le cnam
Pays de la Loire

RÉGION FORMATION accélérateur d'emploi
Région PAYS DE LA LOIRE

Possibilité d'obtention par la VAE AU CNAM

Certificat de spécialisation Analyste de données massives Analyste Big Data / *Data scientist*

Crédits : 27 ECTS - code diplôme : CS5900A

Formation à distance

Public concerné et conditions d'accès

Informaticiens, mathématiciens ou statisticiens ayant un niveau ingénieur ou master (par exemple après un diplôme d'ingénieur en informatique CYC9101A ou CYC9105A, ou le Master MEDAS) et exerçant en entreprise. Formation supérieure en mathématique (algèbre linéaire, analyse). Connaissances en bases de données, en programmation, en statistique et analyse des données.

Métiers et débouchés

Ce certificat offre la possibilité à des informaticiens, mathématiciens ou statisticiens de suivre une formation professionnelle pluridisciplinaire pour acquérir les compétences propres à l'exercice du métier émergent de *data scientist* également appelé «analyste big data».

Compétences visées

Alliant des compétences en mathématiques, statistique, informatique, visualisation de données, le *data scientist* / analyste *big data* est capable de stocker, rechercher, capter, partager, interroger et donner du sens à d'énormes volumes de données structurées et non structurées, produites en temps réel et provenant de sources diverses.

Conditions de délivrance du certificat

Obtenir une note supérieure ou égale à 10 à toutes les UE proposées ainsi qu'au projet professionnel.

Calendrier

L'année est organisée en 2 semestres : semestre 1 (S1) d'octobre à février/mars et semestre 2 (S2) de février/mars à juin.

- **Parcours diplômant**

Le cursus est proposé selon une programmation permettant d'optimiser la durée de la formation, compatible avec une activité professionnelle.

- **Unités d'enseignement « à la carte »**

Vous avez toute liberté pour effectuer votre choix parmi l'ensemble des unités d'enseignement (UE) qui vous sont proposées.

Consultez les plannings des UE proposées par le **www.cnam-paysdelaloire.fr** *rubrique Inscriptions. D'autres UE proposées à distance sont disponibles sur le réseau Cnam. Renseignez-vous auprès de nous.*

Les cours

- **cours à distance via Internet :** autoformation avec accompagnement par un enseignant(e) (en individuel ou collectif). Utilisation de supports numériques (documents pdf, documents sonorisés, vidéos interactives, quiz d'autoévaluation…) et échanges en classes virtuelles par visioconférence (en direct ou en différé), messagerie, forums, chat…

Les tarifs

Ils sont consultables sur **www.cnam-paysdelaloire.fr** rubrique Inscriptions.

Contacts

Angers • 02 41 66 10 66 • angers@cnam-paysdelaloire.fr
Cholet • 02 41 66 05 26 • cholet@cnam-paysdelaloire.fr
La Roche/Yon • 02 51 44 98 28 • laroche@cnam-paysdelaloire.fr
Laval • 02 43 26 22 37 • laval@cnam-paysdelaloire.fr
Le Mans • 02 43 43 31 30 • lemans@cnam-paysdelaloire.fr
Nantes • 02 40 16 10 95 • nantes@cnam-paysdelaloire.fr
Saint-Nazaire • 02 40 90 50 00 • saint-nazaire@cnam-paysdelaloire.fr

Programme

STA211	**Entreposage et fouille de données**	9 CR
NFE204	**Bases de données documentaires et distribuées**	6 CR
RCP216	**Ingénierie de la fouille et de la visualisation de données massives**	6 CR
UASB03	**Projet certificat analyste de données massives**	6 CR

Les unités d'enseignement (UE) correspondent à des crédits européens : 4, 6 ou 8 crédits. 1 crédit correspond à environ 10h d'apprentissage : cours magistral, exercices dirigés, travail sur projet etc.
(CR : crédits)

STA211 Entreposage et fouille de données

Pré-requis : NFA008 et connaissances en analyse des données et méthodes descriptives (sinon suivre STA101).
Modèles prévisionnels et systèmes de gestion de l'entreprise : structures spécifiques des bases de données de Data warehouse (star schema) - OLAP
Méthodologies générales : Méthodologies de Data Mining
Pré-traitement des données : Analyses de la qualité des données - Techniques d'appréhension des valeurs manquantes ou aberrantes - Techniques de construction de bases de travail (agrégations, etc.)
Données et techniques de fouille : *Méthodes non supervisées :* Cartes de Kohonen, Règles d'association – *Méthodes supervisées :* Rappels de théorie de l'apprentissage, Arbres de décision, forêts aléatoires, SVM, Réseaux de neurones, deep learning – *Méta-algorithmes :* boosting, bagging – *Fouille dans de nouveaux types de données et méthodes associées :* Données textuelles - Données multivues - Images et Multimedia.
Outils : Environnements freeware : Weka, Tanagra, R, Python - Outils spécifiques : SAS-EM, SPAD - Data Mining et bases de données : OLAP Business Object.

NFE204 Bases de données documentaires et distribuées

Prérequis: Bonnes connaissances en bases de données, architectures des systèmes informatiques, pratique de la programmation.
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.

RCP216 Ingénierie de la fouille et de la visualisation de données massives

Pré-requis: NFE204 et STA211
Vous êtes encouragés à évaluer votre capacité à suivre cette UE en répondant au questionnaire en ligne accessible sur http://cedric.cnam.fr/vertigo/Cours/RCP216/preambule.html
Introduction : applications, typologie des données, typologie des problèmes
- Approches : réduction de la complexité, distribution
- Passage à l'échelle de quelques problèmes fréquents : Recherche par similarité, systèmes de recommandation - Classification automatique - Fouille de données textuelles - Fouille de flux de données - Apprentissage supervisé à large échelle - Fouille de graphes et réseaux sociaux - Visualisation d'information : historique, applications, outils - Enjeux perceptifs de la visualisation d'information : couleurs, formes, immersion, lecture - Techniques de représentations : graphes, hiérarchies, lignes de temps - Techniques d'interaction : association focus/contexte, distorsion, filtrage.

Le cours est complété par des TP permettant de mettre en pratique des techniques présentées. Pour la partie fouille de données, les TP seront réalisés à l'aide de Apache Spark. Pour le travail sur le projet, l'auditeur devra installer le logiciel Spark (gratuit) sur un ordinateur personnel de capacité suffisante, suivant les instructions disponibles en ligne.

UASB03 Projet certificat analyste de données massives

Pré-requis : avoir validé les UE STA211, NFE204 et RCP216
Le projet consiste à choisir et traiter un sujet d'analyse de données présentant potentiellement une problématique de passage à l'échelle. Ce sujet devra s'appuyer sur un jeu de données, disponible soit dans le cadre professionnel, soit par une source n'imposant pas de limitation en termes de droits d'utilisation. Le choix du jeu de données est validé par l'accord du responsable de l'UA. Les sujets proposés devront être différents de ceux ayant déjà fait l'objet d'évaluation dans les autres UE du certificat.
Le travail devra couvrir les aspects suivants :
1. Choix d'un système de stockage passant à l'échelle par distribution (base relationnelle distribuée, système « NoSQL », moteur de recherche, etc.)
   - chargement du jeu de données choisi dans ce système,
   - étude des connecteurs avec les autres composants logiciels du projet (R, Spark...) ;
2. Analyses exploratoires, prétraitement, études préalables (normalisation, nettoyage des données, gestion de données manquantes, agrégation...) ;
3. Choix d'une méthode d'analyse adéquate et mise en œuvre au moins en partie avec Spark ;
4. Une partie visualisation si elle est utile à la compréhension de l'analyse et des résultats.

Le rapport (25 à 30 pages maximum pour la partie rédigée) devra inclure :
- une introduction, contextualisant le problème étudié, présentant de façon claire et précise l'objectif de l'étude ainsi que les données utilisées,
- un développement, présentant la démarche choisie, les résultats obtenus et leur interprétation,
- une conclusion, synthétisant l'apport de l'étude vis-à-vis du problème étudié.

Il n'est pas imposé d'effectuer des expérimentations en vraie grandeur sur des données massives stockées dans un système distribué, même si une telle expérimentation est bien entendu bienvenue si vous en avez l'opportunité et les moyens. En revanche, le rapport doit inclure une étude argumentée de la scalabilité de la solution analytique envisagée. Cette étude propose typiquement une architecture globale articulant le système de stockage, des composants analytiques, et l'intégration de ces composants dans une plate-forme de calcul distribuée passant à l'échelle et couplée au système de stockage. En d'autres termes, il s'agit d'une mise en œuvre des compétences acquises respectivement en NFE204, STA211 et RCP216. La capacité de cette architecture à soutenir une forte croissance de la volumétrie des données par un ajout, en proportion, de ressources de stockage et de calcul, sans dégradation de la performance globale, doit être justifiée. Les parties importantes du code ou de ses dérivés (graphiques ou aides à l'interprétation) devront être incorporées au texte. Les programmes/scripts/codes utilisés seront placés dans une archive zip.

Service communication Cnam Pays de la Loire – 24 mai 2019