1] Basics of Data science & Big data

Data :- Raw facts & figures such as numbers, text, or images that can be collected & analyzed for insights.

DS :- The study of extracting meaningful insights from data using scientific methods, algorithms & systems.

Big data :- Massive & complex datasets that are difficult to process using traditional methods due to their size and diversity.

—x— x —x— x—

Need of data science

1. **Extract useful insights** – Helps in identifying trends and patterns from complex data
2. **Predictive analytics** – Enables forecasting future outcomes using historical data
3. **Automation** – Automates processes like anamoly detection, recommendation sys., etc.
4. **Personalization** – Enhances user experience by delivering personalized services.
5. **Efficient data management** – Organizes, processes, and cleans large amounts of data for better usage.

Need of Big Data.

1. **Handling large scale data.** – Helps manage & analyze massive datasets from various sources.
2. **Real time analysis.** – Enables processing and analyzing data as it's generated.
3. **Competetive advantage.** – Provide deeper insights, giving buisnesses an edge over competitors.
4. **Operational efficiency** – Streamlines operations by optimizing processes & reducing costs.
5. **Innovation** – Fuels new product development & innovations.

Applications

1. **Health care** – Used to build sophisticated medical instruments to detect & cure diseases.
2. **Gaming** – Video & computer games are created by using data science.
3. **Image Recognition** – Identifying patterns in image & detecting object in image is one of the most popular data science application
4. **Logistics**
5. **Product future market trends**

Difference betn Data science & Big Data.

Aspect	Data science	Big data.
defn	study of extracting meaningful insights from data by scientific methods, algorithm & system.	It is the massive & complex data which can't proccessed by using traditional methods.
Primary focus	Analyzing data to extract meaningful info & insight	Handling, storing & processing large volumes of data efficiency.
Data size	It can work from small to large data	Focus is on very large data
Tools used	Python, R, Jupyter Notebooks, Tensorflow, scikit-learn, pandas, SQL.	Hadoop, Spark, HDFS, NoSQL.
Key goals	Understanding data patterns, making predictions & deriving actionable insights	Efficient storage, processing & retrieval of massive data sets for future use.

2) Data explosion.

→ The essence of computer applications is to store things in real-world into computer system in the form of data.

→ The large scale of data is rapidly generated & stored in computer system which is called data explosion.

→ Data is generated automatically by mobile device computers, think facebook, search queries, GPS location, directions & image capture.

→ Sensors also generate data, including medical data & commerce location-based sensors.

→ The phenomenon of exponential multiplication of data that gets stored in termed as Data explosion.

The data world is governed by three fundamental trends are

Reasons of Data Explosion.

(1) Buisness model transformation

→ The buisness are required to produce more data related to product & provide services to cater each sector & channel of customer.

(2) Globalization.

→ Globalization is an emerging trend in buisness where organization start operating on international scale.

→ Variety & different formats of data is generated due to globalization.

(3) Personalization of services.

→ To enhance customer service, the form of one-to-one marketing in the form of personalization of service is opted by customer

→ customer expects communication through

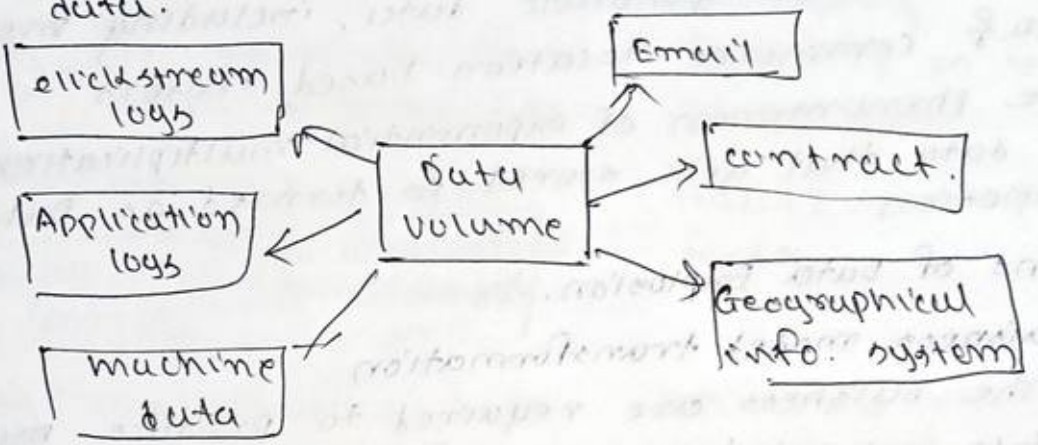
5 V's of Big Data

- Volume
- Velocity
- Variety
- Veracity
- Value

(1) Volume

→ volumes of data are larger than conventional relational database infrastructure can cope with

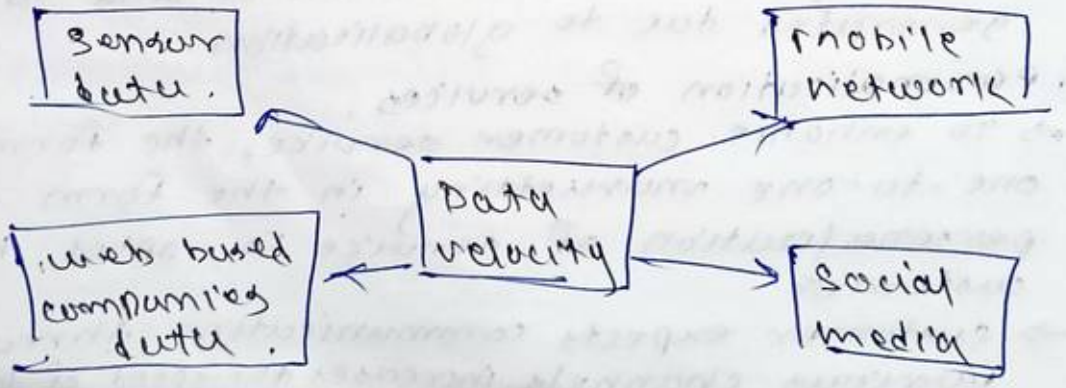
→ It consist of terabytes or petabytes of data.



(2) Velocity.

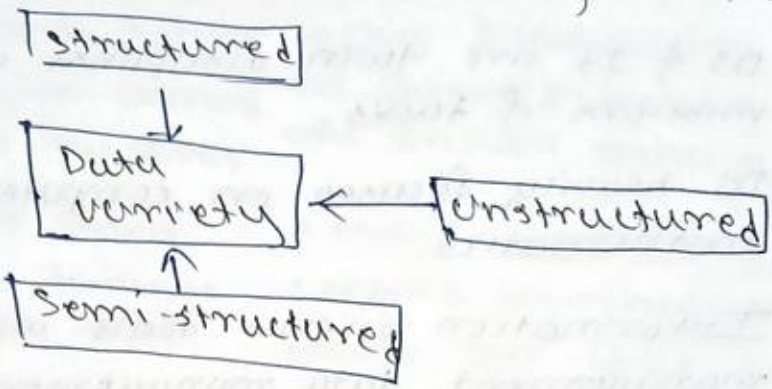
→ The terme velocity refers to the speed of generation of data.

→ How fust data is generated & processed to meet demands, determines potential in data



③ Variety.

→ It refers to heterogeneous sources & the nature of data, both structured & unstructured



④ Value

→ It represents the business value to be derived from big data.

→ The ultimate objective of any big data project should be to generate some sort of values for the company doing all the analysis.

⑤ Veracity

→ Veracity refers to the trustworthiness & quality of data.

→ It ensures data accuracy by evaluating it's credibility, integrity & messiness as data is often sourced from multiple origins

→ High veracity mean more reliable insights.

Relationship between Data science & info. science.

→ Data science is interdisciplinary field combining mathematics, statistics, information science & computer science.

→ DS & IS are twin disciplines with similar missions & tasks.

→ DS heavily focuses on computer science & mathematics.

→ Information science deals with knowledge management, data management & interaction design.

→ It involves collection, storage, retrieval & use of information & technologies related to managing recorded information & knowledge.

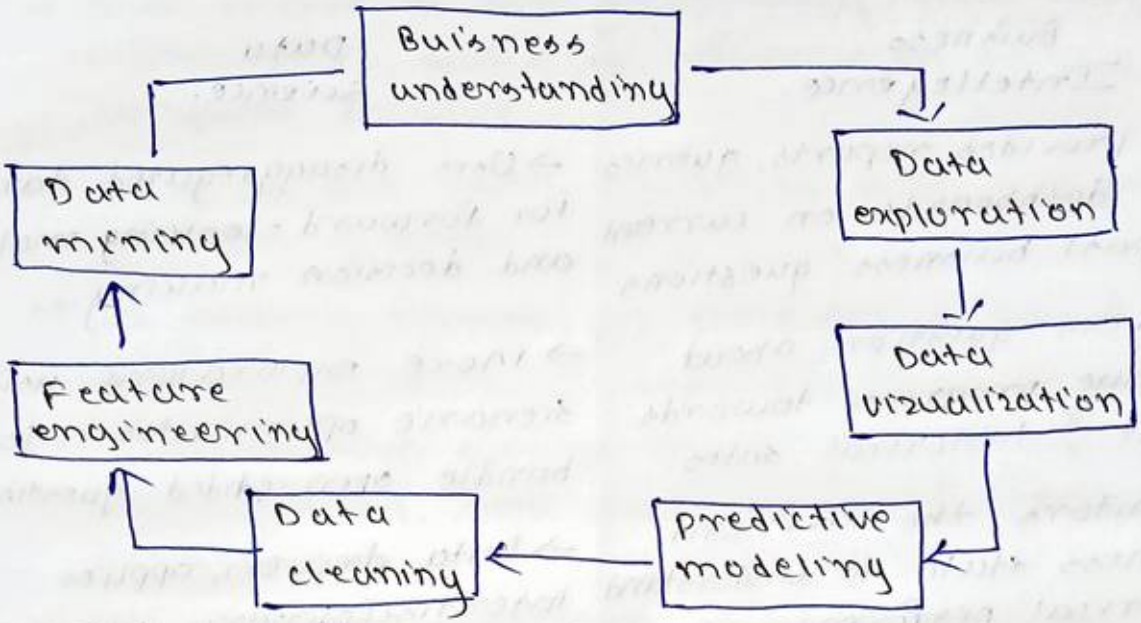
Buisness Intelleyence V/s Data science.

Buisness Intelleyence.	Data science.
→ It provides reports, queries and dashboards on current or past buisness questions	→ Uses disaggregated data for forward-looking analysis and decision making.
→ Answers question about revenue, progress towards target & historical sales.	→ more exploratory, using scenario optimization to handle open-ended questions.
→ Monitors the current buisness state to understand historical performance	→ Data driven, applies interdisciplinary sciences to extract insights & meaning.
→ Designed for static, highly structured data	→ Handles high-speed, high-volume, multi-structured data from various sources.

Cloud computing V/s Big Data.

Cloud Computing	Big Data.
→ provides resources on demand	→ Handles huge amount of data & generate insights
→ Refers to internet services from Saas, Paas, Iaas	→ Refers to structured, semi or unstructured data
→ Used to store data and information on remote servers	→ Describes huge volumes of data & information.
→ Economical due to low maintainance	→ cost effective ecosystem
→ vendors: Google, AWS, IBM, Apple, IBM	→ vendors: cloudera, Apache, mapR.
→ Focus: providing computer resources & services via network connection	→ Focus: Solving problem with large amounts of data generation & processing

Data Science life cycle.



(1) Buisness understanding :-
→ Understanding the buisness problem & objectives

(2) Data Exploration :-
→ Examining data to understand it's structure, content & patterns

(3) Data Visualization :-
→ Representing data in graphical formats to identify trends & insights.

(4) Predictive modelling :-
→ Building models to make predictions based on data patterns.

(5) Data Cleaning :-
→ Removing inconsistencies, errors or missing values in the data.

(6) Feature engineering :-
→ Creating new features or transforming existing data to improve model performance.

(7) Data mining :-
→ Extracting valuable patterns, trends from data

Data

→ Data is raw, unorganized info. like numbers, words or descriptions, which are fundamental for producing meaningful info.

→ Data can include printed papers, bank passbooks, student attendance or salary sheets.

→ Technology convert data into convinient digital form's such as emails, e-books, videos & images

→ Data itself has no inherent meaning until processed into information, often stored in digital formats (0 & 1's)

→ Info. System help store, manage & process raw data to generate useful info., essential for organization in today's digital age.

Data Types :- → Structured
→ Unstructured.

① Structured data.

→ Organized in rows & column's for easy retrieval and processing.

→ Easily identifiable & stored in database like DBMS.

→ Common example are database tables with a clear structure

→ Searchable by data type & efficiently organized

→ Readable by both computer & humans.

(2) Unstructured data.

→ Lacks a specific format ; not organized in rows & columns.

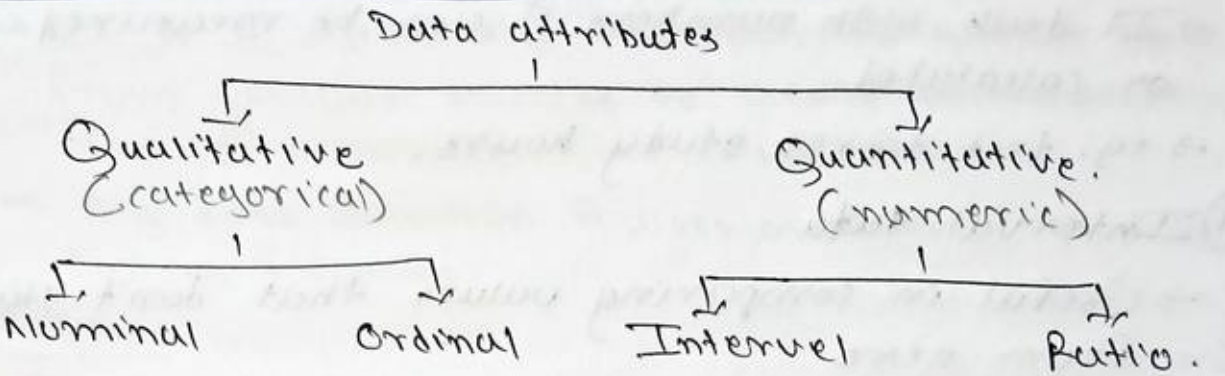
→ Examples include, texts, emails, audio, video, and images.

→ Makes up over 80% of organizational data, but hard to extract useful info.

→ Has no predefined structure, format or sequence.

→ Unpredictable & flexible due to no structural binding.

Qualitative & Quantitative data.



(i) Qualitative data.

→ provides non-measurable, descriptive information.
→ Often involves categories like gender, economic status or preferences.
→ Cannot be expressed as number & includes descriptions.

——— OR ———

→ Describes qualities that can't be measured
→ Focuses on descriptions rather than numerical values
→ Example, economic status, religious preferences

(i) Nominal data.

→ It helps to group data but doesn't imply any order or rank.
→ Used in statistical models as identifiers.
→ Numbers used as tags (male = 1, female = 0)

(ii) Ordinal data.

→ Represents data with a specific order or ranking
→ categories have a meaningful order but no fixed difference betn ranks.
→ Example: University ranking (1st, 2nd, 3rd)

Quantitative data

→ It deals with numbers & can be measured or calculated.

→ eg. test scores, study hours.

(i) Interval data

→ Useful in comparing values that don't start from zero

→ Doesn't allow for ratios because there's no absolute zero.

→ Quantifies differences & allows for calculating mean & median.

→ eg. time on a clock.

(ii) Ratio data.

→ Includes zero, ~~for~~ making calculations like multiplication / division meaningful.

→ eg. weight, height, age.

1) Data collection

→ It is a systematic approach to gather info. from various sources to create a comprehensive view of a subject.

→ Big data collection focuses on data from human & devices.

Types

(i) Network data :-
→ Gathered from all kinds of n/w like social media, info n/w, internet, mobile n/w.

(ii) Real-time data :-
→ Data generated by online streaming platforms.

(iii) Transactional data :-
→ captured during online purchases.

(iv) Geographical data :-
→ Location data from devices, vehicles & objects.

(v) Natural language data :-
→ Mostly from voice searches or text queries entered into devices connected to the internet.

(vi) Time series data :-
→ Related to the observation of trends over time.

Data Wrangling :

→ It is the process of converting raw data into usable format through cleaning organizing and transforming.

Tasks in data wrangling :

→ Merging datasets
→ Handling missing data
→ Removing outliers
→ Standardizing inputs.

Goals :

→ Producing reliable, usable data for business decisions
→ Saving analyst time by creating well-organized data models
→ Ensuring consistency & security in data warehousing.

Key steps in Data wrangling.

① Discovering
→ Discovering what's in the data to decide how to analyze it
② structuring
→ organizing data for easier computation.
③ Cleaning
→ Fixing errors, handling null values & standardizing data for better quality.
④ Enriching
→ Adding new data or features that improve dataset's usefulness
⑤ Validating
→ Ensuring data is consistent & accurate
⑥ Publishing
→ making the clean data available for analysis.

Benefits of Data wrangling.

→ Enhances data usability by converting it into a compatible format.

→ Speeds up the creation of data flow.

→ Integrates various types of information and sources.

→ facilitates the processing & sharing of large data volumes.

→ Improves efficiency in data-driven decision-making.

Data cleaning

→ Real world data is often incomplete, noisy & inconsistent.

→ It involves tasks like data acquisition, filling missing values, converting data formats.

→ It is the first step in preprocessing to ensure reliable, usable data.

(i) Missing values

→ Missing data can lead to unreliable outputs.

→ Techniques for handling missing data include :-

- Ignoring the tuple
- Filling missing values manually.
- Using a global constant.
- Using the attribute mean.
- Using the mean for all samples in same class.
- Using the most probable value.

(ii) Noisy data

→ Noisy is a random error or variance in measured variable.

→ Smoothing techniques like :

- Binning (mean, medians or boundaries of bins)
- Regression (linear or multiple)
- Outlier analysis (like clustering)

Data Integration & Transformation

① Data Integration :-

→ It combines data from multiple sources to create a unified data store.

→ Techniques like metadata, correlation analysis, conflict detection, and resolving semantic heterogeneity help smooth the integration process

→ It is essential because it provides a unified view of scattered data & maintains data accuracy

→ common issues in data integration.

① Entity identification problem

- matching real-world entities from diff. source is challenging
- one source might use "customer_id" while another uses "customer_number"
- Schema integration using metadata can resolve this.

② Redundancy

- Redundant data, such as attributes that can be derived from others, can cause issues.
- correlation analysis helps detect interdependent data & eliminate redundancy.

(2) Data transformation

→ It refers to converting or consolidating data into formats that are suitable for data mining.

→ common transformation techniques include:

(i) **Smoothing** :- It removes noise from data through methods like binning, regression

(ii) **Aggregation** :- Summarizing data to simplify it.

(iii) **Generalization** :- transforming low-level data into higher-level concept using hierarchies.

(iv) **Normalization** :- Scaling data attributes so they fit within a specified range.

(v) **Attribute construction** :- creating new attributes based on existing on to assist in the mining process.

→ Normalization involves scaling the value of an attribute to fall within specific range

→ using methods like, min-max normalization z-score normalization & decimal scale normalization.