Chapter 3
Data Warehousing and Analytical Processing

Outline

- Data warehouse
- Data warehouse modeling: schema and measures
- OLAP operations

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
    - Ex. Hotel price: differences on currency, tax, breakfast covered, and parking
  - 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

- Independence
  - A physically separate store of data transformed from the operational environment
- Static: 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

- OLTP: Online transactional processing
  - DBMS operations
  - Query and transactional processing
- OLAP: Online analytical processing
  - Data warehouse operations
  - Drilling, slicing, dicing, etc.

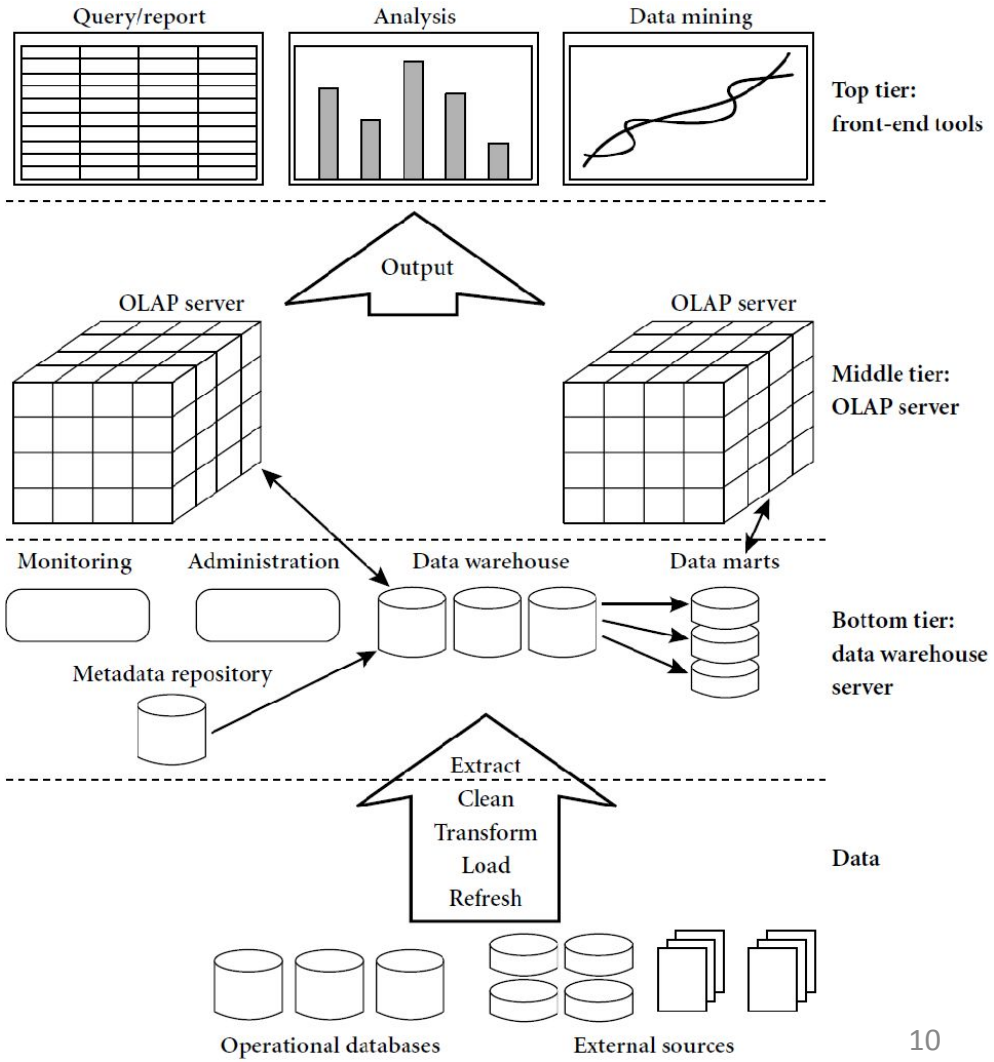
	OLTP	OLAP
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
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: A Multi-Tiered Architecture

- Top Tier: Front-End Tools
- Middle Tier: OLAP Server
- Bottom Tier: Data Warehouse Server
- Data



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

Extraction, Transformation, and Loading (ETL)

- 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 indicies and partitions
- Refresh
  - propagate the updates from the data sources to the warehouse

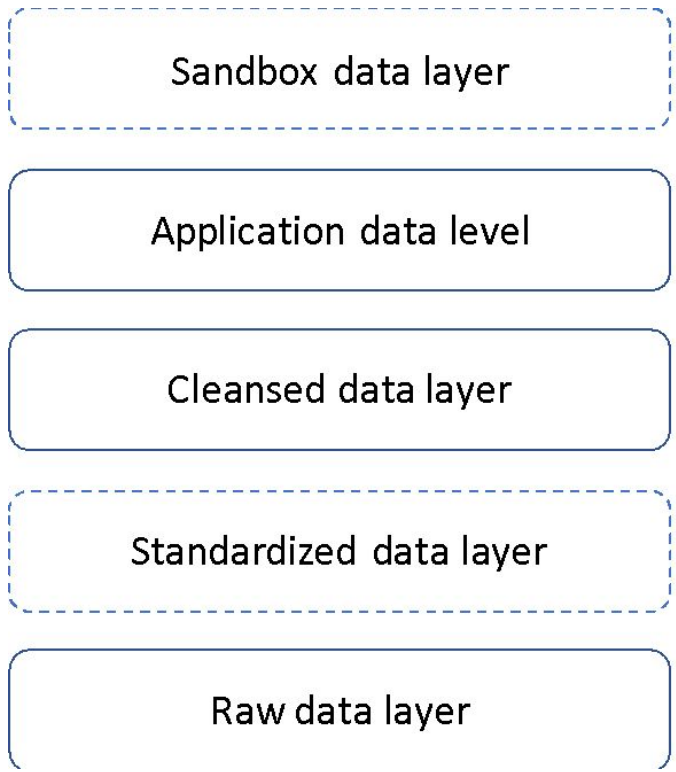
Three Data Warehouse Models

- Enterprise warehouse
  - Collects all of the information about subjects spanning the entire organization
- Data Mart
  - A subset of corporate-wide 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

Data Lake

- A data lake is a centralized repository storing all structured and unstructured data at any scale in an organization
- Data is stored as is, without having to first structure the data, and run different types of analytics—from dashboards and visualizations to big data processing, real-time analytics, and machine learning to guide better decision

Layers of Storage



Enriching data by data scientists and analysts

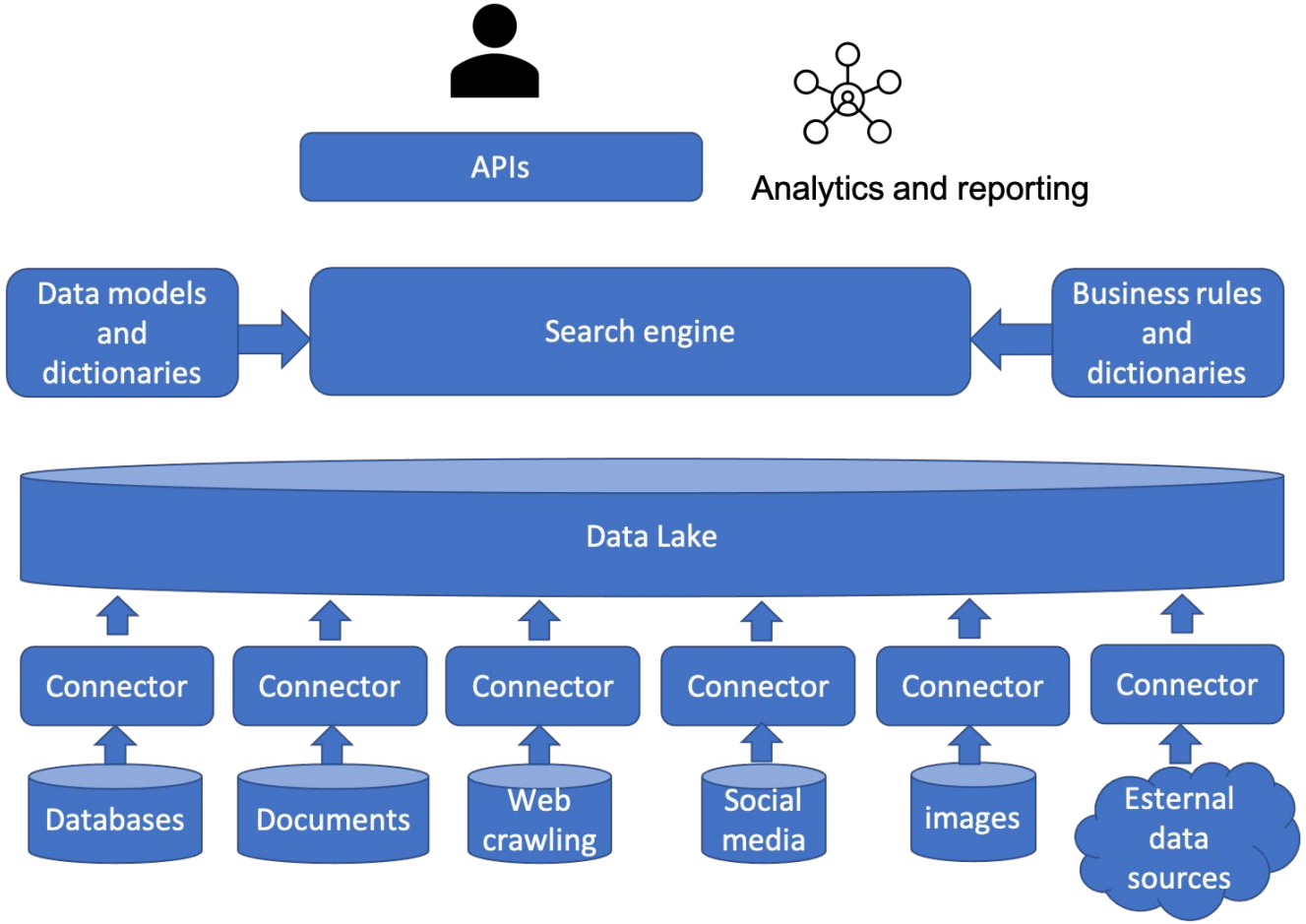
Implementing business logics

Cleansing and transforming data

Facilitating efficient data transferring and cleansing

Loading raw data

Conceptual Architecture



Data Lakehouses

- “A data lakehouse is a new, open data management architecture that combines the flexibility, cost-efficiency, and scale of data lakes with the data management and ACID transactions of data warehouses, enabling business intelligence (BI) and machine learning (ML) on all data.” (Databricks)

Outline

- Data warehouse
- Data warehouse modeling: schema and measures
  - Data cube: a multidimensional data model
  - Schemas for multidimensional data models: stars, snowflakes, and fact constellations
  - Concept hierarchies
  - Measures: categorization and computation
- OLAP operations
- Data cube computation
- Data cube computation methods

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
- Data cube: A lattice of cuboids
  - 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.

Table 3.1 2-D view of sales data according to *time* and *item*.

	location = "Vancouver"			
time (quarter)	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

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

ke to view the data according to *time* and *item*, as well as *location*, for the cit
oronto, and Vancouver. These 3-D data are shown in Table 3.2. The 3-D dat

Table 3.1 2-D view of sales data according to *time* and *item*.

	location = "Vancouver"			
time (quarter)	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

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

Table 3.2 3-D view of sales data according to *time*, *item*, and *location*.

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

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

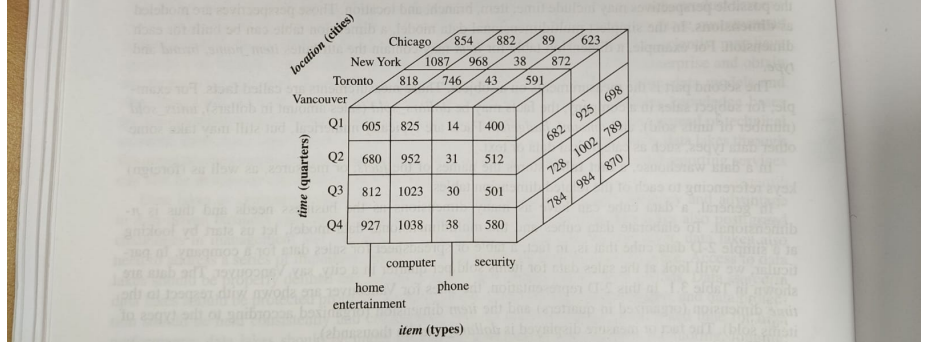


FIGURE 3.4

A 3-D data cube representation of the data in Table 3.2, according to *time*, *item*, and *location*. The measure displayed is *dollars_sold* (in thousands).

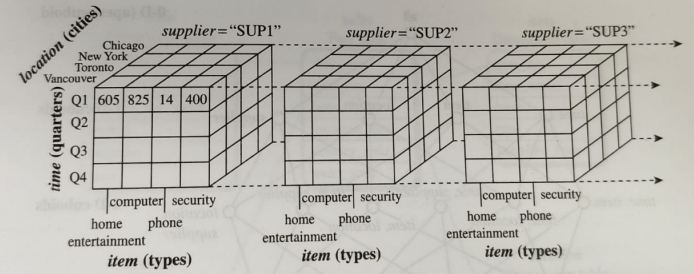
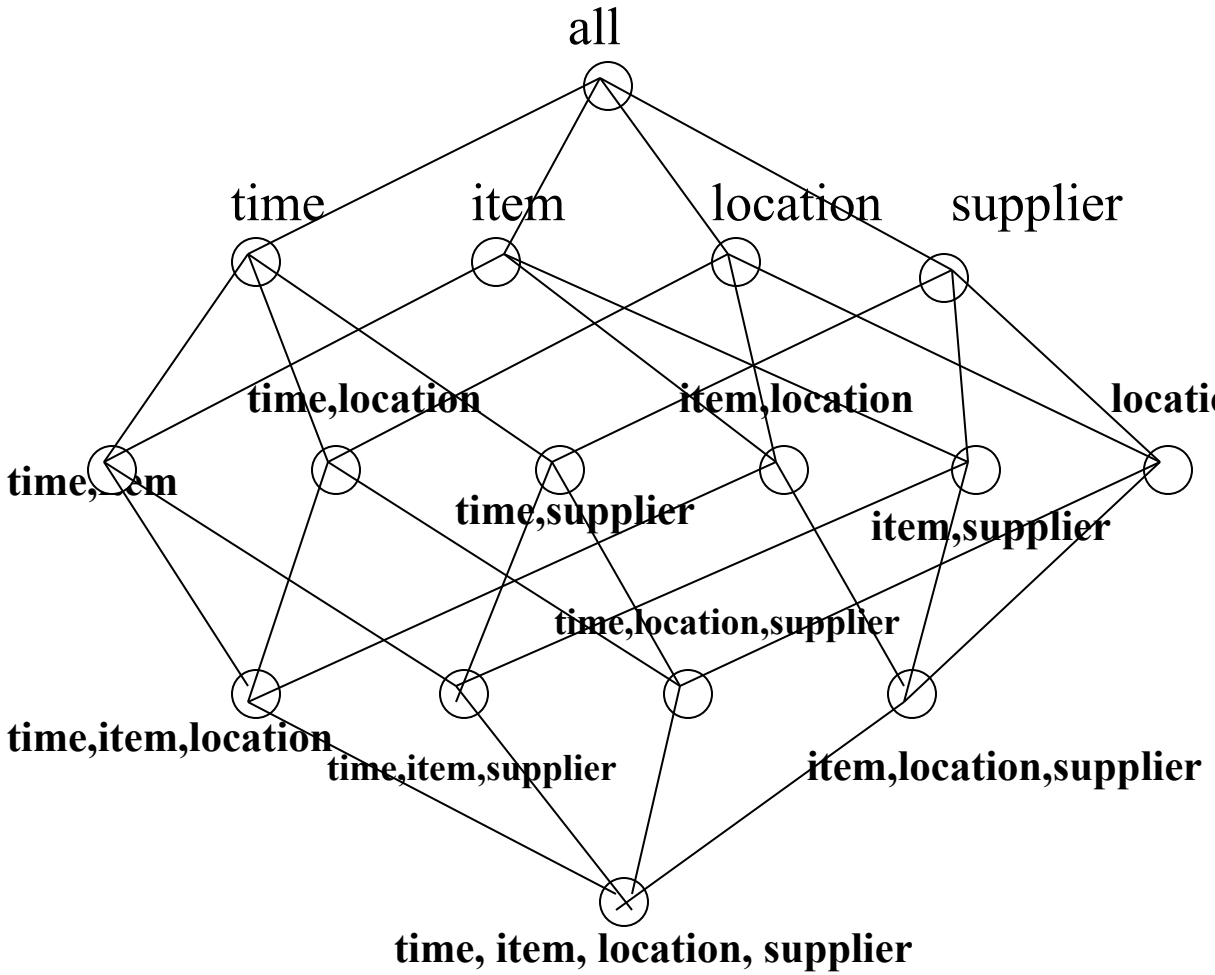


FIGURE 3.5

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.

Each group-by can be represented by a

Data Cube: A Lattice of Cuboids



0-D (*apex*) cuboid

1-D cuboids

2-D cuboids

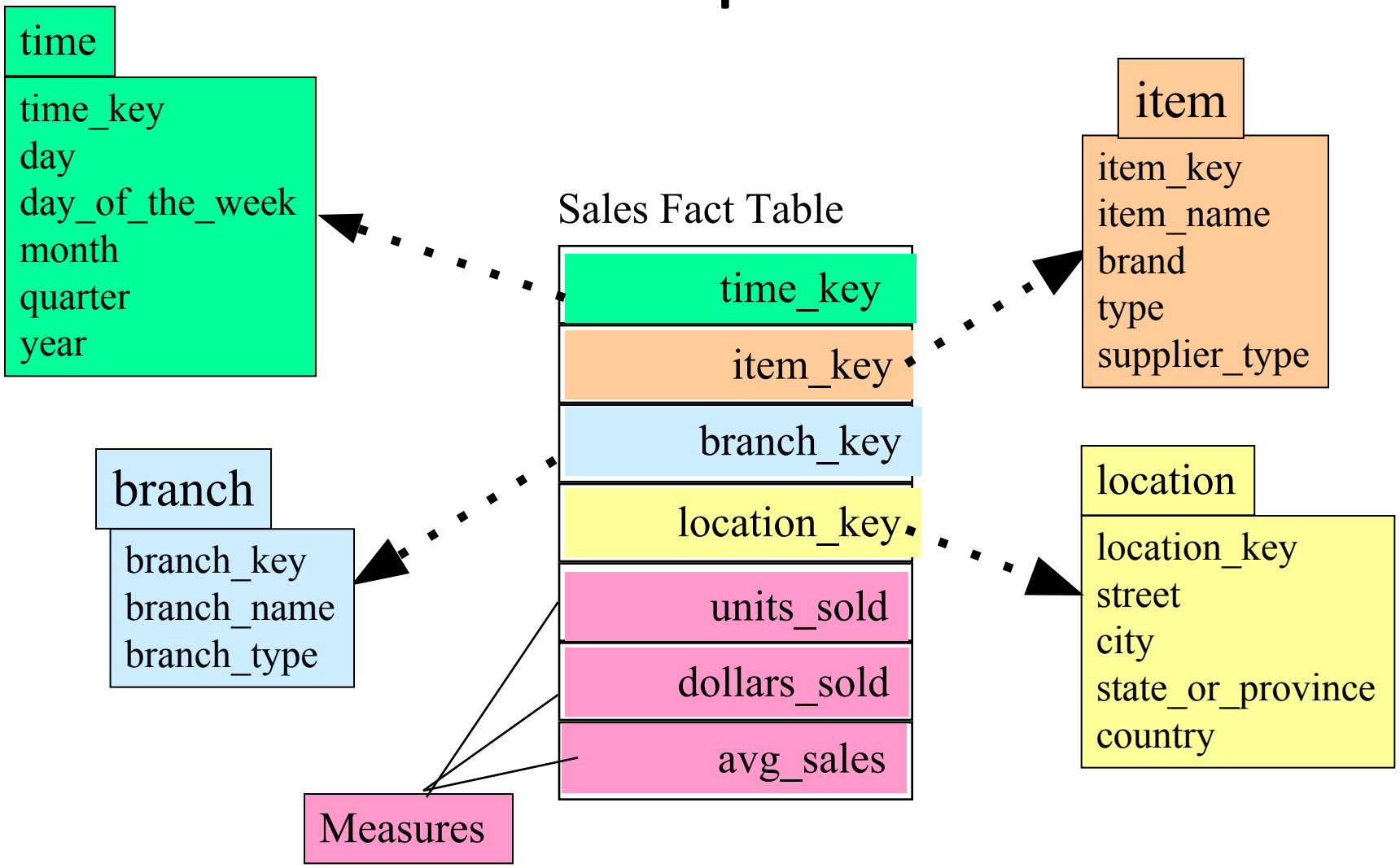
3-D cuboids

4-D (*base*) cuboid

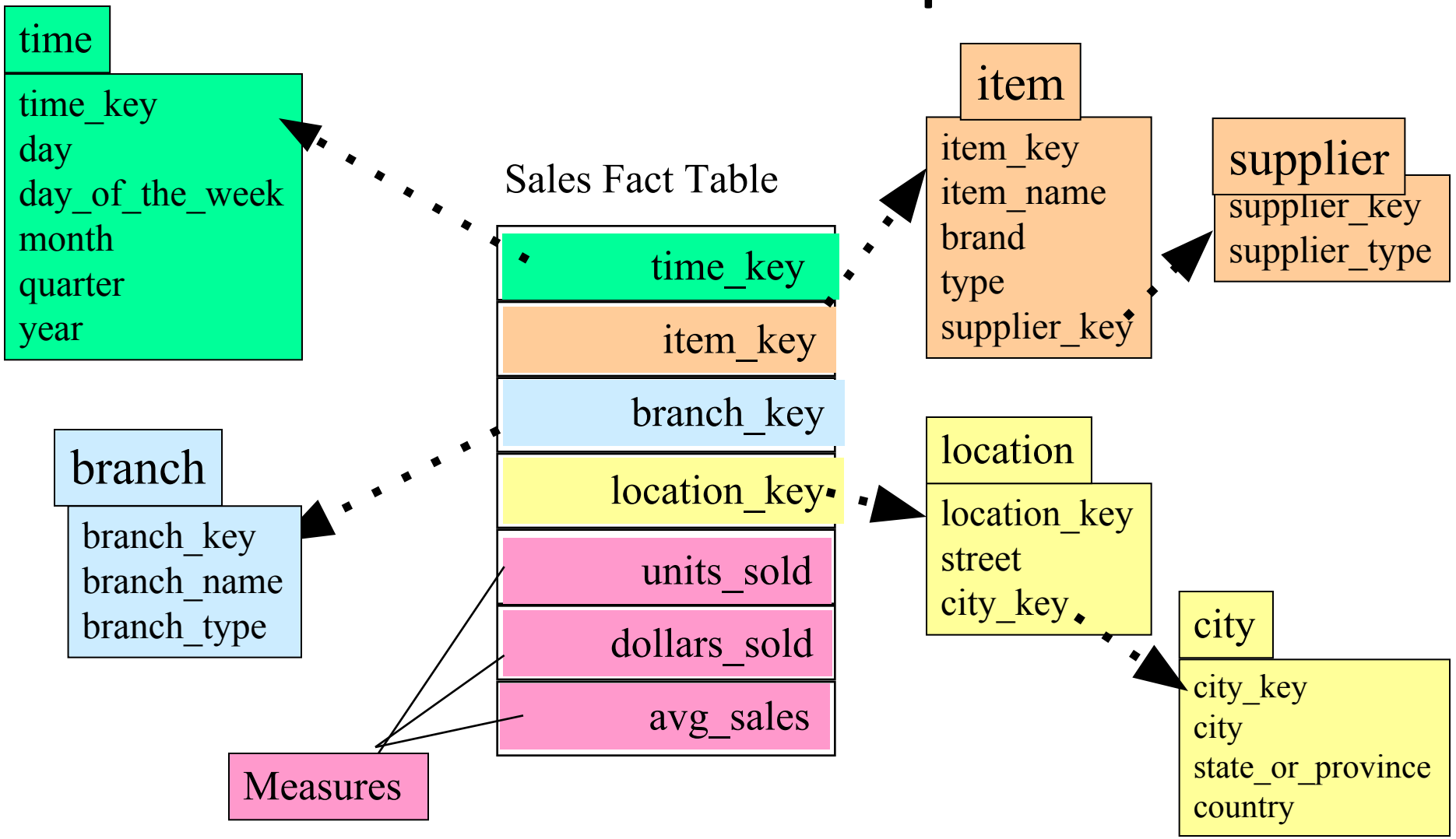
Conceptual Modeling of Data Warehouses

- 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

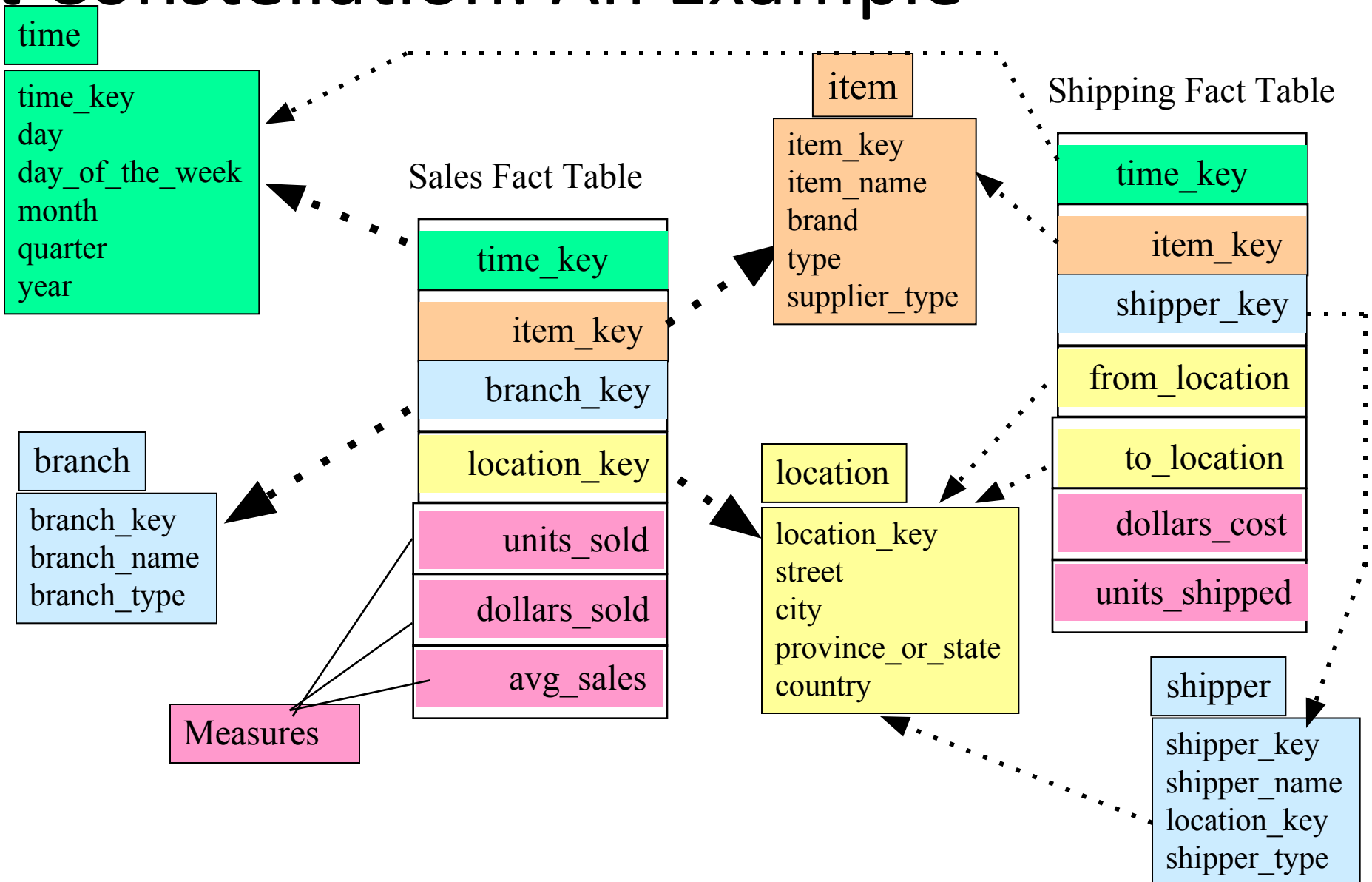
Star Schema: An Example



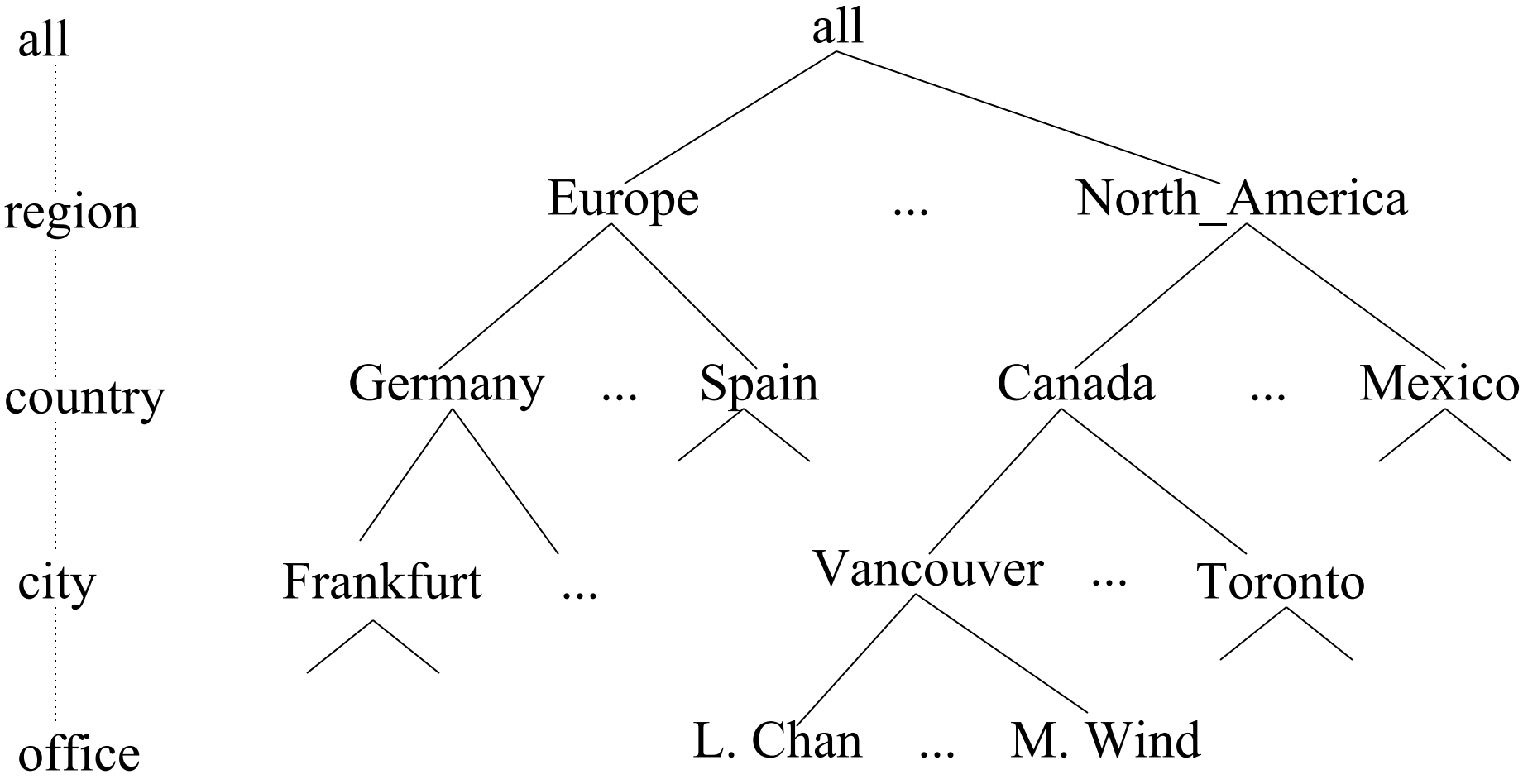
Snowflake Schema: An Example



Fact Constellation: An Example



A Concept Hierarchy for a Dimension (location)

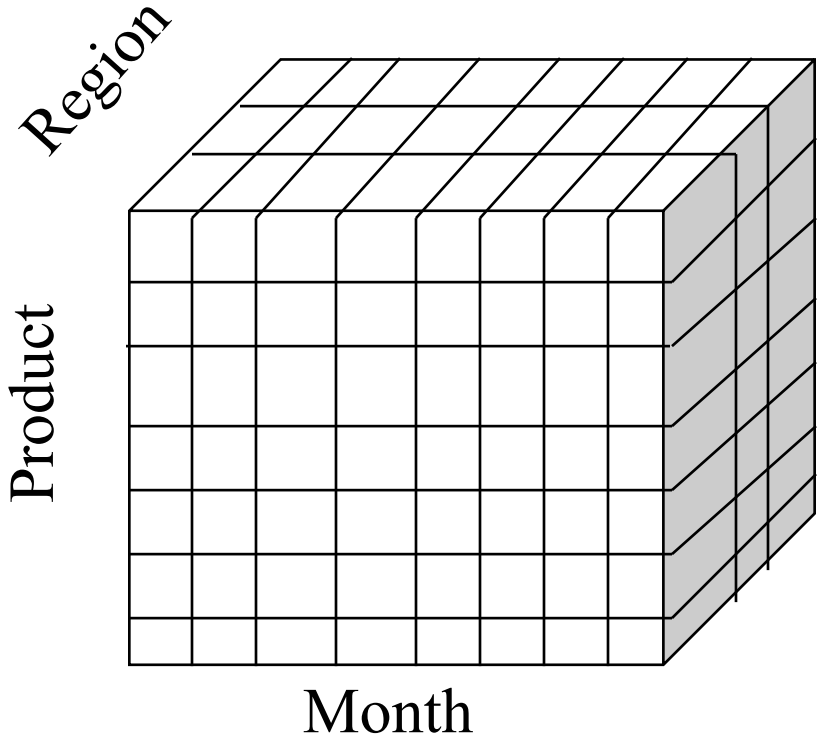


Data Cube Measures: Three Categories

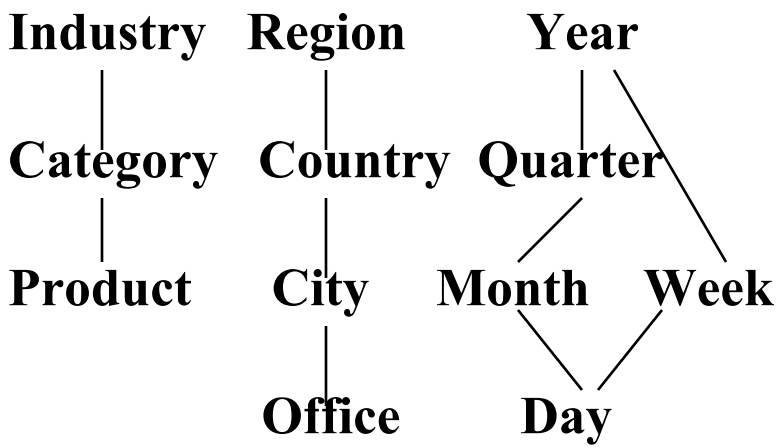
- Distributive: if the result derived by applying the function to n aggregate values is the same as that derived by applying the function on all the data without partitioning
    - E.g., count(), sum(), min(), max()
- Algebraic: if it can be computed by an algebraic function with M arguments (where M is a bounded integer), each of which is obtained by applying a distributive aggregate function
    - avg(x) = sum(x) / count(x)
    - Is min_N() an algebraic measure? How about standard_deviation()?
- Holistic: if there is no constant bound on the storage size needed to describe a subaggregate.
    - E.g., median(), mode(), rank()

Multidimensional Data

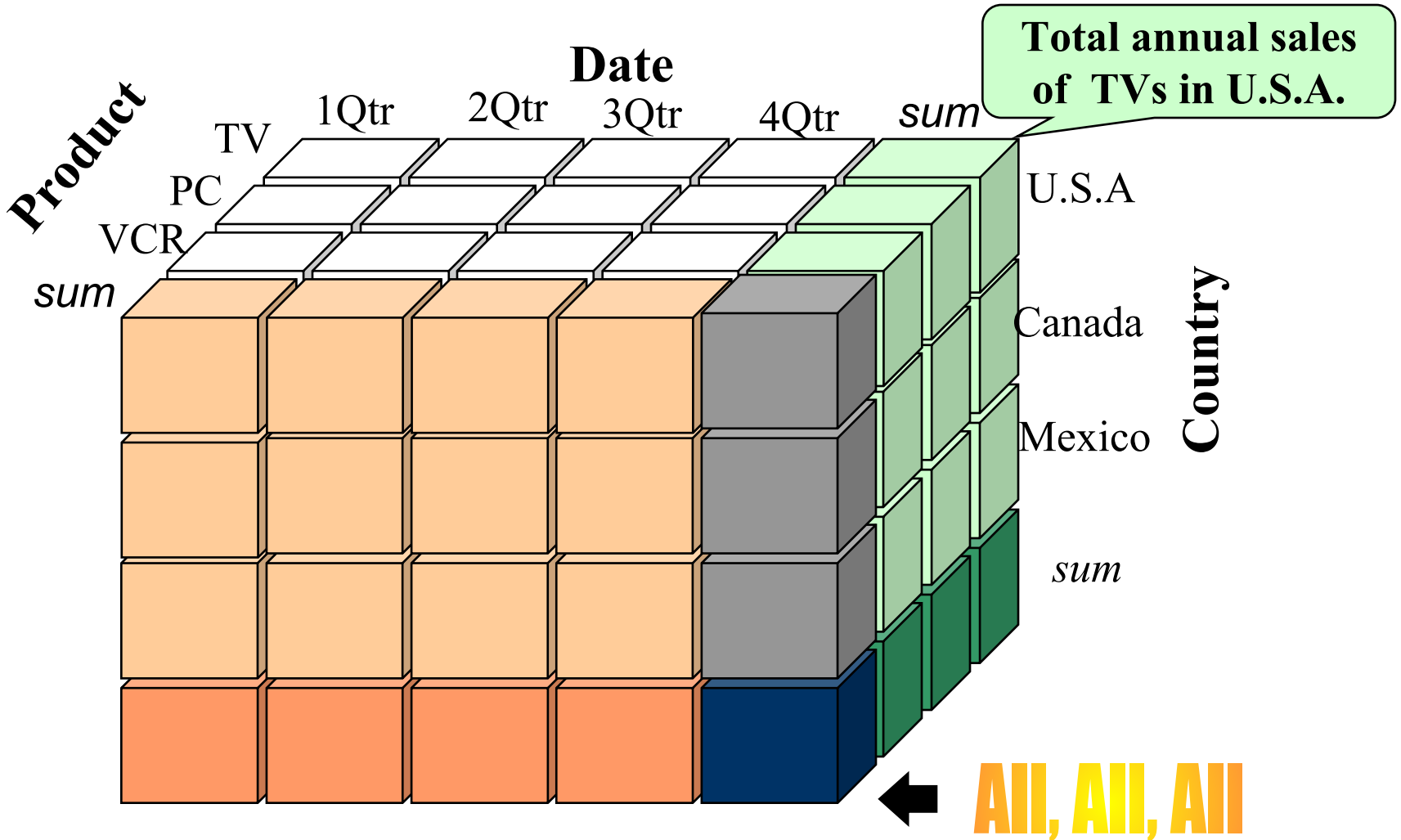
- Sales volume as a function of product, month, and region



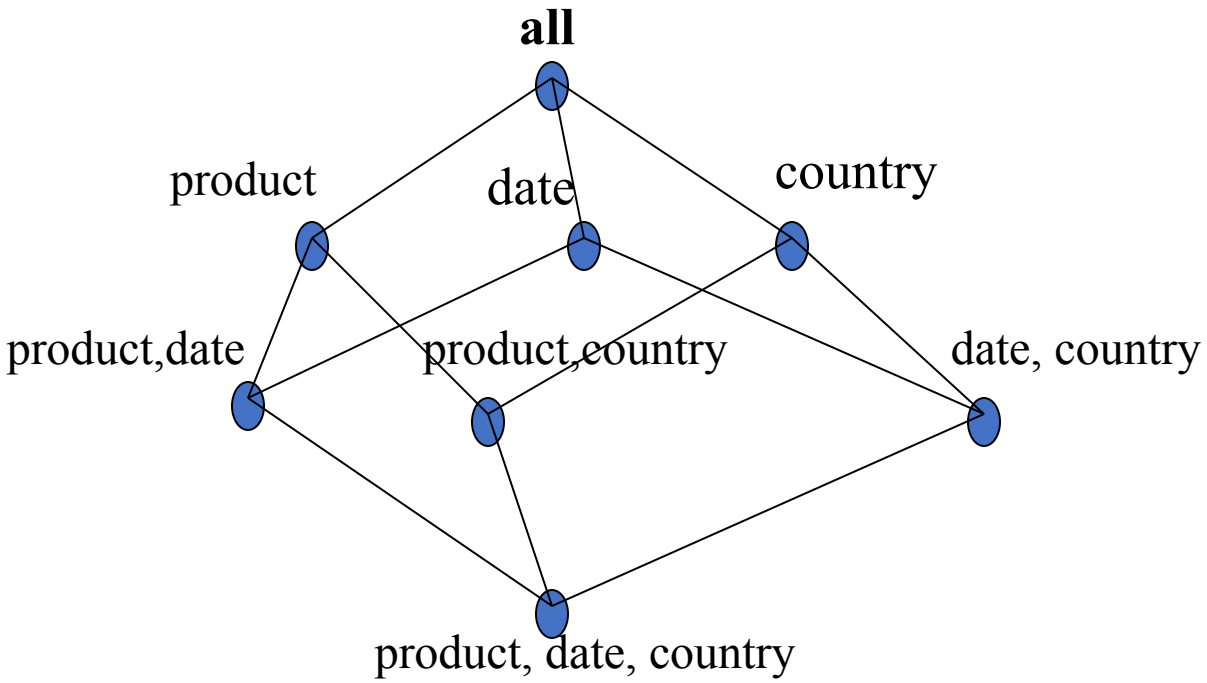
Dimensions: *Product, Location, Time*
Hierarchical summarization paths



A Sample Data Cube



Cuboids Corresponding to the Cube



0-D (*apex*) cuboid

1-D cuboids

2-D cuboids

3-D (*base*) cuboid

Outline

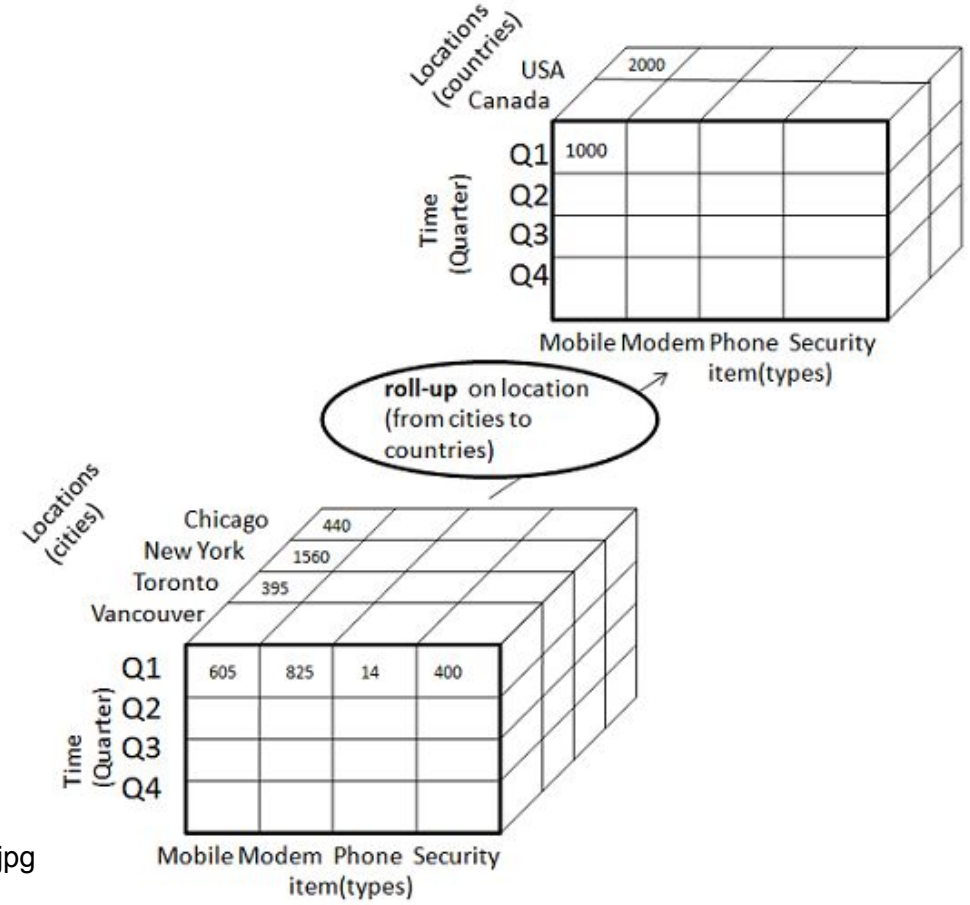
- Data warehouse
- Data warehouse modeling: schema and measures
- OLAP operations
  - Typical OLAP operations
  - Indexing OLAP data: bitmap index and join index
  - Storage implementation: column-based databases
- Data cube computation
- Data cube computation methods

Online Analytic Processing (OLAP)

- Conceptually, we may explore all possible subspaces for interesting patterns
- Some fundamental problems in analytics and data mining
  - What patterns are interesting?
  - How can we explore all possible subspaces systematically and efficiently?
- Aggregates and group-bys are frequently used in data analysis and summarization
  - SELECT time, altitude, AVG(temp)
  - FROM weather GOUP BY time, altitude;
  - In TPC, 6 standard benchmarks have 83 queries, aggregates are used 59 times, group-bys are used 20 times
- Online analytical processing (OLAP): the techniques that answer multi-dimensional analytical (MDA) queries efficiently

OLAP Operations

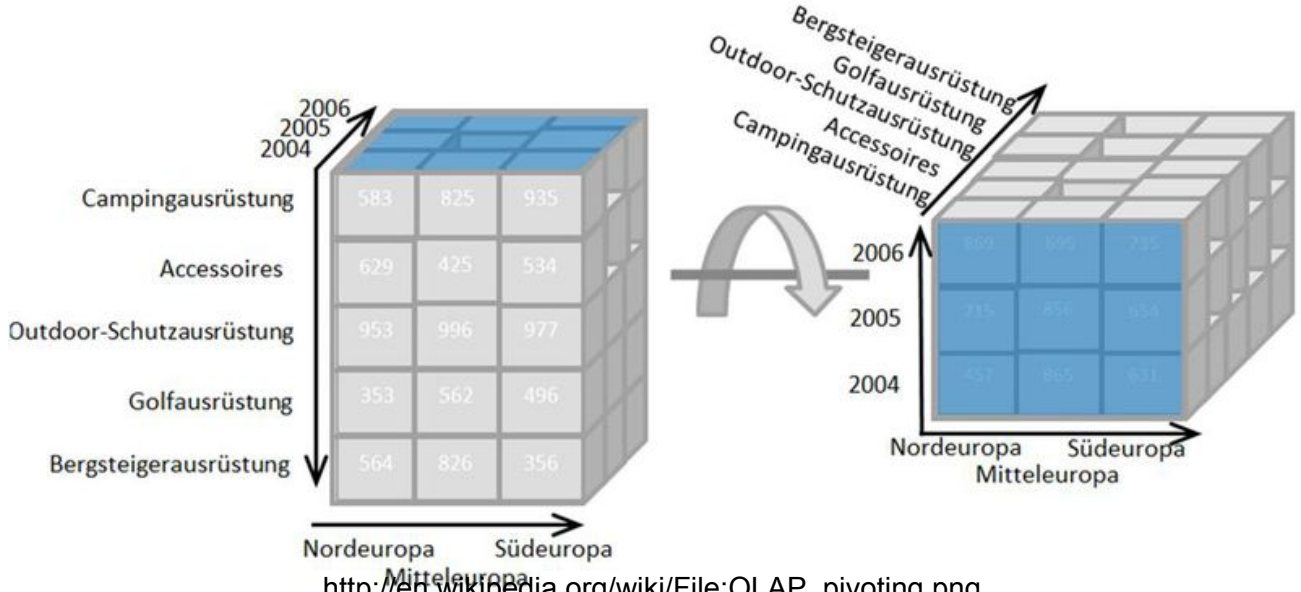
- Roll up (drill-up): summarize data by climbing up hierarchy or by dimension reduction
  - (Day, Store, Product type, SUM(sales) ⍰ (Month, City, *, SUM(sales))
- Drill down (roll down): reverse of roll-up, from higher level summary to lower level summary or detailed data, or introducing new dimensions



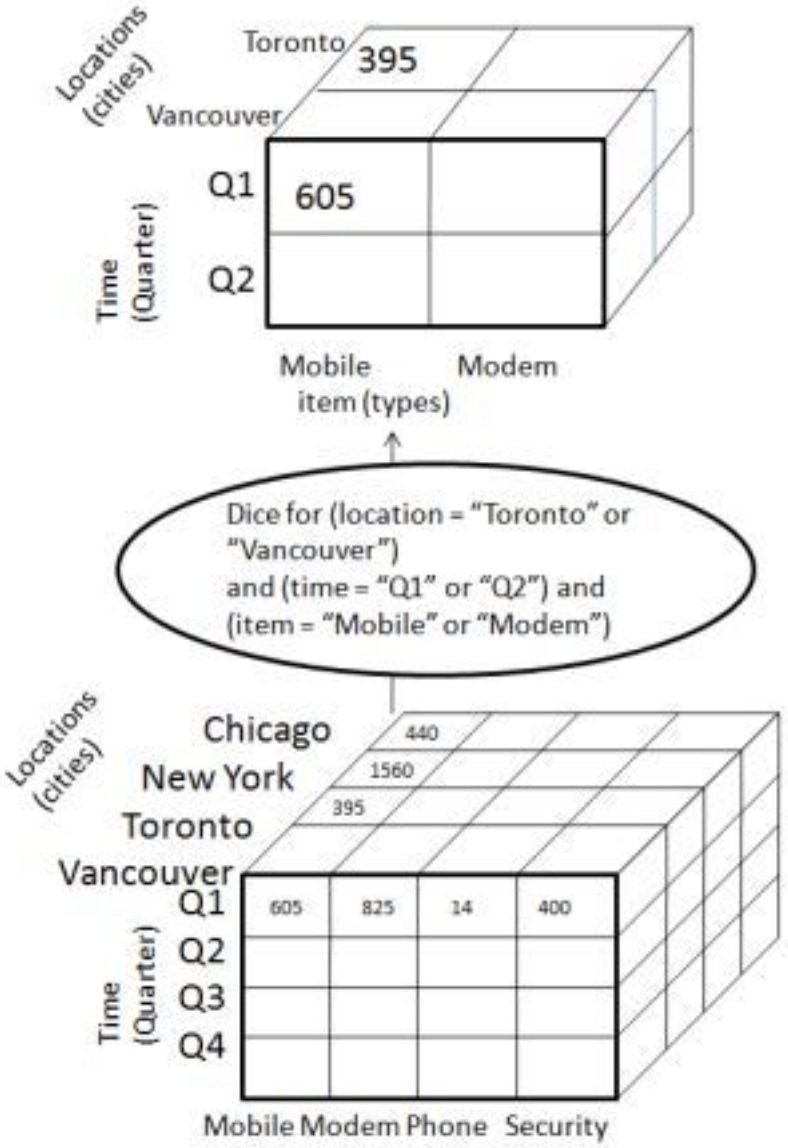
http://www.tutorialspoint.com/dwh/images/rollup.jpg

Other Operations

- Dice: pick specific values or ranges on some dimensions
- Pivot: “rotate” a cube – changing the order of dimensions in visual analysis



http://en.wikipedia.org/wiki/File:OLAP_pivoting.png



http://www.tutorialspoint.com/dwh/images/dice.jpg

Typical OLAP Operations