CSI4142: Data Science

Topic 1:
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)

Overview of topic

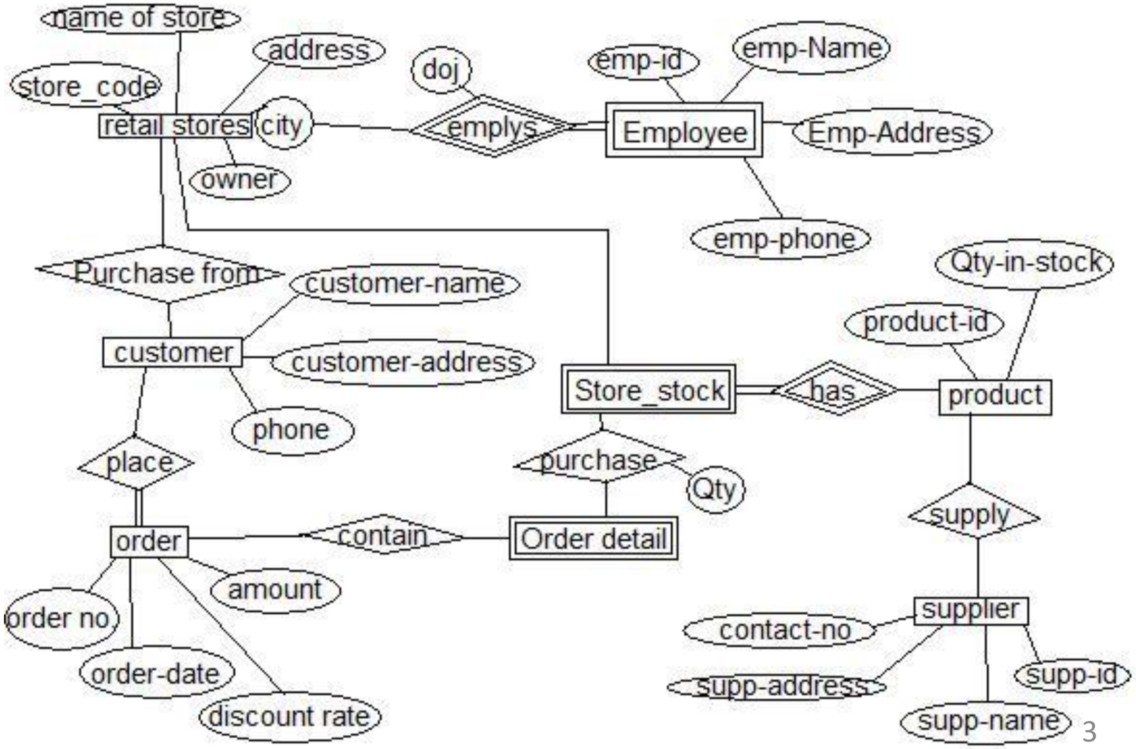
1. Supporting decisions:

   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

Recall from CSI2132: Online Transaction Processing (OLTP)

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



Online Transaction Processing (OLTP)

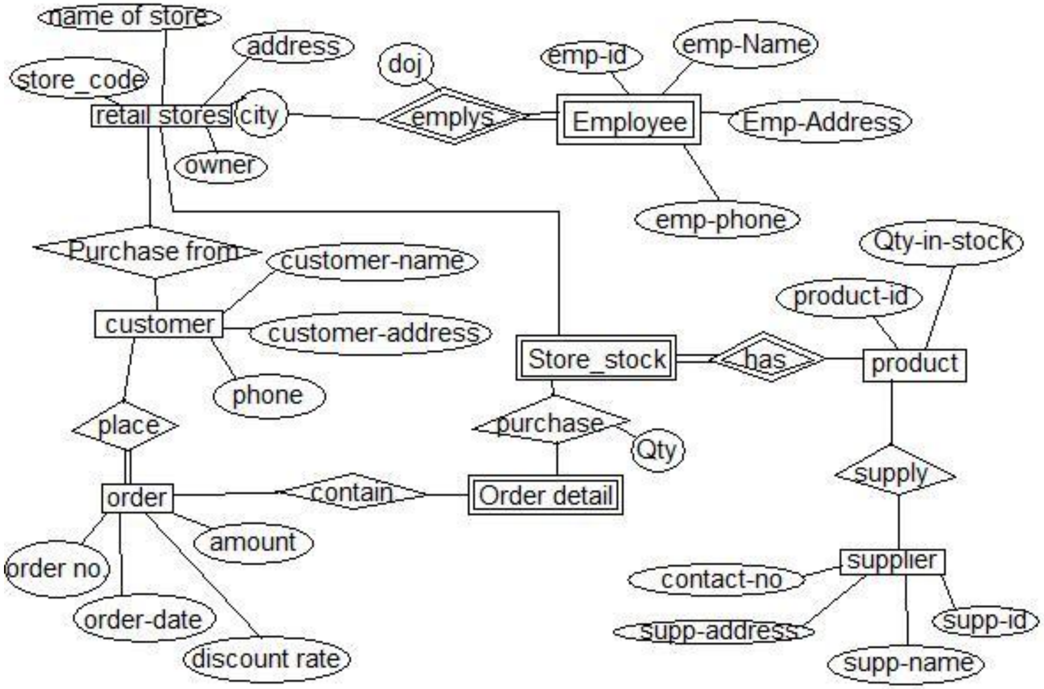
Operations/Transactions:

- INSERT
- DELETE
- UPDATE
- QUERY

DBMS:

- Concurrency control
- Recovery
- Security

- Etc.

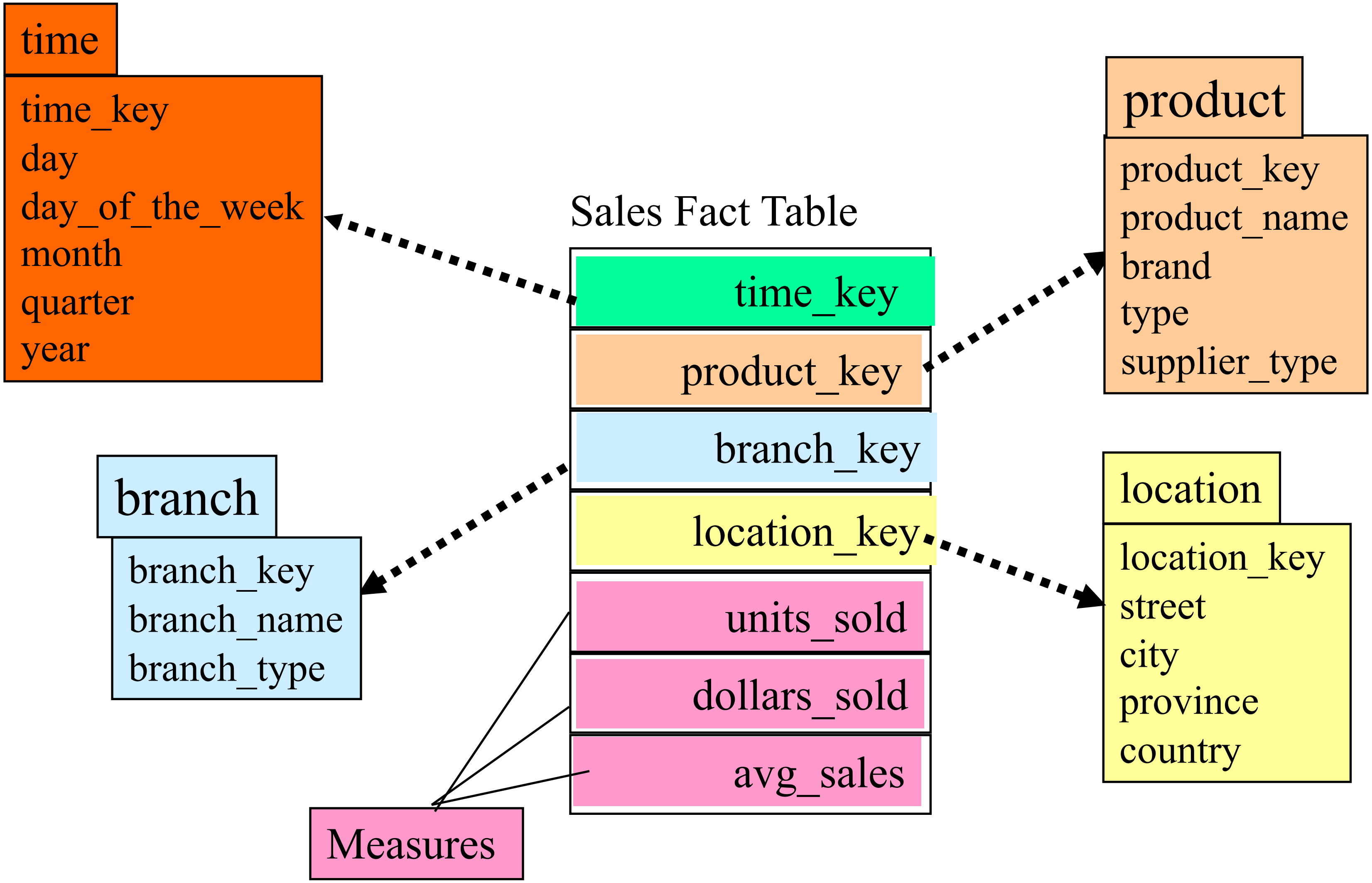


Case Study: Clothing Store with 10 Branches in Ontario (OLTP)

- Open 9h30

- Close 21h00

Online Analytic Processing (OLAP)



time

time_key
day
day_of_the_week
month
quarter
year

Sales Fact Table

time_key
product_key
branch_key
location_key
units_sold
dollars_sold
avg_sales

product

product_key
product_name
brand
type
supplier_type

branch

branch_key
branch_name
branch_type

location

location_key
street
city
province
country

Measures

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

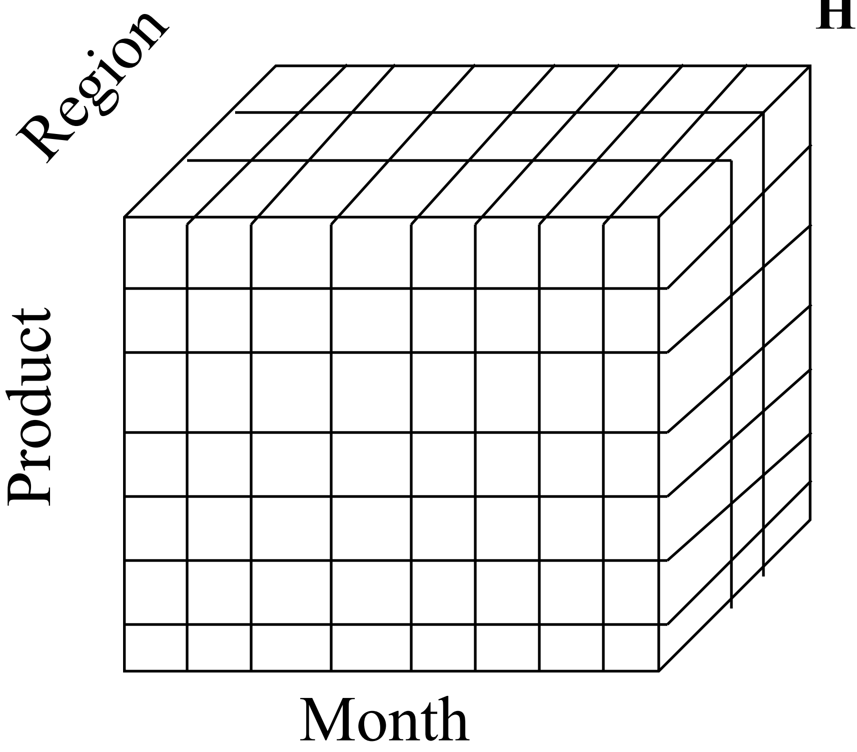
- Open 9h30

…..

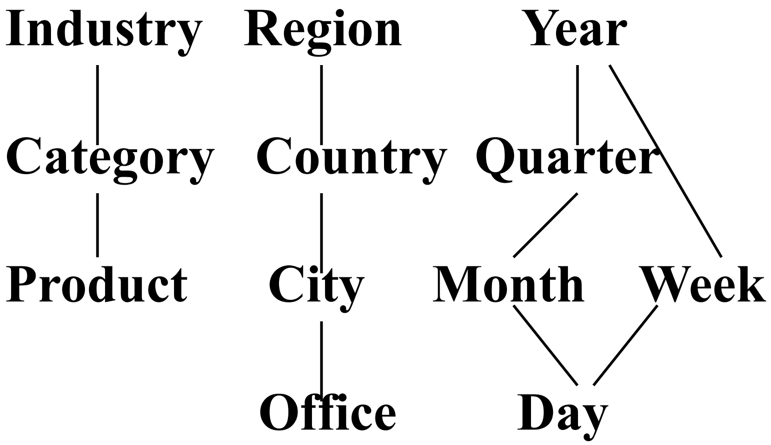
- Close 21h00
- Data staging at 23h30

Multidimensional Data

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



Dimensions: Product, Location, Time
Hierarchical summarization paths



Industry Region Year

Category Country Quarter

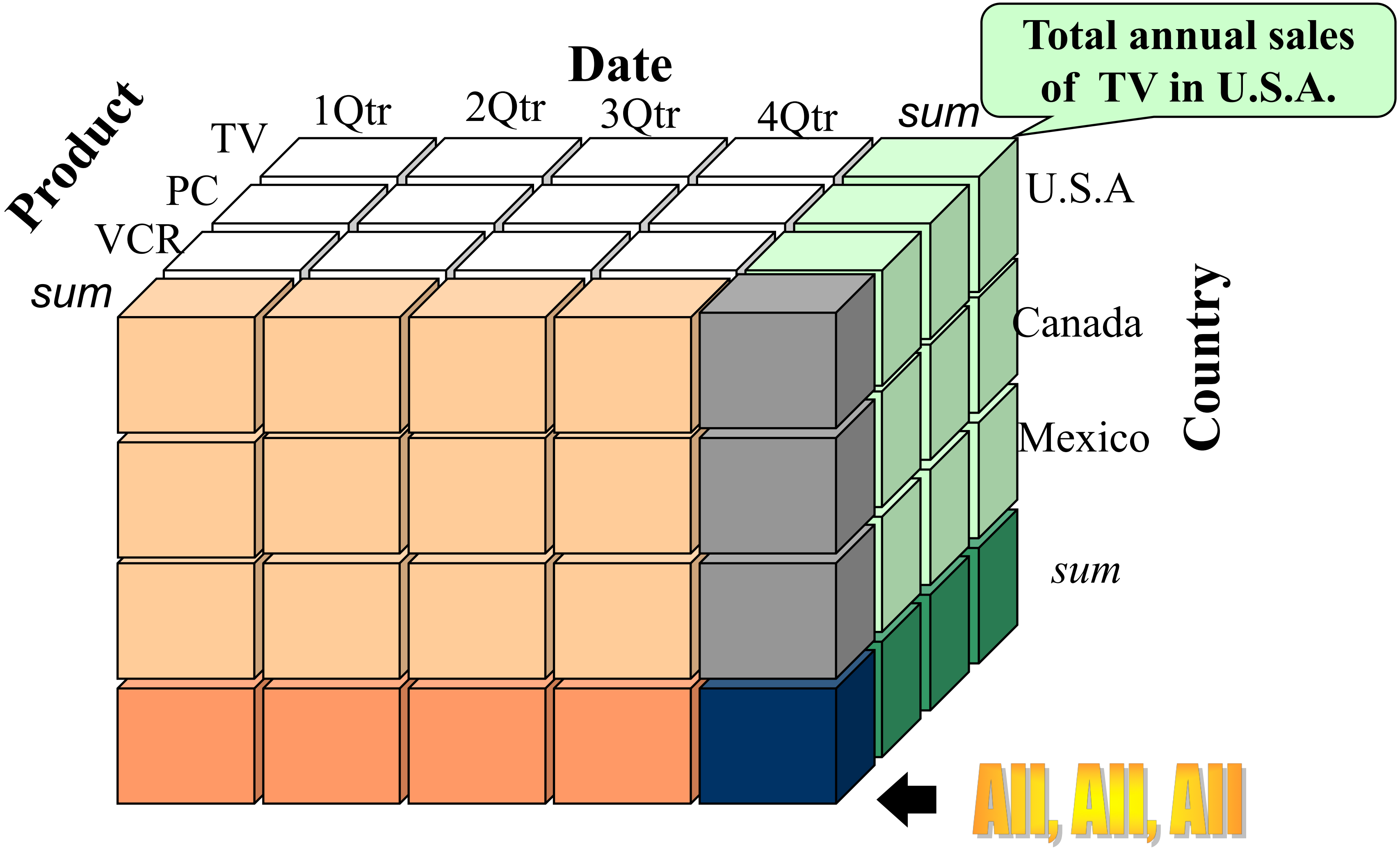
Product City Month Week

Office Day

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

A Sample Data Cube



Total annual sales of TV in U.S.A.

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

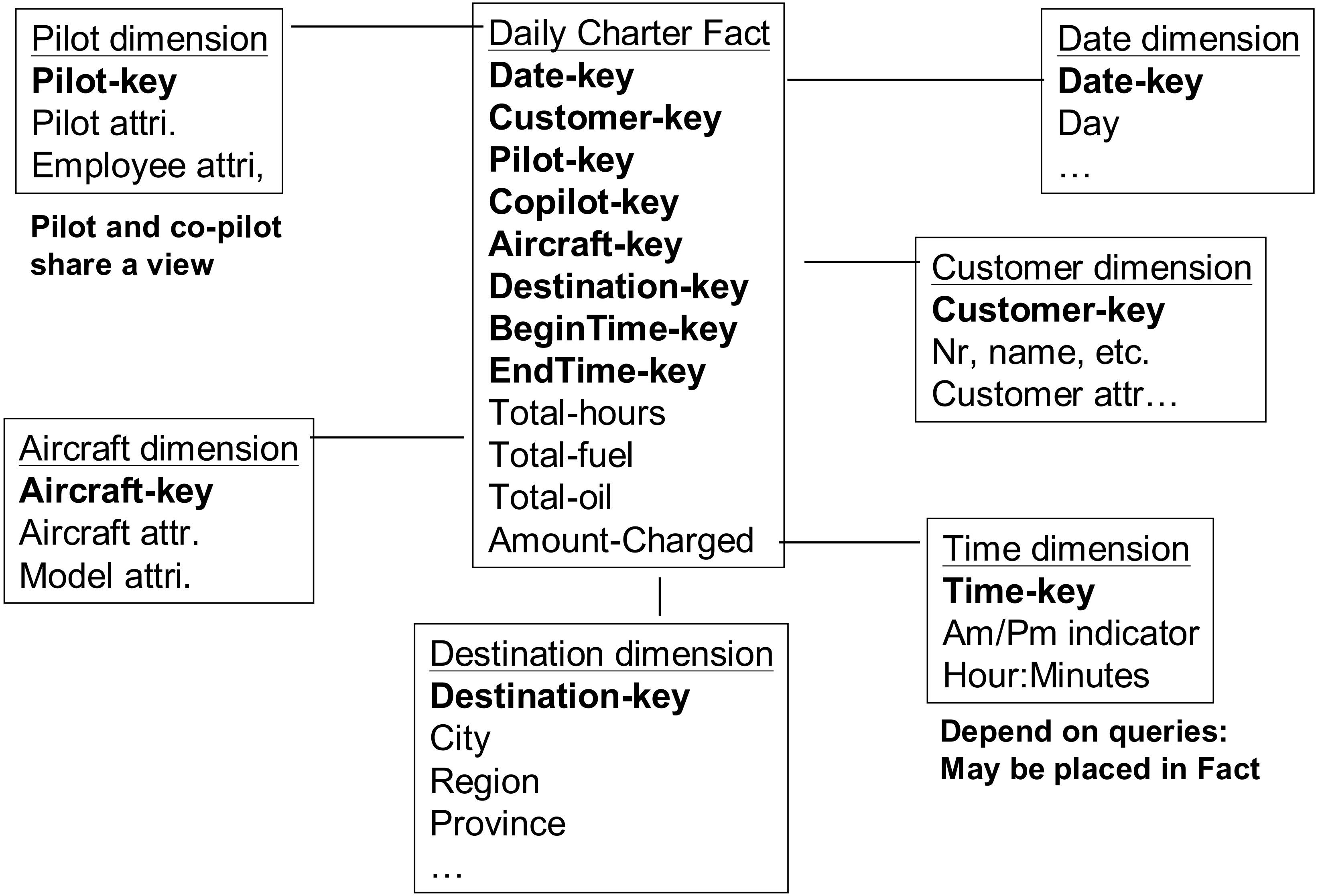
Data Warehouse—Subject-Oriented

- 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

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.

Charter flights



Daily Charter Fact
Date-key
Customer-key
Pilot-key
Copilot-key
Aircraft-key
Destination-key
BeginTime-key
EndTime-key
Total-hours
Total-fuel
Total-oil
Amount-Charged

Pilot dimension
Pilot-key
Pilot attri.
Employee attri,

Pilot and co-pilot share a view

Aircraft dimension
Aircraft-key
Aircraft attr.
Model attri.

Destination dimension
Destination-key
City
Region
Province
…

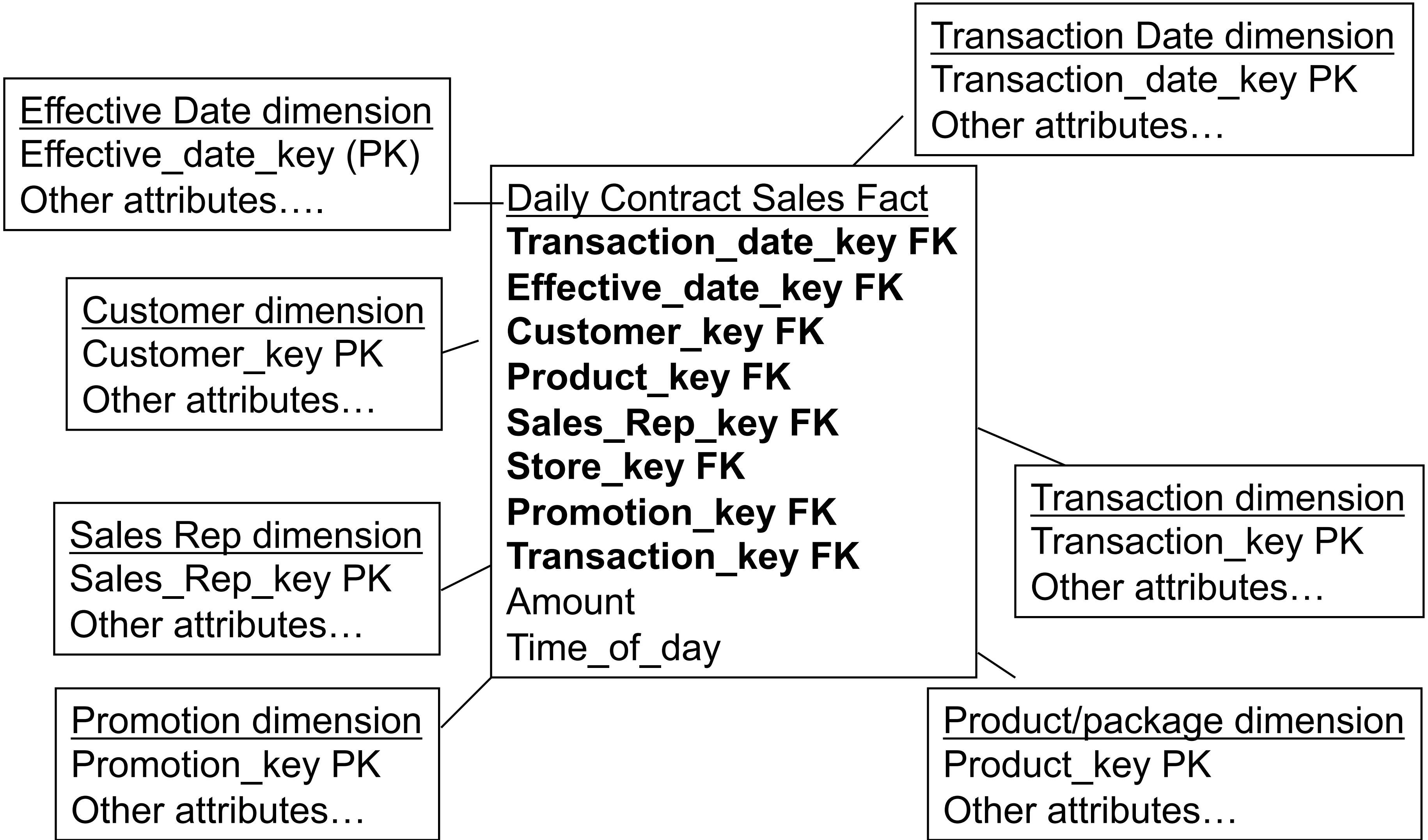
Date dimension
Date-key
Day
…

Customer dimension
Customer-key
Nr, name, etc.
Customer attr…

Time dimension
Time-key
Am/Pm indicator
Hour:Minutes

Depend on queries:
May be placed in Fact

Mobile phone contract sales



Transaction Date dimension
Transaction_date_key PK
Other attributes…

Effective Date dimension
Effective_date_key (PK)
Other attributes….

Daily Contract Sales Fact
Transaction_date_key FK
Effective_date_key FK
Customer_key FK
Product_key FK
Sales_Rep_key FK
Store_key FK
Promotion_key FK
Transaction_key FK
Amount
Time_of_day

Customer dimension
Customer_key PK
Other attributes…

Sales Rep dimension
Sales_Rep_key PK
Other attributes…

Transaction dimension
Transaction_key PK
Other attributes…

Promotion dimension
Promotion_key PK
Other attributes…

Product/package dimension
Product_key PK
Other attributes…

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.

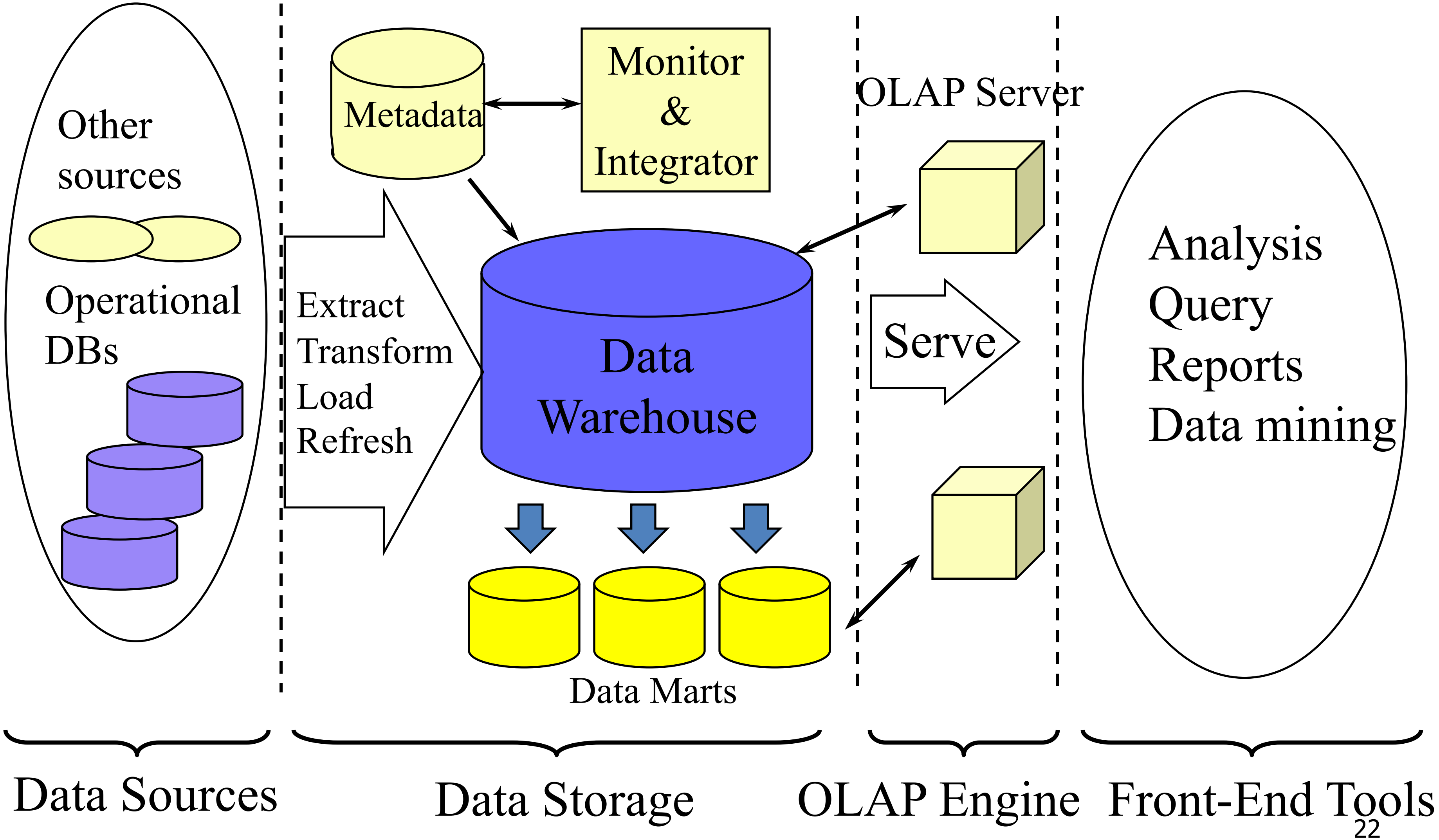
Data Warehouse—Time Variant

- 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”

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*

Data Warehouse: A Multi-Tiered Architecture



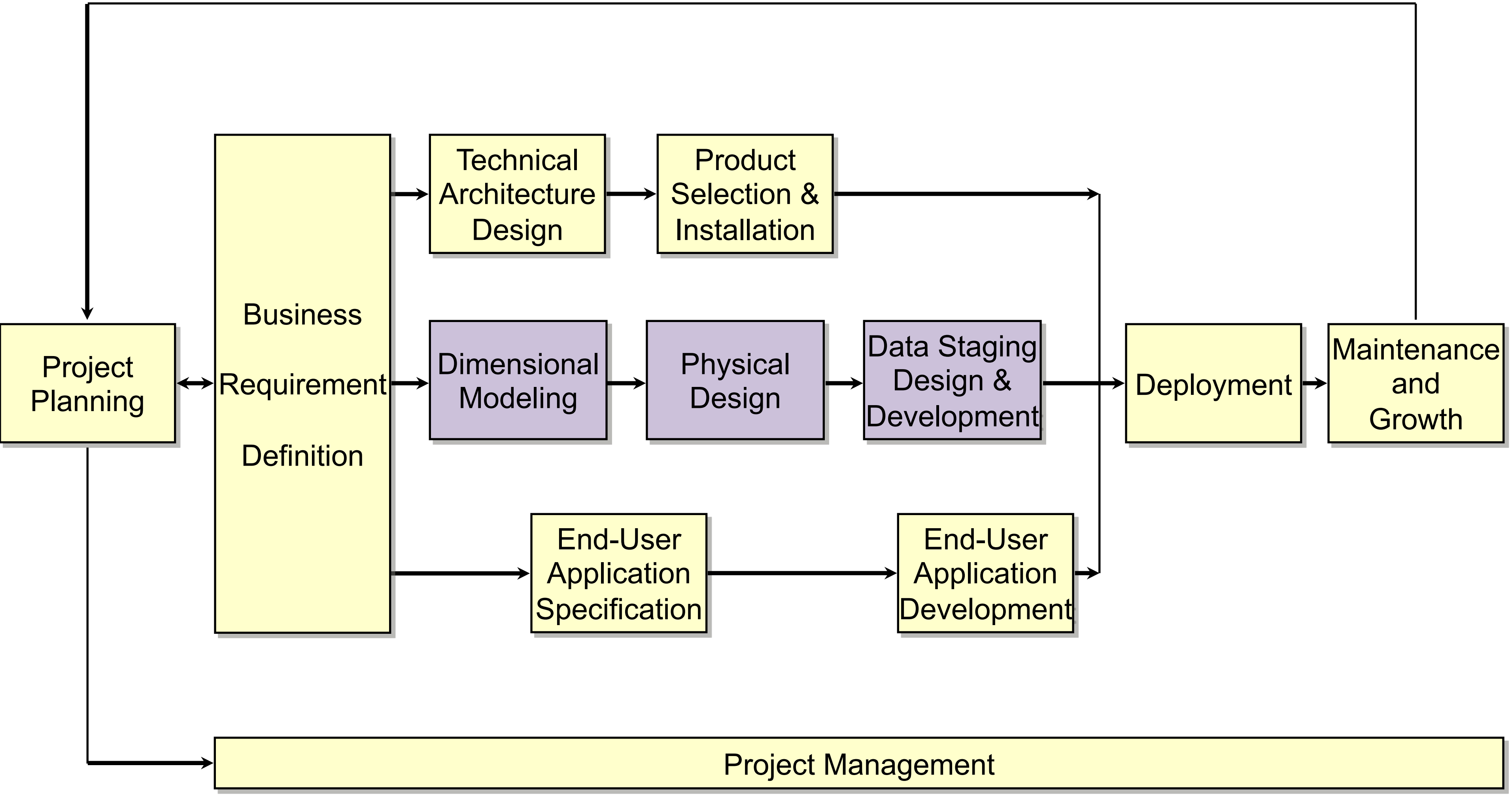
Building a Data Warehouse

Business Life Cycle Toolkit

Kimball et. al.

(http://decisionworks.com/)

Business Dimensional Lifecycle: Kimball et. al.


Project Planning
Business Requirement Definition
Technical Architecture Design
Product Selection & Installation
Dimensional Modeling
Physical Design
Data Staging Design & Development
End-User Application Specification
End-User Application Development
Deployment
Maintenance and Growth
Project Management

Data Track:

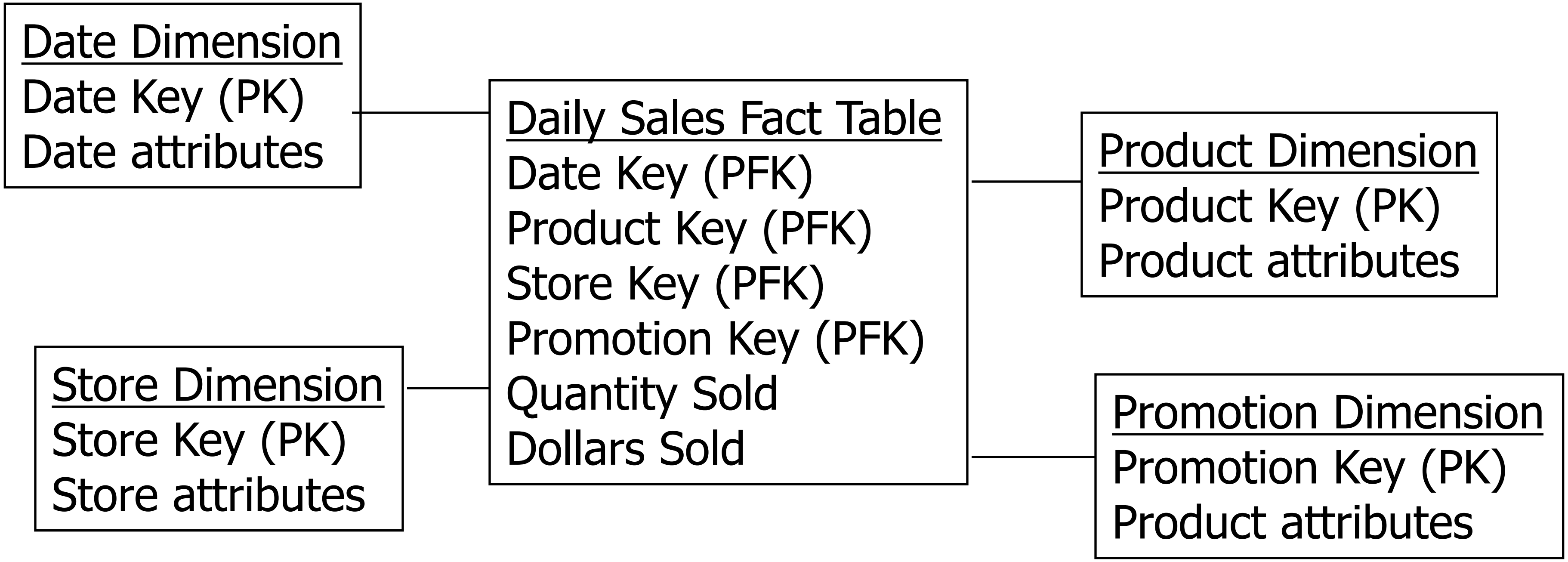
Steps to create a single data mart

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

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

Dimensional Model Components: The classic example



Date Dimension
Date Key (PK)
Date attributes

Daily Sales Fact Table
Date Key (PFK)
Product Key (PFK)
Store Key (PFK)
Promotion Key (PFK)
Quantity Sold
Dollars Sold

Product Dimension
Product Key (PK)
Product attributes

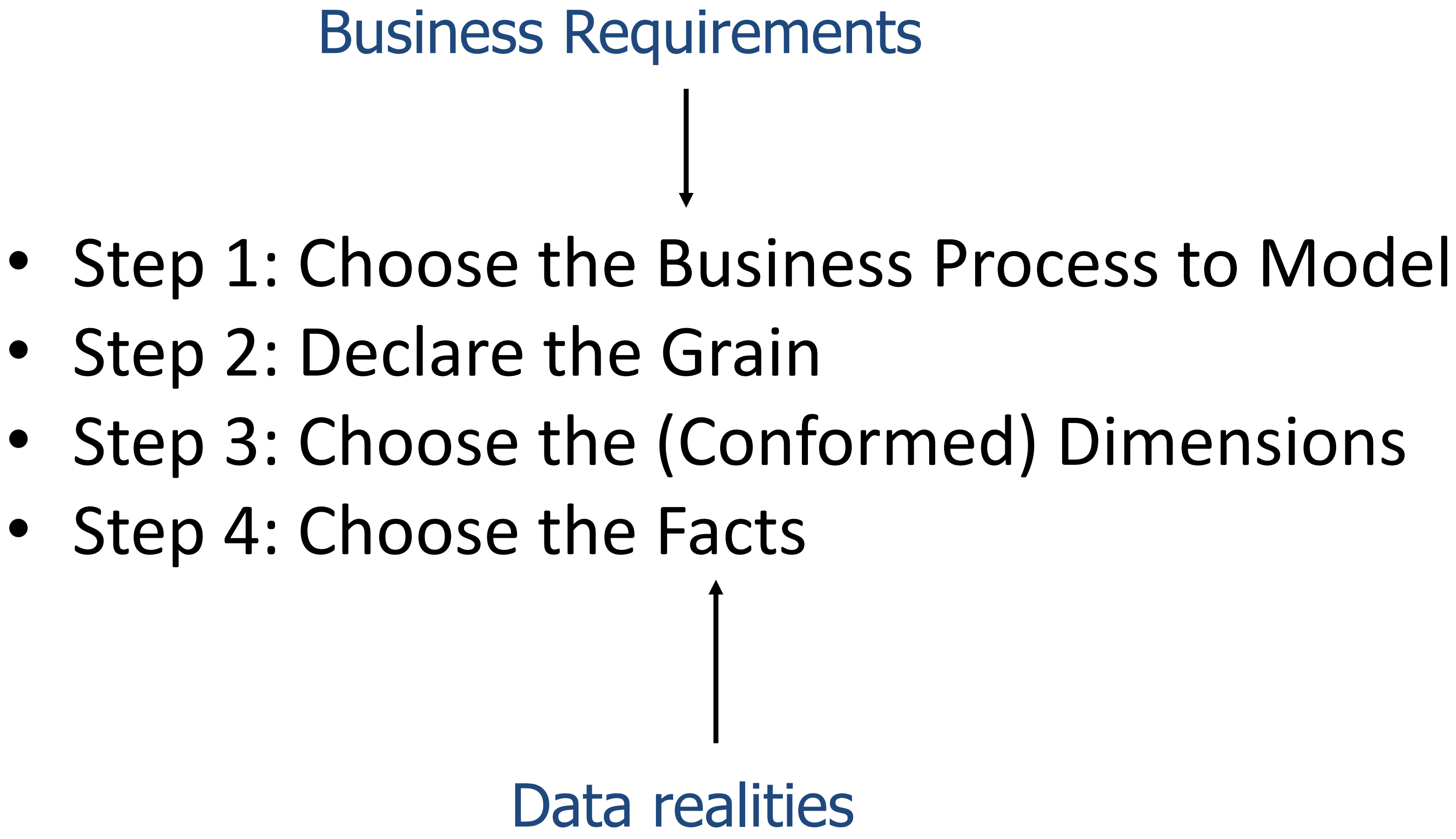
Store Dimension
Store Key (PK)
Store attributes

Promotion Dimension
Promotion Key (PK)
Product attributes

The Star Join Schema

PFK is shorthand for “Primary and Foreign Key”

Four-Step Method to Designing an Individual Dimensional Model



Business Requirements

- Step 1: Choose the Business Process to Model
- Step 2: Declare the Grain
- Step 3: Choose the (Conformed) Dimensions
- Step 4: Choose the Facts

Data realities

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

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

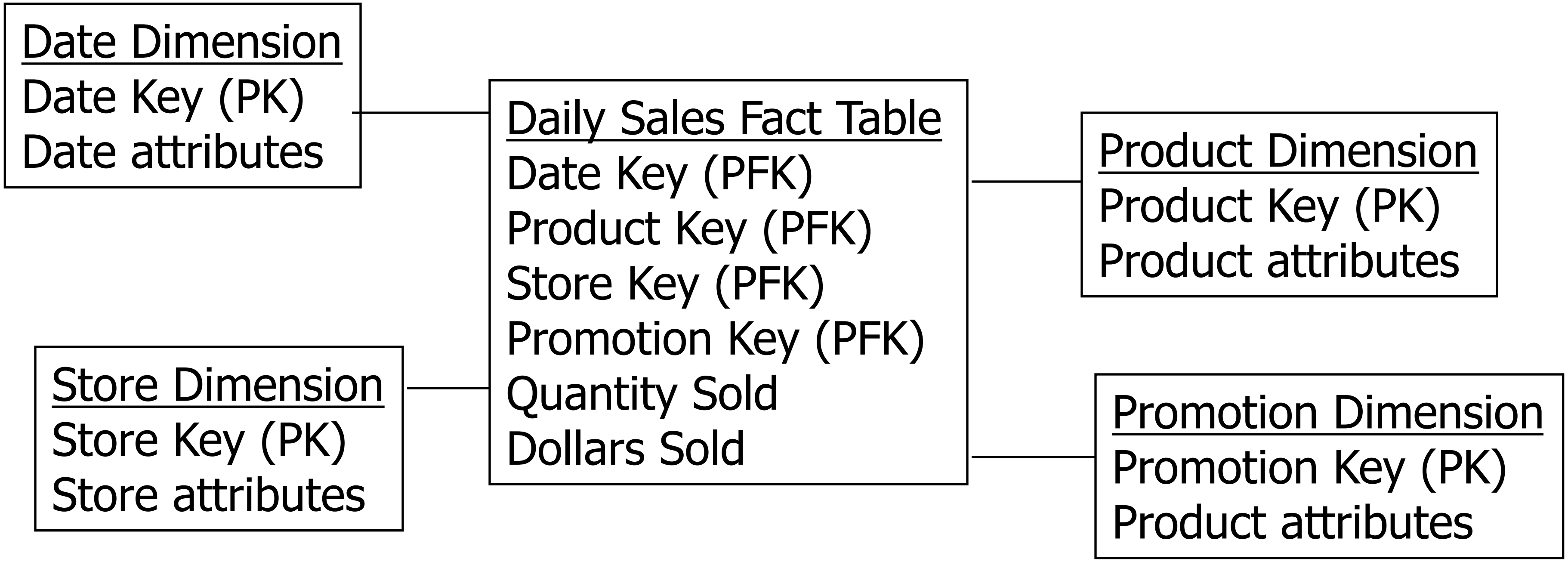
Step 3: Choose the Dimensions

- 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)

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

Dimensional Model Components: The classic example explained



Date Dimension
Date Key (PK)
Date attributes

Daily Sales Fact Table
Date Key (PFK)
Product Key (PFK)
Store Key (PFK)
Promotion Key (PFK)
Quantity Sold
Dollars Sold

Product Dimension
Product Key (PK)
Product attributes

Store Dimension
Store Key (PK)
Store attributes

Promotion Dimension
Promotion Key (PK)
Product attributes

The Star Join Schema

PFK is shorthand for "Primary and Foreign Key"

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.

Sales:
The Product dimension

<u>Product dimension</u>
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

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, …

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

Sales:
Adding the facts/measures

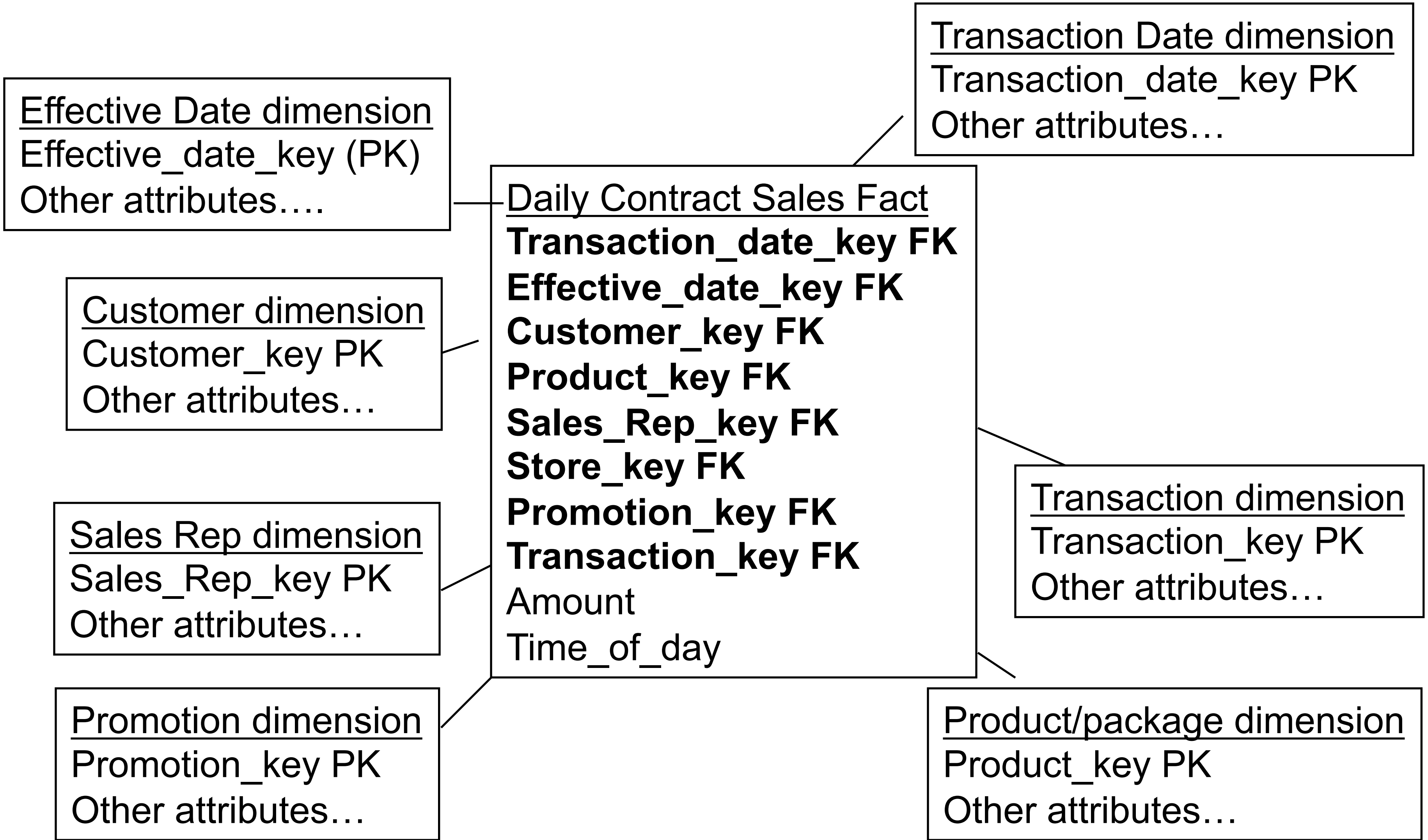
Daily Sales Fact Table
Date Key (PFK)
Product Key (PFK)
Store Key (PFK)
Promotion Key (PFK)
Quantity_sold
CA$_revenue
CA$_cost
Customer_count

Dimensions
Date
Product
Store
Promotion

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

-

Mobile phone contract sales



Transaction Date dimension
Transaction_date_key PK
Other attributes…

Effective Date dimension
Effective_date_key (PK)
Other attributes….

Daily Contract Sales Fact
Transaction_date_key FK
Effective_date_key FK
Customer_key FK
Product_key FK
Sales_Rep_key FK
Store_key FK
Promotion_key FK
Transaction_key FK
Amount
Time_of_day

Customer dimension
Customer_key PK
Other attributes…

Sales Rep dimension
Sales_Rep_key PK
Other attributes…

Transaction dimension
Transaction_key PK
Other attributes…

Promotion dimension
Promotion_key PK
Other attributes…

Product/package dimension
Product_key PK
Other attributes…

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

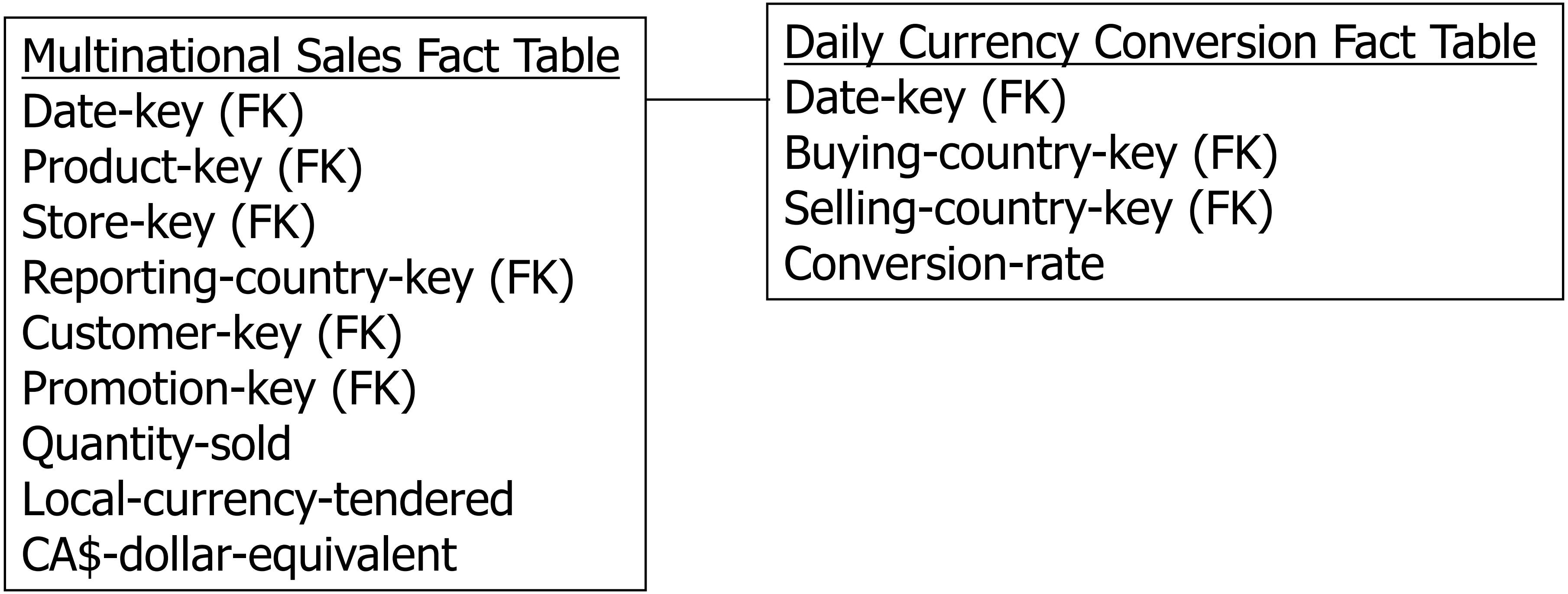
More about dimensions: Role-playing Dimensions

- **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

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

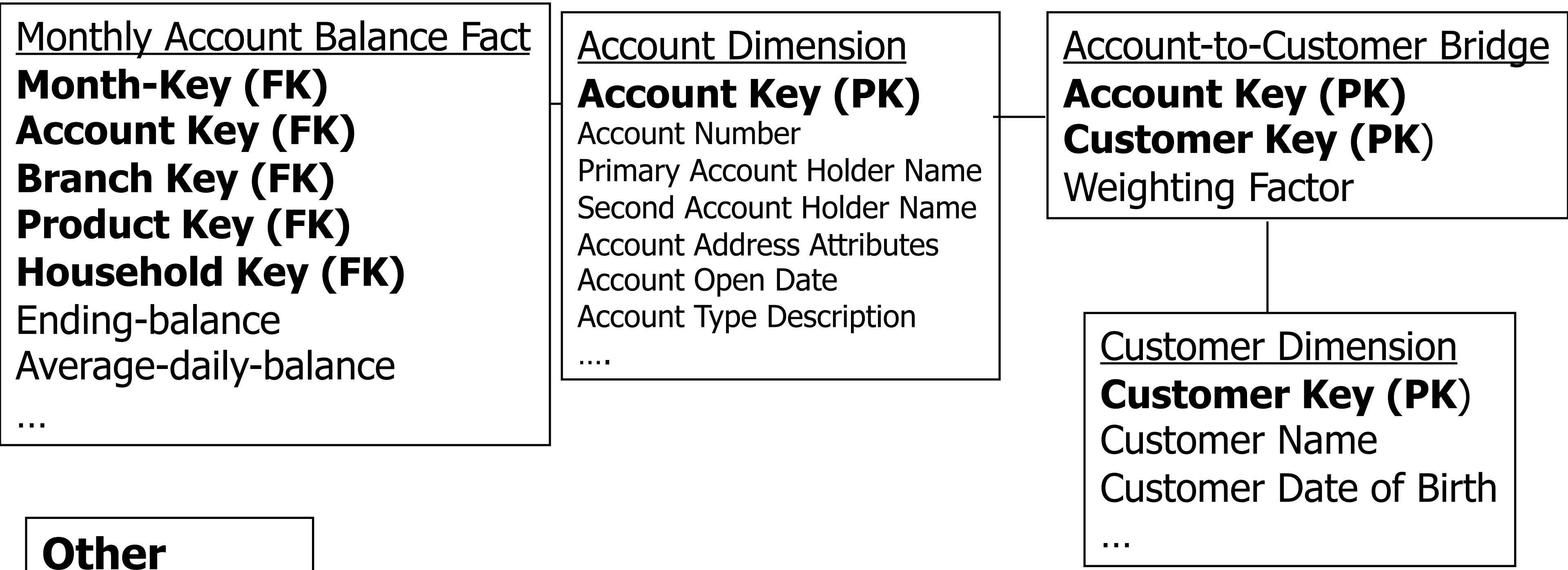
Daily Currency Conversion Fact Table
Date-key (FK)
Buying-country-key (FK)
Selling-country-key (FK)
Conversion-rate

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.

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



Monthly Account Balance Fact
Month-Key (FK)
Account Key (FK)
Branch Key (FK)
Product Key (FK)
Household Key (FK)
Ending-balance
Average-daily-balance
…

Account Dimension
Account Key (PK)
Account Number
Primary Account Holder Name
Second Account Holder Name
Account Address Attributes
Account Open Date
Account Type Description
….

Account-to-Customer Bridge
Account Key (PK)
Customer Key (PK)
Weighting Factor

Customer Dimension
Customer Key (PK)
Customer Name
Customer Date of Birth
…

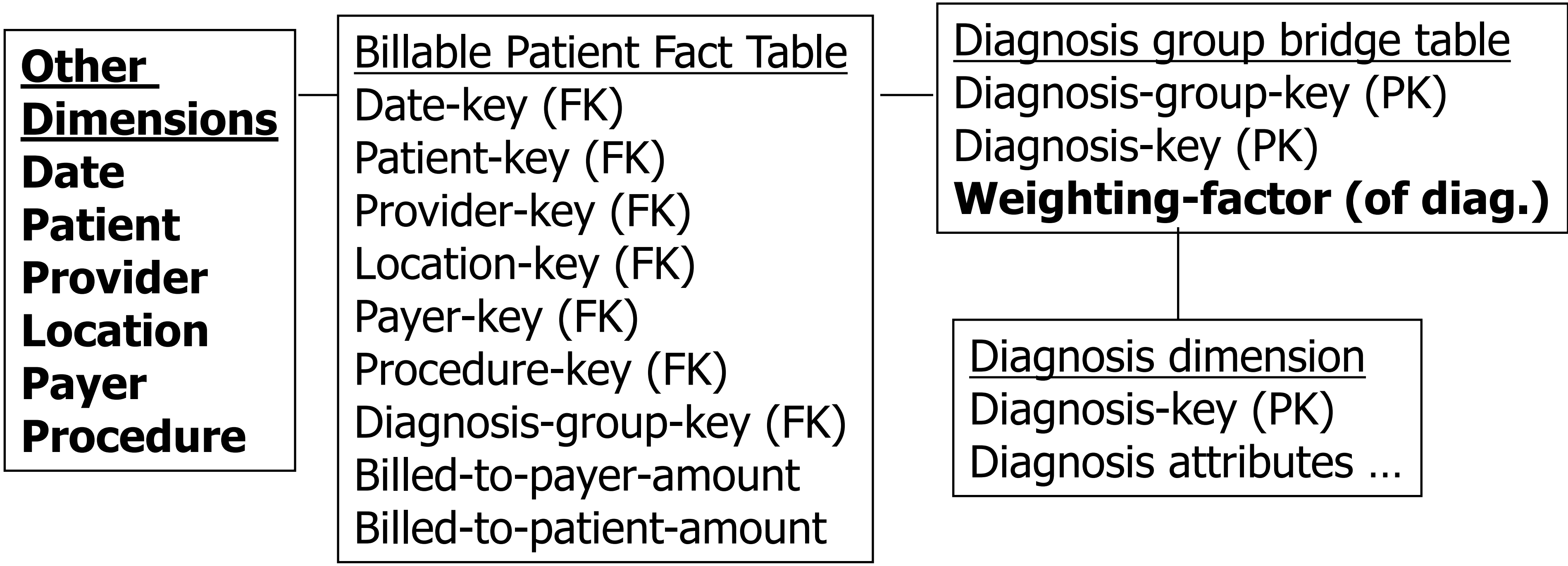
Other
Dimensions
Month
Branch
Product
Household

Here the weighting factor (measure) is the customer's contribution to the account (e.g. income).

Joe: 4,000, 0.4   Sue: 4,000, 0.4   John: 2,000 0.2

More Many to Many Dimensions

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



Other Dimensions
Date
Patient
Provider
Location
Payer
Procedure

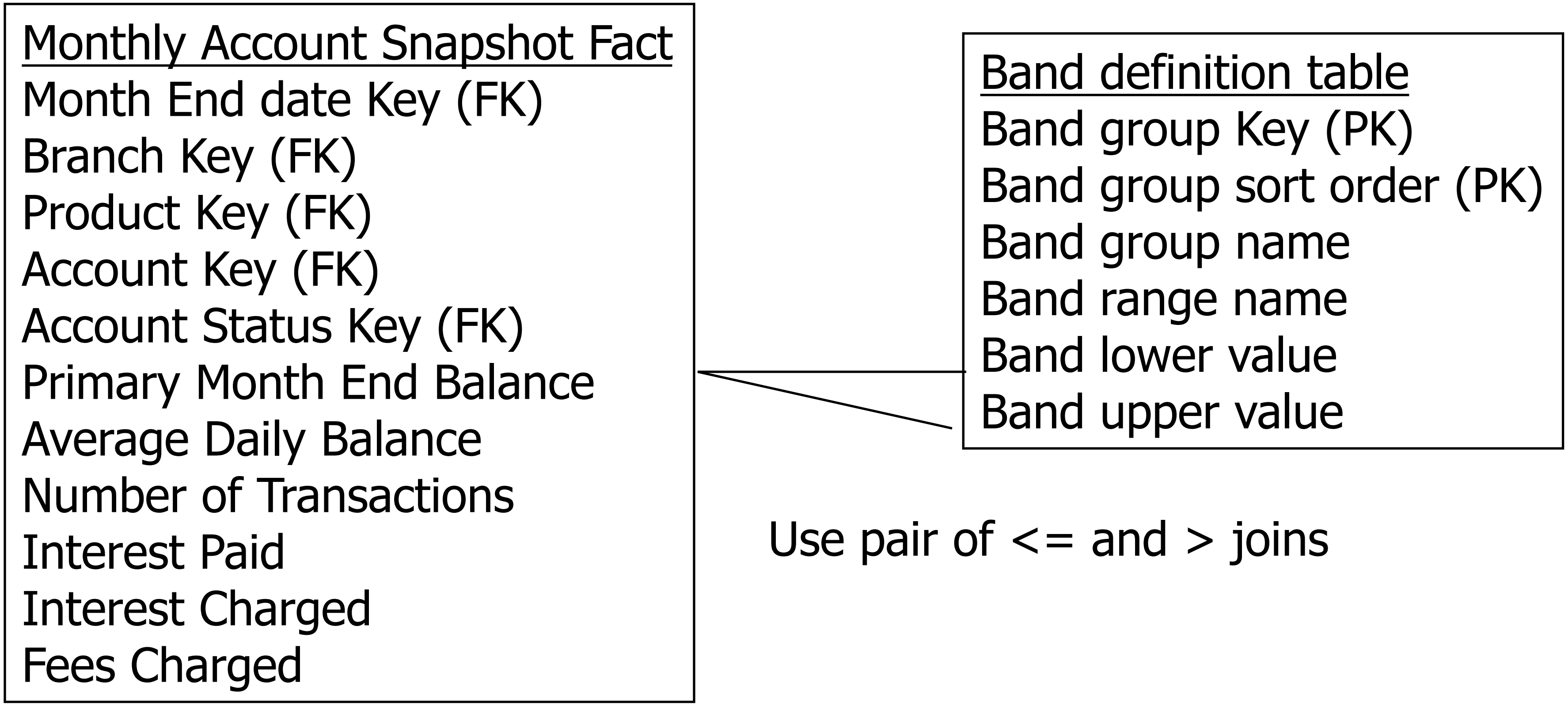
Billable Patient Fact Table
Date-key (FK)
Patient-key (FK)
Provider-key (FK)
Location-key (FK)
Payer-key (FK)
Procedure-key (FK)
Diagnosis-group-key (FK)
Billed-to-payer-amount
Billed-to-patient-amount

Diagnosis group bridge table
Diagnosis-group-key (PK)
Diagnosis-key (PK)
Weighting-factor (of diag.)

Diagnosis dimension
Diagnosis-key (PK)
Diagnosis attributes …

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 <= and > joins

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

<u>Band definition table</u>
Band group Key (PK)
Band group sort order (PK)
Band group name
Band range name
Band lower value
Band upper value

Modeling Time…

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

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

Other design approaches

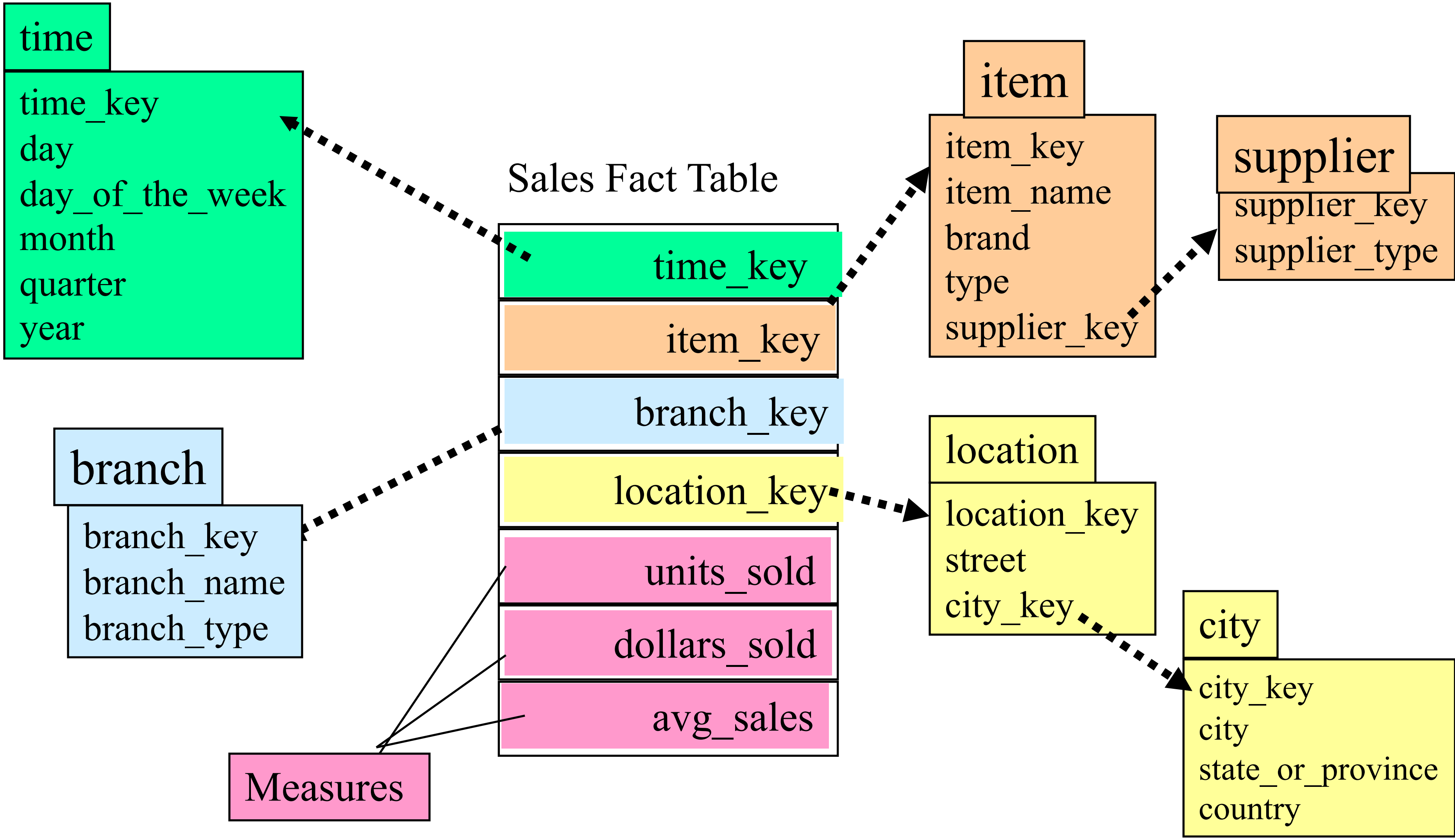
Snowflaking (avoid as far as possible)

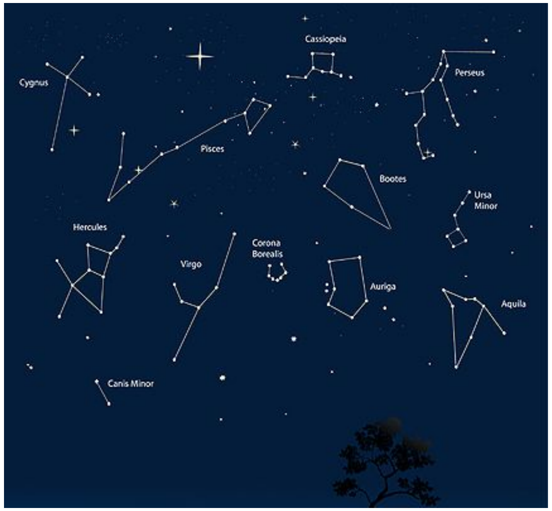
Galaxies (a way to model multiple interconnected Data Marts)

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!)

Example of **Snowflake Schema**

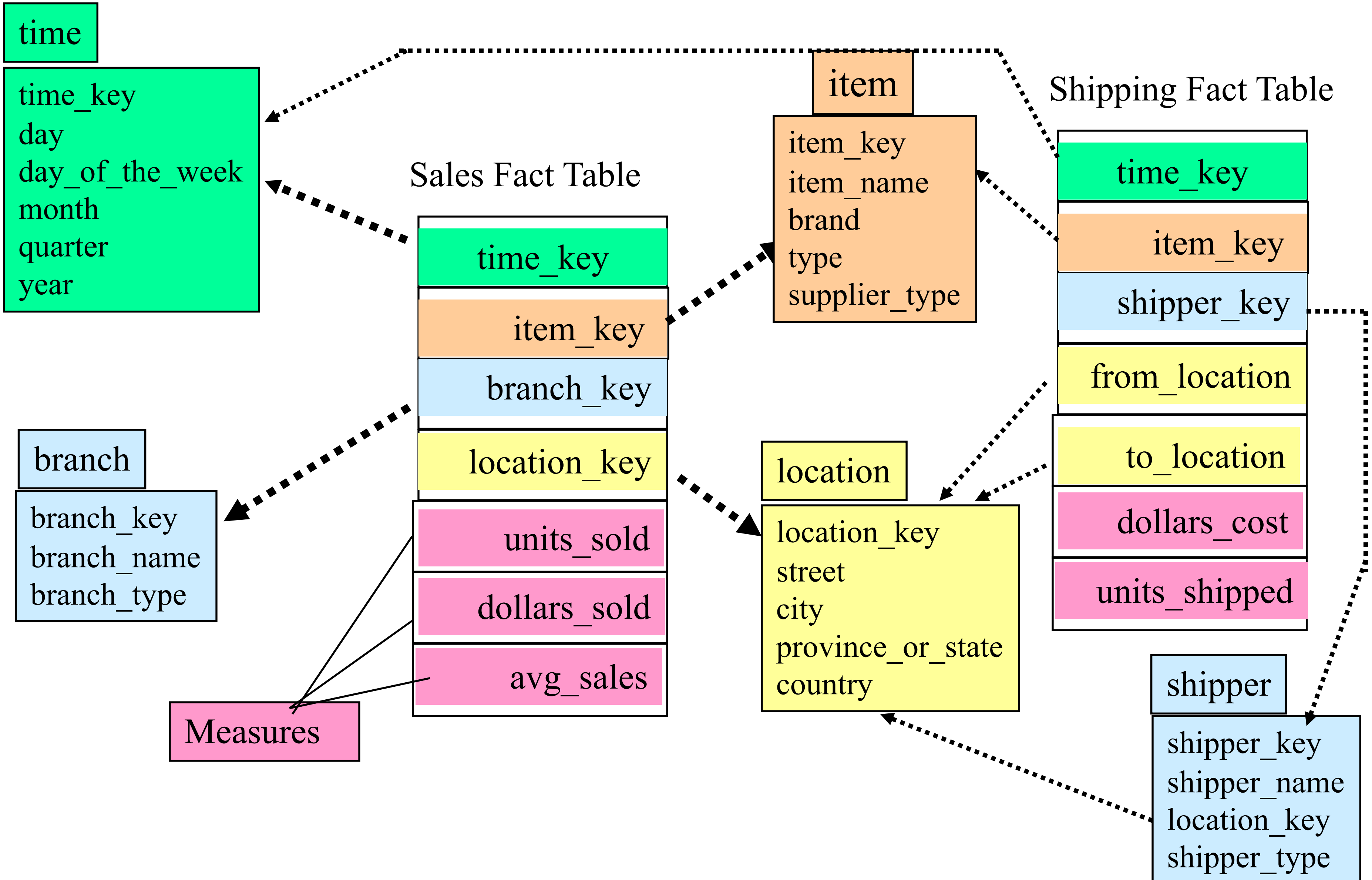




Galaxies or Fact Constellations

- Data Marts “share” dimensions

Example of **Fact Constellation**



time

time_key
day
day_of_the_week
month
quarter
year

Sales Fact Table

time_key
item_key
branch_key
location_key
units_sold
dollars_sold
avg_sales

Measures

branch

branch_key
branch_name
branch_type

item

item_key
item_name
brand
type
supplier_type

location

location_key
street
city
province_or_state
country

Shipping Fact Table

time_key
item_key
shipper_key
from_location
to_location
dollars_cost
units_shipped

shipper

shipper_key
shipper_name
location_key
shipper_type

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

Next…

Physical Database Design