Shri Vile Parle Kelavani Mandal's

DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING

(Autonomous College Affiliated to the University of Mumbai)

NAAC Accredited with "A" Grade (CGPA : 3.18)

A.Y. 2022-2023

Name: Siddharth Unny

SAP ID.: 60004200080

Subject: Data Mining and Warehousing Laboratory

Batch: A4

Division: A

Branch: Computer Engineering

LAB EXPERIMENT NO.: 2

Aim:

To build a Data Warehouse for a given problem statement by performing the following:

1. **Making information package diagram**
2. **Design dimensional data model i.e., Star schema, Snowflake schema and Fact Constellation schema (if applicable)**

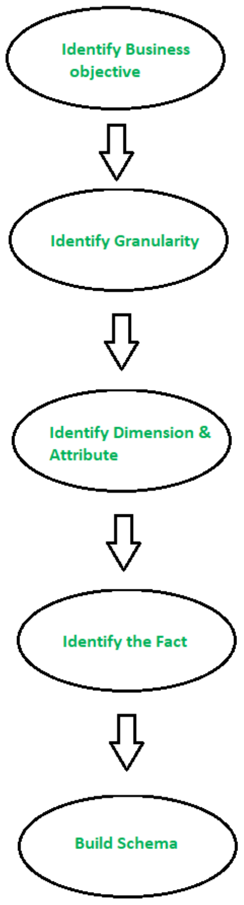
Theory:

Dimensional Data Modelling:

Dimensional Data Modelling is one of the data modelling techniques used in data warehouse design.

The main goal is to improve the data retrieval.

The concept of Dimensional Modelling was developed by Ralph Kimball which is comprised of facts and dimension tables. Since the main goal of this modelling is to improve the data retrieval so it is optimized for SELECT OPERATION. The advantage of using this model is that we can store data in such a way that it is easier to store and retrieve the data once stored in a data warehouse. Dimensional model is the data model used by many OLAP systems.



The figure shown depicts the various steps for dimensional data modelling.

These steps are as follows:

Step 1 - Identifying the business objective:
The first step is to identify the business objective. Sales, HR, Marketing, etc. are some examples as per the need of the organization. Since it is the most important step of Data Modelling the selection of business objective also depends on the quality of data available for that process.

Step 2 - Identifying Granularity:
Granularity is the lowest level of information stored in the table. The level of detail for business problem and its solution is described by Grain.

Step 3 - Identifying Dimensions and their attributes:
Dimensions are objects or things. Dimensions categorize and describe data warehouse facts and measures in a way that supports meaningful answers to business questions. A data warehouse organizes descriptive attributes as columns in dimension tables. For Example, the data dimension may contain data like a year, month and weekday.

Step 4 - Identifying the Fact:

The measurable data is held by the fact table. Most of the fact table rows are numerical values like price or cost per unit, etc.

Step 5 - Building of Schema:

We implement the Dimension Model in this step. A schema is a database structure. There are two popular schemes: Star Schema and Snowflake Schema.

Information Package Diagram (IP Diagram):

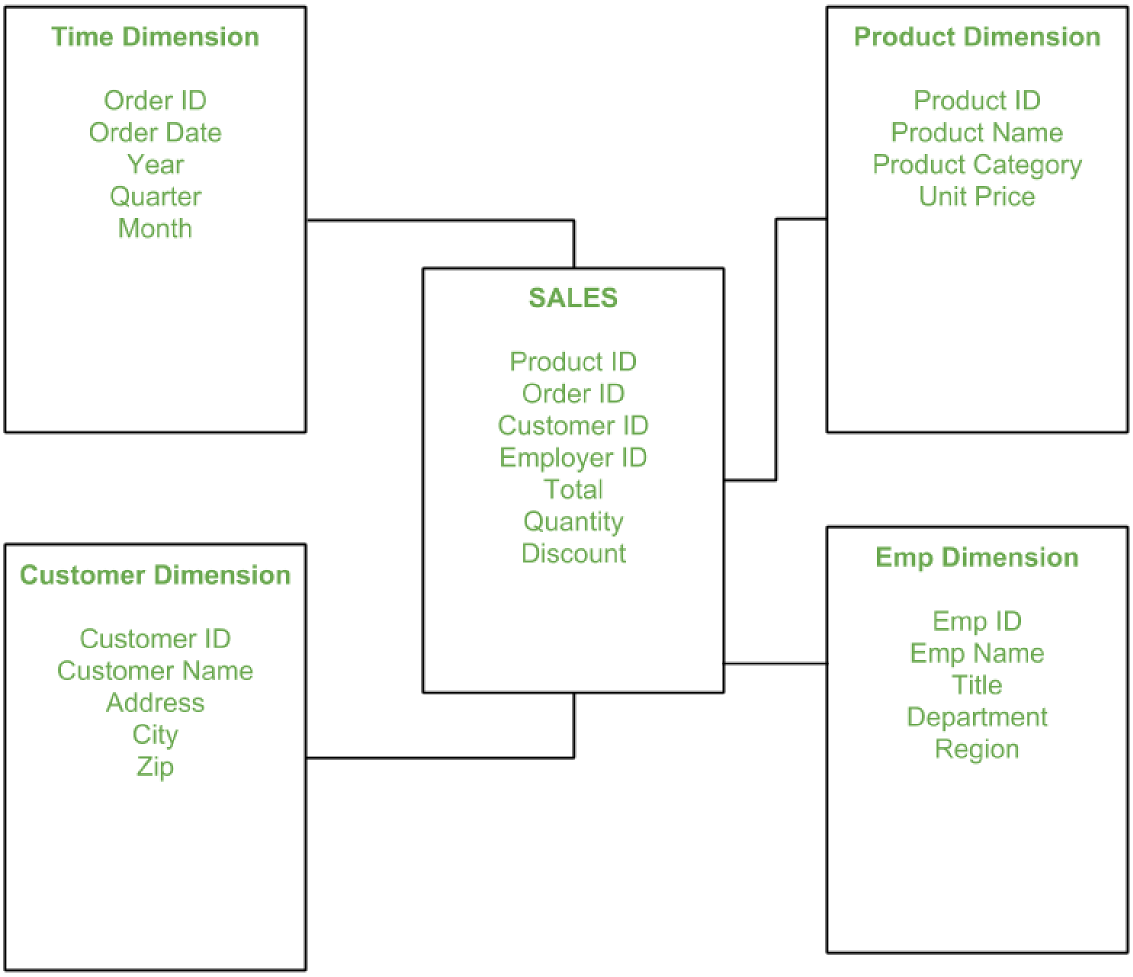
The presence of information package diagrams in the requirements definition document is the major and significant difference between operational systems and data warehouse systems. Remember that information package diagrams are the best approach for determining requirements for a data warehouse.

The information package diagrams crystallize the information requirements for the data warehouse. They contain the critical metrics measuring the performance of the business units, the business dimensions along which the metrics are analysed, and the details of how drill-down and roll-up analyses are done.

Star Schema:

Star schema is the fundamental schema among the data mart schema, and it is simplest. This schema is widely used to develop or build a data warehouse and dimensional data marts. It includes one or more fact tables indexing any number of dimensional tables. The star schema is a necessary cause of the snowflake schema. It is also efficient for handling basic queries.

It is said to be star as its physical model resembles to the star shape having a fact table at its centre and the dimension tables at its peripheral representing the star’s points. Below is an example to demonstrate the Star Schema:



Model of Star Schema:

In Star Schema, Business process data, that holds the quantitative data about a business is distributed in fact tables, and dimensions which are descriptive characteristics related to fact data. Sales price, sale quantity, distant, speed, weight, and weight measurements are few examples of fact data in star schema.

Often, A Star Schema having multiple dimensions is termed as Centipede Schema. It is easy to handle a star schema which have dimensions of few attributes.

Advantages:

1. Simpler Queries
   Join logic of star schema is quite cinch in comparison to other join logic which are needed to fetch data from a transactional schema that is highly normalized.
2. Simplified Business Reporting Logic:
   In comparison to a transactional schema that is highly normalized, the star schema makes simpler common business reporting logic, such as of reporting and period-over-period.

3. Feeding Cubes:
   Star schema is widely used by all OLAP systems to design OLAP cubes efficiently. In fact, major OLAP systems deliver a ROLAP mode of operation which can use a star schema as a source without designing a cube structure.

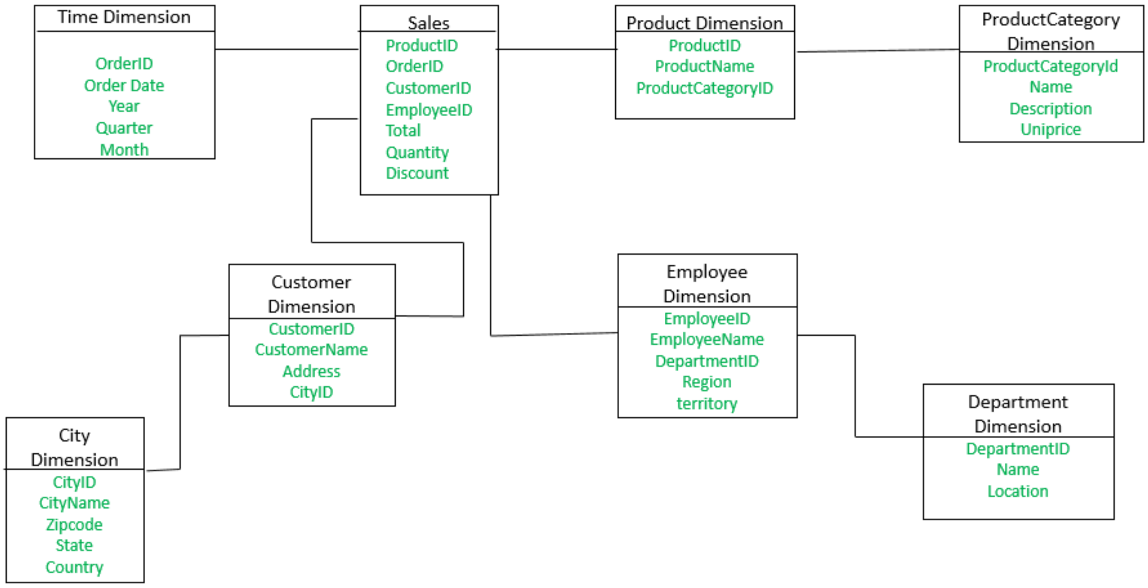
Disadvantages:

1. Data integrity is not enforced well since in a highly de-normalized schema state.
2. Not flexible in terms if analytical needs as a normalized data model.
3. Star schemas don’t reinforce many-to-many relationships within business entities – at least not frequently.

Snowflake Schema:

The snowflake schema is a variant of the star schema. Here, the centralized fact table is connected to multiple dimensions. In the snowflake schema, dimensions are present in a normalized form in multiple related tables. The snowflake structure materialized when the dimensions of a star schema are detailed and highly structured, having several levels of relationship, and the child tables have multiple parent tables. The snowflake effect affects only the dimension tables and does not affect the fact tables.

Example:



The main difference between star schema and snowflake schema is that the dimension table of the snowflake schema is maintained in the normalized form to reduce redundancy. The advantage here is that such tables (normalized) are easy to maintain and save storage space. However, it also means that more joins will be needed to execute the query. This will adversely impact system performance.

Snowflaking:

The snowflake design is the result of further expansion and normalized of the dimension table. In other words, a dimension table is said to be snowflaked if the low-cardinality attribute of the dimensions has been divided into separate normalized tables. These tables are then joined to the original dimension table with referential constraints (foreign key constrain).

Generally, snowflaking is not recommended in the dimension table, as it hampers the understandability and performance of the dimension model as more tables would be required to be joined to satisfy the queries.

Characteristics:

The dimension model of a snowflake under the following conditions:

1. The snowflake schema uses small disk space.
2. It is easy to implement dimension that is added to the schema.
3. There are multiple tables, so performance is reduced.
4. The dimension table consists of two or more sets of attributes that define information at different grains.
5. The sets of attributes of the same dimension table are being populated by different source systems.

Advantages:

There are two main advantages of snowflake schema given below:

1. It provides structured data which reduces the problem of data integrity.
2. It uses small disk space because data are highly structured.

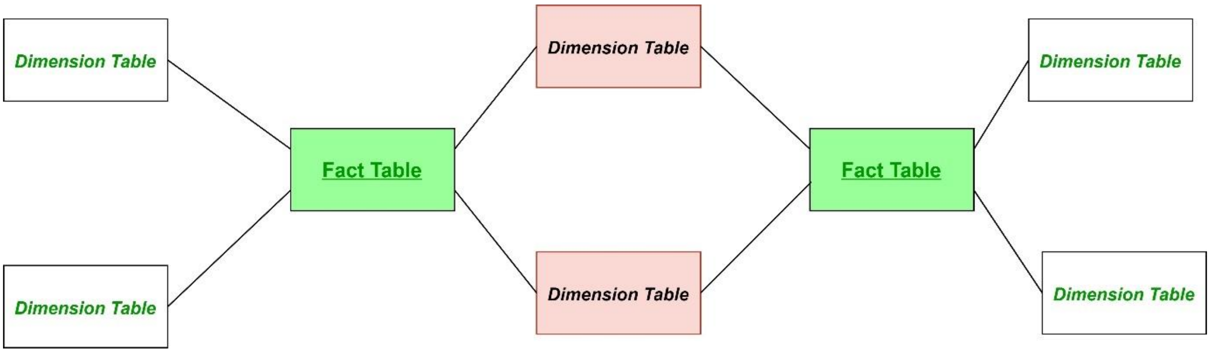
Disadvantages:

1. Snowflaking reduces space consumed by dimension tables but compared with the entire data warehouse the saving is usually insignificant.
2. Avoid snowflaking or normalization of a dimension table, unless required and appropriate.
3. Do not snowflake hierarchies of one dimension table into separate tables. Hierarchies should belong to the dimension table only and should never be snowflakes.
4. Multiple hierarchies that can belong to the same dimension have been designed at the lowest possible detail.

Fact Constellation Schema:

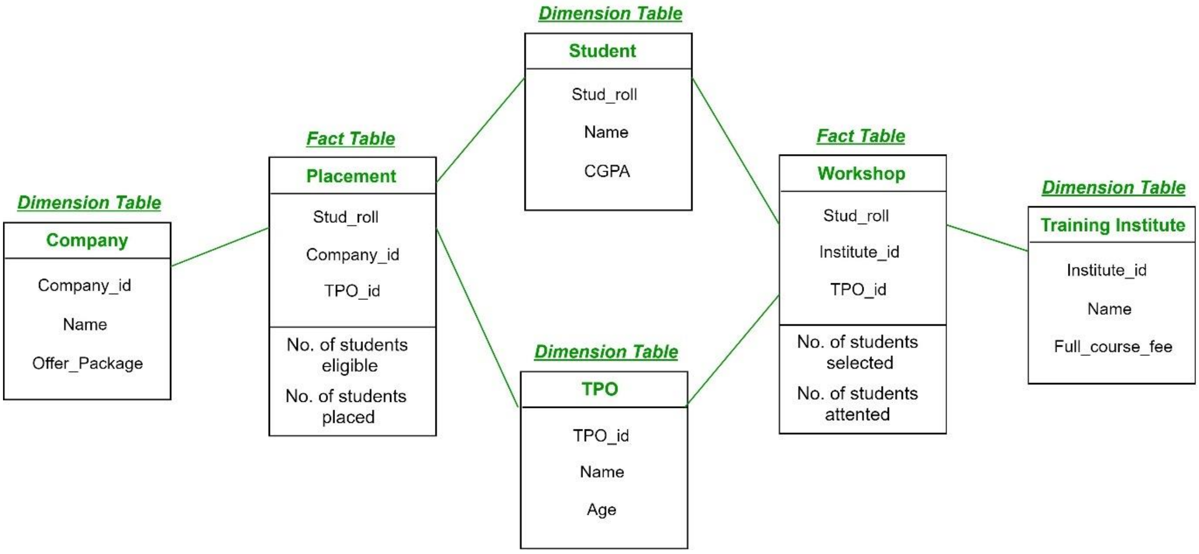
Fact Constellation is a schema for representing multidimensional model. It is a collection of multiple fact tables having some common dimension tables. It can be viewed as a collection of several star schemas and hence, also known as Galaxy schema. It is one of the widely used schema for Data warehouse designing and it is much more complex than star and snowflake schema. For complex systems, we require fact constellations.

General Structure is as follows:



Here, the pink coloured Dimension tables are the common ones among both the star schemas. Green coloured fact tables are the fact tables of their respective star schemas.

Example:



<u>Advantages</u>:

1. Provides a flexible schema.

<u>Disadvantages</u>:

1. It is much more complex and hence, hard to implement and maintain.

Factless Fact Tables:

Factless facts are those fact tables that have no measures associated with the transaction. Factless facts are a simple collection of dimensional keys which define the transactions or describing condition for the time period of the fact. Factless fact tables are important dimensional data structures use to convey transactional information which contain no measures. These tables are occasionally necessary for capturing important dimensional relationships which are critical to the meeting the defined business reporting requirements.

Snowflake vs Star Schema:

Snowflake schema dimension tables are normalised, whereas star schema dimension tables are not. Snowflake schemas require less storage space for dimension tables but are more complex. Star schemas only join the fact and dimension tables, resulting in simpler, faster SQL queries. Snowflake schemas are easier to maintain because they contain no redundant data. Snowflake schemas are appropriate for data warehouses, whereas star schemas are appropriate for data-marts.

Snowflake vs Fact Constellation Schema:

In reality, a constellation schema is a type of multidimensional model. The dimension tables in the fact constellation schema are shared by several fact tables. At any given time, a constellation schema may contain more than one star schema. The planetarium schema, unlike the snowflake schema, is more difficult to use because it has multiple numbers between tables. In contrast to the snowflake schema, the constellation schema uses extremely complex queries to access data from the database.

Snowflake vs Factless Schema

Factless fact tables are only used to establish relationships between different dimensions of elements. Tables are also useful for describing events and coverage, implying that nothing has happened. It is frequently used to represent many-to-many relationships. Thus, unlike snowflake schema we cannot mention any additional facts other than an abbreviated key.

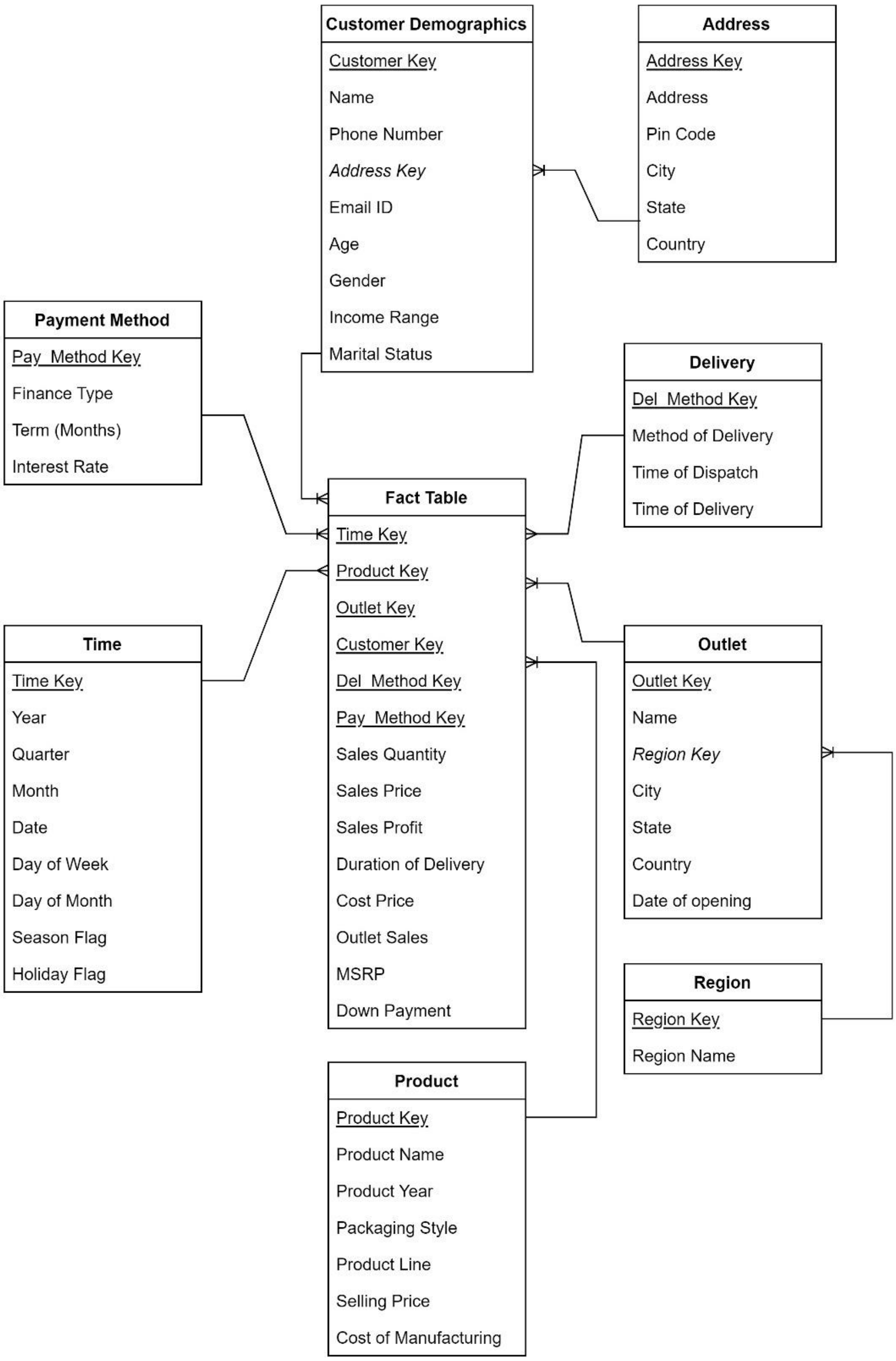
Information Package Diagram:

Time	Product	Outlet	Customer Demographics	Delivery	Payment Method
Time ID	Product ID	Outlet ID	Customer ID	Del_Method ID	Pay_Method ID
Year	Product Name	Name	Name	Method of Delivery	Finance Type
Quarter	Product Year	Region	Phone Number	Time of Dispatch	Term (Months)
Month	Packaging Style	City	Address	Time of Delivery	Interest Rate
Date	Product Line	State	Email ID		
Day of Week	Selling Price	Country	Age		
Day of Month	Cost of Manufacturing	Date of Opening	Gender		
Season Flag			Income Range		
Holiday Flag			Marital Status		
Facts: Sales Quantity, Sales Price, Sales Profit, Duration of delivery, Cost price, Outlet Sales, MSRP, Down Payment					

We have chosen IKEA warehouse as our topic of study.

IKEA, home furnishings retailer that was the world's largest seller of furniture in the early 21st century, operating more than 300 stores around the world. Due to a large customer base, IKEA will need a comprehensive data warehouse to meet the analytical needs of the company. We have designed an information package that contains dimensions of Time, Product, outlet, Customer demographics, delivery and payment method. These dimensions deal with customer interactions and will help the company in making critical decisions such as locations of new stores, availability of in demand products, addition of new payment methods, ensure quick delivery, etc.

Snowflake Schema:



For designing our schema, we used snowflake schema as it covers all the needed details in our dimension tables. In a snowflake schema, we break down individual dimension tables into logical subdimensions. This makes the data model more complex, but it can be easier for analysts to work with, especially for certain data types. Hence here, we have further created subdimensions of Address and Region fields as creating a sub dimension of address field helped in making the customer demographic more concise and reduces redundancy. Most of the time each city has IKEA only one IKEA store hence creating subdimension of region field reduces the region count, thus normalizing the data.

Conclusion:

In this experiment we demonstrated various steps of dimensional modelling. We constructed the Information Package Diagram and Snowflake Schema for IKEA warehouse, which was our chosen topic of study. We first observed the processes executed in an IKEA warehouse, and after understanding the requirements of the management team of IKEA, we formulated and IP diagram with appropriate facts and levels of granularity. Next, we converted this IP diagram to a snowflake schema, since there was a need to snowflake the Address and Region fields in the Customer Demographics and Outlet dimension tables respectively. In this way, we could represent all the identified facts effectively. Thus, we can conclude that dimensional data modelling is an efficient way to identify and represent information required to build a data warehouse.