Conception de bases de données

Introduction aux data warehouses : la modélisation en étoile

dwh1.pdf

utc Formation

Stéphane Crozat

7 juin 2017

Table des matières

Cours

I

A. Data warehouses et bases de données décisionnelles

Objectifs

Comprendre ce qu'est et à quoi sert un *data warehouse*.
Comprendre les différences entre un *data warehouse* et une base de données transactionnelle.

1. Décisionnel

Définition

« *Le système d'information décisionnel est un ensemble de données organisées de façon spécifiques, facilement accessibles et appropriées à la prise de décision [...].*

La finalité d'un système décisionnel est le pilotage d'entreprise.

*Les systèmes de gestion sont dédiés **aux métiers** de l'entreprise [...].*

*Les systèmes décisionnels sont dédiés **au management** de l'entreprise [...].*

(Goglin, 2001, pp21-22) »

Synonymes : informatique décisionnelle, *business intelligence*, BI

2. Data warehousing

Définition : Définition historique de Inmon

« *A data warehouse is a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management's decisions. The data warehouse contains granular corporate data.*

(Inmon, 2002, p31) »

Définition

Un data warehouse (DW✯) est une base de données construite par copie et réorganisation de multiples sources (dont principalement le système transactionnel de l'entreprise), afin de servir de source de données à des applications décisionnelles :

- il agrège de nombreuses données de l'entreprise (**intégration**) ;
- il mémorise les données dans le temps (**historisation**) ;
- il les organise pour faciliter les requêtes de prise de décision (**optimisation**).

(Goglin, 2001, p27) [(Goglin, 2001)]

Synonymes : entrepôt de données, base de données décisionnelle

Fondamental

L'objectif du data warehouse est de permettre des requêtes sur de grands ensembles des données, la plupart du temps sous forme d'agrégats (`GROUP BY`) afin d'en obtenir une vision synthétique (propre à la prise de décision).

Remarque

Le data warehouse dédié au décisionnel est séparé du système transactionnel dédié à la gestion quotidienne.

Complément : Voir aussi

Data warehouse et data mart - p.28

3. Différence entre un DW et un système transactionnel

BD transactionnelle

Une base données classique est destinée à assumer des **transactions** en temps réel :

- Ajout, mise à jour suppression de données
- Questions sur des données identifiées ou questions statistiques

Datawarehouse

Un DW✯ est uniquement destiné à l'exécution de **questions statistiques** sur des données statiques (ou faiblement dynamiques).

PRIMITIVE DATA/OPERATIONAL DATA	DERIVED DATA/DSS DATA
• application oriented	• subject oriented
• detailed	• summarized, otherwise refined
• accurate, as of the moment of access	• represents values over time, snapshots
• serves the clerical community	• serves the managerial community
• can be updated	• is not updated
• run repetitively	• run heuristically
• requirements for processing understood a priori	• requirements for processing not understood *a priori*
• compatible with the SDLC	• completely different life cycle
• performance sensitive	• performance relaxed
• accessed a unit at a time	• accessed a set at a time
• transaction driven	• analysis driven
• control of update a major concern in terms of ownership	• control of update no issue
• high availability	• relaxed availability
• managed in its entirety	• managed by subsets
• nonredundancy	• redundancy is a fact of life
• static structure; variable contents	• flexible structure
• small amount of data used in a process	• large amount of data used in a process
• supports day-to-day operations	• supports managerial needs
• high probability of access	• low, modest probability of access

Un changement d'approche, extrait de (Inmon, 2002, p15)

4. Implémentation du DW avec un SGBDR

Fondamental

Les deux problématiques fondamentales des DW★ sont l'optimisation et la simplification : comment rester performant et lisible avec de très gros volumes de données et des requêtes portant sur de nombreuses tables (impliquant beaucoup de jointures) ?

On utilise massivement :

- **Les vues concrètes : Un data warehouse procède par copie depuis le ou les systèmes transactionnels**
- **La dénormalisation : Un data warehouse est hautement redondant**

Fondamental

Le caractère statique du data warehouse efface les inconvénients de ces techniques lorsqu'elles sont mobilisées dans des systèmes transactionnels.

Rappel

Dénormalisation - p.29
Vues concrètes - p.29

5. Modélisation logique de données en étoile

Définition : Le modèle en étoile

Le **modèle en étoile** est une représentation fortement **dénormalisée** qui assure un haut niveau de performance des requêtes même sur de gros volumes de données.

Exemple

Exemple de modèle dimensionnel en étoile

Complément : Modèle en flocon

Le modèle en flocon est aussi un modèle dénormalisé, mais un peu moins que le modèle en étoile : il conserve un certain niveau de décomposition pour chaque dimension prise isolément.

Complément : Voir aussi

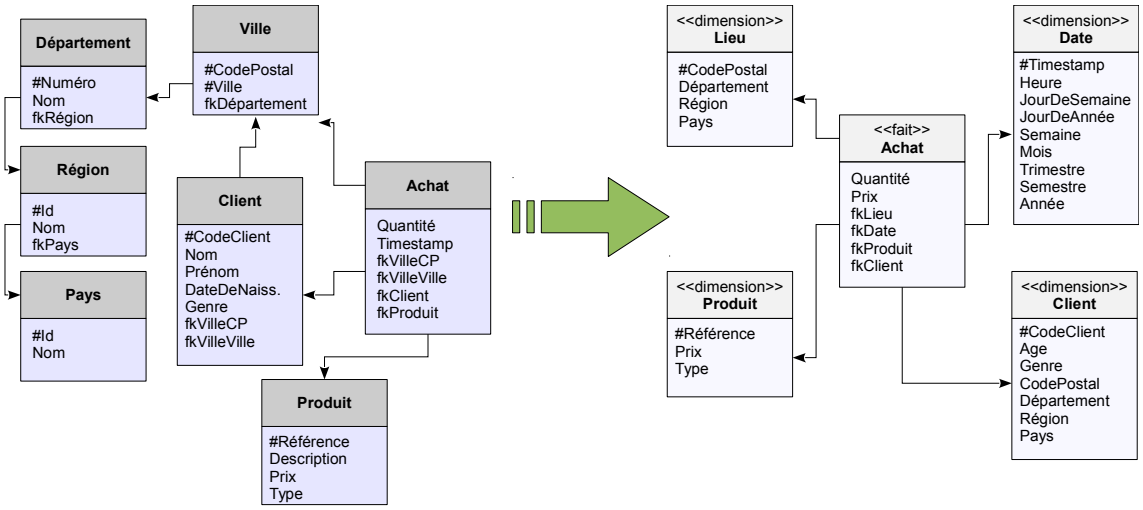
Modélisation en flocon - p.31

6. Extraction Transformation Loading

Définition : ETL

L'ETL (*Extraction Transformation Loading*) est le processus de copie des données depuis les tables des systèmes transactionnels vers les tables du modèle en étoile du data warehouse.

Exemple



Exemple de modèle dimensionnel en étoile

Remarque

Les tables du modèle dimensionnel peuvent être vues comme des **vues concrètes** sur le systèmes transactionnel, à la nuance que des transformations (correction d'erreur, extrapolation...) peuvent avoir été apportées dans le processus ETL.

B. Pour aller plus loin

Complément

https://stph.scenari-community.org/dwh/[1]

1 - https://stph.scenari-community.org/dwh/

Exercice

II

A. Étude de cas : Fantastic

Cet exercice constitue une simulation de modélisation dimensionnelle basée sur une étude de cas.

1. Approche générale de modélisation

Rappel

La modélisation en étoile

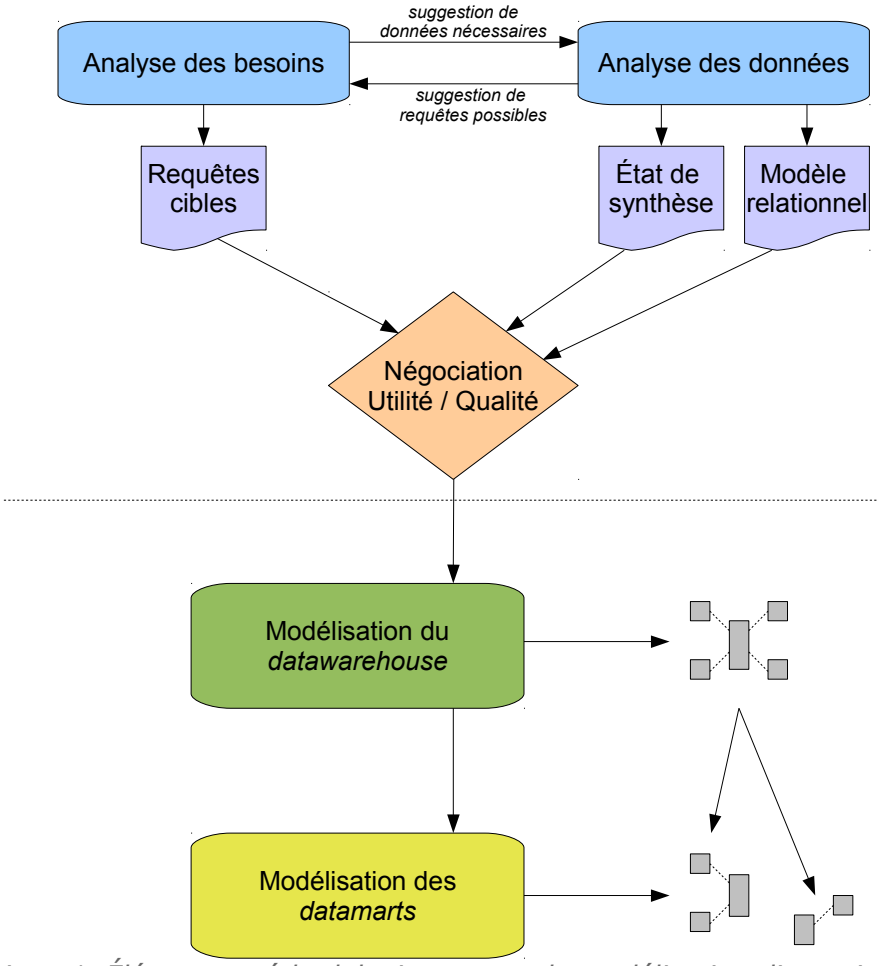
Fondamental

Un modèle dimensionnel est le résultat :

- **d'une analyse des besoins : ce que je souhaite étudier.**
- **d'une analyse des données disponibles : ce que je peux étudier.**

Méthode : Méthode générale de modélisation

1. Analyse des données
   a. Étude des sources de données (quantification, analyses générales)
   b. Qualification des données (qualité et intérêt)
   c. Intégration logique des données (simulation d'un schéma relationnel virtuel)
   d. Normalisation du schéma virtuel en 3NF pour en avoir une vue cohérente
2. Analyse des besoins clients
   a. Exprimer les besoins sous la forme de requêtes décisionnelles
   b. Réaliser les vues hiérarchiques pour chaque requête
3. Sélectionner les requêtes qui seront effectivement réalisables en fonction des données disponibles
4. Conception du data warehouse et des data marts
   a. Séparer les requêtes en fonction de la granularité de la table des faits (grain fin des ventes, grain plus grossier du ticket de caisse, etc.)
   b. Créer un data warehouse intégrant toutes les requêtes de grain fin
   c. Extraire un data mart par niveau de grain supérieur et/ou pour des thématiques particulières nécessitant par exemple une pré-agrégation



Graphique 1 Éléments méthodologiques pour la modélisation dimensionnelle

Remarque

Il est généralement intéressant de paralléliser les tâches d'analyse des besoins et d'analyse des données.

En particulier il est inutile d'aller trop loin dans l'expression de besoins que l'on sait **a priori** impossibles à satisfaire pour cause d'absence de donnée ou d'absence de donnée exploitable.

Rappel : Informations

Il est conseillé de conserver certains champs d'information dans le modèle dimensionnel, même s'ils ne seront pas exploités pour les calculs ou les agrégats.

Cela permettra par exemple d'identifier des enregistrements, comme les désignations de produits.

On pourra noter en italique ces champs dans le modèle dimensionnel.

2. Présentation du cas

a) Cas Fantastic : Problème posé

Vous travaillez en tant qu'ingénieur spécialisé dans les systèmes décisionnels au siège de l'entreprise française "Fantastic".

L'entreprise "Fantastic" vend principalement des ouvrages de divertissement de type science fiction, thriller, policier... Elle dispose pour cela de plusieurs magasins de vente dans les centres des grandes villes en France.

La direction de l'entreprise souhaite faire une étude large sur les ventes de l'année passée afin de prendre des orientations stratégiques nouvelles : ouverture de nouveaux magasins, fermeture ou transfert de magasins mal implantés, extension territoriale à de nouveaux départements français, réorganisation des directions, réorientation du marketing, élargissement ou réduction du catalogue, etc.

Fondamental

La question posée est donc : quels sont les facteurs sur lesquels l'on pourrait jouer pour augmenter les ventes ?

Elle vous charge dans ce cadre de mettre en place une solution logicielle permettant d'intégrer les données pertinentes et de pouvoir les interroger efficacement sous des angles divers.

Notons que bien entendu, la direction éclairée de l'entreprise ne compte pas se fier à ces seuls facteurs de ventes pour prendre ses décisions, mais bien privilégier les facteurs sociaux et territoriaux, en dialoguant avec ses salariés et ses clients, pour maintenir sa mission culturelle et son rôle d'entreprise citoyenne. Votre posture d'ingénieur est bien entendu de se préoccuper de ces dimensions fondamentales, même si elles seront largement ignorées dans le cadre de cet exercice à vocation essentiellement technique. Elle pourront néanmoins être brièvement abordées en marge de vos rapports d'analyse.

b) Cas Fantastic : Données disponibles

Attention : Installation

Le cas repose sur une installation spécifique, nous utiliserons les termes génériques suivantes :

- ***oracle.univ.fr* le nom d'un serveur Linux hébergeant une instance Oracle**
- ***prof* le nom d'un schéma Oracle géré par l'enseignant**
- ***etuXXX* le nom d'une série de schémas Oracle disponibles pour les étudiants**
- ***prof* le nom d'un compte Linux géré par l'enseignant sur *oracle.univ.fr***
- ***profdir* le chemin du compte de *prof* sur le serveur Linux**
- ***etuXXX* le nom d'une série de comptes Linux disponibles pour les étudiants**
- ***www.univ.fr/fantastic* le nom d'un site Web géré par l'enseignant**

Exemple : Installation UTC

En 2016, pour le cours AI07 à l'UTC l'installation est la suivante :

- `sme-oracle.sme.utc` (serveur Oracle)
- *ai07* (schéma Oracle et compte Linux enseignant)
- */volsme/users/ai07* (chemin du compte Linux enseignant)
- *ai07aXXX* (schémas Oracle et comptes Linux étudiants)
- *https://stph.scenari-community.org/ai07/fantastic*[2] (site Web de mise à disposition des données)

Catalogue des livres

Une base Oracle contient le catalogue complet de l'entreprise que chaque magasin a à sa disposition.

- Cette base, composée d'une seule table publique `catalogue`, est disponible sur le serveur Oracle sous le schéma *prof*.

Fichier des ventes

Un fichier *Fantastic* contient une consolidation de l'ensemble des ventes de l'année passée réalisées dans chaque magasin.

2 - https://stph.scenari-community.org/ai07/fantastic

- Ces données sont disponibles sous la forme d'un fichier CSV dans un répertoire du serveur `oracle.univ.fr` : `profdir/data`
- La structure du fichier est : Numéro de ticket, date de ticket, produit, magasin (comme c'est rappelé dans la documentation HTML associée)

Fichier des magasins

Un fichier ODS géré par la direction marketing contient pour chaque magasin l'organisation des rayonnages : `marketing.ods`

- Le responsable des ventes de chaque département décide de l'organisation des rayonnages des magasins de son département.
- Il existe 3 types de rayonnage : par Auteur (A), par Année (Y), par Éditeur (E)
- Le fichier est déposé sur *www.univ.fr/fantastic/data1*

Données géographique sur les départements

Un stagiaire a trouvé sur Internet un fichier permettant de connaître la population de chaque département, présageant que cette information sera utile.

- Le stagiaire parvient à trouver une information un peu datée qui pourra suffire sous la forme d'un fichier CSV : `departementsInsee2003.txt`.
- Le fichier ainsi que sa documentation en HTML est déposé sur *www.univ.fr/fantastic/data1*

Méthode

Inspecter les données pour chaque source :

1. Se connecter à la base Oracle avec SQL Developer et inspecter le schéma de la table Oracle.
2. Ouvrir un terminal et se connecter en `ssh` au serveur.
   Utiliser la commande Unix `more` pour regarder les premières lignes du fichier *Fantastic*.
3. Récupérer le fichier CSV `departementsInsee2003.txt` et l'ouvrir avec un éditeur de texte.
4. Récupérer le fichier ODS et l'ouvrir avec un traitement de texte.

c) Exemple d'inspection des données

i Fichier "Fantastic"

Documentation

Relevé des ventes de livres Fantastic pour l'année 2015

1. Ticket number
2. Date
3. ISBN
4. Store

Commande 1

```
1  head -10 Fantastic
```

```
1  "142282000001";2015-10-10;"769";"M143"
2  "0"; ;"9782841724680";"M119"
3  "082229000003";2015-08-18;"9782266186315";"83"
4  "082229000003";2015-08-18;"9782266186315";"M83"
5  "082229000003";2015-08-18;"9782266186315";"M83"
6  "028093000004";2015-04-04;"9780765341600";"M29"
7  "115072000005";2015-03-14;"9782290348789";"M116"
8  "040187000006";2015-07-07;"9782702141045";"M41"
```

```
"031002000007";2015-01-03;"552";"M32"
"055114000008";2015-04-25;"9782207301999";"M56"
```

Commande 2

```
wc -l Fantastic
```

```
512018
```

ii Fichier "departementsInsee2003.txt"

Documentation

Liste des départements français, 2003

1. **Department** Départements français métropolitains
2. **DptName** Nom du département
3. **Population** Population du département en 2003

Commande 1

```
head -10 departementsInsee2003.txt
```

```
"01";"Ain";"529378"
"02";"Aisne";"552320"
"03";"Allier";"357110"
"04";"Alpes-de-Haute-Provence";"144809"
"05";"Hautes-Alpes";"126636"
"06";"Alpes-Maritimes";"1022710"
"07";"Ardèche";"294522"
"08";"Ardennes";"299166"
"09";"Ariège";"142834"
"10";"Aube";"301388"
```

Commande 2

```
wc -l departementsInsee2003.txt
```

```
95
```

iii Fichier "marketing.ods"

Documentation

Rayonnage : Le responsable des ventes de chaque département décide de l'organisation des rayonnages des magasins de son département. Il existe 3 types de rayonnage : par Auteur (A), par Année (Y), par Éditeur (E)

Ouverture du fichier dans LibreOffice (extrait)

Magasin	Département	Rayonnage		Rayon Bestseller
M62	1	E	Editor	1
M101	2	A	Author	1
M122	4	A	Author	0
M130	6	A	Author	1
M139	6	A	Author	1
M93	6	A	Author	1
M119	8	A	Author	0
M69	8	A	Author	0
M13	10	E	Editor	0
M10	13	A	Author	1

(153 lignes, dont l'entête)

iv Table "catalogue"

Commande 1

```
1  DESCRIBE catalogue
```

```
1  Name       Null     Type
2  ---------- -------- -------------
3  REF        NOT NULL NUMBER(38)
4  ISBN       NOT NULL VARCHAR2(13)
5  TITLE      NOT NULL VARCHAR2(255)
6  AUTHORS    NOT NULL VARCHAR2(255)
7  LANGUAGE            VARCHAR2(3)
8  PUBDATE             VARCHAR2(25)
9  PUBLISHER           VARCHAR2(255)
10 TAGS                VARCHAR2(255)
11 GENRE               VARCHAR2(255)
```

Commande 2

```
1  SELECT * FROM catalogue
2  WHERE ROWNUM < 11;
```

REF	ISBN	TITLE	AUTHORS	LANGUAGE	PUBDATE	PUBLISHER	TAGS	GENRE
2778	721	[Anita Blake-4]Narcisse Enchaîné	Hamilton,Laurell Kaye	fra	2001-01-01T23:00:00+00:00	?	Fantastique, Roman Fantastique	Fantastic
2833	779	[La Communauté du Sud-1] Quand le danger rôde	Harris,Charlaine	fra	2001-01-01T23:00:00+00:00	?	Fantastique	Fantastic
315	141	[Assassin Royal-07] Le prophète blanc	Hobb,Robin	fra	2001-01-01T21:00:00+00:00	?	Fantasy	Fantastic
2883	828	[Assassin Royal-08] La secte maudite	Hobb,Robin	fra	2001-01-01T21:00:00+00:00	?	Fantasy	Fantastic
2770	712	[Merry Gentry-1] Le baiser des ombres	Hamilton,Laurell Kaye	fra	2000-01-14T23:00:00+00:00	?	Fantastique	Fantastic
339	233	[Anita Blake-9]Papillon d'Obsidienne	Hamilton,Laurell Kaye	fra	2000-01-01T23:00:00+00:00	?	Fantastique, Roman Fantastique	Fantastic
2882	827	[Les Aventuriers de la mer-07] Le seigneur des Trois Règnes	Hobb,Robin	fra	2000-01-01T21:00:00+00:00	?	Fantasy	Fantastic
317	143	[Les Aventuriers de la mer-08] Ombres et flammes	Hobb,Robin	fra	2000-01-01T21:00:00+00:00	?	Fantasy	Fantastic
2904	848	[Les Aventuriers de la mer-09] Les marches du trône	Hobb,Robin	fra	2000-01-01T21:00:00+00:00	?	Fantasy	Fantastic
259	214	[Robert Langdon-1]Anges et démons	Brown,Dan	fra	1999-12-31T21:00:00+00:00	?	Policier, Thriller Scientifique, BROWN, Langdon	Crime

Commande 3

```
1  SELECT COUNT(*) FROM catalogue;
```

```
1  1443
```

3. Étude des données

Objectifs

Savoir faire une étude des données existantes

a) Étude séparée des sources données

Méthode

Pour chaque source de données identifiée on proposera une synthèse avec les informations suivantes :

- Nom, origine précise, nombre de lignes de la source
- Nom, type et description des colonnes
- Qualité
  - 0 : données inexploitables
  - 1 : données peu exploitables (traitements incertains)
  - 2 : données exploitables après traitements
  - 3 : données exploitables sans traitement
- Utilité
  - 0 : données sans intérêt
  - 1 : données utiles pour la documentation uniquement
  - 2 : données utiles a priori
  - 3 : données utiles avec certitude
- Commentaires.

Exemple

Nom	Description	Type	Qualité	Utilité	Commentaire
isbn	Identifiant international d'un livre publié	char(3) et char(13)	3	3	Certaines ISBN ne sont pas corrects (3 caractères au lieu de 13) ; clé de référencement dans le fichier de ventes
titre	Titre du livre	varchar(255)	2	1	
auteur	Auteur(s) du livre	varchar(255)	1	2	Peut contenir plusieurs auteurs ; certains auteurs sont inscrits différemment d'un livre à l'autre
...					

Tableau 1 Table Oracle oracle.utc.fr/schema.table [1000 lignes]

b) Étude intégrée des sources de données

Méthode

Afin d'avoir une vision globale de l'ensemble des données, il est conseillé de rétro-concevoir :

- une représentation relationnelle des données telles qu'elles existent
- une représentation relationnelle des données idéalisées (telles qu'elles existeraient si elles étaient normalisées dans une même BD)
- une représentation conceptuelle en UML

4. Cas Fantastic : Étude des données

[1h]

Afin de réaliser votre travail, l'entreprise vous met à disposition les données suivantes.

Dans le contexte de cet exercice, les données vous sont livrées *a priori*, notez que dans un contexte réel, vous aurez la plupart du temps à rechercher vous même les données qui peuvent exister.

Données disponibles

Question 1

[*Solution n°1 p 21*]

Établissez le modèle **relationnel** sous-jacent aux données présentes.

Indice :

Pour initier une connexion ssh : `ssh user@serveur`

Question 2

Étudiez les données dont vous disposez et proposez une synthèse des données disponibles pour chaque source.

Indice :

Pour compter les lignes d'un fichier texte sous Linux, on peut utiliser la commande `wc -l`

Question 3

[*Solution n°2 p 21*]

Afin de clarifier les données et leur organisation, rétro-concevez un modèle relationnel **normalisé** en troisième forme normale unifiant toutes les données grâce à l'identification de clés primaires et l'expression de clé étrangères.

Question 4

Rétro-concevez le modèle **conceptuel** en UML correspondant au modèle relationnel normalisé (un modèle UML peut être plus facile à utiliser ensuite).

5. Étude des besoins utilisateurs

Objectifs

Savoir formaliser une étude de besoins sous forme de requêtes multidimensionnelles

a) Requête décisionnelle

Définition : Requête décisionnelle

Une requête décisionnelle exprime toujours la mesure d'une quantification de **faits** par rapport à des **dimensions**, sous une forme du type : "Quelle a été la quantité de ... en fonction de ...".

Synonyme : Vue, requête multidimensionnelle

Fondamental

- **Les faits sont des grandeurs que l'on cherche à mesurer (prix, quantité...)**
- **Les dimensions sont des axes d'analyse (date, lieu, produit, personne...)**

Syntaxe

```
quantité de faits
/ dimension1
/ dimension2
...
```

Exemple

"Quelle a été la quantité de produits vendus en fonction des départements et des mois de l'année."

```
quantité de produits vendus
/ département
/ mois
```

b) Rapport

Définition : Rapport

La réponse à une requête décisionnelle est un rapport, généralement sous une forme tabulaire ou graphique.

Synonyme : État

Exemple

01													02												03											
J	F	M	A	M	J	J	A	S	O	N	D		J	F	M	A	M	J	J	A	S	O	N	D		J	F	M	A	M	J	J	A	S	O	N	D
04													05												06											
J	F	M	A	M	J	J	A	S	O	N	D		J	F	M	A	M	J	J	A	S	O	N	D		J	F	M	A	M	J	J	A	S	O	N	D
07													08												09											
J	F	M	A	M	J	J	A	S	O	N	D		J	F	M	A	M	J	J	A	S	O	N	D		J	F	M	A	M	J	J	A	S	O	N	D

Exemple de rapport

Méthode

Les besoins des utilisateurs s'expriment en général plus facilement sous la forme d'exemples de rapports recherchés.

c) Hiérarchie

Définition : Hiérarchie

Une hiérarchie est un ensemble de paramètres d'étude de granularité croissante appartenant à une même dimension au sein d'un modèle décisionnel.

Exemple

```
quantité de produits vendus
/ lieu (département, région)
/ date (jour, mois, trimestre, semestre)
```

Remarque

Une même dimension peut contenir plusieurs hiérarchies.

6. Cas Fantastic : Étude des besoins

[30 min]

À partir de l'étude des besoins sommaire effectuée par un de vos collègues, et en fonction des données disponibles, exprimer les **requêtes cibles** de votre système.

Le contexte de l'exercice ne permet pas de dialoguer réellement avec des utilisateurs, **en situation réelle il faudra développer cette phase de recueil des besoins des utilisateurs**. Vous pourrez amendez l'indicateur d'utilité des données en fonction de cette étude.

Question 1

[*Solution n°3 p 21*]

« *La direction marketing est en charge de l'implantation des magasins dans les départements et de l'organisation des rayonnages (type de rangement et présence de rayons spécifiques pour les best-sellers). Elle cherche à savoir si l'organisation du rayonnage des magasins a une influence sur les volumes ventes, et si cela varie en fonction des jours de la semaine ou de certaines périodes de l'année. Elle voudrait également savoir si certains magasins ou départements sont plus dynamiques que d'autres.* »

Question 2

[*Solution n°4 p 22*]

« *La direction éditoriale se demande si certains livres se vendent mieux à certaines dates et/ou dans certains magasins ou départements. Elle aimerait également savoir si certains auteurs ou éditeurs se vendent mieux, et s'il existe un lien entre l'ancienneté des livres et les ventes. Elle se demande aussi si certaines périodes sont plus propices que d'autres à l'écoulement des livres les plus anciens.* »

7. Modélisation en étoile

a) Table des faits

Définition

« *A row in a fact table corresponds to a measurement. A measurement is a row in a fact table. All the measurements in a fact table must be the same grain.* (Kimball, Ross, 2008, p.17) »

« *Fact tables express the many-to-many relationships between dimensions in dimensional models.*(Kimball, Ross, 2008, p.19) »

Remarque

La table des faits est (dans la plupart des cas) la table la plus volumineuse (avec le plus grand nombre de lignes) du modèle.

Exemple

Daily Sales Fact Table
Date Key (FK) Product Key (FK) Store Key (FK) Quantity Sold Dollar Sales Amount

Exemple de table des faits (Kimball, Ross, 2008, p.17)

Fondamental : Faits additifs et numériques

« *The most useful facts are numeric and additive.* (Kimball, Ross, 2008, p.17) »

Méthode : Granularité minimale

« *Preferably you should develop dimensional models for the most atomic information captured by a business process. Atomic data is the most detailed information collected; such data cannot be subdivided further.* (Kimball, Ross, 2008, p.34) »

Méthode : Granularité des data marts

Pour un data mart on peut pré-agréger sur un grain plus gros que le data warehouse : des colonnes d'agrégation (somme, moyenne, compte...) peuvent alors apparaître pour rendre compte statistiquement d'informations perdues à l'agrégation.

b) Table des dimensions

Définition

« *Dimension tables are the entry points into the fact table. [...] The dimension implement the user interface to the data warehouse.* (Kimball, Ross, 2008, p.20) »

Exemple

Product Dimension Table
Product Key (PK)
Product Description
SKU Number (Natural Key)
Brand Description
Category Description
Department Description
Package Type Description
Package Size
Fat Content Description
Diet Type Description
Weight
Weight Units of Measure
Storage Type
Shelf Life Type
Shelf Width
Shelf Height
Shelf Depth
... and many more

Exemple de table de dimension (Kimball, Ross, 2008, p.20)

Conseil : Intelligibilité

« *The best attributes are textual and discrete. Attributes should consist of real words rather than cryptic abbreviations.* (Kimball, Ross, 2008, p.20) »

8. Cas Fantastic : Modélisation

À partir de l'étude des données et des besoins utilisateurs, proposez un modèle **dimensionnel** pour un **data warehouse** permettant de traiter la question générale qui a été posée.

Question 1

[*Solution n°5 p 22*]

Proposez une modélisation dimensionnelle **en étoile** pour chaque contexte d'usage (directions marketing et éditoriale) :

1. identifiez la table des faits
2. identifiez les dimensions en intégrant les vues exprimées (utilisez le rapport qualité/utilité pour décider des données que vous souhaitez conserver)

Question 2

[*Solution n°6 p 23*]

Intégrer vos deux sous-modèles pour proposer le modèle de votre **data warehouse**.

Vous pourrez augmenter vos dimensions de données disponibles bien que non identifiées a priori lors de l'analyse des besoins.

Question 3

[*Solution n°7 p 23*]

Établissez les métadonnées du modèle dimensionnel : décrivez chaque données (type précis, description...) ; identifiez la clé primaire de chaque dimension, ainsi que les informations descriptives.

Question 4

[*Solution n°8 p 24*]

Proposez la représentation logique finale de votre modèle dimensionnel, en affichant :

- la table des faits et les dimensions
- l'ensemble des attributs
- les clés primaires et étrangères
- les types de données
- les hiérarchies

Solution des exercices

> Solution n°1 *(exercice p. 15)*

```
catalogue(isbn,titre,auteur,langue,parution,editeur, genre)
data(num,magasin,date,isbn)
marketing(dpt,rayonnage,ray_nom,magasin,rayon_bs, rayon_recent)
dpt(dpt,nom,pop)
```

> Solution n°2 *(exercice p. 15)*

```
auteur(#num,nom,prenom)
langue(#langue)
editeur(#editeur)

catalogue(#isbn,titre,fkauteur=>auteur,langue=>langue,parution,editeur=>editeur,genre)
data(num,magasin=>magasin,date,isbn=>catalogue)
magasin(#magasin, dpt=>dpt,rayonnage,ray_nom,rayon_bs, ray_recent)
dpt(#dpt,nom,pop)
```

> Solution n°3 *(exercice p. 17)*

Requêtes brutes

```
Quantite
    / rayonnage
    /  rayon_bs
    / rayon_recent
```

```
Quantite
    / semaine
     / mois, trimestre
     / jds
    / rayonnage
    /  rayon_bs
```

```
Quantite
    / magasin, dpt
```

Requêtes organisées par dimension

```
Quantite
    / date (semaine)
    / date (mois, trismestre)
```

```
4       / date (jds)
5      / magasin (rayonnage)
6      / magasin (rayon_bs)
7      / magasin (recent)
```

```
1  Quantite
2      / magasin (dpt)
```

> Solution n°4 *(exercice p. 17)*

Requêtes brutes

```
1  Quantite
2      / produit
3      / magasin, departement
4       / date, semaine
5       / date,  mois, trismestre
```

```
1  Quantite
2      / auteur
3      / editeur
```

```
1  Quantite
2      / jds
3      / jds, semaine
4      / parution
```

Requêtes organisées par dimension

```
1  Quantite
2      / produit
3      / magasin (departement)
4       / date (semaine)
5       / date (mois, trismestre)
```

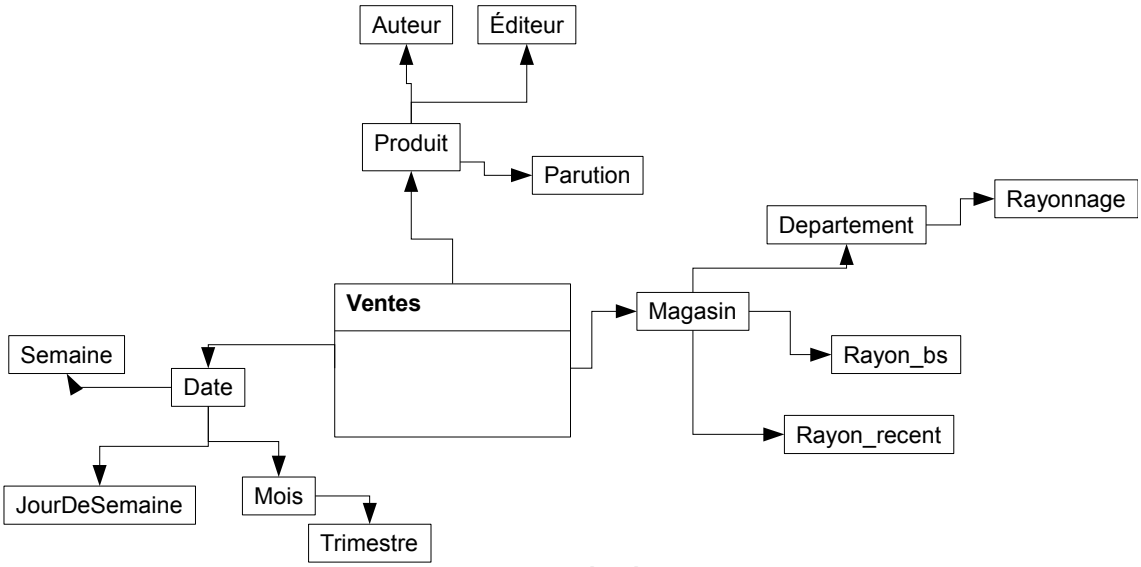
```
1  Quantite
2      / produit (auteur)
3      / produit (editeur)
```

```
1  Quantite
2      / date (jds)
3      / date (semaine)
4      / produit (parution)
```

> Solution n°5 *(exercice p. 19)*

Vue direction éditoriale

> Solution n°6 *(exercice p. 19)*



Vue intégrée

> Solution n°7 *(exercice p. 20)*

#ISBN	Char(13)	Clé primaire
Langue	Char(3)	Langue du livre sur trois lettres
Auteur	Varchar(50)	Auteur (Nom, Prénom) ; seul le premier auteur est conservé
Editeur	Varchar(50)	Type de produit
Parution	Char(4)	Année de parution sur 4 chiffres

Description de la dimension Produit (DataWarehouse)

#NumMag	Char(4)	Mxxx
Rayon_bs	0..1	0 : pas de rayon BS / 1 : un rayon BS
Dpt	Char(2)	Numéro du département
Population	1..20	Nombre d'habitants du département en centaine de milliers
Rayonnage	{A,Y,E}	Type de rayonnage dans le département

Description de la dimension Magasin (DataWarehouse)

#Date	Date (yyyy-mm-dd)	Date d'achat (clé primaire)
Mois	1..12	Mois d'achat
Trimestre	1..4	Trimestre d'achat
Semaine	1..52	Semaine d'achat
JDS	Lundi..Dimanche	Jour de la semaine d'achat

Description de la dimension Date (DataWarehouse)

> Solution n°8 *(exercice p. 20)*

Modèle dimensionnel du data warehouse Fantastic

Signification des abréviations

- BD Base de Données
- BI Business Intelligence
- DW Data Warehouse

Bibliographie

[(Goglin, 2001)] Goglin J.-F. (2001, 1998). *La construction du datawarehouse : du datamart au dataweb*. Hermes, 2ème édition.

[(Inmon, 2002)] Inmon W.-H. (2002, 1990). *Building the data warehouse*. Wiley Publishing, 3rd edition.

[(Kimbal, Ross, 2003)] Kimball R., Ross M. (2003). *Entrepôts de données : guide pratique de modélisation dimensionnelle*. Vuibert.

[(Kimball, Ross, 2008)] Kimball R., Ross M. (2008, 2002). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. Wiley Publishing, second edition.

[(Kimball et al., 2008)] Kimball R., Ross M., Thornthwaite W., Mundy J., Becker B. (2008, 1998). *The Data Warehouse Lifecycle Toolkit*. Wiley Publishing, second edition.

Index

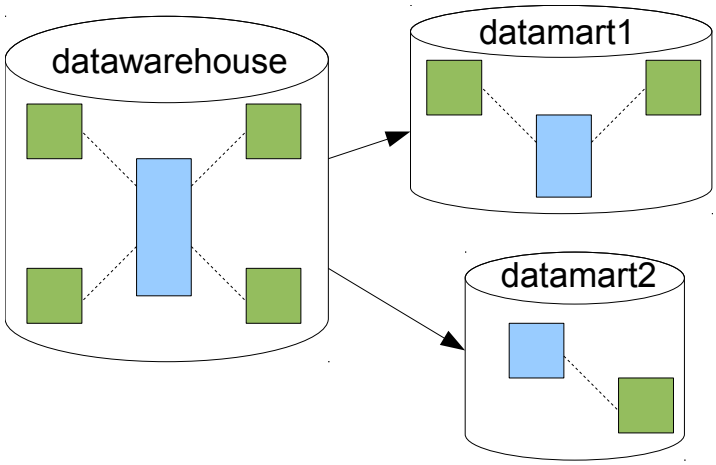
Contenus annexes

- Data warehouse et data mart

Un data warehouse et un *data mart* se distinguent par le spectre qu'il recouvre :

- Le *data warehouse* recouvre l'ensemble des données et problématiques d'analyse visées par l'entreprise.
- Le *data mart* recouvre une partie des données et problématiques liées à un métier ou un sujet d'analyse en particulier

Un *data mart* est fréquemment un sous-ensemble du *data warehouse* de l'entreprise, obtenu par extraction et agrégation des données de celui-ci.



Graphique 2 Datawarehouse et datamarts

Pourquoi des data marts ?

Les *data marts* sont destinés à pré-agréger des données disponibles de façon plus détaillée dans les *data warehouse*, afin à traiter plus facilement certaines questions spécifiques, critiques, etc.

Exemple : Ticket de caisse

Si un *data warehouse* enregistre un ensemble de ventes d'articles avec un grain très fin, un *data mart* peut faciliter une analyse dite de **ticket de caisse** (co-occurrence de ventes de produits par exemple) en adoptant un grain plus grossier (le ticket plutôt que l'article).

Complément : Ticket de caisse

La lecture de Entrepôts de données : guide pratique de modélisation dimensionnelle [(Kimbal, Ross, 2003)] est recommandée pour la modélisation dimensionnelle des tickets de caisse (en particulier pages 31-60 sur la grande distribution).

- Dénormalisation

Rappel

La normalisation est le processus qui permet d'optimiser un modèle logique afin de le rendre non redondant. Ce processus conduit à la fragmentation des données dans plusieurs tables.

Définition : Dénormalisation

Processus consistant à regrouper plusieurs tables liées par des références, en une seule table, en réalisant statiquement les opérations de jointure adéquates.

L'objectif de la dénormalisation est d'améliorer les performances de la BD⋆ en recherche sur les tables considérées, en implémentant les jointures plutôt qu'en les calculant.

Remarque : Dénormalisation et redondance

La dénormalisation est par définition facteur de redondance. A ce titre elle doit être utilisée à bon escient et des moyens doivent être mis en œuvre pour contrôler la redondance créée.

Méthode : Quand utiliser la dénormalisation ?

Un schéma doit être dénormalisé lorsque les performances de certaines recherches sont insuffisantes et que cette insuffisance à pour cause des jointures.

Attention : Inconvénients de la dénormalisation

La dénormalisation peut également avoir un effet néfaste sur les performances :

- **En mise à jour**

  Les données redondantes devant être dupliquées plusieurs fois.
- **En contrôle supplémentaire**

  Les moyens de contrôle ajoutés (*triggers*, niveaux applicatifs, etc.) peuvent être très coûteux.
- **En recherche ciblée**

  Certaines recherches portant avant normalisation sur une "petite" table et portant après sur une "grande" table peuvent être moins performantes après qu'avant.

Fondamental : Redondance et bases de données

La redondance volontaire est autorisée dans une base de données à trois conditions :

1. **avoir une bonne raison d'introduire de la redondance (améliorer des performances dans le cas de la dénormalisation)**
2. **documenter la redondance en explicitant les DF responsables de la non 3NF**
3. **contrôler la redondance par des mécanismes logiciels (triggers par exemple)**

- Vues concrètes

Un moyen de traiter le problème des requêtes dont les temps de calcul sont très longs et les fréquences de mise à jour faible est l'utilisation de vues concrètes.

Définition : Vue concrète

Une vue concrète est un stockage statique (dans une table) d'un résultat de requête.

Un accès à une vue concrète permet donc d'éviter de recalculer la requête et est donc aussi rapide qu'un accès à une table isolée. Il suppose par contre que les données n'ont pas été modifiées (ou que leur modification est sans conséquence) entre le moment où la vue a été calculée et le moment où elle est consultée.

Une vue concrète est généralement recalculée régulièrement soit en fonction d'un événement particulier (une mise à jour par exemple), soit en fonction d'un moment de la journée ou elle n'est pas consultée et où les ressources de calcul sont disponibles (typiquement la nuit).

Synonymes : Vue matérialisée

Complément : Voir aussi

Vue matérialisée sous Oracle - p.30

- Vue matérialisée sous Oracle

Introduction

Oracle 9i propose une extension au LDD SQL pour créer explicitement des vues matérialisées plutôt que la gestion manuelle classique consistant à créer une table et à l'alimenter manuellement avec une requête de type INSERT.

Rappel

Une vue matérialisée est un stockage statique (dans une table) d'un résultat de requête. Il permet donc d'anticiper des requêtes complexes en pré-calculant tout ou partie du résultat.

Vues concrètes - p.29

Syntaxe : Déclaration de vue matérialisée

```
1  CREATE MATERIALIZED VIEW nom_vue
2    [PARTITION BY ...]
3    BUILD IMMEDIATE -- quand la construire (alt. : DEFERED)
4    REFRESH FORCE -- mode de rafraîchissement (alt. : COMPLETE, FAST)
5    ON DEMAND -- quand la rafraîchir (alt. : COMMIT)
6    ENABLE QUERY REWRITE  -- utilisable pour l'optimisation automatique de requêtes
   (alt. : DISABLE)
7    AS
8    SELECT ...
```

Syntaxe : Rafraîchissement manuel de la vue

Si le rafraîchissement de la vue peut être programmé, il est néanmoins parfois nécessaire de procéder à un rafraîchissement manuel.

```
1  EXECUTE dbms_mview.refresh('nom_vue');
```

Optimisation des requêtes

La déclaration d'une vue matérialisée dans Oracle, permettra à l'optimiseur de l'utiliser pour effectuer la requête correspondante plutôt que d'interroger les tables sources.

Attention

Oracle ne pourra pas toujours inférer automatiquement que des vues matérialisées sont utilisables à la place des tables originales. Il est donc nécessaire dans ce cas de réécrire soi-même les requêtes concernées. En pratique c'est souvent le cas si les vues sont complexes.

Raffraîchissement de la vue matérialisée

La déclaration d'une vue matérialisée sous Oracle 9i permet également de planifier automatiquement les modalités de rafraîchissement de la vue, plutôt que de devoir gérer un rafraîchissement uniquement manuel.

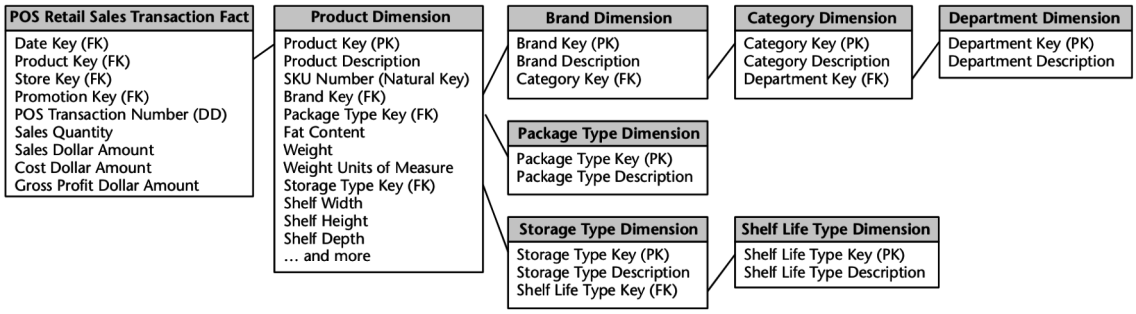
Remarque

Une vue matérialisée peut être indexée ou partitionnée comme toute table Oracle.

- Modélisation en flocon

Définition

Un modèle en flocon est un modèle pour lequel chaque dimension est représentée avec plusieurs tables. Il est donc plus normalisé (moins redondant) qu'un modèle en étoile.



Exemple de dimension représentée en flocon (Kimball, Ross, 2008, p.55)

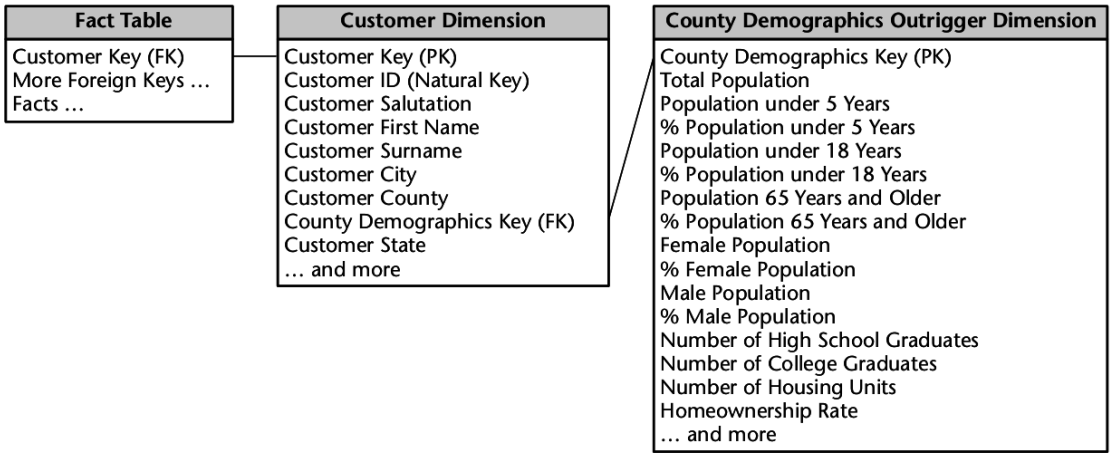
Attention

Les représentation en flocon sont déconseillées en général (Kimball, Ross, 2008, p.56) [(Kimball, Ross, 2008)]:

- **Le modèle en flocon est plus complexe et son appréhension par l'utilisateur difficile**
- **Les performance en requête sont diminuées par les jointures**
- **Le gain en espace disque est faible (les dimensions sont peu volumineuses relativement aux faits)**

Méthode

La normalisation partielle est préconisée lorsqu'il y a des répétitions très nombreuses sur un grand nombre de colonnes (Kimball, Ross, 2008, p.153) [(Kimball, Ross, 2008)].



Exemple de cas pertinent de représentation en flocon d'une dimension (Kimball, Ross, 2008, p.153)