DATA WAREHOUSES

Marek Maurizio

E-commerce, winter 2011

DATAWAREHOUSES AND OLAP

DATA WAREHOUSES

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

DATA WAREHOUSES

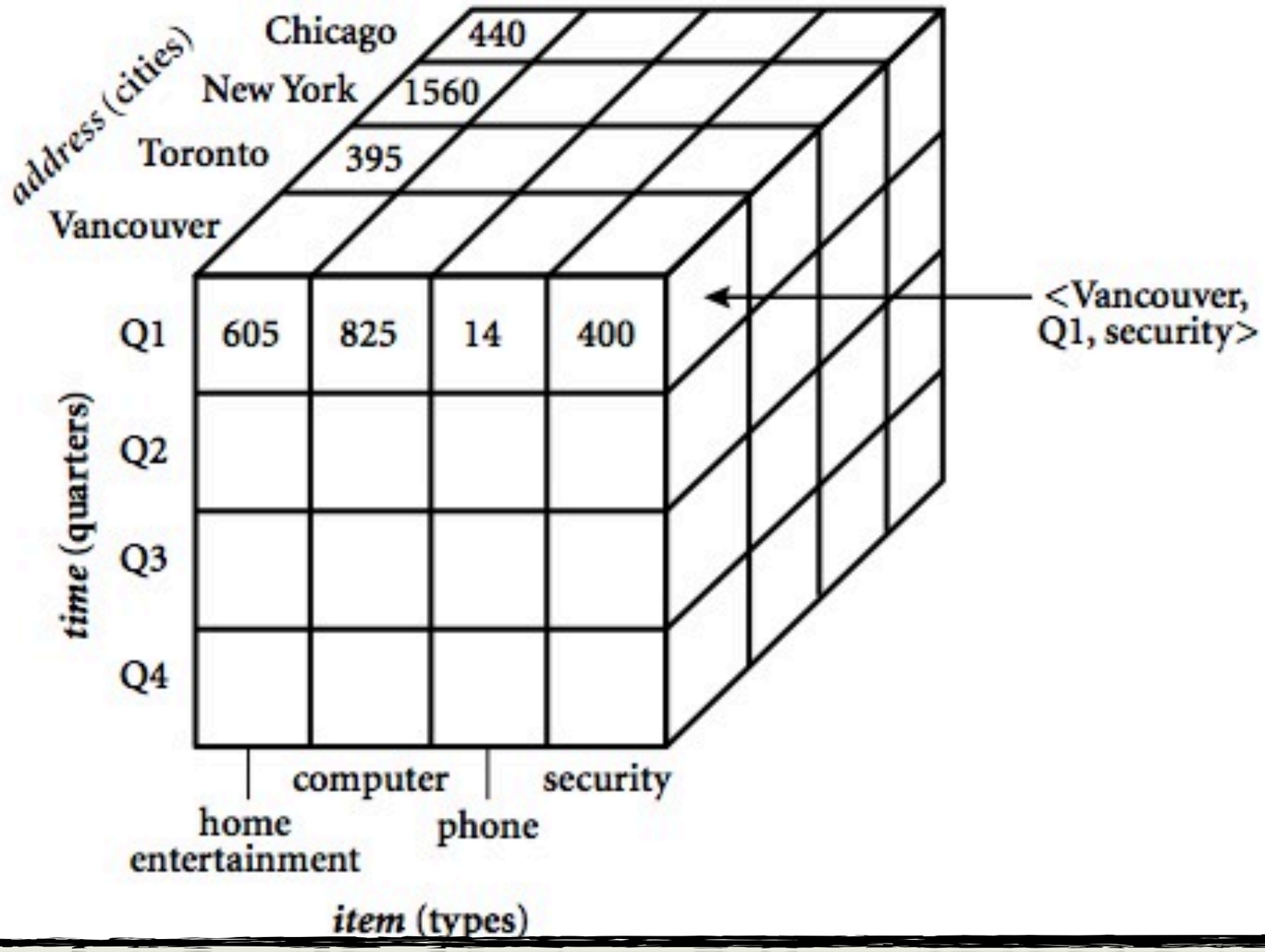
- To facilitate decision making, the data in a data warehouse are organized around major subjects
- Historical data, usually summarized

DW DATA REPRESENTATION

- A data warehouse is usually modeled by a multidimensional database structure, where each dimension corresponds to an *attribute* or a set of attributes in the schema, and each cell stores the value of some *aggregate measure*, such as count or sales amount

- The actual physical structure of a data warehouse may be a *relational data store* or a *multidimensional data cube*.

- A data cube provides a multidimensional view of data and allows the precomputation and fast accessing of summarized data

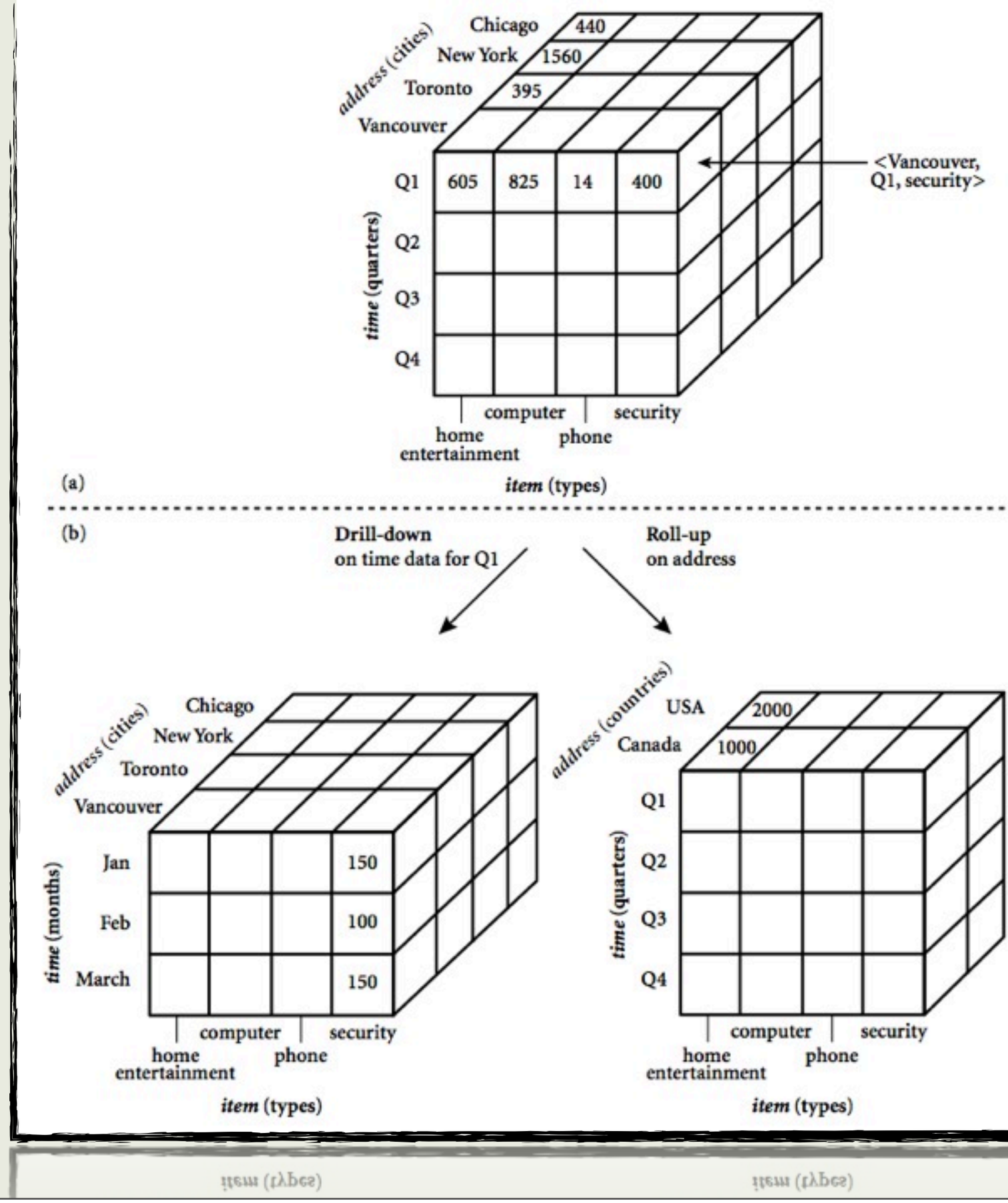

address (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
Q1
Q2
Q3
Q4
605
825
14
400
time (quarters)
<Vancouver,
Q1, security>
computer
security
home
entertainment
phone
item (types)

"I have also heard about data marts. What is the difference between a data warehouse and a data mart?"

A data warehouse collects information about subjects that span an entire organization, and thus its scope is enterprise-wide. A data mart, on the other hand, is a department subset of a data warehouse. It focuses on selected subjects, and thus its scope is department-wide

OLAP

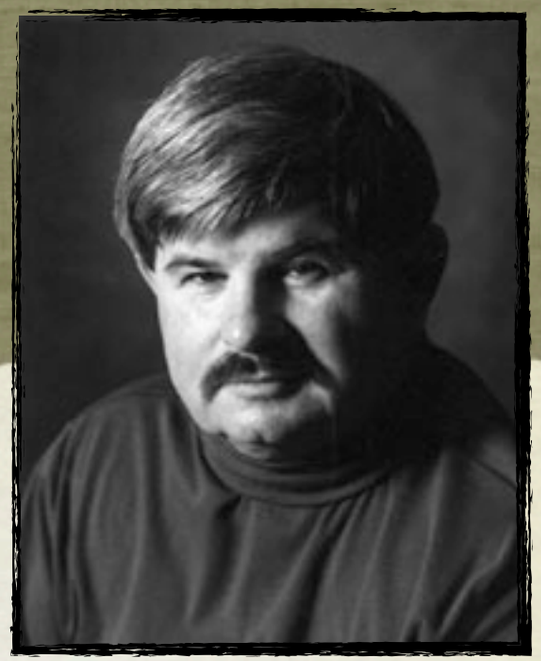
- data warehouse systems are well suited for on-line analytical processing, or OLAP
- OLAP operations use background knowledge regarding the domain of the data being studied in order to allow the presentation of data at different levels of abstraction
- Examples of OLAP operations include drill-down and roll-up, which allow the user to view the data at differing degrees of summarization


Chicago
New York
Toronto
Vancouver
address (cities)
440
1560
395
Q1
Q2
Q3
Q4
time (quarters)
605
825
14
400
<Vancouver, Q1, security>
home entertainment
computer
phone
security
item (types)
(a)
(b)
Drill-down on time data for Q1
Roll-up on address
Jan
Feb
March
time (months)
150
100
150
USA
Canada
address (countries)
2000
1000

“Then, what exactly is a data warehouse?”

Data warehouses have been defined in many ways, making it difficult to formulate a rigorous definition. Loosely speaking, a data warehouse refers to a database that is maintained separately from an organization’s operational databases

WHAT IS A DW?



William H. Inmon

- *"A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision making process"*

- four keywords, subject-oriented, integrated, time-variant, and nonvolatile

SUBJECT-ORIENTED

- A data warehouse is organized around major subjects, such as customer, supplier, product, and sales

- Rather than concentrating on the day-to-day operations and transaction processing of an organization, a data warehouse focuses on the modeling and analysis of data for decision makers

- Simple and concise view around particular subject issues

INTEGRATED

- A data warehouse is usually constructed by integrating multiple heterogeneous sources, such as relational databases, flat files, and on-line transaction records

TIME-VARIANT

- Data are stored to provide information from a historical perspective

- Every key structure in the data warehouse contains, either implicitly or explicitly, an element of time

NON-VOLATILE

- A data warehouse is always a physically separate store of data trans- formed from the application data found in the operational environment

- a data warehouse does not require transaction processing, recovery, and concurrency control mechanisms

- requires only *initial loading* and *access of data*

"How are organizations using the information from data warehouses?"

Many organizations use this information to support business **decision-making activities**, including:

- increasing **customer** focus, which includes the analysis of customer buying patterns (such as buying preference, buying time, budget cycles, and appetites for spending);
- repositioning **products** and managing product portfolios by comparing the performance of sales by quarter, by year, and by geographic regions in order to fine-tune production strategies;
- managing customer **relationships**;

data warehousing

has become popular in industry

DATA WAREHOUSES

VS

OPERATIONAL DATABASE SYSTEMS

OLTP

- 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

OLAP

- Data warehouse systems serve users or knowledge workers in the role of data analysis and decision making
- 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

Feature	*OLTP*	*OLAP*
Characteristic	operational processing	informational processing
Orientation	transaction	analysis
User	clerk, DBA, database professional	knowledge worker (e.g., manager, executive, analyst)
Function	day-to-day operations	long-term informational requirements, decision support
DB design	ER based, application-oriented	star/snowflake, subject-oriented
Data	current; guaranteed up-to-date	historical; accuracy maintained over time
Summarization	primitive, highly detailed	summarized, consolidated
View	detailed, flat relational	summarized, multidimensional
Unit of work	short, simple transaction	complex query
Access	read/write	mostly read
Focus	data in	information out
Operations	index/hash on primary key	lots of scans
Number of records accessed	tens	millions
Number of users	thousands	hundreds
DB size	100 MB to GB	100 GB to TB
Priority	high performance, high availability	high flexibility, end-user autonomy
Metric	transaction throughput	query throughput, response time

NOTE: Table is partially based on [CD97].

"why not perform on-line analytical processing directly on such databases instead of spending additional time and resources to construct a separate data warehouse?"

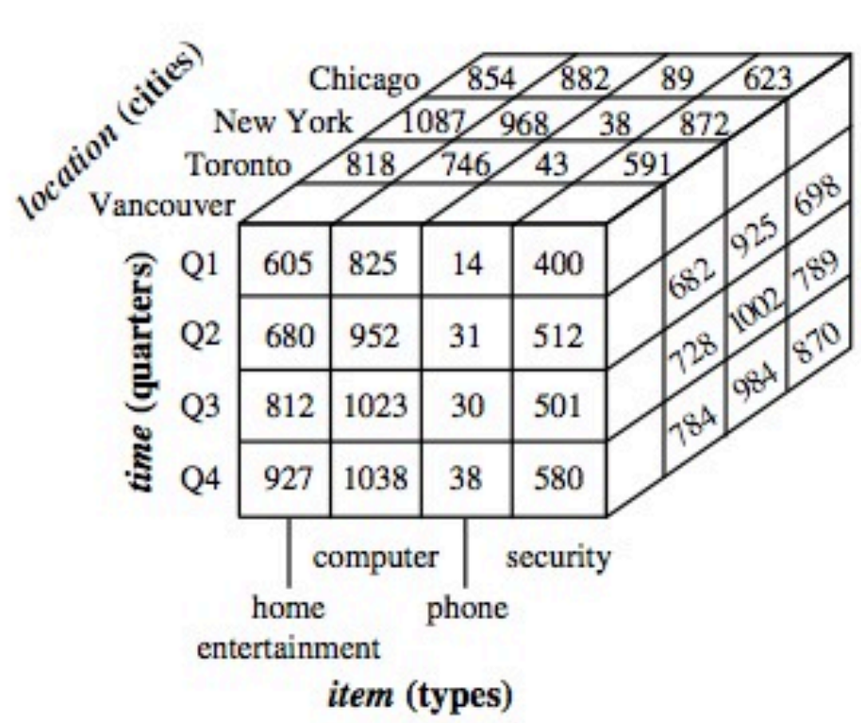
Difficult to have high performances on both tasks, different query languages.

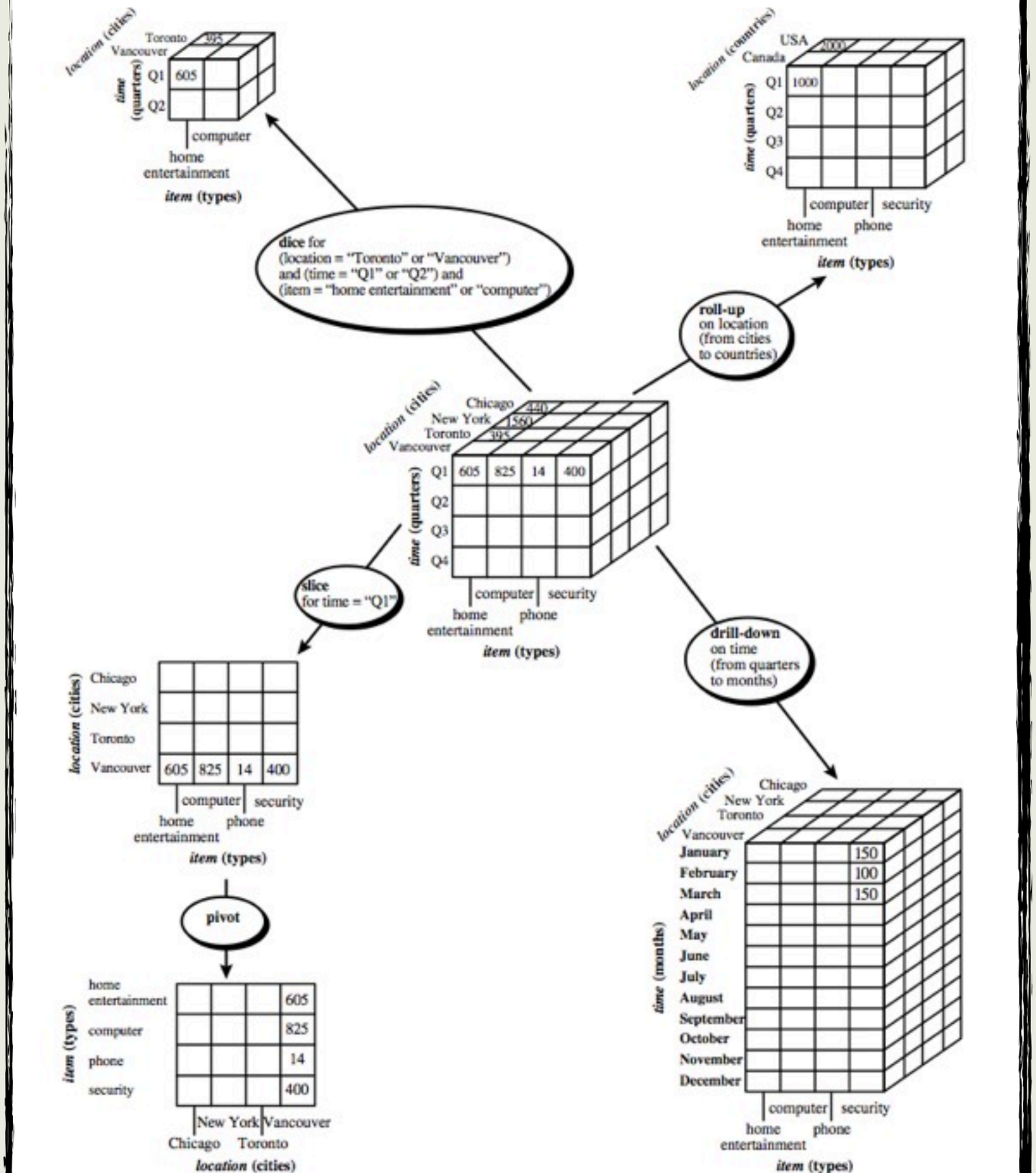
However vendors of operational relational database management systems are beginning to optimize such systems to support OLAP queries. As this trend continues, the separation between OLTP and OLAP systems is expected to decrease

MULTIDIMENSIONAL DATA MODEL

- **Data cube:** allows data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts.
- In general terms, dimensions are the perspectives or entities with respect to which an organization wants to keep records
- A multidimensional data model is typically organized around a central theme, like sales, for instance
- Although we usually think of cubes as 3-D geometric structures, in data warehousing the data cube is n-dimensional

	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





Examples of typical OLAP operations on multidimensional data.

OLAP OPERATIONS

- Operations on multidimensional data: roll-up, drill-down, slide-and-dice, pivoting
- Statistical operations and other measures
  - Different from a statistical database (SDB)

"How do data warehousing and OLAP relate to data mining?"

Typically, the longer a data warehouse has been in use, the more it will have evolved, from generating simple reports to complex knowledge discovery.

Three kinds of data warehouse applications: information processing, analytical processing, and data mining.

INFORMATION PROCESSING

- querying, basic statistical analysis, and reporting using crosstabs, tables, charts, or graphs. A current trend in data warehouse information processing is to construct low-cost Web-based accessing tools that are then integrated with Web browsers.

ANALYTICAL PROCESSING

- Analytical processing supports basic OLAP operations, including slice-and-dice, drill-down, roll-up, and pivoting. It generally operates on historical data in both summarized and detailed forms.

DATA MINING

- Data mining supports knowledge discovery by finding hidden patterns and associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools.

"Do OLAP systems perform data mining? Are OLAP systems actually data mining systems?"

The functionalities of OLAP and data mining can be viewed as disjoint: OLAP is a data summarization/aggregation tool that helps simplify data analysis, while data mining allows the automated discovery of implicit patterns and interesting knowledge hidden in large amounts of data.

SUMMARY

- What is a data warehouse
- Basics on dw data model
- How dw are related to data mining