Hadoop을 이용한 스마트 자동차 서비스용 빅 데이터 솔루션 개발

라이오넬¹ · 장종욱²*

Addressing Big Data solution enabled Connected Vehicle services using Hadoop

Lionel Nkenyereye¹ · Jong-Wook Jang²*

¹Department of Computer Engineering, Dong-Eui University, Busan 614-714, Korea
²*Department of Computer Engineering, Dong-Eui University, Busan 614-714, Korea

요 약

자동차 진단 데이터의 양이 증가함에 따라 자동차 에코시스템의 액터는 스마트 자동차에서 수집된 데이터에 따라 새로운 서비스를 시뮬레이션 하거나 설계하기 위하여 실시간으로 분석을 해야 하는 어려움에 직면하게 된다. 본 논문에서는 자동차에서 생성된 막대한 양의 자동차 내장 진단 데이터를 처리하고 분석하는데 필수적이고 심오한 해석을 제시하는 빅 데이터 솔루션에 관한 연구를 하였다. Hadoop 및 그 에코시스템은 자동차 소유자에 대한 새로운 서비스 제공을 위해 자동차 에코시스템의 액터에 의해 사용될 수 있는 막대한 데이터 및 전달된 유용한 결과를 처리하기 위해 개발된 것이다. 지능형 교통시스템이 안전성 보장, 속도로 인한 사고로 인하는 상해 및 충돌의 비율 감소 등에 관여함에 따라, 자동차 진단 데이터 기반의 빅 데이터 솔루션 개발을 통해 향후 실시간 결과 감시, 여러 스마트 자동차에서의 데이터 수집, 수집된 데이터에 대한 신뢰성 있는 처리 및 용이한 저장을 실현화하게 된다.

ABSTRACT

As the amount of vehicle’s diagnostics data increases, the actors in automotive ecosystem will encounter difficulties to perform a real time analysis in order to simulate or to design new services according to the data gathered from the connected cars. In this paper, we have conducted a study of a Big Data solution that expresses the essential deep analytics to process and analyze vast quantities of vehicles on board diagnostics data generated by cars. Hadoop and its ecosystems have been deployed to process a large data and delivered useful outcomes that may be used by actors in automotive ecosystem to deliver new services to car owners. As the Intelligent transport system is involved to guarantee safety, reduce rate of crash and injured in the accident due to speed, addressing big data solution based on vehicle diagnostics data is upcoming to monitor real time outcome from it and making collection of data from several connected cars, facilitating reliable processing and easier storage of data collected.

키워드: 커넥티드 카, 하둡, 하이브 오픈소스, 빅데이터, 진단서비스

Key word: Connected car, Hadoop project, HIVE open source, Big data Problem, on-board diagnostics, MapReduce
I. INTRODUCTION

The information technology authorizes information sharing from vehicles, roads infrastructure’s sensors, vehicle to vehicle, vehicle-to infrastructure, road condition [1]. Based on data received from ITS (Intelligent Transportation System), the third party interested in automobile ecosystem such as car manufacturers, repair shops, road and transportation authorities will continuously support a multitude of applications as for instance, monitoring performance of vehicles sold in the market by leveraging reporting of DTC (Diagnostic Troubles codes), status diagnosis information, emergency services management to locate and help victims of accidents or injured persons, traffic transportation authorities to increase safety, accident prevention, car repairs to analyze vehicle diagnosis in real time in case the vehicle breaks down.

However, data from connected cars will serve as a source to the actors in automotive ecosystem to afford value added services to the car owners. It seems that as long as data from connected vehicles increases, a big solution is expected to be implemented by car manufacturers or transportation authorities in order to process it and make available a reliable database of outcomes to all interested in connected car field. Apache Hadoop is nowadays suitable to process a huge amount of data. In this paper, we present a study of big data solution using Hadoop to process connected vehicle’s diagnostics data and the results from processing will allow connected vehicle services.

II. RELATED STUDIES

2.1. Internet of vehicle and geographical navigation application program from Google

The evolution of IoT (Internet of Things) has brought one line many little nodes plainly exists today [2]. Google has considered the vehicle as one of the big node comprise by a lot of little node. Google’s Waze (the world’s largest community-based traffic and navigation application) division gathers traffic information from all enabled Waze geographical application based on smartphone out and distribute that information to all car’s owner who has opted to use Google’s Waze.

Like other GPS (Global Positioning System) software it considers from users’ driving times to provide routing and real-time traffic updates. People can report accidents, traffic jams, speed and police traps, and from the online map editor, they can update roads, landmarks, house numbers, etc. [3].

The Google’s Waze GPS application involves only car owners or drivers and requires to upload their current reports accidents in order to allow them to make their optimal navigation routing decisions. This appear to be a weakness of the connected car vehicle platform that enables sharing vehicle sensors data and allows others services providers interest to automotive industry to design and propose their respectively services to the car owners. The proposed solution in this paper highlights acquisition of vehicles on board diagnostics data, real time processing using Apache Hadoop. The outcomes from processing are stored back to the relational database or generated in CVS (Comma Separated Values) format, so actors in automotive ecosystem can access them via internet and provide new services to the car owners.

2.2. Apache Hadoop platform

Hadoop is an open source framework for writing and running distributed applications that process large amounts of data [4]. Data is broken up and distributed across the cluster and as much as possible, computation on a piece of data taken place on the same machine where the piece of data resides.

The Hadoop projects that are used in this paper to keeping processing vehicle sensor data are: HDFS (Hadoop Distributed File System) a distributed file system that runs on large clusters of commodity clusters, Apache Hive a distributed data warehouse that manages data stored in and make possible for analysts with strong SQL (Structured Query Language) skills.
2.3. Big data context

The big data refers to a collection of data sets that relational databases have are not able to store for example a few terabyte of unstructured data from a variety of sources can considered as big data [4].

Effectively analysis of data is considered as a new way to increase productivity and also accelerate new services which may be proposed to car owners. The evolution of IoT (Internet of Things) has brought one line many little nodes.

III. SYSTEM DESIGN

3.1. Design of the solution to handle vehicle diagnostic data

The solution proposed is related to use Hadoop framework. After vehicle diagnostic software based on android performs an upload of vehicle diagnostics data, to the data centers, Hadoop framework will process them and the outcomes will be stored on the web server on which third party can access them. Fig. 1 shows the design of solution proposed based on Hadoop.

3.2. Acquisition of vehicle diagnostics data

The acquisition of vehicle diagnostics data consists of OBD (On-Board Diagnostics) scan tool, an android smartphone and a database (MySQL database) known to us as a RDBMS (Relational Database Management System). OBD scan tool interacts with the ECU (Engine Control Unit), while the android device is used as the client to handle data acquisition and data transmission, the database stores vehicles diagnostics data.

The HTTP (Hypertext Transfer Protocol) in this paper is used to complete the communication between android and web server. The auto scanner tool used to connect with OBD interface starts reading OBD-PIDs (On-board Diagnostics Parameters IDs) codes [5]. The profile of the driver as for instance car owner name, car owner password, and valid vehicle identification based on Bluetooth MAC (Machine Access Address) of the OBD scan tool are authenticated on the database, and then sent back to the car owner. The failure of authentication leads to a new car owner registration. Fig. 2 below summarizes the design of the acquisition of vehicle diagnostics data to the remote data centers.

3.3. Processing of vehicle diagnostics data using Hadoop framework

The processing of the data acquired on the remote database is divided in four phases: import data from MySQL to Hadoop clusters, loading data from HDFS to Apache Hive, analysis through Hadoop MapReduce framework, upload outcome files in CVS format from HDFS to the web server.
1) Import data from MySQL to Hadoop data file

In this phase, the process consists of importing data from MySQL into Sqoop. Sqoop is an open-source tool that allows users to extract data from a relational database into Hadoop for further processing[6].

2) Loading data from HDFS to Hive

After Apache Sqoop keeping parallelizing import across multiple mappers, the next step consists of loading data into a Hive table using Apache Sqoop. The import data into hive relies on sake of efficiency that has a post processing step where Hive table is created and loaded. When the data is loaded into HIVE from HDFS directory, Hive moves the sqoop replication table which is viewed as a directory into its warehouse rather than copying data.

3) Analysis using Map Reduce framework

Hadoop acts as a distributed computing system that holds a distributed file system across Hadoop cluster. It relies to a distributed computational, MapReduce, which coordinates and runs jobs in the cluster to operate on a part of the overall processing task in parallel. Some of this analysis needs to be performed by using customs map and reduce scripts. In conjunction with HiveQL. (HiveQL is Hive’s query language, a dialect of SQL) It is heavily influenced by MySQL[6], the developers can write SQL scripts to perform historical analysis over data loading in Apache Hive warehouse. SQL is great for many analyses, and it has the huge advantage of being very well known in the industry, so Hive is well placed to integrate with business intelligence tools (ODBC for example).

4) Save the outcome to the web server

In this paper, when the final result from processing is available, Apache Hive is able to copy the outcome to the web server, thereafter Sqoop can exports it back to the MySQL database.

IV. IMPLEMENTATION AND ITS RESULTS

The purpose of the big data solution is to process vehicle diagnostics data, save the outcomes into a MySQL database and on the web server on which data is remained available for the third party (actors in automotive ecosystem) for their further needs. Therefore, we set up a plan of required useful outcome information. For each useful information, we develop and execute a MapReduce jobs on single node. Fig. 3 describes in summary the result expected for each job according to the vehicle diagnostics data and on which event the outcomes may be used to fulfill car owners’ needs or services.
Hadoop을 이용한 스마트 자동차 서비스용 빅 데이터 솔루션 개발

4.2. Processing using Hadoop platform

Apache Hadoop open source and its projects are used to process the on-board diagnostics data and make available useful information to the third party as described in (see the Fig. 3). When on-board diagnostics data are uploaded to the database, Apache Sqoop performs a replication import of data required to run Map Reduce jobs. Before the import can be initiated, Sqoop uses JDBC(Java Database Connectivity) to examine the table to import [6]. Apache Hive open source has an important role especially for storing vehicle diagnostics data unto a relational database. Sqoop generates a Hive table based on it originally definition known as metadata and at the same time stores data on HDFS. Data are imported into Apache Hive data warehouse and stored on HDFS as well.

One of the most key of Map Reduce jobs implementation is to take the vehicle on-board diagnostics data replicated from MySQL database and stored into Hive warehouse; process them according to the useful information we design and then store it to the MySQL database or generate it into CSV, JSON(JSON,JavaScript Object Notation) formats that can be readable by users. HiveQL is used to execute Map Reduce jobs. Fig. 6 shows our UDF(User-Defined Functions) pseudocode using HiveQL query. In this paper, we consider that the value of the RPM(Rate of Revolution of a Motor) for example helps to compute vehicle’s movement by taking in consideration current data and value of RPM after every 3 minutes. Fig. 7 and Fig. 8 show how Hadoop converts HiveQL queries into a set of Map Reduce jobs.

그림 4. OBD-PID를 요청하기 전 클라우드에서 운전자 프로필 검증화면
Fig. 4 Validation car user profile on cloud before requesting OBD-PIDs

그림 5. 운전자가 클라우드에 저장된 on-board 진단 데이터 요청 화면
Fig. 5 Request of on-board diagnostics data saved on the cloud by the car owner

그림 6. HiveQL 쿼리 사용한 캐릭터스 기능 정의의 구조
Fig. 6 Map Reduce function definition structure using HiveQL query

그림 7. 캐릭터스 작업 집합으로 하둡 converted에 HiveQL 쿼리 실행
Fig. 7 HiveQL query runned on Hadoop converted into a set of Map Reduce jobs
V. CONCLUSION AND FUTURE WORK

With the development of Hadoop platform project, now is possible to build big data solution using open source projects integrated with Hadoop. In this paper, a study of big data solution for processing data from vehicles using Hadoop and making final results available, allows accessing useful information via web services to the third party such as car manufacturer, transportation and road operators, car dealers, police, emergency services has been conducted on a single node cluster. The outcome obtained from various Map Reduce functions managed after executing HiveQL query indicate favorable results in term of time taken.

Future work will center on implementing and evaluating performance on cloud platforms for instance EC2(Amazon Elastic compute Cloud).

ACKNOWLEDGMENT

This work was supported by Dong-eui University Grant.(No.2014AA324), and Nurimaru R&D(No. 2014XC005).

REFERENCES


