

Bases de Données Avancées : Introduction

Thomas Gerald

September 13, 2024

Laboratoire Interdisciplinaire des Sciences du Numérique – LISN, CNRS

thomas.gerald@lisn.upsaclay.fr

Un système de gestion de bases de données ?

- Stockage des données
- Ajouter des données
- Rechercher des informations

Quel type (dans ce cours)

- Sur des données structurées
- Avec le modèle relationnel !!

id [PK] integer	title character varying	release_date character varying	budget bigint	revenue bigint	popularity double precision	runtime integer	rating double precision	original_language integer	belongs_to_collection integer
2	Ariel	1988-10-21	[null]	[null]	3.860491	69	3.2369531849577897	55	[null]
3	Varjoja paratlisissa	1986-10-16	[null]	[null]	2.29211	76	3.1755501064722202	55	[null]
5	Four Rooms	1995-12-09	4000000	4300000	9.026586	98	3.079564818455892	0	[null]
6	Judgment Night	1993-10-15	[null]	12136938	5.538671	110	3.841763756945689	0	[null]
11	Star Wars	1977-05-25	11000000	775398007	42.149697	121	3.6605905006418484	0	10
12	Finding Nemo	2003-05-30	94000000	940335536	25.497794	100	2.6721787709497207	0	137697
13	Forrest Gump	1994-07-06	55000000	677945399	48.307194	142	3.326441784548422	0	[null]
14	American Beauty	1999-09-15	15000000	356296601	20.726578	122	3.4308065226972233	0	[null]
15	Citizen Kane	1941-04-30	839727	23217674	15.811921	119	2.72704	0	[null]
16	Dancer in the Dark	2000-05-17	12800000	40031879	10.684806	140	3.7955110396256764	0	498
17	The Dark	2006-01-26	[null]	[null]	5.691508	87	3.9522270706405402	0	[null]

Objectifs

- Être capable d'interroger, créer et définir une base de données relationnelle (rappel ?)
- Savoir manipuler les expressions d'algèbre relationnelle (rappel ?)
- Connaître les algorithmes utilisés dans les SGBD
- Estimer le coût d'une requête, savoir optimiser les accès à un SGBD
- Connaître les solutions des SGBD pour préserver l'intégrité des données

Objectifs

- Être capable d'interroger, créer et définir une base de données relationnelle (rappel ?)
- Savoir manipuler les expressions d'algèbre relationnelle (rappel ?)
- Connaître les algorithmes utilisés dans les SGBD
- Estimer le coût d'une requête, savoir optimiser les accès à un SGBD
- Connaître les solutions des SGBD pour préserver l'intégrité des données

Faire votre propre SGBD ?

Objectifs

- Être capable d'interroger, créer et définir une base de données relationnelle (rappel ?)
- Savoir manipuler les expressions d'algèbre relationnelle (rappel ?)
- Connaître les algorithmes utilisés dans les SGBD
- Estimer le coût d'une requête, savoir optimiser les accès à un SGBD
- Connaître les solutions des SGBD pour préserver l'intégrité des données

Faire votre propre SGBD ?



En tous cas réaliser certaines briques nécessaires

Partie 1 : SQL (Structured Query Language)

- Le SQL c'est quoi ?
- La syntaxe SQL
- Introduction à l'algèbre relationnelle

Les différents chapitres/parties

Partie 1 : SQL (Structured Query Language)

- Le SQL c'est quoi ?
- La syntaxe SQL
- Introduction à l'algèbre relationnelle

Partie 2: Optimisation (algorithmique)

1. Représentation interne des données
2. Indexer l'information
3. Algèbre relationnelle
4. Implémentation des opérateurs
5. Plan d'exécution

Les différents chapitres/parties

Partie 1 : SQL (Structured Query Language)

- Le SQL c'est quoi ?
- La syntaxe SQL
- Introduction à l'algèbre relationnelle

Partie 2: Optimisation (algorithmique)

1. Représentation interne des données
2. Indexer l'information
3. Algèbre relationnelle
4. Implémentation des opérateurs
5. Plan d'exécution

Partie 3 : Concurrency (algorithmique)

1. SGBD multi-utilisateurs (sérialisation...)
2. Performances et parallélisation des calculs
3. Reprise sur panne (si assez de temps)

Heures de cours

- 20 heures de cours (CM) : 3 séances de 3h (dont Partiel) et 11 de 1h
- 22 heures de TD/TP : 2 * 11

Information et contenu

- Cours et TD : Ajout des cours après le CM (idem pour les TDs)
- Corrections : En fin de chapitre !!!
- Information : Sur la plateforme E-Campus (2024 - 2025 Bases de données avancées - MIAGe)

Partiel et Examen :

- Interrogations écrites (20%) : (au moins deux notées)
- Partiel (30%) : Sur la partie modèles relationnels et optimisation
- Examen (50%) : Sur la partie concurrence (mais aussi optimisation)

Introduction & Objectifs du cours

SGBD (Système de gestion de bases de données)

Il s'agit du système qui gère la base de données :

- Stockage des données
- Accès aux données
- Planification pour l'exécution des requêtes
- Gestion des utilisateurs
- ...

SGBD (Système de gestion de bases de données)

Il s'agit du système qui gère la base de données :

- Stockage des données
- Accès aux données
- Planification pour l'exécution des requêtes
- Gestion des utilisateurs
- ...

Différents types de données

- Des "tables" (postgres, mySQL)
- Des dictionnaires, json (MongoDB)
- Des graphes (Neo4j)
- Des vecteurs...

SGBD (Système de gestion de bases de données)

Il s'agit du système qui gère la base de données :

- Stockage des données
- Accès aux données
- Planification pour l'exécution des requêtes
- Gestion des utilisateurs
- ...

⚠ SQL et NoSQL

- **BD SQL** : Utilisation de requête SQL pour interagir avec la BD (modèle relationnel)
- **BD NoSQL** : Not Only SQL, utilisation de fonctionnalités sortant du cadre du modèle relationnel

Différents types de données

- Des "tables" (postgres, mySQL)
- Des dictionnaires, json (MongoDB)
- Des graphes (Neo4j)
- Des vecteurs...

En Licence : Initiation à la base de données

- Utilisation de la base de données avec le langage SQL (revue en première partie)
- Création d'une base de données, requêtes, vues, triggers...

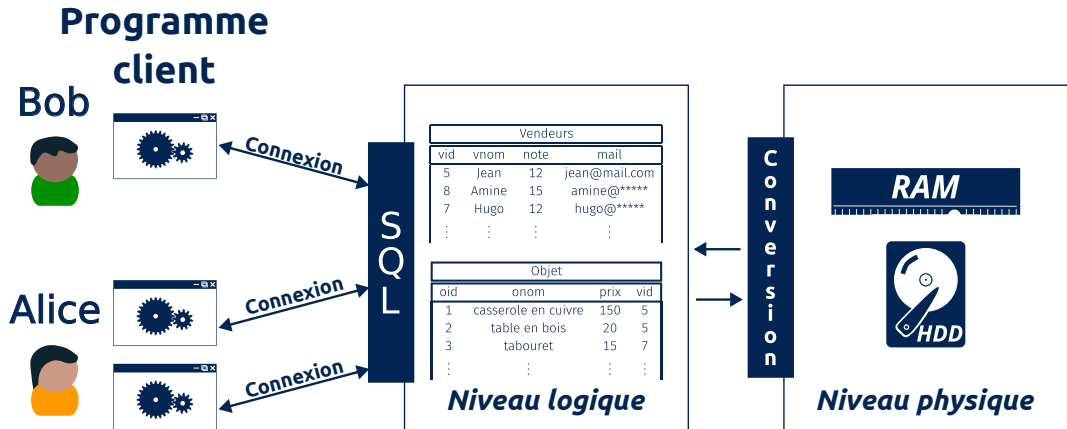
Dans ce cours

- Fonctionnement d'un SGBD
- Moteur de requêtes (optimisation)
- fiabilité
- Concurrency

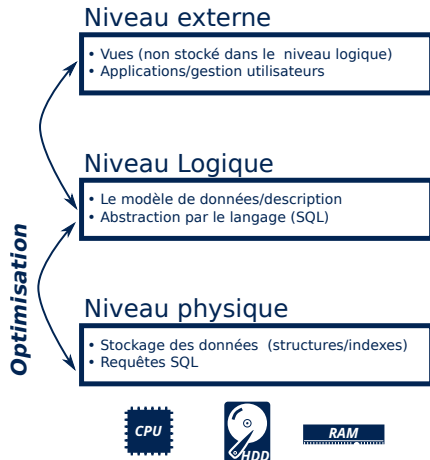
Objectifs d'un SGBD :

- Efficacité (**performances**) - retrouver, ajouter des informations rapidement
- Persistance (**persistence**) - stockage sur des supports persistants (disque dur)
- Fiabilité (**reliability**) - la reprise sur panne (coupure de courant)
- Simplicité d'utilisation (**convenience**) - abstraction pour interrogation
- Multi-utilisateurs/concurrence **concurency** - gestion des transactions simultanées
- **Sécurité** (à priori pas dans ce cours) - restriction d'accès

Un SGBD, c'est quoi ?



Les différents niveaux d'abstraction



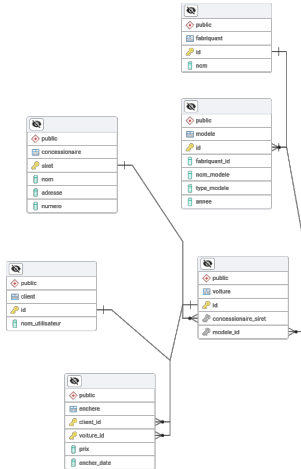
Les différents niveaux:

Correspondent à des niveaux d'abstraction

- Indépendant les uns des autres
- Hiérarchie → proximité avec le matériel

Stockage des données

Première partie : Le modèle relationnel



Objectifs

- Structurer, définir sa base de données
- Interroger une base de données avec du SQL
- Les requêtes avec l'algèbre relationnel
- etc...

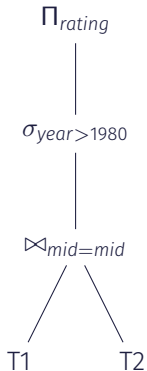
→ Créer, organiser, manipuler les bases de données relationnelles

```
SELECT *  
FROM Vendeurs V  
WHERE V.note = 12  
       OR V.note = 8
```

Objectifs

- Structurer définir sa base de données
- Interroger une base de données avec du SQL
- Les requêtes avec l'algèbre relationnel
- etc...

→ Créer, organiser, manipuler les bases de données relationnelles



Objectifs

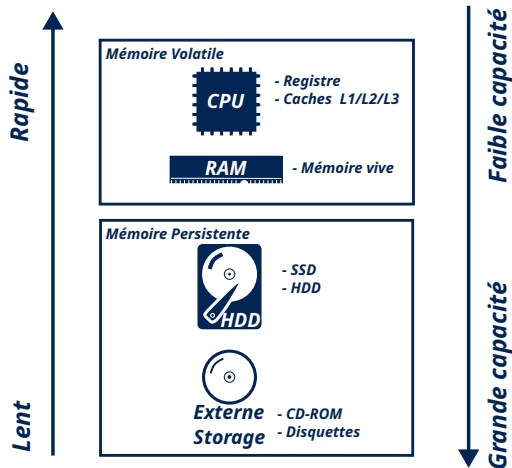
- Structurer définir sa base de données
- Interroger une base de données avec du SQL
- Les requêtes avec l'algèbre relationnel
- etc...

→ Créer, organiser, manipuler les bases de données relationnelles

Seconde partie : Optimisation

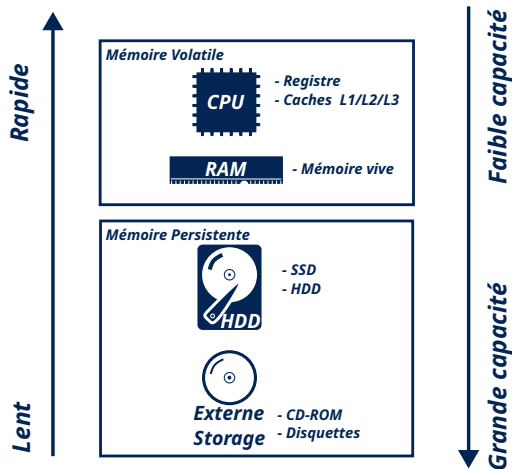
Hiérarchie des temps d'accès (approximatif)

1. Mémoire du processeur (cache),
 $\approx 370\text{Go/s}$ à 2300Go/s
2. Mémoire vive, 20Go/s à 60Go/s (débit théorique $\text{largeur_bus} \times \text{frequence}$)
3. HDD/SSD, 100Mo/s à 8Go/s (Différence importante si données séquentielles ou non)
4. CD-ROM, quelques Mo/s



Les capacité de la mémoire

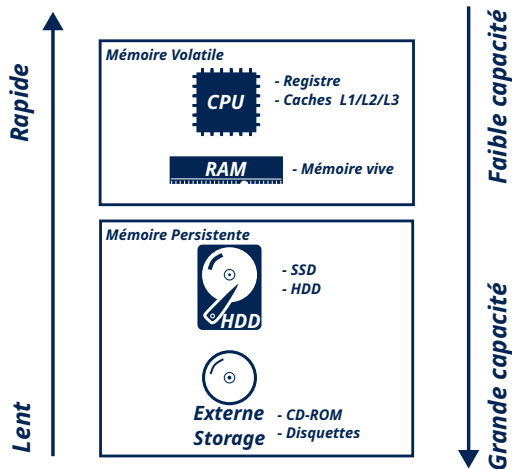
1. Mémoire du processeur (cache), quelques Ko à plusieurs dizaine de Mo
2. Mémoire vive, 4Go à 128Go
3. HDD/SSD, quelques Go à plusieurs To
4. CD-ROM/DVD/Blu-ray, 700Mo à 100Go



Les capacités de la mémoire

1. Mémoire du processeur (cache), quelques Ko à plusieurs dizaines de Mo
2. Mémoire vive, 4Go à 128Go
3. HDD/SSD, quelques Go à plusieurs To
4. CD-ROM/DVD/Blu-ray, 700Mo à 100Go

→ Minimiser les accès aux disques !!!



Requête

SELECT user FROM vendor WHERE



Optimisation de la requête
Plan d'exécution

Execution des opérateurs

Accès aux données/Index

Gestion du cache

Gestionnaire de disque

Objectifs

- Comment sont stockées les données ?
- Quelles sont les structures d'accès aux données ?
- Comment est sélectionné un plan d'exécution ?
- etc...

Soit deux enseignants *Alice* (A) et *Bob* (B), deux notes de cours C_1 et C_2 et M la moyenne.
Quand un enseignant modifie une note il doit mettre à jour la moyenne.
Que se passe t-il si les enseignants mettent à jour leur note de la manière suivante ?

1. A enlève 2 points à C_1
2. A lit la note C_2
3. B ajoute 4 points à C_2
4. B lit la note C_1
5. B met à jour la moyenne
6. A met à jour la moyenne

Prenons $C_1 = 14$ et $C_2 = 8$

Quelles sont les valeurs de C_1 , C_2 et M après l'exécution des instructions précédentes ?

Comment garantir le bon fonctionnement de requêtes exécutées en parallèles?

Comment garantir le bon fonctionnement de requêtes exécutées en parallèles?

Objectifs

- Les garanties d'un SGBD (ACID)
- Execution des transactions en parallèles
- Sécurité dans les SGBDs
- Intégrité des données
- etc...

Comment garantir le bon fonctionnement de requêtes exécutées en parallèles?

Objectifs

- Les garanties d'un SGBD (ACID)
- Execution des transactions en parallèles
- Sécurité dans les SGBDs
- Intégrité des données
- etc...

Si le temps !

Reprise sur panne (que faire si le système est débranché ?)

Quelques ressources pour ce cours

Ressources

- **Cours en ligne de Phillipe Rigaux** : Ce cours reprend les grands principes que nous verrons dans ce cours !!!

<http://sys.bdpedia.fr/intro.html>

- **Database Management Systems (Ramakrishnan, Gehrke)** : Livre en anglais contenant la majeure partie des concepts abordés dans le cours

<https://github.com/Premiumlab/Database-Management-System-Ramakrishnan-Gehrke-3rd-Edition>

- **Documentation postgresSQL** : Documentation et wiki du SGBD postgresSQL, voici quelques pages liés au cours :

- *Les pages* : <https://www.postgresql.org/docs/current/storage-page-layout.html>

- *Les index Btree* : https://en.wikibooks.org/wiki/PostgreSQL/Index_Btree

- *Les index par Hash* : https://en.wikibooks.org/wiki/PostgreSQL/Index_Hash

- **E-CAMPUS**

<https://ecampus.paris-saclay.fr/course/view.php?id=154859>