Data warehousing Components

What Is Data Warehousing?

- Data Warehousing is an **architectural construct of information** systems that provides users with **current and historical decision support information** that is hard to access or present in traditional operational data stores.

- It is a blend of technologies and components aimed at effective integration of operational databases into environment that enables strategic use of data.

The need for data warehousing

- Business perspective
  - In order to survive and succeed in today’s highly competitive global environment:
    - Decisions need to be made quickly and correctly
    - The amount of data doubles every 18 months, which affects response time and the sheer ability to comprehend its content
    - Rapid changes

Business Problem Definition

- The business problems solved by data warehousing and complementary technologies provide organizations with a **sustainable competitive advantage.**
- An decision support of business applications helps to take decisions about all aspects of their business which includes
  - Customer retention
  - Sales and customer service
  - Marketing
  - Risk assessment and fraud detection

Business Problem Definition

- Data warehousing classify the business problems into
  - Retrospective analysis
  - Predictive Analysis
  - Retrospective analysis :Focuses on the present and past events.
  - Example: Analysis of the performance of the sales organization for the last 2 years across different geographic regions, demographics, and types of products
- Predictive analysis: Focuses on predicting certain events or behaviour based on historical information.
  - Example: predictive model which describes the attrition rates of their customers and define steps that reduce it
- This technique further classified as classification, clustering and segmentation, Associations and sequencing

Operational and Infomational Data

Operational Data:

– Focusing on transactional function such as bank card withdrawals and deposits

- Detailed
- Updateable
- Reflects current

– It answers such questions as “How many gadgets were sold to a customer number 123876 on September 19? ”

Operational and Infomational Data

Informational Data:

– Focusing on providing answers to problems posed by decision makers

- Summarized
- Nonupdateable

– “What three products resulted in the most frequent calls to the hotline over the past quarter?”

What is a Data Warehouse?

- Defined in many different ways, but not rigorously.
  - A decision support database that is maintained separately from the organization’s operational database
  - Support information processing by providing a solid platform of consolidated, historical data for analysis.
- “A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision-making process.”—
- Data warehousing:
  - The process of constructing and using data warehouses

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,

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
  - But the key of operational data may or may not contain “time element”

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*

OLTP Vs. OLAP

- OnLine Transaction Processing systems(**OLTP**) : Task of online operational database systems which manages transaction-oriented applications on the Internet is to perform OLTP.
- Cover day to day activities such as Purchasing, Inventory, manufacturing, banking etc.
- Online Analytical Processing systems(**OLAP**): DW 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 different users.

OLTP vs. OLAP

Feature	OLTP	OLAP
Characteristic	operational processing	Information Processing
Orientation	Transaction	Analysis
Summarization	Primitive, highly detailed	Summaried, consolidated
View	Detailed, Flat relational	Summarized,consolidated
Acess	Read/write	Mostly read/ Batch(update)

OLTP vs. OLAP

	OLTP	**OLAP**
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

To summarize ...

OLTP Systems are used to *“run”* a business

The Data Warehouse helps to *“optimize”* the business

Datawarehouse Models

- There are three data warehouse models:
  - Enterprise Warehouse
  - Data Mart
  - Virtual Warehouse

Enterprise Warehouse

- Enterprise Warehouse:
  - Collects all information abouts subjects spanning to entire organization
  - Corporate wide data integration obtained from operational system or external information
  - Contains detailed data as well as summarized data
  - Range in size of few gigabytes to hundreds of gigabytes
  - Implemented using mainframes, computer super servers or parallel architecture platforms
  - Needs extensive business modeling takes year to design

Data Marts

- Data marts are presented as an inexpensive alternative to a data warehouse subset of corporate wide data
- Scope confined to specific related subjects
- Data mart is a physically separate store of data and is normally resident on separate database server.
- Can be implemented in weeks
- Complexity lies integrating it in long run

Virtual Warehouse

- A virtual warehouse is a set of views over operational databases.
- For efficient query processing only possible summary views may be materialized
- Easy to build but requires excess capacity on operational database servers.

Data cube-Multidimensional data model

- Datawarehouses and OLAP tools are based on multidimensional data model
- A data cube, such as sales, allows data to be modelled and viewed in multiple dimensions.
- It is defined by **dimensions** and **facts**.
- Dimensions: Entities
- Suppose ALLELETRONICS create a *sales* data warehouse with respect to dimensions
  - Time
  - Item
  - Location
- Dimensions may have a table associated

Data Cube

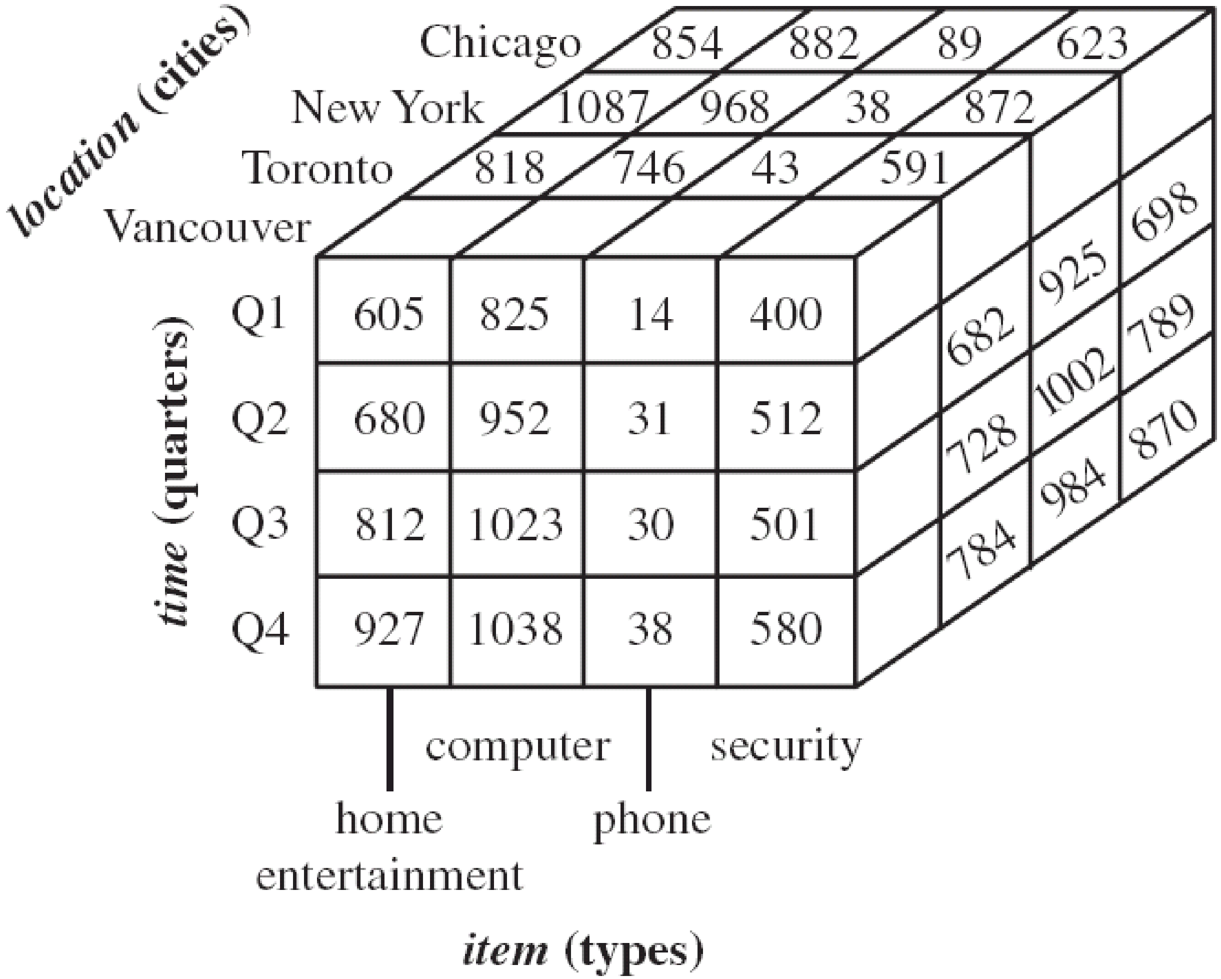
A multidimensional data model is organized around central theme which is represented as **Fact table.**

Facts are numeric measures, quantities helps to analyze relationships between dimensions.

Fact table contains the name of the facts or measures as well as the keys of the related dimension tables.

Cube can be defined as 3-D geometric structures

Data Cube



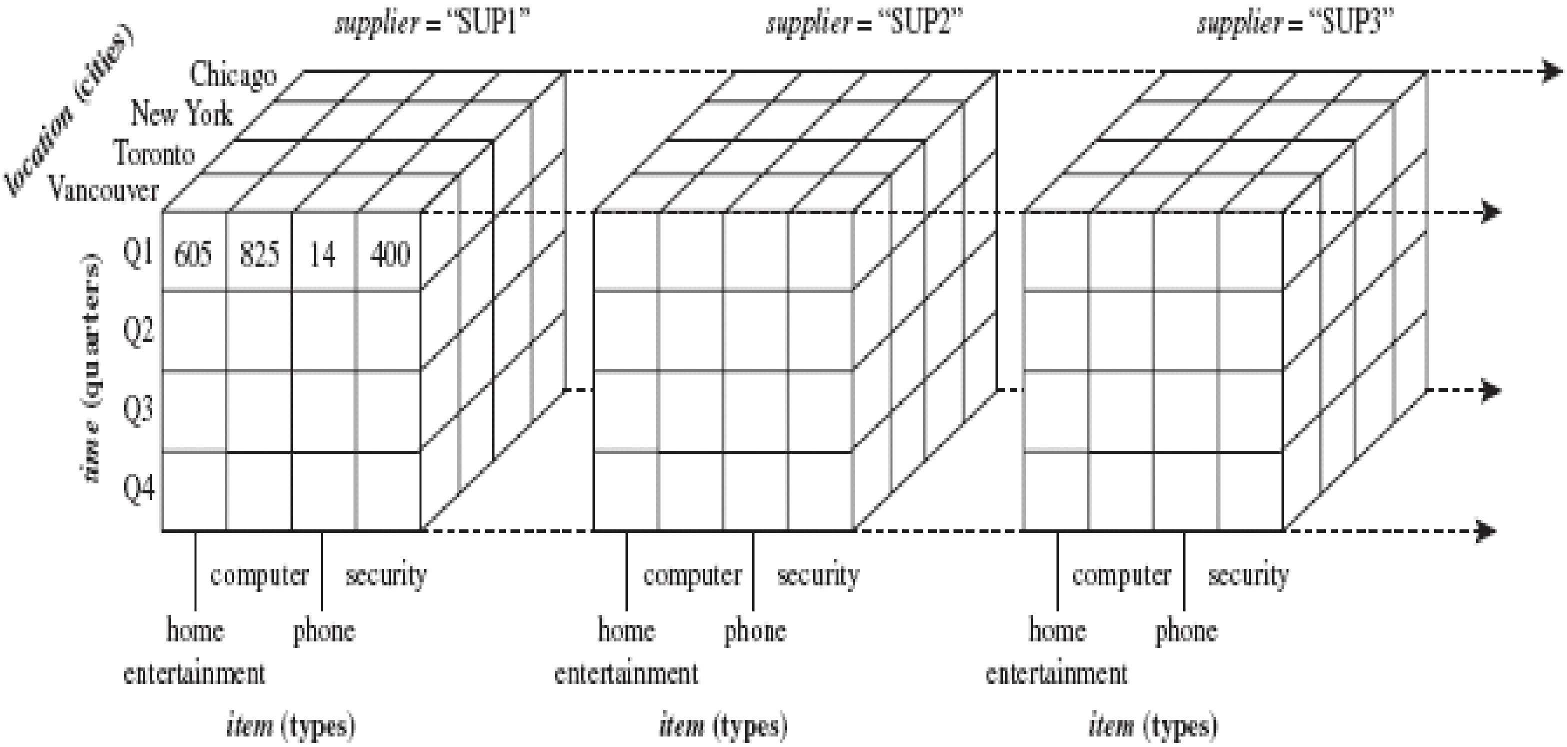
	home entertainment	computer	phone	security
Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	30	501
Q4	927	1038	38	580

- **Item,time and location**

Data Cube

- Suppose ALLELETRONICS create a *sales* data warehouse with respect to dimensions
  - Time
  - Item
  - Location
  - Supplier

Data Cube


supplier = "SUP1"
supplier = "SUP2"
supplier = "SUP3"
location (cities)
Chicago
New York
Toronto
Vancouver
time (quarters)
Q1
Q2
Q3
Q4
605
825
14
400
computer
security
home entertainment
phone
item (types)
computer
security
home entertainment
phone
item (types)
computer
security
home entertainment
phone
item (types)

Cuboid

Given a set of dimensions we can generate a cuboids for all possible subsets of the given dimensions.

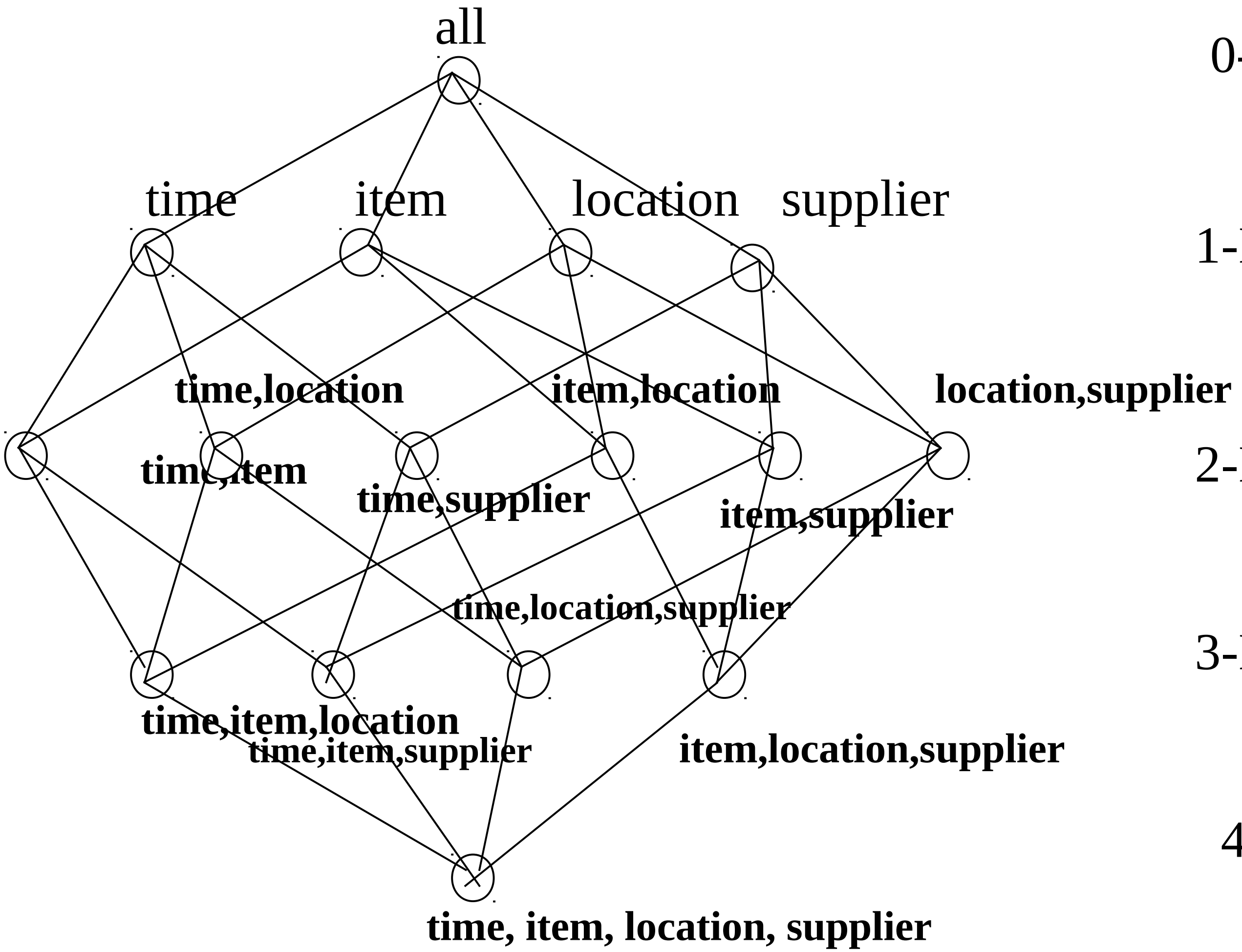
Lattice of cuboids: shows different level of summarization

Lattice also referred as **data cube**

The cuboid that holds the lowest level of summarization is called as base cuboid

The apex cuboid or 0-D cuboid holds the highest level of summarization

Data Cube: A Lattice of Cuboids



0-D (*apex*) cuboid

1-D cuboids

2-D cuboids

3-D cuboids

4-D (*base*) cuboid

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):

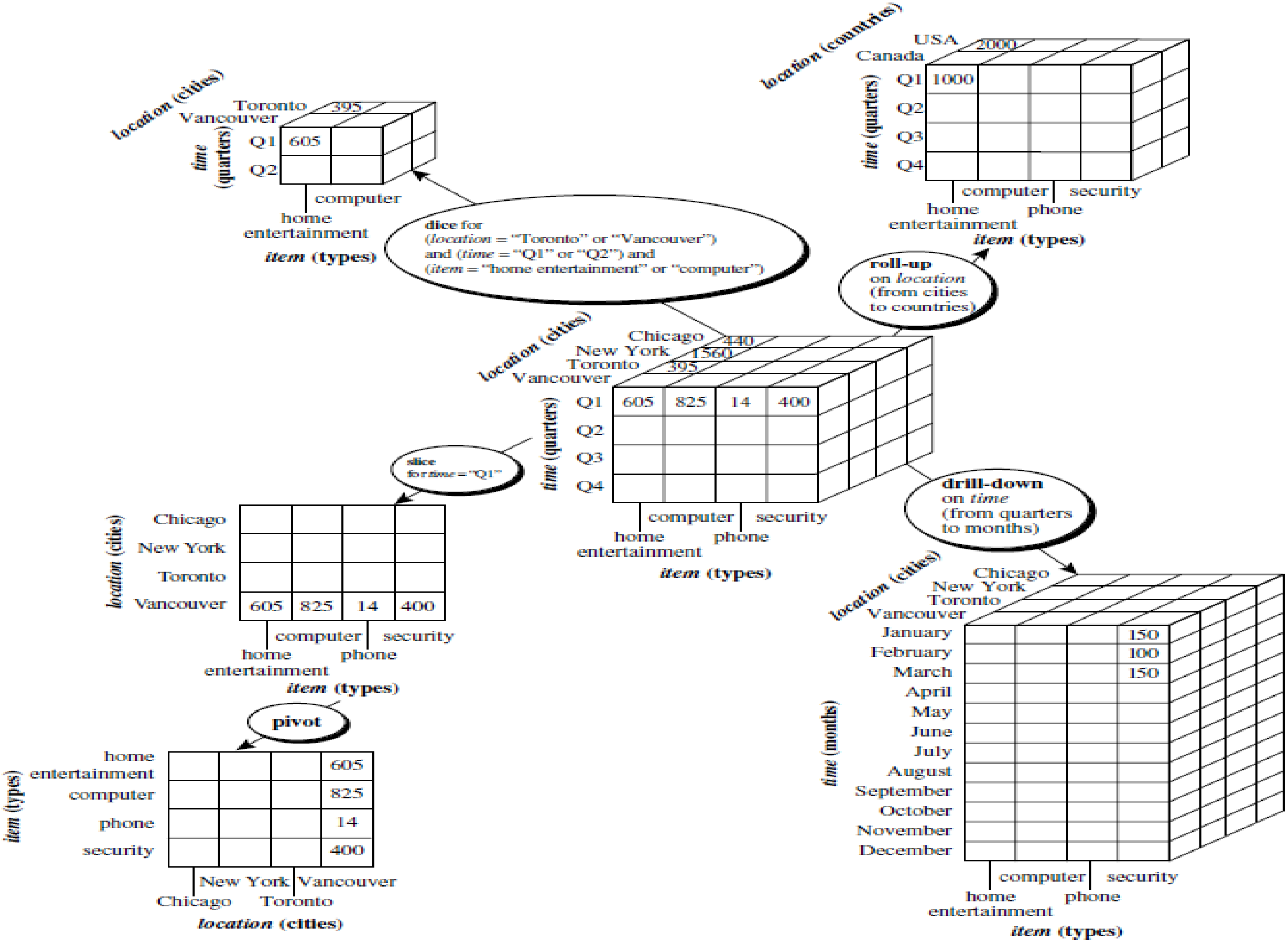
reorient the cube, visualization, 3D to series of 2D planes

Other operations

Drill across: involving (across) more than one fact table

Drill through: through the bottom level of the cube to its back

OLAP Operations



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