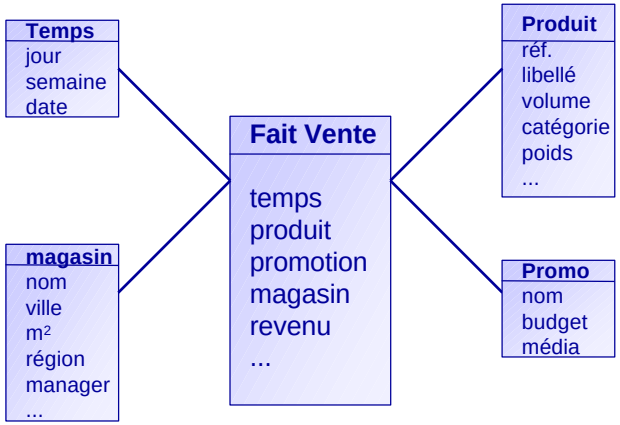
Le Data Warehouse



Plan

- Les concepts et l ’architecture
- Les techniques de modélisation
- L’alimentation du data warehouse
- Les systèmes OLAP
- L ’offre décisionnelle de Microsoft : SQL Server 7
- Les aspects économiques
- Conclusion et perspectives
- Références bibliographiques
- Glossaire
- Annexes

Le contexte

- Besoins
  - contexte de mondialisation
  - concurrence : l ’entreprise doit savoir anticiper
  - besoin d ’informations pertinentes
- Problème
  - données orientées production
- Première définition
  - Un data warehouse est un lieu de stockage intermédiaire des données issues des applications de production, dans lesquelles les utilisateurs finaux puisent avec des outils de restitution et d’analyse.

Les objectifs

- Accès aux informations
- Cohérence des informations
- Analyse multidimensionnelle
- Outils de requêtes, d ’analyse et de présentation
- Publication de données ayant déjà servi

- ***Le data warehouse ne peut remédier à la mauvaise qualité des données sources***

Quelques exemples

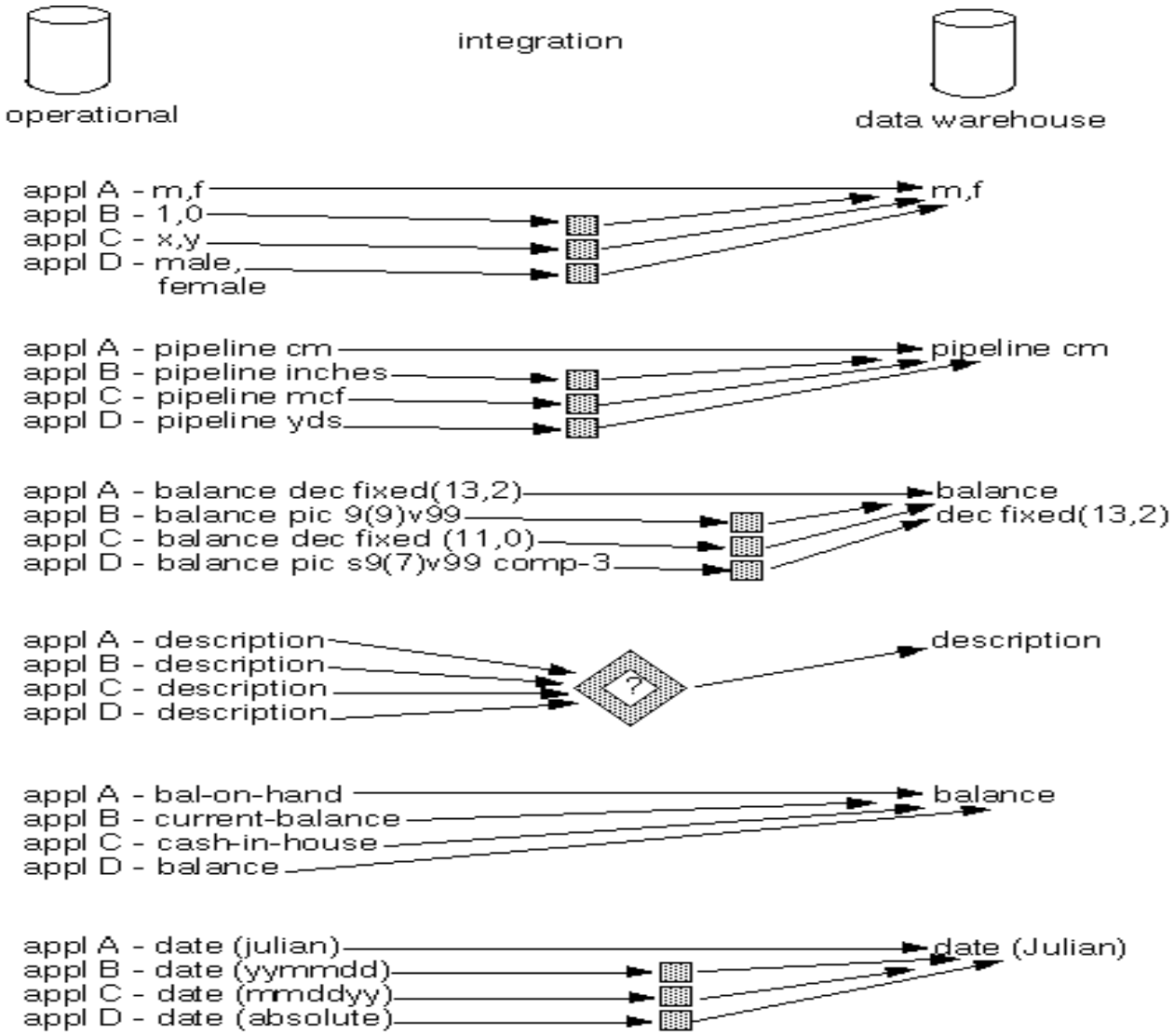
- Secteurs privilégiés
  - Automobile, télécommunications, distribution, assurances et banques …
- Exemples
  - Application d'aide à la décision chargé d'améliorer la rentabilité et la réactivité au Crédit Lyonnais
  - Système stratégique d'aide à la production mondiale chez un équipementier
  - Application de suivi commercial sur des postes nomades
  - Application de pilotage médico-économique dans un hôpital

Définition

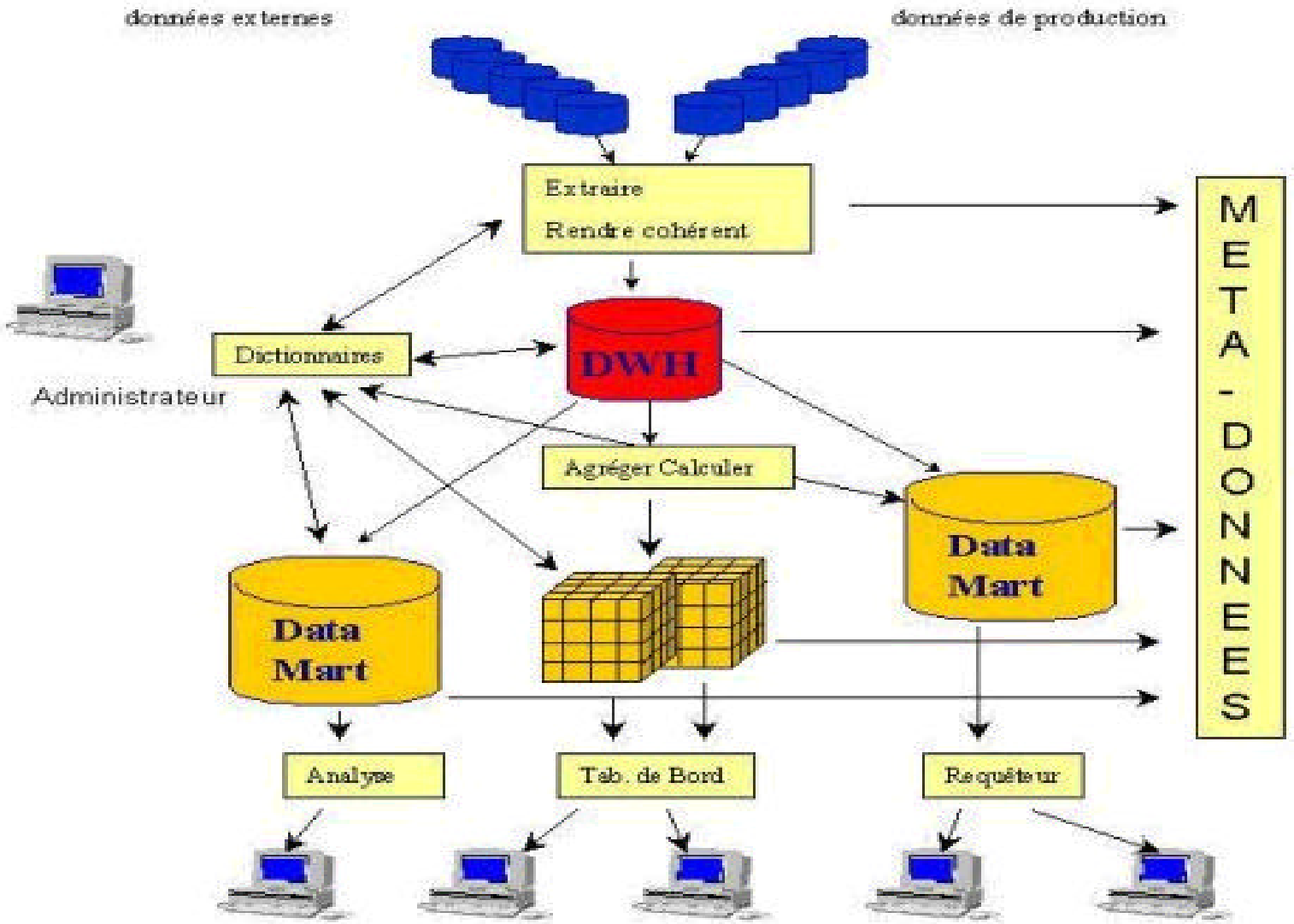
Bill Inmon

- Un data warehouse est une collection de données
  - orientées sujet
    - regroupées par centre d ’intérêt de l ’utilisateur
  - intégrées
    - cohérentes en terme de codage et de représentation
  - historisées
    - conservation de l ’historique des valeurs prises
  - non volatiles
    - une même requête exécutée à 2 moments différents fournira la même réponse
  - organisées pour le support d ’un processus d ’aide à la décision

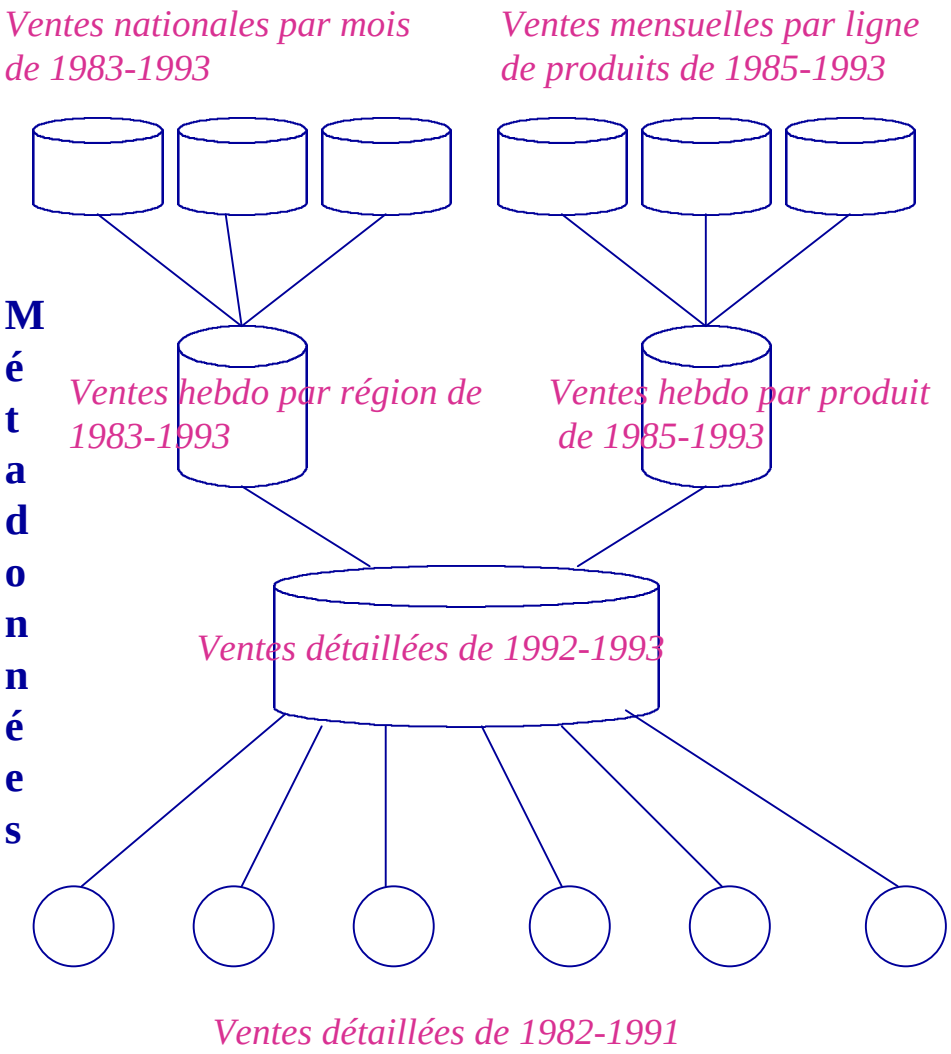
Données intégrées

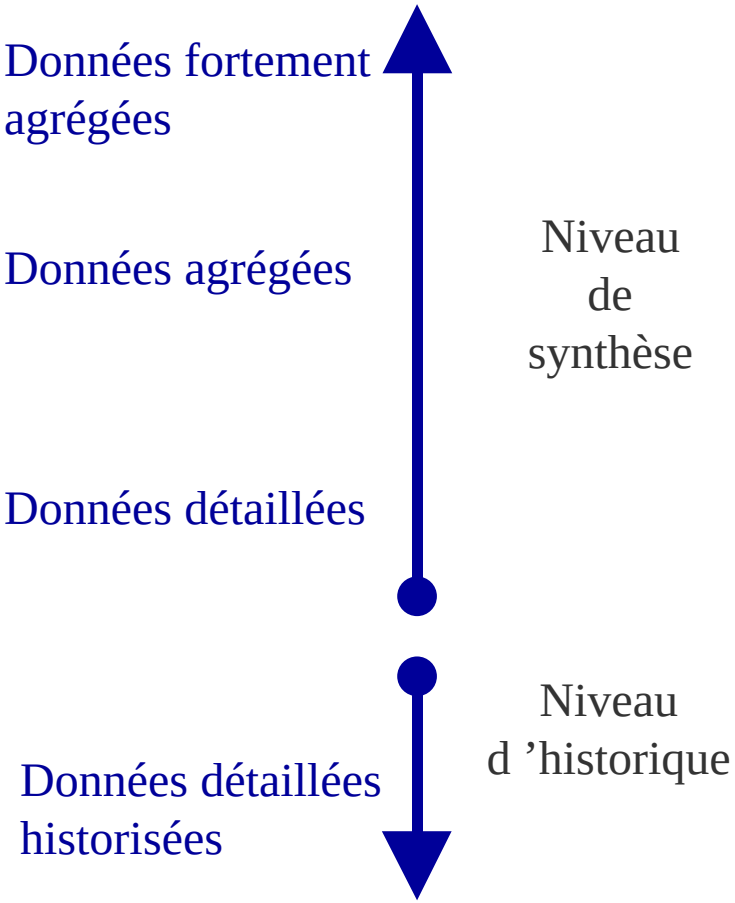


Le schéma de principe d'un data warehouse



Les types de données





Comparaison OLTP vs Décisionnel

	OLTP	Décisionnel
Données	détaillées	dérivées
Accès données	lectures, écritures	lecture seule
Requêtes	Simples, prévisibles	complexes, imprévisibles
Tps de réponse	<1sec, reproductible	secondes à minutes
Priorité	performance, fiabilité	souplesse, autonomie
Cohérence	microscopique	globale
Transactions	petites, nombreuses	grosses, 1 par jour

La modélisation des systèmes OLTP

- Basée sur le modèle entité-relation
- Hautes performances
  - Limitation de la redondance
  - Optimisation du modèle pour privilégier les requêtes fréquentes
- Orientation processus
- Faible lisibilité pour les utilisateurs finaux

La modélisation des systèmes décisionnels

- Lisibilité du point de vue de l'utilisateur
- Performance au chargement des données
- Performance à l'exécution des requêtes
- Facilitation de l 'administration du Data Warehouse
- Evolutivité

Le modèle relationnel

- Le modèle relationnel normalisé
  - Complet mais peu lisible
  - Certains indicateurs doivent être calculés à chaque requête
  - Requêtes complexes
  - Limité aux systèmes décisionnels simples

- Le modèle relationnel dénormalisé
  - Orienté besoins
  - Dénormalisation, agrégats
  - Moins complet mais plus lisible
  - Gain de performances relatif : tables volumineuses

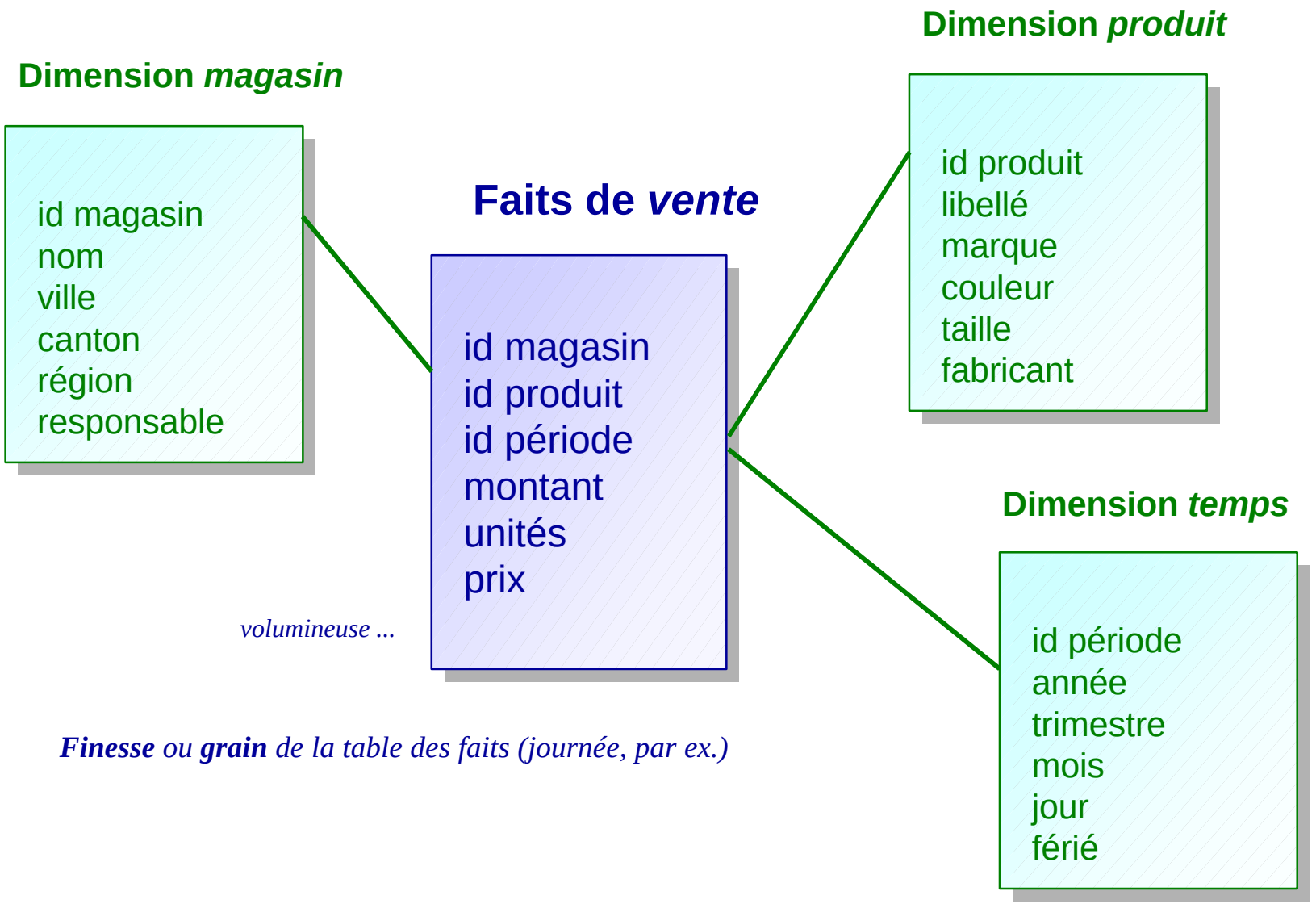
Le modèle dimensionnel

- Ensemble d ’indicateurs de performance (les faits) et d ’axes d ’étude (les dimensions)

- Exemple de faits
  - pour une entreprise : le chiffre d ’affaire
  - pour un hôpital : le taux d ’occupation des lits

- Exemple de dimensions
  - pour le chiffre d ’affaire : la période, le client, le produit
  - pour le taux d ’occupation des lits : la période, le service

Le schéma en étoile



volumineuse ...

***Finesse** ou **grain** de la table des faits (journée, par ex.)*

Les faits

- Numériques
- Valorisés de façon continue (*continuously valued*)
  - prenant une valeur à l ’intérieur d ’une grande fourchette
  - contrairement aux dimensions
- Additifs
  - pour synthétiser (additionner) de grandes masses de chiffres

- Identifiant de la table des faits
  - Clé multiple : concaténation des clés de chacune des dimensions d’analyse.

- Tables des faits éparses (*sparse*)
  - évitant les zéros signifiant « rien à signaler »

Les dimensions

Caractérisées par des attributs

- textuels
- discrets
  - propriétés constantes
  - contrairement aux faits
- utilisés comme contraintes (filtres) ou en-têtes dans les rapports

- *enjeux majeurs de la modélisation ...*

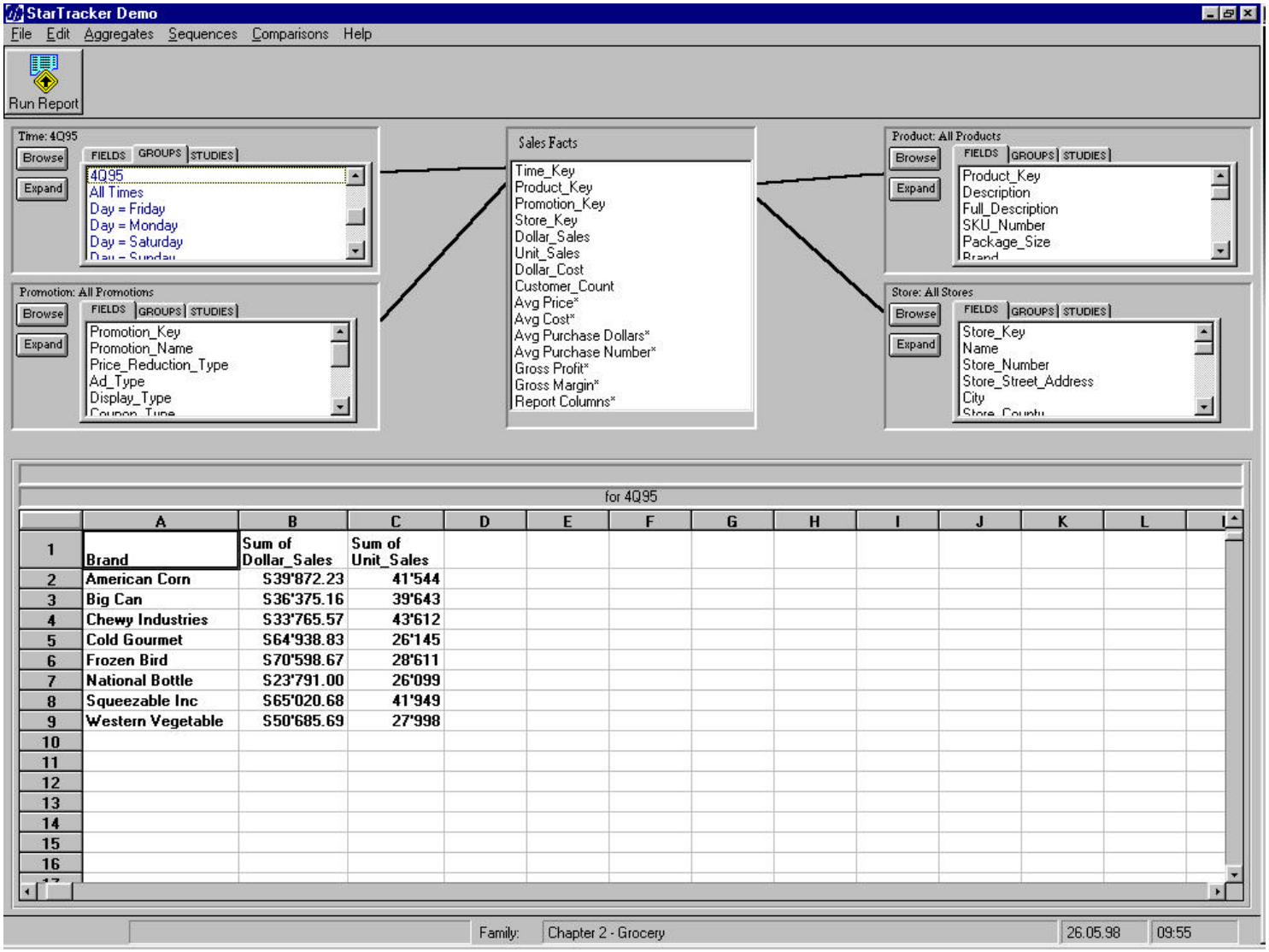
La dimension *Temps*

- ***le type SQL « date » standard est insuffisant***
  - *peut éventuellement être la clé de la table (et la clé étrangère dans la table des faits)*
- chaque jour peut être caractérisé par
  - le jour de la semaine (lundi, …)
  - le numéro du jour du mois (1, 2, …)
  - s 'il est le dernier jour du mois (O/N)
  - le numéro du jour (calendrier julien à partir d 'une date donnée)
  - le numéro de semaine dans l 'année (1, 2, … 52)
  - le numéro du mois (1, 2, …12)
  - le mois (janvier, février, …)
  - le trimestre (1er, 2ème, …)
  - la période fiscale (1Q98, 2Q98, …)
  - s 'il est férié ou non
  - la saison (printemps, été, …)
  - un événement (final de foot, …)

id temps

jourSemaine
noJourMois
dernJour
noJour
noSemaine
noMois
mois
trimestre
périodeFisc
férié
saison
événement

Exemple : Trouver toutes les marques de produits vendues au cours du mois d ’octobre 1995 et présenter le total des ventes et le total des unités vendues



SQL

```
SELECT [Time_SQL], [Time_KEY], [Time_TYPE] FROM [Time_GP] WHERE
   [Time_GRP_NAME] = '4Q95' ORDER BY [Time_KEY]
```

```
SELECT [Product].[Brand],
        Sum([Sales Fact].[Dollar_Sales]) as Col1,
        Sum([Sales Fact].[Unit_Sales]) as Col2
 FROM  [Sales Fact], [Time], [Product]
 WHERE [Sales Fact].[time_key]=[Time].[time_key]
 AND    [Sales Fact].[product_key]=[Product].[product_key]
 AND    [Time].[Fiscal_Period] IN ('4Q95')

 GROUP BY [Product].[Brand]
 ORDER BY [Product].[Brand]
```

colonnes

Jointures
fait & dimensions

Contrainte
filtre sur dimensions

rupture

L’alimentation du Data Warehouse (1/6)

- Deux modes de fonctionnement de l’entrepôt :
  - en ligne : exécution des requêtes des utilisateurs
  - hors ligne : alimentation
- Alimentation = migration et préparation des données provenant des systèmes opérationnels
- Etape très importante (80% du budget)
- Optimisation et automatisation :
  - Outils d’extraction-alimentation (ETL)
  - Configuration miroir

L’alimentation du Data Warehouse (2/6)

- La découverte des données
- L’extraction des données
- La transformation des données
- Le transfert des données

L’alimentation du Data Warehouse (3/6)

- La découverte des données : identifier les données utiles
  - Adresse complète ou code postal?
  - L’âge du client est-il nécessaire?

- L’extraction des données :
  - Difficulté liée à l ’hétérogénéité des données sources
  - Ne charger que les données modifiées ou crées :
    - Analyse des transactions des systèmes de production
    - Examens des fichiers opérationnels
    - Intérêt des outils d ’extraction : mécanisme de « changed data capture »

L’alimentation du Data Warehouse (4/6)

- La transformation des données : homogénéiser les données
  - L’épuration
    - Le filtrage des données aberrantes ou sans signification
    - Le dédoublonnage des données redondantes
  - Le formatage : conversion au format cible
  - La synchronisation des clés

- Le transfert des données
  - Le transfert de fichiers
  - Le transfert de base à base : plus lent, - sécurisé, transf. - cplexes

L’alimentation du Data Warehouse (5/6)

- Les outils d ’extraction et d ’alimentation (ETL) :
  - automatisation de l ’alimentation

- Les familles d ’ETL :
  - Première génération (1990) : les générateurs de code
    - Warehouse Manager de Prism Solutions
    - Passport de Carleton
    - ETI Extract d ’ETI : puissant, sophistiqué mais cher (>1 M de francs la licence)
  - Deuxième génération : les moteurs d ’extraction de données
    - Powermart d ’Informatica
    - Datastage d ’Ardent Software
    - Genio de Leonard's Logic : 300 000 francs
    - Inconvénient : présence de goulets d ’étranglement

L’alimentation du Data Warehouse (6/6)

- Les familles d ’ETL (suite) :
  - La troisième génération : les solutions globales
    - DTS compris dans SQL Server 7 (Microsoft)
    - Redbrick (Informix)
    - Les outils de data hub :
      - Constellar Hub (Constellar)
      - Paseo (Cimm Informatique)

Les systèmes OLAP (1/7)

- OLAP = OnLine Analytical Processing

- Système d ’analyse rapide d’information multidimensionnelle partagée
  - pas de programmation nécessaire
  - la plupart des réponses sont fournies en 5 secondes
  - toutes les données sont accessibles
  - vue multidimensionnelle : cubes ou pyramides
  - conditions de sécurité et de confidentialité jusqu’au niveau cellule du cube

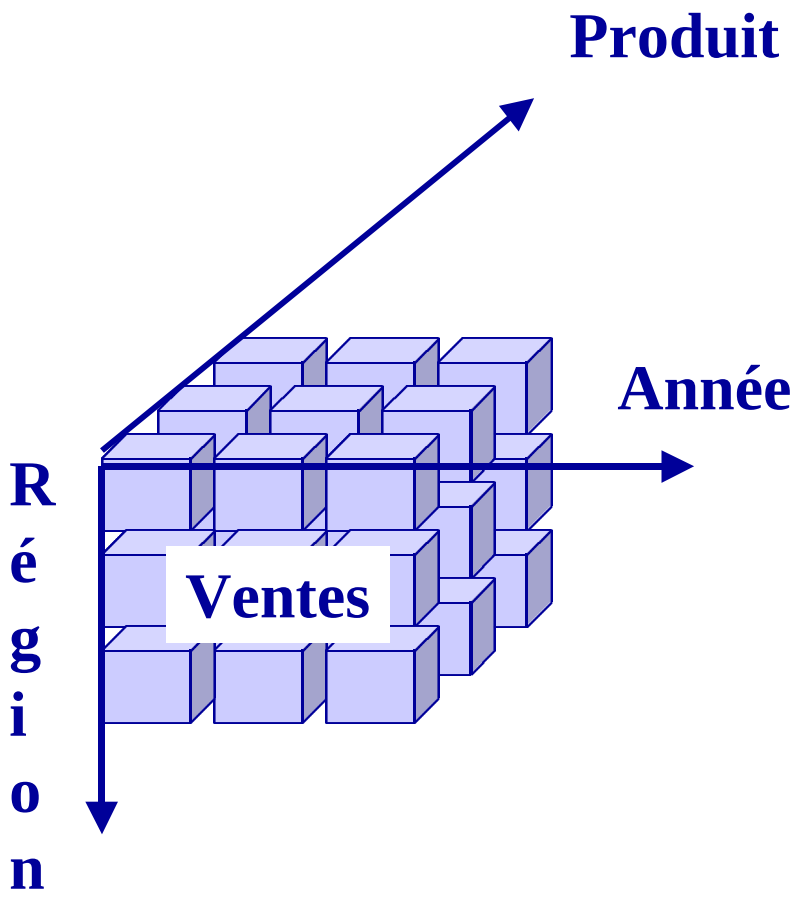
Les systèmes OLAP (2/7)

Dimension : axe d ’analyse

Cellule : indicateurs numériques

Drill-Down : analyse descendante

Slicing and Dicing : analyse transversale



Les systèmes dérivés d ’OLAP

Les systèmes OLAP (3/7)

- Le concept OLAP a fait de nombreux petits
- Les systèmes MOLAP (Multidimensional OLAP) : bases réellement multidimensionnelles.
- Les systèmes ROLAP (Relational OLAP) : bases relationnelles classiques organisées pour réagir comme une base OLAP.
- Les systèmes HOLAP (Hybride OLAP) : compromis entre les deux concepts précédents .
- Les systèmes DOLAP (Desktop OLAP) : bases OLAP hébergées sur le poste client.
- Etc.

Les outils MOLAP

Les systèmes OLAP (4/7)

- Base de données multidimensionnelle
- Avantages : performance
- Inconvénients :
  - Volumétrie importante en raison des agrégations systématiques
  - Volume maximum gérable : 10 Gigaoctets
  - Structure de stockage propriétaire
  - Modélisation préalable des besoins
- Marché :
  - Arbor (EssBase), Kenan (MultiWay), Oracle (expressServer)
  - Pilot (LightShip), Plaaning Science (Gentium)
  - TM1 Software (Table Manager)

Les outils ROLAP

Les systèmes OLAP (5/7)

- Base de données relationnelle
- Simulation d ’un SGBD multidimensionnel
- Pas d ’agrégation systématique
- Plus lent que MOLAP pour des petits volumes
- Supportent de très gros volumes
- Marché :
  - Informix (Metacube), Oracle (Discovere 2000)
  - Information Advantage (Axsys)
  - Micro Strategy (DSS Agent), Platinum (ProdeaBeacon)
  - if Solution (Star Tracker)

Les outils HOLAP

Les systèmes OLAP (6/7)

- Compromis entre MOLAP et ROLAP
- Données souvent consultées sont stockées dans une base MOLAP
- Les autres données sont stockées dans une base relationnelle
- Marché :
  - Oracle (Express)
  - SAS Institute (MDDB)
  - IBM (DB2 OLAP Server)
  - Holistic Systems (Holos)

Les outils DOLAP

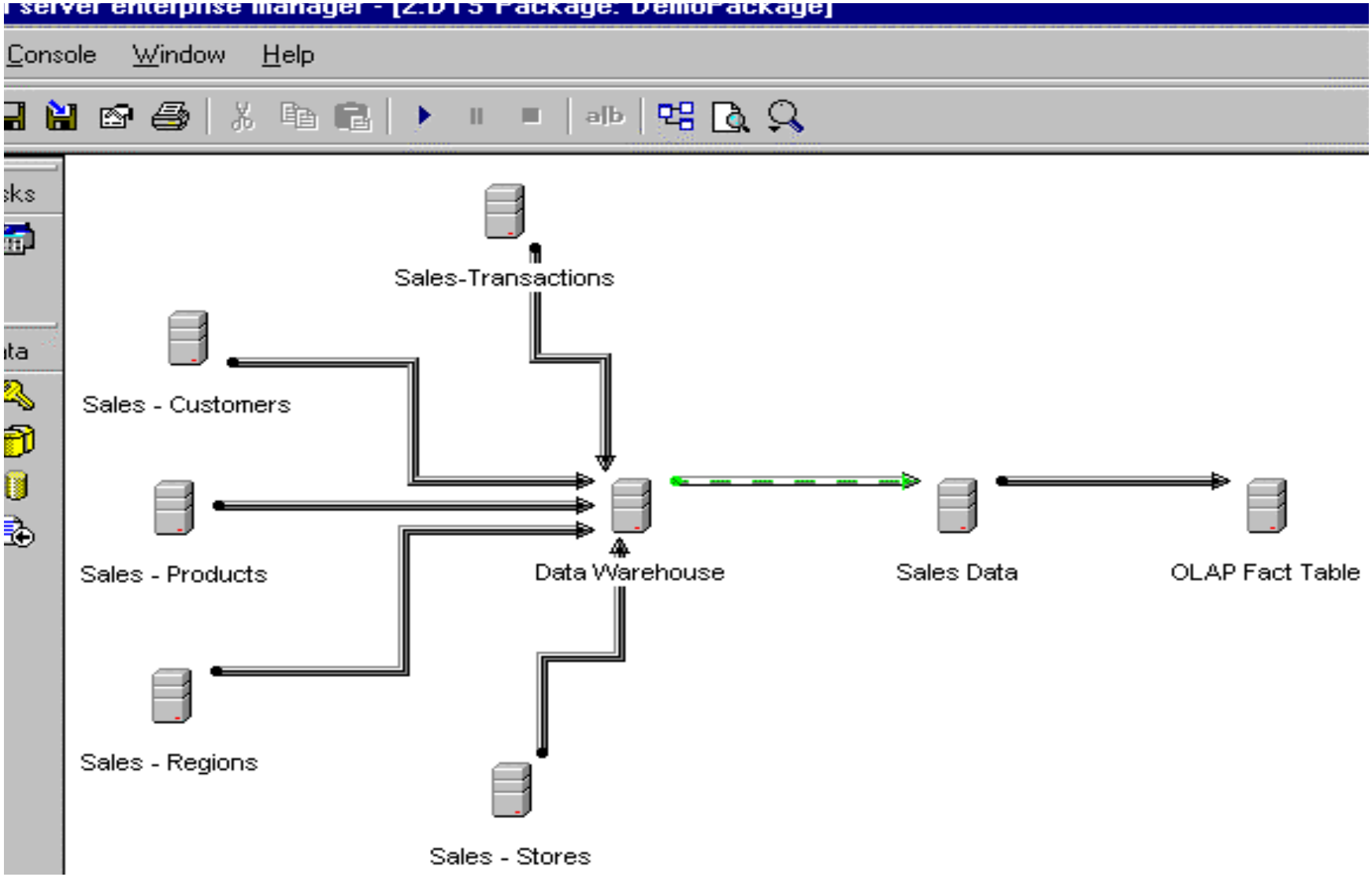
Les systèmes OLAP (7/7)

- OLAP de bureau
- Base OLAP hébergée sur le poste client
- Très rapide
- Marché :
  - Andyne (Pablo)
  - Business Object
  - Cognos (PowerPlay)
  - Dimensional Insight (Cross Traget)
  - Speedware (Media)

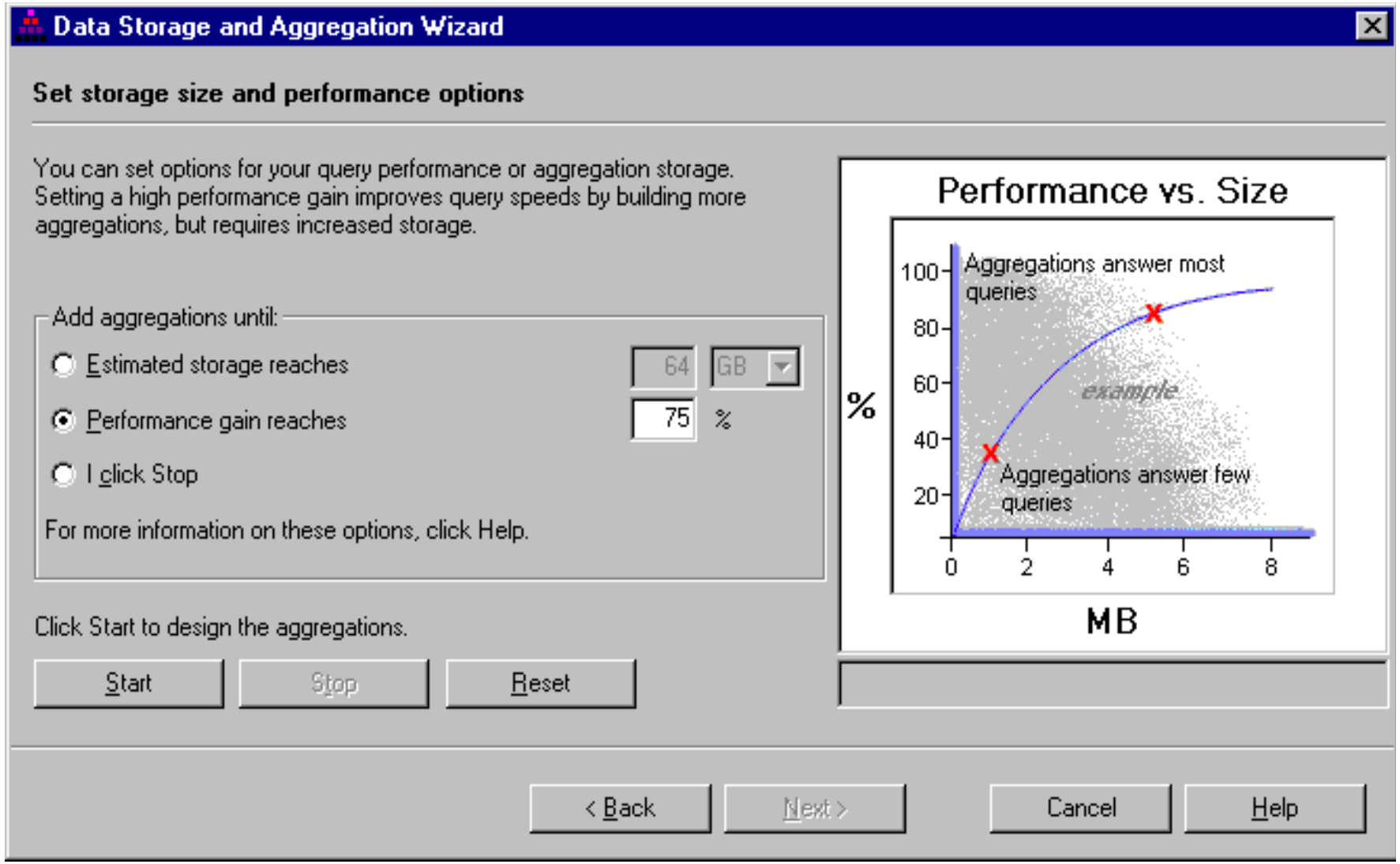
L ’offre décisionnelle de Microsoft

- Intérêts
  - diminution des coûts : démocratiser les technologies du DWH
  - intégration d ’outils d ’acquisition de données et analyse multidim.
  - standardisation des métadonnées
- SGBDR SQL SERVER 7
- Outil d ’extraction et d ’alimentation : DTS (Data Transformation Services)
- OLAP Services
- Les outils clients
  - Excel 2000 et outils tiers
  - Pivot Table Service : struct.intermédiaire de stockage permettant W en déconnecté
- Le référentiel : Microsoft Repository

SQL Server 7 : Data Transformation Services



SQL Server 7 : Optimisation de l ’agrégation



Les aspects économiques

- Investissement moyen : 2,2 millions de dollars
- Retour sur investissement à 3 ans
  - en moyenne = 400%
  - > 40% pour 90% des entreprises
  - >1000% pour 13% des entreprises
- Durée de rentabilisation
  - <3 ans pour 65% des entreprises
- Gains en information et en efficacité
  - Une augmentation des ventes grâce à un marketing mieux ciblé
  - Une amélioration des taux de rotation des produits
  - L'élimination des produits obsolètes
  - Une réduction des rabais et remises diverses
  - Une meilleure négociation au niveau des achats

Conclusion et perspectives

- Deux tendances actuelles
  - datamarts et dataweb

- La construction du data warehouse est un processus long et difficile.
- Construction progressive par datamarts
  - avantages : rapide
  - inconvénient : risque de cohabitation de datamarts incohérents

- Dataweb :
  - ouverture du data warehouse au web

Références

- Bibliographiques :
  - ***Le Data Warehouse - Le Data Mining*** de J.M. Franco et EDS-Institut Promothéus Ed. Eyrolles (juillet 1999)
  - ***La construction du Data Warehouse : du datamart au dataweb*** de J.F. Goglin Ed. Hermès (septembre 1998)
  - ***Entrpôts de données : Guide pratique du concepteur de « data warehouse »*** de R. Kimball Ed. International Thomson Publishing (janvier 1997)
- Presse informatique
  - ***SQL Server 7.0 de Microsoft : Le décisionnel en ligne de mire*** Informatiques Magazine (11 décembre 1998)
  - ***Datamart : mode d 'emploi*** Informatiques Magazine (10 janvier 1998)
  - ***Le Crédit Lyonnais allie web et décisionnel*** Informatiques Magazine (21 mai 1999)
  - ***Le data warehouse optimise le duty free*** Informatiques Magazine (1 juillet 1998)
  - ***Trois critères pour appréhender les solutions H-Olap et Rolap*** Informatiques Magazine (15 mai 1998)
  - ***Six solutions completes de data Web*** Informatiques Magazine (6 novembre 1998)

Références

- Documentation technique :
  - ***SQL Server 7.0 de Microsoft***
- Adresses Internet :
  - ***http://pwp.startenic.com/larryg/index.html*** : beaucoup de conseils intéressants pour construire un Data Warehouse
  - ***http://www.prismsolutions.com*** : serveur de la société de Bill Inmon, spécialisé sur la problématique de mise en œuvre du Data Warehouse

Glossaire (1)

- Agrégation
  - **Partitionnement horizontal d'une relation selon des valeurs d'attributs suivi d'un regroupement par une fonction de calcul (somme, moyenne, min,max,comptage)**
- Cluster
  - **Architecture matérielle permettant la coopération de plusieurs machines pour une instance de SGBD par partage des disques. Environnement à haute disponibilité**
- Data Mart
  - **Base de données orientée sujet mise à disposition des utilisateurs dans un contexte décisionnel décentralisé**
- Data Mining
  - **Ensemble des technologies avancées suceptibles d ’analyser l ’information d ’un Data Warehouse pour en tirer des tendances, pour segmenter l ’information, ou pour trouver des corrélations dans les données**

Glossaire (2)

- Data Warehouse
  - **«Entrepôt de données ». Base de données spécifique au monde décisionnel et destinée principalement à analyser les leviers « business » potentiels. Un Data Warehouse est** intégré, orienté sujet, et contient des données non volatiles
- Dimension
  - **Axe d'analyse associé aux indicateurs; correspond la plus souvent aux sujets d'intérêt du Data Warehouse**
- Drill down/Drill up
  - **Mécanisme permettant de se déplacer dans une structure multidimensionnelle, en allant du global vers le détail (***drill down***) ou vice versa (***drill up***)**
- EIS (Executive Information System)
  - **outils d ’aide à décision**
- Index
  - **Structure annexe pointant sur les données d ’une table à partir des valeurs d ’une colonne ou d ’un ensemble de colonnes de cette même table et utilisée pour accélérer la recherche des données**

Glossaire (3)

- Index binaire
  - **tableau de bits faisant correspondre les indices de valeur 1 aux lignes de la table contenant une valeur donnée pour la colonne indexée**
- Infocentre opérationnel
  - **Collection de données destinées à l ’aide à la décision orientées sujet, intégrées, volatiles, actuelles, organisées pour le support d'un processus de décision ponctuel en support à une activité particulière**
- Intégrité
  - **Ensemble de contraintes appliquées aux mises à jour d ’une base de données permettant de garantir leur cohérence**
- Jointure
  - **Rapprochement entre deux tables par comparaison de valeurs sur la base d'un attribut commun**
- Métabase
  - **Ensemble de tables systèmes utilisées par les SGBD pour stocker la description des objets utilisateurs (tables, vues, droits, procédures stockées…) d ’une base**

Glossaire (4)

- Métadonnée
  - **Information décrivant une donnée. Dans un contexte de Data Warehouse, elle qualifie une donnée précisant par ex. sa sémantique, les règles de gestion associées, sa source, son format...**
- Middleware
  - **Outil logiciel de connectivité. Dans un contexte décisionnel, il est situé entre les outils d'aide à la décision et la base de données décisionnelle. Un bon Middleware permet de conserver l'indépendance de ces deux types de composants**
- Modèle de données
  - **Schéma d ’une base. Le modèle décrit les tables, les attributs, les clés, les contarintes d ’intégrité. Le modèle relationnel décrit des tables à deux dimensions(ligne et colonne). Le modèle multidimensionnel ne limite pas le stockage des données dans l ’espace**
- Modèle en étoile
  - **Technique de modélisation dimensionnelle, consistant à distinguer physiquement les tables de faits des tables de dimensions. La table de faits est placée au centre du modèle, les tables de dimensions gravitant autour. Ce modèle représente visuellement un étoile**

Glossaire (5)

- Modèle en flocon
  - **Technique de modélisation dimensionnellle, dérivée de la modélisation en étoile, dont la représentation visuelle s ’apparente à un flocon. Dans ce modèle, les tables de dimensions sont dénormalisées, c.a.d. dénuées de redondances**
- Modèle dimensionnel (ou multidimensionnel)
  - **Technique de modélisation consistant à modéliser une base décisionnelle à partir de l'identification des faits à analyser et des dimensions d'analyses qui leur sont associées**
- Modèle relationnel
  - **Technique de modélisation consistant à décomposer une base de données en entité et en relations corrélant ces entités**
- MOLAP
  - **Multidimensional On Line Analytical Processing (voir OLAP)**
- Multidimensionnel (SGBD)
  - **Caractérise une base de données dédiée au décisionnel, stockant les données sous forme d ’un tableau multidimensionnel. Ces SGBD sont une alternative aux SGBD relationnels**

Glossaire (6)

- OLAP (On Line Analytical Processing)
  - **Caractérise l'architecture nécessaire à la mise en place d'un système d'information décisionnel. S'oppose à OLTP. Le terme OLAP désigne souvent les outils d'analyse s'appuyant sur des bases de données multidimensionnelles. On parle alors également d'outils MOLAP (pour les bases multidimensionnelles), par opposition aux outils ROLAP (pour les bases relationnelles)**
- OLTP (On Line Transactionnel Processing)
  - **Type d'environnement de traitement de l'information dans lequel une réponse doit être donnée dans un temps acceptable et consistant**
- Référentiel
  - **Structure de stockage des métadonnées. Un référentiel fédère ces métadonnées. On distingue le** *Data Warehouse Repository***, fédérant les métadonnées de la base décisionnelle, de** *l'Enterprise Repository***, qui inclut en théorie toutes les métadonnées de l'entreprise, aussi bien transactionnelles que décisionnelles**
- ROLAP (Relational On Line Analytical Processing)
  - **Architecture nécessaire à la mise en place d ’un système multidimensionnel en s ’appuyant sur les technologies relationnelles**

Glossaire (7)

- SGBD multidimensionnel
  - **Caractérise une base de données dédiée au décisionnel, stockant les données sous la forme d'un tableau multidimensionnel (cube). Ces SGBD sont une alternative aux SGBD relationnels. (voir aussi SIAD)**
- Structure multidimensionnelle
  - **stockant les données sous la forme d'un tableau multidimensionnel. Ces SGBD sont une Caractérise une base de données dédiée au décisionnel alternative aux SGBDR (voir SIAD également)**
- SIAD (Système Interactif d ’Aide à la Décision)
  - **Environnement permettant de stocker et de structurer l ’information décisionnelle. Ce terme désigne souvent les bases de données multidimentionnelles. L ’arrivée des concepts de Data Warehouse fait perdre de l ’importance à ce terme, qui fait fortement référence à une technologie spécifique**