Introduction to Big Data
Lecture 1

Outline

Data and Information 01

Data Units 02

What is Small Data 03

What is Big Data 04

Small data vs Big Data 05

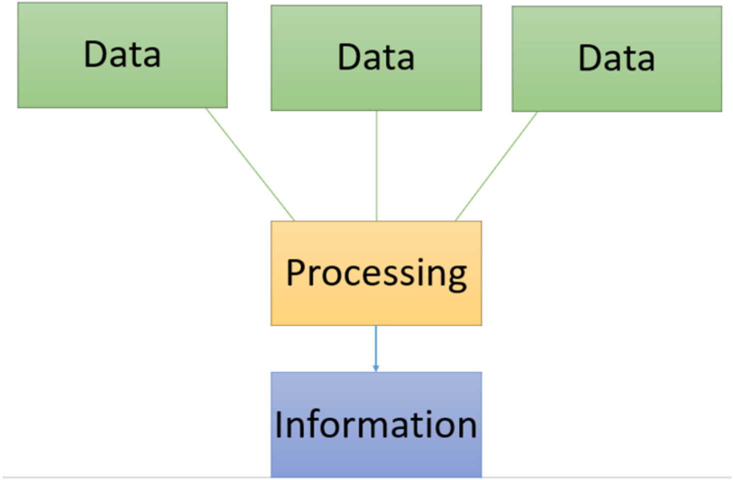
How Big data comes 06

Types of Big Data 07

Big data Characteristics 08

Data and information

required to be processed to make it meaningful
Data is a raw and unorganized fact that



Data

Data

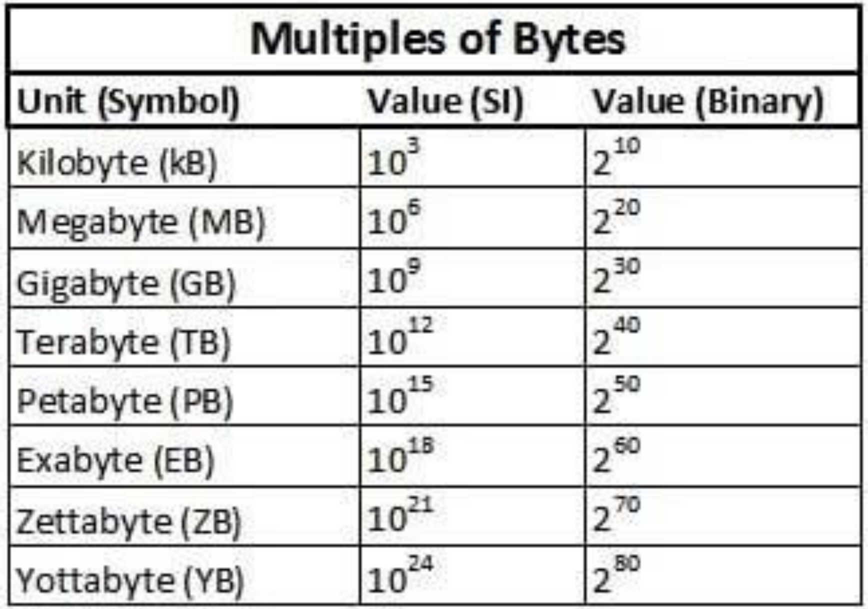
Data

Processing

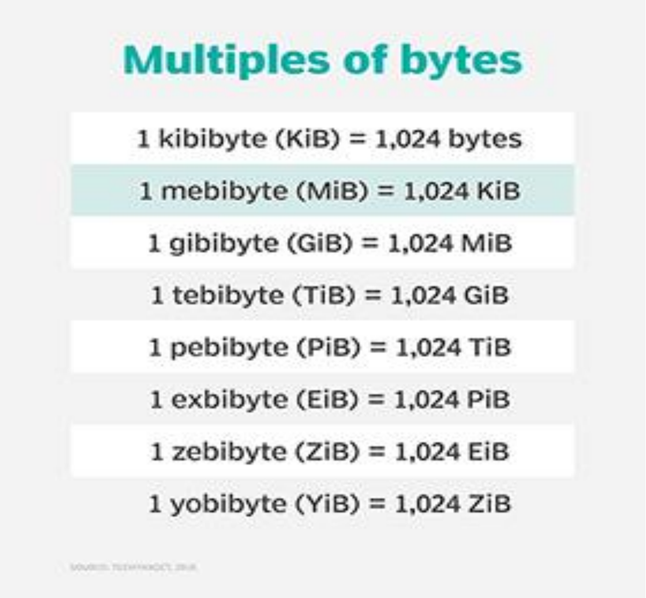
Information

Information is a set of data which is processed in a meaningful way according to the given requirement.

Data Units



Multiples of Bytes		
Unit (Symbol)	**Value (SI)**	**Value (Binary)**
Kilobyte (kB)	10^3	2^{10}
Megabyte (MB)	10^6	2^{20}
Gigabyte (GB)	10^9	2^{30}
Terabyte (TB)	10^{12}	2^{40}
Petabyte (PB)	10^{15}	2^{50}
Exabyte (EB)	10^{18}	2^{60}
Zettabyte (ZB)	10^{21}	2^{70}
Yottabyte (YB)	10^{24}	2^{80}



Multiples of bytes

1 kibibyte (KiB) = 1,024 bytes

1 mebibyte (MiB) = 1,024 KiB

1 gibibyte (GiB) = 1,024 MiB

1 tebibyte (TiB) = 1,024 GiB

1 pebibyte (PiB) = 1,024 TiB

1 exbibyte (EiB) = 1,024 PiB

1 zebibyte (ZiB) = 1,024 EiB

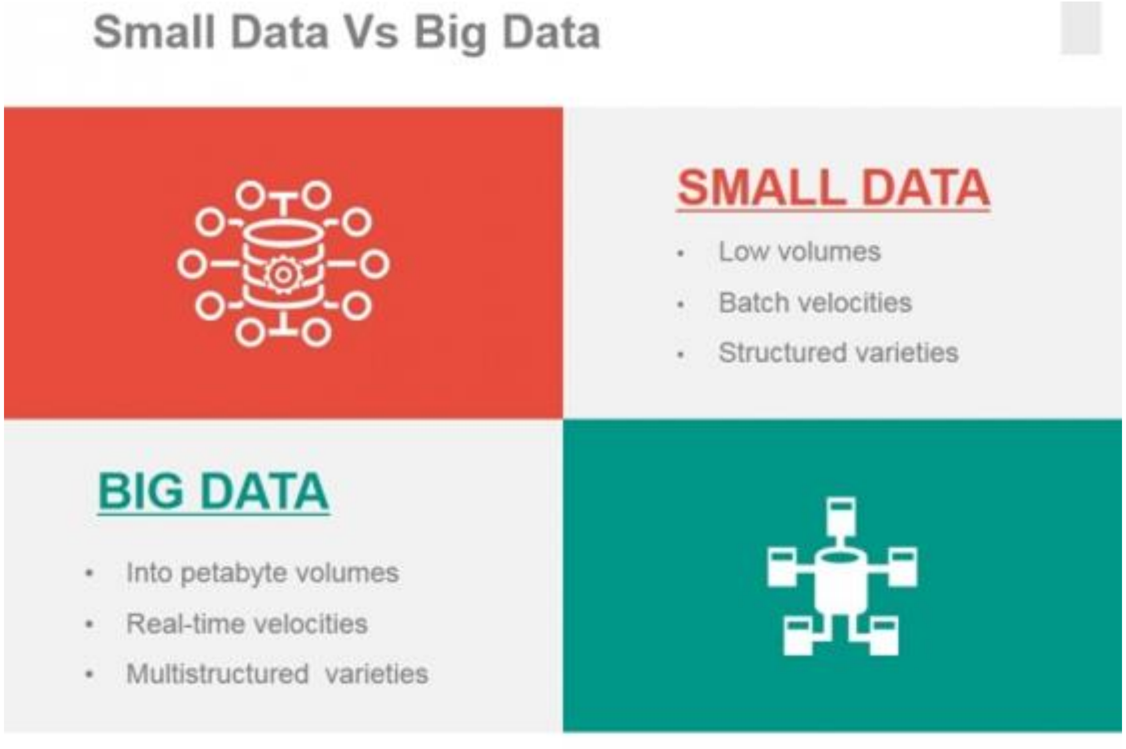
1 yobibyte (YiB) = 1,024 ZiB

[illegible]

What is Big Data?

Big data is the term for collection of data sets so large and complex that it becomes difficult to process using on-hand database system tools or traditional data processing applications

Small data vs Big data



Small Data Vs Big Data

SMALL DATA

- Low volumes
- Batch velocities
- Structured varieties

BIG DATA

- Into petabyte volumes
- Real-time velocities
- Multistructured varieties

The term "**big data" is about machines and "small data" is about people**.

How Big Data Comes

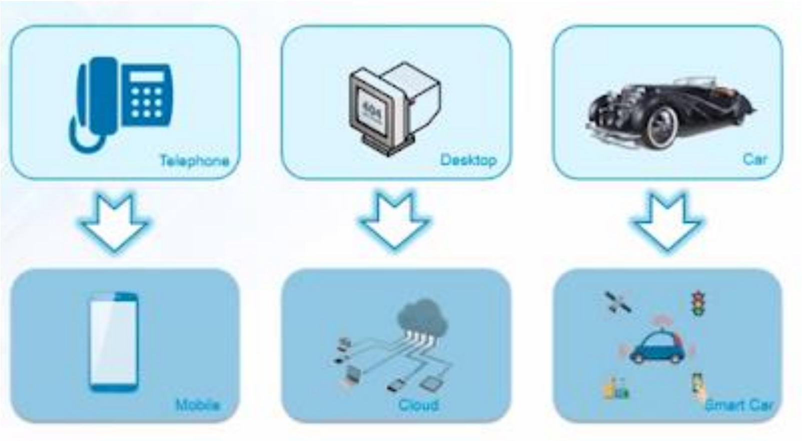
The below are the reasons behind the big data comes in picture:

1. Evolution of technology
2. IOT(Internet Of Things)
3. Social Media
4. Other factors

How Big Data Comes

1)Evolution of technology:

- Earlier we had landline phones, But nowadays,we have android,IOS smartphones, to make our life smarter. so just think, for each operation which we perform on smartphones, generates a data, that resides somewhere
- Desktops are the source to handle operations, i mean to store and process using storage devices like floppy,discs,taps,..etc. But in these days, Hard disks,cloud storage plays a vital role.
- Earlier , we are in the hand of Analog storage, but these days almost of Digital storage. and also about the evolution of car, self driving car,



How Big Data Comes

2)IOT(Internet Of Things):

IOT connects physical device to Internet and makes device smarter.

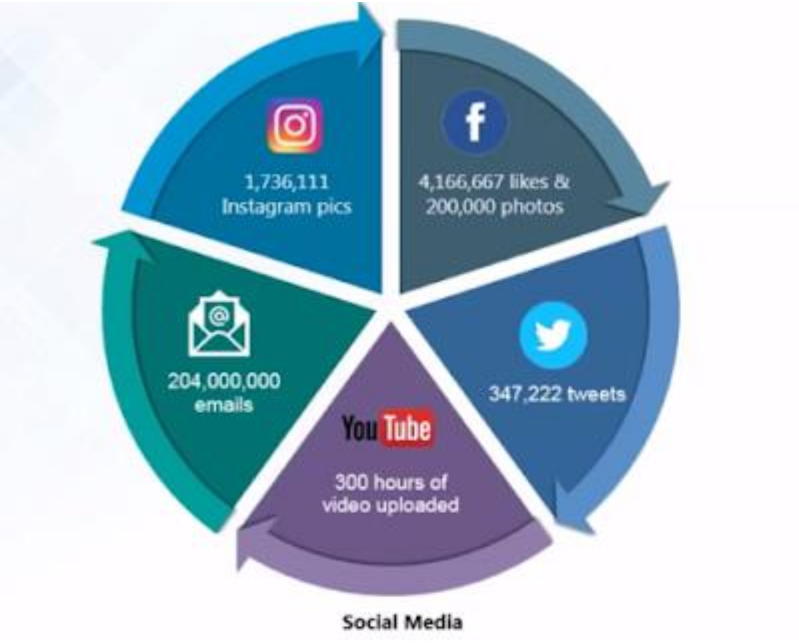
Example:
Smart TV's, Smart Ac's, Smart Car's etc.,

How Big Data Comes

3)Social Media:

Data generation on social media sites,

- Facebook likes,videos,photos,tags,comments etc.,
- Tweeter tweets,
- Youtube video uploads
- Instagram pics,
- Emails



How Big Data Comes

4)Other Factors:

- Retail
- Banking & Finance,
- Media & Entertainment
- Health care,
- Education areas,
- Government,
- Transportation, Insurance etc.

Retail
Banking & Finance
Insurance
Media & Entertainment
BIG DATA
Transportation
Healthcare
Government
Education

Types of Big Data

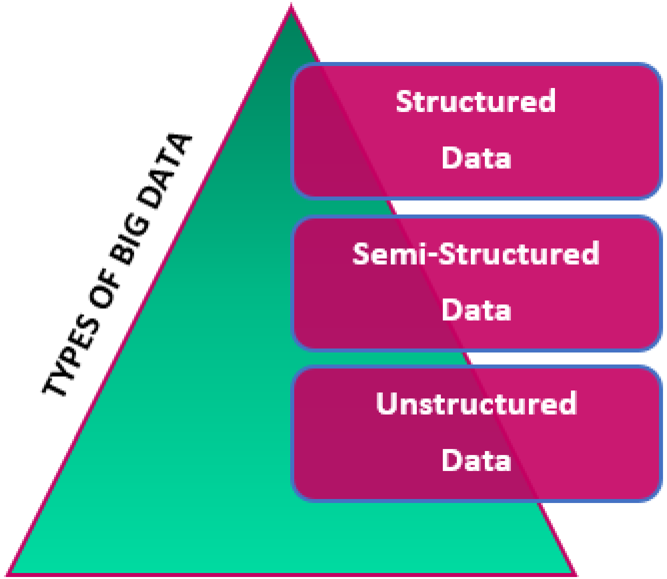
1. Structured data

- As the name suggests, this kind of data is structured and is well-defined. It has a consistent order that can be easily understood by a computer or a human. This data can be stored, analyzed, and processed using a fixed format. Usually, this kind of data has its own data model.
- You will find this kind of data in databases, where it is neatly stored in columns and rows. Two sources of structured data are:
  - **Machine-generated data** – This data is produced by machines such as sensors, network servers, weblogs, GPS, etc.
  - **Human-generated data** – This type of data is entered by the user in their system, such as personal details, passwords, documents, etc. A search made by the user, items browsed online, and games played are all human-generated information.
- For example, a database consisting of all the details of employees of a company is a type of structured data set.

Types of Big Data

Big Data could be of three types:

- Structured
- Semi-Structured
- Unstructured



Types of Big Data

2. Unstructured data

- ❑ Any set of data that is not structured or well-defined is called unstructured data. This kind of data is unorganized and difficult to handle, understand and analyze. It does not follow a consistent format and may vary at different points of time. Most of the data you encounter comes under this category.
- ❑ For example, unstructured data are your comments, tweets, shares, posts, and likes on social media. The videos you watch on YouTube and text messages you send via WhatsApp all pile up as a huge heap.

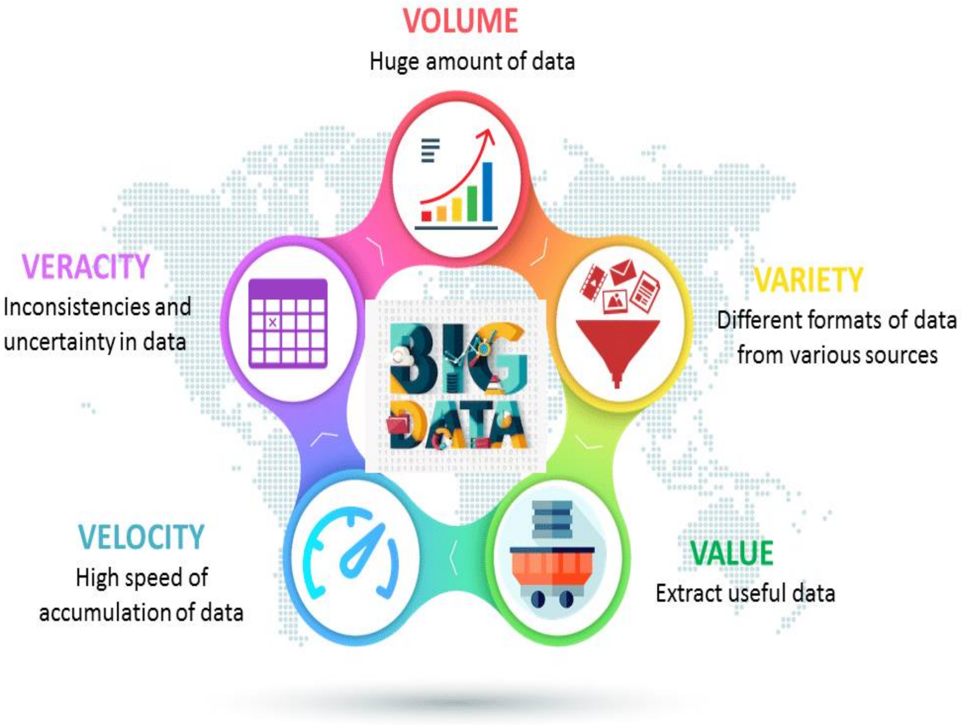
3. Semi-structured data

- ❑ This kind of data is somewhat structured but not completely. This may seem to be unstructured at first and does not obey any formal structures of data models such as RDBMS. For example, NoSQL documents have keywords that are used to process the document.
- ❑ CSV files are also considered semi-structured data.of unstructured data.

Big Data Characteristics

5 V's of Big Data

- Volume
- Veracity
- Variety
- Value
- Velocity


VOLUME
Huge amount of data
VERACITY
Inconsistencies and uncertainty in data
VARIETY
Different formats of data from various sources
BIG DATA
VELOCITY
High speed of accumulation of data
VALUE
Extract useful data

Big Data Characteristics

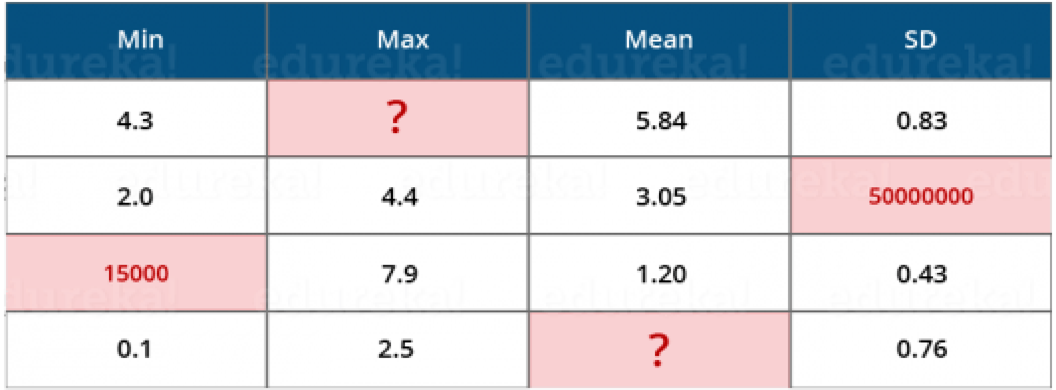
- **VOLUME** - Volume refers to the ‘amount of data’, which is growing day by day at a very fast pace. The size of data generated by humans, machines and their interactions on social media itself is massive. **Facebook** can generate approximately a **billion** messages, **4.5 billion** times that the "**Like**" button is recorded, and more than **350 million** new posts are uploaded each day. Big data technologies can handle large amounts of data

Amount of data being generating and generated.

Big Data Characteristics

- **Veracity** - Veracity refers to the data in doubt or uncertainty of data available due to data inconsistency and incompleteness. In the image below, you can see that few values are missing in the table. Also, a few values are hard to accept, for example – 15000 minimum value in the 3rd row, it is not possible. This inconsistency and incompleteness is Veracity.

Uncertainty and inconsistencies in the data, i.e., The quality of captured data can vary greatly, affecting accurate analysis.

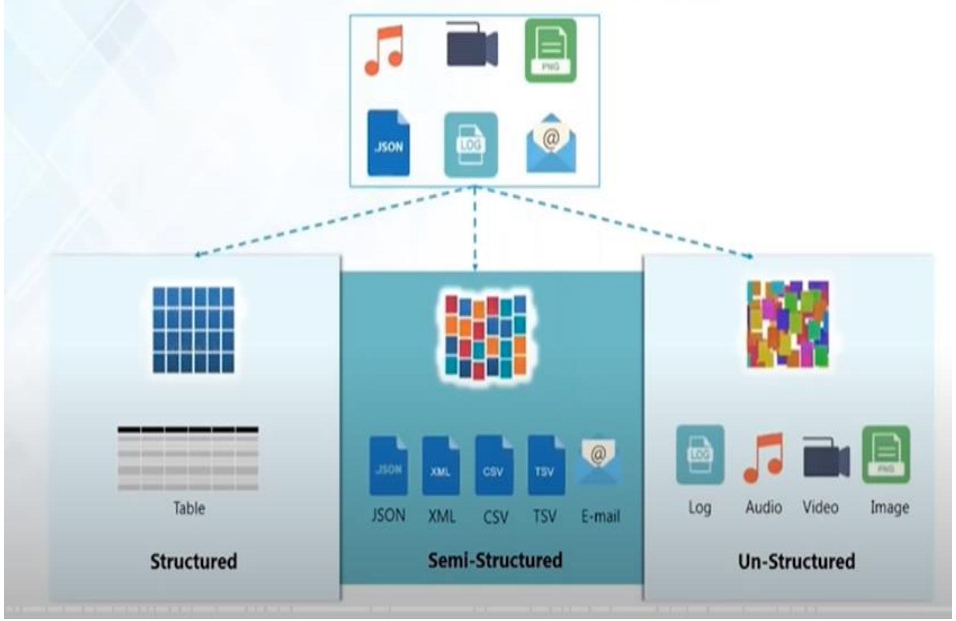


Min	Max	Mean	SD
4.3	?	5.84	0.83
2.0	4.4	3.05	50000000
15000	7.9	1.20	0.43
0.1	2.5	?	0.76

Big Data Characteristics

- **Variety** - As there are many sources which are contributing to Big Data, the type of data they are generating is different. It can be structured, semi-structured or unstructured. Hence, there is a variety of data which is getting generated every day. Earlier, we used to get the data from excel and databases, now the data are coming in the form of images, audios, videos, sensor data etc. as shown in image. Hence, this variety of unstructured data creates problems in capturing, storage, mining and analyzing the data.

Different kinds of data , that is being generated from various sources

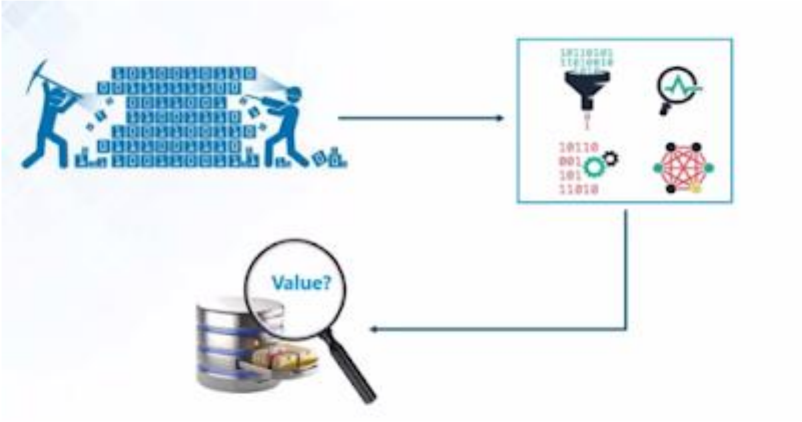


Big Data Characteristics

- **Value**- Among the characteristics of Big Data, value is perhaps the most important. No matter how fast the data is produced or its amount, it has to be reliable and useful. Otherwise, the data is not good enough for processing or analysis. Research says that poor quality data can lead to almost a 20% loss in a company's revenue.

  Data scientists first convert raw data into information. Then this data set is cleaned to retrieve the most useful data. Analysis and pattern identification is done on this data set. If the process is a success, the data can be considered to be valuable.

Mechanism to bring the correct meaning out of data

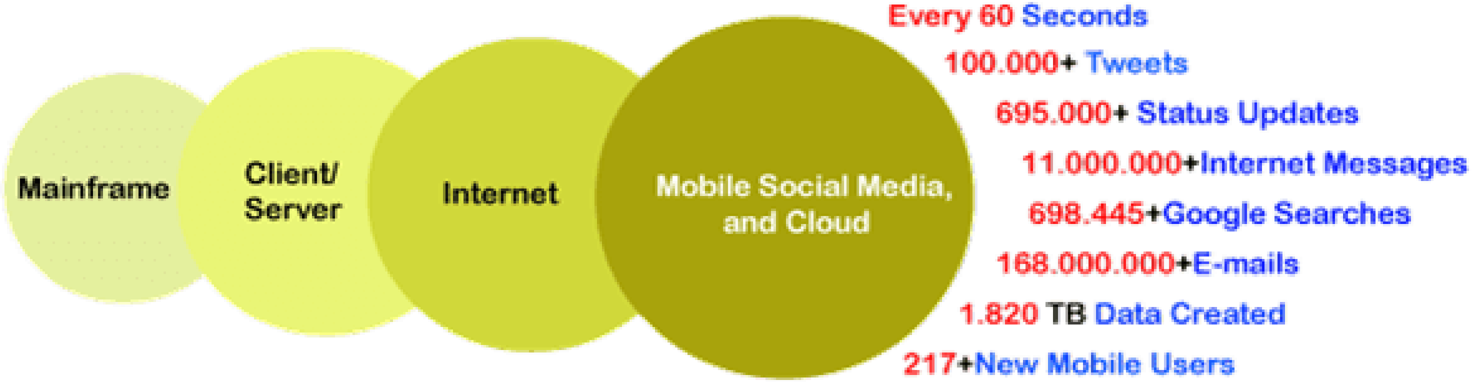


Big Data Characteristics

- **Velocity-** The term ‘velocity’ refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data.

  Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks, and social media sites, sensors, Mobile devices, etc. The flow of data is massive and continuous.

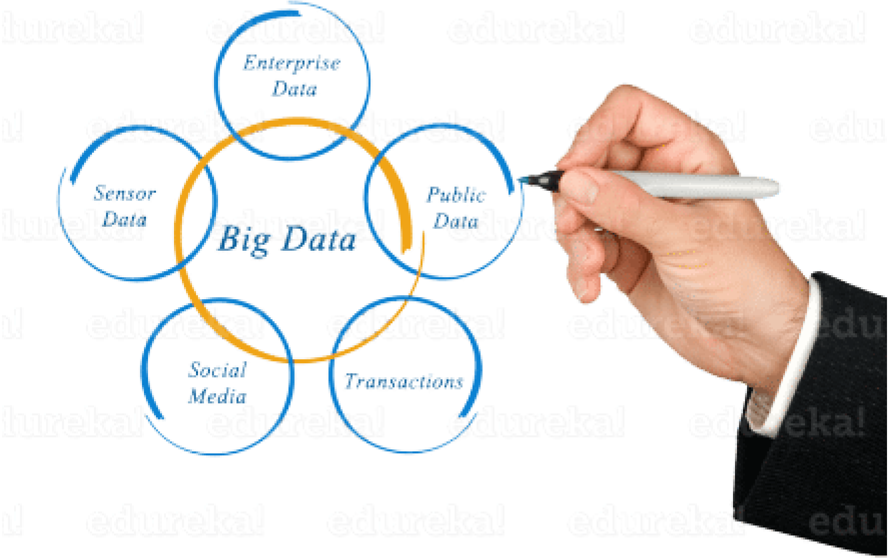
Data is being generated in an alarming rate



Every 60 Seconds
100.000+ Tweets
695.000+ Status Updates
11.000.000+Internet Messages
698.445+Google Searches
168.000.000+E-mails
1.820 TB Data Created
217+New Mobile Users

Examples of Big Data

Daily we upload millions of bytes of data. 90 % of the world's data has been created in last two years.



Thank You