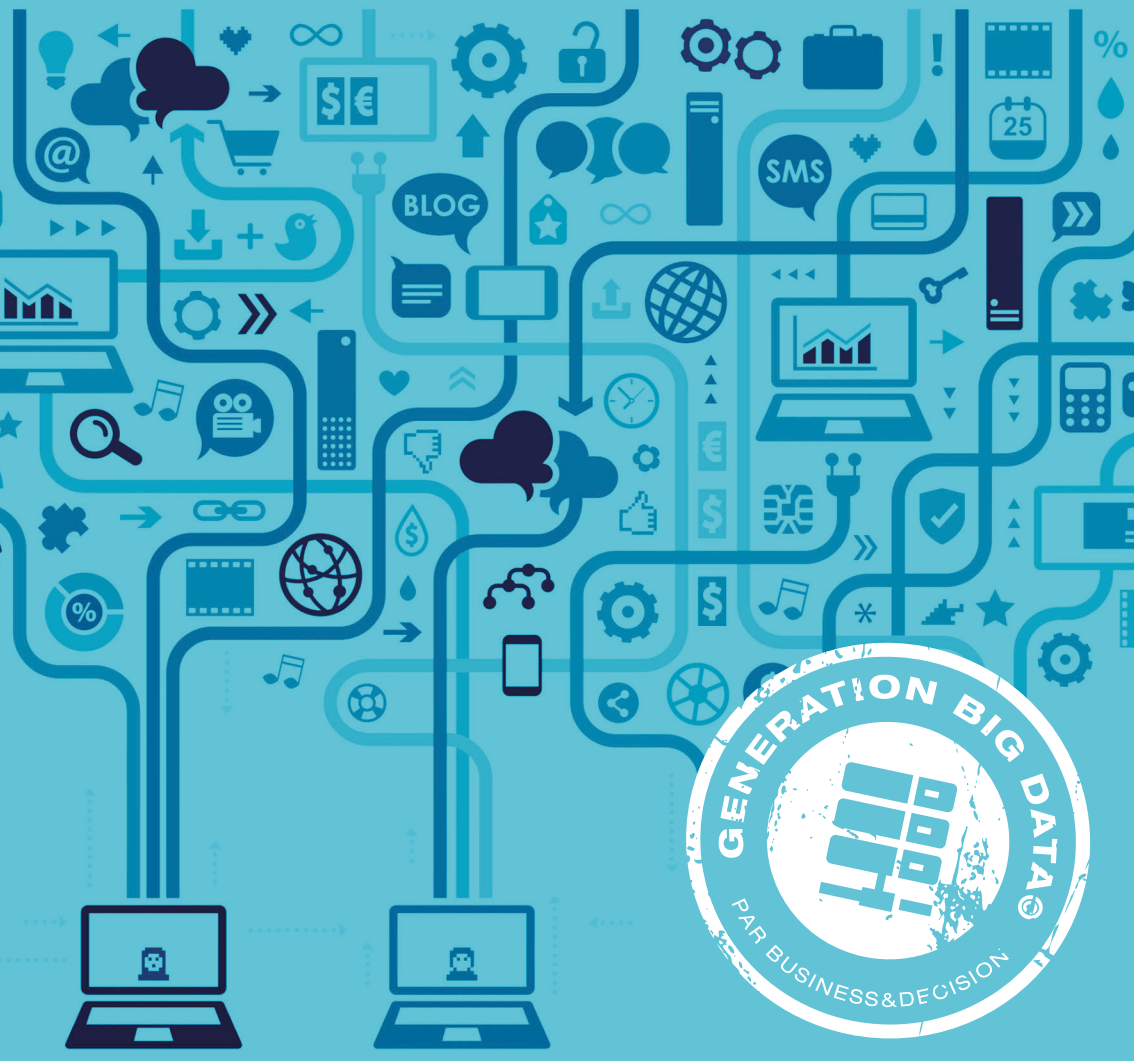
Business & Decision

LIVRE BLANC

DU BIG DATA AU BIG BUSINE$$

LIVRE 1 : Phénomène de mode ou facteur de performance ?



CHAPITRE 01

CHAPITRE 02

CHAPITRE 03

CHAPITRE 04

CHAPITRE 05

CHAPITRE 06

CHAPITRE 07

CHAPITRE 01
Big Data, contexte et fondements

Pourquoi ce livre ?

La littérature sur le Big Data est abondante. Cette abondance est symptomatique d'un élan dont l'importance est perçue fortement par l'ensemble du marché, non seulement en France, mais dans le monde. Toutefois, même quand elle est de qualité, cette littérature reste assez descriptive et focalisée sur la dramatisation d'enjeux quasi apocalyptiques, reliés à la profusion exponentielle des volumes de données et de leurs sources. Cette approche ne permet pas de comprendre les véritables enjeux des Big Data ni comment les entreprises peuvent en tirer parti.

Même si les prévisions sont délicates, nous avons la conviction que l'impact des Big Data sur l'avenir des entreprises et de la Société civile sera fort, polymorphe et en constante reconfiguration. C'est donc en faisant rapidement leurs premières armes sur les Big Data, qui en sont encore à leurs prémices, que les entreprises pourront s'approprier le phénomène et apprendre, jour après jour, à en tirer parti.

L'objectif de ce Livre Blanc est de donner aux entreprises, les premières clés de lecture qui permettront aux lecteurs de sortir de la mythologie associée aux Big Data pour les replacer dans leur contexte propre et les aborder comme un outil puissant de développement de la performance.

Nous espérons ainsi permettre au lecteur de poser, voire de valider, les premières orientations d'une intégration sereine et maitrisée du Big Data à l'écosystème de son Entreprise.

[1] *Ceci d'autant plus que les Big data impliquent de nouvelles formes de raisonnements, qui embrassent notamment les formes de raisonnements inductifs (cf. page 8). On peut sans grand risque parler des Big data comme d'une nouvelle philosophie et une nouvelle façon de penser le marketing.*

[2] *http://trends.levif.be/economie/actualite/entreprises/les-big-data-posent-probleme-aux-marketers/article-4000606787740.htm*

Évolution de la pensée managériale

La vogue du Big Data représente beaucoup pour le monde de l'entreprise. Au-delà d'une simple mode, il s'agit d'une véritable révolution du mode de pensée, un apport crucial dans la panoplie managériale qui va profondément et réellement changer la face du monde du business. Le marketing est le principal impacté, mais il ne faut pas sous-estimer les conséquences de la maîtrise de ce sujet sur le monde des études, de la gestion, en passant par celui de la production, du Supply Chain Management et de la R&D… (la liste serait trop longue).

Cette révolution profonde des modes de pensée est cependant mal servie par cette abondante littérature qui a tendance soit à expurger le vocabulaire trop technique de cette nouvelle discipline afin d'en masquer la complexité[1], au point de la rendre incompréhensible, soit au contraire à rentrer trop profondément dans cette complexité et son vocabulaire quadruplement technique (métier, technologie, bases de données et statistiques) et de perdre le lecteur. Le résultat, c'est que le lecteur pourrait penser, à tort, que le sujet des Big Data est soit trop générique et donc peu innovant, soit trop innovant pour que l'entreprise de tous les jours puisse en profiter.

En conséquence, les Big Data, pour paraphraser le journal de tendances trends.be[2], « posent problème à l'entreprise » alors qu'elles devraient au contraire être perçues comme une solution. Le but de ce livre, le premier d'une série de 6 est au contraire de décrire simplement et clairement les impacts et les usages des Big Data sans appauvrir le discours, mais aussi en explicitant son jargon afin de rendre accessible à tous les bénéfices de ces nouveaux outils.

Ce premier livre est dédié au phénomène général des Big Data et sera ensuite détaillé dans chacune de ses composantes. Chacun des chapitres de ce livre fera par la suite l'objet d'un nouveau livre, selon le plan suivant :

- Le livre 2 sur les données, carburant essentiel des Big Data ;
- Le livre 3 sur les usages des Big Data ;
- Le livre 4 sur les architectures & les algorithmes qui sous-tendent le Big Data ;
- Le livre 5 sur les déclinaisons en métiers des Big Data ;
- Le livre 6 sur la confidentialité des données, la protection des utilisateurs et l'éthique.

Ainsi, nous proposons à la fois une vue globale (dans ce livre) et détaillée (dans les autres livres) des Big Data, afin de les rendre accessibles à tous les professionnels qui veulent mettre à profit cette nouvelle approche et ses outils pour leur business.

Définition des Big Data[3]

Le terme de Big Data (parfois appelées « données massives » en français, mais nous éviterons d'utiliser cette traduction peu réussie) désigne une nouvelle discipline qui se situe au croisement de plusieurs domaines : statistiques, technologie, base de données et métiers (marketing, finance, RH, etc.).

Cette nouvelle discipline a été rendue possible grâce à une puissance technologique qui a rendu possible des choses qui jusque là n'étaient que théoriques. Ces choses dont on parle ici, sont principalement liées à deux enjeux : le volume des données et leur complexité.

---

[3] *Le, la ou les Big Data ? Big Data est un nom anglais (littéralement « grosses données » et ne nécessite pas d'être mis au féminin ni au masculin. Le mot « data » étant le pluriel latin de Datum, nous avons décidé de garder ce nom au pluriel dans ce document.*

Ainsi, le Big Data a pour objectif d'exploiter des volumes de données qui sont en croissance exponentielle et qui deviennent difficiles à travailler avec des outils classiques de gestion de base de données ou de gestion de l'information[4]. Elle a aussi pour objectif de traiter rapidement des données complexes.

Si l'on se réfère aux travaux de *the 451 group* et *Gartner*, on trouve la formulation suivante ; Les big Data visent à tirer un avantage concurrentiel des méthodes de collecte, d'analyse et d'exploitation des données qu'on ne pouvait utiliser jusqu'à présent du fait des contraintes économiques, fonctionnelles et techniques liées aux volumétries, à la vitesse de traitement et à la variété des données à considérer.

Rupture ou évolution ?

Les Big Data sont parfois présentées comme un phénomène en rupture complète avec ce qui ce qui a pu se faire jusqu'à aujourd'hui en terme d'aide à la décision, ou au contraire, comme une simple évolution des organisations et des systèmes décisionnels. La question est plus importante qu'il n'y paraît, et ne se réduit pas à un simple problème sémantique. Cette importance, pour l'entreprise se traduit par le fait qu'en fonction de la réponse, les scénarios mis en place seront probablement très différents.

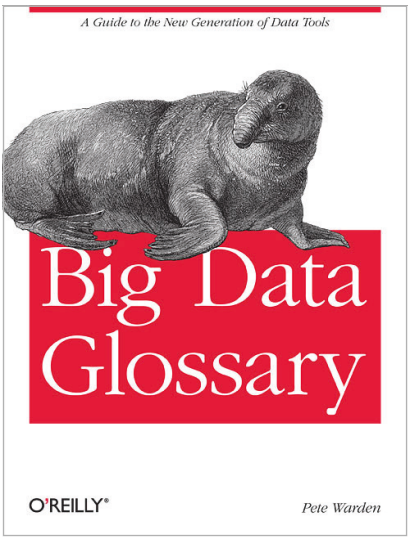
La genèse des Big Data

La genèse des Big Data est en soi porteuse de sens et explique la spécificité de ce domaine ; si le terme de Big Data a été utilisé pour la première fois par le cabinet d'analyse *Gartner* en 2008, on peut cependant faire remonter la genèse des Big Data à beaucoup plus loin. Dans un sens, celles-ci naissent avec l'essor de l'informatique, et comme toutes les innovations, il a fallu un certain temps pour que le concept se généralise et se raffine au fil du temps.

Gil Press fait même remonter les origines de cette nouvelle discipline à une date plus que lointaine[5] (1944 !). Mais sans aller jusque-là, et même si la paternité de l'invention du terme « Big Data[6] » fait l'objet de débats assez techniques, on peut facilement remonter au début de 2001, selon le cabinet d'analyse *Gartner*, pour trouver les premiers écrits sur les fameux 3V (Volume, Vélocité et Variété), prédisant l'explosion de la donnée et la naissance d'une nouvelle forme de traitement de celle-ci.

Il faut préciser également que les Big Data sont aussi l'aboutissement de la démarche de Data Mining, en vogue dans les années 1995-2000, elle même issue de deux écoles (ou tendances) assez anciennes que sont la statistique d'un côté et l'intelligence artificielle d'un autre.

La fonction crée l'organe



La lecture de l'excellent glossaire des Big Data de O'Reilly pourrait-elle mettre tout le monde d'accord en faisant remonter leur genèse non aux analystes qui décrivent le phénomène, mais aux

[4] *Voir également cette définition complète dans l'encyclopédie ouverte Wikipedia dont nous nous sommes inspirés : http://fr.wikipedia.org/wiki/Big_data*

[5] *http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data*

e-commerçants et autres acteurs du Web collaboratif qui ont contribué à sa création.

En quelque sorte, comme ce fut le cas avec le fameux Web 2.0 de O'Reilly en 2004, le terme descriptif de ce phénomène est arrivé après le phénomène lui-même, comme très souvent dans la sphère digitale.

Le cloud computing par exemple, est né de la surcapacité en termes d'hébergement des grands sites Web (Amazon, Google, e-Bay, Microsoft,...) et des architectures originales qu'ils avaient mises en place afin de faire face aux afflux de leurs visiteurs et des besoins en « élasticité » de leurs hébergements ; de la même manière, les Big Data sont nées des surplus de données induits par les usages des internautes sur des sites comme Amazon, Yahoo! et bien plus tard, Google – toujours les mêmes protagonistes – qui ont permis des usages marketing originaux, qui n'étaient pas prévus au départ[7].

Un mécanisme vertueux s'est mis en place, qui a permis à des entreprises puissantes et riches d'investir dans des technologies pour répondre dans un laps de temps très court aux besoins de leurs utilisateurs. La bulle internet a rendu possible une expérimentation en temps réel avec des ressources illimitées... le rêve de tout chercheur en quelque sorte !

Ainsi la logique est-elle bousculée, la donnée vient avant l'usage, l'entreprise avant la recherche, la fonction, en quelque sorte, crée l'organe ... et les usages induits par ces avancées technologiques sont nombreux, comme nous le décrivons dans notre chapitre dédié aux usages (cf. Chapitre 3 : page 5)

Avant de penser en termes de rupture ou de continuité, présentons les nouveautés qu'apporte ou qu'induit le phénomène Big Data en les classant dans 4 thèmes : les données, les usages, les méthodes de travail et les outils.

Le traitement des signaux faibles (P. Cahen)

Si les Big Data permettent de tirer partie des signaux faibles, il ne faudrait pas en tirer la conclusion que la détection des signaux faibles est un sujet nouveau. Voyons avec Philippe Cahen, auteur du livre « le marketing de l'incertain » comment les signaux faibles ont un impact sur le marketing, et pourquoi ils sont si importants.

Qu'est-ce au juste qu'un signal faible ? Philippe Cahen nous livre sa définition dans son ouvrage « tout savoir sur... le marketing de l'incertain » :

« Un signal faible est une information paradoxale de réflexion [...]. Un signal faible n'est pas un petit fait porteur d'avenir. Ce serait trop simple en effet, voire naïf, d'imaginer que l'on trouve tout cru des informations sur l'avenir. C'est la réflexion qu'il provoque qui est porteuse d'avenir. Large part est laissée à l'intuition pour déceler puis interpréter les signaux faibles ».

« Le futur est un saut dans l'inconnu, un inconnu allant du sympathique rassurant à l'intolérable que l'on voudrait fuir. Il est en effet particulièrement rare de vivre ce qui a été prévu. L'inverse est plutôt la règle ».

Les Big Data sont un moyen d'alimenter la réflexion, et l'action, autour de ces signaux faibles, en partant d'hypothèses que l'on peut vérifier en faisant des croisements entre données et comportements.

Source : « Le Marketing de l'incertain » par Philippe Cahen, éditions Kawa, 2012

[6] *Voir cet article sur le blog du New York Times : http://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story*

[7] *Cf. l'article de Lise Gasnier sur Solucom Insight http://www.solucominsight.fr/2013/08/auw-origines-du-big-data*

Le Projet de Big Data

Un projet du Big Data est un ensemble constitué de 4 composantes.

Il est articulé autour des technologies du Big Data, matériels et logiciels, bien sûr, mais induit également une approche méthodologique particulière, que nous évoquerons brièvement dans ce document et plus en détail dans les prochains ouvrages.

La troisième composante est juridique, car il convient de maîtriser le cadre juridique particulier lié aux données qu'on manipule et aux types d'utilisation souhaités.

Enfin, il est une composante sociale importante qu'il convient de prendre en compte dans un projet Big Data. Quelle est la capacité de nos sociétés et de chaque groupe de personne ou individu à accepter la circulation et l'utilisation de ses données personnelles. Si l'on ne veut pas exposer son projet et plus généralement ce domaine d'application, il appartiendra aux entreprises de s'autoréguler et aux législateurs de s'adapter à ces nouveaux contextes et ces nouvelles possibilités qu'offrent les technologies.

CHAPITRE 02
Les données

Comme le nom l'indique, les données sont bel et bien le fondement et la matière brute du phénomène Big Data ; il est donc naturel que nous commencions par là.

Aux données structurées, gérées dans les applications informatiques traditionnelles (ERP, CRM, SCM, etc.) sont venues se greffer en quelques années de nombreuses autres données, souvent nommées « données non structurées » ou « semi-structurées »[8] :

- les messages électroniques (e-mails et de plus en plus les messages instantanés), la numérisation de tous les documents contractuels, les saisies et les traces déposées sur les sites Internet, les conversations avec les centres d'appels ;

- viennent ensuite les données associées à la mobilité : identifiants (numéro d'identification IMEI, carte SIM, UID…), les historiques de navigation, les positions géolocalisées et même les préférences d'utilisateurs ;

- ensuite, en augmentation constante et exponentielle, les données générées par les objets connectés : machines, voitures et compteurs « intelligents », Set top box (passerelles Internet des opérateurs, Box des câblo-opérateurs etc.), capteurs, données issues de la domotique, des systèmes biométriques personnels ;

- enfin et surtout, les données créées et échangées hors des circuits traditionnels de communication de l'entreprise, via le Web social.

Ces données seront considérées comme non structurées dans la mesure où elles vont nécessiter une transformation plus complexe avant de délivrer leur signification.

Qu'il s'agisse d'une image ou d'un son, d'un sentiment ou d'un texte dans une langue quelconque, de géolocalisation ou de capteurs, on comprend aisément le besoin d'algorithmes puissants nécessaires pour un traitement et notamment si c'est en temps réel.

[8] *Exemples de données semi structurées : messages mail, log etc.) ; et non structurées : photo, vidéo, son.*

Ces nouveaux types de données peuvent avoir vocation à enrichir les autres types de données mais elles peuvent constituer dans certains cas, le cœur de l'information à traiter. Cela va dépendre de l'industrie concernée et du processus impacté, comme nous le verrons plus tard dans le chapitre consacré aux usages du Big Data.

Il est clair que ces données doivent se rattacher aux référentiels et aux données déjà en place et qu'il convient donc avant de s'attaquer au Big Data d'avoir structuré et traité la masse des données traditionnelles de l'entreprise, à savoir, selon le sujet d'application : les données transactionnelles, les tickets de caisse, les données de campagnes, les données de navigation, les données issues des capteurs, des sondes, des outils de mesure, des outils d'analyses statistiques ou de visites, de comptage, d'alerte, etc.

À ce stade, il convient de préciser que les Big Data n'ont pas pour vocation de traiter systématiquement la totalité des données disponibles d'un domaine. Tenter de le faire serait contre productif et aboutirait à des projets hasardeux, d'une complexité inouïe, et pour tout dire, inexploitables.

On a souvent oublié dans le passé le vieil adage qui dit que « les arbres ne montent pas au ciel ». Malgré l'enthousiasme ambiant, il est raisonnable de ne pas partir d'une hypothèse pour laquelle il n'y aurait pas de contraintes et de limites.

Si on admet donc, comme principe de base, qu'on ne cherchera pas à traiter toutes les données d'un domaine systématiquement, mais plutôt à se focaliser sur celles dont on a besoin et qu'il est logique de traiter, se pose ensuite la question de savoir où commencer.

Existe t-il une priorité dans le traitement de la donnée, une ou plusieurs sources de données plus prioritaires que d'autres ? La réponse précise à cette question dépend en partie de l'objectif qu'on cherche à atteindre et en partie des données disponibles.

Prenons un exemple dans le domaine des Big Data appliquées au marketing : comment choisir entre une démarche de personnalisation et de ciblage des messages marketing et, une approche de mesure ou d'amélioration de la notoriété ou de la e-réputation d'une marque ? La réponse à la question posée, dans ces deux cas, sera bien différente selon l'ordre dans lequel les sujets auront été traités.

Dans le premier cas c'est l'historique des achats d'un client qui va prendre le plus de place dans votre analyse. Ses achats passés en disent en effet très long sur son comportement d'achat (pour celui qui sait l'analyser). De même pour ses préférences de marque, ses usages, ses besoins, etc. Vous chercherez ensuite à prendre en compte les données d'interaction avec la marque, les données de navigation Web, les informations issues des campagnes marketing, et ainsi de suite.

Dans le deuxième cas, si votre objectif est d'analyser la réputation d'une marque, vous allez prioritairement chercher à interpréter les informations en provenance des réseaux sociaux, des forums, et plus généralement de ce qui se dit sur le Web au sujet de cette marque.

On comprend ainsi pourquoi il est important de se fixer un objectif de départ dans tout projet de Big Data, et que le choix de la priorité du traitement des données découlera de celui-ci.

La règle qui se dégage de notre observation du terrain est que moins

les données sont structurées, plus elles vont nécessiter un traitement important afin de les transformer en connaissance actionnable via un processus de reformatage, qui en outre, se doit d'être intelligent. Il faut donc d'abord s'intéresser aux données structurées disponibles et s'assurer qu'on les exploite bien. Dans un deuxième temps et progressivement, on enrichira la démarche en rajoutant des données non-structurées et les algorithmes intelligents qui vont avec.

Il faut noter par ailleurs, que la quantité de données disponibles ne donne pas de réelle indication de la complexité du traitement nécessaire ; à l'inverse, leur richesse, leur degré de fiabilité et leur structure (ou l'absence de structure) vont être des facteurs beaucoup plus importants à prendre en compte.

Les 3 V des Big Data

Ces nouvelles sources de données sont caractérisées par ce qu'on a coutume d'appeler les 3 V[9] :

• **V comme Volume** : en augmentation annuelle de plus 50%, le volume de données disponibles croit de manière exponentielle. Le croisement de ces données entre elles étant à la base de pertinence de l'information générée, la volumétrie des données est explosive.

• **V comme Variété** : à la diversité des formats (Texte, Photo, Vidéo, Son, Log technique..) s'ajoute une grande variété de fournisseurs internes et externes, objets ou personnes... La variété porte également sur les usages possibles associés à une donnée brute (par exemple un même fichier son généré sur un plateau téléphonique pourra servir à créer un fichier texte [application de speech-to-text] ou à échantillonner la voix en vue d'une reconnaissance vocale ultérieure).

• **V comme Vitesse** : à l'obsolescence rapide d'une partie de ces données issues du temps réel et des médias sociaux (données comportementales ou données exprimant un sentiment[10]), s'ajoute la nécessité d'intégrer au plus vite d'autres données pour générer une information de première fraîcheur.

Cela nous amène à apporter une première précision. Il existe deux grandes familles de projets de Big Data. Celle qui traite de données en temps réel et celle qui travaille sans cette contrainte. Ces deux familles de projets induisent des approches différentes, des architectures techniques différentes, des outils et des données différentes.

Il est facile de se rendre compte qu'un projet de recommandation d'achats en temps réel sur un site de e-commerce et un projet d'analyse comportemental des achats en magasins ne sont pas complètement alignés en termes d'objectifs et donc de moyens à mettre en œuvre.

Dans tous les cas, pour cerner ces problèmes, nous envisagerons une approche en deux temps : l'expérimentation et l'industrialisation.

L'Expérimentation ou « build » correspond à la validation du cas d'usage, la spécification et la mise en forme des données et à la première analyse de celles-ci par un Data Scientist[11]. Cette première analyse permettra l'élaboration de différents modèles prédictifs qui pourront aboutir à une mise en production, c'est à dire à l'automatisation de ces modèles.

Pour cette phase, notre recommandation est de ne pas encore investir dans une architecture mais de préférer une plateforme « as a service » qui permettra d'adapter les besoins au fur et à mesure de l'avancée de l'expérimentation et donc de la maturité des besoins.

---

[9] *Pour la paternité des 3V et les nombreux prétendants à leur invention, voir l'article de Doug Laney : « Deja VVVu: Others Claiming Gartner's Construct for Big Data » : http://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data*

[10] *Appelées aussi « analyse de tonalité » en français, équivalent de « sentiment analysis » en anglais.*

[11] *Voir notre chapitre sur les nouveaux métiers des Big Data à la page 21 pour plus de détails sur ce nouveau profil.*

La problématique autour des modèles prédictifs est complexe et nous y reviendrons en détail dans les chapitres suivants. On pourra déjà retenir à ce stade qu'il existe deux types de modèles, les modèles auto-apprenants qui font appel à des algorithmes d'intelligence artificielle d'une part, et les modèles prédictifs traditionnels basés sur des algorithmes statistiques d'autre part.

Bien que les modèles auto-apprenants nécessitent moins de temps de préparation et d'analyse initiale des données, ils ne dispensent pas pour autant de faire appel à un « Data Scientist » ou un « Data Miner ».

La phase d'expérimentation nous apparaît aujourd'hui comme absolument incontournable dans la méthodologie d'implémentation du Big Data. Ceci est lié notamment à la maturité de ses 4 composants précédemment décrits : Métiers, Technologie, Algorithme et Données.

Une fois l'approche stabilisée, on entre dans la deuxième phase, dite d'Industrialisation ou « run » qui permet d'exploiter des modèles Big Data dans un format automatisé relié au système d'information, donc aux référentiels et aux données de l'entreprise de manière fluide et évolutive.

Pour une phase de « build », qui représente 80 % de l'effort à fournir, dans un environnement « temps réel », nous travaillerons sur des échantillons et nous aurons néanmoins besoin de faire intervenir un ou plusieurs « Data Scientists », pour travailler sur les données « à froid », en utilisant un ou plusieurs logiciels de Data Mining (quels que soient leurs types).

Quand il n'y a pas de temps réel, la phase d'Expérimentation peut se pratiquer sur des environnements de données proches du réel.

La phase d'industrialisation ou « run » nécessite également certains tuning entre exploitation de données en temps réel ou pas. On doit décider ce qui sera traité dans un processus interactif avec les données de l'utilisateur, dans un mode temps réel ; et ce qui pourra passer dans un processus de calcul décalé en mode « batch[12] ».

Il n'est pas possible de réaliser tous les traitements en temps réel car la plupart du temps, ceux-ci requièrent une réponse immédiate, et il est donc impossible de parcourir toute une base de données pour réaliser une analyse complète.

En revanche, les données « chaudes » peuvent être traitées en temps réel à condition de s'appuyer très largement sur les agrégats et résultats qui ont pu être définis préalablement en mode « batch ». On parle alors d'adaptation du modèle prédictif au temps réel.

Les cas qui demandent un traitement véritablement et intégralement en temps réel sont donc en réalité très rares.

Au delà des 3 V : les 5 V

À cette caractérisation classique, sont venus s'ajouter 2 autres « V » qui nous paraissent importants :

• **V comme Véracité** : les données issues des applications centrales du système d'information sont limitées en nombre mais maîtrisées en termes de cohérence, et de niveau qualité. A l'opposé, des données publiques touchant à l'expression de sentiment ou au comportement, peuvent être abondantes mais soumises à des prismes ou des déformations. Dans l'usage qui en sera fait il faudra pouvoir neutraliser ces phénomènes sans pour autant modifier la donnée d'origine. La gestion des critères de véracité des données manipulées est donc une caractéristique induite du projet Big Data. La fiabilité des

[12] *Ou « traitement par lots » en français, c'est-à-dire à l'opposé du mode temps réel, le traitement des données qui ont été préalablement déportées sur un espace de stockage.*

données est devenue un critère essentiel, car l'expression GiGo[13] « garbage in garbage out » s'applique plus que jamais aux Big data. A tel point qu'est maintenant née l'expression « Right Data » par opposition aux Big Data, trop « Big » et pas assez « Right ».

• **V comme Valeur** : s'il est difficile de juger à priori de la valeur d'une donnée élémentaire, il est de bon sens de s'attacher à intégrer des sources de données susceptibles de générer une information dont la valeur ajoutée est avérée. Attention toutefois à ne pas tomber dans un schéma réducteur : Une source de données sans usage interne peut avoir une valeur monétisable pour un partenaire. Une autre source de données peut être a priori sans valeur et s'avérer dans le cadre d'un rapprochement, être porteuse d'un signal discriminant.

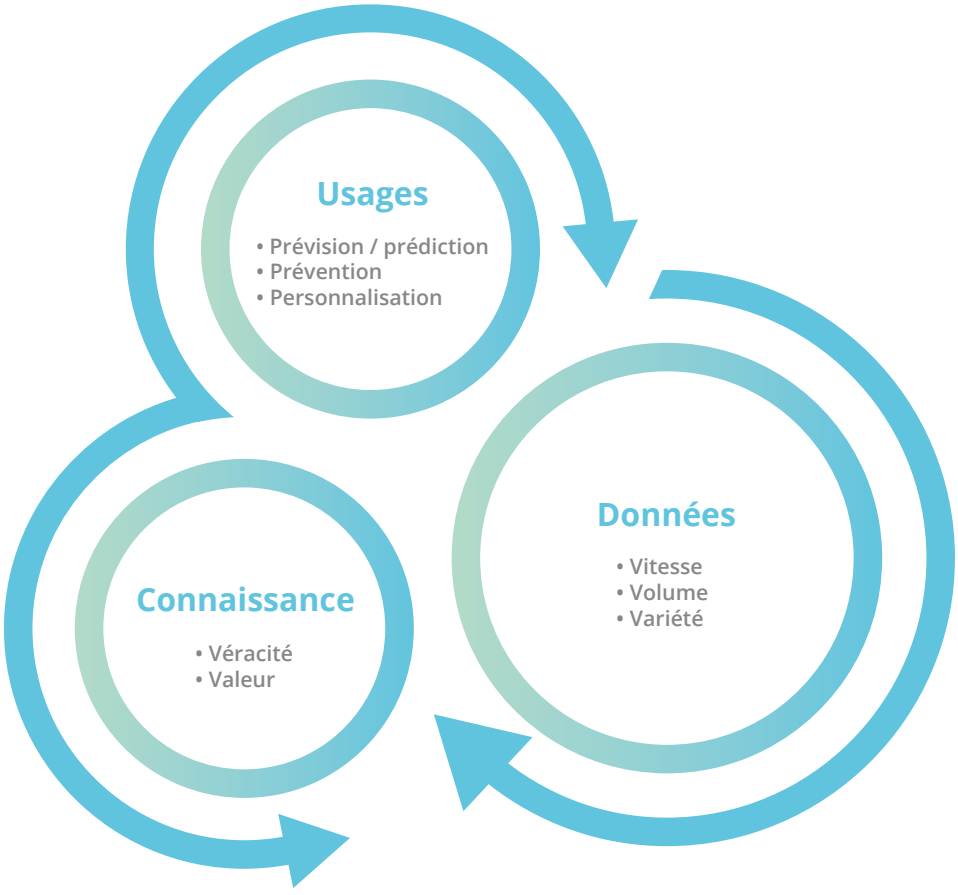


Figure 1. *Les Big Data ne se résument pas aux 3 V, très descriptifs. On peut leur adjoindre 2 V supplémentaires qui permettent de qualifier les données et enfin et surtout, les 3 P qui décrivent la destination des Big Data.*

[13] *Cf. http://en.wikipedia.org/wiki/Garbage_in,_garbage_out*

Au-delà des 5 V, les 3 P

Au delà de cette approche des 3V qui sont devenus 5, utile mais très descriptive, nous pouvons apporter un éclairage supplémentaire en termes de destination des Big Data au travers de ce nous avons appelé les 3P : Prévision, Personnalisation et Prévention, qui soulignent de manière originale le rôle joué par les Big Data dans certains cas d'usage particulièrement pertinents.

• P comme Prévision en vue de l'anticipation

Une grande partie des usages du Big Data porte sur la notion de prévision. Comment exploiter les données pour mieux anticiper ? Comment engranger suffisamment de connaissances pour pouvoir prévoir la demande, les problèmes, les comportements, les goûts, etc. En allant trop loin, on touche évidemment au problème éthique que l'on abordera ultérieurement. En restant trop prudente, une entreprise peut se mettre en difficulté par rapport à ses concurrent et à l'évolution de son marché. Pour prendre un exemple, les Big Data permettent de mieux comprendre le client et ses attentes, en réalisant le croisement des données venant du décisionnel (BI ou Business Intelligence) ou du CRM analytique, (descriptives du comportement de mon client sur les canaux traditionnels), les données de navigation (descriptives du comportement de mon client sur les canaux digitaux web et mobile), et les données captées sur les réseaux sociaux. Ainsi, par leur entremise, je vais pouvoir collecter, agréger et réconcilier l'ensemble des données captées pour y appliquer des modèles d'analyse et de prévision permettant de fournir pour chaque client, des scores d'appétence ou d'attrition d'une finesse inégalée et des recommandations pertinentes et personnalisées ;

• P comme Personnalisation

Une deuxième famille d'usages porte sur la capacité de personnaliser au niveau le plus fin l'interface que propose la solution. Ainsi, au delà d'une composante prévisionnelle, il s'agit plutôt d'une connaissance approfondie d'un environnement qui permet de configurer tout le système spécifiquement pour un groupe de personnes voire un individu. Prenons ici aussi, un exemple. Les Big Data permettent d'imaginer un site web complètement personnalisé selon la personne et le contexte de connexion. Dans ce cas, au moment même de la connexion, le système va comprendre « qui est là », avec un degré de certitude variable et rapidement, « pourquoi il est là », avec également une incertitude possible. En fonction de cela, la page va se construire de manière automatisée, les contenus vont s'adapter, les processus vont se différencier. Le système va analyser en temps réel les interactions avec le client et les croiser avec les informations organisées en base (voir ci-dessus). La solution émet en fonction de ces analyses et de ces interactions, une personnalisation en temps réel ou une offre sur-mesure ;

• P comme Prévention

La troisième famille d'applications porte sur la Prévention. Ici on entend se servir du Big Data pour identifier un risque, un danger et si possible, le prévenir. Ainsi, au delà de la notion prévisionnelle, l'objectif ici est de définir ce qu'est le risque ou ce qui représente un danger potentiel. Prenons quelques exemples. Les Big Data permettent d'identifier des comportements de fraude et d'appliquer en temps réel un schéma de traitement adapté. Dans ce dernier cas, la solution Big Data permet d'appliquer aux données collectées, des modèles d'analyse et de prévision permettant de définir des schémas fins de comportement potentiellement frauduleux et de mettre en place le moteur de règles permettant en temps réel de détecter des comportements conformes à un schéma défini et d'y adapter le workflow de traitement adapté. Ce schéma de fonctionnement ne se limite pas à la sécurité, mais peut se concevoir dans le cadre d'applications liées à la santé et à la prévention des risques par exemple.

Données, informations, connaissance et sagesse... Quelle différence, quelle valeur ?

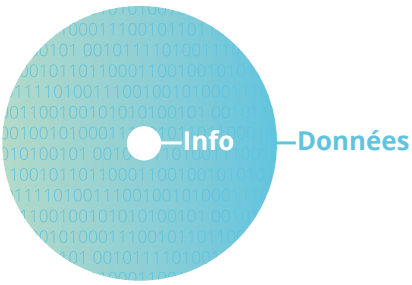


Figure 2. *La donnée est dissociée de l'information, elle nécessite d'être raffinée avant d'être considérée, si possible, comme une véritable information*

La différenciation entre données, informations, connaissance et sagesse a été bien codifiée par Russell Ackoff[14], un théoricien du système et professeur en changement des organisations ; celui-ci a classé le contenu tel qu'il est interprété par l'esprit humain en cinq catégories différentes :

1. **les données** : qui se placent au niveau du symbole ;
2. **l'information** : qui se réfère aux données qui peuvent être traitées de façon à devenir utiles ; fournit des réponses aux questions suivantes « qui, quoi, où et quand » ;
3. **la connaissance** : qui est relative à l'information et au traitement des données par l'esprit humain ; la connaissance répond à la question « comment » ;
4. **la compréhension** : c'est-à-dire la prise en compte du « pourquoi » ;
5. **la sagesse** : l'étape ultime, résultat de l'évaluation de la compréhension.

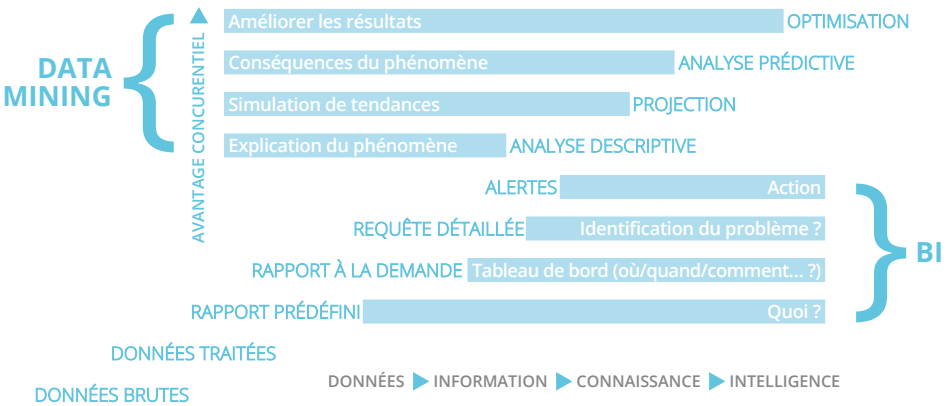


Figure 3. *Les Big Data sont parfaitement adaptées à la théorie d'Ackoff sur la donnée et l'information.*

[14] *Nous nous sommes inspirés librement du texte de Bellinger, Castro et Mills à l'adresse : http://www.systems-thinking.org/dikw/dikw.htm*

Ackoff indique que les quatre premières catégories se réfèrent au présent et au passé, à un savoir fini. La cinquième catégorie se réfère au futur et incorpore la vision et la conception. Avec la sagesse, les individus peuvent imaginer le futur plutôt que de se contenter de comprendre le présent et le passé. Mais arriver à cette dernière catégorie n'est pas aisé, et impose d'être capable de trouver son chemin au travers des quatre premières.

Cette approche linéaire du savoir, séquentielle et progressive est transformée par l'approche des Big Data, même si on peut admettre qu'Edgar Morin[15], au travers de son Introduction à la pensée complexe règle déjà son compte à cette vision déterministe du savoir. Avec les Big Data, on peut littéralement toucher du doigt la Méthode pensée par Edgar Morin.

L'accumulation des données ne voulait rien dire, mais l'usage en a décidé autrement.

À la base, l'accumulation de données, notamment sur Internet et les réseaux sociaux, n'apporte pas de valeur intrinsèque, car dans un mode de fonctionnement traditionnel, la donnée non structurée n'est pas l'information. Les données brutes, non raffinées, non croisées, dans ce cas classique, ne peuvent être exploitées sans un travail préalable, souvent considérable.

Le phénomène des Big Data vient bousculer la vision traditionnelle du monde de la donnée – sans pour autant rendre obsolète la gestion des données de référence (MDM[16] ou Master Data Management en anglais) – et introduisant ainsi une notion d'imperfection et d'incertitude (voir le cartouche sur le marketing de l'incertain).

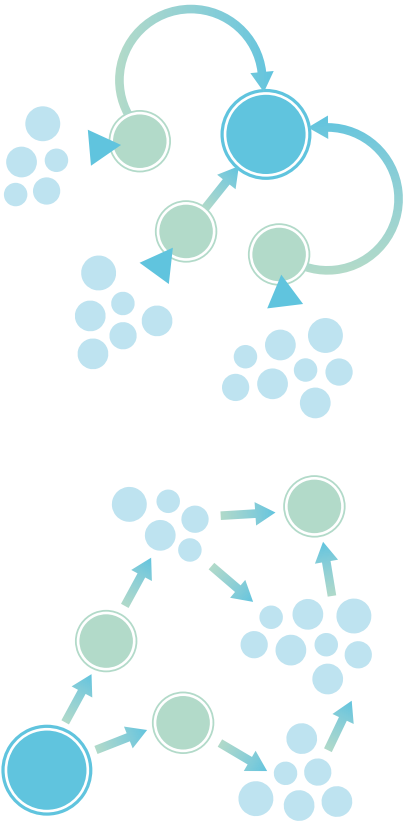


Figure 4. *La démarche déductive classique, en haut, (raffinage, croisement, déduction) et la méthode d'inférence, en bas, typique dans les démarches de Big Data où l'on part d'une hypothèse, qui amène à croiser des données puis à les combiner pour arriver à d'autres hypothèses… qui finissent par donner de l'information.*

Les Big Data font largement appel à la statistique et à l'intelligence artificielle. Un des gros atouts de la statistique étant de s'accommoder naturellement de la notion d'incertitude sur les données. Cette incertitude ne signifie cependant pas qu'il est possible de travailler avec des données d'une qualité trop inférieure.

Mais on peut admettre une certaine tolérance d'incertitude sur les données et donc sur les résultats. Ainsi les résultats produits dans le cadre des Big Data pourront être caractérisés par un indice de robustesse.

[15] *Voir le résumé de la Méthode d'Edgar Morin à l'adresse : http://fr.wikipedia.org/wiki/La_Méthode_(Edgar_Morin)#La_connaissance_de_la_connaissance*

[16] *http://www.piloter.org/business-intelligence/mdm.htm*

Ce qu'on observe sur le terrain, c'est qu'une robustesse de 95 % (soit 95 % de chances que le résultat soit juste), est très acceptable, et une telle donnée sera donc considérée comme fiable. La prise en compte de ce degré d'incertitude autant dans les données que dans les résultats, est donc essentielle dans la mise en place d'un projet de Big Data.

La mesure de cette incertitude se fait au travers d'outils de mesure dont le plus connu est celui dit de « p-value ».

Alors que, a priori, la donnée brute n'a aucune importance, par son croisement et sa validation statistique, elle finit par prendre de la valeur au fur et à mesure, non au travers de la déduction, mais au travers de l'induction et de l'inférence statistique. Sans aller jusqu'à l'intelligence artificielle (où la machine décide de la validité d'une donnée), avec le Big Data, c'est le croisement statistique des données qui finit par créer de la valeur et du sens. C'est ainsi que l'on peut considérer que la donnée donne naissance à l'information, même si cela peut paraître à l'origine contre-intuitif.

les Big Data, moteur du marketing de l'incertain

Une des caractéristiques des Big Data, c'est que les données qui les sous-tendent ne sont pas toujours des données certaines. Les informations recueillies sont si nombreuses et issues de tant de sources, qu'elles permettent des combinaisons qui rendent ces données, *a posteriori*, de plus en plus certaines. Ce sont des signaux faibles, qui, petit à petit, vont devenir des signaux forts, et quand le signal devient suffisamment fort, il va être enfin possible de s'en servir, soit pour alimenter des systèmes en temps réel, soit des machines, soit des études, des identifications d'anomalies, des certifications de cas clients ou d'opportunités et actionner les leviers ou des processus dans l'entreprise…

Lire la suite de l'interview de Patrick Bensabat sur le Marketing de l'incertain et les Big Data sur le blog des Big Data de Business & Decision : www.blog.businessdecision.com

CHAPITRE 03
Les nouveaux usages induits par les Big Data

L'apport des Big Data est de venir chercher dans les données disponibles dans l'entreprise en interne, ou dans son écosystème, le moyen de générer des usages innovants et des facteurs de performance.

S'intéresser aux gisements de données existantes inexploitées

Traditionnellement, les entreprises exploitent avant tout les données **structurées** dans des bases de données relationnelles et associées aux applications de gestion : ERP, CRM... Les autres données, de type bureautique, e-mails, enregistrements audio... ne sont pas partagées ni croisées avec les données structurées, pour enrichir l'information mise à disposition pour la prise de décision.

Un usage des données souvent limité à une application : les données existantes au sein de l'entreprise restent aujourd'hui cantonnées à un mono-usage dans le silo où elles ont été créées et souvent limités à l'usage premier qui les a générés. Par exemple :

- une donnée de compteur pour l'établissement d'un relevé ;
- une donnée journal (log) pour un exploitant informatique ;
- une donnée de géolocalisation pour un push d'offre marketing.

Pourtant, le croisement de ces données élémentaires avec d'autres données du système d'information pourrait apporter des informations d'une grande valeur à un grand nombre d'utilisateurs.

Prenons un exemple : Les sites Internet génèrent des données techniques (logs, tags...) destinés à l'administration technique ou à l'optimisation des sites. Ces données contiennent des informations sur les clicks réalisés par les visiteurs de ces sites Internet ; en rapprochant

ces données techniques des données de connaissance client stockées en base, on entrevoit vite une équation gagnante :

Figure 5. *Comment mieux utiliser les données de l'entreprise.*

En réconciliant finement les deux sources de données (au niveau de la personne, au plus tard à J+1), nous disposons d'une information enrichie permettant de proposer une offre adaptée au client au moment où il en exprime le besoin.

Peut-on réaliser un tel projet sans penser Big Data ? Dans un sens, oui cela est possible, mais au prix d'un accroissement de la complexité et des coûts de développement, et avec la nécessité d'adapter les fonctionnalités à chaque nouvelle hypothèse obligeant à repenser le cheminement du client.

Il est clair que la construction de grands entrepôts transverses de type « Datawarehouse » ont permis de commencer à « dé-siloter » les données. En ce sens, la Business Intelligence a été une étape préalable à la venue du Big Data. Elle a permis aux entreprises de comprendre l'importance des regroupements de données autour de référentiels structurants tels que le client ou le produit.

L'approche était néanmoins fondamentalement différente de part la nature des données et leur volume bien sûr mais aussi de part la philosophie de mise en œuvre de la solution. En effet, pour des raisons liées aux contraintes techniques et financières, les bonnes pratiques de la Business Intelligence consistent à organiser les données par rapport à l'utilisation envisagée. On doit d'abord comprendre la finalité avant de structurer la base de données afin de la rendre optimale. Or les données évoluent très rapidement et de ce fait, chaque mise à niveau du datawarehouse devient une remise en question de la stratégie d'organisation de l'information.

Avec le Big Data, il y a eu un changement de paradigme.

La multiplication des données brutes internes disponibles

Nous avons évoqué dans le point précédent que le développement de l'Internet a généré un grand nombre de données supplémentaires : adresses physiques (IP) des visiteurs, données remontées via les cookies (ou son remplaçant le fingerprinting), données techniques sur le fonctionnement du site (logs), données de statistiques Web (Web analytics/tags) pour garder la trace de toutes les pages visitées, les zones cliquées et de tous les événements, y compris les tests de réactivité des clients à un scénario de navigation (A/B testing).

Plus récemment, le développement des Smartphones et de l'Internet mobile a encore accru le volume de données générées : numéro d'appel, positions géographiques, horodatage de l'activité…

Exemple 1 (page 20) : comment ACCOR a utilisé les Big Data pour accroître ses ventes

Accor, opérateur hôtelier mondial a une présence dans 92 pays. La société est à la tête de 3 500 hôtels répartis sur 15 enseignes et accueille 230 millions de visites par an sur ses sites Internet ; 10 millions de membres font partie de son club d'affaires pour clients réguliers.

Le problème : des clients autonomes qui suivent un parcours très personnel et complexe

50 % de l'activité d'Accor est réalisée en direct au travers de ses canaux de distribution centraux. Cette activité repose sur une offre très large, dont la croissance est tirée par les agences de voyage en ligne.

Le marché du tourisme est totalement « digitalisé » : le client fait son marché en comparant, en choisissant ce qui lui convient et dans un parcours « digital » autonome. L'expérience client, propre à chacun, s'exécute au travers de divers canaux ; elle est déterminante dans le fait qu'un client décide de retourner dans un hôtel où il a déjà séjourné.

Il est donc crucial pour augmenter la performance commerciale de l'entreprise, de connaître intimement les clients et leurs comportements afin de pouvoir leur proposer l'offre la plus adaptée à leurs souhaits, centres d'intérêt ainsi que leurs expériences passées. Cette approche a pour but d'accroître de manière significative l'efficacité des dispositifs marketing d'offre, en contenu comme en délai.

Comment Fidéliser les clients pour accroître le taux de remplissage des hôtels

Pour ce faire, l'hôtelier a dû approfondir la connaissance intime de ses clients, quelque soit le canal utilisé.

Cela lui a permis de proposer des offres en « temps réel » en exploitant l'expérience et les préférences de chaque client, pour toutes les marques du groupe Accor. La connaissance du client est ensuite partagée avec tous les acteurs de la relation client pour mieux répondre à ses attentes

Cela a été possible grâce à la mise en œuvre, en 9 mois à peine, d'un outil de CRM 1 to 1 qui permet de construire une base de connaissance de chaque client, en le liant à une solution de marketing temps réel afin d'optimiser les ventes via les canaux digitaux et les centres d'appels, tant en local, dans les hôtels, qu'en central, pour l'ensemble des marques et enseignes du groupe Accor.

L'information client est enrichie des données issues de l'outil de recommandation d'offres en temps réel. Accor déploie progressivement cette solution dans toutes les unités du Groupe.

Du Big Data au Big Business

La mise en place d'indicateurs de performance a permis de valider les résultats et le retour sur investissement de cette solution de marketing 1 to 1.

- Accor, grâce à ce système de marketing personnalisé, diffuse 1 200 000 recommandations d'offres personnalisées par jour ;

- La base de données clients est passée de 20 à 35 millions de contacts ;
- Le taux de clic sur les invites diffusées sur les pages du site internet ont été multiplies par deux grâce à la personnalisation des messages ;
- Les taux de clic, ratios de conversion, la mesure du cycle de vie de chaque client, et bien d'autres indicateurs sont mesués et consignés dans les tableaux de bord;

Ce projet de Big Data appliqué au marketing a permis de lancer une véritable offre de marketing 1 to 1 sur un marché de masse en alliant intelligence logicielle et action humaine.

Il s'agit d'un des projets les plus ambitieux d'Europe dans le domaine du tourisme, un cas d'école de « transformation digitale ». Ce projet est une réalisation Business & Decision.

L'utilisation d'une carte SIM, d'une connexion Bluetooth, ou de tout autre protocole n'étant pas limitée aux personnes physiques, l'Internet des objets va générer également un volume de données considérable dans un grand nombre de secteurs : Automobile, Santé, Distribution d'énergie…

Selon Michel Lévy-Provençal, le nombre d'objets connectés dans le monde, principalement des ordinateurs, des téléphones et des tablettes est estimé à 5 milliards, en 2015, ils devraient atteindre 15 milliards et 50 milliards en 2020[17]. Il fait aussi mentionner l'arrivée en masse du paiement sans contact et, notamment, de iBeacon qui va faire donner à cette tendance une importance croissante.

La profusion de données externes, publiques ou achetables

Les Réseaux Sociaux (Twitter, Facebook, LinkedIn etc.) génèrent eux-aussi des volumes de données considérables. Accessibles publiquement via des API (Application Program Interface[18]), des programmes d'interface qui permettent de siphonner les données d'une application et de les réinjecter dans une autre, ces données « sociales » peuvent constituer une source d'information pour des entreprises dont la réputation est ainsi exposée au grand jour : Biens de consommation, Agroalimentaire, Luxe, Distribution et peut-être demain Assurance sont des secteurs qui disposent ainsi d'une manne de données externes abondante.

Considérées comme « non structurées » ces données, en grande partie textuelles, embarquent également de la vidéo et des photos. Si elles sont abondantes, ces données non structurées (commentaires, avis…) ont une pertinence intrinsèque limitée. Les « signaux faibles » qu'elles génèrent n'auront de vraies valeurs que croisées à d'autres données[19].

La problématique à laquelle l'entreprise est confrontée est l'expansion des champs applicatifs générés par la maîtrise de ces nouveaux types de données. Si on s'intéresse par exemple à la reconnaissance faciale, qui est l'une des priorités de nombre des géants de l'internet. Le potentiel de création de valeur des applications liées à cette donnée justifie les investissements colossaux qui sont réalisés actuellement. Dès qu'elle sera maîtrisée, il faudra pour l'entreprise concernée, mettre à jour tous ses référentiels et bon nombre de ses applications marketing si elle ne veut pas se retrouver à la traine face à de nouveaux entrants ou des concurrents plus prompts à réaliser des investissements dans ce domaine.

[17] *Cf. cet article sur le site de SFR et cette interview de Michel Lévy-Provençal : http://bit.ly/sfrlevyp*

[18] *http://encyclopedia2.thefreedictionary.com/Application+Program+Interface*

[19] *Il est à noter cependant que des progrès notables restent à faire dans le domaine de l'exploitation de ces données multimédia.*

Exemple 2 : collecte de données sur une flotte de véhicules pour un grand opérateur de services aux particuliers européen

Un grand opérateur de services aux particuliers en Europe recourt aux Big Data pour mieux gérer sa flotte de véhicules.

À la base il y a la collecte des données sur la flotte et ses tournées : relevés de compteurs kilométriques en fin de mois, des cartes d'essence des chauffeurs, les données issues des révisions au garage, et depuis peu, les capteurs implantés dans les véhicules électriques. **Toutefois, en raison des déclarations approximatives, ces données peu fiables restaient jusque-là inexploitées.**

Le Client a donc fait appel à Business & Decision pour concevoir et mettre en œuvre un prototype sur Hadoop[20]. Les données citées plus haut sont traitées sur la plateforme Hadoop pour être fiabilisées, puis injectées dans l'outil de Business Intelligence QlikView[21] afin de restituer des indicateurs d'éco-conduite permettant d'améliorer la gestion de la flotte.

Convaincue par ce premier essai, la DSI de l'opérateur envisage de promouvoir cette technologie auprès des différentes entités de son Groupe.

Les données ouvertes ou « Open Data »

L'Open Data regroupe un ensemble de données publiques gratuites mise à disposition par les organismes publics comme l'Etat ou les collectivités locales, il peut être un choix stratégique pour des entreprises comme la RATP qui met ainsi à disposition, progressivement, une partie de ses données de fréquentation[22].

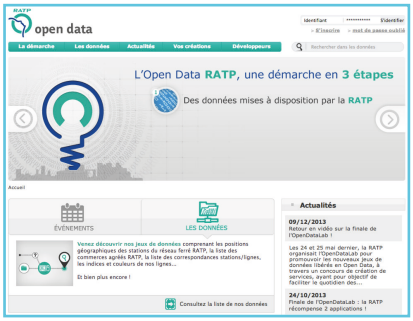


Figure 6. *data.ratp.fr, la plateforme d'Open Data de la RATP.*

Monétiser ses données

Quand on parle de données internes ou externes, il est important d'avoir à l'esprit la monétisation possible de ces nouvelles données : les données de géolocalisation captées par un opérateur téléphonique, par exemple, ont-elles une valeur pour des secteurs qui doivent s'intéresser de près à la mobilité de leurs clients (Assurance, Voyagiste, etc.) ?

Les données fournies par un routeur domestique de connexion Internet (« Box »), par exemple, peuvent fournir nominativement, à un instant donné, des informations sur qui regarde quelle publicité, données précieuses pour les annonceurs ou les publicitaires.

[20] *Pour des explications sur Hadoop et Map Reduce, voir la page 20.*

[21] *Voir le site : www.qlik.com/fr*

[22] *Sur la conversion progressive de la RATP à l'Open Data, voir cet article sur le blog du Monde : http://data.blog.lemonde.fr/2013/01/09/stations-desertes-temperatures-quand-la-ratp-ouvre-ses-donnees/ agrémenté d'exemples de ce qui peut être fait avec ces données.*

La question qui se pose est comment va s'autoréguler ce marché naissant de la monétisation des données. À terme, chaque entité peut devenir un fournisseur et un consommateur de données. A la manière de ce qu'on anticipe sur l'énergie, il faudra créer des systèmes sophistiqués qui permettront de véhiculer et de commercialiser les données entre tous ces possibles intervenants aux multiples casquettes.

Le croisement de données tous azimuts

De ce qui précède nous pouvons retenir deux choses.

D'une part, la variété et le volume de données est une réalité qui ne fera que s'affirmer dans les prochaines années.

Ensuite, que chaque donnée élémentaire peut répondre à différents usages et besoins en informations internes à l'entreprise ou dans le cadre de la coopération avec d'autres entreprises.

A ce stade, il est important de prendre conscience de deux éléments complémentaires.

La multiplication des croisements de données va s'accélérer. Le croisement de données pertinentes pour des usages maitrisés peut être source d'opportunités; on l'a présenté sur le cas de 2 types de données : CRM et navigation Internet. Mais des croisements multiples avec d'autres sources de données seront peut-être pertinents en terme de création d'une information de valeur.

Le croisement de données multiples devient en effet une nécessité à partir du moment où on intègre à l'ensemble des données non structurées porteuses de signaux faibles : plus on pourra recroiser la pertinence d'une donnée de ce type avec d'autres données de même type, meilleure sera la qualité de l'information produite.

L'échelle de temps pour le croisement de données se réduit considérablement, et pour certains usages se rapproche du temps réel. En effet, la volatilité de certaines données captées nous incite à les exploiter au plus près du moment où elles ont été produites.

C'est dans ce contexte que se posera de la manière la plus évidente, le problème de la confidentialité de l'information. L'information prend de l'amplitude quand elle est constituée de données issues d'univers différents. Quand j'achète un billet d'avion, les données générées prennent beaucoup de valeur pour un hôtelier, un assureur, un loueur de voiture, etc. Et plus le temps passe, moins l'information a de la valeur. Il s'agit donc ici, de savoir comment exploiter l'information dans les règles de l'art, les contraintes juridiques et dans le cadre d'un parcours client efficace pour en tirer le meilleur parti. Mais comment le client peut-il contrôler ou autoriser ces échanges de données ?

Peut-on envisager de tirer avantage de toutes ces données sans se poser la question d'une approche spécifique dite de Big Data ? Quand on dit Big Data, il faut entendre au-delà de la technologie, une approche avec toutes ses composantes : technologiques, méthodologiques, juridiques et sociales. Cela est probablement possible, mais se ferait au prix d'investissements matériels, logiciels et humains très importants.

CHAPITRE 04
Les architectures et les algorithmes

Les spécificités d'architecture matérielle en Big Data

La genèse du massivement parallèle

Les technologies du massivement parallèle sont à l'ordre du jour depuis 20 à 25 ans. Les Informaticiens ont adopté différentes façons de traiter leurs problèmes techniques : au travers des machines vectorielles, avec des mainframes, avec des serveurs, puis en ajoutant de plus en plus de processeurs, et enfin aujourd'hui, avec des processeurs qui ont de plus en plus de cœurs, ou encore des machines regroupées en clusters.

Aujourd'hui, cette parallélisation est devenue la solution standard. Le massivement parallèle, qui nécessitait des investissements matériels conséquents (avec 8 ou 16 processeurs, on payait cela l'équivalent de 100 000 euros pour un matériel qui n'avait qu'une durée de trois ans, et qui au-delà était obsolète), notamment si on désirait ensuite ajouter de nouveaux processeurs ; sans compter la captivité des clients vis-à-vis de leurs fournisseurs, une fois les premiers investissements réalisés. En cas de changement technologique, il fallait recourir à un processus de migration qui servait à valider la manière dont les programmes tournaient, avec quelle performance, et cette migration avait un coût : la montée en charge (« scalability » en anglais) n'était donc que relative.

Ce sont les infrastructures cloud computing (IaaS[23]) qui ont rendu possible les Big Data

Tout ceci a changé avec l'arrivée du cloud computing. Désormais, le client a la possibilité de monter facilement en charge (principe d'élasticité dans le cloud) car on a rendu les machines universelles et extrêmement simples.

Cette standardisation des infrastructures a permis de réaliser des calculs multiprocesseurs, parallèles, sur des machines et des systèmes d'exploitation standard, très faciles d'accès. Ce n'est pas la puissance de calcul des serveurs qui a changé, c'est la façon de monter en charge.

[23] *IaaS : Infrastructure as a Service, c'est-à-dire la capacité d'acheter de l'infrastructure déportée et de la consommer à la demande, de la même manière que ce que l'on fait pour les logiciels SaaS (Software as a Service).*

Auparavant, en cas d'obligation de montée en charge sans avoir le budget nécessaire, on revoyait ses ambitions à la baisse.

C'est le cloud computing qui a rendu possible les Big Data. Avec cette facilité de montée en charge, on peut mettre sur ces machines virtuelles des systèmes qui sont développés et conçus pour les traitements parallèles.

Les technologies Big Data immergées dans le cloud

Si l'on prend l'exemple de Cassandra de Facebook, il s'agit d'une base de données standard, dont la spécificité est qu'elle a été conçue pour s'intégrer sur des clusters, que l'on gère dans le cloud. C'est donc bien le cloud qui est l'élément déclencheur de ce phénomène Big Data en termes technologiques. Cassandra, en cas de besoin, va provisionner des machines supplémentaires. Cette opération de provisionnement est aussi possible par le simple envoi d'un SMS sur une machine système qui déclenche une action d'administration du serveur.

Ce qui est vrai de Cassandra est aussi vrai pour Hadoop qui est un « Framework » logiciel[24] de traitement massivement parallèle ; il est donc logique qu'on ait une plate-forme massivement parallèle sous-jacente qui la fasse tourner.

Hadoop est-il la panacée des Big Data ?

Le modèle Hadoop a des capacités de montée en charge indéniables. Est-il pour autant adapté à toutes les problématiques des Big Data ? On peut, sans crainte de se tromper, affirmer le contraire.

Certains traitements statistiques, comme par exemple l'évaluation des profils des clients-types d'une base de données, nécessitent d'accéder à quasiment l'intégralité de la base au sein du même algorithme et s'accommodent assez mal de la méthode Map/Reduce[25]. On peut certainement effectuer un modèle de scoring en utilisant Map/Reduce, mais la mise à jour d'une classification comportementale, par exemple, nécessitera de centraliser à nouveau la donnée pendant toute la durée du calcul.

Cela limite quelque peu l'usage de Hadoop dans le cadre de traitements, en particulier ceux destinés au Marketing, mais ne l'interdit pas pour autant. Il est en effet concevable de faire les traitements lourds en mode batch dans des bases SQL qui viendront en soutien de bases Hadoop plus orientées temps réel.

Il faut comprendre que les différentes structures de base seront amenées à coexister et qu'il n'existe pas aujourd'hui de modèle de stockage qui soit la panacée en Big data.

Les spécificités d'architecture logicielle en Big Data

D'un point de vue logiciel, les spécificités des Big Data découlent de ce qui précède. À une époque, les calculs parallèles étaient réservés à une communauté réduite de scientifiques, de développeurs et de spécialistes très pointus, notamment dans le domaine des jeux vidéo. Ces spécialistes étaient formés dans des écoles spécialisées ou au travers de cursus de formation dédiés. C'était une connaissance qui ne se diffusait que très peu.

Le calcul parallèle s'est démocratisé

Aujourd'hui, les choses ont complètement changé : le cloud computing, puis les outils comme Map Reduce, Hadoop et Cassandra, tous massi-

[24] *Un framework logiciel est un ensemble méthodologique et d'outillage lié à un langage de programmation. Cf. http://fr.wikipedia.org/wiki/Framework*

[25] *Pour l'explication de Map/Reduce, voir la page 19 de ce document.*

vement parallèles, obligent tout le monde informatique à se confronter à ce qu'est un développement de calcul parallèle, voire distribué. Nous sommes donc dans une phase de transition où il faut que tous les informaticiens développent ces nouvelles compétences.

Tout développeur qui veut pratiquer le Web, du reporting temps réel, des analyses de Twitter ou de tonalité (« sentiment analysis » en anglais) est obligé de comprendre ce qu'est un algorithme distribué, d'avoir des notions de maths appliquées, d'analyse numérique, et qu'ils sachent ce qu'est un algorithme. Les cours d'algorithmique, il y a encore quelques années, étaient facultatifs dans les écoles, aujourd'hui ils sont devenus incontournables et obligatoires.

Map Reduce : un modèle de programmation en deux temps

Map Reduce est un modèle de programmation, un algorithme qui est principalement utilisé pour manipuler beaucoup de données qui sont distribuées sur des clusters ou des machines parallèles. Ce processus se déroule en deux étapes :

Premier temps : Map pour découper le problème en morceaux

Le problème est d'abord découpé en morceaux plus ou moins gros ; on va ensuite dédier chaque machine à un sous-problème particulier. Chaque machine va elle-même redécouper ce sous-problème en sous-sous-problèmes, et de manière récursive, on arrive à ce que chaque machine traite une toute petite partie du problème.

Deuxième temps : Reduce pour regrouper les morceaux

Chaque nœud va utiliser la deuxième étape (Reduce) qui consiste à faire un calcul manière indépendante (Shuffle) et à le remonter vers la machine qui lui a donné l'ordre de faire ce calcul, jusqu'à revenir au nœud initial.

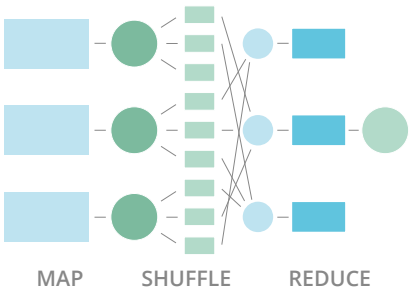


Figure 7. *Comment Map Reduce fonctionne (librement inspiré d'un schéma du blog sqlauthority.com)*

Avec cette méthode, on découpe les problèmes soit pour arriver un problème suffisamment petit pour pouvoir être traité par l'algorithme, soit pour s'adapter à la puissance de la machine.

Tout cela remonte enfin et donne le calcul final. Ce principe est commun à un grand nombre d'algorithmes de traitement des données.

Les limites de cette méthode

Certains problèmes ne se prêtent cependant pas à cette méthode. Il faudra dans ce cas les reformuler, ou trouver un nouvel algorithme du même type qui pourra traiter le problème en question.

Les spécificités des bases de données en Big Data

Jusqu'à maintenant, Nous avions tendance à stocker toutes les données de la même façon sans réfléchir à leur pertinence, dans des bases de données relationnelles qui sont plutôt orientées « lignes ».

Les bases de données lignes

Comme leur nom l'indique, dans ce type de bases de données, chaque enregistrement est égal à une ligne (c'est le cas par exemple pour les bases de données de type Oracle, DB2, SQL, SQL Server). La raison à cela en était soit le manque de choix, soit le manque de performance, ou de maintenabilité. C'est-à-dire des raisons étrangères au problème à résoudre. Le format ligne est également appelé format « tableau » et est comparable à ce qu'on trouve dans les tableurs Excel.

	Produit	Description	Prix au kg en euros	Durée du SAV en années
1	chou	légume vert	1	2
2	carotte	légume orange	10	5
3	navet	légume sans goût	15	10

Figure 8. *Dans un modèle ligne, même spécifique, une valeur pour chaque colonne ; exemple ici avec une base de données produits*

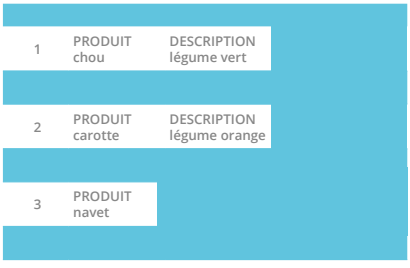


Figure 9. *Au contraire, avec une base de données colonne, comme dans l'exemple ci-dessus, le nombre de colonnes peut varier pour chaque enregistrement.*

Les bases de données colonnes

Il existe aussi des bases de données dites « colonnes », pour lesquelles, au lieu de stocker des lignes, on va sélectionner une colonne en particulier, en fonction de l'algorithme qui nous intéresse. Dans le domaine du décisionnel, on modélise des entrepôts de données (« data warehouses » en anglais), avec des modèles flocons ou étoiles qui permettent ensuite de faire des recherches très rapide d'enregistrement.

Cela se modélise très bien avec des bases de données colonnes (avec des produits comme SAP et Sybase IQ) où dans le monde Big Data, Hbase[26], qui est une base de données « colonnes ».

Les bases de données orientées documents

Enfin, on trouve les bases de données orientées documents (exemple MongoDB) : dans ce cas, on va stocker uniquement des couples clé/valeur («key/value» en anglais) en vrac. Ce sont des bases de données très déstructurées, où on stocke tout en vrac et où les algorithmes de recherche de canalisation des données vont utiliser les spécificités de la base de données pour améliorer la performance.

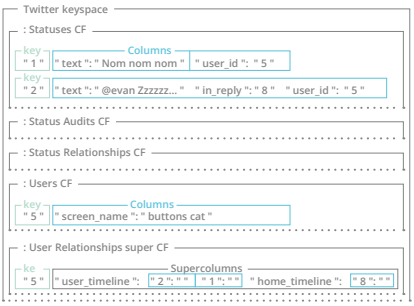


Figure 10. *Evan Weaver décrit comment Cassandra traite les enregistrements de données dans sa base clé/valeur.*[27]

[26] *Cf. http://www.journaldunet.com/developpeur/outils/comparatif-des-bases-nosql/hbase.shtml*

[27] *http://blog.evanweaver.com/2009/07/06/up-and-running-with-cassandra*

Pas de prêt-à-porter dans les bases de données

Le choix de la base de données dépend des besoins. Dans une base de données relationnelle, une recherche ou une sélection en fonction d'un champ donné, par exemple dans une base de données clients, tous les clients qui s'appellent « Michael », est une opération très peu coûteuse.

Par contre, si je veux faire beaucoup d' « inserts » avec des « jointures »[28] entre une base de données clients et une base de données des produits achetés dans le passé pour faire des statistiques, c'est un peu moins performant. Ce même genre d'opérations sur une base colonne est quasi-instantané car elles sont faites pour cela.

Beaucoup d'outils, chacun spécialisé dans un domaine

Les recherches en plein texte mettront à mal les bases de données en colonnes ou en lignes très rapidement. A l'opposé, avec MongoDB, c'est l'inverse, la recherche plein texte y est très rapide.

Le choix de la base de données dépend donc de ce que l'on recherche à y faire. En dehors de cela, il y a aussi des bases de données qui cherchent à faire la synthèse de tous ces mondes.

On trouve déjà à ce jour, entre 50 et 110 de ces bases hybrides qui sont très régulièrement utilisées, et qui seront choisies en fonction des problèmes à traiter, sans compter les nouveaux formats comme le format Graph développé par SAP/INFINITE INSIGHT[29] spécialement adapté à l'analyse des interactions des réseaux sociaux.

C'est là une des spécificités et des complexités du monde des Big Data : il y a beaucoup d'outils, mais le parti pris de la communauté a été de préférer les outils qui sont très spécialisés, simples voire même simplistes, mais qui sont les meilleurs dans chacune de leurs catégories.

Architecture Big Data uniquement pour du Big Data ?

C'est une question dont la réponse est donnée par le marché. Les entreprises ont déjà commencé à adapter certains composants de ces architectures pour améliorer les performances de leur système d'information. L'enjeu ici est iso-fonctionnel, c'est à dire qu'il s'agit de refaire la même en chose en mieux et en moins cher. Par mieux, on entend plus rapide, plus de temps réel et par moins cher, on souhaite profiter de la profusion d'entreprises innovantes qui viennent challenger (avant d'être rachetées ?) les leaders établis du secteur des nouvelles technologies. Les cibles sont les systèmes les plus coûteux et les plus complexes, les ERP, le CRM, la BI, etc...L'approche est, elle, purement technique.

[28] *« insert (SQL) » : http://fr.wikipedia.org/wiki/Insert_(SQL) « jointure » ou « join » en anglais : http://fr.wikipedia.org/wiki/Jointure_(informatique)*

[29] *Autrefois appelé KXEN : Voir http://en.wikipedia.org/wiki/KXEN_Inc.*

CHAPITRE 05
les métiers des Big Data

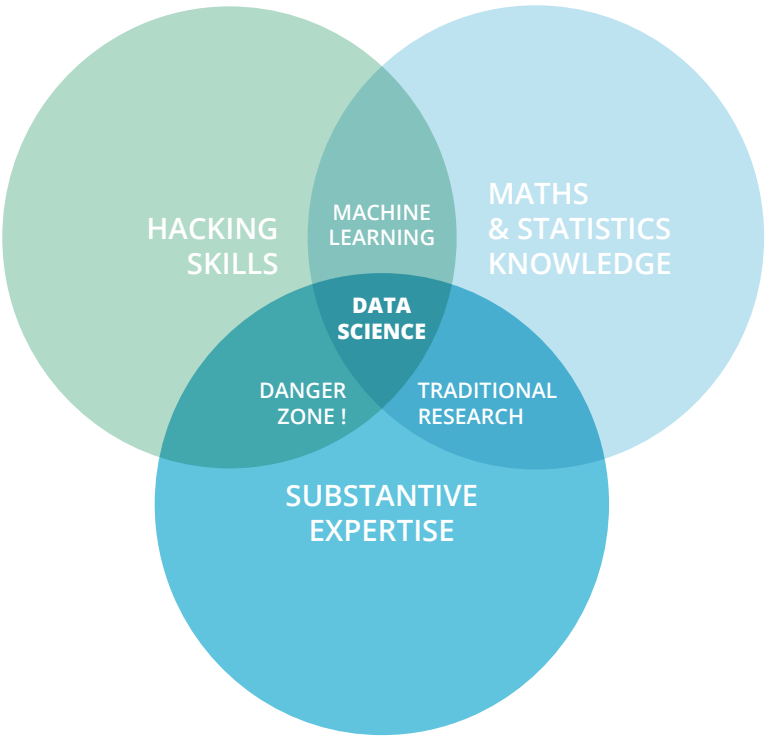


Figure 11. *La science des données (selon Drew Conway)[30] insiste sur la combinaison de connaissances de fond (substantive expertise), de talents techniques (hacking skills) et de connaissances en mathématique et en statistique. C'est la combinaison de ces expertises qui permet d'éviter les écueils décrits par Conway. Nul doute qu'il faille former des experts…*

[30] *Le diagramme de Venn des data sciences de Drew Conway : http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram*

Le retour de l'EIM (Entreprise Information Management)

L'EIM est à nouveau à l'ordre du jour. Le devoir de réflexion que nous nous sommes imposés dans le cadre de ce livre blanc nous oblige à nous pencher à nouveau sur cette notion inventée par Gartner, ainsi qu'à prendre un peu de recul par rapport aux Big Data. Ce terme d'EIM, devenu désuet pendant un temps, revient sur le devant de scène.

Les initiatives autour de la donnée dans les entreprises sont souvent le sous-produit de la mise en œuvre d'un ERP (SAP, JD Edwards ou autre), qui révèle des problèmes de gestion des données de référence (MDM).

Les clients sont devenus beaucoup plus mâtures sur ce sujet. Le besoin de gérer la donnée de référence n'est plus à démontrer : même si cette maturité reste très relative, car les progiciels de gestion de la donnée de référence ont une dizaine d'années ou un peu plus. Si les clients sont de plus en plus attentifs sur ce point, ils ignorent en revanche comment mettre en œuvre ce chantier.

Comment mettre en œuvre ce chantier ?

La mise en œuvre de l'initiative autour de la donnée, c'est vrai du MDM mais aussi du Big Data, nécessite d'avoir une vision rigoureuse de la donnée. Cela a plusieurs impacts :

D'abord, ceci nécessite de connaître ses données, identifier les gisements de données sur un périmètre.

Ensuite, cela implique qu'on ait des données de quantité et de qualité (le MDM implique que la donnée de référence soit le point de vérité unique dans le système d'information). Le MDM concerne les produits, les clients, la RH...

Enfin et surtout, cela exige la mise en place d'une gouvernance : cela va se traduire par l'énonciation des devoirs et responsabilité des différents acteurs (dans et hors de l'organisation) à chaque étape du cycle de vie de la donnée.

Le résultat en est la valorisation de la donnée car celle-ci est un capital de l'entreprise dans le sens où elle a une valeur intrinsèque.

La difficulté de mise en œuvre de ce type de chantier est due au nécessaire changement de l'organisation pour tenir compte de la donnée ; la plupart des entreprises découvrent le niveau de qualité de leurs données au cours de ce processus, et à l'issue de ce constat, mettent en place une gouvernance des données. Cette approche est menée par des opérationnels, le sponsoring des patrons de l'entreprise.

Valoriser la donnée en la rendant aux métiers

Pour sortir du « bla-bla marketing », il va falloir démontrer la valeur de cette donnée. C'est là qu'on va parler de vision (Quoi ?) et stratégie (Comment ?) et métriques (Combien ?).

La gouvernance évoquée plus haut a pour but de redonner la main aux métiers sur ces données. La problématique que l'on retrouve systématiquement chez les clients, c'est que cette initiative de la donnée est essentiellement une initiative informatique. Au fur et à mesure de l'accroissement de la maturité des entreprises, la responsabilité de la donnée doit être rendue aux métiers. Le vrai rôle de l'informatique est de faire en sorte que les gens qui ont la maîtrise de la donnée passent du temps à l'analyser, et non à la préparer.

Tout ceci est complémentaire du Big Data

Tout ce qui précède est utile afin de comprendre les nouveaux métiers des Big Data, qui ne sont pas, comme dans le cas du MDM cité plus haut,

des métiers informatiques, mais des métiers business. *Gartner*, dans sa hiérarchisation des nouveaux métiers du Big Data, montre clairement que ceux-ci sont bien rattachés aux métiers et non à l'informatique.

Les nouveaux métiers du Big Data

Nous pouvons, pour simplifier, répertorier 4 catégories de métiers si non liés directement au Big Data, tout au moins rendus visibles et attractifs grâce à la prise de conscience des enjeux liés aux data, créée par le Big Data :

- le CDO (Chief Data Officer) ;
- le Data Stewart : c'est l'administrateur des données ;
- le Data Scientist : c'est celui qui analyse la donnée à l'aide d'outils statistiques et datamining complexes ;
- le Data Analyst : c'est celui qui analyse les données pour ses besoins métiers propres.

1. LE CDO

C'est encore, à ce jour en France, un poste qui n'existe pas beaucoup dans la réalité. Mais c'est un concept qui monte et qui va se développer. Il est au niveau des décideurs de l'entreprise (C-Level), et il participe au comité exécutif. Son rôle est multiple :

D'abord, contribuer à la stratégie de l'entreprise en s'appuyant sur les données, leur gestion, éventuellement en maintenant le niveau de qualité de celles-ci;

Puis diffuser la connaissance en interne des données (dans les grandes entreprises il y a des gisements de données dans toutes les entités, mais personne n'a de vision transversale). Optimiser les processus clés du métier via la consommation de la donnée.

Enfin, construire une équipe avec des profils différents pour réaliser cet objectif.

Le CDO a suffisamment de pouvoir pour passer outre toutes les résistances au changement et mettre en œuvre une stratégie de changement dans l'entreprise. Ce n'est pas tout : le CDO a aussi une dimension externe, notamment dans le cas où l'entreprise partage des données avec l'extérieur. Il est aussi acteur dans les secteurs fortement normés et/ou réglementés (comme les normes GSI dans la distribution par exemple).

Les CDO peuvent devenir acteurs de ces normes, de la fédération des différents intervenants dans un secteur d'activité, pour mettre en place une norme, faciliter le partage des données etc. En l'absence de nomination de CDO, en France, c'est de facto le DSI qui en remplit le rôle.

Un métier où la communication est primordiale

Le rôle du CDO est aussi le mettre en place une organisation opérationnelle, ainsi que la gouvernance au sein d'un « Governance Board » ; le mot faisant peur, on a tendance à le remplacer par celui de « Governance council ».

Par ailleurs, en filigrane de tous ces métiers, il y a l'importance de la communication. Car ces nouveaux métiers (pas seulement celui du CDO mais les deux autres également) doivent beaucoup communiquer afin de prouver le bien fondé de chacune des actions entreprises autour de ces données.

Le paysage va donc immanquablement évoluer lors des cinq prochaines années.

La question en suspens est de savoir si le CDO est un poste de transition. En fait, cela dépend du DSI (CIO) et de l'évolution de son poste, notamment si celui-ci n'est pas capable de dépasser les tâches liées aux infrastructures. Il reste néanmoins une incertitude sur ce point.

2. Le Data Stewart

Ils ont une connaissance de la donnée, et la travaillent quotidiennement, même si ce n'est pas nécessairement un travail à temps plein. C'est le poste le plus bas dans l'organisation de ces nouveaux métiers, c'est un « faiseur ». Il doit dépendre du CDO, car il fait partie d'une communauté, il n'est pas seul. Dans les grandes entreprises, il y aura un Data Stewart par territoire. Si c'est une organisation matricielle, on y rajoutera une dimension métier. Les Data Stewart sont responsables de la mise en œuvre de la stratégie sur le terrain, ils vont appliquer la gouvernance décidée par le CDO, et veiller à ce qu'elle soit suivie ; de même pour ce qui est des bonnes pratiques et des cycles de vie.

3. Le Data Scientist

Le Data Scientist est celui qui produit la valeur de la donnée. Il part de données fiables (grâce au travail du Data Stewart), et il a les outils pour le faire. Les Big Data sont récentes, mais les problèmes liés aux Big Data existent depuis longtemps. C'est la technologie récente qui permet de traiter les plus gros volumes et aussi de faire ce travail en temps réel. Le Data Scientist est un expert aux multiples compétences. Il maîtrise les outils statistiques et le datamining pour pouvoir manipuler les données à sa guise ; il connaît suffisamment bien l'industrie pour laquelle il opère afin de tenir compte de ses enjeux dans ses recherches ; enfin, il est capable de comprendre les finesses d'un processus pour pouvoir se poser les bonnes questions tout en suggérant des pistes de réponses.

Par exemple, pour une entreprise dans le secteur des Télécoms, la problématique du Churn ou Rétention des Clients est cruciale, car la concurrence est rude et le produit souvent pas assez différencié. Les mécanismes liés à la compréhension nécessitent de la part du Data Scientist, une compréhension du secteur qui a ses spécificités, une maitrise des enjeux liés au Churn et une maîtrise des outils de statistiques et de Datamining les plus couramment utilisés par ces industries.

On le voit, le Data Scientist est souvent un mouton à 5 pattes, doté d'un cursus de formation avancé et d'une expérience professionnelle avérée.

4. Le Data Analyst

Le Data Analyst est également quelqu'un qui produit la valeur de la donnée. Il réceptionne une partie du travail du Data Scientist et le rapproche des autres reportings et des autres données qu'il a en sa possession pour pouvoir faire son travail. Il utilise des outils de dashboarding, de visualisation de l'information et d'exploration de l'information assez proches de ceux qu'il utilise pour de la Business Intelligence mais il les applique différemment en fonction des données qu'on met à sa disposition et des enjeux métiers auxquels il doit faire face.

Le Data Analyst n'est pas un technicien, c'est un professionnel des métiers qui a une sensibilité à la donnée. Par exemple, chez un client dans les arômes en Suisse (Givaudan), il est important voire crucial, de rassurer les clients sur la fiabilité et les risques liés aux produits. C'est dans ce contexte une obligation réglementaire. L'utilisation des données permet de stabiliser et de pérenniser l'activité. Des outils traditionnellement issus de la Business Intelligence existe et sont utilisés par des Data Analysts. Ils sont aujourd'hui complétés par des solutions de type Big Data et c'est le Data Analyst qui sera en mesure de consolider les différentes sources de données et de faire évoluer ses méthodes afin de faciliter et d'améliorer ses processus.

De la « punition » aux perspectives de carrière

Ce travail sur la donnée va servir à optimiser les processus métiers et à améliorer les facteurs de performance (KPI) de ces métiers.

Ces nouveaux métiers offrent désormais des perspectives de carrière et ne sont plus une « punition ». Ces postes étaient traditionnellement considérés comme des « placards » ; ce n'est plus vrai aujourd'hui.

Les grandes écoles (citons HEC, ENSAE, Essec) proposent aujourd'hui des formations autour de la donnée. HEC a intégré une dimension donnée dans son programme de MBA, c'est un signal fort. Et cela s'adresse à des gens qui sont amenés, par cette formation, à remplir des rôles importants dans les organisations dans lesquelles ils travaillent.

Le lien entre MDM et Big Data

À l'Entreprise Information Management Summit 2013 à Londres organisé par *Gartner*, le cabinet d'analyse a insisté sur la nécessité d'étendre cette maturité et cette structuration autour de la donnée de référence (MDM) vers la donnée semi ou non structurée. C'est cela qui va faire le lien entre MDM et Big Data.

MDM et Big Data sont tout à fait complémentaires. C'est un cas d'usage qu'on retrouve dans la littérature, mais qui en réalité n'est pas encore véritablement mis en œuvre en France. Si le Big Data est synonyme de foisonnement et de chaos (souvent créateur), le MDM est ce qui permet de mettre de l'ordre dans tout cela, permettant de rendre structuré quelque chose qui ne l'est pas, de le classer et de l'organiser.

Exemple d'illustration

Comment réaliser et livrer la vision 360° du client, le but ultime de tout marketeur ? Pour cela on a besoin de plusieurs types de données de référence, géré par le MDM : cela va être le client (numéro de sécurité sociale, Siret/Siren si c'est du B2B, adresse e-mail, autres données de contact, etc.). Ensuite, il y a l'historique du client, sa segmentation : c'est la Business Intelligence (BI). C'est-à-dire qu'on va tisser des liens entre la donnée de référence et la donnée relationnelle, la donnée de fait, celle de ses achats.

Le Big Data vient donc enrichir ce capital initial avec des données comportementales, les accointances du consommateur avec différentes marques et ses réseaux en général, notamment à travers les divers réseaux et médias sociaux. On a donc besoin de ces différents types de données pour mettre en place une vision métier.

Le travail de Big Data ne peut pas se concevoir en dehors de cette réalité, même si les sources (Facebook, Google, Amazon etc.) ont une existence propre. L'intelligence est dans les liens entre les données, c'est ce qui va transformer multiplier la valeur de la donnée.

Exemple d'illustration de la distribution

La tendance actuelle n'est plus au multicanal, mais à l'omnicanal. Pour mettre cela en place, cela implique une bonne connaissance du client, pour pouvoir lui proposer de la personnalisation, du MDM, pour bien le connaître, de la BI et enfin des réseaux sociaux.

Si le distributeur connaît bien ses produits, qu'il fournit l'information pertinente quel que soit le canal, et qu'il connaît bien ses stocks (via le BI et le lien notamment vers les ERP), il est capable de personnaliser l'offre son client, en tirant des liens entre ces différents niveaux de données..

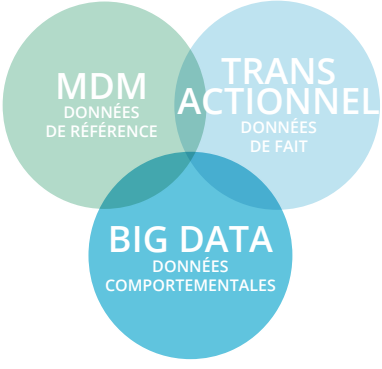


Figure 12. *3 types de données et la fusion des 3 domaines*

CHAPITRE 06
Big Data ou Big Brother ?

Les Big Data ne laissent personne indifférent, soit pour évoquer en elles le futur de l'informatique et un nouvel ordre économique, soit pour les fustiger et les pointer du doigt comme dérive orwellienne d'une société scientiste devenue incontrôlable.

Il ne s'agit d'ailleurs pas que d'un débat d'idée, car le jugement rendu par l'Union Européenne à l'encontre du géant Google en juin 2014 est symbolique d'un malaise profond dans une société où une minorité d'utilisateurs très agissante, lance un débat démocratique qu'il n'est pas question ici d'éluder, ni même de remettre en cause. L'avènement des Big Data est bel et bien une nouvelle donne économique, même si elle ne vient pas bousculer tout sur son passage non plus. Et cette nouveauté ne vient que renforcer les devoirs moraux et d'éthique des entreprises, dont l'engagement citoyen devient aujourd'hui une obligation à défaut d'être un choix.

En quelque sorte, les débats qui surgissent aujourd'hui autour des Big Data sont très similaires à ceux qui ont eu lieu vers les débuts du Web au début des années 1990.

Côté pile : l'espoir d'un secteur dynamisant qui vient irriguer toute l'économie

Lors d'un récent matin de l'économie organisé par le Journal du Dimanche et dont Business & Decision était co-sponsor, Philippe Oddo de la Banque éponyme a déclaré en guise d'introduction : "le Big data c'est surtout la collecte et le traitement de l'information pour anticiper au mieux ce qui va se passer dans tous les secteurs et particulièrement dans celui de l'analyse financière". Avec une pareille entrée en matière, il n'y a pas de doute possible ... les Big Data sont l'avenir de la high tech mais aussi de beaucoup de secteurs plus traditionnels. Le fait que l'assertion vienne d'un homme du métier est un signal fort.

Côté face : le débat sur la vie privée

Ce sujet est véritablement au cœur de l'économie (notamment dans la « Presse qui possède énormément

de données » selon Denis Olivennes), mais elle est au cœur des débats sur la « vie privée ». Et d'ajouter que dans le cadre de l'Open Internet Project[31], une présentation effrayante a été faire par Laurent Alexandre (chirurgien fondateur de Doctissimo) sur le projet totalitaire des acteurs dominants des technologies de l'information, avec un coup de projecteur assez fort sur Google.

Un air de déjà vu

Cette approche par les deux extrêmes des technologies n'est pas nouvelle et réapparaît à chaque innovation, dans la High Tech et même au-delà. Le premier exemple connu étant, au Royaume Uni, celui des Luddites du 19e siècle, au moment de l'introduction des métiers à tisser.

Tout nouveau saut technologique, aussi petit soit-il, génère avec lui son lot de techno-scientisme et, en même temps et à l'opposé, une opposition acharnée et irrationnelle la plupart du temps.

Il est hors de question cependant, de tenter de faire abstraction des questions d'éthique, bien au contraire. C'est même l'entreprise qui doit prendre les devants en matière d'éthique pour lever les ambiguïtés qui pourraient se faire jour dans l'esprit de leurs utilisateurs et encore plus de leurs détracteurs.

Les Big Data peuvent en elles-mêmes porter la réponse au problème de la confidentialité

Si le problème de la confidentialité des données privées n'est pas né avec les Big Data, force est de constater par contre que celles-ci l'ont exacerbé.

Mais si d'aucuns peuvent en effet penser que le danger des Big Data peut être induit par la technologie, cette dernière, bien utilisée, peut aussi apporter une réponse à ce problème de confidentialité.

Le Réseau Tor, par exemple, « propose un navigateur qui permet de brouiller les données » ; aujourd'hui cela est assimilable à un travail de pirate a-t-il précisé, « mais demain il y aura aussi un marché pour cela ».

La régulation est peut être dépassée en bien par la technologie et les entrepreneurs. Peut-être que demain on paiera pour protéger ses données.

Le salut n'est peut-être pas en effet à venir de l'Etat, incapable de réglementer des outils avec un attirail juridique trop lent et trop rigide pour s'adapter à la mouvance des technologies de l'information.

The Open Internet Project (Site et association manifeste anti Google)

[31] *Article du JDD paru le 8 février 2014 : http://www.lejdd.fr/Economie/Entreprises/Laurent-Alexandre-La-strategie-secrete-de-Google-apparait-652106*

Il n'y a pas d'équivalence entre espionnage et Big Data

Enfin et surtout, dans l'esprit du grand public et dans les critiques, souvent hâtives, faites des Big Data, il y a ce raccourci abusif entre Big Data, Big Brother et les grands réseaux sociaux… sans parler de la NSA[32].

Il convient de différencier l'usage des Big Data faits par ces acteurs et les objectifs des entreprises qui veulent utiliser les Big Data pour optimiser leurs ventes et leurs résultats dans le respect de leurs clients. C'est ce dernier segment qui est l'immense majorité des utilisateurs des Big Data et c'est bien à celui-ci que nous nous intéressons ici et dans nos travaux quotidiens.

Par ailleurs, nous sommes confrontés à un double paradoxe aujourd'hui : d'une part la volonté de l'utilisateur qui désire des informations pertinentes et personnalisées mais en même temps, qui ne veut pas être surveillé. La preuve en est que les consommateurs, depuis déjà quelque temps, considèrent comme spam tous les messages qui lui déplaisent.

D'autre part, on observe que les utilisateurs ont aussi le désir de préserver leur vie privée et en même temps de la dévoiler largement sur les réseaux sociaux. C'est même ce phénomène là qui a amené Mark Zuckerberg à s'exprimer sur la fin de la vie privée, sans doute un peu hâtivement[33].

Du point de vue de l'annonceur, ce débat n'est pas neutre non plus et doit forcer les entreprises à prendre position en matière d'éthique de manière évidente et si possible anticipée. Avant tout, il y a une barrière nette à dresser entre exploitation anonyme et statistique des données (même à titre de recommandation personnalisée) et espionnage de la vie privée. En fin de compte, la différence entre les deux n'est pas technique mais humaine et éthique, et la réponse à ce problème se doit elle aussi d'être humaine et éthique[34].

Il n'y a pas, en matière de Big Data, de chemin médian : soit l'annonceur pratique les Big Data de manière éthique, soit il pratique l'intrusion et le spam et il outrepasse ses droits ; avec les conséquences qu'on imagine sur sa réputation si la démarche en vient à être dévoilé et publiquement critiquée.

Mais tout compte fait, et une fois ce problème d'éthique résolu sans ambiguïté par les annonceurs, il est clair que l'amélioration de l'expérience client passe par une connaissance client plus intime ; en fin de compte, toutes les entreprises se mettront un jour au Big Data.

---

[32] *Ces débats sont déjà largement traités dans la sphère publique et sont avant tout affaire d'opinion. Il ne convient donc pas que nous les traitions ici. A savoir également, que l'union européenne a déjà pris des dispositions contre les géants du Web et notamment Google dans le cadre de ses actions pour le droit à l'oubli. Celui-ci déchaînant également les passions, dans les deux sens. Ces sujets sont évoqués ici mais n'y seront pas abordés dans les détails.*

CHAPITRE 07
Comment passer des Big Data au Big Busine$$

Que retenir des Big Data ?

En conclusion, que faut-il retenir des Big Data ? Certainement qu'avant toute chose, il convient de comprendre ce que c'est, qu'il ne faut pas conclure hâtivement et péremptoirement que rien n'a changé, ni au contraire que rien n'est plus comme avant. La réalité est tout autre.

Avant toute chose, il faut définitivement se départir de certains mythes : tout vouloir analyser en Big Data est purement utopique. De la même façon, l'idée qu'il faut tout stocker dans le but « d'en faire quelque chose un jour » est tout autant fantaisiste.

D'abord, stocker des données inutiles coûte cher, en Big Data encore plus que dans des projets plus traditionnels de Business Intelligence. Mais surtout, le stockage doit être pensé dès le départ, en vue des traitements à effectuer. Comme nous l'avons vu précédemment, il existe des traitements statistiques impossibles à réaliser si les données sont stockées dans certains formats décentralisés. L'utilisation de ces formats interdit donc certains types d'exploitation.

Sans évoquer les cas extrêmes, certains traitements Big Data prennent du temps, un temps qui augmente souvent exponentiellement avec le volume des données.

Pour obtenir des temps de réponse acceptables, on devra donc souvent faire un compromis entre la quantité de données prises en compte et le temps dont on dispose pour le traitement. Le fait de maintenir certains calculs intermédiaires à jour dans des entrepôts de données orientés « métier » peut aussi considérablement aider à écarter ce problème et à réduire les temps de calcul à des niveaux acceptables.

Si on est capable d'intégrer tous ces points il sera possible de passer du Big Data au « Big Business ».

[33] *Cf. http://bit.ly/privateZuck pour son intervention sur le sujet fin 2009. La phrase avait fait beaucoup réagir sur les médias sociaux comme ailleurs, forçant Facebook à donner des preuves de respect des données privées à plusieurs reprises ... sans bien convaincre personne au final.*

[34] *Cf. la charte d'éthique du groupe Orange, clairement orientée en ce sens : http://oran.ge/SMYku4*

Les 10 points clés

Alors quels sont les 10 points à retenir si vous voulez réussir vos projets de Big Data ?

1. L'objectif de départ est important

Si votre objectif n'est pas clair, vous risquez non seulement de vous tromper d'outils mais également de pénaliser votre projet en termes de temps passé et de ressources consommées. Concentrez-vous sur la maturation du cas d'usage et l'identification des données plutôt que d'investir dans une infrastructure technique. Utilisez des plateformes prêtes à l'emploi externalisées en attendant.

2. La notion d'incertitude

Une des évolutions les plus significatives des Big Data par rapport au travail plus traditionnel sur les données est le management de l'incertitude. Ceci ne veut pas dire que rien n'est planifié ni que les projets de Big Data se lancent sans préparation, bien au contraire. Cela veut dire cependant – et c'est particulièrement vrai en marketing – que le projet de Big Data doit prendre en compte cette incertitude dès sa conception, et fonctionner sur un modèle itératif avec possiblement de l'auto apprentissage.

3. Le Big Data est multi compétences

Les Big Data ne sont pas affaire de robots. Elles sont avant tout le résultat du croisement de l'automatisation et de la technologie et de l'intelligence humaine. Pour qu'elles fonctionnent bien et fournissent des résultats à la hauteur des espérances, elles nécessitent de nouveaux profils, au croisement de différentes disciplines : informatique, bases de données, statistiques, intelligence artificielle et enfin et surtout, les connaissances métier (marketing, finance, logistique, etc.).

4. C'est un phénomène majeur, aussi important que le CRM en son temps

La quasi totalité des détracteurs des Big Data vous diront tout de go qu'il ne s'agit pas de quelque chose de nouveau, mais plutôt d'un épiphénomène, d'une mode passagère qui disparaîtra comme elle est venue. C'est mal comprendre l'histoire et le marketing des technologies, car les Big Data trouvent leurs racines dans de nombreuses années d'efforts et de tâtonnements (marketing, financiers, managériaux) qui finissent par donner leurs fruits aujourd'hui du fait du développement des technologies et de leurs usages (infrastructures, logiciels, généralisation des usages, prédominance du Web sociale et omniprésence des réseaux). Tout ceci fait que les questions fondamentales des métiers trouvent enfin réponse après 15 à 20 ans de travaux et d'essais-erreurs. C'est un signe qui ne trompe pas, celui de l'arrivée à maturité et de la convergence des moyens qui rendront possibles de nouvelles avancées spectaculaires.

5. L'impact des Big Data sur les organisations est significatif

Non seulement car de nouveaux métiers sont apparus, pour lesquels, les formations sont encore largement à créer. Mais également car les organisations « métier » des entreprises sont fortement refondées : savoirs, approches, décisions et méthodes sont profondément bouleversés. Nous ne ferons plus de marketing, pour prendre cet exemple, comme nous en faisons aujourd'hui, même si ce changement va prendre quelques années. Si vous faites partie d'une organisation métier, commencez-donc dès à présent à vous intéresser à la donnée car votre métier va être bouleversé.

6. On peut faire des Big Data dès maintenant car les technologies sont disponibles

Les Big Data ne sont pas non plus une prévision, elles sont déjà disponibles, ici et maintenant, même si leur paysage évolue à grande vitesse. Bon nombre des technologies utilisées dans le cadre des Big Data ont en effet été inventées et popularisées par les géants du Web (Google et Yahoo! font partie des pionniers) et sont désormais mises à disposition de tous ceux qui sont capables de les mettre en œuvre.

7. La donnée est probablement la matière première la moins connue et la moins comprise

La distinction entre système d'information (l'ensemble des processus et des organisations entre les données, leur naissance, leur vie, leur traitement et leur archivage) et système informatique (la mécanique matérielle et surtout logicielle qui permet de faire tourner l'ensemble et de traiter la donnée) est un grand classique. La donnée est encore aujourd'hui un espace largement méconnu des responsables « métier » qui considèrent encore les systèmes informatiques comme des formules magiques capables de transformer le business sans effort.

Or, la donnée est capricieuse, et elle requiert beaucoup de travail. Son importance croissante dans une société où l'informatisation est omniprésente, dans tous les secteurs, force à changer la perception de cette donnée par l'utilisateur. Beaucoup reste encore à faire pour que ce changement soit totalement abouti.

8. Un projet Big Data se gère différemment

Les Big Data ne sont pas une mode ni le simple changement de nom du datamining. Elles ont leur vocabulaire, leurs professionnels, leurs méthodes, leurs algorithmes, et leurs approches projets spécifiques.

Nous l'avons évoqué dans ce livre blanc, un projet Big Data a ses spécificités. Au delà de l'approche technique, il induit des méthodologies particulières, un cadre juridique adapté et un bonne mesure des impacts sociaux.

Un apprentissage sera donc nécessaire ; il ne s'agit d'ailleurs pas d'un savoir fini, les Big Data sont en reconfiguration constante.

9. Les Big Data sont incontournables, et toutes les entreprises en feront, comme l'Internet

Les véritables innovations suivent toujours un schéma d'adoption à peu près similaire, fort bien décrit par Geoffrey Moore dans « Crossing the Chasm », son best seller des années 90. Les Big Data en sont à un point d'inflexion qui fait que leur adoption se généralise et sort du cercle fermé des géants du Web et des médias sociaux qui les ont inventées. Il est désormais possible d'appliquer ces techniques et ces approches aux entreprises faisant partie des secteurs plus traditionnels. Nous n'en sommes qu'au tout début.

10. Les Big Data ne se limitent pas au temps reel

Même si Hadoop est une grande innovation, les Big Data ne se limitent pas à Hadoop et encore moins au temps réel. Certains de ces usages sont en effet adaptés à de grands volumes de données et nécessitent, selon les cas, des traitements déportés. C'est la bonne combinaison des différentes approches et techniques qui fera la qualité et le résultat d'un projet de Big Data.

C'est en intégrant ces différents points qui font les spécificités des vraies Big Data, que les entreprises de demain pourront passer des Big Data au Big Business.

Ont contribué à ce livre blanc :
Patrick Bensabat, Didier Gaultier, Michael Hoarau, Bruno Laug et Yann Gourvenec

Pour en savoir plus, nous vous conseillons de vous rendre sur notre blog expert des Big Data :

blog.businessdecision.com