Introduction to Multidimensional Data Model

Tanzim Hossain
Lecturer
Department of Software Engineering
Daffodil International University

Online Transaction Processing (OLTP) Systems

OLTP

- The full form of OLTP is Online Transaction Processing.
- OLTP is an operational system that supports **transaction-oriented** applications.
- It administers the day-to-day transaction of an organization.

OLTP

- OLTP or Online Transaction Processing is a type of data processing that consists of executing a number of **transactions** occurring **concurrently**—online banking, shopping, order entry, or sending text messages, for example.
- These transactions traditionally are referred to as **economic** or **financial** transactions, recorded and **secured** so that an enterprise can access the information anytime for accounting or reporting purposes.

OLTP

- In the past, OLTP was limited to real-world interactions in which something was exchanged–money, products, information, request for services, and so on.
- But the definition of transaction in this context has expanded over the years, especially since the advent of the internet, to encompass any kind of digital interaction or engagement with a business that can be triggered from anywhere in the world and via any web-connected sensor.

OLTP

- It also includes any kind of **interaction** or **action** such as downloading pdfs on a web page, viewing a specific video, or automatic maintenance triggers or comments on social channels that maybe critical for a business to record to serve their customers better.

Transactions

- The primary definition for transactions—economic or financial—remains the foundation for most OLTP systems.
- So, online transaction processing typically involves inserting, updating, and/or deleting small amounts of data in a data store to collect, manage, and secure those transactions.

Transactions

- Typically a web, mobile, or enterprise application tracks all those interactions or transactions with customers, suppliers, or partners and updates them in the OLTP database.
- This transaction data stored in the database is critical for businesses and used for reporting or analyzed to use for data-driven decision making.

OLTP

- OLTP is basically focused on **query processing**, maintaining **data integrity** in **multi-access environments** as well as effectiveness that is measured by the **total number of transactions per second**.

Online Analytical Processing (OLAP)

OLAP

- Online analytical processing (OLAP) is software technology you can use to analyze business data from different points of view.
- Organizations collect and store data from multiple data sources, such as websites, applications, smart meters, and internal systems.
- OLAP combines and groups this data into categories to provide actionable insights for strategic planning.

OLAP

- For example, a retailer stores data about all the products it sells, such as color, size, cost, and location.
- The retailer also collects customer purchase data, such as the name of the items ordered and total sales value, in a different system.
- OLAP combines the datasets to answer questions such as which color products are more popular or how product placement impacts sales.

Multidimensional Data Representation

Why Multidimensional Data

- The multidimensional data model supports data representation and operations specifically tailored for **decision support processing** in data warehouses.

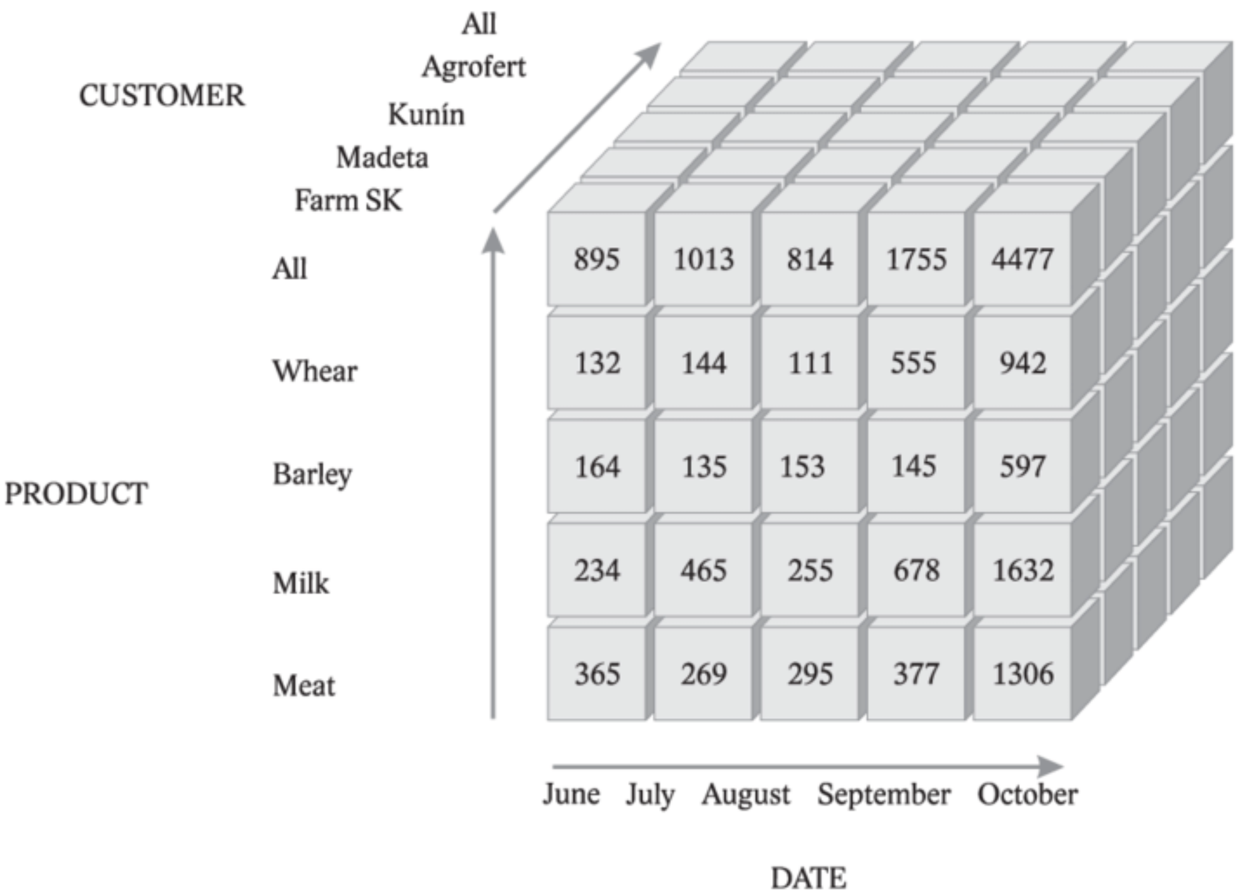
Why Dimensional Modeling?

- Dimensional modelling is widely accepted as the preferred technique for presenting analytic data because it addresses two simultaneous requirements:
  - Deliver data that’s **understandable** to the **business users**.
  - Deliver fast **query performance**.

An Example

- Imagine an executive who describes her business as, "We sell **products** in various **markets** and measure our performance over **time**."
- Dimensional designers listen carefully to the emphasis on product, market, and time.
- Most people find it intuitive to think of such a business as a cube of data, with the edges labelled product, market, and time.

An Example



CUSTOMER
All
Agrofert
Kunín
Madeta
Farm SK
PRODUCT
All
Whear
Barley
Milk
Meat
895
1013
814
1755
4477
132
144
111
555
942
164
135
153
145
597
234
465
255
678
1632
365
269
295
377
1306
June
July
August
September
October
DATE

An Example

- Imagine slicing and dicing along each of these dimensions.
- Points inside the cube are where the measurements, such as sales volume or profit, for that combination of product, market, and time are stored.

Another Example

- Consider a company that sells electronic products in different parts of the United States.
- In particular, the company markets four different printer products (mono laser, ink jet, photo, and portable) in five different states (California, Washington, Colorado, Utah, and Arizona).
- In order to store daily sales data for each product and each location in a relational database, you need Table, which consists of three columns (Product, Location, and Sales) and 20 rows (four instances of Product times five instances of Location).

Relational Data

Product	Location	Sales
Mono Laser	California	80
Mono Laser	Utah	40
Mono Laser	Arizona	70
Mono Laser	Washington	75
Mono Laser	Colorado	65
Ink Jet	California	110
Ink Jet	Utah	90
Ink Jet	Arizona	55
Ink Jet	Washington	85
Ink Jet	Colorado	45
Photo	California	60
Photo	Utah	50
Photo	Arizona	60
Photo	Washington	45
Photo	Colorado	85
Portable	California	25
Portable	Utah	30
Portable	Arizona	35
Portable	Washington	45
Portable	Colorado	60

Problems with Relational Data

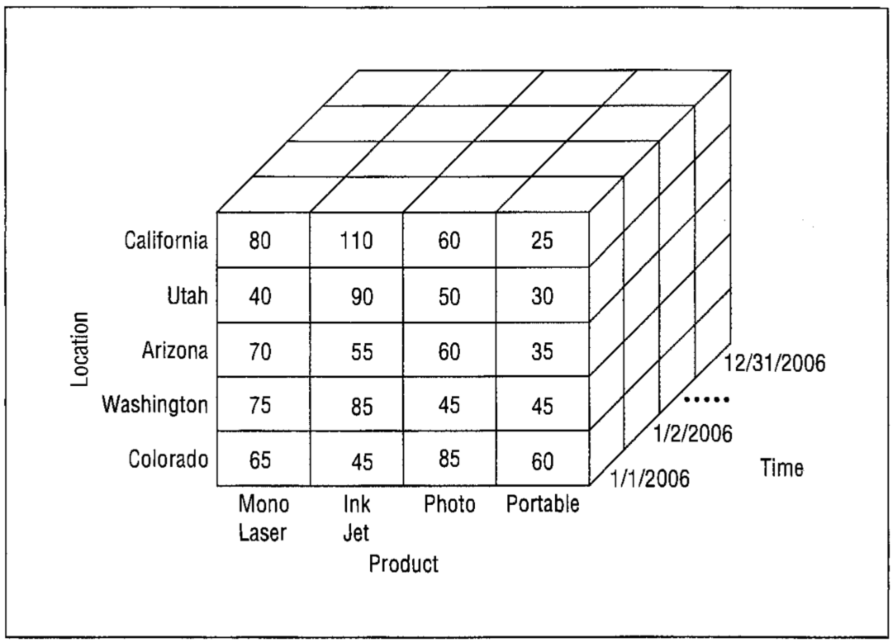
- The representation of Table can be complex and unwieldy.
- First, imagine that the company wishes to add a fifth product (say, color laser).
- In order to track sales by states for this new product, you need to add five rows, one each for each state.
- Second, note that the data in Table represents sales data for a particular day (for example, August 10, 2006).

Problems with Relational Data

- In order to store the same data for all 365 days of 2006, you need to add a fourth column to store the sales date, and duplicate the 20 rows for each date 365 times to yield a total of 7,300 rows.
- By the same token, if you wish to store historic data for a period of 10 years, you need 73,000 rows.
- Each new row must contain the product, state, and date values.

Solution: Multidimensional Data

Location	Product Mono Laser	Ink Jet	Photo	Portable
California	80	110	60	25
Utah	40	90	50	30
Arizona	70	55	60	35
Washington	75	85	45	45
Colorado	65	45	85	60



Multidimensional Data Advantages

- The multidimensional representation is simple to understand and extend.
- For example, adding a fifth product category requires an additional column to the right of Table.
- Adding dates requires a third dimension called Time, resulting in a three-dimensional arrangement.

Multidimensional Data Advantages

- The multidimensional representation also provides a convenient representation of summary totals.
- Each dimension in a cube can accommodate totals (row totals, column totals, depth totals, and overall totals) that a user can identify easily.

Multidimensional Data Advantages

- In addition to advantages in usability, the multidimensional representation can provide in creased retrieval speed.
- Direct storage of multidimensional data obviates the need to convert from a table representation to a multidimensional representation.

Multidimensional Data Disadvantages

- However, the multidimensional representation can suffer from excessive storage because many cells can remain empty.
- Even with compression techniques, large multidimensional tables can consume considerably more storage space than corresponding relational tables.

Multidimensional Terminologies

Data Cube

- A data cube or hypercube generalizes the two-dimensional and three- dimensional representations shown in the previous section.
- A data cube consists of cells containing **measures** (numeric values such as the unit sales amounts) and **dimensions** to label or group numeric data (e.g., Product, Location, and Time).
- Each dimension contains values known as **members**. For instance, the Location dimension has five members (California, Washington, Utah, Arizona, and Colorado).

Data Cube

- Both dimensions and measures can be **stored** or **derived**.
- For example, purchase date is a stored dimension with purchase year, month, and day as derived dimensions.

Dimension Details

- Dimensions can have hierarchies composed of levels.
- For instance, the Location dimension may have a hierarchy composed of the levels country, state, and city.
- Likewise, the Time dimension can have a hierarchy composed of year, quarter, month, and date.
- Hierarchies can be used to **drill down** from higher levels of detail (e.g., country) to lower levels (e.g., state and city) and to **roll up** in the reverse direction.

Dimension Details

- Although hierarchies are not essential, they allow a convenient and efficient representation.
- Without hierarchies, the Location dimension must contain the most detailed level (city).
- However, this representation can be difficult to compute aggregates across the dimension.
- Alternatively, the Location dimension can be divided into separate dimensions for country, state, and city resulting in a larger data cube.

Multiple Hierarchies

- For flexibility, dimensions can have multiple hierarchies.
- In a dimension with multiple hierarchies, usually at least one level is shared.
- For example, the Location dimension can have one hierarchy with the levels country, state, and city, and a second hierarchy with the levels country, state, and postal code.

Measure Details

- Cells in a data cube contain measures such as the sales values.
- Measures support numeric operations such as simple arithmetic, statistical calculations, and simultaneous equations.
- A cell may contain one or more measures.
- For example, the number of units can be another measure for the sales data cube.
- The number of nonempty cells in a multidimensional cube should equal the number of rows in the corresponding relational table.

Measure Details

- Derived measures can be stored in a data cube or computed from other measures at run time.
- Measures that can be derived from other measures in the same cell typically would not be stored.
- For example, total dollar sales can be calculated as total unit sales times the unit price measures in a cell.
- Summary measures derived from a collection of cells may be stored or computed depending on the number of cells and the cost of accessing the cells for the computation.

Data Cube Operations

Slice

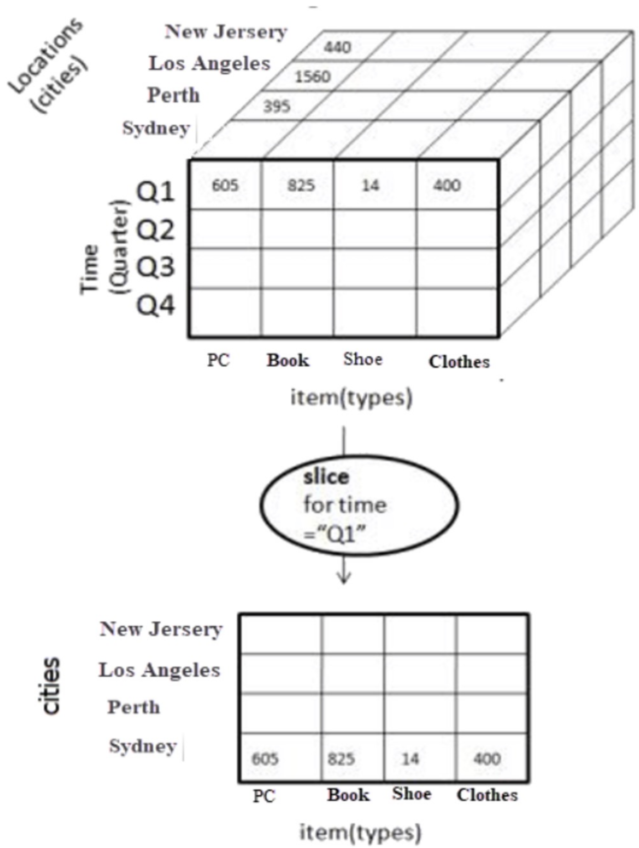
- Because a data cube can contain a large number of dimensions, users often need to focus on a subset of the dimensions to gain insights.
- The slice operator retrieves a subset of a data cube similar to the restrict operator of relational algebra.
- In a slice operation, one or more dimensions are set to specific values and the remaining data cube is displayed.

Slice

	Product			
Location	Mono Laser	Ink Jet	Photo	Portable
California	80	110	60	25
Utah	40	90	50	30
Arizona	70	55	60	35
Washington	75	85	45	45
Colorado	65	45	85	60

Figure shows the data cube resulting from the slice operation where Time = 1/1/2006 and the other two dimensions *(Location* and *Product)* are shown.

Slice



Locations (cities)
New Jersery
Los Angeles
Perth
Sydney
440
1560
395
Time (Quarter)
Q1
Q2
Q3
Q4
605
825
14
400
PC
Book
Shoe
Clothes
item(types)
slice for time ="Q1"
cities
New Jersery
Los Angeles
Perth
Sydney
605
825
14
400
PC
Book
Shoe
Clothes
item(types)

Slice-Summarize

- A variation of the slice operator allows a decision maker to summarize across members rather than to focus on just one member.
- The slice-summarize operator replaces one or more dimensions with summary calculations.
- The summary calculation often indicates the total value across members or the central tendency of the dimension such as the average or median value.

Slice-Summarize

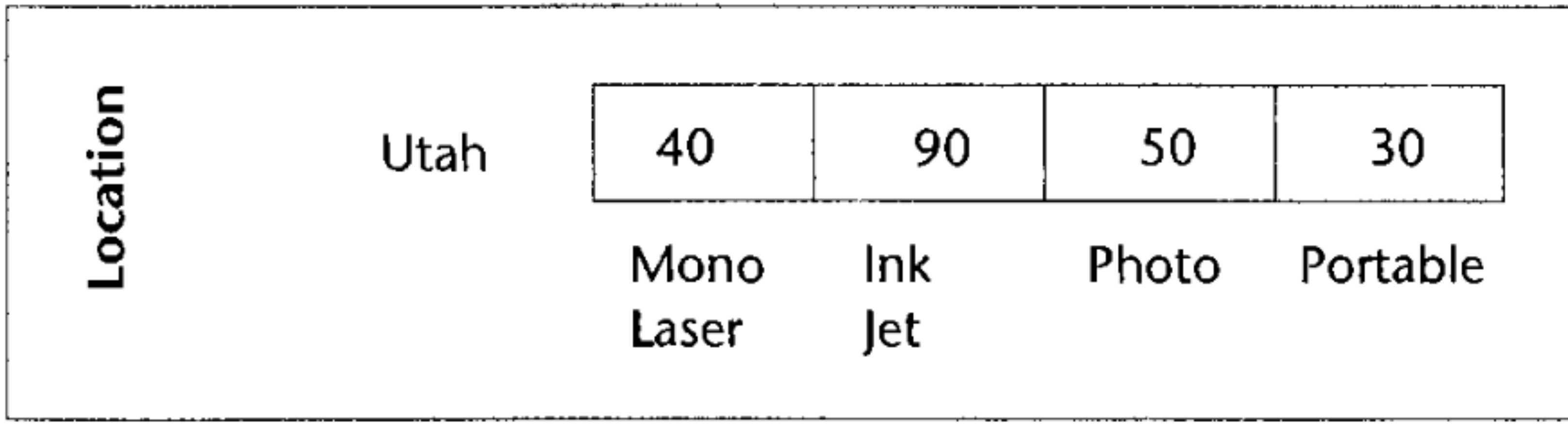
Location	Time			
	1/1/2006	1/2/2006	...	Total Sales
California	400	670	...	16,250
Utah	340	190	...	11,107
Arizona	270	255	...	21,500
Washington	175	285	...	20,900
Colorado	165	245	...	21,336

For example, Figure shows the result of a slice-summarize operation where the Product dimension is replaced by the sum of sales across all products. A new column called Total Sales can be added to store overall product sales for the entire year.

Dice

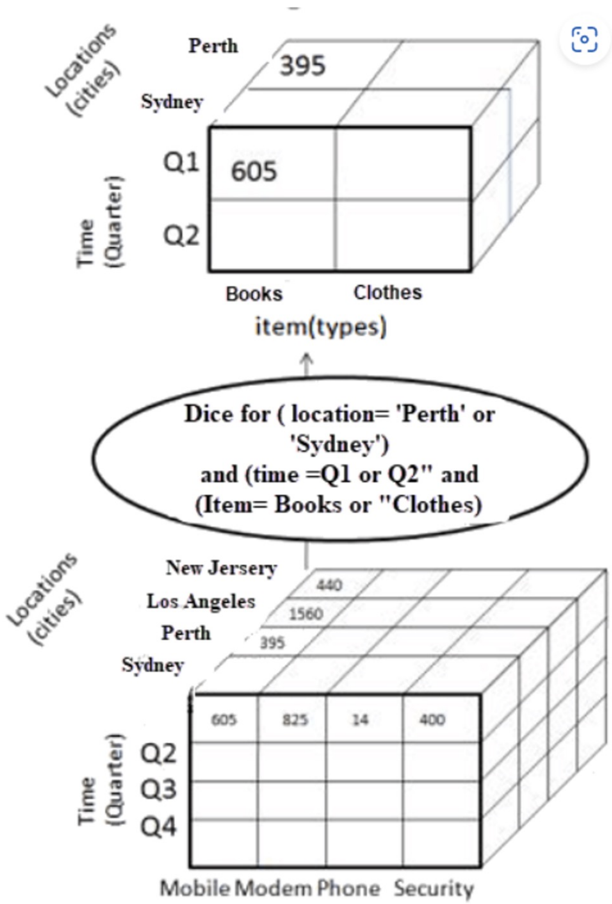
- Because individual dimensions can contain a large number of members, users need to focus on a subset of members to gain insights.
- The dice operator replaces a dimension with a subset of values of the dimension.

Dice



For example, Figure shows the result of a dicing operation to display sales for the State of Utah for January 1, 2006. Dicing typically follows a slicing operation and returns a subset of the values displayed in the preceding slice. It helps focus attention on one or more rows or columns of numbers from a slice.

Dice



Locations (cities)
Perth
Sydney
395
Q1
605
Q2
Time (Quarter)
Books
Clothes
item(types)
Dice for (location= 'Perth' or 'Sydney')
and (time =Q1 or Q2" and
(Item= Books or "Clothes)
Locations (cities)
New Jersery
Los Angeles
Perth
Sydney
440
1560
395
605
825
14
400
Q2
Q3
Q4
Time (Quarter)
Mobile Modem Phone Security

Drill-Down

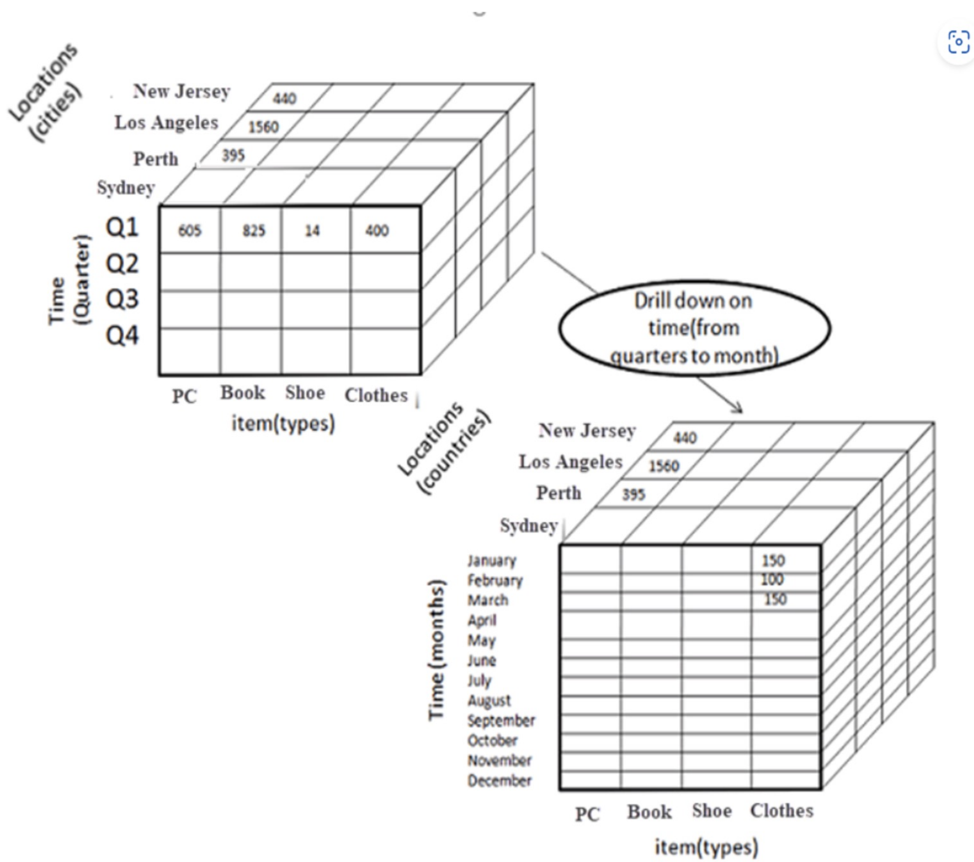
- Users often want to navigate among the levels of hierarchical dimensions.
- The drill-down operator allows users to navigate from a more general level to a more specific level.

Drill-Down

	Product			
Location	**Mono Laser**	**Ink Jet**	**Photo**	**Portable**
California	80	110	60	25
+ Utah				
Salt Lake	20	20	10	15
Park City	5	30	10	5
Ogden	15	40	30	10
Arizona	70	55	60	35
Washington	75	85	45	45
Colorado	65	45	85	60

Figure shows a drill-down operation on the State of Utah of the Location dimension. The plus sign by Utah indicates a drill-down operation.

Drill-Down

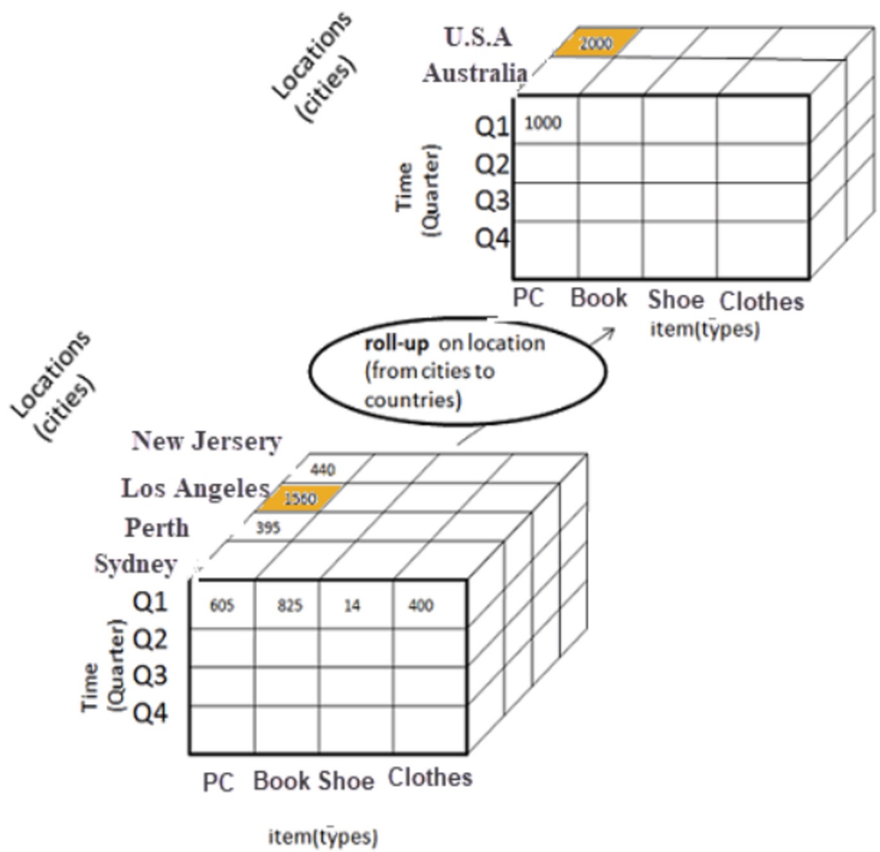


Locations (cities)
New Jersey
Los Angeles
Perth
Sydney
440
1560
395
Time (Quarter)
Q1
Q2
Q3
Q4
605
825
14
400
PC
Book
Shoe
Clothes
item(types)
Drill down on time(from quarters to month)
Locations (countries)
New Jersey
Los Angeles
Perth
Sydney
440
1560
395
Time (months)
January
February
March
April
May
June
July
August
September
October
November
December
150
100
150
PC
Book
Shoe
Clothes
item(types)

Roll-Up

- Roll-up (also called drill-up) is the opposite of drill-down.
- Roll-up involves moving from a specific level to a more general level of a hierarchical dimension.
- For example, a decision maker may roll up sales data from daily to quarterly level for end-of-quarter reporting needs.

Roll-Up

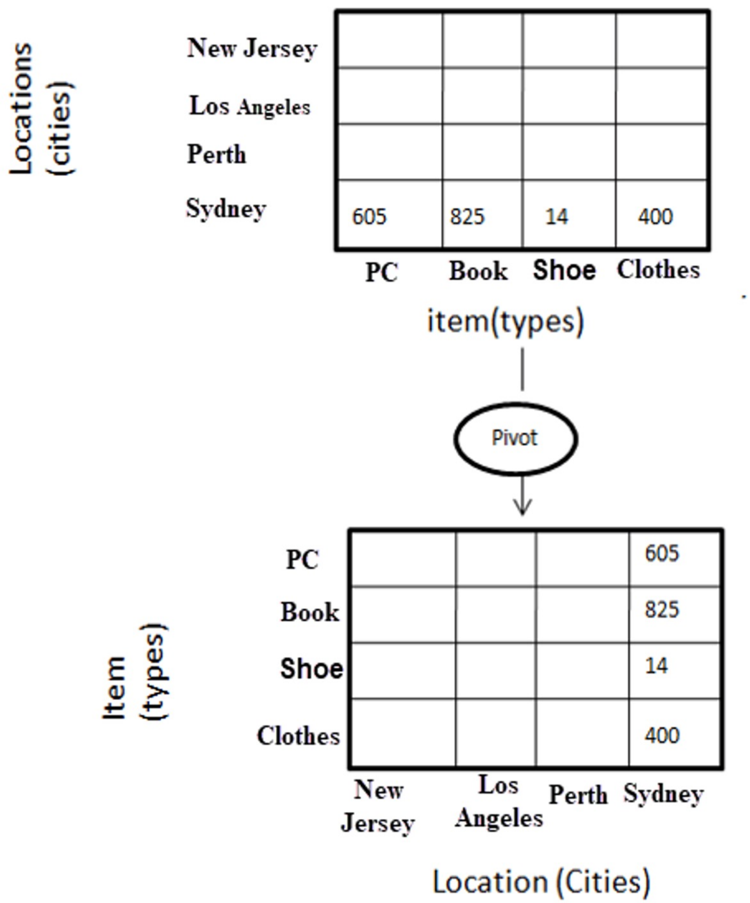


Locations (cities)
U.S.A
Australia
2000
1000
Q1
Q2
Q3
Q4
Time (Quarter)
PC
Book
Shoe
Clothes
item(types)
roll-up on location (from cities to countries)
Locations (cities)
New Jersery
Los Angeles
Perth
Sydney
440
1560
395
605
825
14
400
Q1
Q2
Q3
Q4
Time (Quarter)
PC
Book
Shoe
Clothes
item(types)

Pivot

- The pivot operator supports rearrangement of the dimensions in a data cube.
- For example, the position of the Product and the Location dimensions can be reversed so that Product appears on the rows and Location on the columns.
- The pivot operator allows a data cube to be presented in the most appealing visual order.

Pivot


Locations (cities)
New Jersey
Los Angeles
Perth
Sydney
605
825
14
400
PC
Book
Shoe
Clothes
item(types)
Pivot
Item (types)
PC
Book
Shoe
Clothes
605
825
14
400
New Jersey
Los Angeles
Perth
Sydney
Location (Cities)

Pivot

- The pivot operator supports rearrangement of the dimensions in a data cube.
- For example, the position of the Product and the Location dimensions can be reversed so that Product appears on the rows and Location on the columns.
- The pivot operator allows a data cube to be presented in the most appealing visual order.

Question

Questions

- What are the advantages of multidimensional representation over relational representation for business analysts?
- Explain why a dimension may have multiple hierarchies.
- What are the advantages of using time-series data in a cell instead of time as a dimension?
- How is slicing a data cube different from dicing?
- What are the differences between drilling-down and rolling-up a data cube dimension?
- How is a pivot operation useful for multidimensional databases?