Data Warehousing and Data Mining
TBC-604(1)

Syllabus

UNIT 1 : **Introduction to Data Warehousing**

Concept of Data Warehouse, DBMS verses data warehouse, Data Marts, Metadata, Multidimensional data model, Multidimensional database, Data warehouse Measures, their categorization and computation, Multi-dimensional database hierarchies.

UNIT 2 : Data Warehouse Architecture

Operations in OLAP, Advantages of OLAP over OLTP, Three-Tier Data Warehouse architecture, OLAP Guidelines, Multidimensional versus Multirelational OLAP , Categories of Tools, OLAP Tools and the Internet

Unit 3: Introduction to Data Mining

Basic Concepts of Data Mining; Data Mining primitives: Task-relevant data, mining objective, measures and identification of patterns, KDD versus data mining, data mining tools and applications.

Data Mining Query Languages: Data specification, specifying kind of knowledge, hierarchy specification, pattern presentation & visualization specification, data mining languages and standardization of data mining, Architectures of Data Mining Systems.

Syllabus

UNIT 4 : **Data Mining Techniques**

Association rules: Association rules from transaction database & relational database, correlation analysis; Classification and predication using decision tree induction. Introduction to Clustering techniques, partition method, and Hierarchical method.

UNIT 5 : Overview of Advanced Features of Data Mining

Mining complex data objects, Spatial databases, Multimedia databases, Time series and Sequence data; mining Text Databases and mining Word Wide Web.

What Is a Data Warehouse?

A single, complete and consistent store of data obtained from a variety of different sources made available to end users in a what they can understand and use in a business context.

The data warehouse is an informational environment

Provides an integrated and total view of the enterprise

Makes the enterprise's current and historical information easily available for decision making

Makes decision-support transactions possible without hindering operational systems

Renders the organization's information consistently

Presents a flexible and interactive source of strategic information

- “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

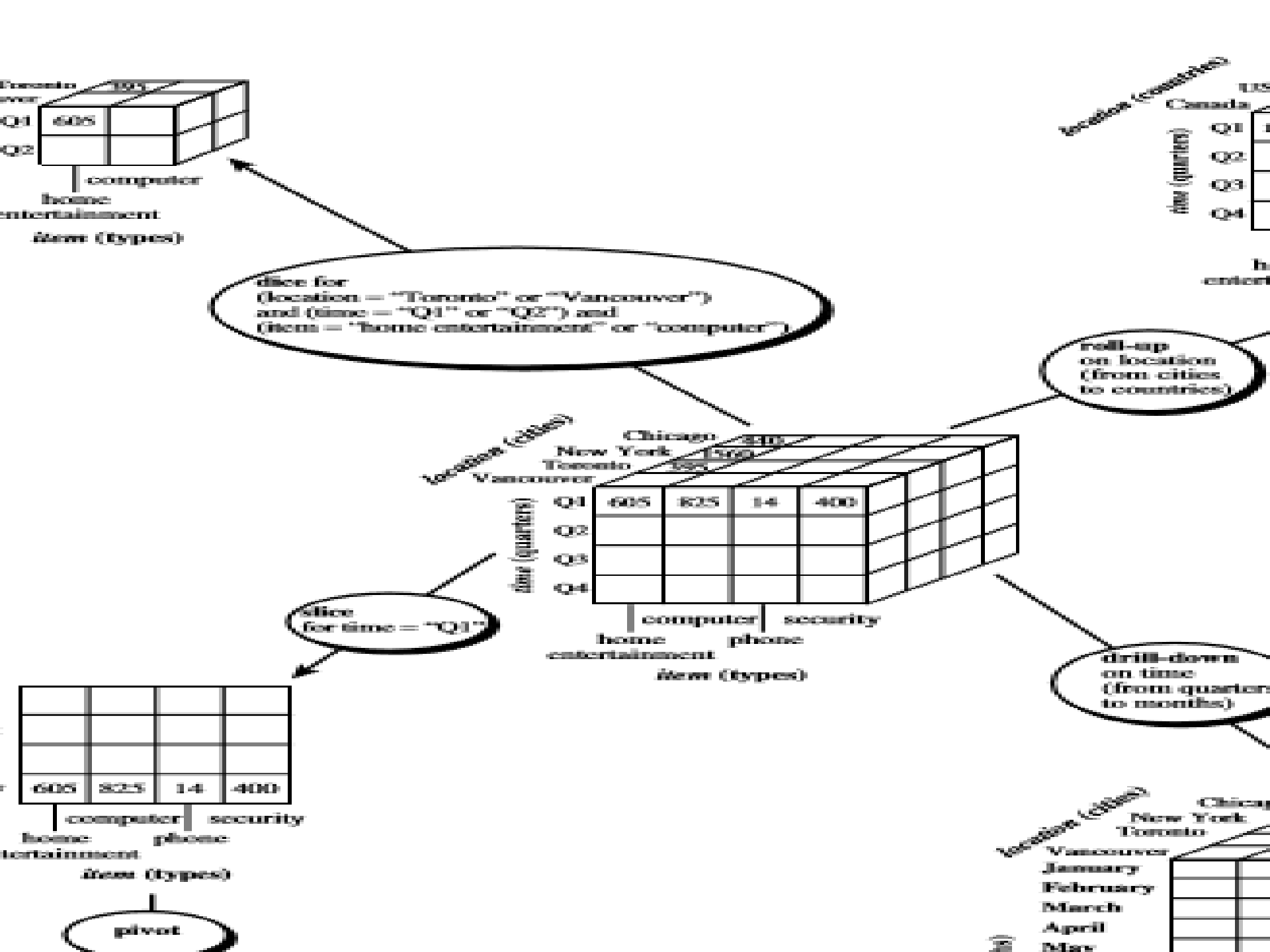
- Data warehousing:
  - The process of constructing and using data warehouses

UNIT 2 : Data Warehouse Architecture

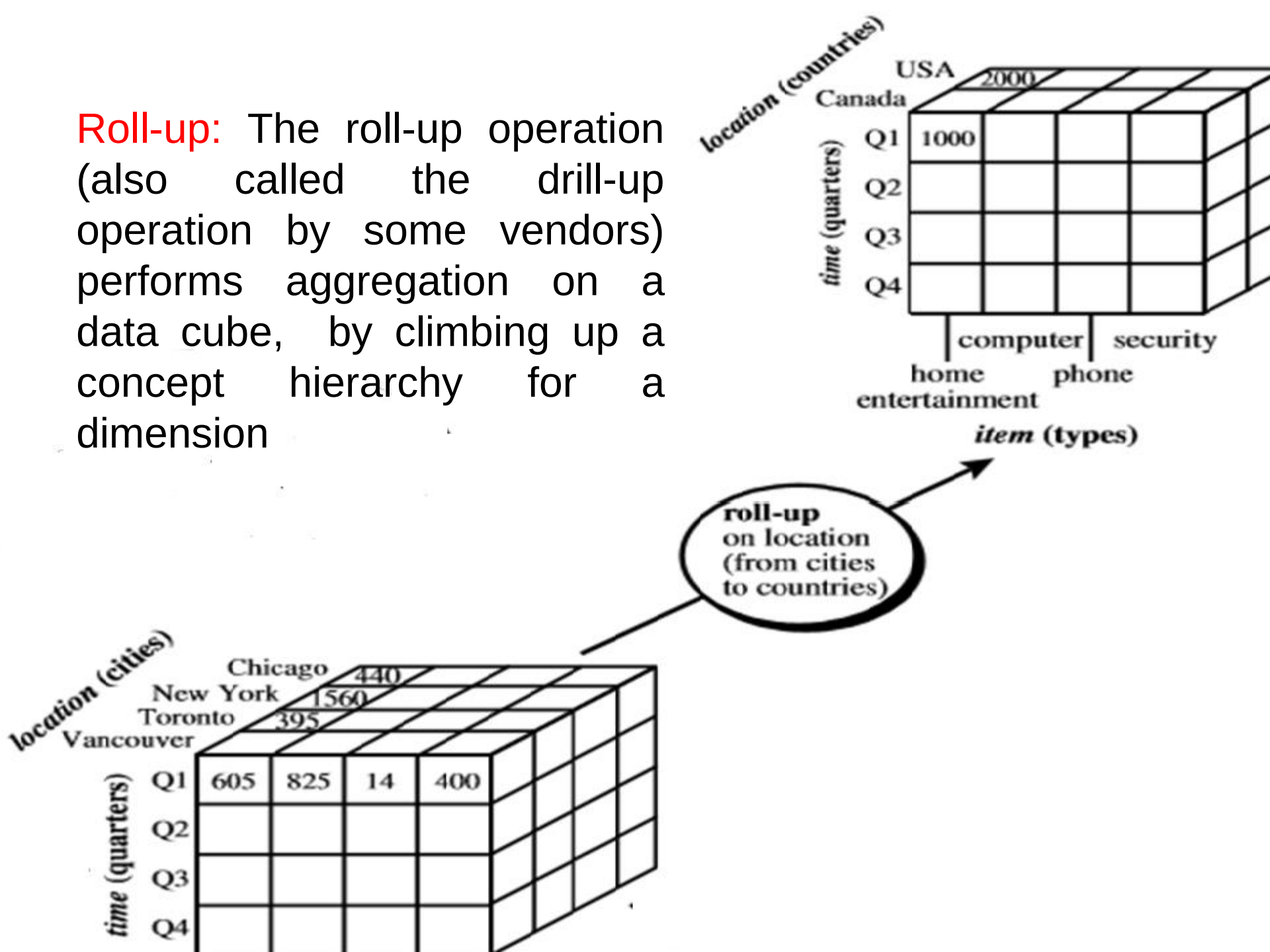
Operations in OLAP, Advantages of OLAP over OLTP, Three-Tier Data Warehouse architecture, OLAP Guidelines, Multidimensional versus Multirelational OLAP , Categories of Tools, OLAP Tools and the Internet

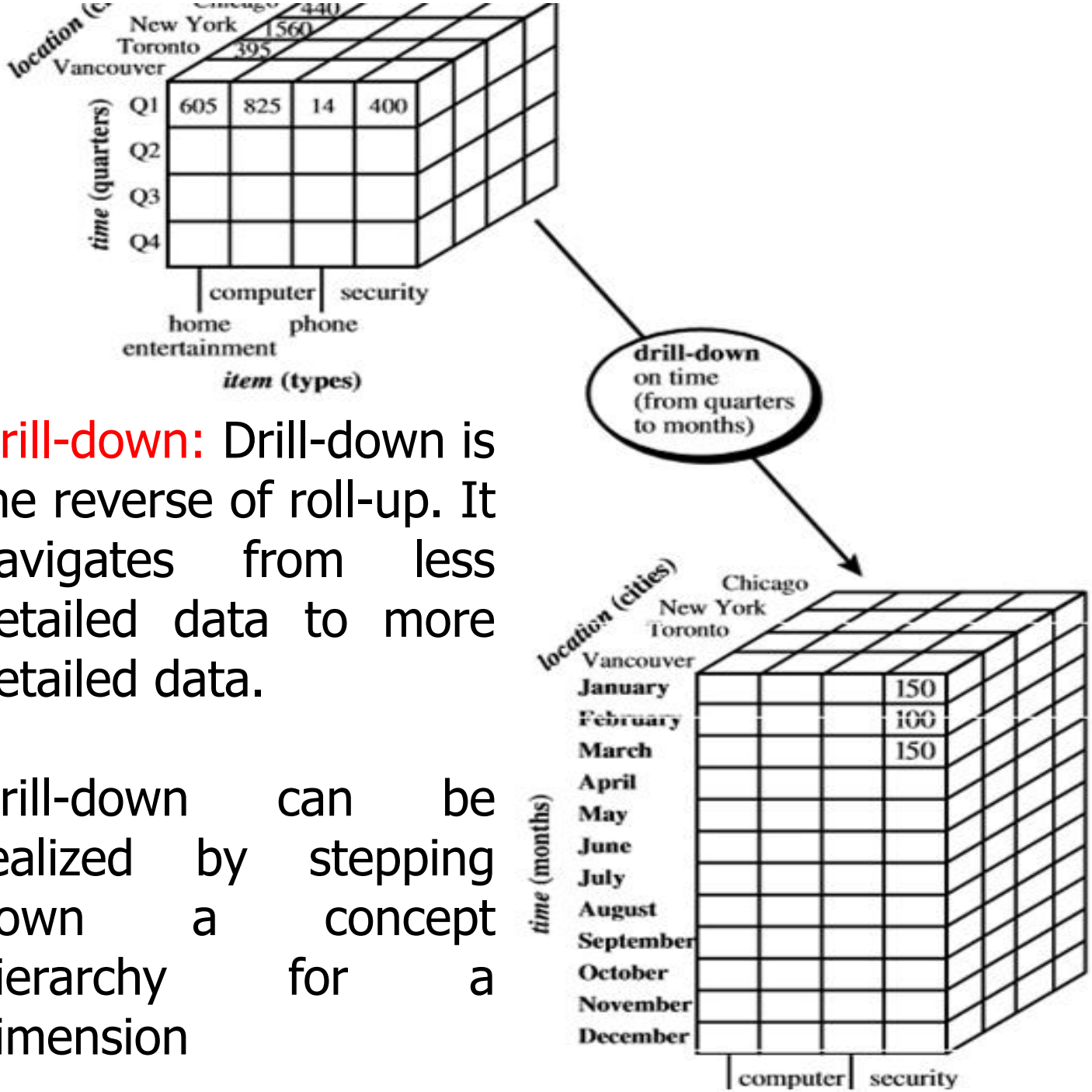
OLAP OPERATIONS

- 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*


dice for
(location = "Toronto" or "Vancouver")
and (time = "Q1" or "Q2") and
(item = "home entertainment" or "computer")
roll-up
on location
(from cities
to countries)
slice
for time = "Q1"
drill-down
on time
(from quarters
to months)
pivot
location (cities)
Chicago
New York
Toronto
Vancouver
time (quarters)
Q1
Q2
Q3
Q4
605
825
14
400
computer
security
home
entertainment
phone
item (types)
location (countries)
USA
Canada
January
February
March
April
May

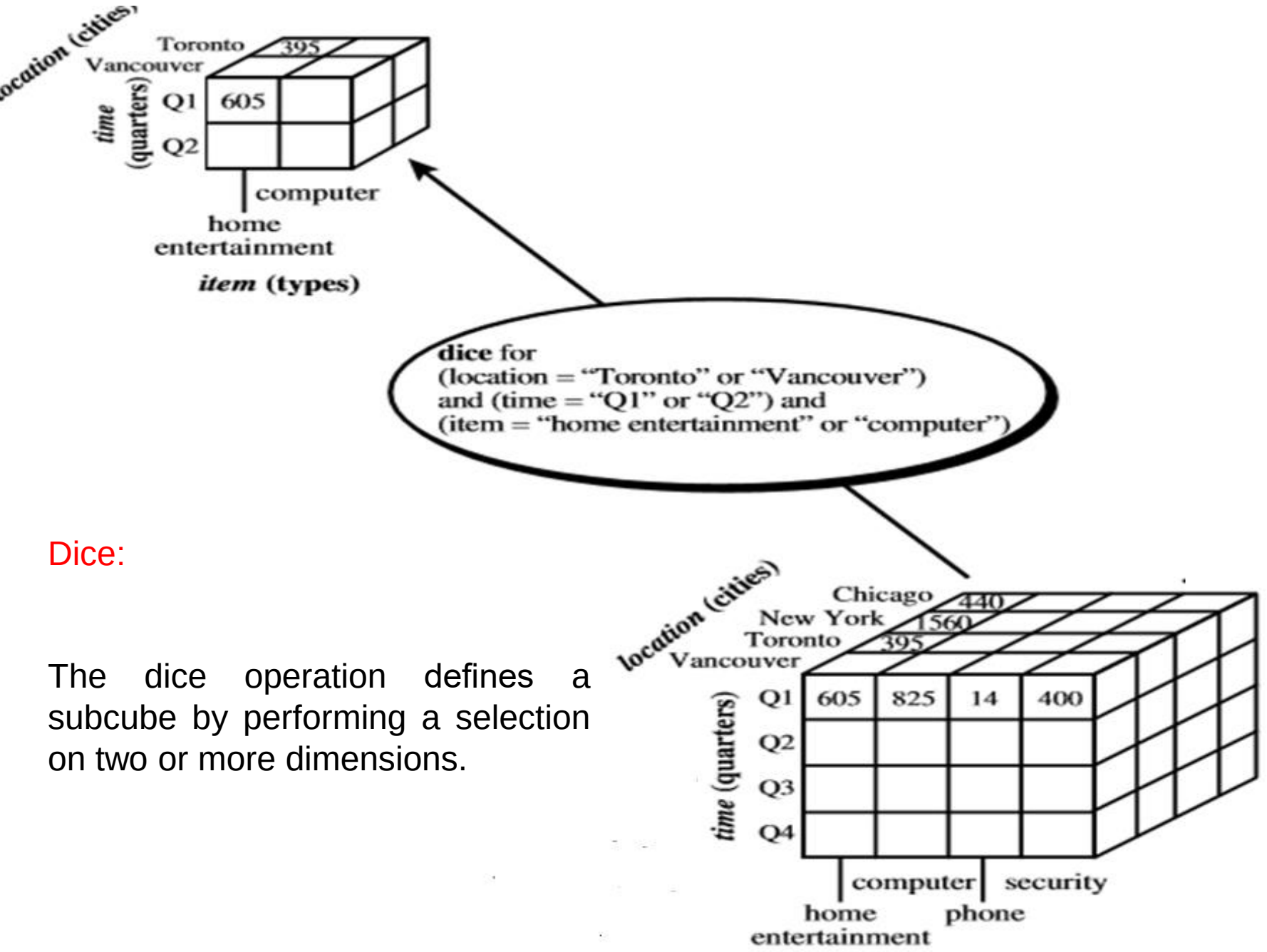
Roll-up: The roll-up operation (also called the drill-up operation by some vendors) performs aggregation on a data cube, by climbing up a concept hierarchy for a dimension





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 stepping down a concept hierarchy for a dimension

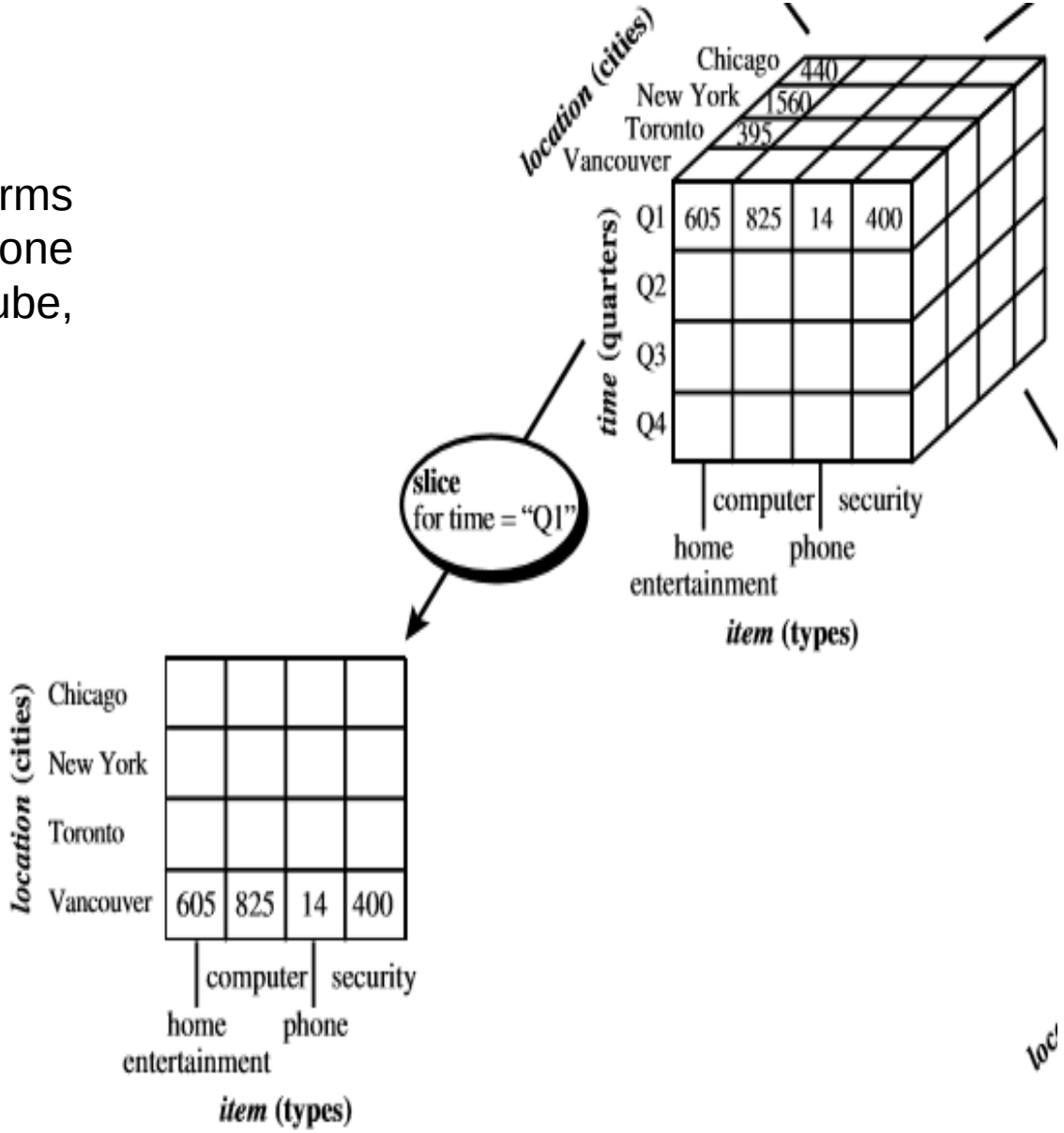


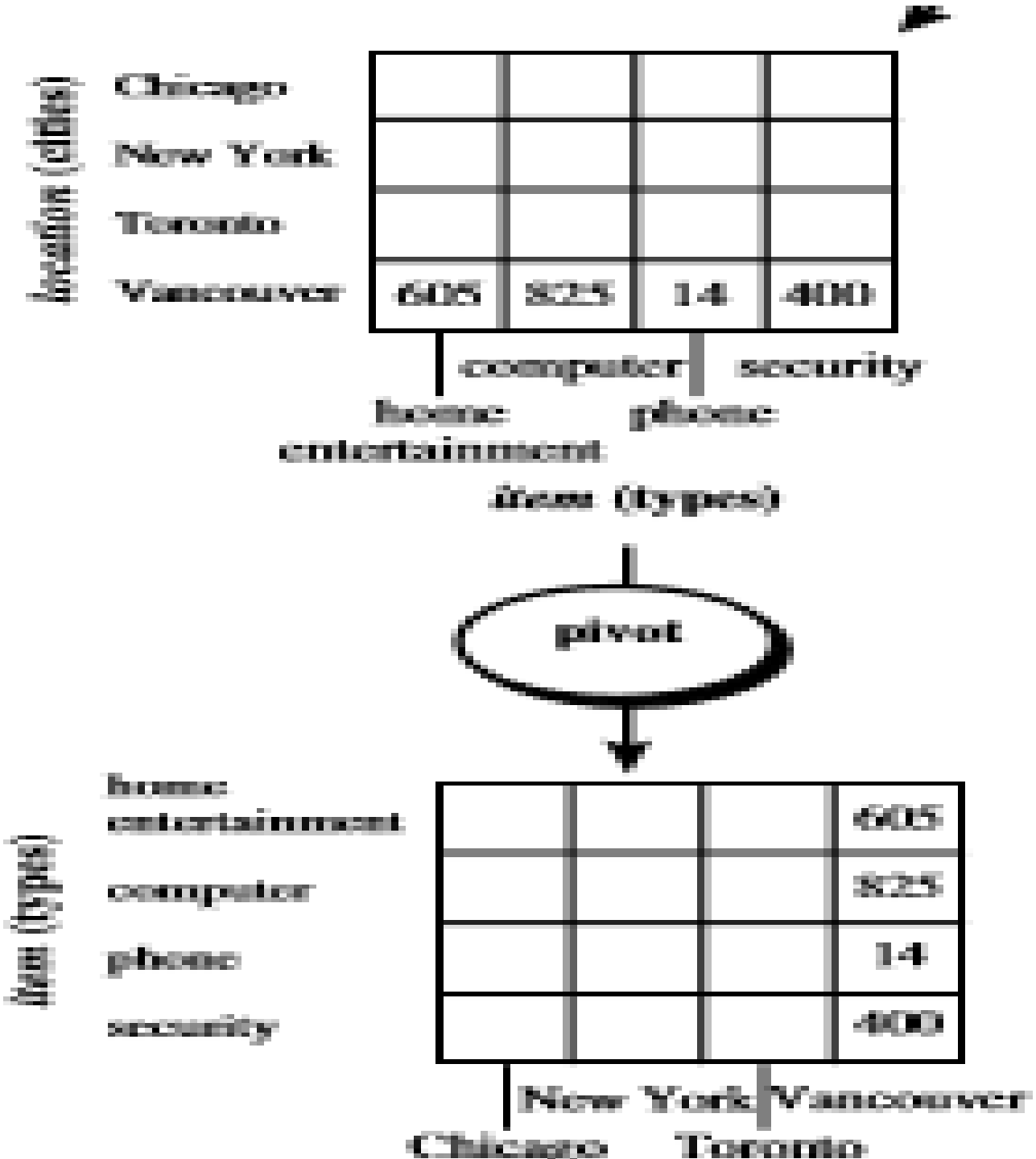
Dice:

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

Slice :

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




location (cities)
Chicago
New York
Toronto
Vancouver
605
825
14
400
computer
security
home entertainment
phone
item (types)
pivot
home entertainment
computer
phone
security
605
825
14
400
New York
Vancouver
Chicago
Toronto

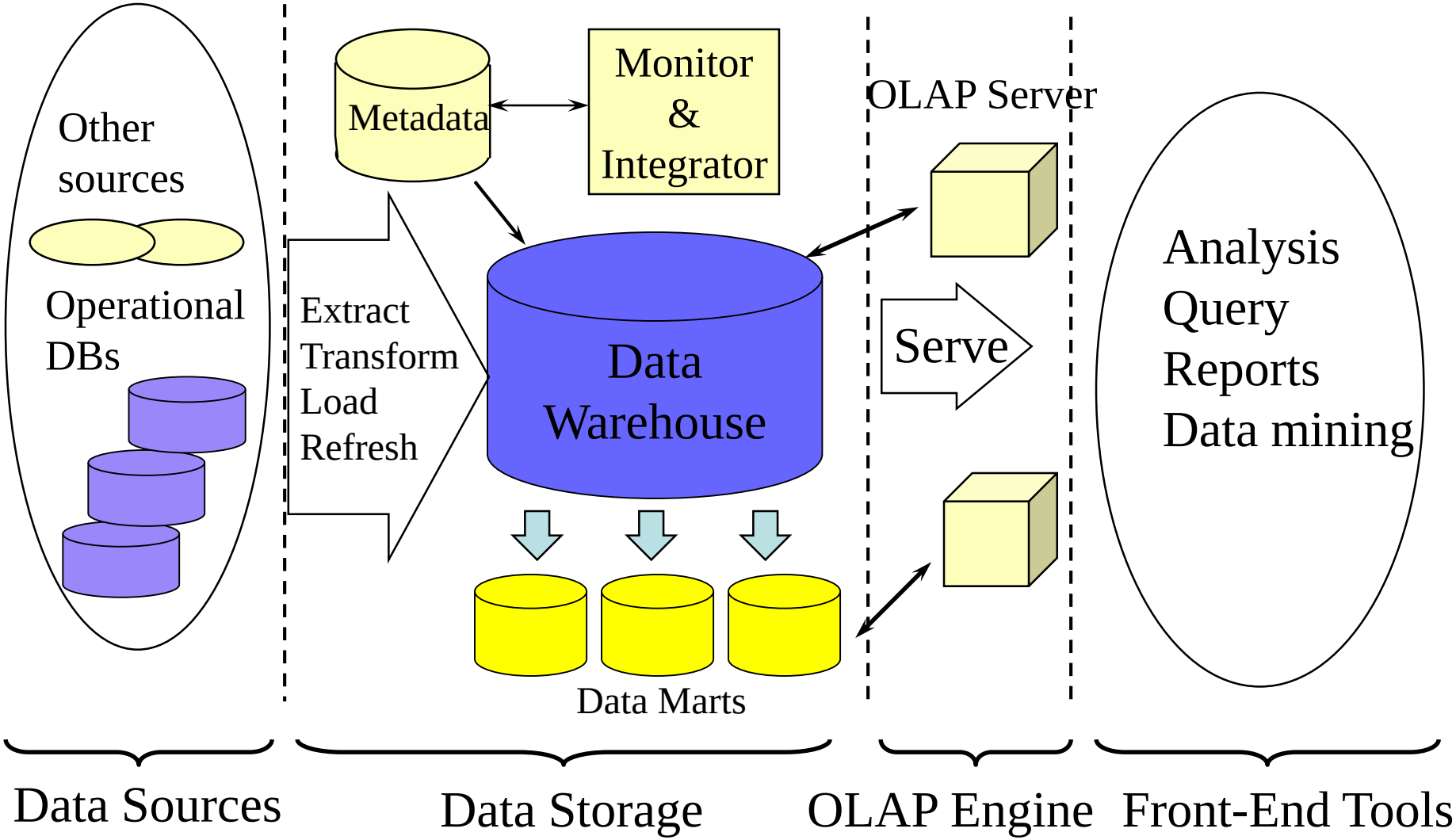
Types of OLAP Servers: ROLAP versus MOLAP versus HOLAP

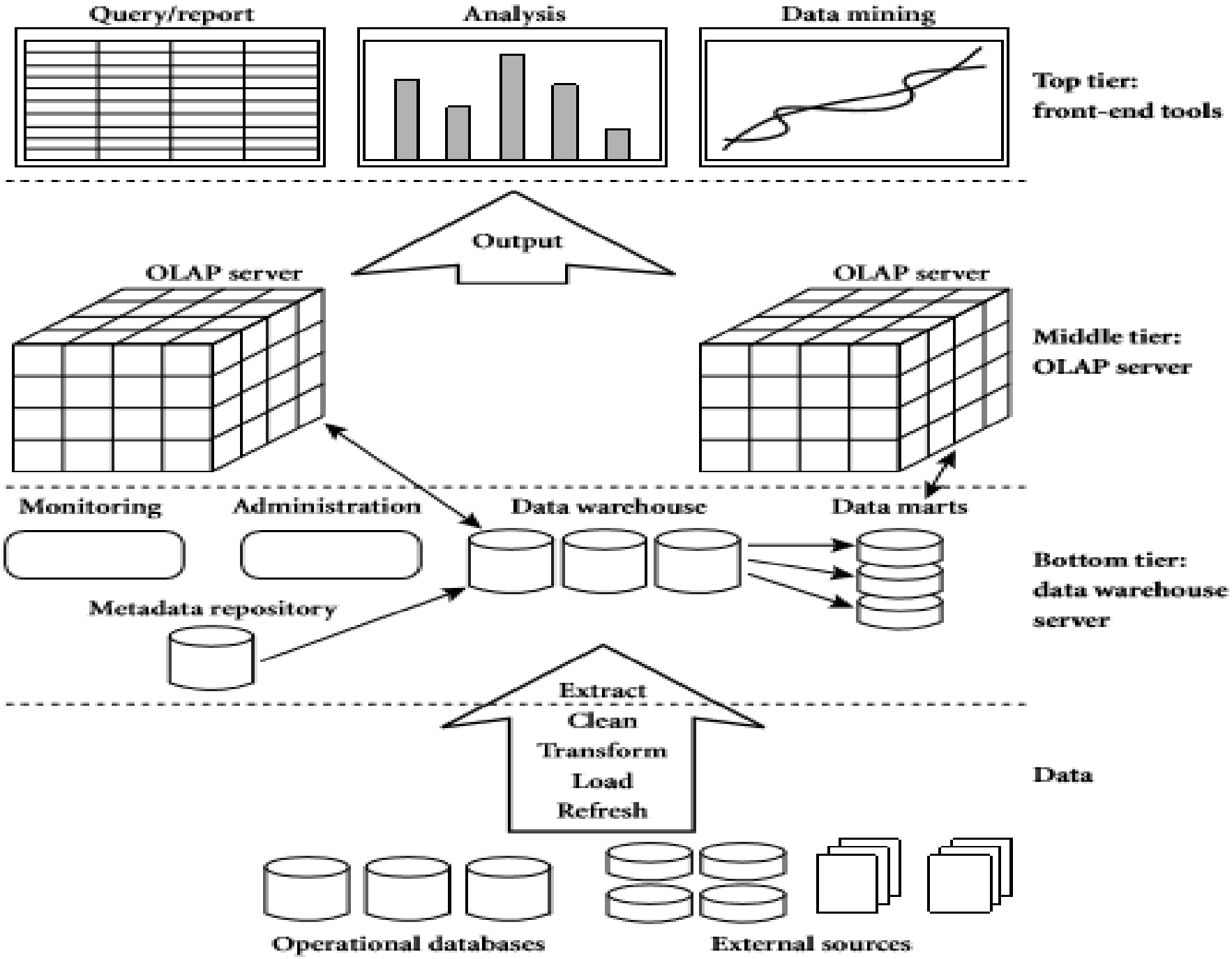
Relational OLAP (ROLAP) servers: These are the intermediate servers that stand in between a relational back-end server and client front-end tools. They use a relational or extended- relational DBMS to store and manage warehouse data, and OLAP middleware to support missing pieces.

Multidimensional OLAP(MOLAP)servers: These servers support multidimensional views of data through array-based multi dimensional storage engines. They map multi dimensional views directly to data cube array structures

Hybrid OLAP (HOLAP) servers: The hybrid OLAP approach combines ROLAP and MOLAP technology, benefiting from the greater scalability of ROLAP and the faster computation of MOLAP.

Data Warehouse: A Multi-Tiered Architecture




Query/report
Analysis
Data mining
Top tier:
front-end tools
Output
OLAP server
OLAP server
Middle tier:
OLAP server
Monitoring
Administration
Data warehouse
Data marts
Bottom tier:
data warehouse
server
Metadata repository
Extract
Clean
Transform
Load
Refresh
Data
Operational databases
External sources

A Three-Tier Data Warehouse Architecture

Data warehouses often adopt a three-tier architecture

1. **The bottom tier** is a warehouse database server that is almost always a relational database system. Back-end tools and utilities are used to feed data into the bottom tier from operational databases or other external sources.

These tools and utilities perform data extraction, cleaning, and transformation as well as load and refresh functions to update the data warehouse.

2. **The middle tier** is an OLAP server that is typically implemented using either

(1)a relational OLAP (ROLAP) model, that is, an extended relational DBMS that maps operations on multidimensional data to standard relational operations; or

(2) a multidimensional OLAP (MOLAP) model, that is, a special-purpose server that directly implements multidimensional data and operations.

The top tier is a front-end client layer, which contains query and reporting tools, analysis tools, and/or data mining tools (e.g., trend analysis, prediction, and so on).

From the architecture point of view, there are three data warehouse models:

the enterprise warehouse,

the data mart,

and the 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
    - 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

```
define cube sales_snowflake [time, item, branch, location]:
        dollars_sold = sum(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
        (supplier_key, supplier_type))
define dimension branch as (branch_key, branch_name, branch_type)
define dimension location as (location_key, street, city
        (city_key, city, province_or_state, country))
```

OLAP

LINE	TOTAL SALES
Clothing	$12,836,450
Electronics	$16,068,300
Video	$21,262,190
Kitchen	$17,704,400
Appliances	$19,600,800
Total	$87,472,140

1 High level summary by product line

2 Drill down by year

LINE	1998	1999	2000	TOTAL
Clothing	$3,457,000	$3,590,050	$5,789,400	$12,836,450
Electronics	$5,894,800	$4,078,900	$6,094,600	$16,068,300
Video	$7,198,700	$6,057,890	$8,005,600	$21,262,190
Kitchen	$4,875,400	$5,894,500	$6,934,500	$17,704,400
Appliances	$5,947,300	$6,104,500	$7,549,000	$19,600,800
Total	$27,373,200	$25,725,840	$34,373,100	$87,472,140

3 Rotate columns to rows

YEAR	Clothing	Electronics	Video	Kitchen	Appliances	TOTAL
1998	$3,457,000	$5,894,800	$7,198,700	$4,875,400	$5,947,300	$27,373,200
1999	$3,590,050	$4,078,900	$6,057,890	$5,894,500	$6,104,500	$25,725,840
2000	$5,789,400	$6,094,600	$8,005,600	$6,934,500	$7,549,000	$34,373,100
Total	$12,836,450	$16,068,300	$21,262,190	$17,704,400	$19,600,800	$87,472,140

Figure 15-3 Simple OLAP session.

The term OLAP was introduced in a paper entitled “ providing On-Line Analytical Processing to User Analysts, “ by Dr. E.F. Codd. The paper published in 1993.

Online analytical processing is a category of software technology that enables an analysts, managers and executives to gain insight into data through fast, consistent, interactive access in a wide variety of possible views of information that has been transformed from raw data to reflect the real dimensionality of the enterprise as understood by the user

OLAP Guidelines

Multidimensional Conceptual View

OLAP system should provide multidimensional conceptual view of the data. This view should be based on the needs of the user and not on the physical data storage.

Transparency
System should be transparent to the user. But data abstraction property should be preserved.

Accessibility
System should provide easy access to the data. Users can access the data through user-friendly interface.

Consistent Reporting Performance
System should provide consistent reporting performance regardless of the complexity of the query or the amount of data being analyzed.

Consistent Reporting Performance
System should provide consistent reporting performance regardless of the complexity of the query or the amount of data being analyzed.

Client-Server Architecture
System should be based on client-server architecture. It has multiple users to access the system at the same time.

Generic Dimensionality
The system should support generic dimensionality. System can handle any number of dimensions and any type of data.

Dynamic Sparse Matrix Handling
System should be able to handle dynamic sparse matrices. System can handle data that is not regularly populated.

Multi-User Support
System should support multi-user access. Multiple users can access and analyze data at the same time.

Unrestricted Cross-Dimensional Operations
System should allow unrestricted cross-dimensional operations. System should allow users to analyze data from different dimensions without restrictions.

Intuitive Data Manipulation
System should provide intuitive data manipulation tools. Users can manipulate and analyze data in a user-friendly manner.

Flexible Reporting
System should provide flexible reporting capabilities. Users can generate reports in various formats and with different levels of detail.

Unlimited Dimensions and Aggregation Levels
System should support unlimited dimensions and aggregation levels. System should can handle any number of dimensions and any level of aggregation

Multidimensional versus Multi relational OLAP

In the MOLAP model, online analytical processing is best implemented by storing the data multidimensional that is easily viewed in a multidimensional way. Here the data structure is fixed so that the logic to process multidimensional analysis ca be based on well defined methods of establishing data storage coordinates. Usually, multidimensional databases (MDDBs) are vendors proprietary systems.

On the other hand , the ROLAP model relies on the existing relational DBMS of data warehouse. OLAP Features are provided against the relational database.

Multidimensional vs. Multirelational OLAP:

MOLAP (Multidimensional OLAP): Stores data in pre-aggregated "cubes" optimized for fast query response, ideal for complex analysis with pre-defined dimensions and hierarchies.

ROLAP (Relational OLAP): Utilizes standard relational databases and SQL queries, offering greater flexibility for ad-hoc analysis but potentially slower performance for complex queries.

As mentioned earlier, multidimensional database management systems are proprietary software systems. These systems provide the capability to consolidate and fabricate summarized cubes during the process that loads data into the MDDBs from the main data warehouse. The users who need summarized data enjoy fast response times from the preconsolidated data.

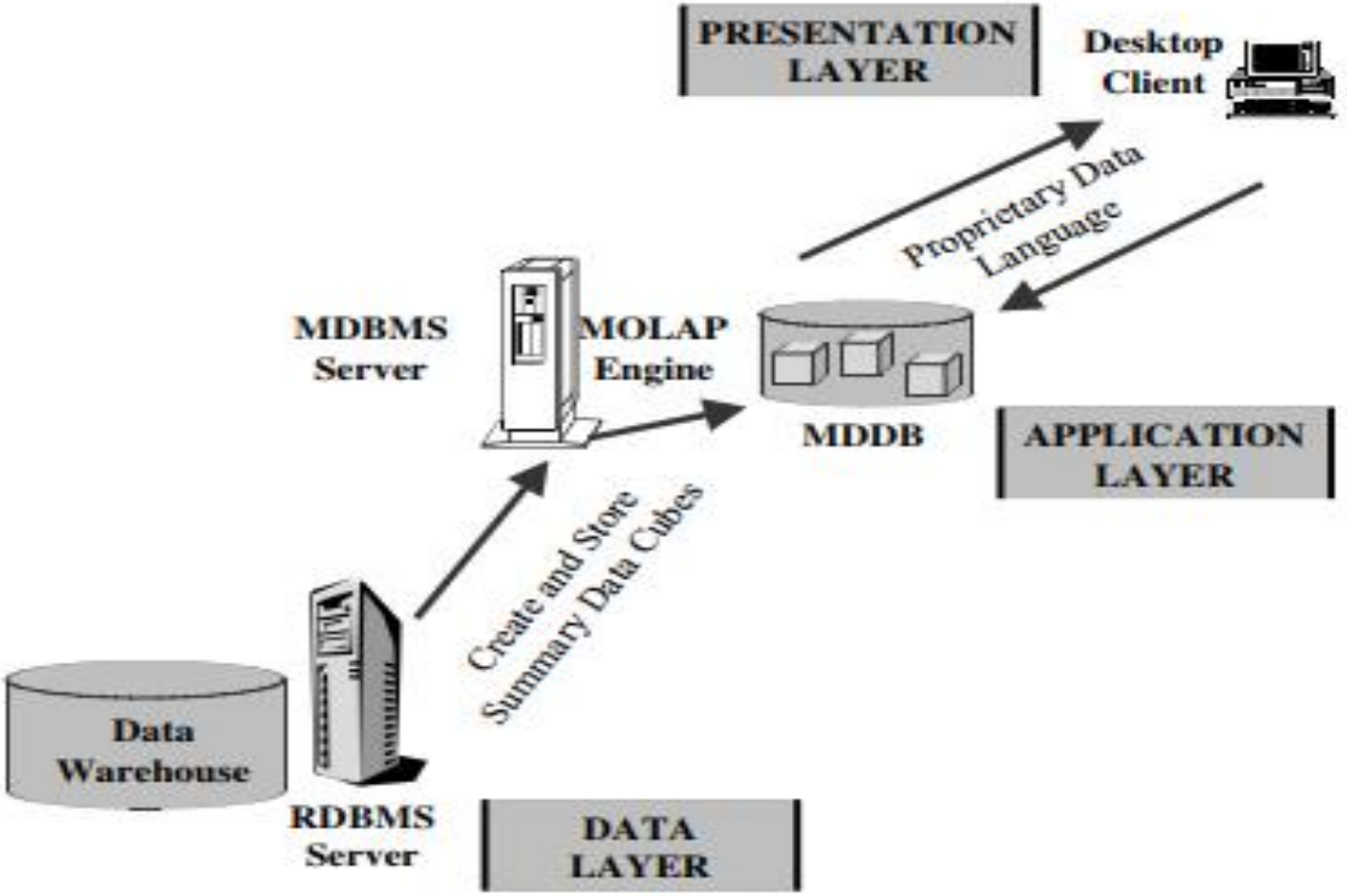


Figure 15-16 The MOLAP model.

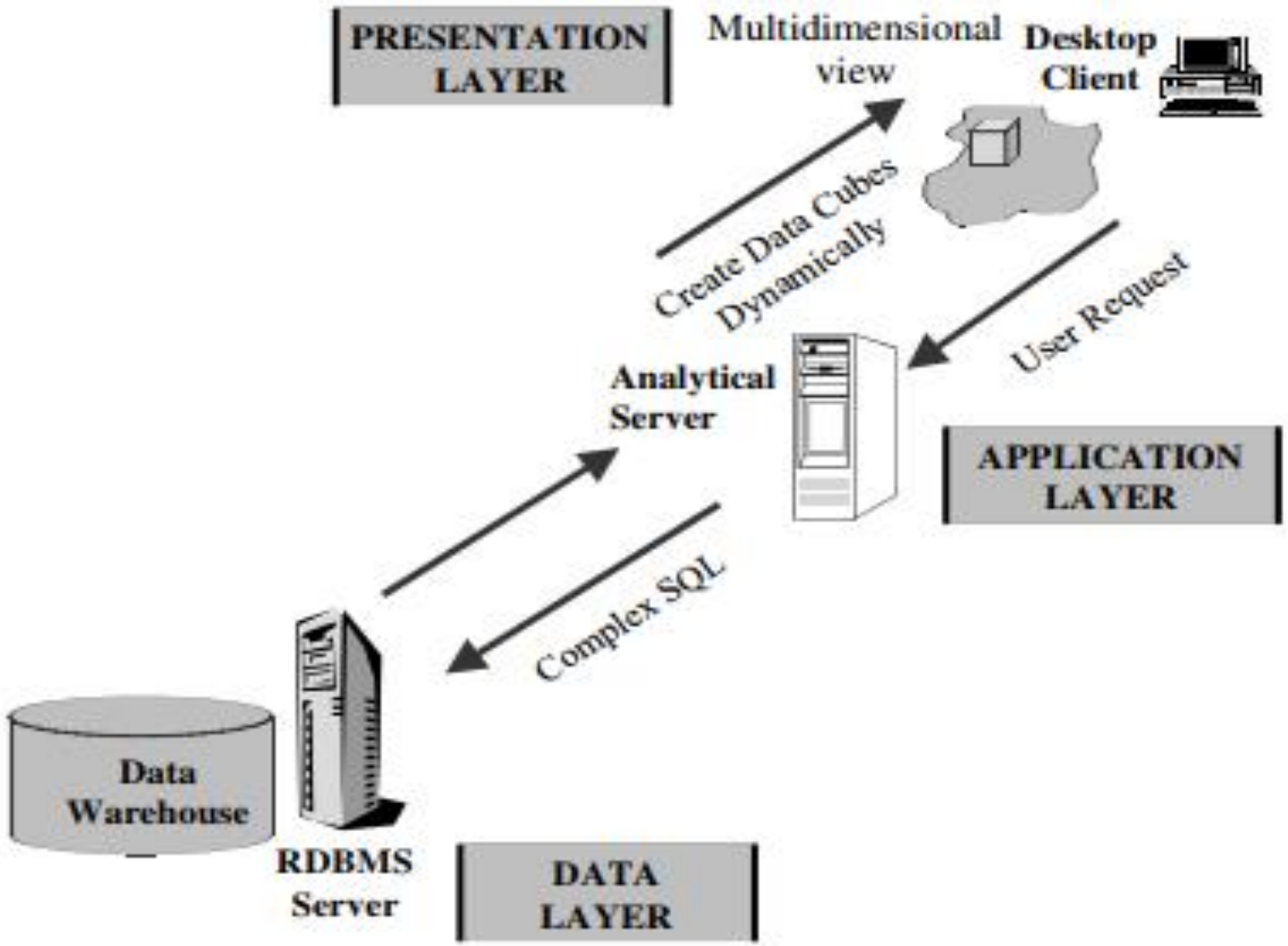


Figure 15-17 The ROLAP model.

ROLAP	MOLAP
ROLAP stands for **Relational Online Analytical Processing.**	While MOLAP stands for **Multidimensional Online Analytical Processing.**
ROLAP is used for large data volumes.	While it is used for limited data volumes.
The access of ROLAP is slow.	While the access of MOLAP is fast.
In ROLAP, Data is stored in relation tables.	While in MOLAP, Data is stored in multidimensional array.
In ROLAP, Data is fetched from data-warehouse.	While in MOLAP, Data is fetched from MDDBs database.
In ROLAP, Complicated sql queries are used.	While in MOLAP, Sparse matrix is used.
In ROLAP, Static multidimensional view of data is created.	While in MOLAP, Dynamic multidimensional view of data is created.

	Data Storage	Underlying Technologies	Functions and Features
ROLAP	Data stored as relational tables in the warehouse. Detailed and light summary data available. Very large data volumes. All data access from the warehouse storage.	Use of complex SQL to fetch data from warehouse. ROLAP engine in analytical server creates data cubes on the fly. Multidimensional views by presentation layer.	Known environment and availability of many tools. Limitations on complex analysis functions. Drill-through to lowest level easier. Drill-across not always easy.
MOLAP	Data stored as relational tables in the warehouse. Various summary data kept in proprietary databases (MDDBs) Moderate data volumes. Summary data access from MDDB, detailed data access from warehouse.	Creation of pre-fabricated data cubes by MOLAP engine. Proprietary technology to store multidimensional views in arrays, not tables. High speed matrix data retrieval. Sparse matrix technology to manage data sparsity in summaries.	Faster access. Large library of functions for complex calculations. Easy analysis irrespective of the number of dimensions. Extensive drill-down and slice-and-dice capabilities.

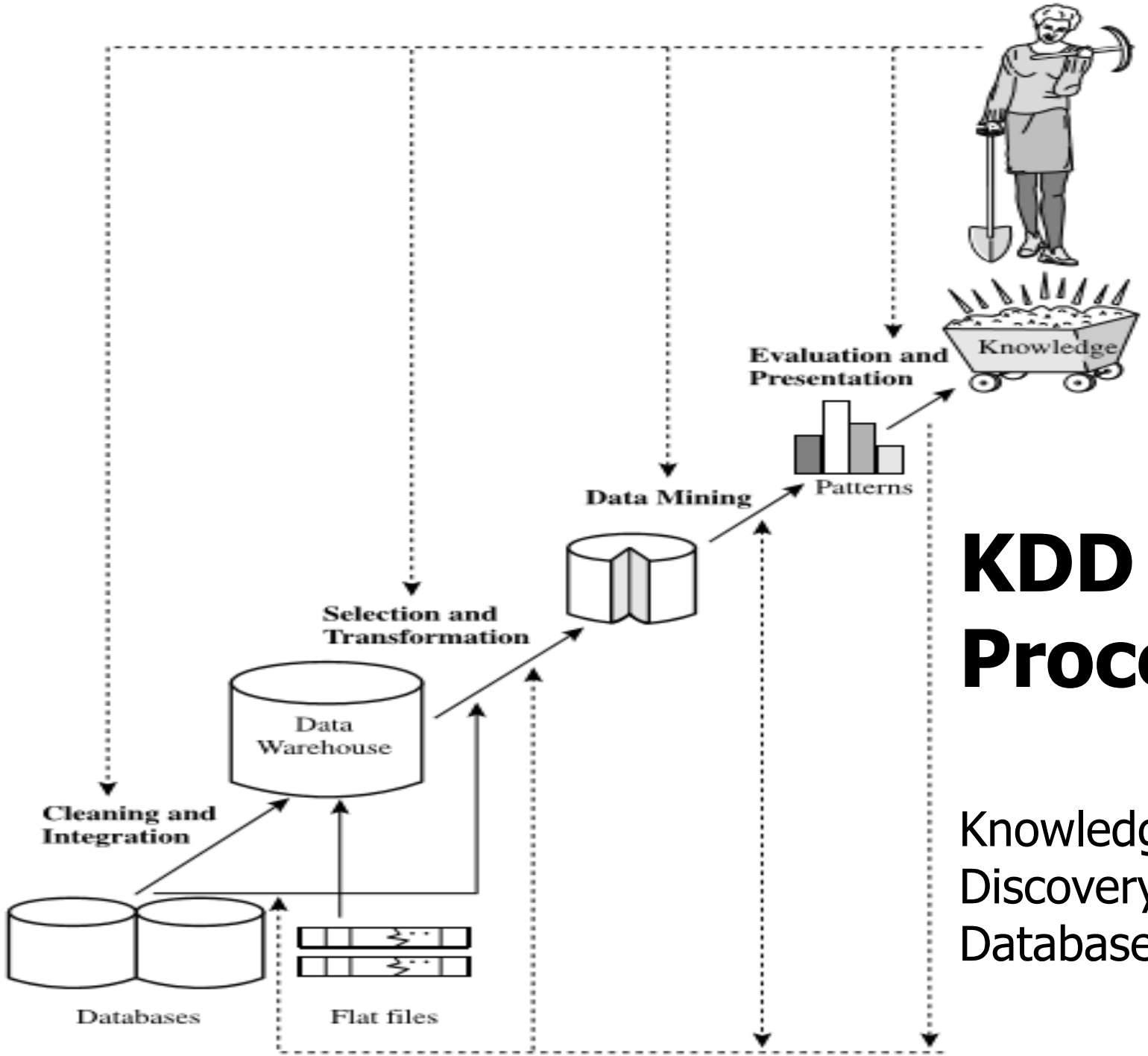
Figure 15-19 ROLAP versus MOLAP.

- **Data mining (knowledge discovery from data)**

  Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data

- **Alternative names**

  Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.



KDD Process

Knowledge Discovery in Databases

Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery from Data.

Steps in KDD

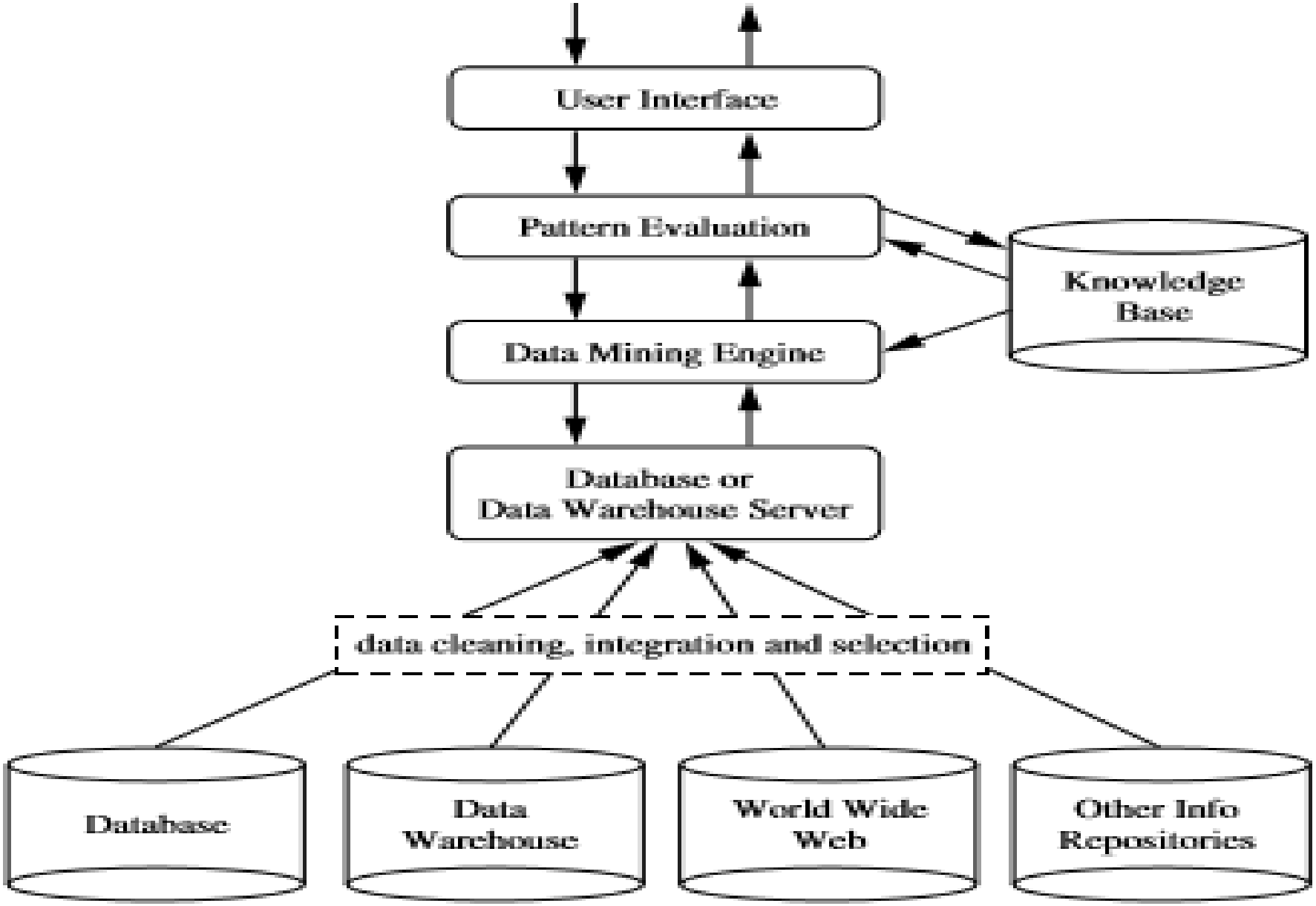
- **Data cleaning** (to remove noise and inconsistent data).
- **Data integration** (where multiple data sources may be combined)
- **Data selection** (where data relevant to the analysis task are retrieved from the database)
- **Data transformation** (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance).

- **Data mining** (an essential process where intelligent methods are applied in order to extract data patterns).

- **Pattern evaluation** (to identify the truly interesting patterns representing knowledge based on some interestingness measures.

- **Knowledge presentation** (where visualization and knowledge representation techniques are used to present the mined knowledge to the user).

Architecture of a Data Mining System



The architecture of a typical data mining system may have the following major components

Database, data warehouse, World Wide Web, or other information repository: This is one or a set of databases, data warehouses, spread sheets, or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the data.

Database or data warehouse server: The database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request.

Knowledge base: This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns.

Such knowledge can include concept hierarchies, used to organize attributes or attribute values into different levels of abstraction.

Knowledge such as user beliefs, which can be used to assess a pattern's interestingness based on its unexpectedness, may also be included.

Data mining engine: This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis.

Pattern evaluation module: This component typically employs interestingness measures and interacts with the data mining modules so as to focus the search toward interesting patterns. It may use interestingness thresholds to filter out discovered patterns.

User interface: This module communicates between users and the datamining system, allowing the user to interact with the system by specifying a data mining query or task