

Comment Le Big Data peut-il rendre une entreprise plus compétitive ?

EVOLUTIONS TECHNOLOGIQUES ET STRATEGIQUES DU
SYSTEME D'INFORMATION

ANTOINE PUYENCHET – JÉRÉMY RAMBAUD – THOMAS ROLLET – HENRI
PINEAU

Table des matières

Table des matières.....	1
Table des illustrations	2
1. Introduction	3
2. BigData – Histoire.....	4
2.1 Les prémisses du Big Data	4
2.2 L'émergence du Big Data.....	5
2.3 L'explosion du Big Data.....	6
3. Le BigData, une histoire de DataScience	8
3.1 DataScience.....	8
3.2 Volume, variété, vélocité.....	9
3.2.1 Volume.....	9
3.2.2 Vitesse.....	9
3.2.3 Variété	10
3.3 Datamining	10
3.3.1 Apprentissage supervisé	10
3.3.2 Arbres de décision	10
3.3.3 Réseau de neurones	11
3.3.4 Apprentissage non supervisé	11
3.3.5 Apprentissage incrémental	12
4. Les étapes du Big Data	13
4.1.1 Recueillir les données	13
4.1.2 Analyser les données	13
5. Enjeux du BigData pour le SI.....	14
5.1 Les enjeux technologiques	14
5.1.1 L'investissement en capacité de stockage.....	14
5.1.2 L'investissement dans l'analyse.....	15
5.2 Les enjeux économiques	15
5.2.1 L'apport du Big Data pour le Marketing.....	15
5.2.2 L'apport du Big Data pour les ressources humaines	15
5.2.3 L'apport du Big Data pour les finances	16
5.2.4 L'apport du Big Data pour la logistique.....	16
5.3 L'impact organisationnel.....	16
5.3.1 La conduite du changement.....	16
5.3.2 L'apparition de nouveaux métiers.....	17
5.4 Synthèse.....	19
6. Enjeux de compétitivité	20
6.1 Améliorer la gestion de la production	20
6.2 Répondre aux normes de traçabilité et de qualité.....	20
6.3 Fidéliser les consommateurs grâce à une meilleure gestion commerciale.....	20

6.4	Identifier les besoins pour mieux innover	21
7.	Cas d'usages pour l'industrie et la ville de demain	22
7.1	Secteur de la distribution.....	22
7.2	Secteur de l'hôtellerie.....	22
7.3	Secteur de l'énergie	22
7.4	Secteur des assurances.....	23
8.	Conclusion.....	24
9.	Bibliographie	25

Table des illustrations

Figure 1 - Croissance de la capacité mondiale de stockage de données et informations	4
Figure 2 - Logo Hadoop.....	6
Figure 3 - Tendance des recherches sur la grippe (2007 - 2008).....	8
Figure 4 - Illustration 3V	9
Figure 5 - Structure d'un neurone artificiel.....	11
Figure 6 - Enjeux du Big Data pour les entreprises	14

1. Introduction

Le Big Data, littéralement « grosses données », parfois appelées données massives, désignent des ensembles de données qui deviennent tellement volumineux qu'ils en deviennent difficiles à travailler avec des outils classiques de gestion de base de données ou de gestion de l'information.

L'explosion quantitative (et souvent redondante) de la donnée numérique constraint à de nouvelles manières de voir et d'analyser le monde. De nouveaux ordres de grandeur concernent la capture, le stockage, la recherche, le partage, l'analyse et la visualisation des données. Les perspectives du traitement des Big Data sont énormes et en partie encore insoupçonnées ; on évoque souvent de nouvelles possibilités d'exploration de l'information diffusée par les médias, de connaissance et d'évaluation, d'analyse tendancielle et prospective (climatiques, environnementales ou encore sociopolitiques, etc.), de gestion des risques (commerciaux, assuranciels, industriels, naturels) et de phénomènes religieux, culturels, politiques, mais aussi en termes de génomique ou méta-génomique, pour la médecine (compréhension du fonctionnement du cerveau, épidémiologie, éco-épidémiologie...), la météorologie et l'adaptation aux changements climatiques, la gestion de réseaux énergétiques complexes (via les smartgrids ou un futur « internet de l'énergie »), l'écologie (fonctionnement et dysfonctionnement des réseaux écologiques, des réseaux trophiques avec le GBIF par exemple), ou encore la sécurité et la lutte contre la criminalité. La multiplicité de ces applications laisse d'ailleurs déjà poindre un véritable écosystème économique impliquant, d'ores et déjà, les plus gros joueurs du secteur des technologies de l'information.

Certains supposent que le Big Data pourrait aider les entreprises à réduire leurs risques et faciliter la prise de décision, ou créer la différence grâce à l'analyse prédictive et une « expérience client » plus personnalisée et contextualisée. C'est ce que nous allons découvrir dans ce document.

Dans ce livre blanc, nous analyserons dans un premier temps ce qu'est le BigData ainsi que son Histoire. Nous étudierons ensuite les différentes notions liées à l'analyse de données et à la DataScience. Nous enchaînerons avec les 2 étapes qui constituent la mise en place de solutions de BigData. Nous terminerons par les enjeux du BigData pour le SI ainsi que les enjeux de compétitivité et les cas d'usages pour l'industrie et la ville de demain. Enfin nous conclurons.

2. BigData – Histoire

2.1 Les prémisses du Big Data

Le BigData part d'un constat réalisé dès les années 1960 par les théoriciens : celui de l'insuffisance des espaces de stockage. Au début des années 1960, Derek Price a constaté qu'il était impossible de suivre le rythme de la recherche scientifique. Les résumés de journaux ont été créés à la fin du XIXe siècle afin de gérer la croissance des bases de connaissances. Ils enregistraient déjà une tendance similaire (**multiplication par 10 tous les 50 ans**) et avaient atteint une « ampleur critique ». Ce n'était donc plus une solution viable pour le stockage ou l'organisation des informations.

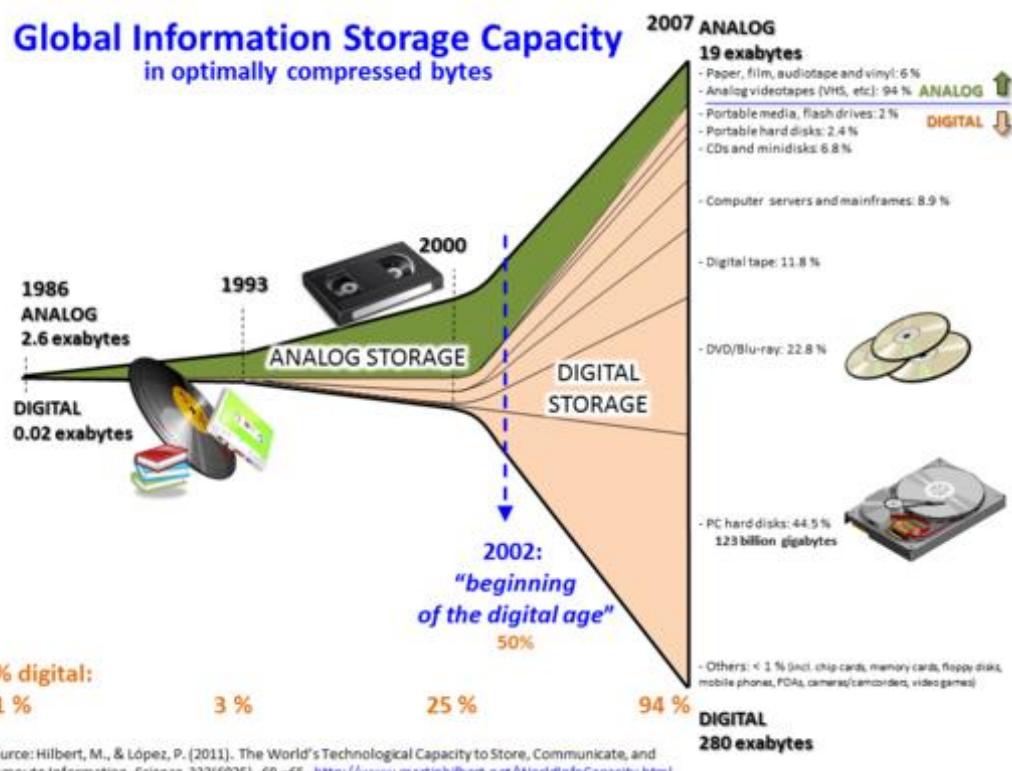


Figure 1 - Croissance de la capacité mondiale de stockage de données et informations

En 1970, Edgar F. Codd, mathématicien diplômé d'Oxford et travaillant dans le laboratoire de recherche d'IBM, a publié un article expliquant comment accéder à des informations stockées dans de grandes bases de données sans connaître la structure ni l'emplacement de ces informations. Jusqu'alors, la récupération des informations nécessitait des connaissances informatiques assez pointues, voire le recours aux services de spécialistes. Cette opération était à la fois chronophage et onéreuse. Aujourd'hui, la plupart des transactions de données de routine (accès aux comptes bancaires, utilisation de cartes de crédit, transactions boursières, réservations de voyage, achats en ligne) emploient des structures reposant sur la théorie de la base de données relationnelle.

Alors que les informations commençaient à se développer plus rapidement, les possibilités de condenser le stockage et l'organisation des données se sont amoindries. Dans son discours intitulé « *Where Do We Go From Here ?* », I.A. Tjomsland a déclaré : « *Ceux qui connaissent les périphériques de stockage se sont rendu compte depuis longtemps que la première loi de Parkinson pouvait s'appliquer à notre secteur : "Les données s'étendent jusqu'à remplir l'espace disponible pour leur stockage". Je pense que de grandes quantités de données sont conservées, car les utilisateurs n'ont aucun moyen d'identifier les données obsolètes. Les inconvénients du stockage des données obsolètes sont moins visibles que les inconvénients de la suppression de données potentiellement utiles.* »

En 1989, Howard Dresner a développé la notion de « Business Intelligence (BI) », terme générique populaire inventé par Hans Peter Luhn en 1958. Selon Howard Dresner, la Business Intelligence (ou l'analyse décisionnelle) désigne « *des concepts et méthodes permettant d'améliorer la prise de décision métier grâce à des systèmes reposant sur des faits* ». Peu après, en réponse au besoin d'une meilleure BI, des entreprises telles que Business Objects, Actuate, Crystal Reports et MicroStrategy ont vu le jour et commencées à proposer des rapports et des analyses de données d'entreprise.

2.2 L'émergence du Big Data

L'afflux d'informations a engendré un nouveau défi en matière de gestion des données, ainsi qu'une augmentation des coûts de publication et de stockage. La gestion des données s'est également avérée plus complexe. Offrant davantage de fonctionnalités, le stockage des données au format numérique est rapidement devenu plus économique que le papier, et des plateformes BI ont commencé à éclore. R.J.T. Morris et B.J. Truskowski ont étudié le stockage des données dans leur article intitulé *The Evolution of Storage Systems* et publié dans l'IBM Systems Journal.

Le terme « Big Data » a fait son apparition dans un article publié par Michael Cox et David Ellsworth, chercheurs à la NASA. Tous deux affirmaient que l'augmentation du volume des données devenait problématique pour les systèmes informatiques de l'époque. C'est ce que l'on a appelé le « *problème du Big Data* ».

Dans son article intitulé « *How much information is there in the world ?* », Michael Lesk déclare « *on compte peut-être quelques milliers de pétaoctets d'informations en tout et pour tout, et la production de bandes et de disques aura atteint ce niveau en l'an 2000. Donc dans quelques années, (a) nous serons en mesure de tout enregistrer (sans suppression d'informations) ; et (b) la plupart des informations ne seront jamais examinées par un être humain* ».

Peter Lyman et Hal R. Varian de l'UC Berkeley ont publié la première étude qui quantifiait, en termes de stockage informatique, le volume total d'informations initiales et nouvelles créées

chaque année dans le monde. Cette étude, intitulée « *How Much Information ?* », a été établie en 1999, année durant laquelle le monde a produit environ 1,5 exaoctet d'informations.

Au cours des années 1990, les fournisseurs d'ERP ont complété les modules de base avec d'autres modules et fonctions, ce qui a donné naissance à des « systèmes ERP étendus ». Les options logicielles et matérielles se sont multipliées. Puis au début des années 2000, des acteurs majeurs du secteur des logiciels ont commencé à fusionner. Oracle et SAP ont été les seules grandes sociétés éditrices d'ERP à survivre à ce grand bouleversement.

2001 voit l'apparition de l'acronyme SaaS dans un article de la division eBusiness de la Software & Information Industry Association (SIIA). La même année, Doug Laney, analyste chez Gartner, a publié un rapport de recherche intitulé *3D Data Management: Controlling Data Volume, Velocity, and Variety*. Encore aujourd'hui, les 3 V constituent les critères globaux du Big Data.

Tim O'Reilly a publié en Septembre 2005 l'article intitulé « *What is Web 2.0?* », dans lequel il affirme que « *les données sont le prochain Intel Inside* ». Toujours selon Tim O'Reilly : « *Comme Hal Varian l'a expliqué au cours d'une conversation personnelle l'an dernier, "SQL est le nouvel HTML". La gestion des bases de données est une compétence essentielle pour les entreprises Web 2.0, tant et si bien que ces applications sont parfois appelées "infoware" et non plus simplement des "logiciels"* ». (Crawford, s.d.) (Michael Cox, s.d.)

2.3 L'explosion du Big Data

Hadoop a été créé en 2006 pour que de nouveaux systèmes sachent gérer l'explosion des données Web. Pouvant être téléchargé, utilisé et amélioré gratuitement, Hadoop constitue un moyen entièrement ouvert de stockage et de traitement des données qui « permet de traiter de manière parallèle et distribuée de grands volumes de données sur des serveurs standard et économiques, capables à la fois de stocker et de traiter les données, et d'évoluer à l'infini ».



Figure 2 - Logo Hadoop

Des chercheurs d'International Data Corporation ont publié un article intitulé *The Expanding Digital Universe: A Forecast of Worldwide Information Growth through 2010*, qui évalue et prévoit le volume de données numériques qui seront créées et reproduites chaque année. Cet article estime que pour la seule année 2006, 161 exaoctets de données ont été créés dans le monde entier, et il prévoit qu'au cours des quatre prochaines années, ce nombre sera multiplié par plus de six (pour atteindre 988 exaoctets). En d'autres termes, ils estiment que le volume d'informations doublerait tous les 18 mois sur ces 4 prochaines années. D'après les rapports de suivi émis en 2010 et 2012, le volume de données numériques créées chaque année a déjà dépassé les prévisions initiales (1 227 exaoctets en 2010 et 2 837 exaoctets de plus en 2012).

L'étude « *How Much Information ? 2009 Report on American Consumers* » du Global Information Industry Center révèle qu'en 2008, « les Américains ont consommé environ 1,3 billion d'heures d'informations, soit en moyenne près de 12 heures par jour. Cette consommation a atteint un total de 3,6 zettaoctets et de 10 845 billions de mots, ce qui correspond à 100 500 mots et 34 gigaoctets en moyenne par jour et par personne ». Un rapport de janvier 2011, intitulé « *How Much Information ? 2010 Report on Enterprise Server Information* », a ensuite estimé qu'en 2008, « les serveurs du monde entier ont traité 9,57 zettaoctets d'informations, soit pratiquement 1022 gigaoctets (soit dix millions de millions). Cela représentait 12 gigaoctets d'informations quotidiennes pour un employé moyen, ou environ 3 téraoctets d'informations par employé et par an. Les entreprises mondiales ont traité en moyenne 63 téraoctets d'informations par an ».

Un article intitulé The World's Technological Capacity to Store, Communicate, and Compute Information et tiré du magazine Science a estimé que la capacité mondiale de stockage des informations avait augmenté de 25 % par an entre 1987 et 2007. Toujours d'après cette source, 99,2 % du stockage des données était analogique en 1986, alors qu'il était à 94 % numérique en 2007. On a donc assisté à un bouleversement complet en à peine 20 ans (en 2002, le stockage numérique a surpassé le stockage non numérique pour la première fois).

Une étude de 2010 a constaté que les déploiements SaaS avaient doublé pour atteindre 15 % en 2009 (contre 7 % en 2008). En 2011, les principales tendances de la Business Intelligence sont bien identifiées, avec en tête de file le cloud computing, la visualisation des données, l'analyse prédictive et le Big Data.

Des scientifiques du McKinsey Global Institute ont publié un article intitulé Big Data: The Next Frontier for Innovation, Competition, and Productivity, dans lequel ils estiment qu'« en 2009, presque tous les secteurs de l'économie des États-Unis possédaient au moins 200 téraoctets de données stockées en moyenne (soit le double de l'entrepôt de données du détaillant américain Wal-Mart en 1999) par entreprise comptant plus de 1 000 employés ». Toujours d'après cette source, les secteurs des valeurs mobilières et des placements étaient les premiers en termes de données stockées par entreprise. Les scientifiques ont calculé que 7,4 exaoctets de données d'origine avaient été sauvegardés par les entreprises, tandis que 6,8 exaoctets de données avaient été sauvegardées par les consommateurs pour la seule année 2010. (Olson, s.d.)

3. Le BigData, une histoire de DataScience

3.1 DataScience

Le web est rempli d'applications dirigées par les données « data-driven apps ». Quasiment tous les sites de e-commerce sont des applications data-driven. Il y a une base de données derrière une interface web et un middleware qui s'occupe de dialoguer avec d'autres bases de données et services (services de paiement par carte, banques, etc...). Mais utiliser ces données n'est pas ce qui définit la science des données. Une application data-driven acquiert sa valeur depuis la donnée elle-même, et crée encore plus de données en résultat.

Un des tous premiers produits de données sur le web fut la base de données CDDB¹ (Compact Disc Database). Les développeurs de cette base se sont rendus compte que chaque CD possédait un identifiant unique basé sur la longueur exacte de chaque piste du CD. Gracenote² a construit une base de données avec ces longueurs de pistes et l'a couplé avec une base de données de metadatas (titres de pistes, artistes, nom d'albums). Toute personne ayant déjà utilisé iTunes pour copier un CD sur un ordinateur s'est servi sans le savoir de cette base de données. Avant de faire quoi que ce soit, iTunes regarde la longueur de chaque piste, l'envoie à la base de données CDDB et récupère le nom de la piste. Cela semble simple, et pourtant c'est un concept révolutionnaire : CDDB voit la musique comme une donnée, par comme un son. Leurs business sont totalement différents de la vente de musique, le partage ou encore l'analyse des goûts musicaux. CDDB est né en convertissant un problème musical en un problème de données.

Google n'est pas en reste dans ce domaine. Par exemple pendant l'épidémie de grippe A (H1N1) de 2009, Google a été capable de suivre la progression de l'épidémie grâce au suivi des recherches liées à la grippe.

Google a été capable de détecter une tendance des recherches à la hausse sur la grippe aviaire environ deux semaines avant les centres de contrôle et de prévention des maladies en analysant les recherches effectuées par la population.

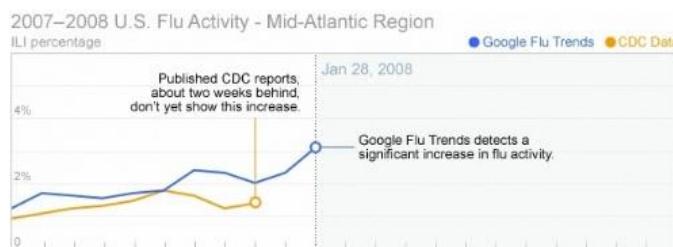


Figure 3 - Tendance des recherches sur la grippe (2007 - 2008)

¹ <https://www.wikiwand.com/en/CDDB>

² <https://www.wikiwand.com/en/Gracenote>

Mais Google n'est pas la seule entreprise qui utilise les données. Facebook et LinkedIn utilisent des algorithmes basés sur les relations pour suggérer à ses utilisateurs de nouvelles relations que ces derniers peuvent connaître. Amazon enregistre vos recherches et les met en corrélation avec celles des autres utilisateurs afin de vous suggérer des produits susceptibles de vous intéresser.

Le point commun de toutes ces applications est que les données collectées depuis les utilisateurs ajoutent de la valeur. C'est le début de la science des données.

Ce domaine qu'est la DataScience est en fait un regroupement de plusieurs domaines portant sur l'utilisation et la manipulation de la donnée. Ce domaine couvre :

- le machine Learning ;
- le DataMining ;
- le Big Data.

(Loukides, 2011) (La science des données : pour qui ? Pourquoi ?)

3.2 Volume, variété, vélocité

3.2.1 Volume

Le volume décrit la quantité de données générées par des entreprises ou des personnes. Le Big Data est généralement associé à cette caractéristique. Les entreprises, tous secteurs d'activité confondus, devront trouver des moyens pour gérer le volume de données en constante augmentation qui est créé quotidiennement. Les catalogues de plus de 10 millions de produits sont devenus la règle plutôt que l'exception. Certains clients gérant non seulement des produits mais aussi leur propre clientèle, peuvent aisément accumuler un volume dépassant le téraoctet de données.

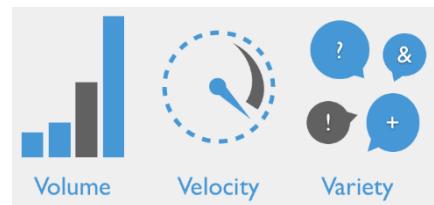


Figure 4 - Illustration 3V

3.2.2 Vitesse

La vitesse décrit la fréquence à laquelle les données sont générées, capturées et partagées. Du fait des évolutions technologiques récentes, les consommateurs mais aussi les entreprises génèrent plus de données dans des temps beaucoup plus courts.

À ce niveau de vitesse, les entreprises ne peuvent capitaliser sur ces données que si elles sont collectées et partagées en temps réel. C'est précisément à ce stade que de nombreux systèmes d'analyse, de CRM, de personnalisation, de point de vente ou autres, échouent.

Ils peuvent seulement traiter les données par lots toutes les quelques heures, dans le meilleur des cas. Or, ces données n'ont alors déjà plus aucune valeur puisque le cycle de génération de nouvelles données a déjà commencé.

3.2.3 Variété

La prolifération de types de données provenant de sources comme les médias sociaux, les interactions Machine to Machine et les terminaux mobiles, crée une très grande diversité au-delà des données transactionnelles traditionnelles. Les données ne s'inscrivent plus dans des structures nettes, faciles à consommer.

Les nouveaux types de données incluent contenus, données géo spatiales, points de données matériels, données de géolocalisation, données de connexion, données générées par des machines, données de mesures, données mobiles, points de données physiques, processus, données RFID, données issues de recherches, données de confiance, données de flux, données issues des médias sociaux, données texte et données issues du Web.

Nos propres objets métiers rapides (inventés il y a 8 ans) préfiguraient cette tendance en permettant aux entreprises d'introduire rapidement de nouveaux objets de données ou de doter les objets existants de nouvelles caractéristiques.

3.3 Datamining

3.3.1 Apprentissage supervisé

En sciences cognitives, l'apprentissage supervisé est une technique d'apprentissage automatique (plus connu sous le terme anglais de machinelearning) qui permet à une machine d'apprendre à réaliser des tâches à partir d'une base d'apprentissage contenant des exemples déjà traités. Chaque élément (item) de l'ensemble d'apprentissage (training set) étant un couple entrée-sortie. De par sa nature, l'apprentissage supervisé concerne essentiellement les méthodes de classification de données (on connaît l'entrée et l'on veut déterminer la sortie) et de régression (on connaît la sortie et l'on veut retrouver l'entrée).

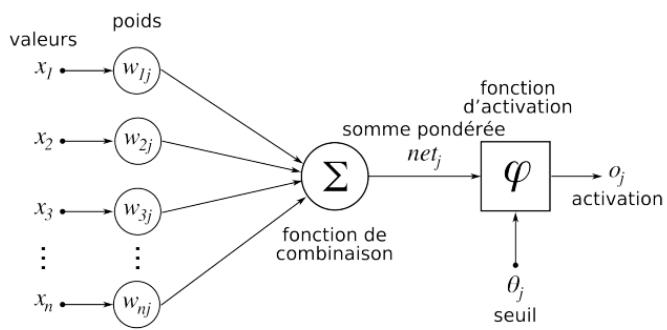
3.3.2 Arbres de décision

Un arbre de décision est, comme son nom le suggère, un outil d'aide à la décision qui permet de répartir une population d'individus en groupes homogènes selon des attributs discriminants en fonction d'un objectif fixés et connu. Il permet d'émettre des Extract, Transform and Load prédictions à partir des données connues sur le problème par réduction, niveau par niveau, du domaine des solutions. Chaque nœud interne d'un arbre de décision porte sur un attribut discriminant des éléments à classifier qui permet de répartir ces éléments de façon homogène entre les différents fils de ce nœud. Les branches liant un nœud à ses fils représentent les valeurs discriminantes de l'attribut du nœud. Et enfin, les feuilles d'un arbre de décision sont ses prédictions concernant les données à classifier. C'est une méthode qui a l'avantage d'être lisible pour les analystes et permet de déterminer les couples discriminants à partir d'un très grand nombre d'attributs et de valeurs.

3.3.3 Réseau de neurones

Un réseau de neurones est un modèle de calcul dont le fonctionnement schématique est inspiré du fonctionnement des neurones biologiques. Chaque neurone fait une somme pondérée de ses entrées (ou synapses) et retourne une valeur en fonction de sa fonction d'activation.

Cette valeur peut être utilisée soit comme une des entrées d'une nouvelle couche de neurones, soit comme un résultat qu'il appartient à l'utilisateur d'interpréter (classe, résultat d'un calcul, etc.).



La phase d'apprentissage d'un réseau de neurones permet de régler le poids associé à chaque synapse d'entrée (on parle également de coefficient synaptique). C'est un processus long qui doit être réitéré à chaque modification structurelle de la base de données traitée.

3.3.4 Apprentissage non supervisé

On parle d'apprentissage non supervisé lorsque l'on cherche à extraire des informations nouvelles et originales d'un ensemble de données dont aucun attribut n'est plus important qu'un autre. Le résultat des algorithmes de data mining non supervisé doit être analysé afin d'être retenu pour un usage ou tout simplement rejeté.

3.3.4.1 Clustering

Le clustering est une méthode statistique d'analyse de données qui a pour but de regrouper un ensemble de données en différents groupes homogènes. Chaque sous-ensemble regroupe des éléments ayant des caractéristiques communes qui correspondent à des critères de proximité. Le but des algorithmes de clustering est donc de minimiser la distance intra-classe (grappes d'éléments homogènes) et de maximiser la distance inter-classe afin d'obtenir des sous-ensembles le plus distincts possible. La mesure des distances est un élément prépondérant pour la qualité de l'algorithme de clustering.

3.3.4.2 Les règles associatives

Les règles associatives sont des règles qui sont extraites d'une base de données transactionnelles (itemset) et qui décrivent des associations entre certains éléments. Cette technique permet de faire ressortir les associations entre les produits de base (les produits essentiels, ceux pour lesquels le client se déplace) et les produits complémentaires, ce qui permet de mettre en place des stratégies commerciales visant à accroître les profits en favorisant, par exemple, les ventes complémentaires. Ces algorithmes permettent de résoudre des problèmes dits de Frequent Set Counting (FSC).

3.3.4.3 Sequence mining

Le sequence mining concerne la détection de motifs dans les flux de données dont les valeurs sont délivrées par séquences. Cette technique est particulièrement utilisée en biologie pour l'analyse de gènes et des protéines mais également afin de faire du text mining, les phrases étant considérées comme des séquences ordonnées de mots.

3.3.5 Apprentissage incrémental

L'apprentissage incrémental permet à une machine d'apprendre par ajout successif d'informations. Pour être considéré comme tel, un système d'apprentissage doit :

- être capable d'apprendre de nouvelles informations à partir de nouvelles données ;
- être capable de se passer des données d'origine pour entraîner le nouveau classeur ;
- préserver le savoir précédemment acquis ;
- être capable de reconnaître de nouvelles classes introduites dans les nouvelles données.

(Calas, 2009)

4. Les étapes du Big Data

4.1.1 Recueillir les données

Les données peuvent être issues de canaux différents, qui ne correspondent pas forcément aux mêmes métiers et aux mêmes services dans chaque entreprise. Les canaux peuvent être digitaux ou non, ou disposer de leur propre application analytique.

Il s'agit alors de centraliser ces données dans un même ensemble :

- Etablir un panorama des canaux de données existantes.
- Mettre en place de nouveaux supports pour recueillir des feedbacks clients complémentaires : questionnaire en ligne ou questionnaire en magasin, application, site web, réseaux sociaux, carte de fidélité...).
- Faire appel à une solution externalisée, ou développer une solution en interne afin de gérer le flux de données.

On peut également distinguer les données internes, que l'entreprise produit et stocke, et les données externes auxquelles elle peut avoir accès.

4.1.2 Analyser les données

Le flux important de données et d'informations peut présenter un risque pour l'entreprise, en noyant les objectifs dans le volume.

Pour éviter ce risque, le client ou le prospect doit être placé au centre de l'analyse : en quoi les données peuvent-elles permettre d'améliorer son expérience d'achat ? De quelles informations a-t-on besoin pour adapter le produit ou le service à ses attentes ou à son comportement ?

Le projet BigData d'une entreprise peut être développé autour de plusieurs axes :

- Concentrer l'effort sur le client et sur le résultat visé.
- Utiliser les données pour renforcer sa compétitivité.

Méler l'analyse statistique et l'analyse prédictive pour affiner les résultats.

5. Enjeux du BigData pour le SI

D'après deux études menées par l'EMC et IDC en 2012, le Big Data est « une réalité émergente au sein des entreprises françaises ». La stratégie « Big Data » s'est imposée comme une des problématiques majeures liées au développement des nouvelles technologies au sein des entreprises. Elle est considérée comme le moteur de l'innovation, de la satisfaction client et de la réalisation de plus grandes marges de profits et permet une meilleure productivité lorsque les décisions sont prises à partir des analyses et des croisements de données.

Une étude menée par McAfee et Brynjolfsson en 2012 a démontré que les entreprises qui ont adopté des techniques avancées en matière d'analyse de données réalisent de meilleurs taux de productivité et de rentabilité que ceux de leurs concurrents.

Le Big Data permet également une meilleure gestion de l'information en terme de classification des informations par priorité pour éviter la surcharge d'informations non pertinentes.

Les entreprises doivent donc faire face à des enjeux technologiques pour assurer les moyens nécessaires au traitement et à l'analyse de la volumétrie de données. Des enjeux organisationnels sont également à prendre en compte pour mener à bien le projet Big data en décidant de la nature des données à traiter ainsi que de la démarche à suivre. Un projet Big Data est pour une entreprise un très grand enjeu économique.

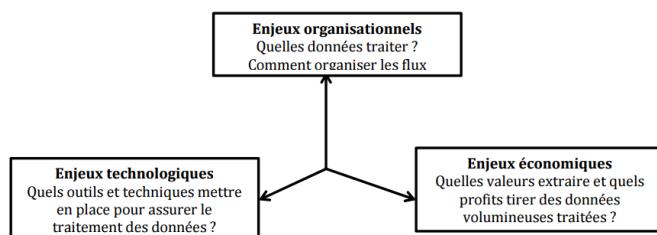


Figure 6 - Enjeux du Big Data pour les entreprises

5.1 Les enjeux technologiques

L'un des enjeux du BigData pour l'entreprise est de s'adapter en investissant dans les nouvelles technologies qui touchent au stockage et à l'analyse de données.

5.1.1 L'investissement en capacité de stockage

Traiter l'augmentation de stockage avec les systèmes actuels reviendrait à dépasser très largement les budgets d'investissements informatiques. Dans l'hypothèse où cela serait possible, il resterait tout de même deux problèmes à résoudre :

- Celui de la place pour stocker les données car les entrepôts de données de l'entreprise seraient très rapidement pleins. Devant le volume de données à traiter, il faudrait ajouter une quantité de disques durs et de baies de stockage tellement importante que l'on se retrouverait en limite de place physique.

- De plus, la dépense énergétique de tout cet équipement ajouterait des dépenses colossales aux budgets de fonctionnement des entreprises.

Les nouvelles technologies Big Data permettent l'évolution du stockage vers des systèmes distribués, grâce à l'infrastructure du Cloud Computing. Cela revient à répartir un même fichier sur plusieurs systèmes et permet de rationaliser les volumes de stockage.

5.1.2 L'investissement dans l'analyse

La valeur des technologies du Big Data provient également des croisements entre diverses sources de données, structurées ou non. L'entreprise doit investir dans ces techniques pour disposer de leur puissance d'analyse. On distingue :

- le Web sémantique, qui permet d'utiliser toutes les ressources de l'Internet et de croiser ces données avec n'importe quelle autre donnée du Web. Il permet ainsi de simplifier la création de valeur à partir des données analysées ;
- les techniques de Datamining (voir 3.3).

5.2 Les enjeux économiques

Surmonter les enjeux organisationnels et technologiques permet de garantir pour les entreprises un retour sur investissement dans un projet Big Data. L'objectif même d'un tel projet est de déployer une stratégie pertinente pour tirer profit de la surabondance exponentielle des données au niveau de différentes fonctions de l'entreprise comme le marketing, les ressources humaines, les finances et la logistique.

5.2.1 L'apport du Big Data pour le Marketing

Les services Marketing ont été les premiers à bénéficier de l'avènement du Big Data. Profitant de toutes les informations disponibles sur le consommateur, les techniques de ciblage sont plus précises et font passer du Marketing de masse au Marketing hyper-personnalisé. Tous les systèmes que nous sommes amenés à utiliser quotidiennement (technologies mobiles, GPS, médias sociaux, transactions en ligne ...) constituent une mine d'informations qui aide les services commerciaux et les services Marketing à mieux connaître le consommateur au point d'anticiper ses besoins.

5.2.2 L'apport du Big Data pour les ressources humaines

Afin de mieux cerner la personnalité des candidats pour un poste, des entreprises ont mis en place des tests qu'elles soumettent aux candidats. La quantité d'information générée par ces tests (temps d'hésitation, ordre des actions choisies, etc.) permet, après une analyse par des algorithmes, de mettre en avant certains traits caractéristiques du candidat. Une autre application du Big Data est l'étude de la corrélation entre certaines réponses dans les tests de personnalité et la performance dans un secteur donné. Cela permet aux entreprises de mieux cibler les profils recherchés et donc d'améliorer la qualité de leur recrutement. (Big data et Ressources Humaines)

5.2.3 L'apport du Big Data pour les finances

Dans la Finance, la Banque et l'Assurance les projets Big Data sont nombreux. Par ailleurs dans les métiers du trading ces projets sont aussi conçus pour identifier les mouvements de fonds suspects. Cela peut être aussi l'occasion de créer de nouveaux services pour les clients. Les banques, par exemple, détiennent souvent des historiques d'opérations sur plusieurs dizaines d'années. L'exploitation de ces données peut être utile pour la banque mais surtout aujourd'hui pour le client de la banque.

5.2.4 L'apport du Big Data pour la logistique

Les transports de marchandises représentent un élément important quant à la logistique d'une entreprise. En effet, s'assurer du bon acheminement au niveau du temps imparti des marchandises permet de garantir une gestion de stock optimale. Egalement cela permet de pouvoir respecter les délais des commandes qui ont été effectuées. Chaque jour il y plus de 150 000 livraisons effectuées avec 90 000 véhicules dans le monde il est donc primordial d'analyser les données de ces transports afin d'obtenir un gain de temps et de réaliser des économies. Bien entendu, l'utilisation de ces données au niveau logistique n'est qu'à ses débuts mais déjà prometteuse.

Les données récoltées sur les trajets précédents permettent d'analyser les comportements des chauffeurs. Egalement, l'analyse se fait sur le trafic existant sur les chemins empruntés comme les heures de pointes. Ainsi, les données apportent des directives sur les chemins à emprunter mais également sur la manière de conduire afin de réduire le temps et les coûts. Connaître les trajets passés est un bon indicateur pour améliorer ses trajets. Mais il ne faut pas oublier les facteurs extérieurs liés aussi bien à la concurrence qu'au trajet en lui-même, avec la possibilité d'un trafic inhabituel dû à un accident de circulation. Il est donc important de pouvoir modifier l'itinéraire d'un camion en s'adaptant au trafic en temps réel. (Molla)

5.3 L'impact organisationnel

L'intégration des nouvelles technologies de Big Data dans l'entreprise va entraîner des changements d'ordre organisationnels. Nous allons en analyser deux types :

- la conduite du changement ;
- l'apparition de nouveaux métiers.

5.3.1 La conduite du changement

L'intégration du Big Data dans une entreprise risque d'entrainer une rupture de son organisation. Elle devra modifier son mode de fonctionnement pour s'adapter à ce nouvel écosystème (arrivée de nouveaux acteurs, renforcement de la concurrence, importance des réseaux sociaux, changement de la relation avec les clients...).

Un manque d'adaptation de l'entreprise pour canaliser le Big Data la mettrait dans des situations à risque comme par exemple une perte de productivité de la part de ses employés ou une dégradation de la réputation de l'entreprise.

Pour réussir cette transformation, il est nécessaire de faire évoluer les mentalités au sein des entreprises. La priorité réside dans la maîtrise des données.

Pour atteindre une efficacité optimale, il faut un partenariat renforcé entre les équipes informatique et marketing qui sont les plus grands acteurs du Big Data au sein de l'entreprise.

5.3.2 L'apparition de nouveaux métiers

Le Big Data a vu l'apparition de nouveaux métiers qui sont devenus incontournables dans ce domaine :

- Chief Data Officer (CDO), qui est le Directeur de la data, le gardien de l'éthique. Il est à la tête d'une équipe spécialisée dans l'acquisition, l'analyse et l'exploitation des données. Sa fonction consiste à la gouvernance de son équipe pour l'approvisionnement des données les plus intéressantes et cohérentes pour l'intérêt de l'entreprise. Il organise le partage de leur analyse avec les directions métiers, et fait respecter l'éthique en matière d'usage de ces informations. Il s'appuie, avec son équipe, sur des connaissances pointues en statistiques, informatique et numérique pour donner des repères à chaque département : marketing, ressources humaines, ingénierie, service qualité, comptabilité et gestion. Un diplôme issu d'écoles d'ingénieurs est de rigueur, ainsi que des compétences et une solide expérience dans les domaines du management, de l'informatique et du marketing sont nécessaires. Le salaire annuel brut, avec 10-15 années d'expérience, est aux alentours de 120 000 euros.
- Business Intelligence Manager dont le travail consiste à faciliter les prises de décision du CDO. Il utilise des nouvelles technologies pour mettre en place des tableaux de bord, des outils de reporting, afin de les intégrer au système informatique et de les rendre accessibles aux utilisateurs au sein de l'entreprise. Ce métier nécessite de solides connaissances en anglais, en informatique et en gestion des données. Comme pour le CDO, un diplôme issu d'écoles d'ingénieurs est de rigueur. Le salaire annuel brut moyen, avec 5-10 années d'expérience, s'élève à 75 000 euros.
- Data Scientist qui est responsable de la collecte, du traitement, de l'évaluation et de l'analyse des données massives, ou Big Data, afin d'optimiser la stratégie de l'entreprise. Son rôle est de créer pour les métiers de l'entreprise des algorithmes qui produisent des informations utiles, notamment afin de proposer aux clients, les produits qu'ils recherchent. Ce sont des profils qui mèlagent des compétences en management, informatique et statistiques. Ils maîtrisent les techniques du datamining, ainsi que les technologies et les outils informatiques des bases de données tels que Hadoop, Java,

MapReduce, Bigtable, NoSQL... Un diplôme issu d'écoles d'ingénieurs est de rigueur. Le salaire annuel brut, avec 3 années d'expérience, est aux alentours de 60 000 euros.

- Data Analyst qui utilise des techniques statistiques et des outils informatiques spécialisés afin d'organiser, de synthétiser et de traduire les informations dont les entreprises ont besoin pour faciliter les prises de décisions. Les spécificités du poste se situent dans le volume des données traitées et la maîtrise des outils spécifiques au Big Data. Son rôle est de faire parler les données et d'en sortir des indicateurs concrets au service de la direction générale pour les exploiter à des fins commerciales. Un diplôme issu d'écoles d'ingénieurs est de rigueur. Le salaire annuel brut, avec 2-3 années d'expérience, est aux alentours de 40 000 euros.
- Le Data Miner qui est le « fouilleur de données », le Sherlock Holmes de la data. Son rôle est de dénicher les informations parmi de multiples données, afin de les rendre exploitables et utiles pour l'entreprise. Il doit disposer d'excellentes compétences en informatique, en statistiques et en business. Il est possible de devenir Data Miner à partir d'une licence en informatique ou en marketing. Il peut évoluer comme Data Analyst puis Data Scientist avec le temps. Le salaire annuel brut, avec 2-4 années d'expérience, est aux alentours de 55 000 euros.
- Master Data Manager, également appelé Gestionnaire des Données, la Data Manager acquiert et organise les informations de l'entreprise en vue de leur exploitation optimale. C'est un expert des données de base qui englobent, d'une part, les données référentielles (liées aux catalogues fournisseurs, clients, articles, etc...) et les métadonnées structurantes (liées aux normes et méthodes réglementaires). Il doit veiller à ce que ces données soient bien conformes et organisées selon les règles de gestion définies et correctement intégrées dans le système d'information exploité par les équipes métier. Le salaire annuel brut, avec 2-4 années d'expérience, est aux alentours de 45 000 euros.
- Data Protection Officer qui est la personne garante de la protection des données personnelles dans l'organisation pour laquelle elle travaille. Ce poste pourrait devenir obligatoire dans toutes les entreprises ayant plus de 250 salariés. Le métier comprend donc un important volet de sensibilisation. Sa fonction est transversale. Son challenge est d'être tenu au courant de tous les projets de traitement des données pour apporter ses préconisations en amont. Il doit réunir des compétences en informatique, en droit, mais aussi en communication. C'est un nouveau métier du digital qui apparaît dans un contexte de forte concurrence où la protection des données est au cœur des problématiques entreprises et représente un enjeu majeur de l'économie. Le salaire annuel brut, avec 1-3 années d'expérience, est aux alentours de 35 000 euros.

Les spécialistes du Big Data sont à la fois très rares et très demandés. On les retrouve principalement au sein de grands groupes des secteurs de la banque, de l'assurance et de la finance, ou chez les opérateurs qui stockent et traitent des données comme les data centers, les fournisseurs d'accès internet et les hébergeurs. Mais la réglementation sur les données évolue rapidement, et les opportunités business data se multiplient, toutes les entreprises se verront entourées, de près comme de loin, de profils semblables à ceux présentés ci-dessus.

(Delort)

5.4 Synthèse

Le Big Data a des enjeux réellement importants pour une entreprise notamment dans la création de valeur. Mais cela a également un impact conséquent sur l'organisation d'une entreprise notamment au niveau des coûts de mises en place et de la formation des différentes parties prenantes concernées. (Engohan) (Karoui, Davauchelle, & Dudezert)

6. Enjeux de compétitivité

6.1 Améliorer la gestion de la production

Développées avec des moyens manuels, de nombreuses entreprises ont besoin aujourd’hui de gagner en productivité. Grâce à l’automatisation de leurs processus opérationnels via la numérisation des ateliers de production, les PMI seront en mesure de mieux piloter leurs coûts et de réaliser d’importantes économies.

Dans cette optique, l’utilisation de logiciels de Gestion de Production Assisté par Ordinateur (GPAO) est indispensable, notamment pour récolter des données relatives au suivi des commandes et au suivi des coûts de production. Les informations concernant le coût de production, que ce soit l’investissement en machines, le prix de la matière première ou le salaire de la main d’œuvre, seront ainsi transmises en temps réel aux collaborateurs compétents qui pourront déterminer les poches d’économies réalisables.

Aussi, le recours à un logiciel ERP complet, dans lequel la gestion de production est intégrée à la gestion commerciale, peut être très utile aux PMI. Elles sont ainsi en mesure de savoir d'où elles partent, où elles vont, et à quel prix ; ce qui leur permet d'alléger leurs chaînes d'achat et de dégager des marges plus conséquentes.

6.2 Répondre aux normes de traçabilité et de qualité

La traçabilité est un enjeu majeur pour les PMI. Toute entreprise industrielle, quels que soient sa taille et son secteur d’activité, est désormais soumise à des réglementations établies par les autorités publiques, et le besoin de traçabilité des matières premières et des produits finis est devenu une contrainte inévitable dans ce secteur mondialisé. Afin de prouver le respect des normes ISO, l’entreprise doit collecter, archiver et classer de nombreuses données quant à son activité, ce que la GPAO assure de manière fiable à travers le suivi intégré d’un certain nombre de critères.

Par ailleurs, en mettant en place des dispositifs précis de suivi et d’analyse de la chaîne de production, la PMI peut être éligible pour l’obtention de labels de qualité, comme les deux labels français « Made in France » et « OFG » (Origine France Garantie). La valeur d’un produit se trouvant aujourd’hui moins dans le prix que dans la qualité pour de nombreux consommateurs, de tels labels sont de véritables arguments commerciaux, dont l’entreprise pourra tirer d’importants bénéfices.

6.3 Fidéliser les consommateurs grâce à une meilleure gestion commerciale

Le Big Data favorisent une vision stratégique de l’entreprise en permettant au dirigeant de mieux connaître ses activités et l’environnement économique dans lequel elles se développent, ainsi que ses clients et leurs comportements. En récoltant, structurant et analysant des données sur les consommateurs, leurs habitudes d’achat, et leurs relations aux produits proposés, la

PMI peut développer des outils de fidélisation efficaces et mieux anticiper les tendances du marché.

6.4 Identifier les besoins pour mieux innover

Une utilisation stratégique et réfléchie des données peut être un puissant levier de croissance et de compétitivité pour les PMI. En étant plus alertes sur les tendances du marché, elles peuvent adapter leurs offres et produits aux besoins et à la demande des consommateurs et innover dans ce sens.

Aussi, la GPAO permet aux dirigeants d'identifier les besoins de leur entreprise en équipement machines, mais aussi de mieux connaître les modifications qui peuvent être apportées aux processus de fabrication pour proposer des produits plus innovants, performants et pérennes. Si de telles informations sont suivies des démarches adéquates, cela aura un impact très positif sur le bilan de la PMI.

Le retour sur investissement en termes de Big Data est double pour les PMI qui, selon une étude de McKinsey, peuvent augmenter leurs marges opérationnelles de 60% tout en réduisant leurs coûts de 10 à 15%. Une bonne gestion de la production est donc indispensable pour les PMI désirant se développer. En utilisant à bon escient les données récoltées par des outils comme la GPAO, l'entreprise bénéficie d'un levier de compétitivité important pour une activité pérenne.

(Sage, 2016)

7. Cas d'usages pour l'industrie et la ville de demain

Cette partie a pour but de présenter certaines applications de le Big Data dans différents secteurs d'activités.

7.1 Secteur de la distribution

La grande distribution a été pionnière dans la mise en œuvre de data warehouses d'entreprise, notamment pour l'analyse des produits vendus aux clients.

L'exploitation de ces données dans la distribution permet actuellement :

- une gestion plus fine et dynamique des prix de vente ;
- une personnalisation des offres pour les programmes de fidélité ;
- un ajustement de l'offre et de la demande, par zone géographique ;
- une gestion du on line multi-canal³.

7.2 Secteur de l'hôtellerie

Le secteur de l'hôtellerie est également très présent dans l'univers du Big Data. En effet, avec la quantité d'informations disponibles sur les individus, en fonction des types d'endroits fréquentés permet de faire d'exploiter les données de la manière suivante :

- une capacité à tester des nouvelles offres, et de les retirer si elles sont peu efficaces (en se servant des avis des clients sur les sites spécialisés) ;
- proposer une offre personnalisée pour chaque client ;
- effectuer une communication personnalisée en fonction des clients.

7.3 Secteur de l'énergie

En France, 25% de l'eau injectée sur le réseau est perdue en fuite et fraudes, le manque à gagner s'élèverait à 2.4 milliards d'euros par an. Grâce au Big Data et aux objets connectés il serait possible de remédier à ce problème et rendre publique les données collectées.

L'exploitation de ces données pourrait permettre :

- de donner des informations en temps réel sur les débits et la qualité de l'eau ;
- de détecter plus rapidement un problème sur le réseau de distribution ;
- d'automatiser les processus de collecte.

C'est notamment « Veolia Habitat Service » qui travaille sur ce type de projets.

³ En marketing, un canal est une interface (physique ou virtuelle) par lequel le client va passer à l'acte d'achat. L'enjeu d'une stratégie multicanal est, comme son nom l'indique, de développer de nouveaux points de contact de vente avec les clients

EDF travaille également à moderniser son réseau et en mettant en place du Big Data avec ses nouveaux compteurs électriques « Linky » qui permettent notamment de transmettre des données sur les consommations en temps réel aux particuliers mais également d'effectuer l'ouverture à distance sans besoin de déplacement d'un technicien.

7.4 Secteur des assurances

Le Big Data est utilisé dans le secteur des assurances surtout pour la gestion des fraudes.

Le principe du Scoring Crédit est utilisé notamment pour analyser les déclarations de sinistre, ces données sont ensuite exploitées pour mettre en avant les incohérences dans les déclarations. Cela permet également à l'aide à la décision poussée sur le terrain au plus près des clients. (Franco, 2014)

8. Conclusion

Le BigData est un concept vieux de quelques dizaines d'année déjà qui est devenu mature ces cinq dernières années. De plus en plus d'entreprises commencent à s'intéresser aux solutions de BigData afin d'améliorer leur compétitivité et de se distinguer par rapport à leurs concurrents. En effet le BigData apporte un avantage concurrentiel à ces dernières en leur permettant d'analyser le comportement de leurs clients et d'agir en conséquence.

La montée du Big Data voit également l'arrivée de nouveaux métiers offrant de nouvelles possibilités sur le marché de l'emploi.

Il ne constitue sans doute pas une révolution de l'ampleur de celles de l'agriculture ou de l'industrie, et à même d'engendrer un nouvel âge d'or. Cependant, en ouvrant à l'induction des données, il permet des transformations de nombreux champs d'activité : politique, social, éducatif, judiciaire, sportif, personnel... avec des aspects juridiques, éthiques ou encore psychologiques à ne pas oublier.

Le BigData est un domaine en plein expansion que les entreprises vont devoir adopter si elles ne veulent pas se faire devancer sur leurs marchés.

9. Bibliographie

(s.d.). Récupéré sur La science des données : pour qui ? Pourquoi ?:

<https://www.youtube.com/watch?v=SCMYCCI1pu4>

Big data et Ressources Humaines. (s.d.). Récupéré sur ArchiBat:

<http://www.archibat.com/blog/big-data-ressources-humaines/>

Calas, G. (2009). *Études des principaux algorithmes de data mining.* Récupéré sur

<http://guillaume.calas.free.fr/data/Publications/DM-Algos.pdf>

Crawford, D. B. (s.d.). *Critical questions for Big Data.* Récupéré sur

<http://www.tandfonline.com/doi/full/10.1080/1369118X.2012.678878>

Delort, P. (s.d.). *Le Big Data.* Récupéré sur <https://www.cairn.info/le-big-data--9782130652113-page-123.htm>

Engohan, T. (s.d.). *Les enjeux du Big Data pour l'Entreprise.* Paris.

Franco, J.-M. (2014, Mars 19). *Big Data : concepts, cas d'usage et tendances.* Récupéré sur

SlideShare: <http://fr.slideshare.net/jmfranco/niort-mars-14>

Karoui, M., Davauchelle, G., & Dudezert, A. (s.d.). *Big Data : Mise en perspective et enjeux pour les entreprises .*

Loukides, M. (2011). *What is Data Science.* O'REILLY.

Michael Cox, D. E. (s.d.). *Application-Controlled Demand Paging for Out-of-Core Visualization.* Récupéré sur

<https://www.nas.nasa.gov/assets/pdf/techreports/1997/nas-97-010.pdf>

Molla, M. (s.d.). *La big data au service de la logistique des transports ?* Récupéré sur

SupplyWeb: <http://www.supplyweb.fr/la-big-data-au-service-de-la-logistique-des-transports/>

Olson, M. (s.d.). *HADOOP: Scalable, Flexible Data Storage and Analysis.* Récupéré sur

http://www.cloudera.com/content/dam/cloudera/Resources/PDF/Olson_IQT_Quarterly_Spring_2010.pdf

Sage. (2016, 01 29). *Les Big Data, un levier de compétitivité essentiel pour les PMI.*

Récupéré sur sage: <https://blog.sage.fr/big-data-levier-competitivite-pmi/>