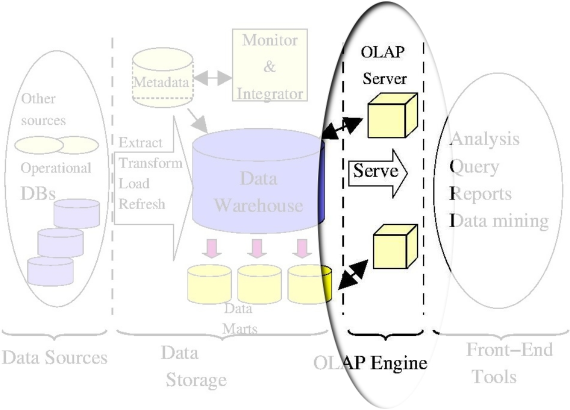
Datawarehouse and OLAP

OLAP

Syllabus, materials, notes, etc.

See http://www.info.univ-tours.fr/˜marcel/dw.html

On-Line Analytical Processing



today

OLAP?

analytical queries

informal model

typical treatments

context

datawarehouses gather

- large volumes of
- homogeneous
- usable
- multidimensionnal
- consolidated

data

how to analyse these data for decision-making purpose?

DW and OLAP

recall that

- OLTP queries are executed on the operational source databases
- the warehouse is refreshed periodically
- OLAP queries are executed on the warehouse data

what is On-Line Analytical Processing?

facilities to

- summarize and synthesize
- consolidate
- browse
- apply formula to

data according to many dimensions

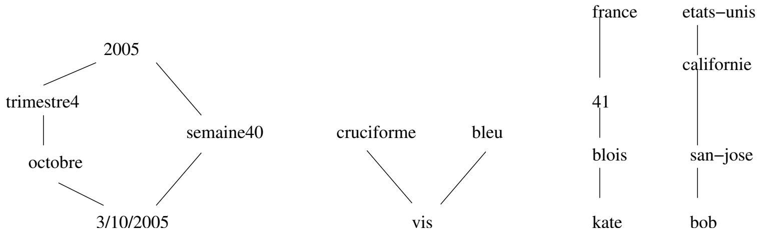
analytical queries

example: a star schema

(sorry, a french datawarehouse)

ventes(codeProduit, date, vendeur, montant)
produits(codeProduit, modèle, couleur)
vendeurs(noms, villes, départements, états, pays)
temps(jours, semaines, mois, trimestres, années)

hierarchies



2005

trimestre4

semaine40

octobre

3/10/2005

cruciforme

bleu

vis

france

etats–unis

californie

41

blois

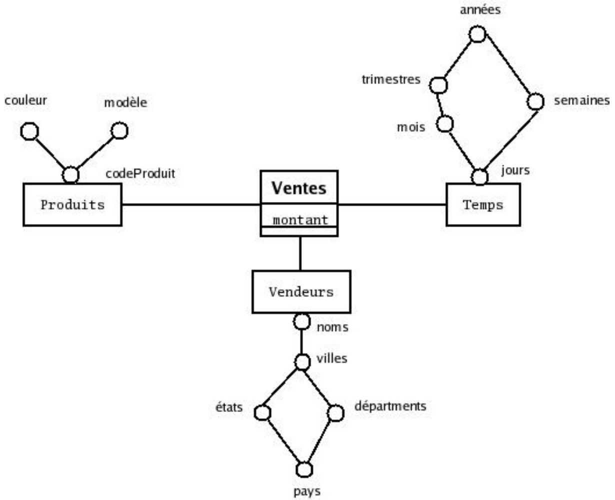
san–jose

kate

bob

conceptual model

Golfarelli (1998)



a typical star join query

```
SELECT     département, modèle, mois, AVG(montant)
FROM       ventes, vendeurs, produits, temps
WHERE      ventes.vendeur = vendeurs.noms
AND        ventes.codeProduit = produits.codeProduit
AND        ventes.date=temps.jours
AND        couleur = ”noir”
AND        années = ”2007”
GROUP BY   département,modèle, mois
HAVING     avg(montant) > 5000
ORDER BY   montant DESC;
```

Analytical query pattern

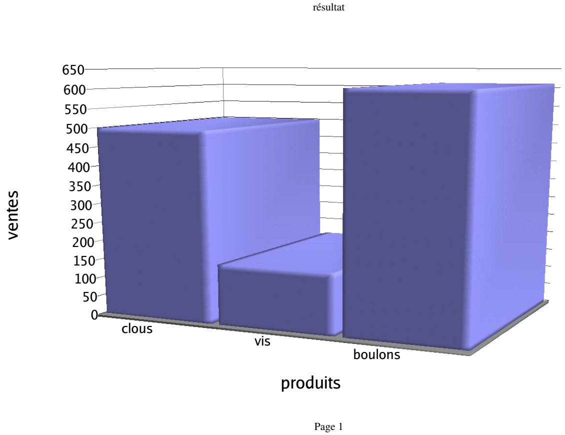
```
SELECT      dimensions, aggregates
FROM        fact table, dimension tables
WHERE       join conditions
AND         fixed-value conditions
GROUP BY    dimensions
HAVING      aggregate conditions
ORDER BY    aggregates;
```

example of a typical analysis

analyzing sales of various products

```
SELECT     modèle, SUM(montant)
FROM       ventes, produits
WHERE      ventes.codeProduit = produits.codeProduit
GROUP BY   modèle ;
```

example of a typical analysis



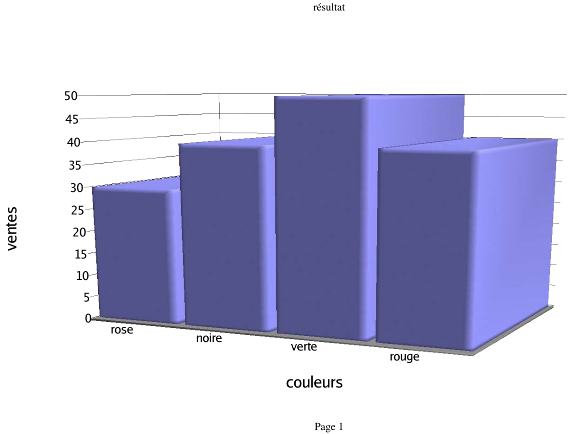
example of a typical analysis

sales of screws (vis) are lower than expected

is it due to one particular color?

```
SELECT     couleur, SUM(montant)
FROM       ventes, produits
WHERE      ventes.codeProduit = produits.codeProduit
AND        modèle = “vis”
GROUP BY   couleur ;
```

example of a typical analysis

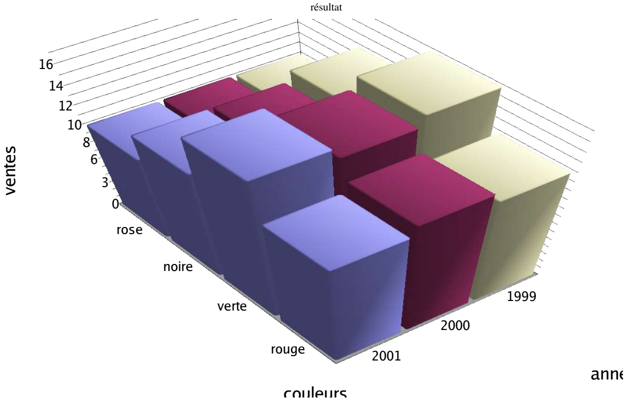


example of a typical analysis

is it for a particular year?

```
SELECT     couleur, années, SUM(montant)
FROM       ventes, produits, temps
WHERE      ventes.codeProduit = produits.codeProduit
AND        ventes.date = temps.jour
AND        modèle = “vis”
GROUP BY   couleur, années ;
```

example of a typical analysis



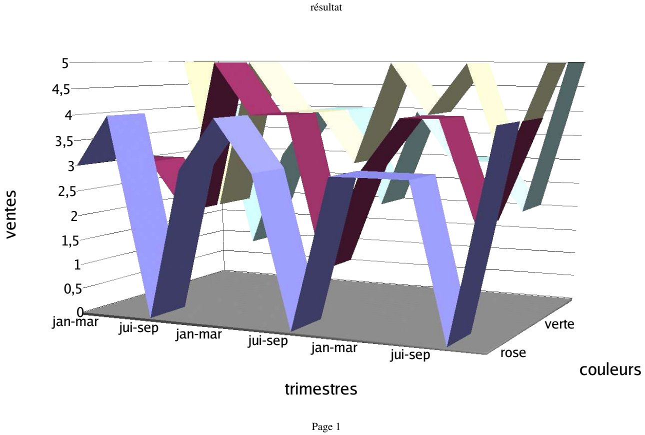
Page 1

example of a typical analysis

or maybe for a particular quarter?

```
SELECT     couleur, trimestre, SUM(montant)
FROM       ventes, produits, temps
WHERE      ventes.codeProduit = produits.codeProduit
AND        ventes.date = temps.jour
AND        modèle = “vis”
GROUP BY   couleur, trimestre ;
```

example of a typical analysis

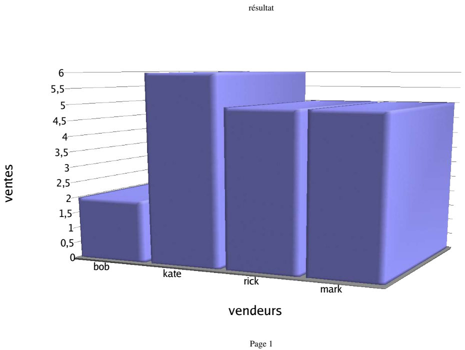


example of a typical analysis

are the salespersons to blame?

```
SELECT      vendeur, somme FROM(
SELECT      trimestre, vendeur, SUM(montant) as somme
FROM        ventes, produits, temps
WHERE       ventes.codeProduit = produits.codeProduit
AND         ventes.date = temps.jour
AND         ventes.vendeur = vendeurs.nom
AND         modèle = “vis”
GROUP BY    trimestre, vendeur)
WHERE       trimestre = “jui-sep”;
```

example of a typical analysis



decision: fire bob :-)

example of a typical treatment

what are the salespersons cumulated sales by month?

```
SELECT     vendeur, mois, CSUM(resultat,vendeur,mois) as cumul
FROM       (SELECT     vendeur, mois, Sum(montant) as resultat
            FROM       ventes, produits, temps
            WHERE      ventes.codeProduit
                       = produits.codeProduit
            AND        ventes.date = temps.jour
            AND        modèle = “vis”
            AND        couleur = “rose”
            GROUP BY   mois, vendeurs)
ORDER BY   mois ;
```

example of a typical treatment

what is the moving average on 2 consecutive days?

```
SELECT      date, montant,
            MAVG(montant,2,date) as moy
FROM        ventes, temps
WHERE       ventes.date = temps.jour
AND         année = 2001
ORDER BY    date ;
```

conclusion: what is the problem?

Chaudhuri & Dayal (Sigmod records, 1997)

supporting spreadsheet-like operations on very large databases

need specific

- data organisation
- access methods
- query languages and aggregation functions
- query optimisation technics
- ...

informal model

model

we need to stay close to the user (analyst) concepts

data are organized

- according to various dimensions
- according to various levels of detail
- into sets

data can be seen as points in a multidimensional space

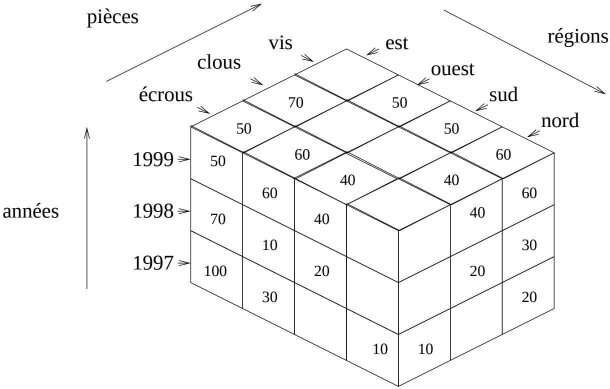
from the relation...

ventes	pièces	régions	années	quantités
	écrous	est	1999	50
	clous	est	1997	100
	vis	ouest	1998	50
	⋮	⋮	⋮	⋮
	écrous	est	total	220
	⋮	⋮	⋮	⋮
	écrous	total	total	390
	⋮	⋮	⋮	⋮
	total	total	total	1200

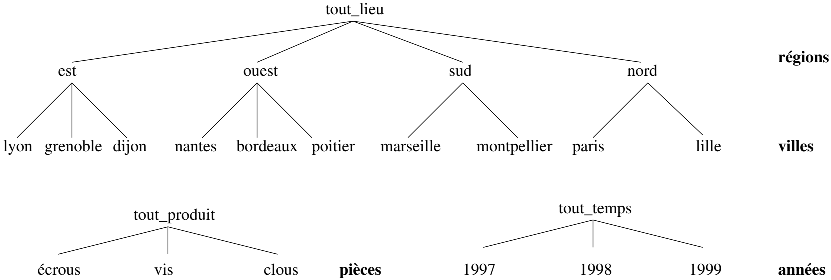
pièce, région, année → quantité

... to the cube

ventes



granularity



tout_lieu

est ouest sud nord

régions

lyon grenoble dijon nantes bordeaux poitier marseille montpellier paris lille

villes

tout_produit

écrous vis clous

pièces

tout_temps

1997 1998 1999

années

terminologie

cube	ventes
cell	écrous, est, 1997, 100
cell reference/position	écrous, est, 1997
measure	100
member (parameter)	est
dimension	lieu
level	régions

benchmark

OLAP council (1999, a bit old now)

dimensions	levels	members	calculated members	detailed data
product	7	$1000 \times x$	0	90%
customer	3	$100 \times x$	0	90%
channel	2	x	0	90%
time	4	$\geq$ 2 years		
scenario	-	2	1	
measures	-	5	5	

benchmark

T^3 project of Microsoft, Unisys, EMC, Knosys (2001)
DW

- 7.7 billion lines
- 8 fact tables
- 1.2 Tb

a MOLAP architecture

- loading, aggregating, indexing, compressing
- a cube of 471 Gb
- 53 hours (40000 rows/second)

benchmark

queries

- 50 users
- 27 different queries/users
- mean waiting time between queries: 30 seconds

mean response time

- 0.02 seconds (warm cache)
- 0.08 seconds (cold cache)

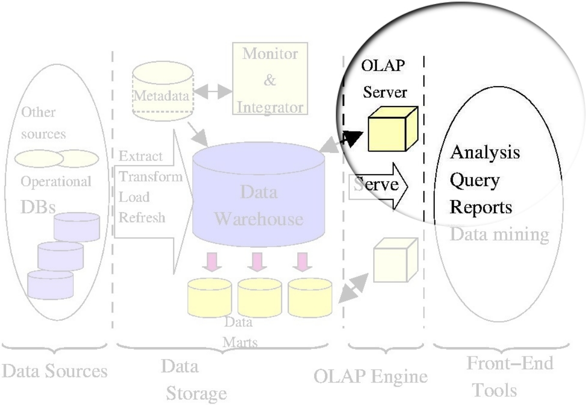
benchmark

TPC-DS from TPC (Transaction Processing Performance Council)
www.tpc.org/tpcds

- constellation schema, 7 fact tables, 24 tables total
- 4 types of dimensions: from no change to fully updatable
- random skewed data
- scale factor: from 1Gb to 100.000 Gb
- 4 types of queries: reporting, ad-hoc, OLAP interactive and data mining
- execution: loading from flat files, concurrent querying sessions # 1, refreshing, concurrent query sessions # 2
- métrics: price/performance, availability, load time, refresh time, query elapsed time

typical treatments

typical treatments



elementary operations

usually 3 categories of operators are distinguished

category	*deals with*
restructuring operators	presentation
granularity operators	level of detail
set/relational operators	filtering

restructuring

reorienting the multidimensional view

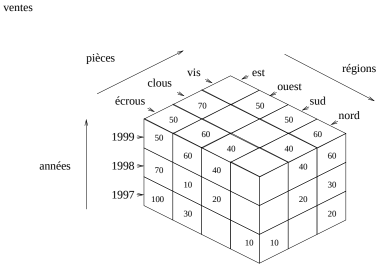
- changing viewpoint
- nesting members
- treating members/measures symetrically

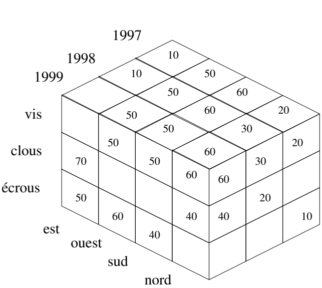
restructuring

properties

- from a cube c to a cube c'
- going from c' to c must be possible
- does not change the information extracted

rotate/pivot





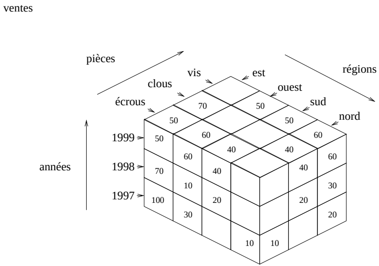
rotate/pivot

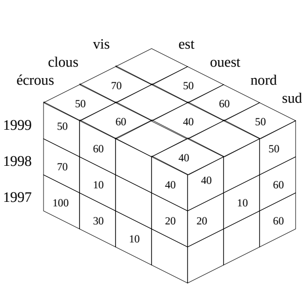
typically viewed with a cross-table

nord	1999	1998	1997
vis	60	30	20
clous	40	20	
écrous			10

vis	1999	1998	1997
est		10	10
ouest	50	50	50
sud	50	60	60
nord	60	30	20

switch


ventes
pièces
régions
écrous
clous
vis
est
ouest
sud
nord
années
1999
1998
1997
70
50
50
50
60
60
40
40
60
50
60
40
40
60
70
10
20
20
30
100
30
20
20
10
10


ventes
vis
est
clous
ouest
écrous
nord
sud
1999
1998
1997
70
50
50
60
60
40
50
50
60
40
50
70
40
40
10
10
60
100
20
20
30
10
60

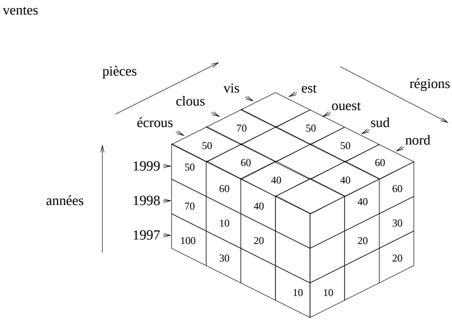
switch

viewed with a crosstab

nord	1999	1998	1997
vis	60	30	20
clous	40	20	
écrous			10

sud	1999	1998	1997
vis	50	60	60
clous		10	
écrous	40	20	

split, nest, push



1. split(régions)
2. nest(pièces, régions)
3. push(années)

1. split(régions)

ventes est	1999	1998	1997
écrous	50	70	100
vis		10	10
clous	70	70	100

ventes ouest	1999	1998	1997
écrous		10	30
vis	50	50	50
clous		10	40

ventes sud	1999	1998	1997
écrous	40	20	
vis	50	60	60
clous		10	

ventes nord	1999	1998	1997
écrous			10
vis	60	30	20
clous	40	20	

2. nest(pièces,régions)

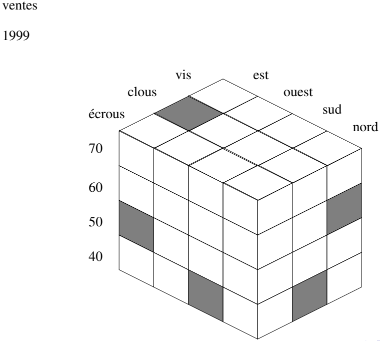
ventes nest		1999	1998	1997
écrous	est	50	70	100
	ouest		10	30
	nord			10
	sud	40	20	
vis	est		10	10
	ouest	50	50	50
	nord	60	30	20
	sud	50	60	60
clous	est	70	70	100
	ouest		10	40
	nord	40	20	
	sud		10	

3. push(années)

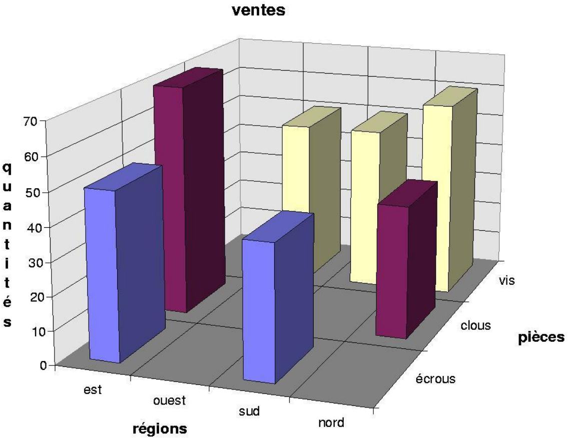
ventes push	est	ouest	nord	sud
écrous	1999 50 1998 70 1997 100	1998 10 1997 30	1997 10	1999 40 1998 20
vis	1998 10 1997 10	1999 50 1998 50 1997 50	1999 60 1998 30 1997 20	1999 50 1998 60 1997 60
clous	1999 70 1998 70 1997 100	1998 10 1997 40	1999 40 1998 20	1998 10

pull

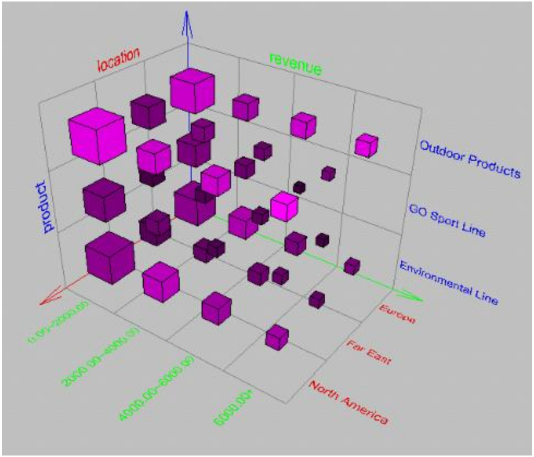
ventes 1999	est	ouest	nord	sud
écrous	50			40
vis		50	60	50
clous	70		40	



visualisation



visualisation



granularity

navigating between the levels of a dimension

- grouping
- aggregating

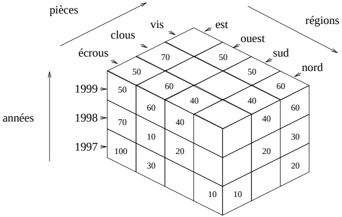
properties

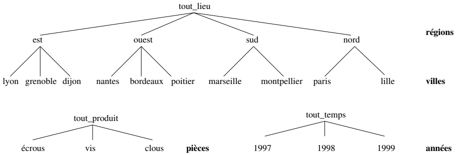
- from a cube c to a cube c'
- but going from c' to c may need more than only c'

roll-up et drill-down

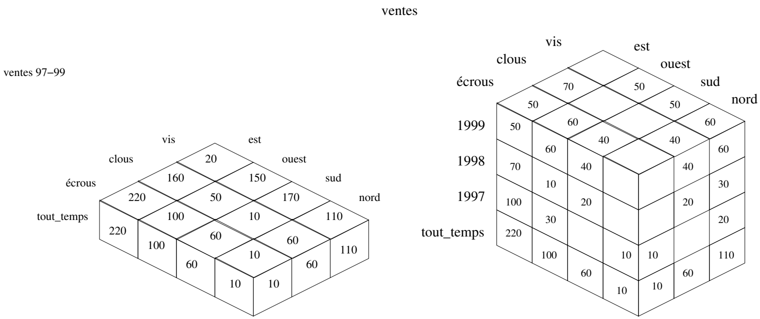
from

ventes





roll-up(années)



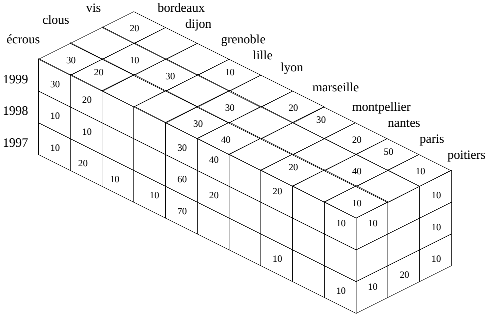
roll-up(années,pièces)

viewed with a crosstab

nord	1999	1998	1997	tout_temps
vis	60	30	20	110
clous	40	20		60
écrous			10	10
tout_produit	100	50	30	180

drill-down(régions)

ventes villes

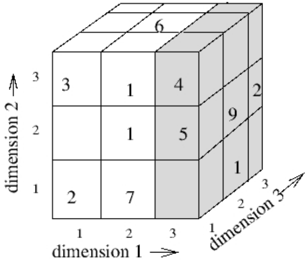


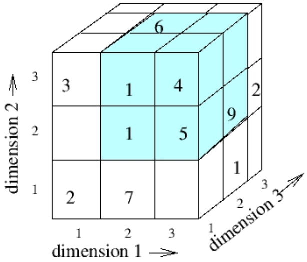
relational/set manipulation

basically the extension of the classical relational operators

well, that might need some adaptation...

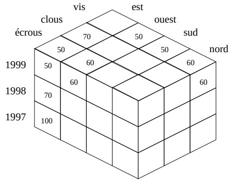
slice and dice





selection

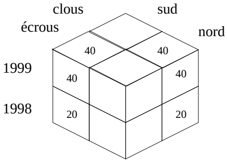
ventes $\geq$ 50



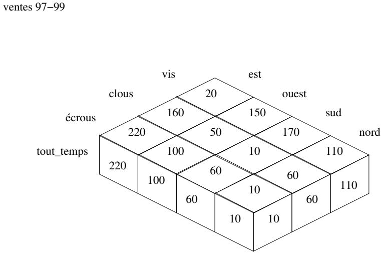
(régions = nord ou régions = sud) et

(pièces = clous ou pièces = écrous) et

(années = 1998 ou années = 1999)



projection



$\pi_{pièces,régions}$

ventes 97-99	est	ouest	sud	nord
écrous	220	100	60	10
clous	160	50	10	60
vis	20	150	170	110

join (drill-across)

ventes 97-99 $\bowtie$

prix	97-99
écrous	1
clous	0.7
vis	0.8

$=$

ventes 97-99	est	ouest	sud	nord
écrous	220 1	100 1	60 1	10 1
clous	160 0.7	50 0.7	10 0.7	60 0.7
vis	20 0.8	150 0.8	170 0.8	110 0.8

the problem with binary operations

prix	97-99
écrous	1
clous	0.7
vis	0.8

$\cup$

prix	97-99
boulons	0.8
forets	1.1
vis	0.7

the problem with binary operations

prix	97-99
écrous	1
clous	0.7
vis	0.8

$\cup$

prix	97-99
boulons	0.8
forets	1.1
vis	0.7

what measure for vis (screws)?

typical treatements

what are the top 10 performing products?

compute the 2 years moving average of sales per regions and parts

given sales for years 1997 to 1999, compute sales forecast for years 2000 to 2002 assuming a yearly 10% increase

typical treatments

data cube operator, Gray & al. 96, Shukla & al. 96
n-dimensionnal generalisation of SQL `GROUP BY`

c_1	jour	ville	ventes
	jour_1	ville_1	v_{11}
	jour_1	ville_2	v_{12}
	jour_2	ville_1	v_{21}
	⋮	⋮	⋮
	jour_q	ville_p	v_{qp}

typical treatments

jour	ville	ventes
$jour_1$	$ville_1$	v_{11}
$jour_1$	$ville_2$	v_{12}
$jour_1$	ALL	v_{1_ALL}
$jour_2$	$ville_1$	v_{21}
⋮	⋮	⋮
ALL	$ville_p$	v_{ALL_p}
ALL	ALL	v_{ALL_ALL}

typical treatments

c_2	jour$_1$	jour$_2$	...	jour$_q$
ville$_1$	v_{11}	v_{12}	...	v_{1q}
ville$_2$	v_{21}	v_{22}	...	v_{2q}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
ville$_p$	v_{p1}	v_{p2}	...	v_{pq}

data cube with hierarchies:
jour $\longrightarrow$ mois $\longrightarrow$ année
ville $\longrightarrow$ région $\longrightarrow$ pays

data cube

c_3	jour$_1$	$\ldots$	mois$_1$	$\ldots$	jour$_q$	mois$_n$	année$_1$	$\ldots$
ville$_1$	v_{11}	$\ldots$	$\sum$	$\ldots$	v_{1q}	$\sum$	$\sum$	$\ldots$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ldots$
ville$_1$	$\sum$	$\ldots$	$\sum$	$\ldots$	$\sum$	$\sum$	$\sum$	$\ldots$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ldots$
ville$_p$	v_{p1}	$\ldots$	$\sum$	$\ldots$	v_{pq}	$\sum$	$\sum$	$\ldots$
ville$_m$	$\sum$	$\ldots$	$\sum$	$\ldots$	$\sum$	$\sum$	$\sum$	$\ldots$
pays$_1$	$\sum$	$\ldots$	$\sum$	$\ldots$	$\sum$	$\sum$	$\sum$	$\ldots$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ldots$

typical treatments

a table per region displaying products whose quantity sold increased from one year to the following

sud		ventes
1999	écrous	40
1998	écrous	20
	clous	10

nord		ventes
1999	vis	60
	clous	40
1998	vis	30
	clous	20

typical treatments

calculated members:

compare this month's performance with last year's same month's performance

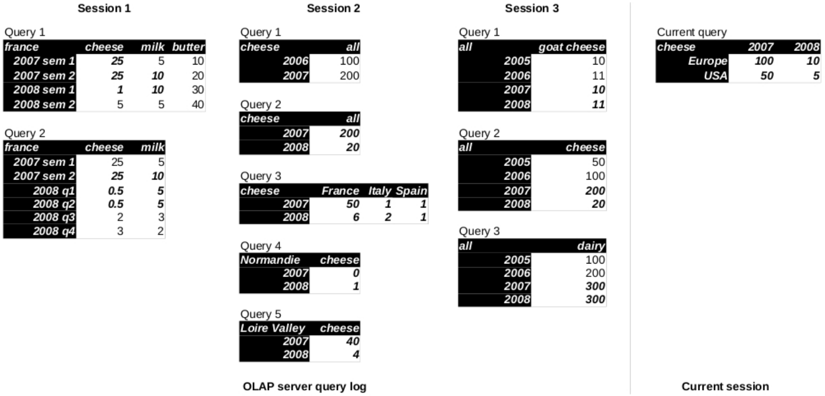
	Mar 2007	Mar 2008	Diff.
result	100	88	12

typical treatments

discovery driven analysis
Sarawagi & colleagues 1999 to 2001, VLDB conference

- explain the difference between these two cells
  - drill-down to the pairs of cells that contribute the most
  - roll-up to see if it holds at less detailed levels
  - if so, are there any exceptions?
- find the cells that deviate the most from an assumption
  - e.g., using maximum entropy principle

discovery driven analysis



Session 1

Query 1

france	cheese	milk	butter
2007 sem 1	25	5	10
2007 sem 2	25	10	20
2008 sem 1	1	10	30
2008 sem 2	5	5	40

Query 2

france	cheese	milk
2007 sem 1	25	5
2007 sem 2	25	10
2008 q1	0.5	5
2008 q2	0.5	5
2008 q3	2	3
2008 q4	3	2

Session 2

Query 1

cheese	all
2006	100
2007	200

Query 2

cheese	all
2007	200
2008	20

Query 3

cheese	France	Italy	Spain
2007	50	1	1
2008	6	2	1

Query 4

Normandie	cheese
2007	0
2008	1

Query 5

Loire Valley	cheese
2007	40
2008	4

Session 3

Query 1

all	goat cheese
2005	10
2006	11
2007	10
2008	11

Query 2

all	cheese
2005	50
2006	100
2007	200
2008	20

Query 3

all	dairy
2005	100
2006	200
2007	300
2008	300

OLAP server query log

Current query

cheese	2007	2008
Europe	100	10
USA	50	5

Current session

conclusion

So far: OLAP manipulations: playing with cubes

Next: how all this could work?