

## CSI4142: Data Science

### Topic 2: Data marts for Analytics Applications

(Slides by HL Viktor ©: material from Kimball and Ross, Chapters 1, 2, 3, 10, 17, 18 and Han Chapter 3)

1

---

---

---

---

---

---

---

---

## Overview of topic

1. Supporting decisions:  
from Online Transaction Processing (OLTP) to Online Analytical Processing (OLAP)
2. Data warehouses defined
3. Data marts defined
4. Business Dimensional Life Cycle of Kimball
5. Creating a data mart:
  - a. Conceptual (Dimensional) modelling
  - b. Physical Design
  - c. Data staging



2

---

---

---

---

---

---

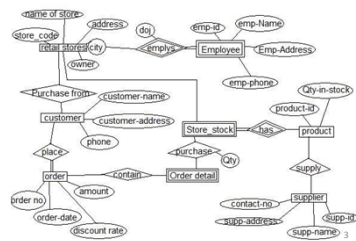
---

---

## Recall from CSI2132:

### Online Transaction Processing (OLTP)

- Entity relationship diagrams
- Relational model (PKs and FKs)
- Records transaction flows




---

---

---

---

---

---

---

---

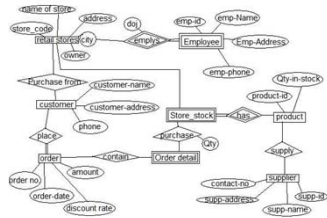
## Online Transaction Processing (OLTP)

Operations/Transactions:

- INSERT
- DELETE
- UPDATE
- QUERY

DBMS:

- Concurrency control
- Recovery
- Security
- Etc.



4

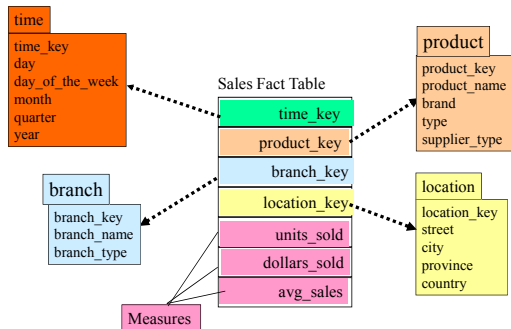
## Case Study: Clothing Store with 10 Branches in Ontario (OLTP)

- Open 9h30

- Close 21h00

5

## Online Analytic Processing (OLAP)



6

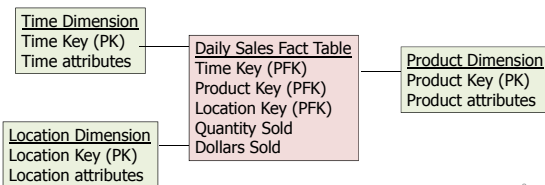
### Case Study: Clothing Store with 10 Branches in Ontario (OLAP)

- Open 9h30
- .....
- Close 21h00
- Data staging at 23h30

7

### Online Analytic Processing (OLAP)

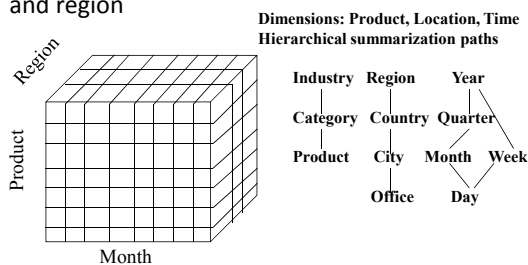
- Data uploaded periodically
- Queries
  - Aggregates (SUM, COUNT, MIN, MAX, AVERAGE)
  - Trends and Icebergs



8

### Multidimensional Data

- Sales volume as a function of product, month, and region



9

### From Tables and Spreadsheets to Data Cubes

- A data warehouse is based on a **multidimensional data model** which views data in the form of a DATA CUBE
- A data cube, such as **SALES**, allows data to be modeled and viewed in multiple dimensions
  - **Dimension tables**, such as **item** (item\_name, brand, type), or **time**(day, week, month, quarter, year)
  - **Fact table** contains **measures** (such as **dollars\_sold**) and keys to each of the related dimension tables

10

---

---

---

---

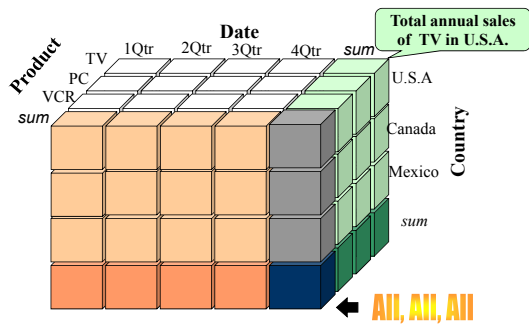
---

---

---

---

### A Sample Data Cube



11

---

---

---

---

---

---

---

---

### OLTP vs. OLAP

	OLTP	OLAP
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

12

---

---

---

---

---

---

---

---

## Why a Separate Data Warehouse?

- High performance for both systems
  - **DBMS**—tuned for **OLTP**: access methods, indexing, concurrency control, recovery
  - **Warehouse**—tuned for **OLAP**: complex OLAP queries, multidimensional view, consolidation
- Different functions and different data:
  - **missing data**: Decision support requires historical data which operational DBs do not typically maintain
  - **data consolidation**: DS requires consolidation (aggregation, summarization) of data from heterogeneous sources
  - **data quality**: different sources typically use inconsistent data representations, codes and formats which have to be reconciled
- Note: There are more and more systems which perform OLAP analysis directly on relational databases

13

---

---

---

---

---

---

---

---

## So, what is a Data Warehouse?

- Definitions:
  - A decision support database that is maintained separately from the organization's operational database
  - Support information processing by providing a solid platform of consolidated, historical data for analysis.
- "A data warehouse is a **subject-oriented, integrated, time-variant, and nonvolatile** collection of data in support of management's decision-making process."—W. H. Inmon
- Data warehousing:
  - The process of constructing and using data warehouses



14

---

---

---

---

---

---

---

---

## Data Warehouse—Subject-Oriented

- Organized around major subjects, such as customer, product, sales
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process



15

---

---

---

---

---

---

---

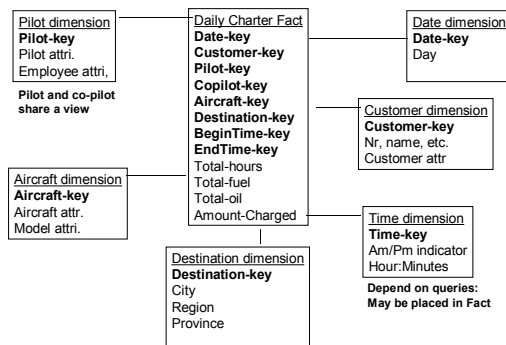
---

## Data marts

- Data warehouse (DW) consists of one or more data mart
- Data mart corresponds to a SUBJECT
- Examples:
  - Insurance Claims
  - Inventory Management : Store and Warehouse
  - Customer Relationships: Frequent Flyers
  - Financial Services: Banking trends
  - Telecommunications: Call tracking
  - Electronic Health Records
  - etc.

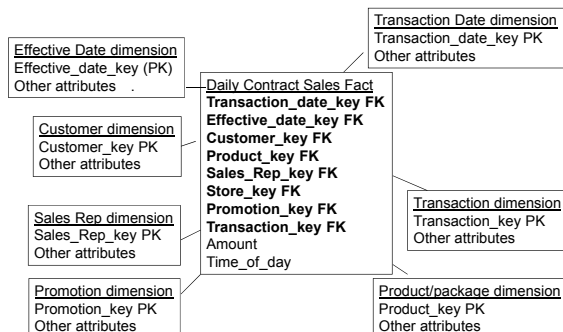
16

## Charter flights



17

## Mobile phone contract sales



## Data Warehouse—Integrated

- Constructed by integrating multiple, heterogeneous data sources
  - relational databases, flat files, on-line transaction records
- Data cleaning and data integration techniques are applied.
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
    - E.g., Hotel price: currency, tax, breakfast covered, etc.
  - When data are moved to the warehouse, it is converted.

19

---

---

---

---

---

---

---

---

## Data Warehouse—Time Variant

- The time horizon for the data warehouse is significantly longer than that of operational systems
  - Operational database: current value data
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
- Every key structure in the data warehouse
  - Contains an element of time, explicitly or implicitly
  - But the key of operational data may or may not contain "time element"

20

---

---

---

---

---

---

---

---

## Data Warehouse—Nonvolatile

- A physically separate store of data transformed from the operational environment
- Operational update of data does not occur in the data warehouse environment
  - Does not require transaction processing, recovery, and concurrency control mechanisms
  - Requires only two operations in data accessing:
    - initial loading of data and access of data

21

---

---

---

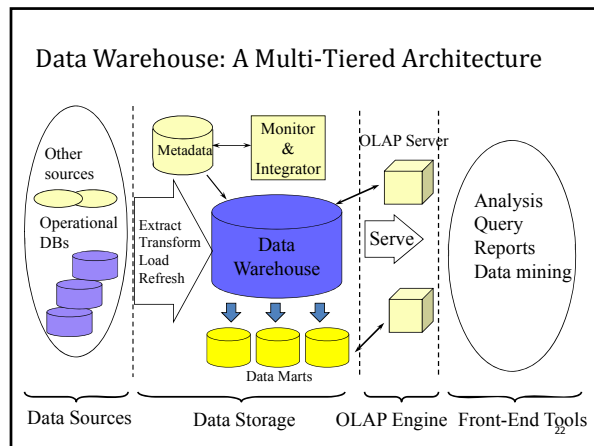
---

---

---

---

---




---

---

---

---

---


---

---

---

### Building a Data Warehouse

Business Life Cycle Toolkit  
Kimball et. al.  
(<http://decisionworks.com/>)



23

---

---

---

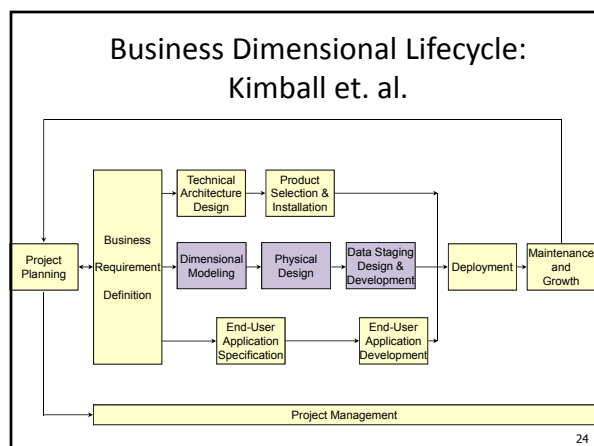
---

---

---

---

---




---

---

---

---

---

---

---

---



## Data Track:

Steps to create a single data mart

1. Dimensional modeling
2. Physical Design
3. Data staging (extract, transform and load)

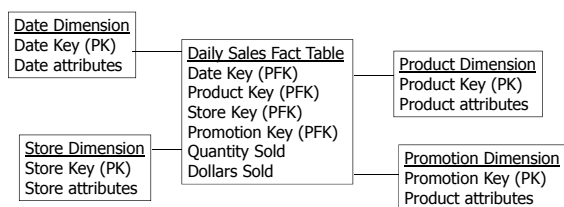
25

## Dimensional Model Components

- **FACT table:** Primary table where numeric **performance measures** for a business process are stored
  - Composite PK from many FKs
  - **Facts:** A business measure (numeric, additive)
- **Dimensional tables**
  - Contains textual description of business
  - MANY dimensional attributes
  - Used to **specify query constraints**

26

## Dimensional Model Components: The classic example

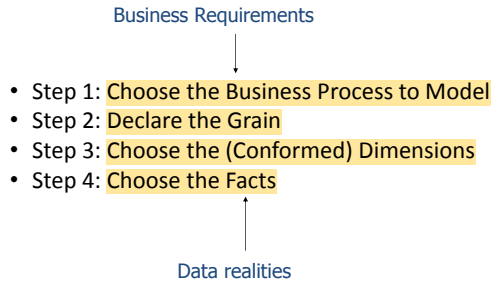


The Star Join Schema

PFK is shorthand for "Primary and Foreign Key"

27

## Four-Step Method to Designing an Individual Dimensional Model



28

---

---

---

---

---

---

---

---

### For example: Reduced time to market

- Average revenue from new products: \$50,000 per month
- After data warehouse: products to market 6 weeks sooner (1.5 month)
- Number of new products per year: 15
- Incremental revenue per year:  
\$50,000 each month x 1.5 months x 15 products  
→ approximate \$1,125,000 incremental revenue per year

29

---

---

---

---

---

---

---

---

### Step 2: Declare the Grain

- Answer the following question: "How do you describe a single row in the fact table?"
  - An individual line item on a customer's retail sales ticket as measured by a scanner
  - A line item on a bill received from a doctor
  - An individual boarding pass to get on a flight
  - An individual phone call made from this phone number
- ALWAYS choose the LOWEST possible (and of course meaningful) grain of each dimension → we want to see the details

30

---

---

---

---

---

---

---

---

### Step 3: Choose the Dimensions

- Answer question: “How do businesspeople describe the data that results from the business process”?
- Determined by grain of fact table
- E.g. Line item fact
  - Order date, customer, produce, order number, etc.
  - Add all possibly relevant dimensions and many describe attribute values (discrete, text-like attributes)

31

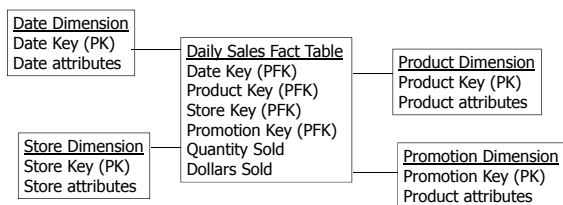
### Step 4: Choose the Facts

- Answer question “What are we measuring?”
- Specific to grain of fact table
- Store additive values: E.g. quantity-sold, taxes, dollars-sold, etc.
- Percentages and ratios: Also store numerator and denominator
- Usually *numeric and additive*

– Some authors refer to facts as measures

32

### Dimensional Model Components: The classic example explained



The Star Join Schema

PFK is shorthand for “Primary and Foreign Key”

33

## Sales: The Date Dimension

- Nearly guaranteed to be in the data mart

### Date Dimension (3650 rows to cover 10 years)

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,  
Last Day in Week Indicator, Last Day in month Indicator,  
Calendar Week Ending Date, Calendar Week Number in Year,  
Calendar Month Number in Year, Calendar Month Name,  
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, etc. 34

---

---

---

---

---

---

---

---

## Sales: The Product dimension

### Product dimension

Product\_key (PK), SKU\_description, SKU\_number, package\_size, brand,  
subcategory, category, department, package\_type, diet\_type,  
weight, weight\_unit\_of\_measure, units\_per\_retail\_case,  
units\_per\_shipping\_case, shelf\_width, shelf\_height, shelf\_depth,  
shelf\_unit\_of\_measure... and many more

### *An example row:*

1000, Green 3-pack Brawny Paper Towers, UPC#142142414, 3-pack,  
Brawny, Paper towers, Paper, Grocery, Bag, No, 300, grams, 100, 3000,  
30, 20, 60, cm,...

- SKU means "stock keeping unit"
- UPC means "universal product codes" → bar code

35

---

---

---

---

---

---

---

---

## Sales: The Store dimension

### Store dimension

Store\_key (PK), store\_name, store\_number, store\_street\_address,  
store\_city, store\_province, store\_zip, sales\_district, sales\_region,  
store\_manager, store\_phone, store\_fax, floor\_plan\_type,  
photo\_processing\_type, financial\_services\_type, first\_opened\_date,  
last\_remodel\_date, store\_sqm, grocery\_sqm, frozen\_sqm, meat\_sqm,  
... and many more

### *An example row:*

2000, Sandy Hill, 121, 10 King Edward Road, Ottawa, Ontario, 1K1 N1H,  
East, Eastern Canada, John Doe, (613) 342 1232, (613) 351 2212,  
Square, 48 hours, none, 1 May 2001, 1 May 2001, 2421, 353, 42, 34,  
...

36

---

---

---

---

---

---

---

---

## Sales: The promotion dimension

### Promotion dimension

Promotion\_key, promotion\_name, **price\_reduction\_type**,  
price\_reduction, price\_reduction\_unit, **ad\_type**,  
**display\_type**, **coupon\_type**, ad\_media\_name, display\_provider,  
promo\_cost, promo\_cost\_unit, promo\_begin\_date, promo\_end\_date,  
..., and many more

### *An example row:*

1000, Brawny paper towels, Discount, 0.30, CA\$, newspaper,  
end\_of\_aisles, none, Ottawa Citizen, store, 20,000, CA\$, 01/09/03,  
07/09/03, ....

### *Another (important) row*

2000, null, null, null, null, ...

Used when there is no promotion on a given day

37

## Sales: Adding the facts/measures

### Daily Sales Fact Table

Date Key (PFK)  
Product Key (PFK)  
Store Key (PFK)  
Promotion Key (PFK)  
**Quantity\_sold**  
**CA\$\_revenue**  
**CA\$\_cost**  
**Customer\_count**

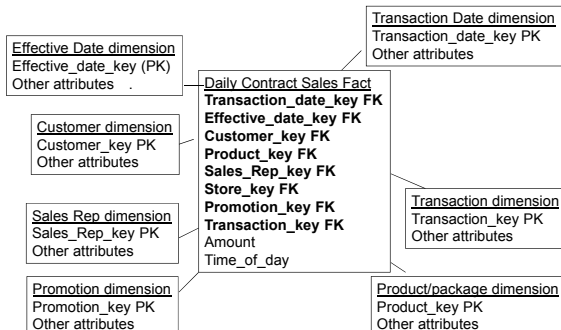
### Dimensions

**Date**  
**Product**  
**Store**  
**Promotion**

- Answer question "What are we measuring"?
- Depend on the grain of the fact table

38

## Mobile phone contract sales



## Creating the dimensional model:

### Types of dimensions

- **Causal**: promotion, contract, deal, etc.
- **Multiple date or timestamp**: date shipped, date received, etc.
- **Degenerate**: ticket number, order number
- **Role-playing**: one table acting in many "views"
- **Status**: account status
- **Audit**: data quality and record lineage ("when the record was loaded for the first time")
- **Junk**: indicators and flags



40

---

---

---

---

---

---

---

---

## More about dimensions: Role-playing Dimensions

### States of customer orders in Shipping Business

- Order date
- Packaging date
- Shipping date
- Delivery date
- Payment date
- Return date
- Refer to collection date
- Order status
- Customer
- Product
- Warehouse

Use a SQL View

41

---

---

---

---

---

---

---

---

## Multinational tracking and multiple units of measure

- Pound versus Kg, meters versus inches
- Time-zones, currency conversions



### Multinational Sales Fact Table

Date-key (FK)  
Product-key (FK)  
Store-key (FK)  
Reporting-country-key (FK)  
Customer-key (FK)  
Promotion-key (FK)  
Quantity-sold  
Local-currency-tendered  
CA\$-dollar-equivalent

### Daily Currency Conversion Fact Table

Date-key (FK)  
Buying-country-key (FK)  
Selling-country-key (FK)  
Conversion-rate

42

---

---

---

---

---

---

---

---

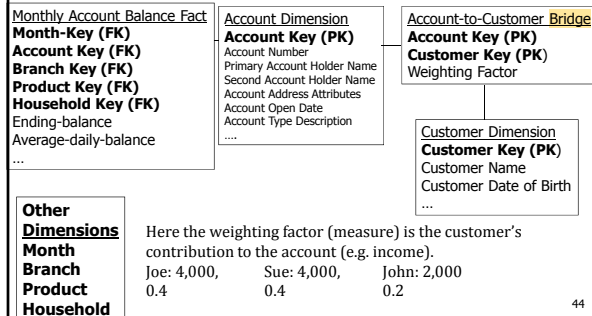
## Many-to-many Dimensions

What about **multivalued dimensions**, where Joe and Sue Smith share a credit card account?

- A **dimension** has 0, 1 or more than 1 value
- Number of values are unknown before creating dimensional model:
  - it (**the dimension**) acts as a **measured FACT**
- E.g. we want to be able to add the values
- **Use bridge**, otherwise we have to add many dimensions
- Useful for easy QUERYING: (e.g. Medical diagnosis)
  - “supply the weighted charges of the combined diagnosis” → amounts add up correctly
  - “supply a report of the cost (impact) of a particular diagnosis for that patient on that day”. E.g. diagnosing a cancerous tumor has a higher impact than the flu.

43

## Modeling the Banking environment: Multivalued Dimensions (M:M)

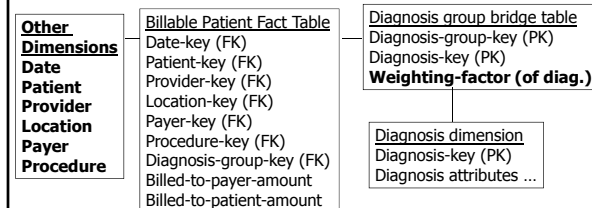


44

## More Many to Many Dimensions



- We use a **group bridge table**, where the weighting factor adds up to 1



45

## Modeling the Banking environment:

### Attribute banding

- Used to answer “banded queries”

#### Monthly Account Snapshot Fact

Month End date Key (FK)  
Branch Key (FK)  
Product Key (FK)  
Account Key (FK)  
Account Status Key (FK)  
Primary Month End Balance  
Average Daily Balance  
Number of Transactions  
Interest Paid  
Interest Charged  
Fees Charged

#### Band definition table

Band group Key (PK)  
Band group sort order (PK)  
Band group name  
Band range name  
Band lower value  
Band upper value

Use pair of  $\leq$  and  $>$  joins

46

---

---

---

---

---

---

---

---

## Attribute Banding:

### Avoid Monster Dimensions

Customer(Cust-key, lastname, firstname, gender, marital status, address, city, postal code, income, age, #children, occupation, etc.)

- Crucial for decision support

#### Band definition table

Band group Key (PK)  
Band group sort order (PK)  
Band group name  
Band range name  
Band lower value  
Band upper value

47

---

---

---

---

---

---

---

---

## Modeling Time...

- What is a month? a day? a week?

Location	Local Time	Time Zone	UTC Offset
<a href="#">Ottawa</a> (Canada - Ontario)	Monday, January 23, 2017 at 11:21:20 am	<a href="#">EST</a>	UTC-5 hours
<a href="#">Sydney</a> (Australia - New South Wales)	Tuesday, January 24, 2017 at 3:21:20 am	<a href="#">AEDT</a>	UTC+11 hours
Corresponding UTC (GMT)	Monday, January 23, 2017 at 16:21:20		

48

---

---

---

---

---

---

---

---



## Other design approaches

Snowflaking (avoid as far as possible)

Galaxies (a way to model multiple interconnected Data Marts)

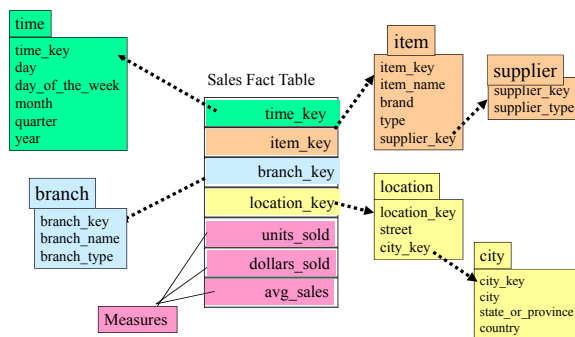
49

## Stars versus Snowflakes

- Snowflaking happens when we choose to **normalize a dimension**
  - e.g. for a so-called “Attribute hierarchies”
- The golden rule: **Avoid** as far as possible
- WHY? (Recall cost of Joins!)

50

## Example of Snowflake Schema



51



## Galaxies or Fact Constellations

- Data Marts "share" dimensions

52

---

---

---

---

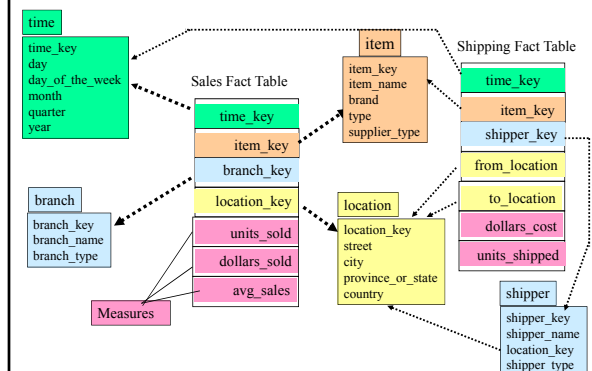
---

---

---

---

## Example of Fact Constellation




---

---

---

---

---

---

---

---

## Summary

- Data marts are designed for decision support
- Data stored over time
- Separate dimension for time/date for easy Analytics (OLAP)
- Dimensional modeling: Star preferred over Snowflake

54

---

---

---

---

---

---

---

---

**Next...**

Physical Database Design

55

---

---

---

---

---

---

---