Lecture Notes Unit-01 Chapter-01 Lecture-01

1.1 Introduction

The talk of the town recently is the Digital Virtual Assistant. The virtual assistants are all soldiers in the tech world's latest battle in digitization. From Siri to Alexa and Google Assistant to Bixby, now they know more than what you know about you [1–3]. The digital life of a person now revolves around these assistants and they are running human beings instead of the other way around. These assistants are now embedded in the operating systems of smartphones reading our data without a person being aware and providing the right information at the right time to us. Sometime, we feel thrilled and magical, the other times; it seems that our privacy is at stake. It can be said that we assent computerized footprints in everything done that involves an advanced activity and these digital footprints accumulate and make huge data. Over this, the measure of machine-created information is quickly growing as well. This growing data is stored and used by the firms to grow their business and is termed as 'Big Data' Concept of 'Big Data' is continually evolving and being reexamined, as it remains the driving power under the umbrella of Artificial Intelligence. Information is created and distributed when our "smart" home devices contact within one another or to their home servers. Industry Machine in plants and processing units far and wide are increasingly furnished with sensors that assemble and transmit data. Be that as it may, what precisely is Big Data and how is it changing our reality, the chapter attempts to answer such questions that are there in everyone's mind. This chapter will provide a broad synopsis of Big Data for executives and managers seeking to implement new technologies in their business processes to avail the optimum usage of loads of data available to them [4, 5]. The chapter is also advantageous for researchers and academicians as it will help them to be abreast of the technologies and market requirements of today. The chapter includes the understanding of Big data together with the current applications for newcomers in the field. It also mentions the historical backdrop of Big data and the emergence of the same. The chapter takes you to the journey from antiquity through the adoption approaches used by the firms together with essential tools and technologies available to organize and manipulate Big data. Discussion on pitfalls and challenges while implementing and using technologies of big data is vital to present in the chapter. In the end, the job market for the current field is presented.

1.1.1 Understanding Big Data

Data is the heart of everything today. It is said that a person is rich if he has data. It is the past we look at, is the present we understand, and is the future to predict. According to the reports provided by various websites that give live stats of the happenings on the internet, shows some astonishing results that happen in a second. In a second around 46% of people are online on the internet. There are around 8093 tweets sent in a second, 853 photos are uploaded on Instagram, a million emails are sent, 74,226 videos are uploaded on youtube and the list goes on. The number of internet users has increased ten folds in the 21st century and 48.5% of users are in Asia. These surprising figures generate tons of data. According to IBM, 2.5 Quintilian bytes of data are made each day. The sources of information in the visible form are social networking platforms, computerized images, motion film, transactions, search engines, global positioning systems, etc. in form of numbers, pictures, videos, emails. This makes a data big and its collection, sorting, organizing, analyzing, and interpreting is termed as 'Big Data' [7, 8]. The data keeps on growing, but it's irrelevant to keep on collecting the data without its meaning. Therefore, extracting relevant real-time data and analyzing it to take out information according to the need of the study becomes vital. Big Data is now in its nascent phase, evolving and emerging itself. If asked, what big data is for you, the answers could be 'terabyte, gigabyte, petabyte, an exabyte of data 'or' anything beyond the human mind can think of'. The answers could be true but big data is much more than this.

If data would have been only one of these i.e. either volume (too large) or variety (mix a structured, unstructured, semi-structured data) or velocity (growing at a fast pace), it would have been relatively easy to hold it. However, data generated with a mix of all three V's complicate the scenario. Above this, the cost of managing the data increase tremendously. The varying veracity and value of data complicate the situation further. Data further generate more data. At a business level, the data makes a cycle and grows bigger and bigger.

Volume: Volume refers to the size or extent of data. As discussed earlier there are a variety of sources of data generation like social media platforms, digital pictures, videos, purchase transactions, search engines, global positioning systems, etc. in form of numbers, pictures, videos, emails, etc. Data is also generated by machines, networks, and human interaction and data itself generates more data. Therefore, the traditional systems of weighing the data in GigaBytes and Terabytes cannot be used these days because of the presence of an enormous volume of data. This data is so huge and is measured in Petabytes and exabytes and it's enthralling if one could analyze this huge data to get some useful insights. Within the advent of Artificial Intelligence and Machine Learning, this has been made possible by companies like Google, Facebook, LinkedIn, etc. to handle such huge data to extract meaningful information.

Velocity: Velocity is the rate at which data is generated and examined. If traditional data is like a bullock cart, big data is like a luxurious fast-paced car. Because of the various sources of data, it’s difficult to have control over such huge data. The reason for such fast-paced data generation is due to the freely available internet by mobile companies to capture the market, moreover, the increased speed of the internet is also another reason that generates and communicates data at a fast pace. There is a variety of digital devices available like smartphones; laptops etc. that even small retail firms are now able to generate data for their business. This data provides information about customers, like the timeline of their location, their buying pattern, their likes and dislikes, their demographics like age, family members, salary, etc. This enables firms to create a real-time analysis and capture the market [10]. *Variety*: Variety refers to structural heterogeneity in the data set. Big data is inclusive of all forms of data, for all kinds of functions, from all sources and devices. If traditional data were such as invoices and ledgers like a small store, big data is the biggest imaginable shopping mall that offers an unlimited variety [11]. The data can have a variety of forms of viz. Form as in texts, graphs, maps, etc., other is Function that defines the human conversations, songs, movies as in social media platforms and source data from machines like mobile phones, tablets, RFID, etc. like sensors from machines. When this data lacks a structure like in texts, images, or audios the data is termed as Unstructured Data or else Structured Data. Spanning a continuation between completely organized and unorganized data, the arrangement

of semi-organized data does not fit in with strict models.

Veracity: IBM framed Veracity as the fourth V, which epitomizes the inconsistency or uncertain nature of data. There is always a chance of error in understanding the pattern even when we predict or analyze the data because of the complex and uncertain nature of human beings or due to the chance of error in collecting data [12]. This misinformation or technical error must be rectified to be put to any great use and are dealt with big data analytical tools. All these four characteristics are not independent and work simultaneously; as per se volume of data may provide variety and vice versa. Volume may lead to veracity. According to the survey of Fortune 1000 companies by Harvard Business Review about the Big Data investments following results came out. The respondents were 1000 CEOs Presidents who were interviewed. Nearly one-half of all executives indicated that they have decreased expenses as a direct result of their investments in big data. Around 30% say that they have started to see the increase in revenue but the other 70% of the population has either started but not seen or not even started to see the revenue. Big transformations take time, and with the advent and further growth of Big data Technologies, it will initiate to give big returns also.

1.1.2 Applications of Big Data Analytics

Analysis of Big data is challenging and empowering. It gives businesses to investigate the insights to generate revenue by tapping the customer base. The New

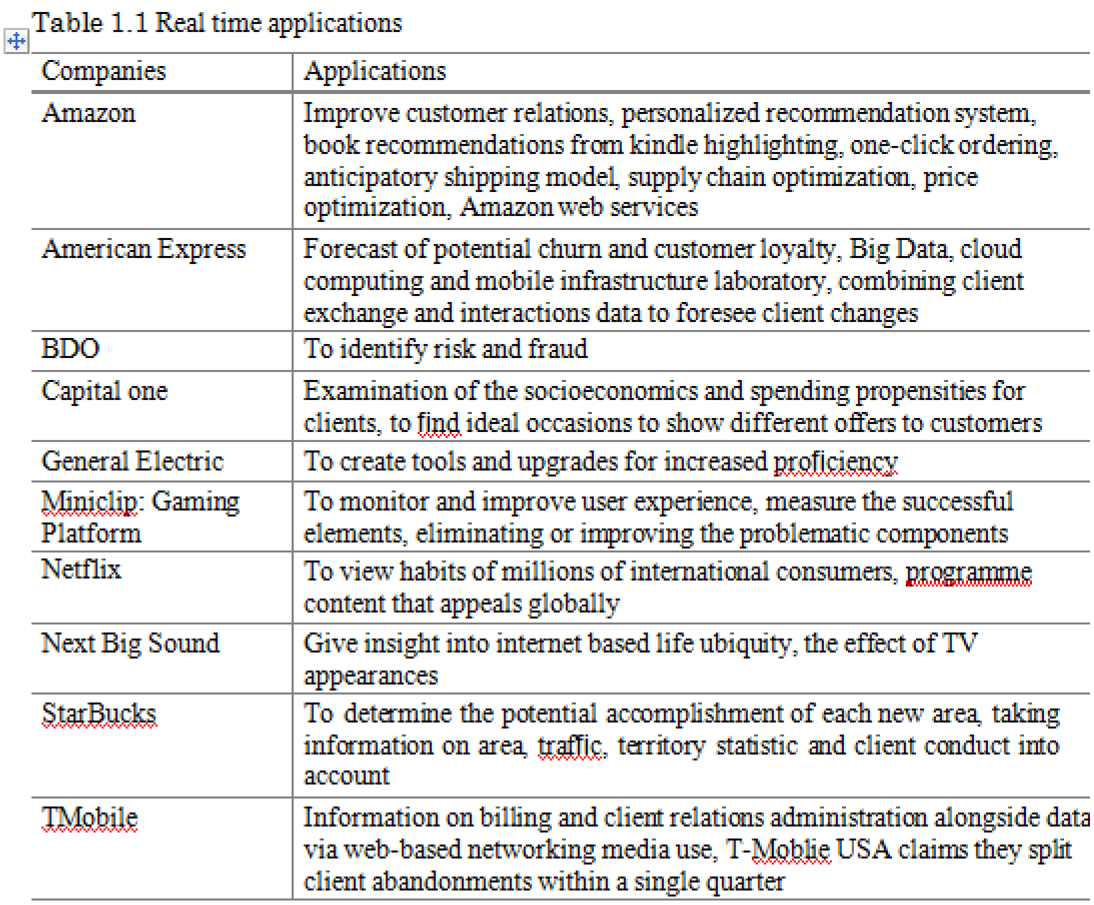

Table 1.1 Real time applications

Companies	Applications
Amazon	Improve customer relations, personalized recommendation system, book recommendations from kindle highlighting, one-click ordering, anticipatory shipping model, supply chain optimization, price optimization, Amazon web services
American Express	Forecast of potential churn and customer loyalty, Big Data, cloud computing and mobile infrastructure laboratory, combining client exchange and interactions data to foresee client changes
BDO	To identify risk and fraud
Capital one	Examination of the socioeconomics and spending propensities for clients, to find ideal occasions to show different offers to customers
General Electric	To create tools and upgrades for increased proficiency
Miniclip: Gaming Platform	To monitor and improve user experience, measure the successful elements, eliminating or improving the problematic components
Netflix	To view habits of millions of international consumers, programme content that appeals globally
Next Big Sound	Give insight into internet based life ubiquity, the effect of TV appearances
StarBucks	To determine the potential accomplishment of each new area, taking information on area, traffic, territory statistic and client conduct into account
TMobile	Information on billing and client relations administration alongside data via web-based networking media use, T-Moblie USA claims they split client abandonments within a single quarter

York Stock Exchange produces around 1 TB of new exchange data every day. Measurement demonstrates that 500 + TBs of new data gets ingested into the databases of web-based life webpage Facebook, consistently [13–15]. This data is mainly produced as far as to photograph and video transfers, message trades, putting remarks, and so on. A single Jet engine can produce 10 + TBs of data in 30 min of flight time. With numerous thousand flights for every day, the age of data comes to up to numerous PBS. The organizations can utilize this data further bolstering their good fortune; automating forms, gaining insight into their objective market, and improving by and large execution using the input promptly accessible. The following are some of the top-notch companies using big data for boosting their brand (Table 1.1).

Big Data is also used to understand the behavior of the customer and targeting them, accepting and optimizing business forms, individual capability and execution improvement, improvising healthcare, improving science and research, enhancing machine and gadget performance, improvising security and law performance, enhancing the traffic flows in and around the cities or countries and financial trading, Forecasting and acknowledging to normal and artificial catastrophes, Preventing crime to name a few more. And with the advent of new technologies, it will keep on giving promising results.

1.2 Emergence and Growth of Big Data Analytics

The evolution of IT, the internet, and globalization has facilitated enormous data and information. It in turn supported the discovery of Big Data. The story of Big data is not new and we may say that it's evolution started at as early as in the 1940s when a Wesleyan University Librarian observed that in every 16 years, libraries of American Universities are doubling in size and by 2040, the Yale University will have over 200 million volumes that would require more than 6000 miles of shelves and 6000 people to manage those shelves. In the 1960s, the author of Book 'Automatic data Compression' stated the "The 'information explosion' noted as of late makes it basic that the capacity necessities of every data are kept to least possible. A completely programmed and fast 3-section blower that could be utilized having "any" assortment of data to extraordinarily lessen moderate outer prerequisites and to expand the rate of data information transmission done by a PC is depicted in this paper [16, 17]. In 1970s Arthur Miller author of 'The Assault on Privacy' stated "Excessively numerous data handlers appear to measure a man by the number of bits of limit constrain his dossier will have." In 1980, Tjomsl said in his discussion of the IEEE symposium that "Those related with capacity gadgets long back understood that Parkinson's First Law might be summarized to portray industries data extends to fill space accessible". I trust such a huge data were kept on the grounds that clients have no chance to get of identifying out of date data; the punishments for storing out of date information were not much evidence that was the punishments for abandoning conceivably valuable information." Another research conducted in Hungary Central Statistics Office measured the volumes of information in bits in 1981. The turning point came when in the 1990s the Digital Storage became more economical than paper storage and challenges to store the data became apparent. The following data gives a summary of the progress of big data according to Forbes. According to Gartner, 72% of the organizations intend to increment the investment in big data investigation however 60% really expressed which were needed by individuals having qualitative profound scientific abilities. Only 22% of the information delivered was acceptable esteem, of that rarely 5% of the information was utilized from examination by trading. EMC examine gauges by 2020, 35% of the information created might take acceptable esteem.

According to IDC, the ending verge of 2020 trading activities on the internet i.e. B2B and B2C activities might surpass 450 billion consistently. The number of IT administrators tracing the information development might surplus by 1.5 times. There may be an interest of 75 additional documents and 10 additional servers. End verge of 2020, 1/third of the information created in this advanced world might live on or go within the cloud. Gartner estimated around 4.9 billion items might be connected through the Internet in ending the verge of 2015 that is required to achieve 25 billion by 2020. Human Genome deciphers prior took 10 years to date within the appearance of big data examination. In 2015, Google turned into the biggest big data organization on the planet which stores 10 billion GBs of information and procedures around 3.5 billion demands each day. Amazon is the organization with the highest amount of servers—the 1,000,000,000 GBs of big data delivered by Amazon and its 152 million clients have put away on in excess of 1,400,000 servers in different information focuses. By the beginning of an advanced period established on information totally, openings in the big data business for PC software engineers, clients, business visionaries, and other IT persons were developing as big data develops tremendously. Organizations that send a troublesome trade demonstrate their strong spotlight on information might be "The Next Big Thing" within the industry. According to the reports of CSC, in 2020, that is not too far, the scenario will look like the Story of Big Data.

Reference:

https://hbr.org/2017/04/how-companies-say-theyre-using-big-data

Maheshwari, A.: Big Data. Mc Graw Hill, New Yor (2017)

Marron, B.A., de Maine, P.A.D.: Automatic data compression. Commun. ACM 10(11), 711–715 (1967)4.https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/ #9e0827d65a18

Sathi, A.: Implementation section (book 1). In: Big Data Analytics: Disruptive Technologies for Changing the Game, 1st ed. MC Press Online (2012)

Job Market: https://www.simplilearn.com/big-data-applications-in-industries-articl

.https://www.simplilearn.com/how-facebook-is-using-big-data-article?source=CTAexp

https://www.icas.com/ca-today-news/10-companies-using-big-data

https://www.bernardmarr.com/default.asp?contentID=1076

Lyman, P., Hal, R.: Varian at UC Berkeley Publish. How Much Information?

.Bryant, R.E., Katz, R.H., Lazowska, E.D.: Big-Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science, and Society

Bohn, R.E., Short, J.E.: How much information? 2009 Report on American Consumers (2009)

Cukier, K.: The economist. A Special Report 2010. Data, data everywhere 14.Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: The McKinsey Global Institute 2011. In: Big Data: The Next Frontier for Innovation, Competition, and Productivity.

Neuman, W.R., Park, Y. J., Panek, E.: Info capacity. Tracking the flow of information into the home (PDF). Int. J. Commun. (2012)

Boyd, D., Crawford, K.: Critical questions for big data. Inf. Commun. Soc.

http://www.wheelhousesearch.com/wp-content/uploads/2014/02/2014-02-19_093536.png

https://www.bernardmarr.com/default.asp?contentID=766

Lecture Notes Unit-01 Chapter-01 Lecture-02

Adoption Approaches

This section provides pointers of execution plans and related challenges while adopting big data analytics. Adoption of big data analytics can provide firms various advantages like it can diminish dormancy by a request of size, providing openness to data in minutes or seconds rather than hours or days, it increases the ability to store data by a request of extent, moving from TBs to PBs, it offers a much lower cost of acquisition and operation, the cost is reduced by an order of magnitude as it requires fewer administrators. There are two approaches to adopting Big Data Analytics into processes viz. the revolutionary approach and the evolutionary or hybrid approach (Fig.).

The revolutionary method includes making a fresh out of the box new Big Data Analytics environment. We move every one of the data to the new environment, and all reporting, modeling, and integration with business forms occur in the new environment. This approach has been received by numerous Greenfield investigation driven organizations. They put their expansive stockpiling in the Hadoop environment and construct an investigation engine on the highest point of that environment to perform orchestration. The conversation layer utilizes the orchestration layer and integrates the outcomes with client-facing forms. The put away data can be investigated using Big Data instruments. This approach has given stunning execution however have required high tooling expenses and abilities.

In a normal evolutionary method, Big Data turns into an input to the present BI stage. The data is amassed and broke down using organized and unstructured instruments, and the outcomes are sent to the data distribution center. Standard modeling and reporting instruments currently approach online networking estimations, use records, and other prepared Big Data things. Ordinarily, this approach requires sampling and processing Big Data to hold the stockroom from the gigantic volumes. The evolutionary approach has been received by developing BI organizations. The design has an ease section edge and in addition minimal effect on the BI organization; however it can't give the noteworthy upgrades seen by the Greenfield administrators. In many cases, the kind of examination and the general end-to-end speed is constrained by the BI environment. This approach advanced effectively by IBM's Information Agenda group puts the AAP design over existing BI infrastructure. All the Big Data courses through AAP, while conventional sources continue to give data to the data stockroom. We build up several integration points to bring data from the distribution center into the examination engine, which would be seen by the data stockroom as a data store. An example of the AAP data would be coordinated back to the data distribution center, while the vast majority of the data would be put away using a Hadoop stockpiling stage for disclosure. The half breed design gives the better of the two universes; it empowers the present BI environment to function as previously while siphoning the data to the AAP engineering for low-inactivity investigation. Depending on the transition achievement and the capacity to advance abilities, the crossover approach gives an important transition to full conversion. Both the revolutionary and the hybrid architectures significantly challenge the data governance function.

REVOLUTIONARY APPROACH
BIG DATA APPROACH
EVOLUTIONARY APPROACH

1.1 Implementation Challenges

This section describes the new set of issues and how to handle them. There are three broad categories of questions that require attention while implementation.

Merging internal and external data: Due to the access of more and more data, it's now possible to firms to tap on consumer's likes and dislikes their demographics and tastes, and commerce needs. It becomes necessary to understand that this information should be fully utilized by organizations to grow their businesses and for that they should merge this data with already present information with them. While merging this newly acquired data, they should be careful and closely monitor how the data is being used and how it is being aggregated. All this occurs as we radically change the rules on data privacy, redefine MDM, and encounter new concerns relating to data quality.

Big Data veracity: Customer data comes from a variety of "biased" samples with different levels of data quality. It is vital to homogenize this data for optimum usage. As it is homogenized, we must establish confidence levels on raw data, as well as aggregations and inferences, in order to understand and remind users of the "biases "built into the sourced data.

Information lifecycle management: This is a lot more data than we have ever encountered before. Our current analytics systems are not capable of ingesting, storing, and analyzing these volumes at the required velocities. How do we store, analyze, and use this data in real-time or near real-time, this is a lot more data than we have ever encountered before. Our current analytics systems are not capable of ingesting, storing, and analyzing these volumes at the required velocities. We may decide to store only samples of the data or use Hadoop for the storage and retrieval of large volumes of unstructured data.

1.2 Big Data Under Considerations

Big Data provides with exceptional ideas and openings, yet it additionally brings up concerns and issues that must be tended to:

Data protection: The Big Data currently created having a considerable measure of data of our private lives, most of them having a privilege to keep personal. Expanding, that were requested to keep harmony among the measure of private information we reveal and the continuance of Big Data-controlled applications and administrations works.

Data safety: Even in the event that we choose we are glad for someone to have our data for a specific reason, would we be able to confide in them to protect it.

Data intolerance: When everything is known, will it turn out to be satisfactory to discriminate against individuals in view of data we have on their lives. We as of now utilize acclaim keeping count to choose which person could acquire funds and coverage is intensely information oriented. We may hope to be investigated and evaluated in more prominent details and responsibility essentially be taken that is not done in the kind that is creating life much troublesome for the individuals who as of now has fewer assets and approaches to data. Overcoming these difficulties is an essential piece of Big Data, and that should be tended to by organizations that need to exploit data. The inability to do as such can leave businesses helpless as far as their reputation, as well as lawfully and financially.

1.3 Big Data Market

According to a CRN Report of 2018, the following would be the companies recommended working for in the Big Data field. The full list can be seen in Industry persons, academia and other important partners mainly concur that big data have turned into a huge distinct advantage in maximums, if not every, kinds of present-day organizations in the course of the most recent couple of years. As big data persists to pervade one's everyday lives, that has been a critical move of the center from the publicity surrounding it to finding genuine incentive in its utilization. While understanding the estimation of big data continues to remain a test, other pragmatic difficulties including funding and rate of profitability and abilities continue to remain at the forefront for various diverse industries that are adopting big data. So, a Gartner Survey for 2015 shows that over 75% of organizations are investing or are planning to invest in big data in the following two years. These findings speak to a huge increase from a comparable overview done in 2012 which indicated that 58% of organizations invested or were planning to invest in big data within the following 2 years. By and large, most organizations have a few objectives for adopting big data ventures. While the essential objective for most organizations is to improve client encounters, different objectives include cost reduction, better focused on marketing, and making existing procedures more proficient. As of late, data ruptures have likewise made upgraded security a critical objective that big data ventures look to incorporate.

Big data and its application in various industries will enable you to more readily acknowledge what your part is or what it is probably going to be later on, in your industry or crosswise over various industries.

Banking and Securities: An investigation of 16 extends in 10 top investment and retail banks show that the difficulties in this industry include: securities misrepresentation early warning, tick examination, card extortion detection, authentic of review trails, enterprise credit hazard reporting, exchange deniability, client data transformation, social examination for trading, IT operations investigation, and IT strategy consistency investigation, among others. Big Data Technologies is heavily used by banks to “Know Your Customer” and fraud mitigation.

Health care Providers: The healthcare area approaches gigantic measures of data yet has been tormented by disappointments in utilizing the data to check the cost of rising medicinal services and by inefficient systems that smother quicker and better human services benefits no matter how you look at it. This is mainly because of the way that electronic data is inaccessible, inadequate, or unusable. Additionally, the social insurance databases that hold well being-related information have made it hard to link data that can show designs helpful in the medicinal field.

Different difficulties identified with big data include the exclusion of patients from the decision-making process and the utilization of data from various promptly accessible sensors.

Education: From an educational point of view, a noteworthy test in the education industry is to incorporate big data from various sources and sellers and to use it on stages that were not intended for the varying data. From a pragmatic point of view, staff and institutions need to take in the new data administration and examination instruments. On the specialized side, there are difficulties to integrate data from various sources, on various stages, and from various sellers that were not intended to work with one another. Politically, an issue of security and personal data protection related to big data utilized for educational designs is a test. Colleges everywhere throughout the world are using Big Data to assess the execution of educators and understudies.

Manufacturing and Natural Resources: Increasing interest for regular assets including oil, farming items, minerals, gas, metals, etcetera has prompted an increase in the volume, many-sided quality, and speed of data that is a test to deal with. So also, extensive volumes of data from the manufacturing industry are undiscovered. The underutilization of this information counteracts the enhanced nature of items, vitality proficiency, unwavering quality, and better net revenues.

Big data has likewise been utilized in solving the present manufacturing challenges and to gain upper hand among different advantages. In the realistic underneath, an investigation by Deloitte shows the utilization of inventory network abilities from big data at present in utilizing and their normal use later on. Apart from the above applications, Big data has its applications in other industries like insurance, transportation, retail and wholesale trade, energy utilities, etc. A few applications of big data by governments, private organizations and individuals include Governments utilization of big data: traffic control, course planning, intelligent transport systems, congestion administration (by predicting traffic conditions), Private part utilization of big, Private part utilization of big data in transport: income administration, mechanical improvements, coordination's and for upper hand (by consolidating shipments and optimizing cargo development) and Individual utilization of big data include: course planning to save money on fuel and time, for movement game plans in tourism and so on.

Lecture Notes Unit-01 Chapter-02 Lecture-01

2.1 Introduction to Big Data, IoT and Hadoop Ecosystem

In current scenario, huge volume of data are produced from several sources viz., health care, government, finance, marketing, social media etc. In this context, developing the Big Data applications has become crucial in recent years and Big Data mechanization have become as a relevant data analysis method for whisking the intelligence within the IoT infrastructure for preferably meeting the aspiration of IoT system. The Big Data has been divided as per the five basic ingredients: volume, variety, velocity, veracity and value. Out of these constituents, volume refers to the size of data, variety represents various types of data from distinct sources, velocity means the data gathered in absolute time. Velocity refers to the ambiguity of the data concerned, value means the beneficial aspects in various industrial and academic fields.

The combination of Big Data with IoT technology have constituted privilege leading to the enhancement of several services aimed for various complex systems such as smart cities, etc.

Various Big Data technologies have evolved for assisting in the processing of huge volume of data that are gathered from distinct sources in intelligent environments. On the contrary, Chen et al. described that the peregrination of IoT and its applications in several domains tend to increase the vast amount of several types of data gathered from different environments. Hence, the combination of Big Data and IoT evolution lead to novel investigative challenges that have still not been known or consigned through the probing association. Basically, as per Akatyev and James IoT has changed the way that technology and human interact. The work done by Rajan et al.details about the IoT generated Big Data employing semantics. Similarly, Khan and Solah detailed the block chain solutions and various open challenges in IoT security. Likewise, Ge et al. surveyed on Big data for Internet of Things (IoT).

Likewise, Hadoop represents an open source processing framework which manages processing of data as well as storage for Big Data implementations. Basically, it is the center of increasing ecosystem of the Big Data technologies for supporting modern analytics initiatives incorporating predictive analytics, machine learning and data mining application prospects. The Hadoop is capable of taking care of several forms of unstructured as well as structured data, thus providing users better flexibility for gathering, processing as well as examining data collected than the data warehouses or relational databases afford.

Basically, Hadoop is made up of modules aimed for carrying out specific tasks.

Following are the four modules associated with Hadoop system:

Distributed File System Module: The dispersed file organization permits data to be contained in a conveniently operable form through huge count of coupled stashed appliances. Moreover, a file organization represents the technique employed through a computer for storing the data. Generally, this gets indicated by the operating system of the corresponding computer.

Map Reduce Module: This module has been named as per the two funda- mental operations it carries out such as: reading data from the database and placing that in suitable format for the analysis purpose as well as carrying out analytical operations

Hadoop Common: It affords the tool desired for the computer of the intended user. Another task is reading data stored beneath the Hadoop file organization

Hadoop Yarn Module: It is the last segment which manages the system resources preserving the data as well as executing the analysis

Different other libraries, functions or features have emerged to be taken into account as component of the Hadoop through ongoing years, however, Hadoop Distributed File organization, Hadoop MapReduce, Hadoop Common and Hadoop YARN represent the major ones.

2.2 Big Data Process

The several Big data technologies incorporate distinct activities, techniques and methods deployed for several purposes. This section examines various existing literature on Big Data processes and considers various activities used for classifying the Big Data approaches employed in IoT. Gandomi and Haider get segregated the Big Data processes under two major phases: (a) data management and (b) data analytics.

Out of these two, data management is aimed at gathering data, storing, cleaning and retrieving it for the process of analysis and preparation. On the contrary, the data analytics is concerned with wresting intuition from the desired data that considers sculpting, interpretation and examination. Pakkala and Paakkonen proffered a Big Data allusion planning which cogitates the data origination as well as data repository as the input and infrastructure for Big Data processes.

The Big Data process involves the following major phases, namely,

Data Extraction
Data Loading
Preprocessing
Data processing
Data Analysis
Data Transformation and Data Visualization

The implication into the transitional organization is hence pursued by data metamorphosis as well as the possible inclusion of meta data before exporting to any other stage in the work flow.

Data Loading: The data loading represents the process that involves taking the transformed data and loading it where the users can access it.

Preprocessing: Data preprocessing is fundamentally a data mining approach which transforms raw data into more recognizable form.Practical data is mostly conflicting, deficient or lack in many features. Real world data may contain many errors. Therefore, for resolving these issues, data preprocessing is a proven method.

It may be considered as a sub set of information processing.

Data analysis involves several strategies while encompassing various techniques in science,business as well as social science domains .

It is a basic direction of many data integration as well as data management tasks like data wrangling, data warehousing, data integrationas well as application integration. Data visualization represents any effort to help people consider the connotation of data through putting it in proper perceptible reference. In this context, trends, patterns and correlations may remain undetected in case of text based data.

Lecture Notes Unit-01 Chapter-02 Lecture-02

1.1 Big Data Ultimatums and Features

The prospecting of Big Data affords several alluring opportunities. But, profes- sionals and investigators are confronting various challenges while exploring Big Data sets while retrieving knowledge and value from inspection of information. The heftiness resides at several levels: data ensnaring, storing, searching, sharing, analysis, administration and fantasy. Moreover, there exists both privacy as well as security issues concerned with distributed data driven applications. Following are the major challenges associated with Big Data:

Volume: Huge volume of data means huge amount of data that are generally indicated as "tonna bytes" to refer the real numerical scale where the data volume becomes impugn specifically in setting domain specific data

Variety: It refers to complexity and more features per data item, the expletive of dimensionality, combinatorial detonation, various more data types, as well as several data formats.

Velocity: It refers to the rate of flow of data items into and out of the system in real time.High rate of data flow may arise threat to thesystem.

Veracity: It is essential for sufficient data to be tested in many distinct hypotheses, huge training samples for more micro-scale model-construction as well as model corroboration

Validity: It refers to quality of data, governance, master data management (MDM) on hefty, various, dispersed as well as divergent, "unclean" data gatherings.

Value: It characterizes the corporate value and latent of big data to revamp the institution from top to bottom.

Variability: It refers to dynamic spatiotemporal data, time series, seasonal, as well as many other type of non-static feature marked in data sources, clients, objects of study, etc.

Venue: It represents discordant distributed data from different platforms gathered out of various owners' systems, having distinct access as well as furcating needs.

Vocabulary: Vocabulary refers to schema, data models, ontologies, semantics, taxonomies, as well as other content- as well as context-based metadata which aim at describing the data's structure, syntax, content, as well as provenance.

Following are few other Big Data ultimatums:

• Big Data Management: Various ultimatums are encountered by Data Scientists while considering with Big Data. One ultimatum is the way to gather, integrate and keep with lesser software as well as hardware requirements [11, 12]. A major ultimatum is Big Data management.

Proper management is required to facilitate extraction of reliable insight while optimizing the expenses incurred. In fact, a proper data management is the bridge for Big Data analytics. Fundamentally, Big Data management refers to cleaning data for reliability, to combine data coming from several sources as well as encoding the data for security and privacy. Big Data management is carried out for ensuring authentic data which is easily available, manageable as well as secured.

• Big Data Cleaning: The crucial ultimatum in Big Data is to handle the com- plexity of Big Data nature such as: velocity, volume and variety [13] and getting processed it in a dispersed environment with mixed implications. On the contrary, data may retain errors, noises or incomplete data etc. The major ultimatum in this context is the way to clean such large data sets and the way to make decision regarding which data is believable and useful.

• Big Data Aggregation: It is essential to combine internal data with external data sources. Internal data refers to the data generated inside the organization and external data includes the third party sources, information regarding market fluctuation, traffic conditions, weather forecasting, etc.

• Imbalanced Big Data: Classifying imbalanced dataset is one of the crucial challenge in case of Big Data. In practical scenarios, real world implementations may generate classes having distinct distributions. Further, the classical learning methodologies are not applicable to data sets that are imbalanced. This is due to the fact that the model architecture is based on global search measures without taking the count of instances. In fact, the global guidelines are fundamentally preferred in lieu of particular guideline.

• Big Data Machine Learning: The main aim of machine learning is to unfold the knowledge while making intelligent decisions. It is employed in various real life applications like recognition systems, autonomous control systems, recommendation engines, data mining and informatics

Basically, machine learning has been segregated into three major domains: Supervised learning, Unsupervised learning and Reinforcement learning. For detailed study of ML types reader may follow Qiu et al.

• Deep learning: In modern era, Deep learning (DL) establishes an exceedingly popular research field in pattern recognition and machine learning. It plays a crucial role in various application prospects of predictive analysis viz. speech recognition, computer vision, natural language processing, etc.

DL is very versatile for resolving learning problems encountered in large data sets.It assists in automatically extracting complex data representations from huge volumes of uncategorized raw data. Furthermore, Deep learning is basically stratified learning and is suitable for simplifying analysis of profuse volume of data.

Further, the crucial characteristics of Big Data can be summarized as illustrated in Table

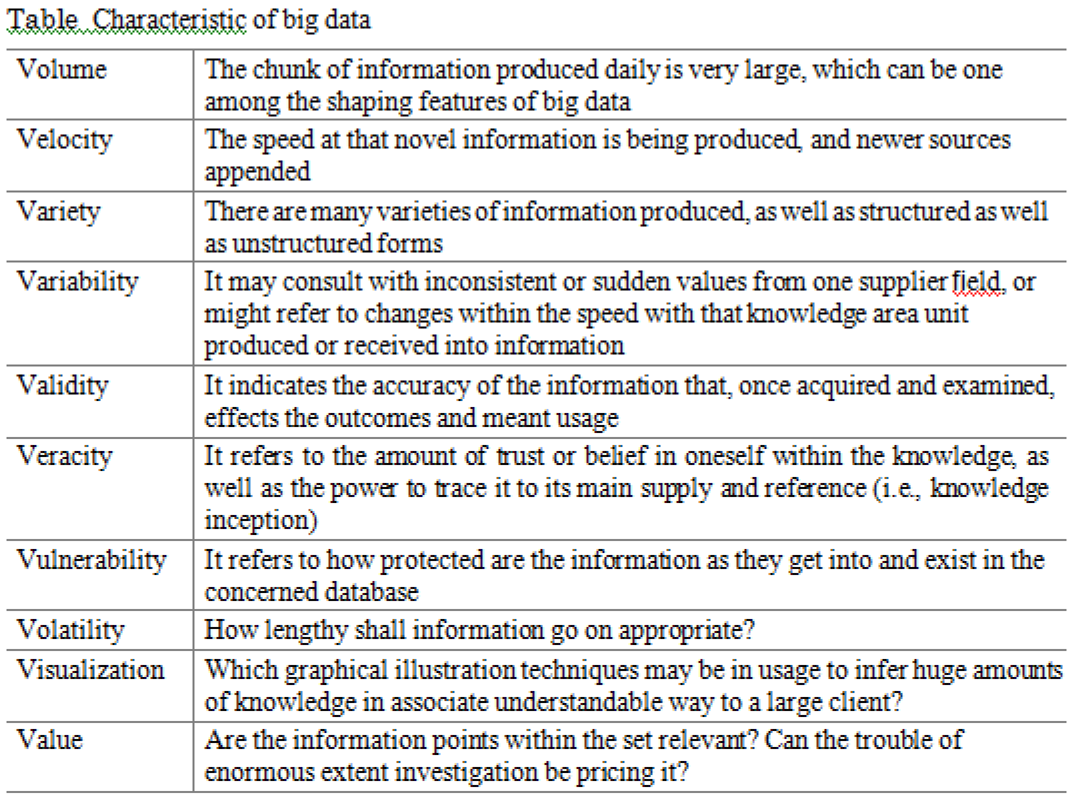

Table Characteristic of big data

Volume	The chunk of information produced daily is very large, which can be one among the shaping features of big data
Velocity	The speed at that novel information is being produced, and newer sources appended
Variety	There are many varieties of information produced, as well as structured as well as unstructured forms
Variability	It may consult with inconsistent or sudden values from one supplier field, or might refer to changes within the speed with that knowledge area unit produced or received into information
Validity	It indicates the accuracy of the information that, once acquired and examined, effects the outcomes and meant usage
Veracity	It refers to the amount of trust or belief in oneself within the knowledge, as well as the power to trace it to its main supply and reference (i.e., knowledge inception)
Vulnerability	It refers to how protected are the information as they get into and exist in the concerned database
Volatility	How lengthy shall information go on appropriate?
Visualization	Which graphical illustration techniques may be in usage to infer huge amounts of knowledge in associate understandable way to a large client?
Value	Are the information points within the set relevant? Can the trouble of enormous extent investigation be pricing it?

2.4 Types of Data Sources

Now a days “patient care activities” create huge amounts of knowledge as an essential byproduct of synergy among the “health-care-system”. Analytic and numerical knowledge will comprehend not simply supplier nothings, however conjointly information from processed supplier order entry (CPOE) and clinical resolution-making software package, laboratory and radiology outcomes (written-reports and therefore the digital imaging files), automatic outcomes from patient observance devices, and registration and monetary knowledge. Such information is also segregated into 2 extreme divisions as follows.

2.4.1 Structured Knowledge

Structured knowledge are generally those which may be arranged as searchable tables, typically produced from a preplanned set of answering decisions. The values is also elite from a menu by practicing as which act with an EHR, for instance, or could embody a listing of investigative or charge codes.

2.4.2 Unstructured Knowledge

Unstructured knowledge comprises of up to 80% of all “health-care” data, and they represent that which can’t be as simply investigated or arranged, like responses got into “free-text fields” by patients or physicians, “narrative-notes”, “hand-written” or “scanned documents”, “images”, etc.

2.4.3 Structured Versus Unstructured Data

The difference between structured and unstructured data is illustrated in below Table

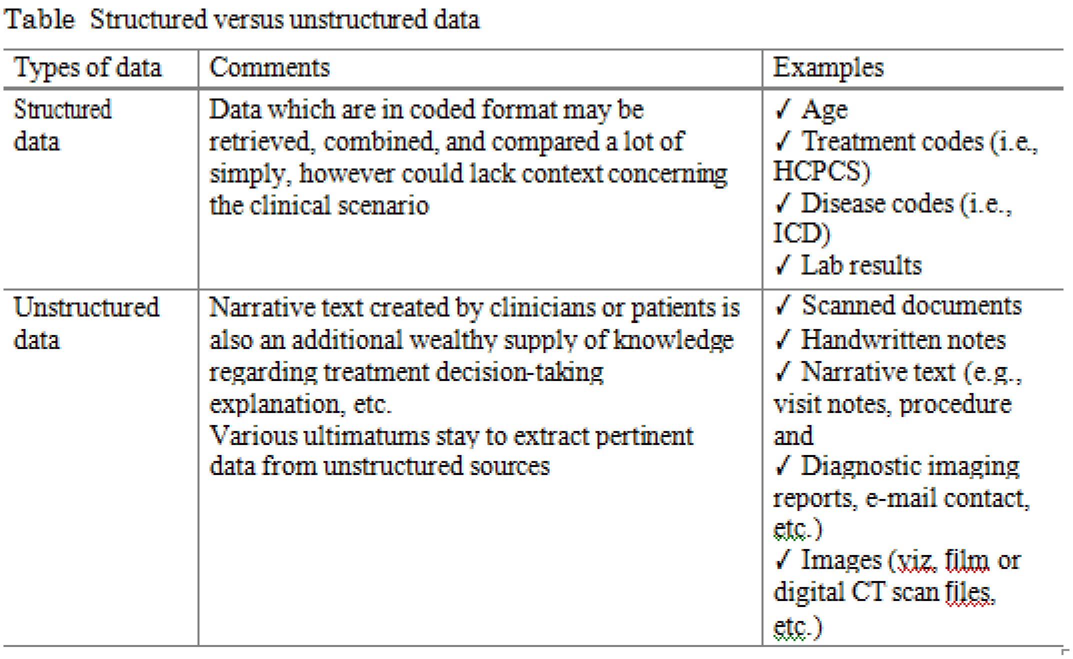

Table Structured versus unstructured data

Types of data	Comments	Examples
Structured data	Data which are in coded format may be retrieved, combined, and compared a lot of simply, however could lack context concerning the clinical scenario	✓ Age ✓ Treatment codes (i.e., HCPCS) ✓ Disease codes (i.e., ICD) ✓ Lab results
Unstructured data	Narrative text created by clinicians or patients is also an additional wealthy supply of knowledge regarding treatment decision-taking explanation, etc. Various ultimatums stay to extract pertinent data from unstructured sources	✓ Scanned documents ✓ Handwritten notes ✓ Narrative text (e.g., visit notes, procedure and ✓ Diagnostic imaging reports, e-mail contact, etc.) ✓ Images (viz. film or digital CT scan files, etc.)

2.5 Real World Applications of Big Data

The word big-data was first coined by NASA researchers in the article in 1997 as it presented the ultimatums to preserve the quantity of data produced as a results of a brand novel, data-demanding variety of process effort. As the motion of information production has magnified, ultimatums on the far side the power to easily store extensive volumes of information became probable, along with the competence to efficiently govern and with success interpret the data collected. The capability to pay off acumen from these huge armory of information has matured to be the recent ultimatum across regulations, incorporating health affliction.

From time-intensive activities, Information creation in health care has historically appeared like destined analytic experiments, with a often alluded gap of seventeen years from the primary distribution of analysis outcomes to pursuit in clinical observe. To contrast this, the Institute of drugs printed the educational attention System, spotlighting the necessity for a brand new en sample to sooner and ceaselessly integrates the most effective proof out of each rigorous clinical analysis with data attained as a universal projection of patient-care to assure novelty, endowment, security and price in health affliction.

The support key of the shift has been the expanding acceptance of “electronic-health-records (EHRs)”, in factor of incited as a results of the HITECH Act, that advertised its uptake &made

public, “multi-stage” commit to tested regulation for significant usage supposed for extending information abduction and distribution, make better clinical advancement, and ultimate trial of improved patient results. Although EHRs exhibit great opportunity to abduct and distribute long-term patient-data across multiple medical care and particular settings, these advantages are difficult to comprehend. Even inside one health-care system make use of a similar EHR, Health Insurance-movability-and-responsibleness-Act (HIPAA) necessities and lack of uniform gathering of data practices still limited period to have the facility to get patient data.

At the population stage, these appliances on their self do not seem to be appropriate to arrange or permit simple retrieval of information for quality coverage or analysis purposes, presently distinctive HER products preponderantly perform as a knowledge abduct appliance and archive for patient-data, without a profound analytic part. While the normal methodology to demonstrate the protection and effectiveness of the ongoing latest cancer medicine can stay the multicenter irregular-controlled trial, huge knowledge platforms offer the chance to accelerate development of complementary strategies of proof generation. The provision of massive knowledge platforms permits the chance to explore influence of inter- ventions likewise as effectiveness, enhancing generalizability and also the involvement of public not habitually taking part in restrained clinical trials, like older grownups and people with scarce infections.

The pivotal to the current amendment is that the appreciation that source and therefore the amount of “real world data” created by patients along cancer area unit currently procuring much more promptly and loosely than data from scientific tentatives will be created. Whereas it’s going like appearing perceptive to use these information, this becomes vital to outline the features of knowledge components and set that fitly devote to authentic experimental enquiry. In August, 2017, the North American country Food and Drug Administration (FDA) promulgated last steerage on the Use of Real-World proof (RWE) to assist regulative Decision-Making for therapeutic appliances, that afforded acumen into the agency’s vista on, however, RWE may be integrated into observation and endorsement decision-making, affording acceptable controls on attributes, efficacy, as well as responsibleness.

Such fields would possibly embrace expedition of contemplative information for getting hypotheses for experimenting in a very intended clinical tentative; genesis of an authentic management association, development of intended repositories with restrained information components to assist later consent and “off-label” observa- tion of instruments and for communal health police work resolutions. The authority acclaimed there may be several authentic inceptions of universal information, like huge easy attempts, or realistic scientific attempts, intended data-based or written account reviews, contemplative info observations, crating narrations, body and soundness consternation prerogatives, electronic health archives, information real- ized as a component of a mutual fettle examination or routine communal vigor police work, and registries.

Although the background of this credential is restricted to medical appliances, some general arguments might doubtless uses the FDA's aspects on RWE within the narcotic evolution and endorsement method moreover. Here, the numerous kinds of knowledge required to realize acumen into the matters, route, and con- sequences of the tumor expertise, varied teams, both public and personal, square measure operating to attach analytical, gynic, monetary, way of living, and different kinds of knowledge created by and for meeks with cancer.

This space can likely still apace extend and develop as results of automation and policy enhancement over ensuing days to come.

A leading stimulus for evolvement during such space is that the Cancer Moon shot drive, started in Jan 2016 by Joe Biden. The National-Cancer-Institute and therefore the presidentially delegated The National-Cancer planning board, a badge Panel was appointed to first assess the state of the science, and so develop rec- ommendations in seven topic-areas, one in all that was associate degree increased knowledge sharing unit. Along side the six different operating teams, the total document of decree was free within the fall of 2016. Besides, Bilal et al.showed the vast use of Big Data in industrial applications.

Forty five Recommendations from the improved knowledge Sharing unit enclosed the necessity to make a Cancer knowledge system to supply the same, in knowledge erudition underpinnings which will hook upnumerous inceptions and kinds of malignancy knowledge, as well as carry out therefore with patient com- mitment, approval, and confidentiality-armaments. The National Cancer Institute had initiated to concern money divulgences through twelvemonth 2019 to assist the implementation of analysis opportunities known by the badge Panel operating teams, assisted partly by funding of the twenty first Century Cures Act. From the Cancer rocket launching cordon bleu Panel increased knowledge sharing working party recommendation. This is a tide of exponential peregrination and aspiration for the conveniences that huge knowledge, knowledge erudition, and real-world knowledge production might produce.

Patients having cancer urgently want novel proof that’s a lot of concretized to own clinical, analytical, and canonic features than ever before the technology is quickly exploring, momentous ultimatums stay to create an exhaustive, practical, and easy medicine knowledge framework that delivers on the assurance of RWE. As nurses, we have a tendency to should currently defend for our patients in a very new method—to be told the way to raise the foremost significant queries from among huge datasets, in order that we will regularly progress towards care of the very best attributes and price for the complete patient, not simply concerned tumors.

Lecture Notes Unit-01 Chapter-02 Lecture-03

2.6 IoT Architecture and Security Challenges

A scenario of a common IoT encompasses different devices along with “embedded sensors” connected via a network. Here the devices are identifiable uniquely and are characterized through lesser memory, lower power as well as limited processing feature. Further, the gateways represent the devices that are employed for connecting IoT devices to the external sphere for distant data provisioning and benefits to IoT customers.

2.6.1 IoT Standards and Protocols

General IoT protocols are employed for messaging, forwarding, authentication and various relevant applications. It incorporates the standards as well as protocols ordinarily employed for Low Rate Wireless Personal Area Network (LW-WPANs) and for Low Power Wide Area Network (LP-WAN).

Basically, Physical Layer and the Medium Access Control (MAC) Layer are the two layers described by IEEE standard 802.15.4. In this context, the physical Layer specification is concerned with exchange through wireless channels having different data rates as well as frequency bands. The specification for MAC Layer is concerned with mechanisms for synchronization and channel access. Moreover, each of the appliances in IoT is solely determined through an IPv6 network address. Further, the Routing Protocol for Low Power Lossy Networks (RPL) is employed to assist WPAN environments. RPL affords communication between single point and multi-points. The application framework in IoT includes “User Datagram Protocol (UDP)” for the purpose of exchange due to limited payload. This is owing to the fact that UDP is much more effective as well as comparatively less complex than “Transmission Control Protocol (TCP)”. Again, “Internet Control Message Protocol (ICMP)” is employed for Control Messages (CMs), viz. specifying not reachable destinations as well as neighborhood discovery. Further, the Constrained Application Protocol (CoAP) helps in Asynchronous message exchange and helps in HTTP mapping .

2.6.2 IoT Security Requirements

Following are some of the relevant parameters needed for secure IoT deployment.

2.6.2.1 Availability of Service

The various incursions over IoT devices need the accouterment of several utilities via the customary denial of service attack. Several techniques incorporating sink- hole attacks, blocking antagonists as well as replay incursions makes use of IoT integrals at numerous layers to depreciate the Quality of Service (QoS).

2.6.2.2 Authentication, Authorization, Accounting

IoT authentication is essential for securing the communication prevailing in IoT. The devices need to be authenticated for privileged ingress to several services. Moreover, the authorization mechanism indicates that the ingress to system or data is afforded to legitimate users. In addition, the reckoning for resource usage, along with informing afford reliability in proper network management.

2.6.2.3 Data Privacy, Data Confidentiality, Data Integrity

A proper encryption mechanism is needed for ensuring data confidentiality since IoT data moves via several lopes in a network. Owing to various amalgamation of services, the data contained in a device is susceptible to privacy and might cause the traducer to affect the data integrity through changing the hoarded data for mis- chievous purpose .

2.6.2.4 Energy Efficiency

The IoT devices are specifically resource stiffed devices and are having reduced power as well as lowered storage. The various incursions on IoT frameworks gives rise to an hike in energy expenditure through flooding the intended network as well as fatiguing the various IoT resources.

2.6.3 Single Points of Failure

A continuing rise of divergent network may lead to a huge number of single points of failures that might destroy the services afforded through IoT. It needs therefore, a tamper-proof environment for IoT devices and fault tolerant networks.

2.7 IoT Code

The UK government has launched a code called IoT Code for consumer Internet of Things (IoT) products. This assists with the hope to make them less liable to hacking. The Security by Design Report published by Digital Culture, Media and Sport (DCMS) analyzes the recent proliferation in IoT market, however, it also portrays the raise in associated risks owing to failure of several vendors for con- sidering security as part of the product development life cycle.

2.8 Conclusion and Future Scope

Basically, Big data and IoT are two pragmatic features in recent scenarios for various types of organizations and they represent a greater challenge to Information Technology (IT). The major features of Big data are concerned with its volume as well as variety including other features, viz. velocity, value, and veracity, etc. Further, Big data and IoT have impact on both public sectors and private sectors and conjointly scientific and technical zones incorporating healthcare, education systems, etc. Now-a-days, IoT forms a large amount of data in vertical silos. On the contrary, an actual IoT is dependent on the actual availability as well as confluence of rich data sets collected out of various systems. The data can be taken from any source and get examined to attain the answer, thus curtailing the cost as well as time. Furthermore, it helps in smart decision making as well as new product development. Besides, Big data assists in deciding the crucial causes behind any issues or failures or defects and detecting fraudulent behavior in real time. In brief, big data concern with data but IoT is about data, devices as well as the connectivity.

Although the technological conditions for big data management have pro- gressed, they are not at all sufficient since they need to be aggregated and handled by desirably trained personnel. In view of such circumstance, the novel technical skills as well as required qualified and trained professionals are needed for recent technical arsenal. Furthermore, the procedure has to be greater efficient, in many cases needing proper revisiting, either in the user's point of view, or from the technical perspectives while taking into account the human issues for operating Big Data as well as IoT.