BIG DATA ANALYTICS FOR IOT

A

Seminar Report submitted to Savitribai Phule Pune University,Ganeshkhind

In Fulfillment for the awards of Degree of Engineering in Computer Engineering

Submitted by

Monali Chavan T120414215

Under the Guidance of

Asst.Prof.B.A.Khivsara

March,2016 - 17

Department of Computer Engineering

SNJB's Late Sau. Kantabai Bhavarlalji Jain College of Engineering, Chandwad Dist:Nashik

SNJB's Late Sau. Kantabai Bhavarlalji Jain College of Engineering, Chandwad Dist:Nashik

Department of Computer Engineering

2016-17

Certificate

This is to certify that the Seminar Report entitled *Big Data Analytics For IOT* submitted by Ms. **Monali Chavan** is a record of bonafied work carried out at SNJB's K. B. Jain College of Engineering during academic year 2016-17, which affiliated to the Savitribai Phule Pune University.

Date:
Place: Chandwad

Seminar Guide
Asst.Prof.B.A.Khivsara

Prof. K.M. Sanghavi **Dr.M.D.Kokate,**
HOD **Principal**

Examiner:..

Acknowledgement

I would like to acknowledge all the people who have been of the help and assisted me throughout my project work. First of all I would like to thank my respected guide Asst.Prof.B.A.Khivsara, Asst. Professor in Department of Computer Engineering for introducing me throughout features needed. The time-to-time guidance, encouragement, and valuable suggestions received from him are unforgettable in my life. This work would not have been possible without the enthusiastic response, insight, and new ideas from her.

I am also grateful to all the faculty members of SNJB's College of Engineering for their support and cooperation.

I would like to thank my lovely parents for time-to-time support and encouragement and valuable suggestions, and thank my friends for their valuable support and encouragement.

The acknowledgement would be incomplete without mention of the blessing of the Almighty, which helped me in keeping high moral during most difficult period.

Monali Chavan

Abstract

Current research on Internet of Things (IoT) mainly focuses on how to enable general objects to see, hear, and smell the physical world for themselves, and make them connected to share the observations. In this we argue that only connected is not enough, beyond that, general objects should have the capability to learn, think, and understand both the physical world by themselves. On the other hand, the future IoT will be highly populated by large numbers of heterogeneous networked embedded devices, which are generating massive or big data in an explosive fashion. Although there is a consensus among almost everyone on the great importance of big data analytics in IoT.[1]

Big data enables real-time systems monitoring, management, optimization and anticipation. In this we present some examples of applications of big data analysis in scenarios of smart cities[14].

Contents

List of Figures

Chapter 1

Introduction

This disruptive technology requires new infrastructures, including hardware and software applications, as well as an operating system; enterprises will need to deal with the influx of data that starts flowing in and analyze it in real-time as it grows by the minute. Thats where big data comes in; big data analytics tools are capable of handling masses of data transmitted from IoT devices that produce a continuous stream of information. But just to differentiate the two, the IoT delivers the information from which big data analytics can draw the information to create the insights required of it.However, the IoT brings data on a different scale, so the analytics solution should accommodate its needs of rapid ingestion and processing followed by an accurate and fast extraction. Solutions like SQream Technologies deliver near real-time analytics on massive sized datasets, and essentially condense a full-rack database into a small server processing up to 100TB, so minimal hardware is required. The next generation analytics database leverages GPU technology, allowing an even further downsizing of the hardware, i.e. big database in the car, or 5 TB on a laptop. This helps IoT companies correlate the growing number of data sets, which helps them get real-time responses and adapt to changintrends, solving the size challenge without compromising on the performance[8].

1.1 Internet Of Things

Internet of Things is a network of physical object or things. embedded with electronics,software,sensors,and network connectivity,which enables these objects to collect and exchange dats. IOT allows objects to be sensed and controlled remotely across existing network infrastructure,creating opportunities for more direct integration between the physical.

The concept of IoT aims to take a wide range of things and turn them into smart objects anything from watches to fridges, cars and train tracks. Products that normally would not be connected to the Internet and able to obtain and process data, are equipped

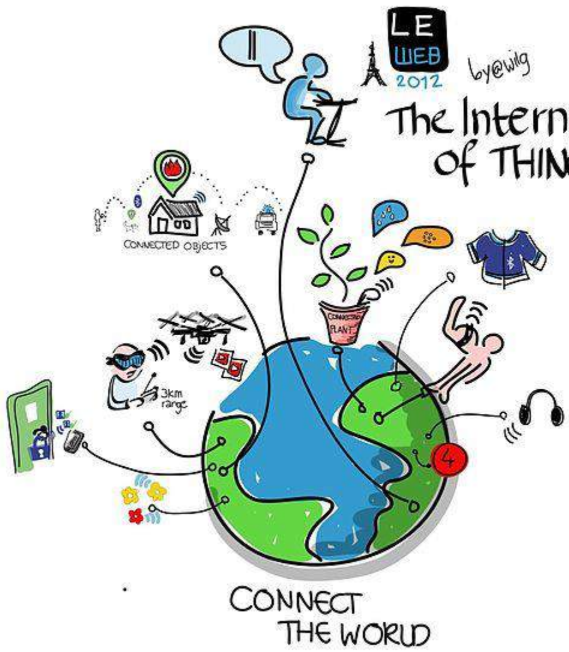

Figure 1.1: OVER VIEW TO INTERNET OF THINGS

with sensors and computer chips for data gathering. However, unlike chips used in PCs, smartphones, and mobile devices, these chips are used mostly for gathering data that indicates customer usage pattern and product performance[8].

The IoT is essentially the means that collects and sends data. Information from IoT devices resides in big data and is measured against it. And IoT will soon touch every aspect of our lives: transportation (cars, smart train tracks and traffic lights), manufacturing, Smart Homes (thermostats and voice activated appliances), and of course consumer goods such as smartphones, wearables, and more. [8].

1.2 Big Data Analytics

Big data analytics is the process of collecting, organizing and analyzing large sets of data (called big data) to discover patterns and other useful information. Big data analytics can help organizations to better understand the information contained within the data and will also help identify the data that is most important to the business and future business decisions. Analysts working with big data basically want the knowledgethat comes from analyzing the data[3].

Big data has existed long before the IoT burst out into the scene to perform analytics; information is defined as big data when it demonstrates the 4 Vs: volume, variety, velocity, and veracity. This equates to a massive quantity of data that can be both structured and unstructured, while velocity refers to the speed of data processing, and veracity determines its uncertainty.[8]

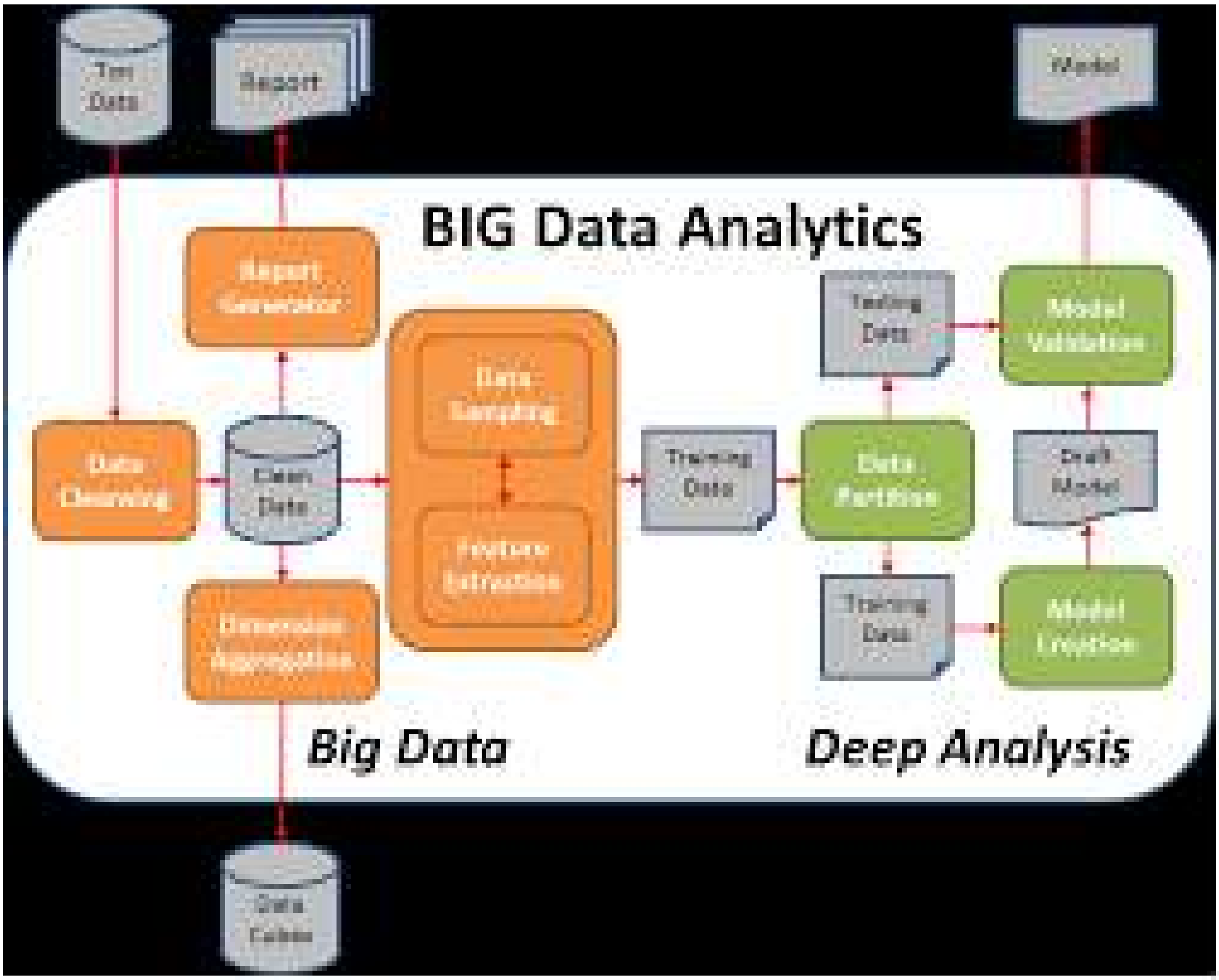

Figure 1.2: OVER VIEW TO BIG DATA ANAYTICS

By now, many companies have decided that big data is not just a buzzword, but a new fact of business life – one that requires having strategies in place for managing large volumes of both structured and unstructured data. And with the reality of big data comes the challenge of analyzing it in a way that brings real business value. Business and IT leaders who started by addressing big data management issues are now looking to use big data analytics to identify trends, detect patterns and glean other valuable findings from the sea of information available to them.

It can be tempting to just go out and buy big data analytics software, thinking it will be the answer to your company's business needs. But big data analytics technologies on their own aren't sufficient to handle the task. Well-planned analytical processes and people with the talent and skills needed to leverage the technologies are essential to carry out an effective big data analytics initiative. Buying additional tools beyond an organization's existing business intelligence and analytics applications may not even be necessary depending on a project's particular business goals.

This Essential Guide consists of articles and videos that offer tips and practical advice on implementing successful big data analytics projects. Use the information resources collected here to learn about big data analytics best practices from experienced users and industry analysts – from identifying business goals to selecting the best big data analytics tools for your organization's needs.[9]

1.3 Challenges of IOT and Big Data

1.Security

As the IoT connects more devices together, it provides more decentralized entry points for malware. Less expensive devices that are in physically compromised locales are more subject to tampering. More layers of software, integration middleware, APIs, machineto-machine communication, etc. create more complexity and new security risks. Expect to see many different techniques and vendors addressing these issues with policy-driven approaches to security and provisioning[6].

2.Complexity, confusion and integration issues:

With multiple platforms, numerous protocols and large numbers of APIs, IoT systems integration and testing will be a challenge to say the least. The confusion around evolving[6].

3.Trust and Privacy:

With remote sensors and monitoring a core use case for the IoT, there will be heightened sensitivity to controlling access and ownership of data. (Note that two recent high-profile security breaches at Target and Home Depot were both achieved by going through third-party vendors= stolen credentials to gain access to payment systems. Partner vetting will become ever more critical.) Compliance will continue to be a major issue in medical and assisted-living applications, which could have life and death ramifications. New compliance frameworks to address the IoT=s unique issues will evolve. Social and political concerns in this area may also hinder IoT adoption[6].

4.Standards and interoperability:

Standards are important in creating markets for new technologies. If devices from different manufacturers do not use the same standards, interoperability will be more difficult, requiring extra gateways to translate from one standard to another. In addition, a company that controls different parts of a vertical market (e.g. the acquisition of data, its integration with other data streams, and the use of those data streams to come up with innovative solutions or to provide services) may dominate a market, stifling competition and creating barriers for smaller players and entrepreneurs. Differing data standards can also tend to lock consumers into one family of products: if consumers cannot easily transfer their data when they replace one device with another from a different manufacturer, they will in effect lose any benefit from the data they have been accumulating over time[6].

5 Data storage and analytics:

A challenge for producers of IoT applications is to clean, process and interpret the vast amount of data which is gathered by the sensors. There is a solution proposed for the analytics of the information referred to as Wireless Sensor Networks.[154] These networks share data among sensor nodes that are send to a distributed system for the analytics of the sensory data. Another challenge is the storage of this bulk data. Depending on the application there could be high data acquisition requirements which in turn lead to high storage requirements. Currently the internet is already responsible for 5 per of the total energy generated[154] and this consumption will increase significantly when we start utilizing applications with multiple embedded sensors[6].

Chapter 2

Literature Survey

The significant increase in connected devices thates due to happen at the hands of the Internet of Things will, in turn, lead to an exponential increase in the data that an enterprise is required to manage. Herees where IoT intersects wonderfully with big data . and where it becomes evident that the two trends fit one another like a glove.the Internet of things gets rolling, stand back,. warns Howard Baldwin, writing for Forbes. ere going to have data spewing at us from all directions - from appliances, from machinery, from train tracks, from shipping containers, from power stations. If that doesnet get you thinking how to handle real-time data feeds, nothing will.. herees a suggestion,. he adds. now..Fact is: the Internet of Things is still in its infancy. It hasnet started to produce an overwhelming deluge of information. But that doesnet mean it wonet. Big Data, on the other hand, has been around for a while; long enough that ites starting to come into its own. Analytics tools designed to handle large, fast-changing volumes of information are gradually becoming accessible to small and midsized organizations, while data science is looked upon as a legitimate and highly-valued- field of study. In short, for most businesses, the timing has never been better to look into the adoption of a big data strategy.[11]

2.1 Technical Standards For IOT

This is a list of technical standards for the IoT, most of which are open standards, and the standards organizations that aspire to successfully setting them[10].

- **Shotr Name : Auto-ID Labs :** Standards :: Networked RFID (radiofrequency identification) and emerging sensing
- **Short Name : GS1 :** Standards for UIDs (unique identifiers) and RFID of fast-moving consumer goods (consumer packaged goods), health care supplies, and other things

- **Short Name : IETF :** Standards that comprise TCP/IP (the Internet protocol suite)
- **Short Name : GS1 :** Standards for UIDs (unique identifiers) and RFID of fast-moving consumer goods (consumer packaged goods), health care supplies, and other things
- **Short Name : MTConnect Institute :** MTConnect is a manufacturing industry standard for data exchange with machine tools and related industrial equipment. It is important to the IIoT subset of the IoT[10].

There are many technologies that enable IoT. Crucial to the field is the network used to communicate between devices of an IoT installation, a role that several wireless or wired technologies may fulfil

2.2 Technologies in IOT

Short Range

- **Bluetooth low energy (BLE) :** Specification providing a low power variant to classic Bluetooth with a comparable communication range.
- **Light-Fidelity (Li-Fi):** Wireless communication technology similar to the Wi-Fi standard, but using visible light communication for increased bandwidth.
- **Near-field communication (NFC) :** Communication protocols enabling two electronic devices to communicate within a 4 cm range.
- **QR codes and barcodes :**Machine-readable optical tags that store information about the item to which they are attached.
- **Radio-frequency identification (RFID):**Technology using electromagnetic fields to read data stored in tags embedded in other items.
- **Thread :** Network protocol based on the IEEE 802.15.4 standard, similar to ZigBee, providing IPv6 addressing.
- **Wi-Fi :** Widely used technology for local area networking based on the IEEE 802.11 standard, where devices may communicate through a shared access point.
- **Wi-Fi:** Direct V Variant of the Wi-Fi standard for peer-to-peer communication, eliminating the need for an access point.

- **Z-Wave :**V Communication protocol providing short-range, low-latency data transfer at rates and power consumption lower than Wi-Fi. Used primarily for home automation.
- **ZigBee :**V Communication protocols for personal area networking based on the IEEE 802.15.4 standard, providing low power consumption, low data rate, low cost, and high throughput[10].

Medium-range wireless

- **HaLow :** Variant of the Wi-Fi standard providing extended range for low-power communication at a lower data rate.
- **LTE-Advanced:** High-speed communication specification for mobile networks. Provides enhancements to the LTE standard with extended coverage, higher throughput, and lower latency[10].

Long-range wireless

- **Low-power wide-area networking (LPWAN):** Wireless networks designed to allow long-range communication at a low data rate, reducing power and cost for transmission.
- **Very small aperture terminal (VSAT):** Satellite communication technology using small dish antennas for narrowband and broadband data[10].

Wired

- **Ethernet:** General purpose networking standard using twisted pair and fiber optic links in conjunction with hubs or switches.
- **Multimedia over Coax Alliance (MoCA):** Specification enabling whole-home distribution of high definition video and content over existing coaxial cabling.
- **Power-line communication (PLC):** Communication technology using electrical wiring to carry power and data. Specifications such as HomePlug utilize PLC for networking IoT devices.[10]

2.3 IoT Development Tools

1. Arduino

Arduino is an open-source prototyping platform based on easy-to-use hardware and software. It is both a hardware specification for interactive electronics and a set of software that includes an IDE and the Arduino programming language. The website explains that Arduino is "a tool for making computers than can sense and control more of the physical world than your desktop computer"[17].

Figure 2.1: AURDINO KIT

2. Eclipse IoT Project

Have you heard of the Lua programming language yet? Eclipse is sponsoring several different projects surrounding IoT. They include application frameworks and services; open source implementations of IoT protocols and tools for working with Lua, which Eclipse is promoting as an ideal IoT programming language. Eclipse-related projects include Mihini, Koneki and Paho[17].

3. Kinoma

Kinoma, a Marvell Semiconductor hardware prototyping platform encompasses three different open source projects. Kimona Create is a DIY construction kit for prototyping electronic devices. Kimona Studio is the development environment that works with Create and the Kinoma Platform Runtime. Kimona Connect is a free iOS and Android app that links smartphones and tables with IoT devices. Github Sparkle Motion: Create an LED world map driven by global Twitter traffic data[17].

4. M2MLabs Mainspring

M2MLabs Mainspring is an open source application framework for building machine to machine (M2M) applications such as remote monitoring, fleet management or smart grid.Its capabilities include flexible modeling of devices, device configuration, communication between devices and applications, validation and normalization of data, long-term data storage, and data retrieval functions. It's based on Java and the Apache Cassandra NoSQL database. M2M applications can be prototyped in hours rather than weeks and finally transfered to a high performance execution environment built on top of a standard J2EE server and the highly scaleable Apache Cassandra database[17].

5. Node-RED

A visual tool for wiring the Internet of Things i.e wiring together hardware devices, APIs and online services in new and interesting ways. Built on Node.js, Node-RED describes itself as "a visual tool for wiring the Internet of Things." It allows developers to connect devices, services and APIs together using a browser-based flow editor. It can run on Raspberry Pi, and more than 60,000 modules are available to extend its capabilities.Contribute to the Node-Red IBM project or on node-red github[17].

2.3.1 Hardware

1. Arduino Yn

Arduino is an open-source electronics platform based on easy-to-use hardware and software. This microcontroller combines the ease of an Arduino-based board with Linux. It includes two processorsthe ATmega32u4 (which supports Arduino) and the Atheros AR9331 (which runs Linux). Other features include Wi-Fi, Ethenet support, a USB port, micro-SD card slot, three reset buttons and more. Project: Ultrasonic Map-Maker using an Arduino Yun- Generates maps based on distance between itself and obstacles autonomously and provides visual feedback[17].

2. BeagleBoard

BeagleBoard offers credit-card sized computers that can run Android and Linux. Because they have very low power requirements, they're a good option for IoT devices. Both the hardware designs and the software they run are open source, and BeagleBoard hardware Big Data Analytics Using IOT SNJB COE (often sold under the name BeagleBone) is available through a wide variety of distributors. Experiment with Linux, Android and Ubuntu and jump-start development in five minutes with the included USB cable.Project: Measuring Temperature with a BeagleBone Black, learn how to connect temperature sensor to a BeagleBone Black[17].

3. Flutter

Flutter is a programmable processor core for electronics projects, designed for hobbysits, students, and engineers.Flutter's claim to fame is its long range. This Arduino-based board has a wireless transmitter that can reach more than a half mile. Plus, you don't need a router; flutter boards can communicate with each other directly. It includes 256-bit AES encryption, and it's easy to use. Github[17].

4. LightBlue Bean Punch Through

The LightBlue Bean is a low energy Bluetooth Arduino microcontroller. Using Bluetooth 4.0, it is programmed wirelessly, runs on a coin cell battery, and is perfect for smartphone controlled projects. With Bean, you can program wirelessly from any of your devices. No more unscrewing screws and ungluing glue. Github[17].

5. Microduino

Microduino presents the worlds smallest series of Arduino-compatible smart modules that are small, flexible, stackable and powerful, and can be used to create a limitless amount of DIY projects. Microduino offers really small boards that are compatible with Arduino. Interactive Projects[17].

6. OpenPicus

OpenPicus is an Italian hardware company who designs and produce Internet of Things system on modules called Flyport. Flyport is open hardware and the openPicus framework and IDE are open software. Its platform and hardware are open source, but its products can be used to create closed source commercial products. The company also offers its development services for hire[17].

2.4 Techniques of big data :

In the current paper, we will not be giving an in depth overview on the tools and techniques, however, we will be giving an overview of the tools and techniques associated with Big Data. This will help the reader get a association with the tools used for Big Data analytics[15].

Techniques :

There are a lot of techniques that could be used when going to start with a project. Some of the tools which have frequent usage are summarized here[15].

Association rule learning:

A set of techniques for discovering interesting relationships, i.e., rules,. among variables in large databases.

Data mining:

One of the most important terms related to data-driven decision making and describes it as searching or .digging into a data file for information to understand better a particular phenomenon..

Cluster analysis:

Cluster analysis is a type of data mining that divides a large group into smaller groups of similar objects characteristics of similarity are not known in advance..

Crowd sourcing:

Crowd sourcing collects data from a large group of people through an open call, usually via a Web2.0 tool. This tool is used more for collecting data than for analyzing it.

Machine learning:

Traditionally computers only know what we tell them, but in machine learning,subspecialty of computer science, we try to craft algorithms that allow computers to evolve based on empirical data.

Text analytics:

A large portion of generated data is in text form. Text Analytics is the process of converting unstructured text data into meaningful data. As with the analytical techniques, there are several software products and available technologies to facilitate big data analytics. Some of the most common will be discussed here.

EDWS :

Enterprise data warehouses are databases used in data analysis.

Visualization products:

One of the difficulties with big data analytics is finding ways to visually represent results. Many new visualization products aim to fill this need, devising methods for representing data points numbering up into the millions. Beyond simple representation visualization can also help in the information search.

Map Reduce :

MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples.

Haddop :

It is an open-source framework that allows to store and process big data in a distributed environment across clusters of computers using simple programming models. Hadoop is an Apachemanaged software framework derived from MapReduce and Big Table.

NoSQL databases :

NoSQL database, also called Not Only SQL, is an approach to data management and database design that's useful for very large sets of distributed data. NoSQL is especially useful when an enterprise needs to access and analyze massive amounts of unstructured data or data that's stored remotely on multiple virtual servers in the cloud. The most popular Big Data Analytics Using IOT SNJB COE NoSQL database is Apache Cassandra. Other NoSQL implementations include SimpleDB, Google BigTable, Apache Hadoop, MapReduce, MemcacheDB, and Voldemort[15].

Usage Area Of Big Data Analytics

Big data is used efficiently in numerous fields. Some of them are listed below[15] :

1. Automotive industry
2. High technology and industry
3. Oil and gas
4. Telecommunication sector
5. Retail industry
6. Packaged consumer products
7. Media and show business
8. Travel and transport sector
9. Social media and online services

2.5 Big Data Tool :

Pentaho Business Analytics:

Pentaho is another software platform that began as a report generating engine; it is, like JasperSoft, branching into big data by making it easier to absorb information from the new sources. You can hook up Pentaho's tool to many of the most popular NoSQL databases such as MongoDB and Cassandra. Once the databases are connected, you can drag and drop the columns into views and reports as if the information came from SQL databases. I found the classic sorting and sifting tables to be extremely useful for understanding just who was spending the most amount of time at my website. Simply sorting by IP address in the log files revealed what the heavy users were doing. Pentaho also provides software for drawing HDFS file data and HBase data from Hadoop clusters. One of the more intriguing tools is the graphical programming interface known as either Kettle or Pentaho Data Integration. It has a bunch of built-in modules that you can drag and drop onto a picture, then connect them. Pentaho has thoroughly integrated Hadoop and the other sources into this, so you can write your code and send it out to execute on the cluster[16].

Karmasphere Studio and Analyst:

Many of the big data tools did not begin life as reporting tools. Karmasphere Studio, for instance, is a set of plug-ins built on top of Eclipse. It's a specialized IDE that makes it easier to create and run Hadoop jobs. I had a rare feeling of joy when I started configuring a Hadoop job with this developer tool. There are a number of stages in the life of a Hadoop job, and Karmasphere's tools walk you through each step, showing the partial results along the way. I guess debuggers have always made it possible for us to peer into the mechanism as it does its work, but Karmasphere Studio does something a bit better: As you set up the workflow, the tools display the state of the test data at each step. You see what the temporary data will look like as it is cut apart, analyzed, then reduced. Karmasphere also distributes a tool called Karmasphere Analyst, which is designed to simplify the process of plowing through all of the data in a Hadoop cluster. It comes with many useful building blocks for programming a good Hadoop job, like subroutines for uncompressing Zipped log files. Then it strings them together and parameterizes the Hive calls to produce a table of output for perusing[16].

Talend Open Studio:

Tableau Desktop is a visualization tool that makes it easy to look at your data in new ways, then slice it up and look at it in a different way. You can even mix the data with other data and examine it in yet another light. The tool is optimized to give you all the columns for the data and let you mix them before stuffing it into one of the dozens of graphical templates provided. Tableau Software started embracing Hadoop several versions ago, and now you can treat Hadoop "just like you would with any data connection." Tableau relies upon Hive to structure the queries, then tries its best to cache as much information in memory to allow the tool to be interactive. While many of the other reporting tools are built on a tradition of generating the reports offline, Tableau wants to offer an interactive mechanism so that you can slice and dice your data again and again. Caching helps deal with some of the latency of a Hadoop cluster. The software is well-polished and aesthetically pleasing. I often found myself reslicing the data just to see it in yet another graph, even though there wasn't much new to be learned by switching from a pie chart to a bar graph and beyond. The software team clearly includes a number of people with some artistic talent[16].

Splunk:

Splunk is a bit different from the other options. It's not exactly a report-generating tool or a Big Data Analytics Using IOT SNJB COE collection of AI routines, although it accomplishes much of that along the way. It creates an index of your data as if your data were a book or a block of text. Yes, databases also build indices, but Splunk's approach is much closer to a text search process[16].

2.6 How the IOT changes Big Data Analytics

Theres plenty of hype surrounding the Internet of Things (IoT). But this is one time the hype underestimates what comes next. IDC says there will be 28 billion sensors in use by 2020, with 1.7 trillion in economic value. The scale, breadth, and business value will exceed anything seen in the past. Ray Kurzweils Singularity is here. IoT will be an order of magnitude bigger than big data in scale and value[18].

Imagine a few billion sensors sending messages 20 times a second or even once a minute. The scale of the data is astonishing. Even Facebook addicts canet talk that much. For many, IoT data volume will be in the petabyte range. Fortunately, the cost of disk storage continues its free fall. Every person on the planet will be touched by sensors in 2025, or sooner. Even in the Outback, the Sahara, and especially grandma. Get ready for digital rivers of data driving new growth use cases in every industry. Sensors and things operate at the edge (Figure 1). . are any item we can attach a sensor to . including you. The edge is where we find Operational Technology (OT). It includes manufacturing plants, cars, electrical grids, and train tracks. The OT engineers and operators have been using sensor data for decades. But now Information Technology (IT) is now pitching in to help out. Gateways are routers and servers that connect the OT to IT systems[18].

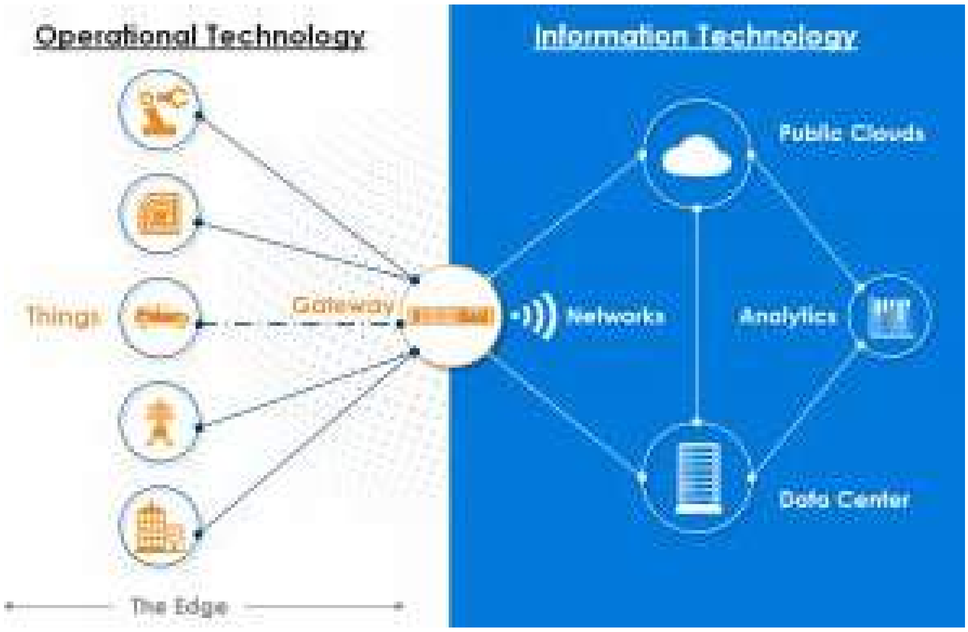

Figure 2.2: HOW IOT CHANGES BIG DATA ANALYTICS

A majority of the ROI comes from analyzing sensor data. Note that analytics are spread throughout IoT systems like chocolate baked into a cake. IoT analytics are collectively called the Analytics of Things (AoT)[18].

While larger enterprises like Coca-Cola, General Electric, and Domino.s Pizza have managed to tap into its value, most businesses will have to wait some time before they can really enjoy the advantages of embedded sensor technology. In the meantime, it.s imperative that those businesses prepare by adopting a big data strategy - and looking into analytics technology. Big Data capacity is, in essence, a prerequisite to tapping into the Internet of Things. Without the proper data-gathering in place, it.ll be impossible for businesses to sort through all the information flowing in from embedded sensors. What that means is that, without Big Data, the Internet of Things can offer an enterprise little more than noise[18].

Chapter 3

Working

As shown in Fig, in this we propose a systematic tutorial on the development of effective algorithms for big data analytics in future IoT, which are grouped into four classes: 1) heterogeneous data processing,2)nonlinear data processing, 3) high-dimensional data processing, and 4) distributed and parallel data processing[21].

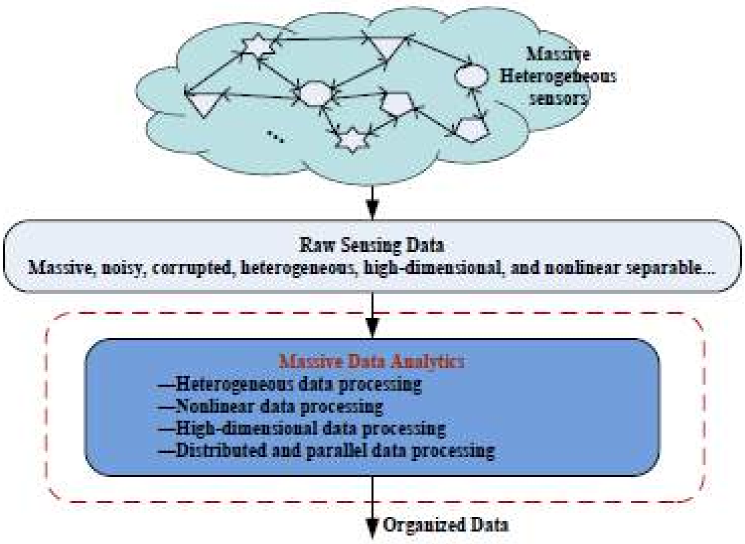

Figure 3.1: FRAMEWORK OF MASSIVE DATA ANALYTICS IN IOT

HETEROGENEOUS DATA PROCESSING

In practical IoT applications, the massive data are generally collected from heterogeneous sensors (e.g., cameras,vehicles, drivers, and passengers), which in turn may provide heterogeneous sensing data (e.g., text, video, and voice). Heterogeneous data processing (e.g., fusion, classification) brings unique challenges and also offers several advantages and new possibilities for system improvement[21].

Mathematically, random variables that characterize the data from heterogeneous sensors may follow disparate probability distributions. Denote zn as the data from the n-th sensor and Z := fzngNn=1 as the heterogeneous data set, the marginals fzngNn=1 are generally non-identically or heterogeneously distributed. In many IoT applications, problems are often modeled as multi-sensor data fusion, distribution estimation or distributed detection. In these cases, joint probability density function (pdf) f(Z) of the heterogeneous data set Z is needed to obtain from the marginal pdfs ff(zn)gNn=1.

For mathematical tractability, one often chooses to assume simple models such as the product model or multivariate Gaussian model, which lead to suboptimal solutions [6]. Here we recommend another approach, based on copula theory, to tackle heterogeneous data processing in IoT. In copula theory, it is the copulas function that couples multivariate joint distributions to their marginal distribution functions, mainly thanks to the following theorem: Sklar Theorem [7]: Let F be an N-dimensional cumulative distribution function (cdf) with continuous marginal cdfs F1; F2; :::; FN. Then there exists a unique copulas function C such that for all z1; z2; :::; zN in [..1;+1]

F(z1; z2; :::; zN) = C F1(z1); F2(z2); :::; FN(zN)

The joint pdf can now be obtained by taking the N-order derivative of (1)

f(z1; z2; :::; zN) =@N@z1@z2 :::@zN C..F1(z1); F2(z2); :::; FN(zN)= fp(z1; z2; :::; zN)c..F1(z1); F2(z2); :::; FN(zN)

where fp(z1; z2; :::; zN) denotes the product of the marginal pdfs ff(zn)gNn =1 and c() is the copula density weights the product distribution appropriately to incorporate dependence between the random variables. The topic on the design or selection of proper copula functions is well summarized in.

Non Linear Data Processing

In IoT applications such as multi-sensor data fusion, the optimal fusion rule can be derived from the multivariate joint distributions obtained in (2). However, it is generally mathematically intractable since the optimal rule generally involves nonlinear operations [9]. Therefore, linear data processing methods dominate the research and development, mainly for their simplicity. However, linear methods are often oversimplified to deviate the optimality. In many practical applications, nonlinear data processing significantly outperforms their linear counterparts. Kernel-based learning (KBL) provides an elegant mathematical means to construct powerful nonlinear variants of most well-known statistical linear techniques, which has recently become prevalent in many engineering applications[21].

Briefly, in KBL theory, data x in the input space X is projected onto a higher dimensional feature space F via a nonlinear mapping as follows:

: X ! F; x 7! (x):

For a given problem, one now works with the mapped data (x) 2 F instead of x 2 X. The data in the input space can be projected onto different feature spaces with different mappings. The diversity of feature spaces provides us more choices to gain better performance. Actually, without knowing the mapping explicitly, one only needs to replace the inner product operator of a linear technique with an appropriate kernel k (i.e., a positive semi-definite symmetric function),

k(xi; xj) := h(xi); (xj)iF; 8xi; xj 2 X: (4)

The most widely used kernels can be divided into two categories: projective kernels (functions of inner product,e.g., polynomial kernels) and radial kernels (functions of distance, e.g., Gaussian kernels).

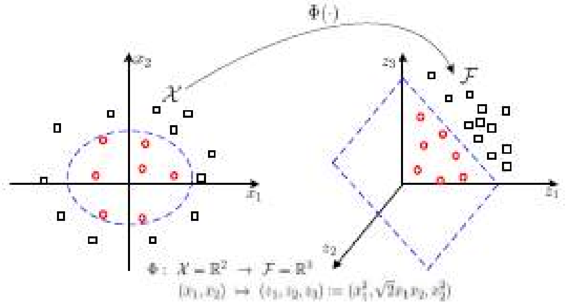

Figure 3.2: NON LINEAR DATA PROCEESING

4Fig shows An introductory binary classification example [10]. By mapping data x = (x1; x2) in 2-D input space X = R2 (left) via nonlinear mapping () onto a 3-D feature space F = R3 (right), the data become linearly separable[21].

HIGH-DIMENSIONAL DATA PROCESSING

In IoT, massive or big data always accompanies high-dimensionality. For example, images and videos observed by cameras in many IoT applications are generally very high-dimensional data, where the dimensionality of each observation is comparable to or even larger than the number of observations. Moreover, in kernel-based learning methods discussed above, the kernel function nonlinearly maps the data in the original space into a higher dimensional feature space, which transforms virtually every dataset to a high-dimensional one[21].

Mathematically, we can represent the massive data in a compact matrix form. Many practical applications have experimentally demonstrated the intrinsic low-rank property of the high-dimensional data matrix, such as the traffic matrix in large scale networks and image frame matrix in video surveillance, which is mainly due to common temporal patterns across columns or rows, and periodic behavior across time, etc.

Low-rank matrix plays a central role in large-scale data analysis and dimensionality reduction. In the following,we provide a brief tutorial on using low-rank matrix recovery and/or completion1 algorithms for high-dimensional data processing, from simple to complex.

1) Low-rank matrix recovery with dense noise and sparse anomalies: Suppose we are given a large sensing data matrix Y, and know that it may be decomposed as

1) Y = X + V;

1) where X has low-rank, and V is a perturbation/noise matrix with entry-wise non-zeros. We do not know the lowdimensional column and row space of X, not even their dimensions. To stably recover the matrix X from the sensing data matrix Y, the problem of interest can be formulated as classical principal component analysis (PCA):

1) min fXg jjXjj subject to jjY .. XjjF ";

1) where " is a noise related parameter, jj jj and jj jjF stands for the nuclear norm (i.e., the sum of the singular values) and the Frobenious norm of a matrix.Furthermore, if there are also some abnormal data A injected into the sensing data matrix Y, we have

1) Y = X + V + A

1) where A has sparse non-zero entries, which can be of arbitrary magnitude. In this case, we do not know the low-dimensional column and row space of X, not know the locations of the nonzero entries of A, and not even know how many there are. To accurately and efficiently recover the low-rank data matrix X and sparse component A, the problem of interest can be formulated as the following tractable convex optimization:

1) min fX;AgjjXjj + jjAjj1 subject to jjY .. X .. AjjF ”;

1) Matrix completion aims to recover the missing entries of a matrix, given limited number of known entries, while matrix recovery aims to recover the matrix with corrupted entries.

1) where is a positive rank-sparsity controlling parameter, and jj jj1 stands for the l1-norm (i.e., the number of nonzero entries) of a matrix.2) Joint matrix completion and matrix recovery: In practical IoT applications, it is typically difficult to acquire all entries of the sensing data matrix Y, mainly due to i) transmission loss of the sensing data from the sensors to the data center, and ii) lack of incentives for the crowdsourcers to contribute all their sensing data. In this case, the sensing data matrix eY is made up of noisy, corrupted, and incomplete observations,

eY := P (Y) = P (X + A + V);

where [M] [N] is the set of indices of the acquired entries, and P is the orthogonal projection onto the linear subspace of matrices supported on, i.e., if (m; n) 2, P(Y) = ym;n; otherwise, P(Y) = 0. To stablyrecover the low-rank and sparse components X and A, the problem can be further formulated as

min fX;Ag jjXjj + jjAjj1 subject to jj P(Y) .. P (X + A + V)jjF ”:

All the problems formulated in (6), (8), and (10) fall into the scope of convex optimization, efficient algorithms can be developed based on the results in[21].

PARALLEL AND DISTRIBUTED DATA PROCESSING

So far, all the data processing methods introduced above are in essence centralized and suitable to be implemented at a data center. However, in many practical IoT applications, where the objects in the networks are organized in an ad hoc or decentralized manner, centralized data processing will be inefficient or even impossible because of singlenode failure, limited scalability, and huge exchange overhead, etc. Now, one natural question comes into being: Is there any way to disassemble massive data into groups of small data, and transfer centralized data processing into decentralized processing among locally interconnected agents, at the price of affordable performance loss? In this subsection, we argue that alternating direction method of multipliers (ADMM), serves as a promising theoretical framework to accomplish parallel and distributed data processing. Suppose a very simple case with a IoT consisting of N interconnected smart objects. They have a common objective as follows[21]

$$\min_{\mathbf{x}} f(\mathbf{x}) = \sum_{i=1}^{N} f_i(\mathbf{x}), \tag{11}$$

where $\mathbf{x}$ is an unknown global variable and f_i refers to the term with respect to the i-th smart object. By introducing local variables $\{\mathbf{x}_i \in \mathbf{R}^n\}_{i=1}^{N}$ and a common global variable $\mathbf{z}$, the problem in (11) can be rewritten as

$$\min_{\{\mathbf{x}_1,\dots,\mathbf{x}_N,\mathbf{z}\}} \sum_{i=1}^{N} f_i(\mathbf{x}_i) \quad \text{subject to } \mathbf{x}_i = \mathbf{z}, \quad i = 1, \dots, N. \tag{12}$$

This is called the global consensus problem, since the constraint is that all the local variables should agree, i.e., be equal. The augmented Lagrangian of problem (12) can be further written as

$$L_\mu(\mathbf{x}_1, \dots, \mathbf{x}_N, \mathbf{z}, \mathbf{y}) = \sum_{i=1}^{N} \left(f_i(\mathbf{x}_i) + \mathbf{y}_i^T(\mathbf{x}_i - \mathbf{z}) + \frac{\mu}{2} \|\mathbf{x}_i - \mathbf{z}\|_F^2 \right). \tag{13}$$

The resulting ADMM algorithm directly from (13) is the following:

$$\mathbf{x}_i^{k+1} := \operatorname{argmin}_{\mathbf{x}_i} \left(f_i(\mathbf{x}_i) + \mathbf{y}_i^{kT}(\mathbf{x}_i - \mathbf{z}^k) + \frac{\mu}{2} \|\mathbf{x}_i - \mathbf{z}^k\|_F^2 \right) \tag{14}$$

$$\mathbf{z}^{k+1} := \frac{1}{N} \sum_{i=1}^{N} \left(\mathbf{x}_i^{k+1} + 1/\mu \mathbf{y}_i^k \right) \tag{15}$$

$$\mathbf{y}_i^{k+1} := \mathbf{y}_i^k + \mu(\mathbf{x}_i^{k+1} - \mathbf{z}^{k+1}). \tag{16}$$

The first and last steps are carried out independently at each smart object, while the second step is performed at a fusion center. Actually, when the smart objects are multi-hop connected, the second step can be replaced by

$$\mathbf{z}_i^{k+1} := \frac{1}{|\mathcal{N}_i|} \sum_{i \in \mathcal{N}_i} \left(\mathbf{x}_i^{k+1} + 1/\mu \mathbf{y}_i^k \right) \tag{17}$$

where Ni denotes the one-hop neighbor set of the i-th object and j j is the cardinality of a set. Eq. means that the second step can also be carried out at each smart object by fusing the local data from one-hop neighbors.This is a very intuitive algorithm to show the basic principle of ADMM[21].

Chapter 4

Application

1.Continuous Monitoring of the Internet of Things as a Source of Big Data :

Continuous monitoring is done in a broad base of disciplines: heart patients have a history of heart problems, inventory goods are monitored using RFID tags, hospital patients are managed using RFID tags,parking place availability is managed using a range of sensors, and so on. Such classic use of sensors is consistent with the historical concepts of the Internet of Things..Continuous monitoring of such data apparently meets the definitional requirements of Big Data..The volume of information deriving from the tags is substantial generating Big Data.. Because there are multiple types of information, there are a variety of different types of information available to monitor. Finally, sensor information operates in real time, speeding the velocity, resulting in a continuous monitoring environment that is classic .Big Data..[7]

2.Continuous Monitoring of Social Media as a Source of Big Data :

There also is continuous monitoring of human-based social media information for different purposes. For example, blogs and other information have been continuously monitored as a means of capturing information about a firm.s reputation and brand (e.g. Spangler et al., 2009). In these types of applications, monitoring of social-media-generated information is done continually. The analysis is based on a wide range of sources and media, and is characterized by large volumes of data, high-velocity data (e.g.millions of Twitter messages every day) and a high variety of different types of data, qualifying this type of potential application as .Big Data..[7]

3 Health Care Applications :

There are several hospitals across the world that use Hadoop to help the hospital staff work efficiently with Big Data. Without Hadoop, most patient care systems could not even imagine working with unstructured data for analysis. Big Data Analytics Using Hadoop SNJBs KBJ COE,chandwad(Nashik)[12]

Childrens Healthcare of Atlanta treats over 6,200 children in their ICU units. On average, the duration of stay in Pediatric ICU varies from a month to a year. Childrens Healthcare of Atlanta used a sensor beside the bed that helps them continuously track patient signs such as blood pressure, heartbeat and the respiratory rate. These sensors produce large chunks of data, which using legacy systems cannot be stored for more than 3 days for analysis.The main motive of Childrens Healthcare of Atlanta was to store and analyze the vital signs. If there is any change in pattern, then the hospital wanted an alert to be generated to a team of doctors and assistants. All this was successfully achieved using Hadoop ecosystem components - Hive, Flume, Sqoop, Spark, and Impala[12]

4.Smart Home :

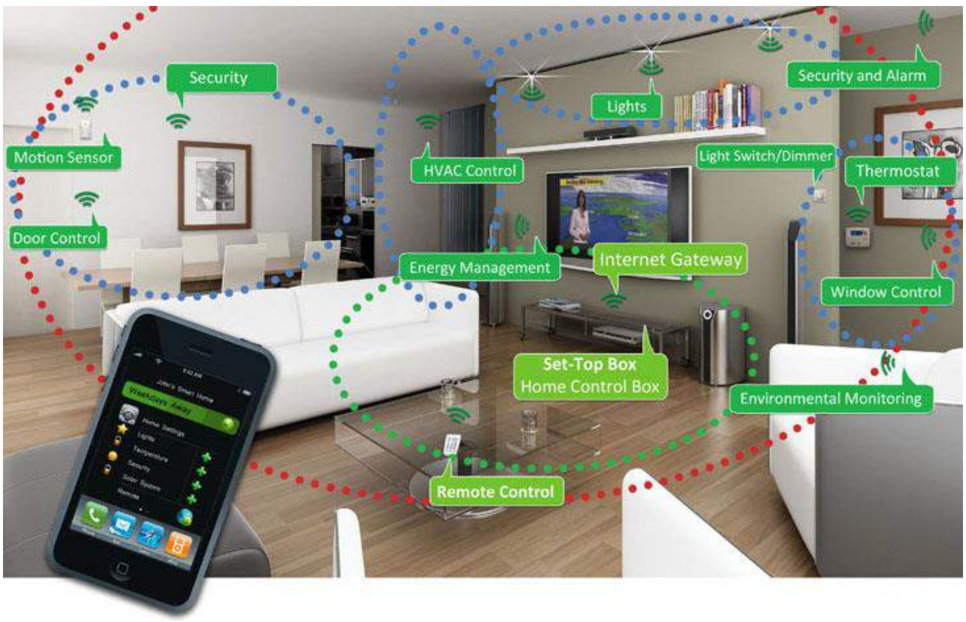

Figure 4.1: SMART HOME

Smart Home clearly stands out, ranking as highest Internet of Things application on all measured channels. More than 60,000 people currently search for the term Home. each month. This is not a surprise. The IoT Analytics company database for Smart Home

Big Data Analytics Using IOT SNJB COE includes 256 companies and startups. More companies are active in smart home than any other application in the field of IoT. The total amount of funding for Smart Home startups currently exceeds 2.5bn dollar This list includes prominent startup names such as Nest or AlertMe as well as a number of multinational corporations like Philips, Haier, or Belkin.

5 Wearables :

Wearables remains a hot topic too. As consumers await the release of Apples new smart watch in April 2015, there are plenty of other wearable innovations to be excited about: like the Sony Smart B Trainer, the Myo gesture control, or LookSee bracelet. Of all the IoT startups, wearables maker Jawbone is probably the one with the biggest funding to date. It stands at more than half a billion dollars[13].

6.Smart City :

Figure 4.2: SMART CITY

Smart city spans a wide variety of use cases, from traffic management to water distribution, to waste management, urban security and environmental monitoring. Its popularity is fueled by the fact that many Smart City solutions promise to alleviate real pains of people living in cities these days. IoT solutions in the area of Smart City solve traffic congestion problems, reduce noise and pollution and help make cities safer.

7.Smart Grid :

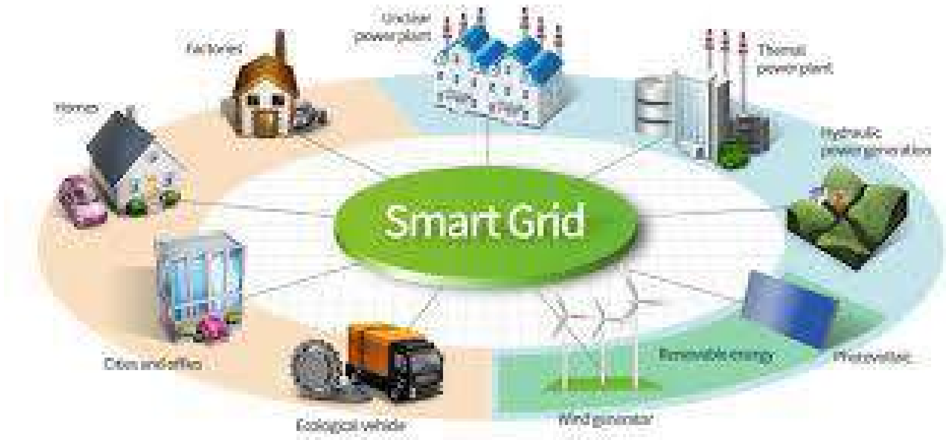

Figure 4.3: SMART GRID

Smart grids is a special one. A future smart grid promises to use information about the behaviors of electricity suppliers and consumers in an automated fashion to improve the efficiency, reliability, and economics of electricity. 41,000 monthly Google searches highlights the concepts popularity. However, the lack of tweets (Just 100 per month) shows that people dont have much to say about it[13].

8.Industrail Internet :

Figure 4.4: INDUSTRIAL INTERNET

The industrial internet is also one of the special Internet of Things applications. While many market researches such as Gartner or Cisco see the industrial internet as the IoT concept with the highest overall potential, its popularity currently doesnt reach the masses like smart home or wearables do. The industrial internet however has a lot going for it.

The industrial internet gets the biggest push of people on Twitter (1,700 tweets per month) compared to other non-consumer-oriented IoT concepts[13].

9.Connected Car :

Figure 4.5: CONNECTED CAR

The connected car is coming up slowly. Owing to the fact that the development cycles in the automotive industry typically take 2-4 years, we havent seen much buzz around the connected car yet. But it seems we are getting there. Most large auto makers as well as some brave startups are working on connected car solutions. And if the BMWs and Fords of this world dont present the next generation internet connected car soon, other well-known giants will: Google, Microsoft, and Apple have all announced connected car platforms[13].

10.Connected Health (Digital health/Telehealth/Telemedicine):

Connected health remains the sleeping giant of the Internet of Things applications. The concept of a connected health care system and smart medical devices bears enormous potential (see our analysis of market segments), not just for companies also for the well-being of people in general. Yet, Connected Health has not reached the masses yet. Prominent use cases and large-scale startup successes are still to be seen. Might 2015 bring the breakthrough[13].

Figure 4.6: CONNECTED HEALTH

Chapter 5

Advantages And Disadvantages

5.0.1 Advantages of IOT

1.Data : The more the information, the easier it is to make the right decision. Knowing what to get from the grocery while you are out, without having to check on your own, not only saves time but is convenient as well.

2. Tracking : The computers keep a track both on the quality and the viability of things at home. Knowing the expiration date of products before one consumes them improves safety and quality of life. Also, you will never run out of anything when you need it at the last moment.

3. Time : The amount of time saved in monitoring and the number of trips done otherwise would be tremendous.

4. Money : The financial aspect is the best advantage. This technology could replace humans who are in charge of monitoring and maintaining supplies[19].

5.0.2 Disadvantages of IOT

1. Compatibility : As of now, there is no standard for tagging and monitoring with sensors. A uniform concept like the USB or Bluetooth is required which should not be that difficult to do.

2. Complexity : There are several opportunities for failure with complex systems. For example, both you and your spouse may receive messages that the milk is over and both of you may end up buying the same. That leaves you with double the quantity required. Or there is a software bug causing the printer to order ink multiple times when it requires a single cartridge.

3. Privacy/Security : Privacy is a big issue with IoT. All the data must be encrypted so that data about your financial status or how much milk you consume isnt common knowledge at the work place or with your friends.

4. Safety : There is a chance that the software can be hacked and your personal information misused. The possibilities are endless. Your prescription being changed or your account details being hacked could put you at risk. Hence, all the safety risks become the consumers responsibility[19].

5.0.3 Advantages of Big Data Analatyics

1.Usability : All cloud storage services reviewed in this topic have desktop folders for Macs and PCs. This allows users to drag and drop files between the cloud storage and their local storage.

2. Bandwidth : You can avoid emailing files to individuals and instead send a web link to recipients through your email.

3. Accesibility : Stored files can be accessed from anywhere via Internet connection.

4. Disaster Recovery : It is highly recommended that businesses have an emergency backup plan ready in the case of an emergency. Cloud storage can be used as a back-up plan by businesses by providing a second copy of important files. These files are stored at a remote location and can be accessed through an internet connection[20].

5.0.4 Disadvantages of Big Data Analytics

1. Data Security : There are concerns with the safety and privacy of important data stored remotely. The possibility of private data commingling with other organizations makes some businesses uneasy. If you want to know more about those issues that govern data security and privacy, here is an interesting article on the recent privacy debates.

2. low Bandwidth : Several cloud storage services have a specific bandwidth allowance. If an organization surpasses the given allowance, the additional charges could be significant. However, some providers allow unlimited bandwidth. This is a factor that companies should consider when looking at a cloud storage provider[20].

Chapter 6

Conclusion

The potential of the IoT appears to be great, despite the range of issues that need to be Addressed. This paper has sought to highlight the IoT concept in general and big data analytics . Based on above ,It can be considered that new research problems arise due to the large scale of devices, the connection of the physical and internet worlds , the openness of the systems of systems, and continuing problems of privacy and security.[6]

Chapter 7

Bibliography

[1] https://www.researchgate.net/publication/258566915_Big_Data_Analytics_in_Future_IOT

[2] http://www.slideshare.net/MohanK umarG/ $internetofthings - iotaseminar - ppt - by - mohankumarg$

[3] http://www.webopedia.com/TERM/ B/big_data_analytics.html

[4] http://www.datamation.com/applications/$why - big - data - and - theinternet - of - things - are - a - perfectmatch.html$

[5] http://www.ibmbigdatahub.com/bl og/internet-things-and-evolvingrole-big-data-analytics-insighteconomy

[6] http://www.ijarcce.com/upload/201 6/march-16/IJARCCE

[7] O'Leary, Daniel E. "BIG DATA, THE =INTERNET OF THINGSAND THE =INTERNET OF SIGNS." Intelligent Systems in Accounting, Finance and Management 20.1 (2013): 53-65.

[8] http://www.huffingtonpost.com/ $philip-kushmaro/the-iot-and-big-damaki_{b1}2116608.html$

[9] http://searchbusinessanalytics.techtarget.com/essentialguide/ $Guide - to - big - data - analytics - tools - trends - and - best - practices$

[10] https://en.wikipedia.org/wiki/Internet_of_things#Architecture Big Data Analytics Using IOT SNJBCOE

[11] http://www.datamation.com/applications/why-big-data-and-the-internet-of-things-are-a-perfect-match.html

[12] https://www.dezyre.com/article/5-healthcare-applications-of-hadoop-and-big-data/85

[13] https://iot-analytics.com/10-internet-of-things-applications/

[14] http://ieeexplore.ieee.org/document/7389091/

[15] http://www.ijaetmas.com/wp-content/uploads/2016/05/IJ16M0506.pdf

[16] http://www.infoworld.com/article/2616959/big-data/7-top-tools-for-taming-big-data.html

[17] http://www.kdnuggets.com/2016/07/open-source-tools-internet-things.html

[18] http://data-informed.com/how-the-internet-of-things-changes-big-data-analytics/

[19] https://e27.co/advantages-disadvantages-internet-things-20160615/

[20] http:// $bigdata - madesimple.com/5 - advantages - and - disadvantages - of - cloud - storage/$

[21] Ding, Guoru, Long Wang, and Qihui Wu. "Big data analytics in future internet of things." arXiv preprint arXiv:1311.4112 (2013).