Bigdata Analytics in Industrial IoT

Bhumi Chauhan and Chintan Bhatt

Abstract In IoT, technology is on an gradually developed both in terms of software and hardware. The high speed with which humans interact with the internet, use social media and interconnect their devices with another device is growing quickly. Due to the interaction between machine to human and machine to machine communication the massive amount of data will be generated and to store this generated data in the database becomes more difficult to store, manage, process and analyses. The data management is a biggest problem in IoT due to the connectivity of billions of devices, objects which are generating big data. With the help of big data technology we can handle that data. However due to the nature of Bigdata it has become important challenge to achieve the real-time capability using the traditional technologies. Big data is some technologies to capture, manage, store, distributed and analyses petabyte or larger sized datasets with highest velocity and different structure. Hadoop is a best platform for structuring Bigdata. It is a best tool for data analysis as it works for distributed big data, Time stamped data, structured, unstructured and semi-structured data, streaming data, text data etc. This paper represents the layer architecture of big data system. In addition, how to use FLUME and HIVE tool for data analysis. For NoSQLdatabse we use Hive which is SQL like query language is used for some analysis and extraction. Flume is used to extract real time data into HDFS.

Keywords Big data · IoT · Industrial IoT · Hadoop · MapReduce · Flume · Hive

B. Chauhan (✉) · C. Bhatt
Department of Computer Engineering, Charotar University of Science and Technology, Changa, Anand 388421, Gujarat, India
e-mail: 15pgce007@charusat.edu.in

C. Bhatt
e-mail: chintanbhatt.ce@charusat.ac.in

N. Dey et al. (eds.), *Internet of Things and Big Data Analytics Toward Next-Generation Intelligence*, Studies in Big Data 30,
DOI 10.1007/978-3-319-60435-0_16

1 Introduction IoT

> When we talk about an Internet of Things [20, 21], its's not just putting RFID tags on some speechless thing so we smart people know where that speechless thing is. It's about embedding intelligence so things become smatter and do more than they were proposed to do.

The basic concepts behind the Internet of Things is that when we discuss about 'things', we refer to objects, devices like wristwatches and medical sensors, actuators etc. These devices are such that they able to connect with users by yielding information from respective environment sensors along with actuators such that users can operate over a span of various interfaces. (see Fig. 1).

When we referring IoT and its technology the concept behind IoT is that the elements are connected with the realistic environment of the internet by reference of sense, monitoring and eventual tracking the objects and their environment. To enable themselves build web application, the developers in hand with the users assemble few components, allotting them sense and networking capabilities for communication and program them into execution by denoting a particular task.

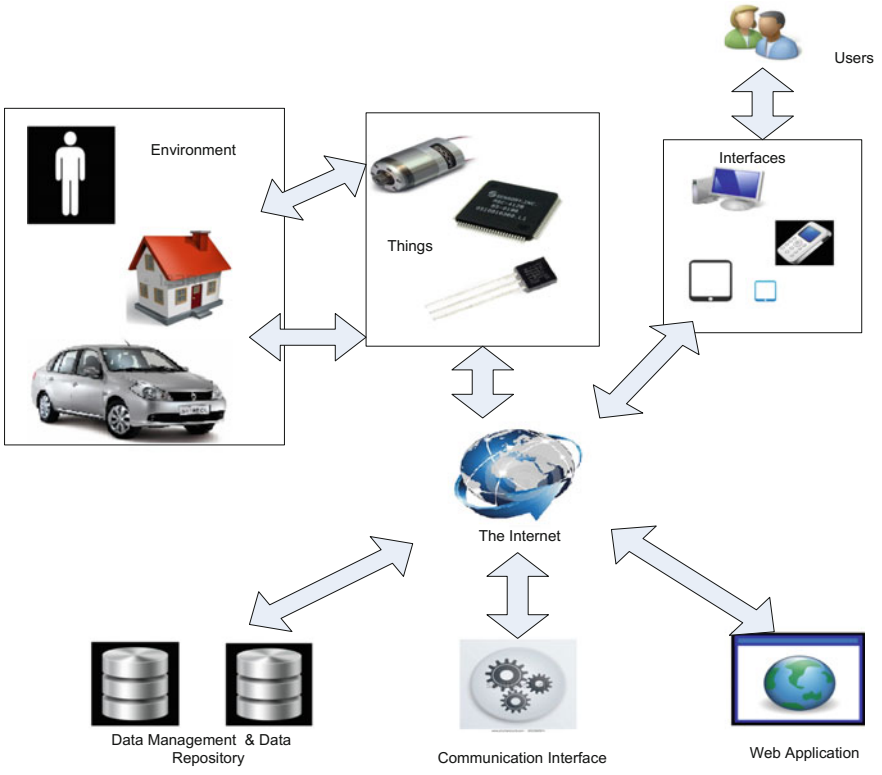

Fig. 1 Structure of 'Internet of Things'

The following are the characteristics of a devices that doing as a component of an IoT networks.

- *Collect and Transmission of data* In environment, the device can sense the things and collecting some information related to it and then transferring to a various device or to Internet also.
- *To Operating the devices based on triggers* It can be operating to abet other devices (e.g., turn on the lights or turn off the heating) base on condition set by you. For example, you can program the device when you enter in room the lights are turn on automatically.
- *To Collecting the information* This is most implement feature for IoT device that they collect the information from the network to the other devices or through the Internet.
- *Communication support* In IoT device there are members of a device network can also nodes of the same network.

The above diagram shows the structure of "Internet of Things". Things are referred as various sensors, actuators and even microcontrollers. These things interact with environment and Internet, thus allowing users to manage their data over various interfaces.

1.1 Industrial IoT

> The Industrial Internet of Things is a network of Physical objects, Systems, platforms and application that contain embedded technology to communicate and share intelligence with each other, the external environment and with people.

The main objectives of IIoT is to improved affordability and Availability of processors, sensors microcontrollers and other technologies are help to facilitate capture and access to real-time information.

GE introduce Industrial Internet (IoT) as a term which means "integrating complex physical machinery together with networked sensors and software. Industrial internet joins fields such as the IoT, Bigdata, machine learning and M2M (Machine to Machine) communication, to collect and analyses data from machines and use if for adjusting operations." [1] Industrial Internet contains three key elements which are represent the following figure (Fig. 2).

The first key element is Intelligent Machines that reparents connected the machines, networks and facilities with advanced controls, sensors and software applications. The second key element is advanced analytics that represent the combination of domain expertise, physical-based analytics, automated and precative algorithms for understanding the operation of systems and machines. The third key element is People at work represent, connecting people at any time and anywhere for supporting intelligent operations, maintains and high service quality and safety. [2, 3]

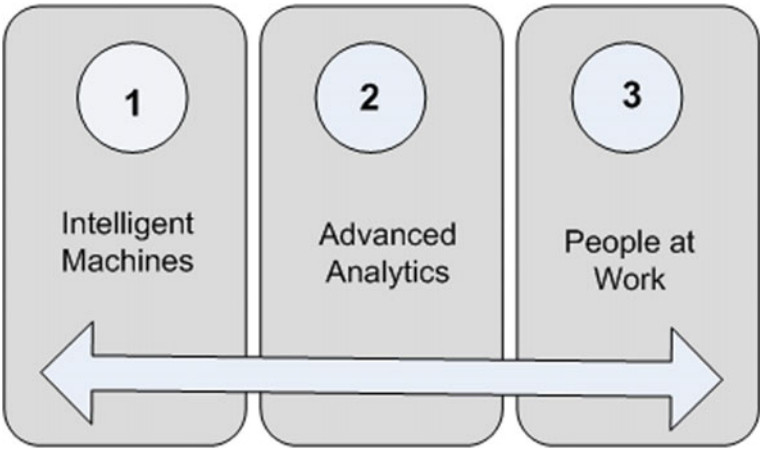

Fig. 2 Key elements of industrial IoT

1.2 IoT Application and Techniques

- **Smart City**

Smart City is one of the strongest application of IoT in subject to World Population's. Smart surveillance, smarter energy management systems, automated transportation, urban security and environmental monitoring and water distribution are some of the real-time examples of IoT for making.

IoT take care of huge issues confronted by mankind living in urban communities like traffic blockage and deficiency of vitality supplies, pollution etc. For example, a product named Cellular Communication if get enabled by smart belly trash then it will send alerts to municipal administrations when a container should be exhausted. By installing sensors and web applications, individuals can undoubtedly see accessible stopping spaces over the city. In additional sensor, can likewise distinguish meter altering issues, general malfunctions and establishment issues in any power framework.

- **IoT in Agriculture**

Nowadays, the demand of Food supply has been extremely raising due to the continued rise in the world's population. In this way, by propelling propelled method and researcher can take greatest yields. Keen cultivating is one of the quickest developing field in IoT. For better quantifiable profit farmers, have begun utilizing important insights from the information to yield. Detecting of soil dampness and supplements, controlling water usage for plant development and deciding customer fertilize are some examples of smart farming of IoT.

- **Smart Retail**

The prospects of IoT in the retail sector is simply amazing. To raise the in-store experiences. IoT gives a great chance to retailers to interface with the customers. Advance mobile phones will make too simple for retailers to get associated with their purchasers at whatever time level out of store to serve them better and also in more efficient way. By this they can undoubtedly track shopper's way through a store to enhance store design and place items appropriately.

- **Energy Engagement**

Power grids won't just be sufficiently keen additionally exceptionally solid for up and coming era. By keeping in that center Smart grid idea pulls in us and getting to be distinctly well known all over world. The principle objective of the Smart grid is to gathering information in a mechanized manner and examining the conduct or power consumers and suppliers for enhancing productivity also financial aspects of power utilize. It also detects sources of power outages more rapidly and at particular household levels like nearby solar panel, making possible distributed energy system.

- **IoT in Healthcare**

Connected healthcare yet remains the resting giant of IoT applications. The concept behind the healthcare system and smart medical devices brings huge prospects together not just in corporate but also in general for well-being of people. Research exhibit IoT in healthcare will set sound standard in coming years. The principle reason for IoT in healthcare is to engage individuals to live more beneficial life by wearing connected gadgets. The gathered information in customized examination will help to analysis individual well-being and give tailor made systems to battle disease.

- **Industrial Internet**

Industrial internet is another powerful application which has already created new buzz in the industrial sector which additionally named as Industrial Internet of Things (IIoT). It is empowering industrial engineering with sensors, software and big data analytics to make extraordinary machines. According to Jeff Immelt, CEO, GE Electric, IIoT is an "Excellent, desirable and investable asset of every organization. Compare to humans, Smart machines are more accurate and precise while communicating through data. In additional this data can help companies to recognize any fault and solving it in more efficient way.

When it comes to control quality and sustainability IIoT holds very strong and remarkable position. As indicated by GE, change industry profitability will create $15 trillion to $20 trillion in GDP worldwide over next 15 years. Application of keeping tracking over goods, continuous data trade about stock among provides and retailers and computerized conveyance will expand the store network effectiveness.

- **Connected Cars**

The internal functions of the vehicles have been optimized by the continual focus the automotive digital technology. Since then for now the attention is biased towards the enhancement of the in-car experience. A connected car is a vehicle which is proficient to upgrade its own operation, upkeep additionally solace of travelers utilizing installed sensors and web availability. Substantial car creators and additionally some valiant new businesses are being chipping away at associated car arrangements. As indicated by the most recent news, the brands like BMW, Apple, Testa and Google are being attempting to acquire new transformation car division.

- **Wearables**

Wearable have an explosive demand in global market companies like Google, Samsung have heavily invested in producing such devices. Wearable gadgets are introduced with sensors and software's which collect information and data about the clients. At that point this all information will be pre-prepared to concentrate fundamental data about client. These devices for the most part cover wellness, well-being and entertainment necessities. These devices mainly cover fitness, health and entertainment necessities. The pre-imperative from IoT innovation for wearable applications is to be exceptionally vitality proficient or ultra-low power and little estimated.

- **Smart Home**

Smart home is the most sought word on Google which is associated with IoT. Let's see how IoT will surprised you by making your routine life simpler and convenient life at your own way by having the real-time example of smart home. Wouldn't you cherish in the event that you could switch on air conditioning before reaching home or turn off lights subsequent to leaving home? Or, on the other hand open the entryways. If someone arrived at your home and want to access even when you are not at home. It will not surprise us if one day smart homes will become a common use in human life as to smart phones. If we can see nowadays, the cost of owning a house is the greatest and unavoidable cost in a property holder's life. So, to make their life smoother, smart home organizations like Nest, Ring, Eco bee and couple of more are wanting to convey a never observed or deal and to set their family rand in Real Estate Sector.

- **Techniques**

To making the IoT possible several technologies are to be used. This section in represent different techniques which are used in IIoT.

The first technologies are identifications. In the IoT system there will be trillions of devices which are connect to internet, to identify these all devices, each one required a unique identification and this is possible by using Ipv6 enabled. The Second technology is, an IoT device needs to sense. It is possible by putting sensors

on the system, here sensor able to measured different aspects of an object. These objects will be able to communicate with another object which are outside the world or other similar object around it. For device identification RFID and QR code are mostly used.

For communicating to each other there are some protocols to be used. By helping this protocol device could be able to collect the data and then sent to another device. If required is sent back to devices with other information also. For this purpose, IIOT use different protocols such as MQTT (Massage Queue Telemetry Transport), XMPP (Extensible Messaging and Presence Protocol), AMQP (Advanced Message Queuing Protocol), CoAP (Constrained Application Protocol) and DDS (Data Distribution Service)

1.3 M2M Communication

M2M Communication is a one type of data communication. It includes one or more entities where the process of communication is done without human interaction. M2M is known as Machine type communication (MTC) in 3GPP (3rd generation partnership project). M2M Communication is varied from the existing communication model It involved

- Less efforts and costs.
- Different market scenarios.
- Communicating terminal in very large number.
- Minor traffic per terminal.

M2M Communication is depending on mobile networks (e.g. CDMA EVDO networks, GSM–GPRS networks). The main responsibility of mobile network is to serve as a transport network in M2M Communication. M2M Communication makes enabling machines such as mobile devices, computers, embedded processors, actuators and smart sensors to communicate with each other, and make some decisions without human interaction [4]. Devices can be communication via wireless or wired network. M2M is a view of the IoT where objects are connected to the whole environment and managed with devices, networks and cloud based servers [5].

The main goal behind to develop M2M for cost reducing, increasing safety and security and achieving productivity gains (Fig. 3).

There are three stages of M2M communication: collection, transmission, and data processing. The data collection stage is the process for collecting the physical data. The transmission stage is the process for transmit the data to the external server. The data processing stage represent dealing and analyzing with data and also provide feedback or controlling the application.

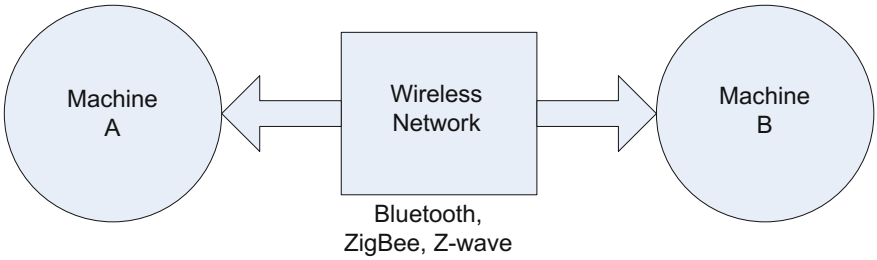

Fig. 3 M2M Communication

Table 1 Area of application of M2M Communication

Area	Application
Health	Remote diagnostics, to monitoring vital signs, supporting handicapped or aged
Tracking & tracing	Traffic information, traffic optimization, pay as you drive, road tolling, order management, navigation
Security	Access control, car security, alarm systems
Metering	Industrial metering, grid control, power, water, gas, heating
Facility management	Building or home or campus automation
Payment	Vending machines, point of sales

1.4 Application of M2M Communication

There are many areas where M2M application used. The following table shows the area in which M2M is used recently (Table 1).

The general requirements of M2M is that, using multiple communication Such that, SMS, IP Access and GPRS. The M2M system will be capable to provide communication between the M2M devices or M2M gateway or network and application domain.

- To deliver the message for sleeping devices. The M2M system will be able to manage the communication between sleeping device.
- M2M system should be support the different types of delivery nodes like any cast, unicast, multicast and broadcast communication modes.
- M2M system should be capable to manage the schedule of messaging and network access also.
- M2M system should be capable to decide communication path based on cost of network, delays or transmission failures when another communication paths exist.

- When number of devices are connected with object then M2M system should be scalable.
- M2M system should be capable to give any notification when it notified of any failure to deliver the message.

1.5 Bigdata and IoT

The use of big data analytics has grown extremely in just the past petty years. At same time, the IoT has introduce the public awareness, people's imagination on what a fully interconnected world can offer. For this reason, where big data and IoT are almost made for each other.

Big data and IoT basically go together. One study predicts that by the year 2020, IoT will be generated 4.4 trillion GB of data in the world. In year 2020, tens of billions of devices and sensors will have some specific type of connection to the internet. All of these devices will be producing, analysing, sharing and transferring data all in real time. Without this data, IoT devices would not have the functions and capable that have caused them to gain so much worldwide observation.

The IoT and big data are so nearly link. For example, wearable technology. Other is fitness band are part of IoT, it taking personal health data, giving the user calculations on recent fitness levels and tracking and noticed them about progress.

2 Bigdata

Newly companies, Government and academia become devoted in the big data. Using Traditional data management techniques are very difficult to process or manage bulky, multifarious and unstructured information. [6] Still, a majority issue for IT analyst and researchers is that this data expansion rate is fast excess their ability to both: (1) analyze is to extract suitable meaning for decision making, to lead and analysis of such datasets and (2) Design proper system to handle the data effectively. Bigdata can be describe using 5V model [7] represent in Fig. (4). This model is an expansion of 3V's model [8] and that includes:

1. *Volume (Size of data)* To generating and collecting huge amount of information (data), data rate take place big. Now-a-days, the data will be producing in the order of zettabytes, and it's expanding almost 50% every year. [9]
2. *Velocity (Streaming Data)* This represent the timeless of Bigdata. Data collection and analyzation both are immediately and timely managed. So, to magnified the use of the financial value of big data.
3. *Varity (Unstructured Data)* This model indicating the version type of data which contain data such as video, audio, webpages, images and text.

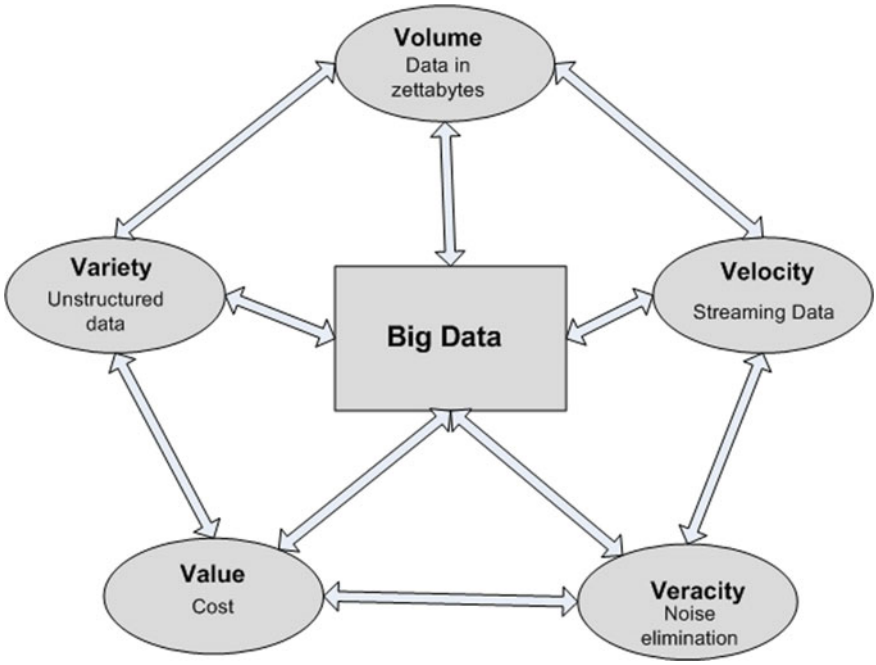

Fig. 4 5V's of big data

4. *Value (Cost of Data)* Data will be producing, collecting, and analyzing from the heterogeneous sources. Now a day the data having some cost. To make budget decision for estimation of data storage cost is very important.
5. *Veracity (Messiness of the data)* It is referring to the uncertainty or distrust around data, which is due to data incompleteness and inconsistency with many types of big data, quality and accuracy are less controllable but big data analytics technology permit us to work with these types of data.

The data will be generating as big impacts on our daily lives as internet has done. Data comes from social network Facebook, twitter, finance for example stock transaction, E-commerce system and specific research and sensor network etc. The given Fig. (5) shows the Layered Architecture of big data system. It is differentiating in three layers from top to bottom that consist infrastructure layer, computing and Application Layer.

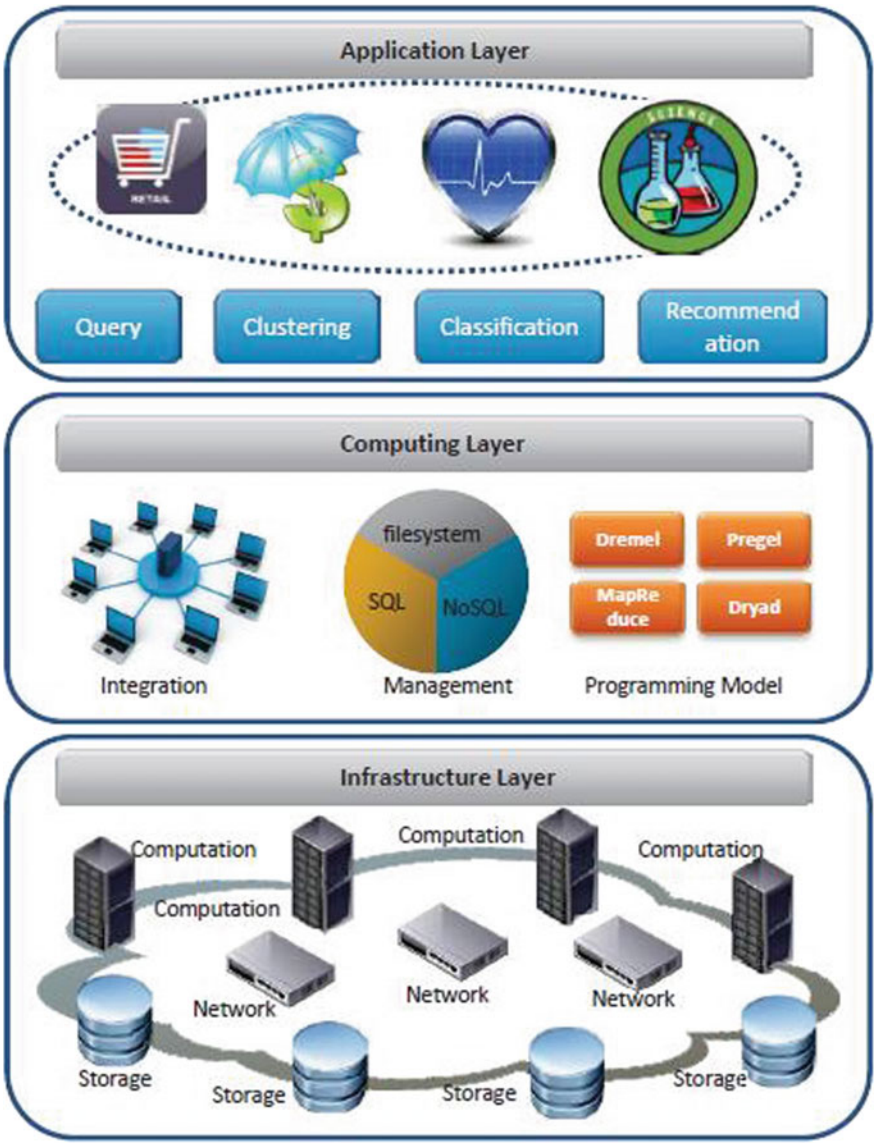

Fig. 5 Layered architecture of big data

2.1 Bigdata Challenges

There are a lot of challenges when manage Bigdata problem, during capturing the data, storing the data, sharing the data and searching the data and analysis and visualization the data. [10, 11]

1. *To Capturing the data* Data can be collected from the different sources like social media, sensors, experiments, meta data and transaction. Because of the different of data sources and the total quantity of data, it is exceptionally hard to gather and integrate information with versatility from conveyed areas. Different issues like this test are data pre-processing, transmission of data and automated created of metadata. [12]
2. *Data Storage* To Store the large data it not requires only a large amount of storage place but also require management of data on large heterogeneous system because of traditional database systems having difficulty to handle big data. As result of the great properties like being schema free, processing a simple API, simple replication, consistency and supporting a colossal measure of information. Apache HBase, Apache Cassandra, NoSQL database this all projects are turn into the core technologies of big data. The apache Hadoop is the most popular implementation open source software. Map Reduce provide automatic scalable and parallelization data distribution across many computers. Data Storage can be classified into several types as: NoSQL DBMS, RDBMS, DBMS based on HDFS and Graph DBMS.
    - *NoSQL DBMS* NoSQL DB is free in schema structure. It can be providing some features, such as distributed storage, horizontal scalability, dynamically schema etc. NoSQL DBMS storage the unstructured data in a key-value model. It is not good for Atomicity, Consistency, Isolation and Durability of data. Also, it can not support some distributed queries. It is new style for data management, that utilize the semantic information represented in ideology when searching the IoT data stores. Thus, some related work has tested to integrated the feature of distributed file system to the IoT data stores. [13]
    - *RDBMS* RDBMS is a basic platform for many structured data storages. RDBMS uses SPARQL queries to get the data on the existing relational stored views and ultra-wrap, which encodes logical representation of each RDBMS as an RDF graph. [14]
    - *DBMS based on HDFS* Here we integrated the HDFS with distributed file repository, which processing huge amount of unstructured file efficiently. In the area of IoT, many data are generated in the XML format, so how to handle these small sized, large-volume XML files becomes a major challenge.
    - *Graph DBMS* Graph DBMS is database. To represent and store the data it uses nodes and edges. Sensor data can be managed efficiently using Graph DBMS. [15] Graph DBMS provides a large graph that consists of large scale nodes and edges and high performance Graph DBMS management system (Table 2).
3. *Data Search* Today data must be available in exact, timely manner and completed and it becomes necessary. [16] At last data are to be use to aside a few minutes. Query optimization have been essential for responding effectively of complicated SQL queries. MapReduce is an exceedingly strong platform. It also provides high level query language such as PigLatin, HiveQL and SCOPE. This

Table 2 Comparison between data storage types

Features	NoSQL DBMS	RDBMS	DBMS based on HDFS	Graph DBMS
Support for Scalability	Yes	Not well	Yes	Yes
Support for Structured data	Not well	Yes	Yes	Use graph structure with nodes, edges
Support for semi and unstructured data	Yes	Not well	Yes	Use graph structure with nodes, edges
Support for ACID	Yes	Not well	Yes	Common
Support for massive & distributed processing	Yes, but not well	Not well	Yes	Yes

profoundly parallel platform, the cost of drag data across modes is capable and there for query optimization and physical structure have a tendency to be analyse components of the architecture.

4. *Sharing Data* Today, data sharing is become more important. In market thousands of data will be generated and absorb important data that is specific to their own business requirements but now it is produced in a way that could be interacted and shared to others requirements of an organization. Researchers facing issues regrading data sharing and presentation in several specific field like medicine, ecology and biology. [17]
5. *Data Visualization* Data visualization is a very challenging task for big data. As indicated to Fisher et al. [19] visual interfaces are reasonable for (1) to recognize patterns in data streams, (2) giving to deal with context by indicating data as a subset of a tremendous bit of data, and represents related variables, (3) data with statistical analysis. Visualization is an extraordinary field in which huge number of datasets are represented to users in visually powerfully irresistible ways with the faith that users would be capable to developed manageable relationships. Visual analytics required to developed many visualization across many datasets.
6. *Data Analysis* The magnetism of system expandability and hardware replication are represent by cloud computing along with MapReduce and Message Passing Interface (MPI) parallel program system suggest one resolution of this challenge by using a distributed approach. [18] Belonging to this, few issues are performance optimization of the software framework of big data analytics and dispersed and scalable data management ecosystem for profound analytics, i.e. the implementation of machine learning, different mining algorithms and statistical algorithms for analysis.

According to some kind of analysis big data tools are classified into (1) Batch analysis, (2) Stream analysis and (3) Interactive analysis. [20] In Batch analysis, data are primarily stored and after that analysed. In stream analysis, to extract the important information and to discover the knowledge from the generated data. In Interactive analysis, allow to users to lunch their own analysis of information.

2.2 Framework, Tools, Techniques, and Programming Language Supported by Big data

The tools of big Data develop for manage and analyse big data. This tools are efficiently process huge amounts of data inside sufficient expired times. The big data uses mostly open source software framework. Big organization give to open sources project and give some profit such as Yahoo!, Twitter, Facebook and LinkedIn. We represent how to deal with big data and their tools, techniques, framework and programming language in this section.

2.3 Big Data Techniques

In Bigdata many techniques are used to analyse the large amount of data. Here we describe some exclusive techniques which are generally use in Bigdata and their projects.

1. *Data Mining* Data mining is multidisciplinary area. It analyses and explore to huge amount of data to discover meaningful patterns. Data mining field use three rules which are Machine Learning, Pattern Recognition rule and Statically rule. Data mining is applicable in security analysis and intelligence, genetics, business and the social and natural sciences. Some techniques consist cluster analysis, association rule mining, classification and regression.
2. *Statistics* This technique are generally used for judgments about relationship between variables could and relationship variables. Variables occurred by chance and variables likely outcome from some kind of understanding causal relationships.
3. *Machine Learning* Machine Learning is the basis method of data analysis. It was well known from pattern recognition. The theory behind the machine learning is that computers can learn to perform specific tasks without being programmed. It has the ability to apply complex mathematical calculation to big data. It's goal to build analytical model automatically and it allows computers to find unread data into model. Right now, machine learning is expanded widely.
4. *Signal Processing* In this research, big data challenges offer extensive opportunities. Where the data are analyses in real-time dictionary learning, principle component analysis, compressive sampling, have already arranged on time/date adaptivity, robustness and dimensionality reduction. The main task of computer science in big data is disputed.
5. *Visualization Techniques* This visualization technique is analysis of high dimensional it is mostly used as sufficient data abstraction tools in searching knowledge, information awareness and making decision process.

2.4 Big Data Tools

According to the kind of analysis big data tools are classified into three types (1) batch analysis, (2) stream processing, (3) interactive analysis. In batch analysis, data are stored after that it is analysed. In stream processing, as early as possible which is analysis and it give results. In interactive analysis allow to users to perform their own analysis of data. Following table represent the taxonomy of the tools, frameworks for applications of big data (Table 3).

- *Based on batch analysis* there are four tools are used such as Google MapReduce, Apache Hadoop, Apache Mahout and Apache Microsoft Dryad.

(1) **Google MapReduce**

It is parallel computing batch oriented java based model. As Master, which responsible for assigning the work to the workers. Master assign to map workers from input data divide into splits. Each worker processed the individual input split and it is also generating key value pairs and write then on disk or in memory (intermediate files). The master gives the location of that files to the reducer workers. Reduce read data, process it and ultimately write data to output files. Map Reduce have three primary features scalability, simplicity and fault tolerance.

Table 3 Tools and framework for big data application

Name	Specific use	Advantages
Batch analysis		
Google MapReduce	Processing of data on large cluster	Fault tolerant, simple, scalable
Apache Hadoop	Platform and Infrastructure	Completeness, scalable, reliabilities, extensibility
Apache Mahout	Machine Learning Algorithm	Scalable
Microsoft Dryad	Platform and Infrastructure	Fault tolerant, good programmability
Stream analysis		
Apache Spark	Use for data processing	Easy to use, fast
Apache Strom	Real-time computing system	Simple, scalable, efficient, fault tolerant, easy to use and operating easily.
Apache S3	Platform for Stream computing	Scalable, extensible, fault tolerant
MOA	Use for Data stream mining	Scalable, extensible
Interactive analysis		
Apache Drill	SQL query engine for Hadoop and NOSQL	Familiar and quick, flexible
Apache BI	Business Intelligence	Big data streaming on Real-time BI
D3	Interactive	Scalable

In a cloud environment MapReduce is a good example for Bigdata processing. It allows develop parallel programs to a newer programmer. MapReduce processing a large amount of data in a cloud. So, that it is preferred for computation model of cloud provides.

MapReduce also have some limitations, MapReduce is suitable only for batch processing job, it is impossible for iterative jobs and models. It became very expensive while implementing iterative map reduce jobs for huge space consumption by each job.

(2) **Apache Hadoop**

Apache Hadoop is used for distributed processing of very huge amount of data sets and it is an open sources framework. Apache Hadoop inspired by Google File System, MapReduce and Google Big Table. It consists processing part called Map Reduce and Hadoop Distributed File System(HDFS) known as storage part that can usually replace SAN devices. The following Fig. (6) shows architecture of Hadoop stack.

Hadoop procures users to distributed queries across multiple datasets on large clusters. HDFS stores the large files across the multiple machines and that files running parallel MapReduce computing on the data. Hadoop support some kind of

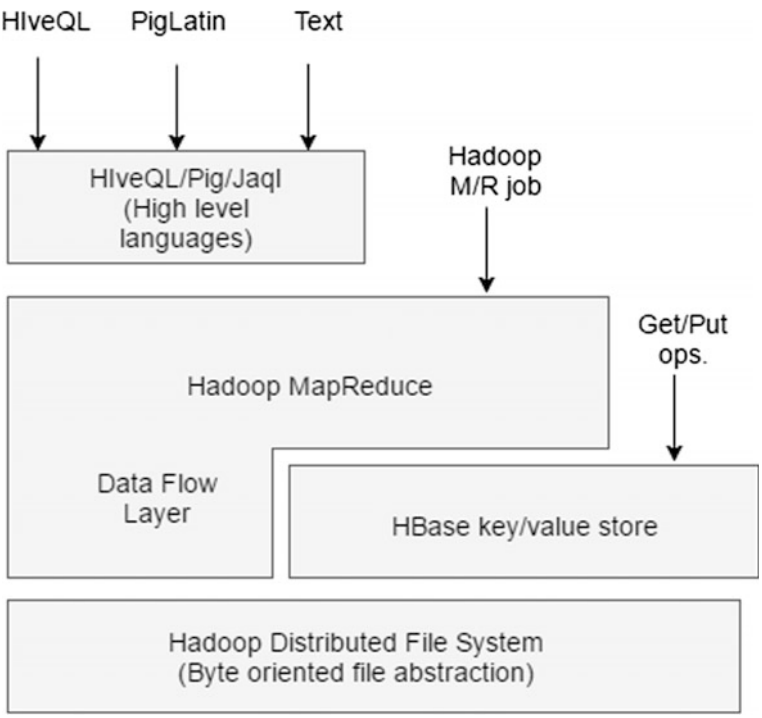

Fig. 6 Hadoop Stack

languages such as for domain specification language like Java and Python, for Query language like SQL, HiveQL and Pig Latin, this all languages Hadoop is suitable for deep analytics. In additional, Hadoop sub-project, HBase and Hive offer data management solution for unstructured and semi-structured data.

Today, Hadoop has risen to be "gold standard" in the industry as a platform that is feasible with scaled data intensive Map Reduction platform, it is highly used for large scale information extraction, Web indexing, Machine Learning, Clickstream and log analysis. Facebook, Twitter, Amazon and LinkedIn this companies used Hadoop technology.

(3) **Apache Mahout**

Apache Mahout is a Machine Learning framework. Apache Mahout supports three utilize cases: Clustering, Classification and Recommendation. It has aims to make intelligent application faster and easier.

(4) **Microsoft Dryad**

Dryad is a framework which allow a software engineer to used the resources of data centre and computer cluster for running data parallel programs. A Dryad programmer can utilize a huge of number of machines, each machine with multiples processors without knowing anything about simultaneous programming. It contains a runtime parallel system called Dryad and two high level programming models.

The process of Dryad is that the dryad programmer write programs sequentially and linked them using one way channels. As a directed graph the computing structured is that the program is known as graph vertices and channels are known as graph edges. It is responsible for generating graph which can synthesize all directed acyclic graph.

Dryad is similar to other frameworks like Google MapReduce or relational algebra. In additional, it handles job mentioning and visualization, fault-tolerance, resource management, scheduling, re execution and job creation and management.

- *Based on Stream Analysis* here we describe four tools they are Apache spark, Apache Strom, Apache S4 and MOA (Massive Online Analysis). This all tools are used for real time streaming in big data. This tools have to ability to handle large amount of data, using compute clusters to balance the workload.

(1) **Apache Spark**

It is large scale data processing engine. The speed of spark is faster than Hadoop Map Reduce in memory. Spark include the stack of libraries such as SQL and Data frames, MLib for Machine Learning, GraphX and Spark streaming. You can join this library consistently in the similar application. Spark can run all around. Using spark standalone cluster mode, you can run it on EC2, on Hadoop Yarn or on Apache Mesos. Spark can access data in HBase, Hive, Cassandra. Spark gives APIs in Python, Scala and Java.

(2) **Apache Strom**

Apache Strom is open sources free distributed real time computation system. Strom doing real time processing what Hadoop did for batch processing. It makes easy to reliable process of data. Strom is very easy and any programming language is used with storm. It is very fast tool for stream analysis. It can be processed millions of tuples per second each node.

(3) **Apache S4**

Apache S4 is a Distributed stream computing platform. It allows programmer to easily develop applications for processing continuous unbounded stream of data. It was initially released by Yahoo! It is fault tolerant stand by server is automatically activated to take over the task. It is extensible on which applications will be easily written and deployed via simple API.

(4) **MOA (Massive Online Analysis)**

MOA is an open source framework for data streaming. It is related to Weka project that allows to build and run experiments of data mining and machine learning algorithms. It includes Classification, Clustering and Frequent item set mining. By combing this two tools S4 and MOA software we can also use for distributed stream mining.

- *Based on interactive analysis* here we describe three tools they are SpagoBI, Apache Drill and D3.

(1) **SpagoBI**

SpagoBI is an open source business intelligence on big data. It offering large range of analytical function, it is an advance data Visualization for geospatial analytics.

(2) **Apache Drill**

Apache Drill is use for analysis of read only data. It is scalable, interactive ad hoc query system. Apache Drill is to able to manage more than 12,000 servers and processed petabyte of data in seconds. It is a framework for data intensive distributed application for interactive analysis of large datasets.

(3) **D3 (Data-Driven Documents)**

D3 stand for Data-Driven Documents. It is a java script library for developing dynamic and interactive data visualization in web browsers. D3 is generally implemented in HTML5, SVG and CSS standard.

3 To Collect, Organize and Analysis the Data

The Internet of Things gather outstanding consideration over the past certain years with implementing latest technology like sensor hardware technology and micro-controller materials that are not expensive. Sensing devices connected to all the elements surrounded such that they can be able to interact with each other without human interaction. To undertake the concept of sensor data is something that is a biggest challenge that will come across IoT.

IoT is creating tremendous volume of data, also rising analytics tools and mechanisms, that are affording that to allow us to operate this all machines absolutely new ways, and efficiently way. The data which are generated from all this resources which are linked to the Internet will apparently of this world to fetch advantages. Here the big data analysis needed to take advantages of it's very high-level engineered knowledge.

This section is describing the research work that how to collect data using Flume, how to organize data in Hadoop and how to analyses data using Hive. Here we use some sensor data like temperature sensor data. We collect the temperature datasets which generate the maximum temperature dataset by the year 1900–2016.

3.1 To Acquire Data

Data has to be collect from various sources. They are listed below.

(1) *Sensor or Machine Generated Data* it is including Smart meters, Call Detail Records, Sensors, Weblogs and trading system data.
(2) *Social Data* Twitter and Facebook. It including customer feedback stream and blogging sites.
(3) *Traditional Organization Data* It including ERP data, Web store truncations and general data. It also includes customer information from CRM systems.

- **Flume**

Flume is distributed open source software. From the many various sources this system collecting, assembly and moving huge number of logs data and store to be centralized data. Here using Flume, we moving and aggregating very huge amount of data around a Hadoop Cluster. In Hadoop cluster Flume is collect log file from all machines and then continues in a HDFS which store centralized. By developing chains of logical nodes, we have to creating data flows and then connect them to source and sinks.

Mostly Flume have three tire architecture: (1) Agent tire, (2) Collector tire, (3) Storage tire. The Agent tier, it has flume agent that collect the data from various sources which is to be moved. Collector tire, includes multiple collectors which collect the data comes from multiple agents and forward it on to storage tier which have file system like HDFS or GFS.

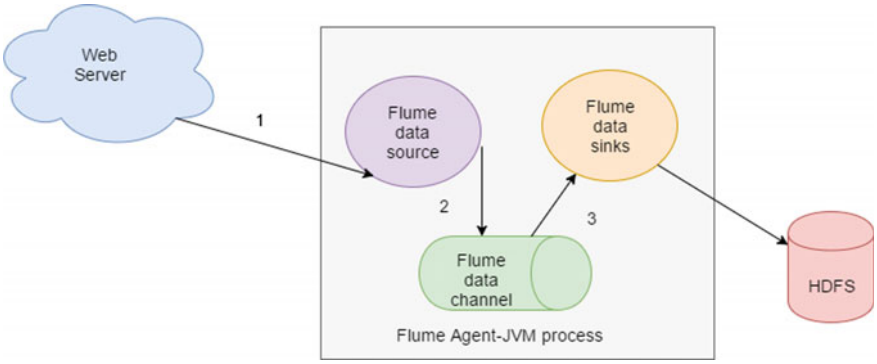

Fig. 7 Working of flume

Flume has three components, *Flume source, Flume channel* and *Flume sink.* Flume agent is a JVM process (Fig. 7).

(1) The above Fig. (3) the working of Flume. The events are generated from the external sources (Web server). The external source sends events to the flume source.
(2) Flume source receives that event which are send by external sources and store it into different channels. The channel performs as a storage system which place the event till it absorbs by the sink.
(3) All events are remove from channel in the sink and store it at HDFS or external database. There will be various flume agents in which case flume sink transfer the events to the flume source of next flume agent in the flow.

3.2 To Organize Data

After collecting data, it has to be organized using a distributed file system it has to be organized by collecting data. Here we have to split this data into fixed size blocks and store it into HDFS file systems. Figure (8) illustrate the architecture of HDFS.

Hadoop distributed file system is a scalable, portable and distributed file system which is written in java. In Hadoop cluster, every cluster has a single name node and many data nodes. Data node has many blocks by default each block have 64 MB size. For communicate with each other clients use RPC call. There are three copies of replication of data are present in every data node. Data nodes are communicating to each other to copy data or to rebalance data or to keep high replication of data. In case, if node goes down the HDFS has able to allowing name node to be failed over to backup.

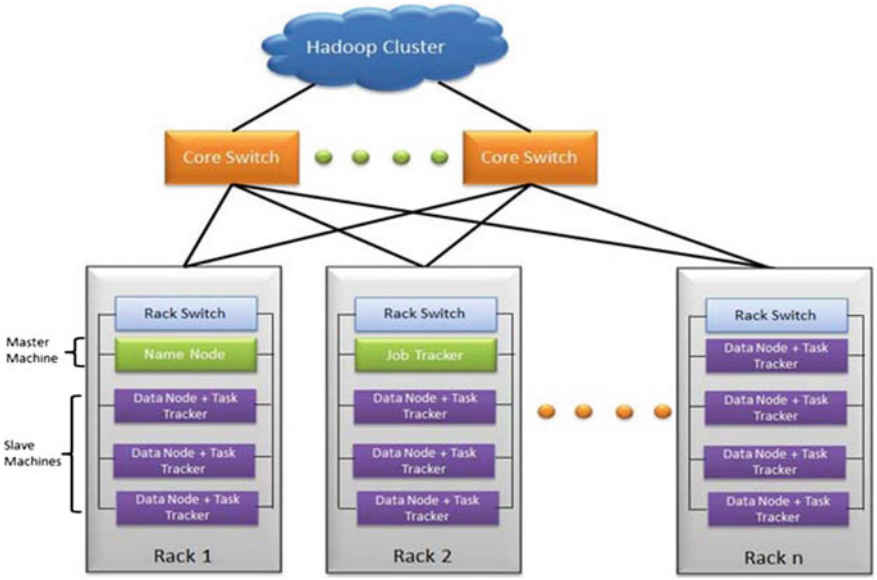

Fig. 8 Architecture of HDFS

Today, automatic failover is also establishing. For this we use a secondary name node which takes the snapshots continuously of primary name node so that is can be active when failure of node occurs.

3.3 To Analyze Data

To analyze data here, we collected the temperature datasets and perform some analysis using Hive. For this purpose, we install Hadoop, Hive on cloud era. The following Fig. (9) represents temperature datasets which are present in HDFS.

- **Hive**

Hive know as a data warehouse to process structure data in Hadoop. It builds on top of the Hadoop and providing query and analyze easily. It was developed by Facebook and it is used by various components for example Amazon Elastic Map Reduce which is used by Amazon. The Architecture of Hive is representing following Fig. (10) it contains many components which are describe in following table (Table 4).

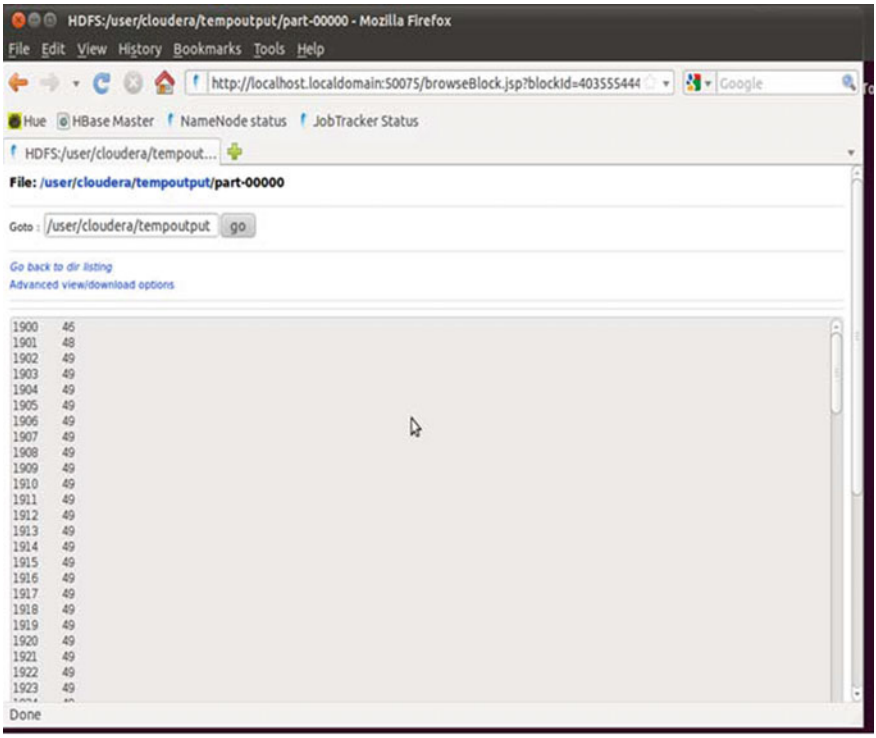

Fig. 9 Showing the temperature from the year 1900–2016

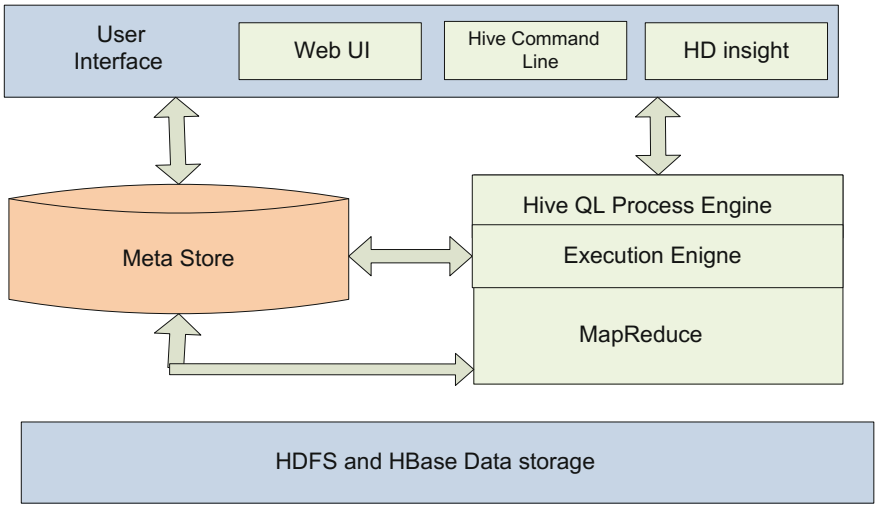

Fig. 10 Architecture of hive

Table 4 Description of Hive components

Component name	Operation
User interface	This component interacts with user and HDFS. The user Interfaces that supports are Hive HD Insight, Hive Command line and Hive web UI
HiveQL process engine	It is use for querying on schema information on the types and it is similar to SQL
Meta store	Hive select respective database server to store the meta data of tables or schema, databases, HDFS mapping and columns in a table and their data types
Execution engine	This component processes the query and produce the results as same as MapReduce results
HDFS or HBase	It is a technique to store data into file system

Fig. 11 Hive screen

There are three different ways to configured Hive

1. By using set command.
2. By hiveconf option in Hive command shell
3. By editing a file hive-site.xml

Here more than ten lacs database record which is analyzed by using Hive to get a temperature. We perform the analysis by using following Hive query language commands (Fig. 11).

- Create Database,

  Create database temp_records;
  Use temp_records;

- To store the records, we create table.
- *Create table temp_record (year INT, year INT);*
- Loading the data into table
- Describe metadata or schema of the table,

Describe temp_record;

- Select data
  *Select*from temp_record;*

4 Open Issues and Challenges in Big Data Analytics

To solve many issues regarding to data cleaning, security, privacy and Communication between systems big data become latest topic in the research field. This all issues are describing briefly in this section.

- **Data Cleaning**

Data Cleaning playing the main role in the data handle, management and analytics and still it has quick development. Furthermore, in the age of Bigdata Data Cleaning consider a major challenge. Because of the growth of volume, variety and velocity of data in so many applications. Here we introduce some issues to be repaired to needs are: (1) Usability, this feature is very important of data management, to interact with machines for human is very challenging task, (2) Tool Selection, the question arise here is that which tool to be use for given database to perform specific task for data cleaning, (3) Rule discovery, another challenge for cleaning data which is very difficult process.

This all issues are extremely hard to deal with when goes to the specific topic of data cleaning. All things considered, given detected errors, there is no right and which aren't right, even for human also.

- **Data Security**

In big data analytics, what way to handle generated data without a security it also biggest challenge of big data analytics. Related to our examination, big data analytics in security issues can ordered into four parts: input, data analysis, output and communication with different framework. In the input part, security issues of big data analytics are to confirm that the sensors won't influenced by the assaults. In the output part and analysis, it could be connected as the security worry of a framework. For communication with rest of the systems, the security concern has been between various external system and big data analytics. Because of issues all issues, security becomes most important issues in big data analytics.

- **Communication Between System**

For the most part, big data analytics systems has been constituted for computing parallels and would be actualized on different systems like cloud computing or else web index or else learning focused, the communication between the other systems and big data analytics would be heavily affect the execution of the entire procedure of KDD. The cost of communication between the system is first challenge of data analytics. The first responsibility of data scientists is that how can there be reduction

in cost of communication. Another open issue on the communication between system is that the data consistency between various systems, operators and modules is also and very important issue, and second open issues of big data analytics is which to make the communication between these frameworks as dependable as could be allowed.

- **Data Privacy**

In big data applications, the data privacy will be big issues specially if an organization does not give the conformation that their personal information's are secure with them the analysis process needs the data so privacy is most concern issues.

5 Conclusions

In IoT, it is very difficult to manage big data which are generated from different sources using traditional database management systems. We also provide a varied diversification of big data challenges which are Capturing, storing, Searching, Sharing, Analyzing and Visualizing. Here we conclude that Analysis is very critical challenge for research relative to big data, since its application is in the areas of earning potential knowledge for big data. We also represent the tools and frameworks which are used mostly and programming language for application of big data.

In addition, to handle big database we use some intensely parallel software. First of all, we collect data from heterogeneous sources which can be done easily by using flume. Then we can be organized data using distributed file system like HDFS. After this we analyzed that data using Hive. Hive can be analyzing a database of ten lacs records in just 35 s. So, using these all components it makes possible to manage and to use big database in an easy and efficient manner.

Acknowledgements Thank you for your cooperation and I would also thank to my guide for his imitable guiding constant monitoring and encouraging support throughout the task. His blessing continual guidance shall help me achieve my goals in journey of my life over which I am.

References

1. General Electric Company. (2014). GE intelligent platforms, Industrial Internet. Retrieved from http://www.ge.com/industrial-internet
2. Industrial Internet Consortium. (2014). Engineering: The first steps. Retrieved from http://www.iiconsortium.org/pdf/IIC-frirst-steps_2014.pdf
3. Annunziata, M., & Evans, P. C. (2012). *Industrial internet: Pushing the boundaries of minds and machines*. Boston: General Electric Co.
4. Watson, D. S., Piette, M. A., Sezgan, O., & Motegi, N. (2014). Machine to Machine (M2M) technology in demand responsive commercial buildings Washington: American Council for an Energy Efficient Economy.

5. Chen, C. L. P., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences, 275,* 314–347.
6. Taila, D. (2013). Clouds for scalable big data analytics. *Computer, 48*(5), 98–101.
7. Lomotey, R. K., & Deters, R. (2014). Towards knowledge discovery in big data. In *Proceeding of the 8th International Symposium on Service Oriented System Engineering IEEE Computer Society*.
8. Lancy, D. (2011). 3-D management: Controlling data volume, velocity and veracity application. *Delivery Strategies*.
9. Fan, W., & Bifet, A. (2012). Mining big data: Current status, and forecast to the future. *SIGKDD Explore*.
10. Chen, P. C. L., & Zhang, C. Y. (2014). Data intensive applications, challenges, techniques and technologies: A survey on big data. *Information science, 275,* 314–347.
11. Alshammari, H., Lee, J., & Bajwa, H. (2015). Improving current Hadoop mapreduce workflow and performance. *International Journal of Computer Applications, 116*(15), 38–42.
12. Hu, H., Wen, Y., Chua, T. S., & Li, X. (2014). Toward scalable systems for big data analytics: A technology tutorial. *IEEE Access, 2,* 652–687.
13. Cure, O., Kerdjoudj, F., Faye, D., Le Due, C., & Lamolle, M. (2013). On the potential integration of an ontology based data access approach in NoSQL stores. *International Journal of Distributed System & Technologies (IJDST), 4*(3), 17–30.
14. Yaish, H., Goyal, M., & Feurelicht, G. (2014). Multi-tenant elastic extension tables data management, proedria computer. *Science, 29,* 2168–2181.
15. Martinez-Bazen, N., Gomez-Villamor, S., & Escale-Claveras, F. (2011) Dax: A high-performance graph database management system. In *Data Engineering workshop (ICDEW), 2011 IEEE 27th International Conference on IEEE 2011* (pp. 124–127).
16. Hameurlain, A., & Morvan, F. (2015). Big data management in the cloud: Evolution or crossroad? In *International Conference: Beyond Databases, Architectures and Structures*. Springer, Cham.
17. Naseer, A., Laera, L., & Matsutsuka, T. (2013). Enterprise big graph. In *46th Hawaii International Conference on System Sciences. IEEE Computer Society* (pp. 1001–1014).
18. Gropp, W., et al. (1996). A high-performance, portable implementation of the MPI message passing interface standard. *Parallel computing, 22*(6), 789–828.
19. Rossi, R., & Hirama, K. (2015). Characterizing big data management. *Issues in Informing Science and Information Technology, 12,* 165–180.
20. Bhayani, M., Patel, M., & Bhatt, C. (2016). Internet of Things (IoT): In a way of smart world. In *Proceedings of the International Congress on Information and Communication Technology* (pp. 343–350).
21. Bhatt, Y,. & Bhatt, C. (2017). Internet of Things in healthCare—Internet of Things and big data technologies for next generation healthcare (pp. 13–33).