Research on the Construction of E-commerce Data Analysis System Based on Hadoop

Yu Chen[1, a], Yi Zhou[1, b]

[a]550417230@qq.com, [b]304810996@qq.com

[1]Chongqing College of Architecture and Technology, Chongqing, China

Abstract: In this paper, the author uses Hadoop ecosystem for data processing, POWER BI for data visualization, and Javaweb technology to develop an e-commerce data analysis system. This system develops four functional modules: traffic data, conversion rate data, user data and order data to analyze e-commerce data, which can effectively help enterprises discover the laws of consumers visiting websites, improve website design or marketing strategies according to these laws, optimize and adjust the operation focus of users' regions and consumer groups, optimize consumers' after-sales service, and then achieve the goal of helping enterprises obtain more ideal economic benefits. Under the background of economic digital era, it is necessary to enhance the competitiveness of small and medium-sized e-commerce enterprises in the e-commerce market, contribute to the GDP growth of China's e-commerce industry, improve people's livelihood, and thus promote the prosperity and development of China's market economy.

Keywords: Hadoop; E-commerce; data analysis; big data

1 Introduction

With the rapid development of social economy and information technology, the concepts of "massive data" and "large-scale data" are common, but their data volume, data complexity and production speed greatly exceed the traditional data form and the processing capacity of existing technical means. Big data technology emerges as the times require and brings great opportunities for industrial innovation to all walks of life. The e-commerce industry described in this paper is a good example.

As of December 2021, the number of online shopping users in China has reached 842 million, an increase of 59.68 million compared with December 2020, accounting for 81.6% of the total number of netizens. As a typical representative of the new format of digital economy, online retailing has continued to grow rapidly, becoming an important force to promote consumption expansion. E-commerce brings a brand-new business model. As an inevitable outcome of the development of market economy, it develops in parallel with the real economy. E-commerce is attached to the Internet, but what is actually done is the physical goods transaction. E-commerce must have actual products. The transaction is completed by online communication between both parties, and then the goods transaction is realized in the form of logistics, and the goods delivery

ICIDC 2022, June 17-19, Qingdao, People's Republic of China

DOI 10.4108/eai.17-6-2022.2322654

is realized offline through logistics [1]. E-commerce is the product of the rapid development of the Internet, with typical interconnection genes. The data increment of e-commerce is geometric, including the transaction (or business) data generated by customers' shopping behavior, click stream data generated by browsing websites, and audio and video data generated in all links of e-commerce activities. These data have become the most valuable resources of e-commerce. From the current environment, big data has become an important tool for the transformation and upgrading of various industries. For e-commerce, big data is not only a marketing tool of e-commerce platform, but also a key tool to help the development of e-commerce. The latest research by BSA Software Alliance shows that in the past few years, the productivity of e-commerce companies that apply big data analysis to their value chains is 5% to 6% higher than that of their competitors, which makes the interest in big data in academia and e-commerce industry surge. Big data analysis contributed 10% or more to 56% of company sales growth. Therefore, 91% of Fortune 1000 companies invest in big data analysis projects, an increase of nearly 85% over the previous year. Big data analysis enables e-commerce companies to use data more effectively and improve conversion rate and decision-making accuracy. From the perspective of e-commerce transaction cost theory, big data analysis can improve market transaction cost efficiency by managing transaction cost efficiency and time cost efficiency. From the perspective of limited resources, big data analysis can improve high-performance business processes and meet a large number of business needs, such as identifying loyal and profitable customers and determining the best price. Therefore, big data analysis was once regarded as a powerful driving force for "employment growth, productivity improvement and consumer surplus increase" [7].

At the same time, when a large number of false data appear in the market environment, it will affect the authenticity of the data, and the reference value of knowledge presented by big data will be reduced accordingly. In fact, most of the enterprises in the whole market are small and medium-sized enterprises or even individuals. E-commerce SMEs often take the experience of operators or the intuition of enterprise decision makers as the main decision-making basis in e-commerce operations. This kind of enterprises are often subject to many factors such as big data technology reserves, funds, data volume and talents, and it is difficult to do data analysis by themselves. Therefore, it is of great significance for individual businesses or small and medium-sized enterprises to build a third-party data analysis and display system platform with good experience [4].

The author believes that according to the above analysis, in order to meet the needs of today's e-commerce platform enterprises, an e-commerce data analysis system based on Hadoop platform combined with PowerBI visualization came into being. The data of this e-commerce data analysis system comes from the merchant's own website server and local database to ensure the authenticity and reliability of the data. By analyzing the data of consumers' visiting, browsing, purchasing and evaluating behaviors in e-commerce, this paper can help businesses intuitively understand the store operation of enterprises and further help them to optimize their business strategies by visualizing charts.

2 Key technologies

2.1 Hadoop

Hadoop is a framework that allows distributed processing of large data sets across computer clusters using a simple programming model. It can be expanded from a single server to thousands of machines, each of which provides local computing and storage. Hadoop operation modes include local mode, pseudo-distributed mode and fully distributed mode. Local mode, that is, stand-alone operation, is only used to demonstrate the official case, not in the production environment. Pseudo-distributed mode also runs on a single machine, but it has all the functions of Hadoop cluster, and one server simulates a distributed environment. Complete distributed mode is a distributed environment composed of multiple servers, most of which are used in enterprise-level production environment. To build a complete distributed mode, you need to prepare three clients (turn off the firewall, static IP, and host name), install JDK configuration environment variables, then Hadoop configuration environment variables, finally configure a single point of the cluster, start ssh configuration and test the cluster. The big data platform developed based on Hadoop usually has the following characteristics. Capacity expansion: It can reliably store and process PB-level data. Hadoop basically uses HDFS as its storage component, with high throughput, stability and reliability. Low cost: data can be distributed and processed by a server farm composed of cheap and universal machines. These server farms can total up to thousands of nodes. High efficiency: By distributing data, Hadoop can be processed in parallel on the node where the data is located, and the processing speed is very fast. Reliability: Hadoop can automatically maintain multiple backups of data, and automatically redeploy computing tasks after task failure. There are many components of Hadoop ecosystem, including yarn resource management, mapreduce offline computing, HDFS file storage, Oozie task scheduling, sqoop data transfer, zookeeper data platform configuration scheduling, etc., as shown in Figure 1 [5].

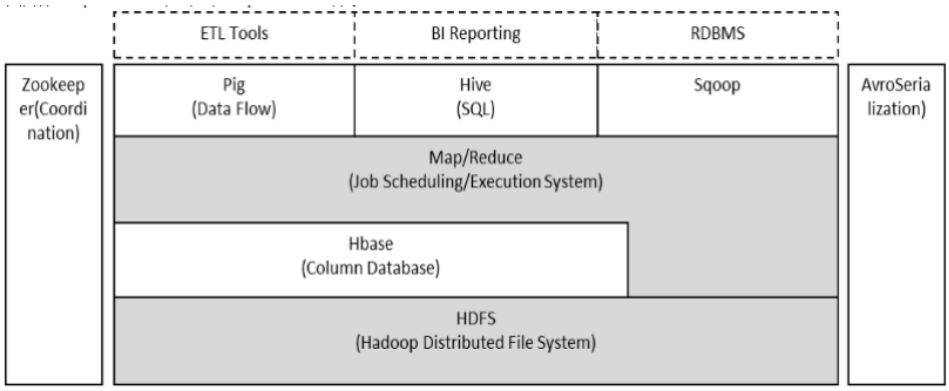

Figure 1: The Hadoop ecosystem components diagram

HDFS: HDFS consists of two types of nodes, one Namenode and several Datanode, and runs in the manager-worker mode. Namenode is the manager of the cluster, responsible for managing the namespace of the file system and maintaining the Metadata of the file system (Metadata, that is, the data such as the location where each data block of each file is stored). Datanode is the worker of the cluster and the storage node of the data. The data will be divided into multiple file blocks with the size of 64MB, distributed on multiple machines, and regularly send its own storage information to Namenode [8].

MapReduce: When Mapreduce calculates files, it will read them line by line and process them in batches. The execution process of Mapreduce task can be divided into map stage and reduce stage. The input data of the Map stage are files stored on HDFS, which will be divided into multiple file blocks, each file block corresponding to a map process. The input data of Reduce stage is the output data of map stage. In the program, these two stages respectively correspond to two functions, namely mapper function and reducer function, and users need to realize the logic of these two functions themselves. Each stage takes key-value pairs as input and output, and users specify data types for them.

2.2 PowerBI

As its name implies, PowerBI is a BI (Business Intelligence) tool, which can generate all kinds of cool reports in a short time. Therefore, it mainly completes the work of making and publishing statements. PowerBI consists of four components: powerquery data query, powerPivot data modeling, PowerBI data interactive display and powerMapExcel map plug-in. PowerBI adopts the drag-and-drop control graphical development mode. PowerBI can grab data from various data sources for analysis. Besides supporting Microsoft's own products such as Excel, SQLServer and other databases such as Oracle, MySQL and IBMDB2, it also supports importing data from R language scripts, Hdfs file system, Spark platform and so on. The e-commerce data analysis system designed in this paper will make use of the visualization function of PowerBI. First, import data from the data source. There are many data source formats supported by Power. Then, data plasticity is carried out in the background area of PowerBI. After the data is imported, it is necessary to determine whether the data column name, data type are correct, whether segmentation is required, whether summary tables are generated, and so on. Finally, draw the report in the chart area of PowerBI. In practice, this step is iterative with the previous step, and we need to continuously plastic the data, and draw various reports based on the data with good plasticity. The flow of PowerBI visualization is shown in Figure 2 [6].

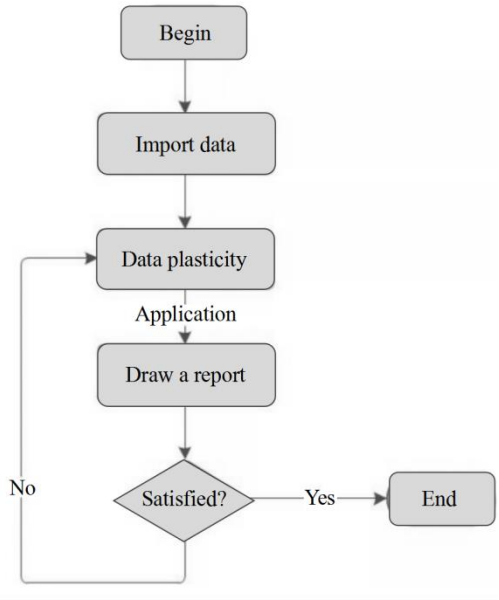

Figure 2: PowerBI visualization flow chart

2.3 Development environment

Hadoop-based e-commerce data analysis system, according to the data volume and overall operation requirements of the system, uses three cluster servers to build Hadoop3.3 cluster to ETL data, as shown in Figure 3. The cluster will be developed under Linux system. This paper selects CentOS-8.2 release version of Linux enterprise operating system. One of the nodes is named Hadoop102 as the master node to store NameNode and DateNode, and the other two nodes are slave nodes, namely Hadoop103 and Hadoop104. Hadoop103 stores HDFS DataNode, hadoop104 stores DateNode and SecondaryNameNode. Hadoop cluster needs to run components including Hive1.2, Mapreduce, Flume1.8, Sqoop-1, etc. In this system, PowerBI desktop Edition is used to visualize the data such as drawing maps, pictograms, circular overlapping maps, line charts, folded column mixed maps and radar maps. By installing OOBC components, Hive data sources are configured in desktop connection data sources for connection [10].

After power bi desktop nodejs vs code, configuring agent npm config set proxy http://usernsme:password@server:port and npm config set https-proxy http://usernsme:password@server:port4. to install pbiviz npm i -g powerbi-visuals-tools, create the installation certificate pbiviz --install-cert. (After the installation is completed, there will be password information of the certificate, which can be used for importing the certificate later.) Finally, open the certificate import wizard pbiviz --install-cert to complete the construction of PowerBI front-end development environment. Through the introduction of the above key technical theories, the overall environment of system development, the configuration of related software and tools are determined, and the technical feasibility of the overall project is also defined.

	hadoop102	hadoop103	hadoop104
HDFS	NameNode DataNode	DataNode	SecondaryNameNode DataNode
YARN	NodeManager	ResourceManager NodeManager	NodeManager

Figure 3: The Hadoop cluster settings

3 Requirements analysis

3.1 Functional requirements

Hadoop-based data analysis system uses appropriate statistical analysis methods to analyze a large amount of collected data, summarize, understand and digest them, in order to maximize the development of data functions and play the role of data. E-commerce data analysis system aims at the records of consumers' visits or purchases on websites or platforms, including geographical location distribution, visiting time, click records, goods purchased, purchase quantity, payment amount and other information, and analyzes their trading behaviors based on operational data, which effectively helps e-commerce enterprises to formulate and optimize targeted e-commerce operation strategies.

3.2 Global design

The main data of Hadoop-based e-commerce data analysis system comes from the server-side data and local database of e-commerce websites operated by enterprises. Use Flume component to obtain server log data and save it in HDFS. Then mapreduce cleans and calculates these unstructured data. In this process, yarn is used to schedule resources. After that, Hive is used to save the data results of structured logs and complete the data processing. Hive and MySQL are connected through sqoop, and finally saved in MySQL relational database. This system uses PowerBI for secondary development to realize web function and visualization of charts, and calls MySQL data for development. As shown in Figure 4, it is the overall framework diagram of the system [9].

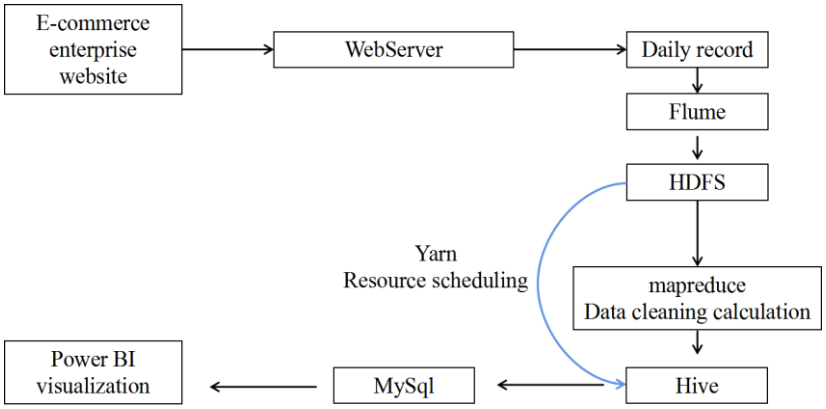

Figure 4: System overall framework diagram

4 Function implementation

Hadoop-based e-commerce data analysis system is user-oriented and positioned as the manager of e-commerce enterprises. According to the needs of users, four functional modules are developed and set up, which are traffic data, conversion rate data, user data and order data. Each functional module adopts data visualization, and from the perspective of rich visualization, it shows the more vivid meanings that data in different scenes can express in many ways. There are six kinds of charts: map, pictographic column chart, circular overlapping chart, line chart, folded column mixed chart and radar chart.

4.1 Traffic data

This module mainly displays the main traffic indicators of the seller's shop, and the traffic data includes the click volume of the e-commerce website shop and the visit volume of each commodity. Generally speaking, website traffic refers to the number of visits of websites, which is used to describe the number of users visiting a website and the number of web pages visited by users. Through traffic analysis, we can find the rules of users visiting websites and improve website design or marketing strategies according to these rules. Traffic analysis can be carried out from two aspects: traffic quantity and traffic quality. Traffic includes UV unique visitors, PV page views, and per capita page views. Traffic quality includes average visit depth, average stay time and bounce rate. The above data will show the trend in a line chart, which will be updated by the system administrator on a weekly basis.

4.2 Conversion rate data

The conversion rate is the ratio of all the people who arrive at the corporate website and make purchases to all the people who arrive at your store. Conversion rate analysis can help enterprises monitor the conversion of users' purchase paths, calculate different conversion rate and churn rate data, and then optimize products or pages. The system counts the conversion rate of each step through the number of visitors n, the number of orders m and the number of payments q. The visitor conversion rate is shown in Formula (1), the order conversion rate is shown in Formula (2), and the payment conversion rate is shown in Formula (3) [3].

$$\text{The visitor conversion rate}=\frac{\text{The number of orders}}{\text{The number of visitors}}$$

$$X=\frac{m}{n} \tag{1}$$

$$\text{The order conversion rate}=\frac{\text{The number of payments}}{\text{The number of orders}}$$

$$X'=\frac{q}{m} \tag{2}$$

$$\text{The payment conversion rate}=\frac{\text{The number of payments}}{\text{The number of visitors}}$$

$$X''=\frac{q}{n} \tag{3}$$

4.3 User data

Analysis users can be simply divided into new users and old users in terms of categories; According to the quality of users, they can be divided into active users, sleeping users, member users and so on. When analyzing users, it is necessary to analyze the repurchase rate to measure the loyalty of users, and the repurchase rate will affect the follow-up strategy of enterprises. If the repurchase rate is not high, it means that there is little dependence on new customers. If the repurchase rate is high, the operation should focus on improving customer loyalty. Moreover, the system will classify consumers by region, and get the visual interface of the map. You can intuitively feel the regional distribution of consumers through the depth of color, and you can see the proportion and number of users in a certain region through mouse hover. By analyzing the distribution of buyers of goods, merchants can plan ahead better in the key arrangement of marketing [2].

4.4 Order data

Order analysis not only refers to the analysis of the paid order quantity, but also the total order quantity, the cancelled order quantity and the complained order quantity, so as to help the merchants analyze the reasons for the success and failure of the order transaction. The e-commerce website operated by the enterprise will make a refund based on the delivery status of the seller, the feedback of the buyer and the business rules, and then generate corresponding records. This data analysis system will collect these data in the background offline state, and the code of the collected data results is shown in Figure 5. Consumers communicate their commodity needs or complaints through e-commerce platform's chat service evaluation, etc. This system will summarize these structured text data, which will help enhance the enterprise's emphasis on customer service. E-commerce companies can make consumers feel valued when purchasing goods or enjoying services, thus providing consumers with high-quality and efficient online purchasing services.

```
CREATE SET TABLE gdw_ _tables.dw_ ems. _dsbrs_ TRANS ,NO FALLBACK,
EMS_ DSBRS_ _TRANS_ ID DECIMAL(18,0) NOT NULL,
EMS_ ESCRW_ PYMNT. _ID DECIMAL(18,0),
EMS_ ORDER_ _ID DECIMAL(18,0),
EMS_ ORDER GROUP_ ID DECIMAL(18,0),
CLIENT_ ID BYTEINT NOT NULL,
SITE_ ID DECIMAL(4,0) NOT NULL,
PAYER_ _ID DECIMAL(18,0),
PAYEE_ JID DECIMAL(18,0),
DISBRS_ TYPE_ ID DECIMAL(4,0) NOT NULL,
CURNCY_ _ID DECIMAL(9,0) NOT NULL,
DSBRS_ _AMT DECIMAL(18,2) NOT NULL,
DSBRS_ STS_ JID DECIMAL(4,0),
PG. _TRANS_ ID DECIMAL(18,0),
RSN_ CD DECIMAL(4,0),
PYMNT_ OPTION_ ID DECIMAL(4,0), .
RELEASE AMT DECIMAL(18,2),
PRIMARY INDEX NUPI_ EMS_ DSBRS_ TRANS ( EMS_ DSBRS_ TRANS_ ID );
```

Figure 5: Refund record code

5 Conclusions

Today's e-commerce is no longer a concept, but a reality deeply rooted in people's hearts, as evidenced by thousands of goods and buyers and sellers. The traditional management relies on people's minds, and the management in the Internet era has spawned more new methods. In this paper, small and medium-sized e-commerce businesses are selected as the target user groups, and based on the actual trading situation of such businesses, the data generated by the businesses in the transactions are statistically analyzed and visually displayed, so as to provide better data-driven decision-making analysis for the businesses from different dimensions and perspectives. Hadoop-based e-commerce system is expected to enhance the reference value of big data to various enterprises' operations, optimize the effectiveness of big data in e-commerce operations, and help enterprises improve their operating efficiency, so as to realize the application of big data technology in e-commerce enterprises' business activities. However, there are some limitations in the research, that is, the limitations of investigating the sample of the research object. Due to the limited number of surveys, it is inevitable that there will be omissions, and it is impossible to show the whole picture of the application of big data in e-commerce operations by enterprises in all industries. In the follow-up research, we can consider industry differences and conduct in-depth research. In addition, the influence of e-commerce operating years on various factors is not considered in this research and analysis. As there are many influencing factors in the process of applying big data in enterprises, follow-up research can explore the influence of other factors on the application of big data.

References

[1] Cai Xiaowei (2021). Analysis of Electronic Business Operation Data. Internet Commerce.04.

[2] He Jiangnan (2021). Research on Big Data Analysis in E-commerce. Computer Knowledge and Technology.09.

[3] Liu Baotai (2019). Discussion on the Application of Visual Analysis of Financial Statements. Accounting Learning.03.

[4] Li Xinpeng (2020). Overview of Big Data Analysis in E-commerce. E-commerce.11.

[5] Ren Kecheng, Li Xiaojiang (2021). Development Strategy of E-commerce Based on Big Data Analysis. China Circulation Economy.03.

[6] Song Xiaoqing, Liu Kunbiao (2020). Application of Business Intelligence Based on Big Data Analysis Technology in E-commerce Data Analysis. Market Modernization.03.

[7] Wang Haiou (2021).Research on Application of Big Data Analysis in E-commerce Marketing. Inner Mongolia Science technology& Economy.11.

[8] Wang Yu (2017). Design and Implementation of Electronic Business Data Visualization System. Xiamen University.06.

[9] Yin Xiankun (2019). Design and Implementation of User-centered E-commerce System in Distributed Environment. Liaoning University.05.

[10] Zheng Jieru (2020). Analysis of the Application Status and Countermeasures of Big Data in Enterprise E-commerce. Journal of Beijing College of Finance and Commerce.04.