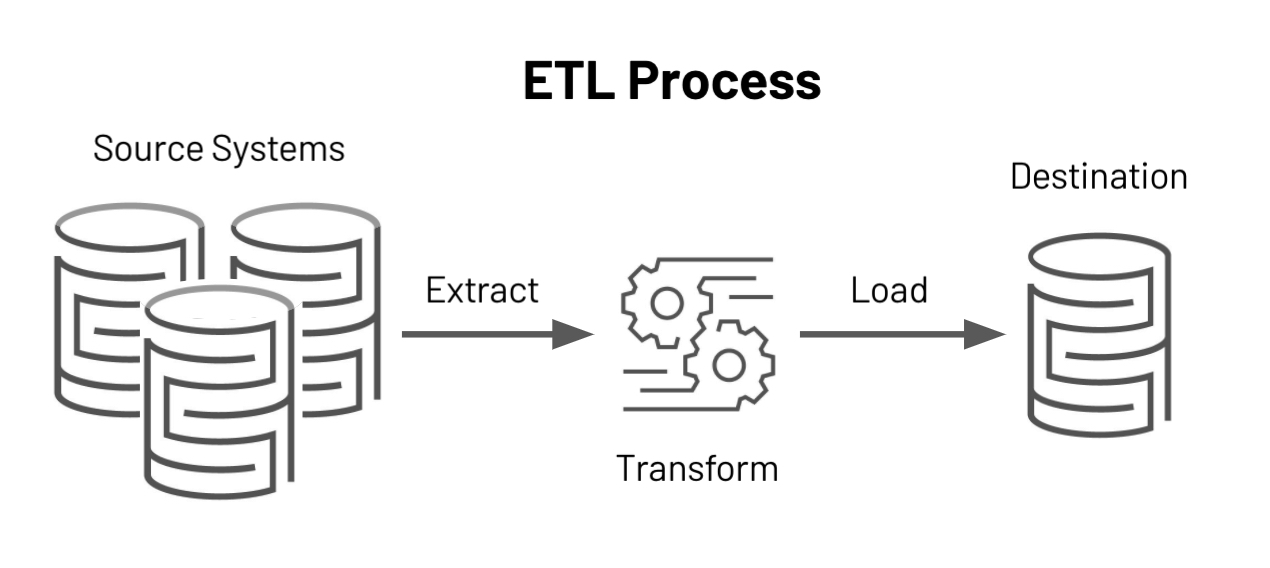
What is ETL? Extract Transform Load.



ETL, which stands for **extract, transform, and load**, is the process data engineers use to extract data from different sources, transform the data into a usable and trusted resource, and load that data into the systems end-users can access and use downstream to solve business problems.

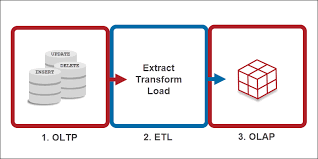
**Source (Jio sales, iPhone sales) -Extract (copy to other db)--> Transform(Aggregation, Normalization) -> Load→ Destination(Data Warehouse).**

**OLTP vs OLAP**

**Online transaction processing (OLTP Database)** captures, stores, and processes data from transactions in real time. [**Using software, you store the real time date**] [Need to fast]

||

ETL - [**The copied data from OLTP is done with ETL, store in data warehouse**

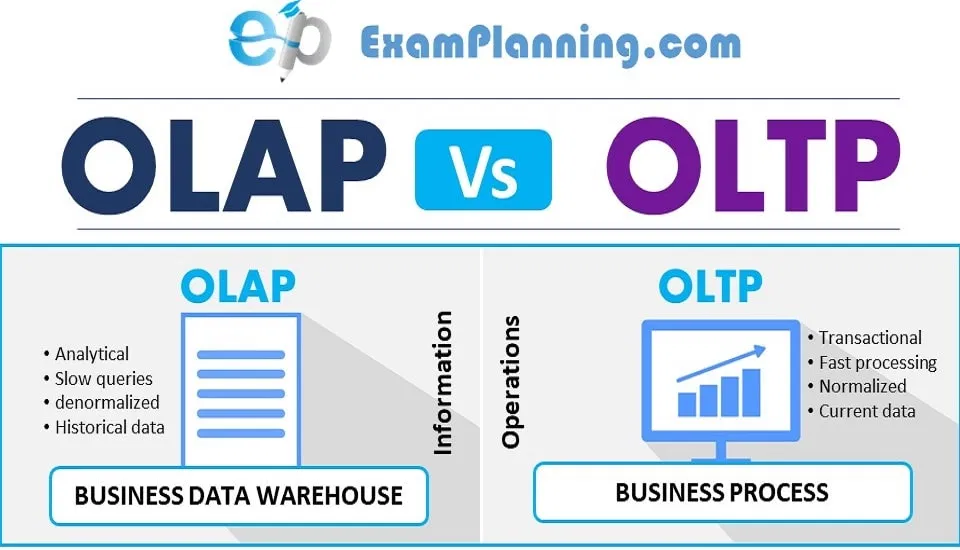
****

||

**Online analytical processing (OLAP Database)** uses complex queries to analyze aggregated **historical data** from OLTP systems.**[ Analytics and Reporting by querying**] [The final data (Data Marts/OLTP cubes is used for analysis].

**The act of performing queries on data warehouse/data cubes is called OLAP.** OLAP is a flexible way for you to make complicated analysis of multidimensional data.

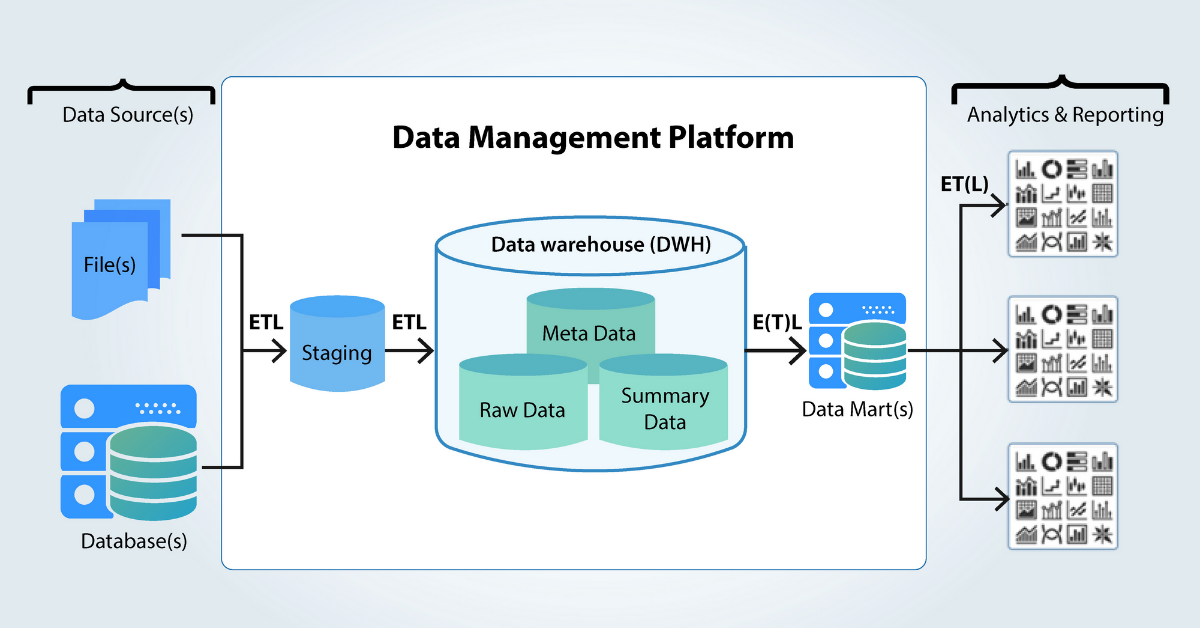
OLTP can’t handle complex data like history of last 5 years sale data. So OLAP is introduced.



**Data Warehouse**

Data warehouse is a database designed for analytical needs(the data is basically ready for analysis). It functions on the basis of OLAP. It is a central location where consolidated data from multiple locations are stored.

**Data Warehousing**: Data warehousing is the act of organizing and storing data in a way to make its **retrieval efficient and insightful.** The data from data ware should have some useful insights or should serve some benefits.



**Metadata- Is the data about the data.**

**Data mart** is used for more **security**, It is a small version of data warehouse like containing data of particular domains, tables. To give access for third party people or operation team like that.

**Types of OLAP CUBES**

**MOLAP: Multidimensional OLAP :** Molap is a form of olap that process and stores the data directly into a multidimensional database.

[A - Excellent performance.

D - Only limited data can be handled.]

**ROLAP : Relational OLAP:** Rolap is a form of olap that perform dynamic multidimensional analysis of data stored in a relational database rather than multidimensional database.

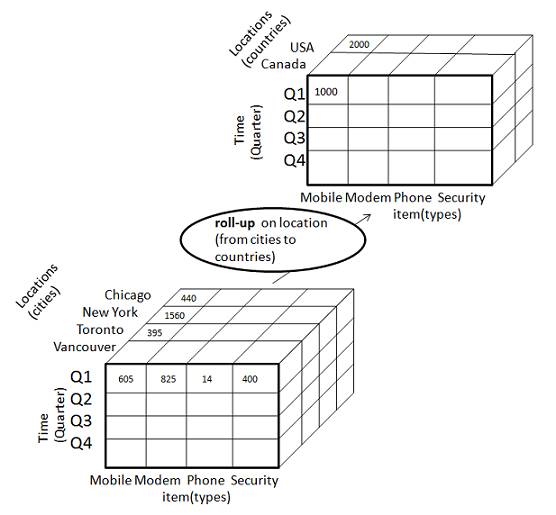
[A - Greater amount of data can be processes.

D - Requires more processing time/disk space.]

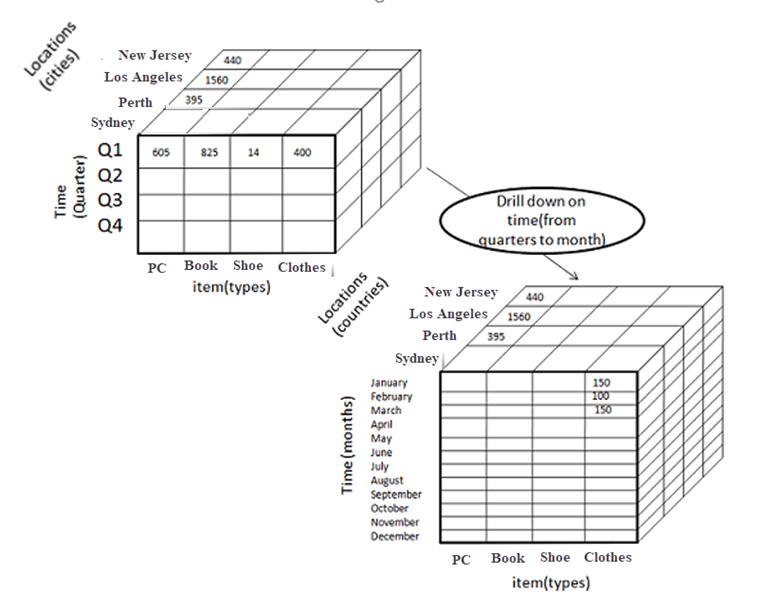
**HOLAP : Hybrid OLAP:** Holap is a combination of the advantages of Molap and Rolap.

**OLAP OPERATIONS**

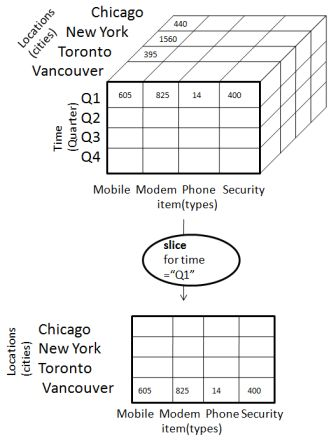
**Roll-Up :** Roll-up perform aggregation on data cube by either climbing up concept hierarchy for a dimension or dimension reduction.



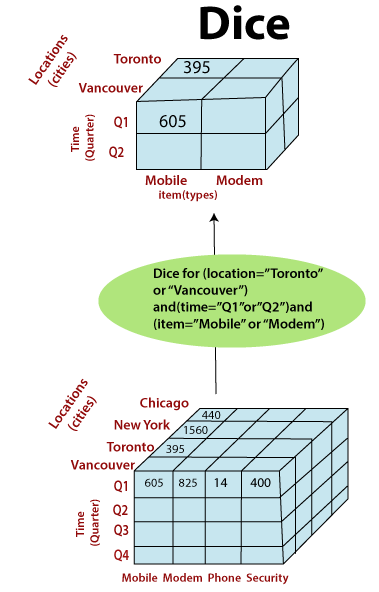
**Drill-down-** Reverse of roll up. Stepping down a concept hierarchy for dimension.

****

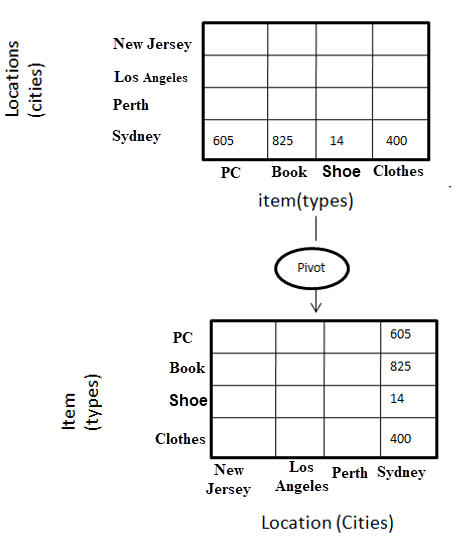
**Slice:** The slice operation provides a new sub cube from one particular dimension in a given cube.



**Dice:** The dice operation provides a new sub cube from two or more dimension in a given cube.



**Pivot:** The pivot operation is also known as rotation operation. It transposes the axes in order to provide an alternative presentation of data.



**DIMENSIONS**

**What is meant by dimensional table?**

**A dimension table or dimension entity is a table or entity in a star, snowflake, or starflake schema that stores details about the facts. For example, a Time dimension table stores the various aspects of time such as year, quarter, month, and day.**

The tables that describe the dimensions involved are called dimension tables in data warehouse.

Dividing a data warehouse project into dimensions provides structured information for analysis and reporting.



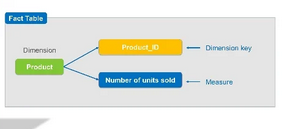
**1st row, subject, 2nd row - dimension, 3rd row - attributes.**

**FACTS AND MEASURES**

A fact is a measure that can be summed, averaged or manipulated.

A fact table contain 2 kinds of data - **dimension key,** and **measure.**

Every dimension table is linked to fact table.



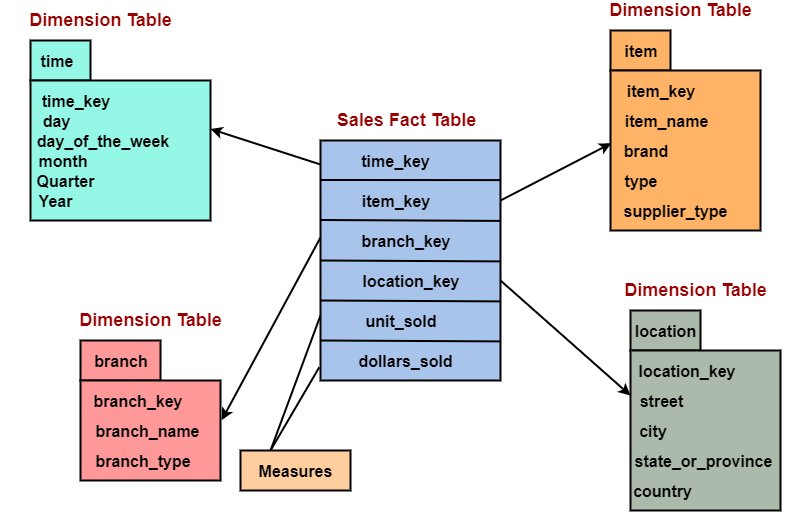
**SCHEMAS**

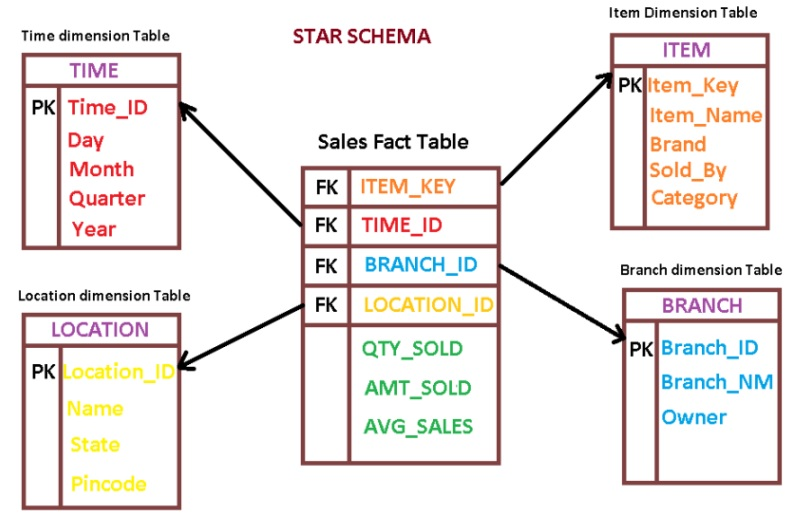
A schema gives the logical description of the entire database.

It gives details about the constraints placed on the tables, key values present and how the key values are linked between the different tables.

A database uses relational model, while a data warehouse uses **STAR, SNOWFLAKE, AND FAST CONSTELLATION** schema.

**START Schema** (foreign keys are in fact table and primary key in dimension tables)

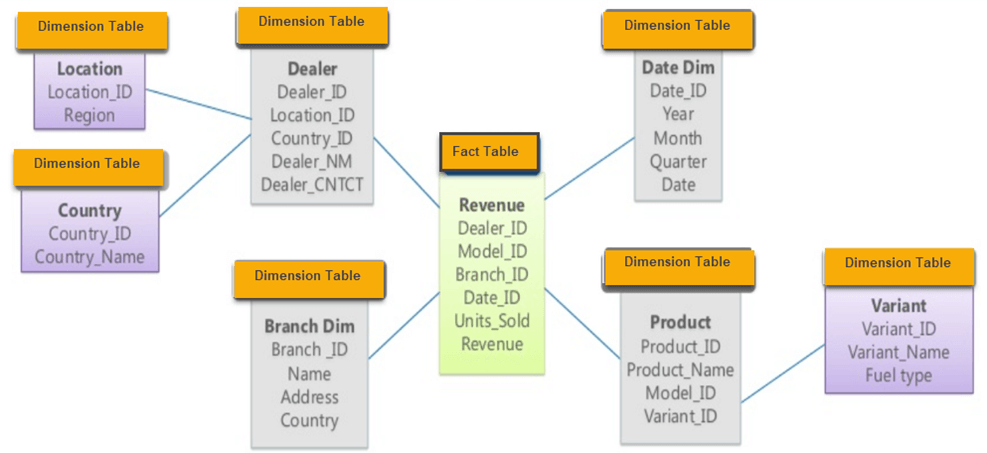




SNOWFLAKE schema

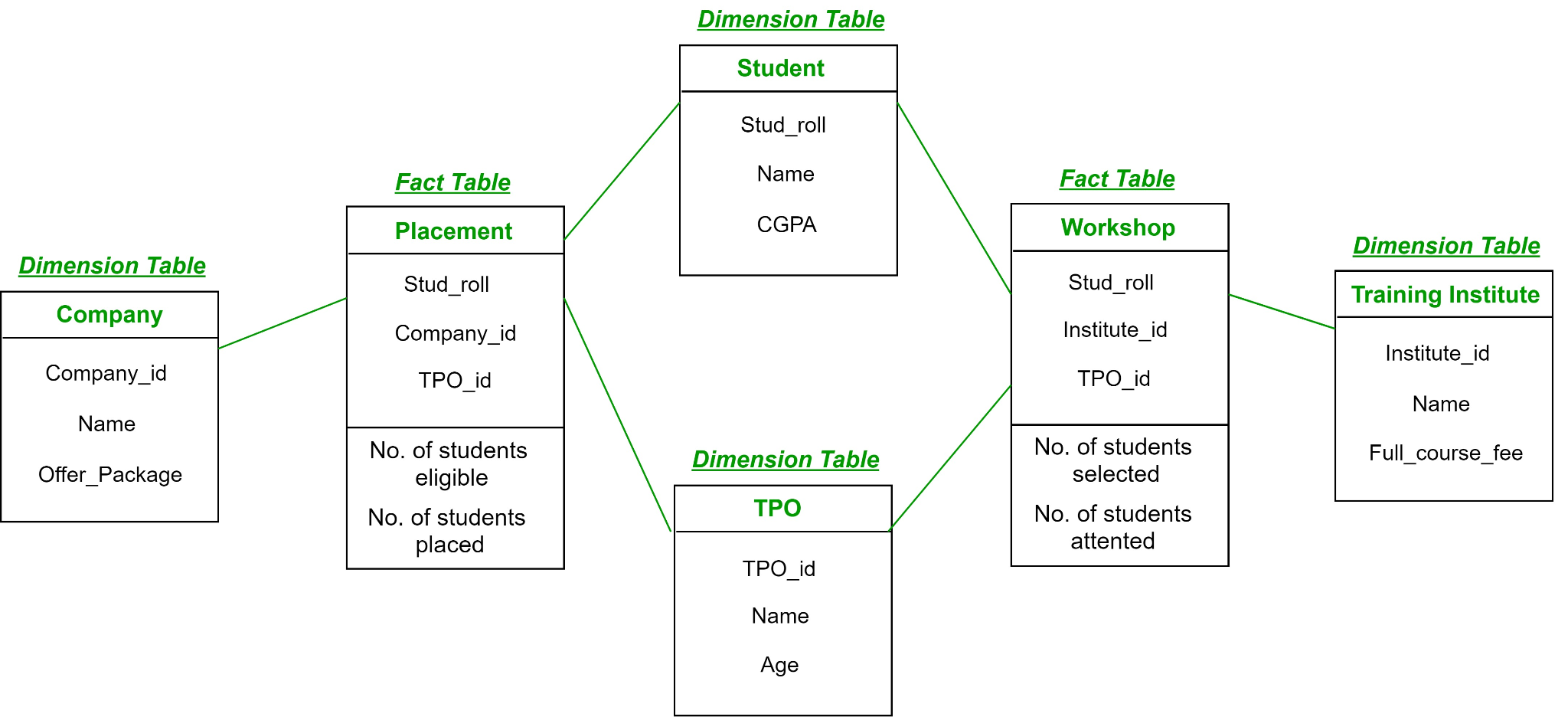
A snowflake schema is a logical arrangement of tables in a multidimensional database that matches the snowflake shape. A Snowflake Schema is an enlarged Star Schema with additional dimensions. After the dimension tables have been normalized, the data is separated into new tables.

Dimension tables in the snowflake schema are normalized(split into additional tables).



**FAST CONSTELLATION/Galaxy** schema

It has more than 1 fact table.

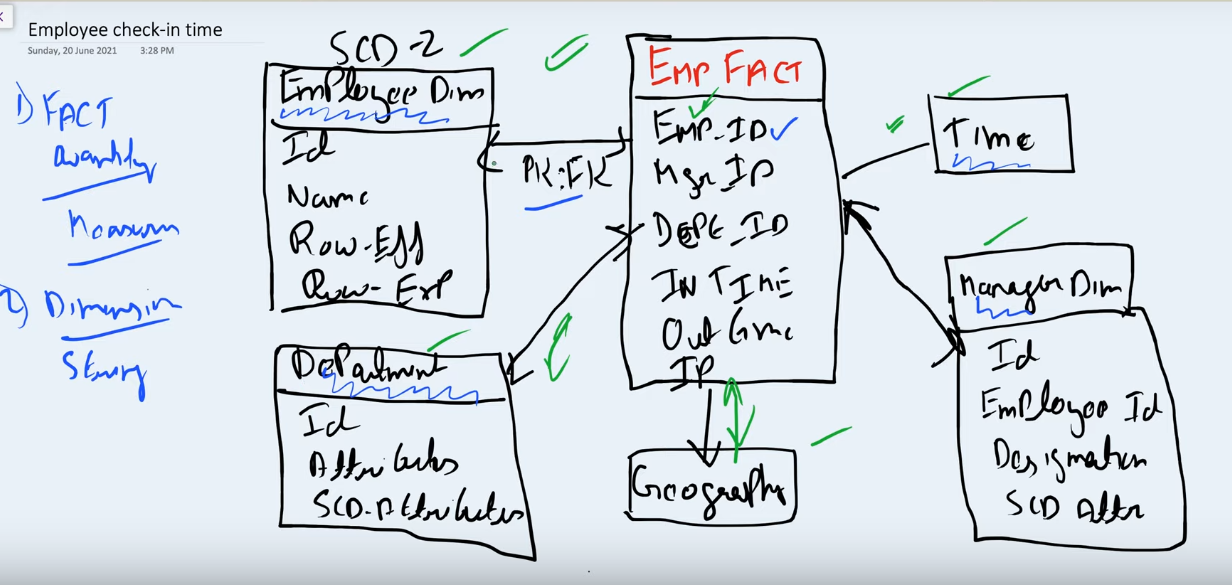


==================================================================

**AWS Redshift** - Amazon Redshift is a fully managed, petabyte-scale data warehouse service in the cloud. Used to perform business analytics. Redshift uses a unique data warehouse architecture called **Massively parallel processing**.

**Data Modeling in Data Ware House**

**Question - Design a data model for employee time spend each day in a office.**

****

**The dimension tables are →** Emploee Dim, Department Dim, Manager Dim, Geographical Dim, Time Dim.

**The Fact Table -** Emp Fact.

**The Metrics are -**

Department wise check in check out.

Geographical wise.

**Drill downs are:** Monthly, Daily, Overall.

**SCD2 dimension table-- is a dimension that stores and manages current and historical data over time in a data warehouse. The purpose of an SCD2 is to preserve the history of changes.**

=====================================================================

DataWarehouse Technologies

**AWS RedShift:**

Amazon Redshift is a cloud-based fully managed petabytes-scale data warehouse By the Amazon Company.

**Google BigQuery:**

BigQuery is a serverless data warehouse that allows scalable analysis over petabytes of data. It’s a Platform as a Service that supports querying with the help of ANSI SQL.

**Snowflake:**

Snowflake is a cloud computing-based data warehousing built on top of the Amazon Web Services or Microsoft Azure cloud infrastructure.

==================================================================

**DBT Tool**

**dbt** is a data transformation *tool* that enables data analysts and engineers to transform, test and document data in the cloud data warehouse.

Dbt only perform transformation, It doesn’t do extraction and loading.

====================================================================

**Data Modelling**

Using Postgres SQL, you can create data models.