

Business Intelligence

Présenté par:

Amal HALFAOUI (Epse GHERNAOUT)

Amal.halfaoui@univ-tlemcen.dz

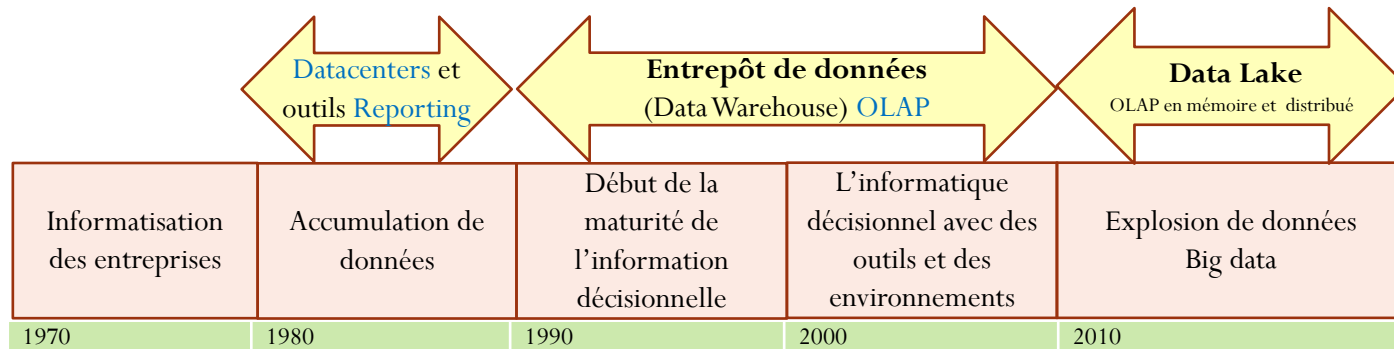
amal.halfaoui@gmail.com

Plan

- Introduction: le processus décisionnel
- Modélisation des Entrepôts de données (DW)
- Approches d'implémentation des serveurs OLAP
- Alimentation d'un DW
- Manipulation des données dimensionnelles
- BI et le Big Data



Naissance de l'informatique décisionnelle (BI)



Fondateurs



Edgar Frank Codd

Ecrit les douze lois du traitement analytique en ligne (1993)



Bill Inmon

Formalise du concept d'entrepôt de données (1994)



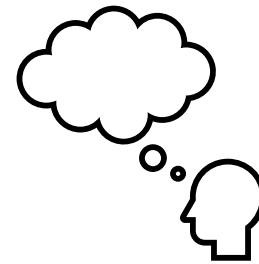
Ralph Kimball

Des premiers travaux sur l'informatique décisionnelle

Problématique

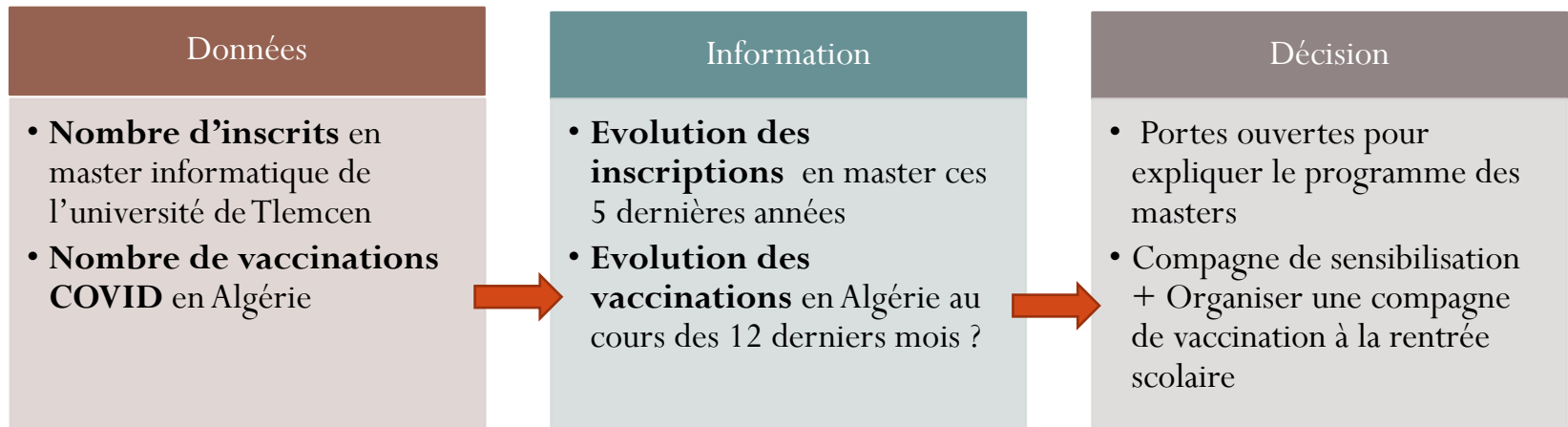
Nous sortons d'une réunion pédagogique (départements + doyen)

- Pourquoi le nombre d'étudiants dans l'enseignement en ligne a baissé?
- Comment a il baissé?
- Dans quelle promotion ?
- À quelle heure du direct du cours ?
- Dans quel type d'unité d'enseignement: cours, TD, TP?
- N'avait-on pas une baisse semblable en décembre ces trois dernières années, chaque année ?



Problématique

Transformer les «**données**» des systèmes d'information opérationnels (base de données) en «**informations**» qui servent à la prise de **décision**



Contexte

- **Besoin**

- ↳ prise de décisions stratégiques et tactiques.

- **Pourquoi**

- ↳ besoin de réactivité.

- **Pour qui**

- ↳ les décideurs (non informaticiens)

- **Comment**

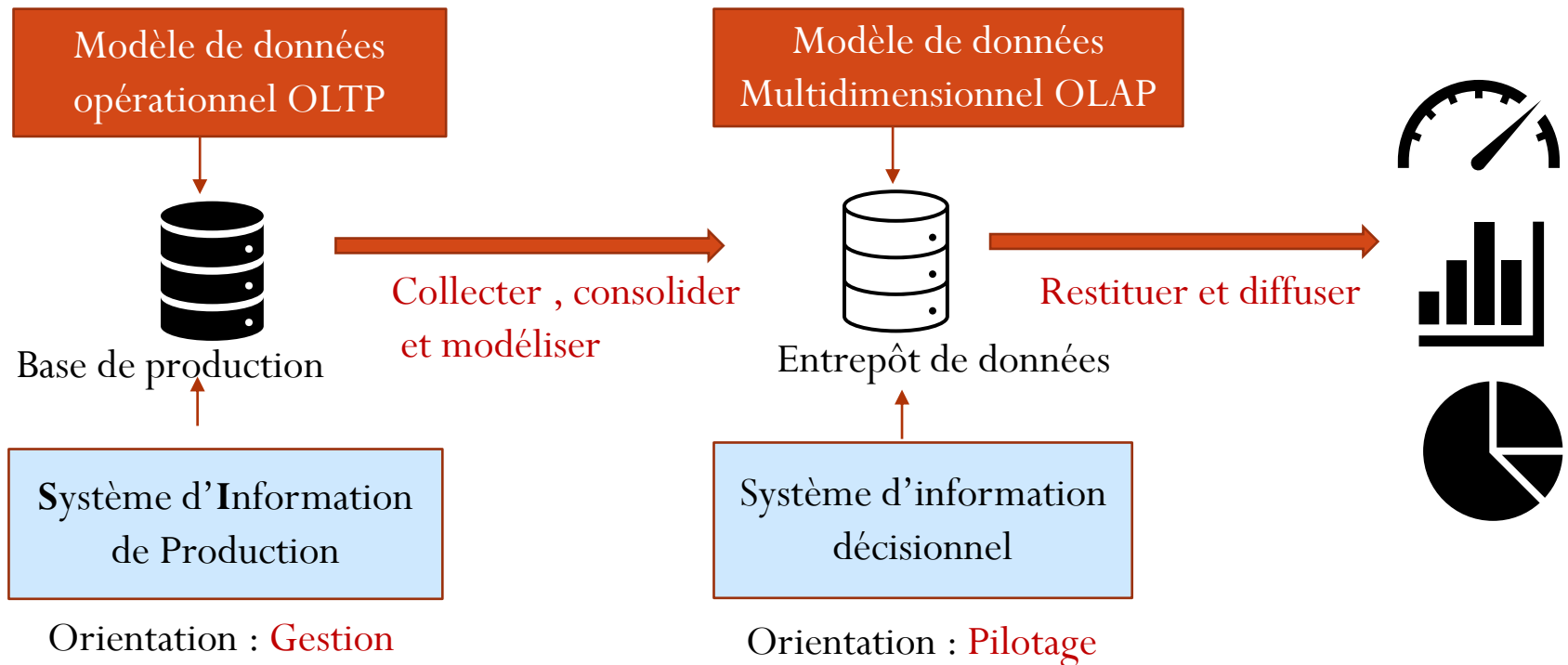
- ↳ répondre aux demandes d'analyse des données, dégager des informations qualitatives nouvelles.

Business Intelligence (BI)

ou bien « l'informatique décisionnelle », ou encore « Decision support system DSS », un ensemble de **solutions informatiques (outils)** permettant aux décideurs (DG, Direction stratégique,..) **l'analyse** et le **requêtage** des données de **l'entreprise**, afin d'en dégager les informations qualitatives nouvelles qui vont fonder des décisions, qu'elles soient tactiques ou stratégiques.

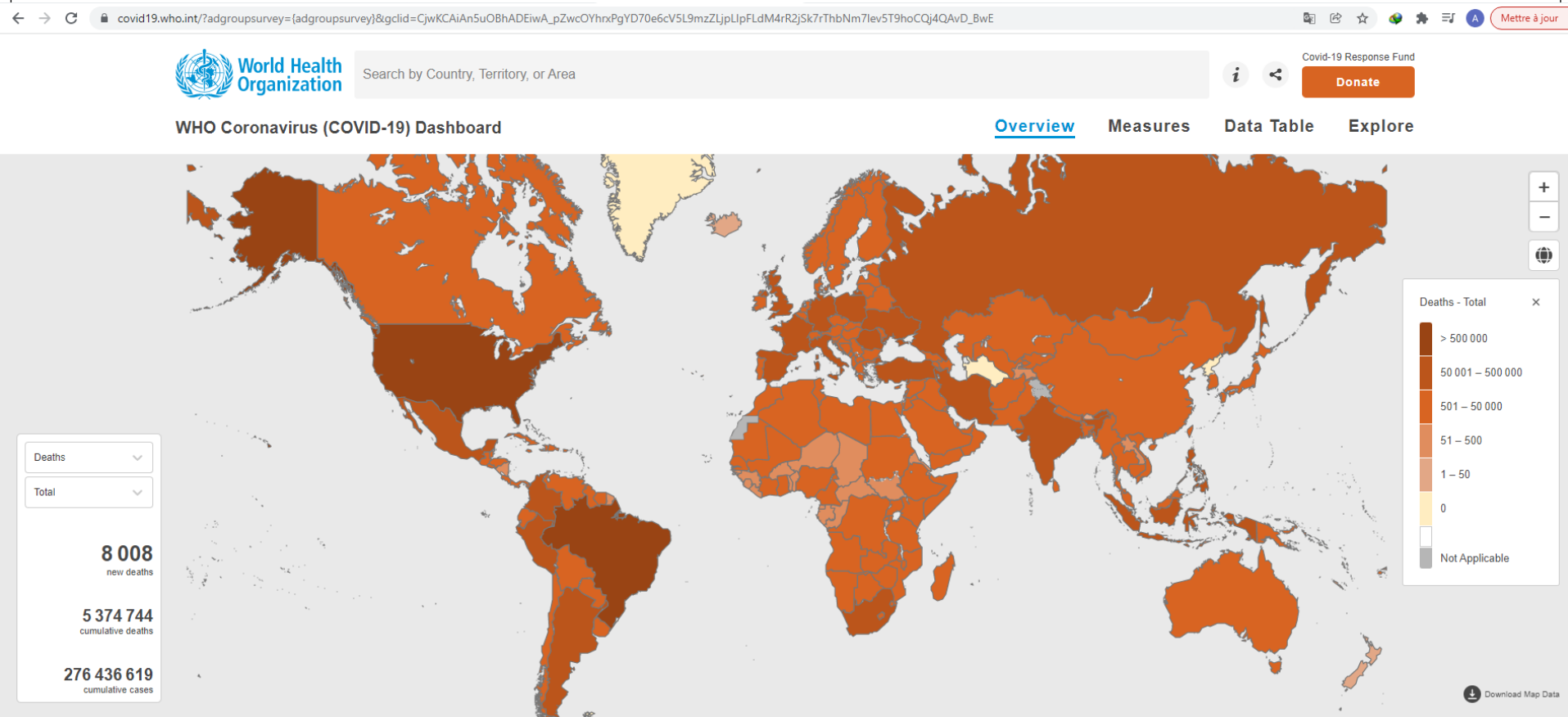
Business Intelligence (BI)

BI permet de **collecter** (les données du système opérationnel-transactionnel), **consolider** (fusionner , éliminer les redondances, les inexactitudes,..), **modéliser**, **restituer** et **diffuser** les données de l'entreprise sous forme de **tableaux de bord** équipés de fonction d'analyses multidimensionnelles



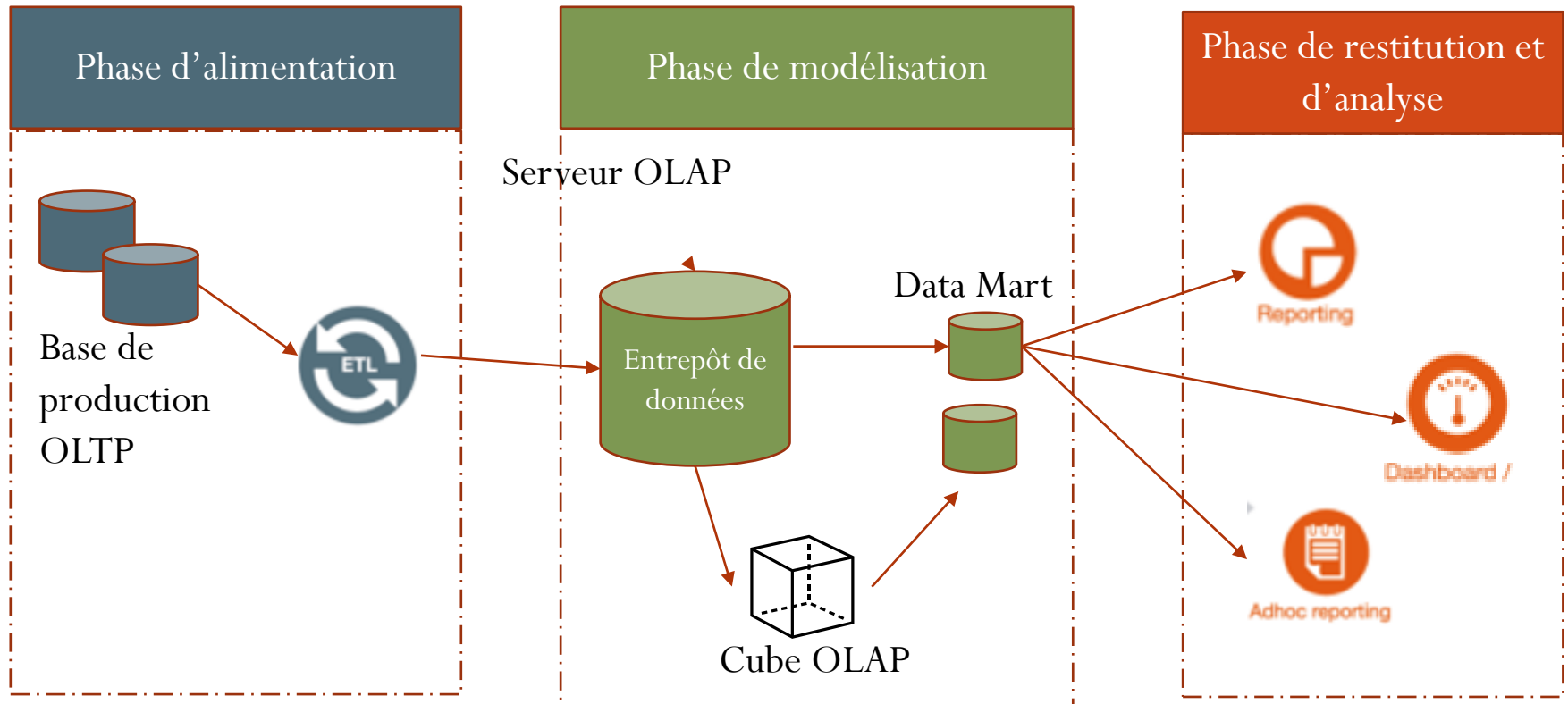
Business Intelligence (BI)

Exemple de tableau de bord



Globally, as of 7:06pm CET, 23 December 2021, there have been 276 436 619 confirmed cases of COVID-19, including 5 374 744 deaths, reported to WHO. As of 23 December 2021, a total of 8 649 057 088 vaccine doses have been administered

La chaîne du processus BI et ses composants



La chaine du Process BI et ses composants

- **Phase d'alimentation:** extraire, transformer et charger dans DW, l'ensemble des données brutes issues des différentes sources de stockage de l'information (bases de données, fichiers plats, applications métier, etc.) outils ETL (Extract, Transform, Load)
- **Phase de modélisation:** stocker et structurer les données (Modèle en étoile ou en flocon de neige: cubes OLAP), dans un espace unifié (le data warehouse) pour qu'elles soient disponibles pour un usage décisionnel. Cette phase peut également être réalisée grâce aux outils d'ETL via des connecteurs qui permettent l'écriture dans le data warehouse.
- **Phase de restitution et d'analyse:** de nouveaux calculs de données en utilisant les outils de reporting, de portails d'accès à des tableaux de bords, d'outils de navigation dans des cubes OLAP (ou hypercubes) ou encore des outils de statistique et de datamining

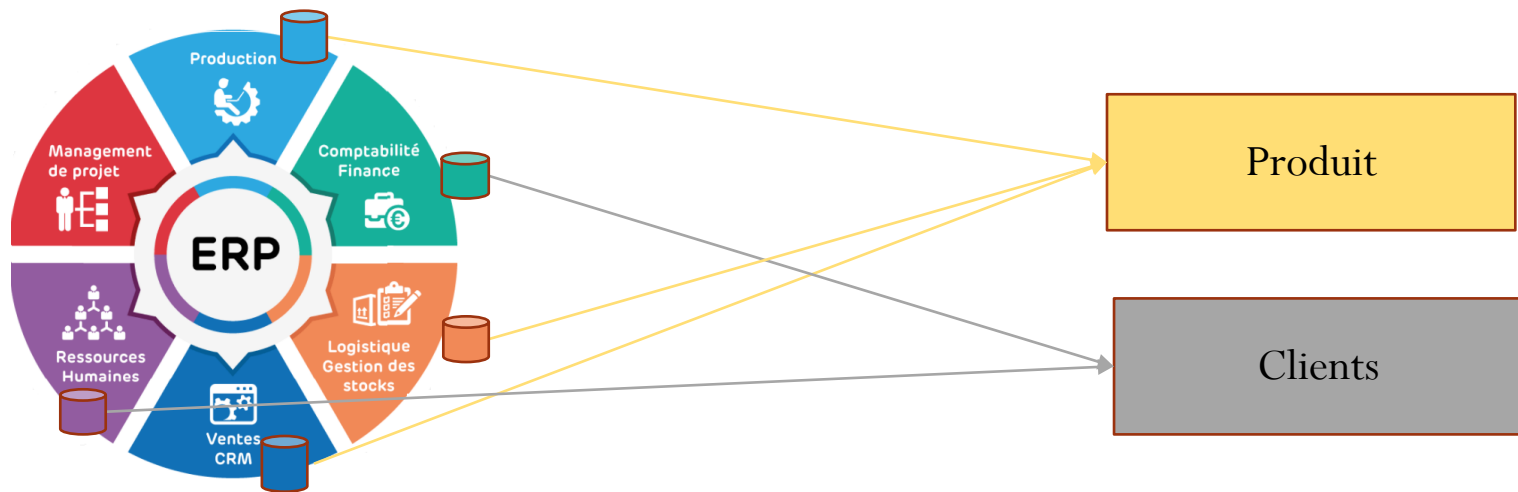
Entrepôt de données

- Data Warehouse (DWH) = BDD décisionnelle
- Collection de données **orientées sujet, intégrées, non volatiles** et **historisées**, organisées pour le support d'un processus d'aide à la décision. (W.H. Inmon 91)
- Base à des fin d'analyse (W.H. Inmon 96)

Caractéristiques des données d'un DW

- **Orientées sujet :**

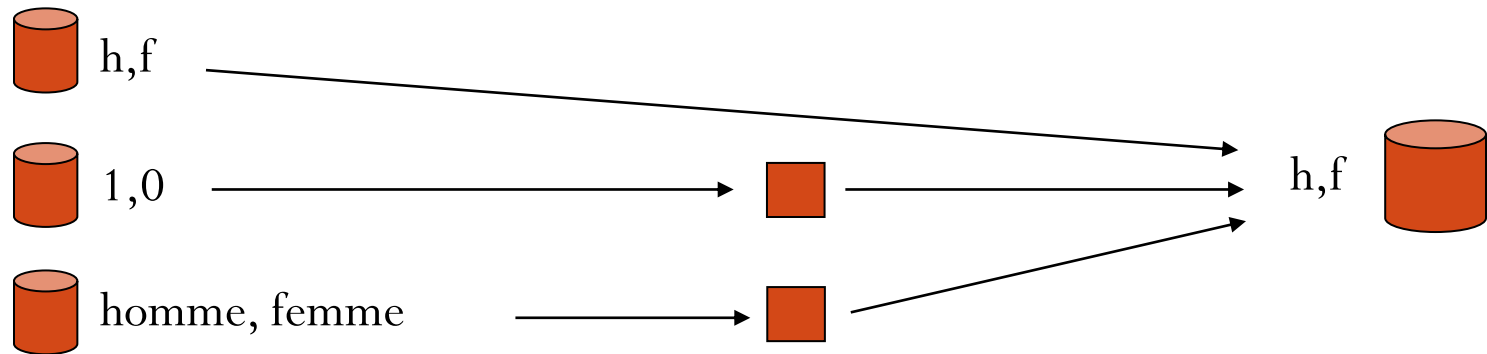
- Organisées autour de sujets majeurs de l'entreprise
- Données pour l'analyse et la modélisation en vue de l'aide à la décision, et non pas pour les opérations et transactions journalières
- Vue synthétique des données selon les sujets intéressant les décideurs



Caractéristiques des données d'un DW

- **Intégrées et consolidé :**

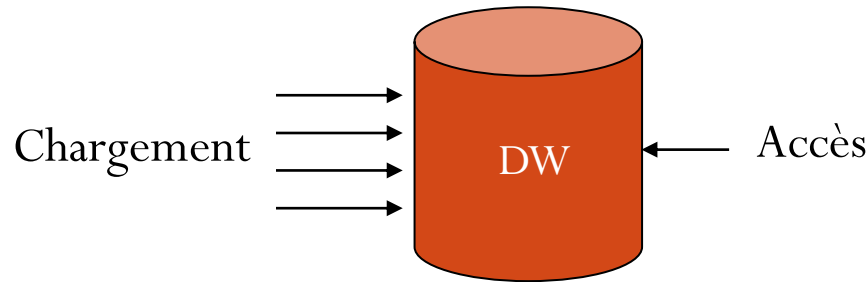
- Intégrer et unifier des sources de données multiples et hétérogènes (BD relationnelles, fichiers, enregistrements de transactions)
- Phase la plus complexe (60 à 90 % de la charge totale d'un projet DW)



Caractéristiques des données d'un DW

- **Non volatiles :**

- Pas de mises à jour des données dans le DW.



- Une même requête effectuée à intervalle de temps, en précisant la date référence de l'information donnera le même résultat.

Caractéristiques des données d'un DW

- **Historisées :**


- Stockage de l'**historique des données**, pas de mise à jour (Conséquence de non volatile)
- Un référentiel temps doit être associé aux données

Image de la base en Mai 2010



Base de
production

Produit		
Produit	Libellé	Packaging
Yaourt	Noix de coco	

Image de la base en Mai 2011

Produit		
Produit	Libellé	Packaging
Yaourt	gourmant noix de coco	

Entrepôt de
données

Temps			Produit			
Code	Année	Mois	N	Produit	Libellé	Packaging
1	2010	Mai	1	Yaourt	Noix de coco	
2	2011	Juillet	1	Yaourt	gourmant de coco	

OLAP (On-Line Analytical Processing)

- Représente l'ensemble des technologies qui, se basant sur une représentation **multi-dimensionnelle** des données, permet aux analystes et décideurs de traiter leurs données de façon **analytique**, **interactive** (sessions), rapide et permettant de voir les données de l'entreprise sous plusieurs angles (**dimensions**).
- OLAP permet de faire l'analyse multidimensionnelle sur l'entrepôt et transforme les données de l'entrepôt en informations stratégiques. Requêtes posés à l'entrepôt

SGBD VS DW (OLTP vs OLAP)

Caractéristiques	Base Opérationnelle	DW
Application	Production	aide à la Décision
Utilisateurs	un département (techniciens, vendeurs)	Transversale (entreprise) (analystes, gestionnaires)
Données	normalisées, non agrégées	dénormalisées, agrégées
Ancienneté de données	Récentes	historisées
Requêtes	simples, nombreuses, régulières, prévisibles, répétitives	complexes, peu nombreuses, irrégulières, non prévisibles
Opérations	Lecture et écriture	Lecture et rafraichissement
Nb tuples invoqués par requête (moyenne)	Dizaines	million

Plan

- Introduction: le processus décisionnel
- **Modélisation des Entrepôts de données (DW)**
- Approches d'implémentation des serveurs OLAP
- Alimentation des DW
- Manipulation des données dimensionnelles
- BI à l'ère des Big Data



Modélisation multidimensionnelle

Exemple : Ventes de produits par années

Tableur simple

Année	Produit	Ventes
2016	Antivols	29 800
2017	Antivols	35 000
2015	Antivols	10 000
2016	Boîtiers de pédalier	1 000
2017	Boîtiers de pédalier	600
2015	Boîtiers de pédalier	500
2016	Casques	17 000
2017	Casques	34 000
2015	Casques	8 300
2017	Casquettes	600
2016	Casquettes	400
2015	Casquettes	500

Tableur croisé (deux dimensions)

	2015	2016	2017
Antivols	10 000	29 800	35 000
Boîtiers de pédalier	500	1 000	600
Casques	8 300	17 000	34 000
Casquettes	500	400	600

Modélisation multidimensionnelle

Exemple : Combinaison de 3 dimensions

Ventes de produits par années d'Alger

	2015	2016	2017
Antivols	10 000	29 800	35 000
Boîtiers de pédalier	500	1 000	600
Casques	8 300	17 000	34 000
Casquettes	500	400	600

Ventes de produits par années de Tlemcen

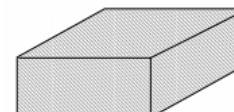
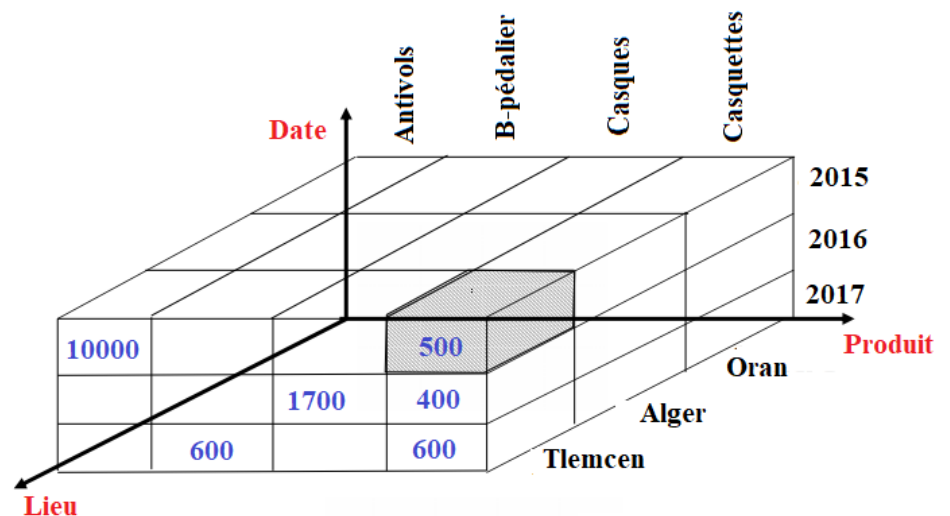
	2015	2016	2017
Antivols	13 300	15 600	27 000
Boîtiers de pédalier	2 300	3 300	5 000
Casques	6 700	3 800	7 500
Casquettes	6 00	500	650

Modélisation multidimensionnelle

Exemple 1 : Combinaison de 3 dimensions

		Oran		
		2015	2016	2017
Antivols	Boîtiers de pédalier	2015	2016	2017
Casques	Casquettes	2015	2016	2017
Antivols	Boîtiers de pédalier	10 000	29 800	35 000
Casques	Casquettes	500	1 000	600
Casques	Casquettes	8 300	17 000	34 000
Casquettes	Casquettes	500	400	600

La troisième dimension est le Lieu



Vente de **casquettes** en 2015 pour Tlemcen

Modélisation multidimensionnelle

Concepts de base

- Nouvelle méthode de conception autour des concepts métiers représente les données comme des faits et des dimensions
Les dimensions ne sont pas forcément normalisées.
 - Que voulez vous analyser ? (**Faits**)
 - Quels sont vos critères d'analyse ? (**Dimensions**)
 - Jusqu'à quel niveau de détail voulez vous aller ? (**Mesures** dans les faits et attributs des dimensions)
- Introduction de nouveaux types de table:
 - Table de faits
 - Table de dimensions
- Introduction de nouveaux modèles: (**ne pas normaliser au maximum**)
 - Modèle en étoile
 - Modèle en flocon

Modélisation multidimensionnelle

Table de Faits

- **Un fait:**

- Tout ce qu'on veut analyser. Il s'est passé quelque chose, et on l'a mesuré selon nos dimensions
- Ce que l'on souhaite mesurer (Mesure) exp : Quantités vendues, montant des ventes...
- correspond à une ligne, dans une table de faits

- **Une Table des faits** est la table principale du modèle dimensionnel

La table de fait contient les valeurs des mesures (les faits) sur le sujet (processus métier) et les clés vers les tables de dimensions (axes d'analyse)

- Associée à un seul processus métier à la fois: exp ventes **ou** l'inventaire **ou** les budgets

Modélisation multidimensionnelle

Table de Faits

- La table Fait peut contenir plusieurs Meseures. Elles donnent les valeurs numériques du fait
- Contient les clés étrangères des axes d'analyse (dimension) exp: Date (id temps), produit (Id Produit) , magasin (id Magasin),

Table de Faits : Ventes				
Id Temps	Id Magasin	Id Produit	Quantité	Prix Unitaire (€)
20111212	35	5	3	10
20111212	56	8	2	25
20111212	22	12	5	5
20111212	35	5	1	10
20111213	56	5	6	15
20111213	56	8	7	20
20111213	22	12	2	5

Modélisation multidimensionnelle

Table de dimension

➤ Les dimensions donnent le contexte du fait

- Axe d'analyse selon lequel vont être étudiées les données observables (faits)
- Contient le détail sur les faits
- Dimension = axe d'analyse
 - Client, produit, période de temps...
- Contient souvent un grand nombre de colonnes
 - L'ensemble des informations descriptives des faits
- Contient en général beaucoup moins d'enregistrements qu'une table de faits

Modélisation multidimensionnelle

Table de dimension : Dimension de TEMPS

- **Commune** à l'ensemble des DW
- Reliée à toute table de faits


Dimension Temps	
Clé de substitution	Clé temps (CP)
Attributs de la dimension	Jour
	Mois
	Trimestre
	Semestre
	Année
	Num_jour_dans_année
	Num_semaine_ds_année

Modélisation multidimensionnelle

Granularité d'une dimension (Hiérarchies)

Granularité d'une dimension : nombre de niveaux hiérarchiques

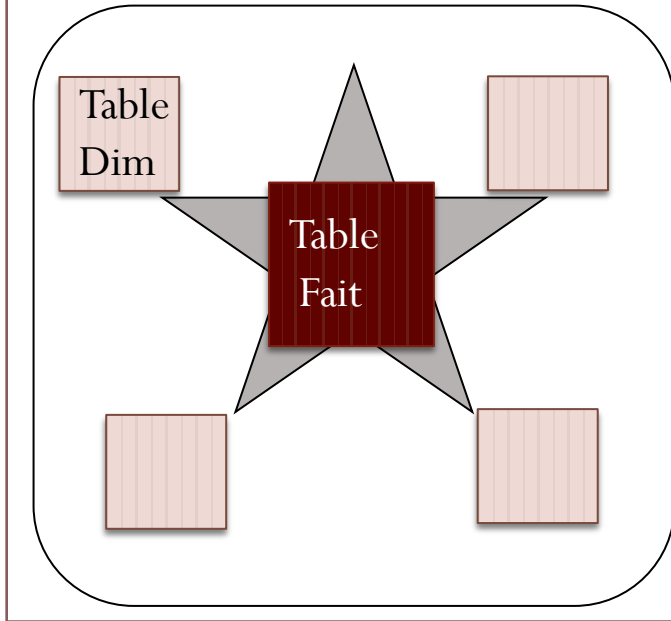
Temps : année — semestre — trimestre — mois - jours



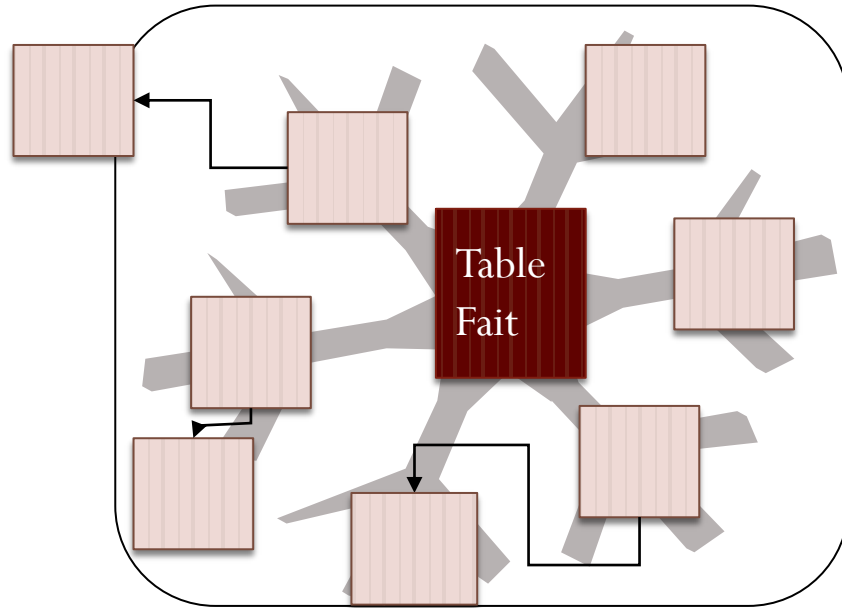
géographie : ville — région — pays



Modélisation multidimensionnelle



Modèle en étoile



Modèle en flocon

Modélisation multidimensionnelle

Modèle en Etoile (R. Kimball)

- Une table de fait centrale et des dimensions
- Les dimensions n'ont pas de liaison entre elles

Avantages:

- Facilité de navigation
- Nombre de jointures limité

Inconvénients:

- Redondance dans les dimensions
- Toutes les dimensions ne concernent pas les mesures

Modélisation multidimensionnelle

Modèle en Etoile (Exemple)

Exemple d'un schéma en étoile

Produit_Dimension

PID	catégorie	type
1	Romans	Livres
2	Enfants	Livres
3	Sciences	Livres
4	CD	Médias
5	DVD	Médias
6	BlueRay	Médias

Fact_Table

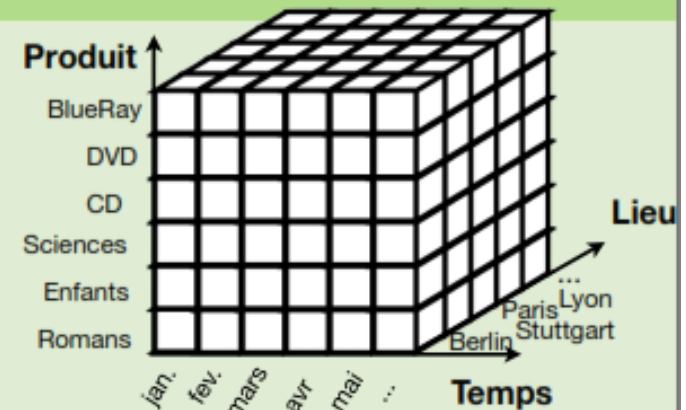
Temps_Dimension

TID	mois	trimestre	année
1	jan10	Q1 2010	2010
2	fev10	Q1 2010	2010
...

Lieu_Dimension

LID	ville
1	Berlin
2	Stuttgart
3	Paris
4	Lyon

PID	TID	LID	#ventes	CD
1	1	1	5	30
1	1	2	5	37
1	1	3	5	45
1	1	4	5	20
2	1	1	2	33
2	1	2	2	35
2	1	3	2	40
2	1	4	2	35
...
1	2	1	3	22
...



Mesures (ventes & chiffre d'affaire) pour des romans en janvier 2010

Toutes les autres combinaisons de catégories et de villes en janvier 2010.

Ici commencent les combinaisons pour février 2010 (suivies par celles des autres mois de l'année).

Modélisation multidimensionnelle

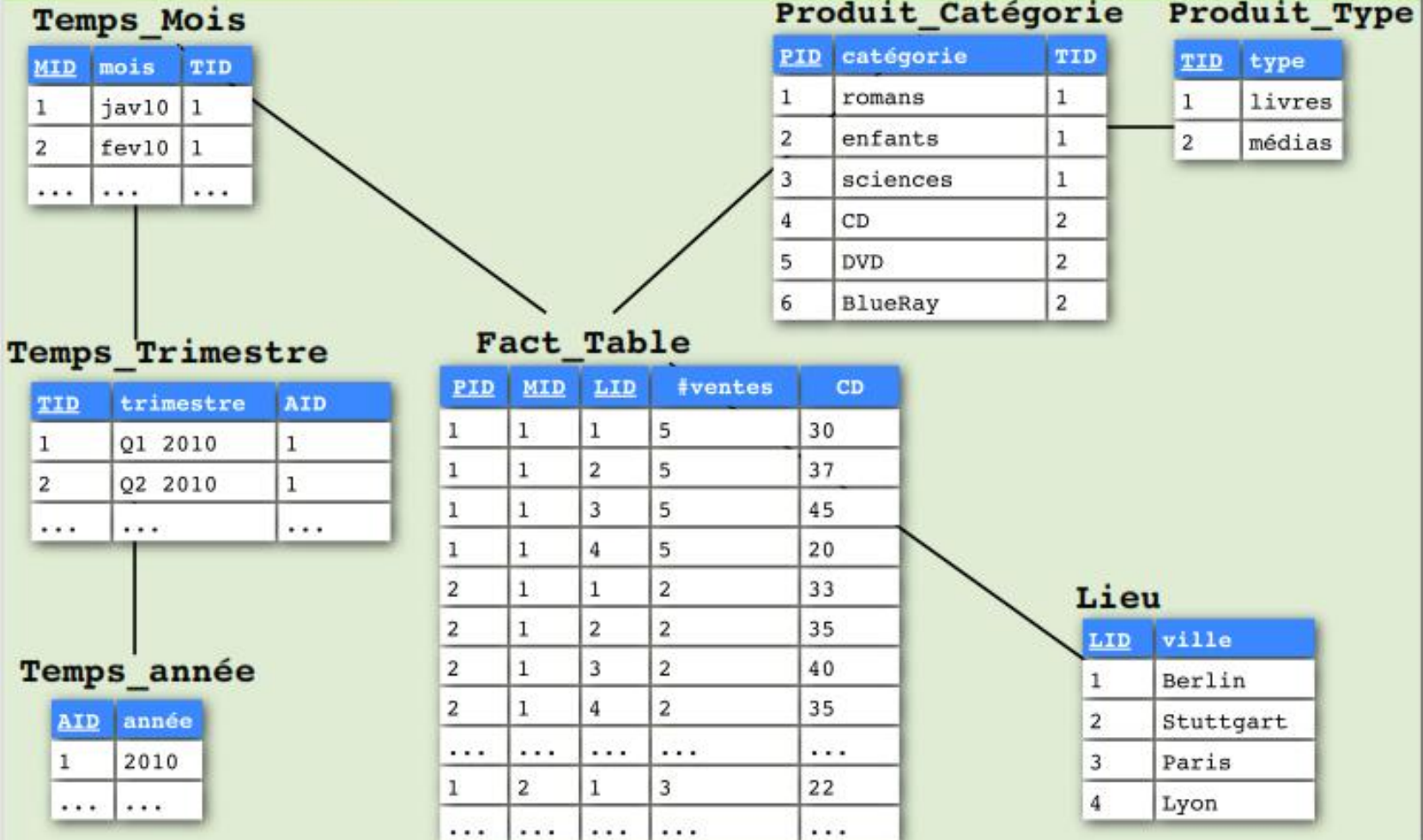
Modèle en Flocon (Bill Inmon)

- Une table de fait et des dimensions décomposées en sous hiérarchies (la dimension est présentée sur plusieurs tables)
- On a un seul niveau hiérarchique dans une table de dimension
- La table de dimension de niveau hiérarchique le plus bas est reliée à la table de fait. On dit qu'elle a la granularité la plus fine
- Avantages:
 - Normalisation des dimensions
 - Économie d'espace disque
- Inconvénients:
 - Modèle plus complexe (jointure)
 - Requêtes moins performantes

Modélisation multidimensionnelle

Modèle en Flocon (exemple)

Exemple d'un schéma en flocon



Modélisation multidimensionnelle

Processus de conception pour une base de données dimensionnelle (R. Kimball)

1 Choisir le processus d'activités opérationnelles à modéliser

- Ex: les commandes, la facturation, les stocks, les ventes, etc...

2. Choisir la granularité du processus d'activité (le niveau de détail fondamental)

- Ex: chaque transaction commerciale, ou une récapitulation quotidienne des transactions par client, ou une balance mensuelle du niveau de stock de chaque article, etc... Chaque « grain » sera une entité de la table des faits.

3. Choisir les dimensions applicables à chaque fait

- Ex: - Pour un fait de vente, le temps, le produit, le client (ou sa localisation), le type de transaction commerciale, la promotion, le statut de cette vente, etc...
- Pour chaque dimension retenue on construira une table de dimension

4. Choisir les mesures appliquées à chaque fait

- Ex: toutes infos retenues pour caractériser un fait deviendront des attributs de la table des faits.

Plan

- Introduction: le processus décisionnel
- Modélisation des Entrepôts de données (DW)
- **Approches d'implémentation des serveurs OLAP**
- Alimentation d'un DW
- Manipulation des données dimensionnelles
- BI et le Big Data



Approches d'implémentation d'OLAP

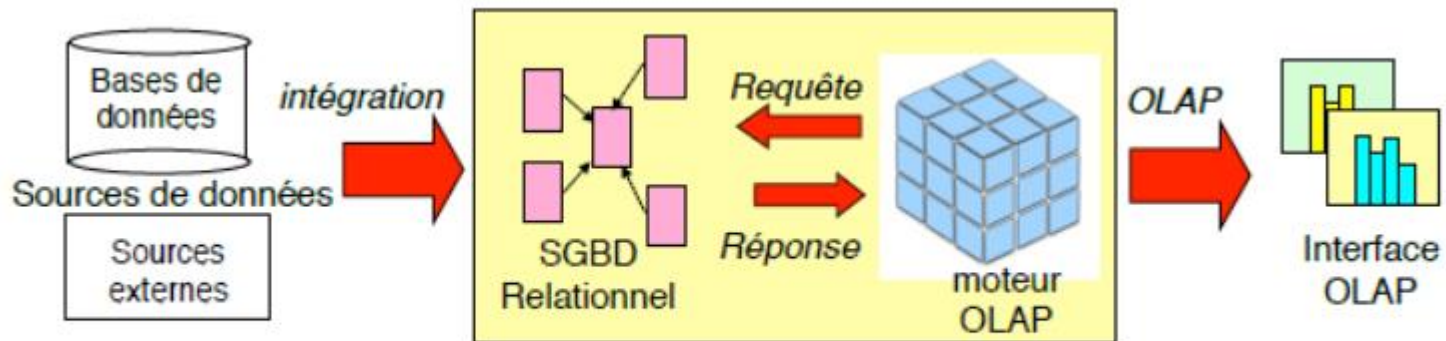


<https://olap.com/which-olap-is-best/>

Approches d'implémentation d'OLAP

- ROLAP (Relationnal OLAP)

- Données stockées dans une base de données relationnelles
- Un moteur OLAP permet de simuler le comportement d'un SGBD multidimensionnel



Source : B. Espinasse

- Plus facile et moins cher à mettre en place
 - Moins performant lors des phases de calcul

Approches d'implémentation d'OLAP

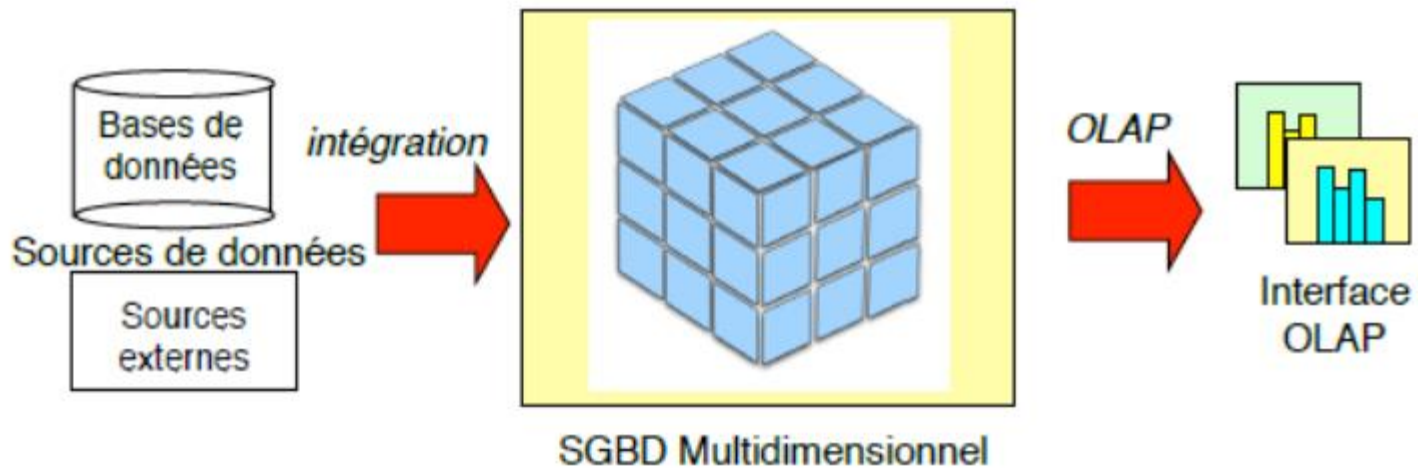
- ROLAP (Relationnal OLAP)

- OLAP sur du relationnel, c'est la technologies la plus utilisée en OLAP car les SGBD relationnels sont très largement répandus
- Plus facile et moins cher à mettre en place
- Moins performant lors des phases de calcul

Approches d'implémentation d'OLAP

- MOLAP (Multi dimensional OLAP)

- Utiliser un système multidimensionnel « pur » qui gère les structures multidimensionnelles natives (les cubes)
- Accès direct aux données dans le cube



Source : B. Espinasse

Approches d'implémentation d'OLAP

- MOLAP (Multi dimensional OLAP)

- MOLAP nécessite le pré-calcul et le stockage des informations du cube,
- Plus difficile à mettre en place
- Formats souvent propriétaires
- Conçu exclusivement pour l'analyse multidimensionnelle
- Mais, il permet des extractions très rapides et optimisées.

Approches d'implémentation d'OLAP

- HOLAP (Hybrid OLAP)

- Les systèmes HOLAP tentent d'exploiter le meilleur des deux techniques:
 - 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)

Quelques serveurs d'OLAP

OLAP Server	Compagnie	MOLAP	ROLAP	HOLAP
Essbase	Oracle	Yes	No	No
IBM Cognos BI	IBM	Yes	Yes	Yes
IBM Cognos TM1	IBM	Yes	No	No
icCube	icCube	Yes	No	No
Infor BI OLAP Server	Infor	Yes	No	No
Jedox OLAP Server (Palo)	Jedox	Yes	No	No
Microsoft Analysis Services	Microsoft	Yes	Yes	Yes
MicroStrategy Intelligence Server	MicroStrategy	Yes	Yes	Yes
Mondrian OLAP server	Pentaho	No	Yes	No
Oracle Database OLAP Option	Oracle	No	Yes	No
SAP NetWeaver BW	SAP	Yes	Yes	No
SAS OLAP Server	SAS Institute	Yes	Yes	Yes

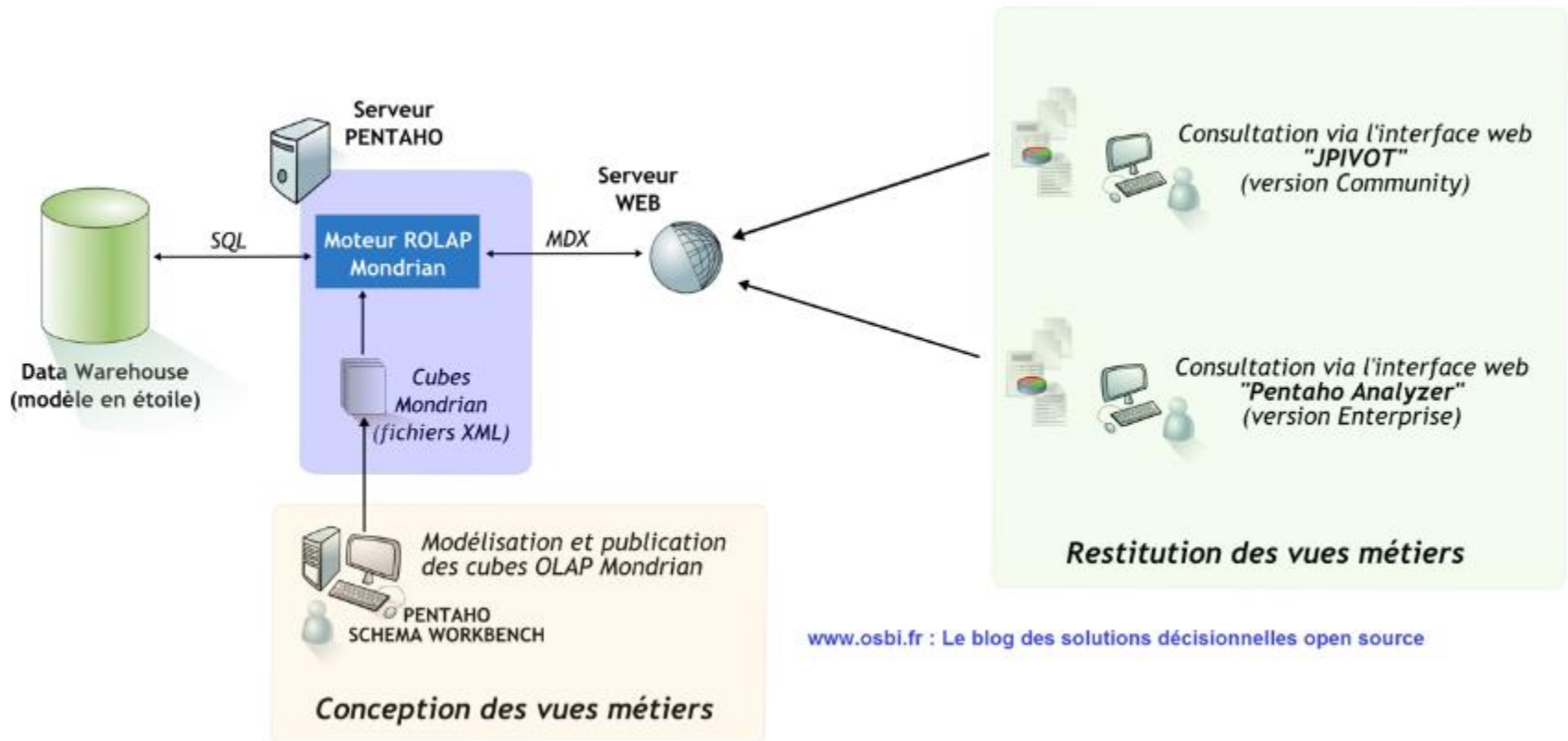
-Comparaison plus complète: https://en.wikipedia.org/wiki/Comparison_of_OLAP_Servers

11/12/2023

- Exemple du Serveur ROLAP: Mondrian
- Définition du schéma logique multidimensionnel

Architecture

- schéma d'architecture de Mondrian et son utilisation au sein de la plate-forme Pentaho.



- schéma d'architecture de Mondrian et son utilisation au sein de la plate-forme Pentaho.

Serveur Mondrian

- fait partie de la catégorie des serveurs ROLAP.
- Le projet Mondrian ainsi que son fondateur Julian Hyde ont rejoint le projet Pentaho sous le nom de *Pentaho Analysis*.
*Mondrian est utilisé par les serveurs BI : **Pentaho, JasperSoft et SpagoBI**.*
- Mondrian est utilisé avec les clients **Saiku, JPivot, JPalo, ou Pentaho Analyzer**
- Mondrian exécute des requêtes utilisant le langage MDX, également utilisé par d'autres moteurs OLAP, tel que celui de Microsoft SQL Server.

Schéma logique Mondrian

- Mondrian s'appuie sur des **schémas XML** pour la définition des cubes. Un *schéma Mondrian* permet donc de définir le modèle logique ainsi que le mapping sur le modèle physique :
 - Le modèle logique décrit les cubes, les dimensions, les hiérarchies, les niveaux et les membres (et plus encore...) sur lesquels vont s'appuyer les requêtes MDX.
 - Le modèle physique correspond à la source de données sur laquelle s'appuie le modèle logique (le modèle en étoile et/ou flocon)
- On peut utiliser l'outil **schema-workbench** pour modéliser le schéma.

Schéma logique Mondrian

- ❑ Le schéma mondrian est un fichier XML

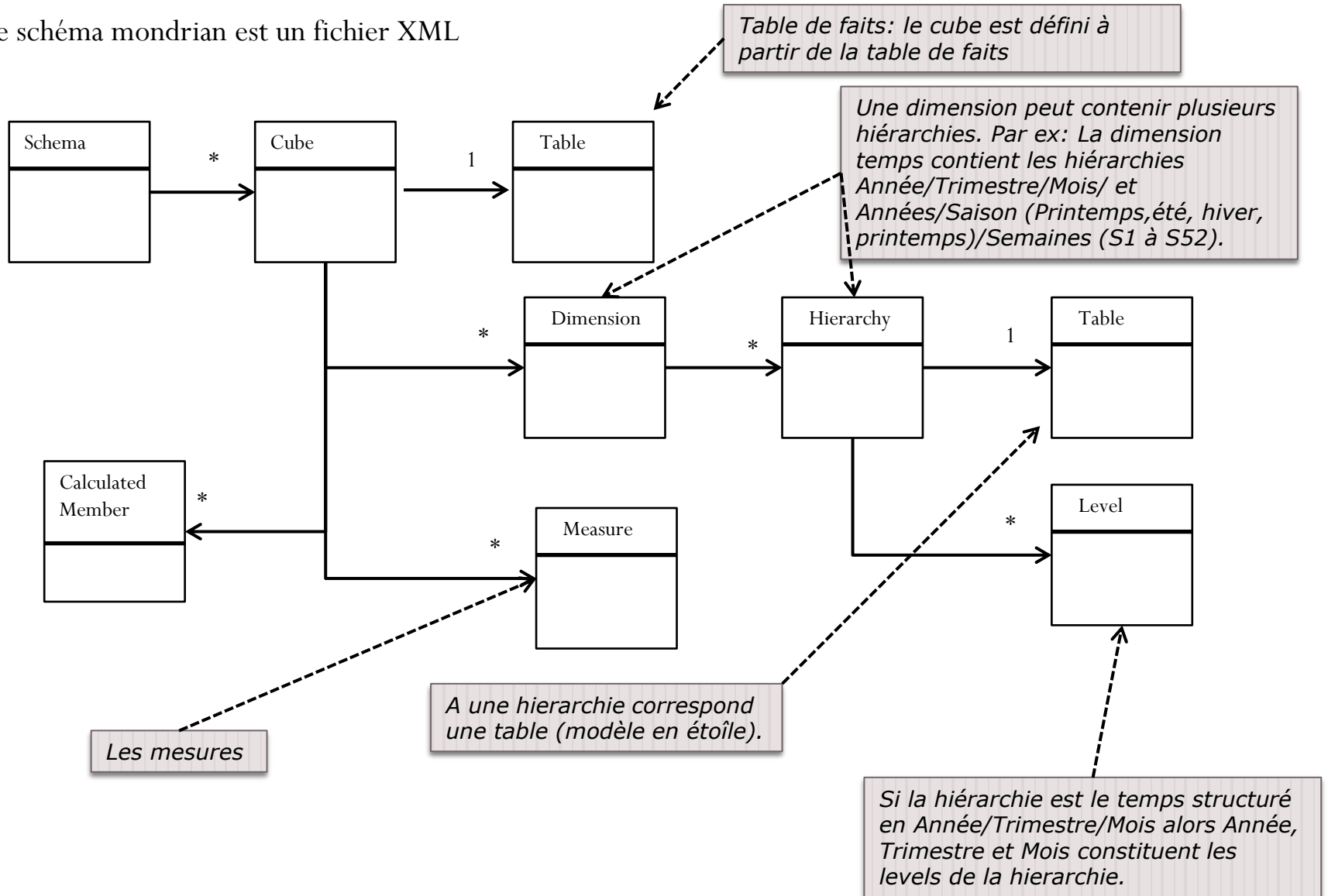


Schéma logique Mondrian (Exemple)

<Schema>

<Cube name="Sales">

Cube sales

<Table name="sales_fact_1997"/>

<Dimension name="Gender" foreignKey="customer_id">

<Hierarchy hasAll="true" allMemberName="All Genders"
primaryKey="customer_id">

<Table name="customer"/>

<Level name="Gender" column="gender" uniqueMembers="true"/>

</Hierarchy>

</Dimension>

<Dimension name="Time" foreignKey="time_id">

<Hierarchy hasAll="false" primaryKey="time_id">

<Table name="time_by_day"/>

<Level name="Year" column="the_year" type="Numeric"
uniqueMembers="true"/>

<Level name="Quarter" column="quarter" uniqueMembers="false"/>

<Level name="Month" column="month_of_year" type="Numeric"
uniqueMembers="false"/>

</Hierarchy>

</Dimension>

<Measure name="Unit Sales" column="unit_sales" aggregator="sum"
formatString="#,###"/>

<Measure name="Store Sales" column="store_sales" aggregator="sum"
formatString="#,###.##"/>

<Measure name="Store Cost" column="store_cost" aggregator="sum"
formatString="#,###.00"/>

<CalculatedMember name="Profit" dimension="Measures"
formula="[Measures].[Store Sales]-[Measures].[Store Cost]"/>

<CalculatedMemberProperty name="FORMAT_STRING" value="\$#,##0.00"/>

</CalculatedMember>

</Cube>

</Schema>

dimension

dimension

Niveaux

les mesures

On a une table de faits, 2 tables de dimensions.
Chaque niveau correspond à une partie de chaque table de dimension

Les mesures sont des attributs de la table de fait.

Plan

- Introduction: le processus décisionnel
- Modélisation des Entrepôts de données (DW)
- Approches d'implémentation des serveurs OLAP
- Alimentation d'un DW
- Manipulation des données dimensionnelles
- BI à l'ère des Big Data



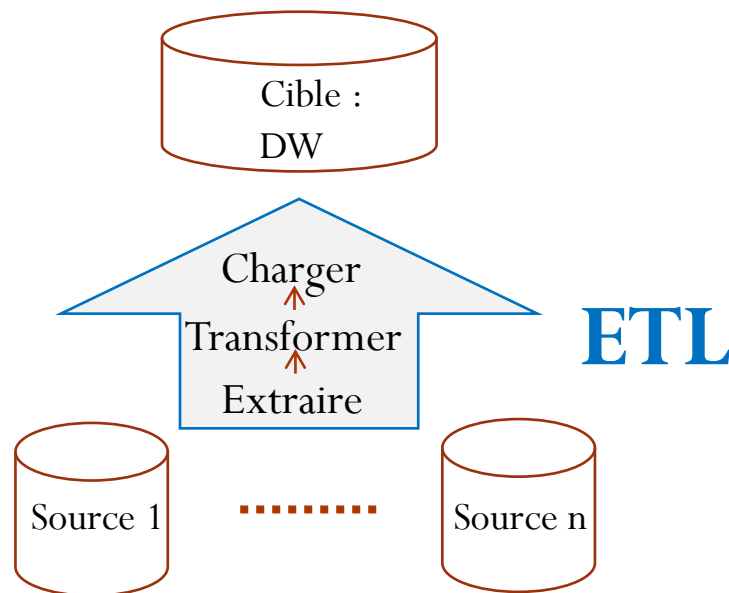
Alimentation/ mise à jour de l'entrepôt

- Besoin d'un système permettant d'automatiser les chargements dans l'entrepôt
- Entrepôt mis à jour régulièrement
- Réalisation et mise en œuvre des systèmes ETL (Extract Transform and Load). Elle constitue **70%** d'un projet décisionnel en moyenne. Ce système est complexe et ne doit rien laisser s'échapper, sous peine d'avoir une mauvaise information dans l'entrepôt, donc des données fausses, donc inutilisables.
- Utilisation d'outils ETL

Alimentation/ mise à jour de l'entrepôt

ETL

Extract-Transform-Load (ETL), extracto-chargeur, (ou parfois : datapumping) est une technologie informatique intergicielle (middleware). C'est une séquence d'opérations portant sur les données qui sont collectés à partir de plusieurs sources, structurées, centralisées dans un référentiel unique



Alimentation/ mise à jour de l'entrepôt

Caractéristiques d'un ETL

- Offre un environnement de développement
- Offre des outils de gestion des opérations et de maintenance
- Permet de découvrir, analyser et extraire les données à partir de sources hétérogènes
- Permet de nettoyer et standardiser les données
- Permet de charger les données dans un entrepôt

Alimentation/ mise à jour de l'entrepôt

Extraction

- Extraire des données des systèmes de production
- Dialoguer avec différentes sources:
 - Base de données,
 - Fichiers,
 - Bases propriétaires
- Utilise divers connecteurs :
 - ODBC,
 - SQL natif,
 - Fichiers plats

Alimentation/ mise à jour de l'entrepôt

Transformation

- Rendre cohérentes les données des différentes sources
- Transformer, nettoyer, trier, unifier les données
- Exemple: unifier le format des dates (MM/JJ/AA , JJ/MM/AA)
- Etape très importante, garantit la cohérence et la fiabilité des données

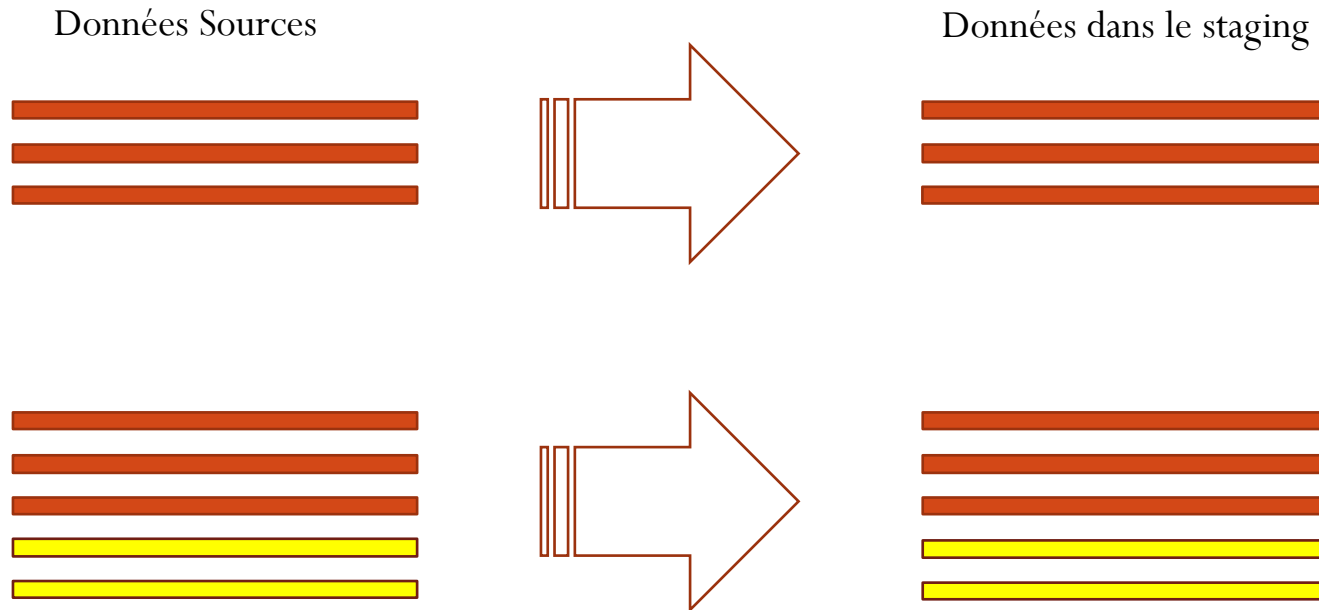
Chargement

- Insérer ou modifier les données dans l'entrepôt
- Utilisation de connecteurs:
 - ODBC,
 - SQL natif,
 - Fichiers plats

Alimentation/ mise à jour de l'entrepôt

Conception d'un ETL

- Deux cas sont à prendre en compte, le **chargement initial** (Full) et les **chargements incrémentiels** (Delta).



Conception d'un ETL

Etudier les sources de données selon lesquelles trois stratégies de chargement peuvent se faire :

- **Push** : le système de production **pousse** les données vers le Staging quand il en a l'occasion.

Inconvénient: Si le système est occupé, il ne poussera jamais les données.

- **Pull** : **Tire** les données de la source vers le Staging.

Inconvénient : peut surcharger le système s'il est en cours d'utilisation.

- **Push-Pull** : La source prépare les données à envoyer et prévient le Staging qu'elle est prête. Le Staging va récupérer les données. Si la source est occupée, le Staging fera une autre demande plus tard.

Alimentation/ mise à jour de l'entrepôt

Les métadonnées

- Données décrivant l'environnement décisionnel.
 - Quelle est cette information ?
 - D'où provient-elle ?
 - Comment est-elle calculée ?
 - De quand date la dernière mise à jour?
 - Quelles sont les précautions d'usage ?..
- Clé de réussite de tout projet décisionnel.
- Assurent l'interopérabilité entre les systèmes

Alimentation/ mise à jour de l'entrepôt

Outils ETL open source

- **Pentaho Data Integration (PDI):** la solution d'ETL Kettle intégrée au sein du projet Pentaho.

Le site de l'éditeur: <http://community.pentaho.com/projects/data-integration/>
<https://sourceforge.net/projects/pentaho/files/Data%20Integration/>

- **Talend Open Studio:** particulièrement complet. Talend Open Studio. Le site : www.talend.com/
- **Enhydra Octopus:** Il se connecte aux bases de données sous JDBC et s'appuie sur un schéma XML. Le site de l'éditeur : www.together.at
- **Clover ETL :** une solution d'intégration des données écrite en Java. Le site de l'éditeur : cloveretl.com
- **Ketl :** Le site de l'éditeur : ketl.org

Plan

- Introduction: le processus décisionnel
- Modélisation des Entrepôts de données (DW)
- Approches d'implémentation des serveurs OLAP
- Alimentation d'un DW
- Manipulation des données dimensionnelles
- BI à l'ère des Big Data



Manipulation des données cube

Exemple : Cube Vente

VILLES				Lyon		Marseille		Paris	
Mesures				CA	Qté vendue	CA	Qté vendue	CA	Qté vendue
PRODUITS	Home Cinéma	DATES	01/02/2015	20 000,00 €	21	6 000,00 €	4	14 000,00 €	22
			04/02/2015	6 000,00 €	13	1 600,00 €	7	7 500,00 €	17
			08/03/2015	1 200,00 €	4	16 000,00 €	7	2 000,00 €	14
			09/03/2015	5 500,00 €	4	1 200,00 €	4	1 500,00 €	1
			15/04/2015	800,00 €	3				
			16/04/2015	1 700,00 €	11	1 870,00 €	6	1 600,00 €	3
			17/04/2015			1 900,00 €	9	600,00 €	1
			22/04/2015			2 000,00 €	3		
			23/04/2015			650,00 €	2		
	24/04/2015				500,00 €	2			
	App. Photo num		01/02/2015	16 000,00 €	30	10 140,00 €	10	9 000,00 €	25
			04/02/2015	2 600,00 €	5	2 000,00 €	6	13 800,00 €	16
			08/03/2015	2 000,00 €	7	20 000,00 €	8	5 000,00 €	21
			09/03/2015	7 500,00 €	5	1 000,00 €	3	1 720,00 €	1
			15/04/2015	1 200,00 €	4				
			16/04/2015	5 000,00 €	30	1 330,00 €	4	3 000,00 €	6
			17/04/2015			1 100,00 €	4	400,00 €	1
			22/04/2015			1 200,00 €	2		
			23/04/2015			850,00 €	3		
	24/04/2015				400,00 €	3			
	Lecteurs DVD		01/02/2015	15 400,00 €	20	4 000,00 €	6	7 000,00 €	15
			04/02/2015	2 900,00 €	5	6 000,00 €	14	1 700,00 €	4
			08/03/2015	4 000,00 €	13	14 000,00 €	5	1 000,00 €	5
			09/03/2015	7 000,00 €	5	3 000,00 €	7	1 780,00 €	1
			15/04/2015	4 000,00 €	8				
			16/04/2015	2 000,00 €	10	2 300,00 €	2	2 400,00 €	3
			17/04/2015			2 200,00 €	10	400,00 €	1
			22/04/2015			400,00 €	1		
			23/04/2015			500,00 €	3		
	24/04/2015				600,00 €	5			

Manipulation des données cube

Opérations sur le cube

- Opération agissant sur la granularité
 - **Forage vers le haut (roll-up):** « dézoomer »
 - Obtenir un niveau de granularité supérieur
 - Utilisation de fonctions d'agrégation
 - **Forage vers le bas (drill-down):** « zoomer »
 - Obtenir un niveau de granularité inférieur
 - Données plus détaillées

Manipulation des données cube

Opérations sur le cube

- Roll UP (Exemple)

- Roll Up sur la dimension Produit

Villes		Lyon		Marseille		Paris	
Mesures		CA	Qté vendue	CA	Qté vendue	CA	Qté vendue
Dates	01/02/15	54 100,00 €	71	20 140,00 €	20	30 000,00 €	62
	04/02/15	11 500,00 €	23	9 600,00 €	27	23 000,00 €	37
	08/03/15	7 200,00 €	24	50 000,00 €	18	8 000,00 €	40
	09/03/15	20 000,00 €	14	5 200,00 €	14	5 000,00 €	3
	15/04/15	6 000,00 €	15				
	16/04/15	8 700,00 €	51	5 500,00 €	13	7 500,00 €	12
	17/04/15			5 200,00 €	23	1 400,00 €	3
	22/04/15			3 600,00 €	6		
	23/04/15			2 000,00 €	6		
	24/04/15			1 500,00 €	3		

Manipulation des données cube

Opérations sur le cube

- Roll UP (Exemple)

- Roll Up sur les deux dimensions Date et Produit

Villes	Lyon		Marseille		Paris	
Mesures	CA	Qté vendue	CA	Qté vendue	CA	Qté vendue
	107 500,00 €	198	101 240,00 €	130	74 900,00 €	157

- Roll Up sur toutes les dimensions

Villes					
Mesures	CA			Qté vendue	
	283 640,00 €			485	

Manipulation des données cube

Opérations sur le cube - drill-down (Exemple)

Villes	Lyon		Marseille		Paris	
Mesures	CA	Qté vendue	CA	Qté vendue	CA	Qté vendue
	107 500,00 €	198	101 240,00 €	137	74 900,00 €	157

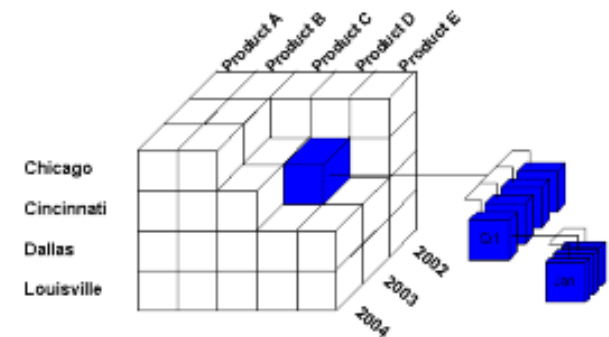
Villes		Lyon	
Mesures		CA	Qté vendue
Dates	01/02/15	54 100,00 €	71
	04/02/15	11 500,00 €	23
	08/03/15	7 200,00 €	24
	09/03/15	20 000,00 €	14
	15/04/15	6 000,00 €	15
	16/04/15	8 700,00 €	51
	17/04/15		
	22/04/15		
	23/04/15		
	24/04/15		

Manipulation des données cube

Opérations sur le cube

- Roll UP et drill down (un autre exemple)

		05 06 07			Dimension Temps						
		Alim.	496	520	255						
05-07		05 06 07									
Fruits	623	Fruits	221	263	139	Fruits	100	121	111	152	139
Viande	648	Viande	275	257	116	Viande	134	141	120	137	116
		05 06 07			Drill down						
		Pomme	20	19	22						
							
		Boeuf	40	43	48						
		Dimension Produit									



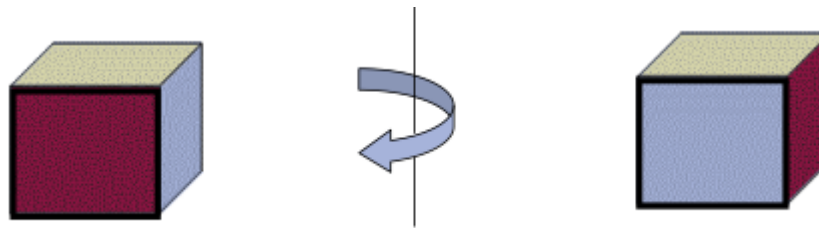
docs.oracle.com

Manipulation des données cube

Opérations sur le cube

- Opération agissant sur la structure

Rotation (rotate): présenter une autre face du cube



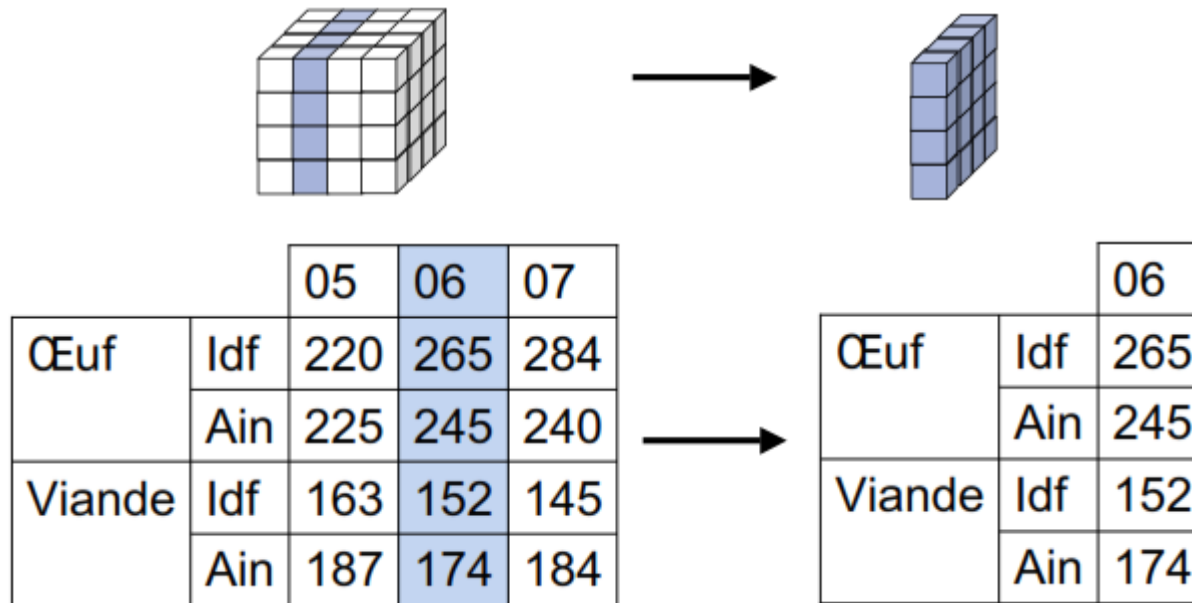
VILLES		Lyon		Marseille		Paris	
Mesures		CA	Qté vendue	CA	Qté vendue	CA	Qté vendue
PRODUITS	Home Cinéma	01/02/2015	20 000,00 € 21	6 000,00 € 4	14 000,00 € 22		
		04/02/2015	6 000,00 € 13	1 600,00 € 7	7 500,00 € 17		
		08/03/2015	1 200,00 € 4	16 000,00 € 7	2 000,00 € 14		
		09/03/2015	5 500,00 € 4	1 200,00 € 4	1 500,00 € 1		
		15/04/2015	800,00 € 3				
		16/04/2015	1 700,00 € 11	1 870,00 € 6	1 600,00 € 3		
		17/04/2015		1 900,00 € 9	600,00 € 1		
		22/04/2015		2 000,00 € 3			
	App. Photo num	23/04/2015		650,00 € 2			
		24/04/2015		500,00 € 2			
		01/02/2015	16 000,00 € 30	10 140,00 € 10	9 000,00 € 25		
		04/02/2015	2 600,00 € 5	2 000,00 € 6	13 800,00 € 16		
		08/03/2015	2 000,00 € 7	20 000,00 € 8	5 000,00 € 21		
		09/03/2015	7 500,00 € 5	1 000,00 € 3	1 720,00 € 1		
		15/04/2015	1 200,00 € 4				
		16/04/2015	5 000,00 € 30	1 330,00 € 4	3 000,00 € 6		
PRODUITS	Lecteurs DVD	17/04/2015		1 100,00 € 4	400,00 € 1		
		22/04/2015		1 200,00 € 2			
		23/04/2015		850,00 € 3			
		24/04/2015		400,00 € 3			
		01/02/2015	15 400,00 € 20	4 000,00 € 6	7 000,00 € 15		
		04/02/2015	2 900,00 € 5	6 000,00 € 14	1 700,00 € 4		
		08/03/2015	4 000,00 € 13	14 000,00 € 5	1 000,00 € 5		
		09/03/2015	7 000,00 € 5	3 000,00 € 7	1 780,00 € 1		
	DATES	15/04/2015	4 000,00 € 8				
		16/04/2015	2 000,00 € 10	2 300,00 € 2	2 400,00 € 3		
		17/04/2015		2 200,00 € 10	400,00 € 1		
		22/04/2015		400,00 € 1			
		23/04/2015		500,00 € 3			
		24/04/2015		600,00 € 5			

PRODUITS		Home Cinéma						App. Photo num					
VILES		Lyon		Marseille		Paris		Lyon		Marseille		Paris	
Mesures		CA	Qté vendue	CA	Qté vendue	CA	Qté vendue	CA	Qté vendue	CA	Qté vendue	CA	Qté vendue
DATES	01/02/2015	20 000,00 €	21	6 000,00 €	4	14 000,00 €	22	16 000,00 €	30	10 140,00 €	10	9 000,00 €	25
	04/02/2015	6 000,00 €	13	1 600,00 €	7	7 500,00 €	17	2 600,00 €	5	2 000,00 €	6	13 800,00 €	16
	08/03/2015	1 200,00 €	4	16 000,00 €	7	2 000,00 €	14	2 000,00 €	7	20 000,00 €	8	5 000,00 €	21
	09/03/2015	5 500,00 €	4	1 200,00 €	4	1 500,00 €	1	7 500,00 €	5	1 000,00 €	3	1 720,00 €	1
	15/04/2015	800,00 €	3					1 200,00 €	4				
	16/04/2015	1 700,00 €	11	1 870,00 €	6	1 600,00 €	3	5 000,00 €	30	1 330,00 €	4	3 000,00 €	6
	17/04/2015			1 900,00 €	9	600,00 €	1			1 100,00 €	4	400,00 €	1
	22/04/2015			2 000,00 €	3					1 200,00 €	2		
	23/04/2015			650,00 €	2					850,00 €	3		
	24/04/2015			500,00 €	2					400,00 €	3		

Manipulation des données cube

Opérations sur le cube

- Opération agissant sur la structure
 - **Tranchage (slicing)**: consiste à ne travailler que sur une tranche du cube. Une des dimensions est alors réduite à une seule valeur

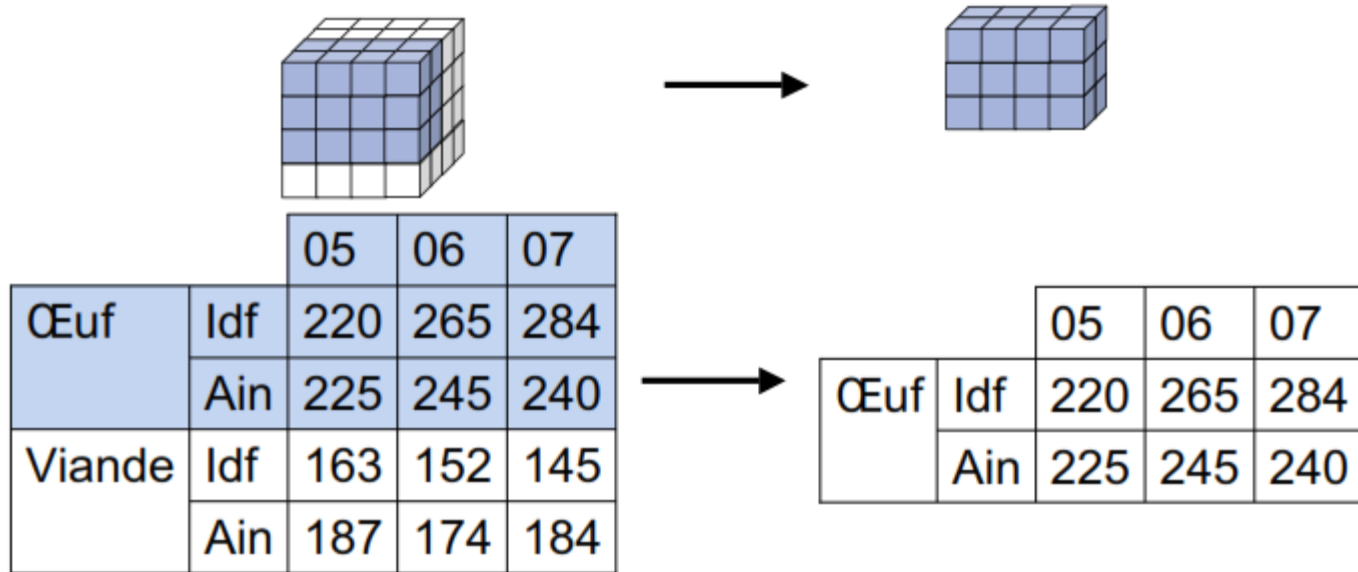


Manipulation des données cube

Opérations sur le cube

- Opération agissant sur la structure

Extraction d'un bloc de données (dicing): ne travailler que sous un sous-cube



Manipulation des données cube

Opérations sur le cube

Exemple : 'Slice' et 'Dice'

Opération de sélection et de projection

Ville = 'Paris' et Dates < '10/03/2008'

Villes		Paris					
Produits		Home Cinéma		App. Photo numériques		Lecteur DVD	
Mesures		CA	Qté vendue	CA	Qté vendue	CA	Qté vendue
Dates	01/02/2008	14 000,00 €	22	9 000,00 €	25	7 000,00 €	15
	04/02/2008	7 500,00 €	17	13 800,00 €	16	1 700,00 €	4
	08/03/2008	2 000,00 €	14	5 000,00 €	21	1 000,00 €	5
	09/03/2008	1 500,00 €	1	1 720,00 €	1	1 780,00 €	1

Manipulation des données cube

Langage de manipulation - Requêtes OLAP-

Deux langages pour manipuler les données

- Extention SQL
- MDX

Manipulation des données cube

Langage de manipulation MDX

- MDX = **MultiDimensional Expressions**: Proposition de Microsoft “OLE DB for OLAP”
- Langage permettant de définir, d'utiliser et de récupérer des données à partir d'objets multidimensionnels
- Permet d'effectuer les opérations décrites précédemment
- Equivalent de SQL pour le monde OLAP
- Origine: Microsoft

Manipulation des données cube

MDX vs SQL

- MDX est fait pour **naviguer** dans les bases **multidimensionnelles** et pour définir des **requêtes** sur tous leurs **objets** (dimensions, hiérarchies, niveaux, membres et cellules) afin d'obtenir (simplement) une **représentation** sous forme de tableaux croisés
- MDX ressemble à SQL par ses mots clé **SELECT, FROM, WHERE**, mais :
 - SQL construit des vues relationnelles
 - MDX construits des vues multidimensionnelles des données
- Analogies entre termes multidimensionnels (MDX) et relationnels (SQL) :

Multidimensionnel (MDX)	Relationnel (SQL)
cube	Table
Niveau (level)	Colonne
Dimension	Plusieurs colonnes liées ou une table de dimension
Mesure (Measure)	Colonne
Membre	Valeur de la colonne d'une ligne

Manipulation des données cube

MDX vs SQL

	SQL	MDX
Structure générale de la requête	SELECT column1, column2,..., columnn FROM table	SELECT axe1 ON COLUMNS, axe2 ON ROWS FROM cube
FROM	une ou plusieurs tables	un cube
SELECT	<ul style="list-style-type: none">- une vue des données en 2 dimensions (lignes (rows) et colonnes (columns))- les lignes ont la même structure définie par les colonnes	<ul style="list-style-type: none">- nb quelconque de dimensions pour former les résultats de la requête.- pas de signification particulière pour les rows et les columns, mais il faut définir chaque axe : axe1 définit l'axe horizontal et axe2 définit l'axe vertical

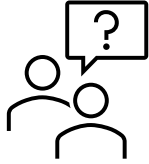
Plan

- Introduction: le processus décisionnel
- Modélisation des Entrepôts de données (DW)
- Approches d'implémentation des serveurs OLAP
- Alimentation d'un DW
- Manipulation des données dimensionnelles
- BI et le Big Data



BI et Big Data

Approche BI classique



Les utilisateurs
d'entreprise
déterminent les
questions à poser

Données structurées
et analyses répétées

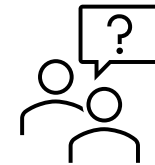


répondre aux questions posées
(analyse **descriptive** ou de
diagnostic)

Approche Big Data Analytics



Données brutes , non structurées,
structurées et analyse itérative



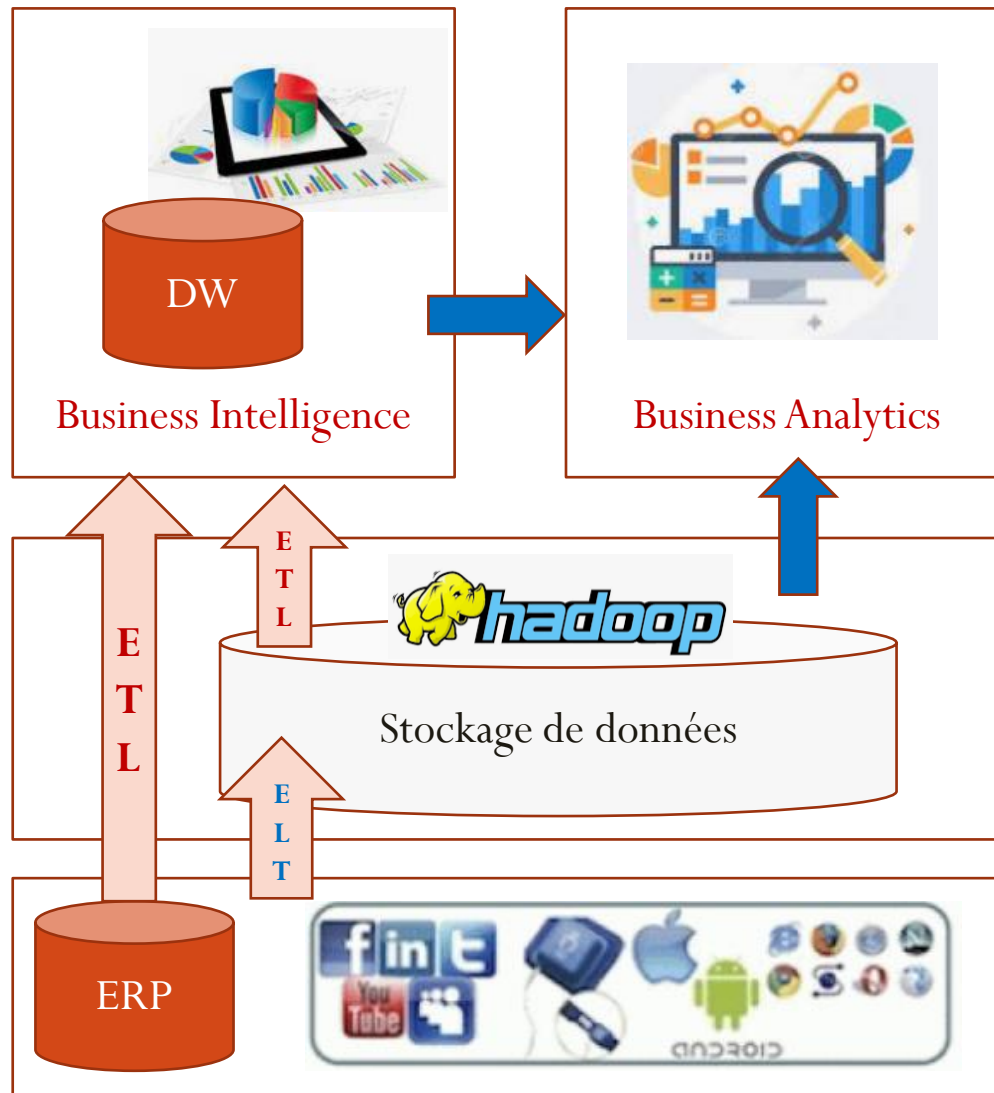
Explorent la données
(Analyse **prédictive**
Analyse **prescriptive**)

BI et Big Data

BI et **big data** permettent de **recupérer** et **traiter** des données pour atteindre de meilleures performances mais **différemment**.

	Big Data	BI traditionnelle
Méthodes d'analyse	Ad-hoc : on ne sait pas encore ce que veulent dire les données et les futures utilisations de ces données.	Préparées : données déjà préstructuré pour effectuer des requêtes déjà prédéfinie pour un résultat escompté
Types de données	Brutes + structurées + Non structurées	Structurées
sources	Opérationnelles mais beaucoup de sources Externes et différentes: stocker massivement.	Opérationnelles
Stockage	Stockage parallèle de données massives (Framework Hadoop)	Data Warehouse
Utilisation	Prédire les nouvelles tendances	Orienter les décisions des managers

BI et Big Data



- Exploitation des données:
 - directe (BAalytics)
 - indirecte (BI)
- Socle Big data :
Intégration en temps réel des flux de données structurées et non structurées, NoSQL et relationnelles
- Données sources :
structurées, internes,
externes, non structurées

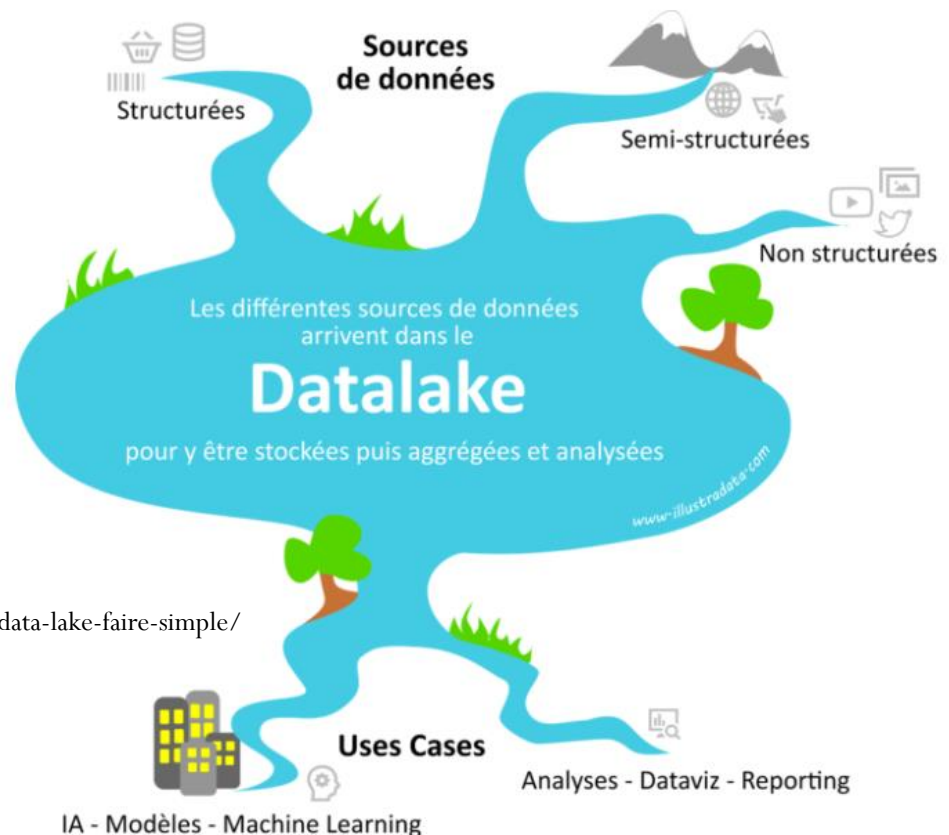
BI et Big Data

Data Lake

Est un système évolutif de stockage et d'analyse de données de tous types (un **réservoir de données brutes**) utilisé principalement par des spécialistes de données (Data scientists, data analysts)



<http://www.illustradata.com/data-lake-faire-simple/>



BI et Big Data

Data lake vs DWH

	DWH	Data lake
Données	Nettoyées	Brutes
Structure	structurées	Brutes + structurées + Non structurées
Sources	Restreintes	multiples
Schéma	En écriture (prédéfini)	en lecture (non prédéfini)
Accès aux données	MDX/SQL	Scripts/Programmes
Intégration de données	ETL	ELT
Utilisateurs	Expert métier	Data Scientists
Analyses	Industrialisées	À la demande