BDC

Unit-1 Types of Digital Data:-

→ classification of Digital Data:-

- → Unstructured data (80%)
- → Semi-structured data (10%)
- → Structured data. (10%)

① Structured data:-

when data conforms to a pre-defined schema/ structure we say it is structured data.

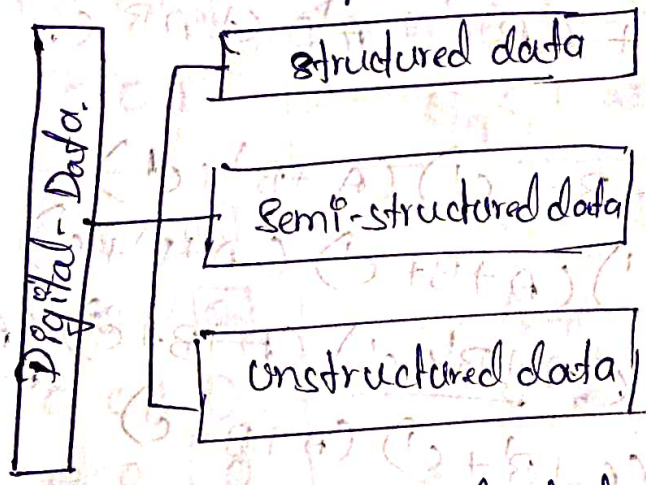
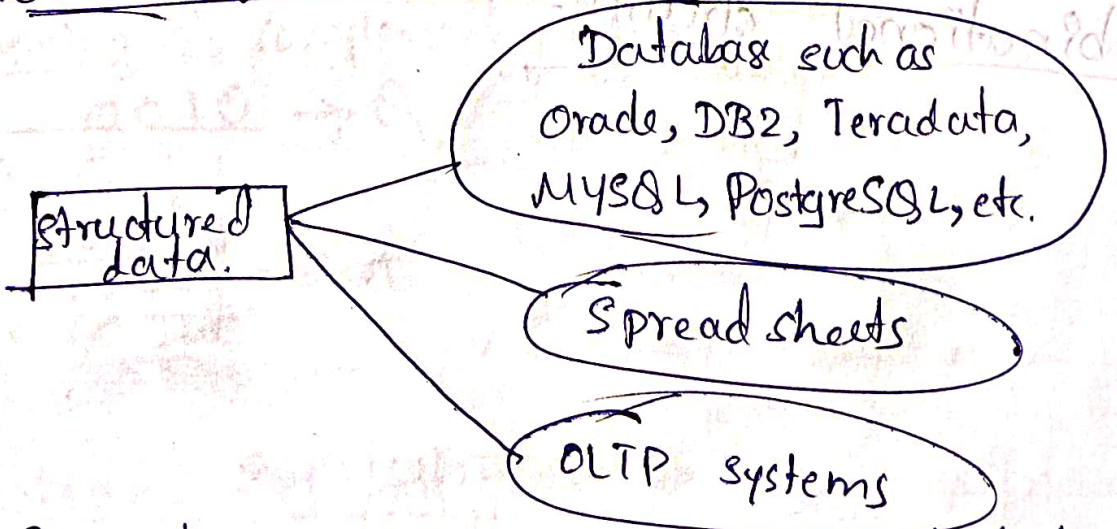


Fig:- Classification of digital data.

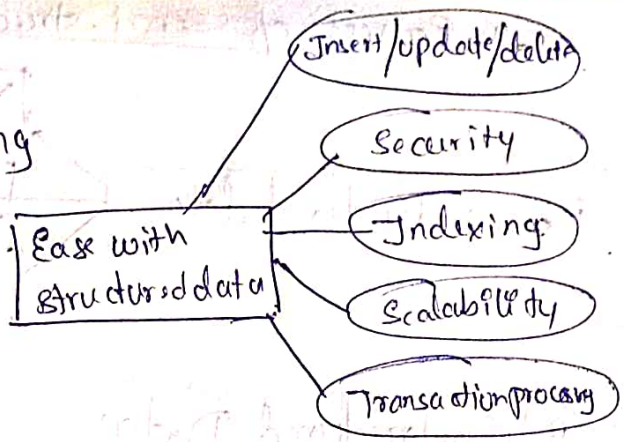
→ Sources of structured data:-



→ Ease of working with structured data:-

- * Insert / Delete / Update
- * Security

* Indexing
* Scalability
* Transaction processing
  * Atomicity
  * Consistency
  * Isolation
  * Durability



(2) Semi-structured data:

Self-describing structure. & It has following features

* It does not conform to the data models that are typically associated with relational databases or any other form of data tables.
* It uses tags to segregate semantic elements.
* Tags are also used to enforce hierarchies of records & fields within data.
* There is no separation b/w the data & the schema. The amount of structure used is dictated by the purpose at hand.
* In semi-structured data, entries belonging to the same class & also grouped together need not necessarily have the same set of attributes. And if at all, they have the same set of attributes, the order of attributes may not be similar & for all practical purposes it is not important as well.

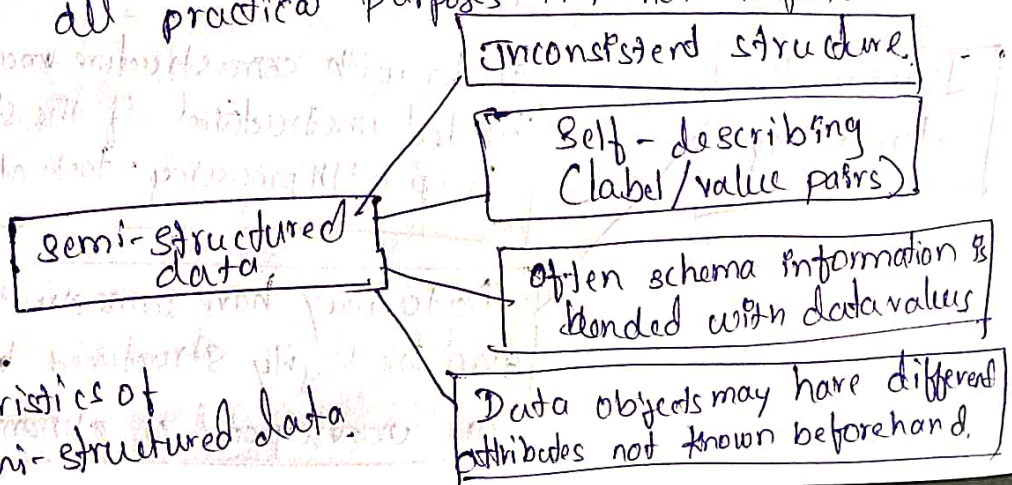
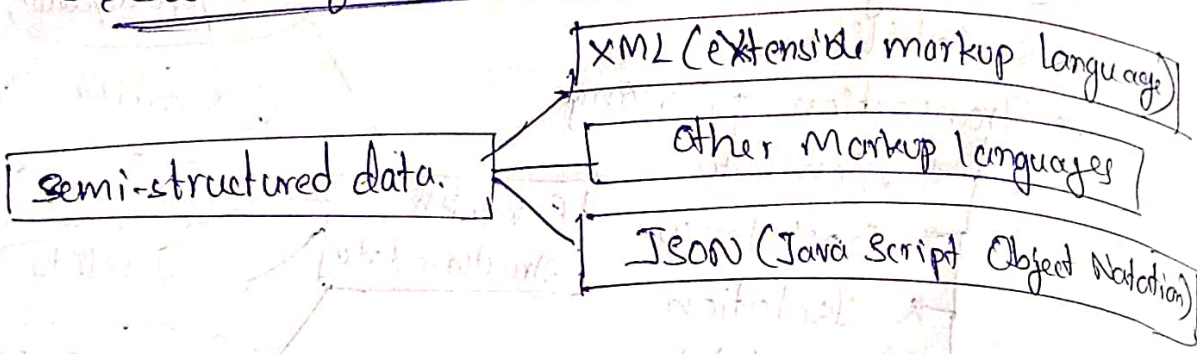


Fig:- Characteristics of semi-structured data.

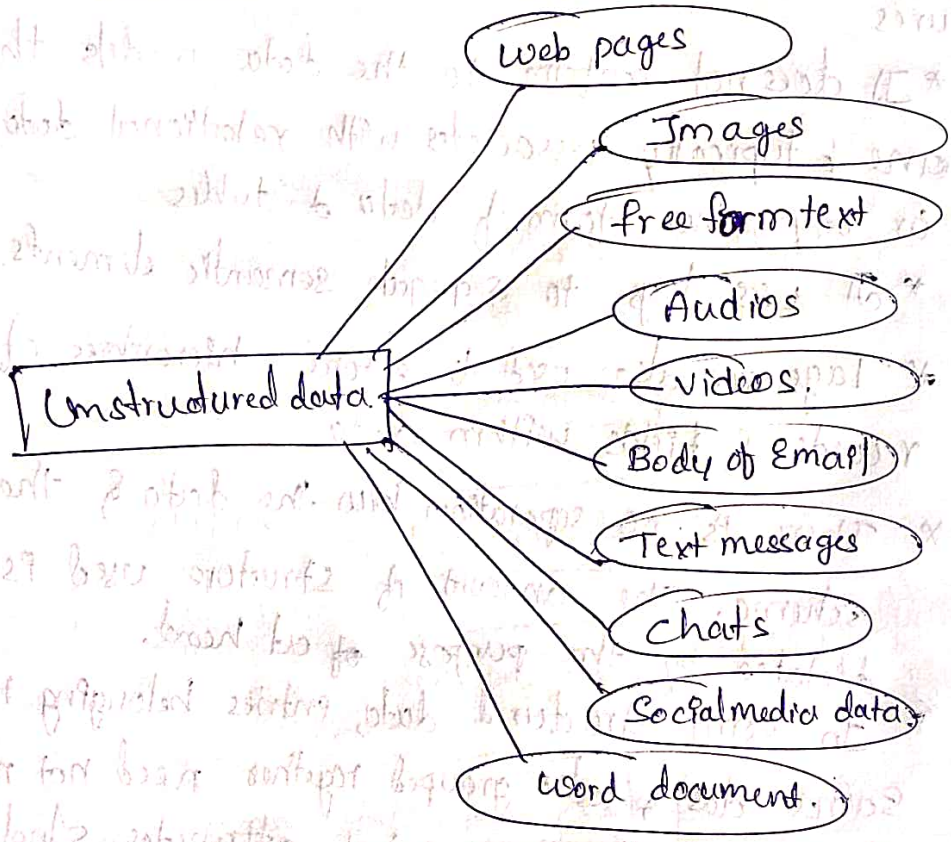
→ Sources of semi-structured Data:-



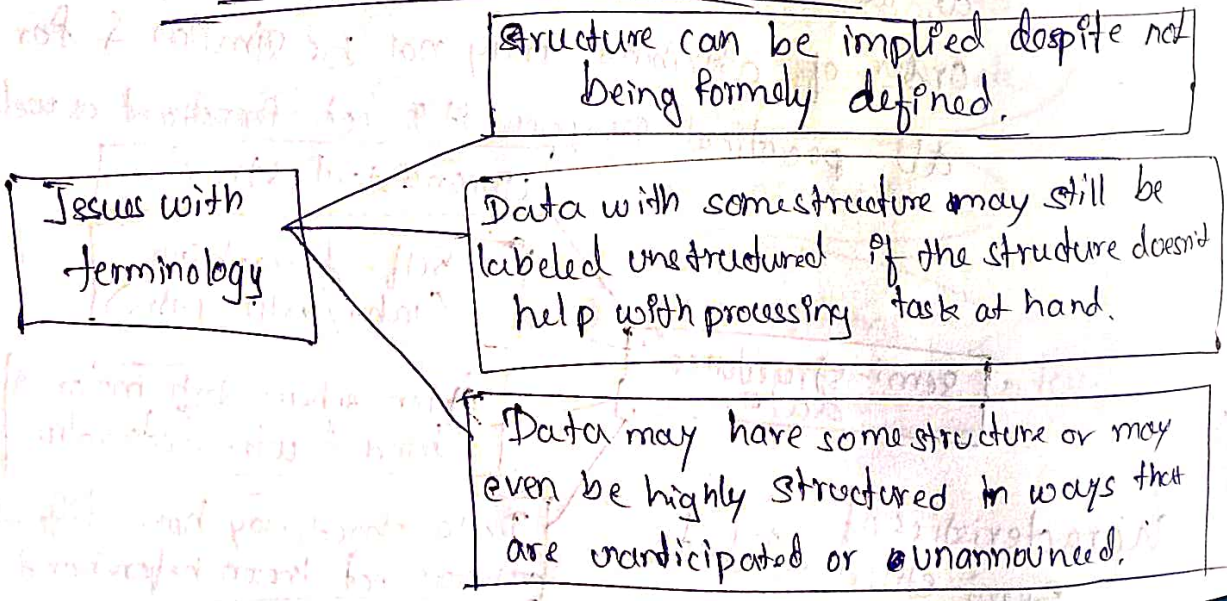
(2) Unstructured Data:-

does not conform to any predefined data model

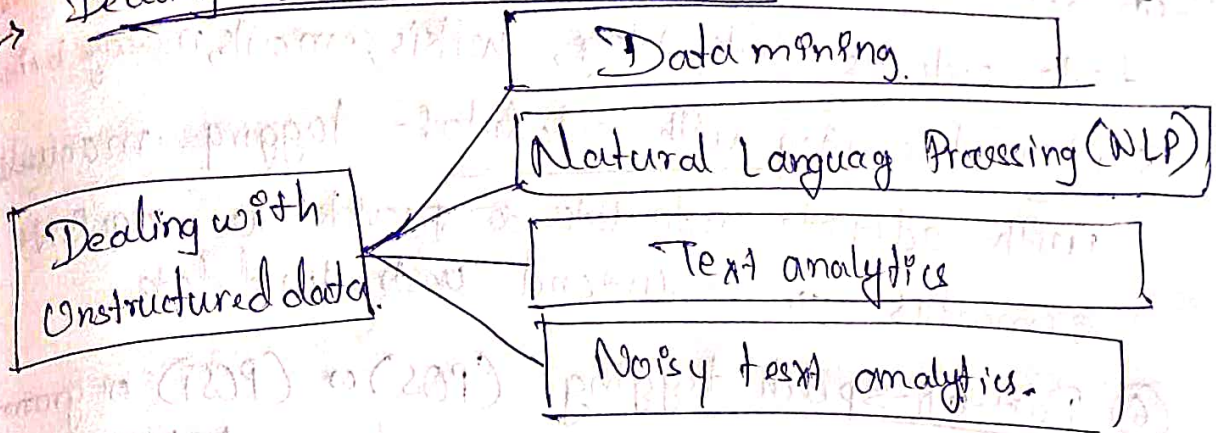
→ Sources of unstructured data:-



→ Issues with Unstructured data:-



→ Dealing with unstructured data:-



① Data mining :- deal with large data sets.

few popular data mining algorithms are as follows:-

* Association rule mining:- (market basked analysis).

Used to determine "what goes with what?".

It is about when you buy a product, what is the other product that your are likely to purchase with it. Ex:- eggs/cheese for breads

* Regression analysis:- helps to predict the relationship b/w 2 variables. The variable whose value needs to be predicted is called dependent variable & variables which are used to predict the value are referred as the independent variables.

* Collaborative filtering:- predicting a user's preference or preferences based on the preferences of a group of users.

② Text analytics or text mining:- Text mining is the process of gleaning high quality and meaningful information from text. (tasks → text categorization, text clustering, sentiment analysis etc.)

③ Natural Language Processing (NLP):- related to human computer interaction.

④ Noisy text analytics:- process of extracting structured

or semi-structured information from noisy unstructured data such as chats, blogs, wikis, emails, message boards etc.

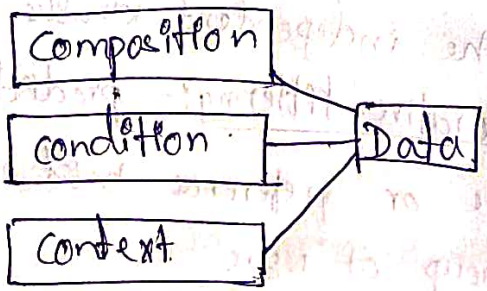
(5) Manual tagging with metadata:- tagging manually with adequate meta data to provide the requisite semantics to understand unstructured data.

(6) Part-of-speech tagging:- (POS) or (POST) or grammatical tagging. Process of reading text & tagging each word in the sentence as belonging to a particular part of speech such as "noun", "verb", "adjective", etc.

(7) Unstructured Information Management Architecture (UIMA):- Opensource platform from IBM. Used for real time content analytics. about Processing text & other unstructured data to find latent meaning & relevant relationship buried therein.

⇒ Introduction to Big Data:-

→ characteristics of Data:-



① Composition:- The composition of data deals with the structure of Data, that is, the source of data, the granularity, the types, & the nature of data as to whether it is static or real-time streaming.

② Condition:- deals with the state of data, that is "Can one use this data as is for analysis?" or "Does it require cleansing for further enhancement & enrichment?"

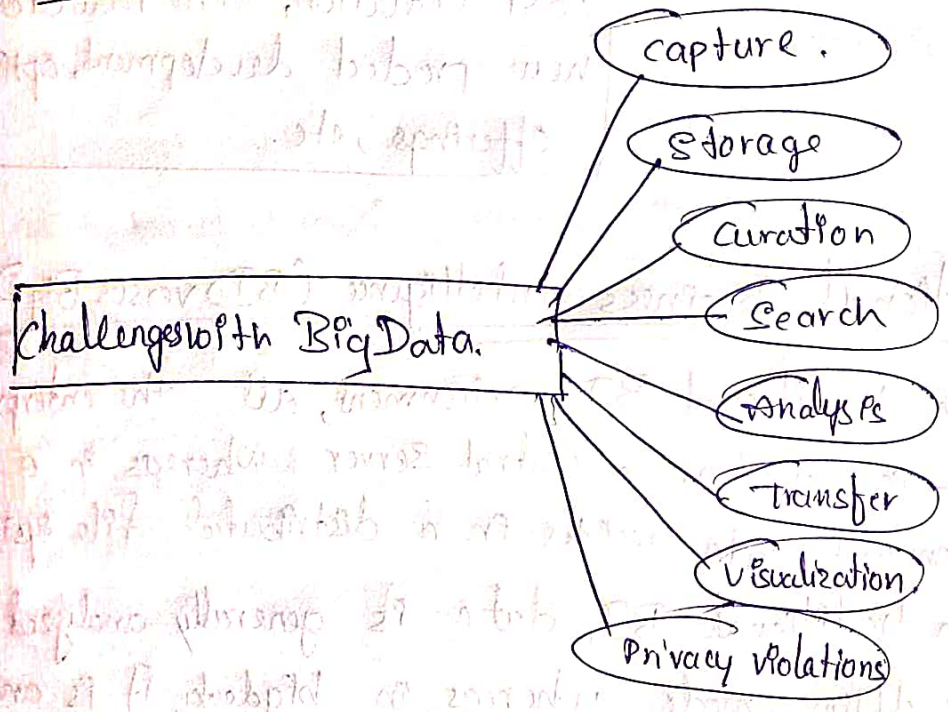
③ Context:- deals with "Where has this data been generated?"

"Why was this data generated?" "How sensitive is this data?" "What are the events associated with this data?" & so on.

→ Definition of Big data:-

Big data is high volume, high velocity, & high-variety information assets that demand cost effective, innovative form of information processing for enhanced insight & decision making.

→ Challenges with Big Data:-



C S² ATVP

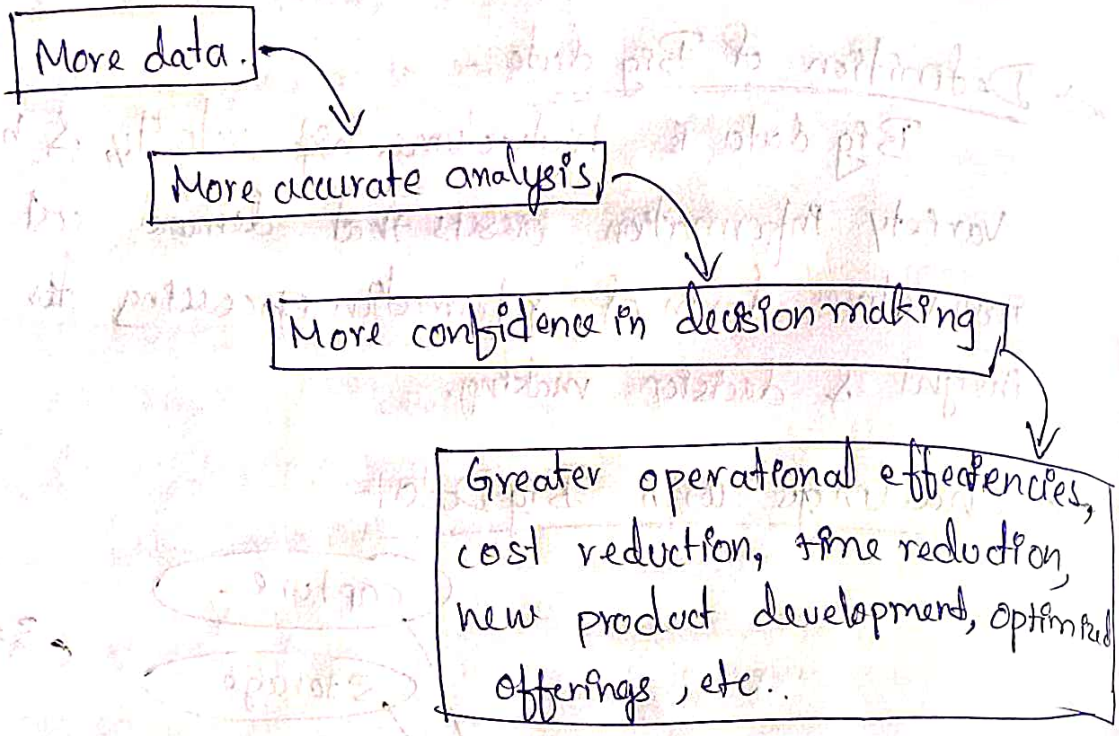
→ Characteristics of Data which are not Definitional Traits of Big Data.

① Veracity & validity:- Veracity refers to biases, noise, & abnormality in data. Validity refers to the accuracy and correctness of the data.

② Volatility:- volatility of data deals with, how long is the data valid?

③ Variability:- Data flows can be highly inconsistent with periodic peaks.

⇒ Need of Big data:-



→ Traditional Business Intelligence (BI) versus Big Data:-

* In traditional BI environment, all the enterprise's data is housed on a central server whereas in a BD environment data resides on a distributed file system.
* In traditional BI, data is generally analyzed in an offline mode whereas in bigdata, it is analyzed in both real time as well as in offline mode.
* Traditional BI is about structured data & it is here that data is taken to processing functions (move data to code) whereas big data is about variety: Structured, semi-structured & unstructured data & here the processing functions are taken to data (move code to data).

⇒ Big data Analytics:-

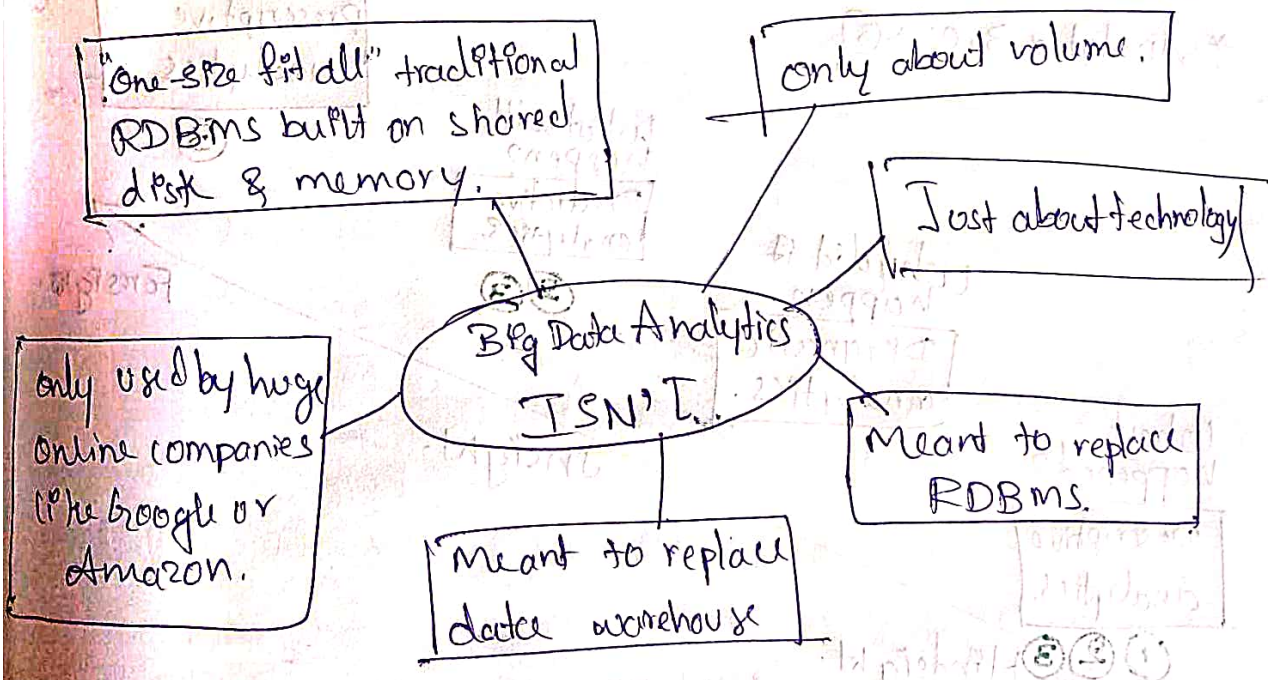
→ what is Big Data Analytics?

Big data Analytics is..

* Technology Enabled Analytics: Data Analytics & visualization tools

* About gaining a meaningful, deeper, & richer insight into your business to steer it in the right direction, understanding the customer's demographics, etc.
* About a competitive edge over your competitors by enabling you with findings that allow quicker & better decision-making.
* A tight handshake blw 3 communities: IT, business users, & data scientists.
* Working with datasets whose volume & variety exceed the current storage & processing capabilities & infrastructure of your enterprise.
* About moving code to data.

→ What big data analytics isn't?



→ Classification of Analytics:-

There are basically two schools of thought
* Those that classify analytics into basic, operationalized, advanced, & monetized.
* Those that classify analytics into analytics 1.0, analytics 2.0, & analytics 3.0.

→ First school of Thought:-

1*. Basic analytics:- This primarily is slicing and dicing of data to help with basic business insights. It's about reporting on historical data, basic visualization, etc.

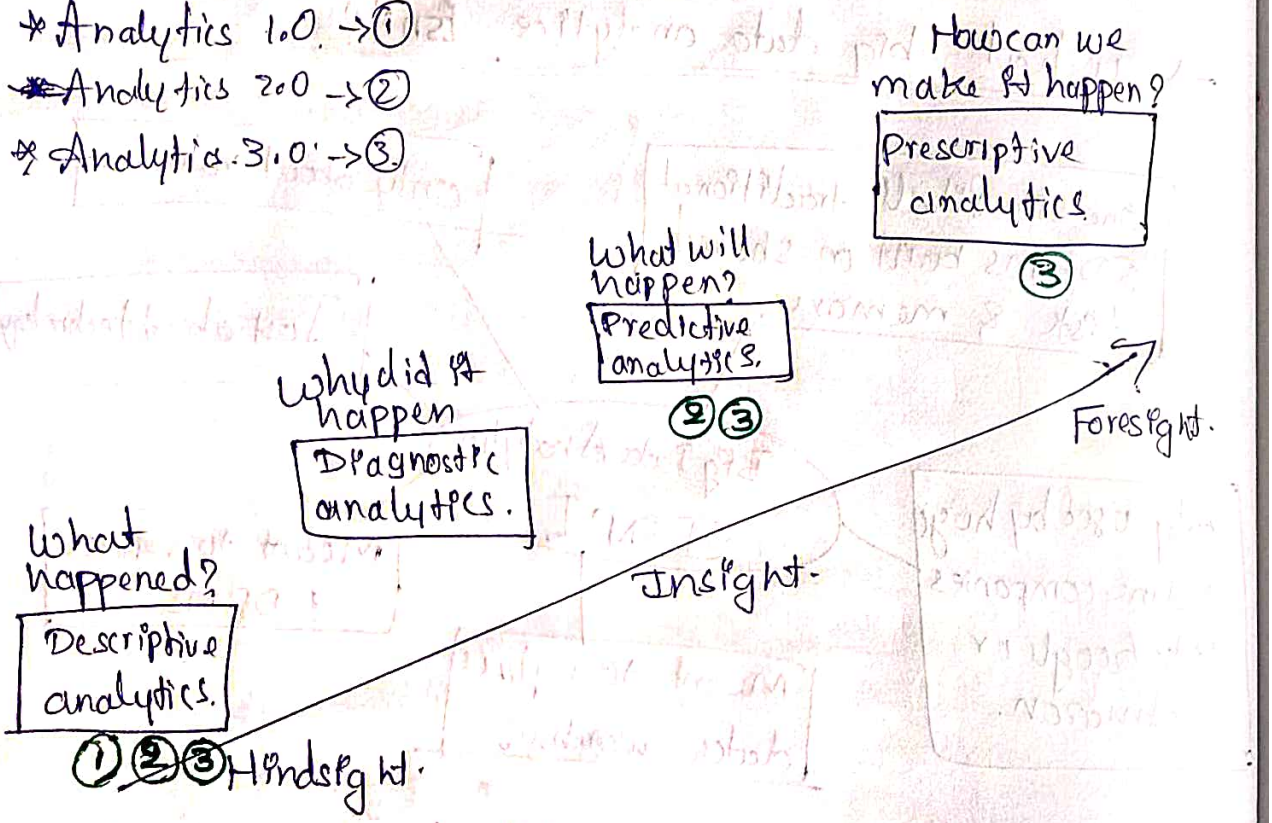
2*. Operationalized analytics:- If it gets woven into the enterprise's business process.

3*. Advanced Analytics:- This largely about forecasting for the future by way of predictive & prescriptive modeling.

4*. Monetized analytics:- use to derive direct business revenue.

→ Second school of Thought:-

* Analytics 1.0 →①
* Analytics 2.0 →②
* Analytics 3.0 →③



→ Greatest challenges that prevent business from capitalizing on big Data:-

* Obtaining executive sponsorship for investments in big data & its related activities.
* Getting the business units to share information across organizational silos.

* Finding right skills (business Analysts & data scientists) that can manage large amounts of structured, semi-structured, & unstructured data & create insights from it.
* Determining the approach to scale rapidly & elastically. In other words, the need to address the storage & processing of large volume, velocity, & variety of big data.
* Deciding whether to use structured or unstructured, internal or external data to make business decisions.
* Choosing the optimal way to report findings & analysis of big data. for the presentations to make the most sense.
* Determing what to do with the insights created from big data.

→ Top Challenges Facing Big Data:-

* Scaling :- Storage is one major concern that needs
* ~~Security~~ to be addressed to handle the need for scaling rapidly & elastically.
* Security:- Most of the NoSQL big data platform have poor security mechanisms when it comes to safeguarding big data.. But Big Data carries credit card info, personal information, & etc..
* Schema:- Rigid schemas have no place. Technology should be able to fit our bigdata & not the other way around.
* Continuous availability:- The big question here is how to provide 24/7 support because almost all RDBMS & NoSQL big data platforms have a certain amount of downtime built in.

* Consistency:- should one opt for consistency or eventual consistency?
* Partition tolerant:- How to build partition tolerant systems that can take care of both h/w & s/w failures.
* Data Quality:- How to maintain data quality - data accuracy, completeness, timeliness etc.? Do we have appropriate metadata in place?

→ Big Data Analytics Important :-

① Reactive-Business Intelligence:-
* BI allows the business to make faster & better decisions by providing the right information to the right person at the right time in the right format.
* It's about analysis of the past or historical data & then displaying the findings of the analysis or reports in the form of enterprise dashboards, alerts, notifications, etc.
* It has support for both pre-specified reports as well as adhoc querying.

② Reactive - Big Data Analytics:-
* the analysis is done on huge datasets but the approach is still reactive as it is still based on static data.

③ Proactive - Analytics:-
* This is to support futuristic decision making by the use of data mining, predictive modeling, text mining, & statistical analysis.
* This analysis is not an big data as it still uses the traditional database management practices on big data & therefor has severe

limitations on the storage capacity & the processing capability.

(4) Proactive- Big Data Analytics:-

* This is sieving through terabytes, petabytes, exabytes of information to filter out the relevant data to analyse:
* This also include high performance analytics to gain rapid insights from big data & the ability to solve complex problems using more data.

→ Basically Available Soft state Eventual Consistency (BASE):-

* where it is used?
→ In distributed computing
* why it is used?
→ To achieve high availability
* How is it achieved?
→ Assume a given data item. If no new updates are made to this given data item for a stipulated period of time, eventually all accesses to this data item will return the updated value.
* What is replica convergence?
→ A sln that has achieved eventual consistency is said to have converged or archieved replica convergence.
* Conflict resolution: How is the conflict resolved?
→ (a) Read repair:- If the read leads to discrepancy or inconsistency, a correlation is initiated. It slows down the read operation.
(b) Write repair:- If the write leads to discrepancy or inconsistency, a correlation is initiated. This will cause the write operation to slow down.
(b) Asynchronous repair:- Here, the correlation is not part of a read or write operation.