Conception d'un DataWarehouse : Etude de cas

Approche générale de Modélisation:

1. Analyse des données

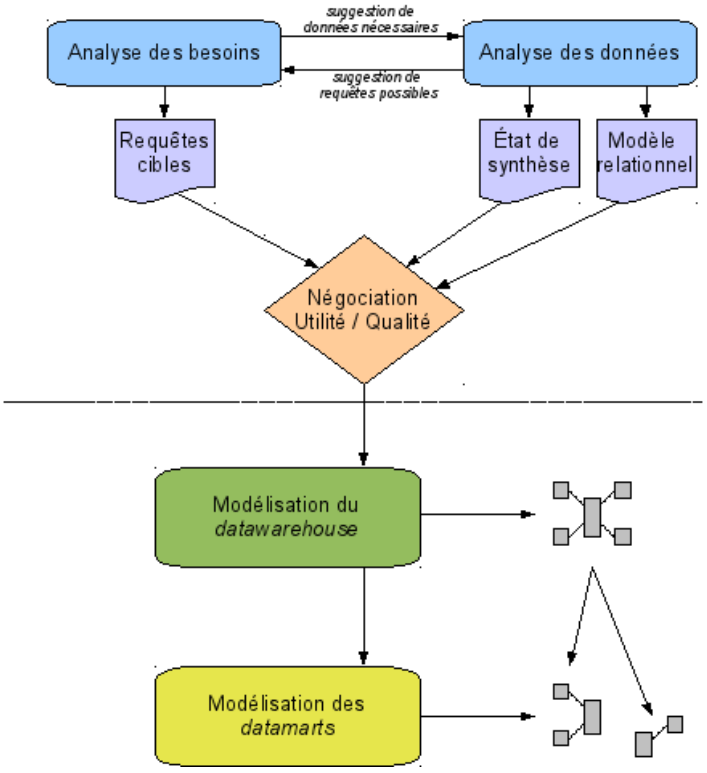
a. Étude des sources de données (quantification, analyses générales)
b. Qualification des données (qualité et intérêt)
c. Intégration logique des données (simulation d'un schéma relationnel virtuel)
d. Normalisation du schéma virtuel en 3NF pour en avoir une vue cohérente

2. Analyse des besoins clients

a. Exprimer les besoins sous la forme de requêtes décisionnelles
b. Réaliser les vues hiérarchiques pour chaque requête
c. Sélectionner les requêtes qui seront effectivement réalisables en fonction des données disponibles

3. 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



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.

Note : 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.

Problème posé dans notre cas

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.

Analyse des données

Données disponibles

Catalogue des livres

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

Cette base est composée d'une seule table publique **catalogue**.

La structure de la table est : **isbn, titre, auteur, langue, parution, editeur, genre**

Fichier des ventes

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

Ces données sont disponibles sous la forme d'un fichier CSV : **data**.

La structure du fichier est : **numTicket**, **date de ticket**, **produit**, **magasin**

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)

La structure du fichier est : **dpt, rayonnage, ray_recent, magasin, rayon_bs**

Données géographique sur les départements

Un stagiaire a trouvé sur Internet un fichier permettant de connaître la population de chaque département.

Le stagiaire parvient à trouver une information un peu datée qui pourra suffire sous la forme d'un **fichier CSV** : departementsInsee2003.txt.

La structure du fichier est : **dpt, nom, pop** (Department, DptName, Population)

Rétro concevoir un Modèle de données normalisé

Modèle initial

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

Modèle normalisé

- auteur **(#num, nom, prenom)**
- langue **(#langue)**
- editeur **(#editeur)**
- catalogue **(#isbn, titre, fkauteur, langue, parution, editeur, genre)**
- data **(num, magasin, date, isbn)**
- magasin **(#magasin, dpt, rayonnage, ray_recent, rayon_bs)**
- dpt **(#dpt,nom,pop)**

Analyse des besoin client

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

Les besoins receuillis :

1- *«* 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. *»*

⇨ On veut mesurer la :

Quantite
- **/ semaine**
- **/ mois, trimestre**
- **/ jds**
- **/ rayonnage**
- **/ rayon_bs**

Quantite
- **/ rayonnage**
- **/ rayon_bs**
- **/ ray_recent**

Quantite
- **/ magasin, dpt**

2- « 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. »

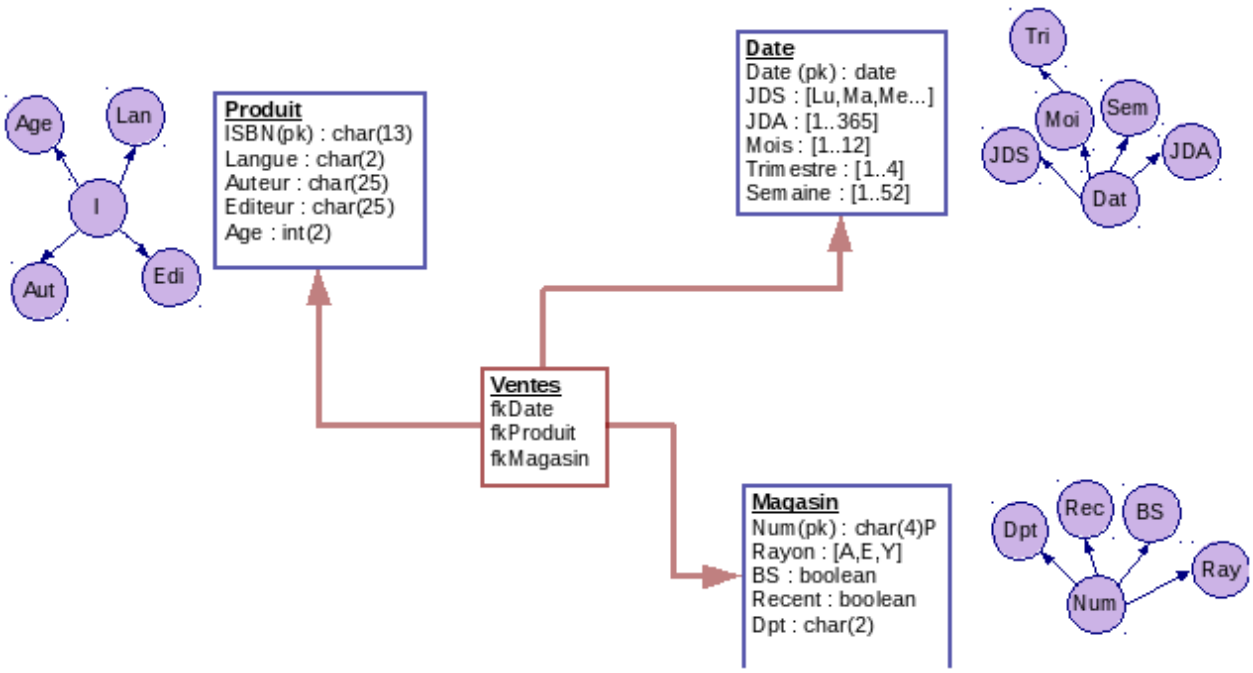
⇨ On veut mesurer la :

Quantite
- **/ produit**
- **/ magasin, dpt**
- **/ date, semaine**
- **/ date, mois, trismestre**

Quantite
- **/ jds**
- **/ jds, semaine**
- **/ parution**

Quantite
- **/ auteur**
- **/ editeur**

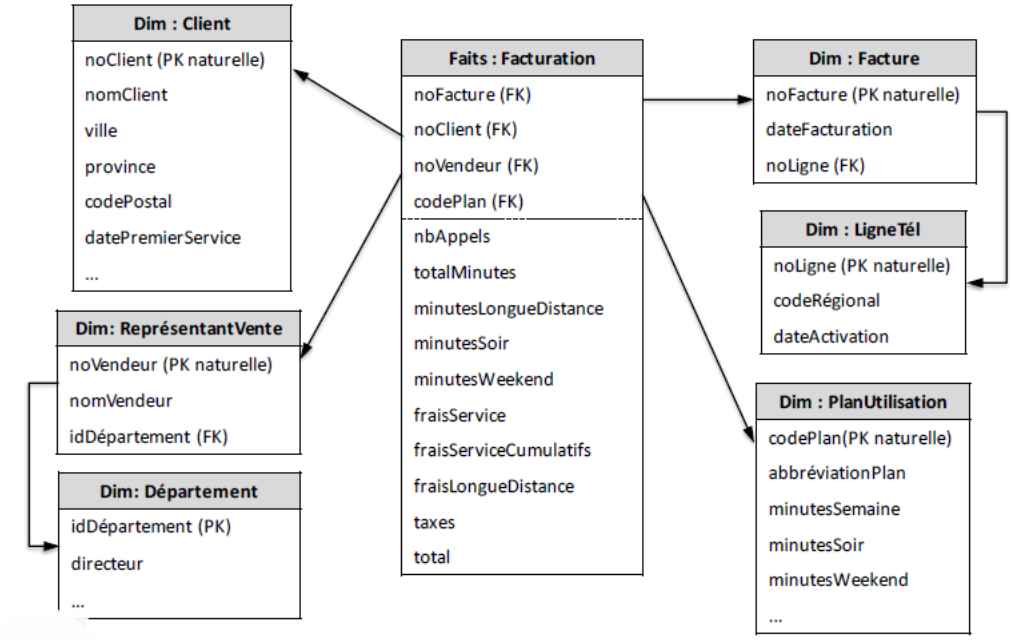
Conception du data warehouse



Cas d'étude en télécommunications

Processus / Dimension	Date	Client	Produit	Plan d'utilisation	Canal de vente	Ligne tél.	Relai	Représentant de vente	Employé	Appel service
Facturation client	X	X	X	X	X	X				
Gestion du trafic d'appels	X	X	X	X		X	X	X		
Inventaire	X		X		X					
Service à la clientèle	X	X	X	X		X			X	X
...										

Modèle initial :



1. Granularité:

• Le grain le plus fin correspond réellement à la facturation d'une ligne d'un client;

• Solution:

- Mettre la clé de la dimension Ligne dans la table de faits.

2. Clés primaires des dimensions:

• Les clés primaires des dimensions doivent être des cl és de substitution (surrogate keys);

• Solution:

- Remplacer les clés primaires par des clés artificielles;
- Lorsque nécessaire, mettre les clés naturelles dans la table de faits comme dimensions dégénérées (DD).

A quoi sert une clé de substitution ?

Remplacer la clé artificielle ou naturelle : Une clé de substitution remplace la clé artificielle en terme d'utilisation, ce n'est plus la clé naturelle qui sera utilisé pour faire les jointures avec les tables de faits ou les autres tables de dimension (niveaux hiérarchiques dans le cas d'une dimension en flocons de neiges);

Compléter l'information : La clé de substitution n'a aucun sens en terme d'affaire, elle est utilisée dans l'entrepôt de données seulement ! et on aura toujours besoin de la clé artificielle ou naturelle dans la dimension pour pouvoir faire la correspondance entre l'élément de dimension (un client par exemple) dans l'entrepôt de données et l'élément de la table des clients dans le système opérationnel.

Les avantages

Performance : Accélère l'accès aux données du moment ou l'on va utiliser un index numérique vu que le type de données de la clé de substitution est numérique.

Indépendance du système source : On ne peut garantir que la clé d'affaire ne change pas dans les systèmes sources.

Historique des changements et granularité infinie : Si l'on désire garder l'historique des changements de la dimension selon certains critères (SCD) nous devons gérer la clé de substitution. Nous nous retrouverons facilement avec plusieurs enregistrements de la même clé d'affaire dans la dimension.

Id	EAN_Code	Product_Name	Brand	Product_Category
1	977147396801	Canon EOS Rebel	Cannon	Camera
2	977147396802	Nikon Coolpix	Nikon	Camera
3	977147396803	Sony Cyber-shot	Sony	Camera
4	977147396804	Olympus XZ-1	Olympus	Camera
5	977147396804	Olympus XZ-1	Olympus	Electronics

3. **Dimension temporelle:**

• La date de facturation est modélisée comme un attribut de Facture au lieu d'être une dimension conforme;

• Solution:

- Créer une dimension à rôles multiples DateFacturation, basée sur la dimension conforme de Date.

4. **Dimensions normalisées:**

• La hiérarchie ReprésentantVente – Département est normalisée, causant des jointures inutiles;

• Solution:

- Mettre les attributs de Département directement dans ReprésentantVente(i.e., dénormaliser).

5. **Attributs non-descriptifs:**

• Certaines dimensions ont des attributs peu informatifs (ex: abbréviationPlan);

• Solution:

- Rajouter des attributs descriptifs pour rendre les données plus compréhensibles aux utilisateurs d'affaires.

6. **Faits non-additifs:**

• Le fait fraisServiceCumulatifs n'est pas additif;

• Solution:

- Retirer cette colonne de la table de fait et calculer les valeurs cumulatives sur demande.

Une mesure peut être additive, semi-additive ou non additive :

Mesure additive - on peut sommer sur toutes les dimensions tout en conservant un sens.

Ex: si l'on considère les sommes des ventes par produits, villes et mois, on peut faire la somme des valeurs des cellules tout en conservant un sens aux données.

Mesure semi-additive - on peut sommer sur certaines dimensions en gardant un sens mais pas sur toutes.

Ex: si on considère un cube décrivant l'état des stocks par ville, produit et mois, il est possible de faire la somme sur les dimensions ville et produit pour connaitre l'état global des stocks pour toutes les villes ou pour tous les produits, mais il n'y a aucun sens à sommer sur la dimension temporelle des mois.

Mesure non additive - on ne peut pas sommer les valeurs des cellules en conservant un sens.

Ex: si le cube contient des moyennes de ventes par mois, produit et ville, il n'y a pas de sens à sommer les valeurs des cellules des villes pour les regrouper en départements.

Sol : Une bonne approche pour les non-additif faits est, si possible, de conserver les composants totalement additifs de la mesure non-additive.