Class 6

Dimensional modeling

Learning Objectives

- Define the notions in dimensional modelling
- Understand the basic design principles
- Study a real-world case study
- Provide a basis for next steps in dimensional modelling

Dimensional modeling vocabulary

- Fact Table
  - Is the primary table in a dimensional model
  - Facts are numeric measurements (values) that represent a specific business aspect or activity
  - Facts can be computed or derived at run-time (metrics).
  - Have two or more foreign keys(FK) that connect to the dimension table's primary keys
    - Satisfy referential integrity
  - Generally has own primary key(called a composite or concatenated key) made up of a subset of the foreign keys
  - Express the many-to-many relationships between dimensions

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

facts (Quantity Sold, Dollor Sales Amount)

Dimensional modeling vocabulary

- Dimension Tables
  - Are integral companions to a fact table
  - Contain the textual descriptors of the business
  - Have many columns or attributes
  - Defined single primary key(PK)
  - We strive to minimize the use of codes in our dimension tables by replacing them with more verbose textual attributes
    - Operational codes often have intelligence embedded in them
  - Typically are highly denormalized
  - Typically are geometrically smaller than fact tables, improving storage efficiency by normalizing or snowflaking
    - Snowflake
      - Brand description and category description replace by brand code and create brand 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

Sample dimension Table

Dimensional modeling vocabulary

- Surrogate Keys
  - rather than operational production codes (:natural keys)
  - are also called as meaningless keys, integer keys, nonnatural keys, artificial keys, synthetic keys
  - are integers that are assigned sequentially as needed to populate a dimension
- Every join between dimension and fact tables should be based on meaningless integer surrogate keys
  - We want to avoid embedding intelligence in the data warehouse keys
    - because any assumptions that we make eventually may be invalidated
  - Queries and data access applications should not have any built-in dependency on the keys
    - because the logic also would be vulnerable to invalidation
- We want to discourage the use of concatenated or compound keys for dimension tables
  - to avoid multiple parallel joins between the dimension and fact tables

Dimensional modeling vocabulary

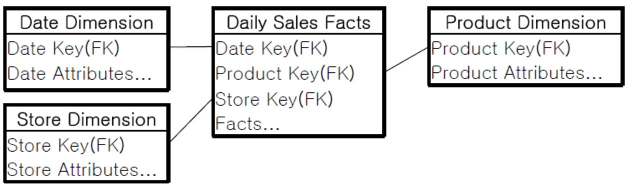
- Surrogate Keys Benefits
  - The surrogate keys buffer the data warehouse environment from operational changes
    - Surrogate keys allow the data warehouse team to integrate data from multiple operational source systems
  - The surrogate key is as small an integer as possible while eusuring that it will accommodate maximum number of rows in the dimension
    - Typically, a 4-byte integer is sufficient to handle most dimension situations
  - The surrogate keys are used to record dimension conditions that may not have an operational code
    - "Date to be Determined" or "Date Not Applicable"
  - Treating the surrogate date key as a date sequence number will allow the fact table to be physically partitioned on the basis of the date key
    - The partitioning is highly effective because it allows old data and new data to be loaded and indexed without disturbing the rest of the fact table

Dimensional modeling vocabulary

- Dimension Table Attributes
  - serve as the primary source of query constraints, groupings, and report labels
    - In a query or report request, attributes are identified as the “by” words
    - Ex) dollar sales by week by brand
  - Key to making the DW usable and understandable
  - The best attributes are textual and discrete
    - Consist of real words

Dimensional modeling vocabulary

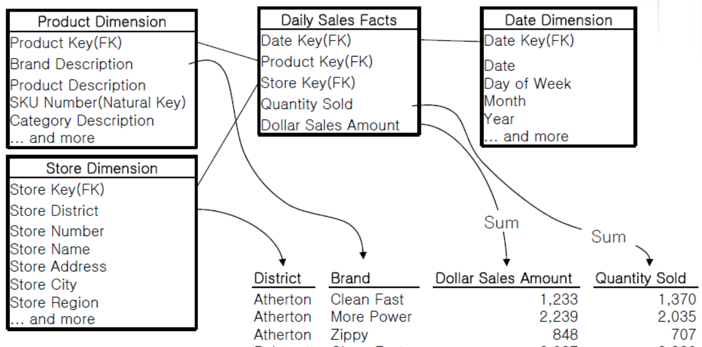
- Bringing Together Facts and Dimensions
  - The fact table consisting of numeric measurements is joined to a set of dimension tables filled with descriptive attributes
  - This characteristic star-like structure is often called a star join schema
  - All dimension are symmetrically equal entry points into the fact table
    - No preferences for any query
  - We Certainly don’t want to adjust our schemas if business users come up with new ways to analyze the business



Fact and dimension Tables in a dimensional model

Dimensional modeling vocabulary

- Bringing Together Facts and Dimensions



District	Brand	Dollar Sales Amount	Quantity Sold
Atherton	Clean Fast	1,233	1,370
Atherton	More Power	2,239	2,035
Atherton	Zippy	848	707
Belmont	Clean Fast	2,097	2,330
Belmont	More Power	2,428	2,207
Belmont	Zippy	633	527

Dragging and dropping dimensional attributes and facts into simple report

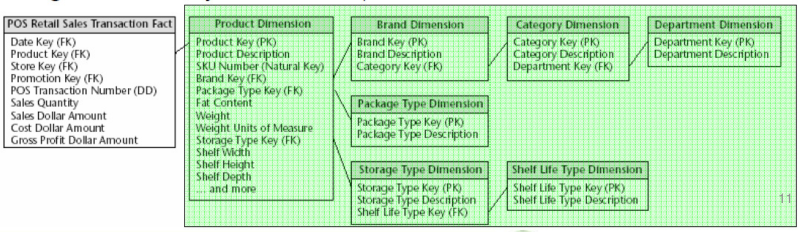
Dimensional modeling vocabulary

- Fact vs Dimension Attribute
  - Fact
    - The field is a measurement that takes on lots of values and participates in calculation
    - Ex) standard cost for a product is fact
      - seems like a constant attribute of the product but may be changed so often that eventually
  - Dimension attribute
    - The field is a discretely valued description that is more or less constant and participates in constraints
  - Occasionally, we can't be certain of the classification

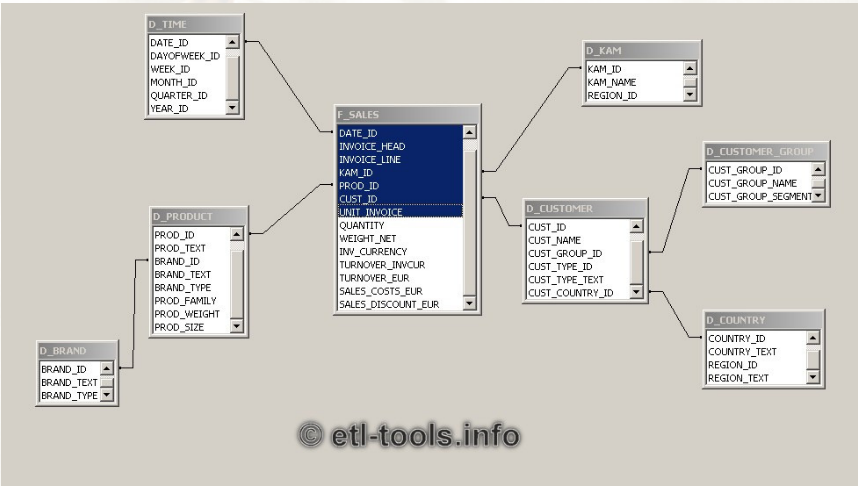
    -> it may be possible to model the data field either way, as a matter of designer's prerogative

Resisting Comfort Zone Urges

- Breaking some traditional modeling rules
  - focused on delivering business value through ease of use and performance, not on transaction processing efficiencies
- Dimension Normalization (Snowflaking)
  - Redundant attributes are removed from the flat, denormalized dimension table and placed in normalized secondary dimension tables
  - While the fact tables in both figures are identical, the plethora of dimension tables is overwhelming
  - Fig 2.12 Partially snowflaked product dimension



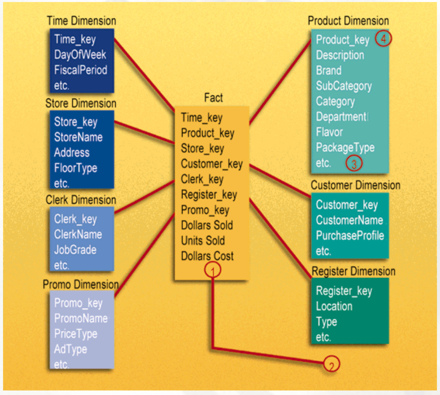
Resisting Comfort Zone Urges

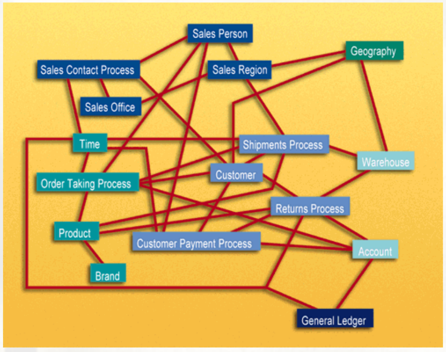


Resisting Comfort Zone Urges

❖ Dimension Normalization (Snowflaking)
- While snowflaking is a legal extension of the dimensional model
- We encourage you to resist the urge to snowflake with ease of use and performance
  - The multitude of snowflaked tables makes for a much more complex presentation
  - Numerous tables and joins usually translate into slower query performance
  - The minor disk space savings associated with snowflaked dimension tables are insignificant
    - Disk space savings grained by normalizing the dimension tables typically are less than 1 percent of the total disk space needed for the overall schema
  - Snowflaking slows down the users' ability to browse within a dimension
  - Snowflaking defeats the use of bitmap indexes

Resisting Comfort Zone Urges





http://www.dbmsmag.com/9708d15.html

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

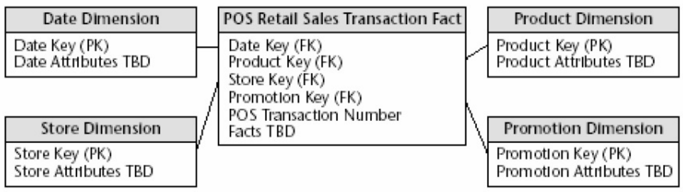
- Brief description of the retail business (2/2)
  - The remaining 5,000 SKUs come from departments such as meat, produce, bakery, or floral
  - While these products don't have nationally recognized UPCs, the grocery chain assigns SKU numbers to them
  - The bar codes are not UPCs, they are certainly 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

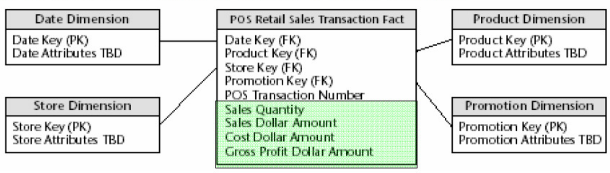
- Dimensional Design Process
  - Step3. Choose the dimensions
    - Fig 2.2 Preliminary retail sales schema



(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

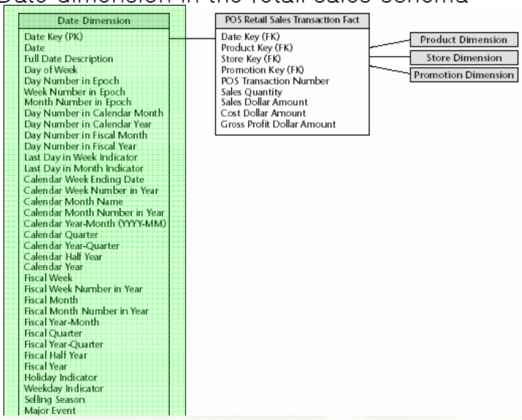


(TBD means to be determined)

- 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

Dimension 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



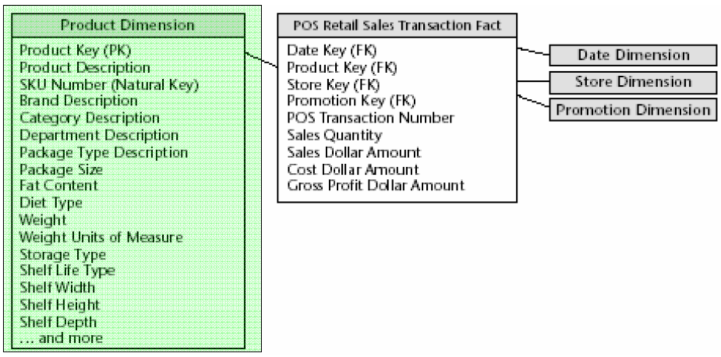
Dimension Table Attributes

- 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

Dimension Table Attributes

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



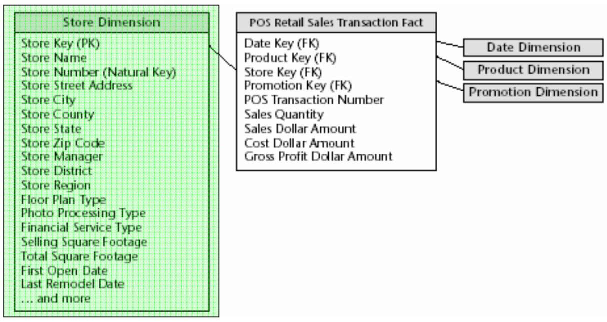
Dimension Table Attributes

- 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

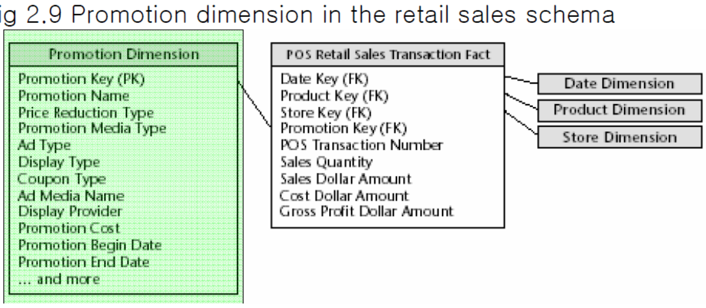
Dimension Table Attributes

- 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



Dimension 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



Dimension Table Attributes

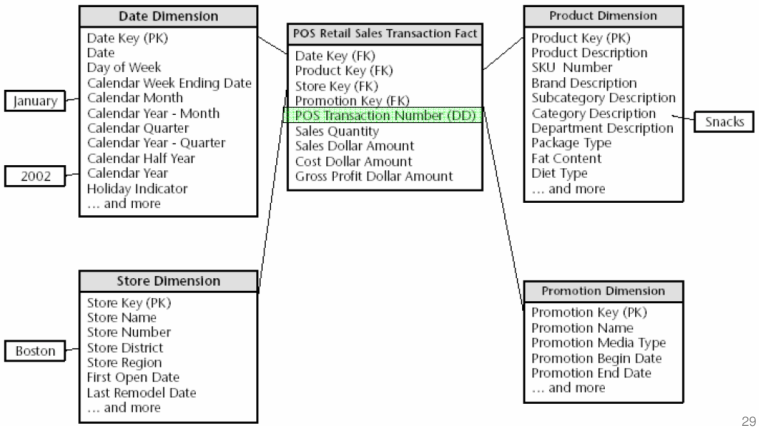
- Promotion Dimension
  - The various possible causal conditions are highly correlated
    - A temporary price reduction is associated with an ad and an end-aisle display
    - Coupons often are associated with ads
- For four major causal mechanisms (price reductions, ads, displays, and coupons)
  - 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

Dimension 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

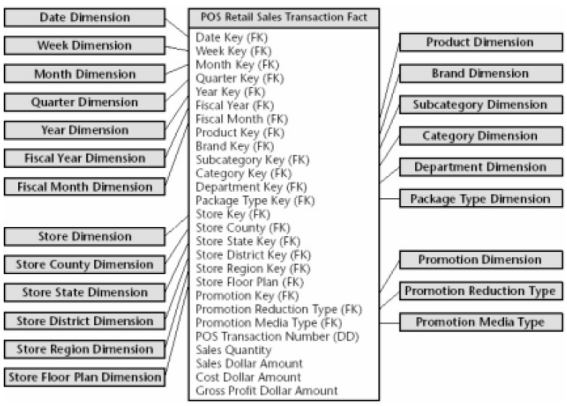
Dimension Table Attributes

- 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