University of Colorado Boulder

CSCI 4502/5502

Data Mining

Fall 2019
Lecture 05 (Sep 10)

Reminders

- Homework 1
  - posted at moodle, CSCI 4502 vs. 5502
  - due at 9:30am, Th Sep 12
  - SUBMIT your attempt before deadline
- Office hours (in-person, Zoom, or by appt)
  - instructor: no regular office this week
  - TA: Tu 4-5pm, ECCR 1B10

Announcements

- Guest lecture on Thursday, Sep 12
  - Prof. Claire Monteleoni on Machine Learning and Environmental Informatics
- Homework 2
  - will be posted at moodle this Thursday
  - due at 9:30am, Th Sep 19
  - SUBMIT your attempt before deadline

Review

- Chapter 3: Data Preprocessing
  - Data preprocessing overview
  - Data cleaning
  - Data integration
  - Data reduction
  - Data transformation and discretization

Data Transformation

- Smoothing: remove noise from data
- Aggregation: summarization
  - e.g., daily sales => monthly, annual sales
- Generalization: concept hierarchy climbing
  - e.g., street => city => state
- Normalization: scale to fall within a range
- Attribute/feature construction: new attributes constructed from existing ones

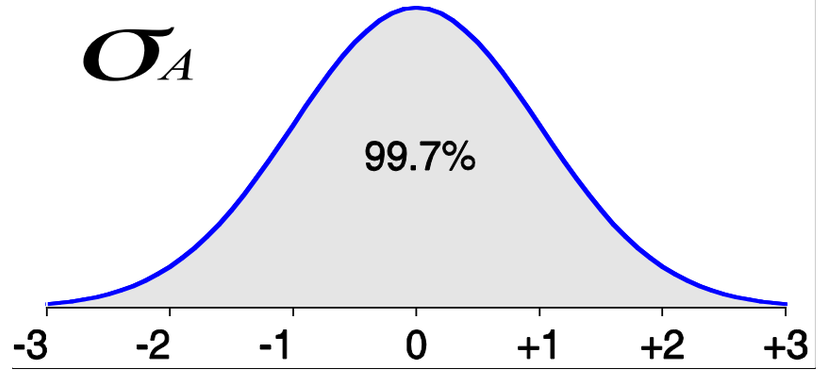
Normalization (1)

- Min-max normalization

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - e.g., income range [$12,000, $98000] normalize to [0.0, 1.0], then $73,600 is mapped to $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$
- Z-score normalization
  - e.g., mean = 54,000
  - stdev = 16,000
  - then $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$

$$v' = \frac{v - \mu_A}{\sigma_A}$$



Normalization (2)

- Normalization by decimal scaling

$$v' = \frac{v}{10^j}$$

  - where j is the smallest integer s.t. Max(|v'|)<1
    - e.g., range [-986, 917]
      - j = 3, divide by 1000
      - -986 => -0.986
      - 917 => 0.917

Discretization

- Three types of attributes
  - nominal: unordered set (e.g., profession)
  - ordinal: ordered set (e.g., military rank)
  - continuous: e.g., integer or real numbers
- Discretization
  - divide continuous range into intervals
  - interval labels used to replace data values
  - supervised vs. unsupervised, split vs. merge

Discretization Methods

- Binning: split, unsupervised
- Histogram analysis: split, unsupervised
- Clustering analysis: split/merge, unsupervised
- Entropy-based discretization: split, supervised
- Interval merging by X^2 analysis:
  - merge, supervised
- Intuitive partitioning:
  - split, unsupervised

Entropy-Based Discretization

- Partition D into D_1 and D_2 at boundary A

$$Info_A(D) = \frac{|D_1|}{|D|} Entropy(D_1) + \frac{|D_2|}{|D|} Entropy(D_2)$$

$$Entropy(D_1) = -\sum_{i=1}^{m} p_i log_2(p_i)$$

- Pick boundary A with minimum $Info_A(D)$
  - "purer" distribution has lower entropy
- Apply recursively to each partition
- Top-down split, supervised (uses class info)

Interval Merge by X^2 Analysis

- Bottom-up merge, supervised
- Merge the best neighboring intervals
  - intervals w/ most similar class distributions
- ChiMerge
  - merge adjacent intervals w/ min X^2 value
    - i.e., class is independent of interval
  - stopping criterion
    - significance, #intervals, inconsistency, ...

Concept Hierarchy Generation

- Categorical data
- Partial/total ordering of attributes
  - street < city < state < country
- Automatic concept hierarchy generation
  - fewer distinct values => higher level
  - e.g., street, city, state, country
  - exceptions
    - e.g., weekday, month, quarter, year

Chapter 4:
Data Warehouse and OLAP

Chapter 5:
Data Cube Technology

What is A Data Warehouse?

- A decision support database that is maintained separately from an 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

Subject-Oriented

- Organized around major subjects
  - e.g., customers, product, sales
- Focus on the modeling & analysis of data for decision making, not on daily operations or transaction processing
- Provide a simple & concise view around particular subject issues by excluding data that are not useful in the decision support process

Integrated

- Integrate multiple, heterogeneous data sources
  - relational database, flat files, on-line transaction records
- Data cleaning and data integration techniques applied
  - ensure consistency in naming, encoding, attribute measures, etc.
  - e.g., hotel price: currency, tax, breakfast, ...

Time-Variant

- Significantly longer time span
  - operational database: current data
  - data warehouse: historical perspective
    - e.g., past 5-10 years
- Every key structure in a data warehouse
  - contains time info, explicitly or implicitly
  - key of operational data may not contain time info

Nonvolatile

- A physically separate store of data transformed from operational environments
- No operational update of data
  - no transaction processing, recovery, concurrency control
- Only two operations in data accessing
  - initial loading of data
  - access of data

Data Warehouse vs.

- Operational DBMS
- OLTP (on-line transaction processing)
  - major task of traditional relational DBMS
  - day-to-day operations
- OLAP (on-line analytical processing)
  - major task of data warehouse system
  - data analysis and decision making

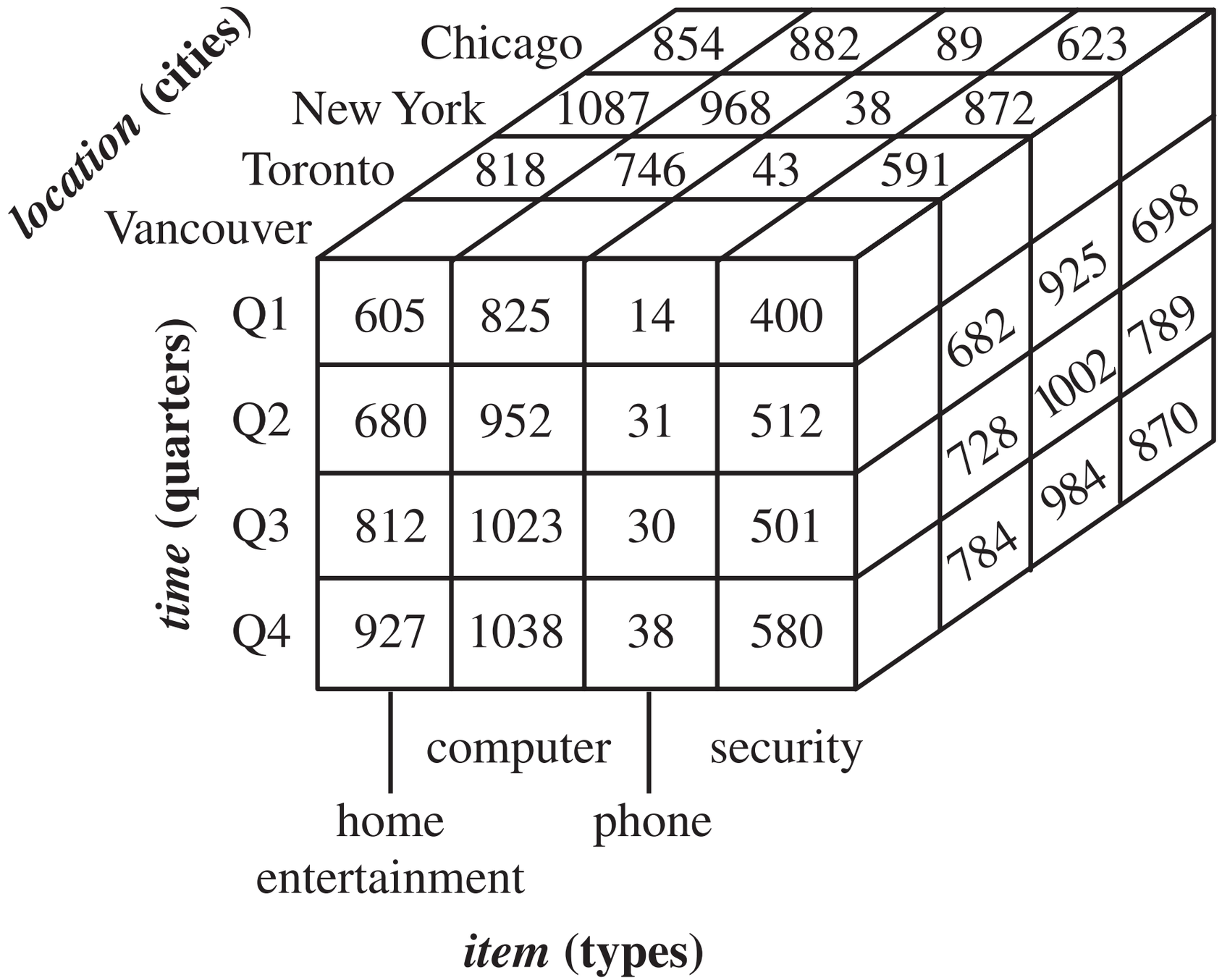
OLTP vs. OLAP

	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

What is A Data Cube?

- Data warehouses and OLAP are based on
  - a multi-dimensional data model
- Data cube
  - allow data to be modeled and viewed in multiple dimensions (e.g., sales)
  - dimensions: e.g., time, item, branch, location
  - facts: numerical measures
    - e.g., items_sold, dollars_sold

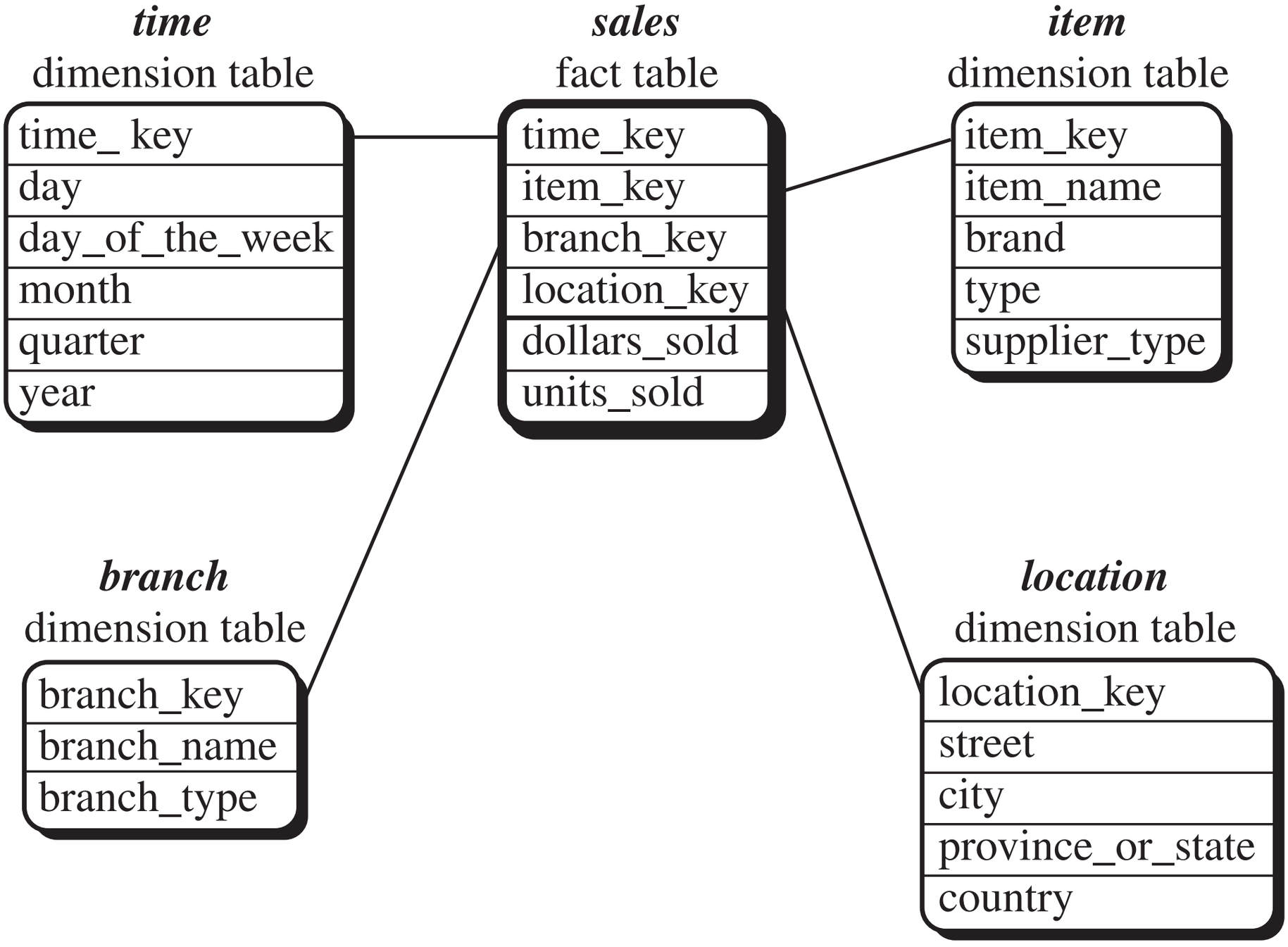
Data Cube Example



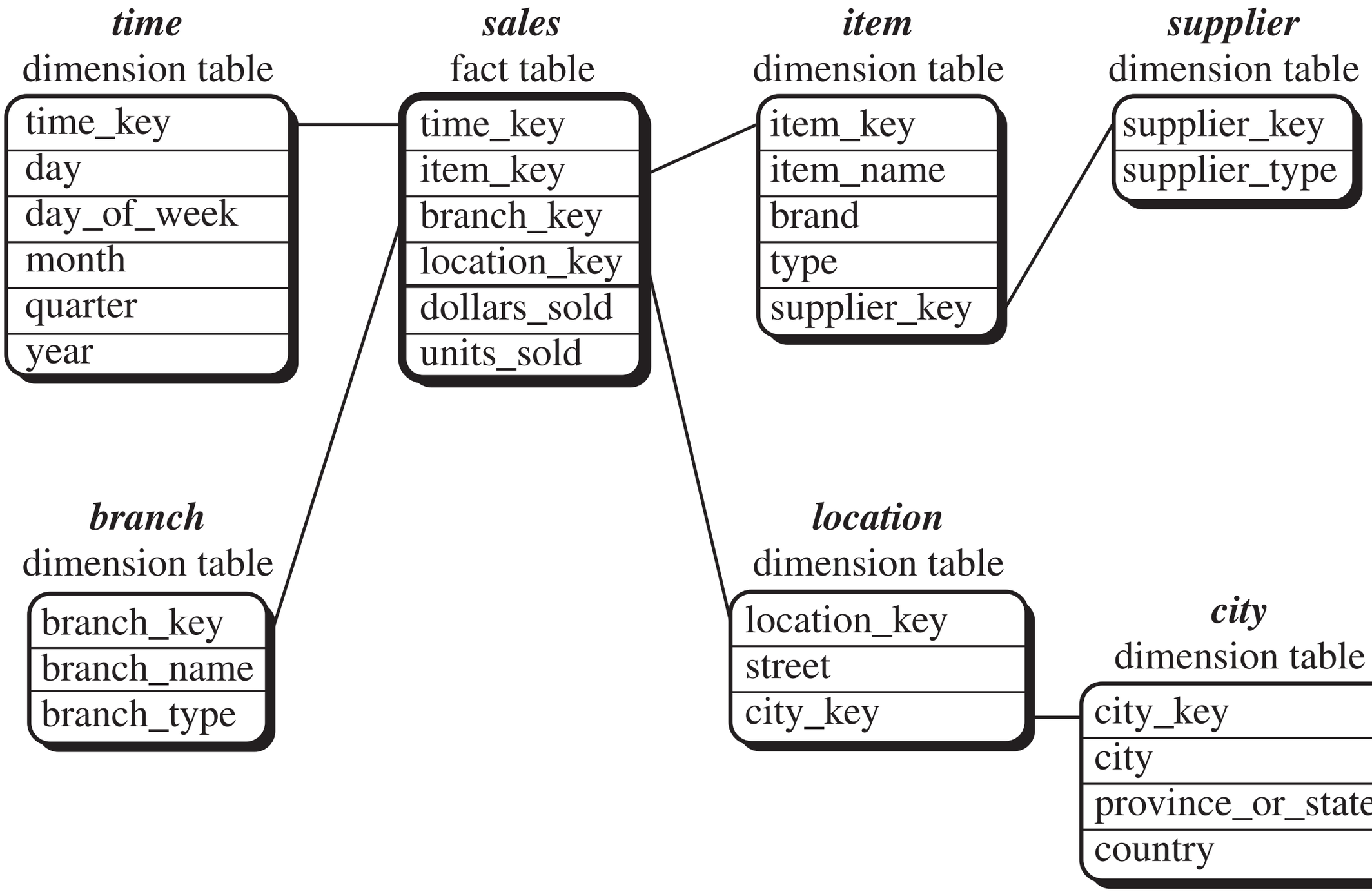
Conceptual Modeling

- Modeling data warehouses
  - dimensions & facts
- Star schema
  - a fact table, a set of dimension tables
- Snowflake schema
  - a fact table, a hierarchy of dimension tables
- Fact constellations
  - multiple fact tables share dimension tables

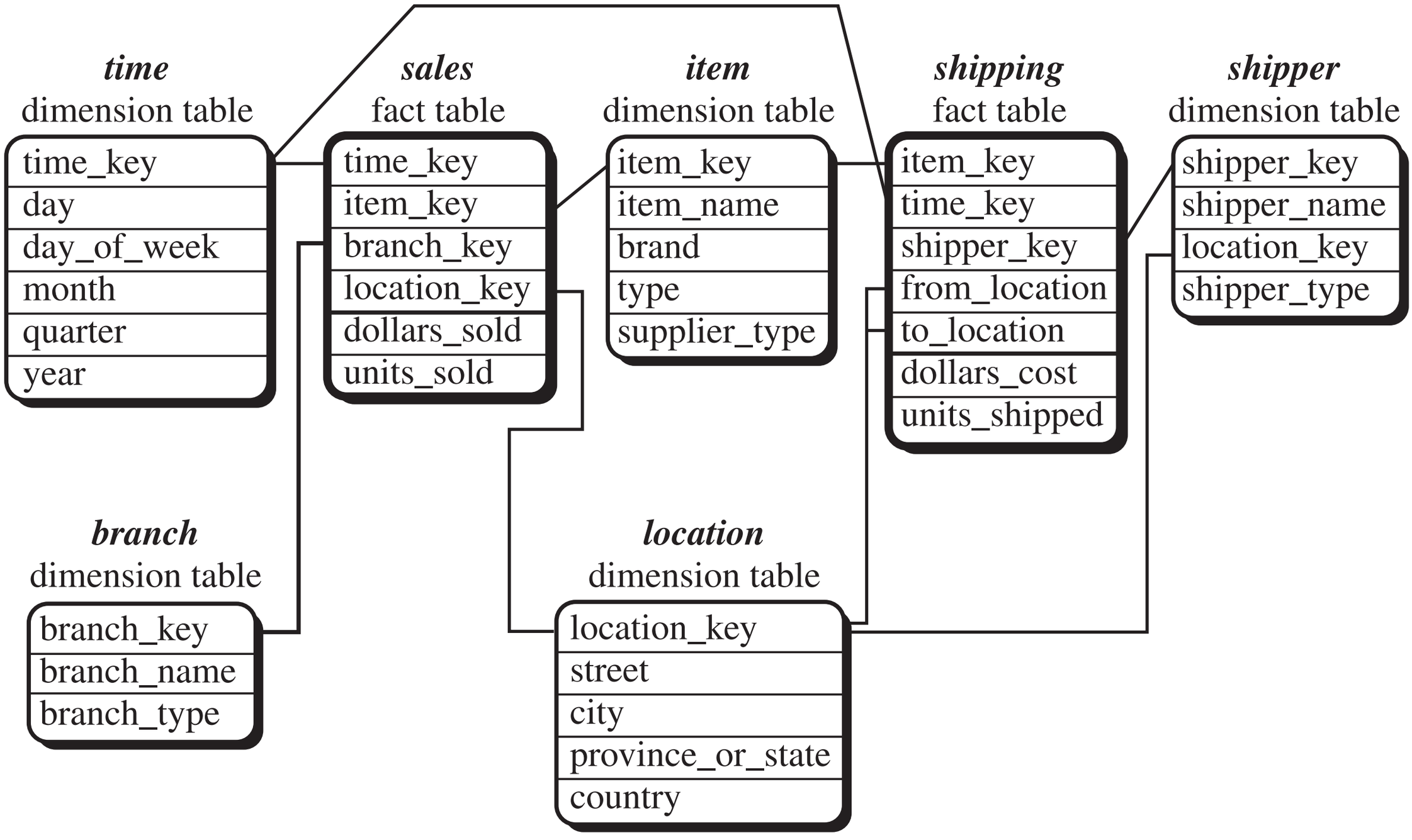
Example of Star Schema



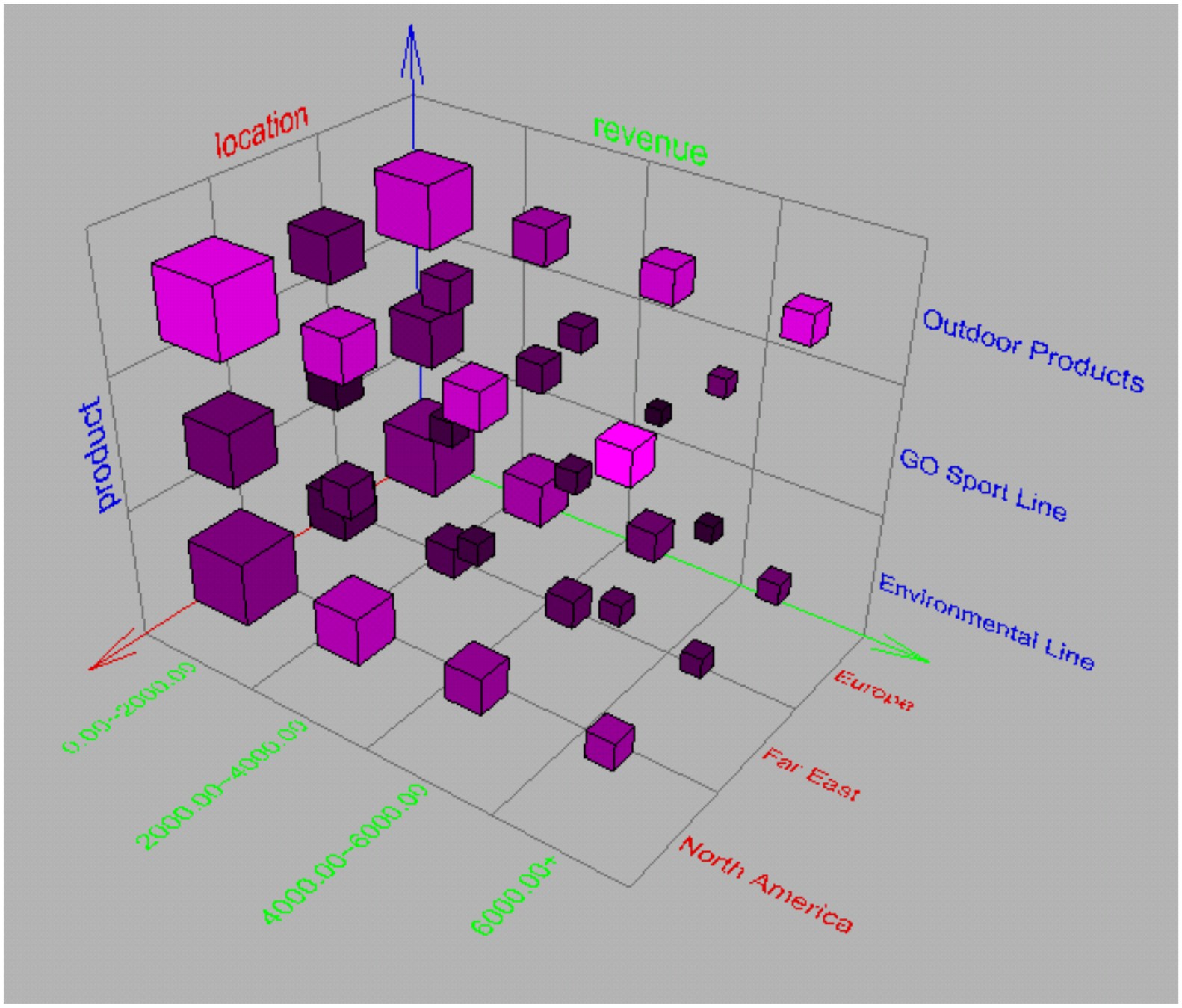
Example of Snowflake



Example of Fact

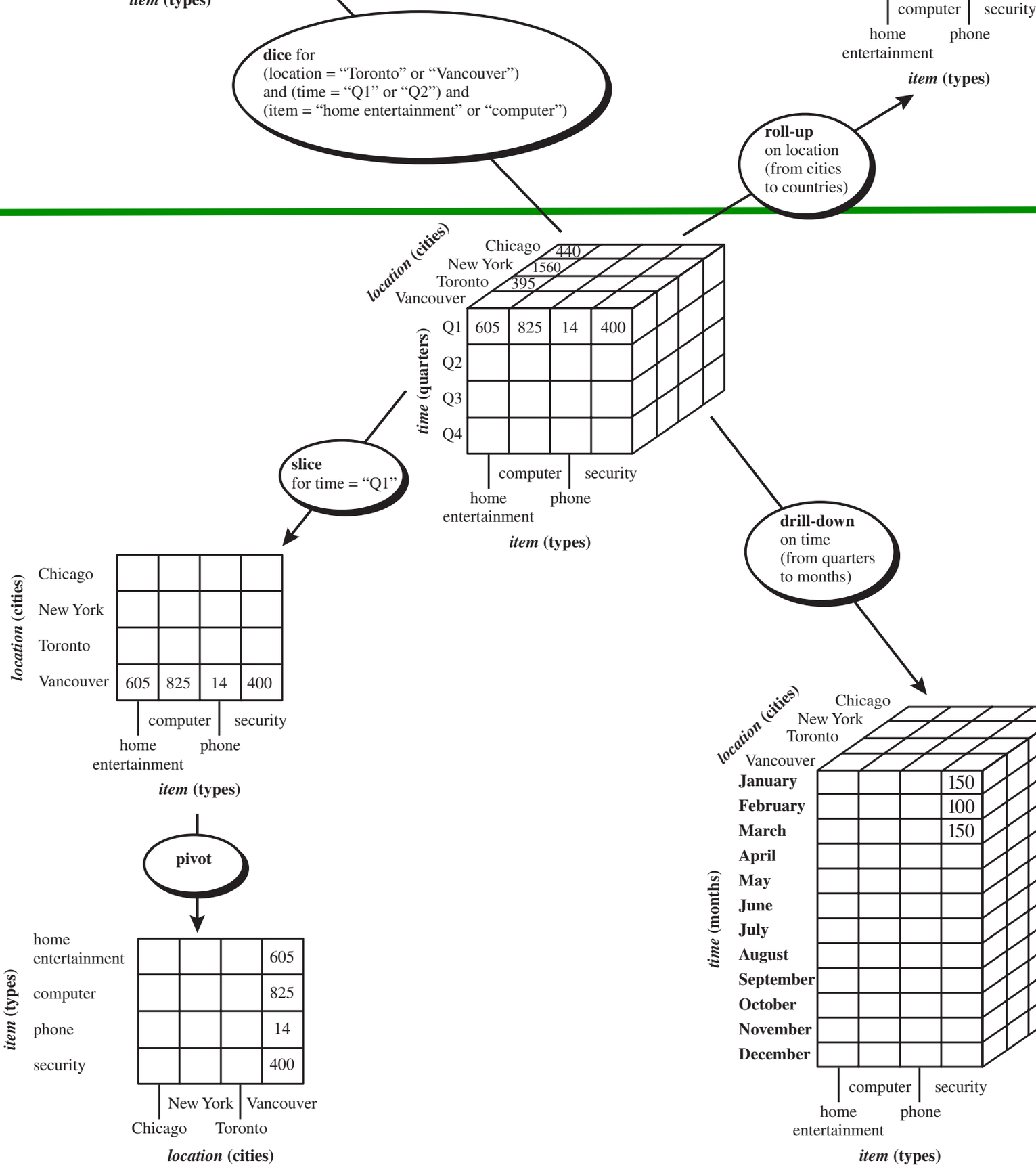


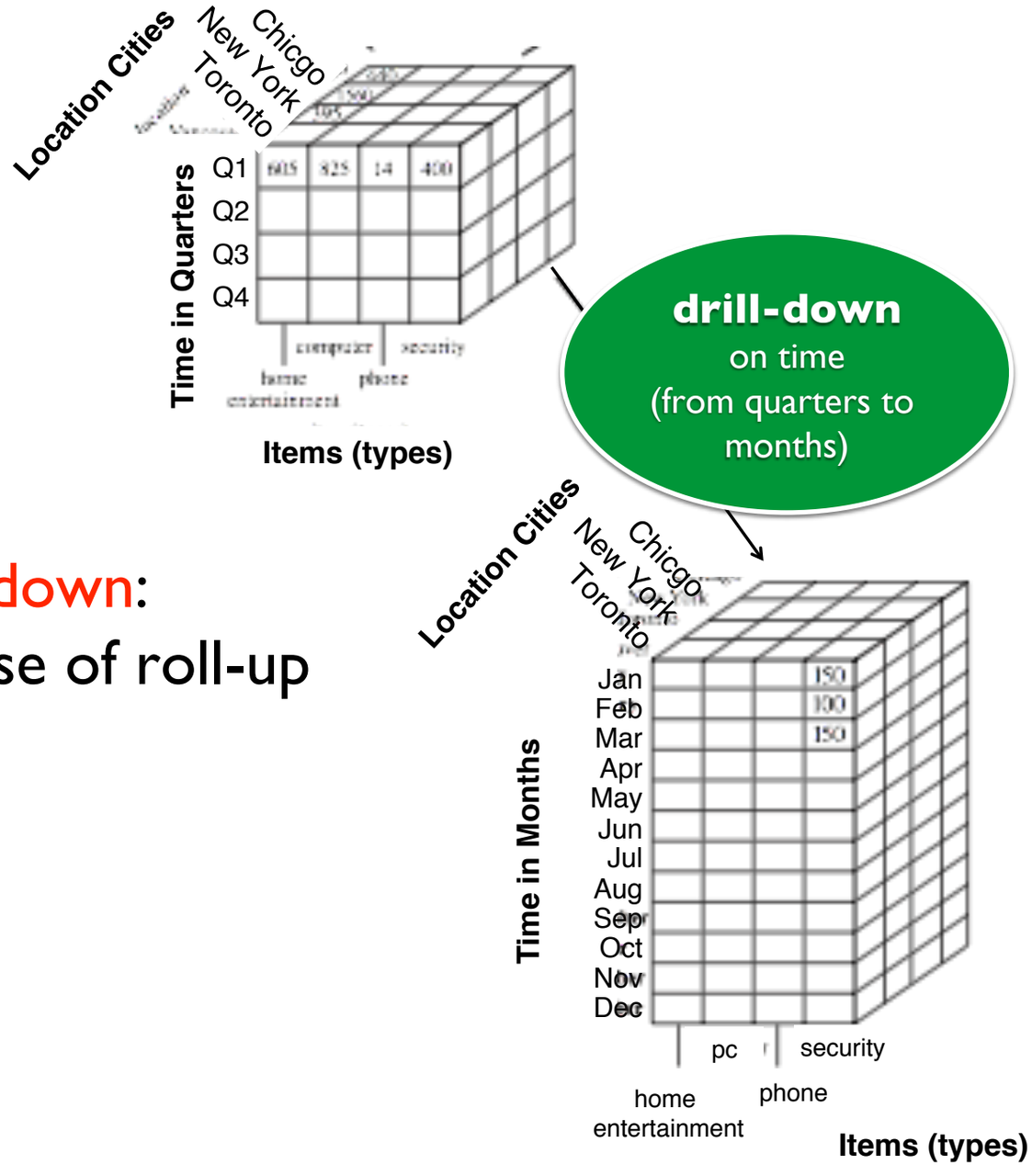
Browsing a Data Cube



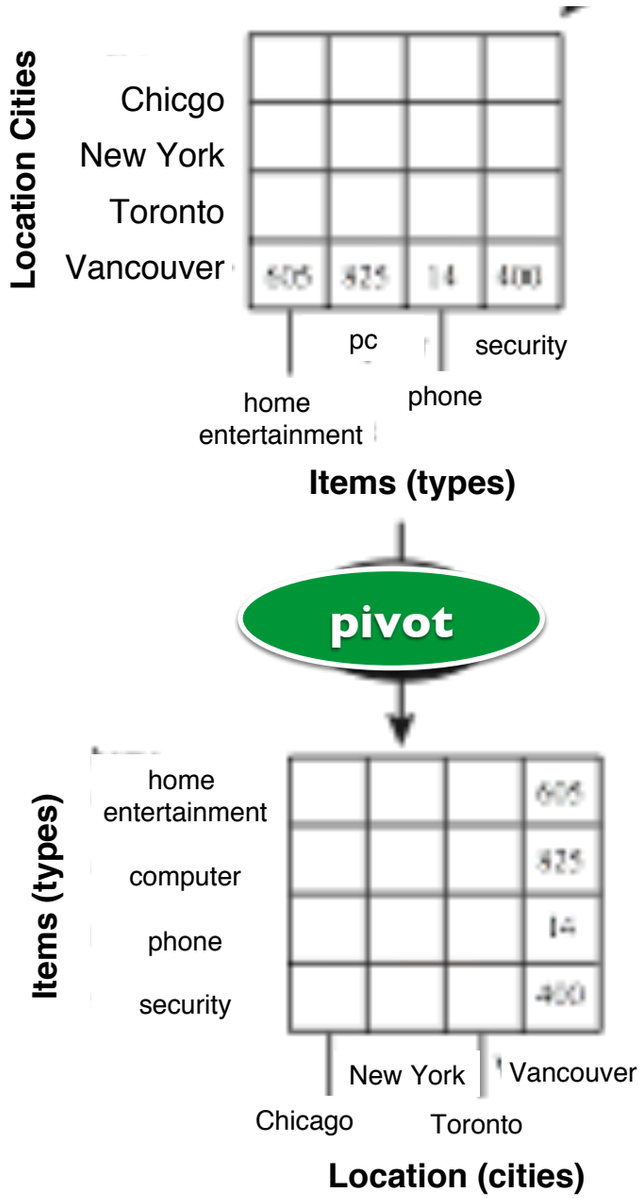
Typical OLAP Operations

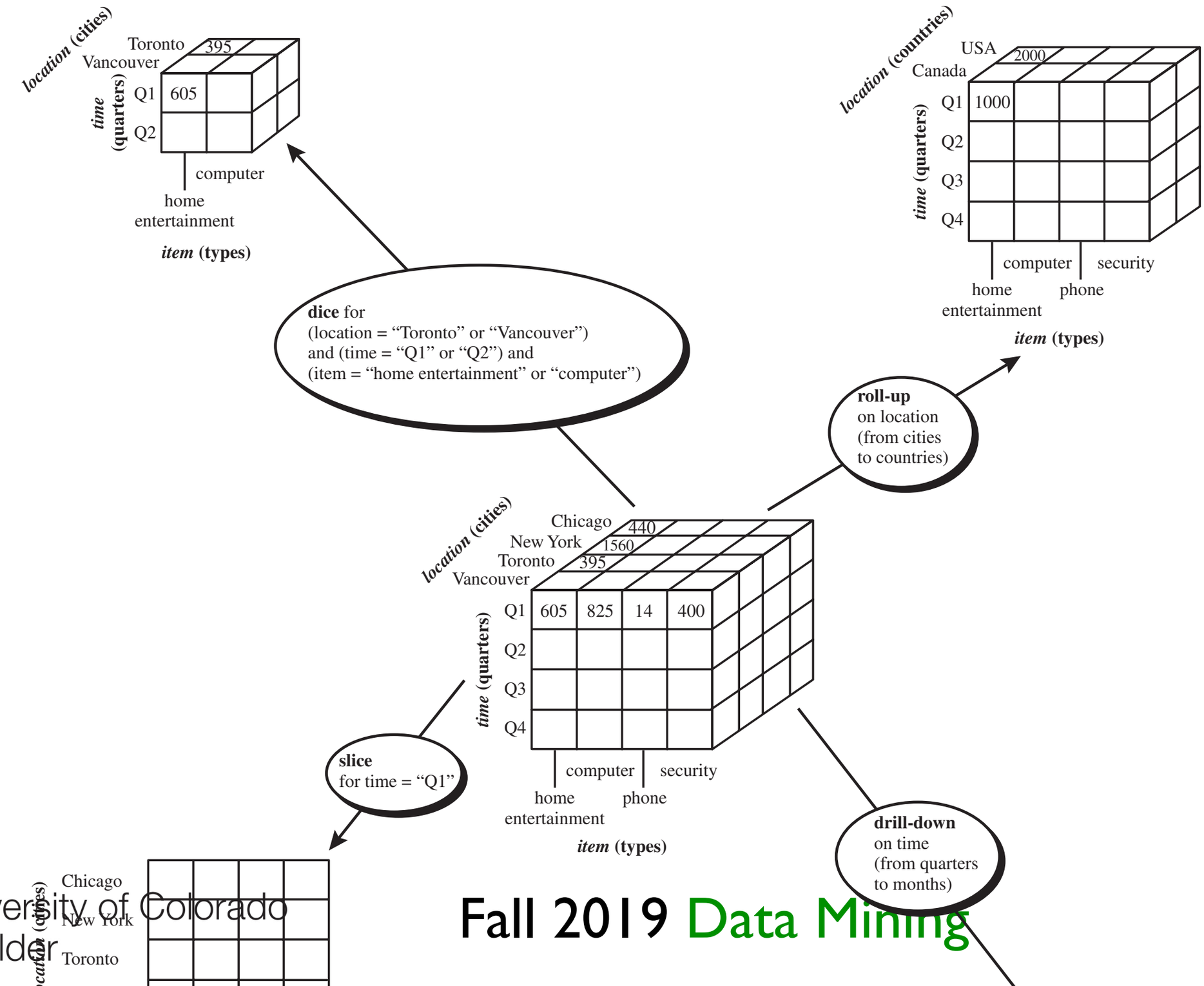
- Roll-up (drill-up): summarization
- Drill-down: reverse of roll-up
- Slice and dice: project and select (sub-cube)
- Pivot (rotate): visualization, 3D to 2Ds
- drill-across: more than one fact tables
- drill-through: to the back-end relational tables

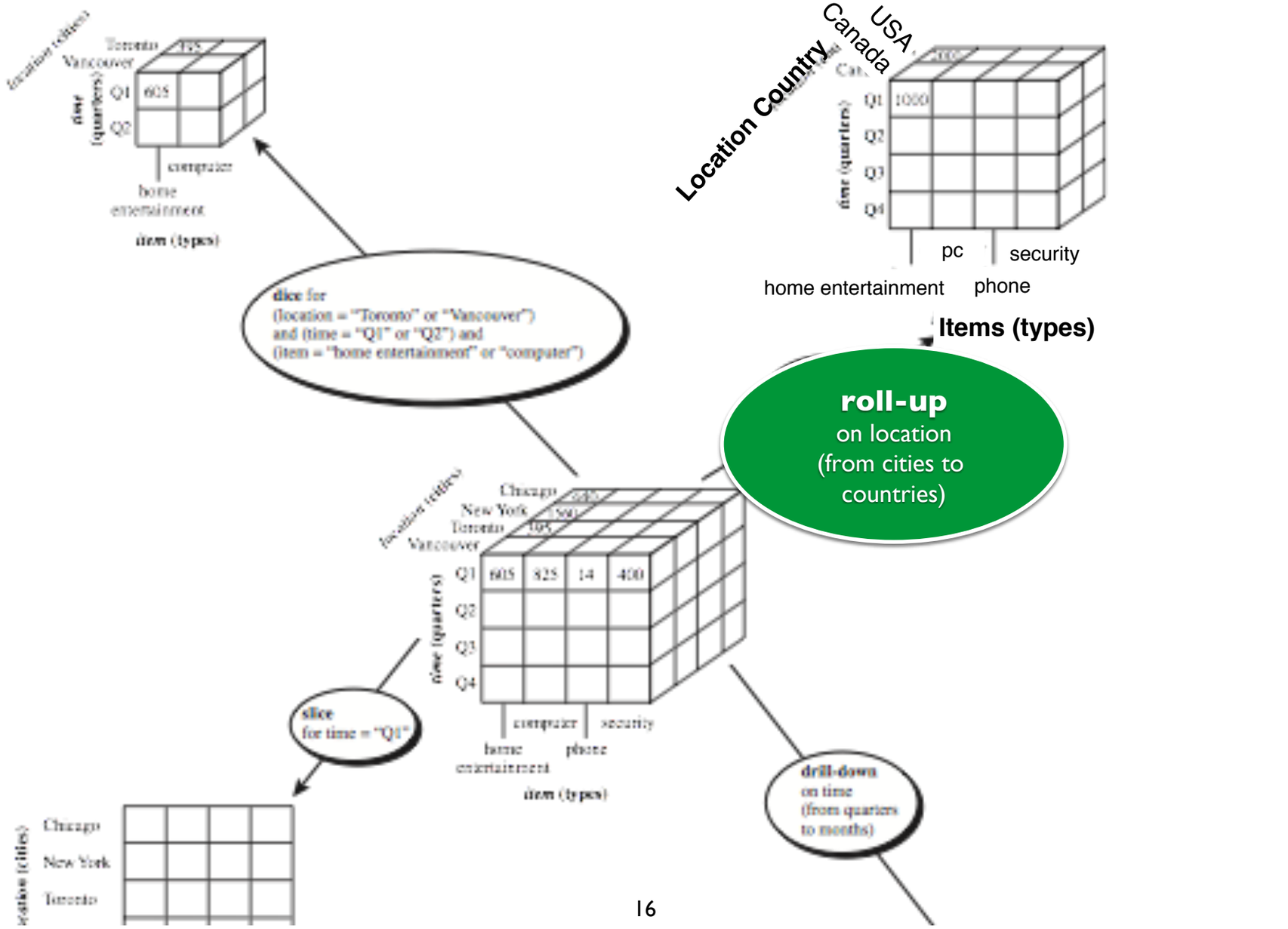

item (types)
dice for
(location = "Toronto" or "Vancouver")
and (time = "Q1" or "Q2") and
(item = "home entertainment" or "computer")
computer
security
home
entertainment
phone
item (types)
roll-up
on location
(from cities
to countries)
location (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395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



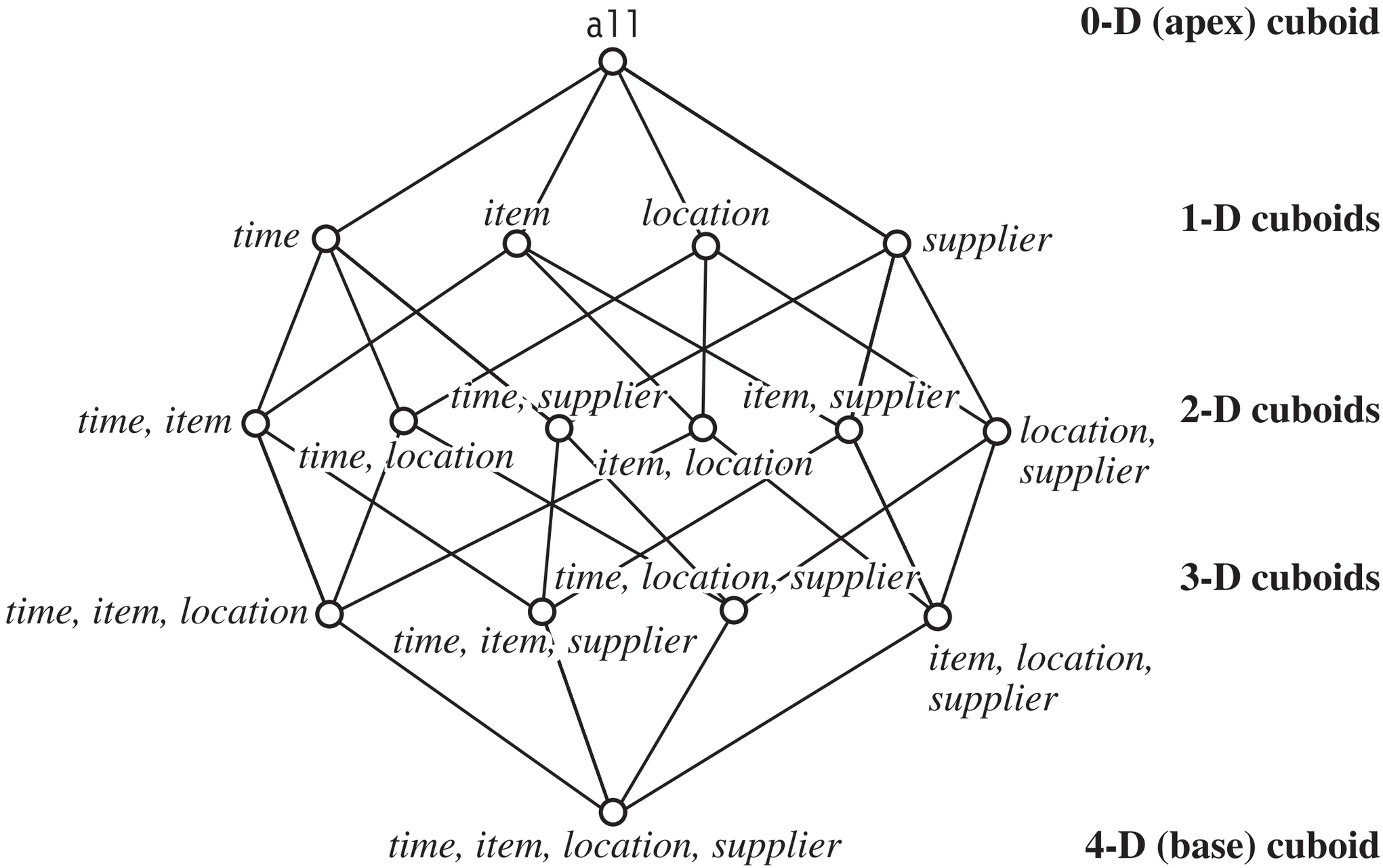
- Drill-down:
  reverse of roll-up


Location Cities
Chicgo
New York
Toronto
Vancouver
605
825
14
400
pc
security
home entertainment
phone
Items (types)
pivot
Items (types)
home entertainment
computer
phone
security
605
825
14
400
New York
Vancouver
Chicago
Toronto
Location (cities)


location (cities)
Toronto
Vancouver
time (quarters)
Q1
Q2
395
605
computer
home entertainment
item (types)
dice for
(location = "Toronto" or "Vancouver")
and (time = "Q1" or "Q2") and
(item = "home entertainment" or "computer")
location (countries)
USA
Canada
2000
1000
Q1
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
605
825
14
400
Q1
Q2
Q3
Q4
time (quarters)
computer
security
home entertainment
phone
item (types)
slice
for time = "Q1"
drill-down
on time
(from quarters
to months)
location (cities)
Chicago
New York
Toronto
Vancouver
605
825
14
400


Location Country
Canada
USA
Q1
Q2
Q3
Q4
pc
security
home entertainment
phone
Items (types)
roll-up
on location
(from cities to
countries)
Chicago
New York
Toronto
Vancouver
605
825
14
400
computer
security
home
entertainment
phone
item (types)
slice
for time = "Q1"
drill-down
on time
(from quarters
to months)
dice for
(location = "Toronto" or "Vancouver")
and (time = "Q1" or "Q2") and
(item = "home entertainment" or "computer")

Cube: Lattice of Cuboids

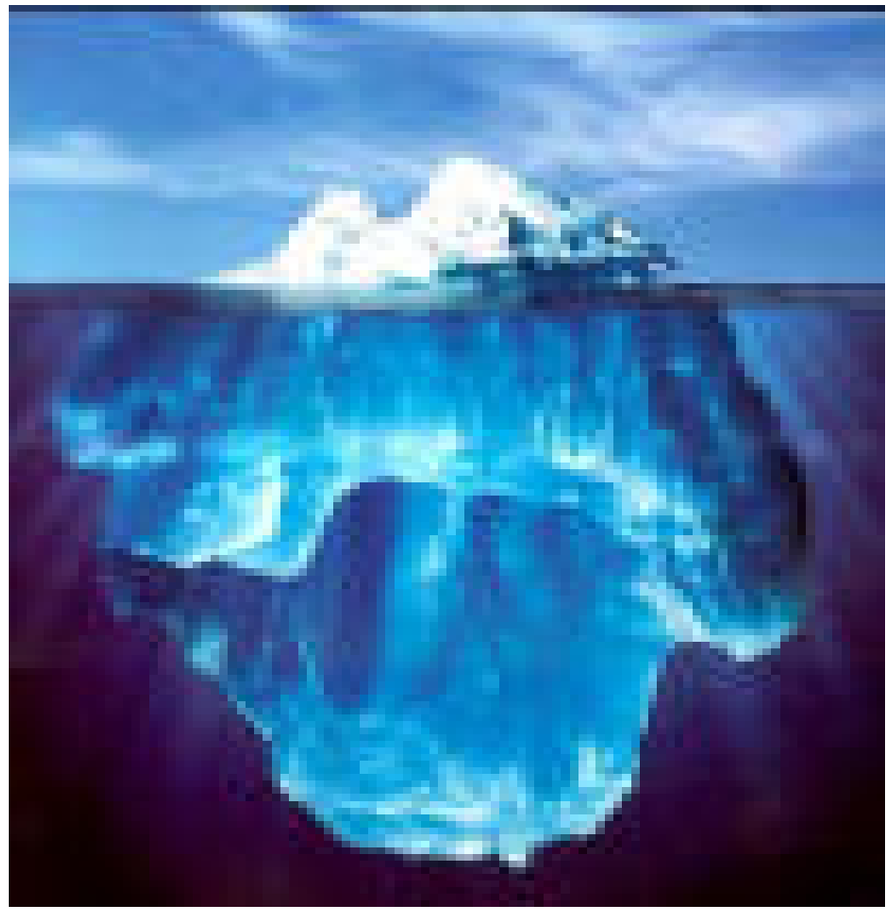


Cuboid Cells

- Cuboid cells
  - base vs. aggregate, ancestor vs. descendant
- E.g., (month, location, customer_group, price)
  - a = (Jan, *, *, 2800)
  - b = (Jan, *, Business, 150)
  - c = (Jan, Toronto, Business, 45)
- Materialization of data cube
  - full, partial, or no materialization

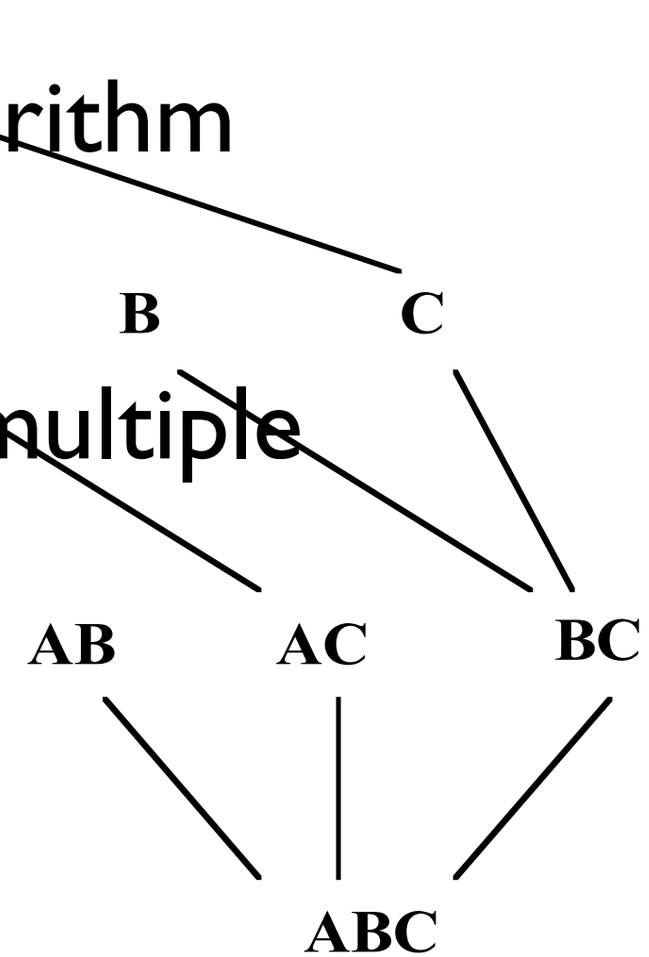
Iceberg Cube

- Only a small portion may be "above the water" in a sparse cube
- Compute only the cuboid cells whose aggregate (e.g., count) is above a threshold
  - minimum support
- Avoid explosive growth of the cube

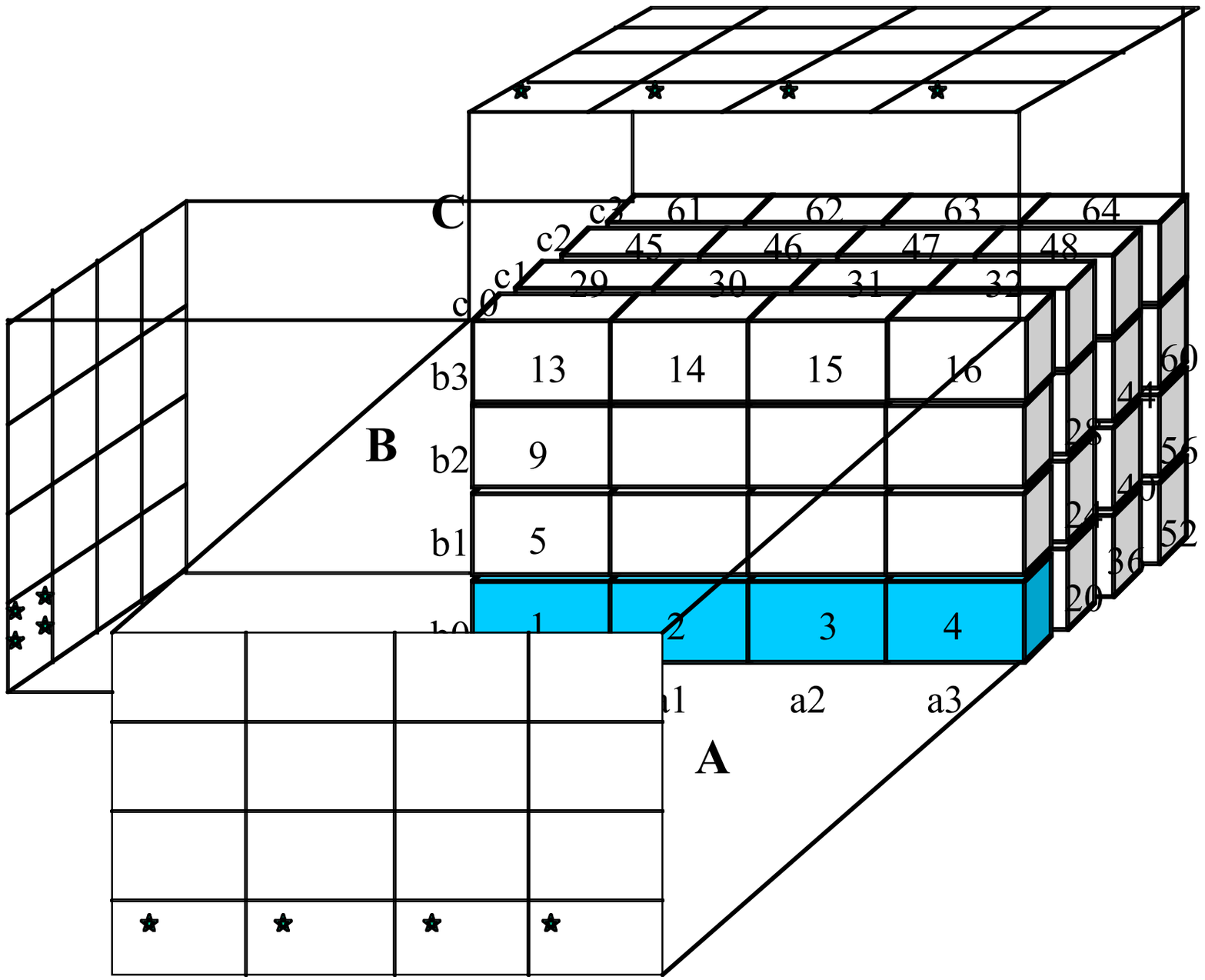


Multi-Way Array Aggregation

- Full cube computation
- Array-based "bottom-up" algorithm
- Multi-dimensional chunks
- Simultaneous aggregation on multiple dimensions
- Cannot do Apriori pruning
- Not for high dimensions

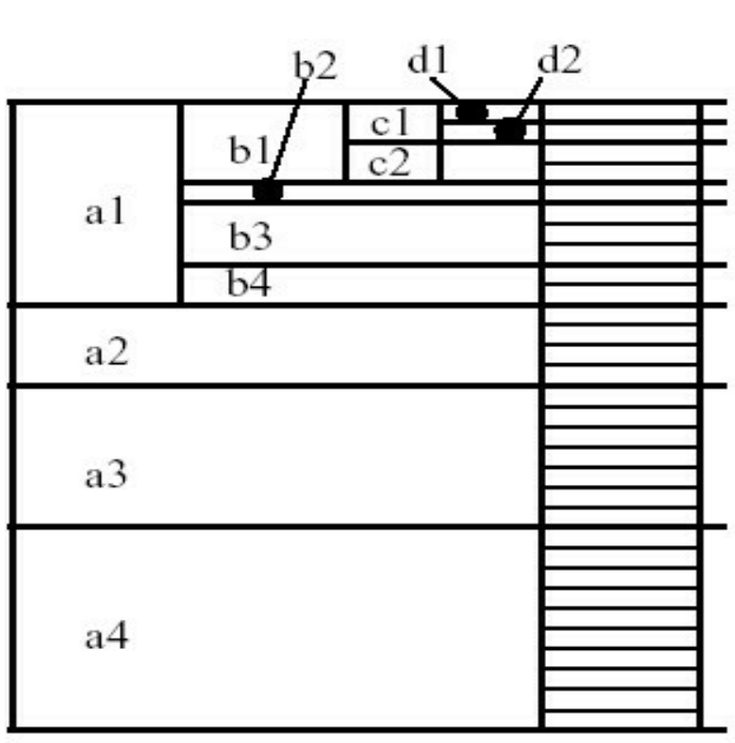


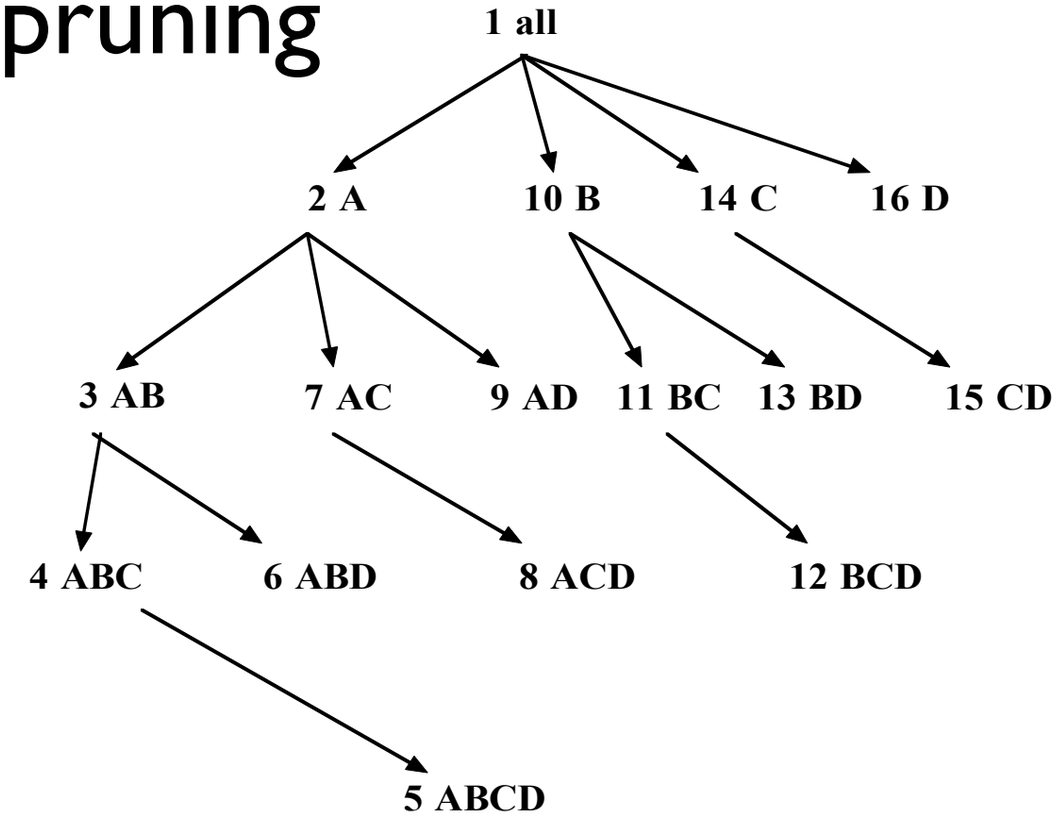
Multi-Way Array Aggregation



Bottom-Up Computation (BUC)

- [Beyer & Ramakrishnan, SIGMOD'99]
- Top-down computation of iceberg cubes
- Divides dimensions into partitions and facilitates iceberg pruning





Summary

- Chap 4 & 5: Data Warehouse, Data Cube
  - what is data warehouse?
  - OLTP vs. OLAP
  - what is data cube?
  - data cube operations
  - data cube computation

Review: Part I

- Chapter 1: Introduction
- Chapter 2: Getting to Know your Data
- Chapter 3: Data Preprocessing
- Chapter 4: Data Warehousing and Online Analytical Processing
- Chapter 5: Data Cube Technology
- Part 2: Core DM Techniques
- Part 3: Mining Complex Data, DM Trends

Course Project

- 40% of overall grade
- A self-contained project related to DM
- Work in teams (3-4 students)
- Pick your own project idea
- Project forum at moodle, online resources
- Discuss w/ instructor & other students
- Start early!!

Choose Your Project

- What are you interested in?
- Who are on your team?
- What data set(s) are available?
- What problem do you want to answer using the data set(s)?
- Why is the problem interesting/important?
- What are the challenges?
- What have been done before? Limitations?

Define Your Project

- Project title
- Project team (4502/5502, expertise, tasks)
- Problem statement
- Motivation, literature survey
- Significance & difference given prior work
- Proposed work (data set, approaches)
- Evaluation: metrics, existing solutions
- Milestones

Course Project Proposal

- Week 7
- Team formation & project identification
  - discuss w/ instructor & other students
- Project forum announcement
  - team, title, brief description
- Project proposal report
- Checkpoint (Week 12)
- Final report (Week 16)