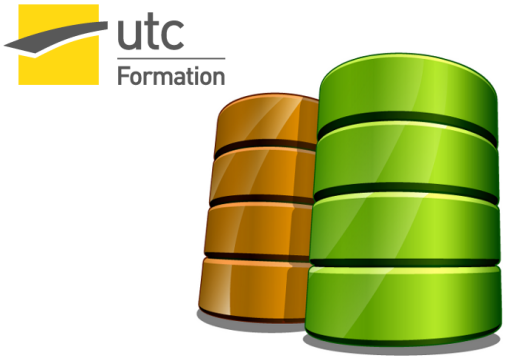
# Éléments avancés pour la modélisation des data warehouses

*Data warehouse*
*http://dwh.crzt.fr*




Stéphane Crozat

1 septembre 2016

# Table des matières

# Objectifs

- Savoir faire un modèle dimensionnel en étoile et en flocon
- Savoir gérer les dimensions dégénérées et attributs de documentation, de segmentation et d'agrégation

# Introduction

- Volume de cours : 1h
- Volume d'exercice : 3h

# Projet Fantastic : Rappel

I

### *Conseil : Pré-requis*

*http://dwh.crzt.fr/mod*

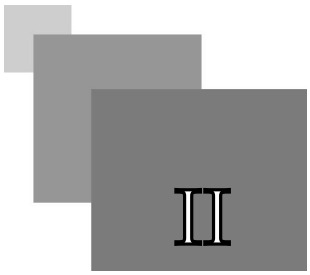
*http://dwh.crzt.fr/etl*

### *Rappel : Problème*

*Projet Fantastique : Problème posé* (cf. p.)

### *Rappel : Données sources*

*Projet Fantastic : Données disponibles* (cf. p.)



# Faits

## 1. Table de faits avec faits et table de faits sans fait

### *Définition : Table de faits avec faits*

En général la table des faits comporte un ou plusieurs attributs représentant des faits, que l'on va sommer sur les dimensions lors de l'analyse.

### *Exemple : Table des fait avec faits*

Par exemple une table des faits représentant des achats de produits pourra contenir la quantité de produits achetés, le chiffre d'affaire de la vente...

`ventes (fk_date, fk_produit, quantite, ca)`

```
1 SELECT sum(v.quantite)
2 FROM ventes v JOIN date d
3 ON fk_date=pk_date
4 GROUP BY d.week
```

### *Définition : Tables de faits sans fait (factless fact table)*

Dans certains cas, on mesure directement dans la table des faits des événements unitaires. Un fait est donc juste un enregistrement dans la table des faits.

Dans ce cas le table des faits ne contient que des clés étrangères, et aucun fait en tant que tel (c'est l'enregistrement qui est le fait).

### *Attention : Analyse en count*

Pour analyser une table de faits sans fait, on ne peut pas utiliser `sum` (il n'y a rien à sommer), on utilise `count` (on analyse le nombre de faits enregistrés).

### *Exemple : Tables de faits sans fait*

On enregistre ici directement des ventes de produits qui sont toujours vendus à l'unité, en vue d'une analyse en quantité (où le prix de vente n'intervient pas).

`ventes (fk_date, fk_produit)`

```
1 SELECT count(*)
2 FROM ventes v JOIN date d
3 ON fk_date=pk_date
```

```
4 GROUP BY d.week
```

### *Méthode : Ajouter une colonne avec la constante 1*

Afin de rendre ce cas plus lisible, il est parfois conseillé d'ajouter une colonne qui contient la valeur 1 pour toutes les lignes. On matérialise ainsi le fait par une valeur, même si elle est toujours la même, et il est de nouveau possible de travailler en somme.

### *Complément*

*http://www.kimballgroup.com/1996/09/factless-fact-tables/*

## 2. Clés artificielles

### *Méthode : Clé artificielles*

> Every join between dimension and fact tables in the data warehouse should be based on meaningless integer surrogate keys. You should avoid using the natural operational production codes. (Kimball, Ross, 2008, p59) *

- Les dimensions doivent être les points entrées dans les faits pour les utilisateurs, donc les clés naturelles n'apporte rien (aucune requête n'est faites directement dans la table des faits, sans jointure)
- L'usage de clés naturelle est plus simple au début, mais plus coûteux sur le long terme : les clés artificielles assurent *l'indépendance aux évolutions futures du système opérationnel* (on rappelle que le data warehouse vise le long terme, au delà de la durée de vie d'une version d'un système opérationnel typiquement)
- Les clés artificielles sont plus performantes (entiers compressés)
- les clés artificielles permettent de gérer les valeurs nulles (date...)
- ...

### *Méthode : OID*

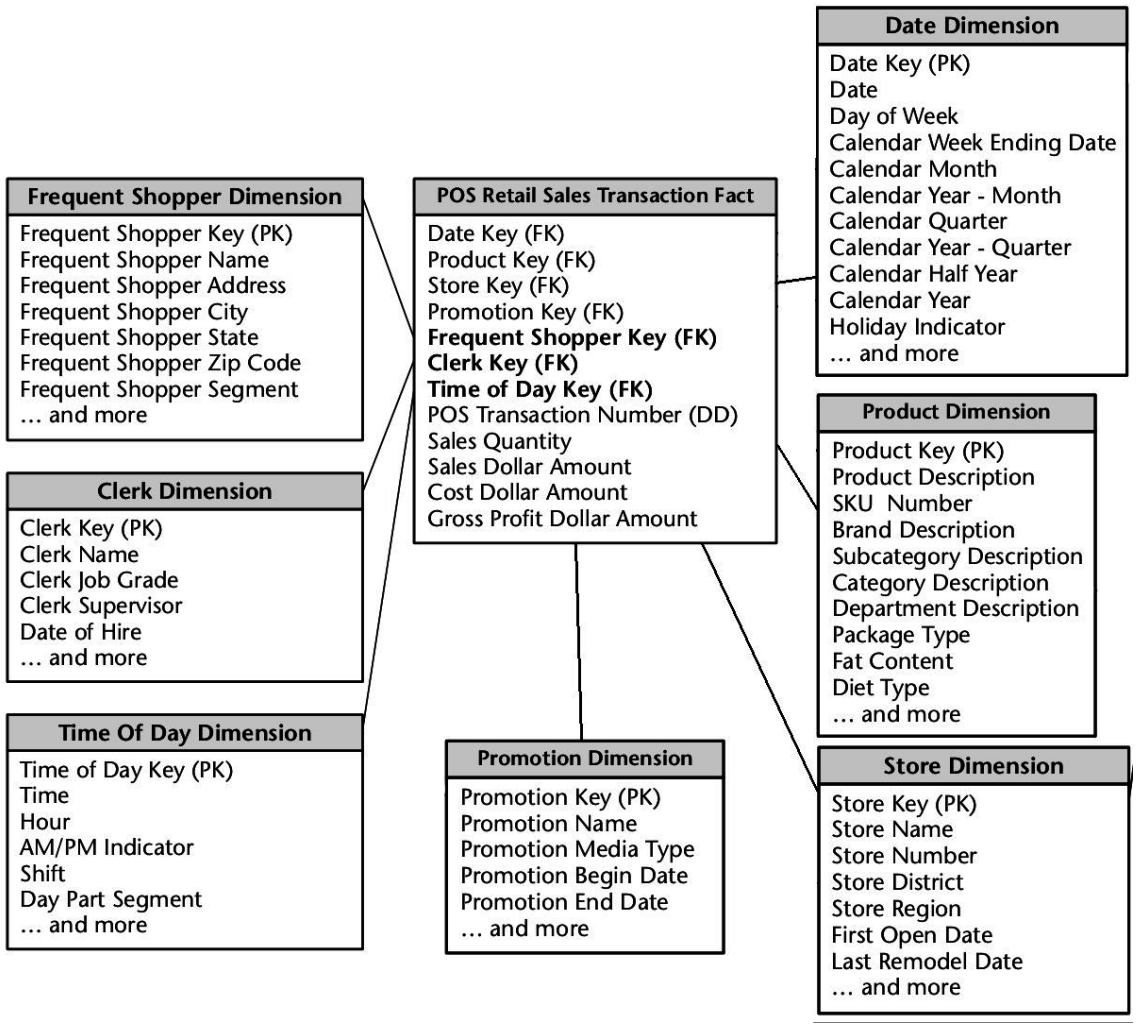
Sous un système relationnel-objet les OID peuvent être utilisés.

### *Méthode*

La mise en place de clés artificielles complique l'ETL et implique la maintenance d'une table de correspondance par exemple.

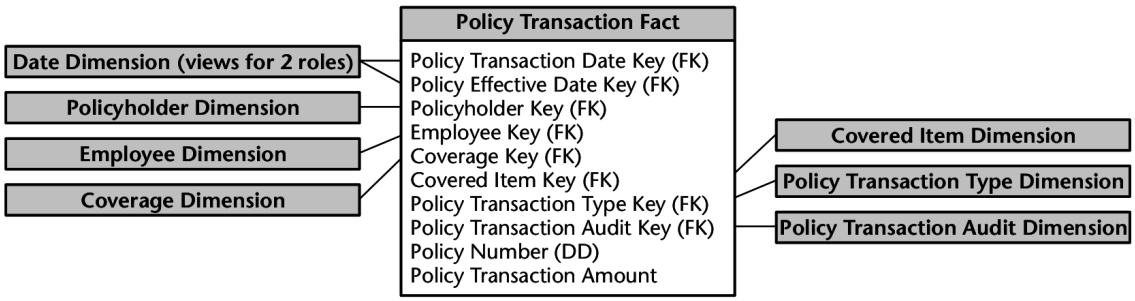
# 3. Exemples de modèles dimensionnels

*Exemple*



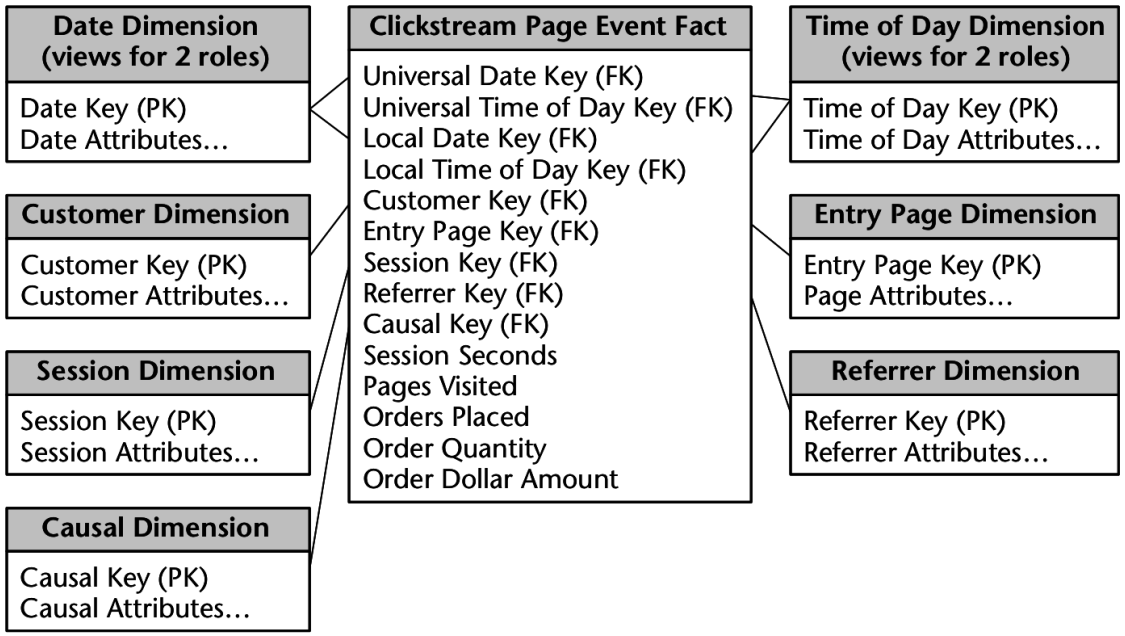

*Exemple de modèle dimensionnel d'analyse de ventes (Kimball, Ross, 2008, p.51-53)*

*Exemple*



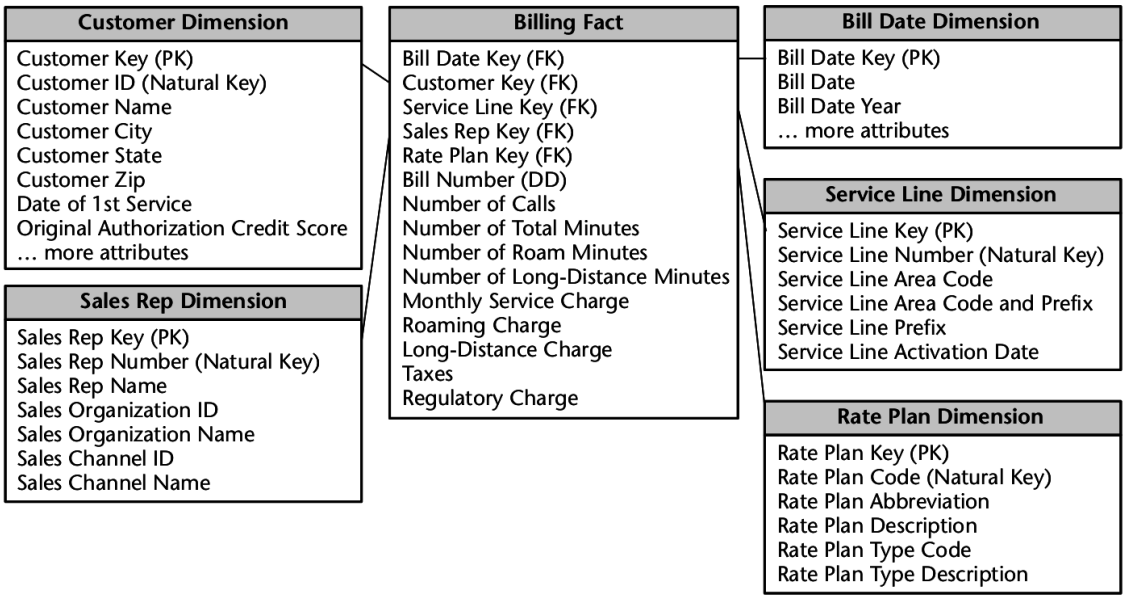

*Exemple de modèle dimensionnel dans le domaine de l'assurance (Kimball, Ross, 2008, p.314)*

*Exemple*



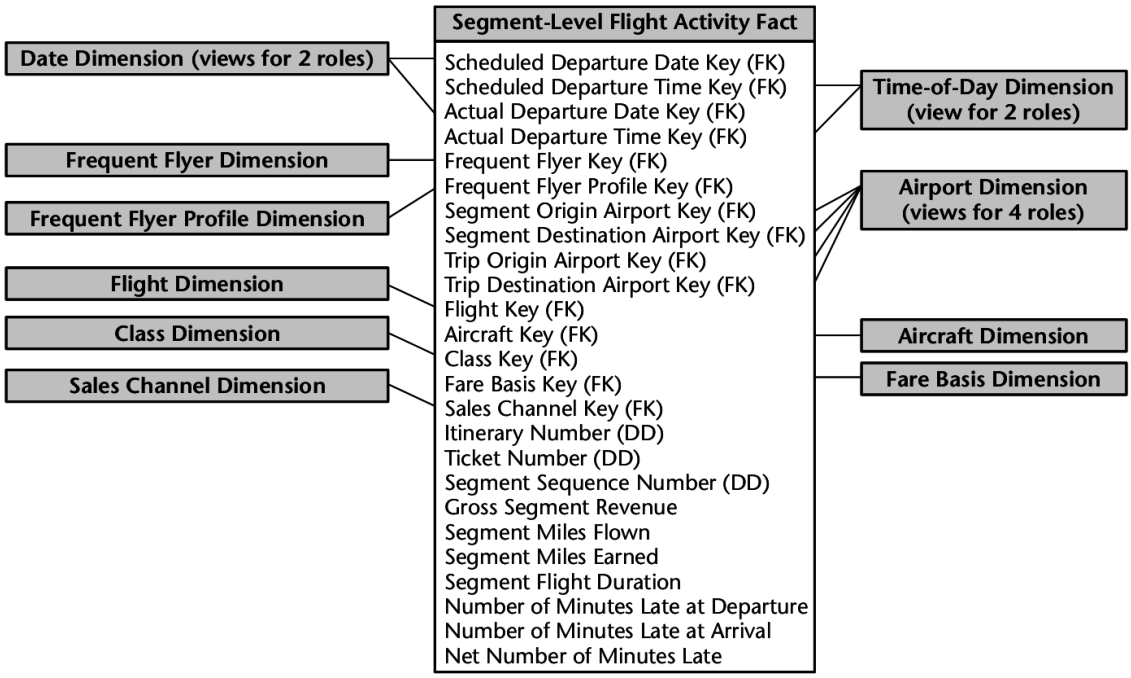

*Exemple de modèle dimensionnel dans le domaine du commerce électronique (Kimball, Ross, 2008, p. 293)*

*Exemple*




*Exemple de modèle dimensionnel dans le domaine des télécommunications (Kimball, Ross, 2008, p. 225)*

*Exemple*




*Exemple de modèle dimensionnel dans le domaine des transports (Kimball, Ross, 2008, p.233)*

# 4. Gestion des valeurs nulles

*Méthode*

> You must avoid null keys in the fact table. A proper design includes a row in the corresponding dimension table to identify that the dimension is not applicable to the measurement. (Kimball, Ross, 2008, p.49) *

*Exemple*

Lorsqu'un client qui ne possède pas de carte de fidélité achète un produit, il n'est pas possible de lier le fait à un client.

On évite de mettre une valeur nulle en ajoutant une valeur "Client sans carte de fidélité" à la dimension.

# 5. Gestion des erreurs

*Méthode*

Il est souvent utile d'ajouter des valeurs dans une dimension afin de gérer les cas d'erreur dans les données.

### *Exemple*

Ces valeurs rendent compte d'une typologie comme par exemple :

- Format invalide
- Valeur tronquée
- Référence inconnue
- ...

### *Remarque*

Si l'on dispose de suffisamment d'information sur la cause de l'erreur, cette cause peut être utilisée.

### *Complément : Voir aussi*

*Gestion des valeurs nulles* (cf. p.11)

## 6. Faits semi-additifs

### *Définition : Semiaddictive facts*

Un fait est semi-additif s'il est additif sur une partie seulement des dimensions du modèle.

> All measures that record a static level (inventory levels, financial account balances, ans measures of intensity such as room temperature) are inherently nonadditive across date dimension and possibly other dimensions. In these cases, the measure may be aggregated usefully across time, for example, by averaging over number of time periods. (Kimball, Ross, 2008, p72) *

### *Méthode*

Pour analyser les faits semi-additifs sur les dimensions sur lesquelles ils ne sont pas additifs, il faut faire des *moyennes*.

### *Exemple*

Modèle :

- Dimension compte
- Dimension date
- Fait (fkCompte, fkDate, solde)

Faits :

- (1,1,50)
- (1,2,100)

La somme (150) ne veut rien dire, la moyenne (75) donne bien le solde moyen du compte.


# Dimensions

## 1. Conception des dimensions

*Conseil*

> Most business processes can be represented with less than 15 dimensions (Kimball, Ross, 2008, p.58) *

*Méthode*

Il ne faut pas dénormaliser la table des faits :

- les dimensions sont indépendantes entre elles ;
- il ne faut pas représenter des hiérarchies différentes dans des dimensions différentes.

En particulier parce :

- pour des raisons d'intelligibilité ;
- et de performance (la table des faits est la plus volumineuse, elle doit être optimisée)

(Kimball, Ross, 2008, p57) *

## 2. Dimension dégénérée

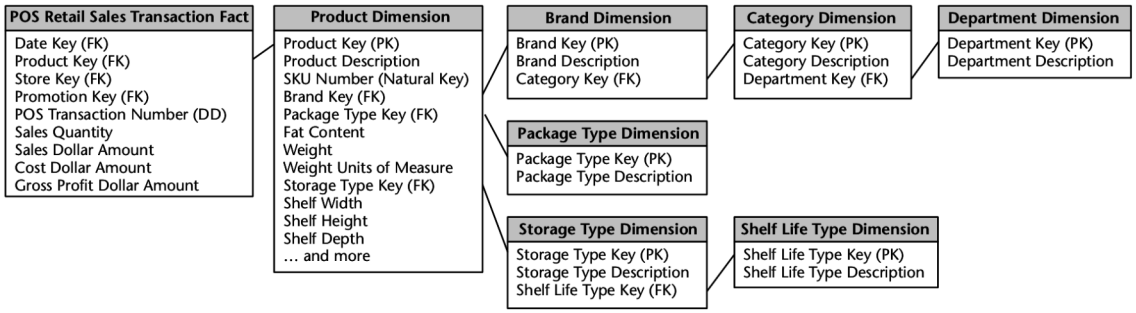
*Exemple*

> Operational control numbers such as order numbers, invoice numbers, and bill-of-lading numbers usually give rise to empty dimensions and are represented as degenerated dimensions (that is, dimension keys without corresponding dimension tables) [...] (Kimball, Ross, 2008, p.50) *

## 3. 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)*

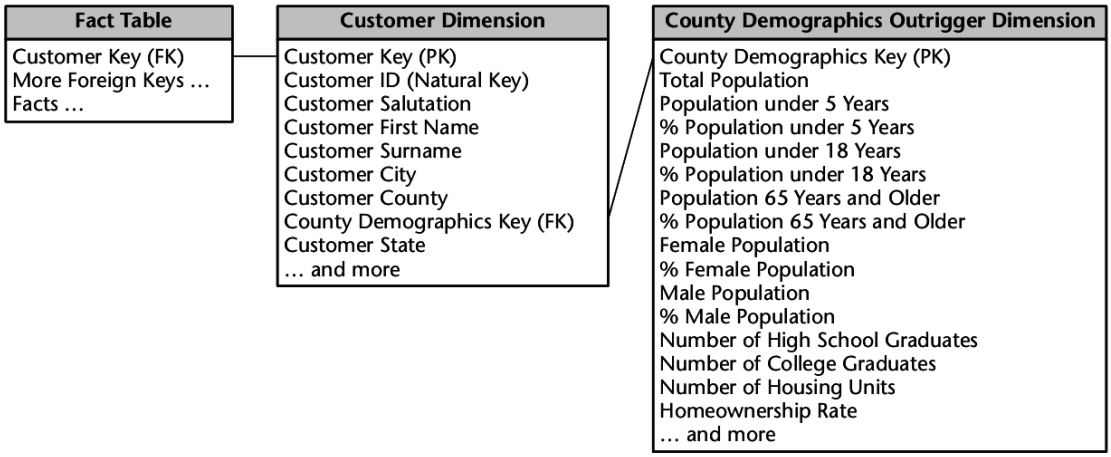
### ⚠ *Attention*

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

- 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) * .




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

# 4. Slow Changing Dimension (SCD)

La gestion des changements dans les dimensions est un enjeu de l'historisation dans le data warehouse.

Ces changements sont généralement lent, on parle de SCD.

## *Méthode*

Il y a 5 grands types de solution (Kimball, Ross, 2008, p.95) * :

- Type 1 : Remplacer la valeur (pas de gestion d'historique)
- Type 2 : Ajouter une nouvelle dimension (multiplication du nombre de lignes)
- Type 3 : Ajouter un attribut (gestion d'un seul niveau d'historique)
- Type 3b : Ajouter plusieurs attributs (changements prévisibles)
- Type 6 (1+2+3) : Combiner les type 1, 2 et 3

## *Exemple*

| Product Key | Product Description | Department | SKU Number (Natural Key) |
|---|---|---|---|
| 12345 | IntelliKidz 1.0 | Education | ABC922-Z |

| Product Key | Product Description | Department | SKU Number (Natural Key) |
|---|---|---|---|
| 12345 | IntelliKidz 1.0 | Strategy | ABC922-Z |

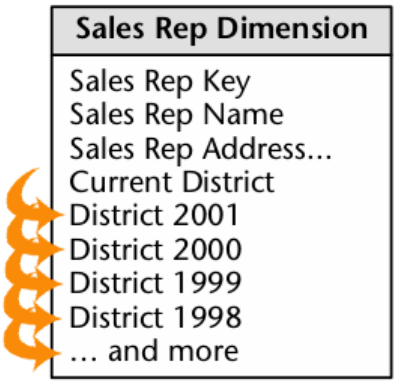
*SCD type 1 (Kimball, Ross, 2008, p.96)*

| Product Key | Product Description | Department | SKU Number (Natural Key) |
|---|---|---|---|
| 12345 | IntelliKidz 1.0 | Education | ABC922-Z |
| 25984 | IntelliKidz 1.0 | Strategy | ABC922-Z |

*SCD type 2 (Kimball, Ross, 2008, p.97)*

| Product Key | Product Description | Department | Prior Department | SKU Number (Natural Key) |
|---|---|---|---|---|
| 12345 | IntelliKidz 1.0 | Strategy | Education | ABC922-Z |

*SCD type 3 (Kimball, Ross, 2008, p.101)*




*SCD type 3+ (Kimball, Ross, 2008, p.103)*

| Product Key | Product Description | Current Department | Historical Department | SKU Number (Natural Key) |
|---|---|---|---|---|
| 12345 | IntelliKidz 1.0 | Education | Education | ABC922-Z |

| Product Key | Product Description | Current Department | Historical Department | SKU Number (Natural Key) |
|---|---|---|---|---|
| 12345 | IntelliKidz 1.0 | Strategy | Education | ABC922-Z |
| 25984 | IntelliKidz 1.0 | Strategy | Strategy | ABC922-Z |

| Product Key | Product Description | Current Department | Historical Department | SKU Number (Natural Key) |
|---|---|---|---|---|
| 12345 | IntelliKidz 1.0 | Critical Thinking | Education | ABC922-Z |
| 25984 | IntelliKidz 1.0 | Critical Thinking | Strategy | ABC922-Z |
| 31726 | IntelliKidz 1.0 | Critical Thinking | Critical Thinking | ABC922-Z |

*SCD type 6 (Kimball, Ross, 2008, p.104)*

# Attributs des dimensions

IV

## 1. Attributs d'analyse

*Définition : Attributs d'analyse*

La majorité des attributs d'une dimension qui servent à l'analyse (ils sont mobilisés dans les GROUP BY).

Synonyme : Attribut de regroupement

*Syntaxe*

Par défaut un attribut mentionné dans le modèle dimensionnel est un attribut d'analyse. Ces attributs sont notés tels quels, sans annotation ni style particulier.

## 2. Attributs de description

*Définition : Attributs de description*

Certains attributs ne sont pas utiles à l'analyse, mais peuvent être conservés dans le modèle, afin d'améliorer la qualité des états, souvent parce qu'ils sont plus explicites pour identifier un enregistrement d'une dimension.

Synonyme : Attribut de documentation

*Exemple : Numéro et nom de département*

Si l'on dispose d'un numéro de département pour l'analyse, le nom peut néanmoins être conservé à des fins d'amélioration des rapports.

*Syntaxe*

Les attributs de description sont notés en italique dans le modèle dimensionnel et/ou annotés de la mention (d).

# 3. Attributs de segmentation

*Définition*

Afin qu'un attribut soit utilisable en analyse, il doit disposer de valeurs discrètes (et non continues) ; et l'échelle de ces valeurs doivent avoir un niveau de précision en adéquation avec les besoins d'analyse.

Lorsque ce n'est pas le cas, l'attribut en question est remplacé par un attribut de segmentation qui projette les valeurs de l'attribut initial dans des segments discrets et utilisables.

*Exemple*

Population d'un département.

*Exemple*

- Age
- Revenus

(Kimball, Ross, 2008, p.151) *

*Méthode*

- Utiliser une segmentation reconnue (habitude socio-économique, pratique de l'organisation...)
- Utiliser une segmentation statistique (parts égales...)

*Syntaxe*

L'attribut est annoté d'un (s) dans le modèle dimensionnel.

# 4. Attributs d'agrégation de faits

*Définition*

Certaines analyses requièrent de regrouper les faits en fonctions de valeurs elles-mêmes issues des faits.

Dans ce cas des attributs d'agrégation des faits sont pré-calculés au sein des dimensions concernées.

*Exemple*

Les produits les plus vendus (livres best-sellers...)

*Exemple*

Les clients "high spender" qui plus dépensent le plus.

(Kimball, Ross, 2008, p.152) *

*Syntaxe*

L'attribut est annoté d'un (a) dans le modèle dimensionnel.

## 5. La dimension date

*Fondamental*

Quasiment tous les data warehouses ont une dimension date.

*Exemple*

| Date Dimension |
| --- |
| Date Key (PK) |
| Date |
| Full Date Description |
| Day of Week |
| Day Number in Epoch |
| Week Number in Epoch |
| Month Number in Epoch |
| Day Number in Calendar Month |
| Day Number in Calendar Year |
| Day Number in Fiscal Month |
| Day Number in Fiscal Year |
| Last Day in Week Indicator |
| Last Day in Month Indicator |
| Calendar Week Ending Date |
| Calendar Week Number in Year |
| Calendar Month Name |
| Calendar Month Number in Year |
| Calendar Year-Month (YYYY-MM) |
| Calendar Quarter |
| Calendar Year-Quarter |
| Calendar Half Year |
| Calendar Year |
| Fiscal Week |
| Fiscal Week Number in Year |
| Fiscal Month |
| Fiscal Month Number in Year |
| Fiscal Year-Month |
| Fiscal Quarter |
| Fiscal Year-Quarter |
| Fiscal Half Year |
| Fiscal Year |
| Holiday Indicator |
| Weekday Indicator |
| Selling Season |
| Major Event |
| SQL Date Stamp |
| … and more |

*Exemple de dimension Date (Kimball, Ross, 2008, p.39)*

# Exercice : Modélisation avancée du data warehouse

V

## [3h]

Afin d'améliorer les analyses du contexte "Fantastic", on va enrichir le modèle dimensionnel et l'implémenter.

## Données supplémentaire

Les données supplémentaires suivantes sont apportés au projet :

- fichier `Prices2014.csv` déposé dans un répertoire du serveur `sme-oracle.sme.utc` : `/home/nf26/fantastic2`
- fichier `Sales2014` déposé dans un répertoire du serveur `sme-oracle.sme.utc` : `/home/nf26/fantastic2`

**Question 1**

*[solution* n°1 p.22*]*

Améliorer le modèle dimensionnel afin d'ajouter :

- le numéro de ticket(rappeler pourquoi c'est une dimension dégénérée) ;
- le fait quantité (que l'on fixe toujours à 1, ou bien que l'on calcule on regroupant les lignes strictement identiques)
- le fait chiffre d'affaire de la vente que l'on récupère du fichier des prix (on fera l'hypothèse que le prix de vente est toujours le prix enregistré dans ce fichier, on pensera à multiplier le prix par la quantité)
- les attributs apportés par les nouvelles données
- des attributs de documentation (nom du département, genre...)
- des attributs de segmentation (population, age de publication...)
- un attribut d'agrégation pour savoir si un livre est un best-seller ou non

**Question 2**

Implémenter le modèle dimensionnel et modifier l'ETL en conséquence.

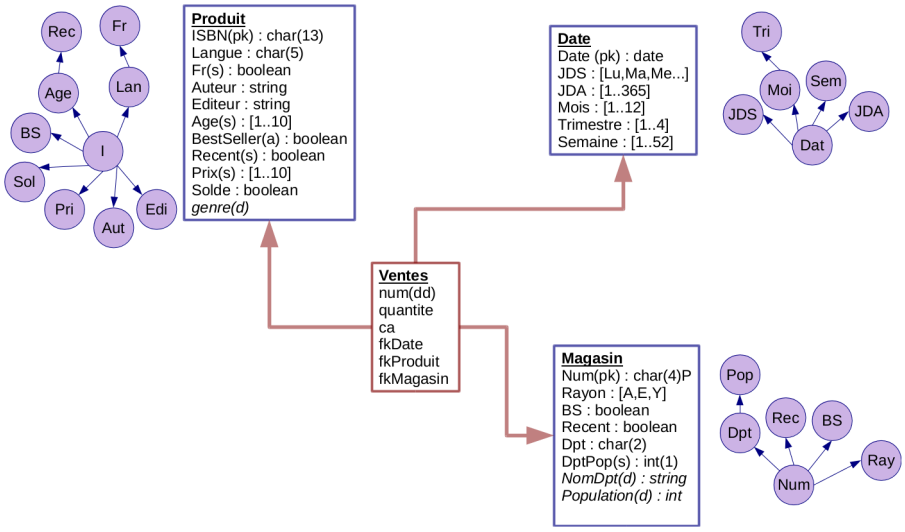
**Question 3**

Expérimenter le nouveau modèle avec des questions en rapport avec les modifications apportées.

# Solutions des exercices

## > **Solution** n°1

Exercice p. 21




*Modèle dimensionnel du DataWarehouse Fantastic*

# Bibliographie

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