EXPERIMENT 6

Subject: DMW
Group Members:
Aryan Parekh - 60004190013
Junaid Girkar – 60004190057
Kanaad Deshpande – 60004190058
Manish Jha - 60004190066

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

Theory
Information Package Diagram
The information package diagram is a novel idea for determining and recording information requirements for a data warehouse. This concept helps us to give a concrete form to the various insights, nebulous thoughts, and opinions expressed during the process of collecting requirements. The information packages, put together while collecting requirements, are very useful for taking the development of the data warehouse to the next phases.
An information package diagram defines the relationships between subject matter and key performance measures. The information package diagram has a highly targeted purpose, providing a focused scope for user requirements. Because information package diagrams target what the users want, they are effective in facilitating communication between the technical staff and the users, indicating any inconsistencies between the requirements and what the data warehouse will deliver.
Facts
Facts and dimensions are data warehousing terms. A fact is a quantitative piece of information - such as a sale or a download. Facts are stored in fact tables, and have a foreign key relationship with a number of dimension tables.
Where multiple fact tables are used, these are arranged as a fact constellation schema. A fact table typically has two types of columns: those that contain facts and those that are a foreign key to dimension tables. The primary key of a fact table is usually a composite key that is made up of all of its foreign keys. Fact tables contain the content of the data warehouse and store different types of measures like additive, non-additive, and semi additive measures.
Fact tables provide the (usually) additive values that act as independent variables by which dimensional attributes are analyzed. Fact tables are often defined by their grain. The grain of a fact table represents the most atomic level by which the facts may be defined. The grain of a sales fact table might be stated as "sales volume by day by product by store". Each record in this fact table is therefore uniquely defined by a day, product and store. Other dimensions might be members of this fact table (such as location/region) but these add nothing to the uniqueness of the fact records. These "affiliate dimensions" allow for additional slices of the independent facts but generally provide insights at a higher level of aggregation (a region contains many stores).
Dimensions
Dimensions are companions to facts, and describe the objects in a fact table. A dimension is thus, a structure that categorizes facts and measures in order to enable users to answer business questions. Commonly used dimensions are people, products, place and time.

In a data warehouse, dimensions provide structured labelling information to otherwise unordered numeric measures. The dimension is a data set composed of individual, non-overlapping data elements. The primary functions of dimensions are threefold: to provide filtering, grouping and labelling.

Star Schema

Star Schema in data warehouse, in which the center of the star can have one fact table and a number of associated dimension tables. It is known as star schema as its structure resembles a star. The Star Schema data model is the simplest type of Data Warehouse schema. It is also known as Star Join Schema and is optimized for querying large data sets.

Characteristics of Star Schema:

- Every dimension in a star schema is represented with the only one-dimension table.
- The dimension table should contain the set of attributes.
- The dimension table is joined to the fact table using a foreign key
- The dimension table are not joined to each other
- Fact table would contain key and measure
- The Star schema is easy to understand and provides optimal disk usage.
- The dimension tables are not normalized. For instance, in the above figure, Country_ID does not have Country lookup table as an OLTP design would have.
- The schema is widely supported by BI Tools

Advantages of Star Schema

- 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.
- In comparison to a transactional schema that is highly normalized, the star schema makes simpler common business reporting logic, such as as-of reporting and period-over-period.
- 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 of Star Schema –

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

Snowflake Schema

Snowflake Schema in data warehouse is a logical arrangement of tables in a multidimensional database such that the ER diagram resembles a snowflake shape. A Snowflake Schema is an extension of a Star Schema, and it adds additional dimensions. The dimension tables are normalized which splits data into additional tables.

Characteristics of Snowflake Schema

- The main benefit of the snowflake schema it uses smaller disk space.
- Easier to implement a dimension is added to the Schema
- Due to multiple tables query performance is reduced
- The primary challenge that you will face while using the snowflake Schema is that you need to perform more maintenance efforts because of the more lookup tables.

Advantages of snowflake schema

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

Disadvantages of snowflake schema

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

Fact Constellation Schema/Galaxy Schema

Fact Constellation Schema is a schema that represents a multidimensional model of tables. This schema is a group of different fact tables that have few similar dimensional tables. It can be represented as a group of multiple star schemas and thus, it is also called a Galaxy schema. Fact schema is the most frequently used schema to design a Data warehouse and also, it is a little complicated than star and snowflake schema model.

Fact constellation schema is a tool of analytical processing via online, which has a huge group of the number of fact tables that share dimensional tables, also aggregated as a group of stars. We can also call it as an extension of the star constellation model.

Characteristics of Fact Constellation Schema

- A fact constellation schema can have multiple fact tables associated with it.
- It is commonly used a schema for designing data warehouses and it is much complicated than any other schema such as star and snowflake schema.
- We can create a fact constellation schema from a star schema by splitting them into one or more, star schemas.
- It is also said that a fact constellation schema can have many fact tables and a shared dimensional table.

Advantages of Fact Constellation Schema

- Tables are subdivided into fact and dimensional to understand the relationship between them.
- It is a flexible schema that makes users use it.
- Here dimensional tables are shared by the number of fact tables.
- It is a normalized form of snowflake and star schema.
- We can access the data in the database using complex queries.

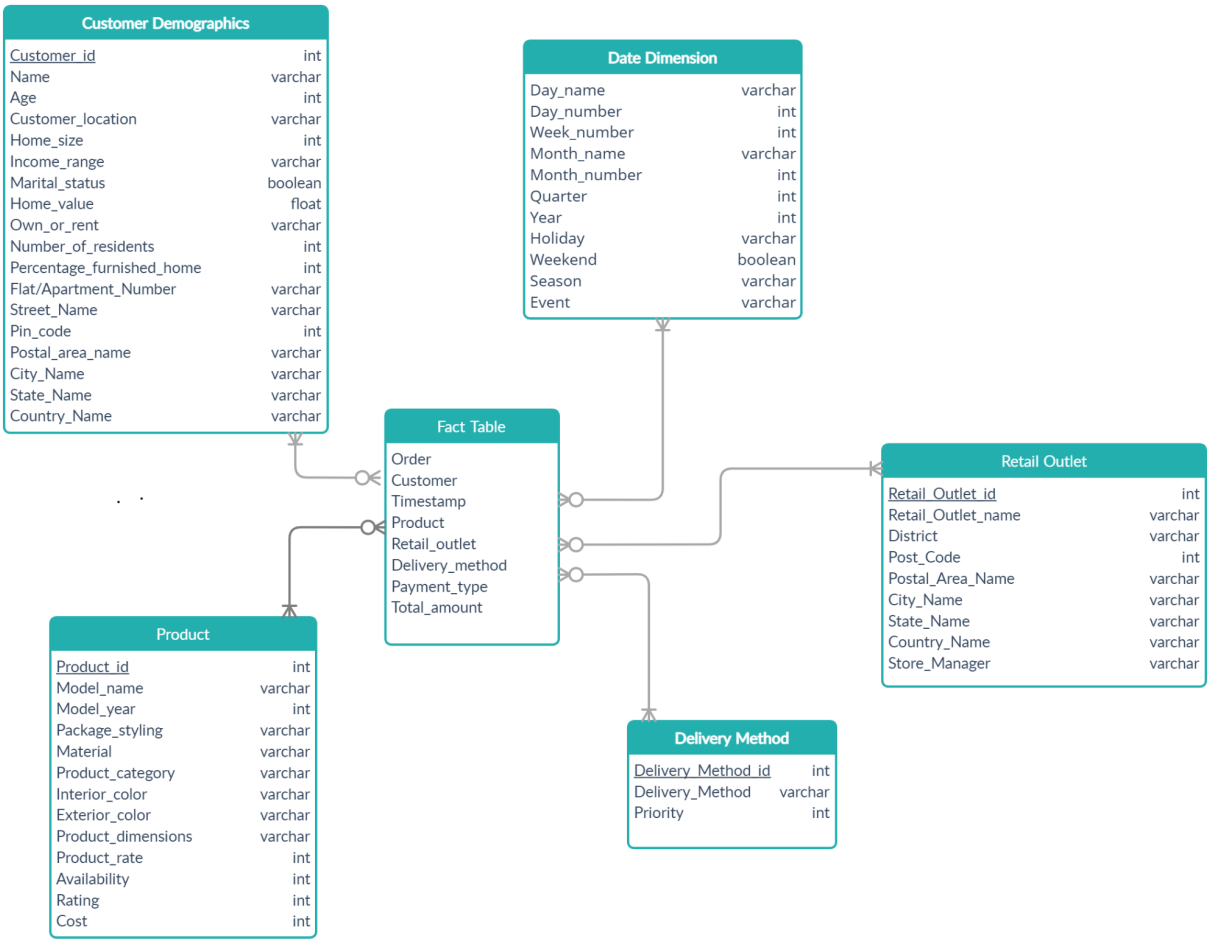
Disadvantages of Fact Constellation Schema

- It is difficult to understand as it is a very complex schema to implement.
- It uses more space in the database comparative to the star schema.
- It has many joins between dimensional and fact tables and thus it is difficult to understand.
- This is difficult to maintain and operate.

1. INFORMATION PACKAGE DIAGRAM

Time	Product	Customer Demographics	Retail Outlet	Delivery Method
Day_name	Model_name	Customer ID	Retail_Outlet_id	delivery_method_id
Day_number	Model_year	Name	Retail_Outlet_name	delivery_method
Week_number	Package_styling	Age	District	priority
Month_name	Material	Customer_Location	Post_Code	
Month_number	Product_category	Home_Size	Postel_Area_Name	
Quarter	Interior_color	Income_Range	City_Name	
Year	Exterior_color	Maritial_status	State_Name	
Holiday	Product_dimension	Home_value	Country_Name	
Weekend	Product_weight	Own_or_Rent	Store Manager	
Season	Availability	Number_of_residents		
Event	Rating	Percentage_furnished_home		
	Cost	Flat/Apartment_Number		
		Street_Name		
		Pin_Code		
		Postal_area_name		
		City_Name		
		State_Name		
		Country_Name		
Facts - Order, customer, timestamp, product, retail outlet, delivery method, payment type, total amount				

2. STAR SCHEMA


Customer Demographics
Customer_id int
Name varchar
Age int
Customer_location varchar
Home_size int
Income_range varchar
Marital_status boolean
Home_value float
Own_or_rent varchar
Number_of_residents int
Percentage_furnished_home int
Flat/Apartment_Number varchar
Street_Name varchar
Pin_code int
Postal_area_name varchar
City_Name varchar
State_Name varchar
Country_Name varchar
Date Dimension
Day_name varchar
Day_number int
Week_number int
Month_name varchar
Month_number int
Quarter int
Year int
Holiday varchar
Weekend boolean
Season varchar
Event varchar
Fact Table
Order
Customer
Timestamp
Product
Retail_outlet
Delivery_method
Payment_type
Total_amount
Retail Outlet
Retail_Outlet_id int
Retail_Outlet_name varchar
District varchar
Post_Code int
Postal_Area_Name varchar
City_Name varchar
State_Name varchar
Country_Name varchar
Store_Manager varchar
Product
Product_id int
Model_name varchar
Model_year int
Package_styling varchar
Material varchar
Product_category varchar
Interior_color varchar
Exterior_color varchar
Product_dimensions varchar
Product_rate int
Availability int
Rating int
Cost int
Delivery Method
Delivery_Method_id int
Delivery_Method varchar
Priority int

Conclusion

In this experiment, we constructed an information package diagram and the star schema for an IKEA warehouse. The information package diagram helped create a system to record information of everything necessary to the IKEA warehouse and the star schema helped visualize all the different relationships in the information package diagram. We learned about various data warehousing terms such as facts. Facts help us analyze dimensional attributes. We also learned about the various schema structures available such as star schema, snowflake schema and constellation schema.