SS G515 - Data Warehousing:
Dimensional Modeling

Dr. Yashvardhan Sharma
Assistant Professor, CS & IS Dept.
BITS-Pilani

Four -Step Dimensional Design Process

Initial definitions

1. Select the business process to model

   Example of business process

   :raw materials purchasing, orders, shipments, invoicing, inventory

2. Declare the grain of the business process

   Example of grain declarations

   :an line item on a bill received from a doctor, an individual boarding pass to get on a flight

3. Choose the dimensions that apply to each fact table row

   Example of dimensions

   :date, product, customer, transaction type, status

4. Identify the numeric facts that will populate each fact table row

   "What are we measuring?" the answer is used to determine the facts

Retail Case Study

Brief description of the retail business (1/2)

- We work in the headquarters of a large grocery chain
- Our business has 100 grocery stores spread over a five-state area
- Each of the stores has a full complement of departments, including grocery, frozen foods, dairy, meat, produce, bakery, floral, and health/beauty aids
- Each store has roughly 60,000 individual products
- The individual products are called stock keeping units (SKUs)
- About 55,000 of the SKUs come from outside manufacturers and have bar codes imprinted on the product package
- These bar codes are called universal product codes (UPCs)
- UPCs are at the same grain as individual SKUs
- Each different package variation of a product has a separate UPC and hence is a separate SKU

Retail Case Study

Brief description of the retail business (2/2)

- The remaining 5,000 SKUs come from departments such as meat, produce, bakery, or floral
- While these products don't have nationally recognized UPCs, the grocery chain assigns SKU numbers to them
- The bar codes are not UPCs, they are certainly SKU numbers
- Our modern grocery store scans the bar codes directly into the point-of-sale (POS) system
- At the grocery store, management is concerned with the logistics of ordering, stocking, and selling products while maximizing profit
- Some of the most significant management decisions have to do with pricing and promotions

Retail Case Study

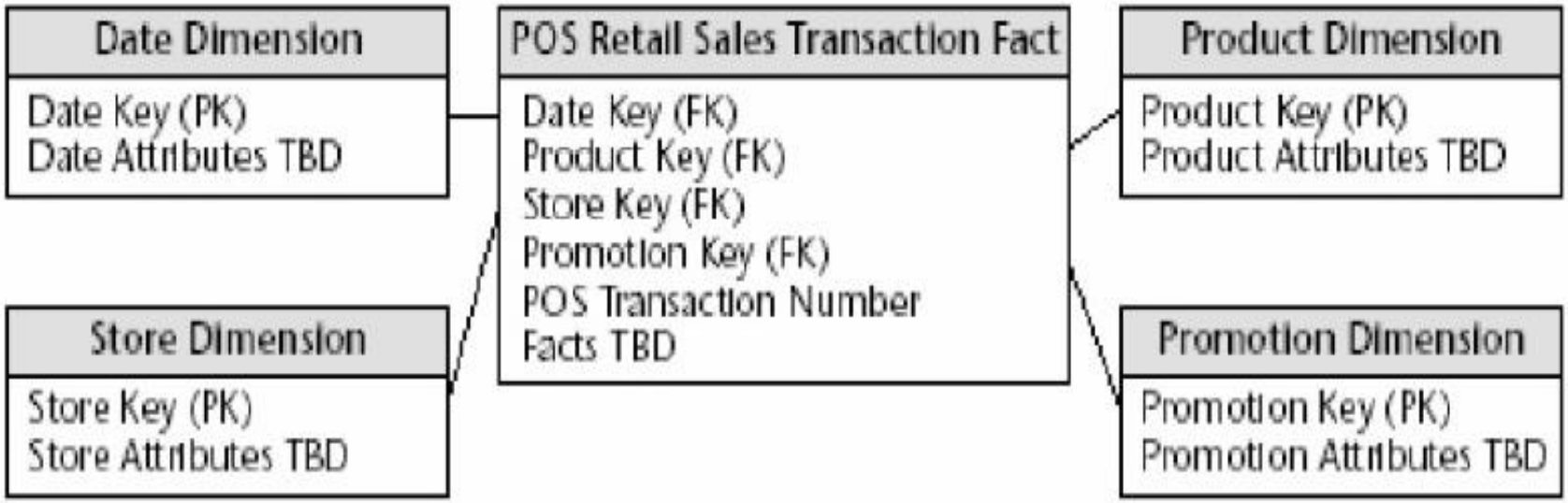
Dimensional Design Process

- Step1. Select the business process
  - "POS retail sales" business process to analyze
  - what products are selling in which stores
  - on what days
  - under what promotional conditions
- Step2. Declare the grain
  - The most granular data is an individual line item on a POS transaction
- Step3. Choose the dimensions
  - Once the grain the fact table has been chosen,
    - The date, product, and store primary dimensions fall out immediately
  - It is possible to add more dimensions to the basic grain of the fact table
    - We can ask whether other dimensions can be attributed to the data, such as the promotion under which the product is sold

Retail Case Study

Dimensional Design Process

- Step3. Choose the dimensions
  - Fig 2.2 Preliminary retail sales schema

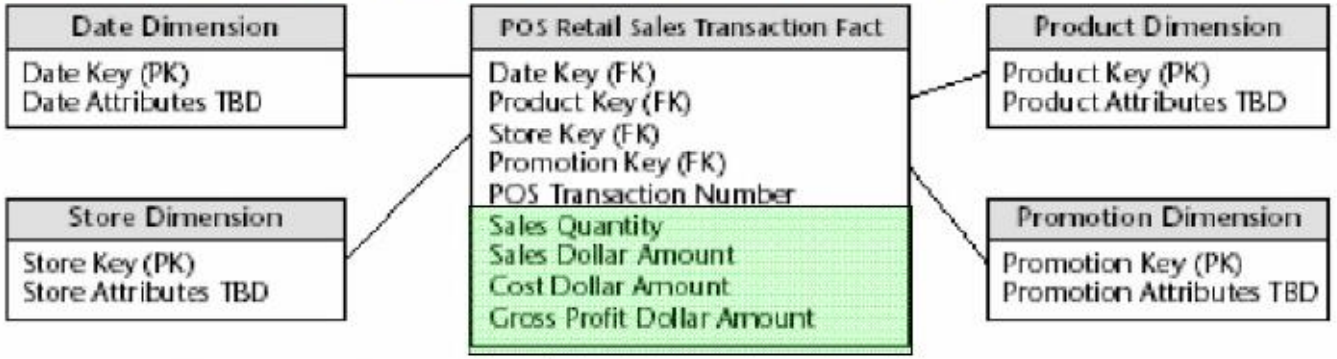


(TBD means to be determined)

Retail Case Study

Dimensional Design Process

- Step4. Identify the facts
  - Fig 2.3 Measured facts in the retail sales schema



(TBD means to be determined)

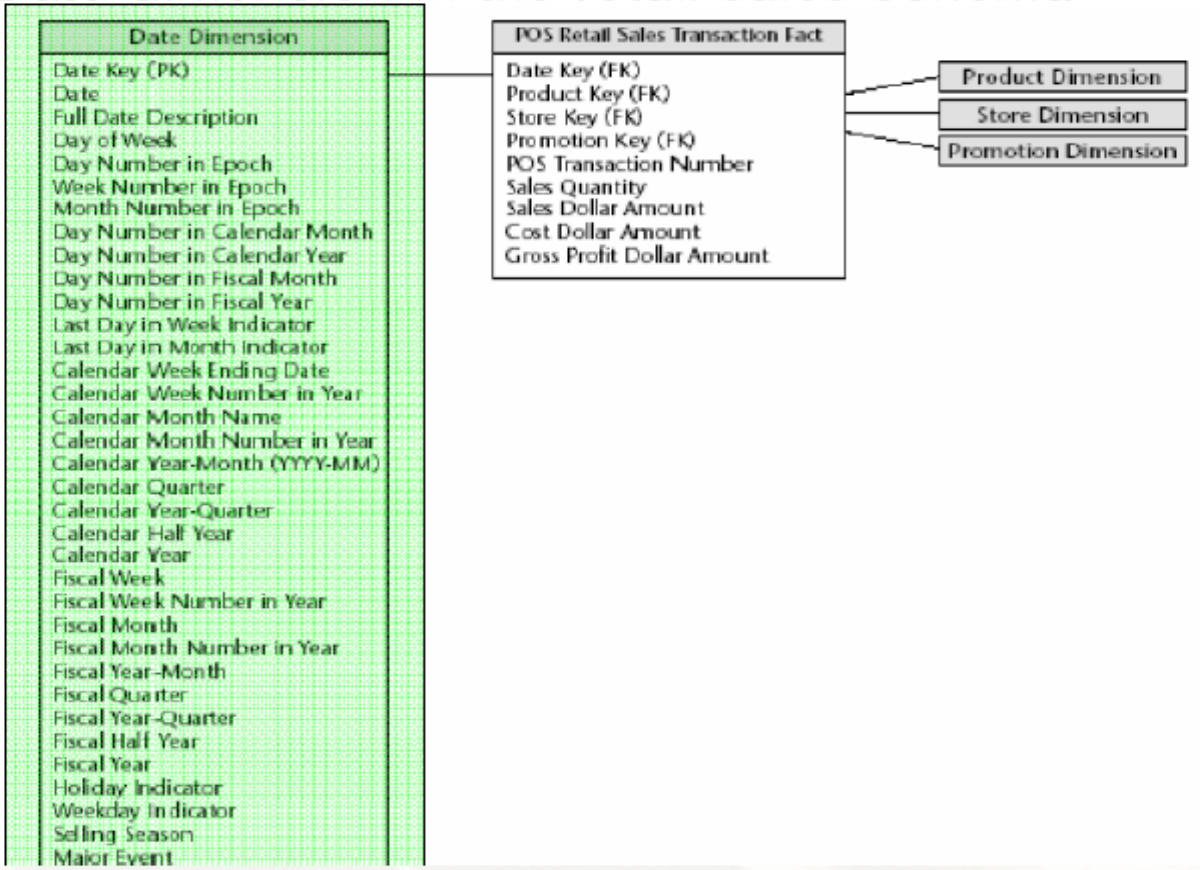
- Sales quantity, sales dollar amount, and cost dollar amount are additive across all the dimensions
- Gross profit is additive across all the dimensions
  - Storing it eliminates the possibility of user error
- Percentages and ratios, such as gross margin, are nonadditive
  - The numerator and denominator should be stored in the fact table
- Unit price is also a nonadditive fact
  - Summing up unit price across any of the dimensions results in a meaningless number

Dimensional Table Attributes

Focus on filling the dimension tables with robust attributes

Date Dimension

- is the one dimension nearly guaranteed to be in every data mart
  - because virtually every data mart is a time series
- Fig 2.4 Date dimension in the retail sales schema



Dimensional Table Attributes

Date Dimension

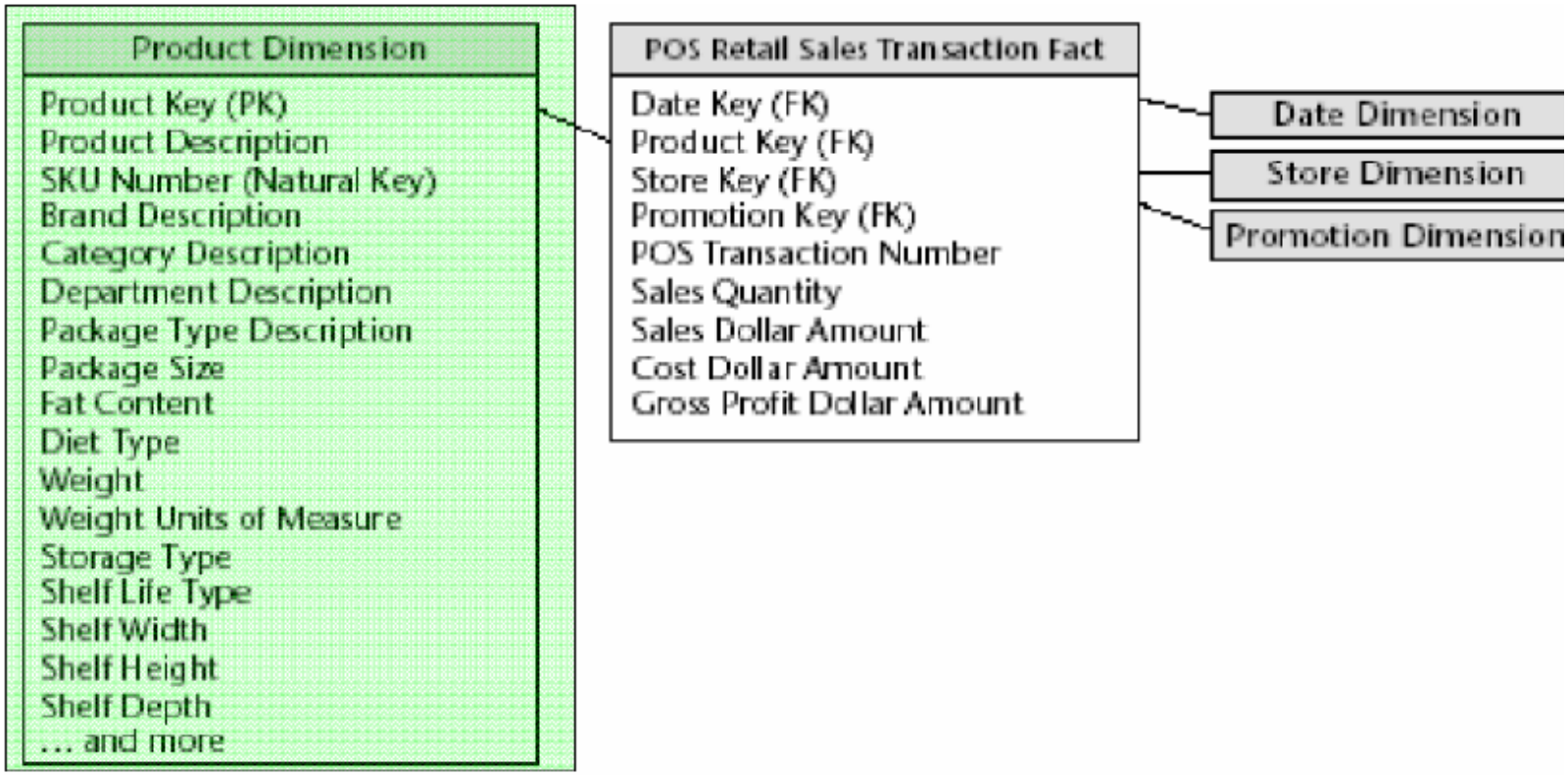
- Fig 2.5 Date dimension table detail

Date Key	Date	Full Date Description	Day of Week	Calendar Month	Calendar Year	Fiscal Year-Month	Holiday Indicator	Weekday Indicator
1	01/01/2002	January 1, 2002	Tuesday	January	2002	F2002-01	Holiday	Weekday
2	01/02/2002	January 2, 2002	Wednesday	January	2002	F2002-01	Non-Holiday	Weekday
3	01/03/2002	January 3, 2002	Thursday	January	2002	F2002-01	Non-Holiday	Weekday
4	01/04/2002	January 4, 2002	Friday	January	2002	F2002-01	Non-Holiday	Weekday
5	01/05/2002	January 5, 2002	Saturday	January	2002	F2002-01	Non-Holiday	Weekend
6	01/06/2002	January 6, 2002	Sunday	January	2002	F2002-01	Non-Holiday	Weekend
7	01/07/2002	January 7, 2002	Monday	January	2002	F2002-01	Non-Holiday	Weekday
8	01/08/2002	January 8, 2002	Tuesday	January	2002	F2002-01	Non-Holiday	Weekday

Dimensional Table Attributes

Product Dimension

- describes every stock keeping unit (SKU) in the grocery store
- Fig 2.7 Product dimension in the retail sales schema



Dimensional Table Attributes

Product Dimension

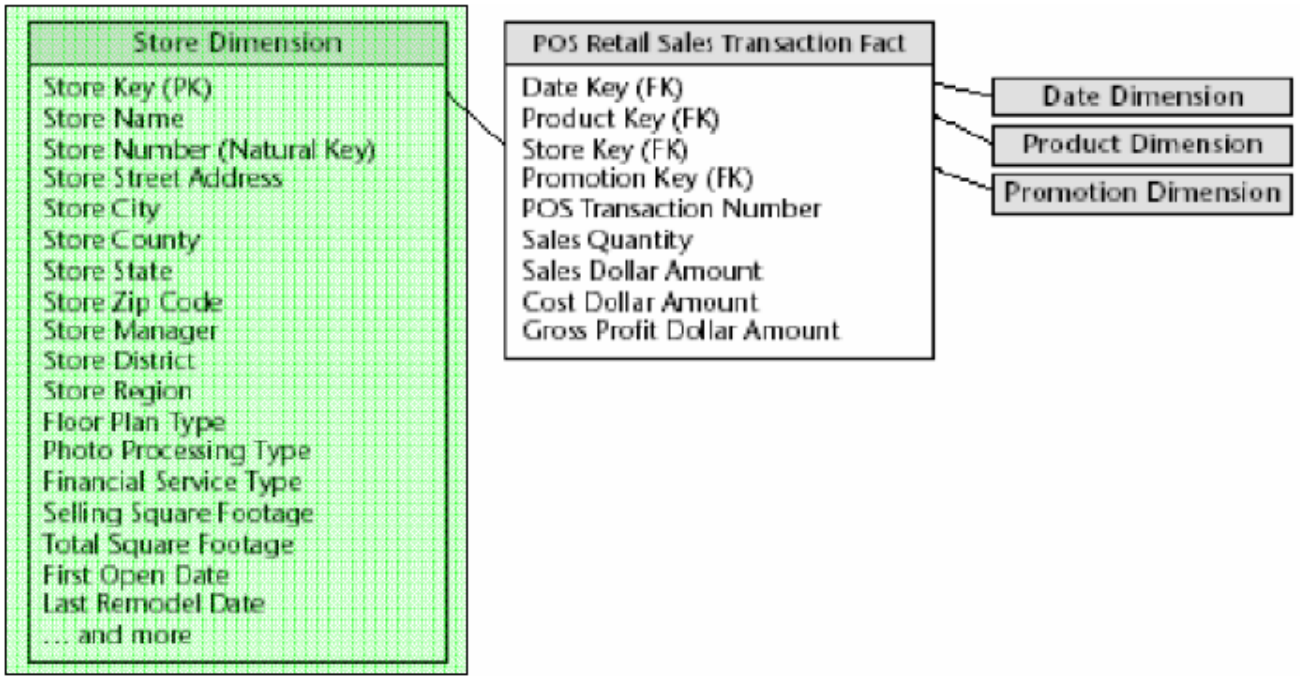
- Fig 2.6 Product dimension table detail

Product Key	Product Description	Brand Description	Category Description	Department Description	Fat Content
1	Baked Well Light Sourdough Fresh Bread	Baked Well	Bread	Bakery	Reduced Fat
2	Fluffy Sliced Whole Wheat	Fluffy	Bread	Bakery	Regular Fat
3	Fluffy Light Sliced Whole Wheat	Fluffy	Bread	Bakery	Reduced Fat
4	Fat Free Mini Cinnamon Rolls	Light	Sweeten Bread	Bakery	Non-Fat
5	Diet Lovers Vanilla 2 Gallon	Coldpack	Frozen Desserts	Frozen Foods	Non-Fat
6	Light and Creamy Butter Pecan 1 Pint	Freshlike	Frozen Desserts	Frozen Foods	Reduced Fat
7	Chocolate Lovers 1/2 Gallon	Frigid	Frozen Desserts	Frozen Foods	Regular Fat
8	Strawberry Ice Creamy 1 Pint	Icy	Frozen Desserts	Frozen Foods	Regular Fat
9	Icy Ice Cream Sandwiches	Icy	Frozen Desserts	Frozen Foods	Regular Fat

Dimensional Table Attributes

Store Dimension

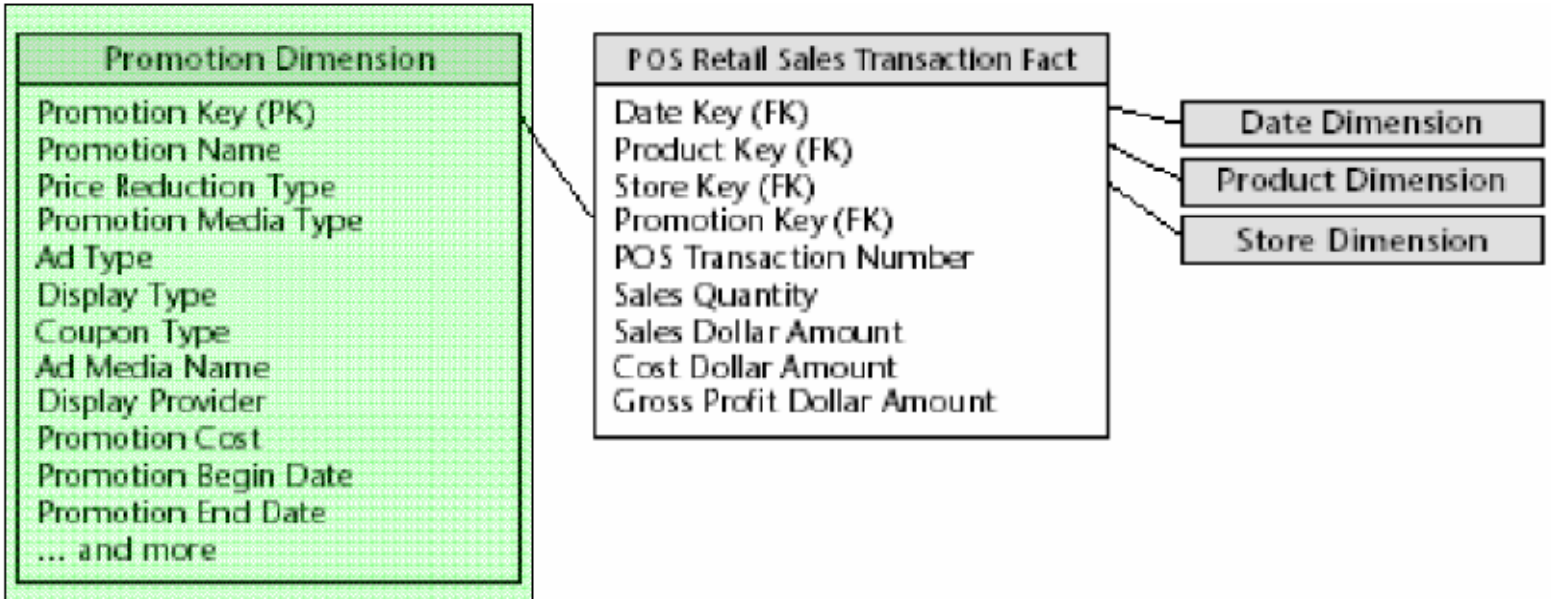
- describes every store in our grocery chain
- is the primary geographic dimension in our case study
  - Each store can be thought of as a location
  - We can roll stores up to any geographic attribute, such as ZIP code county, and state in the United States
- Fig 2.8 Store dimension in the retail sales schema



Dimensional Table Attributes

Promotion Dimension

- is potentially the most interesting dimension in our schema
- describes the promotion conditions under which a product was sold
  - Temporary price reductions, end-aisle displays, newspaper ads, and coupons
- is often called a causal dimension (as opposed to a casual dimension)
  - It describes factors thought to cause a change in product sales
- Fig 2.9 Promotion dimension in the retail sales schema



Dimensional Table Attributes

Promotion Dimension

- The various possible causal conditions are highly correlated
  - A temporary price reduction is associated with an ad and an end-aisle display
  - Coupons often are associated with ads

For four major causal mechanisms (price reductions, ads, displays, and coupons)

- The tradeoffs in favor of keeping the four dimensions together
  - The combined single dimension can be browsed efficiently to see how the various causal mechanisms are used together
- The tradeoffs in favor of separating the four causal mechanisms into distinct dimension tables
  - The separated dimensions may be more understandable to the business community

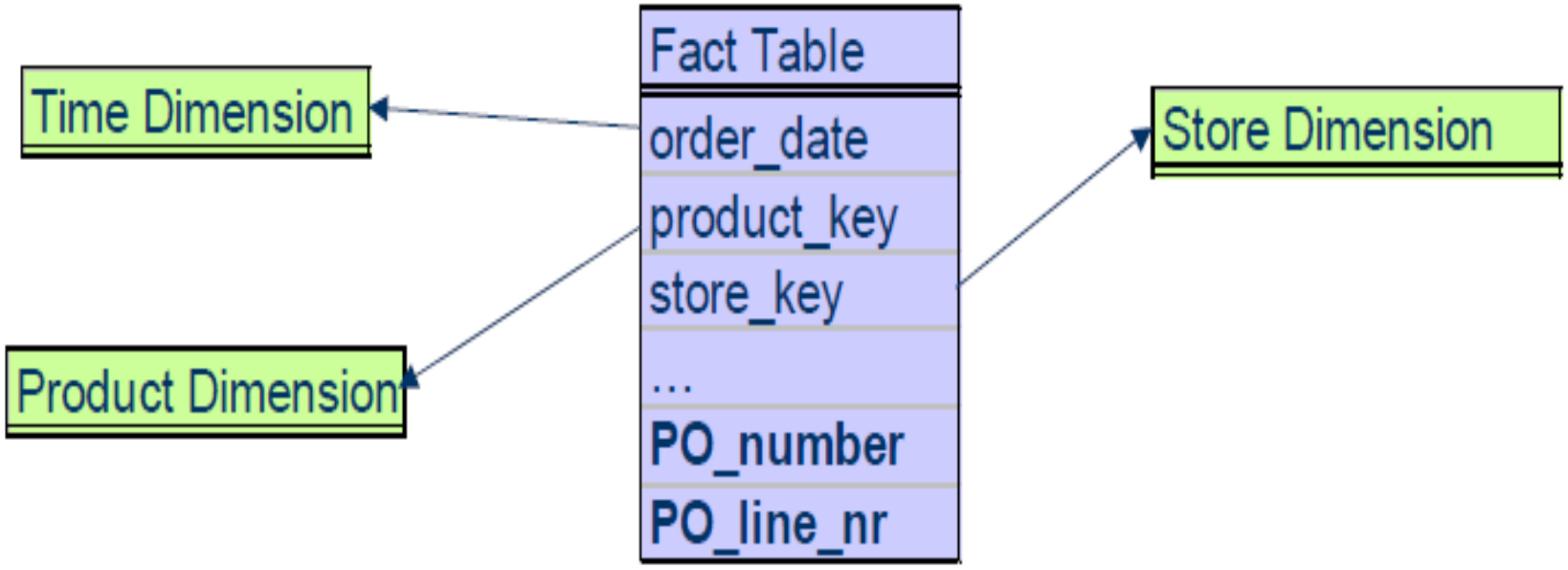
Dimensional Table Attributes

Degenerate Transaction Number Dimension

- Degenerate Dimension (DD)
  - The resulting dimension is empty
- POS transaction number
  - The natural operational ticket number, such as the POS transaction number, sits by itself in the fact table without joining to a dimension table
- Degenerate Dimensions are very common
  - When the grain of a fact table represents a single transaction or transaction line item
- Degenerate Dimensions often play an integral role in the fact table's primary key
  - In this case study, the primary key of the retail sales fact table consists of the degenerate POS transaction number and product key

Degenerate Dimension

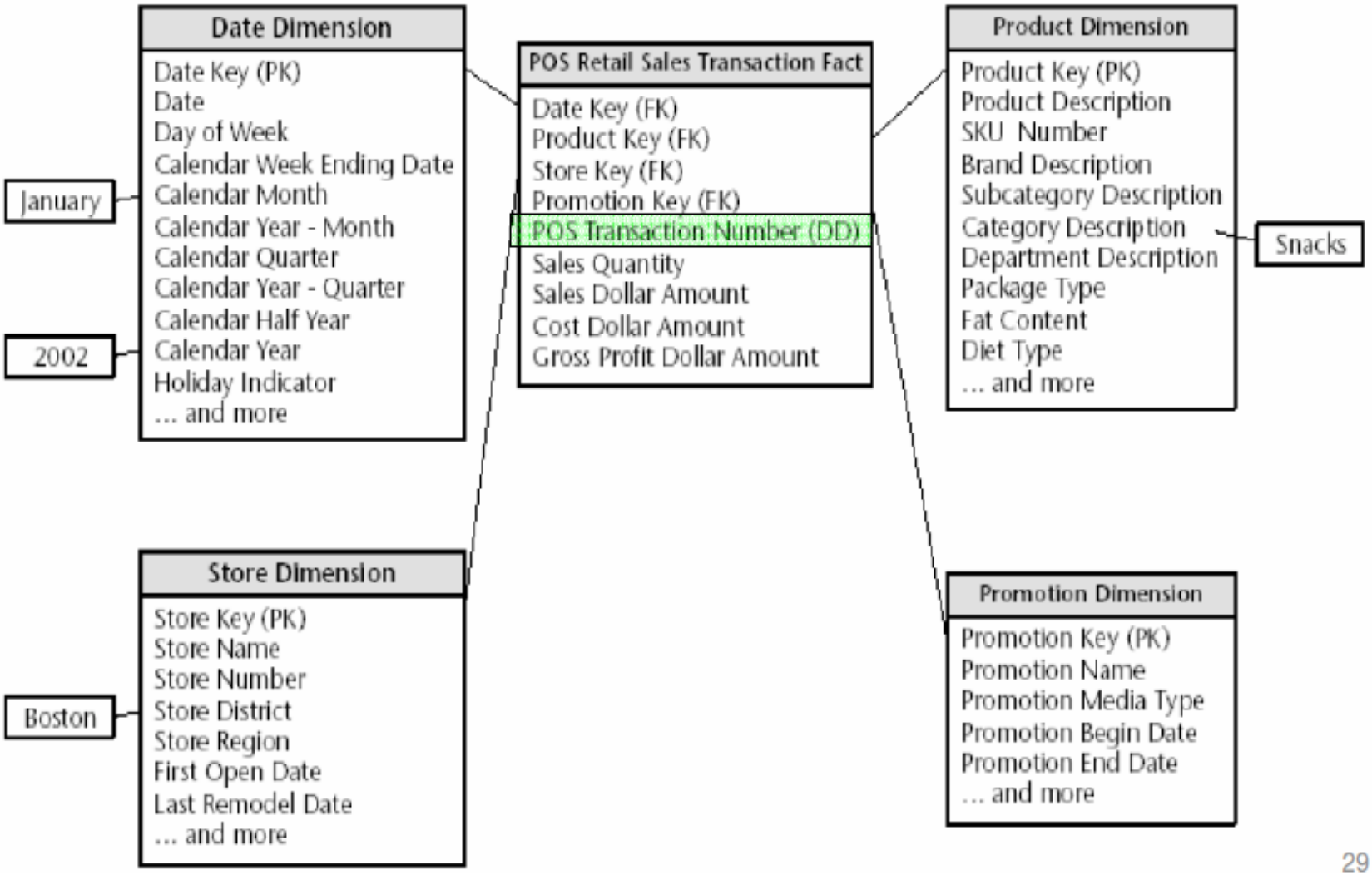
- A degenerate dimension is represented by a dimension key attribute(s) with no corresponding dimension table
- Occurs usually in line-item oriented fact table design



Dimensional Table Attributes

Degenerate Transaction Number Dimension

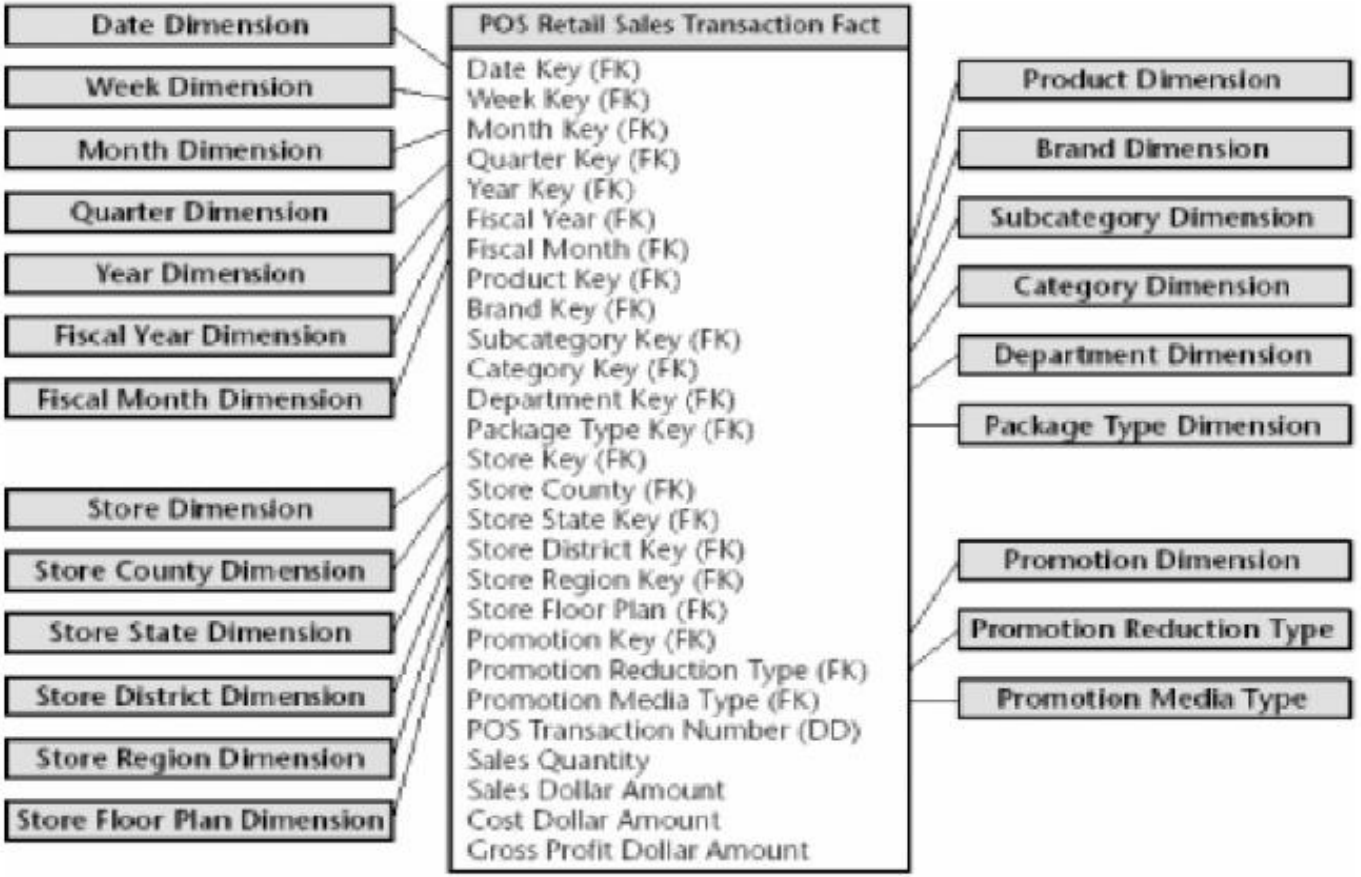
- Fig 2.10 Querying the retail sales schema



Denormalized Fact Tables

Too Many Dimensions

- Fig 2.13 Centipede fact table with too many dimensions



- Centipedes fact tables appear to have nearly 100 legs
- The compact fact table has turned into an unruly monster that joins to literally dozens of dimension tables

Denormalized Fact Tables

Too Many Dimensions

- Designing a fact table with too many dimensions leads to significantly increased fact table disk space requirements
- The numerous joins are an issue for both usability and query performance
- Most business processes can be represented with less than 15 dimensions in the fact table
- If our design has 25 or more dimensions, we should look for ways to combine correlated dimensions into a single dimension
  - Perfectly correlated attributes, such as the levels of a hierarchy, as well as attributes with a reasonable statistical correlation, should be part of the same dimension

Dimensional Modeling Myths

1. Dimensional models and data marts are for summary data only
2. Dimensional models and data marts are departmental, not enterprise, solutions
   - Data marts are process-centric, not department-centric
3. Dimensional models and data marts are not scalable
4. Dimensional models and data marts are only appropriate when there is a predictable usage pattern
5. Dimensional models and data marts can't be integrated and therefore lead to stovepipe solutions
   - Most certainly can be integrated if they conform to the DW bus architecture