What is normalization/Database normalization?

Normalization is the process of restructuring a relational database in accordance with a series of normal forms in order to reduce data redundancy

Redundancy is nothing but data repetition/duplication

Most commonly used normal form is 3NF

More normalized => more number of tables (If we increase the normal form then number of tables will also increase)

One thing to observe here if the tables are normalized we will have more number of small tables

Scenario1: perform DML operations on normalized tables

If we have to insert/update/delete data from those tables they will be faster since they were small tables with less columns

Scenario2: Perform Data retrieval (reads/selects) on normalized tables (Analyze data)

We will have to join data from multiple smaller tables and select statements will be slow

What are OLTP systems and purpose?

OLTP (Online Transactional processing) Systems are used to “run” a business (requires only current data and all the tables will be in normalized form).

OLTP systems Designed to handle massive DML (Insert, update, delete) operations. Tables will be in normalized form.

What is Data Warehouse?

Data warehouse= Database that is used for reporting and data analysis with data retrieval (select queries) being key and

Helps in decision making process

Data Warehouse (DW) helps to “optimize” the business (requires historical data and most of the dimension tables will be de normalized

Data warehouse will be designed with either star schema or snowflake schema primarily

What is Data mart?

Data mart is a subject oriented database which supports the business needs of individual department within an enterprise. Data mart is a subset of Data warehouse.

Difference between data mart and data warehouse?

Data warehouse is maintaining the total organization of data. Whereas data mart is maintains only segment of business area(particular subject) example Sales, Purchase..

What is star and what is snowflake? which schema design is normalized/denormalized?

Star schema is a logical database design which contains a centrally located fact table and all dimension tables are directly connected to fact.

Snow Flake Schema: In a Star schema database design, if the dimension table is split into a one or more dimension tables which results in Normalization.

i.e couple of dimension tables will not be directly connected to fact and will be connected to dimension table

Most of the projects will use star schema and its preferred since it involves lesser joins and queries will be faster since all tables are de normalized

ETL = Extract, Transform & Load

When to go with Start Schema when to go with Snowflake schema design?

Whenever your dimension table is growing fast (receiving large volumes of data) then we can normalize that particular dimension table into one more dimension table

(i.e. one dimension table will be split into one more dimension so inserting data will be faster)

What is dimension? What is Fact?

Dimension table contains/stores attributes about a specific business process (Contains detailed information)

Fact table consists of the Foreign keys and measurements, metrics or facts of a business process (contains summarized numerical data and historical data)

What are types of facts?

Additive fact: Additive Fact can be aggregated by simple arithmetical additions along all dimensions

Semi-Additive fact: Semi additive fact can be aggregated simple arithmetical additions along with some other dimensions.

Non-additive fact: Non-additive fact can’t be added at all.

Fact less Fact: It doesn’t contain facts or measures but contains only foreign keys no measures.

What are different types of Dimension tables?

Conformed Dimension

A conformed dimension is the dimension that is shared across multiple data mart or subject area. Company may use the same dimension table across different projects without making any changes to the dimension tables.

Conformed dimension example would be Customer dimension, i.e. both marketing and sales department can use Customer dimension for their reporting purpose.

Junk Dimension

A junk dimension is a grouping of typically low cardinality attributes, so you can remove them from main dimension.

You can use Junk dimensions to implement the rapidly changing dimension where you can use it to stores the attribute that changes rapidly.

For example, attributes such as flags, weights, BMI (body mass index) etc

Degenerated Dimension

A degenerated dimension is a dimension that is derived from fact table and does not have its own dimension table.

For example, receipt number does not have dimension table associated with it. Such details are just for information purpose.

Role Playing Dimension

Dimensions which are often used for multiple purposes within the same database are called role-playing dimensions.

For example, you can use a date dimension for “date of sale”, as well as “date of delivery”, or “date of hire”.

What is Surrogate Key?

It is a system generated sequence number and it is treated as primay key in the dimension table.

Surrogate Key is a series of sequential numbers assigned to be a primary key for the table.

It will not have any business meaning

What is Natural Key/Business key? Eg: customer id, loan id... (It will have business meaning)