UN DU D'ANALYSTE BIG DATA EN FORMATION CONTINUE COURTE, AU NIVEAU L3

Jean-Michel POGGI[1], Charles BOUVEYRON[2], Georges HEBRAIL[3] et François-Xavier JOLLOIS[4]

TITLE

A diploma of university (DU) "Big Data Analyst" in lifelong training, at level L3

RESUME

Nous présentons le diplôme d'université (DU) Analyste Big Data, délivré depuis cette année par le département STID de l'IUT de l'université Paris Descartes. D'un volume global de 150h, réservé aux apprenants en formation continue courte, au niveau L3, il constitue une voie de diplomation originale dans ce domaine émergent. Constitué de 5 modules, le DU est articulé autour de deux modules plutôt dédiés aux méthodes informatiques, deux plutôt statistiques qui font la part belle aux données de type « open data » et à la fouille des réseaux sociaux, et un dernier module dédié aux enjeux cruciaux concernant la qualité et la confidentialité des données. Il s'agit d'orienter fortement vers la mise en œuvre des outils liés à ce sujet émergent. Ainsi plus d'une moitié des intervenants sont issus du monde économique et industriel, en collaboration avec une équipe académique mélangeant statisticiens et informaticiens.

Mots-clés : *Big Data, formation continue, licence.*

ABSTRACT

We present the diploma of university (DU) "Big Data Analyst", starting this year and delivered by the STID department of IUT Paris Descartes. This 150-class-hour diploma is available for learners in lifelong training with at least an undergraduate level (L3 in France). It introduces an innovative way to certify essential skills in the emergent domain of Big Data. The diploma contains 5 modules. It is organized in two modules dedicated to computing methods, two models focused on statistical techniques, which give a good place to open data and social network analysis, and one module concerns with the crucial stakes of data quality and privacy. Another originality of this diploma is the strong incorporation of implementation tools, such that at least half of the teachers come from industry.

Keywords*: Big Data, lifelong training, bachelor.*

1. Introduction et positionnement de la formation

L'actualité est riche d'articles, d'analyses et de projets gouvernementaux sur le phénomène « Big Data » qui constitue le défi majeur en statistique et informatique décisionnelle des prochaines années. Les données accumulées dans les systèmes d'information sont un capital qu'il faut chercher à valoriser en leur appliquant différents traitements informatiques au sein desquels les méthodes statistiques jouent un rôle central.

[1] LMO, Univ. Paris-Sud Orsay et Univ. Paris Descartes, jean-michel.poggi@parisdescartes.fr
[2] MAP5, Univ. Paris Descartes, charles.bouveyron@parisdescartes.fr
[3] EDF R&D, georges.hebrail@edf.fr
[4] LIPADE, Univ. Paris Descartes, francois-xavier.jollois@parisdescartes.fr

Cette évolution significative des métiers de la statistique et de l'informatique décisionnelle nous a conduit à mettre en place, à l'IUT Paris Descartes, une formation courte et diplômante (diplôme d'Université).

Elle prolonge une évolution des formations en IUT débutée en 2001 avec la Licence Professionnelle « Décision et Traitement de l'Information – Data Mining », puis actée par l'émergence du décisionnel dans l'acronyme STID. De plus, elle ouvre le département à la formation continue courte.

Aussi, même si de très nombreuses filières universitaires et des cursus de formation continue des grandes écoles apparaissent avec le mot-clé « Big Data » dans l'intitulé ou au moins dans l'un des modules, bien peu sont d'un volume global de 150h, réservé aux apprenants en formation continue courte, au niveau L3. Ce diplôme constitue ainsi une voie de diplomation originale dans ce domaine émergent.

2. Objectifs et apprenants

En reprenant l'analyse du cabinet Gartner[5], les grands défis à surmonter dans le « Big Data » sont les « 3V » :

- la prise en compte de données très volumineuses (Volume) ;
- le traitement de données arrivant sous forme de flux continus en temps réel (Vélocité) ;
- l'hétérogénéité des provenances et des types des sources de données (Variété).

On trouve parfois dans les publications un quatrième V (Véracité) qui s'attaque à la qualité des données et même un cinquième relatif à la Valeur des données comme un capital à faire fructifier.

Il en résulte aujourd'hui le développement de nouveaux outils répondant à ces défis, aussi bien au niveau du stockage des données que de leur traitement statistique, dans une perspective décisionnelle.

Ce Diplôme Universitaire propose un complément de formation aux nouveaux concepts et outils pour le Big Data, pour des professionnels ayant une formation de base en statistique et informatique décisionnelle (bases de données, statistique, fouille de données). Le DU permettra aux apprenants d'évoluer vers des postes au sein de projets Big Data dans les entreprises, les administrations et les collectivités territoriales, ainsi qu'accompagner au niveau technique ces entités dans les évolutions liées à la révolution digitale.

Ce diplôme d'Université s'adresse à des **salariés ou à des adultes en formation continue ou en reprise d'études**, souhaitant valider et compléter des acquis professionnels dans le domaine du traitement de l'information en plan de formation, en congé individuel de formation ou en autofinancement.

Pour pouvoir candidater, il faut avoir un niveau équivalent Bac+2 avec des compétences en statistique et informatique décisionnelle. La procédure d'admission comprend une sélection sur dossier (admissibilité), puis éventuellement un entretien d'admission.

[5] http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

3. Intervenants

Les quatre auteurs de cet article se partagent la coordination pédagogique du DU, sous la responsabilité du premier auteur.

Il ne s'agit pas ici de donner une liste nominative des intervenants (disponible sur le site du diplôme) mais simplement de mentionner les équilibres disciplinaires et le caractère à la fois académique et professionnel de la formation.

On trouve en effet, en commençant par les académiques : 2 Professeurs (26e section du CNU[6]) et 1 Maître de conférences (27e section du CNU[7]) de l'Université Paris Descartes qui portent le diplôme, ainsi que 2 Maîtres de conférences (26e et 27e) de l'Université Lyon 2 et une Maîtresse de Conférences (27e) de l'Université de Tours. Du côté des professionnels, l'équipe compte 1 Chercheur Senior de EDF R&D qui copilote la formation, 3 Ingénieurs-chercheurs d'EDF[8] R&D, 1 Chargé de recherche de l'Ifsttar[9], 2 Conseils Indépendants en management et protection des données personnelles, un Directeur Data Intelligence de Capgemini[10] Consulting, le Président et un des responsables de Celeonet[11] et un Architecte de réseaux de OCTO Technology[12].

4. Organisation et vue d'ensemble de la formation

Pour le rendre compatible avec une activité professionnelle et pour permettre la nécessaire maturation des acquis, il se déroule sur **6 mois**, de janvier à juin**,** à raison de **5 modules** de 4 jours chacun (1 jour correspond à 7h de formation), précédé d'une demi-journée de présentation et suivi d'un séminaire final de synthèse d'une journée.

L'ensemble totalise donc 150 heures de formation. De niveau **Licence**, il s'agit d'orienter fortement vers la mise en œuvre des outils liés à ce sujet émergent. Chaque module alterne **enseignements théoriques, travaux dirigés et travaux pratiques** sur des outils du marché Hadoop, NoSQL et R.

Le DU se déroule avec un rythme de **2 jours de cours (jeudi et vendredi) toutes les 2 semaines (c'est-à-dire qu'une semaine avec cours est suivie d'une semaine n'en comportant pas)**, afin d'en faciliter le suivi par des professionnels en entreprise. Chaque module se déroule ainsi sur 4 semaines et s'appuie sur des outils support et des partenariats avec des professionnels du secteur. Dans chaque module sont aussi présentées des applications d'un ou plusieurs domaines parmi lesquels les télécommunications, la finance, l'énergie, les réseaux sociaux, la relation-client et le marketing.

Les frais de formation (hors droits universitaires) s'élèvent à 3 000 euros.

Constituée de 5 modules, elle est articulée autour de deux modules plutôt dédiés aux méthodes informatiques (les modules 1 et 2), deux plutôt statistiques (les modules 3 et 4) et un

[6] Section Mathématiques appliquées (incluant la statistique) du conseil national des universités (CNU).
[7] Section Informatique du conseil national des universités (CNU).
[8] EDF : https://www.edf.fr/
[9] IFSTTAR : http://www.ifsttar.fr/
[10] Capgemini : https://www.fr.capgemini.com/
[11] Celeonet : http://www.celeonet.fr/
[12] OCTO Technology : http://www.octo.com/fr

dédié à la qualité et la confidentialité des données (le module 5). La Figure 1 en présente l'architecture et le contenu.

MODULE 1 - STOCKAGE MASSIF DE DONNEES
- Bases de données relationnelles VLDB
- Bases de données mémoire
- Stockage dans le cloud
- Stockage dans l'éco-système Hadoop
- Environnement de type « data lab » (data mining agile) et outils de fouille de données passant à l'échelle

MODULE 2 - DECISIONNEL TEMPS REEL
- Sources de données en flux : capteurs, salles de marché…
- Complex Event Processing (CEP) : notion de flux de données, requêtes sur flux, fenêtrage …
- Tableaux de bord dynamiques, alarmes
- Méthodes de fouille de flux de données : modèles incrémentaux et adaptatifs (clustering, classification, prévision)
- Visualisation de flux de données

MODULE 3 - FOUILLE DE DONNEES COMPLEXES
- Sources de données complexes : séries temporelles, données spatiales, texte
- Pré-traitements des données complexes (transformations, indexation, signatures…)
- Méthodes de fouille de séries temporelles et fonctionnelles, de données spatiales, de textes (clustering, classification)
- Visualisation de données complexes
- Applications

MODULE 4- FOUILLE DU WEB ET DES RESEAUX SOCIAUX
- Sources et formats de données du web
- Extraction de données à partir du web
- Structures de grands graphes, propriétés statistiques de ces graphes
- Algorithmes de fouille des graphes des réseaux sociaux (statique, dynamique)
- Visualisation de graphes

MODULE 5 - QUALITE, SECURITE ET CONFIDENTIALITE DES DONNEES
- Gestion de la qualité des données : concepts et mise en œuvre
- Outils pour la qualité des données : Master Data Management, audit de bases de données
- Sécurisation de l'accès aux données
- Gestion des données personnelles (CNIL), anonymisation des données personnelles

FIGURE 1 – *Présentation synthétique du DU d'Analyste Big Data en formation continue courte, au niveau L3*

Dans le paragraphe suivant, nous complétons notre présentation du DU en précisant la stratégie pédagogique mise en place et permettant de décliner le programme.

5. Programme détaillé et stratégie pédagogique

Fournissons quelques détails sur le contenu des modules, en commençant par les modules plutôt dédiés aux méthodes informatiques et celui dédié à la qualité et la confidentialité des données, et en passant ensuite aux deux modules plutôt dédiés aux méthodes statistiques.

5.1 Les modules à dominante informatique et le module centré sur les données

Le module 1 dédié au stockage massif de données (voir Chang *et al.* (2008) et Dean & Ghemawat (2008)) propose en premier lieu de faire le point sur les outils nouveaux dédiés au Big Data (fichiers distribués, bases de données réparties, cloud computing, NoSQL, Hadoop...). Ensuite, il comprend une mise en situation sur la base d'un cloud par apprenant, hé-

bergé par notre partenaire Celeonet. Dans cette partie, la plus importante du module (sur deux jours), les apprenants vont évaluer et comparer une solution classique (MySQL, 2015) avec une solution nouvelle (Hadoop, 2015), sur le même jeu de données. Enfin, la dernière journée est dédiée à une introduction au logiciel R et surtout à l'interrogation d'une base NoSQL (MongoDB, 2015).

Le module 2 passe en revue sources et modèles de données en flux (capteurs, salles de marché…), introduit les concepts, les modèles de données et les langages de requêtes pour les données arrivant sous la forme de flux continu (Complex Event Processing - CEP). Sur le plan des méthodes statistiques, le cours montre comment les méthodes classiques doivent être revisitées pour être rendues incrémentales et adaptatives. Sont traitées les méthodes exploratoires (ex. : clustering) ainsi que les méthodes décisionnelles (ex. : arbres de décision, naive bayes). La notion de « concept drift » ainsi que la méthodologie d'évaluation des méthodes décisionnelles dans un cadre de flux de données sont également abordées. On pourra se reporter à Gama (2010), Aggarwal (2007) et Luckham (2002). Le module comporte des mises en œuvre des concepts du cours à l'aide d'un logiciel de type « CEP ».

Le module 5 est dédié à la qualité, la sécurité et la confidentialité des données. En effet, le « big data » s'appliquant à des données qui n'ont pas été collectées initialement pour être analysées, la maîtrise de leur qualité est un enjeu crucial (Berti-Equille, 2012). Un deuxième enjeu important est la sécurité des données et le respect des contraintes juridiques liées à la protection des données personnelles (Desgens-Pasanau, 2013). Des spécialistes et professionnels de ces domaines apportent leur concours pour exposer tant les fondements théoriques que les solutions du marché permettant d'apporter des réponses concrètes dans les applications. Des études de cas sont travaillées avec les apprenants.

5.2 Le module « Fouille de données complexes » (géographiques, textuelles ou temporelles)

Le module 3 débute par une journée d'études de cas en fouille de données complexes dans des secteurs économiques divers pour exhiber des métiers et des problèmes où la variété des données est manifeste. Puis le module consacre un jour à chaque type de données : géographiques, textuelles (Ibekwe-Sanjuan, 2007) et temporelles (Hyndman & Athanasopoulos, 2014). Les outils de description, d'analyse et de visualisation sont présentés au travers de nombreux exemples (dans le domaine du géomarketing, de la presse en ligne au travers d'un journal, ses blogs, forums et chats ou encore de la consommation électrique, respectivement) en s'attachant aux spécificités du type de données concerné et en veillant à la prise en main effective d'outils logiciels utilisant le langage de programmation R.

5.3 Le module « Fouille du web et des réseaux sociaux »

Le module 4 est dédié aux données libres d'accès, appelée communément « open data », et aux données de types réseaux (réseaux sociaux, réseaux de communications, réseaux de gènes…).

Ces dix dernières années ont été riches en développements autour de ces deux types de données aujourd'hui omniprésentes qui présentent des difficultés intrinsèques d'analyse.

Deux journées sont tout d'abord consacrées aux données de type réseau. Un premier objectif est d'apprendre à reconstruire et visualiser un réseau à partir d'un graphe ou de données transactionnelles. La méthode statistique « latent space model » (LSM) permet notamment d'accomplir cette tâche (Hoff *et al.*, 2010). Le modèle LSM suppose que la probabilité d'observer un lien entre deux nœuds est inversement proportionnelle à la distance des nœuds dans un espace latent, qui est donc à estimer. Le second objectif de ces deux journées est de pouvoir segmenter (classifier) les nœuds d'un réseau. La méthode du « Stochastic Block Model » (SBM) est présentée pour ce faire (Nowick & Snijders, 2001). Le modèle SBM suppose que la probabilité de lien entre deux nœuds dépend uniquement des groupes latents des nœuds considérés.

Les deux autres journées de ce module sont consacrées aux « open data » et leur visualisation. Il est notamment question de l'importation des données aux formats JSON et XML. La visualisation de ce type de données avec des outils web tels que D3JS est également abordée. L'ensemble de ce module est illustré avec le logiciel statistique R et ses paquets d'interaction avec les outils web.

Plusieurs jeux de données réelles sont utilisés pour mettre en œuvre les techniques vues en cours. A titre d'exemple, les méthodes d'analyse de réseaux sont appliquées au jeu de données des communications emails de la compagnie Enron[13].

5.4 Le choix de R

Aujourd'hui, nous assistons à une révolution du « Big Data » qui est train de s'opérer. Les solutions logicielles sont donc en forte évolution, et la mise en œuvre de la manipulation et de la fouille de données se fait sous la forme de l'écriture de programmes ou de scripts (ex. : Python, Scala, Java, R…). Le choix de R nous a semblé un bon compromis car il commence à être très utilisé dans le monde industriel, il est ouvert et évolutif et permet d'accéder à des méthodes récentes ; enfin, il existe des versions passant à l'échelle sur de gros volumes de données (ex. : Revolution Analytics).

Ce choix ne s'est pas fait sans beaucoup d'hésitations dans la phase de conception du diplôme, avant de faire le choix de R (R Development Core Team, 2012) mais outre les arguments évoqués ci-dessus, notre position est confortée par le fait qu'il existe désormais de nombreuses connexions entre R et les outils dédiés au traitement de données massives (Hadoop, MongoDB, Spark...).

6. Les deux premières promotions

Même s'il est encore trop tôt pour faire un bilan, il faut noter que malgré une publicité tardive et limitée, le DU n'ayant été adopté par les instances de l'université qu'en juin 2014, la promotion 2015 comptait 15 apprenants.

Sa composition frappe par sa diversité tant en termes de formation initiale, pour moitié à dominante informatique et pour moitié à dominante statistique, qu'en termes de branches et métiers concernés.

[13] Les données originelles sont disponibles à l'adresse http://www-2.cs.cmu.edu/~enron.

L'analyse des retours d'expérience à la fin de la session 2015 montre une grande satisfaction quant à la richesse du contenu, une stimulation réelle et un appétit à la fois pour les nouvelles techniques, ce qui était attendu, mais aussi pour les nouveaux enjeux concernant la qualité et la confidentialité des données. De surcroît, les cours de chacun des modules ont été l'occasion d'échanges réels permettant de concrétiser les concepts présentés, en les confrontant souvent aux différents contextes spécifiques des branches et des métiers des apprenants.

La composition de la promotion d'apprenants est variée de nombreux points de vue :

- le niveau de leur formation initiale : 1 DUT puis 6 LP Info ou STID, 1 Licence Eco-gestion, 3 Maitrises (dont Eco-gestion et MIAGE), 1 DESS, 2 ingénieurs et 1 Docteur en informatique ;
- l'expérience qui peut être plutôt informatique ou bien statistique. Les fonctions occupées sont, par exemple : Chargé de projet marketing, Chargé d'études statistiques, Data Manager, Analyste ainsi que Consultant, Cadre Ingénieur d'études Technique mais aussi un Enseignant-chercheur en Informatique ;
- les branches dans lesquelles les apprenants ont acquis leur expérience. On peut noter quelques entreprises ou organisations : bien sûr Softcomputing, IBM France, ATOS, mais aussi Crédit agricole Ile de France, MGEN et la Mutualité Française, ainsi que Bouygues Telecom, France Business School, SNEEP (argus de l'automobile), l'APHP, ou encore la Française des jeux ;
- l'ancienneté professionnelle globale de 3 ans à 30 ans, avec deux modes autour de 6 ans et 12 ans.

La deuxième promotion, qui a démarré son travail en janvier 2016, présente des caractéristiques semblables en termes de taille, de variété et d'équilibres que la précédente.

Remerciements

Les auteurs remercient les deux rapporteurs anonymes de leurs remarques et suggestions qui ont contribué à clarifier et améliorer la première version du manuscrit.

Lien vers le site du DU

http://www.iut.parisdescartes.fr/DIPLOMES/Autres-diplomes/Diplome-d-Universite-Analyste-Big-Data

Bibliographie

[1] Aggarwal, C. C. (Ed.) (2007), *Data Streams: Models and Algorithms, Advances in Database Systems*, Springer.

[2] Berti-Equille, L. (2012), *La qualité et la gouvernance des données au service de la performance des entreprises*, Hermès.

[3] Chang, F., J. Dean, S. Ghemawat, W. C. Hsieh, D. A.Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber (2008), Bigtable: A distributed storage system for structured

data, *ACM Transactions on Computer Systems (TOCS)*, **26**(2), 4.

[4] Dean, J. and S. Ghemawat (2008), MapReduce: simplified data processing on large clusters, *Communications of the ACM*, **51**(1), 107-113.

[5] Desgens-Pasanau, G., F. Naftalski, and S. Revol (2013), *Informatique et Libertés - Enjeux, risques, solutions et outils de gestion*, Editions Lamy.

[6] Gama, J. (2010), *Knowledge Discovery from Data Streams*, Chapman and Hall/CRC.

[7] Hadoop : Open-source software for reliable, scalable, distributed computing, https://hadoop.apache.org/, 2015.

[8] Hoff, P. D., A. E. Raftery, and M. S. Handcock (2002), Latent space approaches to social network analysis, *Journal of the American Statistical Association*, **97**(460), 1090-1098.

[9] Hyndman, R.J. and G. Athanasopoulos (2014), *Forecasting: principles and practice*, Éditeur OTexts.

[10] Ibekwe-Sanjuan, F. (2007), *Fouille de textes : méthodes, outils et applications*, Hermès.

[11] Luckham, D. (2002), *The Power of Events. An Introduction to Complex Event Processing in Distributed Entreprise Systems*, Addison-Wesley Professional.

[12] MongoDB, Document-Store NoSQL database, http://www.mongodb.org/, 2015.

[13] MySQL : Base de Données Open Source, https://www.mysql.fr/, 2015.

[14] Nowicki, K. and T. A. B. Snijders (2001), Estimation and prediction for stochastic blockstructures, *Journal of the American Statistical Association*, **96**(455), 1077-1087.

[15] R Development Core Team (2012), R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org.