MIT-WPU
॥ विश्वशान्तिर्ध्रुवं ध्रुवा ॥

Dr. Vishwanath Karad
MIT WORLD PEACE
UNIVERSITY | PUNE
TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS

Unit III: Data Warehousing

Syllabus

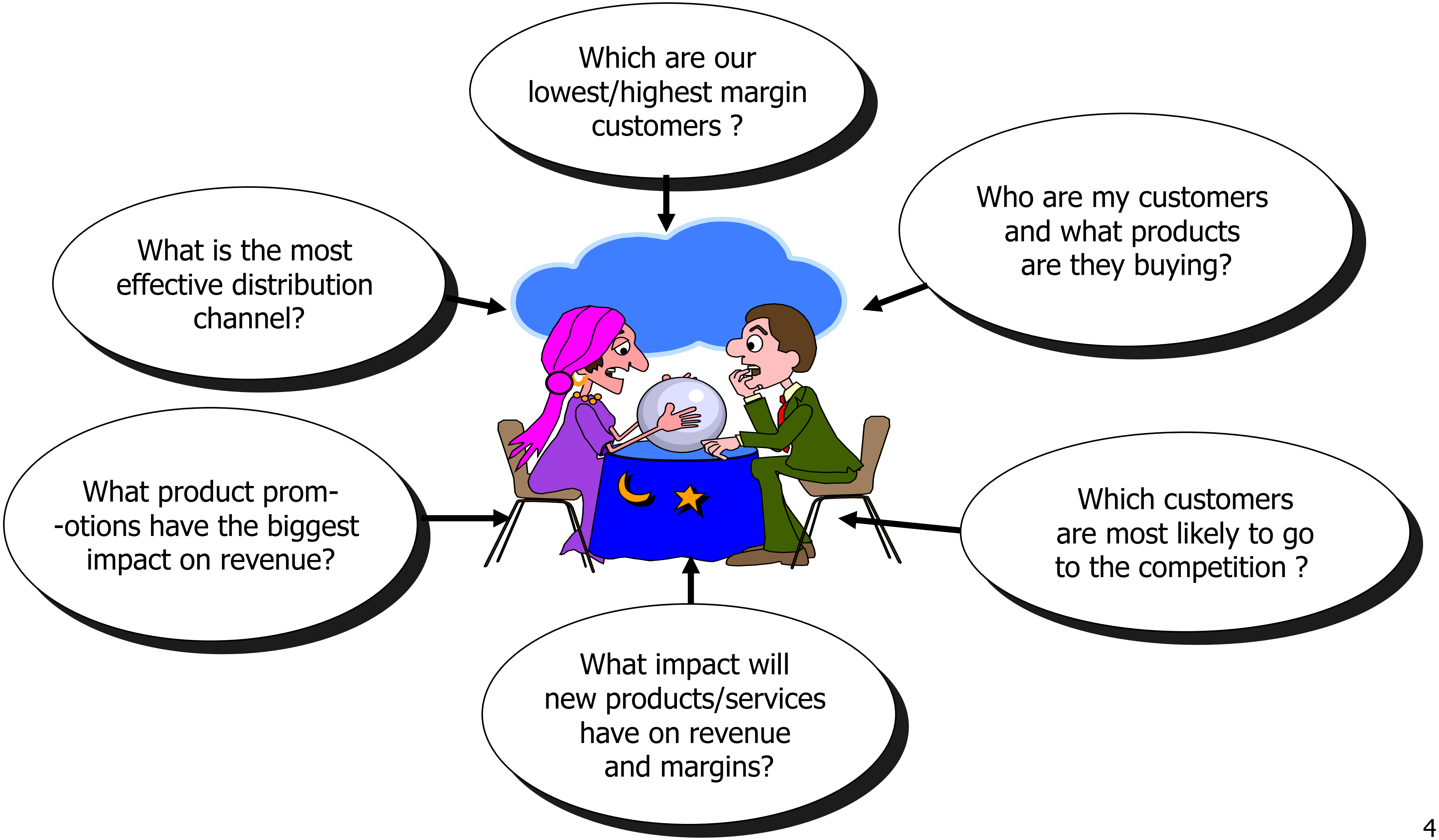
Data Warehouse

Data warehouse concepts, Data Warehouse Modeling: A Multidimensional Data Model, Data Warehouse Design: Stars, Snowflakes, and Fact Constellations Schemas, Data warehouse – design and usage, implementation, architectural components, Role of Metadata, Dimensional Modelling, Dimensions: The Role of Concept Hierarchies, Measures: Their Categorization and Computation, Materialized views.

UNIT II-Syllabus

- **DATA WAREHOUSE**
- Introduction to Data Warehouse
- OLTP and OLAP
- Data Warehouse architecture
- Data Warehouse Modeling: A Multidimensional Data Model
- Data Warehouse Design: Stars, Snowflakes, and Fact Constellations Schemas
- Dimensions: The Role of Concept Hierarchies
- Measures: Their Categorization and Computation
- Typical OLAP Operations, ROLAP, MOLAP,
- Materialized views
- Integration of Data Warehouse with other technologies

Scenario:A producer wants to know….



Which are our lowest/highest margin customers ?

Who are my customers and what products are they buying?

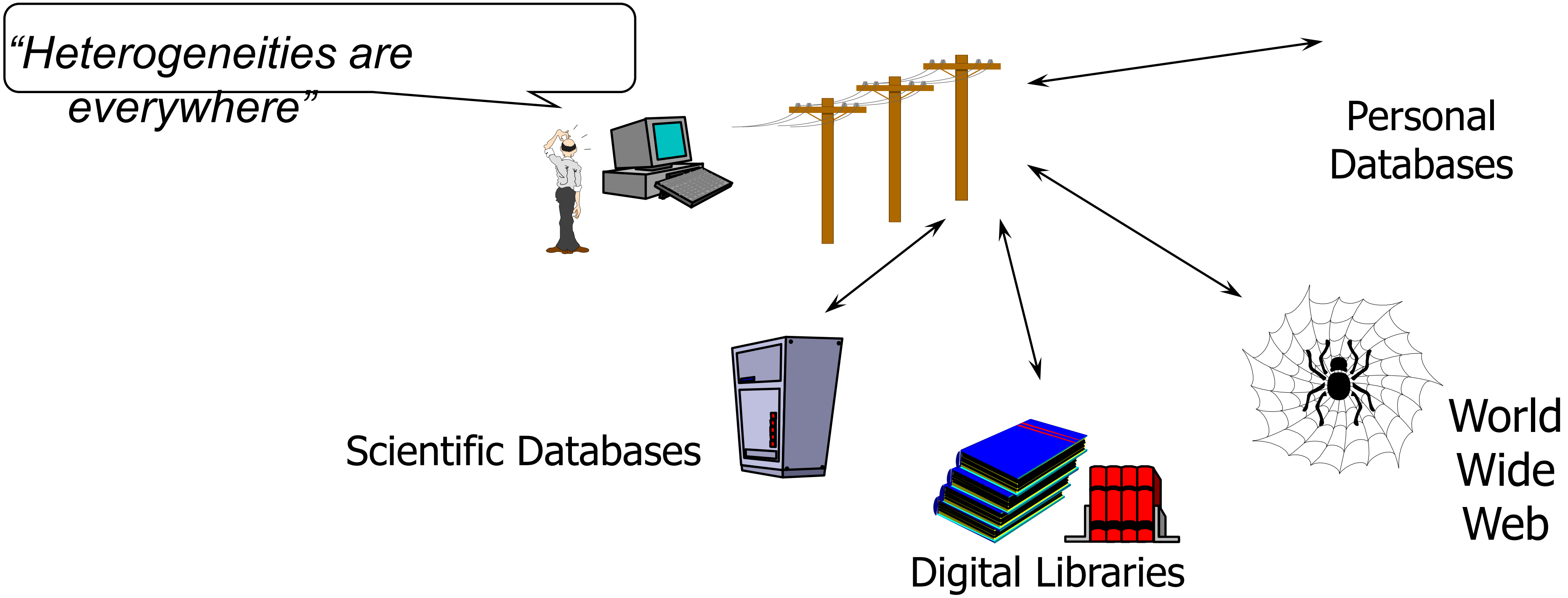
What is the most effective distribution channel?

What product prom- -otions have the biggest impact on revenue?

Which customers are most likely to go to the competition ?

What impact will new products/services have on revenue and margins?

Problem: Heterogeneous Information Sources



“Heterogeneities are everywhere”

Personal Databases

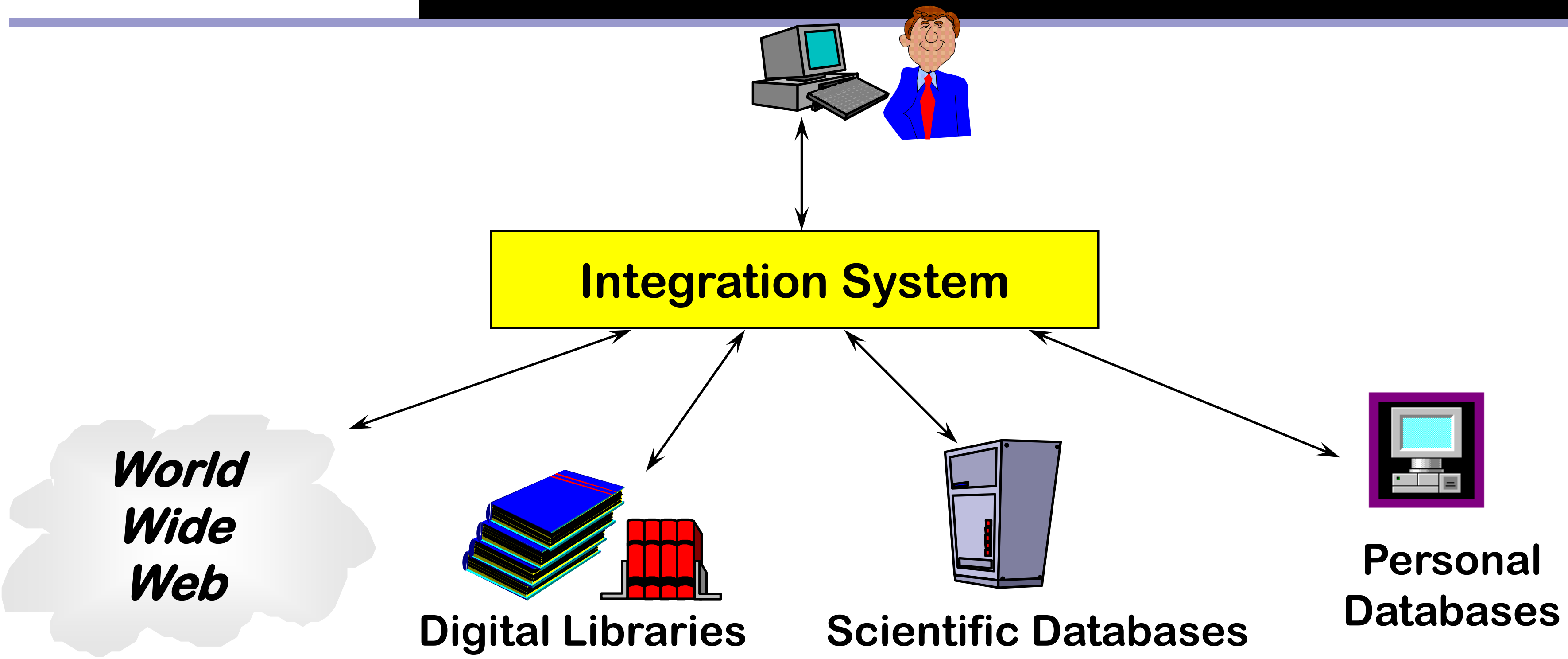
Scientific Databases

Digital Libraries

World Wide Web

- Different interfaces
- Different data representations
- Duplicate and inconsistent information

Unified Access to Data



- Collects and combines information
- Provides integrated view, uniform user interface
- Supports sharing

Data, Data everywhere yet ...

- Can’t find the data needed
  - data is scattered over the network
  - many versions, subtle differences
- Can’t get the data needed
  - need an expert to get the data
- Can’t understand the data found
  - available data poorly documented
- Can’t use the data found
  - results are unexpected
  - data needs to be transformed from one form to other

What is Data Warehouse?

- A single, complete and consistent store of data obtained from a variety of different sources, made available to end users in a way they can understand and use in a business context. [Barry Devlin]

- A decision support database that is maintained separately from the organization’s operational databases.

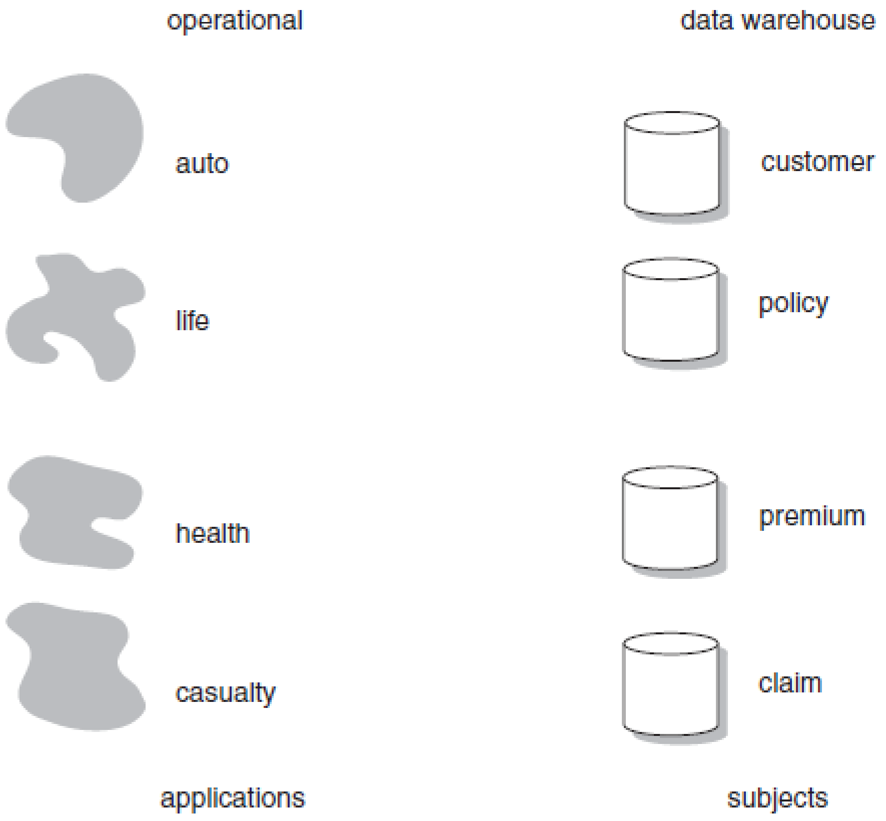
- Support information processing by providing a solid platform of consolidated, historical data for analysis

- “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 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

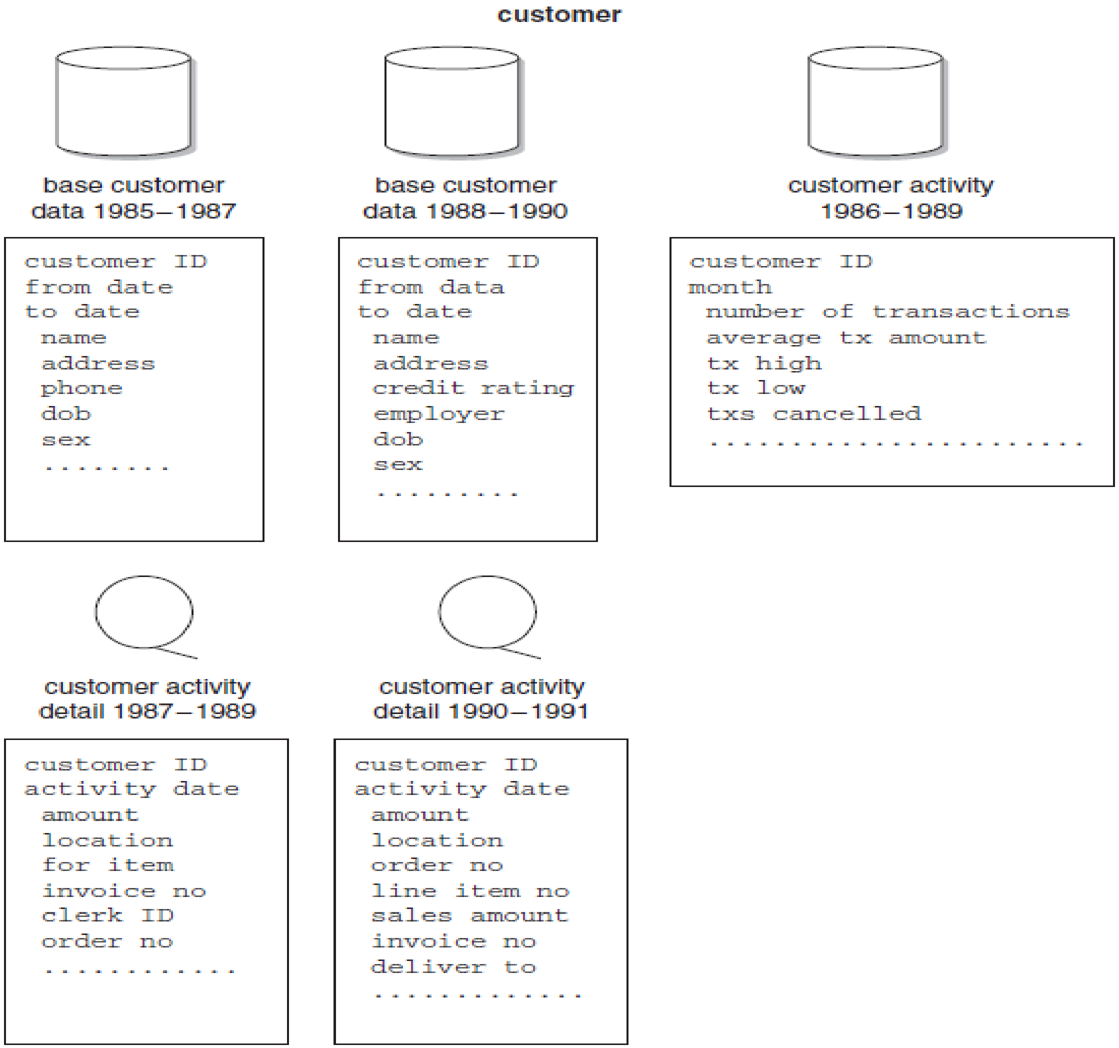
Data Warehouse -Subject-Oriented



Example :
Subject area—customer

- There is a base table for customer information as defined from 1985 to 1987.
- There is another for the definition of customer data between 1988 and 1990.
- There is a cumulative customer activity table for activities between 1986 and 1989.
- Each month a summary record is written for each customer record based on customer activity for the month.
- There are detailed activity files by customer for 1987 through 1989 and another one for 1990 through 1991.
- The definition of the data in the files is different, based on the year.
- All of the physical tables for the customer subject area are related by a common key.
- Figure shows that the key—customer ID—connects all of the data

Example :
Subject area—customer



customer

base customer data 1985–1987

```
customer ID
from date
to date
 name
 address
 phone
 dob
 sex
 ........
```

base customer data 1988–1990

```
customer ID
from data
to date
 name
 address
 credit rating
 employer
 dob
 sex
 .........
```

customer activity 1986–1989

```
customer ID
month
 number of transactions
 average tx amount
 tx high
 tx low
 txs cancelled
 .......................
```

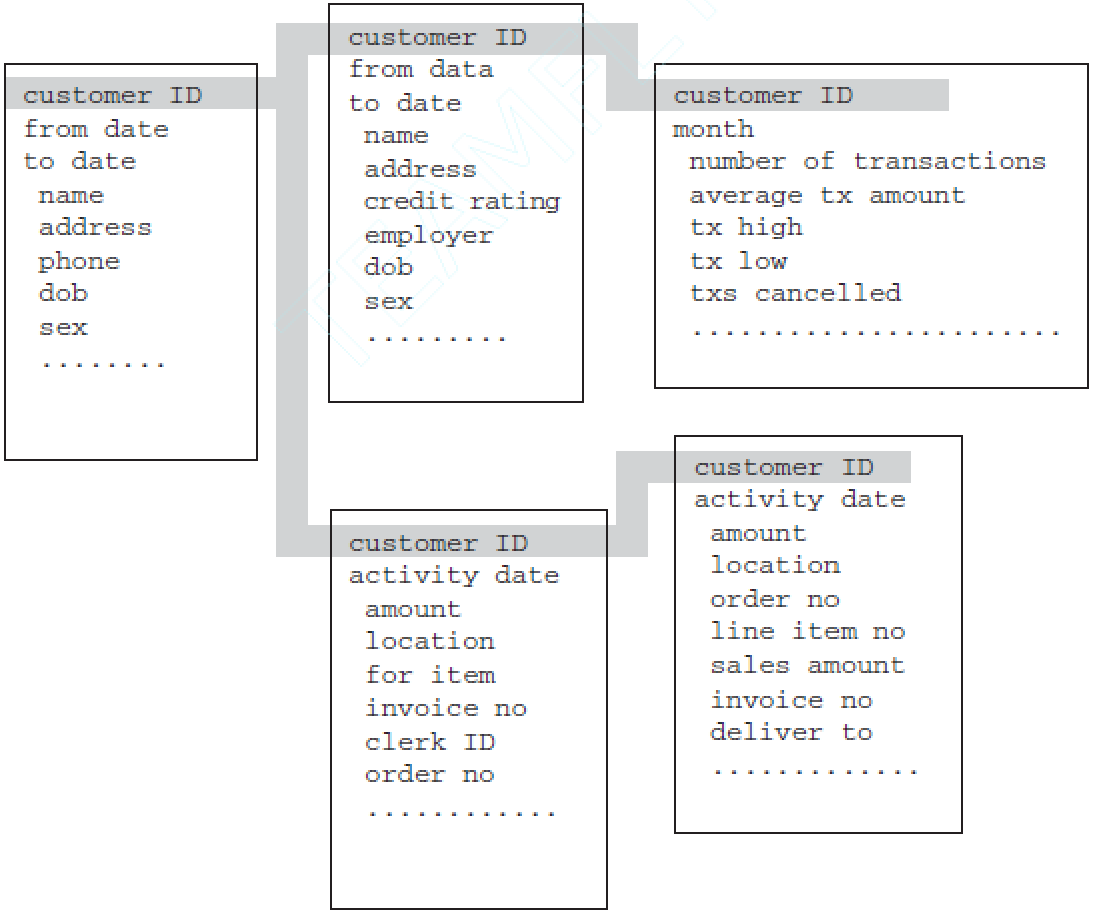
customer activity detail 1987–1989

```
customer ID
activity date
 amount
 location
 for item
 invoice no
 clerk ID
 order no
 ............
```

customer activity detail 1990–1991

```
customer ID
activity date
 amount
 location
 order no
 line item no
 sales amount
 invoice no
 deliver to
 .............
```

Example :
Subject area—customer



customer ID
from date
to date
name
address
phone
dob
sex
. . . .

customer ID
from data
to date
name
address
credit rating
employer
dob
sex
. . . .

customer ID
month
number of transactions
average tx amount
tx high
tx low
txs cancelled
. . . .

customer ID
activity date
amount
location
for item
invoice no
clerk ID
order no
. . . .

customer ID
activity date
amount
location
order no
line item no
sales amount
invoice no
deliver to
. . . .

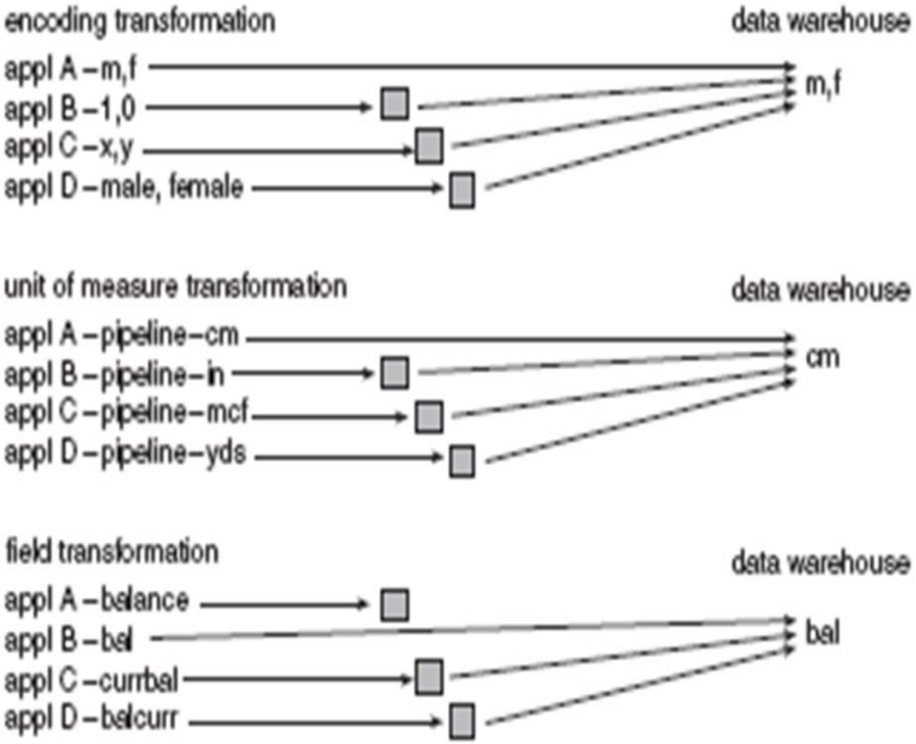
Data Warehouse—Integrated

- 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 is moved to the warehouse, it is converted.

Data Warehouse—Integrated

- Across multiple applications there is no application consistency in encoding, naming conventions, physical attributes, measurement of attributes, and so forth.
- Each application designer has had free rein to make his or her own design decisions.

Data Warehouse—Integrated



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—Time Variant

- Different environments have different time horizons. A time horizon is the parameters of time represented in an environment.
- The collective time horizon for the data found inside a data warehouse is significantly longer than that of operational systems.
- A 60-to-90-day time horizon is normal for operational systems;
- A 5-to-10-year time horizon is normal for the data warehouse.
- As a result of this difference in time horizons, the data warehouse contains *much* more history than any other environment

Data Warehouse—Time Variant

operational
data warehouse

operational

- time horizon—current to 60–90 days
- update of records
- key structure may/may not contain an element of time

data warehouse

- time horizon—5–10 years
- sophisticated snapshots of data
- key structure contains an element of time

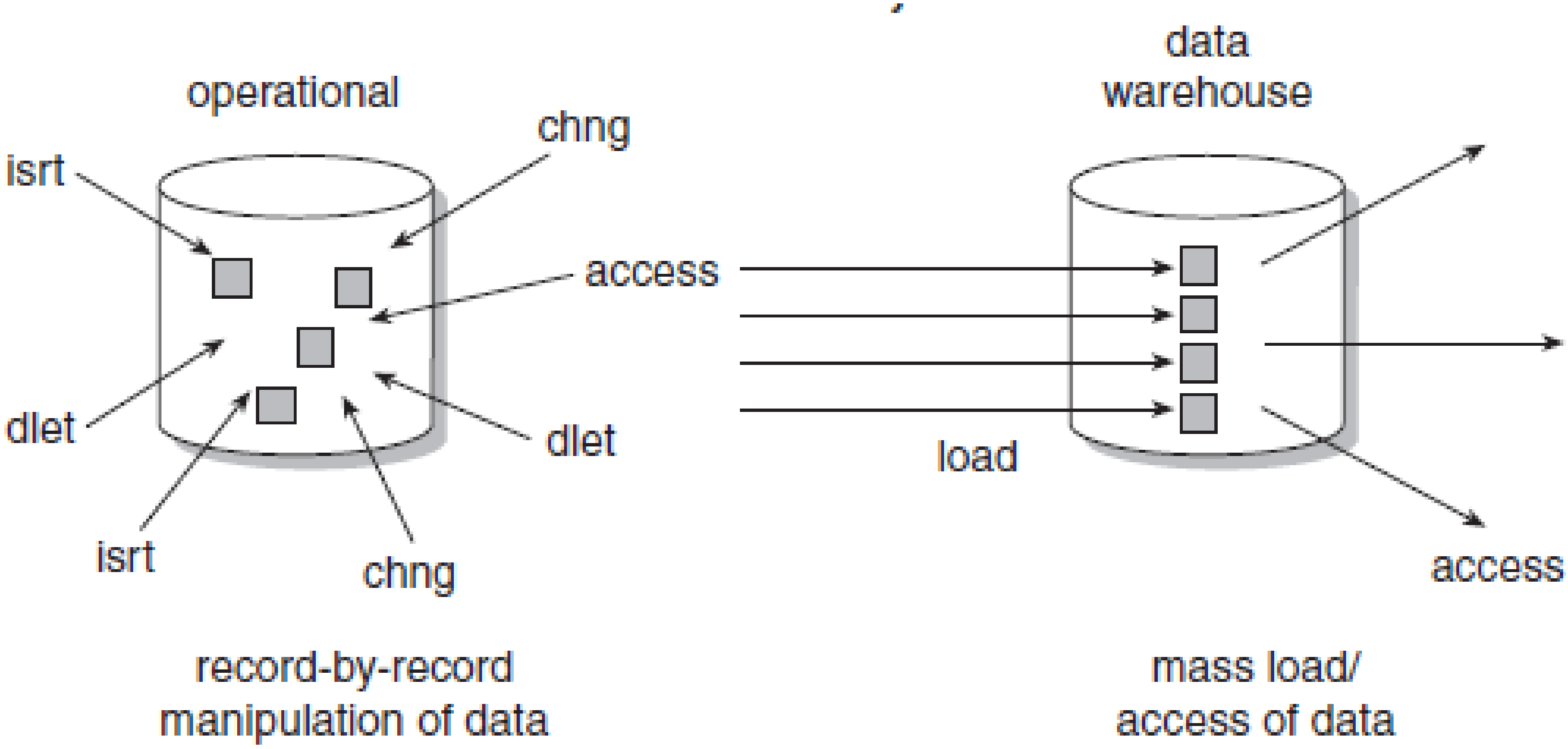
Data Warehouse—Nonvolatile

- 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*

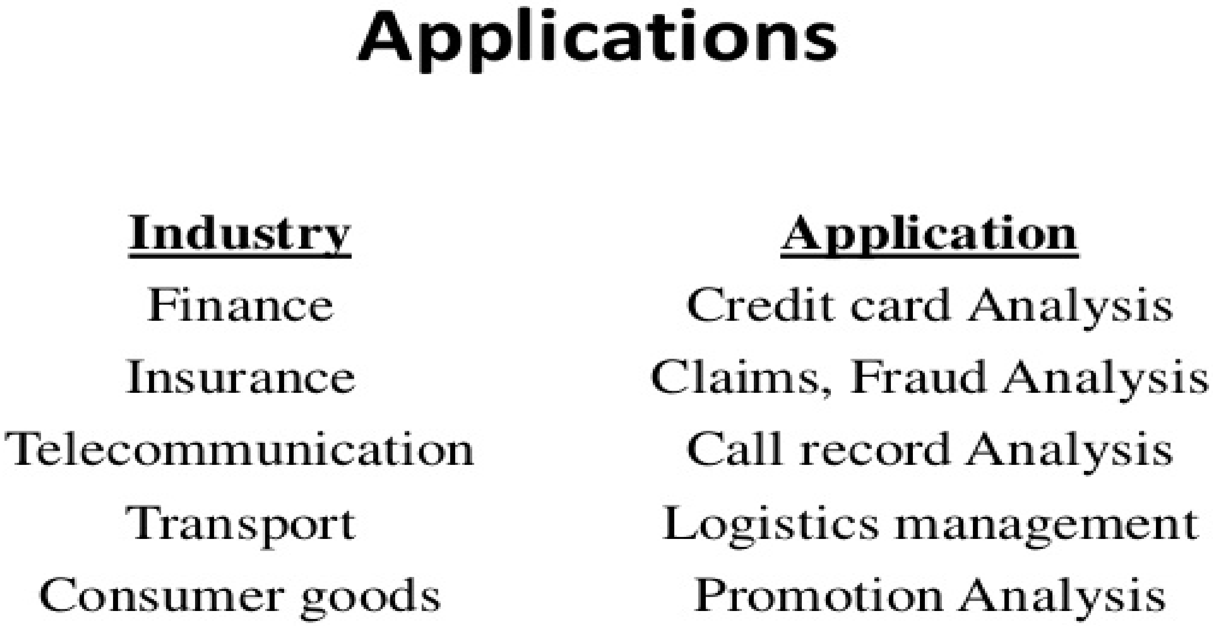
Data Warehouse—Nonvolatile

- Operational data is regularly accessed and manipulated one record at a time.
- Data is updated in the operational environment as a regular matter of course, but data warehouse data exhibits a very different set of characteristics.
- Data warehouse data is loaded (usually en masse) and accessed, but it is not updated (in the general sense).
- Instead, when data in the data warehouse is loaded, it is loaded in a snapshot, static format.
- When subsequent changes occur, a new snapshot record is written.
- In doing so a history of data is kept in the data warehouse.

Data Warehouse—Nonvolatile



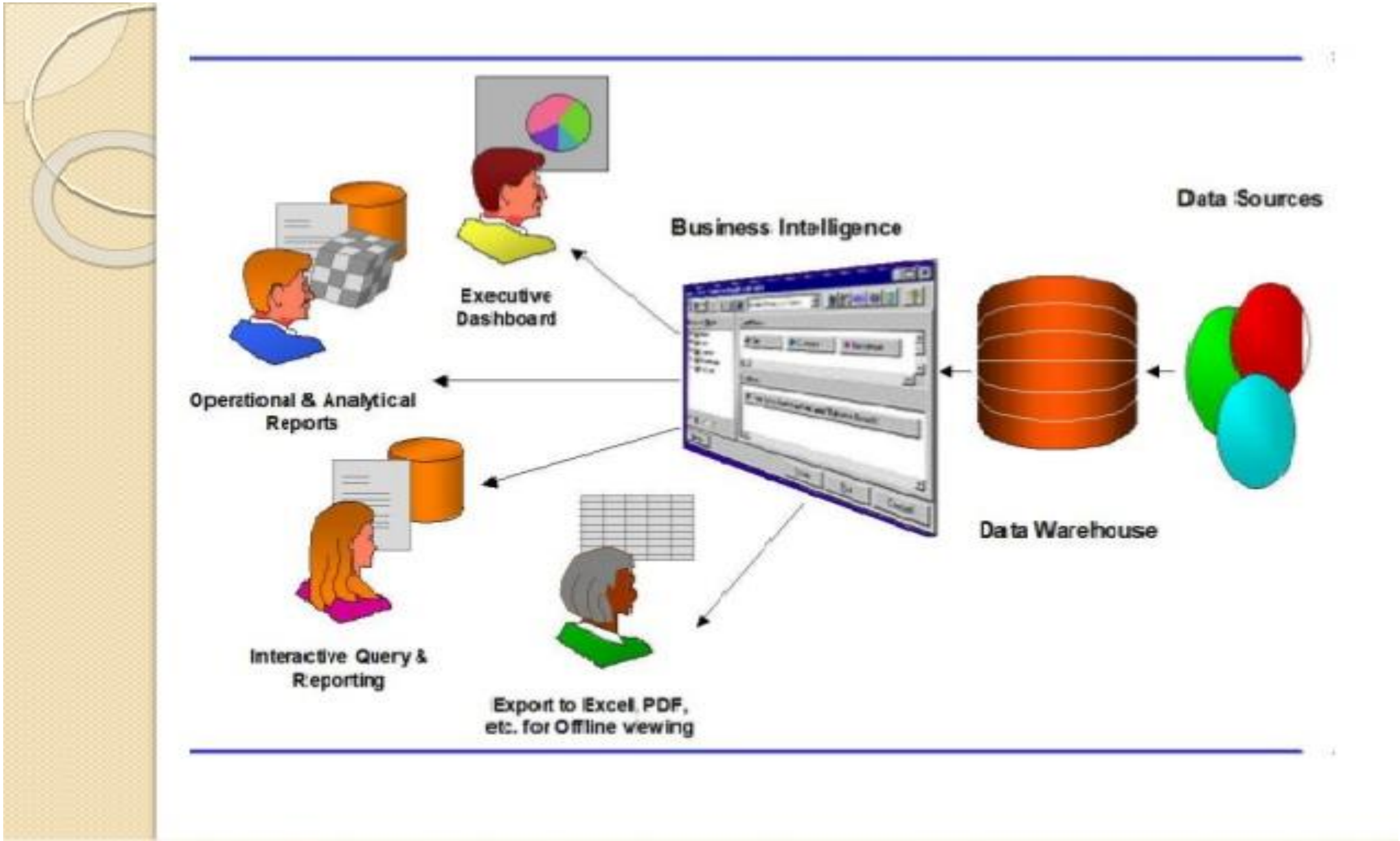
Data Warehouse



Applications

Industry	Application
Finance	Credit card Analysis
Insurance	Claims, Fraud Analysis
Telecommunication	Call record Analysis
Transport	Logistics management
Consumer goods	Promotion Analysis

Data Warehouse:Users


Data Sources
Business Intelligence
Executive Dashboard
Operational & Analytical Reports
Interactive Query & Reporting
Export to Excel PDF, etc. for Offline viewing
Data Warehouse

Data Warehouse vs. Operational DBMS

- OLTP (on-line transaction processing)
  - Major task of traditional relational DBMS
  - Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.
- OLAP (on-line analytical processing)
  - Major task of data warehouse system
  - Data analysis and decision making
- Distinct features (OLTP vs. OLAP):
  - User and system orientation: customer vs. market
  - Data contents: current, detailed vs. historical, consolidated
  - Database design: ER + application vs. star + subject
  - View: current, local vs. evolutionary, integrated
  - Access patterns: update vs. read-only but complex queries

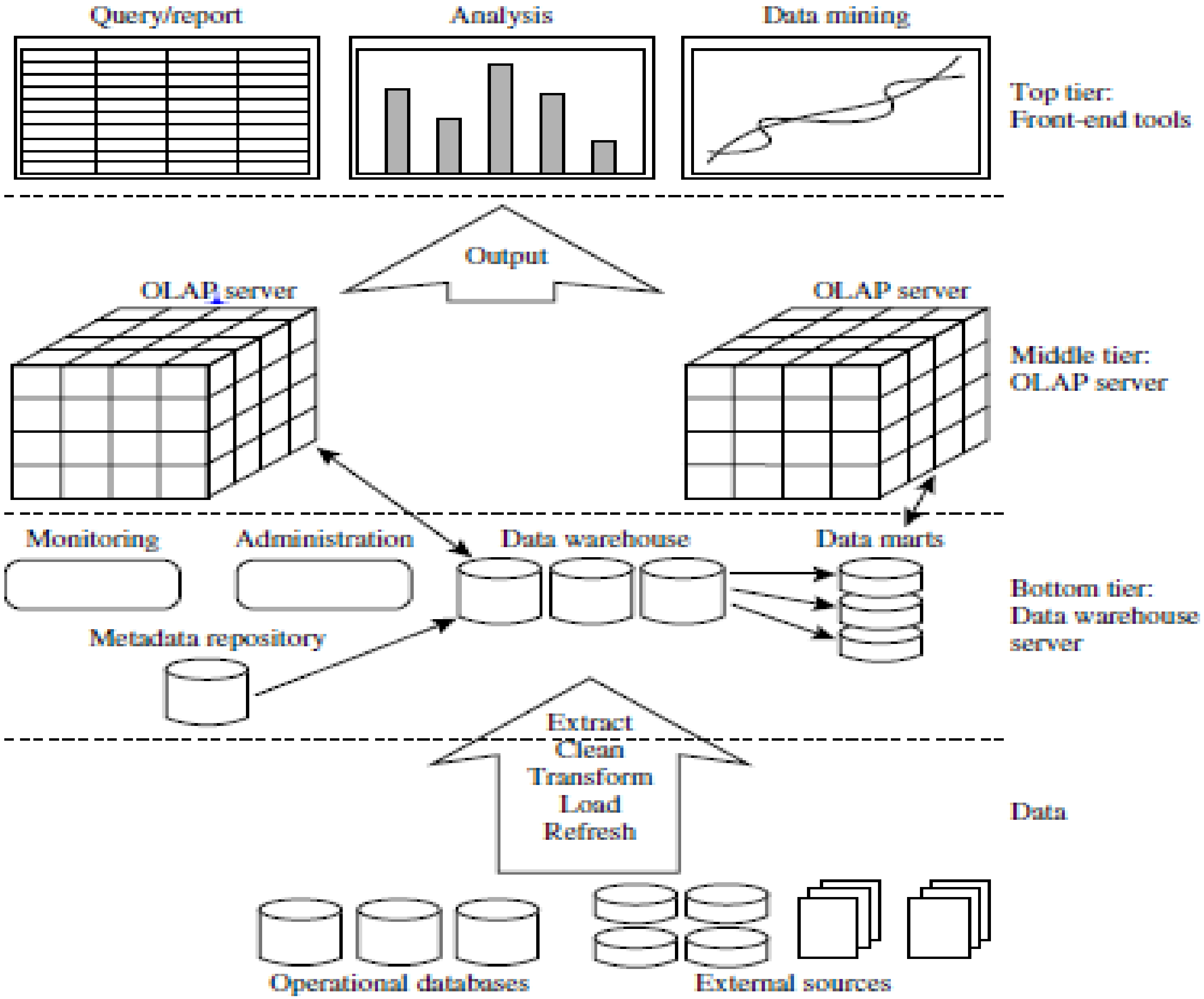
OLTP vs. OLAP

	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

Why 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

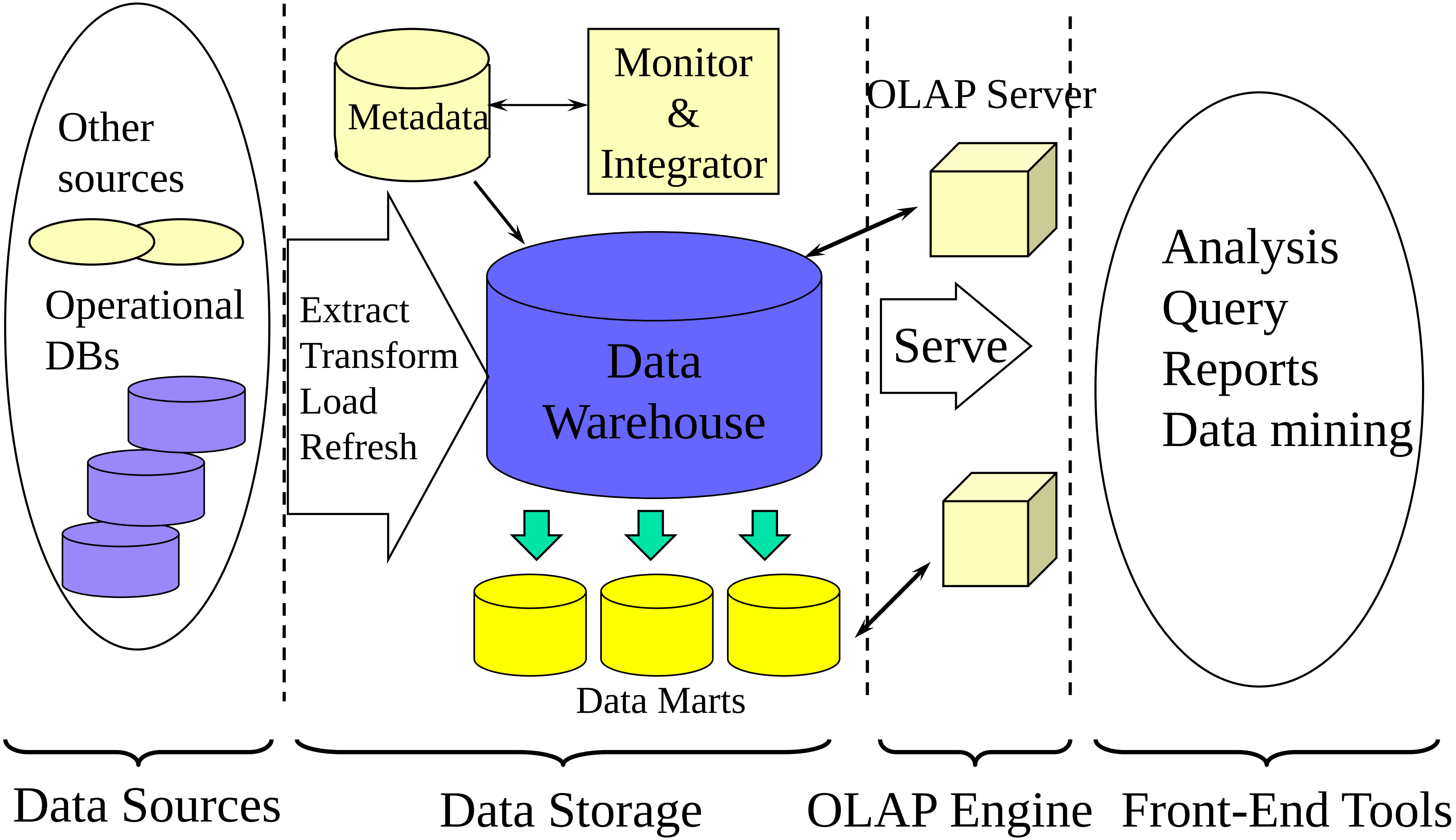
DataWarehousing: A Three-tire Architecture



Data Warehouse Models

- Enterprise warehouse
  - collects all of the information about subjects spanning the entire organization
- Data Mart
  - a subset of corporate-wide/organizational data that is of value to a specific groups of users. Its scope is confined to specific, selected groups, such as marketing data mart
    - Independent vs. dependent (directly from warehouse) data mart
- Virtual warehouse
  - A set of views over operational databases
  - Only some of the possible summary views may be materialized
- Cloud-based data warehouse
  - They can typically perform complex analytical queries much faster because they are massively parallel processing (MPP).

Data Warehouse: A Multi-Tiered Arhitecture



Data Warehouse Back-End Tools and Utilities

- **Data extraction**
  - get data from multiple, heterogeneous, and external sources
- **Data cleaning**
  - detect errors in the data and rectify them when possible
- **Data transformation**
  - convert data from legacy or host format to warehouse format
- **Load**
  - sort, summarize, consolidate, compute views, check integrity, and build indices and partitions
- **Refresh**
  - propagate the updates from the data sources to the warehouse

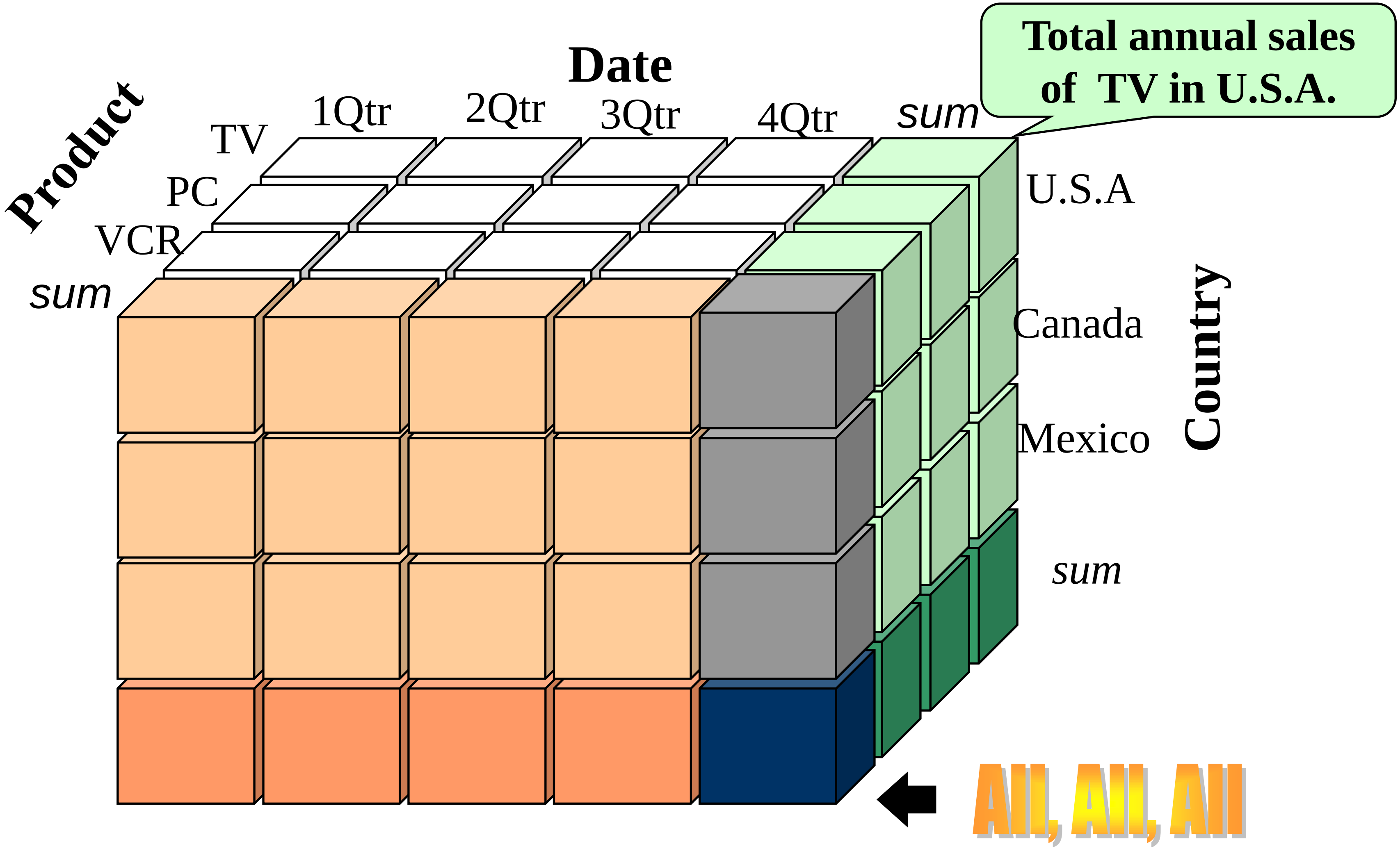
Metadata Repository

- Meta data is the data defining warehouse objects. It stores:
- **Description of the structure of the data warehouse**
  - schema, view, dimensions, hierarchies, derived data defn, data mart locations and contents
- **Operational meta-data**
  - data lineage (history of migrated data and transformation path), currency of data (active, archived, or purged), monitoring information (warehouse usage statistics, error reports, audit trails)
- **The algorithms used for summarization**
- **The mapping from operational environment to the data warehouse**
- **Data related to system performance**
  - warehouse schema, view and derived data definitions
- **Business data**
  - business terms and definitions, ownership of data, charging policies

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
- In data warehousing literature, an n-D base cube is called a base cuboid. The top most 0-D cuboid, which holds the highest-level of summarization, is called the apex cuboid. The lattice of cuboids forms a data cube.

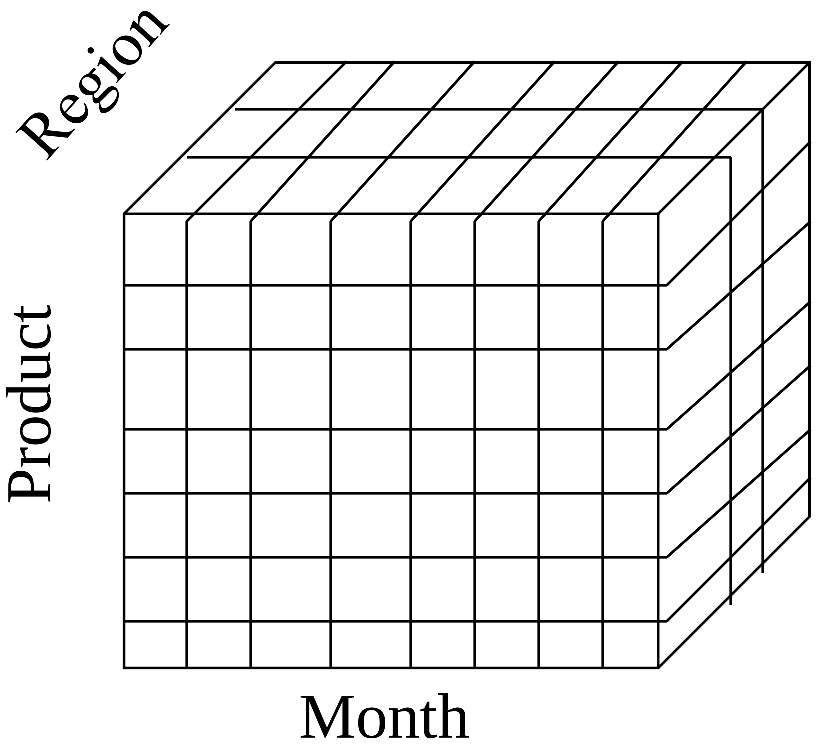
A Sample Data Cube for Sales



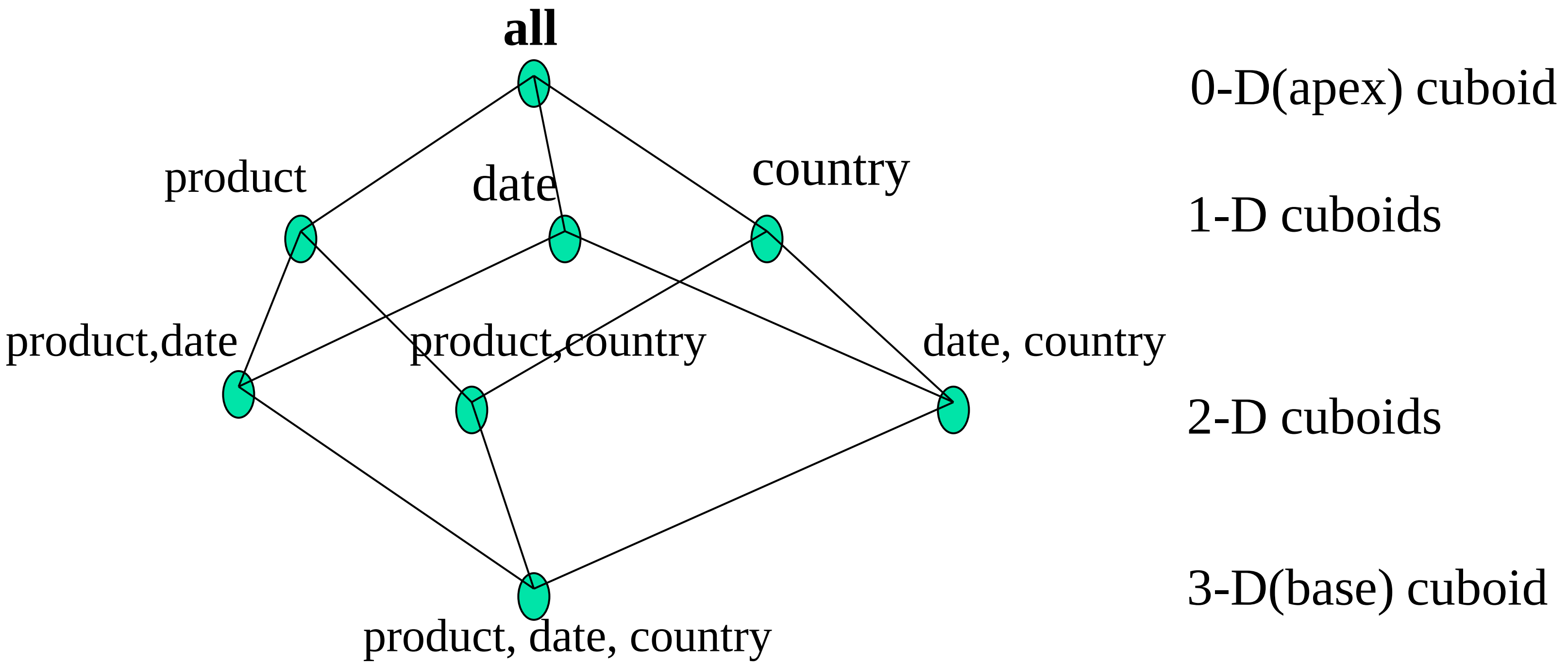
Multidimensional Data

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

Dimensions: Product, Location, Time



Cuboids Corresponding to the Cube



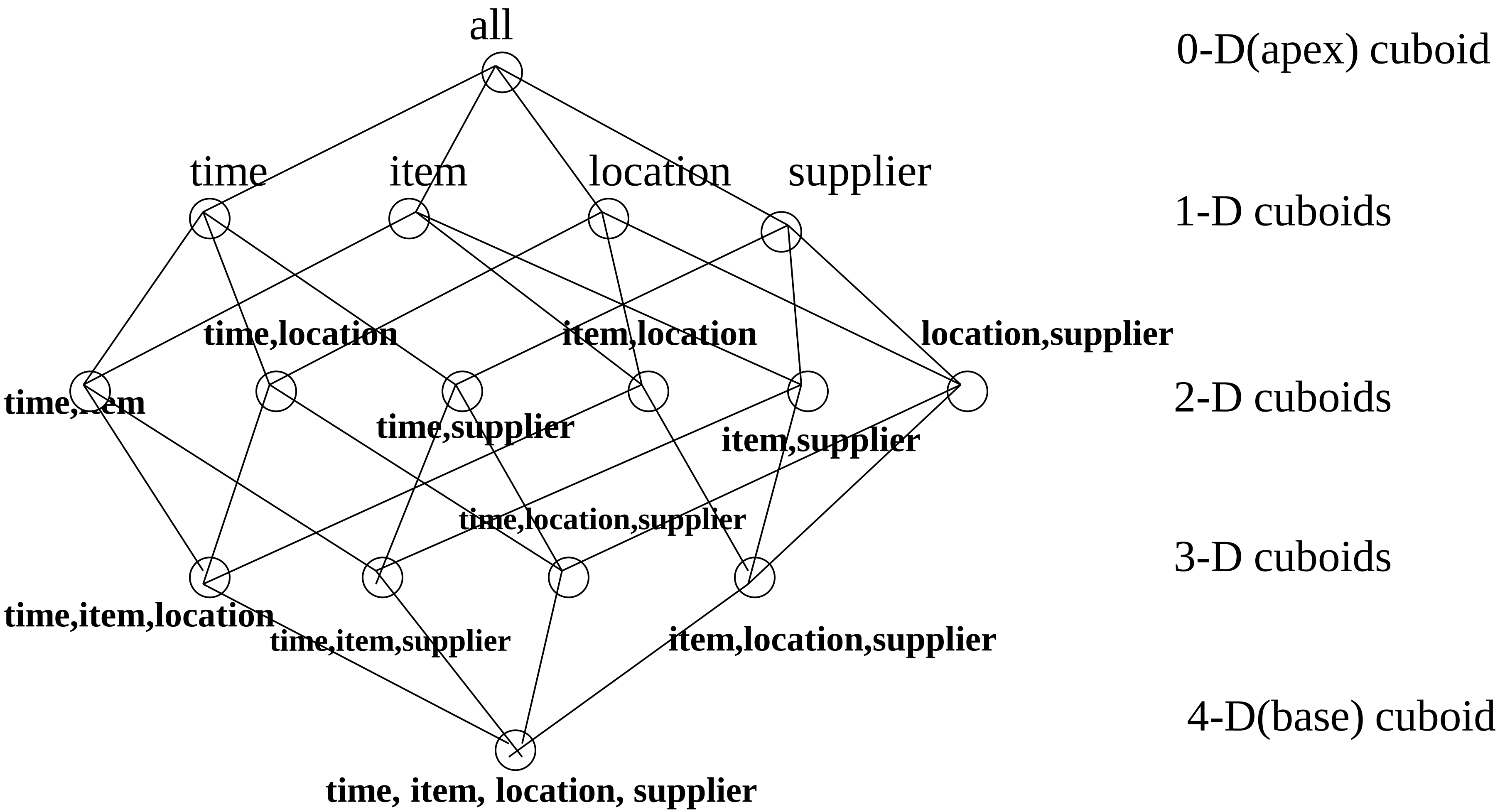
The **base cuboid contains all three dimensions**,. It can return the total sales for any combination of the three dimensions.

The **apex cuboid, or 0-D cuboid, refers to the case where the group-by is empty**. It contains the total sum of all sales.

The base cuboid is the least generalized (most specific) of the cuboids.

The apex cuboid is the most generalized (least specific) of the cuboids, and is often denoted as all.

Cube: A Lattice of Cuboids



N-D View of Data : Cuboid

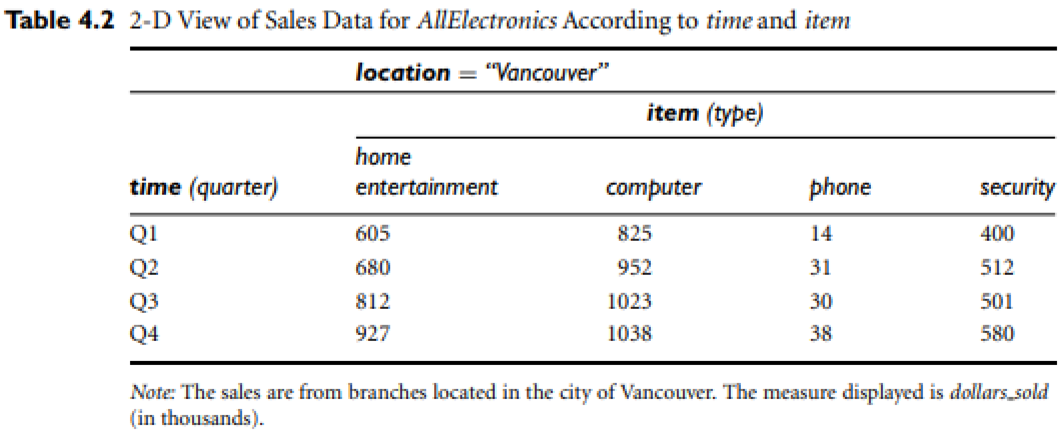


Table 4.2 2-D View of Sales Data for *AllElectronics* According to *time* and *item*

	location = "Vancouver"			
	item *(type)*			
time *(quarter)*	*home entertainment*	*computer*	*phone*	*security*
Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	30	501
Q4	927	1038	38	580

Note: The sales are from branches located in the city of Vancouver. The measure displayed is *dollars_sold* (in thousands).

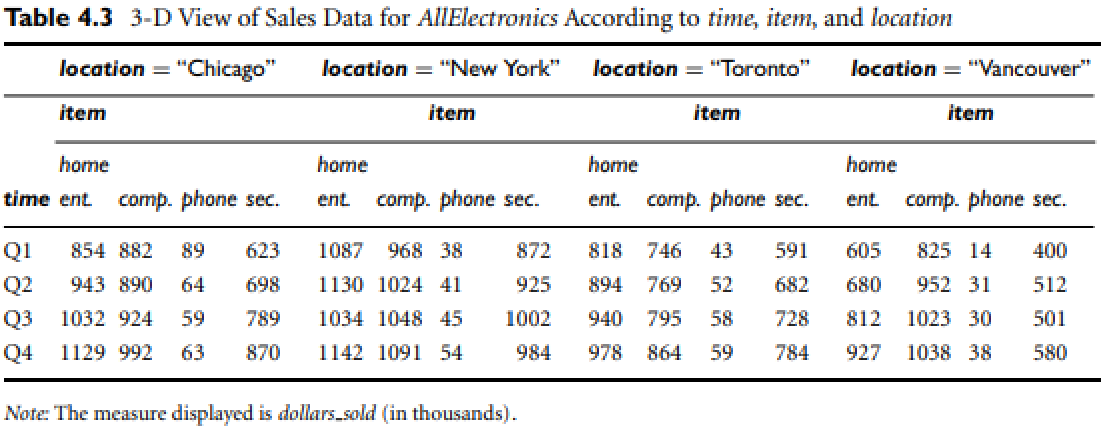


Table 4.3 3-D View of Sales Data for *AllElectronics* According to *time, item,* and *location*

	location = "Chicago"				***location*** = "New York"				***location*** = "Toronto"				***location*** = "Vancouver"			
	item				***item***				***item***				***item***			
time	*home ent.*	*comp.*	*phone*	*sec.*	*home ent.*	*comp.*	*phone*	*sec.*	*home ent.*	*comp.*	*phone*	*sec.*	*home ent.*	*comp.*	*phone*	*sec.*
Q1	854	882	89	623	1087	968	38	872	818	746	43	591	605	825	14	400
Q2	943	890	64	698	1130	1024	41	925	894	769	52	682	680	952	31	512
Q3	1032	924	59	789	1034	1048	45	1002	940	795	58	728	812	1023	30	501
Q4	1129	992	63	870	1142	1091	54	984	978	864	59	784	927	1038	38	580

Note: The measure displayed is *dollars_sold* (in thousands).

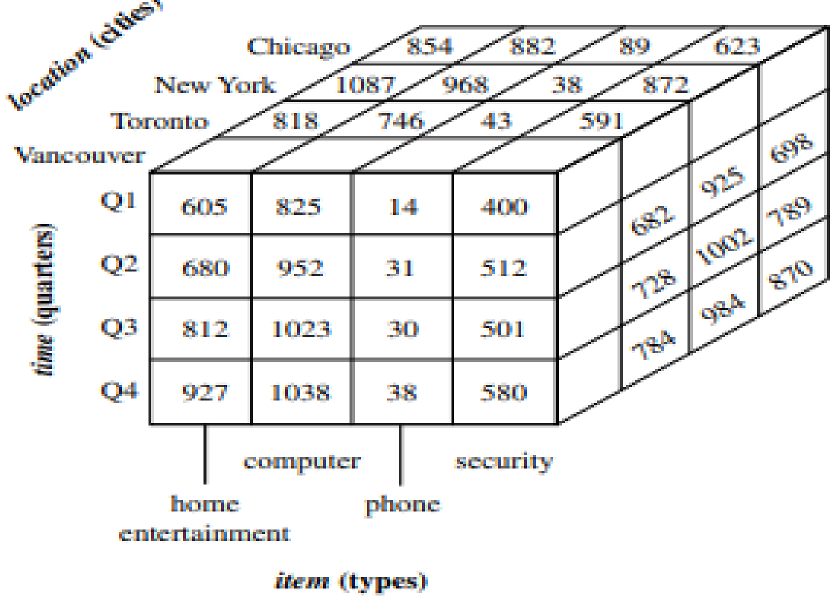


Figure 4.3 A 3-D data cube representation of the data in Table 4.3, according to *time*, *item*, and *location*. The measure displayed is *dollars_sold* (in thousands).

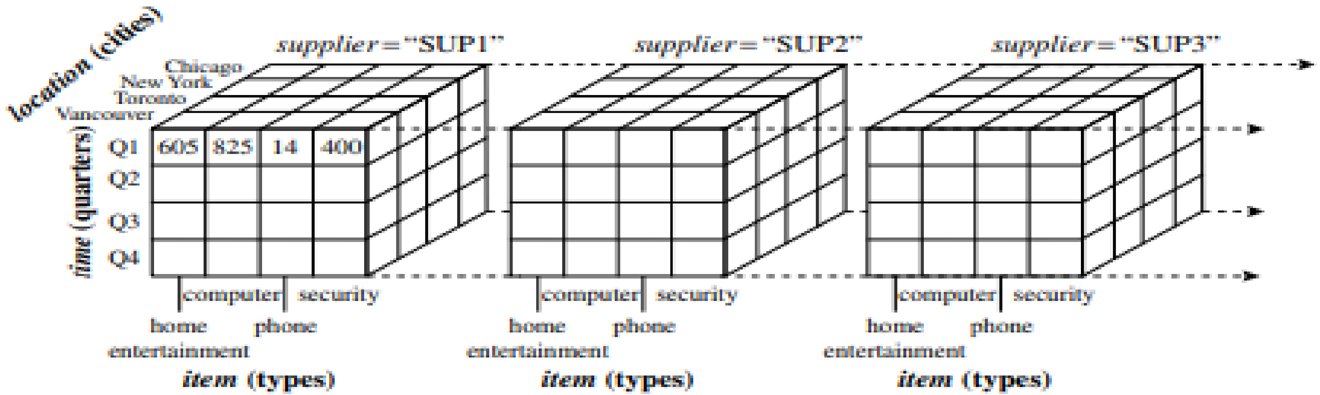


Figure 4.4 A 4-D data cube representation of sales data, according to *time*, *item*, *location*, and *supplier*. The measure displayed is *dollars_sold* (in thousands). For improved readability, only some of the cube values are shown.

Design Requirements

- Design of the DW must directly reflect the way the managers look at the business
- Should capture the measurements of importance along with parameters by which these parameters are viewed
- Must facilitate data analysis, i.e., answering business questions

Dimensional nature of business data

- Managers think of business in terms of business dimensions
- **Marketing vice president** is interested in revenue numbers broken down by

  -month

  -division

  -customer

  -demographic

  -sales office

  -product version

  -plan.

These are the business dimensions

Dimensional nature of business data

- **Marketing Manager** business dimensions are

  -product

  -product category

  -time(day,week,month)

  -sale district

  -distribution channel

Dimensional nature of business data

- **Financial Controller** business dimensions are

  -budget line

  -time(month,quarter,year)

  -district

  -division

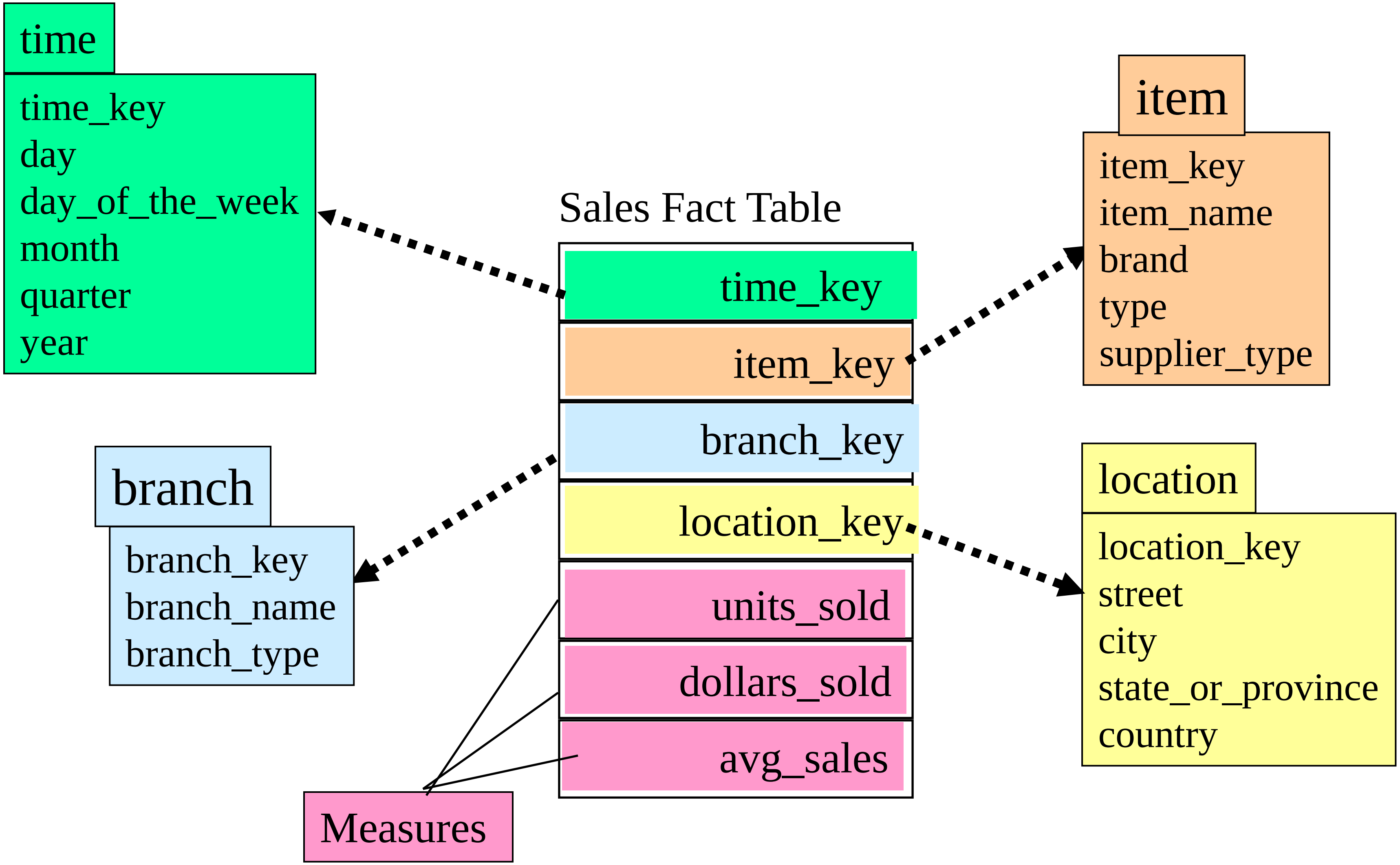
Modeling of Data Warehouses

- Modeling data warehouses: dimensions & measures
  - Star schema: A fact table in the middle connected to a set of dimension tables
  - Snowflake schema: A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake
  - Fact constellations: Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation

Star Schema

- A single fact table and a single table for each dimension
- Every fact points to one tuple in each of the dimensions and has additional attributes
- Does not capture hierarchies directly
- Generated keys are used for performance and maintenance reasons

Example of Star Schema



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

item

item_key
item_name
brand
type
supplier_type

branch

branch_key
branch_name
branch_type

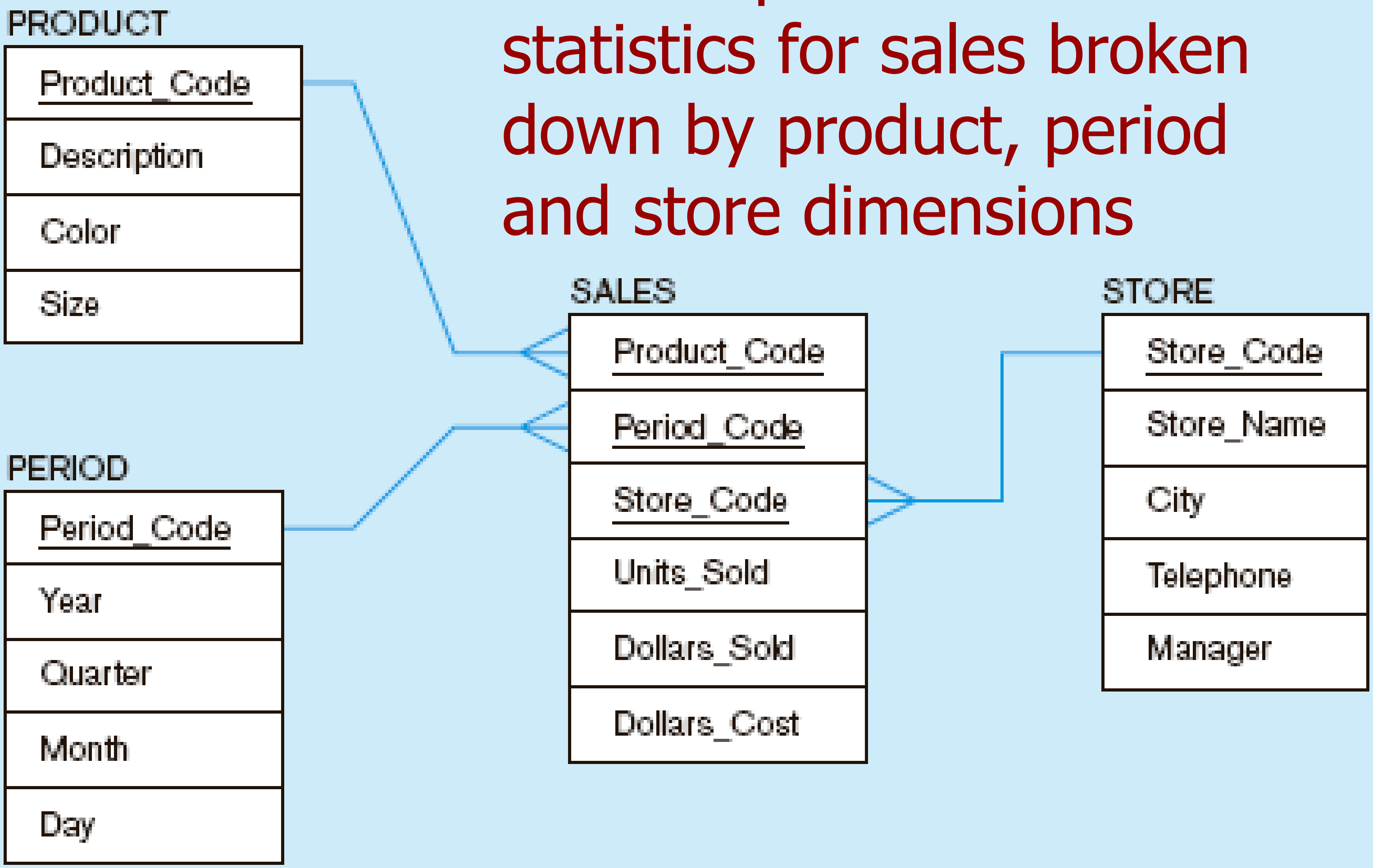
location

location_key
street
city
state_or_province
country

Measures

Star schema example

Fact table provides statistics for sales broken down by product, period and store dimensions



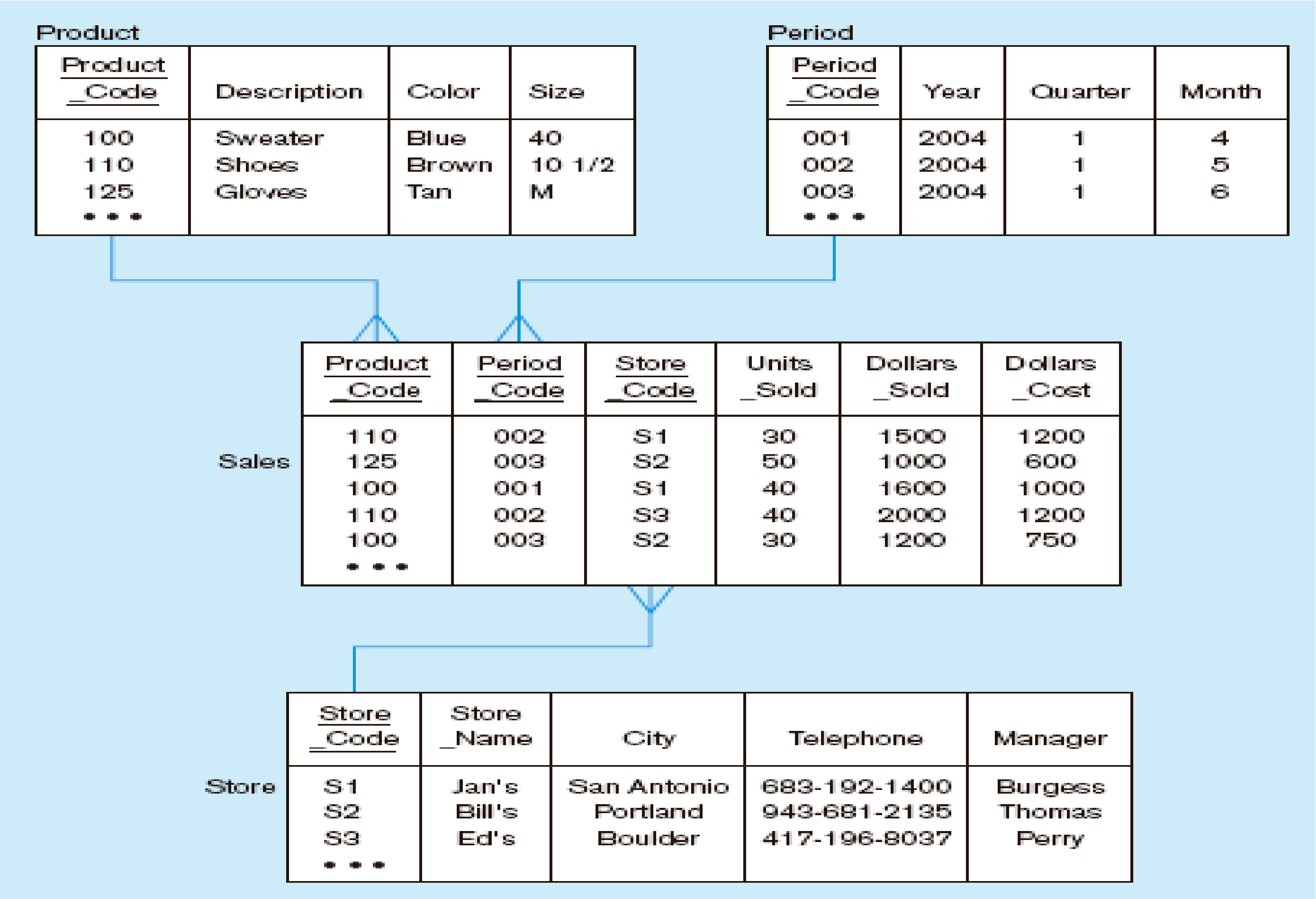
PRODUCT
Product_Code
Description
Color
Size

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	2004	1	4
002	2004	1	5
003	2004	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
• • •				

STAR Schema keys

- Primary Keys-Uniquely identifies a a record in dimension table

- Surrogate Keys- System generated sequence numbers
  - Do not have any built in meanings.
  - Operational system keys are stored as additional attributes

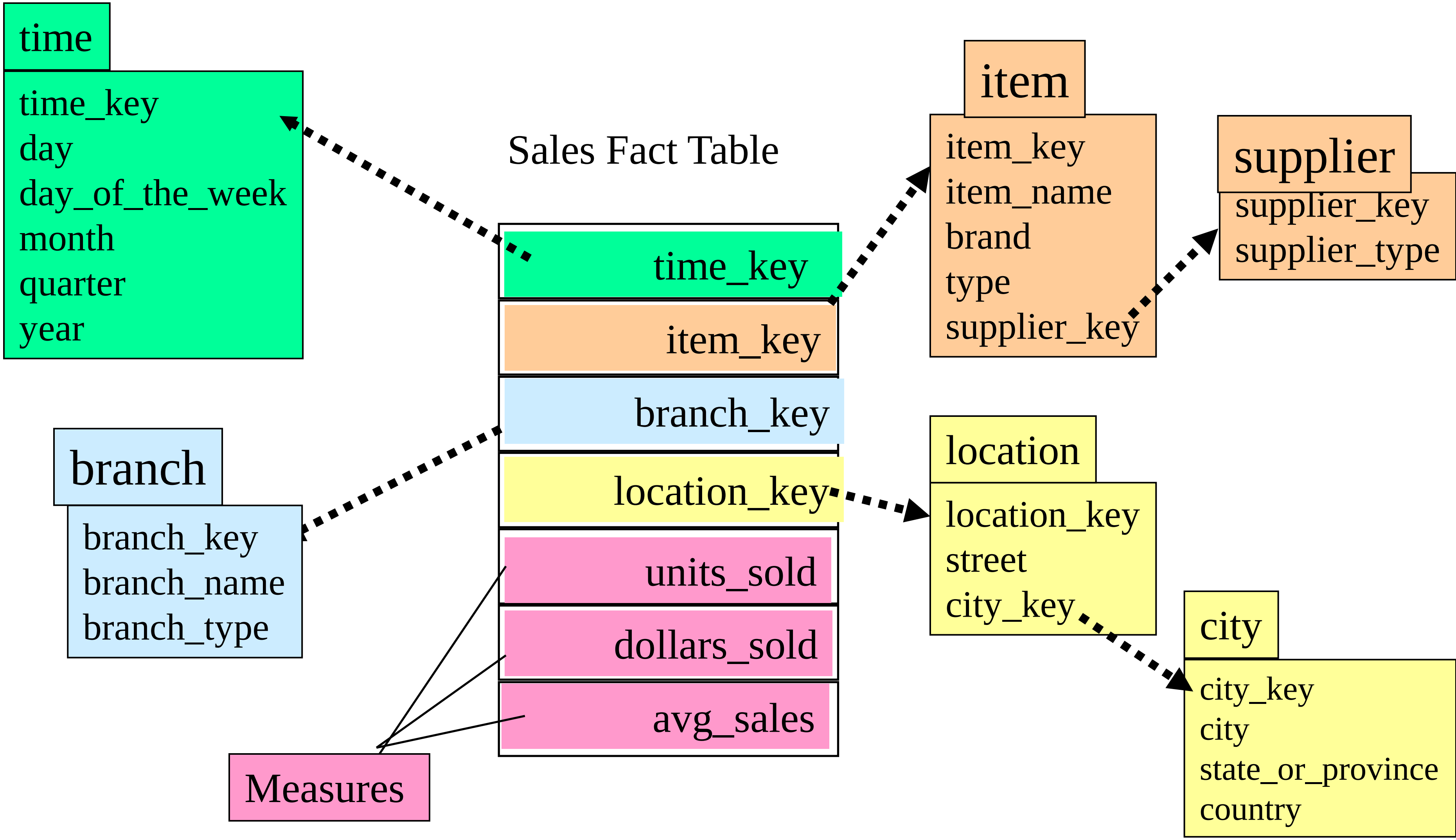
Pros & Cons of STAR Schema

- **Star schema pros**
  - simple design;
  - fast read and queries;
  - easy data aggregation
  - easy integration with OLAP systems and data cubes.
- **Star schema cons**
  - redundant data makes for larger storage on disk;
  - potential for data abnormalities, errors and inconsistencies;
  - slower queries;
  - limited flexibility on non-dimensional data.

Snowflake Schema

- A snowflake schema database is similar to a star schema in that it has a single fact table and many dimension tables.
- For a snowflake schema, each dimension table might have foreign keys that relate to other dimension tables.
- A snowflake schema is more normalized

Example of Snowflake Schema



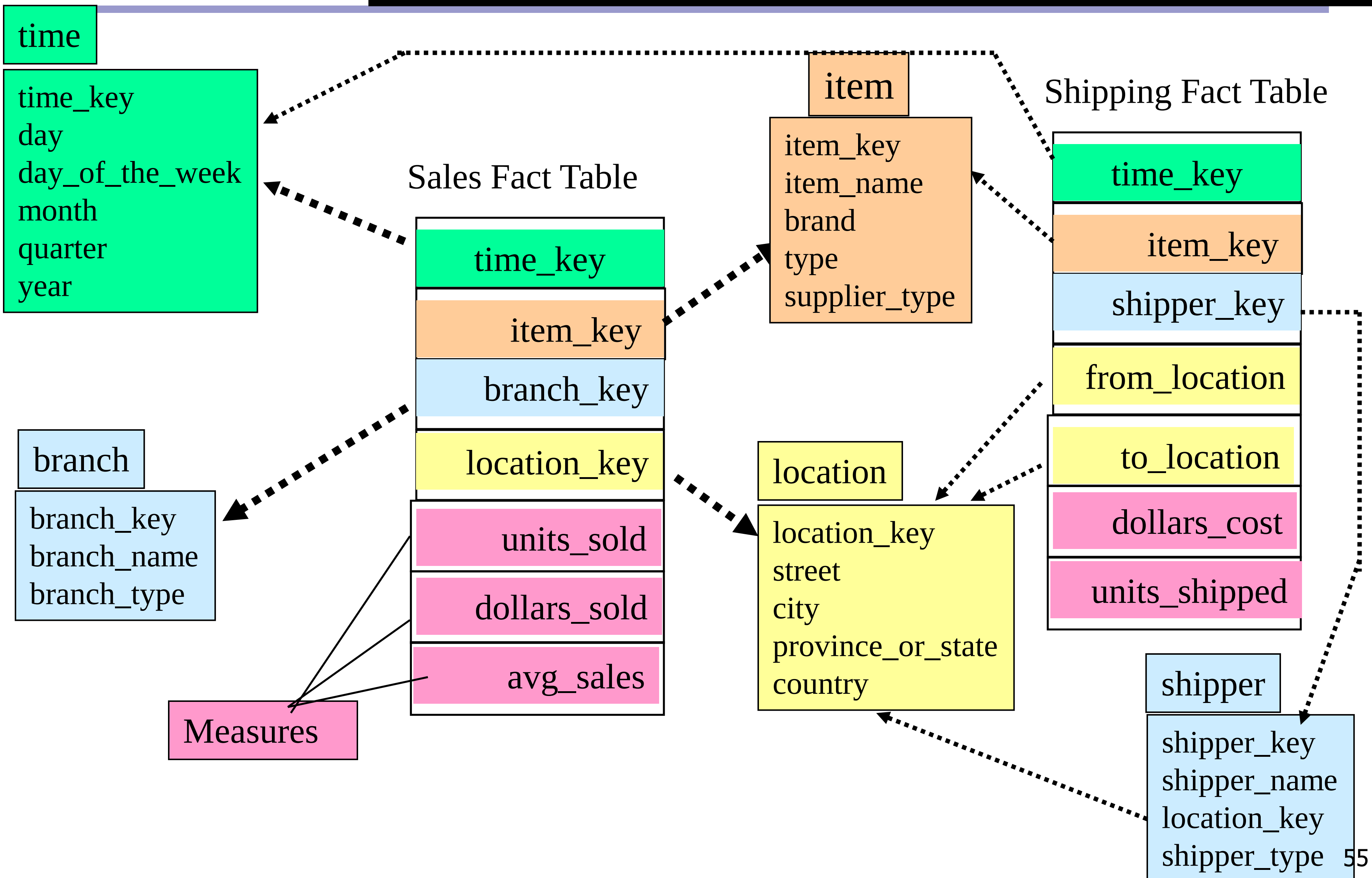
Snowflake Schema

- Advantages
  - Memory save
  - Normalized structures easier to update and maintain
- Disadvantages
  - Increases complexity
  - Browsing difficult
  - Degrades performance because of additional joins

Fact Constellation Schema

- Some applications are complex in nature and require multiple fact tables or sharing of dimension tables.
- This schema is one of the widely used data warehouse design methodology and is also called **Galaxy schema**

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

item

item_key
item_name
brand
type
supplier_type

Shipping Fact Table

time_key
item_key
shipper_key
from_location
to_location
dollars_cost
units_shipped

branch

branch_key
branch_name
branch_type

location

location_key
street
city
province_or_state
country

shipper

shipper_key
shipper_name
location_key
shipper_type

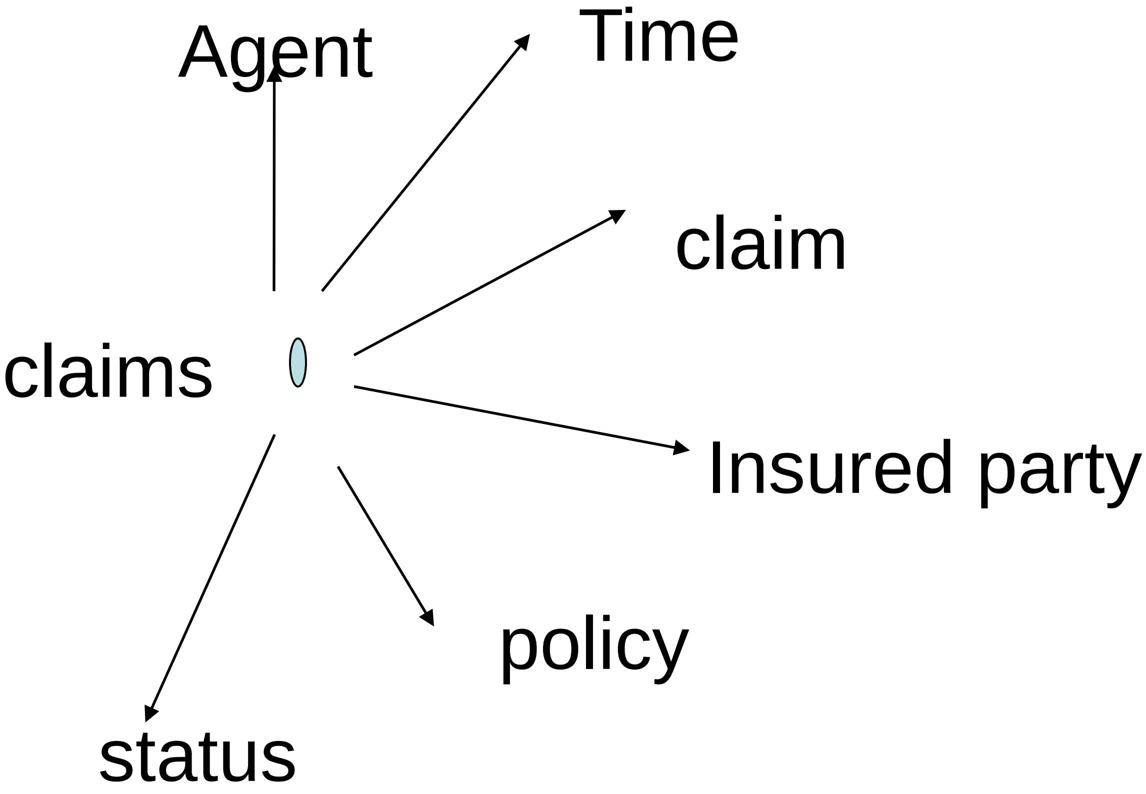
Pros & Cons of Fact Constellation Schema

- Pros:
  - It fact constellation schema offers more flexibility.
  - Multiple fact tables are explicitly assigned to dimension tables.
- Cons:
  - The structure is more complex and sophisticated.
  - It is difficult to maintain.
  - The number of aggregations are high in constellation.

Datawarehouse : business dimensions

- Identify dimensions for building a DW for claim analysis of insurance business.

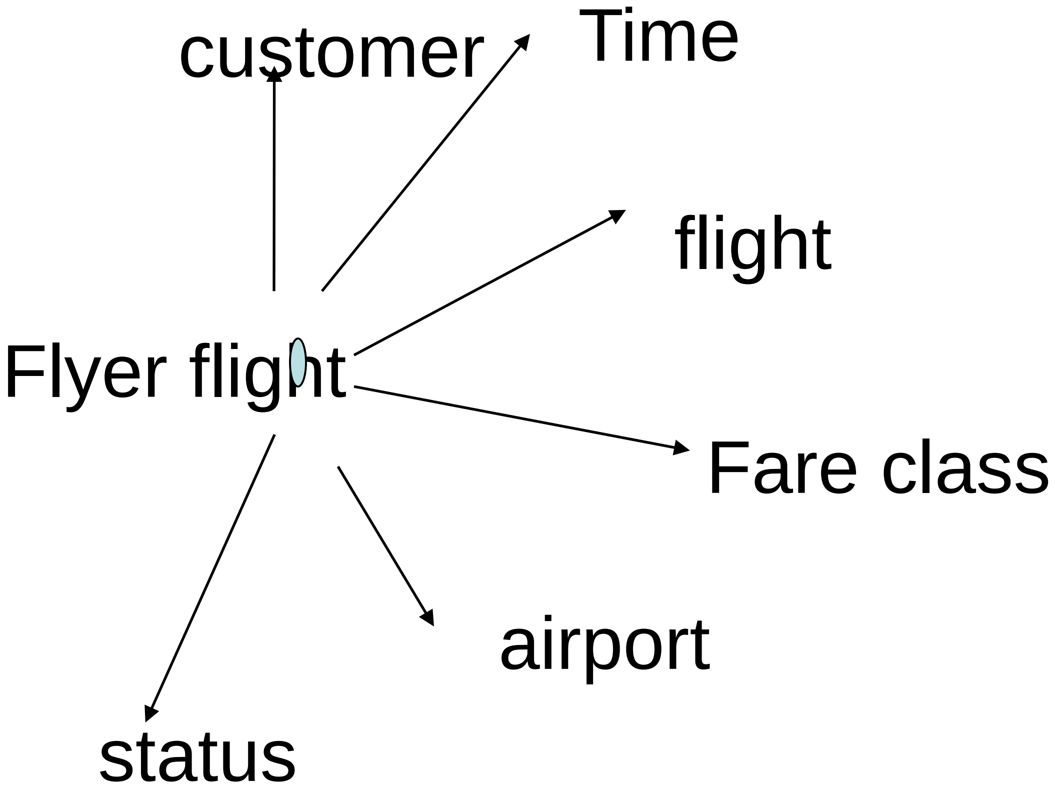
Example of business dimensions



Datawarehouse : business dimensions

- Identify dimensions for building a DW for flyer flight analysis of Airline Company

Example of business dimensions



customer

Time

flight

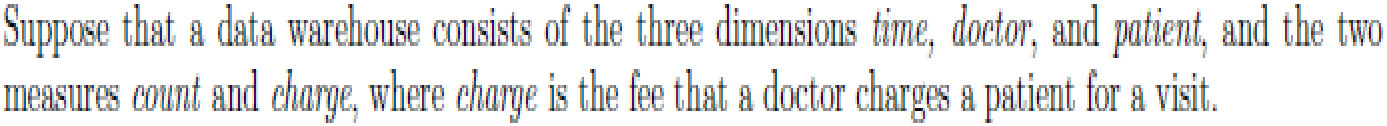
Flyer flight

Fare class

airport

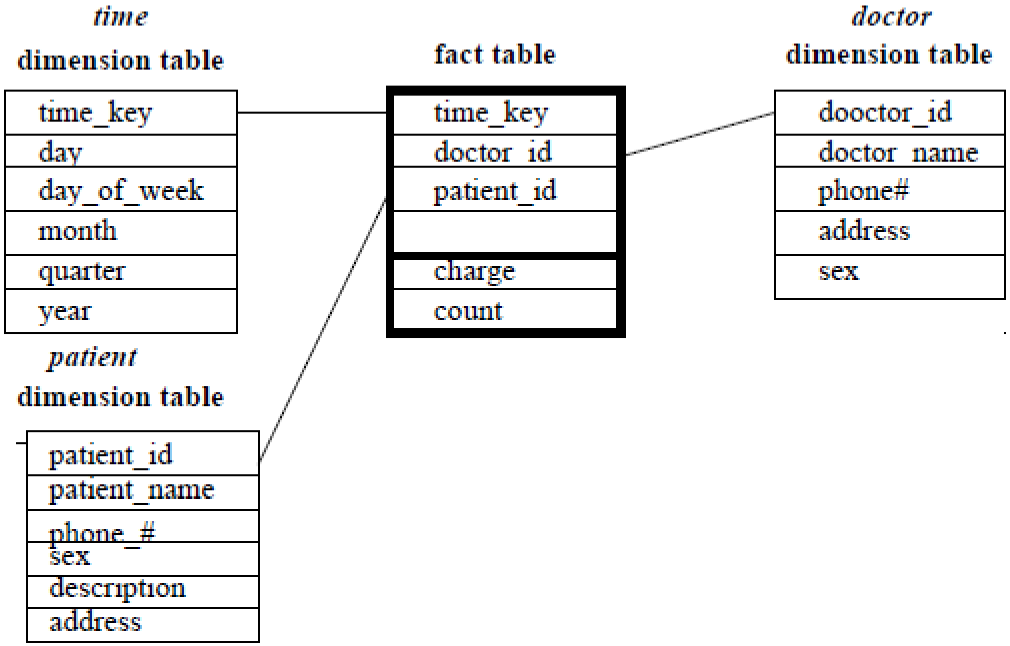
status

Exercise



Suppose that a data warehouse consists of the three dimensions *time*, *doctor*, and *patient*, and the two measures *count* and *charge*, where *charge* is the fee that a doctor charges a patient for a visit.

STAR Schema



time
dimension table

- time_key
- day
- day_of_week
- month
- quarter
- year

fact table

- time_key
- doctor id
- patient_id
- charge
- count

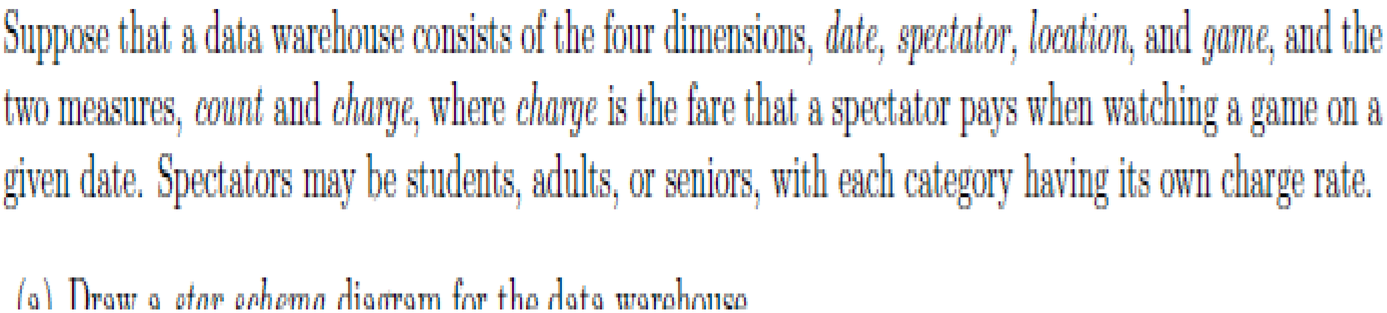
doctor
dimension table

- dooctor_id
- doctor name
- phone#
- address
- sex

patient
dimension table

- patient_id
- patient_name
- phone_#
- sex
- description
- address

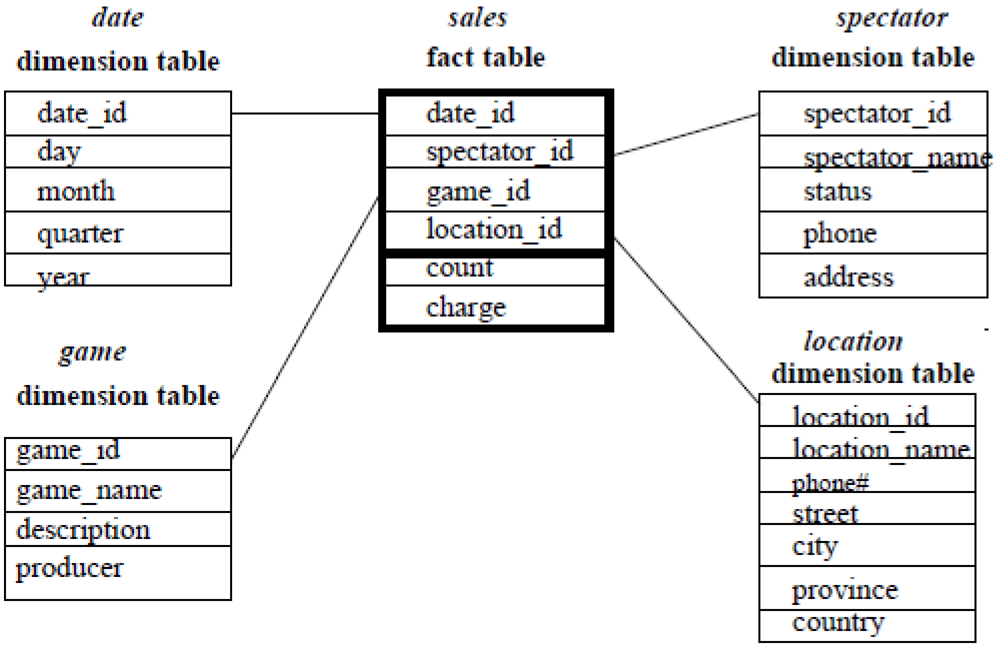
Exercise



Suppose that a data warehouse consists of the four dimensions, *date*, *spectator*, *location*, and *game*, and the two measures, *count* and *charge*, where *charge* is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own charge rate.

(a) Draw a *star schema* diagram for the data warehouse.

STAR Schema



date
dimension table

- date_id
- day
- month
- quarter
- year

sales
fact table

- date_id
- spectator_id
- game_id
- location_id
- count
- charge

spectator
dimension table

- spectator_id
- spectator_name
- status
- phone
- address

game
dimension table

- game_id
- game_name
- description
- producer

location
dimension table

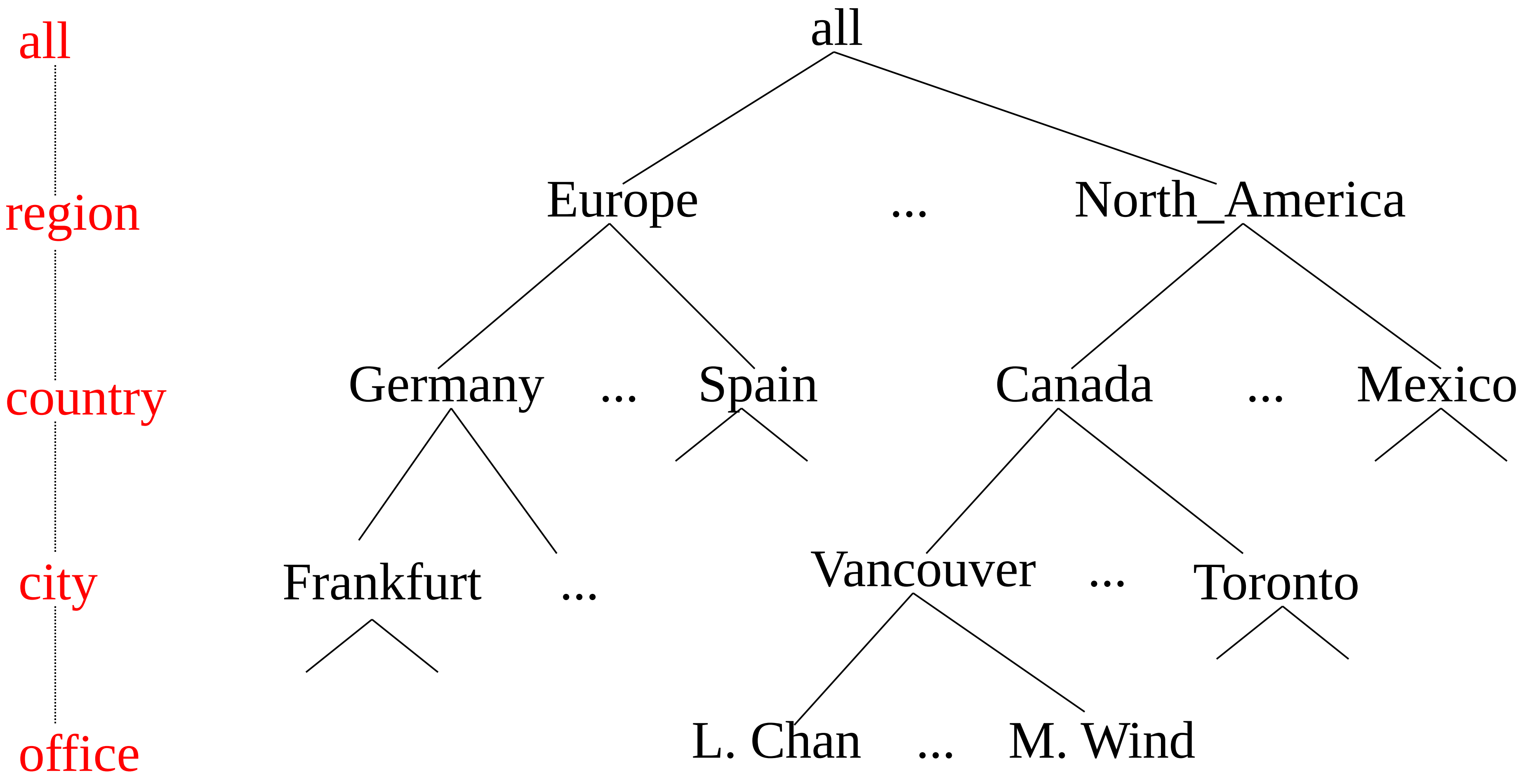
- location_id
- location_name
- phone#
- street
- city
- province
- country

Figure 3.3: A star schema for data warehouse of Exercise 3.5.

Measures of Data Cube: Three Categories

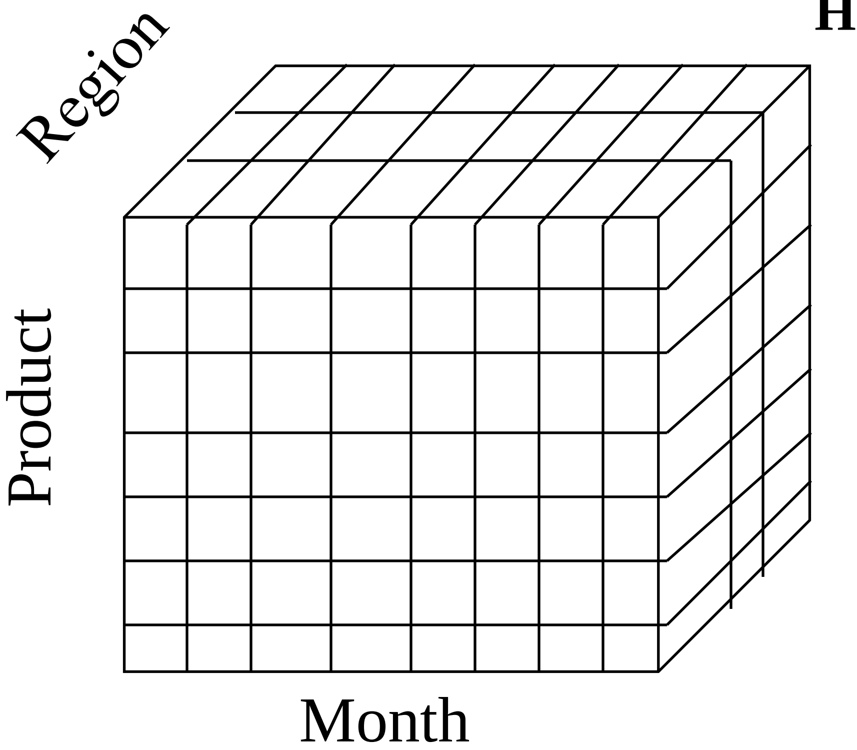
- Distributive:
- An aggregate function is distributive if it can be computed in a distributed manner
- if the result derived by applying the function to *n* aggregate values is the same as that derived by applying the function on all the data without partitioning
    - E.g., count(), sum(), min(), max()
- Algebraic: if it can be computed by an algebraic function with *M* arguments (where *M* is a bounded integer), each of which is obtained by applying a distributive aggregate function
    - E.g., avg()=sum()/count(), min_N(), standard_deviation()
- Holistic: if there is no constant bound on the storage size needed to describe a subaggregate.
- There does not exist an algebraic function with M arguments (where M is a constant) that characterizes the compu tation E.g., median(), mode(), rank()

A Concept Hierarchy: Dimension (location)

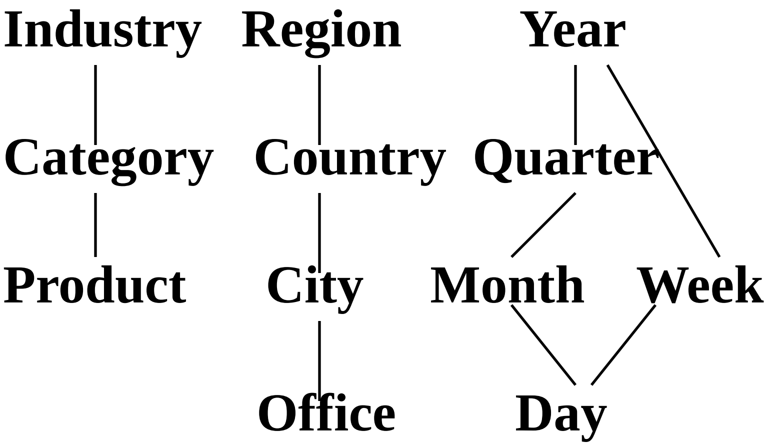


Multidimensional Data

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



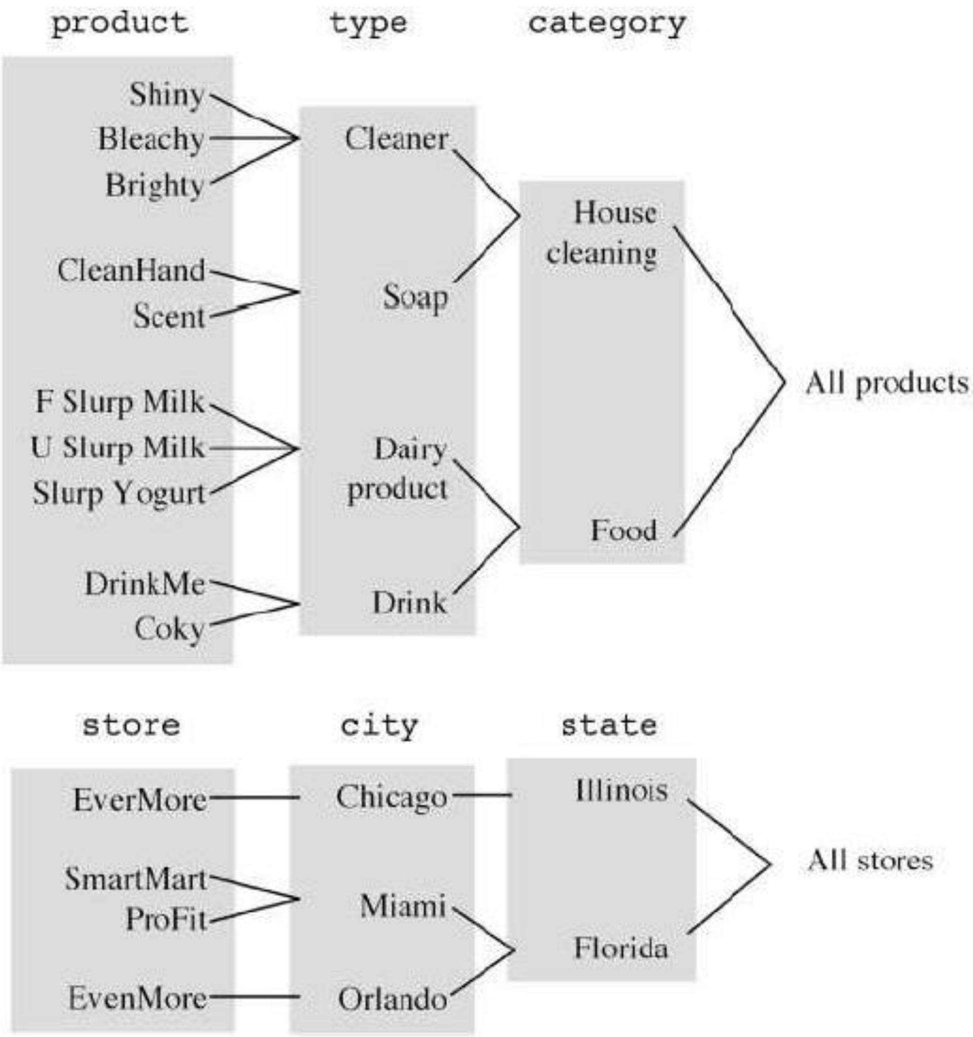
Dimensions: Product, Location, Time
Hierarchical summarization paths



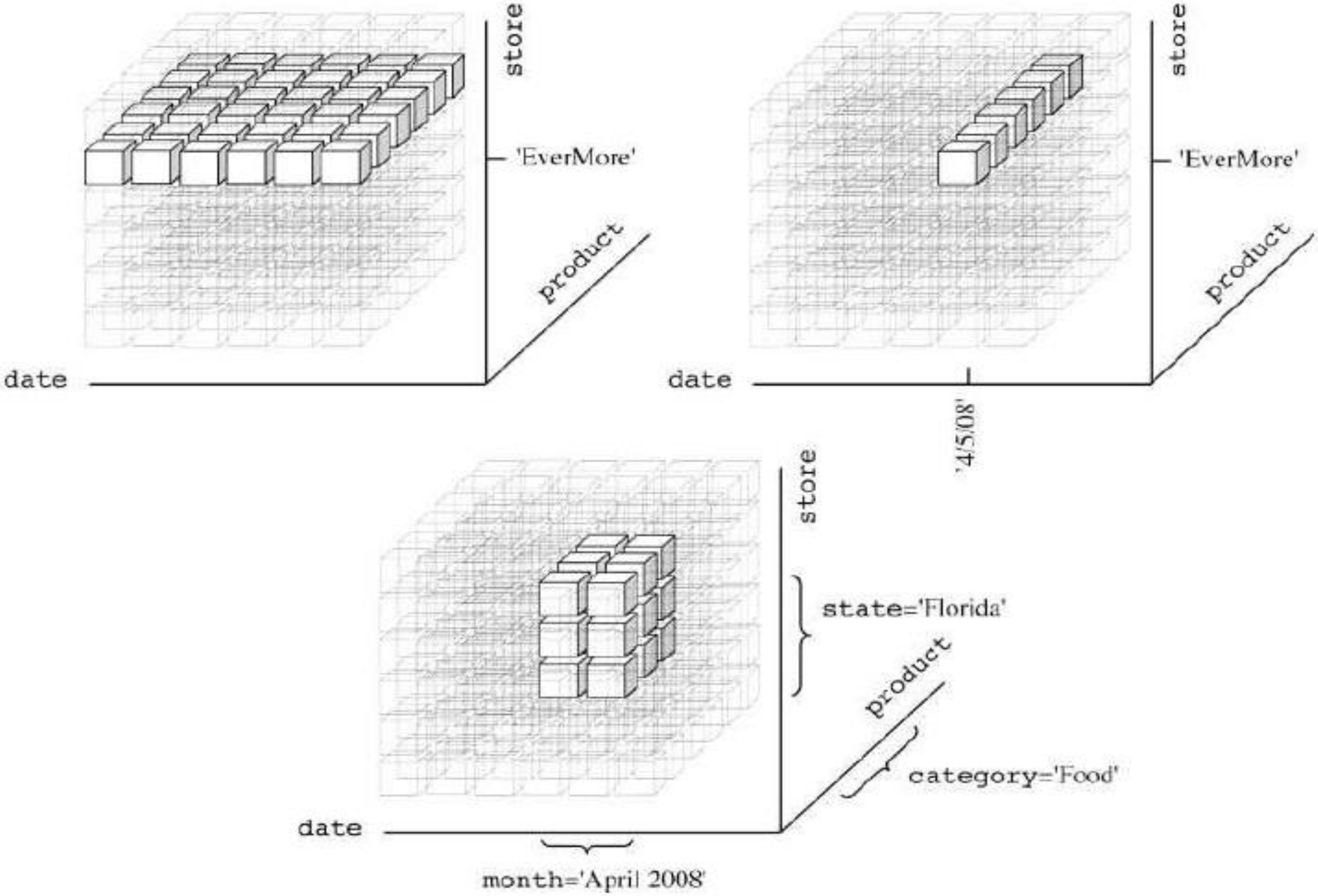
Typical OLAP Operations

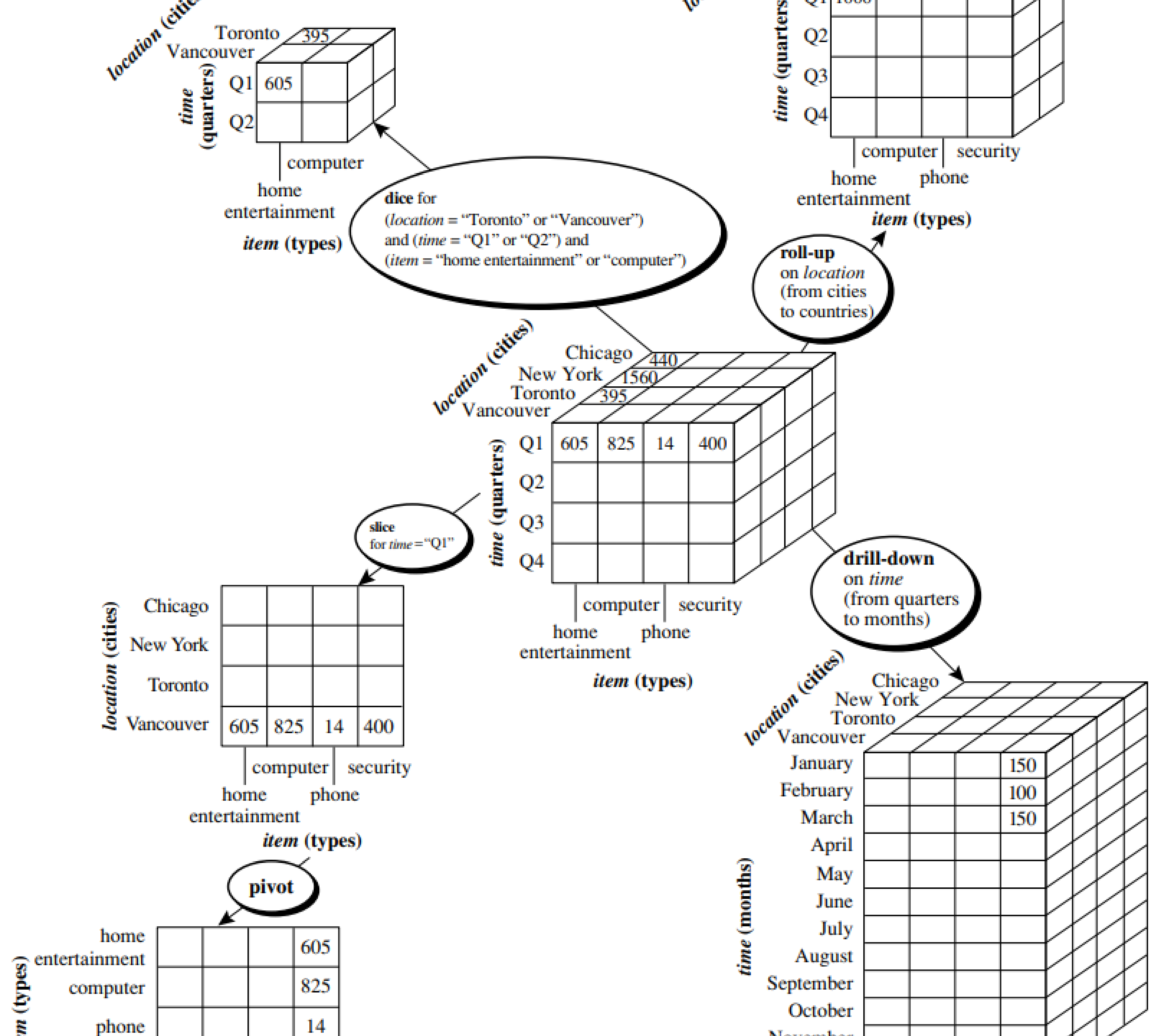
- Roll up (drill-up): summarize data
  - The roll-up operation (also called the drill-up operation by some vendors) performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction.
- Drill down (roll down): reverse of roll-up
  - Drill-down is the reverse of roll-up. It navigates from less detailed data to more detailed data.
  - *from higher level summary to lower level summary or detailed data, or introducing new dimensions*
- Slice and dice: *project and select*
- Pivot (rotate):
  - *reorient the cube, visualization, 3D to series of 2D planes*

Roll up (drill-up):



Slice and dice




location (cities)
Toronto
Vancouver
395
Q1
605
Q2
time (quarters)
computer
home entertainment
item (types)
dice for
(location = "Toronto" or "Vancouver")
and (time = "Q1" or "Q2") and
(item = "home entertainment" or "computer")
Q1
1000
Q2
Q3
Q4
time (quarters)
computer
security
home entertainment
phone
item (types)
roll-up
on location
(from cities
to countries)
location (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
Q1
605
825
14
400
Q2
Q3
Q4
time (quarters)
computer
security
home entertainment
phone
item (types)
slice
for time = "Q1"
Chicago
New York
Toronto
Vancouver
location (cities)
605
825
14
400
computer
security
home entertainment
phone
item (types)
pivot
home entertainment
computer
phone
605
825
14
item (types)
drill-down
on time
(from quarters
to months)
location (cities)
Chicago
New York
Toronto
Vancouver
January
February
March
April
May
June
July
August
September
October
November
150
100
150
time (months)

Question

Let the query to be processed be on {*brand, province_or_state*} with the condition "*year = 2004*", and there are 4 materialized cuboids available:

1) {*year, item_name, city*}

2) {*year, brand, country*}

3) {*year, brand, province_or_state*}

4) {*item_name, province_or_state*} where *year = 2004*

Which should be selected to process the query?

Design of Data Warehouse: A Business Analysis Framework

- Four views regarding the design of a data warehouse
  - Top-down view
    - allows selection of the relevant information necessary for the data warehouse
  - Data source view
    - exposes the information being captured, stored, and managed by operational systems
  - Data warehouse view
    - consists of fact tables and dimension tables
  - Business query view
    - sees the perspectives of data in the warehouse from the view of end-user

Data Warehouse Design Process

- Top-down, bottom-up approaches or a combination of both
  - Top-down: Starts with overall design and planning (mature),Useful where the business problems that must be solved are clear and well understood.
  - Bottom-up: Starts with experiments and prototypes (rapid), useful in the early stage of business modeling and technology development
  - Combined
- From software engineering point of view
  - Waterfall: structured and systematic analysis at each step before proceeding to the next
  - Spiral: The spiral method involves the rapid generation of increasingly functional systems, with short intervals between successive releases. This is considered a good choice for data warehouse development, especially for data marts, because the turnaround time is short, modifications can be done quickly, and new designs and technologies can be adapted in a timely manner.

Top Down Approach

- **Advantages Of Top Down Design :**
  - It is easier to maintain Top Down Design
  - Provides consistent dimensional views of data across data marts, as all data marts are loaded from the data warehouse.
  - This approach is robust against business changes. Creating a new data mart from the data warehouse is very easy.
  - Initial cost is high but subsequent project development cost is lower
- **Disadvantages Of Top Down Design :**
  - It represents a very large project and the cost of implementing the project is significant.
  - It is time consuming and more time required for initial set up
  - Highly skilled people required for set up

Bottom-Up approach

- **Advantages Of Bottom Up Design :**
  - This model contains consistent data marts and these data marts can be delivered quickly.
  - The data marts are created first to provide reporting capability
  - It is easier to extend the data warehouse as it can easily accommodate new business units. It is just creating new data marts and then integrating with other data marts.
  - This Approach take less time. Initial set up is very quickly
- **Disadvantage of Bottom Up Design :**
  - Initial cost is low but each subsequent phase will cost same
  - The positions of the data warehouse and the data marts are reversed in the bottom-up approach design.
  - It is difficult to maintain and often redundant and subject to revisions

DW Design Process

1. Choose a business process to model (e.g., orders, invoices, shipments, inventory, account administration, sales, or the general ledger). If the business process is organizational and involves multiple complex object collections, a data warehouse model should be followed. However, if the process is departmental and focuses on the analysis of one kind of business process, a data mart model should be chosen.

2. Choose the business process grain, which is the fundamental, atomic level of data to be represented in the fact table for this process (e.g., individual transactions, individual daily snapshots, and so on).

3. Choose the dimensions that will apply to each fact table record. Typical dimensions are time, item, customer, supplier, warehouse, transaction type, and status.

4. Choose the measures that will populate each fact table record. Typical measures are numeric additive quantities like dollars sold and units sold.

Data Warehouse Usage

- Three kinds of data warehouse applications
  - Information processing
    - supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts and graphs
  - Analytical processing
    - multidimensional analysis of data warehouse data
    - supports basic OLAP operations, slice-dice, drilling, pivoting
  - Data mining
    - knowledge discovery from hidden patterns
    - supports associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools

OLAP Server Architectures

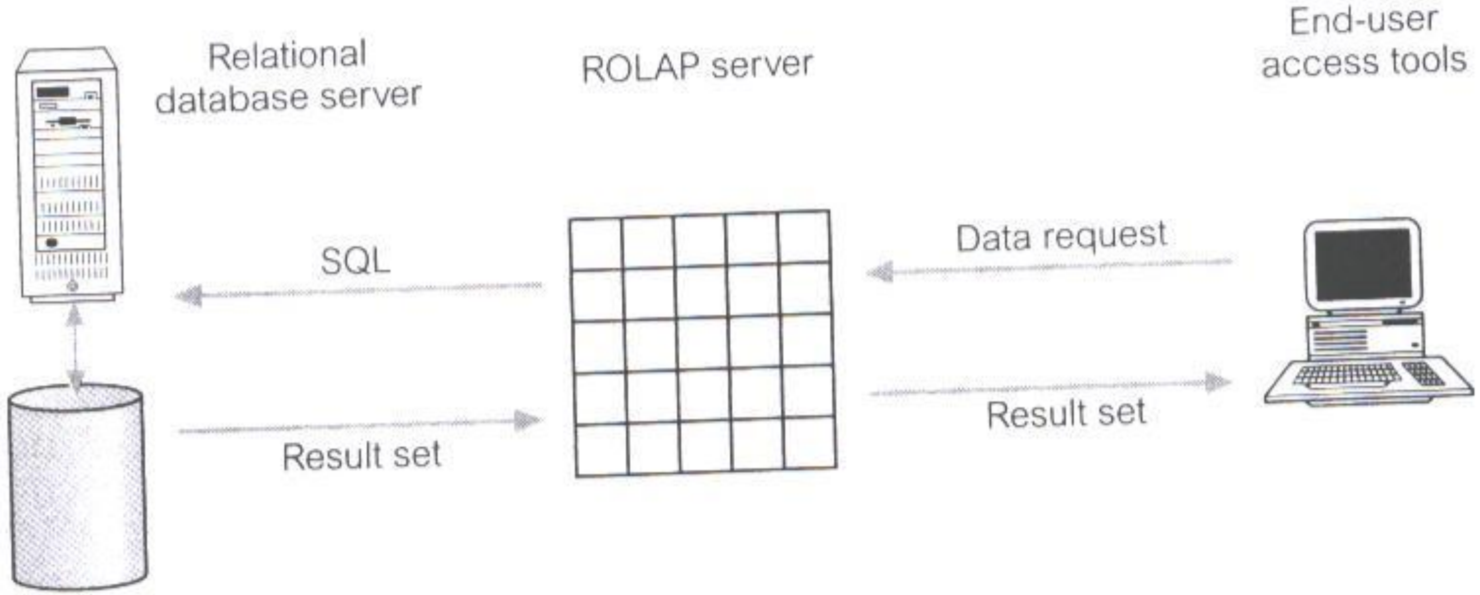
- Relational OLAP (ROLAP)
- Multidimensional OLAP (MOLAP)
- Hybrid OLAP (HOLAP) (e.g., Microsoft SQLServer)
- Specialized SQL servers (e.g., Redbricks)
  - Specialized support for SQL queries over star/snowflake schemas

Relational OLAP (ROLAP)

- These are the intermediate servers that stand in between a relational back-end server and client front-end tools. They use a relational or extended-relational DBMS to store and manage warehouse data.

- ROLAP technology tends to have greater scalability than MOLAP technology

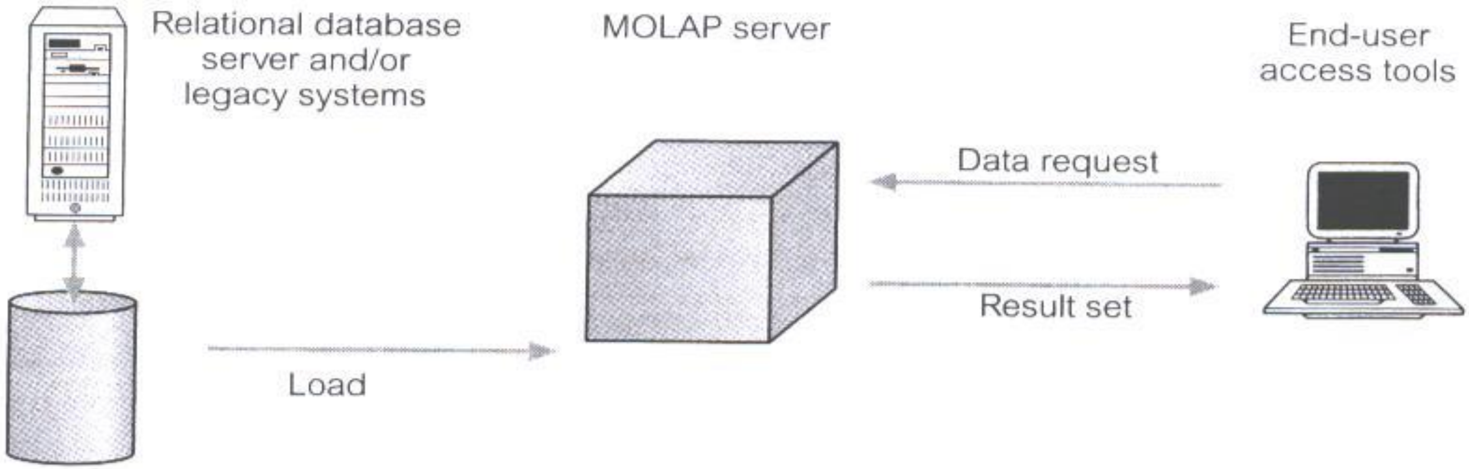
Relational OLAP (ROLAP)



Multidimensional OLAP (MOLAP)

- MOLAP tools use specialized data structures and multi-dimensional database management systems (MDDBMS) to organize, navigate, and analyze data.
- Sparse array-based multidimensional storage engine that minimize the disk space requirements through sparse data management.
- Fast indexing to pre-computed summarized data

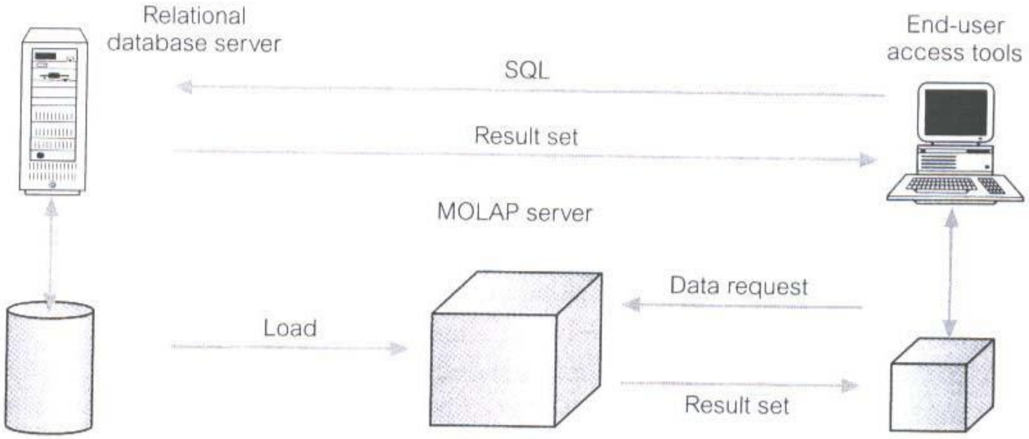
Multidimensional OLAP (MOLAP)



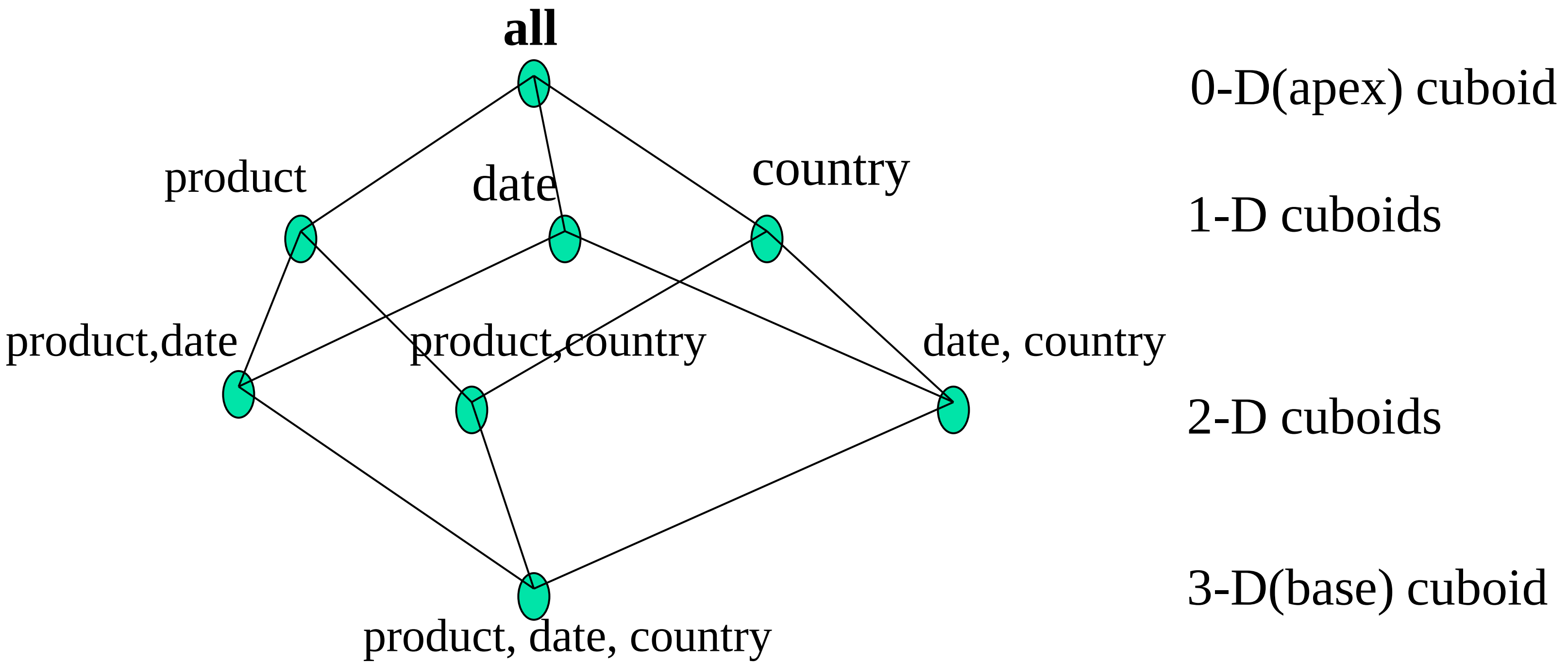
Hybrid OLAP (HOLAP)

- The hybrid OLAP approach combines ROLAP and MOLAP technology, benefiting from the greater scalability of ROLAP and the faster computation of MOLAP.

Hybrid OLAP (HOLAP)



Cuboids Revisited



The **base cuboid contains all three dimensions**,. It can return the total sales for any combination of the three dimensions.

The **apex cuboid, or 0-D cuboid, refers to the case where the group-by is empty**. It contains the total sum of all sales.

The base cuboid is the least generalized (most specific) of the cuboids.

The apex cuboid is the most generalized (least specific) of the cuboids, and is often denoted as all.

Efficient Data Cube Computation

- Data cube can be viewed as a lattice of cuboids
  - The bottom-most cuboid is the base cuboid
  - The top-most cuboid (apex) contains only one cell
  - How many cuboids in an n-dimensional cube with L levels?

$$T = \prod_{i=1}^{n} (L_i + 1)$$

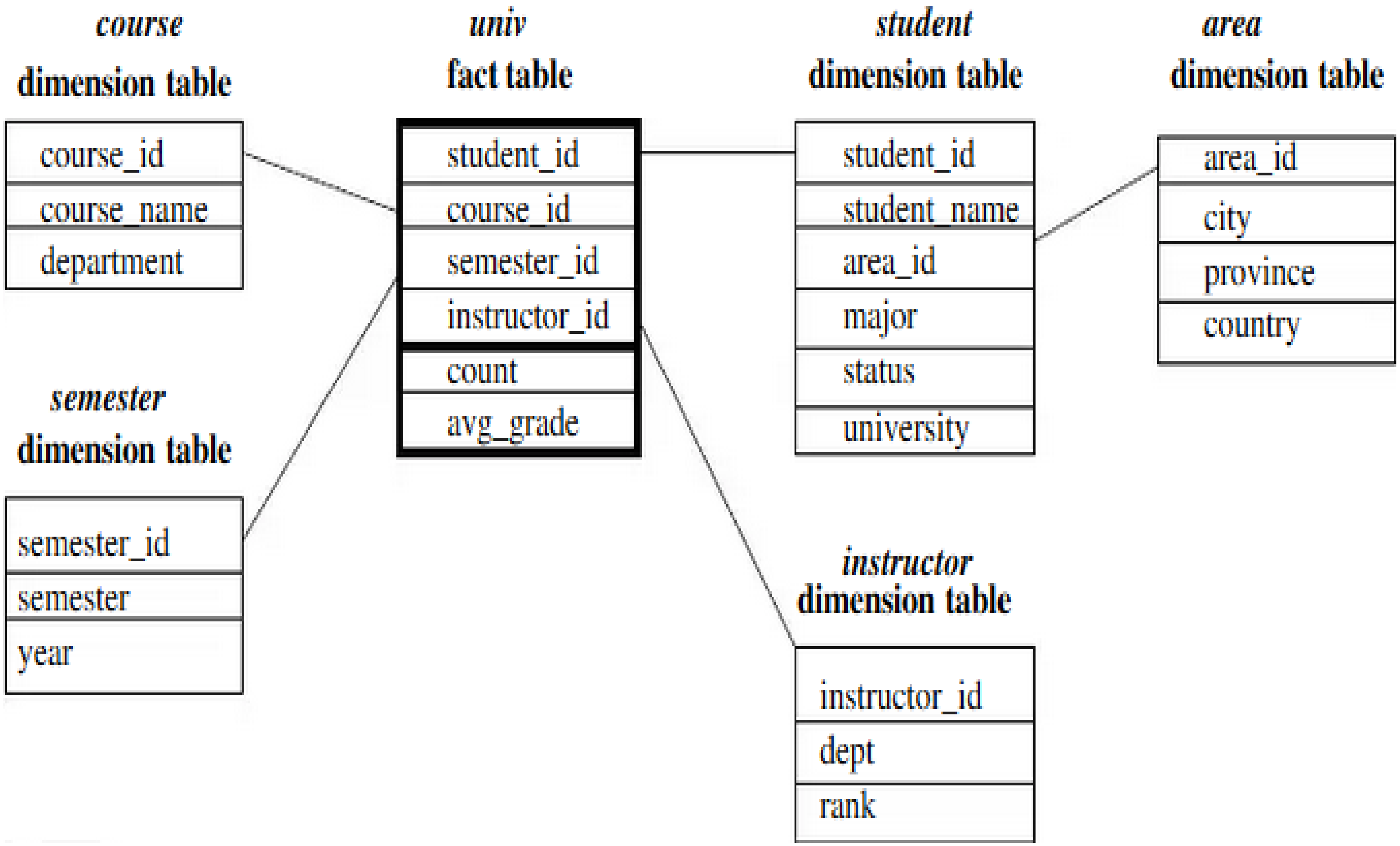
  - Ex. If the cube has 10 dimensions and each dimension has five levels (including all), the total number of cuboids that can be generated is $5^{10} \approx 9.8 \times 10^{6}$
- Materialization of data cube : Precomputation of Cuboids
  - Materialize every (cuboid) (full materialization), none (no materialization), or some (partial materialization)
  - Selection of which cuboids to materialize
    - Based on size, sharing, access frequency, etc.

Case Study

- **Suppose that a data warehouse for Big University consists of the following four dimensions: student, course, semester, and instructor , and two measures count and avg grade . When at the lowest conceptual level (e.g.,for a given student, course, semester, and instructor combination), the avg grade measure stores the actualcourse grade of the student. At higher conceptual levels, avg grade stores the average grade for the givencombination.**

  (a) Draw a snowflake schema diagram for the data warehouse.

Snowflake Schema diagram



Question

- Starting with the base cuboid [student,course,semester,instructor], what specific OLAP operations (e.g., roll-up from semester to year) should one perform in order to list the average grade of CS-branch courses for each Big University student.

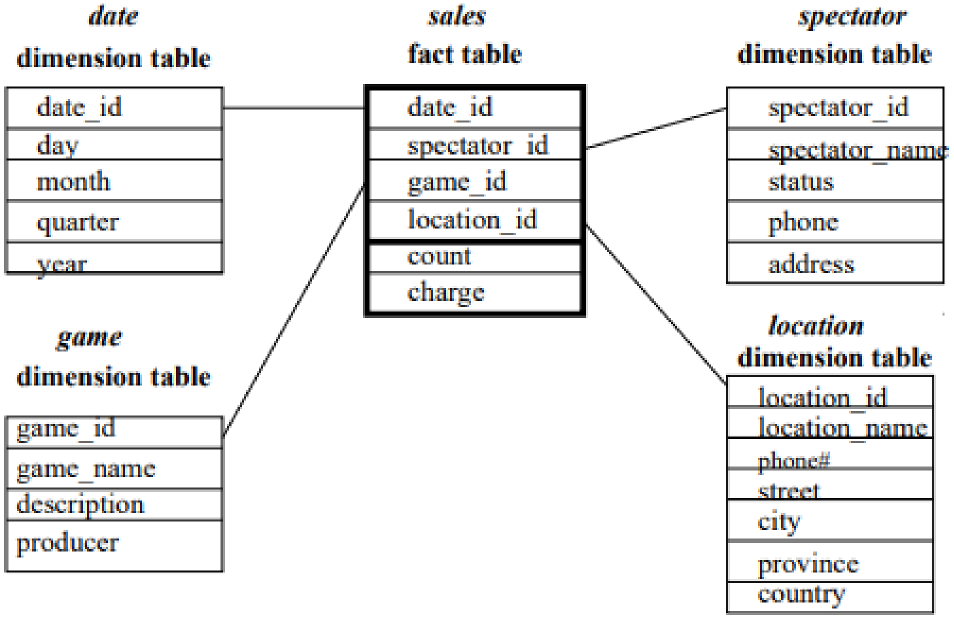
Answer

- Roll-up on course from course id to department .
- Roll-up on student from student id to university.
- Dice on course, student with department="CS" and university = "Big University".

- **(c) If each dimension has five levels (including all), such as student < major < status < university < all,how many cuboids will this cube contain (including the base and apex cuboids)**
- This cube will contain 5^4= 625 cuboids

Questions

- Starting with the base cuboid **[date, spectator, location, game]**, what specific OLAP operations should one perform in order to list the total charge paid by student spectators at GM Place in 2004?



date dimension table
date_id
day
month
quarter
year

sales fact table
date_id
spectator_id
game_id
location_id
count
charge

spectator dimension table
spectator_id
spectator_name
status
phone
address

game dimension table
game_id
game_name
description
producer

location dimension table
location_id
location_name
phone#
street
city
province
country

Solution

- Roll-up on date from date id to year.
- Roll-up on game from game id to all.
- Roll-up on location from location id to location name.
- Roll-up on spectator from spectator id to status.
- Dice with status="students", location name="GM Place", and year=2004.

- Thank You