http://www-adele.imag.fr/~donsez/cours

Systèmes d'information décisionnels
(Data Warehouse / Data Mining)

E. GRISLIN-LE STRUGEON

Université de Valenciennes, ISTV

Emmanuelle.Grislin@univ-valenciennes.fr

D. DONSEZ

Université Joseph Fourier, IMA

Didier.Donsez@imag.fr

1996-2006

Plan

- 1. Introduction
  - » Problématique- Le Système d'Information - La Suite Décisionnelle
- 2. L'Entrepôt de Données
  - » Extraction des données - Constitution de l'entrepôt - Modélisation
- 3. Les Bases Multidimensionnelles
  - » Analyse multidimensionnelle - OLAP - Data Marts
- 4. La Restitution des Informations
  - » Data Mining
- 5. La Gestion de Projet Data Warehouse
- 6. Les outils
- 7. Perspectives du Data Warehouse
- 8. Conclusion et Bibliographie

1. Introduction - Problématique

- Objectif
  - » Améliorer les performances décisionnelles de l'entreprise
- Comment ?
  - » en répondant aux demandes d'analyse des décideurs
- Exemple
  - » clientèle : Qui sont mes clients ? Pourquoi sont-ils mes clients ? Comment les conserver ou les faire revenir ? Ces clients sont-ils intéressants pour moi ?
  - » marketing, actions commerciales : Où placer ce produit dans les rayons ? Comment cibler plus précisément le mailing concernant ce produit ?
  - » ...

1. Introduction - Problématique

- Une grande masse de données :
  - Distribuée
  - Hétérogène
  - Très Détaillée
- A traiter :
  - Synthétiser / Résumer
  - Visualiser
  - Analyser
- Pour une utilisation par :
  - des experts et des analystes d'un métier
  - NON informaticiens
  - NON statisticiens

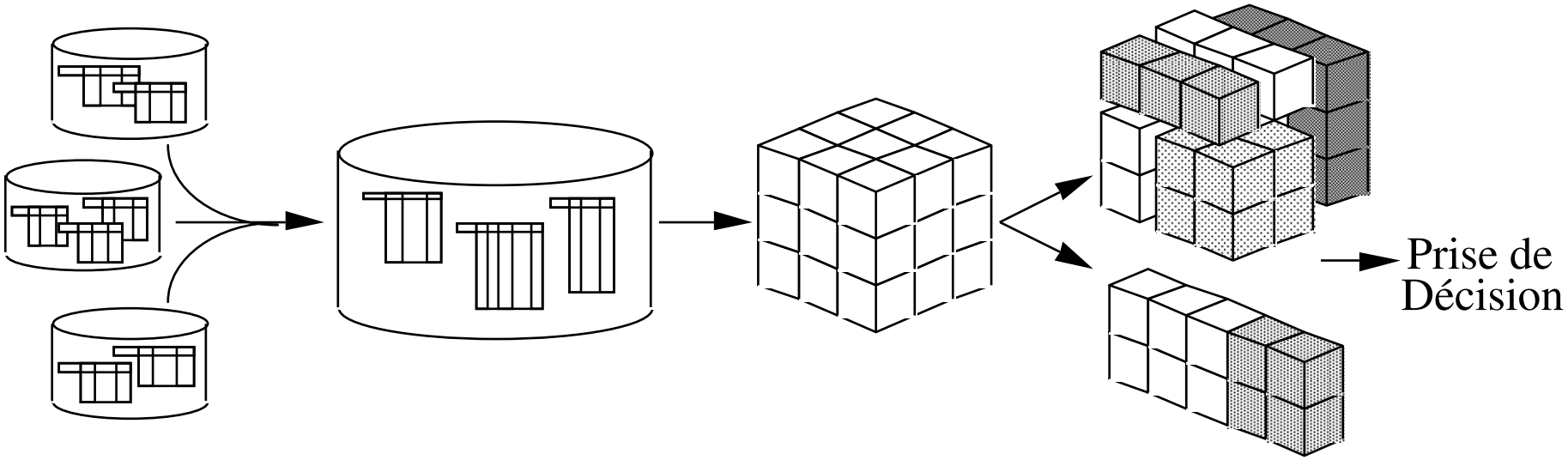
1. Introduction - Le système d'information

Moyen d'atteindre ces objectifs :

Le Data Warehouse, un système d'information dédié aux applications décisionnelles

- En **Aval** des bases de production
  (ie bases opérationnelles)
- En **Amont** des prises de décision
  - » basé sur des indicateurs (*Key Business Indicators* (KBI))

1. Introduction - La Suite Décisionnelle



1. Introduction - Utilisation

- Mailing
  - » amélioration du taux de réponse
- Banque, Assurance
  - » déterminer les profils client
    - Risque d'un Prêt, Prime plus précise
- Commerce
  - » ciblage de clientèle
  - » déterminer les promotions
  - » aménagement des rayons (2 produits en corrélation)

1. Introduction - Utilisation

- Logistique
  - » adéquation demande / production
- Santé
  - » épidémiologie (VIH, Amiante, ...)
- Econométrie
  - » prédiction de trafic autoroutier
- Ressources Humaines
  - » adéquation activité / personnel

Déclinaisons métiers du Décisionnel

- SPM (Strategic Performance Management)
  - » Déterminer et contrôler les indicateurs clé de la performance de l'entreprise
- FI (Finance Intelligence)
  - » Planifier, analyse et diffuser l'information financière. Mesurer et gérer les risques.
- HCM (Human Capital Management)
  - » Aligner les stratégies RH, les processus et les technologies. Modéliser la carte des RH (Ressources Humaines)
- CRM (Customer Relationship Management)
  - » Améliorer la connaissance client, Identifier et prévoir la rentabilité client. Accroître l'efficacité du marketing client.
- SRM (Supplier Relationship Management)
  - » Classifier et évaluer l'ensemble des fournisseurs. Planifier et piloter la stratégie Achat.

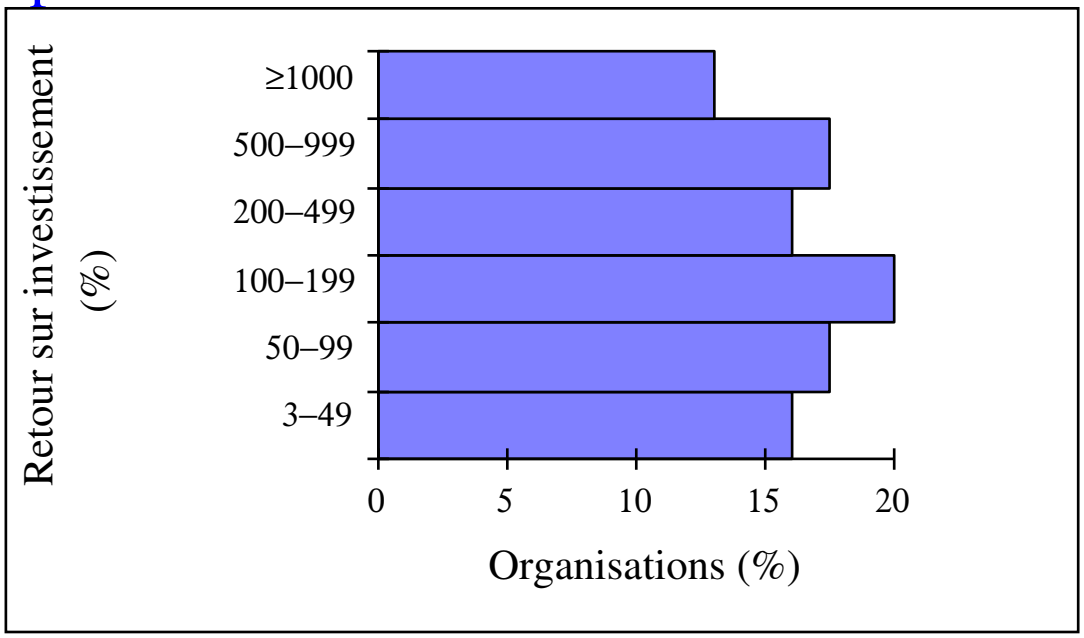
1. Introduction - RSI

D'après une enquête de l'IDC auprès de 45 organisations ayant un Data Warehouse en fonctionnement *(fin 1995-1996) :*

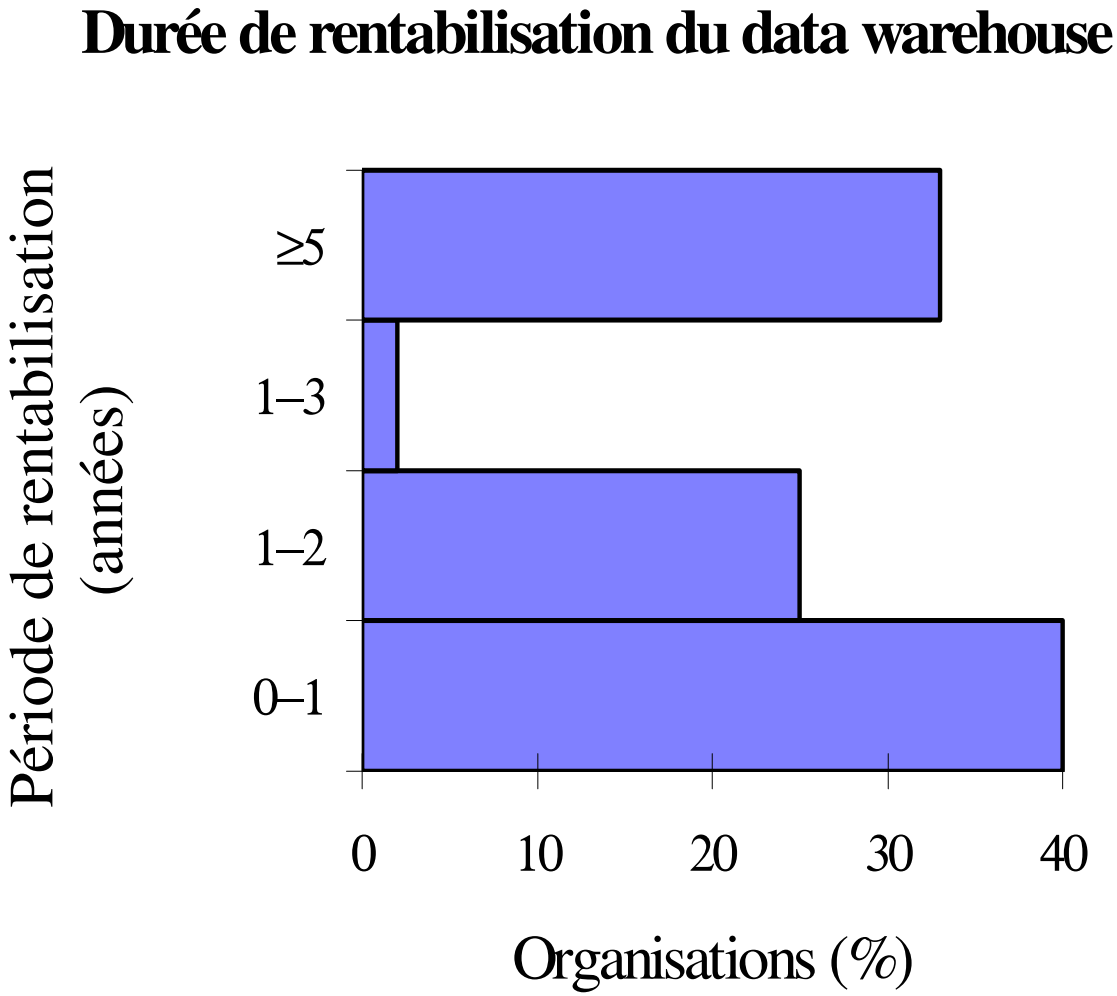
- 90% des entreprises ont un RSI au moins égal à 40%
- 50% ont un RSI supérieur à 160%
- 25% ont un RSI supérieur à 600%

Retour sur investissement du datawarehousing

RSI moyen = 401%
RSI médian = 167%



1. Introduction - Rentabilisation



1. Introduction - Rentabilisation

- Constat: orientation marché (client, techno, produit)
  - » Stratégies proactive meilleur que des stratégies réactives
  - » Cf livre de David Gotteland

2. L'Entrepôt de Données (Data Warehouse)

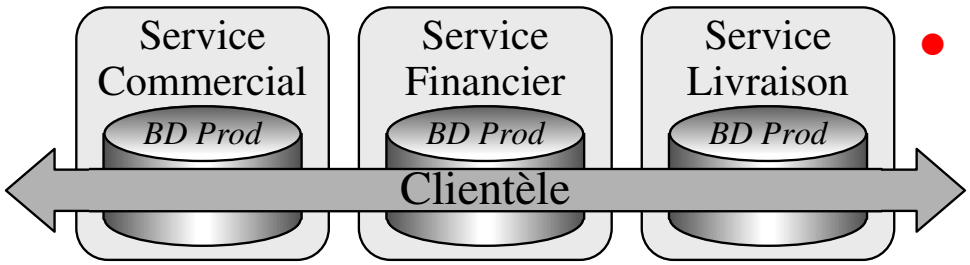
- Définition de Bill Inmon (1996)

  «Le Data Warehouse est une collection de données orientées sujet, intégrées, non volatiles et historisées, organisées pour le support d'un processus d'aide à la décision.»

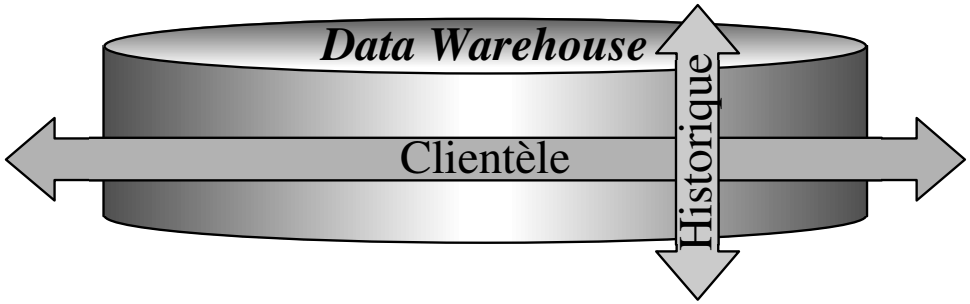
- Principe
  - Base de Données utilisée à des fins d'analyse.
  - Caractéristiques :
    - orientation sujets («métiers»)
    - données intégrées
    - données non volatiles
    - données datées

2. L'Entrepôt de Données (Data Warehouse)

- Objectif
  - » Retrouver une information historique et transversale à l'entreprise
- Comment
  - »Fédérer/Regrouper l'ensemble des données de l'entreprise



- Données réparties
- Vue «au-jour-le-jour»

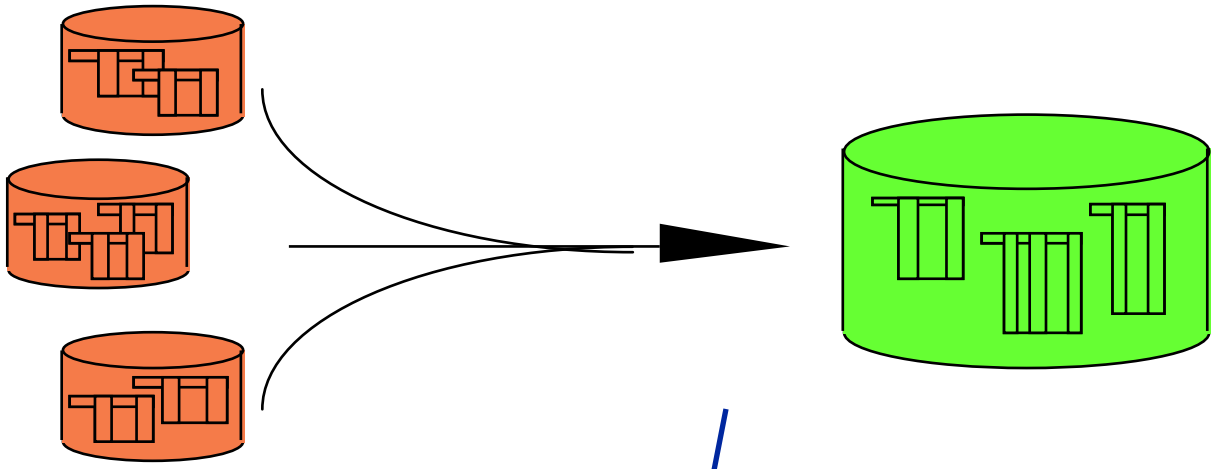


- Recoupements d'informations
- Vue sur l'évolution des informations

2. DW - OLTP versus DW

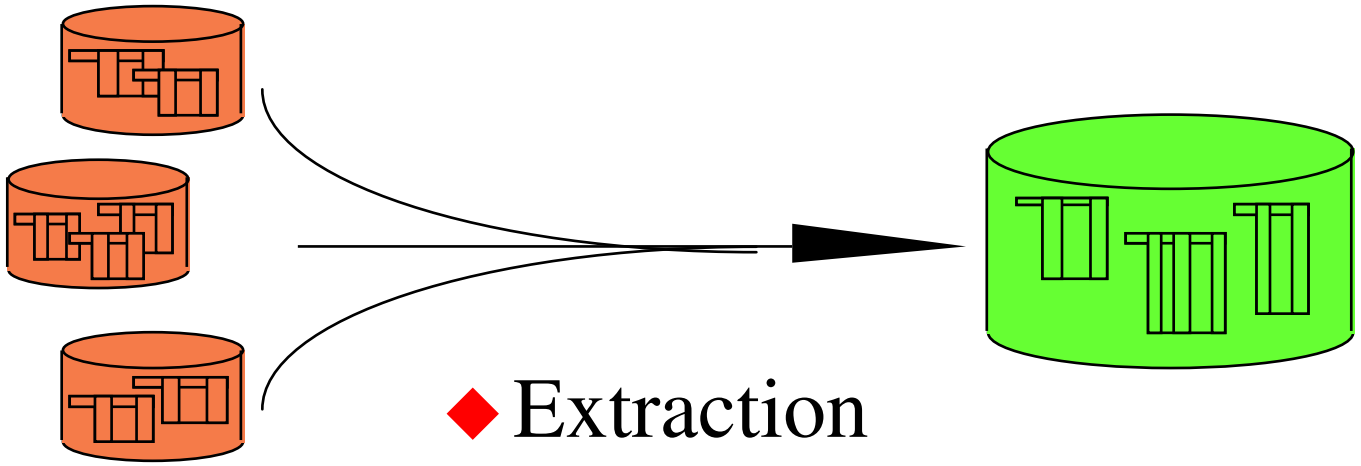
	Bases de Production (OLTP)	Entrepôt de Données (DW)
Données	•atomiques •orienté application •à jour •dynamiques	•résumés •orienté sujet •historiques •statiques
Utilisateurs	•employés de bureau •nombreux •concurrents •mises à jour •requêtes prédéfinies •réponses immédiates	•analystes •peu •non concurrents •interrogations •requêtes " one-use" •réponses moins rapides
	•accès à peu de données	•accès à beaucoup d'information

2. DW - OLTP → DW



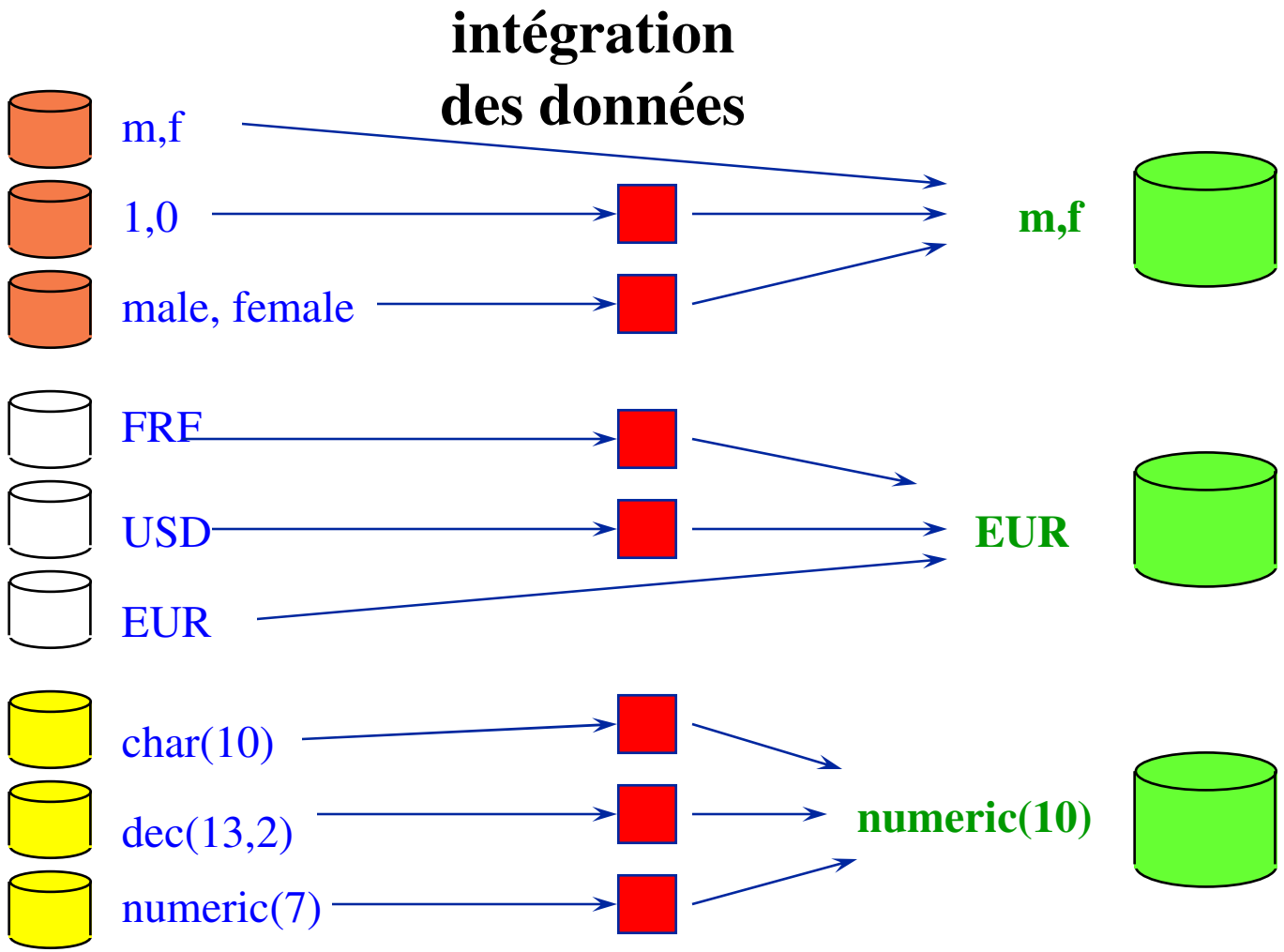
- Données de production :
  - SGBD et supports physiques hétérogènes
  - Qualité inégale des données
  - Représentations hétérogènes
- Objectif d'obtention de données :
  - centralisées
  - fiables
  - interprétables

2. DW - Alimentation (ETL) du DW



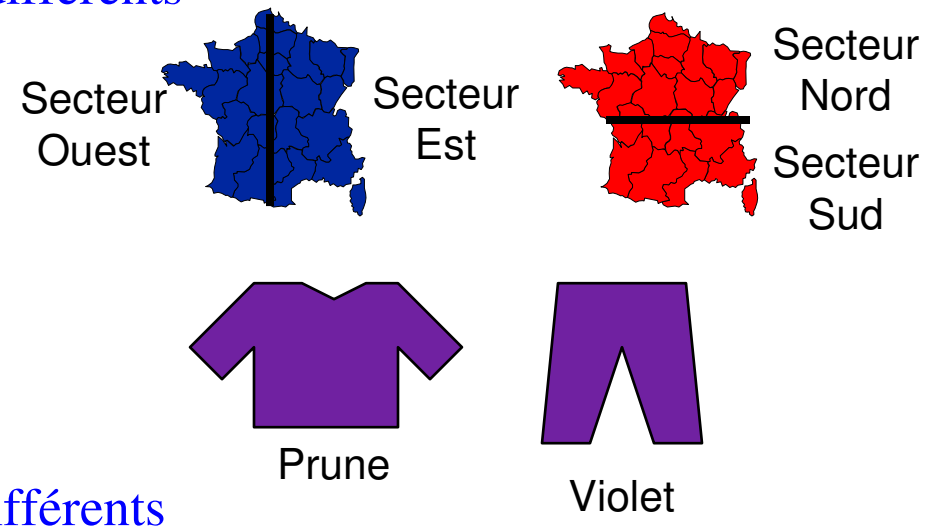
- Extraction
- Transformation
  - filtrer
  - trier
  - homogénéiser
  - nettoyer
  - ...
- Chargement (Loading)

2. DW - Transformations



DW - Transformation

- Existence de plusieurs sources
  - ➢ non conformité des représentations
    - découpages géographiques différents
    - codage des couleurs
    - identification des produits différents
      - produits en vrac
  - ↙ difficulté de comparaison des sources de données
- Mise en conformité nécessaire



2. DW - Constitution de l'entrepôt

- Extraction des données
  - » Besoin d'outils spécifiques pour :
    - accéder aux bases de production (requêtes sur des BD hétérogènes)
    - améliorer la qualité des données : «nettoyer», filtrer, ...
    - transformer les données : intégrer, homogénéiser
    - dater systématiquement les données
- Référentiel
  - » La métabase contient des métadonnées :
    des données sur les données du D.W.
    - quelles sont les données «entreposées», leur format, leur signification, **leur degré d'exactitude**
    - les processus de récupération/extraction dans les bases sources
    - la date du dernier chargement de l'entrepôt
    - l'historique des données sources et de celles de l'entrepôt
- Méthodologie : sera vu plus loin

2. DW - Stockage

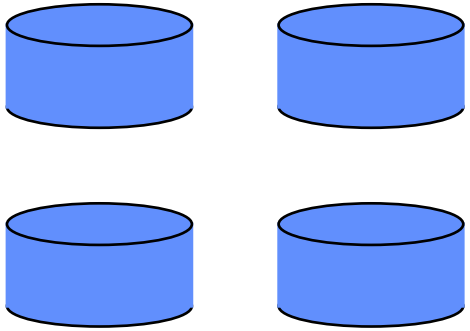
- Optimisation
  - » besoin de synthèse ⇨ agrégation des données

  vs

  - » besoin de détails ⇨ conservation des données détaillées
- Notion de granularité
- Structures
  - » directe simple
  - » cumul simple
  - » résumés roulants : structure généralement choisie

2. DW - Stockage

- Structure directe simple



- pas d'accumulation
- rafraîchissement sur une longue période

JANVIER 2003

J Adams 123 Main Street
P. Anderson 456 High Street
K Appleby 10 A Street
L Azimoff 64 N Ranch Rd
.....

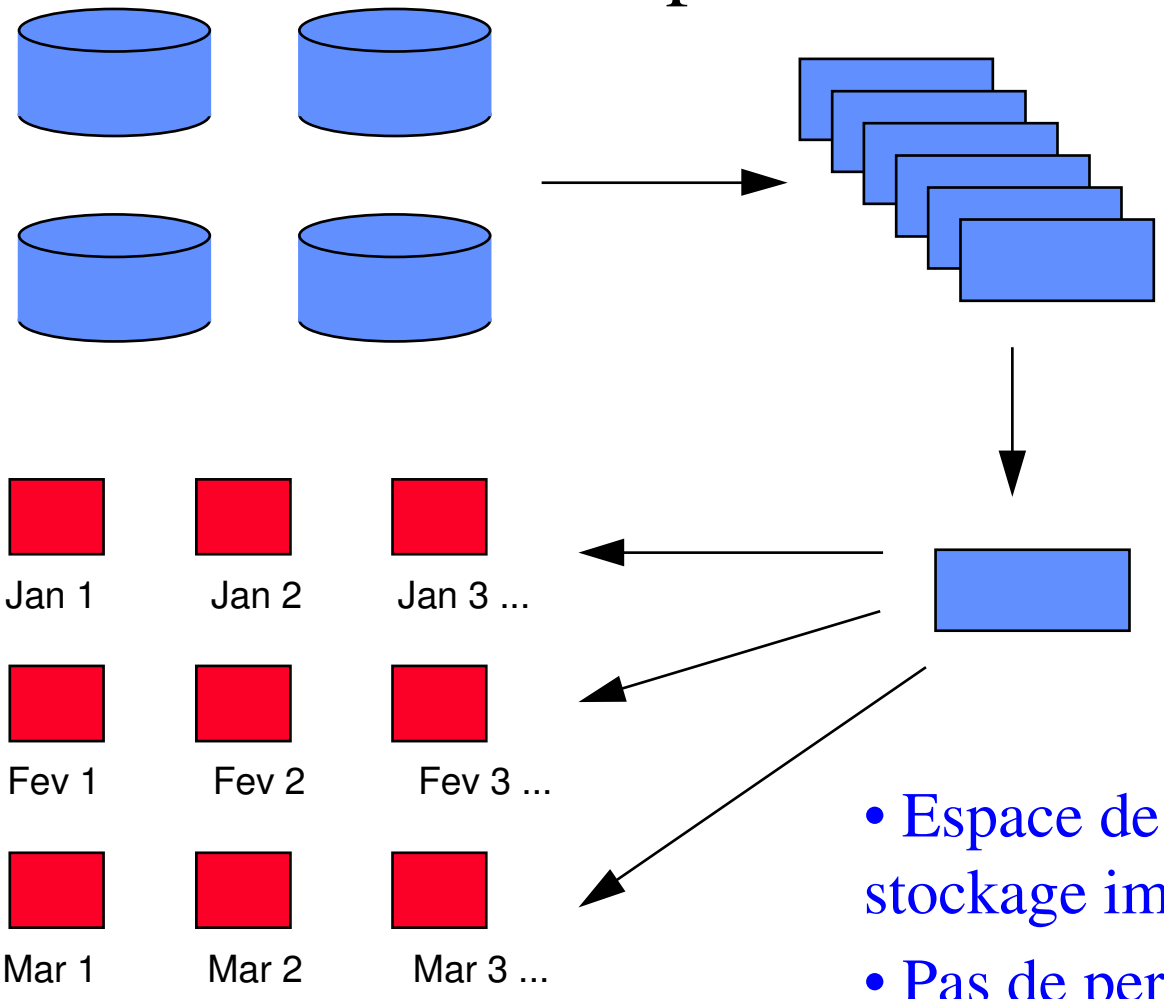
FEVRIER 2003

J Adams 123 Main Street
K Appleby 10 A Street
L Azimoff 64 N Ranch Rd
W Abraham 12 Hwy 9

J Adams	Jan-pres	123 Main street
W Abraham	Feb-pres	12 Hwy 9
P. Anderson	Jan-Jan	456 High Street
.....		

2. DW - Stockage

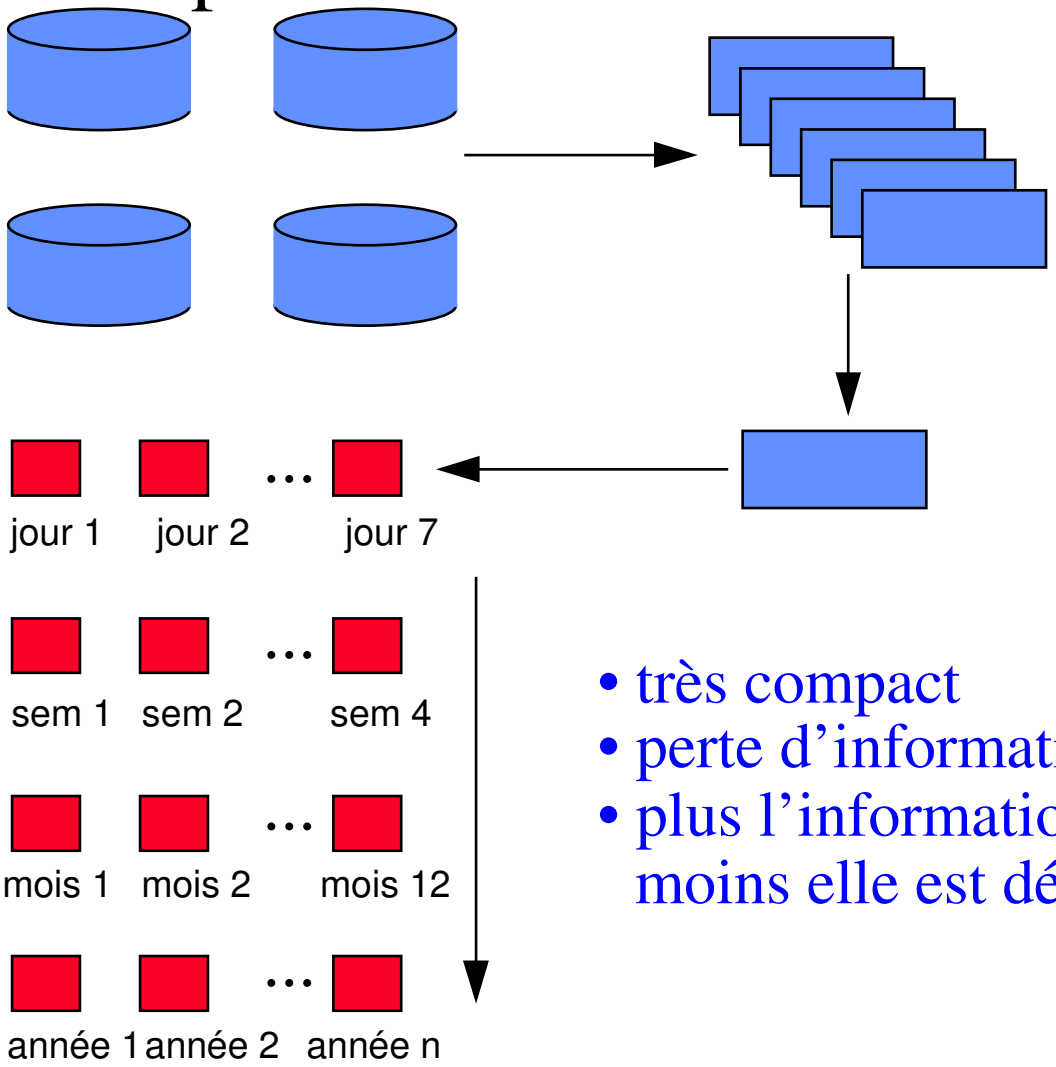
- Structure de cumul simple



- Espace de stockage important
- Pas de perte de détail

2. DW - Stockage

- Structure par résumés roulants



- très compact
- perte d'information
- plus l'information vieillit, moins elle est détaillée

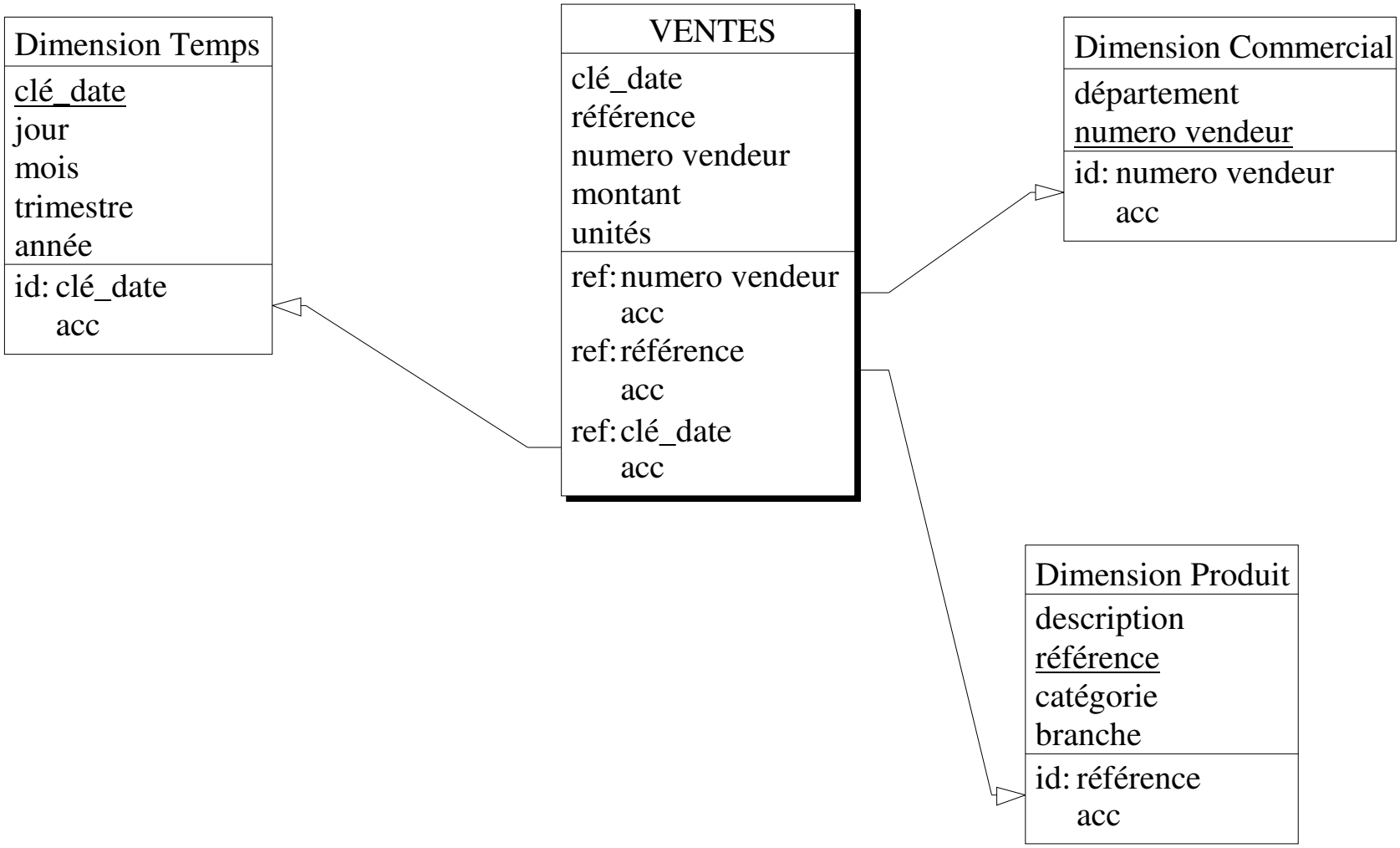
2. DW - Modélisation

- Schéma entités-relations (classique)

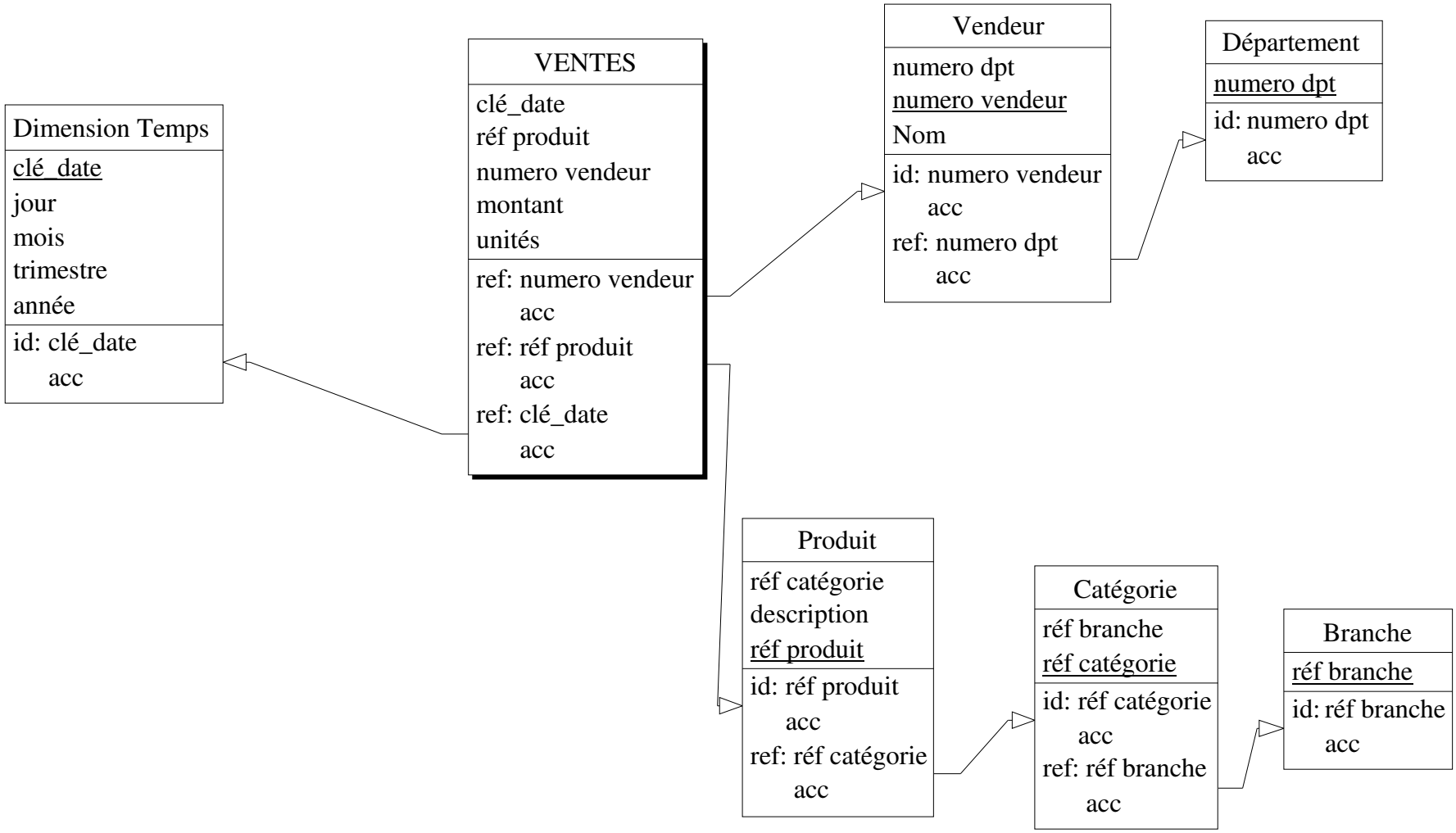
- Schéma en étoile (star schema)
- Schéma en flocon (snowflake schema)

  - » tables de faits : nombreux champs, tables centrales
  - » dimensions : peu de champs,
    permettent d'interpréter les faits

2. DW - Schéma «étoile»



2. DW - Schéma «flocon»



2. DW - Rappel sur les Aggrégats

- Exemple de BD
  - Produit(GENCOD, Designation, Marque, Nature, PrixAchat, PrixReventeConseille)
  - Vente (GENCOD, NMAG, Date, Qte, PrixVente)
  - Magasin(NMAG, Enseigne, Adresse, Ville, Dept)
  - Nat2Cat(Nature, Categorie)
  - Cat2Ray(Categorie, Rayonnage)
  - Dep2Reg(Dept, Region)
- Exercice
  - Donnez les clés primaires et les clés étrangères

2. DW - Questions et Requêtes

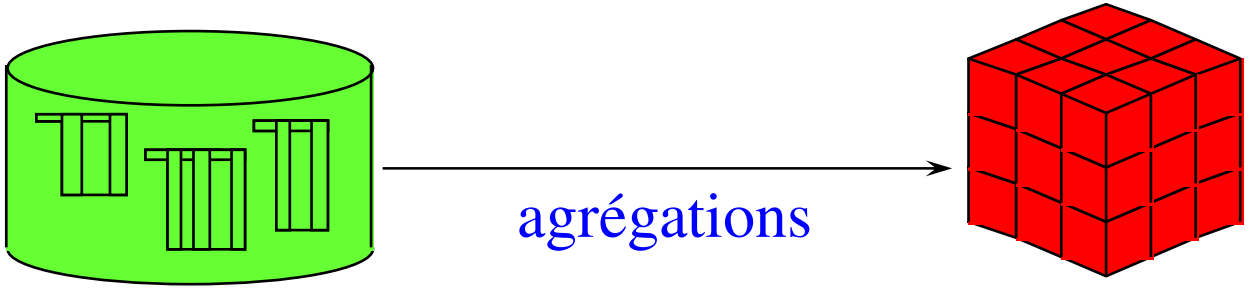
- Montant totale des ventes par ville et par produit
  - select ville, produit, sum(qte*prixvente)
    from vente, produit, magasin
    where produit.GENCOD = vente.GENCOD and
    vente.NMAG = magasin.NMAG
    group by ville, produit
- par région et par catégorie
  - select region, categorie, sum(qte*prixvente)
    from vente, produit, magasin, dep2reg, nat2cat
    where produit.GENCOD = vente.GENCOD and
    vente.NMAG = magasin.NMAG
    and produit.nature = nat2cat.nature
    and magasin.dept = dep2reg.dept
    group by region, categorie

2. DW - Questions et Requêtes

- par région et par catégorie et par année
  - ```
    select   region, categorie, semestre(date), sum(qte*prixvente)
    from     vente, produit, magasin, dep2reg, nat2cat
    where    produit.GENCOD = vente.GENCOD and
             vente.NMAG = magasin.NMAG and
             produit.nature = nat2cat.nature and
             magasin.dept = dep2reg.dept
    group by region, categorie, year(date)
    ```
  - Remarque : year(date) n'est pas toujours disponible

- par région et par catégorie en 2000
  - ```
    select   region, categorie, sum(qte*prixvente)
    from     vente, produit, magasin, dep2reg, nat2cat
    where    produit.GENCOD = vente.GENCOD and
             vente.NMAG = magasin.NMAG and
             produit.nature = nat2cat.nature and
             magasin.dept = dep2reg.dept and
             year(date) = 2000
    group by region, categorie
    ```

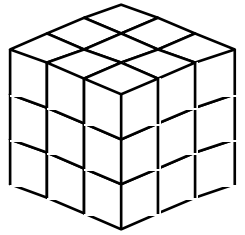
3. BM - D.W. → Base Multidimensionnelle



- Analyse multidimensionnelle
  - » capacité à manipuler des données qui ont été aggrégées selon différentes dimensions
    - ex. : analyse des ventes /catégorie de produit → 1 dim.
      + /année → 2 dim.
      + /département commercial → 3 dim.
      + / zone géographique → 4 dim.
      ....

3. BM - L'Analyse MultiDimensionnelle

- Objectif
  - » obtenir des informations déjà aggrégées selon les besoins de l'utilisateur : simplicité et rapidité d'accès
- HyperCube OLAP
  - » représentation de l'information dans un hypercube à N dimensions



- OLAP (On-Line Analytical Processing)
  - » fonctionnalités qui servent à faciliter l'analyse multidimensionnelle : opérations réalisables sur l'hypercube

3. BM - Glossaire OLAP

- Dimension
  - Temps, Produit, Géographie, ...
- Niveau : hiérarchisation des dimensions
  - Temps :
    - Année, Semestre, Trimestre, Mois, Semaine, ...
  - Produit :
    - Rayon, Catégorie, Nature, ...
  - Géographie :
    - Région, Département, Ville, Magasin
- Membre d'un Niveau
  - Produit::Rayon
    - Frais, Surgelé, ..., Liquide
  - Produit::Rayon.Catégorie
    - Frais.Laitage, ..., Liquide.Vin
  - Produit::Rayon.Catégorie.Nature
    - Frais.Laitage.Yaourt, ... , Liquide.Vin.Champagne

3. BM - Glossaire OLAP

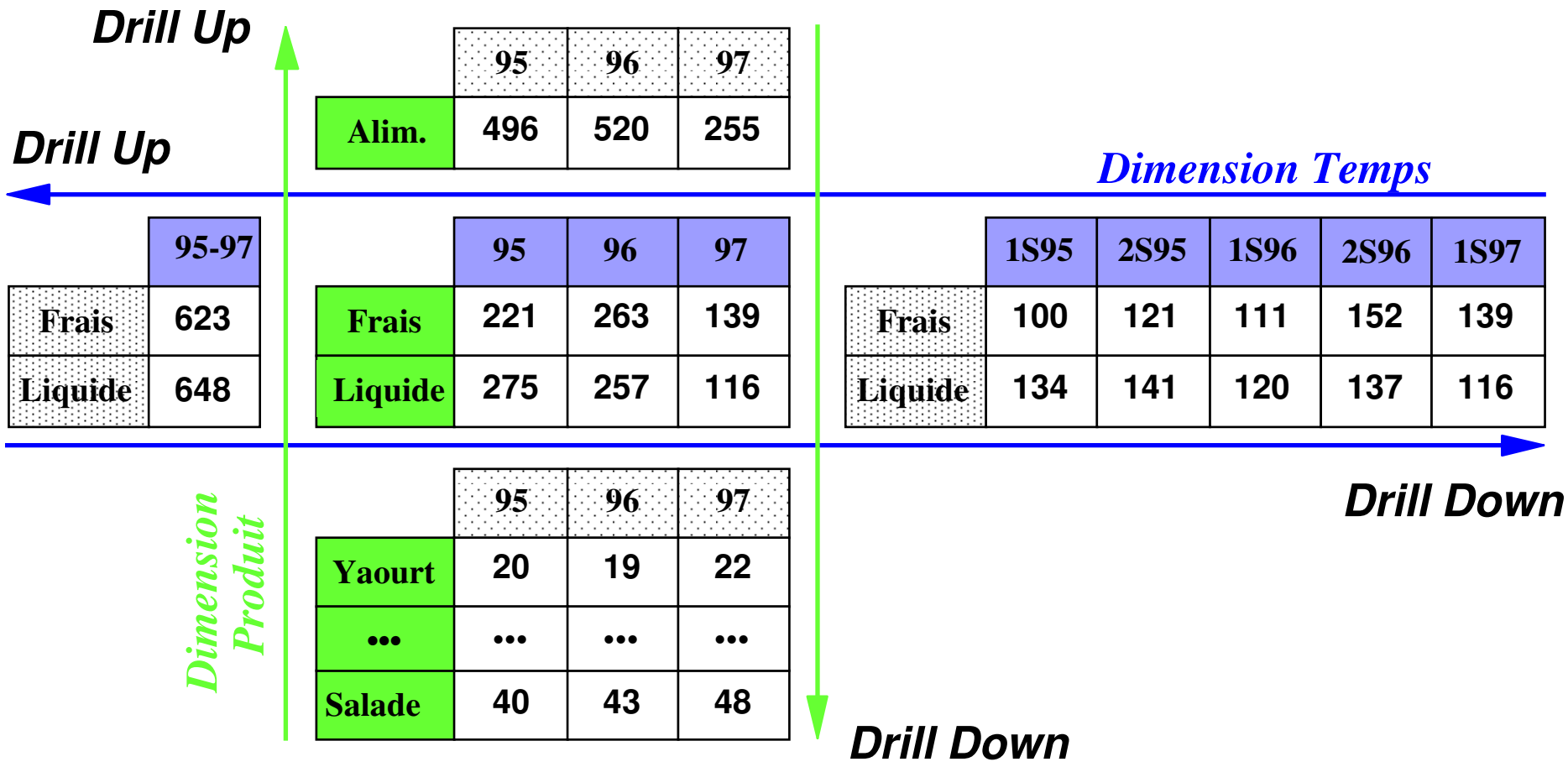
- Cellule
  - » intersection des membres des différentes dim.
- Formule
  - » calcul, expression, règle, croisement des dim.
    - Somme(Qte), Somme(Qte*PrixVente), Moyenne(Qte*(PrixVente-PrixAchat)), ...

3. BM - Opérations OLAP

- But
  - » Visualisation/Utilisation
    d'un fragment de l'Hypercube
- Opérations OLAP
  - » Drill Up / Drill Down
  - » Rotate
  - » Slicing
  - » Scoping

3. BM - Opérations OLAP - Drill Up/Down

vue synthétique / vue détaillée



	95-97
Frais	623
Liquide	648

	95	96	97
Alim.	496	520	255

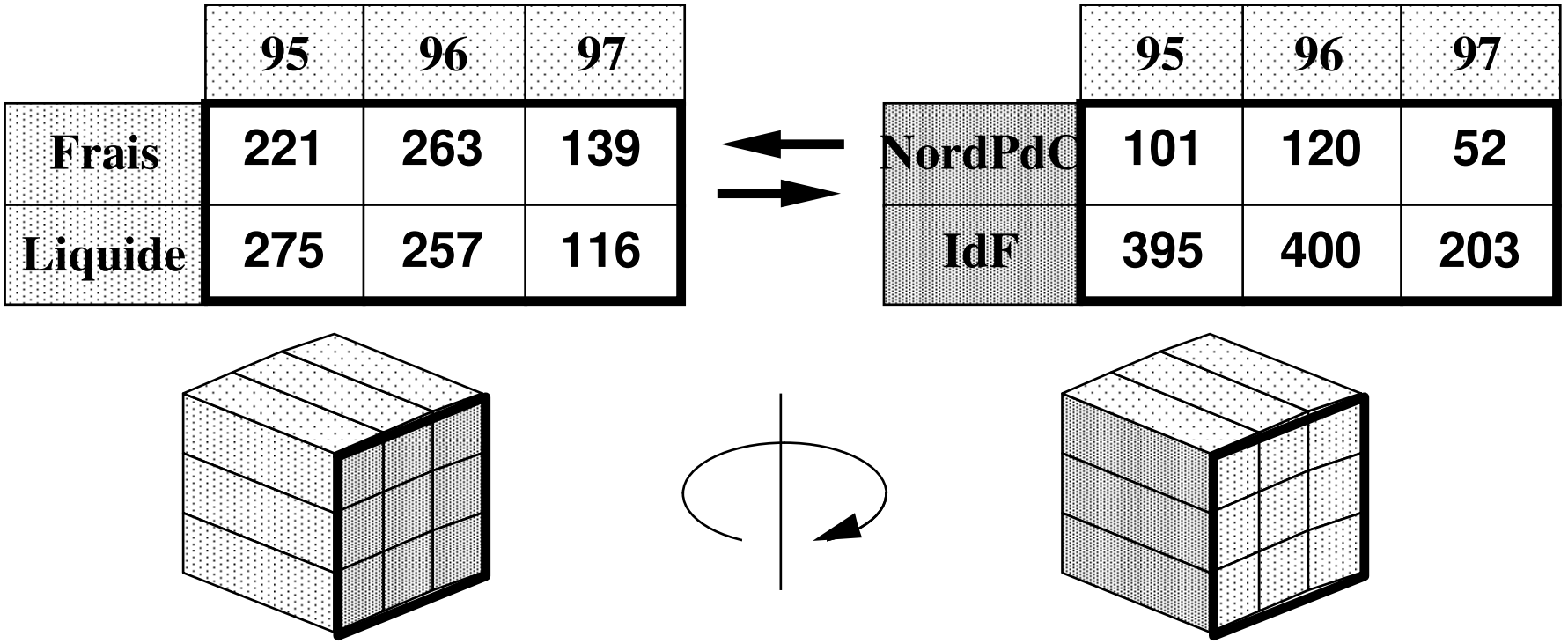
	95	96	97
Frais	221	263	139
Liquide	275	257	116

	1S95	2S95	1S96	2S96	1S97
Frais	100	121	111	152	139
Liquide	134	141	120	137	116

	95	96	97
Yaourt	20	19	22
...	...	...	...
Salade	40	43	48

3. BM - Opérations OLAP - Rotate

	95	96	97
Frais	221	263	139
Liquide	275	257	116



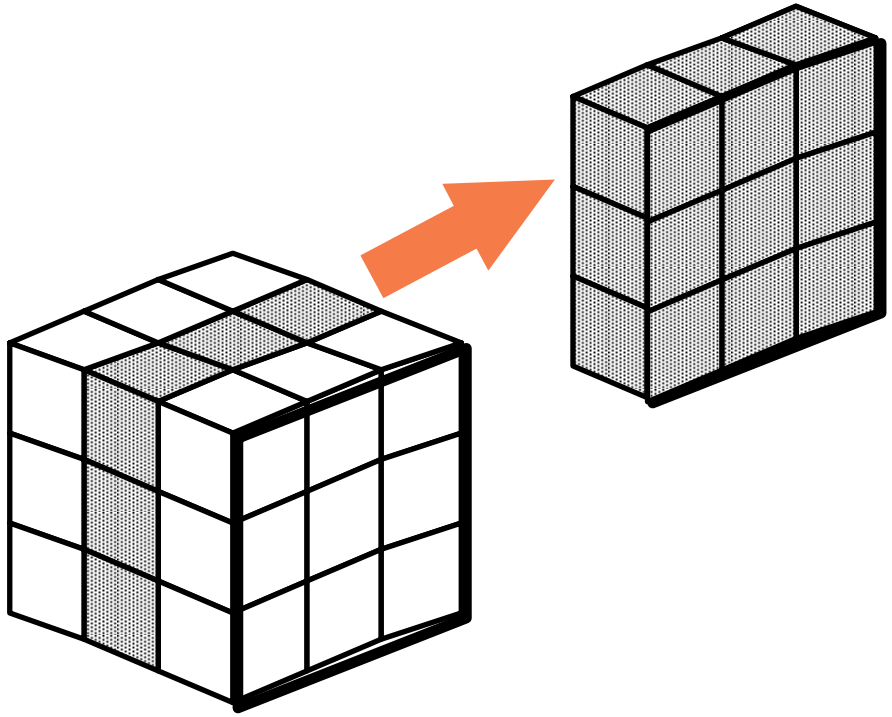
	95	96	97
NordPdC	101	120	52
IdF	395	400	203

3. BM - Opérations OLAP - Slicing

		1995	1996	1997
Frais	IdF	220	265	284
	Province	225	245	240
Liquide	IdF	163	152	145
	Province	187	174	184

→

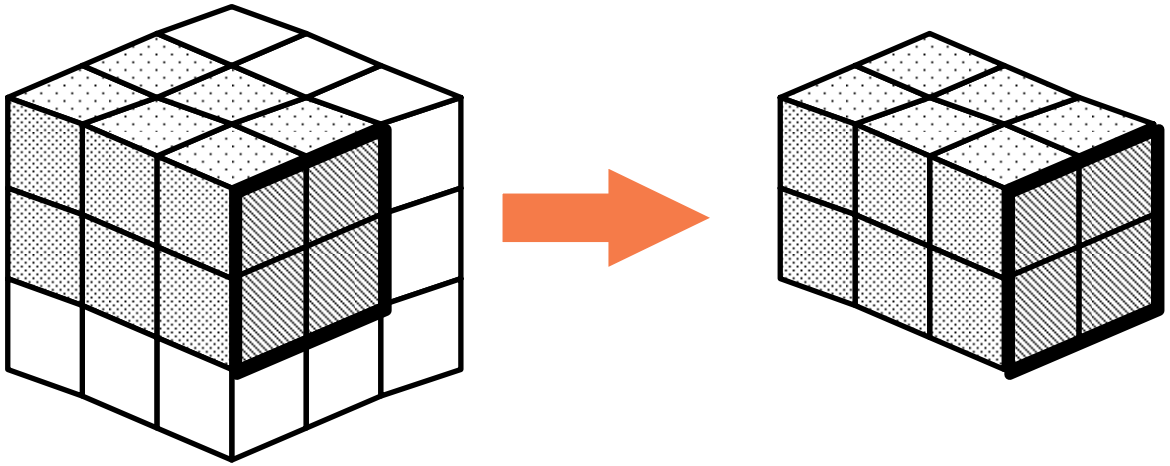
		1996
Frais	IdF	265
	Province	245
Liquide	IdF	152
	Province	174



3. BM - Opérations OLAP - Scoping

		1995	1996	1997
Frais	IdF	220	265	284
	Province	225	245	240
Liquide	IdF	163	152	145
	Province	187	174	184

		1995	1996
Frais	IdF	220	265
	Province	225	245

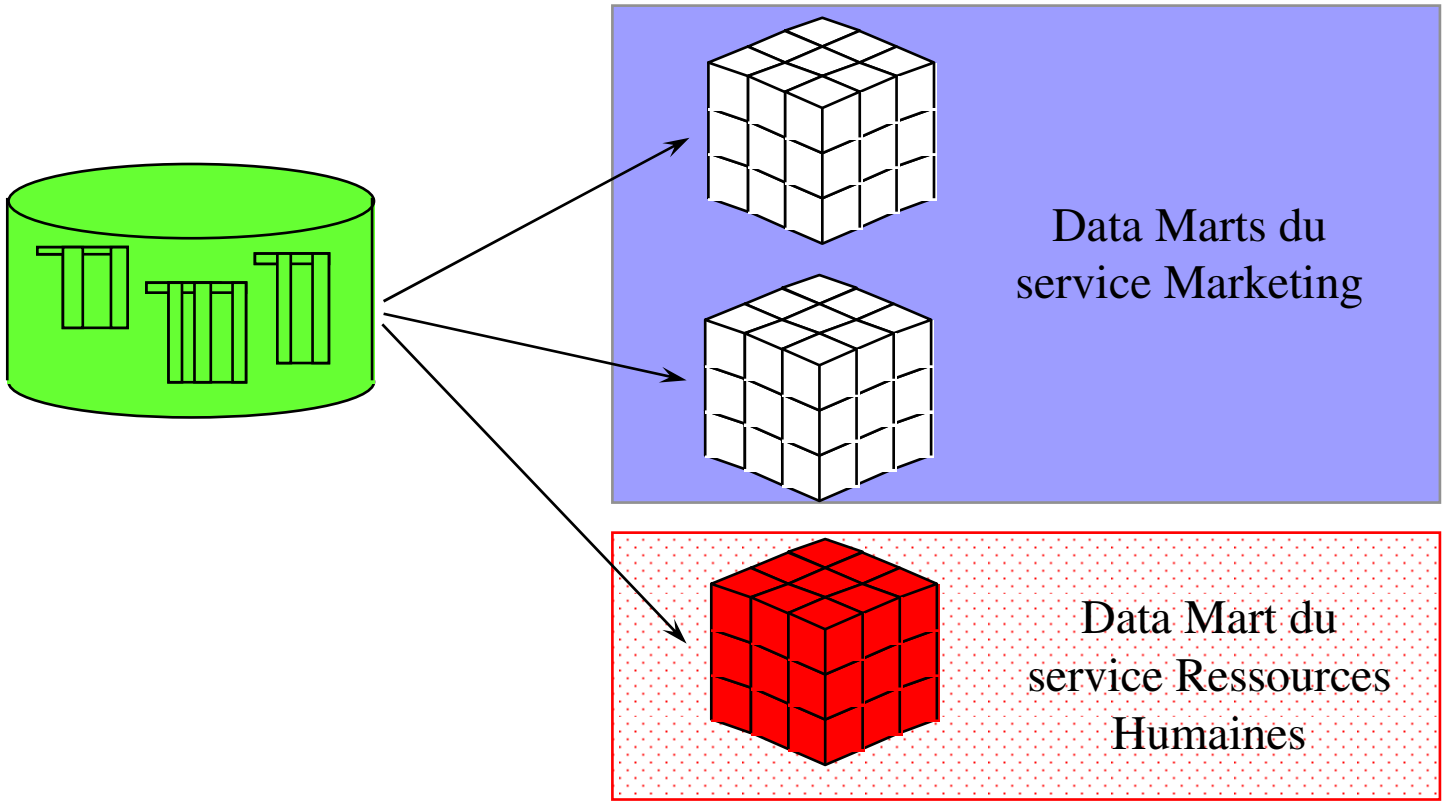


3. BM - OLAP

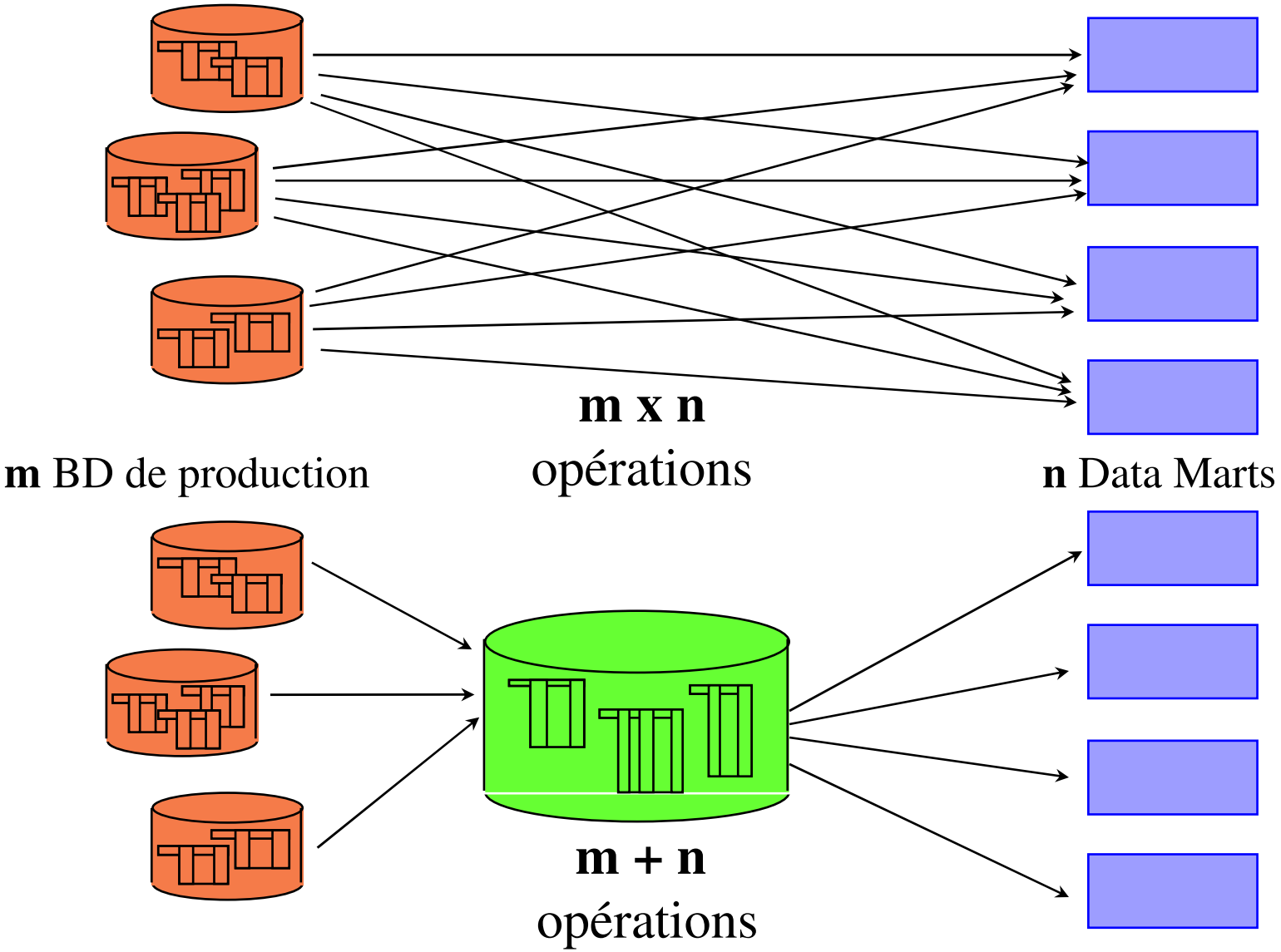
- Constitution de l'Hypercube
  - » Administration
  - » Définition des Dimensions / Niveaux / Membres
    - — Automatique, Manuel, Configuration Métier
- Serveurs OLAP / Clients OLAP
  - » Le client utilise une partie de l'hypercube qu'il cache
  - » Le serveur calcule, stocke l'hypercube et permet son partage.
- Stockage
  - » M-OLAP : accède à une base multidimensionnelle
    - + rapidité
  - » R-OLAP : accède à une base relationnelle
    - + mise à jour
  - » H-OLAP : hybride, multidimensionnel avec accès au niveau le + bas à une base relationnelle

3. BM - Orientation métier : les Data Marts

- Data Mart
  - » vue partielle et orientée métier sur les données du D.W.
  - » à chacun son ensemble d'hypercubes OLAP



3. BM - Un D.W., des Data Marts

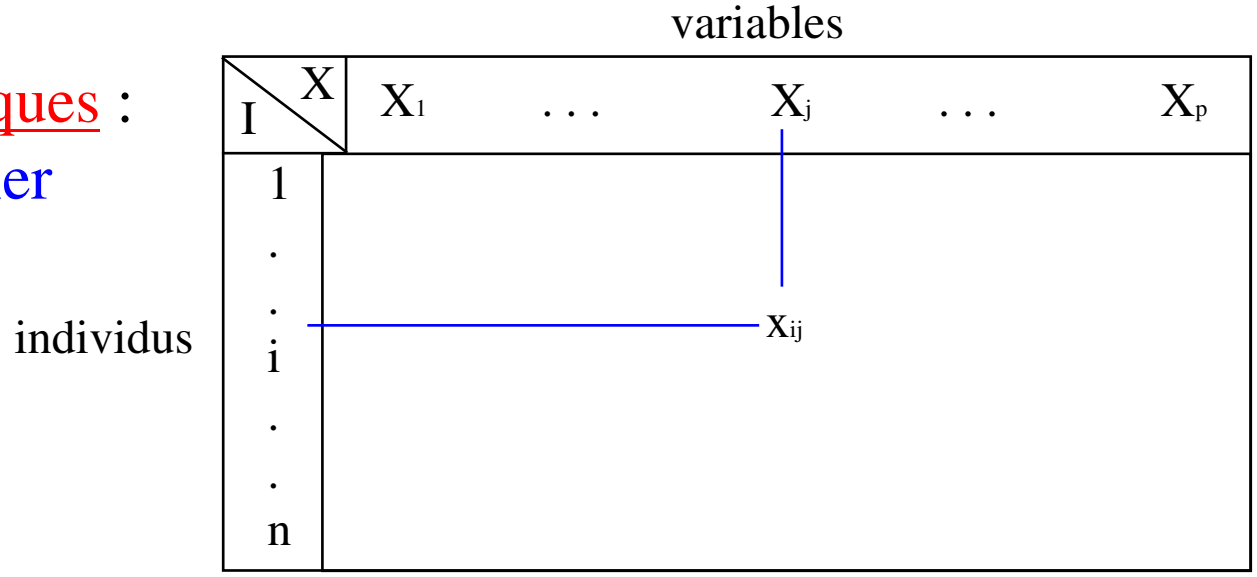


4. Restitution des informations

- Requêteurs
  - » donne une réponse à une question plus ou moins complexe (type SQL)
- EIS (Executive Information Systems)
  - » outils de visualisation et de navigation dans les données
    - statistiques + interfaçage graphique
- Applications spécialisées (ad-hoc)
  - » applications développées spécialement pour les besoins de l'entreprise
- Data Mining
  - » outils évolués de prédiction, simulation, ...

4. Restitution des informations

Techniques statistiques :
utilisées pour vérifier
des hypothèses



- 2 types de variables : quantitatives et qualitatives
- autres caractéristiques possibles des variables :
  - » temporelle
  - » disjonctive (logique , booléenne)
  - » à réponses multiples
  - » catégorique (par catégorie) non ordonnée vs. de rang (ordre sur les données)
  - » de classes (intervalles de valeurs)

4. Restitution des informations

- Recodage de données sur 1 variable
  - » pour normaliser, avoir des ordres de grandeur comparables
- Ex. :
  - » x_i utilisée pour avoir (x_i - moyenne$_{xi}$)
  - » x_i ⟶ (x_i - moyenne$_{xi}$) / e , avec e écart-type de l'échantillon
  - » x_i ⟶ log (x_i) pour limiter l'impact des valeurs exceptionnelles
  - » x_i ⟶ son rang dans l'échantillon
  - » répartition des x_i en classes d'amplitude ou de fréquence équivalente : x_i ⟶ sa classe Cj
  - » x_i ⟶ 0 ou 1 : création d'un tableau logique
  - » date ⟶ durée
  - » données géographiques ⟶ coordonnées, distances

4. Restitution des informations

- Recodage de données sur plusieurs variables
- Ex. :
  - ratios (%) : montant / total
  - fréquences : fréquence de x_{ij} = valeur v par rapport à l'ensemble des valeurs prises par x_{ij}
  - tendance : mesure d'une variation
  - combinaisons (linéaires ou non) : formules de calculs combinant plusieurs données

    Ex. : revenu résiduel = revenu - (charges + x*nb d'adultes + y*nb d'enfants)

4. Restitution des informations

étudiants	Note1	Note2	(Note1-moy1) écart note1 / moy	(Note2-moy2) écart note2 / moy	(ecart 1) puis 2	(ecart 2) puis 2	écart1 * écart2
A	16	9	6	0	36	0	0
B	8	7	-2	-2	4	4	4
C	4	8	-6	-1	36	1	6
D	15	9	5	0	25	0	0
E	9	8	-1	-1	1	1	1
F	19	10	9	1	81	1	9
G	2	11	-8	2	64	4	-16
H	15	12	5	3	25	9	15
I	3	8	-7	-1	49	1	7
J	9	12	-1	3	1	9	-3
	10,00	**9,40**	**0**	**0**	**322**	**30**	**2,3**
	moyenne	moyenne	somme	somme	somme	somme	somme/n = co-variance
					32,2	**3**	**0,24**
				somme/nb individus =	variance	variance	coef corrélation = covariance/(ecart-type1*ecart-type2)
				racine(variance) =	**5,7**	**1,7**	
					écart-type	écart-type	

Conclusions :

- Matière 1 plus «risquée» : différenciation importante dans les notes
- Matière 2 : - de risque mais ne permet pas d'obtenir bcp de points supplémentaires
- Classification des étudiants + aisée avec les notes 1 (nuage de points)
- Pas de corrélation entre les notes des 2 matières

4. Restitution des informations

- Similarité : coïncidences positives ou négatives
  - » Ex. sur le tableau (from Lefébure et Venturi):

• calcul des coïncidences
• calcul des indices de similarité entre BC et CD, BC et GR, CD et GR

	barre céréale	crème dessert	gâteau de riz
chocolat	OUI	NON	OUI
beurre	NON	NON	OUI
liquide	NON	OUI	NON
parfum mandarine	NON	NON	OUI
emballage métal	NON	OUI	OUI
mini-dose	OUI	OUI	NON
sucre	OUI	OUI	OUI
riz	OUI	NON	OUI
édulcorant	NON	NON	OUI
colorant	NON	NON	OUI

  - » Indices de similarité (3 formules différentes):
    - Russel : nb de coïncidences positives / nb de comparaisons
    - Jaccard : nb de coïncidences positives / (nb de comparaisons - nb de coïncidences négatives)
    - Sokal : nb de coïncidences positives et négatives / nb de comparaisons

4. Data Mining

- OLAP vs Data Mining
  - OLAP : l'utilisateur cherche à confirmer des intuitions
    - ex. : «A-t-on vendu plus de yaourts en Région Parisienne qu'en Bretagne en 2003 ?»
  - Data Mining : l'utilisateur cherche des corrélations non évidentes
    - ex. : «Quelles sont les caractéristiques de l'achat de yaourts ?»

4. Data Mining

- Principe
  - » Creuser une mine (le DW) pour rechercher un filon (l'information)
  - » Evolution par rapport aux statistiques «classiques»
- Objectifs
  - » Prédiction (What-if)
    - — ex. demande de prêt
  - » Découverte de Règles Cachées (corrélations)
    - — ex. bière + couches
  - » Confirmation d'hypothèses
- Entrées
  - » Fichiers Texte, Feuille de Calcul (SYLK, XLS)
  - » Slice/Scope d'un HyperCube OLAP

4. Restitution des informations

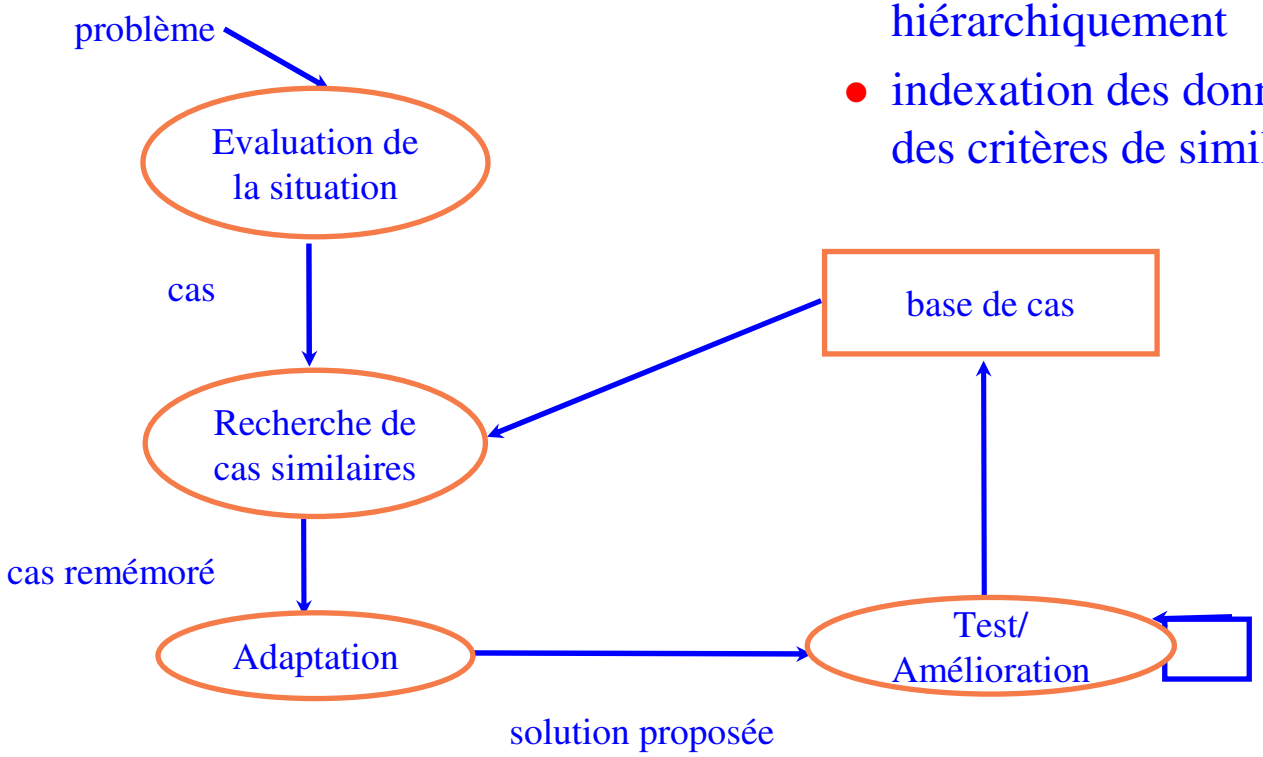
- Recherche des exemples les plus proches
  - » Raisonnement à base de cas
  - » Agents intelligents

- Analyse de données : construction d'un modèle
  - » réseaux de neurones
  - » arbres de décisions
  - » ...

4. Restitution des informations - RBC

- Raisonnement à base de cas (RBC ou CBR)
  - » résolution de problèmes par comparaison avec problèmes similaires déjà rencontrés

- la base de cas est structurée hiérarchiquement
- indexation des données : pondération des critères de similarité

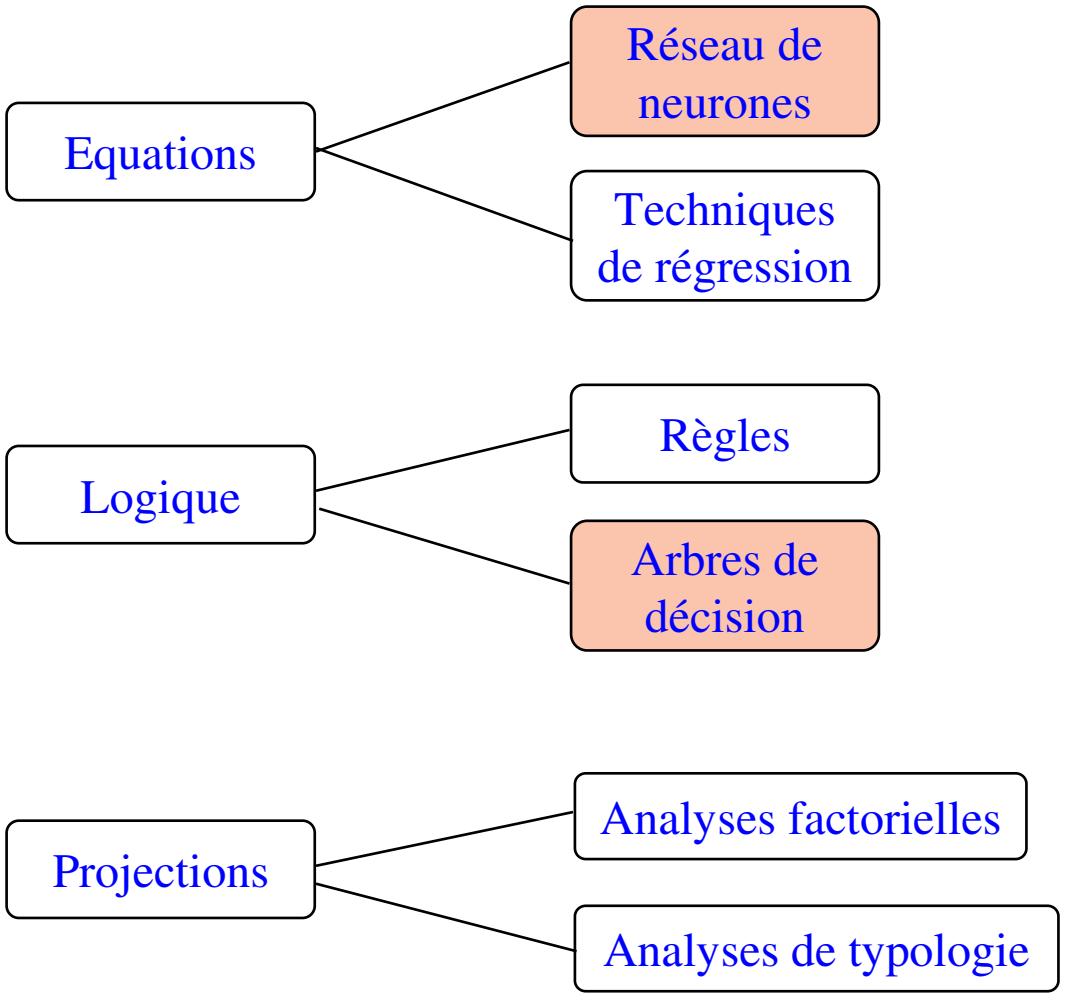


4. Restitution des informations - Agents

- Agents intelligents ou Knowbots
  - » entités logicielles capables d'agir de manière autonome dans un environnement informatique hétérogène
  - » personnalisation de l'information par apprentissage d'un «profil» utilisateur
  - » utilisation sur internet, agents commerciaux électroniques

4. Restitution des informations

- Analyse de données



4. Techniques de Data Mining

Arbres de Décision

- Principe :
  - » division de la population par groupes dont les individus partagent une caractéristique commune
  - » construction à partir d'une base d'exemples
  - » recherche de la caractéristique la plus discriminante à chaque étape (classification automatique)
  - » variables discrètes
- Résultat : mise en évidence de corrélations
  - » enchaînement hiérarchique de règles logiques sous forme d'un «arbre»

4. Techniques de Data Mining

» exemple : le mailing, le contact téléphonique

base d'exemples

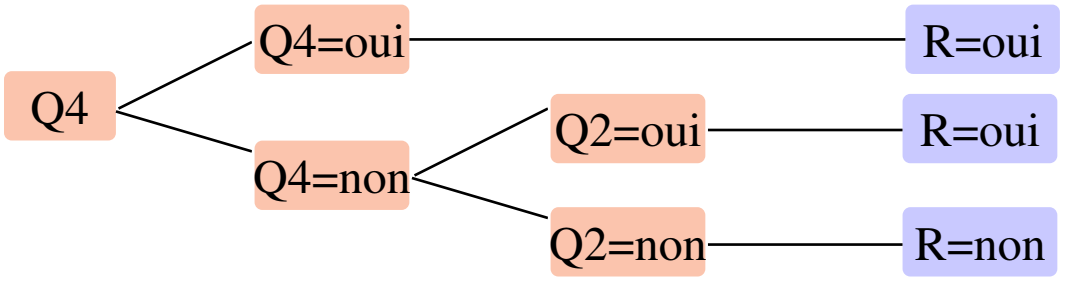
	Question	E1	E2	E3	E4	E5	E6
Q1	Connaît l'école	oui	oui	non	oui	non	non
Q2	A eu un stagiaire	oui	non	non	non	non	non
Q3	A embauché un ancien étudiant	oui	non	oui	non	oui	oui
Q4	Verse la taxe	non	oui	oui	non	non	non
Q5	A participé à un événement	oui	oui	oui	oui	oui	oui
R	Rendez-vous	oui	oui	oui	non	non	non

4. Techniques de Data Mining

» exemple : le mailing, le contact téléphonique

base d'exemples

	Question	E1	E2	E3	E4	E5	E6
Q1	Connaît l'école	oui	oui	non	oui	non	non
Q2	A eu un stagiaire	oui	non	non	non	non	non
Q3	A embauché un ancien étudiant	oui	non	oui	non	oui	oui
Q4	Verse la taxe	non	oui	oui	non	non	non
Q5	A participé à un événement	oui	oui	oui	oui	oui	oui
R	Rendez-vous	oui	oui	oui	non	non	non



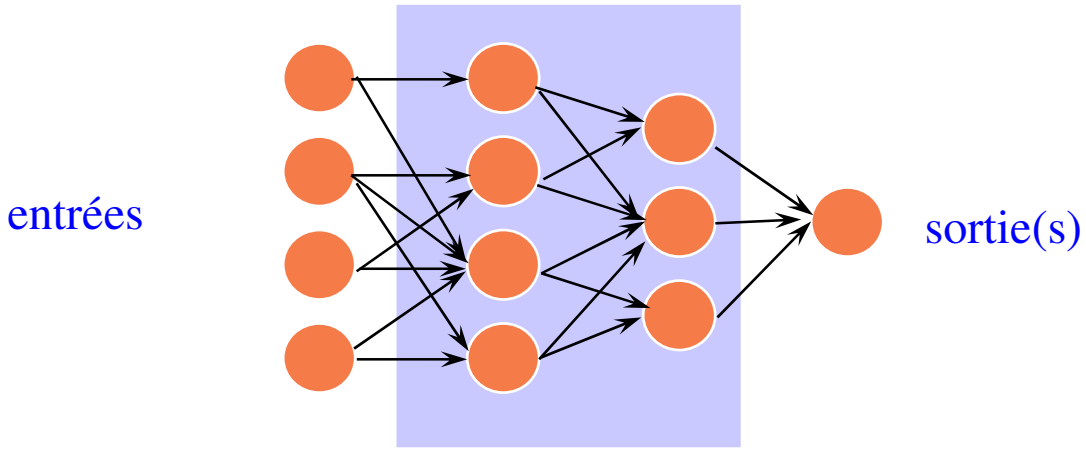
4. Techniques de Data Mining

Réseaux de Neurones

- Principe :
  - » neurones = petits modules de calcul organisés en «couches» constituant un réseau
  - » activation et apprentissage
    - activation d'un neurone par ceux de la couche amont
    - sortie fonction plus ou moins complexe des entrées
    - apprentissage à partir d'une base d'exemples : si telles entrées alors telles sorties attendues
    - renforcement des chemins les plus parcourus
- Résultat
  - » Création d'un modèle reposant sur les données existantes par un réseau apprenant

4. Techniques de Data Mining

- Techniques les plus utilisées
  - » MultiLayer Perceptron, RadialBasis Function, Kohonen Network
- Données numériques
- Prédiction / Simulation
  - » ex. : le prêt bancaire

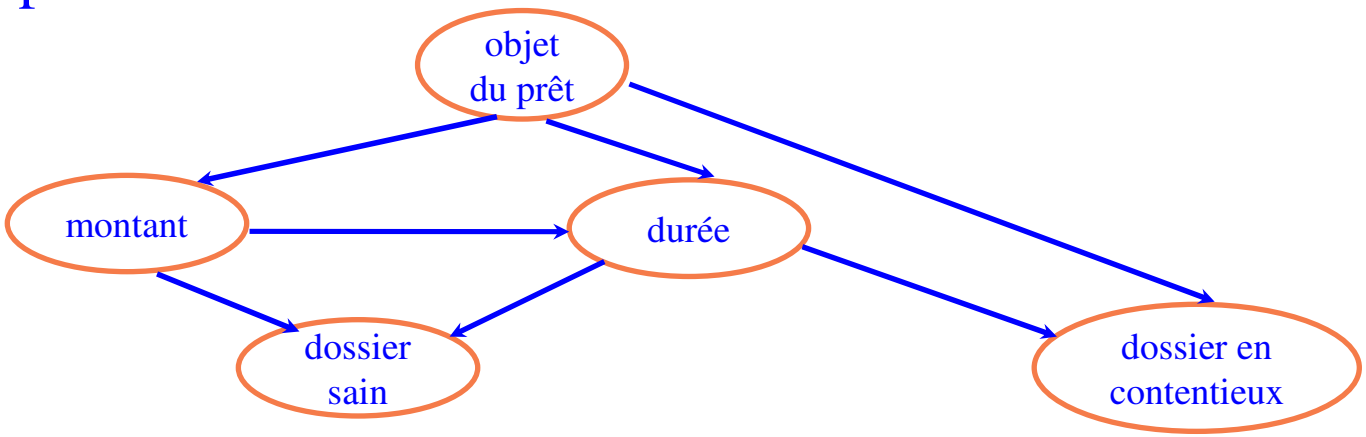


4. Techniques de Data Mining

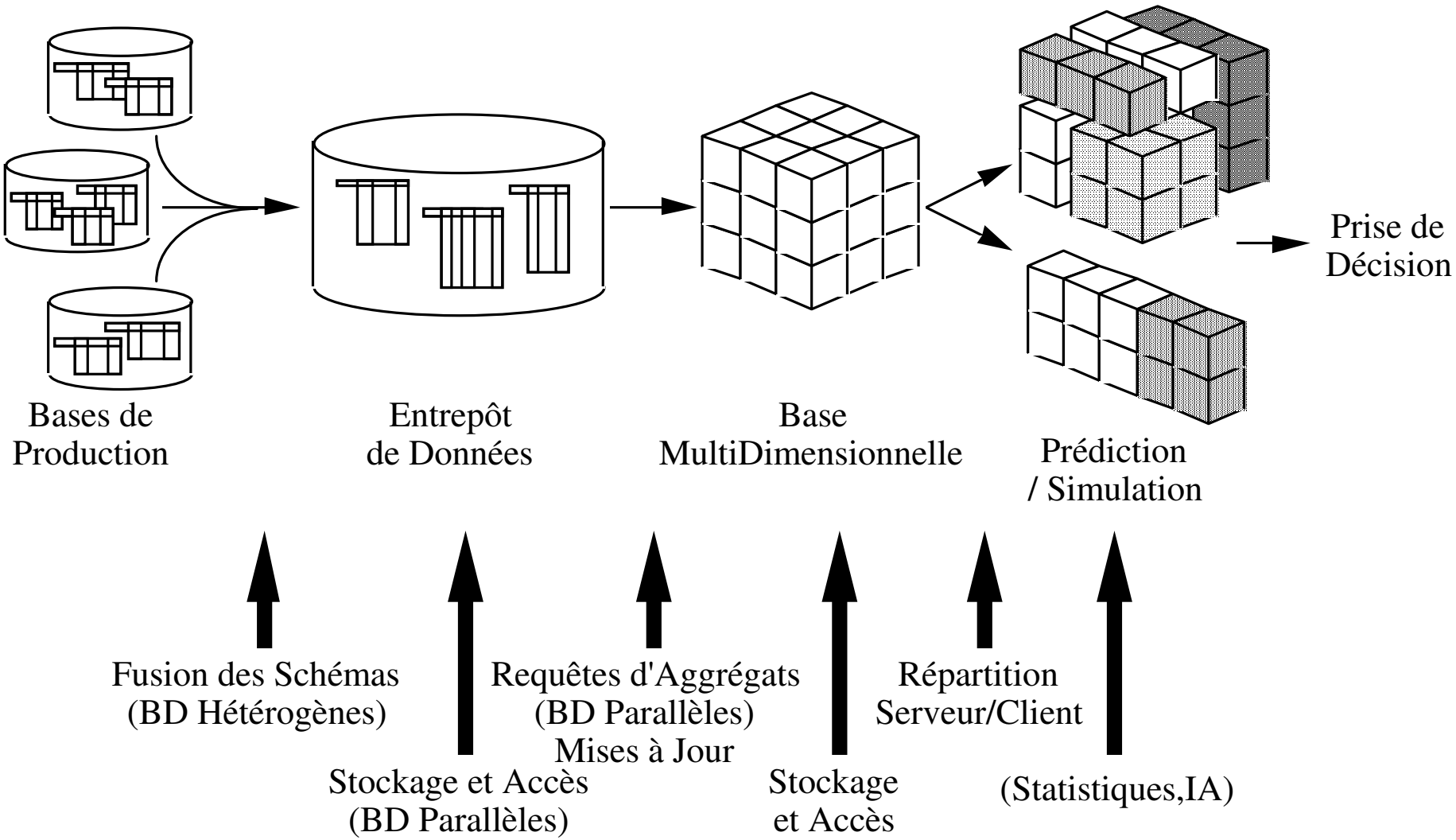
- Les algorithmes génétiques
  - principes de sélection, reproduction et mutation génétiques
  - convergence vers les solutions les meilleures (les plus adaptées) par conservation des bons individus / chromosomes aux générations suivantes tout en gardant une population identique en volume
  - utilisation :
    - optimisation de grilles de score : modification des paramètres d'une régression logique,
    - optimisation d'arbres de décision : isoler les variables les plus pertinentes pour expliquer un comportement,
    - optimisation de réseaux de neurones : modification des poids des liaisons

4. Techniques de Data Mining

- Les réseaux bayésiens
  - but : associer une probabilité d'apparition d'un événement étant donnée la connaissance de certains autres événements
  - graphe orienté dans lequel les noeuds représentent des variables et les arcs, les dépendances entre ces variables
  - probabilités associées aux variables et aux liens de dépendance

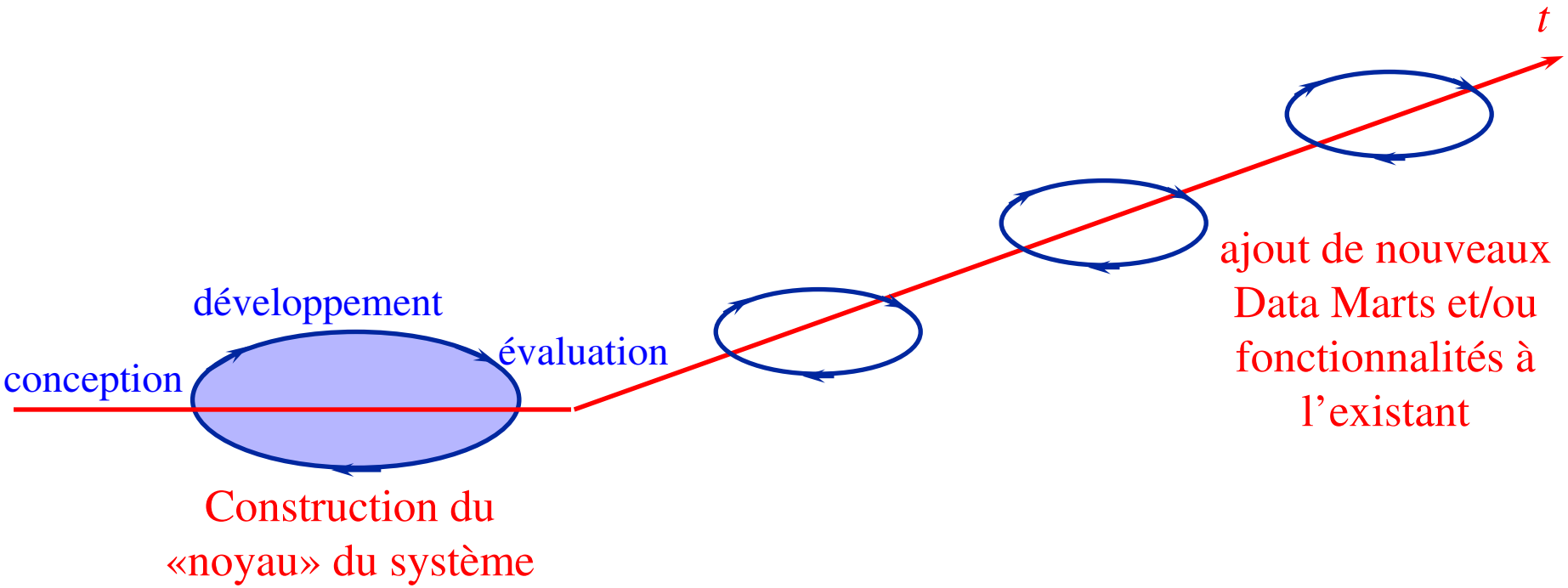


4. Synthèse


Bases de
Production
Entrepôt
de Données
Base
MultiDimensionnelle
Prédiction
/ Simulation
Prise de
Décision
Fusion des Schémas
(BD Hétérogènes)
Stockage et Accès
(BD Parallèles)
Requêtes d'Aggrégats
(BD Parallèles)
Mises à Jour
Stockage
et Accès
Répartition
Serveur/Client
(Statistiques,IA)

5. Gestion de projet Data Warehouse

- Chaque Data Warehouse est unique
- Tâche complexe et ardue
- Construction itérative
  - Focalisations successives sur un ensemble de besoins

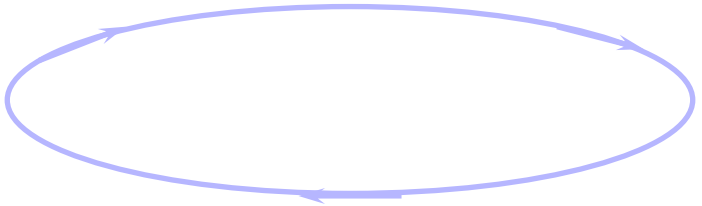


5. Les acteurs

- Le «sponsor»
  - » membre de la direction, soutient le projet
- Le comité utilisateur
  - » différentes catégories (regroupement par besoins)
  - » des représentants
- Les administrateurs du système d'information
  - » très importants (connaissance des données)
  - » maintenance future du Data Warehouse
- L'équipe de conception
  - » souvent : consultants externes

5. Cycle de vie

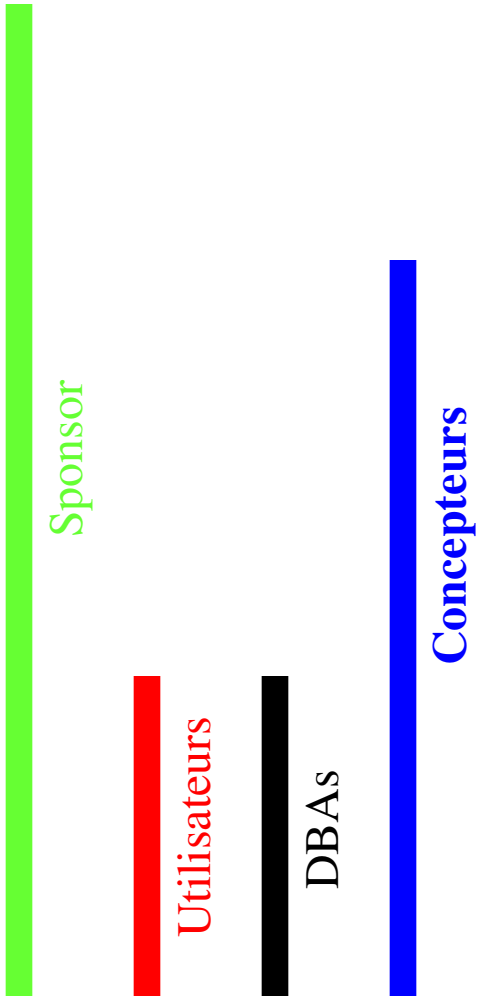
- Justification du projet
- Itérations :
  - » Conception
  - » Développement (prototypage)
  - » Evaluation



- Tests et Mise en exploitation
- Evaluation et évolution

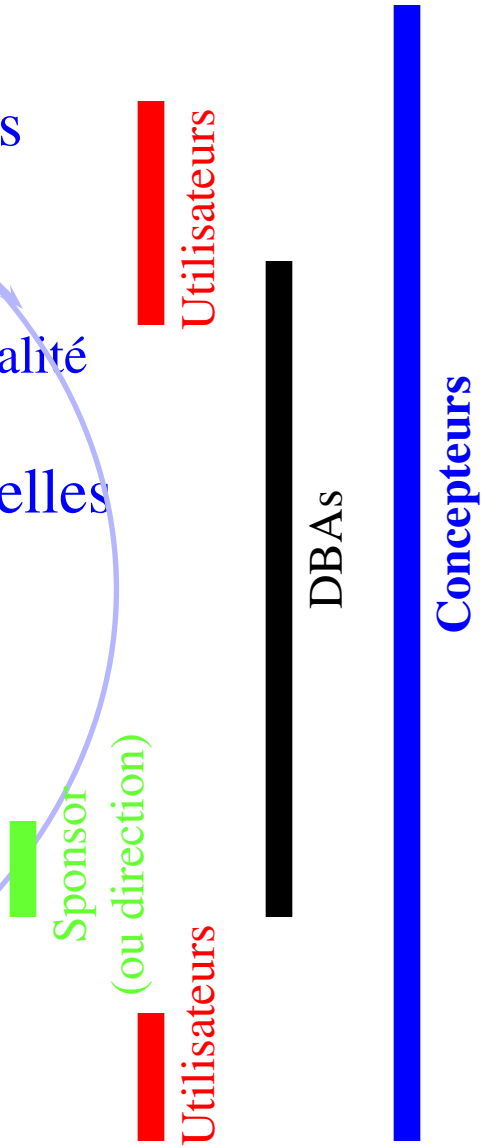
5. Justification du projet

- objectif, retours attendus
- choix de l'équipe de conception
  - » appel éventuel à un intervenant extérieur
- choix du ou des domaine(s) cibles
  - » correspondant au(x) premier(s) Data Mart(s)
- constitution du comité utilisateurs et de l'équipe de DBAs
- planification



5. Cycle de prototypage

- Analyse
  - » besoins des utilisateurs, difficultés actuelles
    - interviews
  - » données de production
    - Rétro-Ingénierie, documentation, évaluation qualité
      ...
  - » existant éventuel en applications décisionnelles
- Modélisation
  - » données
  - » traitements
- Choix techniques
- Développement de prototype
- Evaluation



5. Recueil des besoins

date de réalisation :
auteur :
utilisateur :

- OBJECTIF PRINCIPAL
  - Qu'attendez-vous principalement du Data Warehouse ?
- DECISIONS
  - Quelles décisions avez-vous à prendre ? (Quoi ?)
  - Quels sont les critères qui influencent la prise de décision ? (Comment ?)
  - Dans quel(s) but(s) les décisions sont-elles prises ? (Pourquoi ?)
- DIFFICULTES ACTUELLES
  - Quelles sont les difficultés actuellement rencontrées dans la prise de décision, difficultés en rapport avec les données ?
    - précision des données (détails, actualisation, vérification)
    - synthèse des données (regroupements)
    - évolution (temps)
    - autres...
- ACTUALISATION DES INFORMATIONS
  - Quels sont les besoins concernant la fréquence de mise à jour des informations proposées par le Data Warehouse ?
- PRESENTATION DES INFORMATIONS
  - Quelles sont vos préférences dans la présentation des informations
    - tableaux, graphiques, ?
  - Type de graphiques : barres-graphes, "camemberts", nuages de points ... ?
  - Existe-t-il une présentation actuelle ou habituelle à conserver ?

5. Analyse des données de production

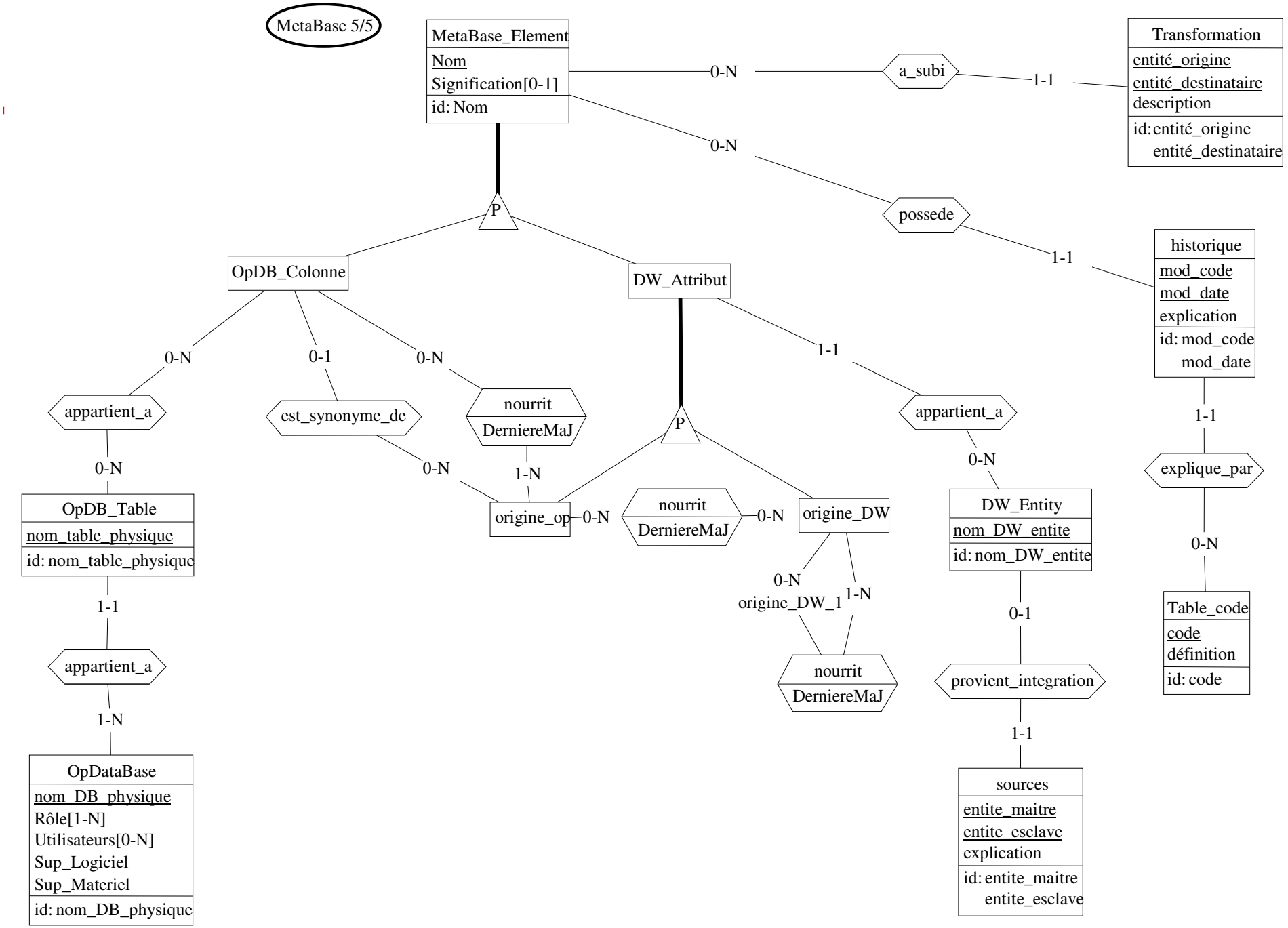
- Identifier les sources de données qui alimenteront le Data Warehouse :
  - » quelles sont les données disponibles
  - » comment accéder à ces données (lieu, système et architecture)
  - » qui les gèrent
  - » leur format
  - » leur signification
  - » leur qualité

➔ méta-données stockées dans la métabase

5. La métabase

*Tout Data Warehouse comporte une **métabase** qui regroupe des méta-données. Les **méta-données** sont utilisées pour stocker des informations à propos des données utilisées par le Data Warehouse.*

- la métabase comprend :
  - un dictionnaire des données : contient les définitions des éléments contenus dans les bases de données et les liens entre eux.
  - l'origine des données : quelle est la base opérationnelle d'origine d'une donnée
  - le flux de données (direction, fréquence)
  - la transformation des données
  - l'historique des données
  - ...


MetaBase 5/5
MetaBase_Element
Nom
Signification[0-1]
id: Nom
0-N
a_subi
1-1
Transformation
entité_origine
entité_destinataire
description
id: entité_origine
entité_destinataire
0-N
possede
1-1
historique
mod_code
mod_date
explication
id: mod_code
mod_date
P
OpDB_Colonne
DW_Attribut
0-N
0-1
0-N
1-1
appartient_a
est_synonyme_de
nourrit
DerniereMaJ
P
appartient_a
1-1
explique_par
0-N
0-N
0-N
1-N
0-N
OpDB_Table
nom_table_physique
id: nom_table_physique
origine_op
0-N
nourrit
DerniereMaJ
0-N
origine_DW
DW_Entity
nom_DW_entite
id: nom_DW_entite
Table_code
code
définition
id: code
1-1
0-N
origine_DW_1
1-N
0-1
appartient_a
nourrit
DerniereMaJ
provient_integration
1-N
1-1
OpDataBase
nom_DB_physique
Rôle[1-N]
Utilisateurs[0-N]
Sup_Logiciel
Sup_Materiel
id: nom_DB_physique
sources
entite_maitre
entite_esclave
explication
id: entite_maitre
entite_esclave

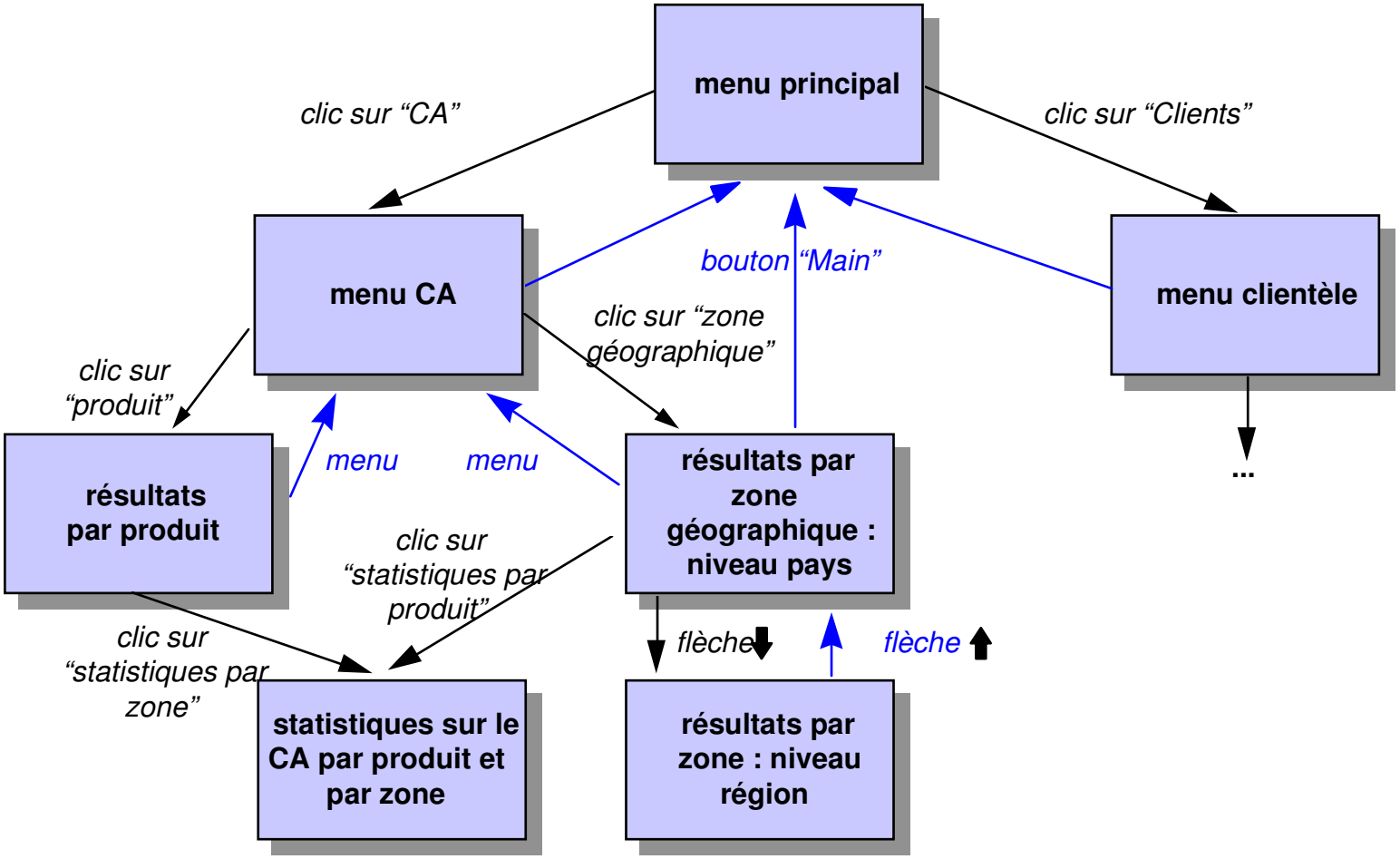
5. Communiquer avec les utilisateurs

- Proposer une maquette de l'interface homme-machine :
  - » contenu des écrans
  - » enchaînement des écrans

➔ *critique par les utilisateurs et recueil des besoins*

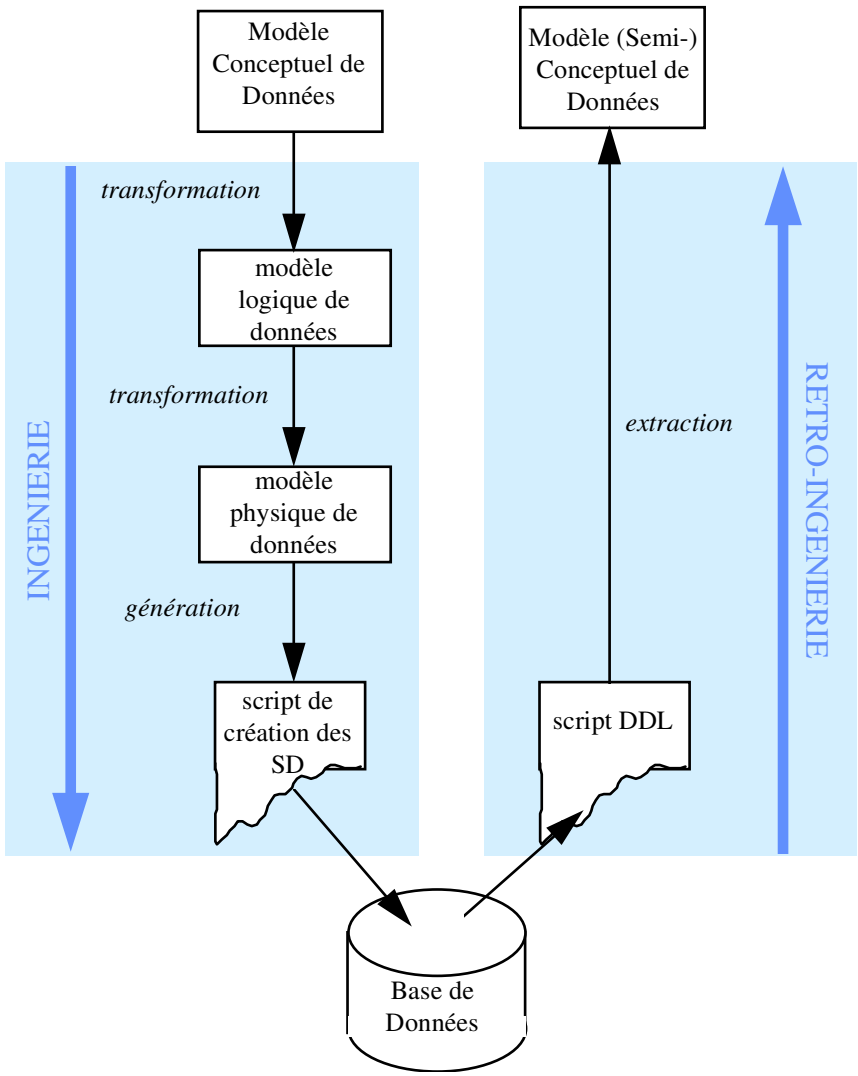
- Support : informatique ou papier
- Privilégier un moyen de communication non technique

5. Exemple d'enchaînement des écrans



5. Rétro-Ingénierie

- Principe
  - » reconstruire les modèles de conception d'une B.D.
- Usage
  - documentation inexistante ou non réactualisée
  - compréhension des données de production en vue de leur intégration
- Outils spécifiques
  - » AGL (Atelier de Génie Logiciel ou CASE)



5. Intégration

- intégrer les MCD obtenus par rétro-ingénierie en un modèle global et homogène

- difficultés :
  - conflit de classification
  - conflit de description
  - conflit de structure

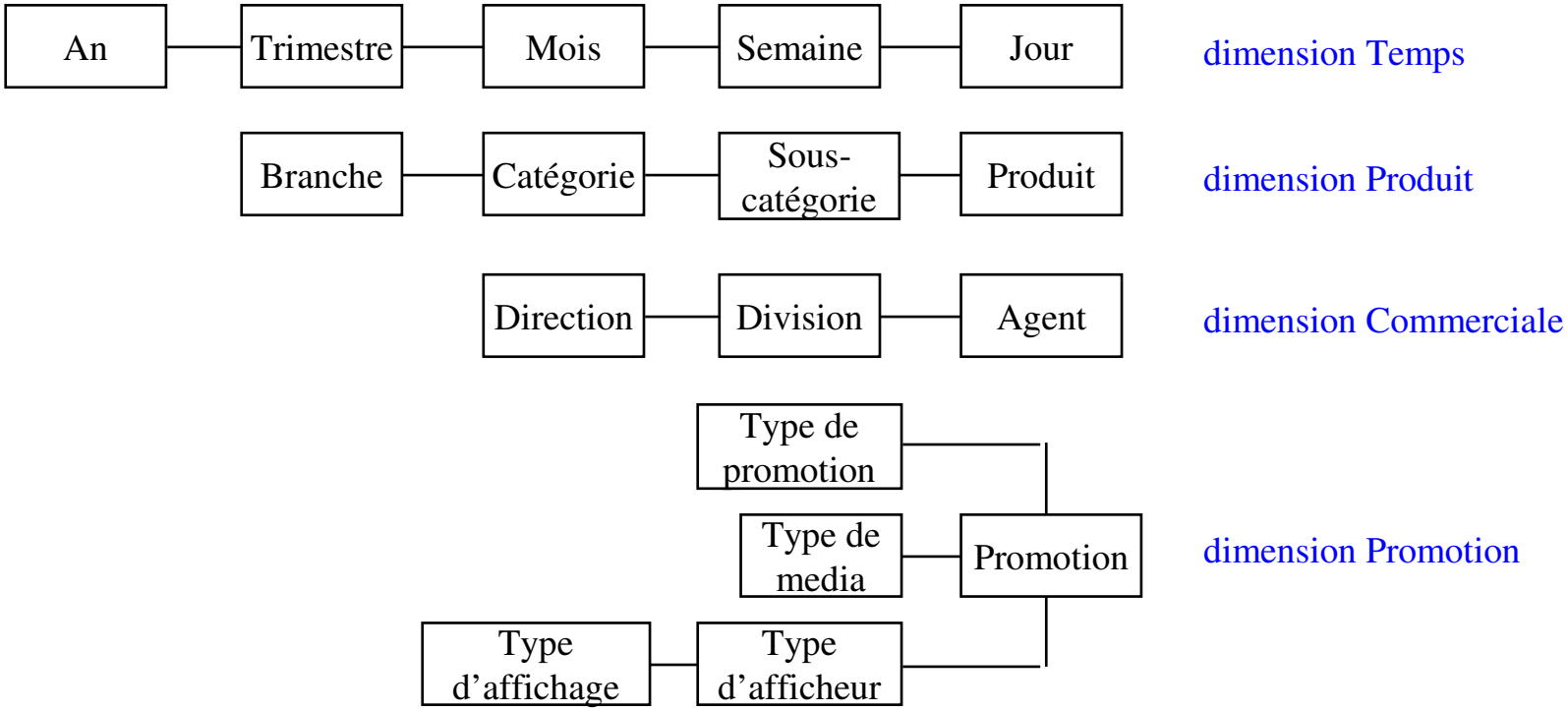
- mémoriser les transformations pour retrouver le lien données opérationnelles / données DW

5. Intégration

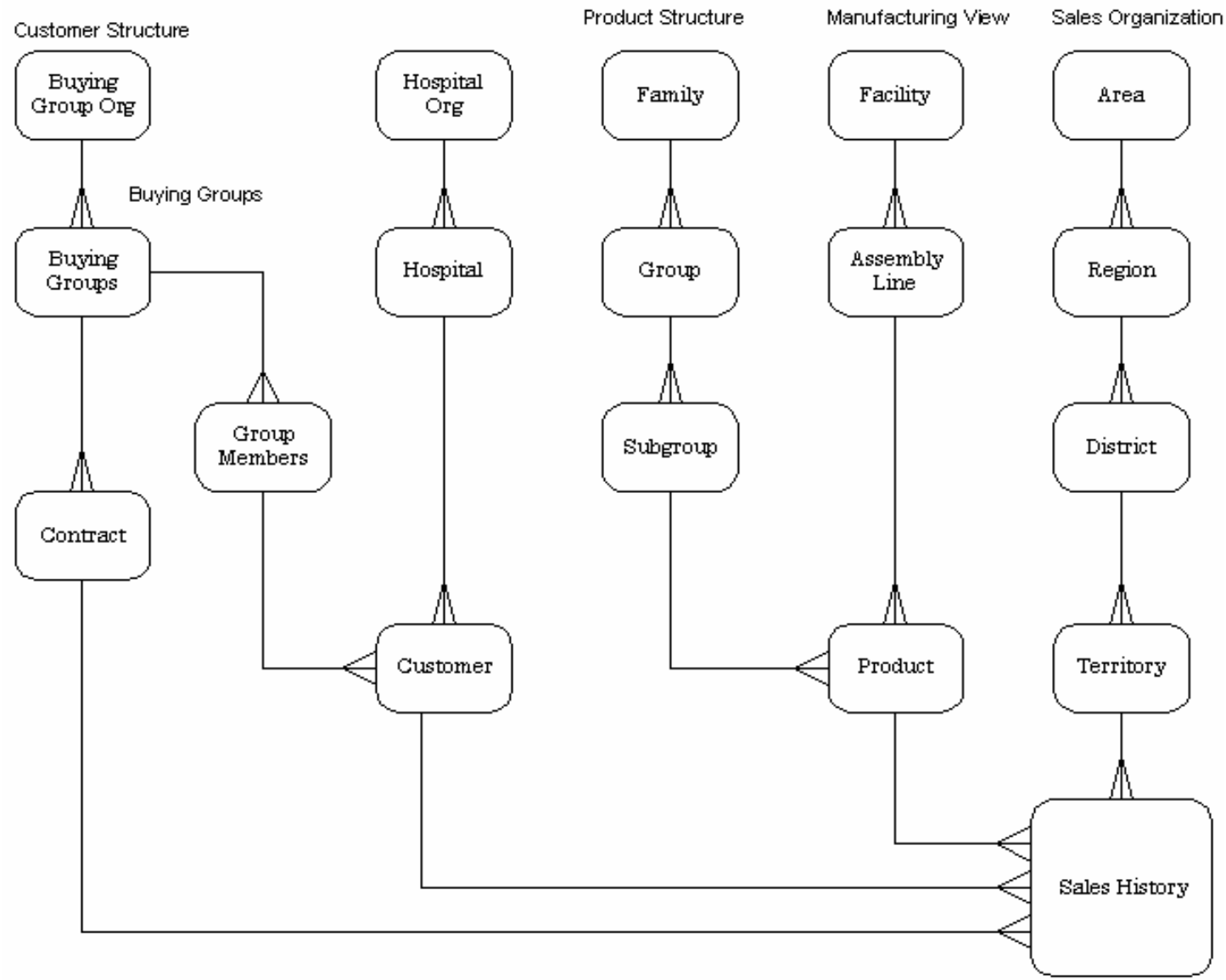
- conflit de classification
  - » objets de sémantiques voisines mais comportant certaines propriétés différentes
  - » Solution : soit établir une relation IS-A, soit opérer une fusion entre les deux objets.
- conflit de description
  - » représentation différente des propriétés, à savoir des identifiants différents, des formats différents d'attributs identiques,... :
  - » Solution : choisir une des deux représentations, la plus logique, la plus cohérente avec le reste du modèle, pour exprimer le résultat de l'intégration.
- conflit de structure
  - » l'attribut d'un modèle correspond à l'entité d'une autre ou un attribut à une association, ou une entité à une association
  - » Solution : passer par une étape de transformation entité/attribut ou entité/association

- Il est très important de mémoriser les transformations opérées afin de garder une trace permettant de retrouver le lien entre un élément du Data Warehouse et les données correspondantes des bases opérationnelles.

5. Modélisation : les dimensions

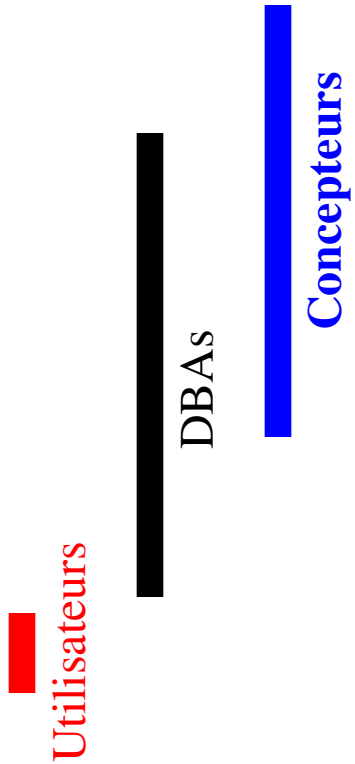


5. Modélisation : les dimensions



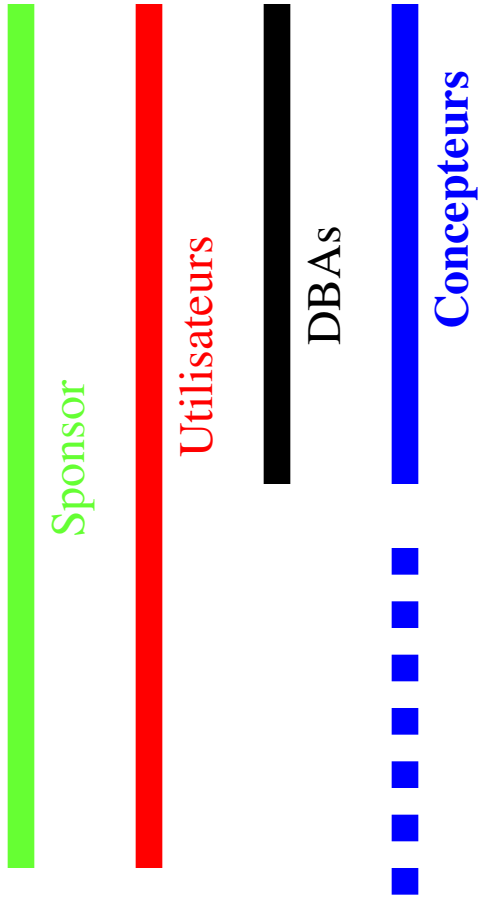
5. Finalisation

- Derniers développements
- Tests
  - » premier chargement du DW sur site
  - » tests
- Mise en exploitation
  - » chargements réguliers
  - » utilisation «au quotidien»



5. Evolution

- Evaluation
  - » mise en œuvre régulière
  - » utilisation
  - » confrontation aux retours attendus
  - » évaluations à différentes échéances
- Evolution
  - » suite du projet :
    - ajout de fonctionnalités ?
    - ajout de Data Marts ?



6. Les outils

- Un marché fragmenté :
  - Constitution du DataWarehouse
  - Stockage
  - Extraction d'Information

6. Constitution du DataWarehouse

- Administration
  - » SourcePoint (Software AG), ISM/OpenMaster (Bull), CA-UniCenter, DataHub (IBM), CPE (SAS), Warehouse Administrator (SAS)
- Extraction et Purification
  - » Warehouse Manager (Prism), Integrity Data Reengineering (Vality), Access (SAS), DataStage (VMark), Génio (Léonard's Logic), InfoRefiner (Platinum), PASSPORT et NATURAL (Software AG), Gentia (Planning Sciences)

6. Stockage

- DataWarehouse
  - » Oracle, Sybase, Informix, Ingres (CA), DB2 (IBM), Tandem, Teradata, ...
- Serveur OLAP
  - » Express (Oracle), Business Objects, Powerplay / Impromptu (Cognos), Adabas (Software AG), Opera (CFI), ALEA (MIS AG), Harry Cube (Adviseurs), Gentia (Planning Sciences), Essbase (Arbor Software), Informix, Pilot, ...

6. Extraction d'Information

- Rétro-ingénierie (Reverse-Engineering)
  - » Business Object, DB-Main
- Browser OLAP
  - » Discoverer (Oracle), ESPERANT (Software AG), InfoBeacon (Platinum), Explorer (Business Objects), le **VCL DecisionCube de Delphi Cl/Sv**
- Arbres de Décision
  - » Alice (ISoft), Knowledge Seeker (Angoss), Chaid (SPSS)
- Réseaux de Neurones
  - » Predict (Neuralware), Neural Connection (SPSS), Previa (Elseware)
- Autres
  - » Mineset (SGI), Darwin (Thinking Machines), Gupta DataMind (basé sur les réseaux d'agents), Discovery Server (Pilot), DSS Agent (Micro Strategy), BusinessMiner (Business Objects), Intelligent Miner (IBM), ...

7. Perspectives du Data Warehouse

- homogénéisation
  - » des outils intégrant les différentes étapes de la suite décisionnelle
- données externes
  - » ouverture à l'internet
- augmentation des volumes de données
- restitution des informations :
  - » nouvelles techniques de data mining
  - » multimédia
- outils de constitution du référentiel
  - » la métabase

8. Bibliographie - Livres

- J.-M. Franco, «Le Data Warehouse / Le Data Mining», Eyrolles, 1997
- J.-M. Franco, S. De Lignerolles, «Piloter l'entreprise grâce au data warehouse», Eyrolles, 2000.
- R. Mattison, «Data Warehousing - Strategies, Technologies and Technics», IEEE Computer Society, 1996.
- W. H. Inmon, «Building the Data Warehouse», ed. Wiley
  - 1ère édition : 1996, 3ème édition: 2002, voir http://www.billinmon.com/
- W. H. Inmon, «Managing the Data Warehouse», ed. Wiley, 1997
- R. Kimball, «Entrepôts de Données», Intl Thomson Pub., 1997.
- Ralph Kimball, Laura Reeves, Warren Thornwaite, « The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses », 800 pages (août 1998), Ed Wiley, ISBN: 0471255475
- Ralph Kimball, Margy Ross, « Entrepôts de données. Guide pratique de modélisation dimensionnelle », 2ème édition (1 janvier 2003), Ed Vuibert, 2-7117-4811-1

8. Bibliographie - Livres

- R. Michalski et al., "Apprentissage symbolique.", Cépaduès, 1993.
- Patrick Becker, Ann Becker, Patrick Naïm, Les Réseaux bayésiens : Modèles graphiques de connaissance, Ed Eyrolles, 1999

Bibliographie

- Surajit Chaudhuri, Umeshwar Dayal: An Overview of Data Warehousing and OLAP Technology. SIGMOD Record 26 (1): 65-74 (1997)

8. Bibliographie - WWW

- http://www.dw-institute.com/
  - The Data Warehouse Institute
- http://pwp.starnetic.com/larryg/
  - Infos dont accès à des livres blancs sur le DW
- http://www.promotheus.eds-fr/themes/dw/
  - Institut Promotheus, thème DW
- http://www.cait.wustl.edu/cait/papers/prism/
  - Société Prisme fondée par W.H. Inmon
- http://www.olapcouncil.org/
  - Outils OLAP
- http://www.valoris.fr/amplitude/j101.htm
- http://www.mediatid.fr/datawarehouse
  - forum sur le Data Warehouse

8. Bibliographie - Recherche

- ACM SIGMOD
- VLDB
- Data Warehousing and Knowledge Discovery (DaWaK)
  - » Conférence scientifique spécialisée
- ACM SIG KDD (Knowledge Discovery and Data Mining)
  - » Conférence scientifique spécialisée
- DOLAP