Enseigner la Statistique pour l'Analyse de Mégadonnées

Philippe Besse [1], Nathalie Villa-Vialaneix[2] & Anne Ruiz-Gazen[3]

[1] *Université de Toulouse, INSA & IMT, UMR CNRS 5219*
[2] *Université Toulouse 1 Capitole, 21 allée de Brienne, 31000 Toulouse*
[3] *INRA, UR 0875 MIAT, BP 52627, F-31326 Castanet Tolosan cedex - France*

Résumé. L'objectif de cette communication est un retour d'expérience sur l'introduction d'un cours ou de notions liés à l'analyse des mégadonnées « *Big Data* » et abordant les notions d'exploration, apprentissage et parallélisme dans ce contexte. Plus précisément, nous décrirons deux cours de ce type que nous avons conçus conjointement mais dont les contextes, contenus et organisations diffèrent. Il s'agit, d'une part, des modules d'*Exploration et Logiciels Statistiques* (4ème année) et d'*Apprentissage Statistique* (5ème année) du cursus « Génie Mathématique et Modélisation » de l'INSA de Toulouse [1] et, d'autre part, d'un cours de « *Multivariate data analysis - Big data analytics* » dispensé en 1ère année des masters « Economics » et « Economics and Statistics » de Toulouse School of Economics [2]. Notre objectif, outre une introduction basique à la problématique enseignée, est de montrer les difficultés, matérielles et pédagogiques, auxquelles se heurte l'enseignant statisticien pour aborder ces concepts et de présenter quelques choix que nous avons faits et la manière dont ces choix ont été reçus par les étudiants.

Mots-clés. *Big Data*, enseignement, *machine learning*, R, Hadoop, calcul parallèle

Abstract. This proposal's aim is to provide a feedback on a new course about "big data", oriented toward data analysis and machine learning. The course also discusses the problems related to parallel computing and scalability. More precisely, we will describe two courses, that were jointly prepared even though their contents are slightly different. The first one was taught for 4-*th* year students of the curriculum "Génie Mathématique et Modélisation" of the INSA de Toulouse [3] and was entitled "Mining and statistical software", and the second one was given for 1-*st* year students of the master "Economics and Statistics" of Toulouse School of Economics [4] during a course entitled "Multivariate analysis - Big data analytics". Our proposal is to present some basic concepts about the topic and to describe the choice that we made, as well as the way the students reacted to these choices.

Keywords. Big Data, teaching, machine learning, R, hadoop, parallel computing

1. http://www.insa-toulouse.fr/fr/formation/ingenieur/specialites/gm.html
2. http://ecole.tse-fr.eu/en/programs/graduate/master-1/economics-and-statistics
3. http://www.insa-toulouse.fr/fr/formation/ingenieur/specialites/gm.html
4. http://ecole.tse-fr.eu/en/programs/graduate/master-1/economics-and-statistics

1 Positionnement

Les dernières années ont connu une modification importante du volume et de la nature des données auxquelles le statisticien est confronté. D'une part, la capacité de stockage des ordinateurs s'est considérablement accrue : en 60 ans d'évolution, on est passé d'une capacité de stockage de quelques mégaoctets au maximum [5] à des disques durs dont la taille est très réduite mais la capacité de stockage peut dépasser plusieurs téraoctets de données, soit 10^6 fois plus que celle des premiers disques durs. Par ailleurs, le coût de moins en moins élevé du matériel et la possibilité d'utiliser des systèmes de stockage en ligne ou sur le *cloud* ont entraîné une modification des systèmes d'information des entreprises : celles-ci stockent des volumes de plus en plus importants, avec des flux continus et massifs de nouvelles données [6]. Celles-ci sont, de plus, souvent de natures diverses, mêlant variables numériques ou catégorielles, fichiers non structurés à des objets plus complexes comme des graphes (données relationnelles) ou des courbes (données fonctionnelles), par exemple. L'ensemble de ces problématiques nouvelles a donné lieu à l'expression « *Big Data* (« Méga Données ») qui repose sur le concept introduit en 2001 par Douglas Laney (groupe Gartner) des « 3V » : *Volume*, *Velocity* (en référence au flux rapide des données) et *Variety* (en référence à leur nature multiple et complexe).

Comme le soulignent [2], cette évolution conduit à repenser le travail du statisticien, qui devient de plus en plus un *data scientist* [7]. L'émergence de ce terme a connu un écho important, à la suite de la déclaration de Hal Varian, chef économiste chez Google, qualifiant la profession de « *sexy job in the next 10 years* » (voir aussi [4]). De fait, l'ère du *Big Data* est aussi l'ère de l'*Open Data*, qui voit notamment l'émergence de site de données ouvertes fournies par des grandes entreprises, sur lesquels sont diffusés des concours publics pour qui souhaite proposer une solution originale de traitement des données mises en ligne [8]. La compétence de *data scientist* est donc une compétence recherchée. [1] souligne qu'il s'agit de former des techniciens et des ingénieurs ayant des connaissances à l'intersection des Mathématiques, de la Statistique et de l'Informatique, avec en outre, éventuellement, des connaissances en *machine learning*, gestion de données et même *business/management*. Pour aider à l'évolution des formations, le rapport [1] dresse un bilan des formations françaises préparant au métier de *data scientist* et fournit une liste indicative de cours qui pourraient être intégrés dans ce type de formations.

L'objectif de cette communication est un retour d'expérience sur l'introduction d'un cours ou de notions liés à l'analyse des mégadonnées, *Big Data*, et abordant les tech-

5. L'ordinateur IBM 350, contenant le premier disque dur avait la taille d'une grande armoire et une capacité de stockage d'environ 3, 75MB.

6. [6] rapportent ainsi qu'en 2010, l'entreprise Facebook© possédait 21 pétaoctets (1PB = 10^3TB) de données avec une augmentation journalière de l'ordre de 12 TB.

7. Un « scientifique des données ».

8. Par exemple, pour la France https://datascience.net//fr/challenge qui contient des données provenant de la SNCF, l'INSEE, AXA... ou bien le célèbre Prix « Netflix » http://www.netflixprize.com qui était doté de 1M$.

niques d'exploration, d'apprentissage et aussi de parallélisme dans ce contexte. Plus précisément, nous décrirons deux approches, que nous avons conçues conjointement mais dont les contextes, contenus et organisations diffèrent. Il s'agit, d'une part des modules d'*Exploration et Logiciels Statistiques* ($4^{\text{ème}}$ année) et d'*Apprentissage Statistique* ($5^{\text{ème}}$ année) du cursus « Génie Mathématique et Modélisation » de l'INSA de Toulouse [9] et, d'autre part, d'un cours de « *Multivariate data analysis - Big data analytics* » dispensé en $1^{\text{ère}}$ année des masters « Economics » et « Economics and Statistics » de Toulouse School of Economics [10]. Notre objectif, outre une introduction basique à la problématique enseignée, est de montrer les difficultés, matérielles et pédagogiques, auxquelles se heurte l'enseignant statisticien pour aborder ces concepts et de présenter quelques choix que nous avons faits et la manière dont ils ont été reçus par les étudiants [11].

2 Rôle du statisticien. Objectifs du cours

La question que nous souhaitons aborder au travers de cette présentation est celle de la place du statisticien dans le *Big Data*, ou plus particulièrement, quels outils mathématiques et statistiques sont nécessaires et comment sont-ils adaptés à ce contexte ? En effet, les premières propositions méthodologiques pour traiter des données volumineuses concernaient essentiellement des opérations simples de tri, comptage ou d'interrogation des données qui ne requiéraient pas de compétences statistiques particulières [5]. Récemment, l'analyse et la modélisation des données massives a connu un intérêt croissant et les environnements logiciels permettant de gérer des grandes bases de données incorporent maintenant fréquemment des librairies de *machine learning* [12] qui incluent quelques approches statistiques standard pour la fouille de données (ACP, k-means, ...) et pour la modélisation (régression linéaire, SVM, CART, forêt aléatoire,...).

Si la nécessité de la présence du statisticien dans le processus semble désormais acquise, la manière d'aborder la formation des étudiants statisticiens, reste ouverte. Une première solution consiste à introduire les méthodes statistiques citées ci-dessous dans un cadre *small data* : c'est évidemment un préalable mais cela ne suffit pas. Des problèmes spécifiques se posent lors du passage à l'échelle et il semble donc important de pouvoir présenter et illustrer, dans un cadre statistique, les notions d'*architecture informatique distribuée* (qui permettent de stocker des données et de gérer des calculs sur plusieurs machines travaillant conjointement) et de *calcul parallèle*. L'enseignant de statistique n'a toutefois pas souvent à sa disposition un gros *cluster* de calcul pour faire travailler ses

9. http://www.insa-toulouse.fr/fr/formation/ingenieur/specialites/gm.html
10. http://ecole.tse-fr.eu/en/programs/graduate/master-1/economics-and-statistics
11. À l'heure d'écriture de cette proposition, les cours INSA viennent de s'achever mais celui de TSE n'a pas encore commencé.
12. par exemple, Mahout [13], projet de la fondation Apache™ pour l'environnement Hadoop ou bien la librairie MLlib [14] de l'environnement Spark [15] qui est aussi un projet de la fondation Apache™.

étudiants. De manière plus réaliste, il a généralement accès à une salle d'ordinateurs de bureau, au mieux multi-cœurs, mais souvent sous système d'exploitation Windows. Compte tenu de ces contraintes, nous présentons dans la section suivante quelques choix pédagogiques que nous avons faits afin de sensibiliser les étudiants à la notion de passage à l'échelle dans l'analyse de mégadonnées.

3 Mise en œuvre pratique

Le logiciel statistique R reste une entrée facile pour aborder l'analyse de données avec les étudiants en statistique et il possède l'avantage d'incorporer une large panoplie de méthodes statistiques, directement ou au travers des nombreux packages supplémentaires qui existent. R permet également, d'aborder, de manière simple, la notion de parallélisation des calculs : la *CRAN Task View* sur le calcul parallèle haute performance avec R[16] fait la liste d'un certain nombre de packages permettant de « pousser les limites du logiciels un peu plus loin ». En particulier, les packages **snow**, **snowfall** ou **foreach** permettent à l'étudiant de concrètement implémenter la parallélisation d'une analyse de données et, si il a à sa disposition un simple ordinateur multi-cœurs, d'en observer les effets en terme de temps de calcul.

R reste toutefois très limité dans la gestion de la mémoire : en effet, les données traitées sont importées dans la mémoire vive (RAM) de la machine, ce qui, dans le cadre du *Big Data* s'avère impossible (la taille des données excédant de beaucoup la taille de la mémoire vive des ordinateurs). Là encore, il semble important de sensibiliser les étudiants à ce problème puis de leur indiquer des solutions alternatives : la première peut être l'utilisation dans R de packages spécifiques à la gestion de la mémoire, comme **bigmemory** qui stocke les objets (comme les matrices volumineuses) sur le disque dur et y accède via des pointeurs.

Mais ces solutions restent limitées comparées aux environnements logiciels qui ont été spécifiquement conçus pour le *Big Data*. Parmi ceux-ci, un des plus populaire est *Hadoop*[17] qui est un environnement logiciel programmé en Java qui, d'une part, incorpore un système de fichiers distribué sur plusieurs machines[18] et d'autre part, inclut des programmes permettant le traitement parallèle des données. En particulier, Hadoop dispose d'une implémentation de l'algorithme *MapReduce* [5] qui est une méthode permettant de réaliser des calculs à grande échelle sur un grand cluster de calcul (voir Figure 1).

16. http://cran.r-project.org/web/views/HighPerformanceComputing.html
17. http://hadoop.apache.org
18. HDFS : Hadoop Distributed File System qui a été conçu à partir du GoogleFS, Google File System.

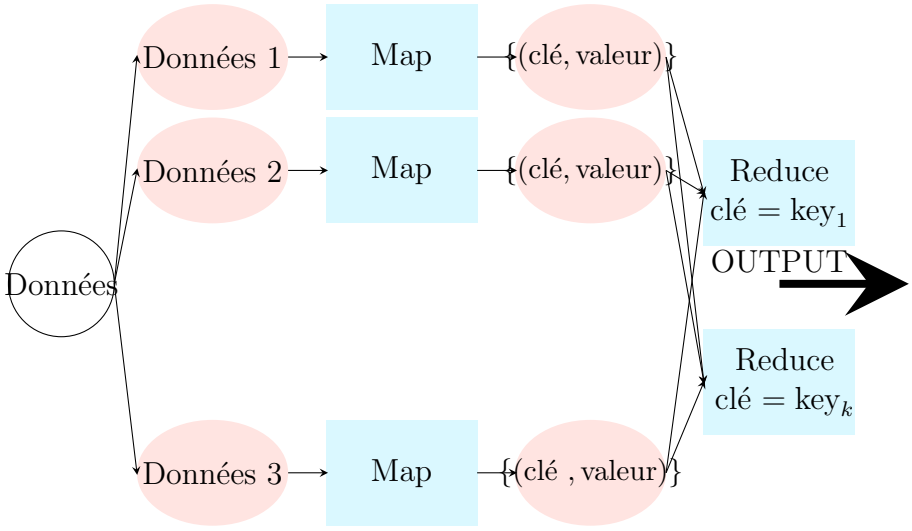

FIGURE 1 – Illustration de l'algorithme *MapReduce*.

Le principe de l'algorithme est de décomposer le calcul en deux grandes étapes : une étape « *Map* » qui traite l'intégralité des données en les découpant en sous-ensembles de petite taille, traités en parallèle. La sortie de l'étape *Map* est un ensemble de paires (clé, valeur) où « clé » est une clé d'indexation et « valeur » est la valeur associée à cette clé. L'étape « *Reduce* » collecte toutes les données correspondant à une valeur donnée de la clé d'indexation et traite chacune des clés en parallèle pour fournir la sortie finale. Il est toutefois rare que les formations standards en statistique mettent à la disposition de leurs étudiants un cluster avec un environnement hadoop fonctionnel. Pour présenter le principe de MapReduce, il est encore possible d'avoir recourt à R et d'utiliser le package **rmr2** [19], qui est une interface entre Hadoop et R, pour initier les étudiants à la manière d'adapter les analyses statistiques classiques pour les faire entrer dans le paradigme de MapReduce. Ceci peut être effectué y compris sans environnement Hadoop fonctionnel car le package inclut une option permettant de faire fonctionner MapReduce sur un environnement local (c'est-à-dire, sans Hadoop). Des exemples d'adaptation de méthodes statistiques classiques pour MapReduce sont proposées sur le blog officiel de Revolution Analytics [20] (voir aussi [3]).

Enfin, pour une mise en œuvre pratique plus efficace et plus réaliste, le langage de programmation Python semble être une bonne alternative pour les enseignants : il partage avec R la simplicité de prise en main mais avec une plus grande efficacité et une

19. Ce package est développé par l'entreprise Revolution Analytics, http://www.revolutionanalytics.com/ et n'est pas disponible sur le CRAN mais à l'adresse https://github.com/RevolutionAnalytics/RHadoop/wiki/rmr.

20. https://github.com/RevolutionAnalytics/rmr2/blob/master/docs/tutorial.md

meilleure gestion de la mémoire. La librairie Scikit-learn [21] embarque un grand nombre de méthodes d'analyse de données (classification, discrimination, régression, sélection de variables, réduction de la dimension, ...).

La communication que nous présenterons sera tournée vers l'illustration de ces différentes possibilités sur des jeux de données de tailles moyennes, compatibles avec leur mise en œuvre en classe et avec la nécessité d'illustrer les problèmes rencontrés lors de l'augmentation de la taille des données à traiter. Nous ferons aussi un retour d'expérience et un bilan sur les réactions des étudiants.

Références

[1] S. Abiteboul, F. Bancilhon, F. Bourdoncle, S. Clémençon, C. de la Higuera, G. Saporta, and F. Soulie-Fogelman. L'émergence d'une nouvelle filière de formation : "data scientist". Technical report, 2014.

[2] P. Besse, A. Garivier, and J.M. Loubes. Big data - Retour vers le futur 3. De statisticien à data scientist. arXiv preprint arXiv :1403.3758, 2014.

[3] P. Besse and N. Villa-Vialaneix. Statistique et big data analytics. Volumétrie, l'attaque des clones. arXiv preprint arXiv :1405.6676, 2014.

[4] T.H. Davenport and D.J. Patil. Data scientist : the sexiest job of the 21st century. *Harvard Business Review*, October 2012.

[5] J. Dean and S. Ghemawat. MapReduce : simplified data processing on large clusters. In *Proceedings of Sixth Symposium on Operating System Design and Implementation (OSDI 2004)*, 2004.

[6] A. Thusoo, Z. Shao, S. Anthony, D. Borthakur, N. Jain, J. Sen Sarma, R. Murthy, and H. Liu. Data warehousing and analytics infrastructure at facebook. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 2010)*, pages 1013–1020, 2010.

21. http://scikit-learn.org