Normalized model vs dimensional model

Relational modeling (Codd, '69)

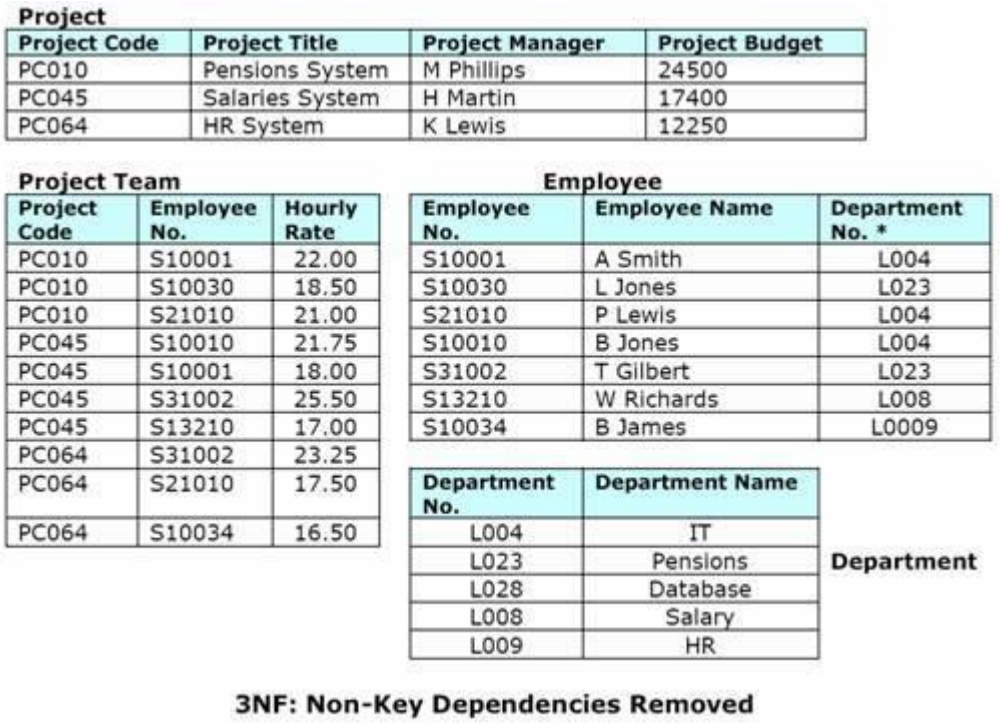
- relation = a set of tuples that have the same attributes
- primary key (a column or group of columns to uniquely identify each row)
- foreign key (a column or group of columns in one table that points to the primary key of another table)
- index = a way of providing quicker access to data (B+ trees, R-trees, bitmaps)

Relational modeling

- Relational operators
  - Union
  - Intersection
  - Difference
  - Cartesian product
  - Selection
  - Projection
  - Join

Relational modeling

- Normalization
  - 3NF



Project

Project Code	Project Title	Project Manager	Project Budget
PC010	Pensions System	M Phillips	24500
PC045	Salaries System	H Martin	17400
PC064	HR System	K Lewis	12250

Project Team

Project Code	Employee No.	Hourly Rate
PC010	S10001	22.00
PC010	S10030	18.50
PC010	S21010	21.00
PC045	S10010	21.75
PC045	S10001	18.00
PC045	S31002	25.50
PC045	S13210	17.00
PC064	S31002	23.25
PC064	S21010	17.50
PC064	S10034	16.50

Employee

Employee No.	Employee Name	Department No. *
S10001	A Smith	L004
S10030	L Jones	L023
S21010	P Lewis	L004
S10010	B Jones	L004
S31002	T Gilbert	L023
S13210	W Richards	L008
S10034	B James	L0009

Department

Department No.	Department Name
L004	IT
L023	Pensions
L028	Database
L008	Salary
L009	HR

3NF: Non-Key Dependencies Removed

  - "[Every] non-key [attribute] must provide a fact about the key, the whole key, and nothing but the key" … ”so help me Codd”

Dimensional modeling (Kimball)

- **Fact tables**
  - the primary table in a dimensional model where the numerical performance measurements of the business are stored
  - The most useful facts are **numeric** and **additive**

Daily Sales Fact Table
Date Key (FK)
Product Key (FK)
Store Key (FK)
Quantity Sold
Dollar Sales Amount

Figure 1.2 Sample fact table.

Dimensional modeling

- **Dimension tables**
  - contain the textual descriptors of the business
  - entry points into the fact table

Product Dimension Table
Product Key (PK)
Product Description
SKU Number (Natural Key)
Brand Description
Category Description
Department Description
Package Type Description
Package Size
Fat Content Description
Diet Type Description
Weight
Weight Units of Measure
Storage Type
Shelf Life Type
Shelf Width
Shelf Height
Shelf Depth
... and many more

Dimensional modeling

- **Bringing together facts and dimensions**

STAR SCHEMA

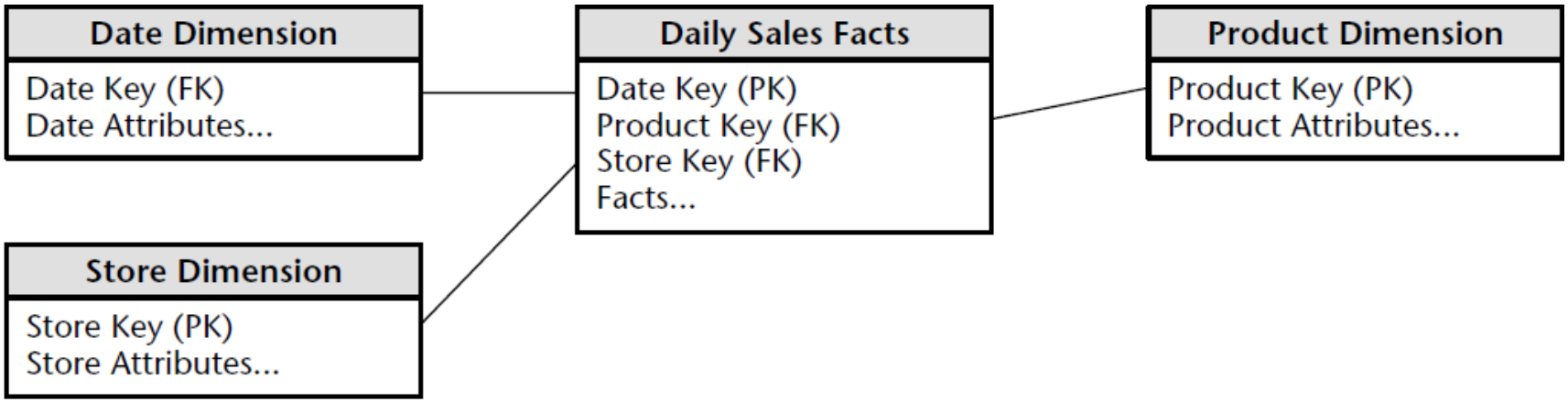


Figure 1.4 Fact and dimension tables in a dimensional model.

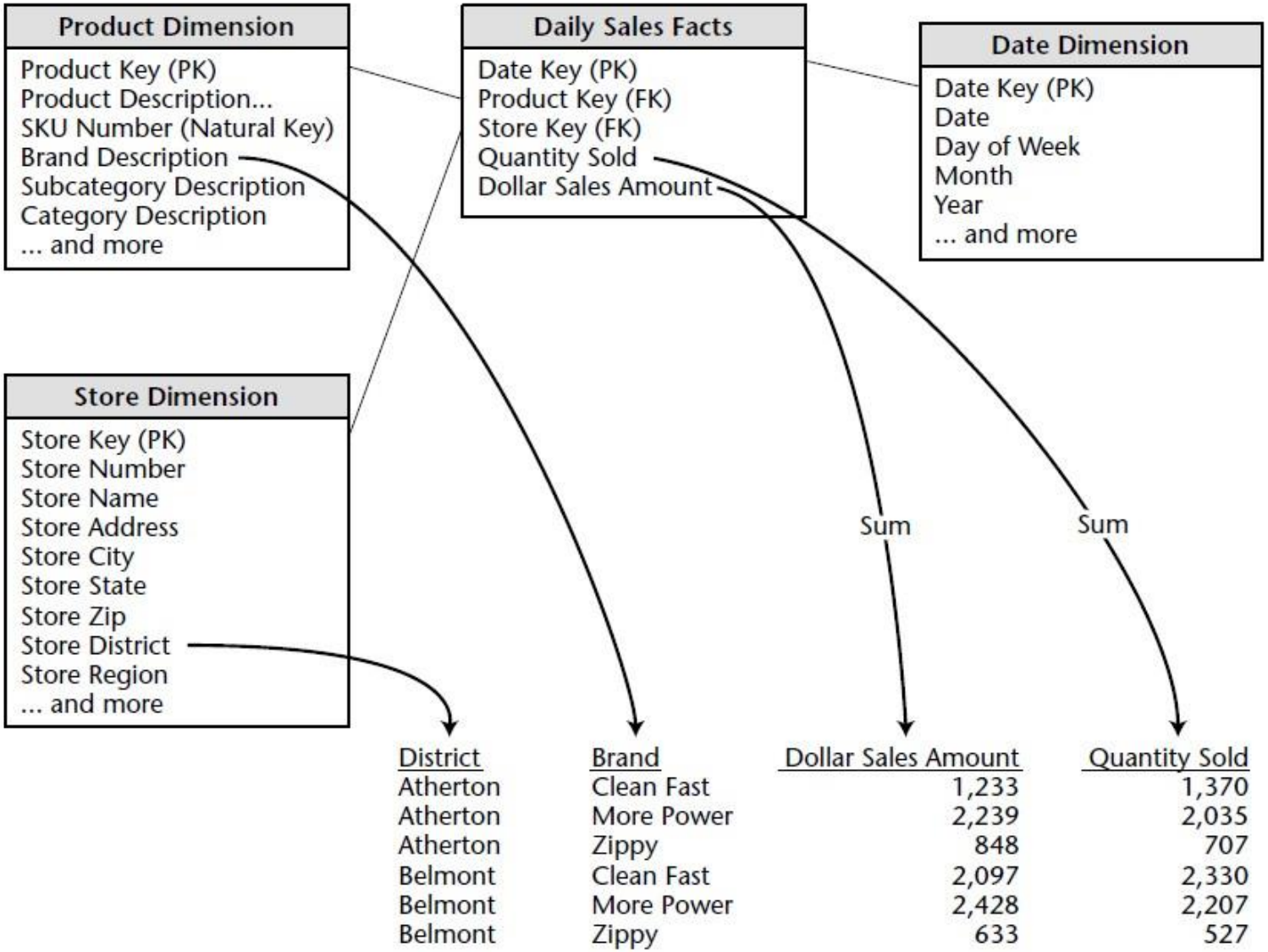


Figure 1.5 Dragging and dropping dimensional attributes and facts into a simple report.

Four-Step Dimensional Design Process

1. Select the business process to model.
2. Declare the grain of the business process.
3. Choose the dimensions that apply to each fact table row.

   "How do business people describe the data that results from the business process?"
4. Identify the numeric facts that will populate each fact table row.

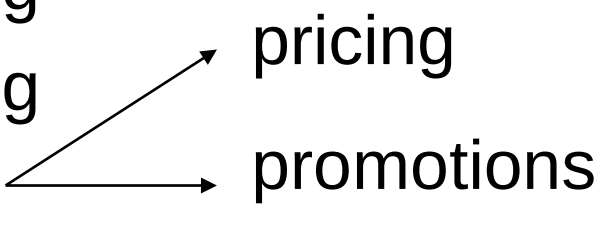
   "What are we measuring?"

Retail Case Study

- 100 grocery stores spread over a five-state area
- departments: including grocery, frozen foods, dairy, meat, produce, bakery, floral, and health/beauty aids
- roughly 60,000 individual products
- data collection:
  - at the cash registers as customers purchase products
  - at the back door, where vendors make deliveries

Retail Case Study

- management is concerned with the logistics of:
  - ordering
  - stocking
  - selling



- while **maximizing profit**

Step 1. Select the Business Process

- **The first dimensional model built should be the one with the most impact—it should answer the most pressing business questions and be readily accessible for data extraction.**

- In our case study, management wants to better understand customer purchases as captured by the POS system

Step 2. Declare the Grain

- **Preferably you should develop dimensional models for the most atomic information captured by a business process.**

- In our case study, the most granular data is an individual line item on a POS transaction

Step 3. Choose the Dimensions

- **A careful grain statement determines the primary dimensionality of the fact table.**

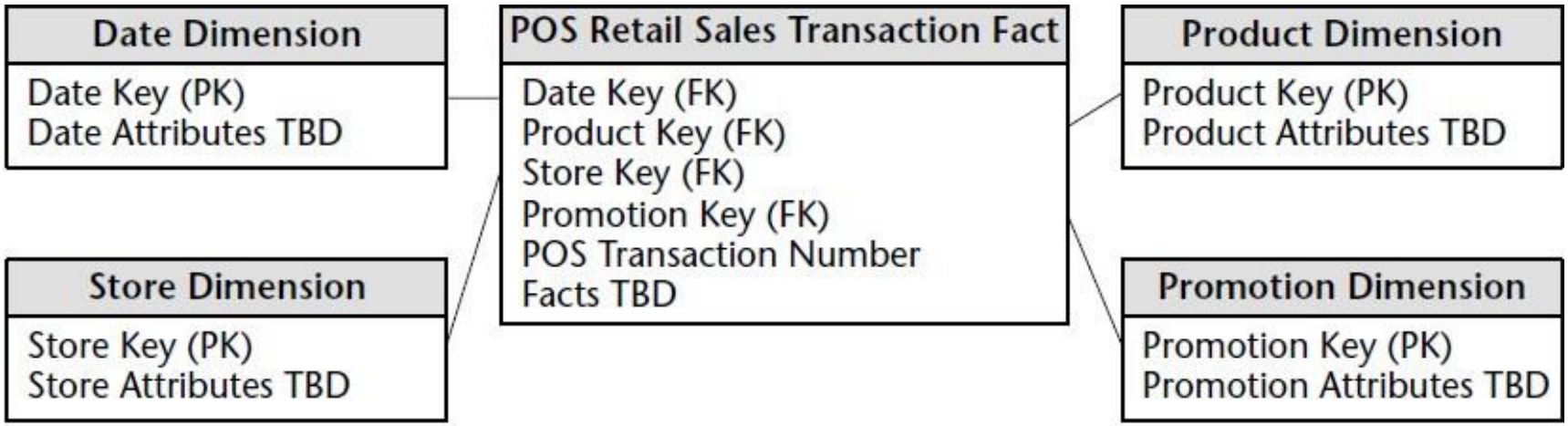


Figure 2.2 Preliminary retail sales schema.

"TBD" means "to be determined."

Step 4. Identify the Facts

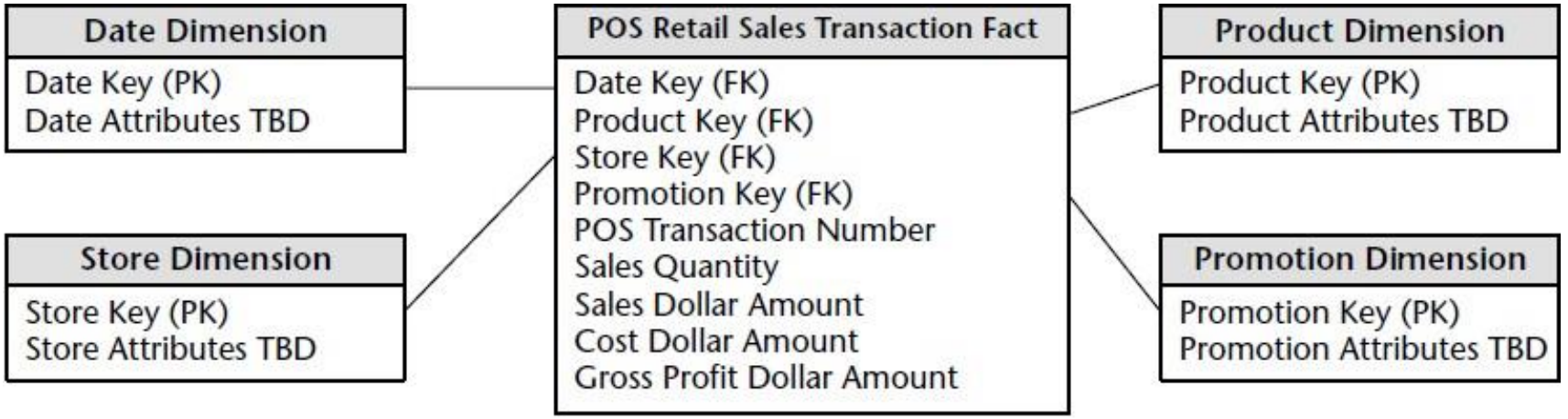


Figure 2.3 Measured facts in the retail sales schema.

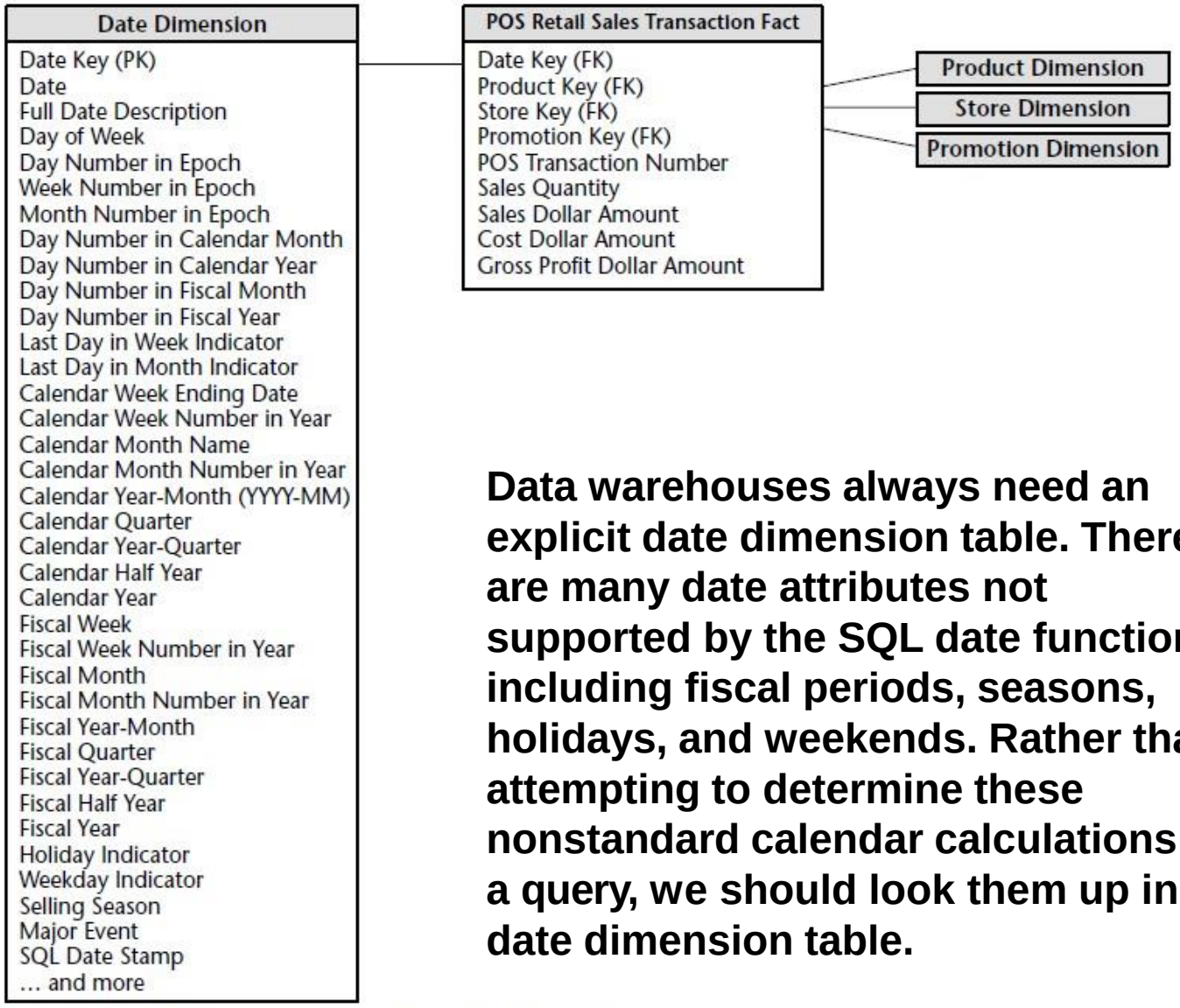


Figure 2.4 Date dimension in the retail sales schema.

Data warehouses always need an explicit date dimension table. There are many date attributes not supported by the SQL date function, including fiscal periods, seasons, holidays, and weekends. Rather than attempting to determine these nonstandard calendar calculations in a query, we should look them up in a date dimension table.

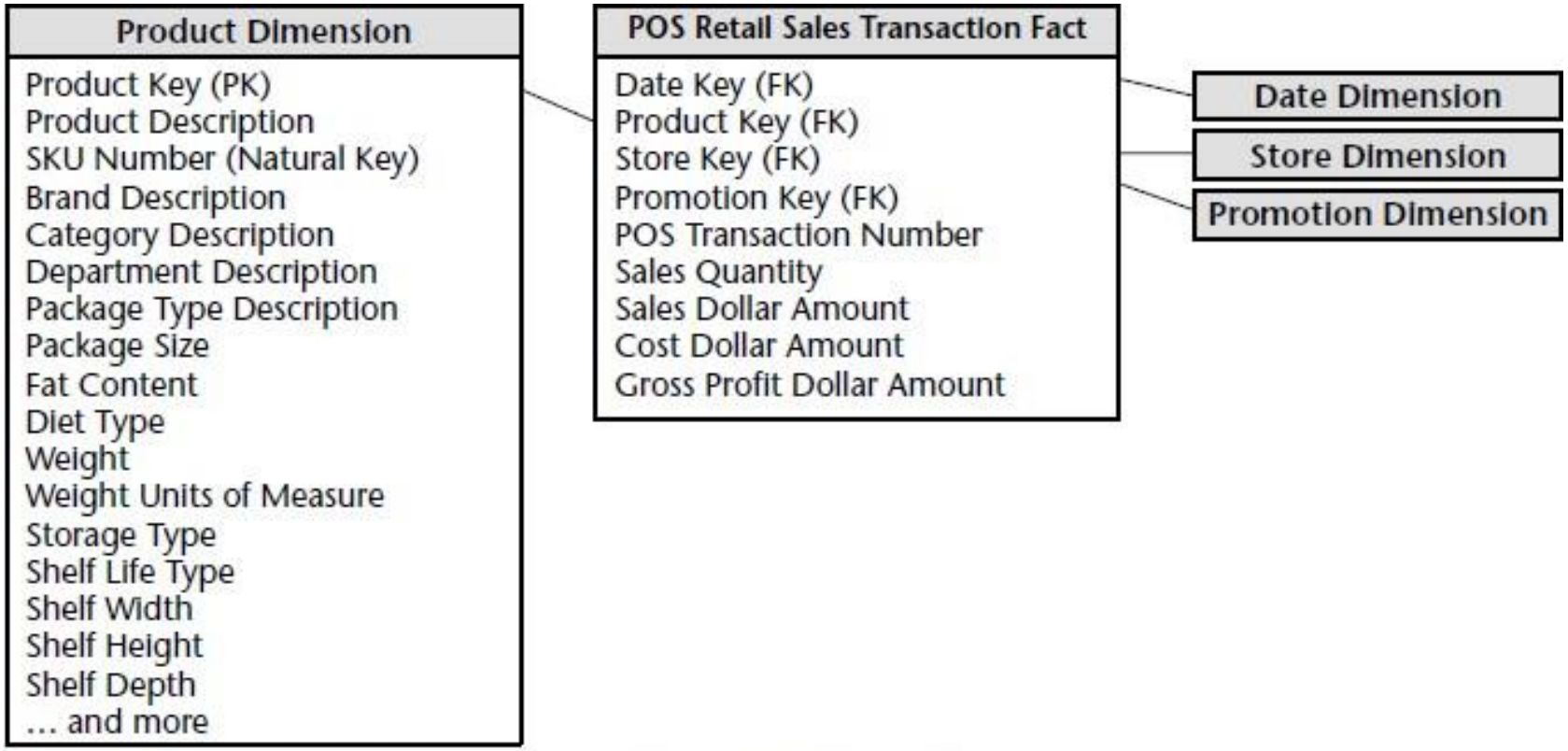


Figure 2.7 Product dimension in the retail sales schema.

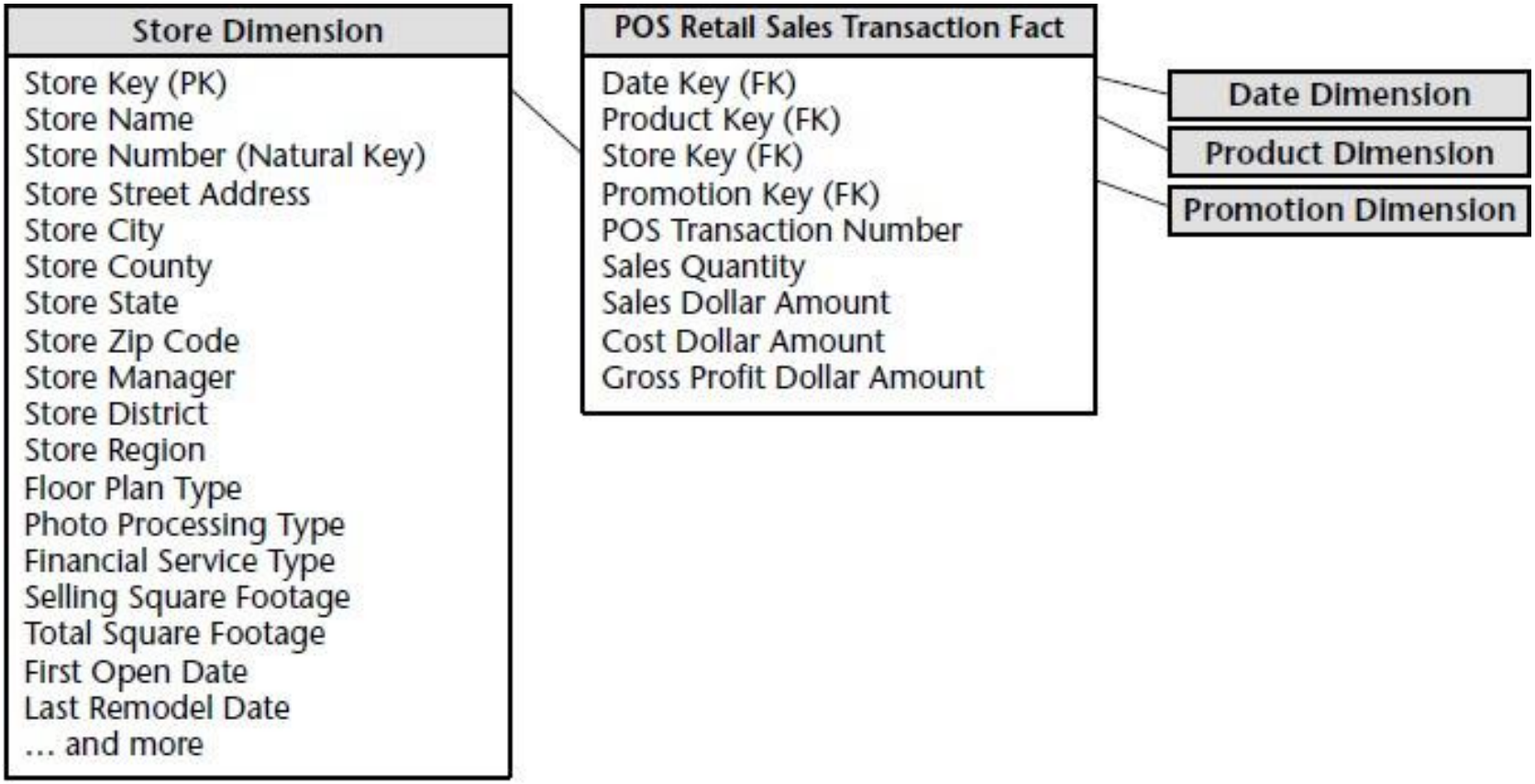


Figure 2.8 Store dimension in the retail sales schema.

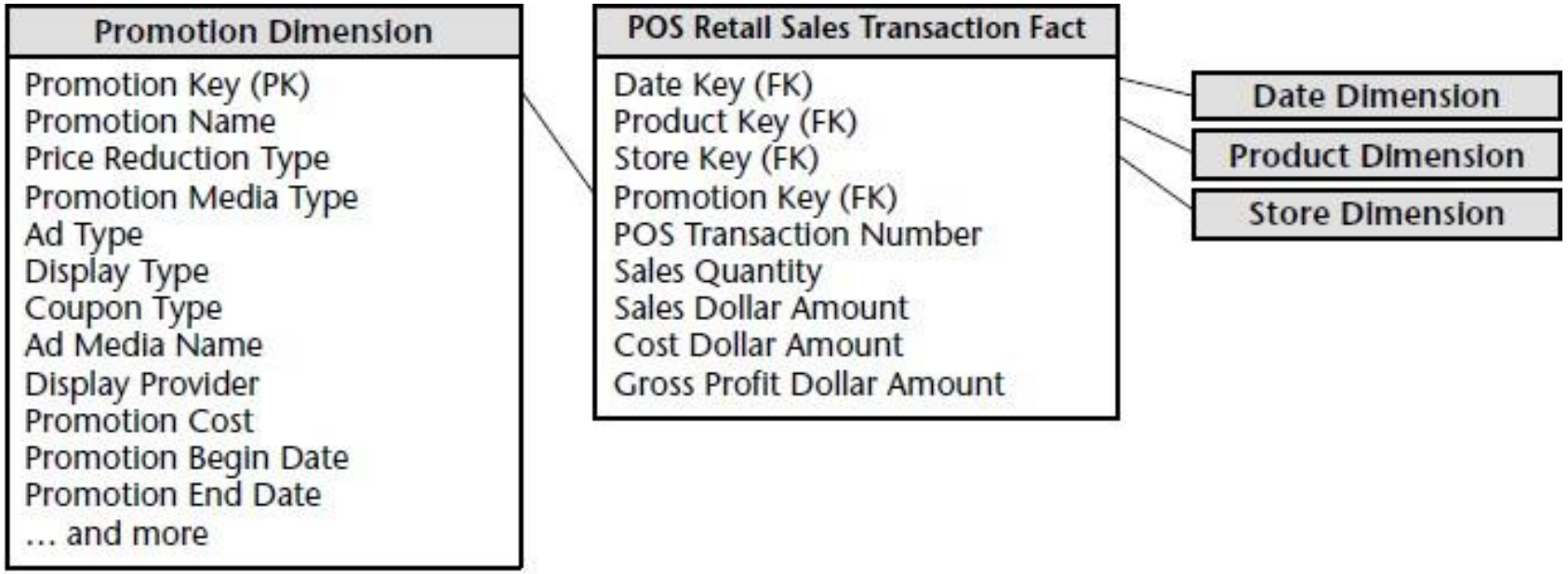


Figure 2.9 Promotion dimension in the retail sales schema.

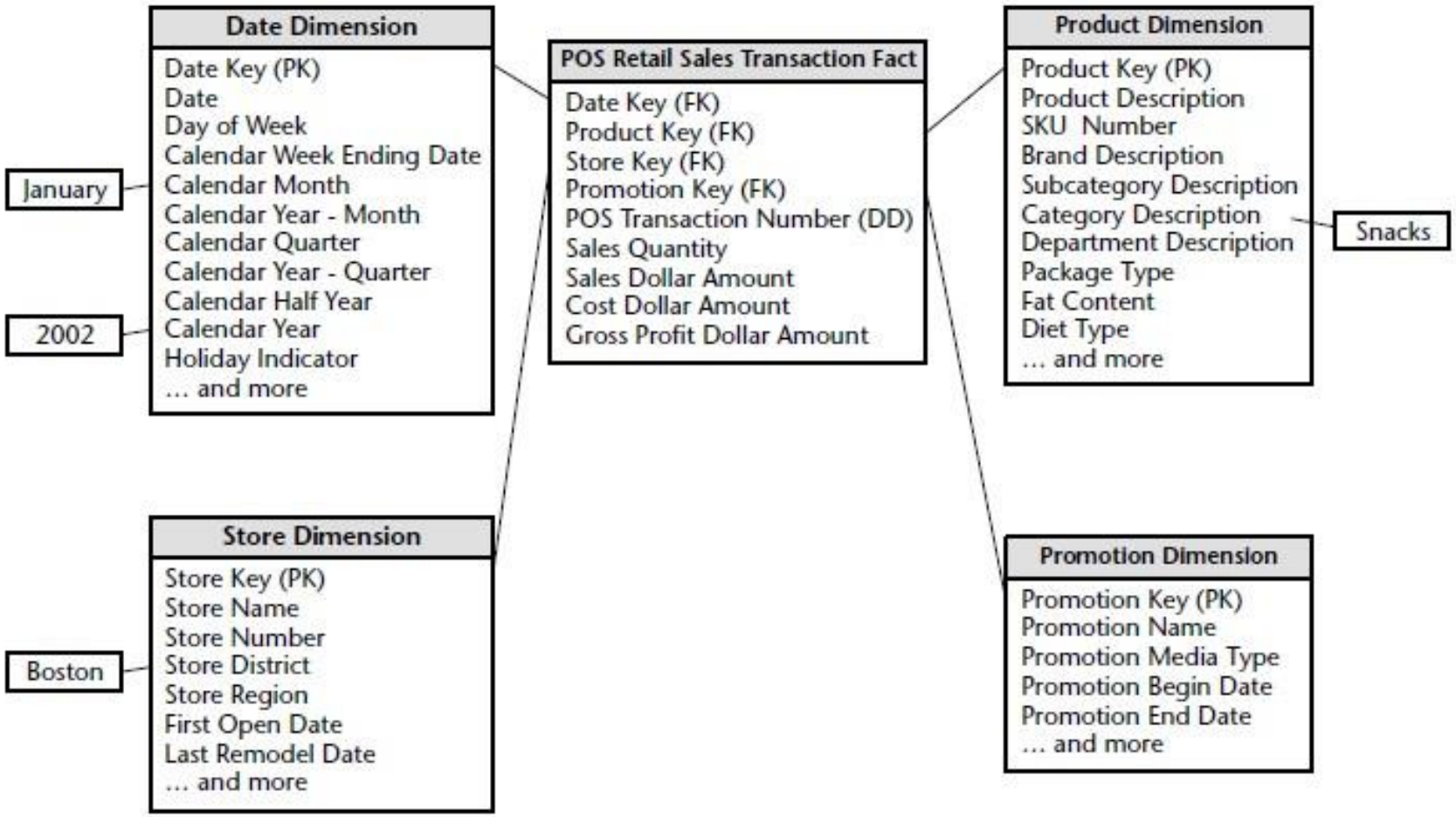


Figure 2.10 Querying the retail sales schema.

Retail Schema in Action

Calendar Week Ending Date	Promotion Name	Sales Dollar Amount
January 6, 2002	No Promotion	22,647
January 13, 2002	No Promotion	4,851
January 20, 2002	Super Bowl Promotion	7,248
January 27, 2002	Super Bowl Promotion	13,798

Calendar Week Ending Date	Super Bowl Promotion Sales Dollar Amount	No Promotion Sales Dollar Amount
January 6, 2002	0	22,647
January 13, 2002	0	4,851
January 20, 2002	7,248	0
January 27, 2002	13,793	0

Dimension Normalization - **Snowflaking**

SNOWFLAKE SCHEMA

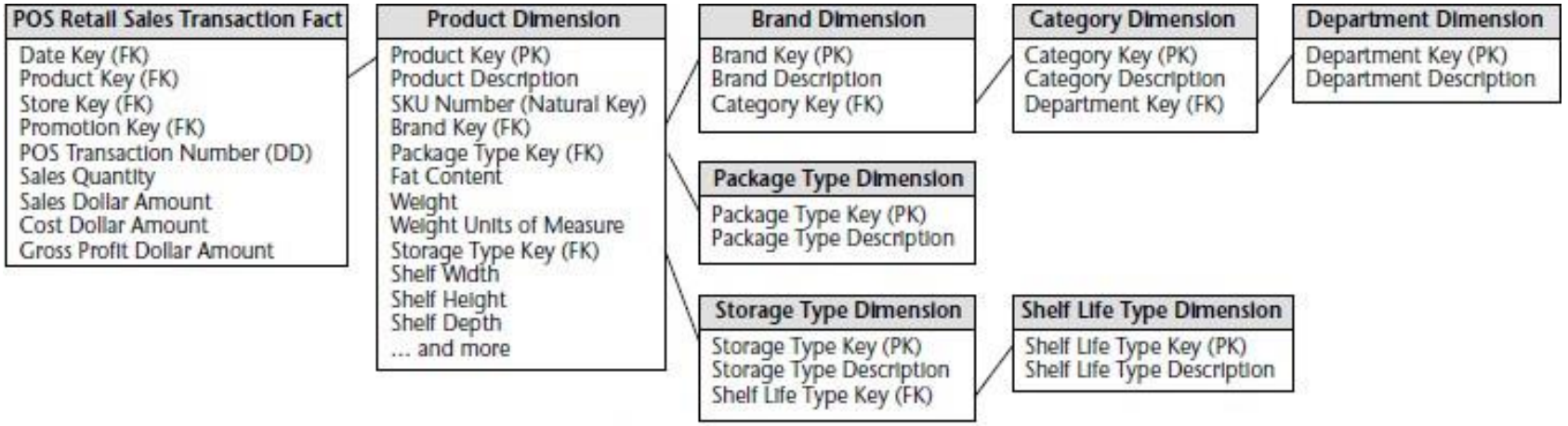


Figure 2.12 Partially snowflaked product dimension.