## Small Data

1970 - Sci EF Codd
Traditional db was designed
RDBMS:- was evolved
→ Structured data → Tables
/excel files → format
not used these days. (only
10% is structured these days)

unstructured - pics, videos, text

- Store in MB, GB, TB

→ Speed/velocity
(speed of data increase is
gradually)

- If x amount of data
is there. 90% of x is
generated recently in last
4-5 years.

- Stored in centralized form
(Structured form)
attendance, library data
- data is locally stored

- Sql server, Oracle,
- Single Node

## (size) Big Data

mostly unstructured (90%)

→ 1000TB → PB → EB
PetaByte Exabyte

FB, google db
4B data is generated daily
More than 2 Billion users
Storage is very high

- Speed of data generated
daily is very high

- Exponentially.
>350 billion pics are
uploaded
>4 million likes daily on
FB

- data centres across the
world. by distributing
globally. Entire traffic
will come at single
place otherwise.

- Hadoop Spark
Mapreduce Bigquery
- Multinode cluster

These are under
apache now and are
Open to use.

⇓

for DS, ML, AI

↓

train - expert s/m

If Amazon: many orders → transactions – how to retain → by big data.

- Big data cant be processed by existing traditional techniques
- Sources :- Stock Exchange (many companies – their own price – their own status) – Big data
  - Social Media (Fb, WA, Instagram) Billions of users – post – their own data. (Generated daily)
  - Video Sharing Portal: You tube → food, travel, etc. many videos – high data
  - Search Engine: Yahoo/google if most searched football player – comes from db
  - Transport: no, owner, distance
  - Banking: we have our account – Transactions

Semi structured data – neither structured nor unstructured – that it can be stored in a Relation
→ JSON, XML is used in this case

Application:
① Call centre (requirements) phn – 2 slots of sim, 2 sim
high data w.r.t. every no → Data is analysed

|R1 | R2 | R3|
D D D

Acc. to this calls are made.
Quality Evaluation – recording is also done.

② Social Media :- like – opinion (Companies take their decision for Promotional Campaigns – Sales ↑ – Profit ↑

②

- Shopping : Amazon, flipkart - Playstore
  variety of items - recommendation
- IOT : Smart devices - connect - functionality
  |
  achieve a goal

Data - sent on cloud - decision

eg Healthcare eg ~~Human~~ (P1) Disease.
Medication is given to all patients
↓
Sensors are used to capture improvement
↓
decision to continue vaccine or not.

- Education :- online exam.
  Particular ~~time~~ - behavior is recorded.
- how many qs u take to solve easily
- how many do u leave difficult (-ve marking)
- easy option - behavioral analysis
  1 hr lecture - how many pay attention
  facial behavior - how much % attended
- Entertainment : Netflix.
  earlier only used in foreign - no indian conten
  ↓
  demand of hindi movies - data is generated.

Challenges of Conventional Systems :-

→ Store → Manage → Analyse
→ Specific Time Interval

① large quantities → impossible to process every
② Must be meaningful & shd be collected in real time
③ Data is collected from Multiple sources

eg Text, Img, Video, Audio, Video => Bring it together. Data shd be categorised properly. Requires lot of manpower + time.

- Time can be long.

(4) Collect correct data (Remove incompleteness/ wrong data else it creates problem.

(5) Comparison of data using statistics /graphs etc (Complicated job)

(6) Managers must be able to access data

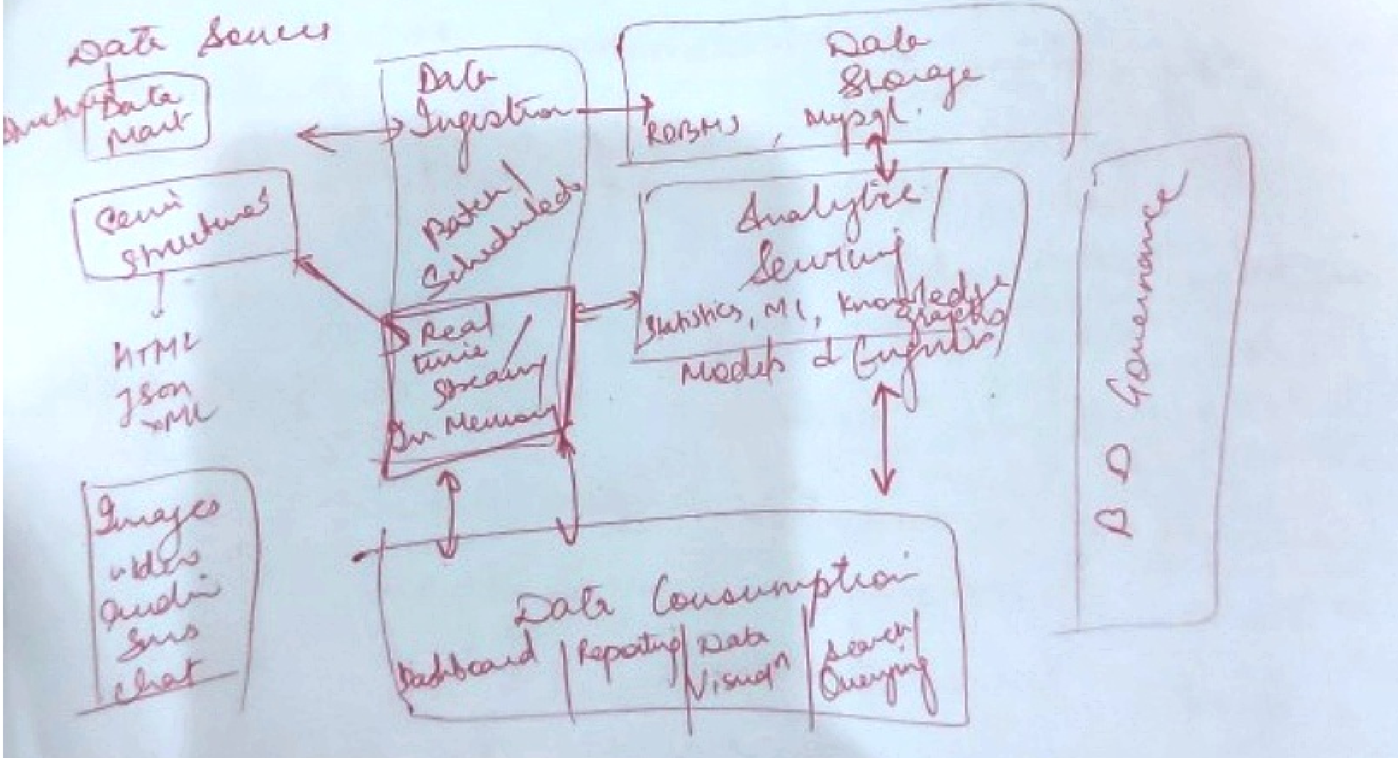
(7) Pressure from top (Pressure on risk managers)

(8) Lack of knowledgeable professionals

To overcome => Intelligent Data Analysis is used

Rederces time, cost & error.

## BIG DATA ARCHITECTURE

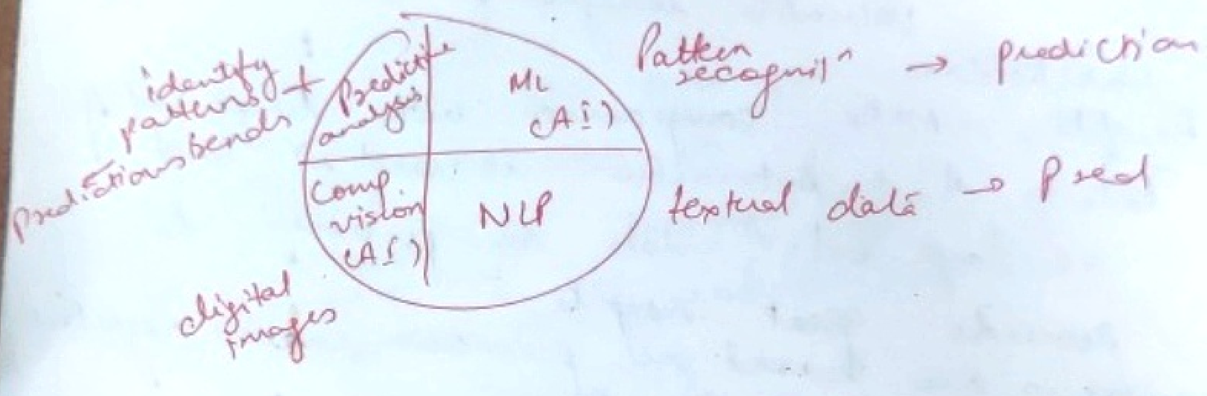


## Technology components

4 components:-

- data capture (social media, sensor readings)
- data storage
- data processing
- data visualizⁿ

## fields



## Technologies

① Apache Hadoop (OS framework) enables distributed processing of large ds on servers. Scalability, CE, flexibility

ML ② Apache Spark OS processing engine (batch & R.T. analytics on large ds) used with Hadoop

③ Apache Flink OS stream processing framework. high speed analytics on live data streams. User friendly API and scalable arch → growing community

④ Presto: OS SQL engine that supports interactive analytics on huge ds stored in multiple s/ms. Distributed query processing arch → low latency & strong performance

⑤ Druid: OS Analytical data storage designed for OLAP queries on event based data (eg log files, clickstreams)

fast aggregations & explorations on large ds due to columnar storage format

Others ⇒ Apache Hadoop, NoSQL db, MapReduce, Cloudera, Hortonworks, IBM Biginsights, MapR, Oracle Big Data Appliance.

Auditing + Analytics :

↑ use of Big data → red^n in Oper^n costs, Improved decision making, high customer retention ↑ Satisf^n rate.

Banking → Manuf.

RT big data → growth + sustainability

Auditing :- Past events → future outcomes. Valuable insights for auditors + stakeholders

Benefits :- Make comparison with large vol of data. AI + Automation is used in b.d.

→ large vol. of data are processed to provide great insights

Decisions are based on previous cases regarding non compliance + fraud!

latest policy changes - upto date.

financial auditors → can adjust their reporting process + spot fraudulent transaction

Perform accurate audits

Before Analytics → use eff. data aggreg^n + mangmt sys.

high quality data + Authentic info^n.

Human error is overcame by automating.

Analytics :- strong Cybersecurity policy to protect data ]

→ change Configurations (outsiders ×)

→ N/W Security (Anti malware)

→ user access

→ Priviledged access mgt.

Appts

① set a goal

② Recruit Competent team ↑

3

Features

① Data wrangling + prep$^n$

before using iterative model.

During model const$^n$

② Data exploration

visualize data to find insights by using interactive dashboards.

③ Scalability

uses less h/w gear & uses less energy → use fast processor + memory

④ Support for various types of Analytics

firms make better business decision

⑤ Version Control :

Keep track & control changes to s/w code

SD Team keep track of changes

⑥ Data Management

Obtain

Store

use data

in cost effective

effective + secure way

optimizes use of data

⑦ Data Integr$^n$ :- combine data → make it Simple

enhance data quality, free resources, lower IT cost

⑧ Data Governance :- Accurate, available usable

Encryption centralised key

(9) Security :- Digital data unauthorized users & intrusions with firewalls, strong user auth^n, IDS

(10) visualiz^n :- data driven env.

Appl^ns:

(1) Monitoring customer spending & shopping behaviour

(2) Rec^n

(3) Smart Traffic sym :-

GPS in Ola, Uber

Siri on Apple, Cortana on windows

google Asst on Android

(4) IOT : Sensors - healthcare

(5) energy sector : in every 15 mins power consump^n to server

Industrial facility → at nyt household can run heavy machinery.

Compliance :- Collect^n - store - processing

↳ requires more resources.

Prev^n of fraudulent activities

- Managing 3^rd parties risk

- helps in managing cost.

- Disclaimer