

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

2^e édition

DUNOD

Toutes les marques citées dans cet ouvrage sont des marques déposées par leurs propriétaires respectifs.

Data Science Studio est une marque déposée de Dataiku.

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

<p>Le pictogramme qui figure ci-contre mérite une explication. Son objet est d'alerter le lecteur sur la menace que représente pour l'avenir de l'écrit, particulièrement dans le domaine de l'édition technique et universitaire, le développement massif du photocopillage.</p> <p>Le Code de la propriété intellectuelle du 1^{er} juillet 1992 interdit en effet expressément la photocopie à usage collectif sans autorisation des ayants droit. Or, cette pratique s'est généralisée dans les établissements</p>		<p>d'enseignement supérieur, provoquant une baisse brutale des achats de livres et de revues, au point que la possibilité même pour les auteurs de créer des œuvres nouvelles et de les faire éditer correctement est aujourd'hui menacée.</p> <p>Nous rappelons donc que toute reproduction, partielle ou totale, de la présente publication est interdite sans autorisation de l'auteur, de son éditeur ou du Centre français d'exploitation du droit de copie (CFC, 20, rue des Grands-Augustins, 75006 Paris).</p>
--	---	--

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

ISBN 978-2-10-075463-2

Le Code de la propriété intellectuelle n'autorisant, aux termes de l'article L. 122-5, 2° et 3° a), d'une part, que les « copies ou reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective » et, d'autre part, que les analyses et les courtes citations dans un but d'exemple et d'illustration, « toute représentation ou reproduction intégrale ou partielle faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause est illicite » (art. L. 122-4).

Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait donc une contrefaçon sanctionnée par les articles L. 335-2 et suivants du Code de la propriété intellectuelle.

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

Préface	III
---------------	-----

Avant-propos	XV
--------------------	----

Première partie – Les fondements du Big Data

Chapitre 1 – Les origines du Big Data	3
1.1 La perception de la donnée dans le grand public	3
1.1.1 <i>La révolution de l'usage</i>	3
1.1.2 <i>L'envolée des données</i>	4
1.1.3 <i>Un autre rapport à l'informatique</i>	4
1.1.4 <i>L'extraction de données ou d'information ?</i>	5
1.2 Des causes économiques et technologiques	5
1.2.1 <i>Une baisse des prix exponentielle</i>	5
1.2.2 <i>Des progrès initiés par les géants du web</i>	6
1.2.3 <i>Où se trouve la frontière du Big Data ?</i>	7
1.3 La donnée et l'information	8
1.3.1 <i>La recherche pertinente</i>	8
1.3.2 <i>Un avantage concurrentiel</i>	8
1.3.3 <i>Des clients plus exigeants</i>	9
1.4 La valeur	9
1.5 Les ressources nécessaires	10

1.6	De grandes opportunités	11
Chapitre 2 – Le Big Data dans les organisations		13
2.1	La recherche de l'Eldorado	13
2.1.1	<i>L'entreprise dans un écosystème</i>	13
2.1.2	<i>Une volonté de maîtrise</i>	14
2.1.3	<i>Des besoins forts</i>	14
2.2	L'avancée par le cloud	14
2.3	La création de la valeur	15
2.4	Les « 3V » du Big Data	15
2.4.1	<i>Le volume</i>	16
2.4.2	<i>La vélocité</i>	16
2.4.3	<i>La variété</i>	16
2.5	Un champ immense d'applications	17
2.6	Exemples de compétences à acquérir	19
2.6.1	<i>Appréhender de nouveaux modèles de traitement des données</i>	19
2.6.2	<i>Maîtriser le déploiement de Hadoop ou utiliser une solution cloud</i>	20
2.6.3	<i>Se familiariser avec de nouvelles méthodes de modélisation</i>	20
2.6.4	<i>Découvrir de nouveaux outils d'analyse de données</i>	21
2.7	Des impacts à tous les niveaux	22
2.7.1	<i>Impacts sur la conception des systèmes</i>	22
2.7.2	<i>Une nécessaire intégration du Big Data dans le SI</i>	22
2.7.3	<i>Un élargissement des champs d'investigation</i>	22
2.7.4	<i>Valorisation de la donnée, pilier de la transformation</i>	23
2.7.5	<i>Un potentiel reposant sur plusieurs composantes SI</i>	23
2.7.6	<i>Une disposition naturelle à partager</i>	24
2.7.7	<i>Toujours plus de métier</i>	25
2.7.8	<i>Conséquences sur l'organisation de l'entreprise</i>	25
2.7.9	<i>Impacts sur les relations entre clients et fournisseurs</i>	26
2.7.10	<i>Implications juridiques</i>	26
2.8	« B » comme Big Data ou Big Brother ?	26
2.8.1	<i>Connaissance client et préservation de la vie privée</i>	26
2.8.2	<i>La lassitude est à notre porte</i>	27

2.8.3	<i>Vers une démarche active</i>	27
Chapitre 3 – Le mouvement NoSQL		29
3.1	Bases relationnelles, les raisons d'une domination	29
3.2	Le dogme remis en question	34
3.2.1	<i>Les contraintes des applications web à très grande échelle</i>	34
3.2.2	<i>Le « théorème » CAP</i>	35
3.2.3	<i>Sacrifier la flexibilité pour la vélocité</i>	37
3.2.4	<i>Peut-on définir ce qu'est une base de données NoSQL ?</i>	39
3.3	Les différentes catégories de solutions	40
3.3.1	<i>Les entrepôts clé-valeur</i>	40
3.3.2	<i>Les bases orientées documents</i>	42
3.3.3	<i>Les bases orientées colonnes</i>	44
3.3.4	<i>Les bases de données orientées graphes</i>	49
3.4	Le NoSQL est-il l'avenir des bases de données ?	50
Chapitre 4 – L'algorithme MapReduce et le framework Hadoop		53
4.1	Automatiser le calcul parallèle	53
4.2	Le pattern MapReduce	54
4.3	Des exemples d'usage de MapReduce	58
4.3.1	<i>Analyse statistique d'un texte</i>	58
4.3.2	<i>Calcul d'une jointure entre deux grandes tables</i>	59
4.3.3	<i>Calcul du produit de deux matrices creuses</i>	62
4.4	Le framework Hadoop	63
4.4.1	<i>Planning des exécutions</i>	64
4.4.2	<i>Tolérance aux pannes</i>	65
4.4.3	<i>Découpage des données en lots</i>	66
4.4.4	<i>Fusion et tri des listes intermédiaires</i>	66
4.4.5	<i>Monitoring des processus</i>	68
4.5	Au-delà de MapReduce	68

Deuxième partie – Le métier de data scientist

Chapitre 5 – Le quotidien du data scientist	73
5.1 Data scientist : licorne ou réalité ?	73
5.1.1 <i>L'origine du terme data scientist et définitions courantes</i>	73
5.1.2 <i>Les compétences clés du data scientist</i>	75
5.1.3 <i>Comment recruter ou se former</i>	79
5.2 Le data scientist dans l'organisation	80
5.2.1 <i>Le data lab – une clé pour l'innovation par la donnée</i>	80
5.2.2 <i>Le data lab – quelle place dans l'organisation ?</i>	82
5.3 Le workflow du data scientist	82
5.3.1 <i>Imaginer un produit ou un service</i>	83
5.3.2 <i>Collecte des données</i>	85
5.3.3 <i>Préparation</i>	86
5.3.4 <i>Modélisation</i>	87
5.3.5 <i>Visualisation</i>	88
5.3.6 <i>Optimisation</i>	89
5.3.7 <i>Déploiement</i>	90
Chapitre 6 – Exploration et préparation de données	91
6.1 Le déluge des données	91
6.1.1 <i>Diversité des sources</i>	92
6.1.2 <i>Diversité des formats</i>	94
6.1.3 <i>Diversité de la qualité</i>	95
6.2 L'exploration de données	96
6.2.1 <i>Visualiser pour comprendre</i>	96
6.2.2 <i>Enquêter sur le passé des données</i>	97
6.2.3 <i>Utiliser les statistiques descriptives</i>	98
6.2.4 <i>Les tableaux croisés dynamiques</i>	99
6.3 La préparation de données	101
6.3.1 <i>Pourquoi préparer ?</i>	101
6.3.2 <i>Nettoyer les données</i>	101
6.3.3 <i>Transformer les données</i>	102
6.3.4 <i>Enrichir les données</i>	103

6.3.5	<i>Un exemple de préparation de données</i>	105
6.4	Les outils de préparation de données	106
6.4.1	<i>La programmation</i>	106
6.4.2	<i>Les ETL</i>	107
6.4.3	<i>Les tableurs</i>	107
6.4.4	<i>Les outils de préparation visuels</i>	107
Chapitre 7 – Le Machine Learning		109
7.1	Qu'est-ce que le Machine Learning ?	109
7.1.1	<i>Comprendre ou prédire ?</i>	109
7.1.2	<i>Qu'est-ce qu'un bon algorithme de Machine Learning ?</i>	114
7.1.3	<i>Performance d'un modèle et surapprentissage</i>	114
7.1.4	<i>Machine Learning et Big Data – sur quoi faut-il être vigilant ?</i>	117
7.2	Les différents types de Machine Learning	119
7.2.1	<i>Apprentissage supervisé ou non supervisé ?</i>	119
7.2.2	<i>Régression ou classification ?</i>	120
7.2.3	<i>Algorithmes linéaires ou non linéaires ?</i>	120
7.2.4	<i>Modèle paramétrique ou non paramétrique ?</i>	120
7.2.5	<i>Apprentissage hors ligne ou incrémental ?</i>	121
7.2.6	<i>Modèle géométrique ou probabiliste ?</i>	121
7.3	Les principaux algorithmes	122
7.3.1	<i>La régression linéaire</i>	122
7.3.2	<i>Les k plus proches voisins</i>	124
7.3.3	<i>La classification naïve bayésienne</i>	125
7.3.4	<i>La régression logistique</i>	126
7.3.5	<i>L'algorithme des k-moyennes</i>	128
7.3.6	<i>Les arbres de décision</i>	129
7.3.7	<i>Les forêts aléatoires</i>	132
7.3.8	<i>Les machines à vecteurs de support</i>	134
7.3.9	<i>Techniques de réduction dimensionnelle</i>	135
7.4	Réseaux de neurones et Deep Learning	136
7.4.1	<i>Les premiers pas vers l'intelligence artificielle</i>	136
7.4.2	<i>Le perceptron multicouche</i>	137
7.4.3	<i>L'algorithme de rétropropagation</i>	141

7.4.4	La percée du Deep Learning	143
7.4.5	Exemples d'architectures profondes	147
7.5	Illustrations numériques	152
7.5.1	Nettoyage et enrichissement des données	153
7.5.2	Profondeur d'un arbre et phénomène de surapprentissage	154
7.5.3	Apport du « feature engineering »	157
7.5.4	Sensibilité de l'algorithme KNN au bruit	161
7.5.5	Interprétabilité de deux modèles	162
7.5.6	Bénéfices de l'approche ensembliste	163
7.6	Systèmes de recommandation	163
7.6.1	Approches type Collaborative-Filtering	164
7.6.2	Approches type Content-Based	169
7.6.3	Approche Hybride	171
7.6.4	Recommandation à chaud : « Multi-armed bandit »	171
Chapitre 8 – La visualisation des données		173
8.1	Pourquoi visualiser l'information ?	173
8.1.1	Ce que les statistiques ne disent pas	173
8.1.2	Les objectifs de la visualisation	176
8.2	Quels graphes pour quels usages ?	177
8.3	Représentation de données complexes	184
8.3.1	Principes d'encodage visuel	184
8.3.2	Principes de visualisation interactive	186

Troisième partie – Les outils du Big Data

Chapitre 9 – L'écosystème Hadoop		193
9.1	La jungle de l'éléphant	194
9.1.1	Distribution ou package	194
9.1.2	Un monde de compromis	195
9.1.3	Les services autour de Hadoop	196
9.2	Les composants d'Apache Hadoop	196
9.2.1	Hadoop Distributed File System	197
9.2.2	MapReduce et YARN	198

9.2.3	<i>HBase</i>	199
9.2.4	<i>ZooKeeper</i>	199
9.2.5	<i>Pig</i>	201
9.2.6	<i>Hive</i>	202
9.2.7	<i>Oozie</i>	202
9.2.8	<i>Flume</i>	202
9.2.9	<i>Sqoop</i>	202
9.3	Les principales distributions Hadoop	203
9.3.1	<i>Cloudera</i>	203
9.3.2	<i>Hortonworks</i>	204
9.3.3	<i>MapR</i>	204
9.3.4	<i>Amazon Elastic MapReduce</i>	205
9.4	Spark ou la promesse du traitement Big Data in-memory	206
9.4.1	<i>L'émergence de Spark</i>	206
9.4.2	<i>De MapReduce à Spark</i>	207
9.4.3	<i>Les RDD au cœur du projet Spark</i>	208
9.4.4	<i>La simplicité et flexibilité de programmation avec Spark</i>	210
9.4.5	<i>Modes de travail en cluster</i>	211
9.5	Les briques analytiques à venir	212
9.5.1	<i>Impala versus Stinger</i>	212
9.5.2	<i>Drill</i>	212
9.6	Les bibliothèques de calcul	214
9.6.1	<i>Mahout</i>	214
9.6.2	<i>MMLib de Spark</i>	215
9.6.3	<i>RHadoop</i>	217
	Chapitre 10 – Analyse de logs avec Pig et Hive	219
10.1	Pourquoi analyser des logs ?	219
10.2	Pourquoi choisir Pig ou Hive ?	220
10.3	La préparation des données	221
10.3.1	<i>Le format des lignes de logs</i>	222
10.3.2	<i>L'enrichissement des logs</i>	222
10.3.3	<i>La reconstruction des sessions</i>	224
10.3.4	<i>Agrégations et calculs</i>	224

10.4	L'analyse des parcours clients	226
Chapitre 11	– Les architectures λ	229
11.1	Les enjeux du temps réel	229
11.1.1	<i>Qu'est-ce que le temps réel ?</i>	229
11.1.2	<i>Quelques exemples de cas réels</i>	230
11.2	Rappels sur MapReduce et Hadoop	231
11.3	Les architectures λ	231
11.3.1	<i>La couche batch</i>	232
11.3.2	<i>La couche de service</i>	233
11.3.3	<i>La couche de vitesse</i>	234
11.3.4	<i>La fusion</i>	235
11.3.5	<i>Les architectures λ en synthèse</i>	236
Chapitre 12	– Apache Storm	239
12.1	Qu'est-ce que Storm ?	239
12.2	Positionnement et intérêt dans les architectures λ	240
12.3	Principes de fonctionnement	241
12.3.1	<i>La notion de tuple</i>	241
12.3.2	<i>La notion de stream</i>	241
12.3.3	<i>La notion de spout</i>	242
12.3.4	<i>La notion de bolt</i>	242
12.3.5	<i>La notion de topologie</i>	243
12.4	Un exemple très simple	244
Conclusion	247
Index	251

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 Moore¹. Qu'il s'agisse de la capacité de stockage, du nombre 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.