Internet of Things Big Data Management and Analytic for Developing Smart City: A Survey and Future Studies

Tuan Anh Vu[1,2,3(✉)], Cong Vinh Phan[4], and Cuong Pham-Quoc[1,2]

[1] Ho Chi Minh University of Technology (HCMUT), Ho Chi Minh City, Vietnam
{vtanh.sdh19,cuongpham}@hcmut.edu.vn
[2] Vietnam National University, Ho Chi Minh City, Vietnam
[3] Faculty of Electronics Technology, Industrial University of Ho Chi Minh City, Ho Chi Minh City, Vietnam
[4] Faculty of Information Technology, Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam
pcvinh@ntt.edu.vn

Abstract. The progress of computing helps us to analyze a large set of data, called big data. Nowadays, there are many Data Analytic Tools for Big Data Analysis, such as Xplenty, Analytics, Microsoft HDInsight, etc.. Big data is a new research field of many research fields as cameras, mobile devices, RFIDs, remote sensing, software log, and wireless sensor network, and IoT (Internet of Things) that is not an exception. Many types of sensors are used to build smart houses, smart cities, and many intelligent things. Smart cities are the future of many countries. A massive data is collected throughout sensors and stored in data centers. It needs to be analyzed in detail to get meaningful information for particular purposes or reduce the database system's size. This paper surveys IoT big data management and analysis. Besides, the study of IoT big data will helps us to find a way to build an intelligent traffic light system in a smart city.

Keywords: Internet of Things · Big Data · Smart traffic light systems · Smart city

1 Introduction

The more progressive social is, the more requirements are. The first need is how to connect some network devices such as modems, switches, and routers for communication in the early days of a computer networking. In human society development, people want more information about cars, devices inside the house (fan, light, door, air controller), and healthcare devices, etc. People use sensors to help things become more intelligent. The sensors get some information about temperature, humidity, speed, action, place, status, situation, etc. Then the data is sent to the data center for analysis or other purposes. Therefore, a smart city

P. Cong Vinh and A. Rakib (Eds.): ICCASA 2021, LNICST 409, pp. 48–55, 2021.
https://doi.org/10.1007/978-3-030-93179-7_4

is researched very much especially in my countries. Our city wants to become better and smarter in 2030. Traffic jam is a real big problem of big cities. The congestion is at a high level and appears usually in corners in rush hours. Traffic police are hard to control traffic jams because of massive corners in the big city (maybe a few thousand corners). The big cities don't have enough traffic police to do this work. It is a waste of time and money. The big cities need an intelligent light system that can solve congestion situations by changing green or red light duration to reduce congestion levels in the corners.

In addition, if the next corner is also in congestion status, cars can't pass through the observed corner. Therefore, the congestion level is higher in the observed corner. Obviously, there is a relationship between the observed corner and next corner. To solve this problem, we need four cameras and controller devices. Besides, we can attach RFID (Radio Frequency Identification) card to emergency vehicles to define when they come to the observed corner. At that time, the observed corner need to turn on the green light first, and then turn back to the red light status after the emergency vehicles already pass through the observed corner. Therefore, the emergency vehicles can save a time of running on the road. Because of massive corners, data of sensors and RFID card are too large and called IoT big data. The traffic controllers analyze it to get much important information. It helps to define the priority of the ways, and the duration of red light and green light. For example, the higher priority the road has, the longer duration the green light has. This is a reason why we need to analyze big data. Next section, we survey IoT big data management and analytic.

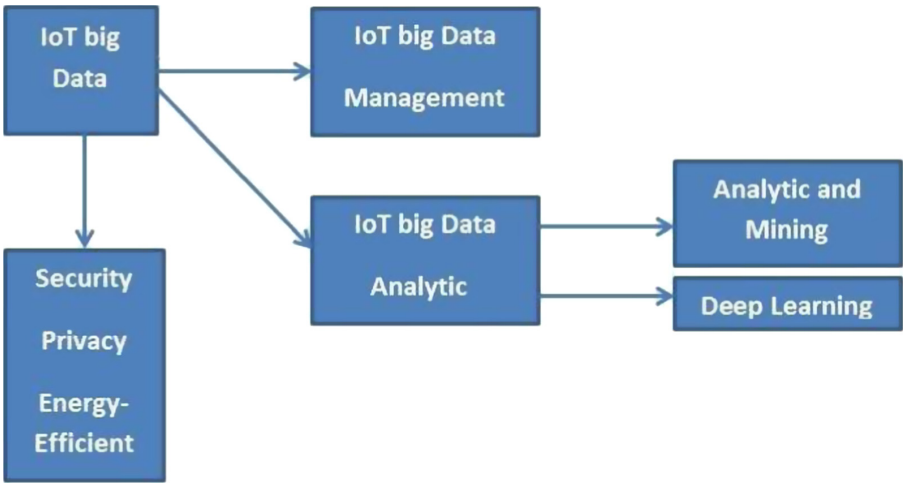

Fig. 1. Overview of IoT Big Data

2 A Survey of IoT Big Data Management and Analytic

In the big data problem of intelligent traffic light systems, the vital thing is classifying the data's position. For example, data belongs to which corner, district, and place. Therefore, we can create a data distribution on Google map. In addition, the time field of data shows congestion's duration. That will also help define place and time of congestion. The congestion status is repeated or not at the same time. The congestion maybe appears every day with higher congestion and more extended time.

The overview of the survey is shown in Fig. 1. The first branch is survey of IoT Big Data Management, the second branch is survey of Analytic, and the remaining one is survey of Security, Privacy, Energy-Efficient. IoT Big Data Analytic is divided into two branches, including Analytic & Mining and Deep Learning.

2.1 A Survey of IoT Big Data Management

Data source of IoT environment has many types and levels. Therefore, we can't compare two types of data due to features such as size, types, and structure. Data includes structured data, semi-structured data, and unstructured data. When data center receives a data flow, it should to classify data due to above features and types. Then we can storage it in database. However, unstructured data is a special type that should be focused on. Because of massive data, we need to scale it down on size.

The paper [13] proposes a Cognitive Oriented Framework (COIB-framework). The framework help to manage data effectively. The framework includes five layers. The first layer, named IoT big-data aggregators, collects all data of IoT sensors network. Then, it divides data into small pieces. This is raw data without encoding or scaling. The raw data is unstructured type and still in same name, scale, and structure level. Anomalies of the data steam can be checked and eliminated due to data fusion operation. The second layer help to categorize cleaned data and save it in clusters according to behaviors and characteristics. The framework can get access to domain and cloud easily. Next layer is HBase storage that have scaling and storing function. In this layer, data is reduced in its size and saved to nodes. The HBase table has two features as key names and data relationship for setting storage nodes. The HBase table can scale big data effectively. The last layer is Iot big-data analysis. It have a tool, named Cognitive CI-Toll, to do a process of knowing data. Then it can decide, plant, and act something with data in next step. The authors believe in success of the framework in the automation environment because of many computing tools. The paper also instructs deploying the framework and cloud computing in the life. However, the paper just give out an ideal without any stimulation or deployment. Therefore, we need to add some features of data like place, number of congestion, priority level of the way. This helps to build intelligent traffic light system so much.

Paper [1] proposes a HEP Framework (Unified Heterogeneous Event Processing Framework) for event data. The event data, collected from sensors and

intelligent objects, is transferred to the framework via Common Event Adapter. Next, the framework has a space to save it according to relationship and XML type. In here, a cache helps data formatting and data processing to be run faster. The data categorizing in before layers also helps data processing effectively in next layer, named Event Processing space. All works should make sure real time constraint because of data streaming. Finally, the data is transferred to end-points. The framework is very suitable for analyzing data of streaming, but the intelligent traffic light system has no data streaming.

The following paper [7] proposes Smart Flood Management Framework. The framework has three layers:

- The first layer is the Internet of Things. The sensor measures water level and speed of water flow, and satellite take some pictures of the water flow. In addition, buoys measure deep factor.
- The second layer is Food Data Storage. This layer has some tasks such us processing, analyzing and reducing. We maybe integrate some tools inside this layer.
- The last layer is the Presentation Layer with Flood Map, where each color presents a level of Flood correspondingly. The authors' setup Food Attributes included Velocity of Catchment, Density of Forests per Acre, and Food Preventing Attributes had Current season, Drainage system, and soil type.

Food Causing Attributes are divided into five levels: the significantly less, the less, the moderate, the high, and the extreme to classify and store big data. In addition, Flood Preventing has the same five levels as >1000, 800–1000, 500–800, 100–500, and <100. Therefore, we can refer to the paper to build three levels of way and congestion, as usual, medium, and high. The considered factors of the system are the car's velocity, congestion level, distance between emergency vehicles and corners, etc. We will explain all things in more detail in future studies section.

2.2 A Survey of IoT Big Data Analytic

In some exceptional cases, we need to make some critical decisions to control the traffic light system. To decide, we should analyze data to help actuators to do some actions. However, big data of the sensors and cameras are a big problem because of the massive traffic lights in the big cities. In the before part, the survey of IoT extensive data management is done already. The management process classifies big data in some attributes for storing and later analyzing. We start discussing the Analytic and Mining section first. The following subsection discusses Deep Learning.

2.2.1 Analytic and Mining

Paper [4] proposes an IoT Big Data Analytic Framework (IBDA framework). The framework is developed on the basics of Python Language and Big Data

Cloudera Platform. They simulate sensors and analyze the data with python code and PySpart tool. Sensors create data and transfer it to HDFS storage. To get it, HDFS need Flume Agent as an agent for get the data Flume Agents is a top-level project of Apache Software Foundation, requiring Java Runtime Environment 1.6, Memory and Disk space, channels or sinks, Directory permissions for reading and writing by the agent. Flume Agents is a data flow unit where events flow is transferred from an external source (like a web server) to the next destination (called hop). The hop is HDFS storage in the next layer. Then the authors use Spark as major software for analyzing data as soon as it arrives into HDFS storage. The last layer is the actuators (turn oxygen pumps, fire alarms, and lights ON and OFF) and Cloudera visualization data. Spark is a unified Analytics Engine for Big Data developed by Apache. Spark also is necessary soft for future researches on building an intelligent traffic light system. Maybe, we also can use Python Code to simulate virtual sensors before applying them in life. Flume Agents is a data flow unit, set by the framework. The external source is a web server where data get out and go to the next-hop named HDFS storage. In the storage, they use Spark to analyze big data. We can apply it to our research on the intelligent light system. We can refer to the paper about sensor stimulation with python code in the lab.

Paper [5] introduces IoTSim as stimulation and analytic software. The IoT-Sime has total five layers. The first layer is the core layer for stimulating Clouds. This layer has network topologies, added sources, added services or already services. It also has a UI structure for users. The network also have delay factors that can be calculated. We can build sensors in the cloud source with some functions as Events Handling. Besides, Datacenter help to manage and analyze data to get results. The second layer is stimulating CloudSim for storing online. The next layer is used for storing data in the system. Data processing is the fourth layer. And the last layer is a layer for coding. The user use this layer to program or code application. We can refer to the paper for simulating the intelligent traffic light system in the lab.

Next, paper [3] explains IoT Fundamentals and IoT Stream Mining Algorithms. The paper also focuses on open-source tools to mine big data such as Spark, Flink, Storm, and Samza. This paper is a good reference for the beginner who needs to be used to data mining tools. Data mining can help to get a map of congestion distribution in the city. Data mining will be discussed in more detail in the future researches part later. The paper introduces some tools to mine big data. Tools, like Spark, Flink, Storm, and Samza, are very popular with users. The beginners should be used to them easily, and we can be the same. We will discuss data mining in a later section.

2.2.2 Deep Learning

The paper [15] proposed a model for deep computing. This is a first step and important thing. The mode integrates tensor (3D data) that have ability to be coded automatically. The paper also instructs calculating errors when the sample

is reconstructed. However, the paper doesn't explain data mining in more detail, and just concern with the results and problem of data.

The paper [5] also proposes a model to do deep learning on data. The paper has the same authors as the paper [15]. The model uses an adaptive distribution that has a rate with a probability named p. There are some attributes been set for the model. The probabilistic model is used for calculating the activating rate, and Maximum Likelihood Estimation is used for estimating parameters of probabilistic distribution to statistic data. The authors used CUAVE and SNAE2 datasets with Matlab R2014 and a high-performance laptop (core i7 and 8G RAM). The DCM results are about 13.3 to 21.3% and are the same with DDCM and ADDCM (about 10.5 to 18.5%). We can refer to the model and apply it to the intelligent light system.

3 Security Privacy and Energy Efficient

Security and Privacy requirements and Innovative Demand of uses conflict in collecting, using, and managing Big Data. Paper [6] is also a survey made by Karen R. Sollins. The author wants to clarify challenges in future researches. Therefore, we will not survey Security Privacy in more detail. The following paper [8] is a request-based, secured, and energy-efficient architecture for handling IoT big data. The nodes have a small battery for actions. The spending energy is due to far distance or near distance between nodes. Therefor, the authors want to save the energy by reducing the distance. Some relay nodes are put in the middle, or use multi-hop system. They have ability to relay data when it is transferred from source to destination. This helps to reduce the distance and save more energy. Besides, the system can turn off passive sensors and turn on them again when they are necessary for actions. Multi-hop system can be used in transferring data to the sink. All of works will increase the lifetime of the system.

4 Future Studies

The survey of IoT Big Data helps to find a way to build an intelligent traffic light system. Basing on all before researches, we need some factors as follows:

– To define some attributes of an intelligent light system, we should have three levels of congestion, speed, and distance, as usual, medium, and high. Suppose the congestion is medium or high level in the next corner. The observed corner should increase more duration of the red light to decrease the congestion of the next corner. The more duration can be from 30 to 60 s according to medium and high levels. However, it still depends on target countries. In addition, the green light needs to be turned on for emergency vehicles and turned back to before status. But, the normal cars can't stop immediately at crossroads, so we need to turn on the yellow light before the red light for about 3 s. That is a reason why we need to know the distance from emergency cars to the corner. This helps the system know the time to turn on the yellow light.

- Using some software such as PySpart, Spark, etc. and some platforms such as COIB, IBDA. It helps build framework and analyze data to create a congestion map. Based on the congestion level in areas, the computer can automatically calculate and define priority levels of ways in the regions.
- The survey helps to decrease spending power so that relay nodes and multihop systems can be applied. In addition, the passive nodes is turned off to decrease spending energy. This increase efficiency and lifetime of the system because of limited energy.

5 Conclusion

The development of IoT is speedy today. In the early days of IoT, the intelligent house is approached very much. The inside devices have become more thoughtful and give more convenience. The other directions research on agriculture, manufacturer, industry, and so on. However, the traffic is also the same field. To build a smart city, we need to develop intelligent traffic light systems. It is a system that can change the duration of red and green lights automatically. The changes depend on the congestion level of the observed corner, congestion of the next corner, and priority level of the way. All attributes are divided into three classes, including usual, medium, and high. The receiving data is categorized and saved to in HDFS. We can refer to the survey papers to use the framework like COIB and IBDA. But, we need to replace 5 attribute level with to 3 attribute level as usual, medium, and high. Next step, some analytic tool are applied to the system. Finally, we can use medium nodes and hops to decrease the distance to help to increase a lifetime of the system. We also refer IoTSim or Python to simulate the intelligent traffic light system in the lab before applying it in life.

Acknowledgments. This research is funded by Ho Chi Minh City University of Technology (HCMUT), VNU-HCM, under grant number BK-SDH-2021-1980918.

References

1. Wang, W., Guo, D.: Towards unified heterogeneous event processing for the Internet of Things. In: 2012 3rd IEEE International Conference on the Internet of Things, pp. 84–91 (2012). https://doi.org/10.1109/IOT.2012.6402308
2. Marjani, M., et al.: Big IoT data analytics: architecture, opportunities, and open research challenges. IEEE Access **5**, 5247–5261 (2017). https://doi.org/10.1109/ACCESS.2017.2689040
3. De Francisci Morales, G., Bifet, A., Khan, L., Gama, J., Fan, W.: IoT big data stream mining. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2016, pp. 2119–2120 (2016). https://doi.org/10.1145/2939672.2945385
4. Bashir, M.R., Gill, A.Q.: Towards an IoT big data analytics framework: smart buildings systems. In: 2016 IEEE 18th International Conference on High-Performance Computing and Communications, IEEE 14th International Conference on Smart City, and IEEE 2nd International Conference on Data Science and

Systems (HPCC/SmartCity/DSS), vol. 1, pp. 1325–1332 (2016). https://doi.org/10.1109/HPCC-SmartCity-DSS.2016.0188
5. Zhang, Q., Yang, L.T., Chen, Z., Li, P., Bu, F.: An adaptive dropout deep computation model for industrial IoT big data learning with crowdsourcing to cloud computing. IEEE Trans. Ind. Inform. **15**(4), 2330–2337 (2019). https://doi.org/10.1109/TII.2018.2791424
6. Sollins, K.R.: IoT big data security and privacy vs. innovation. IEEE Internet Things J., Special Issue on Security and Privacy Protection for Big Data and IoT. ISSN 2327-4662 CD 2372-2541
7. Sooda, S.K., Sandhuab, R., Singla, K., Chang, V.: IoT, big data and HPC based smart flood management framework. Sustain. Comput. Inform. Syst. **20**, 102–117 (2018). https://doi.org/10.1016/j.suscom.2017.12.001
8. Ahad, M.A., Biswas, R.: Request-based, secured and energy-efficient (RBSEE) architecture for handling IoT big data. J. Inf. Sci. 1–12 (2018). https://doi.org/10.1177/0165551518787699
9. Hajiheydari, N., Talafidaryani, M., Khabiri, S.: IoT big data value map: how to generate value from IoT data. In: 5th International Conference on e-Society, e-Learning and e-Technologies, pp. 98–103 (2019). https://doi.org/10.1145/3312714.3312728
10. Saheb, T., Izadi, L.: Paradigm of IoT big data analytics in the healthcare industry: a review of scientific literature and mapping of research trends. Telemat. Inform. 70–85 (2019). https://doi.org/10.1016/j.tele.2019.03.005
11. Misra, N.N., Dixit, Y., Al-Mallahi, A., Bhullar, M.S., Upadhyay, R., Martynenko, A.: IoT, big data and artificial intelligence in agriculture and food industry. IEEE Internet Things J. (Early Access) 1–18 (2020). https://doi.org/10.1109/JIOT.2020.2998584
12. Atitallah, S.B., Driss, M., Boulila, W., Ghézala, H.B.: Leveraging deep learning and IoT big data analytics to support the smart cities development: review and future directions. Comput. Sci. Rev. **38** (2020). https://doi.org/10.1016/j.cosrev.2020.100303
13. Mishra, N., Lin, C.-C., Chang, H.-T.: A cognitive adopted framework for IoT big-data management and knowledge discovery prospective. Int. J. Distrib. Sens. Netw. (2015). https://doi.org/10.1155/2015/718390
14. Zeng, X., Garg, S.K., Strazdins, P., Jayaraman, P.P., Georgakopoulos, D., Ranjan, R.: IOTSim: a simulator for analysing IoT applications. J. Syst. Archit. **72**, 93–107 (2017). https://doi.org/10.1016/j.sysarc.2016.06.008
15. Zhang, Q., Yang, L.T., Chen, Z.: Deep computation model for unsupervised feature learning on big data. IEEE Trans. Serv. Comput. **9**(1), 161–171 (2016). https://doi.org/10.1109/TSC.2015.2497705
16. Tuan Anh, V., Cuong, P.Q., Cong Vinh, P.: Context-aware mobility based on π-calculus in internet of thing: a survey. In: Vinh, P.C., Rakib, A. (eds.) ICCASA/ICTCC -2019. LNICST, vol. 298, pp. 38–46. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-34365-1_4
17. Anh, V.T., Cuong, P.Q., Vinh, P.C.: Context-aware mobility in internet of thing: a survey. EAI Endorsed Trans. Context-Aware Syst. Appl. **6**(16), e3 (2019). https://doi.org/10.4108/eai.13-7-2018.158875