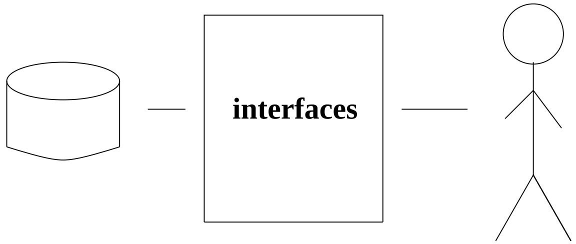
Manipulations de données multidimensionnelles

et langages de règles

16 décembre 1998

Patrick Marcel

On-Line Transactional Processings



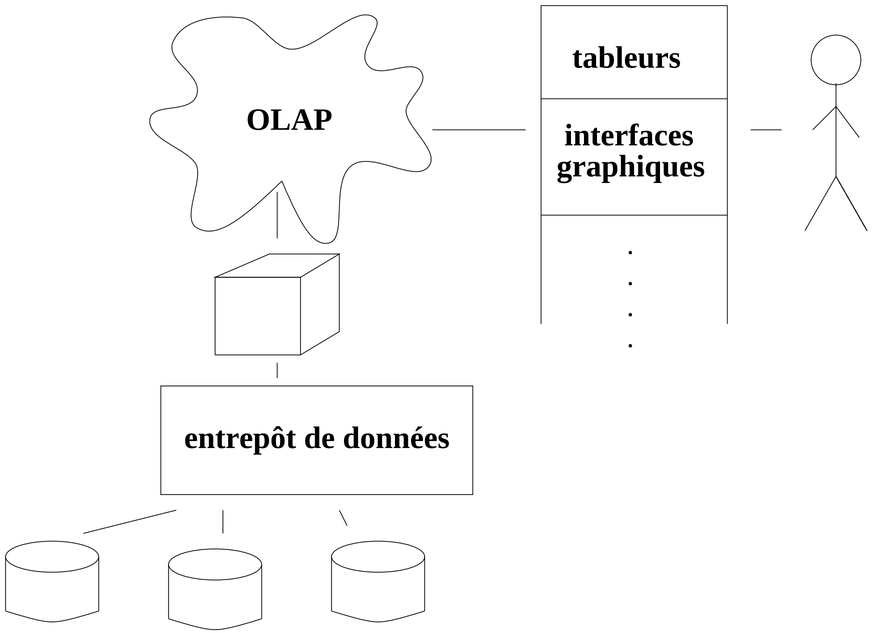
ventes	pièces	régions	années	quantités
	écrous	est	1996	50
	écrous	sud	1995	20
	clous	est	1994	100
	vis	ouest	1995	50
	⋮	⋮	⋮	⋮

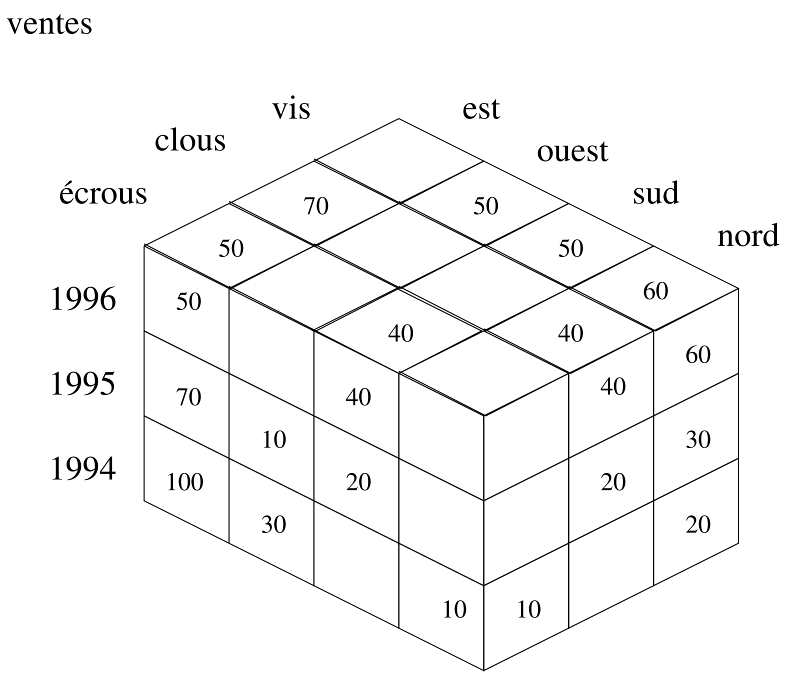
traitements transactionnels en ligne

processus simples et nombreux

mise à jour des données

On-Line Analytical Processings



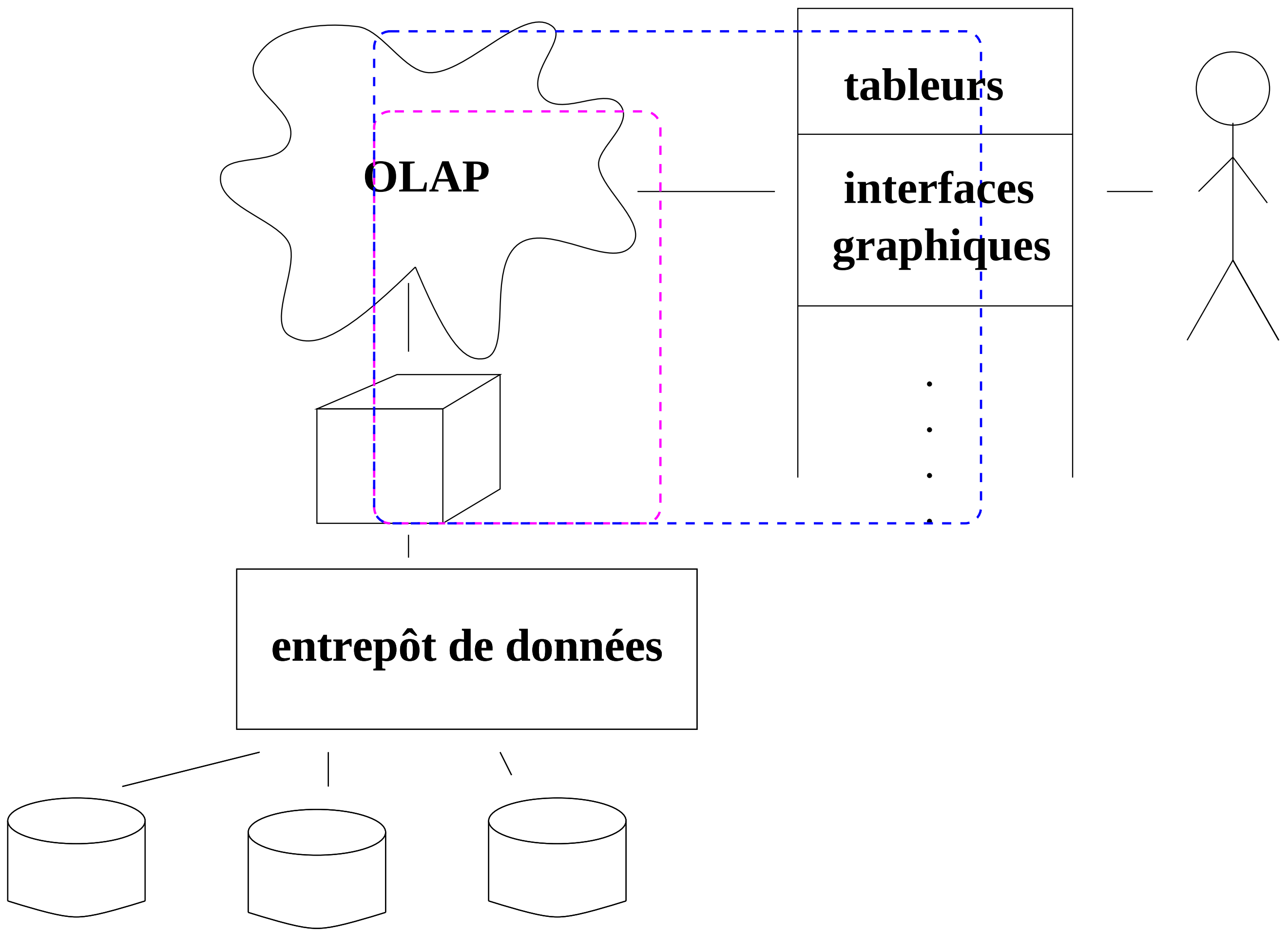


traitements analytiques en ligne

processus complexes et peu nombreux

aide à la décision

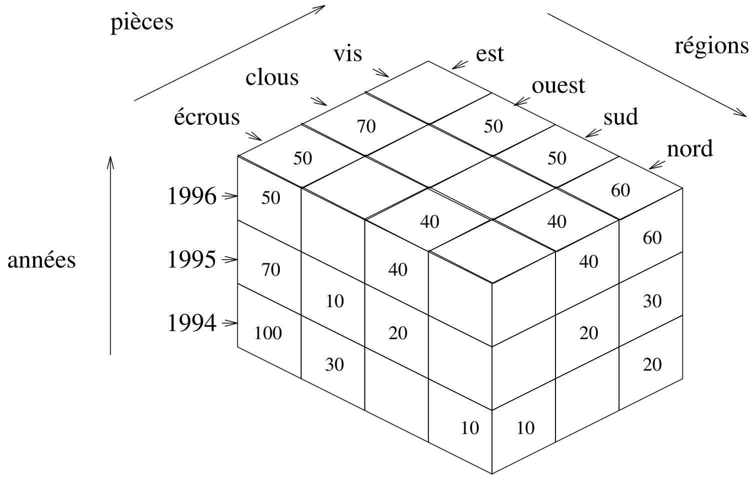
Plan

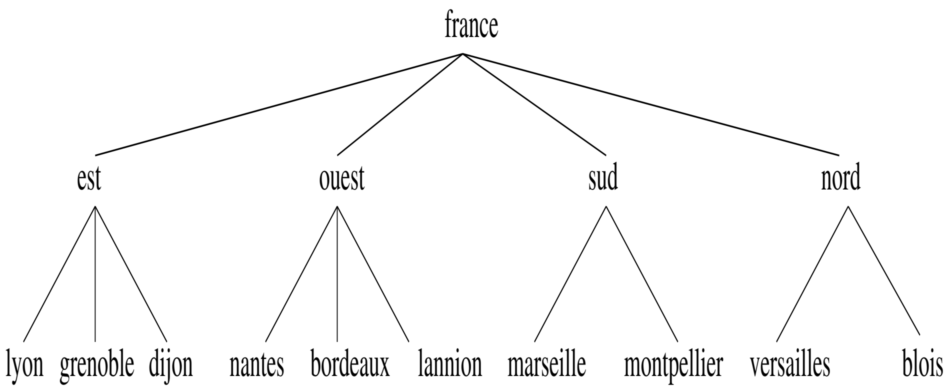


Nouvelle vision de l'information

ventes

pièces	régions	années	quantités
écrous	est	1996	50
écrous	sud	1995	20
clous	est	1994	100
vis	ouest	1995	50
⋮	⋮	⋮	⋮
écrous	est	total	220
écrous	sud	total	60
⋮	⋮	⋮	⋮
écrous	total	total	390
⋮	⋮	⋮	⋮
total	total	total	1200



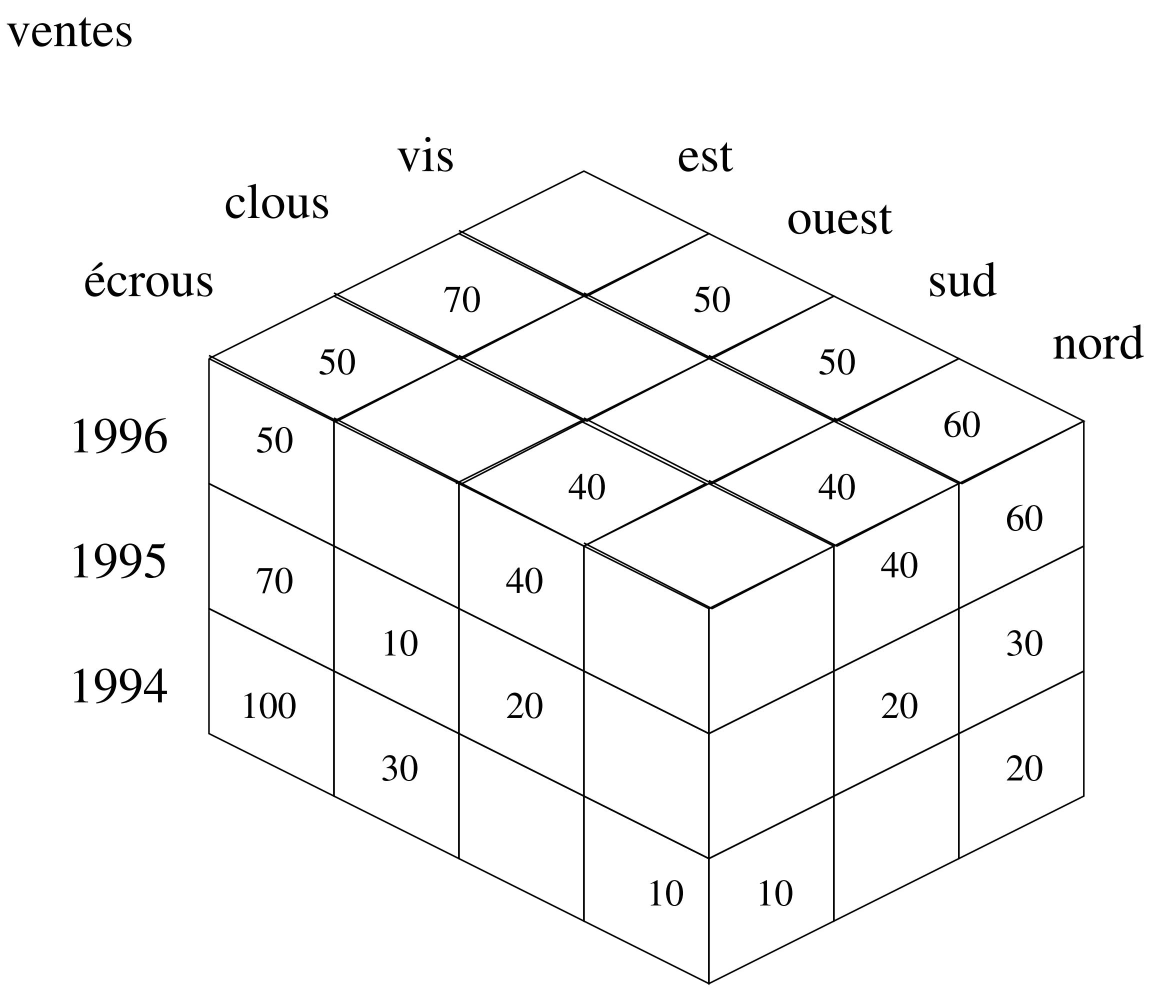


Nouvelle vision des traitements

3 catégories d'opérations élémentaires

catégorie	*concerne*
restructuration	représentation
granularité	niveau de détail
ensembliste	extraction

Le cube ventes



Flexibilité du schéma

opération élémentaire “split”

ventes est	1996	1995	1994
écrous	50	70	100
vis		10	10
clous	70	70	100

ventes ouest	1996	1995	1994
écrous		10	30
vis	50	50	50
clous		10	40

ventes sud	1996	1995	1994
écrous	40	20	
vis	50	60	60
clous		10	

ventes nord	1996	1995	1994
écrous			10
vis	60	30	20
clous	40	20	

Membres “complexes”

opération élémentaire “nest”

ventes nest		1996	1995	1994
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	

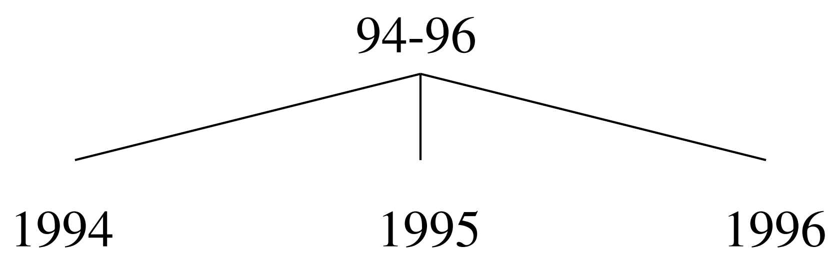
Symétrie membres/mesures

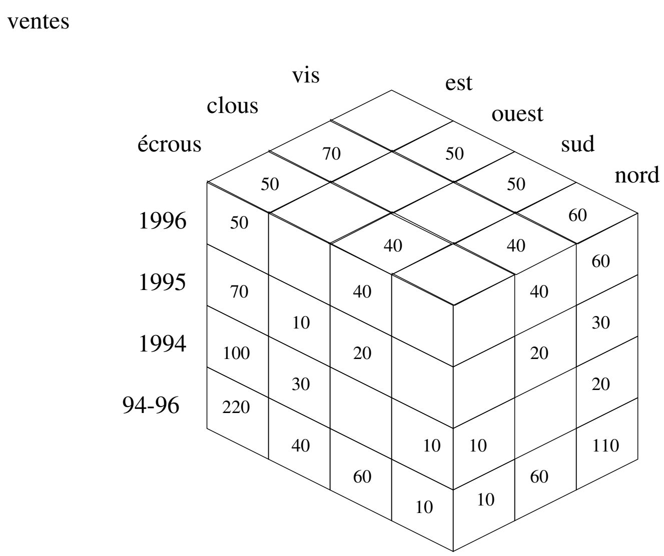
opération élémentaire "push"

ventes push	est	ouest	nord	sud
écrous	1996 50 1995 70 1994 100	1995 10 1994 30	1994 10	1996 40 1995 20
vis	1995 10 1994 10	1996 50 1995 50 1994 50	1996 60 1995 30 1994 20	1996 50 1995 60 1994 60
clous	1996 70 1995 70 1994 100	1995 10 1994 40	1996 40 1995 20	1995 10

Groupements

opération élémentaire “roll-up”





Synthèse

	Gyssens & al. 1996	Agrawal & al. 1996	Li & Wang 1996	Cabibbo & Torlone 1997	Gyssens & Laksh. 1997
type	algèbre	algèbre	algèbre	calcul	algèbre calcul
push	non	oui	non	oui	oui
split	oui	non	non	non	non
nest	non	non	oui	non	oui
groupements	non	fonction externe	oui	oui	fonction externe

un cube est une fonction

proximité du cadre relationnel

Exemple de requête typique

une table par région contenant les produits dont la quantité vendue a augmenté d'une année sur l'autre...

sud		ventes
1996	écrous	40
1995	écrous	20
	clous	10

nord		ventes
1996	vis	60
	clous	40
1995	vis	30
	clous	20

Exemple de spécification algébrique

Algèbre de Gyssens & Lakshmanan, VLDB 97

$$unfold^{lignes}_{années,produits}($$
$$fold^{produits}($$
$$fold^{années}($$
$$\pi_{années,produits,quantité}(sud))))$$

où le cube *sud* est défini par :

$$\sigma_{régions=sud}($$
$$\pi_{années,régions,produits,quantité}($$
$$\sigma_{années=a+1 \wedge quantité>q}($$
$$\rho_{quantités,régions,produits,années \rightarrow q,r,p,a}(ventes)$$
$$\times ventes)))$$

. . . et pourquoi pas ça?

$$
\begin{aligned}
\mathsf{R}(\mathsf{A}_2\cdot\mathsf{P},\mathsf{ventes})\mathbf{:}\langle \mathsf{Q}_2\rangle \leftarrow\ & \mathsf{ventes}(\mathsf{P},\mathsf{A}_2,\mathsf{R})\mathbf{:}\langle \mathsf{Q}_2\rangle,\\
& \mathsf{ventes}(\mathsf{P},\mathsf{A}_1,\mathsf{R})\mathbf{:}\langle \mathsf{Q}_1\rangle,\\
& \mathsf{Q}_2 > \mathsf{Q}_1,\\
& \mathsf{succ}(\mathsf{A}_1,\mathsf{A}_2).
\end{aligned}
$$

Idée de base

une cellule : ventes(écrous,1996,ouest) :$\langle$50$\rangle$

inspirations :

monovaluation	dépendances fonctionnelles	(e.g., COL)
restructuration	langages d'ordre supérieur	(e.g., Hilog)
granularité	langages logiques objets	(e.g., F-logic)

Restructurations

split

$$\text{ventes}\cdot\text{R}(\text{P},\text{A}) :\langle \text{Q}\rangle \leftarrow \text{ventes}(\text{P},\text{A},\text{R}) :\langle \text{Q}\rangle.$$

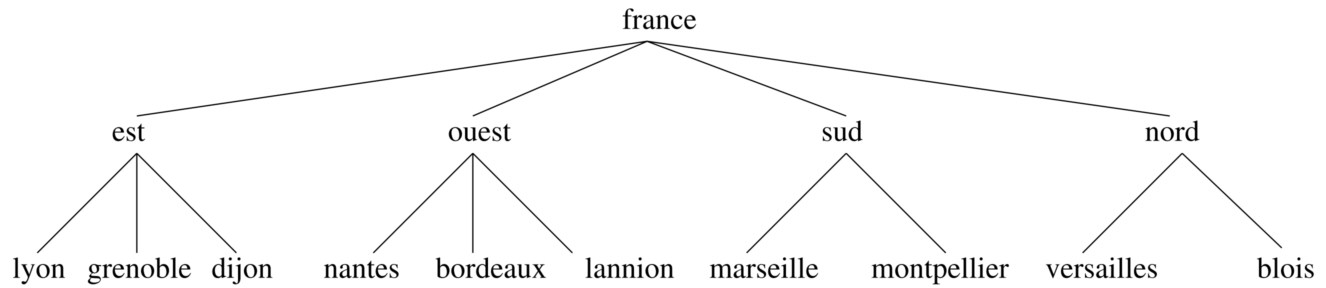
nest

$$\text{ventesNest}(\text{P}\cdot\text{R},\ \text{A}) :\langle \text{Q}\rangle \leftarrow \text{ventes}(\text{P},\text{A},\text{R}) :\langle \text{Q}\rangle.$$

push

$$\begin{aligned}\text{VentesPush}(\text{P},\text{R}):\langle \text{A}_1,\text{Q}_1,\text{A}_2,\text{Q}_2,\text{A}_3,\text{Q}_3\rangle \leftarrow\ & \text{ventes}(\text{P},\text{A}_1,\text{R}):\langle \text{Q}_1\rangle,\\ & \text{ventes}(\text{P},\text{A}_2,\text{R}):\langle \text{Q}_2\rangle,\\ & \text{ventes}(\text{P},\text{A}_3,\text{R}):\langle \text{Q}_3\rangle,\\ & \text{A}_1 < \text{A}_2,\\ & \text{A}_2 < \text{A}_3.\end{aligned}$$

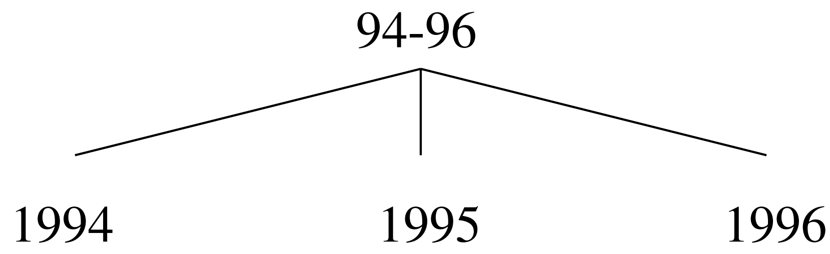
Granularité



description des groupements

in(écrous,pièces), in(vis,pièces), in(1994,94-96), in(lyon,est), in(est,france), . . .

Roll-up



20=sum(ventes(vis,94-96,est))

$$\{\text{ventes(vis,1995,est)}:\langle 10\rangle,\ \text{ventes(vis,1994,est)}:\langle 10\rangle\}$$

roll-up sur la dimension des années

$$\begin{aligned}\text{ventes(P,94-96,R)}:\langle T\rangle \leftarrow\ & \text{T=sum(ventes(P,94-96,R))},\\ & \text{in(P,pièces)},\\ & \text{in(R,france)}.\end{aligned}$$

Prévisions

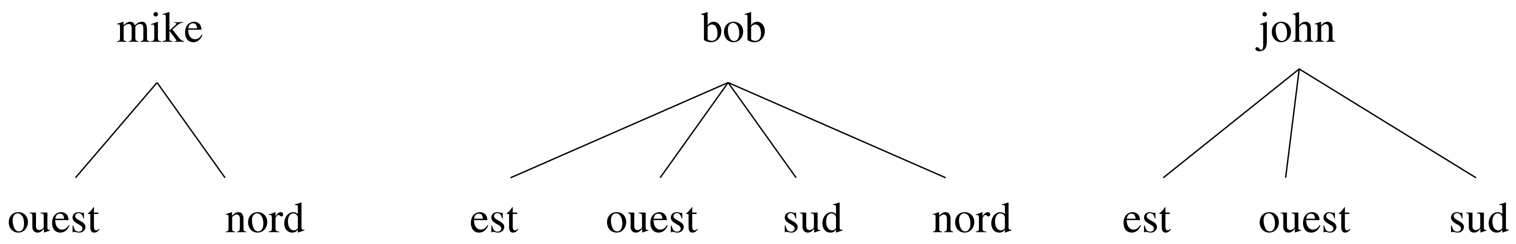
$$
\begin{aligned}
\text{ventes}(P,A_2,R){:}\langle Q_2\rangle \leftarrow\ & \text{ventes}(P,A_1,R){:}\langle Q_1\rangle,\\
& A_2 \text{ is } A_1 + 1,\\
& Q_2 \text{ is } Q_1 + Q_1/10,\\
& A_2 \leq 2000,\\
& A_1 \geq 1996.
\end{aligned}
$$

$\{$ventes(clous,1997,est):$\langle 77\rangle$, ventes(clous,1997,nord):$\langle 44\rangle$,
ventes(vis,1997,sud):$\langle 55\rangle$, ventes(clous,1998,est):$\langle 84.7\rangle$,
ventes(clous,1998,nord):$\langle 48.4\rangle$, $\ldots$ $\}$

Groupements ad-hoc

responsables	est	ouest	sud	nord
vis	john	mike	bob	mike
clous	bob	john	bob	mike
écrous	john	bob	john	bob

$\mathsf{in(R,N)} \leftarrow \mathsf{responsables(P,R)}\ \mathsf{:}\langle \mathsf{N} \rangle.$



Groupements ad-hoc

$$\text{résultat96(P,N)} :\langle \text{S} \rangle \leftarrow \text{S} = \text{sum(ventes(P,1996,N))}, \\ \text{responsables(P,\_)} :\langle \text{N} \rangle.$$

résultat96	john	bob	mike
vis	50	50	110
clous		70	40
écrous	90		

Opérateur "data cube"

Gray et al. 96, Shukla et al. 96

généralisation n-dimensionnelle de l'opérateur "group-by" de SQL

c	jour_1	jour_2	$\ldots$	jour_q
vendeur_1	v_{11}	v_{12}	$\ldots$	v_{1q}
vendeur_2	v_{21}	v_{22}	$\ldots$	v_{2q}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
vendeur_p	v_{p1}	v_{p2}	$\ldots$	v_{pq}

$\text{in}(\text{jour}_1, \text{mois}_1)$,
$\text{in}(\text{jour}_2, \text{mois}_1)$,
$\vdots$
$\text{in}(\text{mois}_1, \text{année}_1)$,
$\vdots$
$\text{in}(\text{vendeur}_1, \text{ville}_1)$,
$\vdots$
$\text{in}(\text{ville}_1, \text{pays}_1)$,
$\vdots$

$\text{niveau}(X) \leftarrow \text{in}(X,Y)$
$\text{niveau}(Y) \leftarrow \text{in}(X,Y)$

Opérateur “data cube”

c(T,C):⟨S⟩ $\leftarrow$ S=sum(c(T,C)),niveau(T),niveau(C).

c	jour_1	$\ldots$	mois_1	$\ldots$	jour_q	mois_n	année_1	$\ldots$
vendeur_1	v_{11}	$\ldots$	$\sum$	$\ldots$	v_{1q}	$\sum$	$\sum$	$\ldots$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ldots$
ville_1	$\sum$	$\ldots$	$\sum$	$\ldots$	$\sum$	$\sum$	$\sum$	$\ldots$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ldots$
vendeur_p	v_{p1}	$\ldots$	$\sum$	$\ldots$	v_{pq}	$\sum$	$\sum$	$\ldots$
ville_m	$\sum$	$\ldots$	$\sum$	$\ldots$	$\sum$	$\sum$	$\sum$	$\ldots$
pays_1	$\sum$	$\ldots$	$\sum$	$\ldots$	$\sum$	$\sum$	$\sum$	$\ldots$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ldots$

Propriétés du langage

cadre formel proche du cadre datalog classique

- sémantique déclarative (type théorie des modèles)
- sémantique opérationnelle (type pppf) équivalente
- classe des traitements traitables sur bases ordonnées

Synthèse

	Gyssens & al. 1996	Agrawal & al. 1996	Li & Wang 1996	Cabibbo & Torlone 1997	Gyssens & Laksh. 1997	Marcel 1998
type	algèbre	algèbre	algèbre	calcul	algèbre calcul	langage de règles
push	non	oui	non	oui	oui	oui
split	oui	non	non	non	non	oui
nest	non	non	oui	non	oui	oui
groupements	non	fonction externe	oui	oui	fonction externe	oui

Tableurs : situation

visualisation, manipulation multidimensionnelle, calcul

niveau	*base de données*	*tableurs*
conceptuel	e.g., modèle E-R	?
logique	e.g., modèle relationnel	?
physique	e.g., Access	e.g., Excel

Feuille souhaitée

	A	B	C	D
1	pièces	quantité	marge	rabais
2	écrous	330		
3	vis	450		
4	clous	360		

Spécifications

entrées initiales

$$\text{pièces(P,quantité):Q} \leftarrow \text{Q=sum(ventes(P,94-96,france)),}$$
$$\text{in(P,piéces)}$$

$$\text{pièces(P,rabais):R} \leftarrow \text{pièces(P,quantité):Q,}$$
$$\text{pièces(P,marge):M,}$$
$$\text{M} \leq 0.1,$$
$$\text{R is Q} \times \text{M}$$

$$\text{pièces(P,rabais):R} \leftarrow \text{pièces(P,quantité):Q,}$$
$$\text{pièces(P,marge):M,}$$
$$\text{M} > 0.1,$$
$$\text{R is Q} \times 0.1$$

Approximation par excès des références

entrées différées

pièces(P,marge):@ $\leftarrow$ pièces(P,quantité):Q.

$$@ \leq 0.1 \rightarrow \text{vrai} \qquad\qquad Q \times @ \rightarrow @$$

cellules candidates

{pièces(écrous,quantité):500, pièces(vis,quantité):1053,
pièces(clous,quantité):5, pièces(écrous,rabais):$@_1$,
pièces(écrous,marge):$@_2$, pièces(vis,marge):$@_3$, ... }

Spécialisation et génération

$$\text{pièces(écrous,rabais):}@_1 \leftarrow \text{pièces(écrous,quantité):500,}$$
$$\text{pièces(écrous,marge):}@_2,$$
$$@_2 \leq 0.1,$$
$$@_1 \text{ is } 500 \times @_2$$

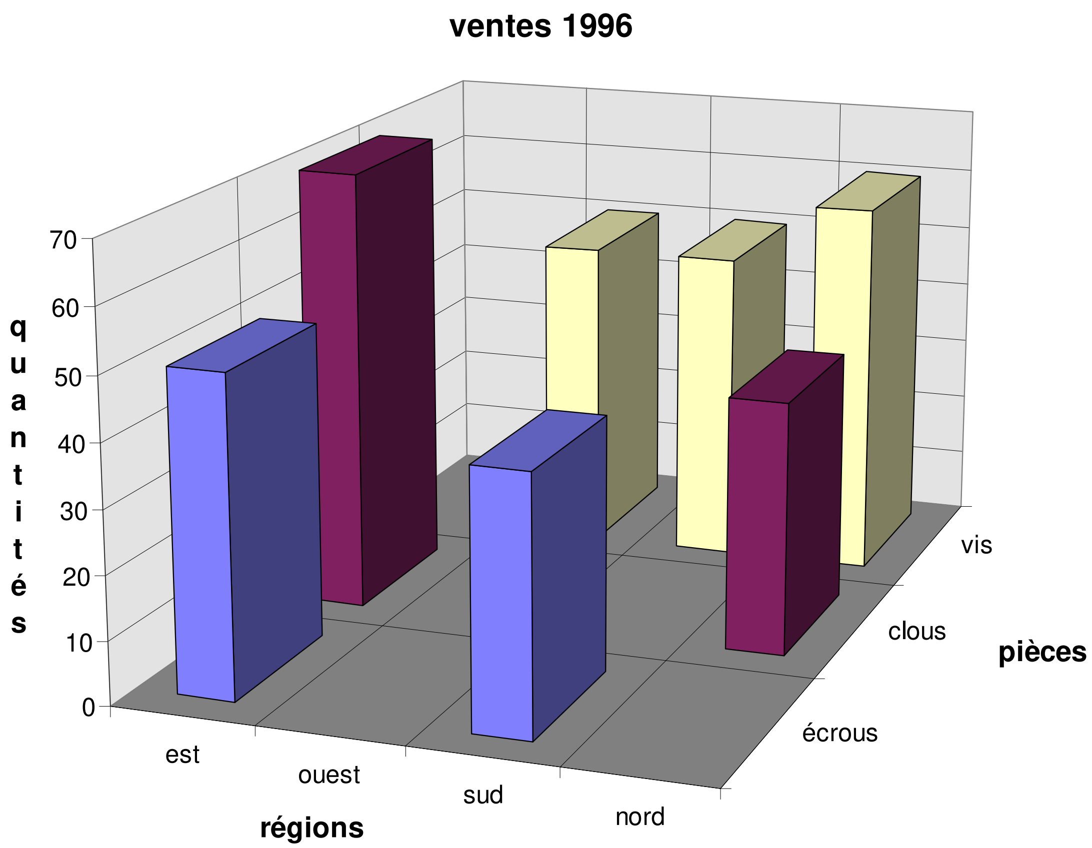
$$\text{pièces(écrous,rabais):}@_1 \leftarrow \text{pièces(écrous,quantité):500,}$$
$$\text{pièces(écrous,marge):}@_2,$$
$$@_2 > 0.1,$$
$$@_1 \text{ is } 500 \times 0.1$$

Formule de la case $D2$

=IF(ISNUM(C2);IF(C2 $\leq$ 0.1;500 $\times$ C2;500 $\times$ 0.1);"indéfini")

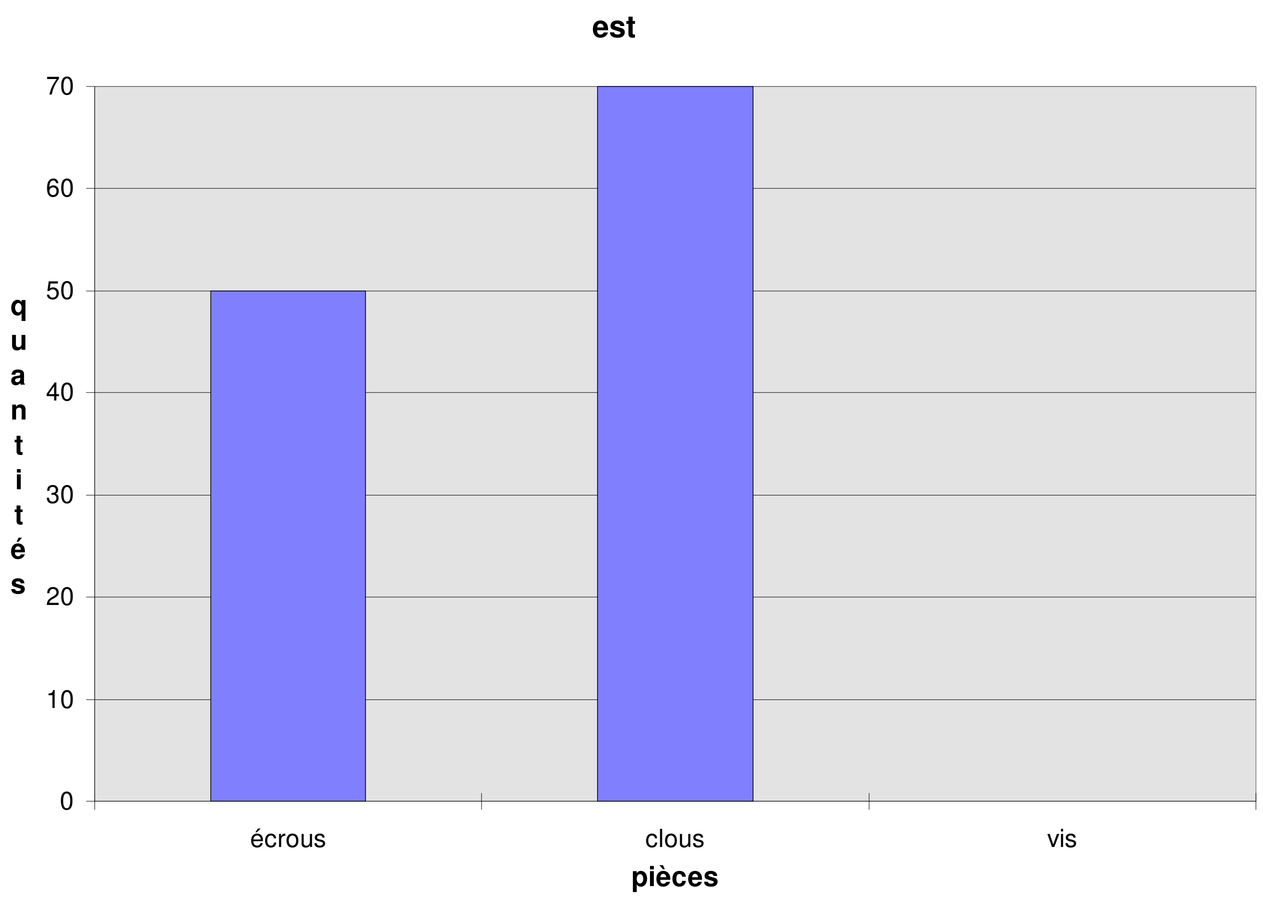
Aspects graphiques

histo3D·ventes1996(P,R,Q) $\leftarrow$ ventes(P,1996,R):Q.



Aspects graphiques

histo2D·est(P,V) $\leftarrow$ histo3D·ventes1996(P,est,V).



Conclusion et perspectives

– descriptions concises de manipulations OLAP complexes

– prise en compte des aspects interfaces

– étude de propriétés OLAP