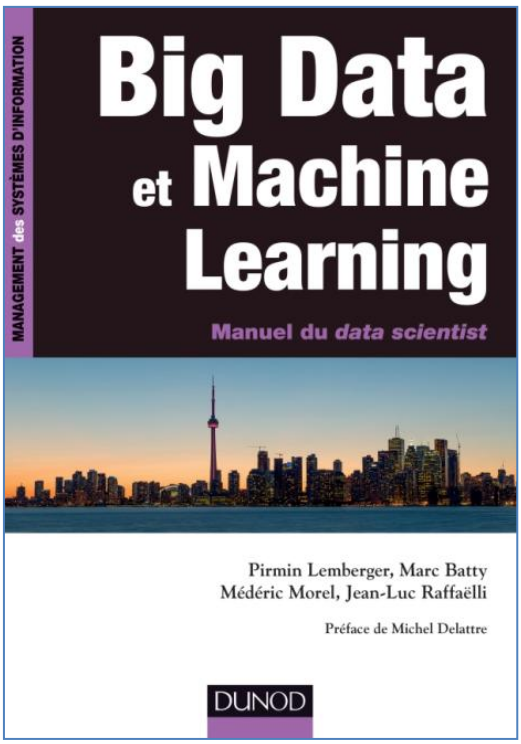
## Big Data et Machine Learning - Manuel du data scientist

**P. Lemberger, M. Batty, M. Morel, J.L. Rafaëlli**

Dunod, 2015.




« Big data », « data science » sont des termes dont la popularité est croissante dixit Google Trends. « Machine learning » revient à la mode. Regroupant ces trois appellations, cet ouvrage s'inscrit à l'évidence dans l'air du temps. Son objectif est de clarifier les différentes notions qui gravitent dans la sphère « big data », un domaine que tout le monde identifie comme porteur de perspectives majeures pour les années à venir. Personnellement, quand je vois un journal de la stature de « Le Monde » consacrer un blog spécifique au « data », je me dis qu'il y a vraiment quelque chose.

Mais recueillir et gérer des données, les exploiter pour en déceler des régularités et en déduire des connaissances est-il si nouveau que cela ? N'est-ce pas là les bases mêmes de ce qu'on appelle l'analyse de données[1], de la modélisation statistique[2], du data mining[3] ? Oui et

[1] La section concernant l'histoire de l'analyse de données est accessible sur la page wikipédia.

non bien sûr. La démarche est ancienne, mais elle s'inscrit aujourd'hui dans un nouveau contexte économique et technologique comme le rappelle Michel Delattre dans la préface. Le coût du stockage et de la CPU connaissent une baisse sans précédent. Dans le même temps, les sources et la circulation des informations connaissent une avancée phénoménale, grâce notamment aux géants du web. Nous sommes dans un cadre sociétal tout à fait nouveau. Tout un chacun devient pourvoyeur de données et, par extension diffuseur de connaissances. A chaque fois que nous cliquons sur une série d'ouvrages sur *Amazon*, nous sommes en train de renforcer leur système de recommandation, qui influencera les choix d'autres internautes. A chaque fois que nous cliquons sur un bouton « j'aime » sur *Facebook*, nous sommes en train de mettre en relation des profils (le notre) et des préférences, voire des comportements. A chaque fois que nous donnons notre avis sur un forum, nous donnons des indications sur nos goûts, sur notre perception des objets qui nous entourent, sur les produits que nous sommes susceptibles d'acheter ou de rejeter. La liste est longue et dépasse le web. Nos téléphones portables en disent plus sur nous-mêmes que notre propre conscience ; les systèmes de navigation GPS n'hésitent pas à nous demander si nous acceptons que notre trajet soit tracé pour que l'éditeur puisse améliorer son algorithme ; etc. Tout cela contribue à la constitution d'une gigantesque masse de données, de sources différentes, de formats différents, de structures différentes. Il est forcément tentant d'en tirer parti.

Tout le monde comprend que pouvoir croiser des informations est créateur de valeur. Mais faut-il tout stocker pour tout analyser ? Qu'est-il pertinent de traiter ? Par quoi commencer ? Comment s'organiser ? Cet ouvrage se propose de nous guider dans la compréhension des enjeux des projets de data science, en traitant des concepts théoriques (nouvelles approches de stockage des données, méthodes d'apprentissage automatique), et en décrivant les technologies et les outils. Il s'adresse aux décideurs informatiques, aux professionnels du décisionnel et des statistiques, aux architectes informatiques, aux responsables métiers.

L'ouvrage est décomposé en **3 grandes parties**. **La première** (chapitres 1 à 4) précise la notion de big data et décrit deux technologies clés : les bases de données NoSQL et le paradigme de programmation MapReduce. **La seconde** (chapitres 5 à 8) est consacrée au

[2] NIST/SEMATECH, « e-Handbook of Statistical Methods », 30/10/2013 ; chapitre 4 « Process Modeling ».
[3] Wikipédia, « Exploration de données », consultée en juin 2015.

métier de data scientist. Quelles sont les compétences qui lui sont associées. Quelle est sa place dans les organisations. Quel type de formation prépare à ce métier. Les auteurs embraient sur plusieurs tâches clés du data scientist : la préparation des données, le machine learning (que je traduirais pas modélisation statistique), et la visualisation. On ne peut s'empêcher de faire un rapprochement avec les étapes du processus data mining (cf. la trame CRISP-DM par exemple). Cela montre bien que data scientist n'est pas un métier créé de toutes pièces mais constitue - en partie, nous y reviendrons lorsque nous détaillerons le chapitre 5 - une évolution du métier de statisticien en entreprise. **La troisième partie** (chapitres 9 à 12) examine le passage à l'échelle en approfondissant les technologies de stockage et de traitement distribués, et l'analyse en temps réel.

Le **chapitre 1** s'intéresse aux origines du big data. Comme évoqué plus haut, nous entrons dans une ère où l'usage de l'informatique et les connexions entre les machines[4] connaissent une envolée sans précédent, nous donnant l'opportunité de collecter une masse d'information énorme. Dans le même temps, les coûts de stockage ont fortement baissé, les technologies adaptées à la gestion des grandes masses de données connaissent une progression considérable. Tout est réuni pour que les différents acteurs des entreprises s'intéressent à la **valorisation** de cette manne qui nous tend les bras. Les auteurs recensent justement quelques opportunités offertes par le big data (section 1.6) : l'analyse et l'assemblage des volumes extrêmes ; l'accès à un éventail de données jamais atteint par le passé ; capter les données en flux continu ; la mise en correspondance - le croisement - de données d'origines variées.

Le **chapitre 2** s'intitule « le big data dans les organisations ». Les auteurs s'appuient sur la caractérisation bien connue des big data par les « 3 V » : volume, oui à l'évidence ; vélocité, les mises à jour se font parfois en flux continu, il est parfois opportun de réaliser des analyses en temps réel ; variété, parce qu'elles sont de sources différentes et se présentent sous de multiples aspects. Ce cadre nécessite de nouvelles compétences à acquérir, notamment pour le statisticien. Des outils et des usages adaptés apparaissent. Il impacte également les métiers, l'organisation de l'entreprise, les modes de décision, les relations

[4] On serait extralucide, on verrait voler au dessus de nos têtes des données, des fichiers binaires... C'est affolant quand on y pense.

avec les clients. L'exemple du « community manager[5] » dans la section 2.7.3 est particulièrement édifiante à cet égard. Vous diffusez un nouveau produit, vous souhaitez cerner l'avis des consommateurs. Plutôt que de procéder par une étude qualitative ou par sondage, vous investissez internet, les réseaux sociaux, les pages des forum. On me dit souvent que les réactions sur internet sont excessives, donc inexploitables. C'est vrai en partie. Mais la répétition de certains mots-clés peut faire la connaissance. Quand un internaute dit « je n'aime pas », la valeur ajoutée est faible. En revanche, lorsque plusieurs personnes mentionnent le conditionnement du produit dans leurs commentaires négatifs, il y a certainement des choses à creuser de côté-là. Mais « community manager » va au-delà d'une simple analyse des informations existantes, il s'agit carrément de fédérer (manipuler ?) un ensemble de personnes. Les dérives peuvent vite arriver. La citation d'Eric Schmidt (PDG de Google) au sujet du respect de la vie privée fait froid dans le dos (page 24) : « Si vous faites quelque chose et que vous ne voulez que personne ne le sache, peut-être devriez-vous déjà commencer par ne pas le faire ». Il ne connaît certainement pas la fameuse formule de Rabelais : « Science sans conscience n'est que ruine de l'âme ». Ceci étant dit, l'âme je ne sais pas, mais pour ce qui est de la ruine, je crois que Google n'a pas trop de soucis à se faire.

Le **chapitre 3** est centré sur les bases de données NoSQL. Les systèmes de gestion des bases de données relationnelles (SGBDR) sont très largement répandues et ont fait leurs preuves. Mais elles répondent à une certaine finalité. Aujourd'hui, la volumétrie augmente toujours de manière considérable, l'usage que l'on fait de certaines bases ne requièrent pas des opérations gourmandes en calcul telles que les jointures, les structures des données à stocker évoluent, les priorités ont également changé. La performance et la disponibilité deviennent des critères clés dans un contexte de stockage distribué des données.

Les bases de données NoSQL répondent aux priorités suivantes (page 36) : distribuer les traitements et le stockage ; donner la priorité aux performances et la disponibilité ; traiter efficacement les données non structurées. Dans les faits, les différentes solutions partagent un certain nombre de points communs (page 37) : elles sont clusterisables et permettent une montée en charge linéaire (deux fois plus de machines = augmentation de la

[5] Voir à ce sujet un interview assez révélateur d'un community manager paru dans Rue89 : « Luc, community manager : ''On s'en prend rapidement plein la gueule'' », 15 juin 2015.

performance dans les mêmes proportions) ; elles sont dépourvus de schémas, trop contraignants ; elles n'offrent pas de jointure ; elles sont dépourvues de transactions.

Les auteurs distinguent deux catégories de bases NoSQL : les bases de données orientées agrégats (BDOA) et les bases orientées graphes (BDOG). Aux premières (BDOA) sont associées principalement les entrepôts clés valeurs, les bases clés-documents, et les bases orientées colonnes. Deux propriétés se démarquent : la souplesse de la structure, et le rôle de la clé dans un système de hachage qui permet un accès très rapide aux enregistrements. La recherche peut poser problème puisqu'il faut a priori scanner toute la base pour retrouver tous les enregistrements relatifs à un terme ou à mot clé mais, avec le mécanisme des index inversés, elle est dans les faits extraordinairement rapide (il suffit de voir à quelle vitesse *Google* répond à nos requêtes).

Les secondes (BDOG) sont surtout adaptées pour stocker les informations qui s'architecturent naturellement sous forme de graphes. Elles se prêtent naturellement aux opérations propres aux graphes (parcours, recherche de chemin entre les nœuds, etc.). On pense naturellement aux données relatives aux réseaux sociaux.

Comme le souligne les auteurs à la fin du chapitre, les bases NoSQL ne viennent pas remplacer les SGBDR. Elles les complètent en répondant à de nouveaux objectifs dans un contexte qui présente de nouvelles caractéristiques.

Le **chapitre 4** s'intéresse à l'algorithme MapReduce et le framework Hadoop. L'idée est de tirer profit d'une infrastructure distribuée constituée de clusters de machines (nœuds). Hadoop est l'implémentation la plus connue à ce jour. L'ouvrage en explicite les principales caractéristiques et les points forts. A mon sens, deux propriétés se détachent : un système de fichiers distribué (HDF : hadoop distributed system) et le patron de programmation MapReduce, qui permet de paralléliser les traitements.

La programmation parallèle est un domaine à part entière de l'informatique. Il s'agit de distribuer les traitements sur plusieurs machines, ou plusieurs processeurs, ou plusieurs cœurs de processeurs. L'idée est d'exploiter efficacement les ressources disponibles en veillant notamment à l'équilibrage des charges c.-à-d. répartir équitablement les calculs pour éviter les temps d'attente. C'est un problème difficile qui nécessite une expertise avancée en développement.

MapReduce propose un schéma de programmation permettant d'organiser la parallélisation. La figure 4.1 de la page 53 décrit clairement les étapes du processus : le système découpe les données en lots, nous n'avons pas prise là-dessus ; nous programmons une fonction « map » qui sera appelée autant de fois qu'il y a de lots ; elle effectue éventuellement des traitements, elle se charge aussi de proposer une autre répartition des données résultantes en blocs associant des clés aux valeurs ; le système réorganise les données en conséquence ; et pour chaque nouveau bloc - nous avons prise sur sa constitution cette fois-ci - la fonction « reduce » que nous programmons est appelée. Dans ce dispositif, notre rôle consiste essentiellement à programmer correctement les fonctions « map » et « reduce ».

MapReduce est censé nous faciliter la programmation distribuée. Ca ne veut pas dire pour autant qu'il résout tous les problèmes. Une certaine expertise est nécessaire pour identifier ce que nous devons intégrer dans les fonctions « map » et « reduce », qu'il faudra ensuite implémenter efficacement. Il concourt néanmoins à rendre la programmation distribuée plus accessible en nous épargnant toute une série de tâches compliquées, notamment la synchronisation des traitements qui constitue un goulot d'étrangement critique en la matière.

Le **chapitre 5** décrit « le quotidien du data scientist ». Il débute la seconde partie de l'ouvrage consacré « **au métier de data scientist** ». J'y suis particulièrement sensible puisque mon travail consiste à former des étudiants en statistique et informatique qui vont se positionner sur le marché du travail. Peuvent-ils investir ce créneau ? De quelle manière ? Quelles sont les enseignements supplémentaires que nous devons introduire pour ce faire ? Mieux cerner le rôle du data scientist dans l'entreprise contribue forcément à mieux situer le contenu des formations qui pourraient leur être dédiées.

Data science, nouvelle discipline ou pas, j'ai l'impression de revivre les controverses autour du data mining à la fin des années 90. Le terme est en vogue à l'évidence[6], après il nous appartient de lui associer des compétences. C'est une bonne manière à mon sens d'identifier ses réelles contributions dans une organisation.

[6] Les auteurs citent notamment l'article de T.H. Davenport et D.J. Patil, « Data scientist : the sexiest job of the 21st century », Hvard Business Review, octobre 2012.

Reprenant une définition communément admise, les auteurs associent trois disciplines de compétence au data scientist. La première dimension est **statistique**. L'objectif est de traiter des données de manière à en faire apparaître des relations, des régularités. C'est la vocation première des statistiques. Le mot clé « **machine learning** » revient souvent aussi aujourd'hui, associant plutôt la discipline à l'informatique. Mais quand on y regarde de plus près, Ronald Fisher n'a-t-il pas déjà investi l'apprentissage supervisé avec ses travaux sur l'analyse discriminante dans les années 30 ? Peut-on vraiment nier que Fisher soit un statisticien ? On peut en dire autant de l'apprentissage non supervisé (classification automatique). A mon sens, le véritable intérêt du rapprochement entre ces deux communautés (statistique et apprentissage automatique) est d'avoir pu mettre dans un pot commun différentes méthodes qui poursuivaient des objectifs identiques, qui présentaient des caractéristiques comparables, mais que l'on n'arrivait pas à positionner les uns par rapport aux autres par pur sectarisme[7].

Une phrase clé a retenu mon attention dans l'ouvrage (page 74) : « Une des différences entre un statisticien et un data scientist dans son acception moderne est que ce dernier accorde moins d'importance à la pureté statistique d'un indicateur ou d'un algorithme qu'à son utilité ''business'' ». Elle peut laisser à penser que les statisticiens sont des gens rigides, attachés avant tout aux propriétés mathématiques des méthodes, alors que les data scientists seraient plus ancrés dans la réalité. C'est un peu caricaturer je trouve. Il y a autant de personnes obtuses en statistiques qu'il y en a en dans tout autre domaine. En revanche, le statisticien ne peut pas tout savoir, c'est une évidence. L'idée de « data lab » qui est avancée plus loin est justement une manière de dépasser ce genre de verrou.

L'**informatique** est la seconde discipline associée au data science. Elle coule de source. Le data scientist doit manipuler des données, stockées de différentes manières, avec des structures variées. Il faut disposer de compétences avancées en informatique pour pouvoir

[7] Etudiant en économétrie, les réseaux de neurones ont gardé pendant longtemps une aura mystérieuse à mes yeux. Lorsqu'enfin, j'ai pu lire des ouvrages suffisamment accessibles (suffisamment bien écrit ?) me permettant de comprendre d'une part qu'il y avait en réalité différents types de réseaux, et que d'autre part, ils avaient des finalités que je comprenait très bien (description, structuration, prédiction à partir de tableaux de données), la barrière à l'entrée a sauté. En étudier les mécanismes n'était plus un problème. Franchement, il n'y a pas plus ou moins de difficultés à étudier la maximisation de la vraisemblance en régression logistique via l'algorithme d'optimisation de Newton Raphson, qu'à analyser le mécanisme de rétrop-progagation du gradient d'un perceptron simple ou multi-couches.

le faire, tant en traitement des bases de données (un data scientist qui ne sait pas faire des requêtes SQL à mon avis doit songer rapidement à une reconversion) qu'en programmation (il doit définir des succession de tâches complexes, il est inenvisageable de les faire et reproduire à la main). Les auteurs insistent d'ailleurs sur l'aspect expérimental du métier, notamment parce qu'il est au cœur de multiples innovations. Il faut un savoir faire informatique certain pour pouvoir se dépêtrer de situations non conventionnelles.

La dimension **métier** est la troisième caractéristique du data science. Elle est évidente pour le professionnel. Personne ne traite des données comme ça dans l'absolu. L'objectif est de répondre à des interrogations concrètes, avec à la sortie des prises de décision qui vont influer sur le fonctionnement ou les résultats de l'entreprise. Il est impératif de tenir compte des éléments de contextualisation spécifiques au domaine étudié. Il faut une connaissance approfondie du métier pour cela.

Trouver des personnes qui réunissent ces caractéristiques est difficile, on parle de « mouton à cinq pattes »[8]. Pour aller plus loin, les auteurs avancent la notion séduisante de **data lab** (page 78). Il s'agit des petites équipes pluridisciplinaires constituées de personnes hautement qualifiées qui réunissent des compétences différentes : des architectes logiciels, des analystes métiers, des data scientist au sens statisticien, des développeurs, des designers web. Plusieurs personnes réunies sont toujours plus efficaces qu'une personne qui fait tout (s'il arrivent à fonctionner en synergie, ce n'est pas toujours facile). D'autant plus que chacun peut être l'aiguillon de l'autre. Un statisticien peut nourrir la réflexion d'un expert métier. Ce dernier peut en retour lui proposer des pistes d'études. En tous les cas, impliquer le métier est absolument nécessaire pour obtenir son adhésion à l'analyse exploratoire des données.

Dans la section 5.3, l'ouvrage décrit le « workflow » du data scientist c.-à-d. les principales étapes d'une étude, partant de la définition des objectifs jusqu'au déploiement de la solution. Le processus n'est pas sans rappeler celui du data mining. Ce qui montre bien que le data science n'est pas une discipline totalement nouvelle mais plutôt une avancée liée, entres autres, à l'évolution de la nature des données et des technologies.

[8] Voir aussi T. Pontoroli, « Le data scientist, un ''mouton à 5 pattes'' au cœur des données », Clubic, avril 2014.

Le **chapitre 6** est consacré à l'exploration et la préparation des données. Le domaine est vaste, il est impossible de le traiter en profondeur en une quinzaine de pages. Les auteurs se contentent de donner des repères, en énumérant les caractéristiques des données (sources, formats, qualité), en décrivant des principales tâches liées à leur exploration rapide (visualisation, croisements), et en survolant la problématique de la préparation des données et des outils associés.

Pourtant la phase de préparation de données est très importante et mobilise près de 80% du temps total d'un projet de data mining[9], plus encore en data science de par la nature disparate des données. Il faudrait en réalité un ouvrage entier pour traiter la question. La multiplicité des sources induisent en particulier de nouveaux défis de fusion de données.

Le **chapitre 7** aborde la question des algorithmes de machine learning. Là aussi le domaine est vaste, il faudrait plusieurs ouvrages pour traiter la question. Consciemment ou pas, les auteurs mettent l'accent sur l'analyse prédictive. Plusieurs algorithmes sont succinctement décrits. Le lecteur dispose d'un panorama sur les méthodes les plus usitées (régression logistique, arbres de décision, forêts aléatoires, etc.).

Un passage très important a retenu mon attention (page 116) : « un des crédos qui sous-tend à l'heure actuelle une bonne part de l'industrie big data affirme que les meilleures performances sont obtenues au moyen d'algorithmes relativement peu sophistiqués qui opèrent sur de grandes bases de données, plutôt que l'inverse… ». Effectivement, les fournisseurs de solutions informatiques mettent l'accent sur la volumétrie. Il suffit de consulter la page d'IBM consacrée au big data pour s'en rendre compte. La première phrase commence par « chaque jour, nous générons 2,5 trillons d'octets de données… » (Définition du Big Data, IBM, consulté le 04/07/2015). Plus de données induit vraiment plus de connaissances ? Est-ce que lancer des algorithmes de machine learning sur des échantillons de données ne serait pas plus bénéfique ? On notera que la statistique inférentielle repose sur cette idée : travailler sur une fraction des données pour inférer (généraliser) sur la totalité. Est-ce que l'industrie du big data va redécouvrir l'inférence statistique bientôt ? La remarque est un peu ironique, je le concède. Les tenants du traitement direct sur de très grandes bases rétorquent souvent que le travail sur échantillons ne permet pas de détecter les phénomènes de niche. L'argument est pertinent.

[9] C. China, « La qualité des données au cœur des projets de data mining », emarketing.fr, n°79, mai 2003.

La section 7.4 enfin présente des études de cas menés à l'aide du logiciel Data Science Studio de la société Dataiku. La lecture de cette partie de l'ouvrage m'a poussé à étudier plus en profondeur les caractéristiques de l'outil dans un tutoriel à part[10].

Le **chapitre 8** aborde la question de la visualisation des données. Le domaine est plus compliqué qu'il n'en a l'air. En effet, la visualisation peut recouvrir des finalités différentes. Les graphiques peuvent servir de support pour mieux faire appréhender à des non statisticiens la teneur des résultats dans les rapports ou dans les présentations. Je le dis souvent à mes étudiants, aucun expert métier n'adhérera à ce que l'on fait s'il ne comprend pas la nature des conclusions que nous émettons.

Mais la visualisation peut agir également en amont, nous aiguiller dans notre recherche de solutions[11]. Notre cerveau dispose de capacités que n'ont pas les indicateurs numériques qui sont forcément réducteurs (un résumé est toujours réducteur, le tout est de savoir jusqu'à quel point). L'analyse exploratoire des données repose aussi sur l'intuition, c'est ce qui la rend si passionnante d'ailleurs. L'analyse en composantes principales par exemple, très populaire en France, correspond justement à ce schéma. Elle donne des résultats (corrélations, etc.), mais elle nourrit aussi la réflexion (existence potentielle de groupes d'observations, leur positionnement relatif, etc.).

Les auteurs proposent quelques principes de bonnes pratiques tirées de la statistique bidimensionnelle. Ils font aussi référence aux travaux de Bertin sur la sémiologie graphique. L'un des objectifs de la visualisation est la stimulation de l'imagination ! (page 160)

Le chapitre 10 présente l'**écosystème Hadoop** et débute la troisième partie relative aux **« outils du big data »**. Hadoop est un projet open source de la fondation Apache dont l'objectif est de développer des outils qui facilitent la construction d'applications scalables distribuées sur des clusters de serveurs bon marché (page 168). Il intègre plusieurs composants énumérés rapidement dans la section 9.2[12].

[10] Tutoriel Tanagra, « Data Science Studio », juin 2015.

[11] W. Jacoby, « Statistical Graphics for Univariate and Bivariate Data », Quantitative applications in the Social Science, vol. 117, Sage University Paper Series, 1997 ; « As a methodological tool, statistical graphics comprise a set of strategies and techniques that provide the researcher with important insights about the data under examination and help guide the subsequent steps of the research process », page 1.

[12] Voir aussi pour la même liste avec une description plus détaillée https://fr.wikipedia.org/wiki/Hadoop

L'installation de Hadoop est une tâche difficile, et surtout les différentes bibliothèques qui la constituent évoluent rapidement. Les versions ne sont pas toujours compatibles entre elles[13]. Heureusement, il existe des distributions clés en main qui permettent d'installer - relativement simplement - un ensemble cohérent. Les auteurs citent notamment Cloudera[14], Hortonworks, MapR et Amazon Elastic MapReduce. Cette dernière va plus loin car elle propose un service en mode cloud.

Les librairies de calcul attirent particulièrement notre attention dans un contexte de fouille de données massives. **Mahout** fournit une librairie Java d'implémentations parallélisables des principaux algorithmes de machine learning. Ces implémentations s'appuient sur le paradigme MapReduce.

Pour dépasser les insuffisances de ce dernier, en particulier sa lenteur et la complexité du code à produire pour les programmeurs. Apache introduit avec Spark[15,16] une alternative in-memory jusqu'à 100 fois plus rapide que le traditionnel MapReduce. Concrètement, les développeurs disposent d'un ensemble de routines (API) de haut niveau en Java, Scala, Python et R leur permettant de simplifier l'écriture du code. La bibliothèque de machine learning **MLlib** s'appuie sur Spark. Elle propose des algorithmes de statistiques, d'analyse prédictive, de classification automatique, de filtrage collaboratif, de réduction de dimension, etc. MLlib est annoncé nettement plus rapide que Mahout.

RHadoop de Revolution Analytics s'addresse spécifiquement aux développeurs sous R. Le logiciel R n'est pas prévu pour traiter de très grandes volumétries. **RHadoop** propose un ensemble de packages (librairies pour R) permettant aux développeurs d'accéder au système de fichier distribué de Hadoop (HDFS : hadoop distributed file system), d'accéder à la base distribuée HBase, et d'implémenter le paradigme MapReduce[17].

[13] Je peux en témoigner, j'ai essayé d'installer Hadoop sur un cluster mono-nœud sous Ubuntu. Je me suis inspiré du tutoriel suivant : http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/. Très rapidement j'ai du improviser. Nombre d'informations étaient obsolètes.

[14] L'installation de la distribution de la distribution Cloudera est décrite dans un de mes tutoriels. Tutoriel Tanagra, « Programmation R sous Hadoop », avril 2015.

[15] S. Leblal, « La fondation Apache réveille Hadoop avec Spark », Le Monde Informatique, Mai 2014.

[16] https://en.wikipedia.org/wiki/Apache_Spark

[17] Voir Tutoriels Tanagra : « MapReduce avec R », février 2015 ; « Programmation R sous Hadoop », avril 2015.

Le **chapitre 10** s'intéresse à l'analyse des logs avec Pig et Hive. Les logs sont des fichiers qui retracent l'activité sur un serveur web. Ils contiennent des informations de connexions telles que les pages visitées, les dates d'accès, l'identité des visiteurs à travers leur numéro IP ou leur cookie, etc. Ces informations peuvent être enrichies en les croisant avec des données externes (sources open data, localisation graphique avec les plages de numéro IP, etc.), nous pouvons également déduire d'autres indications au moyen de divers calculs (durée de connexion, distribution des pages visitées sur une période, etc.).

L'analyse des logs n'est pas nouvelle. Elle est mise en œuvre depuis que les serveurs web existent. L'originalité ici est l'utilisation des outils Pig et Hive. Ils permettent de créer des jobs MapReduce sur Hadoop sans qu'il soit nécessaire de les codes explicitement, en utilisant des langages similaires aux procédures stockées (Pig) ou au SQL (Hive) (page 188). La volumétrie n'est plus un goulot d'étranglement pour le traitement batch de fichiers logs qui peuvent être de taille considérable.

Le **chapitre 11** présente les architectures λ dédiées au traitement en temps réel. En effet, dans un contexte de forte vélocité des données, avec des mises à jour fréquentes, la rapidité de traitement est un atout fort. Elle devient cruciale dans certains domaines telles que la lutte contre la fraude. Pour répondre à cette attente, une architecture spécifique à 3 couches a été imaginée (pages 199 à 205).

La **couche batch** est lancée périodiquement, elle fonctionne selon le mode traditionnel des applications big data. La **couche de service** a pour rôle de rendre exploitable les résultats calculés en batch. La **couche de vitesse** est chargée de traiter les nouvelles de données au fur et à mesure de leur arrivé. Elle fonctionne en continu et maintient sa propre base de données qui est de taille limitée, sachant que ces mêmes données sont également envoyées aux bases sources du traitement batch. Lorsque ce dernier est déclenché, intégrant les données récentes, les résultats mis à jour sont rendus disponibles par la couche de service, les données de la couche de vitesse sont supprimées.

La cohérence du dispositif repose sur le mécanisme de fusion. Lorsqu'une nouvelle requête arrive, le système interroge et consolide les résultats issus des bases des couches de vitesse et de service. Nous sommes sûrs de disposer des informations les plus récentes.

Le **chapitre 12** est consacré à Apache Storm. Il s'agit d'un framework qui permet d'organiser les traitements temps réel sur les flux de données. Il fournit tous les mécanismes de base de la couche de vitesse. Un exemple de comptage de mots à l'intérieur de fichiers texte est utilisé pour expliquer le mode opératoire de Storm.

Bien évidemment, il est impossible de tout aborder en profondeur dans un ouvrage généraliste. Les deux derniers chapitres 11 et 12 nous aident surtout à appréhender les codes de compréhension du temps réel.

En **conclusion**, je dirais que le principal mérite de cet ouvrage est de nous proposer une vision globale d'un domaine aussi dynamique que le big data. Chacun a tendance à voir midi à sa porte. Pour ma part, j'ai été surtout intéressé par les parties relatives aux technologies ; je conçois aisément qu'un décideur soit plus curieux par rapport à la place du big data dans les organisations ; l'étudiant en statistique serait plutôt attiré par la partie consacrée au métier de data scientist, il doit pouvoir se situer lorsqu'il présentera devant un recruteur. La trame cohérente proposée par l'ouvrage nous aide à nous positionner.