le cnam
Nouvelle-Aquitaine

Fiche récapitulative

NFE204 | Bases de données documentaires et distribuées

51
Total d'heures d'enseignement

6
Crédits ECTS

Date non définie
Début des cours prévu

Programme

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

Objectifs : aptitudes et compétences

Objectifs :

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.

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

Délais d'accès

Le délai d'accès à la formation correspond à la durée entre votre inscription et la date du premier cours de votre formation.

- UE du 1er semestre et UE annuelle : inscription entre mai et octobre
- UE du 2e semestre : inscription de mai jusqu'à mi-mars

Exemple : Je m'inscris le 21 juin à FPG003 (Projet personnel et professionnel : auto-orientation pédagogique). Le premier cours a lieu le 21 octobre. Le délai d'accès est donc de 4 mois.

Planning

Légende:

Cours en présentiel

Cours 100% à distance

Mixte: cours en présentiel et à distance

Centre de formation	Prochaine session*	Modalité	Tarif individuel
100% à distance	2023/2024 : Date non définie		De 0 à 1.020 €

*Selon les UEs, il est possible de s'inscrire après le début des cours. Votre demande sera étudiée pour finaliser votre inscription.

Modalités

Modalités pédagogiques :

Pédagogie qui combine apports académiques, études de cas basées sur des pratiques professionnelles et expérience des élèves. Équipe pédagogique constituée pour partie de professionnels. Un espace numérique de formation (ENF) est utilisé tout au long du cursus.

Modalités de validation :

examen, projet, travaux pratiques

Tarif

Mon employeur finance	1.020 €
Pôle Emploi finance	510 €
Je finance avec le co-financement Région	Salarié : 156 €
Je finance avec le co-financement Région	Demandeur d'emploi : 124,80 €

Plusieurs dispositifs de financement sont possibles en fonction de votre statut et peuvent financer jusqu'à 100% de votre formation.

Salarié : Faites financer votre formation par votre employeur

Demandeur d'emploi : Faites financer votre formation par Pôle emploi

Votre formation est éligible au CPF ? Financez-la avec votre CPF

Si aucun dispositif de financement ne peut être mobilisé, nous proposons à l'élève une prise en charge partielle de la Région Nouvelle-Aquitaine avec un reste à charge. Ce reste à charge correspond au tarif réduit et est à destination des salariés ou demandeurs d'emploi.

Pour plus de renseignements, consultez la page Financer mon projet formationopen_in_new ou contactez nos conseillers pour vous accompagner pas à pas dans vos démarches.

Passerelles : lien entre certifications

- CS5900A - Certificat de spécialisation Analyste de données massives

Avis des auditeurs

Les dernières réponses à l'enquête d'appréciation de cet enseignement :

↓ Fiche synthétique au format PDF

Taux de réussite

Les dernières informations concernant le taux de réussite des unités d'enseignement composant les diplômes

↓ Taux de réussite