Date.....

BIG DATA

UNIT - 1

B → Business
I → Intelligence
G → Gathering of data

data having intelligence for growth of business.

Ex Email - Beyond 25MB data will be Big data which is difficult to store or process with existing system.

A 2d 5 year is also generating data

Sources
- Sensor
- CC cam
- social networking sites
- stalk Exchange
- Insurance
- Manual

If data is generating rapidly, we need to take care of it.

What is Big Data?

↳ Data which is large in size.

↳ Data represent quantitative or qualitative values that are variable and usually come in tables, with rows and columns, in tree like format such as a set of nodes with parent children relationships or a graph with a set of interconnected nodes.

↳ It is a collection of large datasets that can't be processed using traditional and management tools efficiently.

Ex Social Media
NYSE
Google Maps
Medical Industry
Transportation

→ "Big data" is high volume, velocity and variety information assets that demand cost effective, innovative forms of information processing for enhanced insight and decision making.

Sources of Big Data

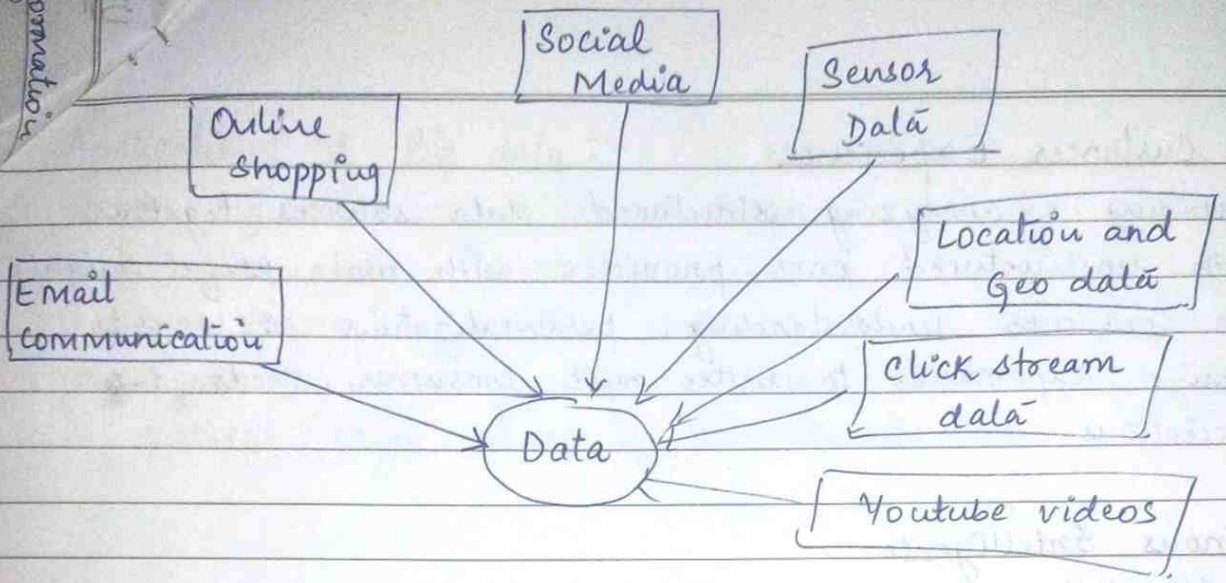
1. Social Networking Sites :-
   Facebook, Google .... These sites generates huge amount of data on day to day basis.

2. E-commerce site :-
   Amazon, FlipKart generates large no. of logs from which users buying trends can be traced.

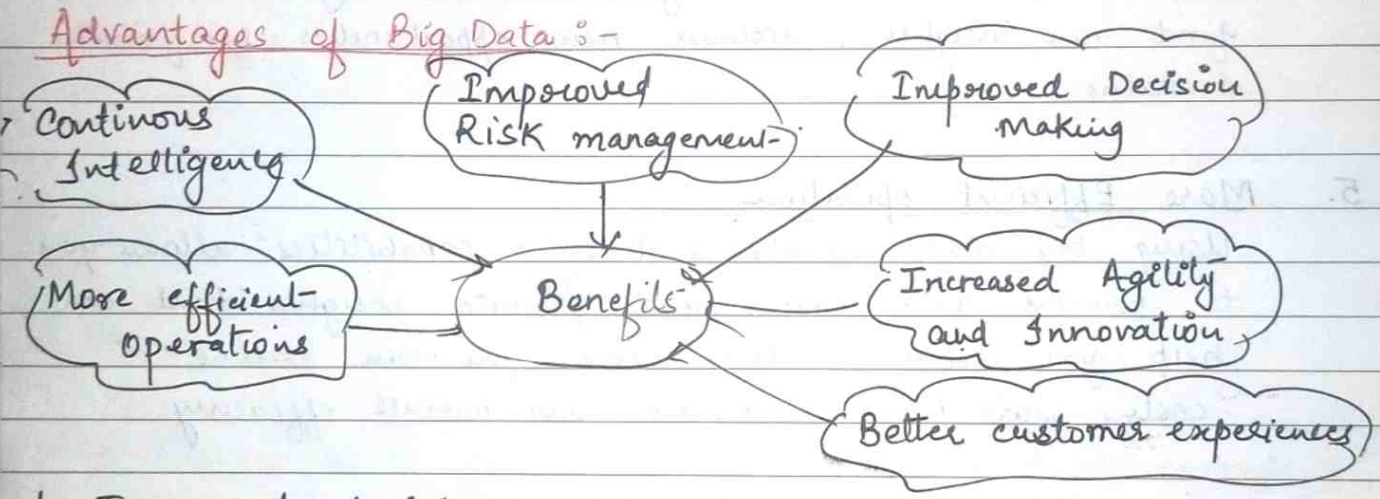
3. Weather Station :- Gives huge data that is stored and manipulated for forecast weather.

4. Telecom company :- Airtel, Vodafone study the user trends & publish their plan.

5. Share Market :- Stock Exchange across the world generates huge amount of data through its daily transaction.



Advantages of Big Data :-



1. **Improved decision making :-**
   ↳ When we manage and analyze big data, we discover patterns and unlock insights that improve & drive better operational and strategic decisions.

2. **Increased Agility & Innovation :-**
   Big data allows to collect & process real-time data points & analyze them to adapt quickly & gain a competitive advantage.
   These insights can guide & accelerate the planning, production & launch of new product, features & updates

3. Better Customer Experiences

↳ Combining & analyzing structured data sources together with unstructured ones provides with more useful insights for consumer understanding, personalization & ways to optimize experiences to better meet consumer needs & expectations.

4. Continous Intelligence

↳ Allows to integrate automated, real-time data streaming with advanced data analytics to continously collect data, find new insights, discover new opportunities for growth & value.

5. More Efficient operations

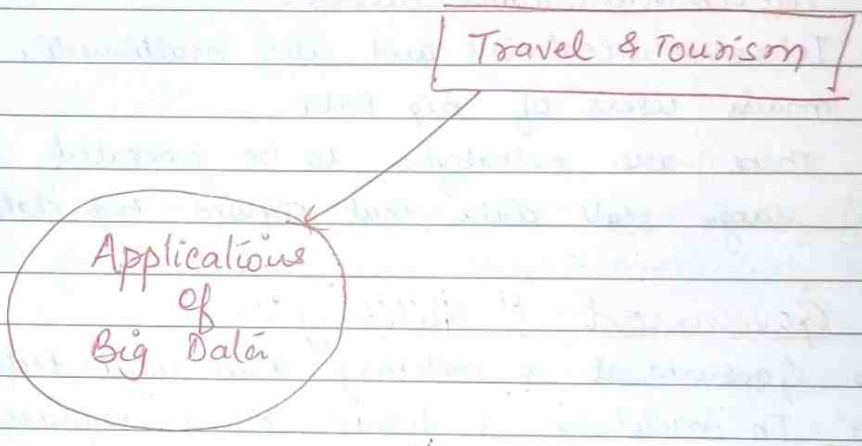
Using big data analytics tools & capabilities allows you to process data faster and generate insights that can help you determine areas where you can reduce costs, save time & increase your overall efficiency.

6. Improved Risk Management

Analyzing vast amount of data helps companies evaluate risk better - making it easier to identify & monitor all potential threats & report insights that lead to more robust control & mitigation strategies

Applications of Big data :-

Term Big data refers to as large amount of complex and unprocessed data. Nowadays company's use Big Data to make business more informative and allows to take business decisions by enabling data scientists and other professionals to analyze large volume of transactional data.



1. Travel & Tourism :-
   - ↳ Enables us to forecast travel facilities requirements at multiple locations, improve business through dynamic pricing & many more.

2. Financial & Banking sector :-
   - ↳ These people use big data technology extensively.
   - ↳ Help banks and customer behaviour on the basis of investment patterns, shopping trends, motivation to invest and inputs that are obtained from personal or financial backgrounds.

3. Healthcare :-

↳ Big data has started making a massive difference in healthcare sector, with the help of predictive analytics, medical professionals & health care personnel.

↳ It can also produce personalized healthcare & solo patients also.

4. Telecommunication Media :-

↳ Telecommunications and the multimedia sector are the main users of Big Data.

↳ There are zettabytes to be generated everyday and handling large scale data that require big data technologies.

5. Government & Military :-

↳ Government & military also used technology at high rates.

↳ In military, a fighter plane requires to process petabytes of data.

↳ Government agencies use Big Data & run many agencies managing utilities, dealing with traffic jams and the effect of crime like hacking and online fraud.

↳ Aadhar Card - Government has a record of 1.21 billion citizens. This vast data is analyzed & store to find things like no of youth in the country.

↳ Big data cannot store in a traditional database, so it stores & analyze data using Big data analytics.

6 E commerce :-

↳ Application of big data.

↳ Maintains relationships with customers that is essential for ecommerce industry.

↳ E-commerce websites have many marketing ideas to retail merchandise customers, manage transactions, & implement better strategies of innovative ideas to improve businesses with Big data.

7 Social Media :-

↳ Largest data generates.

↳ statistics have shown that around 500+ terabytes of fresh data generated from social media daily, particularly on Facebook.

↳ Single activity on social media site generates many stored data & gets processed when required.

↳ Data stored is in terabytes it takes a lot of time for processing. Big data is a solution to the problem.

8. Academia :-

↳ Numerous online educational courses to learn from.

↳ Academic institutions are investing in digital courses powered by Big Data technologies to aid all-round development of budding learners.

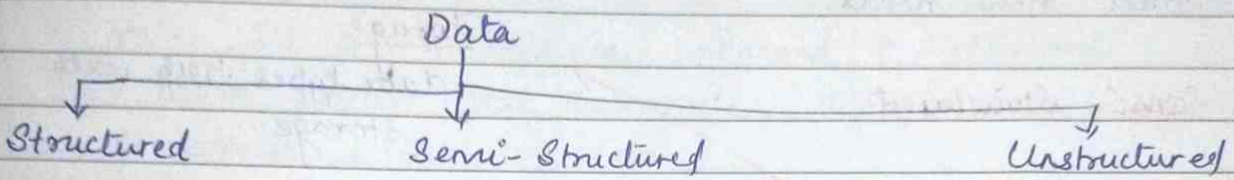
Big Data Analysis Tools and Software :-

↳ Tools that are used to store and analyze a large no. of data sets and processing these complex data are known as big data tools.

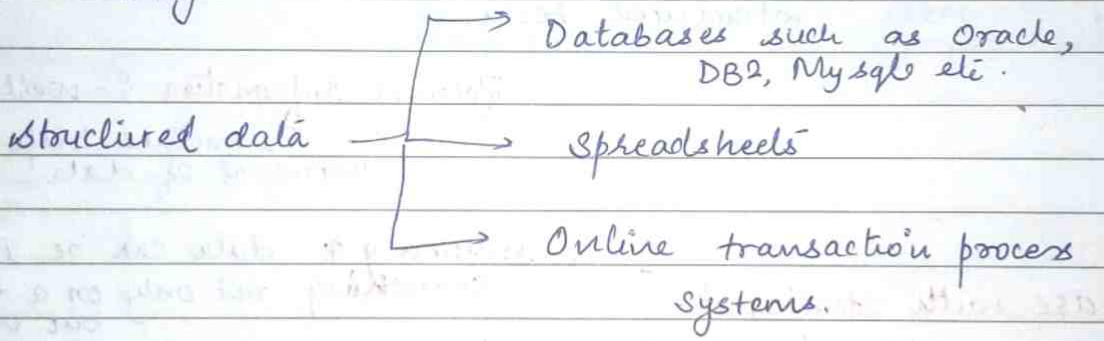
1. Xplenty
   ↳ Data analytic tool for building a data pipeline by using minimal codes in it.
   ↳ Offers wide range of solutions for sales, marketing, and support.
   ↳ With the help of its interactive graphical interface, it provides solutions for ETL, ELT.
   ↳ Best part is its low investment in hardware & software & it offers support via email, chat, telephonic ...
   ↳ Platform to process data for analytics over cloud and seggregates all data together.

   Features
   - Rest API : User can possibly do anything by implementing API.
   - Flexibility : data can be sent, and pulled to databases, warehouses & salesforce.
   - Data security : offers SSL/TSL encryption.
   - Deployment : offers integration apps for both cloud & in-house

Types of Big Data :-



1. **Structured Data :-**
   - ↳ data that resides in a fixed field within a record.
   - ↳ type of data most familiar to our everyday lives.
   - ↳ Structured data is also called relational data.
   - ↳ data split into multiple tables to enhance integrity of data by creating a single record to depict an entity.



- ↳ Structured data is easy to enter, query & analyze.
- ↳ All of the data follows same format.
- ↳ Business data of an ecommerce website can be considered to be structured data.

Cons of structured data :-

- ↳ Structured data can only be leveraged in cases of predefined functionalities.
- ↳ data is stored in data warehouse with rigid constraints & a definite schema. Any change in requirem

would mean updating all of that structured data to meet new needs.

Semi-structured

Ease with structured data

- **Storage** — data types help with storage
- **Scalability** — not an issue with increase in data
- **Security** — data is secured
- **Update and delete** — Updation & deletion is easy due to structured form.

Ease with structured data retrieval

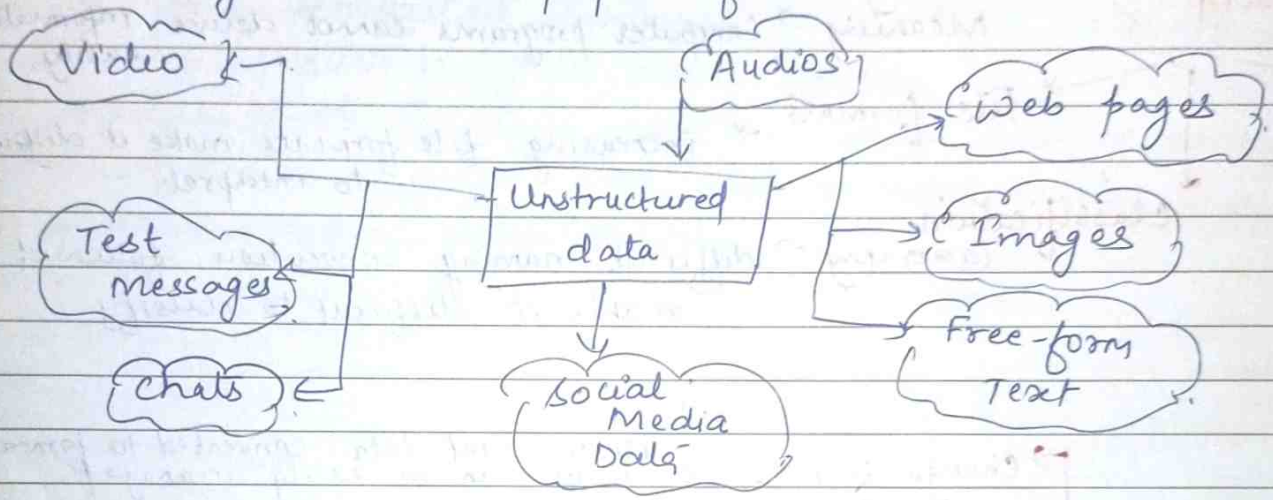
- **Retrieve Information** :- well defined structure helps in easy retrieval of data
- **Indexing & searching** — data can be indexed not only on a text string but other attributes as well
- **Mining data** :- data can be easily mined & knowledge can be extracted
- **BI operations** — BI works well with structured data

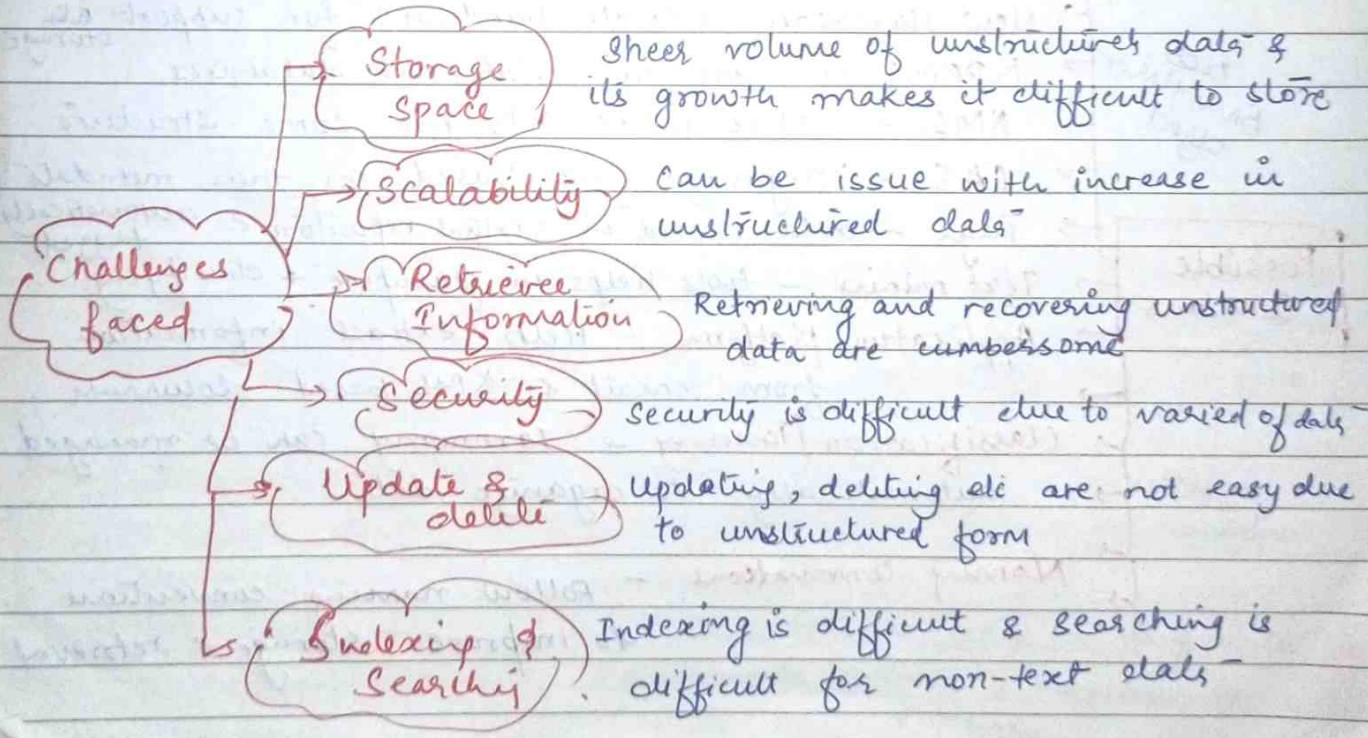
Unstructured data :-

↳ data doesn't adhere to any definite schema or set of rules. Its arrangement is unplanned & haphazard.

↳ Photos, videos, text documents, log files can be generally considered unstructured data.

↳ Also known as "dark data" because it cannot be analyzed without proper software tools.





Other challenges

- Interpretation - data is not easily interpreted
- Tags - data grows, it is difficult to tag.
- Indexing - Algorithm construction is difficult for indexing.
- Deriving Meaning → Computer programs cannot derive information easily.
- File formats → increasing file formats make it difficult to interpret
- Taxonomy → different naming convention followed makes it difficult to classify.

Possible solutions

- Change formats → Unstructured data converted to formats which can be easily managed
- New Hardware → Create hardware for support- ate storage.
- Blob Binary large object
  - RDBMS → store in relational databases.
  - XML - store in XML to give some structure
  - CAS → Organize files based on their metadata
  - Tags - data stored in virtual repositories & automatically tagged
- Text mining - tools helps in grouping & classifying.
- Application platforms - Help extract information from email & KML based documents.
- Classification / Taxonomy → Taxonomy can be managed automatically to organize data.
- Naming conventions - Follow naming conventions to improve storage & retrieval.

3. Semi Structured -

↳ data is ~~to~~ not bound by any rigid schema for data storage & handling.

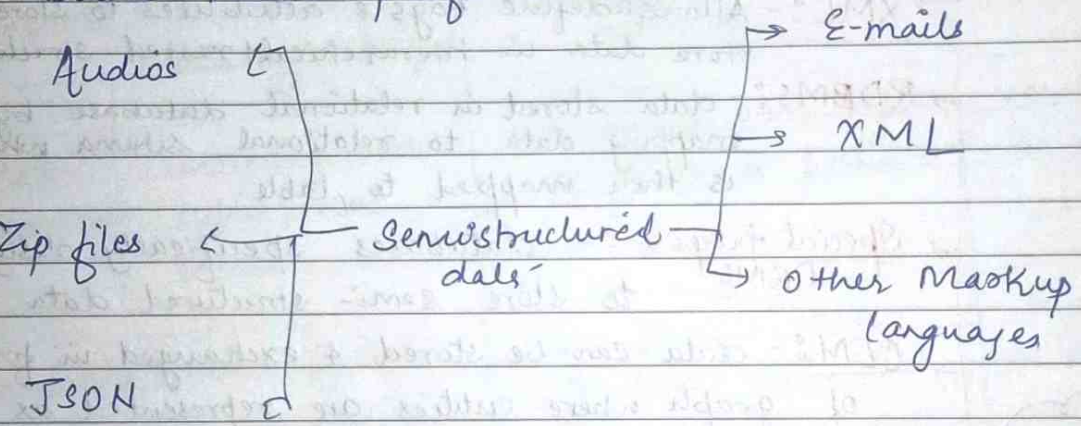
↳ data is not in relational format & is not neatly organized into rows & columns

↳ Doesn't need SQL rather called as NoSQL data.

↳ data serialization language is used to exchange semi-structured data across systems.

↳ often used to store metadata about business process.

↳ Information typically comes from external sources such as social media platforms



Challenges with Semi Structured Data

- Storage Cost : Storing data with schemas ↑ cost.
- RDBMS : Semistructured data cannot be stored in tables.
- Irregular & Partial Structure - Some elements have extra info. while others not
- Implicit Structure - In many cases, structure is implicit. Interpreting relationships is difficult.
- Evolving schemas - Schemas keep changing with requirements making it difficult to capture in database.

Challenges

- Distinction between schema & data - Vague distinction between schema & data exists at times making it difficult to capture data.
- Flat files - data stored in flat files difficult to search
- Incomplete / Irregular Structures - Extracting structure when there is none & interpreting relations existing in structure is difficult task

Possible Solutions

- XML :- Allows to define tags & attributes to store data ~~store data in hierarchical/nested structure~~.
- RDBMS :- data stored in relational database by mapping data to relational schema which is then mapped to table.
- Special purpose DBMS :- databases specifically designed to store semi-structured data
- OEM :- data can be stored & exchanged in form of graph where entities are represented as objects which are vertices in graph.
- Indexing :- Indexing data in graph enables quick search

Page.....

Difference between structured, semi-structured & unstructured

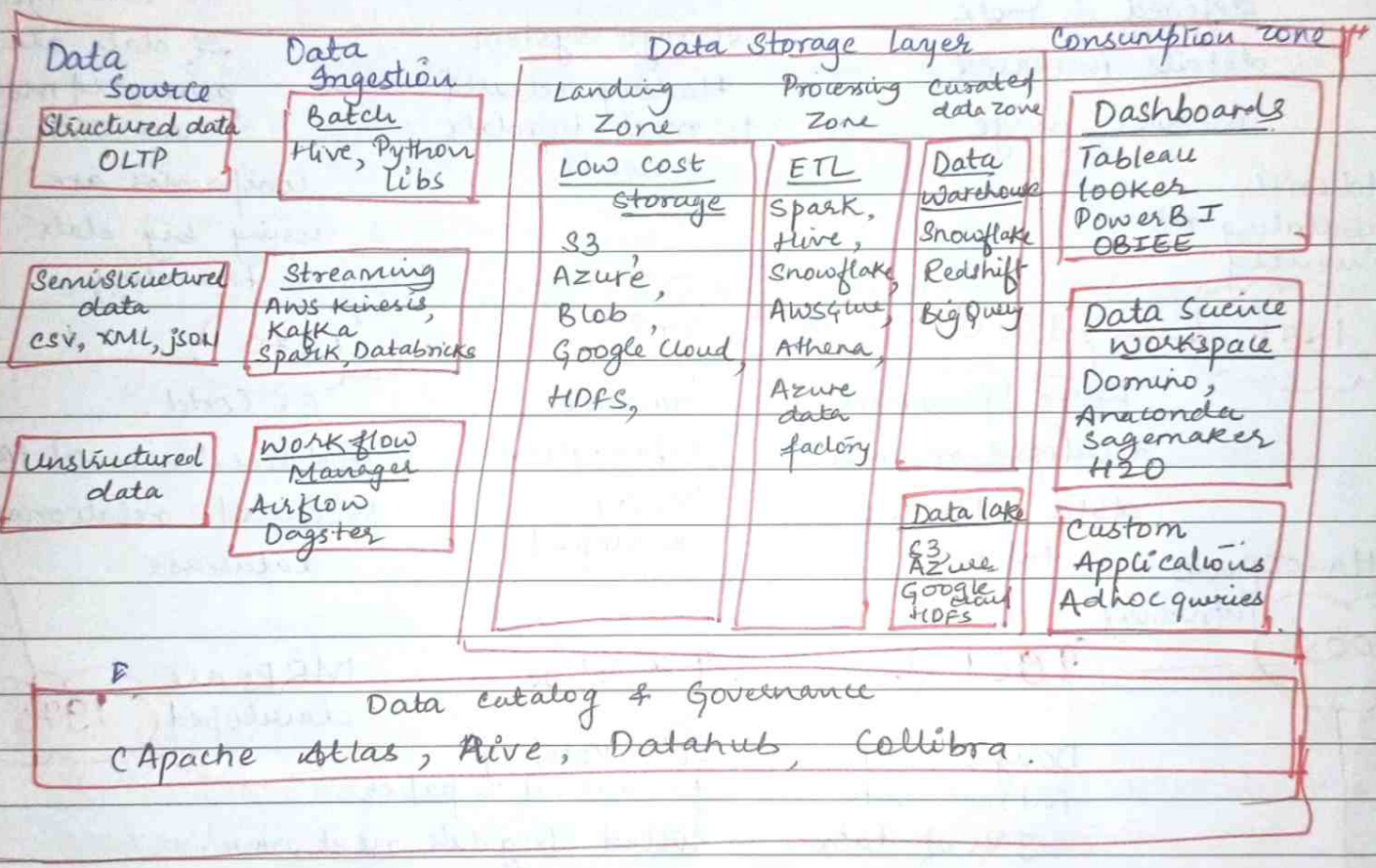
Properties	Structured Data	Semi Structured Data	Unstructured Data
Technology	Based on relational database table	Based on XML	Based on characters and binary data
Transaction management	Matured transaction and various concurrency technique	Transaction is adapted from DBMS not matured	No transaction management and no concurrency
Version management	versioning over tuples, row, tables	versioning over tuples or graph is possible	versioned as a whole.
Flexibility	Schema dependent and less flexible	More flexible than structured data but less flexible than unstructured	More flexible and there is absence of schema.
Scalability	Difficult to scale DB schema	Scaling is simpler than structured data	More scalable
Robust	very robust	New technology, not spread	
Query Performance	Structured query allow complex joining	Queries over anonymous nodes are possible	only textual queries are possible

Big Data Platform :-

↳ Integrated computing solution that combines numerous software systems, tools & hardware for big data management.

↳ One stop architecture that solves all data needs of a business regardless of volume & size of data at hand.

↳ Enterprises are increasingly adopting big data platforms to gather tons of data & convert it into structured.



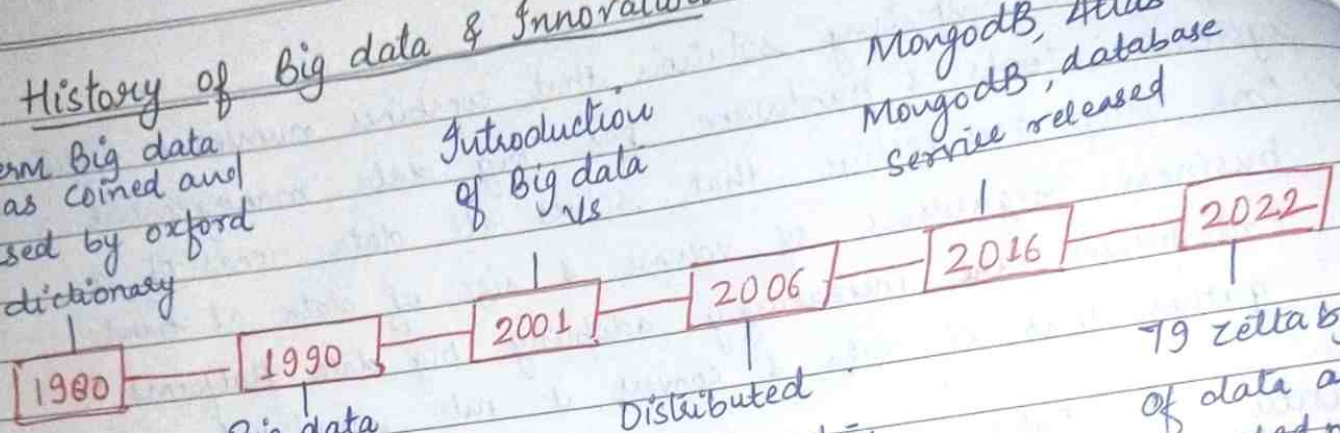
Platforms

↳ Hadoop Delta Lake Migration platform

↳ open source platform by Apache Foundation

↳ used to manage & store large data set at a low cost & great efficiency

History of Big data & Innovation



1980 — Term Big data was coined and used by oxford dictionary

1990 — Big data defined in more details increased internet usage

2001 — Introduction of Big data Vs

2006 — Distributed storage system Hadoop released to meet big data needs

2016 — MongodB, Atlas MongodB, database service released

2022 — 79 zettabytes of data are generated more than 53% of companies are using big data technology.

1881 — Hollerith Tabulating m/c invented

1928 — Fritz Pfleumer developed magnetic data storage on tape

1948 — Shanon's Information Theory developed

1970 — EF Codd, mathematician at IBM presented relational database.

1976 — MRPs are developed

1998 — John Mashey presented a paper titled "Big data.. next wave".

2001 — Doug Laney introduced 3 Vs of data

2005 — Hadoop was introduced

2014 — Companies started moving ERP to cloud

② Data Catalogy Platform :-

↳ Single self service environment to users to help them with find, understand & trust data source.

↳ Helps to discover new data sources

↳ use catalogy discovery to find which data fits their needs

↳ Provides comprehensive, security, data governance & solution to protect data

③ Data Ingestion Platform :-

↳ First step for data coming from various sources

↳ Data is priorilized & categorized

↳ Importing & loading data into system

④ IoT Analytics Platform :-

↳ Provides wide range of tools to work upon big data.

⑤ Big Data Integration & Management Platform :-

↳ Elixir Data provides highly customizable solution for Enterprises.

↳ Provides flexibility, Security & stability for an Enterprise application to deploy public cloud.

⑥ ETL Data Transformation Platform :-

↳ Can be used to build pipelines & even schedule the running of the same for data transformation.

Essential Components of Big Data Platform :-

There are so many essential components which are given as:-

① Data Ingestion, Management, ETL, & Warehouse

↳ Provides resources for effective data management and data warehousing.

2. Stream Computing :-
↳ Helps compute the streaming data used for real time analytics.

3. Analytics / Machine learning :-
↳ features for advanced analytics & machine learning.

4. Integration :- integrate big data from any source with ease.

5. Data Governance - provides comprehensive security, data governance, and solutions to protect data.

6. Provides Accurate data - Helps bunisess to make right decision by providing accurate info.

7. Scalability - Helps scale application to analyze all climbing data; offers scalable storage capacity.

8. Price optimization :- provides insight for B2C & B2B enterprises which helps bussiness to optimize the prices.

9. Reduced Latency :- with warehouse, analytics tools & efficient data transformation, reduce data latency & increase throughput.

Common platform :-

1. Apache Hadoop :-
   - ↳ Open source programming architecture & server software
   - ↳ store and analyze large dataset very fast.
   - ↳ uses commodity hardware in a clustered computing environment.

2. Cloudera :-
   - ↳ big data platform based on Apache's Hadoop.
   - ↳ Handle huge volumes of data
   - ↳ Enterprises store over 50 petabytes in this platform data warehouse.
   - ↳ its dataflow enables real-time processing.

3. Amazon web services :-
   - ↳ popular as AWS
   - ↳ Another Hadoop based big data ~~dat~~ platform
   - ↳ Hosted as cloud environment.
   - ↳ Using Amazon EMR, enterprises easily set up & scale, other big data platforms like Spark, Aeslio.

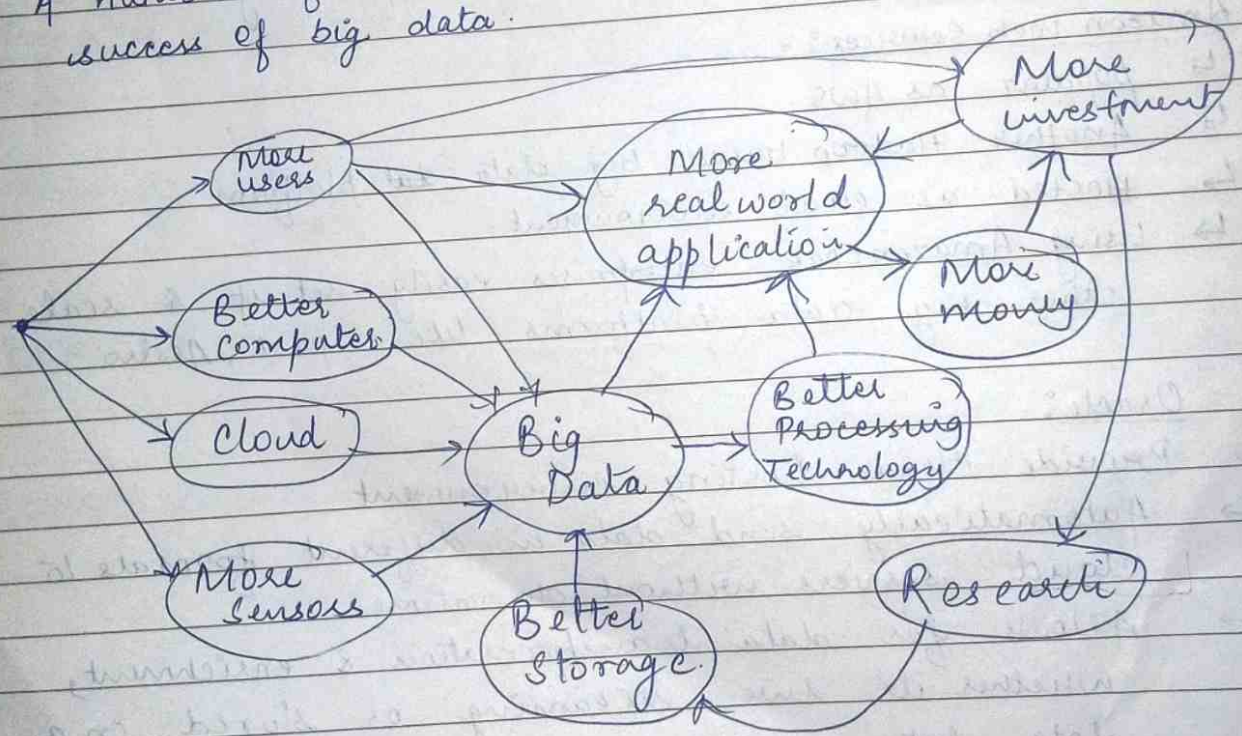
4. Oracle :-
   - ↳ Provide cloud hosting environment
   - ↳ Automatically send data in different formats to cloud servers without downtime.
   - ↳ Allows for data transformation & enrichment, whether its live streaming or stored in a data lake.

..., performance and options of public on a

5. Snowflake :-
↳ data warehouse for storing, processing & analyzing data
↳ top cloud hosting frameworks & integrated with new SQL query engine.

6. Apache Storm :-
↳ Brainchild of Apache Software Foundation.
↳ used in real-time analytics & distributed processing.
↳ supports all programming languages.

Drivers of Big Data :-
A number of business drivers are the core of success of big data.



1) Digitization of society :-

↳ Most people spend 4-6 hours per day consuming and generating data through variety of devices & social applications.

2) Plummeting (Reducing) Technology Costs :-

↳ Costs of data storage & processors keed reducing, makes possible for small businesses & individuals to become involved with Big data.

3) Connectivity through Cloud computing :-

↳ organizations want to process massive quantities of data do not have to invest in large quantities of IT infrastructure. Instead they can license the storage & processing capacity they need.

4) Increased Knowledge about data Science :-

↳ Knowledge & education about data science has greatly professionalized & more information becomes available every day.

↳ data analysis & statistics is becoming a popular subject among students.

5) Social Media Applications :-

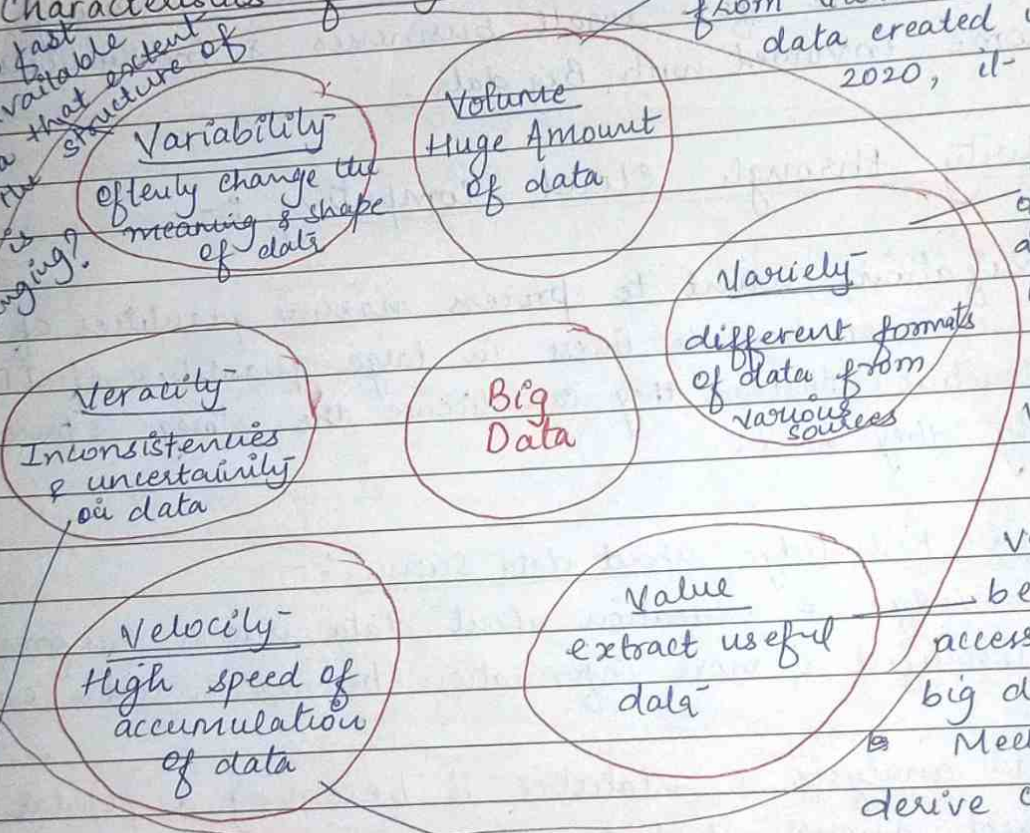
↳ Social media data can be used to identify customer preferences for product development, target new customers, behaviours, preferences and opinions of public on a

scale that has never been known before.

6. Upcoming internet of Things (IoT) :-
↳ increasingly gaining popularity as consumer goods providers start including smart sensors in household appliances.

Characteristics of Big data :-

How fast data is generated on daily basis from various sources. In 2016, data created was 8ZB by 2020, it rises to 40ZB



Combination of data types that are being dumped into system. emails, PDFs, photos, audios, videos etc.

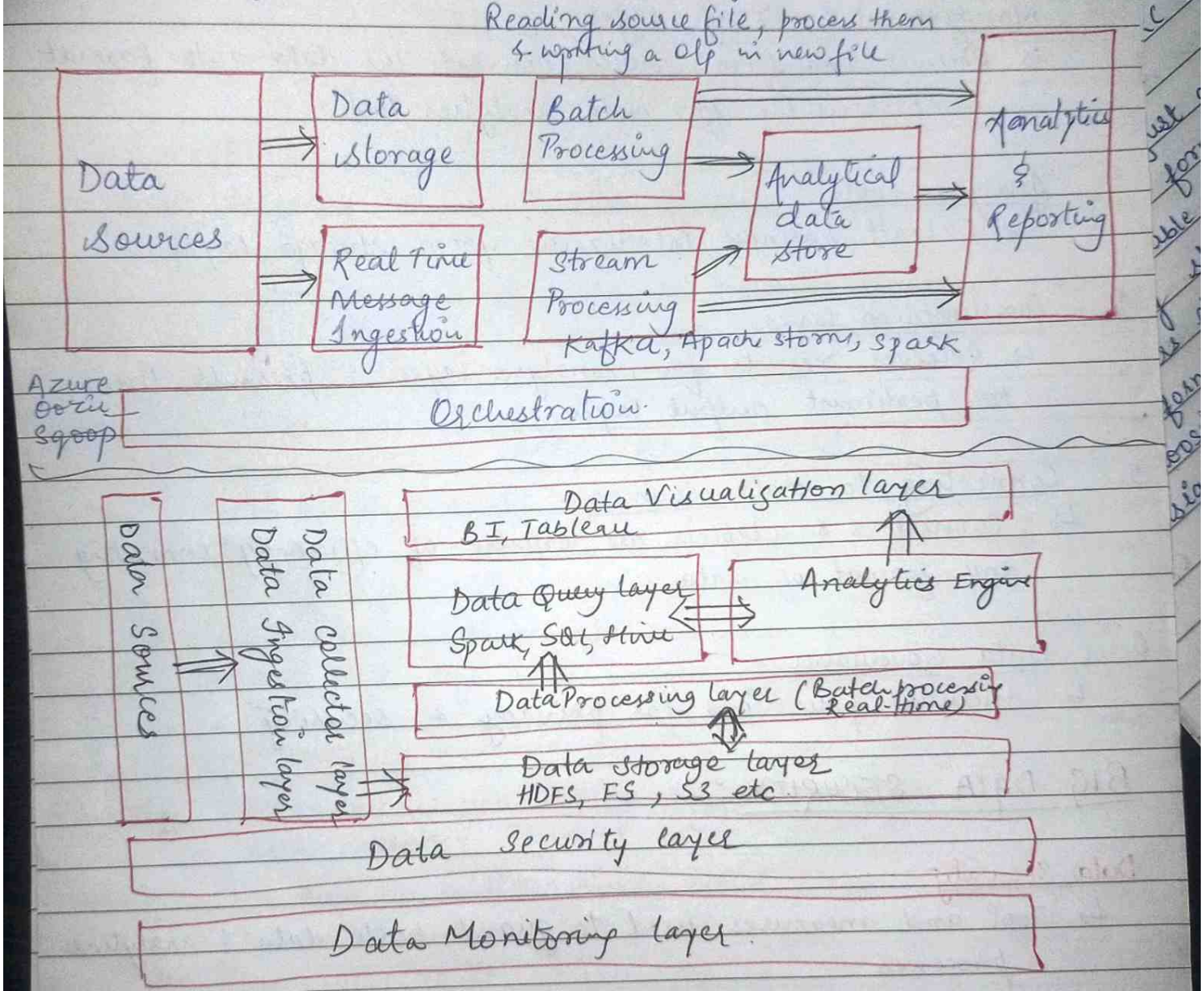
Value that can be derived from accessing & analysing big data.
Mechanism to derive correct meaning of data.

discrepancies found in data

Refers to speed of generation of data. How fast data is generated & processed to meet the demands.

Data Architecture :-

Big data architecture is designed to handle the ingestion, processing, and analysis of data that is too large or complex for traditional database system.



1. Big Data Sources layer - datawarehouses, RDBMS, IoT devices
   ↳ Big data environment can manage both batch processing & real-time processing of big data sources

2. Management & storage layer -
   ↳ Receives data from source, converts the data into format comprehensible for data analytics tool.

3. Analysis tool
   ↳ Extract business intelligence from storage layer.

4. Consumption layer
   ↳ Receives results from analysis layer & presents them to pertinent output layer.

5. Connecting to data sources
   ↳ Connectors & adapters are capable of efficiently connecting any format of data.

6. Data Governance
   ↳ Includes provisions for privacy & security

BIG DATA SECURITY :-

Data Security
↳ Tool and measures used to guard both data & analytics processes.

Main purpose of big data security is to provide protection against attacks, thefts, & other malicious activities.

Data Security Technologies :-

- Encryption
- User Access Control
- Physical Security
- Centralized Key Mangement.

1. **Encryption :-** - DES, RSA
   - Secure a massive volume of data, different types of data.
   - Can be user-generated or machine-generated code.
   - Encryption tool along with different analytics toolsets format or code the data.

2. **User Access Control :-** MAC, RBAC (oracle)
   - Most basic security tool.
   - Automated strong user access control is a must for organizations.
   - Automation control manages complex user control levels that protect big data platform against the inside attack.

3. **Physical Security :-**
   - Built in when you deploy big data platform in your own centers.
   - Can also be built around cloud provider's data center security.
   - Video Surveillance, security logs.

4. Centralized Key Management — on-demand key delivery, logging, OTP, PIN
↳ Applied in Big data environments, especially on those having wide geographical distribution.
5. Intrusion detection & prevention - IPS behind the firewall.

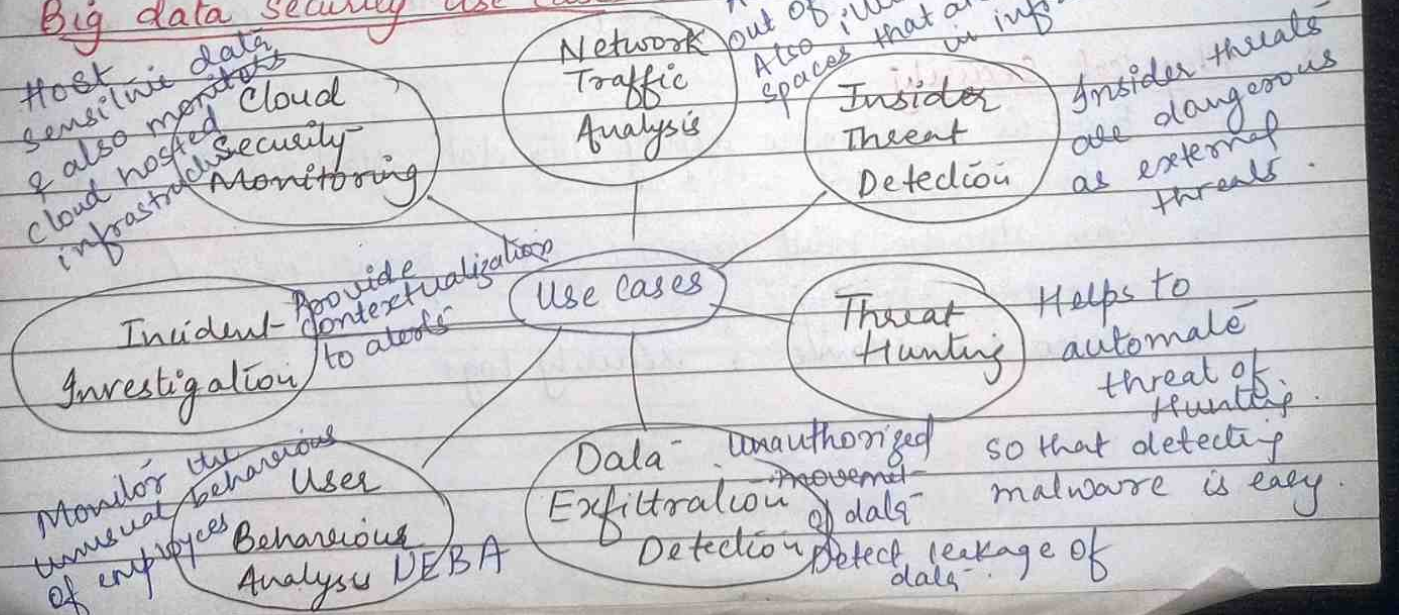
Some of the companies practicing Big data securities are; -

1. Cloudwick
↳ KDAP (Cloudwick Data Analytics Platform) is a managed security hub that integrates security features from multiple analytics toolsets & machine learning projects.

2. IBM
↳ IBM Security Guardium to monitor Big data and NoSQL environments.

3. Gemalto
↳ Gemalto SafeNet protects Big data platforms.

Big data Security Use Cases :-



Data Security Issues :-

- Access Controls
- Non relational data Stores
- Storage
- Endpoints
- Real-time security
- data mining solutions

1. Access Controls
   - ↳ Permission to exchange data should be permitted to authenticated users only.
   - ↳ Needs to be such that it would not get hacked by attackers, hackers or by any malicious activities.

2. Non-relational data stores
   - ↳ Like NoSQL usually lack security by themselves

3. Storage
   - ↳ we store data on multiple tiers
   - ↳ Storage depends on business needs in terms of performance & cost.
   - ↳ High-priority data is generally stored on flash media.

4. Endpoints
   - ↳ Security solutions that usually draw logs from endpoints will need to validate the authenticity of those endpoints

5. Real-time Security / Compliance tools :-
   ↳ generate a large amount of information.
   ↳ Key to find a way to ignore false or rough information.

6. Data Mining solutions :-
   ↳ Find a pattern that suggests business strategies.
   ↳ Ensure security from both internal & external threats.

Big Data Privacy & Ethics :-

↳ Private customer data & identity should remain private.
↳ Share private information should be treated confidentially
↳ Customers should have transparent view.
↳ Big data should not interfere with human will
↳ Big data should not institutionalize unfair biases.

~~Reporting vs Analysis~~ :-