Data Mining

Ch2. Introduction to Data Warehousing

Er. Bidur Devkota
Gandaki College of Engineering and Science
Pokhara 16

DATA CUBES

- We all are familiar with Relational Database
- Relational databases include tables and fields which are joined together by keys.
- Relational databases are great – businesses cannot run without them.
- They are optimized to **store** information into a system in a cohesive manner.
- But, **NOT** optimized for **getting** the information out of the system.
- **Data Cubes** serve for such purpose

DATA CUBES

Example: “How much profit did we make selling Wai Wai Noodles to Iceland last year?”

- In business decision, such query are often too frequent.
- The problem with a relational database, is to answer that type of question, we need to get information that is scattered across many different tables.
- In a typical database, to answer this question we may need to combine data from the:
  - Customer table
  - Country/Region table
  - Item table
  - Sales Invoice Line
  - Sales Credit Memo Line
  - Sales Invoice Header
  - Sales Credit Memo Header
- Finally, mash up and extract the data to get the information that we need.
- **Implications**: slow and resource intensive process.

DATA CUBES

Example: “How much profit did we make selling Wai Wai Noodles to Iceland last year?”

- Optimal Solution must be **easy** and **quick**: **Data Cubes.**
- Other related question must also be answered with it:
  - Who were these customers?
  - Are sales growing or shrinking?
  - Did sales fluctuate month-over-month?
  - Who was our top salesperson?
  - Could we put that person to a better use?

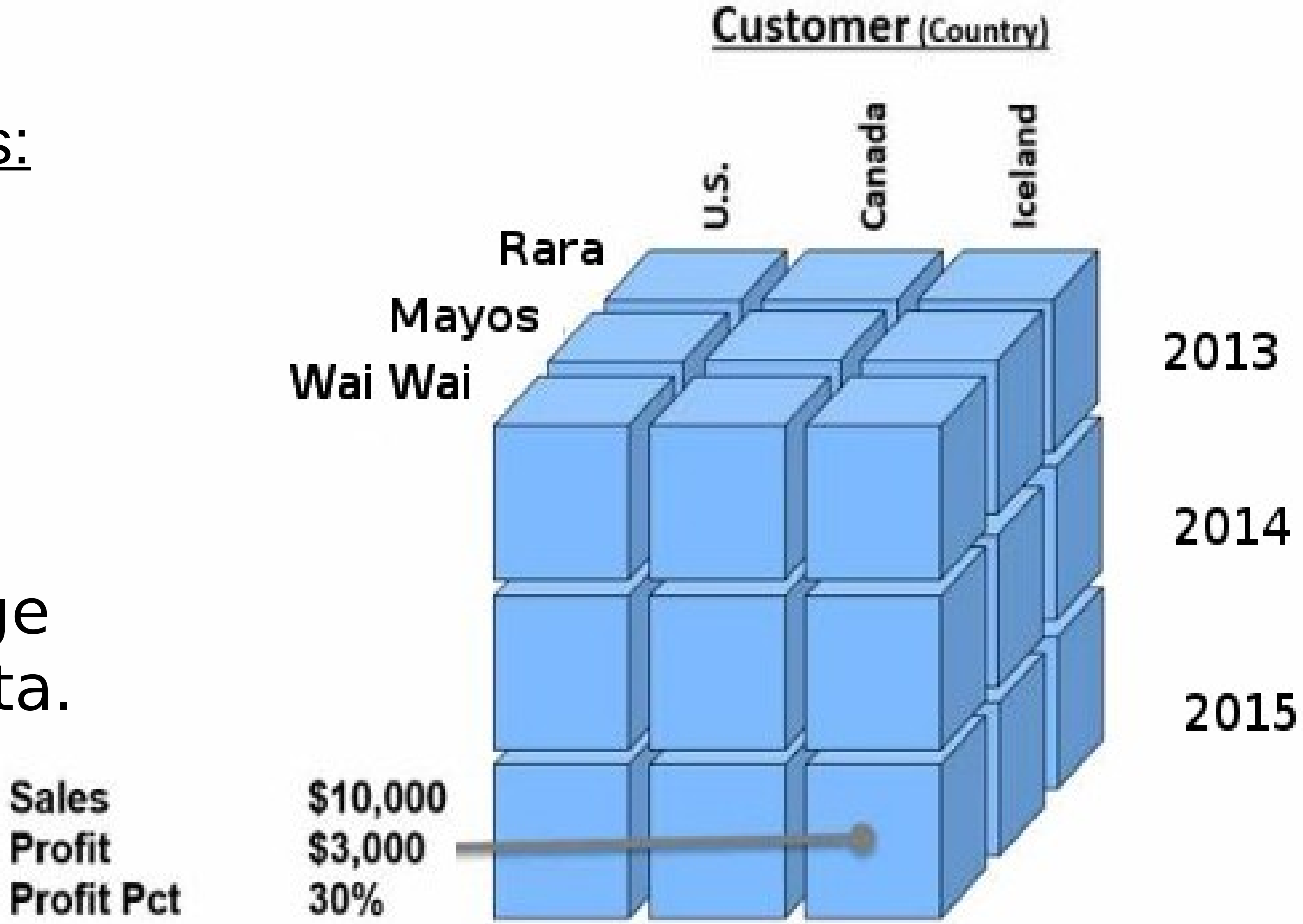
Cubes reorganizes a copy of the data so that such information can be accessed easily and quickly.

DATA CUBES

multi-dimensional way of organizing data

Example

- three dimensions: Products, customers (by country), posting date (year).
- near instant analysis of large amounts of data.



- To see profit for **Wai Wai** in **Iceland** in **2015**, all of that data is put together in the **Cube**

DATA CUBES

Now, more into Data Cubes!

- Data cube is a structure that enable OLAP to achieves the multidimensional functionality.
- The data cube is used to represent data along some measure of interest.
- Data Cubes are an easy way to look at the data (allow us to look at complex data in a simple format).
- Although called a "cube", it can be 2-dimensional, 3-dimensional, or higher-dimensional.

Dimensions And Measures

- data cubes have categories of data called **dimensions** and **measures**.
- **measure**
  - represents some fact (or number) such as cost or units of service.
- **dimension**
  - represents descriptive categories of data such as time or location.

Dimensions And Measures

- data cubes have categories of data called **dimensions** and **measures**.
- **measure**
  - represents some fact (or number) such as cost or units of service.
- **dimension**
  - represents descriptive categories of data such as time or location.

Data Cubes Concepts

- Three important concepts associated with data cubes :
    1. Slicing.
    2. Dicing.
    3. Rotating.

Slicing

- the term slice most often refers to a two- dimensional page selected from the cube.

- subset of a multidimensional array corresponding to a single value for one or more members of the dimensions not in the subset.

Slicing

- the term slice most often refers to a two- dimensional page selected from the cube.

- subset of a multidimensional array corresponding to a single value for one or more members of the dimensions not in the subset.

Slicing

- the term slice most often refers to a two- dimensional page selected from the cube.

- subset of a multidimensional array corresponding to a single value for one or more members of the dimensions not in the subset.

Dicing

- A related operation to slicing .
- in the case of dicing, we define a subcube of the original space.
- Dicing provides you the smallest available slice.

Dicing

- A related operation to slicing .

- in the case of dicing, we define a subcube of the original space.

- Dicing provides you the smallest available slice.

Rotating

- Some times called pivoting.
- Rotating changes the dimensional orientation of the report from the cube data.
- For example …
  - rotating may consist of swapping the rows and columns, or moving one of the row dimensions into the column dimension
  - or swapping an off-spreadsheet dimension with one of the dimensions in the page display

Rotating

- Some times called pivoting.
- Rotating changes the dimensional orientation of the report from the cube data.
- For example …
  - rotating may consist of swapping the rows and columns, or moving one of the row dimensions into the column dimension
  - or swapping an off-spreadsheet dimension with one of the dimensions in the page display

Dimensions

- represents descriptive categories of data such as time or location.
- Each dimension includes different levels of categories.

Dimensions

- represents descriptive categories of data such as time or location.
- Each dimension includes different levels of categories.

Categories

- is an item that matches a specific description or classification such as years in a time dimension.

- Categories can be at different levels of information within a dimension.

Categories

- is an item that matches a specific description or classification such as years in a time dimension.
- Categories can be at different levels of information within a dimension.

Categories

- is an item that matches a specific description or classification such as years in a time dimension.
- Categories can be at different levels of information within a dimension.

Categories

- is an item that matches a specific description or classification such as years in a time dimension.

- Categories can be at different levels of information within a dimension.

measures

- The measures are the actual data values that occupy the cells as defined by the dimensions selected.
- Measures include facts or variables typically stored as numerical fields.

measures

- The measures are the actual data values that occupy the cells as defined by the dimensions selected.
- Measures include facts or variables typically stored as numerical fields.

Computed versus Stored Data Cubes

- The goal is to retrieve the information from the data cube in the most efficient way possible.
- Three possible solutions are:
  - Pre-compute all cells in the cube.
  - Pre-compute no cells.
  - Pre-compute some of the cells.

Computed versus Stored Data Cubes

- The goal is to retrieve the information from the data cube in the most efficient way possible.
- Three possible solutions are:
  - Pre-compute all cells in the cube.
  - Pre-compute no cells.
  - Pre-compute some of the cells.

Computed versus Stored Data Cubes

- The goal is to retrieve the information from the data cube in the most efficient way possible.
- Three possible solutions are:
  - Pre-compute all cells in the cube.
  - Pre-compute no cells.
  - Pre-compute some of the cells.

representation of Totals

- A simple data cube does not contain totals.
- The storage of totals increases the size of the data cube **but** can also decrease the time to make total-based queries.

- A simple way to represent totals is to add an additional layer on *n* sides of the *n*-dimensional data cube.

representation of Totals

- A simple data cube does not contain totals.
- The storage of totals increases the size of the data cube **but** can also decrease the time to make total-based queries.

- A simple way to represent totals is to add an additional layer on *n* sides of the *n*-dimensional data cube.

REFRENCES

- Jiawei Han and Micheline Kamber, *Data Mining: Concepts and Techniques*, 2nd Edition, 2006, Morgan Kaufmann.
- Data Mining By Pieter Adriaans, Dolf Zantinge
- http://www2.cs.uregina.ca/~dbd/cs831/notes/dcubes/dcubes.html
- Data Cube Presentation,2011, Mohammed Siddig Ahmed, Sudan University
- http://projects.cs.dal.ca/panda/datacube.html
- http://blogs.jetreports.com/2014/05/28/olap-cubes-101/