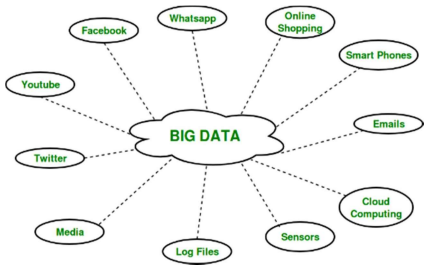
Big Data

Dawn of the bigdata

The word dawn means: the beginning of a phenomenon or period of time, especially one considered favourable.

Over the past 20 years, data has increased in a large scale in various fields. According to a report from International Data Corporation (IDC), in 2011, the overall created and copied data volume in the world was 1.8ZB, which has increased by nearly nine times within 5 years. Such figure will double at least every other 2 years in the near future.1

Big data has attracted considerable interest from industry, academia, and government agencies. For example, issues on big data are often covered in public media, including The Economist, New York Times, and National Public Radio.



First, the latest advances of information technology (IT) make it more easily to generate data. Second, the collected data is increasingly growing, which causes a problem of how to store and manage such huge, heterogeneous datasets with moderate requirements on hardware and software infrastructure. Third, in consideration of the heterogeneity, scalability, real time, complexity, and privacy of big data, we shall effectively " mine" the datasets at different levels with analysis, modeling, visualization, forecast, and optimization techniques, so as to reveal its intrinsic property and improve decision making.

Def: Definition

Big data refers to the datasets that could not be perceived, acquired, managed, and processed by traditional IT and software/hardware tools within a tolerable time.

In 2010, Apache Hadoop defined big data as "datasets which could not be captured, managed, and processed by general computers within an acceptable scope.

Big data is a collection of large datasets that cannot be processed using traditional computing techniques.

Challenges:

The challenge includes capturing, curating, storing, searching, sharing, transferring, analyzing and visualization of this data.

The three different formats of big data are:

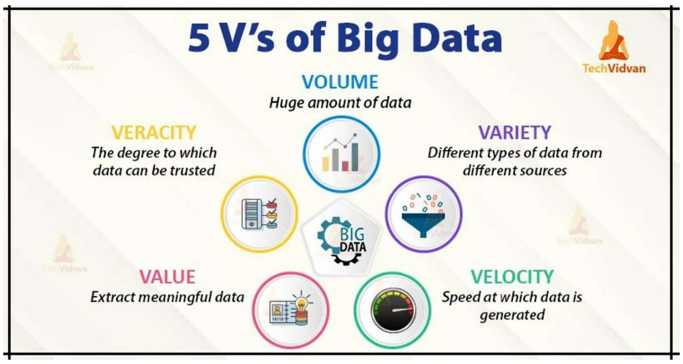
- ***Structured:*** Organised data format with a fixed schema. Ex: RDBMS
- ***Semi-Structured:*** Partially organized data which does not have a fixed format. Ex: XML, JSON
- ***Unstructured:*** Unorganized data with an unknown schema. Ex: Audio, video files, etc.

Advantages of BD:

- Cost Savings
- Better decision-making
- Better Sales insights
- Increased Productivity

- Improved customer service.

5V's of big data: Features



characteristics of Big Data:

1. Volume:
- The name 'Big Data' itself is related to a size which is enormous.
- Volume is a **huge amount of data**.
- To determine the value of data, size of data plays a very crucial role. If the volume of data is very large then it is actually considered as a 'Big Data'. This means whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data.
- Hence while dealing with Big Data it is necessary to consider a characteristic 'Volume'.

2. Velocity:
- Velocity refers to the **high speed of accumulation of data.**
- In Big Data velocity data flows in from sources like machines, networks, social media, mobile phones etc.
- There is a massive and continuous flow of data. This determines the potential of data that how fast the data is generated and processed to meet the demands.
- *Example:* There are more than 3.5 billion searches per day are made on Google. Also, FaceBook users are increasing by 22%(Approx.) year by year.

3. Variety:
- It refers to **nature of data** that is structured, semi-structured and unstructured data.
- It also refers to heterogeneous sources.
- Variety is basically the arrival of data from new sources that are both inside and outside of an enterprise. It can be structured, semi-structured and unstructured.

4. Veracity:
- It refers to **inconsistencies and uncertainty in data**, that is data which is available can sometimes get messy and quality and accuracy are difficult to control.
- Big Data is also variable because of the multitude of data dimensions resulting from multiple disparate data types and sources.
- *Example:* Data in bulk could create confusion whereas less amount of data could convey half or Incomplete Information.

5. Value:
- After having the 4 V's into account there comes one more V which stands for Value.

- Data in itself is of no use or importance but it needs to be converted into something valuable to extract Information. Hence, you can state that Value is the most important V of all the 5V's.

Big Data challenges : challenges

- ***Sharing and Accessing Data:***
  - Perhaps the most frequent challenge in big data efforts is the inaccessibility of data sets from external sources.
  - It include the need for inter and intra- institutional legal documents.
  - Accessing data from public repositories leads to multiple difficulties.
  - It is necessary for the data to be available in an accurate, complete and timely manner.
- ***Privacy and Security:***
  - It is another most important challenge with Big Data.
  - Most of the organizations are unable to maintain regular checks due to large amounts of data generation. However, it should be necessary to perform security checks and observation in real time because it is most beneficial.
  - There is some information of a person which when combined with external large data may lead to some facts of a person which may be secretive and he might not want the owner to know this information about that person.
- ***Analytical Challenges:***
  - There are some huge analytical challenges in big data which arise some main challenges questions like how to deal with a problem if data volume gets too large?
  - Or how to find out the important data points?
  - Or how to use data to the best advantage?
- ***Technical challenges:***
  - **Quality of data:**
    - When there is a collection of a large amount of data and storage of this data, it comes at a cost.
    - For better results and conclusions, Big data rather than having irrelevant data, focuses on quality data storage.
  - **Fault tolerance:**
    - Fault tolerance is another technical challenge and fault tolerance computing is extremely hard, involving intricate algorithms.
    - Nowadays some of the new technologies like cloud computing and big data always intended that whenever the failure occurs the damage done should be within the acceptable threshold that is the whole task should not begin from the scratch.
  - **Scalability:**
    - Big data projects can grow and evolve rapidly. The scalability issue of Big Data has lead towards cloud computing.
    - It leads to various challenges like how to run and execute various jobs so that goal of each workload can be achieved cost-effectively.
    - It also requires dealing with the system failures in an efficient manner.

cloud prelimanries

Cloud:

Cloud computing refers to the on demand availability of computing resources over internet. These resources includes servers, storage, databases, software, analytics, networking and intelligence over the Internet and all these resources can be used as per requirement of the customer. In cloud computing customers have to pay as per use. It is very flexible and can be resources can be scaled easily depending upon the requirement. Instead of buying any IT resources physically, all resources can be availed depending on the requirement from the cloud vendors. Cloud computing has three service models i.e Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS).

Examples of cloud computing vendors who provides cloud computing services are Amazon Web Service (AWS), Microsoft Azure, Google Cloud Platform, IBM Cloud Services etc.

Characteristics of Cloud Computing :

- On-Demand availability
- Accessible through a network
- Elastic Scalability
- Pay as you go model
- Multi-tenancy and resource pooling.

Advantages of Cloud Computing :

- Back-up and restore data
- Improved collaboration
- Excellent accessibility
- Low maintenance cost
- On-Demand Self-service.

Disadvantages of Cloud Computing :

- Vendor lock-in
- Limited Control
- Security Concern
- Downtime due to various reason
- Requires good Internet connectivity.

bigdata vs cloud

Difference between Big Data and Cloud Computing :

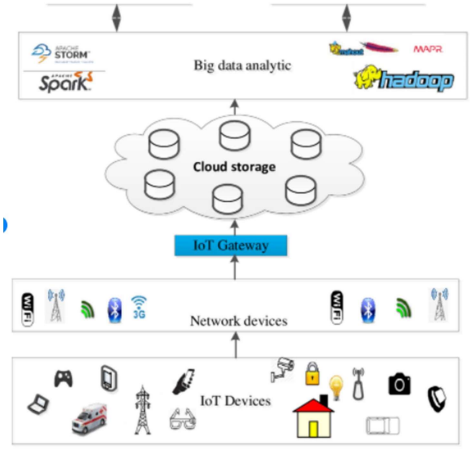
S.No.	BIG DATA	CLOUD COMPUTING
01.	Big data refers to the data which is huge in size and also increasing rapidly with respect to time. def	Cloud computing refers to the on demand availability of computing resources over internet.
02.	Big data includes structured data, unstructured data as well as semi-structured data. types	Cloud Computing Services includes Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS).
03.	Volume of data, Velocity of data, Variety of data, Veracity of data, and Value of data are considered as the 5 most important characteristics of Big data.	On-Demand availability of IT resources, broad network access, resource pooling, elasticity and measured service are considered as the main characteristics of cloud computing. kolichinatha service
04.	The purpose of big data is to organizing the large volume of data and extracting the useful information from it and using that information for the improvement of business.	The purpose of cloud computing is to store and process data in cloud or availing remote IT services without physically installing any IT resources.
05.	Distributed computing is used for analyzing the data and extracting the useful information.	Internet is used to get the cloud based services from different cloud vendors.
06.	Big data management allows centralized platform, provision for backup and recovery and low maintenance cost.	Cloud computing services are cost effective, scalable and robust.
07.	Some of the challenges of big data are variety of data, data storage and integration, data processing and resource management.	Some of the challenges of cloud computing are availability, transformation, security concern, charging model.
08.	Big data refers to huge volume of data, its management, and useful information extraction.	Cloud computing refers to remote IT resources and different internet service models.
09.	Big data is used to describe huge volume of data and information.	Cloud computing is used to store data and information on remote servers and also processing the data using remote infrastructure.
10.	Some of the sources where big data is generated includes social media data, e-commerce data, weather station data, IoT Sensor data etc.	Some of the cloud computing vendors who provides cloud computing services are Amazon Web Service (AWS), Microsoft Azure, Google Cloud Platform, IBM Cloud Services etc.

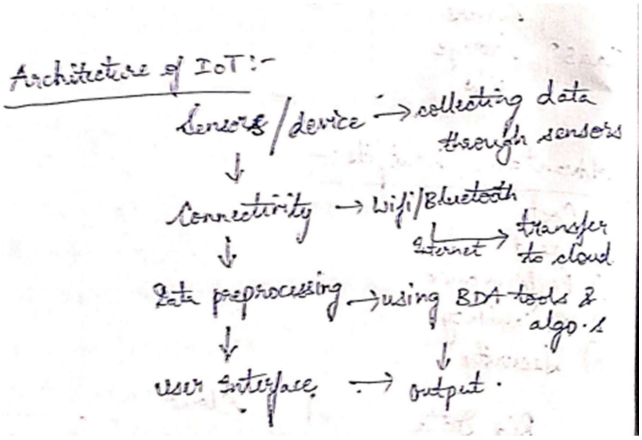
iot prelimanaries

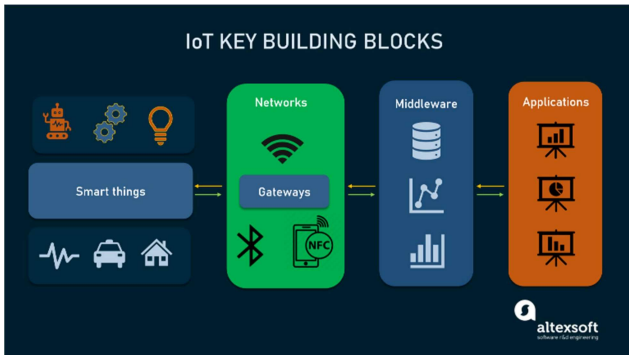
IoT(Internet of Things):

The Internet of Things (IoT) describes the network of physical objects—"things"—that are embedded with sensors, software, and other technologies for the purpose of connecting and exchanging data with other devices and systems over the internet.

IoT Architecture:







These elements make up the backbone of any IoT system upon which effective, multi-layered architecture can be developed. Most commonly, these layers are:

- the **perception layer** hosting smart things;
- the **connectivity or transport layer** transferring data from the physical layer to the cloud and vice versa via networks and gateways;
- the **processing layer** employing IoT platforms to accumulate and manage all data streams; and
- the **application layer** delivering solutions like analytics, reporting, and device control to end users.

iot vs bigdata

Differences between IoT and Big Data.

No	IOT	BIG DATA
1	IoT is a global system of interrelated computing devices that are able to sense, collect, and exchange data over the Internet.	Big Data is described as large sets of data generated from a variety of sources that are so large to process using traditional techniques.
2	The concept is to provide interconnection between devices to create a smart environment thereby making machines smart enough to bypass human intermediaries.	The concept is to find insights in new and emerging types of data and content that lead to better decisions and strategic business moves.
3	IoT collects, analyzes, and processes data streams in real-time without any delay to make control decisions in an effective manner.	The data streams are not subjected to processing real-time and there is a delay between when the data is collected and when it is processed.
4	IoT involves analyzing machine-generated data such as sensors in home appliances and so on.	Big Data deals with human-generated data such as social media usage, photos, and videos, etc
5	IoT is about simultaneously collecting and processing data to make real-time decisions.	Big data is more into collecting and accumulating huge data for analysis afterward.
6	Using IoT you can track and monitor assets like trucks, engines, HVAC systems, and pumps. You can correct problems as you detect them.	With big data, you can analyze all the information you have about failures and start to uncover the root causes.