Chawda and Thakur, J Comput Sci Syst Biol 2018, 11:4
DOI: 10.4172/jcsb.1000281

Journal of
Computer Science & Systems Biology

Research Article Open Access

Analysing Big Data in VANET via HADOOP Framework

Rahul Kumar Chawda and Ghanshyam Thakur*

Department of Information Technology, Maulana Azad National Institute of Technology Bhopal, Bhopal, Madhya Pradesh, India

Abstract

Objectives: Illustration of features of big data which make the Vehicular Ad Hoc Network communication more accurate and precise. The open framework like Hadoop and Map Reduce which are used in big data for managing, storing and accessing of information is also provided.

Methods/Statistical analysis: The technology of big data is continuously growing and with its rapid increase it is gaining the attention of the researchers. The data is analyzed and outputted in a form that it helps in making quick responses and action in real time environment like Vehicular Ad hoc network. Big data helps in gaining the insight view of the stored, operational and altered data, to improve the traffic conditions. When the Vehicular Ad Hoc Network and the big data are combined, it helps in maintaining the large amount of traffic triggers very easily as the data mining process in big data helps to make quick decisions on the basis of statistics or graph, which are the result of analysis of data.

Findings: Big data and Hadoop cannot be compared because these two are reciprocal to each other. Big data can be considered as a problem and Hadoop can be a solution to it. The combination of Hadoop and Big data in Vehicular Ad Hoc Network provides services useful for number of applications.

Application/Improvements: A lot of applications can be made in future which helps in making the big data analysis much easier and helps in making the on-road condition more secure.

Keywords: Vanet; HADOOP; Big data; Application unit; Road side unit; Secure file

Introduction

Big data

Big data is an emerging technology in today era which arise many challenges for industry, academia and other commercial organizations. It refers to the datasets which becomes so huge that it hard to handle, control, store, share, extract and visualize. It is basically a fresh topic which is increasing and wide spreading very fast and is needed for smart management. Big data can be both organized and unorganized that is so big that it is very difficult to process it using the old-style database and software methods. The data on the network is growing day by day from different medium like social networking sites, organizations etc. [1]. A dataset can be referred as big data if it is tough to capture and analyze. As in today world there are lot of different mediums through which data is collected like sensor networks, scientific experiments, telescopes and high throughput devices which give rise to the data at higher rate. The off the shelf methods are not satisfactory for storing and analysing the data. Sometimes science is lagging a lot behind to find out valuable information from the gigantic volume of data. Due to the technologies in the big data, the way the business management work has been a lot changed. Data intensive computing has tools which help in tackling with problems of big data [2].

The big data can be précised into 4 V's:

- **Volume:** The huge volumes of data is gathered and analysed. Distributed cloud storage is the solution for storing such information. The security challenges occur due to huge volume of data and applying encryption and decryption norms will reduce the performance.
- **Velocity:** The data extraction should be without delay as the data is gathered and processed in the real time. It is very difficult to maintain the short response time while taking into account the security and privacy.
- **Variety:** Data can be of many different ranges such as voice, web analytical data, text, images and facial data etc. Due to this variety it becomes very difficult to maintain the security and integrity of data.
- **Veracity:** It is basically the fake data which can take place of the original one to deceive the user from obtaining the correct information. It is mainly due to vast data which is difficult to look upon [3].

Frame work of big data and it uses:

The framework consists of many layers which contains distinct applications used:

- Data sources: Many different data sources can be there like sensors, mobile, enterprise databases (SAP, oracle) and web.
- Data Management: The main aim is extraction of semi-structured and unstructured data and manages it with different techniques like file system, data cleaning, data storage, data security, etc.
- Data Analytics: Using the methods like data mining, machine learning, statistics and network analysis the data analysis is done.
- Access and application: The data or statistic is then directed through application to machine. These actions which shows the result on machine are done Via, Secure File Transfer Protocol (SFTP), File Transfer protocol and Java Message Service (JMS) [4,5].

***Corresponding author:** Dr. Ghanshyam Thakur, MSc, Department of Information Technology, Maulana Azad National Institute of Technology Bhopal, Bhopal, Madhya Pradesh, India, Tel: +919584479833; E-mail: ghanshyamthakur@gmail.com

Received May 07, 2018; **Accepted** June 11, 2018; **Published** June 20, 2018

Citation: Chawda RK, Thakur G (2018) Analysing Big Data in VANET via HADOOP Framework. J Comput Sci Syst Biol 11: 249-253. doi:10.4172/jcsb.1000281

VANET (Vehicular Ad Hoc Network)

VANET is classified as an integral part of the mobile ad-hic network as it aids in improving the road safety measures and provides the users travellers ease of comfort. The VANET has taken the limelight towards it in the area of wireless and mobile communication, as it differs a lot from the MANET architecture. The field of ITS i.e., Intelligent Transport System is growing at higher speed as the development in VANET is growing exponentially. There exist various standards for wireless access in VANET like 2G, 3G, 4G and Wi MAX. Moreover, the current topologies and routing protocol of MANET are used for making the routing decision of vehicles in a highly mobile environment. There are many routing protocols that are used for this decision making in vehicular environment (Figure 1) [6].

The communication between the moving devices and Road Side Unit is achieved through a wireless medium called WAVE. This technique of communication gives a range of data to the vehicles and drivers and enables the safety of the road. The main parts of the whole VANET system are Application Unit (AU), On-Board Unit (OBU) and Road Side Unit (RSU). Mainly the RSU is the main communication medium between different moving vehicles to transfer the information and it hosts the application which delivers service sand OBU is a client or peer device which take that services given by RSU [7]. In this case RSU is a service provider whereas OBU is a user. Every vehicle of the network is equipped with OBU and set of actuators and sensors which collect and process information. After processing of gathered information, it is then sent to the other moving vehicles in the network or RSU [8-10].

On-Board Unit (OBU)

It is the wave device which is generally mounted over the vehicle for the exchange of information between the RSU and vehicles. It contains Resource Command Processor (RCP) which has the memory to read and write information and a user interface. The OBU links with the RSU through a wireless medium on the IEEE 802.11p radio frequency channel. The important tasks of the OBU are wireless audio access, geographical routing, reliable message transfer, network congestion, IP mobility and data security.

Application Unit (AU)

It is a part which is equipped within the moving device that uses the application given by the RSU using the capabilities of On-board Unit of better communication. It can be used for safety applications, running the internet by PDA. The main core role of the Application unit is that it communicates with the network through the On-Board Unit which do the entire network functioning and mobility.

Road Side Unit (RSU)

It is the communication device usually has static location along the road side like the junctions and parking spaces as depicted in Figure 2. It is wholly equipped with the network device which has the very less range of communication which has base to IEEE 802.11p radio technology. The road side unit can be extended with the other functioning like safety applications installed in the RSU for low bridge warning, accident warning. Road side unit gives the OBU access to the internet while in the range of the network. The data can be also being sent from on network range to other via the OBU of the other network.

Vehicle to Vehicle (V2V): In this architecture the communication occurs between the vehicles. The transmitter and receiver of the information are through the vehicle. Hence, the collection and distribution of the information is done in the network for fast delivery.

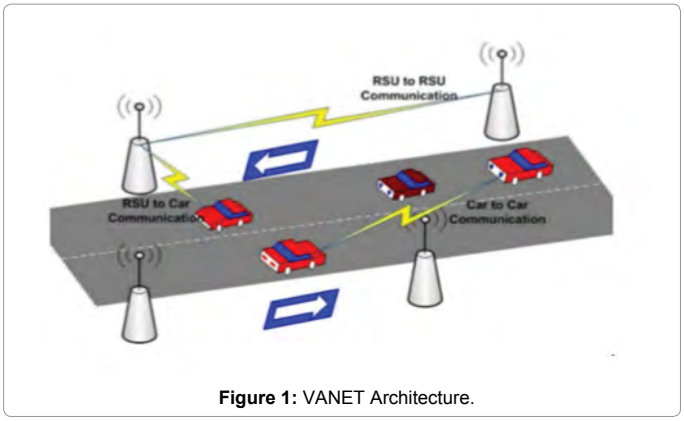

Figure 1: VANET Architecture.

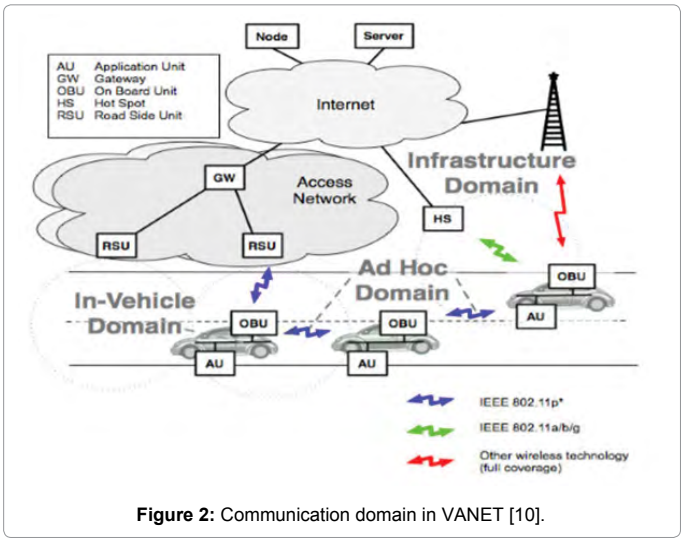

Figure 2: Communication domain in VANET [10].

Vehicle to Infrastructure (V2I): In this architecture the RSU which is static infrastructure alongside the road used for taking the information from the vehicles and delivering it to other vehicles (Figure 3).

Hybrid: This is mainly the mixture of both the Vehicle to Vehicle V2V and Vehicle to Infrastructure V2I.

Challenges of VANET 10

• Multi-Hop data delivery is a difficult task due to recurrent loss of the range of the network and high mobility,

• The information gathering such as speed limit, traffic conditions and information about obstacle is difficult,

• The routing of the packets should be given more importance as to have minimum delay.

VANET in Big Data

In the speed of the modern era, every organization is converting the information into the digital format i.e., in form of graphs. After each interval new changes are done to the statistics according to the newly acquired information. Many different companies are gathering and converting the raw data into the form for analysing and decision making and thus becoming sources for Big Data resources. Now all the

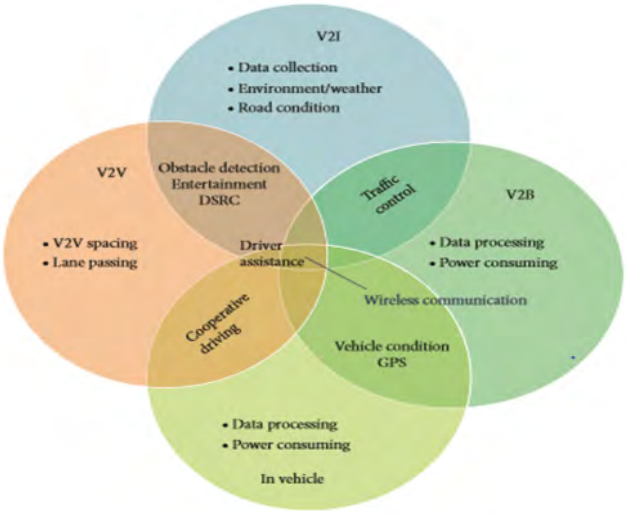

Figure 3: Functions of communication types.

commercial organization are now going for digital representation of data. Some of the areas which have given its contribution in big data are health care, transportation, Retail, Networking, Transportation and Retail, Entertainment media, Video Surveillance and Government sector.

It helps in targeting the possible users and giving them the services, they wanted. There are many challenges in big data such as scalability, privacy, security, storage and processing [11]. Big data is now widely growing in the transportation area i.e., VANET. As large amount of data is transferred in the real the environment full management, storage and control is needed to take accurate and correct decisions related to the congestion, traffic and triggering alerts [12].

Real time data: As the communication in the VANET network is always fast and real, and huge amount of data is transferred over the network with minimum delay. Large table are required for management of the data as they are updated at regular intervals. Mapping of data is very much needed as data is continuously being generated from distinct sources which are also known as Big Data in Real Time.

Variable network density: It refers to the variability in the network data due to variable data sources which continue to send data in the network. Hence data in the communication can be structured or unstructured.

Dynamic topology and mobility modelling: The nodes in the VANET network are highly mobile in nature which results in topology change in a very small interval. The data in the network is generated at very fast rate and even if the topology change occurs data is send via other communication devices through the shortest route. The rates of generation of data, time at which input data is processes and the frequency at which at which it is delivered all are included [13].

High computation ability: Global positioning system (GPS) enabled devices upsurge the computational ability of nodes. Precise value is needed to forecast the decisions of routing.

Giving the address of the neighbouring node and infrastructure support: The main prevailing application is to identify the particular vehicle at some location with basic infrastructure support. So, the neighbouring nodes need to be trustworthy [14].

HADOOP

It is a free platform which computes huge amount of data for the development and execution of different distributed applications. In this method is specified that distribute distributed in a cluster full of machines. HADOOP gives distributed storage known as HADOOP distributed file system. Moreover, it also provides distributed computing with the help of a programming model called Map Reduce. It is basically a framework for processing of big data. In relational databases the processing of structured data is very easy. It is the platform for big data processing (applications, services etc). This whole project is developed by Apache Software Foundation in Java to support variety of applications that depend on particular data. It permits the applications to have access to nodes and data in the network. Google's Map Reduce and Google File System method stimulate to work with HADOOP. HADOOP is used extensively at many commercial platforms. The two key modules of HADOOP are:

- HDFS,
- Map Reduce [15].

HADOOP distributed file system

A cluster of HDFS is working in the pattern of master-worker which contains two types of nodes:

- NameNode (The master),
- Number of DataNodes (Workers).

Name Node manages the file system namespace. It maintains the file system tree and metadata of the whole files in the tree. Data Nodes acts the work horses of the file system. These Data Nodes gather and recover blocks when they are asked to by clients and the data is conveyed back to Name Node at equal intervals with the collection. The data which is needed to be replicated is decided by Name Node. The normal by default size of block is 64 MB is a HDFS with the replication degree of 3. It refers to that one actual data has 2 more copies. The factor of replication can be altered according to the wish. It has a replication policy by default that each Data Node consist of at most on replica of every block. Each rack consists of 2 replicas of every particular block. Sufficient number of replicas is there and after particular time they are checked for over-replication or under replication. In the over-replication, name Node will delete that replica and in the other case under replication the Name Node creates one. When the user has to read the file from the HDFS, Name Node tells it where the particular block and its replica reside in the database. After that, user can read the block from the closest node. For writing the data into the file the user need to take permission from the Name Node again. After writing the content into the node, the block is passed to other nodes for replication. In between each operation Data Node sends a Heartbeat message to the Name Node telling that Data Node is still working. In the case, if the Name Node does-not get the message from the Data Node it is viewed as out of service and at that time Data Node replicas are copied to other nodes.

Map reduce: It a programming model which is linearly scalable and developed by Google. Due to which it easily processes the large amount of parallel data on number of computers. It mainly works on the basis of following two functions.

MAP function: This function takes the key values and convert it into set of data/values or zero.

Reduce function: It takes one unique key and group the connected

values into a unique value set. It provides ease of use and scalability.

The Summarization of Map Function is as Follows

• It takes the input by the client and generates the set of intermediate pair of values.

• The library of Map Reduce connects all the intermediate values which are linked to one unique key and forwards them to reduce function.

• Map (k1, v1).

• List (K2, v2).

The Summarization of Reduce Function is as Follows

• It receives an intermediate key and the values for that key from map function.

• The values are merged to have the smaller set.

• Generally, two values are generated by reduce function i.e., zero or one.

• Through iterator, the intermediate values are send to the reduce function of user.

• Reduce (k2, list (v2)) (k2, v3).

• The domain of the output key values is same as intermediate values.

Hadoop Map Reduce framework of programming models for effortlessly writing applications. It processes large amounts of information in-parallel on big in fault tolerant manner. The term Map Reduce actually mentions to the following two distinct tasks that Hadoop method perform [16,17].

• The Map Task: This the initial part that takes input data and changes it into a set of data in which specific elements are fragmented down into tuples that can be key pairs.

• The Reduce Task: This part takes the output from a map task as required data and integrating data tuples into a minor set of tuples. After the map task the reduce task is always performed.

Normally the file-system stores both the input and the output. The main framework monitors the scheduling tasks, monitoring and re-execution of failed tasks [18].

Algorithm

Preparation of map input: System separates the input into k number of pieces and as an output it gets n number of workers on the machines [19].

Running of user defined map code: Every key-value pair is send to a user-defined Map function and the key-value generated is stored in the memory. On definite intervals, key-value pairs are printed to the disk when it is divided into R regions.

Map output is put to reduce processors: From the local disk of map worker the protected data is read by the reduce worker. The sorting of data is done after the process of reading as the data of similar key are merged together.

Run the reduce code: Over the sorted data reduce worker iterates and finds the unique intermediate key, if encountered it passes the key and its corresponding values to reduce function.

Final output is produced.

Background

Numerous different establishments are accumulating and transforming the unprocessed data into processed form for further analysis and decision making and therefore, becoming sources for Big Data resources. Big data offers several real time opportunities, when it is merged with a wireless medium. Here, in this research work, the utmost benefits of big data within the vehicular ad hoc network has been reviewed.

Firstly, a brief study of Big Data has been given and then complete summary of VANET has been provided along with its utilization with Big Data. VANET along with Big Data has contributed a lot in the different sectors such as health care, transportation, Retail, Networking, Transportation and Retail, Entertainment media, Video Surveillance and Government sector. As large amount of data is transferred in the real the situation, full management, storage and control is required to take precise and accurate decisions interrelated to the congestion, traffic and initiating warning signs. This can be accomplished by using the data mining process in big data which will helps to make quick decisions on the basis of statistics or graph, and a better control can be achieved for the on-road safety. This will eventually lead to decrease in the road accidents and will ensure the safer journeys.

In this paper, the ever-increasing data i.e., the big data has been analysed within the VANET and for the integration purpose the Hadoop Framework has been utilized. The whole framework of the Hadoop has been defined in this paper including its working pattern i.e., Master-Slave pattern. Big data and Hadoop cannot be compared because these two are reciprocal to each other. Big data can be considered as a problem and Hadoop can be a solution to it. The combination of Hadoop and Big data in VANET provides a number of services that can be useful for a lot of applications.

A lot of researches have already been carried by a number researcher till now. Many researchers tried this concept but still there is not any robust solution. A study of the existing literature will definitely help to overcome the drawbacks of the previous researcher and formulating a novel robust solution that will give answers to all the problems that we are facing today.

Related Work

A comparison has been provided by examining co-operative with non-cooperative nature of game players. BCG (Bayesian coalition game) and LA (Learning Automata) has been used for analysing the problem. It is being assumed that the LA is being stationed on the vehicles as game player. For every action, given by automata, a reward/penalty from the surroundings can be achieved by which the automaton updates its probability of its action vector for the actions to be taken in future [20].

Asolution, named, V-PADA (vehicle-platoon-aware data access) has been given. The proposed solution pre-fetches the concerned data and send the buffered data to another vehicle in advance for accessing the data. For that particular aim, a protocol of vehicle platooning has been designed for identifying formation of platoon and then predicts the splits. A component of data management is then designed for guiding the platoon members for replicating and pre-fetching the appropriate data for achieving less data access and more data availability [21].

An analysis of threat analysis and architecture for security in VANET has been provided. The author has defined few design

decisions with more technical suggestions. The security protocol that provides and protects privacy has been shown with their efficiency and robustness. Has defined different results for finding the best strategy implementation in utilizing Hadoop map reduce for the distributed indexing and for analysing the value for practical usage of DTPS by evaluation with same tools. To be precise, the main aim of this research is the valuable search engine development mostly aimed at big data indexing as a main part for the future service of e-discovery. Examine and discuss the solution of big data that may be leveraged for addressing few VANET emerging challenges. The author has defined the big data solution that needs to be leveraged for addressing few of the promising challenges of VANET.

Conclusion

Big Data has reached to heights in every domain. And when it is combined with the wireless network (VANET), it gives a wide range of services and provides a lot of applications to the users like location-based search on the real time data, transfer of data through nodes in real time without delay, maintaining and modifying the database after every new interaction takes place between the nodes. In the vehicular ad-hoc network the decision of traffic controlling, congestion removal and triggering alerts are taken through the statistics generated by analysing the figures of big data coming from different sources in the network. A lot of applications can be made in future which helps in making the big data analysis much easier and helps in making the on-road condition more secure.

References

1. Lin K, Luo J, Hu L, Hossain MS, Ghoneim A (2017) Localization based on Social Big Data Analysis in the Vehicular Networks. IEEE Transactions on Industrial Informatics 13: 1932-1940.
2. Ekbia H, Mattioli M, Kouper I, Arave G, Ghazinejad A, et al. (2015) Big data bigger dilemmas A critical review. Journal of the Association for Information Science and Technology 66: 1523-1545.
3. Mittelstadt BD, Daniel B, Floridi L (2016) The ethics of big data: Current and foreseeable issues in biomedical contexts. Science and Engineering Ethics 22: 303-341.
4. Chen M, Mao S, Liu Y (2014) Big data. A Survey Mobile Networks and Applications 19: 171-209.
5. Chen CLP, Zhang CY (2014) Data-intensive applications challenges techniques and technologies: A survey on Big Data. Information Sciences 275: 314-347.
6. He Y, Yu FR, Zhao N, Yin H, Yao H, et al. (2016) Big data analytics in mobile cellular networks. IEEE Access 4: 1985-1996.
7. Sultan AS, Doori MMA, Bayatti HAA, Zedan H (2014) A comprehensive survey on vehicular Ad Hoc network. Journal of Network and Computer Applications 37: 380-392.
8. Wang Y, Li F (2009) Vehicular ad hoc networks. In Guide to wireless ad hoc networks, pp: 503-523.
9. Hartenstein H, Laberteaux LP (2008) A tutorial survey on vehicular ad hoc networks. IEEE Communications Magazine 46: 164-171.
10. Wenshuang L, Li Z, Zhang H, Wang S, Bie R (2015) Vehicular ad hoc networks architectures research issues methodologies challenges and trends. International Journal of Distributed Sensor Networks, p: 11.
11. Zeadally S, Hunt R, Chen YS, Irwin A, Hassan A (2012) Vehicular ad hoc networks (VANETS): status, results and challenges. Telecommunication Systems 50: 217-241.
12. Nestor BC, Santos AR, Espinoza PM, Velazco JA, Cass AM, et al. (2016) Traffic congestion detection system through connected vehicles and big data. Sensors 16: 599-599.
13. Punam B, Vinita J (2014) Use of big data technology in vehicular ad-hoc networks. In IEEE International Conference on Advances in Computing, Communications and Informatics.
14. Ghafourian GSAA, Hemmatyar AMA, Kavousi K (2016) A network model for vehicular ad hoc networks: an introduction to obligatory attachment rule. IEEE Transactions on Network Science and Engineering 3: 82-94.
15. Yao X, Mohamed F, Alarabi ML, Eldawy A, Yang J, et al. (2017) Spatial Coding-based Approach for Partitioning Big Spatial Data in Hadoop. Computers & Geosciences 106: 60-67.
16. Kyuseok S (2012) MapReduce algorithms for big data analysis. Proceedings of the VLDB Endowment 5: 2016-2017.
17. Darji A, Waghela D (2014) Parallel Power Iteration Clustering for Big Data using Map Reduce in Hadoop. International Journal of Advanced Research in Computer Science and Software Engineering 4: 1357-1363.
18. Grolinger K, Hayes M, Higashino WA, L'Heureux A, Allison DS, et al. (2014) Challenges for Map Reduce in big data. IEEE World Congress on Services.
19. He Y, Yu FR, Zhao N, Yin H, Yao H, et al. (2016) Big data analytics in mobile cellular networks. IEEE Access 4: 1985-1996.
20. Kumar N, Misra S, Rodrigues PCJJ, Obaidat MS (2015) Coalition games for spatio-temporal big data in internet of vehicles environment: A comparative analysis. IEEE Internet of Things Journal 2: 310-320.
21. Yang Z, Cao G (2011) V-PADA Vehicle-platoon-aware data access in VANETs. IEEE Transactions on Vehicular Technology 60: 2326-2339.