Introduction to Data Warehousing

Data Warehouse

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

- Data warehouse refers to a database that is maintained separately from an organization's operational databases.

"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 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`

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*

Organizations use the information from Data Warehouses

1. **Increasing customer focus, which includes the analysis of customer buying patterns**

   such as buying preference, buying time, budget cycles, and appetites for spending

2. Repositioning products and managing product portfolios by comparing the performance of sales by quarter, by year, and by geographic regions

3. Analyzing operations and looking for sources of profit

4. Managing the customer relationships, making environmental corrections, and managing the cost of corporate assets.

Differences between 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.

- Covers most of the day-to-day operations of an organization.

  Ex: purchasing, manufacturing, banking, payroll, registration, and accounting

- In Data warehouse systems, 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.

Difference between OLTP and OLAP

1. **Users and system orientation:**

- OLTP system is customer-oriented and is used for transaction and query processing by clerks, clients, and information technology professionals.

- OLAP system is *market-oriented and is used for data analysis by knowledge* workers, including managers, executives, and analysts.

2. Data contents:

- OLTP system manages current data that, typically, are too detailed to be easily used for decision making.

- OLAP system manages large amounts of historical data, provides facilities for summarization and aggregation, and stores and manages information at different levels of granularity.

3. Database Design:

- OLTP system usually adopts an entity-relationship (ER) data model and an application-oriented database design.

-OLAP system typically adopts either a star or snowflake model and a subject oriented database design.

4. View:

- An OLTP system focuses mainly on the current data within an enterprise or department, without referring to historical data or data in different organizations.

- OLAP systems deal with information that originates from different organizations, integrating information from many data stores.

5. Access patterns:

- OLTP system consists mainly of short, atomic transactions.

  Such a system requires concurrency control and recovery mechanisms.

- Accesses to OLAP systems are mostly read-only operations.

- Data warehouses store historical rather than up-to-date information.

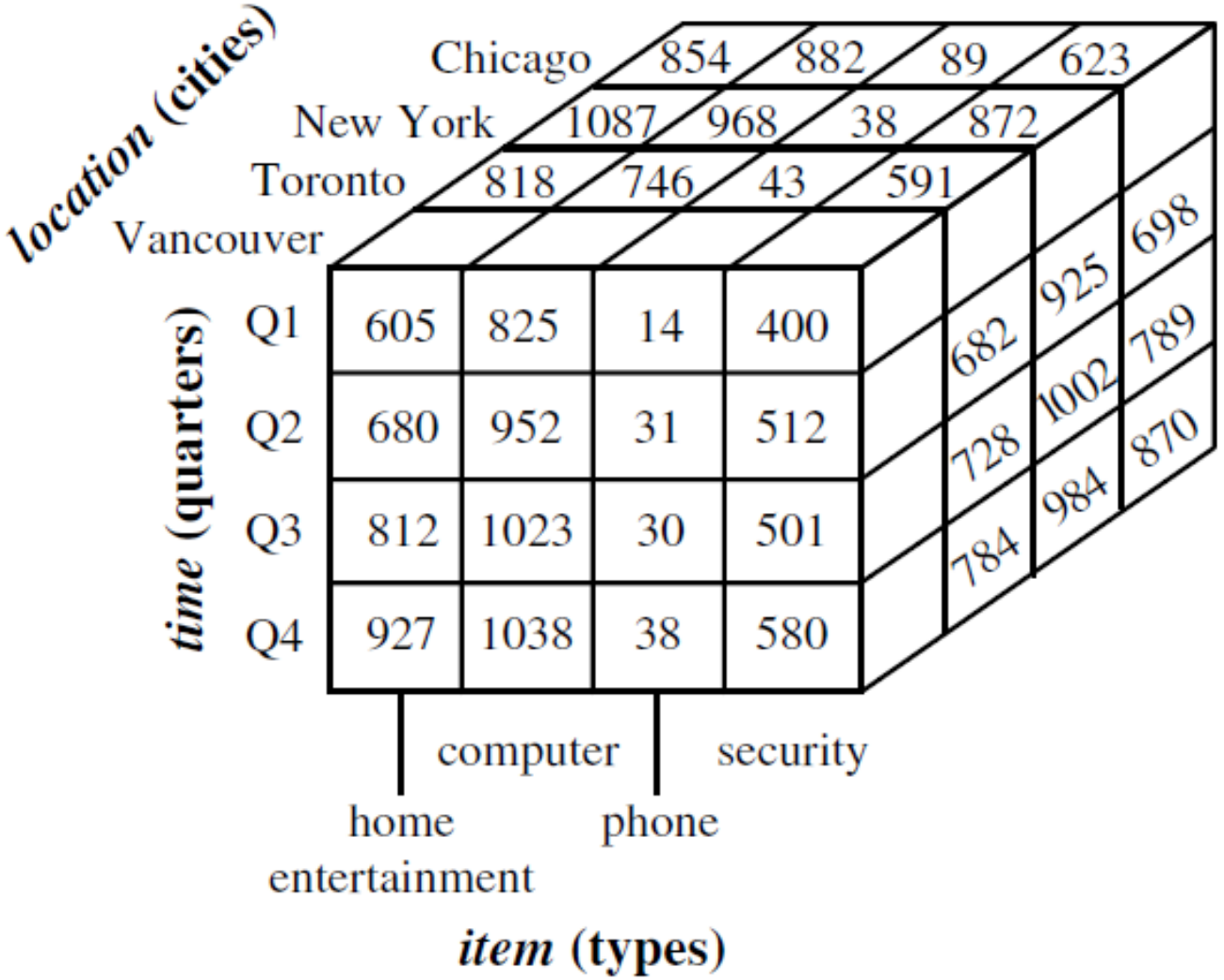
Data Model

A 2-D view of sales data for *AllElectronics* according to the dimensions *time* and *item*, where the sales are from branches located in the city of Vancouver. The measure displayed is *dollars_sold* (in thousands).

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

Table 3.3 A 3-D view of sales data for *AllElectronics*, according to the dimensions *time*, *item*, and *location*. The measure displayed is *dollars_sold* (in thousands).

	location = *"Chicago"*				***location*** = *"New York"*				***location*** = *"Toronto"*				***location*** = *"Vancouver"*			
	item				***item***				***item***				***item***			
	home				*home*				*home*				*home*			
time	*ent.*	*comp.*	*phone*	*sec.*	*ent.*	*comp.*	*phone*	*sec.*	*ent.*	*comp.*	*phone*	*sec.*	*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


location (cities)
Chicago
854
882
89
623
New York
1087
968
38
872
Toronto
818
746
43
591
Vancouver
time (quarters)
Q1
605
825
14
400
Q2
680
952
31
512
Q3
812
1023
30
501
Q4
927
1038
38
580
682
925
698
728
1002
789
784
984
870
computer
security
home entertainment
phone
item (types)

Data Cube

Stars, Snowflakes, and Fact Constellations: Schemas for Multidimensional Databases

- **Schema** gives a logical description of entire database.

  It gives details about how key values are linked with different tables.

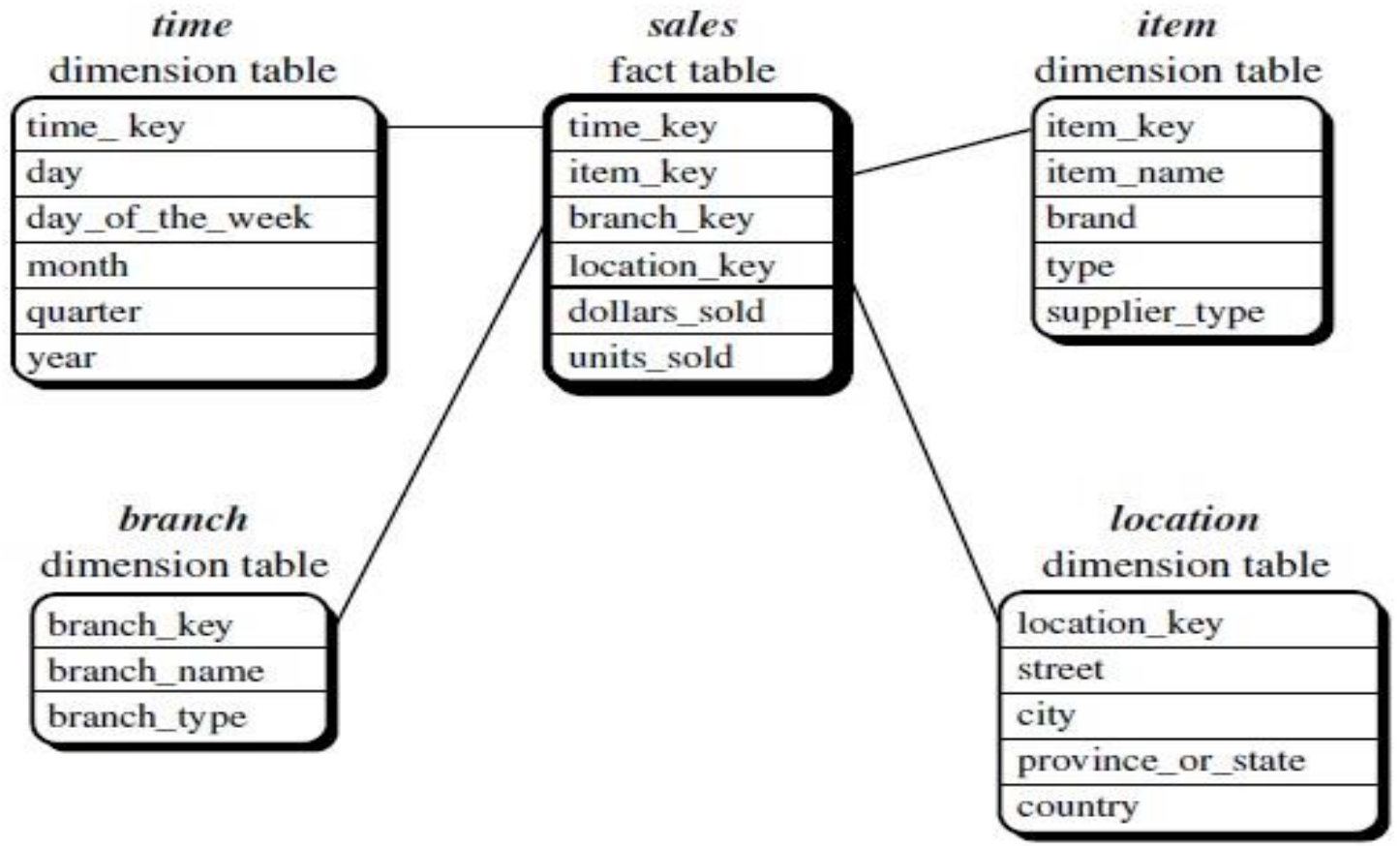
- A data warehouse requires a concise, subject-oriented schema that facilitates on-line data analysis, such as **multidimensional model.**

- **A model in Data Warehouse can exist in the any form of**

  - Star schema,
  - Snowflake schema
  - Fact constellation schema

Star schema:

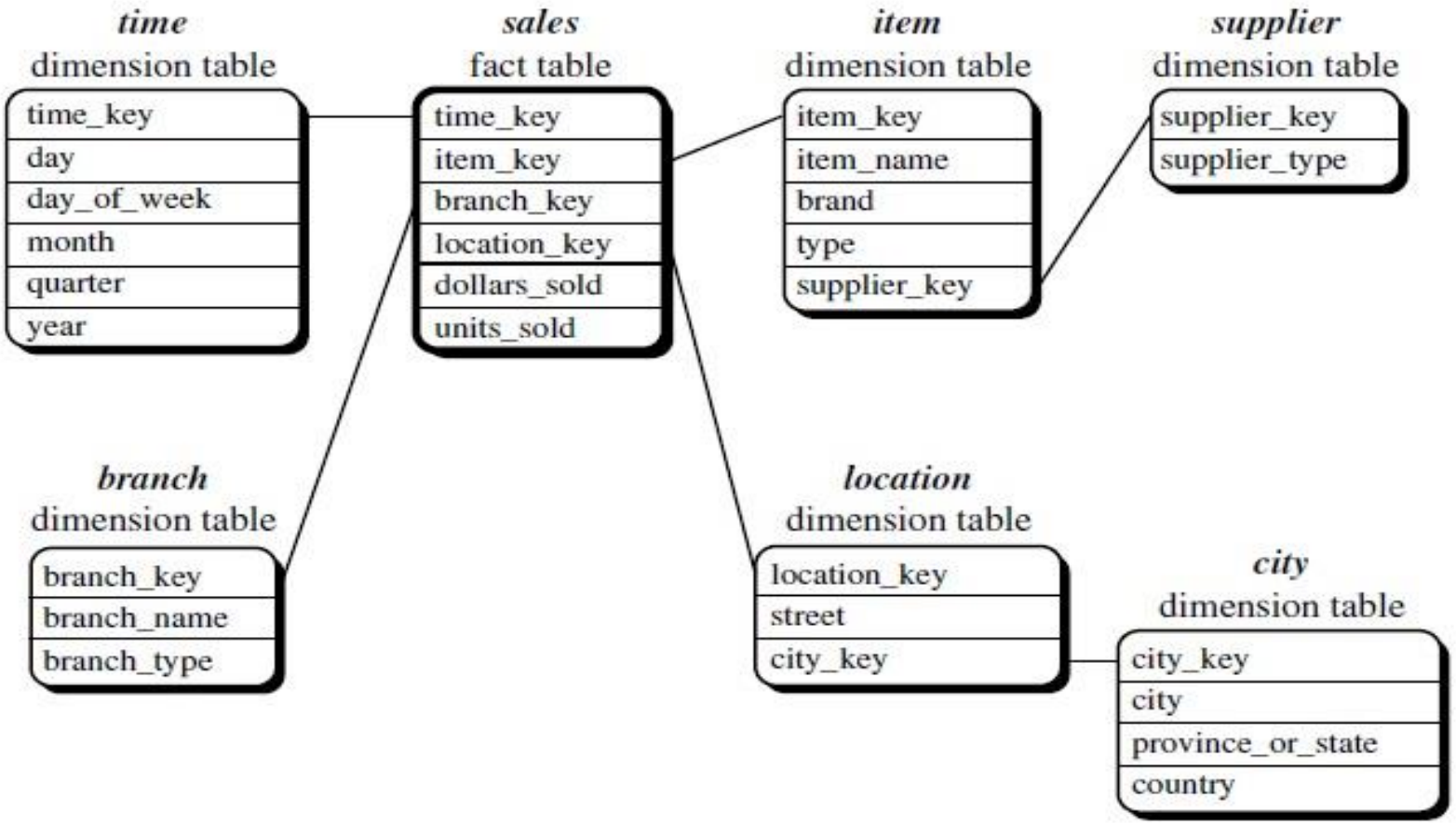
The star schema, in which the data warehouse contains **a large central table (fact table)** containing the bulk of the data, with no redundancy, and **a set of smaller attendant tables (dimension tables)**, one for each dimension.



Star schema of a data warehouse for sales.

Snowflake schema:

The snowflake schema is a variant of the star schema model, where some **dimension tables are normalized**, thereby further splitting the data into additional tables.



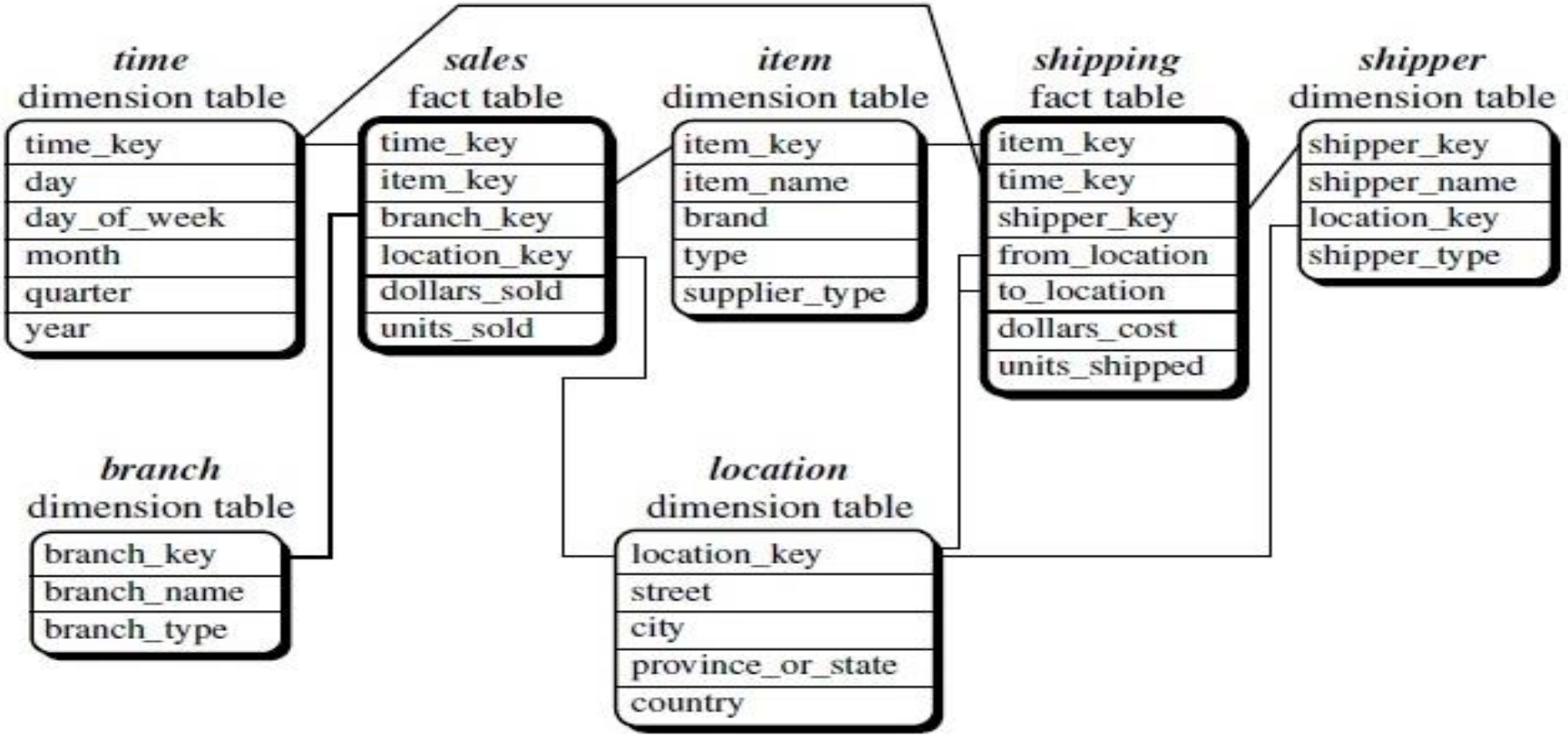
Snowflake schema of a data warehouse for sales.

Difference between the snowflake and star schema models:

	Star Schema	Snowflake Schema
1	Not Normalized	Normalized
2	Data Dependency and Redundancy	Less Data Dependency and Redundancy
3	No need of Join	More Joins are required to execute query

Fact constellation:

Multiple fact tables to share **dimension tables**. This schema can be viewed as a collection of stars, and it is called **a galaxy schema or a fact constellation.**



Fact constellation schema of a data warehouse for sales and shipping.

In data warehousing,

- A data warehouse collects information about subjects that span the entire organization, such as customers, items, sales, assets, and personnel, and thus its scope is **enterprise-wide.**

- So, the **fact constellation schema** is commonly used, since it can model multiple, interrelated subjects.

- A **data mart** is a department subset of the data warehouse that focuses on selected subjects, and thus its scope is **department wide.**

- So, **star or snowflake schema** are commonly used, although the star schema is more popular and efficient.

Measures: Their Categorization and Computation

- A data cube measure is a **numerical function** that can be evaluated at each point in the data cube space.

- A measure value is computed for a given point by **aggregating the data corresponding** to the respective dimension-value pairs defining the given point.

1. Distributive

- An aggregate function is *distributive if it can be computed in a distributed* manner.

- If the data are partitioned into *n sets and we apply the function* to each partition, resulting in *n aggregate values.*

- *If the result derived by applying the* function to the *n aggregate values is the same as that derived by applying the function* to the entire data set (without partitioning), then the function can be computed in distributive manner.

- **Ex: sum(), count(), max() , min()**

2. Algebraic:

An aggregate function is *algebraic if it can be* ***computed by an algebraic*** **function with** ***M arguments (where M is a bounded positive integer),*** *each of which* is obtained by applying a **distributive** aggregate function.

Ex: avg(), max_n(), min_n()

Avg() = sum()/count()

3. Holistic:

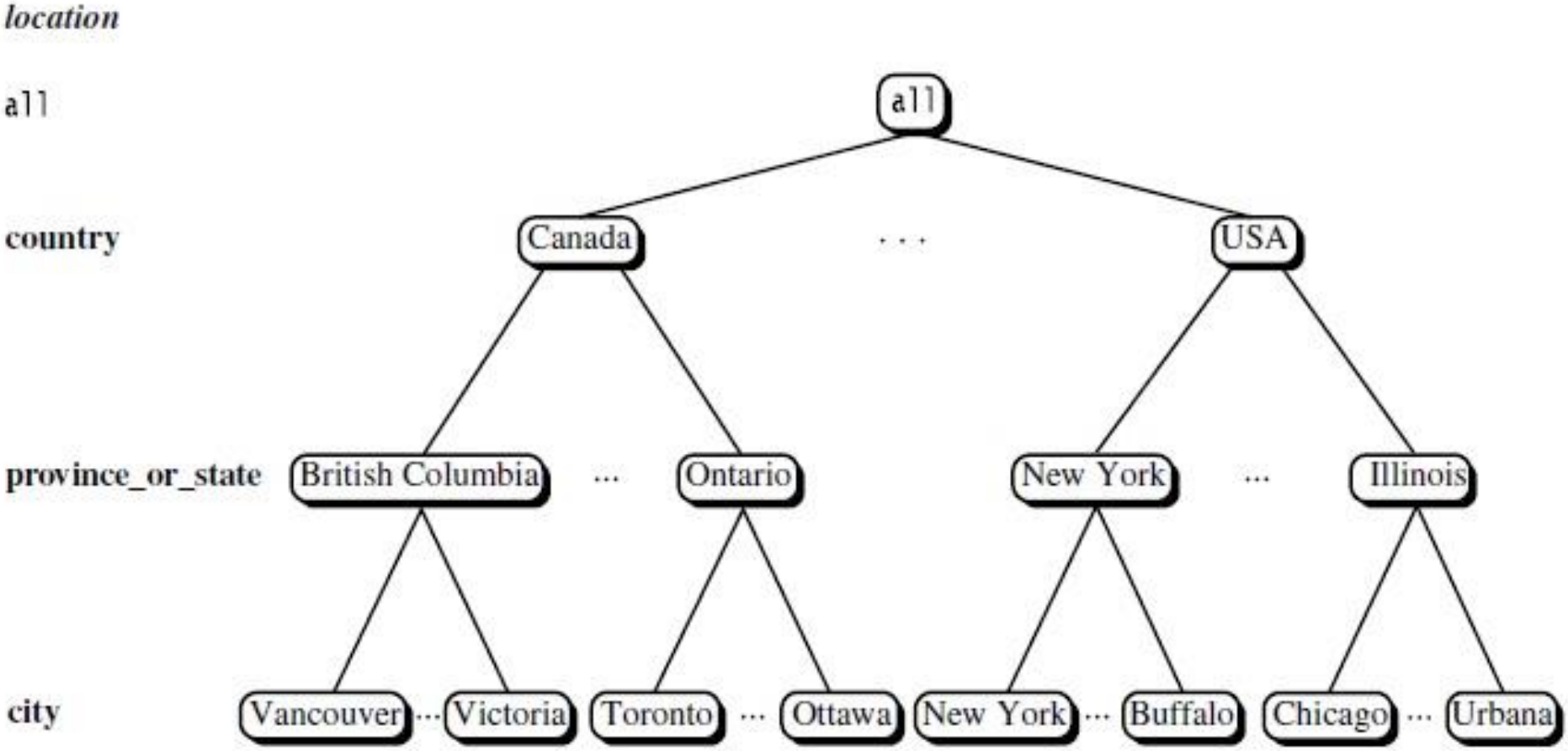
- An aggregate function is *holistic if there is no constant bound on the storage* size needed to describe a subaggregate.

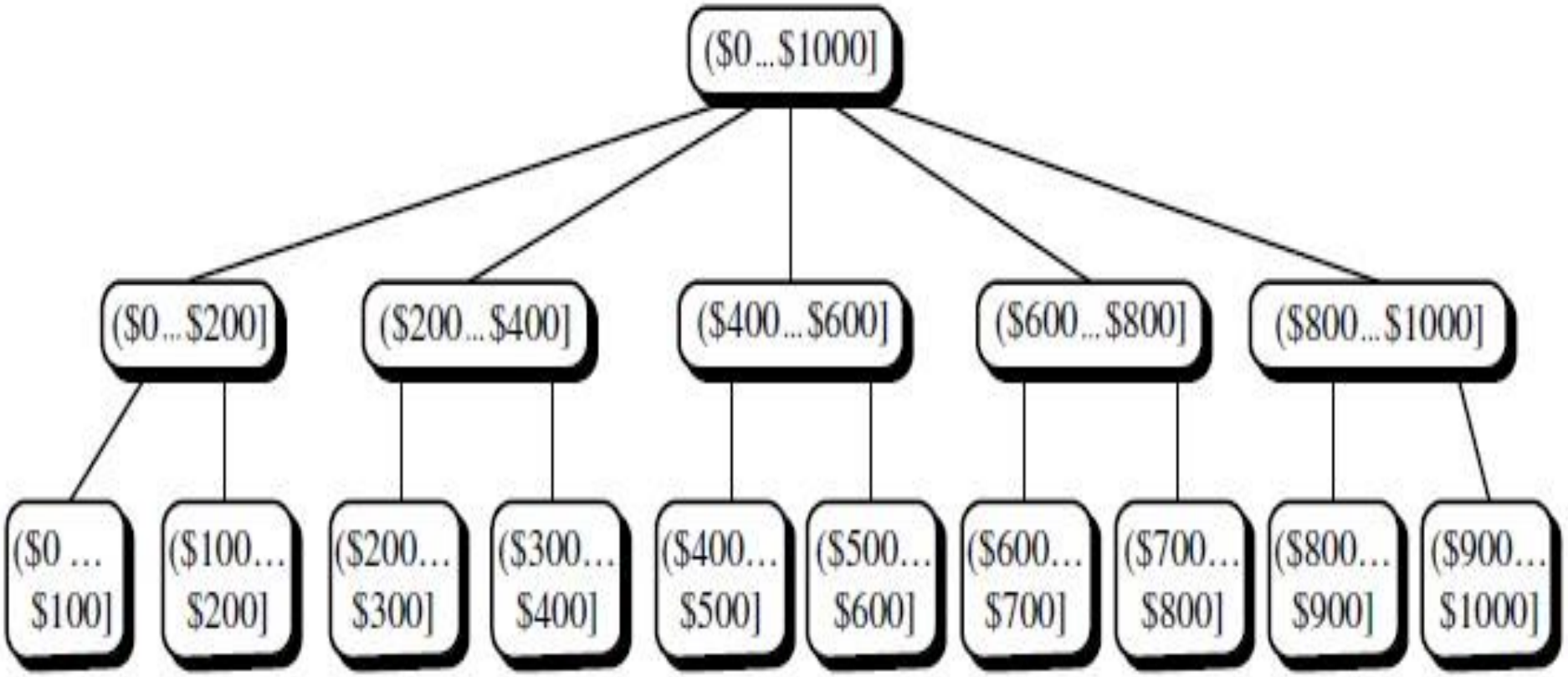
- There does not exist an algebraic function with *M arguments (where M is a constant) that characterizes the computation.*

Ex: median(), mode()

Concept Hierarchies

- A concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher-level, more general concepts.

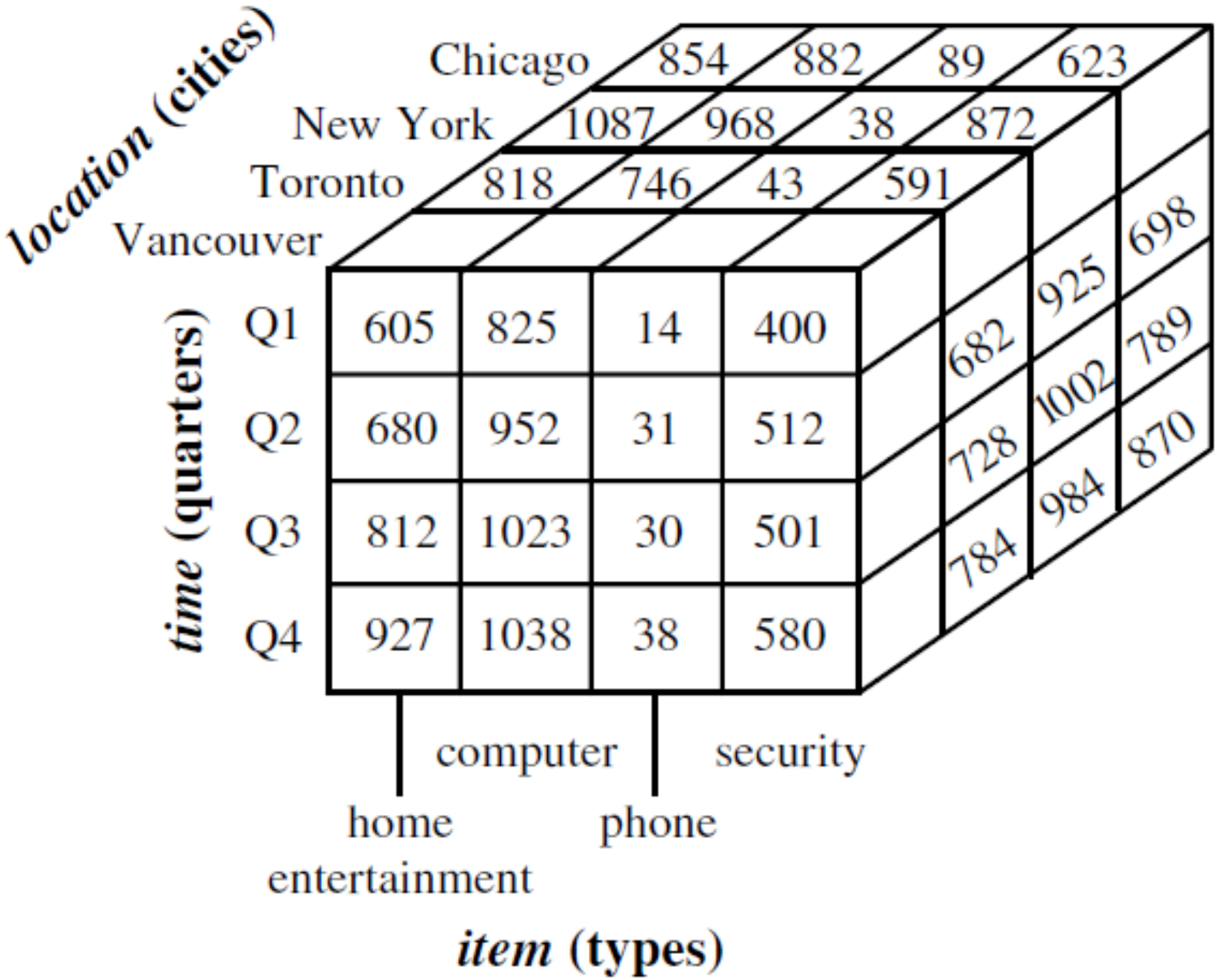

location
all
country
province_or_state
city
all
Canada
...
USA
British Columbia
...
Ontario
New York
...
Illinois
Vancouver
...
Victoria
Toronto
...
Ottawa
New York
...
Buffalo
Chicago
...
Urbana

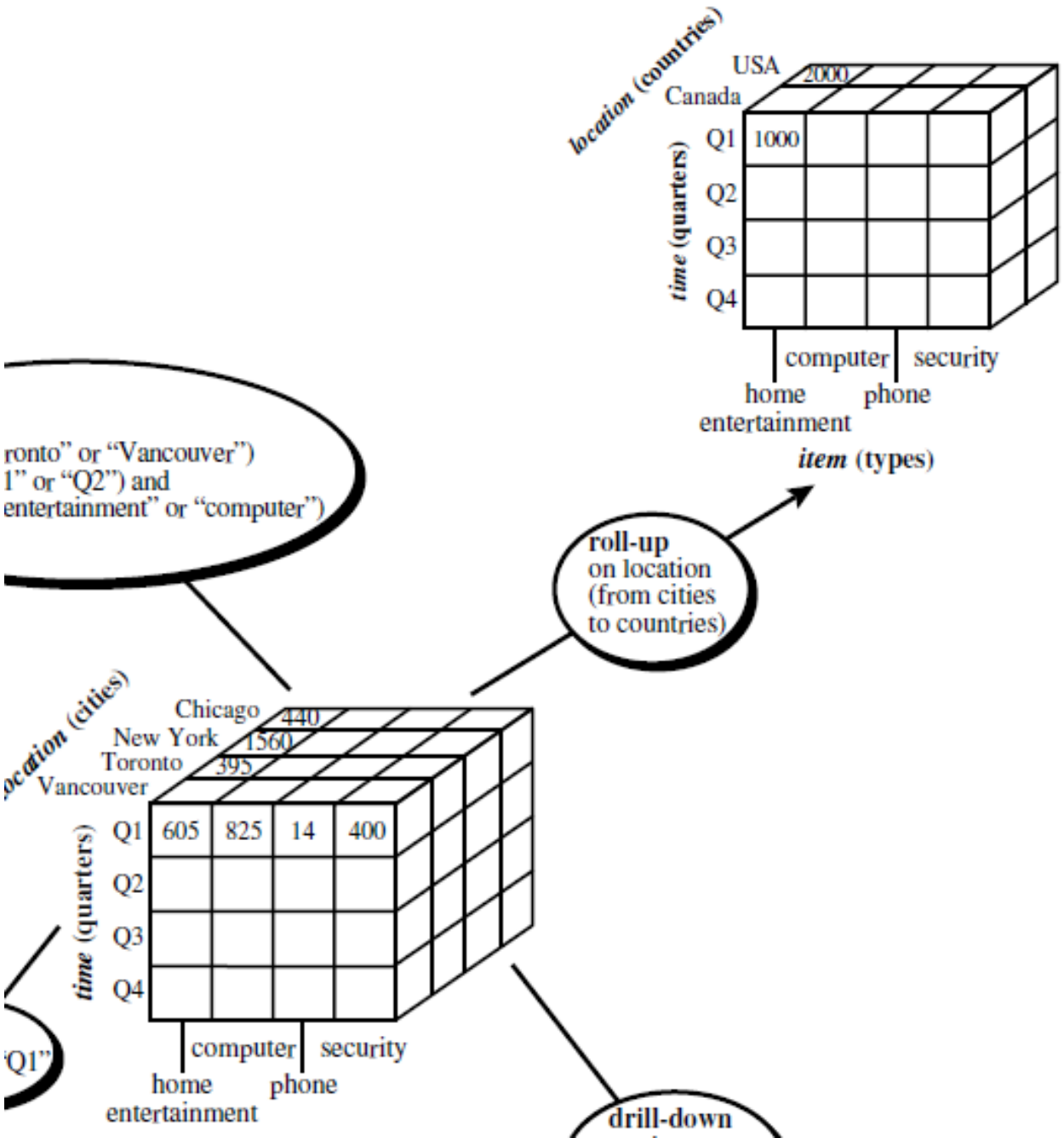


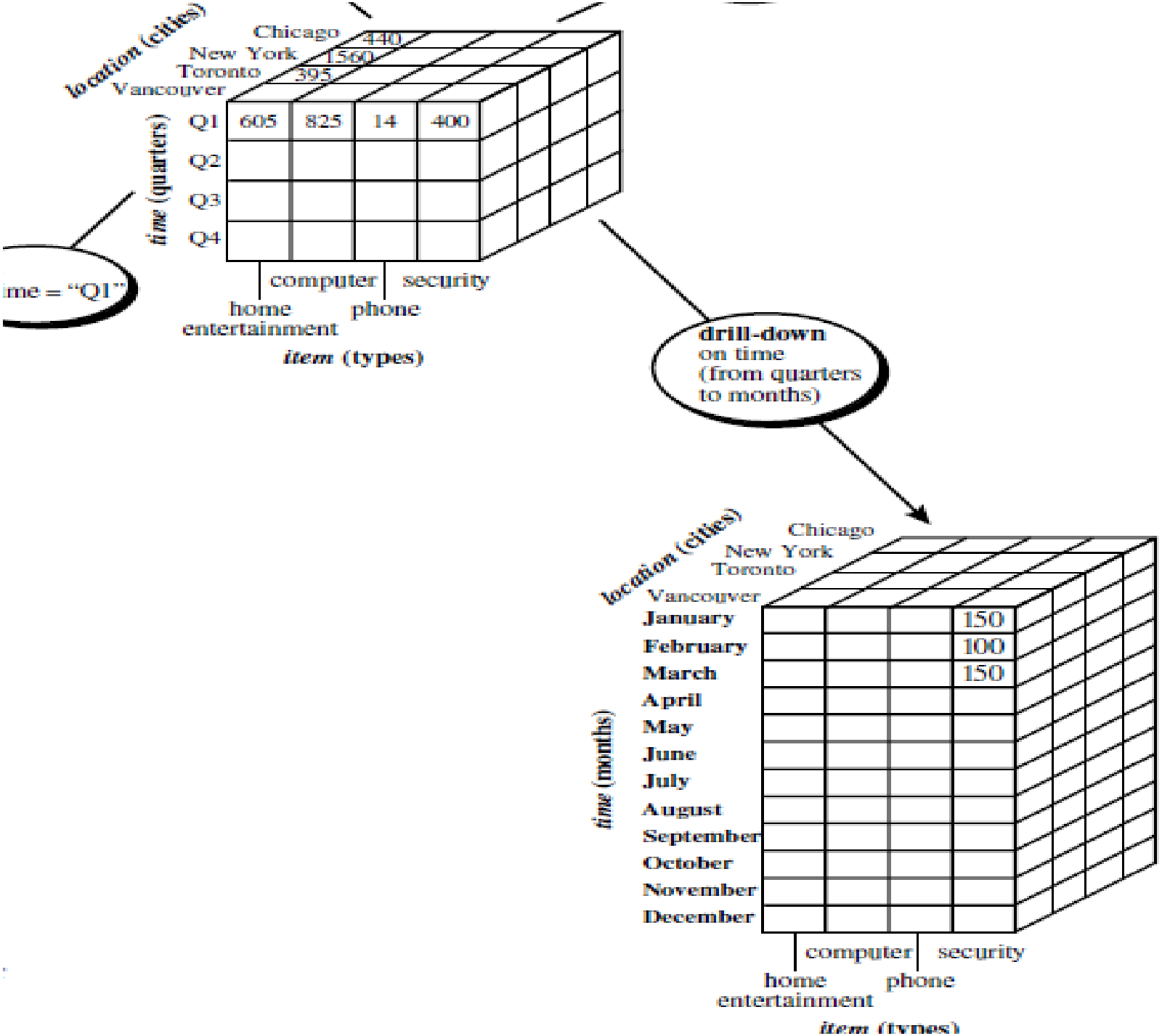
A concept hierarchy for the attribute *price*.

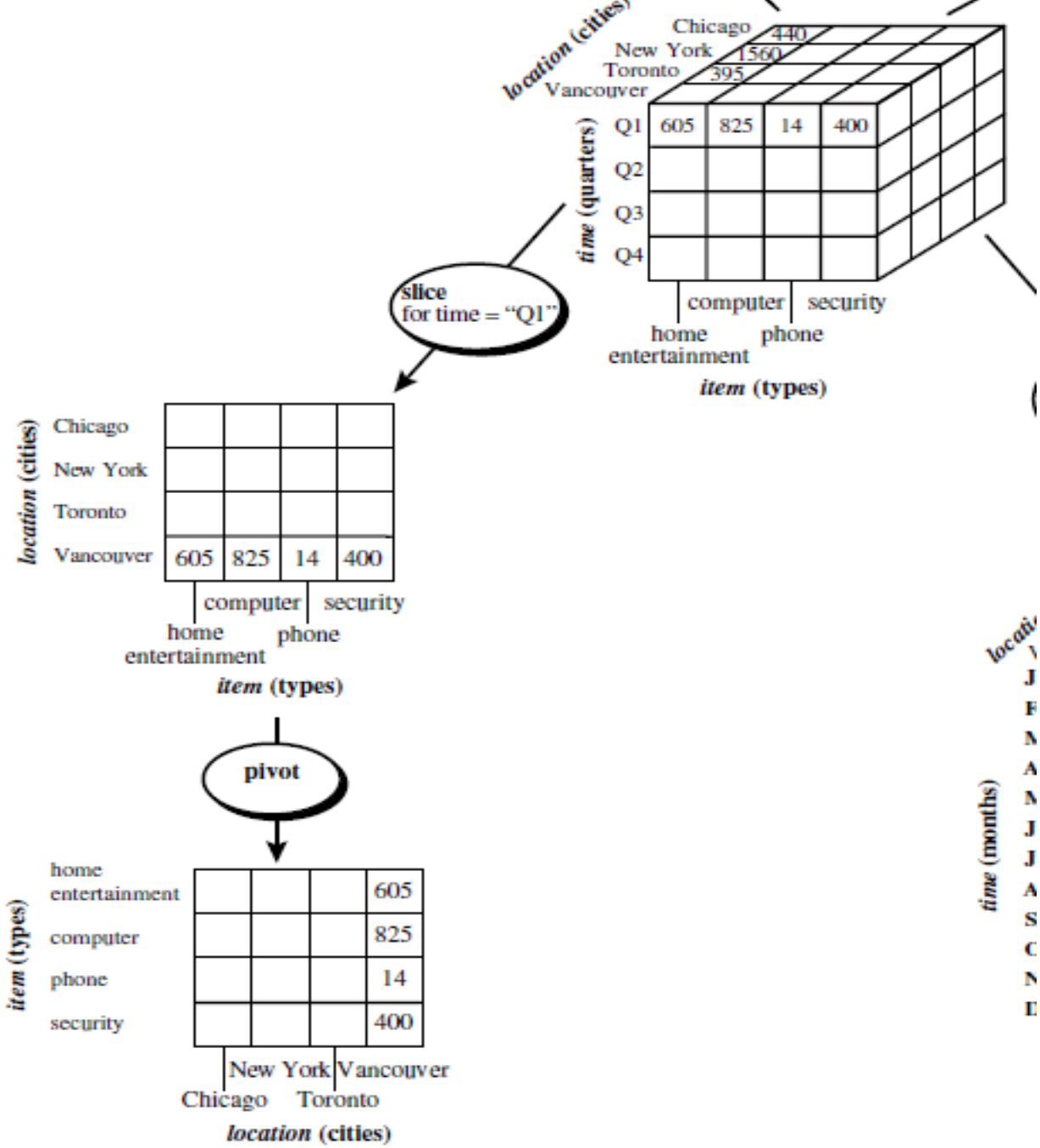
OLAP Operations in the Multidimensional Data Model

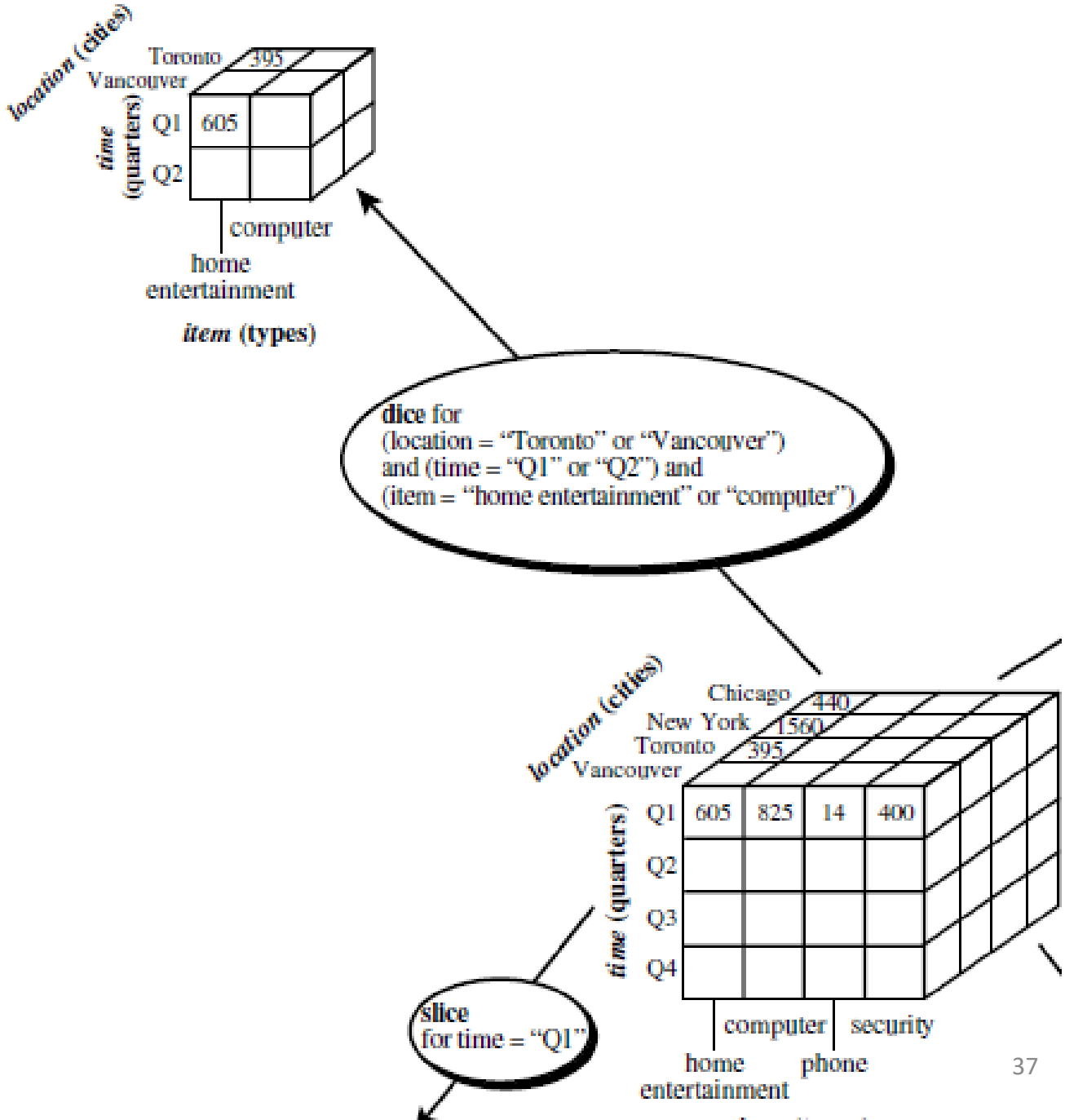
1. Roll-up:
2. Drill-down:
3. Slice and dice:
4. Pivot (rotate):


location (cities)
Chicago
854
882
89
623
New York
1087
968
38
872
Toronto
818
746
43
591
Vancouver
698
925
682
1002
789
728
870
984
784
time (quarters)
Q1
605
825
14
400
Q2
680
952
31
512
Q3
812
1023
30
501
Q4
927
1038
38
580
computer
security
home entertainment
phone
item (types)

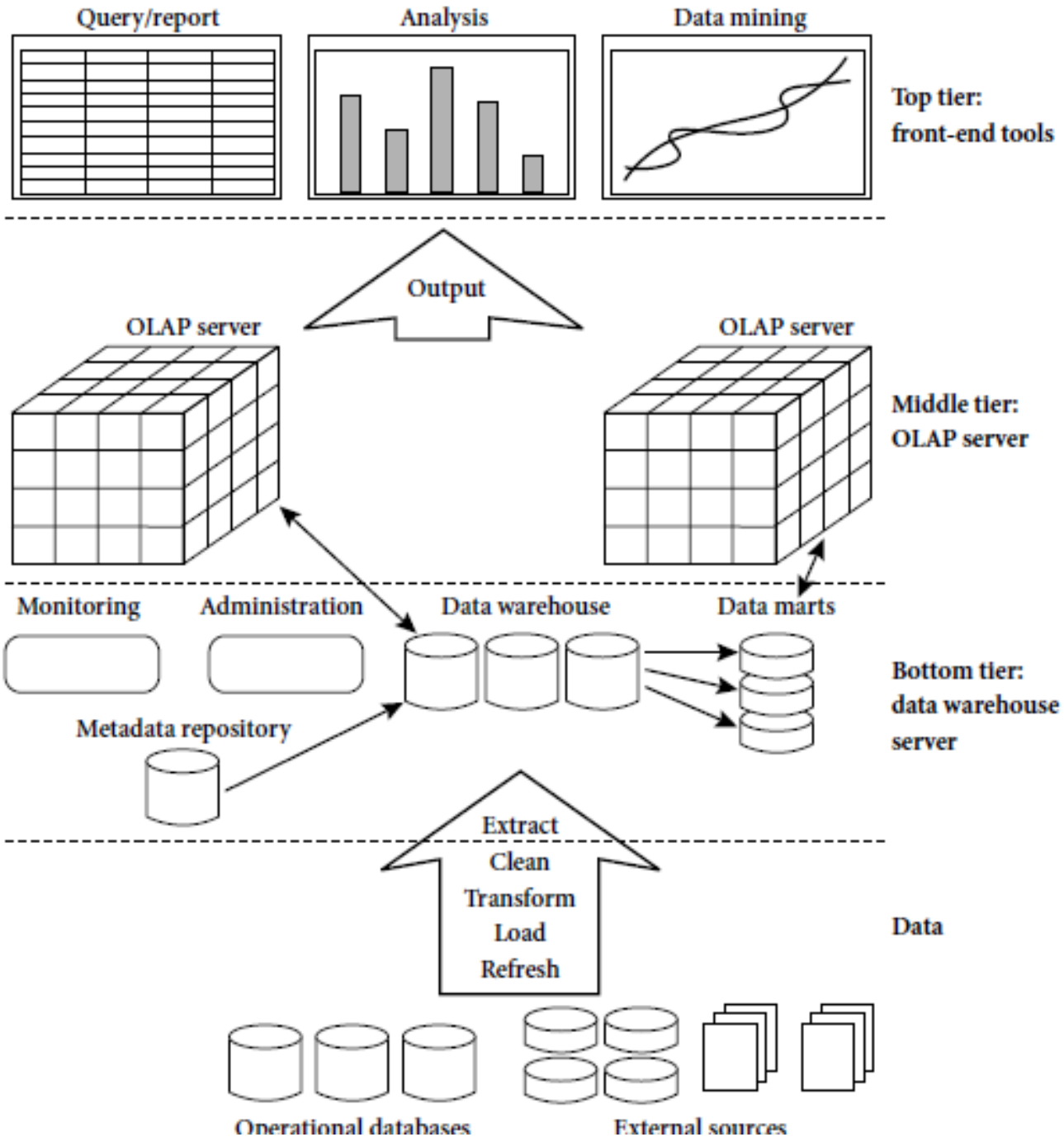

location (countries)
USA
Canada
2000
1000
time (quarters)
Q1
Q2
Q3
Q4
computer
security
home
entertainment
phone
item (types)
ronto" or "Vancouver")
1" or "Q2") and
entertainment" or "computer")
roll-up
on location
(from cities
to countries)
ocation (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
605
825
14
400
time (quarters)
Q1
Q2
Q3
Q4
computer
security
home
entertainment
phone
Q1"
drill-down


location (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
time (quarters)
Q1
Q2
Q3
Q4
605
825
14
400
computer
security
home entertainment
phone
item (types)
me = "Q1"
drill-down on time (from quarters to months)
location (cities)
Chicago
New York
Toronto
Vancouver
January
February
March
April
May
June
July
August
September
October
November
December
150
100
150
time (months)
computer
security
home entertainment
phone
item (types)


location (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
time (quarters)
Q1
Q2
Q3
Q4
605
825
14
400
computer
security
home entertainment
phone
item (types)
slice
for time = "Q1"
location (cities)
Chicago
New York
Toronto
Vancouver
605
825
14
400
computer
security
home entertainment
phone
item (types)
pivot
item (types)
home entertainment
computer
phone
security
605
825
14
400
New York
Vancouver
Chicago
Toronto
location (cities)
time (months)


location (cities)
Toronto
Vancouver
395
time (quarters)
Q1
605
Q2
computer
home entertainment
item (types)
dice for
(location = "Toronto" or "Vancouver")
and (time = "Q1" or "Q2") and
(item = "home entertainment" or "computer")
location (cities)
Chicago
440
New York
1560
Toronto
395
Vancouver
time (quarters)
Q1
605
825
14
400
Q2
Q3
Q4
computer
security
home entertainment
phone
slice
for time = "Q1"

A Three-Tier Data Warehouse 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

The bottom tier is a warehouse database server:

- Back-end tools and utilities are used to feed data from operational databases or other external sources.

- Tools and utilities perform data extraction, cleaning, and transformation.

- Load and refresh functions to update the data warehouse.

The middle tier is an OLAP server

- Implemented using either
  (1) a relational OLAP (ROLAP) model
  (2) a multidimensional OLAP (MOLAP) model

- The top tier is a front-end client layer

  - Contains query and reporting tools, analysis tools, and/or data mining tools

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

1. Enterprise warehouse:

- Information about subjects spanning the entire organization.

- Contains detailed data as well as summarized data.

- Size from a few gigabytes to hundreds of gigabytes, terabytes, or beyond.

- Implemented on traditional mainframes, computer super servers, or parallel architecture platforms

- Requires extensive business modeling

2. Data mart:

- A data mart contains a subset of corporate-wide data that is of value to a specific group of users.

- Scope is confined to specific selected subjects. For example, Marketing data mart

- The data contained in data marts tend to be summarized.
- Implemented on low-cost departmental servers

- Measured in weeks rather than months or years

3. Virtual warehouse:

- Virtual warehouse is a set of views over operational databases.

- For efficient query processing, only some of the possible summary views may be materialized.

Data Warehouse Back-End Tools and Utilities

- **Data extraction:** Typically gathers data from multiple, heterogeneous, and external sources.
- **Data cleaning:** Detects errors in the data and rectifies them when possible.
- **Data transformation:** Converts data from legacy or host format to warehouse format.
- **Load**: Sorts, summarizes, consolidates, computes views, checks integrity, and builds indices and partitions.
- **Refresh**: Propagates the updates from the data sources to the warehouse.

Metadata Repository

- Metadata are data about data.

- When used in a data warehouse, metadata are the data that define warehouse objects.

Metadata repository should contain the following:

- A description of *the structure of the data warehouse*
- *Operational metadata*
- *The algorithms used for summarization*
- *The mapping from the operational environment to the data warehouse*
- *Data related to system performance*
- *Business metadata*

Reference

1. J. Han, M. Kamber - “Data Mining Concepts and Techniques”, Morgan Kaufmann, 3rd Edition.

Thank You