ĐẠI HỌC ĐÀ NẴNG
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG VIỆT - HÀN
VIETNAM - KOREA UNIVERSITY OF INFORMATION AND COMMUNICATION TECHNOLOGY
한-베정보통신기술대학교

Nhân bản – Phụng sự – Khai phóng

Chapter 2: Conceptual Data Warehouse Design

Data Warehouse

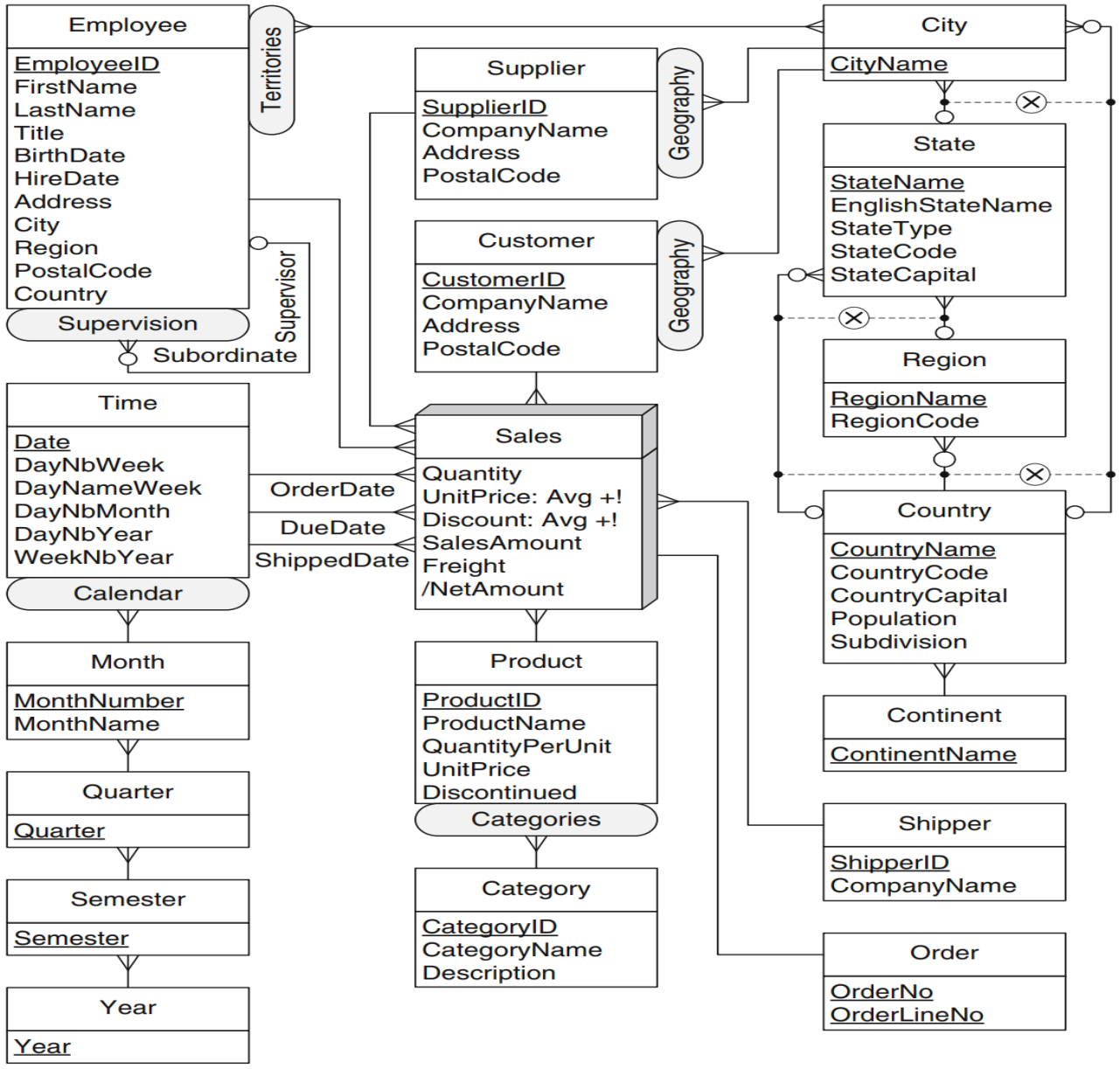
➔ **Conceptual modeling DW** performs a transformation of the *semi-formal business requirements specification* into a formalized *conceptual multidimensional schema* with the highest-level relationships between the different entities.

➔ Characteristics of the **conceptual data model in the DW**:

- It contains the essential entities and the relationships among them.
- No attribute is specified.
- No primary key is specified.

➔ The formalization results in a ***graphical multidimensional diagram***, which comprises ***fact schemata*** with their related ***measures*** and ***dimension hierarchies***.

Example

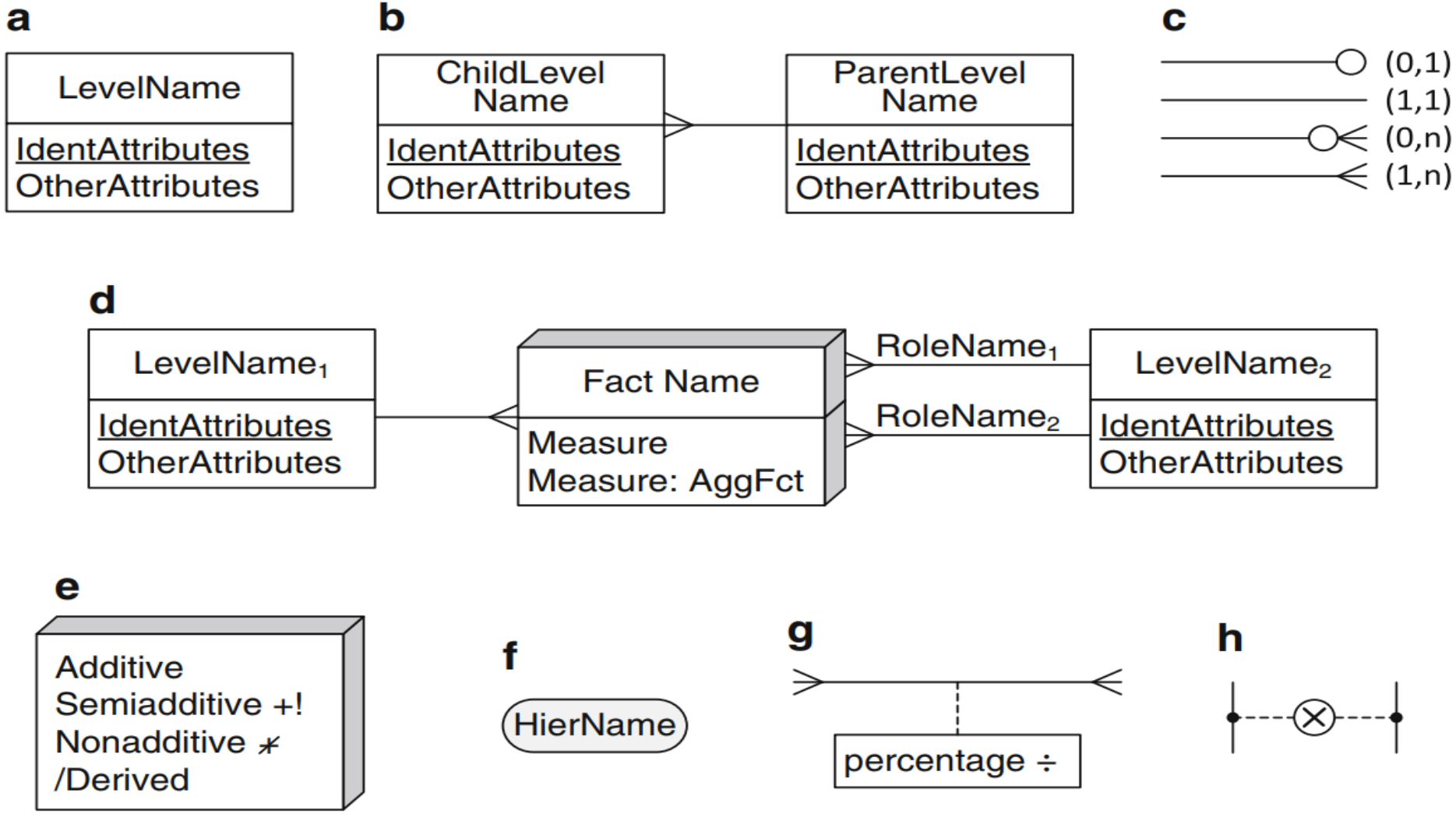


Conceptual schema of the Northwind data warehouse

Conceptual Modeling of Data Warehouses: Terminologies

- A **schema** is composed of ***a set of dimensions*** and ***a set of facts (measures)***.
- A **fact** consists of quantifying ***values stored in measures*** and a qualifying ***context*** which is determined ***through dimension levels***.
- A **dimension** is composed of either ***one level or one or more hierarchies***.
- A **hierarchy** is in turn composed of ***a set of levels.***
- A **level** is ***analogous to an entity type*** in the ER model.
- **Instances of a level** are called ***members.***
- A level has a set of **attributes** that describe the ***characteristics of their members*** (additional information related to a dimension level).

Example



Notation of the MultiDim model. (**a**) Level. (**b**) Hierarchy. (**c**) Cardinalities. (**d**) Fact with measures and associated levels. (**e**) Types of measures. (**f**) Hierarchy name. (**g**) Distributing attribute. (**h**) Exclusive relationships

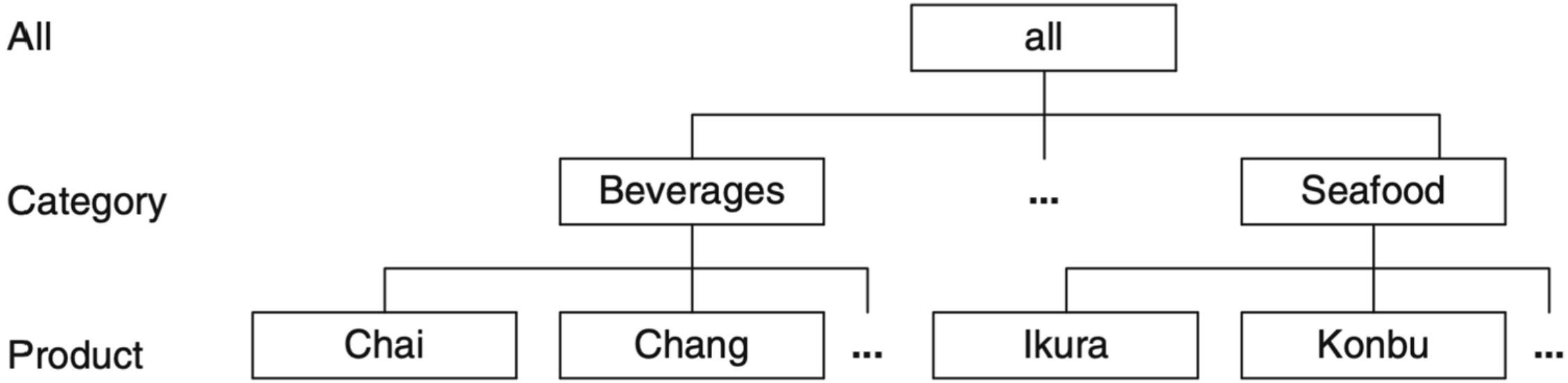
➔ are key elements in analytical applications, since they provide the means to represent the data under analysis at different abstraction levels.

➔ There exist many kinds of hierarchies.

- Balanced Hierarchies
- Unbalanced Hierarchies
- Generalized Hierarchies
- Alternative Hierarchies
- Parallel Hierarchies
- Nonstrict Hierarchies

A **balanced hierarchy** has only one path at the schema level, where all levels are mandatory.

Example: balanced hierarchy in which there is the same number of levels from each individual product to the root of the hierarchy.



Members of a hierarchy Product → Category

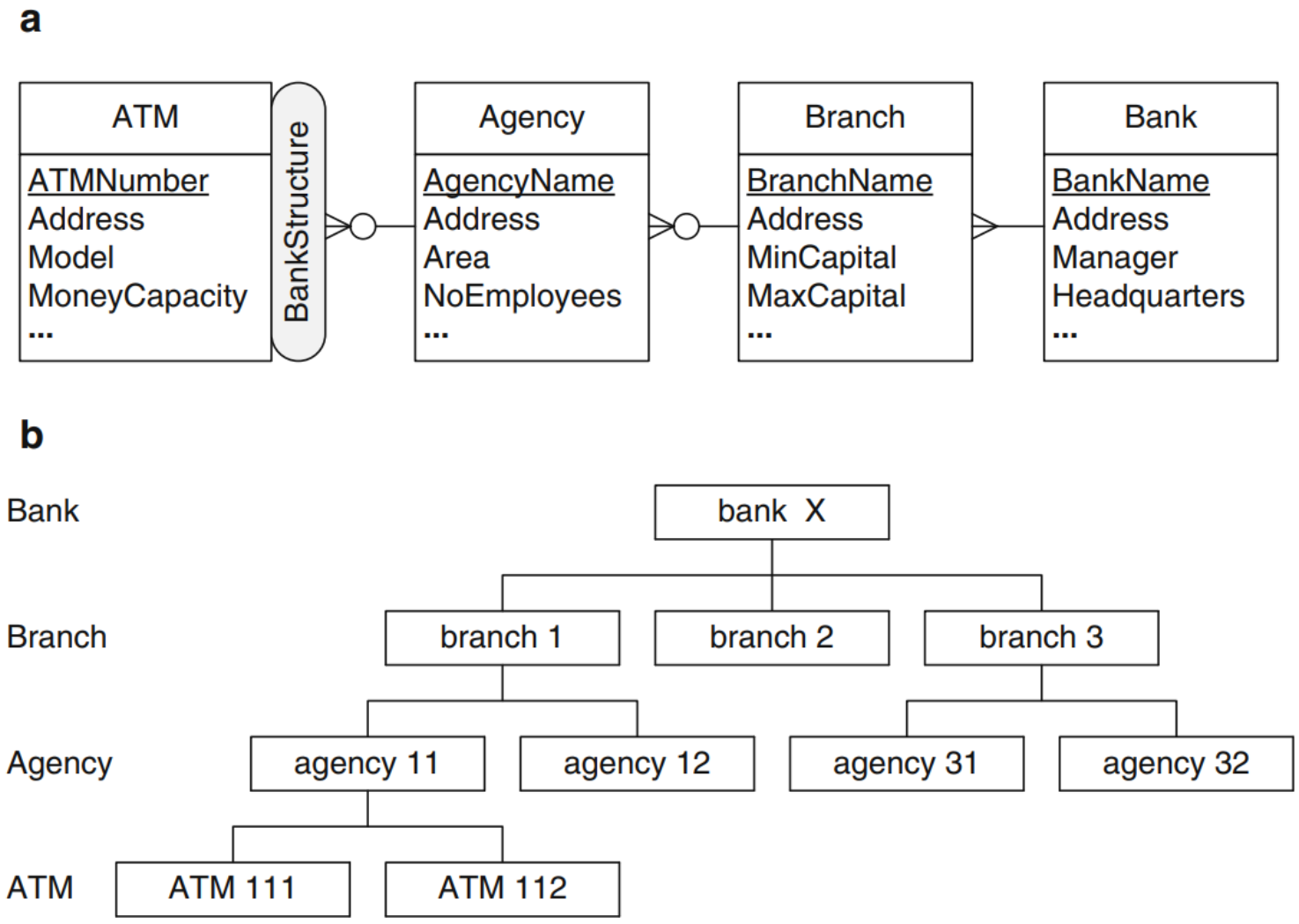
- ➔ **An unbalanced hierarchy** has only one path at the schema level, where at least one level is not mandatory.
- ➔ At the instance level, there can be parent members without associated child members.
- ➔ Unbalanced hierarchies include a special case that is recursive

hierarchies, also called parent-child hierarchies where the same level is linked by the two roles of a parent-child relationship.

Example

a. A hierarchy schema in which a bank is composed of several branches, where a branch may have agencies; further, an agency may have ATMs.

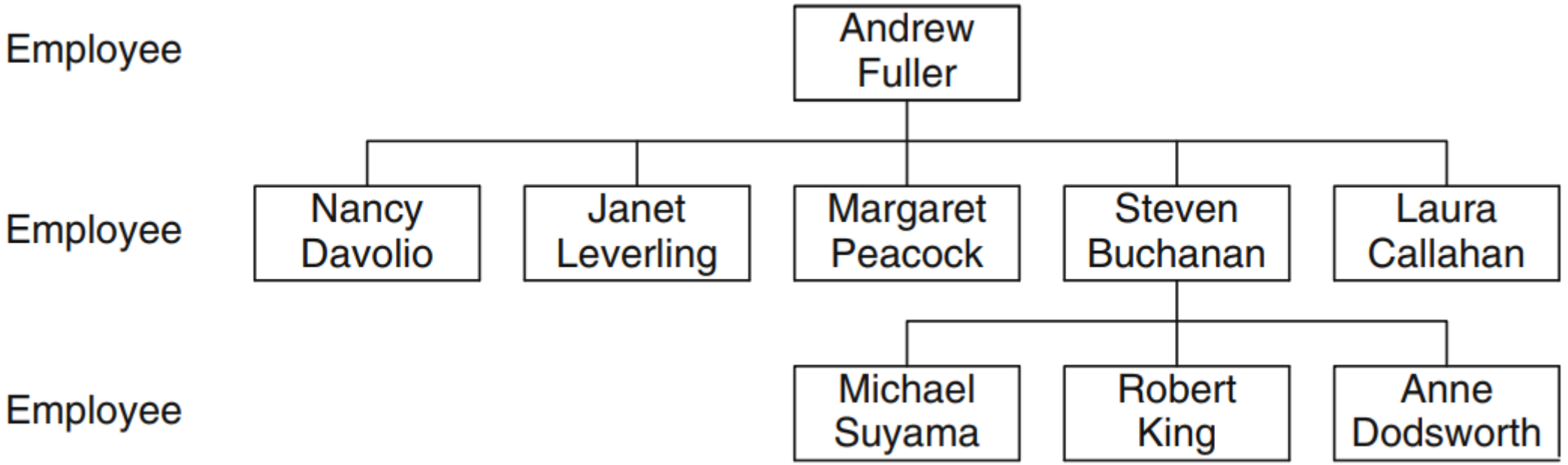
b. A branch with no agency and several agencies with no ATM.



An unbalanced hierarchy. (**a**) Schema. (**b**) Examples of instances

Example

An unbalanced hierarchy where employees with no subordinate will not have descendants in the instance tree.

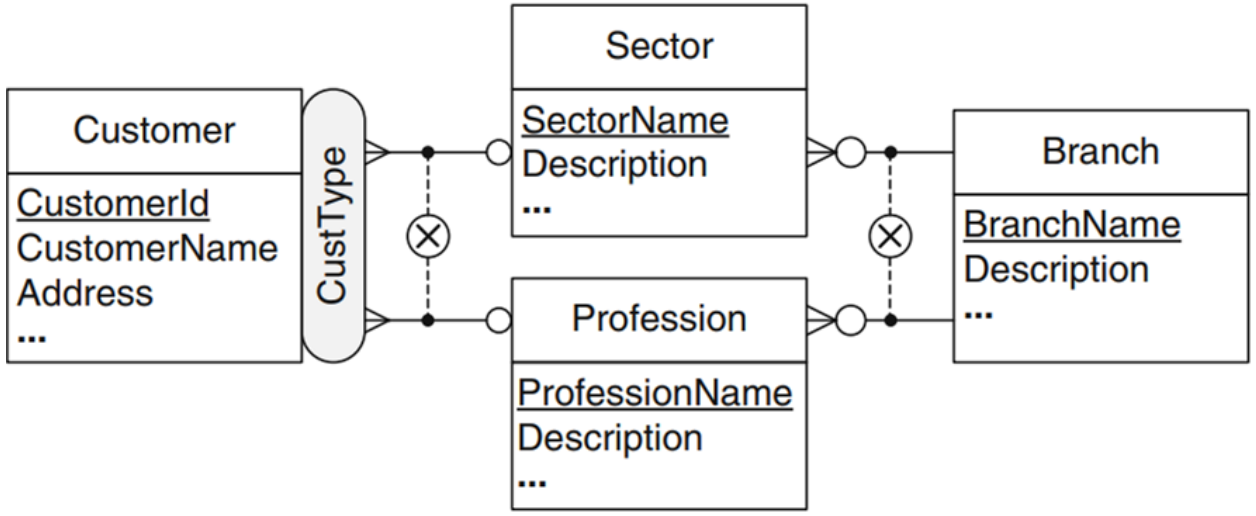


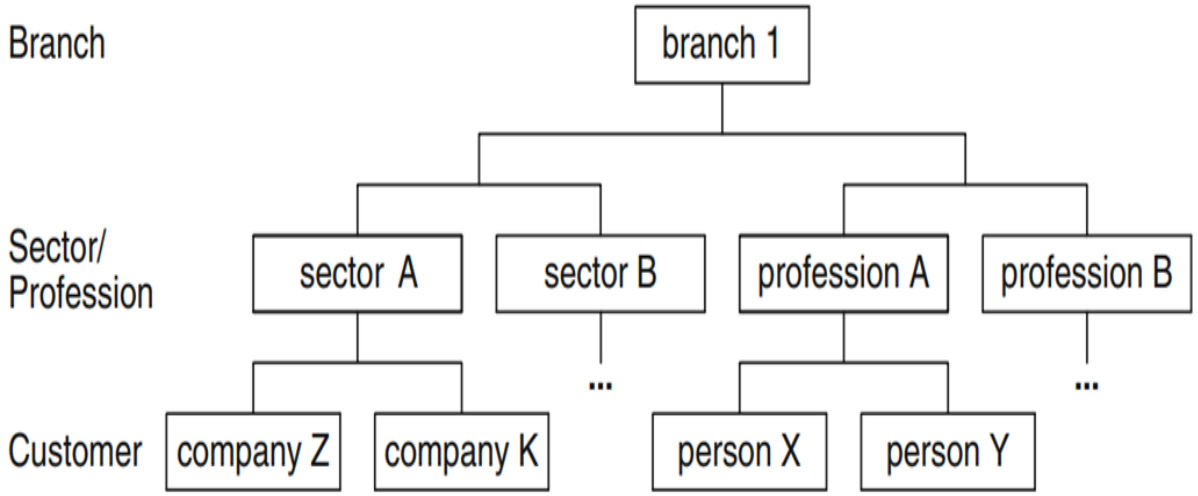
Hierarchies are called **generalized hierarchies** when the members of a level are of different types.

Example:

Customers can be either companies or persons. measures pertaining

to customers must be aggregated differently according to the customer type, e.g. for companies the aggregation path is Customer → Sector → Branch, while for persons it is Customer → Profession → Branch





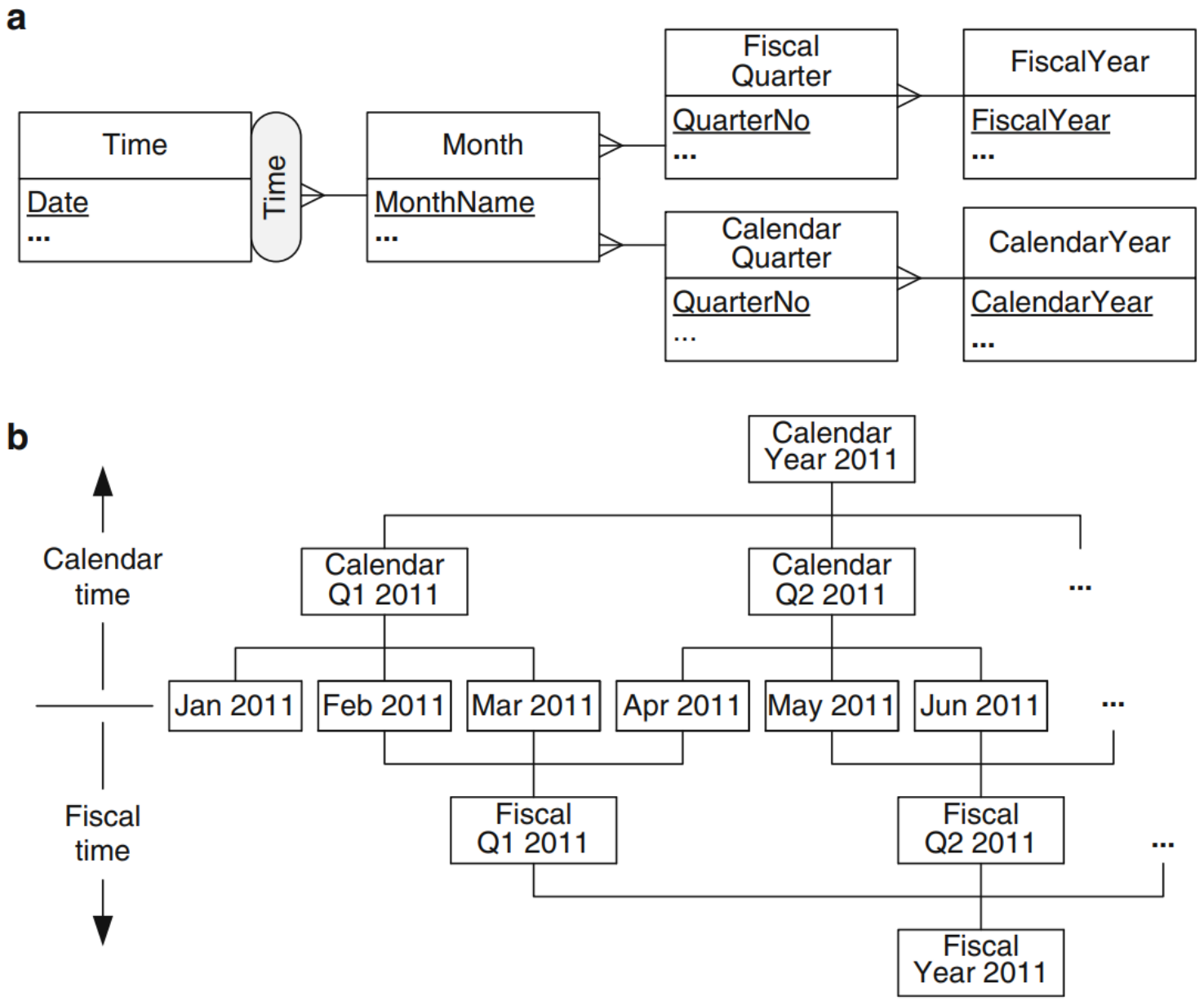
Alternative Hierarchies

- **Alternative hierarchies** represent the situation where at the schema level, there are several nonexclusive hierarchies that share at least the leaf level.
- A child member is associated with more than one parent member and these parent members belong to different levels.
- Alternative hierarchies are needed when we want to analyze measures from a unique perspective (e.g., time) using alternative aggregations.
- In a generalized hierarchy, a child member is related to only one of the paths, whereas in an alternative hierarchy, a child member is related to all paths, and the user must choose one of them for analysis

Example

a. The Time dimension includes two hierarchies corresponding to different groupings of months into calendar years and fiscal years.

b. an instance of the dimension where it is supposed that fiscal years begin in February.

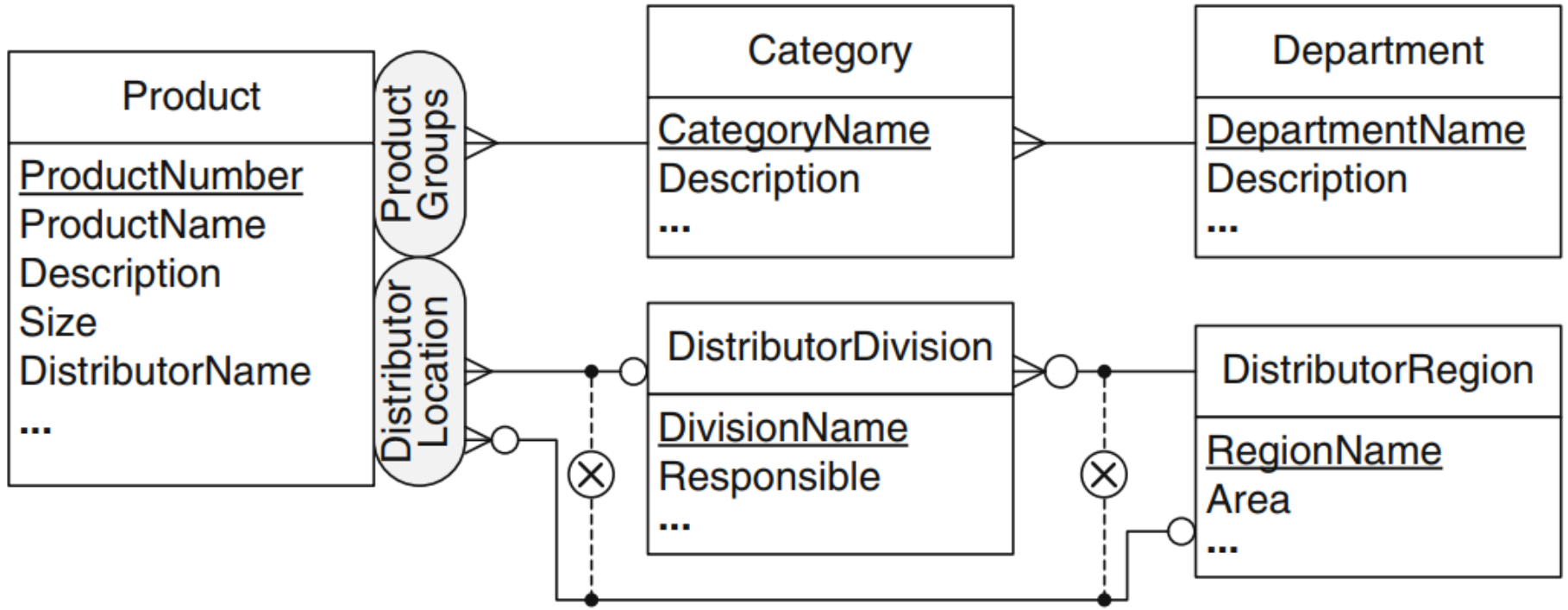


➔ **Parallel hierarchies** arise when a dimension has several hierarchies associated with it, accounting for different analysis criteria.

➔ The component hierarchies may be of different kinds.

➔ Parallel hierarchies can be dependent or independent depending on whether the component hierarchies share levels.

An example of a dimension that has two parallel independent hierarchies. The hierarchy ProductGroups is used for grouping products according to categories or departments, while the hierarchy DistributorLocation groups them according to distributors' divisions or regions.

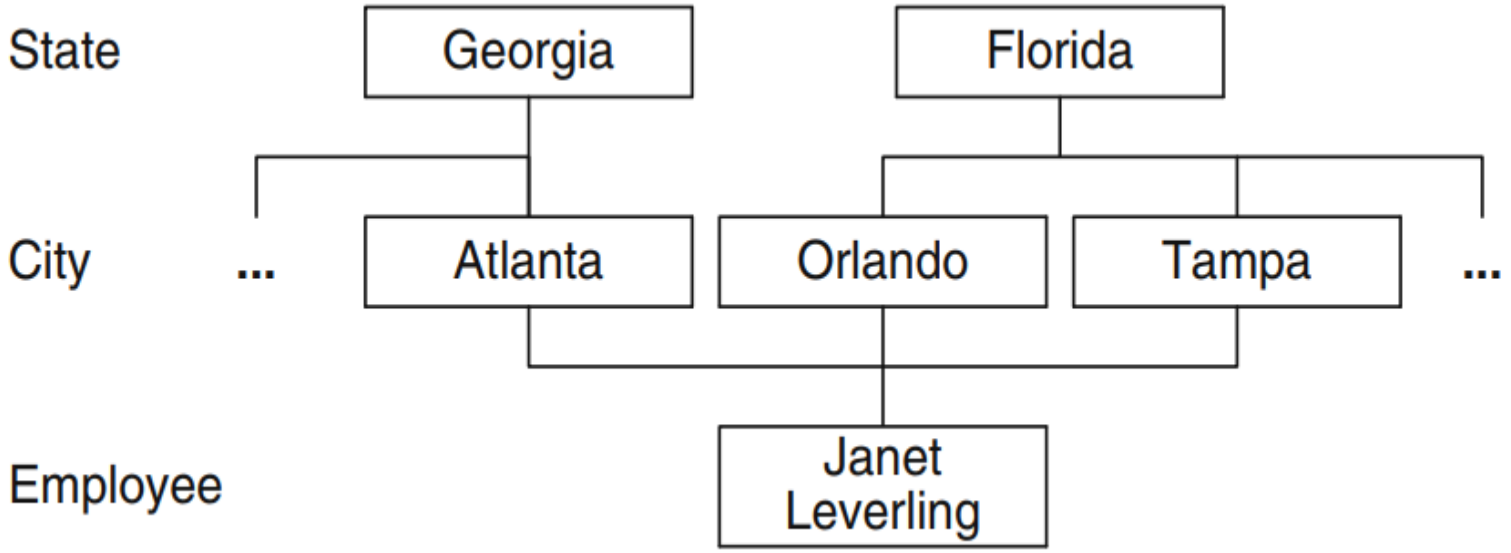


➔ A hierarchy that has at least one many-to-many relationship is called **nonstrict**; otherwise, it is called strict.

➔ The fact that a hierarchy is strict or not is orthogonal to its kind

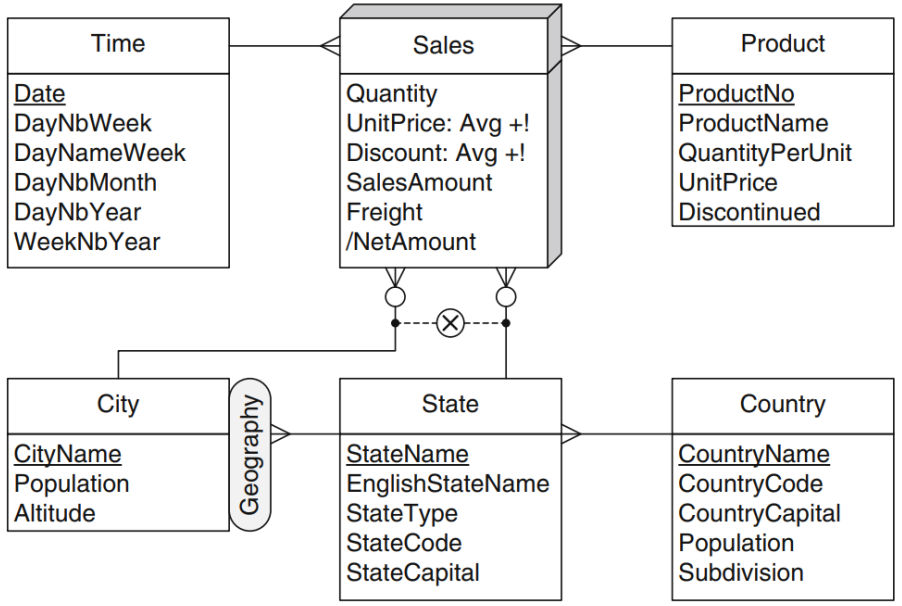
Example:

a nonstrict hierarchy where an employee may be assigned to several cities.

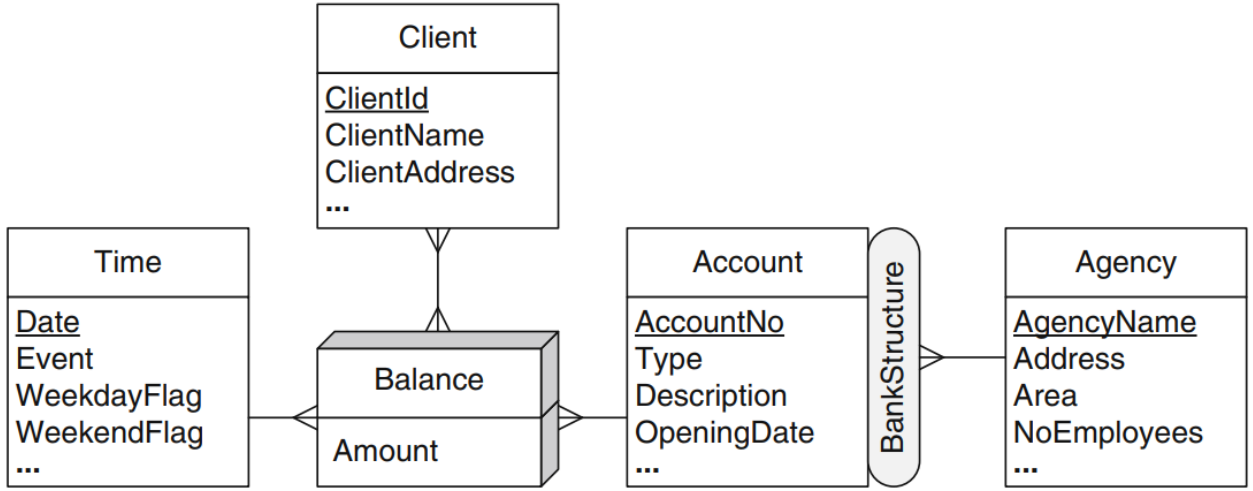


➔ measures are captured at multiple granularities.

Example: sales for the USA might be reported per state, while European sales might be reported percity

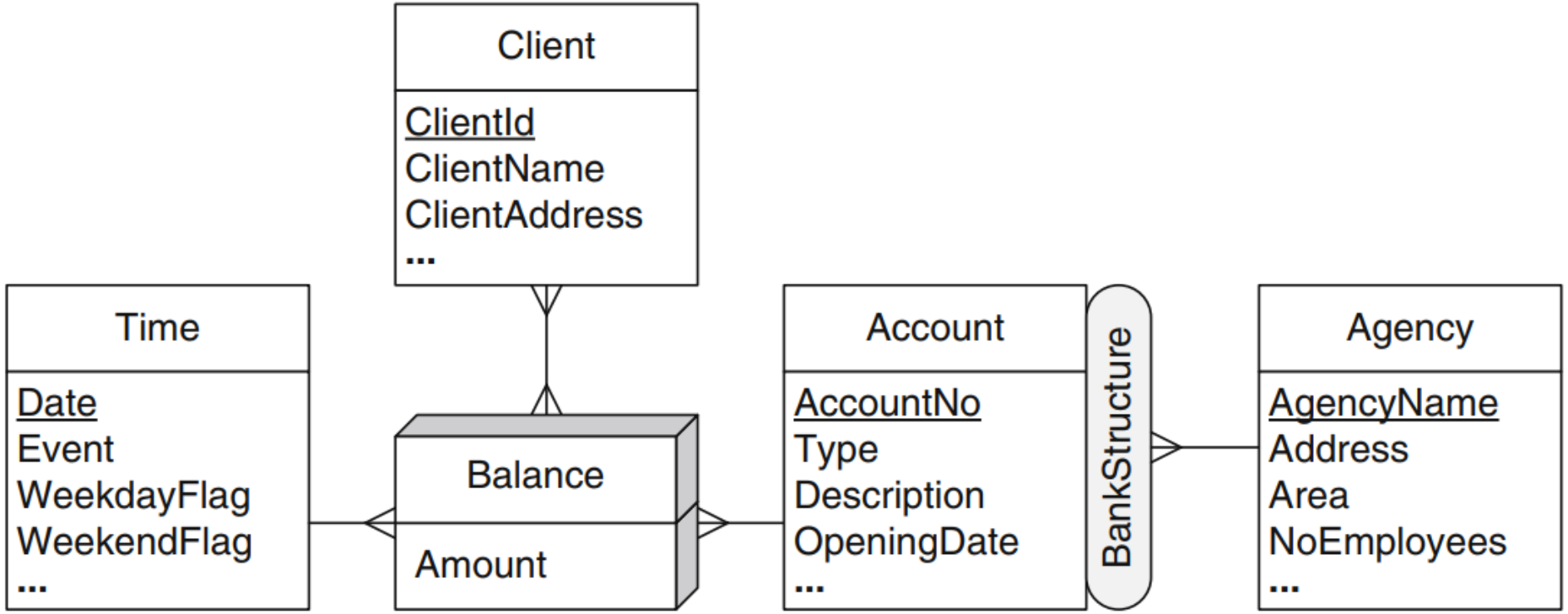


Example: consider a medical data warehouse for analyzing patients, where there is a diagnosis dimension with levels diagnosis, diagnosis family, and diagnosis group.



Many-to-Many Dimensions

➔ In a many-to-many dimension, several members of the dimension participate in the same fact member.



Multidimensional schema for the analysis of bank accounts

Querying the Northwind Cube Using the OLAP Operations

- **Query 4.1.** Total sales amount per customer, year, and product category.
- **Query 4.2.** Yearly sales amount for each pair of customer country and supplier countries.
- **Query 4.3.** Monthly sales by customer state compared to those of the previous year.
- **Query 4.4.** Monthly sales growth per product, that is, total sales per product compared to those of the previous month.
- **Query 4.5.** Three best-selling employees.
- **Query 4.6.** Best-selling employee per product and year.
- **Query 4.7.** Countries that account for top 50% of the sales amount.
- **Query 4.8.** Total sales and average monthly sales by employee and year.
- **Query 4.9.** Total sales amount and total discount amount per product and month.

Querying the Northwind Cube Using the OLAP Operations

- **Query 4.10.** Monthly year-to-date sales for each product category.
- **Query 4.11.** Moving average over the last 3 months of the sales amount by product category.
- **Query 4.12.** Personal sales amount made by an employee compared with the total sales amount made by herself and her subordinates during 1997.
- **Query 4.13.** Total sales amount, number of products, and sum of the quantities sold for each order.
- **Query 4.14.** For each month, total number of orders, total sales amount, and average sales amount by order.
- **Query 4.15.** For each employee, total sales amount, number of cities, and number of states to which she is assigned.

Enjoy the Course…!