DOI 10.29042/2018-3966-3970

Helix Vol. 8(5): 3966- 3970

Data Analysis of Driving Events on Processing Sensor Data through Hadoop Ecosystem

[*1]Pooja M. Gupta, [2]Shrikant D. Zade
[1]M.Tech Scholar, [2] Associate Prof. (CSE), PIET, Nagpur
Email: 910poojagupta@gmail.com, cdzshrikant@gmail.com

Received: 09th July 2018, Accepted: 14th August 2018, Published: 31st August 2018

Abstract
Moreover, Insurance Company decided the premium according to the risk of accidental ratio, average speed, driving patterns of owner of a vehicle. Minimum speed of the vehicle should be 30-50, onwards it considered as more chances of a accidental stage. In past papers the hardware were used to detect the speed and driving patterns of a vehicle by GSM/GPS modem. SD card consist of each data record within a certain limitation. But there is no certain analysis on this input files generated from the hardware used. Hence in this paper the main task is to process input data file in hadoop ecosystem for analysis. This will speed up the process of tracking the vehicle, mapping and calculating the mileage and various operations of vehicle. This application consume large amount of data and this data should not be loss and it should be manageable with historical and current data for this hadoop is the best feature for better analysis and fault tolerant. Other than this we can process number of applications for processing through hadoop ecosystem. But here we have taken vehicular data which is used to analyze the each drivers of vehicle in the form of speed, date and time. It is also used to derive driving patterns of the vehicle and each driver and also help to manage the speed of a vehicle and to analyze the accidental ratio.

Keywords: Hadoop Ecosystem, SD Card, GPS/GSM Data File.

Introduction

Big Data

Big Data is used for collections of extremely large dataset which increases day by day which may be analysed computationally to find the patterns, different association-relationship, trends mostly related to human interactions and behaviour. It is difficult to process Big Data using traditional data management systems[1].Therefore, Hadoop is a framework introduced by Apache Software Foundation to solve Big Data management and processing challenges.

1. Year over year, the large number of client and variety of devices increasing to serve the facilities.
2. The amount of data and having different variety collected continuously due to expand use of social, mobile, and embedded technologies.
3. In a competitive marketplace the need of large amount of data for deep understanding of business to maintain stability[2].

The source data which is going to be process further, collected from the geolocation through GPS/GSM modem, Microcontroller and SD Card. The source data can be from any application but we have taken vehicular data for analysis of driving behaviour and each drivers records with various operations like fatigue of drivers, risk analysis and mileage calculations. The data is stored into a SD card in which the data is in the form of date time, longitude and latitude and speed[3]. The GPS/GSM modem is placed inside the vehicle which is connected to the battery itself. This Hardware can be used in any vehicle truck bus, car etc. for the analysis of driving patterns of each vehicle with location. This data is update and stored into SD Card in every 63 sec. The text file generated with all data is secreted into SD card will be also in text format[3]. Ramesh Gardi, Ankita Chavan, et al. proposed land vehicle application on android platform. The purpose of designing computer software is based on certain operations while a land vehicle tracking system need to place the electronic device in vehicle. The vehicle information can be viewed on electronic map via internet or specialized software[4]. Transportation is a very important shared resource that enabling efficient and effective use of resources. GSM modem and GPS unit can be installed on a vehicle and used to track its location. This system is located on the bus and GSM modem communicates through SMS which is connected to a server and phone[5].
GPS/GSM modem is used to carry out and processing the further operations of vehicular events like speed, location, mileage. The data stored into a SD card is in large amount which is in unstructured form[9]. First the data is collected and prepare the RDBMS for the monthly or weekly analysis of driver events. The monthly data is collected with drivers id and corresponding name of driver. Here data is managed in structured form. For processing the drivers activity, speed range, driving patterns and analysis of mileage, accidental ratio, Fatigue of drivers using Hadoop ecosystem. This kind of big data is very much essential for business perspective.

Why Hadoop? Hadoop is used to transform, store and process data throughout enterprise. According to analysts. In the world, unstructured data is about

80% until Hadoop. In any systematic way it was unusable [7].

Hadoop Basic Concepts

APACHE HADOOP

- The best software solution for distributed computing of large datasets is Apache Hadoop.
- It helps to implement HDFS and Map-Reduce [8].
- This is useful for filtering and aggregating the data. e.g. analysis of web server i.e. logs file is used to find page or URL.

2. Overview of HDFS

HDFS is same as DFS (Distributed file system) and differ in many aspects. HDFS's write-once-read-many model. It is used to avoid simultaneous occurrences, to simplify data comprehensibility and to enable high-throughput access. HDFS returns safe condition of event when failure occurs and provide instant results[7].

Definitions /Acronyms

Data Node:

A Data Nodes stores data in the [HadoopFilesystem]. A function of this file system is to collect more than one Data Node with replicate data.

Name Node:

For HDFS, Name Node work as a directory name-space manager and "inode-table".

Secondary Name Node:

The Secondary Name node connects with the Primary Name node. It takes the snapshots of the Primary Name node's directory information and this is saved to local/remote directories.

Map Reduce:

In hadoop, Map Reduce is a programming model. It is used as a software framework. For writing applications which rapidly process parallel amount of data and compute large cluster nodes [10]

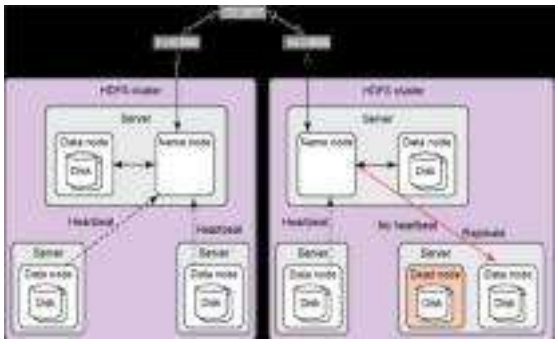

Fig 1: HDFS Architecture

Accordingly, in the cluster it creates number of blocks and then distribute them. It is reliable and able to be retrieved faster. A block size of HDFS is 128 MB[8].

Materials and Methods

Collection of data

This field shows the actual data which is coming from geolocation through GPS/GSM modem stored into SD card. The text file generated into SD card contains large amount of data vehicular events in every minute. It represents the longitude latitude for location and speed according to time and date. The monthly or weekly data can be distributed as per the need and analysis of driver's behaviour for further operation. The source data is loaded as geolocation data and provide to hadoop ecosystem for further operations like driver mileage, driving patterns, location ,speed analysis and Risk analysis etc. This represents the results with position, date, UTC time and speed information stored into SD card.

Data Analysis

The real time data is stored into a SD card with time, date, longitude, latitude and Speed is considered as a source data for further analysis. The data coming from geolocation is in unstructured form so the ETL(Extraction, Transformation and loading) these three basic attributes are used to make it structured. The source file where the driver id, name, source and destination are listed used as an input file. The overall work needed the hadoop framework and using the several methodologies as follows:

1. YARN

YARN is the brain of Hadoop Ecosystem. It allocate resources and scheduling tasks and performs all the processing activities. There are two important components like

1. Resource Manager
2. Node Manager

Resource Manager is work as a node in a processing department. It works on receiving the requests and passes this requests to corresponding node Managers for processing. Every data node consist of Node Manager. It executes the task of each single data node.

Resource Manager has two components

1. Schedulers
2. Applications Manager

2. MAP REDUCE

The Hadoop uses Map and Reduce concepts to process volumes of data-sets. A Map-Reduce program consist of two main components: Mapper and Reducer [10].

A map reduce job works on data set into independent chunks which is processed parallel manner by using map tasks. After sorting the output of the map tasks it becomes input to the reduce task and both input and output tasks are stored in a file system.

Keys and values: In this value and key cannot be defined by itself. Every value associated with a key. Related values are associate with key.

3. APACHE PIG

Apache pig supports pig Latin language, which uses command like SQL. pig latin code of 10 lines = approx. Map-Reduce Java code of 200 lines.

At the back end of pig job executes map-reduce. The pig Latin converts to Map Reduce. It perform set of Map Reduce jobs sequentially and it has a quality of dealing with ideas e.g. black box. Yahoo developed

the PIG. It process and analyzed the large amount of data using ETL(Extraction, Transformation and Loading).

In PIG, first is the load command which loads the data. It performs the functions like grouping, joining, filtering, sorting, of data etc. After getting results either save on screen or can store in HDFS.

4. APACHE HIVE

Hive provides the data summarization, query and analysis. HIVE is worked as data warehousing component. It uses SQL-like interface to perform various operations like writing and reading, managing large volume of data in a distributed environment.

SQL + HIVE = HQL

The Hive consist of query language known as Hive Query Language (HQL). This is similar to SQL. It has two basic components: Hive-command line, ODBC(Java database Connectivity) /JDBC (Object Database Connectivity) driver.

Hive is known as command line interface. This executes HQL commands only. To establish connection from data storage JDBC and ODBC are used. Hive consist of scalable data. It serves both Interactive query processing and Batch query processing.

5. SQOOP

Sqoop is a Data ingesting service. Flume and Sqoop are differentiate as:

1. Unstructured data or semi-structured data can be ingests into HDFS by flume only.
2. Structured data can import and export from RDBMS to HDFS or vice versa by using sqoop.

After submit Sqoop command, the main task is divided into the subtasks. Subtask get map to imports part of data into Hadoop Ecosystem. Whole data can be import by MAP task. In a similar way Export also works. It is mapped the group of data in the form of chunks from HDFS by Map Tasks, when job is submit. In the form of structured data these chunks are export to a destination. Hence all these chunks of data get received in combine at destination. It is stored most of the cases in RDBMS(MYSQL).

6. FLUME

A service which to helps to move semi-structured and unstructured data into HDFS can be done by Flume. It provides a reliable and distributed solution. It performs the functions like collection, aggregation and moving large data sets. It helps to process online continuously moving data coming from various sources like network traffic, email messages in HDFS.

Visualization

The process of data visualization is a graphical representation of large data volume. It helps to monitor or analyze the drivers behaviour and different data sheets, table, charts and maps. This process can be done through the hadoop Cloudera but there is some UI problems so this can be used for the same purpose. Data received from Hive is send to Tableau[3]. Tableau is used as business Intelligence tool. It is used for visual analysis of data. Users are able to create distributive, interactive and sharable dashboard. It introduced graphical representation for the distribution of trends, variations and density of the data. Tableau work on the files, relational tables to analyze and process the data. Using Tableau enables to calculate all requirement and can generate various reports. The reports is represented in various formats like charts, diagrams, printed reports etc using data files.

Results and Discussion

Load

In this phase the sensor data which extracted from the various vehicle get stored into text file as excel and the text file get processed. For this the captured sensor data is first stored into the HDFS file in

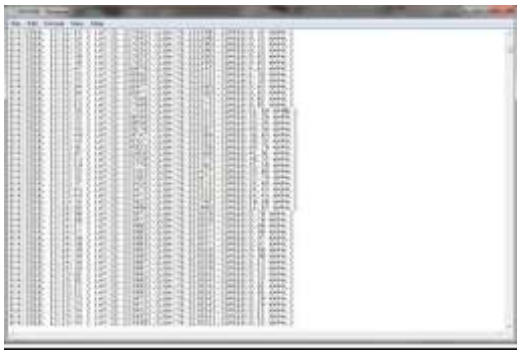

Fig 2: Text file Generated from Geolocation (Source Data)

hadoop. Place sensor in the vehicle to trace the driving speed, change in speed, current location, time taken to reach from source to destination etc. The data from this sensor will be stored in the system or RDBMS which is in JSON format. Were as geolocation data which is in unstructured format is stored in local spool directory.

Refine

The data from trucks sensors i.e. Sensor attached to breaks, sheering's etc. which is in JSON format stored in RDBMS can be transferred to hive using Sqoop. It is a command-line interface application for transforming data between relational database and Hadoop. Apache Hive is a data warehouse software project. It provides data summarization, query and analysis. The unstructured data coming from GPS i.e. geolocation data is loaded first to HDFS to hive by using flume. Apache Flume is a service. It provides data ingestion mechanism for transportation of large amounts of streaming data sets and collecting aggregation example like log files, events from various sources to a centralized data store.

Fig 3: Data Block for Query Analysis with Log Files

The above window showing the main home page of a Hive where we are able to fire query and further processing with all table shown on left side of windows and represent the executable SQL query, how to see the output results of respective table.

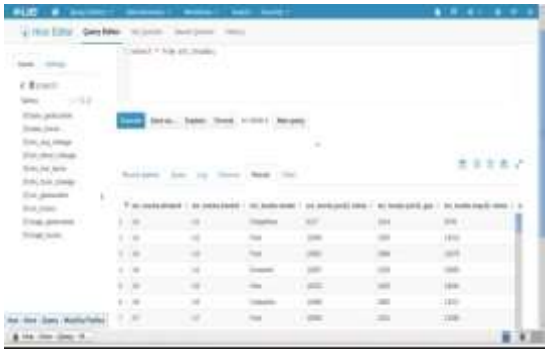

Fig 4: Refine the Data Sheet on Monthly Basis

Above window shows the records of drivers with driver id, distance travel on monthly basis which can analyze for further processing.

Processing

Data is processed in hive using HIVE query and fetched data for different parameters. We will now remove/discard unwanted data this process is called cleansing.

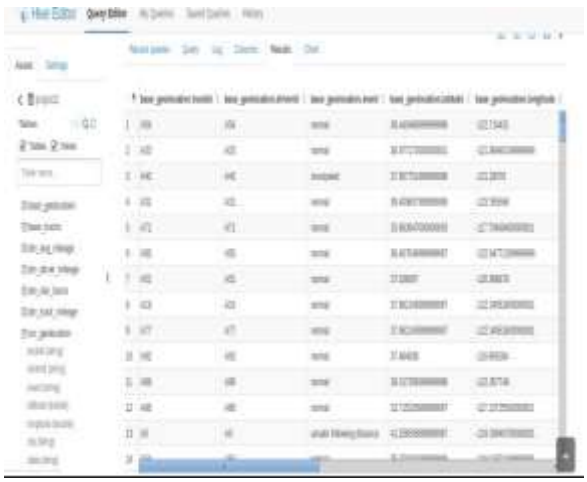

Fig 5: Output of Source Geolocation Data

The above window showing the source geolocation table with Driver id, truck id, longitude and latitude and calculate the risk factor as a category normal, over speed when vehicle speed ranges greater than 40 or 50.

Visualize

Visualization is the process of representing the operational into graphical form on the basis of comparison, charts, tables, maps etc. It is possible to visualize into hadoop as well as tableau.

Fig 6 : Risk Analysis

This window shows the accidental ratio after driving of each driver which helps to analyze the driving behavior of each driver and for processing the operations of driving events. Below windows shows the graphical representation of each drivers speed with respect to Driver's id.

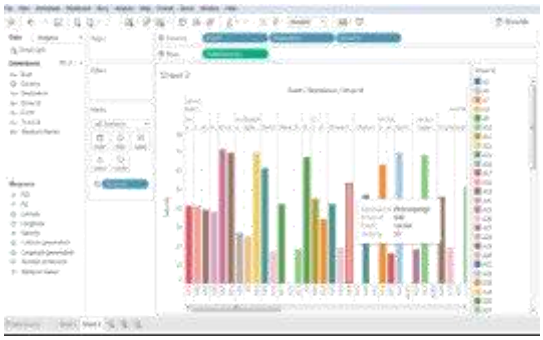

Fig 7: Graphical Representation of Data Analysis

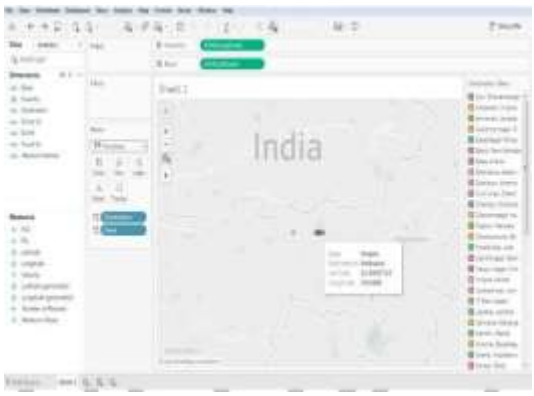

Fig 8: Location Analysis on Map

This window represents the longitude and latitude over a map to find the location.

Single Node Cluster

When Name Node, Data Node, Job Tracker and Task Tracker all essential daemons run on the same machine then is known as Single node cluster. The default replication factor is 1 in single node cluster. A single node cluster is used to simulate a full cluster like environment. To test hadoop applications and unlike the Stand-alone mode. HDFS is accessible in this node. Single node Hadoop means setting up all the Name Node, Data Node, Resource Manager and Node Manager on a single machine. This is used for studying and testing purposes [9].

Future Scope

Multi –node cluster can be used to distribute the same file operation in multiple system at the time. While in a Multi-node cluster. There are more than one Data Node & Node Manager running. These daemons are running on different machines. In organization for analyzing data multi-node cluster is practically used. In real time when we deal with petabytes of data, it needs to be distributed across hundreds of machines to be processed.

Conclusion

To get familiar with end to end implantation of a big data project using various tools in Hadoop ecosystem. The company wants to use this data to better understand risk. It can used in a several business objective to minimize the time. It analyses the data in large amount with specific results and to implement machine learning and implementation of several frameworks and projects for enhancing the better results in shorter period of time. The data can be analyzed in the form of pie charts, graph, bar graphs.

References

[1]Omesh Kumar, Abhishek Goyal, and Visualization: A novel approach for Big Data analytics, IEEE, 2016.

[2]X.Wu, X. Zhu, G-Q. Wu, W. Ding, "Data mining with big data", Knowledge and Data Engineering IEEE Transactions on, vol. 26,no. 1, pp. 97-107, Jan 2014.

[3]Unpublished but accepted by ICECA Pooja gupta, shrikant zade, "vehicle data monitoring system using GPS/GSM modem and PIC microcontroller", ICECA 2018–1338.

[4]Parvesh Kumar and Shri Krishan Wasan "Comparative Study of K-Means, Pam and Rough K-Means Algorithms Using Cancer Datasets", International Symposium on Computing, Communication, and Control (ISCCC 2009).

[5] N. Vijayalashmy, V. Yamuna, G. Rupavani, A. Kannaki@VasanthaAzhagu," GNSS based bus monitoring and sending SMS to the passengers," International Journal of Innovative Research in Computer and Application Engineering, Vol. 2, Special Issue 1, March 2014.

[6]Nitin Sawant, Himanshu Shah, Big Data Application Architecture Q & A. A Problem Solution Approach, 2013.

[7]https://www.tutorialspoint.com/hadoop/hadoop_quick_guide.htm

[8] Mohsen Marjani , Fariza Nasaruddin , Abdullah Gani , Ahmad Karim , Ibrahim Abaker Targio Hashem , Aisha Siddiqa, And Ibrar Yaqoob, "Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges", March 29, 2017,unpublished.

[9]Abid Khan, Ravi Mishra, "GPS-GSM based tracking system," International Journal of Engineering Trends and Technology, Vol. 3, Issue 2, pp: 161-164, 2012.

[10]Y. Lee, W. Kang, H. Son, "An Internet traffic analysis method with MapReduce", Network Operations and Management Symposium Workshops (NOMS Wksps) 2010 IEEE/IFIP, pp. 357-361, 2010.

[11]https://www.ibm.com/developerworks/library / wa-introhdfs/index.html

