Metro OD Matching and Statistical Research Based on AFC Data

Ailin Jiang¹, Qianqian Sun¹, Zhen Peng¹², Xinyi Wei¹, Qiang Huang¹²,* and Ming Tong¹

¹Sichuan Agricultural University School of Information Engineering, Sichuan, Ya’an, 625014, China
²“Agricultural Information Engineering”Key Laboratory of Sichuan Agricultural University, China

*Corresponding author email: 19623556@qq.com

Abstract. With the development of the smart city system, it is more and more important to improve China's rail transit system and create a convenient traffic environment for passengers. This article will use the AFC data of Chongqing Regional Rail Transit for two months for OD (origin-destination) matching and passenger flow statistics. The main work includes AFC (Automatic Fare Collection) data cleaning, OD information matching, data conversion, passenger flow statistics and data visualization under the big data environment. At the same time, in the construction of the data platform, three schemes are currently adopted. One is to adopt the traditional method of java-Oracle when the amount of data is small. The second is to use the classic method of Hadoop-Hive when the amount is large. The third is to use parallel computing Maxcompute for large amounts of data to improve data calculation speed. Through the experimental comparison of the three schemes, the exact matching of the site, line, and time and passenger flow statistics, in terms of speed, the parallel computing method is used to calculate the fastest data. This paper based on the AFC data of metro OD matching and statistical research can help the transformation of urban traffic structure from extensive to fine and provide support for urban transportation planning and decision-making.

Keywords: Regional Rail Transit; AFC data; OD matching; Passenger flow statistics; Hadoop-Hive.

1. Background Introduction

With the extensive application of China's public transportation IC card system and vehicle-mounted GPS, the subway system has accumulated a large amount of data. As the basic data of regional rail transit characteristics research, OD [1] (Origin-Destination) people statistics can display the OD passenger flow at each site, each line, and each section, reflect the distribution of people, and analyse the dynamic characteristics of the number of people in real time. This article takes the subway IC card data [2] in Chengdu and Chongqing as an example, the whole process of data is completed in about 20 seconds per day and it spends about 10 minutes for 30 days per month. First, the data is cleaned and sorted comprehensively, and then the subway OD matching is implemented [3]. The core is the OD site, line, and time matching. After the OD matching is completed, the site number of the OD is converted to a name to make the results more intuitive. Furthermore, the number of OD people is counted. Within a certain time, granularity, the number of passengers in the OD interval is counted, and finally all the data results are cleaned. Get effective and usable data for urban transportation construction services [4].
2. Problem Analysis
This article performs data processing on the IC card data of Chengdu and Chongqing for two months (Sept. and Oct. 2018). The data is about 24G, totalling about 200 million pieces. There are about three problems. First, the amount of data is large, and it is difficult to clean the data; second, there are many types of data and the data redundancy is large; third, the OD matching and passenger traffic statistics take a long time, and the work efficiency is low. Among them, IC card data records about 100 million pieces a month, 25.63 million pieces of data a week, and an average of 3.66 million pieces a day, with an average of about 42.36 card swiping records per second, including a total of 235 subway card swipe data, 7 types of ticket. The equipment system can accept more than 300 ticket counts per second, and the data upload and transportation are almost every 2 ~ 3 milliseconds. Data fields include 11 fields: ticket number, transaction date, transaction time, ticket type, transaction event, transaction station, pre-transaction balance, transaction amount, post-transaction balance, interline transaction, card counter, device number, and data reception time. The large number and variety of big data also have the problem of noisy and redundant data, which makes the speed of OD matching and passenger traffic statistics change slow, the efficiency becomes lower.

3. Architecture Design
In view of the existing problems, this paper carried out three data processing architectures and gradually optimized them. Finally, the method with the best effect and fast data processing was obtained, which brought convenience and technical support for subsequent research. The three architectures mainly include Java- Oracle; Java-Hive; Alibaba Cloud Maxcompute.

In this article, the OD table uses the Oracle database and Navicate for Oracle[5]. The tool provides a clear interface design, which perform intuitive and efficient SQL execution. It takes about 14 minutes to test the data OD pairing of Chongqing AFC for one day. However, when scheduling and processing massive amounts of data, the Oracle database has a high CPU occupancy rate, fast memory capacity consumption, and large IO overhead on the system disk. Therefore, the second method is used-using the Hadoop Hive framework to store and process massive data.Hadoop is a storage computing platform under cloud computing, and Hive is a data warehouse based on Hadoop, which can use Hadoop's HDFS[6] and MapReduce to provide services for massive data. The architecture diagram is shown in Figure 1. Chengdu-Chongqing regional rail transit IC card data is uploaded to the hive data warehouse through a local file system. It provides high-fault tolerance and high-throughput storage services for massive IC card data.Hive conducts interactive large-scale query and analysis operations, queries OD sites, lines, and time of related OD tables, and uses the redundant backup mechanism in HDFS to prevent data loss[7].Count the number of OD people, and automatically call Spark SQL, a parallel computing framework at runtime, then process the node data, and finally integrate the calculation results by MapReduce and return to the master node.The data processing platform based on Hadoop + Hive[8] architecture can complete tasks such as data processing and calculation faster. the OD pairing of the data for one day in Chongqing is tested for about 3 minutes. However, the IO interactive query performance of the Hive data warehouse is poor, so Alibaba Cloud Maxcompute was tried to make up for this shortcoming of Hive.Big data computing service Maxcompute[9] is a fast and fully managed EB-level data warehouse analysis solution with high performance, good security and low cost. This article uses Maxcompute to test the sample data. Based on the new generation SQL engine, it provides IDE for new data development, Maxcompute Studio, which greatly, reduces the service load, realizes the storage and calculation of massive OD data. Sex. Test the OD pairing of Chongqing's one-day data for about 20 seconds. However, Maxcompute is very different from the SQL used by Oracle and the HSQL used by Hive. There is also room for improvement in standard SQL support.

4. OD Matching

4.1. Data Processing and Storage
The data processing needs to be performed twice. For the first data clean up, the collected IC card data is stored in the structure table of the database in txt format, and multiple parallel SQL scripts are executed at once. After illegal data inspection, preliminary specification, and basic Data cleaning. The processed
data is stored in the Hive data warehouse, and the regional rail transit data is divided into independent data blocks, which are distributed and stored in the HDFS distributed system at the bottom of Hadoop, which can alleviate the storage pressure of massive data. Name Node, Data Node and other nodes automatically make redundant backups to prevent data loss due to node errors. After the OD pairing is completed, the data is cleaned for the second time, which is directly completed in the Hive data warehouse. Deletes the data rows whose inbound time is earlier than the outbound time, and the overlapping data lines ensure the accuracy of the data. This paper uses a collection framework to solve the problem of one inbound corresponding to multiple outbound and improve the accuracy of matching. After cleaning, MapReduce automatically returns the result.

**Hadoop**

**Spark(without sort)**

**MapReduce**

**Figure 1.** Distributed architecture diagram.

**Figure 2.** Site passenger flow chart in September 2018.

**Figure 3.** Workflow of station number conversion.

### 4.2. OD matching

The ID information pairing of AFC data of Chengdu-Chongqing Rail Transit distribution principle is to separate the inbound and outbound data into two tables, and to match the outbound table data with the closest time by circulating the inbound data table. When performing Od pairing, the od matrix is used to reduce the global od mode and enhance the visual design. The master machine and three slaves of the Hadoop / Hive big data storage and processing module are called to perform job scheduling and task distribution. The first step is HDFS's Name Node and Data Node to collaborate to complete efficient data analysis. Based on the inbound and outbound of transaction events, the entire data table is divided into inbound and outbound tables. The inbound table
fields include inbound events, inbound numbers, and inbound time, and the outbound table fields include outbound events, outbound numbers, and outbound time. The second step is to automatically submit Hive SQL statements in Hive to parse into MapReduce tasks. The Spark SQL parallel computing framework automatically performs data parallel operations to achieve high-speed processing. Hive conducts interactive large-scale query and analysis operations. The flow chart is shown in Figure 2. Finally, reduce generates the data processing results and completes the OD pairing based on the OD table. When data is exported, the data will be stored locally in csv format, which effectively completes the process of big data processing, analysis and calculation.

4.3. Data Conversion and Passenger Flow Statistics

4.3.1. Data Conversion. Regional rail transit data has the characteristics of complexity and diversity. This article will convert the station number to its corresponding station name, to make the results presented are more intuitive. by using a comparison table containing all station numbers and station names. Divide the station numbers in the table OD so that the station numbers match the inbound and outbound numbers in the comparison table one by one. In this experiment, Maxcompute is used to fully control the amount of OD data and schedule the conversion of all data.

4.3.2. Passenger flow statistics. Directly analyze the number and proportion of people who choose to take urban rail transit on the Hadoop-Hive data processing platform, and establish a unified data format to achieve the integration and sharing of OD data between multiple systems. Due to the difference in the length of the section, the size of the station, the number of stations in the middle, the transfer route, etc. the daily passenger flow in each section is also different. Transfer is a problem in counting OD passenger flow. Different passengers may choose different transfer stations, different transfer routes, and different transfer times. Based on the above considerations, this paper uses two constrained parameters, OD site and time granularity, as OD statistical parameters. Real-time analysis and calculation of regional traffic OD data of 5 minutes, 30 minutes, 1 hour and 1 day, and record the original card swiping amount of each line of the OD table. Clean the OD initial matching data again, sort the data according to the inbound station, outbound station, passenger flow, time granularity, and convert the date object in the data into a timestamp type[10].

This paper presents the traffic volume for the week of September 2018 and the traffic volume for Sept. and 10 years 2018, with a time granularity of 1 day. its visualization is shown in figures 4,5 and 6. as shown in figure 4, the station number in the figure is separated by 10,000 passenger flow, and the station is arranged in descending order of its passenger flow. But in view of too many platforms, there is not too much display here. A representative Caojiawan, Changhe and other platforms; B, C, D and 24 representative of the passenger flow level of various platforms. X on behalf of pebbles, Guanyin Bridge platform and other platforms. From the overall data of the week, 1-7 September 2018, Total traffic at each site is 10, 652, 882 persons; The highest passenger flow at Guanyin Bridge, Total 420, 801 passengers. Guanyin Bridge Station is the busiest. Guanyin Bridge is the largest pedestrian street in western Chongqing. It's an iconic business district, radiates the population of the city.

In the one-day time granularity, the monthly passenger volume statistics of September and October 2018 show that the number of stations per month is separated by 30,000 passengers in two months, and the passenger volume of stations in the interval is arranged from small to large. The A in the September picture represents Cao Jiawan and other platforms; B, C, D and 24 platforms representing the passenger flow level of various platforms. X on behalf of pebbles, Guanyin Bridge platform and other platforms. From the overall data of the week, 1-7 September 2018, Total traffic at each site is 10, 652, 882 persons; The highest passenger flow at Guanyin Bridge, Total 420, 801 passengers. Guanyin Bridge Station is the busiest. Guanyin Bridge is the largest pedestrian street in western Chongqing. It's an iconic business district, radiates the population of the city.

In the one-day time granularity, the monthly passenger volume statistics of September and October 2018 show that the number of stations per month is separated by 30,000 passengers in two months, and the passenger volume of stations in the interval is arranged from small to large. The A in the September picture represents Cao Jiawan and other platforms; B, C, D and 24 platforms representing the passenger flow levels. Small stone stations have the most passenger traffic, Total 1992, 310. Most passenger traffic is within the A-J range, Between 10,000 and 110,000 passengers. Guanyin Bridge Station is the busiest. Guanyin Bridge is the largest pedestrian street in western Chongqing. It's an iconic business district, radiates the population of the city.

In the one-day time granularity, the monthly passenger volume statistics of September and October 2018 show that the number of stations per month is separated by 30,000 passengers in two months, and the passenger volume of stations in the interval is arranged from small to large. The A in the September picture represents Cao Jiawan and other platforms; B, C, D and 24 platforms representing the passenger flow levels. Small stone stations have the most passenger traffic, Total 1992, 310. Most passenger traffic is within the A-J range, Between 30,000 and 300,000 passengers. The site is the busiest, has a great relationship with its geographical location. The small stone site is the central station of Chongqing Hongya Cave, Chongqing Jiefangbei and Yangtze River cableway. Figure 6 is similar to Figure 5, but the site within the station number in the figure is slightly different. figure 6, A representative cao jiawan, wangjiazhuang and other platforms; B, C, D and 26 representative of the passenger flow level of various platforms. X in October 2018 passenger flow
reached 54000263; The pebble station has the highest passenger flow, Total 2436738. The traffic volume of each station increased in October. The National Day holiday in September, October is the peak of passenger flow. According to these two months of passenger flow, As you can see, The passenger flow of Caojiawan, Wangjiazhuang and Huangmaoping is relatively small. At the intersection, Shapingba, Y the red flag channel, It is highly consistent with the actual situation.

![Figure 4. Site passenger flow chart from September 1 to 7, 2018.](image)

![Figure 5. Site passenger flow chart in September 2018.](image)

![Figure 6. Site passenger flow chart in October 2018.](image)

4.4. Result Analysis
Through the cleaning and analysis of the AFC data for two months, three points can be made clear: First, the accuracy rate of obtaining dynamic OD information for metro IC data is high, which is consistent with the actual situation; Second, the combination of Hadoop computing framework and Hive data warehouse uses parallel computing to greatly improve the speed of calculation and reduce the time of OD matching and passenger traffic statistics; Third, for the statistics of passenger flow, to a certain extent, it can be estimated whether the number of passengers on each line and each station is saturated, reflecting the number of transportation services provided by the entire rail transit network for the general public, which plays an important role in rail builders’ effect.

5. Conclusion
Based on the OD matching and statistical research of Metro IC card data, this paper improves the availability and accuracy of using Metro IC data to obtain dynamic OD information, and provides strong basic support and scientific decision-making basis for the construction of refined subway and high-speed rail. Due to the large amount of data, it leads to slow operation, insufficient storage space and other problems, and more energy and money are spent in optimizing equipment and improving the system, we can further improve the value of this kind of data to the research and application of traffic planning and management in the future by optimizing the matching calculation time, improving the statistical accuracy of data, and refining the time division standard.
Acknowledgments
For the output of this paper, I would like to thank Prof. Huang Qiang and all the team members for their support and help, as well as the financial support of the National Key R&D Program (2017YFB1200702) and the experiment provided by the Sichuan Agricultural University Key Laboratory of "Agricultural Information Engineering" surroundings. This paper is the collective wisdom and belongs to the work of all the people who provide help.

References
[1] Yi Kunnan, Li Zhichun. An effective method for calculating OD matrix from road traffic volume and its application [J]. Operations research and management, 2002, (5). 65-70.
[2] Ma Xiaolei, Liu congcongcongcong, Liu Jianfeng, et al. Study on Calculation of boarding stops based on bus IC card data [J]. Transportation system engineering and information, 2015, (4). 78-84 doi:10.3969/j.issn.1009-6744.2015.04.012.
[3] Jiang Xi, Feng Jiaping, Jia Feifan, Sun Jiping, Li Chunxiao. Modeling and simulation method for large-scale urban rail network passenger flow distribution deduction [J]. Journal of the China Railway Society, 2018, 40 (11): 9-18.
[4] Zhang Lingrui, Yang Yang. Research on improved BPR road impedance function in freight OD matrix inversion [J]. Highway transportation technology, 2016, (7). 125-129144 doi:10.3969/j.issn.1002-0268.2016.07.020.l.
[5] Hao Lin, Zhou Guangwei, Li Junfeng. Analysis of Oracle database backup and recovery strategy [J]. Digital communication world, 2019, (5.57).
[6] Chen Zhong, Fan Kaiyong, Rao Hongbo. Design of distributed transportation big data storage analysis platform based on Hadoop [J]. Computer Programming Skills and Maintenance, 2018, (12): 111-113.
[7] Jia Li, Du Hao, Zhao Yiming, Liang Xiaobin, Wu Si. Design of data mining and service system based on Hadoop distributed architecture [J]. Digital Technology and Application, 2019, 37 (2): 176-178.
[8] Chen Jiao, Zhu Yan, Ding Guofu. Research on hive access control technology in big data environment [J]. Software guide, 2018, (12). 187-190196 doi:10.11907/rjdk.181337.
[9] Li Xinrui. Research on the application and development of MaxCompute SQL [J]. Information Technology, 2017, (5).
[10] Zhang Jinfei, Xue binqiang. State estimation of networked systems based on rolling horizon random packet loss [J]. Electronic design engineering, 2020,28 (6): 121-124129 doi:10.14022/j.issn1674-6236.2020.06.026.