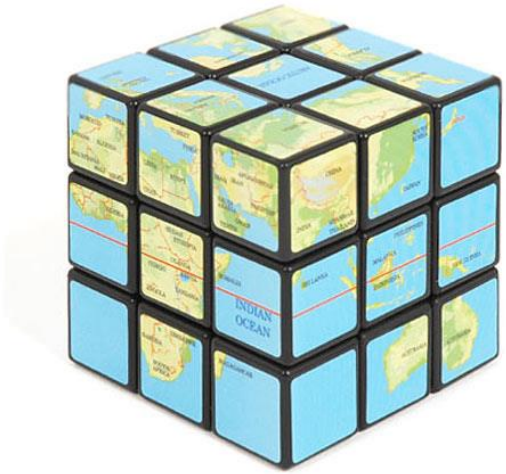
Informatique décisionnelle

Spatial OnLine Analytical Processing (SOLAP)

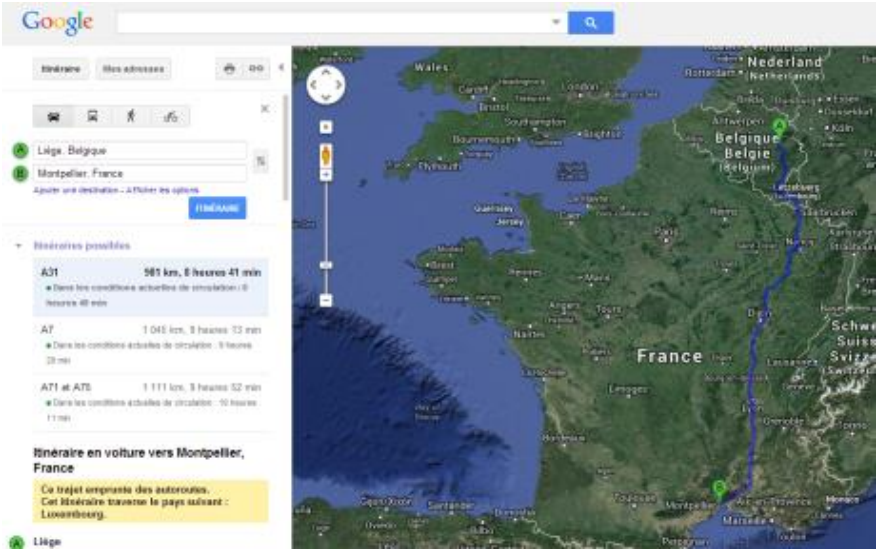


1. Introduction: « Big Data »

1.1 Aspect transactionnel des données (*OnLine Transactionnal Processing*)

» Interrogation et alimentation de base de données au quotidien par un grand nombre d'utilisateurs

» Sources très diverses



1.2. Aspect décisionnel des données (*OnLine Analytical Processing*)

» Le volume des données numériques croit exponentiellement

» Archivage et analyse des données transactionnelles par un petit groupe d'utilisateurs dans un but décisionnel

OLTP	OLAP
« *On Line Transactional Processing* »	« *On Line Analytical Processing* »
Accès à une information précise sur base de critères de recherche	Analyse de grands jeux de données à différents niveaux de granularité
Requêtes simples et nombreuses	Requêtes complexes et sporadiques
Nécessite des mises à jour fréquentes	Nécessite peu de mises à jour (outil d'archivage)
Pas de redondance (données normalisées)	Redondance autorisée
Approche conceptuelle « entités-associations »	Approche conceptuelle « multidimensionnelle »

1.3. Informatique décisionnelle (« *Business Intelligence* »)

» Collecte, consolidation et analyse de données afin d'aider les entreprises dans le processus de prise de décision

» Approche décisionnelle (OLAP) ≠ approche transactionnelle (OLTP)

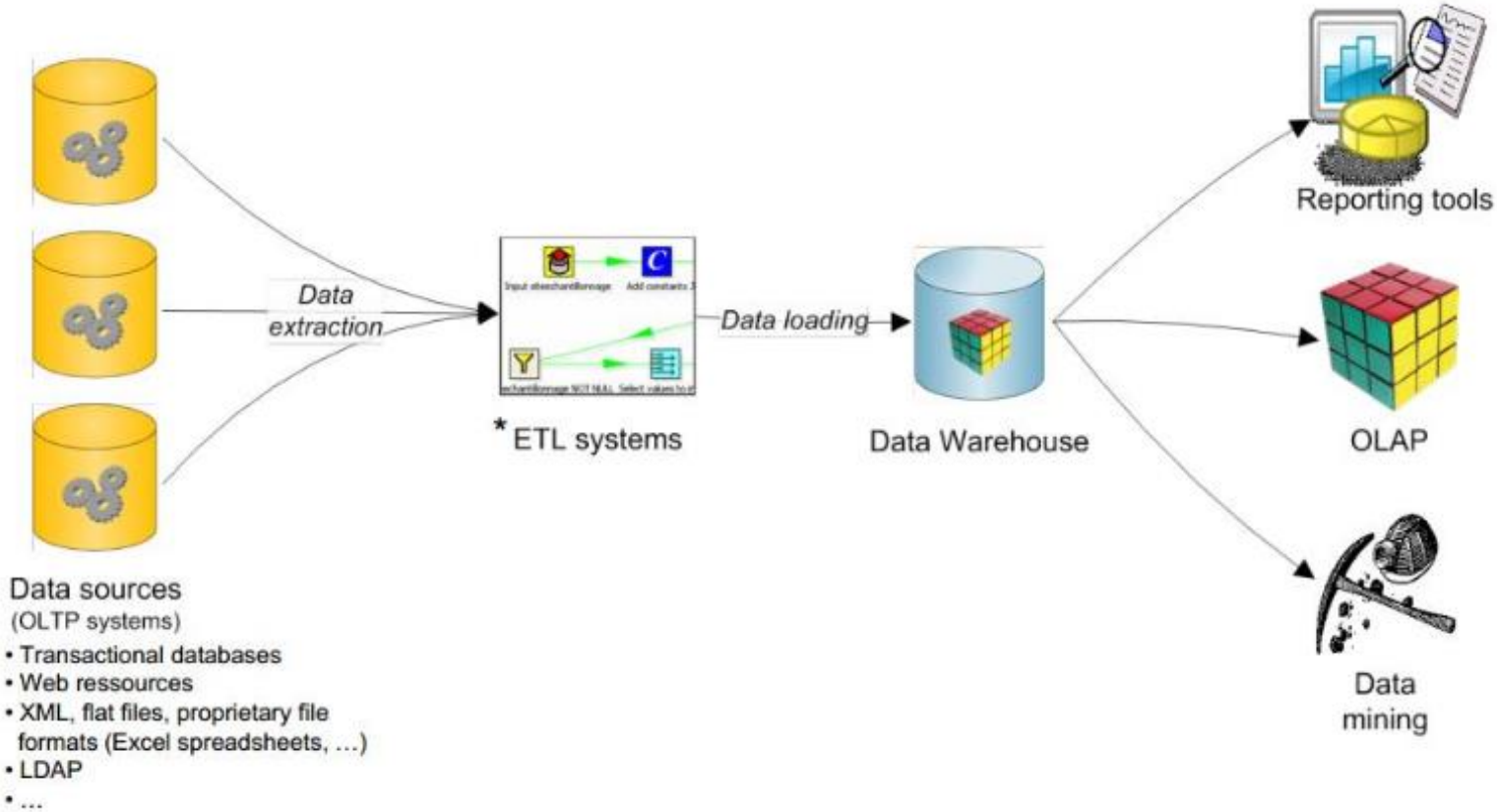
» Le cœur d'une architecture BI est l'entrepôt de données (« *data warehouse* »)

- L'entrepôt présente une structure multidimensionnelle

» Le serveur OLAP (« *On Line Analytical Processing* ») permet à un utilisateur d'extraire simplement et rapidement de l'information synthétisée hors de l'entrepôt

- à différents niveaux de granularité des dimensions (agrégations des données)
- via des tableaux et des graphiques interactifs

» Un serveur SOLAP (« *Spatial OLAP* ») permet une navigation dans un entrepôt de données spatiales via des cartes interactives (en plus des tableaux et graphiques) et combine l'OLAP aux fonctionnalités d'un SIG



Architecture BI, d'après Badard et al, 2009

» Intérêt du multidimensionnel

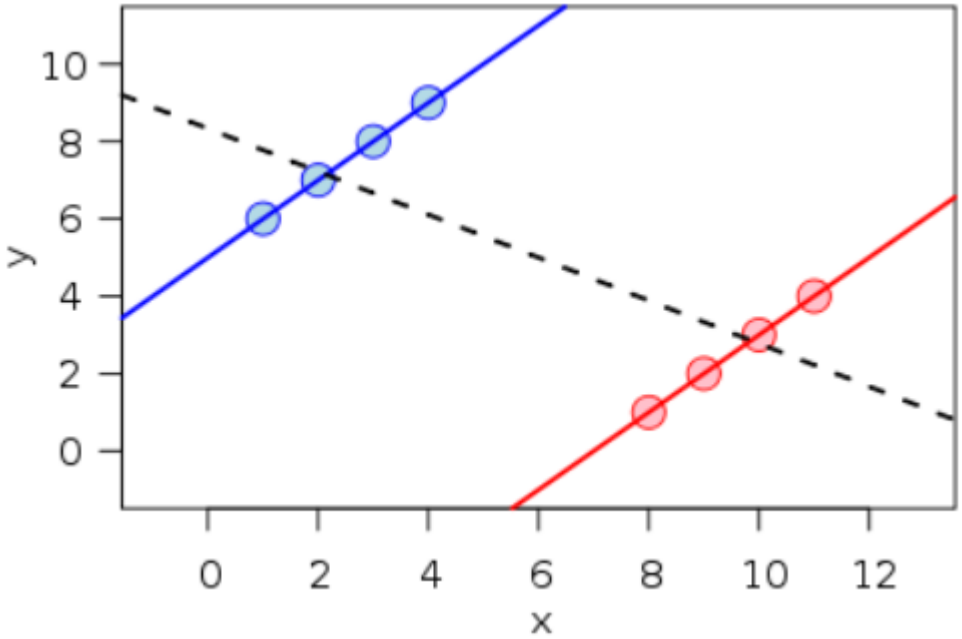


Illustration du paradoxe de Simpson
(d'après Wikipedia, 2017)

2. Entrepôt de données

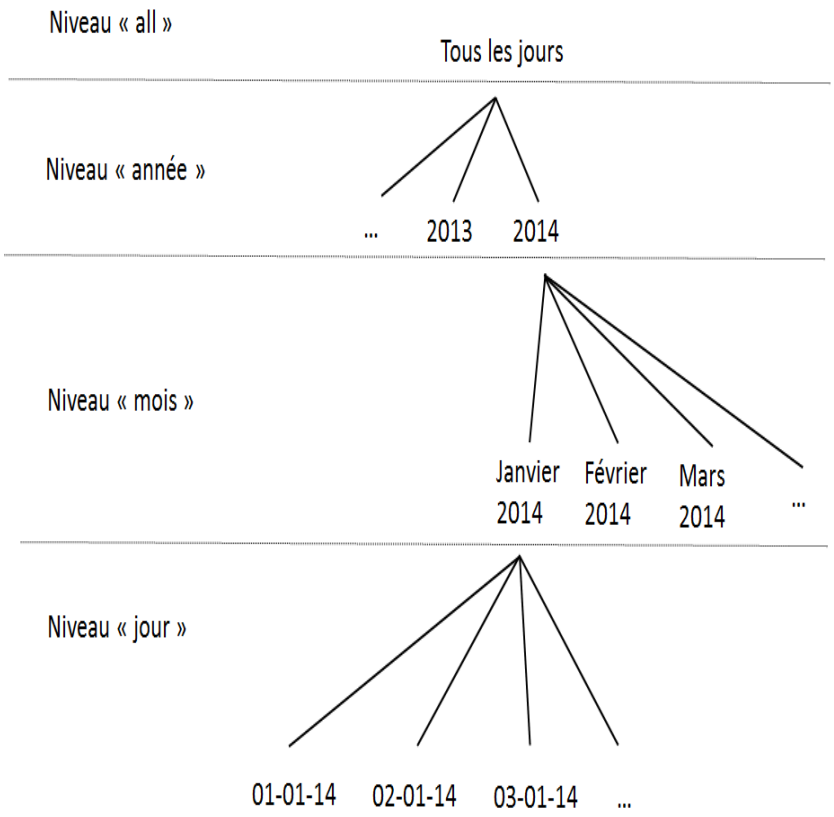
2.1. Dimension

» Une dimension est un ensemble d'éléments appelés membres organisés en plusieurs niveaux hiérarchiques

» Les membres du niveau le plus bas sont appelés membres détaillés

» Propriétés d'une hiérarchie « idéale »

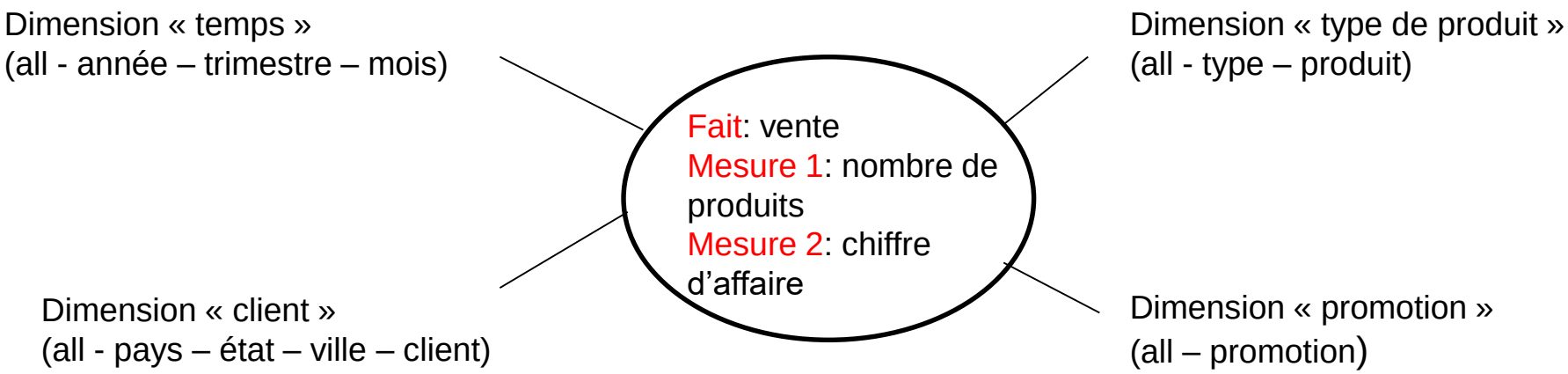
- stricte: un membre enfant a strictement un seul parent
- symétrique: seuls les membres du niveau le plus détaillé n'ont pas d'enfants
- couvrante: aucun membre ne saute de niveau
- statique: ne varie pas en fonction des autres dimensions



Exemple de dimension temporelle hiérarchisée

2.2. Schéma en étoile

- » Conceptuellement , la structure multidimensionnelle d'un entrepôt de données est décrite par un schéma en étoile
- » Un fait auquel est associé une ou plusieurs mesure(s) est l'élément central du schéma et les dimensions sont les branches de l'étoile
- » Un fait est défini par un membre de chaque dimension
- » Un fait détaillé est décrit uniquement par des membres détaillés
- » Les mesures sont soit stockées dans l'entrepôt (fait détaillé), soit calculées par agrégation (fait non-détaillé)



Exemple de schéma en étoile

2.3. Cube de données

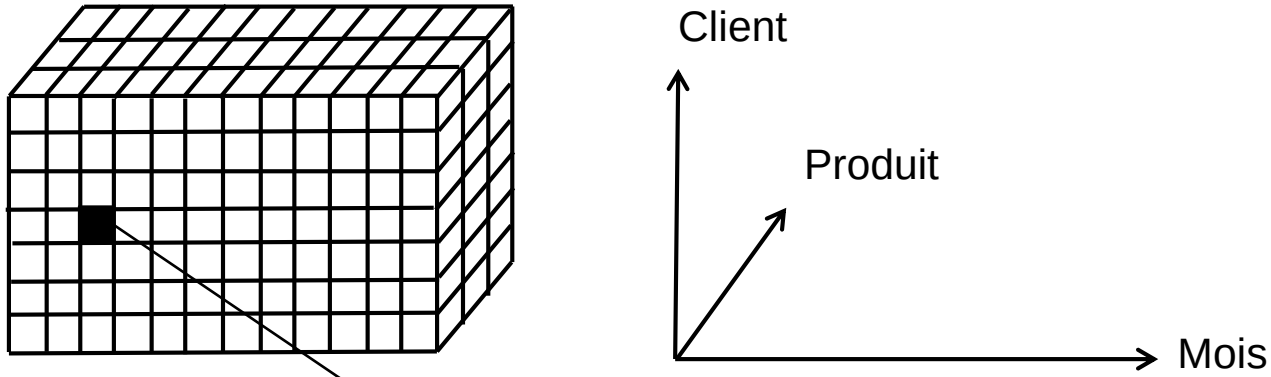
» Cube de données = instance d'un schéma en étoile

» Les cellules du cube de base contiennent les mesures des faits détaillés (mesures atomiques)

- L'ensemble des faits détaillés est le produit cartésien des dimensions au niveau le plus fin (ensemble de membres au niveau le plus fin)

» Un cube de données à plus de 3 dimensions est aussi appelé « hypercube de données »

» La valeur d'une cellule du cube est une mesure et la coordonnée d'une cellule selon un axe d'analyse est un membre de dimension



Exemple de cube de données

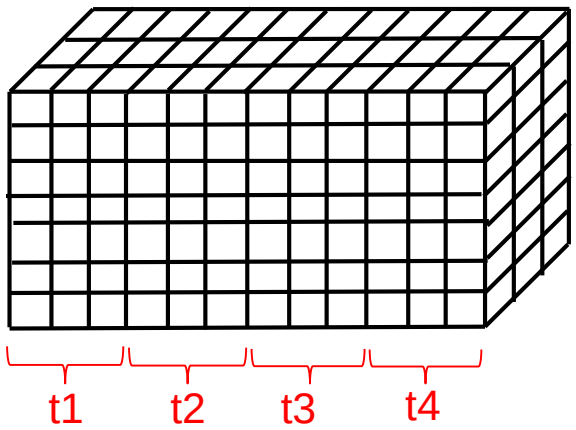
Membres [Client: Dupont, Produit: guitare, Mois: mars 2017]
Mesures [Chiffre d'affaire: 1000 €, Quantité: 1]

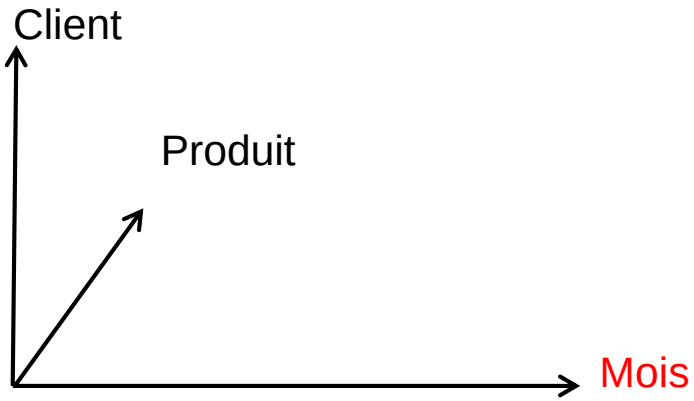
2.4. Cuboïde de données

» Un cuboïde de données contient les faits pour un niveau non-détaillé de une ou plusieurs dimension(s) (faits non-détaillés)

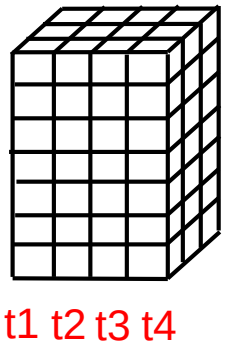
» Certains cuboïdes sont pré-calculés dans l'entrepôt pour optimiser les requêtes

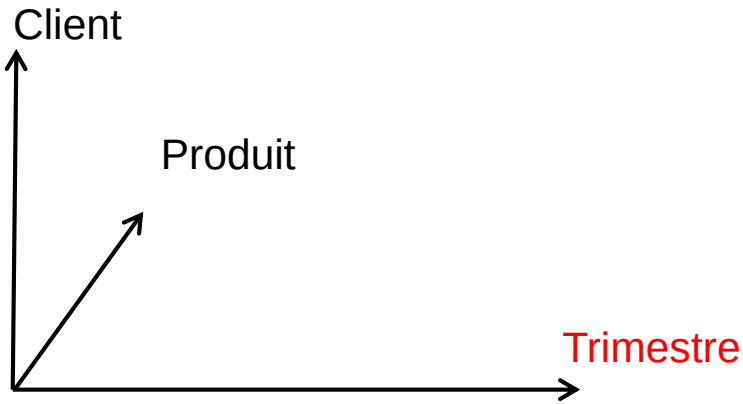
Cube de base





Cuboïde





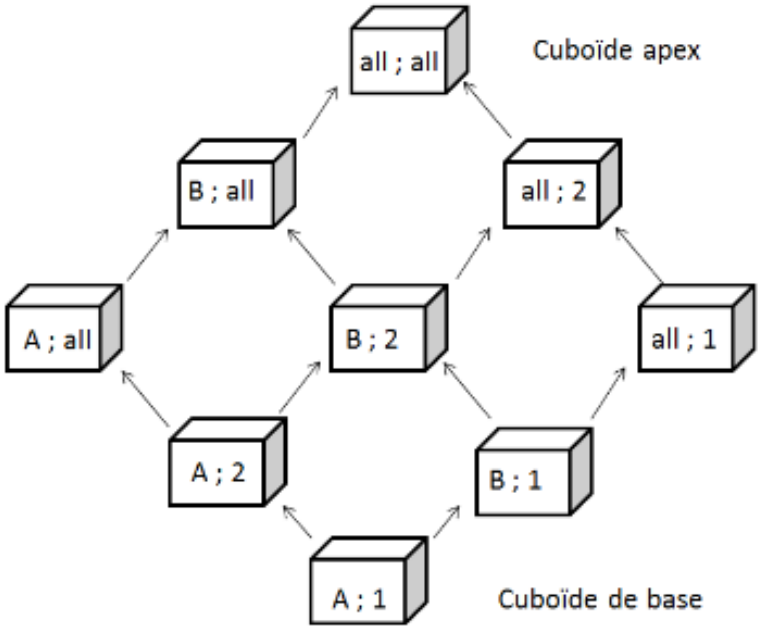
Exemple de cuboïde de données

2.5. Treillis de cuboïdes

» Il est théoriquement possible de pré-calculer tous les cuboïdes d'un entrepôt via un treillis de cuboïdes…

… mais leur nombre croît exponentiellement avec le nombre de dimensions !

» Pour n dimensions comprenant L niveaux, le nombre de cuboïdes vaut: $\prod_{i=1}^{n} L_i$



Exemple de treillis de cuboïdes comprenant deux dimensions {A,B,all} et {1,2,all}

3. Serveur OLAP

3.1. Généralités

» Rôle du serveur OLAP: offrir à l'utilisateur une navigation intuitive dans un cube de données

» Définition par l'acronyme FASMI (*BI Verdict*) : « *Fast Analysis of Shared Multidimensional Information* »

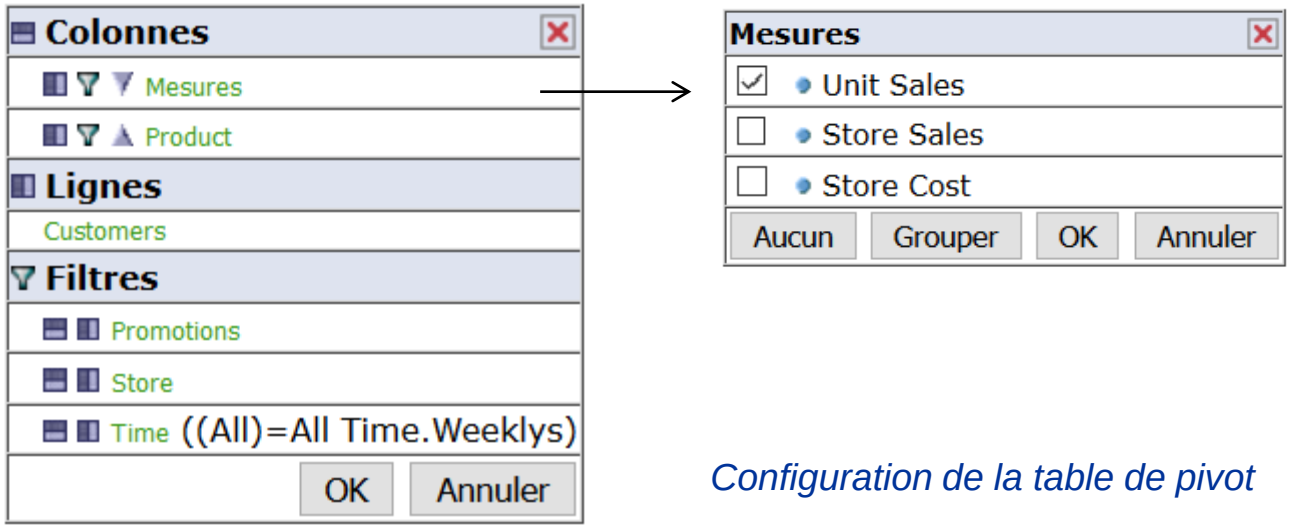
» Interfaces: affiche un nombre limité de dimensions

- Tableaux (table de pivot)
- Graphiques

» Opérations typiques

- Configuration de l'interface: ajout/suppression des dimensions et mesures visibles
- Opérations de forage (changement de cuboïde)
  - « *Roll Up* » : passage au niveau supérieur d'une hiérarchie dimensionnelle
  - « *Drill Down* » : passage au niveau inférieur d'une hiérarchie dimensionnelle
- Opérations de coupe (« tranche de cube »)
  - « *Slice* »: fixe les mesures sur un ou plusieurs membre(s) de dimension
  - « *Dice* »: supprime certains membres de dimension

3.2. Exemples d'opérations sur une table de pivot (Serveur OLAP Mondrian)



Configuration de la table de pivot

	Mesures
	Unit Sales
	Product
Customers	All Products
All Customers	266 773

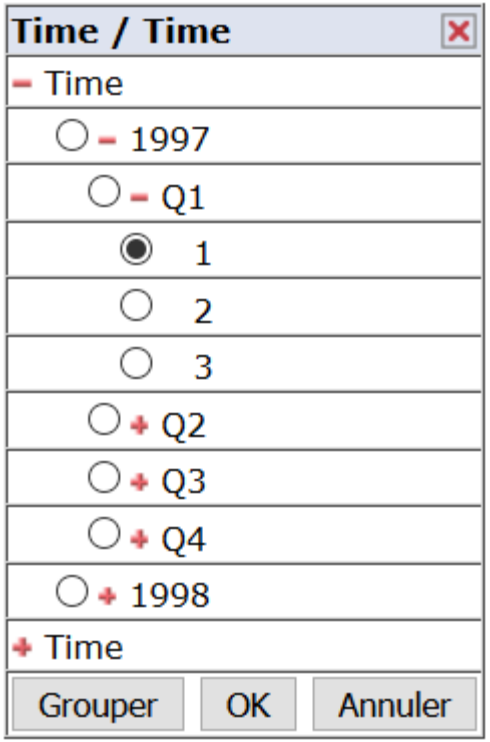
Table de pivot initiale (cube apex)

	Mesures			
	Unit Sales			
	Product			
Customers	-All Products	+Drink	+Food	+Non-Consumable
+All Customers	266 773	24 597	191 940	50 236

« Drill down » sur la dimension « Product »

	Mesures			
	Unit Sales			
	Product			
Customers	-All Products	+Drink	+Food	+Non-Consumable
-All Customers	266 773	24 597	191 940	50 236
-USA	266 773	24 597	191 940	50 236
+CA	74 748	7 102	53 656	13 990
+OR	67 659	6 106	48 537	13 016
-WA	124 366	11 389	89 747	23 230
+Anacortes	766	82	541	143
+Ballard	2 559	214	1 899	446
+Bellingham	758	68	548	142
+Bremerton	12 177	1 160	8 796	2 221
+Burien	2 783	251	2 065	467
+Edmonds	2 041	166	1 534	341
+Everett	2 690	208	1 953	529
+Issaquah	2 063	203	1 462	398
+Kirkland	2 249	247	1 598	404
+Lynnwood	2 192	201	1 599	392
+Marysville	2 232	193	1 581	458
+Olympia	12 570	1 066	9 173	2 331
+Port Orchard	12 399	1 128	9 013	2 258
+Puyallup	11 802	1 040	8 501	2 261
+Redmond	2 105	137	1 519	449
+Renton	2 212	225	1 568	419
+Seattle	1 885	168	1 381	336
+Sedro Woolley	713	58	498	157
+Spokane	23 591	2 238	16 925	4 428
+Tacoma	10 885	986	7 779	2 120
+Walla Walla	2 203	191	1 622	390
+Yakima	11 491	1 159	8 192	2 140

« Drill down » sur la dimension « Customers »



« Slice » sur la dimension « Time »
(Mois de Janvier 1997)

	Mesures			
	Unit Sales			
	Product			
Customers	All Products	Drink	Food	Non-Consumable
All Customers	21 628	1 910	15 604	4 114
USA	21 628	1 910	15 604	4 114
CA	5 377	524	3 843	1 010
OR	6 909	574	4 939	1 396
WA	9 342	812	6 822	1 708
Anacortes	81	8	61	12
Ballard	189	19	140	30
Bellingham	37	5	26	6
Bremerton	908	53	659	196
Burien	185	8	155	22
Edmonds	126	13	98	15
Everett	227	9	180	38
Issaquah	189	28	131	30
Kirkland	214	25	148	41
Lynnwood	118	15	86	17
Marysville	124	16	88	20
Olympia	980	86	702	192
Port Orchard	748	63	567	118
Puyallup	969	73	709	187
Redmond	219	15	166	38
Renton	129	15	80	34
Seattle	191	12	153	26
Sedro Woolley	67	9	53	5
Spokane	1 592	160	1 145	287
Tacoma	920	66	673	181
Walla Walla	146	4	113	29
Yakima	983	110	689	184

Slicer: [Month=1]

	Mesures			
	Unit Sales			
	Product			
Customers	-All Products	+Drink	+Food	+Non-Consumable
-All Customers	21 628	1 910	15 604	4 114
-USA	21 628	1 910	15 604	4 114
+CA	5 377	524	3 843	1 010
+OR	6 909	574	4 939	1 396
+WA	9 342	812	6 822	1 708

Slicer: [Month=1]

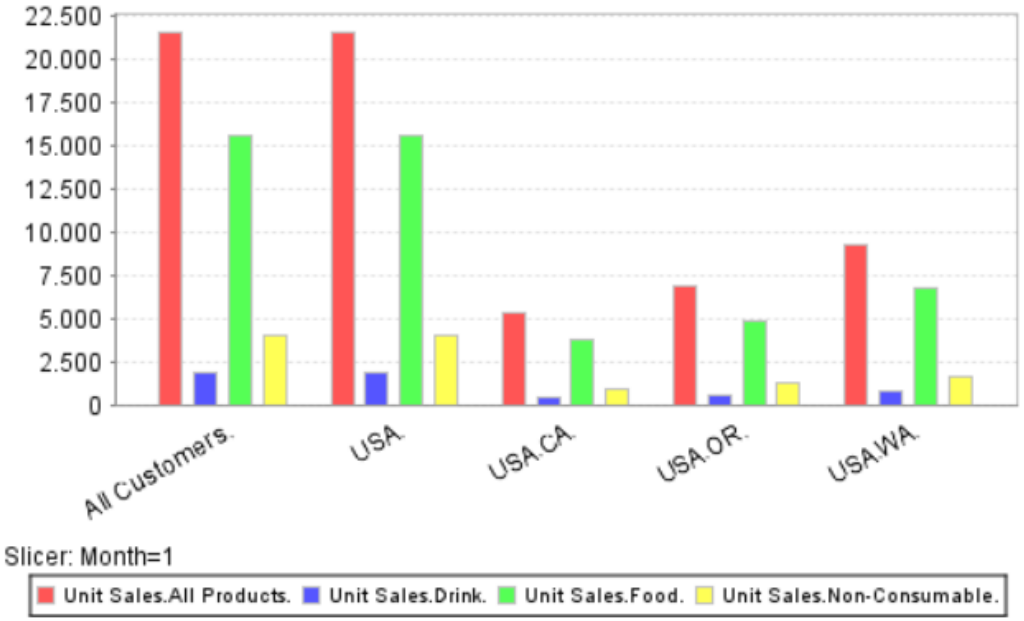
« Roll Up » sur la dimension « Customers »

		Customers				
Mesures	Product	-All Customers	-USA	+CA	+OR	+WA
Unit Sales	-All Products	21 628	21 628	5 377	6 909	9 342
	+Drink	1 910	1 910	524	574	812
	+Food	15 604	15 604	3 843	4 939	6 822
	+Non-Consumable	4 114	4 114	1 010	1 396	1 708

Slicer: [Month=1]

Inversion lignes et colonnes

		Customers				
Mesures	Product	-All Customers	-USA	+CA	+OR	+WA
Unit Sales	-All Products	21 628	21 628	5 377	6 909	9 342
	+Drink	1 910	1 910	524	574	812
	+Food	15 604	15 604	3 843	4 939	6 822
	+Non-Consumable	4 114	4 114	1 010	1 396	1 708



Génération d'un graphique représentant les dimensions de la table de pivot

4. Architectures OLAP

4.1. ROLAP

» « *Relational OLAP* » : l'entrepôt de données est géré par un SGBD relationnel (Kimball, 1996)

» Le serveur OLAP interprète la structure multidimensionnelle de l'entrepôt et gère les requêtes côté utilisateur

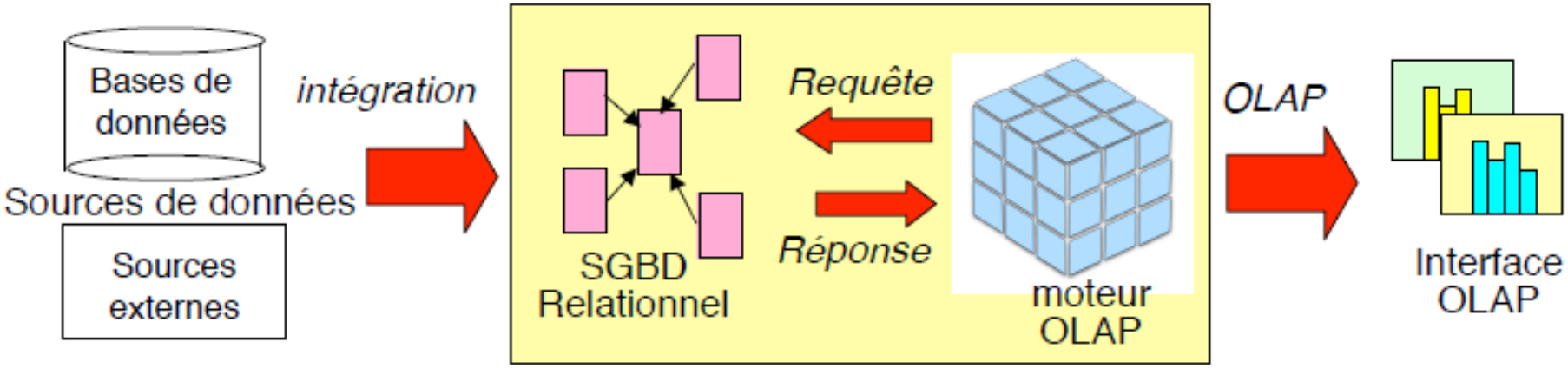
» Architecture la plus populaire

» Avantages :

- Grande capacité de stockage
- Grand choix d'outil exploitant la technologies relationnelle
- Structure logique simple
- Efficace pour des cubes « non-denses »

» Désavantages : requêtes lentes car les données ne sont pas directement stockées sous formes de cubes (ou hypercubes)

» Exemple d'outil ROLAP *Open Source*: Mondrian



Architecture ROLAP (d'après Espinasse, 2014)

» Modélisation relationnelle du schéma conceptuel en étoile

- Schéma logique en étoile
  - La table des faits (centrale) stocke les faits détaillés et leurs mesures
    - Peu contenir un très grand nombre d'enregistrements (augmente exponentiellement avec le nombre de dimensions)
    - Gestion de la « faible densité » : les faits associés à des mesures nulles ne sont pas stockés
  - La table des faits est reliée à des tables de dimensions
    - Une table de dimension contient également les niveaux hiérarchiques de cette dernière

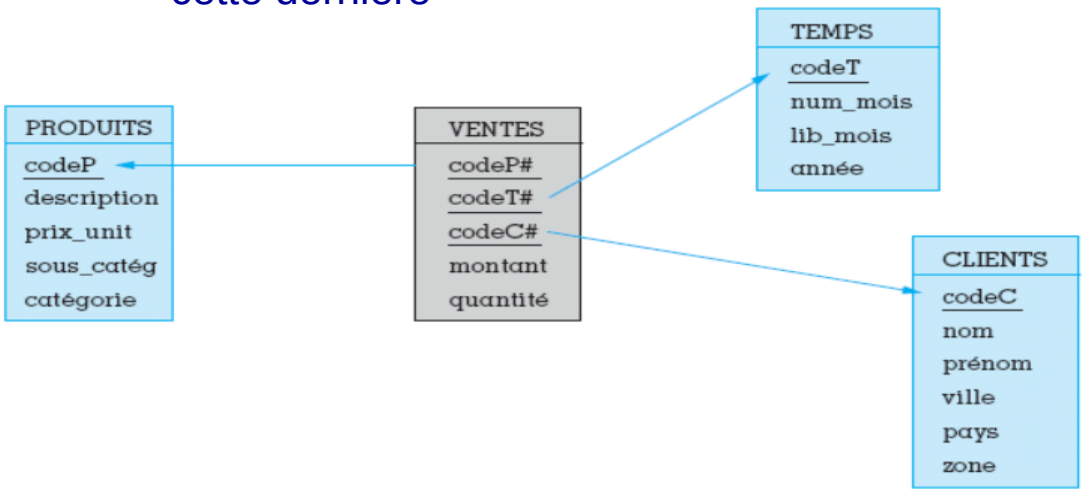


Schéma logique en étoile (d'après Chrisment et al, 2005)

- Schéma logique en flocon
  - Variante du schéma logique en étoile
  - Une table par niveau de dimension
  - Modèle présentant moins de redondance sur la définition des dimensions...

  … mais ne fonctionne pas avec une hiérarchie non-couvrante

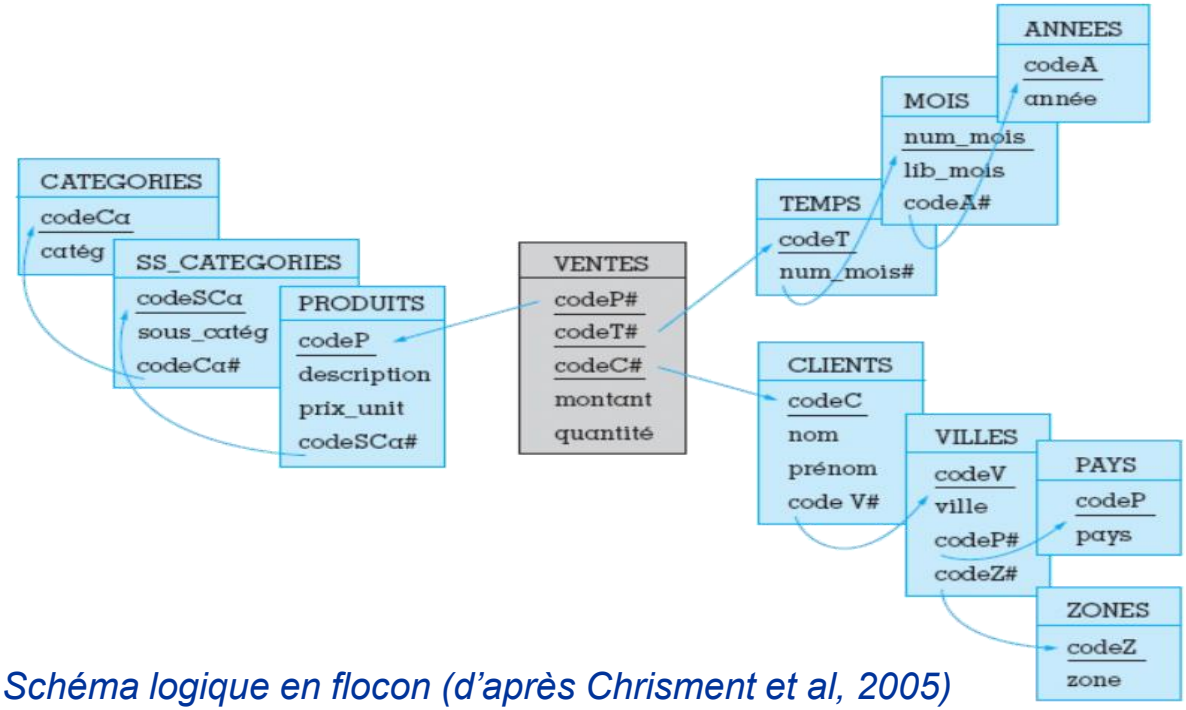


Schéma logique en flocon (d'après Chrisment et al, 2005)

- Autres types de schémas logiques:
  - Schéma en constellation
    - plusieurs tables des faits partageant des tables de dimensions
    - permet la modélisation des hiérarchies dynamiques
  - Table dégénérée:
    - entrepôt contient une seule table stockant les faits et les dimensions
    - présente énormément de redondance
    - structure la plus simple pour un entrepôt de données

» Le langage MDX (*MultiDimensional Expressions*)

- Proposé par Microsoft en 1997 dans l'extension OLAP du SGBD *SQL Server*
- Formulation plus simple des requêtes OLAP par rapport au SQL
- Aujourd'hui, langage de référence de l'OLAP
  - Langage du serveur OLAP Mondrian

```
select Crossjoin({[Measures].[Unit Sales]}, [Product].[All Products].Children) ON COLUMNS,
  [Customers].[All Customers].[USA].Children ON ROWS
from [Sales]
where [Time].[1997]
```

	Mesures		
	Unit Sales		
	Product		
Customers	Drink	Food	Non-Consumable
CA	7 102	53 656	13 990
OR	6 106	48 537	13 016
WA	11 389	89 747	23 230

Slicer: [Year=1997]

Exemple de requête MDX

4.2. MOLAP

» « *Multidimensional OLAP* » : L'entrepôt est physiquement géré de manière multidimensionnelle

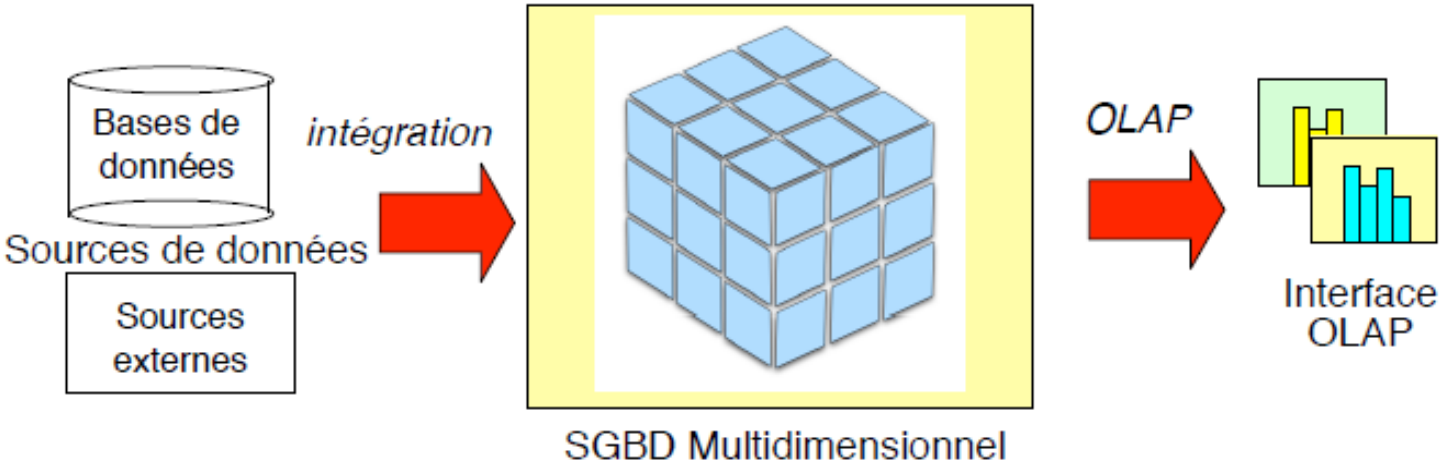
» Premier outil OLAP : Essbase (Codd, 1992)

» Les données sont directement stockées dans des tableaux (ou cubes) multidimensionnels

» Avantage : les données ne nécessitent pas de conversion relationnelle / multidimensionnelle et les requêtes sont donc plus rapides que dans un ROLAP

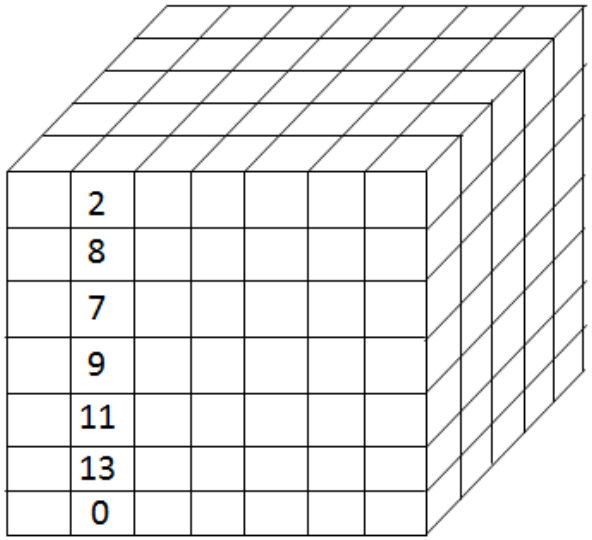
» Désavantages :

- Mauvaise gestion de la « faible densité » ➔ le nombre de dimensions est limité
- Moins de choix dans les outils existants (la plupart des solutions sont propriétaires)

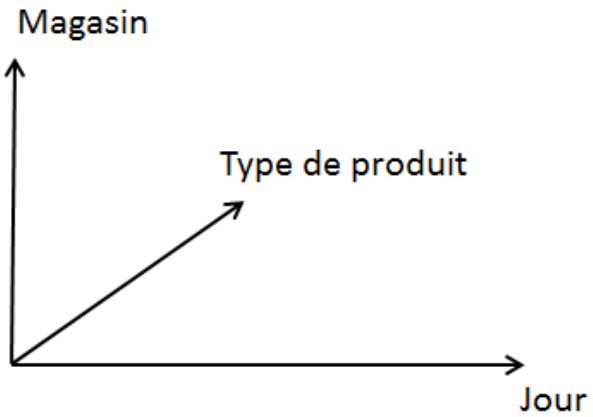


Architecture MOLAP (d'après Espinasse, 2014)

» Chaque fait (cellule d'un cube MOLAP) contient une mesure qui est indexée par les valeurs des membres dimensionnels qui le définissent



Quel est le nombre de ventes pour le produit « raquette de pingpong », le 02-01-14 dans l'ensemble des magasins ?



Agrégation des cellules
[1, 0, 0], [1,1,0], [1,2,0], [1,3,0], [1,4,0], [1,5,0] et [1,6,0]:
0 + 13 + 11+ 9 + 7 + 8 + 2 = 50

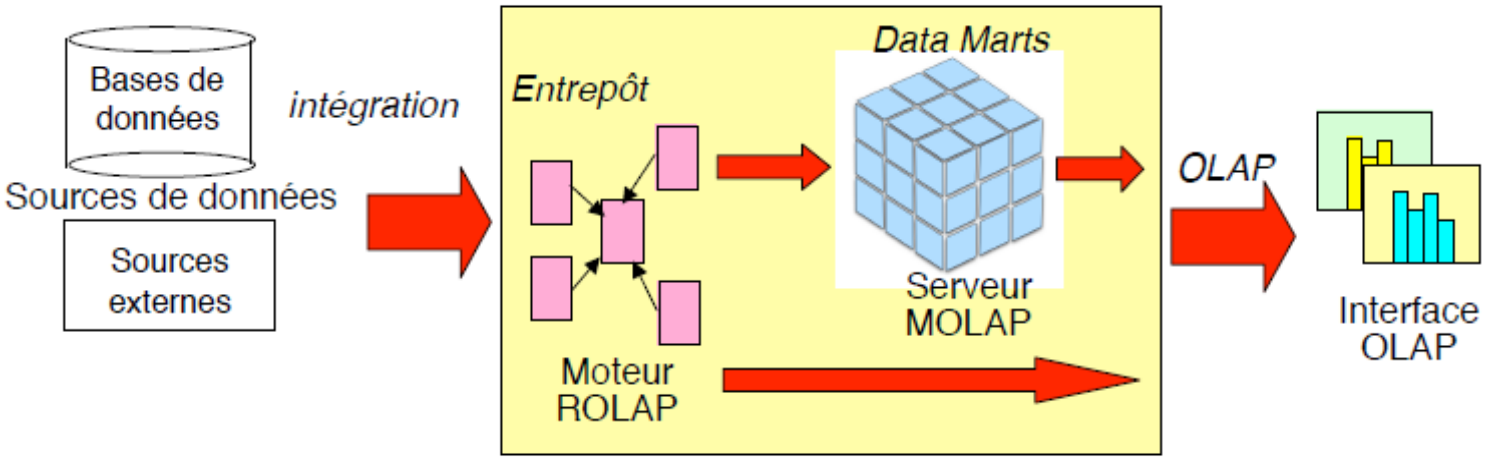
Exemple de requête sur un cube MOLAP

– **4.3. HOLAP**

» « *Hybrid OLAP* »: architecture exploitant à la fois les avantages du ROLAP et du MOLAP

» Un SGBD relationnel stocke toutes les données du système et un moteur ROLAP exploite directement ces données

» Certains cuboïdes construits à partir du SGBD relationnel sont dupliqués en MOLAP (« *datamarts* »)



Architecture HOLAP (d'après Espinasse, 2014)

5. SOLAP

– *5.1. Généralités*

» *OLAP + SIG = SOLAP*

» *« une plateforme visuelle conçue spécialement pour supporter une analyse spatio-temporelle rapide et efficace à travers une approche multidimensionnelle qui comprend des niveaux d'agrégation cartographiques, graphiques et tabulaires » (Bédard, 1997)*

» Un serveur SOLAP exploite un entrepôt de données spatialisées

- En général:
  - à travers le modèle vectoriel
  - pour la représentation de phénomènes spatialement discrets
  - architecture ROLAP

» La représentation cartographique permet la visualisation de dimensions géographiques (ou « dimensions spatiales ») et/ou de mesures spatiales

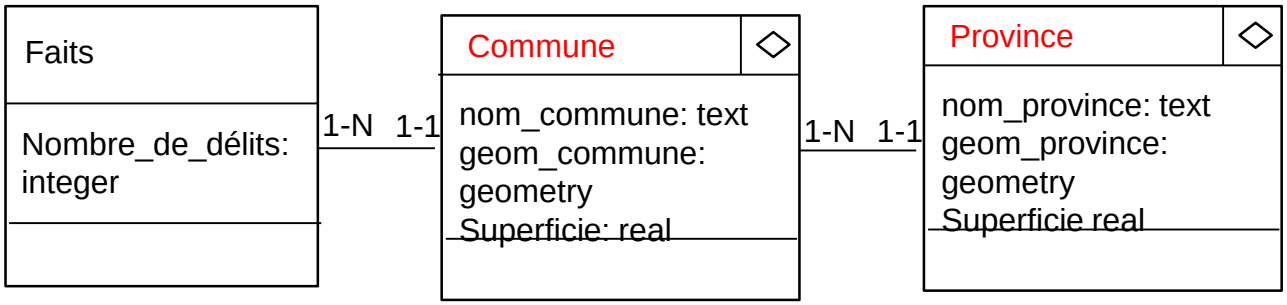
» Application dans de nombreux domaines: geomarketing, aménagement du territoire, agriculture, crime mapping, etc.

5.2. Spatialisation de l'entrepôt de données (vectoriel)

» Dimension géographique

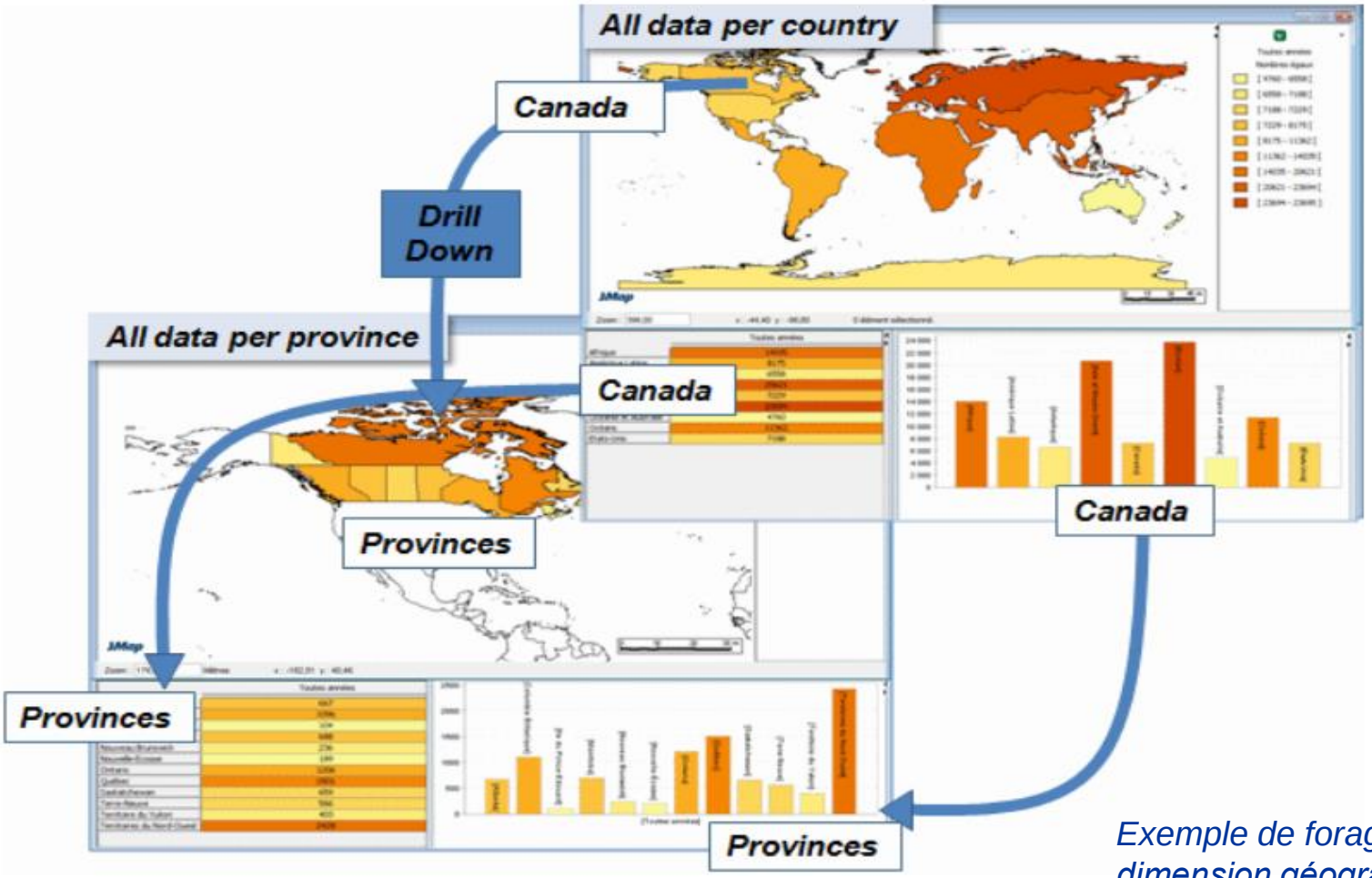
- Membre géographique: membre de dimension auquel est associée une géométrie (vectorielle)
  - Caractérisé par un aspect sémantique et un aspect géométrique
  - Exemple: Nom_commune: Liège

- Une dimension est géographique lorsqu'au moins un niveau de la dimension est composé de membres géographiques:
  - Exemple: le niveau « commune » de la dimension « entités administratives »



Représentation UML d'une dimension géographique

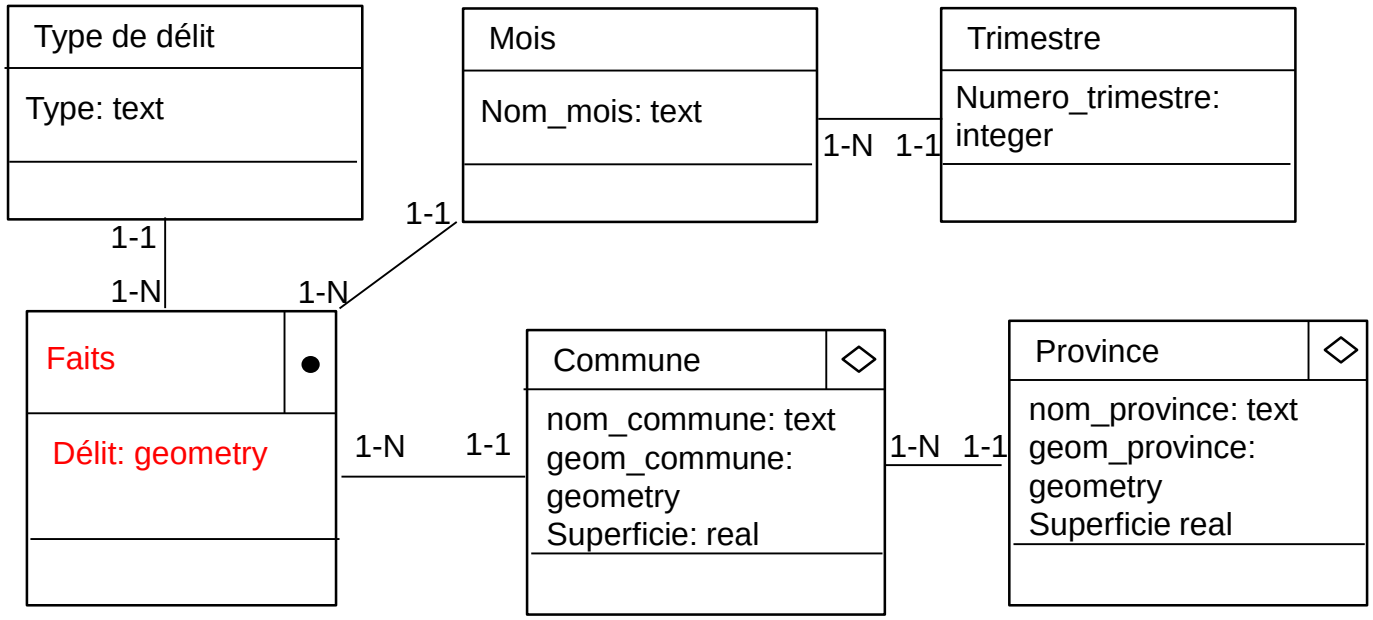
- L'interface cartographique représente des faits géographiques (impliquant une dimension géographique)
  - Les mesures sont affichées en exploitant les règles de symbolisation de cartographie thématique
- Operations SOLAP appliquées sur dimensions géographiques:
  - « *Spatial Drill Down* » / « *Spatial Roll Up* »
  - « *Spatial Slice* » / « *Spatial Dice* »
- Une dimension géographique peut toujours être représentée dans un tableau ou un graphique (aspect sémantique uniquement)



Exemple de forage spatial sur une dimension géographique avec l'outil Map4Decision (d'après Bédard et all, 2009)

» Mesure spatiale

- La mesure est une entité vectorielle
  - Nécessite des fonctions d'agrégations spatiales particulières :
    - Union, enveloppe convexe, intersection, *etc.*
  - Exemple: des délits représentés par des points



Représentation UML d'une mesure spatiale

- La mesure est une valeur numérique résultant d'un opérateur d'analyse spatiale (mesure dérivée)
  - Exemples: superficie, distance cumulée, densité, *etc*.

5.3. Solutions SOLAP

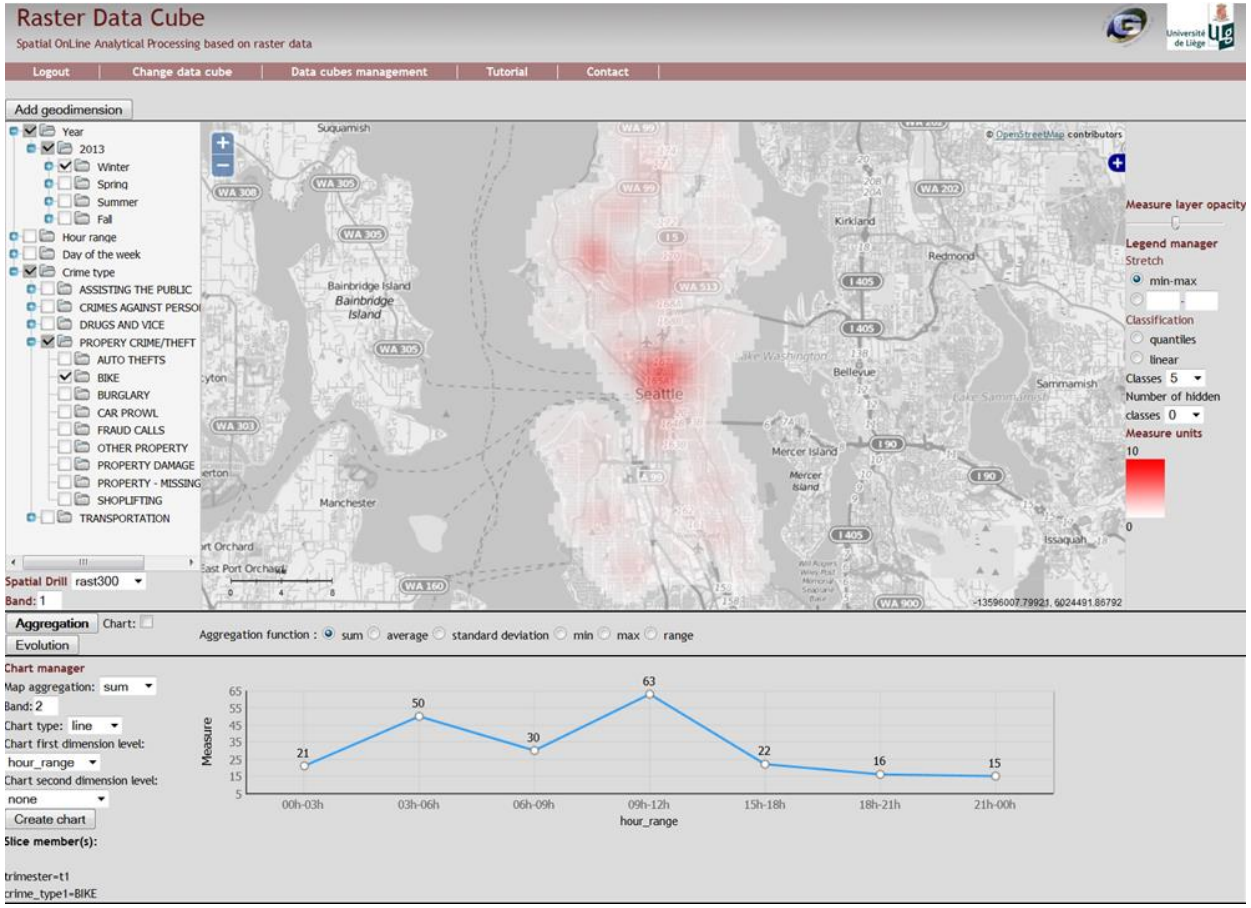
» Solution **OLAP dominant**

- On ajoute des fonctionnalités cartographiques basiques à un outil OLAP
  - Fonctionnalités classiques de navigation cartographique : zoom, pan...
  - Pas de forage spatial
  - Pas de modification de la symbolisation des cartes
  - Une seule dimension géographique
  - Un seul type d'entité spatiale
  - Interactions entre la carte et interfaces tabulaires/graphiques limitées
- Exploitation du langage MDX
- Solution à privilégier lorsque la composante spatiale est jugée « secondaire »

Exemple de solution OLAP dominant: Oracle BI Server + Oracle Java Viewer (d'après Proulx, 2009)

» Solution **SIG dominant**

- On ajoute des fonctionnalités OLAP à un SIG
- Exploitation d'un SGBD spatial tel que PostgreSQL/PostGIS
- Programmé directement en SQL et non en MDX
  - Pas de table de pivot
- Exploitation poussée des fonctionnalités SIG :
  - Navigation cartographique
  - Analyse spatiale
  - Symbolisation
  - Gestion de tous les types d'entités vectorielles
  - Gestion du modèle raster
  - Gestion des phénomènes spatialement continus (interpolations)
  - Gestion de plusieurs dimensions géographiques
- Solution à privilégier lorsque la composante spatiale est jugée importante



Exemple de solution SIG dominant exploitant des données raster: prototype « Raster Data Cube » (Kasprzyk, 2015)

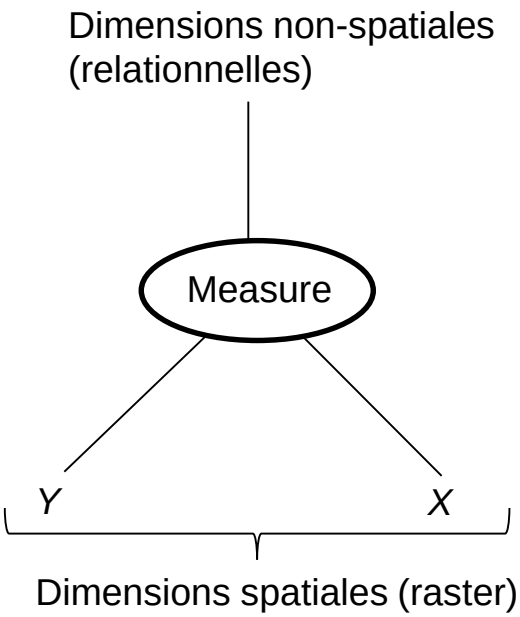


Schéma en étoile d'un entrepôt de données raster (Kasprzyk, 2015)