🙄

REMOVING DUPLICATES IN DATA

WHAT IS DUPLICATE ?

A duplicate in a database is a record that is an exact copy of another record. Duplicates can occur for a variety of reasons, such as:

- **Human error:** A user may accidentally enter the same data twice.
- **System errors:** A system error may cause the same data to be entered twice.
- **Data migration:** When data is migrated from one system to another, duplicate records may be created.

Duplicates can cause a number of problems, including:

- **Data inconsistency:** Duplicate records can lead to data inconsistency, which means that the same data may have different values in different places. This can make it difficult to track and manage data.
- **Reduced performance:** Duplicate records can reduce the performance of a database, as the database has to search through more data to find the correct record.
- **Security risks:** Duplicate records can increase the risk of security breaches, as attackers can use duplicate records to gain access to sensitive data.

In Snowflake, as we know it doesn’t enforce primary key so you can fill duplicate value in the primary key.

To remove DUPLICATE we use keyword `DISTINCT` in sql command.

```
SELECT DISTINCT * FROM TABLE_NAME;
```

IF 2 WHOLE ROW IS SAME TO SAME THEN THEY ARE DUPLICATE.

WE CAN ALSO REMOVE DUPLICATE USING SELF JOIN TOO.

Don’t use `<>` not equal to in self join because if we consider a table with self join having `<>` it means A —> B & B —> A, is right but it is a duplicate, as A is related to B and B is related to a so not equal to will give this type of duplicate value in self join better use `>` and `<`