Welcome Students.

My paper is Data Mining.

This paper is for the 5th semester of BSc Computer Science.

Today, in this session we will be dealing with module number 9.

In this module we are going to cover the topics from Unit 3, Datawarehousing and OLAP, which includes Data warehouse Schema: Star, Snowflake, and Fact Constellation.

The outline for the today's session is :

We will be learning about different Datawarehouse Schema.

The first one is Star Schema.
2nd is Snowflake Schema.
3rd is Fact Constellation Schema.

At the end of this session, the learning outcomes will be:

The student will be able to understand the concept of schema.

Describe the schemas which are used in datawarehouse.

Describe a Star Schema cited with an example.

Describe a Snowflake Schema cited with an example and describe a Fact Constellation Schema cited with an example.

So let's start with the first topic that is (i.e.) Datawarehouse Schema.

The term schema in database refers to the organization of data as a blueprint of how the database is constructed.

It is a logical description of the entire database.

It includes the name and description of records of all the record types,including all associated data items and aggregates.

Like a database, a datawarehouse also requires to maintain a schema.

Schemas in  Data Warehouses are a collection of database objects including tables, views, indexes and synonyms.

Data warehouse environment usually transforms the relational data model or two dimensional model into multidimensional models.

The most popular data model for a data warehouse is a multi dimensional model which can exist in the form of Star Schema, Snowflake schema, or Fact Constellation schema.

So let's learn the first data warehouse schema that is Star Schema.

In Star Schema, each dimension is represented with only one dimension table.

This dimension table contains the set of attributes.

Figure one shows the Star Schema for company Sales.

As you can see on your screen, this is a Star Schema of Sales Data warehouse wherein you have your fact table.

You have your dimension tables.

Sales are considered along 4 dimensions that is time, item, branch and location.

So your fact table that is your central table for your sales will have dimension tables : time, branch, item and location.

Each of these dimension tables will have attributes like for time, it is time_key, day, day_of_the_week, month, quarter, year.

Similarly, for branch dimension table you have branch_key, branch_name, branch_type.

For item dimension table we have item_key, item_name, brand, type, supplier_type and location dimension table have location_key, street, city, province_or_state and country.

Each of these dimension tables are connected to your fact tables via a key that is time_key, item_key, branch_key, location_key and the measures that is dollars_sold and units_sold.

So this schema contains a central fact table for Sales that contains keys to each of the four dimensions along with two measures. That is dollars_sold and units_sold.

There is a fact table that is there at the center.

It contains measures such as dollars_sold and the keys to each of the four related dimension tables.

The links between the fact table at the center and the dimension tables in the extremities form a shape like a star as you can see in this figure.

A fact table typically has two types of columns that is foreign keys to dimension tables and measures those that contain numeric facts.

Dimensions are organized into hierarchies.

For example, you have a dimension table represented as day, day_of_the_week, month, quarters, and year.

Each dimension has only one dimension table, and each table holds a set of attributes.

For example, the location dimension table contains the attribute set {location_key, street, city, province_or_state and country}.

The second data warehouse schema is Snowflake Schema.

Some dimension tables in the Snowflake Schema are normalized.

When we say normalization, it helps us in removing the redundant data.

The normalization splits up the data into additional tables.

Unlike Star schema, the dimension table in the Snowflake schema are normalized.

For example, the location dimension table in the Star schema is normalized and split into dimension table, namely city.

This is a figure which will represent the Snowflake Schema of a Sales data warehouse.

You have a fact table for your Sales and dimension tables that is time, branch, item and location.

Now location dimension table is further split into city dimension table via a city_key that is your foreign key to your location dimension table.

So all this attributes that is location dimension table has attributes that is location_key, street, city_key..

Similarly, city dimension table has attributes city_key, city, province_or_state and country.

Snowflake Schema is a type of star schema, but a more complex model.

Snowflaking is a method of normalizing the dimension tables in a Star schema.

The normalization eliminates redundancy and therefore it becomes easy to maintain and save the storage space.

The result is more complex queries and reduced query performance.

The third data warehouse schema is Fact Constellation Schema.

A fact constellation schema has multiple fact tables.

It is also known as Galaxy Schema.

The following figure shows two fact tables, namely sales and shipping.

This figure represents Fact Constellation schema of sales and shipping data warehouse wherein you can see that there you have two fact tables that is sales and shipping.

You have your dimension tables that is time dimension table, branch dimension table, item dimension table, location dimension table and shipper dimension table which is connected to your fact tables that are shipping and sales via a foreign key.

The Sales fact table is the same as that in the star schema.

The shipping fact table has five dimensions, namely item_key, time_key, shipper_key, from_location, to_location.

The shipping fact table also contains two measures, namely dollars_sold and units_sold.

It is also possible to share dimension tables between fact tables.

For example, time, item and location dimension tables are shared between the sales and shipping fact table.

Here are my references.

Thank you.