BIG DATA ET MACHINE LEARNING

Les concepts et les outils de la data science

Chez le même éditeur

Machine Learning avec Scikit-Learn
2e édition
Aurélien Géron
288 pages environ
Dunod, 2019

Deep Learning avec Keras et TensorFlow
2e édition
Aurélien Géron
392 pages environ
Dunod, 2019

Introduction au Machine Learning
Chloé-Agathe Azencott
240 pages
Dunod, 2018

BIG DATA ET MACHINE LEARNING

Les concepts et les outils de la data science

Pirmin Lemberger
Directeur scientifique
chez onepoint x weave

Marc Batty
Cofondateur de Dataiku

Médéric Morel
Cofondateur et CEO de Mapwize

Jean-Luc Raffaëlli
Architecte d'entreprise
au sein du groupe La Poste

3e édition

DUNOD

Illustration de couverture : © Dmitry Rukhlenko – 123RF

11 rue Paul Bert, 92240 Malakoff
www.dunod.com
ISBN 978-2-10-079037-1

TABLE DES MATIÈRES

DEUXIÈME PARTIE
Le métier de data scientist

TROISIÈME PARTIE
Les outils du Big Data

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 start-up 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 depuis le milieu des années 2000 ?

- ✓ 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 relationnel 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.

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'information 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 depuis 2010. 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.

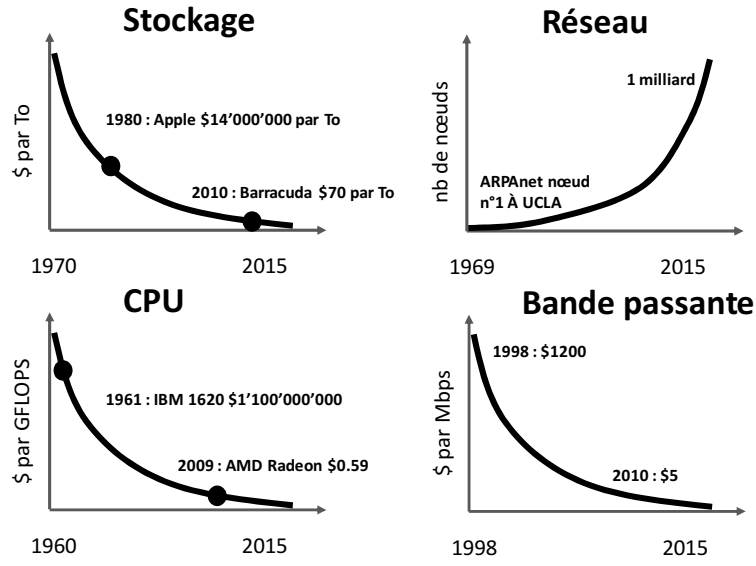
1. 2 DES CAUSES ÉCONOMIQUES ET TECHNOLOGIQUES

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

Depuis le début des années 2000, 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.

Figure 1.1 – Évolution des prix des ressources IT au cours des dernières décennies.



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.

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.

1.2.2 Des progrès initiés par les géants du web

Pour bénéficier de ces ressources de stockage colossales, les géants du web ont du développer pour leurs propres besoins de nouvelles technologies notamment en matière de parallélisation des traitements opérant sur des volumes de données se chiffrant en plusieurs centaines de téraoctets.

L'un des principaux progrès en la matière est venu de Google qui, pour traiter les données récupérées par les *crawlers* de son moteur de recherche et indexer la totalité du web, a mis au point un modèle de conception qui permet d'automatiser la parallélisation d'une grande classe de traitements. C'est le célèbre modèle **MapReduce** que nous décrirons en détail au chapitre 4.

De nombreuses avancées en génie logiciel développées à cette occasion ont par la suite essaimé dans la communauté open source qui en a proposé des systèmes équivalents gratuits. Le système de traitement parallèle **Hadoop Apache** est le principal exemple de ce transfert de technologie vers le monde open source. L'écosystème qui l'accompagne, notamment les systèmes de base de données non relationnelle comme *HBase*, le système de fichiers distribués *HDFS*, les langages de transformation et de requêtage *Pig* et *Hive* seront décrits au chapitre 9.

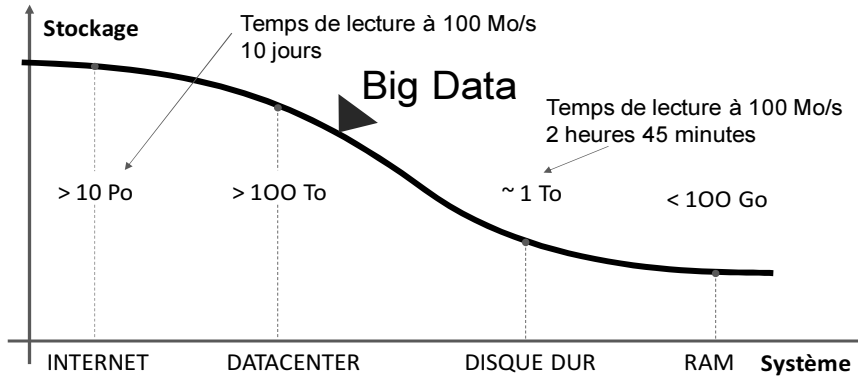
Face aux nouvelles exigences de montée en charge et de disponibilité, une nouvelle classe de système de gestion de base de données non relationnelle a émergé. On désigne habituellement ces systèmes au moyen du sigle NoSQL. Nous les décrirons en détail au chapitre 3 après avoir expliqué pourquoi et dans quelles circonstances ils se sont progressivement imposés face aux bases relationnelles.

1.2.3 Où se trouve la frontière du Big Data ?

Donner une définition exacte d'un terme aux contours aussi flous que le Big Data est délicat. On peut toutefois s'en faire une bonne idée en considérant les différents ordres de grandeur d'espaces de stockage représentés sur la figure 1.2.

On s'accorde généralement pour situer la frontière des volumes qui relèvent du Big Data à partir du moment où ces données ne peuvent plus être traitées en un temps « raisonnable » ou « utile » par des systèmes constitués d'un seul nœud. À titre d'exemple si l'on doit traiter ou analyser un téraoctet de données, qui correspond à la taille d'un disque dur standard, en quelques minutes il faudra impérativement recourir à une mise en parallèle des traitements et du stockage sur plusieurs nœuds. Selon cette définition le Big Data n'est pas uniquement lié à la taille mais à la vitesse des traitements. Nous y reviendrons au chapitre 2 lorsque nous parlerons des trois « V » : volume, vitesse et variété.

Figure 1.2 – Quelques ordres de grandeur d'espaces de stockage ainsi que la frontière approximative du Big Data.



Venons-en à quelques exemples et contre-exemples.

◆ *Quelques exemples qui ne relèvent pas du Big Data*

- ✓ Un volume de données que l'on peut traiter au moyen d'une fiche Excel.
- ✓ Des données que l'on peut héberger dans un seul nœud d'une base de données relationnelle.
- ✓ Les données qui sont « chères » à produire telles que celles qui sont collectées par sondage ou par recensement ou produite par un organisme tel que l'INSEE. L'idée ici est que les données qui relèvent du Big Data sont typiquement créées pour un coût quasi nul.
- ✓ Les données issues de capteurs physiques comme ceux de l'Internet des objets à venir.
- ✓ Les données publiques librement disponibles au téléchargement. Là encore on se place dans une perspective où ce qui relève du Big Data ne pourra être téléchargé au moyen d'une seule connexion Internet même à très haut débit.

◆ *Quelques exemples qui relèvent du Big Data*

- ✓ Les volumes de données qu'il n'est pas possible de stocker ou de traiter avec les technologies traditionnelles que les SGBDR ou pour lesquelles le coût d'un tel traitement serait prohibitif.
- ✓ Les données de logs transactionnelles d'un site web d'une grande enseigne de distribution. Nous examinerons en détail ce cas de figure au chapitre 10.
- ✓ Le trafic d'un gros site web.
- ✓ Les données de localisation GSM d'un opérateur téléphonique sur une journée.
- ✓ Les données boursières échangées quotidiennement sur une grande place financière.

1. 3 LA DONNÉE ET L'INFORMATION

1.3.1 La recherche pertinente

L'exemple le plus fréquent donné en illustration de l'usage des données est celui des moteurs de recherches qui constituent aujourd'hui un bon point de départ pour un parcours ciblé sur Internet.

Au milieu des années 1990, plus l'utilisateur fournissait de mots clés pour sa recherche et plus le nombre de réponses augmentait à l'écran. Dans la logique de l'utilisateur, cela était incohérent la plupart du temps, puisque l'ajout de mots clés devait permettre au contraire de retrouver l'article recherché avec plus de précision. Il manquait un élément essentiel à ces moteurs, c'était la capacité de saisir ce que l'utilisateur avait « en tête » lorsqu'il cherchait une information. On parle alors de recherche contextuelle.

1.3.2 Un avantage concurrentiel

La principale avancée de Google (et dans une moindre mesure de Yahoo!), qui explique en partie son succès, est d'avoir très tôt travaillé sur la compréhension de l'univers utilisateur, certes à des fins publicitaires mais également pour apporter de la performance dans la recherche et se distinguer de la concurrence.

Cette réussite brillante de Google a métamorphosé la relation entre l'utilisateur et les entreprises. Dans de nombreux domaines, une prise de conscience de l'énorme potentiel offert par le traitement de l'information est venue encourager le déploiement de nouvelles méthodes d'analyse comportementale et de processus qui étaient auparavant l'apanage quasi exclusif des grands acteurs du web.

1.3.3 Des clients plus exigeants

L'information pertinente est donc celle conditionnée (et attendue) par l'utilisateur dans de grands nombres de solutions. On dit souvent que 80 % de la réponse attendue est déjà présente dans le cerveau de l'utilisateur. Cela fait partie désormais de la culture de base de la relation client / systèmes d'information.

C'est d'autant plus frappant que le rapport avec le progiciel a grandement changé : la plupart des entreprises exigent maintenant des éditeurs de solutions logicielles que l'utilisateur final puisse disposer de plus en plus de liberté dans la constitution de ses tableaux de bords.

De manière désormais innée, les clients de solutions informatiques savent qu'un logiciel qui s'adapte à l'environnement de travail maximise les chances de fournir une information pertinente et à forte valeur. À l'inverse, si le contexte d'utilisation ne peut pas être modifié, c'est à la solution informatique d'être capable de comprendre plus finement ce qui est recherché.

Autre progrès visible dans l'analyse de la donnée, le « stocker tout » cohabite avec le « stocker mieux ». L'idée sous-jacente est de *vectoriser* la donnée, c'est-à-dire de

remplacer les données complexes par une description simplifiée et de se concentrer sur le sens. Cette rationalisation de l'information n'est cependant pas toujours aisée à mettre en œuvre.

Dans le cadre particulier de la connaissance du client, l'analyse comportementale et la recherche d'opportunités ou de nouveaux produits, nécessitent désormais la mise en relation de différents univers (potentiel financier du client, contexte économique, perception de l'image de l'entreprise, etc.). La complexité exposée plus tôt devient donc multidimensionnelle et à la limite de ce qu'un cerveau humain (ou même plusieurs) peut résoudre.

L'utilisation et l'analyse de la donnée sont donc à relier avec une finalité. Des résultats pertinents ayant du sens avec le métier sont conditionnés à la mise en place d'un ensemble d'informations prêt à être exploité de manière ciblée. Une information ciblée et satisfaisante devient enfin porteuse d'une valeur attendue. L'information extraite des données constitue un socle intéressant permettant des analyses ultérieures.

1.4 LA VALEUR

Parallèlement à la croissance régulière des données internes des entreprises, la montée en puissance des grands du web (Google, Amazon, LinkedIn, Twitter…) a révélé le potentiel immense des données externes à l'entreprise. Cette « masse » de données constitue la base des opportunités de demain. Ces acteurs réussissent à dégager des *business models* à forte rentabilité en nettoyant et en exploitant ces données pour se recentrer sur la pertinence de l'information isolée au service de leur stratégie.

Très progressivement, les entreprises et les éditeurs réalisent qu'il faut tirer profit de ce « trésor » et mettre en application les enseignements des récents progrès dans l'analyse des données. Dans cette perspective, les acteurs traditionnels de l'IT (IBM, Oracle, HP…) se sont lancés à leur tour bénéficiant au passage des travaux de R&D des grands du web dont la plupart des produits sont sous licences open source.

Le marché du Big Data adresse une vaste étendue de besoins :

- ✓ La mise en relation de données d'origine très diverses.
- ✓ La prise en charge de volumes jusqu'alors problématiques pour les systèmes d'information.
- ✓ La performance des traitements en termes de vitesse et de pertinence fonctionnelle.

Quant aux champs d'investigation possibles, les solutions sont capables de passer de la recherche sous différentes échelles : de l'information « macro », illustrant des tendances par exemple, à la recherche de l'information à l'échelle « micro », facteur déclenchant d'une décision très précise.

Les valeurs sont donc multiples : de la valeur stratégique à la valeur tactique et opérationnelle, en passant par les champs d'expérimentation pour les opportunités de demain encore non explorées.

1. 5 LES RESSOURCES NÉCESSAIRES

Dans l'histoire récente des systèmes d'information, les bases de données ont été créées pour rationaliser, structurer et catégoriser les données. Cet ensemble de solutions a connu un vif succès, et les années 1980 ont vu naître des bases de données de plus en plus complexes sans que l'on prenne le temps de simplifier leurs structures, avec des environnements de plus en plus gourmands. L'échelle de temps des utilisateurs ne permettait pas de se repencher sur ce qui avait été mis en place : les machines de la génération suivante répondaient seulement à chaque fois au supplément de ressources requises.

De nombreux constructeurs ont été ravis de vendre de la puissance de calcul et des infrastructures toujours plus puissantes. Toutefois la plupart des spécialistes sont restés prudents : certains d'entre eux ont rappelé tous les bienfaits de l'optimisation des traitements et de la rationalisation des systèmes d'information, conscients qu'une limite des ressources allait prochainement imposer de revoir les besoins et les solutions architecturales proposées.

L'apparition de phénomènes non linéaires dans les bases de données, pourtant optimisées, a constitué une première alerte. Des architectes ont constaté des phénomènes troublants, comme un doublement de volume de la base qui entraînait un temps de traitement multiplié par 5, 10 voire 100. Les *capacity planners* ont commencé à observer des courbes exponentielles entre le nombre de demandes et les temps de réponse. Les engorgements de systèmes complexes et inadaptés ont confirmé que, pour certains traitements, il fallait radicalement revoir la manière avec laquelle les données étaient traitées.

Ainsi les structures traditionnelles ont montré leurs limites face à cet engouement de « consommation de la donnée » et il a bien fallu reconnaître que la progression de la performance passait inévitablement par la simplification et la délégation des processus des traitements à des machines optimisées.

Ces évolutions ont été là aussi très diverses :

- ✓ De nombreux efforts ont été apportés aux structures matérielles en permettant des exécutions encore plus rapides, avec un recours massif à la mémoire, plutôt qu'à l'accès direct au stockage.
- ✓ La parallélisation est venue donner une réponse à ce besoin en s'inscrivant dans la simplification des traitements imbriqués et la délégation à des nœuds de traitements isolables[1].
- ✓ Les nouveaux moteurs de requête (*Not only SQL* ou NoSQL) se sont affranchis de certaines imperfections du SQL et ouvrant la porte à d'autres modes d'accès[2].

1. Un cas particulier important de mécanisme de parallélisation, l'algorithme MapReduce sera décrit en détail au chapitre 4.

2. Le sujet sera abordé en détail au chapitre 3.

- ✓ Le recours à des algorithmes d'apprentissage statistiques a relancé l'exploration de données à isopérimètre, comme cela est souvent le cas dans le domaine du décisionnel[1].

1.6 DE GRANDES OPPORTUNITÉS

Les possibilités offertes par le Big Data sont immenses :

- ✓ Analyser une grande variété de données, assembler des volumes extrêmes.
- ✓ Accéder à un éventail de données jamais atteint par le passé.
- ✓ Capter la donnée en mouvement (c'est-à-dire très changeante), même sous forme de flux continu.
- ✓ Compiler de grands ensembles et les mettre en correspondance en les recoupant.
- ✓ Découvrir, expérimenter et analyser des données cohérentes, tenter des rapprochements.
- ✓ Renforcer des associations entre données, intégrer et construire des ensembles performants et industriels.

En résumé

Le Big Data correspond bien à une réalité de l'usage de la donnée et de ce qui est désormais attendu d'un système d'information.
Le Big Data, par son changement d'échelle et les réponses architecturales comme fonctionnelles qui y sont associés, est effectivement une rupture d'approche dans l'analyse de l'information.
La valeur est devenue l'élément central de l'approche Big Data dans des dimensions qui n'ont pas plus rien à voir avec les « anciennes » extractions de l'informatique décisionnelle.

1. C'est le sujet des quatre chapitres de la partie 2.

Le Big Data dans les organisations

Objectif

L'ambition clairement formulée du Big Data a rendu les entreprises impatientes d'atteindre le Graal de la valeur tant vanté par les acteurs du web. Ce chapitre décrit les principaux changements inhérents à la révolution numérique dans les entreprises et comment elles doivent se réorganiser aussi bien en interne qu'en externe pour optimiser l'accès à l'information.

2. 1 LA RECHERCHE DE L'ELDORADO

2.1.1 L'entreprise au cœur d'un écosystème

Le paysage dans lequel gravite l'entreprise a radicalement changé. Tant dans les informations qu'elle requiert que les informations qu'elle doit fournir pour exister dans cet écosystème économique et social où elle se doit d'être présente et réactive pour se développer. La notion de proximité numérique avec les clients est inscrite dans de nombreux plans stratégiques. Cette véritable rupture culturelle a bousculé les repères de l'entreprise tout comme ceux de ses clients.

À l'ère du numérique, l'entreprise devient de plus en plus ouverte aux flux d'informations, de façon directe ou indirecte (ne serait-ce que par la contribution de ses collaborateurs). L'entreprise doit de plus en plus contribuer numériquement au devenir citoyen, de la même manière que, au début des années 1990, ses actions pouvaient contribuer au développement de l'économie nationale. Progressivement, le réel enjeu des données est compris non seulement comme un moyen de simplifier le contact avec ses partenaires, ses prospects ou ses clients mais également comme le moyen d'acquérir des données nécessaires à la compréhension des grandes tendances sociétales.

2.1.2 Une volonté de maîtrise

De plus en plus d'entreprises cherchent à exploiter ces données dans de multiples domaines : personnalisation de la relation client, marketing ciblé, traçabilité des parcours et retours clients, pilotage et valorisation de l'image sur les réseaux sociaux, optimisation des processus logistiques… ces données constituent un gisement d'informations au potentiel quasi illimité, à condition de pouvoir les maîtriser. Or les outils traditionnels ne savent pas répondre à ce besoin.

Formatées sur le traitement de données dites structurées, les applications, qui jusqu'alors donnaient satisfaction, sont devenues impuissantes devant la diversité des données à traiter : origines diverses sous forme de texte, mais aussi de sons, d'images, de vidéos, de sites web et de blogs aux formats très variés.

2.1.3 Des besoins forts

Les éditeurs SI, qui ont compris ce bouleversement et ces nouveaux besoins, ont commencé à s'intéresser au calcul scientifique pour traiter toutes ces données avec le maximum d'efficacité. Côté marketing, de nouvelles méthodes ont permis de stocker indifféremment des formats multiformes, presque de manière brute, principalement pour offrir des possibilités de recouvrements avec des données issues de processus complètement différents (réseau des boutiques, parcours sur le web, fréquentation des sites, échanges téléphoniques, dossiers clients, etc.). Cela a permis d'établir des regroupements d'informations totalement nouveaux, dans des domaines fonctionnels anciennement éloignés (canaux d'informations multiples, expérience UI web, proximité physique…) et de constituer des profils opportunistes sur les clients, sur leur potentiel, leurs projets et leurs envies.

La presse décrit fréquemment des entreprises qui ont assemblé des données de plusieurs sources et qui peuvent désormais proposer à leurs clients un ciblage des offres (réduction, abonnement) en fonction de leur localisation, par exemple, à proximité d'une galerie commerciale et cela, grâce au Big Data.

2. 2 L'AVANCÉE PAR LE CLOUD

Un des facteurs importants dans l'émergence du Big Data est le *cloud computing*, qui a grandement simplifié l'accès à des infrastructures performantes. Basé sur des ressources ajustables, disponibles dans des délais très courts, par durée identifiée et à un coût plus adapté, le cloud computing a ouvert de nombreuses portes aux projets innovants en abaissant considérablement le coût du ticket d'entrée sur ces solutions novatrices. L'accès à la performance, possiblement temporaire, est devenu plus aisé et souvent avec des latences matérielles et contractuelles moindres.

Place est faite aux expérimentations et aux POC (*Proof of Concept*) sur plateformes performantes qui permettent d'essayer certains traitements ou certains recoupements et transformations, sans (trop) ponctionner le budget de la DSI.

Lieu de convergence des données d'origines très diverses (météo, localisation, économie, tendances multiples), le cloud a fait entrer l'entreprise dans un univers de croisements et d'études de données rarement tentés par le passé, en combinant l'analyse de la mesure, plus ou moins poussée, pour ensuite aborder celui fascinant de la prédiction.

2. 3 LA CRÉATION DE LA VALEUR

Deux manières de créer de la valeur grâce au Big Data :

- ✓ **La conception de nouveaux produits**. Aujourd'hui, les grands acteurs du web sont des exemples suivis par des centaines de start-up en Europe et aux États-Unis. Les conditions de déploiement de ces nouveaux produits dans les entreprises que nous connaissons autour de nous ne sont pas toujours équivalentes aux start-up américaines, mais leurs success-stories servent d'exemples et constituent des leviers déclencheurs à la transformation numérique. Les domaines impactés sont immenses et, même dans des filières de métiers fortement informatisés, ces produits peuvent apporter une étendue de vision et de justesse très appréciée, comme dans le domaine bancaire ou celui de l'assurance en fournissant le produit ou le service réellement attendu par le public.
- ✓ **L'angle analytique en élargissant le champ d'investigation à des données qui n'étaient pas accessibles sur d'autres outils**. Par exemple, la mise en relation de données disponibles sur des réseaux plus personnels (réseaux sociaux) pour mieux analyser l'écosystème autour du client, son réseau, ses influences et celui des ventes potentielles en le reliant avec le contexte géographique et la proximité du réseau des magasins. En résumé, il s'agit de construire les conditions favorables pour assurer le meilleur accueil du produit auprès du client.

Le déploiement de tels outils ne peut se faire sans effort. L'émergence du Big Data en entreprise doit impérativement bénéficier de changements culturels profonds afin qu'elle soit un succès (la section 2.7 reviendra sur le sujet).

La promotion des valeurs telles que celles de l'agilité et la flexibilité est indispensable, alors que souvent elles sont opposées aux modes d'organisations pyramidales actuels. L'innovation portée par ces approches Big Data doit reposer sur l'ambition et la confiance des structures managériales.

Les cycles courts de test doivent stimuler la créativité dans les équipes et de nombreux efforts doivent être maintenus pour sortir de la culture de l'échec trop souvent présente dans les entreprises françaises. C'est à ce prix que de nouvelles opportunités s'offriront aux équipes s'attelant aux défis du Big Data.

2. 4 LES « 3V » DU BIG DATA

Afin de mieux cerner les caractéristiques du Big Data, des spécialistes d'IBM ont proposé trois propriétés caractéristiques à des degrés divers : il s'agit du *volume*, de la *vélocité* et de la *variété*. On les appelle communément les **3V**. D'autres dimensions sont fréquemment ajoutées, mais cette présentation ciblera ces trois propriétés.

2.4.1 Le volume

La question du volume a déjà été évoquée dans la section 1.2.3, nous n'y reviendrons donc pas ici.

L'une des technologies aujourd'hui couramment utilisées pour faire face aux besoins inhérents au volume est le framework Hadoop. Le sujet sera abordé en détail aux chapitres 4 et 9. Disons simplement que le stockage est assuré par un système de fichiers distribué appelé HDFS et que la parallélisation des traitements exploite un pattern appelé MapReduce. Tous les problèmes de parallélisation ne peuvent exploiter l'algorithme MapReduce mais, lorsque c'est le cas, celui-ci offre une scalabilité quasi infinie. Hadoop est aujourd'hui un programme de la fondation Apache. La version actuelle de Hadoop réutilise la gestion distribuée des ressources développées pour exécuter MapReduce afin de permettre l'exécution d'autres schémas de calcul, plus généraux.

Hadoop est aujourd'hui disponible dans le cloud ; la solution Elastic MapReduce d'Amazon est un exemple.

2.4.2 La vélocité

◆ *Au cœur du Time To Market*

La vélocité ou vitesse de circulation des données a connu une évolution similaire à celle du volume au sein des entreprises. Contexte économique et évolution de la technologie, les entreprises se trouvent de plus en plus cernées par un flux continuel de données, qu'il soit interne ou externe.

Rien de bien nouveau pour certains métiers comme celui de la finance, où le *trading*, par exemple, tire de la vélocité un avantage concurrentiel fondamental qui permet chaque jour de prendre une longueur d'avance dans la décision, source de gains financiers. Le temps réel est déjà là, bien installé, depuis plusieurs années dans ces familles de métiers. Le sujet sera abordé en détail aux chapitres 11 et 12.

◆ *Au service des clients*

Le paramètre vélocité est déterminant dans les situations où il s'agit de tenir compte en temps réel des souhaits exprimés par un client et de l'état de disponibilité des stocks pour fournir le meilleur service possible. Dans d'autres situations, on pourra utiliser des données de géolocalisation pour guider un client vers un magasin moins exposé aux problèmes de circulation ou pour lui offrir des services de proximité.

2.4.3 La variété

L'une des grandes ambitions du Big Data est de proposer le recoupement d'informations ou la mise en correspondance de données d'origines très diverses. Ce serait même l'une des principales sources de valeur. Toutefois ces données sont souvent de qualité très variable d'une origine à l'autre mais, une fois rapprochées, elles viennent constituer une vision plus ouverte et plus large que celles des référentiels. Pourtant, il ne faut pas se leurrer, parmi les 3V c'est le seul pour lequel il n'existe

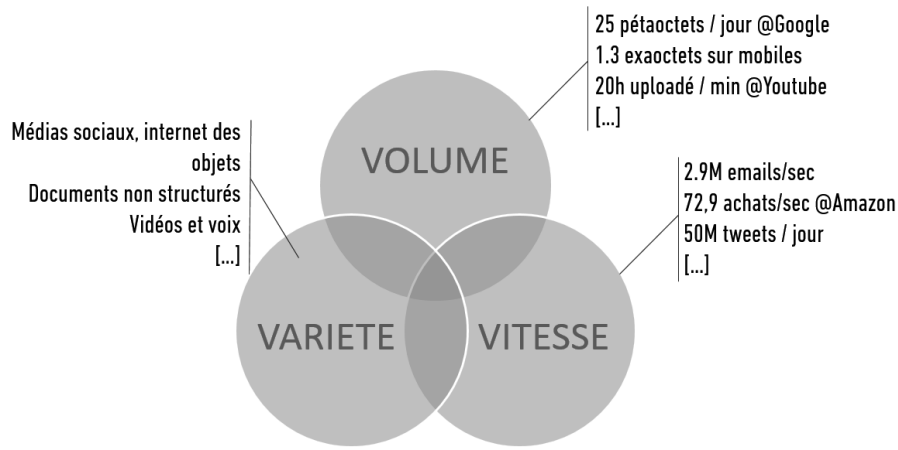
à l'heure actuelle aucune méthode universelle. C'est sans nul doute le problème le plus épineux car le traitement et le recoupement de données de sources et de formats variés demandent une étude adaptée à chaque situation.

Un exemple typique d'utilisation de données hétéroclites est celui d'un croisement entre des données contenues dans un CRM (gestionnaire de la relation client), des données de géolocalisation et des données extraites d'un réseau social qui, collectivement, permettront d'enrichir un profil utilisateur avec des informations à caractère affectif très souvent corrélées au déclenchement d'un acte d'achat.

Pour faire face à cette diversité de structures, certains systèmes NoSQL (voir le chapitre 3) utilisent des schémas de données qui s'écartent du modèle relationnel classique. Les bases orientées graphes ou les bases orientées colonnes en sont des exemples.

La prise en compte de la diversité des formats de données fera typiquement partie de la phase de préparation des données, le sujet du chapitre 6.

Figure 2.1 – Le Big Data à l'intersection des trois « V ».



2.5 UN CHAMP IMMENSE D'APPLICATIONS

Voici quelques exemples d'applications du Big Data en entreprise. La liste est loin d'être exhaustive :

✓ **Analyser les données en mouvement**

- La surveillance d'un grand nombre de nœuds de données (cloud, hébergement, traitements répartis).
- La sécurité électronique et la détection de fraude dans un contexte de banque ou d'assurance.
- Le suivi en temps réel et à forte réactivité de clients (commerce de détail, téléassistance).
- Le pilotage avancé de systèmes complexes (recherche en astronomie et en physique des particules, domaine spatial, forces armées).
- L'analyse des logs (voir chapitre 10) et surveillance des processus.

- Le tracking des identifiants, connexions et pistage des transports (avec le RFID par exemple).
- La logistique avancée avec ou sans contact physique.

✓ **Analyser une grande variété de données**

- Le décodage et l'analyse des humeurs sur les réseaux sociaux ou dans la blogosphère.
- L'alignement avec la stratégie d'entreprise.
- Le suivi de phénomènes propagés (santé : épidémies, intrusions, ingénierie sociale, terrorisme).
- L'analyse multimédia à partir de formats vidéo, audio, texte, photos (sécurité, traductions, médiamétrie).

✓ **Traiter un volume conséquent de données**

- Rapprochement d'objets métiers : produits, clients, déploiement, stocks, ventes, fournisseurs, lieux promotionnels.
- Détection de fraudes, repérage de clients indélicats ou manipulateurs.
- Toxicité, défense des intérêts d'un groupe, *feedback* de crise.
- Management du risque, prise de décision sensible appuyée de modèles prévisionnels.
- Analyse de contexte, d'environnement.

✓ **Découvrir et expérimenter**

- Approcher la notion de désir ou de sentiment déclencheur d'action.
- Cerner l'entourage social d'un client, les conditions favorables.
- Expérimenter l'impact d'un nouveau produit, son ressenti.
- Mesurer l'efficacité d'une stratégie de conquête, mesurer des écarts ou des erreurs de positionnement d'image.
- Profilage de nouveaux comportements.
- Identification de nouveaux indicateurs d'influence.

2. 6 EXEMPLES DE COMPÉTENCES À ACQUÉRIR

Les paragraphes qui suivent illustrent quelques exemples de compétences qu'une société pourra être amenée à acquérir pour faire du Big Data. Il serait vain de vouloir en donner une liste exhaustive mais les exemples qui suivent aideront à s'en faire une idée.

2.6.1 Appréhender de nouveaux modèles de traitement des données

Nous avons déjà évoqué le pattern MapReduce, un des algorithmes historiques du Big Data puisqu'il est au cœur des premières versions de Hadoop. Même si comme nous le verrons au chapitre 9 ce modèle est supplanté par d'autres, plus flexibles, il continue à jouer un rôle structurant dans de nombreux traitements asynchrones.

Si son principe est simple (comme nous le verrons au chapitre 4), les détails en revanche sont plus ardus et exigeront, pour être pleinement maîtrisés, un effort

d'apprentissage significatif de la part des développeurs. L'expérience des pionniers comme Facebook et Yahoo! a amplement démontré cette difficulté. Il est vraisemblable cependant qu'une majorité des développeurs métiers pourra se dispenser d'une connaissance approfondie de ces aspects algorithmiques. En effet, pour bénéficier du parallélisme massif de MapReduce, ils n'auront que rarement à implémenter explicitement l'algorithme MapReduce[1] mais utiliseront plutôt des abstractions de haut niveau, essentiellement des langages de scripts (comme Pig) ou des langages pseudo SQL (comme Hive QL) plus familiers.

Dès lors la question se pose de savoir s'il est possible de faire l'impasse complète sur les concepts de MapReduce. Hélas la réponse est non, et c'est là que résident la subtilité et les risques. Comme souvent dans l'IT, on ne peut pas oublier complètement une complexité algorithmique qui a été masquée. Une connaissance élémentaire des mécanismes sous-jacents reste nécessaire pour faire face aux situations où les automatismes dysfonctionnent ou dès lors que les performances exigent des optimisations.

Dans le même ordre d'idées, il faudra veiller à ce que les plans d'exécution MapReduce générés par Hive ne soient pas trop complexes pour être exécutables et testables en un temps raisonnable. Liées aux ordres de grandeur inhabituels des quantités de données traitées, c'est tout un jeu d'intuitions nouvelles à acquérir par les concepteurs.

Quel est alors l'effort à consentir pour maîtriser la programmation MapReduce explicite (sans scripts) sous Hadoop ? Une durée comprise entre six mois à un an ne semble pas surestimée s'il s'agit d'acquérir une expérience significative. Ces compétences sont aujourd'hui détenues par quelques petites sociétés spécialisées et sont rarement développées en interne bien qu'il existe des cursus de formation dédiés.

Pour ce qui concerne les langages de plus haut niveau comme Pig et Hive QL, on peut estimer à deux ou trois semaines le temps de parvenir à un niveau de compétences suffisant, vu la proximité avec les langages existants comme SQL. En donnant ces estimations, nous présupposons qu'un coach soit à disposition pour assurer la formation initiale des développeurs ou alors qu'une cellule spécialisée et dédiée, fonctionnant en mode R&D, soit mise en place localement dont la mission sera de monter en compétence sur ces sujets et, à terme, de former le reste des troupes.

2.6.2 Maîtriser le déploiement de Hadoop ou utiliser une solution cloud

À l'appréhension des concepts et des langages de scripts spécifiques à MapReduce, il convient d'ajouter la maîtrise du déploiement d'une plateforme qui implémente ce modèle. Le rôle d'un tel framework consiste à masquer l'énorme complexité technique qu'implique un traitement massivement parallèle par un cluster de machines non fiables : planification des tâches en fonction des ressources du cluster, réplication

1. L'implémentation des fonctions Map et Reduce comme on le verra au chapitre 4.

des données, réexécution des tâches en cas d'échec, garantie de très haute disponibilité, etc. Un outil prédomine actuellement, il s'agit de Hadoop, un framework libre développé ces dernières années sous l'égide de la fondation Apache.

Bien qu'il existe d'excellents ouvrages sur le sujet[1], le déploiement « *on premise* » d'un cluster Hadoop et son optimisation restent un travail d'experts hautement qualifiés. Dès lors, l'alternative pour une DSI qui souhaiterait mettre en œuvre Hadoop, consistera à choisir de faire appel à un prestataire spécialisé ou alors à utiliser un déploiement de Hadoop dans le Cloud chez un fournisseur PaaS. La solution Elastic MapReduce d'Amazon Web Services constitue ici une référence. Dans ce dernier cas, pour les très gros volumes de données, interviendront des considérations de coûts de transferts qu'il ne faudra pas prendre à la légère. La maturité des plateformes PaaS et la disponibilité de tutoriels pédagogiques devraient rendre accessibles le paramétrage et le déploiement d'applications à la plupart des DSI. Quelques semaines d'expérimentation devraient suffire pour permettre à une équipe IT chevronnée de maîtriser l'exploitation de Hadoop en mode PaaS.

2.6.3 Se familiariser avec de nouvelles méthodes de modélisation

Traiter d'énormes quantités de données non structurées exige non seulement de nouveaux algorithmes et de nouvelles infrastructures de calcul, mais impose également de repenser l'organisation des données elles-mêmes. Dans ce nouveau contexte, les enjeux de performance priment le plus souvent sur ceux d'une stricte intégrité référentielle, contrairement aux principes qui s'appliquent aux bases relationnelles. Par ailleurs, le caractère non structuré des données rend caduque la notion de schéma prédéfini. De nouvelles catégories de base de données répondent à ces exigences, comme les bases NoSQL que nous décrirons au chapitre 3. Ce type de bases pourra jouer aussi bien le rôle de source que de destination pour les jobs MapReduce. L'expérience montre qu'elles s'avèrent particulièrement bien adaptées aux données hiérarchiques ou structurées en graphe qui sont monnaie courante dans un contexte Big Data. L'archétype ici est celui d'une *webtable* qui répertorie les attributs d'une collection de milliards de pages web indexées par leur URL. À la panoplie des savoir-faire déjà évoqués, il convient donc d'ajouter de nouvelles techniques de modélisation. À cette occasion, certains réflexes anciens devront être désappris, les processus de dé-normalisation et de duplications, autrefois prohibés, étant désormais à l'origine même des gains en performance et des capacités à monter en charge linéairement.

Une compréhension fine des algorithmes ou des applications qui exploitent les données sera nécessaire pour pouvoir les modéliser. Contrairement aux systèmes relationnels la structure optimisée de ces nouveaux modèles dépend fortement des applications qui les utilisent (voir chapitre 3).

---

1. Tom White, *Hadoop: The Definitive Guide*, 3rd Edition. O'Reilly, 2012.