UNIVERSITY OF TEXAS ARLINGTON

CSE4334/5334 Data Mining

Multi-dimensional Data Analytics: OLAP, Data Cube

Fall 2020

Chengkai Li

(Slides courtesy of Jiawei Han, Micheline Kamber and Jian Pei. Data Mining: Concepts and Techniques, 3rd ed.)

What is Data Warehouse?

- “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—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

- 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

Data Cube

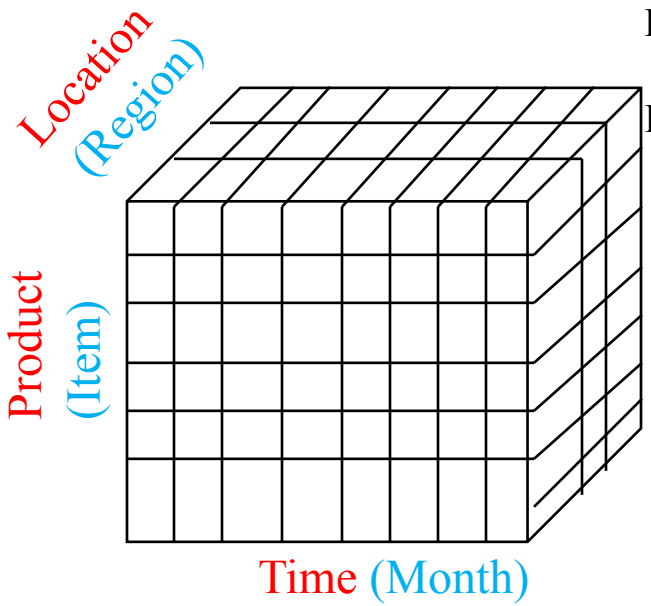
- A data warehouse is based on a multidimensional data model which views data in the form of a data cube

- A data cube contains aggregates of measure values, on various combinations of dimensions, and furthermore, with various levels of aggregation on individual dimension.

- 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.

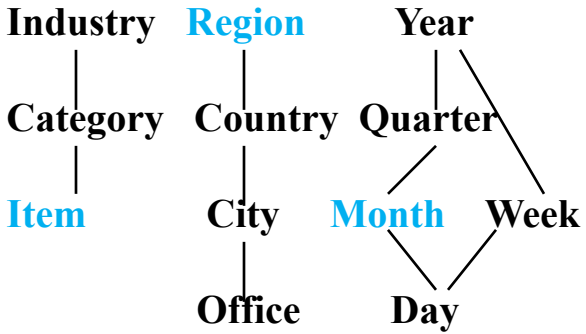
A 3-D Cuboid

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

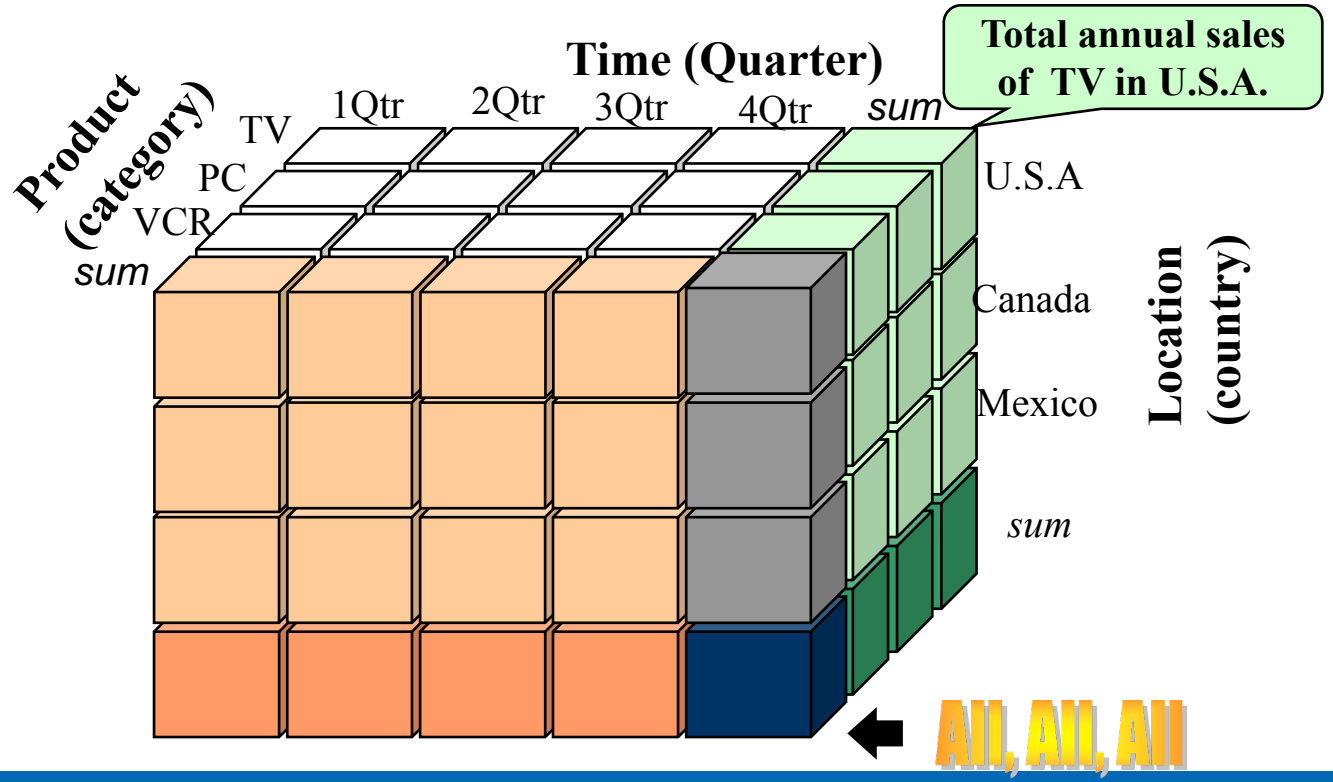


Dimensions: Product, Location, Time

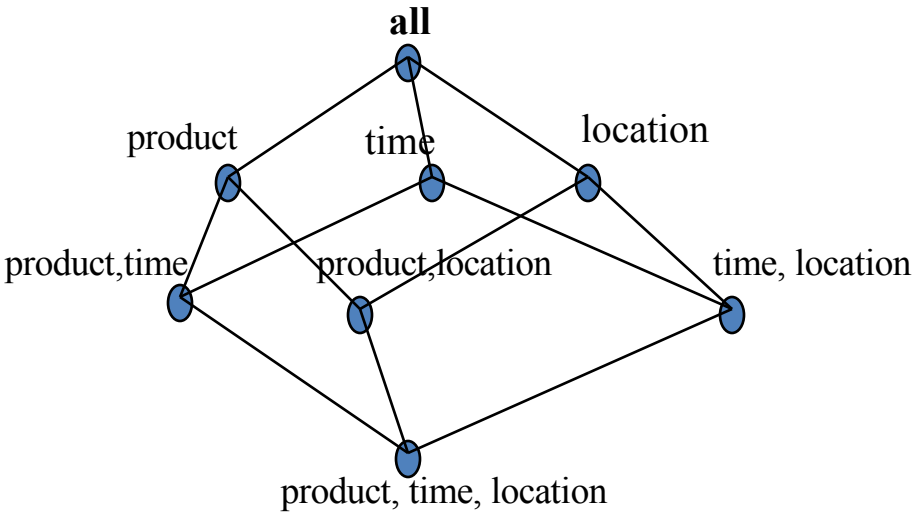
Hierarchical summarization paths



An Example of Data Cube



Data Cube: A Lattice of Cuboids



0-D(apex) cuboid

1-D cuboids

2-D cuboids

3-D (base) cuboid

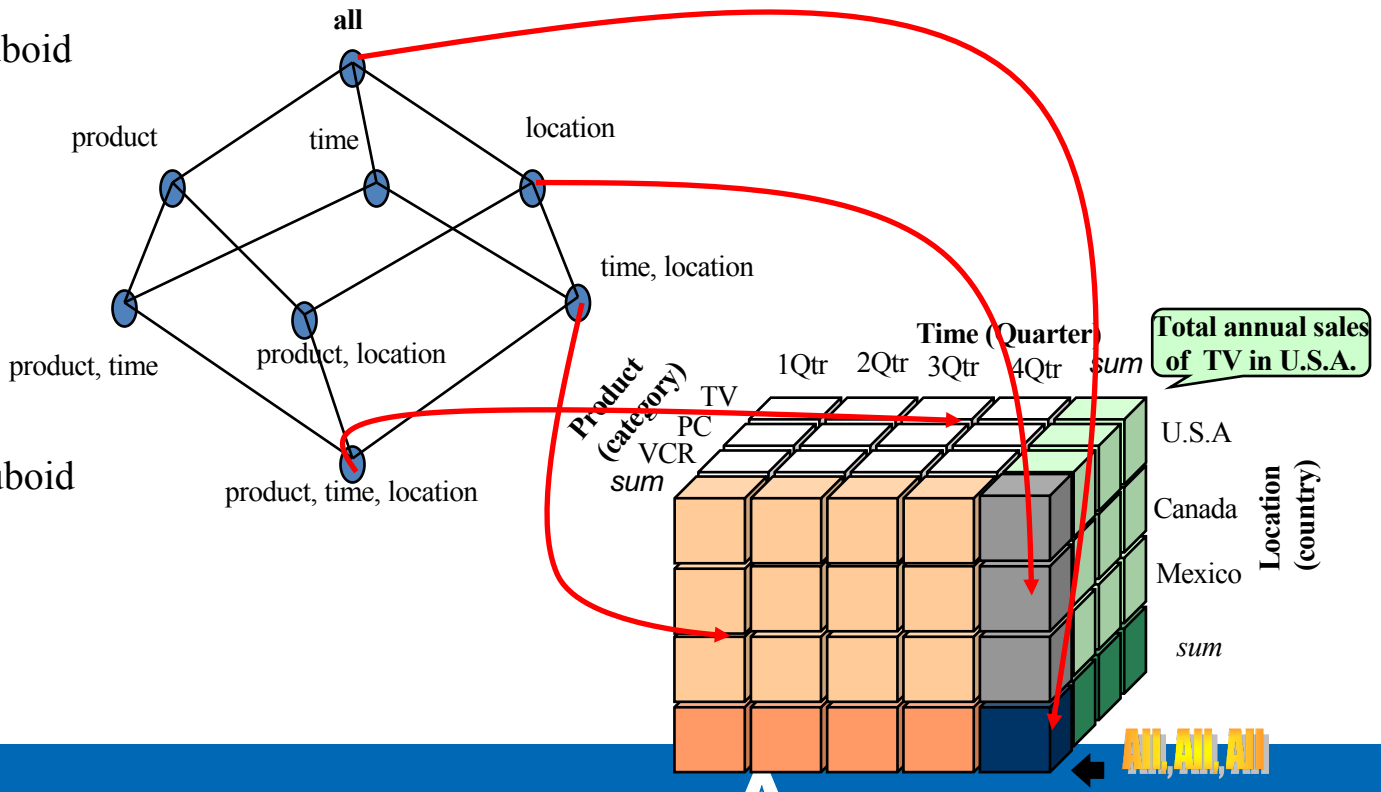
Cuboids

0-D(apex) cuboid

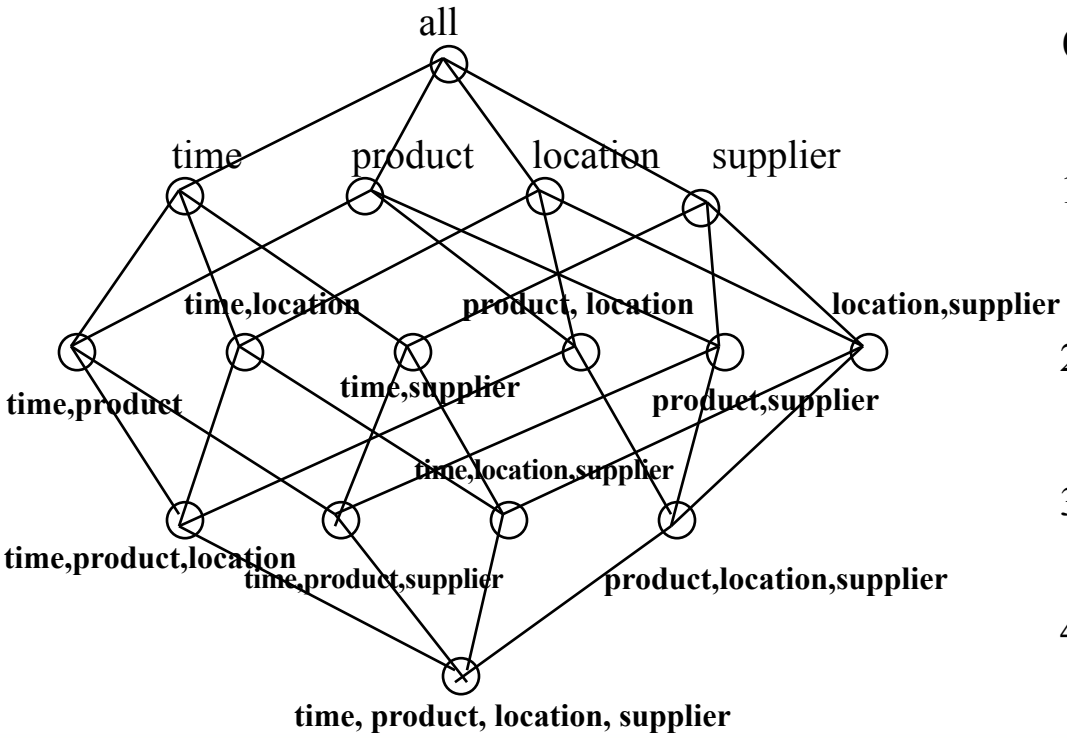
1-D cuboids

2-D cuboids

3-D (base) cuboid



A 4-D Data Cube



0-D(apex) cuboid

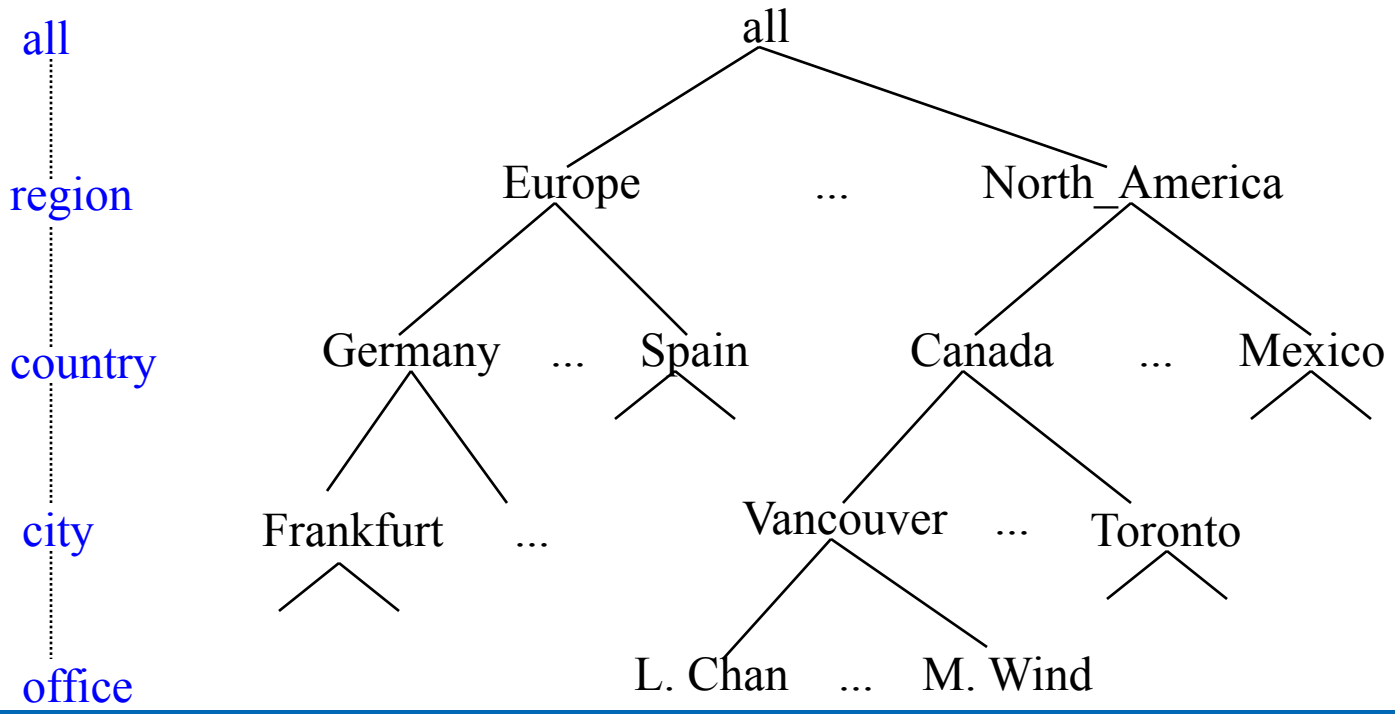
1-D cuboids

2-D cuboids

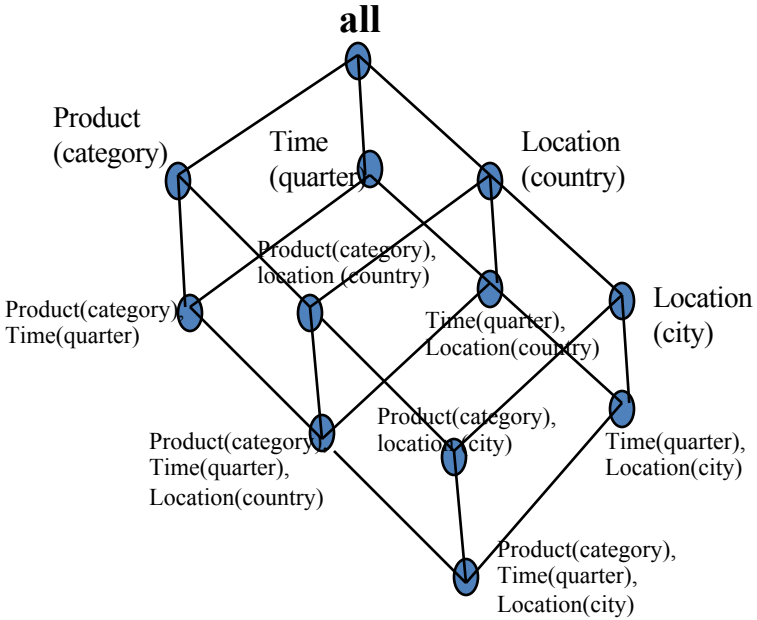
3-D cuboids

4-D(base) cuboid

A Concept Hierarchy on Location Dimension



Concept Hierarchy in Data Cube



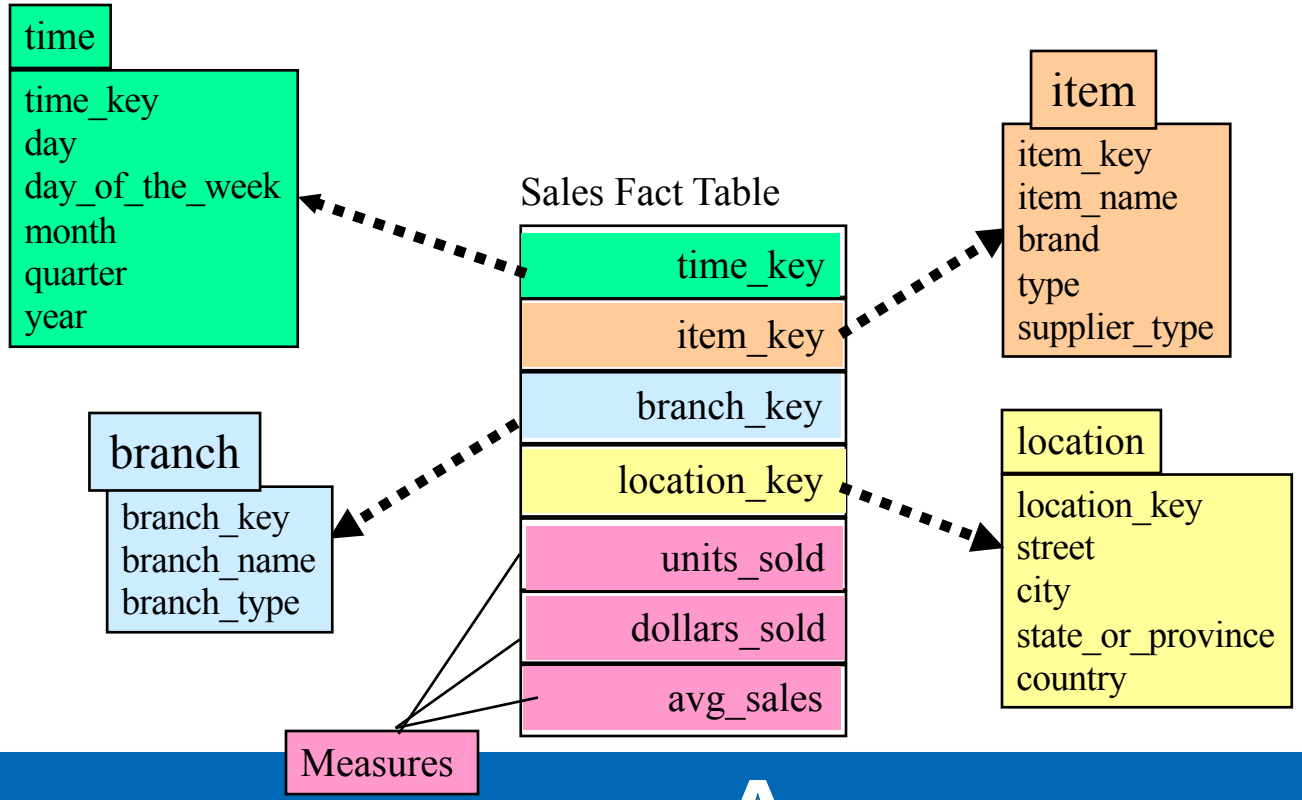
Conceptual Schema Design

- Dimensions & Measures
  - Dimension tables, such as product (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

Conceptual Modeling of Data Warehouses

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

Example of Star Schema



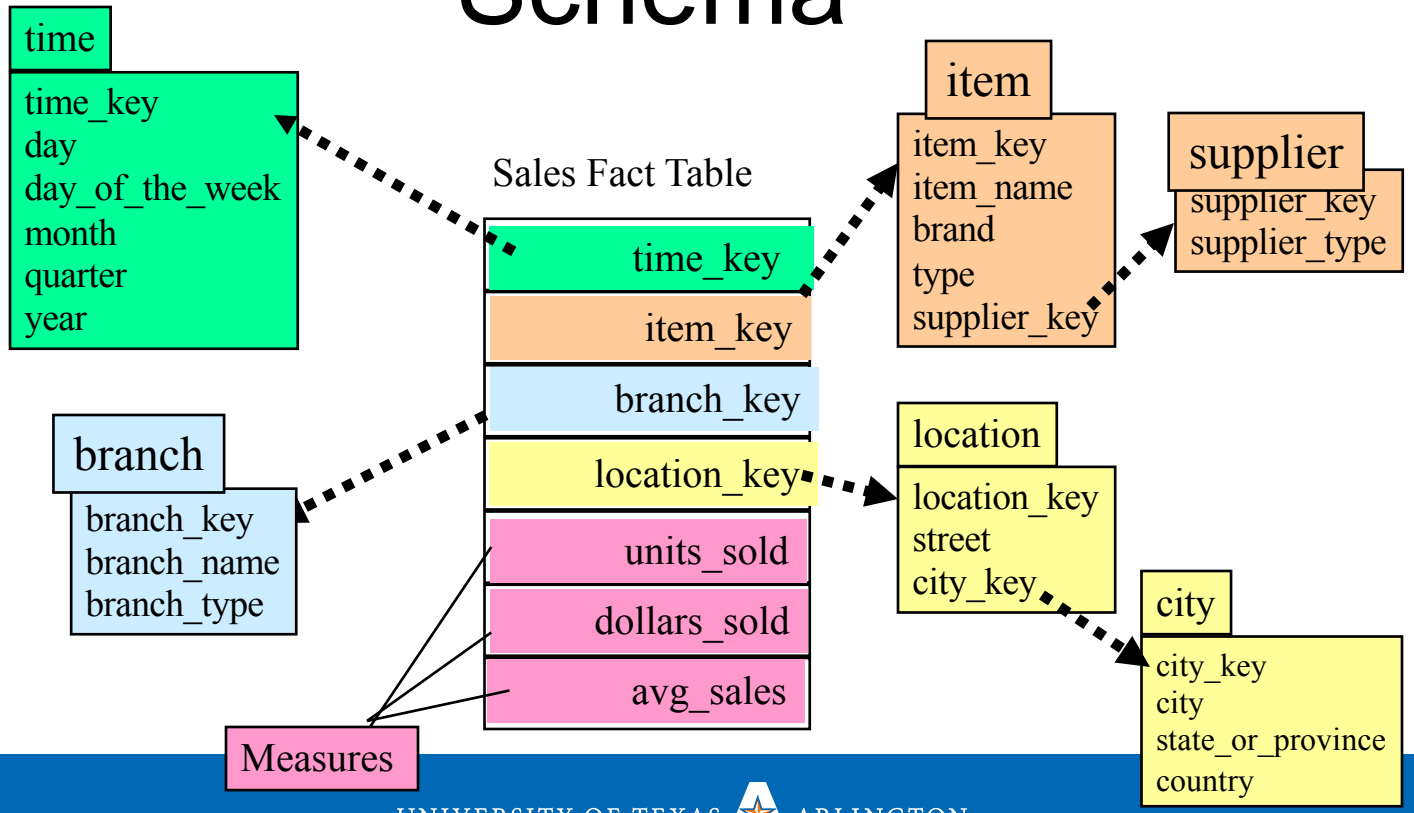
Conceptual Modeling of Data Warehouses

- 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

It provides explicit support of hierarchy

- Easier to manage the dimension
- Can be less efficient (due to join) than star schema

Example of Snowflake Schema

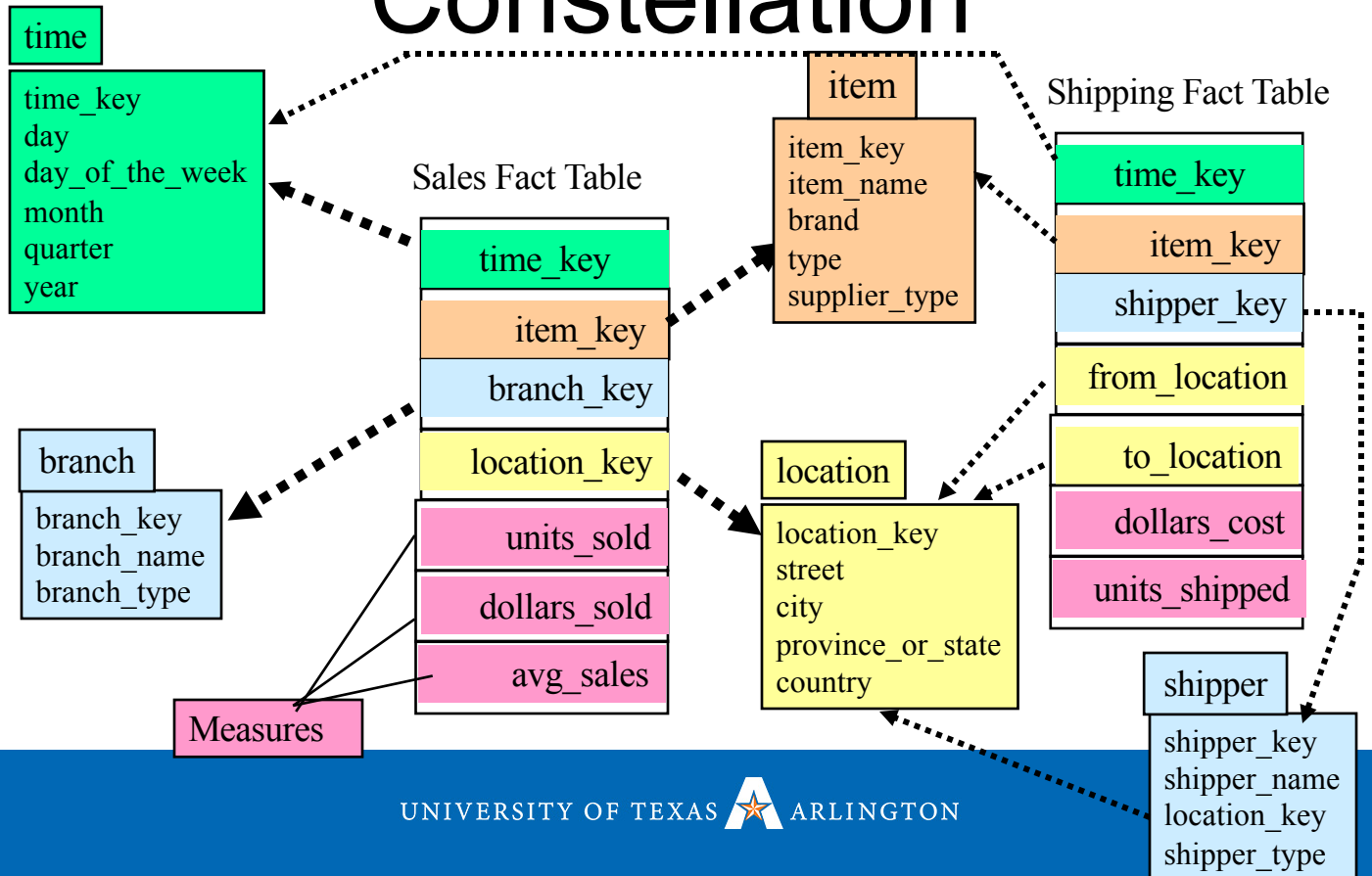


Conceptual Modeling of Data Warehouses

- Fact constellations: Multiple fact tables share dimension tables, viewed as a collection of stars,

  therefore called galaxy schema or fact constellation

Example of Fact Constellation



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*

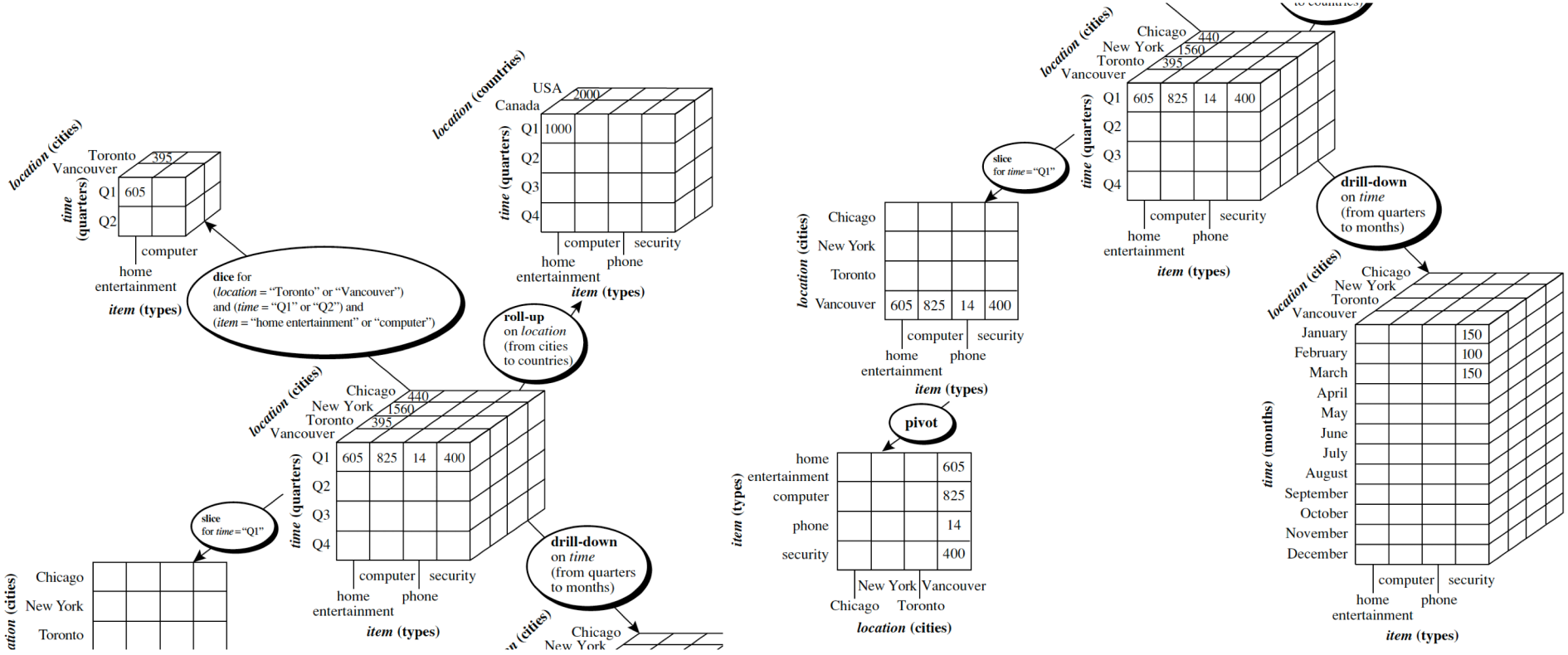
Roll up and Drill Down

- Roll up: increasing the level of aggregation
  - further aggregating along one more dimension
  - or further aggregating along the hierarchy of one dimension

- Drill down: decreasing the level of aggregating

It is like traversing in the lattice of cuboids.

Typical OLAP Operations



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?

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

- 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.

Measures of Data Cube: Three Categories

Consider aggregating a two dimensional set of values $\{X_{ij} \mid i = 1, \ldots, I;\ j = 1, \ldots, J\}$. Aggregate functions can be classified into three categories:

Distributive: Aggregate function $F()$ is distributive if there is a function $G()$ such that $F(\{X_{i,j}\}) = G(\{F(\{X_{i,j} \mid i = 1, \ldots, I\}) \mid j = 1, \ldots J\})$. `COUNT()`, `MIN()`, `MAX()`, `SUM()` are all distributive. In fact, $F = G$ for all but `COUNT()`. $G =$ `SUM()` for the `COUNT()` function. Once order is imposed, the cumulative aggregate functions also fit in the distributive class.

Algebraic: Aggregate function $F()$ is algebraic if there is an M-tuple valued function $G()$ and a function $H()$ such that $F(\{X_{i,j}\}) = H(\{G(\{X_{i,j} \mid i = 1, \ldots, I\}) \mid j = 1, \ldots, J\})$. Average(), standard deviation, MaxN(), MinN(), center_of_mass() are all algebraic. For Average, the function $G()$ records the sum and count of the subset. The $H()$ function adds these two components and then divides to produce the global average. Similar techniques apply to finding the N largest values, the center of mass of group of objects, and other algebraic functions. The key to algebraic functions is that a fixed size result (an M-tuple) can summarize the sub-aggregation.

Holistic: Aggregate function $F()$ is holistic if there is no constant bound on the size of the storage needed to describe a sub-aggregate. That is, there is no constant M, such that an M-tuple characterizes the computation $F(\{X_{i,j} \mid i = 1, \ldots, I\})$. Median(), MostFrequent() (also called the Mode()), and Rank() are common examples of holistic functions.