LE CERTIFICAT « ANALYSTE DE DONNÉES MASSIVES » DU CONSERVATOIRE NATIONAL DES ARTS ET MÉTIERS

Ndèye NIANG, Gilbert SAPORTA, Michel CRUCIANU et Philippe RIGAUX [1]

TITLE

Big Data Analysis : the CNAM Specialization Certificate

RÉSUMÉ

Dans cet article nous présentons le certificat de spécialisation « Analyste de données massives » du Conservatoire National des Arts et Métiers (CNAM). Après une présentation du CNAM et ses missions, le contenu pédagogique de la formation, le public auquel elle s'adresse et ses spécificités qui la distinguent des enseignements classiques de statistique et d'informatique sont présentés. Enfin, deux brefs témoignages sur l'apport de cette formation aux professionnels des entreprises sont joints.

Mots-clés : *données massives, formation continue, apprentissage numérique, Big Data, NoSQL, visualisation.*

ABSTRACT

We present the CNAM certificate dedicated to Big Data Analysis. The paper shortly presents the CNAM and its missions, the content of the training units, the public it is aimed for and how its specificities distinguish it from traditional courses in statistics and computer science. We conclude with two brief testimonies on how this certificate matches the expectations of companies confronted to Big data issues.

Keywords: *big data, lifelong learning, machine learning, NoSQL systems, visualization.*

1 La formation continue de *Data Scientists*

La nécessité de développer en France une filière de formation de *data scientists* a été fortement affirmée par différents acteurs tant politiques, industriels, qu'universitaires depuis quelques années. Pour reprendre les termes du rapport Hermelin et Bourdoncle (2014), « la « science des données » (*data science*) est la science qui sous-tend l'ensemble des technologies du « *Big Data* », un domaine qui va susciter une demande rapidement importante de profils connus sous le nom de « *data scientist* », à même d'appréhender à la fois les méthodes statistiques, mathématiques et informatiques dans le cadre d'un contexte métier particulier. Le système éducatif français doit donc s'adapter en conséquence ».

« La formation de *data scientist* est considérée comme une des principales actions de nature à développer l'écosystème *Big Data* en France parmi les mesures visant à renforcer l'offre de

[1]CEDRIC Conservatoire National des Arts et Métiers, ndeye.niang_keita@cnam.fr, gilbert.saporta@cnam.fr, Michel.Crucianu@cnam.fr, Philippe.Rigaux@cnam.fr

technologies et/ou services *Big Data*. Elle doit permettre de s'assurer que le système éducatif français forme suffisamment de « *data scientist* », profil qui va être au cœur de la nouvelle révolution industrielle que représente le numérique. Les formations devront être à la fois des formations courtes, de type formation continue ou professionnalisante, des mastères spécialisés, et des formations d'ingénieur de niveau bac +5 plus classiques. » (voir Abiteboul *et al.* (2014)).

L'objectif de former rapidement des milliers de *data scientists* ne peut être atteint uniquement par les filières de formation initiale, d'où l'importance de la formation continue.

Le certificat de spécialisation « Analyste de données massives » du CNAM a été créé en 2014-2015 pour répondre à la demande en formation des personnels en place dans les entreprises. Il s'adresse essentiellement à des informaticiens, mathématiciens, statisticiens, ingénieurs engagés dans la vie professionnelle qui souhaitent évoluer ou se reconvertir dans le nouveau domaine du *Big Data* ou données massives. Ce certificat s'inscrit donc parfaitement dans la lignée des rapports précédemment cités et dans les missions du conservatoire. Après une brève présentation du CNAM, nous détaillons le public auquel s'adresse le certificat, son contenu pédagogique, ses modalités d'enseignement, les moyens mis à disposition. Enfin, deux témoignages d'élèves en formation sont joints à l'article avant la conclusion et les perspectives.

2 Le Conservatoire National des Arts et Métiers

Comme l'indique le site web[2] de présentation du CNAM : « Le CNAM est le seul établissement d'enseignement supérieur français dédié à la formation des adultes, placé sous la tutelle du ministère en charge de l'Enseignement supérieur. Il donne à chacun les moyens de se former, à tout moment de sa vie, sur place au CNAM, dans son entreprise ou à distance. Le Conservatoire a été créé par la Convention en 1794 sur proposition de l'abbé Henri Grégoire « pour perfectionner l'industrie nationale ». Le CNAM concourt à la diffusion de l'innovation technologique et des savoirs scientifiques ainsi qu'à la promotion de l'esprit de création. Il est aujourd'hui doté du statut de grand établissement à caractère scientifique, culturel et professionnel. »

Le CNAM pilote un réseau de 29 centres régionaux et de 158 centres d'enseignement, dont le siège est à Paris. S'y ajoutent trois centres à l'étranger (Côte d'Ivoire, Liban et Maroc). Le CNAM offre des formations développées en étroite collaboration avec les entreprises et les organisations professionnelles afin de répondre au mieux à leurs besoins et à ceux de leurs salariés. Les trois missions principales du CNAM sont :

- La formation tout au long de la vie, s'adressant à des auditeurs dont 50% sont des femmes, pour la plupart en activité professionnelle, en alternance ou en reconversion et dont l'âge moyen est 34 ans. Le CNAM propose aussi la validation par acquis de l'expérience (VAE).
- La recherche : depuis la création du Conservatoire en 1794 et de ses premiers laboratoires de recherche en 1852, le CNAM s'est toujours attaché à développer une interaction soutenue entre formation et recherche. Aujourd'hui, le CNAM possède 21 laboratoires de recherche et une école doctorale (ED) de site (ED Abbé Grégoire), 3 ED co-accréditées (Edite, Spiga, SMI) et 6 ED en partenariat et 7 projets d'excellence (Labex). Les statisticiens du CNAM sont regroupés dans l'équipe MSDMA (Méthodes Statistiques de *Data*

[2] `http://presentation.cnam.fr/`

Mining et Apprentissage) du laboratoire CEDRIC (Centre d'Études et de Recherche en Informatique et Communications).

- La diffusion de la culture scientifique et technique à travers, en particulier, le Musée des arts et métiers avec 250 000 visiteurs par an et 80 000 objets, une bibliothèque avec 160 000 volumes, 40 000 lecteurs ainsi que 300 événements et conférences.

3 Le certificat de spécialisation « analyste de données massives »

S'inscrivant dans la ligne des missions du CNAM, ce diplôme d'établissement s'adresse essentiellement à des personnes engagées dans la vie professionnelle qui souhaitent évoluer, ou se reconvertir dans le nouveau domaine des données massives. Il leur offre la possibilité de suivre une formation professionnelle pluridisciplinaire pour acquérir les compétences propres à l'exercice du métier émergent de *data scientist* orienté spécifiquement vers les données massives. Cette formation vient en complément du master de statistique du CNAM qui existe depuis 2005 et formait déjà des *data scientists*, avant que cette dénomination ne soit en vogue, mais sans la spécificité des données massives. Alliant des compétences en mathématiques, statistique, informatique, visualisation de données, l'analyste de données massives doit être 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.

3.1 Le public et les conditions d'inscription et d'étude

Le certificat est accessible à des informaticiens, mathématiciens ou statisticiens ayant un niveau ingénieur ou master et exerçant en entreprise. Au cours de ces deux années d'existence, nous avons enregistré un plus grand nombre de candidats au profil plutôt informatique. Ceci peut s'expliquer par le fait que ce profil est plus concrètement touché par la révolution des données ; ce sont ces informaticiens qui ont en charge la gestion des données (stockage, structuration, extraction…), les statisticiens intervenant ensuite pour leur analyse selon le schéma traditionnel.

Il n'y a pas d'inscription spécifique au certificat ; les candidats doivent demander un agrément pour suivre les trois unités d'enseignement (UE) composant le certificat. Il s'agit des UE : Entreposage et fouille de données (code STA211, voir section 3.3.1), Bases de données documentaires et distribuées (code NFE204, voir section 3.3.2) et Ingénierie de la fouille et de la visualisation de données massives (code RCP216, voir section 3.3.3). Cette autorisation d'inscription leur est accordée après une vérification des prérequis correspondant à une formation supérieure en mathématique, précisément en algèbre linéaire et analyse, en statistique et analyse des données, ainsi qu'à des connaissances en bases de données et en programmation.

Les inscriptions ont lieu en septembre et en février pour les UE semestrielles. Il est recommandé aux élèves de suivre les UE dans un ordre respectant les contenus pédagogiques détaillés ci-dessous en section 3.3.

Les effectifs par unité d'enseignement varient entre environ 80 et 100 élèves en présentiel répartis sur les deux semestres ou pour un semestre en formation à distance.

Les enseignements sont dispensés hors temps de travail (HTT), en cours du soir ou en enseignement à distance (uniquement pour NFE204 pour le moment). Les supports de cours dispensés en présentiel sont mis à disposition des élèves sur une plateforme d'enseignement ou sur le site web dédié à chaque UE. Cela permet d'offrir une formule « hybride » souple adaptée aux élèves engagés professionnellement qui peuvent facilement « rattraper » un cours qu'ils auraient manqué suite à des contraintes professionnelles.

La spécificité du hors temps de travail a un impact direct sur l'organisation de la formation ; il est ainsi difficile pour une grande partie des élèves de réaliser le certificat en un an. En effet, le volume horaire disponible en HTT ne permet pas de couvrir en une année l'ensemble des unités d'enseignement, vues la pluridisciplinarité de la formation, la richesse du contenu pédagogique et la part importante de travail personnel.

Conformément à la vocation sociale du CNAM, le coût de la formation supporté par un élève en inscription individuelle est faible : au droit forfaitaire d'inscription annuelle de 150 € s'ajoutent 14 € par crédit soit 378 € pour les 27 crédits du certificat (tarif 2015-2016).

3.2 Moyens humains et matériels

Les enseignants-chercheurs permanents du CNAM et d'autres universités (Paris 13 par exemple) ont la majeure charge des enseignements, mais nous faisons appel à de nombreux intervenants issus du monde professionnel conformément aux usages du CNAM. Ainsi, une partie des enseignements de l'UE STA211 est réalisée en collaboration avec des professionnels d'entreprises telles que AXA, WASABI ANALYTICS, CISCO, SAS, SPAD-COHERIS qui viennent exposer aux élèves leur expérience de la mise en œuvre dans leur travail quotidien de quelques-unes des méthodes et techniques enseignées ou non.

Le CNAM dispose de salles dédiées aux travaux pratiques équipées d'ordinateurs de bureau, avec 2 To (teraoctects) de disque, utilisés comme serveurs nécessaires pour stocker des données issues de sites tels que *dumps wikimedia* ou d'autres données libres (*open data*). Sur ces machines sont installés les logiciels nécessaires à la mise en œuvre des méthodes étudiées dans les différentes UE. Ces logiciels sont pour la plupart gratuits (R, Weka, Spark…) ou rentrent dans le cadre de licences ou conventions en partenariat avec le CNAM (SAS, SPAD…).

Pour garantir une certaine qualité dans l'encadrement des TP, en particulier pour les UE NFE204 et RCP216, deux sessions sont proposées dans l'année permettant ainsi de suivre chaque semestre universitaire une quarantaine d'élèves répartis en deux groupes.

3.3 Le contenu pédagogique

Le certificat est composé de 3 unités d'enseignement et d'un projet professionnel de synthèse (unité d'activité ou UA) décrits dans les sections suivantes. Le certificat de spécialisation s'acquiert par capitalisation en obtenant une note supérieure ou égale à 10 à toutes les UE proposées ainsi qu'au projet professionnel. L'évaluation des UE se fait selon des modalités qui peuvent différer selon les UE : examen, projet de programmation, analyse d'un jeu de données, exposés, suivi de travaux pratiques.

3.3.1 UE STA211 : Entreposage et fouille de données (9 ECTS)

UE d'entrée du certificat, elle dispense des enseignements sur les méthodes et techniques d'apprentissage statistique et numérique permettant le traitement multidimensionnel de données de grandes dimensions. L'UE débute par un rappel sur les concepts fondamentaux de la statistique et de l'analyse des données indispensable pour brièvement mettre à niveau les élèves de provenance variée. Ensuite, le programme couvre les différents aspects du *data mining* ou fouille de données.

Après une présentation de la méthodologie générale de la fouille de données, les concepts et méthodes théoriques pour la partie statistique sont abordés. Les cartes de Kohonen et règles d'association sont présentées pour les méthodes non supervisées. Ensuite, après un rappel de théorie de l'apprentissage, les arbres de décision, forêts aléatoires et réseaux de neurones sont étudiés pour ce qui concerne les méthodes supervisées.

Une partie importante de l'UE est consacrée à la fouille dans de nouveaux types de données et méthodes associées : données en blocs, données évolutives, données images et multimédia, données textuelles, données symboliques, réseaux sociaux et données du web. Les quatre derniers types de données sont présentés par des intervenants extérieurs pour environ 20% du volume horaire de l'UE (60 heures).

Les aspects informatiques de la fouille de données sont abordés avec la gestion des bases de données, en particulier la construction de *data warehouse* ou entrepôt de données.

Nous abordons également la phase de pré-traitement des données avec des méthodes d'analyse de la qualité des données, les techniques d'appréhension des valeurs manquantes ou aberrantes, de construction de bases de travail (agrégations, etc.).

Finalement, l'agrégation de modèles à travers les méta-algorithmes tels que le *bagging* et le *boosting* est aussi étudiée.

Deux types d'outils informatiques sont employés : des environnements *freeware* ou libres comme R, Weka, Tanagra, et des outils statistiques spécifiques comme SAS-EM, SPAD, ainsi que BO pour la partie bases de données OLAP.

C'est une UE à 9 crédits avec une grande part de travail personnel pour revoir les méthodes de base et surtout leur mise en œuvre, car beaucoup d'auditeurs ont eu des cours souvent théoriques mais pas de connaissances pratiques de l'analyse des données : analyse factorielle, régression linéaire multiple, le traitement de la multicolinéarité (régression PLS, Ridge), classification automatique…

En outre, certains des élèves du CNAM peuvent avoir besoin d'une mise à niveau liée à l'évolution des méthodes et outils qu'ils ont eu à utiliser lors de leur formation initiale ou à l'apparition de nouvelles méthodologies et algorithmes. Chaque méthode du cours est illustrée avec des outils spécifiques (SAS, SPAD, …) ; des packages R sont aussi indiqués. Les élèves doivent se les approprier de façon autonome.

Le volume horaire est de 60 heures en présentiel dont 10 heures consacrées aux professionnels, 30 heures de cours et 20 heures de travaux pratiques (TP) et travaux dirigés (TD). Le temps consacré par les élèves à cette UE en dehors des heures de présence est estimé à environ 30 heures et varie selon leur profil préalable de formation.

3.3.2 UE NFE204 : Bases de données documentaires et distribuées (6 ECTS)

L'objectif de cette unité d'enseignement est la compréhension des défis liés à la croissance des volumes de données à traiter, et des caractéristiques communes aux outils récents qui proposent des solutions à ces défis, plus ou moins assimilables aux systèmes dits « NoSQL » (*Not only SQL*).

Le déroulé général du cours présente des modèles de données alternatifs au modèle relationnel (rassemblés par simplicité sous le terme de « modèles documentaires »), les méthodes de recherche pour des collections de tels documents, et le passage à l'échelle de très grands volumes par distribution des données et des traitements.

Les élèves acquièrent à l'issue de ce cours un panorama des nouvelles technologies liées à cette évolution : systèmes NoSQL, systèmes distribués, recherche d'information.

Cette unité d'enseignement est ainsi consacrée en grande partie à la gestion de données documentaires, non-structurées ou semi-structurées. L'essentiel de ces données est accessible sous forme de « documents » souvent dénués de structure connue (documents images, vidéos, documents Office, etc.) ou d'une structure très souple (documents XML, JSON). De plus, le volume des données considérées implique la mise en place d'infrastructures à grande échelle. De nouveaux systèmes de gestion de données renonçant à certaines fonctionnalités fortes (transactions, langage d'interrogation) des bases relationnelles, au profit du passage à l'échelle, sont fortement orientés vers la distribution dans des environnements de type « *cloud* », et leur conception varie selon l'objectif visé (accès temps réel, ou traitements analytiques). Après une étude des principes généraux des systèmes NoSQL, quelques systèmes sont examinés : MongoDB, CouchDB, Cassandra pour les bases distribuées ; ElasticSearch et Solr pour les moteurs de recherche ; Hadoop, Spark, et Flink, les moteurs de calculs distribués, etc.

Le volume horaire est de 60 heures en présentiel réparties en 30 heures de cours et 30 heures de TP. Pour la formation à distance, le temps consacré en particulier au projet d'évaluation de l'UE est estimé à 20 heures environ mais peut varier en fonction des élèves.

3.3.3 UE RCP216 : Ingénierie de la fouille et de la visualisation de données massives (6 ECTS)

Cet enseignement s'intéresse à l'impact des caractéristiques des données massives (volume, variété, vélocité) sur les méthodes de fouille de données. Sont examinées les approches actuelles qui permettent de faire passer à l'échelle les méthodes de fouille, en insistant sur les spécificités des opérations de fouille en environnement distribué. Les caractéristiques mentionnées sont ensuite considérées de façon plus spécifique pour certains problèmes fréquents dans le traitement des données massives.

Sont ainsi abordés les systèmes de recommandation et la recherche efficace par similarité, la classification automatique et l'apprentissage supervisé sur une plate-forme distribuée, les opérations spécifiques au traitement des données textuelles souvent hétérogènes, les implications de la vélocité sur la fouille de flux de données, l'analyse de grands graphes et de réseaux sociaux. L'UE s'intéresse ensuite au rôle de la visualisation et de l'interaction, non seulement dans la présentation des résultats mais aussi dans les opérations de fouille de données.

Le cours est complété par des travaux pratiques (TP) permettant de mettre en œuvre de façon directe des techniques présentées. Pour la partie fouille de données, les TP sont réalisés avec Apache Spark. Le logiciel Processing est employé dans la partie visualisation.

Le volume horaire est de 60 heures en présentiel réparties en 30 heures de cours et 30 heures de TP. Comme pour les autres UE, le temps de travail consacré à cet enseignement varie en fonction des élèves.

UE finale du certificat, elle nécessite de bonnes connaissances mathématiques et statistiques générales et une maîtrise de méthodes statistiques pour la fouille de données étudiées dans l'UE STA211. Cette UE exige par ailleurs une connaissance de techniques de gestion de données massives faiblement structurées et de techniques de passage à l'échelle par distribution (UE NFE204), ainsi que la maîtrise d'au moins un langage de programmation et la capacité à utiliser le système d'exploitation linux.

3.3.4 UA : projet final (6 ECTS)

Les élèves doivent réaliser un projet en fin de cycle, choisi au préalable avec un enseignant au début du deuxième semestre. Ce projet donne lieu à un rapport écrit d'un trentaine de pages hors annexes, ainsi qu'une soutenance orale.

Le projet consiste à mettre en œuvre une méthode d'analyse particulière avec des techniques présentées dans les unités d'enseignement. Il peut être réalisé dans le cadre de l'activité professionnelle avec les environnements informatiques auxquels l'élève est susceptible d'être confronté. 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 et l'interprétation des résultats. Si le cadre de l'activité professionnelle ne permet pas (ou pas totalement) d'accéder à un environnement complet, il est possible de recourir à des jeux de données publics, et de les analyser dans un contexte informatique fourni par le CNAM.

L'UA est co-encadrée par les enseignants responsables des 3 UE de cours, qui interviennent respectivement pour vérifier et valider les aspects étudiés au cours du projet : choix des données, choix du système de stockage et des méthodes d'accès, choix de la méthode analytique et mise en œuvre, et enfin évaluation de la capacité à passer à l'échelle.

Le temps consacré par les élèves à cette UA dépend de leur parcours de formation préalable, du sujet, de leur situation professionnelle au moment de la réalisation de l'UA, etc. Les cours prennent fin au début du mois de juin ; la soutenance prévue à la fin du mois de septembre laisse ainsi aux élèves une période assez longue pour la réalisation du projet final. En pratique, ils peuvent commencer à préparer le projet à tout moment de la formation (par exemple les élèves sollicitent les enseignants pendant, ou plus souvent après, la présentation des cours en lien avec le sujet qu'ils souhaitent traiter).

4 Deux témoignages

Nathalie RAMOS, actuaire, 25 ans d'expérience dans le monde des assurances :

J'ai suivi la première année (2014-2015) de formation au Certificat d'analyste en données massives du CNAM. Cette formation m'a aidée à préparer mon certificat de Data Science pour l'Actuariat (2015-2016 première année également), et de progresser, très rapidement, vers un poste de responsable Data Science dans un cabinet d'actuaires.
Mon objectif est de faire évoluer mon métier d'actuaire vers de nouveaux postes et de nouveaux métiers que vont engendrer les techniques offertes par l'innovation, le digital et les traitements des données de masse. Le certificat du CNAM a pu m'ouvrir les portes à ces perspectives.
Cette formation, en somme, m'a permis de rediriger ma carrière vers de nouvelles aventures professionnelles dans les domaines qui me passionnent : l'informatique, la statistique et l'actuariat.

Aurélie LE CAIN, R&D Program Manager chez Essilor :

Après une thèse en mathématiques appliquées spécialité statistique, j'ai intégré en 2012 un pôle de mathématiciens en R&D chez Essilor, pour développer des algorithmes et maintenir des applications de calcul. Je travaillais essentiellement avec mes collègues mathématiciens.
En 2015, j'ai suivi à titre individuel la formation Big Data du CNAM dans le but d'accompagner Essilor dans l'essor du numérique. Cette formation, bien qu'intense à raisons de trois soirs par semaine, a été très stimulante dans mon métier.
L'ouverture que m'a apportée cette formation, notamment la découverte de nouveaux concepts, m'a permis de proposer de nouvelles idées. En particulier, l'une d'elles s'est concrétisée par l'ouverture d'un projet que je gère, impliquant cinq personnes de métiers différents.
Par ailleurs, j'ai intégré une équipe dans un projet Big Data dans lequel je suis responsable de l'analyse des données, me menant à me former à de nouveaux outils et nouveaux langages.
Cette formation pointue et complète m'a ainsi permis de concilier mon intérêt pour les sujets transverses, pour le travail en équipe et ma curiosité.

5 Conclusions et perspectives

Le certificat « analyste de données massives » positionné à bac+5 s'adresse à un public avec déjà un certain nombre d'acquis, en particulier professionnels, conformément à la mission de formation tout au long de la vie du CNAM. Il offre une formation pluridisciplinaire avec un contenu pédagogique riche alliant les enseignements théoriques et pratiques et un contact avec les entreprises exerçant dans les métiers des données massives. Sur sa première année d'existence, nous ne disposons pas encore d'évaluations formelles mais les témoignages très positifs d'élèves ainsi que la forte augmentation du nombre de candidats et d'inscrits montrent que le certificat est en phase avec la demande.

Nous envisageons la possibilité d'ouvrir le certificat sous forme de stages intensifs en journée, en accord avec les employeurs, sur une durée donc plus courte. Cela permet de répondre à la réalité du temps de l'entreprise qui n'est pas toujours le même que celui de l'enseignement. Nous étudions également la possibilité d'une formation en alternance accessible au niveau

bac+3.

Enfin, le certificat est destiné à être déployé dans les autres centres du réseau du CNAM en France et à l'étranger avec des enseignants locaux et le partage de ressources à distance.

Références

[1] Abiteboul S., F. Bancilhon, F. Bourdoncle, S. Clémencon, C. de la Higuera, G. Saporta et F. Soulie-Fogelman (2014), L'émergence d'une nouvelle filière de formation : « data scientists », *http://www.abiteboul.com/DOCS/14.DataScience.pdf*.

[2] Hermelin, P. et F. Bourdoncle (2014), La feuille de route Big data, *http://www.economie.gouv.fr/big-data-feuille-route-en-action/*.