OntarioTech Science

Kourosh Davoudi
kourosh@uoit.ca

Week 2: Data Exploration (OLAP)

CSCI 4150U: Data Mining

Data Mining: Data Exploration

OLAP

On-Line Analytical Processing (OLAP)

- On-Line Analytical Processing (OLAP) was proposed by E. F. Codd, the father of the relational database.
- Relational databases put data into tables, while OLAP uses a multidimensional array representation.
  - Such representations of data previously existed in statistics and other fields
- There are a number of data analysis and data exploration operations that are easier with such a data representation.

Creating a Multidimensional Array (Example: Iris data)

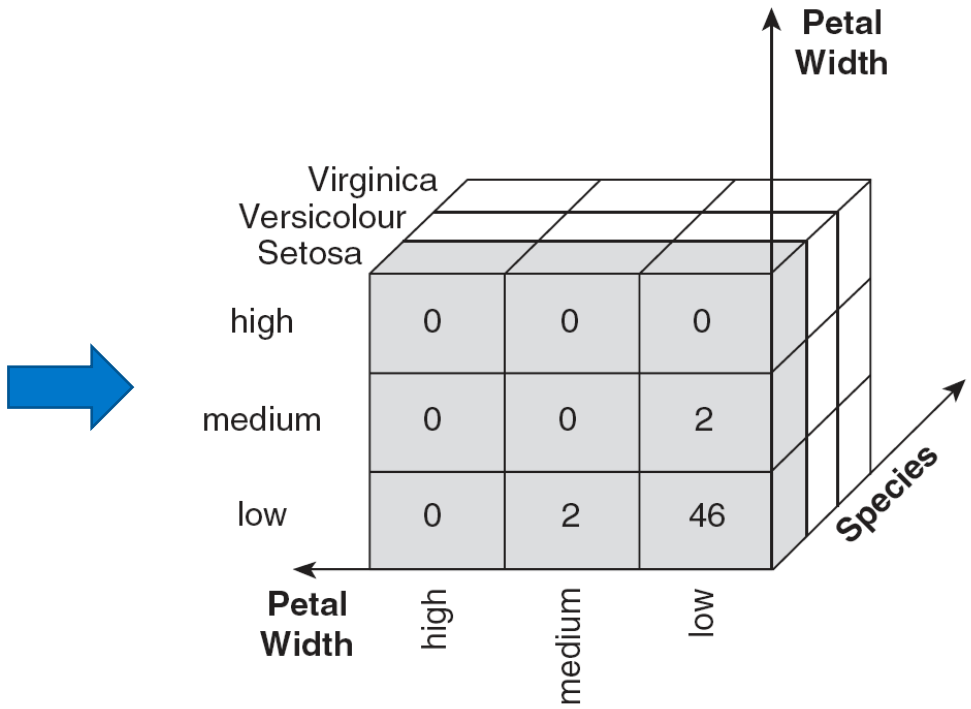
- We show how the attributes, petal length, petal width, and species type can be converted to a multidimensional array
  - First, we discretized the petal width and length to have categorical values: low, medium, and high

Petal Length	Petal Width	Species Type	Count
low	low	Setosa	46
low	medium	Setosa	2
medium	low	Setosa	2
medium	medium	Versicolour	43
medium	high	Versicolour	3
medium	high	Virginica	3
high	medium	Versicolour	2
high	medium	Virginica	3
high	high	Versicolour	2
high	high	Virginica	44

Example: Iris data (continued)

- Each unique tuple of petal width, petal length, and species type identifies one element of the array.
- This element is assigned the corresponding count value.
- All non-specified tuples are 0.

Petal Length	Petal Width	Species Type	Coun
low	low	Setosa	46
low	medium	Setosa	2
medium	low	Setosa	2
medium	medium	Versicolour	43
medium	high	Versicolour	3
medium	high	Virginica	3
high	medium	Versicolour	2
high	medium	Virginica	3
high	high	Versicolour	2
high	high	Virginica	44



Creating a Multidimensional Array (General Procedure)

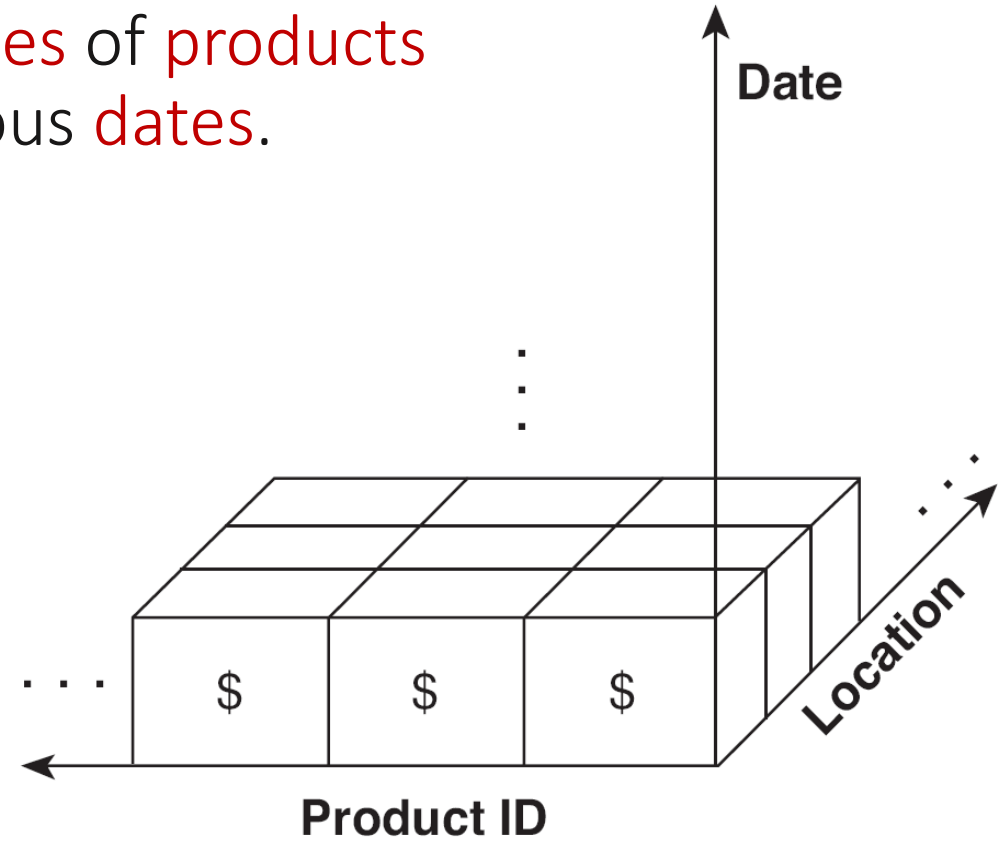
- Converting tabular data into a multidimensional array:
  - Identify which attributes are to be the dimensions and which attribute is to be the target attribute
    - Attributes used as dimensions must have discrete values
    - Values of target variable appear as entries in the array
    - The target value is typically a count or continuous value
    - Can have no target variable at all except the count of objects that have the same set of attribute values
  - Find the value of each entry in the multidimensional array by summing the values (of the target attribute) or the count of all objects that have the attribute values corresponding to that entry.

OLAP Operations: Data Cube

- The key operation of a OLAP is the formation of a data cube
- A data cube is a multidimensional representation of data, together with all possible aggregates.
- By all possible aggregates, we mean the aggregates that result by:
  - selecting a proper subset of the dimensions and summing over all remaining dimensions.

Data Cube Example

- Consider a data set that records the sales of products at a number of company stores at various dates.
- This data can be represented as a 3 dimensional array
- There are 3 two-dimensional aggregates (3 choose 2), 3 one-dimensional aggregates, and 1 zero-dimensional aggregate (the overall total)



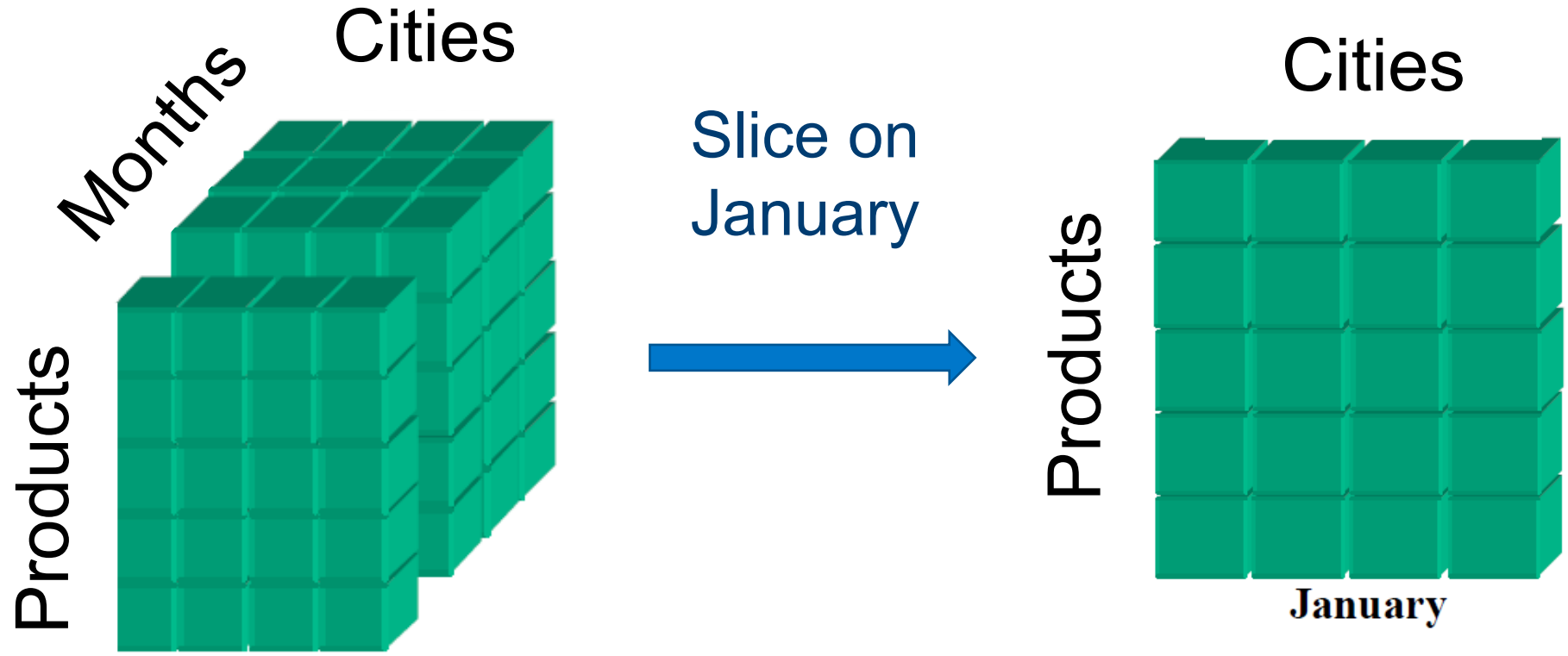
Data Cube Example (continued)

- The following figure table shows one of the two dimensional aggregates, along with two of the one-dimensional aggregates, and the overall total

product ID \ date	Jan 1, 2004	Jan 2, 2004	...	Dec 31, 2004	total
1	$1,001	$987	...	$891	$370,000
⋮	⋮			⋮	⋮
27	$10,265	$10,225	...	$9,325	$3,800,020
⋮	⋮			⋮	⋮
total	$527,362	$532,953	...	$631,221	$227,352,127

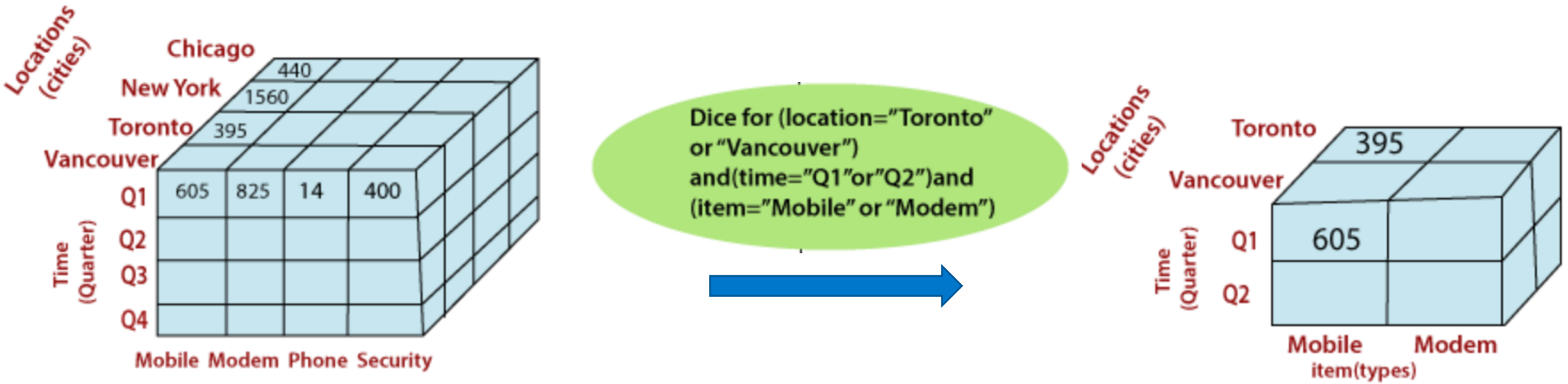
OLAP Operations: Slicing

- Slicing is selecting a subset of cells from the entire multidimensional array by specifying a specific value for one/more dimensions.



OLAP Operations: Dicing

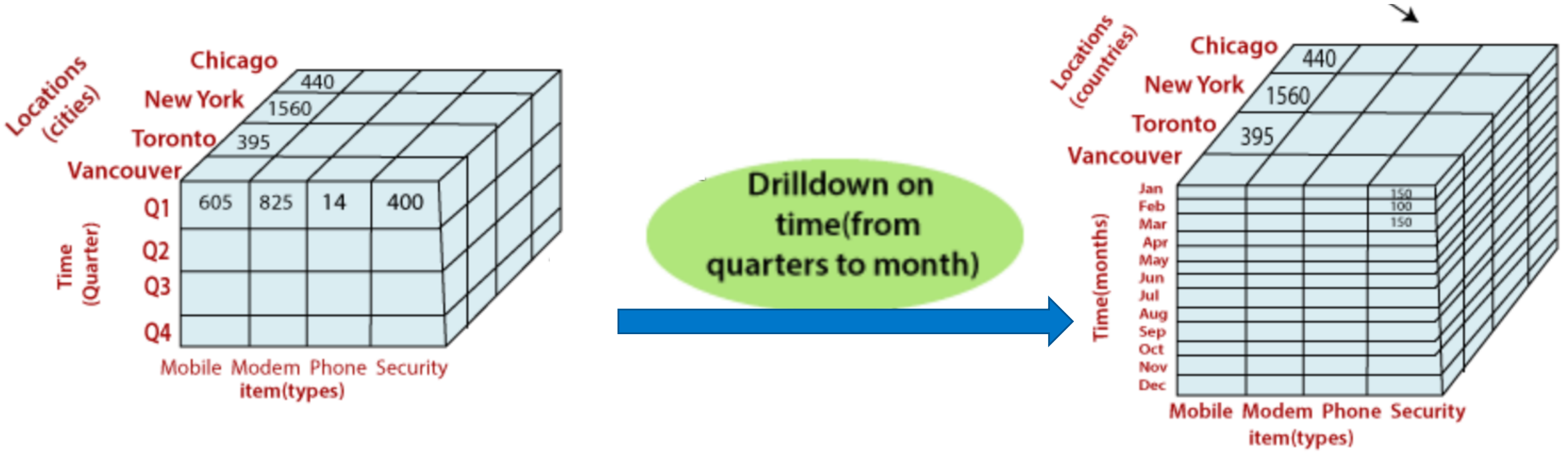
- Dicing involves selecting a subset of cells by specifying a range of attribute values for dimensions.
  - This is equivalent to defining a subarray from the complete array.



Example: Roll-up

https://www.javatpoint.com/olap-operations

Example: Drill-Down



https://www.javatpoint.com/olap-operations

OLAP Operations: Roll-up and Drill-down

- Attribute values often have a hierarchical structure.
  - Each date is associated with a year, month, and week.
  - A location is associated with a continent, country, state (province, etc.), and city.
  - Products can be divided into various categories, such as clothing, electronics, and furniture.
- Note that these categories often nest and form a tree or lattice
  - A year contains months which contains day
  - A country contains a state which contains a city

OLAP Operations: Roll-up and Drill-down

- This hierarchical structure gives rise to the **roll-up** and drill-down operations.
  - For sales data, we can aggregate (roll up) the sales across all the dates in a month.
  - Conversely, given a view of the data where the time dimension is broken into months, we could split the monthly sales totals (drill down) into daily sales totals.
  - Likewise, we can drill down or roll up on the location attributes.