Partie 1

Data Warehouse

Entrepôts de données

Chapitre 3 : Analyse OLAP (Traitement analytique en ligne)

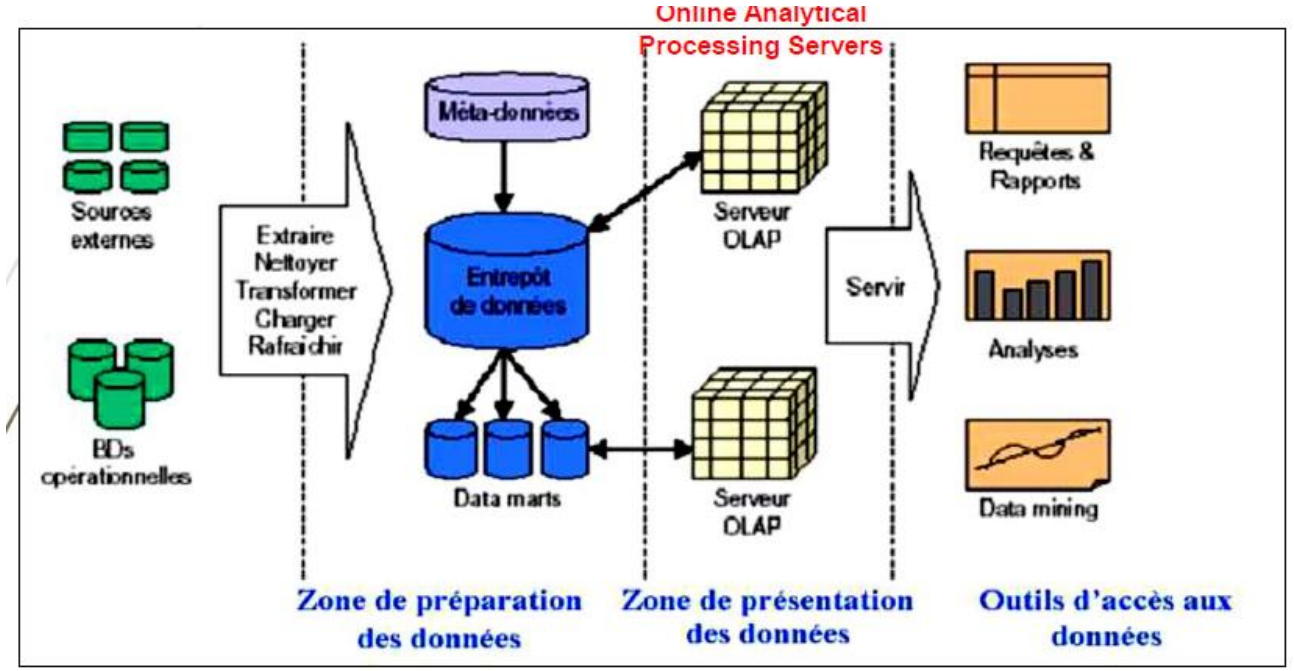
Dr H. EL BOUHISSI Epse BRAHAMI

Plan

- Introduction
- Analyse OLAP
- Principes de l'analyse OLAP
- Fonctions des systèmes OLAP
- Opérations d'analyse OLAP

Introduction

- Le modèle multidimensionnel présente une vue statique des données.
- Il a besoin d'être manipulé pour extraire des informations nécessaires à la prise de décision.
- L'exploitation des données multidimensionnelles peut se faire par divers outils (reporting, OLAP, fouille de données).



Analyse OLAP

Le Terme OLAP a été proposé par Codd (1993).

E.F. Codd

1923-2003

OLAP est un acronyme pour « Online Analytical Processing ». OLAP effectue une analyse multidimensionnelle des données d'entreprise et offre la possibilité de calculs complexes, d'analyses de tendances et de modélisation de données sophistiquées.

OLAP permet aux utilisateurs finaux d'effectuer des analyses de données dans de multiples dimensions, leur fournissant ainsi les informations et la compréhension dont ils ont besoin pour prendre de meilleures décisions.

Analyse OLAP

OLAP est un système de traitement analytique en ligne. La base de données OLAP stocke les données historiques entrées par OLTP. Il permet à l'utilisateur de visualiser différents résumés de données multidimensionnelles. Avec OLAP, vous pouvez extraire des informations d'une base de données volumineuse et les analyser en vue d'une prise de décision.

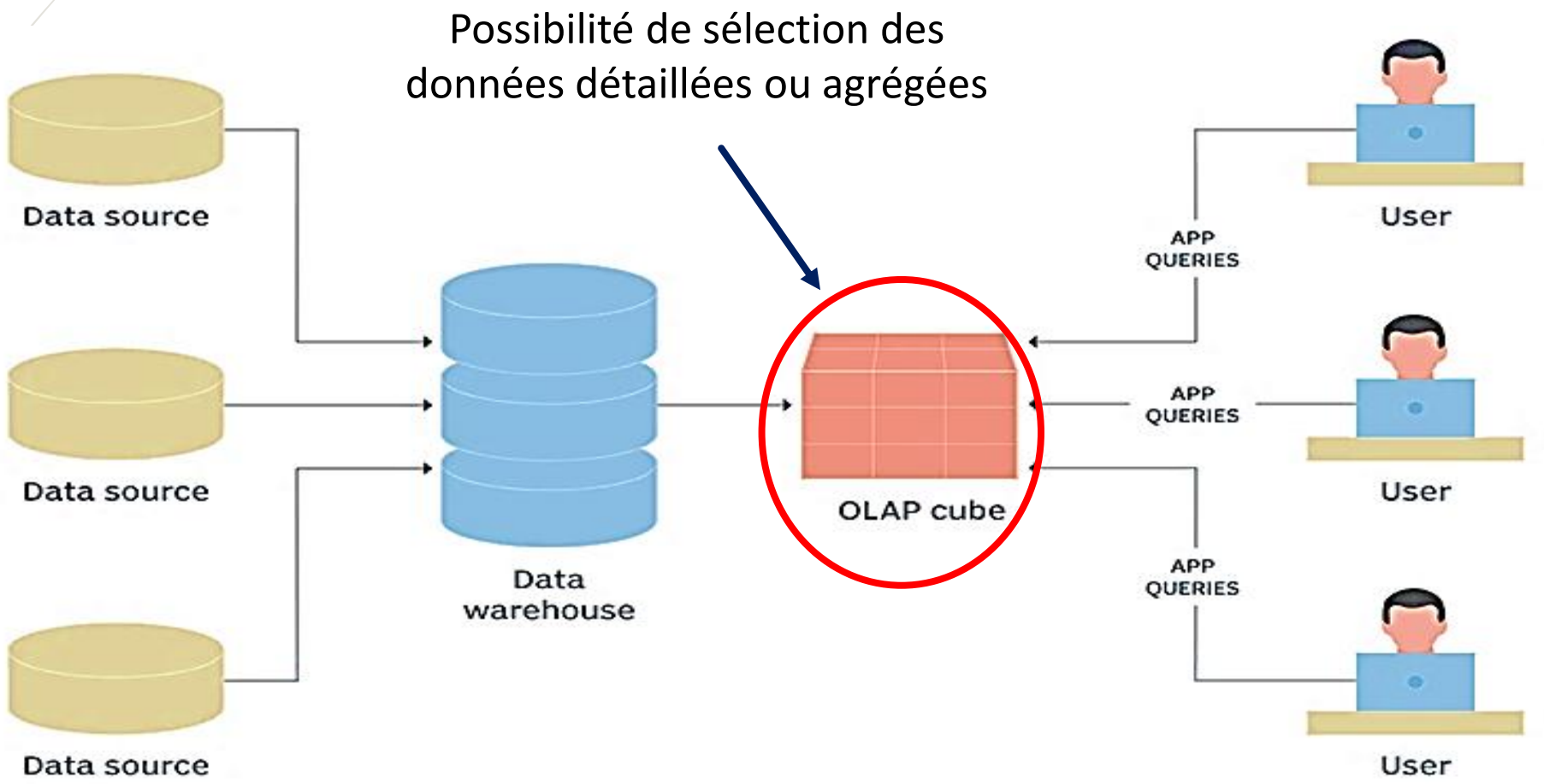
Rappel → *OLTP est un système de traitement des transactions en ligne. L'objectif principal du système OLTP est d'enregistrer la mise à jour, l'insertion et la suppression en cours de transaction. Les requêtes OLTP sont plus simples et plus courtes et nécessitent donc moins de temps de traitement, ainsi que moins d'espace.*

OLTP VS OLAP

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 »

Le processus OLAP

Comment les données sont préparées pour une analyse analytique en ligne (OLAP)



Fonctions d'un serveur OLAP

- Présenter une vue multidimensionnelle des données.
- Présenter les hiérarchies d'analyse.
- Permettre le partage de données.
- Connexion aux supports de restitution (feuilles de calcul Excel, …).
- Calcul des agrégats.
- Navigation souple dans les données

Cube OLAP

- Cube de données = instance d'un schéma en étoile
- Les cellules du cube de base contiennent les mesures des faits détaillés (mesures atomiques)
- Un cube de données à plus de trois 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

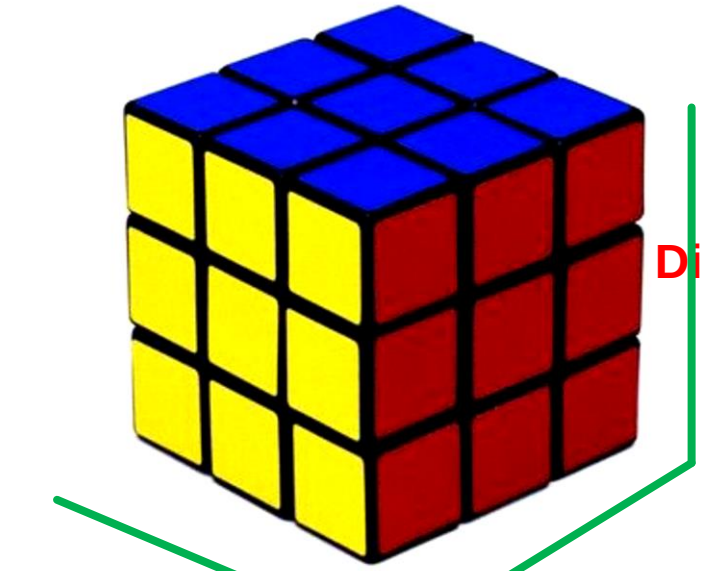
Cube OLAP

Un cube OLAP est une structure de données multidimensionnelle stockant les faits comme des mesures , indexées par plusieurs dimensions.

Graphiquement, limité à trois dimensions, au-delà de trois, difficile à schématiser.

Chaque cellule d'un cube représente la mesure ou valeur quantitative d'un fait sur le croisement de plusieurs dimensions.

L'intérêt d'un cube OLAP est d'offrir à l'utilisateur la capacité de faire des analyses multidimensionnelles ou des agrégations par axe de dimension dans l'espace.



Cube OLAP

Un cube est composé de cellules organisées en groupes de mesures et en dimensions. Une cellule représente l'unique intersection logique dans un cube d'un membre de chaque dimension du cube.

Il n'est pas nécessaire que toutes les cellules d'un cube contiennent une valeur ; il peut exister dans un cube des intersections qui ne contiennent pas de données. Ces intersections, appelées cellules vides, sont même fréquentes dans les cubes, car une intersection entre un attribut de dimension et une mesure à l'intérieur d'un cube ne contient pas nécessairement un enregistrement correspondant dans une table de faits.

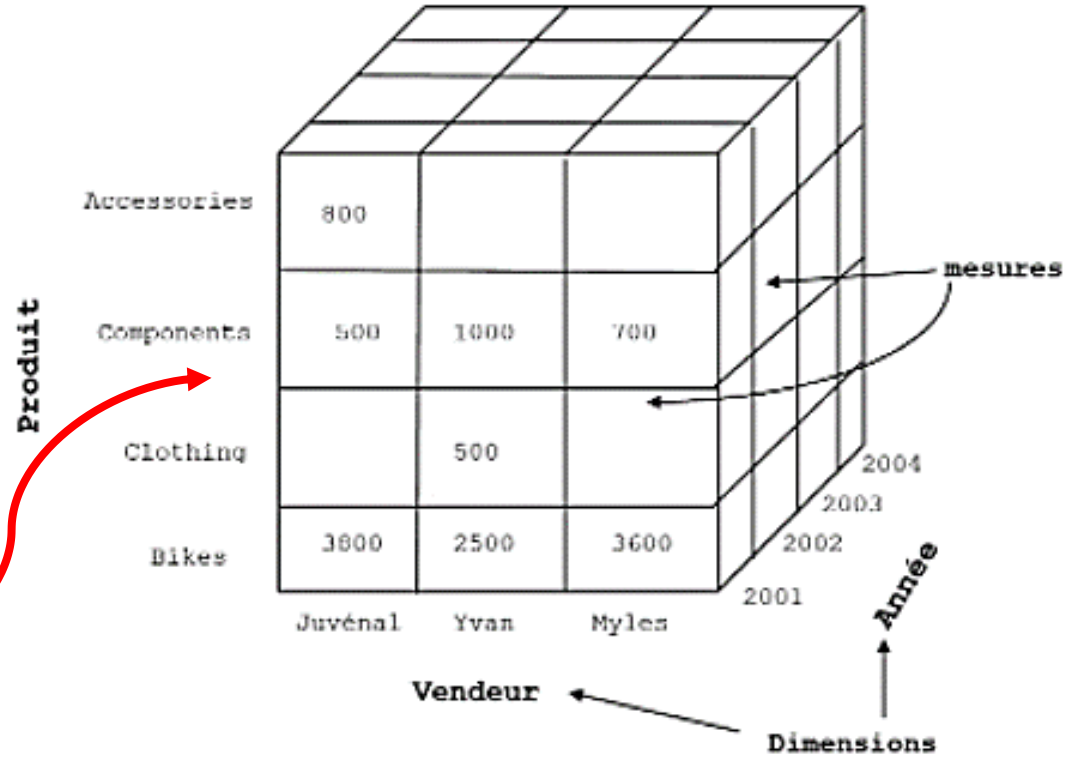
Le ratio de cellules vides dans un cube au nombre total de cellules d’un cube est souvent appelé éparsité d’un cube.

Exemple d'un cube OLAP (1)

Vendeur	Produit	Date de vente	Prix de vente
Juvénal	Accessories	01/04/2001	800
Myles	Bikes	09/05/2001	1400
yvan	Clothing	02/02/2002	500
yvan	Components	02/03/2002	1000
Juvénal	Bikes	15/03/2002	1800
Juvénal	Bikes	10/03/2003	2000
Myles	Components	12/10/2003	700
Myles	Bikes	25/12/2003	2200
Juvénal	Components	10/01/2004	500
…	……	……	……
yvan	Bikes	15/11/2004	2500

Nous souhaitons calculer la somme des ventes par produit et par année

Représentation du tableau en cube OLAP. Les 3 colonnes de catégories deviennent des dimensions dans le cube, tandis que le prix devient la mesure, la valeur correspondant au croisement des 3 dimensions simultanément



Exemple d'un cube OLAP (2)

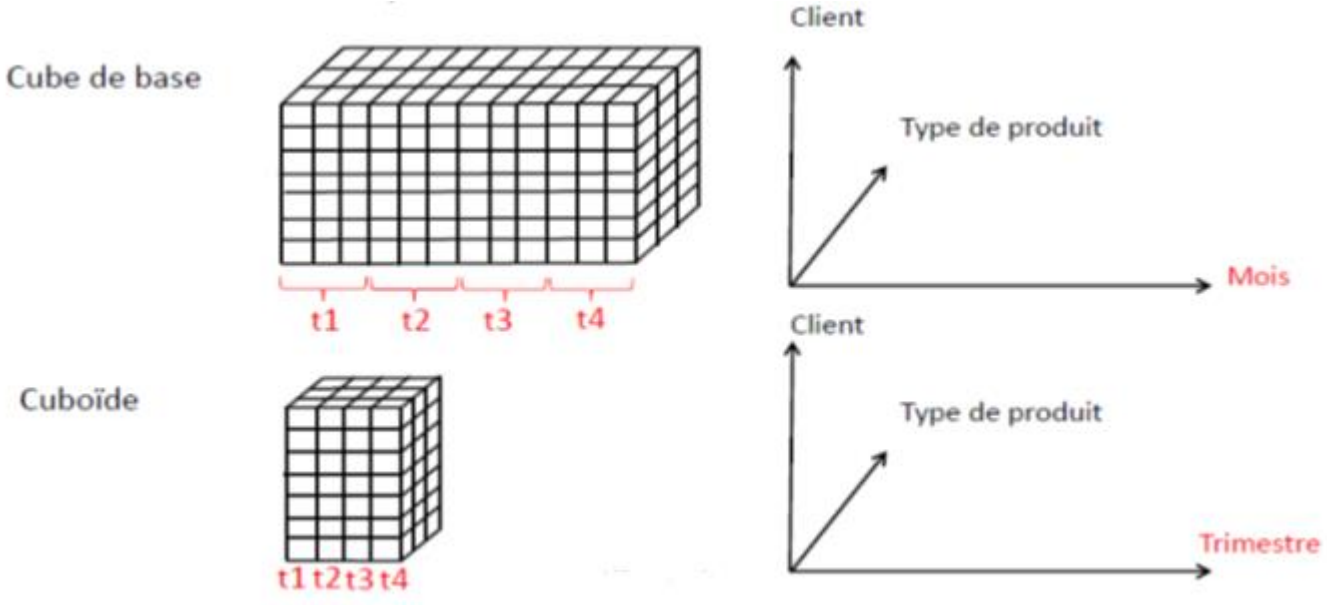
Dans le cube, nous avons trois dimensions : Année, Vendeur et Produit ; la mesure c'est le prix de vente. Pour rendre le cube fonctionnel, on applique à ses cellules (donc à la mesure) une fonction d'agrégation, cette fonction peut être soit une somme, une moyenne, un maximum ou un minimum, toute fonction qui regroupe ou compare un ensemble de données.

Pour faciliter la représentation graphique du cube, nous avons présenté la somme des mesures uniquement sur la première face, mais en réalité, toutes les cellules comportent des valeurs pour chaque recoupement de dimensions et répondent à une requête bien précise. Par exemple la cellule qui est à l'intersection de l'attribut « *Bikes* » de la dimension Produit et de l'attribut « Juvénal » de la dimension Vendeur a pour somme de ventes 3800; ceci correspond à la requête chiffre d'affaire réalisé par le Vendeur Juvénal sur les Produits « Bikes » depuis 2001.

Cependant, certains croisements de dimensions peuvent ne pas avoir de valeur (agrégation nulle dans ce cas), c'est ce qui explique que certaines cellules du cube soient vides.

Cuboïde de données

- Un cuboïde de données contient les faits à un niveau non détaillé
- Certains cuboïdes sont pré-calculés dans l'entrepôt pour optimiser les requêtes
- En pratique, entièrement décomposer un cube de données en cuboïdes est impossible car ils sont trop nombreux (un cuboïde par combinaison possible de niveaux de dimension)



Fonctionnement des systèmes OLAP

Pour faciliter ce type d'analyse, les données sont recueillies à partir de multiples sources de données et stockées dans des ***Data Warehouse***, puis nettoyées et organisées en cubes de données.

Chaque cube OLAP contient des données classées par dimensions (telles que les clients, la région géographique de vente et la période de temps) dérivées par tables dimensionnelles dans les Data Warehouse.

Les dimensions sont ensuite complétées par les membres (tels que les noms de clients, les pays et les mois) qui sont organisés de manière hiérarchique. Les cubes OLAP sont souvent pré-résumés dans toutes les dimensions afin d'améliorer considérablement le temps de requête par rapport aux bases de données relationnelles.

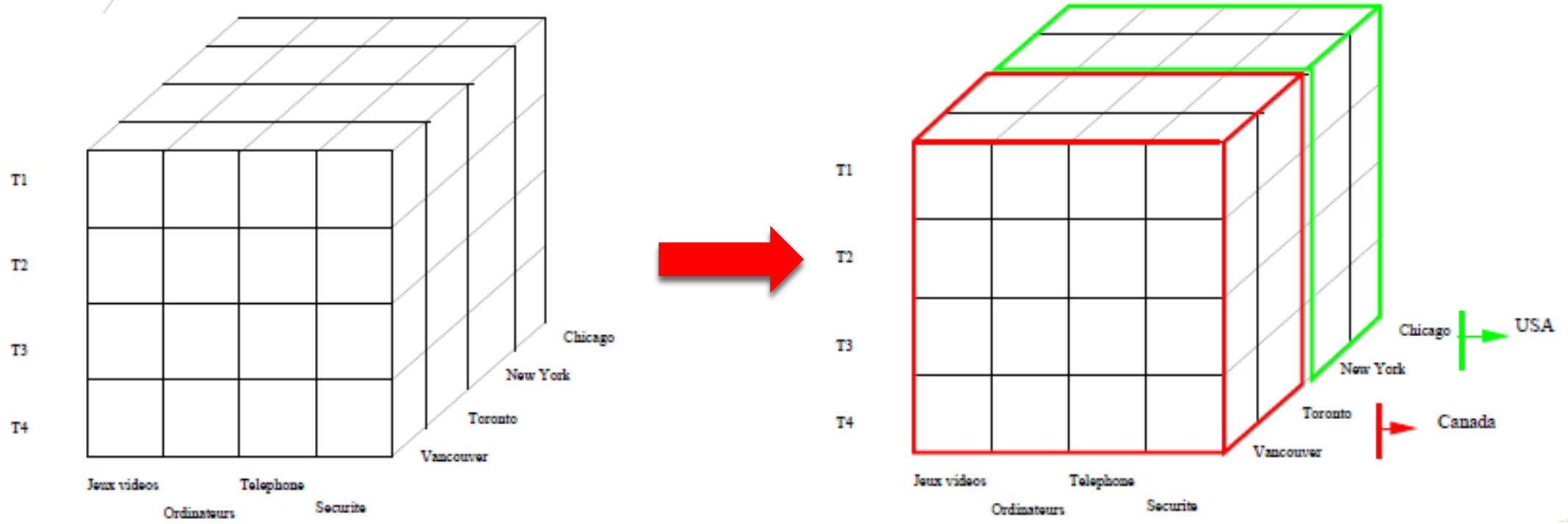
Opérations d'analyse OLAP

Les systèmes OLAP sont conçus pour repérer les intersections entre ces multiples dimensions. Les analystes peuvent ensuite effectuer quatre types d'opérations d'analyse OLAP à partir de ces bases de données multidimensionnelles :

- **ROLL-UP**
- **DRILL-DOWN**
- **SLICE**
- **DICE**

Opérations d'analyse OLAP : ROLL-UP

Compresser une dimension (hiérarchie) / supprimer une dimension.
Passer d'un niveau de détail à un niveau moins détaillé.

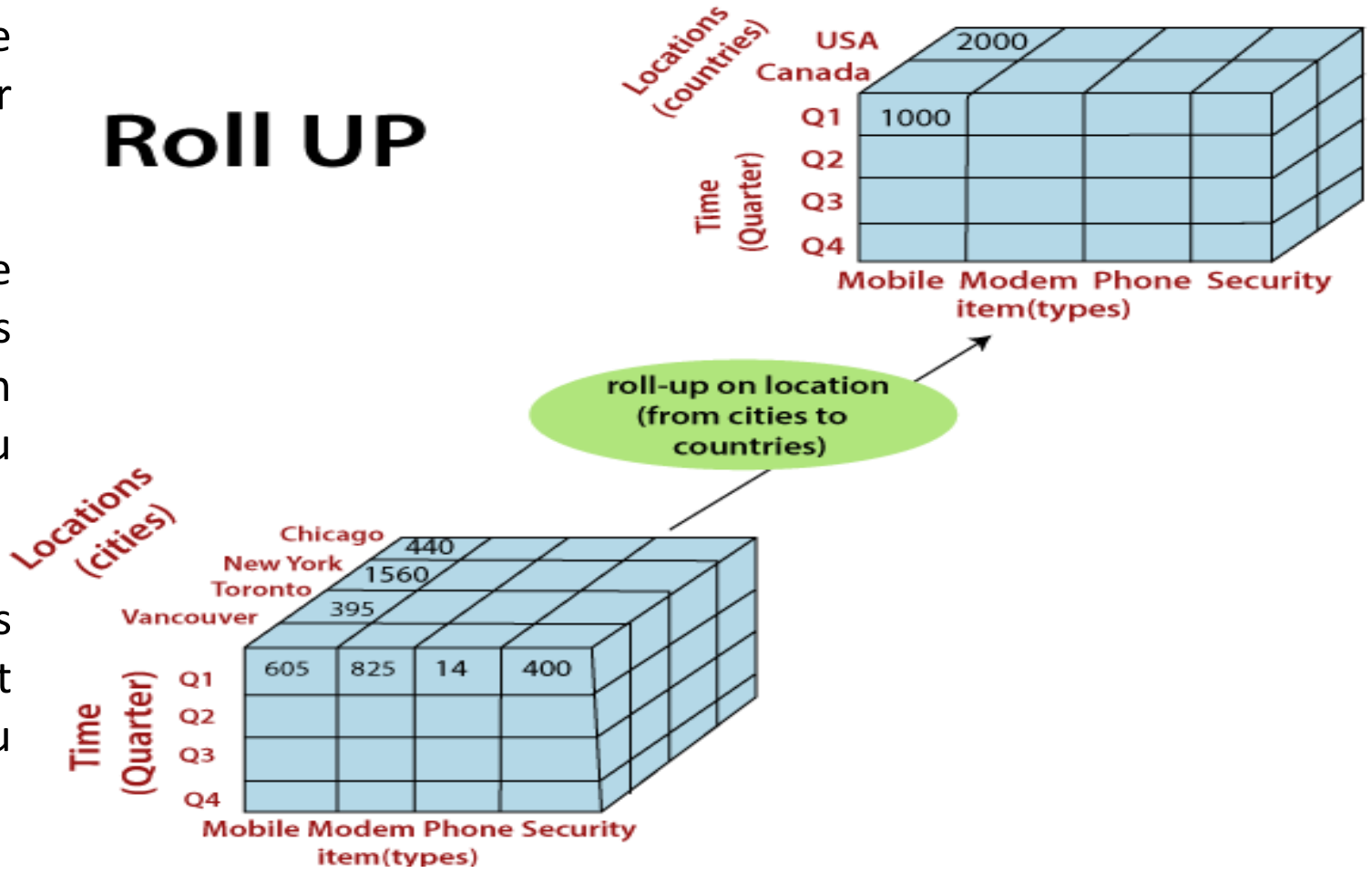


Opérations d'analyse OLAP : ROLL-UP

•Le roll-up est effectué en remontant une hiérarchie de concepts pour l'emplacement de la dimension.

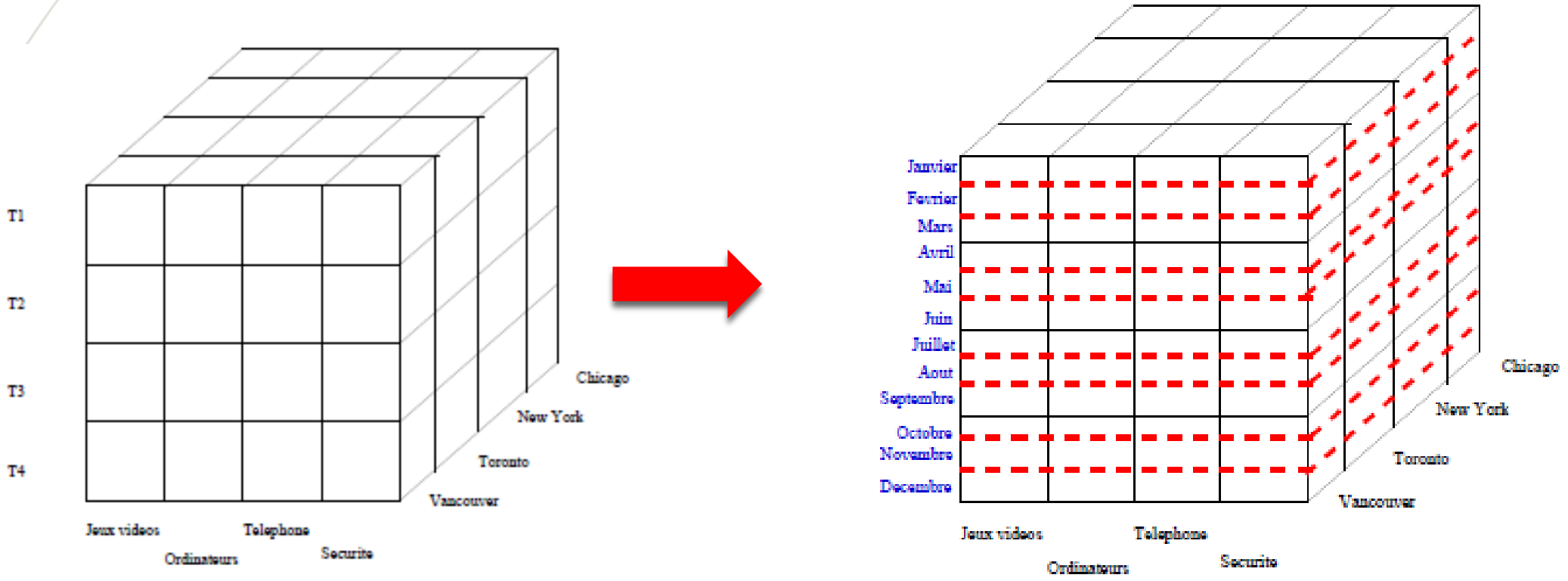
•Initialement, la hiérarchie conceptuelle était "rue < ville < province < pays". Lors du roll up, les données sont agrégées en remontant la hiérarchie de localisation du niveau de la ville au niveau du pays.

•Les données sont regroupées par villes plutôt que par pays. Lorsque le roll-up est effectué, une ou plusieurs dimensions du cube de données sont supprimées.

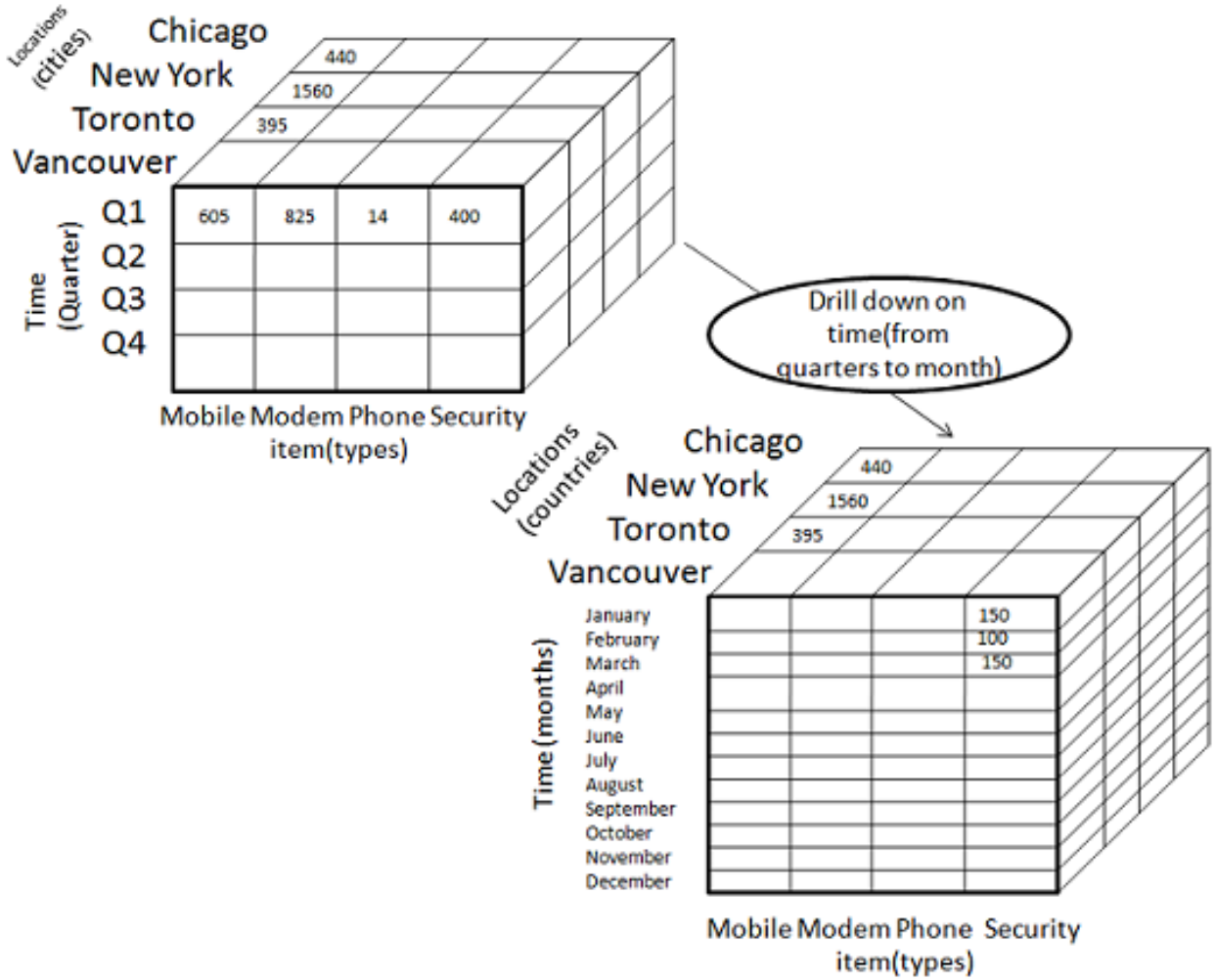


Opérations d'analyse OLAP : DRILL-DOWN

Étirer une dimension (hiérarchie) / ajouter une dimension.
Passer d'un niveau de détail à un niveau plus détaillé



Opérations d'analyse OLAP : DRILL-DOWN



•L'éclatement est effectué en descendant dans une hiérarchie de concepts pour la dimension temps. Initialement, la hiérarchie conceptuelle était "jour < mois < trimestre < année".

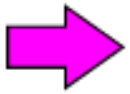
•Lors du Drill down , la dimension temps est descendue du niveau du trimestre au niveau du mois.

•Lors du drill-down, une ou plusieurs dimensions du cube de données sont ajoutées. Cela permet de naviguer des données les moins détaillées aux données les plus détaillées.

Opérations d'analyse OLAP : ROLL-UP- DRILL-DOWN

- Add up amounts by day, product
- In SQL: SELECT date, sum(amt) FROM SALE
  GROUP BY date, prodId

sale	prodId	storeId	date	amt
	p1	c1	1	12
	p2	c1	1	11
	p1	c3	1	50
	p2	c2	1	8
	p1	c1	2	44
	p1	c2	2	4



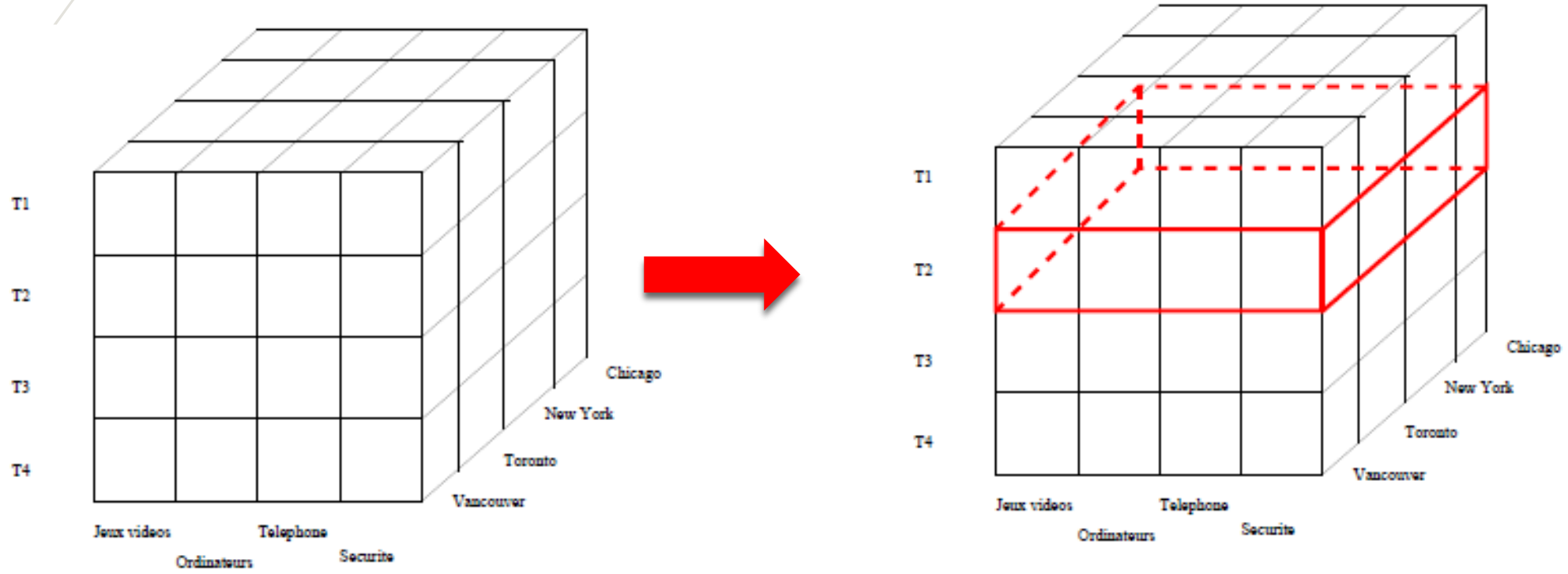
sale	prodId	date	amt
	p1	1	62
	p2	1	19
	p1	2	48

rollup →

← drill-down

Opérations d'analyse OLAP : SLICE

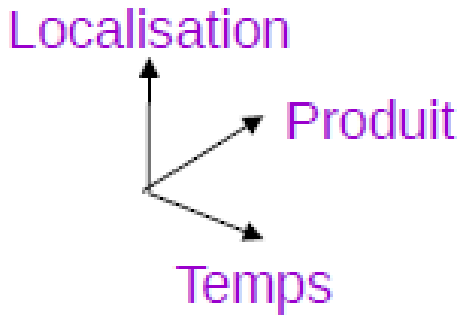
Projection selon une dimension : Couper une tranche / fixer la valeur d'une dimension.

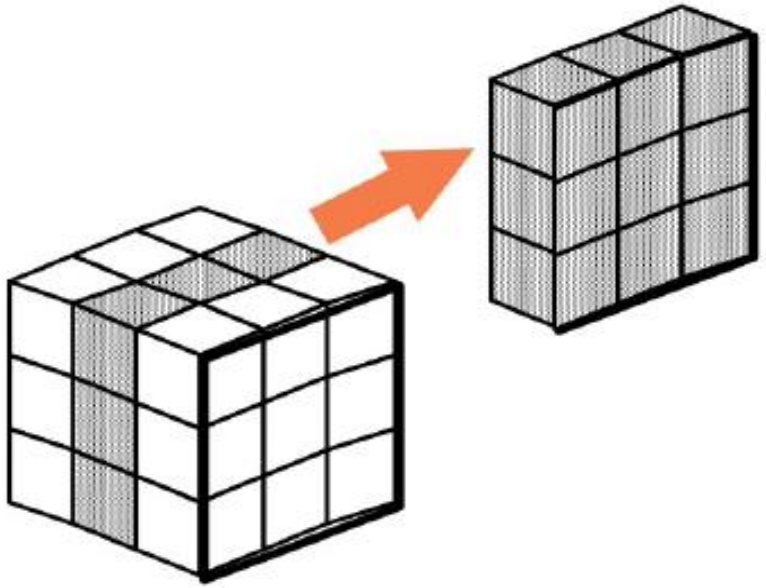


Opérations d'analyse OLAP : SLICE

		1995	1996	1997
Frais	IdF	220	265	284
	Province	225	245	240
Liquide	IdF	163	152	145
	Province	187	174	184

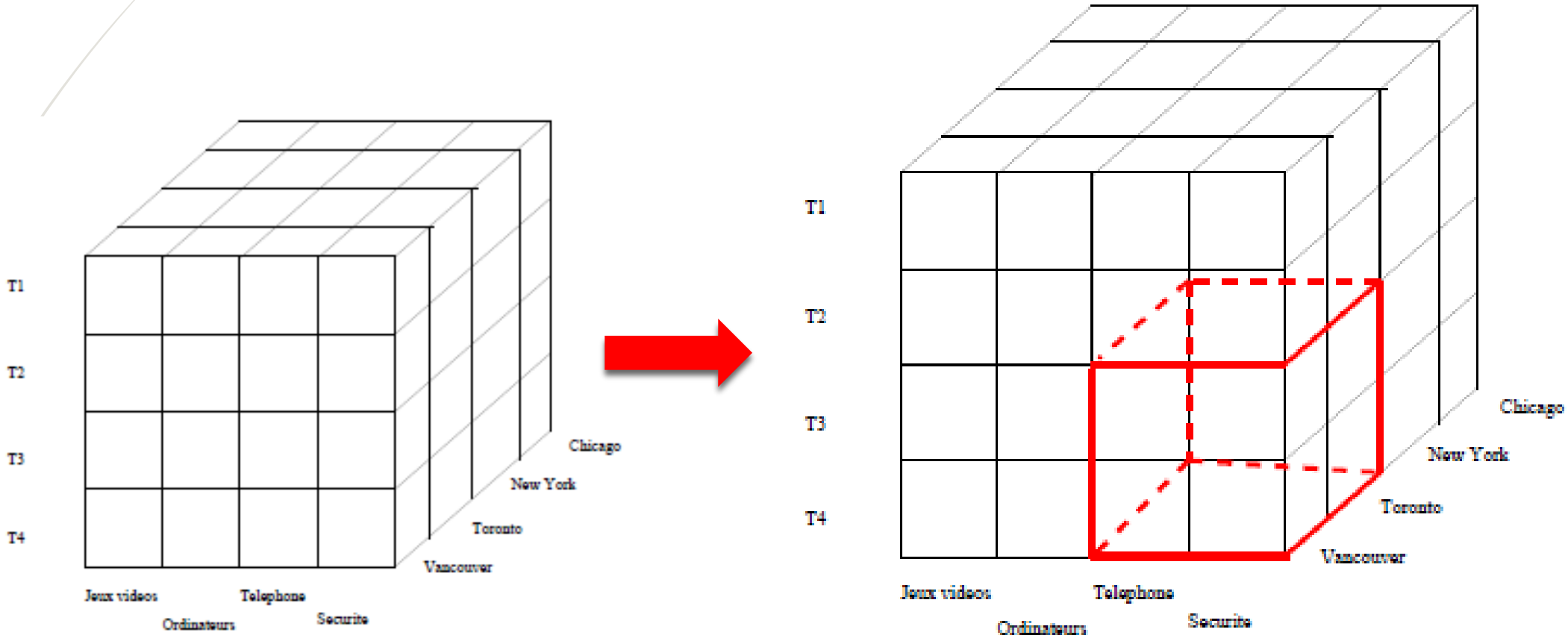
		1996
Frais	IdF	265
	Province	245
Liquide	IdF	152
	Province	174





Opérations d'analyse OLAP : DICE

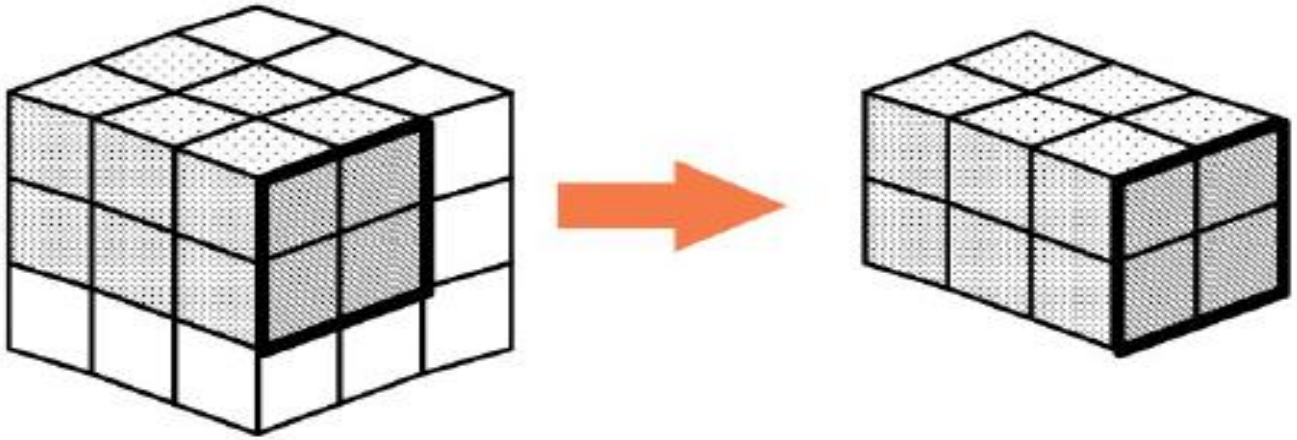
Sélection / Restriction : Isoler un sous cube.



Opérations d'analyse OLAP : DICE

		1995	1996	1997
Frais	IdF	220	265	284
	Province	225	245	240
Liquide	IdF	163	152	145
	Province	187	174	184

		1995	1996
Frais	IdF	220	265
	Province	225	245



Solutions techniques

Parmi les différentes solutions disponibles sur le marché, on compte deux solutions principales. Il s'agit **d'Hyperion Solution Essbase, et Oracle Express Server, tous deux appartenant à Oracle.**

Les produits OLAP sont généralement conçus pour des environnements regroupant plusieurs utilisateurs. De fait, **le coût des logiciels dépend du nombre d'utilisateurs**. Il faut également prendre en compte le coût du serveur associé.

Différents types de systèmes OLAP

Quel type d'implémentation choisir ?

Types de systèmes OLAP: Approche ROLAP

ROLAP : **Relationnal OLAP** : opérations traduites en SQL.

L'entrepôt de données est géré par un SGBD relationnel

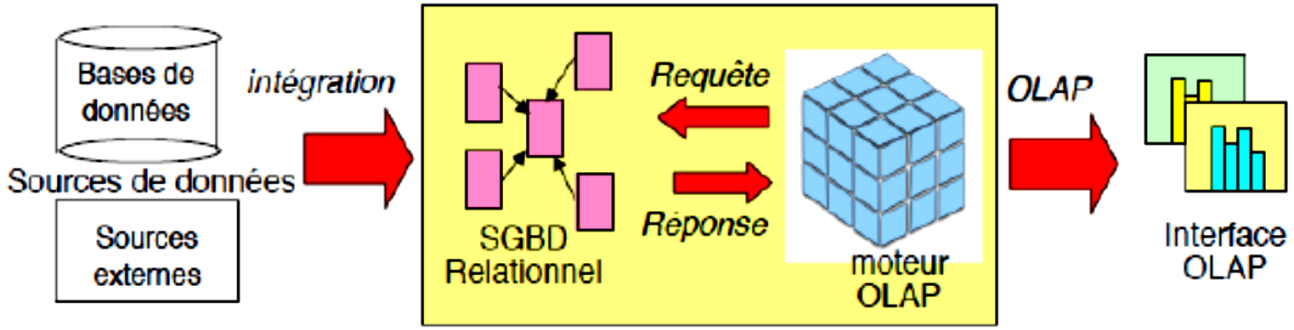
Le serveur OLAP interprète la structure multidimensionnelle de l'entrepôt et gère les requêtes côté utilisateur

Avantages

- Facilité et faible coût de mise en œuvre
- Stockage de gros volumes de données
- Evolution facile

Inconvénients

- Performance (jointures)
- Reformatage nécessaire des résultats pour les utilisateurs finaux



Types de systèmes OLAP: Approche MOLAP

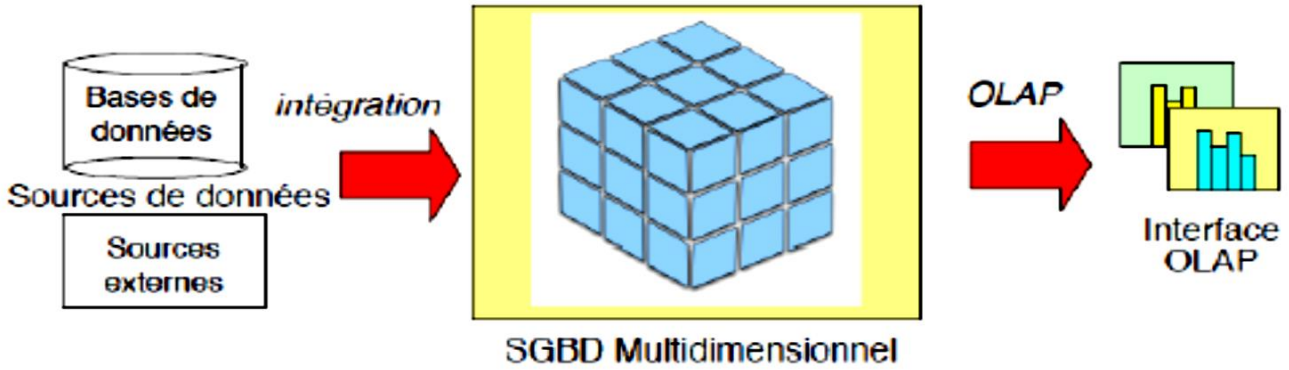
MOLAP : **Multidimensionnal OLAP** : Stockage natif des cubes dans des tableaux multidimensionnels.

Avantages

- Calculs d'agrégats rapides

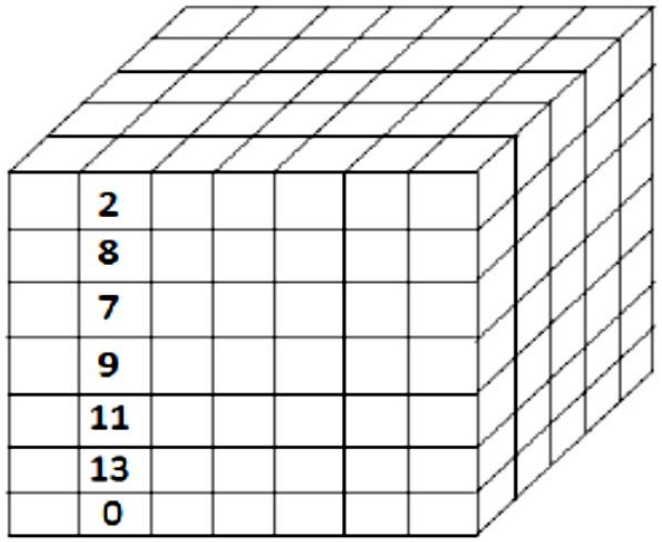
Inconvénients

- Difficulté de mise en œuvre, systèmes majoritairement propriétaires
- Volume de données limité, Problème d'éparsité des cubes
- Redondance des données avec l'entrepôt source

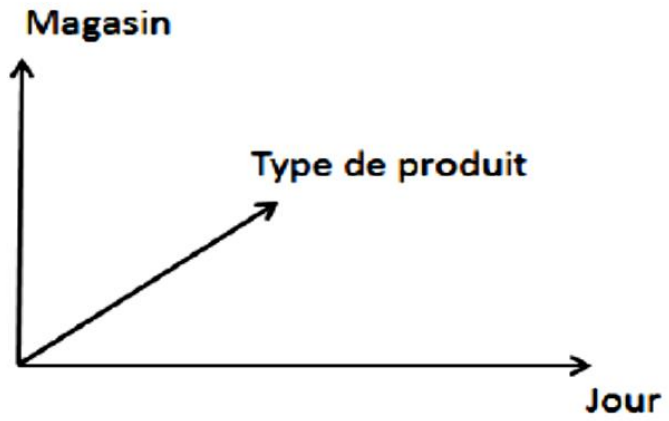


Exemple de cube MOLAP

Chaque fait (cellule) 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

ROLAP VERSUS MOLAP

	ROLAP	MOLAP
Signifier	Relational Online Analytical Processing	Multidimensional Online Analytical Processing
Forme de données	Les données sont stockées sous forme de tables relationnelles.	Les données sont stockées dans le grand tableau multidimensionnel composé de cubes de données.
Accès	Accès lent	Accès plus rapide.
Vue	ROLAP crée une vue multidimensionnelle des données de manière dynamique.	MOLAP stocke déjà la vue multidimensionnelle statique dans les MDDB.
Technologies	Utilise des requêtes SQL complexes pour extraire des données de l'entrepôt principal.	Le moteur MOLAP crée des cubes de données pré-calculés et pré-fabriqués pour les vues de données multidimensionnelles. La technologie de matrice fragmentée est utilisée pour gérer la dispersion des données.
Stockage & Récupération	Les données sont stockées et extraites de l'entrepôt de données principal.	Les données sont stockées et extraites à partir des MDDBs de la base de données propriétaire.

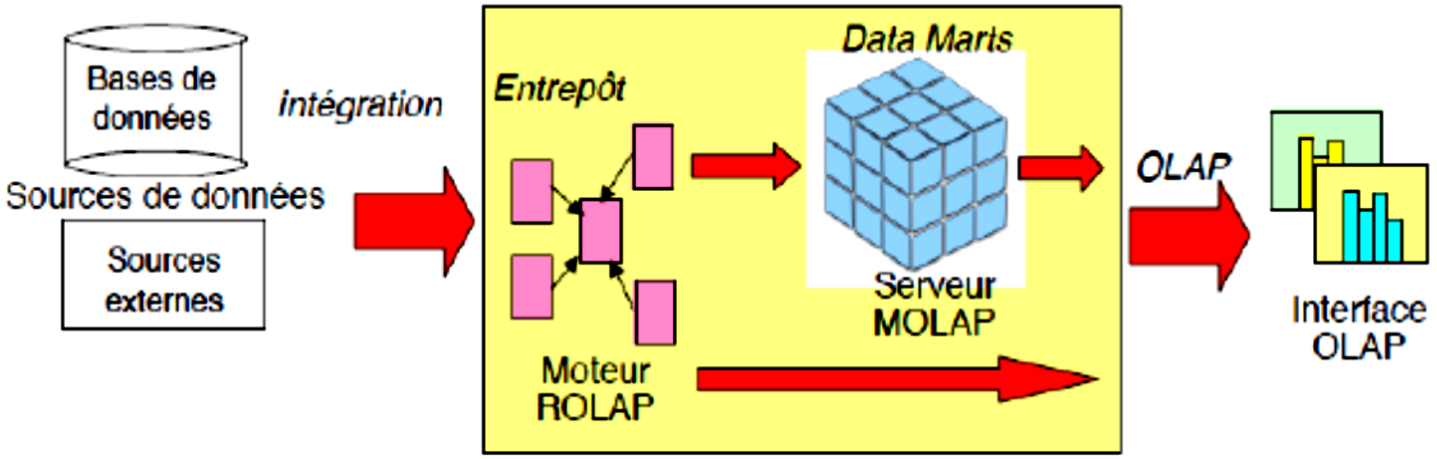
Lequel choisir entre **ROLAP** et **MOLAP** ? Il dépend des performances et de la complexité de requête. **MOLAP** est le choix d'un utilisateur s'il veut une réponse plus rapide.

Types de systèmes OLAP: Approche HOLAP

HOLAP : HYBRID OLAP

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 »)



Exercice

L'entreprise Renault désire construire un entrepôt de données Ventes pour suivre l'évolution de ses ventes de voitures par modèle, par concessionnaire et par année. Elle s'intéresse particulièrement au nombre de voitures ainsi qu'au prix total de voitures vendues selon les trois axes modèle, concessionnaire et année.

N°	*concessionnaire*	*modèle*	*année*	Nbre de Ventes
1	Croix Rousse	Clio	2000	1000
2	Croix Rousse	Clio	2001	1500
3	Croix Rousse	Twingo	2000	1000
4	Croix Rousse	Twingo	2001	5000
5	Croix Rousse	Espace	2002	1000
6	Gerland	Clio	2001	1200
7	Gerland	Twingo	2002	500
8	Mermoz	Twingo	2000	1500
9	Mermoz	Espace	2001	500
10	Bron	Clio	2001	1000
11	Bron	Clio	2002	500
12	Bron	Twingo	2001	700
13	Bron	Twingo	2002	1000
14	Bron	Espace	2002	500

Représentation relationnelle de l'entrepôt Ventes

Exercice

Proposer une représentation multidimensionnelle (MOLAP) pour observer l'évolution des ventes sur les trois axes. Le cube obtenu est-il éparse ? Argumenter.

	2000			2001			2002		
	Clio	Twingo	Espace	Clio	Twingo	Espace	Clio	Twingo	Espace
Croix Rousse	1 000	1 000		1 500	5 000				1 000
Gerland				1 200				500	
Mermoz		1 500				500			
Bron				1 000	700		500	1 000	500

Eparsité : **E**

E – 14 / 36 = 38,89% → Le cube est éparse.

Exercice

Proposer une représentation multidimensionnelle pour observer l'évolution des ventes sur les deux axes concessionnaire et modèle uniquement (MOLAP).

	Clio	*Twingo*	*Espace*
Croix Rousse	2 500	6 000	1 000
Gerland	1 200	500	
Mermoz		1 500	500
Bron	1 500	1 700	500

Exercice

Calculer l'opérateur CUBE avec l'approche ROLAP (les données résultats sont stockées dans la même table relationnelle que les données sources).

N°	concessionnaire	modèle	année	Nbre de Ventes
1	Croix Rousse	Clio	2000	1 000
2	Croix Rousse	Clio	2001	1 500
3	Croix Rousse	Twingo	2000	1 000
4	Croix Rousse	Twingo	2001	5 000
5	Croix Rousse	Espace	2002	1 000
6	Gerland	Clio	2001	1 200
7	Gerland	Twingo	2002	500
8	Mermoz	Twingo	2000	1 500
9	Mermoz	Espace	2001	500
10	Bron	Clio	2001	1 000
11	Bron	Clio	2002	500
12	Bron	Twingo	2001	700
13	Bron	Twingo	2002	1 000
14	Bron	Espace	2002	500
15	*ALL*	Clio	2000	1 000
16	*ALL*	Clio	2001	3 700
17	*ALL*	Clio	2002	500
18	*ALL*	Twingo	2000	2 500
19	*ALL*	Twingo	2001	5 700
20	*ALL*	Twingo	2002	1 500
21	*ALL*	Espace	2001	500

Le langage MDX (Multi Dimensional eXpression)

- Langage de requêtes sur les bases de données OLAP, il comprend des instructions de manipulation de données et des instructions de définition de données.
- Développé par Microsoft en 1997.
- Adopté par les plus importants éditeurs de solutions BI.
- Langage de requête analogue au rôle de SQL pour les bases de données relationnelles. C'est aussi un langage de calcul avec une syntaxe similaire à celle des tableurs.
- Navigation dans les bases multidimensionnelles.

MDX VERSUS SQL

- Mots clé : SELECT, FROM, WHERE mais leurs sémantiques sont différentes.
- SQL : construction de vues relationnelles.
- MDX : construction de vues multidimensionnelles des données.

Multidimensionnel (MDX)	Relationnel (SQL)
Cube	Table
Niveau (`Level`)	Colonne (chaine de caractère ou valeur numérique)
Dimension	plusieurs colonnes liées ou une table de dimension
Mesure (Measure)	Colonne (discrète ou numérique)
Membre de dimension (*Dimension member*)	Valeur dans une colonne et une ligne particulière de la table

Structure générale d'une requête

SQL	MDX
SELECT column1, column2, ..., columnn FROM table	SELECT axis1 ON COLUMNS, axis2 ON ROWS FROM cube
Clause From (source de données) : **Une ou plusieurs tables**	Clause From (source de données) : **Un cube**

Structure générale d'une requête

- Syntaxe générale d'une requête MDX :

```
SELECT [<specification d'un axe>
        [, <spécification d'un axe>...]]
FROM [<spécification d'un cube>]
 [WHERE [<spécification d'un filtre (slicer)>]]
```

- Parenthèses en MDX :
  - { } : Ensemble des éléments servant à la création d'une dimension du résultat de la requête
  - ( ) : Sélection de tuples dans la clause WHERE
  - [ ] : Représentation d'espaces, de caractères spéciaux et d'interprétation non numérique de chiffres.

NB: [ ] : optionnels, sauf si le nom contient des caractères espace, des chiffres, ou est un mot-clé MDX

La clause SELECT

Indication des résultats que l'on souhaite récupérer par la requête

- **en SQL :**

Une vue des données en 2 dimensions : lignes et colonnes.

Les lignes ont la même structure définie par les colonnes.

- **en MDX :**

Nombre quelconque de dimensions pour former les résultats de la requête

On parle axe pour éviter confusion avec les dimensions du cube

Pas de signification particulière pour les lignes et les colonnes mais il faut définir chaque axe : axe1 définit l'axe horizontal et axe2 définit l'axe vertical

```
SELECT {Paris , Berlin} ON ROWS
       {[Q1] , [Q2].CHILDREN} ON COLUMNS
FROM CubeSales
WHERE (MEASURES.SalesAmount ,
       Time.[2014] ,
       Product.Product)
```

Exemple - Select

Soit la requête MDX suivante : *(Q=Quarter)*

```
SELECT  {Paris, Berlin} ON ROWS
        {[Q1], [Q2].CHILDREN} ON COLUMNS
FROM    CubeSales
WHERE   (MEASURES.SalesAmount,
         Time.[2014],
         Product.Product)
```

Agrégation de la mesure « SaleAmount » avec la fonction SUM function

Sélection de la dimension Time (2014 seulement)

Sélection de la dimension Product (all product)

Résultat :

	Q1 2014	April 2014	May 2014	June 2014
Paris	12,567	3,360	5,450	4,570
Berlin	12,567	3,360	5,450	4,570
…	…	…	SalesAmount.values	…

Fin

Dr. H. EL BOUHISSI BRAHAMI