Big Data & IOT : A Blended Approach

- **Ms. Nimisha Manan, Research Scholar, Deptt. of Mathematics, Patliputra University, Patna email id : nimisha.manan@gmail.com**

- **Dr. Krishnandan Prasad, Associate Professor, Deptt. of Mathematics, T.P.S. College, Patliputra University, Patna email id: knpd1962@gmail.com**

Abstract : *The rapid advancement of technology and communication made possible by the Internet has improved connectivity between various machines and sensor-based devices. The term "Internet of Things" (IoT) refers to the network of machines or other objects connected via the internet. The Internet of things is being used to connect a variety of wearable technology, including smartwatches, autos, home appliances like washing machines, doors, door locks, lighting, etc. Big data is generated daily in large amounts by these sensor devices. This information can be analysed to provide solutions to a variety of everyday issues. This paper examines various Big data tools and strategies that can be applied to IoT frameworks. It also demonstrated a method for using Big Data to analyse IoT data sets intelligently. The various Big-data analytics platforms are thoroughly explained, and it is made clear which one is appropriate for IoT data.*

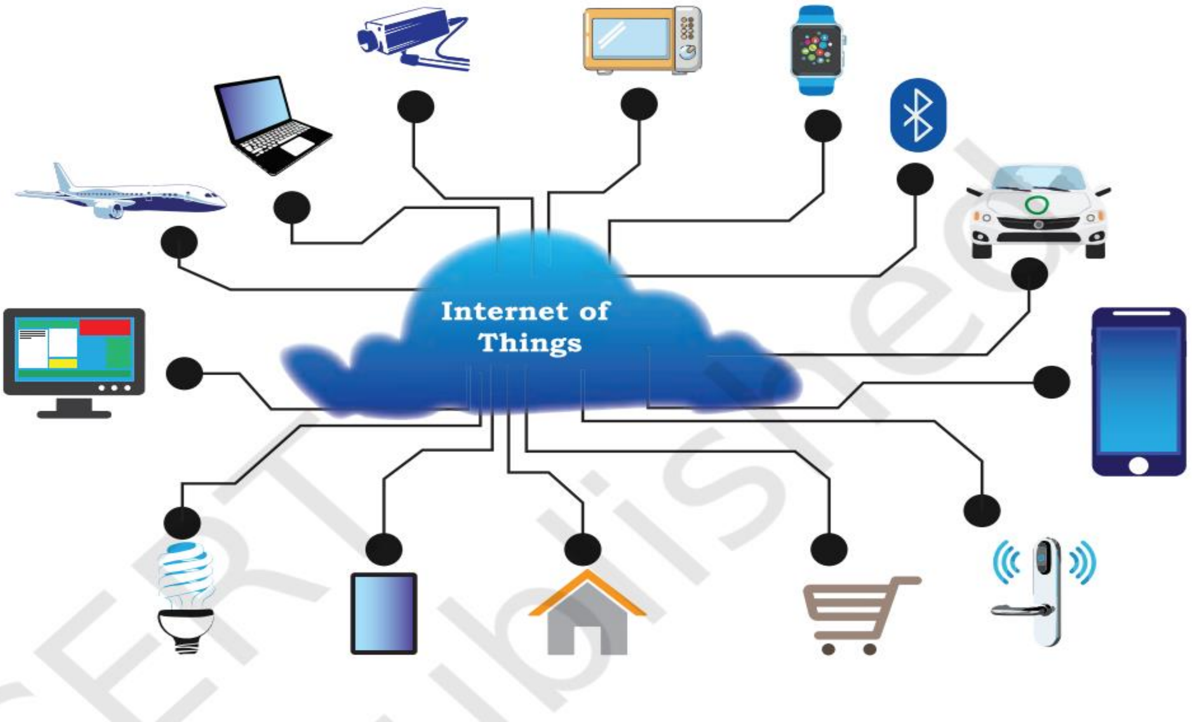
Keywords: *Big data, Frameworks, Internet of Things (IoT), Architecture, Big Data Analytics (BDA)*

Introduction

“In recent years, two rapidly developing technology areas have been the Internet of Things (IoT) and Big Data. Big Data and the Internet of Things (IoT) are complementary concepts. The IoT's basic premise is that nearly every thing /device will be assigned an IP address and be connected to the others. The effectiveness of data gathering mechanisms will now be evaluated in light of the billions of connected devices that will be creating enormous volumes of data is likely to face difficulties. The real-time or nearly real-time communication of data about the "linked things" is one of the IoT's key characteristics. The challenge is achieving this on a large scale (for example, between tens of thousands and tens of millions of objects). IoT has four key distinguishing characteristics:
a) large data size (TBs to PBs); b) high data flow, change, and processing speeds (OLTP, OLAP, and OLTP-like) and analytics c) A variety of structured and unstructured data, a variety of data models and query languages, a variety of data sources, and a variety of data veracity. The challenges of IoT with Big Data are discussed in this review paper, along with the requirements, technologies employed, problems with data security, challenges, etc.

Both Big-Data and IoT are growing quickly. All economic and technological fields are being impacted by this most recent development. IoT device data are crucial in the process of turning raw data into knowledge. Applying the proper big data analytics techniques to the raw data will enable you to do this. Volume, variety, and velocity are the three characteristics of Big Data that Gartner has identified. IoT gathers data in many formats and from various sources, which is why it is referred to as heterogeneous data. IoT can gather information from the healthcare sector, smart homes, smart traffic management, railroads, aeroplanes, weather forecasting systems, and agricultural. IoT data lacks structure and is randomly distributed. One can uncover a secret pattern, a fresh pattern, or both by using the appropriate big data analysis approaches.

IOT



The term "Internet of Things" (IoT) refers to a situation where anybody, anything, anytime, anywhere, any service, and any network are connected. Researchers offer many IoT definitions and architectures. IoT refers to a network of interconnected machines or things (computers, mechanical, or digital devices) that can connect these machines or things without any human intervention. It involves machine-to-machine (M2M) communication. IoT is defined and structured differently by researchers. IoT is a network of connected objects or machines (computers, mechanical devices, or digital gadgets) that may connect these objects or machines without any intervention from humans. It is an M2M (Machine to Machine) communication procedure.

A variety of architectural shapes have been suggested by various academics. The simplest IoT design, represented below, has three Layers and is labelled as follows:

Application Layer
Network Layer
Perception Layer

Figure 2 Three layered IoT architecture

- **Perception Layer:** This lowest layer is referred to as Perception Layer. It is employed to gather data.

- **Network Layer:** A connecting point between the application and perception layers is established using this intermediary layer.

- **Application Layer:** This layer delivers services and is utilised for processing data obtained from the previous two tiers.

Additionally, Fig. 3 depicts an alternative IoT architecture. Middleware and a gateway are added to the prior architecture in this design. The architecture has five layers. The layers are listed below.

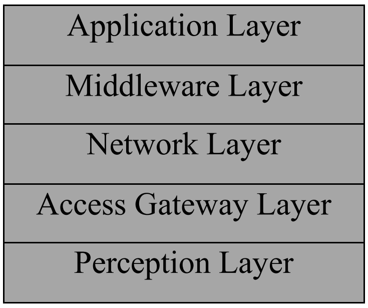


Figure 3 Five layered IoT architecture

- **Perception Layer:** The perception layer is often referred to as the edge layer.

- **Access Gateway Layer:** This layer controls the transmission of messages or data between IoT devices.

- **Network Layer:** This layer functions in a similar manner to the layer mentioned before. In IoT systems, it also aids in message transmission between sender and recipients.

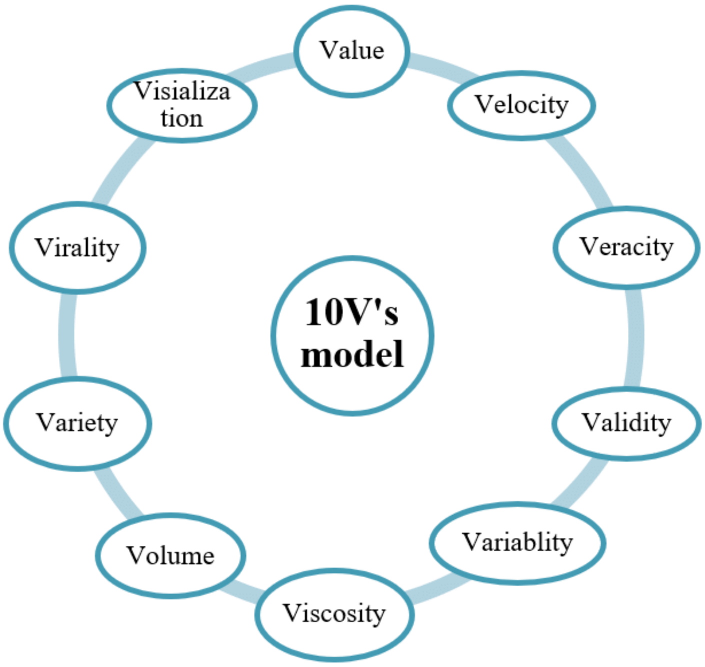
- **Middleware Layer:** This layer acts as a bridge between various software and hardware. It aids in forming a flexible alliance between hardware and its applications.

- **Application Layer:** In the three-layer architecture, this layer offers the same services as the third layer. It exists on top of all other levels. It is used to evaluate all the data provided by the layers behind it.

Big Data

Big data pertaining to businesses using internet services is now widely available. Examples include the hundreds of Petabytes (PB) of data handled by Google, the 10 PB of data logged by Facebook each month, the 10 PB of data analysed and processed by Baidu, and many more.

The Internet of Things (IoT) paradigm relies on sensors to gather and send data globally. These Sensors produce a growing amount of data, which tends to aggregate into a sizable heterogeneous dataset. This data must be processed and stored in a way that preserves its quality. Existing IT businesses must upgrade their designs and infrastructures in order to retain the volume and relationships of such enormous data. To disclose the intrinsic qualities of this heterogeneous data and enhance decision-making, new mining, analysing, modelling, visualising, and forecasting tools are required. For a detailed explanation of the term and definition, see big data. Let's examine the V’s model. Doug Laney, an analyst at META (now Gartner), introduced the 3V's model, which described the various opportunities and problems brought on by the massive volume of data produced by sensors. After then, IDC developed the four V's model in response to developments in the big data field in 2011. Further developments have allowed scientists to access 10V's of Big Data. We have the following in 10V's model:



- **Figure 4 10 V’s model of Big Data**

- **Volume:** The most important V in the V’s model is volume. Big data is described. Wide-ranging, diversified data are being generated as data production devices proliferate. Our conventional data processors and approaches cannot handle such a big volume of heterogeneous data.

- **Velocity:** The velocity of large incoming data from numerous devices is represented. In reality, this velocity is a crucial component of big data. Velocity refers to the rate at which different machines generate data through a network. Social media is one of the most prevalent examples of data creation speed. It generates a wide range of data. Everybody is now concerned with posting the hottest updates about oneself (on Twitter, Instagram, WhatsApp, etc.).

- **Variety:** According to the definition big data is a significant amount of diverse data. Therefore, big data's most important characteristic is variety. Today, there are many distinct types of data (structured, semi-structured, or unstructured) spread throughout data producing devices. Sometimes the format of the data gathered may differ from what is intended. The data processing could be complicated by this unexpected format. Any organisation that wants to avoid these issues needs a data storage system that can analyse and process any type of data, regardless of its structure.

- **Value:** Big Data is often produced as a result of continuous data generation. This information is useless until or unless it appears to be valuable. Therefore, the value of the data is undoubtedly a crucial component of big data. The useful data that various devices supply to the analyst or data scientist today is the foundation of big data analytics, which has now become a crucial component of society. Big data doesn't necessarily have to be valuable.

- **Veracity:** In this case, the amount of data is not relevant. It is a component of the easily comprehensible data that Big Data offers to its customers. The removal of "dirty data" is a best practise for any organisation handling a lot of data to prevent system buildup.

- **Validity:** Information needs to be exact and correct in order to be used in the future. If an organisation hopes to base future decisions on the data gathered by the devices, it should validate the data. Therefore, it is thought that validity is a crucial component of large data.

- **Variability:** This covers the accuracy and usefulness of the data. Velocity is thought to be a component of viscosity. It refers to the period of time between the sender and the receiver when sending or receiving data.

- **Viscosity:** Viscosity is considered as a part of velocity. It is used to describe the delay or lag-time which occurs between the sender and receiver during data transmission.

- **Virality:** It describes the data speed. This property has checks on the data speed with which sender and receiver access data from different devices.

- **Visualisation:** Big data is symbolically represented by this feature. Finding hidden patterns is made easier with visualisation. For any huge data query, these hidden patterns aid in decision-making. Big data plays an important role in decision-making thanks to visualisation.

Reliable software systems are necessary for managing such a large amount of data. In order to guarantee the software's quality, software testing is essential.

IOT AND BIG DATA INTEGRATION

Everything is integrated with technology in the modern way of living. IoT is expanding quickly across many industries. IoT comprises of devices that gather data, and these devices link with the outside world using this data. We can use this information to solve a variety of research problems, therefore it is helpful. Various big data analysis tools and methodologies can be useful for analysing this data. Big Data and IoT are seen as two sides of the same coin. IoT and big data analytics are related.

Relationship between Big Data Analytics and IoT

Due to the inclusion of multiple sensors and objects during data collection, IoT data differ significantly from normal data. IoT data is a heterogeneous type of data that is growing quickly and includes noise and variation. The number of data points generated by IoT devices reached 4.4 trillion by 2020. These tools can also gather real-time data, which changes every millisecond, look at it, send it, analyse it, and share it.

Big Data Analytics will play a crucial role in handling this redundant, diverse, and erratic data. Big data is used to store this enormous amount of data using a variety of storage methods and then analyse it to get specific results.

It has been determined from numerous studies that huge IoT data has three characteristics that prove its compatibility with the big data paradigm:

i. It is made up of numerous terminals, all of which produce a tonne of raw data.

ii. IoT device-generated raw data can take on any form, but is typically unstructured.

iii. IoT device generated raw-data is meaningless if not analysed.

Steps for IoT Big Data Processing

Four steps are generally used to manage IoT Big data, and they are listed below

i. The initial stage in managing IoT data sources is to manage IoT sensor devices, which have sensors that communicate with one another through various applications and produce highly unstructured, semi-structured, or structured data.

ii. The second step involves the collection and storage of Big IoT data, which is data produced by various IoT devices. This information is based on Gartner's 3V model. This IoT data is transformed into distributed and shareable Big data files in a big data storage system .

iii. Subsequently, it uses several analytical tools, like as Hadoop, MapReduce, Spark, and many others—more of which are covered in the following section—to analyse the data.

iv. The final step generates and displays to the user the report associated with the injected data.

Different Big Data Analytics Platforms for IO

Big Data Analytics requires certain tools and methods to convert IoT structured, semi-structured, and unstructured data into complete or metadata form for further analysis. These tools employ algorithms that look for patterns, correlations, and trends across different types of data.

1. Apache Hadoop

A platform that is open-source is Apache Hadoop. It serves as a storage facility for a substantial amount of raw data. It is capable of Big Data Analytics. Apache Hive, the Hadoop kernel, Map-Reduce, and HDFS (Hadoop Distributed File System) make up this common framework. Libraries in Hadoop make advantage of a straightforward programming model. The data is stored in HDFS, and it is distributedly processed by Map-Reduce. Data may be copied and spread over N distinct nodes thanks to the Map-Reduce architecture and HDFS combo.

The Master node and Slave node form the foundation of Hadoop. The master node assists in breaking the problem down into smaller problems. Then, various slave nodes are assigned to these subproblems. Following that, the output of all the slaves' sub-problems is gathered by the master nodes.

2. Apache Spark

Although it is just as open-source as Apache Hadoop, it is used to get around Map-Reduce's drawbacks, including as fault tolerance and linear scalability. It offers swiftness, usability, and comprehensive analytics. The analysis of graphs and ETL are combined by libraries. It offers in-the-moment analysis.

3. Dryad

For both parallel and distributed data sets, it functions as a data flow graph. Even if they don't know concurrent programming, a user can operate several machines at once. It is effective at managing cluster failures, creating graphs, allocating jobs to free computers when they become available, scheduling available machines that are available for allocation, etc.

4. Apache Drill

It is utilised in a distributed system for big IoT data analytics. It works with a variety of query languages. It can manage thousands of servers at once. Map-Reduce is used for analysis, and HDFS is used for storage.

5. Storm

Significant data processing is done with it. Real-time data is used, and this data should be distributed and fault-tolerant. In the same way as Hadoop clusters do, it creates a data cluster. Furthermore, it functions as a worker node and a master node.

6. Splunk

It combines cloud computing and big data. The user can search, analyse, and keep track of the data via a web interface. Indexing machine-generated structured and unstructured data is helpful. As a result, it is helpful for IoT Big data-sets. It is an intelligent system that supports the investigation of current, commercial data.

7. Jaspersoft

Real-time data analysis is accomplished with this open-source programme. Data from many platforms, including Mongo DB, Cassandra, and Redis, are visualised. It can produce effective HTML reports.

8. Apache Mahout

It is open-source data analytics software, meaning there is no need for a licence. It is employed for robotic learning. It is utilised to put several machine learning techniques into practise. It is used by large corporations, including Google, Yahoo, Amazon, IBM, Twitter, and Facebook, to construct scalable machine learning algorithms.

9. 1010 Data

It is composed of database columns. It works with semi-structured data. Large-scale infrastructure is supported by it. For loading, manipulating, and extracting data, it is not regarded as adequate. It offers high-end analytical services, such as statistical analysis and optimisation.

10. Cloudera Data Hub

It serves as a Data Hub for various businesses. It is used primarily for IoT-based data analytics and processing.

11. SAP-Hana

For massive IoT data analytics, it is used for in-memory addressing transactions. It provides answers to numerous massive unstructured IoT data problems. SAP-Hana includes libraries for R tool support, text analysis, and spatial processing. as the foundation for its analytical functions. It can serve as a focal point for comprehensive IoT-based data analysis. It offers security, excellent performance, and data access management. Because it lacks hardware, it must rely on a third party for processing.

12. HP-HAVEn

Hadoop Autonomy Vertica Enterprise (HAVEn), a new product from HP. This platform is used for Big IoT data analytics by several HP systems. It is used for analysing large amounts of data in a columnar database. Parallel processing is available.

13. Hortonworks

A platform built on Hadoop is called Hortonworks. It is utilised for big IoT data analytics. It is an upgraded version of Hive and is open-source software. It cannot reduce the number of nodes in a group.

14. Pivotal Big Data Suite

It is set up, tested, and put into use on a public cloud. It comes with a single licence. Pivotal aids in extremely parallel processing. It can perform predictive analytics on IoT data, but this data should be stored in HDFS.

15. Infobright

The analysis of machine-generated data, such as Internet of Things data, is appropriate for it. Up to 50 TB of data can be analysed simultaneously. It is compatible with large-scale data-based platforms like Hadoop. It is a columnarly constructed tool with the ability to automatically index data and skip rows.

Conclusion

The Internet of Things (IoT) has now grown to be a substantial source of Big Data, which is useless if not properly analysed. In relation to the Internet of Things, this study focuses on the Big Data scenario. It provides an overview of the IoT's architecture and fundamental principles. In the form of a 10 V's model, it provides a more detailed breakdown of the Gartner 3 V's model for big data. The relationship between IoT, Big Data, and Analytics is made easier to understand for the reader by this study. It acquaints the reader with several Big Data analytics tools that can handle varied IoT datasets. Readers will be able to choose the best platform for their unique challenges once they have read his paper and are aware of the various platforms.

REFERENCES

[1] M. Beyer, ``Gartner says solving `Big Data' challenge involves more than just managing volumes of data," Tech. Rep., AaltoDoc, Aalto Univ., 2011.

[2] R. Mital, J. Coughlin, and M. Canaday, ``Using big data technologies and analytics to predict sensor anomalies," in Proc. Adv. Maui Opt. Space Surveill. Technol. Conf., Sep. 2014, p. 84.

[3] N. Golchha, ``Big data-the information revolution," Int. J. Adv. Res., vol. 1, no. 12, pp. 791_794, 2015.

[4] Y. Wang, L. Kung, W. Y. C. Wang, and C. G. Cegielski, "An integrated big data analytics-enabled transformation model: Application to health care," Inf. Manage., vol. 55, no. 1, pp. 64–79, Jan. 2018.

[5] R. Khan, S. Khan, R. Zaheer & S. Khan, "Future Internet: The internet of things architecture, possible applications, and key challenges," In Proceedings of international conference on

frontiers of information technology, pp. 275-260, 2012.

[6] A. Ilapakurti, J. S. Vuppalapati, S. Kedari, S. Kedari, C. Chauhan, and C. Vuppalapati, “iDispenser #x2014; Big Data Enabled Intelligent Dispenser,” in 2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService), pp. 124–130, 2017.

[7] Y. Wang, L. Kung, and T. A. Byrd, “Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations,” Technol. Forecast. Soc. Change, vol. 126, pp. 3–13, Jan. 2018.

[8] M. Marjani, F. Nasaruddin, A. Gani, A. Karim, I.A.T. Hashem, A. Siddiqa, I. Yaqoob “Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges,” IEEE Access, vol. 5, pp. 5247–5261, 2017

[9] E. Ahmed, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Imran Khan, Abdelmuttlib Ibrahim Abdalla Ahmed, Muhammad Imran, Athanasios V. Vasilakos, “The role of big data analytics in Internet of Things,” Computer Networks, vol. 129, pp. 459–471, Dec. 2017.

[10] T. O. Center: Introducción a Hadoop y su ecosistema. http://www.ticout.com/blog/2013/04/02/introduccion-a-Hadoop-y-su-ecosistema/

[11] Acharjya, D.P., Ahmed, K., “A survey on Big Data analytics: challenges, open research issues, and tools.” in Int. J. Adv. Comput. Sci. Appl. Vol.7, issue 2, pp. No.- 511–518, 2016.

[12] F. Constante Nicolalde, F. Silva, B. Herrera, and A. Pereira, “Big Data Analytics in IoT: Challenges, Open Research Issues and Tools,” in Trends and Advances in Information Systems and Technologies, Cham, 2018, pp. 775–788.

[13] A. S. Foundation: Spark 0.8.0: This document gives a short overview of how Spark runs on clusters, to make it easier to understand the components involved, 2014, https://spark. apache.org/docs/0.8.0/cluster-overview.html

[14] V. Morabito, “Managing change for big data driven innovation,” in Big Data and Analytics. Springer, 2015, pp. 125–153.

[15] A. Bhardwaj, S. Bhattacherjee, A. Chavan, A. Deshpande, A. J. Elmore, S. Madden, and A. G. Parameswaran, “Datahub: Collaborative data science & dataset version management at scale,” arXiv preprint arXiv:1409.0798, 2014.

[16] F. Farber, S. K. Cha, J. Primsch, C. Bornh¨ovd, S. Sigg, and W. Lehner, “Sap hana database: data management for modern business applications,” ACM Sigmod Record, vol. 40, no. 4, pp. 45–51, 2012.

[17] S. Burke, “Hp haven big data platform is gaining partner momentum,” CRN [online] http://www. crn.com/news/applications-os/240161649, 2013.

[18] (2019, Accessed on 3rd December) Hortonworks. [Online]. Available: https://hortonworks.com/

[19] Y. Zhuang, Y.Wang, J. Shao, L. Chen, W. Lu, J. Sun, B.Wei, and J. Wu, “D-ocean: an unstructured data management system for data ocean environment,” Frontiers of Computer Science, vol. 10, no. 2, pp. 353–369, 2016. [Online]. Available: http://dx.doi.org/10.1007/s11704- 015-5045-6

[20] D. Slezak, P. Synak, J. Wr ´oblewski, and G. Toppin, “Infobright analytic database engine using rough sets and granular computing,” in Granular Computing (GrC), 2010 IEEE International Conference on. IEEE, 2010, pp. 432–437.

[21] Isard, M., Budiu, M., Yu, Y., Birrell, A., Fetterly, D. Dryad, “distributed data-parallel programs from sequential building blocks” in ACM SIGOPS Oper. Syst. Rev. 41, pp. No.- 59–72, 2007.

[22] Kelly, J.: Apache Drill Brings SQL-Like, Ad Hoc Query Capabilities to Big Data (2013). http://wikibon.org/wiki/v/Apache_Drill_Brings_SQL-like,_Ad_Hoc_Query_Capabilities_to_Big_Data

[23] C.L.P., Chen, C.Y. Zhang, “Data-intensive applications, challenges, techniques, and technologies: a survey on Big Data.” In Inf. Sci. 275, pp. no. -314–347, 2014.

[24] G. Ingersoll, “Introducing apache mahout: Scalable, commercial-friendly machine learning for building intelligent applications,” White Paper, IBM Developer Works, pp. no. - 1- 8, 2009.

[25] A. Verma, “Internet of Things and Big Data - Better Together,” Whizlabs Blog, 01-Aug-2018. [Online]. Available: https://www.whizlabs.com/blog/iot-and-big-data/. [Accessed: 11-Mar-2020].

[26] “Integrating IoT with Big Data, a Revolutionary Step,” Experfy Insights. [Online]. Available: https://www.experfy.com/blog/integrating-iot-with-big-data-a-revolutionary-step. [Accessed: 11-Mar2020].

[27] C.-W. Tsai, C.-F. Lai and A. V. Vasilakos, “Future Internet of Things: open issues and challenges,” Wireless Netw, vol. 20, no. 8, pp. 2201–2217, Nov. 2014, DOI: 10.1007/s11276-014- 0731-0.

[28] M. Chen, S. Mao, and Y. Liu, “Big Data: A Survey,” Mobile Netw Appl, vol. 19, no. 2, pp. 171–209, Apr. 2014, DOI: 10.1007/s11036-013-0489-0.