UNIVERSITÉ LUMIÈRE LYON 2

ICOM INSTITUT DE LA COMMUNICATION

Big data management
- partie -
Entrepôts de données

Master 2 Data Mining
Année 2016-2017
Jérôme Darmont

http://eric.univ-lyon2.fr/~jdarmont/

Actualités du cours

http://eric.univ-lyon2.fr/~jdarmont/?page_id=3144

http://eric.univ-lyon2.fr/~jdarmont/?feed=rss2

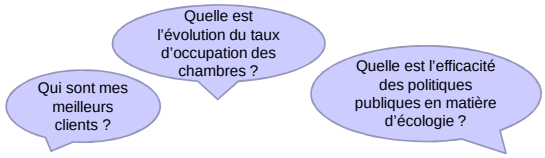
https://twitter.com/darmont_lyon2 *hashtag* #dmed

Plan du cours

- Introduction : le processus décisionnel
- Modélisation conceptuelle des entrepôts
- Modélisation logique des entrepôts
- Mise en œuvre d'un entrepôt de données
- Analyse en ligne (OLAP)

BI or not BI?

- Informatique décisionnelle (*business intelligence*) : à l'usage des décideurs
  - Accéder rapidement et simplement aux informations stratégiques
  - Donner du sens aux données
  - Donner une vision transversale des données d'une organisation
  - Extraire, grouper, organiser, agréger corréler les données



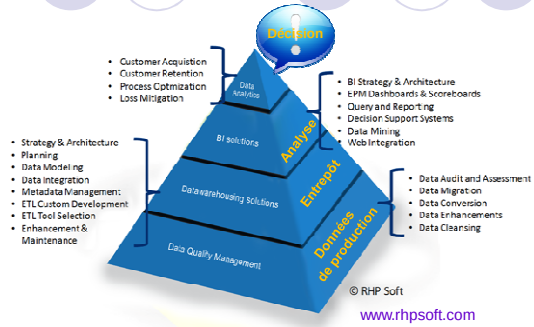
Problématique

- Données disponibles
  - Volumineuses
  - Hétérogènes
  - Très détaillées

- Traitement
  - Synthétiser/résumer
  - Visualiser
  - Analyser

- Utilisateurs
  - Non informaticiens
  - Non statisticiens

Système d'information décisionnel



www.rhpsoft.com

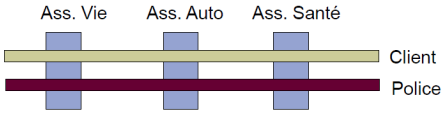
Entrepôt de données, la définition

Un entrepôt de données est une 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, 1991

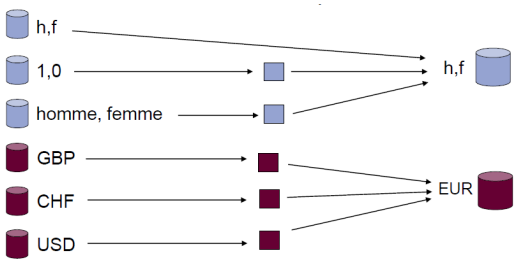
Orientées sujet

- Agrégation des informations de différents métiers
- Pas de prise en compte de l'organisation fonctionnelle des données



Lydie Soler, AgroParisTech/INRA

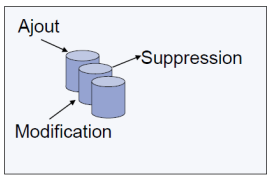
Intégrées



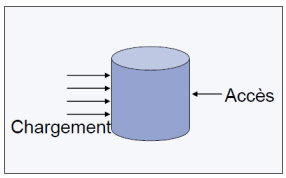
Lydie Soler, AgroParisTech/INRA

Non volatiles

Base de données opérationnelle



Entrepôt de données



Lydie Soler, AgroParisTech/INRA

On Line Transaction Processing *vs.* **On Line Analytical Processing**

Historisées

Base de données opérationnelle

Image de la base en Mai 2005

Répertoire

Nom	Ville
Dupont	Paris
Durand	Lyon

Image de la base en Juillet 2006

Répertoire

Nom	Ville
Dupont	Marseille
Durand	Lyon

Entrepôt de données

Calendrier

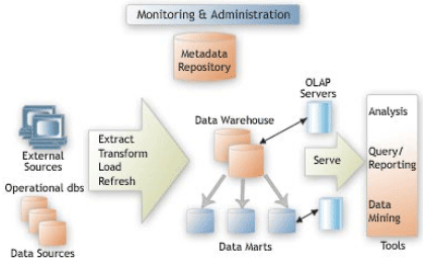
Code	Année	Mois
1	2005	Mai
2	2006	Juillet

Répertoire

Code	Année	Mois
1	Dupont	Paris
1	Durand	Lyon
2	Dupont	Marseille

Lydie Soler, AgroParisTech/INRA

Processus d'entreposage de données



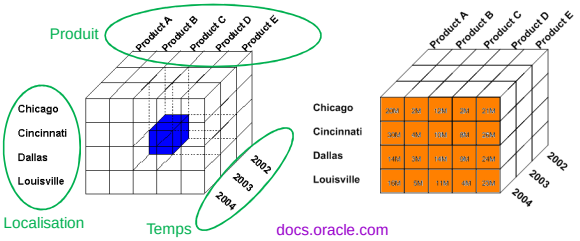
www.kahassoc.com

Plan du cours

- ✓ Introduction : le processus décisionnel
- Modélisation conceptuelle des entrepôts
- Modélisation logique des entrepôts
- Mise en œuvre d'un entrepôt de données
- Analyse en ligne (OLAP)

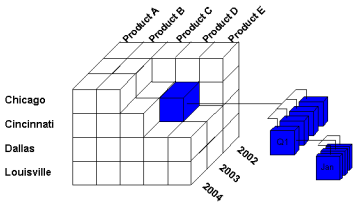
Métaphore du cube de données

- Fait : sujet d'analyse
  - Ensemble de mesures
- Dimensions : axes d'analyse



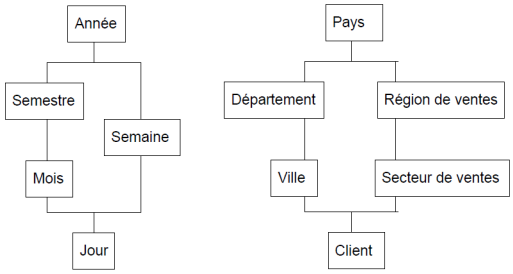
docs.oracle.com

Hiérarchie de dimension



docs.oracle.com

Hiérarchies multiples



Elsa Nègre, Université Paris Dauphine

Attributs de dimensions

- Paramètres
  - Définissent les niveaux hiérarchiques
- Attribut faibles
  - Descriptifs

DIM PRODUIT
Nom
Type
Gamme
Prix unitaire
Couleur

Schéma en étoile

Modèle dénormalisé

+ Limitation des jointures

- Redondance

DIM PRODUIT
Nom
Type
Gamme
Prix unitaire
Couleur

FAIT VENTE
Chiffre d'affaire

DIM LOCALISATION
Ville
Région
Pays

DIM TEMPS
Jour
Mois
Trimestre
Année

Schéma en flocon de neige

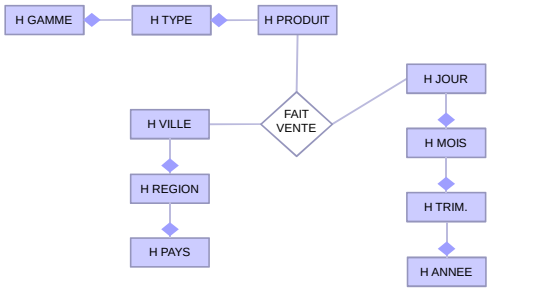
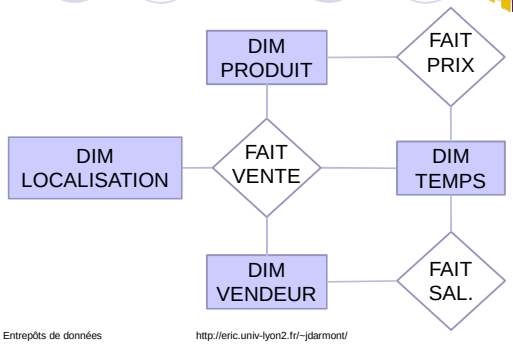


Schéma en constellation



Modèle multidimensionnel de Rizzi et Golfarelli

Plan du cours

- ✓ Introduction : le processus décisionnel
- ✓ Modélisation conceptuelle des entrepôts
- Modélisation logique des entrepôts
- Mise en œuvre d'un entrepôt de données
- Analyse en ligne (OLAP)

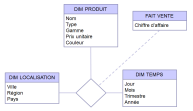
Quelle approche pour l'OLAP ?

olap.com/which-olap-is-best/

Approche ROLAP

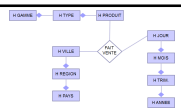
- Relational OLAP : Stockage de l'entrepôt dans une base de données relationnelle
  - Faits, dimensions ou niveaux hiérarchiques : tables
  - Analyse OLAP : requêtes SQL99 (GROUP BY CUBE…)
- Avantages
  - Facilité et faible coût de mise en œuvre
  - Stockage de gros volumes de données
  - Evolution facile
- Inconvénients
  - Performance (jointures)
  - Reformatage nécessaire des résultats pour les utilisateurs finaux

Étoile ROLAP



- DIM_PRODUIT(IDproduit, Nom, Type, Gamme, PrixUnitaire, Couleur)
- DIM_LOCALISATION(IDloc, Ville, Région, Pays)
- DIM_TEMPS(IDtemps, Jour, Mois, Trimestre, Année)
- FAIT_VENTE(IDproduit#, IDloc#, IDtemps#, ChiffreAffaire)

Flocon ROLAP



- H_PRODUIT(IDproduit, Nom, PrixUnitaire, Couleur, IDtype#)
- H_TYPE(IDtype, NomType, IDgamme#
- H_GAMME(IDgamme, NomGamme)

DIM_PRODUIT

- H_VILLE(IDville, NomVille, IDrégion#)
- H_REGION(IDrégion, NomRégion, IDpays#)
- H_PAYS(IDpays, NomPays)

DIM_LOCALISATION

- H_JOUR(IDjour, Jour, IDmois#)
- H_MOIS(IDmois, Mois, IDtrim#)
- H_TRIMESTRE(IDtrim, Trimestre, IDannée#)
- H_ANNEE(IDannée, Année)

DIM_TEMPS

- FAIT_VENTE(IDproduit#, IDville#, IDjour#, Chiffre_affaire)

Index binaires

ID	Titre	Année	Genre
1	Brazil	1984	Science Fiction
2	Underground	1995	Drame
3	Easy Rider	1969	Drame
4	Psychose	1960	Drame
5	Annie Hall	1977	Comédie
6	Jurasic Park	1992	Science Fiction
7	Metropolis	1926	Science Fiction
8	Manhattan	1979	Comédie
9	Smoke	1995	Comédie

Relation Film

Index binaire sur l'attribut Genre

N-uplet	9	8	7	6	5	4	3	2	1
Science Fiction	0	0	1	1	0	0	0	0	1
Drame	0	0	0	0	0	1	1	1	0
Comédie	1	1	0	0	1	0	0	0	0

Index binaires et entrepôts

- Avantages
  - Faible coût de stockage
  - Rapides en lecture, pas d'accès aux données pour :
    - Requêtes de comptage
    - Opérations bits à bits
- Inconvénient
  - Peu performants si mises à jour nombreuses
    - Rafraîchissement des index

Opérations sur index binaires

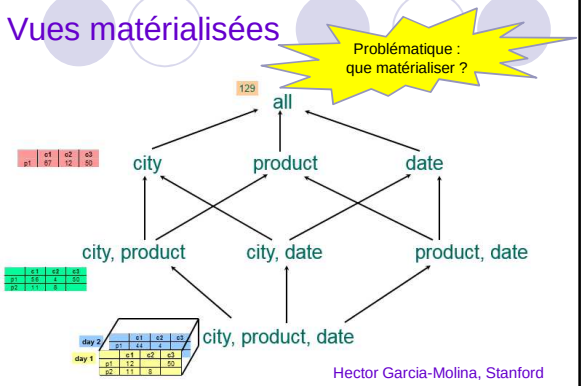
- Nombre de comédies
  Compter le nombre de 1 dans le bitmap associé

N-uplet	9	8	7	6	5	4	3	2	1
Science Fiction	0	0	1	1	0	0	0	0	1
Drame	0	0	0	0	0	1	1	1	0
Comédie	1	1	0	0	1	0	0	0	0

- Nombre de comédies en 1995
  AND entre les deux bitmaps correspondants et comptage

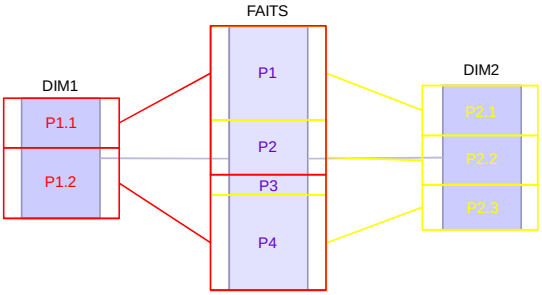
N-uplet	9	8	7	6	5	4	3	2	1
Comédie	1	1	0	0	1	0	0	0	0
1995	1	0	0	0	0	0	0	1	0
AND	1	0	0	0	0	0	0	0	0

Vues matérialisées



Hector Garcia-Molina, Stanford

Fragmentation (horizontale dérivée)



Approche MOLAP

- Multidimensional OLAP : Stockage natif des cubes dans des tableaux multidimensionnels

Dimensional view on gross revenue

		Product group			
		Books	News papers	Magazines	Maps
Country	France	500	452	124	35
	Spain	852	634	236	85
	Austria	632	234	963	45
	Belgium	459	325	456	96

www.wikibrl.info

- Avantage
  - 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
  - Rafraîchissement limité (reconstruction périodique complète)

Compression de cube

Store	Customer	Product	Price
S1	C2	P2	$70
S1	C3	P1	$40
S2	C1	P1	$90
S2	C1	P2	$50

Table 1: Fact Table for cube Sales

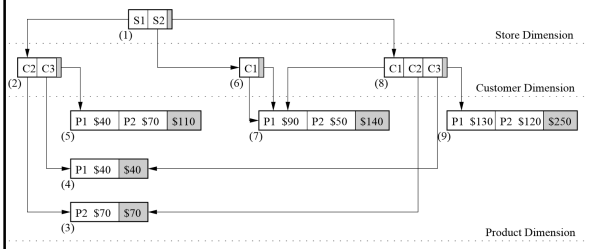


Figure 1: The Dwarf Cube for Table 1

Yannis Sismanis et al., 1992

Approche HOLAP

- Hybrid OLAP :
  - Stockage de l'entrepôt dans une base de données relationnelle
  - Stockage des données agrégées dans des cubes MOLAP
- Avantages
  - Bon compromis coût/performance sur de gros volumes de données
  - Exploite les fonctionnalités de SQL
  - Cube connecté à l'entrepôt relationnel

- Inconvénients
  - Difficulté de mise en œuvre
  - Pas aussi rapide que MOLAP
  - Passage à l'échelle moins facile qu'en ROLAP

Approche HTAP

Gartner 2014

- Hybrid Transaction / Analytical Processing :
  - SGBD en mémoire vive
  - Traitements OLTP et OLAP simultanés

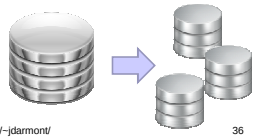
- Avantages
  - Calcul distribué rapide des requêtes
  - Pas de redondance des données
  - Informations transactionnelles rendues disponibles rapidement dans les modèles décisionnels
  - Unification des tables relationnelles et des modèles décisionnels
- Inconvénient
  - Modification drastique des architectures décisionnelles

Plan du cours

- ✓ Introduction : le processus décisionnel
- ✓ Modélisation conceptuelle des entrepôts
- ✓ Modélisation logique des entrepôts
- Mise en œuvre d'un entrepôt de données
- Analyse en ligne (OLAP)

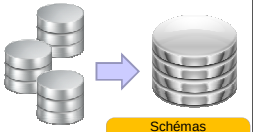
Approche top-down (Inmon)

- Conception intégrale de l'entrepôt *a priori*
  - Magasins de données (*datamarts*) extraits de l'entrepôt
- Avantages
  - Vision conceptuelle globale de l'entrepôt
  - Architecture intégrée
  - Normalisation des données, absence de redondance
- Inconvénients
  - Difficulté de mise en œuvre
  - Manque d'évolutivité



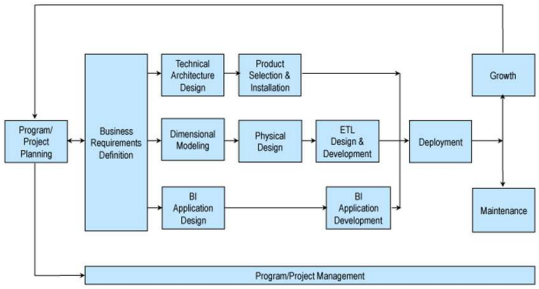
Approche bottom-up (Kimball)

- Construction incrémentale de l'entrepôt
  - L'entrepôt de données est une union de magasins de données
  - Notion de bus décisionnel et de dimensions conformes
- Avantages
  - Simplicité de mise en œuvre
  - Résultats rapides
- Inconvénient
  - Problèmes d'intégration des magasins de données



Schémas multidimensionnels, magasin de données

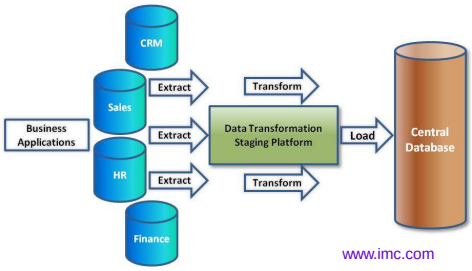
Cycle de vie d'un entrepôt



holowczak.com

Alimentation de l'entrepôt

Extract, Transform, Load



www.imc.com

Extraction

- Sources de données variées
  - Bases de données opérationnelles
  - Fichiers
  - Logs
  - Web…

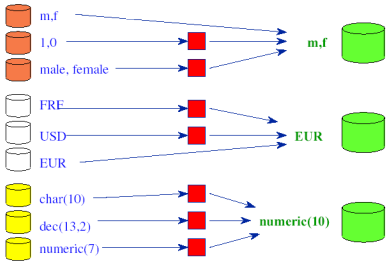
- Stratégies de rafraîchissement de l'entrepôt
  - Push : déclencheurs dans les sources
  - Pull : requêtage des sources
- Périodicité du rafraîchissement
  - Contrainte : ne pas perturber les opérations OLTP

Transformation

- Unification des données
  - Noms des attributs
  - Types (ex. précision numérique)
  - Formats (ex. dates)
  - Unités de mesure

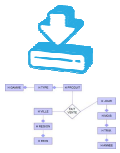
- Nettoyage des données
  - Vérification des contraintes d'intégrité
  - Suppression des doublons
  - Traitement des valeurs manquantes
  - Détection des valeurs erronées ou incohérentes

Transformation



Didier Donsez, Université Grenoble 1

Chargement



- Politiques de chargement
  - Complet / incrémental
  - En ligne / hors ligne
- Mises à jour des dimensions
  - Ecrasement de l'ancienne valeur
  - Versionnement
  - Traitement particulier des dimensions à évolution rapide
- Rafraîchissement des index et vues matérialisées
- Oubli des données anciennes
  - Suppression
  - Agrégation

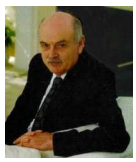
Plan du cours

- ✓ Introduction : le processus décisionnel
- ✓ Modélisation conceptuelle des entrepôts
- ✓ Modélisation logique des entrepôts
- ✓ Mise en œuvre d'un entrepôt de données
- Analyse en ligne (OLAP)

Algèbre OLAP

- OLAP = navigation interactive dans un cube de données
  « Spreadsheet on steroids » (T.B. Pedersen)
- Opérateurs ensemblistes
  - Projections et restrictions classiques
- Opérateurs de restructuration
  - Changement de point de vue
  - Réorientation selon les dimensions
- Opérateurs liés à la granularité
  - « Zoom » et « dézoom »

E.F. Codd



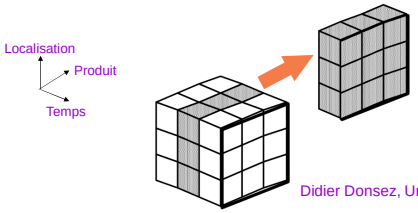
1923-2003

Slice (projection selon une dimension)

		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



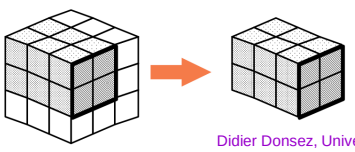
Didier Donsez, Université Grenoble 1

Dice (restriction)

		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

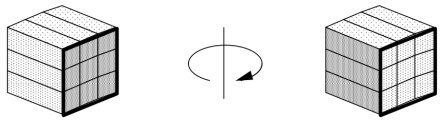


Didier Donsez, Université Grenoble 1

Rotate (pivot)

	95	96	97
Frais	221	263	139
Liquide	275	257	116

	95	96	97
NordPdC	101	120	52
IdF	395	400	203



Didier Donsez, Université Grenoble 1

Rollup/Drill down (forages)

Rollup

Rollup

	95	96	97
Alim.	496	520	255

Dimension Temps

	95-97
Frais	623
Liquide	648

	95	96	97
Frais	221	263	139
Liquide	275	257	116

	1S95	2S95	1S96	2S96	1S97
Frais	100	121	111	152	139
Liquide	134	141	120	137	116

Drill down

Dimension Produit

	95	96	97
Yaourt	20	19	22
...	...	...	...
Salade	40	43	48

Drill down

Didier Donsez, Université Grenoble 1

SQL *vs.* MDX

SQL	MDX
Table(s)	Cube
Attribut	Niveau hiérarchique
Attributs liés ou table(s) de dimension	Dimension
Attribut	Mesure
Valeur d'attribut dimension	Membre de dimension

SQL	MDX
SELECT attribut(s)	SELECT axe(s)
FROM table(s)	FROM cube
WHERE condition(s)	WHERE filtre(s)

SQL *vs.* MDX

SQL	MDX
SELECT COUNT(*) AS [Internet Order Quantity], SUM(SalesAmount) AS [Internet Sales Amount] FROM FactInternetSales	SELECT { [Measures].[Internet Order Quantity], [Measures].[Internet Sales Amount] } ON COLUMNS FROM [Internet Sales]

Results | Messages

	Internet Order Quantity	Internet Sales Amount
1	60398	29358677.22

Messages | Results

Internet Order Quantity	Internet Sales Amount
60,398	$29,358,677.22

www.mssqltips.com

SQL *vs.* MDX

SQL	MDX
SELECT d1.Education AS [Customer Education Level], COUNT(*) AS [Internet Order Quantity], SUM(f1.SalesAmount) AS [Internet Sales Amount]	SELECT { [Measures].[Internet Order Quantity], [Measures].[Internet Sales Amount] } ON COLUMNS, { ([Customer].[Education].[Education].MEMBERS) } ON ROWS
FROM FactInternetSales f1, DimCustomer d1 WHERE f1.CustomerKey = d1.CustomerKey	FROM [Internet Sales]
GROUP BY d1.EnglishEducation	

Results | Messages

	Customer Education Level	Internet Order Quantity	Internet Sales Amount
1	Bachelors	18144	9900142.76
2	Graduate Degree	10603	5460560.25
3	High School	10320	4638026.07
4	Partial College	16623	7723542.88
5	Partial High School	4708	1636405.26

Messages | Results

	Internet Order Quantity	Internet Sales Amount
Bachelors	18,144	$9,900,142.76
Graduate Degree	10,603	$5,460,560.25
High School	10,320	$4,638,026.07
Partial College	16,623	$7,723,542.88
Partial High School	4,708	$1,636,405.26

www.mssqltips.com

SQL *vs.* MDX

SQL	MDX
SELECT g.Country, sum(rs.OrderQuantity) as OrderQuantity	SELECT [Measures].[Reseller Order Quantity] ON ROWS, [Geography].[Country].[Country].MEMBERS ON COLUMNS FROM [Adventure Works])
FROM FactResellerSales rs, dimReseller r, dimGeography g WHERE r.ResellerKey = rs.ResellerKey AND g.GeographyKey = r.GeographyKey	FROM (
AND g.City IN ('Melbourne', 'Sydney', 'Seattle', 'New York')	SELECT {[Geography].[City].[Melbourne], [Geography].[City].[Sydney], [Geography].[City].[Seattle], [Geography].[City].[New York]} ON ROWS
GROUP BY g.Country	FROM [Reseller Sales])

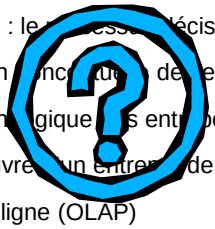
	Country	OrderQuantity
1	United States	3617
2	Australia	1217

	Reseller Order Quantity
Australia	1,217
United States	3,617

geekswithblogs.net/darrengosbell/

Plan du cours

- ✓ Introduction : le processus décisionnel
- ✓ Modélisation conceptuelle des entrepôts
- ✓ Modélisation logique des entrepôts
- ✓ Mise en œuvre d'un entrepôt de données
- ✓ Analyse en ligne (OLAP)

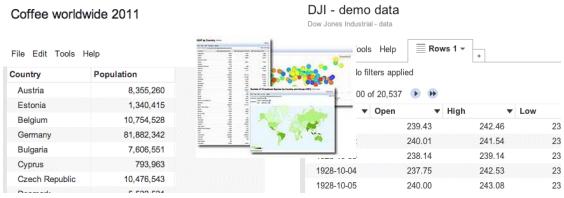


What next? Big data warehouses!

- Volume
  - Optimisation/parallélisation des agrégations
  - OLAP dans le nuage
- Variété
  - Nouveaux modèles multidimensionnels et opérateurs d'agrégation
  - Entrepôts NoSQL
- Vélocité
  - Travailler en mémoire : gare à l'explosion dimensionnelle
  - Fonctions d'oubli
- Véracité
  - Qualité des données sources
  - Sécurité des données entreposées

Du tableur collaboratif…

Google Fusion tables



…au décisionnel mobile

Utilisateur non expert

- Analyses à la demande
- Requêtage façon moteur de recherche
- *Drill beyond*
- Quasi temps réel
- Collaboratif

Technos

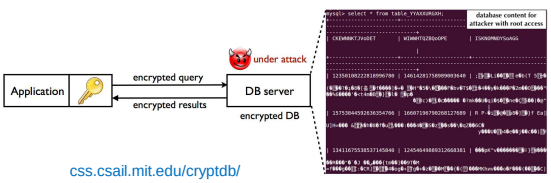
- Entrepôts de données / OLAP
- Web sémantique
- Gestion de documents
- Fouille de données
- Ingénierie sociale

Problèmes ouverts

Modèle de données avancé, algèbre de requêtes
Equilibre entre partage et confidentialité des données
Formalisation de l'intelligence collaborative, interfaces
Modèle économique

Confidentialité de données décisionnelles partagées dans le nuage

Travail basé sur CryptDB
Réalisé avec S. Sobati Moghadam et G. Gavin



css.csail.mit.edu/cryptdb/

BI4people : le décisionnel pour tous