Data Mining: Data Warehousing and on-line Analytical processing

Introduction to Data Mining

by

Eng. Asma’a Hassan

Learning Objectives

After completing this lecture, students should be able to:

- Define data warehouse
- Differentiate between operational & transactional Database Systems.
- Explain A multi-dimensional data model
- Demonstrate Data warehouse architecture
- Describe data warehouse implementation issues.

Data Warehousing and OLAP Technology

- **What is data warehouse?**
- Difference between operational & transactional Database Systems.
- A multi-dimensional data model
- Data warehouse architecture
- data warehouse implementation.

What is 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- Non-volatile

- 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 Warehousing and OLAP Technology

- What is data warehouse?
- **Difference between operational & transactional Database Systems**.
- A multi-dimensional data model
- Data warehouse architecture
- data warehouse implementation.

Data Warehouse (OLAP) vs. Operational DBMS (OLTP)

- 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

Parameters	OLTP	OLAP
Purpose	It is a system for processing large volumes of real-time transactional data.	It is a system for the multidimensional analysis of consolidated business data.
Usage	It is used for adding, deleting, or updating databases to keep the data up-to-date.	It is used to make business decisions through queries and complex analyses of large amounts of data.
Focus	The system is more focused on transactional data maintenance and less on data analysis.	The system is focused on data analysis and not on maintaining day-to-day transactions.
Data Source	OLTP sources data from traditional database management systems.	OLAP has multiple data sources, which include real-time and historical databases, including OLTP.
Data Type	The data consists of a large number of short transactions.	The system processes large volumes of data from multiple sources.
Processing Time	Very low processing time at the scale of a few milliseconds.	Depending on the query, processing time is not as fast as OLTP systems and may range from a few seconds to hours.
Query	Related to adding, deleting, and updating data.	Related to data analysis.
Availability	OLTP systems are available round-the-clock and updated frequently to maintain data integrity.	OLAP systems don't need to be updated so frequently since their functions are analytic in nature.
Normalization	Data tables are normalized.	Data tables are not normalized.
Backup	Requires constant backup and recovery.	Can be backed up less frequently.
User volume	Supports large user volume simultaneously.	Accommodates multiple users but doesn't have a large user volume like OLTP.
Operations	Allows both read and write operations.	Usually supports read-only operations.
Process	Processes day-to-day data quickly.	Processes analytical queries consistently and at a fast pace.

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 Warehousing and OLAP Technology

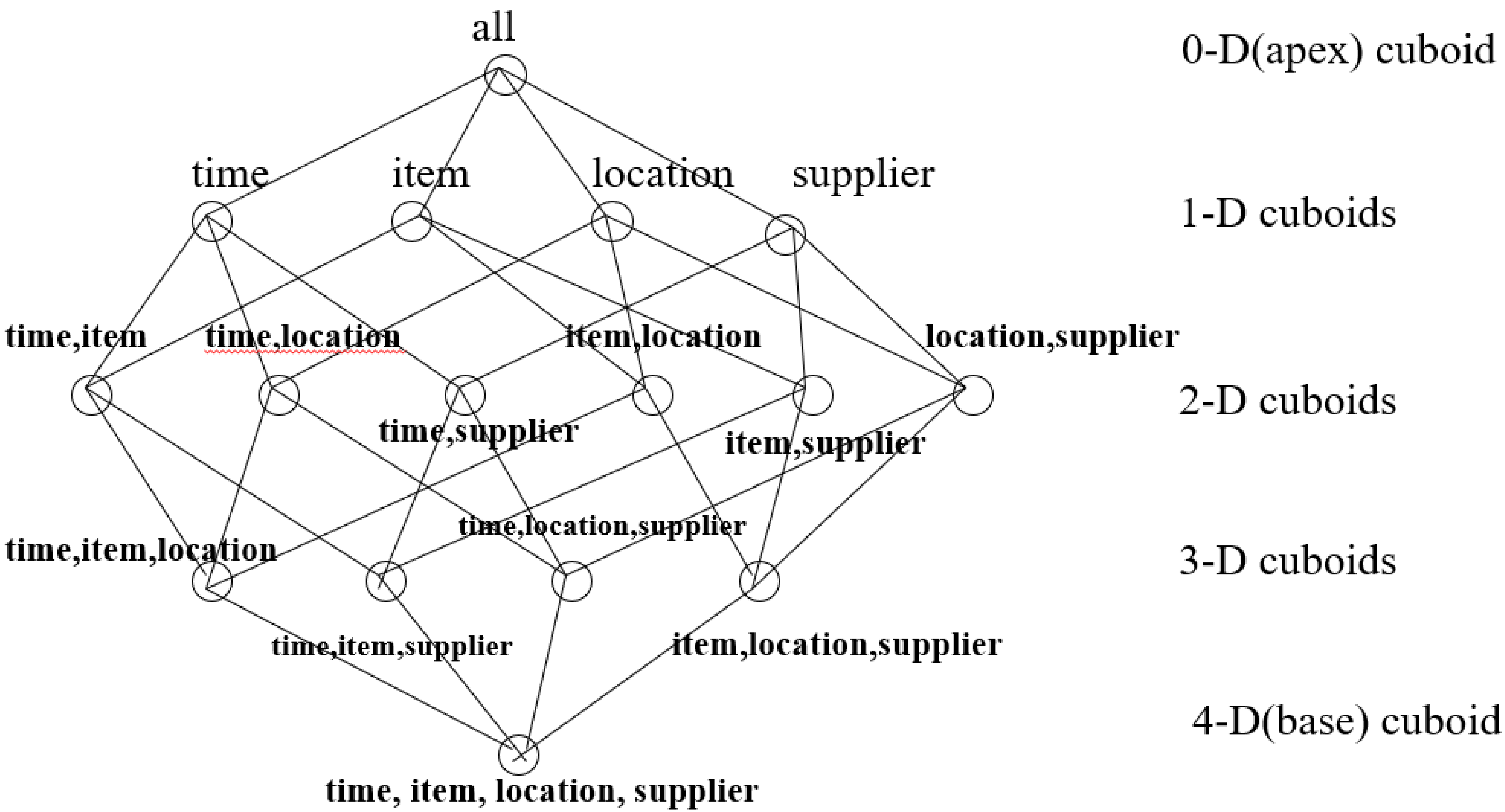
- What is data warehouse?
- Difference between operational & transactional Database Systems.
- **A multi-dimensional data model**
- Data warehouse architecture
- data warehouse implementation.

Multi Dimensional Data Model

- Data Cube: (base cube, apex cube, concept of hierarchies)
- Schemas: (Star, Snowflakes, Fact Constellations)
- OLAP Operations: (Roll up, Drill down, Slice & Dice, Pivot)

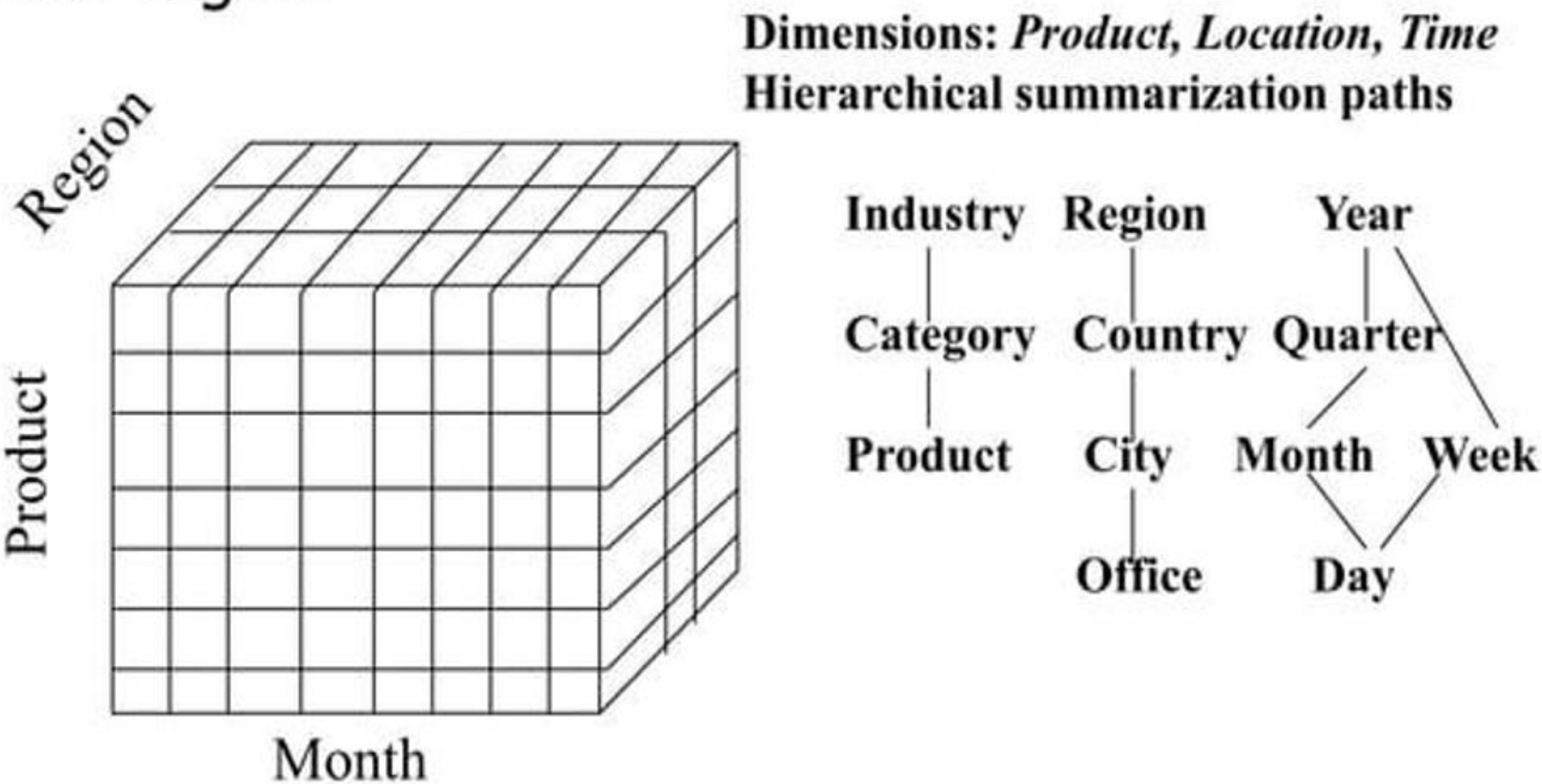
From Tables and Spread sheets to Data Cubes

- 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.

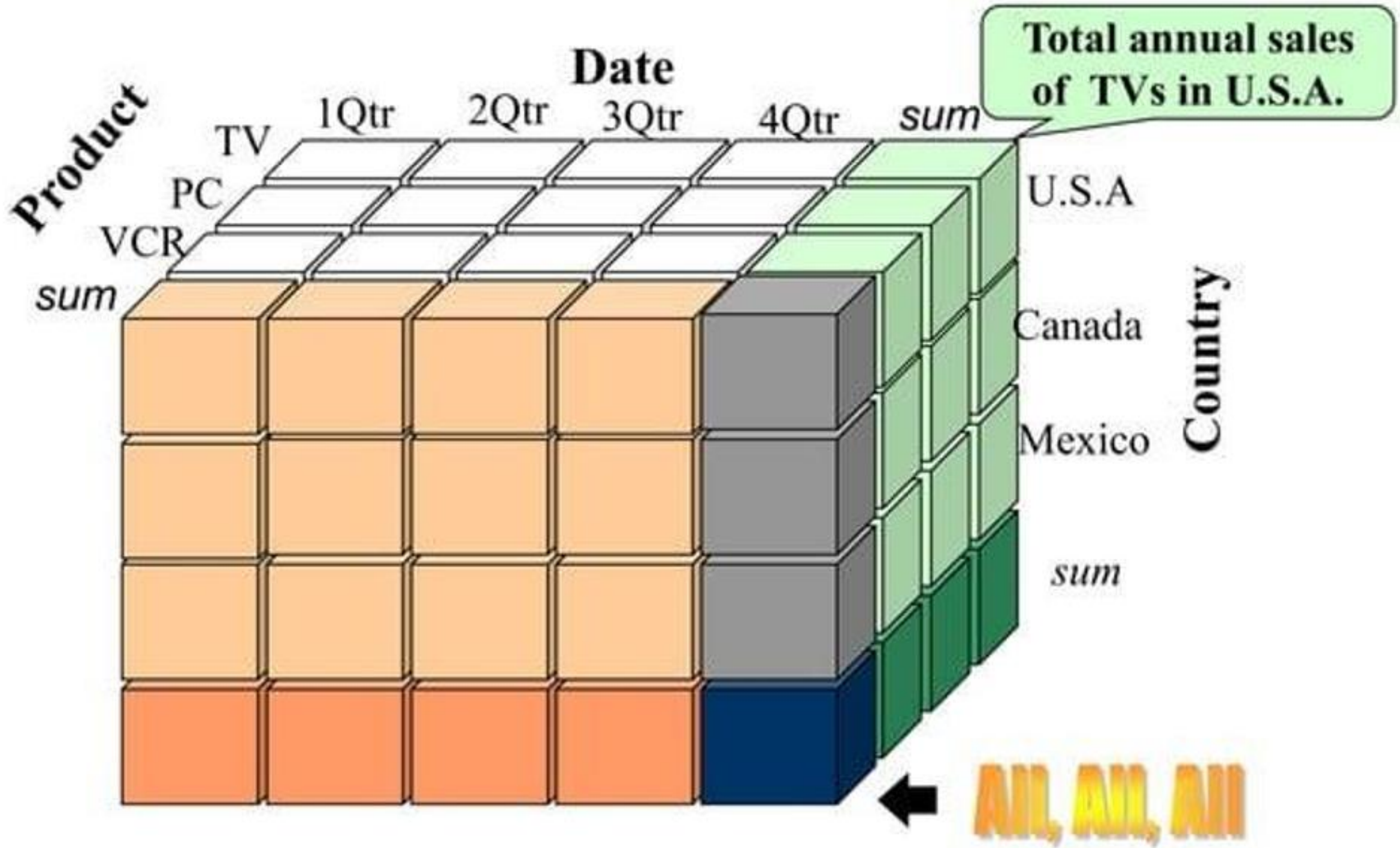

all
0-D(apex) cuboid
time
item
location
supplier
1-D cuboids
time,item
time,location
item,location
location,supplier
2-D cuboids
time,supplier
item,supplier
time,location,supplier
time,item,location
3-D cuboids
time,item,supplier
item,location,supplier
4-D(base) cuboid
time, item, location, supplier

Multidimensional Data

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

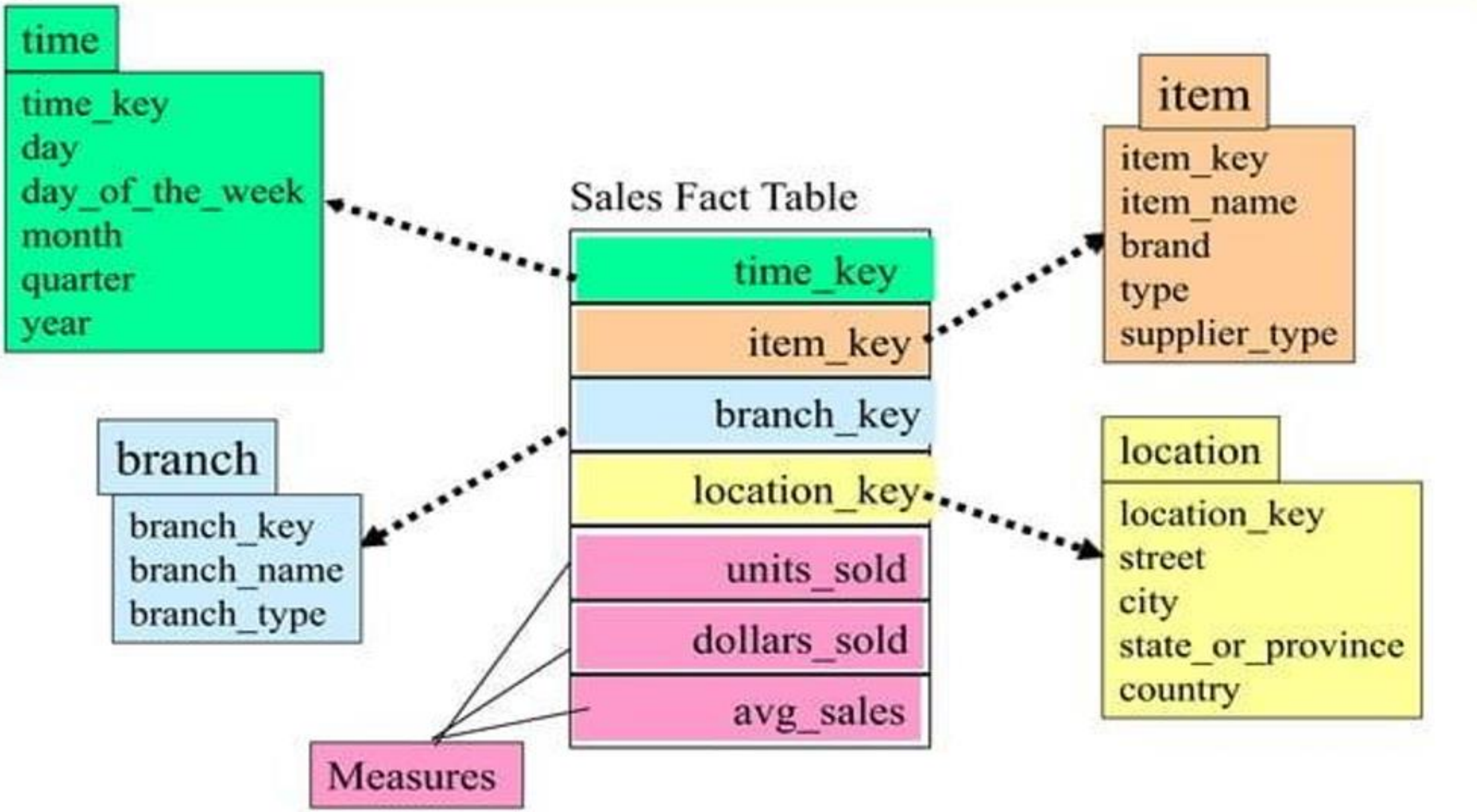


A Sample Data Cube



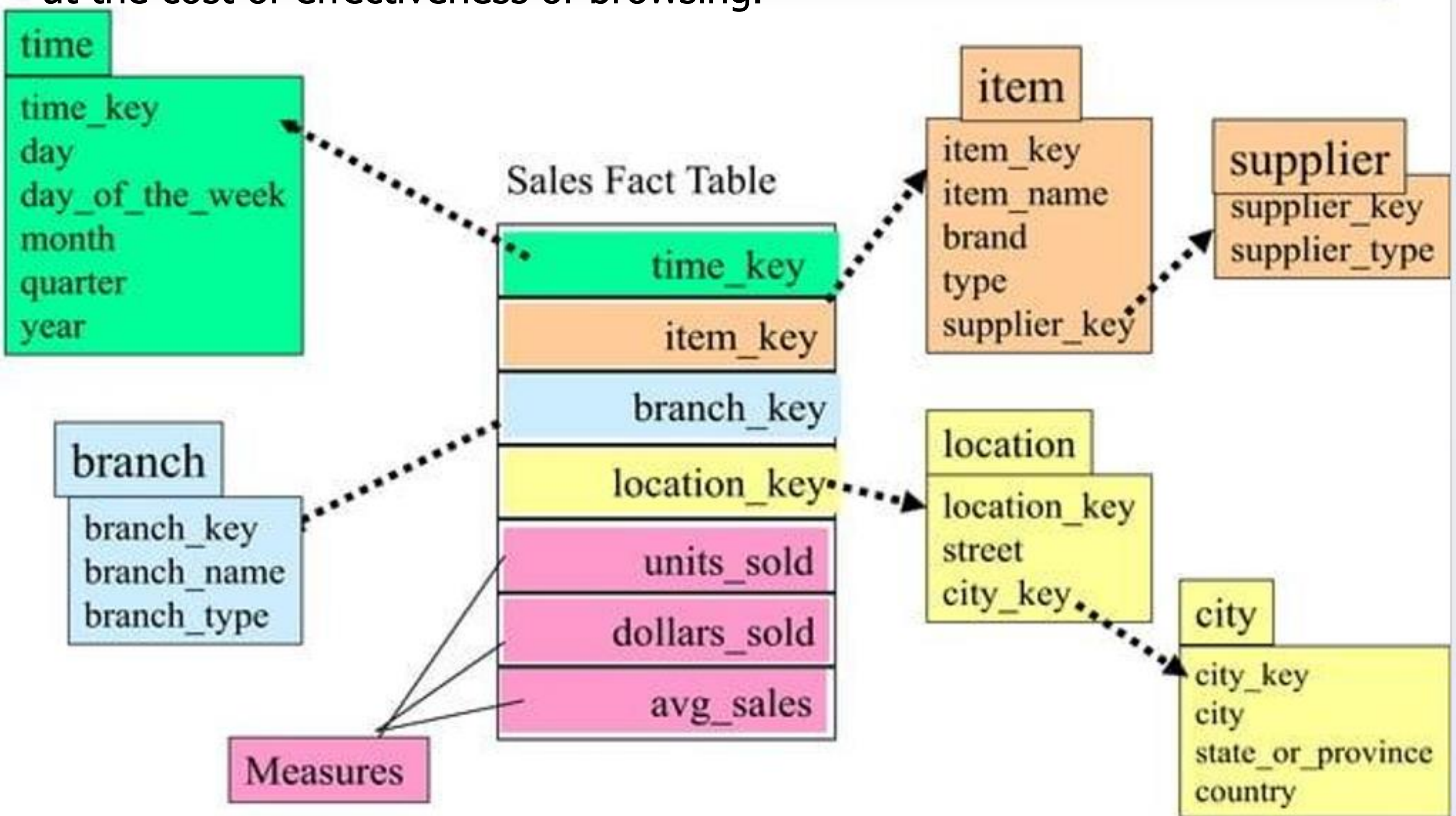
Star Schema

A fact Table in the middle connected to a set of dimension tables



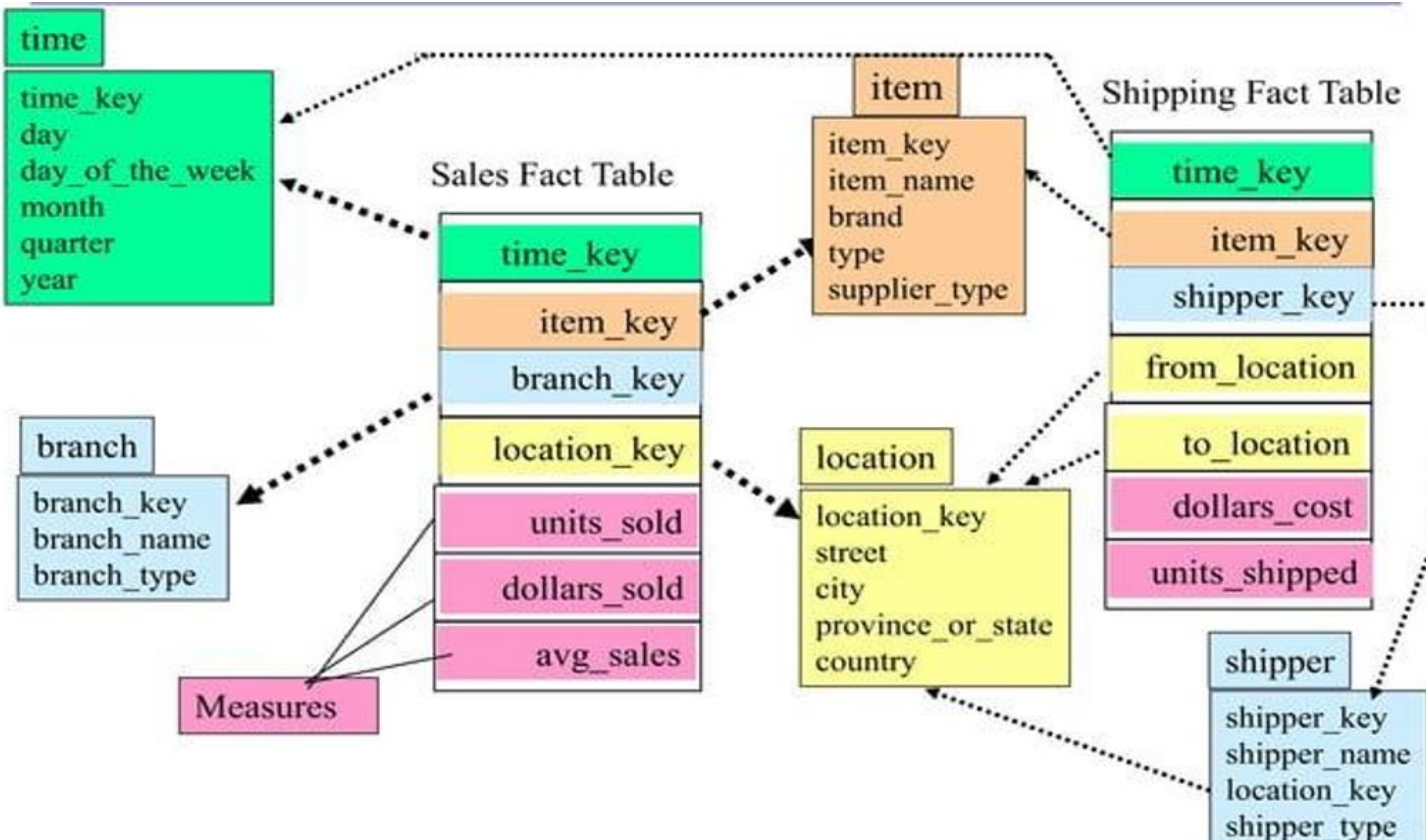
Snowflake Schema

Some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake. Reduces redundancy, however at the cost of effectiveness of browsing.



Fact Constellation

Multiple fact tables share dimension tables, viewed as a collection of stars. (Galaxy schema).



Cube Defintion syntax (BNF) in DMQL

- Cube Definition (Fact Table)

  define cube <cube_name> [<dimension_list>]: <measure_list>

- Dimension Definition (Dimension Table)

  define dimension <dimension_name> as (<attribute_or_subdimension_list>)

- Special Case (Shared Dimension Tables)
  - First time as "cube definition"
  - define dimension <dimension_name> as <dimension_name_first_time> in cube <cube_name_first_time>

Defining Star Schema in DMQL

- Cube Definition (Fact Table)

  define cube <cube_name> [<dimension_list>]: <measure_list>

- Dimension Definition (Dimension Table)

  define dimension <dimension_name> as (<attribute_or_subdimension_list>)

- Special Case (Shared Dimension Tables)
  - First time as “cube definition”
  - define dimension <dimension_name> as <dimension_name_first_time> in cube <cube_name_first_time>

Defining Snowflake Schema in DMQL

- Cube Definition (Fact Table)

  define cube <cube_name> [<dimension_list>]: <measure_list>

- Dimension Definition (Dimension Table)

  define dimension <dimension_name> as (<attribute_or_subdimension_list>)

- Special Case (Shared Dimension Tables)
  - First time as "cube definition"
  - define dimension <dimension_name> as <dimension_name_first_time> in cube <cube_name_first_time>

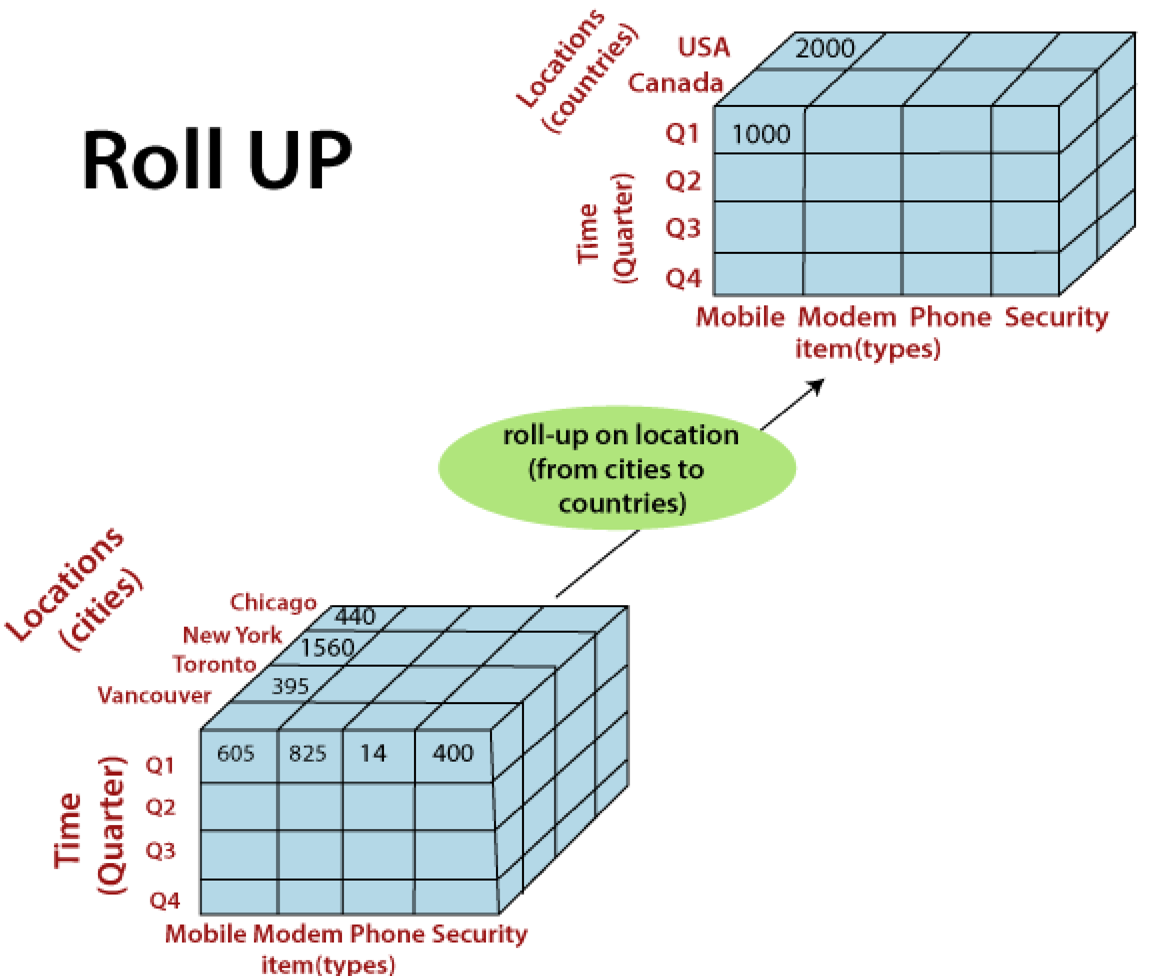
Defining Fact Constellation in DMQL

```
define cube sales [time, item, branch, location]:
        dollars_sold = sum(sales_in_dollars), avg_sales = avg(sales_in_dollars),
          units_sold = count(*)
define dimension time as (time_key, day, day_of_week, month, quarter, year)
define dimension item as (item_key, item_name, brand, type, supplier_type)
define dimension branch as (branch_key, branch_name, branch_type)
define dimension location as (location_key, street, city, province_or_state, country)
define cube shipping [time, item, shipper, from_location, to_location]:
        dollar_cost = sum(cost_in_dollars), unit_shipped = count(*)
define dimension time as time in cube sales
define dimension item as item in cube sales
define dimension shipper as (shipper_key, shipper_name, location as location in
  cube sales, shipper_type)
define dimension from_location as location in cube sales
define dimension to_location as location in cube sales
```

Typical OLAP(software tool) Operations

- Roll up (drill-up):
- Perform aggregation on a data cube by
  - Climbing up a concept hierarchy for a dimension
  - Dimension reduction summarize data

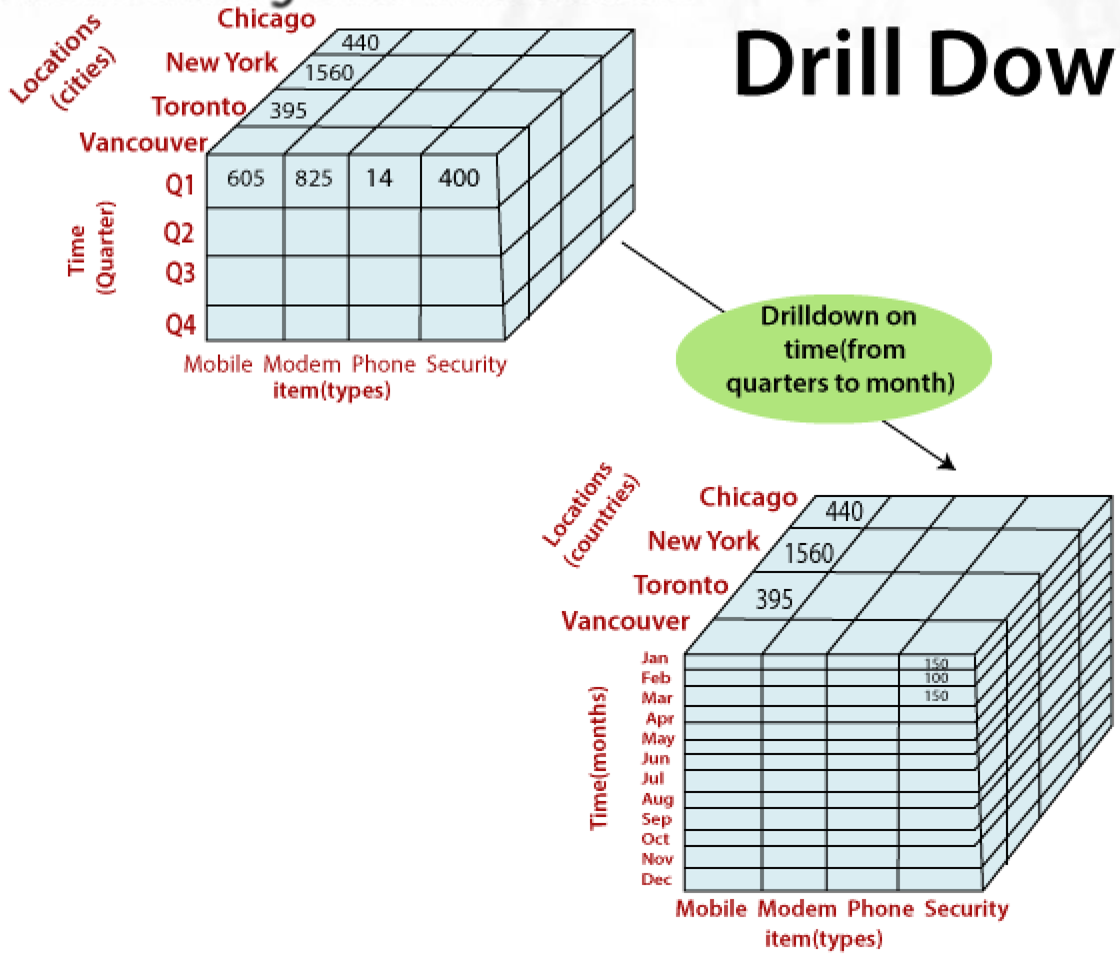
Roll UP



Typical OLAP(software tool) Operations

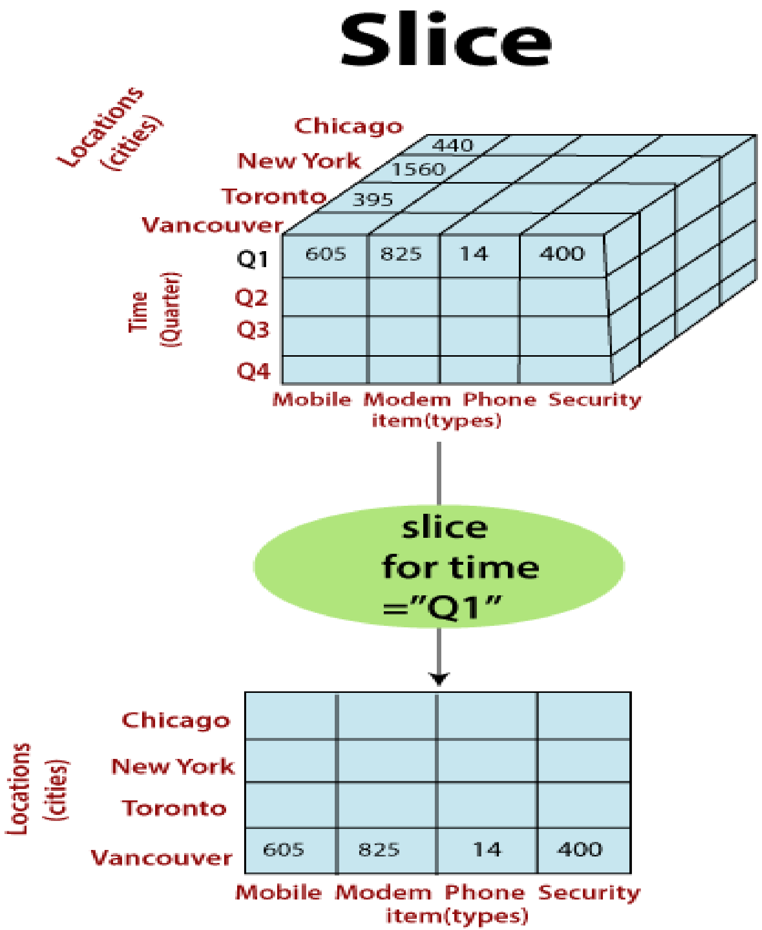
- Drill down (roll down): reverse of roll-up
  - *from higher level summary to lower level summary or detailed data, or introducing new dimensions*

Drill Down



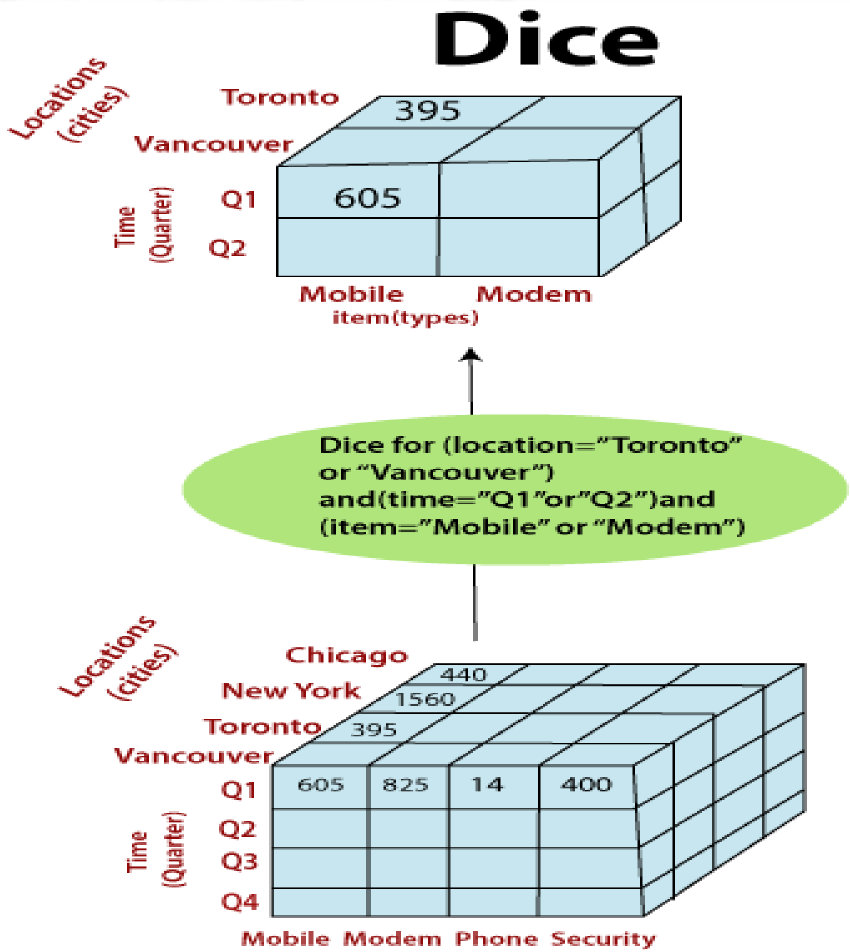
Typical OLAP(software tool) Operations

- Slice
- *The slice operation performs a selection on one dimension of the given cube, resulting in a sub-cube*



Typical OLAP(software tool) Operations

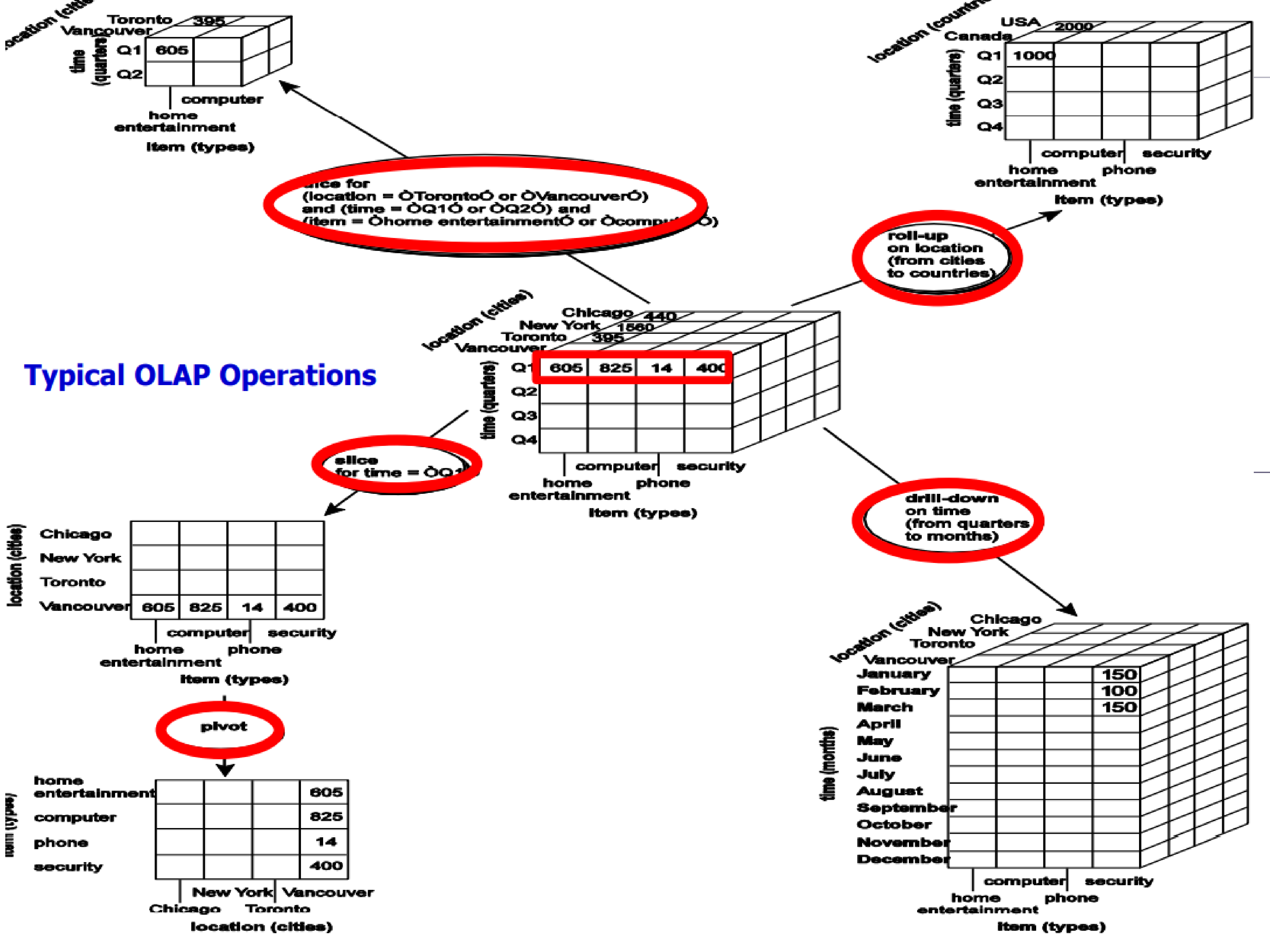
- Dice:
  - *The dice operation defines a sub-cube by performing a selection on two or more dimensions*



Typical OLAP(software tool) Operations

- Pivot (rotate):
  - *Visualization operation that rotate the data axes in view in order to provide an alternative presentation of the data.*

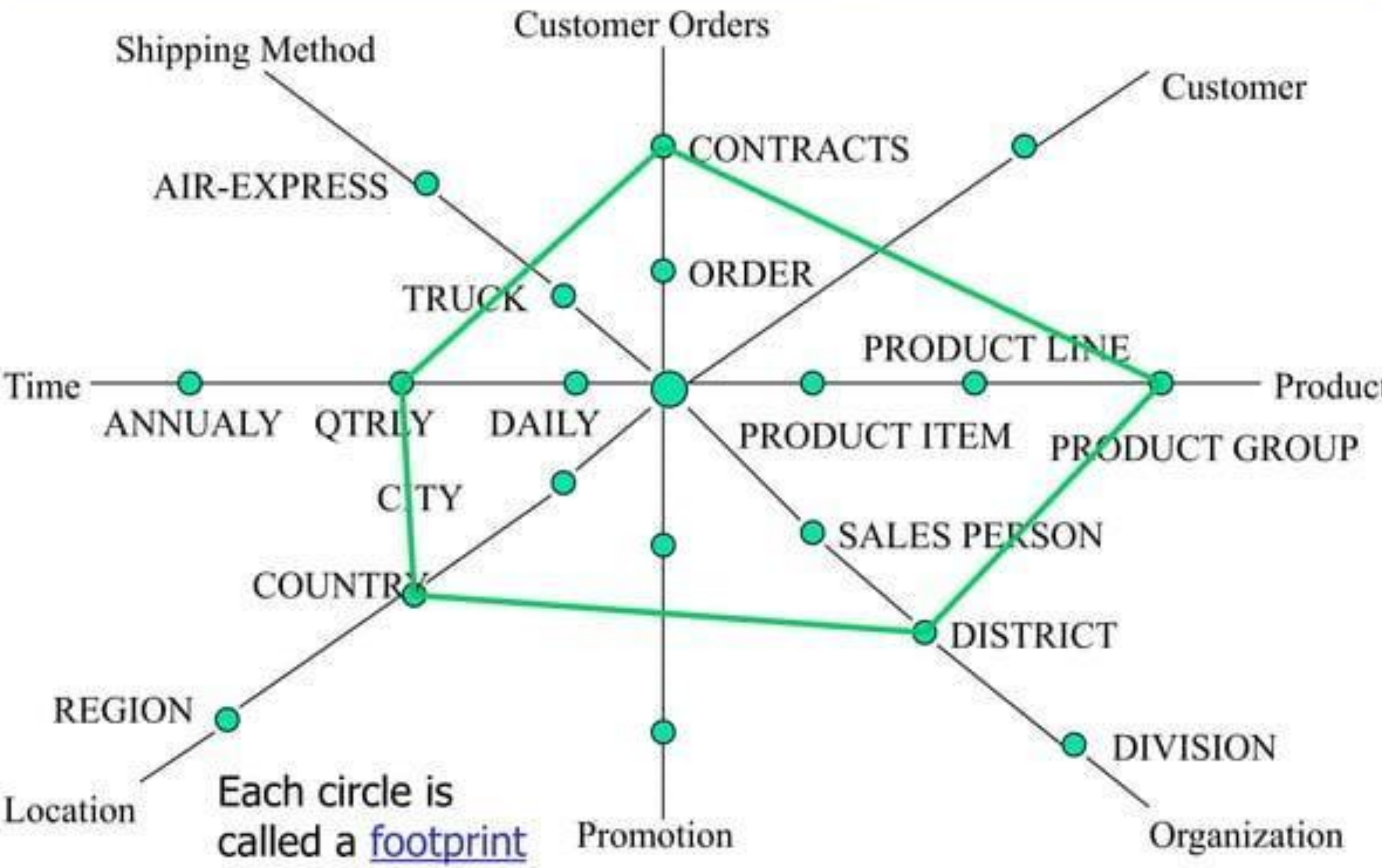
Typical OLAP Operations



A Star-Net Query Model

- Querying multidimensional DBs
- Consists of radial lines originated from central point
- Each line represent a concept hierarch for dimension
- Each abstraction level in the hierarch-footprint

A Star-Net Query Model



Each circle is called a footprint