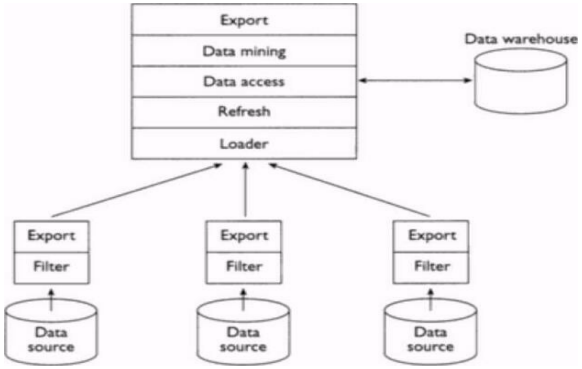
Analyse des données

Entrepôts de données

Traitement des données

- Types de traitement de données
  - OLTP (*On Line Transaction Processing*)
  - OLAP (*On-line Analytical Processing*)
- Types de bases de données
  - Transactionnelles
    - Usagers nombreux
    - Dynamique (scintillant)
    - Maintenir l'état de données actuel
    - Critiques (très chargé)
  - Entrepôts de données
    - Peu d'usagers (analyseurs)
    - Relativement stabile
    - Maintenir l'histoire des données (les états différents pendant une période)
    - Pas chargé

Architecture



Architecture

- Composant du source des données
  - Filtre – séparer et valider la cohérence des données exportées
  - Exporte – transmettre de portions de données en moments précis.
- Composant de l'entrepôt
  - Chargeur – initialiser et stoker les données initiales
  - Rafraîchir – télécharger les portions
  - Accès aux données
  - Data mining – analyse des données
  - Exporte – vers autres entrepôts. Former une hiérarchie des entrepôts

Schéma relationnel

- Etoile

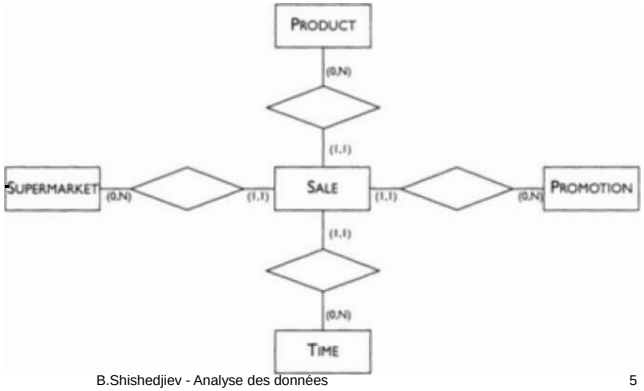
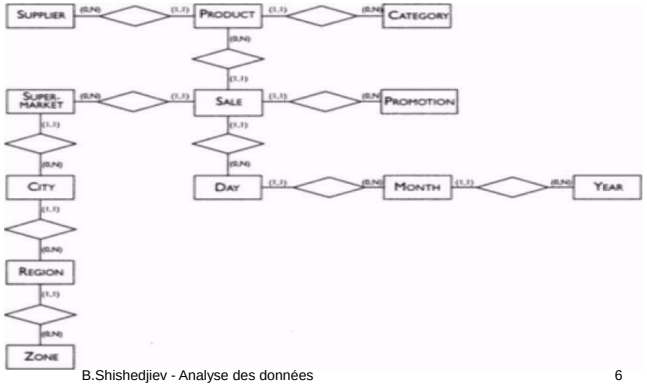


Schéma relationnel

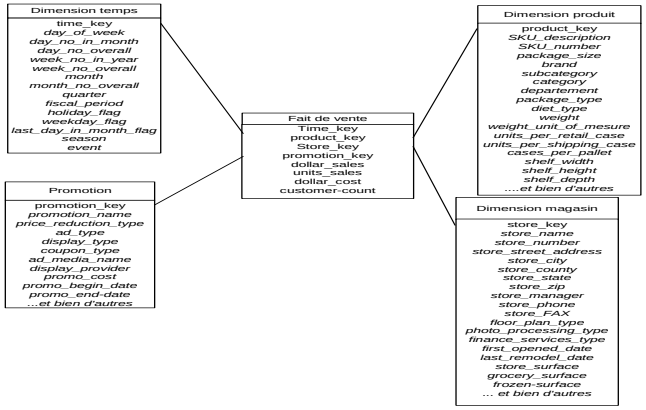
- Flocon de neige



Conception d'un entrepôt

- Etapes
  - Choisir les activités à modeler
  - Choisir la granularité de l'activité.
  - Choisir les dimensions qui identifient chaque enregistrement du tableau des fait
  - Choisir les faits stockés dans le tableau des faits (les valeurs, les unités)

Exemple – chaîne des supermarchés



Conception d'un entrepôt

- Types des faits – les valeurs numériques continues sont les plus utiles
  - Additives
  - Semi-additives
  - Non additives
- Tableaux des dimensions – non-normalisés
  - Attributs – nombreux, textuels, discrets.

Conception d'un entrepôt

- Recommandations
  - Utiliser des faits additive dont les valeurs sont numériques continues
  - Le tableau des faits est normalisé
  - Les dimensions ne sont pas normalisées. Le profit de normalisation est < 1%
  - Concevoir soigneusement les attributs de la dimension. Le plus souvent ils sont textuels et discret. Ils sont utilisés comme en-têtes et des sources de contraintes dans les réponses aux utilisateurs

Conception d'un entrepôt

- ***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*

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

Opérations pour l'analyse des données

- **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
```

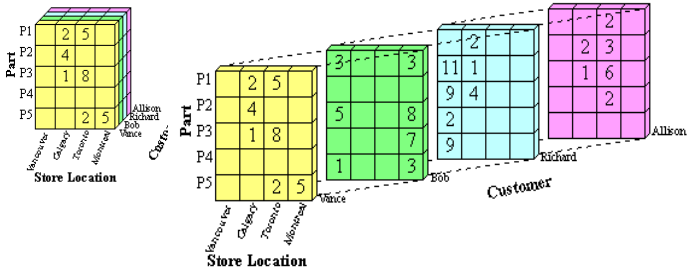
	Feb	Mar	Apr
Oil	5K	5K	7K
Pasta	45K	50K	51K

Cube des données

Le cube est utilisé de présenter les donner en respect certaine mesure d'intérêt. Bien que appelé un "cube", il peut être à deux dimensions, en trois dimensions, ou ultérieure-dimensionnelle. Chaque dimension représente une certaine attribut dans la base de données et de les cellules dans le cube de données représentent la mesure d'intérêt.

```
select Customer,  Part,  Location, sum(Sales)
from Sales S, Customers C, Locs L, Parts P
Where S.CustNo = C.CustNo and
S.LocNo=L.LocNo and
P.Partno=S.PartNo
group by Customer, Year, Location
with cube
```


Cube des données



Cube des données

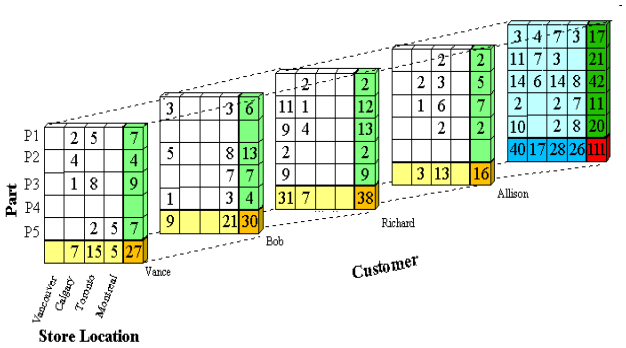
- Representation cube

Combination	Count
{P1, Calgary, Vance}	2
{P2, Calgary, Vance}	4
{P3, Calgary, Vance}	1
{P1, Toronto, Vance}	5
{P3, Toronto, Vance}	8
{P5, Toronto, Vance}	2
{P5, Montreal, Vance}	5
{P1, Vancouver, Bob}	3
{P3, Vancouver, Bob}	5
{P5, Vancouver, Bob}	1
{P1, Montreal, Bob}	3
{P3, Montreal, Bob}	8
{P4, Montreal, Bob}	7
{P5, Montreal, Bob}	3
{P2, Vancouver, Richard}	11

Combination	Count
{P3, Vancouver, Richard}	9
{P4, Vancouver, Richard}	2
{P5, Vancouver, Richard}	9
{P1, Calgary, Richard}	2
{P2, Calgary, Richard}	1
{P3, Calgary, Richard}	4
{P2, Calgary, Allison}	2
{P3, Calgary, Allison}	1
{P1, Toronto, Allison}	2
{P2, Toronto, Allison}	3
{P3, Toronto, Allison}	6
{P4, Toronto, Allison}	2

Cube des données

- Totales – les valeurs ANY ou ALL ou NULL



Cube des données

- 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
With cube
```


Cube des données

- Le cube entier

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
ALL	Pasta	North	54K
ALL	Pasta	Centre	53K
ALL	Pasta	South	42K
Feb	Pasta	ALL	48K
Mar	Pasta	ALL	50K
Apr	Pasta	ALL	51K
ALL	Pasta	ALL	149K
ALL	ALL	ALL	149K

Cube des données

- Roll-up - supprimer une dimension

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

Product.Name	Zone	sum(Qty)
Pasta	North	54K
Pasta	Centre	53K
Pasta	South	42K

Cube des données

- Drill down – additionner une dimension

Time. Month	Product.Name	sum(Qty)
Feb	Pasta	48K
Mar	Pasta	50K
Apr	Pasta	51K

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

Data mining (Analyse des données)

- **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 données: 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

Fouille de données (Data mining)

- **Problèmes de data mining**
  - Découvrir les règles d'association
  - Discrétisation
  - Classification

Data mining

- 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

Data mining

- 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.

Transaction	Date	Articles	Qté	Prix
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

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

Data mining

- 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)
```