Data Warehouse Schemas

Definition of Data Warehouse

A data warehouse is a *subject-oriented*, *integrated*, *time-varying*, *non-volatile* collection of data that is used primarily in organizational decision making.

Data Warehouse models

A data warehouse is usually modeled by a multidimensional database structure, where each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure.

The three data warehouse schemas are:

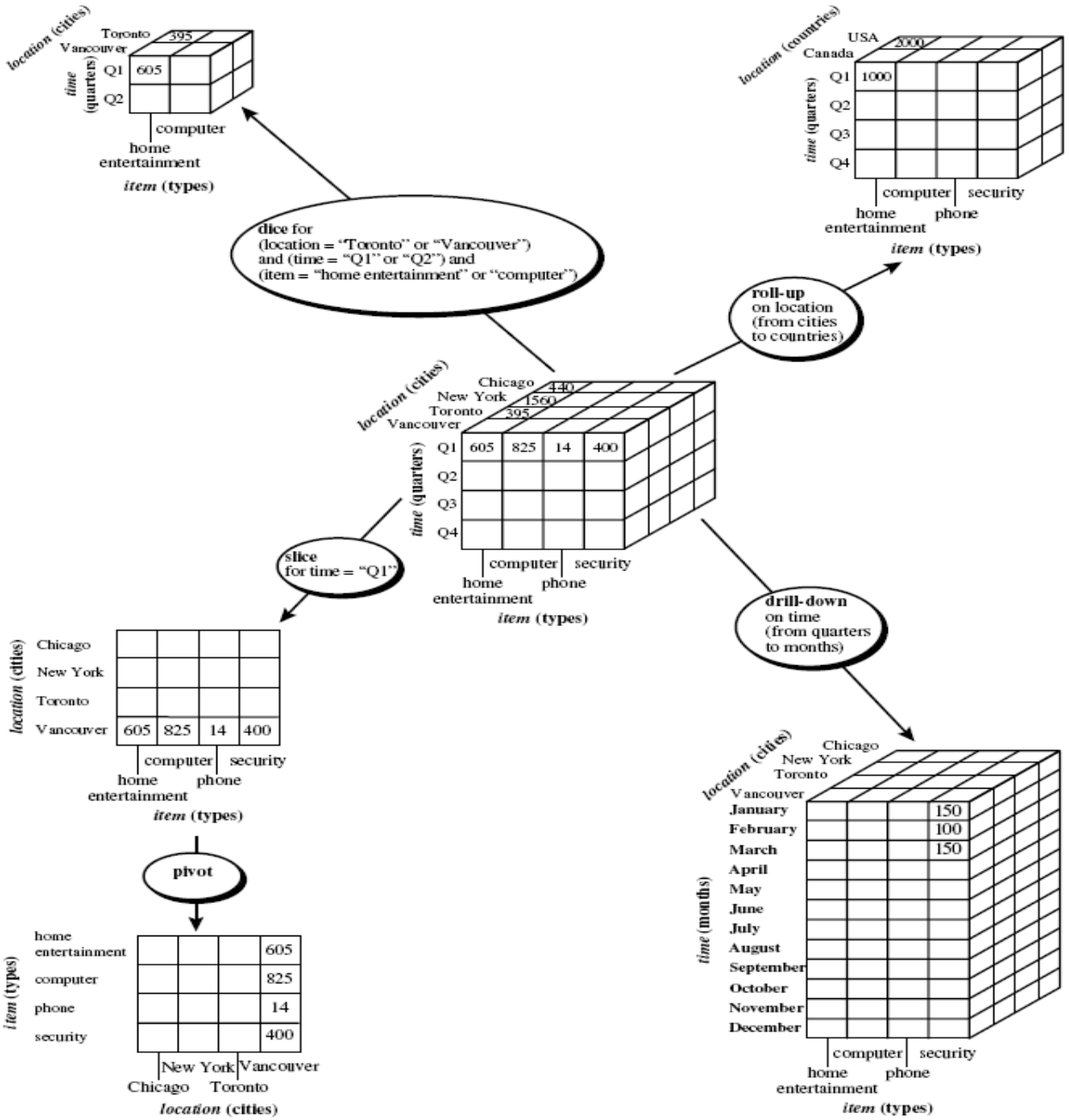
1. Star Schema
2. Snowflake Schema
3. Fact Constellation Schema

OLAP operations

OLAP operations:

1. Drill down
2. Roll up
3. Slice
4. Dice
5. Pivot

OLAP operations



OLAP operations

- Roll-up: The roll-up operation (also called the drill-up operation) performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction.

- Drill-down: Drill-down is the reverse of roll-up. It navigates from less detailed data to more detailed data. Drill-down can be realized by either stepping down a concept hierarchy for a dimension or introducing additional dimensions.

OLAP operations

- Slice: The slice operation performs a selection on one dimension of the given cube, resulting in a subcube.

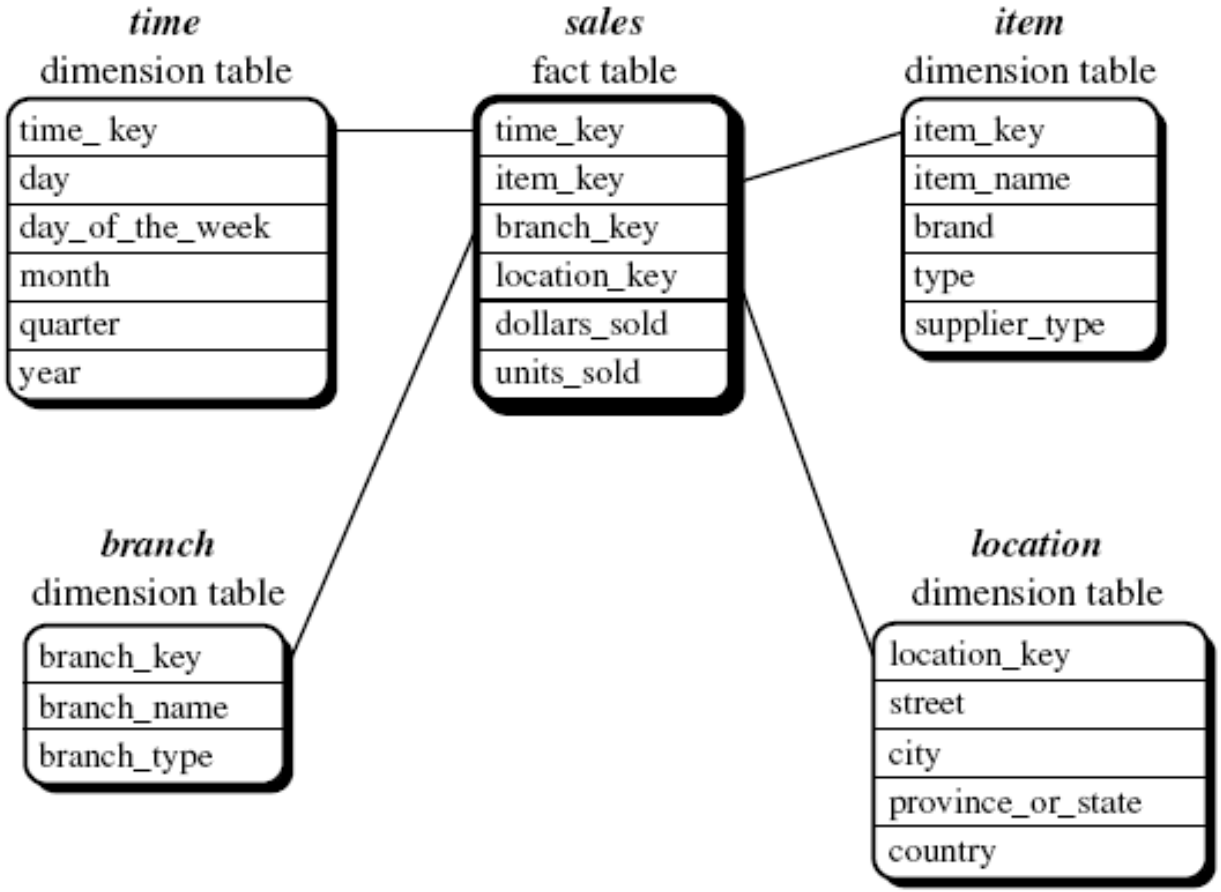
- Dice: The dice operation defines a subcube by performing a selection on two or more dimensions.

- Pivot: Pivot (also called rotate) is a visualization operation that rotates the data axes in view in order to provide an alternative presentation of the data.

Star Schema

- The most common modeling paradigm is the star schema, in which the data warehouse contains (1) a large central table (fact table) containing the bulk of the data, with no redundancy, and (2) a set of smaller attendant tables (dimension tables), one for each dimension.

Star Schema illustration



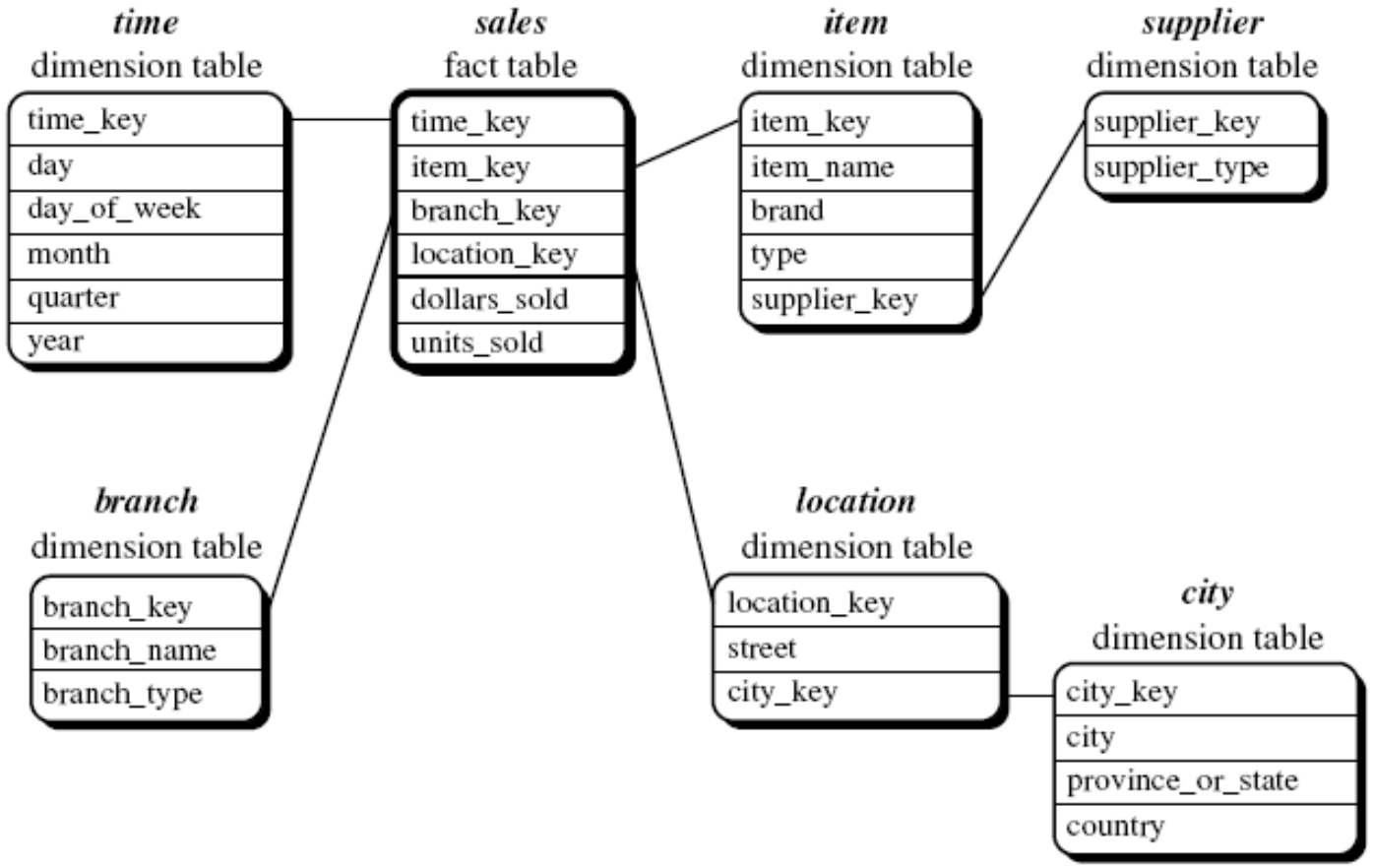
Star schema of a data warehouse for sales

Snowflake Schema

- The snowflake schema is a variant of the star schema model, where some dimension tables are normalized, thereby further splitting the data into additional tables.

- The major difference between the snowflake and star schema models is that the dimension tables of the snowflake model may be kept in normalized form to reduce redundancies.
  - Such a table is easy to maintain and saves storage space.
  - The snowflake structure can reduce the effectiveness of browsing, since more joins will be needed to execute a query. Consequently, the system performance may be adversely impacted.
  - Hence, although the snowflake schema reduces redundancy, it is not as popular as the star schema in data warehouse design.

Snowflake schema illustration

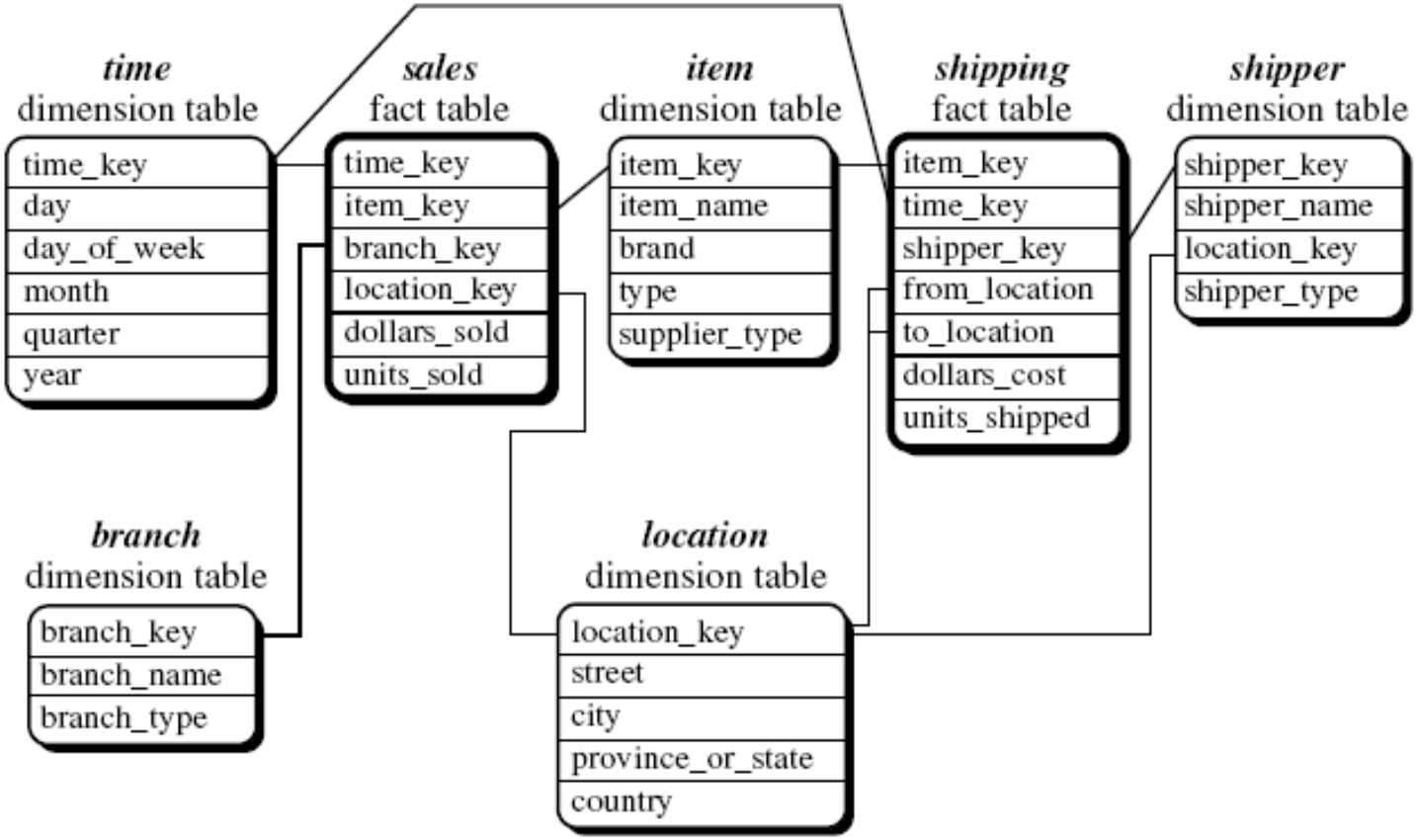


Snowflake schema of a data warehouse for sales

Fact Constellation Schema

- A fact constellation schema allows dimension tables to be shared between fact tables. This kind of schema can be viewed as a collection of stars, and hence is called a galaxy schema or a fact constellation.

Fact constellation schema illustration



Fact constellation schema of a data warehouse for sales and shipping

Data Mart

- A data warehouse collects information about subjects that span an *entire organization*, and thus its scope is *enterprise-wide*. A data mart, is a department subset of a data warehouse. It focuses on selected subjects, and thus its scope is *department-wide*.

- For data warehouses, the fact constellation schema is commonly used, since it can model multiple, interrelated subjects. For data marts, the star or snowflake schema are commonly used.