Data Mining for Business

Introduction to Data Warehouse

Dr. Shipra Maurya
Department of Management Studies
IIT (ISM) Dhanbad
Email: shipra@iitism.ac.in

Dimensional Data Model

- “Dimensional modeling is a design technique for databases intended to support end-user queries in a data warehouse” – Ralph Kimball – father of dimensional modeling

- Dimensional model is a data structure technique optimized for data warehousing tools.

- It comprises of **fact** and **dimension** tables.

- It arranges data in such a manner that it is easier to retrieve information and generate reports so basically it is optimized for select operations

- Dimensional data model is used by many OLAP systems

Elements of Dimensional Data Model

Fact:

- A fact, also called a measure, is a measurable metric which is described by the dimensions such as the sale amount or order quantity.
- There are usually many more dimensions than facts.
- Example -

Sales_fact
Store_key
Product_key
Salesperson_key
Time_key
Sale_amount
Quantity

Sales amount and quantity are fact

Elements of Dimensional Data Model

Dimension:

- Dimensions describe business events like the sale of a product.
- They are what users would want to sort, group and filter on like dates, customer id, store id, etc.
- Example -

Store_dimension
Store_key
Store name
City
State
Zip_code
Region

Elements of Dimensional Data Model

Attributes:

- The attributes are the various characteristics of the dimension.
- Attributes are used to search, filter or classify facts.
- Dimension tables contain attributes.
- Example -

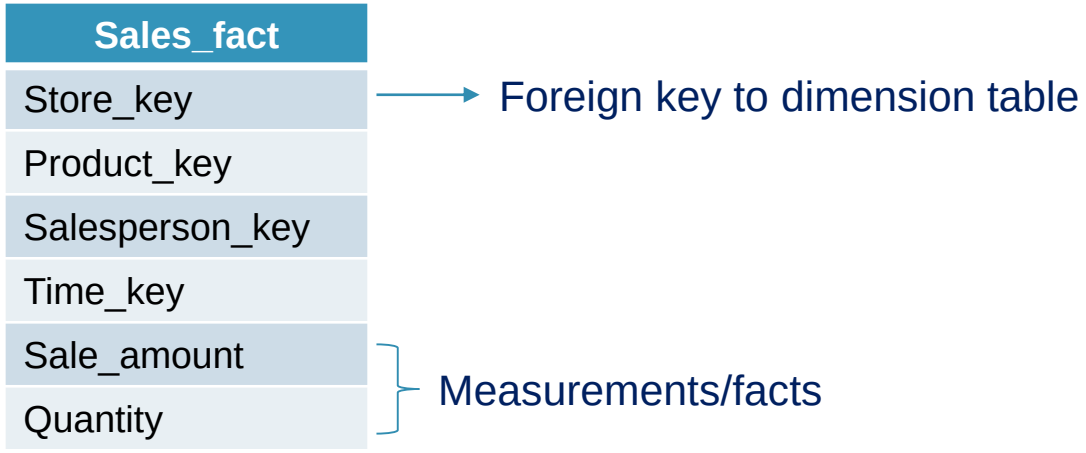
Store_dimension
Store_key
Store name
City
State
Zip_code
Region

Attributes

Elements of Dimensional Data Model

Fact Table:

- A fact table is a primary table in a dimensional model.
- A fact table contains:
  - Observations or events
  - Foreign key to dimension table
- Example -



Elements of Dimensional Data Model

Dimension Table:

- A dimension table describes the observations and event and contains dimensions of a fact
- They are joined to fact table via a foreign key
- Dimension tables are de-normalized tables
- Dimensions offer descriptive characteristics of the facts with the help of their attributes.
- Example -

Store_dimension
Store_key
Store name
City
State
Zip_code
Region

Steps of Dimensional Modeling

1. Identify the business objective
2. Declare/identify the grain (lowest level of detail)
3. Identify the dimensions and its attributes
4. Identify the facts
5. Build the schema – A schema is the database structure (arrangement of tables) – Star schema/Snowflake schema/Galaxy schema

Star Schema

- Each dimension is presented by only one dimension table. The dimension tables are de-normalized and fact table is normalized
- Dimensions are directly linked to the fact table
- Usually the dimension keys are NOT keys from source systems, rather they are generated by the data warehouse load process and they are called surrogate keys (artificial keys)
- Surrogate keys (integer) are used to uniquely identify a row in a dimensional table
- The attributes of the dimension determine the granularity called the grain of the facts, i.e., how detailed are the measures.
- **Most widely used in the industry**

Star Schema Example

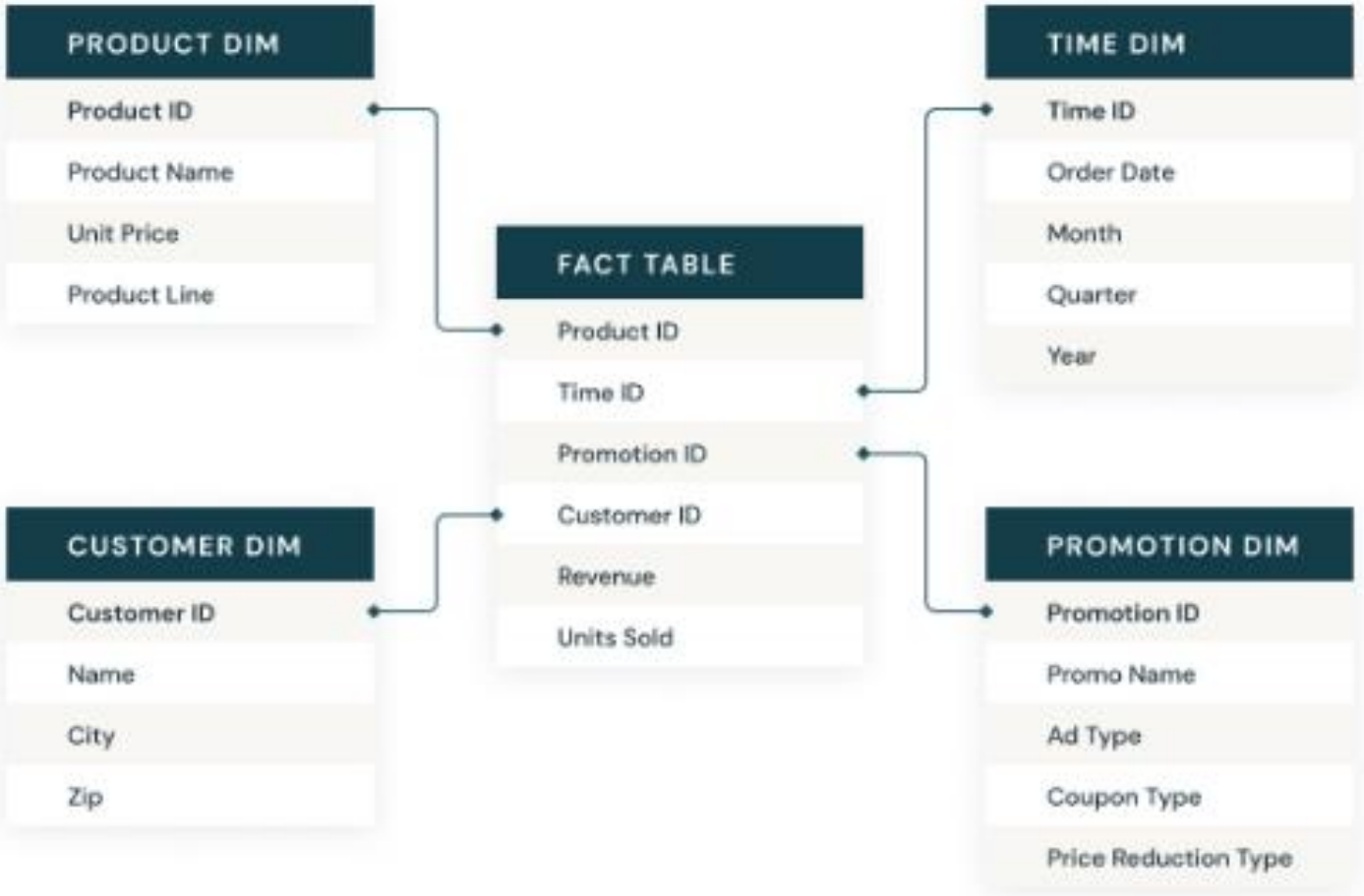


Image Source: Internet

Why Surrogate keys?

- Required to implement history of slowly changing dimensions
- Avoids conflicts among backend application keys
- Insulates the data warehouse from backend application changes
- Different backend applications may use different columns as the dimension key

Snowflake Schema

- It is the variant of Star schema
- Dimensions are present in a normalized form in multiple related tables
- The snowflake effect only affects the dimensional tables and does not affect the fact table
- One dimension table can be connected to another dimension table

Snowflake Schema Example

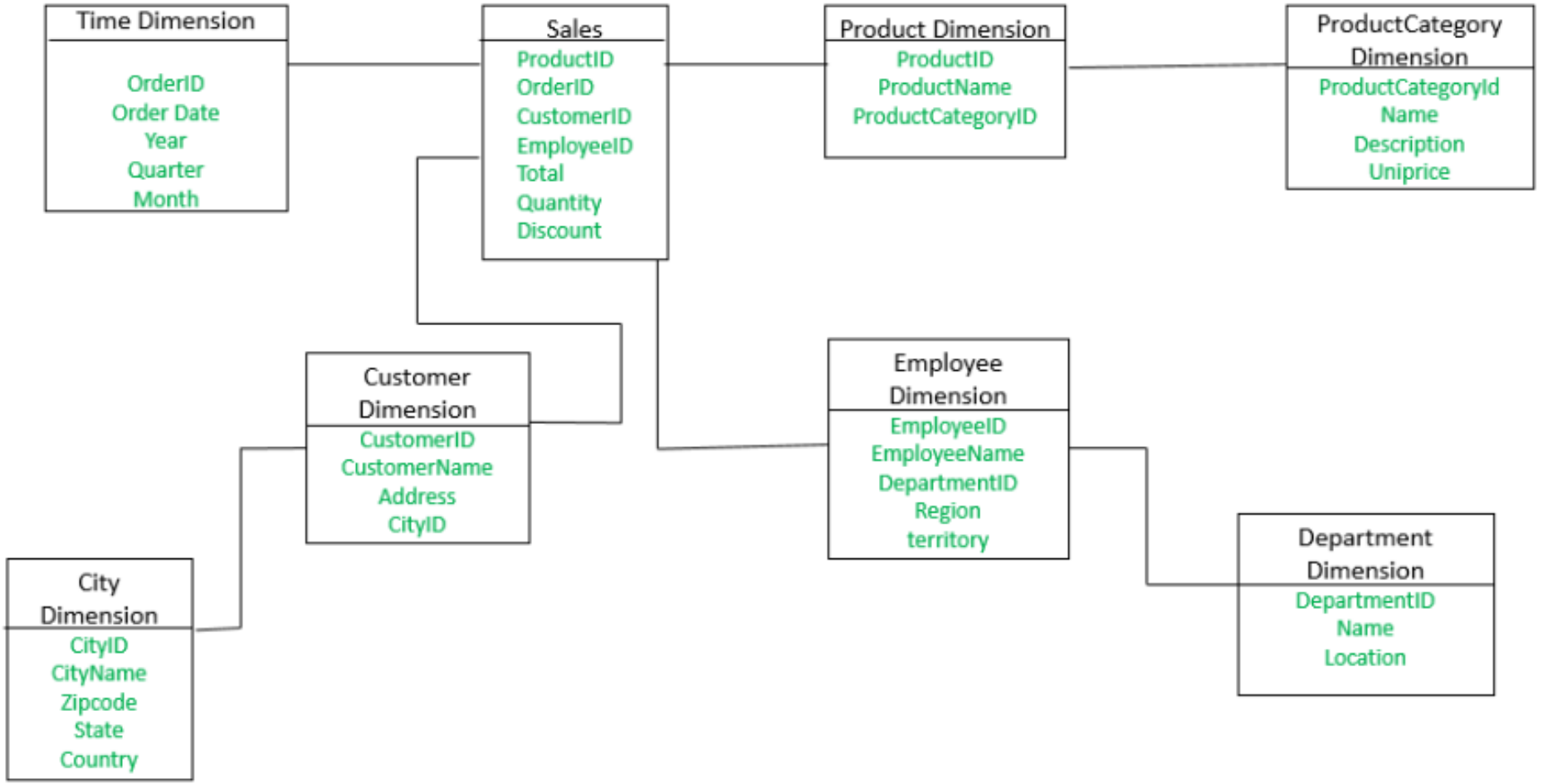


Image Source: Internet

Galaxy (Fact Constellation) Schema

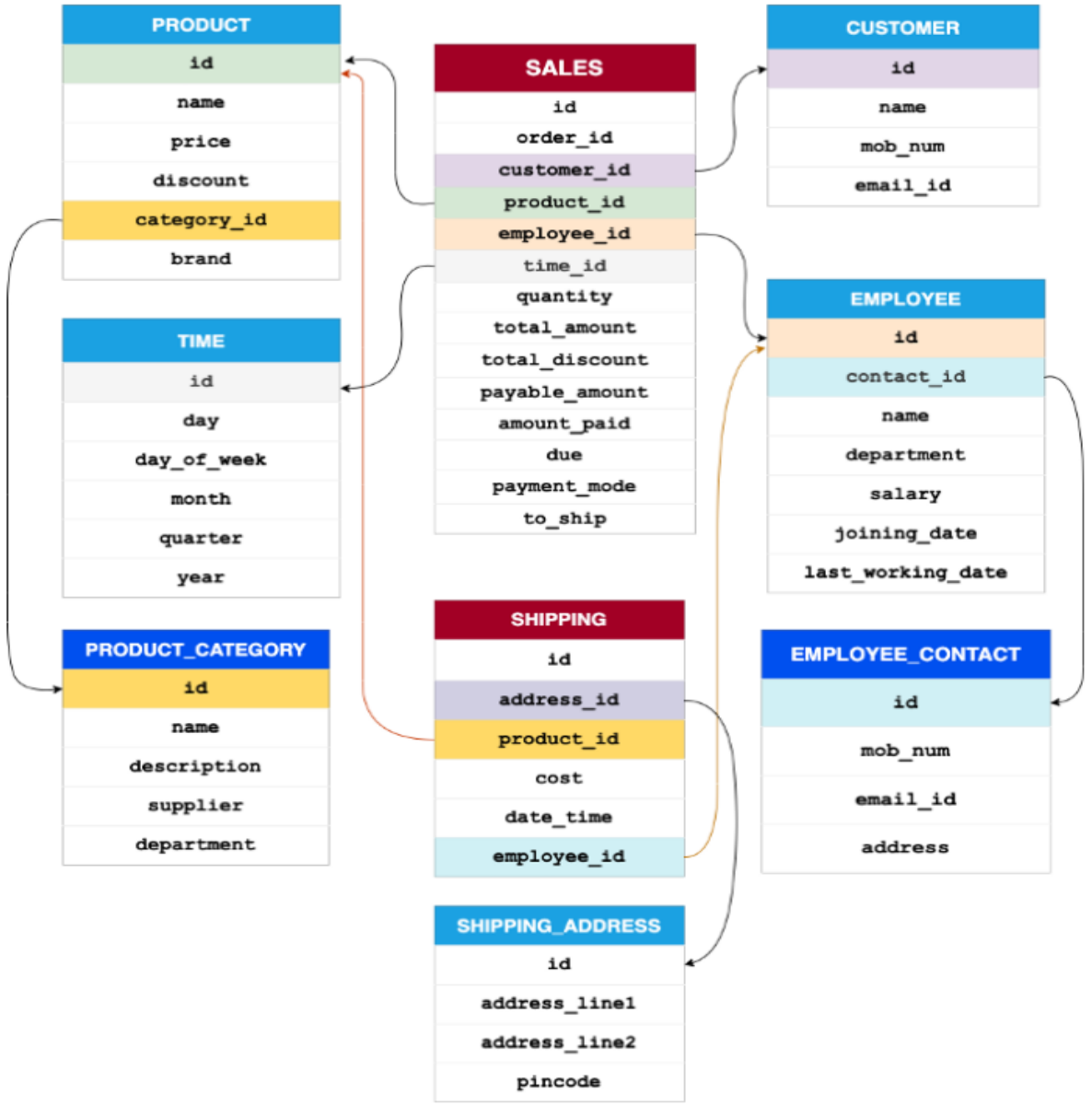
- It has more than one fact table
- It can be an extension of either star schema or snowflake schema
- The same dimension table can be shared between more than one fact table
- It is one of the **widely used schema for Data warehouse designing** and it is much more complex than star and snowflake schema. For complex systems, we require fact constellations

Advantages – provides a flexible schema

Disadvantages – much more complex and hence hard to implement and maintain

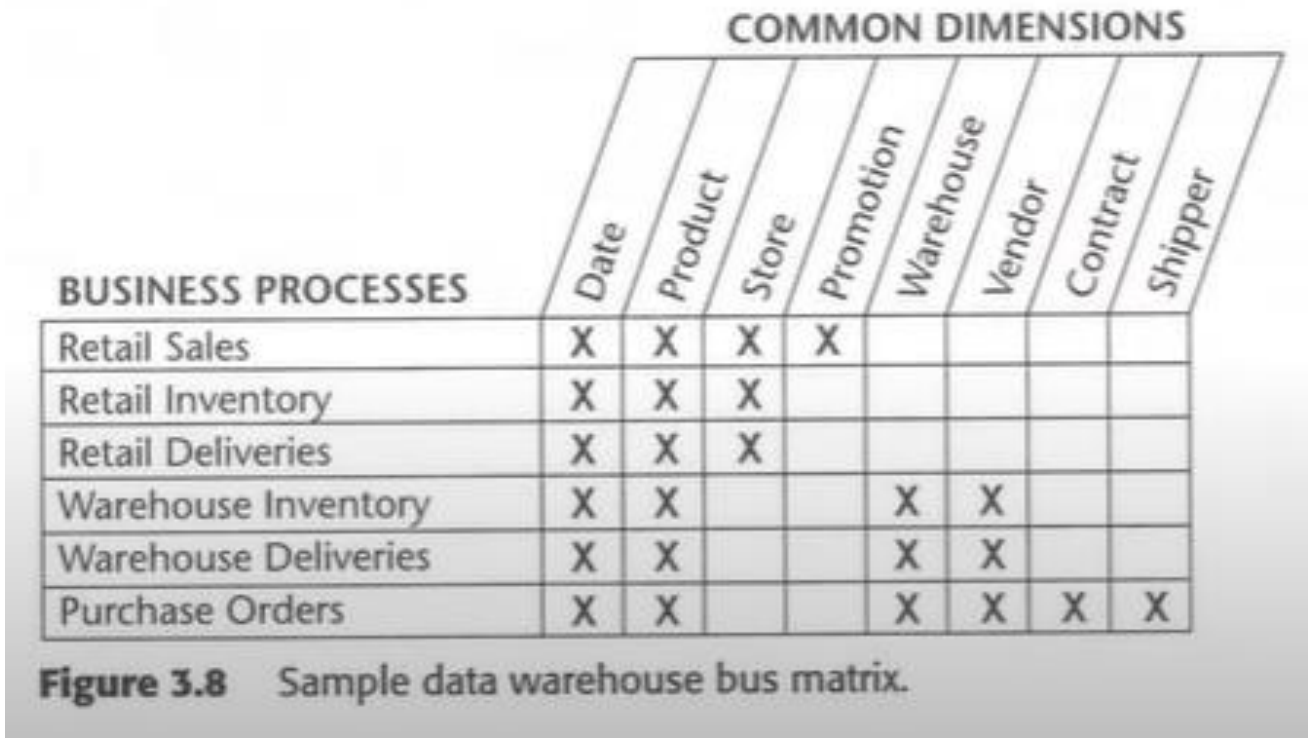
Galaxy (Fact Constellation) Schema Example

- Sales and Shipping are fact tables
- Others are dimension tables



The Bus Matrix

- Utilized to find the common dimensions



BUSINESS PROCESSES	Date	Product	Store	Promotion	Warehouse	Vendor	Contract	Shipper
	COMMON DIMENSIONS							
Retail Sales	X	X	X	X				
Retail Inventory	X	X	X					
Retail Deliveries	X	X	X					
Warehouse Inventory	X	X			X	X		
Warehouse Deliveries	X	X			X	X		
Purchase Orders	X	X			X	X	X	X

Figure 3.8 Sample data warehouse bus matrix.

Source: Agile data warehouse design by Lawrence Corr

The 7 W's of Data Warehouse Design

The dimensional model should describe below questions for the business process:

- What?
- When?
- Where?
- Who?
- How?
- How many?
- Why?

The 7 W's of Data Warehouse Design – User story

- An efficient way to get the required details
- Intuitive for business users

Source: Agile data warehouse design by Lawrence Corr

Thank you!

Schema

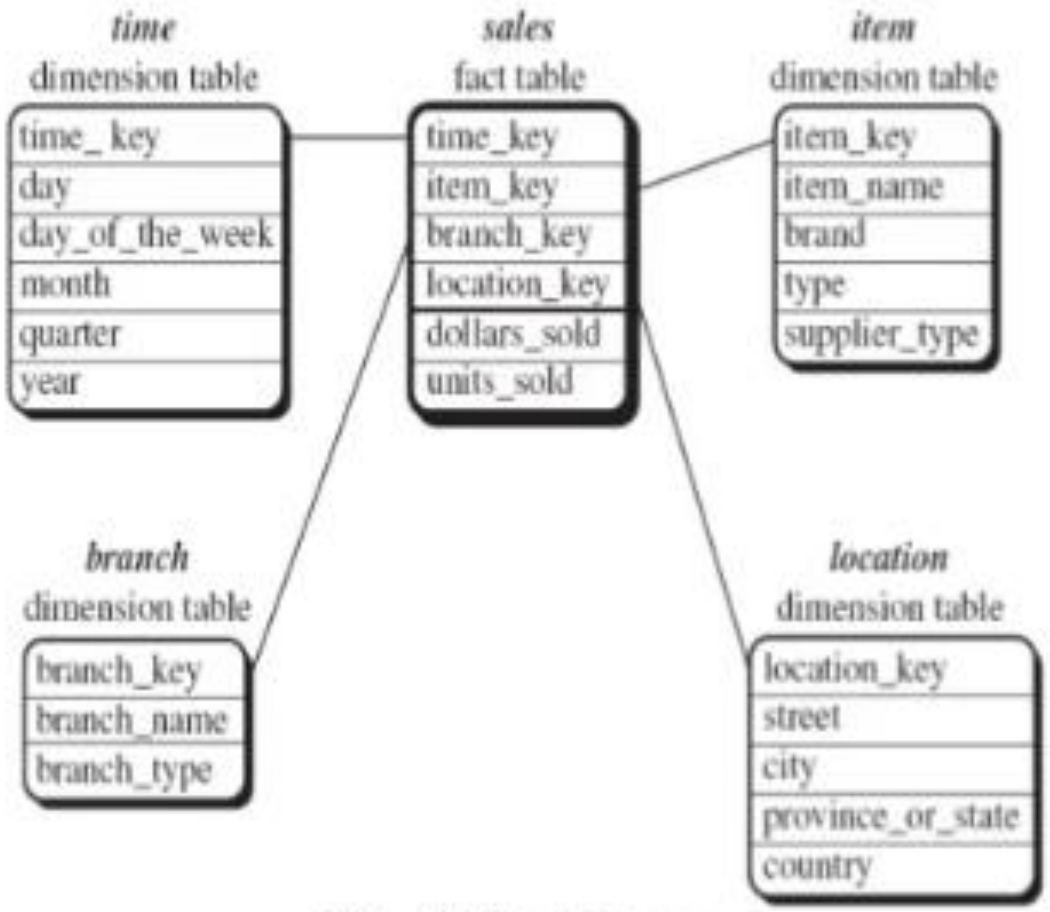


Fig. 5. Star Schema

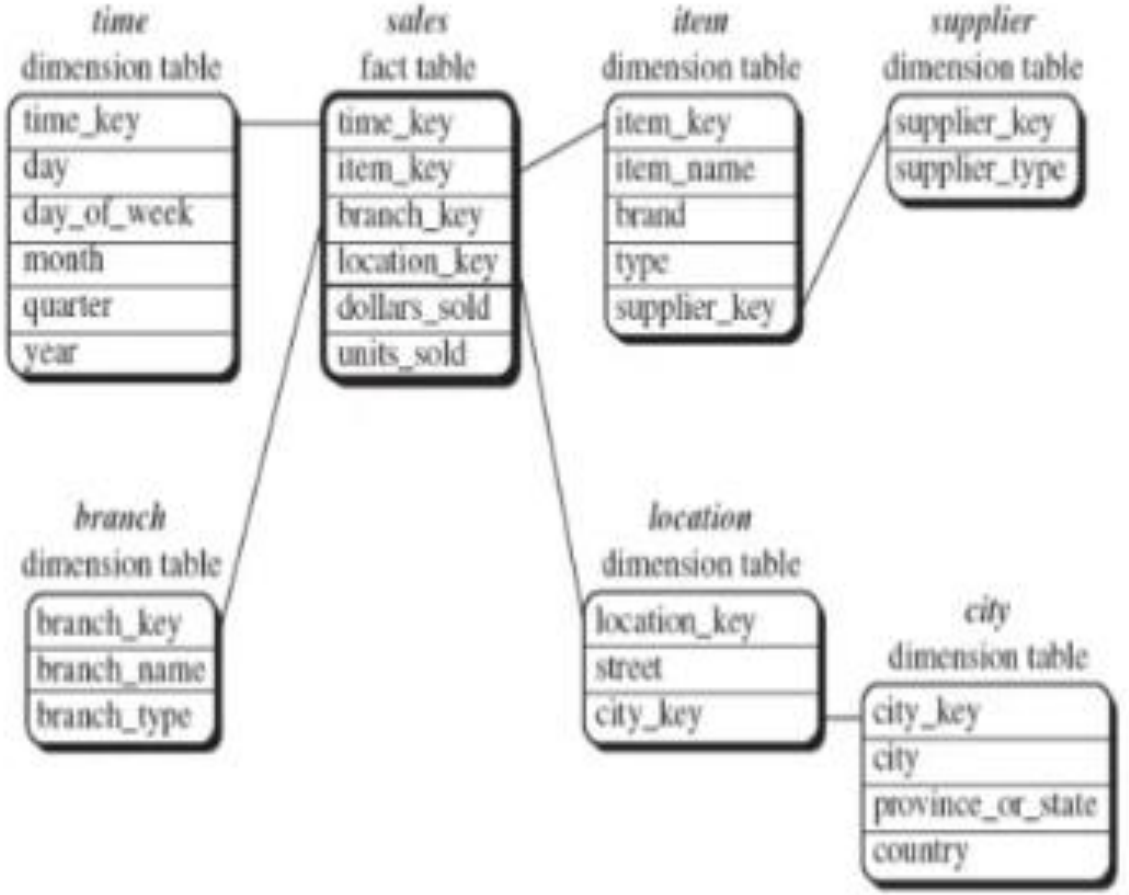


Fig. 6. Snowflake Schema

Schema

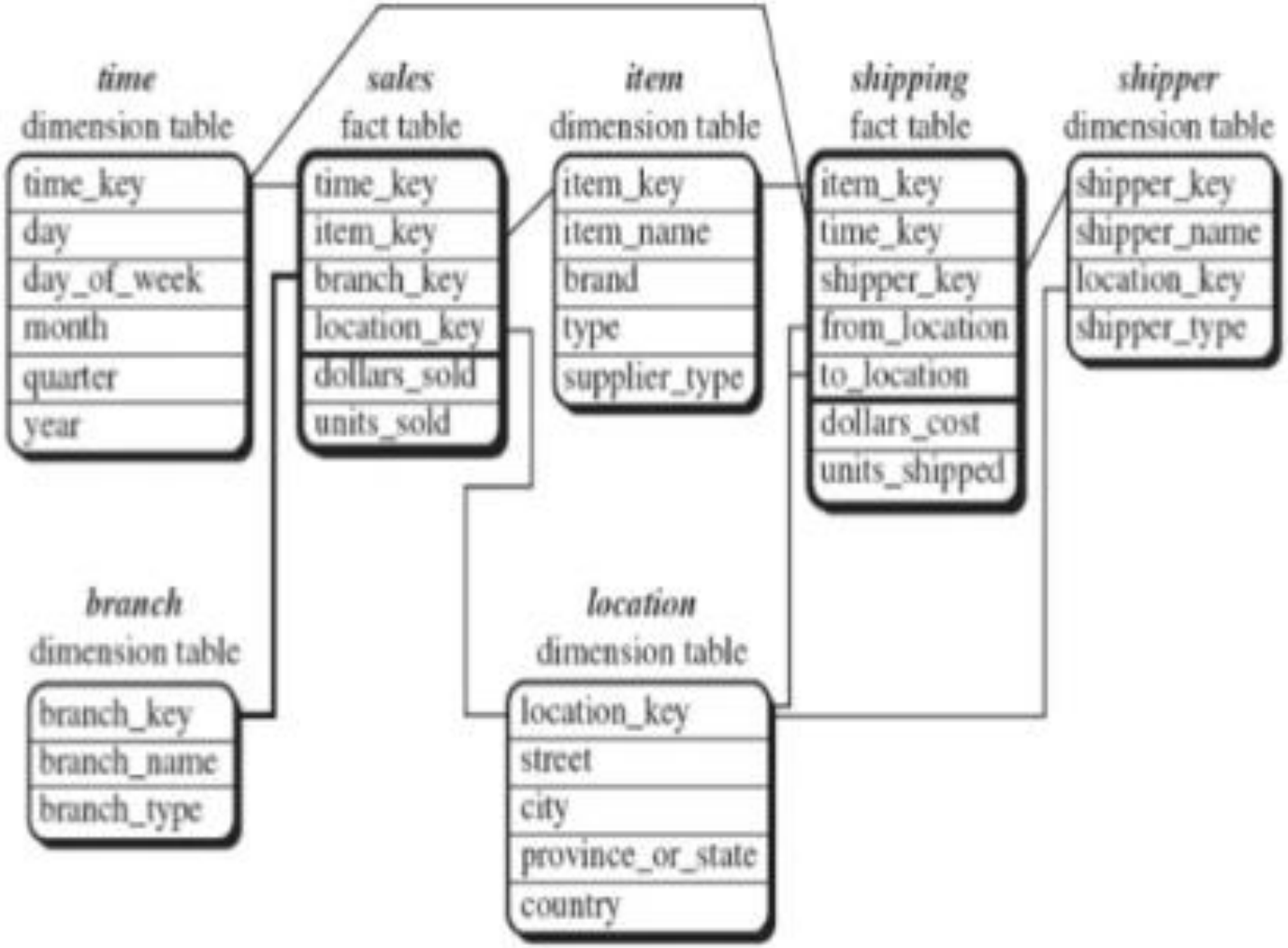


Fig. 7. Fact constellations