Big Data Management in IoT Systems

Avinash Kumar Chaurasia
Computer Science
Paderborn University
Paderborn, Germany
avinashk@mail.uni-paderborn.de

***Abstract*—The Internet of Things (IoT) is an evolutionary concept that aims to connect billions of physical items around the world that are all connected to the internet and collecting and sharing data. The ever-increasing number of IoT devices and the exploding amount of data consumed in IoT shows how the amplification of Big Data intersects with that of IoT. Moreover, the data management task in a constantly expanding network has given rise to significant concerns regarding data collection efficacy, data processing and, analytics. In this paper, I have addressed the challenges arising in four dimensions of Big data namely, volume, variety, velocity and, veracity due to the humongous amount of data generation. This paper also reviews the two approaches, IoT Big data analytics presented for effective Big data management and analytics, and deep learning methodology for data reduction on edge cloud. After performing the comparative study of the two approaches closely, an adapted approach is suggested by merging the above two proposals, which can overcame the challenges not addressed in the first approach and can provide robust data management and analysis of IoT data.**

***Index Terms*—IoT, Big Data, Data Analytics, Big Data Management, IoT Big Data Analytics, Deep Learning, Machine Learning, sensors.**

I. Introduction

The term Internet of things (IoT) came into existence when Kevin Ashton used the term as the title of a presentation in 1999. It is a networking model that is driving the new era of universal connectivity, computing and, communication [1]. IoT has expanded rapidly and will continue to do so in the upcoming years. Besides, It is worth noting that an enormous amount of data is also getting accumulated every day by the internet from numerous sources, for instance, web searches, social media platforms such as Facebook, Instagram, WhatsApp and, many more. IoT simulates this process of data generation by linking the internet with smart devices (sensors), engaging its users with a melange of services and, gathering various kinds of data at the same time [2].

Over the last few years, there have been comprehensive efforts from academicians, service providers, network operators, and standard development organizations to situate the cutting-edge innovations in IoT from notions to fully functional products, hence driving the markets for technology. Recent research on IoT is primarily concerned with connecting the objects for information sharing and facilitating the general objects to observe the physical world for themselves. However, emphasis should be placed on the upcoming IoT which will have a significant count of multifarious networked embedded devices, creating enormous data in a volatile way, thus converting it into a Big Data problem. Furthermore, the collected Big Data will not have any importance unless it is efficiently interpreted and analyzed [1]. This is the point where data analytics come into the picture. Data analytics methods aid in analyzing large volumes of data procured from various data sources like IoT devices and applications. It also provides a mechanism for automated processing of data operations by enhancing the several data operations [3].

IoT devices are generating about 4.4 trillion of data, and the task to oversee that fraction of data is quite difficult. These devices are expanding at a rapid rate and their connection to the internet creates a significant impact on data generation. Moreover, It is simply not that correlation between IoT and Big data is responsible for the complication in data management of IoT devices but the matter of fact is that it is occurring in different types of domains. The process of Big Data management is an intricate task, and data analytics is being utilized to provide the proper management and assure the finest level of accessibility [3]. The interconnection between IoT, Big data, and analytics as depicted in Fig 1 creates a wide range of opportunities for businesses to exercise exponential growth. To summarize it in an unembellished fashion, IoT is a data source, but Big Data is a data analysis platform that provides predictive analyses to foresee future problems and provide solutions [4].

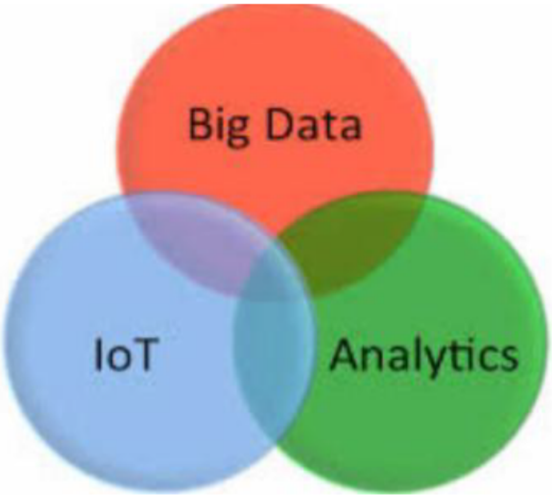

Fig 1: Intersection of IoT, Big Data and Analytics [1]

The purpose of this study is to emphasize the association between Big Data in IoT and challenges occurring due to the collection of heterogeneous IoT data, also called Big Data. This paper also covers the current state of the art for various data management approaches for IoT devices. Additionally, this paper also reviews the two approaches presented for effective data management and analysis in IoT devices and proposes a solution to overcome the challenge of data reduction which was not addressed in the first approach.

II. Problem Analysis

Big Data consolidates the stowing of data, collected from various technology nodes. As a consequence, IoT networks create several heterogeneous data, noise, and some superfluous data, and the network runs based on the analysis of this generated data. Big data corporations try to minimize the likelihood of errors and sustain meticulous decision-making as IoT devices deteriorate with time [2]. Multiple instances in the field of industrial IoT were shown by Wetzkar et al. in [5], where they faced problems in recognizing, interpreting, and troubleshooting failures. There should be automated data acquisition and spontaneous error rectification. The four V's model, which is well-known in the world of big data, contains four dimensions. The four Vs have the following characteristics:

- Volume: size of data.
- Variety: data from a number of different repositories, domains, or types.
- Velocity: rate of information processing
- Veracity: correctness of the data

Few researchers removed veracity from the above-mentioned issues of Big data, thus making it as 3V's, and some have increased the issues by adding Value, Validity to the existing model. Ishwarappa and Anuradha considered 5Vs' in [6], but Khan in [7] considered 10Vs' as Big Data challenges by including Value, Validity, Variability, Viscosity, Viability, and Volatility as new issues. The IoT industry suffers from various challenges in the four dimensions of Big Data which are addressed in the following subsections:

A. Volume

Most of the big establishments put their money in cutting-edge databases, data management firms, distributed systems, and cloud storage for storing digital information. It is also important that the data produced and gathered from various IoT devices are studied, stored, and imparted to other nodes. Performing these operations has also become problematic since the traditional databases are not convenient for handling the quantity of data that is being generated [2]. The prodigious volume of data has made the data processing and integration challenging, and also requires big storage concurrently, which becomes a financial burden for the companies [2].

B. Variety

Big data is created by simultaneously collecting the target data from a variety of sources. IoT data consists of data from various types of sensors, components such as mp3, mp4, radio signals, and many more. This type of auxiliary information is called a meta-data that supports contextualizing the knowledge. Managing the data of this diversity is an arduous task. Therefore, the extracted data must be aligned in proper context with meta-data which ought to connect upcoming data collections intuitively. An additional concern when taking the current state-of-art of IoT and change in techniques into account is the capability of storage software to adjust to these modifications in IoT, such as change of video quality or configuration in sensors [2].

C. Velocity

IoT devices generate data at a significantly high rate. This high intensity of data generation and data gathering is turning out to be problematic since the data must be managed forthwith to provide a way for the new data to breeze in. Furthermore, the velocity of data is dynamic and changes with time. For instance, sales of a company peak over a certain amount of time. Gandomi and Haider give insights into the relevance of time in [8]. To evade data loss and system outage, it is imperative to have sustainable planning, processing capacities, and storage space for the events where data is getting changed with time. While it could be exorbitant to achieve such a degree of computational power, it must be scheduled beforehand to boost the revenue of an organization [2].

D. Veracity

IoT devices like wireless sensors can face certain problems such as system fault, communication error since they do not have mechanisms to check margins of errors. As such, the appropriate storage, precision, and comprehensiveness of data are paramount, to achieve the truthness of data, which is fundamental for many business decisions [2]. As a result, distinguishing between reliable and unreliable data is critical.

To find an optimal solution for Big Data management, I conducted a comparison study on Bashir and Gill's IoT big data analytics (IDBA) framework [9] and Ghosh's deep learning approach for Edge-Cloud Data Analytics in IoT devices [10]. Both the mentioned approaches are sturdy and capable enough to efficiently manage the data processing and storage problem on IoT devices in real-time. After closing analyzing the two approaches,I evaluated the two techniques and found that the IDBA framework fell short of meeting the problem of processing large amounts of heterogeneous data from numerous IoT devices. Therefore, I have proposed an adapted approach in the My Proposal section to mitigate the problem of processing Big Data in the IDBA framework in an efficient manner.

III. State of the Art

Sezer et al. [11] suggested an augmented framework that combines various technologies like semantic web, Big Data, Internet of Things (IoT). Following a study of the essential conditions for the proposed framework, an abstract design of the predicted IoT system is given based on the analysis findings. The proposed framework consists of five layers, namely data acquisition, extract-transform-load (ETL), semantic rule reasoning, learning and, action. The top layer, data acquisition can be considered as the input layer to the framework since it is responsible for collecting data from different sources. The second layer, ETL, assists in the transformation of data gathered from various types of sensors by supplying sensor drivers, and the third layer, semantic-rule reasoning, supports the reasoning engine in making conclusions from the data received from the ETL layer. Then, there is the learning layer which excerpts numerous features from the data and develops machine-learning models. At last, the Action layer imparts predetermined actions for the output of its preceding learning layer [11] [12].

Lee et al. [13] presented an IoT-based cyber-physical system, which has the provision of information analysis and knowledge attainment to strengthening productivity in various industries. The suggested system emphasizes industrial data analytics and incorporates many data analytics components in the form of reconfigurable and indistinguishable modules to meet diverse business requirements. Moreover, The authors present a new context intelligence framework for big data mining that can manage information services based on sensors, locations, and unstructured data [13] [12].

Rathore et al. [14] put forward the idea of a system that can take care of several challenges in a smart city environment, for instance reducing costs for collecting data generated by smart devices, allowing objects to respond concerning context, and procuring foresight from data in case the data is gathered and processed in actual time. The system suggested by the authors has four-tier architecture. The system has a bottom to up approach where the bottom tier is accountable for data generation and collection, intermediate tier 1 allows transmission among sensors, relays, base stations, and the internet, the intermediary tier 2 is in charge of data administration and computation utilizing the Hadoop framework, while the top tier provides data analytics techniques and aids in the generation of results. The execution results shown by the proposed system are much more scalable and efficacious in the context of throughput and processing time than the contemporary systems [14] [12].

IV. comparative Study

A. IoT Big data analytics

An impressive Big Data analytics framework called IoT big data analytics (IDBA) is presented by Bashir and Gill [9], which can render the challenges of storing and analyzing Big Data emanating from smart buildings. IoT Sensors, Big Data Management, and Data Analytics are the three key interconnected components of the given framework. Furthermore, the proposed framework can monitor and control the oxygen level, luminosity, and smoke/hazardous gases of the smart building to refine the user experience, comfort, safety and, health [9].

The workflow of the IDBA framework analytics process is depicted in Fig 2. Following steps are required to implement the IDBA framework:

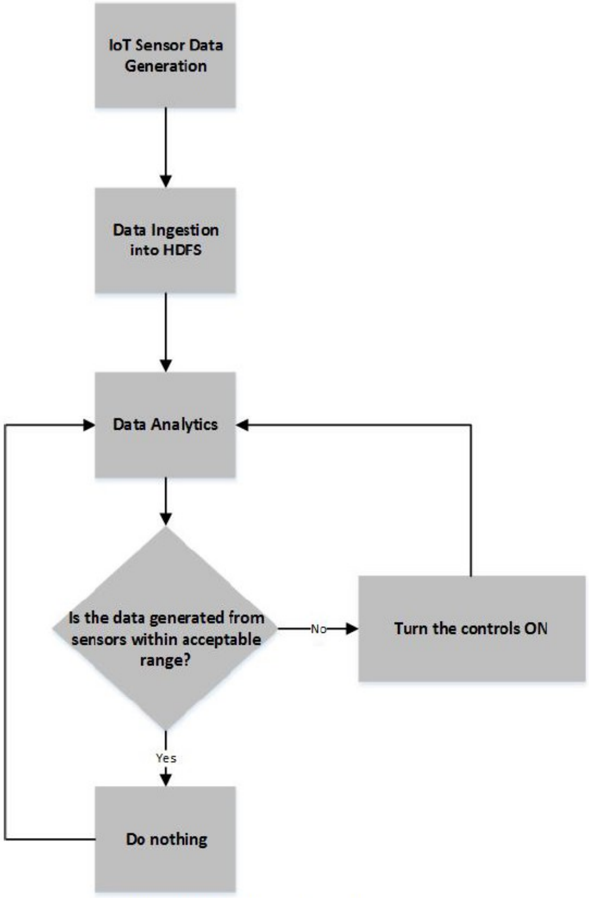
IoT Sensor Data Generation
Data Ingestion into HDFS
Data Analytics
Is the data generated from sensors within acceptable range?
No
Turn the controls ON
Yes
Do nothing

Fig 2: Workflow of IDBA [9]

- Firstly, data is generated using Python code to simulate five virtual oxygen sensors, five virtual smoke detectors, and five virtual luminosity sensors. Apache Flume is then used to send the data to the TCP port [9].
- Secondly, the huge amount of data generated due to the sensors is stored in Cloudera HDFS (Hadoop Distributed File System) using Apache Flume, which provides a robust and coherent model to store large volumes of data [9].
- Thirdly, the authors utilize PySpark scripts (Analytics), which provide the mechanism to analyze the heterogeneous data generated from oxygen, smoke detector and, luminosity sensor. Based on the analysis of data, some measures are being taken, for instance, if the oxygen concentration level is within the users or residents predefined comfortable range, no action is taken; otherwise, an oxygen pump is turned on and remains in the same mode until the level of oxygen measured from the same sensor exceeds the specified threshold level. Similarly, smoke detectors and luminosity detectors are turned ON in case they detect hazardous gas and low luminosity levels respectively [9].

B. Deep Learning: Edge-Cloud Data Analytics for IoT

Ghosh presents a successful solution to combine cloud and edge computing for IoT data analytics in [10] and proposes a deep learning-based approach for data reduction on the edge with machine learning on the cloud. Following are some of the important concepts used in the implementation of this approach:

1) Deep Learning: Deep Learning is a subclass of Machine Learning techniques that allows computer models with several processing layers to learn multiple degrees of abstraction for data representations [15]. This kind of learning employs data representations rather than explicit data: data is turned into hierarchical abstract representations that allow for the learning of useful information, which are then exercised in ML tasks. Autoencoders are used in this approach, to reduce the data sent to the cloud [10].

2) Dimensionality Reduction: Dimensionality reduction in machine learning refers to the methods that reduce the number of features or attributes in a data set and this method is important when we are working on large data sets. Furthermore, Principle Component Analysis (PCA) is commonly used for linear dimensionality reduction technique whereas Autoencoders are widely used for non-linear transformations [10].

As shown in Fig 3, the encoder part of the autoencoder resides on the edge cloud to curtail the data dimensions, and the decoder part is on the cloud to reconstruct the original data from the abstract representation. This way, the high dimensional data are first reduced to a smaller number of dimensions when they approach the edge node. The reduced data is then sent to the cloud where they can be used for machine learning (ML) tasks straightway or the original data can be restored with the help of a decoder placed on the cloud and then further used for ML tasks. The authors used two use cases: in the first use case, autoencoders are used for reducing the data on the edge and the machine learning (ML) tasks are executed on reduced data on the cloud. In the second use case, the autoencoders carried out the same task of reducing data on the edge but, the reduced data is reconstructed to original data before performing the ML tasks on the cloud [10]. The purpose of implementing the mentioned use cases is to compare the accuracy of reduced data with that of original data after performing the ML tasks on them [10].

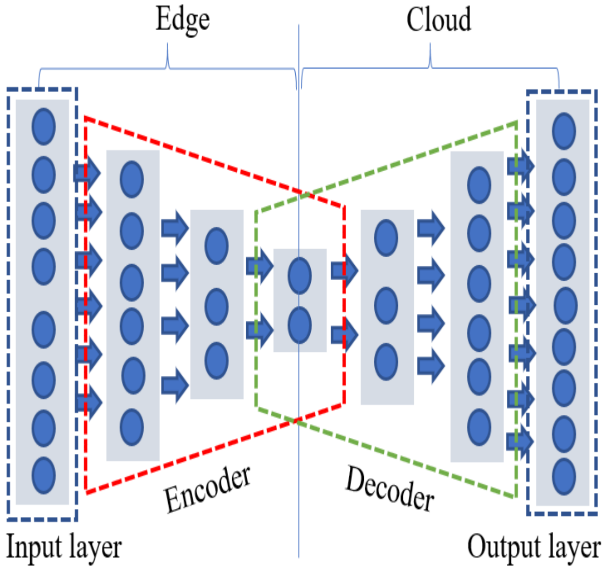

Fig 3: Autoencoder for Dimensionality Reduction in Edge-cloud Architecture [10]

The experiments are conducted on human activity recognition from smartphones. The encoder reduces the features from 561-265-128 to 265-128-64 features. It can be observed that data is reduced from 11.2MB for original data to 5.754MB for 256 features, 2.877MB for 128, and 1.4386MB for 64 features. Hidden layers are added to improve the classification accuracy. Mean Squared Error (MSE), accuracy and, Mean Absolute Error (MAE) were used to evaluate the classification. The results of the two methodologies are highly encouraging, showing that even a 50% reduction in data did not have a significant influence on classification accuracy, while a 77% reduction only resulted in a 1% change. [10].

C. Comparison of two approaches

The first approach presented by Bashir and Gill [9] tried to mitigate the data storage management problem in IoT by pushing the huge data into Cloudera HDFS instead of traditional databases like RDBMS that are usually not scalable to process large data sets. HDFS provides a scalable and resilient solution to a Big data problem. On the other hand, the deep learning approach presented by Ghosh [10] is focused on minimizing the size of IoT data by reducing the data on edge using autoencoders and implementing machine learning on the cloud with the reduced data, to classify the accuracy of data. The first approach utilizes data analytics to analyze the sensors data and provides a decision-making mechanism for smart buildings whereas the second approach uses edge IoT analytics for analyzing the human behavior using smart mobiles.

V. My Contribution

IDBA framework presented by Bashir and Gill [9] is very effective in overcoming the data storage problem in IoT devices. However, it did not properly address the challenge of processing various heterogeneous data on HDFS, which are coming from different IoT sensors. HDFS provides a coherent environment to process large amounts of data and runs on commodity hardware. However, it has certain disadvantages like slow processing speed and latency due to its distributed and parallel mechanism of processing large data files. Therefore, I suggest solving this problem by merging the IDBA framework with the deep learning approach of Ghosh [10] by storing the data on edge cloud instead of HDFS. By employing the deep learning approach from Ghosh [10], we can perform data reduction on edge, and later the compressed data will be put on the cloud to perform the data analytics process for better decision making in IoT sensors. Since the data will be reduced on edge itself, the data analytics process performed on the cloud will be much faster and robust. My proposed workflow for the adapted IDBA is shown in Fig 4. Furthermore, the implementation of the adapted IDBA model can be divided into three steps:

- Firstly, Python code can be used to simulate the data from five virtual oxygen sensors, five virtual smoke detectors, and five virtual and luminosity sensors. The sensors generate the data after every ten seconds, which leads to the generation of huge data. [9] This data can be put on edge servers. Afterward, edge servers can perform part of the computation on the device itself or on a node close to the source of the node, which reduces the data sent to the cloud and consequently, reduces latencies and improves response time [10].
- Secondly, the Big Data collected at the edge will be used for data reduction. Autoencoders will encode the data by compressing it, which will reduce the size of huge data significantly [10]. The reduced data is then sent to the cloud for further data analytics tasks.

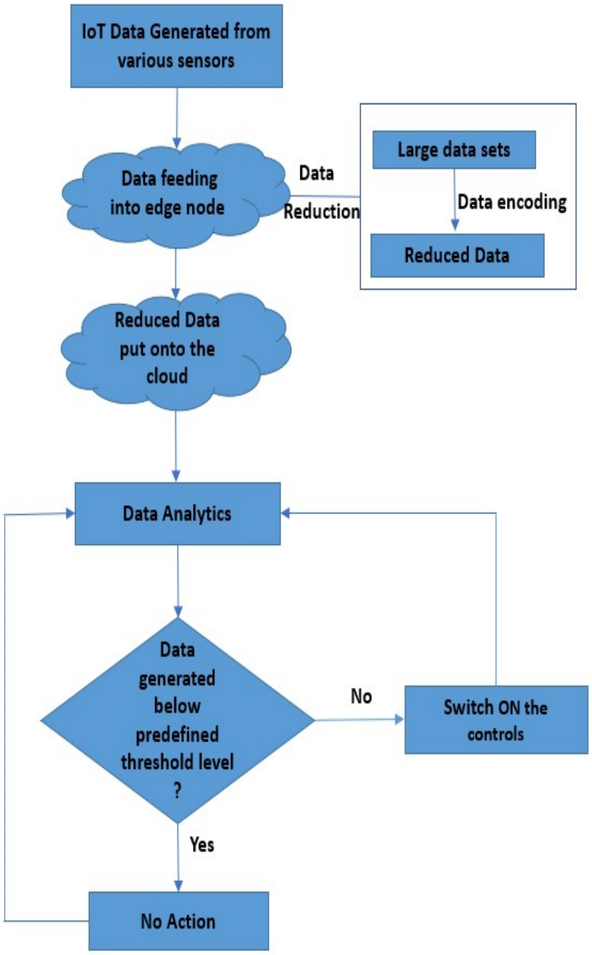
IoT Data Generated from various sensors
Data feeding into edge node
Data Reduction
Large data sets
Data encoding
Reduced Data
Reduced Data put onto the cloud
Data Analytics
Data generated below predefined threshold level ?
No
Switch ON the controls
Yes
No Action

Fig 4: Adapted Workflow of IDBA

- Thirdly, data analytics processes will be carried out on the cloud to analyze the reduced data. Necessary actions like turning oxygen, fire, and luminosity ON in case the data goes below the predefined threshold [9].

VI. Conclusion

The proliferation of IoT devices has resulted in massive volumes of data, which is anticipated to continue. IoT has emerged as one of the biggest sources of Big Data, which is unavailing if not managed and analyzed properly. In this paper, I have addressed the challenges attributing to Big Data in IoT devices and done a comprehensive review on two approaches, providing adequate solutions for managing Big data in IoT using robust technologies like Cloudera Hadoop Distributed File System and Deep Learning respectively. Based on the analysis of two proposals, I have proposed to merge the two approaches to mitigate the problem of processing huge data on the Cloudera HDFS, which was not addressed in the first approach. The proposed methodology will lead to effective management and analysis of Big Data in IoT. My proposal is limited to IoT devices in smart buildings and smartphones but the approach can be extended to other paradigms like smart cities and smart homes, which also contain sensors for oxygen, heat, and luminosity.

References

[1] B Sobhan Babu, T Ramanjaneyulu, I Lakshmi Narayana, and K Srikanth. Data management in iot using big data technologies and tools.

[2] Shivanjali Khare and Michael Totaro. Big data in iot. pages 1–7, 07 2019.

[3] Owais Khalid and Suntharalingam Senthilananthan. A review of data analytics techniques for effective management of big data using iot. In *2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA)*, pages 1–10, 2020.

[4] Nick McKenna. Explaining the relationship between iot, big data and cloud computing.

[5] Ulf Wetzker, Ingmar Splitt, Marco Zimmerling, Kay Römer, and Carlo Alberto Boano. Troubleshooting wireless coexistence problems in the industrial internet of things. 08 2016.

[6] Ishwarappa Kalbandi and J. Anuradha. A brief introduction on big data 5vs characteristics and hadoop technology. *Procedia Computer Science*, 48:319–324, 12 2015.

[7] Nawsher Khan, Mohammed Alsaqer, Habib Shah, Gran Badsha, Aftab Abbasi, and Solmaz Salehian. The 10 vs, issues and challenges of big data. pages 52–56, 03 2018.

[8] Amir Gandomi and Murtaza Haider. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35:137–144, 04 2015.

[9] Muhammad Rizwan Bashir and Asif Qumer Gill. 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; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pages 1325–1332, 2016.

[10] Ananda M. Ghosh and Katarina Grolinger. Deep learning: Edge-cloud data analytics for iot. In *2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)*, pages 1–7, 2019.

[11] Omer Berat Sezer, Erdogan Dogdu, Murat Ozbayoglu, and Aras Onal. An extended iot framework with semantics, big data, and analytics. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 1849–1856. IEEE, 2016.

[12] Ejaz Ahmed, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Imran Khan, Abdelmuttlib Ibrahim Abdalla Ahmed, Muhammad Imran, and Athanasios V Vasilakos. The role of big data analytics in internet of things. *Computer Networks*, 129:459–471, 2017.

[13] C.K.M. Lee, C.L. Yeung, and M.N. Cheng. Research on iot based cyber physical system for industrial big data analytics. In *2015 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pages 1855–1859, 2015.

[14] M Mazhar Rathore, Awais Ahmad, Anand Paul, and Seungmin Rho. Urban planning and building smart cities based on the internet of things using big data analytics. *Computer networks*, 101:63–80, 2016.

[15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.