Data Warehouse and On-Line Analytical Processing (OLAP)

Data Warehouse and OLAP

- **Data warehouses** generalize and consolidate data in multidimensional space.
  - The construction of data warehouses involves data cleaning, data integration, and data transformation and can be viewed as an important preprocessing step for data mining.

- Data warehouses provide **on-line analytical processing (OLAP)** tools for the interactive analysis of multidimensional data of varied granularities, which facilitates effective data generalization and data mining.
  - Many other data mining functions, such as association, classification, prediction, and clustering, can be integrated with OLAP operations to enhance interactive mining of knowledge at multiple levels of abstraction.

What is a Data Warehouse?

- **Data warehousing** provides architectures and tools for business executives to systematically organize, understand, and use their data to make strategic decisions.

- **Data warehouses** have been defined in many ways:
  - A **decision support database** that is maintained **separately** from an organization's operational databases.
  - Data warehouses support **information processing** by providing a solid platform of consolidated, historical data for analysis.

- **Data warehousing**:
  - The process of constructing and using data warehouses.

Major Features of a Data Warehouse
Subject-Oriented

Four major features of a 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.

Subject-Oriented:

- A data warehouse is organized around major subjects, such as customer, product, sales.
- A data warehouse focuses on the modeling and analysis of data for decision makers, not on daily operations or transaction processing.
- A data warehouse provides a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

Major Features of a Data Warehouse
Integrated

Integrated:

- A data warehouse is constructed by integrating multiple heterogeneous data sources such as relational databases, flat files, on-line transaction records.
- Data cleaning and data integration techniques are applied to ensure consistency in naming conventions, encoding structures, attribute measures, etc.
  - When data is moved to the warehouse from operational databases, it is converted.

Major Features of a Data Warehouse
Time-Variant

Time-Variant:

- The time horizon for a 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 structure of operational data may or may not contain “time element”

Major Features of a Data Warehouse
Nonvolatile

Nonvolatile:

- A data warehouse is a physically separate store of data transformed from the operational environment.
- Operational update of data does not occur in a 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

Operational Database Systems and Data Warehouses

- The major task of on-line **operational database systems** is to perform on-line transaction and query processing.
  - These systems are called **on-line transaction processing (OLTP)** systems.
  - They cover most of the day-to-day operations of an organization, such as purchasing, inventory, banking, payroll, registration, and accounting.
- **Data warehouse systems**, on the other hand, serve users or knowledge workers in the role of data analysis and decision making.
  - Such systems can organize and present data in various formats in order to accommodate the diverse needs of the different users.
  - These systems are known as **on-line analytical processing (OLAP)** systems.

OLTP vs. OLAP

Users and system orientation:

- An OLTP system is ***customer-oriented*** and is used for transaction and query processing by clerks, clients, and information technology professionals.
- An OLAP system is ***market-oriented*** and is used for data analysis by knowledge workers, including managers, executives, and analysts.

Data contents:

- An OLTP system manages **current data** that, typically, are too detailed to be easily used for decision making.
- An OLAP system manages large amounts of **historical data**, provides facilities for summarization and aggregation.

Database design:

- An OLTP system usually adopts an ***entity-relationship (ER)*** data model and an ***application-oriented*** database design.
- An OLAP system typically adopts either a ***star*** or ***snowflake*** model and a ***subject oriented*** database design.

OLTP vs. OLAP

View:

- An OLTP system focuses mainly on the ***current data***.
- In contrast, an OLAP system often spans ***multiple versions of a database schema***, due to the evolutionary process of an organization.

Access patterns:

- The access patterns of an OLTP system consist mainly of ***short, atomic transactions***.
  – Such a system requires concurrency control and recovery mechanisms.
- However, accesses to OLAP systems are mostly ***read-only operations,*** although many could be complex queries.

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: Decision support 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 many systems which perform OLAP analysis directly on relational databases

A Three-Tier Data Warehouse Architecture

- Back-end tools and utilities are used to feed data into the **bottom tier** from operational databases or other external sources.

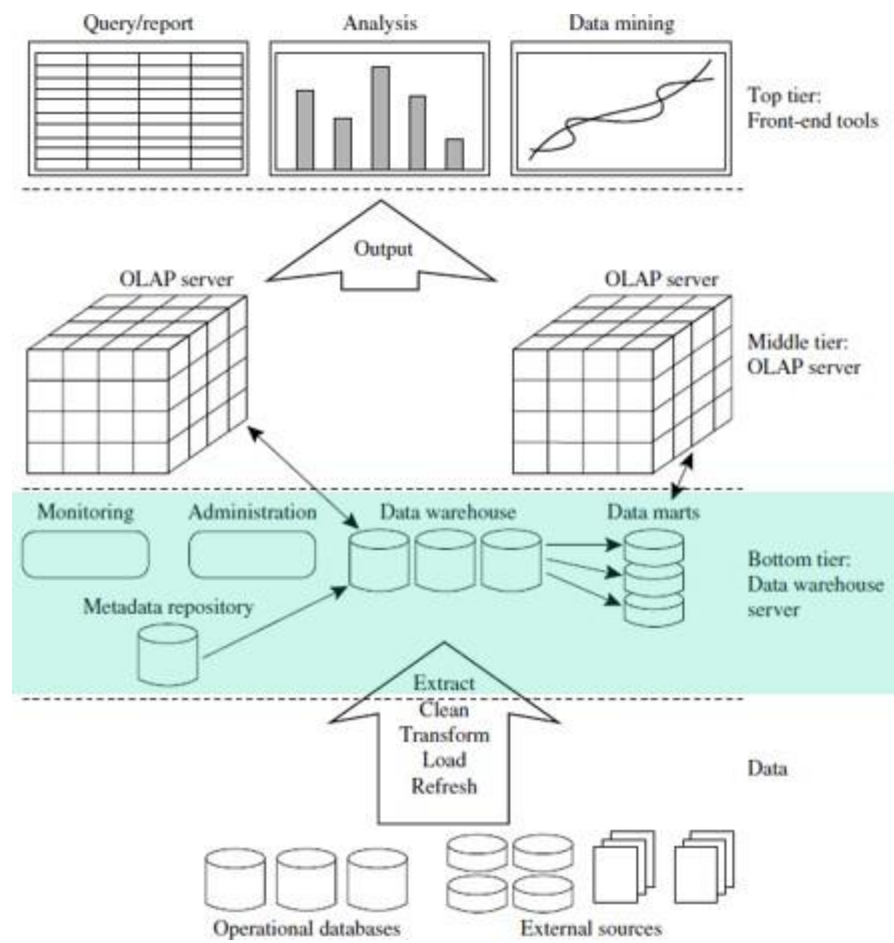
- **Data extraction:** get data from multiple, heterogeneous, and external sources.

- **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.
- A **metadata repository** contains:
  - A description of the data warehouse structure:
    - schema, view, dimensions, hierarchies, derived data definitions, 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

Three Data Warehouse Models

- From the architecture point of view, there are three data warehouse models: enterprise warehouse, data mart and virtual warehouse.

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

Virtual Warehouse

- A set of views over operational databases
- Only some of the possible summary views may be materialized

Multidimensional Data Model: Data Cube

- Data warehouses and OLAP tools are based on a **multidimensional data model**.
- This model views data in the form of a **data cube**.
- A data cube allows data to be modeled and viewed in **multiple dimensions**.
  - It is defined by *dimensions* and *facts*.
  - Dimensions are the perspectives or entities with respect to which an organization wants to keep records.
    - Each dimension may have a table associated with it, called a dimension table, which further describes the dimension. For example, a dimension table for item may contain the attributes item name, brand, and type.
  - Facts are numerical measures.
    - Examples of facts for a sales data warehouse include dollars sold units sold

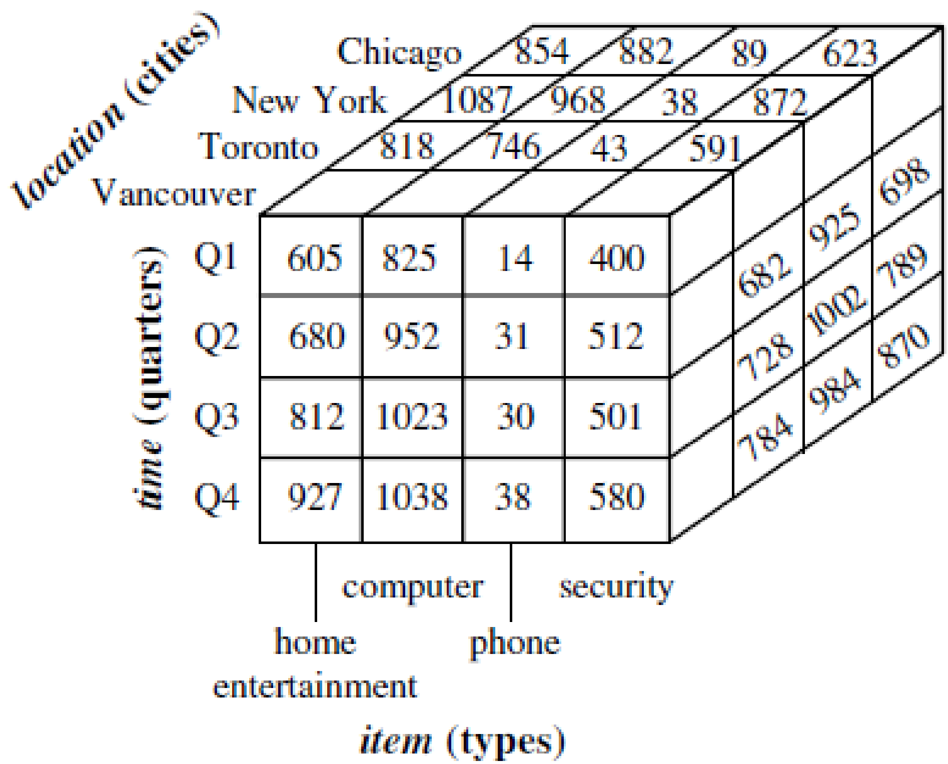
Data Cube: A 2-D Data Cube

- Although we usually think of cubes as 3-D geometric structures, in data warehousing the data cube is n-dimensional.
- A 2-D data cube:
  - **dimensions** *time* and *item*, the measure displayed (**fact**) is *dollars_sold*.
  - In this 2-D representation, the *sales* for Vancouver are shown with respect to the *time* dimension (organized in quarters) and the *item* dimension (organized according to the types of items sold).

	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

Data Cube: A 3-D Data Cube

- A 3-D data cube representation of the data according to the dimensions **time**, **item**, and **location**. The measure displayed is ***dollars_sold***.

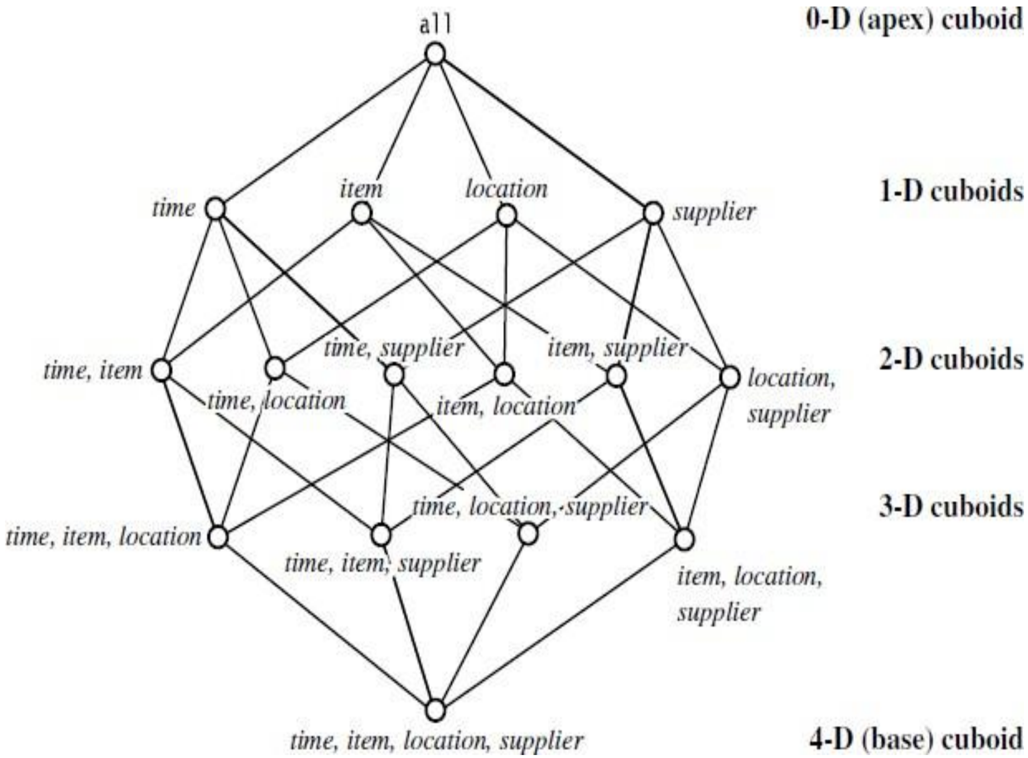


From Tables to Data Cubes

- 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 Lattice of Cuboids

- A lattice of cuboids, making up a 4-D data cube for the dimensions *time*, *item*, *location*, and *supplier*.
  - Each cuboid represents a different degree of summarization.



- **0-D cuboid** which holds highest level of summarization, is called **apex cuboid**.
  - This is the total sales summarized over all four dimensions.
  - The **apex cuboid** is typically denoted by **all**.

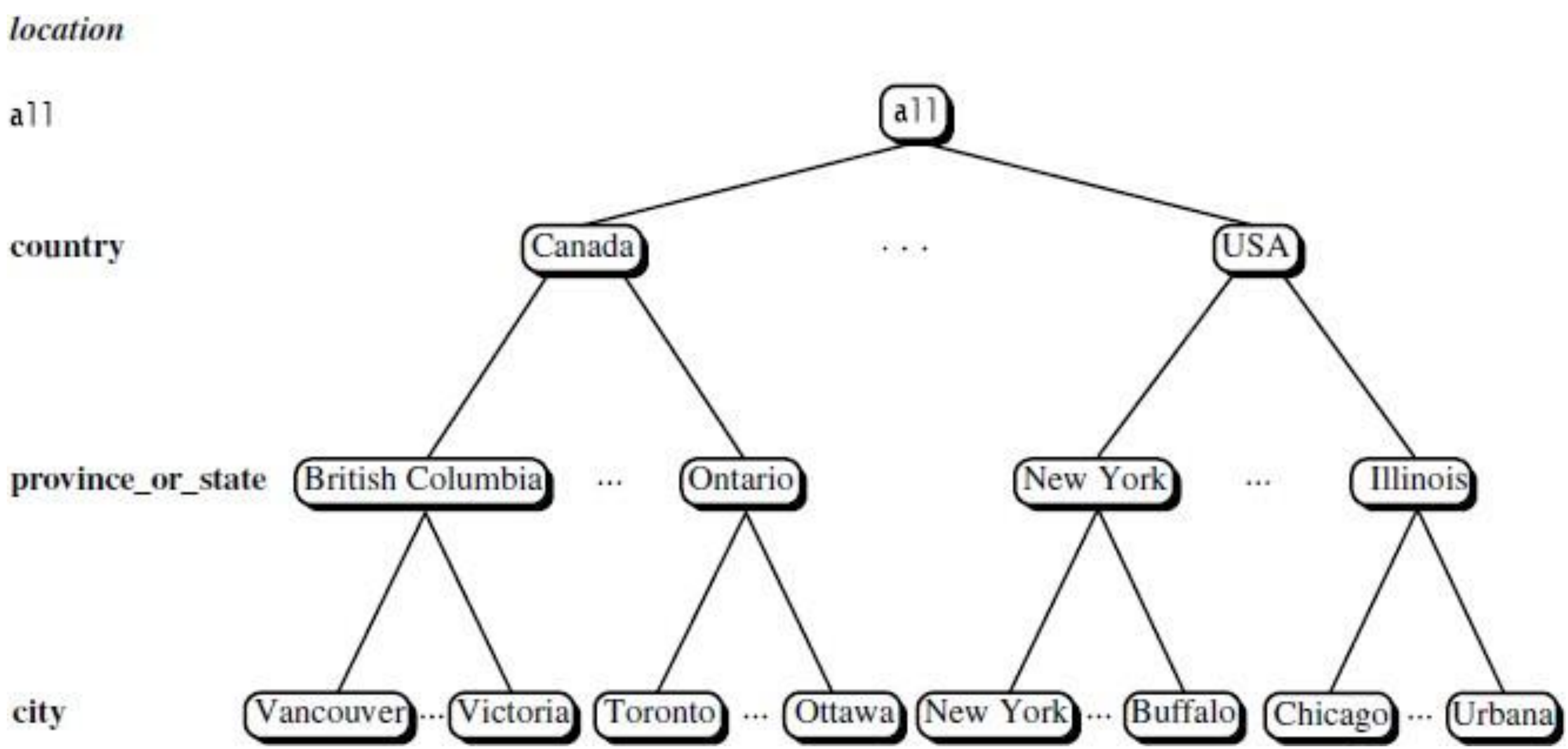
- A **3-D (nonbase) cuboid** for ***time, item, location***, summarized for all suppliers.

- The cuboid that holds the lowest level of summarization is called the **base cuboid.**
  - base cuboid for ***time***, ***item***, ***location***, and ***supplier*** dimensions

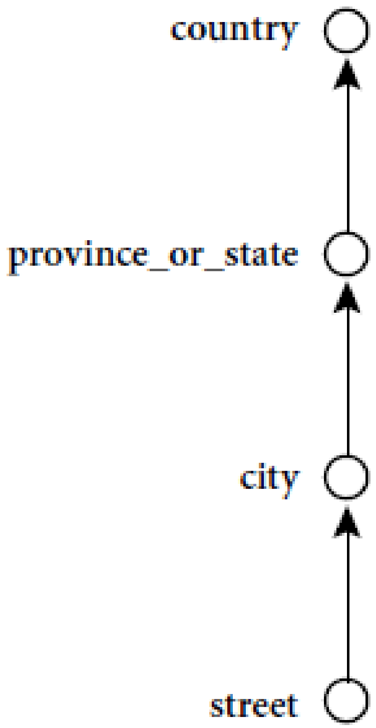
Concept Hierarchies

- A ***concept hierarchy*** defines a sequence of mappings from a set of low-level concepts to higher-level, more general concepts.
- Many concept hierarchies are implicit within the database schema.
- Concept hierarchies may be provided manually by system users, domain experts, or knowledge engineers, or may be automatically generated based on statistical analysis of the data distribution.
- A concept hierarchy that is a total or partial order among attributes in a database schema is called a *schema hierarchy*.
- Concept hierarchies may also be defined by discretizing or grouping values for a given dimension, resulting in a *set-grouping hierarchy*.
  - A total or partial order can be defined among groups of values.

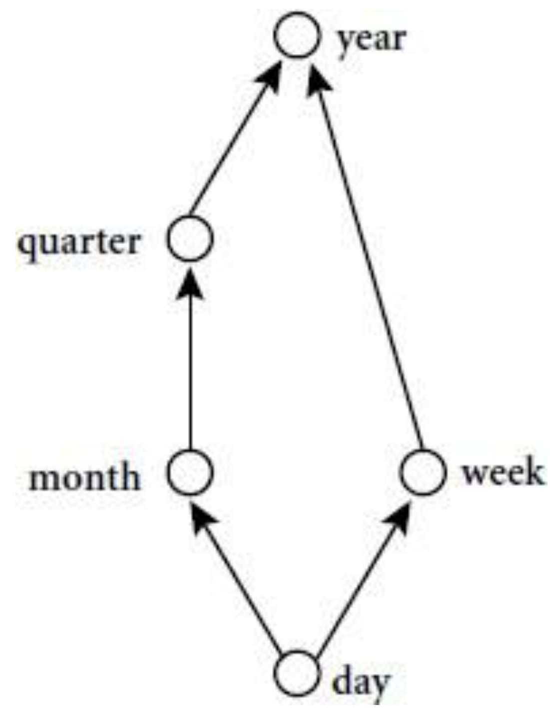
Concept Hierarchies – A concept hierarchy for the dimension location



Concept Hierarchies: Hierarchical and lattice structures of attributes in warehouse dimensions

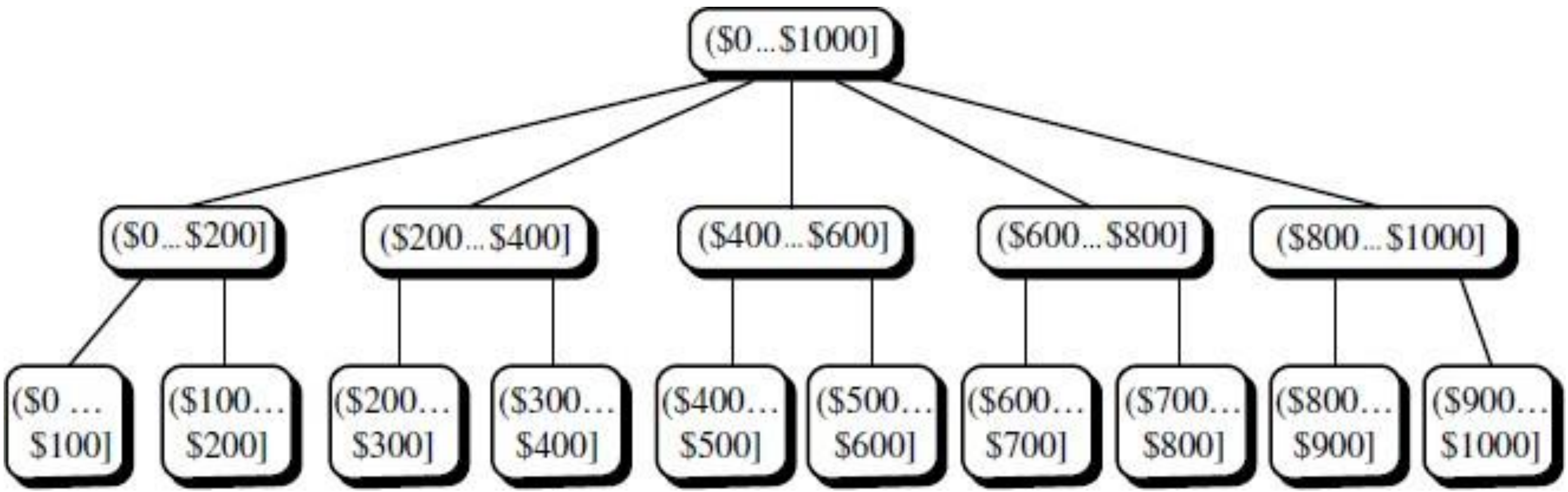


a hierarchy for *location*



a lattice for *time*

Concept Hierarchies:
A concept hierarchy for the attribute price



OLAP Operations

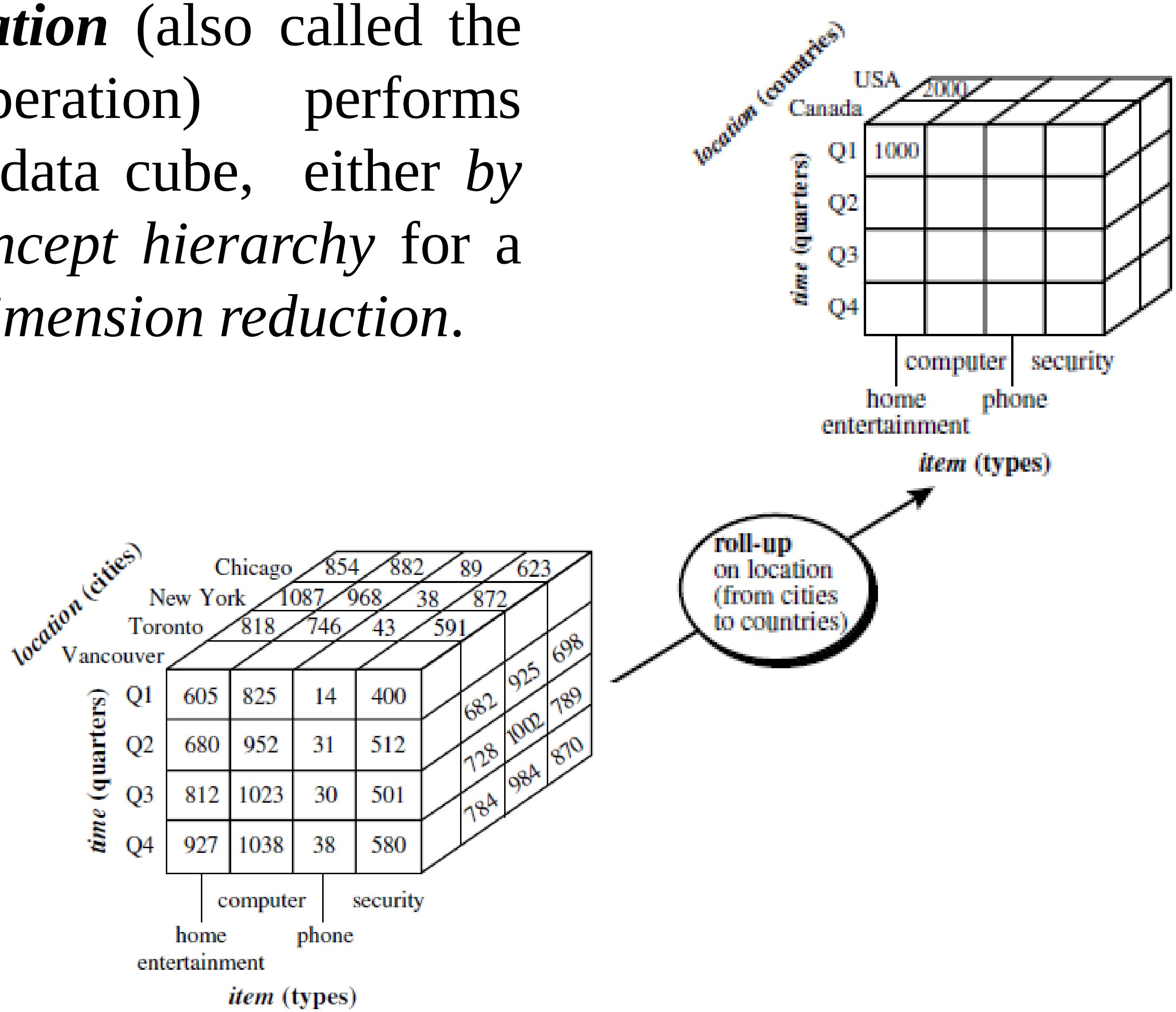
- How are concept hierarchies useful in OLAP?
- In the multidimensional model, data are organized into multiple dimensions, and each dimension contains multiple levels of abstraction defined by concept hierarchies.
- This organization provides users with the flexibility to view data from different perspectives.
- A number of OLAP data cube operations exist to materialize these different views, allowing interactive querying and analysis of the data at hand.

Typical OLAP Operations

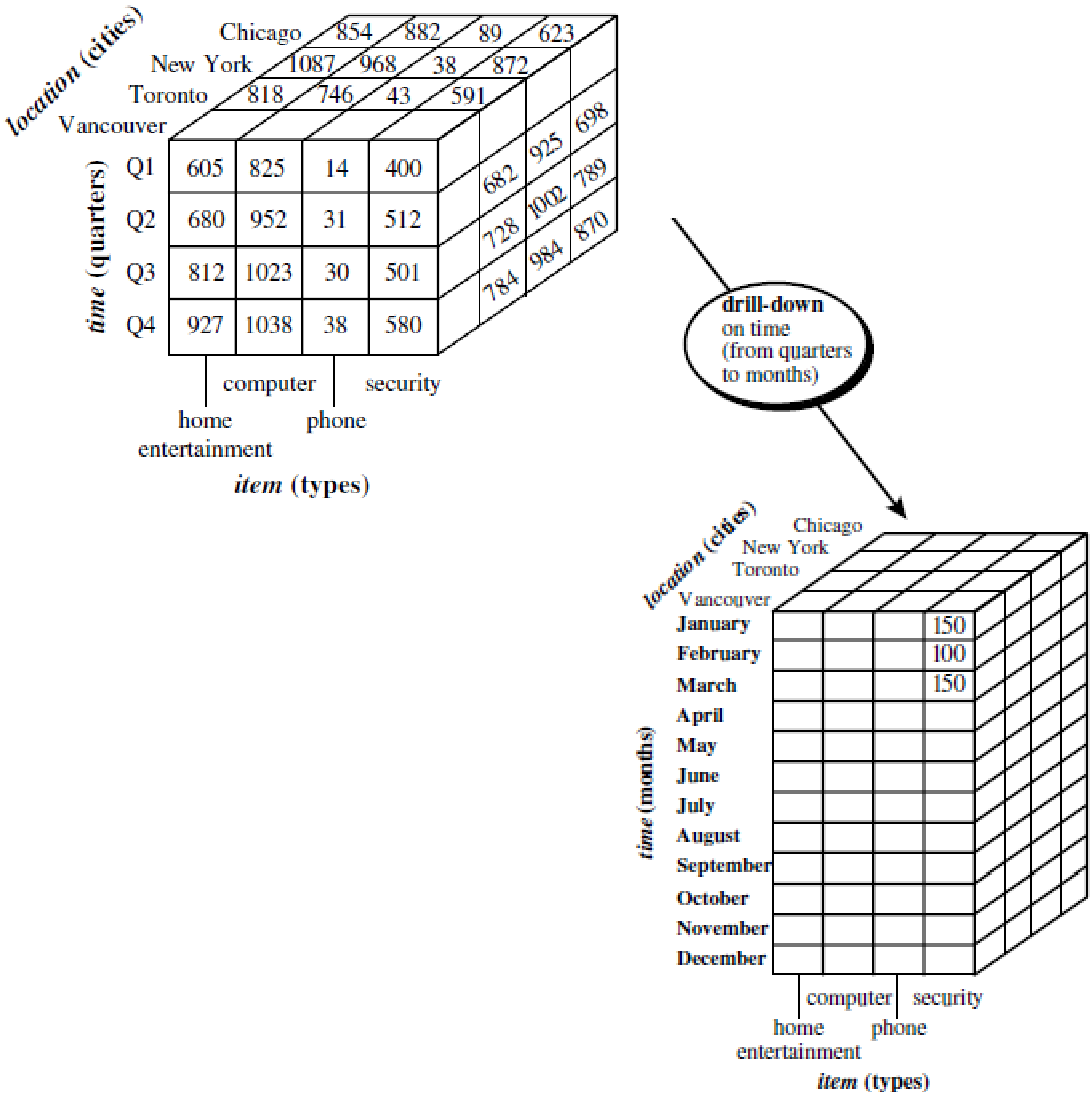
- **Roll up (drill-up):** summarize data
  - by climbing up hierarchy or by dimension reduction
- **Drill down (roll down):** reverse of roll-up
  - from higher level summary to lower level summary or detailed data, or introducing new dimensions
- **Slice and dice:** project and select
- **Pivot (rotate):**
  - reorient the cube, visualization, 3D to series of 2D planes

OLAP Operation: Roll-up

- The ***roll-up operation*** (also called the **drill-up** operation) performs aggregation on a data cube, either *by climbing up a concept hierarchy* for a dimension or *by dimension reduction*.



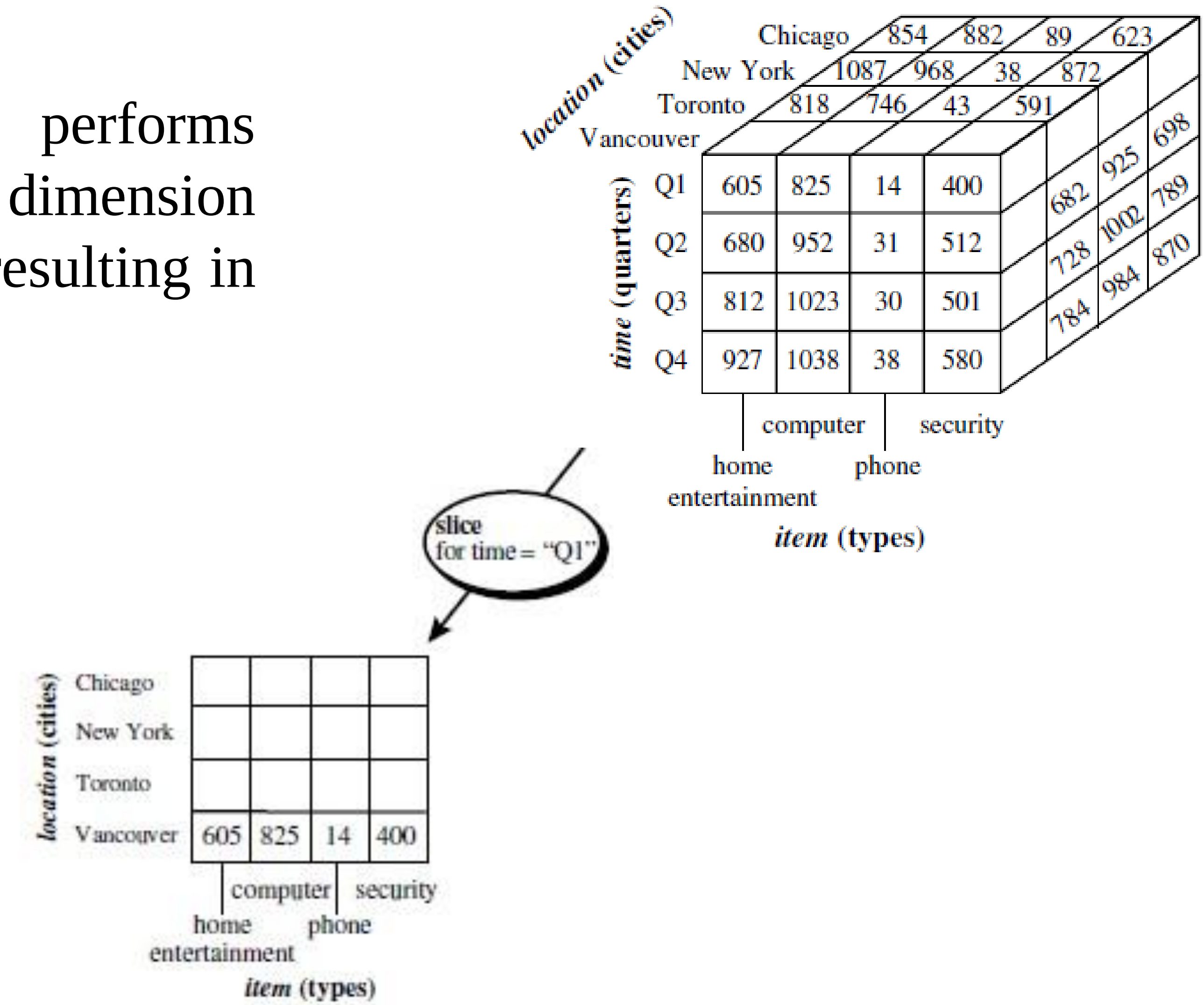
OLAP Operation: Drill-down



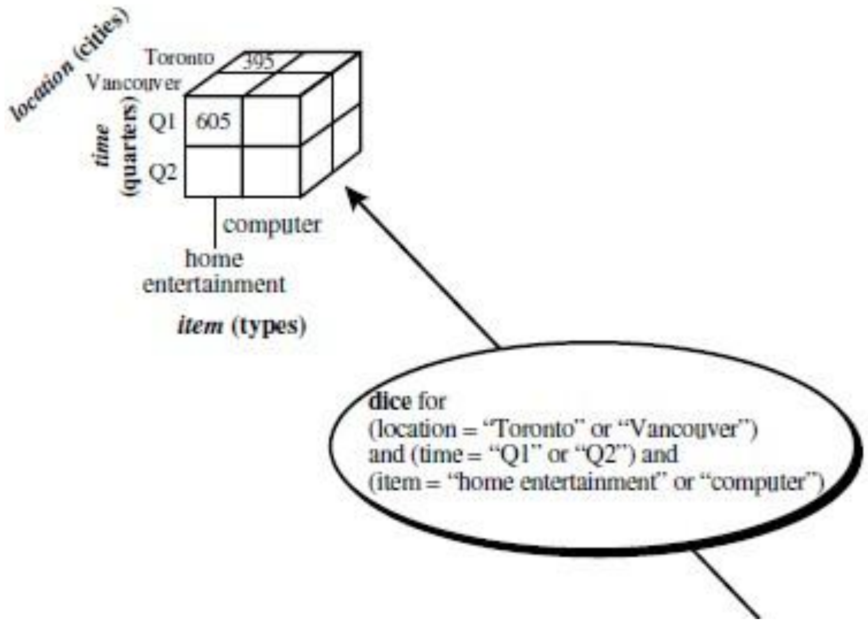
- ***Drill-down*** is the reverse of roll-up. It navigates from less detailed data to more detailed data.
- Drill-down can be realized by either stepping down a concept hierarchy for a dimension or introducing additional dimensions.

OLAP Operation: Slice

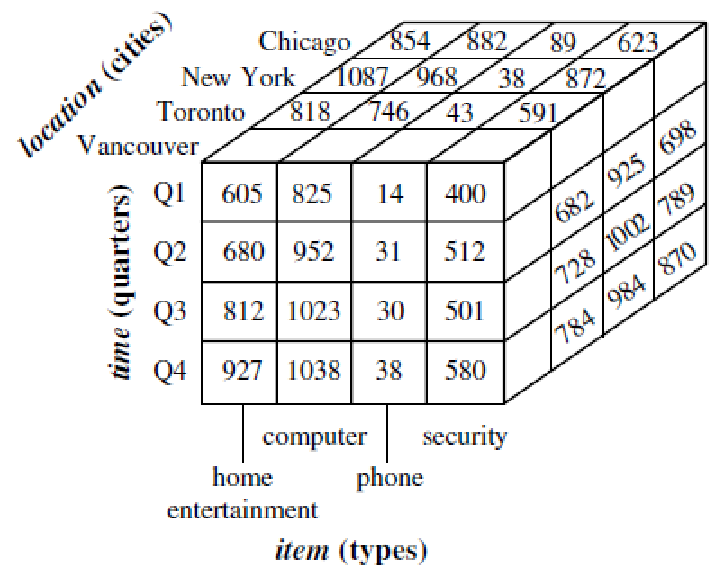
- The ***slice operation*** performs a selection on one dimension of the given cube, resulting in a subcube.



OLAP Operation: Dice

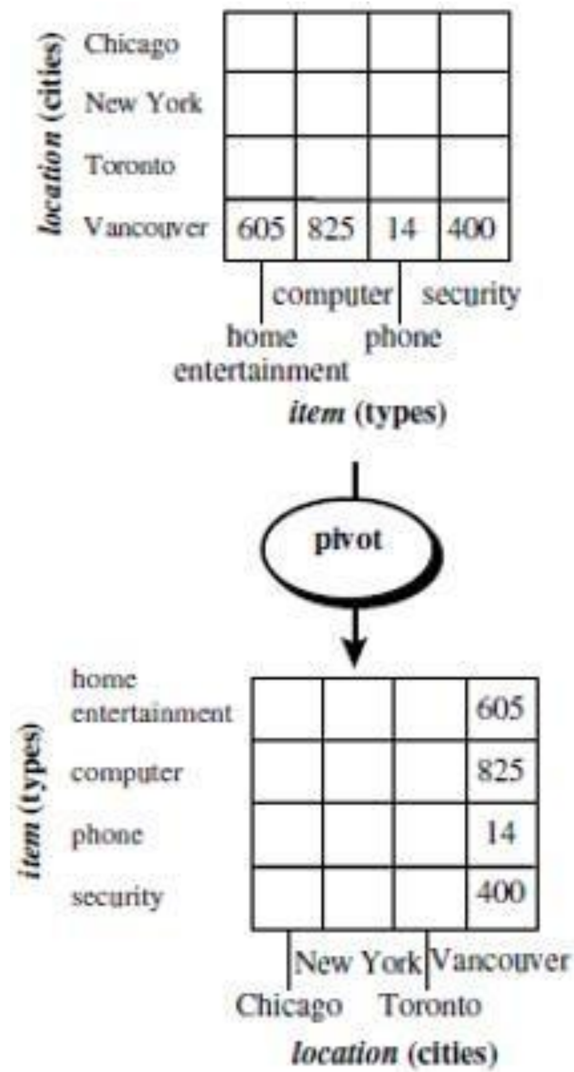


- The ***dice operation*** defines a subcube by performing a selection on two or more dimensions.



OLAP Operation: Pivot

- ***Pivot (rotate)*** is a visualization operation that rotates the data axes in view in order to provide an alternative presentation of the data.



Data Warehouse Usage

- 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 Warehouse and OLAP: Summary

- A **data warehouse** is a *subject-oriented*, *integrated*, *time-variant*, and *nonvolatile* collection of data organized in support of management decision making.
  - Several factors distinguish data warehouses from operational databases.
  - Because the two systems provide quite different functionalities and require different kinds of data, it is necessary to maintain data warehouses separately from operational databases.
- A ***multidimensional data model*** is typically used for the design of corporate *data warehouses*.
  - A multidimensional data model can adopt a *star schema*, *snowflake schema*, or *fact constellation schema*.
  - The core of the *multidimensional model* is the data cube, which consists of a large set of *facts* (or *measures*) and a number of *dimensions*.
  - Dimensions are the entities or perspectives with respect to which an organization wants to keep records and are hierarchical in nature.
- A **data cube** consists of a lattice of cuboids, each corresponding to a different degree of summarization of the given multidimensional data.

Data Warehouse and OLAP: Summary

- **Concept hierarchies** organize the values of dimensions into gradual levels of abstraction.
- **On-line analytical processing (OLAP)** can be performed in data warehouses using the multidimensional data model.
  - Typical OLAP operations include roll-up, drill-down,, slice-and-dice, pivot (rotate), as well as statistical operations such as ranking and computing moving averages and growth rates.
- Data warehouses often adopt **a three-tier architecture**.
  - The bottom tier is a warehouse database server, which is a relational database system.
  - The middle tier is an OLAP server, and
  - The top tier is a client, containing query and reporting tools.
- A data warehouse contains **back-end tools and utilities** for populating and refreshing the warehouse.
  - data extraction, data cleaning, data transformation, loading, refreshing, and warehouse management.
- Data warehouse **metadata** are data defining the warehouse objects.
  - A metadata repository provides details regarding the warehouse structure, data history, the algorithms used for summarization, mappings from the source data to warehouse form.