nanjing University, 2020.

[6] Zhou Weikun Research on big data analysis of e-commerce user behavior based on spark [D]. Guangdong University of Technology, 2019.

[7] He Xiaojun, Sun Kang. Spark-based e-commerce user behavior analysis system [J]. Information Technology and Informatization, 2021(11):95-97.