Gestion de Masse de Données (GMD)

Introduction

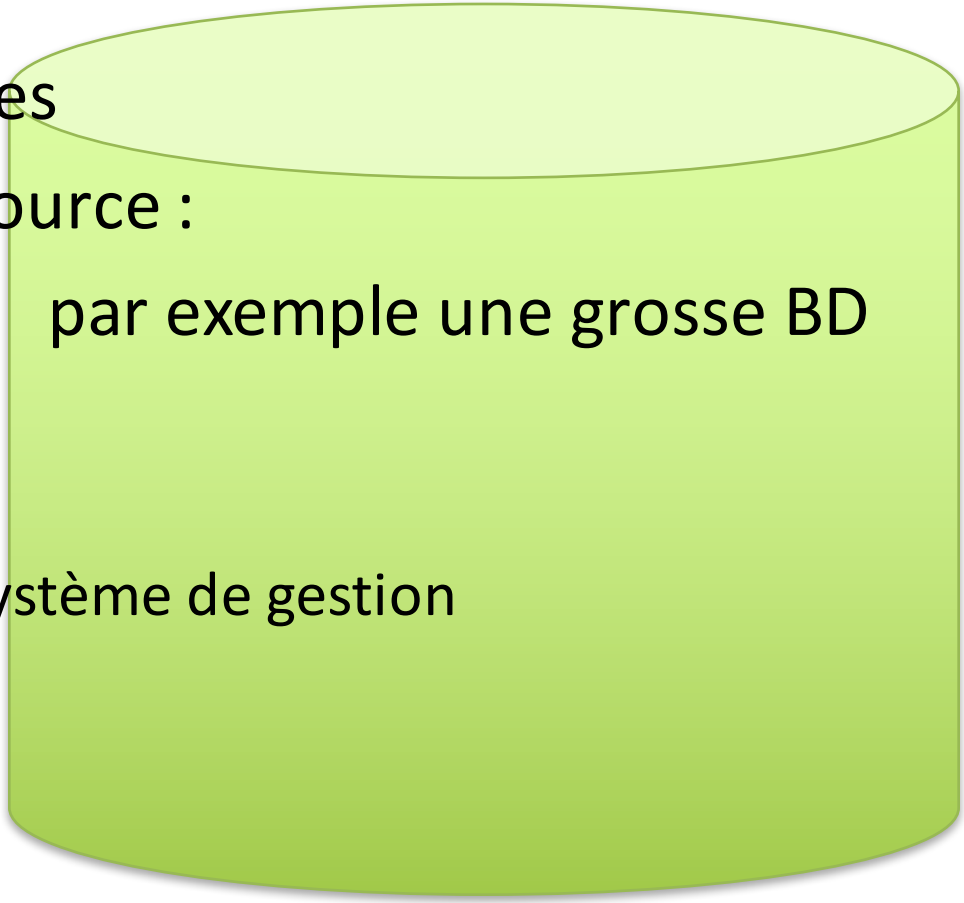
Présentation du module

- Problématique informatique
- Objectifs
- Plan
- Organisation du module (CM/TP/Projet/Evaluation)

GMD : problématique (1/4)

- CAS 1 : Une masse de données

  regroupée dans une seule source :



  - problème **d'optimisation** du système de gestion
    - *ex1 : création d'indexes*
    - *ex2 : division de tables*

GMD (un peu) + Module "BD Avancées" (beaucoup)

Limites de taille des sources de données

- Systèmes de fichiers

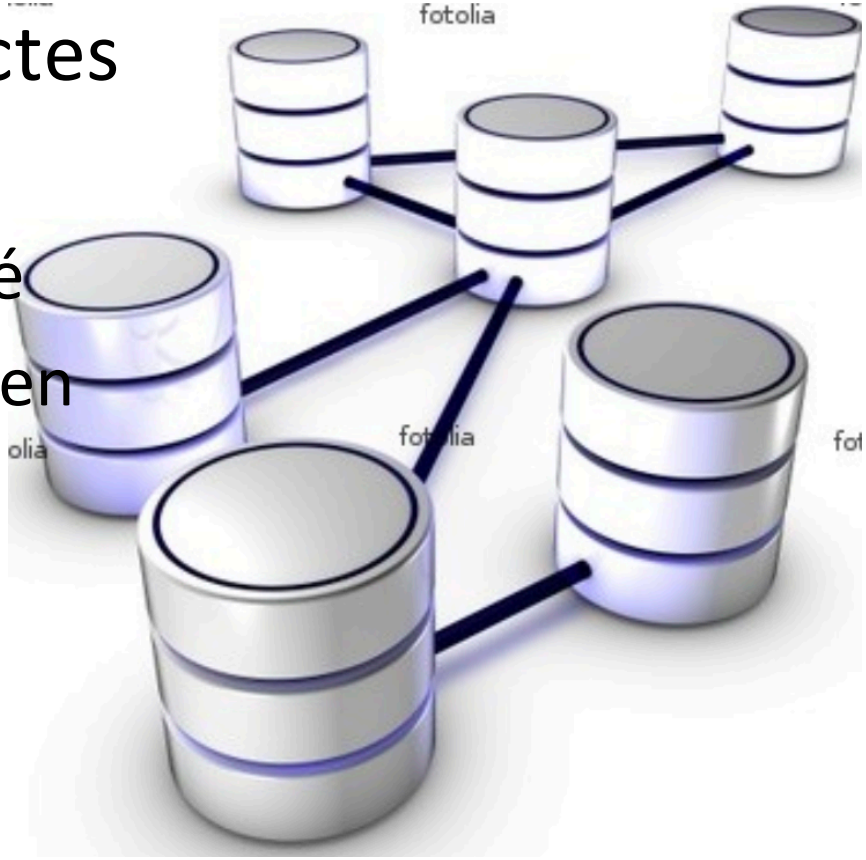
	taille max d'1 fichier	*taille max d'1 volume*	
– *FAT 32*	4GB	2TB	
– *HFS Plus*	8EB	8EB	(OS10.3.9)
– *NTFS*	16EB	16EB	
– *ext4*	16TB	16TB	

- Bases de données

	taille max d'une table	*taille max d'une base*
– *Excel 2010*	65 536 ligne x 256 col	
– *Access 2010*	2GB	2GB
– *MySQL 5*	256/64TB	Illimité (en théorie)
– *Oracle 10*	4GB x block size(*i.e.* 4KB)	4GB x block size

GMD : problématique (2/4)

- CAS2 : des données réparties dans de nombreuses sources de données distinctes
  - le volume de données à manipuler peut être très élevé
  - c'est le cas le plus fréquent en pratique



Problème : les sources de données sont **hétérogènes**

GMD : problématique (3/4)

- Les sources sont **hétérogènes**
  - en terme de contenu
  - en terme de localisation physique

    ex : locale ou distante
  - en terme d'accès

    ex : parsing, appel de Web service, requête relationnelle
  - en terme de qualité
  - en terme de syntaxe (format de données)

    ex : XML, schéma relationnelles, fichier CSV, textes
  - en terme de sémantique (sens associé aux données)

    ex : polysémie, synonymie

Quelques "*définitions*"

- **Source** ou ressource de données

  ex : BD, fichier, site web, corpus de texte

- **Élément** d'une source de données

  ex : un n-uplet, une ligne d'un tableau, une page web, un texte

- Système d'intégration

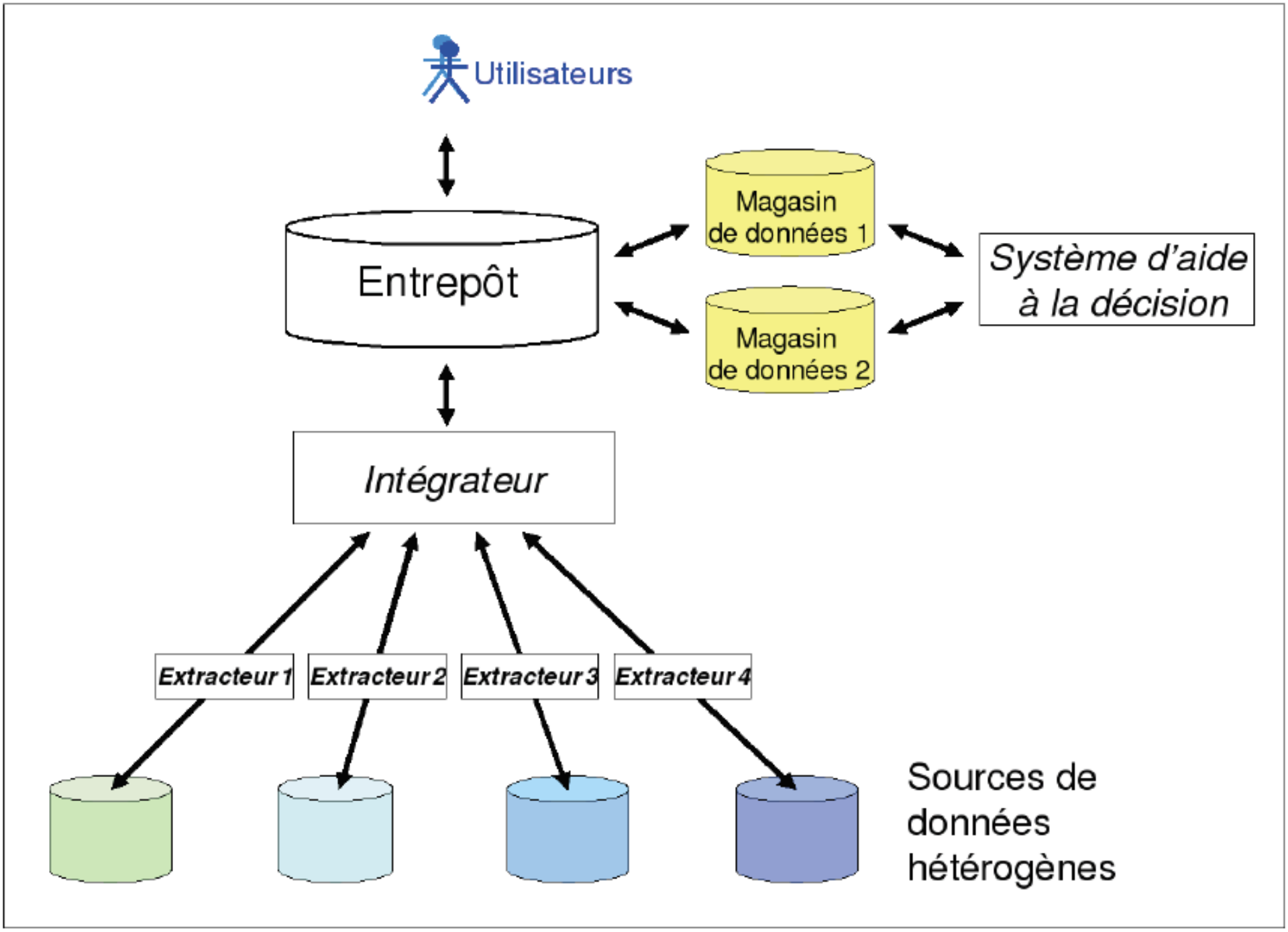
  système qui permette d'interroger de façon uniforme et transparente des source de données hétérogènes

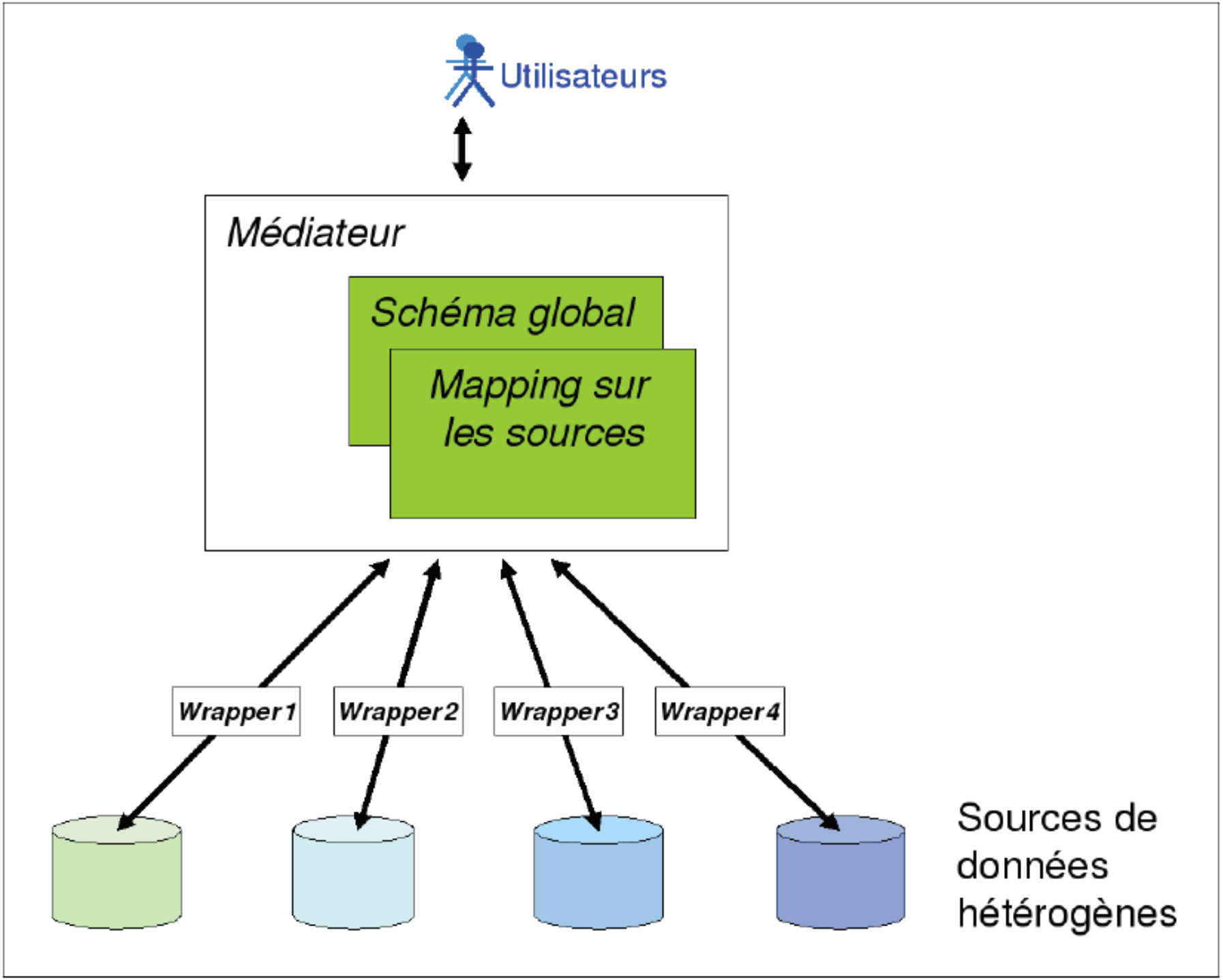
- Mise en correspondance (ou mapping)

GMD : problématique (4/4)

- Le vrai défi pour manipuler des masses de données, c'est de pouvoir utiliser différentes sources hétérogènes ensembles, càd de les intégrer

- On distingue deux grandes approches (et deux type d'architectures associés) pour l'intégration de données :
  - matérialisée
  - dématérialisée (ou fédérée ou à médiateur)


Utilisateurs
Entrepôt
Magasin de données 1
Magasin de données 2
Système d'aide à la décision
Intégrateur
Extracteur 1
Extracteur 2
Extracteur 3
Extracteur 4
Sources de données hétérogènes


Utilisateurs
Médiateur
Schéma global
Mapping sur les sources
Wrapper1
Wrapper2
Wrapper3
Wrapper4
Sources de données hétérogènes

GMD : objectifs

Vous donner des clés pour

1) accéder à
2) extraire
3) transformer
4) intégrer

des données de sources hétérogènes

de façon efficace

de façon cohérente

Un cours pour des ingénieurs

GMD : organisation du module

- 4 Cours/ 6 TP
- 10 Séances de projet
- Evaluation
  - Examen écrit de 2h, documents autorisés (30/03 à 8h!!)
  - Projet (+ soutenance)

GMD : plan du cours

I. Accéder et extraire des données
   ① dans un système de gestion de fichier,
      à partir de fichiers textes
   ② par un service Web
      à partir de fichier XML
   ③ par utilisation d'une API de programmation
      dans une BD relationnelle
II. Transformer les données
   ① par rapport à un schéma global ou une ontologie
   ② pour gérer les données manquantes ou bruitées
III. Regrouper et interroger les données
   ① de façon matérialisée (entrepôts, cubes de données, NoSQL)
   ② de façon dématérialisée (systèmes médiateurs)

GMD : projet noté

- Groupes de 3
- 10 séances encadrés (TP) prévues
- implémenter un système dématérialisé d'intégration de données
- Soutenances:
  - démo, motivation des choix techniques