Data Warehousing

Outline

- Data warehouse
    - Operational databases versus data warehouse
    - Characteristics
    - Design
    - Building
- Online analytical processing (OLAP)

OLAP Demo

- Using Excel as a client
  - http://office.microsoft.com/en-us/excel-help/demo-explore-adventure-works-in-excel-by-using-an-olap-pivottable-report-HA010288281.aspx
  - Alternate link: http://blip.tv/file/4515898

Goals

- Goal 1 – support day-to-day operations
  - e.g., Handle order processing
- Goal 2 – provide management with information to make more informed decisions and plan for the future
  - e.g., What were the sales volumes by region and product category for the last year?

- Need different solution for each goal

Categories of Business Systems

- Transaction processing systems (TPS)
  - Application systems used by company employees for everyday operational tasks
    - Online Transaction Processing (OLTP)
    - E.g., sales, manufacturing, customer support
  - Employ production databases

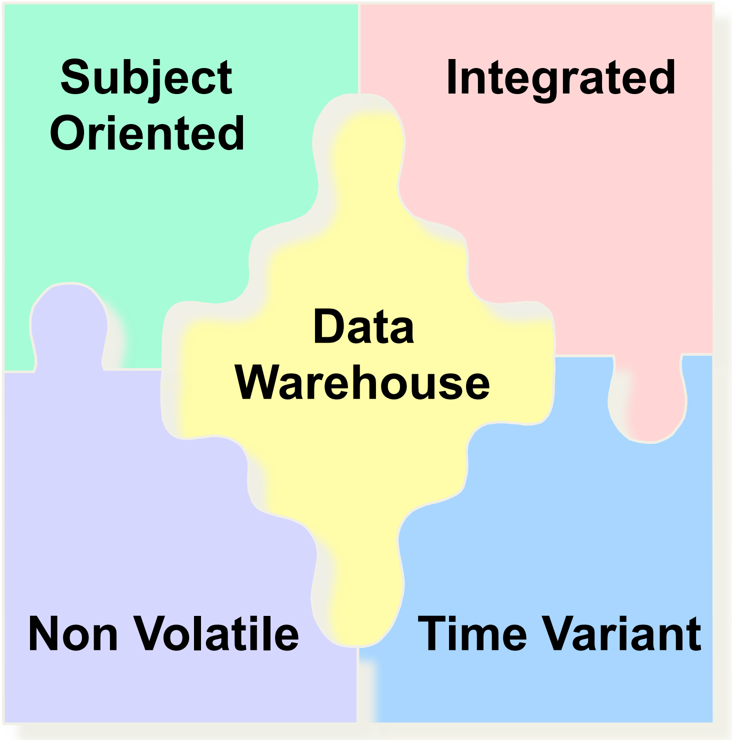
Categories of Business Systems

- Decision support systems (DSS):
  - Systems specifically designed to aid managers in decision-making tasks
    - E.g., budgeting, forecasting, planning
  - Employ data warehouses and/or data marts
  - Require analytical capabilities
    - E.g., data mining, online analytical processing (OLAP)
    - Also called business intelligence (BI) systems

What is a data warehouse?

- “A data warehouse is a *subject oriented* *integrated* *nonvolatile* *time variant* collection of data in support of management’ s decisions.”

Inmon 1992



What is a data warehouse?

- "A data warehouse is a collection of corporate information, derived directly from operational systems and some external data sources. Its specific purpose is to support business decisions, not business operations. This is what a data warehouse is all about, helping your business ask "What if" questions."

Corey and Abbey 1997

Data Warehouse

- Subject-oriented:
  - Data is organized around subject areas
    - What the business wants to talk about
  - e.g., finance, manufacturing, marketing, sales, HR, legal, shipping
- Integrated:
  - Data is collected from several transactional databases
    - Integrated to provide a unified picture of subject over time
  - Data from different databases transformed into common schema, measurement, data type, etc.
- Time variant:
  - Data identified with particular time period
  - Allows analysis of trends
- Non-volatile:
  - Data is stable
  - New data added, but data rarely changed (old data may be removed)

Data Warehouse Example

- Wal-Mart’s RetailLink system:
  - Gives suppliers full access to WM’s sales and inventory data in real-time for collaborative planning, forecasting, and replenishment (CPFR)
  - Powered by NCR’s Teradata servers:
    - Runs 30+ business applications
    - Supports 18,000+ users (WM managers)
    - Handles 120,000 queries/week
    - Receives 8.4 million updates/minute (transactions) at peak-time

Database vs. Data Warehouse

Operational Databases	Data Warehouse
Supports transaction processing systems used in everyday business operations	Supports decision support systems used for managerial decision making
Data stored in relational format	Data stored in multidimensional format
MB/GB in size	Terabytes in size
No specific support for time-series (archive old data)	Supports time-series/periodicity
Good for data input/output (non-aggregate), but poor for accessing large quantities of data (e.g., aggregate)	Poor for data input/output (non-aggregate), but good for accessing large quantities of data (e.g., aggregate)
Exists independently	Aggregated from operational databases
No special analytical operations supported	Supports special analytical operations such ROLLUP and CUBE

Operational Databases

- ❑ Transaction-oriented with frequent updates
- ❑ Processes often operate on small subsets of data
- ❑ Speed matters
- ❑ Designed to control redundancy
- ❑ Administered as a unit
- ❑ High availability required
- ❑ Stable structure and variable contents
- ❑ *Supports day-to-day operations ...*

Inmon 1992

Data Warehouses

- ❑ Analysis-oriented and “read-only”
- ❑ Processes often use large amounts of data
- ❑ Relaxed performance constraints
- ❑ Redundancy is “a fact of life”
- ❑ Separately administered units
- ❑ Relaxed availability requirements
- ❑ Flexible structure
- ❑ *Supports analysis and decision making ...*

Inmon 1992

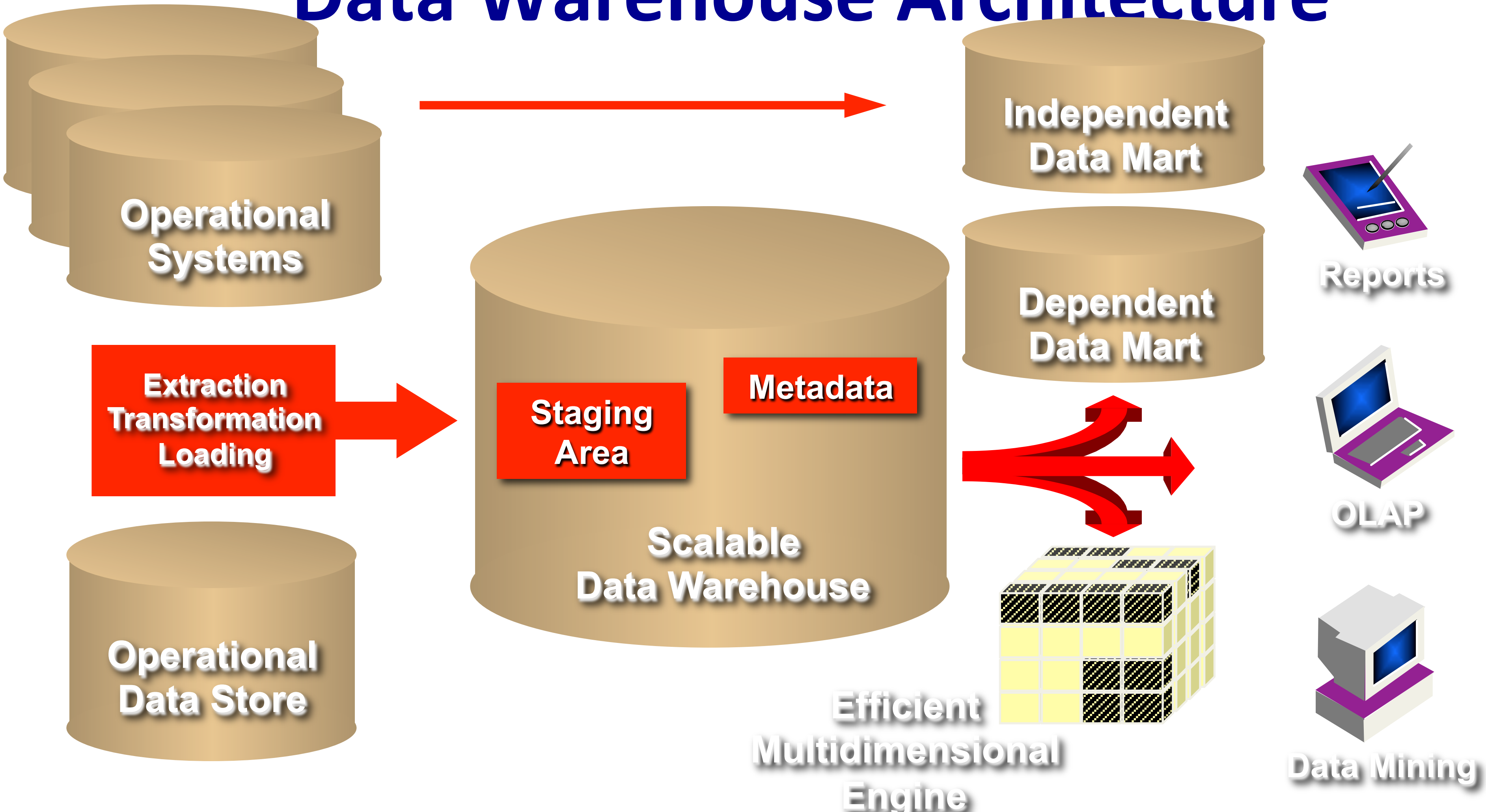
Characteristics of DW Data

- Aggregated:
  - Data may be aggregated by some business dimension (e.g., products, regions, months/years)
  - Depends on how fine-grained data may be needed
    - Transaction level, hourly, daily, etc.
- Historical:
  - Data updated at some time interval: weekly, monthly, etc.
  - Data stored by weeks, months, etc. for historical comparison and trend analysis
- Denormalized:
  - Used to improve query performance (fewer joins)

Data Warehouse vs. Data Marts

- Enterprise Data Warehouse (EDW)/Corporate Data Warehouse (CDW):
  - Large-scale data repository
  - Incorporates aggregated historical data for an entire company, division, or business unit
  - Built around many subjects
  - Supports wide range of decision tasks
- Data marts:
  - Small-scale data repository serving the needs of one department
  - Based on a limited number of subjects (sometimes one)
  - Constructed from few transactional databases or a subset of EDW data

Data Warehouse Architecture



Data Warehouse Components

- Source or Operational System
  - An operational system of record whose function it is to capture the transactions of the business.

- Data Staging Area
  - A storage area and set of processes that clean, transform, combine, de-duplicate, household, archive, and prepare source data for use in the data warehouse.

Data Warehouse Components

- Dimensional Model
  - A specific discipline for modeling data that is an alternative to entity-relationship (E/R) modeling.

- Business Process
  - A coherent set of business activities that make sense to the business users of our data warehouse.

Data Warehouse Components

- Data Mart
  - A logical subset of the complete data warehouse.
  - *Independent or dependent?*

- Data Warehouse
  - The queryable source of data in the enterprise.

Data Warehouse Components

- On-Line Analytic Processing (OLAP)
  - The general activity of querying and presenting text and number data from data warehouses, as well as a specifically dimensional style of querying and presenting that is exemplified by a number of OLAP vendors.

Data Warehouse vs. Data Marts

- Which is done first:
  - Top-down development: EDW/CDW is created first, from which data is extracted to create one or more DMs
  - Bottom-up approach: Build independent DMs as needed, overall EDW built later from existing DMs

Dimensions & Measures

- Dimensions – way to categorize your data
  - How the business wants to talk about the subject
  - E.g., product, location, time, customer, supplier
  - Similar to viewing a 3D model from different angles
    - Sales by products over time
    - Sales by region over time

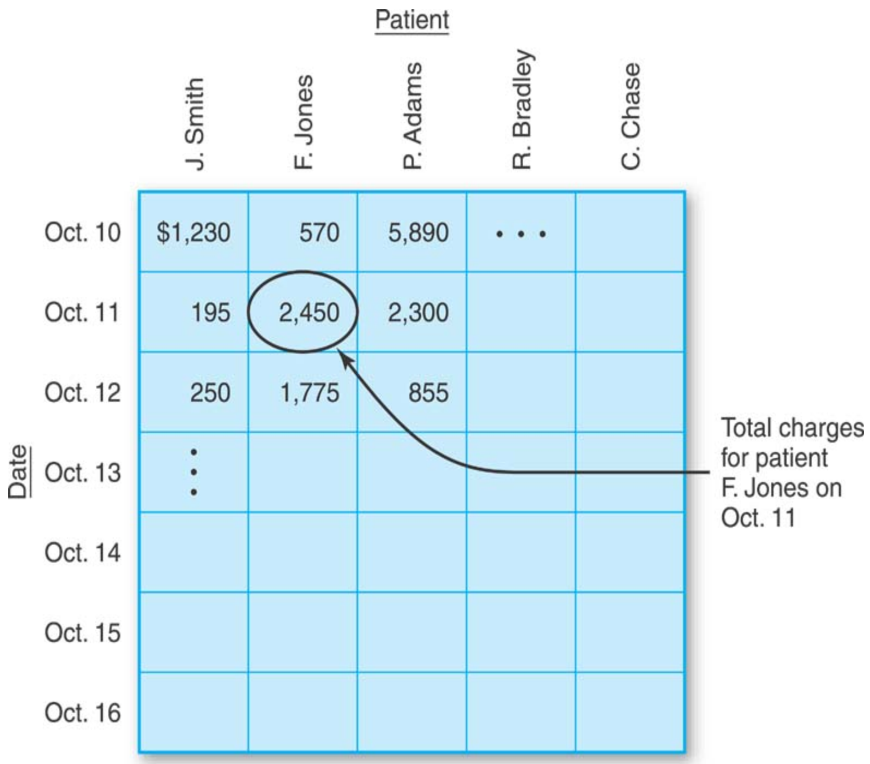
- Measures – provide meaning to dimensions
  - Quantitative – e.g., Revenue, cost, quantity, etc.

Dimensions & Measures

	2008	2009	2010
Product A	$53,493	$41,402	$23,093
Product B	$32,439	$43,202	$56,492
Product C	$85,231	$99,403	$123,403

	2008	2009	2010
Region A	$1,212,329	$1,513,202	$1,245,992
Region B	$998,322	$882,223	$748,232
Region C	$459,349	$758,983	$1,938,231

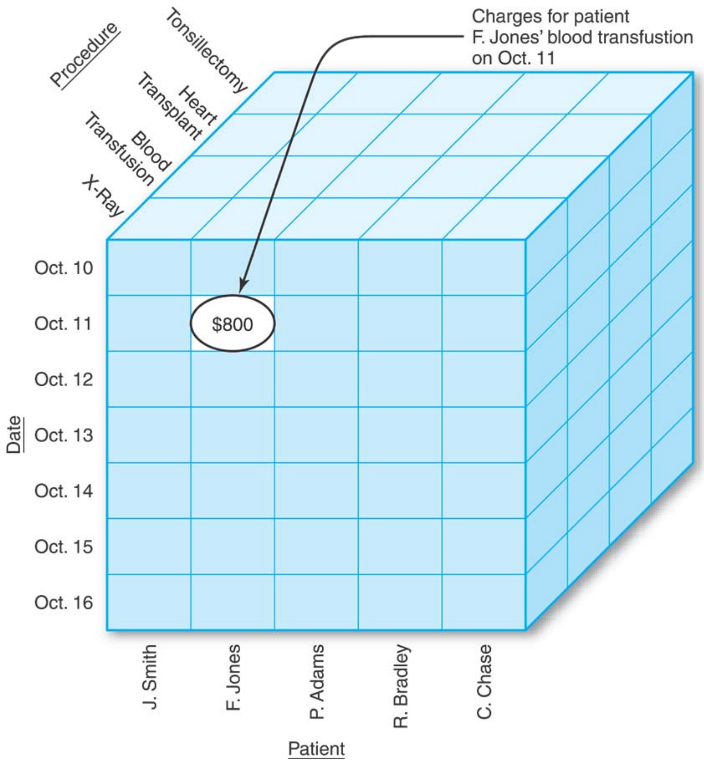
Dimensions of a Data Warehouse



Patient

Date	J. Smith	F. Jones	P. Adams	R. Bradley	C. Chase
Oct. 10	$1,230	570	5,890	• • •	
Oct. 11	195	2,450	2,300		
Oct. 12	250	1,775	855		
Oct. 13	⋮				
Oct. 14					
Oct. 15					
Oct. 16					

Total charges for patient F. Jones on Oct. 11



Charges for patient F. Jones' blood transfustion on Oct. 11

Two-dimensional data warehouse

Three-dimensional data warehouse

Data warehouses can have four or more dimensions

Hierarchies

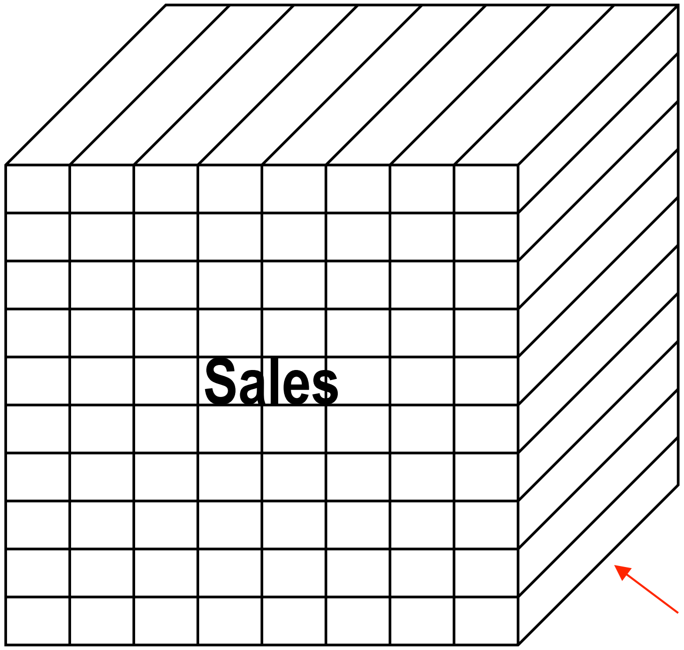
- Organize business dimensions into hierarchies
  - Sales area by country, region, state, county, city
  - Time grouped by year, quarter, month, day

- Drill-down analysis – extracting data from higher to lower hierarchy
- Slice and dice – extracting data from two hierarchies

Hierarchies

Sales area (hierarchy):

- Region: Northeast
    - State: NY
        - Area: NYC
        - Area: Albany
        - Area: Buffalo
        - Area: Long Island
    - State: NJ
    - State: PA
- Region: Midwest
- Region: West



Time (hierarchy):

- Year: 1995
    - Quarter: Q1
        - Month: January
            - Day: 01
            - Day: 02
        - Month: February
    - Quarter: Q2
- Year: 1996
- Year: 1997

Products (hierarchy):

- By product lines
- By responsibility centers
- By work centers

Drill-down: Overall sales figures for NY vs. sales figures for NYC, Albany, Buffalo, etc.
Slice and dice: Sales of individual product lines in NYC vs. Albany, vs. Buffalo, etc.

Designing a Data Warehouse

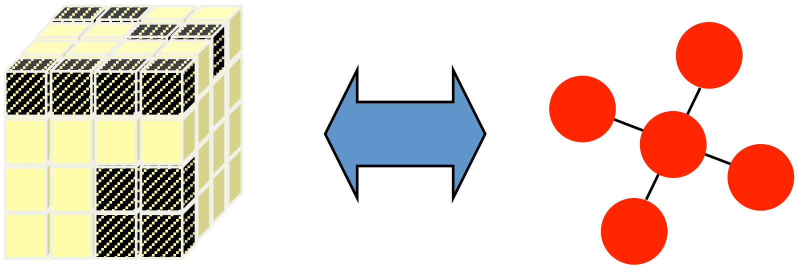
- Data warehouses called multidimensional databases
  - ❑ Often stored in relational database

- Design
  - ❑ Star schema:
    - Two components:
      - Fact table: Sales (subject)
      - Dimension tables: Time Period, Salesperson, Products
  - ❑ Snowflake design:
    - Same as start schema but, dimensions may lead to other dimensions

Star Schemas

- The *dimensional model* or *star join schema* is characterized by a large central fact table with supporting dimension tables arranged in a radial pattern around the fact table.

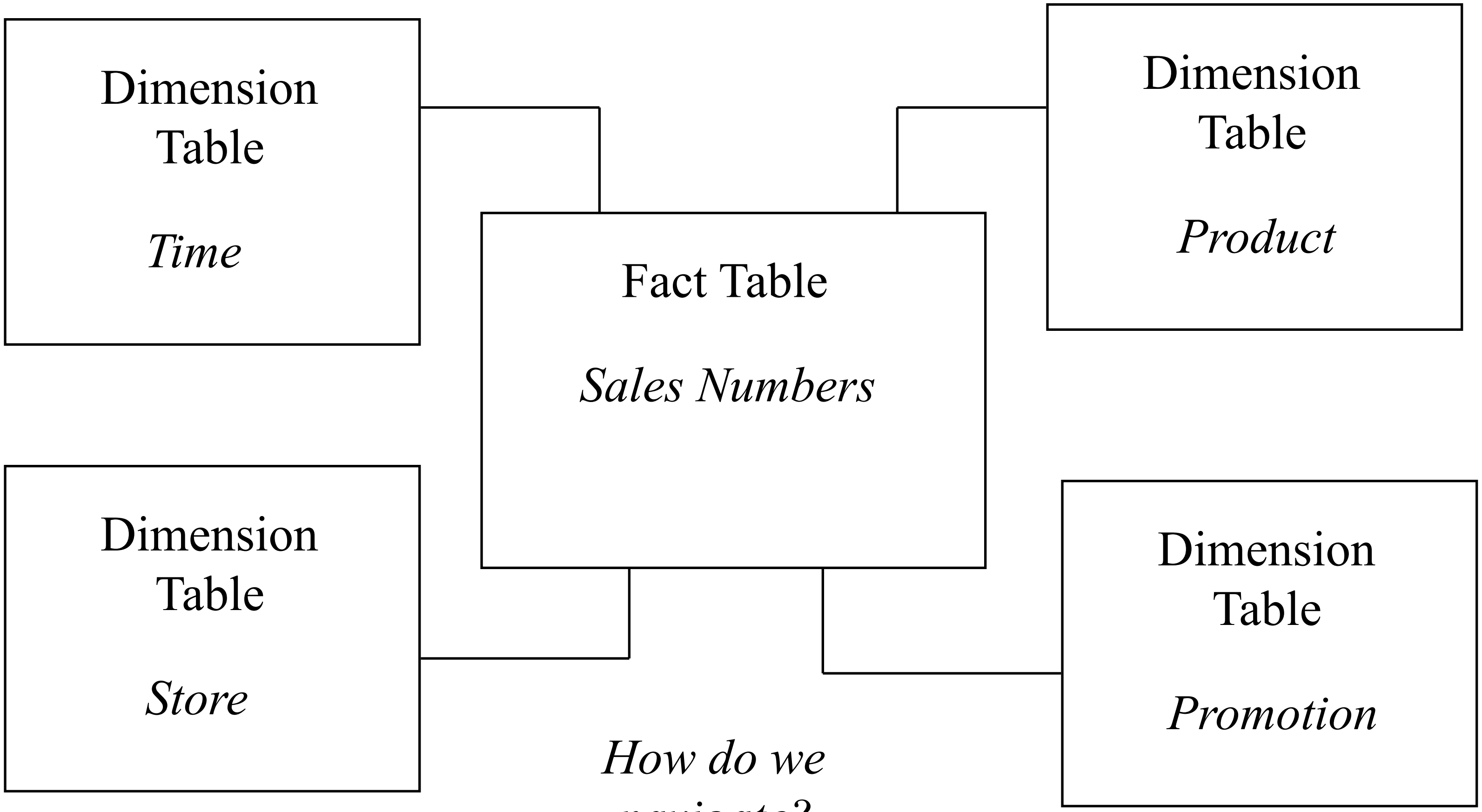
- Fact Tables
- Dimension Tables



Design Considerations in Star Schema

- Fact table:
  - Should contain quantitative time-period data
  - Granularity: what level of detail should you store in fact table?
    - Transactional (finest level) versus aggregated (summarized)
  - Finer grain provides better analysis capability, but requires more rows and hence, slower performance
- Dimension table:
  - Should be denormalized to maximize performance
- Relationship:
  - 1:N relationship between fact and dimension tables

Star Schema



Dimension Table

Time

Dimension Table

Product

Fact Table

Sales Numbers

Dimension Table

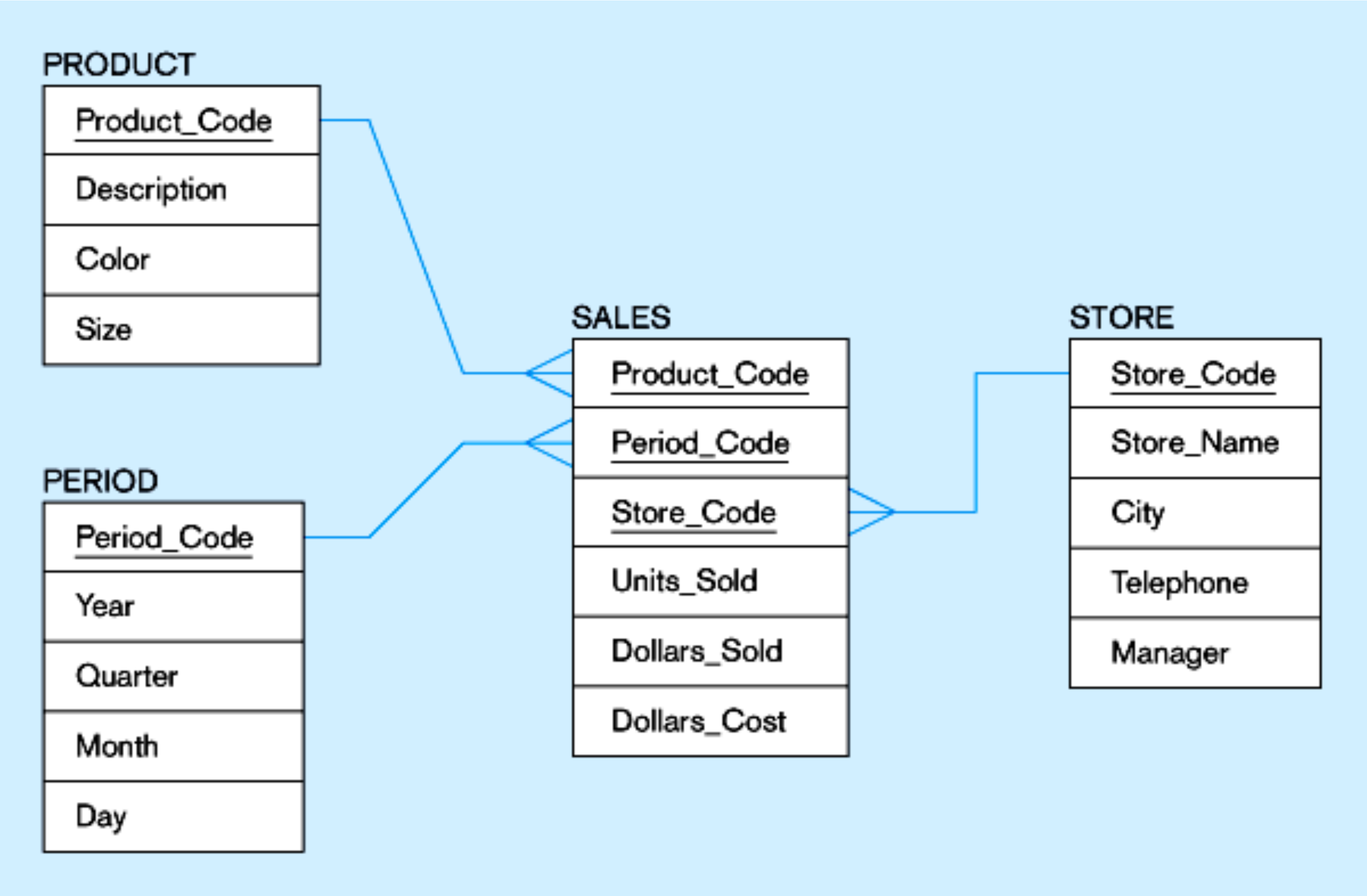
Store

Dimension Table

Promotion

How do we navigate?

Star Schema Example



PRODUCT

Product_Code

Description

Color

Size

PERIOD

Period_Code

Year

Quarter

Month

Day

SALES

Product_Code

Period_Code

Store_Code

Units_Sold

Dollars_Sold

Dollars_Cost

STORE

Store_Code

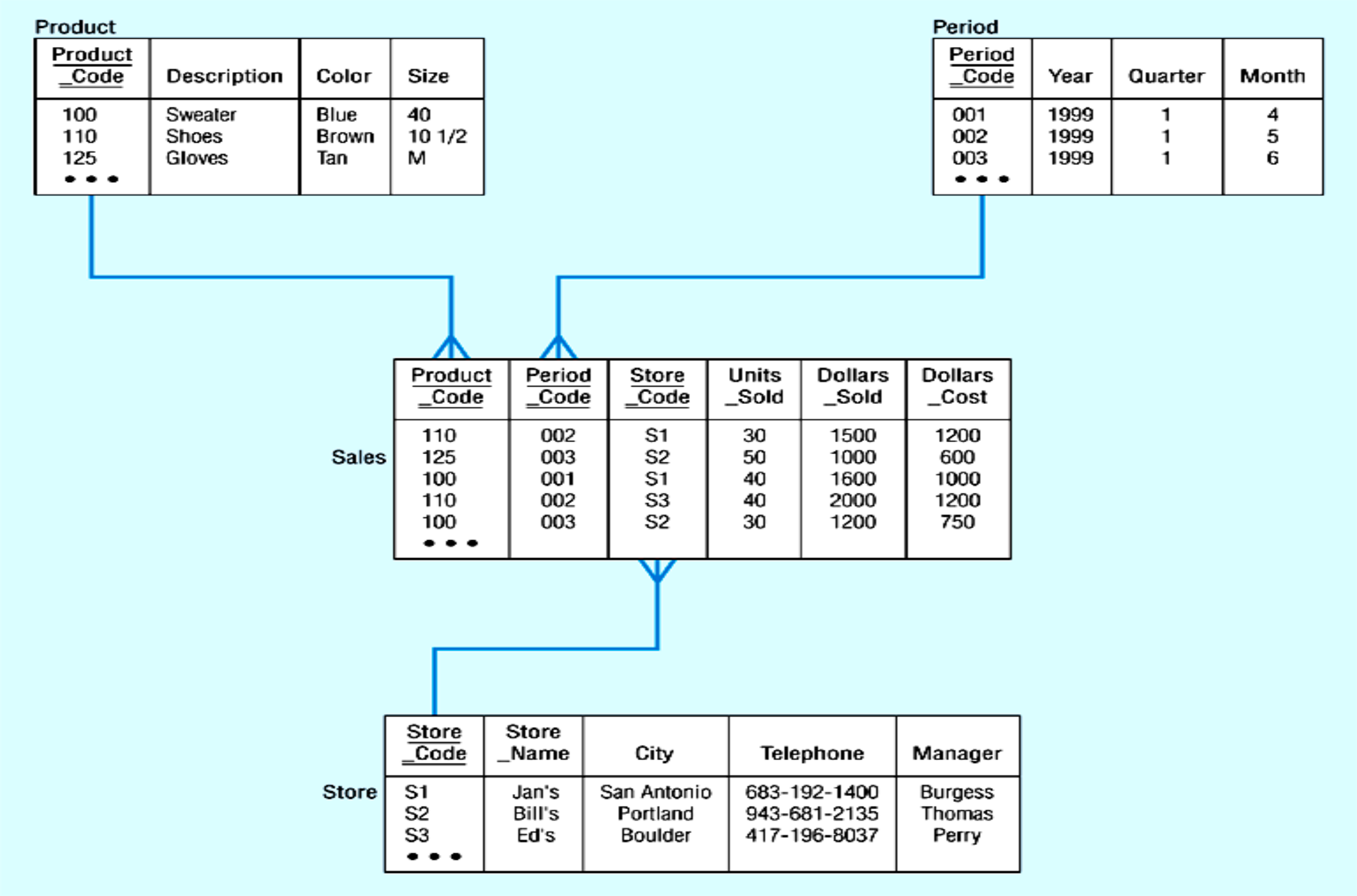
Store_Name

City

Telephone

Manager

Star Schema with Sample Data



Product

Product _Code	Description	Color	Size
100	Sweater	Blue	40
110	Shoes	Brown	10 1/2
125	Gloves	Tan	M
• • •			

Period

Period _Code	Year	Quarter	Month
001	1999	1	4
002	1999	1	5
003	1999	1	6
• • •			

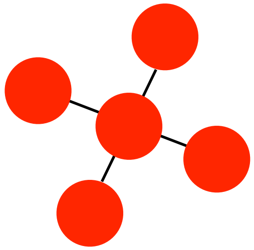
Sales

Product _Code	Period _Code	Store _Code	Units _Sold	Dollars _Sold	Dollars _Cost
110	002	S1	30	1500	1200
125	003	S2	50	1000	600
100	001	S1	40	1600	1000
110	002	S3	40	2000	1200
100	003	S2	30	1200	750
• • •					

Store

Store _Code	Store _Name	City	Telephone	Manager
S1	Jan's	San Antonio	683-192-1400	Burgess
S2	Bill's	Portland	943-681-2135	Thomas
S3	Ed's	Boulder	417-196-8037	Perry
• • •				

Fact Tables



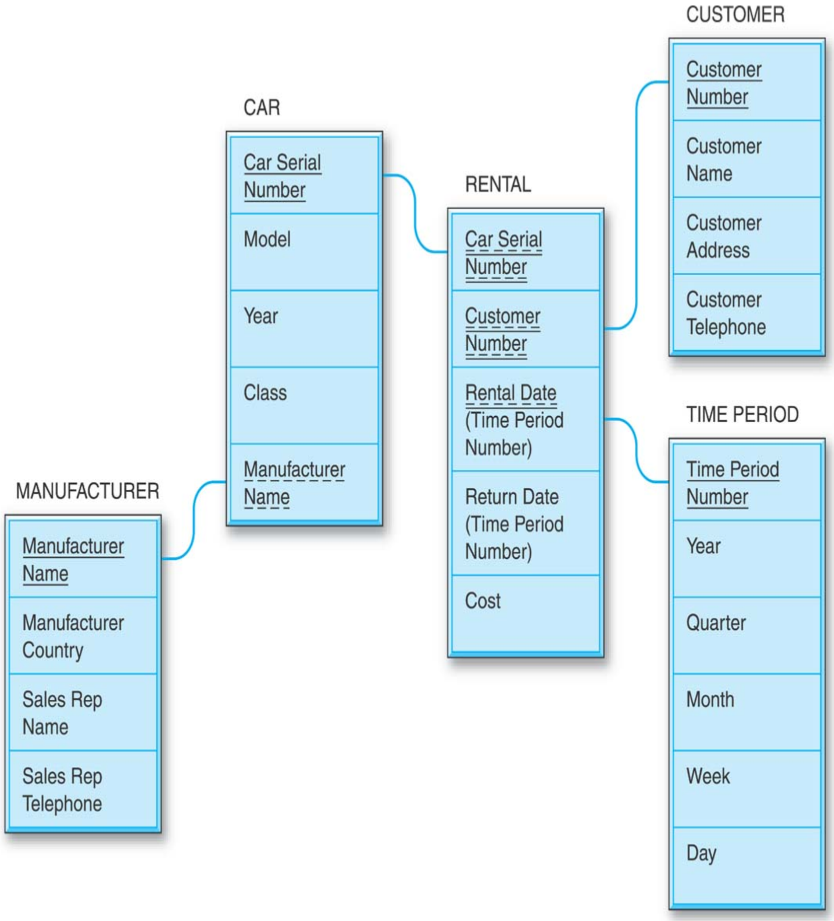
- Fact tables store important business measurements at the intersection of all the radiating dimensions.
- The best and most useful facts are
  - numeric
  - continuously valued
  - additive
- The huge number of rows must be “compressed” to form a useful summary.
- Fact tables may be sparse.
- Facts are sampled.

Dimension Tables

- Store discrete categories that identify the granularity of the measurements in fact tables.
- May have many attributes to provide a rich query environment.
- The best attributes are
  - ❑ Textual
  - ❑ Discrete
  - ❑ Used as the source of constraints and row headers in the user's answer set
- Are constant (or slowly changing).

Designing a Data Warehouse

- Snowflake design:
  - One dimension table leads to another dimension table



CUSTOMER
Customer Number
Customer Name
Customer Address
Customer Telephone

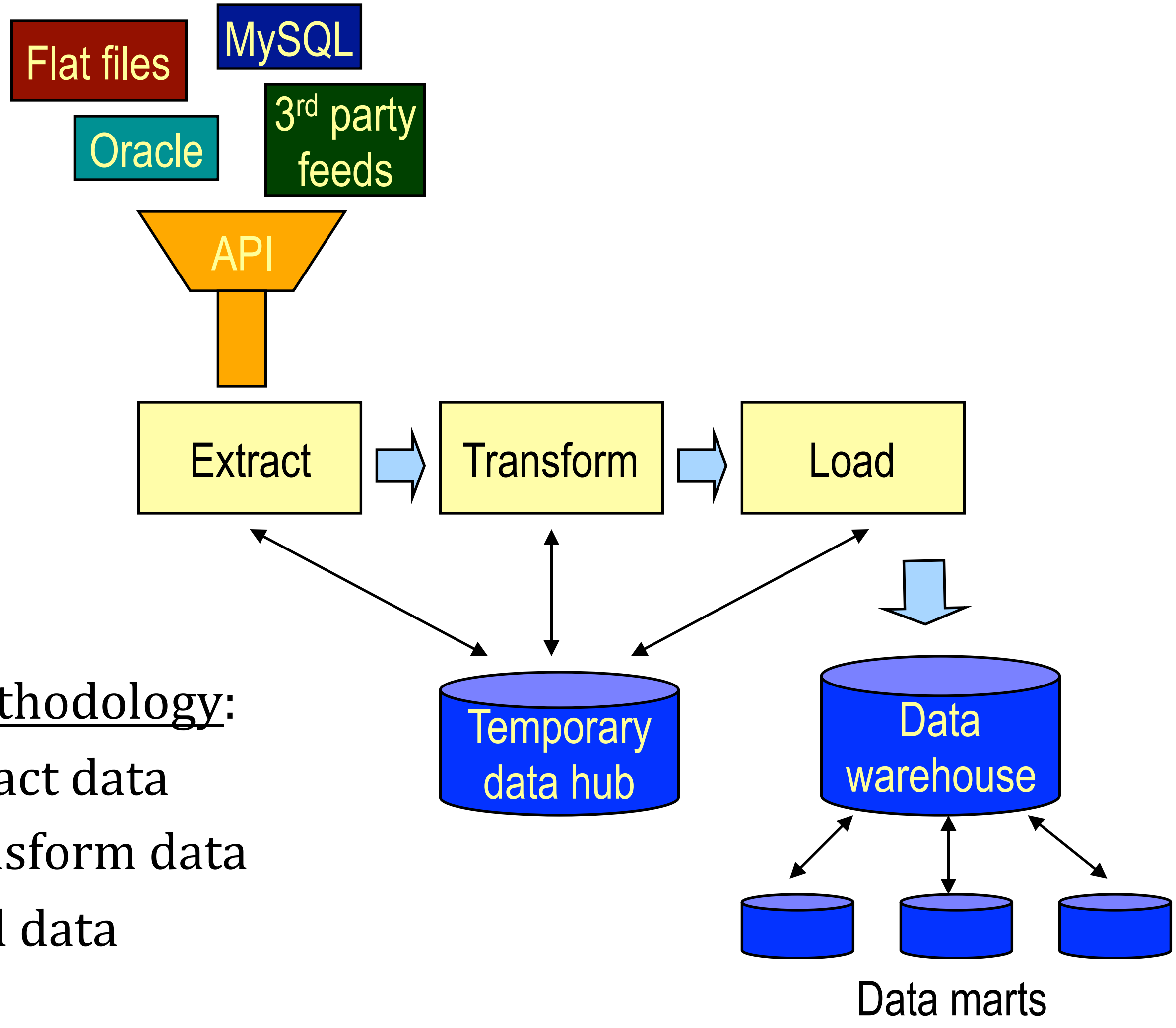
CAR
Car Serial Number
Model
Year
Class
Manufacturer Name

RENTAL
Car Serial Number
Customer Number
Rental Date (Time Period Number)
Return Date (Time Period Number)
Cost

TIME PERIOD
Time Period Number
Year
Quarter
Month
Week
Day

MANUFACTURER
Manufacturer Name
Manufacturer Country
Sales Rep Name
Sales Rep Telephone

Building a Data Warehouse



<u>ETL Methodology</u>:

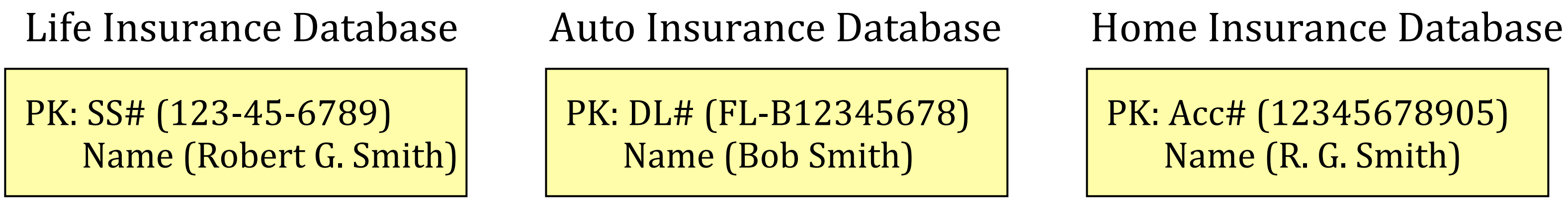
- Extract data
- Transform data
- Load data

Extraction

- Copying relevant data from variety of transactional databases
- May occur at regular intervals (e.g., weekly, monthly) to add new data
- Use API (application programming interfaces) to obtain data from incompatible databases, flat files, text documents, etc. as needed

Transformation/Cleansing

- Data extracted from transactional databases must be cleaned ("scrubbed") and transformed before loading into a DW
- Format differences across different tables/databases must be reconciled
- Missing or misspelled data values must be resolved
- Erroneous data are identified using application programs, and scrutinized / corrected by DW analysts using system-generated exception reports
- Transaction-level data is aggregated by business dimensions
- Key step in DW construction since DW is very sensitive to data errors



Life Insurance Database

PK: SS# (123-45-6789)
Name (Robert G. Smith)

Auto Insurance Database

PK: DL# (FL-B12345678)
Name (Bob Smith)

Home Insurance Database

PK: Acc# (12345678905)
Name (R. G. Smith)

Challenges of Data Reconciliation

Loading

- Extracted, cleaned, and transformed data is loaded into DW at a predetermined data refresh frequency
    - Hourly, daily, weekly, etc…

Data Cleaning Example

Sale Table

	Book Number	Customer Number	Date	Price	Quantity
1	426478	03480	Feb 19, 2011	32.99	1
2	077656	18575	Feb 19, 2011	19.95	201
3	365905	06837	Feb 19, 2011	24.99	3
4	645688	21359	Feb 20, 2011	49.50	1
5	474640	15367	Feb 34, 2011	3200.99	1
6	426478	08362	Mar 03, 2011	32.99	2
7	276432	03480	Mar 04, 2011	30.00	1
8	365905	12738	Mar 04, 2011	24.99	1
9	276432	06837	Mar 05, 2011	30.00	5
10	327467	18575	Mar 12, 2011	-32.99	2
11	426478	06837	Mar 15, 2011	32.99	1

Data Cleaning Example

Customer Table

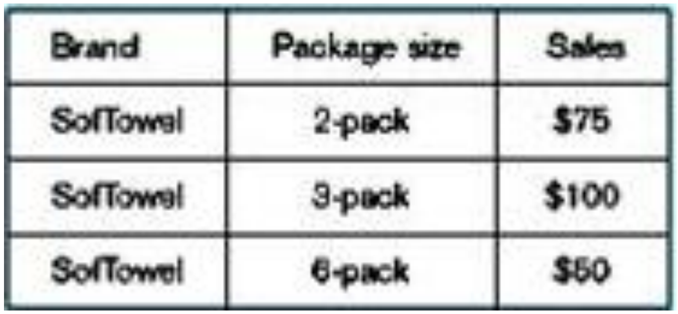
	Customer Number	Customer Name	Street	City	State	Country
1	02847	Mervis	123 Oak St.		TN	USA
2	03185	Gomez	345 Main Ave.	Columbus	OH	USA
3	03480	Taylor	50 Elm Rd.	San Diego	CA	USA
4	06837	Stevens	876 Leslie Ln.	Raleigh	NC	USA
5	08362	Adams	1200 Wallaby St.	Brisbane		Australia
6	12739	Gomez	345 Main Ave.	Columbus	GA	USA
7	13848	Lucas	742 Ave. Louise	Brussels		Belgium
8	15367	Tailor	50 Elm Rd.	San Diego	CA	USA
9	15933	Chang	48 Maple Ave.	Toronto	ON	Canada
10	18575	Smith	390 Martin Dr.	Columbus	RP	USA
11	21359	Sanchez	666 Ave. Bolivar	Santiago		Chile

Using a Data Warehouse – OLAP

- Online analytic processing (OLAP):
  - A decision support approach based on viewing data by dimensions
  - Well suited for multidimensional data hierarchies
  - Pre-compute data and store in cubes for fast response time
- OLAP techniques:
  - Drill-down: Retrieving finer levels of data detail (roll up opposite direction)
  - Slice: Data subset based on a single value of one dimension
  - Pivot or Rotation: Interchanging data dimensions in a slice

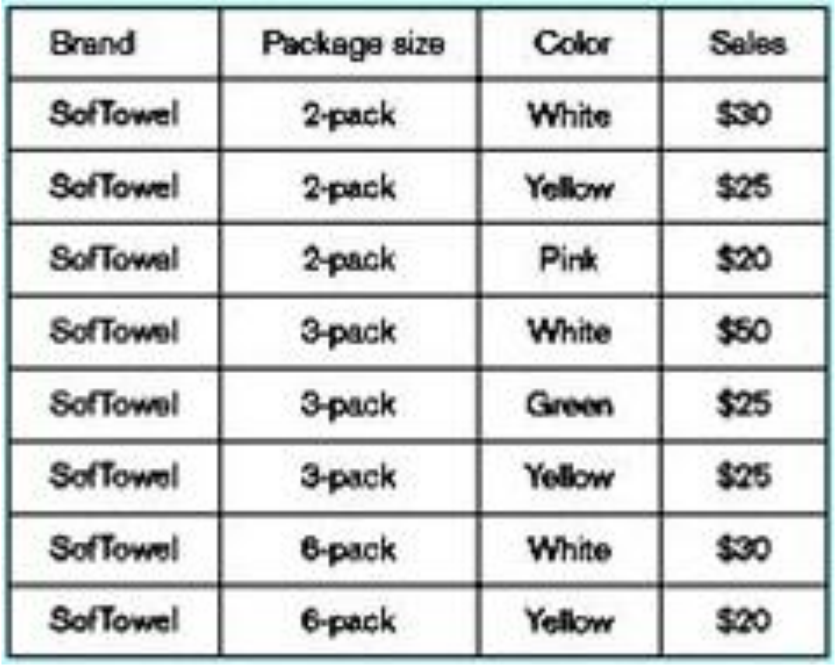
Drill Down

Drill-down by Package Size



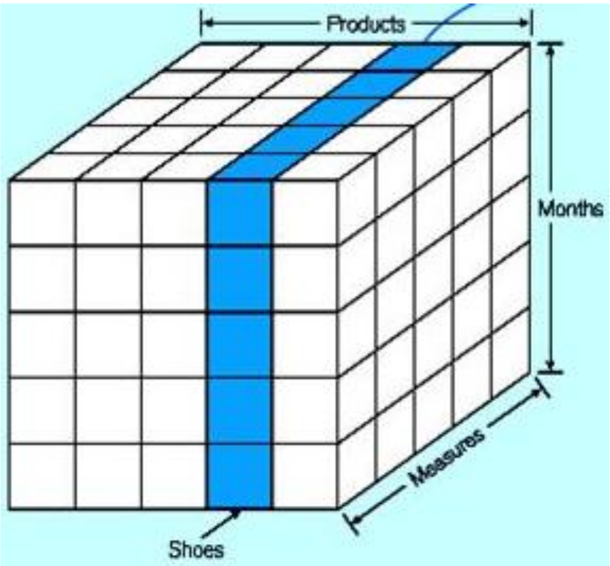
Brand	Package size	Sales
SofTowel	2-pack	$75
SofTowel	3-pack	$100
SofTowel	6-pack	$50

Drill-down by Package Size and Color

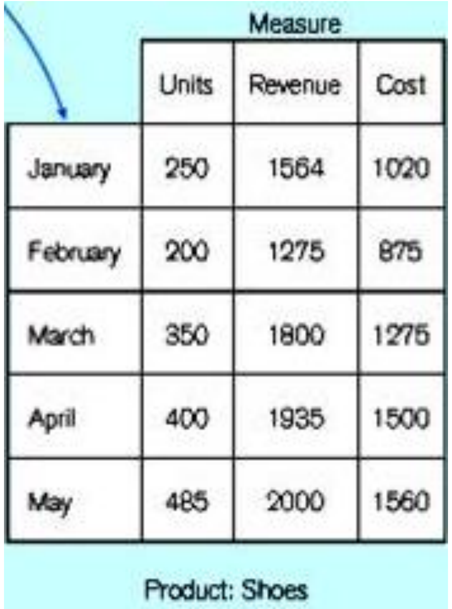


Brand	Package size	Color	Sales
SofTowel	2-pack	White	$30
SofTowel	2-pack	Yellow	$25
SofTowel	2-pack	Pink	$20
SofTowel	3-pack	White	$50
SofTowel	3-pack	Green	$25
SofTowel	3-pack	Yellow	$25
SofTowel	6-pack	White	$30
SofTowel	6-pack	Yellow	$20

Slice



Slice operation →

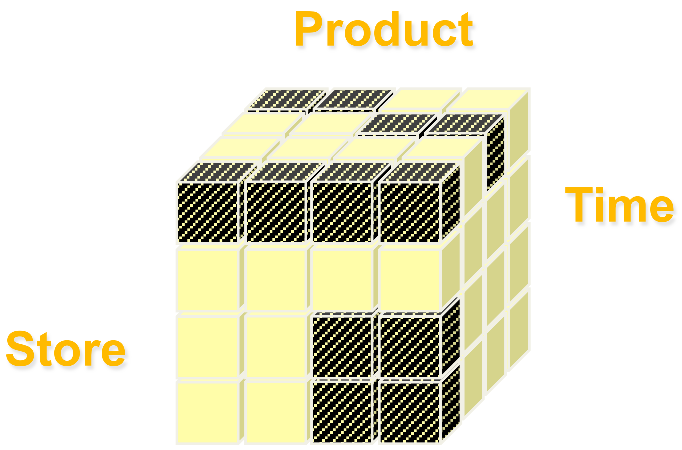


	Measure		
	Units	Revenue	Cost
January	250	1564	1020
February	200	1275	875
March	350	1800	1275
April	400	1935	1500
May	485	2000	1560

Product: Shoes

Star Join Queries

- SELECT p.brand, sum(f.dollars), sum(f.units)
- FROM salesfact f, product p, time t
- WHERE f.productkey = p.productkey
- AND f.timekey = t.timekey
- AND t.quarter = ‘1 Q 1995’
- GROUP BY p.brand
- ORDER BY p.brand



Kimball 1996

Challenges in Data Warehousing

- Data cleaning and finding more “dirty” data than expected
- Coordinating the regular appending of new data from transactional databases to the data warehouse
- Managing very large databases

More Information

- Oracle 11g Data Warehousing Guide
  http://download.oracle.com/docs/cd/B28359_01/server.111/b28313/toc.htm
- Courses
  - ❑ CSCI 4957– Data & Text Mining