## Star Schema vs Snowflake schema-

|  |  |
| --- | --- |
| **STAR SCHEMA** | **SNOWFLAKE SCHEMA** |
| Star schema is partially normalized.  Kind of denormalized, but not completely denormalized. | Snowflake schema is normalized completely. So only the dimensions split into multiple tables until no redundancy in the dimensional table, no repetition of values (except for identifier values, such as id’s). |
| Star schema leads to simple query writing. | Snowflake is complex query as there are many lookup tables. |
| Query execution is faster | Query execution is slow. |
| Takes more space | Takes less space |
| Data redundancy is more | No data redundancy |
| Easier to implement | Complex to implement. |
| Data integrity is more at risk | It is easy here, as the data is normalised. If any data change happens, update will require at only one place. But in star it requires at many places as it is denormalized. |

## Data Integrity-

Data integrity is a concept and process that ensures the accuracy, completeness, consistency, and validity of an organization's data. By following the process, organizations not only ensure the integrity of the data but guarantee they have accurate and correct data in their database.

## Data Normalization-

1NF – Each cell should contain single value.

2NF - All non-key attributes are fully functional and dependent on the primary key

3NF – No transitive relationship

## Data mart-

It is a subset of data warehouse. And it only depends on one functional area of data warehouse.

## Data sharding-

It is a process of splitting the large table into multiple machines. It mainly gets used in,

* Parallelism
* Reduce the out of storage in the machine

Doing the data sharding of small table is overhead of network.

## Sharding vs indexing-

Indexing is a way to store column values in a data structure aimed at fast searching. This speeds up a search tremendously compared to a full table scan since not all rows will have to be examined. You should consider having indices on the columns in your WHERE clauses.

Sharding is a technique to split the table up between different machines. This makes it possible for parallel resolution of queries. For example, half the table can be searched on one machine and the other half on another machine. This will in some cases make it possible to increase the performance by adding more hardware, especially for large tables.