Big Data Analytics

Huge amt of existing data. Knowledge every day

Big Data
- unstructured — images, messages, videos, audio
- structured (rows & columns) — dbms
- semi-structured — HTML, JCC scripts, XML, JSON

80% of data is unstructured

10% & 10% semi & structured respectively

Advantages of structured data

1) insert / update / delete
2) security
3) indexing — ease operation of retrieval
4) scalability — store large amt. of data
5) transaction processing

ACID properties

A — Atomicity — values should be simple
C — consistency — DB should be consistent
I — Isolation — each & every transactions are executed
D — Durability — changes should be made available until transaction is done.

Semi structured data

All features are available
- data in form of key value pair
- structure of data is inconsistent
- data objects will have different attributes
- schema info along with data values

unstructured data

unstructured data —Processing→ structured data —operations→

ways of conversion
- text analytics
- Data mining
- NLP (Natural language processing)

Scanned by CamScanner

techniques for extracting pattern | interpreting

① Data mining

association rule Mining

regression analysis — independent — dependent } relation

collaborative filtering
(preference of single user using group of users)

Characteristics of Big data

① composition - types & nature of resources of the data available

② condition - state of the data (whether directly the data can be used or requires cleaning or preprocessing

③ context - how the data is generated? how sensitive the data is? whether its updated or old data.

Evolution of Big data:

→ complex data or simple data

challenges of Big data:

① whether the data that is being generated is useful

whether we are able to remove noise from the data & use it

② cost effective? whether it can be upgraded

③ period of keeping data ie whether the data is useful for short term or long term

④ Implementation of Big data

⑤ securing of BD, transfer data & search the data, store the data, capture the data.

3 Vs of Big data — veracity - trust data value

volume (large amt of data available)

velocity (how speed it is generated, from src to dest how fast data can travel)

variety (different kinds of data that is being generated, structured, semi structured & unstructured)

difference b/w traditional Business Intelligence & BD

Business intelligence	Big data
- data is present in a central server.	- data is distributed i.e distributed file s/y
- analysed offline	- analysed both offline & in real time
- only structured form of data is taken as i/p	- all forms of data is taken as i/p

challenges of Big data

1. scale - it should be possible to scale, the size can be anything but we must store it
2. security
3. schema - proper schema must be supported
4. continuous available (data must be available 24x7)
5. Consistent data.
6. Partition tolerant (robust s/y)
7. Data quality, completeness, accurate, timeliness