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.

- **Souces 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.

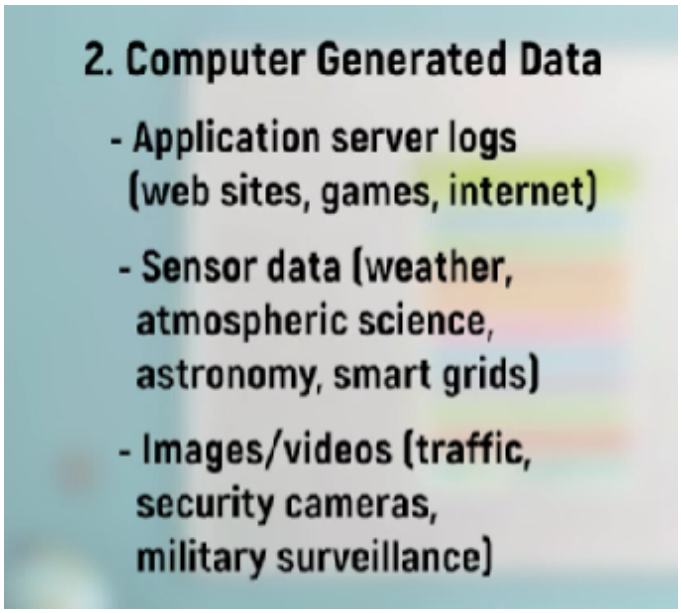
- **Tools :**

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

2)Visualization : Excel,Tablu ,Power BI

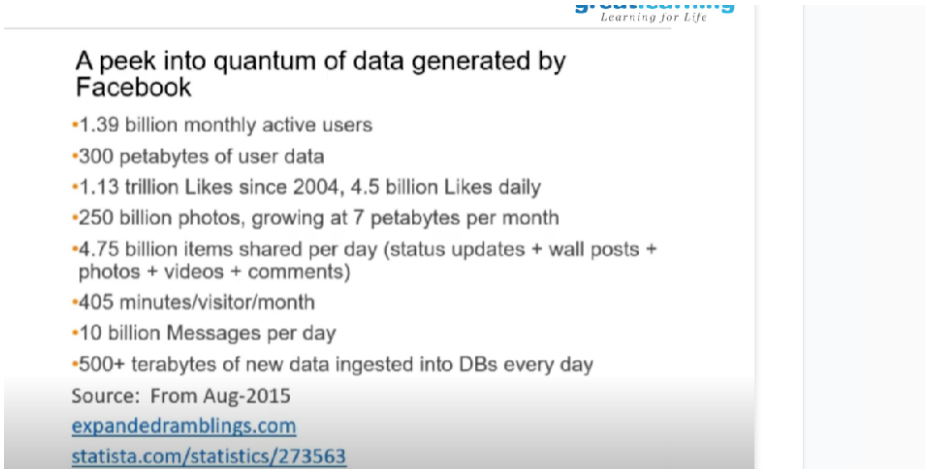
3)Hadoop : It is java based framework that is used to process and manage large datasets. Hadoop -> Solution [process and store large amounrt of data -HDFS -Mapreduce]

Types of data :



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

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



- 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 indivisual review. and suppose anouther person went to same shop looking that review but that person is not satisfied with that restarent ..in such cases inconsistant and ambiguious data is there because the first person hv not mentioned minute details like time of order and what specific item that person ordered it makes data inconsistant.

5.**Variability :** exammple is sensor thermometer because it capture 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 milions of rupess to extract insight and making reports now is it valueable.if we are not able to ecxtract insights then it impact on revennue of buisness.

- 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.

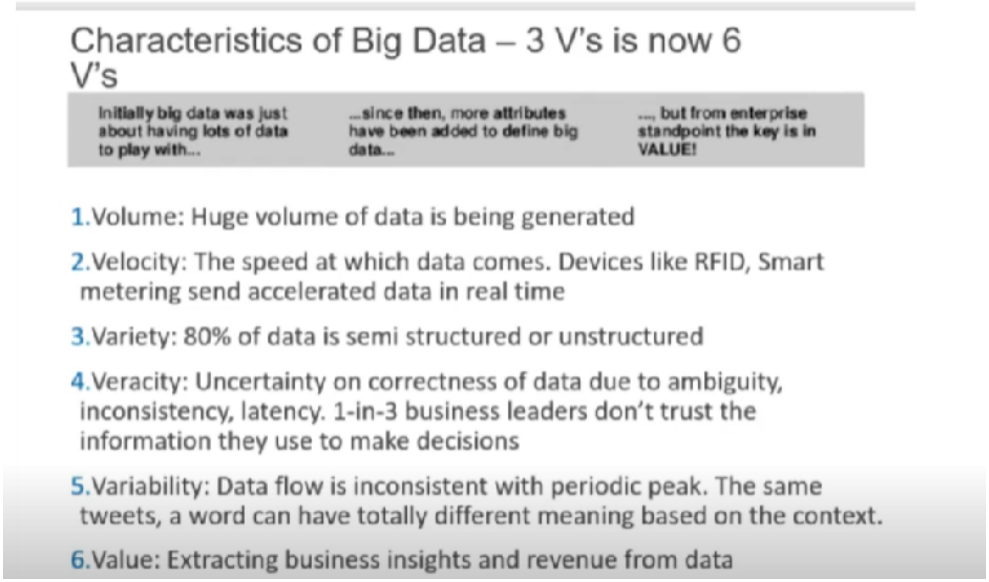
- Solution

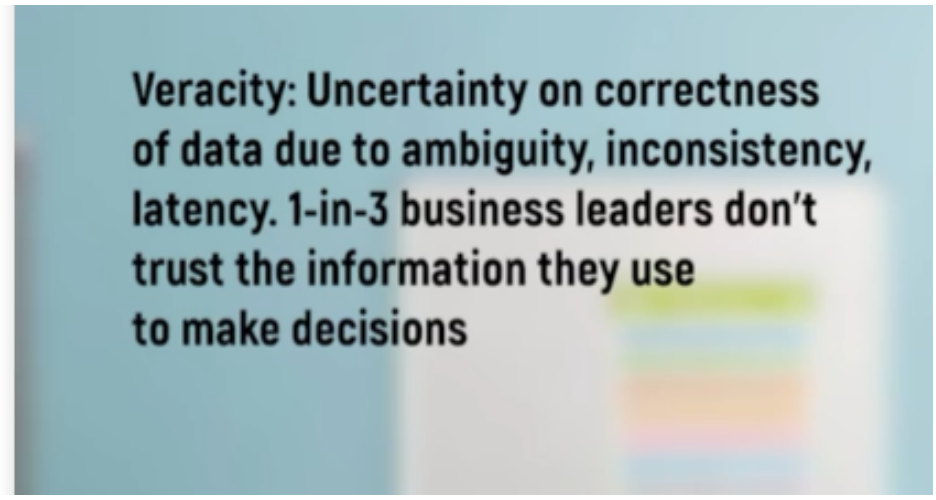
Storage: This huge amount of data, Hadoop uses HDFS (Hadoop Distributed File System) which uses commodity hardware to form clusters and store data in a distributed fashion. It works on Write once, read many times principle.

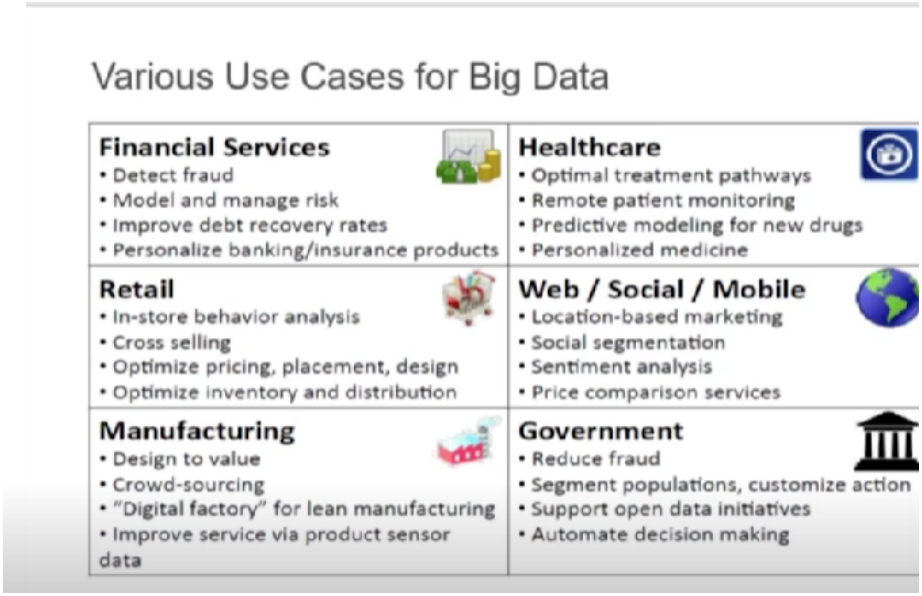
Processing: Map Reduce paradigm is applied to data distributed over network to find the required output.

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

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







Big Data Touch