

Base de données

S. Lèbre
slebre@unistra.fr

Université de Strasbourg, département d'informatique.

Présentation du module

- Contenu général
 - Notion de bases de données
 - Fondements / Conception
 - Utilisation : langage SQL (Oracle)
- Organisation
 - Séances : 14 heures de cours
 - TD : 12h, TP : 8h.
 - Evaluation
 - Contrôle continu (TP noté)
 - Examen
- Documents de cours disponibles sur :
<http://lsiit-cnrs.unistra.fr/bfo-fr/index.php/Cours>

Plan

1. Qu'est-ce qu'une base de données ?
 1. Définitions et enjeux
 2. Système de Gestion de Bases de Données (SGBD)
 3. Système complet de base de données
 4. Architecture des base de données
2. Le modèle logique de base de données
➡ **Modèle relationnel**
3. Algèbre relationnelle

Plan

4. Utilisation d'un SGBD : le langage SQL
 1. Langage de Manipulation des Données (LMD) : Requêtes simples, expressions et fonctions, sous-requêtes, jointures : **select, update, delete.**
 2. Langage de Définition des Données (LDD) : types de colonnes, manipulations de tables : **create, drop, alter.**
5. Normalisation
6. Modélisation de base de données relationnelles
 1. Le modèle Entités-Associations (E/A) (modèle conceptuel)
 2. Traduction en modèle relationnel (modèle logique)

Chapitre I

Qu'est-ce qu'une base de données ?

1. Définitions et Enjeux

Définitions

- Base de Données (BD) ou Database (DB)
 - C'est un ensemble de **données** structuré et organisé permettant le stockage **d'informations** utiles à l'entreprise



Ne pas confondre !

- Système de Gestion de Base de Données (SGBD) ou Database Management System (DBMS)
 - Ensemble de logiciels permettant l'exploitation d'une base de données (ajout, mise à jour, recherche ...).

Enjeux

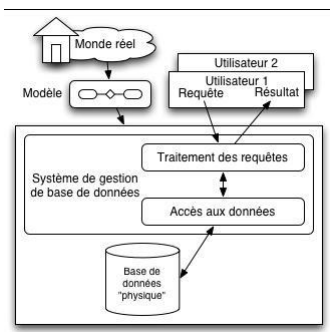
➡ Développement de concepts, méthodes et algorithmes spécifiques.

- Gérer les données en mémoire secondaire (*i.e.* disques durs)
 - Les bases de données doivent être capables de gérer volumes des données de + en + importants (plusieurs dizaines de téra-octets)
 - Pérennité des données
- Multi-utilisateurs
 - Internet : plusieurs dizaines de milliers d'utilisateurs dans un contexte d'exploitation changeant .
 - Gestion multi-agents : entre plusieurs magasins.
- Aide à la décision (OLAP = OnLine Analytical Processing)

"Les gagnants seront ceux qui restructurent la manière dont l'information circule dans leur entreprise." (Bill Gates)

2. Système de Gestion de Bases de Données (SGBD)

Illustration du fonctionnement d'un SGBD



Objectifs du SGBD

- Un SGBD doit répondre à plusieurs objectifs
 - Objectifs principaux :
 - Langage de manipulation
 - Indépendance données/SGBD
 - Administration aisée des données
 - Objectifs secondaires :
 - Intégrité et cohérence des données
 - Redondance contrôlée des données
 - Efficacité d'accès aux données
 - Partage des données
 - Sécurité des données

Langage de manipulation

- L'utilisateur doit pouvoir utiliser un langage de requête sur la BD, il doit être :
 - Simple à écrire
 - Déclaratif (dire ce que l'on veut sans préciser la façon de l'avoir)
 - Optimisé automatiquement par le processeur de requête
- Exemple : langage SQL (Structured Query Language).

Indépendance données/SGBD

Idéalement, les applications ne doivent pas dépendre de la manière dont les données sont stockées ou de comment on y accède.

- Indépendance **physique**
 - Un changement au niveau physique (comment les données sont stockées) ne doit pas engendrer une modification de l'application.
- Indépendance **logique (Plus difficile à atteindre)**
 - un changement au niveau logique (objets, tables, colonnes, rangées, ...) ne doit pas engendrer une modification de l'application.

Administration facilitée

- Le SGDB doit comprendre des outils pour :
 - Définition et modification de la description de données (schéma)
 - Vérifier l'adéquation entre la structure physique et le besoin des applications
 - Suivre les performances et faire des statistiques
- Gestion des « méta-données »
 - Données propres au fonctionnement du SGBD (utilisateur , droits ...)

Cohérence des données

- Une donnée doit respecter un **format**, un **type** ou un **domaine** de valeur.
- Intégrité des données
- Exemple :
 - Date invalide, âge négatif...
 - Panne pendant une modification de la BD

Redondance contrôlée

- Eviter un maximum les répliquions de données
 - Perte de place
 - Difficulté de Mise à jour (MAJ) des informations (→ source d'erreurs)
- Possibilité de redondance gérée par le SGBD
 - Afin d'optimisation des performances
 - Aide à l'analyse
 - Sauvegarde

Efficacité d'accès aux données

- Performances importantes
 - Réponse simultanée à différentes requêtes
- Partage des ressources
- Mécanismes d'accès
 - Authentification et droits
 - Verrou
 - Transaction
- Problème principal : E/S disque
 - Utilisation du cache et de la mémoire (RAM)

Partage des données

- Accès multi utilisateur **mais** chaque application doit accéder aux données comme si elle était la seule à accéder à la BD.
- Gestion de l'accès disque
 - FCFS (« first come, first serve »)
 - Minimisation du temps d'attente
 - SSTF (« shortest seek time first »)
 - Minimisation des mouvement mécaniques
 - Maximisation du débit

Sécurité des données

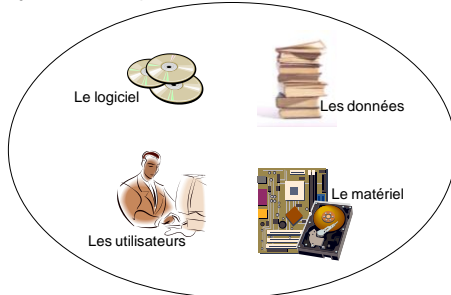
- Protection contre les accès non autorisés
 - Authentification
 - Gestion de droits (Accès, Opérations)
- Protection des données
 - Reprise sur panne
 - Sauvegarde / exportation / importation
 - Journalisation
 - Redondance matérielle ...

Quelques SGBD connus

- Il existe de nombreux systèmes de gestion de bases de données, en voici une liste non exhaustive :
 - **MySQL** : <http://www.mysql.org/> (domaine public)
 - **PostgreSQL** : <http://www.postgresql.org/> (domaine public)
 - **Oracle** : <http://www.oracle.com/> (Oracle Corporation)
 - **IBM DB2** : <http://www-306.ibm.com/software/data/db2/>
 - **Microsoft SQL** : <http://www.microsoft.com/sql/>
 - **Sybase** : <http://www.sybase.com/linux>
 - **Informix** : <http://www-306.ibm.com/software/data/informix/>

3. Système complet de BD

Système complet de BD



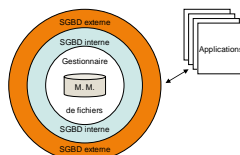
Les données

- **Données volatiles**
 - Variables du programme
 - Disparaissent à la fin de l'exécution
 - Stockées en mémoire principale (RAM)
- **Données persistantes**
 - Ne sont pas liées au programme
 - Persistent à la fin de l'exécution
 - Stockées en mémoire de masse (disque dur, cd-rom, bande, ...)

Ne pas confondre !

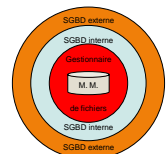
Le logiciel : le SGBD

- C'est le SGBD, le chef d'orchestre.
- Gère le partage et la cohérence des données
- Permet l'indépendance des données grâce à un modèle en 3 couches :
 - Le gestionnaire de fichiers
 - Le SGBD interne
 - Le SGBD externe



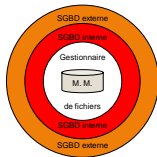
Le gestionnaire de fichiers

- Il gère la **structure de mémorisation** (fichiers)
- Il gère le **stockage** des données dans les supports physiques
- Gestion des **accès** (index, clé ...)



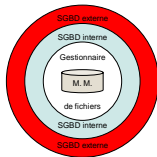
Le SGBD interne

- Il gère les **données stockées dans les fichiers**
- **Regroupe** les données sous formes d'objets
- Gère les **liens** entres ces objets
- Optimise l'**accès** à ces objets
- Gère les **conflits** d'accès
- **Coordination** et suivi des processus



Le SGBD Externe

- **Interprète et analyse** les requêtes
- **Transforme** les requêtes en primitives internes
- **Met en forme** et présente les données aux applications
- Contrôle les **droits d'accès**



Les utilisateurs

- **Programmeur d'applications**
 - Responsable de l'écriture des programmes utiises par la base de données
 - Cobol, C, C++, Java, VB, C#, PHP, ...
- **Utilisateur final**
 - Interagit avec la BD à travers une application ou une interface spécifique du SGBD



Les utilisateurs

- **Administrateur des données (AD)**
 - Doit bien connaître l'entreprise et ses besoins
 - Il décide des types de données à stocker
 - Il décide de la politique de manipulation et de maintenance des données
- **Administrateur de la base de données (ABD)**
 - Il implante la BD
 - Il met en place les techniques nécessaires pour répondre à la politique et aux choix de l'AD
 - Il est responsable du suivi des performances du système.



Le matériel

- Essentiellement les mémoires

Types	Temps d'accès	Gestion	Prix
Bandes	Minutes	Humaine	<1 €/Go
CD / DVD	100 ms	Humaine	<1 €/Go
Disque Dur	10 ms	Humaine	<1 €/Go
RAM	10 ns	Compilateur	> 100 €/Go
Cache	0,5 ns	Matériel	>20k €/Go

Le matériel

- Accès à la mémoire
 - RAM
 - Chaque octet possède une adresse
 - Accès octet / octet ou mot / mot (1 mot = 32 ou 64 bits, selon le processeur)
 - Disque dur
 - Les octets sont groupés par blocs (« cluster ») de 512 – 4096 octets.
 - Ecriture d'un octet :
 - Lire le bloc en question, le charger en mémoire
 - Modifier l'octet dans la mémoire
 - Ecrire le bloc sur le disque



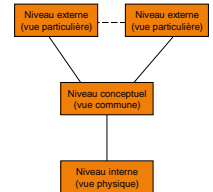
4. Architecture des BD

- a) Architecture Fonctionnelle
- b) Architecture opérationnelle

a) Architecture Fonctionnelle

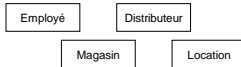
- 3 niveaux de description des données ont été définis par la norme ANSI/SPARC

- **Le niveau externe**
Niveau relatif aux utilisateurs
Manière dont ils voient les données
- **Le niveau conceptuel**
Niveau intermédiaire
Représentation « abstraite » de l'ensemble de la BD
- **Le niveau interne**
Niveau relatif à la mémoire physique
Manière dont les données sont réellement enregistrées



Le niveau externe

- Pour un utilisateur la vue externe est LA base de données
 - Exemple : le service comptabilité de « Top Vidéo »



- Chaque vue externe est décrite par un schéma (définition des types d'enregistrement)
- Exemple : l'enregistrement externe « Employé »

NumEmp (6 Char)	Nom (String)	Salaire (Réel)
--------------------	-----------------	-------------------



Le niveau conceptuel

- La vue conceptuelle est une vue abstraite
- Elle doit représenter un modèle des données de l'entreprise
- Défini par un schéma conceptuel
- Plus qu'un assemblage de vue externe contient plus de propriétés
 - Droits
 - Règles d'intégrité
 - Notion de flux de données (pas dans les SGBD actuels)



Le niveau interne

- La vue interne est une représentation de bas niveau de l'ensemble de la BD
- Suppose un espace d'adressage infini, dépend du système choisi
- Décrit par un schéma interne comprenant :
 - Les champs mémoires
 - Mais aussi les index



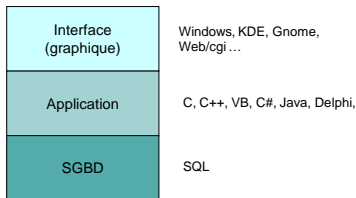
Les liens

- Il y a 2 liens qui définissent les correspondances entre les niveaux
- **Lien conceptuel / interne**
 - Correspondance entre les 2 couches (traduction)
 - Si on modifie le niveau interne il faut ajuster le lien pour que le niveau conceptuel soit inchangé.
- **Lien externe / conceptuel**
 - Correspondance entre la couche Externe et Conceptuelle
 - Possibilité de fusion de plusieurs données conceptuelles en un champs externe.
 - En cas de modification d'une couche il faut adapter le lien de correspondance



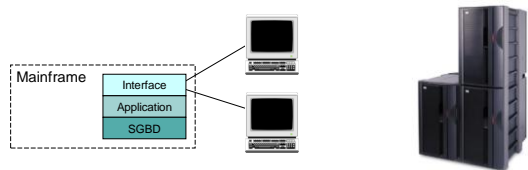
b) Architecture opérationnelle (Architecture globale du système)

- On peut voir le système en 3 couches :



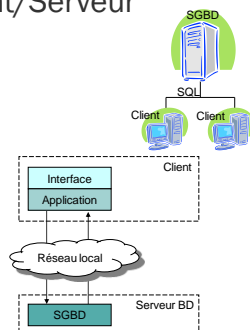
Architecture Mainframe (1960)

- Les 3 couches sont implémentées sur la même machine
- Systèmes propriétaires, non standardisés
- Terminaux passifs



Architecture Client/Serveur

- Apparue avec :
 - Serveur puissant
 - Ordinateur de bureau répandu au sein de l'entreprise
 - Réseaux locaux rapides
- Avantage
 - Technologie standard
 - Milieu hétérogène



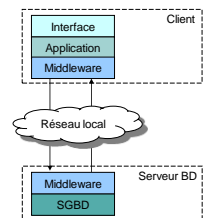
Architecture Client/Serveur 2 tiers

Avantages

- Architecture simple
- Données centralisées
- Interface utilisateur riche

Inconvénients

- Client lourd
- Logiciel spécifique au serveur
- Limité à l'entreprise



Exemple de middleware : ODBC

- Open Data Base Connectivity
 - Interface de programmation défini par Microsoft
 - Permet la communication des clients Windows et pratiquement tous les SGBD du marché



Architecture client/serveur 3 tiers

- Empilement de trois couches, étages, niveaux ou strates dont le rôle est clairement défini :

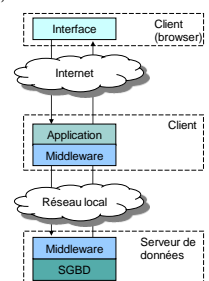
- la **présentation** des données
- le **traitement** métier des données
- l'**accès aux données** persistantes

Avantages

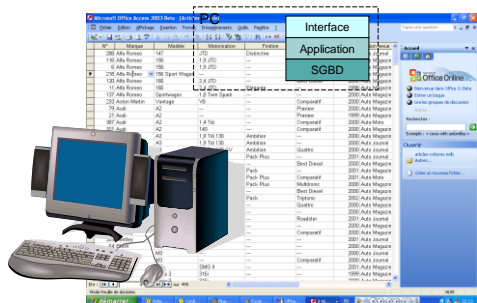
- Client léger (navigateur)
- Facilité de maintenance
- Spécialisation des serveurs

Inconvénients

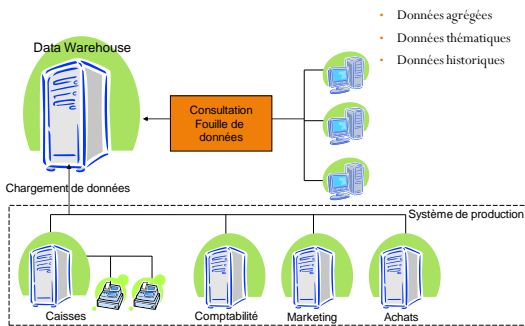
- Architecture plus complexe à mettre en place



Architecture BD personnelle



BD décisionnelles, entrepôt de données



Entrepôts de données et OLAP

- OLAP (OnLine Analytical Processing)
 - Architecture et outils d'analyse d'entrepôts de données
 - Analyse multidimensionnelle des données
- Exemple : les articles dans un super marché
 - Groupes d'articles (alimentaire / non-alimentaire)
 - Temps d'achat (jour, mois, année, période marketing)
 - Géographique (Magasin, département, région, pays)
- Nécessite :
 - Des performances exceptionnelles du SGBD
 - Des agrégations (redondance!)
 - Des outils de navigation et de visualisation puissants