DATA MINING

CHAPTER 4 – DATA WAREHOUSING & OLAP

Dr. Ahmed Said

CHAPTER 4: DATA WAREHOUSING & OLAP

- Data Warehouse: Basic Concepts
- Data Warehouse Modeling: Data Cube
- 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
    - 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

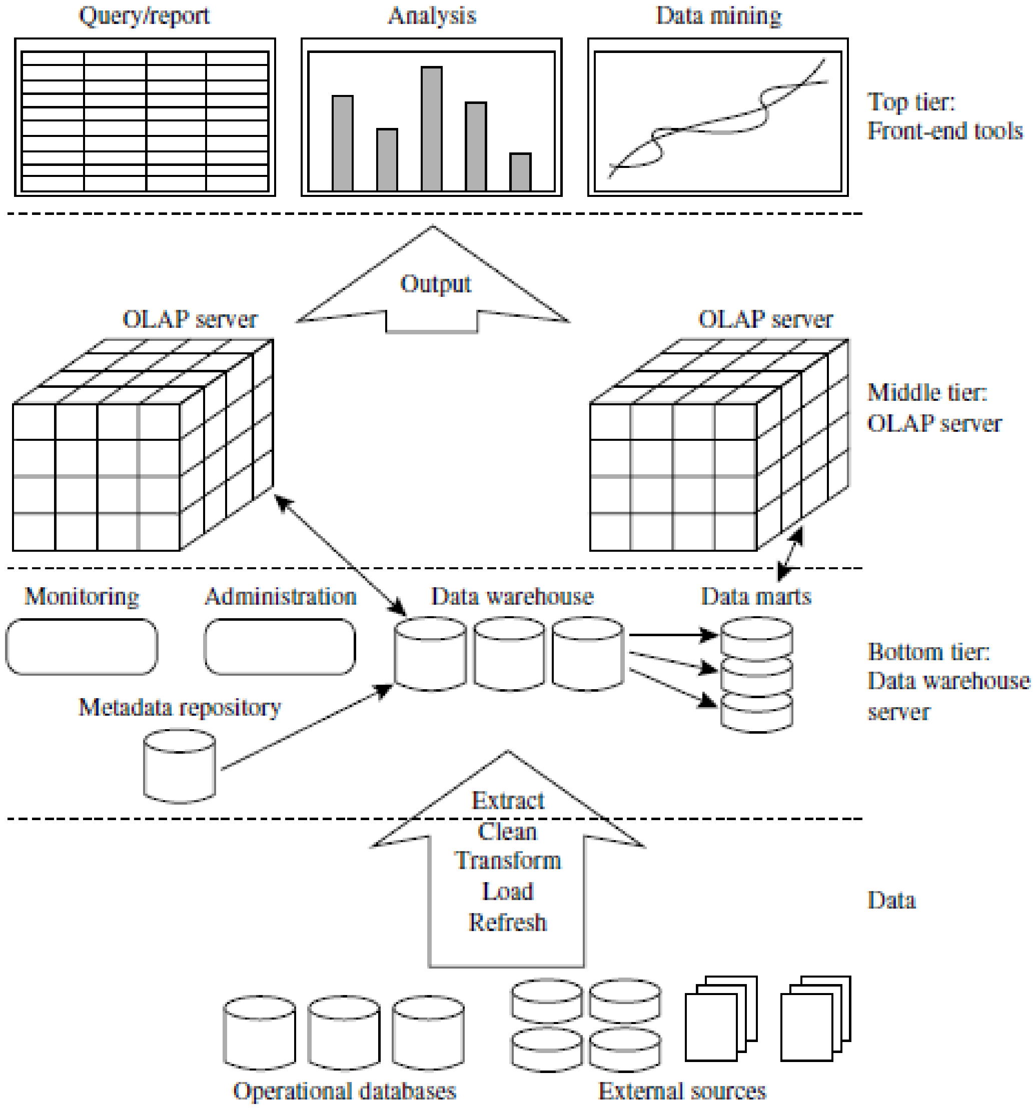
OLTP VS. OLAP

	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
Usage	Repetitive	Ad-hoc
Access	Read/Write Index/Hash On Prim. Key	Lots Of Scans
Unit Of Work	Short, Simple Transaction	Complex Query
# Of Records Accessed	Tens	Millions
# Of 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



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

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 define, 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

CHAPTER 4: DATA WAREHOUSING & OLAP

- Data Warehouse: Basic Concepts
- Data Warehouse Modeling: Data Cube
- 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 topmost **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 MODEL

2-D View of Sales Data for *AllElectronics* According to *time* and *item*

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

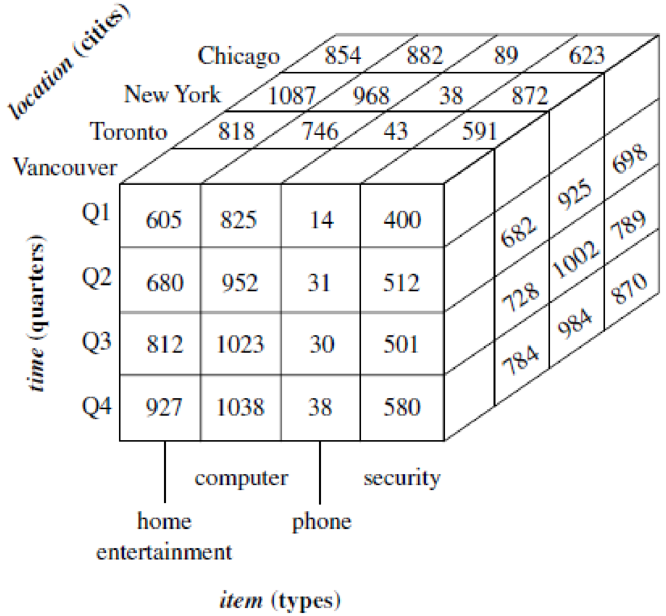
	location = "Chicago"				**location** = "New York"				**location** = "Toronto"				**location** = "Vancouver"			
	item				**item**				**item**				**item**			
time	home ent.	comp.	phone	sec.	home ent.	comp.	phone	sec.	home ent.	comp.	phone	sec.	home ent.	comp.	phone	sec.
Q1	854	882	89	623	1087	968	38	872	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

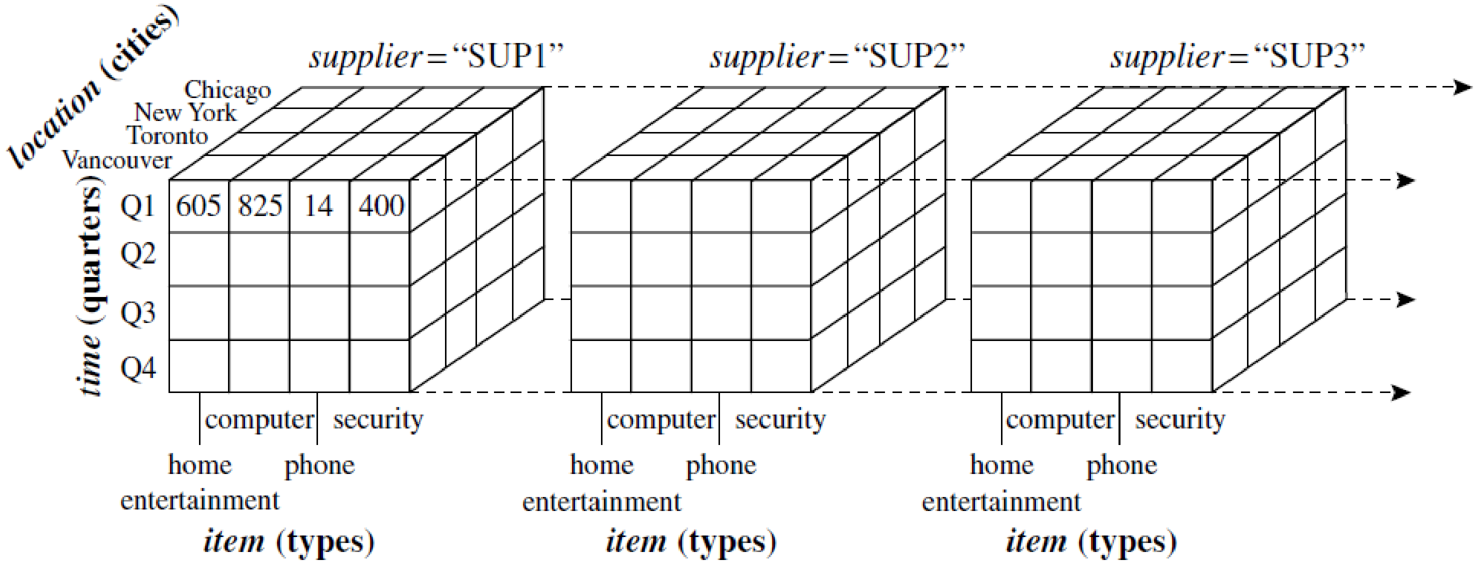
Note: The measure displayed is *dollars_sold* (in thousands).

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

DATA CUBE: A MULTIDIMENSIONAL DATA MODEL

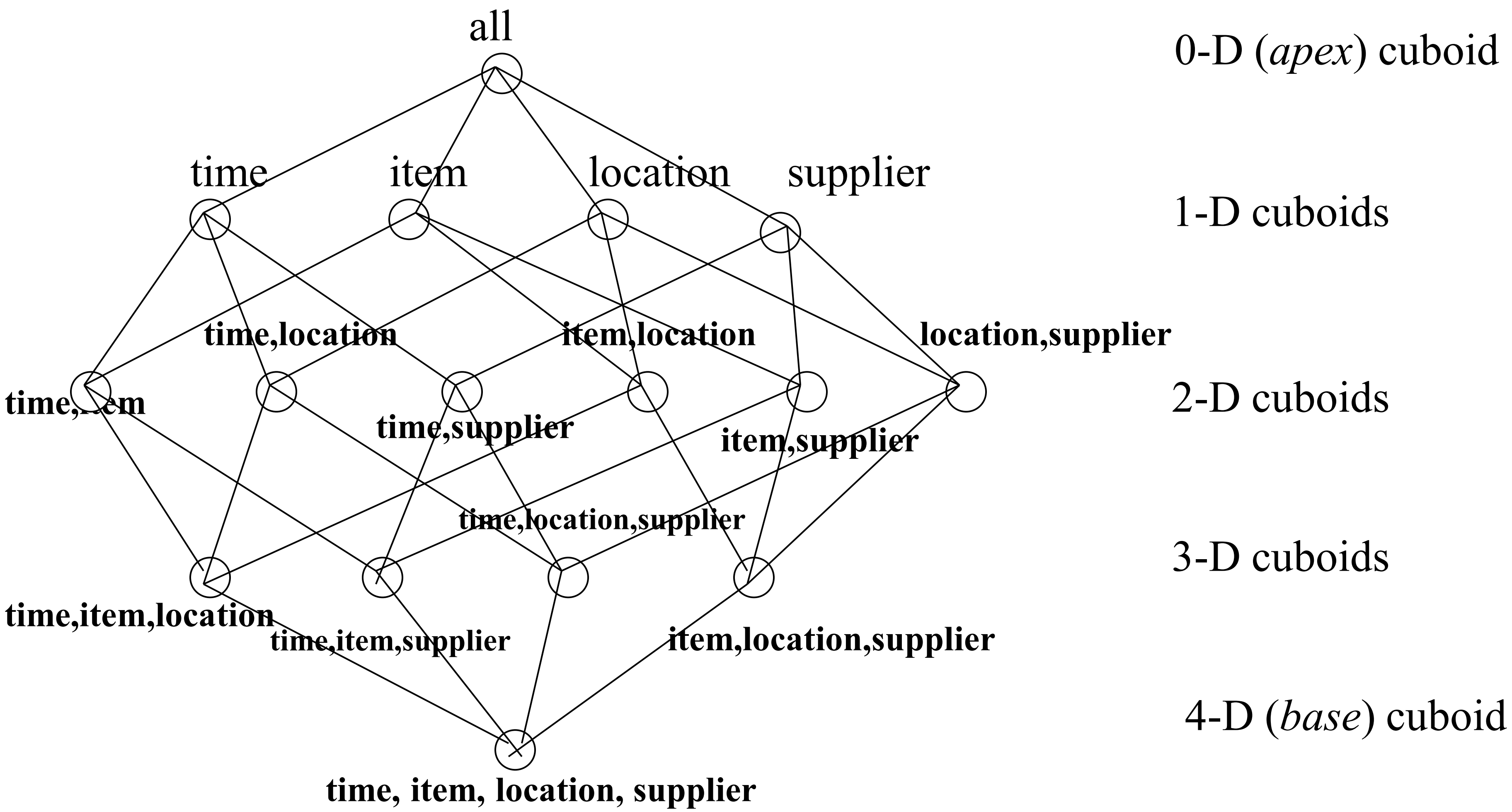
A 3-D data cube representation of the data in the previous table, according to *time*, *item*, and *location*. The measure displayed is *dollars sold* (in thousands).





A 4-D data cube representation of sales data, according to *time*, *item*, *location*, and *supplier*. The measure displayed is *dollars sold* (in thousands).

CUBE: A LATTICE OF CUBOIDS



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

OLTP EXAMPLE

customer

cust_ID	name	address	age	income	credit_info	category	...
C1	Smith, Sandy	1223 Lake Ave., Chicago, IL	31	$78000	1	3	...
...	...	...	...	...	...	...	...

item

item_ID	name	brand	category	type	price	place_made	supplier	cost
I3	hi-res-TV	Toshiba	high resolution	TV	$988.00	Japan	NikoX	$600.00
I8	Laptop	Dell	laptop	computer	$1369.00	USA	Dell	$983.00
...	...	...	...	...	...	...	...	...

employee

empl_ID	name	category	group	salary	commission
E55	Jones, Jane	home entertainment	manager	$118,000	2%
...	...	...	...	...	...

branch

branch_ID	name	address
B1	City Square	396 Michigan Ave, Chicago, IL
...	...	...

purchases

trans_ID	cust_ID	empl_ID	date	time	method_paid	amount
T100	C1	E55	03/21/2005	15:45	Visa	$1357.00
...	...	...	...	...	...	...

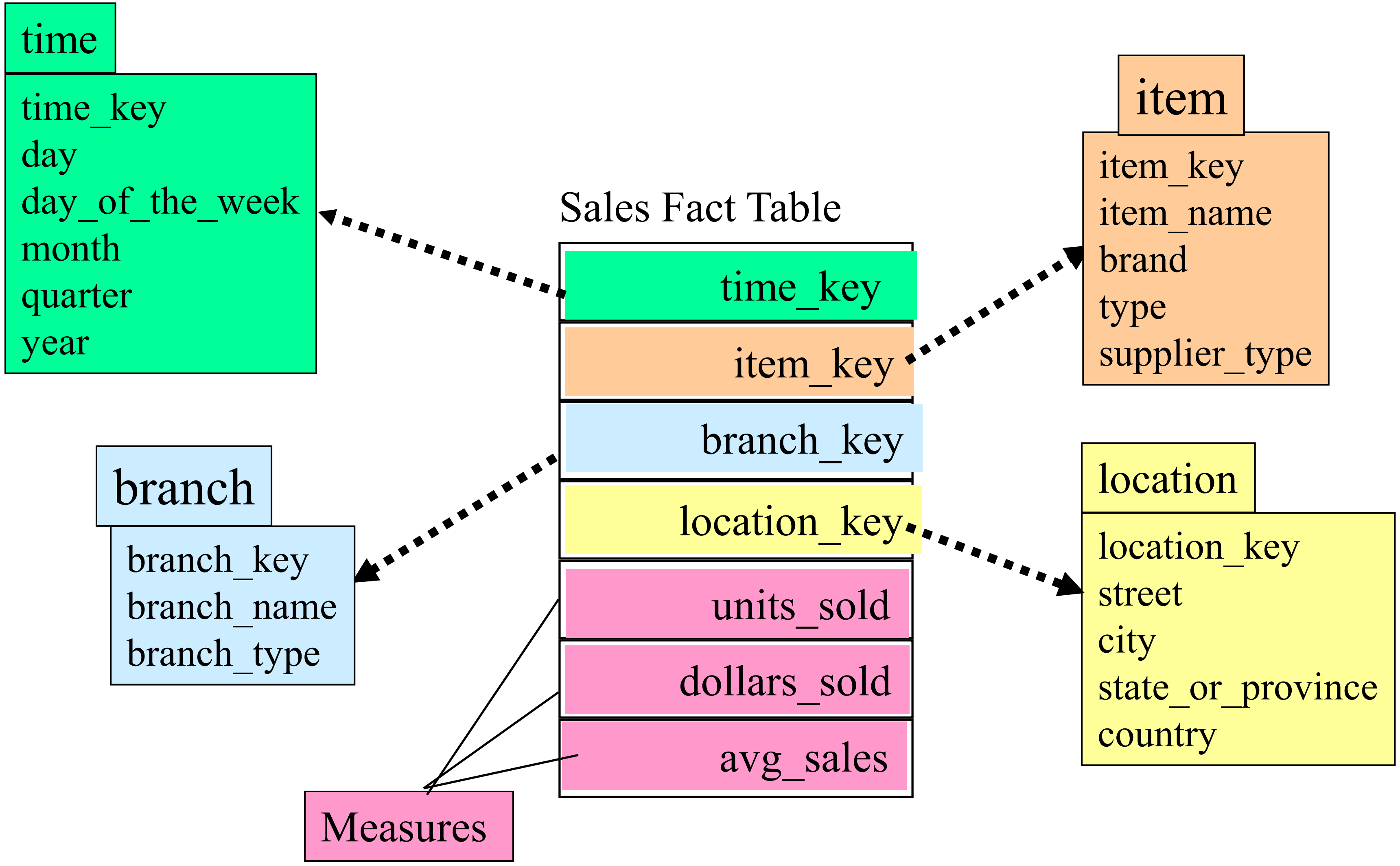
items_sold

trans_ID	item_ID	qty
T100	I3	1
T100	I8	2
...	...	...

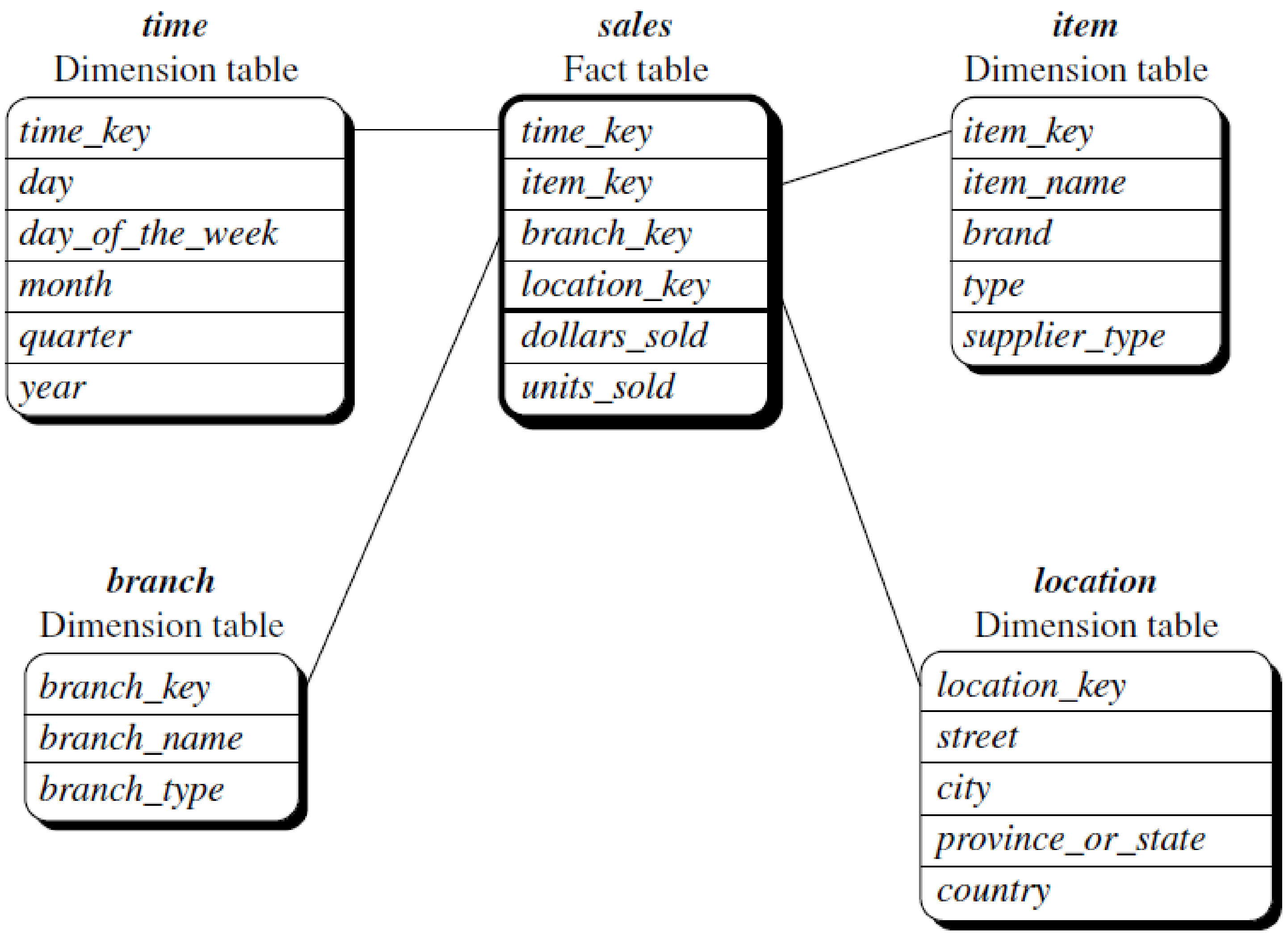
works_at

empl_ID	branch_ID
E55	B1
...	...

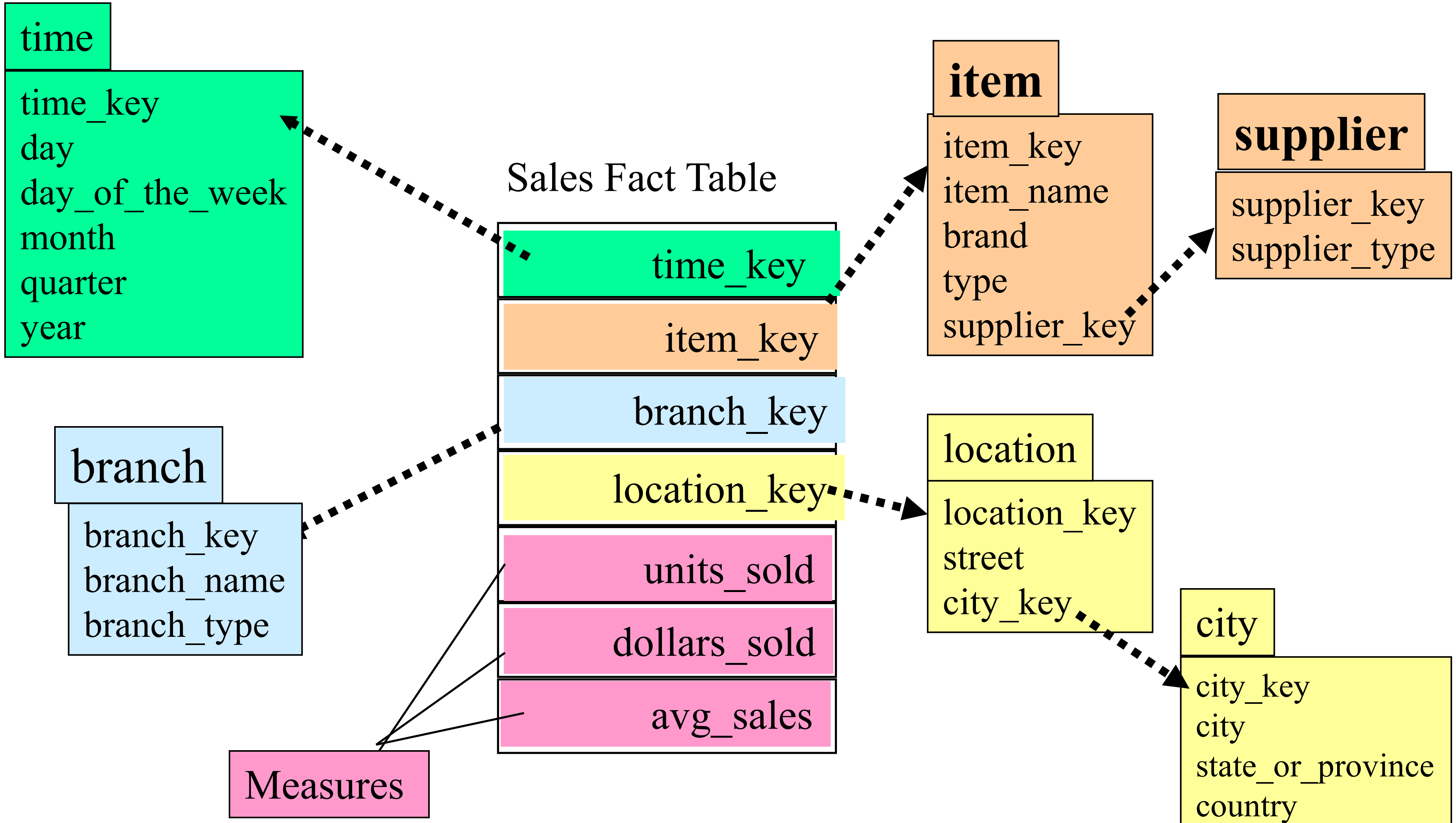
EXAMPLE OF STAR SCHEMA



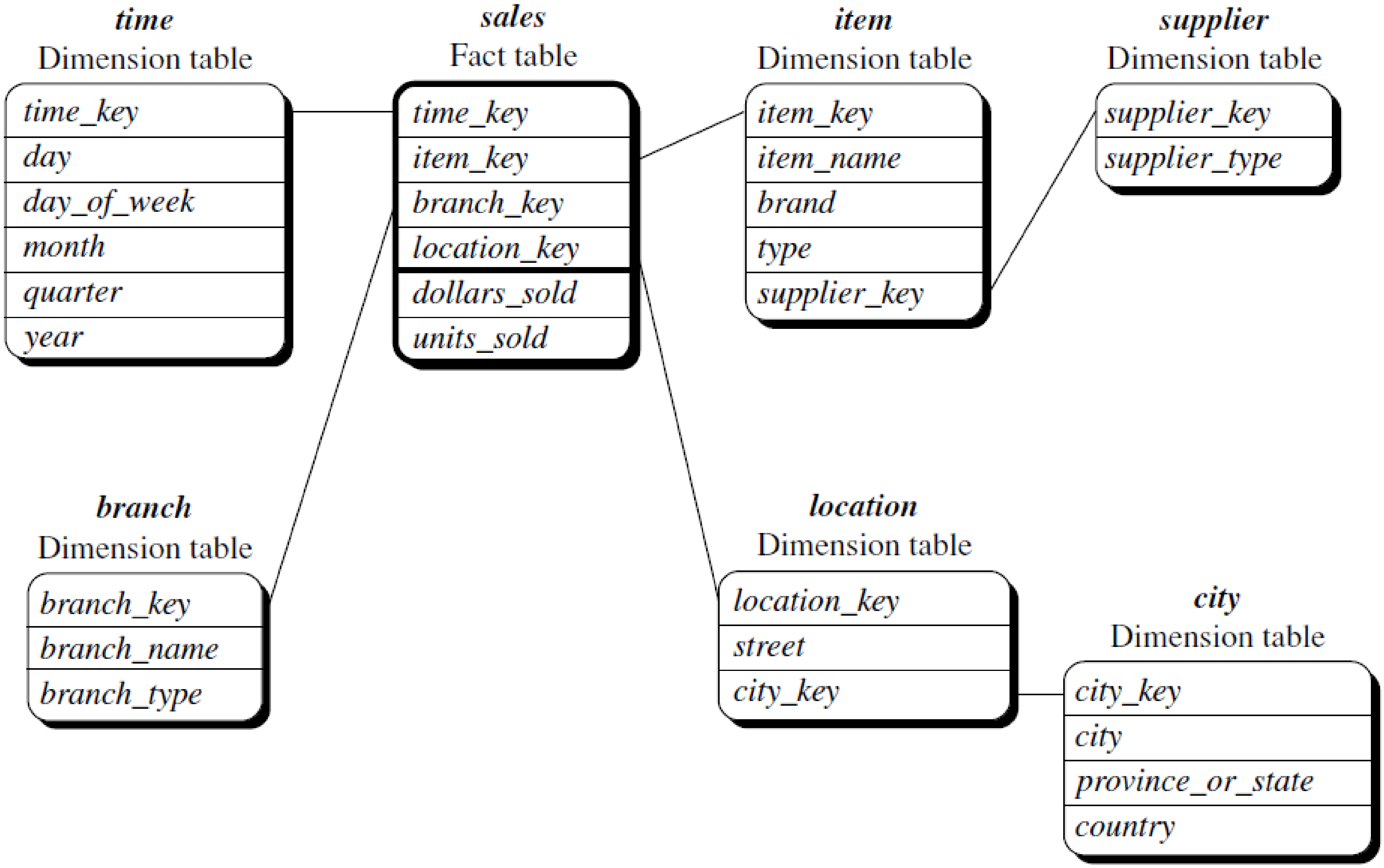
EXAMPLE OF STAR SCHEMA



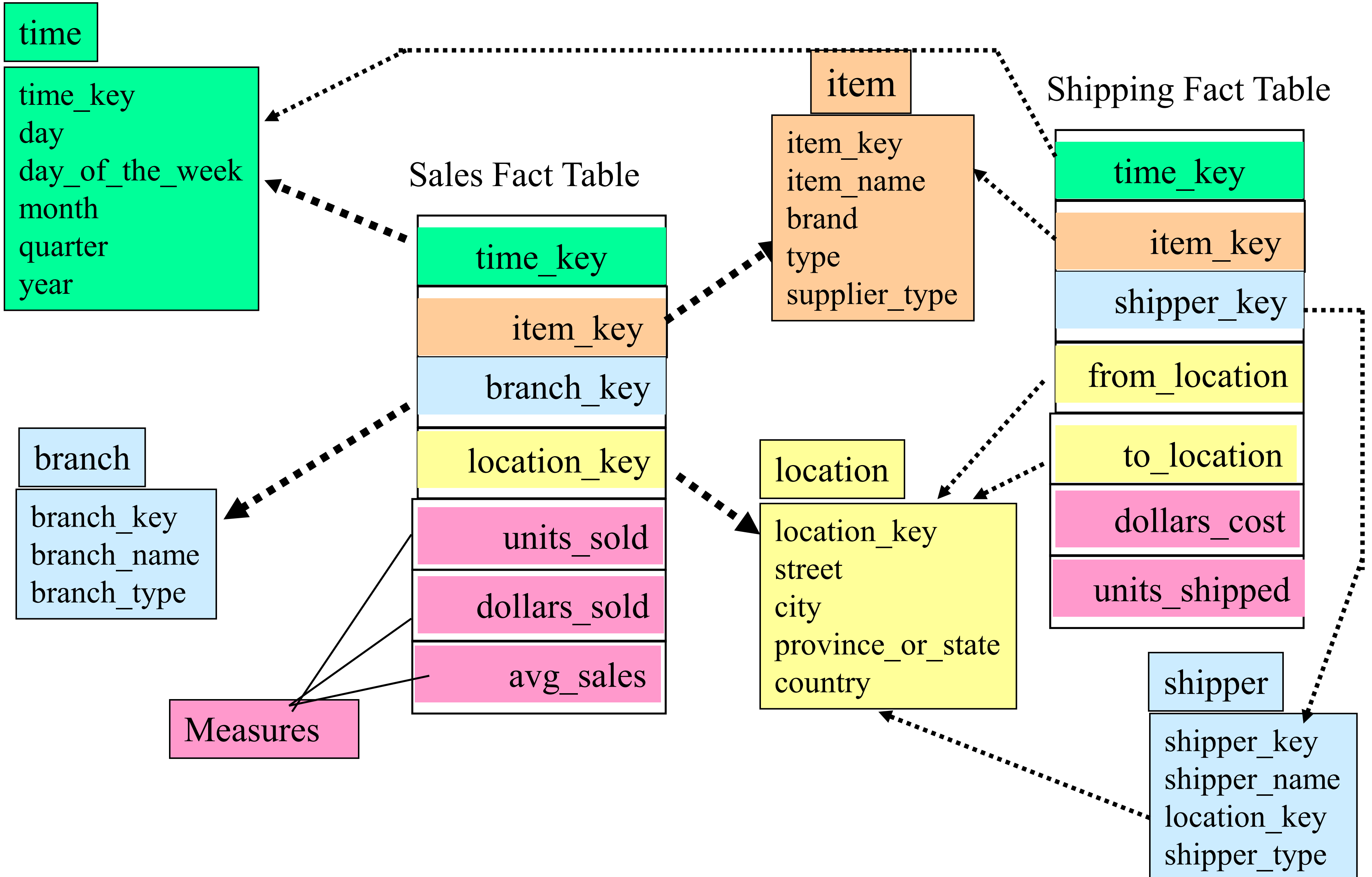
EXAMPLE OF SNOWFLAKE SCHEMA



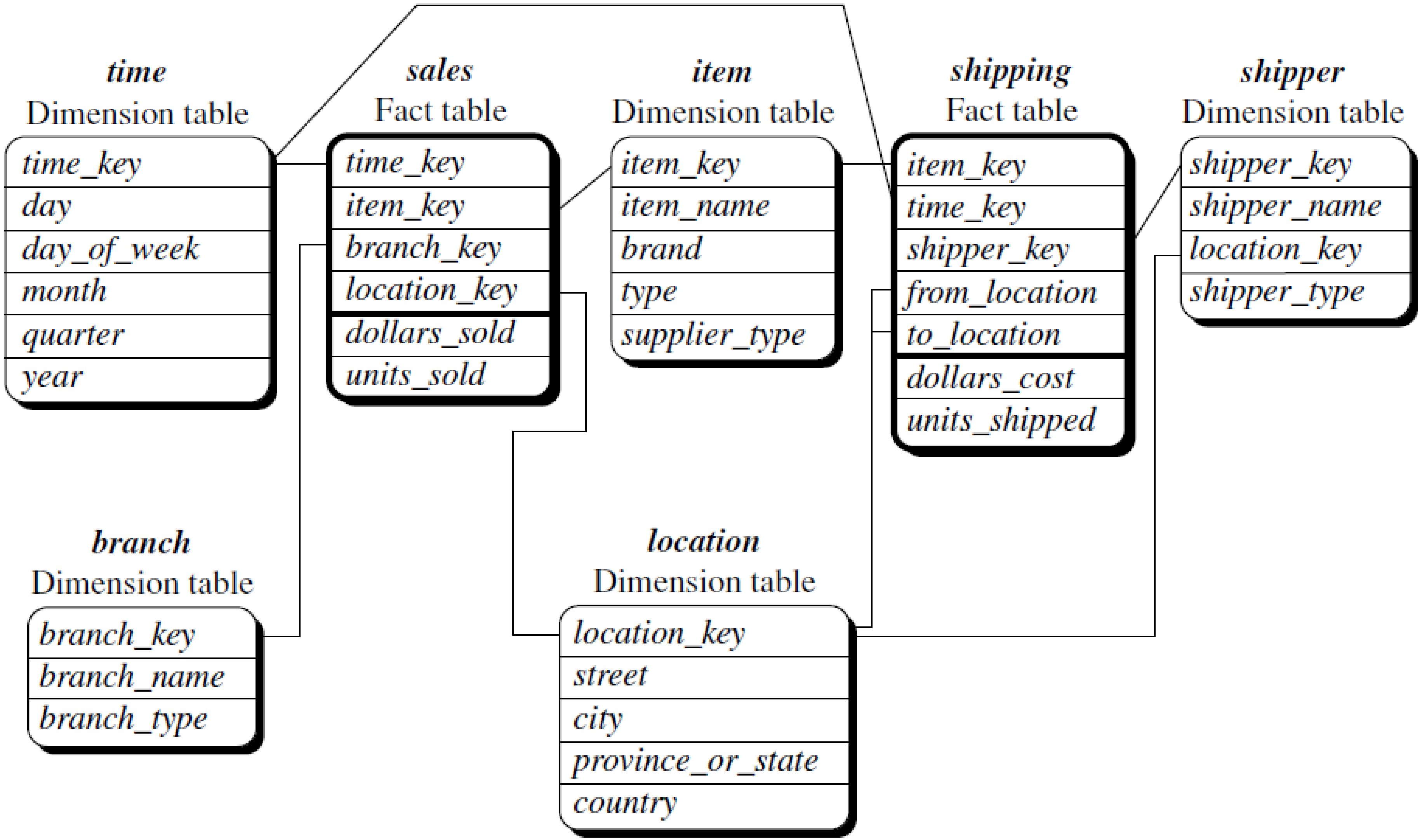
EXAMPLE OF SNOWFLAKE SCHEMA



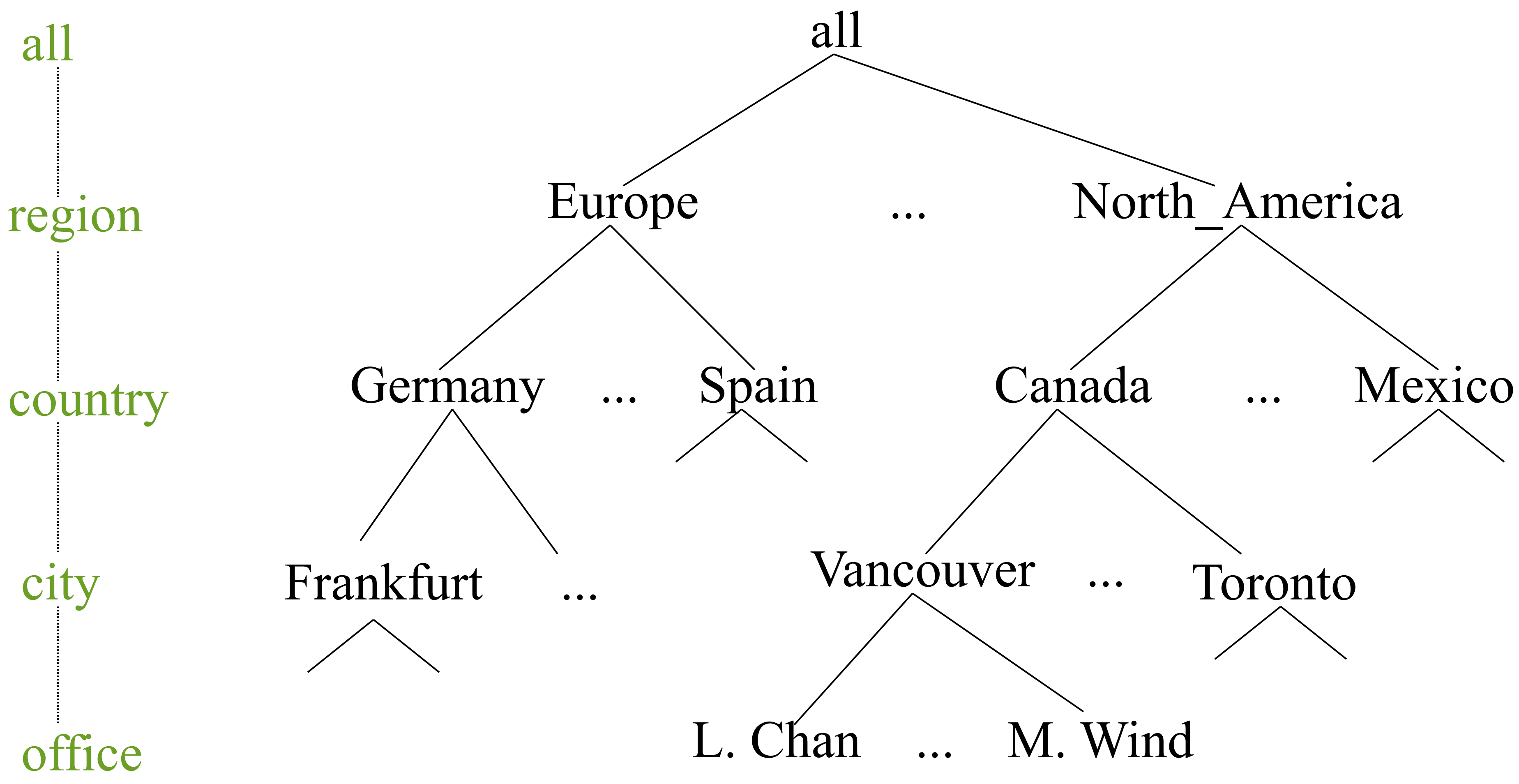
EXAMPLE OF FACT CONSTELLATION



EXAMPLE OF FACT CONSTELLATION



A CONCEPT HIERARCHY: DIMENSION (LOCATION)

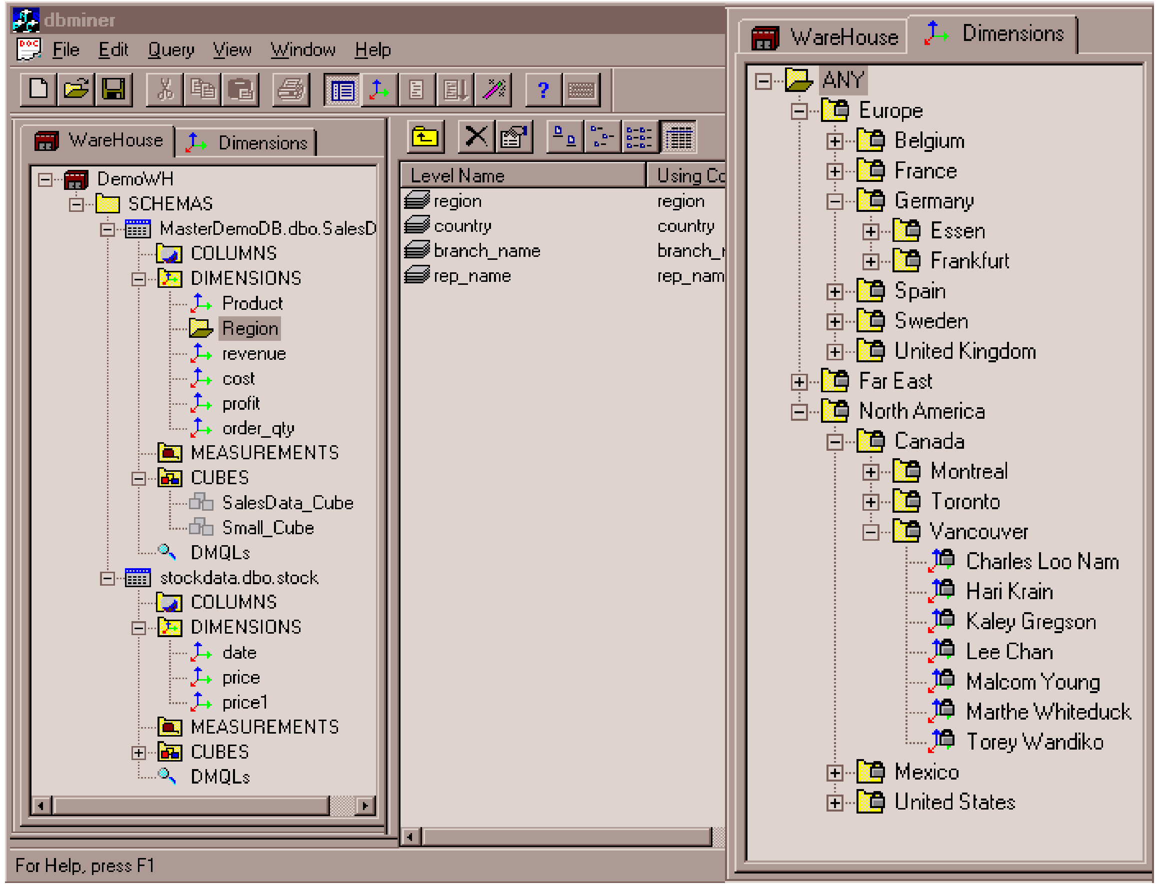


DATA CUBE MEASURES: THREE CATEGORIES

- **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(), 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
  - E.g., avg(), min_N(), standard_deviation()
- **Holistic**: if there is no constant bound on the storage size needed to describe a subaggregate.
  - E.g., median(), mode(), rank()

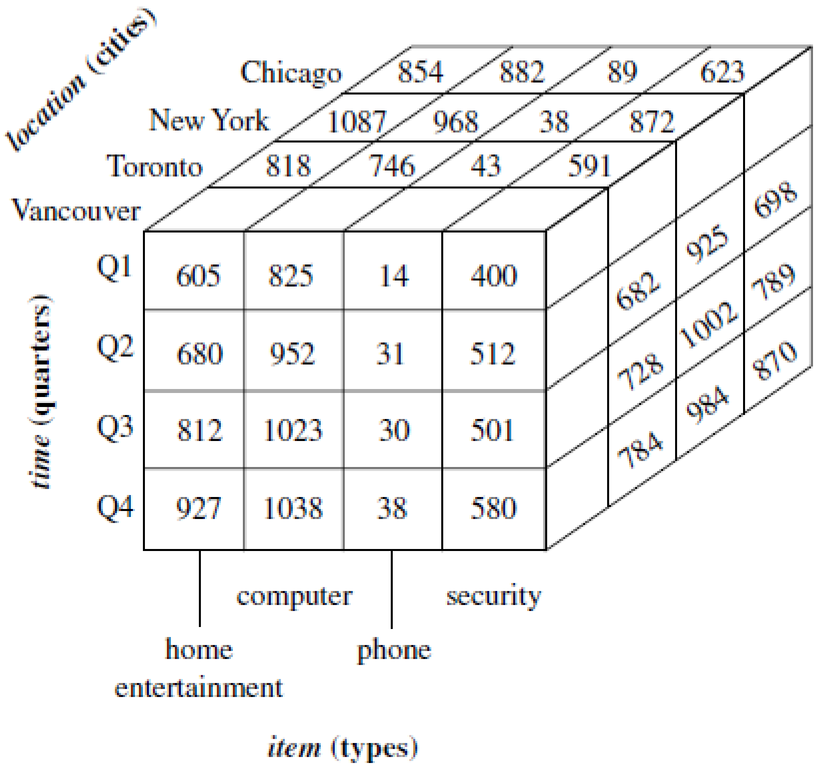
VIEW OF WAREHOUSES AND HIERARCHIES

- Specification of hierarchies
- Schema hierarchy
  - day < {month < quarter; week} < year
- Set_grouping hierarchy
  - {1..10} < inexpensive



MULTIDIMENSIONAL DATA

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



Dimensions: ***Product, Location, Time***

Hierarchical summarization paths

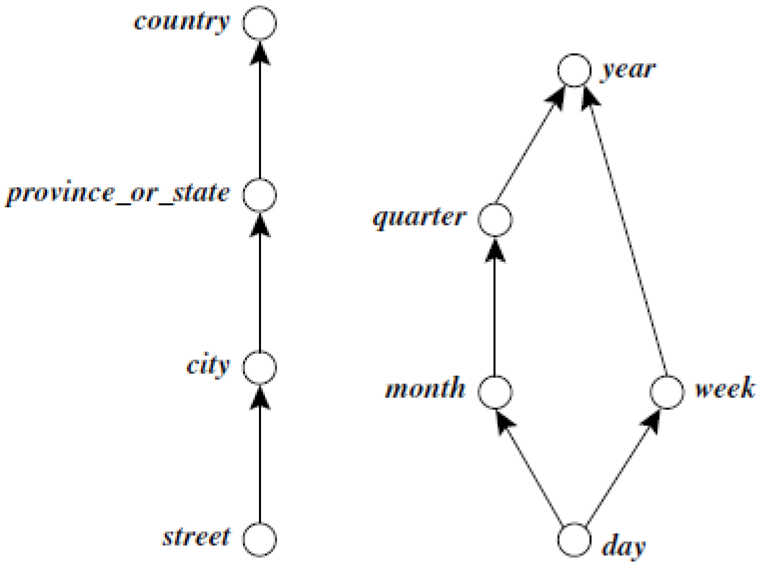
Industry

↑

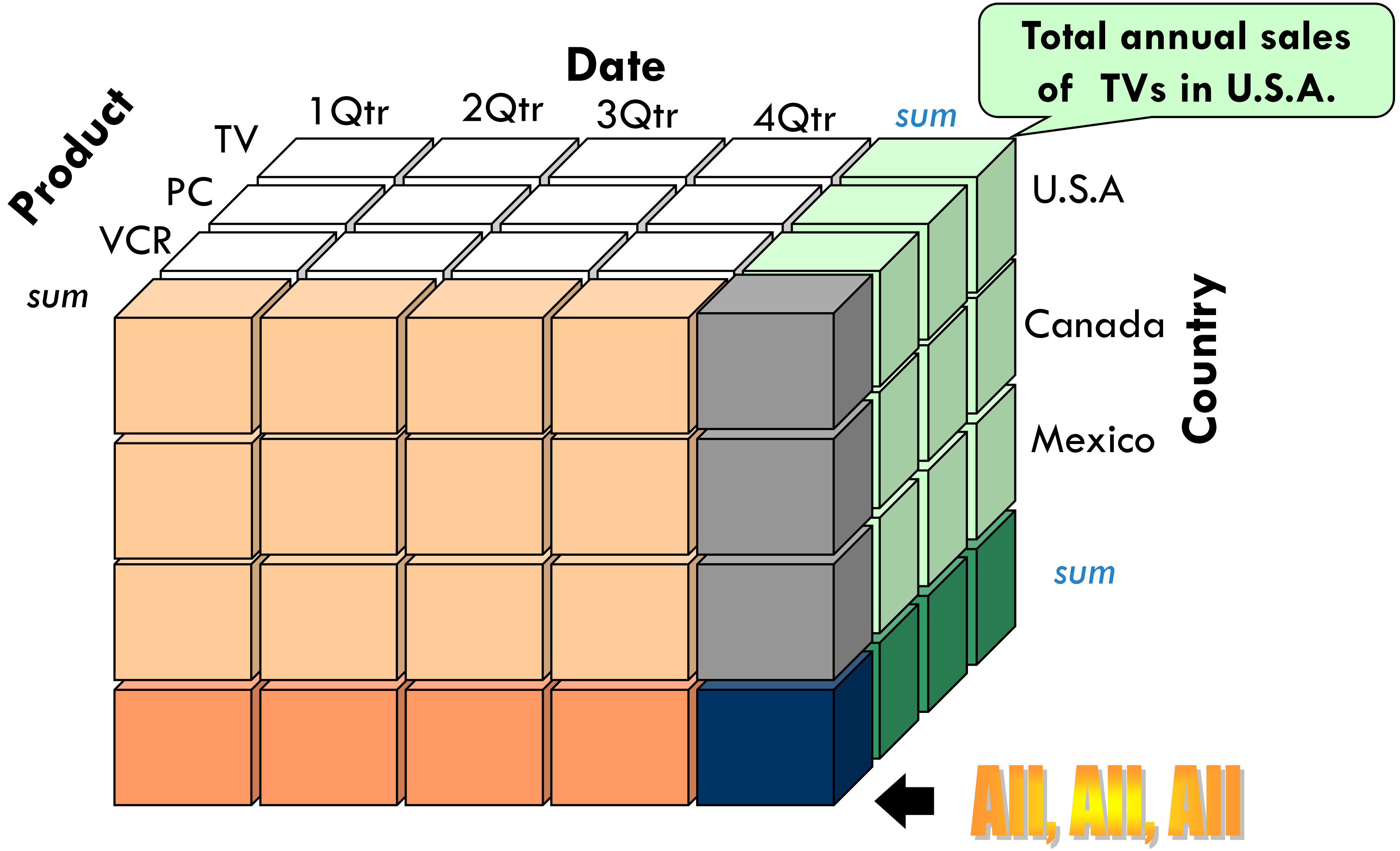
Category

↑

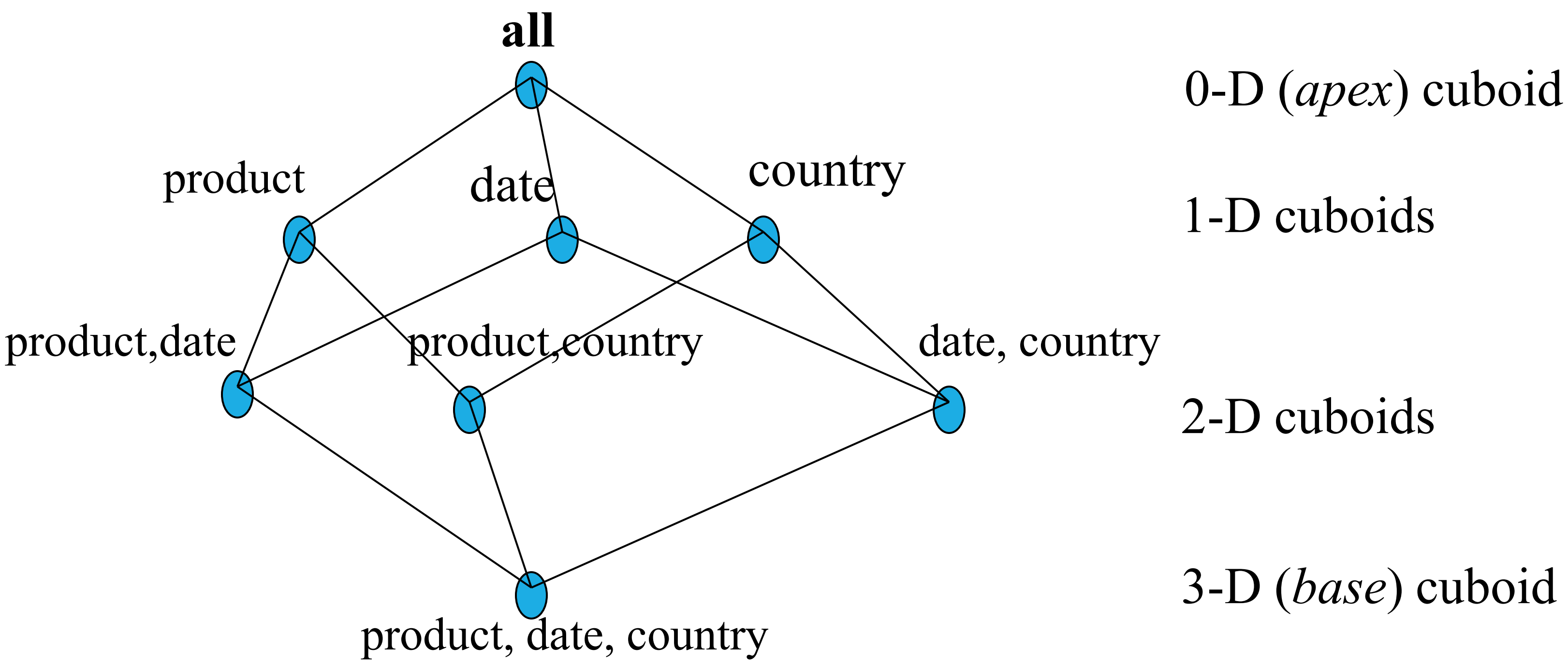
Product



A SAMPLE DATA CUBE



CUBOIDS CORRESPONDING TO THE CUBE

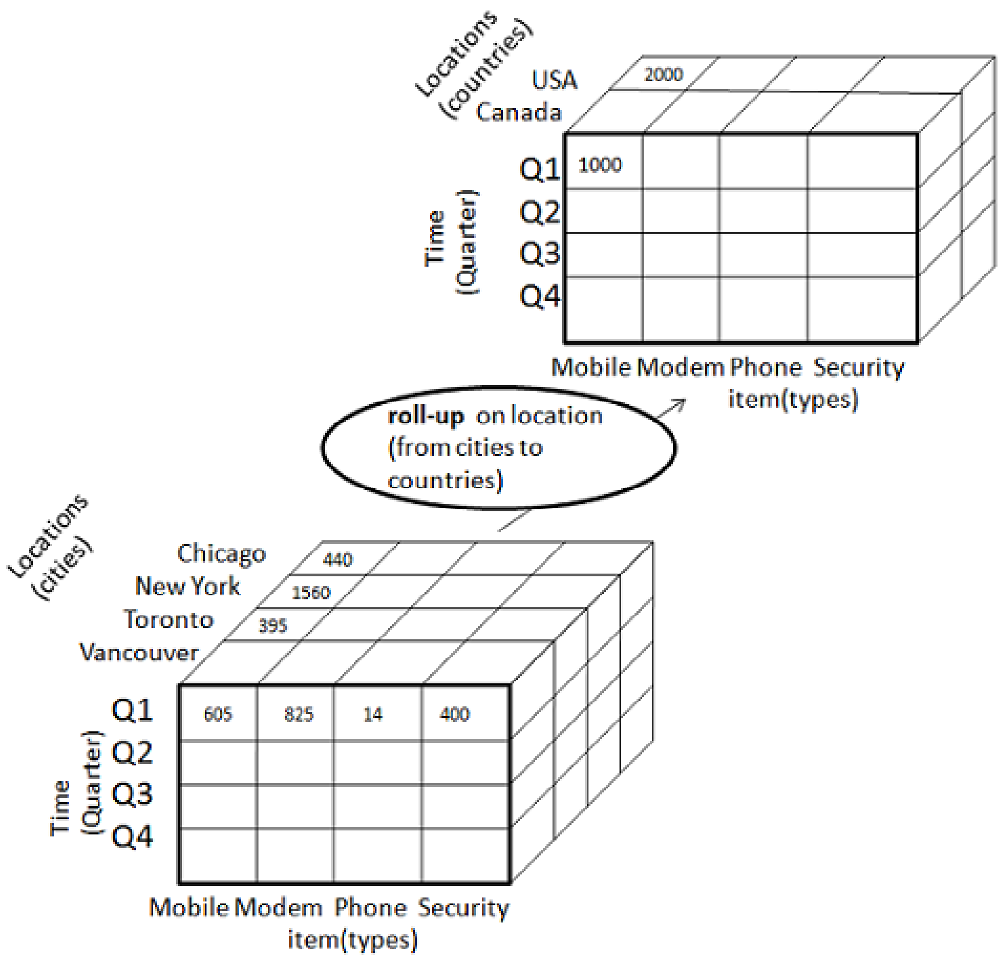


TYPICAL OLAP OPERATIONS

- **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)

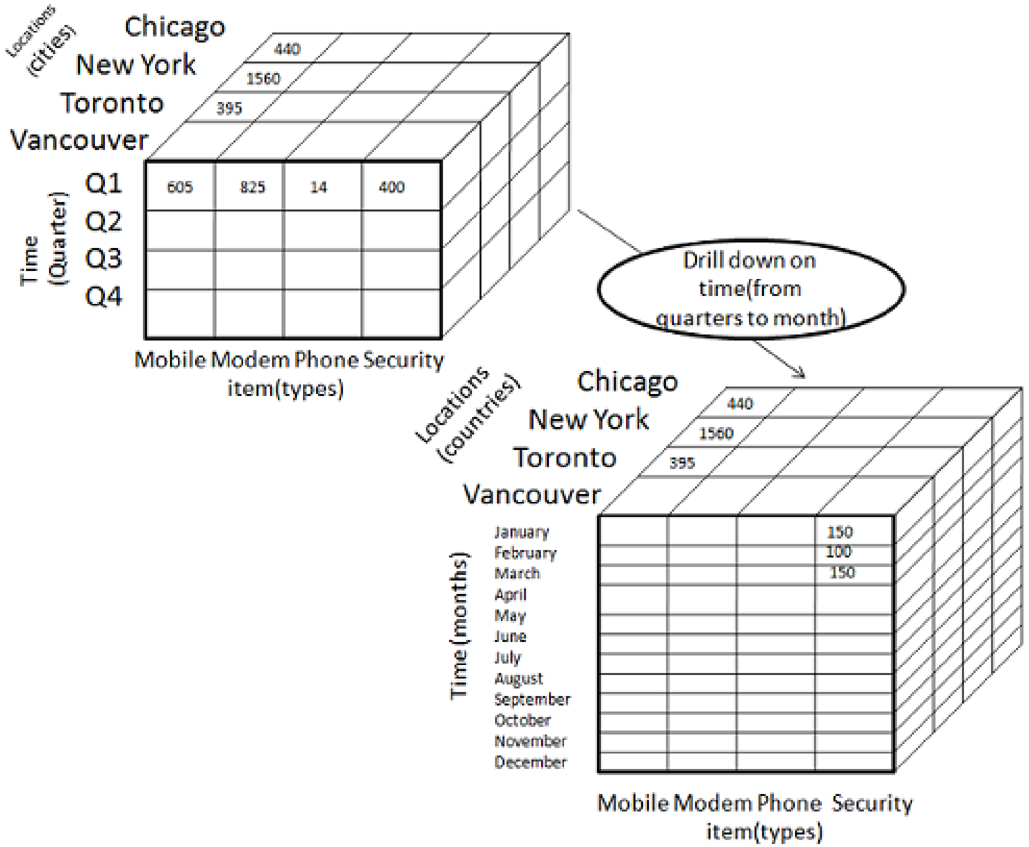
ROLL-UP

- Roll-up performs aggregation on a data cube in any of the following ways:
  - By climbing up a concept hierarchy for a dimension.
  - By dimension reduction.



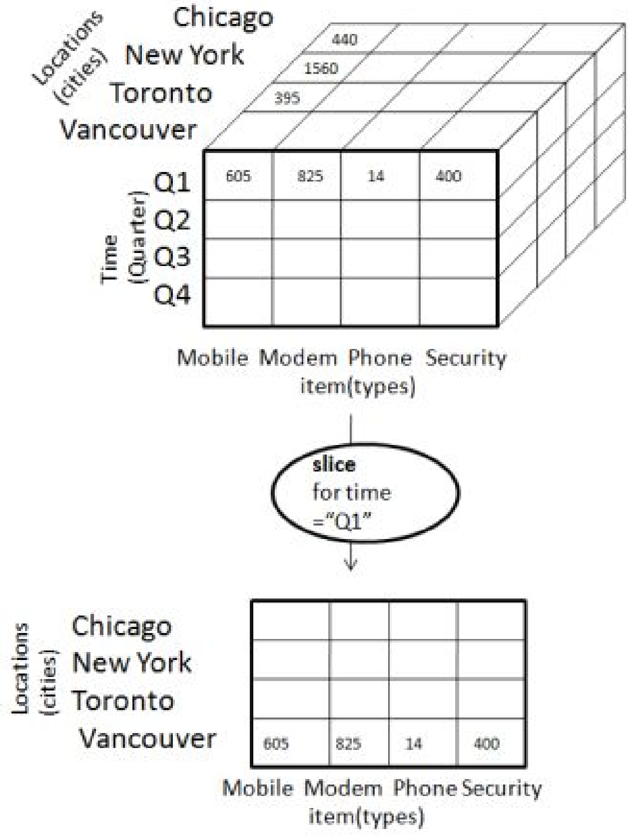
DRILL-DOWN

- Drill-down is the reverse operation of roll-up. It is performed by either of the following ways:
  - By stepping down a concept hierarchy for a dimension.
  - By introducing a new dimension.



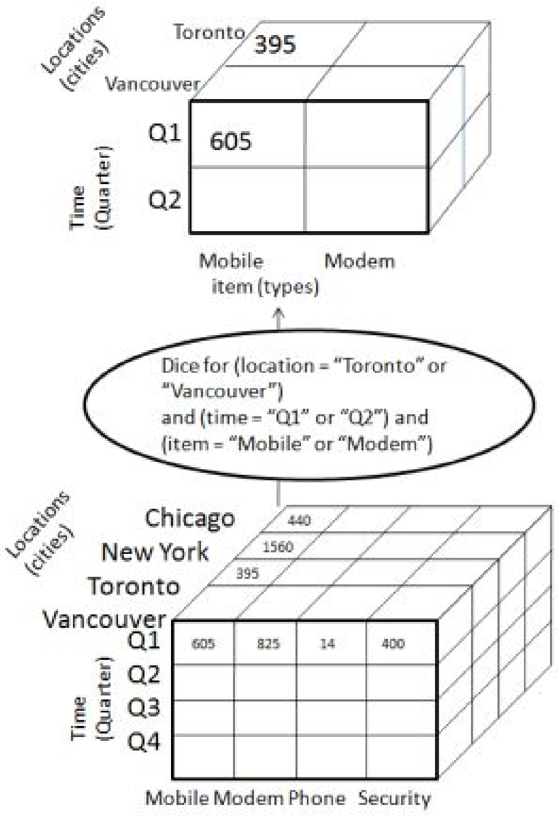
SLICE

- The slice operation selects one particular dimension from a given cube and provides a new sub-cube. Consider the following diagram that shows how slice works.



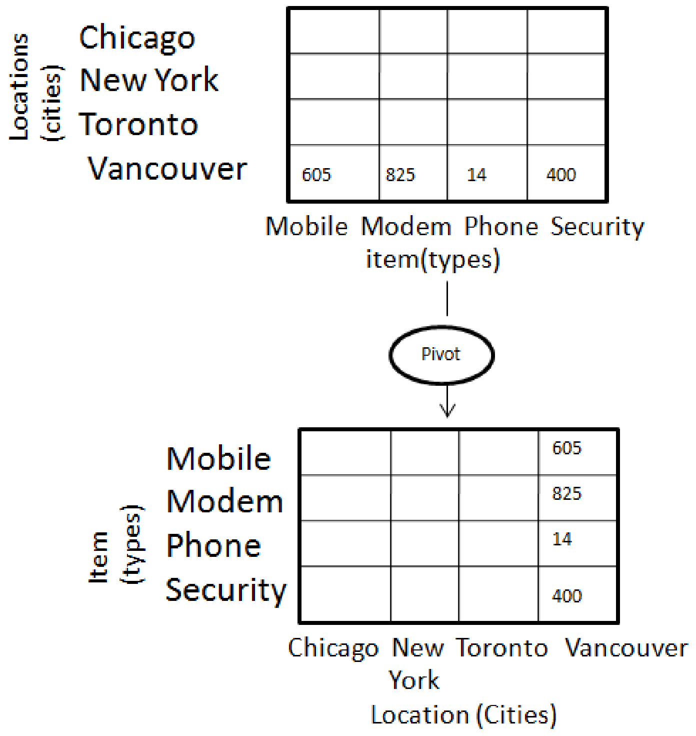
DICE

- Dice selects two or more dimensions from a given cube and provides a new sub-cube. Consider the following diagram that shows the dice operation.

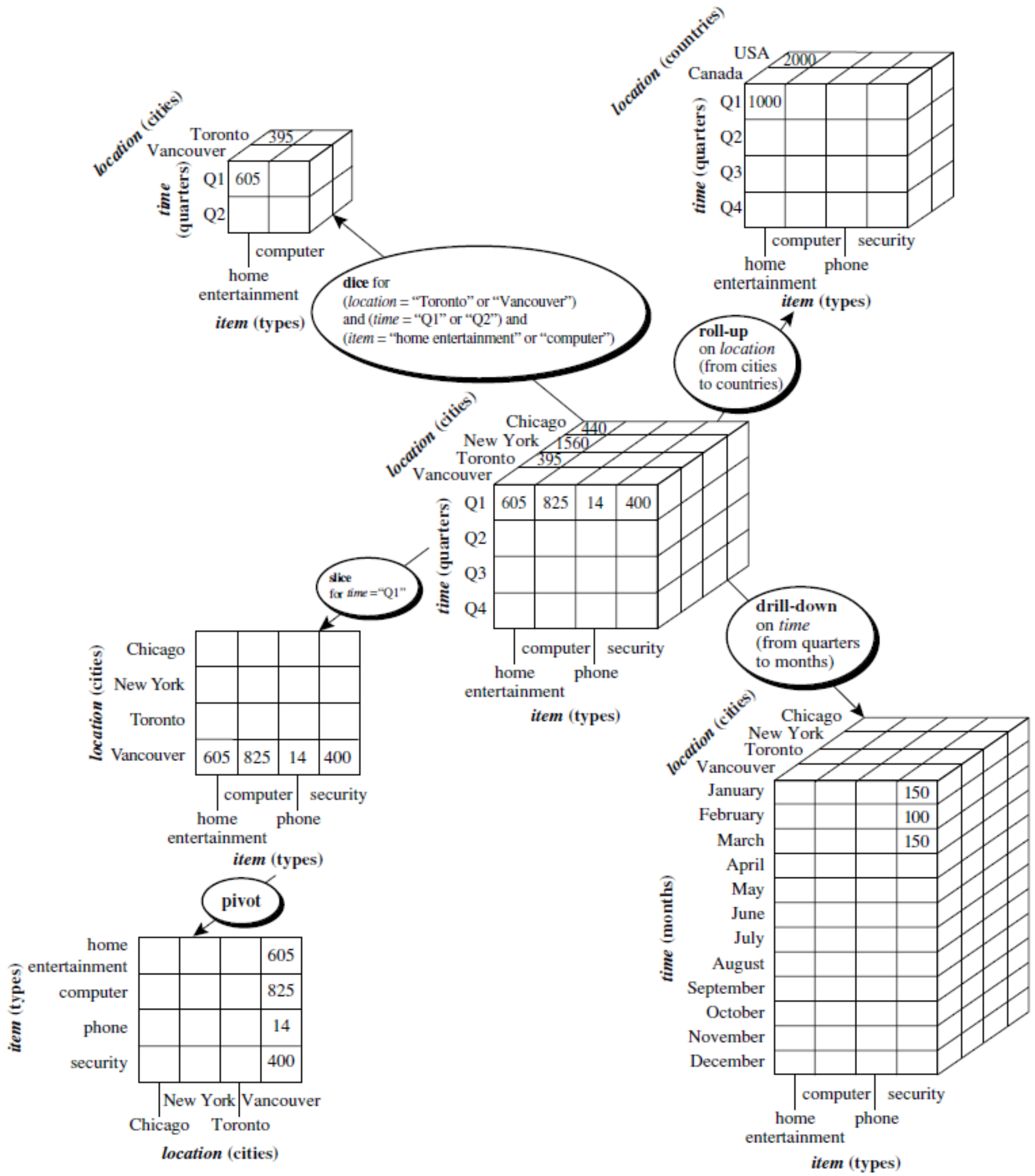


PIVOT (ROTATE)

- The pivot operation is also known as rotation. It rotates the data axes in view in order to provide an alternative presentation of data. Consider the following diagram that shows the pivot operation.



TYPICAL OLAP OPERATIONS



End of Chapter 4

THANK YOU