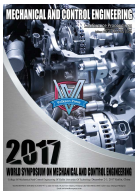
ISBN: 978-1-948012-06-5

Contents List available at VOLKSON PRESS

World Symposium on Mechanical and Control Engineering (WSMCE)



VEHICLE DATA PROCESSING AND ANALYSIS PLATFORM BASED ON SPARK

Xiaolan Xie[1]*, Tianwei Yuan[2,3], Xiao Zhou[4]

[1]College of Information Science and Engineering, Guilin University of Technology, Guilin, Guangxi Zhuang Autonomous Region, China
[2]Guangxi Universities Key Laboratory of Embedded Technology and Intelligent Information Processing Guilin University of Technology), China
[3]College of Information Science and Engineering, Guilin University of Technology, Guilin, Guangxi Zhuang Autonomous Region, China
[4]College of Mechanical and Control Engineering, Guilin University of Technology, Guilin, Guangxi Zhuang Autonomous Region, China
**Corresponding Author Email: 1191948476@qq.com*

ARTICLE DETAILS

Article History:

Received 02 october 2017
Accepted 06 october 2017
Available online 11 november 2017

Keywords

Vehicle, Spark, Hbase, Platform

ABSTRACT

Based on the Spark big data analysis platform, using Yarn management resource scheduling problem, using Hbase as the storage mode of distributed data. Through the mass of car data, dig out the key factors that affect the safety of car performance, and show in the form of visualization. An effective scheme is put forward for the maintenance and fault detection of the user's vehicle. The experimental results show that the analysis method based on Spark platform can quickly, effectively and accurately analyze the key information and play a guiding and analytical role for users.

1. Introduction

In the process of maintenance of the vehicle, a large amount of data will be generated, but the data cannot collate statistics to guide the future situation. And the maintenance of complex and large amount of data characteristics, the traditional data processing methods have been unable to meet the needs of the development of the vehicle industry [1-4]. With the development of cloud computing and big data, it is more advanced and reliable, which provides an effective solution for the vehicle industry. Using cloud computing and data mining technology, the scattered and unstructured data generated in the vehicle market is stored and excavated, and the specific analysis and calculation are carried out at any time [5]. Through cloud computing and big data technology, you can analyze and predict the data generated the vehicle market, will make the enterprise decision-making more accurate, release more hidden value after the vehicle market.

2. DATA HANDLING PLATFORM

The data processing platform is built with spark+yarn+hbase mode, which can deal with a large number of vehicle data effectively and quickly.

2.1 Spark

Apache Spark is a big data processing framework. It is fast, easy to use and can complex analysis. Spark allows program developers to use DAG to develop complex multi-step data pipelines. It also supports memory data sharing, so that different jobs can process the same data [6]. Spark runs on the existing Hadoop distributed file system (HDFS) to provide additional enhancements. It supports the Hadoop V2 YARN cluster.

2.2 Yarn

YARN is a resource management system in Hadoop 2, and YARN is still a Master/Slave structure in general. In the framework of resource management, Resource Manager as Master, Node Manager as Slave, and Resource Manager is responsible for the unified management and scheduling of resources on each Node Manager [7]. The advantage, you run many kinds of frameworks over YARN, and it can manage and distribute the resource of frameworks. Make them share a cluster, which can greatly reduce operation and maintenance costs and hardware costs.

2.3 Hbase

HBase is a built on HDFS distributed column storage system development based on Google Big Table model, a typical key/value system. It is characterized by a large scale, a table can have tens of billions of rows, millions of columns, each row has a sorted primary key and any number of columns, columns can be dynamically increased according to the needs of the same table in different rows can have a different column. It has the advantages of large data storage, large amount of data and high concurrency operation, and it is very simple to read and write data random read and write operations. It has good fault tolerance and scalability, and can be extended to hundreds of nodes.

3. SYSTEM DESIGN

3.1 Data preprocessing and analysis

Due to uncontrollable factors that may occur, data missing, data errors. Need to pretreatment of the input information to ensure that the input information standards, reliable, convenient for post-processing and analysis to ensure the accuracy of the results.

Example of vehicle data:

Table 1: Vehicle Data

Number	Vehicle	King Pin Inclination	Camber	Toe-in	caster	problem	mileage
1	Car1	10	15	6	7	1011	50km

1 represents problems, 0 represents no problems

3.2 Spark Cloud Platform

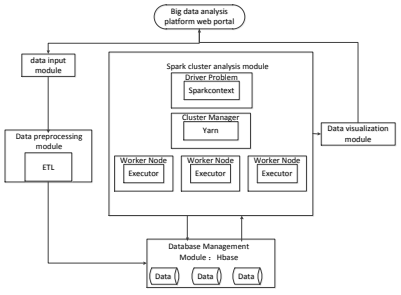


Figure 1: Spark Cloud Platform

3.3 Methodology

Algorithm1: Calculate the number of deviation of the parameters of the same model

Table 2: Algorithm1

Algorithm1
Input fault vehicle detection data
Use filter() to filter data with vehicle problem 1
Use map() to set key= vehicle model, value= vehicle problem
Use reducebykey() to increase the number of vehicles appearing on the same road condition
Loops Step2, 3, 4.
Output the number of vehicle problems according to different types of vehicles

Algorithm2: The influence of driving distance on vehicle problem based on spark

Table 2: Algorithm2

Algorithm2
Input fault vehicle detection data
Use map() to set key= distance, value= vehicle problem
Use sortByKey () to sort by travel distance
Use distinct() to remove duplicate data from vehicles
Output data using map(), map set key= vehicle problem value=1
Use reducebykey() to increase the number of occurrences
Output the Influence of driving distance on different vehicle problems

4. PERFORMANCE EVALUATION

4.1 Analysis of vehicle parameters affected by vehicle type

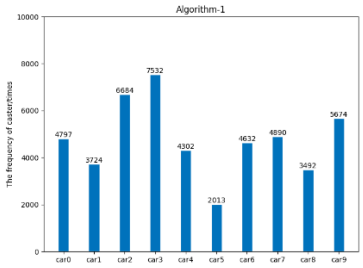


Figure 3: Analysis of Vehicle Parameters

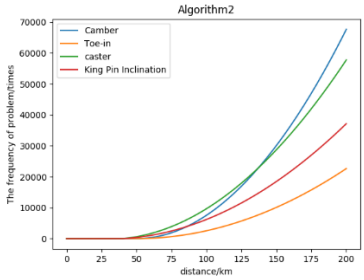


Figure 4: Analysis of Distance

The figure is the statistics of the 100,000 data, respectively, calculated the number of different models appear the number of issues. The use of the platform 100,000 data calculation time of 1 minute, the calculation results as shown. Through this method, a large number of possible shortcomings of different brands can be quickly counted.

5. CONCLUSION

This paper puts forward the analysis of mass vehicle data based on Spark platform. Under this platform, it can quickly and reliably analyze and process the vehicle data according to the existing data in Hbase, and get the expected results quickly. Has a certain degree of scalability, can dig out more effective data and information, and plays a guiding role for vehicle maintenance in the future.

ACKNOWLEDGEMENTS

This research work was supported by the National Natural Science Foundation of China (Grant No.61762031), Guangxi Key Research and Development Plan, GuangXi key Laboratory of Embedded Technology and Intelligent Information Processing.

REFERENCES

[1] Zaharia, M., Chowdhury, M., Franklin, M.J., Shenker, S., Stoica, I. 2010. Spark: cluster computing with working sets. Usenix Conference on Hot Topics in Cloud Computing USENIX Association,10-10.

[2] Vavilapalli, V.K., Murthy, A.C., Douglas, C., Agarwal, S., Konar, M., Evans, R., Graves, T., Lowe, J., Shah, H., Seth, S., Saha, B., Curino, C., O'Malley, O., Radia, S., Reed, B., Baldeschwieler, E. 2013. Apache Hadoop YARN: yet another resource negotiator. Symposium on Cloud Computing (pp.5). ACM.

[3] Chowdhury, M., Zaharia, M., Ma, J., Jordan, M.I., Stoica, I. 2011. Managing data transfers in computer clusters with orchestra. Acm Sigcomm Conference ACM, 98-109.

[4] Lee, B., Riche, N.H., Karison, A.K., Carpendale, S. 2010. SparkClouds: visualizing trends in tag clouds. IEEE Transactions on Visualization and Computer Graphics 16 (6), 1182.

[5] Ewen, S., Tzoumas, K., Kaufmann, M., Markl, V. 2012. Spinning Fast Iterative Data Flows. Proceedings of the Vldb Endowment, 5 (11), 1268-1279.

[6] Wirtz, G.P., Brown, S.D., Kriven, W.M. 2016. Ceramic coatings by anodic spark deposition. Materials and Manufacturing Processes, 6 (1), 87-115.

[7] Lu, R., Lin, X., Zhu, H., Shen, X. 2009. Spark: A New Vanet-Based Smart Parking Scheme for Large Parking Lots. INFOCOM (pp.1413-1421). IEEE.