EKSPLORASI DATA

Apa itu eksplorasi data ?

- **Merupakan eksplorasi pendahuluan terhadap data untuk pemahaman yang baik mengenai karakteristiknya**
- Tujuan utama eksplorasi data di antaranya:
  - Membantu pemilihan tools yang tepat untuk proses preprocessing atau analisis
  - Mempermudah user untuk mengenali pola yang tidak tergambar oleh tools analisis data
- Berhubungan dengan Exploratory Data Analysis (EDA) yang diperkenalkan oleh ahli statistik John Tukey

Teknik-teknik Eksplorasi Data

- Summary statistics
- Visualisasi
- Online Analytical Processing (OLAP)
- Clustering dan anomaly detection → menurut EDA Tukey

Contoh Dataset Iris

- Merupakan dataset mengenai bunga iris yang sangat populer dan banyak digunakan teknik eksplorasi data.
  - Sumber : UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html
  - From the statistician Douglas Fisher
  - Terdapat 3 tipe bunga (kelas)
    - Setosa
    - Virginica
    - Versicolour
  - 4 Atribut
    - Sepal width
    - Sepal length
    - Petal width
    - Petal length

Iris setosa

Iris Versicolor

Iris Virginica

Summary Statistics

- Merupakan angka-angka yang meringkaskan properti dari data seperti:
  - Frekuensi, lokasi dan sebaran
    - Contoh: lokasi – mean (nilai tengah)
      sebaran- standar deviasi

Frekuensi, Mode, Persentil

- Penggunaan Frekuensi dan mode untuk data kategoris
- Frekuensi dari nilai atribut adalah persentase jumlah kemunculan nilai pada data set
  - For example, given the attribute ‘gender’ and a representative population of people, the gender ‘female’ occurs about 50% of the time.
- Mode sebuah atribut adalah nilai atribut yang paling sering muncul
- Persentil adalah nilai dalam skala seratus yang menunjukkan distribusi sama atau lebih
  - berguna untuk data kontinu

Mean dan Median

- Mean adalah is the most common measure of the location of a set of points.
- Mean sangat sensitif terhadap outlier.

$$\text{mean}(x) = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

Range dan Variance

- Range adalah selisih antara nilai max dan min
- variance atau standar deviasi merupakan pengukuran untuk melihat persebaran data point.

$$\text{variance}(x) = s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})^2$$

- Karena variance juga sensitif terhadap outliers, sehingga ada perhitungan lain:

$$\text{AAD}(x) = \frac{1}{m}\sum_{i=1}^{m}|x_i - \overline{x}|$$

$$\text{MAD}(x) = median\Big(\{|x_1 - \overline{x}|, \ldots, |x_m - \overline{x}|\}\Big)$$

$$\text{interquartile range}(x) = x_{75\%} - x_{25\%}$$

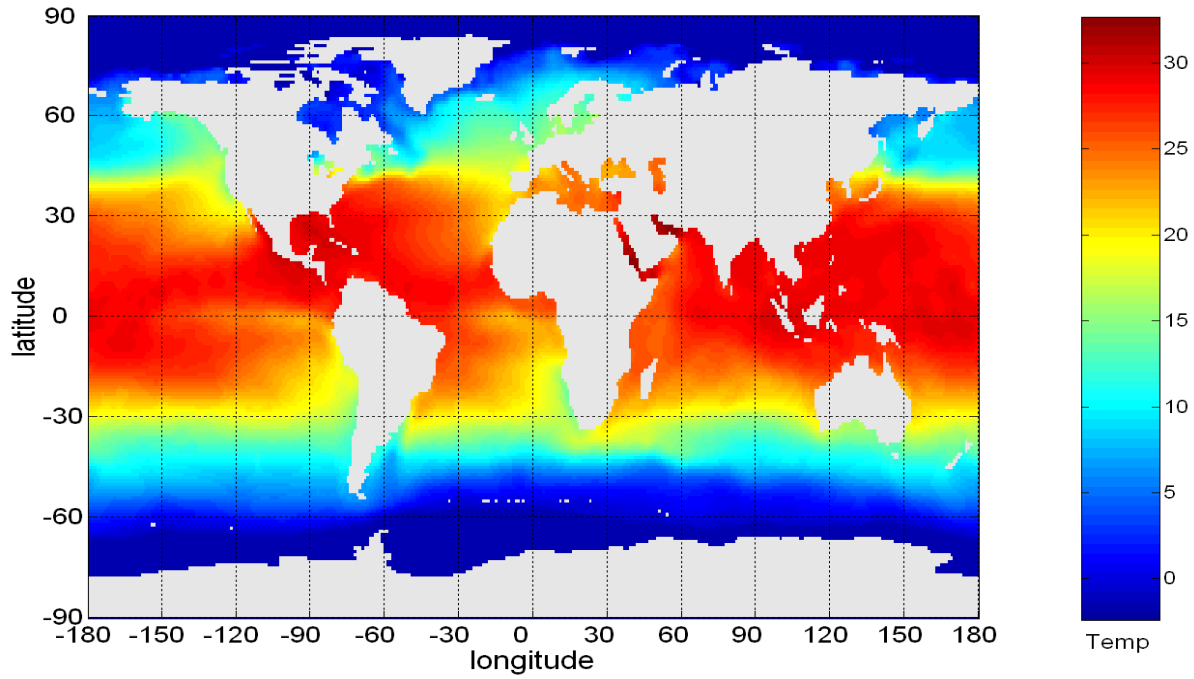
Visualisasi

Visualisasi adalah konversi data menjadi bentuk gambar atau tabular sehingga karaktristik data dan hubungan antar item data atau atribut dapat dianalisis atau dilaporkan

- Visualisasi merupakan salah satu teknik eksplorasi data yang paling powerful
  - Manusia lebih mudah menganalisis data dengan jumlah besar melalui representasi visual
  - Dapat mendeteksi pola umum dan tren
  - Dapat mendeteksi outliers dan pola khusus (unsual)

Contoh: Suhu permukaan laut

- Gambar berikut menunjukkan suhu permukaan laut pada Juli 1982
  - 10 ribu data point diringkas dalam satu gambar



Representasi

- Adalah pemetaan informasi menjadi format visual
- Objek data, atribut, dan hubungan antar objek data diterjemahkan menjadi elemen grafis seperti point, garis, bentuk , dan warna.

- Contoh:
  - Objek direpresentasikan sebagai points
  - Nilai atributnya direpresentasikan sebagai posisi points atau karakteristik points, misal warna, ukuran, dan bentuk
  - Jika posisi digunakan, maka hubungan antar point (outlier atau tidak) akan terlihat jelas misal jika point mengumpul atau tidak

Arrangement

- Merupakan penempatan penampakan elemen visual sehinga pemahaman data bisa lebih baik
- contoh:

	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	1	0	0	1
3	0	1	0	1	1	0
4	1	0	1	0	0	1
5	0	1	0	1	1	0
6	1	0	1	0	0	1
7	0	1	0	1	1	0
8	1	0	1	0	0	1
9	0	1	0	1	1	0

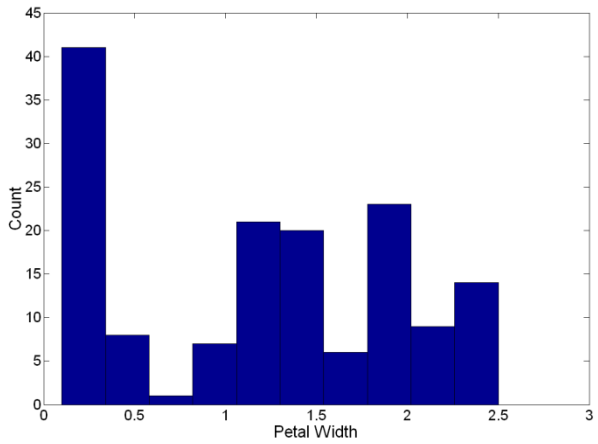
	6	1	3	2	5	4
4	1	1	1	0	0	0
2	1	1	1	0	0	0
6	1	1	1	0	0	0
8	1	1	1	0	0	0
5	0	0	0	1	1	1
3	0	0	0	1	1	1
9	0	0	0	1	1	1
1	0	0	0	1	1	1
7	0	0	0	1	1	1

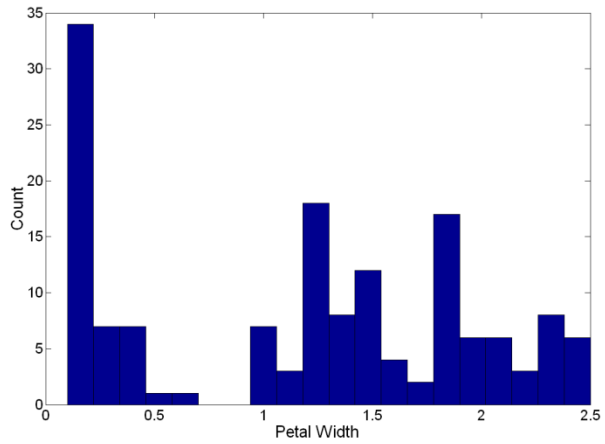
Seleksi

- Penghilangan atau de-emphasis objek dan atribut tertentu
- Seleksi akan melibatkan pemilihan subset atribut
  - Pengurangan Dimensi
- Seleksi juga akan melibatkan pemilihan subset objects

Teknik Visualisasi : Histograms

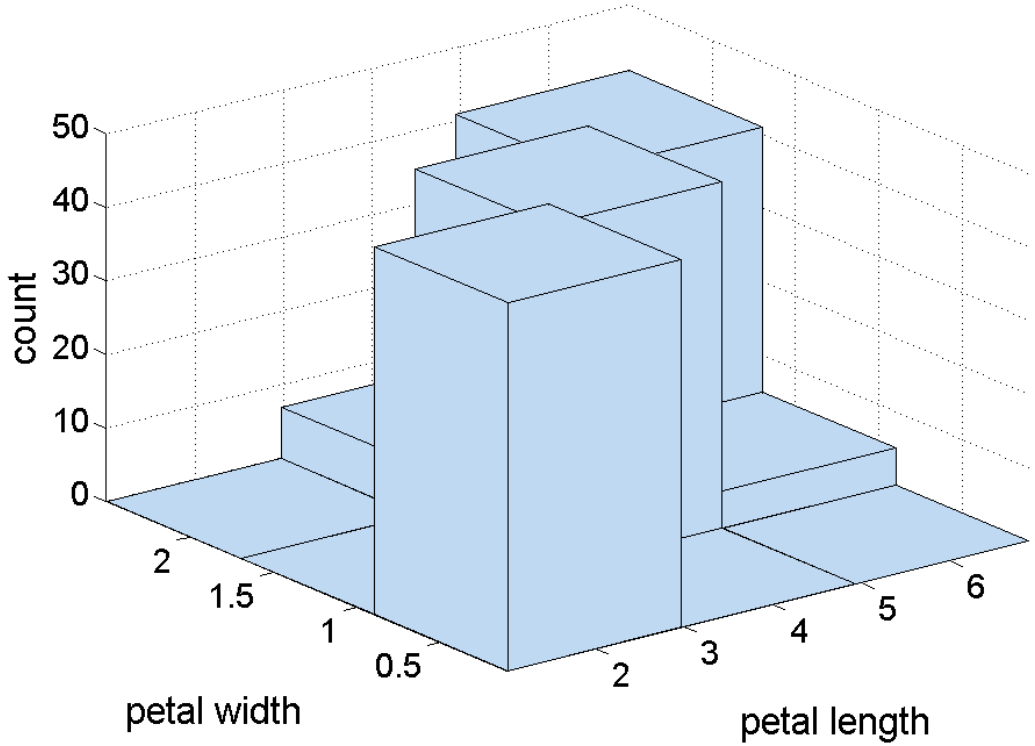
- Histogram
  - Digunakan untuk menunjukkan distribusi nilai suatu variabel
  - Membagi nilai menjadi bins dan menunjukkan bar plot jumlah objects di setiap bin.
  - Tinggi setiap bar mengindikasikan jumlah objek
  - Ukuran histogram tergantung jumlah bins
- Contoh: Petal Width (10 dan 20 bins)





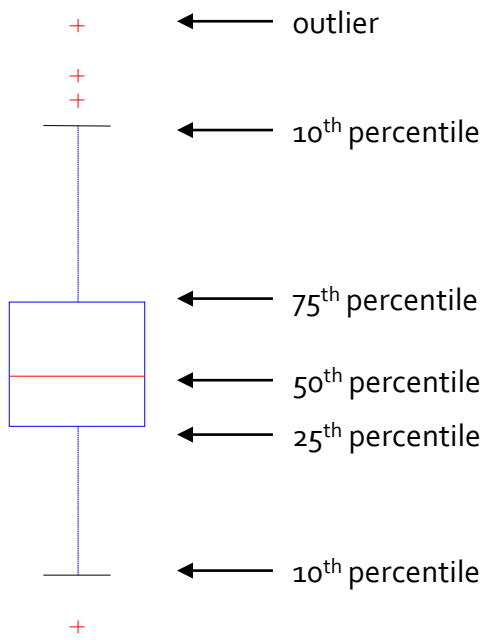
Histogram 2 Dimensi

- Menampilkan joint distribution dari nilai dua atribut
- contoh: petal width dan petal length
  - What does this tell us?



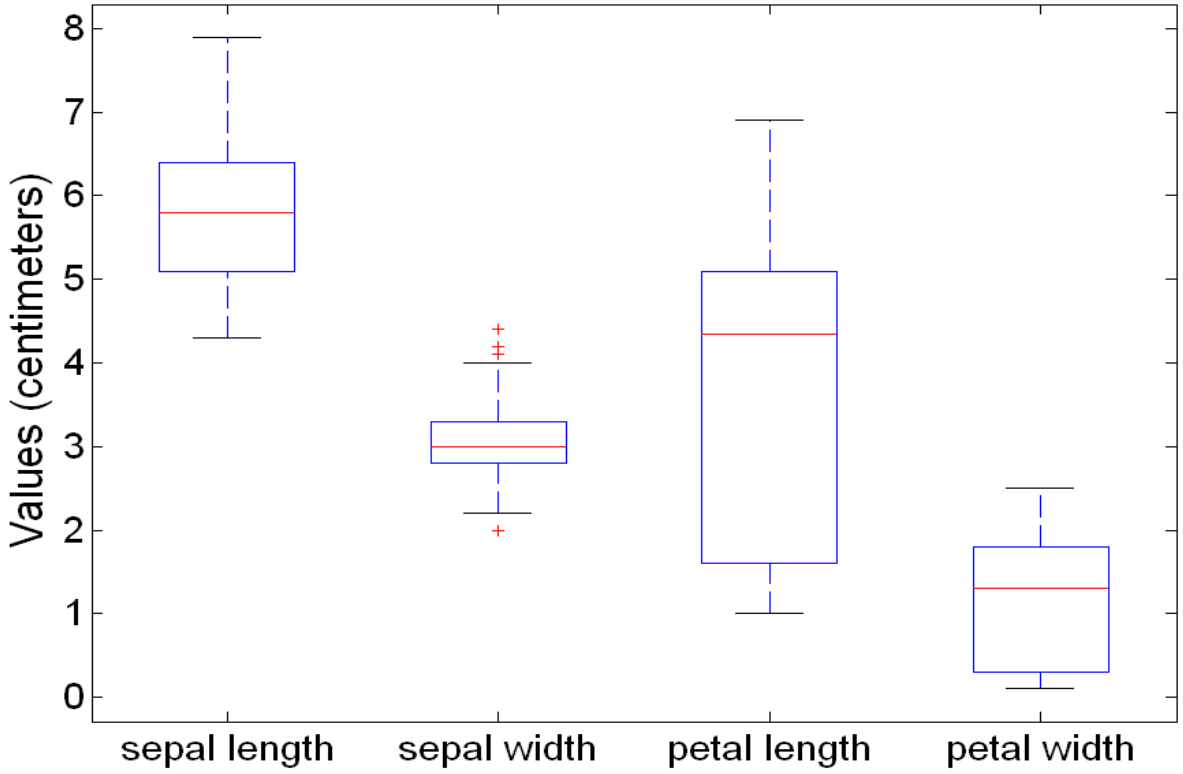
Teknik Visualisasi : Box Plots

- Box Plots
  - Ditemukan oleh J. Tukey
  - Cara menampilkan distribusi data



Contoh Box Plots

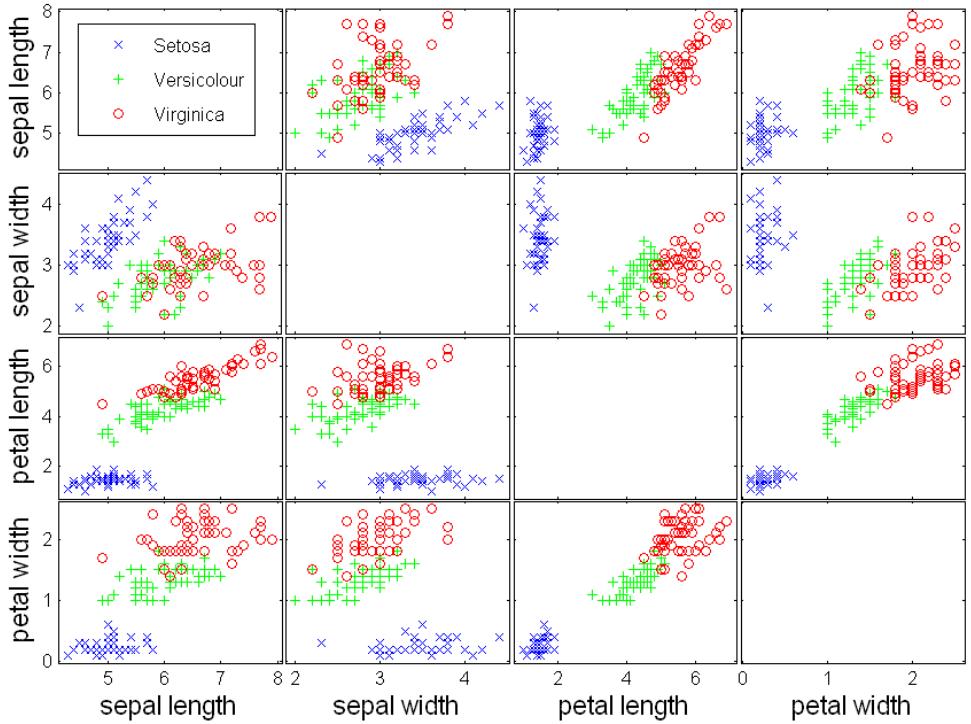
- Box plots dapat digunakan untuk membandingkan atribut



Teknik Visualisasi : Scatter Plots

- Scatter plots
  - Nilai Atribut menentukan posisi
  - Biasanya dalam 2 dimensi namun bisa juga dalam 3 dimensi
  - Sebaiknya menggunakan tampilan yang berbeda dari segi ukuran, bentuk, dan warna untuk representasi objek sehingga dapat dengan mudah dianalisis

Scatter Plot Atribut Iris



OLAP

- On-Line Analytical Processing (OLAP) diperkenalkan oleh E. F. Codd (Bapak Relational database)
- Relational databases menempatkan data menjadi tabel, sedangkan OLAP menggunakan representasi multidimensional array
- Dengan representasi data tsb, analisis dan ekplorasi data dapat dengan mudah dilakukan

Membuat Multidimensional Array

- 2 langkah mengkonversikan data tabel menjadi multidimensional array.
  - *pertama*, identifikasi atribut yang akan menjadi dimensi dan atribut yang menjadi target yang mana nilainya muncul sebagai entries pada multidimensional array.
    - Atribut yang digunakan sebagai dimensi harus memiliki nilai diskrit
    - Nilai target biasanya count atau nilai kontinu mis., harga barang
    - Bisa saja variabel target tidak ada kecuali jumlah objek yang memiliki nilai atribut yang sama
  - *Kedua*, cari nilai setiap entry pada multidimensional array dengan menjumlahkan nilai (dari atribut target) atau hitung seluruh objek yang memiliki nilai atribut yang berkorespondensi dengan entri tsb.

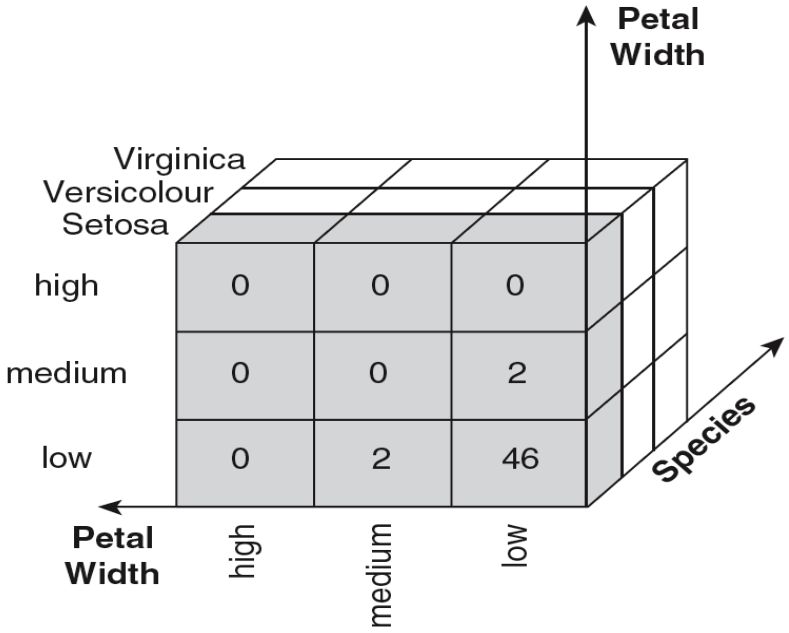
Contoh: Data Iris

- Atribut petal length, petal width, dan species type dapat dikonversi menjadi multidimensional array
  - *Pertama*, diskritkan petal width dan petal length sehingga memiliki nilai kategoris : *low*, *medium*, dan *high*
  - Didapat tabel sbb- (lihat atribut count)

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

Contoh: data Iris (lanjt)

- Tiap tuple unik dari petal width, petal length, dan species type mengidentifikasikan satu elemen dari array.
- Elemen ini dihubungkan dengan nilai count.
- Semua non-specified tuples =0.



Contoh: data Iris (lanjt)

- Slices dari multidimensional array dapat terlihat pada cross-tabulations
- Informasi apa yang bisa didapat dari tabel ini?

		Width		
		low	medium	high
Length	low	46	2	0
	medium	2	0	0
	high	0	0	0

Iris-setosa

		Width		
		low	medium	high
Length	low	0	0	0
	medium	0	43	3
	high	0	2	2

Iris-virginica

		Width		
		low	medium	high
Length	low	0	0	0
	medium	0	0	3
	high	0	3	44

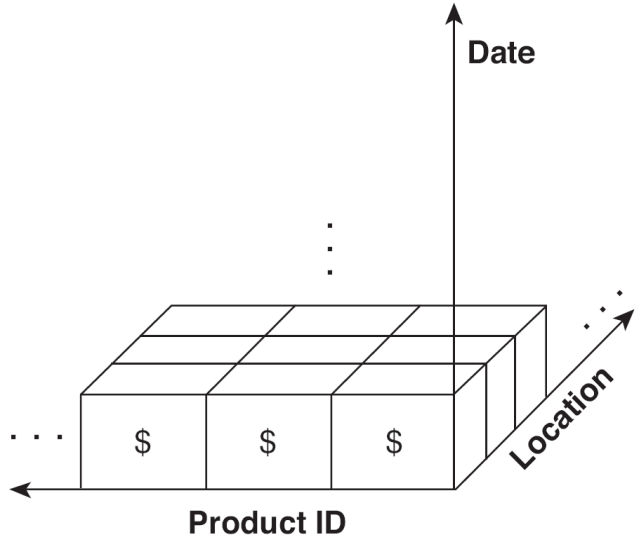
Iris-versicolor

Operasi OLAP : Data Cube

- Operasi kunci OLAP adalah formasi data cube
- data cube merupakan representasi multidimensional data, dengan semua kemungkinan agregasinya

Contoh Data Cube

- Misalkan terdapat data mengenai penjualan produk pda perusahaan dagang pada waktu yang berbeda-beda
- Data ini dapat direpresentasikan menjadi 3 dimensional array
- Ada 3 two-dimensional aggregates (3 choose 2), 3 one-dimensional aggregates, dan 1 zero-dimensional aggregate (total keseluruhan)



Contoh Data Cube (lanjt)

- Gambar tabel berikut menunjukkan satu dari 2 dimensional aggregates, 2 dari one-dimensional aggregates, dan total keseluruhan

		date				
		Jan 1, 2004	Jan 2, 2004	...	Dec 31, 2004	total
product ID	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

Operasi OLAP : Slicing dan Dicing

- Slicing adalah memilih group cells dari seluruh multidimensional array dengan menentukan nilai specifik untuk satu atau lebih dimensi
- Dicing merupakan pemilihan subset cells dengan menentukan range nilai atribut.
  - Ekivalen dengan menentukan subarray dari complete array.
- Dalam prakteknya, kedua operasi tsb dapat juga dilakukan dengan aggregasi beberapa dimensi.

Operasi OLAP : Roll-up dan Drill-down

- Nilai Atribute biasanya memiliki struktur hirarki.
  - Setiap tanggal diasosiasikan dengan tahun, bulan dan minggu
  - Lokasi diasosiasikan dengan benua, negara, propinsi dan kota
  - Produk dapat dikategorikan menjadi beberapa seperti pakaian, elektronik dan furnitur
- Perhatikan bahwa kategori tsb bersarang dan membentuk tree
  - Tahun terdiri dari bulan yang terdiri dari hari
  - Negara terdiri dari porpinsi yang terdiri dari kota

Operasi OLAP : Roll-up dan Drill-down

- struktur hirarki memungkinkan dilakukannya operasi roll-up dan drill-down.
  - Untuk data penjualan, dapat diaggregate (roll up) penjualan selama satu bulan
  - Sebaliknya, jika ada data dimana dimensi waktunya dibagi menjadi bulan, maka dapat dilakukan split penjulaan bulanan total (drill down) menjadi total penjualan harian.
  - Demikian juga, drill down atau roll up pada lokasi atau atribut product ID.