A Big Data Analytics Architecture for the Internet of Small Things

Moneeb Gohar, Syed Hassan Ahmed, Murad Khan, Nadra Guizani, Awais Ahmed, and Arif Ur Rahman

The increase in data rates generated by IoST is escalating exponentially. After attempting to analyze and store the massive volume of IoST data using existing tools and technologies, the SK Telecom Company of South Korea realized the shortcomings immediately. The current article addresses some of the issues and presents a big data analytics architecture for its IoST.

Abstract

The SK Telecom Company of South Korea recently introduced the concept of IoST to its business model. The company deployed IoST, which constantly generates data via the LoRa wireless platform. The increase in data rates generated by IoST is escalating exponentially. After attempting to analyze and store the massive volume of IoST data using existing tools and technologies, the South Korean company realized the shortcomings immediately. The current article addresses some of the issues and presents a big data analytics architecture for its IoST. A system developed using the proposed architecture will be able to analyze and store IoST data efficiently while enabling better decisions. The proposed architecture is composed of four layers, namely the small things layer, infrastructure layer, platform layer, and application layer. Finally, a detailed analysis of a big data implementation of the IoST used to track humidity and temperature via Hadoop is presented as a proof of concept.

Introduction

Big data is progressively growing due to its ability to accomplish a variety of tasks such as designing urban plans, developing and reusing software and products, and interacting with the Internet of Things (IoT). In fact, industry visionaries are taking a greater interest in the Internet of Small Things (IoST) concept than academia; this is likely due to the advantages brought on by data analysis, which can be reused for many purposes. The term big data is related to a specific type of data with three major properties: data is huge in volume and generated from heterogeneous formats, and requires high-speed processing. Several researchers are currently working on designing specific methods and techniques to process the data in real time. Many challenges prevent large data processing in real time since many processing techniques must consider and analyze many aspects of the data. Big data is normally produced by many online and offline sources including email, video and audio chatting, the number of clicks, network traffic, flight routes, and so on [1]. The data generated by these sources is accumulated in a database, which is further processed with high-end software such as Hadoop, GraphX, and Spark [1]. Hadoop is efficient at processing offline data, while Spark and GraphX are mostly used for processing online data up to certain limits. However, storing, analyzing, confining, and normalizing big data is still a challenging job, and it is quite difficult to accomplish even simple tasks like processing and managing data in real time. Apart from these challenges, there are other challenges faced by researchers. For example, presenting visual data (e.g., graphs, charts, summaries, and tables) to decision makers requires more than simply the presentation materials.

Big data has a very close relationship with wireless sensor networks (WSNs) that collect, sense, and accumulate data. These days most of the data are generated by sensors embedded with smart devices in a smart environment such as a smart home, smart cities, e-health, and WiFi enabled transportation systems [2, 3]. The sensors used for various smart city operations generate enormous amounts of data. For example, sensors installed on roads that are used to count the number of vehicles can generate data with thousands and millions of entries during high traffic periods. Data from parking slots in a city can generate data with an unlimited number of entries per day. Similarly, there are many other examples available in smart homes such as tracking gas, water, and electricity consumption from devices installed in a home. Converting such a huge amount of data into a scientific and normalized form can lead to many challenges. The findings of a literature review conducted as part of this work show that there are some approaches available that can address the challenges to some extent. For example, a system was developed that tracks data from a variety of smart city sources and analyzes it with sophisticated tools such as the Hadoop ecosystem [4]. The authors designed a system in which various sensors are deployed to collect data from various sources such as traffic, environment, and smart home data. The data are then analyzed with the Hadoop ecosystem, which is further used to design future urban projects. Similarly, another system was developed by the same authors to analyze geo-satellite data that detects rivers in various parts of the world. A divide-and-conquer approach is used to split the data into groups and then target each group separately [5]. However, all these approaches typically process the data offline. Therefore, a system that can process large amounts of data in real time and support

Digital Object Identifier:
10.1109/MCOM.2018.1700273

Moneeb Gohar and Arif Ur Rahman are with Bahria University; Syed Hassan Ahmed is with the University of Central Florida; Murad Khan is with Sarhad University; Nadra Guizani is with Purdue University; Awais Ahmed is with Yeungnam University.

the decision making process in day-by-day business is highly desirable. Software such as Spark or GraphX can be used to tackle such problems, particularly when dealing with multi-format datasets. Thus, the heterogeneity of the data, which must be processed and stored in a database with great care, is one of the challenging tasks. Designing a database for heterogeneous data is a complicated job. There are a variety of vendors and firms that have come up with sophisticated mechanisms based on relational databases to process data using analytical modeling [6]. This type of specialized software is available in the market in various forms and can be used in a third partitioned environment. Most of the techniques, models, and methods used to filter the data happen in the acquisition phase. However, filtering such a huge amount of data in real time requires optimized and high-processing tools.

SK Telecom Company deployed a massive IoST via a LoRa platform that constantly generated data. There is an exponential increase in the data rate because of the large number of sensors/devices involved in the IoST; the devices also generate data at a very high speed and in a variety of formats. There are many shortcomings to the existing tools because it is not currently possible to process and analyze the complex data generated by IoST in real time. The current article presents a system architecture that can be followed to develop systems that can cope with the robust data generated by IoST. The architecture focuses on the analysis of data for decision making purposes. The architecture consists of four layers, which are explained later in the article. The rest of the article is organized as follows. In the following section, we describe the state-of-the-art research efforts that have been carried out to analyze big data w.r.t IoT. Then we propose a novel architecture for IoST and big data. Following, that we provide system evaluations, while the final section concludes our article.

State of the Art

The use of big data analytics is rapidly increasing as it accomplishes tasks such as designing urban environments, managing traffic, and updating e-health. Using big data analytics in an efficient way is a lofty goal because the data processing techniques still need to be optimized before implementing the existing techniques to solve the challenges and issues encountered by big data. In addition, smart cities are designed based on many concepts ranging from the abstraction level to a specific set of services. Scientists, engineers, architects, and researchers are working hard to come up with a standardized system architecture. For instance, an architecture consisting of many services that a smart city needs which is based on testbeds was proposed in [7]. An architecture based on IoT and smart services was tested in the city of Santander. The system was designed in such a way that a researcher and experimenter can use and test the platform in various urban environments. Similarly, many other systems have been proposed since the Santander architecture was implemented. Many provide planning and design tools for smart environments. These architectures include many resources such as mobility and communication services, security and surveillance systems, large-scale systems, and much more. The testbeds architecture used in the Santander smart city planning tool solves many challenges and issues. The Santander city planning tool was based on big data analytics methods and started a new era of scientific research. The data generated in a smart city using IoT have several characteristics. There are three characteristics that are very common to most of the data: volume, variety, and velocity (3 Vs). The datasets are always generated at high speed and are always available in heterogeneous form. For example, the data generated in a smart home of a smart city can be from various services such as electricity, gas, water, surveillance, and security systems [8]. In addition, maintaining such heterogeneous data is a challenging job. It is quite difficult to process such a huge amount of heterogeneous data in finite time. Similarly, the data generated in real time always needs sophisticated and efficient algorithms and tools to process it in real time. The IoT is not mature enough to be used as a generic standard for generating big data. The recent IoT development is limited to specific domains, and the current techniques of analyzing big data are mature in one area; however, the same method may not be implementable in another area. Thus, the heterogeneity of data remains a challenge. Data management applications in areas such as water management, pollution control, and energy consumption do not guarantee a better solution in the entire smart environment [9]. One of the possible solutions is to develop a system of IoT integrated with WSNs. This can lead us to define a generic architecture; however, we have a long way to go until we reach this goal, that is, a generic big data analytical model.

Data collection via sensors from existing cities can lead us to solve many challenges and issues present in the current literature and real-time environments. However, collecting data from existing cities and converting it to an understandable format is one of the main challenges highlighted in various research studies. Authors have claimed that a system is needed to efficiently convert data to a meaningful format using formal methods. Research carried out by Khan *et al.* uses big data analytics in combination with the Web of Things (WoT). The authors proposed an architecture that follows the case of smart home appliances linked to the web using a RESTful application programming interface (API) [10]. The collected data are converted to a meaningful format using a gateway installed in the management station of the smart home and web. However, the technique used to convert the data from one format to another is not well defined. Additionally, some techniques using a Hadoop ecosystem have been proposed for data processing in recent literature [11]. For example, a divide-and-conquer technique is used in [5] to process data in frames using the Hadoop Ecosystem. Similarly, Yet Another Resource Negotiator (YARN) is used to perform the resource management duties incorporating a separation between infrastructure and programming model. Most of the existing systems have many functionalities that are divided into different layers. Each layer is designed in a way that enables information to be handled using different approaches. For example, data collection is performed through

The use of big data analytics is rapidly increasing as it accomplishes tasks such as designing urban environments, managing traffic, and updating e-health. Using big data analytics in an efficient way is a lofty goal because the data processing techniques still need to be optimized before implementing the existing techniques to solving the challenges and issues encountered by big data.

The outcome of all the discussion is that a generic system is needed that can efficiently process data considering all the aspects of big data. Therefore, the development of a system that can efficiently process big data and report results to improve the daily lives of human beings in a smart city environment is highly desirable.

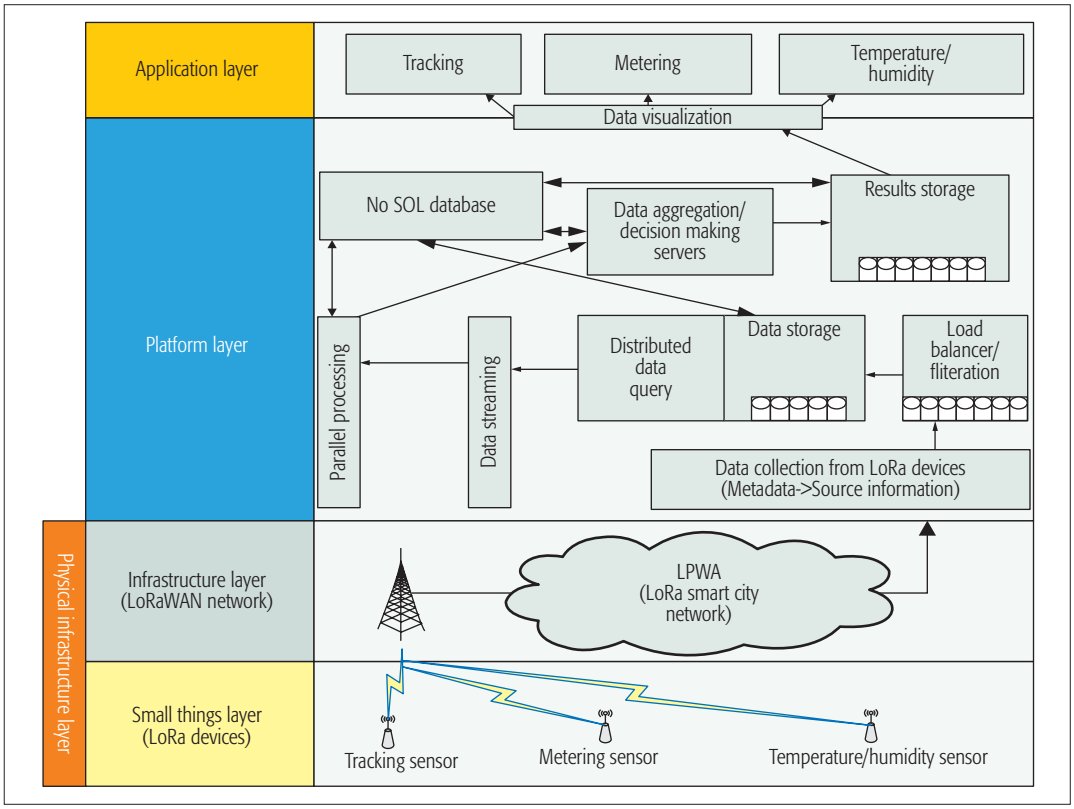

Figure 1. Big data analytics architecture for IoST.

IoT brokers, while storing the data is performed using IoT agents. A server system is designed to process the data collected from sensors deployed in various parts of a city. Finally, the proposed system achieves high throughput by integrating all systems onto a single platform. Similarly, various other systems have been proposed by researchers in the last couple of years. A couple of systems called SCOPE [12] and FIRMWARE [13] use big data analytics and follow a layered architecture. However, these systems are not openly available to researchers and engineers. The outcome of all the discussion is that a generic system is needed that can efficiently process data considering all the aspects of big data. Therefore, the development of a system that can efficiently process big data and report results to improve the daily lives of human beings in a smart city environment is highly desirable.

Proposed Architecture of Big Data Analytics for the Internet of Small Things

The proposed architecture consists of four layers, namely the small things layer (STL), infrastructure layer (IL), platform layer (PL), and application layer (AL). The basic principle followed by the development of the system architecture was simplicity. The architecture should not become very complex, which may not ease deployment. The architecture involves the use of multiple software products as well as a variety of hardware (i.e., sensors and devices used for data collection). The integration of all the tools (software) should be easy and does not require high-level expertise. The architecture should also be easily extendable, and the addition of more modules should not be an issue.

The proposed architecture is presented in Fig.1. Each layer of the architecture is explained next. The self-explanatory flowchart supporting the working of the proposed architecture is depicted in Fig. 2.

Small Things Layer

The IoST layer contains LoRa devices that track and measure temperature, and humidity via a set of sensors. The LoRa devices generate multiple data streams, which results in enormous amounts of IoST data. Each device generates data in a specific format, which includes information about the units, that is, Centigrade, Fahrenheit, location information, unique identification of devices, and so on. The LoRa gateway collects all the information and sends it to the server for further processing in the platform layer.

Infrastructure Layer

In this layer, there are multiple gateways that receive data from the installed sensors that are used for a variety of purposes. LoRa uses the Internet to connect the devices.

Platform Layer

In the PL, the incoming IoST data is initially collected using traditional data collection methods. We assume that the IoST will have a lot of metadata containing source information like device/sensor ID, location information, and device/sensor type. At this layer, various issues are handled including the IoST data redundancy, removal of noise, and fixing minor errors. Moreover, preprocessing is performed using Max-Min normaliza-

tion. Various thresholds limit values are set based on which the dataset is evaluated. Hence, all the unnecessary IoST data is discarded by using filtration algorithms like the Kalman filter. The filtration algorithms are dependent on the type of data collected and the type of issue that needs to be resolved. However, most of the algorithms use the collection of some simple statistics to identify data anomalies and clean bad data points or filter out the noise. Since the necessary IoST data, which is collected from different LoRa devices, needs load balancing in the storage servers, the proposed architecture implements the round-robin (RR) load balancing technique with the integration of the Least Slack Time algorithm (LST) that equally balances the load among dif ferent storage servers and separates redundant LoRa device data. The data storage is based on NoSQL databases such as MongoDB, Neo4j, and FlockDB to store IoST big data and index final and intermediate results. Different distributed data query mechanisms are required to generate queries because of the complex nature of the data; that is, data are collected from various types of sensors for various purposes. After querying, the IoST data is ready for processing. The server intelligently processes IoST data based on web standard specifications. This processing is also based on NoSQL databases. Furthermore, the IoST data being generated by the same LoRa devices are aggregated using the divide-and-conquer approach. After aggregation, the results are stored and maintained using various storage mechanisms including HDFS, HBASE, and HIVE. This is also based on NoSQL databases.

Application Layer

In the AL, the IoST data is visualized by the users via various techniques. The visualization techniques must be easy to understand and help in decision making. The techniques may include simple tables, bar charts, and graphs or complex but meaningful coloring schemes. The scale and units must be easy to understand.

Data Analysis and System Evaluation

The proposed system is implemented using Spark and GraphX with a single-node Hadoop setup on an UBUNTU 14.04 LTS coreTMi5 machine with a 3.2 GHz processor and 4 GB memory. For real-time traffic, we generated Pcap packets from the datasets by using Wireshark libraries and retransmitted them into the developed system. Hadoop-pcap-lib, Hadoop-pcap-side, and Hadoop Pcap input libraries are used for network packet processing and generating Hadoop Readable formats (sequence file) at the collection and aggregation step so that the data can be processed by Hadoop and GraphX. GraphX is used to build and process graphs with the goal of making smart transportation decisions. We have considered the massive volume of data from [14, 15]. The intensity of the traffic varies between times on the same road. The intensity analysis of the various times of day helps administrators manage and make a proper plan for the traffic at that time.

Initially, the analysis is performed on Aarhus city traffic. The speed analysis on the intensity of traffic is carried out as shown in Fig. 3a. When the intensity of traffic is higher, that is, more vehicles

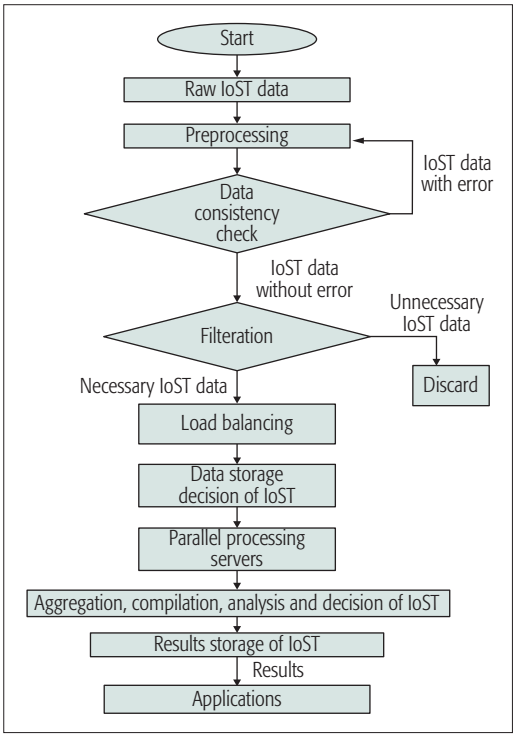

Figure 2. Flowchart of the big data analytics architecture for IoST.

are on the road between two points, the average speed of the vehicles is greater than the actual speed. The drop of vehicles on the road results in a rise in average speed. We can easily notice a higher number of cars (25–30). The average speed is very low at various times of the day, shown in red; a lower intensity (0–10) is shown as a blue line, showing that the average speed of the vehicles is much higher. There are also some abnormalities with the lower number of vehicles, and the average speed is also lower. This might be because of construction on roads or some other incidents. Normally, the distance is conserved to measure the time to reach the destination. However, we observe that the number of vehicles and the average speed also affects the time to get to the destination. Figure 3b shows the blockage of one of the roads in Aarhus. Based on the proposed scheme, the average speed of the vehicles is too low, even when there are a minimum number of vehicles on the road. Most of the road blockage happens during the early morning hours on different days. This is because road construction and commuting occur in the morning. Similarly, we can easily perceive that the increase in the number of vehicles on the road results in more time to reach another point. More traffic on the road reduces the average speed of the vehicles, which results in more time to reach the destination. Because of this phenomenon, we take real-time traffic information to calculate the shortest and quickest path between source and destination rather than only the distance information.

Figure 4a shows the percentage of humidity inside a home. Humidity plays an important role in user behavior when the user is doing physical exercise or any other activity. Moreover, if there is an increase in humidity, the usage of electricity

In the application layer, the IoST data is visualized by the users via various techniques. The visualization techniques must be easy to understand and help in decision making. The techniques may include simple tables, bar charts, and graphs or complex but meaningful coloring schemes. The scale and units must be easy to understand.

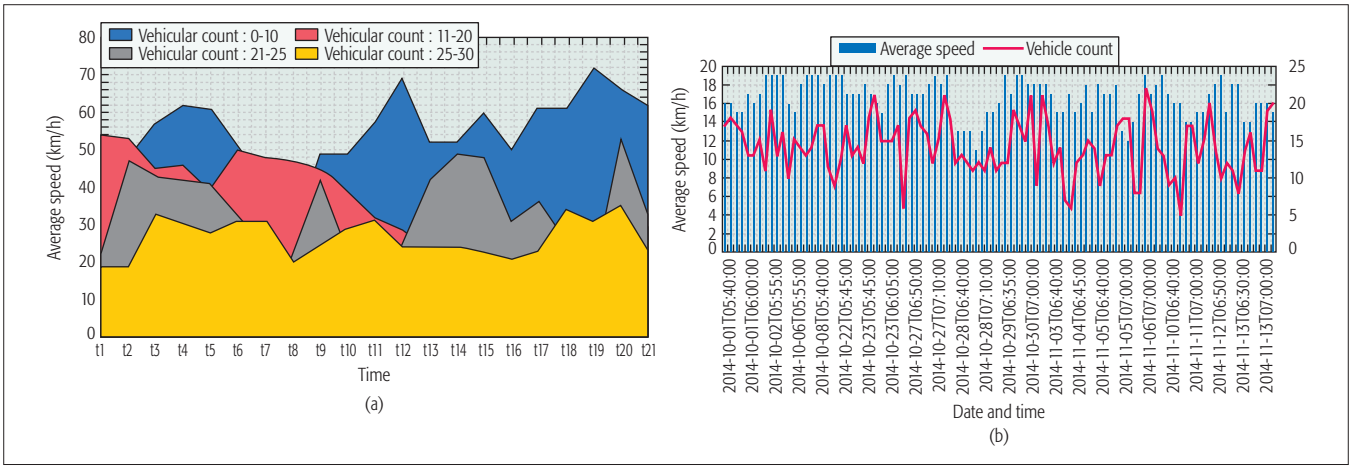

Figure 3. The average speed of a vehicle at a different date and time: a) the average speed of a vehicle; b) average speed during various dates and times.

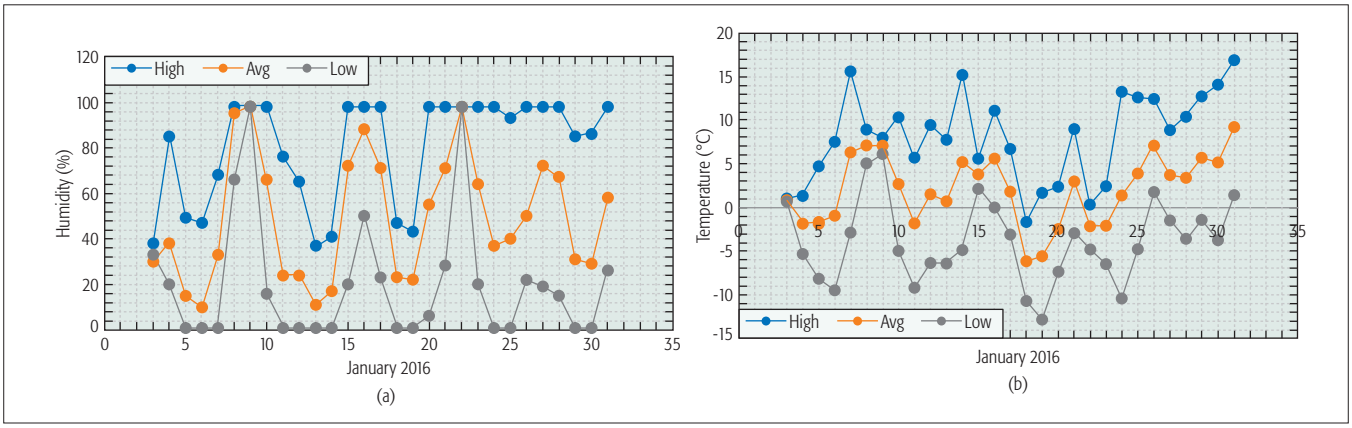

Figure 4. Room temperature vs. outdoor temperature: a) humidity inside home; b) outdoor temperature.

also increases. In this case, the proposed scheme exploits the phenomena of the learning mechanism. Sensors measure humidity, and this data is transferred to our proposed scheme, which records the humidity level. Our proposed scheme considers several readings, and thus creates one threshold during the month of December 2016. Based on previous knowledge, the proposed scheme will predict the weather for the month of January 2016. Thus, the user will react accordingly if humidity is increased or decreased, as shown in Fig. 4a. Similarly, the same technique is followed for outside temperatures, as shown in Fig. 4b.

Since the main contribution of the work is the processing of large graphs to achieve smart transportation, the system is evaluated for efficiency in the throughput (in megabytes per second) and the response time (in milliseconds). The size of the dataset was increased to analyze the system throughput effects and the process of monitoring the efficiency results associated with throughput. We noticed that with the increase in dataset size, the system throughput also increased, as shown in Fig. 5a. To sum up, we conclude that the throughput is directly proportional to the data rate. This is because of the parallel processing of large graphs on the Hadoop ecosystem. When the dataset is larger, the Hadoop system partitions the data into chunks and processes them in parallel. We examine the throughput for larger datasets, for example, 5345 MB. The throughput for this dataset is much better than other systems. This is the major achievement of the system: with an increase in data size, the throughput also increases. However, for a smaller dataset, for example, less than 100 MB, the use of Hadoop is not efficient.

The effect of processing time on increases of the graph is also examined while evaluating the efficiency of the system. We tested the system by increasing the number of nodes and number of edges from zero to 100,000, as shown in Fig. 5b. The massive increase in the number of edges and nodes results in a gradual increase in the processing time to build the graph. Moreover, even for 100,000 nodes and edges, the processing time is quite low (i.e., less than 1000 ms). Therefore, based on the efficiency results, we can say that the system performs well and in real time if it is developed using Spark and GraphX on a Hadoop ecosystem.

Concluding Remarks

In this article, we present a big data analytics method for deployment of an IoST via a LoRa for smart cities. Analyzing and storing a massive volume of IoST data by using the current tools and technologies brings about several challenges. Therefore, this article proposes to use an architecture for analyzing big data that comes from IoST. The proposed architecture can be used to develop systems able to store and analyze the IoST data efficiently and help us make better decisions.

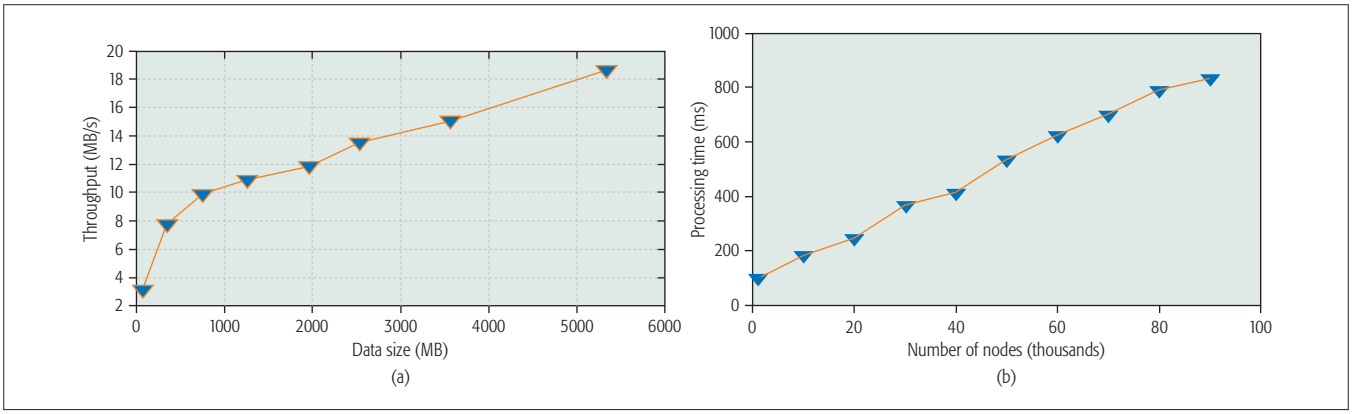

Figure 5. Efficiency of the proposed system architecture: a) throughput of the system, depending on data size; b) graph generation time depending on the number of edges.

A detailed analysis of IoST big data for tracking humidity and temperature is provided in the article using Hadoop.

References

[1] C. Eaton *et al.*, *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, McGraw-Hill, 2012.

[2] S. H. Bouk *et al.*, "Named-Data-Networking-Based ITS for Smart Cities," *IEEE Commun. Mag.*, vol. 55, no. 1, Jan. 2017, pp. 105–11.

[3] A. Al-Fuqaha *et al.*, "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications," I*EEE Commun. Surveys & Tutorials*, vol. 17, no. 4, June 2015, pp. 2347–76.

[4] M. M. Rathore *et al.*, "Urban Planning and Building Smart Cities Based on the Internet of Things Using Big Data Analytics," *Computer Networks*, vol. 101, June 2016, pp. 63–80.

[5] A. Ahmad, A. Paul, and M. M. Rathore, "An Efficient Divide-and-Conquer Approach for Big Data Analytics in Machine-to-Machine Communication," *Neurocomputing*, vol. 174, Jan. 2016, pp. 439–53.

[6] P. Chandarana and M. Vijayalakshmi, "Big Data Analytics Frameworks," *Proc. Int'l. Conf. Circuits Syst. Commun. Inf. Technol. Appl. (CSCITA)*, Apr. 2014, pp. 430–34.

[7] L. Sanchez *et al.*, "SmartSantander: IoT Experimentation over a Smart City Testbed," *Computer Networks*, vol. 61, Mar. 2014, pp. 217–38.

[8] M. Khan, B. N. Silva, and K. Han, "Internet of Things Based Energy Aware Smart Home Control System," *IEEE Access*, vol. 4, Oct. 2016, pp. 7556–66.

[9] N. Maisonneuve *et al.*, "Citizen Noise Pollution Monitoring," *Proc. 10th Int'l. Conf. Digital Government Research: Social Networks: Making Connections between Citizens, Data, and Government*, Puebla, Mexico, May 17–20, 2009, pp. 96–103.

[10] M. Khan, B. N. Silva, and K. Han, "A Web of Things-Based Emerging Sensor Network Architecture for Smart Control Systems," *Sensors*, vol. 17, no. 2, Feb. 2017, p. 332.

[11] B. Cheng *et al.*, "Building a Big Data Platform for Smart Cities: Experience and Lessons from Santander," *IEEE Int'l. Congress on Big Data*, New York, Aug. 2015.

[12] "SCOPE: A Smart-City Cloud-Based Open Platform and Ecosystem"; http://www.bu.edu/hic/research/scope/; accessed Aug. 15, 2017.

[13] "FIWARE Open Source Platform"; https://www.fiware.org/2015/03/25/fiware-a-standard open-platform-for-smart-cities/, accessed Aug. 15, 2017.

[14] S. Bischof *et al.*, "Semantic Modeling of Smart City Data, Position Paper in W3C Workshop on the Web of Things: Enablers and Services for an Open Web of Devices," Berlin, Germany, 25–26 June 2014.

[15] R. Tönjes *et al.*, "Real Time IoT Stream Processing and Large-Scale Data Analytics for Smart City Applications, Poster session," *Proc. Euro. Conf. Networks and Commun.*, Italy, June 2014.

Biographies

Moneeb Gohar received a Ph.D. degree from Kyungpook National University (KNU), Korea, in 2012. From 2012 to 2014, he worked as a postdoctoral researcher at the Software Technology Research Center, KNU From 2014 to 2016, he worked as a foreign assistant professor with the Department of Information and Communication Engineering, Yeungnam University. Currently, he is a senior assistant professor with the Department of Computer Science at Bahria University, Islamabad, Pakistan.

Syed Hassan Ahmed [S'13, M'17] s a postdoctoral fellow with the University of Central Florida, Orlando. Previously, he completed his B.S in computer science from KUST, Pakistan, and a combined Master's/Ph.D. degree from KNU, South Korea, in 2012 and 2017, respectively. So far, he has authored/co-authored over 100 international publications, including journal articles, conference proceedings, book chapters, and two books. His research interests include sensor and ad hoc networks, vehicular communications, and future Internet.

Murad Khan completed his Ph.D. degree at SCSE, KNU. He has published over 40 international conference and journal papers. In 2016, he was awarded the Qualcomm innovation award at KNU for designing a smart home control system. He was also awarded the Bronze Medal in ACM SAC 2015, Salamanca, Spain, for his distinguished work on multi-criteria-based handover techniques. Currently, he is an assistant professor at Sarhad University, Peshawar, Pakistan.

Nadra Guizani is a Ph.D. student and graduate lecturer in the Electrical and Computer Engineering Department at Purdue University. Her research work is in data analytics and prediction, and access control of disease spread data on dynamic network topologies. Research interests include machine learning, mobile networking, large data analysis, and prediction techniques. She is also an active member in both the Women in Engineering Program and ECE.

Awais Ahmad received his Ph.D. in computer science and engineering from KNU. He is currently working as an assistant professor in the Department of Information and Communication Engineering, Yeungnam University. Since 2013, he has published more than 55 international journal and conference papers and book chapters in various IEEE, Elsevier, and Springer journals and leading conferences. His current research interests include big data, the Internet of Things, the Social Internet of Things, and human behavior analysis using big data.

Arif Ur Rahman received a doctoral degree from the University of Porto, Portugal. He is currently working as an assistant professor in the Department of Computer Science, Bahria University. His areas of interest include information retrieval, digital preservation, and architecture design for the Internet of Things.