Bases de données documentaires et distribuées, http://b3d.bdpedia.fr

Le sujet, en bref

Documents, collections massives, stockage et calcul distribués

Représentation d'entités sous forme de **documents** pour la gestion de **collections** à très grande échelle dans des environnements distribués.

Quatre thèmes traités successivement :

- **modélisation** sous forme de **document**, structuré, semi-structuré, non structuré ;
- **recherche d'information** dans de très grandes collections ;
- **distribution du stockage** pour maintenir de très grands volumes de données ;
- **distribution des calculs**, notamment pour les méthodes analytiques.

Des mots-clés à comprendre (et à critiquer) : données structurées/non structurées, bases NoSQL, Cloud, BigData, moteurs de recherche, etc.

Point de départ : bases relationnelles

Les **bases relationnelles** = applications gérant des informations **structurées** et **régulières** : applications de "gestion", Web, mobiles.

- Une modélisation normalisée
- Un langage (SQL) très bien défini, normalisé lui aussi
- De très bonnes performances, obtenues **automatiquement**
- Une gestion robuste de la concurrence d'accès

Attention à bien apprécier ce qu'on gagne / perd en passant au "NoSQL" (on gagne marginalement, on perd beaucoup)

Révisions nécessaires ? Voir `http://sql.bdpedia.fr` et `http://sys.bdpedia.fr`

Plan du cours

Quatre parties, $\approx$ 3 semaines chacune, 1 jalon à la fin de chaque partie

- **Représentation de documents textuels** : les formats XML et JSON ; la modélisation de documents, les langages de manipulation
- **Recherche d'information dans les bases documentaires** : moteurs de recherche, index, algorithmes
- **Stockage distribué**. Systèmes distribués ; NoSQL
- **Calcul distribué**. MapReduce, et au-delà avec Spark et Flink

Connaissances pratiques (en cours, en TP, chez vous)

- Des systèmes "NoSQL" ; illustré avec MongoDB, CouchDB et Cassandra
- Des moteurs de recherche . avec ElasticSearch
- Systèmes distribués ; avec MongoDB, ElasticSearch et Cassandra.
- Calculs distribués ; avec Hadoop, Spark, Flink

Prérequis

- **Compréhension des bases relationnelles**, soit au moins la conception d'un schéma, SQL, ce qu'est un index et des notions de base sur les transactions.
  $\Rightarrow$ si nécessaire, voir `http://www.bdpedia.fr`
- **Une aisance minimale dans un environnement de développement**. Editer un fichier, lancer une commande, ne pas paniquer devant un nouvel outil, savoir résoudre un problème avec un minimum de tenacité, etc.
  $\Rightarrow$ il est utile (ais pas **obligatoire**) reproduire les exemples donnés.

Il ne s'agit pas **d'apprendre** des systèmes, mais de comprendre
des principes via la pratique (limitée)

Environnement matériel et logiciel

Pour la mise en pratique :

- au Cnam, tout vous est fourni ;
- Cnam à distance, ou en auditeur libre, il vous faut une machine dotée d'au moins 8 GO de mémoire (Linux, Mac/OS ou Windows).

Tous les logiciels utilisés sont libres de droits. Nous travaillons essentiellement avec

- ElasticSearch pour l'indexation de documents.
- MongoDB et Cassandra
- Spark et (un peu) Flink

L'installation de ces outils (et leur utilisation) peut se faire avec **Docker**, un gestionnaire de systèmes distribués virtuels.

Comment travailler

Le cours est découpé en **chapitres**, couvrant un sujet bien déterminé, et en **sessions**.

L'unité de travail est la session, correspondant à 1-2 heures de travail personnel

Pour assimiler une session vous pouvez combiner les ressources suivantes :

- **La lecture** du support en ligne, également disponible en PDF ou en ePub.
- **Le suivi du cours**, en vidéo ou en présentiel.
- **Le test des exemples de code** fournis dans chaque session.
- **La réalisation des exercices** proposés en fin de session.

Pendant les visios : reprise (rapide) des principes, et nous faisons les quiz.

Mener votre travail personnel

Vous devez maîtriser le contenu des sessions **dans l'ordre où elles sont proposées**.

Commencez par lire le support, jusqu'à ce que les principes vous paraissent clairs.

Suivez la visio, pour une synthèse et un média complémentaire

Reproduisez les exemples de code, **sans y passer trop de temps**

Cherchez à résoudre les problèmes par vous-mêmes au besoin : c'est le meilleur moyen de comprendre.
Finissez par les exercices. Faites les vous-mêmes, avant de regarder la solution.
Préparez les quiz proposés en fin des principaux chapitres.