BI = Business Intelligence
Master Data-Science
Cours 4 - OLAP

Ludovic DENOYER - ludovic.denoyer@lip6.fr

UPMC

15 février 2015

Plan

- Vision générale
- ETL
- Datawarehouse
- OLAP
- Reporting
- Data Mining

Entrepôt de données

Les entrepôts de données (data warehouse) sont :

- Orientés sujet
  - Les données sont organisées par sujet ou *faits* (ex : clients, produits, ventes, etc.).
  - Les données sont organisées selon des dimensions
- Intégrés
  - Les données, qui proviennent de diverses sources hétérogènes, sont consolidées et intégrées dans l'entrepôt.
- Historiques
  - Les données ont très souvent une composante temporelle (ex : date et heure d'une transaction).
- Non-volatiles
  - Une fois insérées dans l'entrepôt, les données ne sont jamais modifiées ou effacées ; elle sont conservées pour des analyses futures.

Modèle dimensionnel

Contrairement aux systèmes opérationnels, le stockage des données dans un DW se fait habituellement sous la forme d'un schéma dimensionnel. Un tel schéma nécessite de définir :

- **des dimensions**
- **des faits**

Dimensions

Les dimensions sont les axes sur lesquels on souhaite baser l'analyse des données : la date, la région géographique, le type de produit, etc...

Modèle dimensionnel

Faits

Les faits sont les données que l'on souhaite analyser : On aura des tables de faits pour les ventes (chiffre d'affaire net, quantités et montants commandés, quantités facturées, quantités retournées, volumes des ventes, etc.) par exemple ou sur les stocks (nombre d'exemplaires d'un produit en stock, niveau de remplissage du stock, taux de roulement d'une zone, etc.), ou peut être sur les ressources humaines (performances des employés, nombre de demandes de congés, nombre de démissions, taux de roulement des employés, etc.).

Stockage de données "dimensionnelles" dans une BD relationnelle

Différents schéma de stockage :

- Schéma en étoile
- Schéma en flocon
- Schéma en constellation

Stockage de données "dimensionnelles" dans une BD relationnelle

Table des faits

La table des faits contient une clef, ainsi que des champs dimension (*foreing keys*) et des champs de mesure. Les champs dimension permettent de relier un fait à ses dimensions, les champs de mesure sont des mesures sur le fait : nombre de vente, etc...

Table des dimensions

Chaque dimension est associée à une table (ou plusieurs dans le cas de dimensions hiérarchiques). Les dimensions contiennent à la fois une clefs, ainsi que des champs descriptifs des dimensions.

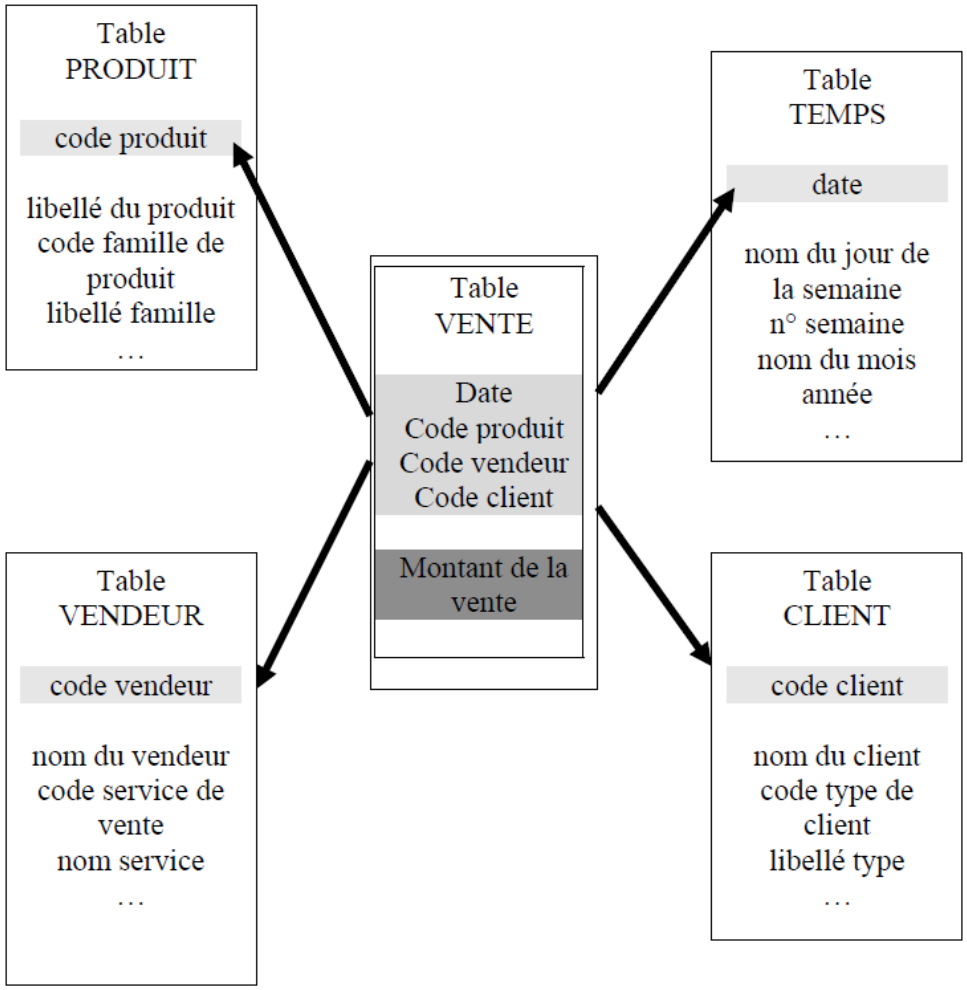
Modèle dimensionnel : étude de cas

Petit énoncé :

Considérons le cas "classique" d'une société de vente de produit. Elle possède des sources de données sur les produits, les ventes et les clients. On propose de concevoir une ED qui permette de fournir le chiffre d'affaires des ventes d'un produit, par date, client, et vendeur, ainsi que toutes les sommations possibles de chiffre d'affaires.

Dessinez le schéma de BD correspondant

Etude de cas



Source : J.-F. Desnos

Définition

OLAP

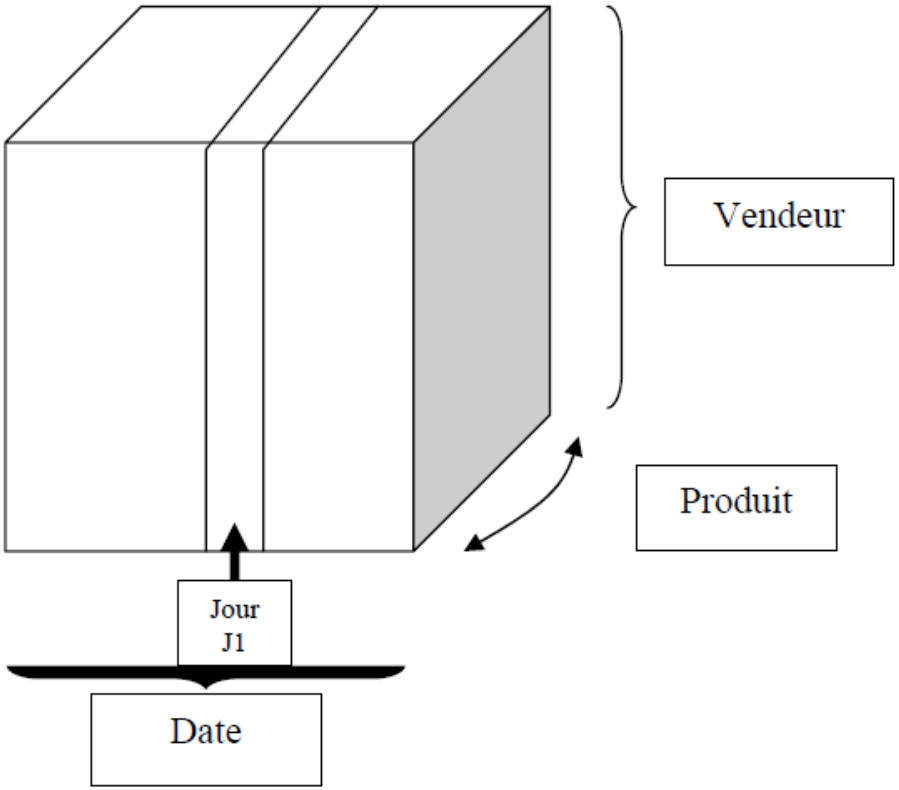
En informatique, et plus particulièrement dans le domaine des bases de données, le traitement analytique en ligne (anglais online analytical processing, OLAP) est un type d'application informatique orienté vers l'analyse sur-le-champ d'informations selon plusieurs axes, dans le but d'obtenir des rapports de synthèse tels que ceux utilisés en analyse financière. Les applications de type OLAP sont couramment utilisées en informatique décisionnelle, dans le but d'aider la direction à avoir une vue transversale de l'activité d'une entreprise.

Source : wikpiedia

OLAP s'oppose au traitement de transactions en ligne *(online transaction processing abr. OLTP)* qui s'inscrit dans un système opérationnel (en production).

Idée générale

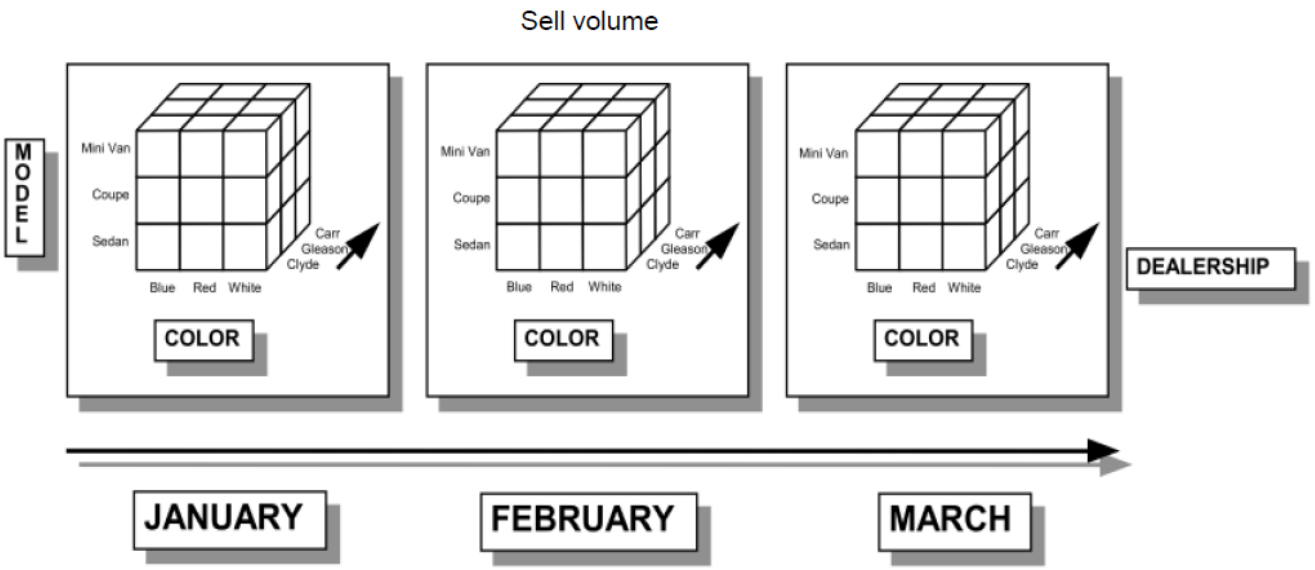
A partir du schéma en étoile précédent, il est possible de construire des tableaux multi-dimensionnels (ou hypercubes) permettant l'analyse des données du DW.



OLAP

OLAP fournit donc un cadre générique permettant l'analyse des données sur plusieurs dimensions. OLAP définit principalement **des opérations génériques** sur les hypercubes permettant l'analyse des données (y compris par des non experts). OLAP repose sur **des technologies** permettant le calcul et la mise à jour des hypercubes. OLAP définit aussi **un lanage de requête** permettant l'interrogation d'un hypercube (langage MDX)

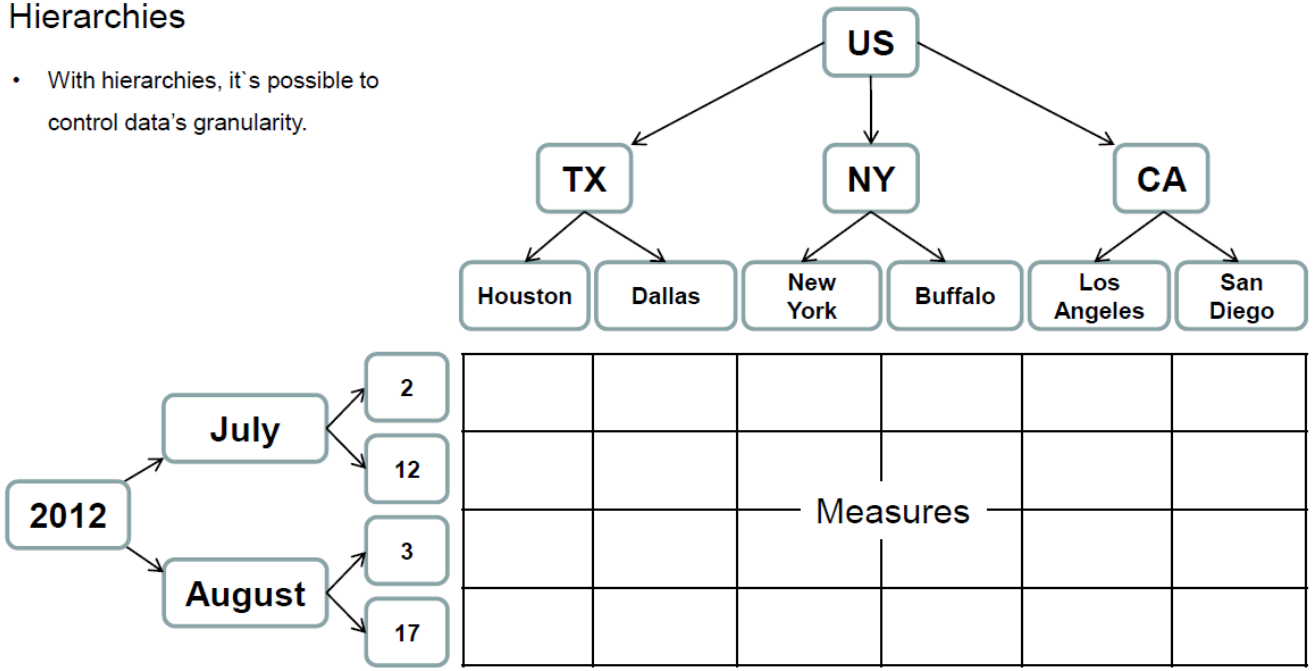
Opérations


Sell volume
MODEL
Mini Van
Coupe
Sedan
Carr
Gleason
Clyde
Blue Red White
COLOR
DEALERSHIP
JANUARY
FEBRUARY
MARCH

Opérations

Hierarchies

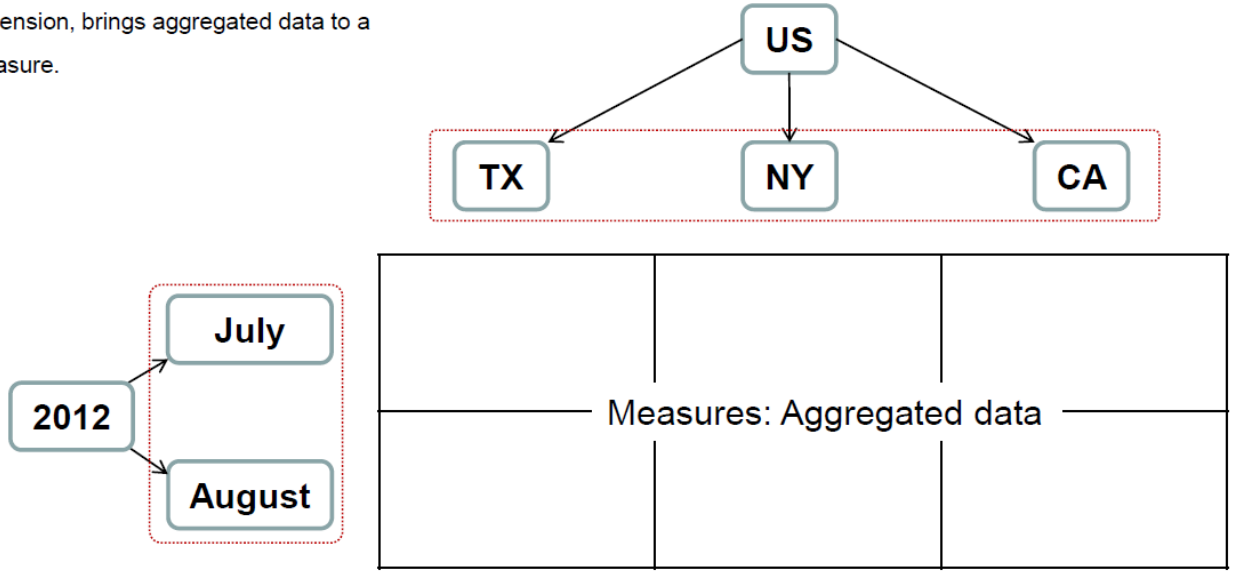
- With hierarchies, it`s possible to control data's granularity.



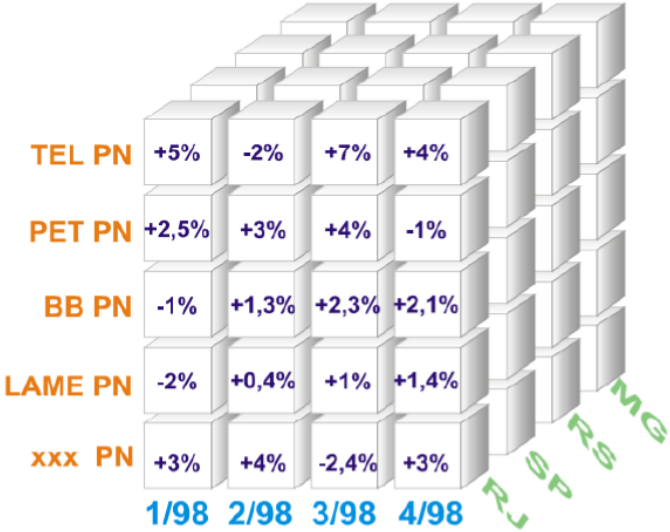
Opérations - Aggregation

Data aggregation

- A query in a higher level of a dimension, brings aggregated data to a measure.



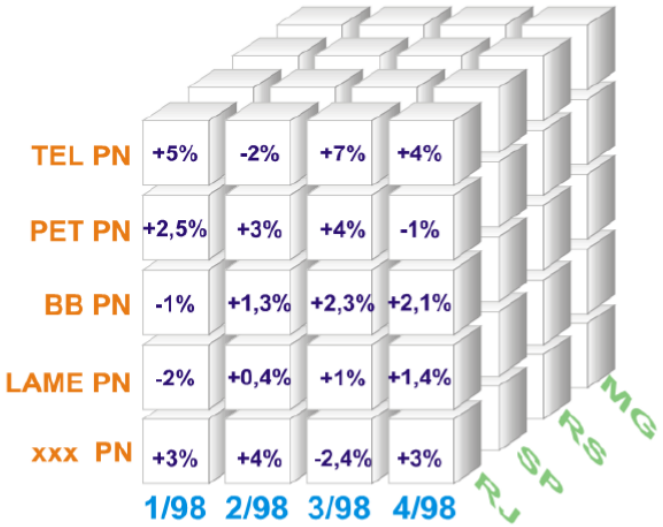
Opérations - Slice



Fixed dimension: State
Value: RJ

Stock	1/98	2/98	3/98	4/98
TEL PN	+5%	-2%	+7%	+4%
PET PN	+2,5%	+3%	+4%	-1%
BB PN	-1%	+1,3%	+2,3%	+2,1%
LAME PN	-2%	+0,4%	+1%	+1,4%

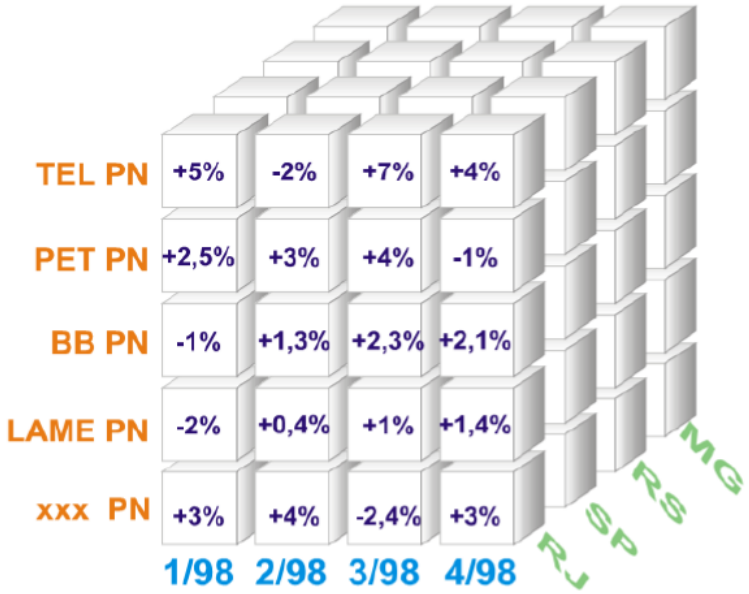
Opérations - Rotation



Fixed dimensions: State, Date, Stock
Values: RJ, 2/98, BB PN

Data	TEL PN	PET PN	BB PN	LAME PN
1/98	5%	2,50%	-1%	-2%
2/98	-2%	3%	1,30%	0,40%
3/98	7%	4%	2,30%	1%
4/98	4%	-1%	2,10%	1,40%

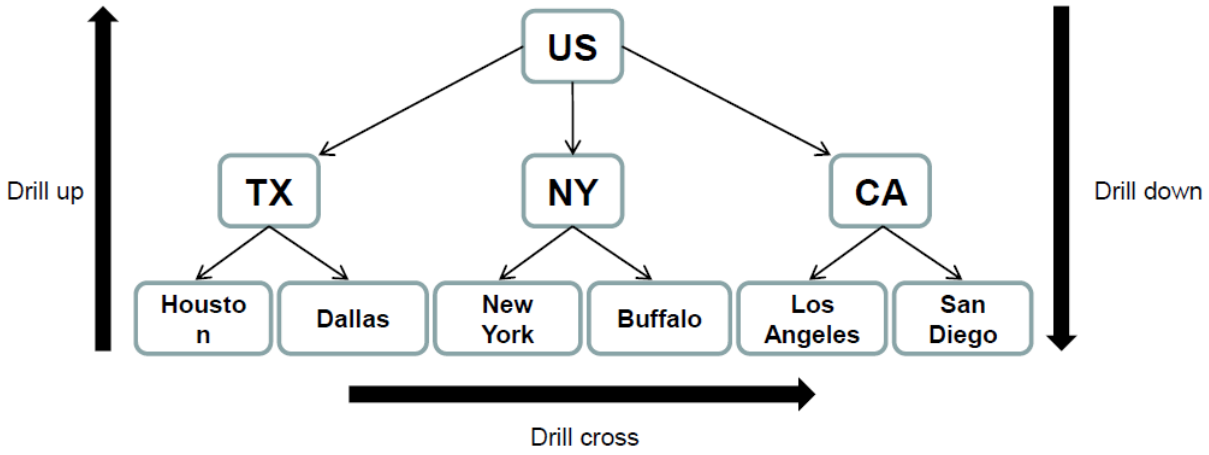
Opérations - Dice



Fixed dimensions: State, Date, Stock
Values: RJ, 2/98, BB PN

Ações	2/98
BB PN	+1,3 %

Opérations - Drill(s)



Différentes Technologies

Plusieurs techinlogies/philosophies sont liées à l'implémentation d'un système OLAP :

- **ROLAP - Relational OLAP :** OLAP sur du relationnel
- **MOLAP - Multidimensional OLAP :** OLAP sur un DW dimensionnel
- **HOLAP - Hybrid OLAP :** Mélange des deux (relationnel si nécessaire)

Mais aussi : Spatial OLAP (SOLAP), Desktop OLAP (DOLAP), ...

Technologies

Nom	Editeur	Technologie
DB2 UDB Server	*IBM*	ROLAP
Oracle9i	*Oracle*	ROLAP
SQL Server 2000	*Microsoft*	ROLAP
DSS	*Microstrategy*	ROLAP
TeraData	*Teradata Corporation*	ROLAP massivement parallèle
Informix Metacube	*Informix*	MOLAP
Essbase	*Arbor Software/Hyperion*	MOLAP
SAS OLAP Server	*SAS*	MOLAP
Metacube	*Informix*	ROLAP
SQL Server	*Microsoft*	HOLAP
MDDB	*SAS Institute*	MOLAP/ROLAP
Oracle Express-server	*Oracle*	MOLAP/ROLAP
DB2 OLAP Server	*IBM*	MOLAP/ROLAP
Crystal	*Seagate Software*	Serveur d'application OLAP unique pour tous les déploiements
PowerPlay	*Cognos*	idem

Source : B. Espinasse

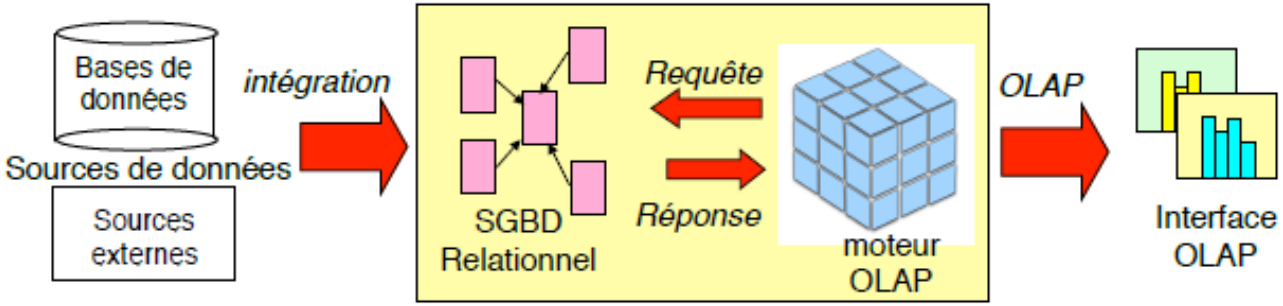
ROLAP

- ROLAP est la technolgoies la plus utilisée en OLAP car les SGBD relationnels sont très largement répandus
- Cependant les SGBDs relationnels "tout seuls" ne sont pas adaptés à des anlayses OLAP $\Rightarrow$ nécessité d'étendre les fonctionnalités d'un SGDB

Un moteur ROLAP :

- Permet de faire les calculs adaptés aux requêtes OLAP sur le SGBD relationnel
- Il permet aussi de générer des requêtes adaptée au schéma de l'entrepôt

ROLAP



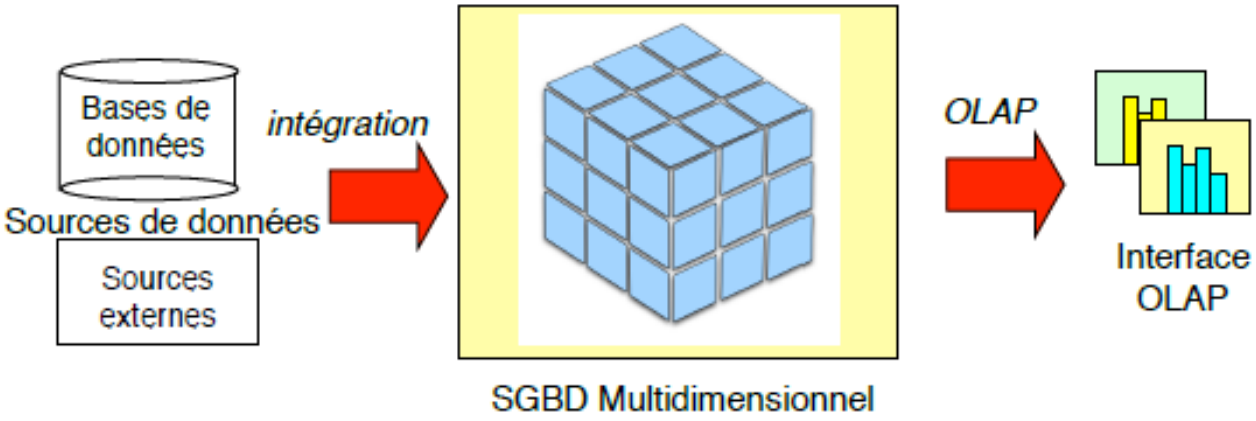
Source : B. Espinasse

MOLAP

La philosophie MOLAP consiste au stockage des données directement dans une structure de cube multidimensionnel :

- MOLAP nécessite le pré-calcul et le stockage des informations du cube,
- mais il permet des extractions très rapides et optimisées.

MOLAP



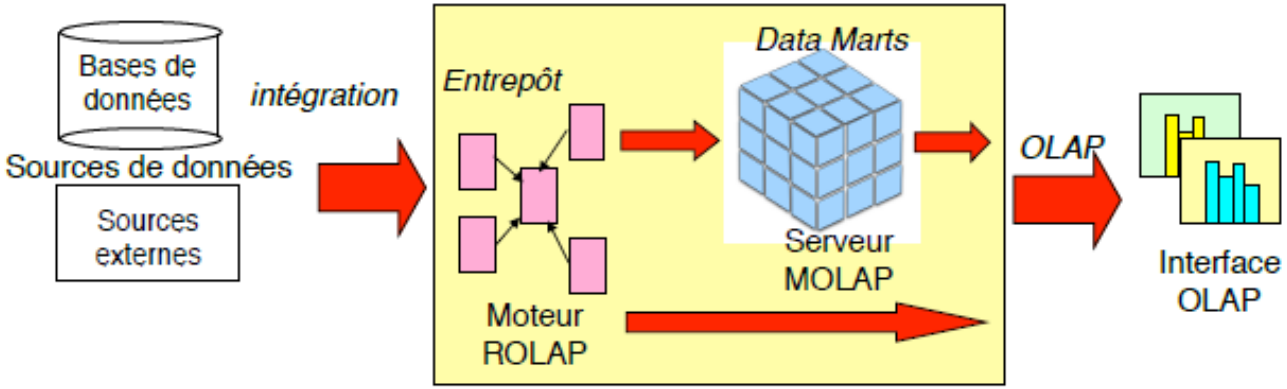
Source : B. Espinasse

HOLAP

Les systèms HOLAP tente d'exploiter le meilleur des deux mondes :

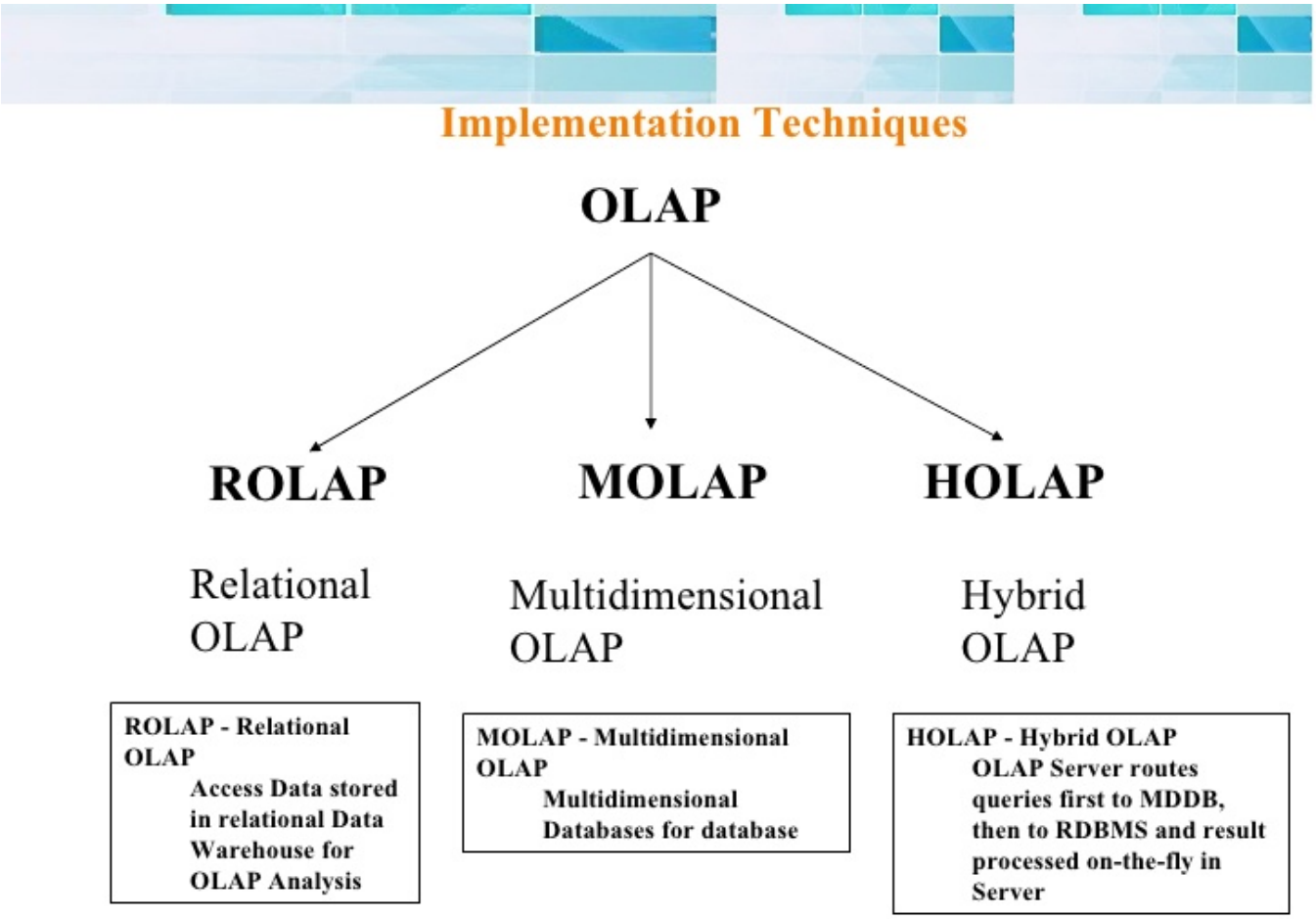
- La structure du moteur SGBD pour le stockage des données détaillées
- Un système de type MOLAP comme structure de données pour un certain nombre de requêtes (données agrégées)

HOLAP



Source : B. Espinasse

HOLAP


Implementation Techniques
OLAP
ROLAP
MOLAP
HOLAP
Relational OLAP
Multidimensional OLAP
Hybrid OLAP
ROLAP - Relational OLAP
Access Data stored in relational Data Warehouse for OLAP Analysis
MOLAP - Multidimensional OLAP
Multidimensional Databases for database
HOLAP - Hybrid OLAP
OLAP Server routes queries first to MDDB, then to RDBMS and result processed on-the-fly in Server

HOLAP

	Data storage	Aggregations storage	Query performance	Latency
MOLAP	Cube	Cube	High	High
HOLAP	Relational database	Cube	Medium	Low (none)
ROLAP	Relational database	Relational database	Low	Low (none)

Mondrian

Nous allons nous intéresser concrètement au système ROLAP Mondrian :

- Serveur OLAP écrit en JAVA
- Supporte le langage MDX
- Supporte l'interface *olap4j* permettant l'utilisation de plusieurs technologies OLAP en JAVA

Attention : ceci est un exemple illustratif....

Mondrian

Mondrian est utilisé pour :

- L'analyse interactive haute performance de grands ou de petits volumes d'informations.
- L'exploration de données multi-dimensionnelles, par exemple l'analyse des ventes par ligne de produits, par région, par période de temps
- Le calcul avancé en utilisant les expressions de calcul du langage MDX
- La transformation de requêtes MDX en Structured Query Language (SQL) pour récupérer des réponses aux requêtes dimensionnelles
- La formulation de requêtes à grande vitesse grâce à l'utilisation des tableaux aggrégés stockés dans le SGBDR

Mondrian

Le schéma OLAP est écrit en XML. Ce XML permet la description de :

- La base physique (relationnelle) sur laquelle seront basés les calculs
- La définition sur cette base des **table de faits** et **tables de dimensions**
- La définition sur cette base des **mesures**
- La définition de tables aggrégées

Mondrian

Le schema physique donne des informations sur la structuration de la BD :

```
<PhysicalSchema>
      <Table name='employee'>
          <Key>
              <Column name='employee_id'/>
          </Key>
      </Table>
      <Table name='store'>
          <Key>
              <Column name='store_id'/>
          </Key>
      </Table>
      <Link source='store' target='employee'>
          <ForeignKey>
              <Column name='store_id'/>
          </ForeignKey>
      </Link>
...
</PhysicalSchema>
```

Mondrian

Les dimensions sont renseignées :

```
<Dimension name='Promotion' table='promotion' key='Promotion Id'>
              <Attributes>
                  <Attribute name='Promotion Id' keyColumn='promotion_id' hasHierarchy='
                  <Attribute name='Promotion Name' keyColumn='promotion_name' hasHierarc
                  <Attribute name='Media Type' keyColumn='media_type' hierarchyAllMember
              </Attributes>
              <Hierarchies>
                  <Hierarchy name='Media Type' allMemberName='All Media'>
                      <Level attribute='Media Type'/>
                  </Hierarchy>
                  <Hierarchy name='Promotions' allMemberName='All Promotions'>
                      <Level attribute='Promotion Name'/>
                  </Hierarchy>
              </Hierarchies>
          </Dimension>
```

Mondrian

Les mesures sont renseignées :

```
<MeasureGroups>
          <MeasureGroup name='Sales' table='sales_fact_1997'>
              <Measures>
                  <Measure name='Unit Sales' column='unit_sales' aggregator='sum' format
                  <Measure name='Store Cost' column='store_cost' aggregator='sum' format
                  <Measure name='Store Sales' column='store_sales' aggregator='sum' form
                  <Measure name='Sales Count' column='product_id' aggregator='count' for
                  <Measure name='Customer Count' column='customer_id' aggregator='distin
                  <Measure name='Promotion Sales' column='promotion_sales' aggregator='s
              </Measures>
              <DimensionLinks>
                  <ForeignKeyLink dimension='Store' foreignKeyColumn='store_id'/>
                  <ForeignKeyLink dimension='Time' foreignKeyColumn='time_id'/>
                  <ForeignKeyLink dimension='Product' foreignKeyColumn='product_id'/>
                  <ForeignKeyLink dimension='Promotion' foreignKeyColumn='promotion_id'/
                  <ForeignKeyLink dimension='Customer' foreignKeyColumn='customer_id'/>
              </DimensionLinks>
          </MeasureGroup>
</MeasureGroups>
```

Mondrian

Les mesures peuvent correspondre à des tables aggrégées :

```
1  <PhysicalSchema>
2  ...
3  <Table name='agg_c_special_sales_fact_1997'/>
4  <Table name='agg_pl_01_sales_fact_1997'/>
5  <Table name='agg_l_05_sales_fact_1997'/>
6  <Table name='agg_g_ms_pcat_sales_fact_1997'/>
7  <Table name='agg_c_14_sales_fact_1997'/>
8  </PhysicalSchema>
```

Mondrian

Plusieurs outils permettant de réaliser ces hypercubes sont disponibles :

- Eclipse
- Pentaho Cube Designer
- Mondrian Workbench

MDX

Définition

Le MDX (de l'anglais Multidimensional Expressions, « expressions multidimensionnelles ») est un langage de requête pour les bases de données OLAP, 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.
Le langage des expressions multidimensionnelles possède une syntaxe appropriée à l'interrogation et manipulation des données multidimensionnelles mémorisées dans un cube OLAP1. Bien qu'il soit possible de traduire certaines expressions dans le langage SQL traditionnel, cela nécessite une syntaxe SQL souvent maladroite même pour des expressions MDX très simples. MDX a été adopté par une large majorité de fournisseur de la technologie OLAP et est devenu un standard de facto pour les systèmes OLAP.
Source : wikipedia

MDX - Exemple

```
SELECT
   { [Measures].[Store Sales] } ON COLUMNS,
   { [Date].[2002], [Date].[2003] } ON ROWS
FROM Sales
WHERE ( [Store].[USA].[CA] )
```

MDX - Exemple

```
SELECT
   { [Measures].[Store Sales] } ON COLUMNS,
   { [Date].[2002], [Date].[2003] } ON ROWS
FROM Sales
WHERE ( [Store].[USA].[CA] )
```

- **Sales** est le cube sur lequel la requête est faite

MDX - Exemple

```
SELECT
   { [Measures].[Store Sales] } ON COLUMNS,
   { [Date].[2002], [Date].[2003] } ON ROWS
FROM Sales
WHERE ( [Store].[USA].[CA] )
```

- **Sales** est le cube sur lequel la requête est faite
- **[Measures].[Store Sales]** et **[Date].[2002], [Date].[2003]** sont les dimensions conservées
- **[Store].[USA].[CA]** est le "slicer"