DATAWARE HOUSE QUICKGUIDE

Tanoy Bhattacharya
tanoybhattacharya@gmail.com

What is data ware Housing?
A data warehouse (DW) is a collection of corporate information and data derived from operational systems and external data sources. A data warehouse is designed to support business decisions by allowing data consolidation, analysis and reporting at different aggregate levels. Data is populated into the DW through the processes of extraction, transformation and loading.

What is the basic four features about Dataware Housing?

1. Subject Oriented
2. Non Volatile
3. Integrated
4. Time Variant

What is Initial Load / Write down the Alternate name?

First Load

What is Delta Load / Write down the Alternate name?

Incremental Load

What are the differences between Inmon Vs Kimball?
Inmon – Top down Approach for enterprise Data warehouse for Strategic solution. However, it is require high cost because development time is high.

Kimball – Bottom up Approach for Dimension modelling as Data warehouse for tactical solution. However, it is require low cost because development time is short if we can compare with Inmon approach.

List Down data ware housing Schema & ETL Product.

- Microsoft SQL Server
- Teradata
- Informatica
- Oracle

What is Business Intelligence & why do we need?
Business intelligence (BI) is a technology-driven process for analysing data and presenting actionable information to help executives, managers and other corporate end users make informed business decisions.
The potential benefits of business intelligence tools include accelerating and improving decision-making, optimizing internal business processes, increasing operational efficiency, driving new revenues and gaining competitive advantage over business rivals. BI systems can also help companies identify market trends and spot business problems that need to be addressed.

What is Data Mining?
Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.

What is differences between Strategic Vs Analytical Business Intelligence & why do we need?

Strategic Business Intelligence provide map, report, scorecard and dashboard to take strategic decision. Analytical Business Intelligence like analytical dashboard to help predictive analysis using OLAP.

What is Relational database?

A relational database is a digital database based on the relational model of data, as proposed by E. F. Codd in 1970. A software system used to maintain relational databases is a relational database management system. Virtually all relational database systems use SQL for querying and maintaining the database

What is Analytical database?

An **analytic database**, also called an **analytical database**, is a read-only system that stores historical data on business metrics such as sales performance and inventory levels. Business analysts, corporate executives and other workers can run queries and reports against an **analytic database**.

Write the differenceDatas between Relational Vs Analytical Database?

Relation is a transactional processing while **Analytical** is an analytical processing system. Business analysts, corporate executives and other workers can run queries and reports against an **analytic database**. Virtually all relational database systems use SQL for querying and maintaining the database

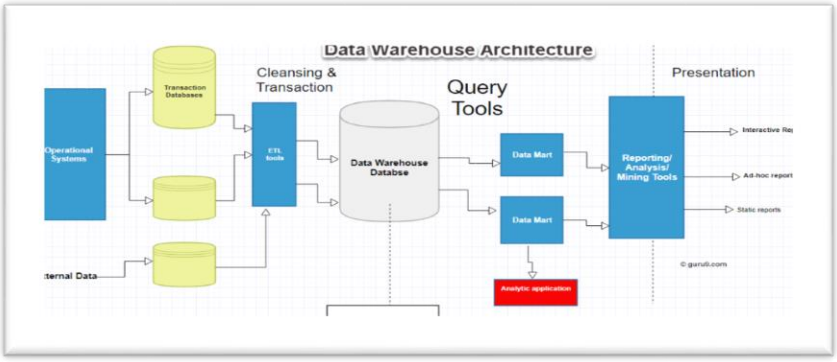
Write the differences between Dataware House Vs Data Mart?

Differences between Data Warehouse and Data Mart. A Data Warehouse is a large repository of data collected from different organizations or departments within a corporation. A data mart is an only subtype of a Data Warehouse. It is designed to meet the need of a certain user group. Data Warehouse. A data warehouse is a big central repository for all of an organization's historical data. ... The size of a data warehouse is typically larger than 100 GB, whereas data marts are generally less than 100GB. Due to the difference in scope, it is comparatively easier to design and use data marts.

What is the differences between Dependent Vs Independent Data Mart?

Dependent data marts draw **data** from a central **data** warehouse that has already been created. **Independent data marts**, in contrast, are standalone systems built by drawing **data** directly from operational or external sources of **data** or both. Hybrid **data marts** can draw **data** from operational systems or **data** warehouses.

Please draw & design Data ware House Architecture Diagram?



What is the Staging Area and why do we need?

A staging area, or landing zone, is an intermediate storage area used for data processing during the extract, transform and load (ETL) **process**. The data staging area sits between the data source(s) and the data target(s), which are often data warehouses, data marts, or other data repositories.

Staging table is a temporary **table** that is **used** to stage the data for temporary purpose just before loading it to the Target **table** from the Source **Table**. As the data resides temporarily, you can do various stuff on that data like ... Normalizing to multiple **tables**. De-Normalizing from multiple to a single **table**.

Specify and explain the Metadata Categories?

Metadata is data that describes other data. Meta is a prefix that in most information technology usages means "an underlying definition or description.".. For **example**, author, date created and date modified and file size **are examples** of very basic document **metadata**.

Purpose: Identification of a core set of **metadata elements** to be used in the development, testing, and implementation of multiple repositories. ... Types of **Metadata**: **Metadata elements** listed in the table are categorized according to three types: descriptive, administrative, and structural.

What is the OLAP Cube and why do we need Cube?

OLAP cube. An **OLAP cube** is a multidimensional database that is optimized for data warehouse and online analytical processing (**OLAP**) applications. An **OLAP cube** is a method of storing data in a multidimensional form, generally for reporting purposes.
In **OLAP cubes**, dimensions categorize data (measures).

What are the OLAP Categories and explain their features?

MOLAP, ROLAP, and HOLAP. Data Warehousing > Concepts > MOLAP, ROLAP, And HOLAP. In the **OLAP** world, there are mainly two different **types**:
Multidimensional **OLAP** (MOLAP) and Relational **OLAP** (ROLAP). Hybrid **OLAP** (HOLAP) refers to technologies that combine MOLAP and ROLAP.

Relational OLAP (ROLAP) –Star Schema based

Considered the fastest growing OLAP technology style, ROLAP or "Relational" OLAP systems work primarily from the data that resides in a relational database, where the base data and dimension tables are stored as relational tables. This model permits multidimensional analysis of data as this enables users to perform a function equivalent to that of the traditional OLAP slicing and dicing feature. This is achieved thorough use of any SQL reporting tool to extract or 'query' data directly from the data warehouse. Wherein specifying a 'Where clause' equals performing a certain slice and dice action.

One advantage of ROLAP over the other styles of OLAP analytic tools is that it is deemed to be more scalable in handling huge amounts of data. ROLAP sits on top of relational databases therefore enabling it to leverage several functionalities that a relational database is capable of. Another gain of a ROLAP tool is that it is efficient in managing both numeric and textual data. It also permits users to "drill down" to the leaf details or the lowest level of a hierarchy structure.

However, ROLAP applications display a slower performance as compared to other style of OLAP tools since, oftentimes, calculations are performed inside the server. Another demerit of a ROLAP tool is that as it is dependent on use of SQL for data manipulation, it may not be ideal for performance of some calculations that are not easily translatable into an SQL query.

Multidimensional OLAP (MOLAP) –Cube based

Multidimensional OLAP, with a popular acronym of MOLAP, is widely regarded as the classic form of OLAP. One of the major distinctions of MOLAP against a ROLAP tool is that data are pre-summarized and are stored in an optimized format in a multidimensional cube, instead of in a relational database. In this type of model, data are structured into proprietary formats in accordance with a client's reporting requirements with the calculations pre-generated on the cubes.

This is probably by far, the best OLAP tool to use in making analysis reports since this enables users to easily reorganize or rotate the cube structure to view different aspects of data. This is done by way of slicing and dicing. MOLAP analytic tool are also capable of performing complex calculations. Since calculations are predefined upon cube creation, this results in the faster return of computed data. MOLAP systems also provide users the ability to quickly write back data into a data set. Moreover, in comparison to ROLAP, MOLAP is considerably less heavy on hardware due to compression techniques. In a nutshell, MOLAP is more optimized for fast query performance and retrieval of summarized information.

There are certain limitations to implementation of a MOLAP system, one primary weakness of which is that MOLAP tool is less scalable than a ROLAP tool as the former is capable of handling only a limited amount of data. The MOLAP approach also introduces data

redundancy. There are also certain MOLAP products that encounter difficulty in updating models with dimensions with very high cardinality.

Hybrid OLAP (HOLAP)

HOLAP is the product of the attempt to incorporate the best features of MOLAP and ROLAP into a single architecture. This tool tried to bridge the technology gap of both products by enabling access or use to both multidimensional database (MDDB) and Relational Database Management System (RDBMS) data stores. HOLAP systems stores larger quantities of detailed data in the relational tables while the aggregations are stored in the pre-calculated cubes. HOLAP also has the capacity to "drill through" from the cube down to the relational tables for delineated data.

Some of the advantages of this system are better scalability, quick data processing and flexibility in accessing of data sources

What is Dimension Table?

Dimensions provide the "who, what, where, when, why, and how" context surrounding a business process event. Dimension tables contain the descriptive attributes used by BI applications for filtering and grouping the facts. Dimension tables are sometimes called the "soul" of the data warehouse because they contain the entry points and descriptive labels that enable the DW/BI system to be leveraged for business analysis.

What are the Dimension Table's Categories?

- **Degenerate Dimensions** - Sometimes a dimension is defined that has no content except for its primary key. This degenerate dimension is placed in the fact table with the explicit acknowledgment that there is no associated dimension table.

- **Role-Playing Dimensions** - A single physical dimension can be referenced multiple times in a fact table, with each reference linking to a logically distinct role for the dimension. The separate dimension views (with unique attribute column names) are called roles.

- **Junk Dimensions** - Transactional business processes typically produce a number of miscellaneous, low-cardinality flags and indicators. Rather than making separate dimensions for each flag and attribute, can create a single junk dimension combining them together.

- **Snowflake Dimensions** - When a hierarchical relationship in a dimension table is normalized, low-cardinality attributes appear as secondary tables connected to the base dimension table by an attribute key. When this process is repeated with all the dimension table's hierarchies, a characteristic multilevel structure is created that is called a snowflake.

- **Outrigger Dimensions** - A dimension can contain a reference to another dimension table.

What is Fact Table?

Facts are the measurements that result from a business process event and are almost always numeric. A single fact table row has a one-to-one relationship to a measurement event as described by the fact table's grain.

What are the Fact Table's Categories?

- **Transaction Fact Tables** - A row in a transaction fact table corresponds to a measurement event at a point in space and time. Atomic transaction grain fact tables are the most dimensional and expressive fact tables; this robust dimensionality enables the maximum slicing and dicing of transaction data.

- **Periodic Snapshot Fact Tables** -A row in a periodic snapshot fact table summarizes many measurement events occurring over a standard period, such as a day, a week, or a month.

- **Accumulating Snapshot Fact Tables** -A row in an accumulating snapshot fact table summarizes the measurement events occurring at predictable steps between the beginning and the end of a process. Pipeline or workflow processes, such as order fulfilment or claim processing, that have a defined start point, standard intermediate steps, and defined end point can be modelled with this type of fact table.

- **Factless Fact Tables** -although most measurement events capture numerical results, it is possible that the event merely records a set of dimensional entities coming together at a moment in time. For example, an event of a student attending a class on a given day may not have a recorded numeric fact, but a fact row with foreign keys for calendar day, student, teacher, location, and class is well defined.

What is the Dimension Attributes, Level, and Hierarchy & Surrogate Key?

A dimension table is designed with one column serving as a unique primary key. This primary key cannot be the operational system's natural key because there will be multiple dimension rows for that natural key when changes are tracked over time. In addition, natural keys for a dimension may be created by more than one source system, and these natural keys may be incompatible or poorly administered. The DW/BI system needs to claim control of the primary keys of all dimensions; rather than using explicit natural keys or natural keys with appended dates, you should create anonymous integer primary keys for every dimension. These dimension surrogate keys are simple integers, assigned in sequence, starting with the value 1, every time a new key is needed. The date dimension is exempt from the surrogate key rule; this highly predictable and stable dimension can use a more meaningful primary key.

Differences between conformed & Degenerated Dimension?

Conformed Dimensions
Dimension tables conform when attributes in separate dimension tables have the same column names and domain contents. Information from separate fact tables can be combined

in a single report by using conformed dimension attributes that are associated with each fact table.

Degenerate Dimensions - Sometimes a dimension is defined that has no content except for its primary key. This degenerate dimension is placed in the fact table with the explicit acknowledgment that there is no associated dimension table.

What is the Slowly Changing Dimension?

Slowly changing dimension. Dimensions in data management and data warehousing contain relatively static data about such entities as geographical locations, customers, or products. Data captured by Slowly Changing Dimensions (SCDs) change slowly but unpredictably, rather than according to a regular schedule.

Explain the SCD 1, 2 & 3 Types?

- Type 1: Retain Original - With slowly changing dimension type 0, the dimension attribute value never changes, so facts are always grouped by this original value.

- Type 1: Overwrite - With slowly changing dimension type 1, the old attribute value in the dimension row is overwritten with the new value; type 1 attributes always reflects the most recent assignment, and therefore this technique destroys history.

- Type 2: Add New Row -Slowly changing dimension type 2 changes add a new row in the dimension with the updated attribute values. This requires generalizing the primary key of the dimension beyond the natural or durable key because there will potentially be multiple rows describing each member.

Differences between Transaction & Cumulative Fact Table?

- Transaction Fact Tables - A row in a transaction fact table corresponds to a measurement event at a point in space and time. Atomic transaction grain fact tables are the most dimensional and expressive fact tables; this robust dimensionality enables the maximum slicing and dicing of transaction data.

- Accumulating Snapshot Fact Tables A row in an accumulating snapshot fact table summarizes the measurement events occurring at predictable steps between the

Please explain the Measure Types?

Additive, Semi-Additive, and Non-Additive Facts
The numeric measures in a fact table fall into three categories. The most flexible and useful facts are fully additive; additive measures can be summed across any of the dimensions associated with the fact table. Semi-additive measures can be summed across some dimensions, but not all; balance amounts are common semi-additive facts because they are additive across all dimensions except time. Finally, some measures are completely non-additive, such as ratios. A good approach for non-additive facts is, where possible, to store

the fully additive components of the non-additive measure and sum these components into the final answer set before calculating the final non-additive fact. This final calculation is often done in the BI layer or OLAP cube.

What is Data Normalization?

Database normalization is the process of restructuring a relational database in accordance with a series of so-called normal forms in order to reduce data redundancy and improve data integrity.

What are the Normailization Form?

- First normal form (1NF) is a property of a relation in a relational database. A relation is in first normal form if and only if the domain of each attribute contains only atomic (indivisible) values, and the value of each attribute contains only a single value from that domain.
- A relation that is in first normal form (1NF) must meet additional criteria if it is to qualify for second normal form. Specifically: a relation is in 2NF if it is in 1NF and no non-prime attribute is dependent on any proper subset of any candidate key of the relation. A non-prime attribute of a relation is an attribute that is not a part of any candidate key of the relation.
- Third normal form (3NF) is a normal form that is used in normalizing a database design to reduce the duplication of data and ensure referential integrity by ensuring that.

What is De – Normalization?

Denormalization is a strategy used on a previously normalized database to increase performance. In computing, Denormalization is the process of trying to improve the read performance of a database, at the expense of losing some write performance, by adding redundant copies of data or by grouping data.

What is the Dimensional Modeling?

Dimensional modeling is part of the Business Dimensional Lifecycle methodology developed by Ralph Kimball which includes a set of methods, techniques and concepts for use in data warehouse design.

What is the Dimension Modeling Process / Stages?

The four key decisions made during the design of a dimensional model include:

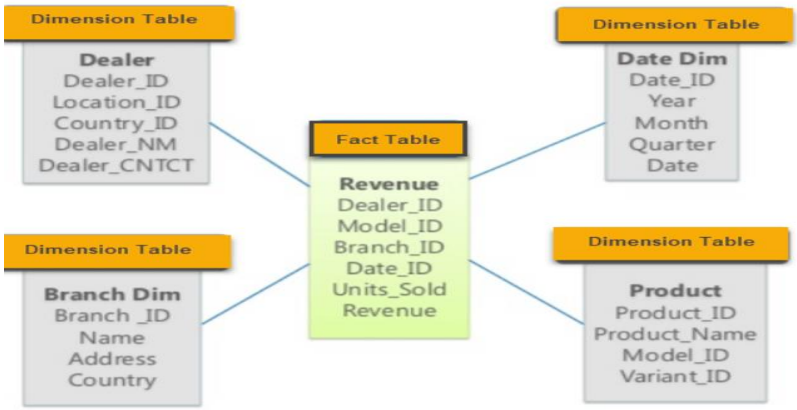
1. Select the business process.
2. Declare the grain.
3. Identify the dimensions.
4. Identify the facts

What are the Dimension Modelling –Schema Types?

- **Star Schema**: A star schema is the one in which a central fact table is sourrounded by denormalized dimensional tables. A star schema can be simple or complex. A simple star schema consists of one fact table where as a complex star schema have more than one fact table.
- **Snow Flake Schema:** A snow flake schema is an enhancement of star schema by adding additional dimensions. Snow flake schema are useful when there are low cardinality attributes in the dimensions.
- **Fact Constellation Schema:** The dimensions in this schema are segregated into independent dimensions based on the levels of hierarchy.
- **Galaxy Schema:** Galaxy schema contains many fact tables with some common dimensions (conformed dimensions). This schema is a combination of many data marts.

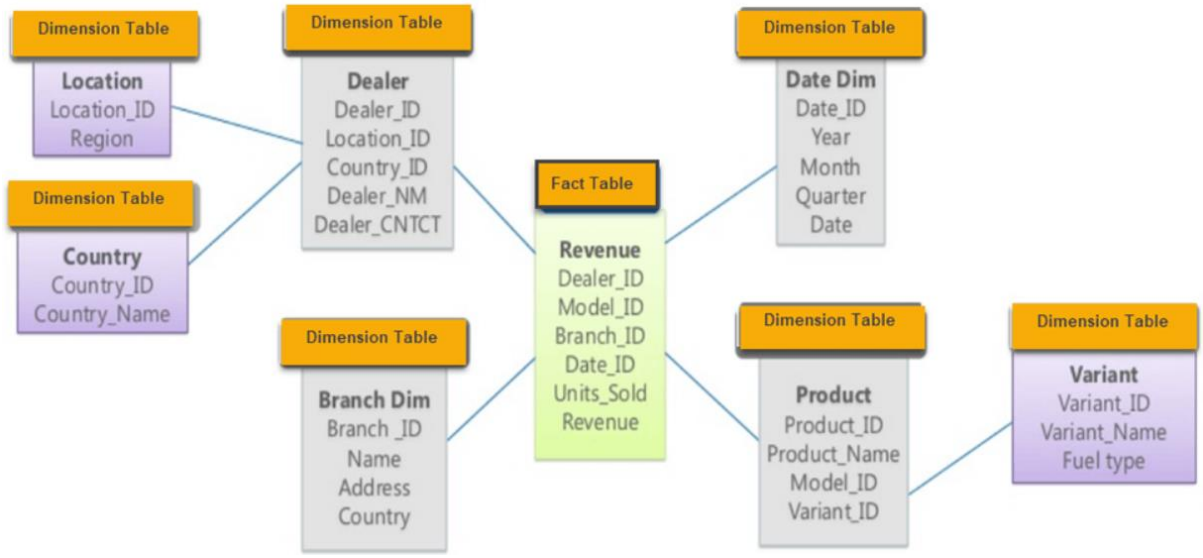
What is the Star Schema?

In computing, the star schema is the simplest style of data mart schema and is the approach most widely used to develop data warehouses and dimensional data marts. The star schema consists of one or more fact tables referencing any number of dimension tables.



What is the Snowflake Schema?

In computing, a snowflake schema is a logical arrangement of tables in a multidimensional database such that the entity relationship diagram resembles a snowflake shape. The snowflake schema is represented by centralized fact tables, which are connected to multiple dimensions.



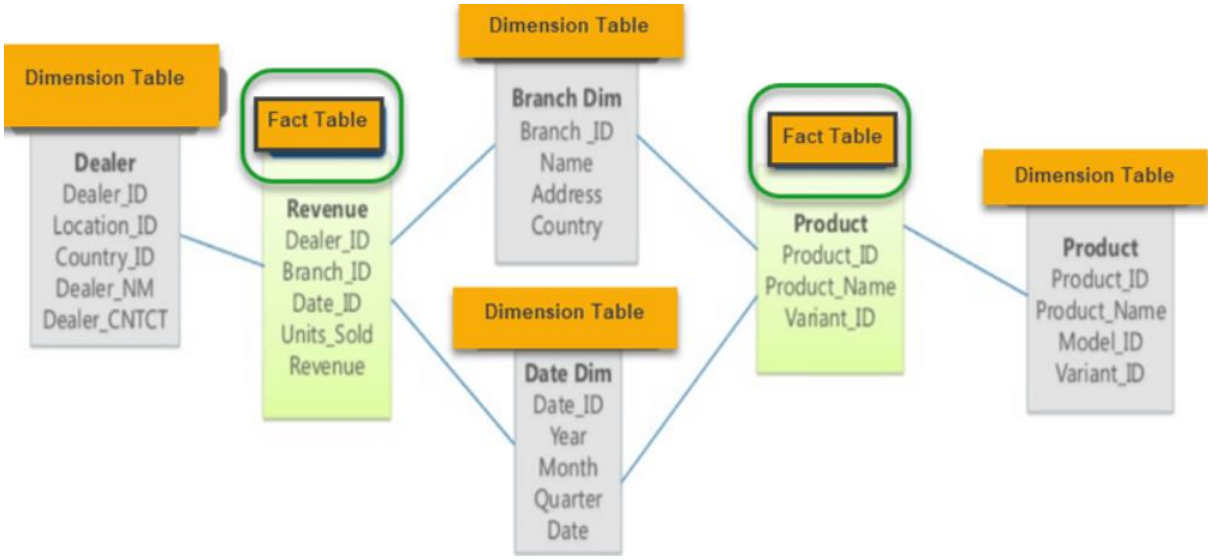
Differences between Star Vs Snowflake Schema?

Star and **snowflake schemas** are similar at heart: a central fact table surrounded by dimension tables. The **difference** is **in** the dimensions themselves. **In** a **star schema** each logical dimension is deformalized into one table, while **in** a **snowflake**, at least some **of** the dimensions are normalized.

Star Schema	Snow Flake Schema
Hierarchies for the dimensions are stored in the dimensional table.	Hierarchies are divided into separate tables.
It contains a fact table surrounded by dimension tables.	One fact table surrounded by dimension table which are in turn surrounded by dimension table
In a star schema, only single join creates the relationship between the fact table and any dimension tables.	A snowflake schema requires many joins to fetch the data.
Simple DB Design.	Very Complex DB Design.
Denormalized Data structure and query also run faster.	Normalized Data Structure.
High level of Data redundancy	Very low-level data redundancy
Single Dimension table contains aggregated data.	Data Split into different Dimension Tables.
Cube processing is faster.	Cube processing might be slow because of the complex join.
Offers higher performing queries using Star Join Query Optimization. Tables may be connected with multiple dimensions.	The Snow Flake Schema is represented by centralized fact table which unlikely connected with multiple dimensions.

Please explain the Galaxy Schema?

A Galaxy Schema contains two fact tables that shares dimension tables. It is also called Fact Constellation Schema. The schema is viewed as a collection of stars hence the name Galaxy Schema.



Sample Data Ware Design