Data Mining:

Concepts and Techniques

(3rd ed.)

— Chapter 4 —

Slides Courtesy of Textbook

Chapter 4:
Data Warehousing and On-line Analytical Processing

- Data Warehouse: Basic Concepts
- Data Warehouse Modeling: Data Cube and OLAP
- Data Warehouse Usage
- Data Warehouse Implementation
- Summary

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

- Constructed for Subjects: Organized around major subjects, such as customer, product
- Constructed for Decision Making: Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing
- Constructed to be Simple and Concise: 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

- Data Integration: Constructed by integrating multiple, heterogeneous data sources
  - relational databases, flat files, on-line transaction records
- Data Preprocessing: 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

- Historic Perspective: 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)
- Time Related: 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"

OLAP : Online Analytic Process
OLTP: Transaction

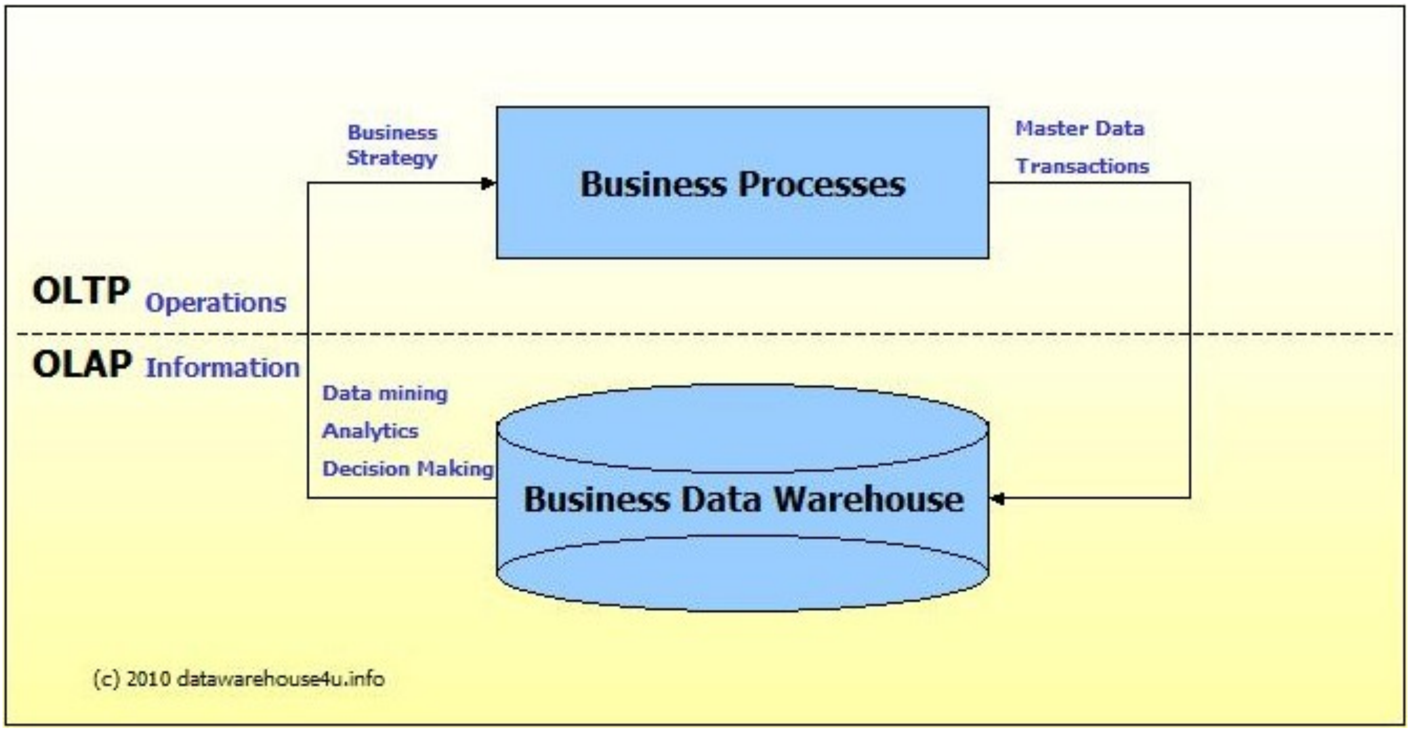
Data Warehouse—Nonvolatile

- Independence: A physically separate store of data transformed from the operational environment. Keep in high performance for both systems (OLTP vs. OLAP)
- Static Status: 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*

OLTP vs. OLAP

- transactional (OLTP) and analytical (OLAP)

Front



Backend

OLTP vs. OLAP

CS411: Transaction update

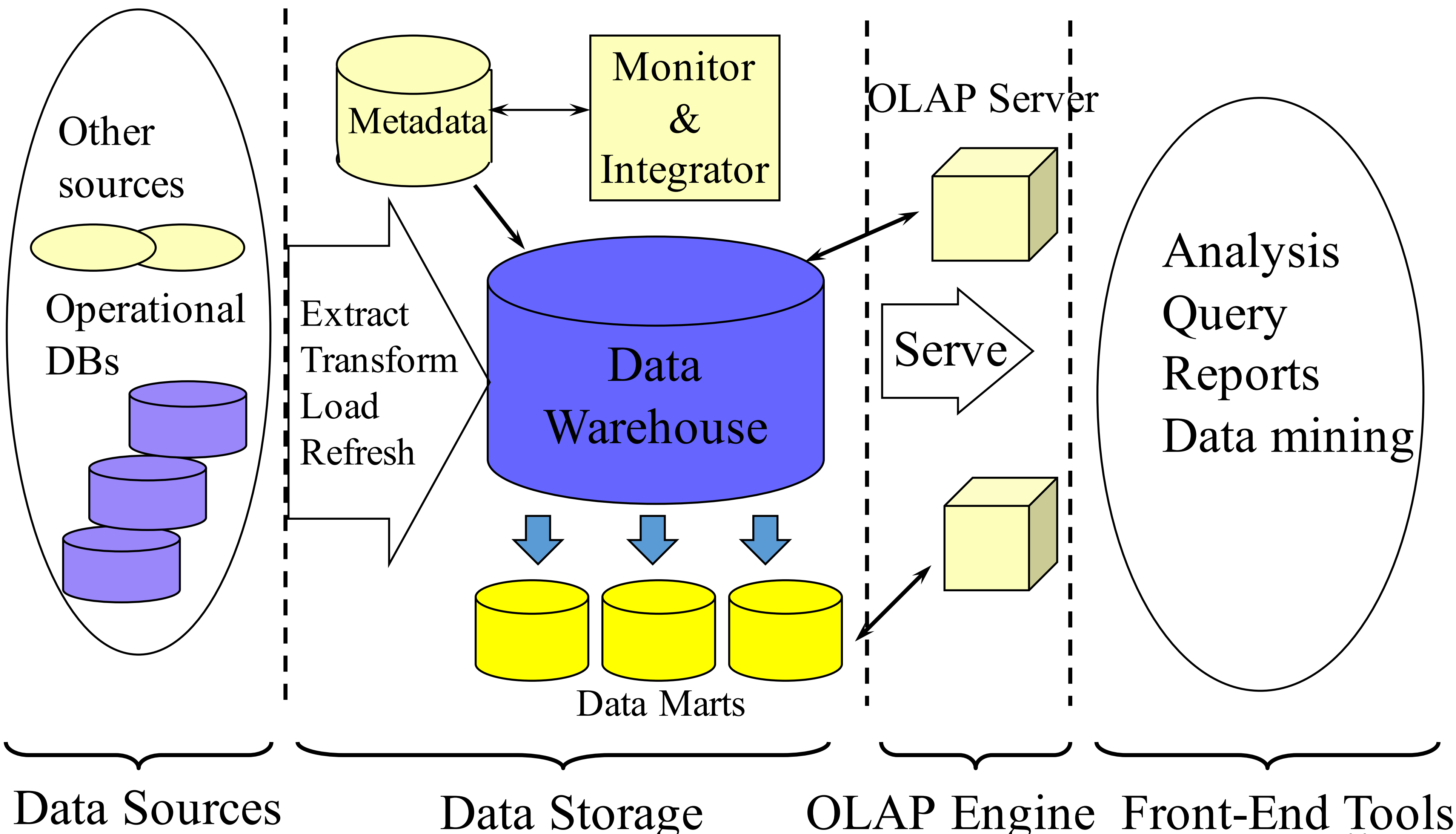
cs412: Analysis Read

	OLTP	**OLAP**
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
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
DB size	100MB-GB	100GB-TB

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: A Multi-Tiered Architecture



Chapter 4:
Data Warehousing and On-line Analytical Processing

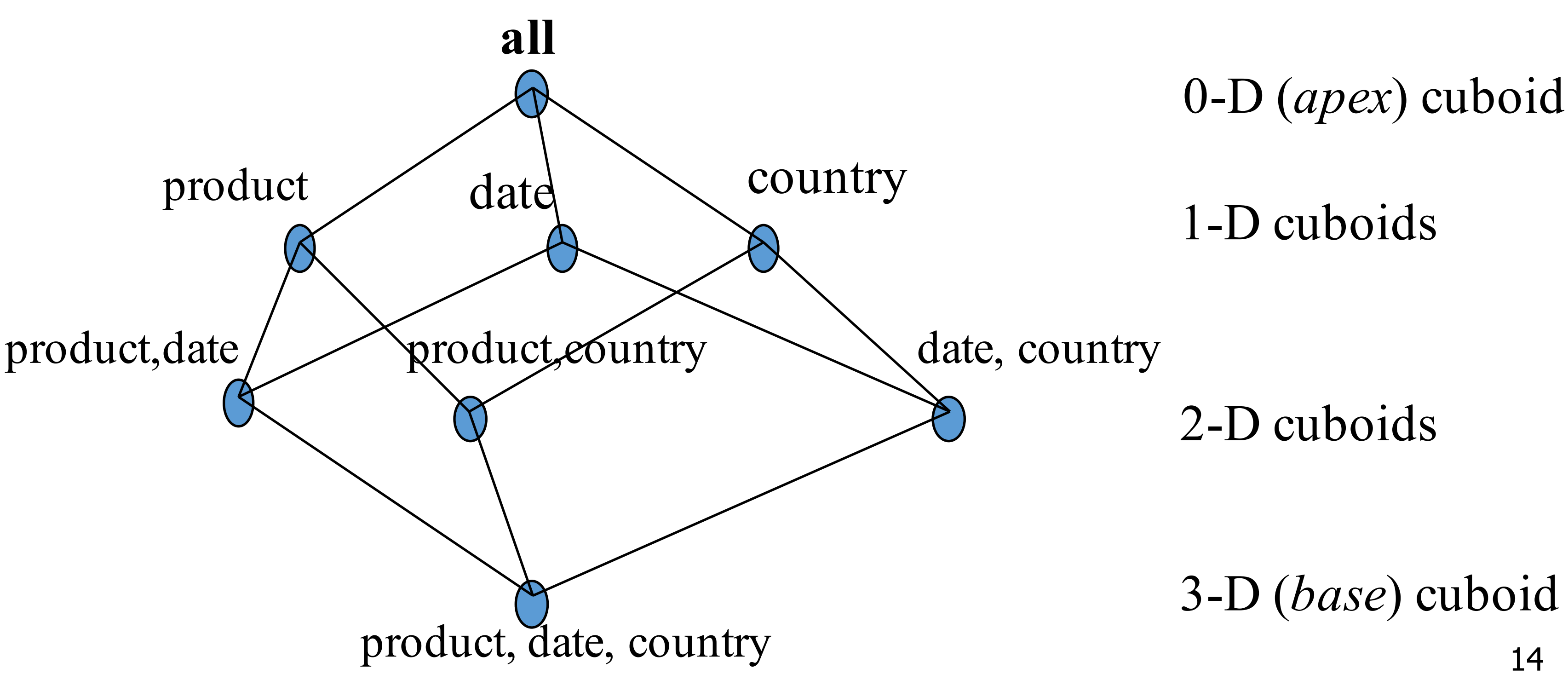
- Data Warehouse: Basic Concepts
- Data Warehouse Modeling: Data Cube and OLAP
- Data Warehouse Usage
- Data Warehouse Implementation
- Summary

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** is a multidimensional generalization of data spreadsheet.
- A data cube, such as sales, allows data to be modeled and viewed in multiple dimensions
  - **Fact table** contains **measures** (such as dollars_sold) and keys to each of the related dimension tables
  - **Dimension tables**, such as item (item_name, brand, type), or time(day, week, month, quarter, year)

Data Cuboid

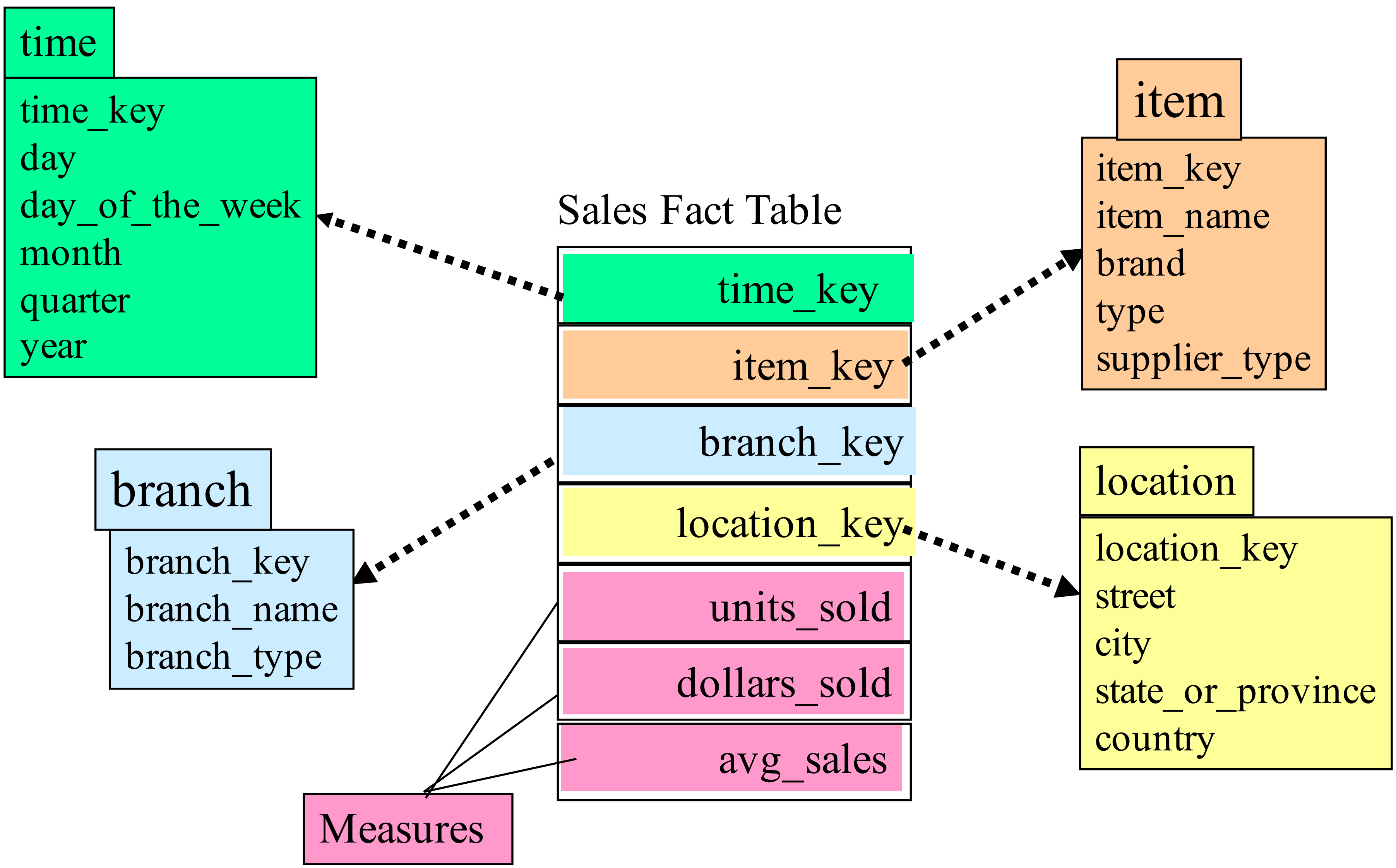
- A **data cuboid** is a subset of data cube.
- In data warehousing literature, an n-D base cube is called a base cuboid. The top most 0-D cuboid, which holds the highest-level of summarization, is called the apex cuboid. The lattice of cuboids forms a data cube.



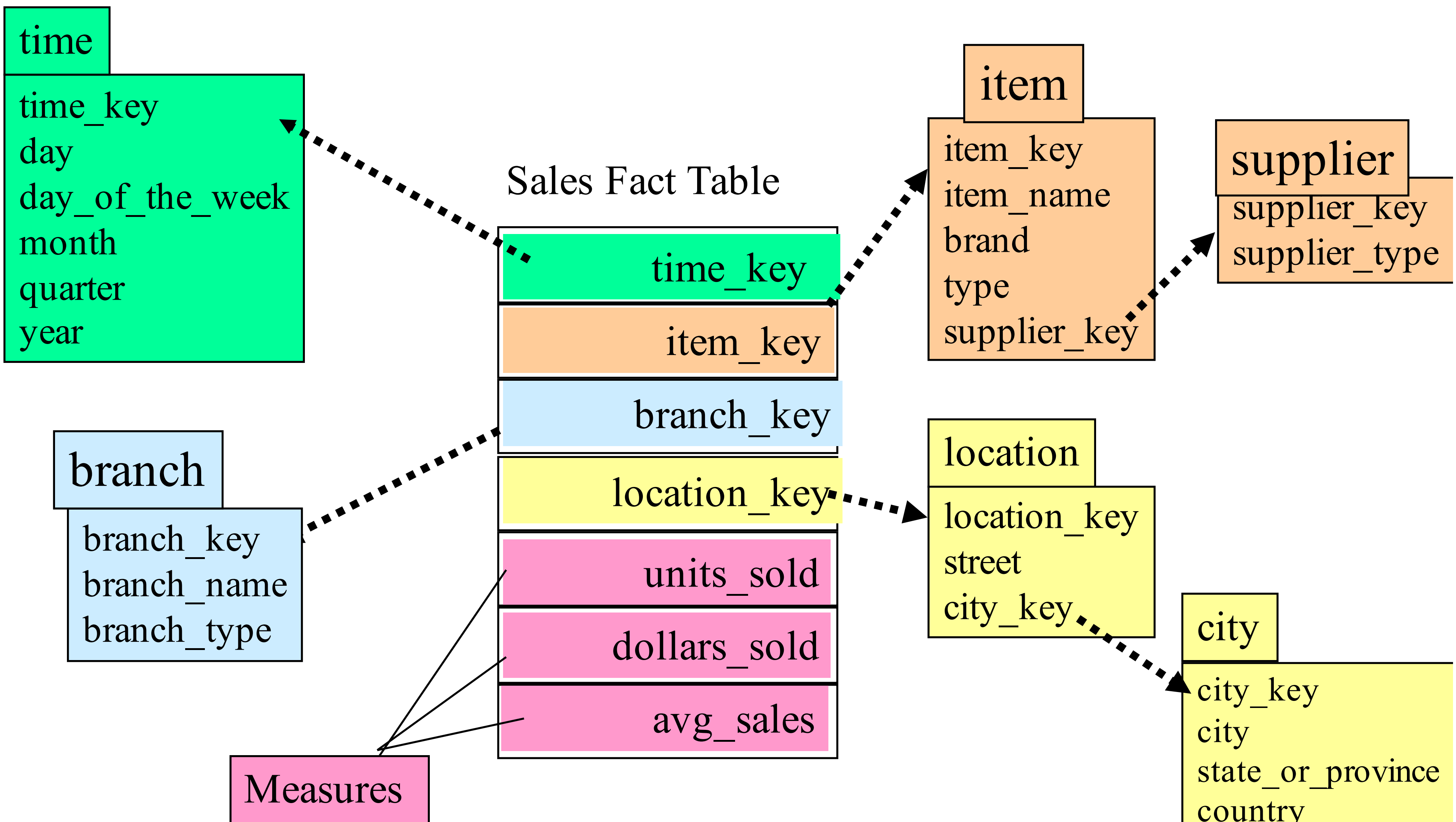
Conceptual Modeling of Data Warehouses

- 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

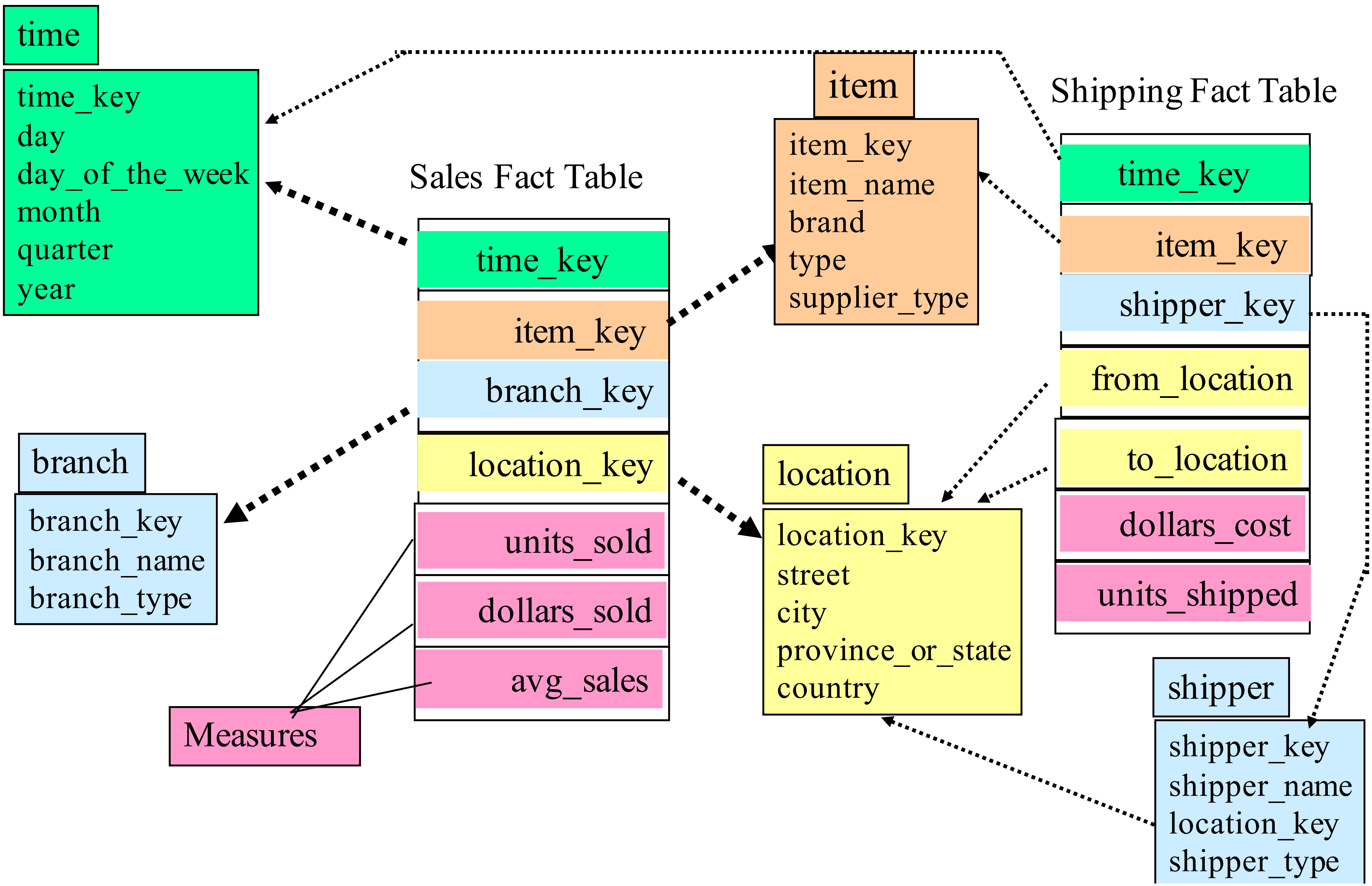
Example of **Star Schema**



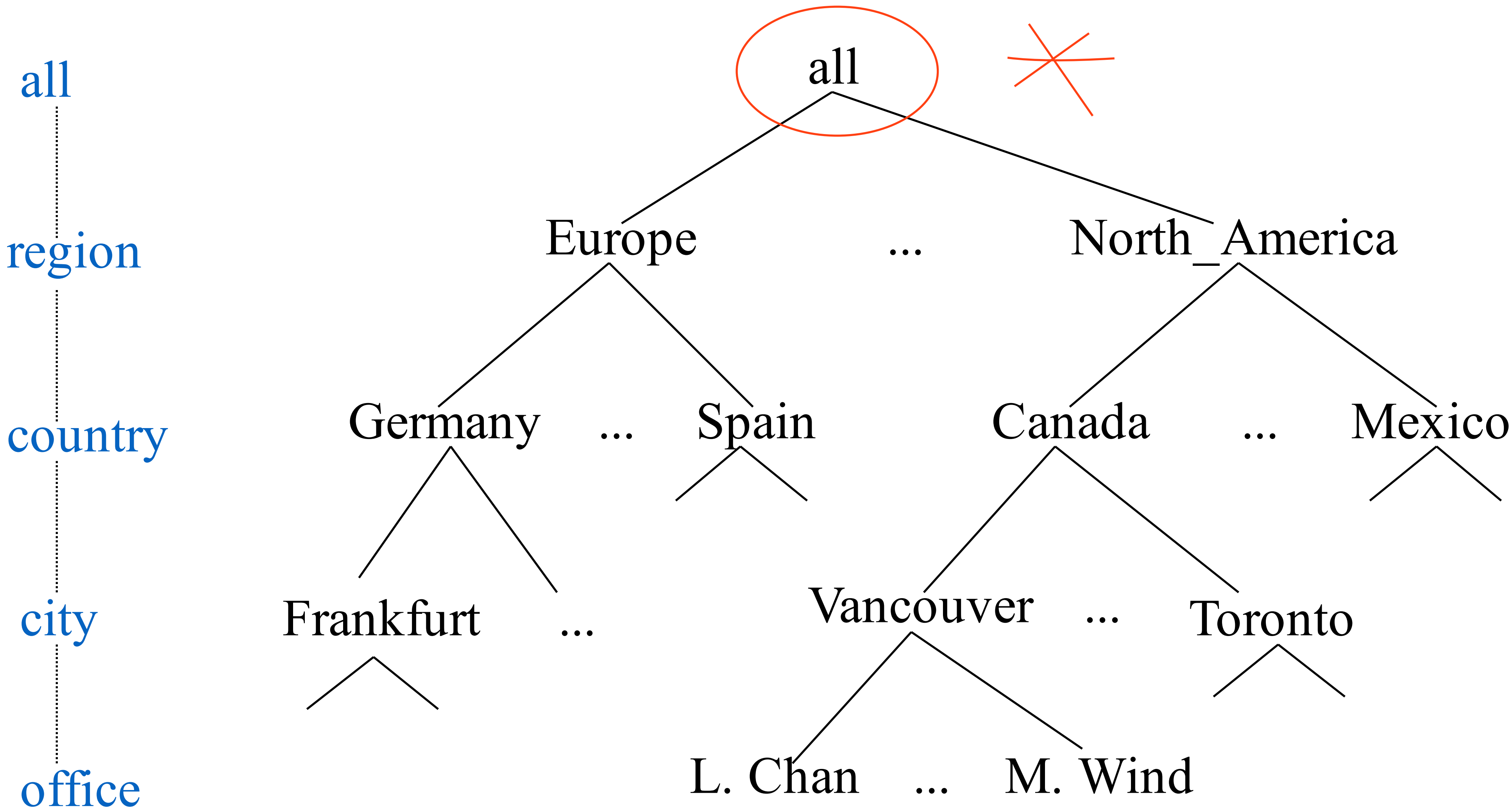
Example of **Snowflake Schema**



Example of **Fact Constellation**



A Concept Hierarchy:
Dimension (location)



Data Cube Measures: Three Categories

- Distributive: 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()
- 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(), Min_N()   [3.0,28] [3.2,40]   [GPA, count]
- Holistic: if there is no constant bound on the storage size needed to describe a subaggregate.
    - E.g., Median(), Mode()

Median 3.2

Median 3.1

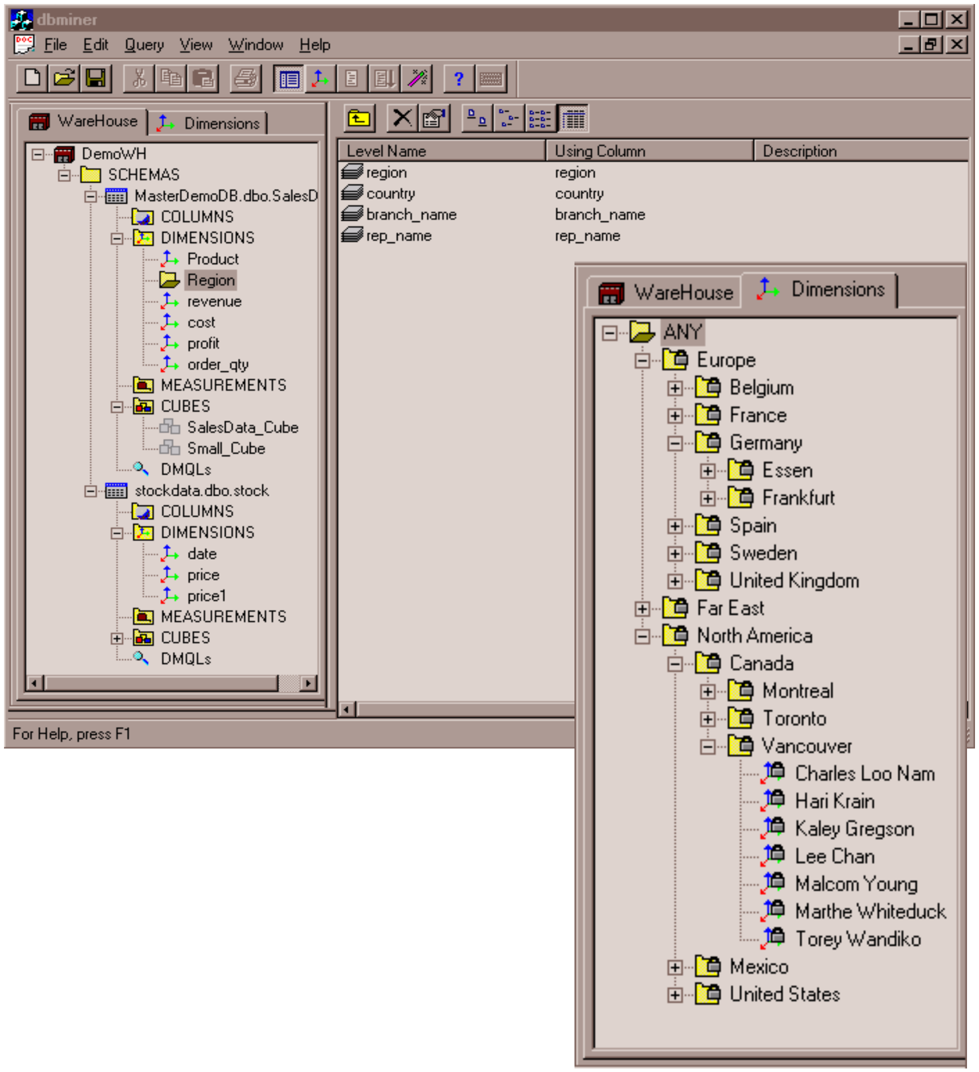
Data Cube Measures: Three Categories

- Evaluate functions below and tell whether they are distributive, algebraic or holistic.
    - Min(), Max()

- What about this?
    - Standard_Deviation()

- What about this?
    - Rank()

View of Warehouses and Hierarchies



Specification of hierarchies

- Schema hierarchy

  day

  < {month < quarter; week}

  < year

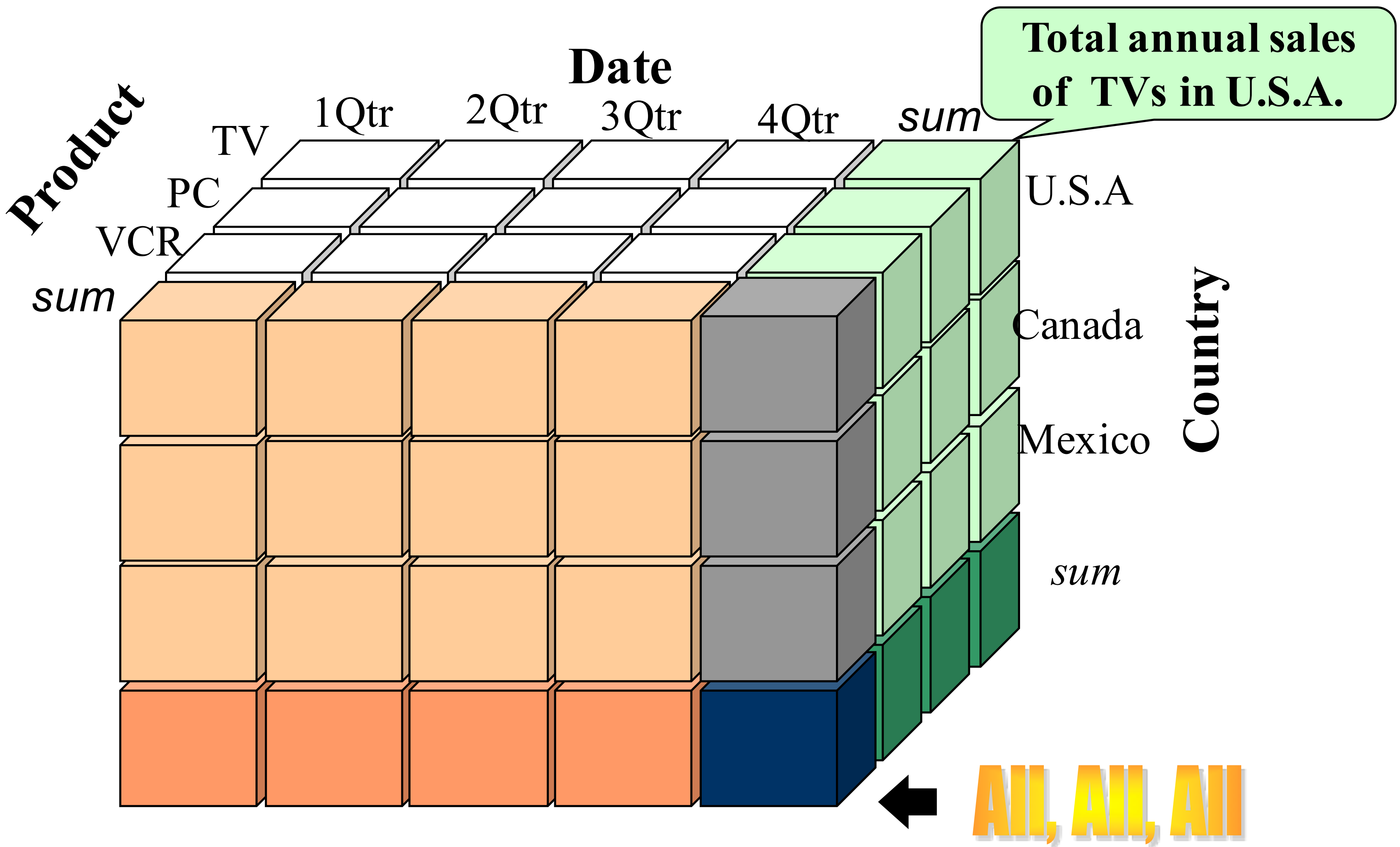
- Set_grouping hierarchy

  {1..10} < inexpensive

Power Pivot

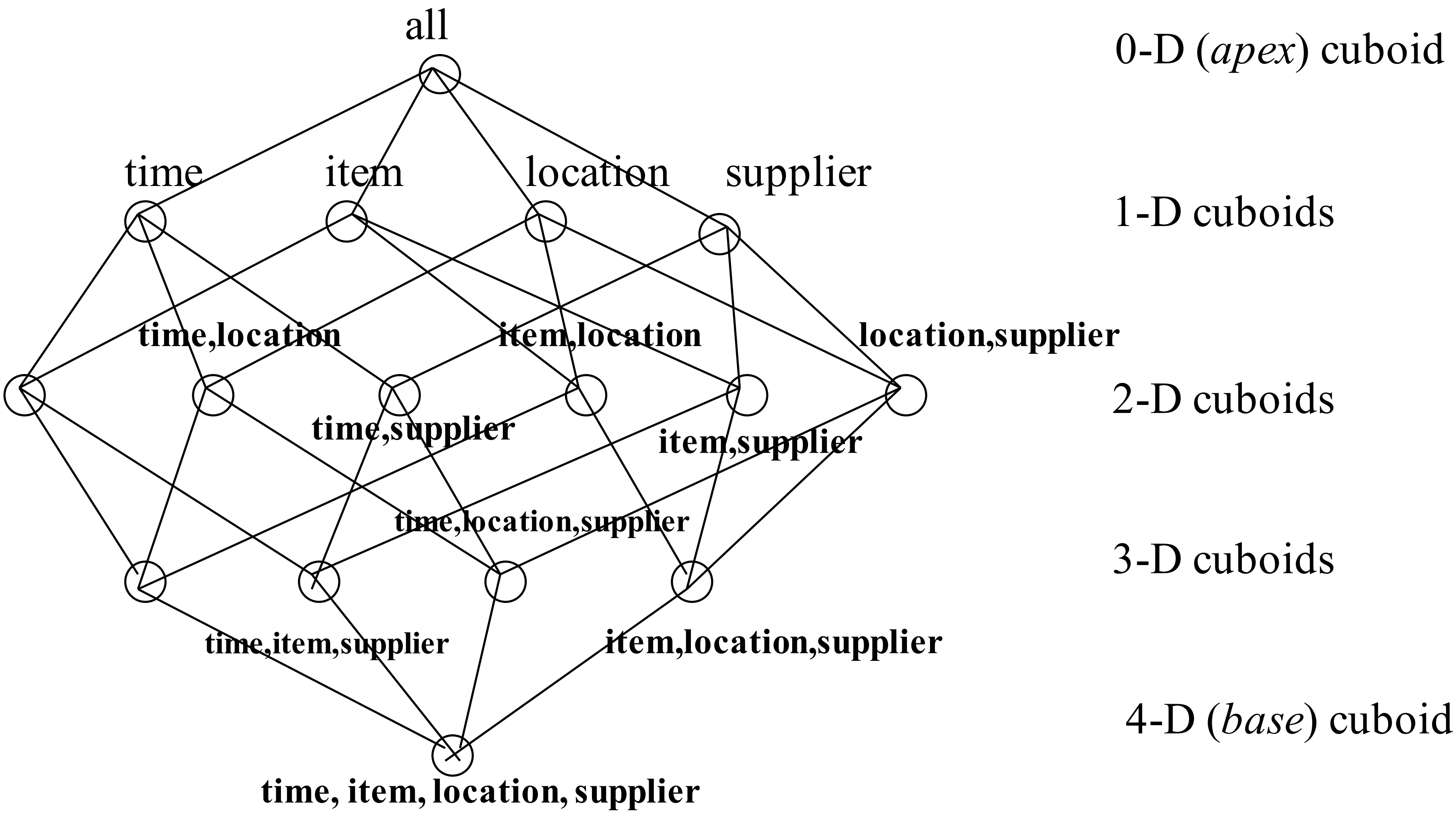
Where can you see such an interface? Excel?

A Sample Data Cube



3 (products) * 4 (date) * 3(countries) = 36

Cube: A Lattice of Cuboids

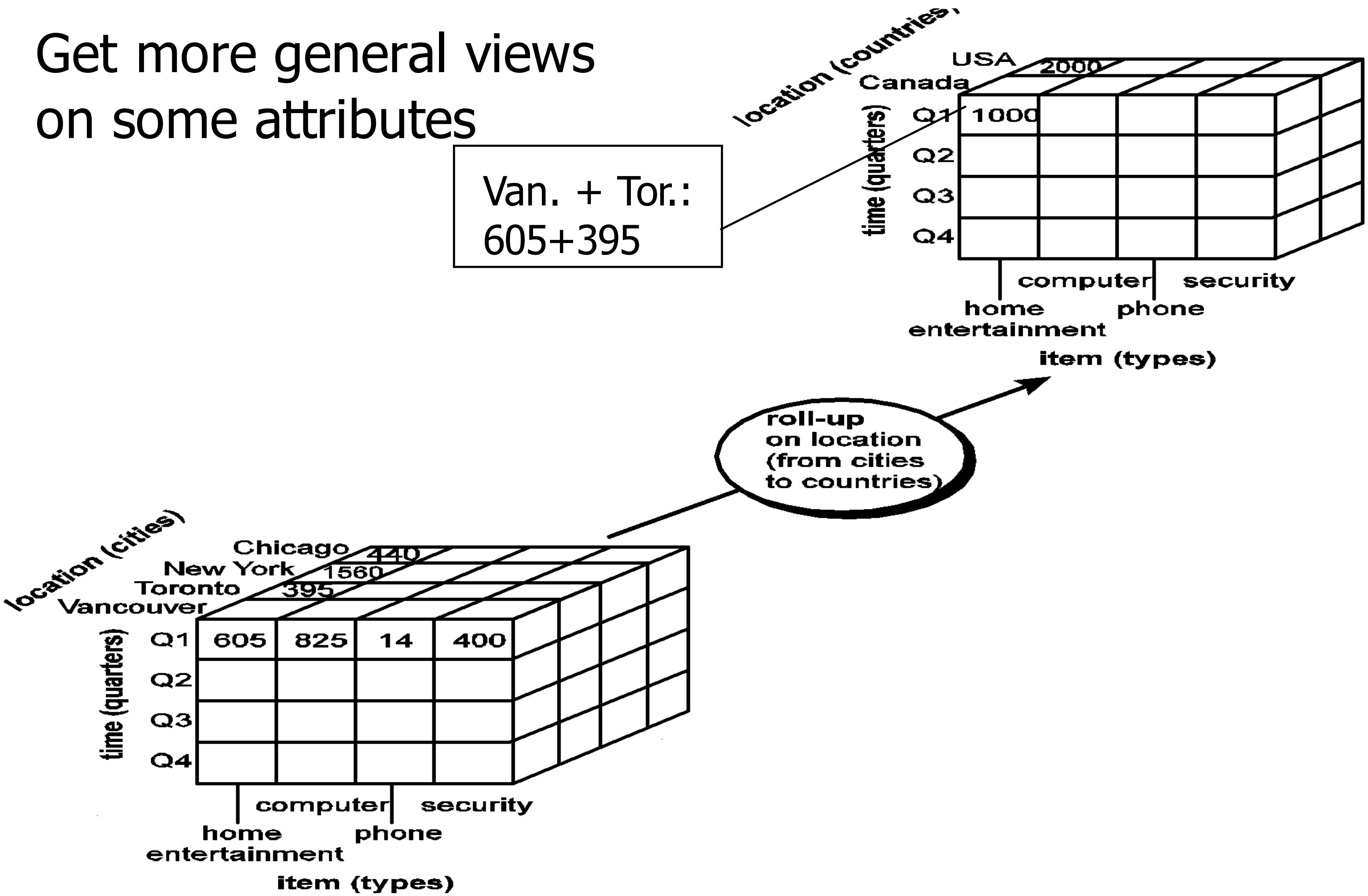


Typical OLAP Operations

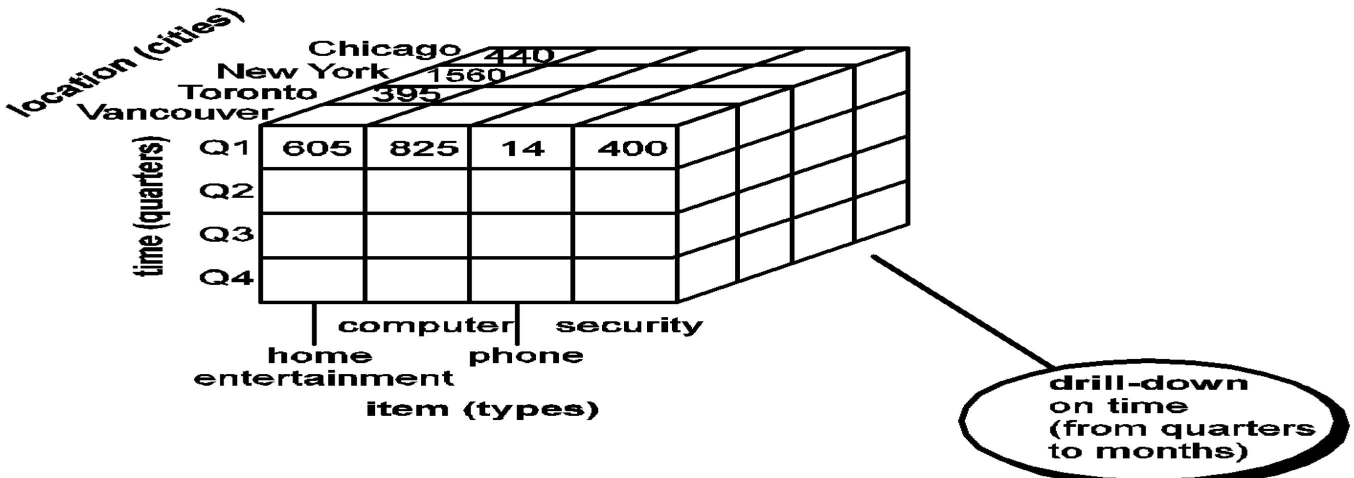
- Roll up (drill-up): summarize data
  - *by climbing up hierarchy or by dimension reduction*
- Drill down (roll down): reverse of roll-up
  - *from higher level summary to lower level summary or detailed data, or introducing new dimensions*
- Slice and dice: *project and select*
- Pivot (rotate):
  - *reorient the cube, visualization, 3D to series of 2D planes*
- Other operations
  - *drill across: involving (across) more than one fact table*
  - *drill through: through the bottom level of the cube to its back-end relational tables (using SQL)*

OLAP Operations: Roll up

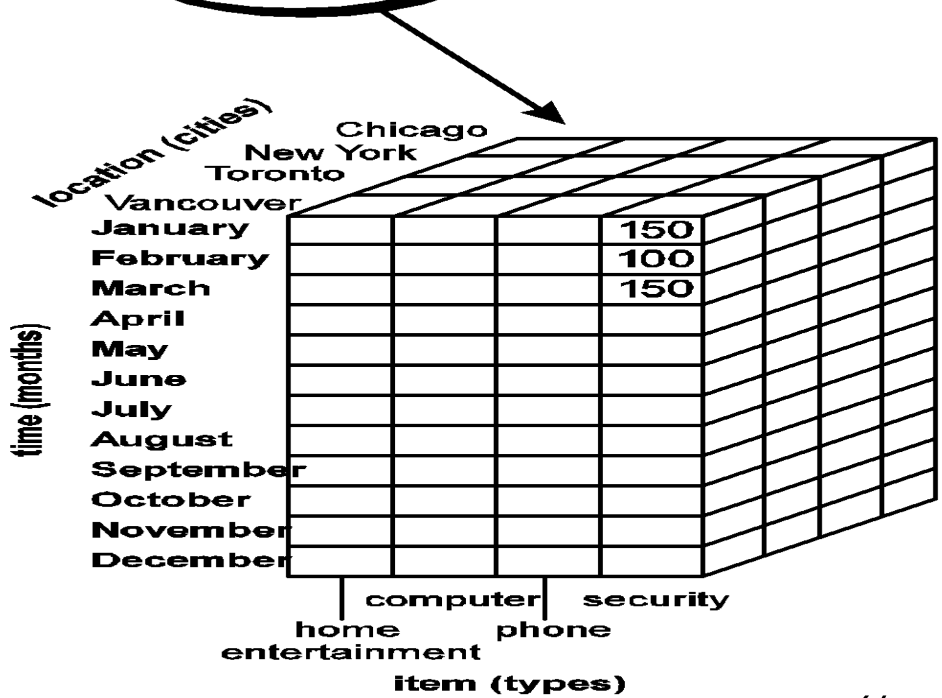
- Get more general views on some attributes



OLAP Operations: Drill down

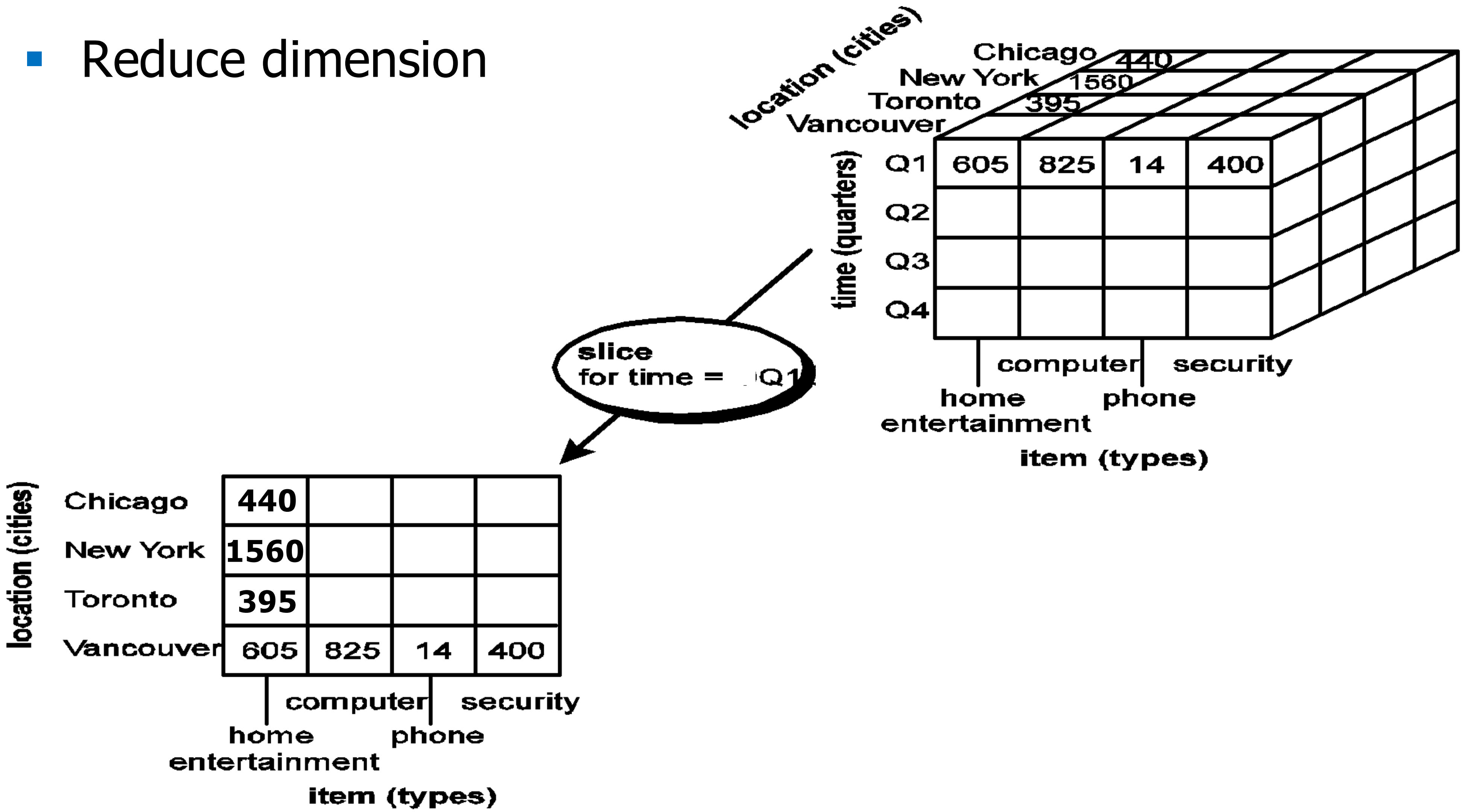


- Get more detailed data

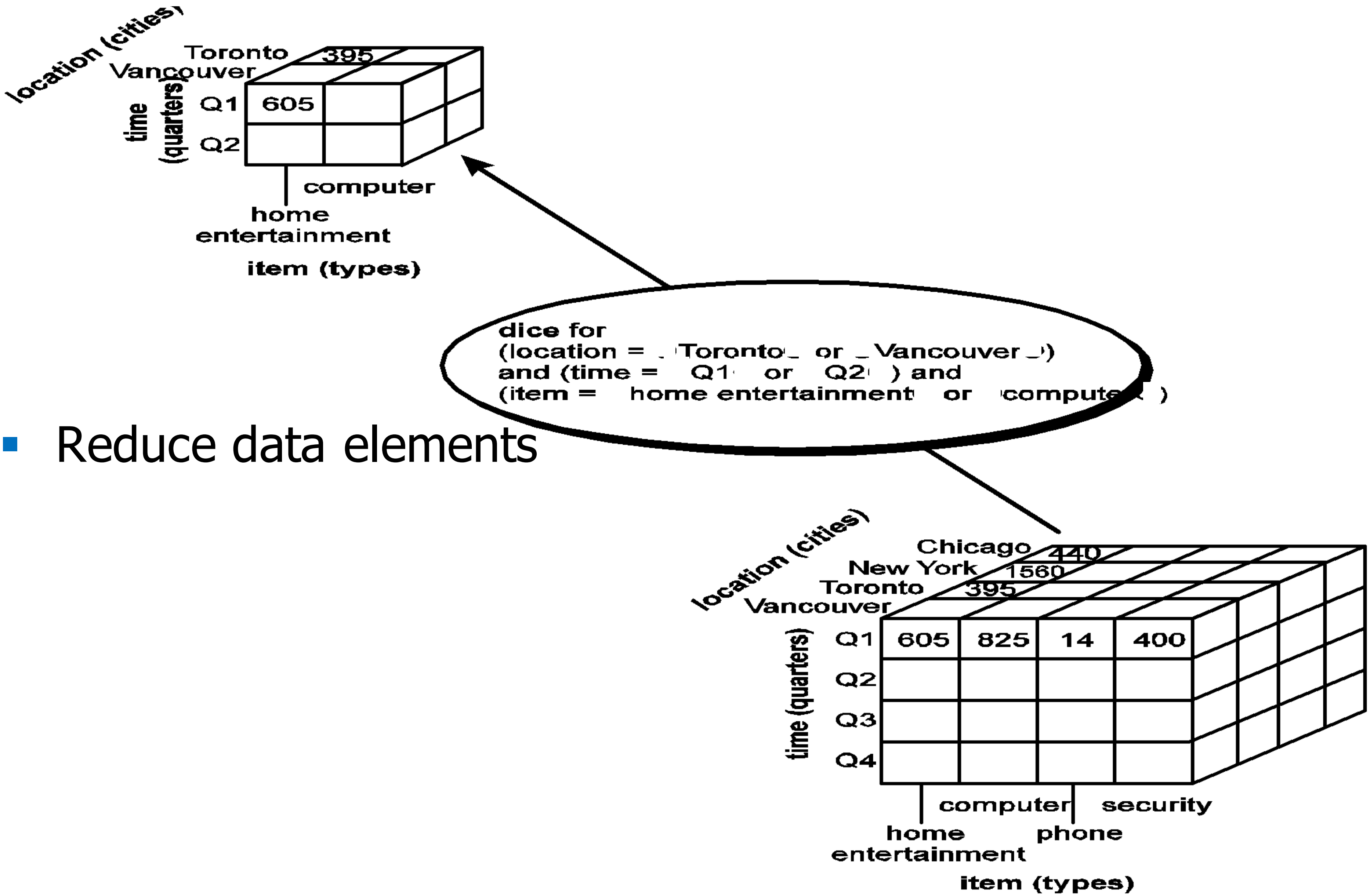


OLAP Operations: Slice

- Reduce dimension



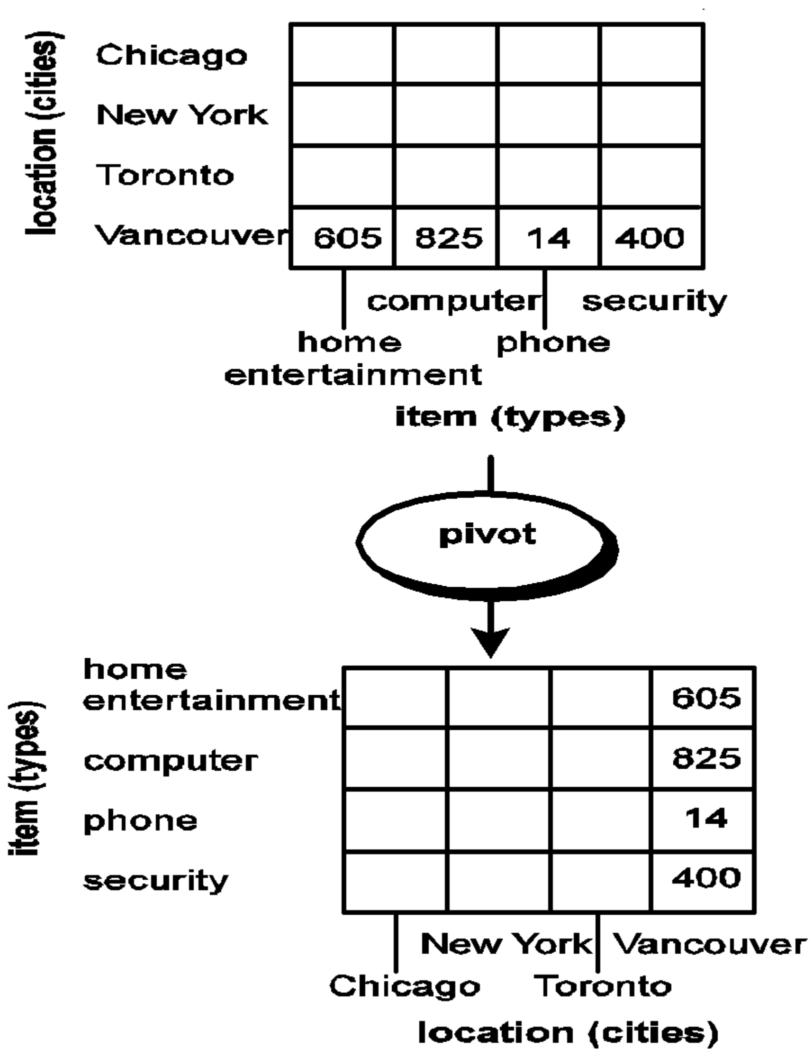
OLAP Operations: Dice



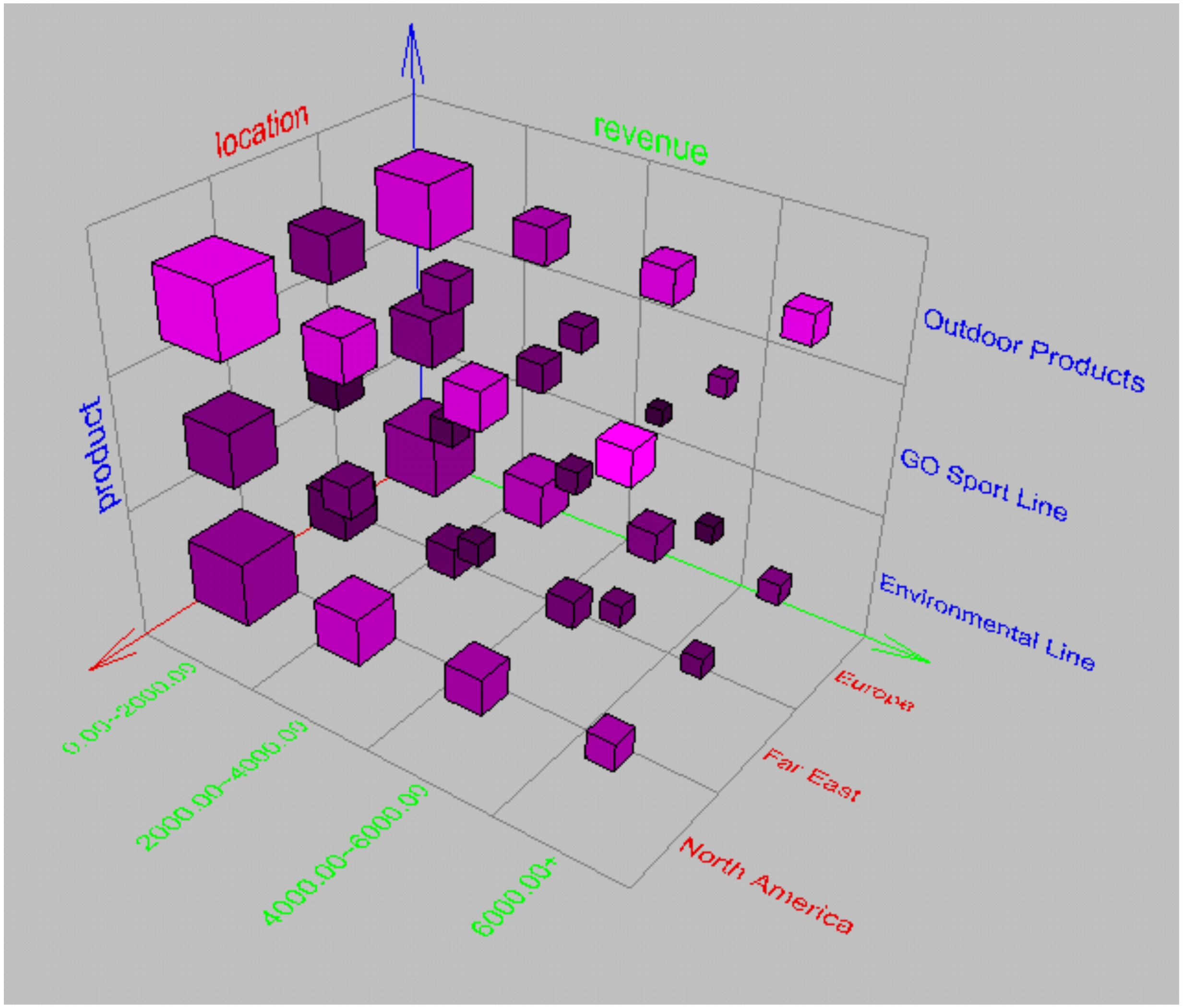
- Reduce data elements

OLAP Operations: Pivot

- Change Perspectives



Browsing a Data Cube



- Visualization
- OLAP capabilities
- Interactive manipulation

Chapter 4: Data Warehousing and On-line Analytical Processing

- Data Warehouse: Basic Concepts
- Data Warehouse Modeling: Data Cube and OLAP
- Data Warehouse Usage
- Data Warehouse Implementation
- Summary

Data Warehouse Usage

- 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

Chapter 4:
Data Warehousing and On-line Analytical Processing

- Data Warehouse: Basic Concepts
- Data Warehouse Modeling: Data Cube and OLAP
- Data Warehouse Design and Usage
- Data Warehouse Implementation
- Summary

Efficient Data Cube Computation

- Motivation: People expect high efficiency when querying data information

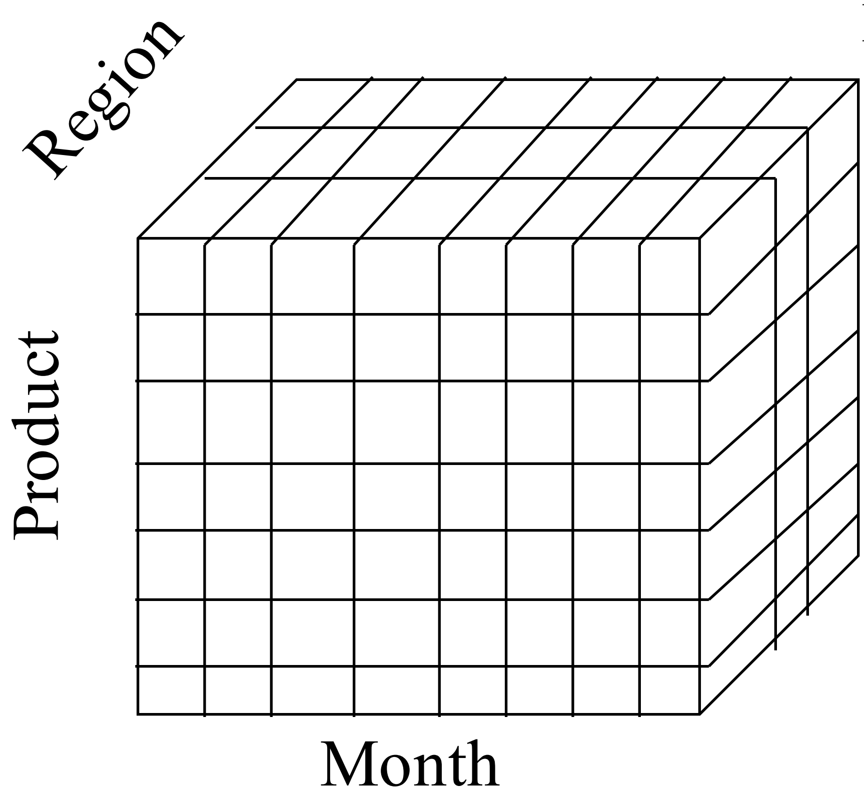
- Data cube can be viewed as a lattice of cuboids
  - The bottom-most cuboid is the base cuboid
  - The top-most cuboid (apex) contains only one cell
  - How many cuboids in an n-dimensional cube with L levels?

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

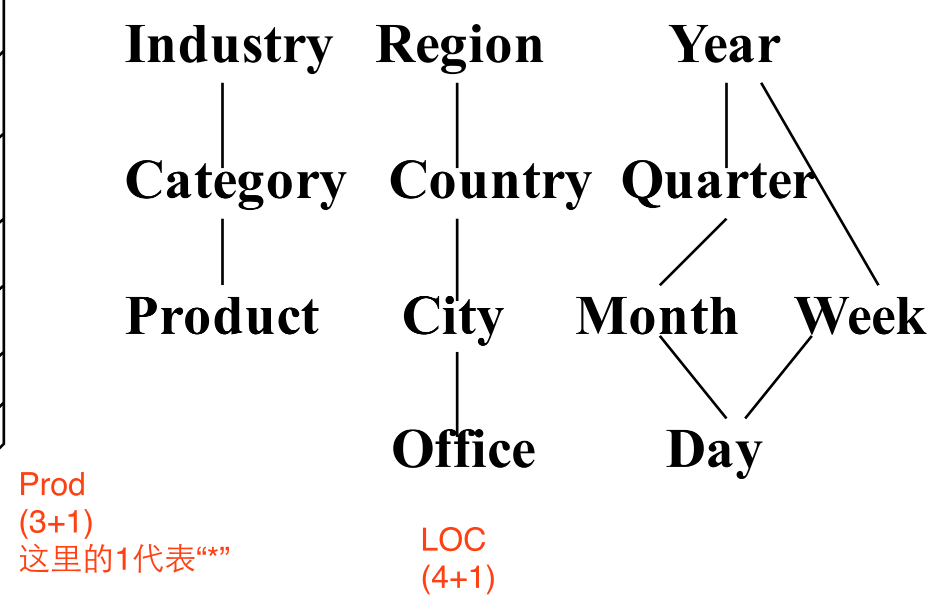
(1+1)*(3+1)*(1+1)

Efficient Data Cube Computation

- How many cuboids are there?
  - Sales volume as a function of product, month, and region



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



Data Cube Materialization

- 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.

The “Compute Cube” Operator

- Cube definition and computation in DMQL

  define cube sales [item, city, year]: sum (sales_in_dollars)

  compute cube sales

- Transform it into a SQL-like language (with a new operator cube by, introduced by Gray et al.’96)

  SELECT item, city, year, SUM (amount)

  FROM SALES

  CUBE BY item, city, year

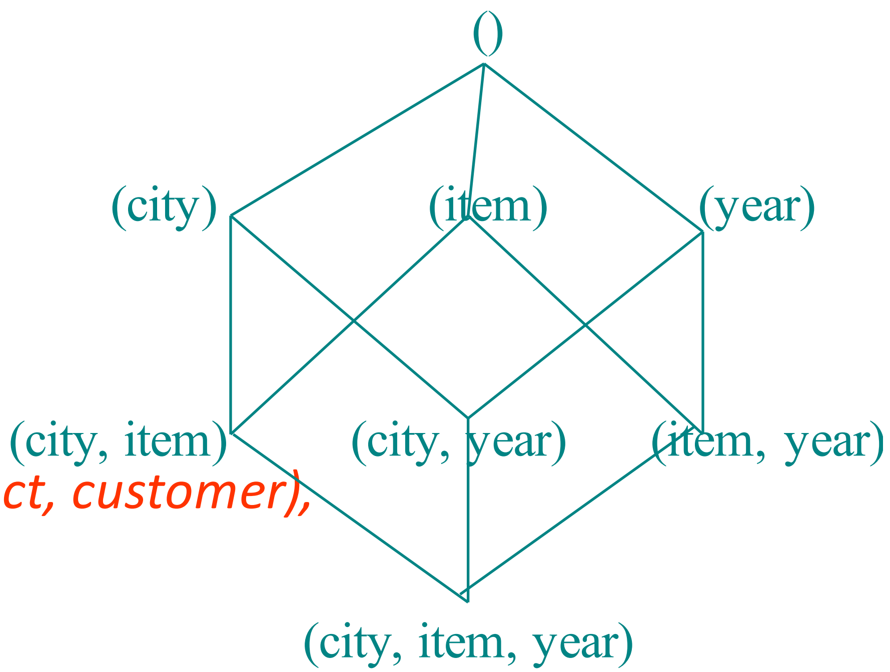
- Need compute the following Group-Bys

  (date, product, customer),

  (date,product),(date, customer), (product, customer),

  (date), (product), (customer)

  ()



Efficient Processing OLAP Queries

- **Determine which operations** should be performed on the available cuboids
  - Transform drill, roll, etc. into corresponding SQL and/or OLAP operations, e.g., dice = selection + projection

    Product, Location, Time

    item - > brand City->state->country

- **Determine which materialized cuboid(s)** should be selected for OLAP op.
  - 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~~*} USA

    3) {*year, brand, province_or_state*}

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

    Which should be selected to process the query?

- Explore indexing structures and compressed vs. dense array structs in MOLAP

Chapter 4: Data Warehousing and On-line Analytical Processing

- Data Warehouse: Basic Concepts
- Data Warehouse Modeling: Data Cube and OLAP
- Data Warehouse Usage
- Data Warehouse Implementation
- Summary

Summary

- Data warehousing: A multi-dimensional model of a data warehouse
  - A data cube consists of *dimensions* & *measures*
  - Star schema, snowflake schema, fact constellations
  - OLAP operations: drilling, rolling, slicing, dicing and pivoting
- Data Warehouse Architecture and Usage
  - Multi-tiered architecture
  - Business analysis design framework
  - Information processing, analytical processing, data mining
- Implementation: Efficient computation of data cubes
  - Partial vs. full vs. no materialization
  - Indexing OALP data: Bitmap index and join index
  - OLAP query processing

Chapter 4: Data Warehousing and On-line Analytical Processing

- Data Warehouse: Basic Concepts
  - (a) What Is a Data Warehouse?
  - (b) Data Warehouse: A Multi-Tiered Architecture
- Data Warehouse Modeling: Data Cube and OLAP
  - (a) Cube: A Lattice of Cuboids
  - (b) Conceptual Modeling of Data Warehouses
  - (c) Stars, Snowflakes, and Fact Constellations: Schemas for Multidimensional Databases
  - (d) Dimensions: The Role of Concept Hierarchy
  - (e) Measures: Their Categorization and Computation
  - (f) Cube Definitions in Database systems
  - (g) Typical OLAP Operations
  - (h) A Starnet Query Model for Querying Multidimensional Databases
- Data Warehouse Usage
  - (a) Design of Data Warehouses: A Business Analysis Framework
  - (b) Data Warehouse Usage
- Data Warehouse Implementation
  - (a) Efficient Data Cube Computation: Cube Operation, Materialization of Data Cubes, and Iceberg Cubes
  - (b) Indexing OLAP Data: Bitmap Index and Join Index
  - (c) Efficient Processing of OLAP Queries
- Summary