Data Mining:
1 Concepts et Techniques
2 Entrepôt de données et technologie OLAP (On-Line Analytical Processing)

Marie Beurton-Aimar

beurton@labri.fr
Cours préparé par Pascal Desbarats, Nicolas Parisey

December 1, 2020

Data Mining (fouille de données)

- Définition:
  Processus ou méthode qui extrait des connaissances "intéressantes "à partir d'une grande quantité de données.
- Référence:
  Data Mining: Concepts and Technics (Han et Kamber)
  http://www.cs.sfu.ca/~han

Au programme

4 séances cours/TD + un projet

- Introduction
- Entrepôts de données (data warehouses) et technologies OLAP
- Règles d'association
- Analyses de clustering

Plan du cours

- Motivation: pourquoi faire du data mining?
- Qu'est-ce que le data mining?
- Data mining: sur quelles données?
- Les fonctionnalités du data mining
- Tous les motifs sont-ils intéressants?
- Classification des systèmes de data mining
- Principaux défis du data mining

Motivation

- Masses de données
  - Outils automatisés de collecte de données
  - Maturité des SGBD
  - Entrepôts de données (data warehouses, information repositories)
- Données vs. connaissances
- Solution: entrepôts de données et data mining
  - Data warehousing et on-line analytical processing (OLAP)
  - Extraction de connaissances (règles, régularités, motifs, contraintes) à partir de grosses bases de données (BD).

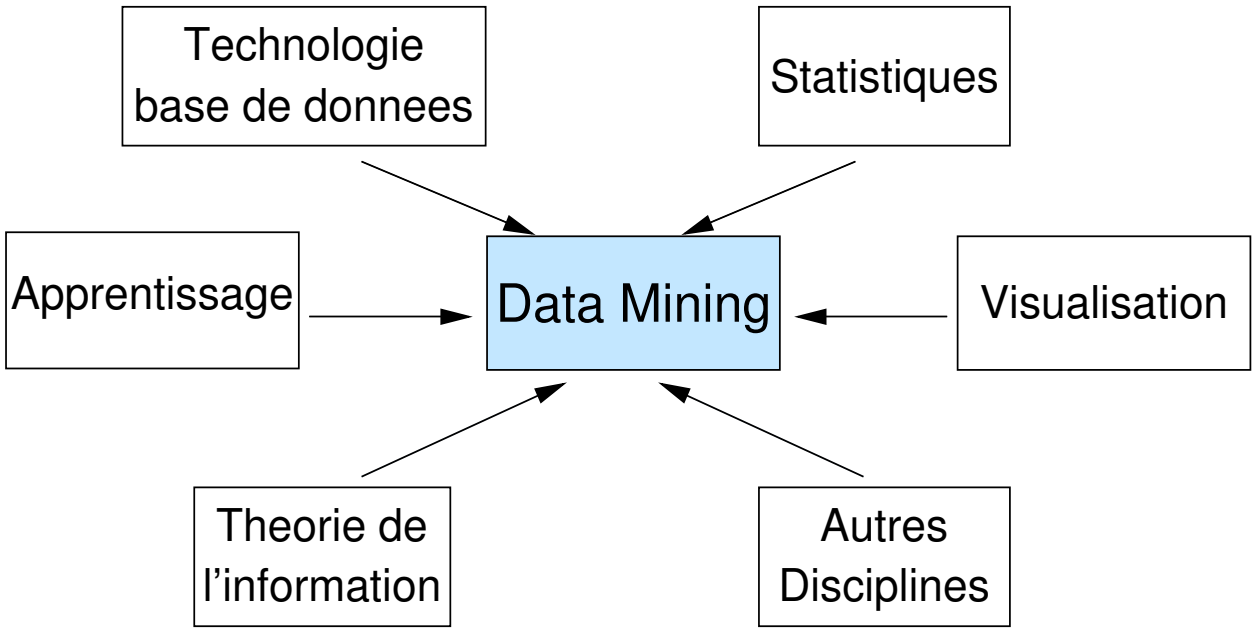
Evolution

- 1960: systèmes de gestion de fichiers, collecte de données, 1ières BD (modèle réseau)
- 1970: émergence du modèle relationnel et de son implémentation
- 1980: SGBD relationnels, modèles avancés (relationnel étendu, OO, déductif, etc) et orientés application (spatial)
- 1990-. . . : data mining et entrepôts de données, multimédia, web

Qu'est-ce que le data mining?

- Data mining (découverte de connaissances dans les bases de données):
  - Extraction d'informations ou de motifs intéressants (non triviaux, implicites, inconnus auparavant et potentiellement utiles) à partir de grandes BD (Extraction de Connaissances à partir de Données)
- Autres appellations:
  - Data mining: est-ce judicieux?
  - Knowledge Discovery (mining) in Databases, knowledge extraction, data/pattern analysis, information harvesting, business intelligence, fouille de données, ...
- Ce qui n'est pas du data mining
  - processus de requête déductive
  - systèmes experts

Data Mining: union de disciplines variées



Pour quoi faire?

- Analyse des BD et d'aide à la décision
  - Analyse du marché: cible marketing, gestion des relations client, analyse du panier de la ménagère, ventes croisées segmentation du marché
  - Analyse de risque: prévisions, fidélisation du client, mises en avant améliorées, contrôle qualité, analyses de compétitivité
  - Détection de fraudes
- Bio-informatique
- Autres applications
  - Text mining: news group, emails, PubMed, documents Web
  - Optimisation des requêtes

Analyse du marché

- Quelles sources de données?
  - Transactions bancaires (CB), coupons de réduction, service clients (plaintes), les études publiques de style de vie
- Cible marketing
  - Trouver des groupes "modèles "de clients qui partagent les mêmes caractéristiques: intérêts, revenus, habitudes de consommation, ...
- Déterminer les profils d'achat des clients au cours du temps
  - Ex: compte-joint après le mariage
- Analyses des ventes croisées
  - Associations/corrélations des ventes entre produits
  - Prédictions basées sur les associations d'information

Analyse du marché (suite)

- Profils clients
  - quels types de clients achètent quels produits? (clustering ou classifications)
- Identifier les besoins des clients
  - Identifier les meilleurs produits pour des clients différents
  - Utiliser la prédiction pour trouver quels facteurs vont attirer de nouveaux clients
- Fournir une synthèse de l'information
  - Rapports multidimensionnels variés
  - Rapports statistiques (tendance générale des données et variation)

Détection de fraudes

- Applications
  - Service de crédit, santé, télécommunications
- Approche
  - Utiliser les données d'historique pour construire des modèles pour les comportements frauduleux puis rechercher par data mining des instances similaires
- Exemples
  - Assurances: détecter les groupes de personnes qui déclarent des accidents/vols pour les indemnités
  - Blanchiment d'argent: détecter les transactions suspectes (US Treasury's Financial Crimes Enforcement Network)
  - Assurance maladie: détecter les patients professionnels et les médecins associés

Bio-informatique

- Sources de données
  - Séquences (ADN, acides aminés)
  - Structures tri-dimensionnelles
  - Puces à ADN
  - Interactions protéiques
  - PubMed
- Applications
  - Prédiction des séquences codantes
  - Prédiction de structure 3D
  - Analyse de données d'expression
  - Prédiction de fonction, interaction, localisation, ...
  - Découverte de motifs sur/sous représentés, répétitions
  - Phylogénie ...

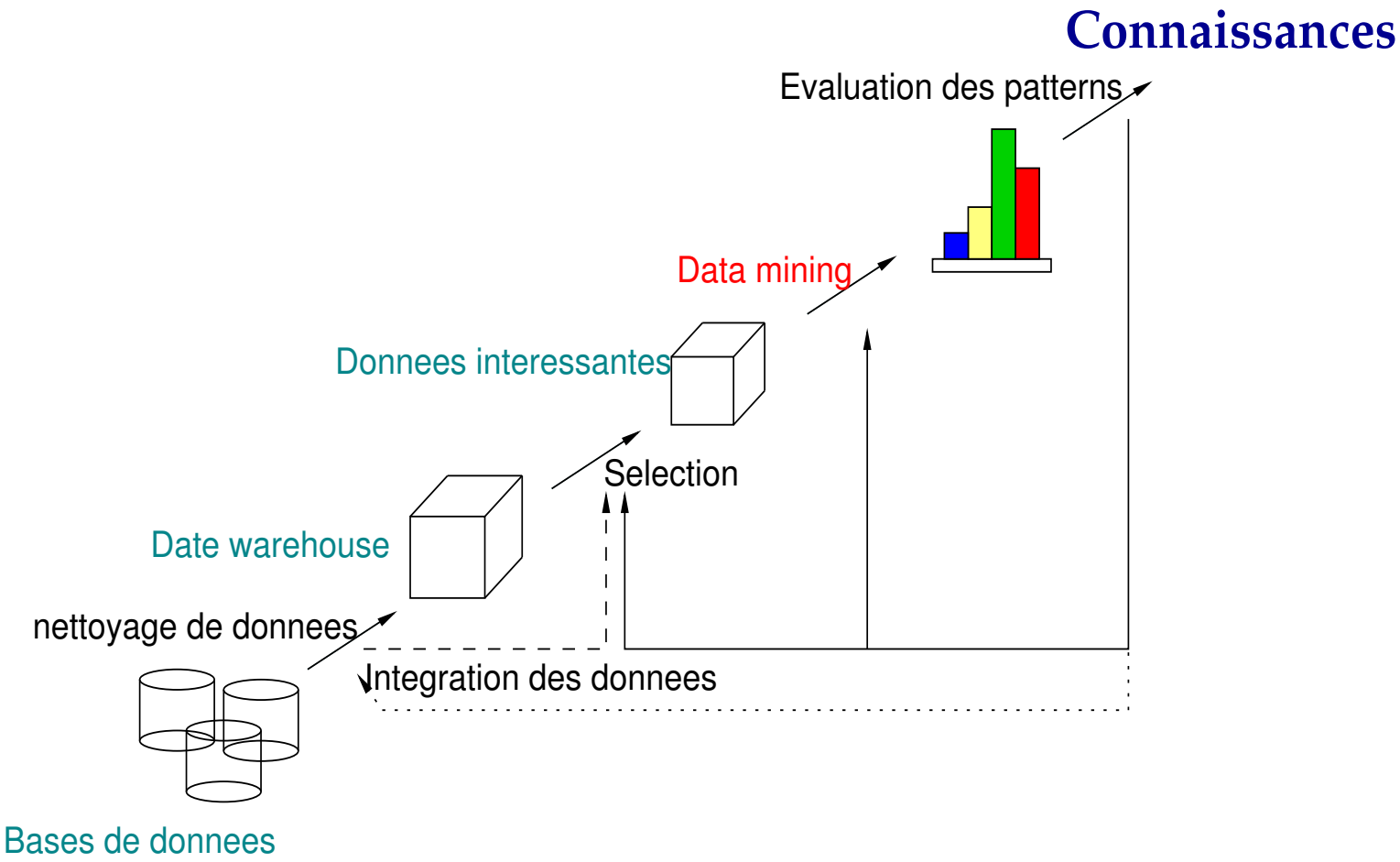
Autres applications

- Astrophysique:
  le laboratoire JPL a découvert 22 quasars en utilisant les techniques du data mining
- Organisation de sites web:
  algorithmes de data mining appliqués aux journaux d'accès aux pages commerciales afin d'identifier les préférences et les comportements des clients, et analyser les performances du web marketing. Ex: IBM a réorganisé son site web.

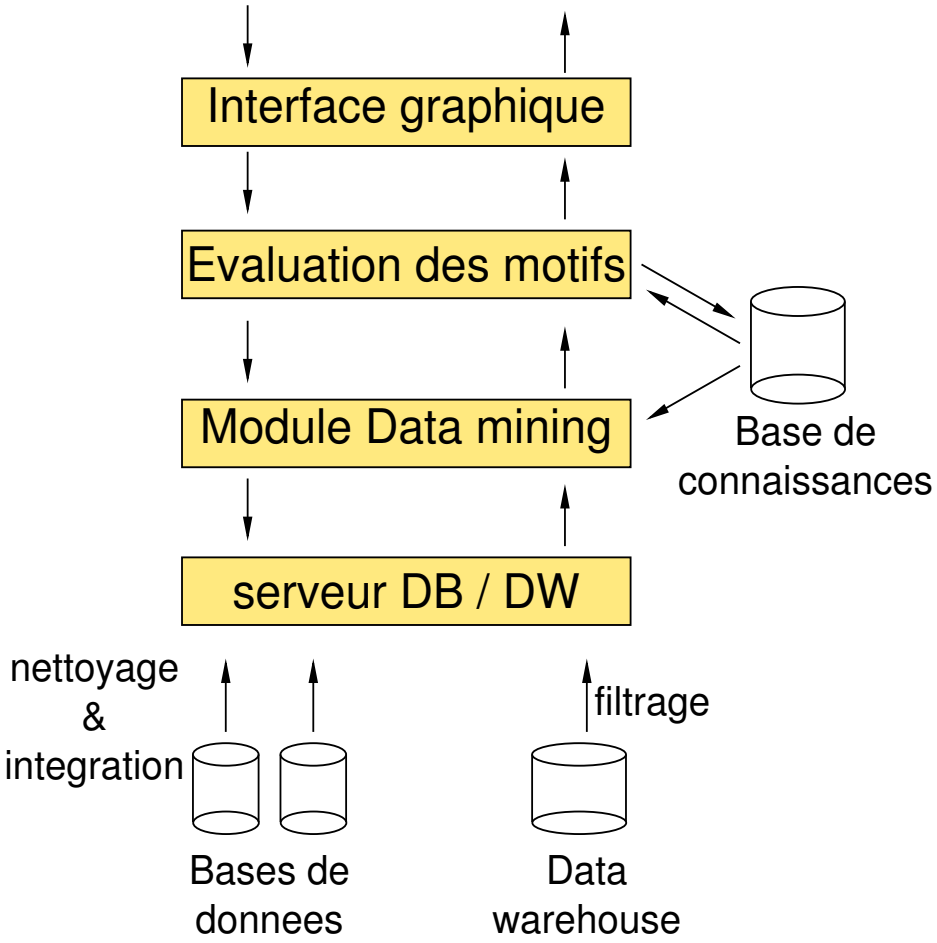
Etapes du processus d'ECD

- Compréhension du domaine d'application:
  connaissances nécessaires et buts de l'application
- Création du jeu de données (sélection des données)
- Nettoyage des données et prétraitement (jusqu'à 60% du travail!)
- Réduction des données et transformation:
  trouver les caractéristiques utiles, dimensionnalité/réduction des variables
- Choix des fonctionnalités du data mining:
  synthèse, classification, régression, association, clustering
- Choix de(s) l'algorithme(s) d'extraction
- Data mining: recherche de connaissances intéressantes
- Evaluation et représentation des connaissances:
  visualisation, transformation, élimination des connaissances redondantes, ...
- Utilisation de la connaissance extraite

Data mining: UN processus dans l'ECD



Architecture typique d'un système de Data mining



Data mining: sur quelles données?

- Fichiers plats
- BD relationnelles
- Entrepôts de données
- BD transactionnelles
- BD avancées:
  - Base de données orientées objets et relationnelles objets
  - Spatiales
  - Données temporelles
  - Textes et multimédia
  - WWW

Les fonctionnalités du data mining

- Description de concepts: caractérisation et discrimination
  - Généraliser, résumer, contraster les données caractéristiques. Ex: régions sèches vs. humides
- Association (corrélation et causalité)
  - Association multidimensionnelle vs. uni-dimensionnelle. Ex:
    - âge(X, "20..29 ")^revenu(X, "20..30K ") $\rightarrow$ achète(X, "PC ") [support = 2%, confiance = 60%]
    - contient(T, "PC ") $\rightarrow$ contient(T, "logiciel ") [1%, 75%]

Fonctionnalités (suite)

- Classification et Prédiction
  - Trouver des modèles (fonctions) qui décrivent et distinguent des classes ou concepts pour la prédiction future
  - Ex: classer les pays par climat
  - Présentation: arbre de décision, règles de classification, réseaux de neurones
  - Prédiction: prédire des valeurs inconnues ou manquantes
  - Démarche:
    - Prendre un échantillon (jeu d'essai) dans lequel chaque objet est associé à une classe
    - Analyser chaque classe (son contenu) pour pouvoir ensuite affecter chaque nouveau objet à une classe particulière

Fonctionnalités (suite)

- Analyse de groupes (clustering)
  - Appelée aussi classification non supervisée
  - pas de classes prédéfinies: grouper les données pour former des classes nouvelles. Ex: familles de protéines basées sur la similarité de séquence
  - Principe: maximiser la similarité intra-classe et minimiser la similarité entre groupes distincts

Fonctionnalités (suite)

- Analyse d'exception
  - Un objet qui se distingue du comportement général des données
  - Une exception peut être considérée comme du bruit mais aussi comme indice de fraude
- Analyse de tendances et d'évolution
  - Tendance et déviation: analyse de régression
  - Découverte de motifs séquentiels, analyse de périodicité
  - Analyses basées sur la similarité
- Autres analyses basées sur des motifs ou sur des statistiques

Les motifs découverts sont-ils tous intéressants?

- Pb: un système de data mining peut générer des milliers de patterns.
  - Approches suggérées: système centré sur l'utilisateur, basé sur des requêtes
- Mesures d'intérêt: un motif est intéressant s'il est:
  - Facile à comprendre par un humain
  - Valide sur de nouvelles données ou données test avec un certain degrès de certitude
  - Nouveau
  - Ou peut servir à valider (ou invalider) une hypothèse utilisateur
- Mesures objective vs. subjective:
  - Objective: basée sur des statistiques et sur les structures des motifs. Ex: support, confiance
  - Subjective: basée sur le point de vue de l'utilisateur. Ex: inattendu, nouveau

Peut-on trouver tous et que les motifs intéressants?

- Trouver tous les motifs intéressants: Complétude
  - Est-ce qu'un système peut trouver tous les patterns intéressants?
  - Association cs. classification vs. clustering
- Trouver que les motifs intéressants: Optimisation
  - Est-ce qu'un système peut trouver seulement les patterns intéressants?
  - Approches:
    - Générer tous les patterns puis les filter
    - Ne générer que les patterns intéressants

Data mining: critères de classification

- Fonctionnalité générale:
  - Data mining descriptif
  - Data mining prédictif
- Vues différentes, classifications différentes
  - Types de BD à fouiller (relationnelles ou OO, texte ou multimédia)
  - Types de connaissances à découvrir (clustering, association)
  - Types de techniques utilisées (automatiques, guidées par l'utilisateur)
  - Types d'applications spécifiques

Data mining: critères de classification (suite)

- Types de BD à fouiller
  - Relationnelle, transactionnelle, orientée objet, relationnelle objet, spatialle, temporelle, texte, multimédia, WWW, ...
- Types de connaissances à découvrir
  - Caractérisation, discrimination, association, classification, clustering, tendance, déviation, analyse d'exception, ...
  - Multiples fonctions intégrées sur différents niveaux
- Techniques utilisées
  - Orienté base de données, entrepôt de données (OLAP), apprentissage automatique (machine learning), statistiques, visualisations, réseaux de neurones, ...
- Applications spécifiques
  - Télécommunication, banque, analyse de fraude, bio-informatique, bourse, Web mining, ...

Principaux défis en data mining

- Méthodologie et interactions utilisateur
  - Différents types de connaissances à extraire des bases de données
  - Prise en compte des connaissances des experts (background knowledge)
  - Langages de requêtes pour le data mining
  - Expression et visualisation des résultats
  - Prise en compte du bruit ou des données manquantes / incomplètes
  - Evaluation des patterns: notion d'intérêt
- Performance et mise à l'échelle
  - Efficacité et mise à l'échelle des algorithmes de data mining
  - Parallélisation, distributivité et possibilitées incrémentales des méthodes de fouilles

Principaux défis en data mining (suite)

- Liés à la diversité des données
  - Données relationnelles et types complexes
  - BD hétérogènes et système global d'information (www)
- Liés aux applications et aux nouvelles connaissances
  - Applications
    - Création d'outils domaine-spécifique
    - Réponse aux requêtes selon domaine
    - Contrôle de processus et aide à la décision
  - Intégration des connaissances découvertes avec celles existantes: problème de fusion des connaissances
  - Protection des données: sécurité, intégrité, et données privées (informatique et libertés)

Bilan

- Data mining: découverte de motifs intéressants à partir de grandes quantités de données
- Evolution naturelle de la technologie des BD, avec de larges applications
- Le processus d'ECD: nettoyage de données, intégration, sélection, transformation, fouille de données, évaluation de motif, et (re)présentation de la connaissance
- L'extraction de connaissances peut être réalisée dans divers types d'entrepôts de données
- Les fonctionnalités du data mining: caractérisation, discrimination, association, classification, clustering, analyse d'exception et de tendance, ...
- Classifications des systèmes de data mining
- Applications importantes du data mining dans la vie courante (finance, consommation, santé,...)

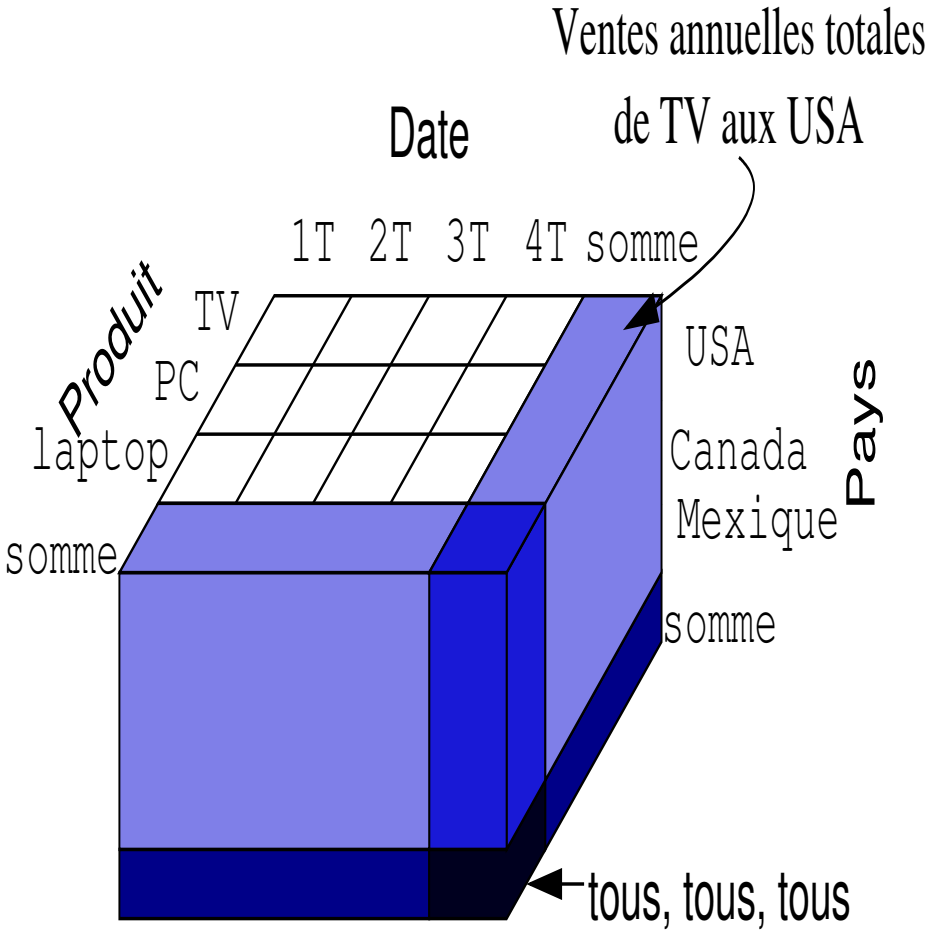
Entrepôt de données (data warehouse) et technologie OLAP (On-Line Analytical Processing) pour le Data Mining

- Un modèle de données multidimensionnel
- Qu'est-ce qu'un entrepôt de données?
- Architecture d'un entrepôt de données
- Implémentation d'un entrepôt de données
- Autres développements basés sur les cubes de données
- De l'entrepôt de données à la fouille de données

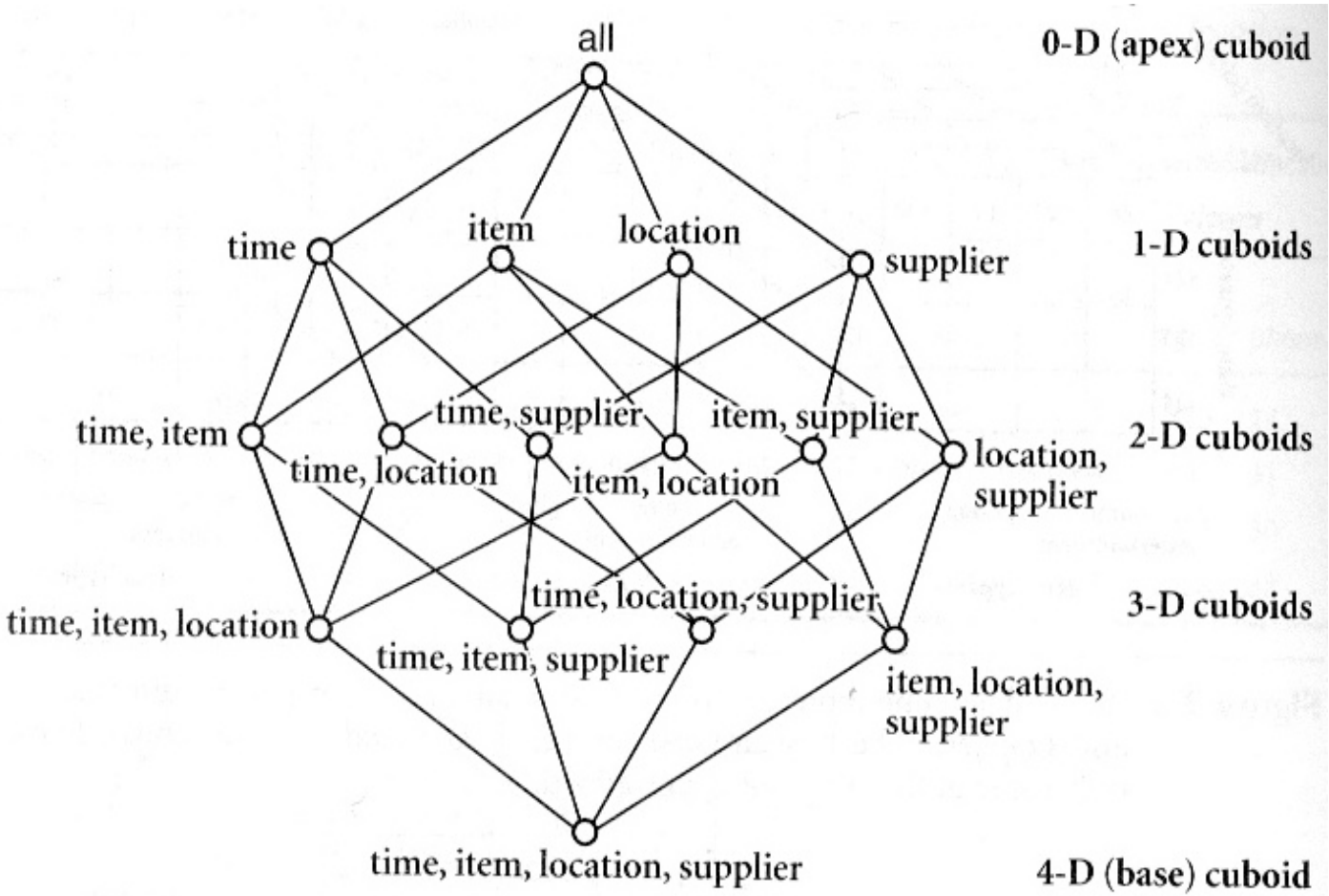
Des tableaux aux cubes de données

- Un entrepôt de données est basé sur un modèle multidimensionnel des données qui représente les données sous la forme d'un cuboïde à n dimensions
- Un cube de données (ex: ventes) permet de voir les données selon plusieurs dimensions:
  - Les tables de dimension: ex: item (nom_produit, marque, type) ou la date (jour, semaine, mois, trimestre, année)
  - La table de faits contient les mesures (ex: unités vendues, prix) et les clefs externes faisant référence à chaque table de dimension
- L'hypercube (cube de cuboïdes) d'un datawarehouse est appellé cube de données (ou OLAP cube). L'implémentation de cet hypercube peut se faire au moyen d'un treillis des cuboïdes d'ordres inférieurs.

Un cube de données



Un cube de données: un treillis de cuboïdes



Entrepôt de données (data warehouse) et technologie OLAP (On-Line Analytical Processing) pour le Data Mining

- Un modèle de données multidimensionnel
- Qu'est-ce qu'un entrepôt de données?
- Architecture d'un entrepôt de données
- Implémentation d'un entrepôt de données
- Autres développements basés sur les cubes de onnées
- De l'entrepôt de données à la fouille de données

Qu'est-ce qu'un entrepôt de données?

- Différentes définition, pas de définition rigoureuse
  - Une BD pour l'aide à la décision, maintenue séparément des bases de données opérationnelles
  - Support pour le traitement de l'information en fournissant une plateforme pour l'analyse de donnés consolidées et historiques
- "Un data warehouse est une collection de données concernant un sujet particulier, variant dans le temps, non volatile et où les données sont intégrées "(W.H. Immon)
- Data warehousing: le processus de construction et d'utilisation d'entrepôt de données

Data warehouse orienté sur un sujet

- Organisé autour d'un sujet bien précis: ex: client, produit, ventes,
- Focalisé sur la modélisation et l'analyse de données pour aider les décideurs, non pas pour des activités quotidiennes ou des transactions,
- Fournissant un aperçu simple et concis concernant un sujet particulier en excluant les données qui ne servent pas à la prise de décision.

Data warehouse intégré

- Construit en intégrant des sources de données multiples et hétérogènes:

  bases de données relationnelles, fichiers plats,
  enregistrement de transactions
- Données nettoyées et intégrées:
  - cela assure la cohérence des conventions de nommage, structures de codage, mesures d'attributs, etc. entre différentes sources de données (ex: prix des hôtels, monnaie (euros, dollars), taxes (incluses ou non), petit-déjeuner compris ou non, . . .)
  - les données sont converties lors de leur importation dans l'entrepôt de données

Data warehouse variant dans le temps

- La période de temps prise en compte par les entrepôts de données est plus importante que pour les systèmes opérationnels
  - BD opérationnelles: valeurs actuelles des données
  - Entrepôt de données: fournit des informations d'un point de vue historique (ex: 5-10 dernières années)
- Chaque élément ou structure contient, implicitement ou explicitement, la notion de date

Data warehouse non volatile

- Un entrepôt physiquement séparé des données transformées depuis l'environnement opérationnel
- Pas de mise à jour des données dans l'entrepôt de données (au sens opérationnel)
  - pas besoin des mécanismes de transactions, reprise après accident, et accès concurrents
  - nécessité de seulement 2 opérations dans l'accès aux données:
    - chargement initial
    - consultation

Data warehouses vs. SGBD hétérogènes

- Intégration "traditionnelle "de données hétérogènes:
  - Utilisation de médiateurs (wrappers) au-dessus des BD hétérogènes
  - Approche orientée requête ("query driven "):
    - Lorsqu'une requête est soumise par un serveur client, un dictionnaire et des méta données sont utilisés pour traduire la requête pour chaque base de données sous-jacente impliquée. Les résultats sont intégrés en une réponse globale.
    - Cela implique des mécanismes complexes de filtrages de l'information et un surcoût en ressources
- Entrepôt de données: haute performance:
  - Les informations provenant de sources de données hétérogènes sont intégrées à l'avance et stockées dans l'entrepôt de données pour des requêtes et des analyses directes.

Data warehouse vs. BD opérationelle

- OLTP (On-Line Transaction Processing)(BD)
  - Tâches principales des BD relationnelles traditionnelles
  - Exécution en temps réel des transactions, pour l'enregistrement des opérations quotidiennes:ex: inventaire, commande, paye, comptabilité
- OLAP (On-Line Analytical Processing)(DWH)
  - Tâches principales du système d'entrepôt de données
  - Traitement efficace des requêtes d'analyse pour la prise de décision
- Caractéristiques distinctes (OLTP vs. OLAP)
  - Orientation vers l'utilisateur et le système: client vs. marché
  - Contenu des données: courantes, détaillées vs. historiques, consolidées
  - Conception de la BD: modèle entité-relation + application vs. modèle en étoile + sujet
  - Vue: courante, locale vs. évolutive, intégrée
  - Mode d'accès: mise à jour vs. lecture seule mais requêtes complexes

OLTP vs. OLAP

	OLTP	**OLAP**
utilisateurs	tout le monde	décideurs
fonction	opérations journalières	aide à la décision
DB design	orienté applications	orienté sujet
données	courantes, mise à jour, relationnel plat	historiques, résumées, multidimensionnelles, intégrées
usage	répétitif	*ad hoc*
accès	read/write, index/hash sur les tables	beaucoup de scans
unité de travail	transactions courtes	requêtes complexes
nbr enregistrements	dizaines	millions
nbr utilisateurs	centaines	dizaines
taille BD	100MB-GB	100GB-TB
métrique	exécution des transactions	temps de réponse aux requêtes

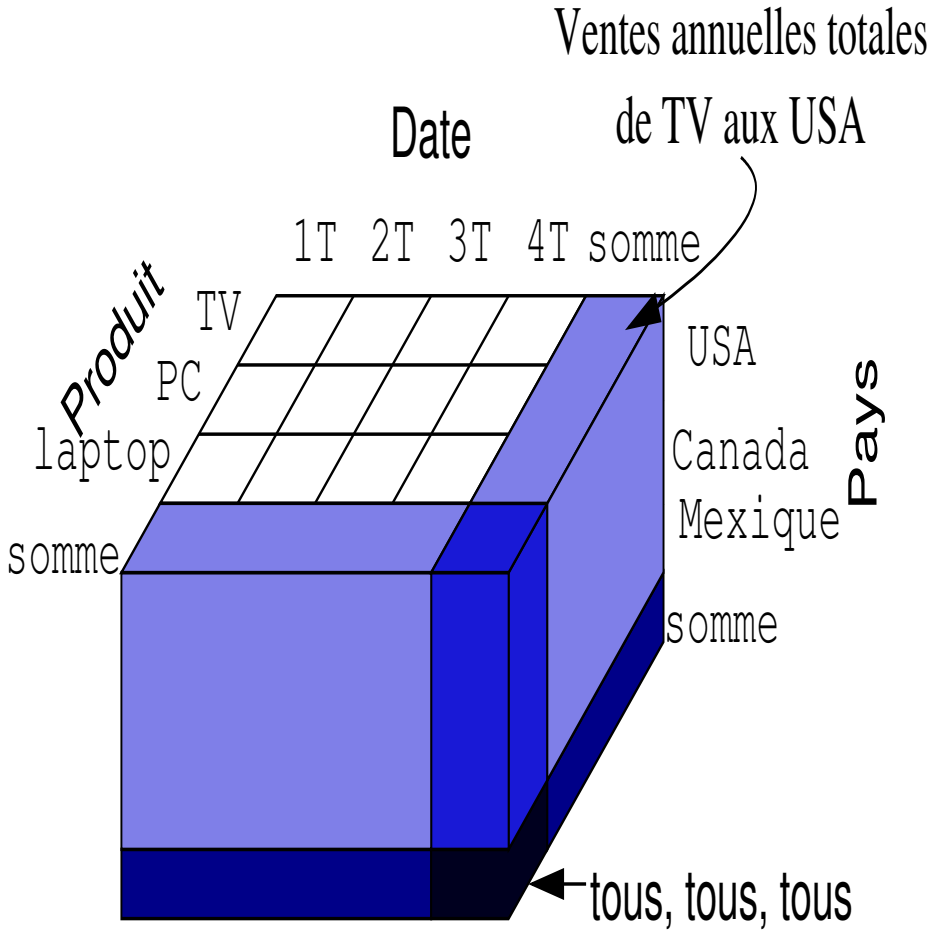
Pourquoi utiliser un entrepôt de données séparé?

- Performances pour les 2 types de systèmes:
  - SGBD optimisé pour l'OLTP: méthodes d'accès, indexation, accès concurrents, reprise après accidents
  - Entrepôt de données optimisé pour OLAP: requête complexe, vue multidimensionnelle, données consolidées
- Des fonctions différentes et des données différentes:
  - Données manquantes: données historiques non maintenues dans les systèmes opérationnels
  - Données consolidées: agrégation, condensation de données hétérogènes
  - Qualité des données: différentes sources utilisent des représentations et des formats différents qui doivent être réconciliés

Entrepôt de données (data warehouse) et technologie OLAP (On-Line Analytical Processing) pour le Data Mining

- Un modèle de données multidimensionnel
- Qu'est-ce qu'un entrepôt de données?
- Architecture d'un entrepôt de données
- Implémentation d'un entrepôt de données
- Autres développements basés sur les cubes de données
- De l'entrepôt de données à la fouille de données

Un cube de données

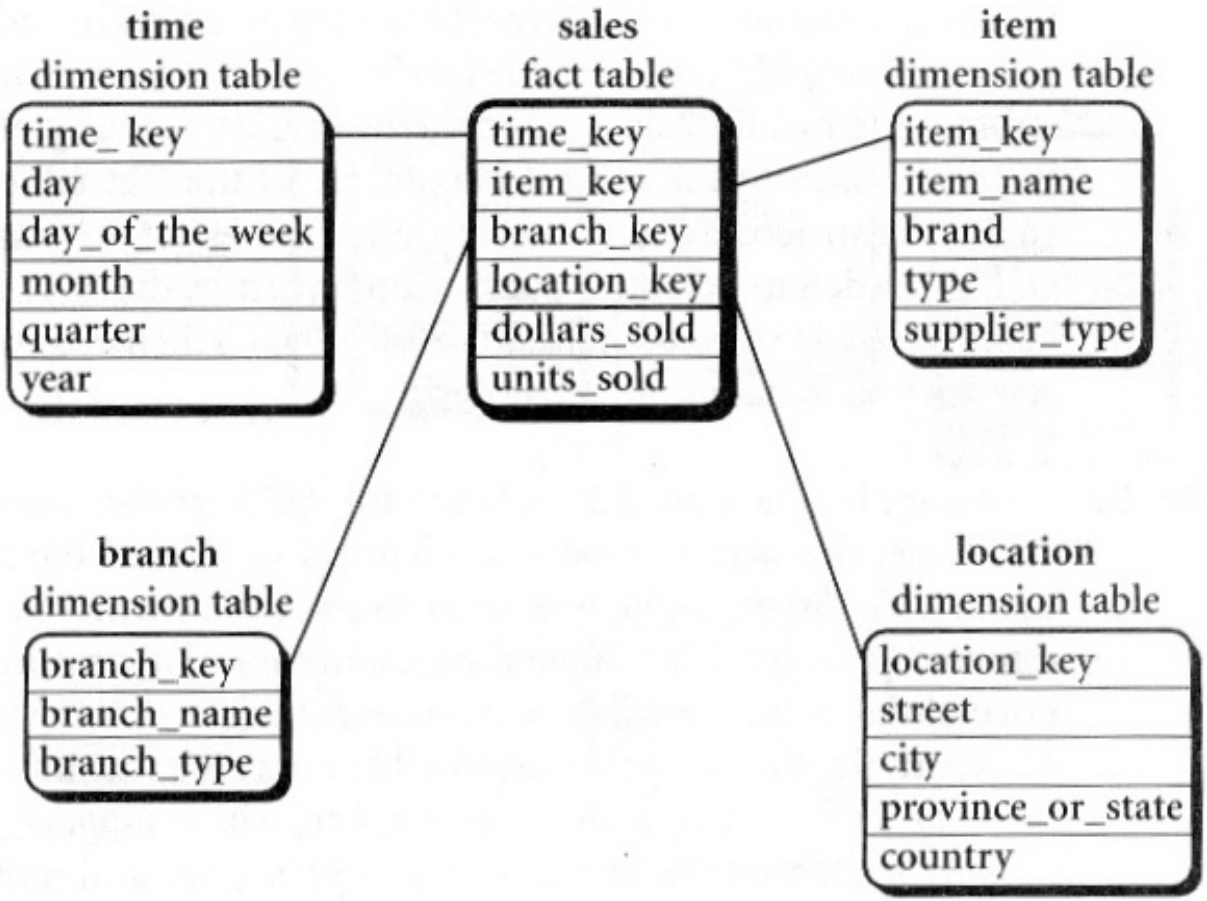


Modélisation conceptuelle des data warehouses

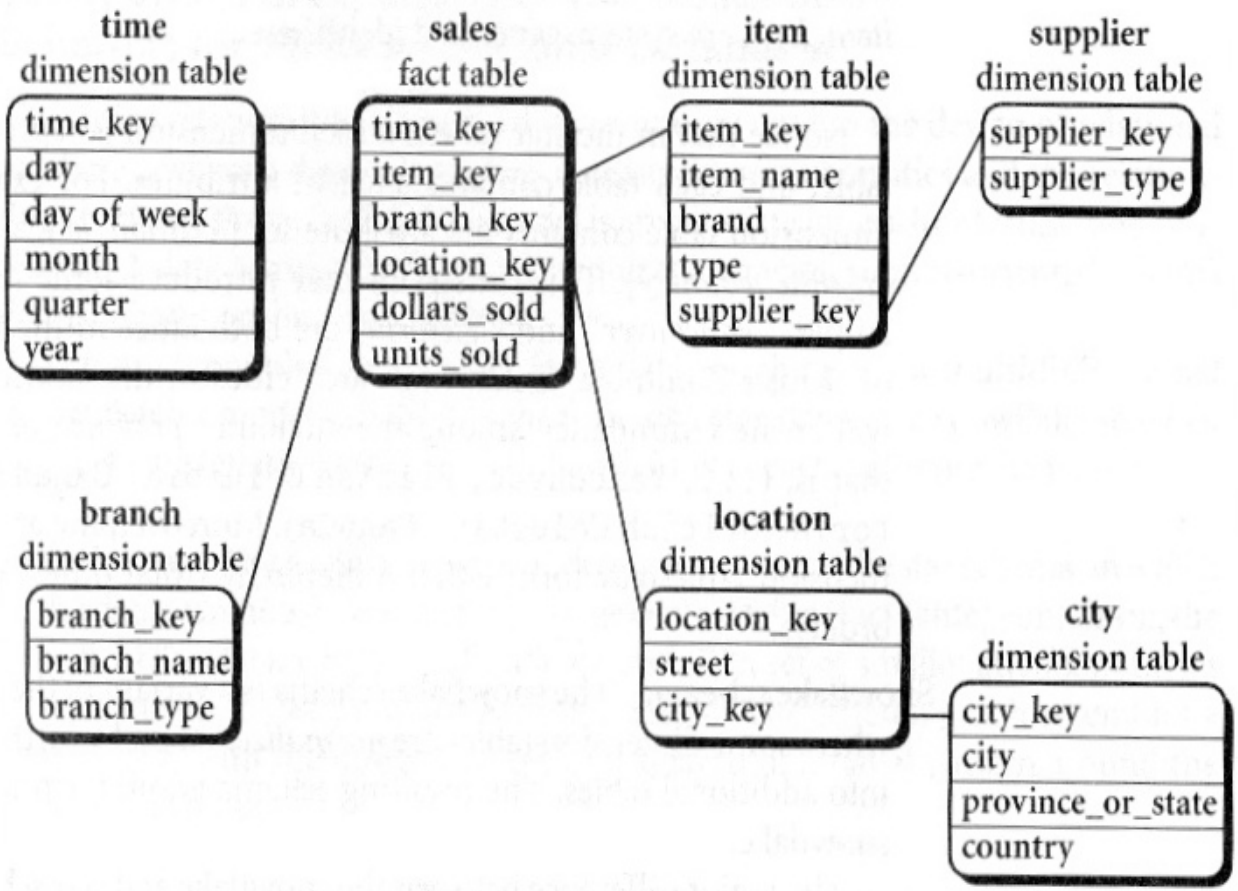
Dimensions et mesures:

- Schéma en étoile: au milieu, une table de faits connectée à un ensemble de tables de dimensions
- Schéma en flocon de neige (snowflake): un raffinement du précédent où certaines tables de dimensions sont normalisées
- Schéma en constellation de faits: plusieurs tables de faits partagent quelques tables de dimension (constellation d'étoiles, schéma de galaxie)

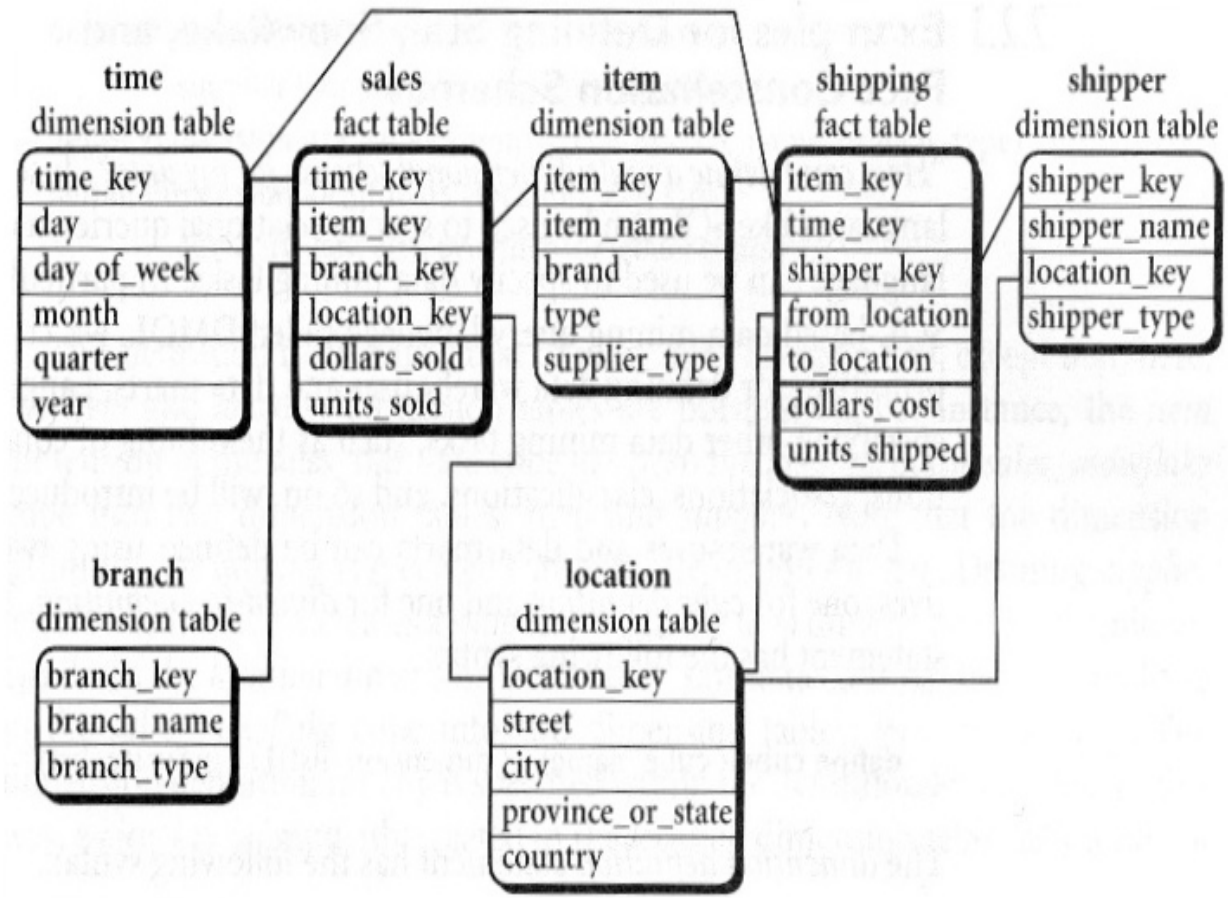
Exemple de schéma en étoile



Exemple de schéma en flocon de neige (Snowflake)



Exemple de schéma en constellation de faits



Data Mining Query Language: un langage pour le data mining

- Définition d'un cube (table de faits)

  `define cube <nom_cube> [<liste_dimensions>] : <liste_mesures>`

- Définition d'une dimension (table de dimensions)

  `define dimension <nom_dimension> as (<liste_attributs_ou_sous-dimensions>)`

- Cas particulier (tables de dimensions partagées)
  - la première fois, comme la définition d'un cube
  - `define dimension <nom_dimension> as <1er_nom_dimension> in cube <1er_nom_cube>`

Définition d'un schéma en étoile avec DMQL

```
define cube ventes_star [temps, item, marque, lieu ]:
     montant_vente = sum(somme),
     moyenne_vente = avg(somme),
     unités_vendues = count()

define dimension temps as
     (Id_temps, jour, jour_semaine, mois, trimestre, année)
define dimension item as
     (Id_item, nom_item, marque, type, type_fournisseur)
define dimension marque as
     (Id_marque, nom_marque,type_marque)
define dimension lieu as
     (ID_lieu, rue, ville, département, pays)
```

Définition d'un schéma en snowflake avec DMQL

```
define cube ventes_snowflake [temps,item,marque,lieu ]:
     montant_vente = sum(somme),
     moyenne_vente = avg(somme),
     unités_vendues = count()
define dimension temps as
     (Id_temps, jour, jour_semaine, mois, trimestre, année)
define dimension item as
     (Id_item, nom_item, marque, type,
          fournisseur(Id_fournissuer, type_fournisseur))
define dimension marque as
     (Id_marque, nom_marque,type_marque)
define dimension lieu as
     (ID_lieu, rue, ville(ID_ville, département, pays))
```

Définition d'un schéma en constellation avec DMQL

```
define cube ventes[temps, item, marque, lieu ]:
      montant_vente = sum(somme),
      moyenne_vente = avg(somme),
      unités_vendues = count()
define dimension temps as
      (Id_temps, jour, jour_semaine, mois, trimestre, année)
define dimension item as
      (Id_item, nom_item, marque, type, type_fournisseur)
define dimension marque as
      (Id_marque, nom_marque, type_marque)
define dimension lieu as
      (ID_lieu, rue, ville, département, pays)
```

Définition d'un schéma en constellation avec DMQL

```
define cube transport [temps, item,transporteur,départ,arrivée ]:
      coût = sum(frais), unités_transportées = cpount()
define dimension temps as
      temps in cube ventes
define dimension item as
      item in cube ventes
define dimension transporteur as
      (Id_transporteur, nom_transporteur, lieu as
      lieu in cube ventes, type_transporteur)
define dimension départ as
      lieu in cube ventes
define dimension arrivée as
      lieu in cube ventes
```

Les mesures: 3 catégories

- Distributive: si le résultat obtenu par une fonction à n valeurs calculées est le même que le résultat de la fonction sur toutes les valeurs

  eg. count(), sum(), min(), max()

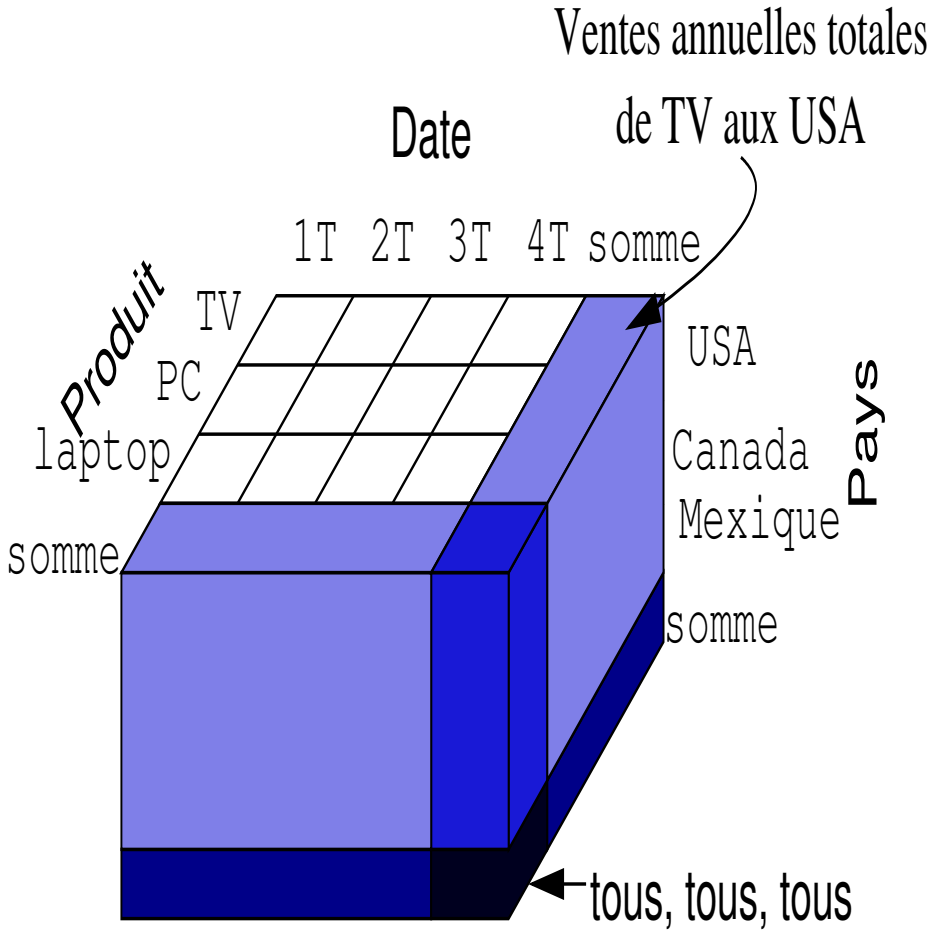
- Algébrique: si elle peut être calculée par une fonction à M arguments, chacun obtenu par une fonction distributive

  eg. avg(), standard_deviation()

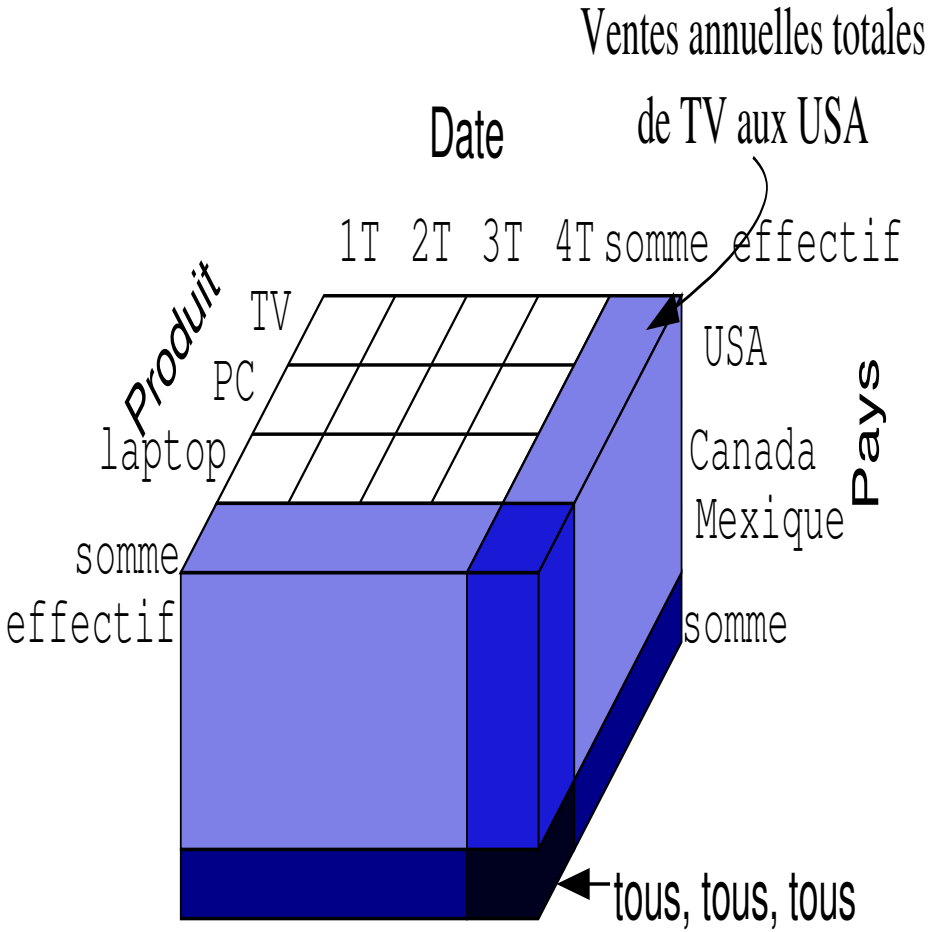
- Holistique: s'il n'y a pas de constante liée à la taille de stockage nécessaire pour décrire un sous-ensemble

  eg. median(), mode(), rank()

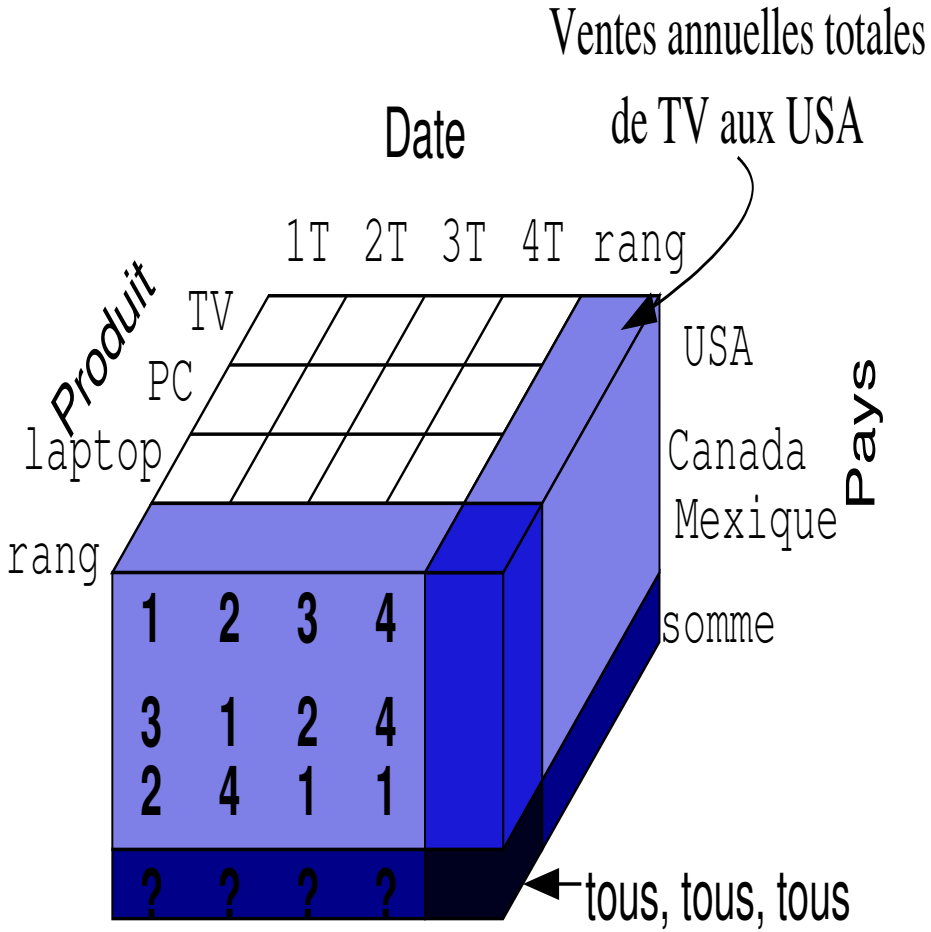
Un cube de données



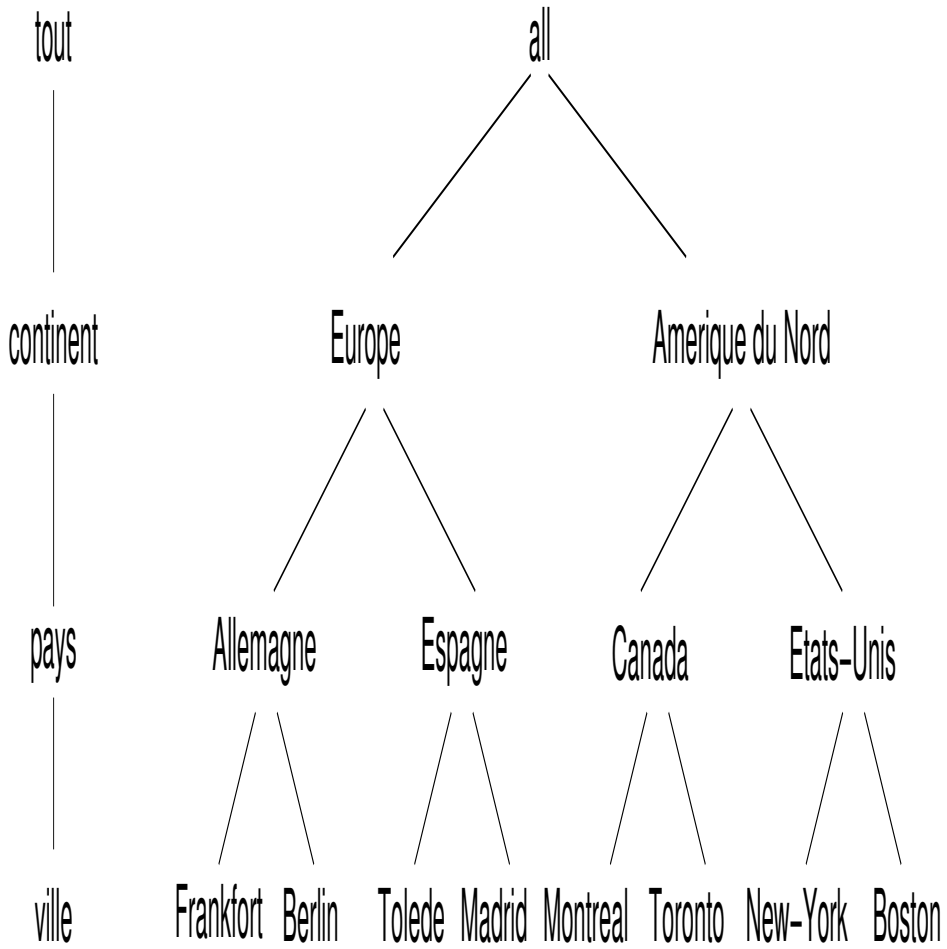
Un cube de données



Un cube de données

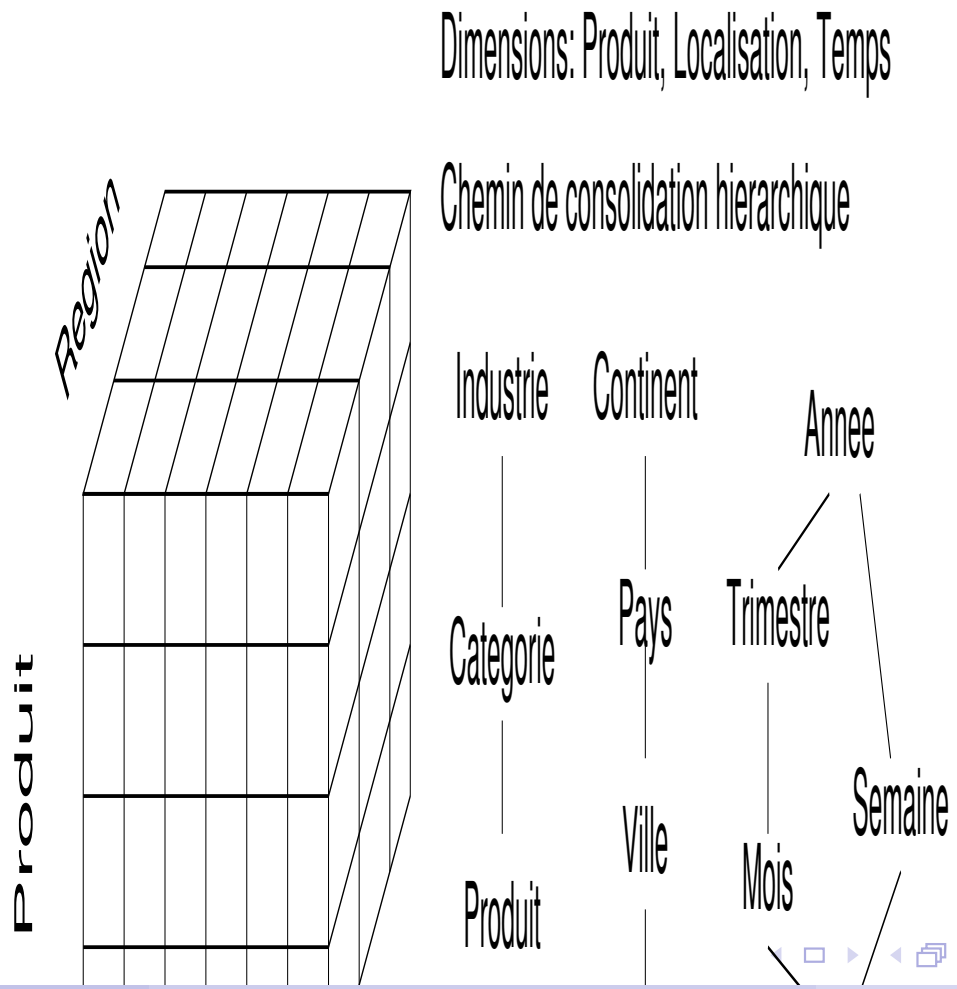


La hiérarchie de concept: la dimension (lieu)



Les données multidimensionnelles

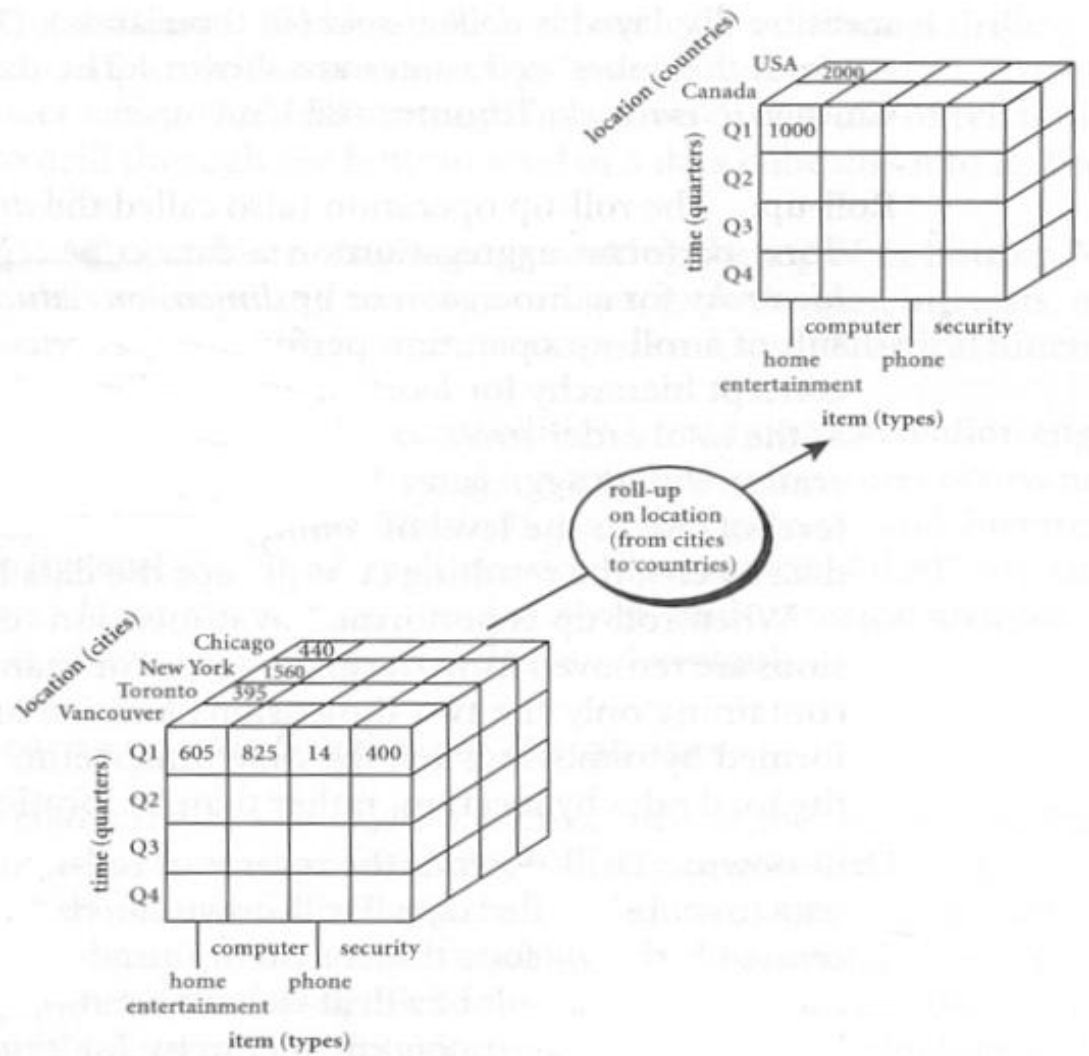
Montant des ventes comme une fonction des paramètres produit, mois, région



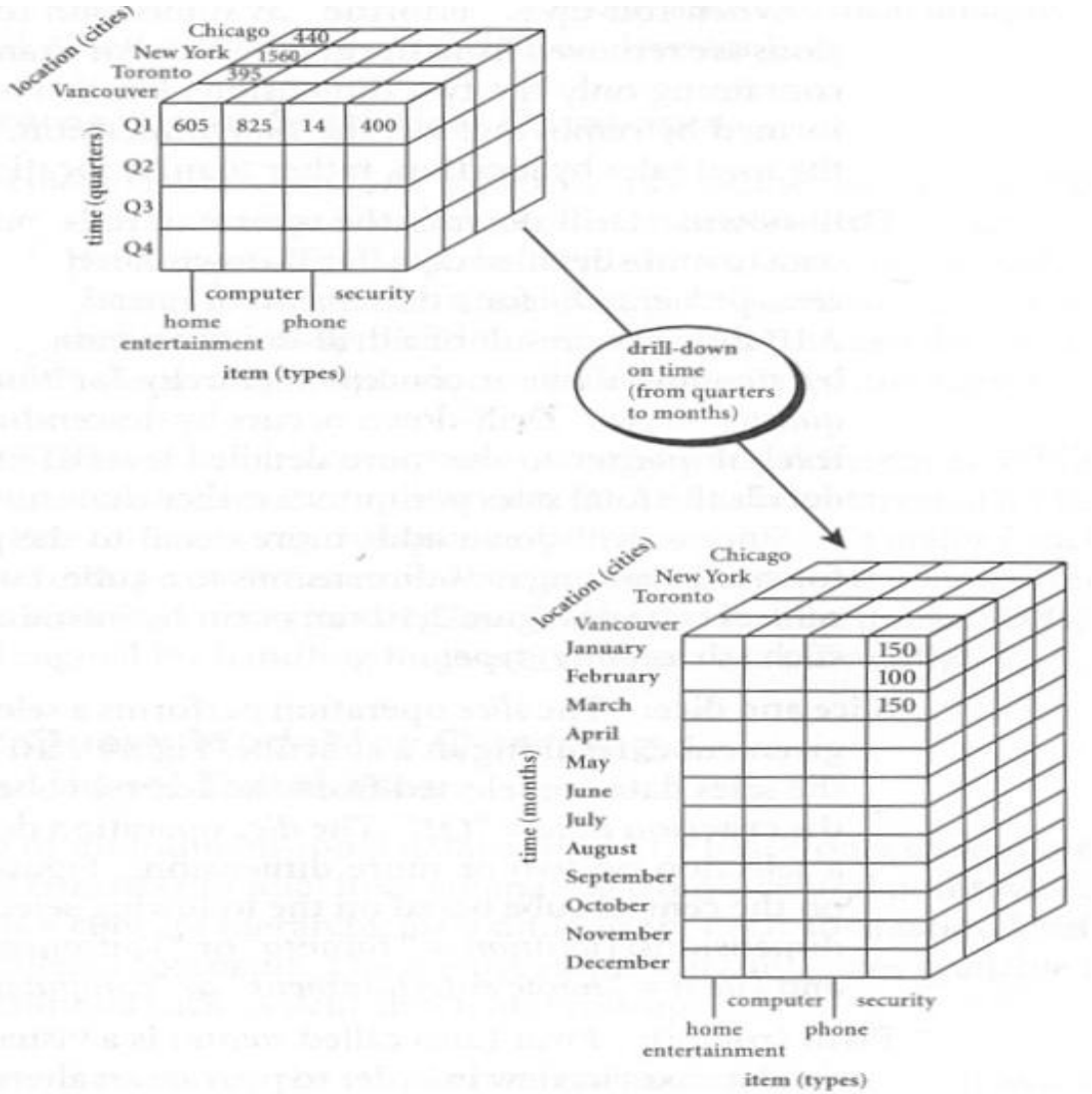
Opérations typiques de l'OLAP

- Roll up (drill up): consolider (résumer) les données
  - Passage à un niveau supérieur dans la hiérarchie d'une dimension
- Drill down (roll down): inverse du roll up
  - Descente dans la hiérarchie d'une dimension
- Slice and dice
  - Projection et sélection du modèle relationnel
- Pivot (rotate)
  - Réorientation du cube pour visualiser le 3D en une série de plans 2D

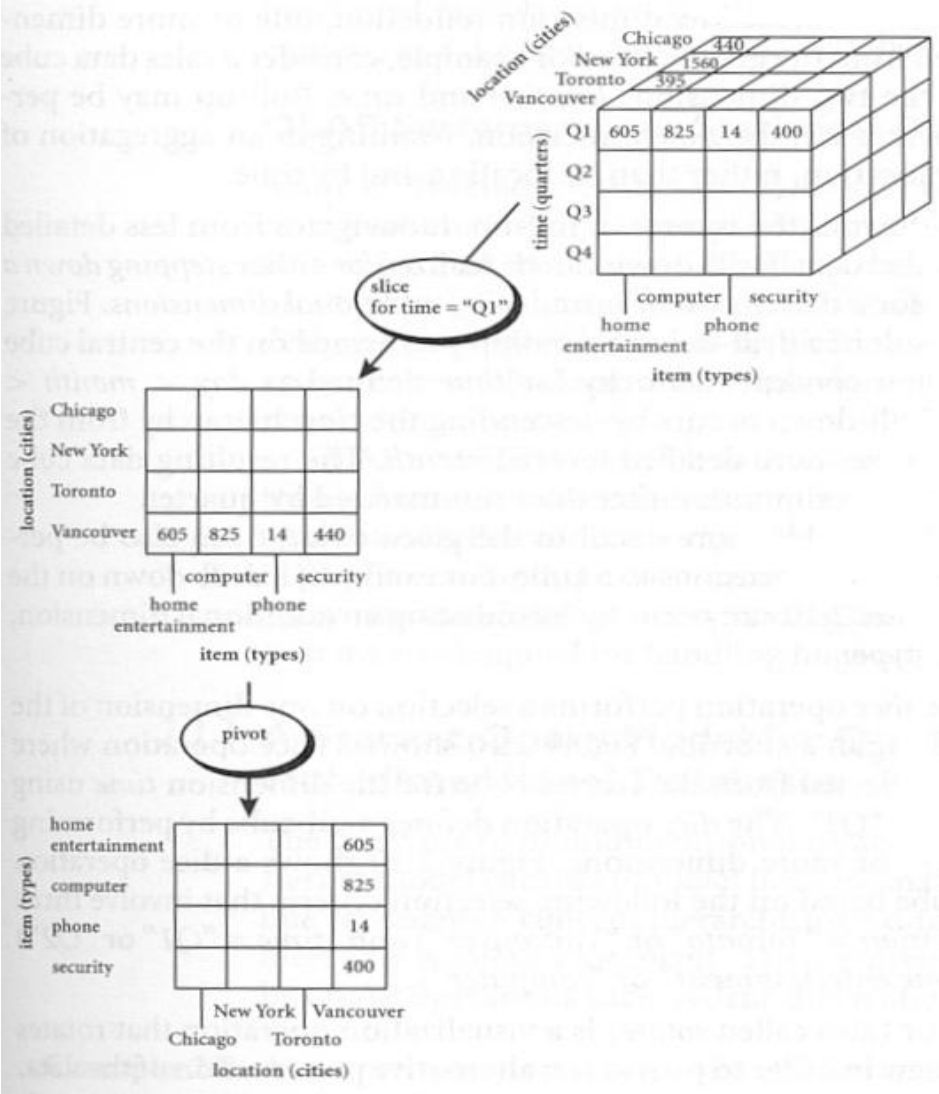
Roll Up


location (countries)
USA
Canada
2000
Q1
1000
Q2
Q3
Q4
time (quarters)
computer
security
home
entertainment
phone
item (types)
roll-up
on location
(from cities
to countries)
location (cities)
Chicago
440
New York
1560
Toronto
395
Vancouver
Q1
605
825
14
400
Q2
Q3
Q4
time (quarters)
computer
security
home
entertainment
phone
item (types)

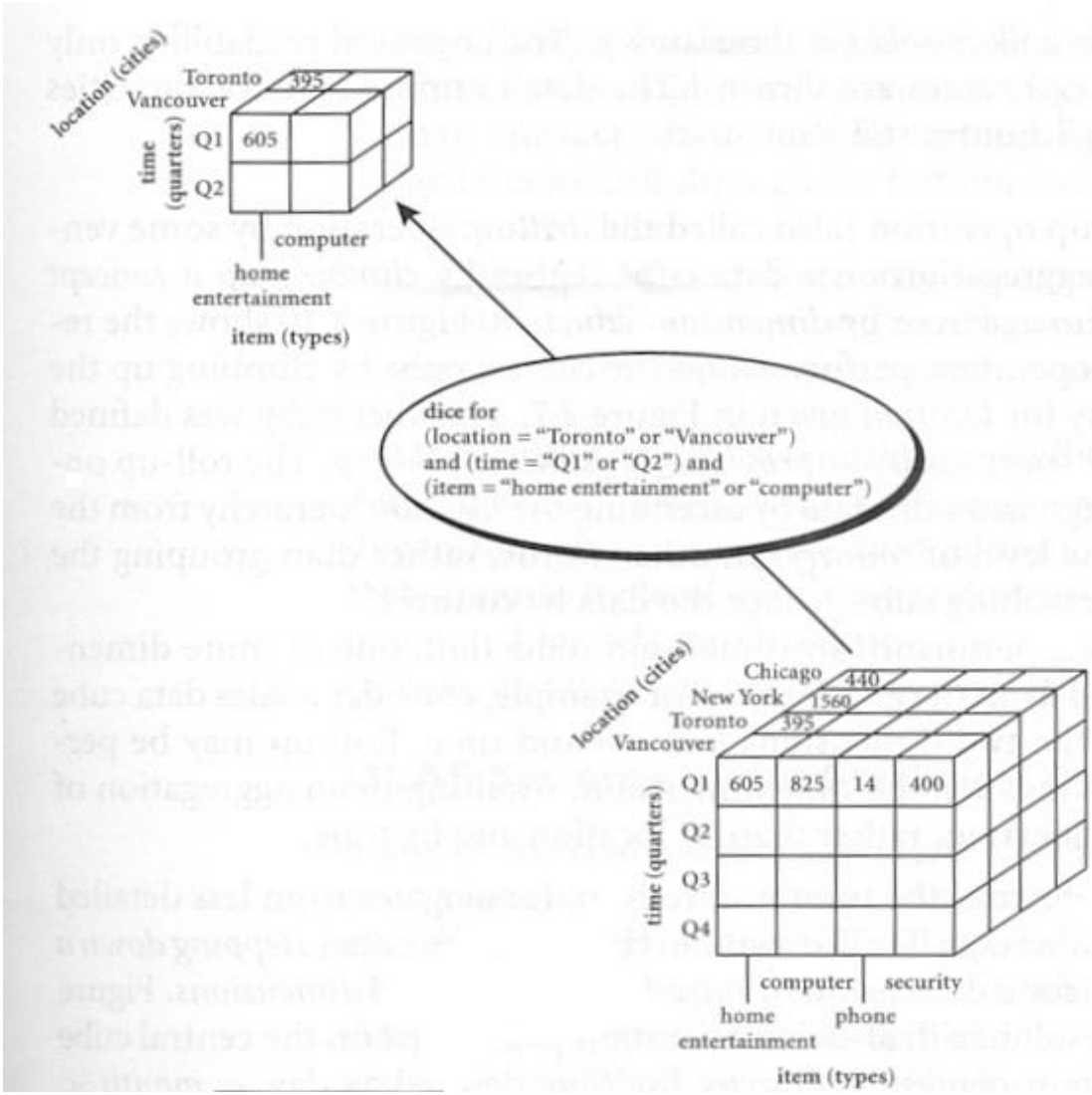
Drill Down


location (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
time (quarters)
Q1
Q2
Q3
Q4
605
825
14
400
computer
security
home
entertainment
phone
item (types)
drill-down
on time
(from quarters
to months)
location (cities)
Chicago
New York
Toronto
Vancouver
January
February
March
April
May
June
July
August
September
October
November
December
150
100
150
time (months)
computer
security
home
entertainment
phone
item (types)

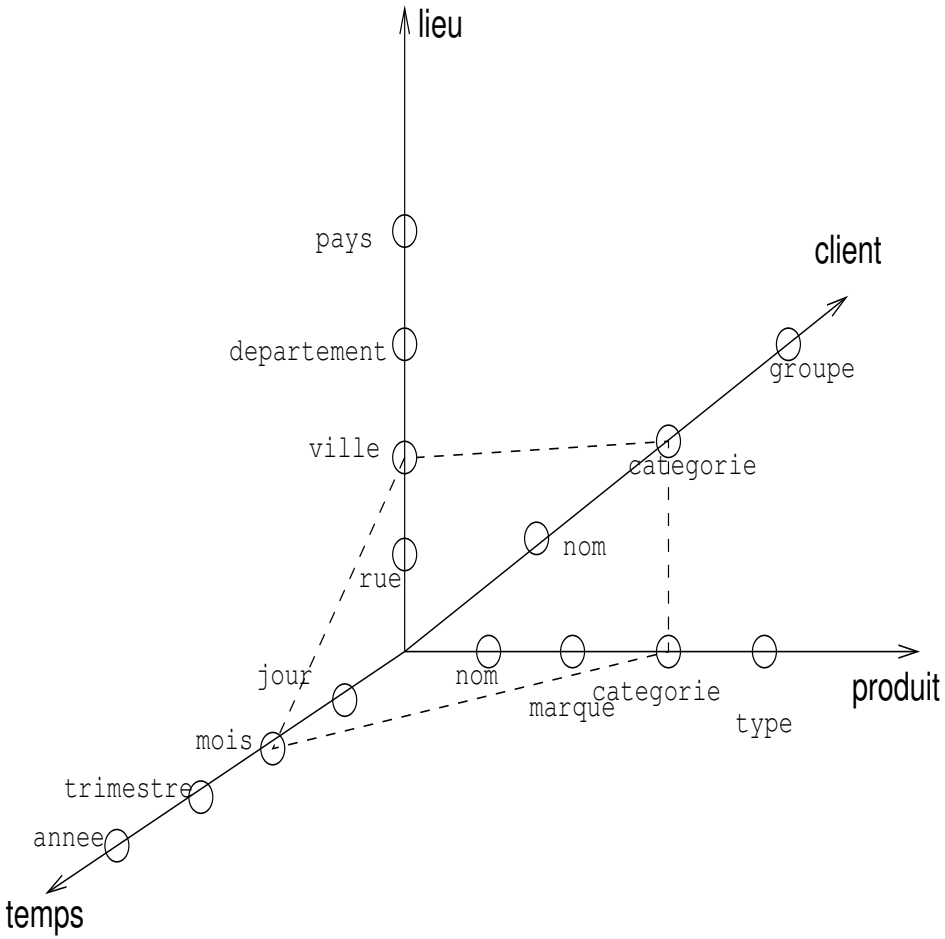
Slice and Pivot



Dice


location (cities)
Toronto
Vancouver
395
time (quarters)
Q1
605
Q2
computer
home
entertainment
item (types)
dice for
(location = "Toronto" or "Vancouver")
and (time = "Q1" or "Q2") and
(item = "home entertainment" or "computer")
location (cities)
Chicago
440
New York
1560
Toronto
395
Vancouver
time (quarters)
Q1
605
825
14
400
Q2
Q3
Q4
computer
security
home
entertainment
phone
item (types)

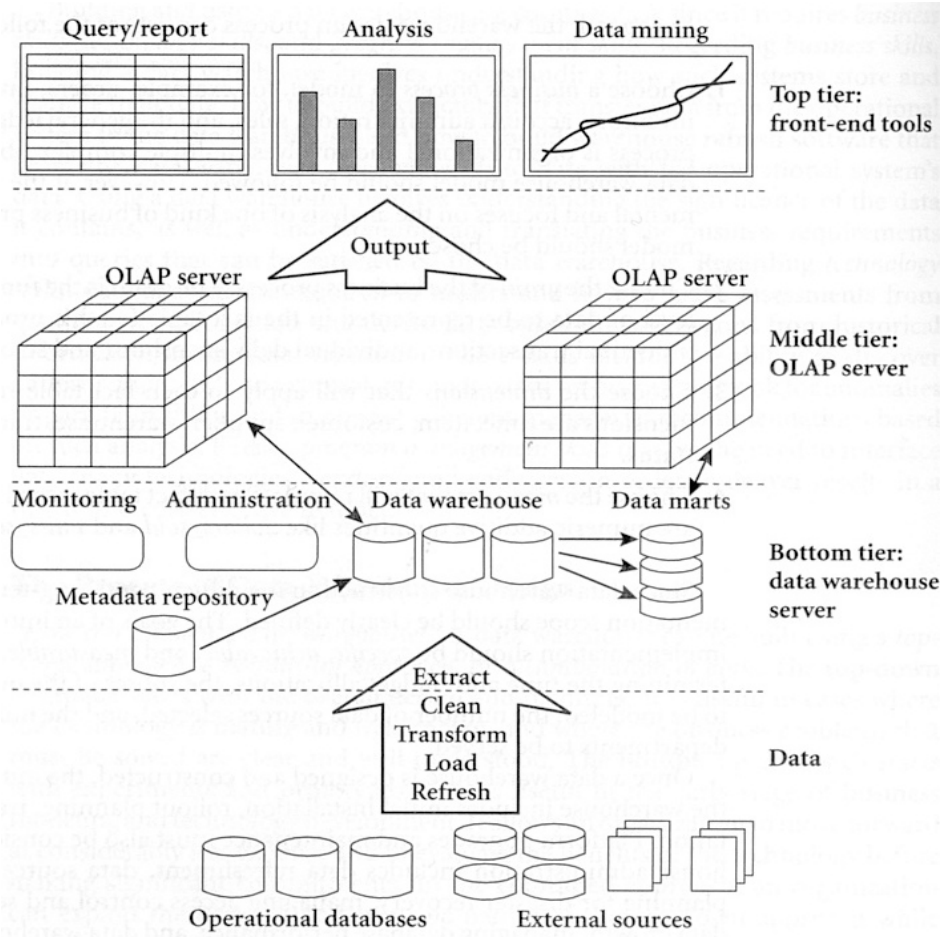
Un modèle pour représenter les requêtes



Entrepôt de données (data warehouse) et technologie OLAP (On-Line Analytical Processing) pour le Data Mining

- Un modèle de données multidimensionnel
- Qu'est-ce qu'un entrepôt de données?
- Architecture d'un entrepôt de données
- Implémentation d'un entrepôt de données
- Autres développements basés sur les cubes de données
- De l'entrepôt de données à la fouille de données

Architecture multi-niveaux



Architectures des serveurs OLAP

- Relational OLAP (ROLAP)
  - Utilise un SGDB relationnel pour stocker les données ainsi qu'un middle ware pour implémenter les opérations spécifiques de l'OLAP
- Multidimensional OLAP (MOLAP)
  - Basé sur un stockage par tableaux (techniques des matrices creuses)
  - Indexation rapide de données précalculées
- Hybrid OLAP (HOLAP)
  - Flexible selon l'utilisateur (eg. bas niveau: relationnel, haut niveau: tableau)
- Serveurs SQL spécialisés
  - support spécialisé pour les requêtes SQL à travers les schémas en étoile ou snowflake

Entrepôt de meta-données

Les meta données sont les données qui définissent les objets de l'entrepôt. Cela peut être:

- la description de la structure de l'entrepôt:
  schéma, vue, dimensions, hiérarchies
- les meta données opérationnelles:
  le suivi des données (historique des données migrées et chemin des transformations), la précision des données (actives, archivées ou nettoyées), le contrôle de l'information (statistiques, rapport d'erreurs)
- les algorithmes utilisés pour consolider
- le mapping d'un environnement opérationnel sur l'entrepôt de données
- les données liées à la performance du système
- les données de travail:
  les termes du milieu et les définitions, la propriété des données

Entrepôt de données (data warehouse) et technologie OLAP (On-Line Analytical Processing) pour le Data Mining

- Un modèle de données multidimensionnel
- Qu'est-ce qu'un entrepôt de données?
- Architecture d'un entrepôt de données
- Implémentation d'un entrepôt de données
- Autres développements basés sur les cubes de données
- De l'entrepôt de données à la fouille de données

Implémentation d'un entrepôt de données

- Opérations sur les cubes: define cube, compute cube, cube by
- Calcul efficace du cube
  - basé sur ROLAP
  - basé sur un tableau
  - bottom up
- Indexation des données OLAP
- Utilitaires

Utilitaires des data warehouses

- Extraction des données
  - récupérer les données à partir de sources multiples, hétérogènes et externes
- Nettoyage des données
  - détecter les erreurs et les rectifier si possible
- Transformation des données
  - convertir les formats
- Chargement des données
  - trier, consolider, calculer les vues, vérifier les contraintes et construire les index
- Rafraîchir les données
  - propager les mises à jour de la source des données vers l'entrepôt

Entrepôt de données (data warehouse) et technologie OLAP (On-Line Analytical Processing) pour le Data Mining

- Un modèle de données multidimensionnel
- Qu'est-ce qu'un entrepôt de données?
- Architecture d'un entrepôt de données
- Implémentation d'un entrepôt de données
- Autres développements basés sur les cubes de données
- De l'entrepôt de données à la fouille de données

Exploration des cubes de données orientée découverte

- Selon l'hypothèse: exploration par l'utilisateur, large espace de travail
- Selon la découverte:
  - Précalculer les données en indiquant les exceptions, guider l'utilisateur dasn l'analyse des données, à tous les niveaux d'aggrégation
  - Exception: significativement différente des valeurs prévues, basées sur un modèle statistique
  - Visualisation: ex: des fonds de couleur sont utilisés pour refléter les degrès d'exception pour chaque cellule
  - Calculer un indicateur d'exception pouvant être chevauchant avec la construction du cube

Exploration des cubes de données menée par la découverte: exemple

Sum of sales	*Month*											
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Total		1%	−1%	0%	1%	3%	−1%	−9%	−1%	2%	−4%	3%

Avg. sales	*Month*											
Item	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Sony b/w printer		9%	−8%	2%	−5%	14%	−4%	0%	41%	−13%	−15%	−11%
Sony color printer		0%	0%	3%	2%	4%	−10%	−13%	0%	4%	−6%	4%
HP b/w printer		−2%	1%	2%	3%	8%	0%	−12%	−9%	3%	−3%	6%
HP color printer		0%	0%	−2%	1%	0%	−1%	−7%	−2%	1%	−4%	1%
IBM desktop computer		1%	−2%	−1%	−1%	3%	3%	−10%	4%	1%	−4%	−1%
IBM laptop computer		0%	0%	−1%	3%	4%	2%	−10%	−2%	0%	−9%	3%
Toshiba desktop computer		−2%	−5%	1%	1%	−1%	1%	5%	−3%	−5%	−1%	−1%
Toshiba laptop computer		1%	0%	3%	0%	−2%	−2%	−5%	3%	2%	−1%	0%
Logitech mouse		3%	−2%	−1%	0%	4%	6%	−11%	2%	1%	−4%	0%
Ergo-way mouse		0%	0%	2%	3%	1%	−2%	−2%	−5%	0%	−5%	8%

Avg. sales	*Month*											
Region	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
North		−1%	−3%	−1%	0%	3%	4%	−7%	1%	0%	−3%	−3%
South		−1%	1%	−9%	6%	−1%	−39%	9%	−34%	4%	1%	7%
East		−1%	−2%	2%	−3%	1%	18%	−2%	11%	−3%	−2%	−1%
West		4%	0%	−1%	−3%	5%	1%	−18%	8%	5%	−8%	1%

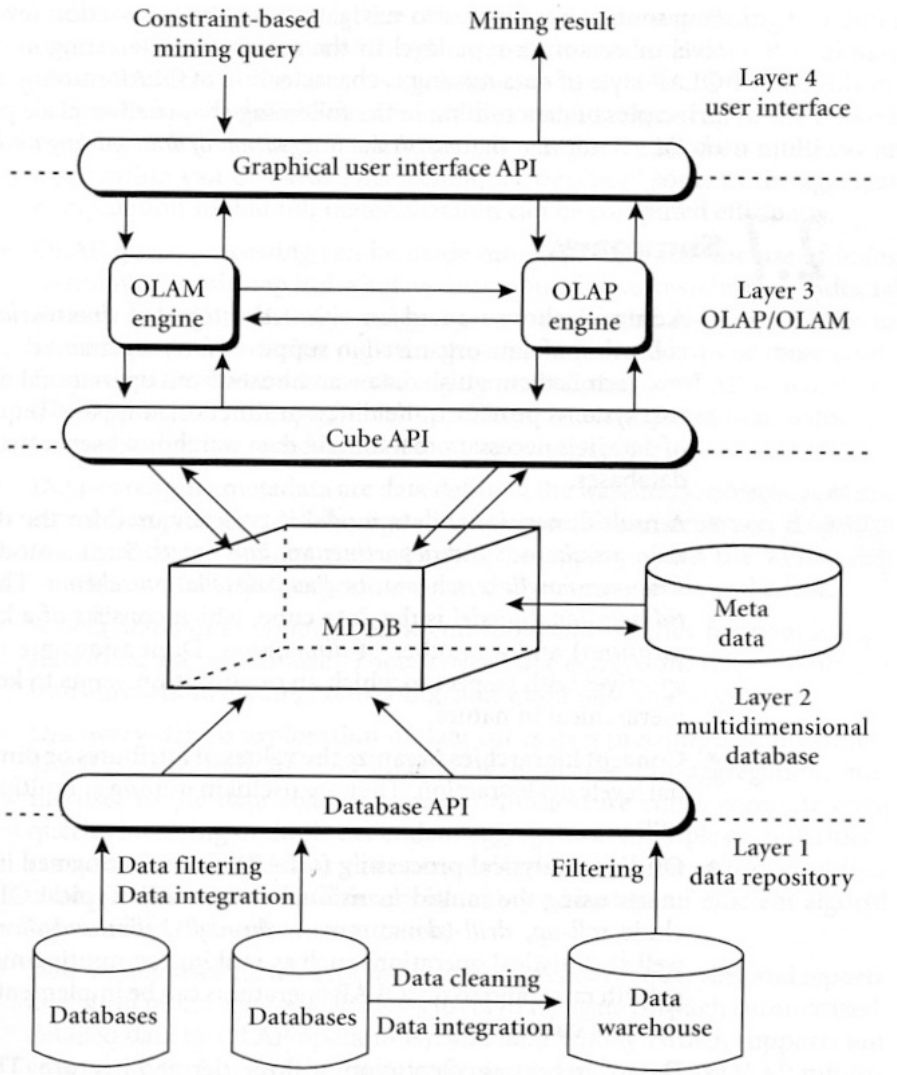
Entrepôt de données (data warehouse) et technologie OLAP (On-Line Analytical Processing) pour le Data Mining

- Un modèle de données multidimensionnel
- Qu'est-ce qu'un entrepôt de données?
- Architecture d'un entrepôt de données
- Implémentation d'un entrepôt de données
- Autres développements basés sur les cubes de données
- De l'entrepôt de données à la fouille de données

De OLAP à On Line Analytical Mining

- Pourquoi le OLAM?
  - Grande qualité de données dans les data warehouses
    - DW contient des données intégrées, consistentes et nettoyées
  - structure disponible du processing d'information entourant les entrepôts de données
    - accès via le web, facilité des services, propagation et utilisaires OLAP
  - Analyse de données basée sur l'OLAP
    - Signification avec le drilling, dicing, pivot, ...
  - Sélection on-line des fonctions du data mining
    - Intégration et échanges des nombreuses fonctions de la fouille, algorithmes et tâches
- Architecture d'OLAM

Architecture OLAM



Conclusion

- Entrepôt de données
  - Données orientées sujet, intégrées, dépendantes du temps, non-volatiles pour la prise de décision
- Modèle multidimensionnel pour les DW
  - Schéma en étoile, snowflake, constellation
  - Un cube de données est décrit par des dimensions et des mesures
- Opérations OLAP: drilling, rolling, slicing, dicing et pivoting
- Architectures OLAP: ROLAP, MOLAP, HOLAP
- Autres développements des technologies cube de données
  - Orienté découverte
  - OLAM