A CASE STUDY: **RETAIL SALES**

ITM-761 Business Intelligence

ดร. สลิล บุญพราหมณ์

Imagine that 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 on its shelves. 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.

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. Since our grocery chain is highly automated, we stick scanner labels on many of the items in these other departments. Although the bar codes are not UPCs, they are certainly SKU numbers.

At the grocery store, management is concerned with the logistics of ordering, stocking, and selling products while maximizing profit. The profit ultimately comes from charging as much as possible for each product, lowering costs for product acquisition and overhead, and at the same time attracting as many customers as possible in a highly competitive pricing environment.

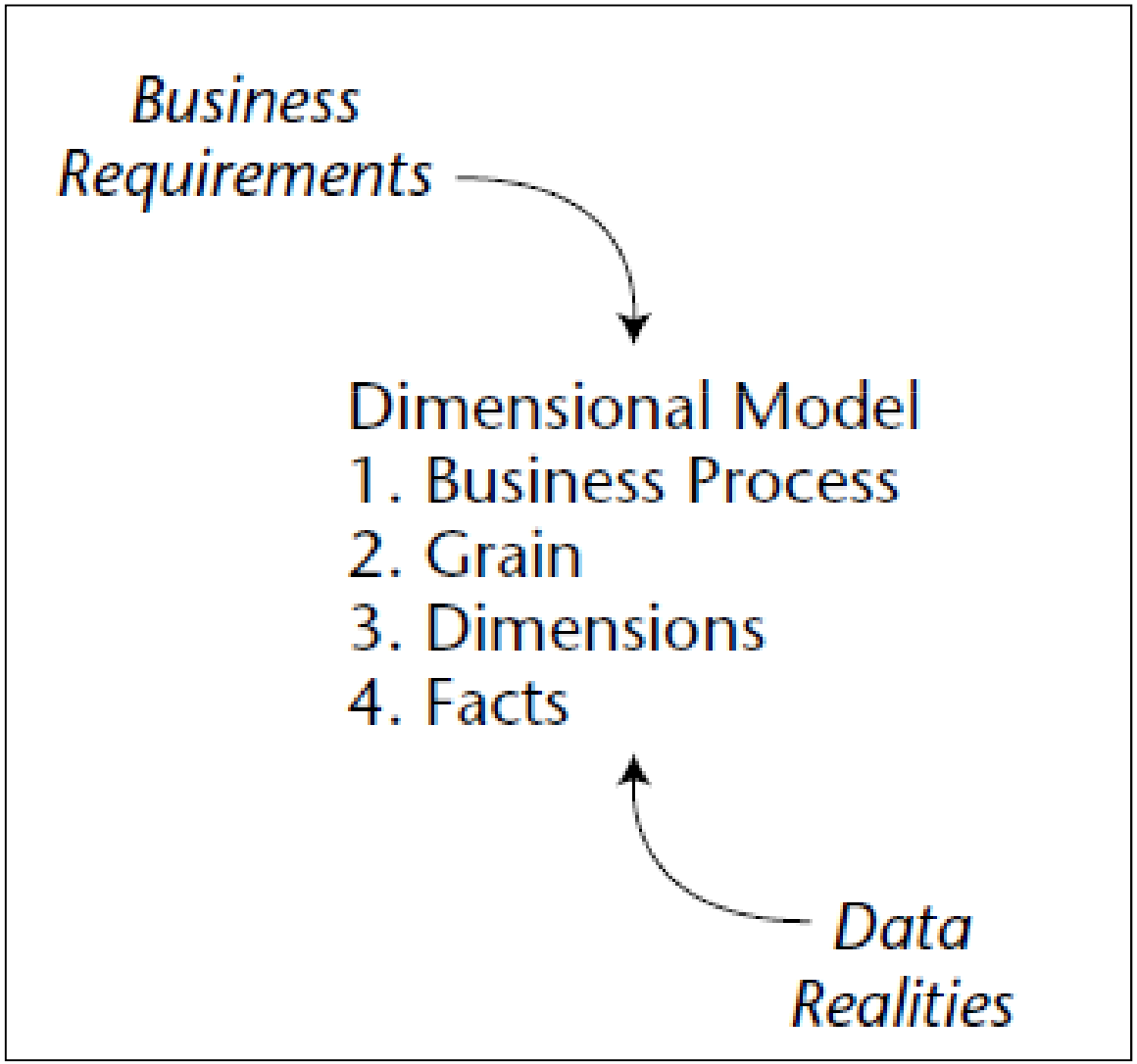
Some of the most significant management decisions have to do with pricing and promotions.

Both store management and headquarters marketing spend a great deal of time tinkering with pricing and promotions. Promotions in a grocery store include temporary price reductions, ads in newspapers and newspaper inserts, displays in the grocery store (including end-aisle displays), and coupons. The most direct and effective way to create a surge in the volume of product sold is to lower the price dramatically

Four-step process for designing dimensional models

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

- Key input to the four-step dimensional design process



Step 1. Select the Business Process

- decide what business process(es) to model by combining an understanding of the business requirements with an understanding of the available data
- management wants to better understand customer purchases as captured by the POS system
- the business process we're going to model is POS retail sales
- This data will allow us to analyze what products are selling in which stores on what days under what promotional conditions

Step 2. Declare the Grain

- Specifying *exactly what an individual fact table row represents.*
- The grain conveys the level of detail associated with the fact table measurements. It provides the answer to the question, “How do you describe a single row in the fact table?”

Example grain declarations include:

- An individual line item on a customer's retail sales ticket as measured by a scanner device
- A line item on a bill received from a doctor
- An individual boarding pass to get on a flight
- A daily snapshot of the inventory levels for each product in a warehouse
- A monthly snapshot for each bank account
- Tackling data at its lowest, most atomic grain makes sense on multiple fronts.
- Atomic data is highly dimensional. The more detailed and atomic the fact measurement, the more things we know for sure.

- the most granular data is an individual line item on a POS transaction
  - they may want to understand the difference in sales on Monday versus Sunday.
  - Or they may want to assess whether it's worthwhile to stock so many individual sizes of certain brands, such as cereal.
  - Or they may want to understand how many shoppers took advantage of the 50-cents-off promotion on shampoo.
  - Or they may want to determine the impact in terms of decreased sales when a competitive diet soda product was promoted heavily.

Step 3. Choose the Dimensions

- Once the grain of the fact table has been chosen, the date, product, and store dimensions fall out immediately

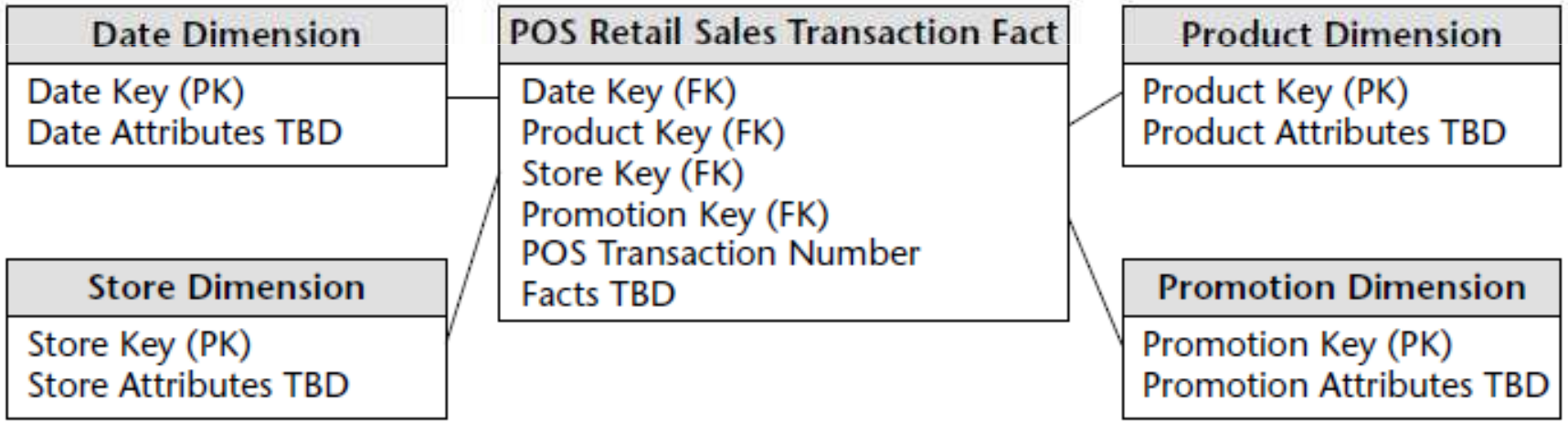


Figure 2.2 Preliminary retail sales schema.

"TBD" means "to be determined."

Step 4. Identify the Facts

- The facts collected by the POS system include the sales quantity (e.g., the number of cans of chicken noodle soup), per unit sales price, and the sales dollar amount. The sales dollar amount equals the sales quantity multiplied by the unit price.
- More sophisticated POS systems also provide a standard dollar cost for the product as delivered to the store by the vendor.

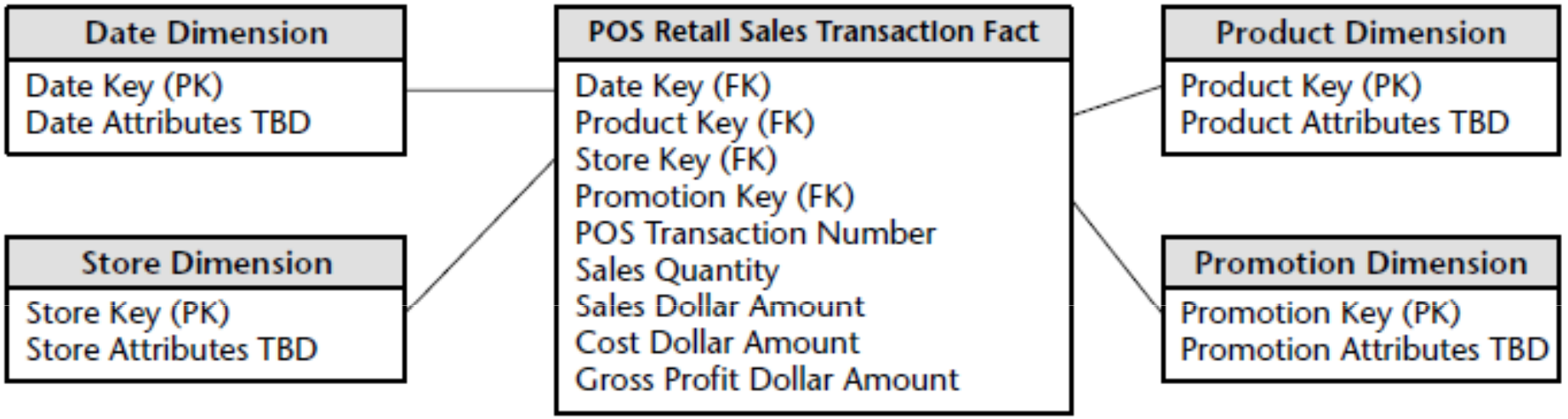
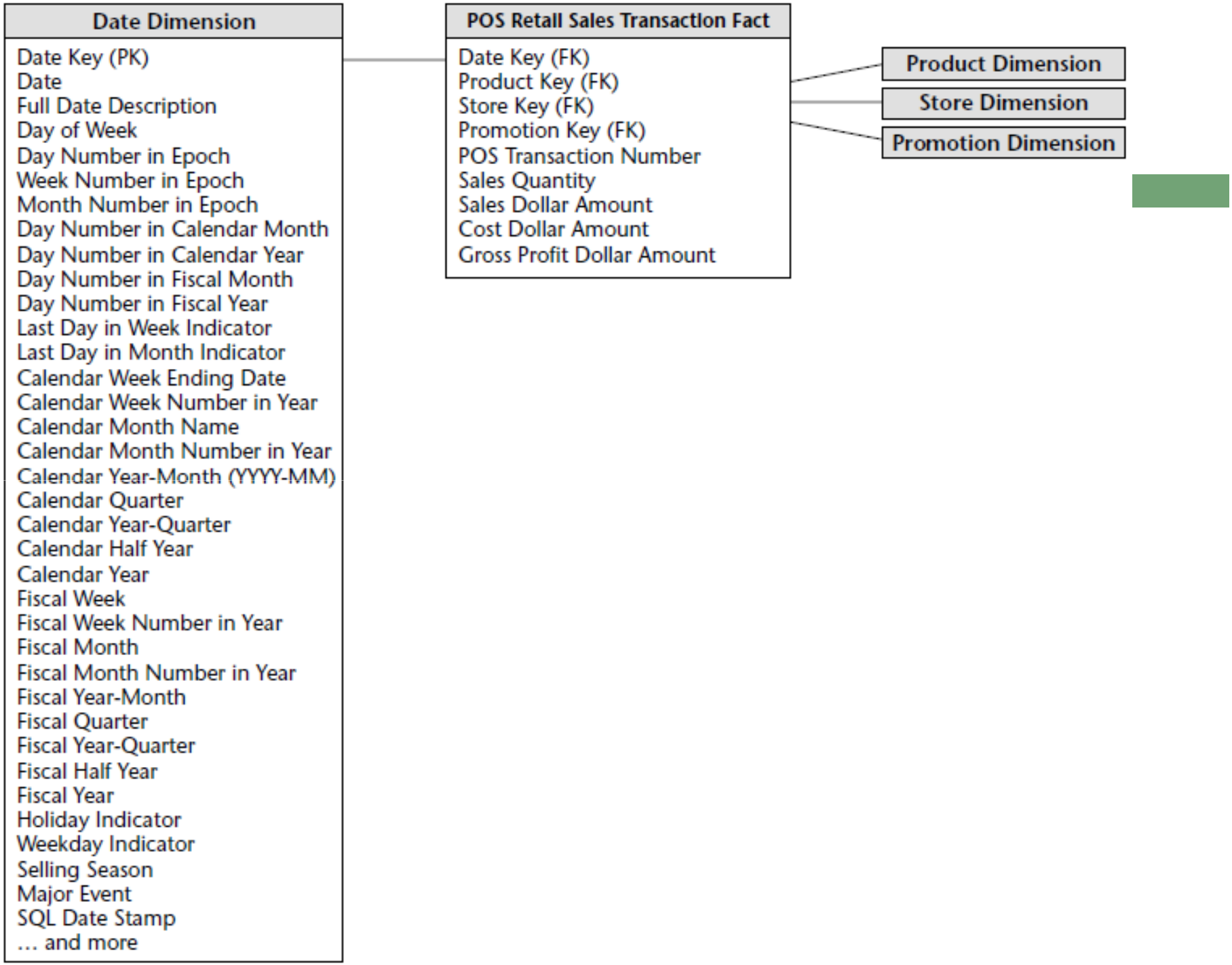


Figure 2.3 Measured facts in the retail sales schema.

Dimension Table Attributes

Date Dimension

- Unlike most of our other dimensions, we can build the date dimension table in advance. We may put 5 or 10 years of rows representing days in the table so that we can cover the history we have stored, as well as several years in the future. Even 10 years' worth of days is only about 3,650 rows, which is a relatively small dimension table.


Date Dimension
Date Key (PK)
Date
Full Date Description
Day of Week
Day Number in Epoch
Week Number in Epoch
Month Number in Epoch
Day Number in Calendar Month
Day Number in Calendar Year
Day Number in Fiscal Month
Day Number in Fiscal Year
Last Day in Week Indicator
Last Day in Month Indicator
Calendar Week Ending Date
Calendar Week Number in Year
Calendar Month Name
Calendar Month Number in Year
Calendar Year-Month (YYYY-MM)
Calendar Quarter
Calendar Year-Quarter
Calendar Half Year
Calendar Year
Fiscal Week
Fiscal Week Number in Year
Fiscal Month
Fiscal Month Number in Year
Fiscal Year-Month
Fiscal Quarter
Fiscal Year-Quarter
Fiscal Half Year
Fiscal Year
Holiday Indicator
Weekday Indicator
Selling Season
Major Event
SQL Date Stamp
... and more
POS Retail Sales Transaction Fact
Date Key (FK)
Product Key (FK)
Store Key (FK)
Promotion Key (FK)
POS Transaction Number
Sales Quantity
Sales Dollar Amount
Cost Dollar Amount
Gross Profit Dollar Amount
Product Dimension
Store Dimension
Promotion Dimension

- Each column in the date dimension table is defined by the particular day that the row represents.
- The day-of-week column contains the name of the day, such as Monday. This column would be used to create reports comparing the business on Mondays with Sunday business.
- The day number in calendar month column starts with 1 at the beginning of each month and runs to 28, 29, 30, or 31, depending on the month. This column is useful for comparing the same day each month. Similarly, we could have a month number in year (1, ... , 12).

- The day number in epoch is effectively a Julian day number (that is, a consecutive day number starting at the beginning of some epoch).
- For reporting, we would want a month name with values such as January. In addition, a yearmonth (YYYY-MM) column is useful as a report column header.

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

- We likely also will want a quarter number (Q1, ... , Q4), as well as a year quarter, such as 2001-Q4. We would have similar columns for the fiscal periods if they differ from calendar periods

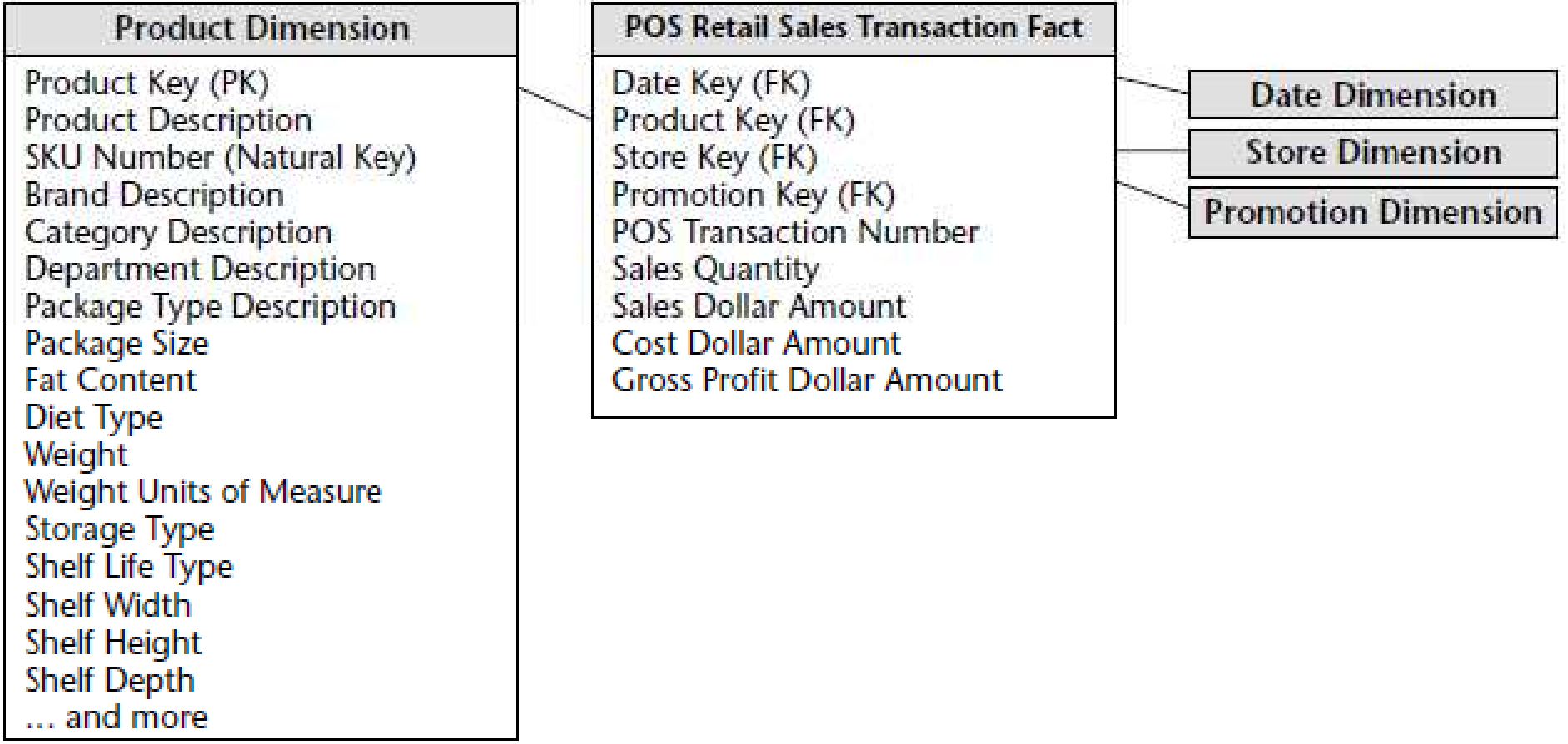
Product Dimension

- The product dimension describes every SKU in the grocery store. While a typical store in our chain may stock 60,000 SKUs,
- when we account for different merchandising schemes across the chain and historical products that are no longer available, our product dimension would have at least 150,000 rows and perhaps as many as a million rows.

- The product dimension is almost always sourced from the operational product master file.
- Most retailers administer their product master files at headquarters and download a subset of the file to each store's POS system at frequent intervals.
- It is headquarters' responsibility to define the appropriate product master record (and unique SKU number) for each new UPC created by packaged goods manufacturers.

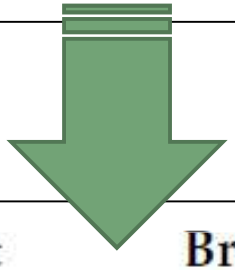
- An important function of the product master is to hold the many descriptive attributes of each SKU.
- The merchandise hierarchy is an important group of attributes. Typically, individual SKUs roll up to brands. Brands roll up to categories, and categories roll up to departments.

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


Product Dimension
Product Key (PK)
Product Description
SKU Number (Natural Key)
Brand Description
Category Description
Department Description
Package Type Description
Package Size
Fat Content
Diet Type
Weight
Weight Units of Measure
Storage Type
Shelf Life Type
Shelf Width
Shelf Height
Shelf Depth
... and more
POS Retail Sales Transaction Fact
Date Key (FK)
Product Key (FK)
Store Key (FK)
Promotion Key (FK)
POS Transaction Number
Sales Quantity
Sales Dollar Amount
Cost Dollar Amount
Gross Profit Dollar Amount
Date Dimension
Store Dimension
Promotion Dimension

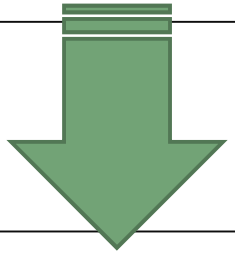
- Many of the attributes in the product dimension table are not part of the merchandise hierarchy. The package-type attribute, for example, might have values such as Bottle, Bag, Box, or Other. Any SKU in any department could have one of these values.
- For example, we could look at all the SKUs in the Cereal category packaged in Bags.

Department Description	Sales Dollar Amount	Sales Quantity
Bakery	$12,331	5,088
Frozen Foods	$31,776	15,565



Department Description	Brand Description	Sales Dollar Amount	Sales Quantity
Bakery	Baked Well	$3,009	1,138
Bakery	Fluffy	$3,024	1,476
Bakery	Light	$6,298	2,474
Frozen Foods	Coldpack	$5,321	2,640
Frozen Foods	Freshlike	$10,476	5,234
Frozen Foods	Frigid	$7,328	3,092
Frozen Foods	Icy	$2,184	1,437
Frozen Foods	QuickFreeze	$6,467	3,162

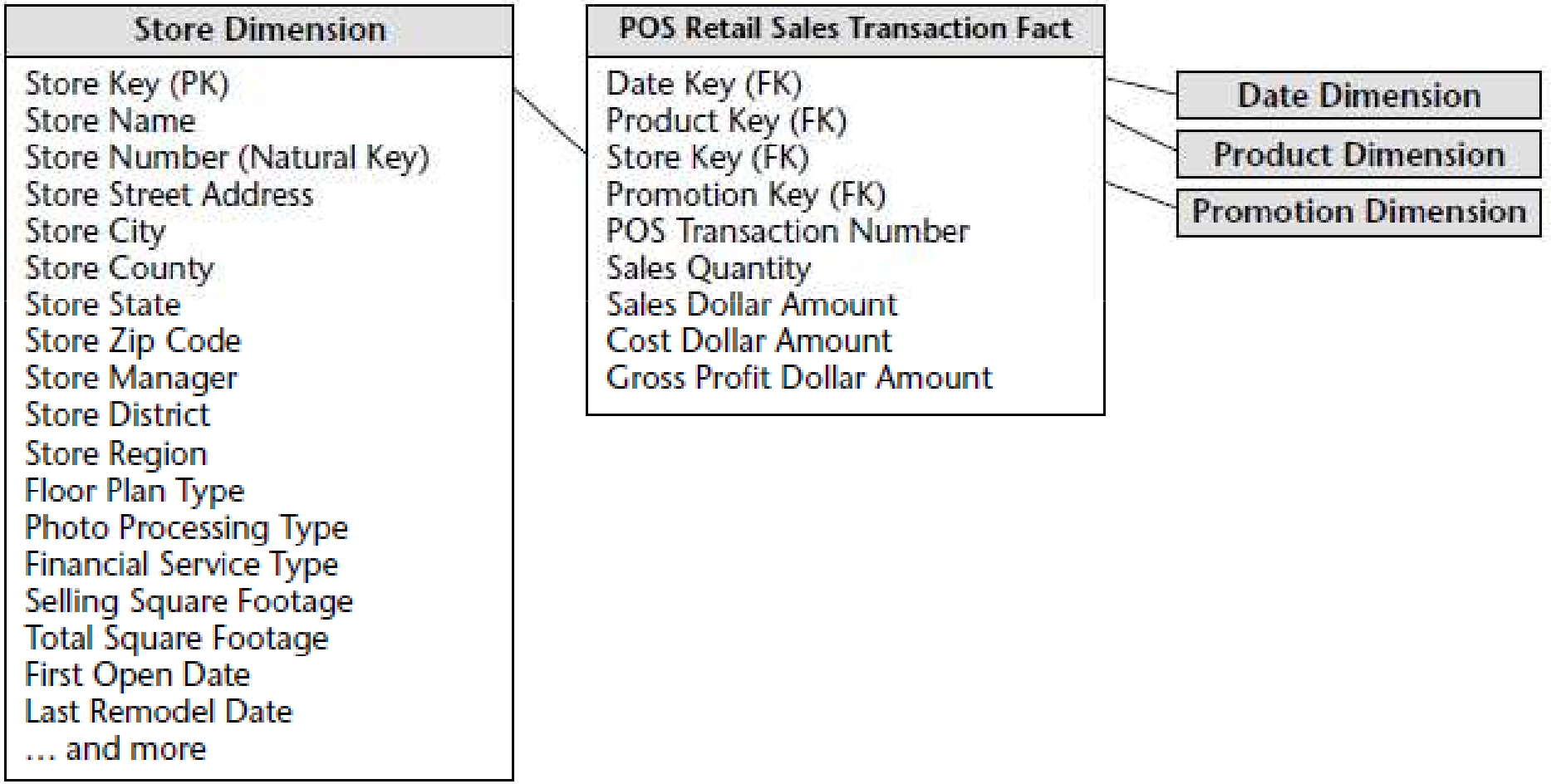
Department Description	Sales Dollar Amount	Sales Quantity
Bakery	$12,331	5,088
Frozen Foods	$31,776	15,565



Department Description	Fat Content	Sales Dollar Amount	Sales Quantity
Bakery	Non-Fat	$6,298	2,474
Bakery	Reduced Fat	$5,027	2,086
Bakery	Regular Fat	$1,006	528
Frozen Foods	Non-Fat	$5,321	2,640
Frozen Foods	Reduced Fat	$10,476	5,234
Frozen Foods	Regular Fat	$15,979	7,691

Store Dimension

- The store dimension is the primary geographic dimension in our case study
- Each store can be thought of as a location. Because of this, we can roll stores up to any geographic attribute, such as ZIP code, county, and state in the United States. Stores usually also roll up to store districts and regions


Store Dimension
Store Key (PK)
Store Name
Store Number (Natural Key)
Store Street Address
Store City
Store County
Store State
Store Zip Code
Store Manager
Store District
Store Region
Floor Plan Type
Photo Processing Type
Financial Service Type
Selling Square Footage
Total Square Footage
First Open Date
Last Remodel Date
… and more
POS Retail Sales Transaction Fact
Date Key (FK)
Product Key (FK)
Store Key (FK)
Promotion Key (FK)
POS Transaction Number
Sales Quantity
Sales Dollar Amount
Cost Dollar Amount
Gross Profit Dollar Amount
Date Dimension
Product Dimension
Promotion Dimension

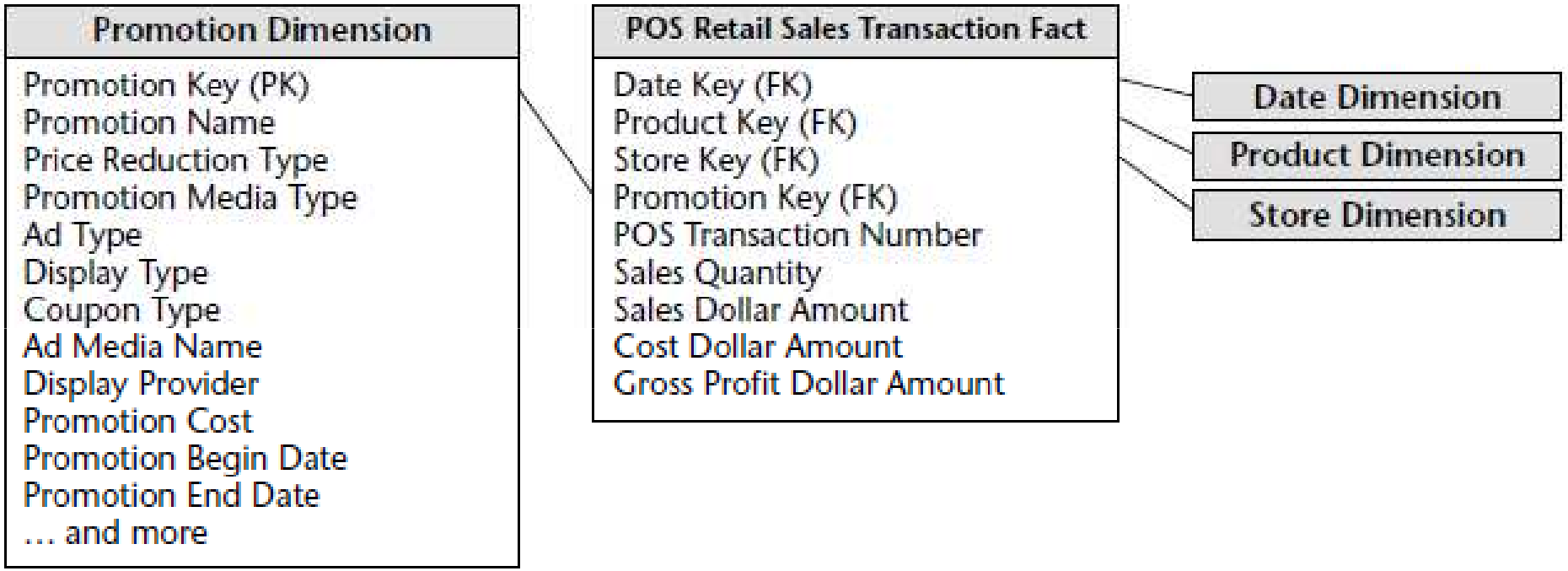
- The floor plan type, photo processing type, and finance services type are all short text descriptors that describe the particular store. These should not be one-character codes but rather should be 10- to 20-character standardized descriptors that make sense when viewed in a pull-down list or used as a report row header.

Promotion Dimension

- The promotion dimension is potentially the most interesting dimension in our schema.
- The promotion dimension describes the promotion conditions under which a product was sold.
- Promotion conditions include temporary price reductions, end-aisle displays, newspaper ads, and coupons.
- This dimension is often called a *causal dimension (as opposed to a casual dimension) because* it describes factors thought to cause a change in product sales

- Managers at both headquarters and the stores are interested in determining whether a promotion is effective or not
- Promotions are judged on one or more of the following factors:
  - Whether the products under promotion experienced a gain in sales during the promotional period. This is called the *lift*. The lift can only be measured if the store can agree on what the baseline sales of the promoted products would have been without the promotion. Baseline values can be estimated from prior sales history and, in some cases, with the help of sophisticated mathematical models.

- Whether the products under promotion showed a drop in sales just prior to or after the promotion, canceling the gain in sales during the promotion (time shifting). In other words, did we transfer sales from regularly priced products to temporarily reduced-priced products?
- Whether the products under promotion showed a gain in sales but other products nearby on the shelf showed a corresponding sales decrease
- Whether all the products in the promoted category of products experienced a net overall gain in sales taking into account the time periods before, during, and after the promotion (market growth)


Promotion Dimension
Promotion Key (PK)
Promotion Name
Price Reduction Type
Promotion Media Type
Ad Type
Display Type
Coupon Type
Ad Media Name
Display Provider
Promotion Cost
Promotion Begin Date
Promotion End Date
… and more
POS Retail Sales Transaction Fact
Date Key (FK)
Product Key (FK)
Store Key (FK)
Promotion Key (FK)
POS Transaction Number
Sales Quantity
Sales Dollar Amount
Cost Dollar Amount
Gross Profit Dollar Amount
Date Dimension
Product Dimension
Store Dimension

- From a purely logical point of view, we could record very similar information about the promotions by separating the four major causal mechanisms (price reductions, ads, displays, and coupons) into four separate dimensions rather than combining them into one dimension

Degenerate Transaction Number Dimension

- The retail sales fact table contains the POS transaction number on every line item row. In a traditional parent-child database, the POS transaction number would be the key to the transaction header record, containing all the information valid for the transaction as a whole, such as the transaction date and store identifier.
- However, in our dimensional model, we have already extracted this interesting header information into other dimensions. The POS transaction number is still useful because it serves as the grouping key for pulling together all the products purchased in a single transaction.

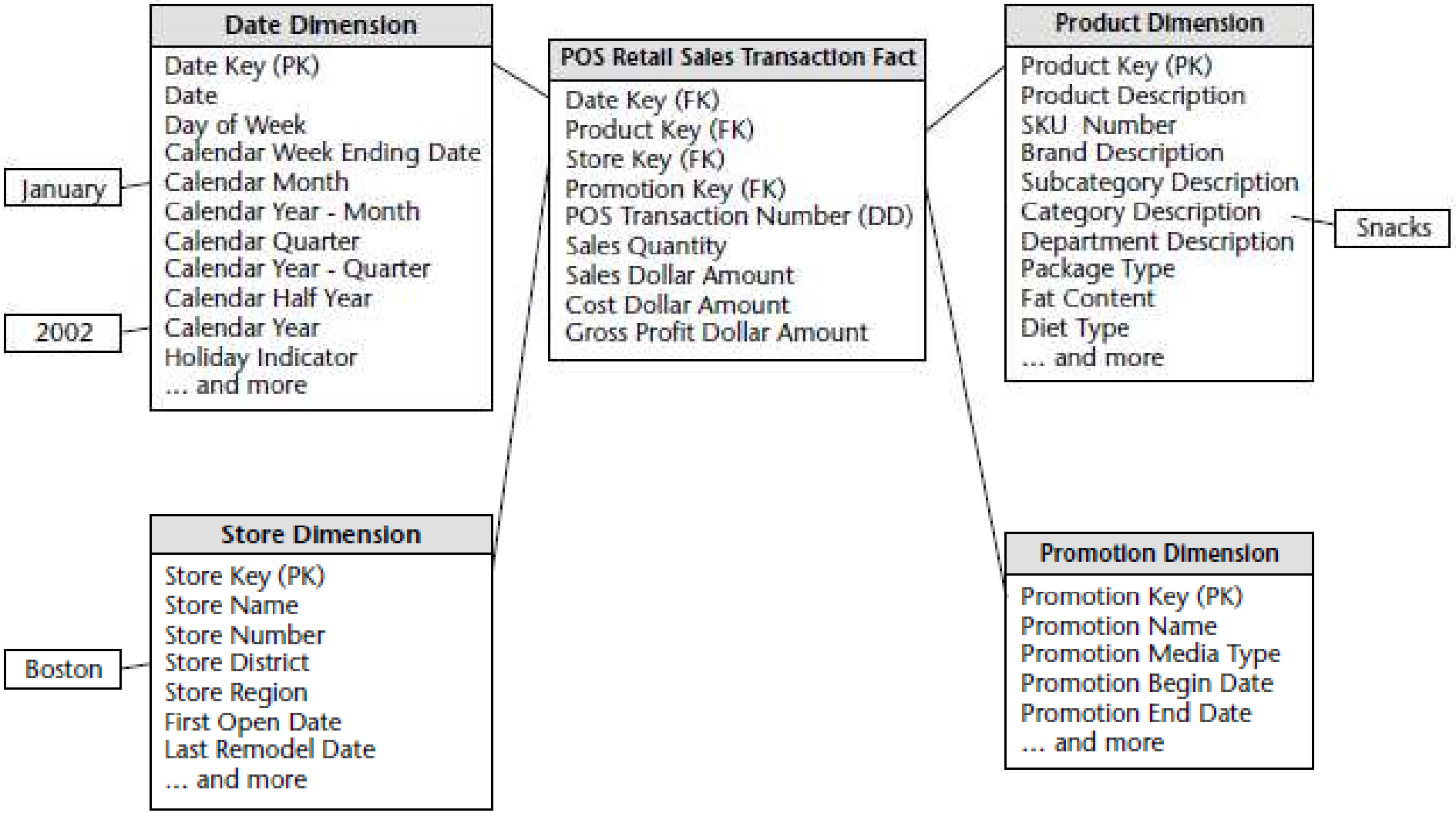
- we refer to the POS transaction number as a *degenerate dimension*
- Degenerate dimensions are very common when the grain of a fact table represents a single transaction or transaction line item because the degenerate dimension represents the unique identifier of the parent.
- Order numbers, invoice numbers, and bill-of-lading numbers almost always appear as degenerate dimensions in a dimensional model

- Degenerate dimensions often play an integral role in the fact table's primary key.
- In our case study, the primary key of the retail sales fact table consists of the degenerate POS transaction number and product key (assuming that the POS system rolls up all sales for a given product within a POS shopping cart into a single line item).
- Often, the primary key of a fact table is a subset of the table's foreign keys. We typically do not need every foreign key in the fact table to guarantee the uniqueness of a fact table row

- Operational control numbers such as order numbers, invoice numbers, and bill-of-loading numbers usually give rise to empty dimensions and are represented as degenerate dimensions (that is, dimension keys without corresponding dimension tables) in fact tables where the grain of the table is the document itself or a line item in the document.

Retail Schema in Action

- With our retail POS schema designed, let's illustrate how it would be put to use in a query environment.
- A business user might be interested in better understanding weekly sales dollar volume by promotion for the snacks category during January 2002 for stores in the Boston district.


Date Dimension
Date Key (PK)
Date
Day of Week
Calendar Week Ending Date
Calendar Month
Calendar Year - Month
Calendar Quarter
Calendar Year - Quarter
Calendar Half Year
Calendar Year
Holiday Indicator
... and more
January
2002
POS Retail Sales Transaction Fact
Date Key (FK)
Product Key (FK)
Store Key (FK)
Promotion Key (FK)
POS Transaction Number (DD)
Sales Quantity
Sales Dollar Amount
Cost Dollar Amount
Gross Profit Dollar Amount
Product Dimension
Product Key (PK)
Product Description
SKU Number
Brand Description
Subcategory Description
Category Description
Department Description
Package Type
Fat Content
Diet Type
... and more
Snacks
Store Dimension
Store Key (PK)
Store Name
Store Number
Store District
Store Region
First Open Date
Last Remodel Date
... and more
Boston
Promotion Dimension
Promotion Key (PK)
Promotion Name
Promotion Media Type
Promotion Begin Date
Promotion End Date
... and more

- If the query tool summed the sales dollar amount grouped by week-ending date and promotion, the query results would look similar to those below.

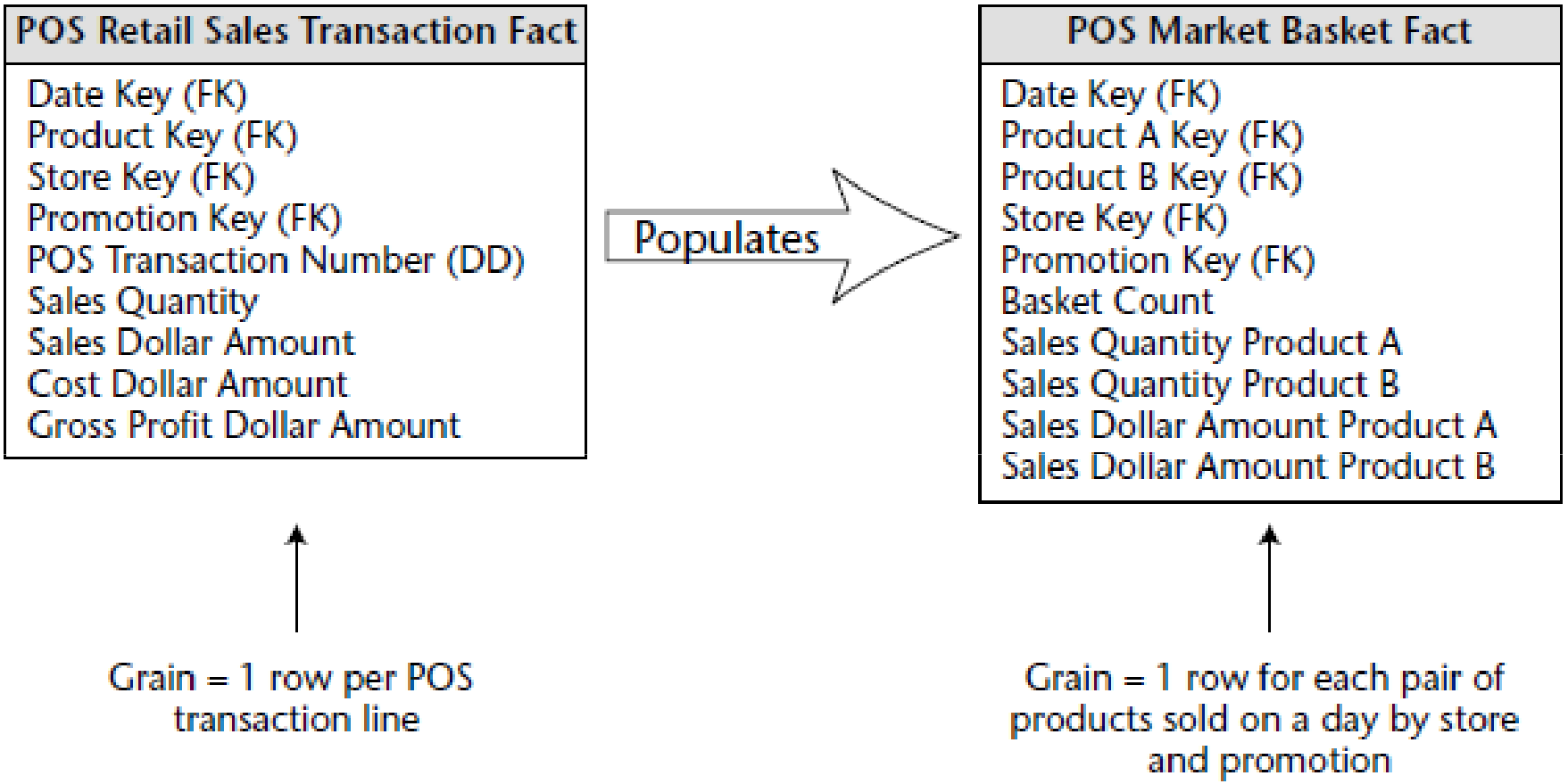
Calendar Week Ending Date	Promotion Name	Sales Dollar Amount
January 6, 2002	No Promotion	22,647
January 13, 2002	No Promotion	4,851
January 20, 2002	Super Bowl Promotion	7,248
January 27, 2002	Super Bowl Promotion	13,798

- If you are using a data access tool with more functionality, the results may appear as a cross-tabular report. Such reports are more appealing to business users than the columnar data resulting from an SQL statement

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

Market Basket Analysis

- Market basket analysis gives the retailer insights about how to merchandise various combinations of items. If frozen pasta dinners sell well with cola products, then these two products perhaps should be located near each other or marketed with complementary pricing.
- The retail sales fact table cannot be used easily to perform market basket analyses because SQL was never designed to constrain and group across line item fact rows.


POS Retail Sales Transaction Fact
Date Key (FK)
Product Key (FK)
Store Key (FK)
Promotion Key (FK)
POS Transaction Number (DD)
Sales Quantity
Sales Dollar Amount
Cost Dollar Amount
Gross Profit Dollar Amount
Populates
POS Market Basket Fact
Date Key (FK)
Product A Key (FK)
Product B Key (FK)
Store Key (FK)
Promotion Key (FK)
Basket Count
Sales Quantity Product A
Sales Quantity Product B
Sales Dollar Amount Product A
Sales Dollar Amount Product B
Grain = 1 row per POS transaction line
Grain = 1 row for each pair of products sold on a day by store and promotion

- The market basket fact table is a periodic snapshot representing the pairs of products purchased together during a specified time period.
- The facts include the total number of baskets (customer tickets) that included products A and B, the total number of product A dollars and units in this subset of purchases, and the total number of product B dollars and units purchased. The basket count is a semi-additive fact.

- Conceptually, the idea of recording market basket correlations is simple, but the sheer number of product combinations makes the analysis challenging.
- If we have N products in our product portfolio and we attempt to build a table with every possible pair of product keys encountered in product orders, we will approach N^2 or N x (N – 1) product combinations

- In other words, if we have 10,000 products in our portfolio, there would be nearly 100,000,000 pairwise combinations.
- The number of possible combinations quickly approaches absurdity when we're dealing with a large number of products. If a retail store sells 100,000 SKUs, there are 10 billion possible SKU combinations
- In order to avoid the combinatorial explosion of product pairs in the market basket fact table, we rely on a progressive pruning algorithm

- We begin at the top of the product hierarchy, which we'll assume is category.
- We first enumerate all the category-to-category market basket combinations. If there are 25 categories, this first step generates 625 market basket rows.
- We then prune this list for further analysis by selecting only the rows that have a reasonably high order count and where the dollars and units for products A and B (which are categories at this point) are reasonably balanced.
- Experimentation will tell you what the basket count threshold and balance range should be

- We then push down to the next level of detail, which we'll assume is brand.
- Starting with the pruned set of combinations from the last step, we drill down on product A by enumerating all combinations of brand (product A) by category (product B).
- Similarly, we drill down one level of the hierarchy for product B by looking at all combinations of brand (product A) by brand (product B).
- Again, we prune the lists to those with the highest basket count frequencies and dollar or unit balance and then drill down to the next level in the hierarchy

- As we descend the hierarchy, we produce rows with smaller and smaller basket counts.
- Eventually, we find no basket counts greater than the reasonable threshold for relevance. It is permissible to stop at any time once we've satisfied the analyst's curiosity

From The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. Ralph Kimball and Margy Ross.