ENTREPÔTS DE DONNEES
ARCHITECTURE ET FONCTIONNALITES

Mokrane Bouzeghoub

http://www.prism.uvsq.fr

UNIVERSITE DE VERSAILLES SAINT-QUENTIN-EN-YVELINES

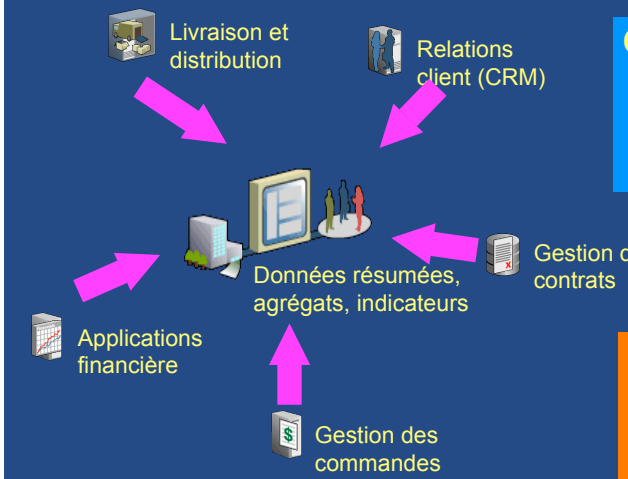
Laboratoire PRISM

CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Sommaire

- Notions de SI opérationnel et de SI décisionnel
- Architecture d'un entrepôt de données
- Différents mode de représentation des données multidimensionnelles
- Opérations sur les données multidimensionnelles
- Problèmes d'hétérogénéité des données
- Chargement et rafraîchissement d'un entrepôt de données
- Méta données et méta modèles de DW
- Autres architectures d'intégration de données

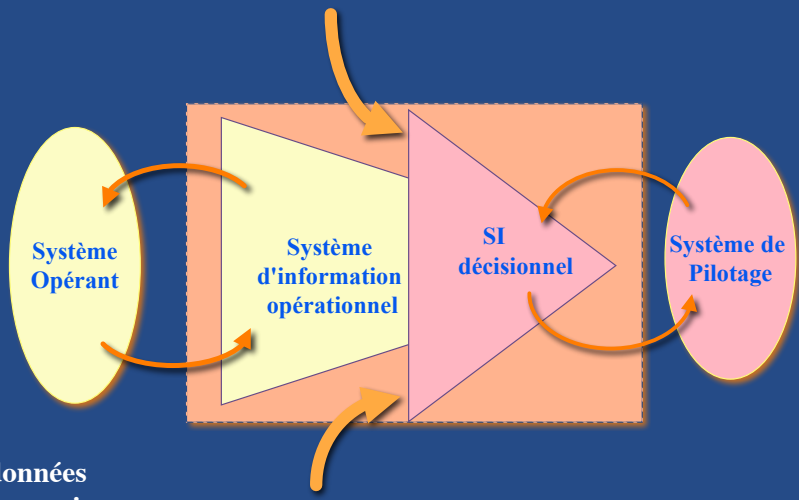
Besoin d'intégration de données



Contexte:
- données distribuées
- sources autonomes
- données hétérogènes
- ...

But
- Réduire le temps d'acquisition des données
- Transparence de l'hétérogénéité
- Flexibilité, évolutivité

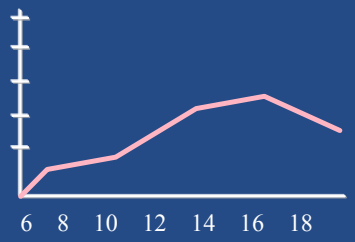
Système d'information opérationnel v.s système d'information décisionnel



- Déluge de données
- Pénurie de connaissances sur ces données

Notion de donnée et notion d'information

Date	Heure	Numéro	Destinat	Durée	Coût
5-3	07:05	00216188	Tunisie	04:08	6.30
5-3	16:12	00216188	Tunisie	08:10	11.50
6-3	09:40	00441216	UK	10:20	16.45
6-3	20:20	04426576	BdRhnes	16:30	8.40



Exemple: un DW dans les télécoms

- Sujets
  - suivi du marché: lignes installées/ désinstallées, services et options choisis, répartition géographique, répartition entre public et différents secteurs d'organisations
  - comportement de la clientèle
  - Comportement du réseau
- Historique
  - 5 ans pour le suivi du marché
  - 1 an pour le comportement de la clientèle
  - 1 mois pour le comportement du réseau
- Sources
  - fichiers nouveaux clients élaborés par les agences régionales
  - fichier facturation de l'entreprise
  - sources externes: études INSEE

Importance du choix du granule et du volume des données

- **Le granule affecte le volume du DW et le type de requêtes**
  - Facturation détaillée dans l'historique

    200 appels/mois, 50 car/appel, sur 2 mois, Total 20 000 car/abonné
  - Sans facturation détaillée

    un seul article de 50 car / abonné dans l'historique

  - Mais les deux choix ne permettent pas la même analyse
- **Le volume du DW détermine le type de serveur de BD à utiliser et les développements à réaliser**
  - AT&T: 20TB pour le suivi des appels sans fil
  - Wal-Mart: BD consommateurs: 70TB

Contraintes

- le DW doit être totalement indépendant des systèmes opérationnels
- les tâches d'alimentation, rafraîchissement et calcul sont totalement asynchrones
- les données provenant des sources internes passent par le réseau
- Les sources externes sont recopiées dans leur globalité
- les modifications sur les sources internes ne sont pertinentes que si elles représentent plus de 1% du volume total des données.
- Les modifications sur les sources externes sont acquises périodiquement une fois par mois.

Requêtes

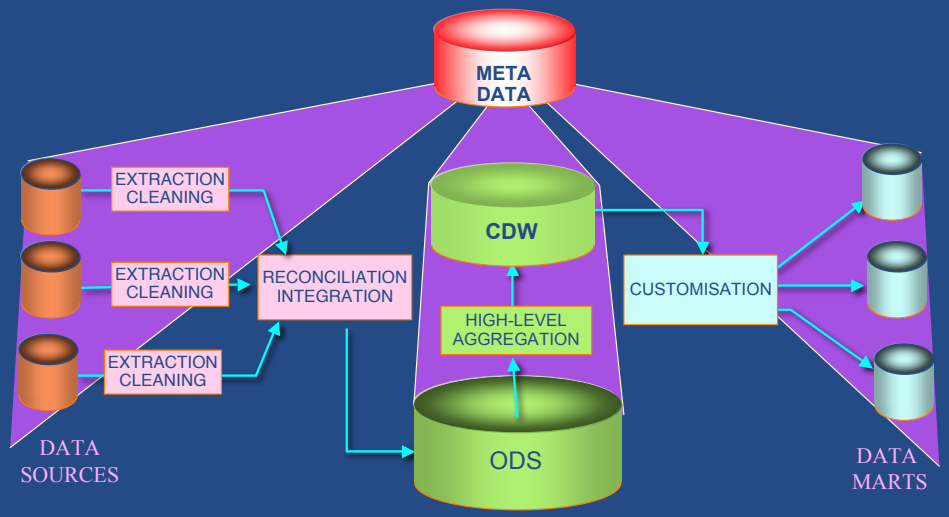
- **Sujet: Comportement clientèle**
  - –Nombre moyen d'heures par client, par mois et par région
  - –Répartition des appels clients sur la semaine
  - –Répartition des appels clients sur la journée
  - –Nombre moyen de numéros appelés représentant 20% d'une facture, 50% d'une facture
  - –Durée moyenne d'une communication urbaine par ville
  - –Durée moyenne d'une communication internationale

Application de production v.s application d'aide à la décision

–Les applications de production sont constituées de traitements factuels concernant les produits, les ressources ou les clients de l'entreprise
- OLTP: On Line Transaction Processing

–Les applications d'aide à la décision sont constituées de traitements ensemblistes réduisant une population à une valeur ou un comportement
- OLAP: On Line Analytical Processing

Architecture d'un DW



Contenu d'un DW

– Des données historisées
- fournies par les sources
- archivées dans l'ODS

– Des données agrégées
- par des fonctions de calcul
- Par des algorithmes de data mining
- Par des techniques de résumés

– Des métadonnées
- décrivant la structure des données de base ou agrégées
- donnant des explications sur la qualité, le mode de dérivation, la duréee de vie, le rafraichissement, etc...

Caractéristiques des données d'un DW

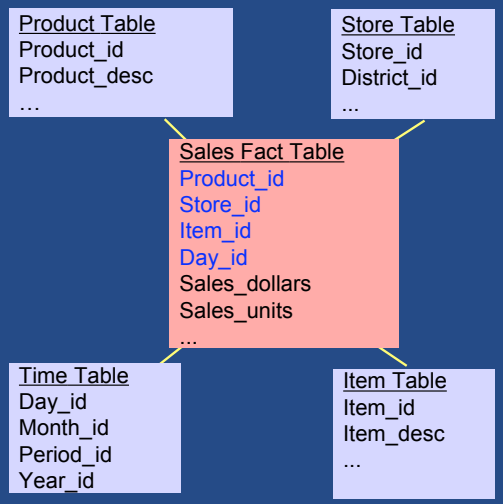
- Orientées sujet
  - Clientèle
  - produit/production
  - marché/transaction
  - politique/stratégique
- Intégrées
  - codage
  - format
  - structure
- Non volatiles
  - pas de mise-à-jour directement
  - chargement en masse

Représentation d'un entrepôt de données

- **Représentation conceptuelle**
  - schémas en étoile
  - en flocons
  - en constellation
- **Représentation logique**
  - tables
  - Cubes
- **Représentation physique**
  - vues matérialisées

Le schéma en étoile

- Une table « faits » par indicateur
- Une table dénormalisée pour chaque dimension
- associations en étoile entre les tables « faits» et « dimensions »



Le schéma en flocons

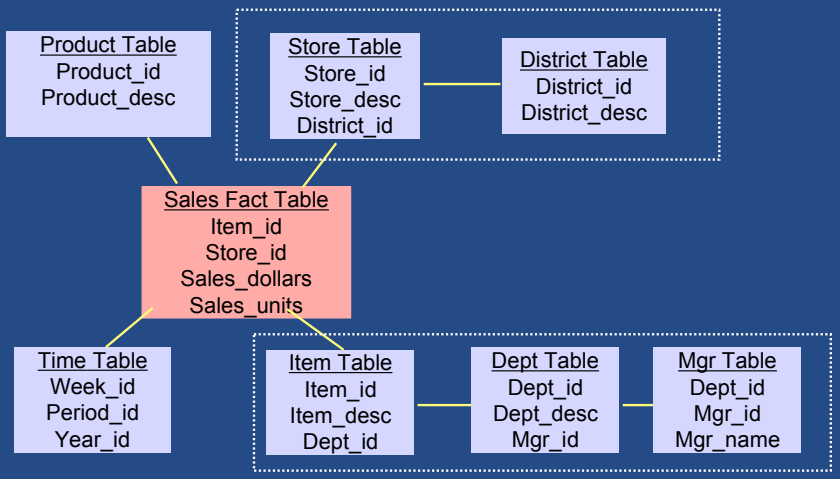
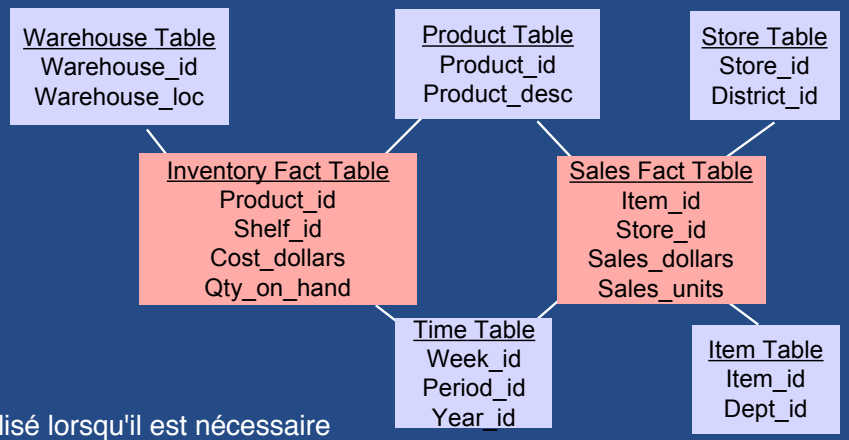


Schéma en constellation



Utilisé lorsqu'il est nécessaire d'avoir plusieurs tables factuelles

Choix des tables « faits » et « dimensions »

- **Analyse des requêtes**
  - attributs « group-by » indiquent les dimensions
  - attributs agrégés indiquent les mesures
  - attributs « where » sont les attributs des tables factuelles ou dimensionnelles

- **Exemple :**

  select **sale.store_id, sale_product_id, sum (sale.price)**
  from **product P, sale S**
  where **P.product_id=S.product_id** and **P.product_desc = « clothes »**
  group by **store_id, product_id**

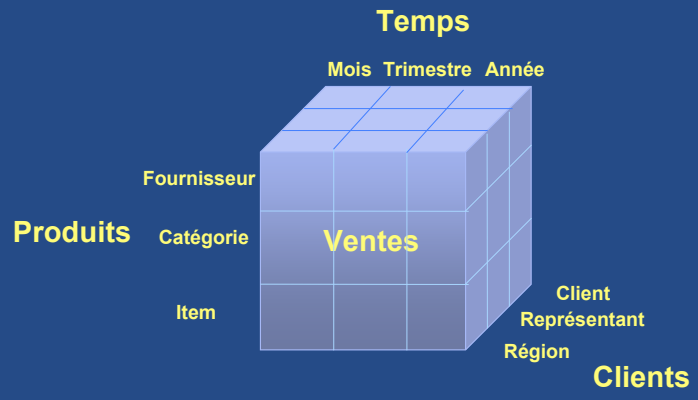
Représentation logique

DuréeMoy	Départ.	Mois	Région
5	Info	Janv	IdF
5	Phys	Janv	IdF
18	Philo	Janv	IdF
7	Droit	Janv	IdF
12	Info	Févr	IdF
8	Phys	Févr	IdF
9	Philo	Févr	IdF
15	Droit	Févr	IdF
18	Info	Mars	IdF
12	Phys	Mars	IdF
22	Philo	Mars	IdF
25	Droit	Mars	IdF

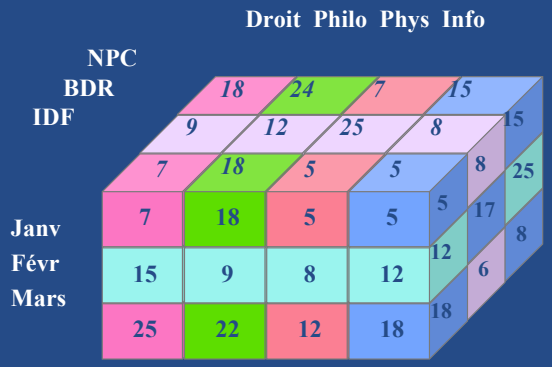
Dimensions: Attributs, Tuples

ou : Départ., Mois, Région

Représentation abstraite et algèbre de cubes

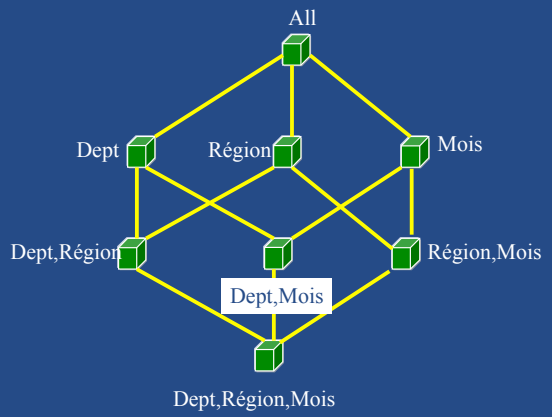


Exemple



Généralisation de la notion de cube

- Etant donné un cube à N dimensions, il est possible de dériver tous les cubes de dimension N-1, N-2,…, N-N, (=> Treillis de 2^n cubes).

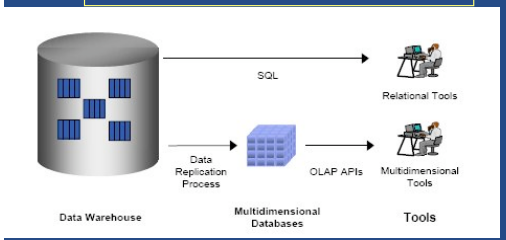


Problématiques

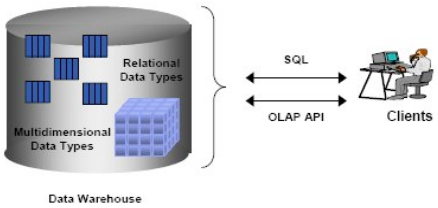
Questions :

- Comment implémenter ces cubes ?
- Comment stocker les données agrégées ?
- Comment accéder aux données par les dimensions des cubes ?

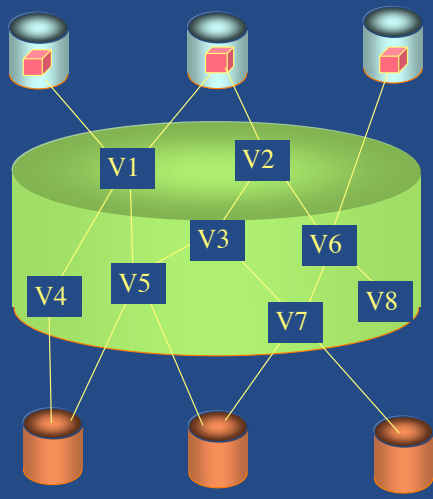
Approche :
Bases de données spécialisée



Approche :
Base de données OLAP-Ready

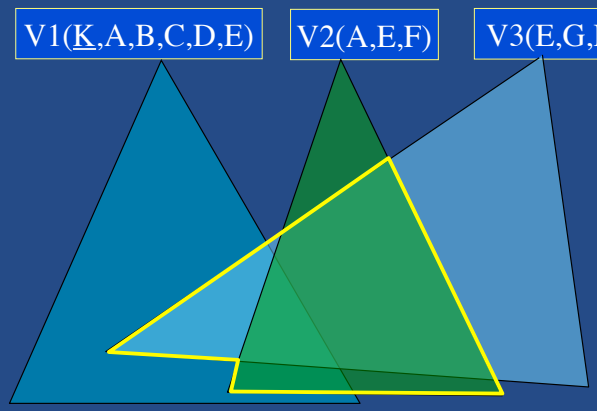


Représentation interne: vues matérialisées



- L'entrepôt de données est perçu comme un ensemble de vues
  - matérialisées
  - ou abstraites

Le choix de la matérialisation des vues



La matérialisation est une technique de cache qui doit optimiser:

- coût d'évaluation
- coût de maintenance
- coût de stockage
- fraîcheur des données

R1(K,A,B,X) R2(Y,C,D,E,W) R3(Y,C,D,E,Z)

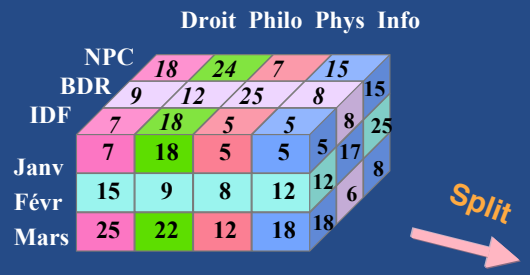
Opérations sur les cubes

- Rotate / Pivot
- Switch
- Split
- Nest / Unest
- Push / Pull

- Roll-up (grain supérieur)
- Drill-down (grain inférieur)
- Slice (Projection)
- Dice (Sélection)

- Jointure
- Union
- Intersection
- Différence

Décomposition (Split)



Droit	IDF	BDR	NPC
Janv	7	9	18
Févr	15	<>	<>
Mars	25	<>	<>

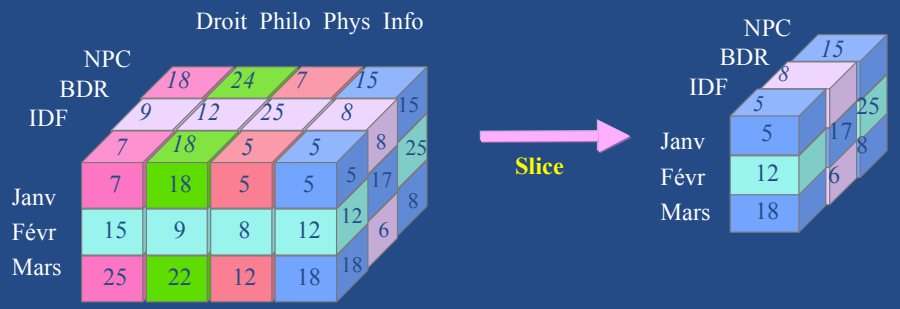
Info	IDF	BDR	NPC
Janv	5	8	15
Févr	12	17	25
Mars	18	6	8

Phys	IDF	BDR	NPC
Janv	5	25	7
Févr	8	<>	<>
Mars	12	<>	<>

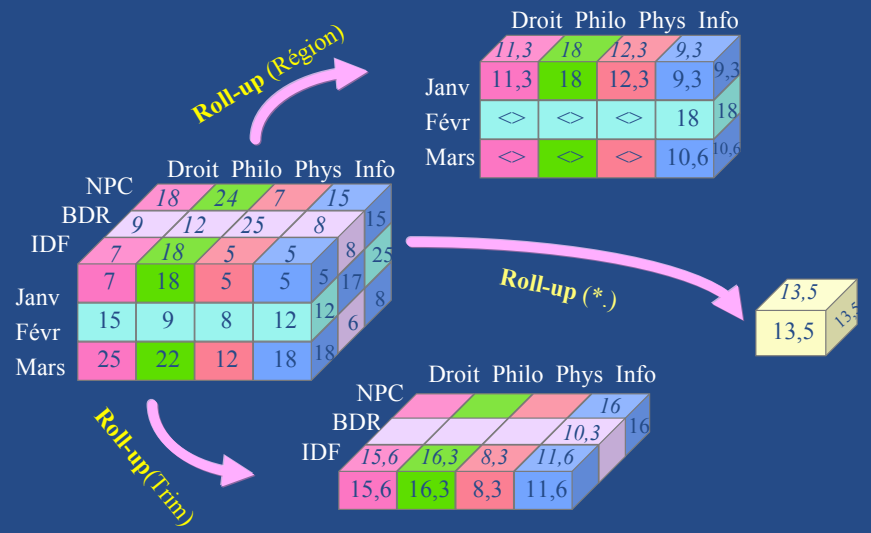
Philo	IDF	BDR	NPC
Janv	18	12	24
Févr	9	<>	<>
Mars	22	<>	<>

Projection (Slice)

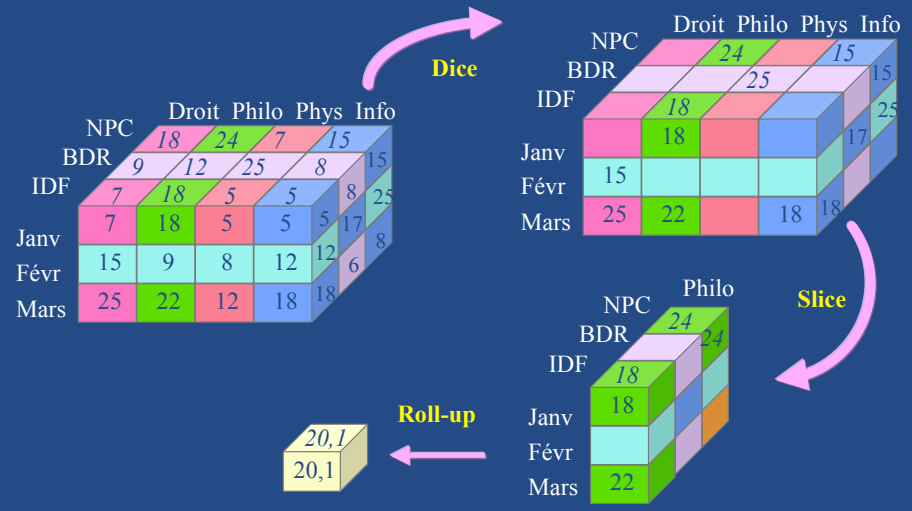
S'applique sur les valeurs d'une dimension



Changement de granule (Roll-up)

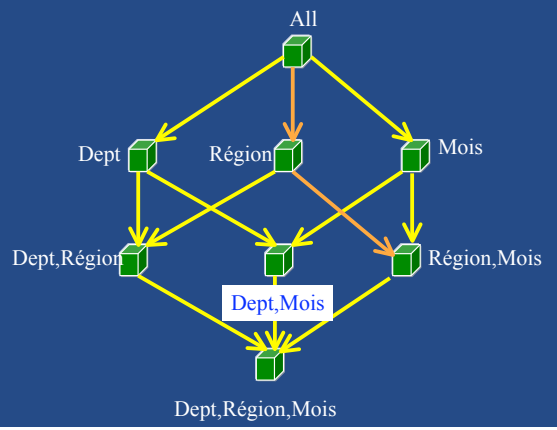


Composition d'opérations (Dice - Slice - Roll-up)



Drill-Down

- Opération inverse de Roll-up.
- S'applique sur le treillis des cubes; fait passer d'un granule élevé à un granule plus détaillé.



Langages OLAP

- Extensions de SQL

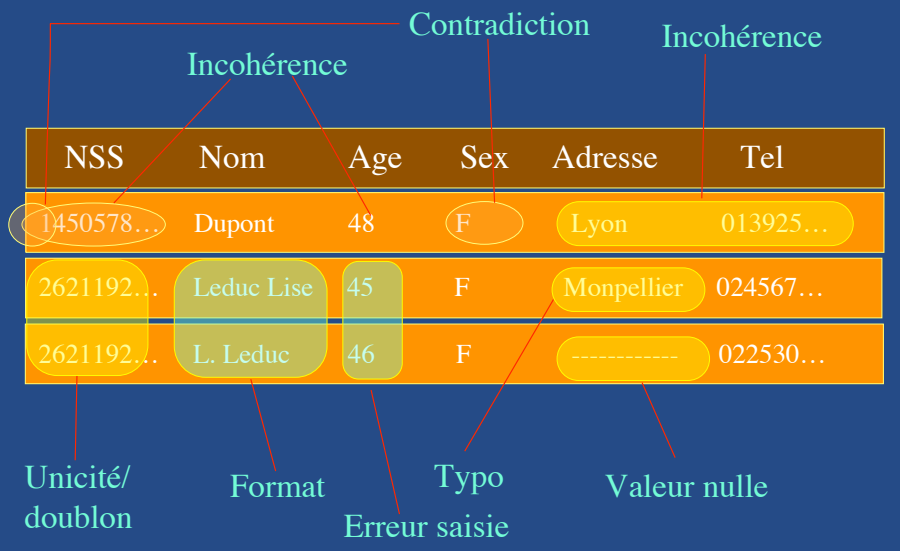
```
select t1.name, t2.annee, sum(t2.montant)
  from region t1, vente t2
  where t1.name_id = t2.region_id
  group by
    cube ( t1.name,t2.annee)
  order by 1,2;
```

Gestion de l'hétérogénéité des données

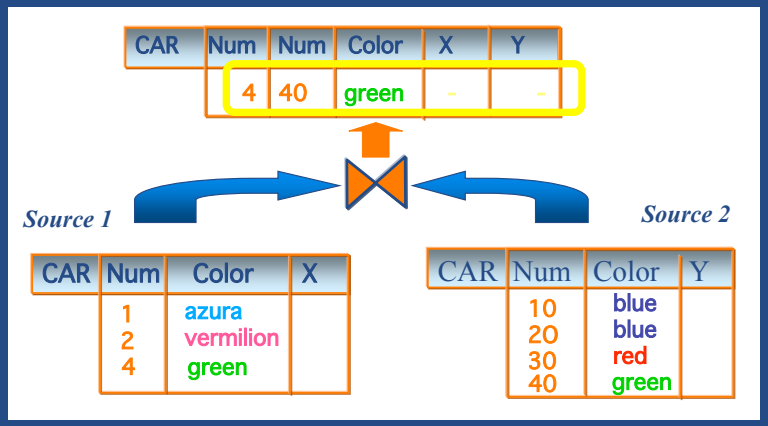
- types d'anomalies
- anomalies mono-sources
- anomalies multisources

- Quelle transformations appliquer?
- Comment organiser/synchroniser ces transformations ?

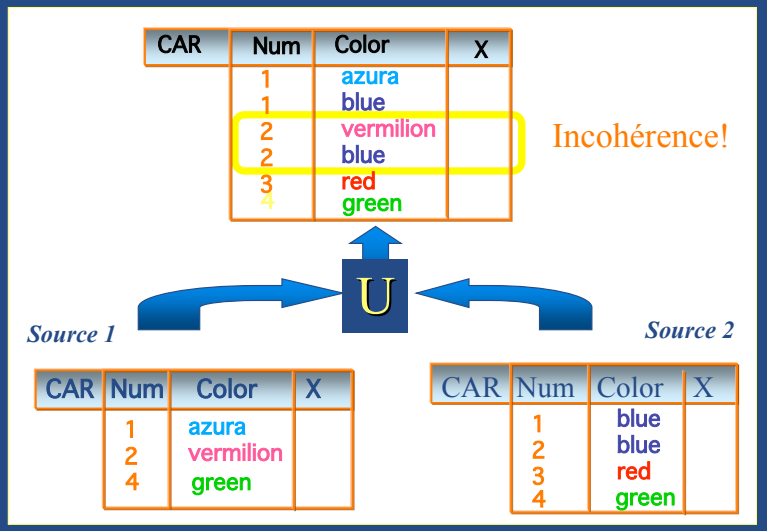
Exemples de problèmes sur une source



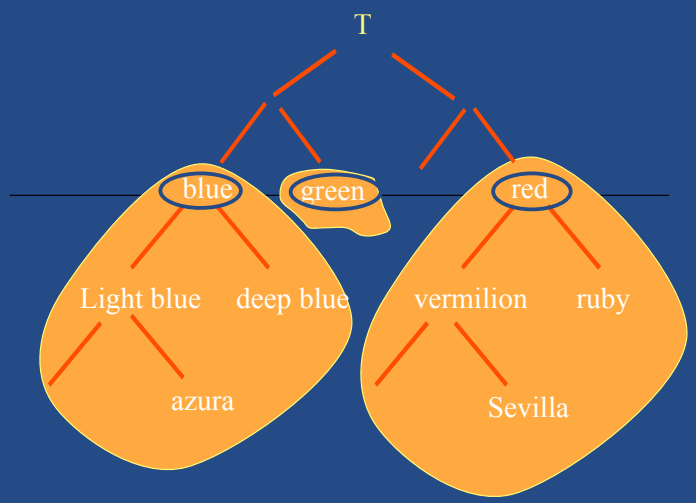
Exemple de problème sur deux sources (jointure)



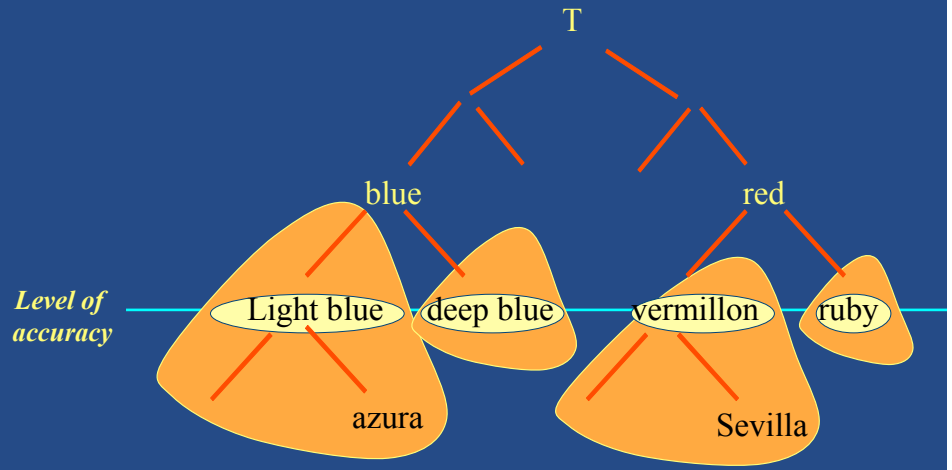
Exemple de problème sur deux sources (union)



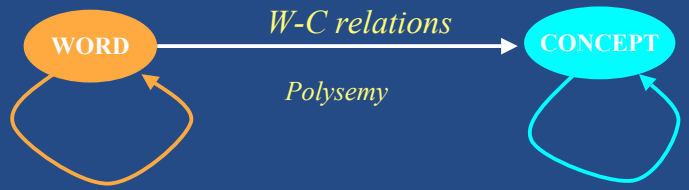
Approche de résolution: Utilisation de graphes conceptuels



Définition du niveau de précision



Ontologie = mots + concepts + liens sémantiques



Lexical Relations

Synonymy
Nominal/Verbal

Semantic Relations

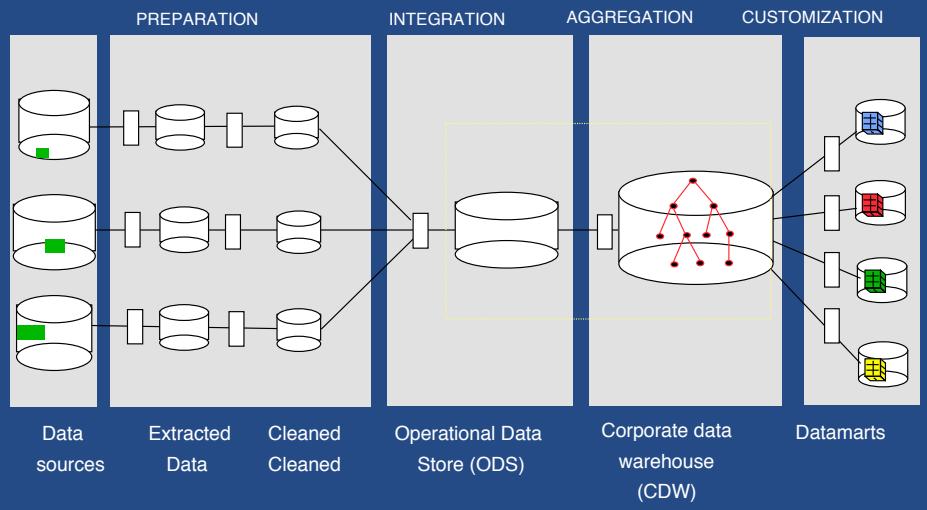
Hyponymy/Hypernymy (isa)
Meronymy/Holonymy (partof)
Casual, Spatial, Attr
Converse

Axiomes: Disjointness, covering, narrow/broader

Fonctionnalités de base des ETL

- **Extraction de données sources (E)**
  - Traitement de l'hétérogénéité des systèmes ⇒ tables sources
  - Extraction d'un cliché des données
  - Extraction des changements survenus depuis la dernière fois
- **Transformation (T)**
  - Graphe orienté acyclique ou programme structuré (type L4G) contenant des opérations de transformations de données
- **Chargement (L)**
  - Chargement du résultat d'un processus de transformation de données dans un système cible (le + souvent une BD)
- **Planification**
  - Regroupement de plusieurs processus de transformation/chargement
  - Planification calendaire ou événementielle de leur exécution

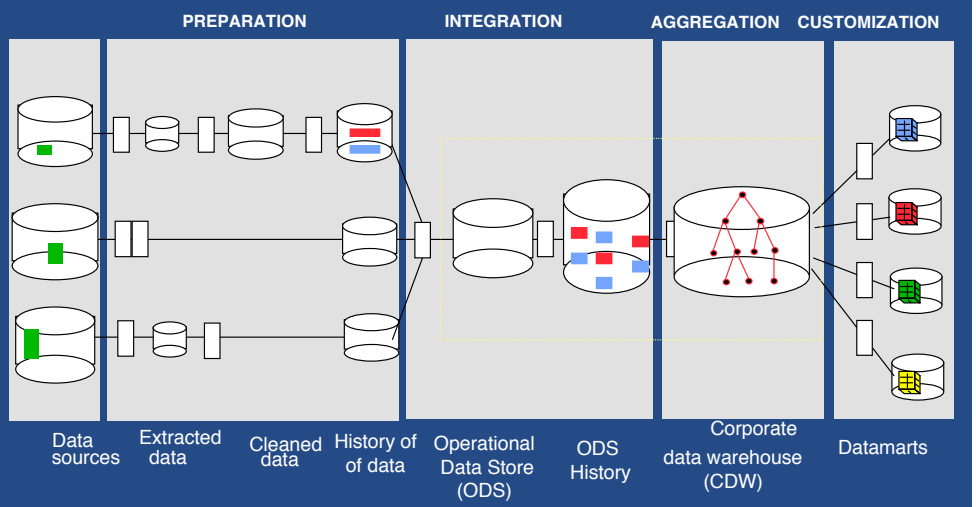
La phase de chargement initial



Caractéristiques

- **C'est la phase ultime de la conception du DW. Avant cette phase le DW n'existe pas pour les utilisateurs**
  - Pas de contrainte sur les temps de réponse
  - Nécessite plus de disponibilité des sources
- **Les 4 phases de chargement peuvent être exécutée séquentiellement ou planifiée dans le temps**
  - avec un certain parallélisme dans la phase préparatoire
- **Le scénarion de chargement initial est défini statiquement.**

La phase de rafraîchissement et de maintenance



Caractéristiques

- **Il peut y avoir un asynchronisme complet entre les différentes activités de rafraîchissement**
  - un niveau de parallélisme élevé dans la phase de préparation
  - chaque sources a sa propre disponibilité (fenêtre d'accès)
  - chaque source a sa propre stratégie d'accès et d'extraction (pull, push)
  - le nettoyage des données peut être monosource et multisource
- **Il n'y a pas un modèle de planification unique pour les activités de rafraîchissement**
  - dépendant de chaque type d'application (utilisateurs)
  - évolutif dans le temps

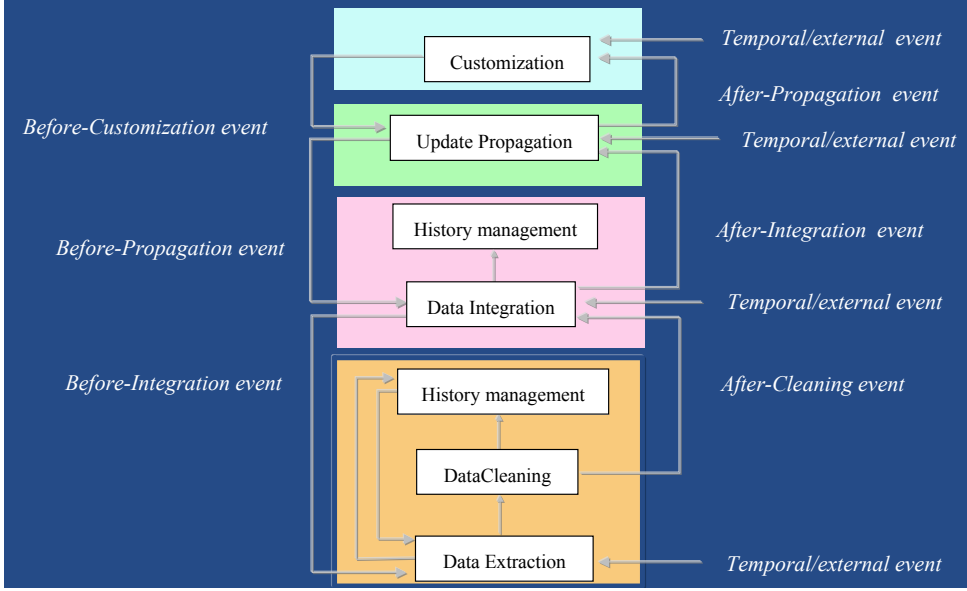
Modèle de planification

- **Un processus de chargement ou de rafraîchissement est vu comme un ensemble d'activités organisées et coordonnées**
  - exécutées par la machine ou par un humain
  - flexible pour subir des modifications
  - évolutif dans le temps
- **Les Workflow sont des modèles adaptés pour la planification de tâches**
  - utilisés pour composer des services web (BPEL)
  - business process (re)engineering (BPR)
  - travail coopératif (CSCW)

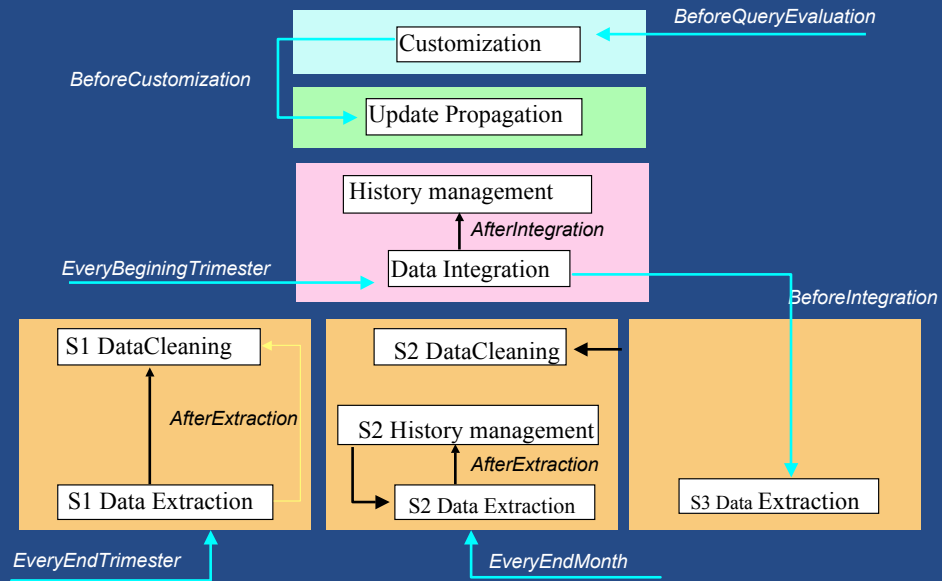
Les Workflows

- **Fournissent un cadre de représentation conceptuel**
  - –facilitant la conception
  - –facilitant la compréhension et l'échange de spécifications
- **Fournissent un modèle de référence**
  - –facilitant le raisonnement
  - –offrant de multiples support pour l'étude de diverses propriétés
    - •des diagrammes d'éta-transition, des réseaux de Petri, des règles actives

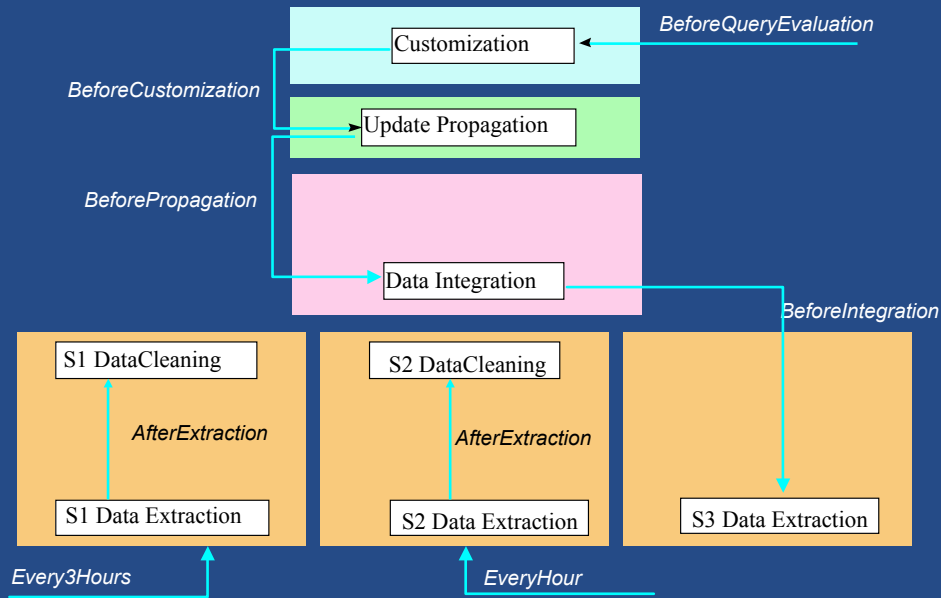
Type de tâches et types d'événements



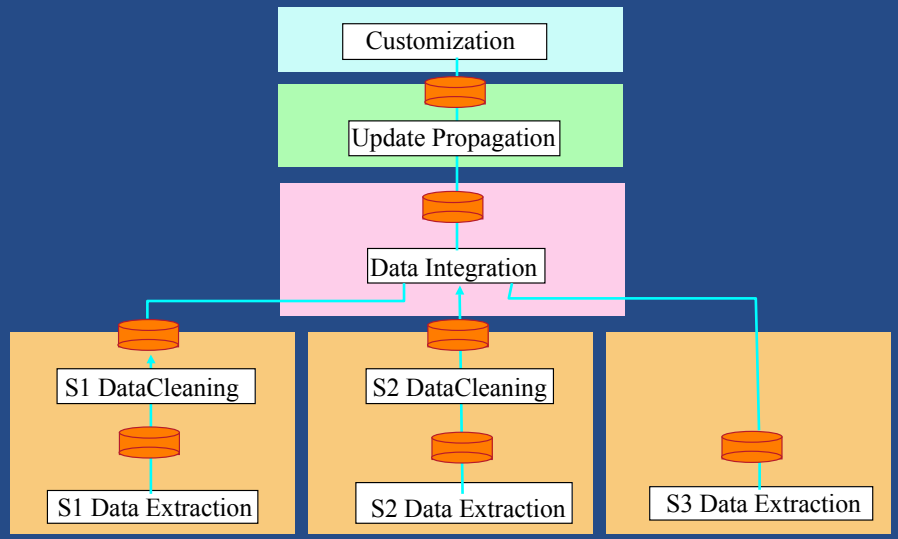
Scénario de rafraîchissement (1)



Scénario de rafraîchissement (2)



Scénario de rafraîchissement (3)



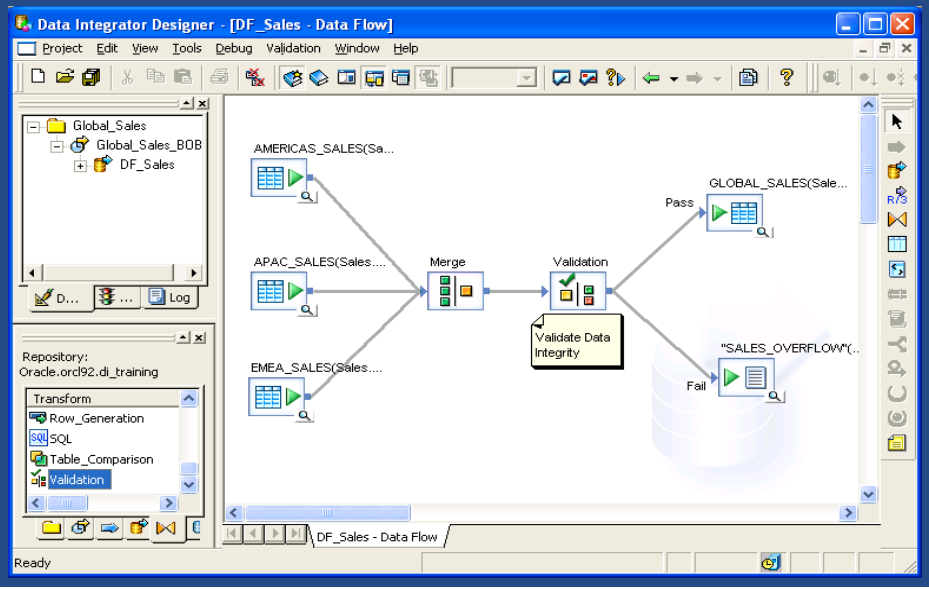
Atouts des workflows

- **Flexibilité**
  - permettent de coordonner des activités avec des sémantiques encapsulées
  - permettent de décomposer/recomposer récursivement les activités
  - permettent une réorganisation dynamique de l'orchestration
- **Intuitivité**
  - facilité d'utilisation
  - facilité de lecture / validation
- **Support d'évaluation de la qualité**

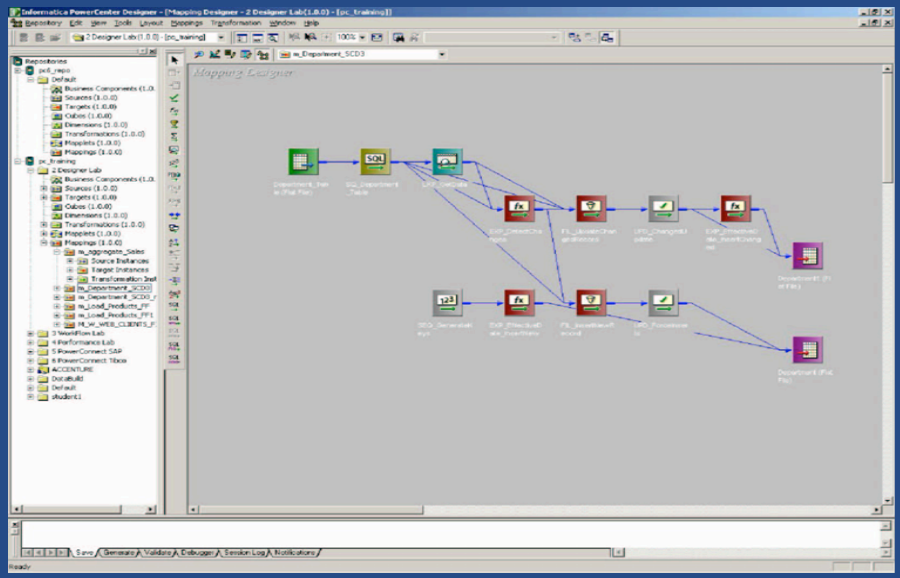
Principaux produits d'ETL

Produit	Transformation	Exécution	Contrôles	Prix
Data Stage	Graphe + langage type Basic; mode pas à pas; saisie interactive	Interprétation; exécution parallèle et flux tendu	Explicites + logs; nettoyage avec Quality Stage	100 K€ +
Informatica	Graphe complet + appels à des procédures; pas à pas	Interprétation; exécution parallèle et flux tendu	Explicites + logs;	100 K€ +
Genio	Langage L4G + appels procédures	Interprétation	Analyse d'impact	50 K€
Sunopsis	Graphe complet + appels à des procédures; pas à pas	Interprétation	Explicites et modules de connaissance	35 K€
Sagent	Graphe complet + appels à des procédures; pas à pas	interprétation	Explicites + logs; Address Cleanser, Merge and Purge	50 K€
BO Data Integrator	Graphe complet + appels à des procédures; pas à pas	Interprétation, parallélisation	Explicites + logs	75 K€ +

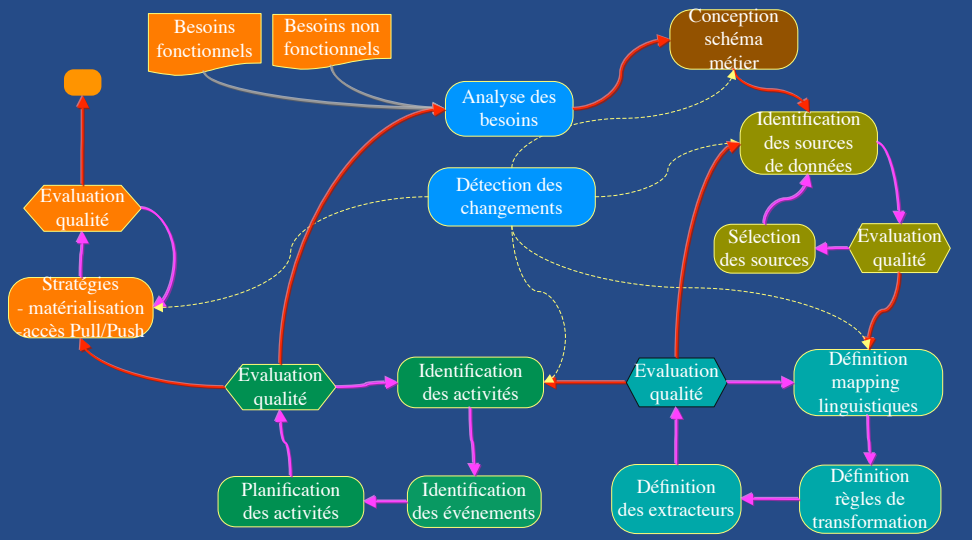
BO Data Integrator



Informatica



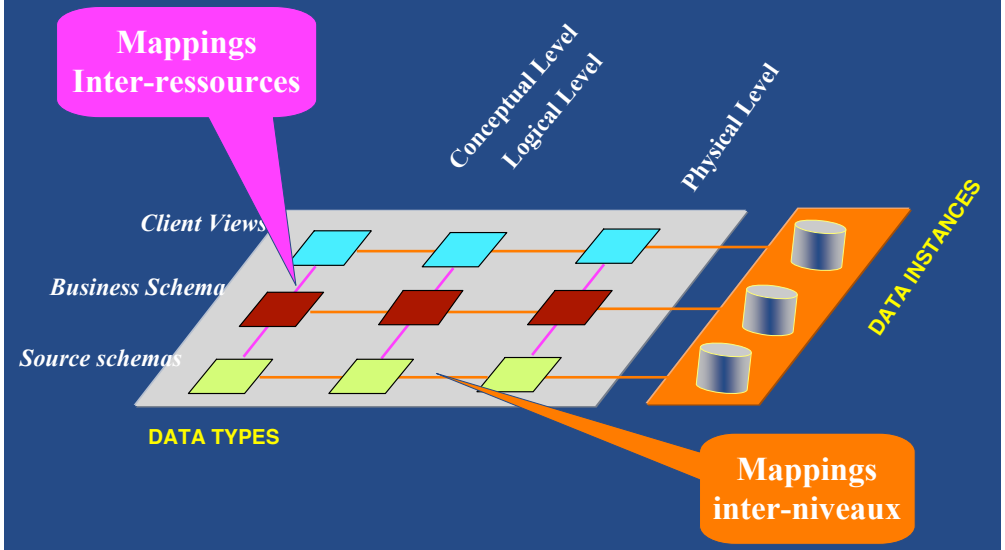
Synthèse: Principales tâches de conception d'un DW



Les métadonnées d'un entrepôt de données

- **Différentes vues des ressources**
  - schémas des sources de données
  - schéma métier (entreprise)
  - schémas clients (utilisateurs)
- **Différents niveaux d'abstraction**
  - conceptuel, logique, physique
- **Mappings inter-niveaux d'abstraction**
  - transformation de schémas
- **Mappings inter-ressources**
  - linguistiques, opérationnels

Vue d'ensemble des méta données



Autres méta données

- Nommage des données (lexique)
- Formattage des données (format caché)
- Contraintes sur les données (explicites ou implicites)
- Historique des évolutions des données
- Statistiques sur l'utilisation des données (fréquence accès/màj, sélectivité)
- Qualité des données
  - …

Importance et utilité des méta données

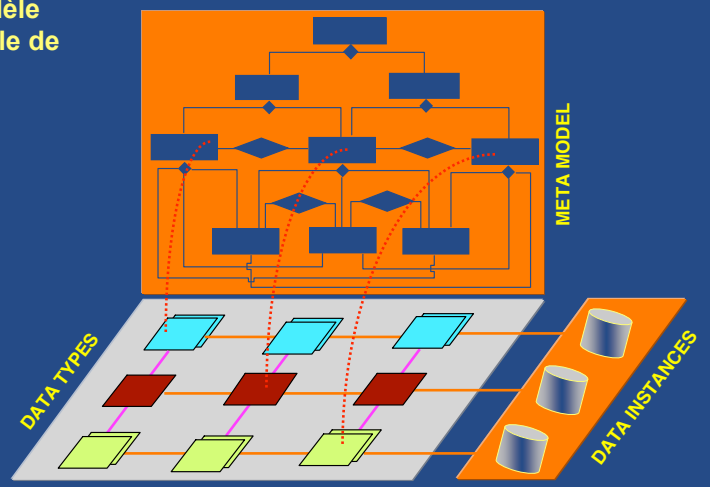
- Richesse des méta données
  - réduit les coûts de développement
  - limite les erreurs d'interprétation des données
- Permettre l'interopérabilité opérationnelle entre plusieurs systèmes hétérogènes
- La gestion efficace des méta données est un outil de maitrise de la complexité des systèmes d'intégration de données

Pourquoi un métamodèle?

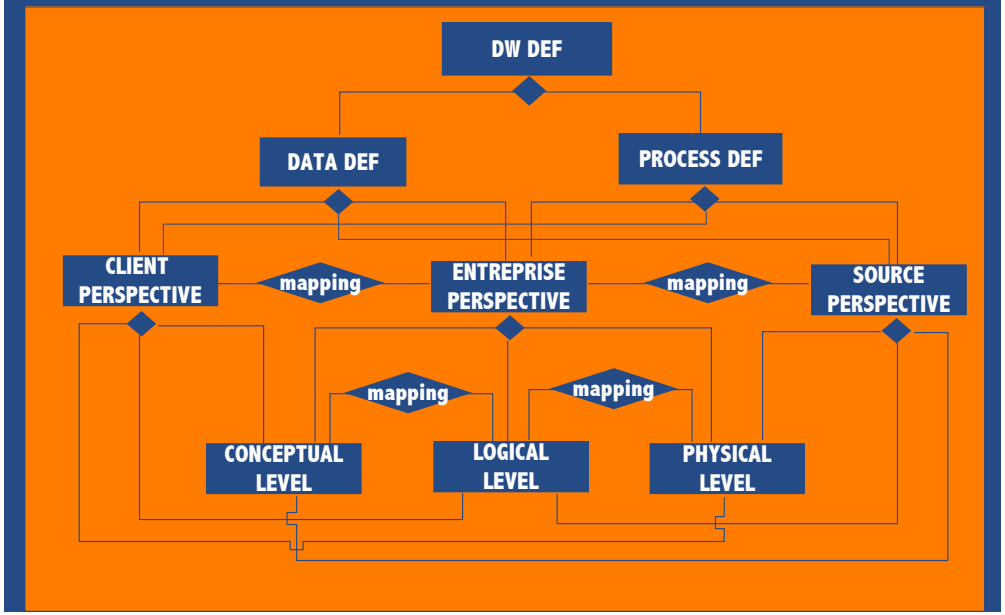
- Fixer la terminologie commune à plusieurs domaines
- Permettre l'interopérabilité opérationnelle entre plusieurs systèmes hétérogènes
- Gérer la complexité en passant à un niveau d'abstraction supérieur
- Fournir un schéma de référence aux méta données (appelé souvent référentiel)

Méta modélisation

Un métamodèle est un modèle de modèles.



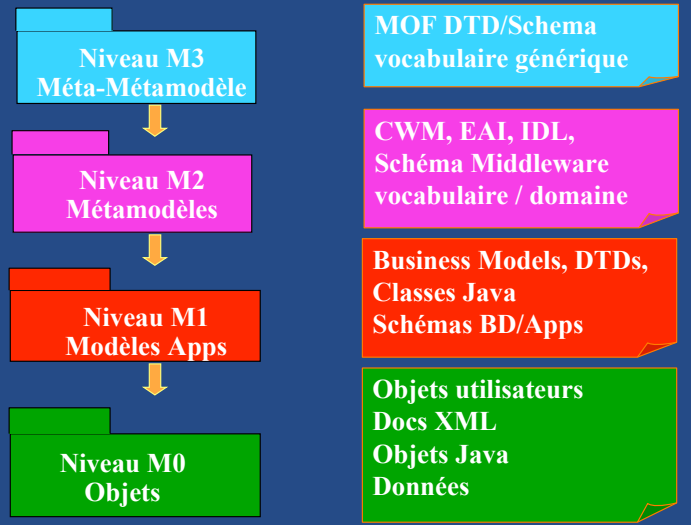
Principales méta classes



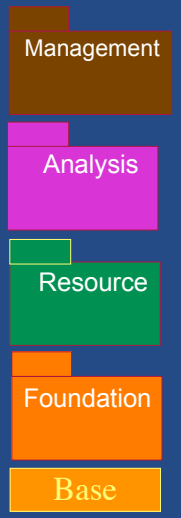
Standards de méta modèles

- IRDS (ANSI) : relationnel
- MOF (OMG) : objet
- CWM (OMG) : sous cas de MOF
- Dublin Core: objet, XML
- …
- Implémentations restreintes
- Support pour l'ingénierie dirigée par les modèles (outils CASE, MDA, MDE)

Le référentiel MOF

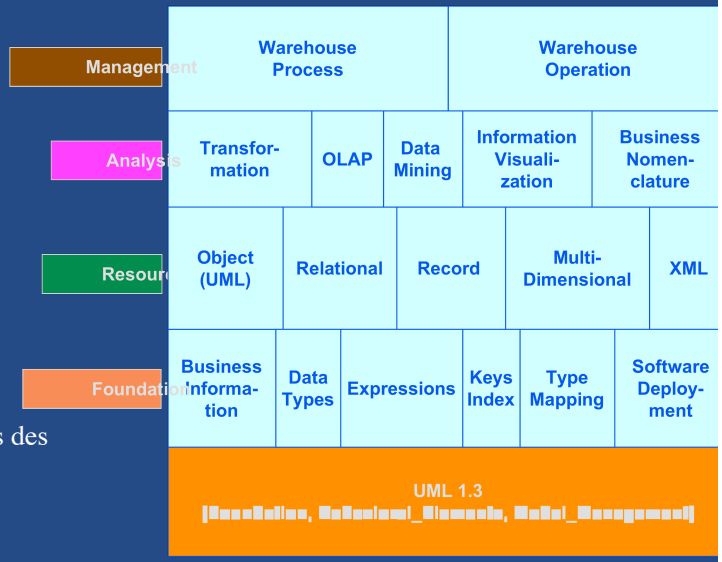


Common Warehouse model (CWM, OMG)



4 couches de description
toutes modélisées en UML
Contributeurs
IBM, Unisys, NCR, Hyperion, Oracle, Genesis, UBS, Dimension EDI...
Sauf Microsoft qui a sa propre offre
Adopté par l'OMG en juin 2000 (Oslo)
différentes implémentations (partielles)

Couche de Base

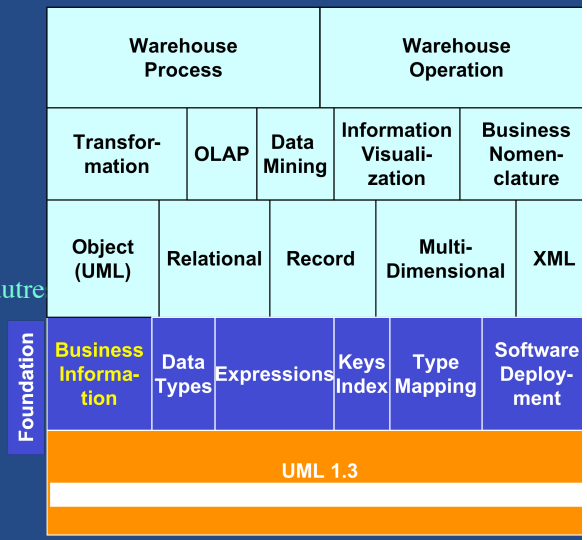


Environnement de modélisation
Notation UML
méta modèle UML étendu aux concepts des DW

Couche Fondation

Représente
- Connaissances métier
- Types de données
- Expressions de calcul
- Clés et les Indexes
- Déploiement de logiciels
- Mapping de types

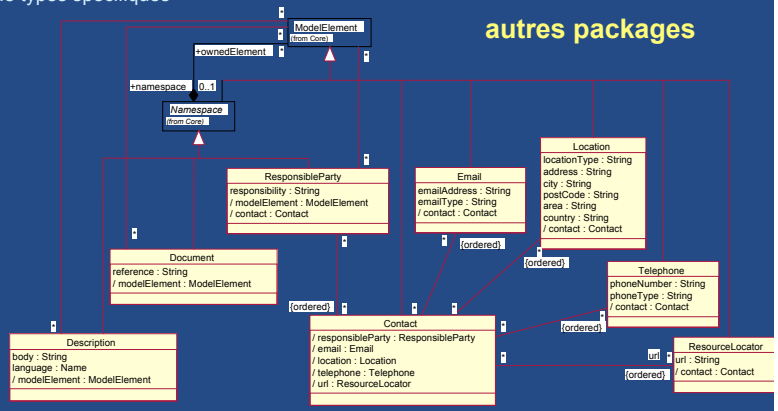
Méta modèles partagés par les autre packages



Fondation: exemple de méta modèle

- **Business Information**
  - Parties responsables et leurs coordonnées
  - Documentation et commentaires généraux
  - Hiérarchies de types spécifiques

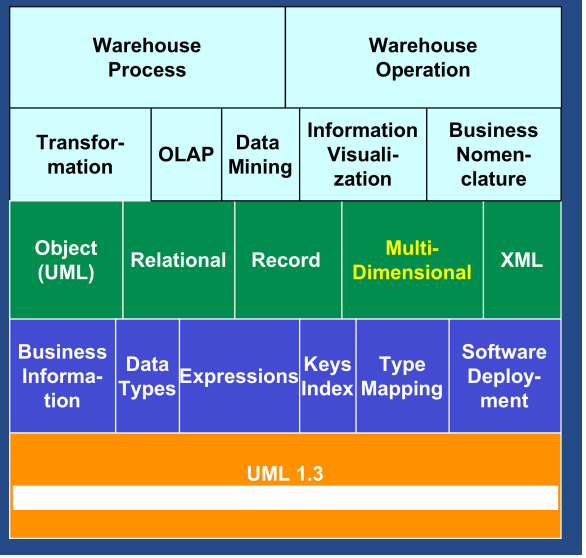
❑**Méta modèles partagés par les autres packages**



Couche Ressources de Données

Décrit les containers de données logiques et physiques
- Sources de données opérationnelles
- Données cibles de l'entrepôt
- Les modèles logiques

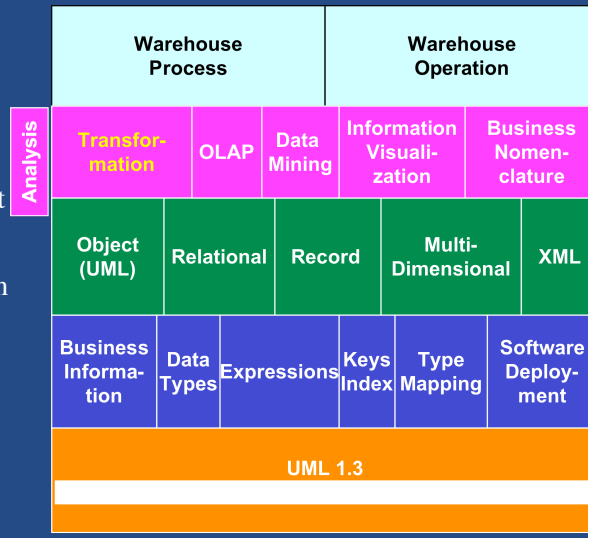
Méta modèles partagés par les autres packages



Couche Analyse de Données

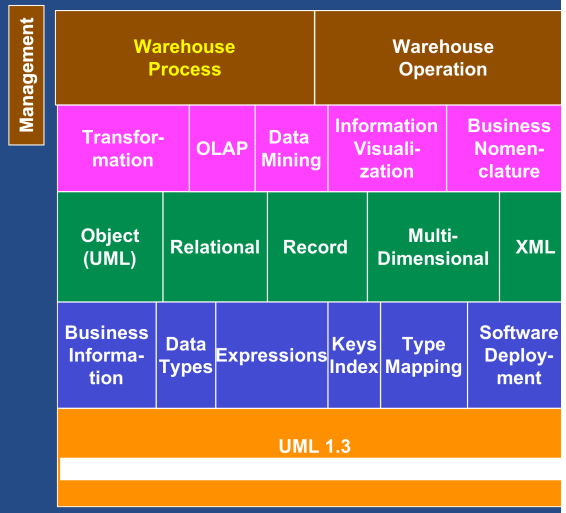
Décrit la production et l'analyse d'informations de décision
Décrit les structures analytiques déployées
Définit les transferts de données et leurs transformations

Détérmine les modes de restitution des données



Gestion de l'entrepôt de données

Orchestration des activités de l'entrepôt
- Process liés à l'entrepôt
  - Enchaînement des activités
  - Transformations des données
  - Evénements déclencheurs
- Opérations de surveillance
  - Métriques



Bilan sur les méta données

- **Les méta données décrivent les SI sur plusieurs dimensions**
  - –Outil de gestion de la complexité
  - –Outil de représentation de l'hétérogénéité
- **Les méta données constituent en elles-mêmes un sujet d'étude et de modélisation (méta modèle, méta modélisation)**
  - –Existence de standards
  - –Fondation des outils CASE et des ETL.

Autres systèmes d'intégration de données

- EAI
- Médiateurs
- Data Grid
- Portails Web
- Systèmes P2P

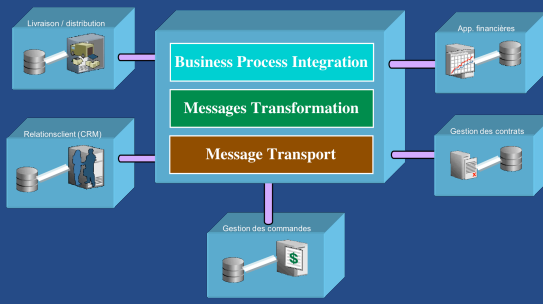
- Nombreux problèmes communs avec les entrepôts de données

EAI: Entreprise Application Integration

Middleware centré sur la communication entre applications

- **Avantages**
  - réduire le pb de latence des données par synchronisation des mises à jour en temps réel
  - utiles pour les applications nécessitant une très grande fraîcheur des données

- **Limites**
  - pas d'intégration de données (nettoyage, transformation, réconciliation, agrégation)
  - peu adaptés aux applications B2B
  - souvent limités à une connectivité entre les produits majeurs (Oracle, SAP, Siebel, People Soft)
    - perpétuent les problèmes de l'intégration en offrant des architectures dont les coûts d'évolution elevés ne permettent plus une intégration flexible des données

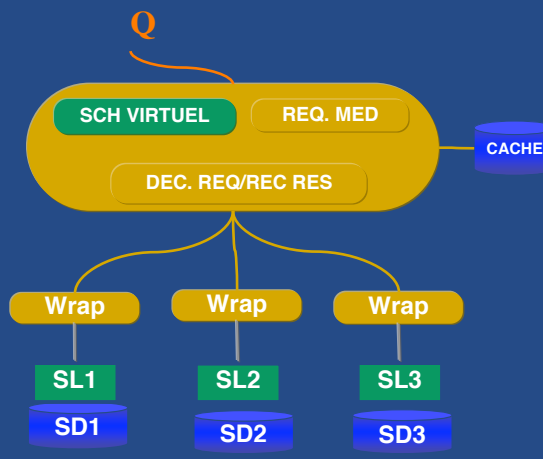


Médiateur

Middleware dédié à l'accès transparent à des sources de données hétérogènes

- **Avantages**
  - grande disponibilité des données
  - fraîcheur des données élevée
  - sources de données relationnelles ou XML

- **Limites**
  - Hypothèses fortes sur la disponibilité des sources de données
  - inadaptée pour des sources de données fortement hétérogènes
    - transformations de données complexes et coûteuses



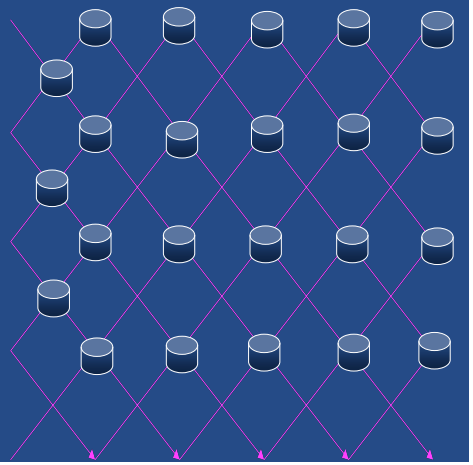
Data Grid

Infrastructure communautaire de services de gestion et de partage de données

- génome et biologie
- astronomie et recherche spatiale
- informations épidémiologiques

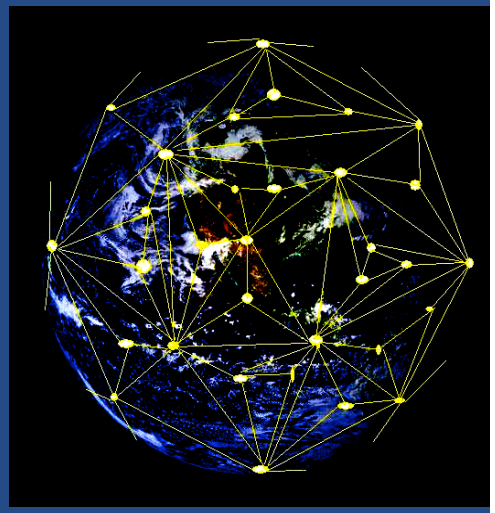
- **Avantages**
  - partage de ressources de calcul
  - partage de ressources d'informations
  - très haute disponibilité
  - composabilité des ressources

- **Limites**
  - droit de propriété
  - confidentialité



Portail Web

- vue uniforme d'informations *agrégées* à partir de sources de données hétérogènes
  - applications existantes
  - bases de données
  - systèmes documentaires
- portail vertical par opposition aux portails horizontaux
  - e.g. Yahoo, Excite, Alta Vista, Lycos
- pour des utilisateurs spécifiques
  - décideurs, employés, clients, fournisseurs
  - intranet ou extranet



- **Problème**
  - difficile d'accéder à toute l'information de l'entreprise