NFE204 - Bases de données documentaires et distribuées

Mis à jour le 24-09-2024

Code : NFE204

Unité d'enseignement de type mixte

6 crédits

Volume horaire de référence (+/- 10%) : **50 heures**

Responsabilité nationale :
EPN05 - Informatique / 1

Contact national :

EPN05 - Informatique
292 rue saint Martin
33.1.13B
75003 Paris
01 40 27 22 64
Florian Gau
florian.gau@lecnam.net

Présentation

Prérequis

Prérequis: M1 ou niveau Bac+4 informatique

Bonnes connaissances en bases de données, architectures des systèmes informatiques, pratique de la programmation
Public: cycle d'ingénieur CNAM, Master M2

Objectifs pédagogiques

Le cours est consacré à le gestion de données massives, non-structurées ou semi-structurées. Le passage à l'échelle de très gros volumes (téraoctets, pétaoctets) peut amener a revoir la modélisation relationnelle qui implique des opérations de jointures assez coûteuses dans un environnement distribué. Cette modélisation est également inadaptée à des données comme les textes, les images, ou un assemblage de plusieurs médias. On s'oriente alors plutôt vers une modélisation sous forme de "documents" souvent dénués de structure connue (e., documents images, vidéos, documents Office, etc) ou d'une structure très souple (documents hypertextes).

Les notions de modèles de données et de langage d'interrogation sont alors à revoir. De plus le volume des données considérées implique la mise en place d'infrastructure à grande échelle typique des systèmes de gestion des données du Web.
Le cours couvre les sujets suivants:

Données peu structurées. Représentation de données complexes et/ou dotée d'une structure variable. Application à la représentation de documents textuels par des langages comme XML ou JSON. Notions essentielles sur la navigation dans une structure de document, le typage de documents, et la gestion de documents dans des bases de données.
Systèmes NoSQL. Des systèmes de gestion de données qui renoncent à certaines fonctionnalités fortes (transactions, langage d'interrogation) des bases relationnelles, au profit du passage à l'échelle, émergent à l'heure actuelle. Ces systèmes sont fortement orientés vers la distribution dans des environnements de type cloud, et leur conception varie selon l'objectif visé (accès temps réel, ou traitement analytiques). La structure des données reprend les principes vus dans la première partie du cours. Nous étudions les principes généraux des systèmes NoSQL, et en étudions certains: MongoDB, CouchDB, Cassandra, etc. Les problèmes de passage à l'échelle, de fiabilité, de sécurité, de reprise sur panne et de cohérence seront évoqués.

La **Recherche d'Information** (RI) consiste à effectuer des recherches sur des ensembles de données peu structurées, en effectuant un classement par pertinence. Avec l'avènement de gros moteurs d'indexation tels que *Google ou Amazon*, les technologies de recherche textuelle devient incontournable et donne un véritable intérêt à toutes ses techniques de stockage et d'index orienté texte.

Stockage distribué. Le volume des données manipulées par les moteurs de recherche, les sites de commerce électronique ou les sites communautaires rassemblant des millions d'utilisateurs, a atteint des niveaux inédits: le téraoctets est un ordre de grandeur courant, bientôt ce sera le pétaoctets. De nouvelles techniques de gestion de ces données massives ont émergé récemment, sous l'impulsion notamment des entreprises (Google, Amazon) directement confrontées aux problèmes liés à ces volumes inédits. L'exposé sera consacré à ces nouvelles techniques, en mettant l'accent sur les solutions s'appuyant sur la distribution du stockage et des traitements dans des parcs de machines extensibles appelés "**Cloud Computing**". Le cours

présente les principales problématiques et méthodes de stockage distribué: réplication, partitionnement, tolérance aux pannes, illustrées par quelques solutions-phares (ElasticSearch, Hadoop, Cassandra, etc).

Calcul distribué. Le stockage distribué est associé à des systèmes permettant de paralléliser les calculs pour traiter en temps raisonnable de très grandes masses de données, notamment à des fins analytiques. Le calcul parallèle à grande échelle est introduit et illustré avec des principes phares comme MapReduce, et des systèmes comme Spark, Hadoop et Flink.

Compétences

Compréhension des défis et des enjeux actuels dans la gestion de l'information, de plus en plus orientée vers l'acquisition et l'analyse de grandes masses de données. Maîtrise des techniques de base concernant ces nouvelles technologies. Systèmes NoSQL, techniques de distribution de données, techniques de recherche d'information.

Programme

Contenu

Modélisation de données peu structurées
- Documents structurés, JSON, XML
- Données web, Open data, services REST

- Bases documentaires: MongoDB, CouchDB, Cassandra

Recherche d'information
- introduction à la recherche textuelle dans les documents, indexation textuelle et Recherche d'Information (IE, Google, Amazon, ...)

- moteur de recherches: ElasticSearch, Solr

Systèmes de stockage distribués

- systèmes distribués, équilibrage, partitionnement, réplication

- cloud, performances, architectures, scalabilité
- illustration concrète avec quelques systèmes NoSQL: MongoDB, Cassandra, ElasticSearch

Systèmes de calcul distribué

- Le paradigme MapReduce

- Systèmes modernes de traitement à grande échelle: Spark, Flink

Modalités de validation

- Examen final

Description des modalités de validation

examen, projet, travaux pratiques

Bibliographie

Titre	Auteur(s)
Web Data Management, Cambdrige Publishing, 2012	Abiteboul, Manolescu, Rigaux, Rousset, Sennelart

Modern Information Retrieval, Addison-Wesely, 1999	R. Baeza-Yates, B. Ribeiro-Neto
Support en ligne: http://b3d.bdpedia.fr	P.Rigaux