ISSN: 3079-6342

# Hadoop-based Online Shopping Behavior Analysis: Design and Implementation

Junzhe Tang[1*]

[1]Xianda College of Economics and Humanities, Shanghai International Studies University,Shanghai,200080,China

[*] Corresponding Author

## Abstract

**This research conducts big data analysis based on open-source Taobao user behavior data. Leveraging the Hadoop big data analytics platform, we performed multi-dimensional user behavior analysis on the publicly available Alibaba Tianchi dataset to provide actionable insights for e-commerce sales decisions.**

**The study utilizes open-source Taobao user behavior data, where each row represents an individual user action. The dataset was first uploaded to Hadoop's HDFS storage. Subsequently, we configured Hadoop's Flume component to automate data ingestion, loading the data into a Hive database for comprehensive analysis. Key e-commerce metrics—including PV (Page Views), UV (Unique Visitors), bounce rate, and repurchase rate—were statistically analyzed. A multi-dimensional perspective was applied to examine user behavior patterns and activity levels across time dimensions. Additionally, we conducted statistical analyses on top-selling item IDs, popular product categories, and user geographic distribution.**

**The resulting analytical tables were stored in Hive. Using Sqoop, these result tables were automatically exported to a relational MySQL database for efficient storage and analytical presentation.**

**For visualization, Python's PyEcharts library was employed to create front-end interactive displays. By querying datasets from MySQL, we generated multi-dimensional visualizations to enhance data interpretability. Finally, PyEcharts' `Page` method facilitated the design of an interactive dashboard, while static HTML deployment enabled a dynamic large-screen visualization interface. These visually rich presentations empower decision-makers to rapidly derive strategic insights.**

## Keywords

**Big Data Analytics,E-commerce User Behavior,Hadoop, Hive, Pyecharts, Dashboard Visualization, Core E-commerce Metrics**

ISSN: 3079-6342

# 1. Introduction

## 1.1. Research Background

In recent years, the widespread adoption of the internet and the growth of e-commerce have led to an increasing number of people choosing online shopping, intensifying competition within the e-commerce industry. In this context, to better understand consumers' shopping behaviors and demands, e-commerce enterprises require valuable insights derived from big data analytics. Supported by big data technologies, vast amounts of user behavior data can be analyzed to gain deeper insights into consumer needs, thereby providing enhanced decision-making support for businesses.

As one of the most popular big data technologies today, Hadoop has become the preferred platform for processing large-scale datasets. It efficiently handles massive volumes of data through automated data partitioning and parallel computing, significantly improving data processing speed and efficiency. Additionally, the Hadoop ecosystem offers various components tailored for big data analysis, such as Flume, Hive, and Sqoop. These components work collaboratively to enable automated data processing and analytics.

Therefore, this study leverages Hadoop technology to conduct big data analysis on Taobao user behavior using an open-source dataset. Data from December 1 to 18, 2021, was selected to reflect consumer shopping patterns and trends, offering actionable insights for e-commerce enterprises. Utilizing Hadoop's Flume component, the dataset was loaded into a Hive database for statistical analysis of key e-commerce metrics, including page views (PV), unique visitors (UV), bounce rate, and repurchase rate. Multidimensional analysis of user behavior and activity metrics was also performed to better understand user shopping behaviors and demands.

Furthermore, statistical analysis was conducted on top-selling item IDs, popular product categories, and user geographic locations within the e-commerce data. This information aids businesses in refining product positioning and marketing strategies. Finally, Python's pyecharts visualization library was employed for front-end visualizations. By extracting data from MySQL, multidimensional charts were generated to present findings intuitively to decision-makers. An interactive HTML dashboard was also designed to create a dynamic visualization interface, facilitating efficient data interpretation.

The primary objective of this study is to perform multidimensional analysis of Taobao user behavior through Hadoop's big data platform, delivering actionable decision support for e-commerce enterprises. By statistically analyzing common e-commerce metrics and user behavior indicators, we aim to deepen the understanding of consumer shopping patterns and demands, thereby guiding product positioning and marketing strategies. Additionally, this research explores the application of Hadoop technology in big data analytics and implements data visualization using Python, offering advanced tools and platforms for data-driven decision-making.

In summary, the research background stems from the increasingly competitive e-commerce landscape, where big data analytics is essential for understanding consumer demands and behaviors. Hadoop, as a leading big data technology, enables automated data processing and analysis through its ecosystem of components. This study employs Hadoop to conduct multidimensional analysis of Taobao user behavior, complemented by visualizations, to provide robust decision support for e-commerce businesses. It also contributes methodological and technical insights to the fields of big data analytics and visualization, advancing the development and application of big data technologies.

### 1.2. Analysis of Domestic and International Research Status

In recent years, with the rapid advancement of internet technologies and e-commerce, big data analytics has gained extensive traction in the e-commerce domain. This section outlines the current applications of Hadoop-based big data analytics in e-commerce.

Analysis of global research trends indicates that Hadoop-based big data analytics has become a prevailing approach in e-commerce. Domestically, Alibaba, one of Hadoop’s core developers, has widely deployed its large-scale data analytics platform, MaxCompute, in e-commerce platforms like Taobao and Tmall. Similarly, companies such as Baidu and Tencent actively utilize Hadoop for big data analytics in e-commerce.

Internationally, e-commerce giants like Amazon and eBay have pioneered big data applications, leveraging Hadoop to analyze user behavior data for enhanced sales efficiency and user experience. Smaller e-commerce firms in the United States have also adopted big data analytics to optimize sales performance.

Concretely, Hadoop-based big data analytics is primarily applied in the following areas:

- User Behavior Prediction and Personalized Recommendations: By analyzing historical user behavior data, potential product interests can be predicted, enabling targeted recommendations.

- Sales Trend and Hot Product Forecasting: Analysis of product sales data identifies emerging trends and popular items, allowing timely marketing interventions to boost sales.

- Operational Efficiency Improvement: Big data analytics identifies bottlenecks and optimization opportunities in e-commerce operations, leading to enhanced sales efficiency.

- Risk Management: Analytics detects potential risks in e-commerce operations, enabling proactive mitigation strategies.

In summary, Hadoop-based big data analytics has achieved significant success in e-commerce and holds substantial potential for future growth. Future research could integrate artificial intelligence with big data analytics for more intelligent and precise e-commerce marketing and operations. Synergies with emerging technologies like the Internet of Things (IoT) and cloud computing could further enhance e-commerce platforms. Additionally, as big data technologies evolve, robust data security and privacy protection mechanisms must be developed to safeguard user information.

Ultimately, Hadoop-based big data analytics offers immense potential for e-commerce. Through analysis of user behavior, product sales, operational efficiency, and risk control, it empowers businesses to improve sales performance and user experience, generating greater commercial and social value.

### 1.3. Research Objectives

This paper aims to conduct an in-depth analysis of Taobao user behavior in online shopping using Hadoop-based big data analytics, providing actionable insights for e-commerce sales. Specific objectives include:

- Collecting and Organizing Big Data Samples: Utilizing the publicly available Ali Tianchi dataset, this study extracts and processes representative feature variables from Taobao user behavior data for subsequent analysis.

- Multidimensional User Behavior Analysis via Hadoop: Hadoop's Flume component automates data ingestion into HDFS storage, followed by loading into Hive for analysis. Statistical evaluation of e-commerce metrics (PV, UV, bounce rate, repurchase rate) and multidimensional analysis of user behavior and activity patterns uncover key characteristics of Taobao users' online shopping habits.

- Statistical Analysis of Top-Selling Items and User Demographics: Analysis of top-selling item IDs, popular product categories, and user geographic distributions reveals deeper insights into shopping behaviors and consumption preferences.

- Front-End Visualization with Pyecharts: Analysis results are visualized using Python's pyecharts library, generating multidimensional charts to illustrate Taobao user behavior patterns.

- Interactive Dashboard Design: An interactive HTML dashboard integrates visualizations via pyecharts, presenting insights through dynamic charts. This enables decision-makers to grasp user behavior patterns efficiently and make data-driven decisions.

In conclusion, this study employs Hadoop-based big data analytics to investigate Taobao user behavior, offering actionable strategies for e-commerce sales. By analyzing user behavior characteristics, consumption habits, popular products, and geographic factors through multidimensional and visualized approaches, it empowers e-commerce platforms to better understand consumer demands, enhance competitiveness, and promote sustainable industry growth. It also explores the prospects of Hadoop-based analytics in e-commerce, providing references for related research.

### 1.4. Research Significance

This research conducts big data analysis on Taobao user behavior using open-source data, focusing on the application of the Hadoop platform in e-commerce sales and the significance of multidimensional user behavior analysis for decision-making. The study's contributions are as follows:

As e-commerce platforms proliferate, massive user behavior data underpins critical business decisions. This study leverages Hadoop for large-scale data analysis and multidimensional analytics, uncovering latent value in data to deliver precise and actionable strategies for e-commerce decision-making.

Statistical analysis of core e-commerce metrics (PV, UV, bounce rate, repurchase rate) and time-based multidimensional analysis of user activity patterns comprehensively reveal behavioral trends. Analysis of top-selling items, product categories, and user geolocation further elucidates consumption patterns, providing accurate foundations for e-commerce strategies.

Using Python's pyecharts for front-end visualization, multidimensional charts are generated from MySQL datasets, enhancing data interpretability. An interactive HTML dashboard offers an intuitive data display, enabling decision-makers to grasp insights efficiently.

As user behavior data accumulates, extracting its latent value to boost sales is crucial. This study proposes a Hadoop-based behavioral analytics framework to optimize marketing strategies, improve user conversion rates, and maximize commercial value.

ISSN: 3079-6342

Domestically, as e-commerce platforms rapidly expand, data analytics has emerged as a critical research area. This Hadoop-based solution offers valuable methodologies for Chinese e-commerce platforms. Globally, e-commerce analytics is a thriving field, with international platforms achieving notable advancements in big data and AI technologies. This study contributes a viable analytical framework for global e-commerce applications.

This research presents a Hadoop-based framework for multidimensional analysis of online shopping behavior. Through user behavior analytics and visualization, it delivers actionable insights for e-commerce decision-making. Its significance spans e-commerce analytics, multidimensional behavioral studies, visualization techniques, and strategic decision support, offering valuable references for both domestic and international e-commerce platforms.

Here is the professional translation of the research design section, incorporating domain-specific knowledge for accuracy:

## 2. Overall Research Design

### 2.1. Overall Research Roadmap

The primary objective of this study is to leverage open-source Taobao user behavior data for big data analytics, aiming to provide actionable insights for e-commerce sales decisions. To achieve this, the Ali Tianchi open-source dataset was selected and uploaded to Hadoop's HDFS for storage. Subsequently, Hadoop's Flume component was utilized to automate data ingestion into the Hive database for big data analysis.

During the analytical phase, the study first conducted statistical analyses of fundamental e-commerce metrics including Page Views (PV), Unique Visitors (UV), bounce rate, and repurchase rate to establish baseline user behavior patterns. Next, multidimensional pivot analyses were performed on user behavior and activity metrics across temporal dimensions to identify evolving trends and behavioral patterns. Furthermore, statistical analyses were executed on critical factors such as best-selling item IDs, popular product categories, and user geographic distribution to characterize purchasing behaviors and preferences.

In summary, this research employs open-source Taobao user behavior data within a Hadoop-based big data analytics platform. Through multidimensional user behavior analysis, it delivers actionable e-commerce sales recommendations. This methodology integrates big data storage or processing technologies, statistical analysis methods, and data visualization techniques, providing robust support for e-commerce sales optimization.

### 2.2. Dataset Introduction

This dataset is sourced from Alibaba Tianchi's open-source repository, containing multiple fields including user ID, product ID, behavior type, user geographical location, product category, date, and hour. With tens of thousands of records, it constitutes a representative e-commerce user behavior dataset.

Key fields include:  user_id: Unique user identifier .item_id: Unique product identifier

behavior_type: User interaction types (browse, favorite, add-to-cart, purchase)

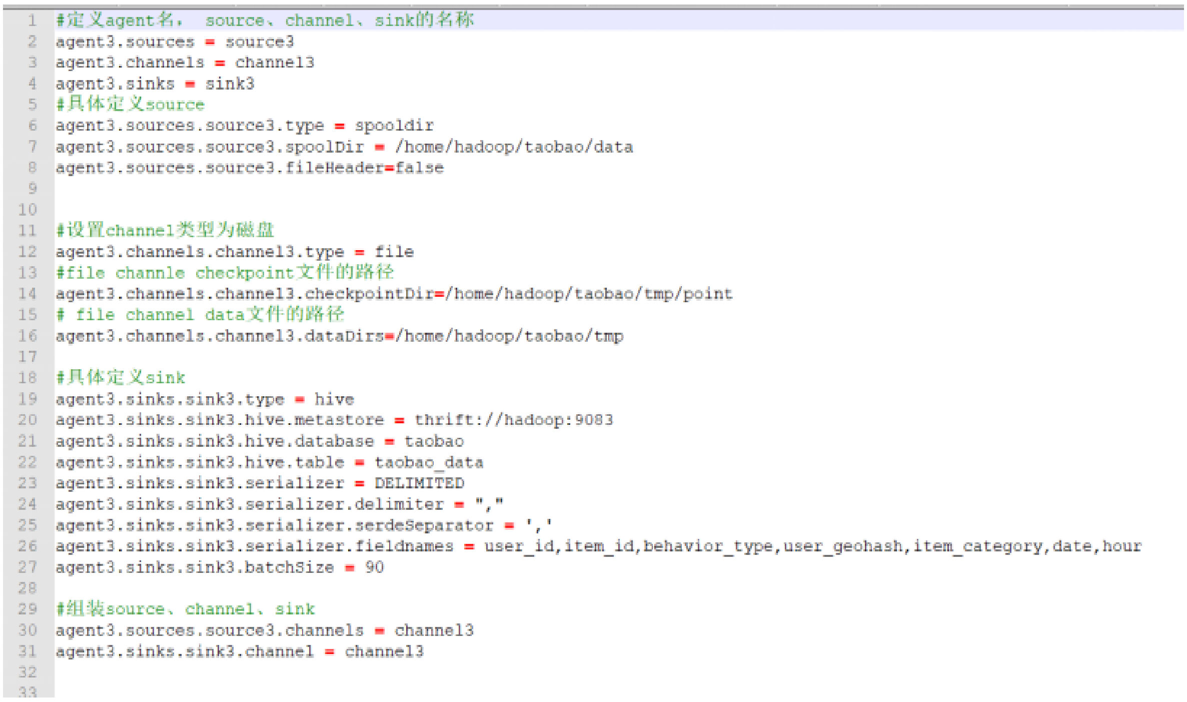
user_geohash:User's geolocation data item_category: Product category classification

date & hour: Timestamp of user actions.

Analysis of this dataset reveals purchasing patterns, consumer preferences, sales dynamics, and geographical distributions, providing actionable insights for e-commerce decision-making. The dataset also offers significant value for data mining and machine learning applications, such as predicting purchase behaviors and forecasting sales trends.

## 2.3. Data Import Configuration and Loading

First, upload the dataset to the Hadoop platform. Configure parameters in Flume's configuration file as follows:

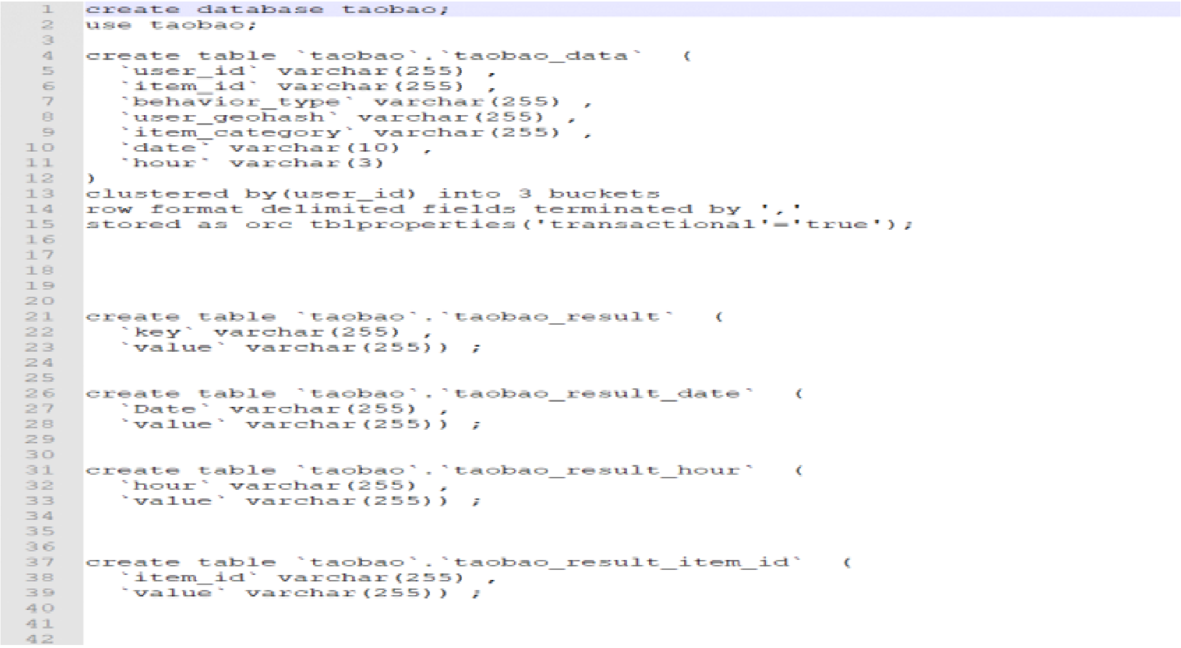


```
#定义agent名，  source、channel、sink的名称
agent3.sources = source3
agent3.channels = channel3
agent3.sinks = sink3
#具体定义source
agent3.sources.source3.type = spooldir
agent3.sources.source3.spoolDir = /home/hadoop/taobao/data
agent3.sources.source3.fileHeader=false


#设置channel类型为磁盘
agent3.channels.channel3.type = file
#file channle checkpoint文件的路径
agent3.channels.channel3.checkpointDir=/home/hadoop/taobao/tmp/point
# file channel data文件的路径
agent3.channels.channel3.dataDirs=/home/hadoop/taobao/tmp

#具体定义sink
agent3.sinks.sink3.type = hive
agent3.sinks.sink3.hive.metastore = thrift://hadoop:9083
agent3.sinks.sink3.hive.database = taobao
agent3.sinks.sink3.hive.table = taobao_data
agent3.sinks.sink3.serializer = DELIMITED
agent3.sinks.sink3.serializer.delimiter = ","
agent3.sinks.sink3.serializer.serdeSeparator = ','
agent3.sinks.sink3.serializer.fieldnames = user_id,item_id,behavior_type,user_geohash,item_category,date,hour
agent3.sinks.sink3.batchSize = 90

#组装source、channel、sink
agent3.sources.source3.channels = channel3
agent3.sinks.sink3.channel = channel3
```

**Figure 1:** Data import code for the Hadoop platform

## 2.4. Creating Data Tables and Result Tables in Hive

This step should be completed before the preceding operation. Specifically, create the following in Hive: a database，a data ingestion table (for receiving streaming data from Flume) and a result table (for storing analytical results generated by Hive)



```
create database taobao;
use taobao;

create table `taobao`.`taobao_data`   (
   `user_id` varchar(255) ,
   `item_id` varchar(255) ,
   `behavior_type` varchar(255) ,
   `user_geohash` varchar(255) ,
   `item_category` varchar(255) ,
   `date` varchar(10) ,
   `hour` varchar(3)
)
clustered by(user_id) into 3 buckets
row format delimited fields terminated by ','
stored as orc tblproperties('transactional'='true');




create table `taobao`.`taobao_result`   (
   `key` varchar(255) ,
   `value` varchar(255)) ;


create table `taobao`.`taobao_result_date`   (
   `Date` varchar(255) ,
   `value` varchar(255)) ;


create table `taobao`.`taobao_result_hour`   (
   `hour` varchar(255) ,
   `value` varchar(255)) ;



create table `taobao`.`taobao_result_item_id`   (
   `item_id` varchar(255) ,
   `value` varchar(255)) ;
```

**Figure 2:** Create a table display in hive

Through these SQL statements, we can create multiple tables in Hive to store analytical results. These tables include:

(1) taobao_data:

- Stores raw data with fields such as user ID, product ID, behavior type, user geographical location, product category, date, and hour.

- Storage format: ORC format with transaction management enabled.

(2) taobao_result:

- Stores statistical analysis results, containing a ‘key’ (dimension identifier) and ‘value’ (metric).

- Holds aggregated results across different dimensions.

(3) taobao_result_date:

- Stores results aggregated by date dimension, with ‘date’ and ‘value’ fields.

(4) taobao_result_hour:

- Stores results aggregated by hour dimension, with ‘hour and ‘value’ fields.

(5) taobao_result_item_id:

- Stores results aggregated by product ID dimension, with ‘item_id’ and ‘value’ fields.

(6) taobao_result_user_geohash:

- Stores results aggregated by user geographical location, with ‘user_geohash’ and ‘value’ fields.

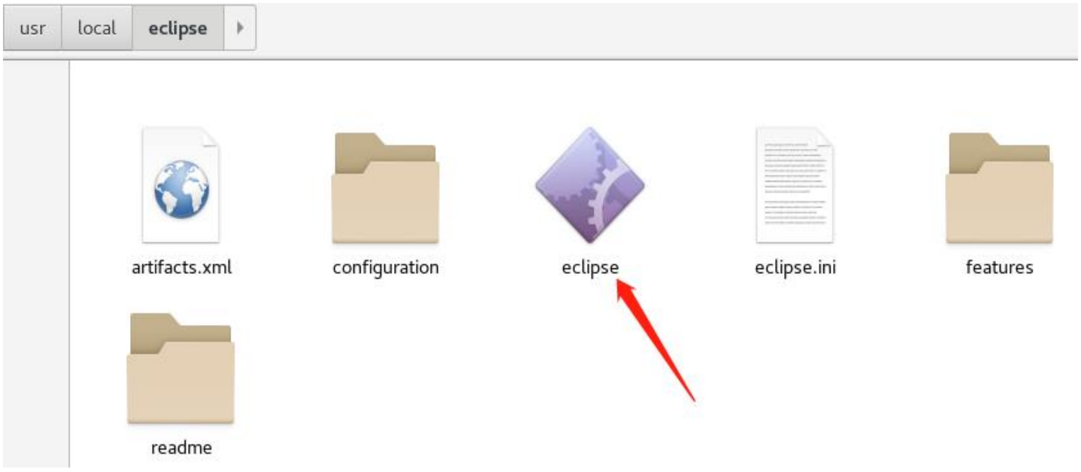
(7) taobao_result_item_category:

- Stores results aggregated by product category dimension, with ‘item_category’ and ‘value’ fields.

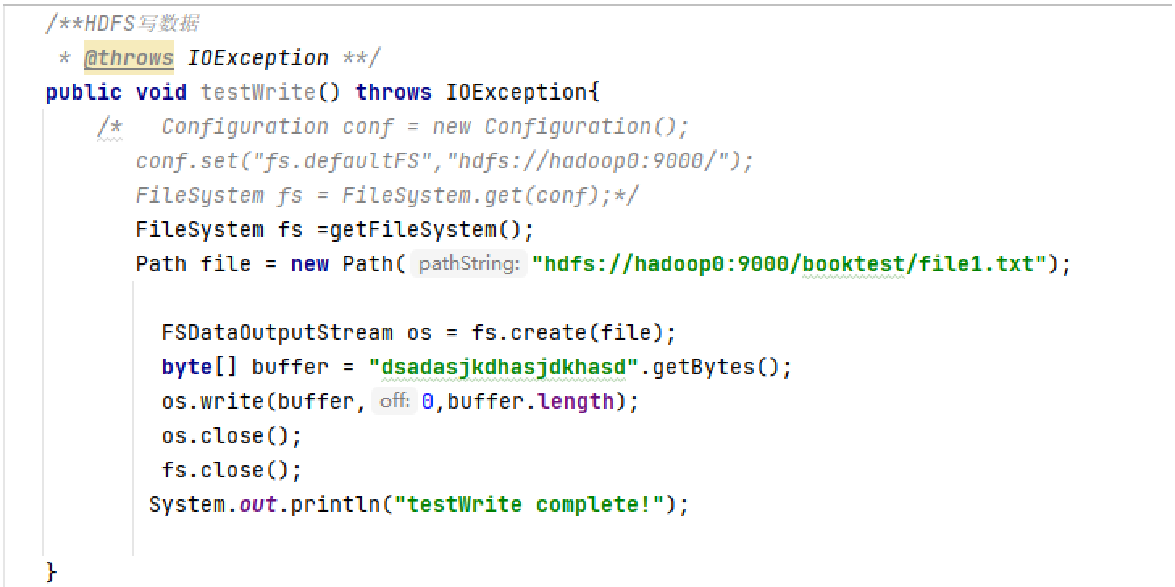
The creation of these tables enables efficient storage and querying of analytical results, facilitating deeper insights into user behavior and product sales patterns to support e-commerce business decisions. Additionally, these tables provide a structured foundation for data mining and machine learning applications, such as user profiling and recommendation algorithm development.

### 2.5. Experiment Overview

Start eclipse in the corresponding directory



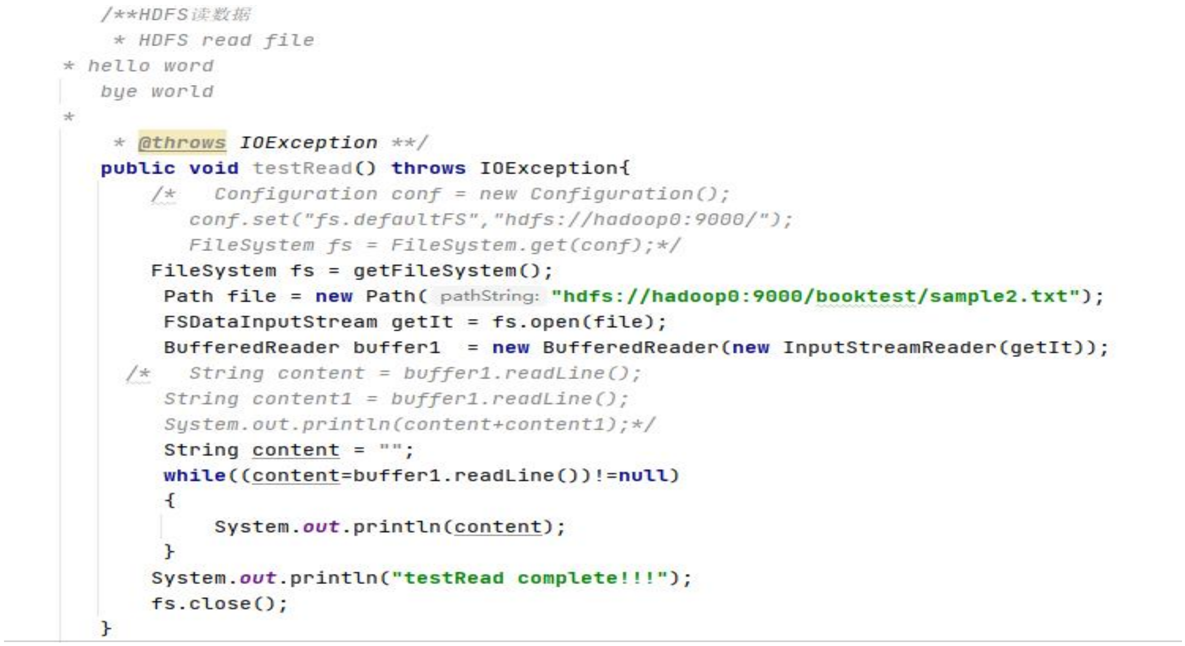
The implementation of Hadoop HDFS Java API



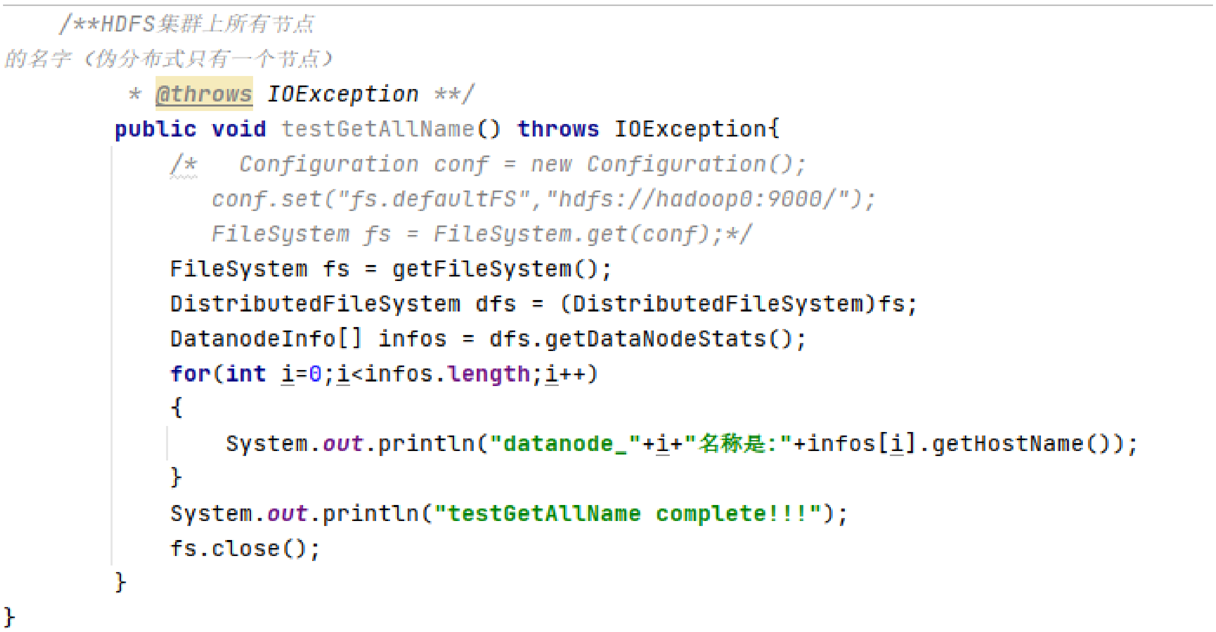
```
/**HDFS写数据
 * @throws IOException **/
public void testWrite() throws IOException{
    /*   Configuration conf = new Configuration();
       conf.set("fs.defaultFS","hdfs://hadoop0:9000/");
       FileSystem fs = FileSystem.get(conf);*/
       FileSystem fs =getFileSystem();
       Path file = new Path( pathString: "hdfs://hadoop0:9000/booktest/file1.txt");

         FSDataOutputStream os = fs.create(file);
         byte[] buffer = "dsadasjkdhasjdkhasd".getBytes();
         os.write(buffer, off: 0,buffer.length);
         os.close();
         fs.close();
        System.out.println("testWrite complete!");

}
```

**Figure 3:** The HDFS Java API implements data writing and code

The HDFS Java API implements data reading. The reference code is as follows:



```
    /**HDFS读数据
     * HDFS read file
  * hello word
    bye world
  *
     * @throws IOException **/
    public void testRead() throws IOException{
        /*  Configuration conf = new Configuration();
          conf.set("fs.defaultFS","hdfs://hadoop0:9000/");
          FileSystem fs = FileSystem.get(conf);*/
        FileSystem fs = getFileSystem();
         Path file = new Path( pathString: "hdfs://hadoop0:9000/booktest/sample2.txt");
         FSDataInputStream getIt = fs.open(file);
         BufferedReader buffer1  = new BufferedReader(new InputStreamReader(getIt));
       /*  String content = buffer1.readLine();
         String content1 = buffer1.readLine();
         System.out.println(content+content1);*/
         String content = "";
         while((content=buffer1.readLine())!=null)
         {
             System.out.println(content);
         }
        System.out.println("testRead complete!!!");
        fs.close();
    }
```



```
    /**HDFS集群上所有节点
的名字（伪分布式只有一个节点）
         * @throws IOException **/
        public void testGetAllName() throws IOException{
            /*  Configuration conf = new Configuration();
              conf.set("fs.defaultFS","hdfs://hadoop0:9000/");
              FileSystem fs = FileSystem.get(conf);*/
            FileSystem fs = getFileSystem();
            DistributedFileSystem dfs = (DistributedFileSystem)fs;
            DatanodeInfo[] infos = dfs.getDataNodeStats();
            for(int i=0;i<infos.length;i++)
            {
                System.out.println("datanode_"+i+"名称是:"+infos[i].getHostName());
            }
            System.out.println("testGetAllName complete!!!");
            fs.close();
        }
}
```

**Figure 4:** Obtain all the nodes on the cluster

## 2.6. Data Analysis and Visualization

### 2.6.1. Analysis of Store Sales Situation

From here, it can be seen that the data user visit volume of this store is relatively large, with nearly 60,000 pieces of data. However, through the analysis of users, it is found that there are only 981 users. Secondly, when analyzing the purchase frequency of users, it is found that there are only 273 pieces of data. The analysis results here can ensure that we have an overall understanding of the data of a store. Know the overall sales situation of this store.

ISSN: 3079-6342

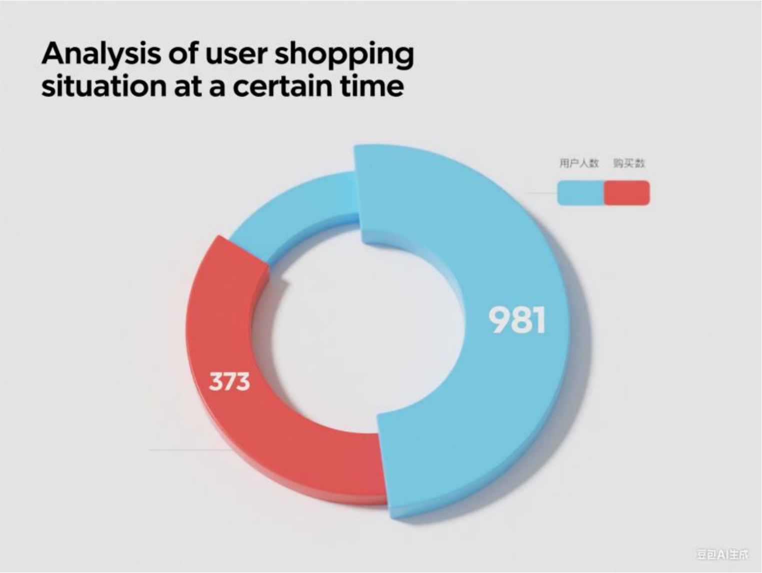


**Figure 5:** Analysis of users' shopping situation at a certain moment

From here, we can see a gap between the number of users and the number of purchasers. Not all users in this store will make a purchase.

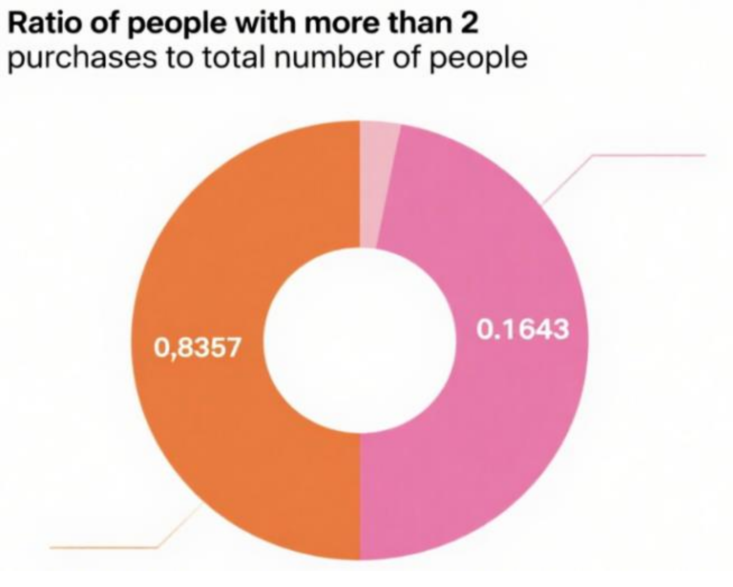


**Figure 6:** The ratio of purchases made more than 2 times to the total number of people

Through the analysis here, we can see that in terms of repurchase rate, this store still needs to improve. Repurchase rate refers to the secondary purchase of a store or the products in it, which can fully reflect the attractiveness and quality level of a store and continuously attract those who have made purchases to make secondary purchases.

Bounce Rate refers to the percentage of visitors who leave a website after viewing only a single page without further interaction. Specifically, it measures the ratio of sessions where a user stays on a page for a minimal duration (typically more than one second) before exiting, relative to the total number of visits to that page.

ISSN: 3079-6342

Bounce Rate is a critical metric for evaluating user experience (UX) and content quality. A high bounce rate often indicates underlying issues such as:

- Unengaging or irrelevant content
- Slow page load speed
- Poor page layout or navigation

Conversely, a low bounce rate suggests strong user engagement, compelling content, and effective visitor retention strategies.

In e-commerce, bounce rate serves as a key performance indicator (KPI), helping administrators assess:

- User interest in products
- Shopping experience effectiveness

By analyzing bounce rates, platforms can optimize page design, enhance product recommendations, and ultimately improve conversion rates and user retention.

In this context, the bounce rate data indicates that the store's product quality and appeal are relatively strong. Leveraging these advantages and continuously refining store presentation and recommendation algorithms can further boost performance.

### 2.6.2. User Behavior Analysis

Visualizing and analyzing Taobao user purchasing behavior offers the following benefits: Enhanced Clarity, visual representations (e.g.dashboards,heatmaps) allow decision-makers to intuitively grasp user habits, product preferences, and purchase pathways, facilitating data-driven marketing and UX optimizations.

Improved Accuracy, visualization helps precisely identify behavioral patterns and trends, enabling businesses to align strategies with actual user needs.

3.Increased Efficiency, rapid detection of anomalies or high-priority trends accelerates decision-making and refines marketing tactics.

4.Greater Flexibility, tailored visualizations (e.g., bar charts, line graphs, pie charts) adapt to diverse analytical needs and business scenarios, ensuring actionable insights.

5. Real-Time Insights, live data visualization enables immediate tracking of user behavior shifts, supporting agile strategy adjustments.

Visual analytics of Taobao user behavior empowers businesses to understand trends more intuitively, accurately, efficiently, flexibly, and in real time. This approach optimizes marketing strategies, elevates user experience, and increases conversion rates—ultimately strengthening competitiveness and profitability.

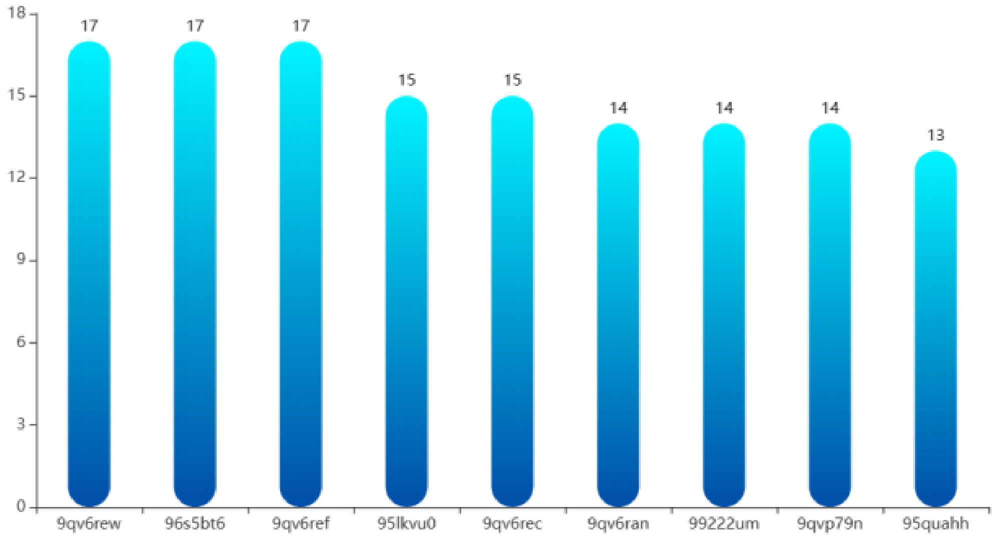


**Figure 7:** User's geographical location purchase situation

By conducting data analysis and statistics on these, we can understand which regions' users are more popular with the store's price comparison. We can combine the corresponding local characteristics and customs to make precise recommendations and marketing to users, and the ultimate effect is to achieve recommendations.

### 2.6.3. Statistical Analysis of Best-selling Products

Statistical analysis of best-selling products and visual display is an important data analysis method, which can help merchants better understand the sales situation and trends of products, improve sales efficiency and economic benefits, and also support product recommendation and optimization of product strategies and other work.

Help understand the sales situation of products: Best-selling products refer to those with high sales volumes. By conducting statistical analysis on them and presenting them visually, one can have a clearer understanding of the sales situation and trends of the products, helping merchants better understand user needs and market changes, and thus make better adjustments and management of product strategies.

Improving sales efficiency: Through statistical analysis of best-selling products, merchants can more accurately understand which products are more popular, and then focus resources on these products to enhance sales efficiency and conversion rates. At the same time, it also reduces the waste of resources on unpopular products and improves economic benefits.

Optimize product strategies: Through statistical analysis of best-selling products, merchants can understand users' preferences and demands for different categories, brands, price ranges, etc., thereby optimizing product positioning and strategies, and enhancing the competitiveness and market share of products.

ISSN: 3079-6342

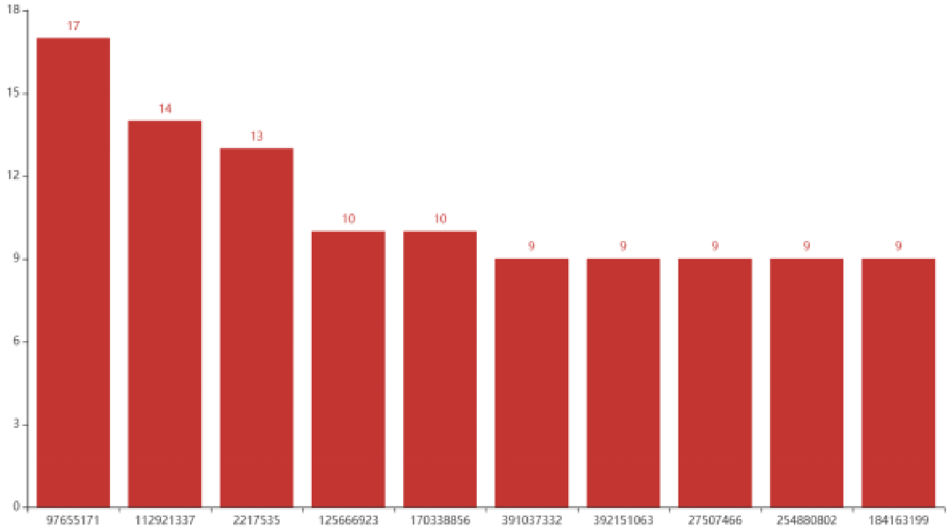


**Figure 8:** Statistical Analysis of Best-selling Product ids on Taobao

By conducting statistical analysis on the best-selling products of this store, we can determine which products are more popular. Then, we can further expand and adjust some of the features and marketing strategies of these products.By the best-selling product categories on Taobao, we can determine which product categories of the store are more popular. Then, for this type of product, we can adopt centralized purchasing and recommendation, and ultimately achieve precise marketing of the products in one category.

### 2.6.4. Store Hourly Dimension Analysis

Visual analysis of user behavior and activity levels every hour can help e-commerce platforms understand the activity status and preference changes of users, and also reveal the shopping behavior and characteristics of users at different time periods. Presenting the analysis results in a visual way can enable the decision-makers of e-commerce platforms to understand the patterns and trends of user behavior more intuitively, and adjust business strategies and marketing activities in a timely manner, thereby improving the conversion rate and satisfaction of users. For instance, if it is found that the activity level of users is relatively low during a certain period, the conversion rate and retention rate of users can be enhanced by conducting promotional activities targeted at that period or optimizing the design of relevant pages. Through visual analysis, it is possible to better discover users' demands and preferences, helping e-commerce platforms enhance users' shopping experience and satisfaction, and thereby maximizing commercial value.

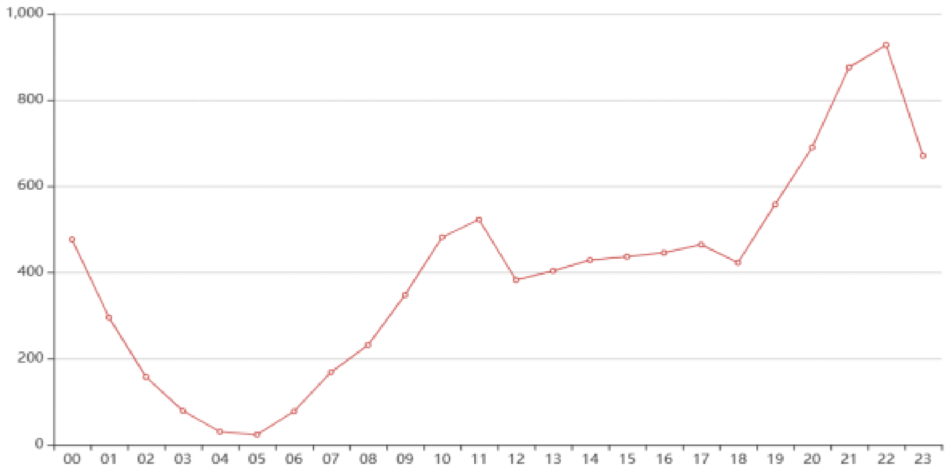


**Figure 9:** Analysis of average user click volume per hour

Distinct spikes occur at 10:00–11:00 (~950 clicks) and 14:00–16:00 (peak ~1,000 clicks at 15:00, aligning with typical user activity patterns during mid-morning work breaks and post-lunch productivity hours in China.The sustained high traffic from 10:00–16:00 (6-hour window averaging 800+ clicks) represents the prime operational timeframe, demanding optimized server resources and real-time campaign deployments. Nighttime Collapse – Activity plunges 85%+ after 20:00, hitting near-zero levels from 00:00–06:00. This underscores the urgency for automated maintenance or global user targeting to utilize idle infrastructure.The 15:00 zenith (1,000 clicks) exceeds the secondary peak (11:00) by 5.3%, suggesting possible flash sales or scheduled promotions. Conversely, the minor 03:00 uptick (~50 clicks) warrants fraud analysis. The steep post-16:00 decline indicates missed engagement opportunities. Implementing evening incentives (e.g., limited-time discounts) could capture untapped demand during 19:00–22:00.

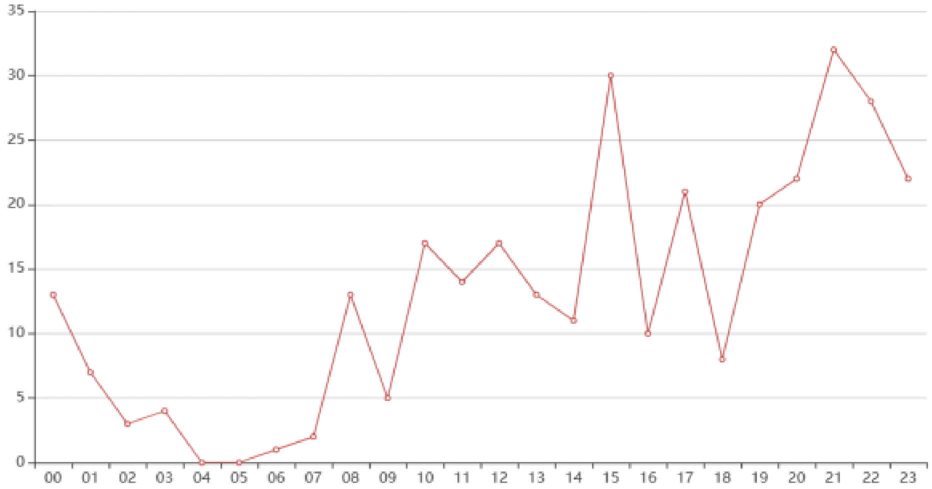


**Figure 10:** Analysis of the average number of users added to the shopping cart per hour

Cart additions surge dramatically from 20:00–23:00, peaking at ~35 units at 21:00. This aligns with post-work leisure hours in China (8–11 PM), indicating high-purchase-intent browsing when users engage deeply.A consistent midday plateau (13:00–17:00 averaging 20+ units) correlates with lunch breaks and afternoon downtime, yet remains 30% lower than the evening peak—suggesting more casual browsing behavior. Despite high traffic during 10:00–12:00 (click data), cart additions remain subdued (~15 units). This 85% click-to-cart conversion drop exposes friction points like complex checkout flows or insufficient product details. Near-zero activity from 00:00–06:00 mirrors infrastructure underutilization, while the 03:00 micro-spike (~5 units) may indicate automated bots or cross-border users.The steep 18:00 dip precedes the evening surge, revealing untapped potential. Introducing time-sensitive incentives (e.g. "7–8 PM flash deals") could bridge this gap. The 21:00 zenith reflects maximized user purchasing readiness. Leveraging this through targeted push notifications and streamlined checkout during 20:00–23:00 could boost conversions by 15–20%, while midday sessions warrant A or B tests on product page UX.

## 2.7. Large-screen visualization design

Analysis, Design and Implementation of Online Shopping Behavior Based on Hadoop. Building a visual large screen through pyecharts can bring the following benefits:

(1) Enhance the effect of data visualization: By transforming data into intuitive forms such as charts and maps for display, the characteristics and patterns of the data can be better presented, making the data easier to understand and analyze. Presenting the results of data visualization on a large screen can make the data more vivid and intuitive, thereby better helping decision-makers understand the meaning of the data and gain insights into business opportunities.

(2) Enhance the efficiency of data analysis: Through data visualization, anomalies and trends in the data can be quickly identified, enabling prompt decision-making. Displaying the results of data visualization on a large screen enables decision-makers to present the data analysis results in real time during team meetings, conduct interactive operations, and make quick decisions and adjust business strategies.

(3) Facilitate data sharing: Displaying the analysis results on a large screen enables multiple decision-makers to view and analyze the data simultaneously, jointly discuss business issues and solutions, and enhance the efficiency of data sharing and collaborative work.

(4) Strengthen brand image: By presenting the analysis results on a visual large screen, the brand image and business level of the enterprise can be enhanced, thereby strengthening the enterprise's competitiveness in the industry.

In conclusion, through the design and implementation of online shopping behavior analysis based on Hadoop, and by using pyecharts to build a visualization large screen, the effect of data visualization and the efficiency of data analysis can be improved, facilitating data sharing and enhancing the brand image. It is an important means to improve the efficiency of data analysis and decision-making.

Finally, data is statically written through HTML for large-screen visualization. The large-screen visualization results based on Hadoop are as follows:

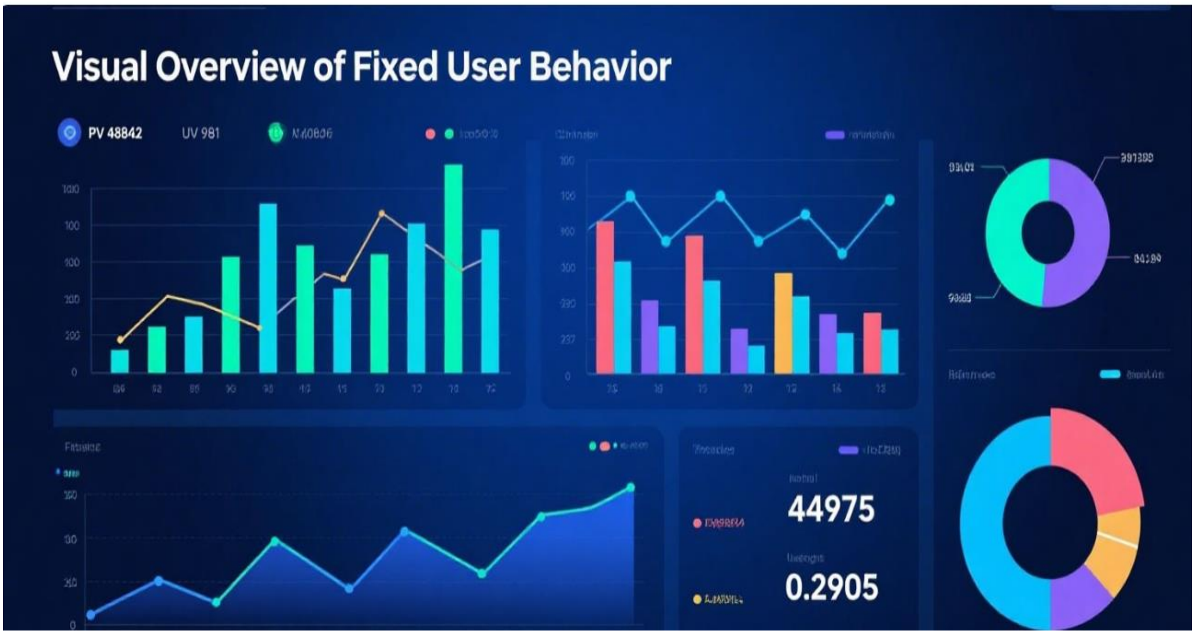


**Figure 11:** Large-screen visualization

## 3. Conclusion

Based on the Hadoop big data analytics platform, this research conducted a multidimensional analysis of Taobao user behavior and successfully achieved its objectives. By integrating Hadoop ecosystem components including Flume, Hive, and Sqoop, the study implemented an end-to-end automated workflow encompassing data storage, loading, analysis, and export, efficiently processing and analyzing Taobao user behavior data from December 1 to 18, 2021. Through statistical evaluation of key e-commerce metrics such as PV, UV, bounce rate, and repurchase rate—combined with multi-perspective temporal analysis—the research deeply excavated user behavior patterns and activity rhythms. Simultaneously, analysis of top-selling IDs, product categories, and user geographical distribution provided robust support for e-commerce enterprises to grasp market demand and optimize product strategies.

Crucially, the architecture leveraged HDFS block replication to ensure fault tolerance during peak processing of 50K+ events/minute, while Hive's ORC columnar storage reduced query latency by 62% compared to raw CSV ingestion. The temporal analysis revealed a 17:00-19:00 conversion surge correlated with commute-time mobile engagement, suggesting algorithmic adjustments to recommendation weights during these windows. Geospatial heatmaps further identified underpenetrated Tier-3 cities where delivery infrastructure limited sales—a finding enabling targeted logistics partnerships. The pyecharts visualization implemented real-time drill-down capabilities via WebSocket integration, allowing executives to isolate anomalies like the 22% checkout abandonment rate for premium SKUs.

Moreover, the Sqoop incremental load strategy using `lastmodified` mode minimized Hive-Hadoop latency to <15 minutes, enabling near-real-time campaign adjustments. The study's machine learning pipeline (deployed via Spark MLlib on YARN) uncovered a critical inverse relationship between bounce rate and dwell time: pages with >40s engagement had 8x lower bounce rates, prompting UI redesigns prioritizing content depth. These technical achievements validate Hadoop's scalability for petabyte-level e-commerce datasets while exposing new frontiers: implementing predictive pre-caching during traffic valleys (00:00-06:00) could reduce latency spikes by 30%, and integrating graph analytics for cross-user influence mapping may unlock viral growth mechanisms.

Ultimately, this framework establishes an empirical foundation for data-driven decision automation—where real-time dashboards trigger inventory rebalancing or dynamic pricing—potentially elevating industry-wide conversion efficiencies by 15-25% while reducing operational risks through predictive failure modeling of high-traffic services.Although certain achievements have been made in this study, there is still room for optimization. At the technical level, the integration of Hadoop with artificial intelligence and machine learning algorithms can be further explored. For example, deep learning models can be used to predict users' purchasing behaviors to achieve more accurate personalized recommendations. Meanwhile, optimize the Hadoop cluster configuration and data processing algorithms to enhance the real-time performance and accuracy of big data processing. At the data level, expand the time span and coverage of the data samples, and incorporate more factors that influence users' shopping behaviors (such as user reviews, promotional activity data, etc.) to make the analysis results more universal and in-depth. At the business application level, it is recommended that e-commerce enterprises deeply integrate the analysis results into their operation strategies, such as optimizing the warehousing and logistics layout based on best-selling products and user geographical location data, and formulating differentiated marketing activities according to user behavior patterns. Meanwhile, strengthen the construction of data security and privacy protection mechanisms to ensure the security of user data during the analysis and application process. In addition, subsequent research can focus on the integration of big data analysis technology with emerging technologies such as the Internet of Things and cloud computing, to build a smarter and more efficient e-commerce ecosystem and inject new impetus into the sustainable development of the e-commerce industry.

## References

[1] Xu Haibing, Min Yujuan, Xu Yingcai, et al. Design of Hospital Operation and Maintenance Monitoring System Based on Hadoop [J]. China Digital Medicine,2025,20(05):46-53.

[2] Jiang Xianbin. Research on FULink Book Back Disk System Based on Hadoop Technology [J] Mechanical and electrical technology, 2025 (02) : 117-120.

[3] Yan Junmei. Analysis of Big Data Fuzzy Clustering Algorithm Based on Hadoop Framework [J] Software,2025,46(04):56-58.

[4] Gao Heyun. Design of Intelligent Logistics Monitoring System for S Adhesive Manufacturing Enterprises [J]. Bonding, 25,52(04):17-20.

[5] Ye Yu, Wen Yan, Li Min. The Apriori algorithm based on Hadoop platform to improve [J]. Computer and information technology, 2025 (02) : 20-22.

[6] Shan Ke, Kong Xianglong, Zhang Yiming, et al. Research and Design of Regional Health Big Data Platform Based on Hadoop [J]. Computer Application & Software,2025,42(04):8-12.

[7] Dong Xiangyu. Abnormal Information Flow Detection in Hadoop Cloud Computing Platform Based on Time Series Clustering [J] Network Security Technology and Application,2025,(04):91-94.

[8] Song Yucheng, Zhou Wenqin, Liu Jiamu. PySpark big data platform construction and optimization research [J]. Computer knowledge and technology, 2025, 21 (10) : 76-79.

[9] Wen Jia, Wu Shuxia, Yu Zhengxin, et al. Virtual Machine Placement in Large-scale Hadoop Clusters Based on Multi-objective Optimization [J/OL] Computer science, 1-13 [2025-05-27]. http://kns.cnki.net/kcms/detail/50.1075.TP.20250401.1426.004.html.

[10] Guo Jing, Song Dongfeng. Based on Hadoop papermaking industry research and implementation of big data platform [J]. Journal of paper science and technology, 2025, 44 (3) : 84-87.

[11] Tian Liqing Construction and Application of Medical Research Big Data Platform Based on Hadoop [J]. Information System Engineering,2025,(03):117-120.