Data Analytics

⚠Overview of Data Analytics

Data analytics: the processing of data to infer patterns, correlations, or models for prediction

Common steps in data analytics:

[1] Gather data from multiple sources into one location

[2] Generate aggregates and reports summarizing data

[3] Build predictive models and use the models for decision making

▲Related Terms [相关术语]:

•Machine learning techniques are key to finding patterns in data and making predictions

•Data mining extends techniques developed by machine-learning communities to run them on very large datasets

•The term business intelligence is a synonym for data analytics

•The term decision support focuses on reporting and aggregation

⚠Data Warehousing [数据仓储]

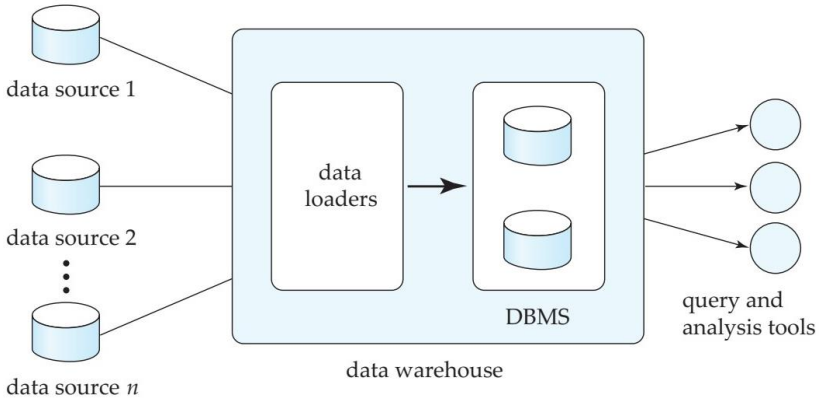
•Data sources often store only current data, not historical data

•Corporate decision making requires a unified view of all organizational data, including historical data

•A data warehouse is a repository (archive) of information gathered from multiple sources, stored under a unified schema, at a single site

Greatly simplifies querying, permits study of historical trends

Shifts decision support query load away from transaction processing systems



▲Design Issues:

♦When and how to gather data:

Source driven architecture: data sources transmit new information to warehouse either continuously or periodically (e.g., at night)

Destination driven architecture: warehouse periodically requests new information from data sources

Synchronous vs asynchronous replication:

•Keeping warehouse exactly synchronized with data sources (e.g. using two-phase commit) is often too expensive

•Usually OK to have slightly out-of-date data at warehouse

•Data/updates are periodically downloaded from online transaction processing (OLTP) systems.

♦What schema to use

Schema integration

♦Data transformation and data cleansing

♦How to propagate updates

♦What data to summaries

▲Multidimensional Data and Warehouse Schemas

Data in warehouses can usually be divided into

Fact tables, which are large

e.g. sales(item_id, store_id, customer_id, date, number, price)

Dimension tables, which are relatively small

Store extra information about stores, items, etc.

Attributes of fact tables can be usually viewed as

Measure attributes

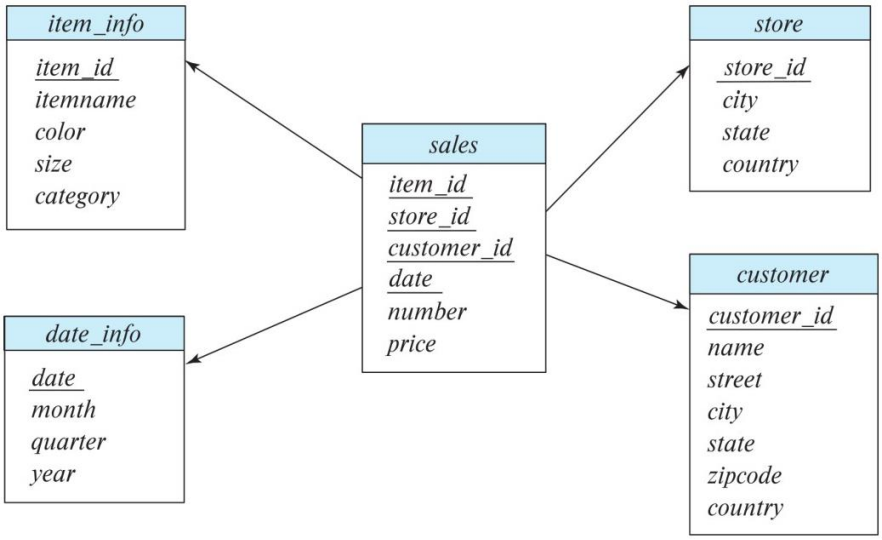
measure some value, and can be aggregated [聚合的] upon

e.g. the attributes number or price of the sales relation

Dimension attributes

dimensions on which measure attributes are viewed

e.g. attributes item_id, color, and size of the sales relation



Resultant schema is called a star schema

More complicated schema structure: Snowflake schema with multiple levels of dimension tables

Typically:

fact table joined with dimension tables and then

group-by on dimension table attributes, and then

aggregation on measure attributes of fact table

Some applications do not find it worthwhile to bring data to a common schema

Data lakes allow data to be stored in multiple formats, without schema integration [结合]

Data in warehouses usually append only, not updated

Data warehouses often use column-oriented storage [面向列的储存]

e.g. a sequence of sales tuples is stored as follows

Values of item_id attribute are stored as an array

Values of store_id attribute are stored as an array,

And so on

Arrays are compressed, reducing storage, IO and memory costs significantly

Queries can fetch only attributes that they care about, reducing IO and memory cost

Data warehouses often use parallel storage and query processing infrastructure

item_name	color	clothes_size	quantity
dress	dark	small	2
dress	dark	medium	6
dress	dark	large	12
dress	pastel	small	4
dress	pastel	medium	3
dress	pastel	large	3
dress	white	small	2
dress	white	medium	3
dress	white	large	0
pants	dark	small	14
pants	dark	medium	6
pants	dark	large	0
pants	pastel	small	1
pants	pastel	medium	0
pants	pastel	large	1
pants	white	small	3
pants	white	medium	0
pants	white	large	2
shirt	dark	small	2
shirt	dark	medium	6
shirt	dark	large	6
shirt	pastel	small	4
shirt	pastel	medium	1
shirt	pastel	large	2
shirt	white	small	17
shirt	white	medium	1
shirt	white	large	10
skirt	dark	small	2
skirt	dark	medium	5

⚠Online Analytical Processing

Online Analytical Processing (OLAP):

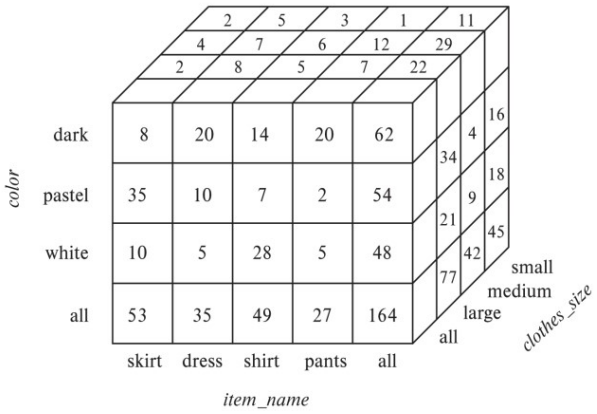
Interactive analysis of data, allowing data to be summarized and viewed in different ways in an online fashion (with negligible delay)

Data that can be modeled as dimension attributes and measure attributes are called multidimensional data.

As an example, given sales relation on clothes shop

sales(item-name, color, size, number)

A data cube is a multidimensional generalization of a cross-tab⬇



▲OLAP Operations

Pivoting [旋转]: changing the dimensions used in a cross-tab

e.g. moving colors to column names

Slicing [切片]: creating a cross-tab for fixed values only

e.g. fixing color to white and size to small

Sometimes called dicing, particularly when values for multiple dimensions are fixed.

Rollup [归纳]: moving from finer-granularity data to a coarser granularity

e.g. aggregating away an attribute

e.g. moving from aggregates by day to aggregates by month or year

Drill down [深化]: the opposite operation - that of moving from coarser-granularity data to finer-granularity data

▲Cross Tabulation of sales by item name and color

clothes_size **all**

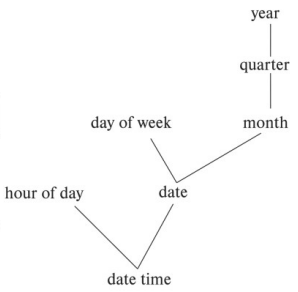
item_name \ *color*	dark	pastel	white	total
skirt	8	35	10	53
dress	20	10	5	35
shirt	14	7	28	49
pants	20	2	5	27
total	62	54	48	164

The table above is an example of a cross-tabulation (cross-tab), also referred to as a pivot-table.

After cross tabulation with hierarchy:

clothes_size: **all**

category	*item_name*	*color*			
		dark	pastel	white	total
womenswear	skirt	8	35	10	53
	dress	20	10	5	35
	subtotal	28	45	15	88
menswear	pants	14	7	28	49
	shirt	20	2	5	27
	subtotal	34	9	33	76
total		62	54	48	164



(a) time hierarchy

(b) location hierarchy

Use the value **all** to represent aggregates in cross-tabs:

item_name	*color*	*clothes_size*	*quantity*
skirt	dark	**all**	8
skirt	pastel	**all**	35
skirt	white	**all**	10
skirt	**all**	**all**	53
dress	dark	**all**	20
dress	pastel	**all**	10
dress	white	**all**	5
dress	**all**	**all**	35
shirt	dark	**all**	14
shirt	pastel	**all**	7
shirt	white	**all**	28
shirt	**all**	**all**	49
pants	dark	**all**	20
pants	pastel	**all**	2
pants	white	**all**	5
pants	**all**	**all**	27
all	dark	**all**	62
all	pastel	**all**	54
all	white	**all**	48
all	**all**	**all**	164

▲Data Mining

♦Data mining is the process of (semi-) automatically analyzing large databases to find useful patterns

♦Part of the larger area of knowledge discovery in databases (KDD)

♦Some types of knowledge can be represented as rules

♦Knowledge is discovered with machine learning techniques on past data to form a model and then used

♦To make predictions for new, unseen instances. It is the exploration and analysis of large amount of data in order to discover valid, novel, potentially useful, and ultimately understandable patterns in data.

Valid [有效的] - patterns hold in general.

Novel [新颖的] - did not know the pattern beforehand.

Useful [有用的] - can devise actions from the patterns.

Understandable [可理解的] - can interpret and comprehend the patterns.

▲Decision Trees

Each internal node of the tree partitions the data into groups based on a partitioning attribute, and a partitioning condition

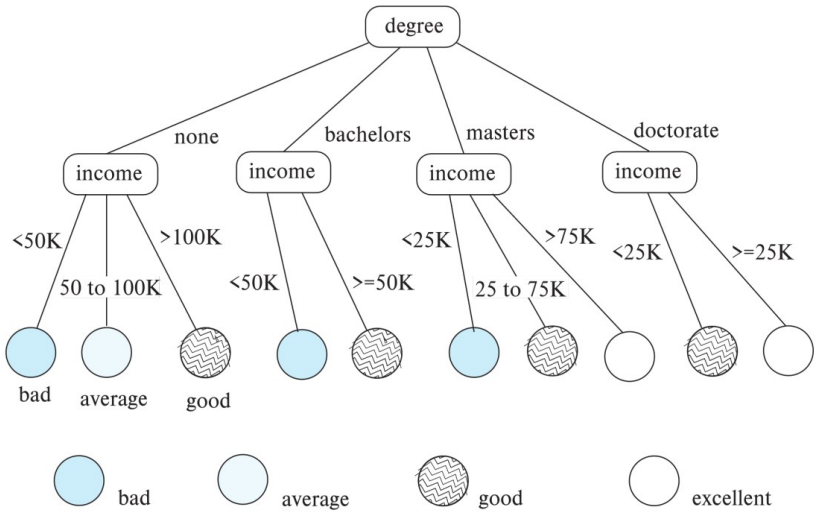
Leaf node:

all (or most) of the items at the node belong to the same class, or

all attributes have been considered, and no further partitioning is possible.

Traverse tree from top to make a prediction

<Decision Tree Classifiers>



Data Mining Tasks:

Classification

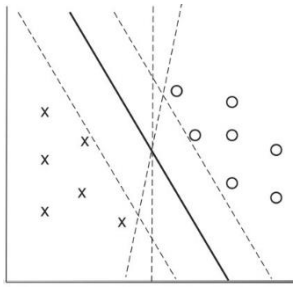
Bayesian classifiers use Bayes theorem:

$$p(c_j \mid d) = \frac{p(d \mid c_j)\, p(c_j)}{p(d)}$$

where

- $p(c_j \mid d)$ = probability of instance d being in class c_j,
- $p(d \mid c_j)$ = probability of generating instance d given class c_j,
- $p(c_j)$ = probability of occurrence of class c_j, and
- $p(d)$ = probability of instance d occuring

Support Vector Machine



SVM example with 2-dimensional case

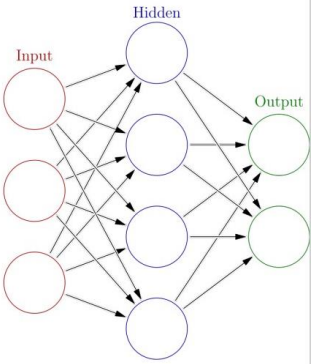
Points are in two classes

Find a line (maximum margin line)

In n-dimensions points are divided by a hyperplane, instead of a line

Neural Network

- Neural network has multiple layers
  - Each layer acts as input to next later
- First layer has input nodes, which are assigned values from input attributes
- Each node combines values of its inputs using some weight function to compute its value
  - Weights are associated with edges
- For classification, each output value indicates likelihood of the input instance belonging to that class
  - Pick class with maximum likelihood
- Weights of edges are key to classification
  - Edge weights are learned during training phase



Deep Neural Network

Deep learning refers to training of deep neural network on very large numbers of training instances

▲Regression

- Regression deals with the prediction of a value, rather than a class.
  - Given values for a set of variables, $X_1, X_2, \ldots, X_n$, we wish to predict the value of a variable Y.
- One way is to infer coefficients $a_0, a_1, a_1, \ldots, a_n$ such that

$$Y = a_0 + a_1 * X_1 + a_2 * X_2 + \ldots + a_n * X_n$$

Regression aims to find coefficients that give the best possible fit

Descriptive Patterns:

▲Association Rules [关联规则]

- Association rules examples
  - *bread* ⇒ *milk*
  - *(DB-Concepts, OS-Concepts)* ⇒ Networks
- Left hand side is called **antecedent**, right hand side **consequent**
- An association rule must have an associated **population**; the population consists of a set of **instances**
  - e.g. each transaction (sale) at a shop is an instance, and the set of all transactions is the population

▲Clustering [聚类]

Clustering intuitively finds clusters of points in the given data such that similar points lie in the same cluster

Some instance of collaborative filtering:

- Goal example: predict what movies/books a person may be interested in, on the basis of
  - Past preferences of the person
  - Preferences of other people
- One approach based on **repeated clustering**
  - Cluster people based on their preferences for movies
  - Then cluster movies on the basis of being liked by the same clusters of people
  - Again cluster people based on their preferences for (the newly created clusters of) movies
  - Repeat above till equilibrium
  - Given new user
    - Find most similar cluster of existing users and
    - Predict movies in movie clusters popular with that user cluster

• • •

Other Types of Mining

Text mining

Application of data mining to textual documents

Sentiment analysis

e.g. learn to predict if a user review is positive or negative about a product

Information extraction

Create structured information from unstructured textual description or semi-structured data such as tabular displays

Entity recognition and disambiguation

e.g. given text with name “Michael Jordan”, does the name refer to the famous basketball player or the machine learning expert?

Knowledge graph

Can be constructed by information extraction from different sources