Quality Assurance in Big Data Analytics: An IoT Perspective

Nicole Ann Fernandes and Rupali Wagh

***Abstract* —Emergence of IoT as one of the key data contributors in a big data application has presented new data quality challenges and has necessitated for an IoT inclusive data validation ecosystem. Standardized data quality approaches and frameworks are available for data obtained for a variety of sources like data warehouses, webblogs, social media, etc. in a big data application. Since IoT data differs significantly from other data, challenges in ensuring the quality of this data are also different and thus a specially designed IoT data testing layer paves its way in. In this paper, we present a detailed review of existing data quality assurance practices used in big data applications. We highlight the requirement for IoT data quality assurance in the existing framework and propose an additional data testing layer for IoT. The data quality aspects and possible implementation models for quality assurance contained in the proposed layer can be used to construct a concrete set of guidelines for IoT data quality assurance.**

***Keywords* — Big Data, Internet of Things (IoT), Data Quality, Data Testing, IoT data Validation, Quality of Service (QoS).**

I. Introduction

IoT or internet of things has not only changed our day to day lives but also revolutionized the entire computing and analytics paradigm. Today IoT is the key contributor in making informed decisions across domains. With these connected devices generating enormous data, seamless integration of this data in a big data application for further analytics is the need of the hour. Since quality data is the backbone of any analytical solution, ensuring the quality of big data is a fundamental task in big data testing. Since the poor data quality may produce inaccurate results, a comprehensive data quality assurance framework is followed for big data testing [1]. The famous V's of big data – volume, variety, velocity, and veracity bring complexities with them. This has been the reason for the inclusion of rigorous data quality check which otherwise was not required in a traditional system [2] data testing.

Paper received October 30, 2018; revised April 4, 2019; accepted May 04, 2019. Date of publication December 25, 2019. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Miroslav Lutovac.

Nicole Ann Fernandes is a postgraduate student, Department of Computer Science, CHRIST (Deemed to be University), Bengaluru, India (e-mail: fernandes.ann@mca.christuniversity.in).

Rupali Wagh is Associate Professor with the Department of Computer Science , CHRIST (Deemed to be University), Bengaluru, India (e-mail: rupali.wagh@christuniversity.in).

In the last decade, we have witnessed the dominance of IoT and today IoT has become a major contributor in the big data application environment. It brings newer complexities in the big data ecosystem. Vastly different sensors from a huge network of connected devices produce data which require careful and systematic preprocessing before actually being fed for analytics. While the wear and tear of the devices/sensors, faulty devices, etc require actions which may be extrinsic to the computing life cycle, but identification of these issues needs to be done intrinsically by analyzing the captured data. IoT is further challenged by security concerns and network issues as they directly impact the reliability and accuracy of data. Thus, the data validation for IoT data goes beyond just data cleaning, aggregation and transformation, and shifts more towards intelligent and machine learning based methods in data testing like ontologies for data abstraction and predictive methods for threat prediction. Since IoT based big data analytics is becoming more and more prevalent, the data quality issues are becoming very significant. Additionally, IoT analytics due to its ubiquitous nature impacts human life largely and hence ensuring the quality of IoT data has become very critical.

In this paper, we discuss major data quality challenges specifically with respect to IoT data. We also elaborate the implementation models used to assure the quality of IoT data and propose an additional IoT data validation layer, which can act as a basis for constructing an IoT inclusive data quality assurance framework for any big data application.

The paper is organized as follows- Section II elaborates a generic big data test framework, section III emphasizes the dominance of IoT data in today's big data applications. Section IV presents data quality challenges with respect to IoT data and various implementation models and methods required for IoT data quality assurance. Section V proposes an additional layer in Big data-IoT framework

II. Big Data Test Framework

The variety and volume of data have become a challenging aspect to databases. With unstructured, structured, semi-structured data being produced every second, data testing is extremely complex. The 4 V's Volume, velocity, variety, and veracity of big data demand the unorthodox form of information that enables magnified insight, decision-making. Big data testing is absolutely dissimilar from general testing scenarios as it involves processing huge data quickly for a business to make better decisions. The primary goal of big data testing is cleaning, masking, monitoring big data but none of these deals with

data validation in a big data framework which lacks the quality of data. Big data testing is verifying data to ensure data transformation, data quality, and automate the regression testing.

Validation of structured and unstructured data in a test environment increases cost and time. Big data testing is based on Extract, Transform and Load (ETL). In the Extract phase test data is uprooted from various sources, traditional databases like relational database management system (RDBMS), the test data and process are verified and in the transformation phase, once the transformation is successful, it is either sent to the data warehouse or deleted. Quality is a major issue and requires a peculiar infrastructure [2]. Data warehouse staging area is a short-term location where data from all sources are recorded. Since data cannot be extracted directly from all databases at the time, therefore, data in the data warehouse is momentary

Quality Assurance (QA) defines whether a product or service meets the specified requirements. Fig. 1 describes various parameters that could cause tangible and intangible losses to an organization due to poor data quality. Unreliable data leads to wastage of resources, business revenues, decisions, productivity, and prevents data from being shared in an organization. Meeting customer requirements is far beyond the reach if data is not validated and accurate. Due to unreliable systems, low-quality data collections, unorganized data, connectivity issues, technical faults between sensors lead to business loss. Data is said to be reliable and consistent when data collected and analyzed remains substantial over time. Data quality parameters, data accuracy, data timeliness, data accessibility, data accountability, data completeness, data scalability, and data security and their significance are discussed in detail in [1], [4].

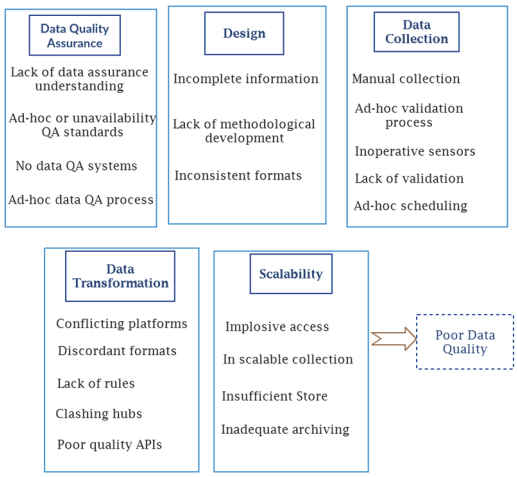

Fig. 1. Data quality concerns in big data environment.

To ensure the quality of data the following big data quality services are generically employed in a big data testing framework [1], [5], [6].

- Data collection: Gathering and quantifying information from various sources.
- Data cleaning: Since data is collected from various sources detecting and correcting untrustworthy, inaccurate, corrupt records data is a major role in big data testing which ensures data quality.
- Data transformation: Process of the transfiguration of dataset from a source data system to the format of a destination data system.
- Data loading: Once the data is transformed it is loaded into a big data repository such as NoSQL big database and Hadoop domain.
- Data analytics: Inspection, modeling, and modification of data into reports, conclusion, supports decision-making.
- Data aggregation: The arrangement of data from a database to develop datasets for data processing.

With the high computing requirement and complexities of the processes in the big data testing framework, test as service (TAAS) is gaining popularity in recent years. TAAS is primarily aimed at providing solutions regarding cost, data and packet loss, and scalability issues of IoT devices and test semantic correctness and functional features remotely [2]. TAAS with IoT testing framework rectifies unnecessary cost, traditional software testing in the development of IoT devices, provides real-world testing and reduces strain on internal resources. With emerging Machine learning methods into software testing [3], software, TAAS is becoming more and more relevant [3].

Existing comprehensive big data quality framework is primarily centered around the data coming from data warehouses, weblogs and social media. Though IoT is an inseparable component of today's big data application, Inclusion of IoT focused data validation is not yet seen as a mandatory element in the framework.

III. IoT Key Contributor of Data in Big Data Application

IoT enables things to actively participate in sharing data with other objects, communication over the network (wired/wireless), recognizing changes and events in other objects where things/object can react inaccurately.

The internet of things helps to connect anything with everything. IoT is connected to cellular services like 30% are phones, 23% tablets, and others are machine-to-machine communication. With the advancement of high-speed internet connection like Broadband connectivity, Google fiber which provides high-speed low latency network.

As shown in Fig. 2, it is projected that IoT will grow about 267 billion in 2020 [7]. IoT generates huge information, this information is analyzed, and resets factors based on the emergency. Sensors help to detect motion; a voice call may be sent through the internet or appropriate altars are sent on devices. With the advancement of technology and the use of sophisticated sensors, IoT generated data reduces human efforts and interaction and improves decision analytics. Real Time Data generated by IoT is highly preferred for decision-making because of its high business value.

IoT generated data is seldom analyzed independently and often exists as one component of the big data analytics ecosystem, Fig. 3. Big data and IoT is used widely across domains to provide diverse solutions. Big data analytics is used to examine huge datasets in order to uncover hidden patterns, customer requirements, market trends, business information, better agriculture planning, reduce the cost of

medical systems and decision-making. There are few domains where IoT and big data analytics has become the norm for the functioning of various processes. Health gadgets with various IoT enabled sensors are becoming the backbone of patient monitoring systems and providing phenomenal support to inefficient customer care [8], [9]. IoT devices are being used to monitor and build patient-centric, remote consultation, to help critical conditioned patients [10]. Smart farming includes technologies like IoT, big data, data mining, machine learning techniques, cloud computing which enables farmers to take actions and better-informed decisions on farming practices. Sensors are used on fields and crops which provides data points on soil conditions, detailed information on wind, water availability and pest infections [9]. Sensors like SHT10, SEN0161, Humidity sensor and Obstacle sensor (ultrasonic) are used on various hardware and software that includes AVR microcontroller atmega 16/32, ZigBee module, Raspberry pi, Dip trace, SinaProg, Raspbian Operating system. Thus, it is now possible to monitor productivity with just a click of a button. Smart homes technologies include a suit of IoT devices, appliances, or systems that connect into a network and can be controlled. IoT and big data fabricate the use of accommodating new devices, appliance, and other technologies. IoT is growing exponentially, Sophisticated sensors and chips are embedded into systems that surround us in a smart home environment which comprise of Temperature sensor, Voice/Sound sensors, an Air composition sensor, Infrared sensors, pressure sensors, Video cameras for surveillance. When an unusual motion takes place, an alert message is sent to the user [11], [12], [13], [14].

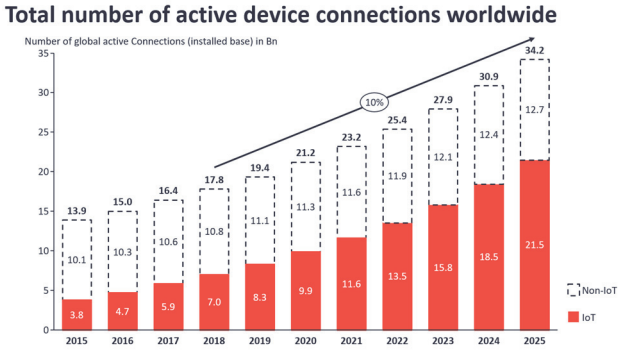

Fig. 2. Worldwide Diversification of IoT Devices, as projected by [7].

Thus, the amount of data generated by connected devices is tremendously huge. Its assimilation in a big data system is further complicated by the variety, time dependency, compatibility, and interpretability.

IV. Quality IoT Data: Challenges

IoT and big data analytics has almost become omnipresent and also brings data challenges along with it. A Huge number of sensors generating an enormously high volume of diverse data requires a multifaceted data quality assurance approach. In this section, we emphasize three main characteristics of data which are essential for producing valid and applicable results namely data reliability and accuracy, data timeliness and data interpretability. We discuss the challenges in ensuring these qualities in IoT data and review the state of art of the solutions provided for them.

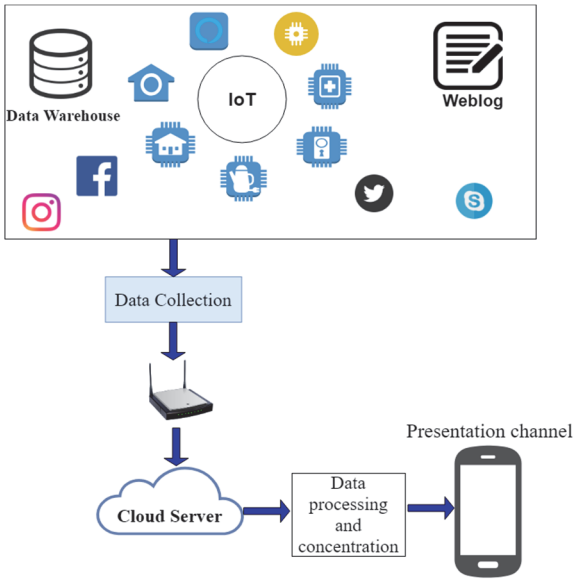

Fig. 3. IoT and Big Data Analytics.

A. *Reliable and Accurate Data – IoT Security*

Security and privacy of data are very crucial to the IoT paradigm. This undoubtedly is the most researched area in the field of IoT, cloud computing and big data because of its high impact on the business value of such systems. Though the solutions to IoT security are based in multiple domains like networks and machine learning, the primary objective is to collect genuine and authentic data. Securing systems is based on a few standard principles: confidentiality, availability, authentication, integrity. Some devices used in IoT have extremely limited storage, battery power, processing rate are unable to cope with the unique security systems and wireless networks are widely used in IoT devices which could lead to packet loss. Security is a widely researched problem in IoT and main security concerns are identified as Eavesdropping, Mac spoofing, Dictionary attack, and Man-in-the-middle attack. [14], [11]. While traditional solutions include encryption and cryptography, a newer research direction based on IoE, internet of entities with blockchain based validation mechanisms is being proposed in the research community [15]. In network security for smart home, domain is proposed in [11] where communication rules for every device are installed in every home router and are further used to filter malicious traffic. The layered architecture of IoT posed challenges in providing end to end privacy and security. Improved privacy preserving the architecture of IoT as proposed in [16] is the need of the hour which is based on the concept of using multiple cloud data stores for preserving privacy. Based on this generic architecture domain specific architecture for more secure data in IoT is also proposed. Application of machine and deep learning approaches for building robust IoT big data applications [5] are effectively used for threat categorization as well as predicting the layer where the threats can surface viz, network services surface/cloud service surface/web application interface, etc.

B. *Data Timelines – Real-Time Data Analytics Models*

With heterogeneous data coming continuously from multiple sources spanning multiple geographic locations, it's difficult to separate valuable data from irrelevant information. IoT big data analytics is further challenged by the need for real-time data updates and its real-time analytics due to the continuous operational state of IoT devices, thus a "Fog Computing" lightweight computing paradigm becomes relevant for IoT. Fog computing is similar to cloud computing which provides temporary storage, services, and application which provides a promising solution for big data applications and IoT. Fog computing is an intermediate layer between cloud computing and data generated from various sources. It reduces the processing time and cost spent on sending huge data to the cloud. As fog nodes analyze all the data that needs to be recorded and delivered into the cloud which is used for prediction or a historical purpose. Fog nodes provide optimization approach for an IoT sensing application which improves data security and reduces data latency, faster response. Fog nodes analyze data with minimum requirements like power and fewer resources by appending an appropriate sensing module. The performance level is reduced as data is uploaded into the fog nodes [17]. Fog computing in IoT can eliminate the dependency on a centralized data center and perform the in-network computation to reduce the latency in computations. This lightweight computation also augments security solutions as it allows lightweight encryption schemes through fog-to-things paradigms [18], [19]. Data generated by sensors and devices are processed efficiently and closer to where the data is originated instead of sending it to a diverse data center as is done by edge computing. A massive amount of data is collected and processed by edge devices locally, stores condemnatory data. Edge computing is closer to end users and provides Quality of Services (QoS) to end users. Edge computing nodes are also called edge/cloudlet servers. Edge servers reduce operating cost, provide real-time analysis, reduce network traffic and improve the performance of applications [20].

C. *Data Interpretability – Semantics of IoT Generated Big Data*

The three V's of big data volume, velocity, and variety are inherently applicable to IoT data. Before integrating this data with other non-IoT data for further analytics, high-level abstraction of the raw IoT data can improve the interpretability of the data. IoT requires algorithms that can analyze data that comes from a variety of sources in real-time. Semantic technologies tend to enhance the abstraction of IoT data through annotation algorithms [17]. The "variety" of IoT data encompasses time series data, streaming data, geographical data, data coming from wearable devices, etc. Providing insights based on these raw values requires a plethora of algorithms. Semantic technologies for interoperability on IoT are one of the latest research field in IoT [14], [21]. Due to the heterogeneity of devices and platforms in any big data and IoT framework, augmenting data with semantics that the data represents can add a very high value to the raw data that accumulates with a very high velocity. Recent paradigms like Resource Description Framework (RDF) are gaining popularity due to the flexibility that they provide in the continuous query processing [22]. Application of semantic annotations of IoT data in healthcare domain is discussed in [23]. The paper shows semantic annotations of the heterogeneous data gathered using IoT devices of patients and physicians to transform the data into RDF. This data is then processed by SPRARQL (SPARQL Protocol and RDF Query Language) facilitating the interoperability across devices. The concept of interoperability is very much relevant in all the domains of IoT and requires standardized data representation formats. These formats essentially describe data as linked objects or entities with characteristics and relationships. Example. Ontologies are required further for knowledge sharing to interpret the data representation [24]. Semantic interoperability can be challenging: integration of multiple data sources, a distinctive ontological point of reference, P2P (peer to peer) communication, semantic discovery of data sources and services. IoT interconnected devices face standardization and reusability issues due to unpredicted faults.

V. IoT Inclusive Quality Assurance Framework for Big Data with IoT

IoT has made a machine to machine communication possible. We propose an additional IoT quality assurance layer before IoT data is integrated with the generic big data application. As shown in Fig. 4, the proposed IoT data validation layer sits on top of the data collection layer. A series of actions proposed in the layer would ensure that the raw IoT data is transformed into suitable abstraction before getting integrated into any new-age analytics model.

As shown in Fig. 4 an IoT data quality validation layer can be included in Big-IoT framework immediately after data collection. Before integrating raw data collected from IoT devices, a series of transformation and quality checks in the proposed layer would facilitate further analysis of this data.

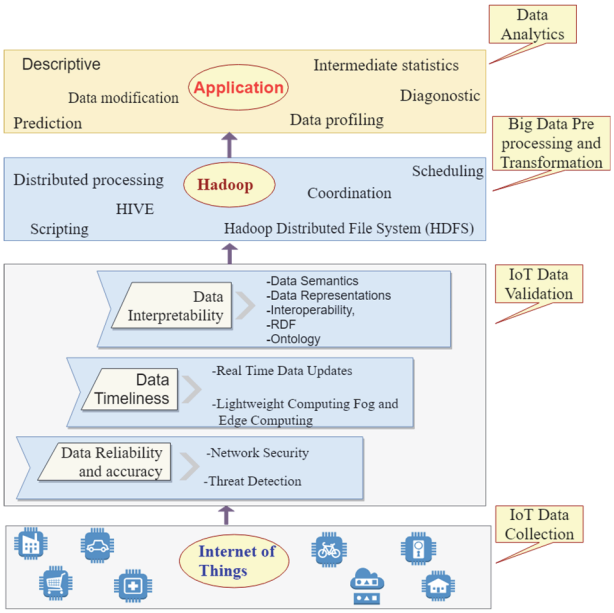

Fig. 4. IoT inclusive quality assurance framework.

Data accuracy and consistency, data timeliness and data usability are very important quality attributes and can affect the performance of an analytics application. Ascertaining these attributes for IoT data requires entirely different approaches and methods. Fig. 5 elaborates the difference between the data quality assurance methods with respect to IoT big data and non IoT big data applications for these above-mentioned quality attributes.

Thus, IoT data needs to undergo various transformations before its assimilation into a big data analytics framework. The data quality validation layer proposed in this study aims to encompass the features of IoT data quality listed in Fig. 5. Based on various processes and methods as mentioned transformations on raw IoT data are performed wherever necessary. Seamless implementation of measures discussed with respect to every challenge mentioned in the preceding section would assure the quality of IoT data which is the primary ingredient of any new-age analytics model. An IoT data validation workflow can be designed based on this proposed validation layer to ensure that the data is ready for integration with other data in the big data ecosystem. This validated IoT data can then be integrated with HDFS, HIVE or any other big data framework for further analysis and interpretation.

Big-IoT Data	Traditional Big Data
Data Accuracy and Consistency Features:-Uninterrupted Network Connections, IoT data Security Implementation Models:- Encryption , Internet of Entities, Block chain, Threat prediction using intelligent techniques **Data Timeliness** Features:- Real Time Data Analytics, Need for Low latency, Cannot be dependent on centralized systems for processing Implementation models:- Lightweight Computing, fog and Edge Computing **Data Usability** Features:-Data Abstraction, Semantically rich data representation for interpretability, Interoperability among heterogeneous devices Implementation models:-Ontology, Semantic annotations and Resource	**Data Accuracy and Consistency** Features:-Correctness of data-attribute pair, noise and outlier removal Implementation Models:- Statistical data cleansing methods, Functional dependency rules, data constraints **Data Timeliness** Features:- Up-to Date data, historic time bound data stored in big data repositories. Implementation models:- Data warehouse, Refresh Mechanisms **Data Usability** Features:-Data Integration, Transformation, Aggregation Implementation models:- Domain and task dependent data transformation workflows, Extract-Transform-Load (ETL) workflows

Fig. 5. Data quality assurance: IoT Big Data vs Traditional Big data.

VI. Conclusion

Data testing is a critically important phase in the development of big data application. IoT is a massive game changer in the modern world where sensors are the heart of IoT and big data. IoT and big data help to connect to devices to generate data to transmit, compile, and run analyses and predict and forecast new future. This paper is an effort to highlight various dimensions of the IoT data quality. The paper also highlights the requirement of a dedicated IoT data pre-processing and validation cycle for IoT data before its integration with other data in Big data IoT paradigm. Authors emphasize a smooth and continuous amalgamation of these additional processes for futuristic IoT big data applications.

References

[1] J. Gao, C. Xie and C. Tao, “Big Data Validation and Quality Assurance -- Issuses, Challenges, and Needs,” *2016 IEEE Symposium on Service-Oriented System Engineering (SOSE)*, Oxford, 2016, pp. 433-441.

[2] N. Elgendy and A. Elragal, “Big Data Analytics: A literature review paper,” P. Pemer (Ed): *ICDM* 2014, LNA 18557, PP.214-227, 2014.

[3] J. Gao, X. Bai, W. Tsai and T. Uehara, "Testing as a Service (TaaS) on Clouds," *2013 IEEE Seventh International Symposium on Service-Oriented System Engineering*, Redwood City, 2013, pp. 212-223.

[4] E. Ahmed *et al*., “The role of big data analytics in Internet of Things,” *Computer Networks*, vol. 129, Part 2, pp. 459-471, 2017.

[5] M. Gudipati, S. Rao, N. D. Mohan and N. K. Gajja, “Big data testing approach to overcome quality challenges,” Infosys publication, vol. 11, pp. 65-72, 2013.

[6] M. Mohammadi, A. Al-Fuqaha, S. Sorour and M. Guizani, “Deep Learning for IoT Big Data and Streaming Analytics: A Survey,” *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 2923-2960, Fourthquarter 2018.

[7] https://iot-analytics.com/state-of-the-iot-update-q1-q2-2018-number-of-iot-devices-now-7b.

[8] P. Verdugo, J. Salvachiua and G. Huecas, “An agile container-based approach to TaaS,” *2017 56th FITCE Congress*, Madrid, 2017, pp. 10-15.

[9] M. Hassanalieragh *et al*., “Health Monitoring and Management Using Internet-of-Things (IoT) Sensing with Cloud-Based Processing: Opportunities and Challenges,” *2015 IEEE International Conference on Services Computing*, New York, NY, 2015, pp. 285-292.

[10] H. Kim *et al*., “IoT-TaaS: Towards a Prospective IoT Testing Framework,” in *IEEE Access*, vol. 6, pp. 15480-15493, 2018.

[11] R. Kumar, *et al.*, “Monitoring system using android App”, *ARPN Journal of engineering and applied sciences*, vol 12, no 19, pp. 5647-5652, October 2017.

[12] C. Bekara, “Security Issues and Challenges for the IoT-based Smart Grid,” Procedia Computer Science, vol. 34, pp. 532-537, 2014.

[13] P. Bhardwaj *et al.*, “A review paper on smart home automation”, International Journal of Scientific Research and Management Studies (IJSRMS), vol. 3, no. 6 pp. 246-250, January 2017.

[14] Z. Khan, Z. Pervez, A. G. Abbasi, “Towards a secure service provisioning framework in a Smart city environment,” *Future Generation Computer Systems*, vol. 77, pp. 112-135, 2017.

[15] M. Sripan, X. X. Lin, P. Petchlorlean and M. Ketcham, “Research and thinking of smart technology,” *International conference on the system and electronic engineering*, December 18-19, 2012.

[16] R. Saia, “Internet of Entities (IoE): a Blockchain-based Distributed Paradigm to Security,” arXiv:1808.08809v1.

[17] A. Čolaković and M. Hadžialić, “Internet of Things (IoT): A review of enabling technologies, challenges, and open research issues,” *Computer Networks*, vol. 144, pp. 17-39, 2018.

[18] C. Mankar *et al*., “Internet of Things (IoT) an Evolution,” *International Journal of Computer Science and Mobile Computing*, vol. 5, no. 3, pp. 772-775, March 2016.

[19] G. Sabarmathi, R. Chinnaiyan, and V. Ilango, “Big Data Analytics Research Opportunities and ChallengesA Review,” *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 6, no. 10, pp. 227-231, October 2016.

[20] W. Yu *et al*., “A Survey on the Edge Computing for the Internet of Things,” in *IEEE Access*, vol. 6, pp. 6900-6919, 2018.

[21] C. Maple, “Security and privacy in the internet of things,” *Journal of Cyber Policy*, vol. 2, no. 2, pp. 155-184, 2017.

[22] S. Pacha, S. R. Murugan and R. Sethukarasi, “Semantic annotation of summarized sensor data stream for effective query processing,” *J Supercomput*, 2017.

[23] P. Murdock ed., “Semantic Interoperability for the web of Things,” DOI: 10.13140/RG2.2.25758.13122, August 2016.

[24] M. Harlamova, M. Kirikova and K. Sandkuhl. “A Survey on Challenges of Semantics Application in the Internet of Things Domain.” *Applied Computer Systems*, vol. 21, pp. 13-21, 2017.