Big Data et Machine Learning

Les concepts et les outils de la data science

Pirmin Lemberger
Data scientist en charge du data lab chez Weave Business Technology

Marc Batty
Cofondateur de Dataiku

Médéric Morel
Cofondateur et CEO de Mapwize

Jean-Luc Raffaëlli
Directeur de projets stratégiques au sein de la DSI du groupe La Poste

*Préface d'***Aurélien Géron**

2e édition

DUNOD

Illustration de couverture : skyline à Shanghai, Chine
© Oleksandr Dibrova – Fotolia.com

© Dunod, 2015, 2016
11 rue Paul Bert, 92240 Malakoff
www.dunod.com

ISBN 978-2-10-075463-2

Préface

En 2006, Geoffrey Hinton et son équipe de l'université de Toronto publièrent un article intitulé *A fast learning algorithm for deep belief nets* (« Un algorithme rapide pour les réseaux bayésiens profonds ») dans lequel ils montraient comment entraîner un système à reconnaître des caractères manuscrits avec une grande précision. Ce système n'atteignait pas tout à fait la performance des meilleurs systèmes spécialisés de l'époque, mais son architecture reposait sur un réseau de neurones « profond », c'est-à-dire organisé en de nombreuses couches superposées (ce que Hinton a appelé le Deep Learning) : cette architecture généraliste portait la promesse de pouvoir s'adapter à toutes sortes de tâches complexes. Les réseaux profonds avaient été délaissés depuis les années 1990 car les chercheurs ne parvenaient pas à les faire fonctionner correctement. Cet article a donc eu l'effet d'un séisme, il a déclenché un regain d'intérêt dans les réseaux de neurones profonds et plus généralement dans tous les algorithmes qui permettent aux machines d'apprendre une tâche simplement à partir d'exemples, ce que l'on appelle le Machine Learning.

En quelques années, la vague du Machine Learning s'est muée en véritable tsunami. Nous utilisons tous quotidiennement le Machine Learning sans même le savoir, par exemple :

- lorsque nous trouvons facilement ce que nous cherchons sur le web ;
- quand un site nous recommande un contenu qui nous plaît ;
- lorsqu'une publicité bien ciblée nous mène à faire un achat ;
- quand on effectue une recherche vocale ou une traduction automatique ;
- ou encore quand notre appareil photo reconnaît les visages de nos proches.

Outre ces applications du quotidien, il ne se passe désormais plus une semaine sans que le Machine Learning soit cité dans les médias pour des applications aussi différentes que :

- le programme AlphaGo, qui a vaincu le champion du monde du jeu de Go à plate couture (4 manches sur 5) ;
- les voitures autonomes, dont beaucoup disent qu'elles révolutionneront la circulation d'ici à quelques années ;

- les robots de la société Boston Dynamics, qui semblent presque vivants ;
- les implants cérébraux, dont les signaux sont interprétés par un système à base de réseaux de neurones artificiels et qui permettent à un tétraplégique de bouger à nouveau sa main.

Après la révolution du mobile et du web, voici donc venue la révolution du Machine Learning… et du Big Data. Lorsqu'on demande à Geoffrey Hinton d'expliquer ce qui lui a permis de réussir là où tant d'autres avaient échoué dix ans auparavant, il répond (avec beaucoup de modestie) qu'entre-temps le volume de données disponibles a augmenté de façon phénoménale, ainsi que la puissance de calcul, l'espace de stockage et les techniques nécessaires pour gérer ce volume de données. Le Big Data serait donc le véritable déclencheur de la révolution du Machine Learning.

C'est ainsi qu'en quelques années les géants du web, de Google à Facebook en passant par Amazon, Apple ou Yahoo!, sont tous passés au Machine Learning. Avec leurs datacenters aux quatre coins de la planète, ces géants disposent d'une puissance de calcul inimaginable. Leurs services sont utilisés par des milliards de personnes chaque jour, et à chacun de nos clics ils récupèrent un peu plus d'information sur nos comportements et nos goûts. On estime que Google, Facebook et Amazon à eux seuls stockent plus de 30 exaoctets de données (en 2016), soit 30 milliards de gigaoctets, répartis sur plusieurs millions de serveurs. D'abord leaders du web, ils sont devenus leaders du Big Data et désormais leaders du Machine Learning.

Mais la révolution du Big Data et du Machine Learning n'est pas réservée aux géants du web. Même si peu d'entreprises parlent encore en exaoctets, beaucoup atteignent les téraoctets (milliers de gigaoctets), voire les pétaoctets (millions de gigaoctets), qu'il s'agisse de logs de sites web ou d'applications, de données de capteurs industriels, de cours de bourse, de statistiques de centrales d'appels, de dossiers clients, etc. Comment exploiter au mieux ces données ? Peut-on prédire quels prospects seront les plus susceptibles d'acheter ? Comment optimiser la qualité d'une chaîne de montage ? Détecter les anomalies de production ? Prédire le cours d'une matière première ? Segmenter les clients et mieux cibler les offres ? Autant de questions auxquelles le Machine Learning peut contribuer à répondre.

Cependant le sujet reste très récent et les compétences rares. Le data scientist est devenu un profil très recherché. Beaucoup de développeurs sont fascinés par les prouesses du Machine Learning et veulent apprendre. Les DSI cherchent la meilleure façon de restructurer leurs systèmes et leurs équipes pour gérer au mieux un volume grandissant de données et parvenir à en extraire toutes les pépites en temps réel. Chacun cherche donc comment tirer parti au mieux de cette révolution (ou comment y survivre). Mais beaucoup se sentent un peu perdus : le domaine est vaste, et on ne sait par où commencer.

J'ai eu la chance de faire partie des relecteurs de la première édition de ce livre, et j'ai été immédiatement séduit par le large panorama qu'il offre : que vous soyez DSI, développeur, chef de projet ou simplement curieux, ce livre vous apportera une vision claire des enjeux et des principaux concepts du Big Data et du Machine Learning. Mais il s'adresse également à un public technique en introduisant les principaux algorithmes. Certes un développeur devra compléter sa lecture par une formation

technique et par la pratique (des conseils à ce sujet ont été rajoutés en fin de livre dans cette édition), mais il aura déjà une bonne compréhension du sujet.

Cette seconde édition que vous tenez entre les mains aborde un certain nombre de thèmes qui étaient absents de la première édition, en particulier les réseaux de neurones et le Deep Learning. En outre, certaines parties ont été remodelées ou approfondies, telles que Spark et les moteurs de recommandations, les Data Layers ou encore la transformation du SI.

J'espère que vous aurez autant de plaisir que moi à lire ce livre, et qu'il vous encouragera à vous lancer dans cette révolution extraordinaire.

Aurélien GÉRON

Consultant en Machine Learning,
ancien responsable de la classification des vidéos chez YouTube
et cofondateur de la société Wifirst

Table des matières

Deuxième partie – Le métier de data scientist

Avant-propos

Pourquoi un ouvrage sur le Big Data ?

Le Big Data est un phénomène aux multiples facettes qui fait beaucoup parler de lui mais dont il est difficile de bien comprendre les tenants et aboutissants. Il est notamment difficile de prévoir quel sera son impact sur les acteurs et sur les métiers de la DSI.

Cet ouvrage se veut un guide pour comprendre les enjeux des projets d'analyse de données, pour appréhender les concepts sous-jacents, en particulier le Machine Learning et acquérir les compétences nécessaires à la mise en place d'un data lab. Il combine la présentation des concepts théoriques de base (traitement statistique des données, calcul distribué), la description des outils (Hadoop, Storm) et des retours d'expérience sur des projets en entreprise.

Sa finalité est d'accompagner les lecteurs dans leurs premiers projets Big Data en leur transmettant la connaissance et l'expérience des auteurs.

À qui s'adresse ce livre ?

Ce livre s'adresse particulièrement à celles et ceux qui, curieux du potentiel du Big Data dans leurs secteurs d'activités, souhaitent franchir le pas et se lancer dans l'analyse de données. Plus spécifiquement, il s'adresse :

- aux décideurs informatiques qui souhaitent aller au-delà des discours marketing et mieux comprendre les mécanismes de fonctionnement et les outils du Big Data ;
- aux professionnels de l'informatique décisionnelle et aux statisticiens qui souhaitent approfondir leurs connaissances et s'initier aux nouveaux outils de l'analyse de données ;
- aux développeurs et architectes qui souhaitent acquérir les bases pour se lancer dans la *data science* ;
- aux responsables métier qui veulent comprendre comment ils pourraient mieux exploiter les gisements de données dont ils disposent.

Des rudiments de programmation et des connaissances de base en statistiques sont cependant nécessaires pour bien tirer parti du contenu de cet ouvrage.

Comment lire ce livre ?

Ce livre est organisé en trois parties autonomes qui peuvent théoriquement être lues séparément. Nous recommandons néanmoins au lecteur d'accorder une importance particulière au chapitre 3 (le mouvement NoSQL) et au chapitre 4 (l'algorithme MapReduce).

La première partie commence par traiter des origines du Big Data et de son impact sur les organisations. Elle se prolonge par la présentation du mouvement NoSQL et de l'algorithme MapReduce.

La deuxième partie est consacrée au métier de *data scientist* et aborde la question de la préparation des jeux de données, les bases du Machine Learning ainsi que la visualisation des données.

La troisième partie traite du passage à l'échelle du Big Data avec la plateforme Hadoop et les outils tels que Hive et Pig. On présente ensuite un nouveau concept appelé architecture λ qui permet d'appliquer les principes du Big Data aux traitements en temps réel.

Travaux pratiques

À plusieurs reprises dans cet ouvrage, le logiciel Data Science Studio est utilisé afin d'illustrer et de rendre plus concret le contenu. Cet outil, développé par la startup française Dataiku, fournit un environnement complet et intégré pour la préparation des données et le développement de modèles de Machine Learning.

Le chapitre 7 est ainsi illustré avec des exemples traités avec Data Science Studio.

Vous pouvez retrouver les jeux de données ainsi qu'une version de démonstration du logiciel à l'adresse suivante : www.dataiku.com/livre-big-data.

Remerciements

Les auteurs tiennent tout d'abord à remercier leurs proches pour leur patience et leur soutien pendant les périodes de rédaction de cet ouvrage. Leur reconnaissance va aussi à Nicolas Larousse, directeur de l'agence SQLI de Paris, à Olivier Reisse, associé et directeur de Weave Business Technology, ainsi qu'à Florian Douetteau, cofondateur et directeur général de Dataiku.

Ils remercient leurs collègues, amis et clients qui ont bien voulu relire l'ouvrage et aider à le compléter par leurs retours d'expérience, en particulier Manuel Alves et Étienne Mercier.

Enfin, les auteurs remercient tout particulièrement Pierre Pfennig pour sa contribution sur le Machine Learning, Jérémy Grèze pour son aide sur la préparation des données et Pierre Gutierrez pour sa participation sur Pig, Hive et les parcours clients.

PREMIÈRE PARTIE

Les fondements du Big Data

Cette première partie décrit les origines du Big Data sous les angles économiques, sociétaux et technologiques. Comment et pourquoi une nouvelle classe d'outils a-t-elle émergé ces dix dernières années ?

- Le premier chapitre explique comment le rattachement de la valeur à l'information en général plutôt qu'aux seules données structurées, et la baisse des coûts des ressources IT de plusieurs ordres de grandeur ont fait progressivement émerger le **Big Data**.
- Le chapitre 2 traite de l'impact du Big Data dans les organisations et présente la **caractérisation dite des 3V**. Il montre en quoi le Big Data n'est pas, loin s'en faut, un défi uniquement technique.
- Le chapitre 3 décrit l'émergence d'une nouvelle classe de systèmes de stockage : les **bases de données NoSQL**. Après avoir analysé les limites du modèle relationnelle classique face aux nouvelles exigences de performance et de disponibilité des applications web à très grande échelle, une classification de ces nouveaux systèmes sera proposée.
- Le chapitre 4 propose un zoom sur **MapReduce**, un schéma de parallélisation massive des traitements, introduit il y a une dizaine d'années par Google, qui est au cœur de beaucoup de projets Big Data. Des exemples d'usage de MapReduce seront décrits. Les limitations de ce modèle seront discutées avant d'esquisser les évolutions futures qui essaient de les surmonter.

1

Les origines du Big Data

Objectif

Au commencement de l'informatique était la donnée. Le Big Data refocalise l'attention sur l'*information* en général et non plus sur la seule donnée structurée ce qui ouvre la voie à des usages inédits. Ce chapitre dresse un premier panorama des origines et des éléments fondamentaux de l'approche Big Data qui permettent d'accéder à la notion de valeur de l'information.

1.1 LA PERCEPTION DE LA DONNÉE DANS LE GRAND PUBLIC

1.1.1 La révolution de l'usage

Depuis le début de l'informatique personnelle dans les années 1980, jusqu'à l'omniprésence du web actuelle dans la vie de tous les jours, les données ont été produites en quantités toujours croissantes. Photos, vidéos, sons, textes, logs en tout genre... Depuis la démocratisation d'Internet, ce sont des volumes impressionnants de données qui sont créés quotidiennement par les particuliers, les entreprises et maintenant aussi les objets et machines connectés.

Désormais, le terme « Big Data », littéralement traduit par « grosses données » ou « données massives » désigne cette explosion de données. On parle également de « datamasse » en analogie avec la biomasse, écosystème complexe et de large échelle.

À titre d'exemple, le site *Planetoscope* (http://www.planetoscope.com) estime à 3 millions le nombre d'e-mails envoyés dans le monde chaque seconde, soit plus de

200 milliards par jour en comptant les spams qui représentent presque 90 % des flux, et ce pour une population d'internautes qui avoisine les 2,5 milliards d'individus. En 2010, le zettaoctet de données stockées dans le monde a été dépassé et on prévoit en 2020 10 Zo (zettaoctets), soit 10 400 milliards de gigaoctets de données déversés tous les mois sur Internet.

1.1.2 L'envolée des données

Dans les domaines des systèmes d'informations et du marketing, la presse et les campagnes de mails regorgent de propositions de séminaires ou de nouvelles offres de solutions Big Data qui traduisent un réel engouement pour le sujet.

Comme mentionné précédemment, ce déluge d'informations n'est pas seulement imputable à l'activité humaine. De plus en plus connectées, les machines contribuent fortement à cette augmentation du volume de données. Les stations de production énergétiques, les compteurs en tout genre et les véhicules sont de plus en plus nombreux à être équipés de capteurs ou d'émetteurs à cartes SIM pour transférer des informations sur leur milieu environnant, sur les conditions atmosphériques, ou encore sur les risques de défaillance.

Même l'équipement familial est concerné par l'intermédiaire d'un électroménager intelligent et capable d'accompagner la vie de tous les jours en proposant des services de plus en plus performants (pilotage du stock, suggestion d'entretien, suivi de régime, etc.).

À cette production et cet échange massif de données s'ajoutent les données libérées par les organisations et les entreprises, désignées sous le nom d'open data : horaires de transports en commun, statistiques sur les régions et le gouvernement, réseau des entreprises, données sur les magasins...

1.1.3 Un autre rapport à l'informatique

Le nombre de données produites et stockées à ce jour est certes important mais l'accélération du phénomène est sans précédent ces cinq dernières années. Cette accélération est principalement due à un changement dans nos habitudes : ce que nous attendons des ordinateurs a changé et la démocratisation des smartphones et des tablettes ainsi que la multiplication des réseaux sociaux encouragent les échanges et la création de nouveaux contenus. La croissance du volume de données produites suit donc des lois exponentielles.

Ce volume impressionnant est à mettre en relation avec la consumérisation de l'informatique et avec les campagnes des grands du web (Apple, Google, Amazon...) visant à encourager un usage toujours croissant de leurs services. Cette hausse de la consommation de leurs services se traduit mécaniquement par une demande croissante de puissance de traitement et de stockage de données qui engendre une obsolescence rapide des architectures IT habituellement utilisées : bases de données relationnelles et serveurs d'applications doivent laisser la place à des solutions nouvelles.

1.1.4 L'extraction de données ou d'information ?

La présence de volumes de données importants est devenue, presque inconsciemment, une valeur rassurante pour les utilisateurs. Ainsi nombre d'entreprises sont restées confiantes sur la valeur de leurs bases de données, considérant que leur seule taille constituait en soi un bon indicateur de leur valeur pour le marketing.

Pour autant ce qui est attendu de ces données, c'est la connaissance et le savoir (c'est-à-dire le comportement d'achat des clients). Il est assez paradoxal de constater que les acteurs qui communiquent le plus sur l'utilisation des données produites sur le web se gardent généralement d'aborder le sujet non moins important du ratio pertinence / volume.

Il y a donc une fréquente confusion entre la donnée et l'information qu'elle contient, le contexte de la création de la donnée et celui de son utilisation qui viennent enrichir à leur tour cette information. C'est là tout le champ d'investigation du Big Data.

1.2 DES CAUSES ÉCONOMIQUES ET TECHNOLOGIQUES

Les deux principales causes de l'avènement du Big Data ces dernières années sont à la fois économiques et technologiques.

1.2.1 Une baisse des prix exponentielle

Durant ces vingt dernières années le prix des ressources IT a chuté de manière exponentielle en accord avec la célèbre loi de Moore[1]. Qu'il s'agisse de la capacité de stockage, du nombreux de nœuds que l'on peut mettre en parallèle dans un data center, de la fréquence des CPU ou encore de la bande passante disponible (qui a favorisé l'émergence des services cloud). La figure 1.1 représente schématiquement ces évolutions.

La mise en place par des géants du web comme *Google*, *Amazon*, *LinkedIn*, *Yahoo!* ou *Facebook* de data center de plusieurs dizaines de milliers de machines bon marché a constitué un facteur déterminant dans l'avènement des technologies Big Data. Ce qui nous amène naturellement au second facteur.

1. La version la plus courante de cette « loi », qui est en réalité plutôt une conjecture, stipule que des caractéristiques comme la puissance, la capacité de stockage ou la fréquence d'horloge doublent tous les 18 mois environ.