Design and Implementation of Hotel Big Data Analysis Platform Based on Hadoop and Spark

Dianwei Chi1*, Chunhua Tang2 and Chen Yin3

1 Artificial Intelligence Academy, Yantai Institute of Technology, Yantai, Shandong, China
2 Shandong Vocational University of Foreign Affairs, Weihai, Shandong, China
3 Shandong Vocational University of Foreign Affairs, Weihai, Shandong, China
*Corresponding author’s e-mail: chidianwei@yitsd.edu.cn

Abstract. In order to provide decision support for user travel and hotel management planning, this paper proposes the design and implementation of a hotel big data analysis platform based on Hadoop and Spark technology, using hotel basic data and user comment data as data sources. First, we design a crawler program to collect basic data and user comment data of city hotels in Qingdao, Shandong Province, and upload them to the hadoop storage platform, use Spark for big data preprocessing, use Hive to build a big data platform data warehouse, and finally use echarts to analyze the results and make decisions perform a visual display.

1. Introduction
With the development of computer networks, large-scale websites and platforms are updated in real time [1], generating a large amount of data. In the context of today's big data [2], all walks of life have accumulated massive amounts of data, which have the characteristics of large data capacity, multiple types, fast data growth, and high value density. Many scholars have also launched research on big data analysis algorithms, analysis modes and analysis software tools [3]. Among them, great progress has been made in the structural model of big data and the theoretical system of data science, the basic theories of big data analysis and mining [4], and the application of big data has also expanded from science, engineering, telecommunications and other fields to all walks of life [5]. In China, many large-scale hotels have their own hotel management systems, providing comprehensive hotel management, hotel reservations, and evaluation services. Due to lack of investment, some small and medium-sized hotels rely on third-party platforms to provide online services. After customers perform operations, the third-party platforms will generate records and save them [6]. The hotel has accumulated a large amount of online basic data and user comment data for a long time. For the hotel industry, how to use big data technology to process and analyze existing data and help hotel practitioners and travel users provide intuitive reference decision-making is a problem that we urgently need to solve. On the one hand, based on user online comment data, it helps hotel practitioners provide intuitive decision support and improve hotel management to maximize profits; on the other hand, it provides data visualization on hotel distribution, hotel satisfaction, and hotel popularity.

In order to solve the above problem, this paper proposes a design and implementation plan for a big data analysis platform based on city hotels.
The platform designed in this paper builds a big data platform data warehouse based on the basic data of city hotels in Qingdao City, Shandong Province and user comment data, and performs statistical analysis. Finally, the analysis results and decisions are displayed in the form of visual charts in the form of Web pages. It can not only provide a reference basis for users to travel and stay in a certain city, but also provide certain decision-making support for hotel practitioners, so that they can understand the distribution of hotels in each district, satisfaction, user travel purpose, etc. in advance in the process of preliminary market research. For example, it is possible to plan hotel types and related supporting facilities for hotel practitioners based on information such as the proportion of user travel types.

2. **Hotel Big Data Analysis Platform Design**

The main framework adopted by the big data analysis platform designed in this paper is the Hadoop ecosystem component and the Spark ecosystem [7] component. The core of the Hadoop framework includes HDFS and MapReduce. Among them, HDFS is NGFS[9] proposed by the founder of Nutch[8] in conjunction with GFS, and later integrated MapReduce. This section will introduce the specific design plan and implementation steps of the hotel big data analysis platform.

### 2.1 Overall Platform Design

The process of a general big data analysis system is: data collection, data storage and cleaning, data analysis, and data visualization. According to the typical process of data analysis, this paper proposes a design scheme suitable for hotel big data platform. First, perform big data analysis and processing on hotel reviews and hotel basic data in Qingdao, Shandong. The data is stored in the hadoop cluster. After data cleaning, a hive data warehouse is constructed, and data analysis is performed based on the hive warehouse, and the analysis results are finally imported into mysql. Finally, a web project based on JavaEE is constructed to visualize hotel data. The design process of the platform is shown in Figure 1:

![Figure 1. Hotel big data platform design process](image)

### 2.2 Platform design process steps

Perform big data analysis and processing on the basic hotel data and user comment data of a certain city. Here we take Qingdao hotel data as an example, and store the final analysis results of the platform in mysql. The overall design is mainly realized from the following steps:

1. Design a crawler to crawl the basic hotel information data and user comment data of a prefecture-level city to form a data source, and the files are stored in csv format.
2. Upload the data source file to the Hadoop platform, use the HDFS of the Hadoop platform to store the data, and use the hdfs module of Python to implement the upload technology.
3. The spark program is used to clean the uploaded data to make it meet the basic data requirements of the big data analysis platform. The main tasks are as follows:

(1) The name of the column of hotel star type in the hotel basic data set is inconsistent. For example, some hotels are called four-star hotels and some are called high-end hotels. "Star" is replaced with "High-end type", "National Tourism Administration rated as a three-star" replaced with "Comfortable", "National Tourism Administration rated as a two-star" with "Economy", and "National Tourism "Five-star rating by the bureau" is replaced with "luxury type".

(2) Since this sub-module of the big data service platform does not perform sentiment analysis on user specific comment content, sentiment analysis is handed over to the sentiment analysis subsystem to process, so the comment content data column is removed here.

(3) Delete all blank lines.

(4) Extract the district/county name from the hotel address and replace the address in the column to provide standard data for district/county hotel distribution statistics.

The cleaned data is stored in two different directories of hdfs. For example, the directory named hotelbasic is used to store basic data, and the name hoteldata is used to store comment data.

4. Construction and analysis of data warehouse. First, create hive external tables based on the storage paths of the two data sets in hdfs, which are the user comment data table (hotel_data) and the hotel basic information table (hotel_basic). Based on the two Hive external tables, the data is analyzed and processed according to the dimensions of the hotel and review information that the user cares about, divided into 5 attention angles, and a hive internal table is created for each of them, as follows:

(1) User impression statistics are also the overall satisfaction of users in the region. Users who have stayed in hotels can also comment and score. The following is the statistics of the user's overall impression based on the user's overall rating of hotels in the region or city. The score of 4.5-5 is excellent, 3.5-4.5 is good, and the score is below 3.5 is bad. The proportion of hotel user ratings is calculated.

(2) Count the top ten hotels with the most online reviews and top ten popular hotels on the Internet. Under normal circumstances, the number of reviews of a hotel can represent the popularity of the hotel. The statistics here are the name of the hotel and the number of reviews.

(3) Statistics on the proportion of different types of tourism. According to the external table hotel_data of user comments hive, the statistics of different types of tourism are carried out, and according to the type of tourism combined with user satisfaction, it provides a reference for users to travel and understands which type of tourism is more suitable for this area.

(4) Hotel star distribution statistics, design hotel star and quantity two attributes, show the proportion of the number of hotels with different star ratings, and provide users at different levels with the number distribution of star hotels.

(5) The distribution of the number of hotels in different districts of the city is presented as a heat map. At the same time, the number of hotels in each area and the average review score need to be displayed.

(6) Use sqoop to export the generated internal table data of hive to mysql database.

5. Data visualization part. Based on the analysis results of the mysql database, it is presented on the web page in the form of graphs. Use SpringBoot+mybatis+Mysql to build the project, IDEA as the development tool, and echarts to support the page chart rendering. In order to improve page loading speed and user experience, ajax asynchronous loading method is adopted. Among them, the statistical analysis part based on the basic hotel data includes: analysis of user travel types, statistics of the number of hotels in various regions, and statistics of hotel star status. The statistical analysis part based on hotel user comment data includes: Internet popular hotels and user satisfaction statistics.

3. Realization of Hotel Big Data Analysis Platform

3.1 Crawler collection module
Since the crawler needs to traverse the collected hotel links when crawling the data, and then
automatically download the information iteratively, we need to analyze and set the format of the page URL so that the crawler can automatically traverse the address list for download.

The number of hotel reviews needs to be spliced, and all hotel reviews JSON addresses can be spliced through the hotel ID and review page number. When collecting, you only need to iterate according to the splicing rule to continuously collect hotel reviews. The crawling of Ctrip's hotel data is carried out here, and the targeted realization of the structure and data transmission characteristics of Ctrip's website is carried out. The specific crawler development is mainly composed of the following parts:

Crawler collection module: Analyze the JSON data collected from the downloader module, and then extract the corresponding data and related link resources.

Data pipeline module: It is mainly for setting the data format of the target site, such as the address of the target site (link), user comments (content), and so on.

Data storage module: Process the Item column extracted and sent by the crawler module through operations such as data cleaning and data persistence, and then store the data in the data column to the specified location.

IP Proxy Module: The function of this module is to collect IP from professional websites that provide free IP proxies to the local area to provide available IP proxies for crawlers to prevent the crawlers from being blocked by the site.

3.2 Data Upload and Preprocessing
Build a hadoop platform cluster, and install hive, mysql and other related software. Upload the csv file corresponding to the Qingdao city hotel dataset and user comment dataset collected by the crawler module of the sentiment analysis subsystem to hdfs storage.

For the work done in the data preprocessing section, refer to the description in section 2.2. Data cleaning is implemented using the Spark framework. The data cleaning code is implemented in the Scala programming language.

3.3 Hive Data Warehouse Construction and Data Analysis
First, create a hive external table based on hotel basic data hotelbasic.csv and hotel review data hoteldata.csv to build a hotel data warehouse.

1) Create a hive external table based on the uploaded hotel user review data hoteldata.csv, the script is as follows:

   ```sql
   create external table hotel_data(user_name string,hotel_name string,trip_type string,time1 string,time2 string ,user_score double) ROW FORMAT DELIMITED FIELDS TERMINATED BY',' LOCATION'/hoteldata';
   ```

2) Create a hive table based on the uploaded hotel basic data hotelbasic.csv, the script is as follows:

   ```sql
   create external table hotel_basic(id string,name string,score double,commentnum int,recommend int,address string,star string,stardetail string)ROW FORMAT DELIMITED FIELDS TERMINATED BY',' LOCATION'/hotelbasic';
   ```

Then, according to hive warehouse data, analyze the data of city hotel data scores, user impressions, the number and scores of hotels in various districts and counties, network popularity, hotel star ratings, and statistics of tourist purpose types of tourists here. Each statistic is built into hive table, and finally imported into the mysql table.

3.4 Data Visualization
Based on the analysis results of the hotel's basic data and user comment data, provide users with a visual display. Based on the mysql database analysis result as the database, using ajax technology to load asynchronously and display it on the web page. The visualization project uses the popular JavaEE framework SpringBoot to quickly build, and the chart presentation part is implemented using echarts.

Taking the statistics of the number of hotels in various regions of Qingdao as an example, the front-end asynchronous request data code is as follows:
$.ajax({
    type: "GET",
    url: "/areastat",
    dataType: "json",
    async: false,
    success: function (result) {
        json = result;
        for (var i = 0; i < result.length; i++) {
            var ob = {name: "", value: ""};
            ob.name = result[i].area_name;
            ob.value = result[i].nums;
            datatemp.push(ob);
        }
    },
});

Among them, the result in the success callback function is the name of each district and the number of hotels in the city queried through mysql.

Load data on the page through echarts. The statistical analysis part based on hotel basic data includes: user travel type analysis, hotel quantity statistics in various regions, and hotel star status statistics. Among them, the tourism type statistics are shown in Figure 2:

![User’s travel type statistics](image)

Figure 2. Visualization of the user’s travel type statistics

The statistical analysis part based on hotel user comment data includes: Internet popular hotels and user satisfaction statistics. Among them, the hotel star distribution statistics are shown in Figure 3:
4. Conclusions
This article proposes the design and implementation of a hotel big data analysis service platform based on Hadoop and Spark. Taking Qingdao hotel data as an example, it specifically introduces the data crawler module, data upload module, data cleaning module, data analysis module and data visualization display. In the era of big data, the use of massive data storage frameworks and parallel processing frameworks is inevitable, so this article uses Hive data warehouse technology in the data analysis module, and the bottom layer uses Hadoop's Mapreduce framework for calculations. The storage framework uses Hadoop's HDFS to implement distributed storage. The data cleaning module has a large workload, so it uses the Spark framework. Spark's memory-based data processing characteristics make it much faster than MapReduce. The visualization module is based on echarts+ajax to realize the asynchronous display of the chart, which improves the user experience.

REFERENCES
[1] Xu Zhihong, Jiang Xinyu, Dong Yongfeng, et al. Research on an improved collaborative filtering algorithm based on Spark[J]. Computer Applications and Software, 2017, 34(5): 247-254, 278.
[2] Chen M, Mao S, Liu Y. Big data: asurvey[J]. Mobile Net-works and Applications, 2014, 19(2): 171-209.
[3] Meng Xiaofeng, Kindly. Big Data Management: Concept, Technology and Challenges [J]. Computer Research and Development, 2013, 50 (01): 146-149.
[4] Cheng Xueqi, Guo Jiafeng, Jin Xiaolong. A review of network information retrieval and mining [J]. Journal of Chinese Information Processing, 2011, 25 (06): 111-117.
[5] Wang Huajin, Cai Qian. Overview of data mining visualization technology [J]. Science and Technology Plaza, 2009, (01): 235-237.
[6] Dong G.Research on lock and dead lock of SQL server database .Computer and Modernization, 2017(8):52-55
[7] Zaharia M, Xin RS, Wendell P, et al. Apache spark: a unified engine for big data processing[J]. Communications of the ACM, 2016, 59(11): 56-65.
[8] Ghemawat S, GobioffH, Leung S T. TheGoogle file system[J]Acm Sigops Operating Systems Review, 2003, 37(5):29-43.
[9] Cafarella M, Cutting D. Building Nutch[J]. Queue,2004,2(2).