

UNIVERSITÉ LUMIÈRE LYON 2

MINES Saint-Etienne

Entrepôts de données

Année 2015-2016
Jérôme Darmont

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

Actualités du cours

 http://eric.univ-lyon2.fr/~jdarmont/?page_id=2519

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

 https://twitter.com/darmont_lyon2 hashtag #edemse

Entrepôts de données <http://eric.univ-lyon2.fr/~jdarmont/> 1

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)

Entrepôts de données <http://eric.univ-lyon2.fr/~jdarmont/> 2

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 corrélés les données

Quelles questions ?

- Qui sont mes meilleurs clients ?
- Quelle est l'évolution du taux d'occupation des chambres ?
- Quelle est l'efficacité des politiques publiques en matière d'écologie ?

Entrepôts de données <http://eric.univ-lyon2.fr/~jdarmont/> 3

Problématique

- Données disponibles
 - Volumineuses
 - Hétérogènes
 - Très détaillées
- Traitement
 - Synthétiser/résumer
 - Visualiser
 - Analyser
- Utilisateurs
 - Non informaticiens
 - Non statisticiens

Entrepôts de données <http://eric.univ-lyon2.fr/~jdarmont/> 4

Système d'information décisionnel



- Customer Acquisition
- Customer Retention
- Process Optimization
- Loss Mitigation
- BI Strategy & Architecture
- EPM Dashboards & Scoreboards
- Query and Reporting
- Decision Support Systems
- Data Mining
- Web Integration
- Data Audit and Assessment
- Data Migration
- Data Conversion
- Data Enhancements
- Data Cleansing

© RHP Soft www.rhpssoft.com

Entrepôts de données <http://eric.univ-lyon2.fr/~jdarmont/> 5

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

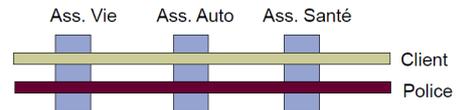
Entrepôts de données

<http://eric.univ-lyon2.fr/~jdarmon/>

6

Orientées sujet

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



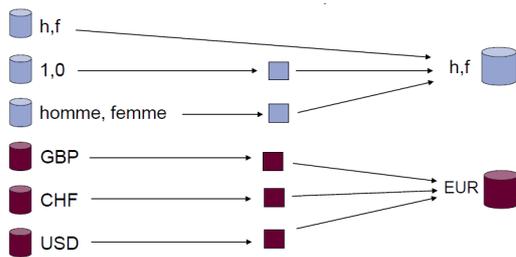
Lydie Soler, AgroParisTech/INRA

Entrepôts de données

<http://eric.univ-lyon2.fr/~jdarmon/>

7

Intégrées



Lydie Soler, AgroParisTech/INRA

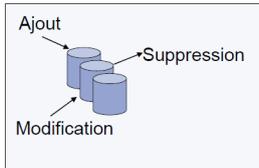
Entrepôts de données

<http://eric.univ-lyon2.fr/~jdarmon/>

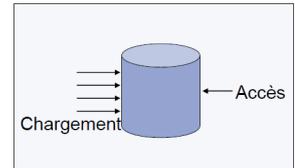
8

Non volatiles

Base de données opérationnelle



Entrepôt de données



Lydie Soler, AgroParisTech/INRA

On **L**ine **T**ransaction **P**rocessing vs. **O**n **L**ine **A**nalytical **P**rocessing

Entrepôts de données

<http://eric.univ-lyon2.fr/~jdarmon/>

9

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			Répertoire		
Code	Année	Mois	Code	Année	Mois
1	2005	Mai	1	Dupont	Paris
2	2006	Juillet	1	Durand	Lyon
			2	Dupont	Marseille

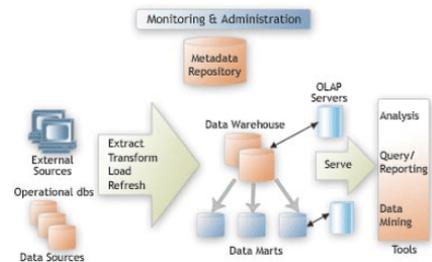
Lydie Soler, AgroParisTech/INRA

Entrepôts de données

<http://eric.univ-lyon2.fr/~jdarmon/>

10

Processus d'entreposage de données



www.kahassoc.com

Entrepôts de données

<http://eric.univ-lyon2.fr/~jdarmon/>

11

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)

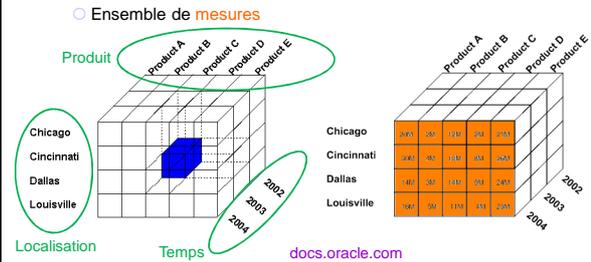
Entrepôts de données

<http://eric.univ-lyon2.fr/~jdamont/>

12

Métaphore du cube de données

- **Fait** : sujet d'analyse
- **Dimensions** : axes d'analyse

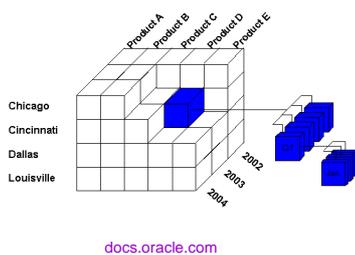


Entrepôts de données

<http://eric.univ-lyon2.fr/~jdamont/>

13

Hiérarchie de dimension

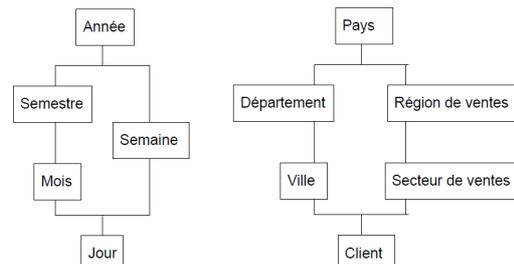


Entrepôts de données

<http://eric.univ-lyon2.fr/~jdamont/>

14

Hiérarchies multiples



Entrepôts de données

<http://eric.univ-lyon2.fr/~jdamont/>

15

Attributs de dimensions

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



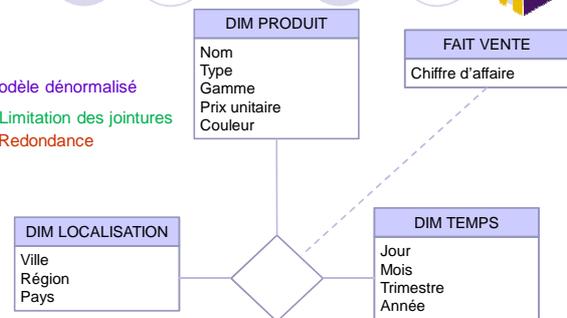
Entrepôts de données

<http://eric.univ-lyon2.fr/~jdamont/>

16

Schéma en étoile

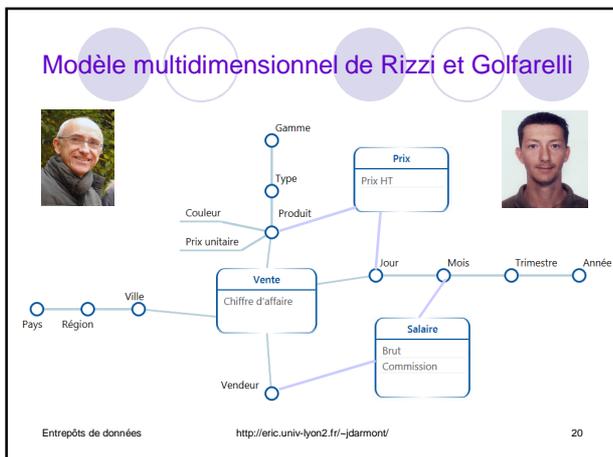
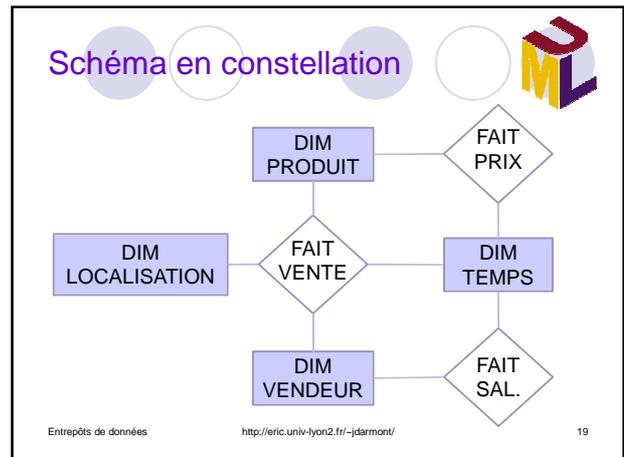
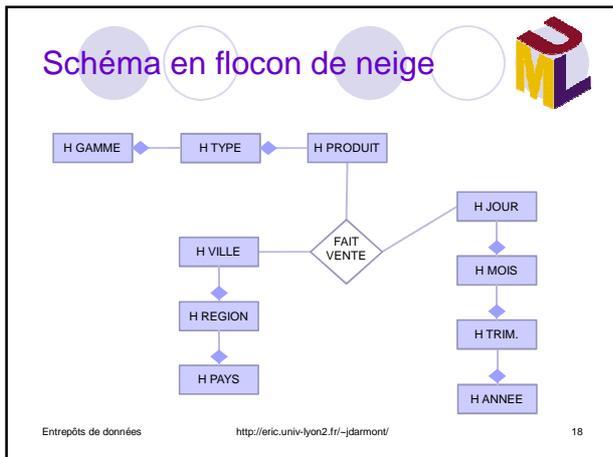
Modèle dénormalisé
 + Limitation des jointures
 - Redondance



Entrepôts de données

<http://eric.univ-lyon2.fr/~jdamont/>

17



- ### 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)
- Entrepôts de données <http://eric.univ-lyon2.fr/~jdarmon/> 21



- ### 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
-
- Entrepôts de données <http://eric.univ-lyon2.fr/~jdarmon/> 23

Étoile ROLAP

- DIM_PRODUIET(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)

Entrepôts de données <http://eric.univ-lyon2.fr/~jdarmon/> 24

Flocon ROLAP

- H_PRODUIET(IDproduit, Nom, PrixUnitaire, Couleur, IDtype#)
- H_TYPE(IDtype, NomType, IDgamme#) DIM_PRODUIET
- H_GAMME(IDgamme, NomGamme)
- H_VILLE(IDville, NomVille, IDrégion#)
- H_REGION(IDrégion, NomRégion, IDpays#) DIM_LOCALISATION
- H_PAYS(IDpays, NomPays)
- H_JOUR(IDjour, Jour, IDmois#)
- H_MOIS(IDmois, Mois, IDtrimestre#)
- H_TRIMESTRE(IDtrim, Trimestre, IDannée#)
- H_ANNEE(IDannée, Année) DIM_TEMPS
- FAIT_VENTE(IDproduit#, IDville#, IDjour#, Chiffre_affaire)

Entrepôts de données <http://eric.univ-lyon2.fr/~jdarmon/> 25

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	Jurassic 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

Entrepôts de données <http://eric.univ-lyon2.fr/~jdarmon/> 26

Index binaires et entrepôts

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

Entrepôts de données <http://eric.univ-lyon2.fr/~jdarmon/> 27

Opérations sur index binaires

- **Nombre de comédies**
Compter le nombre de 1 dans le bitmap associé
- **Nombre de comédies en 1995**
AND entre les deux bitmaps correspondants et comptage

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

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

Entrepôts de données <http://eric.univ-lyon2.fr/~jdarmon/> 28

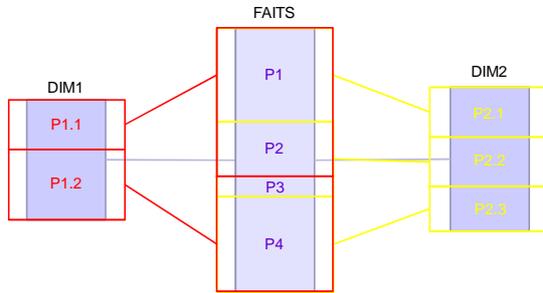
Vues matérialisées

Problématique : que matérialiser ?

Hector Garcia-Molina, Stanford

Entrepôts de données <http://eric.univ-lyon2.fr/~jdarmon/> 29

Fragmentation (horizontale dérivée)



Entrepôts de données <http://eric.univ-lyon2.fr/~jdarmon/> 30

Approche MOLAP

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

Dimensional view on gross revenues

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

- **Avantage**
 - Calculs d'agrégats rapides
- **Inconvénients**
 - Difficulté de mise en œuvre, systèmes majoritairement propriétaires
 - Volume de données limité
 - Problème d'éparité des cubes
 - Redondance des données avec l'entrepôt source
 - Rafraîchissement limité (reconstruction périodique complète)

Entrepôts de données <http://eric.univ-lyon2.fr/~jdarmon/> 31

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

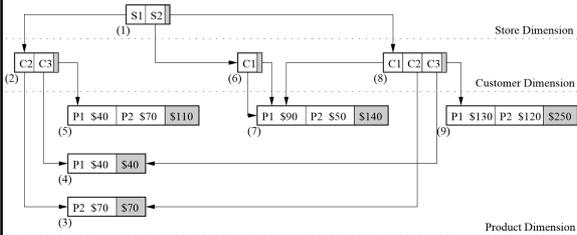


Figure 1: The Dwarf Cube for Table 1

Yannis Sismanis et al., 1992

Entrepôts de données <http://eric.univ-lyon2.fr/~jdarmon/> 32

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



Entrepôts de données <http://eric.univ-lyon2.fr/~jdarmon/> 33

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



Entrepôts de données <http://eric.univ-lyon2.fr/~jdarmon/> 34

Plan du cours

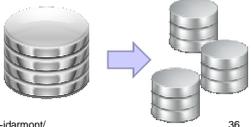
- ✓ Introduction : le processus décisionnel
- ✓ Modélisation conceptuelle des entrepôts
- ✓ Modélisation logique des entrepôts
- Mise en œuvre d'un entrepôt de données
- Analyse en ligne (OLAP)

Entrepôts de données <http://eric.univ-lyon2.fr/~jdarmon/> 35

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



Entrepôts de données <http://eric.univ-lyon2.fr/~jdamont/> 36

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

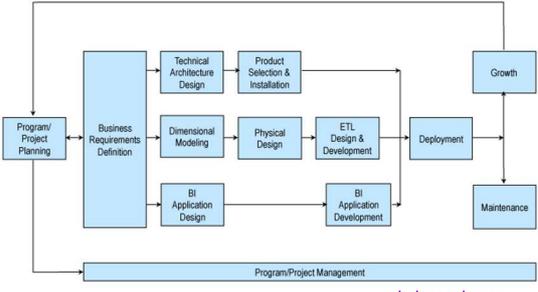


Schémas multidimensionnels, magasin de données



Entrepôts de données <http://eric.univ-lyon2.fr/~jdamont/> 37

Cycle de vie d'un entrepôt

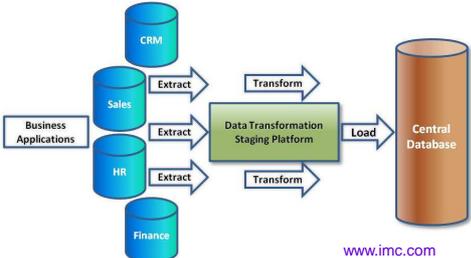


holowczak.com

Entrepôts de données <http://eric.univ-lyon2.fr/~jdamont/> 38

Alimentation de l'entrepôt

Extract, Transform, Load



www.imc.com

Entrepôts de données <http://eric.univ-lyon2.fr/~jdamont/> 39

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




Entrepôts de données <http://eric.univ-lyon2.fr/~jdamont/> 40

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




Entrepôts de données <http://eric.univ-lyon2.fr/~jdamont/> 41

Transformation

Didier Donsez, Université Grenoble 1

Entrepôts de données <http://eric.univ-lyon2.fr/~jdarmon/> 42

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

Didier Donsez, Université Grenoble 1

Entrepôts de données <http://eric.univ-lyon2.fr/~jdarmon/> 43

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)

Didier Donsez, Université Grenoble 1

Entrepôts de données <http://eric.univ-lyon2.fr/~jdarmon/> 44

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

Didier Donsez, Université Grenoble 1

Entrepôts de données <http://eric.univ-lyon2.fr/~jdarmon/> 45

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

Entrepôts de données <http://eric.univ-lyon2.fr/~jdarmon/> 46

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

Entrepôts de données <http://eric.univ-lyon2.fr/~jdarmon/> 47

Rotate (pivot)

	95	96	97		95	96	97
Frais	221	263	139	NordPdC	101	120	52
Liquide	275	257	116	IdF	395	400	203

Didier Donsez, Université Grenoble 1

Entrepôts de données <http://eric.univ-lyon2.fr/~jdamont/> 48

Rollup/Drill down (forages)

		95	96	97		1895	2895	1896	2896	1897			
Alim.	496	520	255	Frais	221	263	139	Liquide	100	121	111	152	139
Yaourt	20	19	22	Liquide	275	257	116	Yaourt	134	141	120	137	116
Salade	40	43	48										

Didier Donsez, Université Grenoble 1

Entrepôts de données <http://eric.univ-lyon2.fr/~jdamont/> 49

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)

www.mssqltips.com

Entrepôts de données <http://eric.univ-lyon2.fr/~jdamont/> 50

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]

www.mssqltips.com

Entrepôts de données <http://eric.univ-lyon2.fr/~jdamont/> 51

SQL vs. MDX

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

www.mssqltips.com

Entrepôts de données <http://eric.univ-lyon2.fr/~jdamont/> 52

SQL vs. MDX

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

geekswithblogs.net/darrogosbell/

Entrepôts de données <http://eric.univ-lyon2.fr/~jdamont/> 53

Plan du cours

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

Entrepôts de données <http://eric.univ-lyon2.fr/~jdamont/> 54

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

Entrepôts de données <http://eric.univ-lyon2.fr/~jdamont/> 55

Du tableur collaboratif...

Google Fusion tables

Coffee worldwide 2011

Country	Population
Austria	8,355,260
Estonia	1,340,415
Belgium	10,754,528
Germany	81,882,342
Bulgaria	7,606,551
Cyprus	793,963
Czech Republic	10,476,543

DJI - demo data

	Open	High	Low	
	239.43	242.46	23	
	240.01	241.54	23	
	238.14	239.14	23	
	1928-10-04	237.75	242.53	23
	1928-10-05	240.00	243.08	23

Entrepôts de données <http://eric.univ-lyon2.fr/~jdamont/> 56

...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

Entrepôts de données <http://eric.univ-lyon2.fr/~jdamont/> 57

Confidentialité de données décisionnelles partagées dans le nuage

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

Application → encrypted query → DB server → encrypted results

DB server (under attack) → encrypted DB

css.csail.mit.edu/cryptdb/

Entrepôts de données <http://eric.univ-lyon2.fr/~jdamont/> 58

BI4people : le décisionnel pour tous

Source documents → NoETL semantic integration / Automatic modeling → Datamart → Automatic visualization / Simple OLAP → 3D visualization

Partners: eric, IRIT, LI, elico, TRIMANE

Entrepôts de données <http://eric.univ-lyon2.fr/~jdamont/> 59