Analyse des données et découverte d'informations

La découverte d'informations consiste à développer des outils de mise en forme des données facilitant leur analyse. Elle repose sur deux aspects :

- projection de données qualitatives sur des espaces vectoriels ("quantification" des données)
- production d'histogrammes dans le but d'analyser la distribution des données dans l'espace de reconstruction

Le but est de dégager :

- des **tendances** (covariables)
- des **modes** de la distribution (présence de plusieurs maxima)

à partir d'un **grand** ensemble de données (chiffre d'affaires, nb de ventes, masse salariale, ...) évoluant dans le **temps** et dans l'**espace**, afin de

- définir des **indicateurs** pertinents
- faciliter la prise de décision.

Vocabulaire anglophone généralement utilisé :

- Business Intelligence (BI)
- Data Warehouses (Entrepôts de données)
- OLAP (Online Analytical Processing) :

"*Unlike Online Transaction Processing (OLTP), where typical operations read and modify individual and small numbers of records, OLAP deals with data in bulk, and operations are generally read-only.*"

Entrepôts de données (Data warehouses) / Magasins de données (Data Mart)

Exemples de grandes masses de données :

- Masses de données (pullulantes) : tickets de caisse, clics web, appels tel, operations bancaires, remboursements no URSSAF, trajets SNCF…
- Données importantes : fichiers de clients, données biométriques, campagnes de mesures, sondages,…
- Données géographiquement localisées (gestion d'un "territoire") : appels tel, centres de production, consommation eau-électricité-gaz, infractions pénales, arrêts maladie, prêts bancaires, allocations chômage, jugements des TGI, accidents du travail…

Remarque : Les transactions marchandes sont un cas typique/fondateur (acte d'achat bien répertorié et enregistrés, livres de comptes, ...)

Cas d'utilisation :

- (qui?) Quels sont les magasins les plus rentables? doit-on ouvrir / fermer des magasins?
- Où doit-on implanter un nouveau magasin?
- Y a-t-il une corrélation entre le lancement d'une campagne publicitaire et les chiffres de vente? quels sont les supports les plus efficaces?
- (qui?) Quelle est la liste des clients à contacter?
- (quand?) De quelle quantité doit-on approvisionner les magasins en fonction de la période de l'année?

Analyse :

- Quels sont les catégories de films/livres les plus fréquemment empruntés?
- Réussite / taux d'embauche / salaire en fonction de la prépa d'origine / sexe / profession des parents

1. Aggrégation

1.1 Opérateurs d'aggrégation

usage : statistique sur les données

principe :

- **opérateur d'aggrégation** : comptage, somme, moyenne, ecart-type (count, sum, mean, avg...)
- données quantitatives (et non qualitatives)
- classes (données qualitatives)

Exemples de requêtes faisant appel aux fonctions d'aggrégation :

Nombre d'élèves par groupe de TD / par prepa d'origine etc..:

```
select groupe_TD , count(num_eleve)
from eleve
group by groupe_TD
```

Donner les chiffres des ventes du magasin pour chaque mois de l'année

```
select mois, sum(montant)
from vente
group by mois
```

Donner le nombre de ventes d'un montant > à 1000 euros pour chaque mois de l'année

```
select mois, count(num_vente)
from vente
group by mois
having montant >= 1000
```

Tester les diaparités salariales entre hommes et femmes

```
SELECT gender, avg( salary )
FROM employee
GROUP BY gender
```

Tester les diaparités salariales selon le niveau d'éducation

```
SELECT education_level, avg( salary )
FROM employee
GROUP BY education_level
```

Problèmes :

- Tester les disparités salariales hommes/femmes en fonction du niveau d'éducation.
- donner le taux de réussite par groupe de matière en fonction de la filière d'origine (MP, PSI, PC, PT, …)
    - plusieurs "GROUP BY"??? ⇒ dimensions
- A quelles heures de la journée la messagerie est-elle la plus sollicitée?
    - définir des intervalles temporels?? créneaux horaires? ⇒ distributions, histogrammes
- Comment se répartissent géographiquement les utilisateurs de la messagerie?
- définir des secteurs géographiques?
    - ⇒ hiérarchies pays > département > région
- Taille du message, has_attachement?
    - ⇒ mesures sur des faits élémentaires

1.2 Faits élémentaires

- Notion de fait élémentaire (*fact*): transaction ou opération localisée dans le temps et dans l'espace

Exemples de "fait":

- Achat/Vente
- Opération bancaire (débit/crédit)
- Consultation (site web)
- Souscription à un contrat d'assurance
- Appel téléphonique
- Inscription

Tous ces faits peuvent être localisés. Des mesures peuvent être effectuées sur ces faits (montant d'une vente, durée d'un appel, montant d'une opération bancaire, ...)

Points clés :

- distinction entre **Dimension** et **Mesure**.
  - Notion de dimension : qui? quoi? où? quand? Comment? : associe des **coordonnées** à l'événement (géographiques, temporelles) et par extension une catégorie.
  - Notion de **mesure(s)** associées à l'événement (exemple : montant de la vente)
- distributions, répartitions, etc... cf analogie proba/stats : événement aléatoire, vecteur aléatoire, ...
  - les événements sont associés par paquets sur des intervalles réguliers ou selon des catégories discretes.
  - Fonctions d'aggrégation : réalise la mesure sur les groupe : somme, comptage, moyenne, min, max, etc...
  - histogramme : nb d'événements observés par secteur sur un maillage régulier de l'espace des coordonnées. Par extension mesure sur ce maillage par une fonction d'aggrégation.

1.3 Cube de données

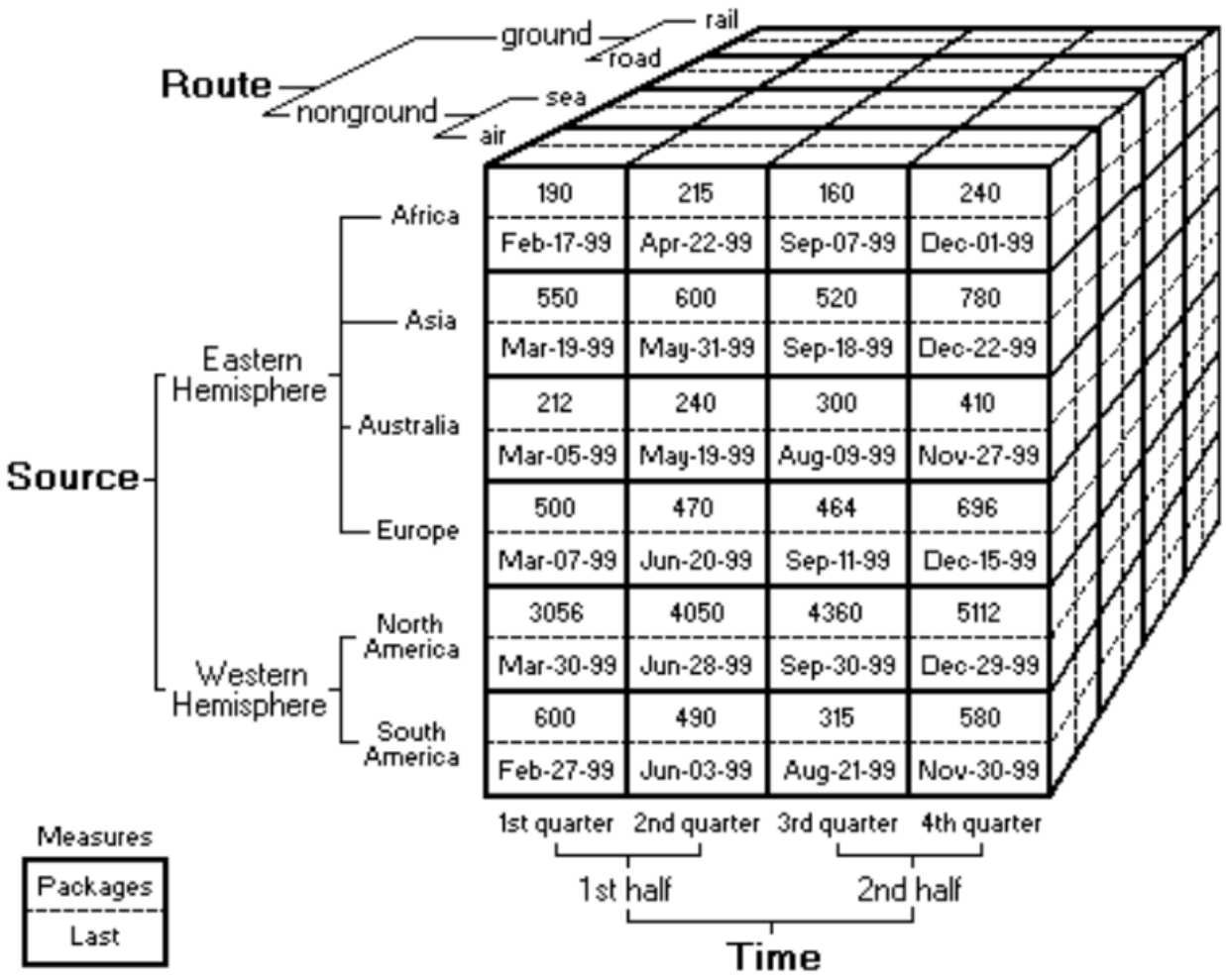
Un cube de données est une structure de données organisée sur le principe des espaces vectoriels. Différents axes sont définis, chaque axe étant associé à une dimension particulière.

- Les dimensions peuvent correspondre à des valeurs discrètes (catégories : type de produit, catégorie de client,...) ou continues (valeurs temporelles ou géographiques, ...).
- Chaque fait est décrit comme un point de l'espace vectoriel. Il est positionné dans une cellule du cube. A ce point sont associées une ou plusieur mesures.
- Le cube est un ensemble de cellules (voir figure), chaque cellule correspondant à un intervalle (sur les axes continus) ou une valeur (sur les axes discrets).

Un élément essentiel du modèle de données est la définition de **hiérarchies** sur les dimensions du cube. Chaque dimension se divise en intervalles et sous-intervalles (pour le continu/ quantitatif) ou en catégories et sous-catégories (pour le discret/qualitatif)

Les hiérarchies sur les différentes dimensions permettent de définir le "niveau de résolution" sur les différentes dimensions.

- On peut ainsi s'intéresser à l'évolution d'une certaine grandeur au cours du temps année par année, trimestre par trimestre ou mois par mois selon le niveau de résolution choisi.
- → Hiérarchie : description arborescente d'intervalles et de sous-intervalles sur une dimension. Implemente differentes granularités sur la dimension considérée.



La structure de cube de données est adaptée pour la réalisation d'histogramme multidimensionnels, selon les axes choisis et le niveau de résolution choisi, à l'aide de fonctions d'aggrégation.

- Histogramme et aggrégation
  - (vue quantitative) comptage/répartition d'événements sur un intervalle (discrétisaton d'une distribution d'événements)
  - (vue qualitative) comptage d'événements par catégorie
  - (vue intermediaire) comptage d'événements par catégories hiérarchisées

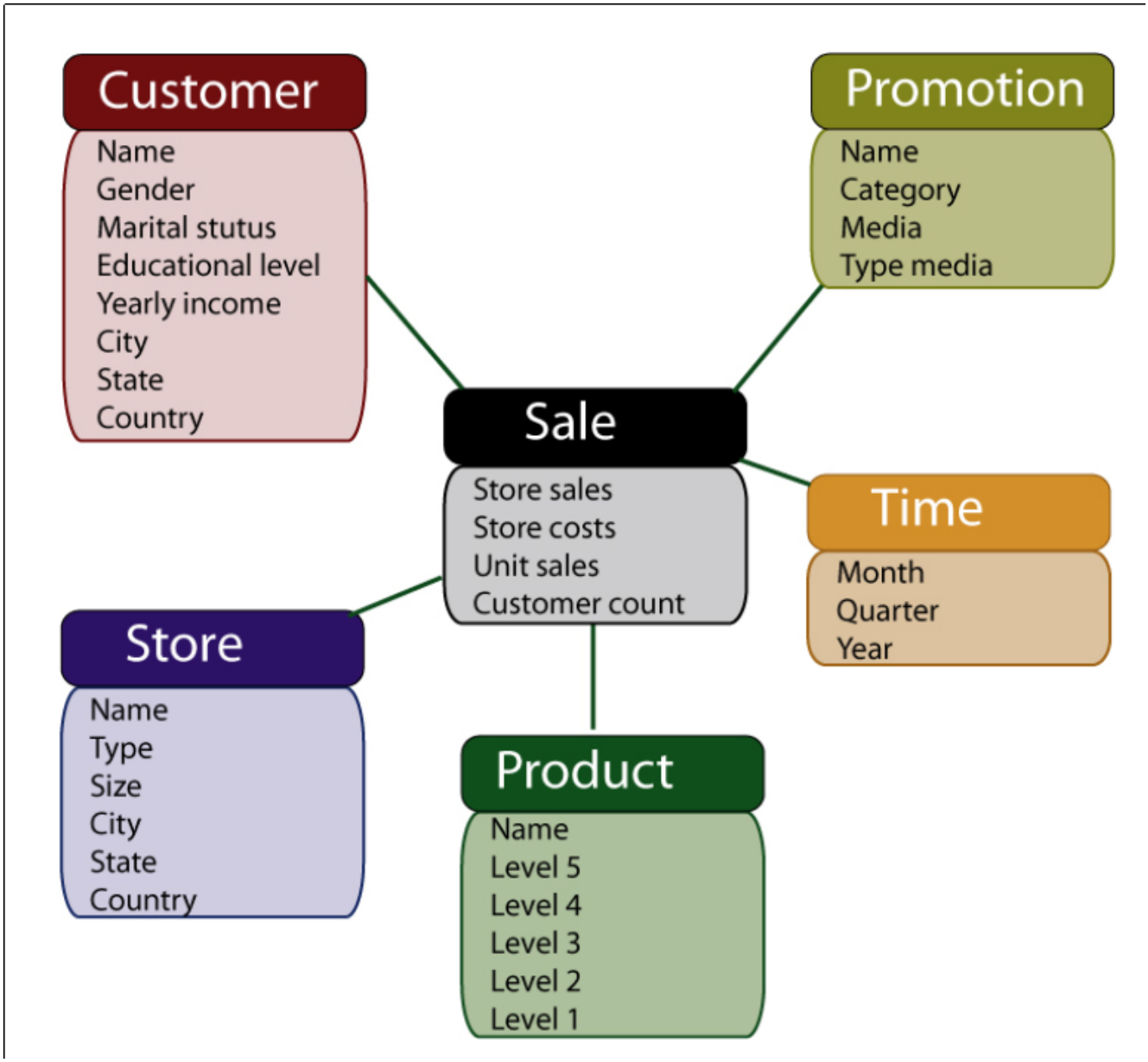
1.4 Modèle de données en étoile

La réalisation d'un cube de données repose en général sur une base de données relationnelle organisée selon un "modèle en étoile".

Le modèle en étoile est une extension des schéma Entité/Association pour lesquels :

- un fait est une association située au centre du schéma. Les attributs de l'association sont les mesures effectuées
- une dimension est une relation participant au fait. Les dimensions sont donc décrites par des attributs (ex : attributs année, trimestre, mois, jour, heure, minute, seconde,...pour une dimension temporelle)

- pour chaque dimension, on décrit une hiérarchie sur les différents attributs de la dimension en définissant un ordre, du particulier au général.



Exemples :

- sur la dimension temporelle : mois ⊂ trimestre ⊂ année
- sur la dimension promotion : nom ⊂ catégorie ⊂ média ⊂ type de média

etc...

2. Mise en oeuvre

Pandas

http://pandas.pydata.org/pandas-docs/stable/10min.html

XMLA / MDX