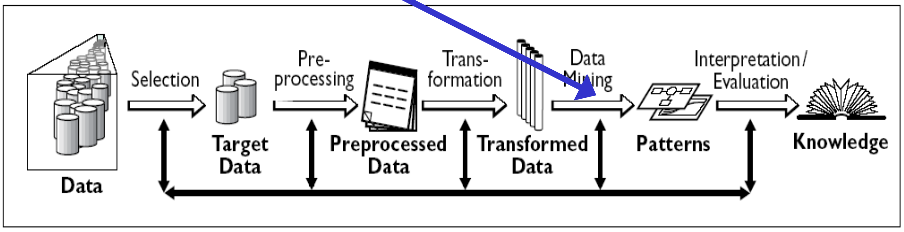
Building Data Cubes and Mining Them

Jelena Jovanovic
Email: jeljov@fon.bg.ac.yu

KDD Process

- KDD is an overall process of discovering useful knowledge from data.
- Data mining is a particular step in the KDD process.



Data Warehouse & OLAP

What is Data Warehouse?

- A repository of information collected through multiple sources and stored under a unified schema at a single site.
- A database that is maintained separately from the organization's operational database.
- Contains collection of consolidated, historical data aimed for online analysis and decision support.
- Focuses on the modeling and analysis of data for decision makers, not on daily operations or transaction processing.
- Provides a simple and concise view of particular subject issues by excluding data that are not useful in the decision support process.

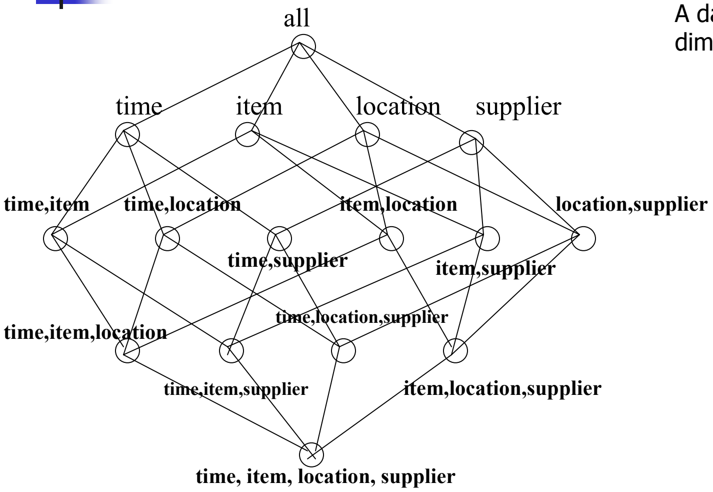
Data Warehouse Usage

- Three kinds of data warehouse applications
  - Information processing
    - supports querying, basic statistical analysis, and reporting using cross-tables, tables, charts and graphs
  - Analytical processing
    - multidimensional analysis of data warehouse data
    - supports basic OLAP operations
  - Data mining
    - knowledge discovery from hidden patterns
    - supports associations, constructing analytical models, performing classification and prediction.

Data Cube

- A data warehouse is based on a multidimensional data model which views data in the form of a data cube.
- A data cube (e.g. sales) allows data to be modeled and viewed in multiple dimensions. It consists of:
  - Dimension tables
    - such as item (item_name, brand, type), or time(day, week, month, quarter, year)
  - Fact table
    - contains measures (such as dollars_sold) and keys to each of the related dimension tables

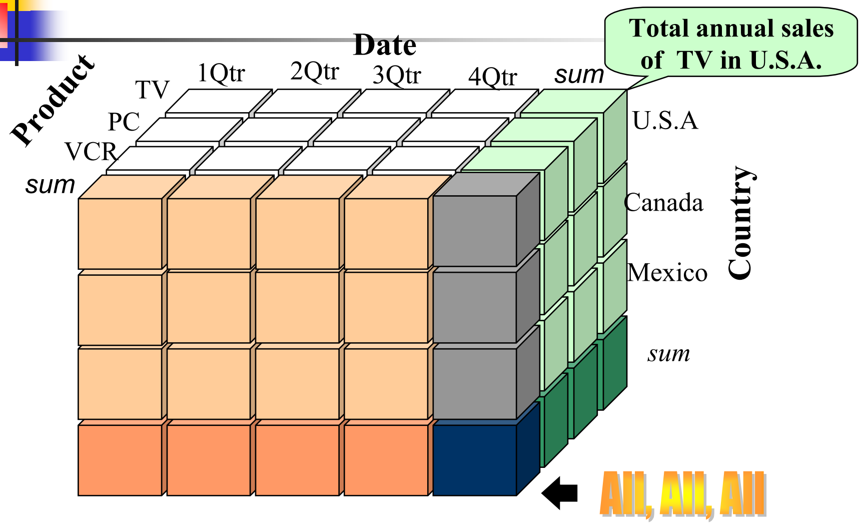
Data Cube



A data cube with 4 dimensions:

- Time
- Item
- Location
- Supplier

A Sample Data Cube



Total annual sales of TV in U.S.A.

OLAP - online analytical processing

- A popular approach for analysis of data warehouses.
- Provides multidimensional data analysis, superior to SQL.
- Main OLAP operations:
  - **Slice**: Extraction of summarized (aggregate) information for a given dimension value, from a data cube.
  - **Dice**: Extraction of a "subcube" or intersection of several slices.
  - **Pivot**: Exchange of rows and columns in a cross-tab table.
  - **Drill Down**: Present data at a more specific level of abstraction.
  - **Roll Up**: Present data at a more general level of abstraction (or granularity).

OLAP Services

- **Microsoft® SQL Server™ OLAP Services** is a middle-tier server for online analytical processing (OLAP), packaged together with the **SQL Server**
- The OLAP Services system includes a powerful server that constructs multidimensional cubes of data for analysis and provides rapid client access to cube information.
- We shall use its **Analyses Manager** to build a data cube

Building Data Cubes Using Analyses Manager

1. Install SQL Server and its OLAP Services
2. Set Up the System Data Source Connection
3. Start Analysis Manager
4. Set Up the Database and Data Source
5. Build a Cube

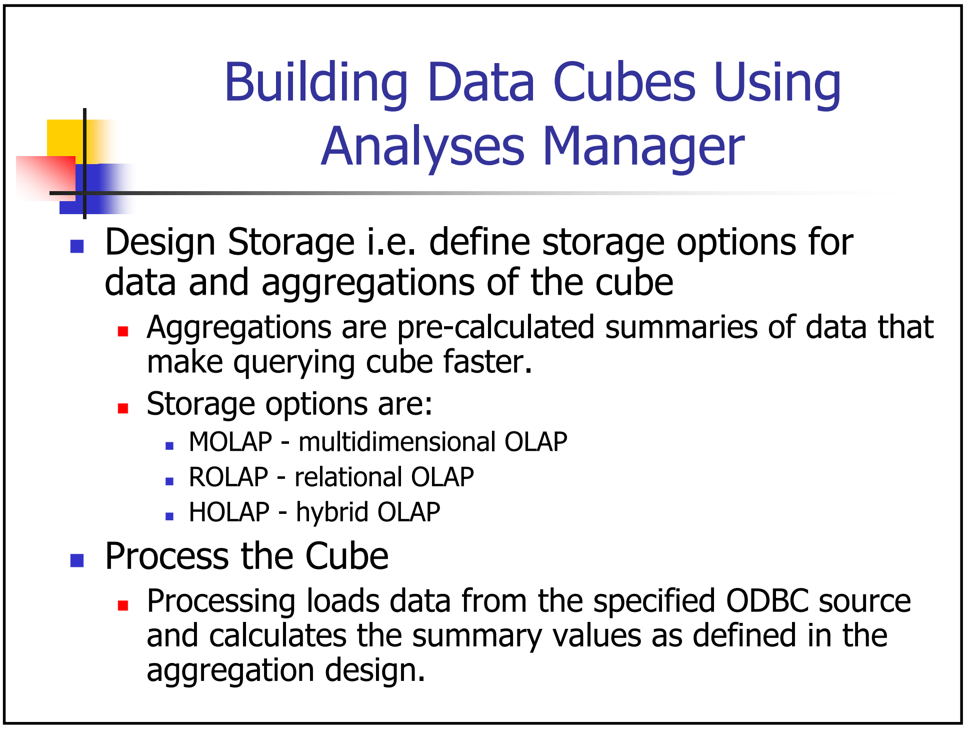
Building Data Cubes Using Analyses Manager

- **Scenario:**
  - Imagine that you are a database administrator working for the FoodMart corporation.
  - FoodMart is a large grocery store chain with sales in the United States, Mexico, and Canada.
  - The marketing department wants to analyze all of the sales by products and customers that were made during the 2009 calendar year.
  - Using data stored in the company's data warehouse, you should build a data cube (Sales) to enable fast response times for marketing analyses.

Building Data Cubes Using

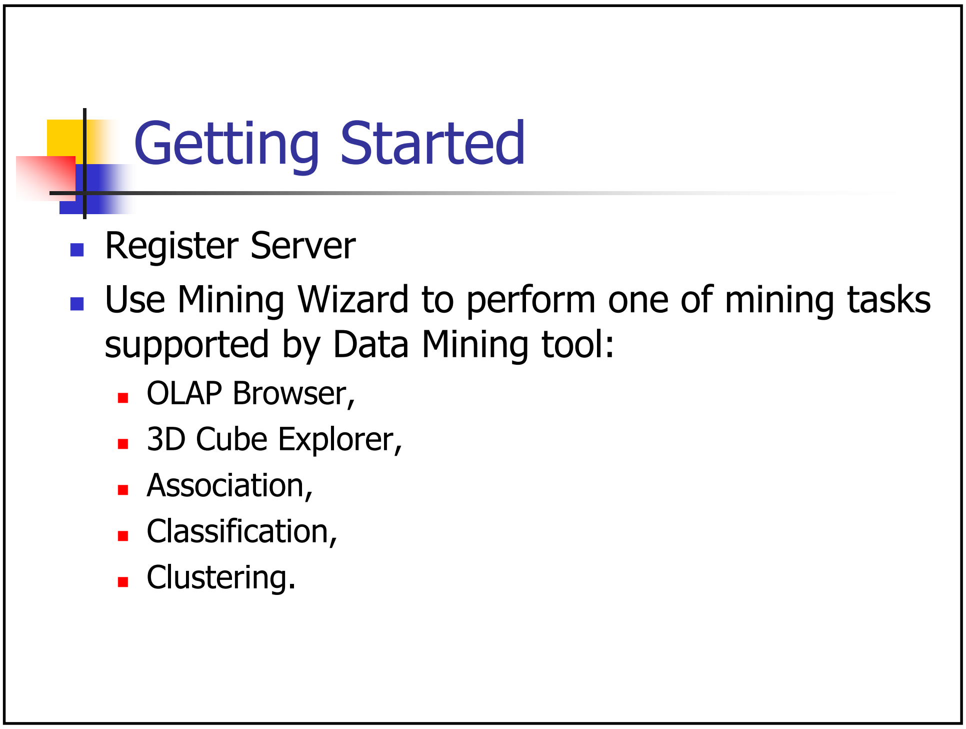
analyze all of the sales by products and customers that were made during the 2009 calendar year

- To build data cube named "Sales" one should specify:
  - One fact table, e.g. sales_fact_2009
  - Dimensions (descriptive business data), e.g.:
    - Time
    - Product
    - Customer
  - Measures (quantitative data), e.g.:
    - Store sales
    - Store cost
    - Unit sales



Building Data Cubes Using Analyses Manager

- Design Storage i.e. define storage options for data and aggregations of the cube
  - Aggregations are pre-calculated summaries of data that make querying cube faster.
  - Storage options are:
    - MOLAP - multidimensional OLAP
    - ROLAP - relational OLAP
    - HOLAP - hybrid OLAP
- Process the Cube
  - Processing loads data from the specified ODBC source and calculates the summary values as defined in the aggregation design.



Getting Started

- Register Server
- Use Mining Wizard to perform one of mining tasks supported by Data Mining tool:
  - OLAP Browser,
  - 3D Cube Explorer,
  - Association,
  - Classification,
  - Clustering.

Associations

- **Association** mining on a set of data looks for values in different dimensions (attributes) that commonly occur together, suggesting an association between them.
- Assume you have two attributes *A* and *B*, then a typical ***association rule*** is given by:

  $A_1 \wedge A_2 \wedge \ldots \wedge A_n \text{ -> } B_1 \wedge B_2 \wedge \ldots \wedge B_m$

  where A_i and B_j are attribute values.

  such a rule can be interpreted as:

  "If A_1 and A_2 ... A_n occur, then it is *often* the case that B_1 and B_2 ... B_m also occur in the same transaction."

Associations

- Three kinds of association are possible:

  1. ***Inter-dimensional association***. Associations among or across two or more dimensions, e.g.

     Customer-Country("Canada") → Product-SubCategory("Coffee").

     i.e. Canadian customers are likely to buy coffee.

  2. ***Intra-dimensional association***. Associations present within one dimension grouped by another one or several dimensions, e.g.

     Within Customer-Country("Canada"):

     Product-ProductName("CarryBags") → Product-ProductName("Tents")

     i.e. Customers in Canada who buy carry-bags, are also likely to buy tents.

Associations

- Three kinds of association are possible:

3. ***Hybrid association***. This method of association combines elements of both inter- and intra-dimensional mining, e.g.

   Within Customer-Country("Canada"):
   Product("Carry Bags") → Product("Tents"), Time("Q3")

   i.e. Customers in Canada who buy carry-bags, also tend to buy tents and do so most often in the 3rd quarter of the year (Jul, Aug, Sep).

- Association rules can span multiple *levels of abstraction*, i.e. conceptual (or hierarchical) levels in a data cube.

  e.g. Customer-Country("Canada") → Product-SubCategory("Coffee")

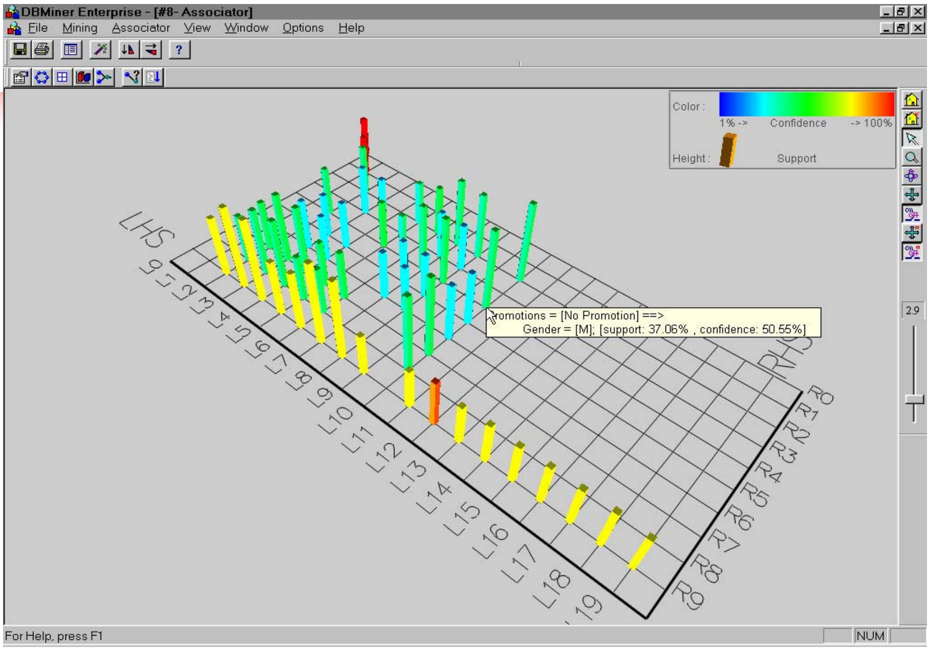
Associations

- Setting constraints
  - Association rules can be focused by specifying one or more constraints
  - A constraint specifies a dimension value that must appear in any association rule subsequently generated.
  - If two or more constraints are chosen, then each generated rule must *contain at least one* of those constraints in either its body or head clause.
  - The user may further require that *only* specified constraints are used in rule generation -> the generated rules will consist only of clauses containing one or more of those constraints.

Associations

- Setting support and confidence thresholds
  - Given the rule A → B:
    - ***support*** is the probability that a transaction contains A ∪ B (designates frequency of the implication).
    - ***confidence*** is the probability that a transaction containing A, also contains B (designates strength of the implication).
  - By adjusting a rule's support and confidence thresholds, one can vary the number of associations found to satisfy it.

Display of Association Rules in Rule Plane Form



Promotions = [No Promotion] ==>
Gender = [M]; [support: 37.06% , confidence: 50.55%]

Classification

- Data Mining Tool’s **Classification Module** analyzes a set of training data (i.e. a set of objects whose class label is known) and constructs a model for each class based on the features in the data.
- Classification rules resulting from the classification process can be used to:
  - classify future data,
  - develop a better understanding of each class in the database.

Classification

- The classification method consists of four steps:
  - Partitioning a relevant set of data into training and testing data.
  - Analysis of relevance of the dimensions involved.
    - determines the relevance of an attribute for the classification. Only a few of the top-most relevant attributes are retained, while the weakly relevant or irrelevant attributes are no longer considered.
  - Construction of the classification (decision) tree.
  - Testing the effectiveness of the classification using the test data set.

Classification

- The Classification module uses two thresholds to facilitate statistical analysis:
  - classification threshold - helps justification of the classification at a node when a significant set of the examples belong to the same class;
  - noise threshold - helps ignore a node in classification if it contains only a negligible number of examples.

Classification

- A classification tree is a hierarchical structure consisting of a set of pie charts and the branches (links) between them.
- Each node indicates the distribution of the classes (classification attribute) at that particular node.
- Changing classification settings user can modify classification results
  - e.g. changing hierarchical levels of dimensions to get a more detailed or more general tree.

Clustering

- A data mining task that maps a data item into one of several categorical classes (or clusters) in which the classes must be determined from the data (unlike classification in which the classes are predefined).
- Data Mining tools typically provide a **Clustering Module** that performs this DM task.
  - Only two cube dimensions can be chosen in a mining session since the clustering space is a 2-dimensional plane.

Clustering

- The clustering algorithm is the K-means method.
- The K-means method takes an input parameter *k,* which indicates the number of clusters the user wants to form.
- Initially, k values (points) are chosen at random from the set of all data points to represent the centre (mean) value of each cluster. Then every other point on the plane is assigned to the cluster it is closest to.
- The "closest cluster" is determined by the shortest distance from a point to the mean value of each cluster, using formula:

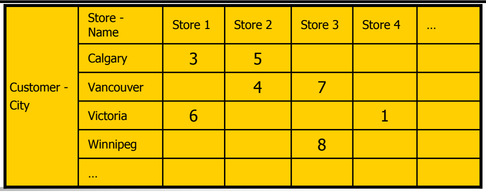
  $d = \surd\ (\alpha(x1 - x2)^2 + \beta(y1 - y2)^2)$

  where α and β are coefficients with a default value of 1.000

Clustering

- Once all points on the plane have been assigned, the centre of each cluster is re-calculated by taking the mean of all points in that cluster. Then new clusters are formed based around the new centers.
- This process repeats until:
  - no further re-distribution of points occurs or
  - a user-specified number of iterations is reached.

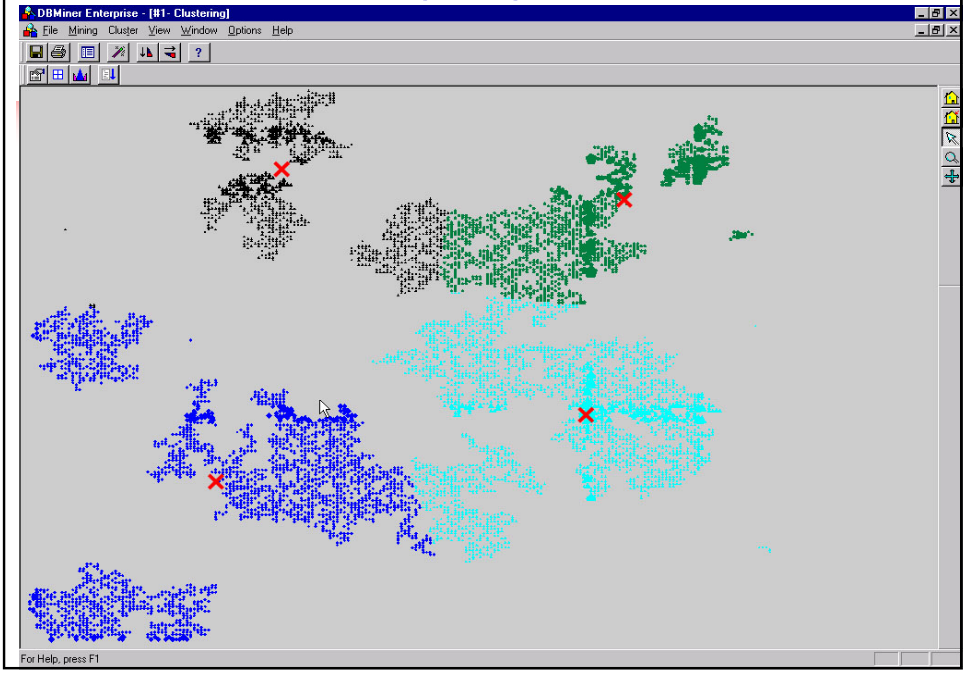
Clustering



	Store - Name	Store 1	Store 2	Store 3	Store 4	…
Customer - City	Calgary	3	5			
	Vancouver		4	7		
	Victoria	6			1	
	Winnipeg			8		
	…					

- The K-means method is designed to run on continuous data, however a majority of data cubes' data is categorical.
  - Problem: how to measure the distance between say, a customer who lives in Calgary and shops at Store 12 and the one who lives in Vancouver and shops at Store 5.
- Data Mining tools handle this problem by creating a table…
- Every non-empty cell in this table appears in the clustering visualization
- The size of the icon used to visualize a cell (e.g. ◆) indicates the count of data tuples in that cell.

Display of Clustering (Segmentation) Results



WWW Resources

- Data sets:

  http://www.mlnet.org/resources/datasets-index.html

- DBMiner Official site:

  http://www.dbminer.com/

Building data cubes and mining them

Jelena Jovanovic
Email: jeljov@fon.bg.ac.yu