Authors Version
Original paper published in International Conference for Emerging Technologies in Computing (iCETiC 2019) proceedings published by Springer Link
Download Link: https://link.springer.com/chapter/10.1007/978-3-030-23943-5_13
https://doi.org/10.1007/978-3-030-23943-5_13
Cite this paper as:
Brohi S.N., Marjani M., Hashem I.A.T., Pillai T.R., Kaur S., Amalathas S.S. (2019) A Data Science Methodology for Internet-of-Things. In: Miraz M., Excell P., Ware A., Soomro S., Ali M. (eds) Emerging Technologies in Computing. iCETiC 2019. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 285. Springer, Cham

A Data Science Methodology for Internet-of-Things

Sarfraz Nawaz Brohi, Mohsen Marjani, Ibrahim Abaker Targio Hashem, Thulasyammal Ramiah Pillai, Sukhminder Kaur, and Sagaya Sabestinal Amalathas

Taylor's University, Selangor, Malaysia
{SarfrazNawaz.Brohi, Mohsen.Marjani, IbrahimAbaker.TargioHashem, Thulasyammal.RamiahPillai, Sukhminder.Kaur, Sagaya.Amalathas}@taylors.edu.my

Abstract. The journey of data from the state of being valueless to valuable has been possible due to powerful analytics tools and processing platforms. Organizations have realized the potential of data, and they are looking far ahead from the traditional relational databases to unstructured as well as semi-structured data generated from heterogeneous sources. With the numerous devices and sensors surrounding our ecosystem, IoT has become a reality, and with the use of data science, IoT analytics has become a tremendous opportunity to perceive incredible insights. However, despite the various benefits of IoT analytics, organizations are apprehensive with the dark side of IoT such as security and privacy concerns. In this research, we discuss the opportunities and concerns of IoT analytics. Moreover, we propose a generic data science methodology for IoT data analytics named as Plan, Collect and Analytics for Internet-of-Things (PCA-IoT). The proposed methodology could be applied in IoT scenarios to perform data analytics for effective and efficient decision-making.

Keywords: Internet-of-Things, Data science, Analytics, Big data.

1 Introduction

There was a time when data communication was a challenge between human beings but nowadays due to revolutionized development in the world of standards and network protocols, communication and data exchange has been possible even among the devices/sensors [1]. These devices represent anything literally from our ecosystem such as a wearables accessories, t-shirts, automobile, keychain, sphygmomanometer, chair, game console, air-conditioner, refrigerator, projector, boiler, smartphone, plants, animals, application platforms, humans beings, and bots "connected with smart sensors" to name a few [2-5]. The communication and data generated by these devices (things) come under the world of Internet-of-Things (IoT). Kevin Ashton coined the word IoT in 1999, and its advancement has been directly proportional to the advancement in the internet technology [2]. According to Gartner, there will be 25 billion internet-connected wired and wireless devices by 2020 and those devices will generate data that could be collected, prepared and analyzed to undertake intelligent decisions [6]. IoT platforms have been deployed in various domains including healthcare, agriculture,

military, food processing sector, energy, security surveillance, and environmental monitoring [7-9]. For example, IoT applications are already serving the community in the weather forecast, monitoring the health and well-being of individuals [10]. The data generated in an IoT environment are processed instantly to enhance the effectiveness and improve the efficiency of the entire service domain. Using IoT applications such as Lenovo smart shoes, one can track and monitor fitness data [11]. Furthermore, the electrical appliances including refrigerators and washing machines can be controlled remotely using IoT. The surveillance cameras installed for security purpose could be remotely monitored [12].

Since data plays an integral role in an IoT environment, IoT data could be considered both as a diamond and as dust. Diamond if it is effectively treated using state-of-the-art data science methodology, tools, algorithms and techniques whereas dust if it is improperly or inappropriately analyzed. An IoT system should be able to gather raw data from various network sources and analyze it to produce knowledge. The field of data science could make IoT platforms more intelligent. Data science is a mixture of diverse scientific domains. It uses techniques such as data mining, machine learning and Big Data Analytics (BDA) to identify new insights and patterns from data [1]. Therefore, IoT BDA aims to assist organizations in achieving a better understanding of data, thus leads to effectual results that could benefit their business processes [13]. However, likewise any technology, IoT has its limitations because IoT devices generate and collect a huge amount of personal data whose management poses severe legal and ethical issues related to security and privacy. The objective of this paper is to enlighten the role of data science in IoT. In order to contribute to the domain of data science and IoT, we have proposed a data science methodology. The proposed methodology will assist the data scientists to perform an accurate analysis of telemetry to seek effective insights and undertake smart decisions. This paper is structured as follows: the relationship between IoT and data science is discussed in Section 2. Section 3 contains the discussion on the opportunities of data science and IoT. The concerns of IoT are discussed in Section 4. We have discussed the stages of the proposed methodology with details in Section 5. Finally, we have discussed the future direction of this research in Section 6.

2 The Amalgamation of IoT and Data Science

Huge amount of data have been generating from IoT devices such as RFIDs, sensors, satellites, business transactions, actuators (such as machines/equipment fitted with sensors and deployed for mining, oil exploration, or manufacturing operations), lab instruments (e.g., high energy physics synchrotron), smart consumer appliances (TV, phone, etc.), and social media as well as clickstreams [14]. Figure 1 illustrates the landscape of IoT and Data Science, in which various applications such as smart transportation, smart home and smart grid, generate data using embedded sensors and objects. These generated data are transferred via networks and stored in the cloud for processing using numerous big data technologies. The data scientists use BDA applications with well-defined data science methods to analyze volumes of structured and unstructured

data with various characteristics generated from IoT devices [15]. BDA is used to extract information that assists in identifying trends, discovering correlations, predicting patterns and undertaking effective decisions [16]. However, since IoT data is mostly collected from sensors, it is different from normal big data regarding characteristics such as extreme noise, heterogeneity, and express evolution. In 2030, the number of sensors will increase by 1 trillion that would eventually upsurge the big data [8] [17].

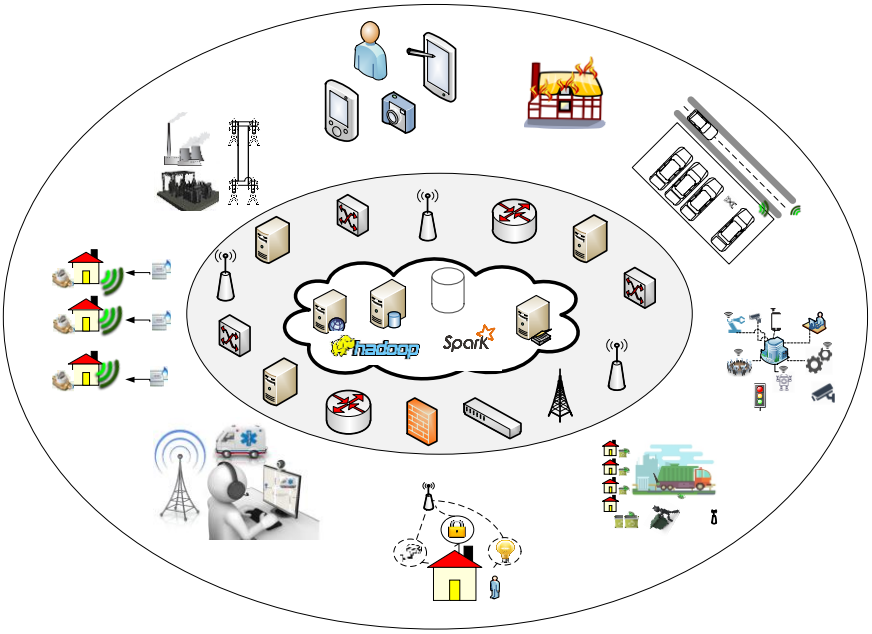

Fig. 1. IoT and data science.

3 Opportunities of IoT and Data Science

IoT is one of the most vital domains of next-generation technology that is obtaining huge attention from the industries widely [18]. IoT technologies offer enhanced data collection, enabling real-time responses, improving the access and control of devices, increasing efficiency and productivity, and connecting technologies [19]. IoT can be considered as a deployment of smart devices which uses data and connectivity [20]. The devices are connected and communicated with each other, and the IoT technologies integrate the collected data from the devices with customer support systems, vendor-managed inventory systems, business intelligence applications, and business analytics tools [18]. The integrated IoT devices produce a huge amount of data rapidly. Hence, data science can play a substantial role in IoT to extract useful information for pattern recognition, trend prediction, and decision-making. Following are some of the opportunities that require IoT and data science to develop more benefits for industry and academia.

3.1 Big IoT Data and Business Analytics

The enormous volume of data is generated by actuators and sensors embedded in IoT machines and devices. This huge amount of data can be transmitted into business analytics and intelligence tools to improve the accuracy of decision-making outcomes. Analyzing markets trends and conditions, and customer behaviors can help business organizations to detect and solve their business issues and increase the level of their customers' satisfaction. Business analytics technologies can be integrated with IoT devices such as wearable health monitoring sensors [18]. This integration provides real-time decision-making possibilities at the source of data. For instance, the health data collected via sensors and monitoring systems such as Humana's Healthsense eNeighbor® remote monitoring system which reports changes in normal activities of its members using in-home sensors can provide opportunities for healthcare providers to analyze the collected data and monitor patients far more regularly and efficiently [18].

3.2 Monitoring and Control System

Monitoring the environmental conditions, the level of energy consumption, and even the performance of equipment require IoT technologies to collect data from available sources and data science to extract useful information for automated controller and managers to monitor the performance and changes of the related objects. Advanced technologies such as smart grid and smart metering offer higher productivity and lower costs by exposing operational patterns, optimizing operations and predicting future changes and trends. One of the well-known IoT monitoring and control Systems is a smart home technology. In this technology, the main intentions are to save energy and also to protect family and property. For instance, the Verizon Home Monitoring and Control network developed remote control applications for home automation using a special wireless communications technology. Users of the applications can monitor and control IoT enabled devices via smartphone, tablet or a computer. They can control the climate, adjust the lights, lock and unlock the doors, manage security systems. The applications also send event notifications to the users automatically. All these functionalities are not possible without analyzing the received data from IoT devices. Another edition of this story is happening in smart cars where IoT technologies are used to monitor and control various parts of smart cars [18].

3.3 Collaboration and Information Sharing

Different types of information sharing can be occurred using IoT technologies. This can be categorized in human-to-human, human-to-things, things-to-human, and things-to-things. For example, in the human-to-human category, communication and sharing information occurs commonly when a manager assigns a task to staffs using IoT enabled mobile devices. When alerts from sensors embedded in a machine are sent to the person in charge of informing about an event like dropping the temperature of the machine, a things-to-human type of information sharing has been happening. Now a user can send a command to the system and react to the alert as a human-to-things type of

collaboration. Sending raw information from a complex machine to a normal user may cause a wrong interpretation. So, the data collected from IoT-enabled devices must be analyzed to take proper actions.

3.4 E-commerce

The real value of IoT for e-commerce platforms is the delivery of intelligent visions which provides new business outcomes. The future of retail is claimed to be e-commerce and shifting to online shopping and marketing is getting the attention of the customers regarding offering more benefits to them. Hence, it is necessary for retailers to adjust their business strategies to embed new technologies such as IoT into their system. Certainly, IoT and big data perform a key role in this ongoing technological disruption. The generated data require to be analyzed to come up with new solutions to improve their business and increase their annual profit. Simultaneously, they should not underestimate the vital impression of their data contribution to gain more benefits by looking for a customized and improved users' shopping experience [20].

3.5 Smart Learning

Activities and behavioral data can be collected from digital sources using IoT devices in various platforms such as social media and online shopping systems. These web-based behavioral data are recorded in different forms such as transactional purchase information or cookies data. IoT devices can observe consumers' habits, preferences, tendencies, and their environments using data science. These IoT enabled devices can learn from the patterns and outcomes extracted from the analytical processes that data science can apply to IoT data. It offers opportunities to markets, providers, and websites to learn more about consumers' needs and interests. This learning process is based on consumers' behaviors in the physical world as opposed to the strictly online world [18].

4 Concerns of IoT: Security and Privacy

Since telemetry travels via several hops in a network, a strong encryption mechanism is essential to guarantee data confidentiality, integrity, and availability. Moreover, the Machine-to-Machine (M2M), Cyber-Physical Systems (CPSs) and Wireless Sensor Networks (WSNs) have progressed as essential components for IoT. Therefore, the security issues related to M2M, CPS, and WSN are rising in relation to IoT. The whole deployment architecture needs to be secured from attacks, which may obstruct the services provided by IoT as well as may pose a threat to privacy, integrity and confidentiality [12]. IoT can bring opportunities for major industries such as healthcare, military, energy, and e-commerce, etc. These opportunities for IoT could also be an encouragement for the hackers to steal a wealth of data generated from IoT sensors due to political and commercial interest [21, 22]. The security of IoT sensors could be violated that could lead to a breach of service integrity [12, 23]. The IoT sensors could retrieve numerous data including the personal information of the users because those sensors can

be integrated into a wide variety of things in our entire ecosystem. The hackers could launch a variety of identity theft attacks on the vulnerable IoT devices for malicious purposes. The ownership of personal data is another concern especially when data is collected without the awareness of the users or with their awareness but without the knowledge of how the data related to them is going to be used and who stays the owner of the data? The European Commission also has doubts regarding data ownership [24]. These challenges related to IoT security and privacy remain the open areas of research. However, efforts have been reported in research and industry standards to make IoT a secure, reliable and trusted platform. Standardization organizations such as IETF and IEEE are also focused on strengthening IoT security by developing necessary communication technologies. These technologies are imperative to enhance IoT reliability and power efficiency. IoT has an extraordinary capability for flexibility and scalability. One of the main goals is to ensure the availability of authentication mechanisms to thwart any attacks, which could compromise the integrity of data and services [23].

5 PCA-IOT: Data Science Methodology for IoT Analytics

Although the IoT and data science are frequently discussed research topics nowadays, to the best of our knowledge and findings, we could not find any paper with the systematic description and application of a data science approach to performing analytics on telemetry. To fulfill the gaps, in this paper we have provided a generic data science methodology named as Plan, Collect and Analytics for Internet-of-Things (PCA-IoT) as shown in Figure 2. The proposed methodology could be applied in IoT scenarios to perform data analytics for effective and efficient decision-making. PCA-IoT initiates with the planning of the project, and it traverses through the collection and analysis of telemetry and ends with the reporting of analytical insights and actions. However, the entire methodology is completely iterative, i.e., there is a possibility to switch backward and forward from one stage to another. For example, a data scientist could switch from analytics to plan stage to modify the initial strategy after the preliminary visualization results. The detailed steps of each stage of the methodology are discussed in the following sub-sections.

5.1 Plan

Since every project has a certain set of goals to achieve, it is imperative for the project to start with the analysis of the requirements. All the stakeholders of an IoT project especially those who require an analytical solution must be involved in the planning stage to ensure that their requirements are being properly understood and analyzed. Furthermore, the main stakeholders such as the domain experts must be involved in every cycle of the project to provide domain knowledge and review and revise the continuous progress as well as the direction of the project to perceive valuable insights and to obtain the required solutions.

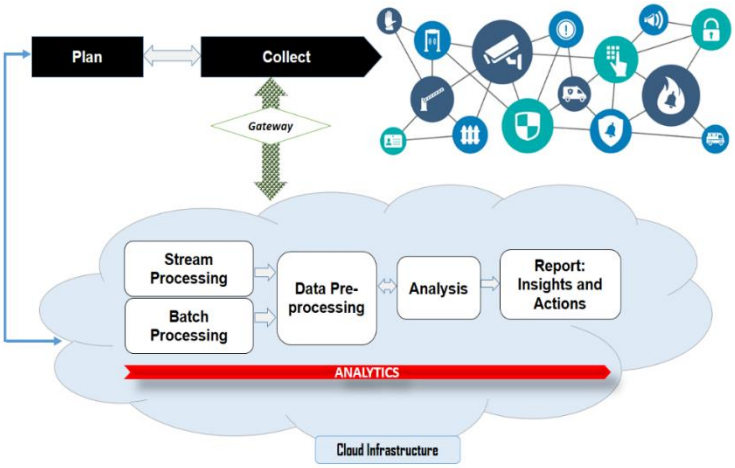

Fig. 2. Data science methodology for IoT analytics (PCA-IoT).

After the successful gathering and analysis of the requirement, a data scientist can formulate the preliminary analytical approaches using statistical techniques and machine learning algorithms to address the problem. With the preliminary findings, team of data scientists, domain experts and appropriate entities from the side of project sponsor could work together to identify and undertake decision on the selection of most suitable analytics tools to be used, algorithms techniques to be applied, the type of models to be generated, and the hosting platform such as in-house or cloud infrastructure. For instance, if the goal is to estimate the relationship between independent and predictor variables, data scientists may choose to generate a regression model. In the planning stage, it is also important to identify the sources of IoT data because telemetry generated from unknown or unreliable sources may lead to inaccurate and invalid analysis.

5.2 Collect

Due to rapidly expanding volume and velocity of telemetry, it would be feasible to perform IoT analytics using third-party cloud services such as Amazon IoT core, IBM Watson IoT, and Azure IoT hub. The gathering of telemetry could initiate after the successful completion of the activities defined at the planning stage. The communication between the IoT hub, i.e., IoT data sources takes places via the gateway which manages all active device connections and implements semantics for multiple protocols to ensure that devices can securely and effectively communicate using various protocols such as MQTT, CoAP, WebSockets, and HTTP. Furthermore, the gateway could apply rules and restrictions to the incoming data using SQL-like statements. A rule can be applied to data from one or many devices. For example, the gateway may filter-out and reject data from certain sensors of the IoT network, or it may accept only certain types of data from specific sensors. The gateway bridge publishes all device telemetry to the cloud that can then be consumed by downstream analytic systems using stream or batch processing.

5.3 Analytics

The data scientist would apply batch-processing techniques to analyze telemetry when analytics takes place on blocks of data that have already been stored over a period. For example, processing all the transactions that have been performed by a major financial firm in a week. However, stream processing will be feasible if real-time analytics is required such as fraud detection and live application monitoring. In an IoT environment, both types of the processing could be useful depending on the requirements and nature of the project especially related to the type of analytics required. Batch processing best fits in the situations where generating real-time analytics results are not the priority and more importance is given to the processing of large volumes of data than to getting fast analytics results. Streaming processing of telemetry can be performed using platforms such as Apache Kafka, Apache Flink, Apache Storm, Apache Samza, etc. whereas batch processing could be performed using Hadoop. Since the sensors can generate inappropriate or null data values, the next step would be to pre-process the telemetry using typical data science approaches such as removing duplication, filter unwanted outliers, handling missing data, etc. Unlike manual data processing in traditional data analytics systems, in an IoT analytics environment, data processing is fast and automated by writing well-defined program codes. During the analysis of data, if data scientists identified that the data needs to further pre-processed, they will switch to pre-processing before performing the analysis. The prepared data is then analyzed using various machine learning and statistical techniques to generate models by considering the steps decided in the project plan. Finally, the models are visualized to perform various analytics such as descriptive, predictive and prescriptive. Due to real-time analytics, organizations, individuals or governments can undertake efficient as well as effective decisions using telemetry.

6 Future Direction

Likewise, any technology such as cloud, big data, and fog computing, etc., IoT has a bright and dark side. However, the research world is currently focused on eliminating the concerns related to IoT to make it as a trusted, reliable and secure platform to seek incredible insights. The research in the field is rapidly increasing, and we could predict that it will continue because data is of high value for the organizations and IoT is the major source for gathering and generating volumes and variety of data. The relationship between the IoT and data science is eternal because to convert data into diamond, analytical approaches are required. However, there are several opportunities to contribute to the areas of IoT and data science. New systems are required to guarantee the security and privacy of users' data and trustworthiness of IoT sensors. Apart from the developments in the world of technology, there is a need to establish new policies, standards, and guidelines for the entire IoT ecosystem to achieve the trust of all the users and to make IoT analytics an opportunity for all types of organizations.

References

1. Mohammad Saeid Mahdavinejad, Mohammadreza Rezvan, Mohammadamin Barekatain, Peyman Adibi, Payam Barnaghi, Amit P. Sheth, Machine learning for internet of things data analysis: a survey, Digital Communications and Networks, Volume 4, Issue 3, 2018, Pages 161-175.
2. Kevin Ashton. That 'Internet of Things'[J]. RFID Journal, 2010.
3. Luigi Atzori, Antonio Iera, Giacomo Morabito. The Internet of Things: A survey. Computer Networks, Vol.54, 2010
4. Mulligan G. The Internet of Things: Here now and coming soon. IEEE Internet Computing, 2010, 14(1) : 35- 36.
5. Rolf H. Weber. Internet of Things–New security and privacy challenges. Computer Law & Security Review, No. 26, 2010.
6. P.P. Ray, A survey on Internet of Things architectures, Journal of King Saud University - Computer and Information Sciences, Volume 30, Issue 3, 2018, Pages 291-319.
7. Xu Da, Li Wu He, Li Shancang, Internet of things in industries: A survey IEEE Trans. Ind. Inf., 10 (4) (2014), pp. 2233-2243.
8. Li S., Tryfonas T., Li H. The internet of things: a security point of view Internet Res., 26 (2) (2016), pp. 337-359.
9. Yuehong Y.I., Zeng Y., Chen X., Fan Y. The internet of things in healthcare: an overview J. Ind. Inf. Integr., 31 (1) (2016), pp. 3-13.
10. M. Rouse, I. Wigmore, Internet of things, 2016. http://internetofthingsagenda.techtarget.com/definition/Internet-of-Things-IoT.
11. B. Heater, Lenovo shows off a pair of intel-powered smart shoes, 2016. https://techcrunch.com/2016/06/09/lenovo-smart-shoes/.
12. Minhaj Ahmad Khan, Khaled Salah, IoT security: Review, blockchain solutions, and open challenges, Future Generation Computer Systems, Volume 82, 2018, Pages 395-411.
13. Mohsen Marjani, Fariza Nasaruddin, Abdullah Gani, Ahmad Karim, Ibrahim Abaker Targio Hashem, Aisha Siddiqa, and Ibrar Yaqoob, 2016. Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges, Volume 5, pp. 5247-5261.
14. Rajiv Ranjan, Dhavalkumar Thakker, Armin Haller, Rajkumar Buyya, 2017. A note on exploration of IoT generated big data using semantics, Future Generation Computer Systems, 76 (2017), 495–498.
15. John B. Rollins, Polong Lin, Alex Aklson, 2017. Data Science Methodology, https://cognitiveclass.ai/courses/data-science-methodology-2/
16. N. Golchha, ‘‘Big data-the information revolution,’’ Int. J. Adv. Res., vol. 1, no. 12, pp. 791–794, 2015.
17. M. Chen, Related Technologies in Big Data. Heidelberg, Germany: Springer, 2014, pp. 11–18.
18. Lee, I., & Lee, K. (2015). The Internet of Things (IoT): Applications, investments, and challenges for enterprises. Business Horizons, 58(4), 431-440.
19. Weinberg, B. D., Milne, G. R., Andonova, Y. G., & Hajjat, F. M. (2015). Internet of Things: Convenience vs. privacy and secrecy. Business Horizons, 58(6), 615-624.
20. R. Rottigni, (2018). Users’ Advantages of Big Data and IoT in E-Commerce. https://readwrite.com/2018/06/05/users-advantages-of-big-data-and-iot-in-e-commerce/
21. Carlo Maria Medaglia, Alex, Ru Serbanati, (2010). An overview of privacy and security issues in the Internet of things. Proc. of 20th workshop on digital communications, 2010.
22. Christoph P. Mayer, (2009). Security and Privacy Challenges in the Internet of Things. Electronic Communications of the EASST.
23. Aakanksha T, Gupta, B. (2018. Security, privacy and trust of different layers in Internet-of-Things (IoTs) framework. Future Generation Computer Systems, Article in Press.
24. Vaclav Janecek, (2018). Ownership of personal data in the Internet of Things. Computer law & security review, 1–14, Article in Press.