Design of a Hadoop Based Data Platform for Auto Aftermarket

Yi Shen

Automotive Engineering institute, Jiangxi University of Technology, Nanchang 330098, China

Keywords: Auto aftermarket; Hadoop; Data platform; Data analysis

Abstract. With the development of China's auto aftermarket, the share of auto aftermarket is going to increase in the future. Due to the unique characterizations of the auto aftermarket, dealers cannot provide adequate choices to customers due limitations in the venue or funds. Additionally, dealers cannot recommend suitable products to customers because the approach to determine their demands is relatively simplified. Moreover, it is complex to maintain the vehicles; therefore, a car owner is not able to make an appropriate maintenance plan and judge the quality of auto spare parts. The three problems bring high security risks. Through in-depth analysis of the status and prospects of the auto aftermarket, this study proposes and implements a Hadoop-based data platform using techniques of big data. The mentioned risks are reduced by the usage of cutting-edge technology for distributed systems and big data analysis techniques. The method can also provide a customized interface, which prepares for ongoing maintenance and the expansion of the platform.

Introduction

Auto aftermarket is a general term of a series of transactions related to usage and service of vehicles after they are sold. It includes maintenance, auto spare parts and the corresponding financial and insurance service system.

According to the National Bureau of Statistics of the People's Republic of China, until the end of 2011, the number of civil vehicles reached 93.5632 million. The number of private cars reached 73.2679 million. Professionals predicted that in the development rate of current sales, the car parc would exceed 200 million by 2016. The high car parc is fundamental to development of the auto aftermarket. However, the current auto aftermarket is not well developed. In the People's Republic of China, most cars are sold by 4S shops. They have many problems in the auto aftermarket:

Firstly, the auto aftermarket does not attract enough interest. Many 4S shops sell auto aftermarket products, such as the maintenance and auto supplies, with the vehicle sales, in order to complete sales tasks assigned by vehicle manufacturers. Additionally, the 4S shops increase the price of the products. This approach reduces the credibility of the prices, and it even affects the reputation of the 4S shops. As a result, the profits come from the sale of a whole vehicle sale, which decreases the importance of the 4S shops. They are easily manipulated by the vehicle manufacturers. This is also why the monopoly of luxury cars can be formed in China.

Secondly, the 4S shops cannot properly analyses customers and their products are not accurately targeted. The sales staffs are intermingled and they do not have methods to segment customers. The selection of products is inappropriate and they do not know how to reasonably determine product prices. Some products are overpriced and some of the products cannot be sold out. Consequently, they do not only occupy the funds of enterprises, but also waste the limited display space in the 4S shops.

Fortunately, some companies have noticed these problems and they build e-commerce platforms, such as Taobao, JD.com and Joyo. They have individually started online sales of auto spare parts. But these e-commerce platforms are not customized for the auto aftermarket. They simply copy the e-commerce business models in other sectors and apply in the auto aftermarket.

At present, no one platform can really solve the mentioned problems in the auto aftermarket. Building a novel platform will produce a promising market and significantly contribute the auto aftermarket.

This study is aiming at solving the problems and building a customized and flexible business platform. It mainly includes four layers, i.e. data service, business service, API service and end user.

Hadoop Methodology

Cloud computing is a convenient, easy to use and on-demand service mode. By dynamically configuring computing resources, cloud services providers can provide users with customized servers, networks, storage, applications and data services. Users can enjoy quick and convenient services with heavy management in conditions of good network. Cloud computing is an integrated product of the traditional computer technologies (parallel computing, multicore computing, distributed computing, grid computing, virtualization, network storage, load balancing, etc.) and Internet technologies. Cloud services providers serve end users with powerful computing based on Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS).

The Hadoop framework is categorized open source cloud computing. This project was created by Doug Cutting and its initial and main purpose was to provide a distributed function for the Nutch search engine. Later, the Hadoop became an independent project. The core components of Hadoop are HDFS and MapReduce. HDFS provides large data storage capacity and MapReduce provides a development and runtime environment for programmers to handle large data. Hadoop is a batch system and it does not guarantee real-time response to requests. Therefore, Hadoop is neither a relational database system nor a structured data storage system.

Because of the advantages in terms of storage and handling of large amounts of data, Hadoop has been increasingly favoured by many well-known companies, such as Yahoo, LinkedIn, Twitter, Facebook, Rackspace, eBay, Amazon, WalMart, Baidu, and Taobao. They have respectively deployed Hadoop clusters to achieve functions, such as recommendation system, financial analysis, data mining, market forecasts and log analysis, etc.

Design of Data Platform

The designed system mainly includes the server and the display system. The server is based on the Linux system and the display system to show part of the specific business uses the WEB architecture. The design and implementation are described in detail in the following.

Distributed storage system is shown in Fig. 1. In this figure, a product item is a document object in the database of MongoDB clusters. The information of clusters is stored in the HDFS distributed file system of Hadoop.

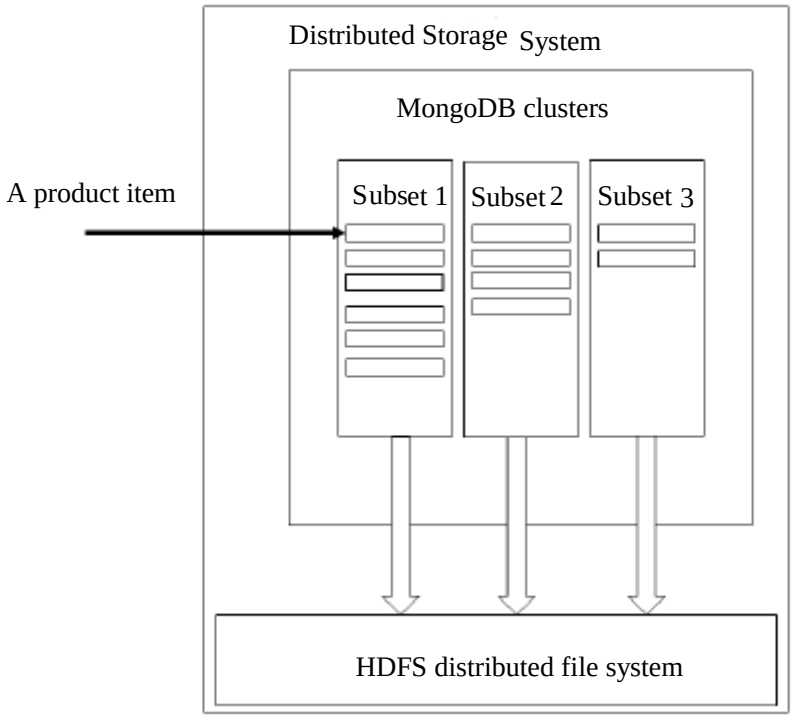

Fig. 1 Framework of Distributed Storage System

Log management system collects the original WEB access logs from various end users. Considering that the amount of data in the log file increases rapidly, this log management system needs to periodically roll logs. Because WEB server runtime log files are in an open state and they are continuously written. When the server is running, the logs cannot be rolled. Additionally, the server must be restarted to open a new log file after the system log files moved or deleted. Therefore, the sever must be rebooted gracefully to ensure that the original log information are written in the new log file. The time interval is critical so that the server finishes its tasks, and writes the data in the original log file. In this paper, the rolling time is 13:00 at 00:00 every day, taking into account the amount of user access and primary access.

Customized traffic system is mainly responsible for realization of business functions, business logic and user permissions management. According to the corresponding API service, this system automatically determines the selected business and ultimately to realize the customized business by preprocessing the message and running the service components. The following introduces the logics of some business:

Fig. 2 shows the flow chart of the commodity Management (add, delete, and change):

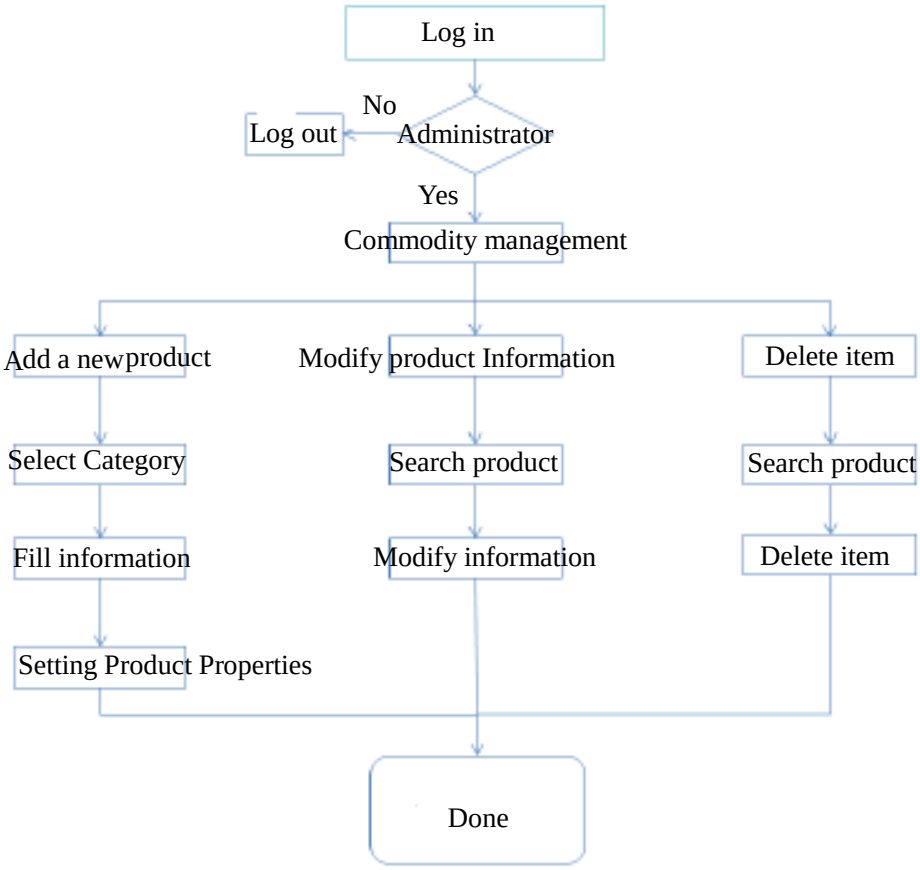

Fig. 2 Flow chart of the commodity management.

Product data platform data analysis techniques

In order to provide results of data analysis, obtain an intelligent network systems of commodity relations, dig out the potential value of the entire platform data generated during operation, improve

product recommendation accuracy and success rates, this design uses a lot of data mining technologies, comprising: a user key attribute analysis, client object analysis, life cycle mining, trade association rule mining technology, etc.

According to the uniqueness of auto aftermarket, the car owner is the ultimate consumer of the whole industry chain. Only after obtaining the key owner attributes, it is possible to analyze key attributes and then develop appropriate product recommendation strategies.

Client object analysis is essential in the realization of intelligent recommendation by the commodity relations network. Analyzing the client object is useful in customer segmentation, analysis of customer changes and their responses to products and services. Thereby, customer satisfaction, loyalty and profit contribution can be accurately calculated and predicted. The merchandise operations strategies can be optimized and eventually increases customer satisfaction and viscosity.

Firstly, an attribute (such as transaction volume, price, cost and associated benefits) is making a split after being tested. Their group is called the correct segmentation. The selected are the attributes that makes the minimum normalized distance in any pair of the correct distance separation. The formulas to split P_1 and P_2 are shown below: the second attribute (such as transaction volume) is tested. The remaining information in the first attribute (price) is subsequently tested. Then add the test of the first attribute and the test of the remaining information of the second attribute. The equations are given as:

$$D(P_1, P_2) = I\left(\frac{p_1}{p_2}\right) + I\left(\frac{p_2}{p_1}\right)$$

where $D(P_1, P_2)$ is the distance between P_1 and P_2; $I\left(\frac{p_1}{p_2}\right)$ is the remaining information in the first attribute after the test of the second attribute. $I\left(\frac{p_2}{p_1}\right)$ is the remaining information in the second attribute after the test of the first attribute. The normalized distance between P_1 and P_2 is the quotient of the distance and the information contained in the intersection of P_1 and P_2.

$$Dn(P_1, P_2) = \frac{D(p_1, p_2)}{I(p_1 \cap p_2)}$$

After formula substitution, the normalized distance is converted into a function of gain information shown as the above equation. In this design, the shortest distance for the proper normalization segmentation is converted to the largest value of the right side of the equation.

$$Dn(Pc, Pv) = 1 - \frac{Gain(Ak, x)}{I(pv \cap pc)} \in [0,1]$$

where $Dn(Pc, Pv)$ is normalized distance between the correct segmentations Pc and Pv. $Gain(Ak, x)$ is the normalized attribute Ak. $I(pv \cap pc)$ is the intersection of the correct segmentations Pc and Pv. Therefore, the attribute selection is transformed to the shortest normalized distance.

Using this distance-based attribute selection method, a superior decision tree can obtained to meet requirements of this design. The decision tree can contribute the life analysis of customers and provide data for market operations strategy.

Conclusions

With the development of e-commerce technologies and auto aftermarket, there will be more abundant auto aftermarket products. Due to limitations in the venue or funds, more dealers will use online and offline operating mode (O2O). The platform designed in this paper uses the large data processing technologies and provides effective and accurate community data. This platform not only increases profits of the manufacturers and dealers, but also ensures the interests of the car owners. The following conclusions can be drawn:

(1) The key questions must be clearly understood. There must be a clear understating in preprocessing the data. Different scenarios require different data formats and different algorithms. In this paper, a clear review in the background and needs play an important role in the success of the design of the platform.

(2) The intelligent deployment of Hadoop distributed clusters is very important. In this design, the number of the final clusters is far more than the number of the test clusters. If the intelligent deployment script was not well written at the start, there would be a lot of work and easily lead to hidden errors.

(3) The big data analysis techniques give more accurate results on large datasets and the algorithms are more complex. Therefore, the amount of data plays an important role in the accuracy and reliability of the platform.

(4) MapReduce of Hadoop provides a convenient programming interface. In this study, the MapReduce programming model saves a lot work in interface development and ensures the accuracy and standardization.

Acknowledgements

This work was financially supported by the key subject building project (vehicle engineering) of Jiangxi University of Technology.

References

[1] Wei Gao. Profitability problems of 4S Shops to be solved [N]. China Times, 2012.06.11 (030 version)

[2] Zhihao Chen. Study on new e-commerce platforms for auto aftermarket [D]. Shanghai: East China University of Technology, 2011

[3] Duane Miller. Driving service business during Car Care Month [J]. AutomotiveNews. Detroit, 2005, Vol.79 (6142): 2

[4] Kenneth Cukier, Viktor Mayer-Schönberger. Big Data: A Revolution That Will Transform How We Live, Work, and Think[M]. Yangyan Sheng, Zhou Tao. Hangzhou: Zhejiang People's Publishing House, 2013: 02-79

[5] Nakada H. , Ogawa H. , Kudoh T. , Stream processing with BigData: SSS-MapReduce [J]. Cloud Computing Technology and Science, 2012, 618-621

[6] Hadoop Quick Start [EB / OL]. [2013.09.12] http://hadoop.apache.org

[7] Zhu Zhu. Study and application of big data processing model based on Hadoop[D]. Beijing: Beijing University of Posts and Telecommunications, 2008

[8] Tom White. Hadoop: The Definitive Guide[M]. Minqi Zhou, Xiaoling Wang. Beijing: Tsinghua University Press, 2011: 41-77

[9] Wen Jiang. Data analysis and application of platforms based on Hadoop[D]. Beijing: Beijing University of Posts and Telecommunications, 2011

[10] Xicheng Dong. Inside Hadoop[M]. Beijing: China Machine Press, 2013: 25-73