A INTROSPECTIVE STUDY ON DATA CENTRIC VIEW OF INTERNET OF THINGS: CHALLENGES AND OPPORTUNITIES

[1]S.S. Boomiga, [2]Dr.V.Prasanna Venkatesan
[1]Research Scholar, [2] Professor
[1]Department of Banking Technology,
[1]Pondicherry University, Puducherry, India

Abstract : IoT is a concept where Things which are uniquely identifiable or addressable connected to Internet that has the ability to send and receive the data via the network based on standard communication protocols. The number of connected smart things is growing remarkably, and it is anticipated to go beyond 26 billion by 2020.Thus the introduction of IoT results in enormous amount of data. Such data needs to be handled in real-time and the processing may be extremely distributed in nature. The objective paper is to study about the challenges in the IoT data which is generated in great volume, velocity and variety in which traditional processing methods and tools has limitations to apply.. This paper reviews the main techniques and current research efforts in IoT from data-centric perspectives. This paper includes the study of data processing steps such as data aggregation, data analysis with various data mining and machine learning algorithms and data visualization in IoT. From the study of various research work a data management framework is also presented to show how data can be handled in IoT environment. Open research challenges for IoT data management are also discussed.

IndexTerms **– internet of things, data processing, aggregation, storage, analytics.**

I. INTRODUCTION

The term "Internet of Things" (IoT) was first used in 1999 by Kevin Ashton in the context of supply chain management [1]. The smart devices in the real world are linked to the Internet by sensors in IoT. This was demonstrated by Aston by using RFID tags in supply chain goods are connected to Internet to count and track goods without the help of human involvement. Nowadays, the Internet of things has become a common term in which the devices, sensors and everyday items can be accessed from anywhere in the world [2].

The Internet of Things (IoT) consist of a network of connected physical objects that can be accessed through the Internet. The connected physical objects are sensors, RFID, that enable objects to sense and communicate [3]. So the IoT refers to uniquely addressable objects that are connected to an Internet-like structure. In order to reach common goal, a variety of physical objects through unique addressing schemes communicate and cooperate with each other using detection technology, internet technology, intelligent computing technology etc. [4]. So things which are uniquely identifiable or addressable connected to Internet that has the ability to send and receive the data via the network based on standard communication protocols.

The data generated by the IoT environment permits people and things to be linked anytime, anywhere with anything and anyone, using Internet- like structure and any service. In order to provide more useful information to the IoT users such as sending alert messages and information, wireless sensor networks and actuators are used. So the data generated by the things must be stored, analyzed [6].

IDC states that, by 2020, the digital data produced by the universe will from 4.4 zettabytes to 44 zettabytes [7]. A Cisco report states that, by 2018, for every year 400 zettabytes of IoT data will be generated [8]. Huge amount of data sets are generated with the growth of Internet, IoT and cloud computing technology in recent years. So for every two years overall information will be doubled as per International Data Corporation IDC2013.The amount of generated is more but does not provide more useful information, so in order to provide more valuable information the data generated must be analyzed quickly and efficiently having high processing capability [9].

Internet of things data are characterized by volume, velocity, variety, veracity and the data generated by IoT is big data in which traditional data processing methods and tools has limitations to apply. New technologies and efficient machine learning and deep learning algorithms are required to process the IoT data in order to bring valuable insights from the data generated from IoT environment. The data generated must undergo various stages during its life cycle including collection, storage and organization, exploration, and presentation of data [10]. The technologies of each data processing steps includes data collection, data aggregation, data storage, and data analysis and data visualization and these concepts are discussed to understand about how data is processed in various stages.

The remainder of the article is organized as follows. Section 2 identifies the sources of data from the IoT environment. Section 3 explains about the IoT data Nomenclature such as IoT data features, data quality and data types. Section 4 reviews the need for aggregation and aggregation techniques and Section 5 discusses about data preprocessing methods and Section 6 focuses on the data models and storage technologies for IoT. Data mining and machine learning algorithms are discussed in Sections 7. Section 8 focuses on need for data visualization. Section 9 proposes a data management framework. Section 10 highlights some major challenges on IoT from the data perspective. Finally, Section 11 offers some concluding remarks.

II.SOURCES OF DATA

The "Thing" in the IoT is the embarkation for an IoT solution. It is characteristically the source of the data. Data are from things. The things generate data for the IoT system. So the following definition contributes the various representations of things. Things can be tangible or intangible objects that have unique identification which is used to recognize the information around the objects and to deliver status of the object [11][12] [13] [14].

Tangible objects can be sensors, RFID tags, GPS, mobile devices, laptops, embedded chips, vehicles, ships, and any apparatus, refrigerators, TVs, vehicles, clothes, food, medicines, books or other devices[15][16]. Intangible objects can be computational processes, software, services, database, data items, data stores, web objects, documents, digital objects. For example, URLs (Universal Resource Locators)are used to identify Web services, while DOIs (Digital Object Identifiers) are used to access documents and other digital objects [17][18].

Thus things are physical or virtual object containing sensors and embedded software to communicate with the external environment [19].Thus the IoT devices collects useful data from the things with the help of various existing technologies.

III. IOT DATA NOMENCLATURE

In this section, the fundamental characteristics of IoT data are recognized and they are classified into three categories, such as Data Features, Data Quality, and Data types. The specific characteristics of each category are identified, and the overall IoT data taxonomy is shown in Fig. 1

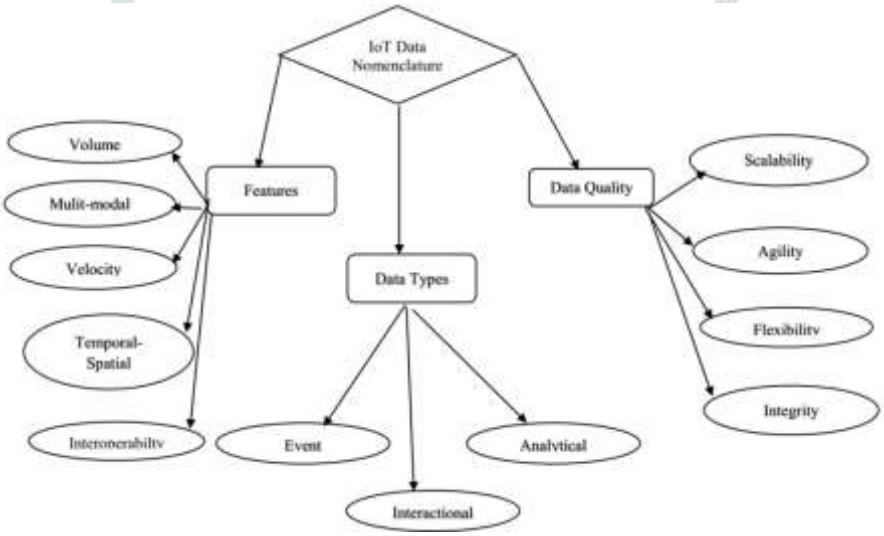

Fig 1. IoT Data Nomenclature

3.1 Features of Data

The IoT is a composite system with a different number of characteristics. Its characteristics may differ from one domain to another. Some of the general and key characteristics of IoT data are as follows [20, 21]:

1. Volume: Large amount of data are produced by the IoT environment. The data generated needs more storage, bandwidth to transmit and adequate processing to get valuable information. Wireless senor networks require large buffer to store the data and transmission of these data reduce the lifetime of the battery power of sensors. In case of in – network processing in sensor networks will lead poor latency.
2. Multidimensional: The IoT data that are composed from many different bases such as Sensor data, RFID which will be having different formats. Several sensors such as temperature, humidity, light, pressure monitors the IoT environments and these data are used by the IoT applications. So an application integrates multi-dimensional data which are attached to a wide array of devices and objects.
3. Velocity: IoT data can be generated rapidly. The flow of data is massive and continuous. The sensor in IoT environment, RFID readers, and other devices generates data fast at a constant rate which must be processed efficiently.
4. Temporal-spatial correlation: In order to describe the dynamic changes of IoT environment, the data from the Internet of things must be collected with respect to dense of sensor objects location over time.
5. Interoperable: IoT data is observed by different numbers of devices RFID readers, cameras, and temperature sensors which has different format and semantics. The data collected has to be collaborated between different applications, which need data sharing

between the applications for the user to get IoT solution. For example, when a patient is in emergency situation, a telemedicine application needs traffic data to in order to get the traffic condition, so that the ambulance can be reached to the hospital in time.
6. Veracity: Some sensors are of low cost which has low accuracy and some data may be sent over unreliable transport which causes data will be corrupted or arrive out of order, be late or missing.

3.2 Data Quality

1. Scalability: The data is generated from various large states of heterogeneous devices including sensors networks, RFID data. More number of devices can be added in a network to enhance the quality of the data by applying data fusion.
2. Agility: IoT systems must deploy real time rules based intelligence at every part of data lifecycle, thereby delivering rapid decisions and actions for an application. IoT solutions must have the ability to understand cause, effects, faults and performance. These ability are obtained from collecting the IoT data, identifying any faults from the sensor data, and issue remote command to the user and the IoT device to rectify the problems occurred in the IoT environment. Thus the IoT data collected must be explored to make quick decisions.
3. Flexibility: IoT application integrates many types of sensors, RFID data ,and command data issued to the actuators .Thus all forms of data are handled by the IoT Systems including both structured and unstructured data.
4. Inconsistency: Since IoT data come from different data sources such as sensors, RFID readers or GPS, inconsistencies exist in IoT data. In case of RFID data, inconsistencies occur because of missing readings of tags at some locations. In case of sensors readings, multiple sensors will monitor the same environment and discover different sensor result. Thus inconsistency leads to inaccuracy and poor precision of data value in IoT data.

3.3 Data types

The IoT data are discovered from different sources, some of these data are discrete in nature and some are continuous and some data are input by humans. These data can be classified as follows
1. Event data: This type of data deals with the measurement of the environment through sensors and RFID when an event occurs. It consists of the data that comes from the 'things' themselves – measures from sensors such as temperature, humidity, acceleration, vibration, speed, video feeds and milliseconds, so there is a high frequency of data creation.
2. Interaction data: The IoT will be used to regulate remote devices. The data discovered from the IoT environment is used to control devices by issuing commands. These data occurs when any two objects interact or interaction between a human and an object. This can be of adjustment or alteration of any value in a device or machine. This type of data occurs less frequently than event data, but it is somewhat more complex.

IV.AGGREGATION

When raw data is transmitted through the network, it needs more energy consumption to transmit the generated data from the thing. So aggregation functions are used to reduce the data size. Thus data aggregation approaches allow coping with huge volumes of data and reducing the size of real-time data streams. The data generated by the IoT environment is large and it leads to more transmission time to transfer the data to the cloud. By aggregating the data, the transmission time and network lifetime can be reduced.

Rongxing et al[22] employed the homomorphic modified Paillier encryption, Chinese Remainder Theorem, and one-way hash chain techniques to aggregate the data . The data from various IoT devices are collected, and passed to the control center where the false injected data is filtered at network edge. Chen et al[23] employed Boneh-Goh- Nissim (BGN) homomorphic encryption to aggregate multifunctional data, supported by average, variance, and one-way ANOVA aggregation techniques.

The aggregation function must be performed closer to the data sources to reduce the communication cost. Thus the data are collected and summarized from various homogeneous or heterogeneous things. So the aggregation points are deployed to improve the efficiency of the IoT system and it should not embrace delay which affects the systems real time performance [5].

The preprocessing for aggregation can be of two types [24]

1. Signal Preprocessing
2. Mathematical/Statistical Preprocessing

4.1. Signal Preprocessing:

An algorithm or hardware circuit is used to filter the unwanted parts of the signal. Thus in the frequency domain the unwanted signal is removed by cutting the signal after or before certain threshold frequency. This method helps to reduce the data size by removing background noise and concentrates on focused data set for further processing. But the problem behind this is sometime data can be missing or outlier data will exist.

4.2. Mathematical/Statistical Preprocessing:

Mathematical preprocessing techniques do not concentrate on the frequency domain, but works on the output produced. The data is aggregated over time by using data windows. Thus data is transmitted over a certain time to the gateway for further processing or to distribute over the network. The following are the various mathematical function used for extracting data over time.

a) Min, max: A minimum and maximum value is set down .The difference between these two values over a sample window are considered for data extraction.

b) Mean, median: In this method last n values are considered over a time window to extract the data.
c) Variance, standard deviation: Variance and standard deviation are used to find the unstableness in the data.
d) Correlation, integration: Correlation and integration are used when data are in multidimensional.

The ability to adapt the algorithms that are used to filter and aggregate data at the edge of the network will be essential for some sensor networks.

V.DATA PREPROCESSING

Internet of things data are generated in high speed generating more data and this vast data must be managed by determining what data can be discarded at the point of collection before it is sent on for processing. For effective processing of the data, the nodes at the outer edge of the network have to do a lot of processing and aggregation before transmitting the results to a central host. This is the place where intelligence is placed out at the edge of the network to filter and manage the flow of machine data .This is a complex process to determine which data is essential that reflects the data needed for an application [25].

The data generated from various sensors from different locations are gathered results in enormous amount and they are in different formats which may be structured and unstructured data. Thus the data collected are redundant and unreliable having much useful information. These data sets require a huge storage and do not fit into traditional databases and it must be preprocessed for storage. Thus the datasets are in raw form with inconsistency and redundancy with much useless information. improve the computational efficiency and to avoid excessive data storage, the generated data must be preprocessed [26].

5.1 Data processing methods

Effective data processing methods must be adopted to challenge the massive data of the system. Some of the data processing methods are discussed below to reduce the data sets size.

5.1.1. Protocol conversion.

The IoT devices which are used to collect the data and operate the machines in the system are manufactured at different ages and from different vendors. So the data collected from various things will be in different protocols and standards and these increase the trouble in operating and maintaining the system.
The solution to this problem is to provide an intelligent gateway to convert the data from several different devices with different protocols, and aggregate overall data by using a standard protocol. In industrial networks Modbus/TCP protocol is used for protocol conversion [27].

5.1.2. Cleaning

Data collected from the IoT environment is normally incomplete, inaccurate, incorrect data of noise or even invalid. Data cleaning is a valuable process that helps to increase the accuracy and efficiency of the data by filling missing values, replacing, modifying, or deleting the dirty or coarse data and smoothing the noise.

The algorithm is effective and simple to treat the noisy data, when the enormous amount of data received frequently at the reception node. In case of missing values, the value is filled with infinite "- ∞" or "unknown" values; otherwise, average of last N values can be used to fill in the missing values. To clean noise, the method named binning can be used, in this method the neighborhoods values are sorted and divided into partitions called bins and then smooth the noise by bin means, bin median and bin boundaries. The methods described are simple, easy to understand and effective [28]. The other methods used to remove noise are clustering, regression combined human and computer inspection.

5.1.3. Data filtering

Filtering refers to the process of reducing noise content from raw data. This means the data needed by the user are defined and the errors are detected and corrected in the given data, to minimize the impact of errors in input data for succeeding analyses. The filters are presented as mathematical formulas or pseudo code so that they can be implemented in a language of choice [27].
Data collected from sensor is very large and redundant and inaccurate. If the unwanted data is not removed the following problem occurs.
a) The network bandwidth will be increased, since the whole data generated from sensors must be transmitted.
b) The data processor workload will be increased, because the processer has to process large number of data.
c) The data storage will be increased, because more unwanted data has to be stored in database.

5.1.4. Data conversion

In Application Layer, different operating system and languages are used; it is difficult to integrate with the system and application, because the data is collected from different sources, at different times and at different globally distributed locations so the data will be of different form and data types. The data collected can be converted into a single format for further processing. XML technology can be used; the sensor node information is converted into XML format [27].

5.1.5. Data Compression

The major challenge in the progress of Internet of Things applications is management of large sensor data. This large volume of data leads to data compression technique in order to deal with the problem of energy-effective usage of transmission,

reducing storage space for tiny sensor devices, and cost-effective sensor analytics. SensCompr is a technique that extracts the useful information from sensor data and adapts the parameters like threshold selection, block-size estimation for Chebyshev compression to yield maximum compression gain while sacrificing irrelevant information loss. Compression is performed in smart meter, healthcare, transportation datasets [29].

5.1.6. Data Fusion

The data collected from various sensors could be incorrect because of various factors in the IoT environment. This is because a sensor node itself collects incorrect data due to failure, geographically distributed sensor data and time based data. Sometimes the neighborhood sensors often generate identical and highly co-related data. Sometimes environmental factors such as pressure, temperature, electromagnetic noise in the monitored area might affect with sensor node reading which could lead to inaccurate readings.

So data fusion is a process of collecting data from multiple sensors and related information. Then the related data are combined and mined to produce more accurate data by removing incorrect and duplicate data. Thus data fusion is a process that consists of algorithms and methods for integrating multisource sensor heterogeneous data for achieving improved accuracy and more specific inferences than that obtained by using only a single sensor [30].

VI. DATA STORAGE

IoT faces a series of challenges with respect to data storage and processing like enormous amount of data generated by sensors data, integrating all IoT data coming from different IoT environment with different format and type, IoT devices generating data rapidly, complicated requirements of data management, etc. A picture-perfect data storage solution is needed to handle the rapidly increasing large voluminous data efficiently.

The data generated by IoT devices, comes in two distinct types [21]:
Unstructured data: Unstructured data does not have recognized or predefined structure. It is unorganized raw data which can be either textual or non-textual. It may be images or videos that are generated from smart phone, camera or other devices. This type of data can be accessed via file name either sequentially or in alphabetical order
Structured: Structured data is highly organized information, which can be stored in a database. These data are like small log- file are captured from sensors. These data are smaller in size and billions of files can create by sensors and accessed.

The data in the IoT environment is earnest if they can be recognized. For example, the sensor data from the IoT environment must be stored with other information such time and geographic locations otherwise it is not valuable. The data generated needs timestamp and the EPC code of the device to identify which device has generated the data [21].

There data generated can be stored in three forms: local, distributed and centralized. [5]
1. Local form: The data is stored in the storage unit of sensor. Flash memories and embedded platforms are installed in sensors to store the data locally. A database management system, StoneDB, form a database to store the sensors data to support queries and mining tasks.
2. Distributed form: The data is stored in more than one server in a network through distributed technologies and dynamically add new attributes to a data record. In a distributed storage system, there exist three types of storage. It can be through block, file and object.
3. Centralized form: The centralized form refers to that the data from the sensors are collected and then it is stored in cloud.

The IoT data can be stored as object storage or Software defined storage for efficient accessing. The IoT data can be stored in cloud platforms using RDBMS, NOSQL DBMS, DBMS based on HDFS, main-memory DBMS, and graph DBMS[21]. The structured data can be stored in the multiple databases such as NoSQL and MongoDB and unstructured data can be stored in file repository (HDFS). Two major issues must be addressed regarding the storage of IoT data: the Location of data storage facilities (the where), and the organization of data storage (the how).

VII. DATA ANALYTICS

Data analytics is the process of converting data into information and knowledge. Depending on the requirement of services, the data collected is analyzed to take specific actions. Today, IoT brings the great research for managing, analyzing and mining data. The data collected from the IoT environment enables to understand the complex environments and used to make better decision, automation, high efficiency, productivity and accuracy.

IoT Data Analytics helps to create valuable information in IoT. The data such as structured, semi-structured, live streaming and historic data collected from IoT devices are analyzed to create specific action for an application to increase the performance of the application .The data generated by the IoT environment is very large and it leads to vast data storage, security problems to store in cloud and greater analytic challenges. To find out the information hidden in the IoT data, data mining algorithms, machine learning algorithms and deep learning algorithms are needed to provide possible solutions.

To produce highly reliable and accurate results, timely analyses of IoT data are a major issue. Data mining and Machine learning algorithm are the best to produce hidden information from IoT data [31].

7.1 Data Mining

Data mining aims to determine the solution that are present in the generated IoT data and thereby increase the efficiency of the system and quality of services [32]. There are various processes and algorithms in data mining, so to select a particular algorithm for a particular IoT system is also a challenge now. Data mining in simple terms can be said as the process of extracting valuable or sensible information from a data warehouse. Effective and efficient mining algorithms are required to mine IoT data streams that are highly dynamic, heterogeneous, uncertain, indefinite and incomplete set. The following are some of the data mining algorithms.

7.1.1. Clustering

Clustering is typically defined as categorizing the data into some sensible, meaningful groups or classes. This helps to achieve an easy perceptive for the users by grouping naturally. The best example for this could be a search engine which is based on clustering, that can categorize endless web pages into news, images, videos, reviews etc. Clustering is an unsupervised learning process. There are various clustering models such as k-Means clustering, k-Medoids clustering, Density based clustering and Hierarchical clustering that can be used depending upon their use.

Clustering methods are divided into 4 major categories such as: [33]

1. Partitioning methods,
2. Hierarchical methods,
3. Density based methods and
4. Grid based methods

7.1.2. Classification

It is a function of data mining that delegates items into categorical labels. It helps us to predict the category of a particular item in a dataset [34]. Classification is a supervised learning process. This classification algorithm can be implemented on different types of data sets and on basis of performance these algorithm also used to detect the natural disasters like cloud bursting, earth quake, etc. This technique is applied for car parking, accident detection, and traffic forecasting services.

7.1.3. Frequent pattern mining

The basic idea of frequent pattern mining is to find patterns such as set of items, subsequences, substructures that occurs frequently in a data set [32].Finding frequent pattern plays a vital role in mining associations, correlations among data. It is a basis for many data mining tasks such as indexing, classification, and clustering, sequential and structural pattern. It helps to find out the hidden information from a set of transactions in a database .Haoshu et al proposed a system to detect frequent trajectory patterns to investigate the production process in manufacturing system by deploying sensors in the manufacturing system [33].

7.1.4. Time Series Analysis

A time series is collection of time-based data objects and it's obtained from scientific and financial applications. The features of time series data include huge data size, high dimensionality, and update continuously. Deep learning algorithms could apply to IoT and Smart city domains in time series analysis [34]

Stock market index value is analyzed in a time series manner. Time series analysis is also used in forecasting, to analyze dependent events; that is to predict future values based on past events. When data points are present in consecutive time interval, time series analysis is applied to extract meaningful related to specific patterns or statistics [35]

7.1.5. Outlier and Anomaly Detection.

The data points which are entirely different from the remaining points in a given data set is called as outlier. The outlier can be obtained based on the distance between the points in the data set. The data points that are most distant from all points will be marked as outliers[36].Paper [37] proposes a model to detect outlier based on sensors placed on different geographical location and sensors reading taken at different time in air pollution data set for a smart city. Zibin[38] etal proposes an algorithm to detect anomaly using support Vector Machine for smart traffic data.

7.2 Machine Learning

Machine learning is a concept in which the machine learns from the programmed code without human intervention. Machine Learning (ML) algorithms can be classified in two groups, supervised and unsupervised techniques. Supervised techniques need a prior knowledge in order to classify the rest of samples. Because the environment completely changes every day in a RFID context (i.e. new objects around the system), the training set may not work for future uses of the system. Hence, for real-time applications it can be a disadvantage since the system requires a previous normalization before obtaining a good classification. However, unsupervised techniques (which do not require training) are suitable in this scenario where the classification does not depend on previous training sets. Algorithms or mathematics plays the most essential role in machine learning; this is the tool to deal with the data. Some of the machine learning algorithms is discussed below.

7.2.1. Bayesian Statistics

Bayesian statistics is a mathematical procedure that applies probabilities to statistical problems. It provides people the tools to update their beliefs in the confirmation of new data.

The data sensed by the sensor are enormous and all the data generated are not important, BP(Belief Propagation),handles these sensor data and controls the sending of useless data [39]. The data sensed must be of high quality to provide good result and

solution. When all the sensors sense data it leads more generation of data and high energy consumption. Lev [41] used multiphase adaptive sensing algorithm with belief propagation protocol (ASBP) to turn on only a smaller number of sensors thereby reducing the energy consumed by the sensor network providing with high data quality.

7.2.2. k-Nearest Neighbors (k-NN)

The k-NN algorithm is a supervised learning algorithm. This is a general classification algorithm which groups k kinds of groups so that the distance between the points inside a group is minimum [39].This algorithm helps to identify missing data by exploiting several neighbour nodes cooperatively by considerering spatial correlation of sensor data than temporal correlation[39] [40]. This algorithm is applicable to the IoT data which is designed for applying in fast online parallel multisensory information processing and Change detection algorithm in multisensory environment [41]. Based on similarity measure, this algorithm stores all available clusters and classifies new clusters.

7.2.3. Neural Networks

Neural network is a concept to process the information. A neural network consists of several hidden layers which helps to find the difference between objects of different groups by training. Knowledge for training is provided by means of input examples called training examples [42].

In case of locating a sensor node (i.e to find out node geographical position) neural networks can be used. Depending upon the frequency signal received from neighbor nodes and the spatial relation between the nodes, a node can be localized. The measurements may include received signal strength indicator (RSSI), time of arrival (TOA), and time difference of arrival (TDOA).By giving multiple training, the location of the node cab be computed by neurons [39].

7.2.4. Support Vector Machines (SVM)

It is supervised machine learning algorithm which uses labeled training model to categorize the data points by learning. So it mainly categorize the data points into two regions separated by a plane and new data points will be placed in one of the two regions[40]. Machine learning techniques like ANN, SVM and multiclass SVM are used to detect abnormal data points using Air quality index [42].In another application, anomaly detection is used to identify the interesting parking locations [43].

7.2.5. Decision Tree (DT)

Decision Tree algorithm is a supervised learning algorithm which is used for building classification or regression model in the form of tree structure for a problem. This algorithm is more appropriate for large data sets. A decision tree is a hierarchical tree structure, in which it breaks dataset into smaller and smaller subsets. Finally the tree structure is developed with nodes. The nodes can be root node, leaf nodes or decision nodes. A decision node can have two or more branches, which perform a test on single value .Leaf node is used to make decision or prediction. The top most node is the root node which is used for best prediction. Thus each node represents assessment on an attribute value, where the outcome of the test is represented by branch and tree leaves are various classes and thus decision tree represents a tree like structure. This algorithm is used in identifying six activities (watching, reading, chatting, sleeping, listening and walking) using wearable's[44].

7.2.6. Principle Component Analysis (PCA)

Principal component analysis aims to analyze the data to identify patterns and finds to reduce the dimensions (for example reduce 2D to 1D, 3D to 2D) of the data set with less loss of information. Thus the necessary information are mined from the data collected and it generates new set of variables called principal components. It can be used for compression, event detection and event recognition. It can be used to deal with high dimensional data. So this algorithm is more applicable for processing data from sensor networks [45].

7.2.7. k-Means Algorithms

k-means algorithm is unsupervised learning algorithm , which solves k-means clustering problem. The algorithm aims to discover clusters in the input data points. The amount of clusters are represented by a variable K. Based on the feature provided, each data point are assigned to one of the K group by working iteratively. The clusters are formed based on feature similarity. This algorithm can be applied to behavioral segmentation, inventory categorization, sorting sensor measurements and detecting anomalies. The advantage of this algorithm is simple implementation and it has linear complexity. This algorithm is used for data analysis based on MapReduce[46]. In [47] the prediction of gestures is achieved by the standard unsupervised machine learning technique K-means clustering.

VIII. DATA VISUALIZATION

Visualization is the only technique which is used to quickly identify the problem and based upon that decision can be made. The data collected from the IoT environment is huge and the information is hidden and they can be represented in a meaningful way by data visualization. Thus data visualization helps to make fast and quick decision with more confident and accurate. Data visualization is not about graphs or charts which shows something interest about the data collected or analyzed .It deals about understanding of data, analyzing data and communicating the data within billions of IoT devices. Any IoT application datasets can be tracked and analyzed with the use of powerful and simple data visualization. Data Visualization is becoming an integral part of IoT, when the analyzed data are translated into a language that is easy to understand process and present on visual language.

IX. DATA MANAGEMENT FRAMEWORK

IoT data management system can be divided into various layers as shown in Fig 2. Layer1 is the Things Layer which works together directly with the interconnected IoT devices and sensors. Data produced by sensors is collected and data is sent to Aggregation layer. Layer 2 is the aggregation layer which is responsible for the process in which the collected data are efficiently summarized the huge volumes of data in real-time. Layer 3 involves Data Pre-processing. IoT data will be in various formats and structures, since Things layers may interact to collect the data with various sources like sensors, RFID readers, GPS, camera etc.

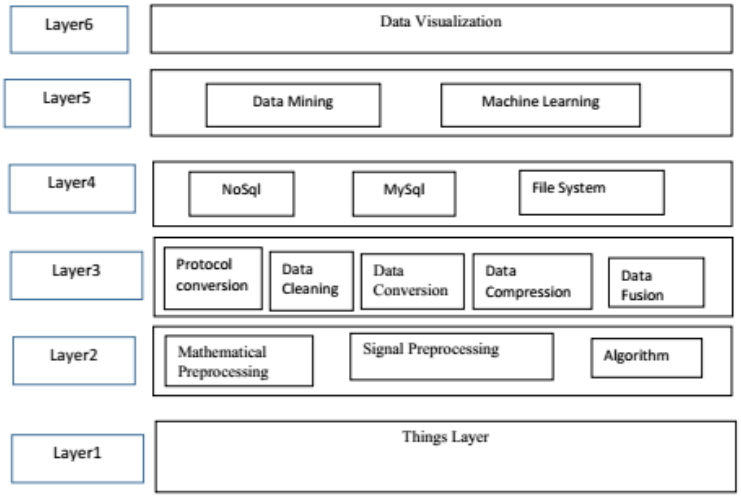

Fig 2 . IoT Data Processing Framework

The IoT data must be stored in unified format, so redundant and missing data are processed and converted into unified format. In order to decrease communication cost, the data compressed in this Layer .Sometime data from the multiple sensors provides duplicate information, so in order to provide precise information a technique called data fusion is applied to the sensed data. Layer 4 is storage-layer which involves the mass storage of produced data to specific data base management systems for later processing and analysis and for query and use. Layer 5 is the Analytic Layer which is responsible to apply data mining techniques or predictive machine learning algorithms to find the hidden information within the pre -processed data. Layer 6 relates with data visualization which is used to quickly identify the problem and based upon that decision can be made.

X. MAJOR CHALLENGES

Challenge #1Need for edge Processing

Raw data generated by the sensors are transported to the cloud which requires high band width to transmit and high storage cost to store the data in cloud. So the data is handled at the edge node or server to reduce the data size and sent to the cloud.

The edge server is similar to cloud server, but its capability is lesser than cloud server. Edge sever can perform computation and data storage and it communicates with the other network by having internet connection. Devices such as base stations, gateways, access points, routers, switches, and video surveillance cameras are some of the edge computing nodes [27]. To reduce communication cost and storage cost, the data must be processed within the edge network.

In the existing cloud computing networks, a middle layer can be built called as edge layer. This edge layer is made up of several edge servers. Edge servers may be available in remote locations or other edge locations. But they are generally available where the sensors reside closer to the sensors. The major advantage of edge processing is to handle the tasks that needs to process immediately without delay, in order to give quick feedback to the end-users [48].

Challenge #2 Distribution of IoT tasks between edge, cloud and things

Edge sever has low computing power and storage capabilities, and it cannot handle huge amount of data like cloud server, but it must be used effectively. So necessary actions must be taken to identify which activities must continue on edge, which activity can be processed by the things or IoT devices and identify other remaining activities which can be moved to cloud. Certain data prioritization techniques can be used to find which IoT data needs to be processed closer to the edge in order to yield reply to the end users with less latency. Other remaining IoT data can be moved to the cloud for further processing and storage.
So while taking the number of IoT jobs requested or created suitable data classification, data prioritization, task scheduling, task and data offloading with limited bandwidth, delay, power consumption must be considered [27].

Challenge #3Security

In IoT things are given unique addresses, and it has the facility to transmit the data over the network. Thus lots of data are collected from IoT environment and they need to be protected from hacking. Collecting and aggregating the IoT data is not only necessary for an IoT application, but providing deployable information and notification on distinct activities must be considered when events lie outside the established regulation. In order to provide solution to this, sophisticated machine learning, artificial intelligence, data mining and big data techniques must be applied to detect any anomaly. Traditional network solutions

such as firewalls do not recognize IoT specific attacks and intrusions, so IoT Security Analytics is essential to identify such attacks.

Challenge #4 Interoperability

In IoT environment, variety of IoT devices (low computing power versus more capable devices), heterogeneous things and homogenous things which has different identification (addressing) are connected to the internet via gateway with different standards and protocols which needs interoperability. So the solution needed is standardization across vendors and a set of Open APIs for developers to overcome interoperability.

Challenge #5 How to store IoT data-Data storage formats

IoT data generated comes in different format such as structured and unstructured. Structured data and unstructured data can't be stored in one database. Because separate schema is used for unstructured data and it won't fit in the database. So the challenge is to permit storage of both types of data supporting faster analytics and providing security [7] and to manage big data so that unstructured data can be accessed quickly [8].

Challenge #6 Where to store the data

The challenge in storing the IoT data depends upon the application. It can be stored either in the cloud or at the edge or the third hybrid approach. A hybrid approach is to use both the edge and the cloud. In this method the data needed far faster response can be stored closer to the objects where the object is generated, and permanent data used for further examination can be stored at cloud. Since persistent data are moved to cloud, it is advantage with respect to storage of vast data in cloud, This approach supports the less cost of data storage and data transmission, since temporal data are stored in edge and remaining data are transferred to cloud for further analysis and queries [5].

Challenge #7 Adaptive data mining/ Machine learning algorithms

Several data mining algorithms and machine learning algorithms such as neural network, artificial intelligence, support vector machine, clustering, regression analysis etc exist. But which mining algorithm must be used for a specific application in real time for analytics to get the best solution is a major challenge.

Challenge #8Focus on multi-modal data

Several sensors are deployed in an IoT environment to continuously observe a number of factors, such as temperature, heat, moisture, motion, speed etc . The data generated by the several sensors are usually in different dimensions. This leads to several problems to tackle with huge amount of mixed data with respect to storage and processing

Challenge #9Analytics at the Edge.

Cloud computing can handle large amount of IoT data generated from IoT environment because of its high computing power, high storage and less cost .Therefore cloud is used in current situation for most of the IoT applications. But certain application demands low latency, high Quality of Service in real time. So for the application requiring less response time, cloud is not suitable. So therefore Analytics can be performed at the edge of the network instead of performing analytics at the cloud.

XI. CONCLUSION

It is foretold that the Internet will consist of trillions of connected computing devices universally in next generation. These nodes will generate huge amount of data and these data is the heart of the IoT architecture and it must be processed in depth to obtain valuable information and to provide a required solution for an application with low latency. Thus the internet of things has a long data processing channel in terms of collecting data, storing the data, and processing the data, and make decisions for an application. This paper put forwards data centric issues on IoT technology. It mainly researches the features and sources of IoT data, data aggregation, data storage, data processing and analysis with machine learning and data mining algorithms. It also provides an overview of some of the challenging areas in IoT with data perspective.

REFERENCES

[1] K. Ashton, 2009. That ''Internet of Things'' thing, RFID Journal.

[2] The internet of things An Overview -Understanding the Issues and challenges of a more connected world .Internet Society , Oct 2015, https://www.internetsociety.org/sites/default/files/ISOC-IoT-Overview-20151014_0.pdf

[3] Internet Of Things Ebook, Big Data Analytics and the Internet of Things, Exploring Enabling Technologies and Industry Opportunities, Datameer. https://www.datameer.com/pdf/eBook-Internet-of-Things.pdf

[4] "Smart Things | Home automation, home security, and peace of mind, "Smart Things, Palo Alto, CA, USA, Sep. 2014. [Online]. Available: http://www.smartthings.com

[5] Mervat Abu-Elkheir, Mohammad Hayajneh and Najah Abu Ali, 2013. Data Management for the Internet of Things: Design Primitives and Solution , Sensors2013, 13, 15582-15612; doi:10.3390/s131115582

[6] Vishakha More, Raghib Nasri, 2016. Application Framework and Data Processing in IoT based Email System,, International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 6, June 2016.

[7] Top Data Storage Challenges,http://www.vicomnet.com/top-data-storage-challenges/

[8]IoT and it's impact on data storage, http://nexiilabs.com/blog/iot-and-its-impact-on-data-storage/

[9] IDC Country brief, The Digital Universe in 2020 https://www.emc.com/collateral/analyst-reports/idc-digital-universe-united-states.pdf

[10] Jing-Song Li, Yi-Fan Zhang and Yu Tian, Medical Big Data Analysis in Hospital Information System , Chapter from the book Big Data on Real-World Applications. Downloaded from: http://www.intechopen.com/books/big-data-on-real-worldapplications

[11] Daniel E. O'Leary ,2013, BIG DATA, THE 'INTERNET OF THINGS' AND THE 'INTERNET OF SIGNS'. Intelligent Systems in Accounting, Finance and Management.20 (1), 53-65

[12] IERC, EUROPEAN RESEARCH CLUSTER ON THE INTERNET OF THINGS, Internet of Things, EU-China Joint White Paper on Internet-of-Things IdentificatioN,Nov,2014, http://www.internet-of-things research.eu/pdf/ IERC _Position_Paper_EU-China_IoT_Identification_Final. pdf

[13] IoT-A. Converged Architectural Reference Model for the IoT v2.0; SAP: Switzerland, 2012

[14].IETF, "The Internet of Things Concept and Problem Statement," 2010, http://tools.ietf.org/id/draft-lee-iot- problem-statement-00.txt

[15]CASAGRAS Project "Final Report, RFID and the Inclusive Model for the Internet of Things," http://www.grifsIproject.eu/data/File/CASAGRAS/FinalReport.pdf

[16] Eleonora Borgia, 2014. The Internet of Things vision: Key features, applications and open issues , Computer Communications54–31

[17] CERP-IoT. "Visions and Challenges for Realising the Internet of Things," European Commission(2010).

[18] Uckelmann, Dieter, Mark Harrison, and Florian Michahelles. 2011.Architecting the Internet of Things. Springer preview available at http://link.springer.com/book/10.1007/978I3-642-19157-2

[19] The Internet of Things: QA Unleashed.https://www.cognizant.com/whitepapers/the-internet-of-things-qa-unleashed-codex1233.pdf

[20] Tingli Li, Yang Liu, Ye Tian, ShuoShen, Wei Mao, 2012 .A Storage Solution for Massive IoT Data Based on NoSQL, IEEE International Conference on Green Computing and Communication

[21]Lihong Jiang, Li Da Xu, , HongmingCai, Zuhai Jiang, Fenglin Bu, and BoyiXu, 2014 ,An IoT-Oriented Data Storage Framework in Cloud Computing Platform, IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, VOL. 10, NO. 2,May

[22] Rongxing Lu, Kevin Heung, ArashHabibi Lashkari, Ali A. Ghorbani, 2017. A Lightweight Privacy-Preserving Data Aggregation Scheme for Fog Computing-Enhanced IoT, IEEE Access, FEBRUARY

[23] L. Chen, R. Lu, Z. Cao, K. Alharbi, and X. Lin,2015.Muda: Multifunctional data aggregation in privacy-preserving smart grid communications" ,Peer-to-Peer Networking and Applications, vol. 8, no. 5, pp. 777–792.

[24] FriederGanz, Daniel Puschmann, PayamBarnaghi, Francois Carrez, 2015. A Practical Evaluation of Information Processing and Abstraction Techniques for the Internet of Things , IEEE , VOL. 2, NO. 4, AUGUST

[25] H. Wang, D. Estrin, and L. Girod, 2003.Preprocessing in a tiered sensor network for habitat monitoring,v EURASIP J. Adv. Signal Process., vol. 2003, no. 4, pp. 392–401.

[26] Shree Krishna Sharma, Xianbin Wang, 2017, Apr.Live Data Analytics With Collaborative Edge and Cloud Processing in Wireless IoT Networks, IEEE Access.

[27] Zhuo Zhou , Min Liu , Feng Zhang , Li Bai , Weiming Shen, 2013 A Data Processing Framework for IoT based Online Monitoring System ,Proceedings of the 2013 IEEE 17th International Conference on Computer Supported Cooperative Work in Design , pp.686-691:

[28] Han J W, Kamber M, "Data mining, concepts and techniques (second edition)," Morgan Kaufmann Press http://www.academia.edu/23187985/Data_Mining_Concepts_and_Techniques_2nd_Edition_Solution_Manual

[29] ArijitUkil, Soma Bandyopadhyay and Arpan Pal, 2015 .IoT Data Compression: Sensor-agnostic Approach , IEEE Access, ,2015 Data Compression Conference

[30] Dr.G.Anandharaj ,Dr.P.Srimanchari , 2016 Unification Algorithm in Hefty Iterative Multi-tier Classifiers for Gigantic Peripatetic Applications Using Data Mining , International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 5, Issue 4, April

[31] FurqanAlama, Rashid Mehmoodb, IyadKatiba, AiiadAlbeshria, 2016. Analysis of Eight Data Mining Algorithms for Smarter Internet of Things (IoT) , Procedia Computer Science 98 (2016) 437 – 442

[32] Chun-Wei Tsai, Chin-Feng Lai, Ming-Chao Chiang, and Laurence T. Yang,2014 . Data Mining for Internet of Things: A Survey, IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 16, NO. 1.

[33] .HaoshuCai, Yu Guo, Wen-An Yang & Kun Lu, 2017. Mining frequent trajectory patterns of WIP in Internet of Things-based spatial-temporal database, International Journal of Computer Integrated Manufacturing Vol. 30, Iss. 12.

[34] Shweta Bhatia, Sweety Patel,2015 .Analysis on different Data mining Techniques and algorithms used in IOT, Int. Journal of Engineering Research and Applications www.ijera.comISSN: 2248-9622, Vol. 5, Issue 11, (Part - 1) Nov, pp.82-85

[35] .Krushika Tapedia, Anurag Manohar Wagh, 2016 . Data Mining for Various Internets of Things Applications", International Journal of Research in Advent Technology (E-ISSN: 2321-9637) Special IssueNational Conference "NCPCI-2016",.

[36] Qin, Yongrui, Quan Z. Sheng, Nickolas J.G. Falkner, Schahram Dustdar, Hua Wang, and Athanasios V. Vasilakos.2016. When things matter: A survey on data-centric internet of things, Journal of Network and Computer Applications 64(2016)137–153

[37] I. Priya Stella Mary , Dr. L. Arockiam, 2017. Detection of outliers in the IoT data using the STCPOD model .International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 4, Issue 10, October

[38] ZibinZheng, Jian Wang, ZiyuZhu, 2011. A General Anomaly Detection Framework for Internet of Things, https://www.cse.cuhk.edu.hk/lyu/_media/paper/zibin-dsn2011.pdf

[39] Yue Xu, Recent Machine Learning Applications to Internet of Things (IoT) ,http://www.cs.wustl.edu/~jain/cse570-15/ftp/iot_ml/

[40] Farshid Hassani, Bijarbooneh, Wei Du, Edith C.-H. Ngai, Xiaoming Fu, and Jiangchuan Liu. 2016. Cloud-Assisted Data Fusion and Sensor Selection for Internet of Things, IEEE INTERNET OF THINGS JOURNAL, VOL. 3, NO. 3

[41] Lev Faivishevsky, Information Theoretic Multivariate Change Detection For Multisensory Information Processing In Internet Of Things, ICASSP 2016: 6250-6254

[42] Raj Jain , Dr. Hitesh Shah , 2017 . An anomaly detection in smart cities modeled as wireless , sensor network.

[43] Yanxu Zheng ,Sutharshan Rajasegarar ,Christopher LeckieSmart , 2014. Car parking: Temporal clustering and anomaly detection in urban car parking ,IEEE

[44].Shu-Yun Lee, Fuchun Joseph Lin ,2017. Situation awareness in a smart home environment, IEEE Access .

[45] Yann-Ael Le Borgne, Sylvain Raybaud, GianlucaBontempi ,2008 .Distributed Principal Component Analysis for Wireless Sensor Networks ,*Sensors* 2008, *8*(8), 4821-4850; doi:10.3390/s8084821

[46] Xin Tao, ChunleiJi, 2015 .Clustering massive small data for IOT.

[47] 2016. Smart Surface: RFID-Based Gesture Recognition Using k-Means Algorithm, IEEE Access

[48] The 4 stages of an IoT architecture, https://techbeacon.com/4-stages-iot-architecture