

Data Warehousing and OLAP

Prepared by Dr. Amjad Qtaish

DS601
Data Mining

Lecture 04



Lecture: #04 Data Warehousing and



Dr. Amjad Qtaish

Faculty of Informatics
Midocean University



Data Mining (DS601)



Data Warehousing and On-line Analytical Processing (OLAP)

- Data Warehouse: Basic Concepts 
- Data Warehouse Modeling: Data Cube and OLAP
- Data Warehouse Design and Usage
- Data Warehouse Implementation
- Data Generalization by Attribute-Oriented Induction
- Summary





- Some popular OLAP server software programs include:

- Oracle Express Server

<https://www.oracle.com/technetwork/database/options/olap/index.html>

- Hyperion Solutions Essbase

- Pentaho (Mondrian)

<https://community.hitachivantara.com/docs/DOC-1009853-mondrian>

documentation <https://mondrian.pentaho.com/documentation/installation.php>



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



What is a Data Warehouse?



☐ Subject-Oriented

- It provides a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process

☐ Integrated

- Constructed by integrating multiple, heterogeneous data sources, relational databases, flat files, on-line transaction records.

☐ 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)



What is a 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*



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



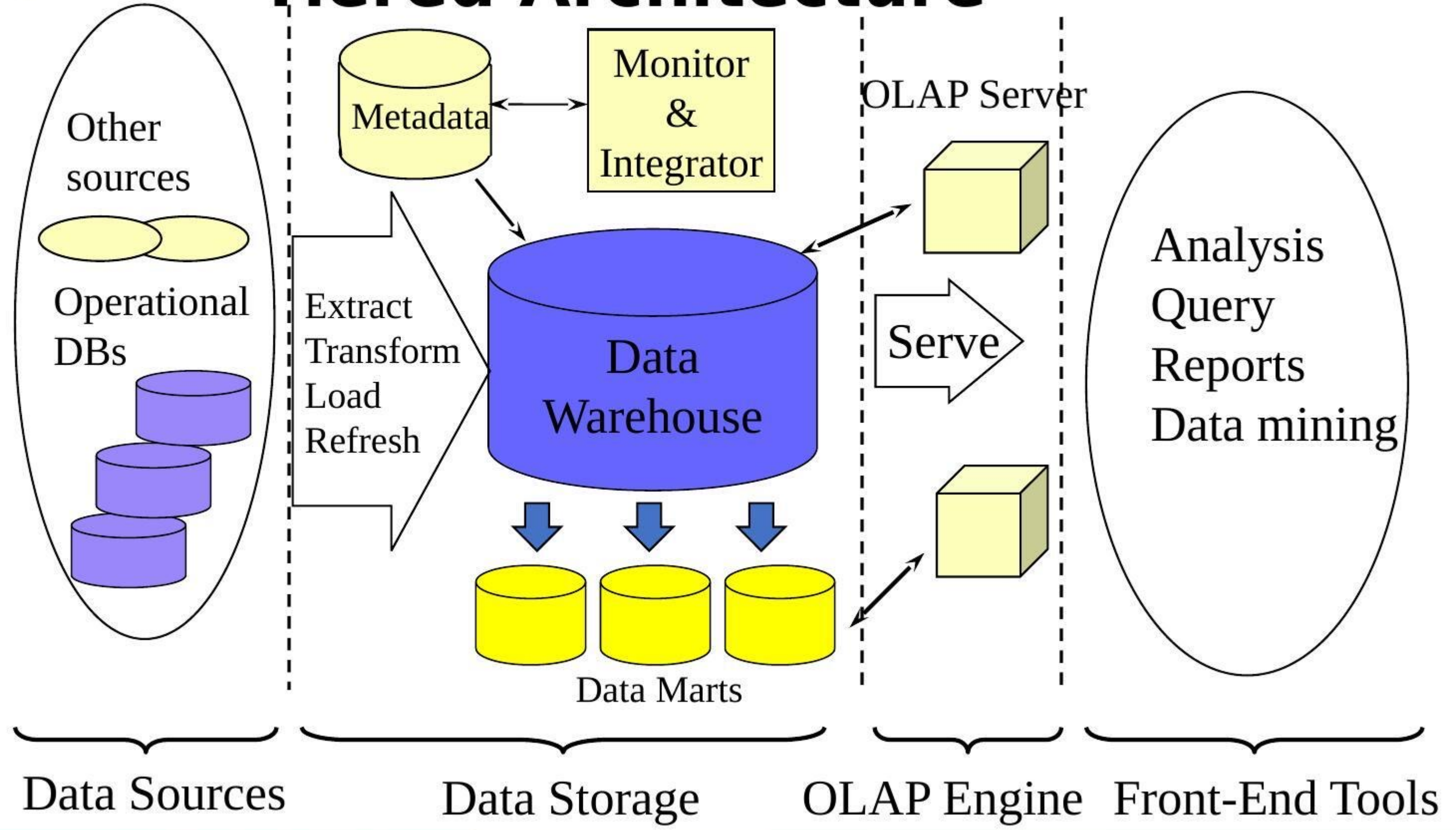
Why a Separate Data Warehouse?



- 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: A Multi-Tiered Architecture





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



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



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



Data Warehousing and On-line Analytical Processing (OLAP)

- Data Warehouse: Basic Concepts
- Data Warehouse Modeling: Data Cube and OLAP
- Data Warehouse Design and Usage
- Data Warehouse Implementation
- Data Generalization by Attribute-Oriented Induction
- 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, 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**.



Data Cube: A Multidimensional Data

2D

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

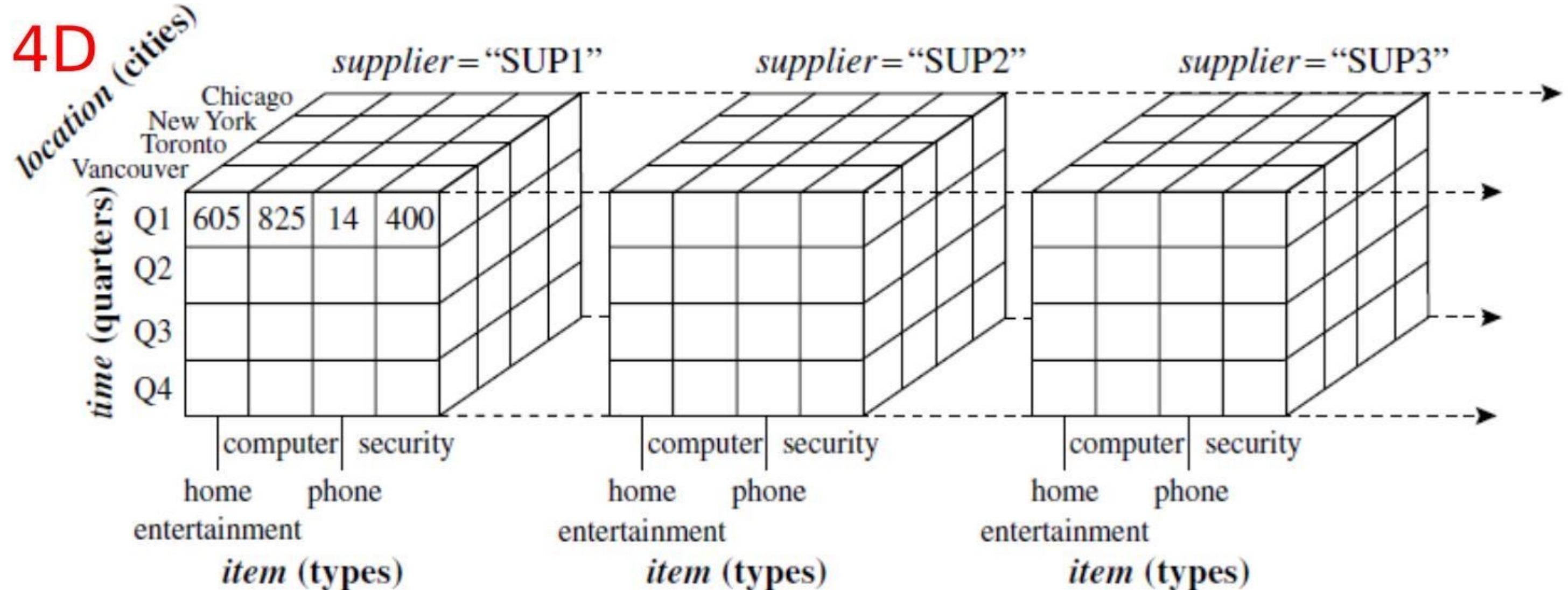
3-D View of Sales Data for *AllElectronics* According to *time*, *item*, and *location*

3D	location = "Chicago"					location = "New York"					location = "Toronto"					location = "Vancouver"				
	item					item					item					item				
	time	ent.	comp.	phone	sec.	time	ent.	comp.	phone	sec.	time	ent.	comp.	phone	sec.	time	ent.	comp.	phone	sec.
	Q1	854	882	89	623	Q1	1087	968	38	872	Q1	818	746	43	591	Q1	605	825	14	400
	Q2	943	890	64	698	Q2	1130	1024	41	925	Q2	894	769	52	682	Q2	680	952	31	512
	Q3	1032	924	59	789	Q3	1034	1048	45	1002	Q3	940	795	58	728	Q3	812	1023	30	501
	Q4	1129	992	63	870	Q4	1142	1091	54	984	Q4	978	864	59	784	Q4	927	1038	38	580

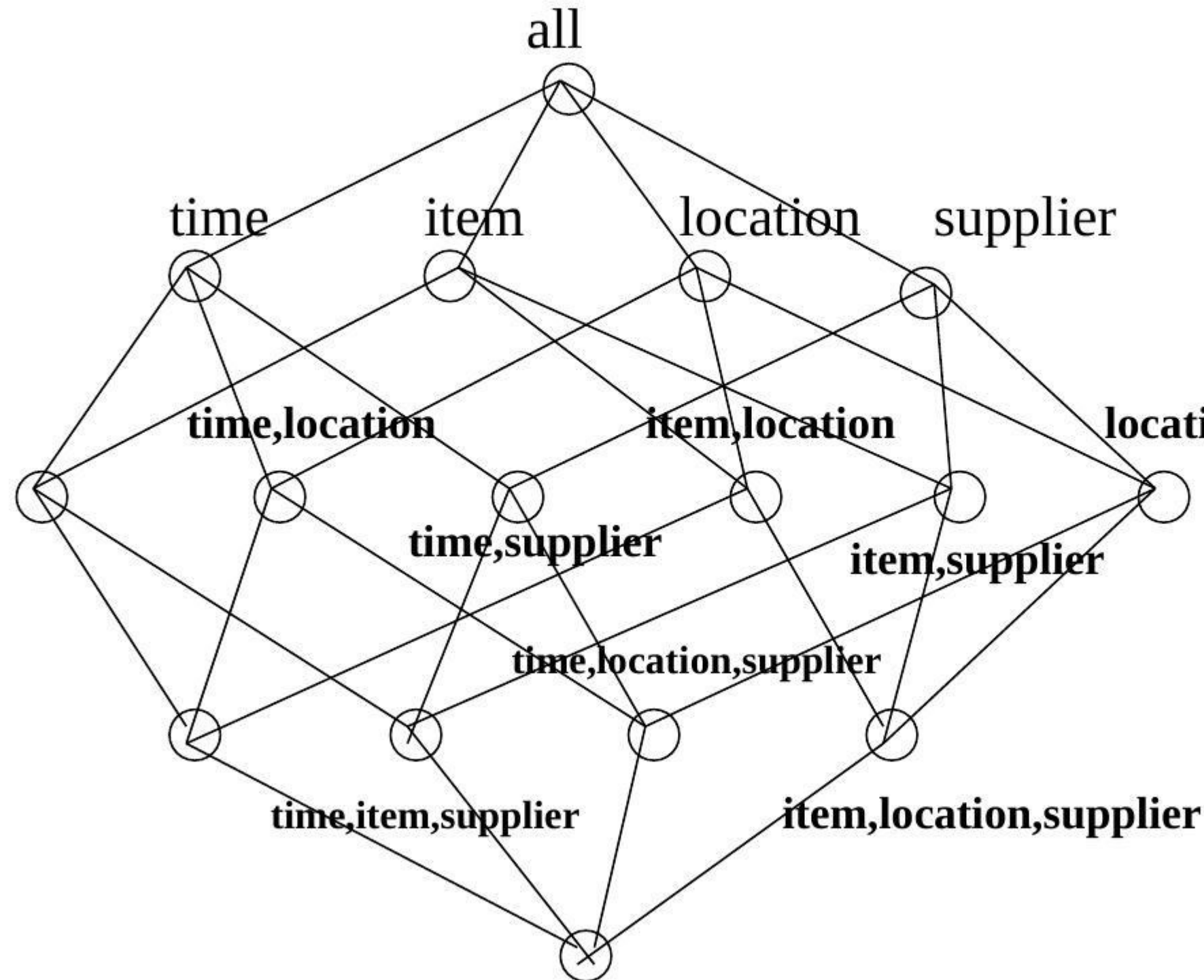
Note: The measure displayed is *dollars_sold* (in thousands).



Data Cube: A Multidimensional Data



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.



0-D (*apex*) cuboid

1-D cuboids

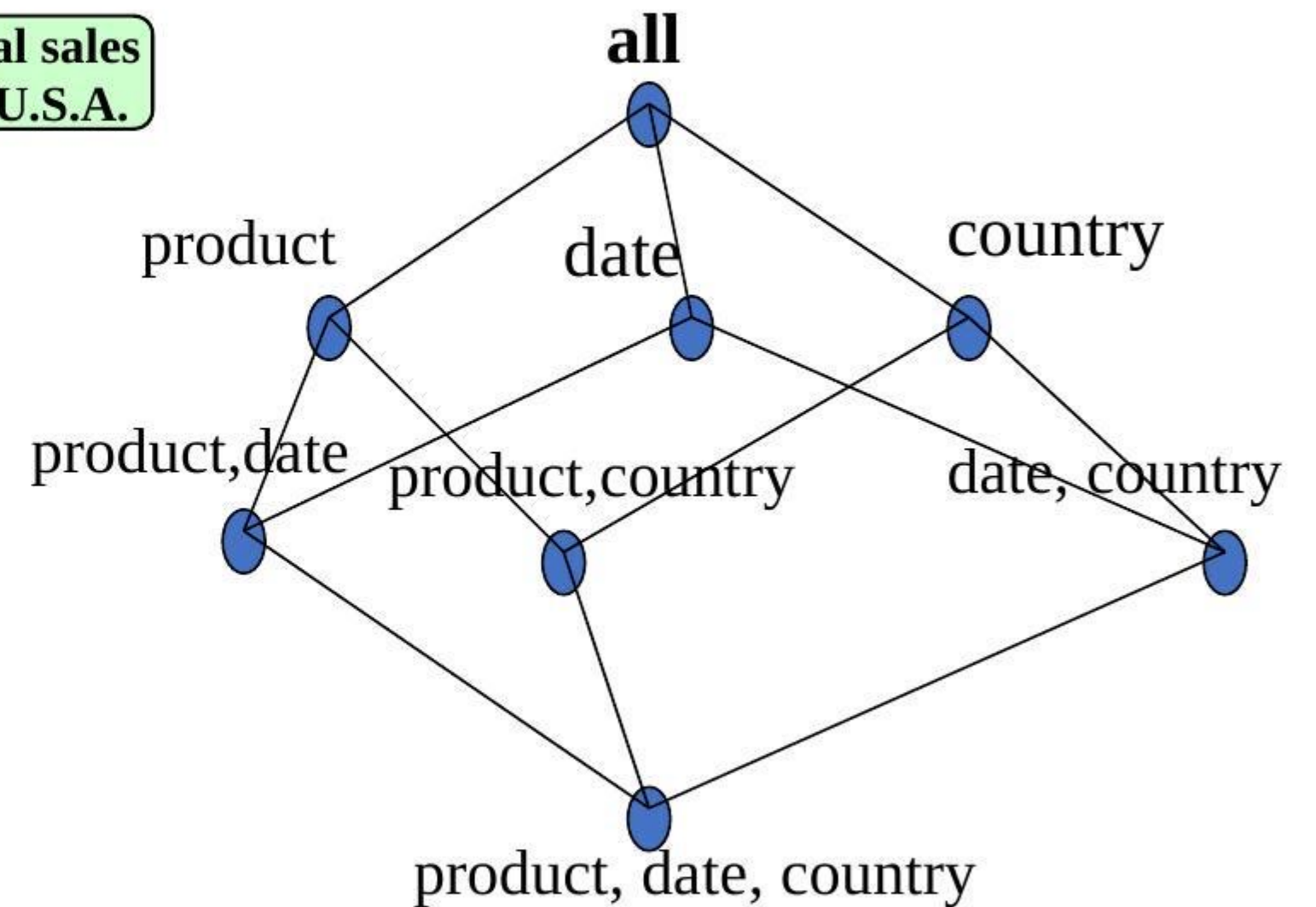
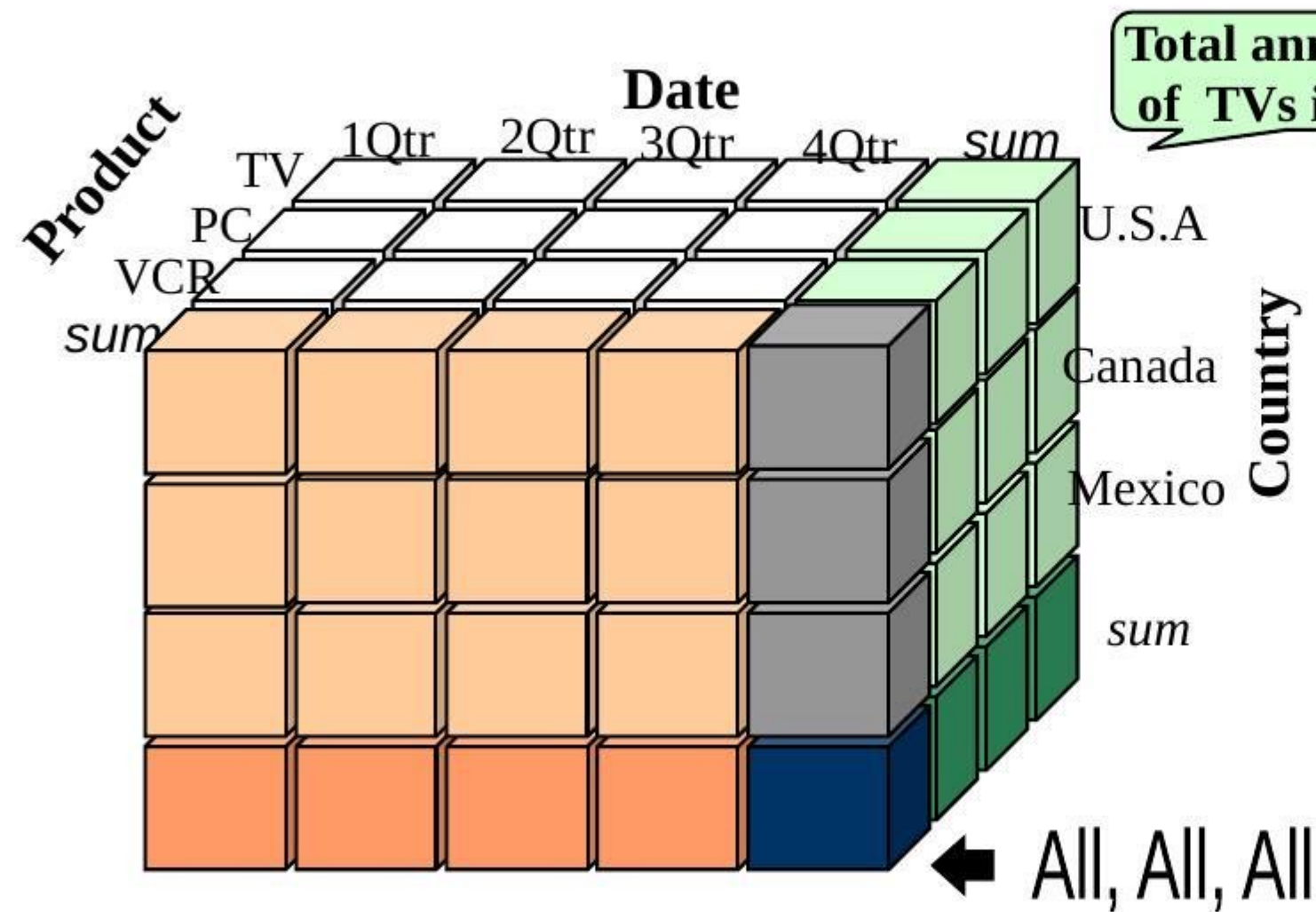
2-D cuboids

3-D cuboids

4-D (*base*) cuboid



A Sample Data Cube





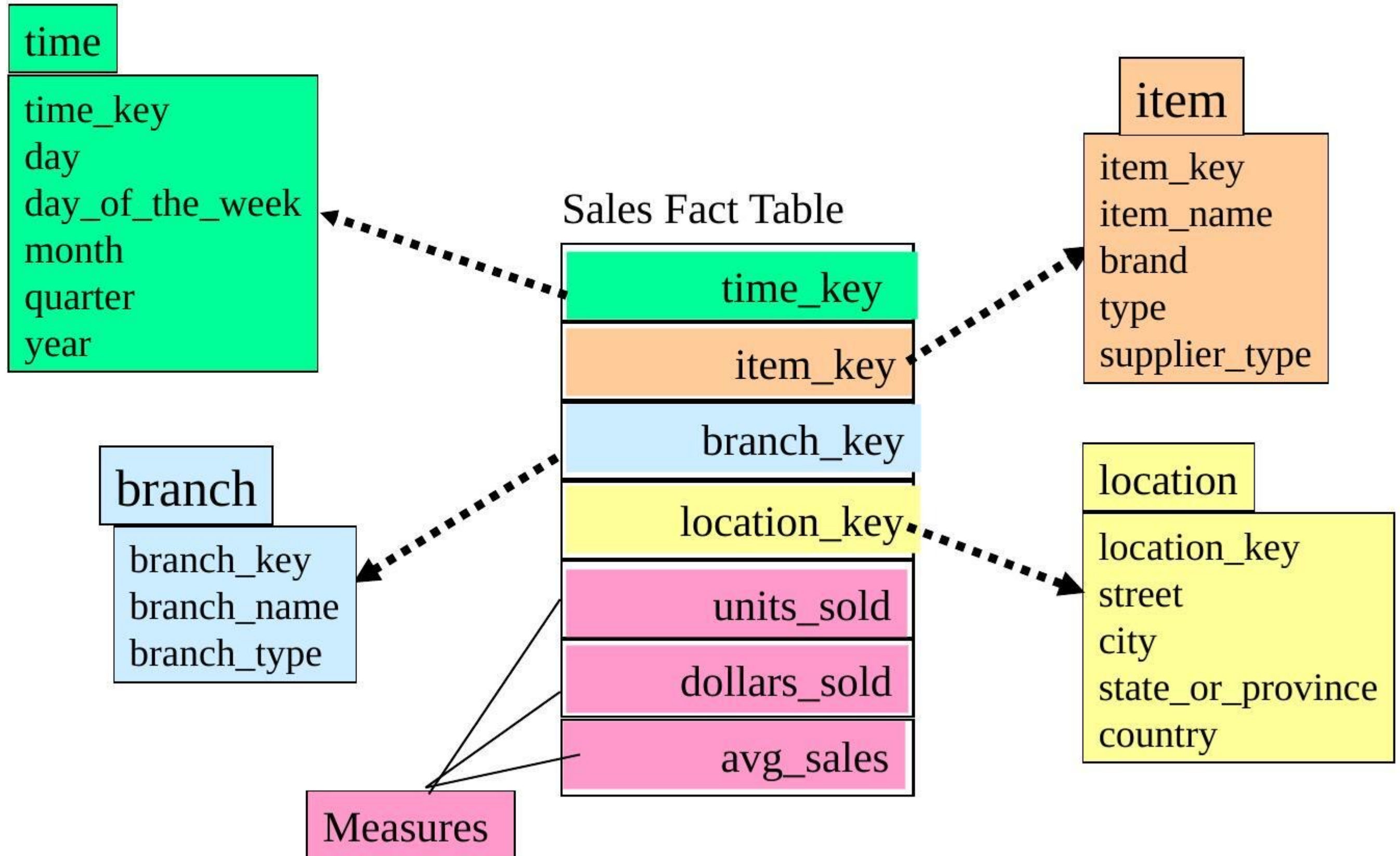
Conceptual Modeling of Data Warehouses

- 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



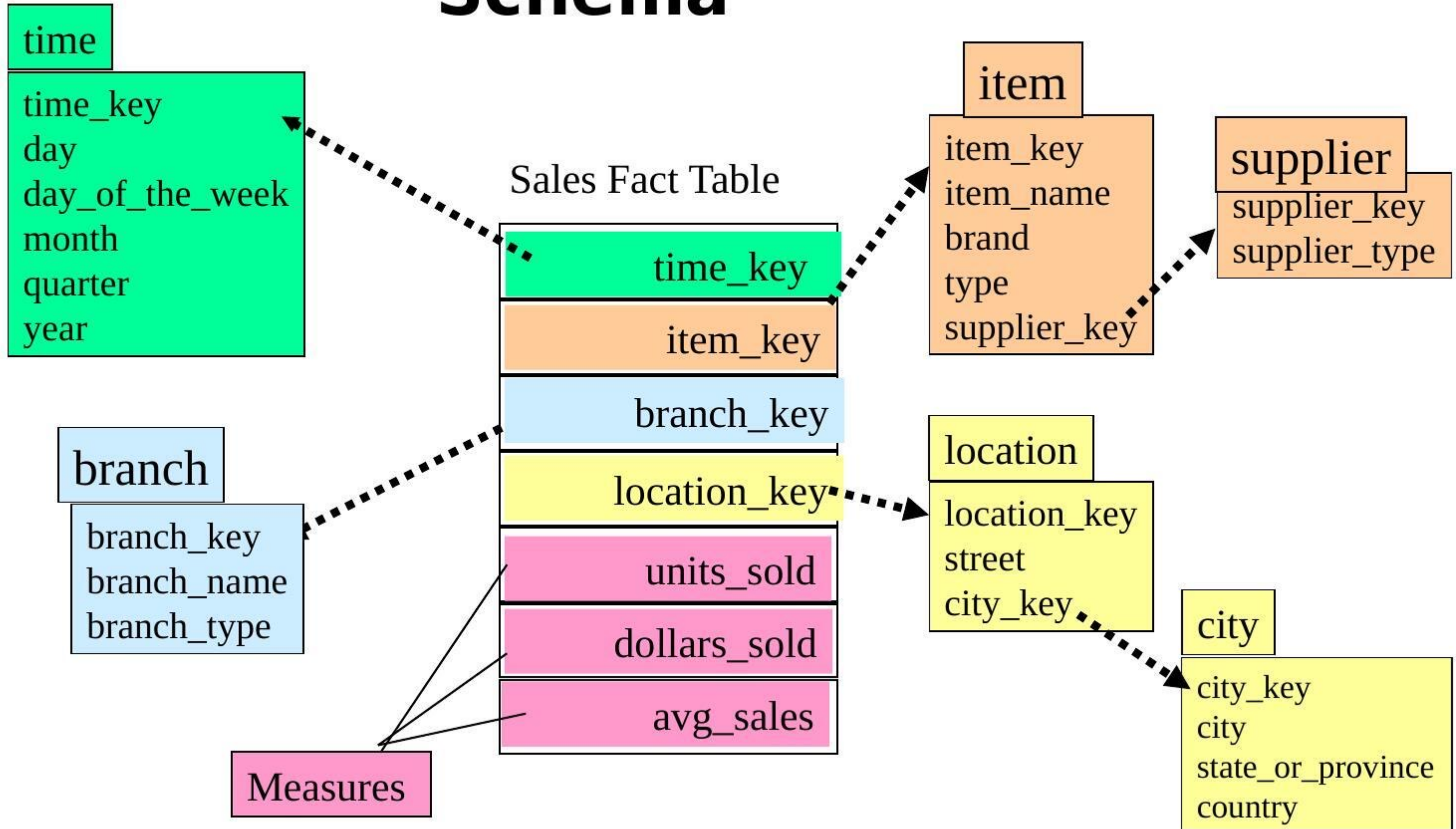


Example of Star Schema



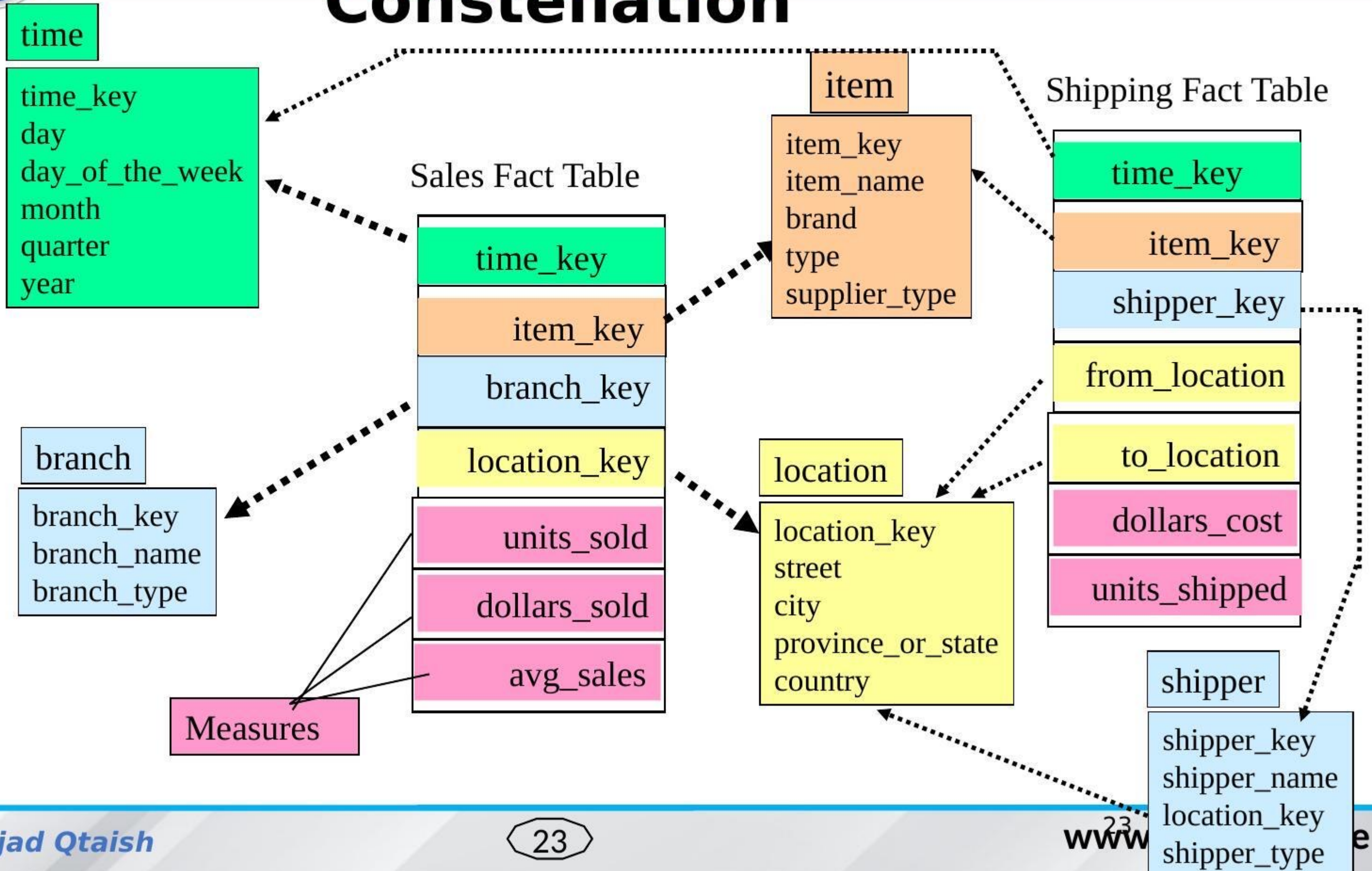


Example of Snowflake Schema



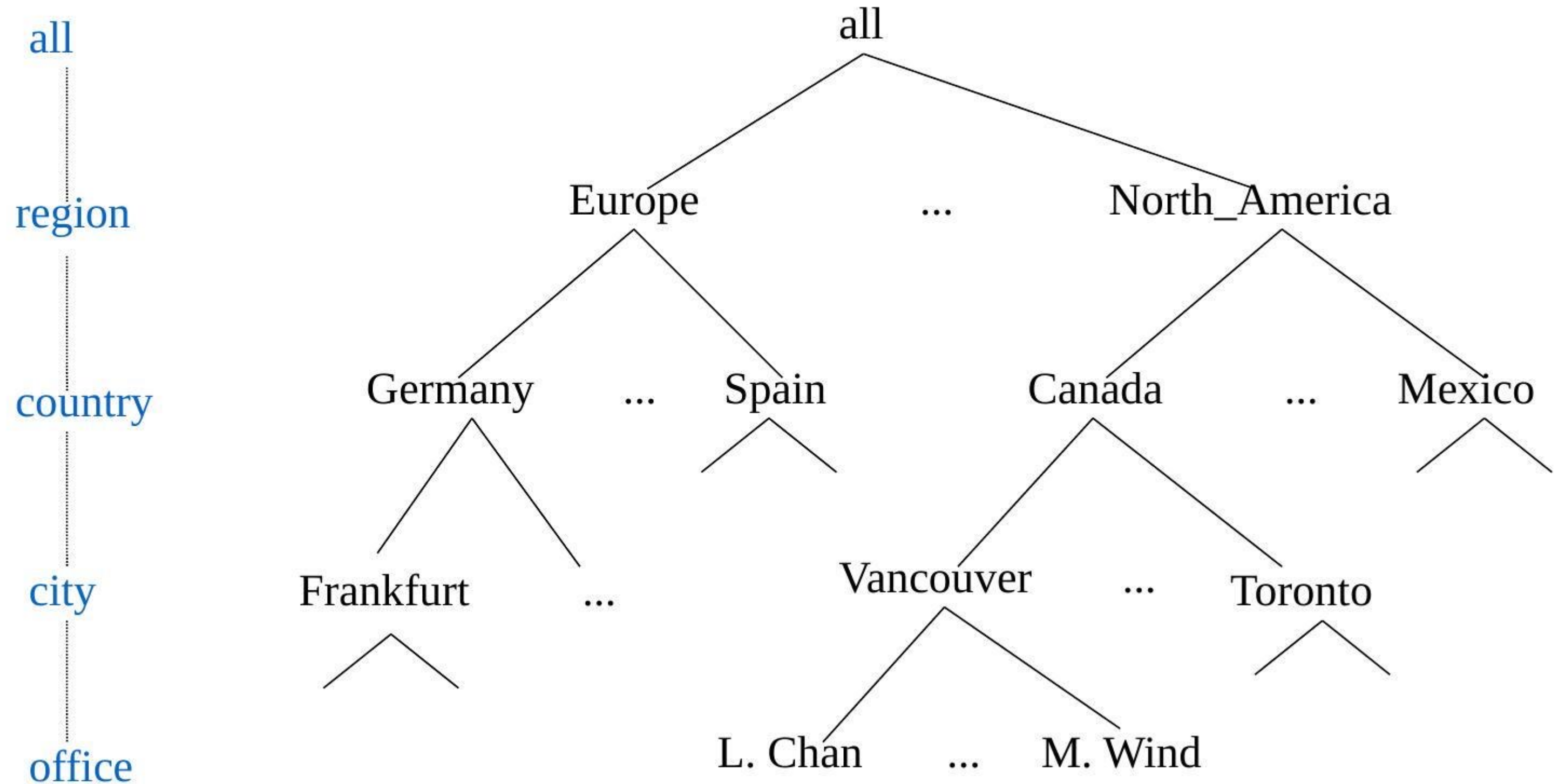


Example of Fact Constellation



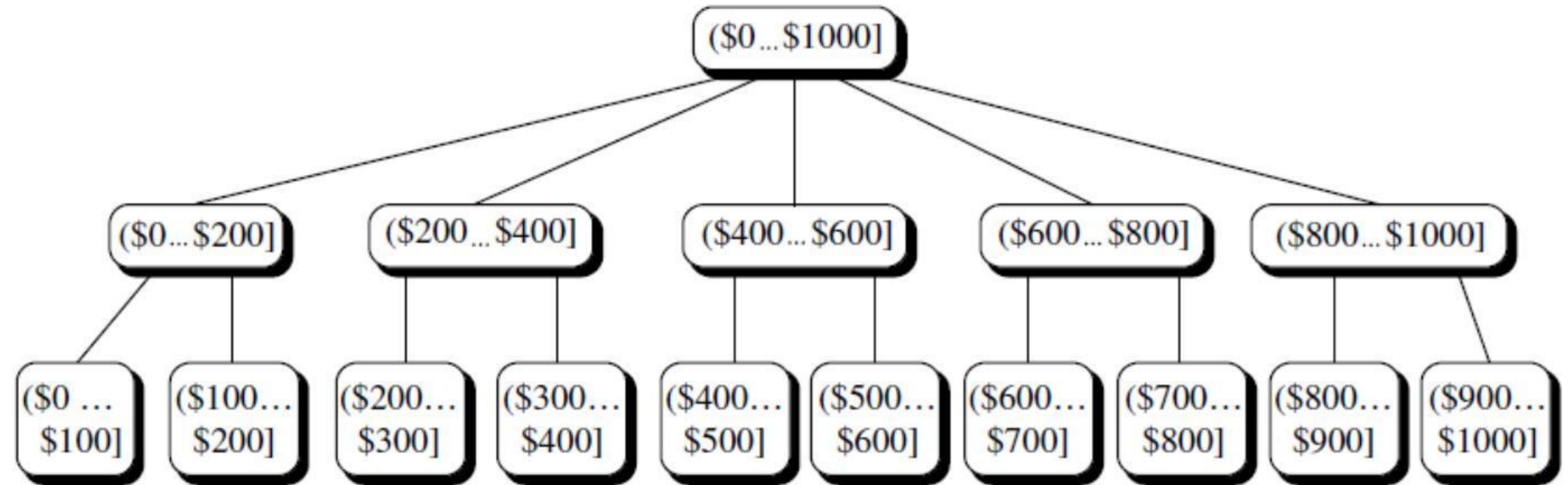


A Concept Hierarchy: Dimension (location)





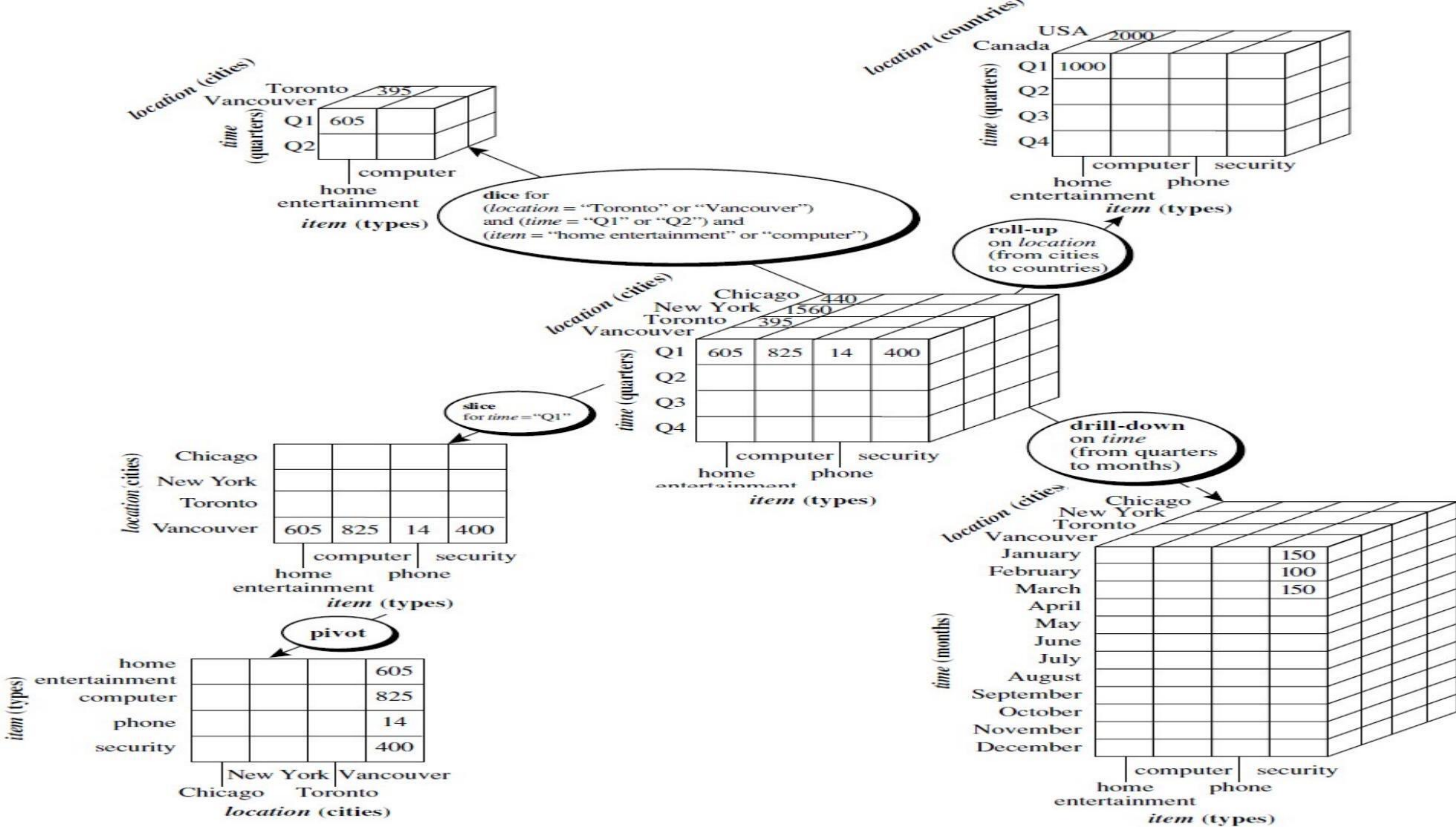
A Concept Hierarchy: Dimension (location)



A concept hierarchy for *price*.



- 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)*





Data Warehousing and On-line Analytical Processing (OLAP)

- Data Warehouse: Basic Concepts
- Data Warehouse Modeling: Data Cube and OLAP
- Data Warehouse Design and Usage
- Data Warehouse Implementation
- Data Generalization by Attribute-Oriented Induction
- Summary





Design of Data Warehouse: A Business Analysis Framework

- Four views regarding the design of a data warehouse
 - **Top-down view**
 - allows selection of the relevant information necessary for the data warehouse
 - **Data source view**
 - exposes the information being captured, stored, and managed by operational systems
 - **Data warehouse view**
 - consists of fact tables and dimension tables
 - **Business query view**
 - sees the perspectives of data in the warehouse from the view of end-user





- 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 and On-line Analytical Processing (OLAP)

- Data Warehouse: Basic Concepts
- Data Warehouse Modeling: Data Cube and OLAP
- Data Warehouse Design and Usage
- Data Warehouse Implementation 
- Data Generalization by Attribute-Oriented Induction
- Summary





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, materialization=precalculation
 - Compute every (cuboid) (**full materialization**), compute no cuboids (**no materialization**), or some cuboids (**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
```

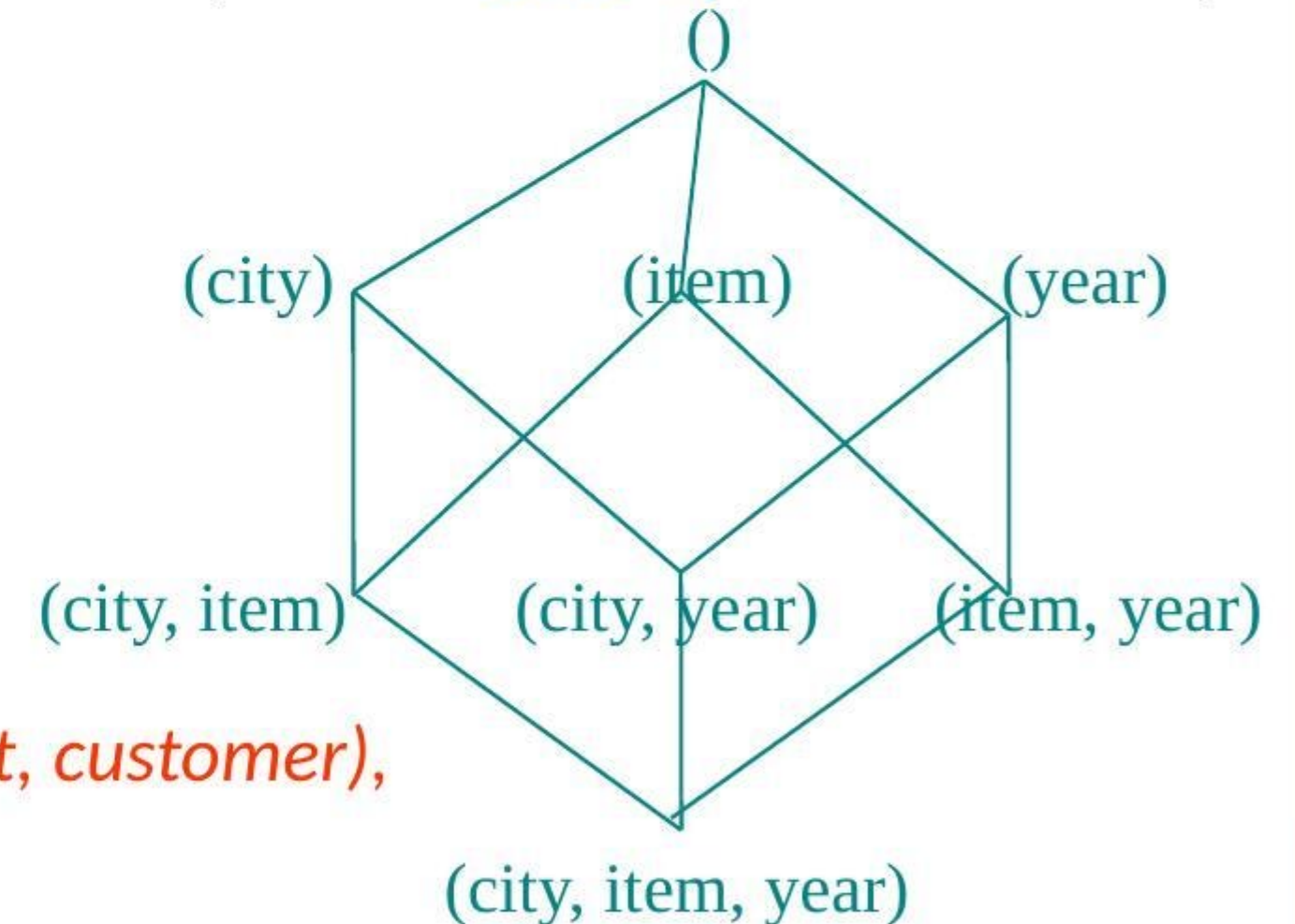
- Need compute the following Group-Bys

```
(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
- **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) $\{year, brand, country\}$
 - 3) $\{year, brand, province_or_state\}$
 - 4) $\{item_name, province_or_state\}$ where $year = 2004$

Which should be selected to process the query?



Data Warehousing and On-line Analytical Processing (OLAP)

- Data Warehouse: Basic Concepts
- Data Warehouse Modeling: Data Cube and OLAP
- Data Warehouse Design and Usage
- Data Warehouse Implementation
- Data Generalization by Attribute-Oriented Induction
- Summary





Attribute-Oriented Induction



- Not confined to categorical data nor particular measures
- How it is done?
 - Collect the task-relevant data (*initial relation*) using a relational database query
 - Perform generalization by attribute removal or attribute generalization
 - Apply aggregation by merging identical, generalized tuples and accumulating their respective counts
 - Interaction with users for knowledge presentation



Attribute-Oriented Induction: An Example

Example: Describe general characteristics of graduate students in the University database



- Step 1. Fetch relevant set of data using an SQL statement, e.g.,
Select * (i.e., name, gender, major, birth_place, birth_date, residence, phone#, gpa)
from student
where student_status in {"Msc", "MBA", "PhD" }
- Step 2. Perform attribute-oriented induction
- Step 3. Present results in generalized relation, cross-tab, or rule forms



Class Characterization: An Example

Initial
Relation

Name	Gender	Major	Birth-Place	Birth_date	Residence	Phone #	GPA
Jim Woodman	M	CS	Vancouver,BC, Canada	8-12-76	3511 Main St., Richmond	687-4598	3.67
Scott Lachance	M	CS	Montreal, Que, Canada	28-7-75	345 1st Ave., Richmond	253-9106	3.70
Laura Lee	F	Physics	Seattle, WA, USA	25-8-70	125 Austin Ave., Burnaby	420-5232	3.83
...	...	...	...	...	...	...	...
Removed	Retained	Sci,Eng, Bus	Country	Age range	City	Removed	Excl, VG,..

Prime
Generalized
Relation

Gender	Major	Birth_region	Age_range	Residence	GPA	Count
M	Science	Canada	20-25	Richmond	Very-good	16
F	Science	Foreign	25-30	Burnaby	Excellent	22
...	...	...	...	...	...	...

Gender \ Birth_Region	Birth_Region		
	Canada	Foreign	Total
M	16	14	30
F	10	22	32
Total	26	36	62



Basic Principles of Attribute-Oriented Induction



- Data focusing: task-relevant data, including dimensions, and the result is the *initial relation*
- Attribute-removal: remove attribute A if there is a large set of distinct values for A but (1) there is no generalization operator on A, or (2) A's higher level concepts are expressed in terms of other attributes
- Attribute-generalization: If there is a large set of distinct values for A, and there exists a set of generalization operators on A, then select an operator and generalize A
- Attribute-threshold control: typical 2-8, specified/default
- Generalized relation threshold control: control the final relation/rule size



Attribute-Oriented Induction: Basic Algorithm



- InitialRel: Query processing of task-relevant data, deriving the *initial relation*.
- PreGen: Based on the analysis of the number of distinct values in each attribute, determine generalization plan for each attribute: removal? or how high to generalize?
- PrimeGen: Based on the PreGen plan, perform generalization to the right level to derive a “prime generalized relation”, accumulating the counts.
- Presentation: User interaction: (1) adjust levels by drilling, (2) pivoting, (3) mapping into rules, cross tabs, visualization presentations.



Presentation of Generalized Results



- Generalized relation:
 - Relations where some or all attributes are generalized, with counts or other aggregation values accumulated.
- Cross tabulation:
 - Mapping results into cross tabulation form (similar to contingency tables).
 - Visualization techniques:
 - Pie charts, bar charts, curves, cubes, and other visual forms.
- Quantitative characteristic rules:
 - Mapping generalized result into characteristic rules with quantitative information associated with it, e.g.,



Data Warehousing and On-line Analytical Processing (OLAP)

- Data Warehouse: Basic Concepts
- Data Warehouse Modeling: Data Cube and OLAP
- Data Warehouse Design and Usage
- Data Warehouse Implementation
- Data Generalization by Attribute-Oriented Induction
- 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, Design, and Usage**
 - Multi-tiered architecture
 - Business analysis design framework
- **Implementation:** Efficient computation of data cubes
 - Partial vs. full vs. no materialization
 - Indexing OALP data: Bitmap index and join index
 - OLAP query processing
- **Data generalization:** Attribute-oriented induction



Thank You