Advanced Data Management Technologies

Unit 6 — Case Studies

J. Gamper

Free University of Bozen-Bolzano
Faculty of Computer Science
IDSE

Acknowledgements: I am indebted to Michael Böhlen and Stefano Rizzi for providing me their slides, upon which these lecture notes are based.

Outline

Outline

The Grocery Store Example/1

- A grocery chain with 500 stores spread over a five-state area.
  - Each of the stores is a typical modern supermarket with a full complement of departments including grocery, frozen foods, dairy, meat, bakery, hard goods, liquor, and drugs.
  - Each store has roughly 60.000 individual products on its shelves.
- The individual products are called stock keeping units (SKUs).
- About 40.000 of the SKU 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 20.000 SKUs come from departments like meat or bakery departments and do not have nationally recognized UPC codes.
  - The grocery store assigns SKU numbers to these products by sticking scanner labels on the items.
  - Although the bar codes are not UPCs they are certainly SKU numbers.

The Grocery Store Example/2

- Data is collected at several places in a grocery store.
  - Some of the most useful data is collected at the cash registers as customers purchase products.
  - Our modern grocery store scans the bar codes directly into the point-of-sale (POS) system.
  - The POS system is at the front door of the grocery store where customer takeaway is measured.
  - The back door, where vendors make deliveries, is another interesting data-collection point.

The Grocery Store Example/3

- At the grocery store, management is concerned with the logistics of ordering, stocking the shelves, and selling the products **while maximizing the profit** at each store.
- 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.
- The most significant decisions have to do with pricing and promotions.
  - Both store management and headquarters marketing spend a great deal of time tinkering with pricing and running promotions.
  - Promotions in a grocery store include temporary price reductions, ads in newspapers and newspaper inserts, displays in the grocery store, and coupons.

Simplified DM Design Process (Kimball and Ross)

- A somehow simplified DM design process consists of the following 4 steps:
  1. Choose the business process(es) to model
     - e.g., Sales,
  2. Choose the granularity of the business process
     - e.g., Items by Store by Promotion by Day.
     - Low granularity is needed.
     - Are individual transactions necessary/feasible?
  3. Choose the dimensions
     - Time, Store, ...
  4. Choose the measures
     - Dollar_sales, unit_sales, dollar_cost, customer_count

Step 1: Choose the Business Process

- A business process is an activity in the organization that typically is supported by a source data management system
  - raw material purchasing, orders, shipments, invoicing, inventory, bank transfers, patient transfers, ...
- Business processes are not necessarily limited to a single department
  - e.g., sales and marketing departments might be interested in the orders
- Focusing on the business process rather than the department avoids duplication of work and keeps data more consistent
- 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

Step 1: Choose the Business Process – Example

- Management wants to better understand customer purchases as captured by the POS system.
- Business process: POS retail sales
- Allows us to analyze:
  - What products are selling?
  - In which stores?
  - On what days?
  - Under what promotional conditions?
  - etc.

Step 2: Choose the Grain of the Business Process

- Preferably develop dimensional models for the most atomic information captured by a business process
  - Not because queries report individual rows, but queries need to cut through the details in very precise ways
- The more detailed/atomic data is, the more things we know
- Atomic data provides maximum analytic flexibility
- Can be constrained and rolled up in every possible way
- It is always possible to declare higher-level grains by aggregation of atomic data; the opposite is not true
- Less granular model is vulnerable to unexpected requests for more details
- Example grain declarations
  - Individual line item on a customer's sales ticket as measured by a scanner
  - An individual boarding pass of a flight
  - A monthly snapshot for each bank account

Step 2: Choose the Grain of the Business Process – Example

- Individual line item on a POS transaction is the most detailed data, and we choose this as grain ($\rightarrow$ event fact)
- Allows a very detailed analysis of sales
  - Difference in sales on Monday vs. Sunday
  - Is it worthwhile to stock so many individual sizes of certain brands?
  - How many shoppers took advantage of the 50-cents-off promotion on shampoo?
  - Impact in terms of increasing sales when a competitive diet soda product was promoted
  - etc.
- Note that none of these queries calls for data from a specific transaction, but require detailed ways to slice data
  - Could not be answered if only aggregated values would be stored, e.g., daily summaries

Step 3: Choose the Dimensions

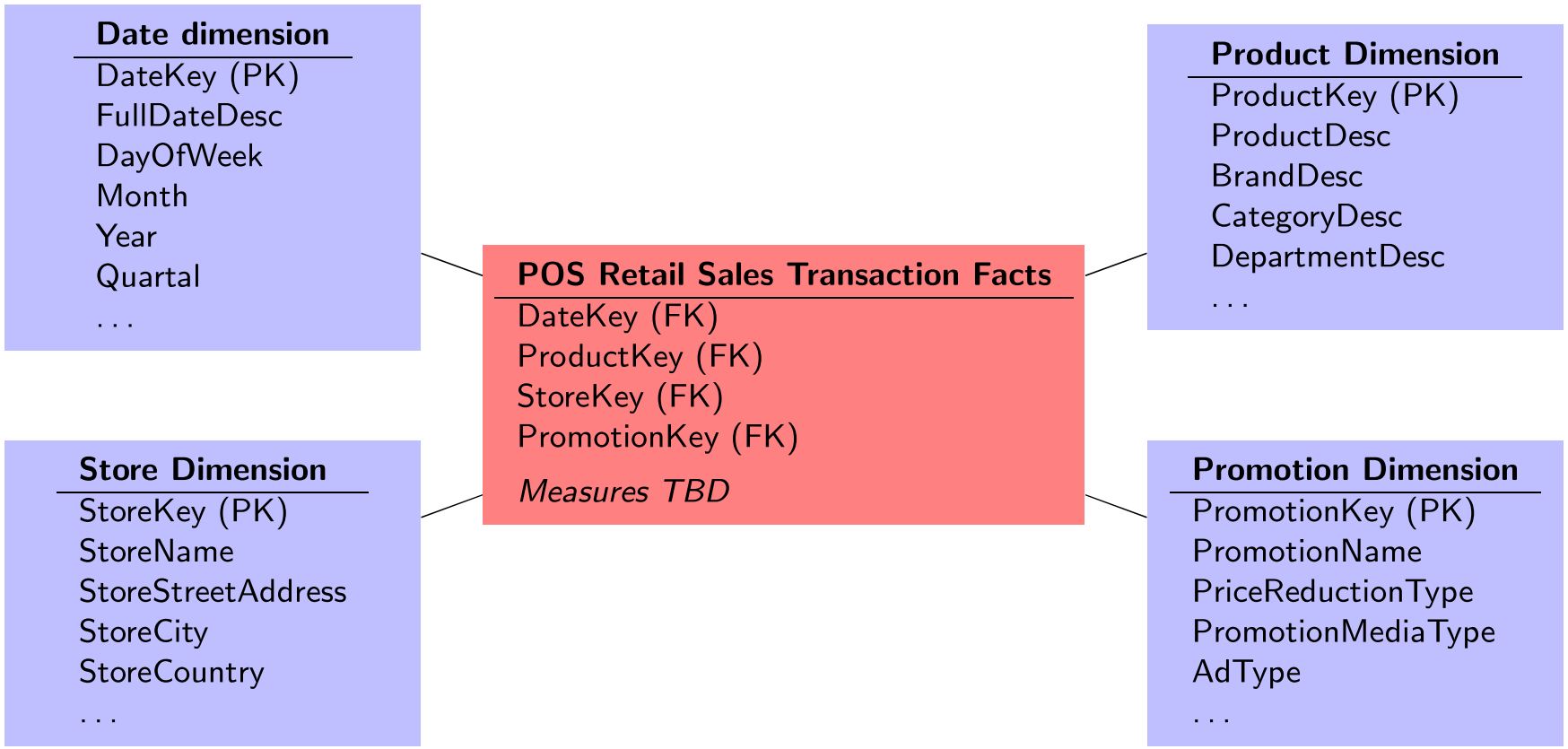
- Dimensions can be derived by answering the question *"How do business people describe the data resulting from the business processes?"*
- "Decorate" fact tables with dimensions representing all possible descriptions of the facts/measures
- A clear grain statement helps to identify the dimensions
- Sometimes a revision of step 2 is required

Step 3: Choose the Dimensions – Example/1

- The Date dimension
  - Explicit date dimension is needed (events, holidays, . . .)
- The Product dimension
  - Hierarchy allows drill-down/roll-up through category, brand, department, etc.
  - Many descriptive attributes (often more than 50)
- The Store dimension
  - Primary geographic dimension to specify location of the store
  - Many descriptive attributes
- The Promotion dimension
  - Used to see if promotions work and are profitable
  - Ads, price reductions, end-of-sale displays, coupons
    - Highly correlated (only 5000 combinations)

Step 3: Choose the Dimensions – Example/2

- Preliminary version of grocery store schema



Step 4: Choose the Measures

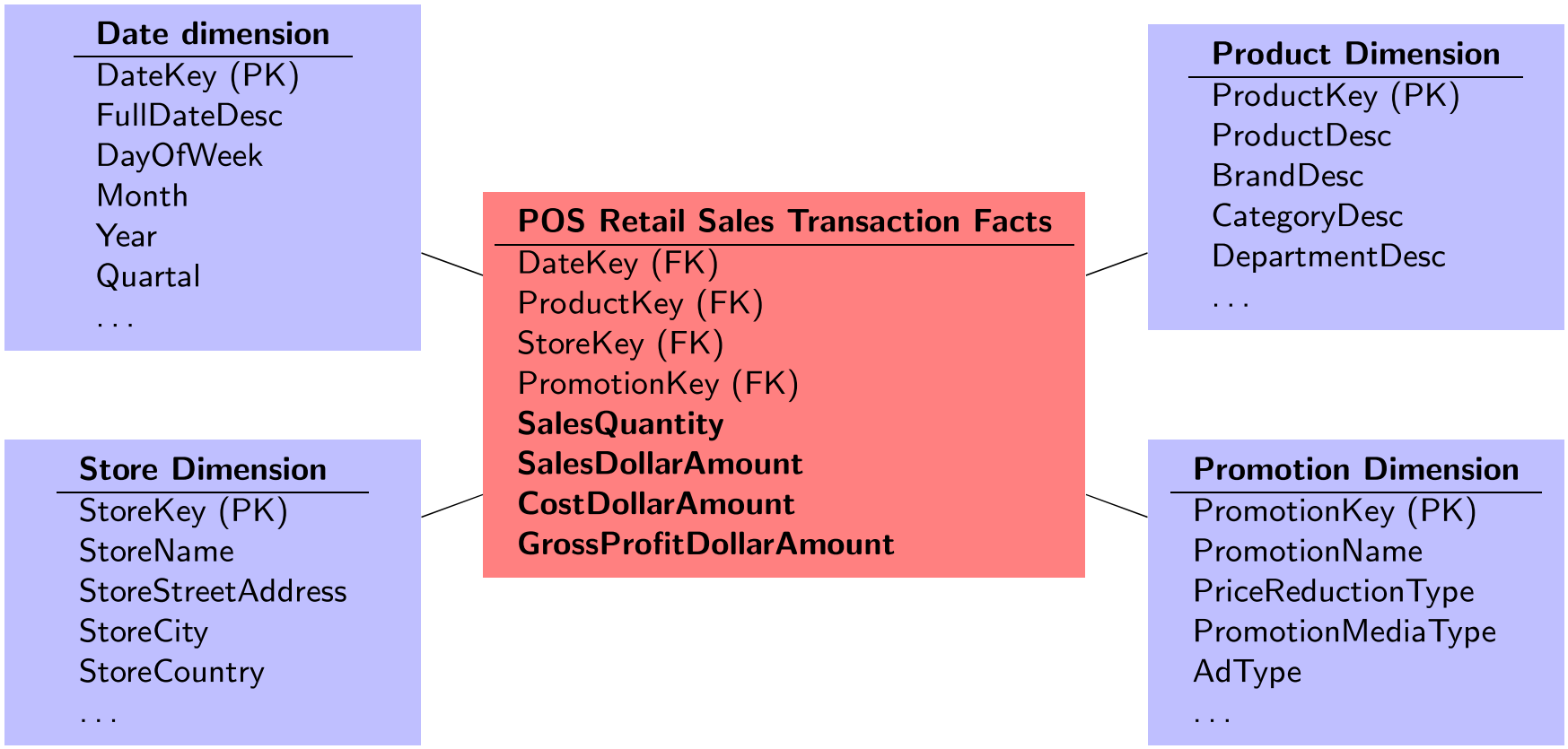
- Identify the business performance measures for the selected processes
- Determined by answering *"What are we measuring?"*
- Measures are determined by the grain declaration
- Revision of the grain might be required

Step 4: Choose the Measures – Example

- Sales quantity
- Sales dollar amount
- Cost dollar amount
- Gross profit
  - Equals to sales dollar – cost dollar
  - Explicit storage avoids user errors

- All measures are additive
- Facts are classified as event facts since each POS transaction is stored (most detailed level)

MD Schema of the Grocery Store Example



Outline

Database Sizing

- Time dimension: 2 years = 730 days
- Store dimension: 300 stores reporting each day
- Product dimension: 30,000 products, only 3,000 sell per day
- Promotion dimension: 5,000 combinations, but a product only appears in one combination per day
- Number of fact records: $730 \times 300 \times 3,000 \times 1 = 657,000,000$
- Number of fields: 4 key + 4 measures = 8 fields
- Total DB size: 657,000,000 × 8 fields × 4 bytes = 21 GB
- Small database by today's standards!

Date Dimension

- Date dimension is present in all DWs
- Can be created in advance
- "Meaningful" values are important for report generation, etc.
  - e.g., Holiday/Nonholiday vs. Yes/No
- Time-of-day a separate dimension
  - Separation keeps both dimensions small
- Date dimension vs. SQL date type
  - Many date attributes are not supported in SQL, e.g., fiscal month
  - Business user is not versed in SQL
- Date dimension is relatively small
  - 10 years = 3,650 rows

Date Dimension
DateKey (PK)
Date
Full Date Description
Day Of Week
Day Number in Epoch
Week Number in Epoch
Day Number in Calendar Month
Day Number in Calendar Year
Last Day in Week Indicator
Last Day in Month Indicator
Calendar Week Ending Date
Calendar Quarter
Calendar Year-Quarter
Calendar Half Year
Calendar Year
Fiscal Week
Fiscal Month
Fiscal Quarter
Fiscal Half Year
Fiscal Year
Holiday Indicator
Weekday Indicator
Selling Season
...

Instance of Date Dimension

Date Dimension Table

DK	Date	FullDateDescription	DayOfWeek	DayNum	HolidayInd	WeekdayInd	...
1	29.09.2013	September 29, 2013	Sunday	29	Nonholiday	Nonweekday	
2	30.09.2013	September 30, 2013	Monday	30	Nonholiday	Weekday	
3	01.10.2013	October 1, 2013	Tuesday	1	Nonholiday	Weekday	
4	02.10.2013	October 2, 2013	Wednesday	2	Nonholiday	Weekday	

Product Dimension

- Description of the products
- >50 attributes is typical for Product dimension
- Concept hierarchy
  - SKU → Brand → Category → Department
  - Many repetitions, but space of dimensions is not critical

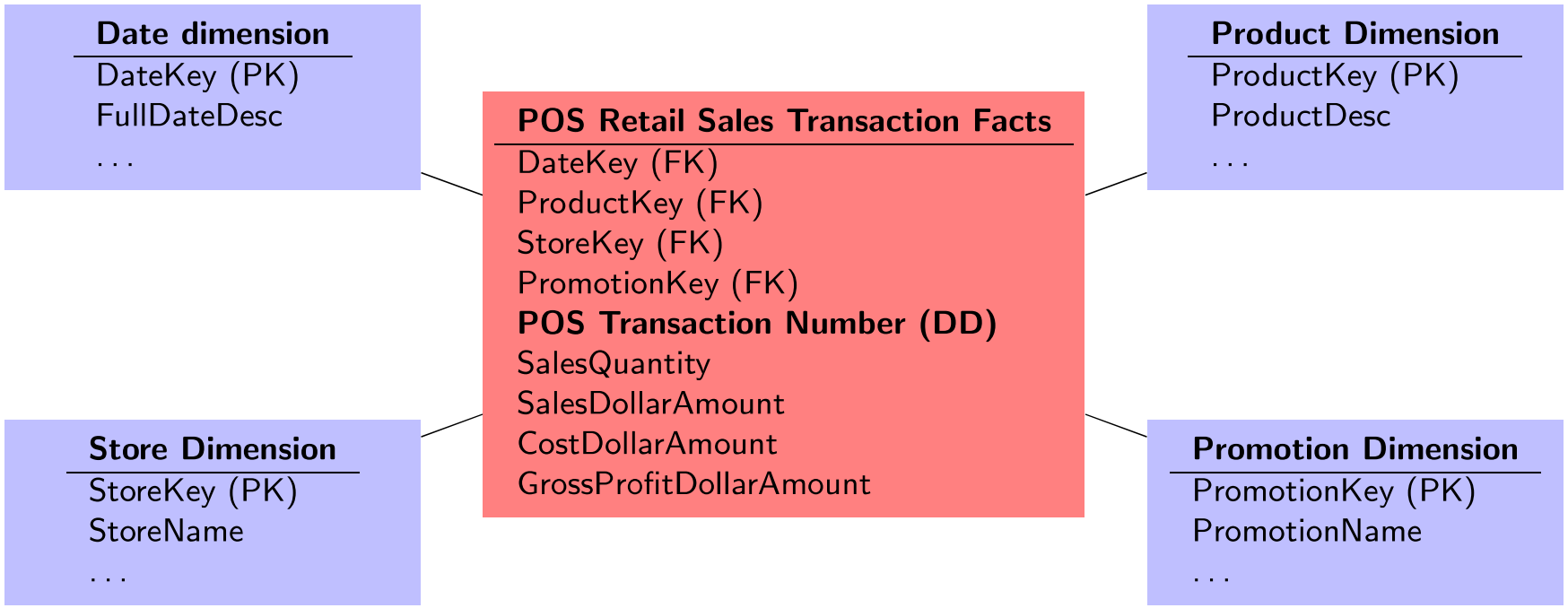
Product Dimension
Product Key (PK)
Product Description
SKU Number
Brand Description
Category Description
Department Description
Package Type Description
Package Size
Fat Content
Diet Type
Weight
Weight Units of Measure
...

Instance of Product dimension table

PK	Product Description	Brand Desc	Cat Desc	Dept Desc	...
1	Baked Well Light	Baked Well	Bread	Bakery	
2	Fluffy Sliced Whole Wheat	Fluffy	Bread	Bakery	
3	Fluffy Light Sliced Whole Wheat	Fluffy	Bread	Bakery	
4	Fat Free Mini Cinnamon Rolls	Light	Sweeten Bread	Bakery	
5	Diet Lovers Vanilla 2 Gallon	Coldpack	Frozen Desserts	Frozen Foods	
6	Light and Creamy Butter Pecan 1 Pint	Freshlike	Frozen Desserts	Frozen Foods	

Degenerate Dimensions

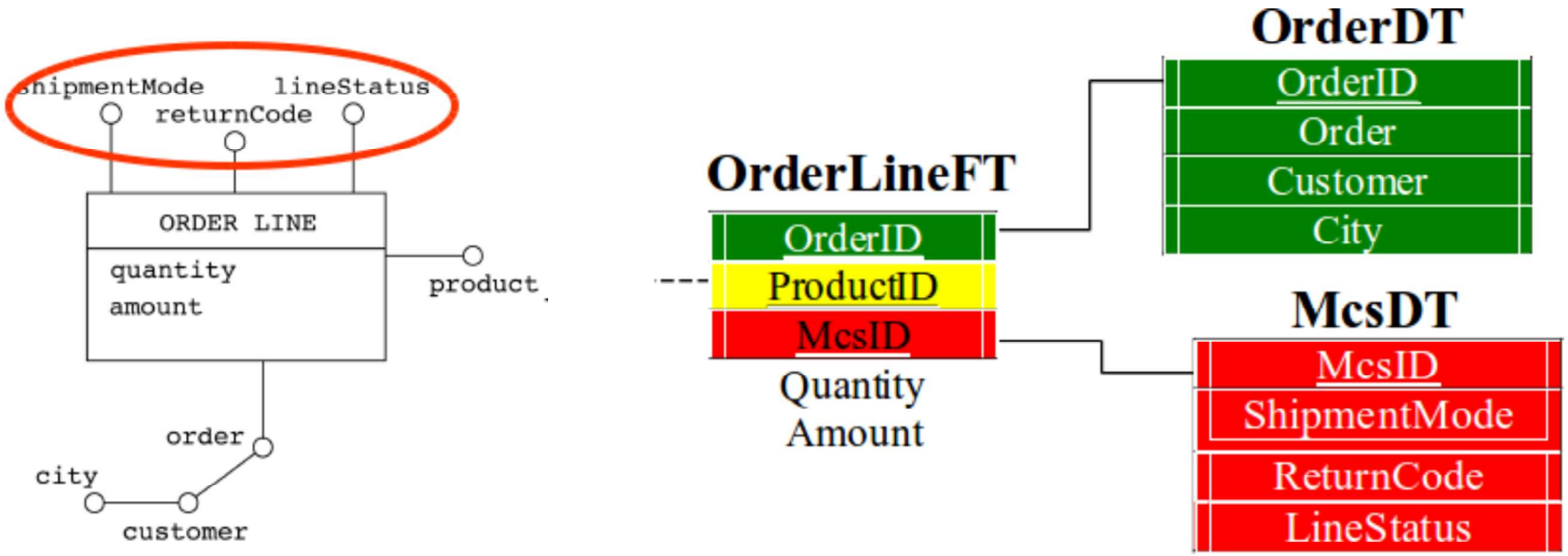
- Degenerate dimensions are "empty"
  - i.e., dimensions with a "hierarchy" of only one attribute
  - Values are directly stored in fact table, no dimension table is needed
- Examples are operational control numbers, e.g., order #, invoice #, POS transaction #, etc.
- Useful to serve as part of primary key in fact table or for grouping
  - e.g, grouping by POS transaction number to retrieve all products purchased in a single transaction



Junk Dimensions

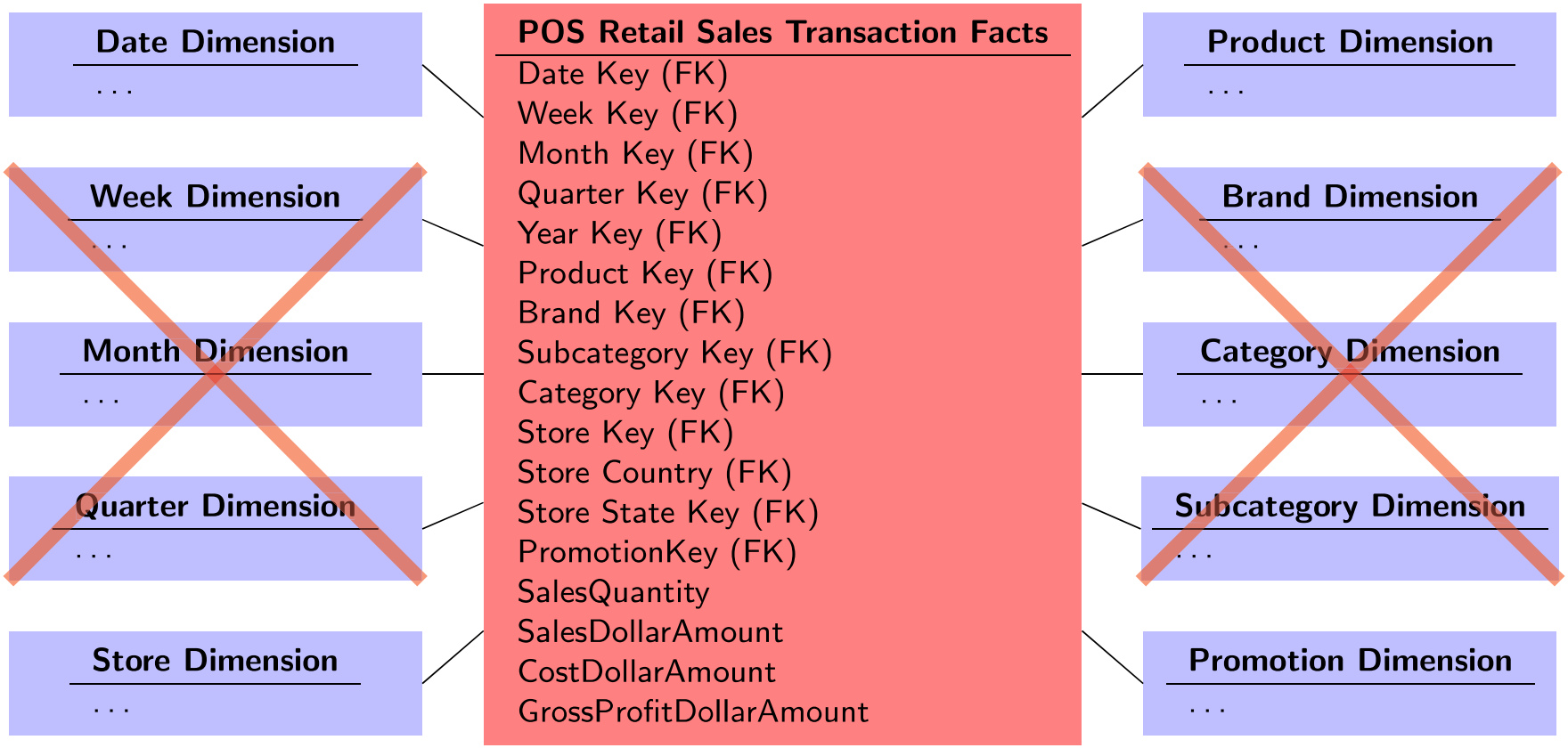
- Junk dimension: a dimension table that stores several degenerate dimensions
  - Within a junk dimension there is no functional dependency (hierarchy)
  - Only feasible if the number of distinct values for the attributes is small

- **Example:** 3 generate dimensions combined in a single junk dimension



How Many Dimensions?

- Dimensions are important to provide detailed information for the analysis!
- But, too many dimensions is bad!
  - A sign that dimensions are not independent, and hence should be combined
  - Significantly increases space requirements of fact table
    - Size of dimension table is not a problem
  - 15 dimensions should normally be enough



Summary/1

- Simplified DW design process by Kimball and Ross consists of 4 steps:
  - Choose business processes, granularity, dimensions, and measures
- Surrogate key should be used instead of operational codes
- Degenerate dimensions are stored in the fact table
  - or stored all together in a junk dimension
- DW sizing
  - dimensions are a small portion of the DW, hence can/should contain as much information as possible
  - fact table determines the size of the DW
- Dimensional model is easy to use, e.g., drag and drop for report generation