DATA SCIENCE

0287

Data: Data is information such as facts and numbers used to analyse something or make decisions.

Data Science: Data science is the field of study, that combines domain expertise, programming skills and knowledge of mathematics and statistics to extract meaningful insights from data.

Types of Data:

1. Qualitative
   - Nominal Data
   - Ordinal Data
2. Quantitative
   - Discrete data
   - Continuous data

Nominal data :- Nominal data is used just for labelling variables, without any type of quantitative value.

Ordinal data :- Ordinal data is data which is placed into some kind of order by their position on a scale

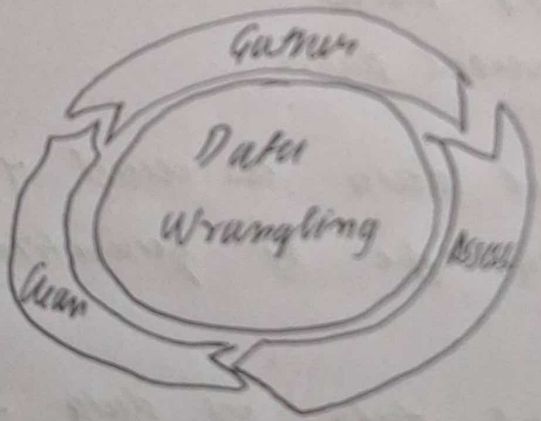
Data source: the location where data is being used to originate from.

Types of Data sources:

- Relational
- Multi dimensional
- Dimensional modelled relational

Data Wrangling

Data wrangling is the process of removing errors and combing complex data sets to make them more accessible and easier to analyze.

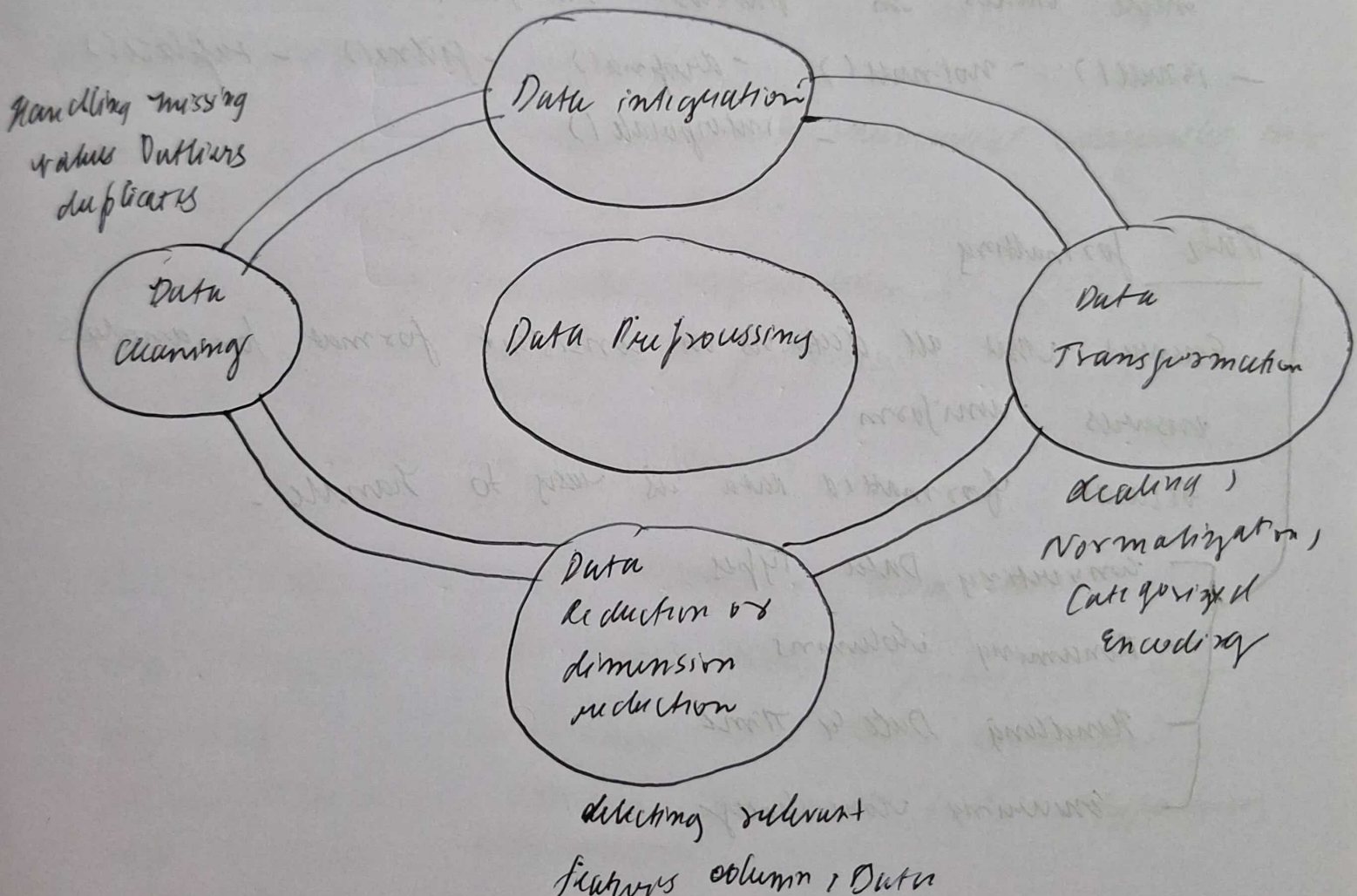


Main concept of Data science is collection of data, design and analysis.

Big Data

Big data refers to significant volumes of data that cannot be processed effectively with the traditional applications that are used.

- Velocity → speed generating data
- Volume → size · KB – PB
- Variety
  - Structured
  - Unstructured
  - Semi structures
- Complex · of large datasets.



DATA CLEANING

involves the systematic identification and correction of inconsistencies and inaccuracies within a dataset.
- handling missing values - removing duplicates
- addressing outliers.

Essential because raw data is messy.

Handling missing values

several useful functions for detecting, removing or replacing null values in Pandas DataFrame.

- isnull() - notnull() - dropna() - fillna() - replace()
- interpolate()

Data formatting

Ensures that all data is in consistent format for analysis
ensures uniform
because formatted data is easy to handle.

- converting Data types
- Renaming columns
- Handling Date & Time
- Ensuring consistency

Data normalization

process used to scale numerical features in a dataset to a standard range.

ensures no single feature to dominate due to weight.

Why?

- Improves Model Performance
- Speeds up training
- Prevents Numerical Instability.

Methods

- Min - Max Normalization
- Z score Normalization (Standardization)

Binning

process of converting continuous numerical variables into discrete intervals or bins.

useful in simplifying the representation of data.

Also helps in identifying patterns

Exploratory Data Analysis (EDA)

Is an approach to analyzing data sets.

crucial step in the data analysis process, helps analysts to understand the data's structure, detect patterns, identify anomalies, test hypotheses and check assumptions.

helps making informed decisions.

Components of EDA

- Descriptive statistics - Summary statistics
- Data Visualization
- Data Cleaning
- Data Transformation
- Correlation Analysis

Correlation

Is a statistical technique that measures the strength and direction of the linear relationship between two quantitative variables

- Pearson Correlation
- Spearman Correlation
- Kendall Correlation