Chapter 17

Data Warehouse

Database System Concepts, 7th Ed.

Data Analytics

- **Data Warehousing**
- Online Analytical Processing
- Data Mining

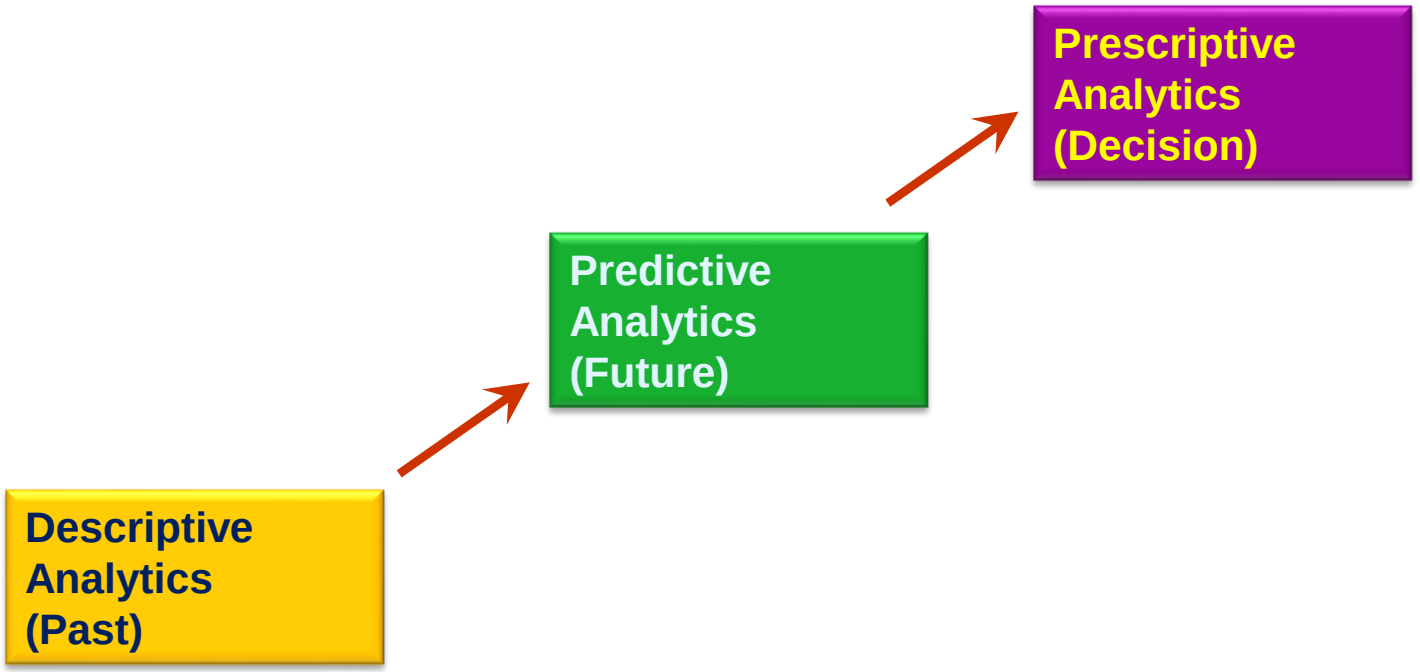
Data Analytics

- **Data analytics** refers to the processing of data to infer patterns, correlations, or models, the results of which are used to drive business decisions
- Predictive models are widely used
  - E.g., use customer profile features (e.g., income, age, gender, education, employment) and past history of a customer to predict likelihood of default on loan
    - use prediction to make loan decision
  - E.g., use past history of sales (by season) to predict future sales
    - to decide what/how much to produce/stock
    - to target customers
- Examples of business decisions:
  - What items to stock?
  - What insurance premium to charge?
  - To whom to send advertisements?

Data Analytics

- **Machine learning** techniques are key to finding patterns in data and making predictions
- **Data mining** extends techniques developed by machine-learning communities to run them on very large datasets
- The term **business intelligence (BI)** is used in a broadly similar sense to data analytics
- The term **decision support** is used in a related but narrower sense to BI, which focuses on reporting and aggregation; the associated systems are **DSS (decision support systems)**

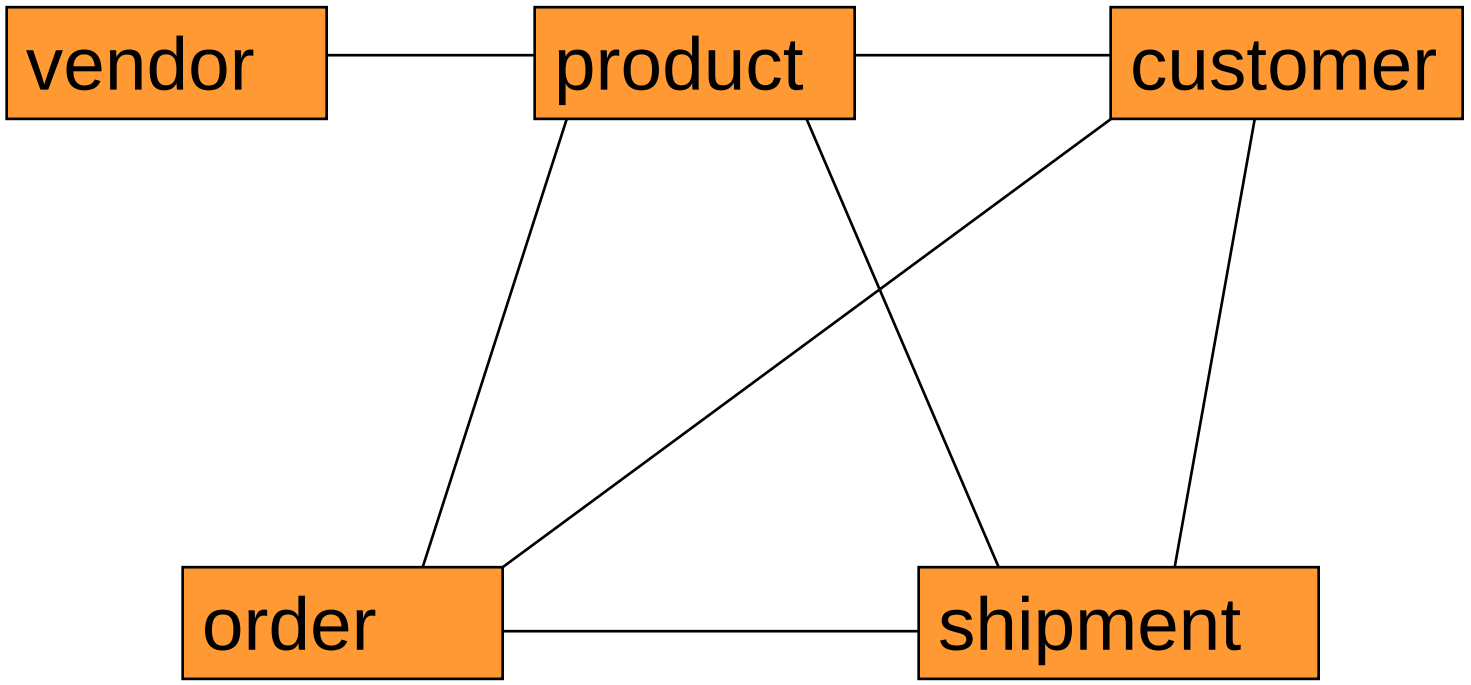
Data Analytics



Multidimensional Data and Warehouse Schemas

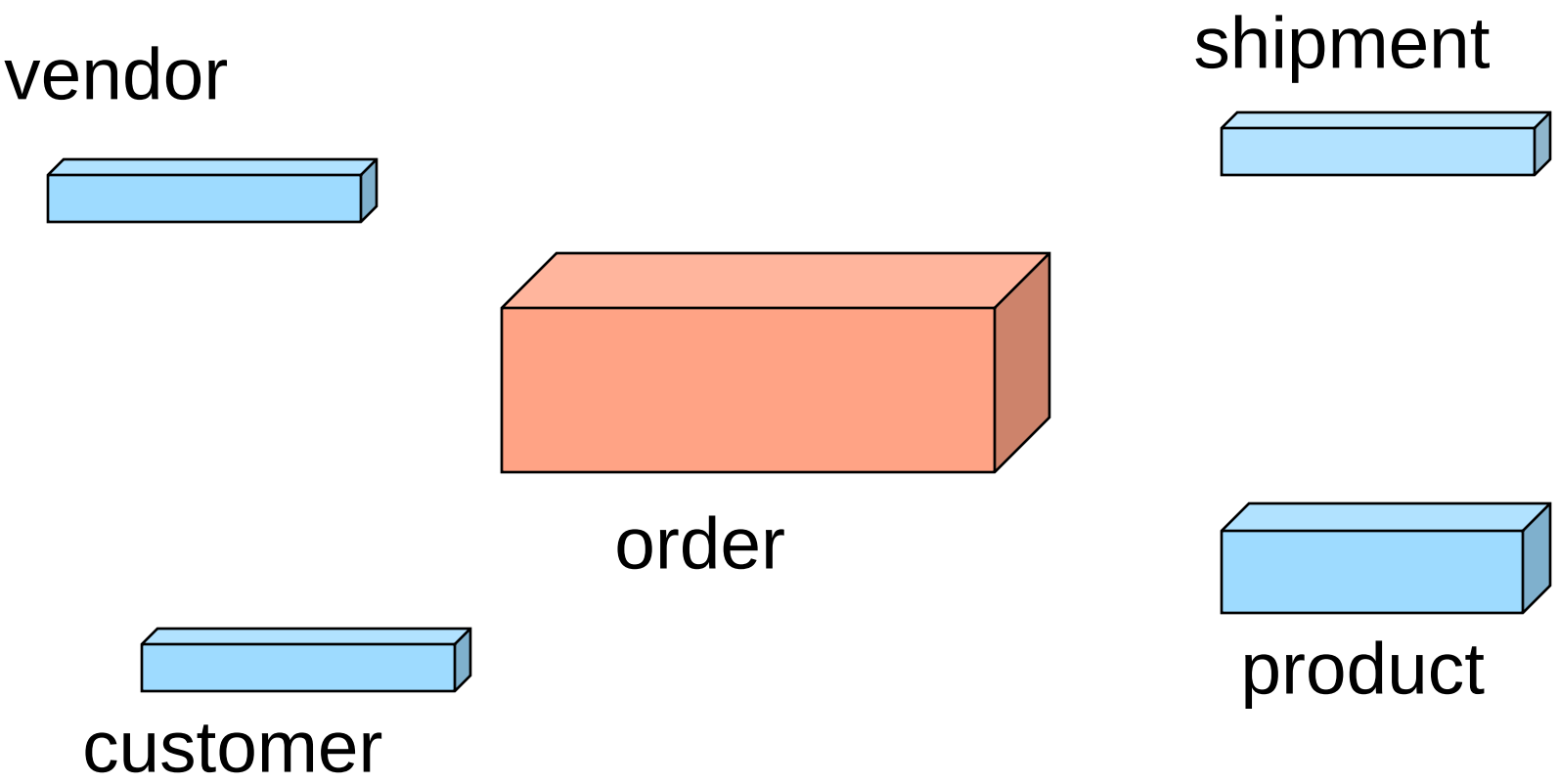
- Data Warehouse often uses a **star schema,** or sometimes a **snowflake schema**
  - fact table joined with dimension tables
  - group-by on dimension table attributes
  - aggregation on measure attributes of fact table
- Some applications do not find it worthwhile to bring data to a common schema
  - **Data lakes** are repositories which allow data to be stored in multiple formats, without schema integration
  - Less upfront effort, but more effort during querying

Limitations of Entity Relationship Modelling

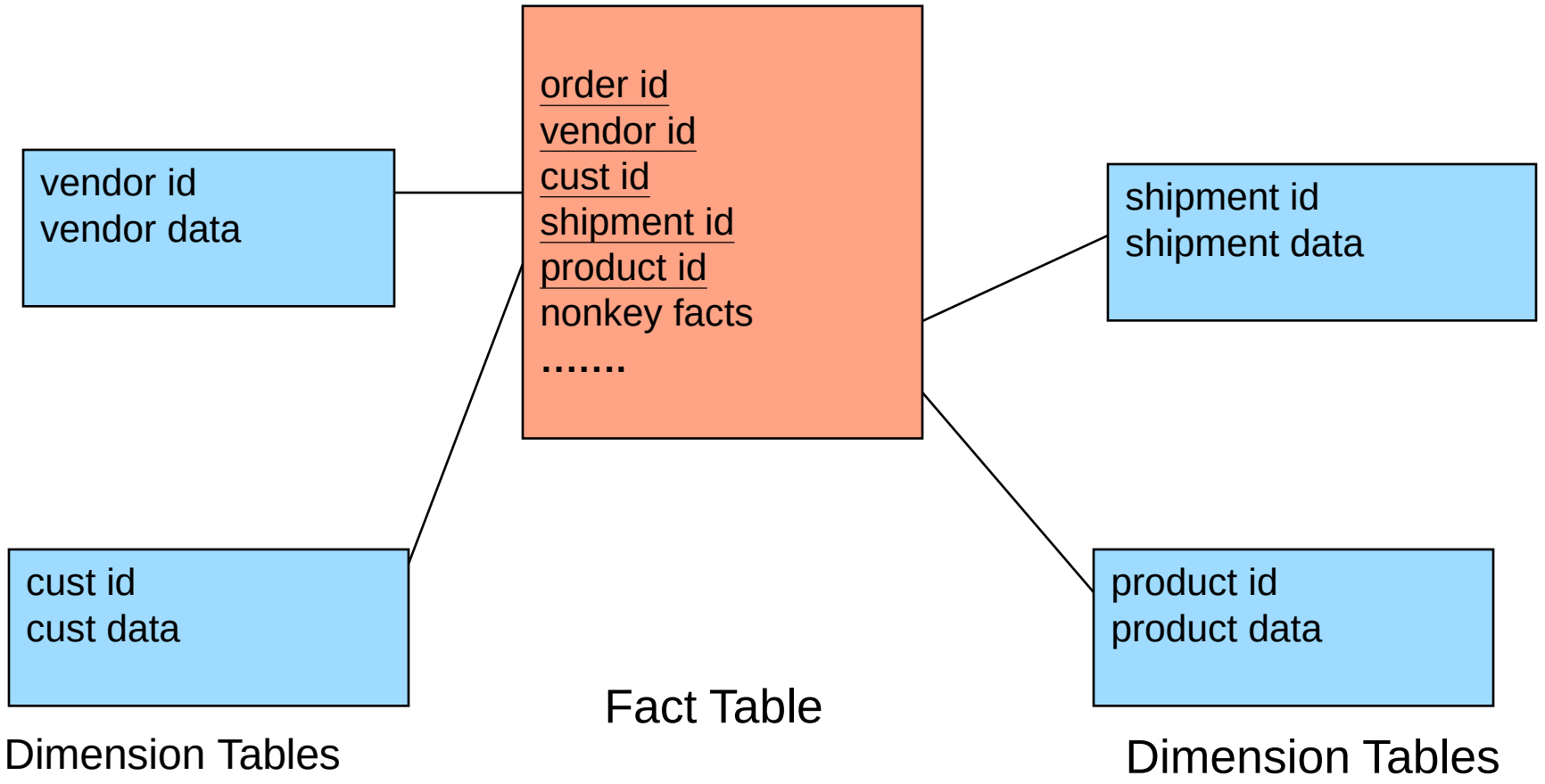


- Very symmetric
- Cannot tell which table is most important or largest
- Cannot tell which tables hold static or dynamic business information
- Joining of any tables is possible by user

A 3-D Perspective: Orders are Much More Numerous



Star Schema: Fact and Dimension Tables



Star Schema

- The centre of the star consists of a fact table and the points of the star are the dimension tables
- A star schema is characterized by very large *fact* tables that contain the primary information in the data warehouse and a number of much smaller *dimension* tables (or *lookup* tables), each of which contains information about the entries for a particular attribute in the fact table
- A *star query* is a join between a fact table and a number of lookup tables
- Each lookup table is joined to the fact table using a primary-key to foreign-key join, but the lookup tables are not joined to each other

Star Schema

- Very asymmetric
- Fact table is the only table that has multiple connections connecting it to other tables
- All other tables have only a single connection attaching them to the central table
- Commonly used in Data Warehouses

Fact Table

- Fact table tends to contain additive facts
- Fact tables have composite keys
- All other tables are dimension tables
- Every combination of key values would give rise to a different record in the fact table
- Fact table is naturally highly normalized

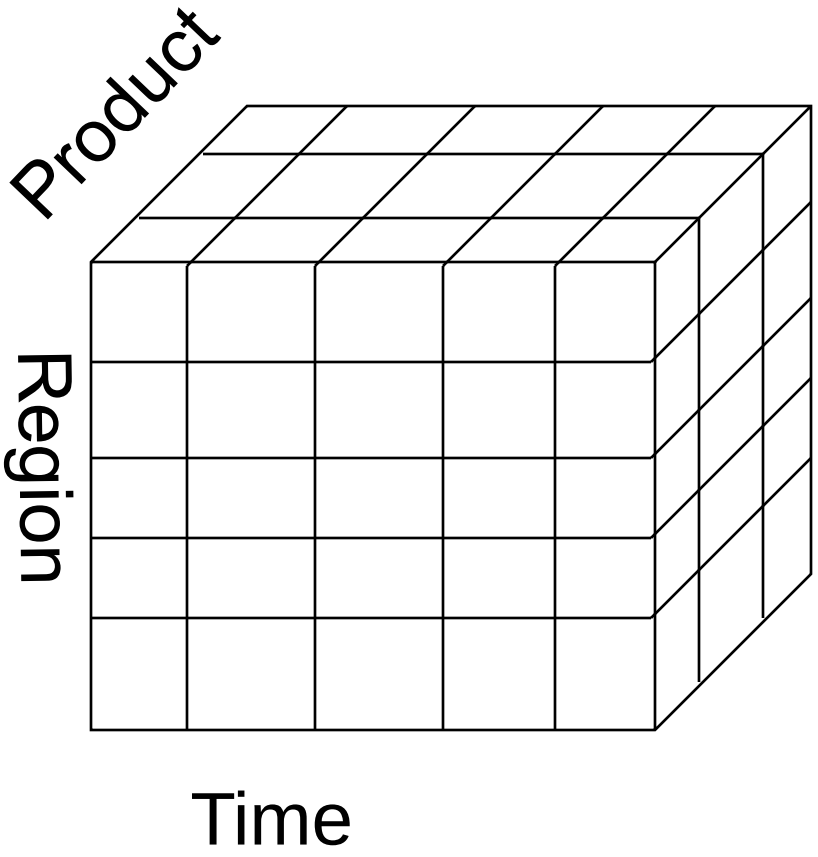
Dimension Tables

- Dimension table tends to contain textual or non-additive facts
- Dimension tables should not be normalized
- Normalized dimension tables destroy the ability to browse

Joins

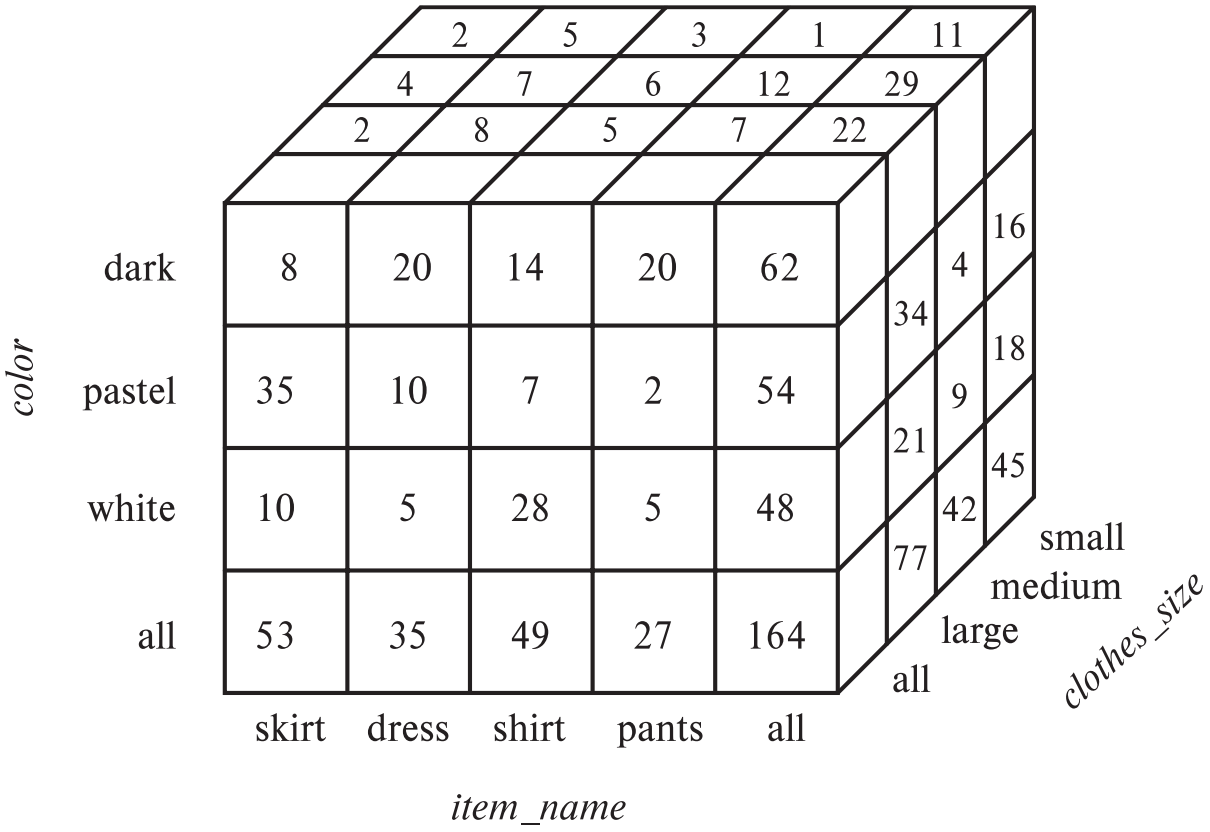
- Generally, there are only a few joins in a dimensional database (typically joining the fact table with one or more of the dimension tables)
- Each of the joins expresses a fundamental relationship between items in the underlying business
- Any joins are in principle possible in an ER database
  - most have little significance

Hypercube/Data Cube View



A *conceptual view* of multidimensional information is a ***hypercube*** (also called ***data cube***). Each combination of keys in the fact table correspond to one of the small cubes (dice)

Hypercube/Data Cube



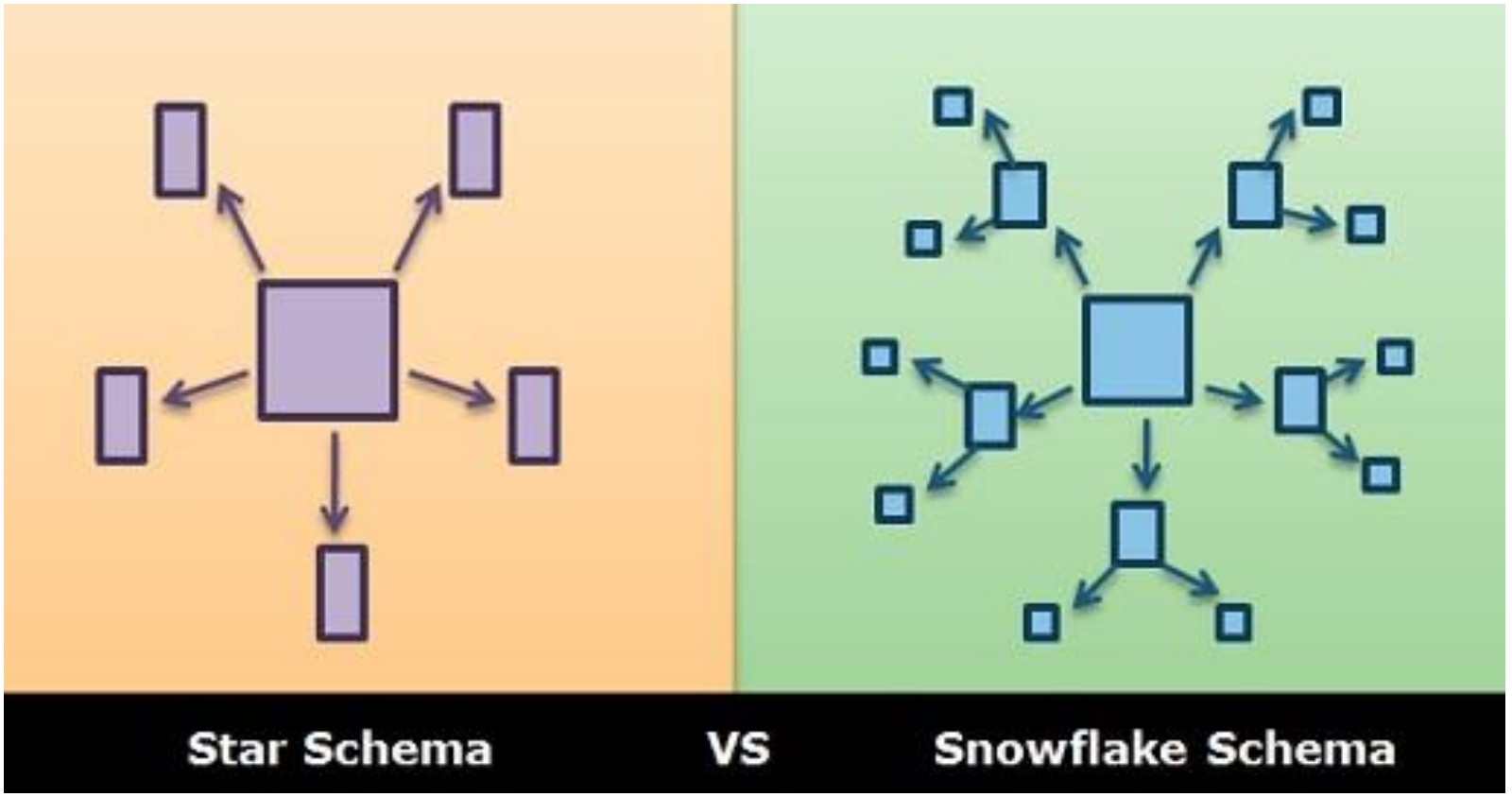
Star Join

- A *star join* is a primary-key to foreign-key join of the dimension tables to a fact table
- The fact table normally has a concatenated index on the key columns to facilitate this type of join
- The main advantages of star schemas are that they:
  - Provide a direct and intuitive mapping between the business entities being analysed by end users and the schema design
  - Provides highly optimized performance for typical data warehouse queries

Snowflake Schema

- The snowflake schema is a more complex data warehouse model than a star schema, and is a type of star schema
- It is called a snowflake schema because the diagram of the schema resembles a snowflake
- Snowflake schemas normalize dimensions to eliminate redundancy
- The dimension data has been grouped into multiple tables instead of one large table

Snowflake Schema



Snowflake Schema

- **Snowflake Schema is Undesirable**
- For example, a product dimension table in a star schema might be normalized into a Product table, a Product_Category table, and a Product_Manufacturer table in a snowflake schema
  - A given product may be a small household product but is encoded as product Category 5 in the Product table, and the separate Product_Category table would also give other information such as weight, colour, safety to children etc. (and a foreign key join on category_id is necessary to obtain such information)
  - The manufacturer information may be stored in a separate table giving details of the manufacturer (and a foreign key join on manufacturer_id is necessary to obtain the full manufacturer information)
- While this saves space, it increases the number of dimension tables and requires more foreign key joins
- The result is more complex queries and reduced query performance

Data and Information

- Information is not easily obtainable from the data
- “How has account activity been different this year from each of the past five years?”
- Not enough historical data are stored to meet DSS requests

Operational vs DSS Data

- Application oriented
- Detailed
- Accurate
- Small amount of data used in a process
- Serves the clerical community
- Frequently updated
- Run repetitively
- Transaction driven
- High availability
- Performance sensitive

- Subject oriented
- Summarized
- Represents values over time
- Large amount of data used in a process
- Serves the executive community
- Not updated
- Run heuristically
- Analysis driven
- Relaxed availability
- Performance relaxed

Data Warehouse

- A data warehouse is a
  - **subject-oriented**
  - **integrated**
  - **non-volatile**
  - **time variant**

  collection of data in support of management's decision

Subject Orientation

Operational

- Focus on applications areas of an organization
- Insurance Company
  - auto
  - life
  - health
  - accident

Data Warehouse

- Focus on major subject areas of an organization
- Insurance Company
  - customer
  - policy
  - premium
  - claim

Integration

- Converting data from several operational databases to data warehouse
- Issues
  - data encoding
  - attribute measurement
  - multiple sources
  - conflicting keys

Non-Volatility

Operational

- Highly changeable
- Lots of updates
- Record by record manipulation

Data Warehouse

- No update
- Mass load and access
- Refreshing the data warehouse

Time Variancy

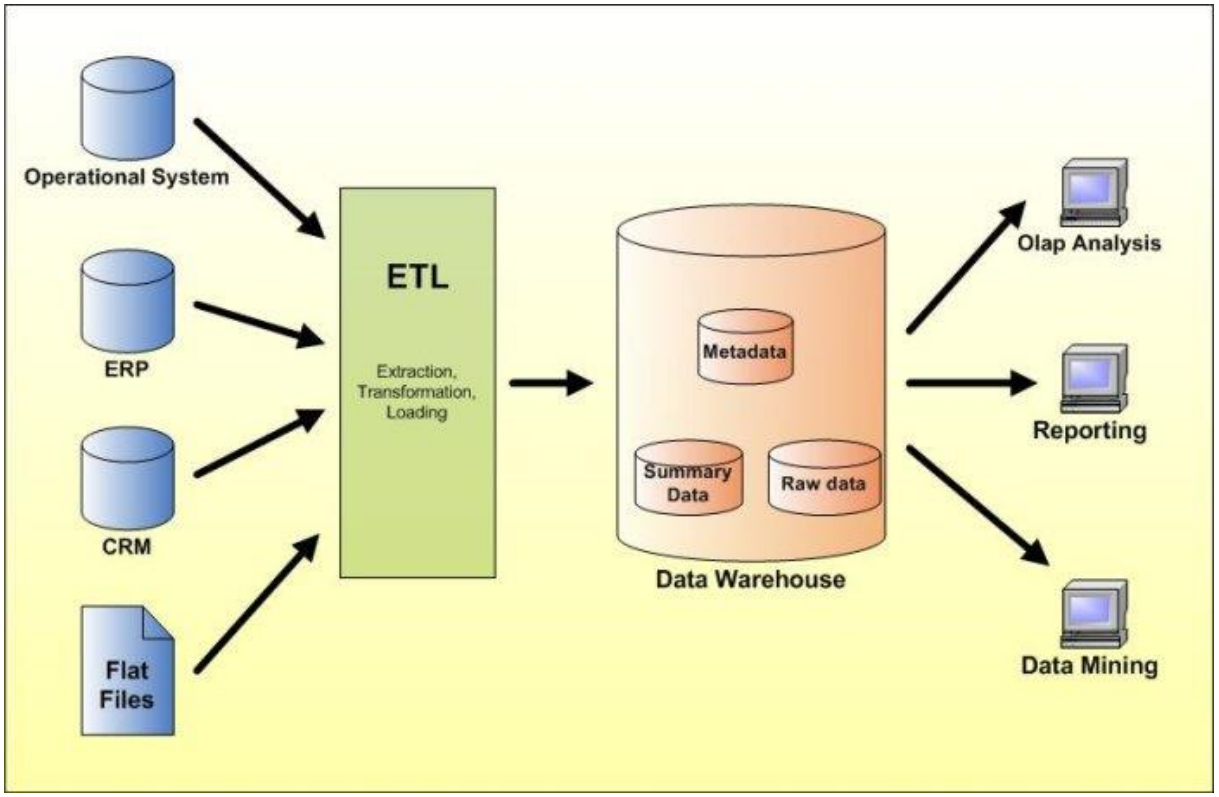
Operational

- Time-horizon: current to 60-90 days
- Accurate as of moment of access
- Key structure may/may not contain an element of time

Data Warehouse

- Time-horizon: 5-10 years
- Sophisticated snapshots of data
- Key structure contains an element of time

Data Warehouse Architecture



- Combines multiple data sources, e.g.,
  - Customer Relationship Management (CRM)
  - Enterprise Resource Planning (ERP)
  - Supply Chain Management (SCM)
- Supports managerial decision making

Data Mart

- A data mart is a simple form of a data warehouse that is focused on a single subject (or functional area), such as Sales or Finance or Marketing
  - Data marts are often built and controlled by a single department within an organization
- Given their single-subject focus, data marts usually draw data from only a few sources
  - The sources could be internal operational systems, a central data warehouse, or external data
- Three basic types of data marts
  - Dependent data marts draw data from a central data warehouse that has already been created
  - Independent data marts, are standalone systems built by drawing data directly from operational or external sources of data or both
  - Hybrid data marts can draw data from operational systems or data warehouses