Business Intelligence 1

Data Warehouse Usage and Design

Slides adapted from Jiawei Han, Micheline Kamber, Jian Pei, (2011), Data Mining: Concepts and Techniques, Third Edition, The Morgan Kaufmann Series in Data Management System.

Advantages of a Data Warehouse

- Advantages of a Data Warehouse
  - Competitive Advantage.
    - Gives managers access to data/information for use in the decision-making process.
  - Data Quality and Consistency.
    - Data stored in a central location for efficient analysis.
    - Data stored in a standard format.
  - Customer Relationship Management
    - Keeps track of the organization's customer base.
  - Tracks Historical Data
    - Allows tracking of trends, patterns and exceptions over time.

Disadvantages of a Data Warehouse

- Disadvantages of a Data Warehouse
  - Extra Workload.
    - Needs a team of specialist personnel to maintain.
  - Data Inflexibility.
    - Stores structured data.
    - Stored in a standard format.
    - Unstructured or semi-structured data not supported.
  - Ownership Concerns.
    - Departments don’t like sharing their data.
    - Departments loose ownership of data
    - Centrally stored data can lead to security issues.

Data Warehouse Usage

- A data warehouse can be use for many applications including:
  - Reporting and ad hoc queries
    - Organisational reports including a variety of graphs and charts, statistical analysis and ad-hoc queries.
  - Multi-dimensional analysis
    - Data view from many dimensions (viewpoints).
    - OLAP operations including slice/dice, drilling, pivoting.
  - Visualisation and Data mining
    - Visualisation using graphs and charts; E.G. Tableau.
    - Data mining using algorithms such as association rules, clustering, classification and prediction to identify trend and patterns.

Design Views of a Data Warehouse

- Different design views
  - Top-down view.
    - Overall view of organizational data requirements.
    - Selection of the relevant data/information.
  - Data source view.
    - Overall view of data being captured, stored and managed by operational systems.
  - Data warehouse view.
    - view of fact and dimension tables.
  - Business query view
    - Overall view of the end-user's data requirements.

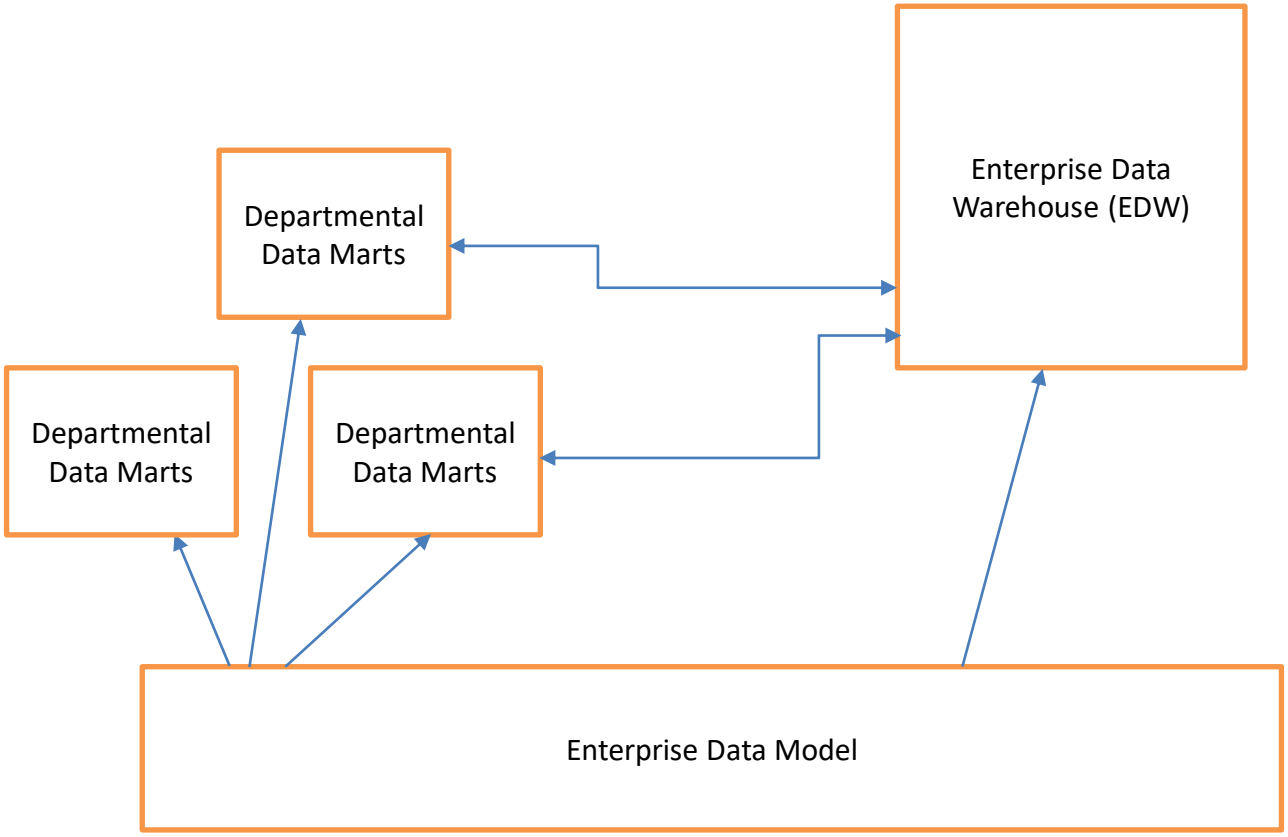
Data Warehouse Design
Top-Down Approach

- Design.
  - Data Warehouse designed for whole organisation.
  - Enterprise Data Warehouse(EDW) built first.
  - Data Marts created as subsets of the EDW.
  - Mature Design.
- Advantages.
  - Systematic solution
  - Minimises integration problems
- Disadvantages.
  - Expensive.
  - Long development time.
  - Lacks flexibility.
  - Costly.

Data Warehouse Design
Bottom-Up Approach

- Design.
  - Starts with experiments and prototypes.
  - Departmental data marts built first.
  - EDW – Combination of departmental data marts
  - Rapid Design.
- Advantages.
  - Design, development and deployment of independent data marts.
  - Flexibly.
  - Low cost.
  - Rapid return on investment.
- Disadvantages.
  - Integration problems.

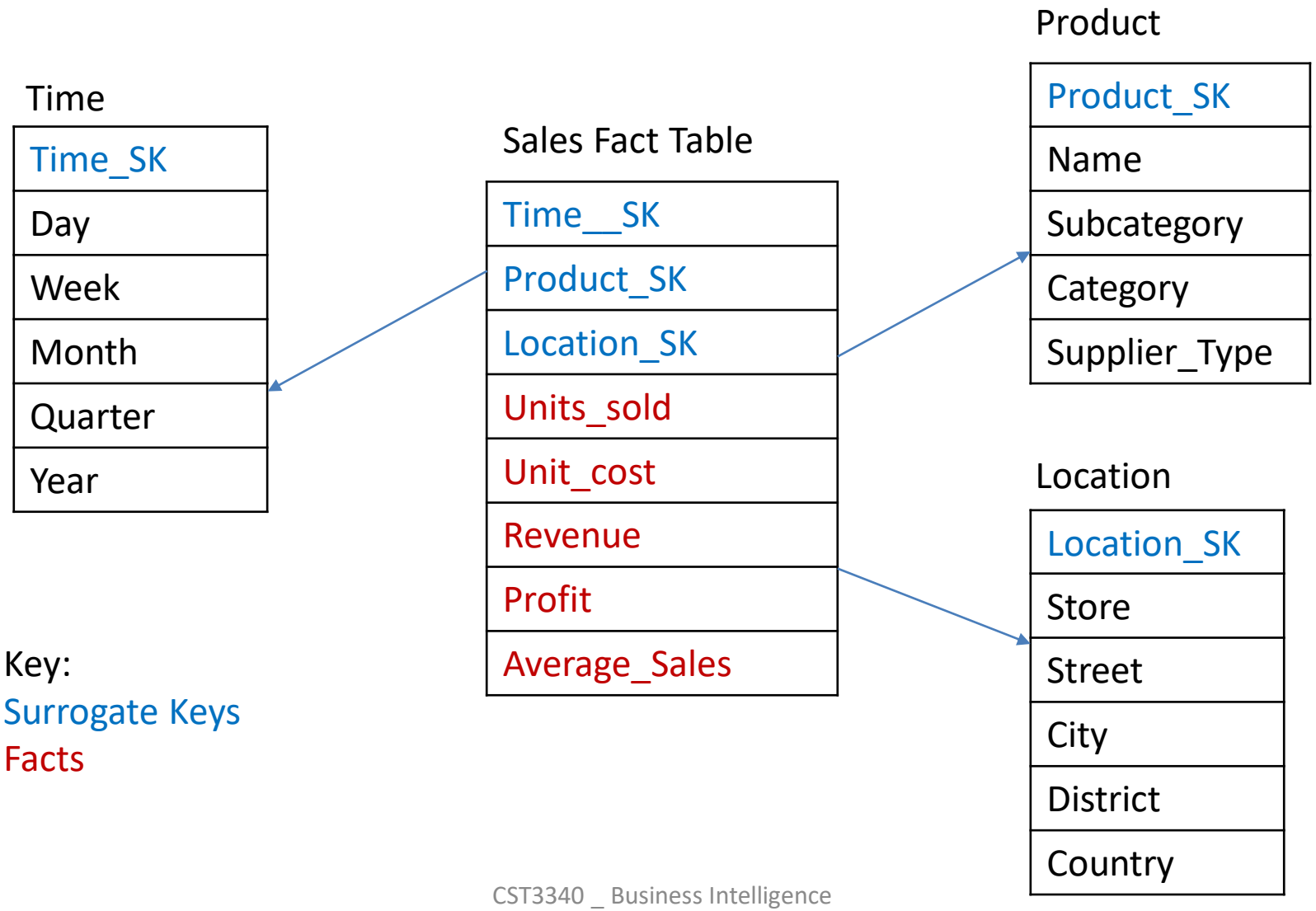
Data Warehouse Development



Data Warehouses – Conceptual Model

- Conceptual Model: uses dimensions & facts
  - Star schema: A single fact table surrounded by a set of dimension tables. Represented by a star shape.
  - Snowflake schema: An extension of a star schema where some dimension tables are split into a set of smaller tables by normalization. Represented by a snowflake shape.
  - Fact constellations (Galaxy): Multiple connected star schemas. Several fact tables share the same dimension tables. Represented as a collection of star shapes.

Example of a Star Schema



Star Schema

- Used to model the data in a Data Warehouse from a decision-makers view of the business.
- Represents a subject e.g. Sales
- One fact table
- Multiple dimension tables
- Allows different views of the business facts
- Allows user to filter, aggregate, drill down & slice and dice the business fact

Fact Tables

- A fact table typically has two types of data:
  - numeric facts (measures) containing data to be analysed.
  - foreign keys linking the dimension tables.
- Facts (measures) can be
  - Detail level data.
  - Data that have been aggregated. E.g. Sum, average etc.
  - Most useful are numeric and additive.
- Each row in a fact table corresponds to an instance of the subject.
- All the measurements in a fact table must be of the same grain which is defined by the dimension tables.

Dimension Tables

- Represent the different views of the business facts (measures).
- Allows users to browse fact data from different angles – e.g. time, item, location.
- Can be used as a filter to minimise the rows of data within a fact table.
- Allow users to aggregate fact data e.g. consider quarterly sales rather than daily sales.
- Allow users to analyse more detailed data e.g. sales at individual stores rather than sale in a particular city.

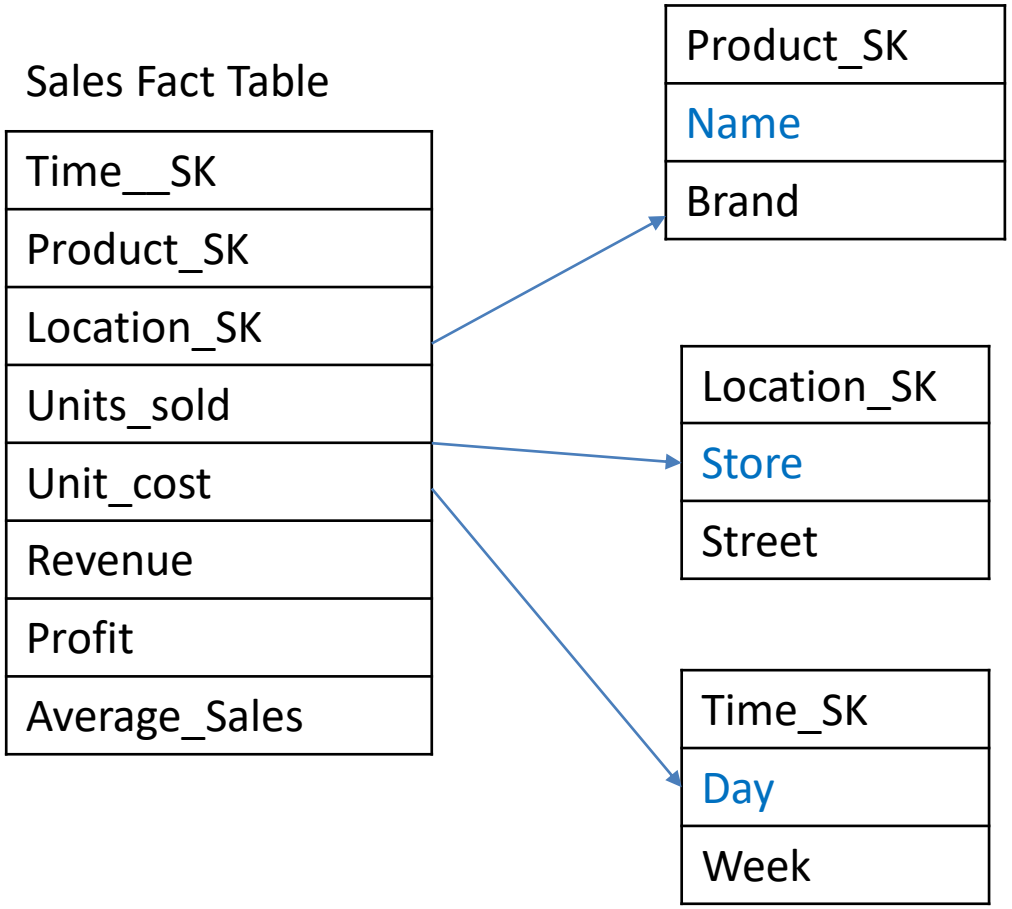
Granularity

- The level of aggregation of the data in the fact table.
- Define by the lowest level of detail in the dimension tables
- E.g. Sales Schema:
  - Time : daily; Location : Individual store; Product : product name.
  - Therefore each row in the sales fact table represents the daily sales of a particular product at individual stores.

Example of a granularity in a Star Schema

Lowest level for each dimension:
Time : daily;
Location : Individual store;
Product : product name.

Therefore each row in the sales fact table represents the daily sales of a particular product at individual stores



Natural keys

- Also known as Production keys, Intelligent keys, Smart keys.
- Natural key can represent the data being stored. E.g. Student Id – M00123456
- Can be imported from the operational systems data.

Surrogate Keys

- Also known as Integer keys, Artificial keys, Non-intelligent keys, Meaningless keys.
- Do not have any meaning about the data.
- Used as the primary keys of the dimension tables.
- Usually generated by the data warehouse as data added to the dimension table.
- Usually sequential numeric numbers.

Surrogate Keys Usage

A surrogate key is used as the unique identifier for the dimension tables.

- Replaces the source data primary keys (business/natural keys)
- Protect against changes in source data systems
- Acts as a buffer between the data warehouse and the source data systems.
- Allows integration from multiple data sources.
- Enable rows that do not exist in the source data.
- Track changes over time (e.g. new customer instances when addresses change)
- Replace text keys with integers for efficiency

Surrogate Keys Usage Cont.

A surrogate key is used as the unique identifier for the dimension tables.

- Appears as foreign keys in the corresponding data warehouse fact table.
- Primary key for the fact table is usually the composite key made up of the foreign keys (surrogate keys) from the dimension tables.
- The fact table may have its own surrogate key.

Advantages of a Surrogate Keys.

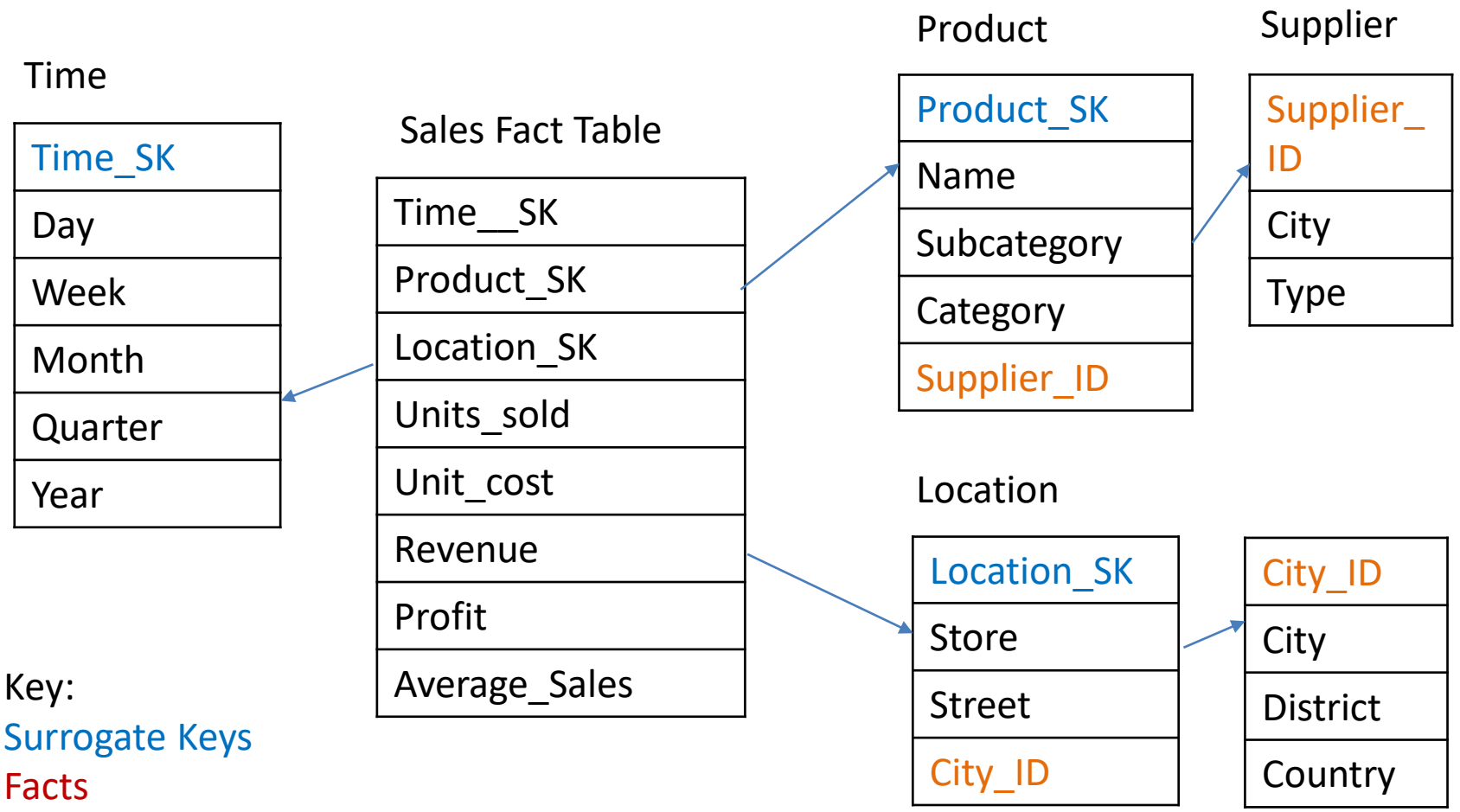
A surrogate key is usually a sequential numeric number.

- Saves storage space.
- Allow for faster joins during data processing,
- Allow for handling slowly changing dimensions.
  - E.g. Allow customers to change billing address. The surrogate key can change while the natural key (Customer ID) remains the same.

Snowflake Schemea

- The snowflake schema is:
  - An extension of the star schema.
  - Dimension tables can be replaced by a set of smaller normalised tables.
  - Allows for more detailed dimensions
  - Reduces storage space
  - Increases processing time (more table joins)

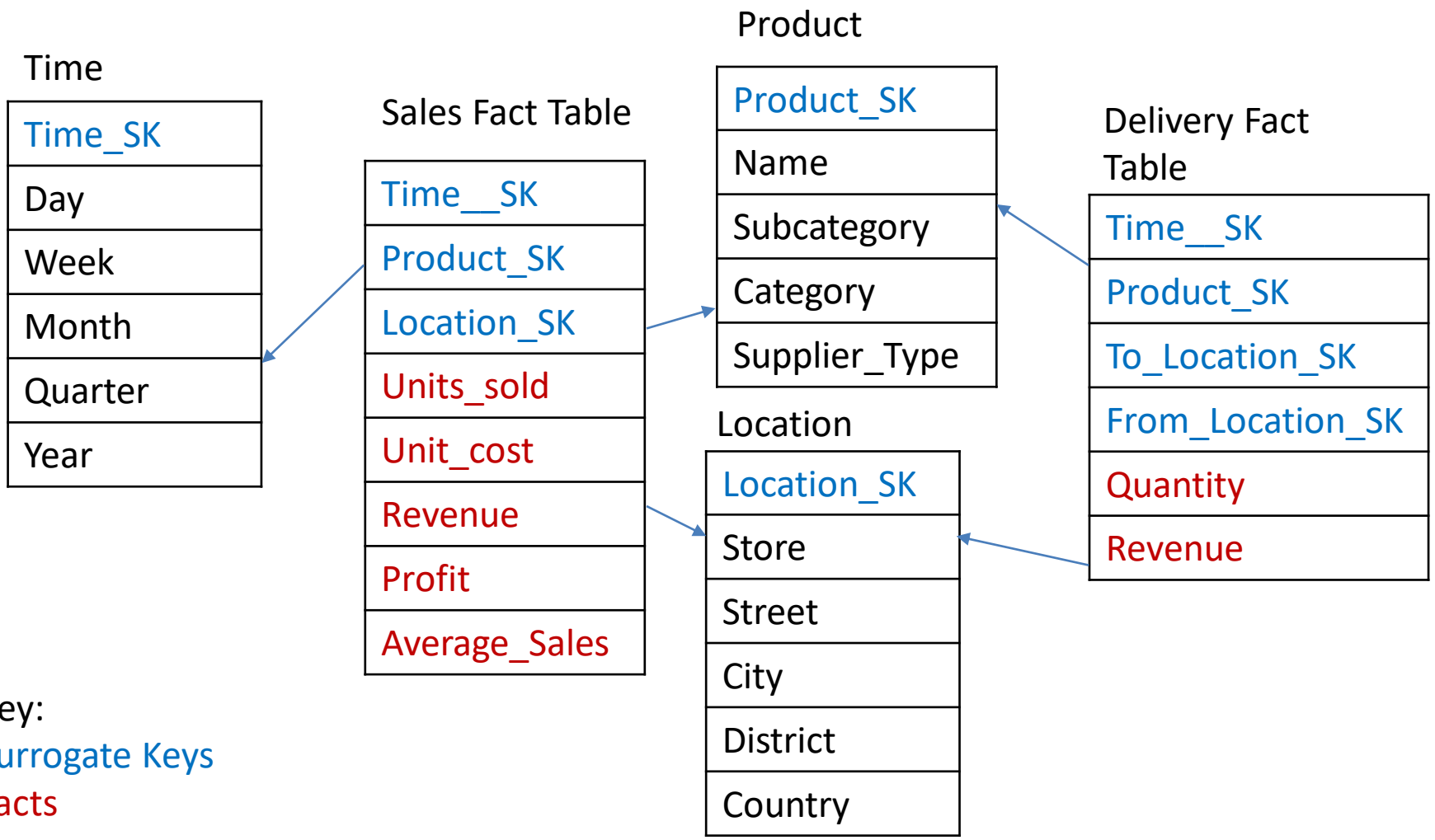
Example of a Snowflake Schema



Fact Constellation

- A fact constellation is:
- Multiple connected star schemas
- Several fact tables share the same dimensions
- Allow a more flexible schema
- More complex queries
- More processing time

Example of Fact Constellation



Key:
Surrogate Keys
Facts

Reading

- Chapter 4, section 4.2:
  - Jiawei Han, Micheline Kamber, Jian Pei, (2011), Data Mining: Concepts and Techniques, Third Edition, The Morgan Kaufmann Series in Data Management System.