Entrepôts de données

Objectifs

Définitions

Les bases transactionnelles

Les entrepôts de données

La dimension temps

La representation de l'histoire:
•par le systèmes transactionels - les bases scintillantes
•par les entrepôts de données - instantanés statiques

*Le modèle dimensionnel(**schéma des jointures en étoile**)*

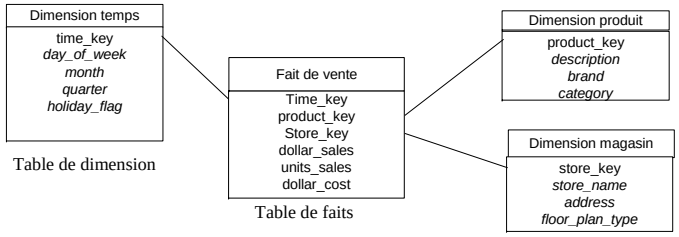


Table de dimension

Table de faits

Définir ce que nous appelons la **finesse** ou le **grain** de la table de faits

La table de faits

La table de faits sert à stocker les mesures de l'activité. Chacune de ces mesures est prise à l'intersection de toutes les dimensions. Les tables de faits sont toujours **éparses** (en anglais *sparse).*

Types des faits :

- Les faits les plus importants et les plus utiles sont numériques, **valorisés de façon continue** (en anglais *contmuously valued)* et ils sont **additifs** (montant des ventes**)**
- **semi-additifs** (précipitation)
- **non additifs** (vitesse du vent, température)

Les tables de dimension

Les tables de dimension servent à enregistrer les descriptions textuelles des **dimensions** de l'activité. Chacune de ces descriptions textuelles concerne un membre de la dimension concernée. Dans une base de données bien conçue, la table de la dimension produit a de nombreux **attributs**. Les attributs les plus intéressants sont **textuels, discrets** et sont utilisés comme source de **contraintes** et **d'en-têtes de ligne** dans le jeu de réponses de l'utilisateur.

Un Exemple : Les magasins d'alimentation

Les étapes du processus de conception:

1. Choisir le processus d'activités modéliser.
2. Choisir le grain du processus d activité.
3. Choisir les dimensions applicables à chaque enregistrement de la table de faits.
4. Choisir les faits mesurés que contiendra chaque enregistrement de la table de faits.

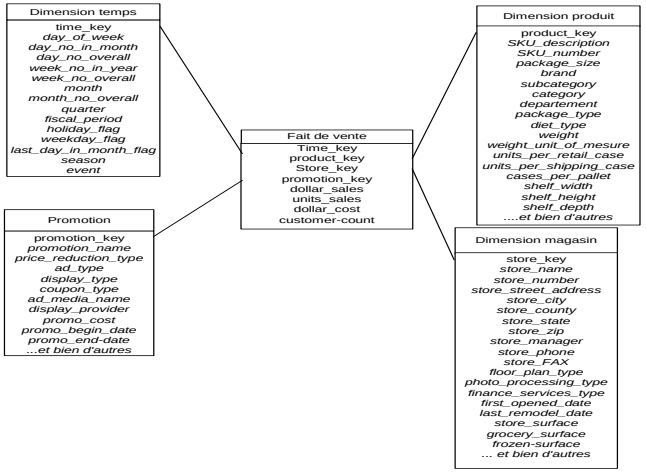
Résister à la normalisation

Ici règnent trois principes :

La table de faits d'un schéma dimensionnel atteint naturellement un haut degré de normalisation.

Tout effort de normalisation des tables d'une base de données dimensionnelle dans le but d'économiser l'espace disque est une perte de temps.

Les tables de dimension ne doivent pas être normalisées mais rester des tables plates. Des tables de dimension normalisées ne sont plus navigables. Les gains d'espace disques permis par la normalisation sont typiquement de l'ordre de un pour cent de l'espace disque total nécessaire à l'ensemble du schéma de données.



Les dimensions

- ***Dimension Temps***
- ***Dimension Produit***
- ***La dimension Magasin***
- ***La dimension Promotion***

Calcul du volume de la base de données pour la chaîne de magasins d'alimentation

Dimension temps : 2 ans x 365 jours = 730 jours Dimension magasin : 300 magasins, enregistrant des ventes chaque jour

Dimension produit : 30.000 produits dans chaque magasin, parmi lesquels 3.000 sont vendus chaque jour dans un magasin donné.

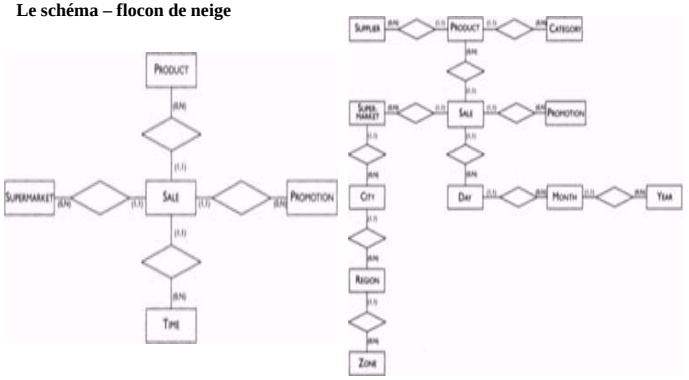
Dimension promotion : un article vendu n'apparaît que dans une seule condition de promotion dans un magasin donné un jour donné

Enregistrements de faits élémentaires - -30 x 300 x 3000 x 1 = *million d'enregistrements*

Nombre de champs de clé = 4; nombre de champs de fait = 4 ; nombre total de champs =8

Taille de la table de faits élémentaires - 657millions x 8 champs x 4 octets - *21 GO*

Le schéma – flocon de neige



Opérations pour l'analyse des données

La forme générale d'une instruction SQL

select D1.C1, ... Dn.Cn, Aggr1(F,Cl), *Aggrn(*F,Cn*)*
from *Fact* **as F,** *Dimension1* ***as*** **D1,...** *DimensionN*
as Dn
where *join-condition* **(F, D1)**
and...
and *join-condition* **(*F*, Dn)**
and *selection-condition*
group by D1.C1, ... Dn.Cn
order by D1.C1, ... Dn.Cn

Exemple:

```
select Time.Month, Product.Name, sum(Qty)
from Sale, Time, Product, Promotion
where Sale.TimeCode = Time.TimeCode
        and Sale.ProductCode = Product.ProductCode
        and Sale.PromoCode = Promotion.PromoCode
        and (Product. Name = ' Pasta' or Product.Name = 'Oil')
        and Time.Month between 'Feb' and 'Apr'
        and Promotion.Name = 'SuperSaver'
group by Time.Month, Product.Name
order by Time.Month, Product.Name
Pivot Time.Month
```

	Feb	Mar	Apr
Oil	5K	5K	*7K*
Pasta	45K	50K	51K

Drill-down et roll-up

Roll-up par mois

Time. Month	Product.Name	sum(Qty)
Feb	Pasta	45K
Mar	Pasta	50K
Apr	Pasta	51K

Product.Name	Zone	sum(Qty)
Pasta	North	54K
Pasta	Centre	50K
Pasta	South	42K

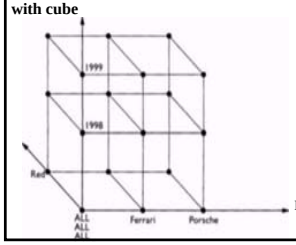
TIme.Monih	Product.Name	Zone	sum(Qty)
Feb	Pasta	North	18K
Feb	Pasta	Centre	18K
Feb	Pasta	South	12K
Mar	Pasta	North	18K
Mar	Pasta	Centre	18K
Mar	Pasta	South	14K
Apr	Pasta	North	18K
Apr	Pasta	Centre	17K
Apr	Pasta	South	16K

Drill-down par zones

Cube des données

Make	Year	Colour	Sales
Ferrari	1998	Red	50
Ferrari	1999	Red	85
Porsche	1998	Red	80

```
select Make, Year, Colour, sum(Sales)
from Sales
where (Make = 'Ferrari' or Make = 'Porsche')
        and Colour = 'Red'
        and Year between 1998 and 1999
group by Make, Year, Colour
with cube
```



Make	Year	Colour	sum(Sales)
Ferrari	1998	Red	50
Ferrari	1999	Red	85
Ferrari	1998	ALL	50
Ferrari	1999	ALL	85
Ferrari	ALL	Red	135
Ferrari	ALL	ALL	135
Porsche	1998	Red	80
Porsche	1998	ALL	80
Porsche	ALL	Red	80
Porsche	ALL	ALL	80
ALL	1998	Red	130
ALL	1999	Red	8S
ALL	ALL	Red	215
ALL	1998	ALL	130
ALL	1999	ALL	85
ALL	ALL	ALL	215


```
select Make, Year, Colour, sum(Sales)
from Sales
where (Make = 'Ferrari' or Make = 'Porsche')
        and Colour = 'Red'
        and Year between 1998 and 1999
group by Make, Year, Colour
with roll up
```

Make	Year	Colour	sum(Sales)
Ferrari	1998	Red -	50
Ferrari	1999	Red	85
Porsche	1998	Red	80
Ferrari	1998	ALL	50
Ferrari	1999	ALL	85
Porsche	1998	ALL	80
Ferrari	ALL	ALL	135
Porsche	ALL	ALL	80
ALL	ALL	ALL	215

Data mining

Le processus de data mining

1. *Compréhension des données*: c'est impossible d'extraire d'information utile sans une bonne compréhension du domaine d'application.
2. Préparation de l'ensemble de donnès: c'est l'identification d'un sous-ensemble des données d'un entrepôt de données pour l'analyse. On doit encore coder les données dans un forme convenable pour l'algorithme de data mining.
3. Découverte des modèles: On essaie de découvrir des modèles répétés de données.
4. Evaluation des données: Cette étape concerne de tirer des implications depuis les modèles découverts en planant les expérimentes et en formulant des hypothèses.

Problèmes de data mining

1. Découvrir les règles d'association
2. Discrétisation
3. Classification

Découverte de règles d'association.

On cherche des modèles réguliers parmi les données comme la présence des deux choses dans un groupe des tuples. Exemples: L'analyse de corbeille des marchandises – de trouves les articles qui sont achetés ensemble. Chaque règle comporte prémisse et conséquence (ski – bâton de ski)

Transaction	Date	Goods	Qty	Price
1	17/12/98	ski-pants	1	140
1	17/12/98	boots	1	180
2	18/12/98	T-shirt	1	25
2	18/12/98	jacket	1	300
2	18/12/98	boots	1	70
3	18/12/98	jacket	1	300
4	19/12/98	jacket	1	300
4	19/12/98	T-shirt	3	25

Mesurer les règles
Support : c'est la partie d'observations qui satisfait la prémisse et la conséquence.
Confidence: c'est la partie d'observations qui satisfait la conséquence parmi les observations qui satisfont la prémisse.

Prémisse	Conséquence	Support	Confidence
ski-pants	boots	0.25	1
boots	ski-pants	0.25	0.5
T-shirt	boots	0.25	0-5
T-shirt	jacket	0.5	1
boots	T-shirt	0.25	0.5
boots	jacket	0.25	0.5
jacket	T-shirt	0.5	0.66
jacket	boots	0.25	0.33
{T-shirt, boots}	jacket	0.25	1
{T-shirt, jacket}	boots	0.25	0.5
{boots. jacket}	T-shirt	0.25	1

Discrétisation

C'est présenter un intervalle continu de valeurs par quelque valeurs discrets, par exemple la tension du sang par basse, normale, haute.

Classification

C'est de cataloguer un phénomène dans une classe prédéfinie. Le phénomène est présenté en général comme une tuple. L'algorithme de classification s'est construit automatiquement en utilisant un ensemble de données qui contient déjà des données classifiées et il est présenté par un arbre de décision.

Exemple: le risque d'une police d'insurence. On suppose qu'il y a un haut risque si le conducteur est moins agé de 23 ans ou la véhicule est de sport ou camion.

Policy(Number,Age,AutoType)