Big Data Analysis

→ It is a type of data whose scale, diversity & complexity require new architecture, techniques, algorithms & analytics to manage it and extract value & hidden knowledge from it.

* **Big data pipeline**: → set of processes that move data from one place to another

Data Sources	Data Integration	Data store	Analyze	Delivery
↳ App	↳ ETL	↳ MDM	↳ Analytics	↳ Dashboards
↳ Mob app	↳ Stream Data integration	↳ Ware house	↳ ML	↳ Reports
↳ Microser			↳ Predictive analytics	↳ Push Notif
↳ IOT Devices		↳ Data Lake		↳ Email
↳ Website				↳ SMS

Collection → Ingestion → Preparation → Computation → Presentation

* **BDA Cycle**:

1) Business Pb Def
2) Data identification
3) " Acquisition & filtering
4) " Extraction
5) Exploratory Data Analysis
6) Data Prep for Modeling & Assessment
7) " Visualization
8) Analysis of Results

* **Data Science**. → multi-disciplinary & emerging field that uses scientific methods, processes, algorithms & systems to extract knowledge & insights from structured & non stru data.

Goal: Turn data into data products & create business value.

Analysis
Analytics

GB → TB → PB → EB → ZB

	Traditional data	Big data
Volume	GB	Constantly updated (PB to TB daily)
Data gene rate	Per hour, day	More rapidly
Struc	Structured	Semi-str, unstru
Data source	Centralized	Fully distributed
" integr	Easy	Difficult
" store	RDBMS	HDFS, NoSQL
Access	Interactive	Batch / real time / near real time

* Characteristics of Big Data :-

1) Volume :
→ Volume of data, which needs to be processed is increasing rapidly.
→ More storage capacity
→ More computation
→ More tools & techniques

exp inc in collec gener data

2) Variety :
→ Various formats, types & structures
Eg text, audio, video, time series, social media data, seq, multi D array, etc
→ Static vs streaming data
→ A single app generate diff types of data
→ to extract knowledge these types of data need to be linked together.

3) Velocity :
→ Data is generated fast & need to be processed fast.
Eg time sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value

→ Big Data is high-volume, high velocity & high variety info assets that demand cost effective, innovative forms of info processing for enhanced insight & decision making.

* Types of Digital Data:

1) Structured: → data stored in organized form.

Sources → Spreadsheets
→ OLTP Sys
→ Database like oracle, DB2, MySql, postgresql, etc.

Ease → ip / up / del
→ Security
→ Indexing / searching
→ Scalability
→ Transaction Processing

2) Semi-Stru: data that does not conform to data model but has some structure.

Eg: emails, XML, HTML, metadata available but not sufficient.

Source → XML
→ other markup lang
→ JSON

Chara → Inconsistent
→ Self-describing (Table / value pairs)
→ Often schema info is blended with data values
→ Data objects may have diff attributes not known beforehand

3) Unstru :- not conform to data model
80 to 90 % of organization's data is in this form
eg chat rooms, ppt, img, videos, letters, etc.

Source → Web pages, img, free-form text, Audios, videos, body of email, Text msg, chats, social media data, word doc

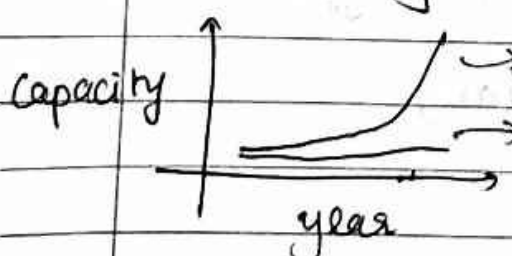
Dealing with Unstru → Data mining
→ NLP
→ Text Analytics
→ Noisy Text "

* Challenges in Big data :-

1) Storing exponentially growing huge datasets
→ Data generated in past 2 yrs is more than prev history in total
→ By 2020, digital data will grow upto 44 ZB
→ " " , new info of 1.7MB created every sec for every person.

2) Processing data having compler structure

3) Processing data faster :



→ Bringing huge amount of data to computation unit becomes a bottleneck.

* Classification of Data analytics :-

1) Descriptive Analytics → What happened ?
   → provide insights into past events
2) Diagnostic " → Why did it happen ?
   → to find cause of outcome
3) Predictive → what will happen next ?
   → leverages historical data & trends to predict future outcomes.
4) Prescriptive → what should be done about it ?
   → based on past events estimate likelihood of diff outcomes.

* Scalable platforms :-

/scale out	/scale in
1) Horizontal ^ Scaling	2) Vertical ^ Scaling
→ It involves distributing the workload across many servers.	→ It involves installing more processors, more memory & faster hlw, within a single server
→ multiple indep mac are added together in order to improve processing capability	→ a single instance of os.
→ multiple instances of os are running on separate mac	Adv → most of slw can easily take adv of vertical scaling
Adv → inc perf in small steps as needed	→ easy to manage & install hlw within a single mc
→ financial investment to upgrade is relatively less	Dis → requires substantial financial investment.
→ can scale out sys as much as needed.	→ sys has to be more powerful to handle future workloads & initially the additional perf is not fully utilized.
Dis → slw has to handle all the data distri & parallel processing complexities.	→ not possible to scale up vertically after certain limit.
→ limited no. of slw are available that can take adv of horizontal scaling	

Horizontal → Peer-to-Peer Network
→ Apache Hadoop
→ Apache Spark
Vertical Scaling platforms → High performance computing clusters
→ Multicore CPU
→ Graphics Processing Unit (GPU)
→ Field Programmable gate arrays (FPGA)

1) Peer-to-Peer Network

→ involve millions of machine connected in a network
→ decentralized & distributed archi where nodes in network serve as well as consume resources
→ oldest distributed computing platform.
→ Message Passing Interface (MPI) for communication
→ Each node can store data instances & scale out is unlimited.

2) Apache Hadoop

→ Open source framework for storing & processing large datasets using clusters of commodity hw
→ can scale up to 100s
→ highly fault tolerant
→ components → ① HDFS ② Yarn

3) Apache Spark

→ developed to overcome disk I/o limitations.
→ ability to perform in-memory computation
→ allows the data to be cached in memory thus eliminating Hadoop's disk overhead limitation for iterative tasks.
→ Supports Java, Scala & python
→ for certain tasks it is 100 times faster than Hadoop Map Reduce.

4) HPC clusters :-
→ blades or supercomputers with 1000s of cores.
→ diff type of disk org, cache, comm mechanisms
→ powerful hlw which is optimized for speed & throughput.
→ not as scalable as Hadoop or Spark but are still capable of processing terabytes of data.

5) Multicore CPU :
→ It refers to one machine having of processing cores, they have shared memory but only one disk
→ No. of cores per chip & no. of operations that a core can perform has increased significantly.
→ Newer breeds of motherboards allow multiple CPUs within a single machine thereby inc parallelism
→ CPUs are responsible for accelerating algos for BDA.

6) GPU :-
→ It is designed to accelerate the creation of images in a frame buffer intended for display d/r.
→ used for graphical operations such as ① video & img editing, ② accelerating graphics-related processing due to their massively parallel archi, ③ recent devel in GPU hlw has given rise to GPGPU,
→ In addition to processing cores, GPU has its own high throughput DDR_5 memory which is faster than DDR_3 memory.

7) ~~FGP~~ FPGA : → highly specialized hlw units for specific app.
→ FPGA can be highly optimized for speed compared to other plat
→ due to customize hlw, development cost is higher
→ On slw side, coding has to be done in HDL with a low-level knowledge of hlw which inc the algorithm development cost.