

Data :

1-bit

4-bit=Nibble

8-bit =1 byte

1024 byte=1 KB

1024 kbyte=1MB

1024 MB=1 GB

1024GB=1TB

1024 TB=1 PB

1024 PB=1 Exabyte

1024 EB=1 ZettaByte

1024 ZB=1 Yottabyte

☁ **Big Data** – is a problem.e.g: sale-->10 million user->5 data points --> 50 million data points.

Lets say we have data of person ->Shubham and 9309944683 now we have this data in my mind and now i am sending msg to shubham means we are processing the data.

Data which are very large in size is called Big Data. Normally we work on data of size MB(WordDoc ,Excel) or maximum GB(Movies, Codes) but data in Peta bytes i.e. 10^{15} byte size is called Big Data. It is stated that almost 90% of today's data has been generated in the past 3 years.

☁ **Sources of Big Data** :

1)Social networking sites: Facebook, Google, LinkedIn all these sites generates huge amount of data on a day to day basis as they have billions of users worldwide.

2)E-commerce site: Sites like Amazon, Flipkart, Alibaba generates huge amount of logs from which users buying trends can be traced.

3)Weather Station: All the weather station and satellite gives very huge data which are stored and manipulated to forecast weather.

4)Telecom company: Telecom giants like Airtel, Vodafone study the user trends and accordingly publish their plans and for this they store the data of its million users.

5)Share Market: Stock exchange across the world generates huge amount of data through its daily transaction.

☂ Tools :

1)Data Analytics :Rapidminer,Tablu,Apache Spark,Power BI,python jupyter notebook,SAS(statistical analysis system),Excel.

2)Visualization : Excel,Tablu ,Power BI

3)Hadoop : It is java based framework that is used to process and manage large datasets.
Hadoop -> Solution [process and store large amount of data -HDFS -Mapreduce]

Types of data :

1. Human Generated Data

- Blogs, Reviews, Emails, Pictures, Scientific Research, Medical Records
- Social Graphs: Facebook, Linked-in, Contacts, Twitter

2. Computer Generated Data

- Application server logs (web sites, games, internet)
- Sensor data (weather, atmospheric science, astronomy, smart grids)
- Images/videos (traffic, security cameras, military surveillance)

- ♦ Even we keep our location turned on it record geospatial data.

- ♣ Structured : formal data model (tabular format) e.g : database
- ♣ Unstructured : not have formal data model e.g : video ,audio
- ♣ Semistructured : xml file

Big Data Sources

Data is pouring from every direction

This flood of data is coming from many sources:

- New York Stock Exchange generates 4-5 TB of data per day
- The Internet Archive stores around 18.5 petabytes of data
- Almost as many cell-phone(6.8 billion) as there are people on this earth(billion)
- YouTube users upload > 48 hours of video every minute
- Twitter 12 terabytes of Tweets every day
- Yahoo:60 PB, eBay:40 PB of data
- Ancestry.com, the genealogy site, stores 10 PB of data
 - All figures are from 2013 or 2014

expandedramblings
Learning for Life

A peek into quantum of data generated by Facebook

- 1.39 billion monthly active users
- 300 petabytes of user data
- 1.13 trillion Likes since 2004, 4.5 billion Likes daily
- 250 billion photos, growing at 7 petabytes per month
- 4.75 billion items shared per day (status updates + wall posts + photos + videos + comments)
- 405 minutes/visitor/month
- 10 billion Messages per day
- 500+ terabytes of new data ingested into DBs every day

Source: From Aug-2015

expandedramblings.com

statista.com/statistics/273563

☂ 3V's of Big Data

1. **Velocity:** The data is increasing at a very fast rate. It is estimated that the volume of data will double in every 2 years.
2. **Variety:** Now a days data are not stored in rows and column. Data is structured as well as unstructured. Log file, CCTV footage is unstructured data. Data which can be saved in tables are structured data like the transaction data of the bank.
3. **Volume:** The amount of data which we deal with is of very large size of Peta bytes.

Now V's Becomes 6

4. **Varacity** : - real life e.g suppose there is person who went to zomato and given review as 5 star and that review that person individual review. and suppose another person went to same shop looking that review but that person is not satisfied with that restaurant ..in such cases inconsistent and ambiguous data is there because the first person has not mentioned minute details like time of order and what specific item that person ordered it makes data inconsistent.

5.**Variability** : example is sensor thermometer because it captures data at certain interval and it has certain error rate.(not whole data is accurate there are some fluctuations)

6.**Value** : is about as we spent millions of rupees to extract insight and making reports now is it valuable.if we are not able to extract insights then it impacts on revenue of business.

✦ Use case :

An e-commerce site XYZ (having 100 million users) wants to offer a gift voucher of 100\$ to its top 10 customers who have spent the most in the previous year.Moreover, they want to find the buying trend of these customers so that company can suggest more items related to them.

✦ Issues :

Huge amount of unstructured data which needs to be stored, processed and analyzed.

✦ Solution

Storage: This huge amount of data, Hadoop uses HDFS (Hadoop Distributed File System) which uses commodity hardware to form clusters and store data in a distributed fashion. It works on Write once, read many times principle.

Processing: Map Reduce paradigm is applied to data distributed over network to find the required output.

Analyze: Pig, Hive can be used to analyze the data.

Cost: Hadoop is open source so the cost is no more an issue.

Characteristics of Big Data – 3 V's is now 6 V's

Initially big data was just about having lots of data to play with...

...since then, more attributes have been added to define big data...

..., but from enterprise standpoint the key is in **VALUE!**

1. **Volume:** Huge volume of data is being generated
2. **Velocity:** The speed at which data comes. Devices like RFID, Smart metering send accelerated data in real time
3. **Variety:** 80% of data is semi structured or unstructured
4. **Veracity:** Uncertainty on correctness of data due to ambiguity, inconsistency, latency. 1-in-3 business leaders don't trust the information they use to make decisions
5. **Variability:** Data flow is inconsistent with periodic peak. The same tweets, a word can have totally different meaning based on the context.
6. **Value:** Extracting business insights and revenue from data

Veracity: Uncertainty on correctness of data due to ambiguity, inconsistency, latency. 1-in-3 business leaders don't trust the information they use to make decisions







Why did Big Data suddenly become so prominent?

Reason why Big Data is Appealing:

1. manage and process a huge amount of data cost efficiently
2. analyzes data in all forms un-/semi-/structured, streaming, historical
3. captures data from fast-happening events in real time
4. handle failure of isolated nodes and tasks assigned to them
5. turn data into actionable insights

greatlearning
Learning for Life

Various Use Cases for Big Data

Financial Services <ul style="list-style-type: none">• Detect fraud• Model and manage risk• Improve debt recovery rates• Personalize banking/insurance products 	Healthcare <ul style="list-style-type: none">• Optimal treatment pathways• Remote patient monitoring• Predictive modeling for new drugs• Personalized medicine 
Retail <ul style="list-style-type: none">• In-store behavior analysis• Cross selling• Optimize pricing, placement, design• Optimize inventory and distribution 	Web / Social / Mobile <ul style="list-style-type: none">• Location-based marketing• Social segmentation• Sentiment analysis• Price comparison services 
Manufacturing <ul style="list-style-type: none">• Design to value• Crowd-sourcing• "Digital factory" for lean manufacturing• Improve service via product sensor data 	Government <ul style="list-style-type: none">• Reduce fraud• Segment populations, customize action• Support open data initiatives• Automate decision making 

Big Data Touch

Challenges of Big Data

1. Store the sheer size of Big Data
2. Process the huge data - No point in just storing big data if we can't process it
3. Handle the variety of data - Big Data is unstructured or semi structured
4. Scalability
5. Cost