UN PARCOURS BIG DATA EN ALTERNANCE DANS UNE LICENCE PROFESSIONNELLE

Sophie DUPUY-CHESSA[1], Sophie LAMBERT-LACROIX[2] et Gaëlle BLANCO-LAINE[3]

TITLE

A Big Data apprenticeship academic curriculum in a professional licence

RESUME

L'IUT2 de Grenoble proposera à la rentrée prochaine un parcours Big Data au sein d'une Licence Professionnelle en informatique «Systèmes d'information et Gestion des données ». La formation qui aura lieu en alternance, visera le métier d'analyste des données, c'est-à-dire la personne qui collecte et organise les données. Elle s'appuie sur un bassin industriel large et varié et un contexte académique pluridisciplinaire.

Mots-clés : *licence professionnelle, alternance, data analyst.*

ABSTRACT

In September, 2016, the IUT2 of Grenoble will propose a course Big Data within a computer science bachelor "Information systems and data Management". The formation will take place in alternation and will aim at data analyst that collects and organizes the data. It leans on a wide and varied industrial area and a multidisciplinary academic context.

Keywords*: bachelor, alternation, data analyst.*

1 Introduction

Suite au rapport publié par le Ministère du redressement productif, c'est-à-dire le Ministère de l'Industrie, en septembre 2013[4] et au vu des nombreux articles publiés dans le domaine, les départements Informatique et STID (Statistique et Informatique Décisionnelle) de l'IUT2 de Grenoble ont mené une réflexion sur l'opportunité de proposer aux étudiants une formation de niveau L3, professionnalisante dans le domaine du Big Data. Cette réflexion aboutira en septembre 2016 à l'ouverture d'un parcours « Big Data » au sein d'une Licence Professionnelle informatique « Système d'Information et Gestion des données ». Cette nouvelle formation ouvrira en alternance et en formation continue grâce au tissu industriel grenoblois.

Cet article décrit le résultat de notre réflexion sur le futur parcours « Big Data » de la licence professionnelle en informatique « Systèmes d'Information et Gestion des données ». Il débute par la présentation du contexte du projet, industriel et académique. Puis il décrit plus en détails la formation en décrivant son rythme et sa maquette.

[1] IUT2 de Grenoble, sophie.dupuy-chessa@iut2.upmf-grenoble.fr
[2] IUT2 de Grenoble, sophie.lambert-lacroix@iut2.upmf-grenoble.fr
[3] IUT2 de Grenoble, Gaelle.Blanco-Laine@iut2.upmf-grenoble.fr
[4] http://www.redressement-productif.gouv.fr/nouvelle-france-industrielle

2 Contexte industriel et académique

2.1 Le métier visé : analyste de données

Le Big Data étant encore une notion assez jeune, les emplois liés dans le secteur informatique peuvent regrouper des profils assez différents. Ainsi on voit apparaître des postes de développeurs ou de consultants spécialisés en Big Data.

Il existe aussi des évolutions aux métiers de l'informatique décisionnelle dont les prérogatives sont encore relativement diverses. Deux métiers sont néanmoins souvent cités dans les offres d'emploi.

Le premier métier est l'analyste de données (data analyst). Il extrait les données par rapport à des objectifs métier grâce à des techniques statistiques et des outils informatiques spécialisés. L'objectif est d'organiser, de synthétiser et de traduire les informations dont les entreprises ont besoin pour faciliter les prises de décision. Les spécificités du poste par rapport au *data miner* plus traditionnel se situe alors dans le volume des données traitées et la maîtrise des outils spécifiques au Big Data. C'est ce métier de niveau licence qui est visé dans le parcours Big Data.

Un autre métier, plutôt de niveau Master, est celui de « scientifique des données » (data scientist) qui peut être vu comme une évolution de l'analyste des données. Comme lui, il recueille les données et fait des rapports, mais il les regarde aussi sous de nombreux angles, doit déterminer ce que les données signifient pour faire ressortir des indicateurs concrets au service de la direction. Il doit avoir une vision plus transverse des grands volumes de données provenant de sources variées.

Les métiers du Big Data sont donc à la convergence de 3 domaines : l'informatique, la statistique et le business. L'*informatique* pour la programmation pour des traitements rapides sur de gros volumes de données, souvent distribués ; la *statistique* pour les capacités d'innovation et de modélisation, et le *business* pour la capacité à interpréter les indicateurs et à les transformer en langage opérationnel.

Dans ce contexte, nous avons envisagé une formation ciblant le métier d'analyste de données qui aborde les aspects techniques, tant informatiques que statistiques du Big Data. Elle s'appuiera sur le tissu industriel grenoblois et sur le contexte académique de l'IUT2 de Grenoble.

2.2 Le contexte industriel grenoblois

Nous avons identifié des entreprises du bassin grenoblois confrontées à la problématique du Big Data. Ces entreprises ont des besoins grandissants pour le dimensionnement, la mise en œuvre, l'adaptation et le suivi de solutions Big Data et Cloud. Elles peuvent être classées de la manière suivante :

- des entreprises qui ont des problématiques de gros volumes de données pas nécessairement textuelles (ex. : images, vidéos) ; ce sont des entreprises de production industrielle ou de commerce en ligne ;
- des éditeurs de solutions logicielles qui proposent une expertise ou des outils spécifiques au Big Data ;

- les agences web qui proposent des solutions d'hébergement ou d'intégration des données ; elles font du data management et/ou du cloud ;
- les Entreprises de Services du Numérique qui développent une compétence Big Data afin de la vendre pour des missions spécifiques à leurs clients.

Dans le cadre du travail sur ce projet, nous avons contacté plusieurs entreprises implantées sur le bassin grenoblois. Un certain nombre d'entre elles nous ont apporté un soutien de principe et certaines envisagent de contribuer activement à la formation en participant aux enseignements ou en recrutant des étudiants en alternance.

2.3 Contexte académique

Le domaine du Big Data est relativement récent, les offres de formation que nous avons identifiées en France sont toutes très récentes (2013). A part un diplôme universitaire (150H) de niveau licence proposé par l'Université Paris Descartes, toutes les formations existantes sont orientées Bac+5 (ou plus) et plutôt dans le domaine de la formation au métier de « Data Scientist ». Notre objectif est de proposer une formation technique au niveau licence afin de couvrir les besoins en analystes des données. Un requis important est alors de couvrir à la fois les domaines de l'informatique et de la statistique.

Aussi, il a été choisi de faire porter le projet par deux des départements alliant les compétences techniques requises : le département Informatique et le département STID de l'IUT2 de Grenoble.

Le département STID a été créé en septembre 1968. Il développe les compétences essentielles pour la gestion des données et leurs traitements statistiques. Il forme également à l'informatique décisionnelle, à la culture d'entreprise et à la communication.

Le département Informatique de l'IUT2 forme principalement des développeurs en informatique. Créé en 1966, il y a presque 50 ans, l'expérience développée au cours des années a permis au département de s'adapter à un milieu professionnel où l'innovation est constante. Ses multiples rapports avec le milieu professionnel permettent de suivre au plus près les évolutions et les besoins du secteur.

Les départements STID et Informatique sont des départements qui relèvent du secteur secondaire, défini par l'INSEE comme « l'ensemble des activités consistant en une transformation plus ou moins élaborée des matières premières ». Cette qualification permet de prendre en compte la nécessité d'effectuer des investissements matériels en termes de matériels et d'infrastructure (salles machines, serveurs, installations réseau…) plus importants que pour les départements du secteur tertiaire.

Il existe actuellement au département Informatique une licence professionnelle spécialité « Métiers de l'informatique : Système d'Information et Gestion des données » qui comporte un parcours SIMO (Systèmes d'Information, Méthodes et Outils). L'objectif est de créer un parcours frère, Big Data, partageant certains cours, mais orienté différemment et qui introduirait une composante statistique importante ainsi que des enseignements d'informatique spécifiques aux problématiques du Big Data. Cette licence offre une couverture large des métiers de la gestion des données dans le cadre des systèmes d'information des entreprises. Le parcours SIMO permet de mener ou d'assister des projets de développement ou d'évolution de systèmes d'information en prenant en compte les aspects techniques, économiques, organisationnels et humains. Le parcours Big Data le complète en s'intéressant à la gestion de gros

volumes de données variées et non structurées, à leur analyse et, enfin, à l'interprétation des résultats. Les deux parcours offrent ainsi des modules communs et un rythme de formation unifié. Les étudiants auront à choisir un des deux parcours.

3 Descriptif de la formation

3.1 Rythme de la formation

La licence professionnelle « Système d'Information et Gestion des données » s'effectue uniquement en alternance. Tous les étudiants travaillent successivement en entreprise et à l'IUT. Ils sont tous salariés en contrat de professionnalisation, en contrat d'apprentissage, ou en formation continue. Les étudiants en formation continue suivent le même rythme d'alternance que les autres étudiants. Ils bénéficient néanmoins d'un suivi particulier. En tant que salariés, tous les étudiants effectuent 35H par semaine.

La formation à l'IUT est organisée en 5 périodes d'enseignement. La fréquence de l'alternance est de une semaine de cours pour une ou deux semaines en entreprise. Les périodes en entreprise permettront de réaliser le travail correspondant au projet tuteuré et au stage.

3.2 Maquette proposée

Le public visé par la Licence professionnelle correspond à des étudiants ayant validé deux années de licence dans les domaines de l'informatique ou de la statistique. Ce sont en particulier les détenteurs d'un DUT Informatique, d'un DUT STID, d'une L2 Maths-Info, ou d'une L2 Info. Au maximum une vingtaine d'étudiants pourront être formés par an. Les prérequis concernent les bases de données relationnelles, l'algorithmique et la programmation, ainsi que des notions de probabilités et de statistique.

Le volume de formation sera d'environ 500 heures. Son contenu vise à conserver un équilibre entre informatique, statistique et professionnalisation. La pédagogie mise en œuvre sera basée sur la pratique et la réalisation de projets. Il est composé de 5 unités d'enseignement :

- **Enseignements fondamentaux (60H)** : ils regroupent 4 modules de 24 ou 12 heures chacun et comportent les enseignements en informatique et en statistique nécessaires pour la suite de la formation ; ainsi seront abordées l'algorithmique et la programmation (24H), les bases de données (12H), l'extraction des données (12H) et les notions élémentaires de statistique (12H).

- **Informatique développement (168H)** : cette unité d'enseignement vise à renforcer les compétences en programmation des étudiants au travers de 7 modules de 20 à 30 heures. En corrélation avec le parcours « Systèmes d'Information Méthodes et Outils » de la licence, elle contient des enseignements de programmation web dans différences technologies. En complément, elle propose 4 modules spécifiques au Big Data.

  Le premier aborde la sécurité des données. Il a pour objectif de savoir mettre en œuvre des briques élémentaires en termes d'authentification, d'autorisation et de protection des données.

La programmation pour l'interprétation des données vise à former les étudiants à l'utilisation de bibliothèques de traitements statistiques au sein de programmes informatiques.

Le module sur l'interrogation des données en noSQL propose de faire découvrir l'univers du monde NoSQL. Il s'agit de comprendre les différences majeures entre le SQL et le NoSQL et d'être capable de proposer une technologie NoSQL adaptée à un problème donné.

Enfin le module d'utilisation de frameworks pour les traitements distribués doit permettre aux étudiants d'acquérir des compétences fondamentales dans les infrastructures d'analyse de données massives de type Map/Reduce.

- **Statistique (162 H)** : cette unité d'enseignement de spécialité introduit les techniques statistiques nécessaires au Big Data. Elle comporte 5 modules de 24H à 30H.

  Le premier concerne les techniques d'analyses exploratoires utilisées dans le traitement des données massives (Analyse en composantes principales (ACP), Factorisation des matrices non-négatives (NFM), Classification Ascendante Hiérarchique (CAH)). Il a pour objectif de comprendre, mettre en œuvre et interpréter les résultats d'analyses exploratoires utilisées dans le traitement des données massives.

  Le module de modélisation statistique (modèles gaussiens, régression logistique...) enseigne la prévision et donc la recherche de modèles optimaux. Il doit permettre aux étudiants de comprendre, mettre en œuvre et interpréter les modèles de régression linéaire simple et multiple, la régression logistique. Les étudiants doivent comprendre la différence entre la mise en place (algorithmes) des méthodes classiques et celles appliquées pour le traitement des données massives.

  Il y a aussi deux modules de fouille de données (data mining) pour introduire les techniques d'apprentissage supervisé ou non, utilisées le plus couramment en fouille de données volumineuses (k-means, arbres de décision, forêts aléatoires, Text et Web mining...). L'accent sera mis sur l'utilisation et la compréhension des méthodes statistiques qui permettent de passer à l'échelle des données massives (c'est-à-dire dont les algorithmes peuvent être distribués) et de certaines librairies associées (MLlib, les packages de R dédiés…).

  Ces enseignements de statistique sont complétés par deux modules d'informatique décisionnelle, ainsi que la visualisation de données complexes.

  Le module d'informatique décisionnelle propose un enseignement classique en conception et en réalisation d'entrepôt de données avec une architecture ROLAP (Relational OLAP).

  Le module de visualisation des données présente les fondements de la visualisation d'information, ainsi qu'un panorama des techniques de l'état de l'art applicables à différents types de jeux de données. Il donne aussi les clés pour la conception de nouvelles techniques de visualisation interactives adaptées à des données et des tâches spécifiques.

- **Environnement (110H)** où sont regroupés des enseignements relatifs à la culture scientifique, sociale et humaine afin de situer le Big Data dans son contexte social et technique. De manière habituelle pour une licence professionnelle, seront proposés des

cours d'expression et communication, d'anglais, de gestion de projet. Un module introductif à la problématique du Big Data ainsi que des séminaires spécifiques complètent les enseignements d'environnement. Ces séminaires, par exemple « Vie privée et sécurité », doivent donner aux étudiants une ouverture sur des problématiques liées au contexte d'intervention mais en dehors du domaine métier technique.

- **Projet tuteuré** qui est un premier retour sur le travail en entreprise où les étudiants devront présenter leurs missions et le contexte de leur travail en entreprise.

- **Stage** qui se déroulera tout au long de l'année durant les périodes d'alternance en entreprise et qui donnera lieu à une présentation à la fin de la formation. Cette présentation devra permettre de mettre en perspective les missions réalisées au cours de l'année au regard des enseignements et des problématiques métier spécifiques au Big Data.

4 Conclusion

Cet article présente la réflexion qui a été menée au sein des départements Informatique et STID de l'IUT2 de Grenoble pour l'ouverture d'un parcours de licence professionnelle « Métier de l'informatique » orienté Big Data. La formation ouvrira en septembre 2016 uniquement en alternance et en formation continue, et accueillera au maximum une vingtaine d'étudiants visant le métier d'analyste des données. Elle sera, bien sûr, amenée à évoluer en fonction des besoins des entreprises et des candidats.