IS 733 Lesson 4

Data Warehousing

Some slides based on those by Witten et al., Han et al., and James Foulds

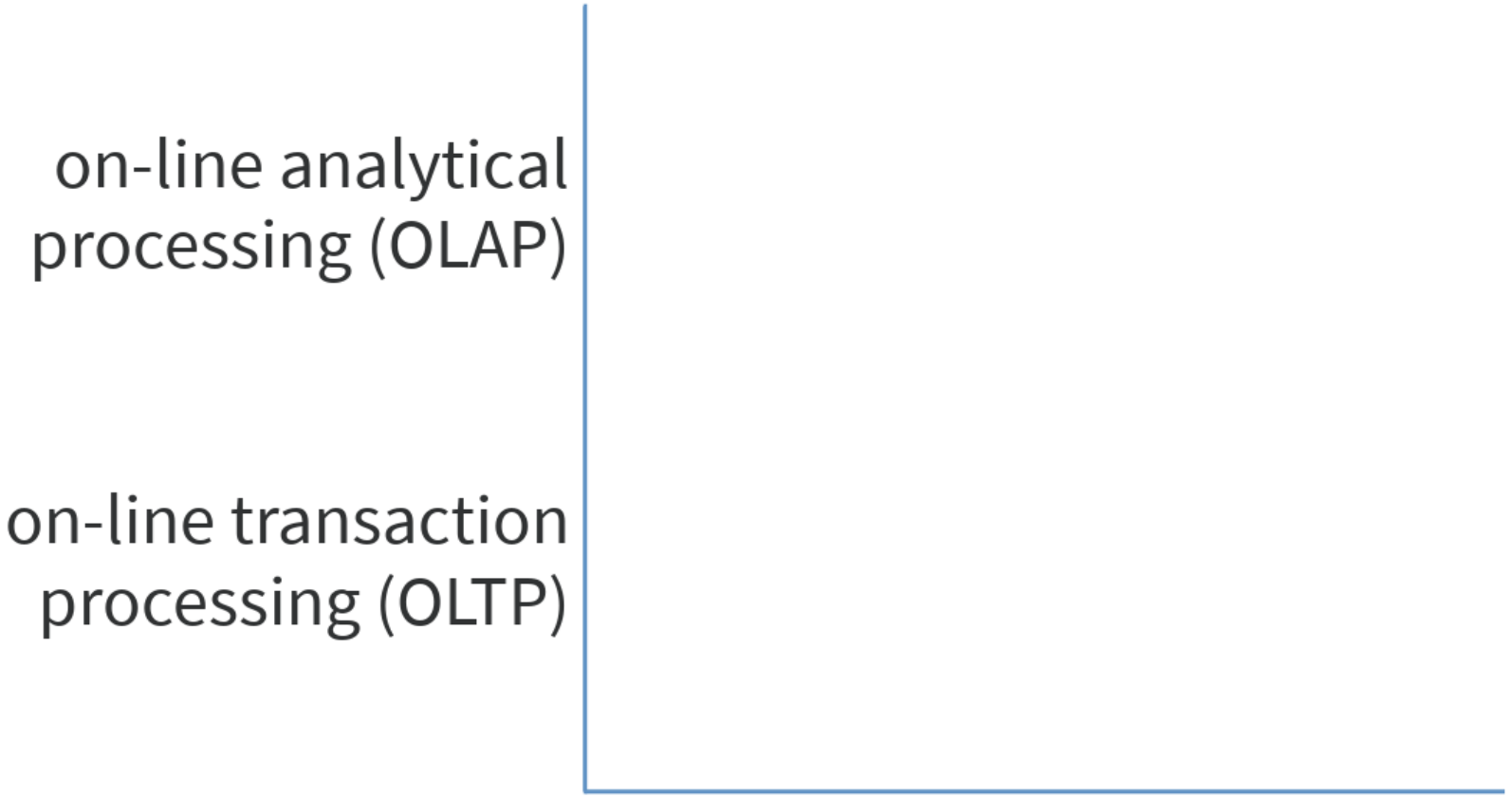
Announcements

- **Homework 2** is up on course website, due 3/2/2021

- Project **proposal due today**, 2/23! Please send this to me by email, nroy@umbc.edu, and CC your teammates

- Reminder: make use of BB discussion forum for asking questions!

Data Warehouses support _____ applications.



According to Chaudhuri and Dayal (1997), which of the following is NOT important for data warehouses to handle effectively, compared to operational databases?

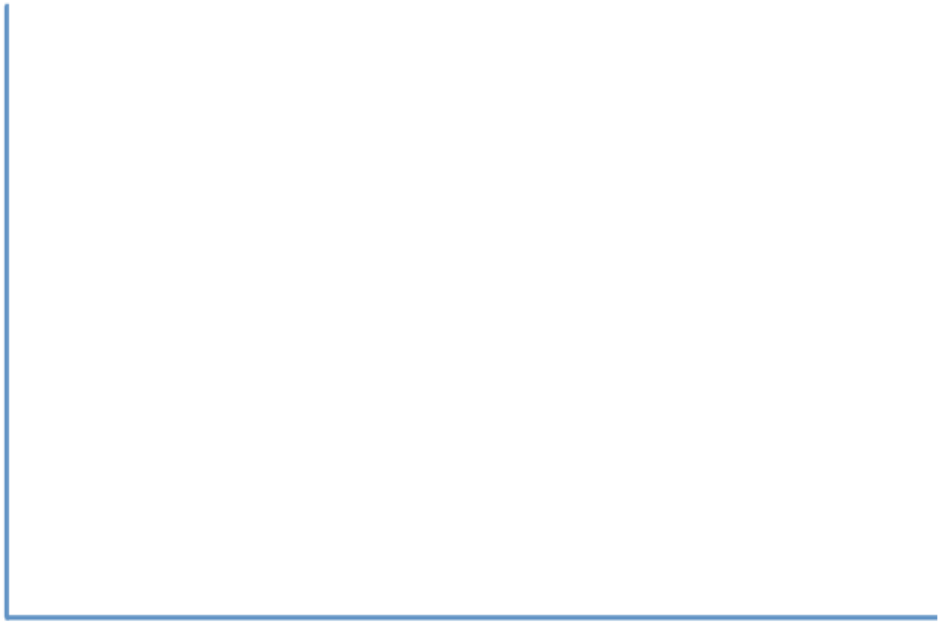
Query intensive workloads, with many complex queries

Managing concurrency conflicts in order to maximize transaction throughput

High query throughput

Short response times for queries

Consolidating data from many heterogeneous sources



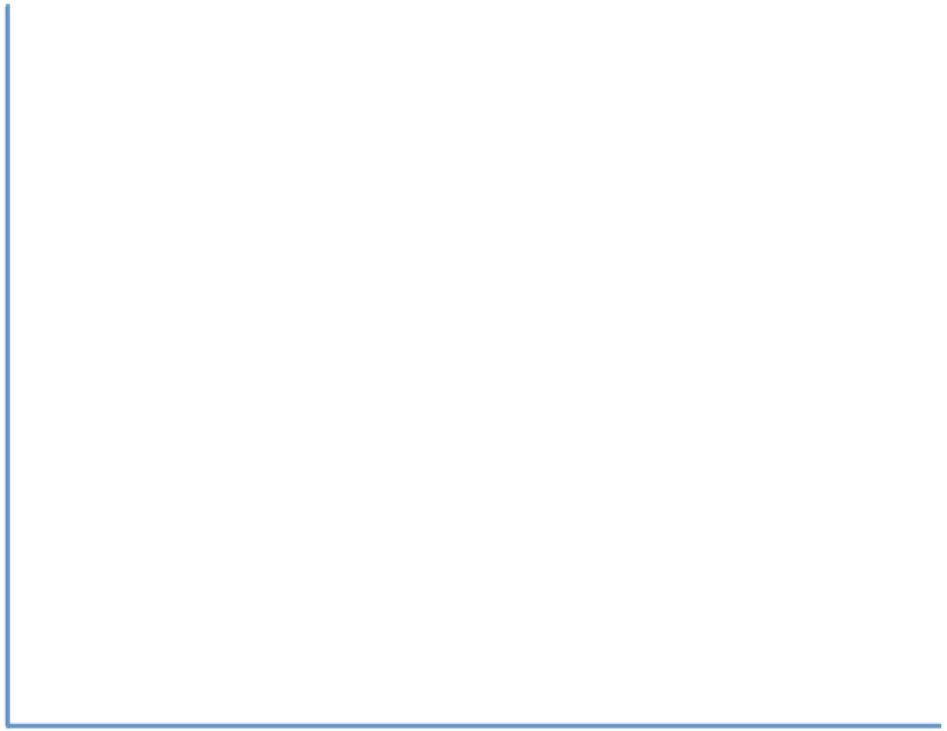
To facilitate complex analyses and visualization, the data in a data warehouse is typically modeled _____.

statistically

financially

transactionally

multidimensionally



Learning outcomes

By the end of the lesson, you should be able to:

- **Enumerate** the key differences between **OLTP** and **OLAP**, and the **corresponding benefits** from the use of a **data warehouse**
- **Compute** the number of **cuboids** in a **data cube**
- **Select** appropriate **OLAP operations** to obtain relevant information from a data warehouse

Data warehousing

Basic concept

What is a Data Warehouse?

- Defined in many different ways, but not rigorously.
  - 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 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—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 vs. Operational DBMS

- OLAP (on-line analytical processing)
  - Major task of data warehouse system
  - Data analysis and decision making

- OLTP (on-line transaction processing)
  - Major task of traditional relational DBMS
  - Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.

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 a Separate Data Warehouse?

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

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

Data warehousing

Data Cubes and OLAP Operations

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.

Flat view of Data

Data about items

home entertainment	computer	phone	security
605	825	14	400

Data about Quarters

Q1	605
Q2	680
Q3	812
Q4	927

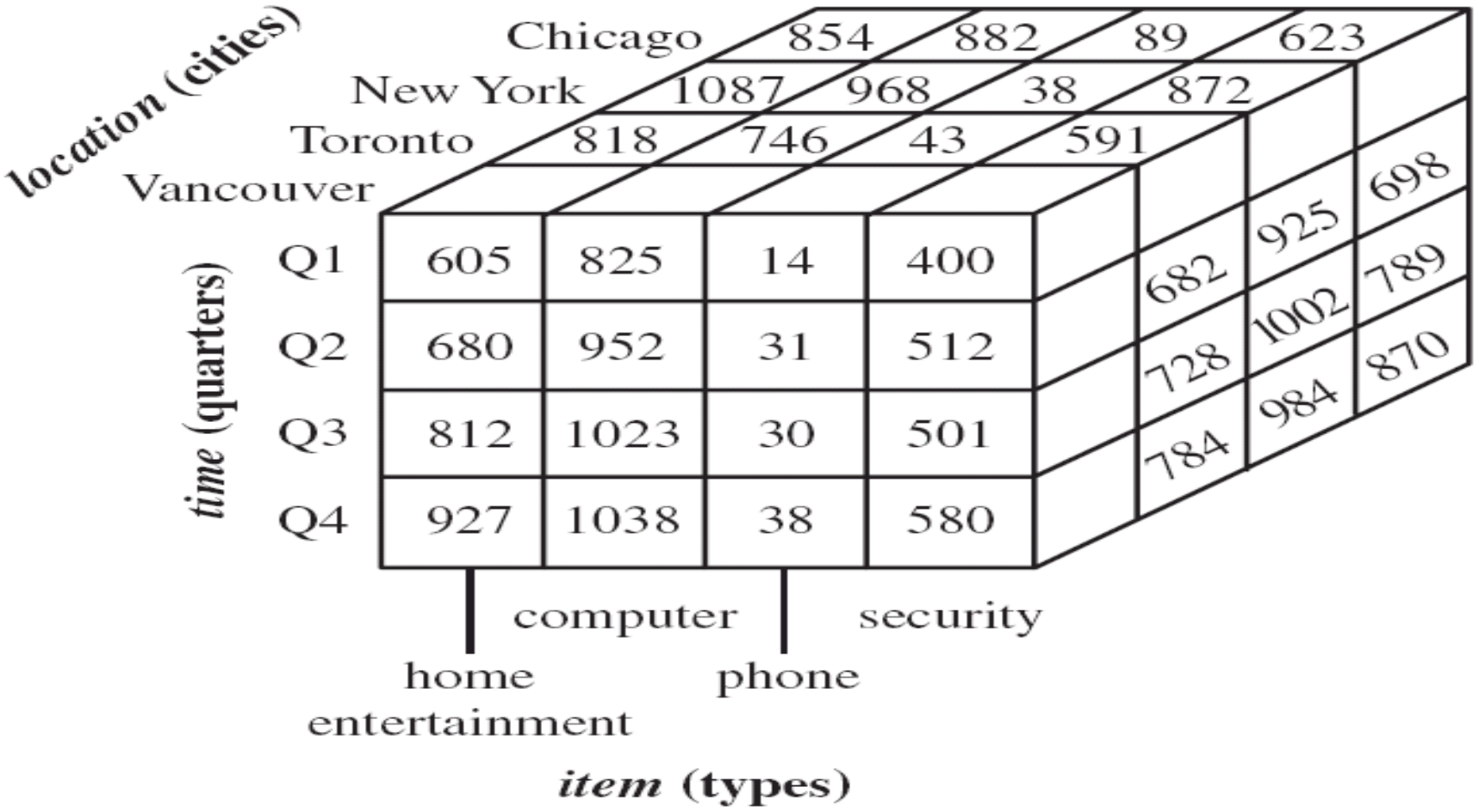
Multidimensional data

time (quarter)	*location* = "Vancouver"			
	item (type)			
	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

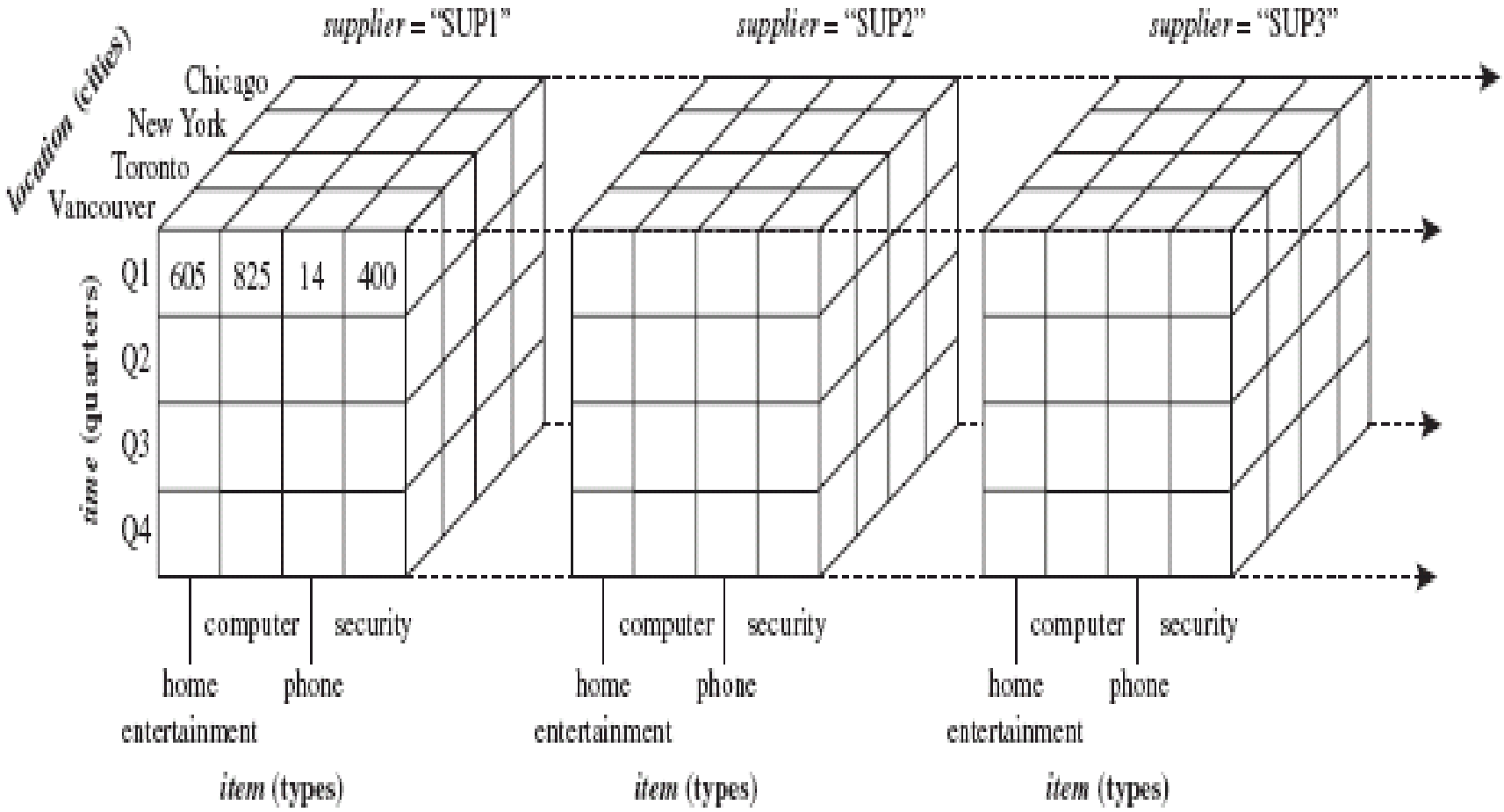
Data for multiple locations

	location = "Chicago"				*location* = "New York"				*location* = "Toronto"				*location* = "Vancouver"			
time	*item*				*item*				*item*				*item*			
	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

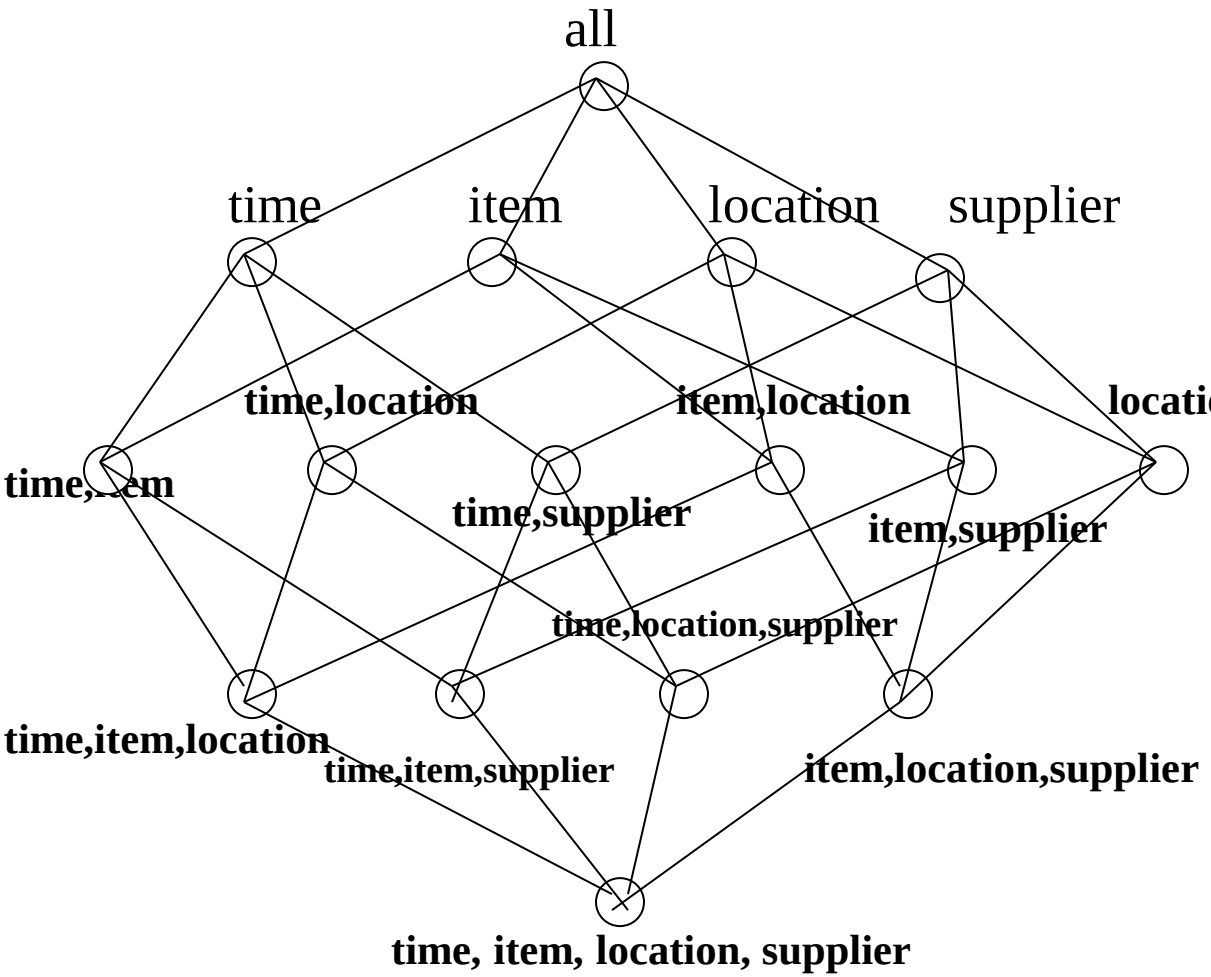
Multidimensional view -3D



Multidimensional view -4D



Data Cube: A Lattice of Cuboids



0-D (*apex*) cuboid

1-D cuboids

2-D cuboids

3-D cuboids

4-D (*base*) cuboid

How many cuboids are there in a data cube with 5 dimensions (e.g. time, item, location, supplier, branch)?

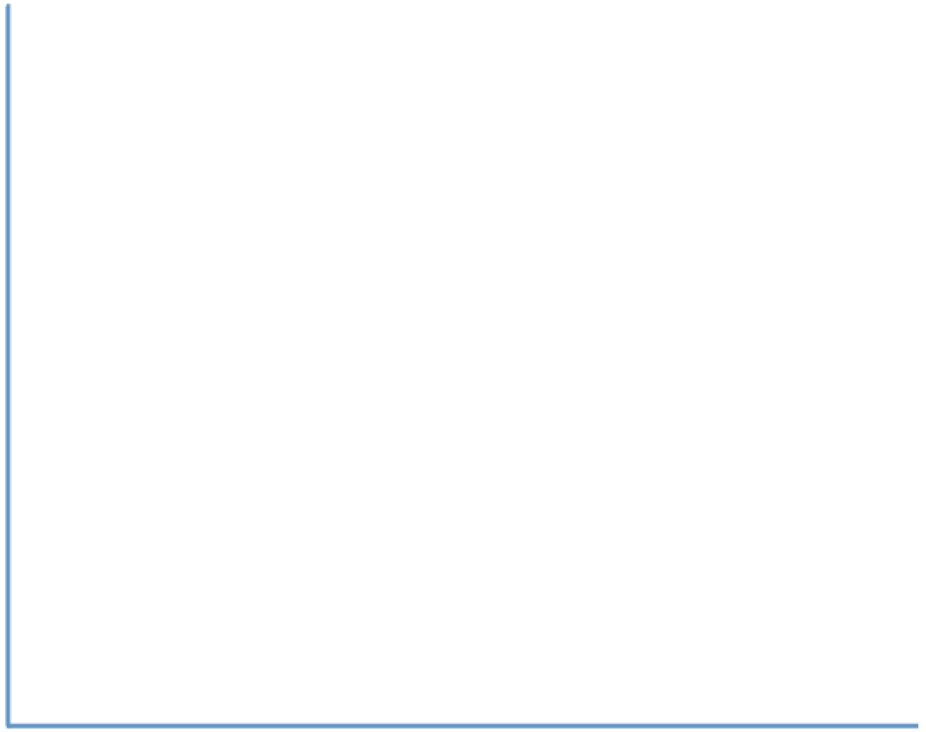
2^5

5^2

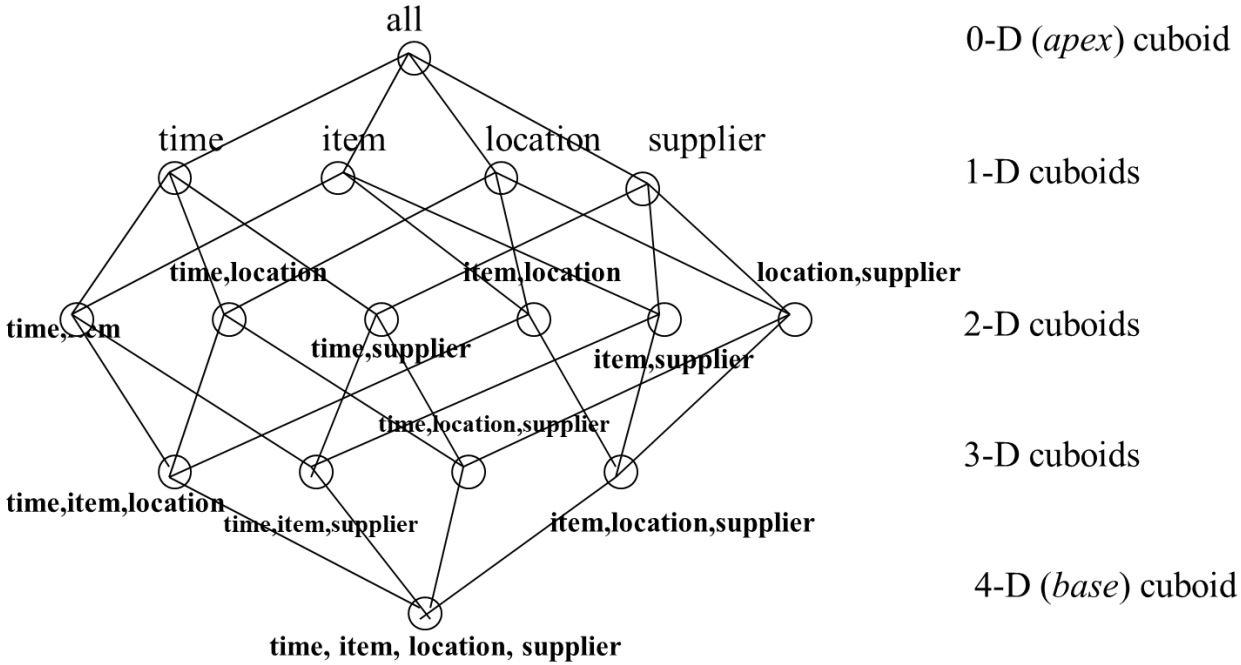
$\binom{5}{2}$

$\binom{2}{5}$

$5!$



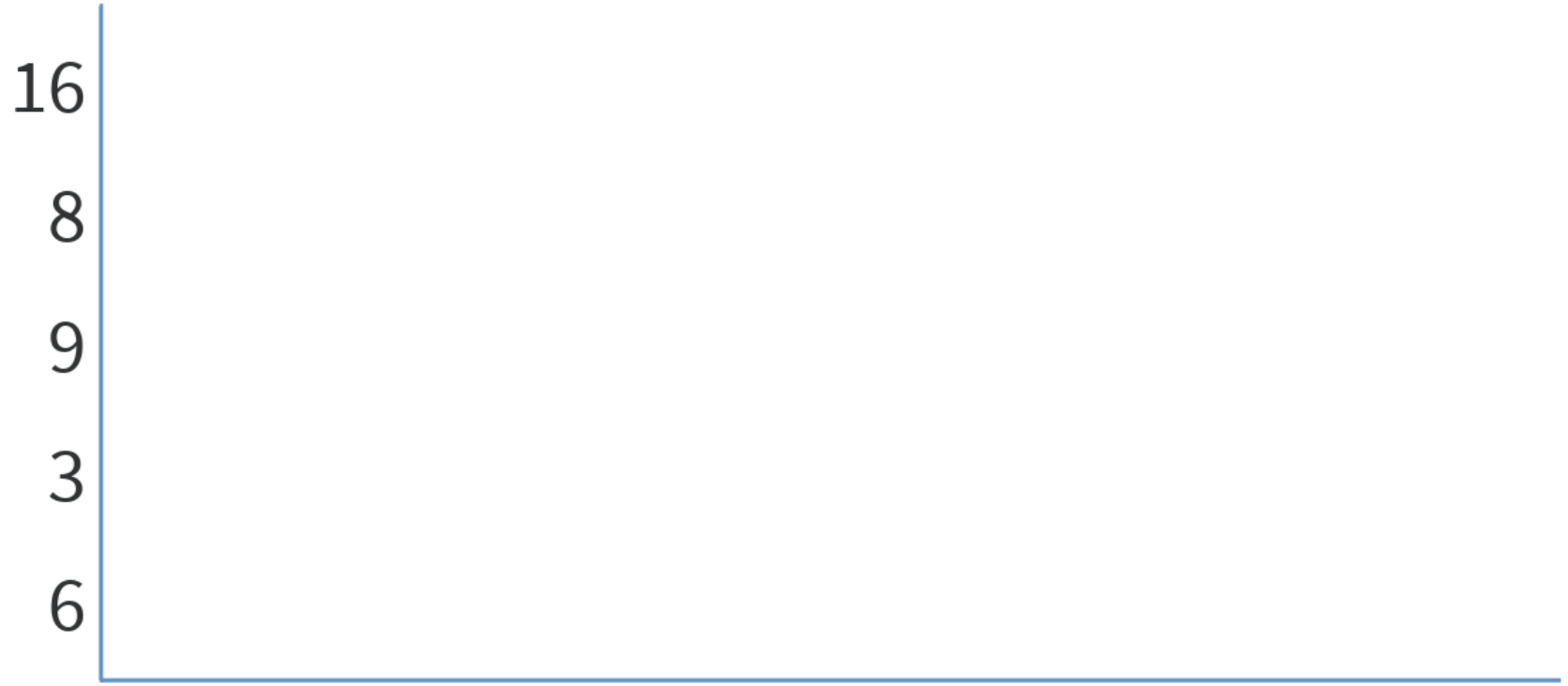
How Many Cuboids in a Data Cube?



With 4 dimensions, there are 16 cuboids (count them!)

In general, there are 2^n cuboids.
To see this, imagine encoding each cuboid with a binary number, e.g. *time, item, supplier= [1, 1, 0, 1].*

How many cuboids are there in a data cube with 3 dimensions (e.g. time, item, location)?



Conceptual Modeling of Data Warehouses

- Modeling data warehouses: **dimensions** & **measures**

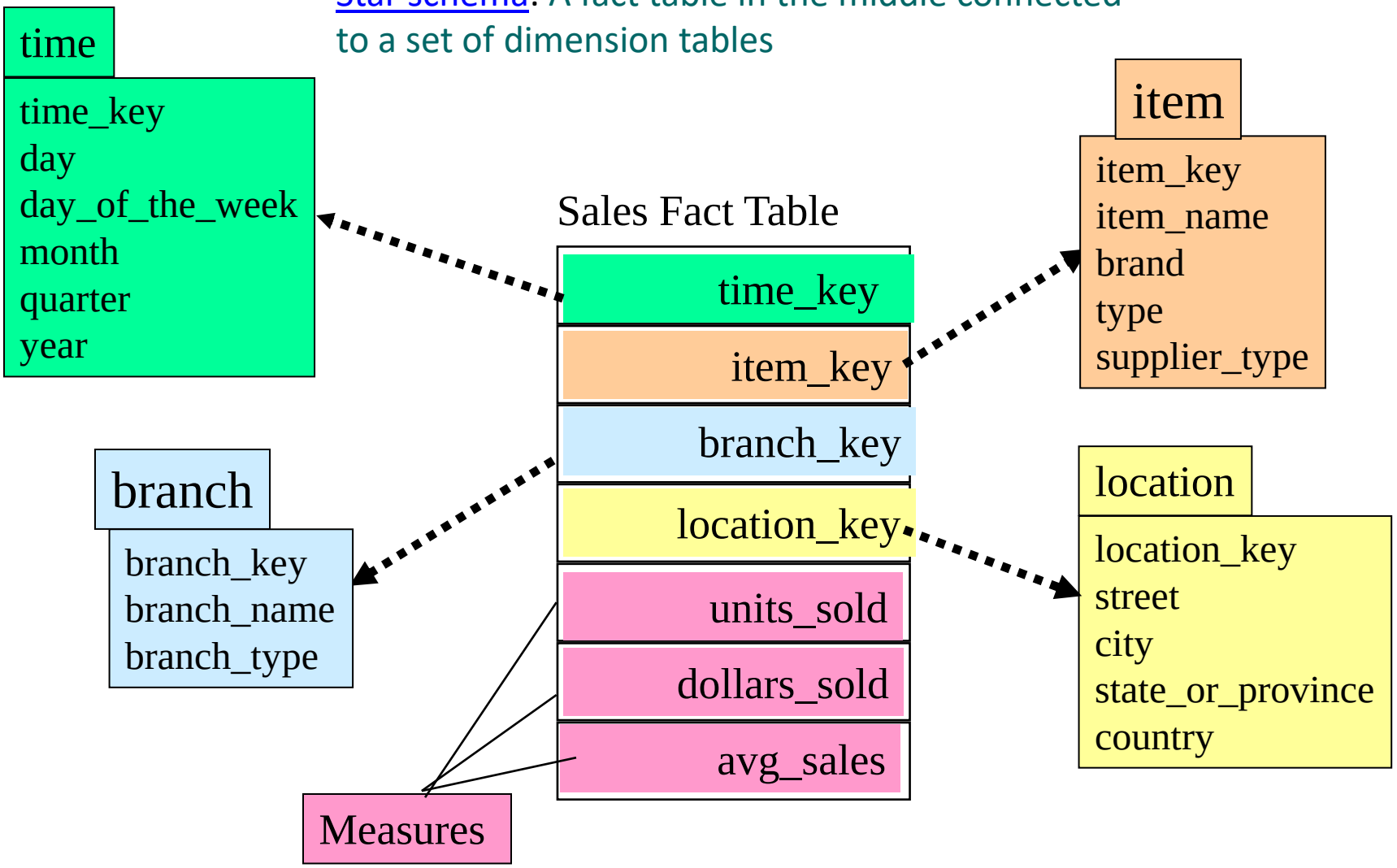
  - **Measures:** the quantities we are interested in (units sold, dollars sold, average sales, etc). Usually a total or an average of something
    - Stored in a **fact table**

  - **Dimensions:** the dimensions on which the measures can vary (time, item type, location, supplier, etc…)
    - Each dimension has a **dimension table** encoding information about that dimension

  - Different database schemas are possible with the fact and dimension tables
    - Star schema, Snowflake schema, Fact constellations

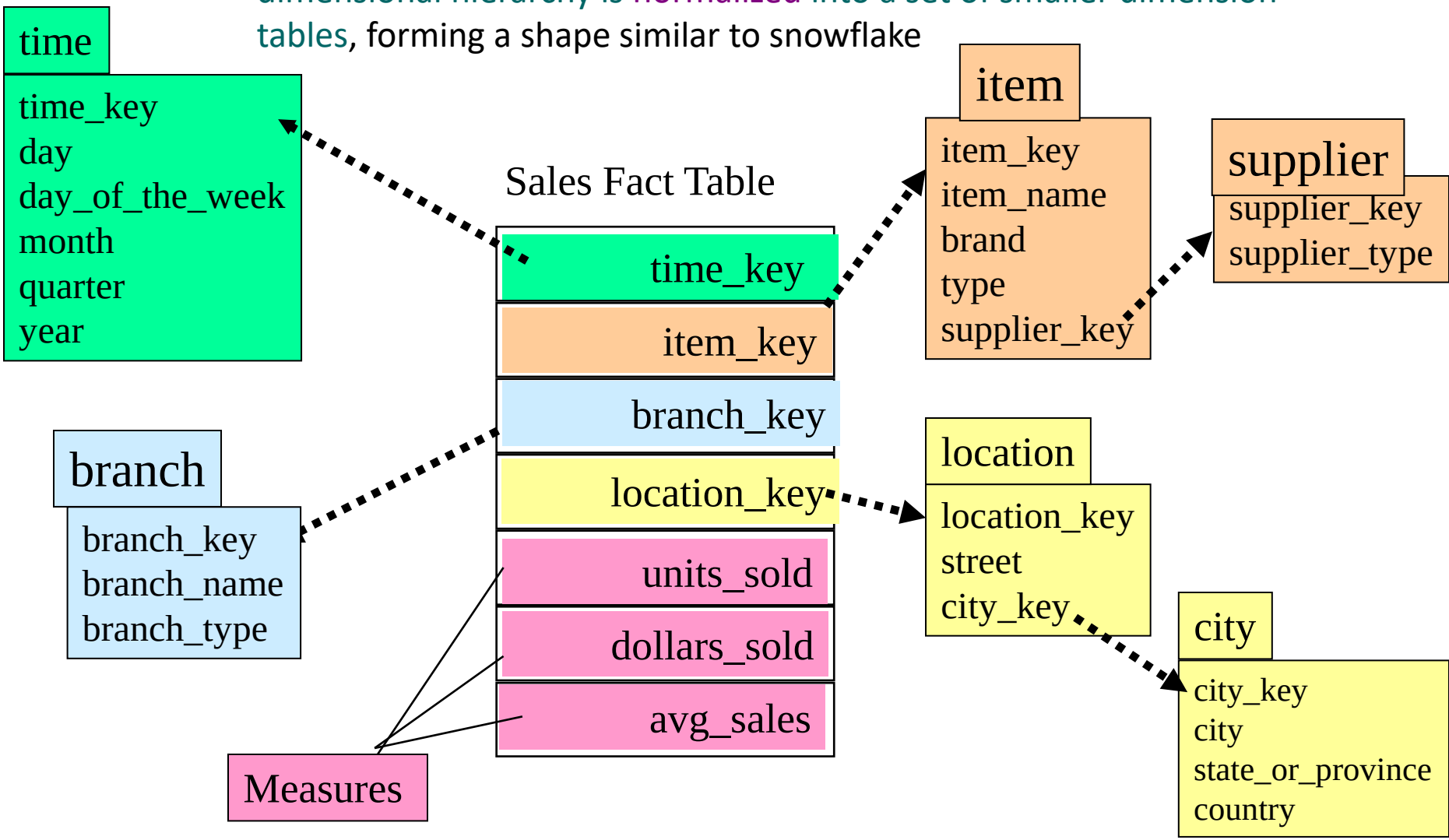
Star Schema

Star schema: A fact table in the middle connected to a set of dimension tables



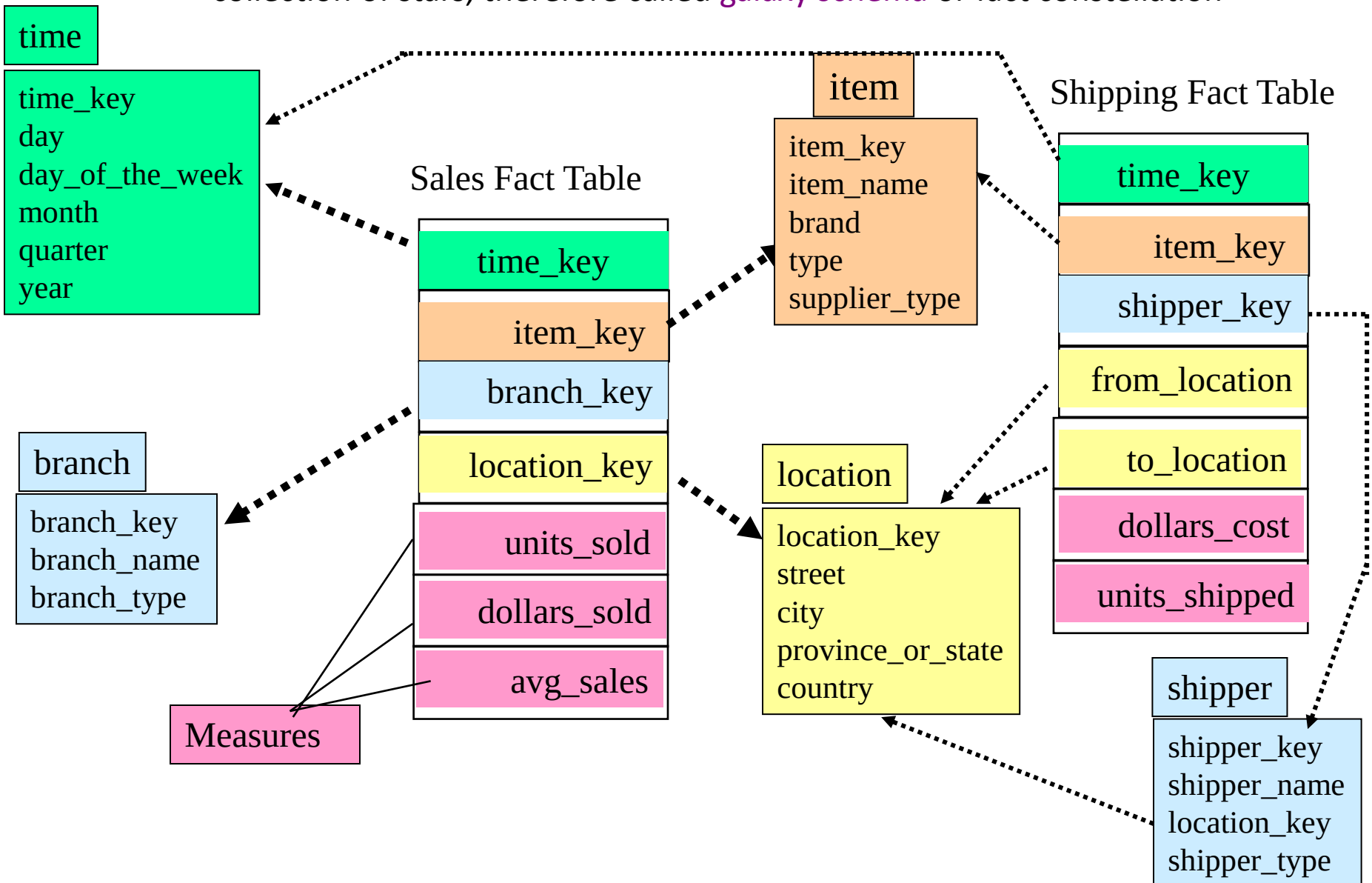
Snowflake Schema

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

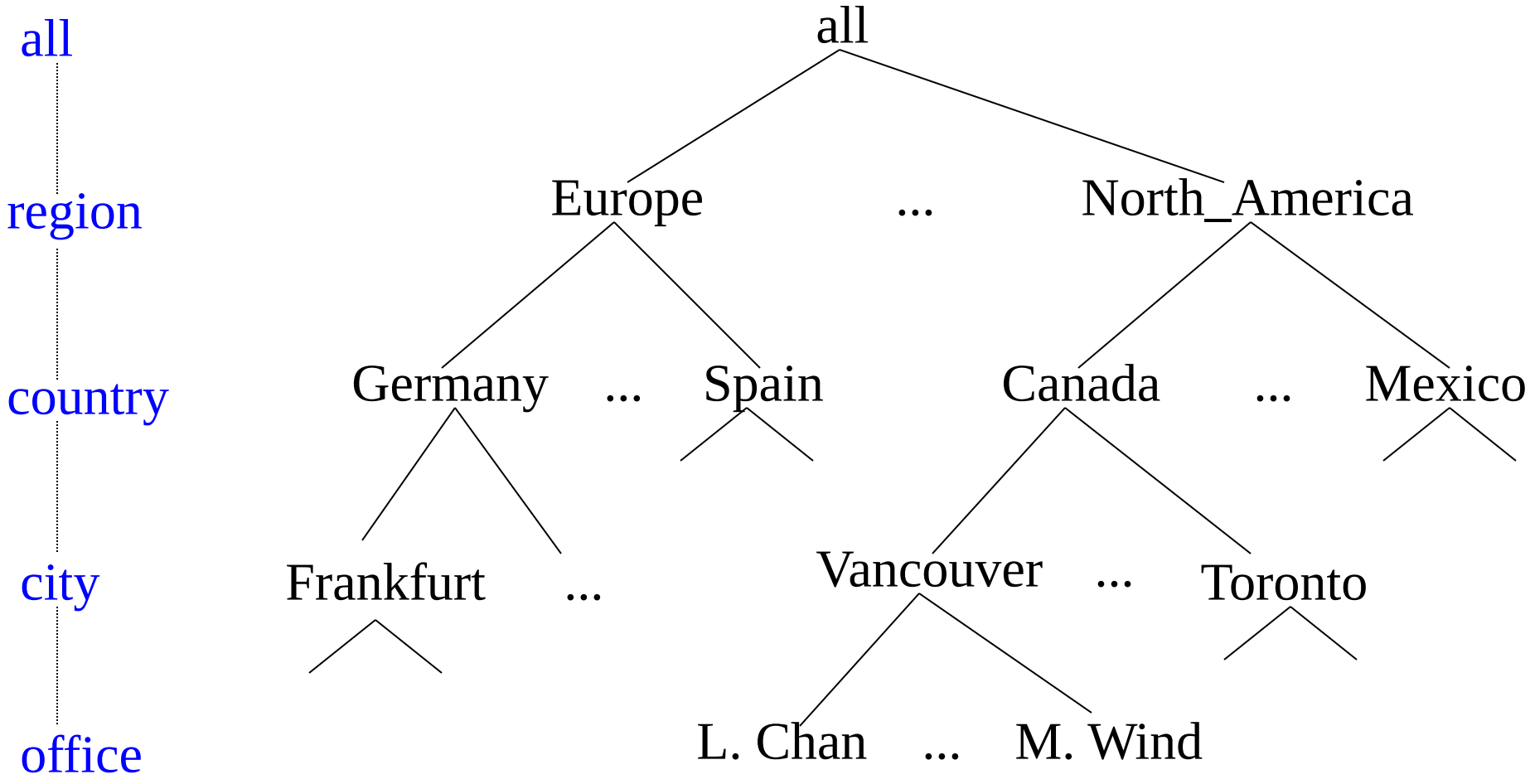


Fact Constellation

Fact constellations: Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation

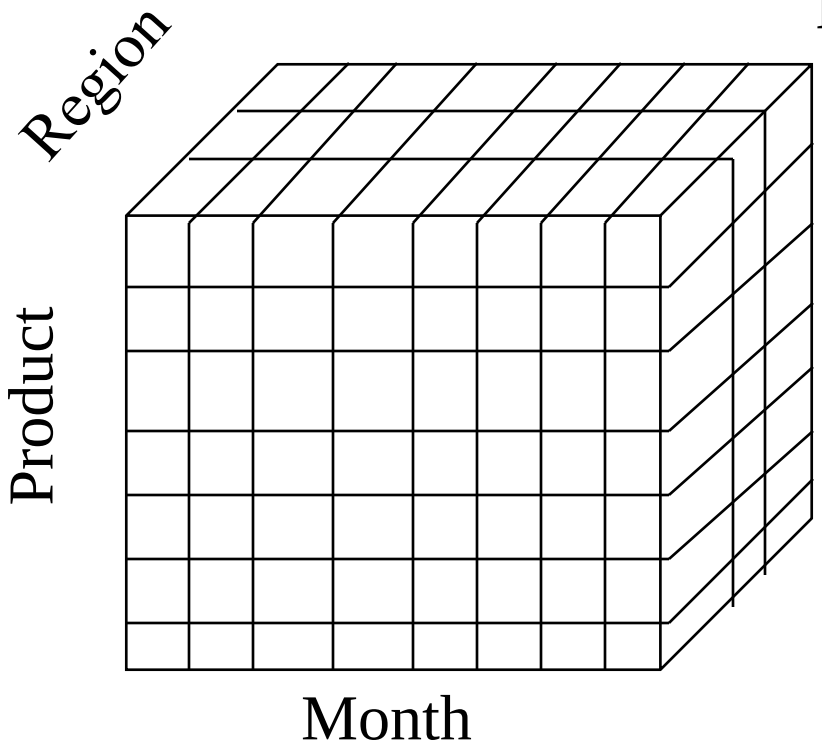


A Concept Hierarchy: **Dimension** (location)

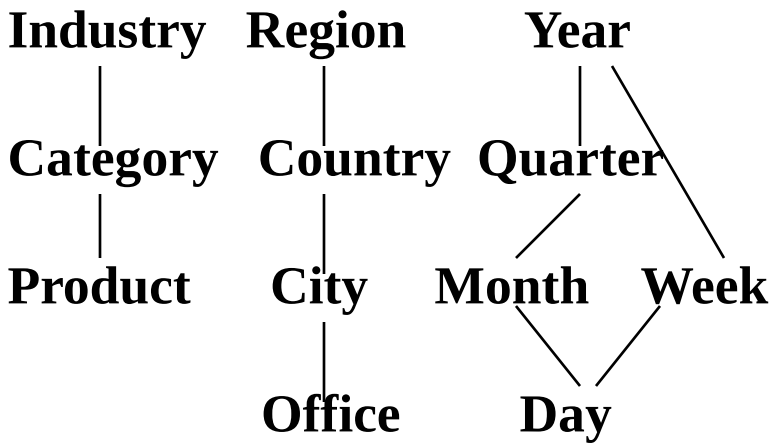


Multidimensional Data

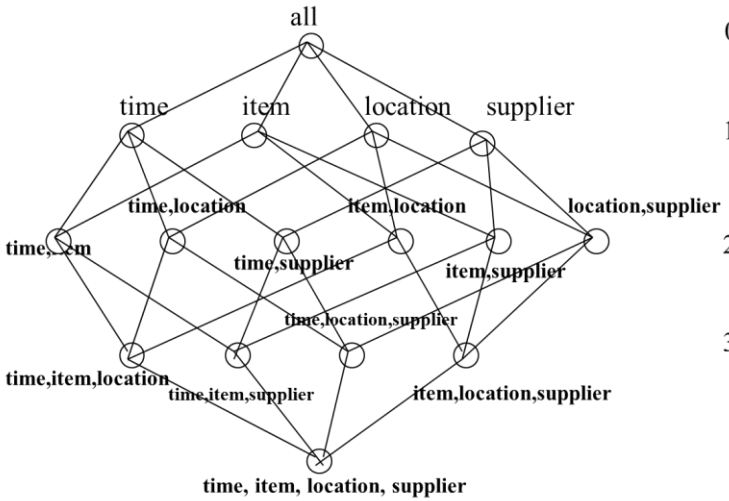
- Sales volume as a function of product, month, and region



Dimensions: *Product, Location, Time*
Hierarchical summarization paths



How Many Cuboids in a Data Cube with a Concept Hierarchy?



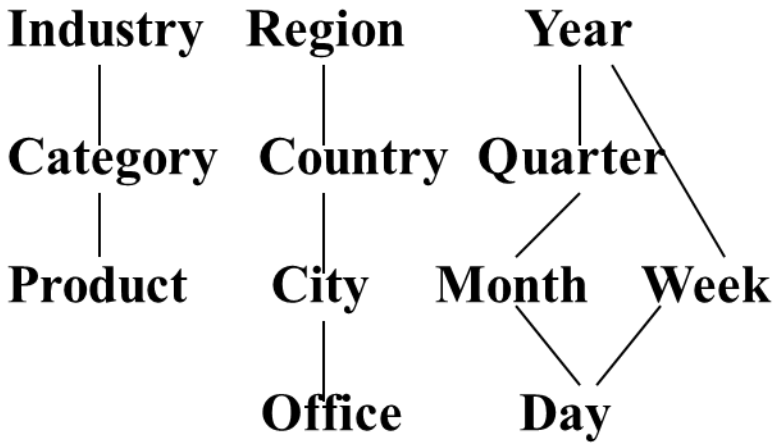
0-D (*apex*) cuboid

1-D cuboids

2-D cuboids

3-D cuboids

4-D (*base*) cuboid

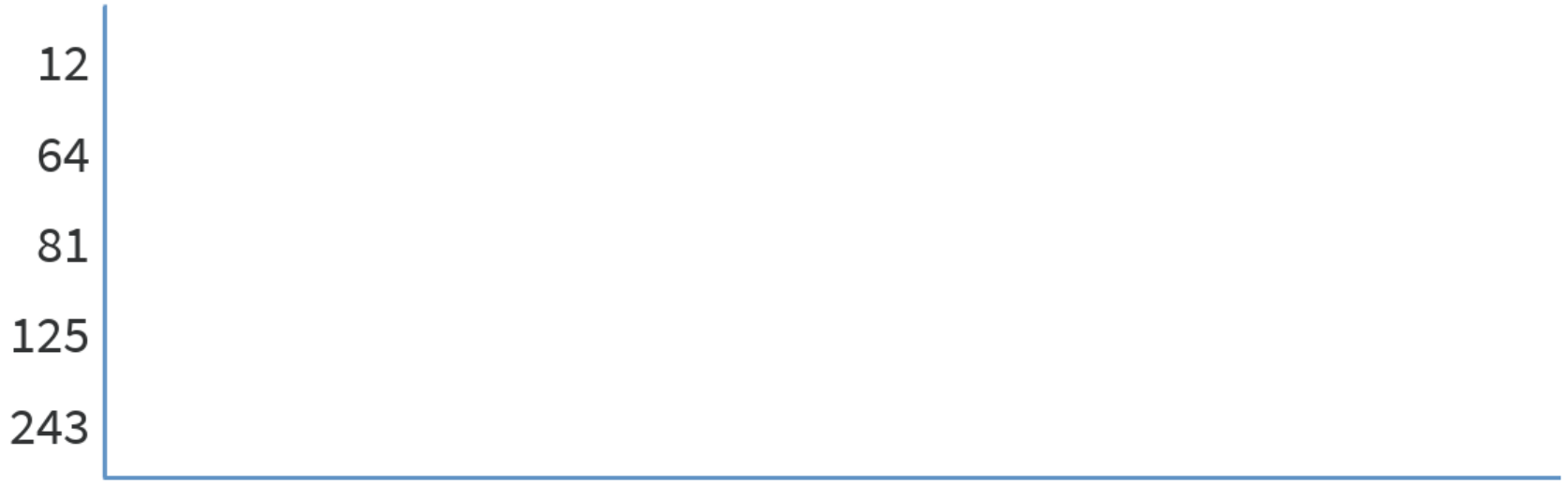


How many cuboids in an n-dimensional cube with L levels?

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

Imagine encoding each cuboid with an array, e.g. *quarter, category, country = [4, 2, 3, 0].* The +1 is for "all", i.e. removing the dimension.

How many cuboids are there in a data cube with 3 dimensions (e.g. time, item, location), and 4 levels of concept hierarchy per dimension (e.g. day, month, quarter, year)?



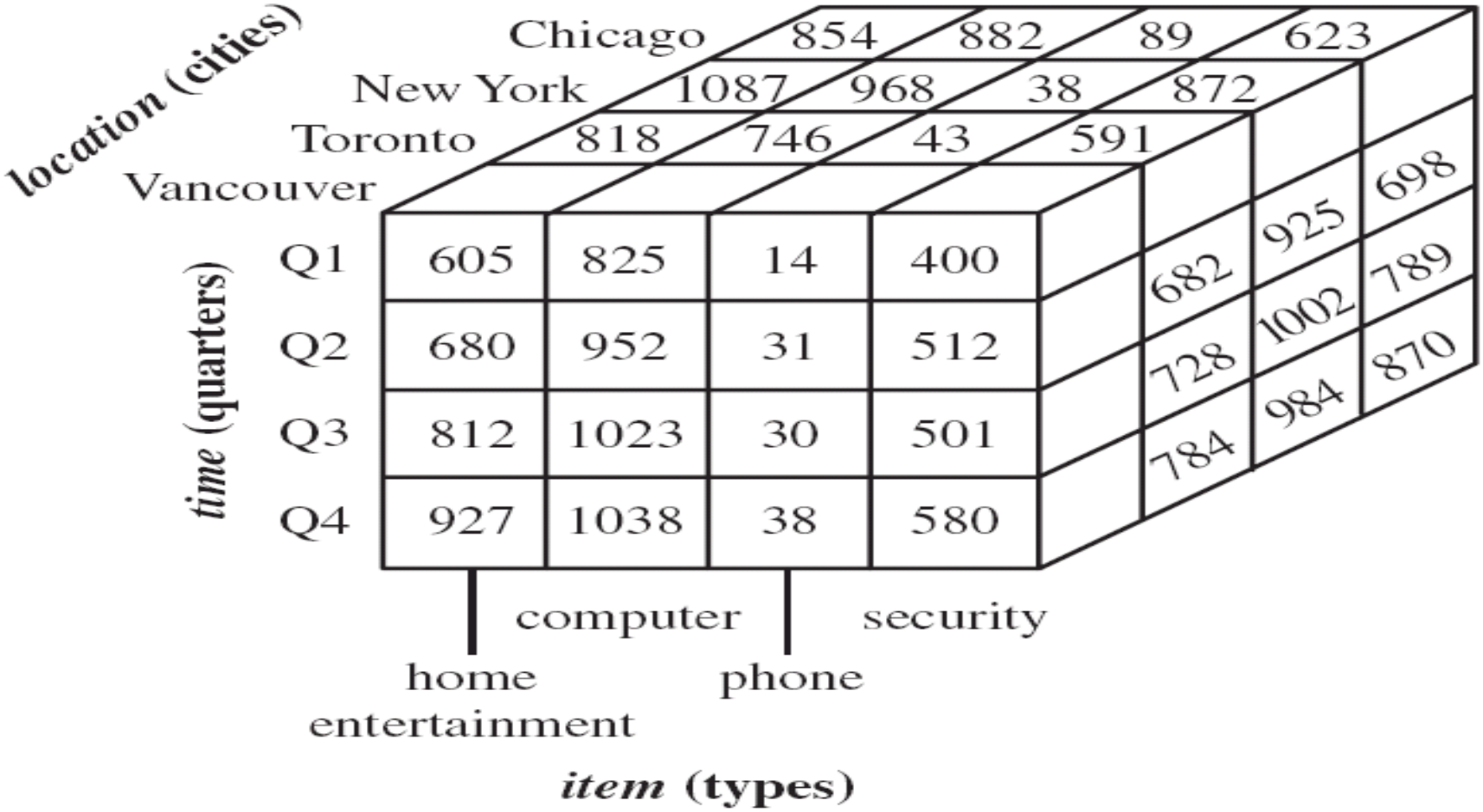
Efficient Data Cube Computation

- Materialization of data cube
  - 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.

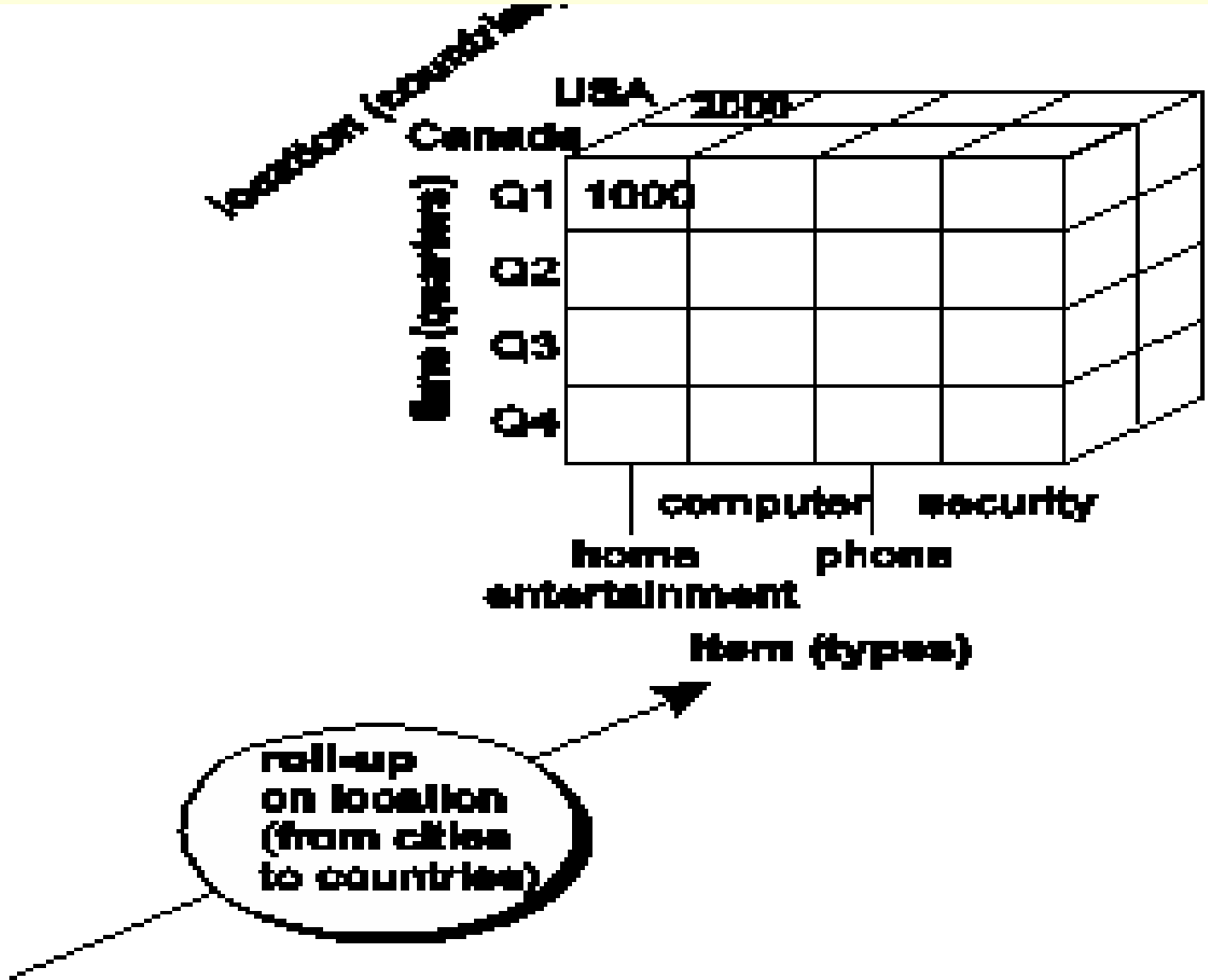
Typical OLAP Operations

- Roll up (drill-up): summarize data
  - *by **climbing up hierarchy** or by **dimension reduction***
- Drill down (roll down): reverse of roll-up
  - *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*

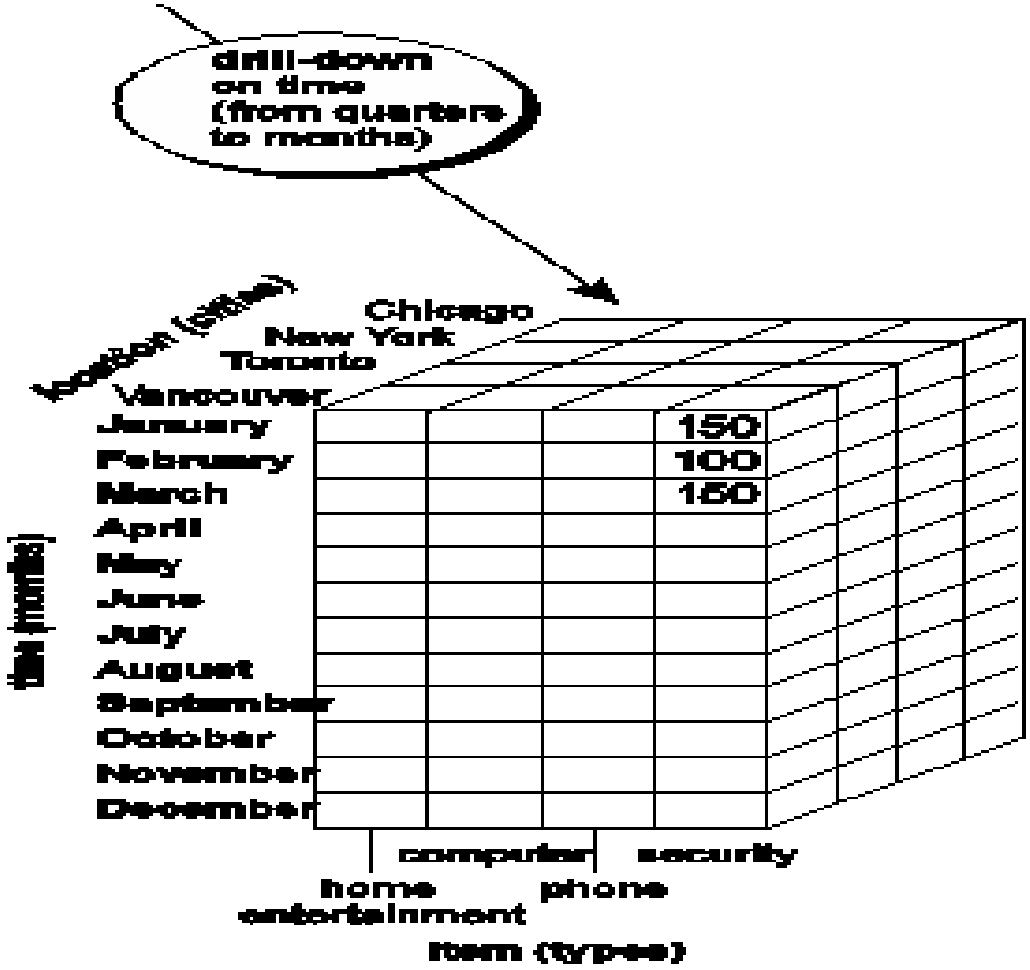
Multidimensional view -3D



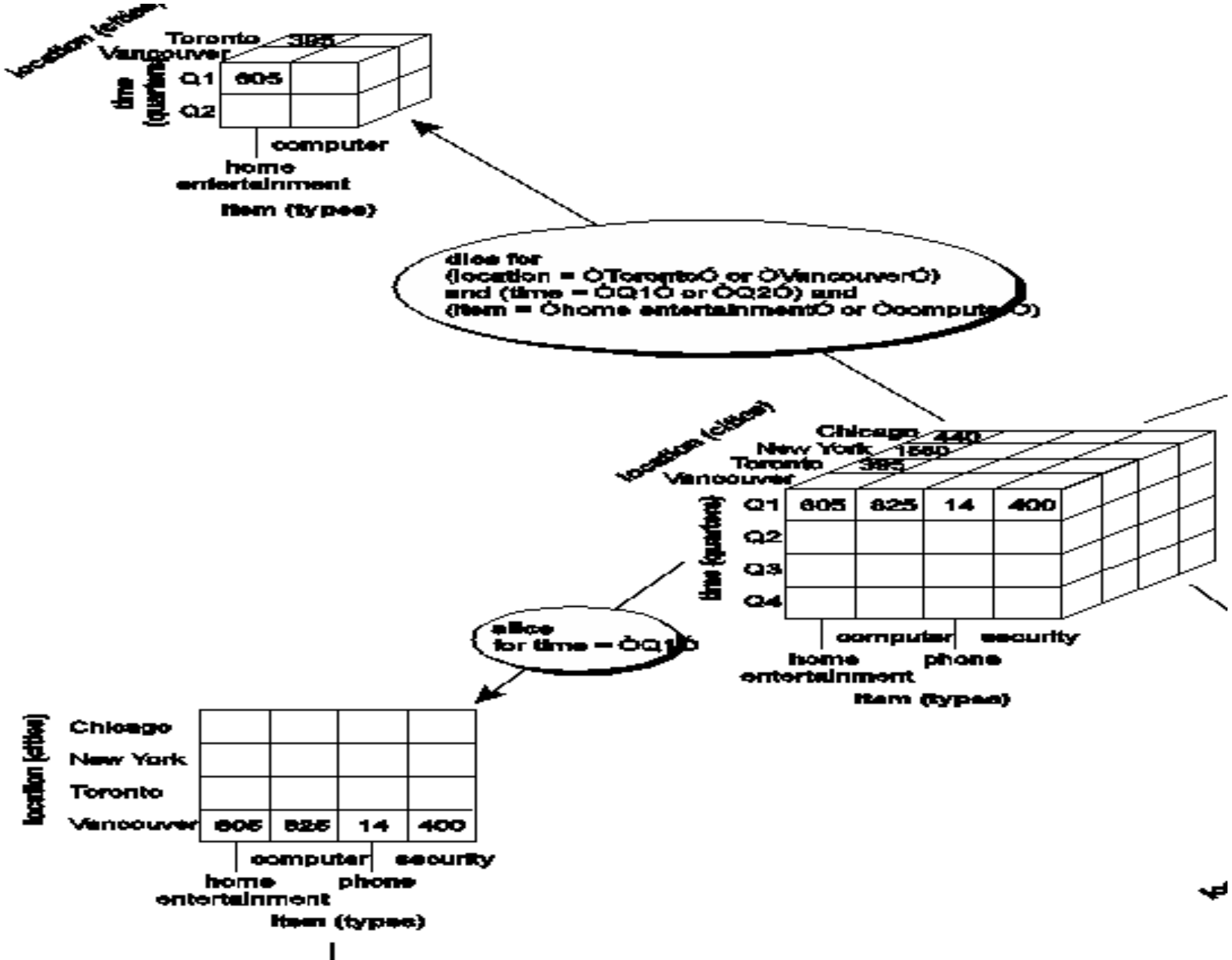
- Roll up (drill-up): summarize data
  - *by **climbing up hierarchy** or by **dimension reduction***



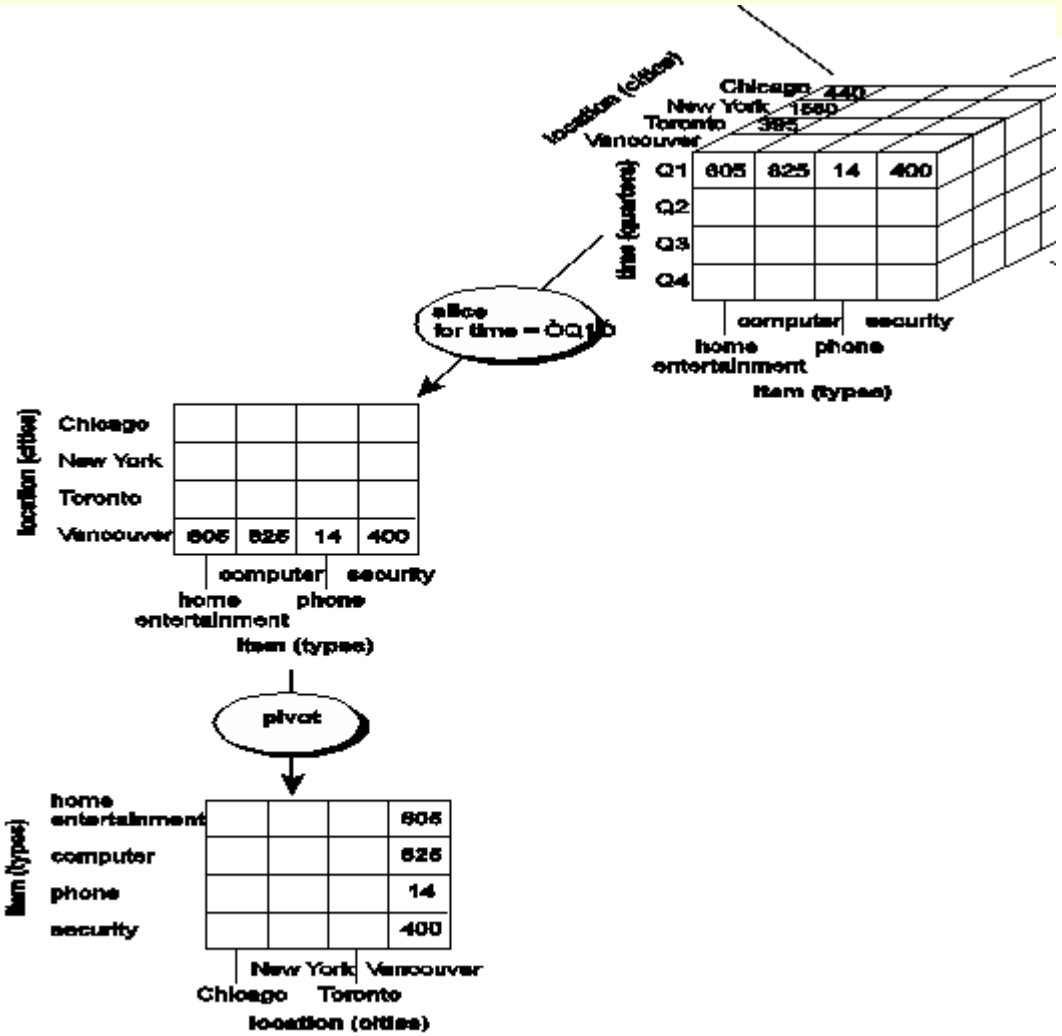
- Drill down (roll down): reverse of roll-up
  - *from higher level summary* ***to lower level*** *summary or detailed data, or* ***introducing new dimensions***



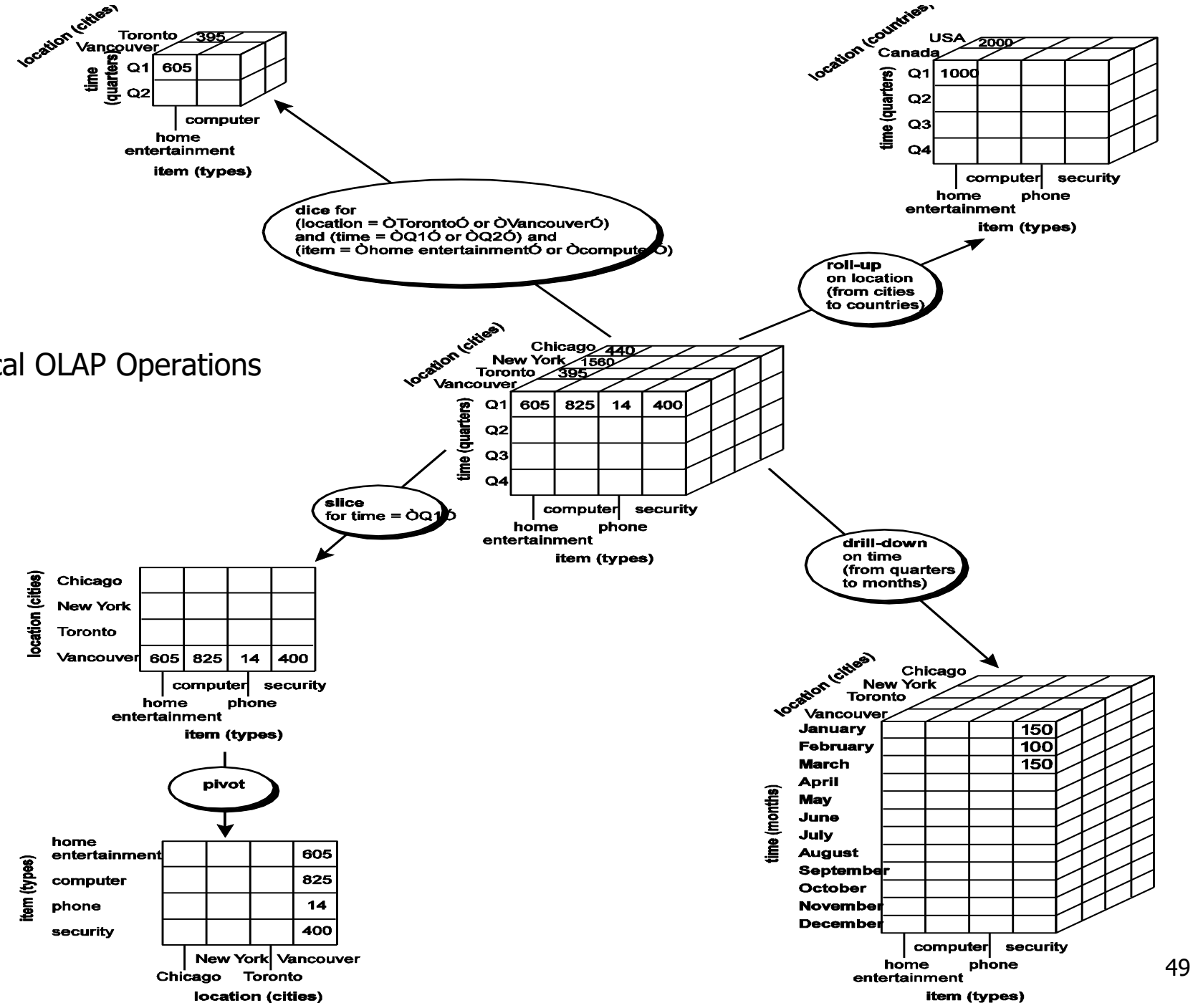
Slice and dice: *project and select*



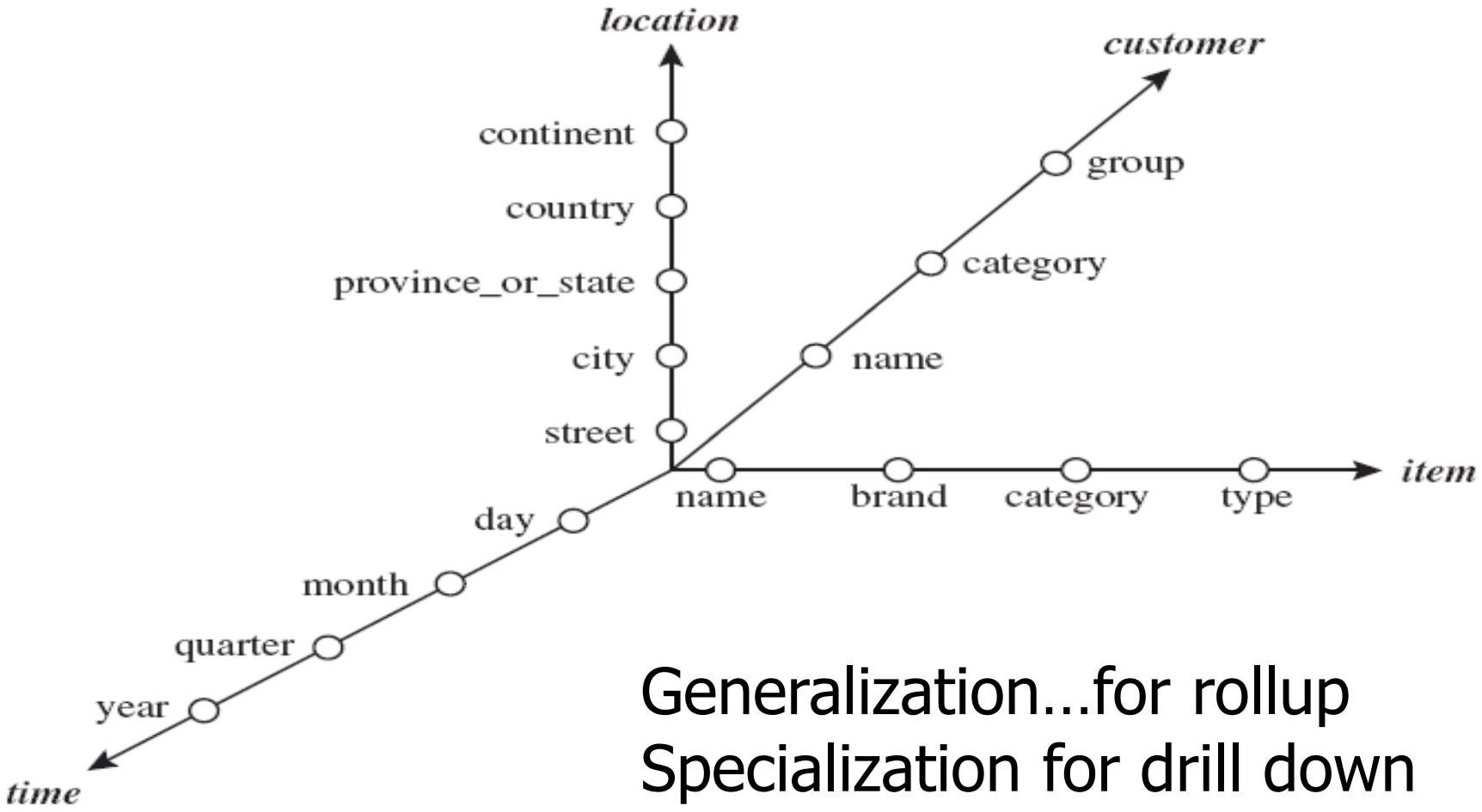
- Pivot (rotate):
  - *reorient the cube, visualization, 3D to series of 2D planes*



Typical OLAP Operations



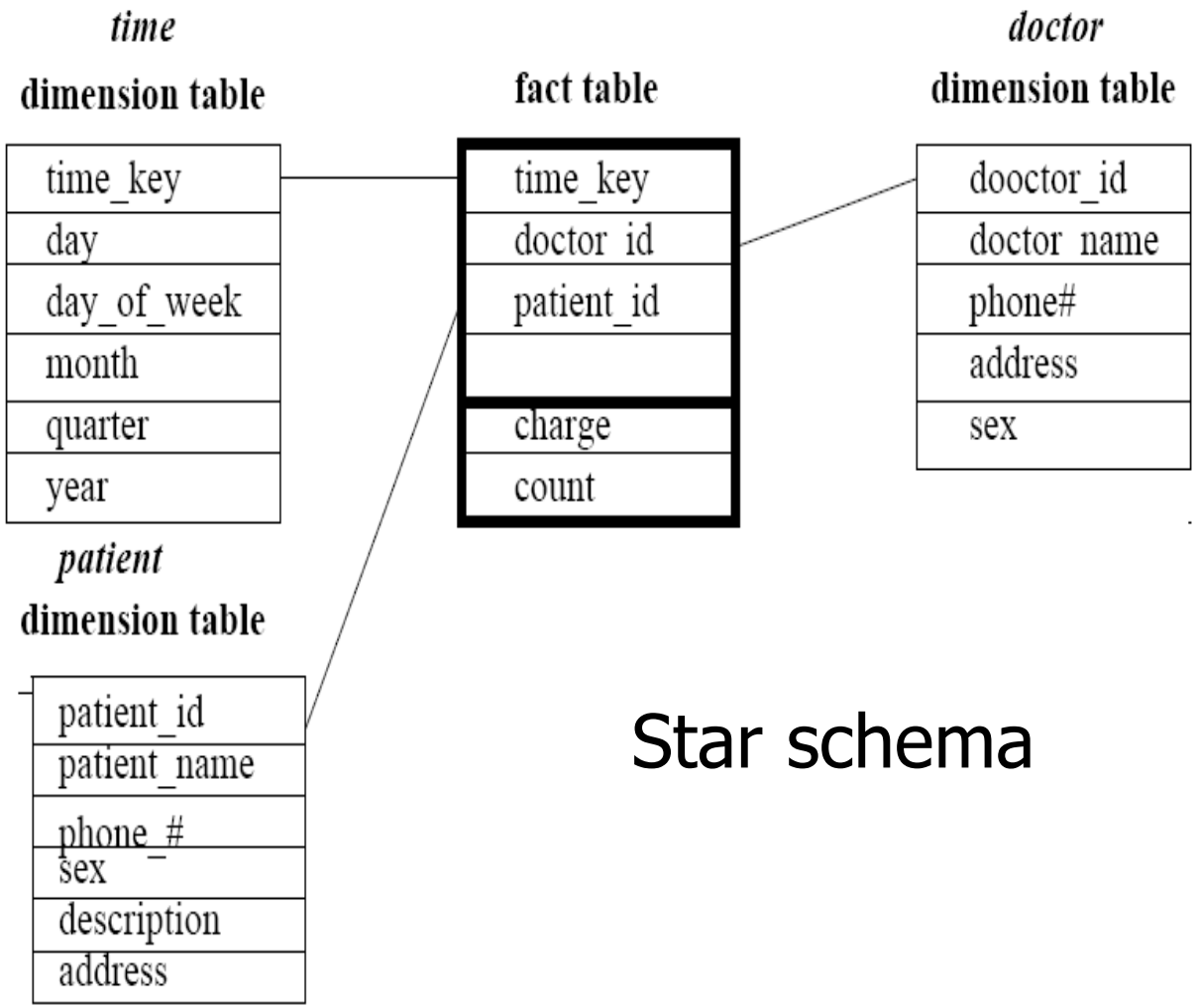
Starnet Query model



Generalization...for rollup
Specialization for drill down

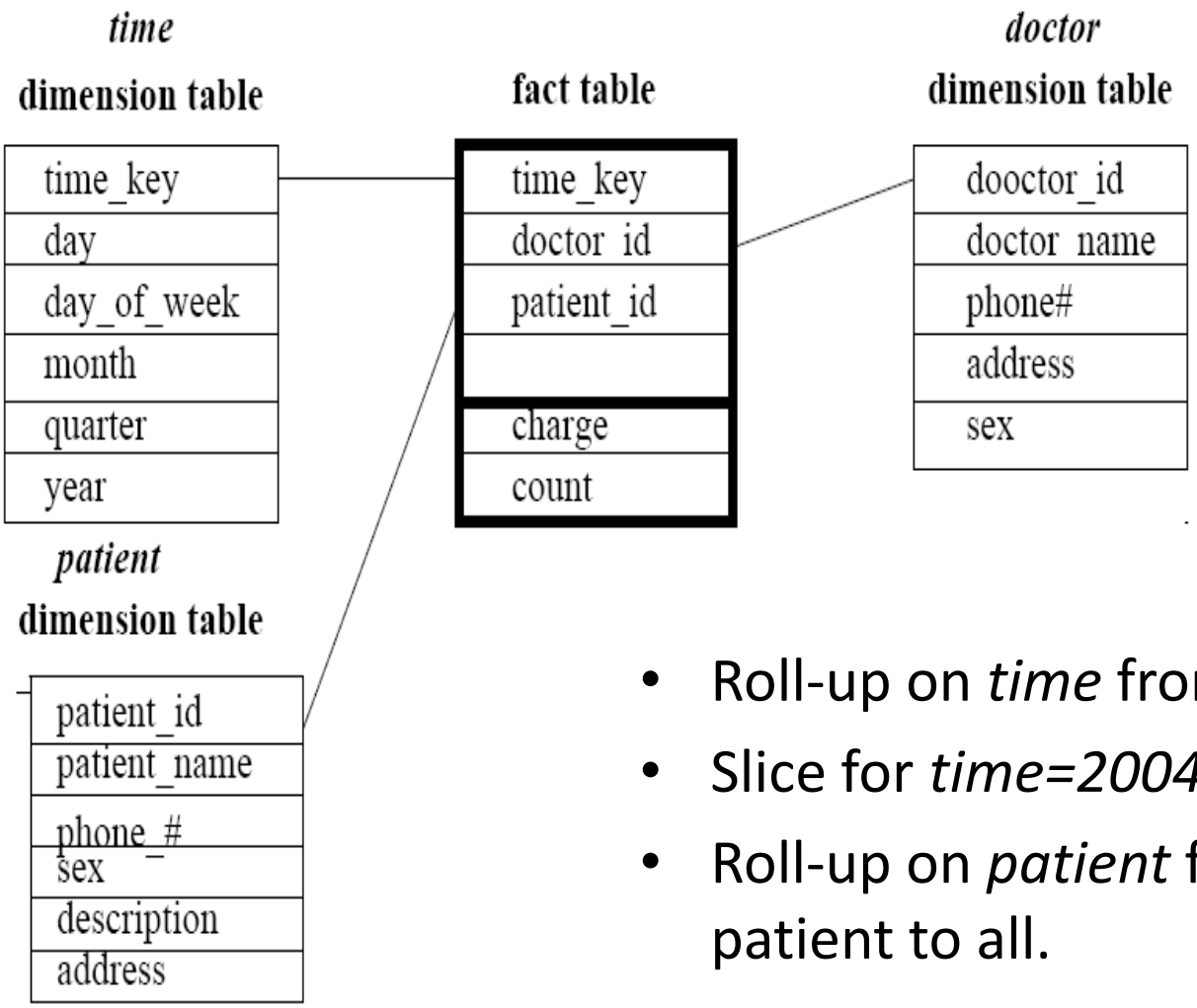
Some examples

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.



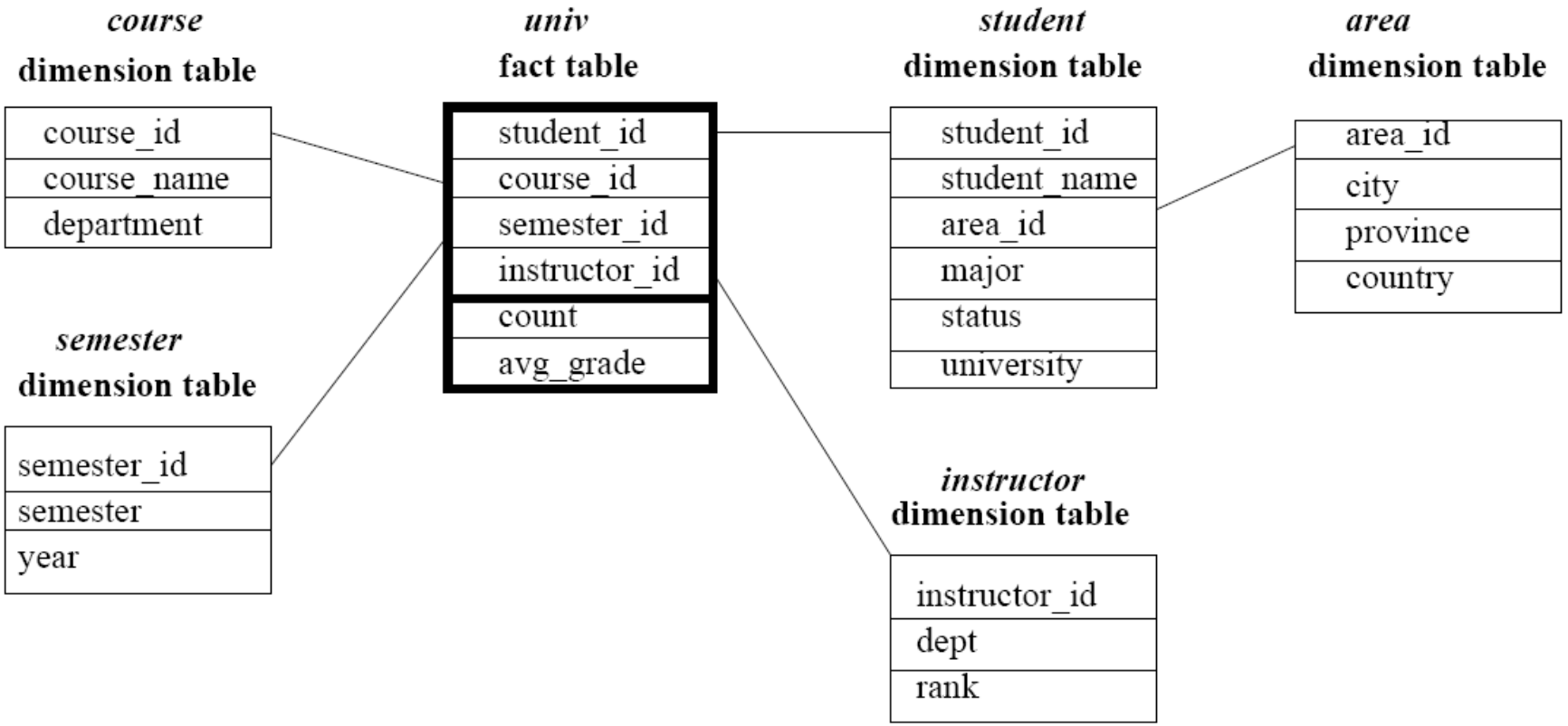
Star schema

Starting with the base cuboid [*day*, *doctor*, *patient*], what specific *OLAP operations* should be performed in order to list the total fee collected by each doctor in 2004?

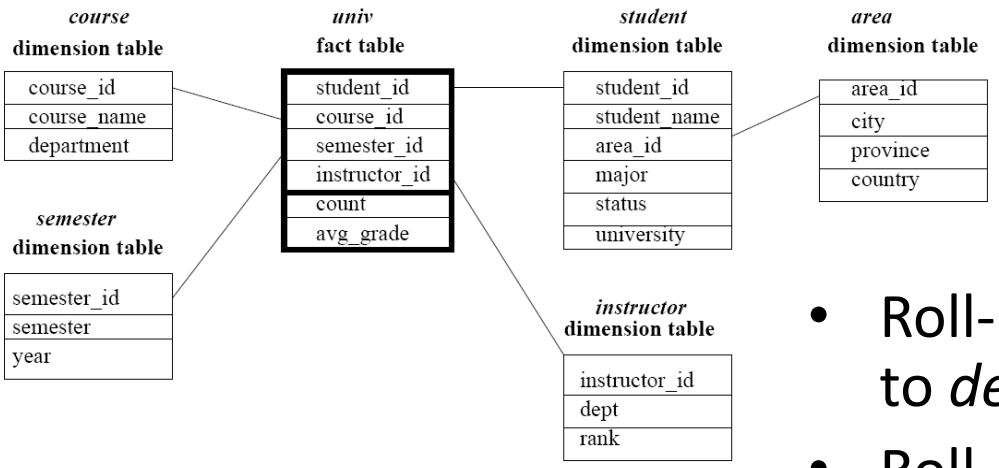


- Roll-up on *time* from *day* to *year*.
- Slice for *time=2004*.
- Roll-up on *patient* from individual patient to all.

(b) 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* courses for each *Big University* student.



(b) 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* courses for each *Big University* student.



- Roll-up on course from *course id* to *department*.
- Roll-up on student from *student id* to *university*.
- Roll-up on semester to *all*
- Dice on course, student with *department=\CS"* and *university = \Big University"*.
- Drill-down on *student* from *university* to *student name*.

Think-Pair-Share: Weather Bureau Data Warehouse

- **Design a data warehouse for a regional weather bureau**. The weather bureau has about 1000 probes, which are scattered throughout various land and ocean locations in the region to collect basic weather data, including air pressure, temperature, and precipitation at each hour. Your design should facilitate efficient querying and online analytical processing, and derive general weather patterns in multidimensional space.
- In your design, consider the following, etc:
  - Fact table(s), dimension tables, schema, measures
  - Concept hierarchies
  - How will you use this data warehouse for data mining and decision support?

Project sharing

- Please get together with your group and discuss, for the next 5 minutes:
  - The goals of your project
  - Why it is important
  - What data you plan to analyze
  - What methods plan to use

- Each group will then report to the class

- If you don’t have a group yet, see me and we’ll see about matching you up.

References (I)

- S. Agarwal, R. Agrawal, P. M. Deshpande, A. Gupta, J. F. Naughton, R. Ramakrishnan, and S. Sarawagi. On the computation of multidimensional aggregates. VLDB'96
- D. Agrawal, A. E. Abbadi, A. Singh, and T. Yurek. Efficient view maintenance in data warehouses. SIGMOD'97
- R. Agrawal, A. Gupta, and S. Sarawagi. Modeling multidimensional databases. ICDE'97
- S. Chaudhuri and U. Dayal. An overview of data warehousing and OLAP technology. ACM SIGMOD Record, 26:65-74, 1997
- E. F. Codd, S. B. Codd, and C. T. Salley. Beyond decision support. Computer World, 27, July 1993.
- J. Gray, et al. Data cube: A relational aggregation operator generalizing group-by, cross-tab and sub-totals. Data Mining and Knowledge Discovery, 1:29-54, 1997.
- A. Gupta and I. S. Mumick. Materialized Views: Techniques, Implementations, and Applications. MIT Press, 1999.
- J. Han. Towards on-line analytical mining in large databases. *ACM SIGMOD Record*, 27:97-107, 1998.
- V. Harinarayan, A. Rajaraman, and J. D. Ullman. Implementing data cubes efficiently. SIGMOD'96

References (II)

- C. Imhoff, N. Galemmo, and J. G. Geiger. Mastering Data Warehouse Design: Relational and Dimensional Techniques. John Wiley, 2003
- W. H. Inmon. Building the Data Warehouse. John Wiley, 1996
- R. Kimball and M. Ross. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. 2ed. John Wiley, 2002
- P. O'Neil and D. Quass. Improved query performance with variant indexes. SIGMOD'97
- Microsoft. OLEDB for OLAP programmer's reference version 1.0. In http://www.microsoft.com/data/oledb/olap, 1998
- A. Shoshani. OLAP and statistical databases: Similarities and differences. PODS'00.
- S. Sarawagi and M. Stonebraker. Efficient organization of large multidimensional arrays. ICDE'94
- OLAP council. MDAPI specification version 2.0. In http://www.olapcouncil.org/research/apily.htm, 1998
- E. Thomsen. OLAP Solutions: Building Multidimensional Information Systems. John Wiley, 1997
- P. Valduriez. Join indices. ACM Trans. Database Systems, 12:218-246, 1987.
- J. Widom. Research problems in data warehousing. CIKM'95.