3CS452 : Data Warehousing and Data Mining Lab

Assignment 3 : To build the data warehouse for X-Mart

Problem description :

X-Mart is having different malls in city, where daily sales take place for various products. Higher management is facing an issue while decision making due to non availability of integrated data they can't do study on their data as per their requirement. So objective is to design a system which can help them quickly in decision making and provide Return on Investment (ROI).

Activity :

- **Identify and Collect Requirements**

We need to interview the key decision makers to know, what factors define the success in the business? How does management want to analyze their data? What are the most important business questions, which need to be satisfied by this new system?

We also need to work with persons in different departments to know the data and their common relations if any, document their entire requirement which need to be satisfied by this system.

Let us first identify the requirement from management about their requirements.

Need to see daily, weekly, monthly, quarterly profit of each store.
Comparison of sales and profit on various time periods.
Comparison of sales in various time bands of the day.
Need to know which product has more demand on which location?
Need to study trend of sales by time period of the day over the week, month, and year?
On what day sales is higher?
On every Sunday of this month, what is sales and what is profit?
What is trend of sales on weekday and weekend?
Need to compare weekly, monthly and yearly sales to know growth and KPI

- **Design the Dimensional Model**

We need to design Dimensional Model to suit requirements of users which must address business needs and contains information which can be easily accessible. Design of model should be easily extensible according to future needs. This model design must supports OLAP cubes to provide "instantaneous" query results for analysts.

Let us take a quick look at a few new terms and then we will identify/derive it for our requirement.

Dimension

The dimension is a master table composed of individual, non-overlapping data elements. The primary functions of dimensions are to provide filtering, grouping and labeling on your data. Dimension tables contain textual descriptions about the subjects of the business.

Let me give you a glimpse on different types of dimensions available like confirmed dimension, Role Playing dimension, Degenerated dimension, Junk Dimension.

Slowly changing dimension (SCD) specifies the way using which you are storing values of your dimension which is changing over a time and preserver the history. Different methods / types are available to store history of this change **E.g**. SCD1, SCD2, and SCD3 you can use as per your requirement.

Let us identify dimensions related to the above case study.

Product, Customer, Store, Date, Time, Sales person

Measure

A measure represents a column that contains quantifiable data, usually numeric, that can be

aggregated. A measure is generally mapped to a column in a fact table. For your information, various types of measures are there. **E.g.** Additive, semi additive and Non additive.

Let us define what will be the Measures in our case.

Actual Cost, Total Sales, Quantity, Fact table record count

Fact Table

Data in fact table are called measures (or dependent attributes), Fact table provides statistics for sales broken down by customer, salesperson, product, period and store dimensions. Fact table usually contains historical transactional entries of your live system, it is mainly made up of Foreign key column which references to various dimension and numeric measure values on which aggregation will be performed. Fact tables are of different types, **E.g.** Transactional, Cumulative and Snapshot.

Let us identify what attributes should be there in our Fact Sales Table.

Foreign Key Column

Sales Date key, Sales Time key, Invoice Number, Sales Person ID, Store ID, Customer ID

Measures

Actual Cost, Total Sales, Quantity, Fact table record count

Design the Relational Database

We have done some basic workout to identify dimensions and measures, now we have to use appropriate schema to relate this dimension and Fact tables.

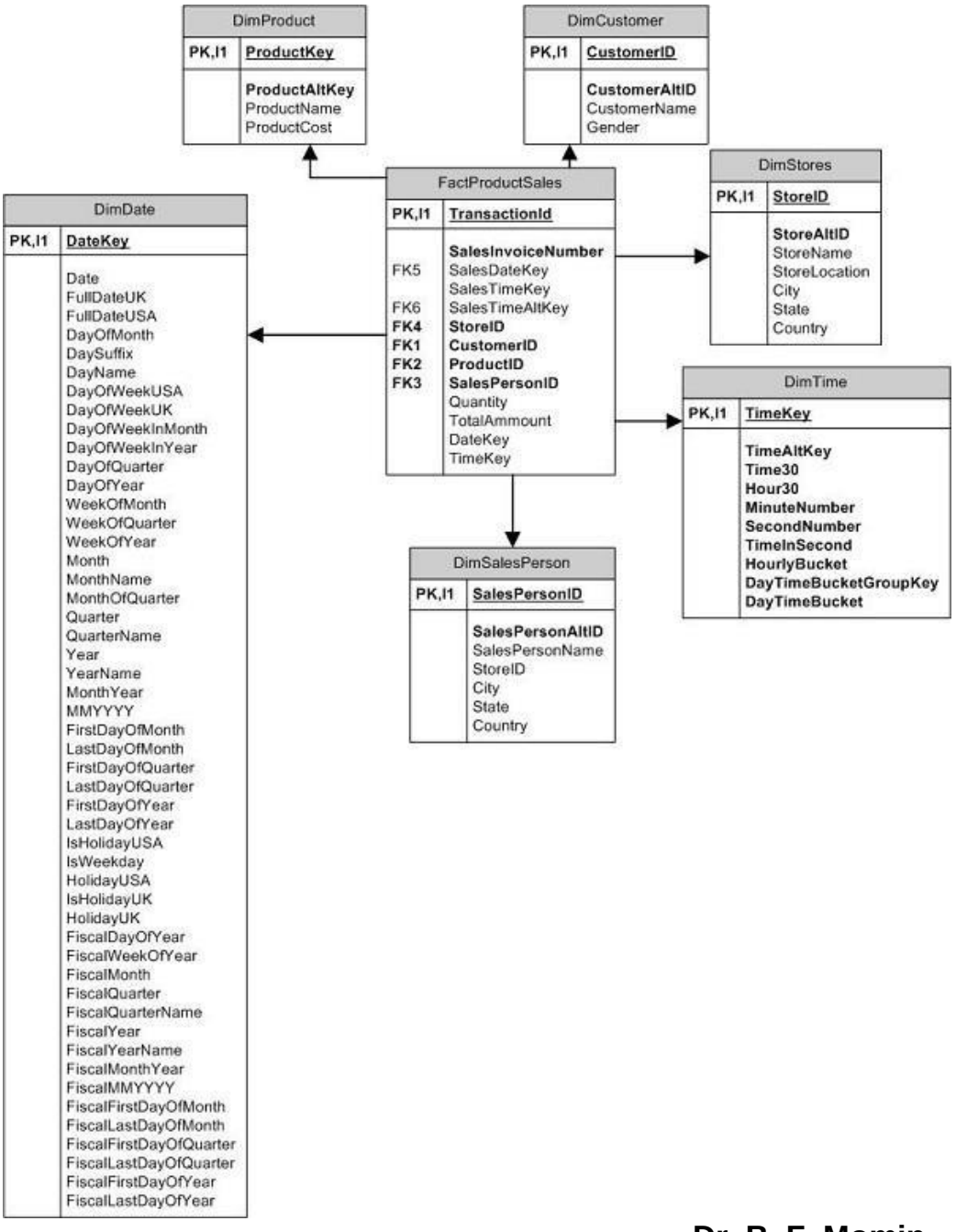
Few popular schemas used to develop dimensional model are as follows:

E.g. Star Schema, Snow Flake Schema, Star Flake Schema, Distributed Star Schema, etc.

In a different article, we will discuss all these schemas, dimension types, measure types, etc., in detail.

Personally, I will first try to use Star schema due to hierarchical attribute model it provides for analysis and speedy performance in querying the data.

Star schema the diagram resembles a star, with points radiating from a center. The center of the star consists of fact table and the points of the star are the dimension tables.



Dr. B. F. Momin
Course Teacher