SS G515 - Data Warehousing

Dr. Yashvardhan Sharma
Assistant Professor, CS & IS Dept.
BITS-Pilani

Design Requirements

- Design of the DW must directly reflect the way the managers look at the business
- Should capture the measurements of importance along with parameters by which these parameters are viewed
- It must facilitate data analysis, i.e., answering business questions

ER Modeling

- A logical design technique that seeks to eliminate data redundancy
- Illuminates the microscopic relationships among data elements
- Perfect for OLTP systems
- Responsible for success of transaction processing in Relational Databases

Problems with ER Model

ER models are NOT suitable for DW?

- End user cannot understand or remember an ER Model
- Many DWs have failed because of overly complex ER designs
- Not optimized for complex, ad-hoc queries
- Data retrieval becomes difficult due to normalization
- Browsing becomes difficult

ER vs Dimensional Modeling

- ER models are constituted to
  - Remove redundant data (normalization)
  - Facilitate retrieval of individual records having certain critical identifiers
  - Thereby optimizing OLTP performance
- Dimensional model supports the reporting and analytical needs of a data warehouse system.

Dimensional Modeling: Salient Features

- Represents data in a standard framework
- Framework is easily understandable by end users
- Contains same information as ER model
- Packages data in symmetric format
- Resilient to change
- Facilitates data retrieval/analysis

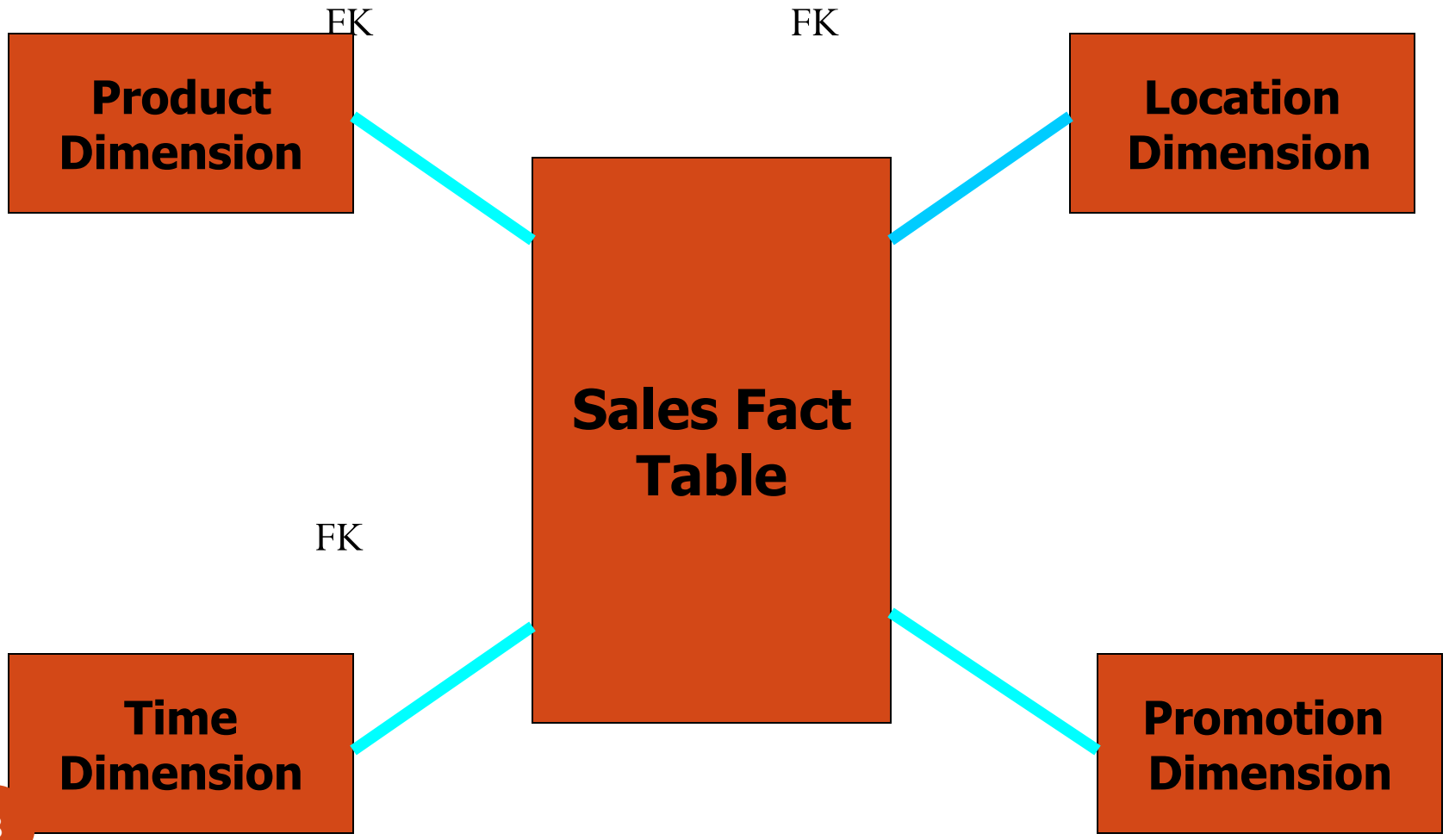
Dimensional Modeling: Vocabulary

- Measures or *facts*
- Facts are "numeric" & "additive"
- For example; Sale Amount, Sale Units
- Factors or *dimensions*
- Star Schemas
- Snowflake & Starflake Schemas

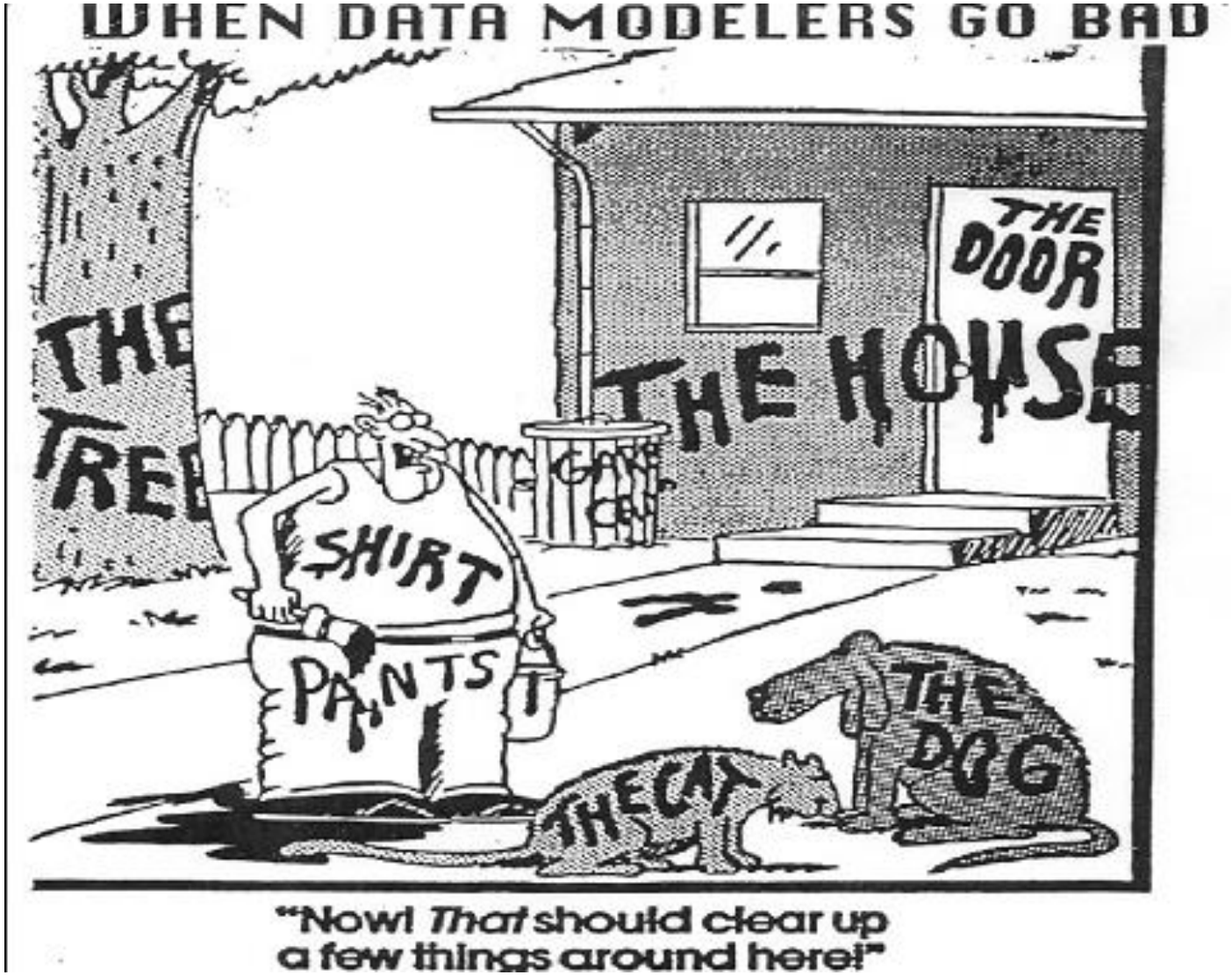
Sales Amt = f (Product,Location,Time)

Fact *Dimensions*

Star Schema


FK
FK
Product Dimension
Location Dimension
Sales Fact Table
FK
Time Dimension
Promotion Dimension

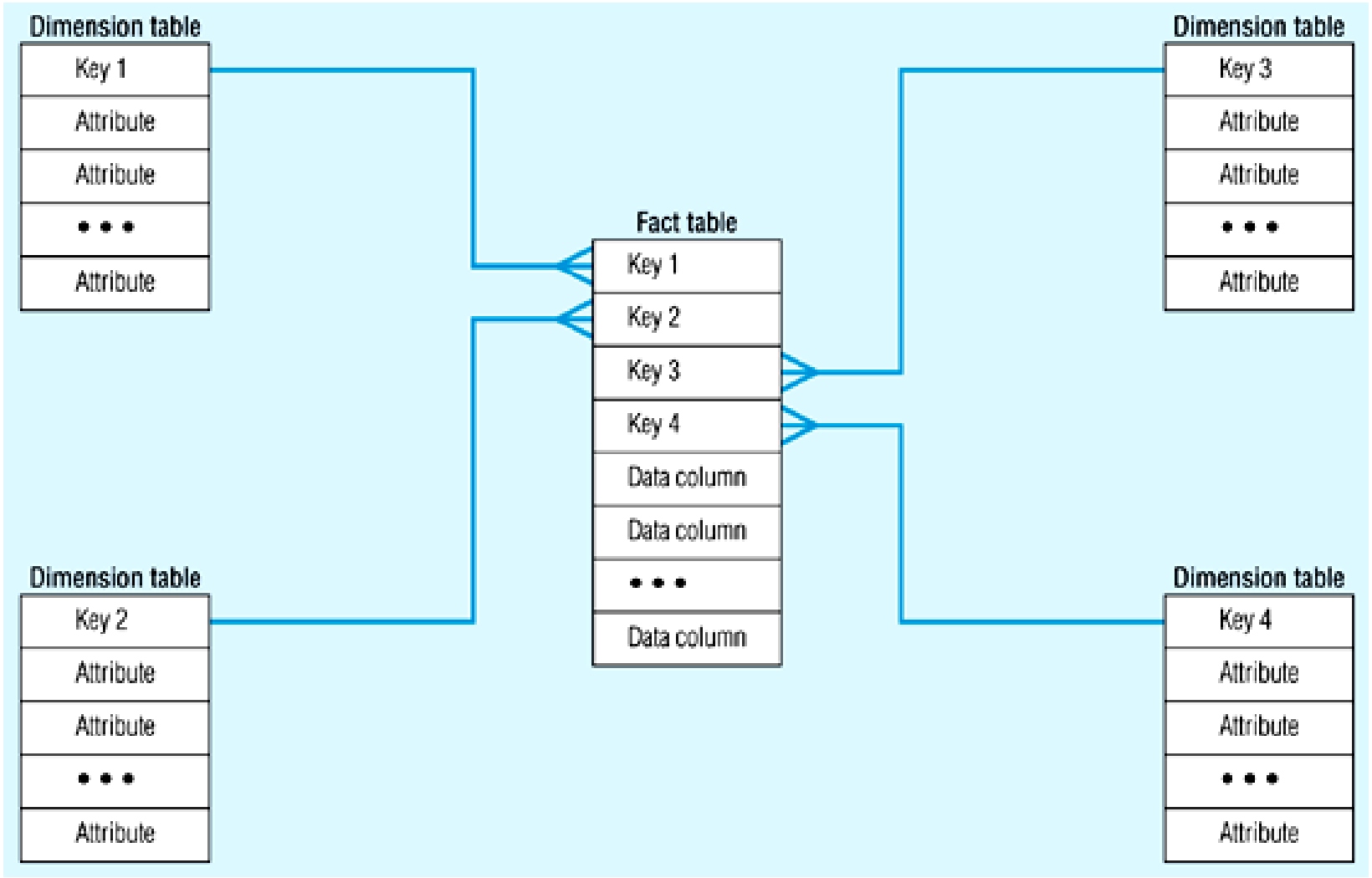
What Is Dimensional Modeling?



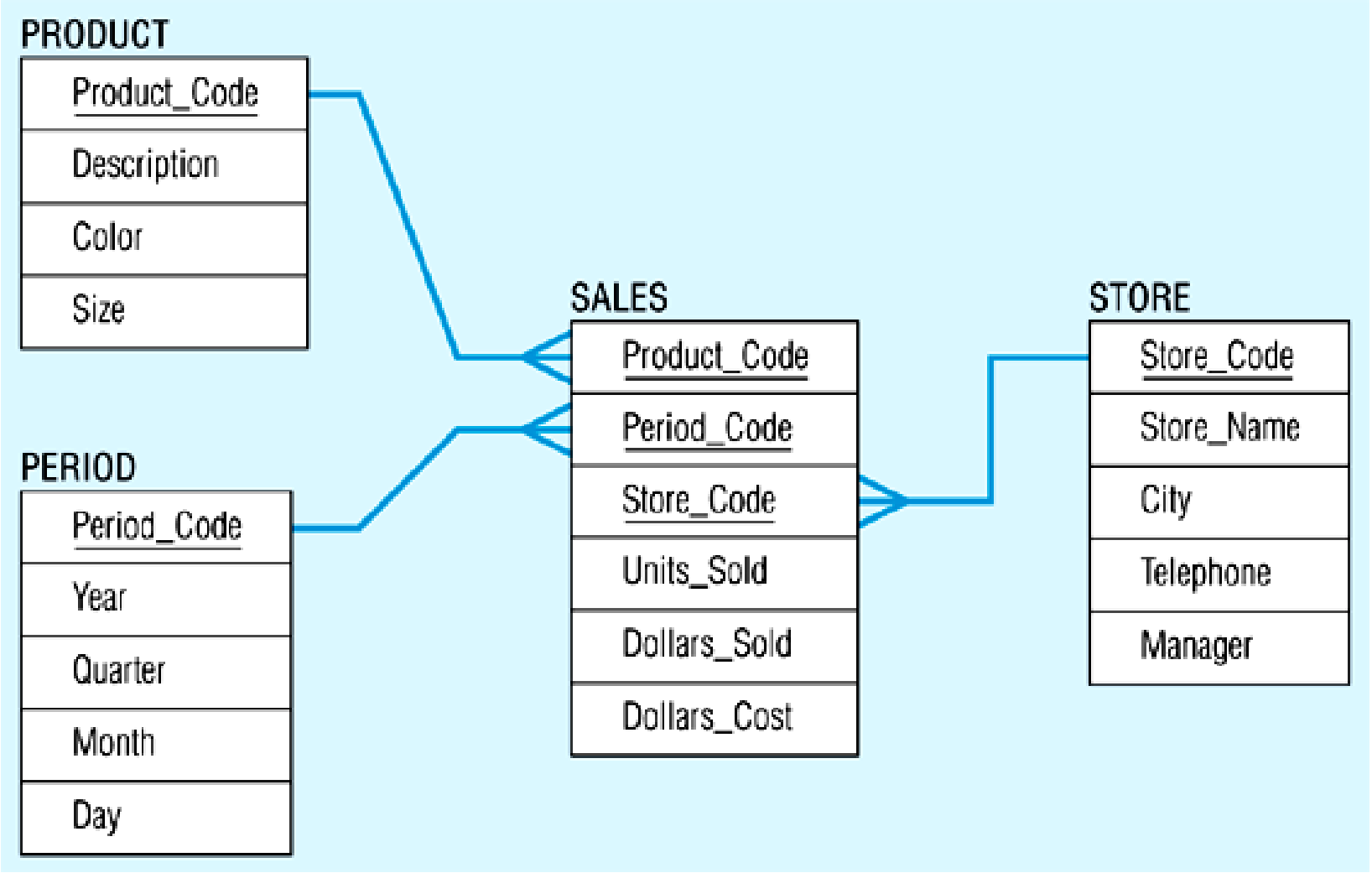
What is Dimensional Modeling (DM)?

- DM is a logical design technique that seeks to present the data in a standard, intuitive framework that allows for high-performance access.
- Can be implemented using a relational or a multidimensional DBMS
- Every dimensional model is composed of one table with a multipart key, called the fact table, and a set of smaller tables called dimension tables.
- Each dimension table has a single-part primary key that corresponds exactly to one of the components of the multipart key in the fact table.
- This characteristic "star-like" structure is often called a star join.

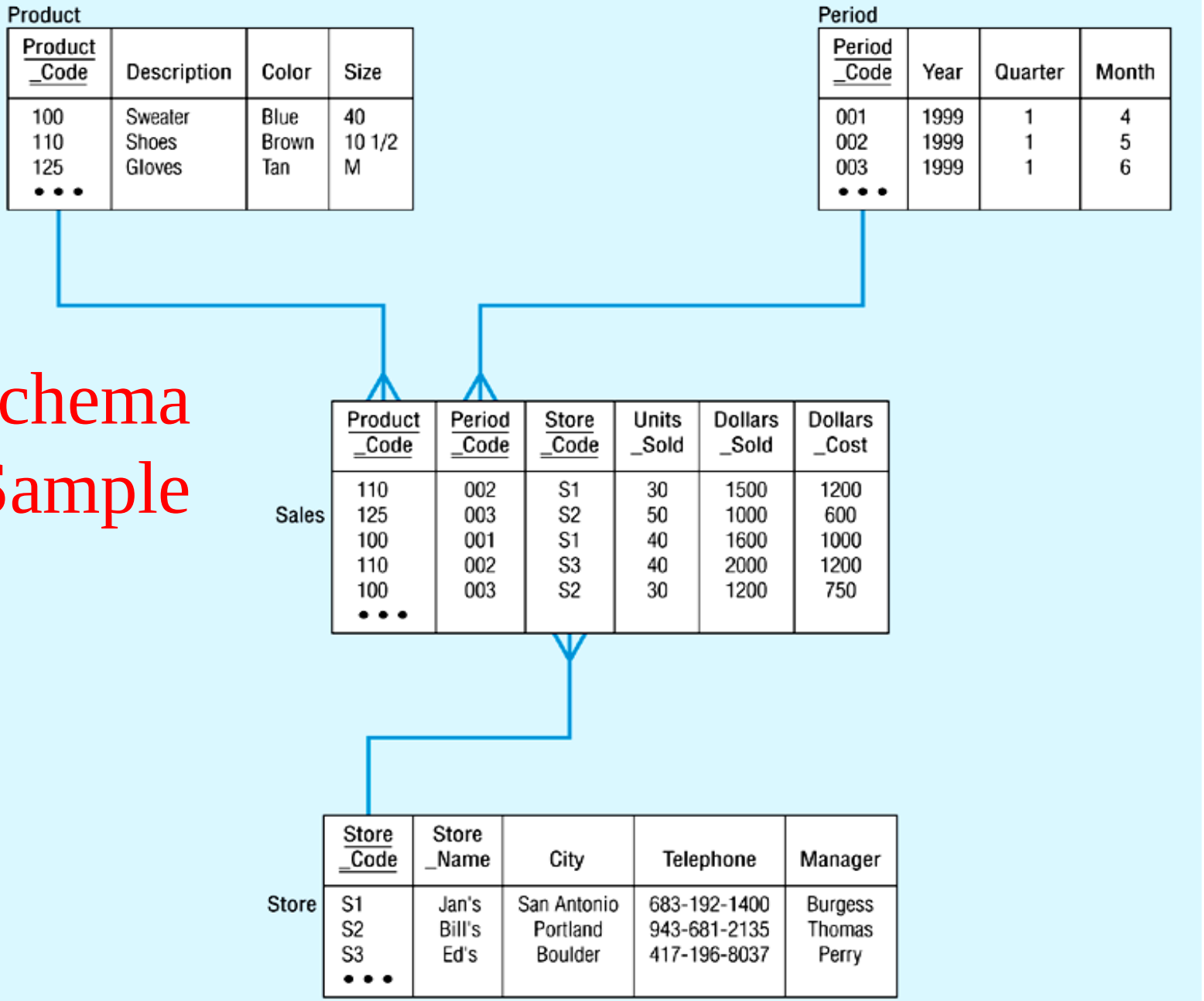
Star Schema (in RDBMS)



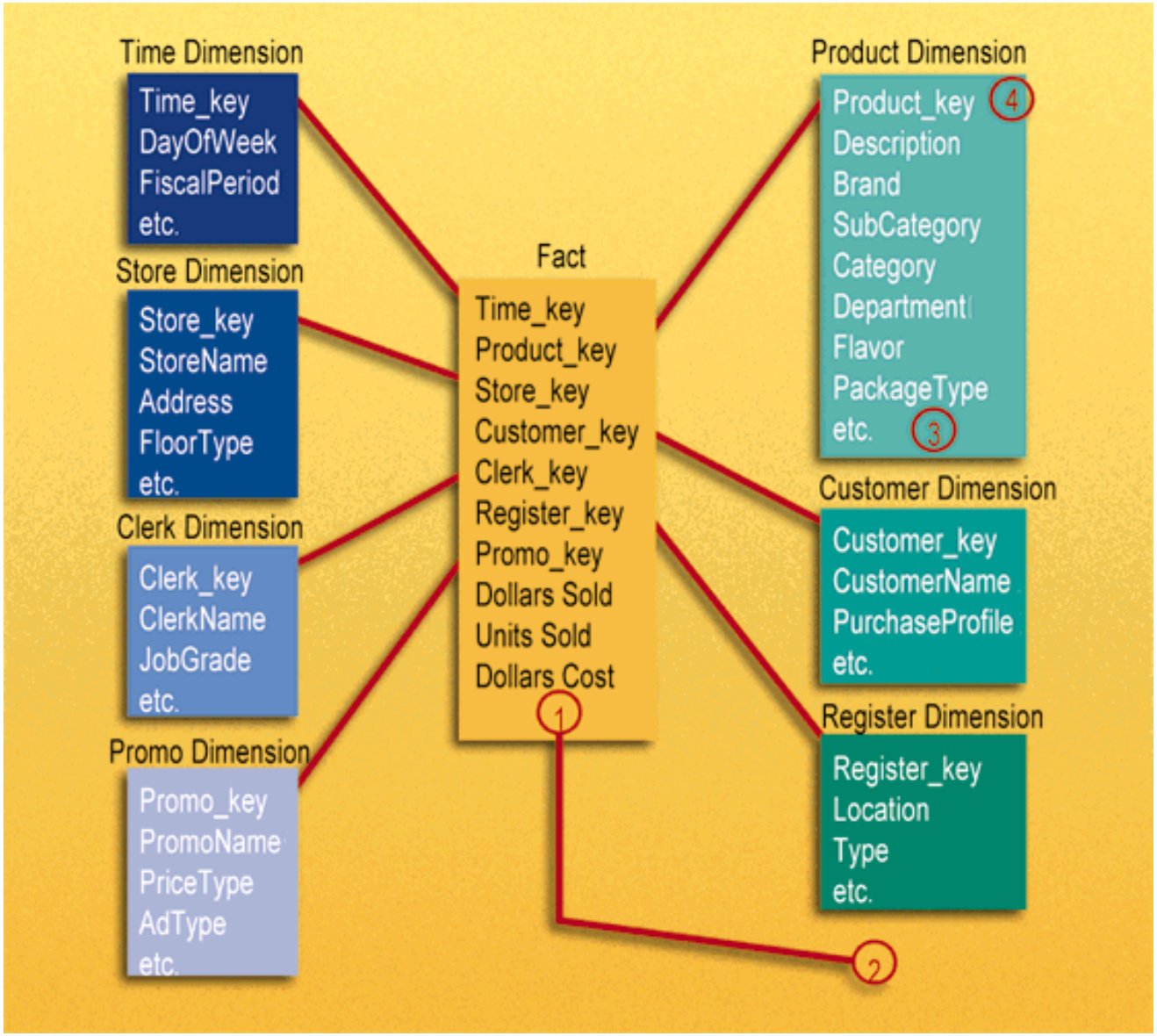
Star Schema Example



Star Schema with Sample Data

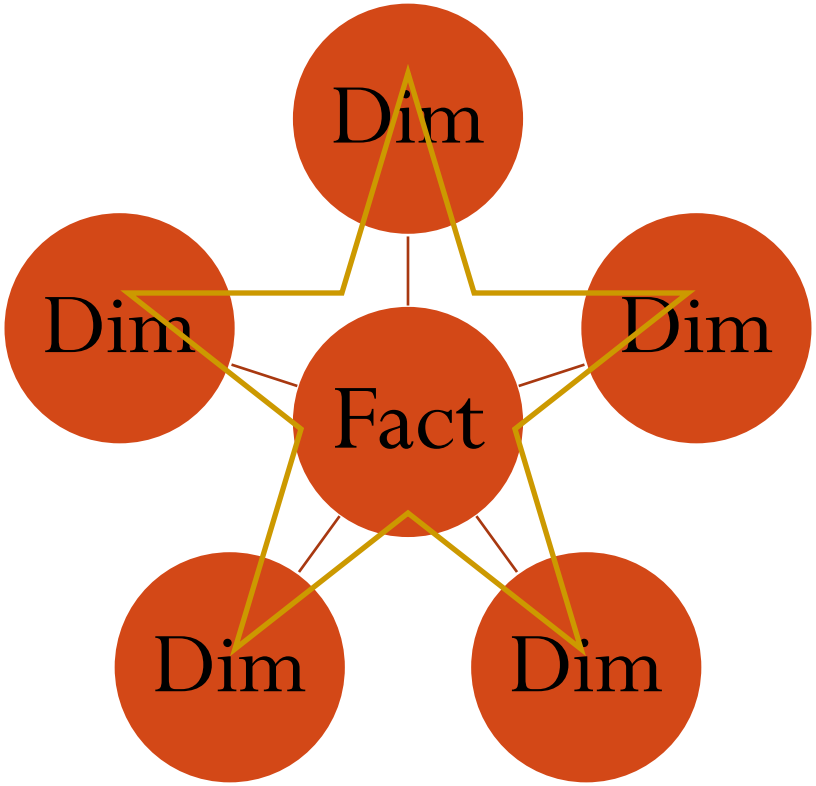


Dimensional Model Example



Star Schema

- Singe data (fact) table surrounded by multiple descriptive (dimension) tables



Dimensional Model: Fact Tables

- A fact table contains information about things that an organization wants to measure.
- A fact table's key is made up from the keys of two or more parents or dimension tables.
- A fact always 'resolves' a many-to-many relationship between the parent, or dimension tables.
- The most useful fact tables also contain one or more numerical measures, or facts, that occur for the combination of keys that define each record.
- Example: the facts are Dollars Sold, Units Sold, and Dollars Cost.

Dimensional model: Fact Tables

- The most useful facts in a fact table are numeric and additive.
- Additivity is crucial because data warehouse applications almost never retrieve a single fact table record; rather, they fetch back hundreds, thousands, or even millions of these records at a time, and often the most useful thing to do with so many records is to add them up.

Fact Tables

- Contains numerical measurements of the business
- Each measurement is taken at the intersection of all dimensions
- Intersection is the composite key
- Represents Many-to-many relationships between dimensions
- Examples of facts

  Sale_amt, Units_sold, Cost, Customer_count

Dimensional Model: Dimension Tables

- Dimension tables contain information about how an organization wants to analyze facts:
  - "Show me sales revenue (fact) for last week (time) for blue cups (product) in the western region (geography)
- Dimension tables most often contain descriptive textual information 'Blue cups', 'Western Region'
- Dimension attributes are used as the source of most of the interesting 'constraints' in data warehouse queries., and they are virtually always the source of the row headers in the SQL answer set.

Dimension Tables

- Contains attributes for dimensions
- 50 to 100 attributes common
- Best attributes are textual and descriptive
- DW is only as good as the dimension attributes
- Contains hierarchal information albeit redundantly
- Entry points into the fact table

Terminology

Dimensions

- The time independent, textual and descriptive attributes by which users describe objects.
- Combining all the attributes including hierarchies, rollups and sub-references into a single dimension is denormalization.
- Often the "by" word in a query or report
- Not time dependent

Facts

- Business Measurements
- Most Facts are Numeric
- Additive, Semi-Additive, Non-Additive
- Built from the lowest level of detail (grain)
- Very Efficient
- Time dependent

Benefits

- Performance (Integer relationships, natural partitioning, Single joins benefit SQL optimizer)
- Source system independence and multiple integration
- Supports Change management
- Usability/Simplicity (easy to read, interpret, join, calculate)
- Presentation (Consistency, Taxonomy, Labeling)
- Reuse (Conformed dimensions reduce redundancy, Role-plays)