ASSIGNMENT NO: 2

Group Members:

Name	GR.No & Roll No.
Vikram Shinde	21920048 (322065)
Pranav Wagh	21810340 (322061)
Samruddhi Desai	21920035 (322063)
Avdhut Sagare	21920143 (322073)

Problem Statement:

Choose a set of business processes like Sales, Customer Services, Accounting, Production, Marketing processes etc. for any organization and design star, snow flake and fact constellation schema. Also using ETL tool, extract data from various sources and perform transform and load operations on data. (Using Power BI)

Objectives:

- To normalize the data into various dimensions table and declare primary key in each.
- To identify the relationships between the dimension tables.
- Create a fact table, identify the relationships and introduce the foreign key for each.
- Using the above identifications, designing star, snowflake and fact constellation schema.

Theory:

1. **ETL Operations:**

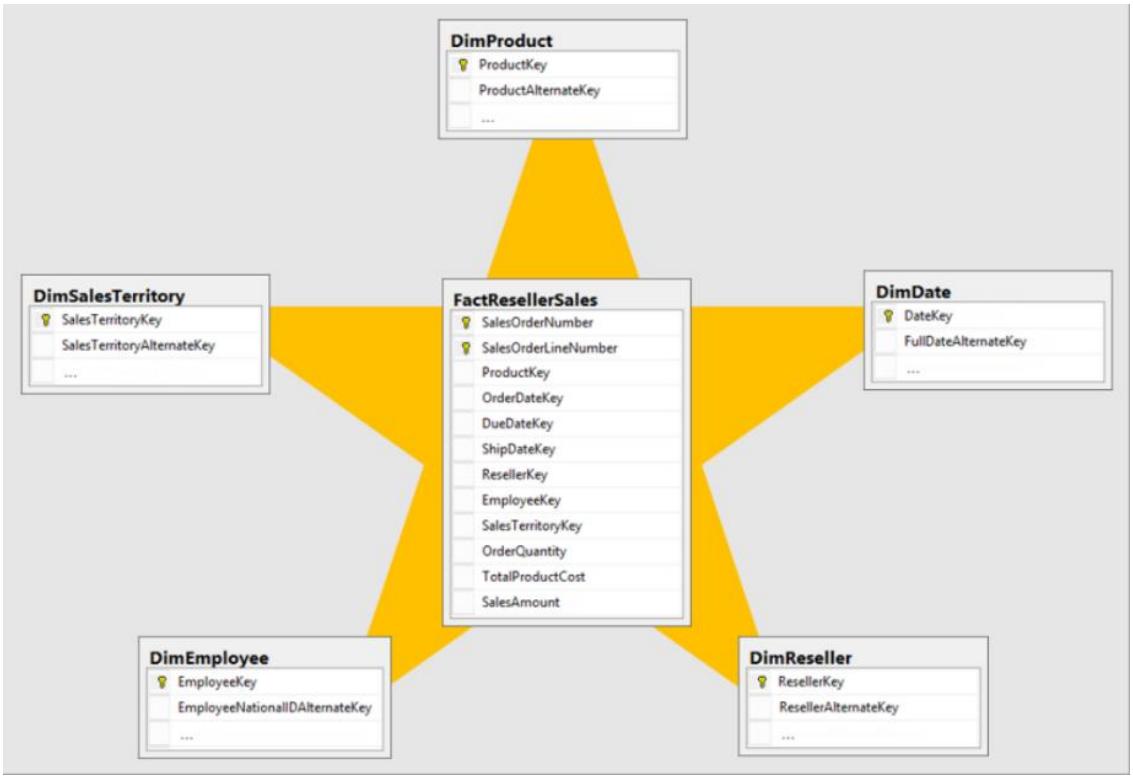
   A. Data Extraction: Get data from multiple, heterogeneous, and external sources.
   B. Data Cleaning: Detect errors in the data and rectify them when possible.
   C. Data Transformation: Convert data from host format to warehouse format.
   D. Load: Sort, summarize, consolidate, compute views, check integrity, build and partitions.
   E. Refresh: Propagate (transmit) the updates from the data sources to the warehouse

2. **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
- The schema is widely supported by BI Tools



3. 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.

Star Schema Vs Snowflake Schema: Key Differences

Star Schema	Snowflake 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 Snowflake schema is represented by centralized fact table which unlikely connected with multiple dimensions.

3.Fact Constellation Schema

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

Characteristics of Fact Constellation Schema:

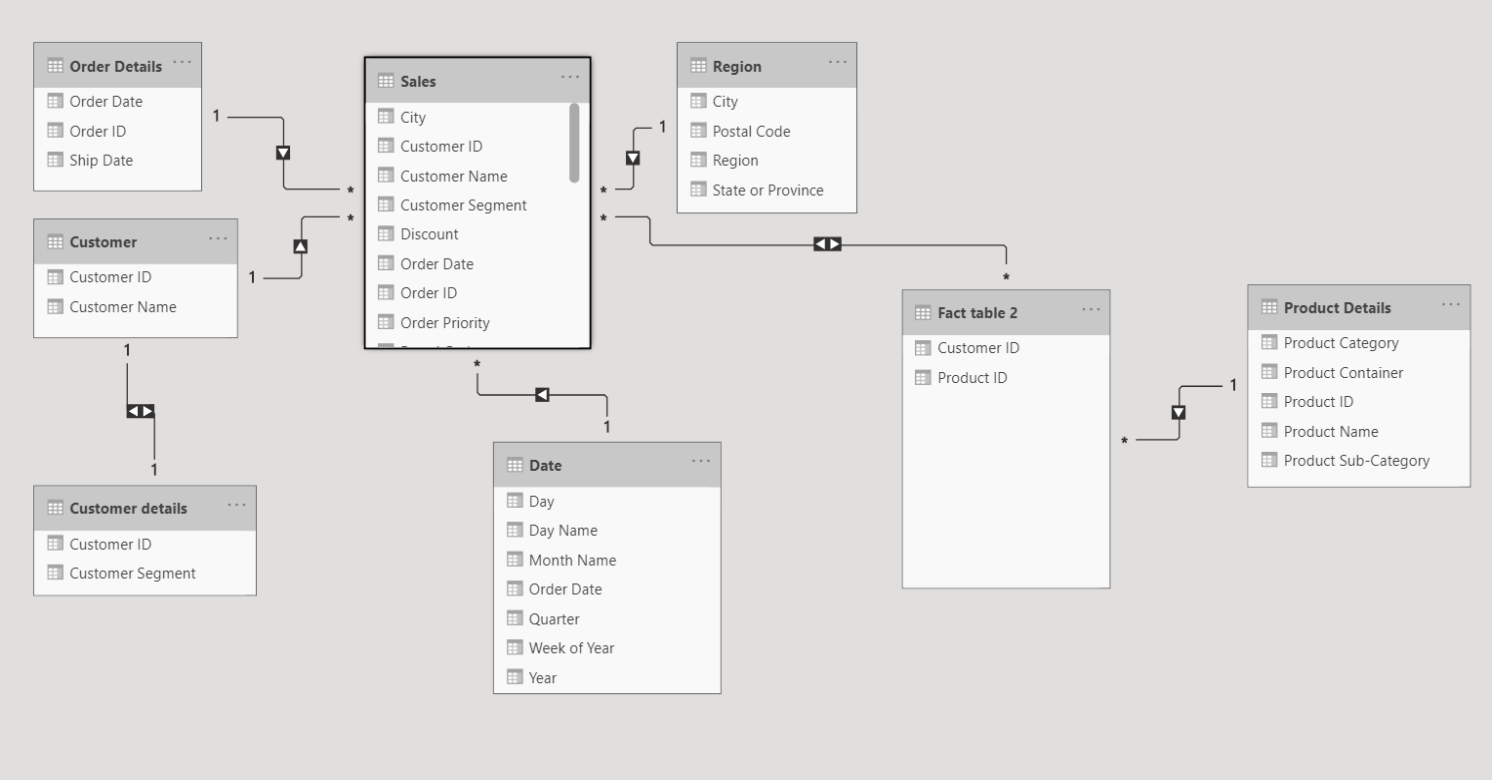
- The dimensions in this schema are separated into separate dimensions based on the various levels of hierarchy.
- For example, if geography has four levels of hierarchy like region, country, state, and city then Galaxy schema should have four dimensions.
- Moreover, it is possible to build this type of schema by splitting the one-star schema into more Star schemes.

- The dimensions are large in this schema which is needed to build based on the levels of hierarchy.
- This schema is helpful for aggregating fact tables for better understanding.

Output:

Dataset used:

Sales Dataset of a company is used. It contains 23 columns and 3000 rows approximately. It can be broadly classified into Dimension tables such as:- Product Details, Customer details, Date, Region, etc.



Inference:

- ➢ “Sales” and “Fact table 2” are the 2 Fact tables constructed.
- ➢ “Order Details”, “Customer”, “Region”, “Date” make up the Dimension tables for Sales Fact table.
- ➢ Thus Star schema is constructed.
- ➢ “Customer Details” is further fetched from Customer Dimension table. Thus Snowflake schema is constructed.
- ➢ Product Details is Dimension table for Fact table 2.
- ➢ Thus Fact Constellation Schema is constructed.