Un DU d'Analyste Big Data en formation continue courte, au niveau L3

Jean-Michel Poggi [1], Charles Bouveyron [2], Georges Hébrail [3], François-Xavier Jollois [4]

[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 et Univ. Paris Descartes, georges.hebrail@edf.fr
[4] LIPADE, Univ. Paris Descartes, francois-xavier.jollois@parisdescartes.fr

Résumé. 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 department STID of the IUT Paris Descartes. This 150-class-hour diploma is available for learners in long-life training with at least an undergraduate level (L3 in France). It introduces an innovative way to certify the 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 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, Long-life training, Bachelor

1. Introduction

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.

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. Les objectifs

En reprenant l'analyse du cabinet Gartner, 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).

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.

3. L’organisation

Ce diplôme d’Université s’adresse à des salariés ou à des adultes en reprise d’études. Pour le rendre compatible avec une activité professionnelle et pour permettre la nécessaire maturation des acquis, il se déroule sur 6 mois, à raison de 5 modules de 4 jours chacun, suivi d’un séminaire final de synthèse.

L'ensemble totalise 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 toutes les 2 semaines, afin de faciliter le suivi par des professionnels en entreprise. Chaque certificat se déroule ainsi sur 4 semaines. Les modules s’appuient 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.

La formation s'adresse à des salariés 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.

Les frais de formation (hors droits universitaires) s’élève à 3 000 euros.

4. La formation

Constituée de 5 modules, elle est articulée autour de deux modules plutôt dédiés aux méthodes informatiques, deux plutôt statistiques et un dédié à la qualité et la confidentialité des données. 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 d'Analyste Big Data
en formation continue courte, au niveau L3

On complète dans le paragraphe suivant en précisant les intervenants et la stratégie pédagogique mise en place permettant de décliner le programme.

5. Intervenants et stratégie pédagogique

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

Le module 1 dédié au stockage massif de données se propose d'une part de faire le point sur les outils nouveaux de type NoSQL, avec une mise en application avec des logiciels importants du domaine. D'autre part, il comprend une mise en situation sur la base d'un cloud par apprenant, permettant l'évaluation et la comparaison de différents supports de stockage et technologies d'accès aux données.

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). Une deuxième partie du module ouvre à la fouille de flux de données et se termine par une journée de mise en œuvre d'un logiciel de type « CEP » au travers d'un TP.

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 et temporelles. Les outils de description, d'analyse et de visualisation sont présentés au travers de nombreux exemples en s'attachant aux spécificités du type de donnée concerné et en veillant à la prise en main effective d'outils logiciels utilisant le langage de programmation R.

Le module 4 est dédié aux données libres d'accès, appelée communément « open data », et aux données de types réseaux. 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. Les deux autres journées de ce module sont consacrées au open data et leur visualisation. L'ensemble de ce module est illustré avec le logiciel statistique R et ces paquets d'interaction avec les outils web.

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. Un deuxième enjeu important est la sécurité des données et le respect des contraintes juridiques liées à la protection des données personnelles. 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.

6. La première promotion

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 voté par les instances de l'université qu'en juin dernier, la promotion compte 16 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.

Les premiers modules de la session 2015 débutée en janvier, montrent une grande richesse, 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.

Lien vers le site du DU

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

Bibliographie

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

[2] Cassandra, Column-Store NoSQL database, URL http://cassandra.apache.org/

[3] Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., Chandra T., Fikes A., Gruber, R. E. Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems (TOCS)*, 26(2), 4, 2008.

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

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

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

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

[8] Hyndman, R.J., Athanasopoulos, G., *Forecasting: principles and practice*, Éditeur OTexts, 291 pages, 2014.

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

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

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

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

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