ISSN No:-2456-2165

The Design and Implementation of E-Commerce Log Analysis System Based On Hadoop

Weiyi Ma
School of Computer Science and Technology
Shandong University of Technology
Zibo City, Shandong Province, China

Dongmei Zhang
School of Computer Science and Technology
Shandong University of Technology
Zibo City, Shandong Province, China

Abstract:- The use of the website will generate a large number of access logs. With the development of the access log collection and processing technology, the access log-based processing can obtain some useful decision data, which can assist the operator in strategic decision-making and monitoring. With the increase of companies and network users, more and more website systems, the amount of data generated is getting larger and larger, and the processing of data is becoming more and more difficult. Based on this problem, the system was developed. The traffic information of the system comes from the user log data collected by the Nginx server and the front-end JS burying point, and then the collected log data is spliced into the URL as parameters. This system mainly uses Hadoop, Mapreduce, HDFS and other technologies, and uses Idea as a development tool. After testing, the system function is simple and convenient, in line with the company's processing data requirements, can produce higher profit value.

Keywords:- *E-commerce log analysis; Hadoop; Flume; HDFS; MapReduce;*

I. INTRODUCTION

A. Subject background and significance

a) Background

This is an era of big data, and discovering the value of big data is the hottest topic of this era. In this era, Internet-based social networks, e-commerce, and mobile services push data from gigabytes to terabytes or even petabytes. What this thesis studies and solves is the storage and structured processing of the problem of "big data", so as to provide computers for computing and analysis of structured data [1].

"Big data" will be an opportunity and challenge for the development of the new era. When the amount of data is sufficient, humans will be able to discover the relationship between objects that used to be irregular. This will be a revolution. Emerging technologies such as artificial intelligence, cloud computing, and machine learning will have unprecedented development based on big data.

b) Meaning

This paper studies a system that can satisfy the data structure cleaning of unstructured log files generated by e-commerce websites. At the same time, the system supports horizontal expansion, which can realize massive data storage and data analysis at a low investment cost. The system requires reliable and permanent storage of data and can expand the analysis module.

This article uses JS embedded points to simulate the generation of website access logs to obtain targeted user behavior information, such as which page the user spends the most time on, repeatedly clicking the navigation bar, and so on. When a large number of users stay in a certain content for a long time, this type of content can be increased to attract users. However, when some users repeatedly click on the navigation bar, it means that they are not interested in the content, or the navigation bar is too complicated, and the user cannot clearly find the content they want to see. It also shows that the keyword search is too outdated and needs to be updated. Through the online behavior of users, formulating promotional strategies, and optimizing the structure of the website can the vitality and competitiveness of the website be maintained. By analyzing these data, some hidden values and shortcomings can be obtained, which can help companies make correct marketing decisions.

B. Introduction to key technologies

a) Introduction to Hadoop

Hadoop is an open source and highly scalable distributed computing framework with a wide range of applications. It mainly has four core modules: Common, HDFS, Map Reduce and Yarn. Hadoop can write calculation models, use distributed server clusters, and perform distributed processing on massive amounts of data according to user-defined calculation logic. Hadoop is widely used in big data storage, analysis and other fields [2]. It can increase storage capacity, reduce costs, and process data efficiently and safely. Most of the massive data processing uses this open source software, which greatly saves development expenditures and improves development efficiency .

b) Introduction to HDFS

HDFS is a distributed system used to store massive amounts of data. It is stored in blocks. Each block has a fixed size and is stored in different racks. Each rack has a backup of blocks on other adjacent racks. HDFS has high reliability. When some files are damaged, they can be found through the heartbeat mechanism and the backup files can be found, so as not to affect the normal operation of the cluster.

c) Introduction to HDFS

An MR program needs to specify the mapper class and reducer class used by the business and the job input path. First start the app master to apply for the map task resource to start the corresponding number of map tasks, read the file according

ISSN No:-2456-2165

to the file format, and form the corresponding KV pair. Pass the corresponding KV pairs to the map() method, and then perform code calculations. Collect the KV pairs output by the map method into the ring buffer, and then overflow to the overflower. After partitioning, sorting, grouping, and then overflowing to On the disk, these files are merged into a single large file, that is, each map task eventually generates an overflow file.

After the App master monitors that there is a map task process task completed, it will start the corresponding number of reduce tasks according to the specified parameters and inform the reduce task to be processed. After the process is started, the App master will inform the location, and then pull the data of the same key. , The generated small files are merged into a large file, and the reduce task is notified of the scope to be processed. After the process is started, the App master will inform the location, and then pull the data of the same key, and the generated small files are merged into one large file. Data with the same key is divided into a group, the reduce method is called for each group of data, and the display result is output. This is the overall process of MR. Through MR, you can perform complex operations on the data to achieve the desired effect.

II. SYSTEM DESIGN

A. *System overall architecture design*

The system is divided into the following modules according to requirements:

- User information analysis module: Analyze the data and count the data under the two dimensions of platform (platform) and date (date): the number of new users, the number of active users, the number of total users, the number of new members , The number of active members, the number of sessions, the length of the eight indicators.
- Browser information analysis module: Analyze the data and count the data under the three dimensions of platform (platform), date (date) and browser (Browser): the number of new users, the number of active users, and the total number of users. Number, number of new members, number of active members, number of sessions, length of sessions, platform information, date information, and browser information.
- Geographical information analysis module: Analyze the data and statistic data under the three dimensions of platform (platform), date (date) and area (area): the number of new users, the number of active users, the number of total users, The number of new members, the number of active members, the number of sessions, and the length of sessions.

The system module structure diagram is shown in Figure 1.

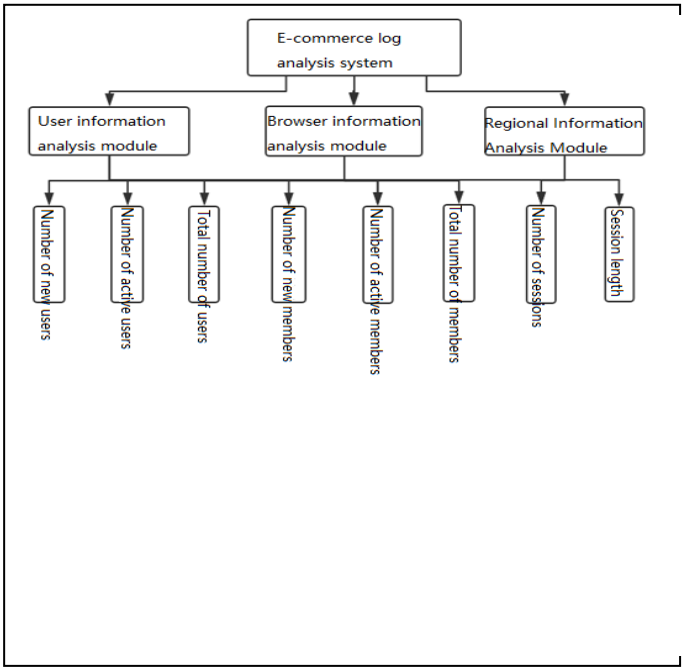

Fig. 1 Camera client business logic flow chart

B. *The design of the system business process*

a) System structure design

Data source: The behaviors such as mouse hover, page dwell time, page click, and browser cache clearing on the front-end page of e-commerce are simulated by JS. Send the front-end JS embedded point data to the Ngnix server, splice the collected data information into the URL in the form of parameters, and Nginx obtains the URL information and saves it to the access.log file on the server.

Data collection: The click data generated on the front-end page is stored in the access.log file of the Nginx server, and the Flume architecture is designed. Flume is used to import the log data in the access.log file to the specified HDFS directory. Import data once a day as the import directory for MapReduce analysis. It is convenient for log analysis based on different dimensions in the future [3].

Flume collection is connected by agents. Each one is composed of three parts, the source collection component, in this article, the data source is connected to the access.log file on the Nginx server to obtain data. Channel transmission channels, the number of Channels can be multiple, the data obtained by Source is passed to Sink through Channel, Channel is not only used for channel, it is like a queue, which can perform temporary storage function. Sink is a sinking component used to connect to an external storage system. This article is used to connect to a directory on HDFS. Through the configuration of these three components, Flume can upload the data in the access.log file to the distributed file system HDFS for data collection. The components of the Flume acquisition system are shown in Figure 2.

ISSN No:-2456-2165

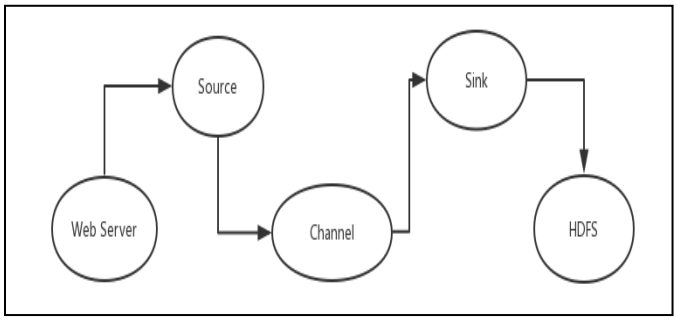

Fig. 2 Flume acquisition system components.

b) Design of Flume acquisition architecture

Name the three components of source, sink, and channel on the agent in Flume to facilitate the call and configuration of the components. Set the type of the component to the exec command type, describe and configure the Source. According to the file name tracking, get the content of the access.log file under the Nginx server. Sink sink target, specify the HDFS external file system.

Configure the sink, specify the HDFS directory to import the collected data, and change the directory every 24 hours. Set the following operation settings for files and directories: the waiting time before file rolling is set to 30 seconds; the size of file rolling is set to 1024 bytes; after 20 event data is written, the file starts to be rolled; every time 5 events are enough, it is written to HDFS ; Format the directory with local time.

After sink sinks, the file type of Sequencefile is generated. Use a Channel to buffer events in memory, set the channel type to Memory, and bind the source and sink to the channel.

- *Data cleaning:* Data cleaning saves the log data in the form of URL into the distributed file system after reasonable splitting, parsing and other operations, and basic preprocessing of the data to facilitate our subsequent statistical analysis. This is data cleaning. . The cleaning process is completed by writing MR code. The log data directly obtained is messy and has various problems, which is not conducive to data analysis. Through data cleaning, we can perform the following operations on log information: remove data without time stamps; filter missing data in fields; crawlers Generated data; parse platform and browser information: platform ID, browser ID, version number; parse URL: get attributes in the table; filter data with incorrect field data format; filter static resource data such as JS and CSS; convert time format; The parsed data is stored in HDFS as formatted data. Data cleaning is an indispensable step for log analysis. It is generally cleaned according to functional requirements to facilitate subsequent data analysis.
- *Data analysis:* The system reads log data information from HDFS, then cleans the data to obtain the required data information, then analyzes the data information, and then saves the results to the MYSQL database.

This article completes the analysis of the following functions.

- Analyze the changes in the number of users, the number of new users every day and the number of total users. New calculation rule for total users: Calculate the number of deduplication of u_ud under each lunch event, and the calculation result is the number of new users. The calculation rule for total users: On the basis of the same dimension, the total number of users = the total number of users on the previous day + the number of new users on the day.
- Analyze the activity level of ordinary users, that is, the number of active users and the number of active members. The calculation rule of active users: the number of active users is equal to the number of u_uid deduplications in all data of the day. The calculation methods for active members and active users are similar. The difference lies in analyzing the number of users visiting the website from different dimensions. Active member calculation rule: in the time dimension, in the pageview event, the number of member_id removed on that day
- Analyze the active level of members, that is, the number of new members and the total number of members. New membership calculation rule: the number of member_id members who visit the website for the first time under the same dimensions for the data of the day. Total number of members = total number of members the day before + number of new members today
- Analyze the conversation, that is, the number of conversations and the length of the conversation. Session analysis calculation rule: The number of sessions is the same as the number of all u_sd. The length of the session is the length of each session when the dimensions are the same, and then the sum of the lengths of all sessions is calculated.
- Analyze regional information. Geographical information analysis is actually to analyze the distribution of active users in different geographic locations and some analysis of bounce rate, so we have to count the following information: the number of sessions, the number of bounced sessions, and the number of active users. Calculate the data of these three indicators at three levels, namely the national level, the provincial level, and the city level. The number of active users, count the number of deduplication of all u_uids in the page view data on the day; the number of sessions is the number of deduplications of u_sd in the pageview event; the number of sessions that jump out is the number of counts where u_sd has only appeared once number.

- *Data results:* The result of log analysis shows the information contained in the log or even the hidden information. Through the form of tables and graphs, it is possible to understand the activity level of users, the proportion of new and old visitors, the number of new members, and the number of new members. The number of new members, the number and duration of sessions, the number of users in different regions, the number of active members, the number of active users, the total number of new users and total members, etc. Through the analysis of this information, we can understand how users feel about the website. Loyalty helps managers make correct

decisions, and uses database tables to sort out this information for other departments to use the data.

III. SYSTEM IMPLEMENTATION

A. System environment construction

a) Hadoop cluster construction

When building a Hadoop distributed cluster, use VMware to install 3 virtual machines and install the corresponding software. Then change the default configuration file. The number of copies needs to be set in the configuration file, and the number set in this system is 3 copies. After this setting is made, each data block in HDFS will be copied 3 copies and stored on the datanode node. In this way, when some of the nodes are down, the integrity of the data will not be affected. When building a hadoop cluster, pay attention to the synchronization time between virtual machines, set SSH password-free login and turn off the firewall.

b) Flume monitor log directory

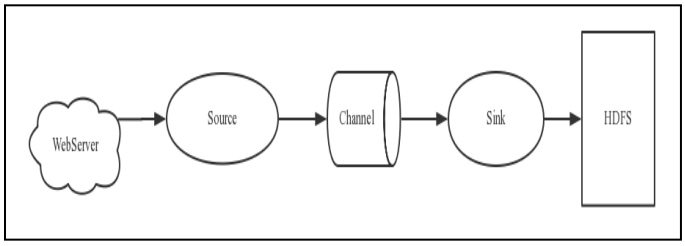

Fig. 3 Flume structure diagram

Configure the Flume component to listen to the specified directory of Nginx, and pull it to HDFS when a new log file is generated. Flume is a message queue project under Apache. Its structure is shown in Figure 3.

Source is an object that collects data. The data generated by the client is specially formatted and encapsulated in an event object, and then the object is pushed into the Channel. Channel is a data pipeline connecting Source and Sink. Events can be temporarily stored in memory or permanently on a local disk until Sink processes the event. Flume provides a large number of channels Memory Channel, JDBC Chanel, File Channel, etc.

MemoryChannel can achieve high-speed throughput, but data integrity cannot be guaranteed. FileChannel guarantees the integrity and consistency of the data. When configuring different FileChannels, it is recommended that the directory set by FileChannel and the directory where the program log file is saved be set to different disks to improve efficiency. Sink pulls data from Channel and stores it in the corresponding persistence system.

B. ETL Module Implementation

First implement the custom data type and implement the Writable interface. This class is used to encapsulate the log data after cleaning [3]. Then use MapReduce to pull log files from HDFS, and divide and filter each line of log data according to the agreed delimiter on the map side. Submit the ip field in the log to Taobao ip library for analysis. The ip library returns a json string containing the ip address information to the local method. This method uses the Ali json parsing object to parse the string into a map set containing the country, province, and city. At the same time, the map side also implements the method of parsing the browser information parameters in the log. This method parses the browser information field to include detailed information such as the browser name and version number. The Map side encapsulates the above data into a custom class and submits it to the reduce side for aggregation and writing out to the HDFS file system. The log data goes through the following steps from the local file system to the specified file system:

- The files that need to be cleaned are fragmented, and if no special settings are made, they are fragmented according to the data block size.
- Each shard is calculated by a maptask process.
- When MapTask pulls data, the formatted input is in the form of key-value pairs, the key is the data offset, and the value is the data text.
- After the data is processed on the Map side, key-value pairs are continuously written to the ring buffer (the format is defined by yourself).
- When the ring buffer is 80% full (settable), this part of the data is locked, the file is overwritten to the local disk, and 20% is reserved for the map to continue writing. When the overflow is written to the file, it is partitioned by key value, and quick sorting is performed at the same time.
- After all the data is overwritten as small files, perform quick sorting and merge into large files by partition. Then press the partition to be pulled into the corresponding temporary buffer.
- The ReduceTask process pulls data from the temporary buffer, aggregates the same partition data from different Map terminals according to the key value, and outputs it to the file system according to the defined output key value format.

a) Parse the IP field

Custom tool class IpAnalyze. In this class, you need to define an internal class to encapsulate the parsed geographic information. In the method ipParser2, send the Http request to the Taobao IP database to request the query of the obtained IP string. When the IP library finds the matching IP address, it will return a json string. First, you need to decode the information, then use the Json object to load the content, and finally get the geographic information through Ali's Json object method and encapsulate it into an internal class.

b) Parse the browser information parameter field

Define the handleParam method. This method divides the obtained parameter field into a string array, and then uses URL decoding on the second element in the array. This element is a URL string containing browser information. The decoded data Return as content.

C. Realization of system functions

Analyze the data and count the data under the two dimensions of platform (platform) and date (date): the number of new users, the number of active users, the number of total users, the number of new members, the number of active members, sessions The eight indicators of the number of sessions and the length of the session.

- The files that need to be cleaned are fragmented, and if no special settings are made, they are fragmented according to the data block size.
- Newly increase the number of users: count the number of duplicates in the uuid field of en=e_l in the log after cleaning.
- Number of active users: count the number of duplicates in the uuid field in all log data of the day.
- Total number of users: the total number of users the previous day plus the number of new users that day, if it is the first day, it is the number of new users that day.
- The number of newly added members: Take out the member_id field of the day and compare it to the member information table in the database, and count the number of member_id that does not exist in the table.
- Number of active members: count the number of memb er_id deduplication of en=e_pv in the log data of the day.
- Number of sessions: Calculate the number of all u_sd fields in the current log file.
- Session length: Each session has an open timestamp and a close timestamp, calculate the time difference of this session, and finally calculate the total value of these lengths.

The result of log analysis shows the information contained in the log and even the hidden information. Through the form of tables and graphs, it is possible to understand the activity level of users, the proportion of new and old visitors, the number of new members, and the number of new members [4]. The number of users, the number and duration of sessions, the number of users in different regions, the number of active members, the number of active users, the total number of new users, the total number of new members, etc. Through the analysis of this information, we can understand how users feel about the website. Loyalty helps managers make correct decisions, and uses database tables to sort out this information for other departments to use the data.

IV. CONCLUSION

This subject implements an e-commerce data log analysis system suitable for small and medium-sized Internet companies [5]. This system uses JS embedded points to generate logs, builds Nginx server to monitor log information under the access.log file, uses flume tool to import the log files from the server to HDFS storage devices, writes Map Reduce code on IDEA to realize the data cleaning process , Through calculation and coding, analyze the number of new users, the number of new members, the number of new users, the number of active members, and the number of active users in different platform dimensions, time dimensions, and browser dimensions. , The duration of the session, the number of sessions and other information. Based on this information, the status of user information can be known, which is convenient for operators to make business decisions, so that economic benefits far exceed development costs.

ACKNOWLEDGMENT

Dongmei Zhang is the corresponding author.

REFERENCES

[1.]Liem Gai Sin,M.Bus and Ria Purnamasari.China E-commerce Market Analysis:Forecasting and Profiling Internet User.IEEE,July 2011,pp.79-82

[2.]QIqi Jianga,Chuan-Hoo Tan.Understanding Chinese online users and their visits to websites:Application of Zip's law.International Journal of Information Management,October 2013

[3.]Ren Chongguang. Research on cloud computing and its key technologies in the field of massive data processing[D]. Nanjing: Nanjing University of Science and Technology 2013

[4.]Tao Xiaoying. Changes to improve the data processing capabilities of data analysis systems in the era of big data[J].Beijing: Information and Communication Technology,2016,1004:16-23.

[5.]Wang Qiang,Li Junjie,Chen Xiaojun,Huang Philosophy,Chen Guoliang.Overview of the construction and application of big data analysis platform[J].Beijing: Integration Technology,2016,502:2-18.