Industrial Engineering Journal
ISSN: 0970-2555
Volume : 50, Issue 12, No. 1, December : 2021

Terminology & Taxonomy for Big Data, Cloud Computing, and Internet of Things Integration

V. L. PADMALATHA
Assistant Professor
vlpadmalatha@gmail.com

A.SANDHYA RANI
Associate professor
sandhyarani1203@gmail.com

BALACHANDRA MEENUGA
Assistant Professor
balachandra.1202@gmail.com

Abstract: The advent of the Internet of Things (IoT) and Cloud Computing as a research area enables researchers with significant problems to address the challenges at the architectural level and network level for seamless connectivity and transformation of noisy data using analytics. This is important given the rapid digitization of data and the rapid advancement of technology. IoT integrates real-world items to connect with each other and send data over the internet, enabling the operation of cyber-physical systems. In the Internet of Things, sensor technology is crucial because it makes it possible to collect data from several sources and store it in the cloud. Additionally, it is a serious worry to execute data analytics on enormous amounts of noisy data in the cloud. In order to address the issue of combining IoT and cloud, this article offers a general architecture to do so. We also go through the language and taxonomy of each technology separately. The article also primarily focuses on the integration of IoT, Cloud Computing, and Big Data and how their widespread adoption will make them important components of the future.

Keywords: Internet of Things, Big data, Cloud Computing, Hadoop.

I. INTRODUCTION:

As technology advances, smart things that are linked to one another over the internet are significantly increasing the data creation. Big data refers to the vast amount of created data that is both organized and unstructured. As cloud computing develops, massive amounts of data may be stored using a pay-as-you-go basis. Big data analytics may be used to evaluate the large quantity of data to help people make better choices or anticipate how the world will develop in the future. Big data, cloud computing, and the internet of things have emerged as the most significant and well-liked paradigms for the creation of intelligent systems. The remainder of the essay is structured as follows. Big data and its taxonomy are discussed in Section II. The cloud and its categorization are shown in Section III. IoT is discussed in Section IV, along with its vocabulary and taxonomy. Big data, cloud computing, and the Internet of Things were integrated in Section V, and this issue was wrapped up in Section VI.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 50, Issue 12, No. 1, December : 2021

II. BIG DATA

Big data is an evolving term that refers to a large and complex set of data like structured and unstructured and it is difficult to process using traditional hardware and software techniques for data analysis. Big data includes thevoluminous amount of information. Some companies say small data is also big data while others say large set of data is big data. Fig. 1: depicts the classification of big data. Big Data We define data as big data when it exhibits the following characteristics:

1. Volume: Volume Refers to the large amount of data generated every second i.e., in Zettabytes, petabytes.

2. Variety: Variety refers to the large dataset may consist of a number of different formats, including spreadsheets, word processing documents, videos, photos, music clips, email/text messages and so on.

3. Velocity: Velocity refers to the speed at which data is gathered. The increasing technology leads to thegeneration of huge amount of data in a second.

4. Veracity: Big Data Veracity refers to the biases, noise, and abnormality in data.

5. Value: Value refers to make sure of getting valued data from big data.

A. Data Types:

Big data has produced varieties of large datasets from different sources contains different structures, scale, density and representation in different domains. The basic idea here is to understand the difference between big data and traditional data [8]. Some examples of big data are:

1. Spatial-Temporal Data: Rapid development of mobile devices, GIS Systems, computer vision applications, wireless systems, online streaming and many other processes a spatial-temporal data has been produced continuously at high speed.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 50, Issue 12, No. 1, December : 2021

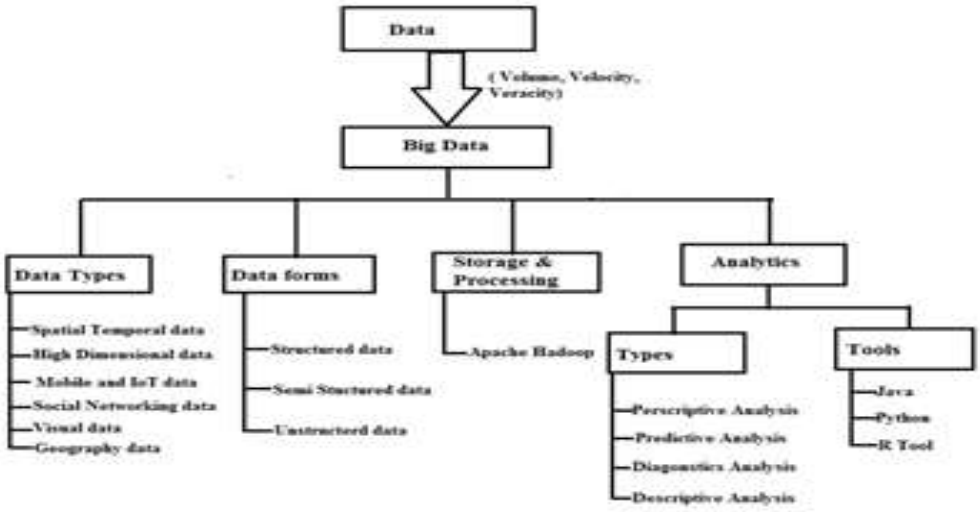

Fig 1: Classification of Big Data

2. Mobile and IoT Data: The data can be collected and exchanged by the network devices like wireless sensor networks, software and accumulators through theinternet. It provides machine to machine communication. Ex: smart grids, smart homes.

3. Social Networking Data: In the era of big data with the continuous development of social networks, the social data has increased extensively [5]. Social data is any type of data that can be collected from social media i.e., Facebook [2], Tweeter etc.

B. Data Forms:

By the different varieties of data, big data can be divided into three data forms. They are:

1.Structured data: The data is said to be structured when theschema is defined**.** The data can be stored in database SQL in table formwithrows and columns. This makeseasy to analyze the insight from large datasets.

2. Semi-Structured data: Semi-structured data isa datathatdoesn'treside in a SQL but it has somepropertiesthatmakeeasier to analyze. NoSQLdatabases are considered as semi-structured. Examples: Json, XML, CSV.

3. Unstructured data: Unstructured data does not have any schema to define. Unstructured data has text and multimedia. Examplesinclude e-mail messages, wordprocessing documents, videos, images, audio files, presentations, web pages etc. Examples: satellite images, social media data, etc.

C. Big Data Storage and processing:

In the era of big data, data originates different varieties of data from different sources. The main problem to be solved is how efficiently the data is to be stored in efficient storage infrastructure that caters all form of data in it and how efficiently the data is processed to get valued data from large datasets. There are many storage and processing models. Here,we address Hadoop because of its advantagesin storage and processing.

1.Apache Hadoop: Apache Hadoop is a framework that gives a scalable, flexible and reliable distributed computing and storage in cheaply and efficiently on a very large data sets for a cluster of systems. Hadoop follows a Master-Slave architecture for the storage, transformation, and analysis of huge datasets. The architecture of Hadoop is shown in Fig 2. In Hadoop architecture, there are two important components that play a vital role. They are

i. Hadoop Distributed File System (HDFS)
ii. Hadoop Map Reduce

i. Distributed storage: HDFS:

HDFS was designed in a distributed fashion and it run on commodity software. HDFS provides high fault tolerance that a file on HDFS is split into multiple blocks and each is stored across multiple machines to rescue the possible data losses. Hadoop Distributed File System (HDFS) stores the data of application and file system data separately on their allocated servers. In Hadoop HDFS, it uses master/slave architecture, that has two critical components namely NameNode and DataNode which plays crucial role. In this architecture, each cluster consists of single NameNode that manages the file system operations and DataNode that manages the data storage on each and every node in a cluster. In HDFS the data can be replicated into multiple DataNodes in order to get reliability.In Hadoop architecture, we use TCP protocol to communicate the DataNode and NameNode with each other. Hadoop architecture [6] works efficiently if it has high throughput hard drivers and high network speed to transfer data. In Hadoop the servers can be added or removed dynamically whenever it needed or any failure occurs. When the DataNode starts up, it informs to the NameNode along with its blocks of nodes that it is ready. When the DataNode goes down, it does not affect to remaining nodes of a cluster because it gets replicated one to perform the operations which were stored in DataNode.

ii.Data Processing: MapReduce

MapReduce is distributed computing or processing paradigm of large datasetsand it uses Java-based programming. It allows massive scalability across numbers of servers in a cluster. It allows parallel processing so it became popular in processing large datasets. MapReduce works in two phases i.e.,

Map phase (Splitting and Mapping): It takes set of data and converts it into another set of data, where elements are broken into tuples.
Reduce Phase (Shuffling and Reducing): It takes the output of map as input and merges similar tuples into asmaller set of tuples.

In Map function,a data is transformed into key-value pairs and then the keys are sorted where a reduce function is applied to merge the values based on the key into a single set. In MapReduce, the overall execution process was controlled by two elements i.e., JobTracker and TaskTracker.Execution of MapReduce starts when the job is submitted to Job Tracker that specifies the map. After getting the job, the job tracker splits based on the input path and select Task Tracker based on their network to the data sources and sends the request to that selected Task Trackers.The Task Tracker reads the data from splits. Map function is invoked which produces key-value pairs in the buffered memory. The memory buffer stores the produce key-value pairs into different reducer nodes by using the combine function.

D. Big Data Analytics:

Nowadays a big data is growing enormously that means varieties of data is being generated rapidly from different sources. To handle or analyze such large datasets was difficult to the traditional data analytics. So, Big data analyticscame into existence. Big data analytics uncover the hidden patterns or derive the insights from large data sets efficiently and effectively. It can help to make fast and better decisions and also it reduces the cost.

a. Big Data Analytics Types:

There are four types of big data analytics that will perform on raw data to get insights. They are

1.Prescriptive Analysis: The prescriptive analysis prescribes that *what action to take*and that make us to derive a solution. Before the decisions are made, this analysis will liberate the effect of future decisions. The main idea behind theprescriptive analysis is to **optimize production and** to make sure that are delivering the right products at in time.

2. Predictive Analysis: The predictive analysis tells "*what will happen".* It can be very useful for the user to know the future cause. This analysis willuse big data to identify the patterns to predict the future.

3. Diagnostics Analysis: The diagnostic analysis determines "*why something happened" and "why did it happen"*. Diagnostics analytics can be categorized by techniques such as drill down, data mining, data correlations and data discovery.

4. Descriptive Analysis: Descriptive analytics tells "*what happened".* This analytics can determine the insights from the raw data as to approach the future.

b. Big Data Analytical Tools:

With the rapid increase of data, to derive insights from the raw data is a challenge. Data analysis is used in almost all domains such as science, social and more. With the increasing need of analytics some tools are designed to analyses the data to get conclusions and other tools generate reports to sum up the conclusion for better data visualization. Data analytics can get accurate results with in the minimum time and lessefforts. [18] Some popular tools that are being currently used for data analytics have been discussed here comprehensively.

1.Java: Java is ascalable, robust and powerful platform, used to build applications that will run on any platform. Java has libraries, API, frameworks, Java Virtual Machine (JVM), the coding can simplify with java and at every level it supports development. Java is object-oriented, it was flexible and extensible.

2. Python: Pythonwas an open source scripting language. It emphasis productivity and code readability. Java libraries such as **Numpy**and **Matplotlib**enables the python to perform analysis. We can store process and analyze large data sets when we use python with Hadoop. In MapReduce, we use python to process the large data sets which are on Hadoop. Python can handle the data for parallel computations. Python is flexible and popular for statistical and data analysis.

3. R: R is a statisticalcomputing and graphics tool, which run on thecommandline like SAS, MATLAB. R is most popular among all data analytics, it executes knowledge analytics and produces graphs, charts, and tables. Rproduces good reporting, analysis, and visualization. R is amore advantages analytical tool than java python.

III. CLOUD COMPUTING:

Cloud computing, that we have learned about from past two decades. The concept behind is, it has inherited the traditional technology and adding new ideas. With the developments of distributed computing, grid computing, application service providers and virtualization, a new computing model came into existence called **cloud computing**. Fig 4. Shows the taxonomyof cloud computing. Before knowing the terminology and taxonomy of cloud computing [4], the evolution of cloud computing. Virtualization is a technology where it virtually creates a multiple devices or resources such as servers from a single system. The main aim of this computing is to make a better use of shared resources from anywhere at any time through their connected devices, in order to achieve high throughput and be able to tackle large computational problems. In virtualization, no elastic storage and computing. So cloud computing overcomes this issues.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 50, Issue 12, No. 1, December : 2021

Cloud computing is the on-demand delivery of computing power, database storage, and other resources through a cloud service platform via theinternet with thepay-as-go model. Cloud computing leverages dynamic resources to deliver a large number of services to its end users, low costs and simplicity to both users and providers. So, this section helps researchers or academia to understand cloud computing.

B. Cloud models:

Cloud computing is growing trend that is impacting all organizations. There are different types of cloud computing that we need to know before taking decisions. There are three types of cloud that organization cloud implement:

1.Public Cloud: Public cloud provides services over the internet and that is open for any type of customers like individuals, enterprises. Here, the data can be stored in provider's data center. Public cloud providers are Windows Azure by Microsoft, AWS by Amazon, AppEngine and Gmail by Google etc.

2. Private Cloud: Private cloud provides services for single organization or specific customers managed either themselves or by others. Here, the data can be stored in own data center so the data can be secured. IBM, HP, Microsoft are some examples of private cloud.

3. Hybrid Cloud: Hybrid cloud is a combination of public and private cloud or on-premise services. In hybrid cloud, it provides scalability like public cloud and security like private cloud. With the benefits derived from both deployment models, the hybrid cloud become more popular nowadays.

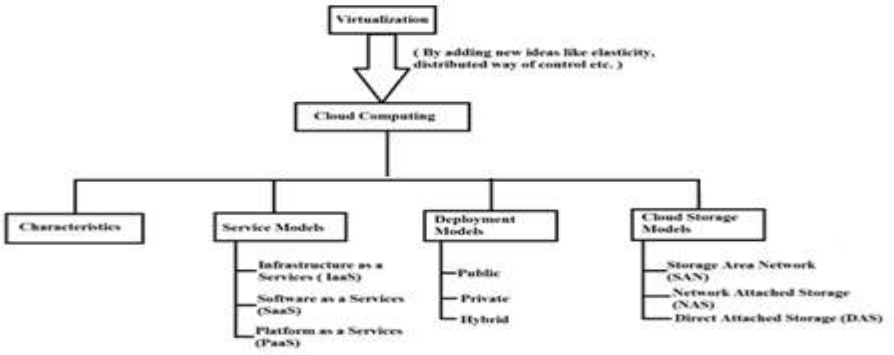

Fig 4: Classification of Cloud Computing

C. Cloud Services Models:

Cloud computing providers offers servicesaccording to different models, of which there are three different approaches of cloud services and these are utilized and consumed by several ways. The three service models are described as follows:

1.Infrastructure as a service(IaaS): Iaas provides computer infrastructure such as virtual machines, storage and networking. Instead of having to purchase software, servers, or network equipment, users can buy these as a fully outsourced service that is billed based on a number of resources consumed. Amazon EC2, Windows, Azure etc., are the examples.

2. Platform as a service(Paas): Paas provides computing platforms which typically includes an operating system, database and webservers. Paas is a framework they can build upon to develop applications using programming languages and deploys in cloud instead of buying hardware and software. AWS Elastic Beanstalk, Windows Azure, force. Com etc., are the examples.

3. Software as a service(SaaS): SaaS provides access to application services installed at a server without worrying about installation, maintenance or coding. We can access and operate through theinternet. Gmail, Salesforce, google docs etc., are the examples.

D. Cloud Storage:

Data storage, to store the data generally hard drives or pen drives are used. There is no guarantee of data security, when the hard drive crashes the data can be lost. To store large data sets reliably, cloud computing come into being. Cloud storage [11] is a computing model in which data is stored on servers and accessed through the internet. Examples are google drive, iCloud, one drive and Dropbox. Cloud storage improves disaster recovery, increases collaboration and save storage space. The cloud storage system cut up into three types based on whether the storage disk is attached directly or through internet.

1.Storage Area Network: Storage Area Network is a high-speed storage network that interconnects all the shared devices to multiple servers to reorganize the servers into independent and high-performance networks, it can easily access the storage device which is attached directly to the server which enables high bandwidth and low latency connections.

2. Direct Attached Storage:DAS is any block device, which is physically connected through the interface called “serial Advanced Technology Attachment” to a host machine. The block devices like hard disk, USB on host machine can be accessed by block numbers, the numbers are stored in file system on top of it in order to get easy access.

3. Network Attached Storage:NAS is accessed over the internet, which is ready to use and mount. NAS was maintained by third-party, they will charge for the usage based on capacity and bandwidth.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 50, Issue 12, No. 1, December : 2021

IV. INTERNET OF THINGS:

Nowadays, theinternet has become prominent in theworld and it has turned out to be ubiquitous. It influences human's life in manyway, so, "Internet of Things" has become more popular. The IoT [7] defined as a paradigm, in which the devices will communicate with other devices by using sensors, actuators, processors and transceivers etc.Fig 5 shows the classification of Internet of Things. IoT is not one technology, combination of other technologies so it is complex and it has several characteristics. Some of the key characteristics of IoT are connectivity, Dynamic Nature and Heterogeneity etc.

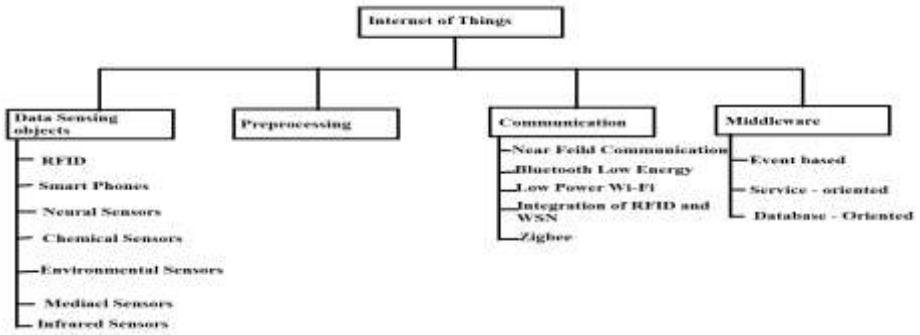

Fig 5: Classification of Internet of Things.

A. Data Sensing Objects:

This characteristic plays a major role because IoT is not possible without sensing. This helps to detect the changes and to generate the data and that data can be stored in remote servers for future use. There are various types of sensors to manageand get the data, some of them are described below:

1.MobilePhone-Based Sensor: with the rapidly increasing usage of smartphones, researchers get IoT solutions from smartphones by embedding sensors depending on the requirement. Some of the sensors are accelerometer, Global Position System and Proximity Sensors etc.

2. Neural Sensors: Neural Sensors detects the neural signals of brain and this technology is called asbrain computer interface. The neurons in brain communicate electronically and createselectric field, which can be measured in terms of various frequencies such as omega, beta, gamma, theta and delta.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 50, Issue 12, No. 1, December : 2021

3. RFID (Radio Frequency Identification): The RFID is attached to the object, which we want to read and the reader reads the data whenever the objects moves. These data which was collected by RFID is transformed into insights in IoT. RFID incorporates RFID tag, which has active and passive types. Active tag contains power source and passive have no power source, which gets power whenever electromagnetic waves are emitted from the reader. The main ideas is that it carries data and data can be read by RFID reader.

There are many other sensors like Medical Sensors, which can be used to measure or monitor the medical parameters of a human body to give frequent feedback to the doctors about the situation. Environmental Sensors measures the environmental parameters such as temperature, pressure and air etc. in the physical world to estimate the environmental changes. Chemical Sensors will detect the substances of chemical and bio-medical.

B. Preprocessing:

In IoT, smart objects collect the enormous amount of data through sensors and it can be stored in cloud to store, analyze and process the data and it provides scalability and flexibility. It won't be sufficient for the characteristics of IoT such as Mobility, Reliable and Power Constraints.To get rid of this problem a mobile cloud computing came out. The MCC also have a problem due to frequently changing of network. So, a concept called Fog Computing be developed, which brings storage and compute resources to the edge of network in which the data can be analyzed, stored and processed before sending it through expensive communication channel. Fog looks same as cloud, but it was nearer to ground. It offers low latency, immediate real-time response, mobility and location awareness.

C. Communication:

IoT is growing rapidly in the world, because of its advantages. IoT is Connecting smart objects which are independent of each other through the internet. IoT has minimum storage capacity because of it various communication challengesinvolved such as addressing and identification of devices, low power and with no data loss communication, and mobility of things. The smart objects or devices are connected through the IP (Internet Protocol), it requires large power and storage from the connected objects. Some devices can communicate through non-IP networks such as Bluetooth Low Energy (BLE), used for short-range communication and it consumes low energy when compared to others. BLE will transfer the small packets of data quickly.,ZigBee and NFC, which is used for short-range communication, in which the devices communicate with each other in centimeters only to transfer any type of data. Two NFC's will communicate with each other using magnetic field. with limited range.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 50, Issue 12, No. 1, December : 2021

V. INTEGRATION OF INTERNET OF THINGS (IOT), CLOUD COMPUTING AND BIG DATA:

In this era, the big data, cloud computing and internet of things became more popular individually. The internet of things enables communication between the "smart things" of a network each other. The things embedded with sensing objects like mobiles phones, RFID, sensors and actuators to accumulate the data from things. The IoT gateway accelerates the data which has been collected from a sensor, render across the sensor protocols and it

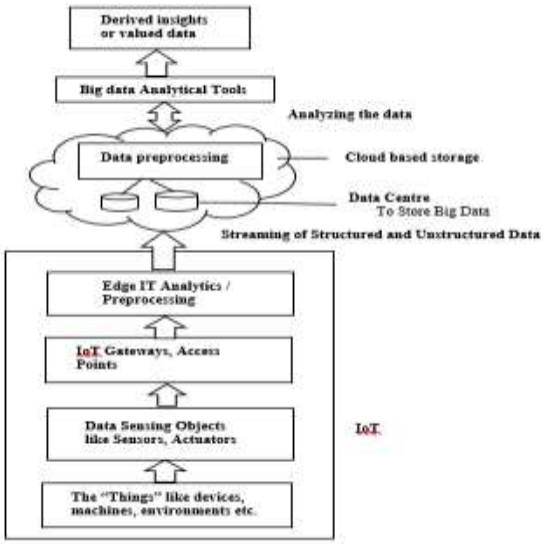

Fig 6: Integration of IoT, Cloud Computing and Big Data

processes the sensor data before it sends to the cloud. The main idea of IoT gateway is to serve as a bridge for both cloud and sensors to transfer the data in both ways i.e., from cloud to things or things to cloud. IoT gateway performs some major functions such as device connectivity, protocol translation, data filtering and processing, security, and updating etc. After building the connection, IoT performs data preprocessing techniques such as reducing high dimensionality, extracting and transforming of dimensional data, domain analysis to store clean data instead of storing raw data in cloud. This data which is taken from data preprocessing used to implement on analytical embedded system. The main sight of data preprocessing is to get predictive featuresof the data and process it, it will increase the power of analytics. Now, the data is stored in cloud. IoT generates a voluminous of data called big data, and that enormous amount of data will handle through cloud.

Cloud Computing provides pay-as-go model to the users to store and process the huge data [14]and it is cost-effective, provides scalability to analyze the big data. The data can be

analyzed by big data analytics to derive insights from the generated data. Internet of Things (IoT), Cloud Computing and Big Data are interrelated with each other. IoT is difficult without cloud and it is difficult without big data. the integration of Internet of Things, Cloud Computing and Big Data Analyticsgives more advantage to the users. By Data preprocessing, most of datahas been filtered and to store only clean data in cloud. Cloud computing is the only technology which can store, filter, process and analyze such large datasets and provides scalable and reliable data. Such huge amount of data can be generated through IoT, Analytical model such as Big Data analytics is used to derive valued data.

VI. CONCLUSION:

This study helps researchers to get a deeper understanding the terminologies and taxonomy involved in integrating IoT and Cloud Computing.Additionally, the generic architecture couldbe modified by addressing the internal layers involved in enabling seamless connectivity of data. At the architectural level, the identification of research challenges at fog Computing and edge analytics may be considered as thefuturescope ofresearch. Integration of these technologies enables smarter communities with better transactions.

REFERENCES:

[1] Z Lv, H song, P Basanta-Vai, Anthony steed, Minho Jo, “Next-Generation Big Data Analytics: State of the Art, Challenges, and Future Research Topics”, IEEE, 2017.
[2] Baratchi, Mitra, Nirvana Meratnia, and Paul JM Havinga. "On the use of mobility data for discovery and description of social ties." Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. ACM, 2013.
[3] BakshiRohit Prasad and Sonali Agarwal, “Comparative Study of Big Data Computing and Storage Tools: A Review”, Internation Journal of database theory and application, 2016.
[4] Bhaskar Prasad Rimal, Eunmi Choi, Ian Lamb, “A Taxonomy and Survey of Cloud Computing Systems”, 2009.
[5] Menon, A. (2012, September). Big data@ facebook. In Proceedings of the 2012 workshop on Management of Big data systems (pp. 31-32). ACM
[6] D Borthakur, “Hadoop Architecture guide”, 2008
[7] PallaviSethi and Smruti R. Sarangi, “Internet of Things: Architectures, Protocols, and Applications”, Journal of Electrical and Computer Engineering, 2017.
[8] Zheng, Yu. "Methodologies for cross-domain data fusion: An overview." IEEE transactions on Big Data 1.1 (2015): 16-34.
[9] Syed Ali Hassan, Sidra Shaheen Syed, Fatima Hussain, “Communication Technologies in IoT Networks”, Internet of Things building blocks and business models, pp: 13-26.
[10] Jyoti Sunil, and Chelpa Lingam. "Reality mining based on Social Network Analysis." In Communication, Information & Computing Technology (ICCICT), 2015 International Conference on, pp. 1-6. IEEE, 2015.

[11] C Wang, Q Wang, K Ren, N Cao, W Lou, "Toward Secure and Dependable Storage Services in Cloud Computing", 2010.