ISSN: 2664-9640

Big Data Processing and Analysis of Taxi Trips Based on Hadoop Technology

Zhisong Peng

School of transportation, Chongqing Jiaotong University, Chongqing 400000, China.

Abstract

In the application of the intelligent transportation industry, the traditional data processing methods have been unable to meet the large amount of data generated in the intelligent transportation, and it is necessary to use efficient technologies and methods to process and analyze the data. The Hadoop platform provides support for the processing and analysis of big data. Based on Hadoop, this paper takes a city's taxi data set as an example, and analyzes the operation characteristics of the city's taxis from the perspective of market share, all-day ride times and travel distance, and provides relevant departments to better understand the operation of taxis. in accordance with.

Keywords

Big data; Hadoop; urban transportation.

1. Introduction

With the rapid development of my country's transportation industry, the amount of data generated from it is becoming larger and more complex. If traditional data analysis methods are used to process data, it will not only waste a lot of time, but also do not have a comprehensive understanding of the data. It wastes the value of these data. In recent years, with the widespread application of computer network, cloud computing, and Internet of Vehicles technologies in the intelligent transportation industry, the massive data generated by urban traffic has shown the 4V characteristics of big data [1]. Therefore, finding out the technologies and methods to efficiently process big data to meet this rapidly developing environment has become an urgent problem to be solved.

Taxis are an important part of urban public transportation and play a very important role in residents' daily travel. The data generated by taxi driving is not only large, but also has a certain representativeness in urban characteristics. Qi et al. used a clustering algorithm to cluster passengers' pick-up and drop-off locations respectively, and excavated the daily activities of residents and the land use properties of urban plots [2]. Tang Yanli et al. studied the travel characteristics of urban taxis based on taxi GPS big data [3]. Based on the Hadoop big data processing platform, Feng Fan used the clustering algorithm to analyze the characteristics of the taxi big data from the passenger travel time and travel space [4]. Li Yongding took the Lanzhou taxi GPS data as the research object, excavated the travel hotspots and travel congestion sections of Lanzhou residents, and conducted a visualization study [5].

In the research of big data and urban transportation, Hu Mingwei and others designed a system of data collection, analysis and information release through Beijing's transportation data, which can provide users with transportation information services [6]. Yue Jianming analyzed its development advantages in the intelligent transportation industry from the perspective of big data [7]. IJ Lee proposed a distributed structure that can process and analyze traffic big data, and took seven years of highway collision data as an example to obtain the collision probability [8].

ISSN: 2664-9640

This paper first briefly introduces Hadoop, and then takes a city's taxi data set published on the Internet as an example, analyzes the data set through the Hadoop platform, and obtains the operation characteristics of the city's taxis and the travel characteristics of passenge. The two most important components in Hadoop are HDFS and MapReduce components.

2. Introduction and Construction of Hadoop Platform

When processing big data, the Hadoop platform is the most commonly used and the most widely used processing platform. It was developed by the Apache Foundation. Using this platform, people can develop distributed programs without understanding the underlying distributed architecture. [9]. HDFS is a distributed storage component that can store data in a distributed manner. It can access data in applications with high throughput and has the characteristics of high fault tolerance. MapReduce provides Map and Reduce processes that can compute large amounts of data. Hadoop has the following characteristics: 1. Reliability. Hadoop considers that the calculation process may fail, so multiple data copies are stored in the calculation, and when the node calculation fails, it will be recalculated. 2. Efficiency: Hadoop adopts distributed parallel computing, and data can be dynamically moved between each node, maintaining the balance of nodes, and making better use of its hardware resources, thereby improving computing speed. 3. Scalability, Hadoop has many components that users can install according to their own needs, and also provides a large number of API interfaces. However, Hadoop also has the disadvantages that it is not suitable for low-latency data access, cannot effectively store a large number of small files, and does not support multiple users to modify files at will.

The hardware used in this experiment is a desktop computer, the operating system is 64-bit Windows 10, the processor is i5-10400, and the memory is 8G. And this article completes the installation and configuration of VMware, Ubuntu, Hbas, Hadoop, MySQL, Sqoop, and Hive. The configuration steps of the experimental environment are shown in Figure 1 below.

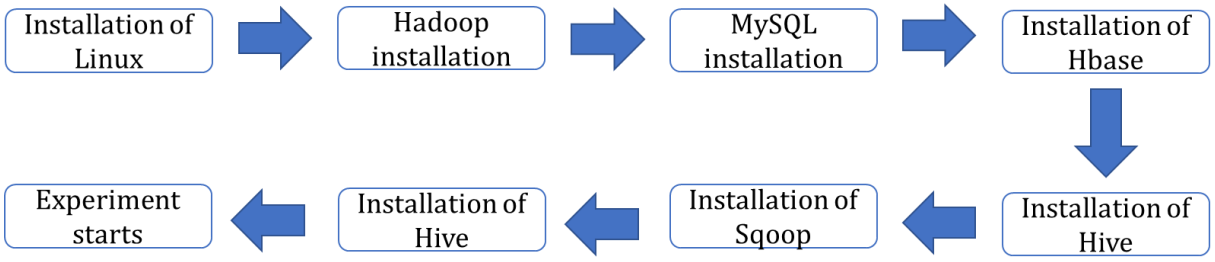

Figure 1: Software installation flow chart

3. Data

This dataset is the big data of taxi trips in a large city. The selected time is from January 1, 2020 to January 31, 2020, with about 6.4 million pieces of data. The data includes the following information: pick-up date/time, pick-up and drop-off location, trip distance, detailed fare, fare category, payment category, and the number of passengers reported by the driver. The table structure of the dataset and the meaning of the fields are shown in Table 1.

Table 1: Field descriptions of the dataset

Field Name	Type of Data	Field Name	Type of Data
VendorID	INT	payment_type	INT
tpep_pickup_datetime	STRING	fare_amount	FLOAT

ISSN: 2664-9640

tpep_dropoff_datetime	STRING	extra	FLOAT
passenger_count	INT	mta_tax	FLOAT
trip_distance	FLOAT	tip_amount	FLOAT
RatecodeID	FLOAT	tolls_amount	FLOAT
store_and_fwd_flag	STRING	improvement_surcharge	FLOAT
PULocationID	INT	total_amount	FLOAT
DOLocationID	INT	congestion_surcharge	FLOAT

4. Operational Characteristics Analysis

4.1. Market Share Analysis

Statistical analysis is made on the number of passengers picked up by the two operators in the data set in January 2020, and the market shares of the two operators and their changing trends are analyzed. According to Figure 2, the market share of operator 1 is much larger than that of operator 2, and the number of taxi orders in the city changes periodically. There are about 4 full cycles in a month.

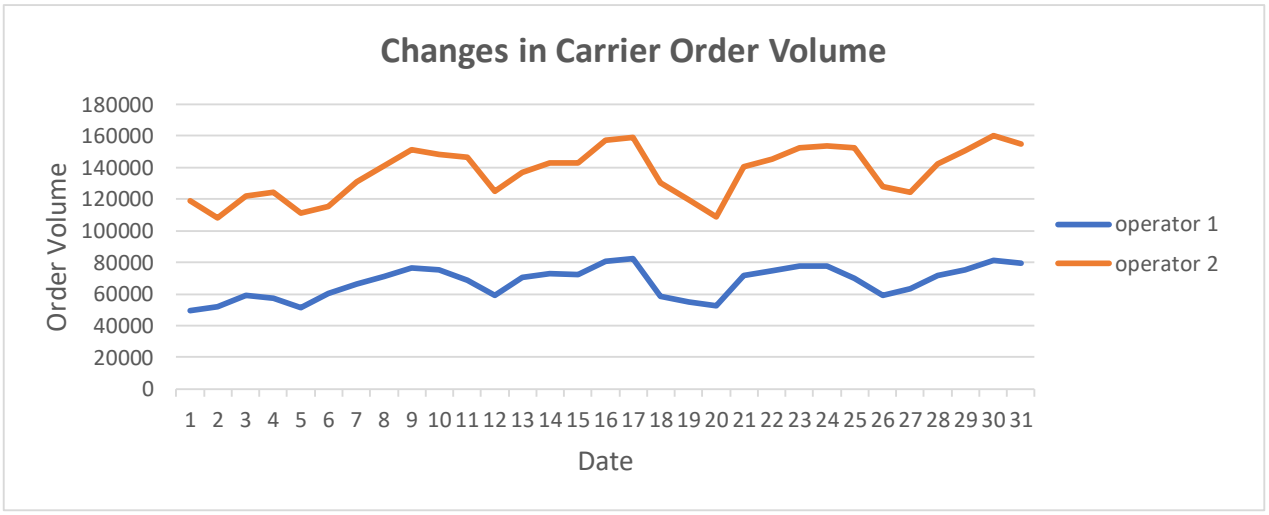

Figure 2 Changes in carrier order volume

4.2. Market Share Analysis

The frequency of passengers taking taxis is shown in Figure 5. It can be seen from the figure that when the travel distance is 0.5-1 mile, the number of passengers taking taxis exceeds 1.4 million times, accounting for about 22% of the total number of trips. , followed by trip distances of 1-1.5 miles, which accounted for about 21% of total trips. Then, as the travel distance continues to increase, the number of taxi rides by passenger vehicles decreases rapidly, because taxis are more expensive for long-distance travel.

ISSN: 2664-9640

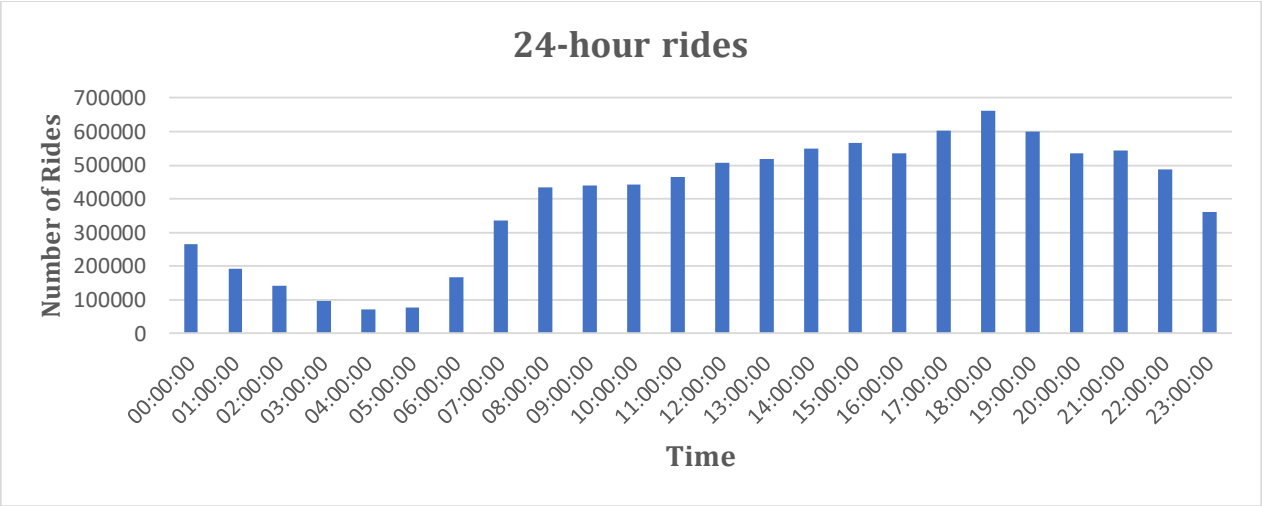

Figure 3 24-hour rides

4.3. travel distance analysis

The frequency statistics of passengers taking taxis are shown in Figure 4. It can be seen from the figure that passengers take the most taxi trips in the distance of 0.5-1 mile, and the number of trips exceeds 1.4 million, accounting for about 22% of the total trips. This is followed by trip distances of 1-1.5 miles, which account for about 21% of total trips. Immediately after the increase in the boarding distance, the number of passengers' rides decreases rapidly. Obviously, taking a taxi is not an economical choice for long-distance travel.

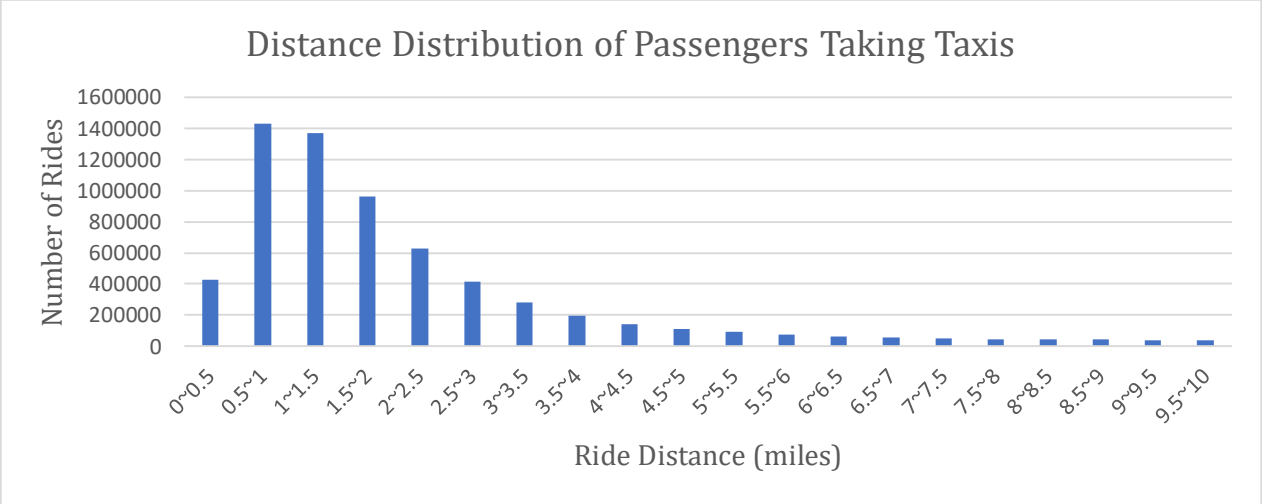

Figure 4 Frequency distribution map of passenger travel distance

5. Conclusion

In the application of the intelligent transportation industry, the traditional data processing methods have been unable to meet the large amount of data generated in the intelligent transportation, and it is necessary to use efficient technologies and methods to process and analyze the data. Based on the Hadoop platform, this paper takes the big data of a city's taxi trips for one month as an example, and analyzes the operation characteristics of the city's taxis from the perspectives of market share, full-day ride times and travel distance, and draws relevant conclusions. This is of great significance for taxi operators to better deploy taxis and for city-related departments to better grasp the travel rules of residents.

ISSN: 2664-9640

References

[1] Ye Liang: Transformation and development of my country's traffic data management application under the background of "big data",Transportation and Transportation (Academic Edition), Vol. 02 (2013),p.65-68.

[2] Z, S, Zhang, et al: Land-Use Classification Using Taxi GPS Traces, IEEE Transactions on Intelligent Transportation Systems, Vol. 14 (2013), p.113-123.

[3] Tang Yanli, Jiang Chao, Zheng Bohong, Li Qianming: Research on passenger travel characteristics of urban taxis based on multi-source data fusion: Yueyang City as an example, Transportation System Engineering and Information, Vol. 02 (2018), p. 45-51.

[4] Feng Fan: *Taxi GPS data analysis system based on Hadoop technology* (MS., Xidian University, China 2016) , p.15.

[5] Li Yongding: *Research on information mining and visualization based on taxi GPS data* (MS., Lanzhou Jiaotong University, China 2021) , p.29.

[6] Hu Mingwei, Miao Lixin, Wang Yunfei: Design Beijing Traffic Flow Data Collection, Processing/ Analysis and Information Release System, Highway Traffic Science and Technology, Vol. 02 (2003), p.77-80.

[7] Yue Jianming, Yuan Lunqu: Big data analysis in the development of intelligent transportation, Productivity Research, Vol. 06 (2013), p.137-138+165.

[8] IJ Lee: Big data processing framework of road traffic collision using distributed CEP, *Asia-Pacific Network Operations and Management Symposium* (Taiwan, September 17-19, 2014), p.1-4.

[9] Batty, Michael: Smart cities, big data, Environment & Planning B Planning & Design, Vol. 39 (2012), p. 191-193.