Entrepôts de données et informatique désicionnelle

Nicolas Lachiche

18 septembre 2009
2 octobre 2009 de 🕘 à 🕘

Table des matières

1 Informatique décisionnelle **3**
1.1 Périmètre . . . 3
1.2 Environnement . . . 3
1.3 Fonctionnalités . . . 3

2 Entrepôts de données **4**
2.1 Architecture d'un système décisionnel . . . 4
2.1.1 Architecture à trois niveaux . . . 4
2.1.2 Sources . . . 4
2.1.3 Entrepôt . . . 4
2.1.4 Magasins (datamart) . . . 4
2.2 Constitution d'un entrepôt . . . 5
2.2.1 Analyse des besoins . . . 5
2.2.2 Modélisation de l'entrepôt . . . 5
2.2.3 Processus de modélisation dimensionnelle proposé par Kimball . . . 5
2.3 Modélisation multidimensionnelle des magasins . . . 6
2.3.1 Nécessité de modèles adaptés . . . 6
2.3.2 Niveau conceptuel . . . 6
2.3.3 Niveau logique . . . 7
2.3.4 Niveau physique . . . 7
2.4 Algèbre multidimensionnelle . . . 11
2.4.1 Table multidimensionnelle . . . 11

3 Fin de la partie retravaillée **11**

1 Informatique décisionnelle

1.1 Périmètre

- Une définition : "timely, accurate, high-value, and actionable business insights, and the work processes and technologies used to obtain them", Swain Scheeps, Business Intelligence for Dummies, Wiley Publishing, 2008.
- Outils d'aide à la décision (DSS)
- Executive/Management/Analysis Information/Decision SystemS

1.2 Environnement

- Entrepôts de données : rassemble les données et prépare les analyses, cf. prochain cours
- Progiciels de Gestion Intégrés (ERP) : unifie toutes les applications de l'entreprise
- Gestion de la relation client (CRM)
- E-commerce

1.3 Fonctionnalités

- Querying and reporting : facilités de consultation et de publication
- OLAP : Online Analytical Processing, tables multi-dimensionnelles, cf. prochain cours
- Tableaux de bords, Indicateurs de performance, Balanced Scorecards

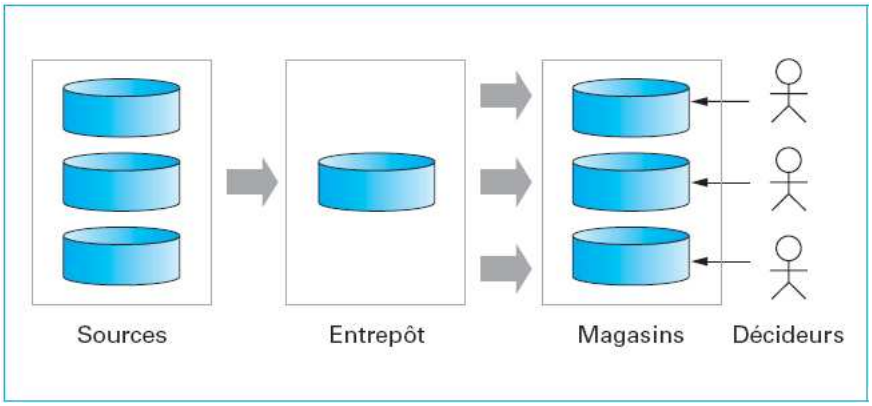

FIG. 1 –

2 Entrepôts de données

2.1 Architecture d'un système décisionnel

2.1.1 Architecture à trois niveaux

(cf figure 1 page 4)

2.1.2 Sources

- sources multiples hétérogènes (BD relationnelles, fichiers, sites web)
- autonomes, internes (BD production) ou externes (clients, fournisseurs, socio-économiques, officielles, etc.)

2.1.3 Entrepôt

- Véritable BD, en général relationnelle
- Alimentée périodiquement à partir des sources
- Données souvent historisées

Rarement utilisé directement par les décideurs :

- Contient plus que nécessaire pour une classe de décideurs
- Structure relationnelle peu adaptée à l'analyse de données

2.1.4 Magasins (datamart)

- Extrait d'un entrepôt
- Destiné à une seule classe de décideurs
- Organisé selon un modèle adapté aux données à analyser, en général multidimensionnel
- Outils de manipulation (requêteurs, tableurs, logiciels d'analyse et de fouille)

2.2 Constitution d'un entrepôt

- Méthodes utilisées pour les traitements transactionnels pas complètement adaptées : destinées à automatiser des traitements répétitifs
- Décisionnels : traitements peu répétitifs, évolutifs
- OLTP : boite noire / OLAP : décideurs accédent eux-mêmes aux données quand ils en ont besoin
- Evolution des besoins et des sources implique une démarche itérative
- Ne pas commencer par un entrepôt complet, mais par un projet pilote attractif et abordable

2.2.1 Analyse des besoins

1. Identifier des classes de décideurs
2. Pour chaque classe
   - Nature et fréquence des analyses effectuées
   - Données permettant ces analyses
   - Localisation des sources de ces données

2.2.2 Modélisation de l'entrepôt

L'entrepôt doit contenir l'ensemble des informations de pilotage extraites des différentes sources. Il est géré par les informaticiens qui en extraient des vues (virtuelles ou matérialisées) -les magasins- destinées aux différentes classes de décideurs.

- Généralement une BD relationnelle
- Normalisation pas nécessaire, car les sources ont déjà supprimé les problèmes liés à la redondance
- Souvent ajout d'une dimension temporelle, historique des données, implique un accroissement considérable du volume
- Choix des fréquences de rafraîchissement des données (global ou plus fin) en concertation avec les décideurs en connaissance du coût

2.2.3 Processus de modélisation dimensionnelle proposé par Kimball

1. Sélectionner le processus d'entreprise à modéliser
2. Déclarer le grain du processus : que représente une ligne de la table de faits ?
   - chaque ligne du ticket de caisse
   - le total de chaque ticket
   - le total des ventes hebdomadaires
   - etc...
3. Choisir les dimensions : comment les gestionnaires décrivent-ils des données qui résultent du processus concerné ?
4. Identifier les faits numériques de la table de faits : que mesurons-nous ?
   - quantité vendue
   - chiffre d'affaire
   - marge
   - etc...

2.3 Modélisation multidimensionnelle des magasins

2.3.1 Nécessité de modèles adaptés

- refléter la vision des analystes
- plusieurs axes : temps, localisation géographique, nomenclature des produits, etc.
- représentation relationnelle mal adaptée

Ventes 2005

Catégorie	Département	Montant
Papeterie	31	150
Papeterie	81	100
Papeterie	82	100
Micro	31	250
Micro	81	200

- Tableau croisé à deux dimensions plus adaptés

2005

	Papeterie	Micro
31	150	250
81	100	200
82	100	

- voire plus de deux dimensions

2005	Papeterie	Micro
31	150	250
81	100	200
82	100	

2004	Papeterie	Micro
31	150	250
81	100	200
82	100	

2003	Papeterie	Micro
31	150	250
81	100	200
82	100	

- Modélisation multidimensionnelle : sujet analysé (fait) comme point d'un espace à plusieurs dimensions. On parle de cube de données (cf figure 2 page 7)

2.3.2 Niveau conceptuel

Description de la base multidimensionnelle indépendamment des choix d'implantation

Fait – Sujet analysé
- un nom
- un ensemble d'attributs appelés mesures ou indicateurs

Dimensions – Axes d'analyse
- Géographique, temporel, produits, etc.
- Chaque dimension comporte un ou plusieurs attributs

Hiérarchie – Attributs d'une dimension organisés suivant des hiérarchies
- Dimension temporelle : jour, mois, année

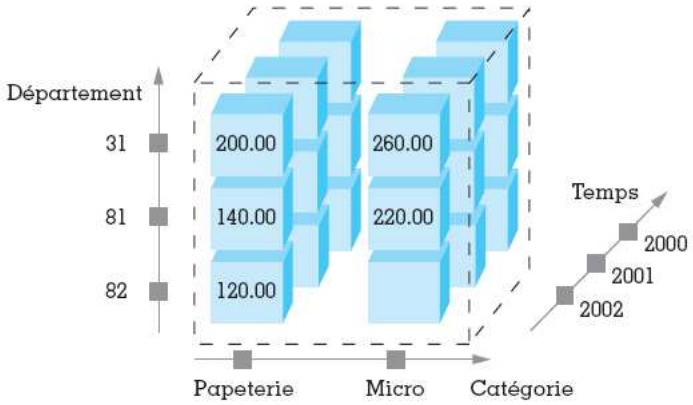

FIG. 2 –

- Dimension géographique : magasin, ville, région, pays
- Dimension produit : produit, catégorie, marque, etc.
- Attributs définissant les niveaux de granularité sont appelés paramètres
- Attributs informationnels liés à un paramètre sont dits attributs faibles

Schéma en étoile (cf figure 3 page 8)

Schéma en constellation (cf figure 4 page 8)

2.3.3 Niveau logique

Description de la base multidimensionnelle suivant la technologie utilisée :
- ROLAP
- MOLAP
- HOLAP

ROLAP – pour chaque dimension, une relation de même nom dont les attributs sont dérivés des paramètres et attributs faibles de la dimension, dont la clé primaire correspond à la granularité la plus fine
- pour chaque fait, une relation de même nom dont les attributs représentent les mesures et des clés étrangères référençant les dimensions liées au fait, dont la clé est la combinaison des clés étrangères ou une clé artificielle

(cf figure 5 page 9)

Schéma en flocon Normalisation des tables de dimension, fait apparaître explicitement les hiérarchies (cf figure 6 page 9)

2.3.4 Niveau physique

Implantation suivant le logiciel utilisé : Oracle 9i

Insuffisance des instructions classiques de SQL – `CREATE TABLE ... AS ...` : recopie physique, à reprendre intégralement lors de l'évolution des sources
- `CREATE VIEW ... AS ...` : recalculé à chaque requête, temps de réponse inacceptable sur les volumes manipulés

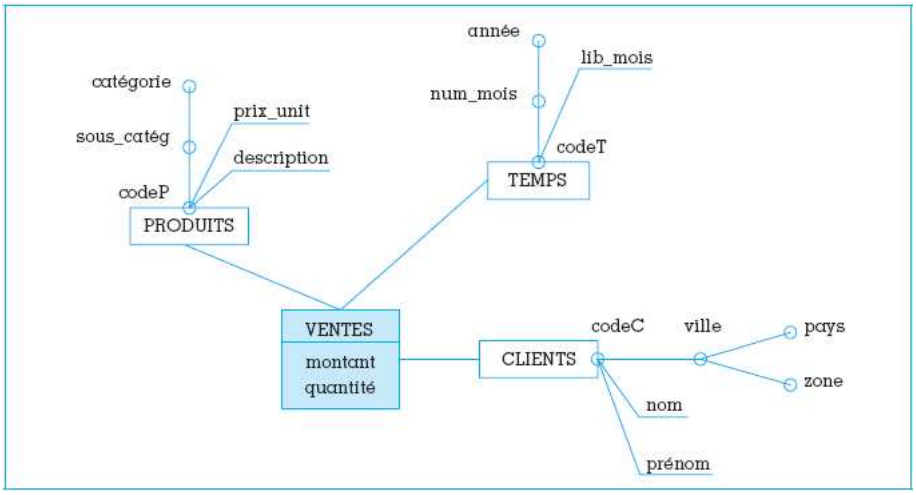
année
lib_mois
catégorie
num_mois
prix_unit
sous_catég
codeT
description
TEMPS
codeP
PRODUITS
VENTES
montant
quantité
codeC
ville
pays
CLIENTS
zone
nom
prénom

FIG. 3 –

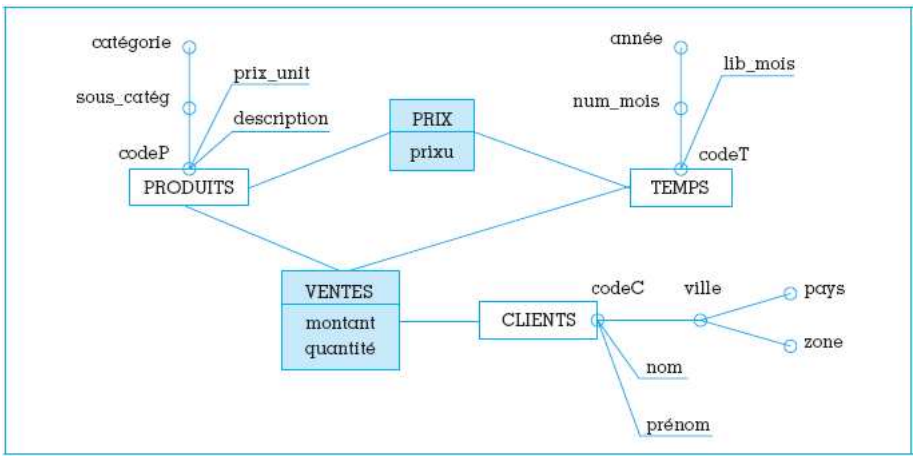
catégorie
année
prix_unit
lib_mois
sous_catég
num_mois
description
PRIX
prixu
codeP
codeT
PRODUITS
TEMPS
VENTES
montant
quantité
codeC
ville
pays
CLIENTS
zone
nom
prénom

FIG. 4 –

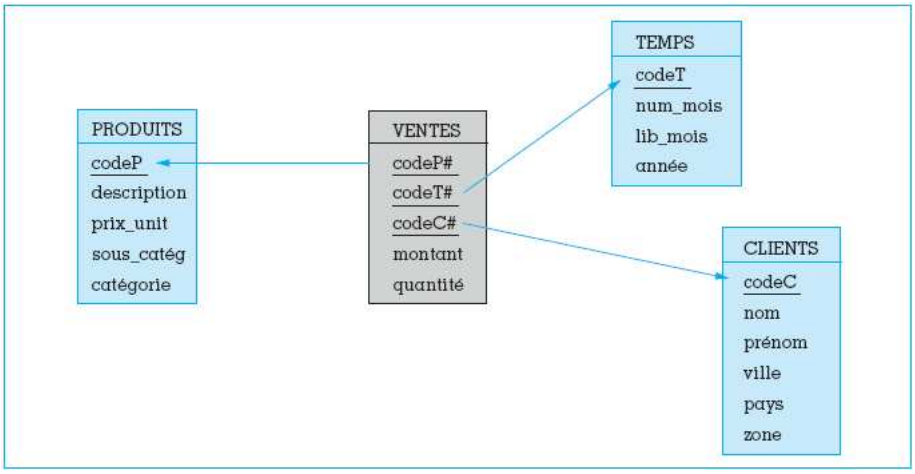
TEMPS
codeT
num_mois
lib_mois
année
PRODUITS
codeP
description
prix_unit
sous_catég
catégorie
VENTES
codeP#
codeT#
codeC#
montant
quantité
CLIENTS
codeC
nom
prénom
ville
pays
zone

FIG. 5 –

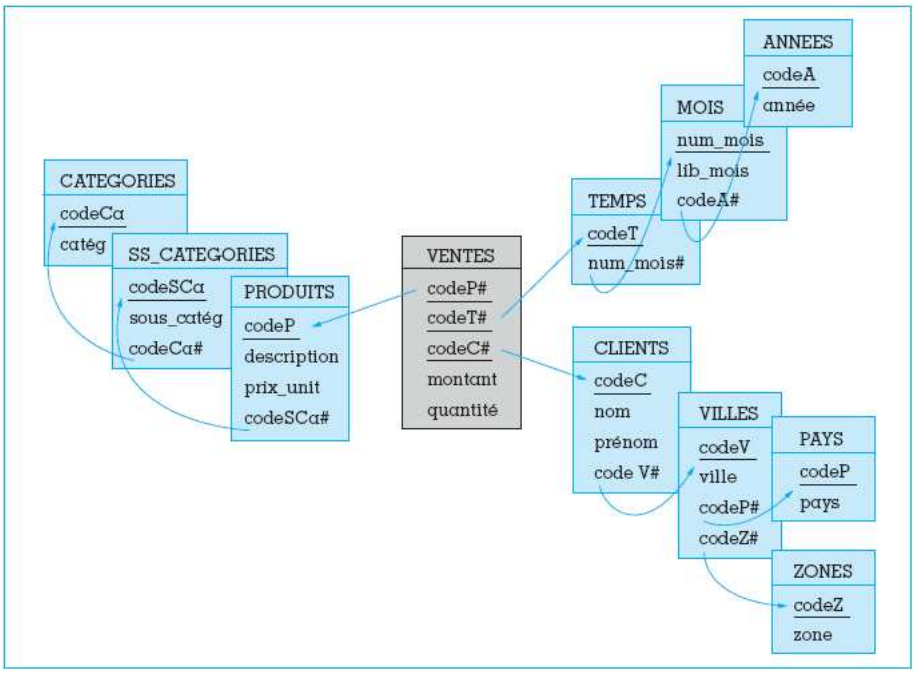
ANNEES
codeA
année
MOIS
num_mois
lib_mois
codeA#
CATEGORIES
codeCa
catég
SS_CATEGORIES
codeSCa
sous_catég
codeCa#
PRODUITS
codeP
description
prix_unit
codeSCa#
TEMPS
codeT
num_mois#
VENTES
codeP#
codeT#
codeC#
montant
quantité
CLIENTS
codeC
nom
prénom
code V#
VILLES
codeV
ville
codeP#
codeZ#
PAYS
codeP
pays
ZONES
codeZ
zone

FIG. 6 –

Vues matérialisées `CREATE MATERIALISED VIEW`

```
BUILD {IMMEDIATE|DEFERRED}
REFRESH {COMPLETE|FAST|FORCE|NEVER} {ON COMMIT|ON DEMAND}
AS SELECT ...;
```

- `IMMEDIATE` : ...
- `DEFERRED` : remplissage ultérieur par `DBMS_MVIEW.REFRESH ()`
- `COMPLETE` : recalcul complet de la vue
- `FAST` : rafraîchissement incrémental
- `FORCE` : `FAST` si possible, sinon `COMPLETE`
- `NEVER` : aucun rafraîchissement
- `ON COMMIT` : rafraîchissement à la fin de chaque transaction modifiant les tables sources
- `ON DEMAND` : par `DBMS_MVIEW.REFRESH`

Dimensions `CREATE DIMENSION <nom_dimension>`

```
  LEVEL <niveau1> IS (<nom_table.nom_attribut1>)
  LEVEL <niveau2> IS (<nom_table.nom_attribut2>)
...
WITH HIERARCHY <nom_hierarchie1>
(
  <niveau_i1> CHILD OF <niveau_i2> CHILD OF ...
)
WITH HIERARCHY <nom_hierarchie2>
(
    <niveau_j1> CHILD OF <niveau_j2> CHILD OF ...
)
...
ATTRIBUTE <niveau_k1> DETERMINES <nom_attribut_x>
ATTRIBUTE <niveau_k2> DETERMINES <nom_attribut_y>
...
```

2.4 Algèbre multidimensionnelle

2.4.1 Table multidimensionnelle

- Présente les valeurs des mesures d'un fait
- en fonction des valeurs des paramètres des dimensions représentées en lignes et en colonnes
- étant données des valeurs des autres dimensions
- correspond à une tranche du cube multidimensionnel

```
TD = (<fait>,{<mesure1>,<mesure2>,...},
  (<dimension1>,{<hierarchie1>,<hierarchie2>,...},
  {<paramètre1>,<paramètre2>,...},{{v1_p1,v2_p1,...},
  {v1_p2,v2_p2,...},...}),
  (<dimension2>,...),
  <predicat_selection>)
```

3 Fin de la partie retravaillée

Exemple

`TD = (Ventes,{montant}, (Temps,{h_annee},{annee},{{2005,2004,2003}}), (Clients,{h_clie`

2005 200 150 300 2004 250 240 260 2003 200 210 220 Opérateurs de transformation de la granularité des données

* Forage o DrillDown o RollUp * Calcul o Cube o Uncube

Forages DrillDown(TD,D,p)

* augmente la granularité des données * forage vers le bas sur la hiérarchie courante de la dimension D jusqu'au paramètre p * ajout d'une ou plusieurs lignes ou colonnes

RollUp(TD,D,p)

* diminue la granularité des données * forage vers le haut sur la hiérarchie courante de la dimension D jusqu'au paramètre p * suppression d'une ou plusieurs lignes ou colonnes

TDVentes1 Allemagne France Espagne 2005 200 150 300 2004 250 240 260 2003 200 210 220 = RollUp(TDVentes2,Clients,pays) TDVentes2 Berlin Hambourg Paris Toulouse Madrid 2005 150 50 100 50 300 2004 160 90 100 140 260 2003 100 100 110 100 220 = DrillDown(TDVentes1,Clients,ville) Calculs Cube(TD,Fnct)

* agrégation par ligne et par colonne * ajout d'une ligne et d'une colonne totaux

UnCube(TD)

* opération inverse * supprime la ligne et la colonne totaux

TDVentes1 Allemagne France Espagne 2005 200 150 300 2004 250 240 260 2003 200 210 220 = UnCube(TDVentes3) TDVentes3 Allemagne France Espagne Total 2005 200 150 300 650 2004 250 240 260 750 2003 200 210 220 630 Total 650 600 780 2030 = Cube(TDVentes1,SUM) Opérateurs de transformation de la structure des données

* Rotation o FRotate o DRotate o HRotate * Permutation o Switch o Order o Nest * Transformation o Push o Pull * Classique o Select o AddM o DelM

Rotations

* changement de sujet d'analyse (rotation de faits) * changement d'axes d'analyse (rotation de dimensions) * changement de vues d'une même dimension (rotation de hiérarchie)

FRotate(TD,F)

* rotation du fait courant pour visualiser les mesures du fait F * partage au minimum les deux dimensions visualisées

DRotate(TD,D1,D2[,Hi])

* rotation de la dimension D1 avec la dimension D2 * précision optionnelle de la hiérarchie à utiliser sur la dimension D2 * positionnement sur le paramètre de granularité maximale de la nouvelle hiérarchie

HRotate(TD,D,H1,H2)

* rotation de la hiérarchie H1 avec la hiérarchie H2 pour la visualisation suivant la dimension D

TDVentes1 Allemagne France Espagne 2005 200 150 300 2004 250 240 260 2003 200 210 220 TDVentes4 Nord Sud Est Ouest 2005 100 120 200 180 2004 230 210 200 140 2003 180 150 190 160 Permutations Switch(TD,D,p,v1,v2)

* permutation des valeurs v1 et v2 du paramètre p avec répercussion sur les valeurs des paramètres de granularité inférieure

Order(TD,D,p,s)

* tri des valeurs du paramètre p * s=ASC/DESC

Nest(TD,D,p1,p2)

* permutation des paramètres p1 et p2 sur la hiérachie courante * le paramètre p1 est imbriqué dans le paramètre p2 *

TDVentes1 Allemagne France Espagne 2005 200 150 300 2004 250 240 260 2003 200 210 220 = Switch(Rollup(TDVentes5,Temps,annee), Clients,pays,France,Allemagne) TDVentes5 Allemagne France Espagne Janvier 2005 50 35 50 2004 40 40 50 2003 40 35 45 Fevrier 2005 25 15 40 2004 20 20 20 2003 15 20 25 ... = Nest(Switch(DrillDown(TDVentes1,Temps,mois), Clients,pays,Allemagne,France), Temps,annee,mois) Transformations Push(TD,D,p)

* conversion du paramètre p en mesure dans le fait courant * p ne doit pas être le paramètre de plus bas niveau! * D est affichée avec au moins deux paramètres

Pull(TD,D,m)

* conversion de la mesure m du fait courant en paramètre de la dimension courante D * le nouveau paramètre est positionné comme granularité minimale des paramètres affichés * le fait courant est visualisé avec au moins deux mesures

TDVentes6 Allemagne France Espagne Berlin Hambourg Paris Toulouse Madrid 2005 150 50 100 50 300 2004 160 90 100 140 260 2003 100 100 110 100 220 = Nest(Pull(TDVentes7,Clients,pays), Clients,ville,pays) TDVentes7 Berlin Hambourg Paris Toulouse Madrid 2005 150, Allemagne 50, Allemagne 100, France 50, France 300, Espagne 2004 160, Allemagne 90, Allemagne 100, France 140, France 260, Espagne 2003 100, Allemagne 100, Allemagne 110, France 100, France 220, Espagne = Push(TDVentes6,Clients,pays) Opérateurs classiques Select(TD,predicat)

* restriction sur les valeurs restituées * le prédicat de sélection porte sur les dimensions et/ou sur le fait

AddM(TD,m) / DelM(TD,m)

* Ajout/suppression de la visualisation d'une mesure

TDVentes1 Allemagne France Espagne 2005 200 150 300 2004 250 240 260 2003 200 210 220

TDVentes8 Allemagne 2004 250, 1000