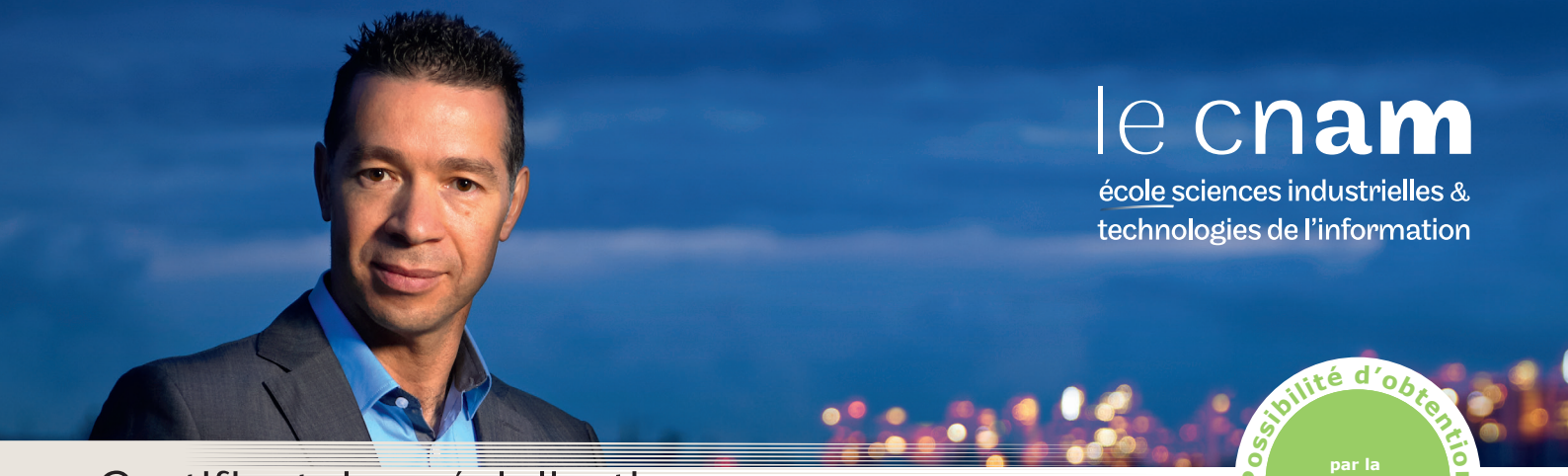

le cnam
école sciences industrielles & technologies de l'information

Possibilité d'obtention par la VAE AU CNAM

Certificat de spécialisation
analyste de données massives

Certificat de spécialisation

Domaines : statistique et informatique

Crédit : 27 ECTS — Code : CS59p1

Niveau d'entrée : bac+5 — Niveau de sortie : bac+5/6

Public concerné et conditions d'accès

Informaticiens, mathématiciens, ou statisticiens ayant un niveau ingénieurs ou master et exerçant en entreprise. Formation supérieure en mathématique (algèbre linéaire, analyse). Connaissances élémentaires en base de données et en statistique linéaire.
L'inscription à l'UA de projet se fait après agrément.

Objectifs pédagogiques

Ce certificat offre la possibilités à des informaticiens, mathématiciens, statisticiens de suivre une formation professionnelle pluridisciplinaire pour acquérir les compétences propres à l'exercice du métier émergent de datascientifique également appellé «analyste big data».
Alliant des compétences en mathématiques, statistique, informatique, visualisation de données ; il 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.

Compétences visées

Donner du sens à d'énormes volumes de données structurées et non structurées, produites en temps réel et provenant de sources diverses.
Maitriser les technologies Hadoop et MadReduce, de passage à l'échelle et le traitement de données d'un nouveau type (textes images, vidéos, etc...) à l'aide de méthodes de data mining et d'apprentissage.

Stage, projet, mémoire

L'auditeur doit réaliser un projet en fin de cycle choisi au préalable avec un enseignant et portant sur un sujet en accord avec son activité professionnelle.
Ce projet donnera lieu à un rapport écrit ainsi qu'une soutenance orale.

Conditions de délivrance du diplôme

Le certificat de spécialisation s'acquiert en obtenant une note supérieure à 10 à toutes les UE proposées ainsi qu'au projet professionnel.

Pré-requis : bac+5			Semestres 1 et 2
STA211 ☾	Entreposage et fouille de données	9 ECTS	Annuel
NFE204 ☾	Bases de données documentaires et distribuées	6 ECTS	Semestre 1
RCP216 ☾	Ingénierie de la fouille et de la visualisation de données massives	6 ECTS	Semestre 2
UASB03	Projet	6 ECTS	Semestre 2

@ Cours disponible en ligne (Île-de-France)

@ Cours disponible en ligne (hors Île-de-France)

☾ Cours du soir (HTT)

ECTS : système européen de transfert et d'accumulation de crédits.

SITI/2013
CS59p1

Conservatoire national des arts et métiers

Préconisation de parcours

Semestre 1

STA211
Entreposage et fouille de données

Crédits : 9 ECTS

Contenu de la formation :
Cette UE est divisée en 5 axes :
-Modèles prévisionnels et systèmes de gestion de l'entreprise. (structures spécifiques des bases de données de Data warehouse)
-Méthodologies générales. (Normes de Data Mining : PMML (Prediction Modelling Markup Language), JDM (Java Data Mining), SQL-MM (multimedia) Méthodologies de Data Mining : SEMMA (SAS), CRISP-DM).
-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, etc.).
-Données et techniques de fouille. (Méthodes non supervisées : Cartes de Kohonen. Règles d'association et séquence mining, etc. Méthodes supervisées : Rappels de théorie de l'apprentissage. Arbres de décision. Réseaux de neurones. Méta-algorithmes : boosting, bagging.
Fouille dans de nouveaux types de données et méthodes associées : Données textuelles. Images et Multimedia. Données symboliques. Réseaux sociaux).
-Outils :
Environnements freeware : Weka, Tanagra, Sodas. Outils spécifiques : SAS® Enterprise Miner, SAP-Analytics, SPAD , DataLab. Data mining et bases de données : OLAP business object.

NFE204
Bases de données documentaires et distribuées

Crédits : 6 ECTS

Contenu de la formation :
Ce cours est consacré à la gestion efficaces de grandes masses de données faiblement structurées, notamment les données documentaires, données multimédia, séries temporelles et autres sources d'information difficilement gérables avec des SGBD relationnels classiques. Ces données se trouvent sur le WEB ou dans les systèmes d'information d'entreprises (boîtes mails, logs des serveurs, collections de documents, etc.), avec une problématique forte de passage à l'échelle pour offrir des temps de recherche très rapide même en présence d'un volume de données très important.
Le cours présente la modélisation et l'interrogation de données semi-structurées (bases XML, bases JSON), l'indexation plein-texte et multimédia, et les techniques de passage à l'échelle par distribution, actuellement en pleine émergence sous l'influence des systèmes proposés par des grands acteurs du WEB comme Google (BigTable), Amazon (Dynamo), Yahoo! (Hadoop), etc.
- Introduction aux données semi-structurées problématique de passage à l'échelle, présentation de quelques systèmes récents (Hadoop, CouchDB, memcached).
- Modélisation et interrogation : XML et ses environnements (XPath, XQuery, XSLT) ; bases de données XML (eXist), bases de données JSon (Couch DB, mongo DB).
- Moteurs de recherche : aquisition de données (crawlers), index plein tête, indexation mutlimédia.
- Passage à l'échelle par distribution : hachage consistant, arbres B distribués, tables de hachage distribuées.
- Techniques d'évaluation parallèles et distribuées : MapReduce.

Semestre 2

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

Crédits : 6 ECTS

Contenu de la formation :
1. Introduction : applications, typologie des données, typologie des problèmes.
2. Approches : réduction de la complexité, distribution.
3. Passage à l'échelle de quelques problèmes fréquents.
a. Recherche par similarité.
b. Recherche par modèle.
c. Classification automatique, jointure par similarité.
d. Fouille de textes.
e. Fouille de graphes (hyper-documents, réseaux sociaux).
4. Visualisation d'information : historique, applications, outils.
5. Enjeux perceptifs de la visualisation d'information : couleurs, formes, immersion, lecture.
6. Techniques de représentations : graphes, hiérarchies, lignes de temps.
7. Techniques d'interaction : associations focus/contexte, distortion, filtrage.
Le cours est complété par des travaux pratiques (TP) permettant la mise en pratique des techniques présentées. Pour la fouille de données, les TP seront réalisées à l'aide de R et Mahout. Pour la partie visualisation, les TP seront effectués avec le logiciel Processing ; une séance d'introduction est réservée à son apprentissage.

UASB03
Projet

Crédits : 6 ECTS

Le projet consistera à mettre en oeuvre une méthode d'analyse particulière avec des techniques présentées dans les unités d'enseignement. Il sera réalisé dans le cadre de l'activité professionnelle de l'élève avec les environnements informatiques auxquels il/elle est susceptible d'être confronté(e). Le travail à faire inclut :
- l'exploitation d'un jeu de données ;
- le choix d'une méthode analytique applicable à ce jeu de données ;
- le choix d'un environnement de stockage et d'exécution d'algorithmes de fouille de données ;
- l'interprétation des résultats.
Les méthodes pour gérer des données dépassant la capacité de mémoire seront également mises en oeuvres car il s'agit d'un problème récurrent pour les data scientists. Les outils à disposition sont par exemple disponibles pour le logiciel R (librairies ff et big memory.
Si le cadre de l'activité professionnelle ne permet pas (ou pas totalement) d'accéder à un environnement complet, il sera possible de recourir à des jeux de données publics et d'utiliser un environnement de travail fourni par le Cnam (Hadoop, Mahout, systèmes NoSQL, etc.).
Le sujet du projet sera défini en commun avec un enseignant et validé par ce dernier sur la base des indications qui précèdent. Les auditeurs mènent ensuite un travail personnel, contrôlé périodiquement par l'enseignant de référence , et débouchant sur un exposé oral et un rapport écrit sur lesquels se base la validation de l'UA.