Ambo University||WC

School of Technology & Informatics

Information Science

Concept Hierarchy

OUTLINES

- Concept Hierarchy
- Typical OLAP Operations
- A Star-Net Query Model
- Data Warehouse Design Process
- Three Data Warehouse Models
- The Complete Data Warehouse System
- Three-Tier Data warehouse Systems
- Data warehousing Processes
- OLAP functionalities on Data warehouses
- Why Data Mining
- Summary

Concept Hierarchy

- Dimensions are organized into concept hierarchies
- A concept hierarchy defines a sequence of mappings from a set of low level concepts to higher-level and more general concepts.
- As shown in the concept hierarchy, each level refers to values of some type.
- The type of hierarchy define ordering which can be partial ordering or total ordering.
- Location dimension can be seen as a total ordering

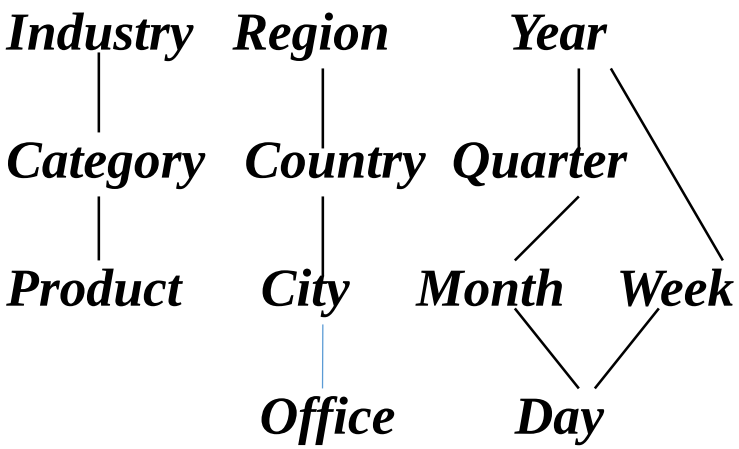
continent → country → Region → Zone → city → kifle ketema → kebele

- Time dimension shows partial ordering

second → minute → hour → day → {month → quarter, week} → year

Concept Hierarchy

- More ordering



Concept Hierarchy

- Many concept hierarchy are implicit within the database schema as location and time are described by the fields shown above.

  define dimension time as (time_key, day, day_of_week, month, quarter, year)

- These are called **schema hierarchy.**

- The concept hierarchy for location is schema hierarchy where as for annual income concept hierarchy may be set as grouping hierarchy
- Concept hierarchies may also be defined by discretizing or grouping values for a given dimension or attribute resulting in a **set-grouping hierarchy**

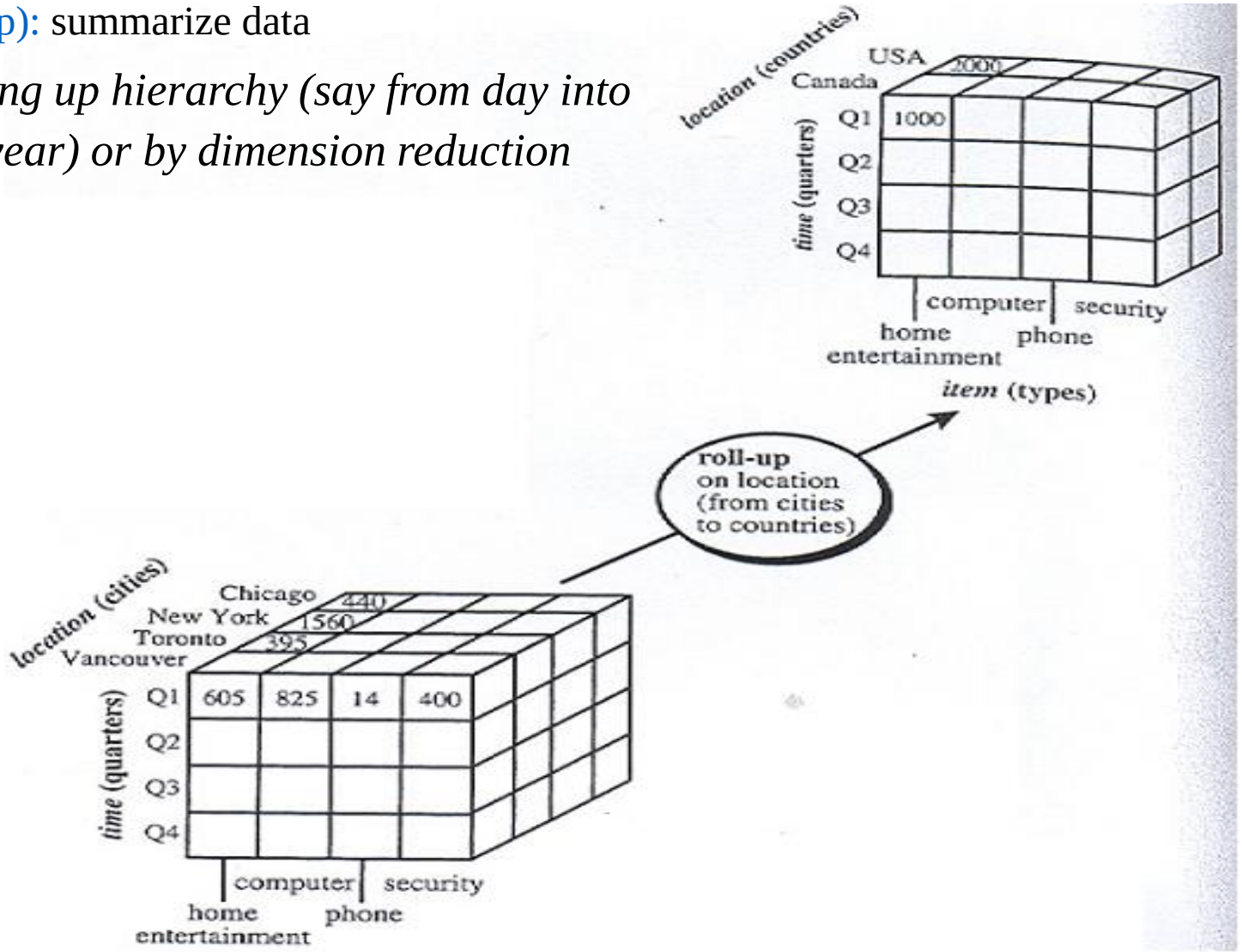
Typical OLAP Operations

- In multidimensional model, data are organized into multiple dimensions, and each dimension contains multiple level of abstraction defined by concept hierarchies.
- This organization provides users with flexibility to view data from different perspectives.

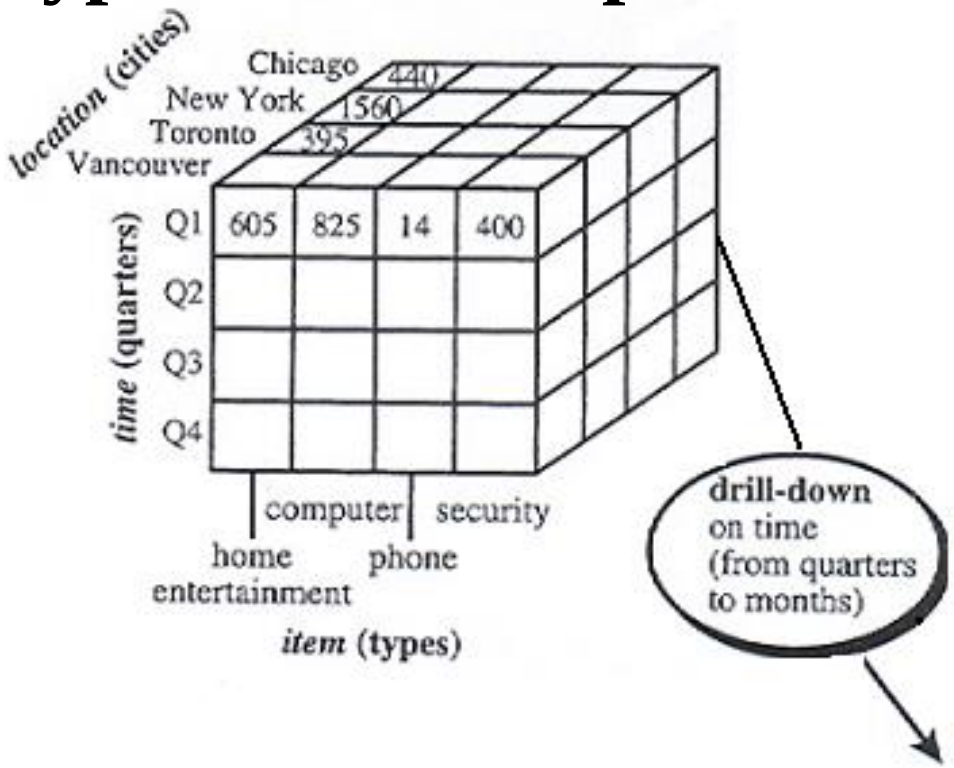
- Different OLAP data cube operations exists to materialize these views:
  - Roll up (drill-up)
  - Drill down (roll down)
  - Slice and dice
  - Pivot (rotate)

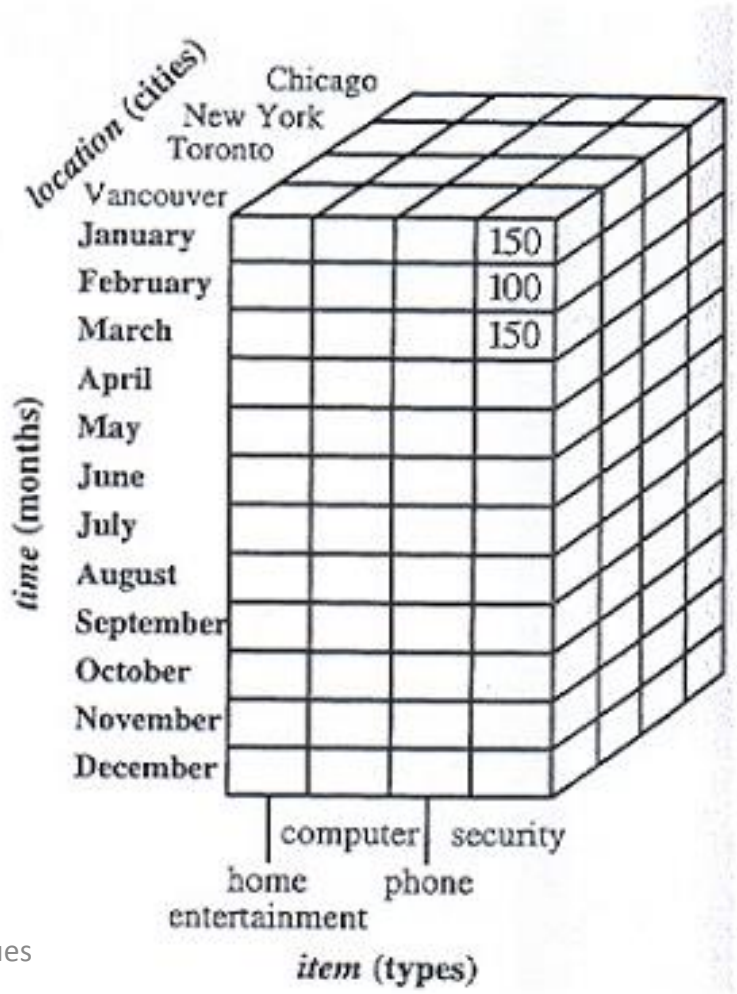
Typical OLAP Operations

- Roll up (drill-up): summarize data
  - *by climbing up hierarchy (say from day into week or year) or by dimension reduction*



Typical OLAP Operations

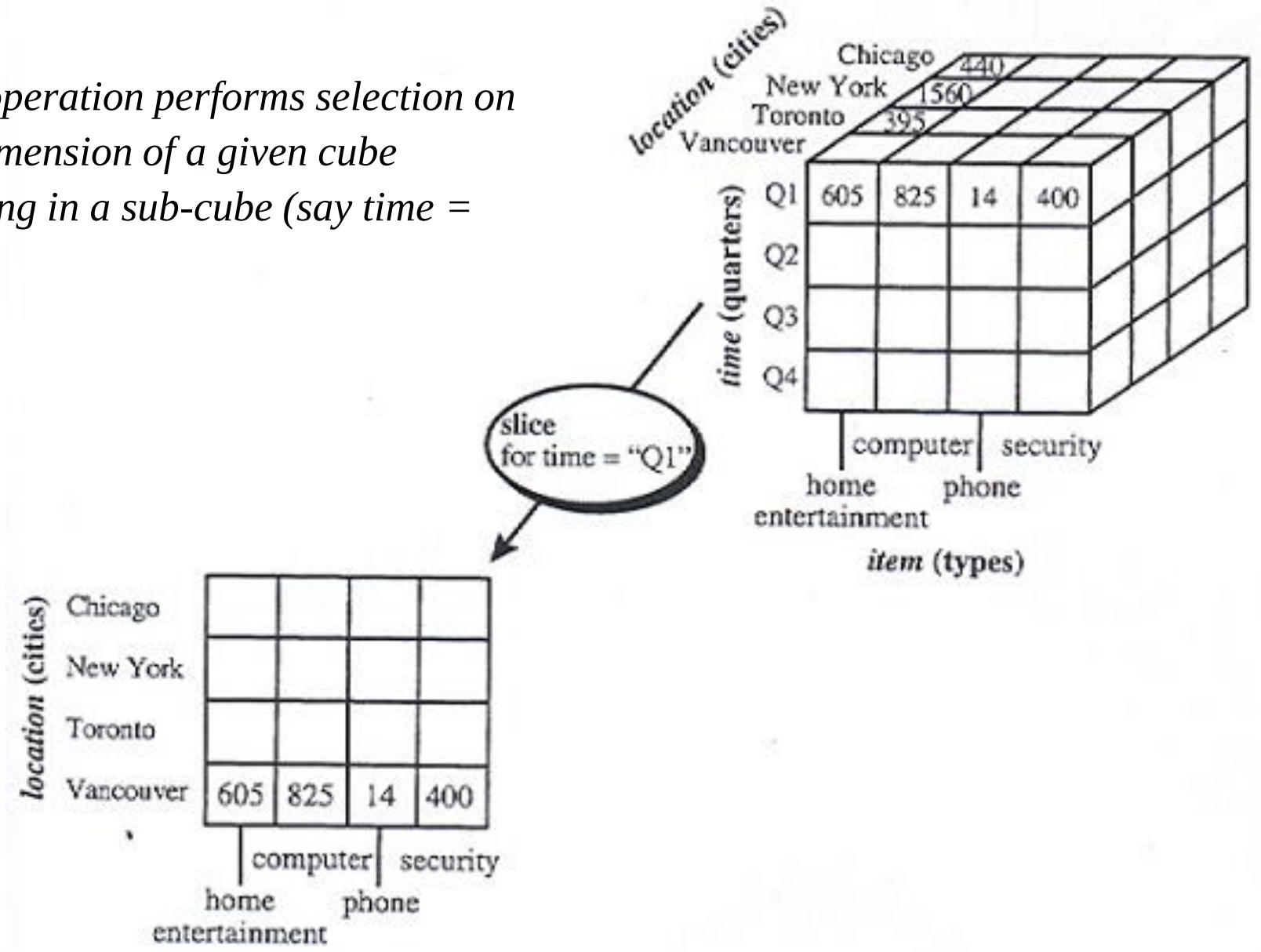




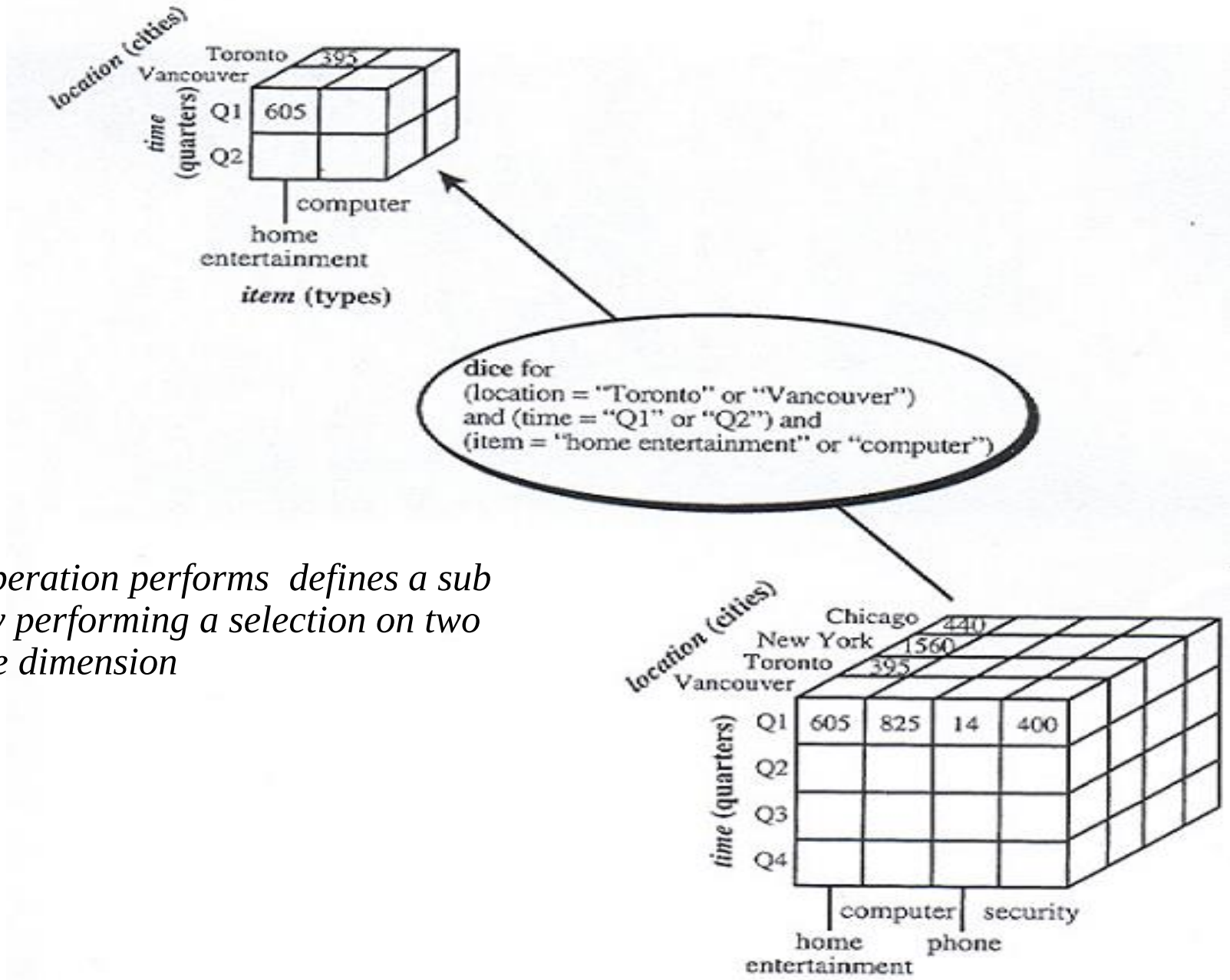
- Drill down (roll down): reverse of roll-up
  - *from higher level summary to lower level summary (say from region to town) or detailed data, or introducing new dimensions*

Typical OLAP Operations

- Slice:
  - *Slice operation performs selection on one dimension of a given cube resulting in a sub-cube (say time = Q1)*



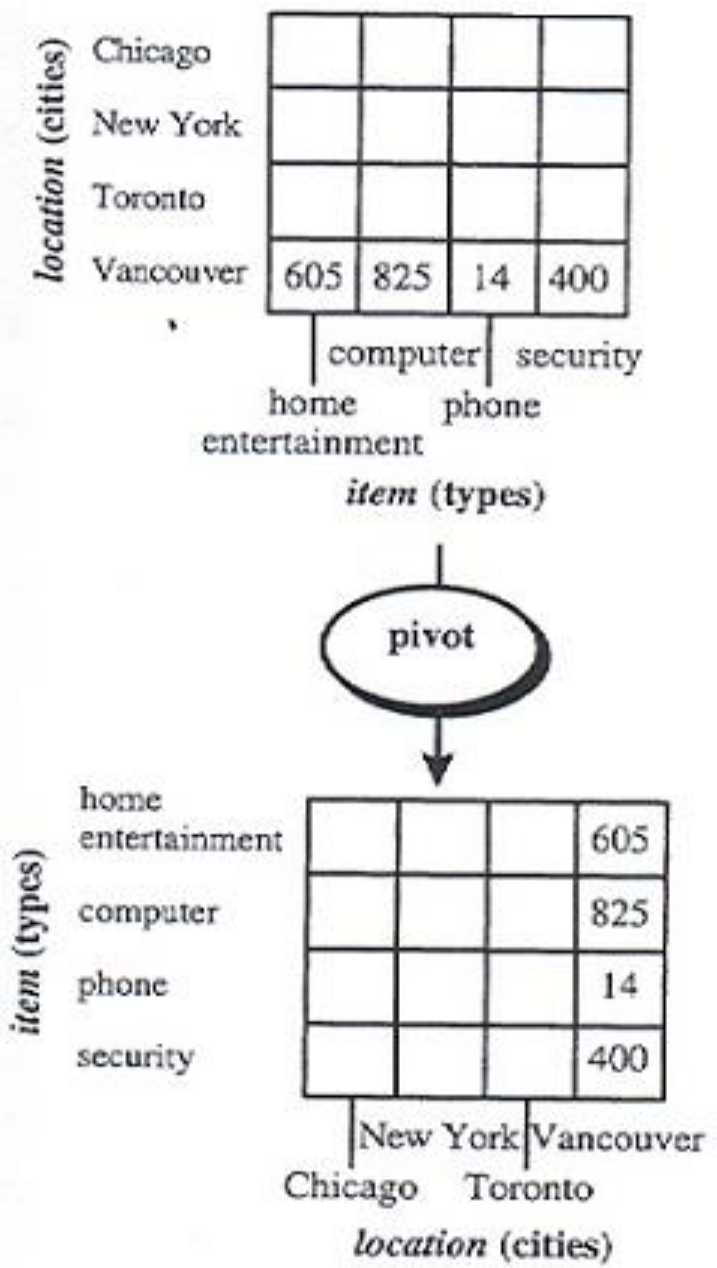
Typical OLAP Operations



- Dice:
  - *Dice operation performs defines a sub cube by performing a selection on two or more dimension*

Typical OLAP Operations

- Pivot (rotate):
  - *reorient the cube, visualization, 3D to series of 2D planes.*



Data Warehouse Design Process

- Data warehouse design process consists of 4 steps
  1. Choosing a business process to model, e.g., orders, invoices, sales, shipment, inventory, account administration, general ledger etc.
  2. Choosing the dimensions that will apply to each fact table record
  3. Choosing the *grain* (*atomic level of data*) of the business process that will be represented in the fact table
  4. Choosing the measure that will populate each fact table record

Three Data Warehouse Models

- From the architecture point of view, there are three data warehouse models described as Enterprise warehouse, Data Mart, or Virtual warehouse

- Enterprise warehouse
  - collects all information about subjects that span the entire organization (*customers*, *products*, *sales*, *assets*, *personnel*)
  - Requires extensive business modeling (may take years to design and build)

Three Data Warehouse Models

- 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
  - For example, a marketing data mart my confine its subject to customer, product and sales
  - Data marts depending on the data source can be dependent or independent
    - **Dependent** data mart are sourced directly from the enterprise data warehouse
    - **Independent** data marts source can be from some operational data sources, external information providers, from data generated locally within a particular department or geographic area

Three Data Warehouse Models

- Virtual warehouse
  - A set of views over operational databases
  - Only some of the possible summary views may be materialized
  - Easy to build but requires excess capacity on operational database servers

Three-Tier Data warehouse Systems

- Data warehouse often adopt three-tier architecture
  - Warehouse database server (The bottom tier)
  - On-Line Analytical Processing OLAP servers (Middle tier)
  - Clients(the top tier)

Three-Tier Data warehouse Systems

- **Warehouse database server (The bottom tier)**
  - Responsible to process the primary data source for the data warehouse
  - The source is usually a relational DBMS, rarely flat files
  - Back end tools and utilities are used to feed data into the middle tier
  - The tools and utilities perform data extraction, cleaning and transformation as well as load and refresh functions to update the warehouse in this tier

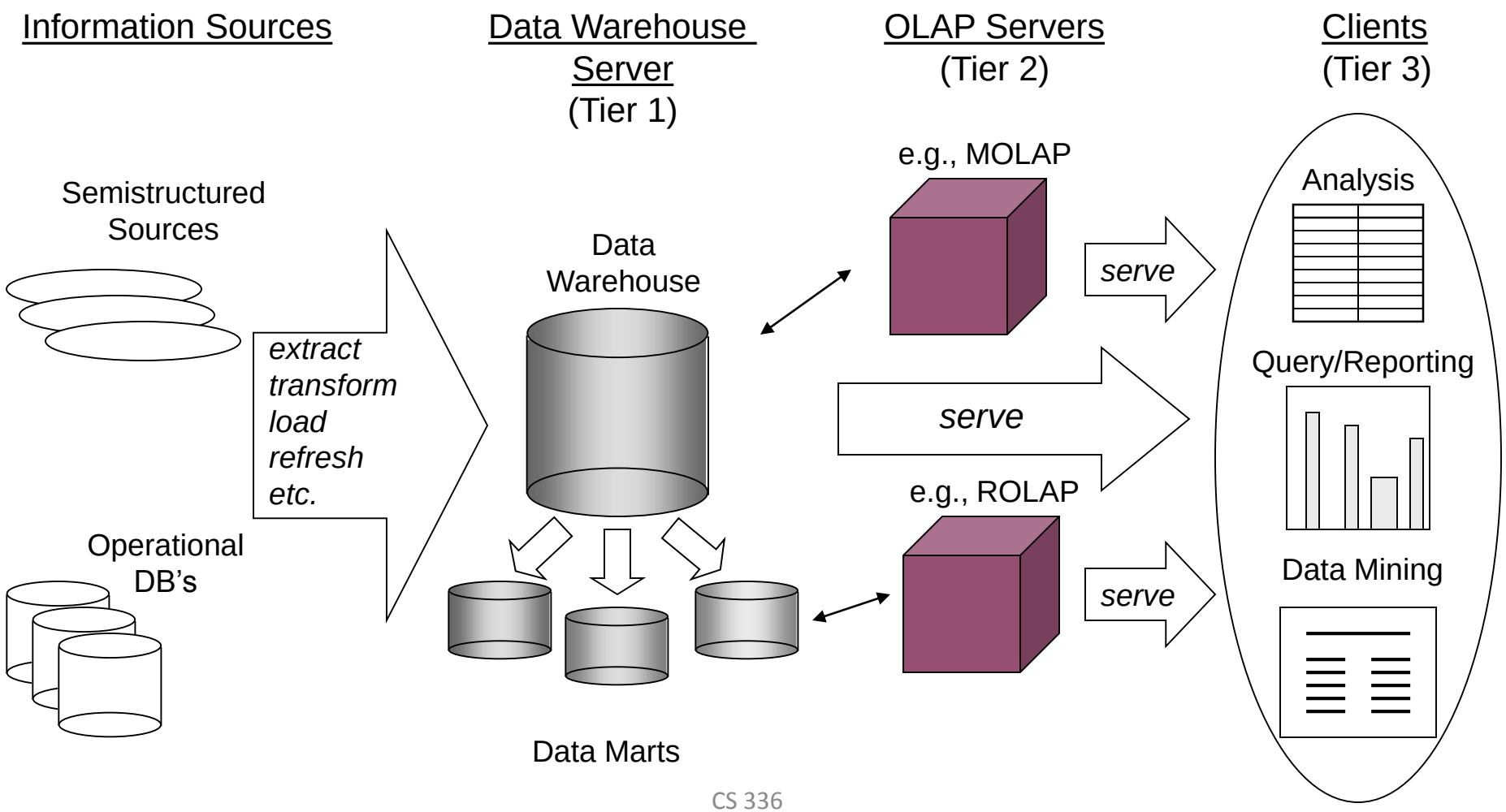
Three-Tier Data warehouse Systems

- **OLAP servers (Middle tier)**
  - Responsible to manipulate the warehouse to get results of the data warehouse OLAP functionalities
  - Implemented either as Relational OLAP (ROLAP) or Multidimensional OLAP (MOLAP)
  - ROLAP: extended relational DBMS that maps operations on multidimensional data to standard relational operators
  - Multidimensional OLAP (MOLAP): special-purpose server that directly implements multidimensional data and operations

Three-Tier Data warehouse Systems

- **Clients(the top tier)**
  - Responsible for any user interaction with the data ware house
  - Includes querying and reporting tools, analysis tools, and data mining tools

The Complete Data Warehouse System



Data warehousing Processes

- Data warehouse technology includes
  - Data cleansing (removing noise and inconsistent data)
  - Data integration (combining multiple data sources into one data warehouse)
  - On-Line Analytical Processing (OLAP)

OLAP functionalities on Data warehouses

- OLAP is analysis technique have functionalities such as
  - Summarization
  - Consolidation
  - Aggregation as well as
  - The ability to view information from different angle

Why Data Mining

- Although Online Analytical processing tools support multidimensional analysis and decision making, additional data analysis tools are required for in depth analysis such as
  - Data classification
  - Clustering and
  - Characterization of data changes over time

- The abundance of data, coupled with the need for powerful data analysis tools has been described as data rich but information poor situation

Summary

- Data mining (Knowledge Discovery in Databases) consists of iterative sequences of seven steps
  1. Learning the application domain:
     - Learn relevant prior knowledge and goals of application
  2. Creating a target data set: data selection
     - Data cleaning and preprocessing: (may take 60% of effort!)
     - Data reduction and transformation:
       - Find useful features, dimensionality/variable reduction, invariant representation
  3. Choosing functions of data mining
     - summarization, classification, regression, association, clustering.

Summary

4. Choosing the mining algorithm(s)
   - Data mining: search for patterns of interest

5. Identify the relevant knowledge by measuring the mining result interestingness
   - Pattern evaluation

6. Present the knowledge to the user
   - knowledge presentation
   - visualization, transformation, removing redundant patterns, etc.

7. Use of discovered knowledge