Name: Chhagan Kumawat
Roll no: 2023020067

Assignment 3

Q1 What is data preprocessing?

→ Data preprocessing :-

Data preprocessing is a crucial step in the data analysis and machine learning pipeline. It involves transforming row data into a clean and usable format.

The main step of data preprocessing include.

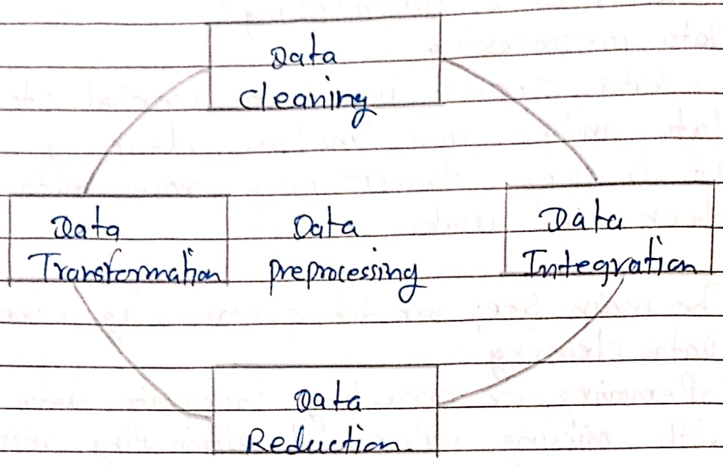
1. Data cleaning :-
Removing or correcting inaccurate data and daling with missing value and eliminating noise

2. Data Integration :-
Combining data from different sources into a coherent dataset.

3. Data Redunction
Reducing the volume of data by aggregating, summerizing or simply compares into form.

4. Data Transformation :-
Normalizing or scaling data to a constistent formate.

The goal of data preprocessing is to improve the quality of the data and make suitable for analysis or modeling.



Data preprocessing.

Q2. What is ETL in the context of Data Integration

⇒ ETL :-

ETL stand for Extract Transform load and it is a process used in data integration to move data from various sourcess to a target data warehouse.

· Data Integration :-

Difference type of data from diffrent Sourcess conheerent together.

Types

1) Type coopling - Actual data Cones together.

2) loose Copling - using interface but not actual data

* Issues of Data Integration :-

1) Schema integration & object matching, Redundancy

2) Reduction & deeduction of data values

3) No change or difficult to change.

* The steps involved in ETL are :-

- Extract : Retrieving raw data from different Sources.
- System such as database APIs or flat files.

- Transform~~at~~ :-

Cleaning transforming & enriching the data to meet the requirements of the data target system.
This include data cleaning, normalization, aggregation & applying business rules.

- Load : loading the tranformed data into target systems which could be a data warehouse, data mart or database.

- ETL :-

ETL is essential for consolidating data form disparate sources ensuring data quality and preparing data for analysis.

Q 3. Why is Data cleaning important during Data preprocessing ?

→ Data cleaning :-

Data cleaning is critical part pre-processing because it ensures the accuracy consistency and reliability of the data.

- Clean data is crucial for several reasons

• Improves Data Quality : Removing errors, inconsistencies and duplicates enhances the overall quality of data.

- Enhance analysis Accuracy :-
  Clean data leads to more accurate & reliable result in data analysis & machine cloaming model.

- Reduce Noise :-
  Eliminating irrelevant or redundant information helps in focusing on meaningful Patterns and insights.

- Increases Efficiency :-
  Clean data reduce the time and effort required to process & analyze the data.

- Improve Model performance :-
  By removing errors and inconsistencies ; data cleaning enhances the performance of machine learning models and other analytical techniquee.

- Enhance Business Decision - making :
  Reliable data ~~&~~ insights, made possible by data cleaning, support informed business decision and strategy development.

Q4. What are some common Data Transformation techniques used in ETL?

→ Data Transformation in ETL involves converting data form its original format to a format that is suitable for analysis & reporting.

- Data Normalization
  Scaling data to a standard range, such [0, 1] or [-1, 1] to ensure consistency.

- Data Aggregation :-
  Summarizing data to a higher level, such as calculating tools, average or counts.

- Data Encoding :-
  Converting data with additional information, such as adding geolocation data based on IP addresses.

- Data filtering :
  Removing irrelevels data.

- Data Sorting :
  ~~Arg~~ Arranging data in particular order.

- Data Cleaning :-

Removing unwanted data ~~or records based on on specific conditions~~. filtering, filling in missing values, and correcting errors to ensure data quality.

- Reshaping :-

Converting data structures, such as pivoting tables or aggregating data, to prepare it for analysis.

Q'5 How does Data loading differ from Data Extraction of Transformation in ETL:

→ Data loading is the final step in the ETL process & differ from data extraction & transformation in fallowing ways:

- Data Extraction :

This is the first step where row data is collected from various source system. The primary focus is on retrieving data without any modification.

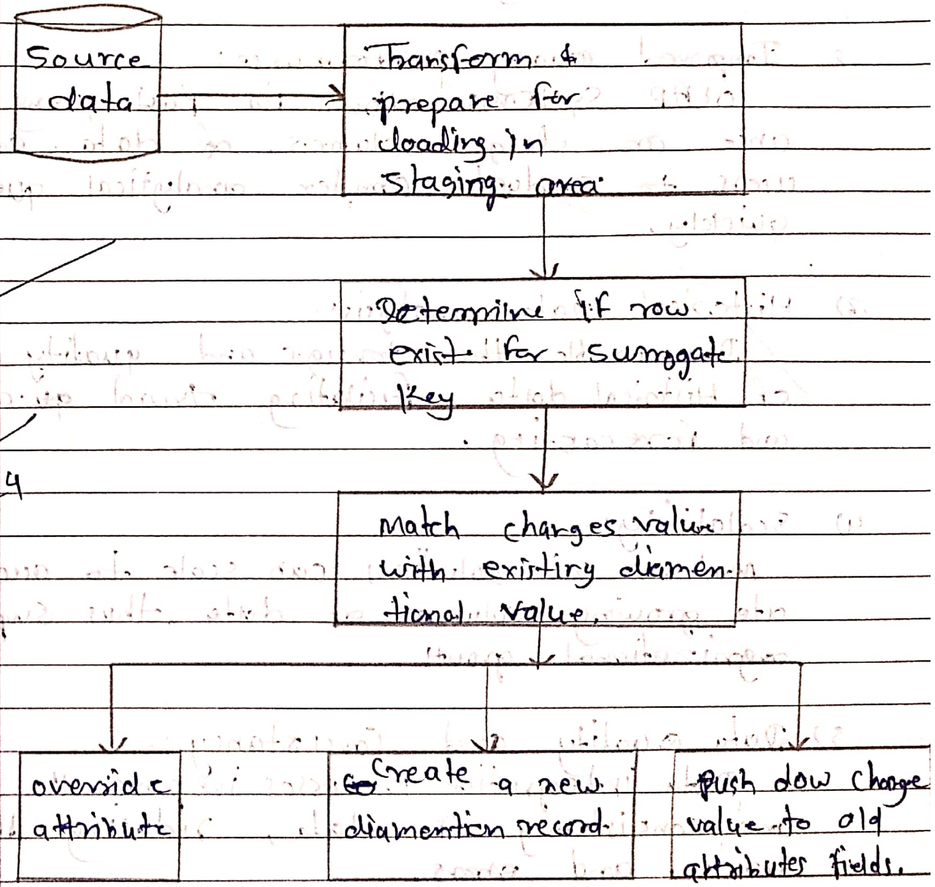
- Data Transformation :

This step involve cleans transforming and enhancing the extraction data to need the reew ~~ret~~ remente of the toraet System It ~~do~~ woudes applying business rules, normalizing data & aggregation information.

16/8/24

- Data loading :-

This is final step where re transformat date if moved into target sytem such as a data warehouse. The focus is an ensuring that the data is accuraticy and efficiency load into the target system for. E


Source data
Transform & prepare for loading in staging area
Determine if row exist for surrogate key
Match changes value with existing diamentional value.
override attribute
Create a new diamention record
Push dow change value to old attributes fields.

Data loading