Data Cleaning

Data cleaning, also known as data cleansing or scrubbing, involves identifying and correcting or removing errors, inaccuracies, and other anomalies in a dataset.

Reason for Data quality issue.

1. *Missing data*: This refers to the absence of values in data fields.
2. *Incorrect data*: Incorrect or dirty data includes erroneous, inaccurate, or invalid values
3. *Duplicate data*: Duplicate or redundant data occurs when multiple instances of the same or similar records exist in a dataset.
4. *Inconsistent data*: Inconsistent data refers to data that deviates from an expected pattern or format.
5. *Outliers:* Outliers are extreme values that significantly differ from the majority of the data points.

Impact of Poor Data Quality on Analytics & Decision-making

- *Inaccurate insights*: Datasets with quality issues can lead to incorrect or biased analytical results.
- *Misinformed decisions*: It can mislead decision-makers, leading to poor judgments and decisions.
- *Reduced trust and credibility*: It can erode confidence in the data analysis process.
- *Inefficient resource allocation*: Messy data can lead to inefficient allocation of resources.
- *Increased costs*: Dealing with it can incur additional costs. Cleaning errors requires time & effort.

SQL Data Cleaning: Key Concepts

SQL used for data cleansing tasks due to its ability to efficiently retrieve, filter, update, and delete data.

- SELECT statement: Retrieves data from one or more tables or views.
- WHERE clause: Filters data based on specified conditions.
- UPDATE statement: Modifies existing data in a table.
- DELETE statement: Removes data from a table.
- DISTINCT keyword: Retrieves only unique/distinct values from a column.
- String functions: TRIM, UPPER, LOWER, and REPLACE
- Aggregate functions: COUNT, SUM, AVG, MAX, and MIN. It is useful for identifying outliers or calculating ranges.

How Data Cleaning done via SQL

- *Removing duplicate records*: using the DISTINCT keyword or by grouping data on specific columns and selecting distinct values.
- *Handling missing values*: You can remove rows with a null value [DELETE] or impute them with valid ones [DEFAULT VALUES].
- *Correcting inconsistent or invalid data*: string functions can standardize and clean messy data.
  - TRIM to remove leading and trailing spaces
  - UPPER or LOWER to convert text to a specific case
  - REPLACE to replace specific characters
- *Data normalization*: Data may have different formats across columns or tables in a database. Need standardize formats. TO_DATE function to convert date strings to a specific date format.
- *Handling outliers:* identify and address outliers by calculating summary statistics and then removing or adjusting values that fall outside an acceptable range.
- *Verifying data integrity*: Ensure integrity using constraints, such as primary key and foreign key constraints, to enforce relationships between tables and prevent invalid data

Key steps involving Data Cleaning in SQL

- *Profiling and assessment*: Understand the data types, structure, quality, and content. Identify quality issues such as duplicate values, inconsistencies, and outliers.
- *Data validation and filtering*: Validate data against predefined rules or criteria. Filter out irrelevant or erroneous records based on specific conditions or constraints.
- *Fixing missing data*: Decide how to handle missing data. Identify rows with null values and decide whether to remove or impute them based on your data cleansing strategy.
- *Standardization and transformation*: Standardize formats, units, or values to ensure consistency.
- *Removing duplicates*: Identify and remove duplicate values from the dataset using SQL's DISTINCT keyword or by grouping datasets and selecting distinct values.
- *Correcting errors*: use functions like TRIM, UPPER, LOWER, or REPLACE to fix inaccurate values, remove extra spaces, convert text cases, or replace specific values.
- *Handling outliers*: Identify outliers. Decide whether to remove outliers or adjust their values based on the context of the data quality project.
- *Data integrity checks and constraints*: Ensure integrity by adding or modifying primary key and foreign key constraints. This helps maintain data relationships and enforce consistency.