

Louvain School of Management

Les critères d'un stockage de données marketing de qualité

Auteur : **Maxime Alliaume**
Promotrice : **Chantal de Moerloose**
Année académique 2020-2021
Master 120 crédits en Sciences de gestion, à finalité spécialisée

Je tiens à remercier le professeur Chantal de Moerloose pour son suivi, ses précieux conseils et sa disponibilité dans le cadre de la rédaction de ce mémoire. Mes pensées vont également à ma famille et mes proches qui m'auront soutenu tout au long de ce projet.

Je remercie par ailleurs l'ensemble des intervenants ayant pris part à la démarche qualitative de ce mémoire, pour leur disponibilité et leur temps de parole.

Table des matières

Chapitre 1 : Introduction	1
Chapitre 2 : Revue de la littérature.....	3
2.1. Introduction au Big Data	3
2.1.1. Définition et caractéristiques du Big Data	3
2.1.2. Enjeux du Big Data pour les entreprises.....	5
2.1.3. Données	6
2.2. Application du Big Data au marketing	8
2.2.1. Contribution du Big Data au marketing	8
2.2.2. Données marketing	9
2.2.3. First-, Second- et Third-party data	11
2.2.4. Règlement général sur la protection des données	12
2.3. Qualité des données	14
2.3.1. Définition et caractéristiques de la qualité des données	14
2.3.2. Impacts de la qualité des données	19
2.3.3. Types d'anomalies	21
2.3.4. Gérer et améliorer la qualité des données	28
i. Processus DMAIC	29
ii. Outils de gestion et d'amélioration de la qualité des données.....	30
2.4. Processus de stockage de données.....	32
2.4.1. Bases de données	32
2.4.2. Focus sur les datacenters : cas de Google	34
i. Économie de ressources.....	34
ii. Économie d'énergie	35
iii. Garantie de sécurité	35
2.5. Revue de la littérature : récapitulatif	37
Chapitre 3 : Méthodologie de l'étude qualitative.....	38
Chapitre 4 : Résultats empiriques	41
4.2. Observations	41
4.2.1. Qualité des données stockées.....	41

i. Observations relatives aux critères identifiés dans la littérature	41
ii. Nouveaux critères identifiés au terme des discussions.....	48
4.2.2. Qualité du processus de stockage	50
i. Observations relatives aux critères identifiés dans la littérature	50
4.3. Confrontation des sous-items associés à chaque critère avec Ali El Maghraoui.....	53
4.4. Conclusions relatives à l'étude qualitative.....	55
4.4.1. Synthèse des observations	57
4.4.2. Résultats et classification des critères par ordre d'importance.....	58
4.5. Limites dans la démarche qualitative.....	60
Chapitre 5 : Réflexion sur l'aspect sociétal du stockage de données.....	61
5.1. Impact environnemental mitigé du stockage des données.....	61
5.1.1. Augmentation de la consommation en énergie des datacenters	61
5.1.2. Amélioration de l'efficacité énergétique du Cloud au regard des datacenters	62
5.2. Impact social du stockage des données.....	62
5.2.1. Remise en question du fonctionnement des entreprises en matière de protection de la vie privée de leurs utilisateurs	63
5.2.2. Méfiance des utilisateurs envers les entreprises	63
Chapitre 6 : Conclusion	64
Bibliographie	67
Annexes	78

Table des annexes

Annexe 1. Datasphère à l'échelle mondiale par année	78
Annexe 2. Mesures courantes de stockage de données	78
Annexe 3. Typologie de données	79
Annexe 4. Trois exemples de droits fondamentaux individuels défendus par le RGPD	79
Annexe 5. Quelques fournisseurs de données tierces (<i>Third-party data</i>)	80
Annexe 6. Guide d'entretien soumis aux interlocuteurs dans la démarche qualitative	81
Annexe 7. Comptes-rendus des entrevues menées dans la démarche qualitative	82
Annexe 7.1. Rencontre avec Jean-Christophe Dewitte	82
Annexe 7.2. Rencontre avec Marguerite Paris	84
Annexe 7.3. Rencontre avec Camille Laimay	85
Annexe 7.4. Rencontre avec Éric de Bellefroid	86
Annexe 7.5. Rencontre avec Ali El	88
Annexe 8. API – Application Programming Interface	91
Annexe 9. Donnée, information, connaissance	91

Chapitre 1 : Introduction

La numérisation croissante constitue l'un des plus grands défis actuels rencontrés par les entreprises. En effet, il serait difficile d'imaginer notre société vivre sans smartphone, ordinateur et connexion Internet. Cette tendance démultiplie toujours plus la quantité de données produites dans le monde. L'organisme spécialiste en statistiques Statista évalue en effet le volume global annuel de données créées à près de 50 zetaoctets (i.e. 50.000 milliards de milliards d'octets) en 2020, soit plus de 20 fois celui cumulé au cours des dix dernières années (Statista, 2021a). Le marketing figure parmi les branches du management les plus impactées par cette révolution numérique. La variété et la quantité des données disponibles rendent effectivement la tâche de leurs utilisateurs particulièrement difficile et impactent la qualité des bases de données concernées. Aussi, la production constante de données encourage à tort ces mêmes utilisateurs à en stocker un maximum dans la mesure du possible. C'est dès lors à se demander quelles sont les données qu'une entreprise doit conserver mais aussi entretenir, car toutes les données ne sont pas bonnes à stocker.

C'est pourquoi ce mémoire s'intitule **Les critères d'un stockage de données marketing de qualité**. Autrement dit, nous nous poserons la question suivante : *“Quels sont les critères qui font d'un stockage de données marketing un stockage de qualité ?”* Avant d'aller plus loin, il est important de décomposer cet intitulé en s'intéressant à chacune de ses composantes majeures, à savoir : donnée marketing, stockage, critère et qualité. Tout d'abord, une donnée marketing désignera toute information explicite relative à un individu (qu'il soit consommateur ou non). Par exemple, son âge, son adresse postale ou de courrier électronique, mais aussi plus largement son comportement, ses réseaux, ses préférences de consommation, etc. Ensuite, le stockage de données renvoie à leur conservation au moyen d'une solution technologique spécialement imaginée pour les contenir dans la durée et les rendre disponibles à la demande (Hewlett Packard, 2021). Par exemple, un disque dur ou un serveur informatique. En outre, un critère se définira comme toute caractéristique ou dimension qui permet de reconnaître une donnée de qualité. Enfin, une donnée sera dite de qualité si celle-ci parvient à répondre aux attentes de ses utilisateurs (Brasseur, 2005). Nous décomposerons ce dernier concept en deux dimensions, à savoir la qualité des données elles-mêmes, mais aussi la qualité de leur processus de stockage.

Afin de répondre à notre problématique, nous restituerons préalablement le contenu existant dans la littérature, ce afin d'établir une première liste de critères de qualité de données. Ensuite, nous confronterons cette liste, finalement très théorique, par le biais d'une étude qualitative menée auprès de cinq experts du domaine. L'idée est alors d'en renforcer le contenu (à la hausse ou à la baisse) en s'intéressant aux expériences pratiques de plusieurs spécialistes. Finalement, la réponse à notre problématique prendra la forme d'un tableau à double entrée dans lequel nous classerons par ordre d'importance les critères relevés communément dans la littérature ainsi qu'au travers de la démarche qualitative.

Ainsi, ce mémoire se structure comme suit. Nous débuterons par un chapitre consacré à la revue de la littérature, ce afin de documenter le sujet d'étude. Nous introduirons entre autres les concepts de Big Data, de données et plus particulièrement de données marketing, pour terminer par la qualité des données. Dans cette section, nous aurons l'occasion d'approcher l'aspect **scientifique** de la thématique de notre mémoire, en corrigeant et complétant un contenu établi après de nombreuses lectures littéraires. Ensuite, le troisième chapitre fera état de la méthodologie abordée dans le cadre de notre étude qualitative. Par ailleurs, nous discuterons des résultats de cette enquête au travers de la quatrième section, dans laquelle nous relèverons, pour chaque critère identifié dans la littérature, les propos de chacun des intervenants. De cette manière, nous pourrions en tirer des conclusions. De plus, nous aborderons dans ce chapitre l'apport **managérial** de notre mémoire. En effet, nous tenterons au travers de cette production de mettre en avant plusieurs points d'attention et de proposer une série de recommandations à destination des entreprises et autres acteurs du stockage en masse de données. Finalement, le cinquième et dernier chapitre (à caractère tout à fait personnel) se concentrera de manière limitée et non exhaustive sur l'aspect **sociétal** de notre sujet de mémoire, très peu couvert dans les sections précédentes. Dans cette partie, nous relèverons quelques impacts sociétaux du stockage des données et tenterons ainsi de mettre le doigt sur ses répercussions environnementales et sociales.

Chapitre 2 : Revue de la littérature

La revue de la littérature a pour finalité de documenter le sujet d'étude. Répartie en trois points, elle se concentrera tout d'abord sur le concept de Big Data à proprement parler, en relevant d'une part de manière introductive ses principes, caractéristiques et les enjeux qu'il couvre actuellement, en s'intéressant d'autre part au rôle qu'il joue dans le marketing. Ensuite, nous aborderons en détails la notion de qualité de la donnée au travers de la seconde partie, qui permettra d'approcher, au cours du dernier point, le concept de stockage de l'information. Nous nous attarderons tout d'abord sur le type de base de données le plus courant dans un contexte Big Data, et illustrerons enfin le concept de *datacenter* au travers du cas de Google en Belgique.

2.1. Introduction au Big Data

Depuis des milliers d'années maintenant, la société assiste à une progression étonnante de la quantité de données produites, de la tablette mésopotamienne à la révolution du digital, en passant par l'avènement des ordinateurs. De plus, la digitalisation accentue cette tendance : l'utilisation des GPS, des smartphones ou des réseaux sociaux par exemple produit une très large quantité de données (*cf.* Annexe 1, p. 78), certaines d'une très grande valeur pour les entreprises. Ces évolutions ont fait naître le concept de Big Data.

2.1.1. Définition et caractéristiques du Big Data

Récemment apparu à la fin du XX^e siècle, le concept de Big Data désigne une quantité très importante de données numériques, de même que les techniques et outils permettant de les exploiter de manière efficace afin de leur apporter du sens (Encyclopædia Universalis, 2021). Appliqué au cas d'une entreprise, le Big Data permet par exemple d'évaluer le succès d'un produit ou d'un service sur un marché, ou encore de mieux comprendre un environnement commercial en vue de construire un ou plusieurs avantages concurrentiels (Davenport, 2012). Par ailleurs, Chen, Chiang et Storey (2012) associent au Big Data trois caractéristiques majeures, plus communément appelées trois *V's* : volume des données, variété des données, et vitesse des données.

Volume. Cette caractéristique fait référence à l'ampleur des données numériques (Gandomi & Haider, 2014). Selon Russom (2011), le volume des données constitue un attribut fondamental

du Big Data. Autrefois mesurés en méga- ou gigaoctets¹, ces volumes de données sont aujourd'hui exprimés en téra- ou pétaoctets (*cf.* Annexe 2, p. 78). Effectivement, les données collectées continuellement et instantanément depuis Internet accroissent rapidement les volumes de données à entreposer (Russom, 2011).

Variété. Elle renvoie à la diversité structurelle d'un ensemble de données (Gandomi & Haider, 2014). D'une part, la variété des sources d'où proviennent les données collectées (Russom, 2011). Par exemple, Internet ou les médias sociaux. D'autre part, la variété des données collectées elles-mêmes (données structurées, semi-structurées ou non structurées, données internes ou externes). Par exemple, du contenu audio ou vidéo, des données personnelles, bibliographiques ou muséologiques. Ce dernier point sera abordé en détails dans une prochaine section dédiée aux types de données.

Vélocité. La vélocité renvoie d'une part à la vitesse à laquelle les données sont générées, d'autre part à la rapidité avec laquelle celles-ci sont analysées et traitées (Gandomi & Haider, 2014). Prenons l'exemple d'une voiture autonome circulant sur la chaussée. Celle-ci dispose de capteurs pour collecter en temps réel les données qui concernent les objets environnants (conducteurs ou obstacles) afin de rouler en toute sécurité. L'analyse des données en temps réel depuis la voiture doit se faire très rapidement, ce afin de prendre les décisions les plus rapides et sûres pour le conducteur (Cheruvu, 2015). Outre cela, la vélocité aide les organisations à comprendre la croissance relative de leur Big Data et la rapidité avec laquelle ces données atteignent les utilisateurs, les applications et les systèmes sources (Chevalier, 2020).

Cette définition axée autour de ces trois grands principes peut être approfondie en suggérant d'autres caractéristiques. Parmi celles-ci, la *véracité* des données, qui met en balance volume et qualité des données. Selon Gandomi & Haider (2014), elle représente en effet le manque de fiabilité de certaines sources de données. Par exemple, les sentiments des consommateurs dans les médias sociaux sont incertains par définition, car ils impliquent un jugement humain. Pourtant, ils peuvent contenir des informations d'une grande valeur. Ainsi, traiter des données imprécises et incertaines constitue une facette supplémentaire du Big Data, gérée à l'aide d'outils et d'analyses développés pour l'exploitation de données incertaines (Gandomi & Haider,

¹ Un octet (ou *byte* en anglais) est une unité de mesure relative à la quantité de données d'un fichier (Encyclopédie Larousse en ligne, 2021).

2014). De plus, Bidoit-Tollu & Doucet (2017) parlent de la *visibilité* des données, qui permet de faire naître du volume de données des structures pertinentes pour l'analyse des informations. Enfin, citons la *valeur* des données (qu'elle soit économique ou scientifique par exemple), qui estime le bénéfice espéré de leur utilisation (Berti-Equille, 2017) : une fois collectées, elles doivent avoir du sens pour l'utilisation que l'on désire en faire.

En somme, croiser une quantité innombrable de données (*volume*) de différents types (*variété*) en temps réel (*vélocité*), tenter de les combiner et les corrélérer constituent la véritable richesse du Big Data. Dans le cas d'une entreprise, cela permet d'affiner la connaissance d'un environnement, d'un client ou d'un marché, mais aussi de révéler des connaissances jusqu'alors inaccessibles et d'en tirer un maximum de profit (Karoui, Devauchelle, & Duzert, 2014).

2.1.2. Enjeux du Big Data pour les entreprises

Pour rappel, le Big Data suppose pour une entreprise d'être en mesure de traiter et analyser en temps réel des flux considérables de données à la forme structurée ou non structurée, issues de sources variées, ce afin d'en extraire de la valeur et d'en tirer du profit le plus rapidement possible (*cf. supra*). Néanmoins, volume et variété des données entraînent des difficultés en matière de gestion et de compréhension de ces informations de la part du marketing (Davoust & Volle, 2014). De nouvelles compétences techniques mais aussi organisationnelles voient ainsi le jour en entreprise. Bon nombre d'organisations investissent dès lors dans le recrutement de *Data Scientists*, autrement dit des profils d'individus ayant à la fois une vision du marché, mais également des compétences d'ordre technique (Davenport & Patil, 2012) et analytique (Davoust & Volle, 2014), en mesure d'identifier la valeur des sources de données diverses, parfois complexes et faiblement contrôlées. Cependant, de tels profils manquent à l'appel sur le marché de l'emploi (Karoui *et al.*, 2014). Les entreprises se voient ainsi contraintes de former leurs propres experts en la matière (Davenport, Barth, & Bean, 2012). De plus, le Big Data présuppose un partage d'informations ainsi qu'une collaboration étroite entre les entités (ou départements) d'une entreprise concernées par la collecte, le traitement et le stockage des données (Day, 2011). Pensons au marketing, à la fonction commerciale et au pôle informatique (Davoust & Volle, 2014). Par ailleurs, Karoui *et al.* (2014) s'attardent sur les enjeux technologiques qu'implique le Big Data. En effet, ce dernier suppose le déploiement de nouveaux outils informatiques suffisamment performants pour pouvoir prétendre à une analyse

de qualité des données, outils qui doivent être intégrés à l'infrastructure informatique existante. En outre, la taille des informations pouvant excéder le pétaoctet (*cf.* Annexe 2, p. 78), les simples disques durs ne suffisent plus : place aux *datacenters*. Cette pratique de stockage de données coûteuse et énergivore (Cappy, 2017) consiste en l'immobilisation d'espaces-serveurs sur un site dédié (Karoui *et al.*, 2014) (*cf. infra*).

2.1.3. Données

Tel qu'abordée ci-avant, la complexité du Big Data s'explique entre autres par la variété des données collectées. Mais comment définir une donnée ? Quels sont les types de données que les entreprises sont amenées à collecter, traiter et stocker ?

Nombre d'ouvrages tentent d'éclaircir la question de la définition du terme *donnée* (*cf.* Annexe 9, p. 91). Nous retiendrons surtout la définition que fournissent Elmasri et Navathe (2016). « Par données, nous entendons des faits connus qui peuvent être enregistrés et qui ont une signification implicite. Par exemple, considérez les noms, les numéros de téléphone et les adresses des personnes que vous connaissez. » (Elmasri & Navathe, 2016, p. 4). Les données constituent le cœur du Big Data, en le sens qu'elles nourrissent les outils qui permettent de les traiter, d'en évaluer la qualité, et de les exploiter. Aussi, Karoui *et al.* (2014) différencient les données selon leur source et leur type (*cf.* Annexe 3, p. 79). Ils identifient en effet deux sources de données. D'une part, celles issues de l'entreprise-même, appelées données internes. Autrement dit, elles représentent les données privées ou exclusives qu'une entreprise détient (Marr, 2015). Citons par exemple les données relatives au stock des produits, le CRM² ou encore l'historique des ventes (Nair & Narayanan, 2012). D'autre part, les données qui existent en dehors de cette même organisation sont appelées données externes. Par exemple, les médias sociaux, les données météorologiques ou encore gouvernementales (Karoui *et al.*, 2014). Marr (2015) précise : ces informations peuvent être de l'ordre public ou privé. Les données publiques sont accessibles à tout le monde – soit en les récoltant librement, soit en payant un tiers pour les recueillir. Les données privées sont celles qu'une entreprise doit collecter et acheter auprès d'une autre organisation ou d'un fournisseur de données tiers (*cf. infra*).

² *Customer Relationship Management* ou CRM renvoie à une solution technologique qui permet de gérer toutes les relations de votre entreprise avec les clients et les clients potentiels (prospects) en centralisant toutes les données importantes qui les concernent sur une plateforme unique (Salesforce, 2021a).

Par ailleurs, Karoui *et al.* (2014) distinguent différents types de données : les données structurées, semi-structurées et non structurées. D'après Marr (2015), les données qui se trouvent dans un champ fixe au sein d'un enregistrement ou d'un fichier défini sont appelées données structurées. Cela inclut les données contenues dans les bases de données relationnelles ou les feuilles de calcul par exemple. En effet, comme l'indique leur nom, les données structurées renvoient à des informations qui suivent un modèle de données prédéfini ou sont organisées d'une manière prédéterminée (Marr, 2015), par exemple dans des colonnes. Les données structurées attribuent un nom à chaque champ d'une base de données et définissent les relations entre ceux-ci. Alwan & Ku-Mahamud (2019) complètent en avançant que celles-ci désignent généralement des données dont la longueur et le format sont prédéfinis. La majorité des experts concluent que cette catégorie de données occupe environ un quart des données disponibles (Alwan & Ku-Mahamud, 2019). Tantôt créées par une machine (ex. données de connexion à un site Internet), tantôt par l'individu (ex. nom, prénom et âge), les données structurées sont par conséquent faciles à saisir, à stocker et à analyser. L'exemple d'une base de données de clients illustre parfaitement le concept de données structurées. En effet, une telle base de données se compose d'une multitude d'informations telles que nom et prénom, date de naissance ou encore adresse de courrier électronique. Si l'on reprend l'argumentation de Marr (2015), toutes ces informations sont bel et bien définies par un champ, c'est-à-dire le titre de la colonne.

Cela étant, toutes les données ne se rangent pas nécessairement dans un champ prédéfini. Marr (2015) estime que 80 % des données qui intéressent les entreprises proviennent d'informations non structurées ou semi-structurées. Par opposition aux données structurées :

Les données non structurées représentent toutes les informations qui ne peuvent pas être facilement classées dans des colonnes, des lignes et des champs. Elles sont généralement constituées de texte, mais peuvent également contenir des données telles que des dates, des chiffres et des faits, ou différents types de données comme des images. Ces incohérences les rendent difficiles à analyser à l'aide de programmes informatiques traditionnels. (Marr, 2015, p. 61)

Alwan & Ku-Mahamud (2019) listent une série d'exemples de données non structurées. Entre autres, celles-ci comprennent les images satellites, les données scientifiques, les photos ainsi que les fichiers vidéo, mais également les textes dans les médias sociaux et sites Internet.

Finalement, les données semi-structurées renvoient aux informations qui ne se trouvent ni dans la classe des données structurées, ni dans celle des données non structurées. Elles sont en réalité un mélange de données non structurées et structurées (Alwan & Ku-Mahamud, 2019). Ce sont des données qui peuvent avoir une certaine structure pouvant être utilisée pour l'analyse, mais qui n'ont pas la structure stricte d'un modèle de données. En d'autres termes, les données n'ont pas de structure fixe. Notamment, une publication sur Facebook ou Twitter peut être caractérisée par son auteur, ses données ou sa longueur (i.e. données structurées), néanmoins le contenu ne sera la plupart du temps pas structuré.

Attention, il serait une erreur de considérer qu'un type de données prévaut sur un autre (Marr, 2015). Ainsi par exemple, il est faux de considérer que les *third-party data* ont plus de valeur que les *first-party data*. Toutes les données sont ainsi bonnes à prendre uniquement dans la mesure où elles ont un sens pour l'entreprise qui en fera usage, ou si celle-ci parvient à leur en donner.

2.2. Application du Big Data au marketing

2.2.1. Contribution du Big Data au marketing

Appliquons la définition du Big Data axée autour de ces trois *V's* au marketing. A l'heure actuelle, les entreprises et les fournisseurs de données tiers (*cf. infra*) suivent et gèrent des bases de données lourdement importantes sur leurs clients (*volume*), souvent accessibles en temps réel (*vitesse*), et offrant aux modèles marketing la possibilité d'adapter les stratégies aux consommateurs au fur et à mesure qu'ils recherchent des informations, comparent les prix ou effectuent des achats (Chintagunta, Hanssens & Hauser, 2016). Ces informations de plus en plus interconnectées se présentent entre autres sous la forme de données numériques, des fichiers texte, audio et vidéo, (*variété*).

Avec l'essor de la digitalisation, l'historique de navigation, les actes d'achat, l'utilisation des appareils mobiles, réseaux sociaux et bien d'autres encore, il est clair que le Big Data aura

permis aux entreprises, depuis quelques années maintenant, des choses autrefois totalement inconcevables. Cette révolution concerne avant tout la manière dont les entreprises commercialisent leurs produits auprès de leur clientèle. A titre d'exemple, les organisations sont aujourd'hui en mesure d'adapter leur communication à leur population cible, et ce à l'échelle très précise de l'individu. Le Big Data permet ainsi aux entreprises de se connecter avec leurs clients, d'enregistrer chaque clic détecté sur leur site Internet, conduisant à une compréhension extrêmement détaillée du comportement d'achat du consommateur (Sathi, 2014), transformant le traditionnel marketing de masse en une forme de marketing (hyper)personnalisé (Karoui *et al.*, 2014). De plus, autrefois confrontés à des échantillons de petite taille, les analystes marketing sont désormais en mesure de collecter une multitude d'informations qui concernent une plus grande étendue d'individus. Avec cette expansion de données, de nouvelles variables marketing ont vu le jour, permettant de mieux définir le besoin du consommateur (Sathi, 2014). Ajoutons à cela un ciblage plus performant, un meilleur retour sur investissement, une meilleure compréhension (parfois anticipation) des tendances commerciales, mais aussi l'avènement du marketing prédictif (Gianelloni & Le Nagard, 2016). En effet, les algorithmes de prédiction du comportement ne sont pas nouveaux, toutefois l'accès aux données facilité par l'émergence du Big Data permet de modéliser et d'anticiper les actions d'un consommateur sur la base de ses précédentes actions (Salesforce, 2021b). L'enjeu pour ces entreprises est alors de conserver tous ces événements et informations, ce afin de les exploiter pour pouvoir en tirer un maximum d'avantages. Néanmoins, afin d'être à même de supporter cette interaction des données en temps réel et d'exploiter ces sources de données que sont Internet, les appareils mobiles et autres objets connectés, ces entreprises n'ont pas d'autre choix que de repenser l'organisation de leurs systèmes d'information.

2.2.2. Données marketing

Dans la continuité de la section précédente dédiée aux contributions du Big Data au marketing, proposons maintenant une définition de *données marketing*. Parmi la multitude de données créées et accessibles directement ou indirectement chaque jour, certaines intéressent les entreprises plus que d'autres. Quelles sont-elles ? En quoi peuvent-elles être utilisées ?

Une donnée marketing est une information qui concerne avant tout un individu, qu'il soit consommateur ou potentiel client (i.e. prospect), son comportement (Bathelot, 2018a), ou ses

réseaux (i.e. réseaux sociaux et relations) (Lambin & de Moerloose, 2021). Chevalier (2020) met en évidence plusieurs types de données marketing : les données démographiques, comportementales, géographiques, psychographiques ou encore relationnelles. Les données démographiques sont des informations sur les caractéristiques de base des individus, telles que l'âge, le genre, le revenu ou le niveau d'éducation (Datarade, 2021). Aussi, les données comportementales sont les informations qui relatent du comportement de l'individu sur la toile, comme le temps passé sur un site Internet ou le nombre de clics sur une publication. Par ailleurs, les données géographiques quant à elles se rapportent aux informations de localisation de l'individu (Chevalier, 2020). Si les données psychographiques renseignent les préférences, les aversions, l'attitude, les traits de personnalité ou encore le sentiment d'un individu (Datarade, 2021), les données relationnelles quant à elles proviennent de « l'analyse des proximités entre réseaux » (Lambin & de Moerloose, 2021, p. 203).

A nouveau, ce sont ces données qui aident les organisations à identifier les tendances qui leur permettent de cibler leurs consommateurs, à répondre aux intentions, intérêts et besoins spécifiques du marché (HubSpot, 2021). Cependant, beaucoup de ces informations sont de l'ordre privé et ne peuvent être collectées et mobilisées sans autorisation de la part de leur propriétaire. En effet, les données des consommateurs sont aujourd'hui protégées par le Règlement général sur la protection des données (*cf. infra*).

Finalement, une base de données marketing comporte des informations de différents types, reprises dans des champs, relatives à chaque client ou consommateur. Attention, ces données peuvent également concerner des produits ou encore des données historiques de campagnes marketing précédemment déployées (Bathelot, 2016). Le plus souvent, celles-ci sont organisées et structurées, ce qui en facilite le traitement (*cf. supra*). Néanmoins, les bases de données actuelles peuvent parfois rassembler des enregistrements qui concernent plusieurs millions d'individus (Courtheoux, 2003), ce qui en complexifie la gestion de la qualité (*cf. infra*). Une fois analysées, ces banques de données serviront de fondement à la stratégie marketing d'une organisation, par exemple via le marketing prédictif³.

³ Le *marketing prédictif* fait appel au Big Data pour développer des prévisions relativement précises sur le comportement futur des clients, et ainsi déterminer quelles actions et stratégies marketing ont le plus de chances de réussir (Salesforce, 2021b).

2.2.3. First-, Second- et Third-party data

Ce point s'intéresse à l'origine des données marketing dont il est question dans la section précédente. En effet, Gertz et McGlashan (2014) confrontent principalement deux concepts : les *first-party data* (ou données propriétaires) aux *third-party data* (ou données de tiers). Ils définissent les premières comme suit. Les *first-party data* sont les données qu'une organisation génère par interaction directe avec ses consommateurs depuis ses propres ressources, CRM et base de données clients : nom, adresse, numéro de téléphone, données d'interaction, informations sur les produits achetés, etc. (Brosche & Kumar, 2014). Elle en est donc propriétaire (Gertz & McGlashan, 2014). Selon eux, ces informations sont plus que précieuses étant donné que les consommateurs concernés se sont déjà engagés, dès lors généralement enclins à se convertir⁴ auprès de l'entreprise. Celle-ci est ainsi en mesure d'identifier les pages qui ont été consultées par un individu, et où cette personne se situe dans le processus d'achat. Les *first-party data* sont ainsi utilisées à des fins d'analyse mais également de ciblage (Groth & Zawadzki, 2014). Elles n'entraînent aucun frais, au-delà du coût de leur collecte. Par ailleurs, plus la quantité de *first-party data* collectées est élevée, plus les organisations se soucieront de la sécurité de ces informations, en veillant à ce que leur utilisation soit conforme aux législations en matière de protection de la vie privée, mais aussi en veillant à ce que ces données ne soient pas exploitées par d'autres entreprises (Gertz & McGlashan, 2014).

Quant aux *third-party data*, celles-ci impliquent des informations qui concernent les utilisateurs, échangées entre autres via des plateformes de gestion des données (Groth & Zawadzki, 2014). Gertz et McGlashan (2014) complètent : ces données sont cette fois-ci à acheter auprès de vendeurs de données. De fait, les places de marché de données (ou *data marketplace*) offrent des possibilités de contact avec de nombreux fournisseurs de données (à l'instar de BlueKai, Exelate, Weborama ou encore Datalogix, qui sont des acteurs du domaine des *third-party data* – cf. Annexe 5, p. 80). Une fois accessibles, ces données peuvent être mobilisées lors d'une campagne marketing (Markarian, Park, & Grether, 2014). Par ailleurs, Gertz & McGlashan (2014) soulèvent quelques conditions relatives aux données de tiers. Par exemple, les fournisseurs de données doivent détailler avec précision la manière dont les

⁴ Au sens marketing, *se convertir* signifie accomplir une action sur un site Internet. Par exemple, effectuer un achat, remplir un formulaire, télécharger un document ou encore s'inscrire à une *newsletter*. Attention, un appel téléphonique ou une visite en magasin constitue également une conversion marketing (Bathelot, 2015).

données ont été recueillies. A nouveau, la collecte des informations doit impérativement se conformer aux lois sur la protection de la vie privée (*cf. infra*).

En bref, les *first-party data* sont des données qui concernent le public d'une organisation que celle-ci collecte directement, par opposition aux *third-party data*, qui sont des données qu'une entité tierce recueille auprès de plusieurs sources et met en vente (Lotame, 2019). A cela, Groth et Zawadzki (2014) mettent en avant l'importance croissante des *second-party data*, qui se caractérisent par une connexion directe entre le propriétaire des données et leur acheteur potentiel. En d'autres termes, il s'agit de données qu'une organisation collecte directement auprès de son public, qu'elle vend personnellement à une autre entreprise.

2.2.4. Règlement général sur la protection des données

Le Règlement général sur la protection des données (ci-après, RGPD) est un texte de loi en vigueur depuis le 25 mai 2018. Il s'applique à toute organisation, publique ou privée (quelle que soit sa taille), qui traite des données personnelles, qu'elle soit établie sur le territoire de l'Union européenne, ou qu'elle cible directement des résidents européens (Journal officiel de l'Union européenne, 2021). En réalité, ce règlement remplace la directive européenne de 1995⁵, relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données (Europa, 2021). Pour rappel, une directive européenne fixe des objectifs à tous les pays de l'Union européenne, laissant libre-cours à chacun d'entre eux dans la mise en place des mesures qui permettront d'atteindre ces objectifs (Europa, 2021). A l'inverse, un règlement européen constitue un acte législatif contraignant qui impose à chaque membre de l'Union européenne de mettre en œuvre son contenu dans son intégralité (Europa, 2021). Le RGPD a pour principale finalité de protéger les libertés et droits fondamentaux des individus (quelques exemples de ces droits sont fournis à l'Annexe 4, p. 79), et en particulier leur droit à la protection des données à caractère personnel (Europa, 2021), qu'il définit comme suit :

[Une donnée à caractère personnel désigne] toute information concernant une personne physique identifiée ou identifiable (personne concernée) ; est réputée

⁵ Directive 95/46/CE du Parlement européen et du Conseil du 24 octobre 1995 (Europa, 2021).

identifiable une personne qui peut être identifiée, directement ou indirectement, notamment par référence à un numéro d'identification ou à un ou plusieurs éléments spécifiques, propres à son identité physique, physiologique, psychique, économique, culturelle ou sociale. (RGPD, 2016, Article 4 p. 33)

Bien que déjà très contraignant, le document législatif autorise les États-membres à introduire des dispositions nationales supplémentaires destinées à préciser davantage l'application des règles qui y sont inscrites (RGPD, 2016). C'est ainsi que les Autorités de protection des données (anciennement Commissions de la protection de la vie privée) ont vu le jour. Organismes de contrôle indépendants, ils s'assurent du bon respect de la législation relative à la protection des données, fournissent des conseils d'experts sur les questions liées à la protection des données, et traitent les réclamations émises suite à des violations des principes du RGPD (Autorité de protection des données, 2021). En Belgique, cette Autorité a été fondée au sein de la Chambre des représentants par la loi du 3 décembre 2017 (Autorité de protection des données, 2021).

2.3. Qualité des données

Au-delà des enjeux technologiques, techniques et organisationnels relevés précédemment, il est un autre enjeu incontournable pour les organisations : la qualité des données. Avec la démultiplication des sources et l'essor des volumes de données disponibles, la qualité de l'information a pris une place de premier plan au sein des entreprises. Commercialisation de nouveaux produits, ajustement au marché impliquent transformation des données, faisant de l'analyse de la qualité des données une étape indispensable (Di Ruocco, Scheiwiler, & Sotnykova, 2012), dans la mesure où les données recueillies constitueront le fondement des décisions (notamment marketing et stratégiques) prise par une organisation. Lorsqu'elle est assurée, la qualité de la donnée peut alors devenir source d'un avantage concurrentiel (Mahanti, 2018).

Afin de saisir toute l'importance de la qualité des données, il faut avant tout en comprendre le sens. Qu'est-ce qu'une donnée de qualité ? Quelles en sont les caractéristiques ? Quels peuvent être les impacts si la qualité de la donnée n'est pas assurée ? Comment entretenir et optimiser la qualité de cette information ?

2.3.1. Définition et caractéristiques de la qualité des données

Une donnée est dite de qualité si celle-ci répond aux attentes des acteurs qui en font usage (Brasseur, 2005). Elle reste essentielle si une entreprise souhaite maximiser l'efficacité de sa base de données marketing ainsi que de la qualité des décisions qui en découlent. Dès lors que des informations inexactes entrent dans la base de données marketing, certains rapports, analyses, et modèles prédictifs ultérieurs seront corrompus (Coussement, Van den Bossche, & De Bock, 2014). Elle constitue ainsi un déterminant potentiel du succès d'une organisation. Toutefois, il est impossible d'atteindre une qualité parfaite des données, dans la mesure où celle-ci coûte cher, mais aussi et surtout car il n'est pas nécessaire qu'une donnée soit parfaite (Maydanchik, 2007). L'important est de s'assurer que celle-ci convienne à l'usage pour lequel elle s'apprête à être utilisée (Eckerson, 2002). Par ailleurs, quels sont les critères qui caractérisent une information de qualité ? La littérature met en avant une série d'indicateurs qui, lorsqu'ils sont remplis, définissent une donnée comme étant de qualité (Batini & Scannapieco, 2016 ; Di Ruocco *et al.*, 2012 ; Eckerson, 2002 ; Loshin, 2011 ; Mahanti, 2018 ; Zhang, Indulska, & Sadiq, 2019).

Complétude. La complétude est la plus fondamentale des caractéristiques de la qualité des données. Elle représente le degré auquel toutes les données attendues sont présentes (Loshin, 2011), et évalue donc la présence ou l'absence d'enregistrements dans la base de données (Mahanti, 2018). L'exhaustivité des données couvre trois sous-caractéristiques (Di Ruocco *et al.*, 2012) : la complétude des entités (ex. entité client ou commande), des attributs (ex. nom, prénom ou montant de la commande), et des occurrences (ex. aucun client ne peut être omis dans la base de données, tout comme aucune commande).

Pertinence. La pertinence peut être définie comme la capacité des données à répondre à l'objectif pour lequel elles doivent être utilisées (Batini & Scannapieco, 2016), ainsi que leur capacité à répondre aux besoins actuels et futurs des utilisateurs (Loshin, 2011). En ce qui concerne les besoins actuels, il s'agit de se concentrer uniquement sur les informations utiles ; quant aux besoins futurs, il s'agit de s'assurer que ces informations soient modulables dans le temps. Par exemple, une année de naissance. Dans les anciennes bases de données, les années sont représentées sur deux positions, plutôt que quatre actuellement (ex. [98] au lieu de [1998]). Or, le passage au second millénaire induit de potentiels conflits entre certains attributs : [1920] ne peut plus être représenté par [20] car, dans la même logique, [2020] porterait le même attribut (Di Ruocco *et al.*, 2012). La pertinence est ainsi une caractéristique très importante : si les données ne sont pas pertinentes, leur valeur diminue considérablement, même si celles-ci semblent exactes, valides, et précises (Mahanti, 2018).

Exactitude & Précision (Accuracy). La précision désigne la mesure dans laquelle les données stockées dans un système reflètent la réalité (Loshin, 2011). En d'autres termes, il s'agit du degré avec lequel les données représentent correctement un objet, une entité, une situation, un phénomène ou encore un événement du monde réel qui est décrit (Batini & Scannapieco, 2016). La précision fait donc référence au degré de détail d'une information. Ainsi, l'orthographe incorrecte des noms de produits ou de clients et de leur adresse postale impactera fortement l'utilisation des données qui en sera faite. Attention, une donnée peut être précisément exprimée, mais s'avérer inexacte (Mahanti, 2018).

Conformité. Cette caractéristique renvoie aux différentes normes standards ou lignes directrices, notamment les métadonnées, qui documentent les données. Nous définirons une métadonnée comme toute information qui permet de comprendre les données elles-mêmes. En

d'autres termes, elle renvoie aux données sur les données. Celles-ci sont donc comparables à des étiquettes structurées d'informations qui caractérisent d'autres données et en facilitent la compréhension, la gestion, et l'utilisation (Furner, 2020). Par exemple, le libellé d'une colonne d'un tableau de données, la longueur ou la définition d'un champ d'information (Mahanti, 2018). Les métadonnées apportent ainsi de la valeur supplémentaire aux données qu'elles décrivent, et elles sont importantes pour les exploiter correctement. Comparer les données à leurs métadonnées permet ainsi d'en mesurer le degré de conformité (Di Ruocco *et al.*, 2012).

Consistance & Cohérence (Consistency). On dit qu'il y a consistance lorsque, pour une entité (ex. un consommateur), les valeurs des données sont parfaitement identiques dans l'ensemble de l'environnement de données (Di Ruocco *et al.*, 2012) : la consistance suppose donc que les données de toute l'entreprise soient synchronisées entre elles. Le format et la présentation des données doivent ainsi être cohérents dans l'ensemble des environnements de données (Batini & Scannapieco, 2016). A nouveau, la consistance des données n'implique pas forcément que celles-ci soient précises et correctes (Mahanti, 2018). Enfin, une mise à jour de données dans un environnement précis de la base de données supposera, pour en assurer la cohérence, que l'information soit corrigée dans l'intégralité des systèmes de données. La cohérence des données quant à elle garantit l'absence de conflits potentiels entre les informations (Mahanti, 2018).

Récence. Cette caractéristique de la qualité des données cherche à déterminer si ces dernières sont exactes vis-à-vis du temps qu'elles représentent (Di Ruocco *et al.*, 2012). En d'autres termes, il est question de déterminer si les données sont temporellement pertinentes (autrement dit, à jour). Une donnée de qualité est donc une donnée dont la valeur est temporellement valable. Cette caractéristique suppose de s'intéresser à la fréquence à laquelle les données varient dans le temps (Mahanti, 2018).

Unicité. L'absence de doublons dans un environnement de données est assurée par la caractéristique de l'unicité des données (Loshin, 2011). A cela, Mahanti (2018) associe la notion de cardinalité, qui renvoie au degré d'unicité des valeurs présentes dans une base de données. Par exemple, un tableau de données à haute cardinalité est un tableau qui contient une large quantité de données uniques (Mahanti, 2018). Pensons au numéro de registre national ou encore un numéro de téléphone d'un consommateur.

Accessibilité & Disponibilité. Cette caractéristique fait référence à la facilité d'accès aux données par les utilisateurs (Di Ruocco *et al.*, 2012). Des données accessibles désignent également des données simples à comprendre, analyser et utiliser (Loshin, 2011). Cela suppose un accès direct aux métadonnées (*cf. supra*) et autres supports nécessaires à la compréhension des données (Mahanti, 2018).

Intégrité. L'intégrité des données renvoie aux relations établies entre des entités de données (ex. entité client ou commande), ainsi qu'à la validité des données au travers de ces relations (Mahanti, 2018). Ce critère se caractérise par la cardinalité d'une relation, qui représente le nombre de fois qu'une entité peut être liée à une autre. La littérature distingue trois types de cardinalité : *one-to-one* (une entité ne peut être liée à une autre qu'à une seule reprise ; ex. ID client unique), *one-to-many* (une entité ne peut être liée à une autre qu'à une seule reprise, qui elle, peut être liée à plusieurs objets de la première entité ; ex. date de naissance), et *many-to-many* (une entité peut être liée à une autre indéfiniment ; ex. un produit peut être acheté par plusieurs consommateurs, de la même manière qu'un consommateur peut acheter plusieurs produits) (Sebastian-Coleman, 2013 as cited in Mahanti, 2018).

A titre de synthèse, le tableau 2.3.1.1 ci-dessous reprend les différents critères couverts ci-avant auxquels sont associées d'une part une série de sous-items qui définissent chacun d'entre eux (première colonne), et d'autre part, une série de questions à se poser pour en évaluer le degré (deuxième colonne).

Macro-critère	Sous-critères	Questions à se poser
<i>Complétude</i>	- Présence de tous les éléments <u>essentiels</u> - Degré de complétude requis	- Toutes les données nécessaires sont-elles présentes ? - Des valeurs essentielles manquent-elles dans les enregistrements de données ? - Tous les éléments de données obligatoires sont-ils enregistrés ?
<i>Pertinence</i>	- Capacité de répondre à l'objectif visé par l'utilisation des données - Profondeur et étendue de données au regard de l'objectif visé	- Les données répondent-elles aux besoins des utilisateurs ? - Les données ont-elles la profondeur et l'étendue suffisantes pour l'objectif visé ?
<i>Exactitude & Précision (Accuracy)</i>	- Représentation précise de la réalité - Niveau de détail requis	- Les données représentent-elles avec précision la réalité ? - Les données sont-elles stockées avec le niveau de détail requis par l'entreprise ?
<i>Conformité</i>	- Formatage de données - Standardisation de données - Correspondance avec les métadonnées	- Les valeurs des données doivent-elles se conformer à des formats spécifiques ? - Si oui, toutes les valeurs doivent-elles être conformes à ces formats ? - Les données correspondent-elles à leurs métadonnées ? - Quelles sont les données qui ne sont pas stockées dans un format standard ?
<i>Consistance & Cohérence (Consistency)</i>	- Homogénéité des représentations dans l'environnement de données - Absence de contradictions	- Des sources de données différentes fournissent-elles des informations contradictoires ou différentes concernant une même entité ? - Les données sont-elles stockées dans le même format dans différents tableaux et systèmes de données ? - Les valeurs sont-elles uniformes dans tous les ensembles de données ?
<i>Récence</i>	- Actualisation de données - Fréquence de mise à jour des données	- Les données sont-elles suffisamment à jour ? - Dans quelle mesure les données sont-elles à jour au moment de leur utilisation ? - La fréquence de mise à jour des données est-elle acceptable pour répondre aux attentes de l'entreprise ?
<i>Unicité</i>	- Représentation unique des données - Absence de répétitions	- Un ensemble de données comporte-t-il des répétitions des mêmes objets de données ? - Des éléments individuels de données de l'ensemble de données sont-ils enregistrés plus d'une fois ? - Y a-t-il de nombreuses représentations inutiles des mêmes objets de données dans un ensemble de données ? - Quels enregistrements de données sont répétés ?
<i>Accessibilité & Disponibilité</i>	- Facilité d'accès - Facilité de compréhension - Facilité d'utilisation - Disponibilité de la documentation sur les données	- Les données sont-elles facilement accessibles, compréhensibles et utilisables ? - Si les données sont disponibles, comment peut-on y accéder ? - Est-il facile de comprendre et d'interpréter les données ? - Des informations supplémentaires, des métadonnées et de la documentation sont-elles disponibles ?
<i>Intégrité</i>	- Présence de tous les liens relationnels essentiels - Validité des liens relationnels dans l'environnement de données - Cohérence de l'évolution des relations entre entités de données	- Quelles sont les données qui manquent de liens relationnels essentiels ? - Les liens relationnels entre les ensembles de données sont-ils des liens valides ? - La structure des données et les relations entre les entités et les attributs sont-elles maintenues de manière cohérente ?

Tableau 2.3.1.1. Qualité des données : macro-critère, sous-dimensions et questions à se poser pour en mesurer le degré

(Batini & Scannapieco, 2016 ; Di Ruocco *et al.*, 2012 ; Eckerson, 2002 ; Loshin, 2011 ; Mahanti, 2018 ; Zhang *et al.*, 2019)

2.3.2. Impacts de la qualité des données

i. Impacts d'une faible qualité de données

Une faible qualité des données peut engendrer de lourdes conséquences. Assimilée à un cancer (Mahanti, 2018), celle-ci peut facilement affaiblir la performance d'une entreprise, voire en provoquer la chute. Avec une qualité faible de données, Eckerson (2002) estime qu'une entreprise n'est pas en mesure de prendre des décisions parfaitement raisonnées. Selon lui, elle ne serait pas en mesure de comprendre ce qui se passe en son sein ou sur le marché, en ne se fiant que sur son simple instinct. Néanmoins, l'évolution rapide et toujours constante du marché rend cette manière de fonctionner très dangereuse, à la fois avec des concurrents agiles et des clients exigeants. Lorsqu'elle n'est pas assurée, la qualité des données peut aussi être la source d'insatisfaction de consommateurs, parfois de leur perte, mais aussi dévaloriser l'image que dégage la marque ou l'entreprise (Di Ruocco *et al.*, 2012). En effet, une entreprise qui perd ses clients fidèles perd également des possibilités de revenus futurs (Eckerson, 2002). Par ailleurs, elle peut engendrer une série de coûts supplémentaires inutiles, pouvant représenter jusqu'à 25 % du budget annuel total d'une organisation (Mahanti, 2018) : perte de temps, coûts opérationnels, perte de revenus, amendes pour non-conformité, etc. Dans son rapport, Eckerson (2002) évalue les coûts totaux annuels des entreprises américaines induits par une faible qualité des données clients à près de 611 milliards de dollars. Il affirme également que 40 % des entreprises ne parviennent pas à atteindre les résultats espérés en raison d'une faible qualité des données. Une organisation peut ainsi investir dans un système de gestion de la relation client (CRM), toutefois si la qualité des données qui s'y trouvent n'est pas considérée comme elle le devrait, le retour sur cet investissement ne serait que marginal.

La figure 2.3.2.1 ci-dessous illustre les résultats d'une enquête menée en 2001 par *The Data Warehousing Institute* (ci-après TDWI) qui s'intéresse aux répercussions que peut amener une faible qualité des données (Eckerson, 2002). Ainsi par exemple, 87 % des participants à ce sondage estiment qu'une faible qualité des données entraîne du temps supplémentaire de travail. De la même manière, 67 % d'entre eux attestent que lorsque la qualité des données est faible, le taux de satisfaction des clients diminue lui aussi.

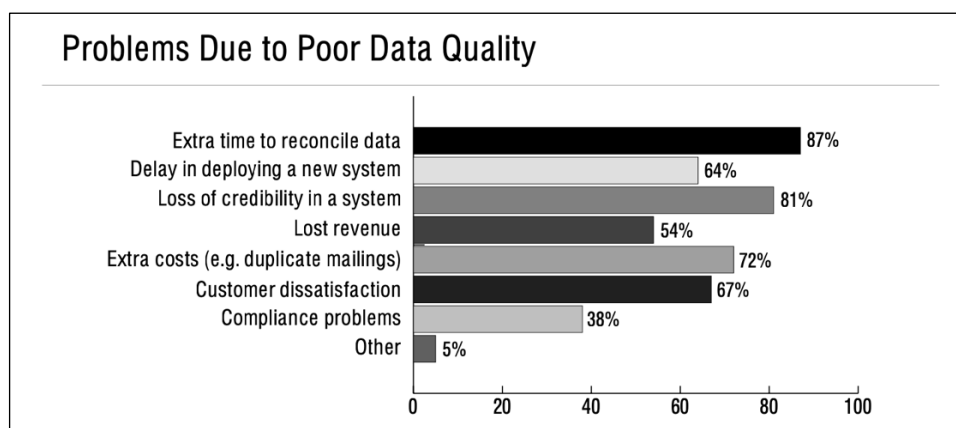


Figure 2.3.2.1. Impacts d'une faible qualité des données. Enquête auprès de 286 répondants, plusieurs réponses possibles au questionnaire (Eckerson, 2002).

En des termes marketing, une faible qualité de données impacte entre autres la capacité d'une organisation à segmenter ses clients et à personnaliser son offre (Mahanti, 2018). Par exemple, des données de mauvaise qualité qui concernent des clients contribuent à gaspiller des sommes d'argent parfois considérables dans des campagnes de vente et de communication dès lors mal ciblées. De la même manière, le manque d'informations précises et complètes sur la relation qu'un consommateur entretient avec une marque ou une organisation (en se fondant par exemple sur les produits qu'il a autrefois achetés, ou de sa propension à acheter d'autres produits) peut nuire aux futures opportunités de vente ou de marketing manquées (Mahanti, 2018). Par ailleurs, une grande partie des programmes de fidélité ont souffert de la mauvaise qualité des données. Ainsi, plus de 70 % des organisations gérant ces programmes ont signalé des incidents liés à des informations clients inexactes (Adamson, 2014, cité dans Mahanti, 2018).

ii. Impacts d'une haute qualité de données

A l'inverse, une entreprise qui accorde beaucoup d'importance à la qualité des données et se donne les moyens pour l'assurer, mais aussi l'entretenir et l'améliorer continuellement, connaîtra les effets inverses que ceux répertoriés ci-avant (Even & Shankaranarayanan, 2009) : gain de temps, prise de décision fiabilisée, augmentation de la satisfaction des clients, réduction des coûts, augmentation des revenus (ventes), création d'avantages concurrentiels supplémentaires, etc. ; en des termes marketing, fiabilisation de la prise de décision stratégique, segmentation encore plus personnalisée, meilleure analyse des données, etc. (Samitsch, 2015). La figure 2.3.2.2 ci-après recense différents avantages qu'apporte une haute qualité des données

soulignés par les répondants de l'enquête du TDWI (*cf. supra*). Ainsi, 13 % d'entre eux confirment une réduction des coûts, tout comme 17 % affirment faire plus facilement confiance aux analyses effectuées sur des données de qualité (Eckerson, 2002).

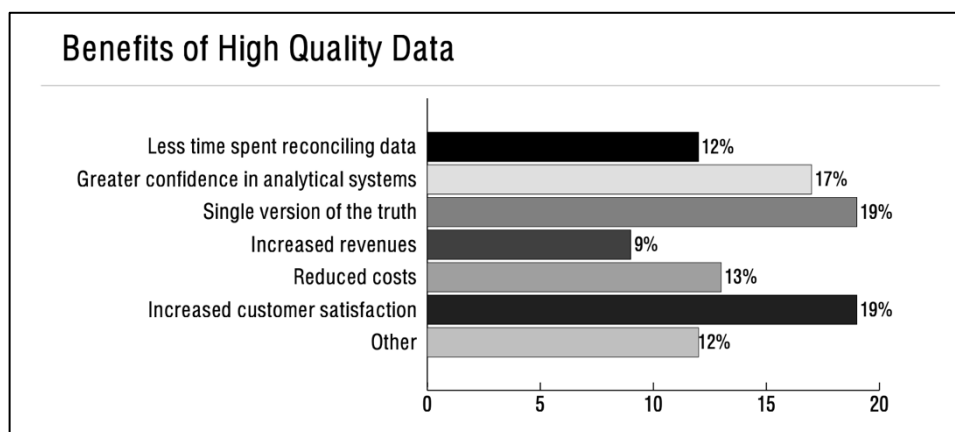


Figure 2.3.2.2. Impacts d'une haute qualité des données. Enquête auprès de 304 répondants, une seule réponse possible au questionnaire (Eckerson, 2002).

2.3.3. Types d'anomalies

Doublons, valeurs manquantes, données inexactes ou obsolètes, informations inutiles ou encore données non standardisées constituent la majorité des anomalies rencontrées dans les bases de données, souvent résultant d'erreurs humaines (Di Ruocco *et al.*, 2012) (*cf.* Figure 2.3.3.1). Citons l'exemple d'un formulaire à remplir sur Internet. Il arrive qu'une information erronée ou incomplète soit renseignée, ou que celle-ci soit correctement renseignée mais au mauvais endroit du formulaire. Ces erreurs s'expliquent souvent par la distraction, le manque de temps ou d'information, ou encore le souci de préserver sa vie privée. En effet, certaines personnes communiquent de fausses adresses électroniques ou numéros de téléphone pour éviter les publicités ou autres communications indésirables (Mahanti, 2018). Une étude menée par le spécialiste en cyber-sécurité américain Symantec (2015) montre d'ailleurs que près d'un consommateur européen sur trois admet avoir fourni de fausses informations en ligne afin de protéger sa vie privée.

En outre, certaines de ces anomalies proviennent du faible contrôle de validité des informations au moment de la saisie (Mahanti, 2018). D'une part, le contrôle du champ dans lequel doit être renseignée l'information. Par exemple, un champ dans lequel doit être inséré un numéro de

téléphone ne peut contenir que des caractères numériques. D'autre part, le contrôle lié à la donnée qui est renseignée. Par exemple, s'assurer que l'information renseignée correspond bien à un numéro de téléphone belge (c'est-à-dire qu'il s'étend bien sur dix positions, valeur de l'indicatif [+32]).

Par ailleurs, les anomalies peuvent provenir de dysfonctionnements et transactions au sein même des systèmes d'information, comme la migration des données d'un environnement vers une autre, les mises à jour informatiques ou encore la purge automatique des données (Gudivada, Apon, & Ding, 2017). Ce dernier processus supprime définitivement des informations d'un espace de stockage ou d'une base de données considérées comme obsolètes (*cf. infra*), ce afin de libérer de l'espace de stockage pour de nouveaux enregistrements. Cette opération est effectuée lorsque la limite de conservation des anciennes données est atteinte. Cela étant, au vu de l'immense volume de données créées aujourd'hui, cette procédure est la plupart du temps automatisée. De ce fait, la purge des données peut parfois poser problème en s'en prenant par exemple à des données toujours et encore pertinentes, ou en purgeant plus (ou moins) de données que prévu. Ces erreurs de suppression ont de lourdes conséquences sur la qualité des données (Loshin, 2011). Ainsi par exemple, supposons que la purge consiste à éliminer toutes les informations qui concernent des clients dont le dernier achat en ligne excède une période de deux ans. Dans le cas où la dernière date d'achat d'un consommateur est faussée dans la base de données (ex. trois années sans acte d'achat), il y a de grandes chances pour que les données qui le concernent soient effacées. Autrement dit, l'entreprise perdrait l'intégralité des données de cet individu (alors que celui-ci s'avérerait un client fidèle) et ainsi tout moyen d'interagir avec lui ultérieurement.

Finalement, il arrive que les contextes de changement organisationnel expliquent la dégradation de la qualité des données. En effet, les fusions et acquisitions d'entreprises par exemple impliquent la consolidation des données (Mahanti, 2018), qui peuvent amener à créer des valeurs en double (affectant ainsi la caractéristique de l'unicité) ou encore des inconsistances. Il en va de même pour les réorganisations (Eckerson, 2002), au cours desquelles des utilisateurs de données se voient contraints de quitter l'entreprise, emportant avec eux toutes leurs connaissances (Loshin, 2011). La qualité des données requiert ainsi une excellente gestion des savoirs.

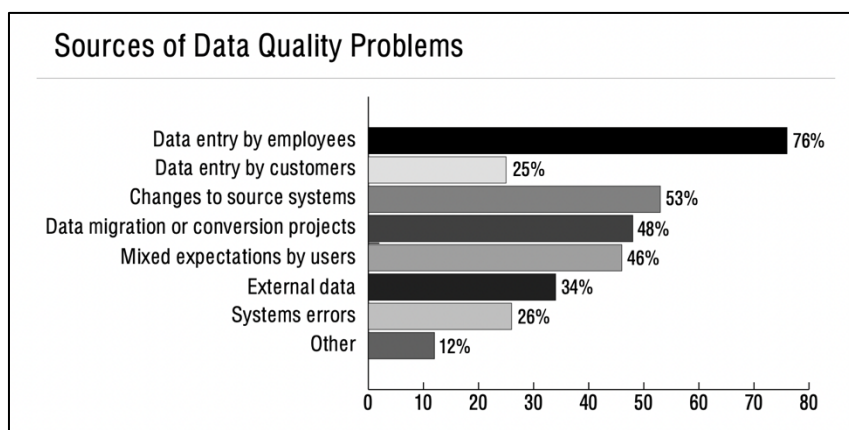


Figure 2.3.3.1 Quelques exemples des sources de problèmes liés à la qualité des données (Eckerson, 2002).

Illustrons ces différents cas à l'aide du tableau 2.3.3.2 ci-dessous, qui regroupe un très petit échantillon (simulé) de consommateurs d'une marque spécialisée dans la fabrication de chaussures en cuir reconvertie dans la conception d'articles de maroquinerie. Ci-après plusieurs informations qui les concernent : leur numéro (unique) de carte de fidélité, leurs nom et prénom, date de naissance, domicile, pointure de chaussure, ainsi que la date à laquelle ils ont adhéré au programme de fidélité de la marque. Ces données s'apprêtent à être mobilisées dans le cadre d'une campagne de communication de la marque vers ses consommateurs.

	A	B	C	D	E	F	G
	ID Fidélité	Nom	Prénom	Naissance	Adresse postale	Pointure	Adhésion
1	17500067	Crespin	Benoit	22/06/1967	248, Albert Ier, Bruxelles	44	08/11/2006
2	43432673	Backer, De	Caroline	03/01/1984	68, rte des Vtrans, Namur	38	25/01/2013
3	1750067	Crespin	Benoit	22/06/1967	Boulevard Albert Ier, 248, BXL	44	08/11/2006
4	00009256	Liesen	Claire	06/02/1936		37	12/08/1989
5	54091472	De Backer	Véronique	07/10/1998	68, Route des Vétérans, Tournai	39	02/02/2015

Tableau 2.3.3.2. Anomalies. Exemple d'un échantillon de clients et de leurs données.

Doublons. Un doublon renvoie à deux ou plusieurs enregistrements créés dans la base de données mais qui représentent une même entité (Di Ruocco *et al.*, 2012). Dans le cas du marketing, une entité consommateur par exemple. Dans le tableau 2.3.3.2 repris ci-avant, les lignes 1 et 3 semblent décrire le même individu, avec des valeurs cependant différentes. Ce type d'anomalie pose souci si les informations devaient être modifiées : comment déterminer quelle ligne est-on supposé modifier ? Aussi, dans l'exemple de la campagne marketing, quelle ligne utiliser dans les phases de ciblage et de communication ?

Valeurs manquantes. Les informations manquantes renvoient à des informations qui ne figurent pas dans la base de données alors que celles-ci étaient pourtant attendues (Di Ruocco *et al.*, 2012). Elles peuvent également être sources de coûts inutiles (Courtheoux, 2003) : le temps utilisé pour pallier ces manquements, le travail supplémentaire à la fois de nettoyage et de synchronisation des données impactent ainsi la performance d'une campagne par exemple. Dans notre exemple, la ligne 4 ne contient pas d'adresse postale (colonne E). Dès lors, comment une organisation est-elle en mesure de communiquer vers ce consommateur ?

Données inexactes. D'une part, les informations inexactes renvoient à des données dépourvues de sens par rapport à la réalité. A nouveau, ce type d'anomalie impactera la performance d'une campagne. Dans l'exemple du tableau 2.3.3.2, supposons que le programme de fidélité de la marque requière que le consommateur soit âgé de 18 ans ou plus. En confrontant les dates de naissance (colonne D) et de fidélisation de la marque (colonne G) de la ligne 5, on constate après calcul que ce consommateur a adhéré au programme de fidélité à l'âge de 16 ans. Ainsi, même si des valeurs renseignées peuvent sembler correctes (les dates sont bien formatées et valides), une évaluation plus précise de ces données peut parfois prouver le contraire. D'autre part, les données inexactes correspondent à des données mal renseignées. L'exemple des lignes 1 et 3 le prouve à nouveau. Ces deux lignes semblent à première vue concerner un seul et même individu (mêmes nom et prénom, date de naissance, etc.). Toutefois, la donnée A3 s'étend sur sept positions (au lieu de huit). Dès lors, la colonne A (seule) donne l'impression qu'il s'agit de deux consommateurs différents. Or, il s'agit d'une simple erreur. Ce type d'anomalie peut être évité à l'aide d'un système de contrôle de la donnée avant importation dans la base de données (*cf. infra*).

Données obsolètes. L'obsolescence des données renvoie à des informations qui ne sont plus valables, qui sont temporellement dénuées de sens, autrement dit périmées. Une donnée obsolète est ainsi une information qui a vieilli et, par conséquent, dont la qualité a diminué (Schwab, 2016). Si des informations comme la date de naissance ou le numéro de registre national ne changeront jamais, d'autres données peuvent quant à elles être mises à jour : domicile, numéro de téléphone, adresse de courrier électronique, etc. La difficulté provient surtout de la large variété des données, qui fait que celles-ci périssent à différentes fréquences. Ainsi, on peut supposer qu'un changement de nationalité se fera moins souvent qu'un changement de numéro de téléphone. De la même manière, la date de décès d'un individu ne sera mise à jour qu'une seule fois (Mahanti, 2018), étant donné que l'on ne décède qu'à une seule reprise. Dans le cas d'une base de données marketing, la durée de vie des informations aura tendance à se détériorer plus rapidement (Inmon, 2009, as cited in Mahanti, 2018) : jusqu'à 2 % des enregistrements d'une base de données qui concernent des clients périssent au bout d'un mois, soit presque un quart de la base de données à l'échelle annuelle, car ces individus meurent, se marient, se séparent, ou déménagent. (Eckerson, 2002). Ainsi, en supposant qu'une base de données clients rassemble quatre millions d'enregistrements de contacts clients, près d'un million de ces données périssent chaque année. En d'autres termes, il s'agit d'un million d'individus qu'une organisation aura tenté de contacter au bout d'un an, en vain. C'est à imaginer les dépenses (dès lors inutiles) qu'auront induites ces communications... Si l'on observe le tableau 2.3.3.2, le consommateur de la ligne 4 est aujourd'hui âgé de 85 ans. Or, rien ne confirme ou n'infirme que cet individu est (ou n'est pas) décédé à l'heure actuelle. Dans l'hypothèse où cette personne a bel et bien disparu, les informations de la ligne 4 deviennent alors obsolètes.

Informations inutiles. Ces données renvoient aux informations qui ont été autrefois collectées, stockées dans des bases de données, mais qui ne sont pas exploitées dans le cadre des activités d'une entreprise. Ces données n'impactent pas la performance d'une campagne marketing, étant donné qu'elles ne sont pas utilisées. En revanche, elles restent de grandes consommatrices d'espace de stockage de l'information (Di Ruocco *et al.*, 2012). Dans le tableau 2.3.3.2, les données relatives à la peinture de chaussure de chaque consommateur (colonne F) n'ont aucun sens dans l'élaboration de la campagne marketing. Ces données étaient nécessaires lorsque l'entreprise était active dans l'industrie du chaussant. Néanmoins à l'heure actuelle, elles ne

sont plus d'aucune utilité, la marque étant reconvertie depuis dans la fabrication d'articles de maroquinerie.

Valeurs non standardisées. La standardisation des données renvoie à la manière dont ces informations sont représentées dans la base de données (Di Ruocco *et al.*, 2012). Une erreur dans la standardisation des valeurs peut amener à une interprétation incorrecte du contenu de la base de données (Di Ruocco *et al.*, 2012). Ainsi, deux enregistrements décrits de manière différente peuvent faire croire qu'ils sont différents, alors qu'ils sont pourtant équivalents. Par exemple, les données B2 et B5 du tableau 2.3.3.2 décrivent le même nom de famille, pourtant représenté de deux manières. Il en va de même pour les données E1 et E3, qui décrivent une adresse postale identique mais exprimée différemment. En bref, il est important de pouvoir faire les rapprochements nécessaires à la standardisation des informations pour ne pas considérer les enregistrements comme distincts mais bien identiques.

En y regardant de plus près, les anomalies relevées ci-avant peuvent être réparties selon le critère qu'elles entravent. Ainsi par exemple, un doublon est une anomalie qui va à l'encontre des principes défendus par le critère d'unicité d'une donnée de qualité. De la même manière, des informations obsolètes remettent en cause le critère de la récence. Enfin, des valeurs non standardisées remettent en question à la fois le critère lié à la consistance et la cohérence (*Consistency*) des données étant donné que les données ne sont pas toujours représentées identiquement (tantôt un numéro de téléphone portable sur 10 positions, tantôt sur 13). Par ailleurs, cette anomalie concerne également le critère lié à l'accessibilité et la disponibilité des données. En effet, des valeurs non standardisées sont plus difficilement accessibles, compréhensibles et utilisables. Le tableau 2.3.3.3 ci-dessous fait état de ces dernières observations en mettant en relation ces anomalies avec chacun des critères de qualité relevés.

Anomalie	Critère(s) concerné(s)
<i>Doublons</i>	Unicité
<i>Valeurs manquantes</i>	Complétude
<i>Données inexactes</i>	Exactitude et précision (<i>Accuracy</i>)
<i>Données obsolètes</i>	Récence
<i>Informations inutiles</i>	▸ Complétude ▸ Pertinence
<i>Valeurs non standardisées</i>	▸ Conformité ▸ Consistance et cohérence (<i>Consistency</i>) ▸ Accessibilité et Disponibilité

Tableau 2.3.3.3. Répartition des anomalies par critère de qualité.

Les anomalies rencontrées dans une base de données peuvent ainsi prendre différentes formes. Il s'agit maintenant de les identifier. Les volumes actuels de données disponibles rendent un examen manuel des données pratiquement et économiquement irréalisable : le signalement et l'identification des anomalies doivent ainsi être hautement automatisés (Courtheoux, 2003), ce afin d'orienter les personnes qui seront en charge de les corriger. Idéalement, ces contrôles doivent être initiés dès l'instant où la base de données est mise à jour (qu'il s'agisse d'un nouvel enregistrement ou bien d'une modification de donnée existante) ; à défaut, à intervalles réguliers. Néanmoins, il est bon de savoir que mesurer la qualité des données mais aussi l'entretenir requièrent des ressources financières importantes (Even, Shankaranarayanan, & Berger, 2010) : tendre vers une qualité élevée pour un grand nombre d'informations implique souvent des coûts technologiques élevés (Even, Shankaranarayanan, & Berger, 2007). Par exemple, investir dans des logiciels d'automatisation de contrôle de base de données. Les dépenses et efforts liés au manque d'insistance sur la qualité de la donnée dépassent de loin les coûts de la recherche de cette qualité, pouvant s'élever de 20 jusqu'à 35 % des recettes opérationnelles d'une organisation (English, 2009). Ainsi par exemple, les coûts associés à une sélection erronée de noms pour une campagne de marketing ciblée auront tendance à dépasser les coûts d'un logiciel de gestion de la qualité (Courtheoux, 2003).

Il existe également des mécanismes de contrôle des données *avant* importation dans la base de données. Cependant, si les données ont déjà été chargées, il est difficile d'identifier rapidement les informations suspectes sans logiciel, puisque l'importation a déjà été effectuée. Dans ce cas, la meilleure solution restera de faire appel à un logiciel de contrôle de la donnée (*cf. supra*). A l'inverse, rejeter toutes les données suspectes n'est pas une option intéressante, car cela réduirait

le nombre d'informations recueillies, et impacterait également la qualité du contenu de la base de données (Courtheoux, 2003).

Par ailleurs, le contrôle de la qualité des données suppose également une vigilance considérable aux valeurs non reconnues ou valeurs qui n'ont jamais été rencontrées par le passé (Courtheoux, 2003). Il est également conseillé de mettre en évidence les valeurs qui sont fréquemment apparues auparavant et qui, subitement, n'apparaissent plus dans la base de données. Courtheoux (2003) insiste en outre sur l'importance des contrôles de la logique de la donnée enregistrée. Par exemple, le montant total des achats d'un consommateur ne peut qu'augmenter au fur et à mesure du temps (l'information deviendrait alors insensée si ce montant diminuait) ; aussi, la date de la transaction la plus récente ne peut qu'être postérieure à la date qui concerne le dernier acte d'achat enregistré.

Ainsi, un ensemble de vérifications strictes basées sur des règles de contrôle constituent une défense extrêmement importante (et efficace) contre les erreurs de manipulation des données marketing non détectées impactant de près ou de loin la qualité des données.

2.3.4. Gérer et améliorer la qualité des données

Si contrôler la qualité des données revient à identifier les possibles erreurs dans une base de données, l'améliorer continuellement est aussi un élément primordial. D'après Mahanti (2018), si ce n'est pas déjà chose faite, la première chose qu'une entreprise a à faire est de construire une stratégie de qualité des données si celle-ci souhaite éviter les incohérences et le manque de fiabilité dans ses systèmes de données. En effet, sans directives liées à la qualité des données, une organisation prend le risque de laisser aux utilisateurs des données de faire appel à leurs propres méthodes pour utiliser, manipuler et gérer les données, en négligeant des procédures fiables (Mahanti, 2018). Une stratégie de qualité des données permet effectivement à une entreprise d'une part d'identifier les objectifs qu'elle veut atteindre, mais aussi de se donner les moyens pour y parvenir. Attention, cette stratégie ne doit pas être considérée comme un projet à long terme mais comme une habitude à faire perdurer au sein d'une entreprise (Eckerson, 2002). Pour cela, la littérature suggère aux plus grandes entreprises de repenser dans la mesure du possible la structure de leur organisation en y intégrant un département entièrement dédié à la maintenance et l'amélioration de la qualité des données, sous la responsabilité d'un *Chief*

Data Officer ; dans le cas des plus petites organisations, il est conseillé de responsabiliser un ou plusieurs individus déjà en place (Batini & Scannapieco 2016 ; Eckerson, 2002 ; Mahanti, 2018). Dès lors, en quoi consiste ce travail ?

i. Processus DMAIC

En guise de réponse, nous retiendrons surtout le processus DMAIC développé par Mahanti (2018), qui résume parfaitement les multiples réponses apportées par la littérature (Batini & Scannapieco 2016 ; Berti-Equille, 2004 ; Courtheoux, 2003 ; Eckerson, 2002 ; Mahanti, 2018). Ce processus se subdivise en cinq phases, qui sont Définir, Mesurer, Aanalyser, Améliorer (Improve) et Contrôler (cf. Figure 2.3.4.1).

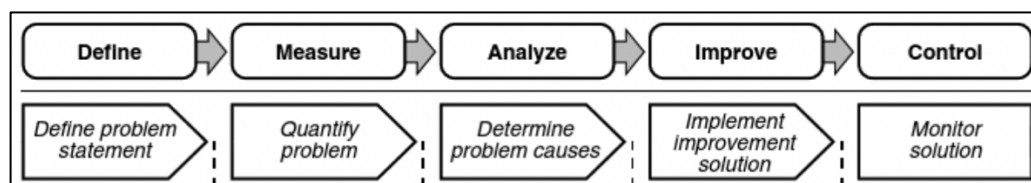


Figure 2.3.4.1. Processus DMAIC (Mahanti, 2018).

Définir. Cette première étape du processus consiste avant tout de prendre conscience de l'existence d'un problème. Puis, il s'agira de définir clairement l'énoncé du problème rencontré, en identifiant les anomalies et leur.s impact.s sur la base de données ainsi que les processus métier concernés (Mahanti, 2018).

Mesurer. Une fois le problème identifié, la deuxième phase permet de mesurer et comprendre l'état des données liées au problème identifié dans la première phase. Comment ? Par exemple, en évaluant les métadonnées (cf. *supra*) ou en procédant au profilage des données (*data profiling*). Cette dernière technique consiste en de multiples analyses qui étudient la structure et le contenu des données (IBM, 2016), dont les résultats seront observés dans la phase suivante (A). Par ailleurs, cette phase de mesure s'effectuera en fonction du seuil de confiance et du degré de conformité définis au préalable par l'organisation (ex. convenir que 85 % des objets de la base de données doivent être associés à une adresse de courrier électronique) (Eckerson, 2002 ; Mahanti, 2018).

Analyser. Cette troisième phase permet d’observer les résultats des différentes analyses menées au cours de la seconde phase, autrement dit d’identifier les anomalies, leur risque, leur coût et le nombre de fois qu’elles surviennent dans la base de données (Mahanti, 2018). Une fois les anomalies repérées et signalées, il est très important d’en comprendre la raison (Berti-Equille, 2017) et la source, mais aussi de les traiter. Enfin, il s’agira de proposer des pistes d’amélioration aux parties prenantes, d’en relever les avantages et inconvénients, d’en analyser les coûts, de sorte que cela ne se produise plus à l’avenir (Berti-Equille, 2004).

Améliorer. Une fois présentées et validées, c’est au travers de cette quatrième étape que les pistes d’amélioration seront concrètement mises en place. Selon les résultats des observations dans la phase précédente, la solution pourra prendre la forme d’une formation des collaborateurs, d’une modification des processus internes ou encore d’une amélioration technique des systèmes de données existants (Batini & Scannapieco 2016).

Contrôler. Cette dernière étape passe en revue la qualité des données suite à l’implémentation de la solution dans la phase précédente (Mahanti, 2018). Elle permettra d’identifier les problèmes liés à la qualité des données, au mieux de les anticiper, grâce au *data monitoring*. Cette pratique évalue la qualité des données en fonction de leur caractéristique (Mohanty, Jagadeesh, & Srivatsa, 2013).

En somme, il convient à ces responsables de la qualité de la donnée de s’approprier ce processus très détaillé afin d’assurer au mieux le maintien et l’amélioration de cette qualité. Aussi, une démarche proactive l’emportera toujours sur une démarche réactive (Mahanti, 2018) : il ne faut pas attendre qu’une anomalie survienne pour agir, celle-ci pouvant parfois amener à de lourdes répercussions.

ii. Outils de gestion et d’amélioration de la qualité des données

Pour terminer, nous relèverons quelques outils commerciaux dédiés à l’entretien et l’amélioration de la qualité des données (*cf.* Tableau 2.3.4.2).

Ainsi, il existe des outils de nettoyage de données permettant de gérer des migrations de données (parfois sources de dégradation de la qualité des données – *cf. supra*) via des

fonctionnalités d'analyse et de transformation des données. Parmi ces outils figure Power Center, programme développé par l'entreprise Informatica (Informatica, 2021). Par ailleurs, l'entreprise britannique Avellino a développé un outil d'analyse de la qualité des données appelé Discovery (Bibard, 2003), qui permet de détecter entre autres les doublons présents dans des bases de données. Cet outil est ainsi spécialisé dans le *data auditing*, tout comme ProfileStage et AuditStage, deux outils complémentaires développés par Ascential (IBM, 2003). Ces outils permettent la détection des défauts de données et mesure la qualité des données après extraction des métadonnées (Berti-Equille, 2004), offrant dès lors un retour sur investissement plus rapide et plus important pour des applications d'entreprise telles que CRM ou entreposage de données. En outre, spécialiste dans le nettoyage et la standardisation des données, Normad1 propose des solutions de restructuration et de normalisation des adresses postales, « qui consiste[nt] à analyser, détecter et, si possible, corriger les erreurs de rédaction d'une adresse (saisie, présente dans un fichier ou en base de données) conformément à la norme postale et en se basant essentiellement sur les référentiels de La Poste [en France]. » (Normad1, 2021). Ainsi par exemple, cette solution optimise et rentabilise les campagnes marketing et de communication. Pour finir, Accurity ou encore DataFlux sont deux outils globaux de gestion de la qualité des données qui tendent à accompagner les entreprises dans la compréhension de leurs données et l'évaluation de la qualité de celles-ci, entre autres afin de renforcer la qualité des prises de décisions (Accurity, 2021 ; SAS, 2021).

Outil	Développeur	Spécialité	Fonctions de l'outil
Discovery	Avellino	<i>Data auditing</i>	Détection entre autres les doublons présents dans des bases de données
AuditStage	Ascential		<i>Data profiling</i>
ProfileStage			
Normad1	Normad1	<i>Data cleansing</i> ⁶	Nettoyage, restructuration et normalisation des adresses postales françaises
Accurity	Accurity	<i>Data monitoring</i>	Surveillance et maintien des normes de qualité des données dans l'ensemble d'un environnement de données
DataFlux	SAS		

Tableau 2.3.4.2. Quelques outils de gestion de la qualité des données. Synthèse.

⁶ *Data cleansing* (ou nettoyage des données) désigne un processus qui supprime ou modifie les données de faible qualité (incorrectes ou mal formatées par exemple) car elles peuvent fournir des résultats inexacts (IBM, 2019).

2.4. Processus de stockage de données

Les concepts de données et base de données ont ainsi été couverts. Aussi, nous avons à plusieurs reprises souligné l'importance de considérer la qualité des données de manière continue. Il est temps de s'attarder quelque peu sur les outils de stockage de ces informations volumineuses. D'une part, quel est le type de base de données le plus rencontré dans un contexte Big Data ? D'autre part, en quoi consiste un centre de données (ou *datacenter*), et quels sont les éléments à considérer pour en assurer la performance ?

2.4.1. Bases de données

Les défis actuels associés au Big Data (notamment récupération rapide, traitement en temps réel, très grand volume de données) supposent des solutions de stockage adaptées (Khan, Liu, Ali, & Alam, 2019). Ces solutions nécessitent d'être toujours en mesure d'accueillir plus de données (évolutivité). Aussi, puisque le Big Data nécessite une analyse instantanée des informations, l'accès aux données doit être effectué fréquemment et efficacement, et ce dans un délai extrêmement court (disponibilité). Par ailleurs, les entreprises doivent veiller à la protection de leurs données (sécurité). Enfin, les solutions Big Data doivent pouvoir communiquer et interagir avec d'autres technologies (intégrité). Ce sont là les quatre critères soulignés par Khan *et al.* (2019), que doivent remplir les bases de données dans un contexte Big Data. Par conséquent, les bases de données dites relationnelles ne suffisent plus. Celles-ci sont essentiellement amenées à gérer des données bien structurées (Oracle, 2021a). Or, l'hétérogénéité des données s'étend de plus en plus à mesure que le temps passe, toujours plus rapidement. Dès lors, quel est le type de base de données qui répond au mieux à ces différents critères ?

Dans la littérature, il semble que les bases de données NoSQL⁷ constituent la solution la plus appropriée. De plus en plus utilisées (Chevalier, 2020), ces bases de données s'avèrent une solution plutôt flexible et peu coûteuse (McFadden, 2013, cité dans Siddiqi, Karim, & Gani, 2017). Sans structure prédéfinie, celles-ci s'adaptent en effet facilement aux données moins structurées (Raghav, Amudhavel, & Dhavachelvan, 2018), provenant notamment des réseaux sociaux, courriers électroniques et autres applications web. Face aux bases de données relationnelles, l'efficacité de la réponse et de la performance des bases de données NoSQL

⁷ NoSQL signifie *Not only Standard Query Language*.

l'emporte (Strohbach, Daubert, Ravkin, & Lischka, 2016). Les bases de données NoSQL se distinguent également par leur simplicité de mise en place (pas besoin de matériel supplémentaire), faisant de celles-ci des outils tout à fait rentables (Strohbach *et al.*, 2016). Toutefois, malgré leur simplicité, il ne faut pas négliger l'importance de la maîtrise des savoirs pour pouvoir en obtenir un résultat efficace. Un outil supplémentaire de visualisation des résultats est d'ailleurs nécessaire pour pouvoir afficher les résultats des requêtes effectuées dans la base de données (Raghav *et al.*, 2018).

On distingue plusieurs types de bases de données NoSQL : (i) *Column Store Database*, (ii) *Document Store Database*, (iii) *Graph Store Database*, et (iv) *Key-Value Store Database* (R. Chen, H. Chen, & Zang, 2010 ; Zhang, Chen, & Tan, 2015, cité dans Raghav *et al.*, 2018). Le premier type de base de données est construit autour d'une infinité de colonnes, au lieu de lignes comme dans la plupart des bases de données relationnelles (Strohbach *et al.*, 2016). Dans ces colonnes, d'énormes volumes de données sont stockés et identifiés grâce à l'attribution de clés uniques (Raghav *et al.*, 2018). Cassandra (Cassandra, 2021) constitue un bon exemple de *Column Store Database*. Dans le second type de bases de données, les données sont structurées. Toutefois, toutes les données ne doivent pas être structurées de la même manière, à l'inverse des bases de données relationnelles (Strohbach *et al.*, 2016). Elles permettent de stocker et récupérer des documents de différents formats tels que XML⁸ ou JSON⁹ par exemple (Raghav *et al.*, 2018). MongoDB (MongoDB, 2021) constitue un bon exemple de *Document Store Database*. Troisièmement, les *Graph Store Database* conservent les informations dans des structures de graphes, ce qui permet entre autres de cartographier des réseaux de données hautement connectées pour en révéler de potentielles relations cachées (Strohbach *et al.*, 2016). Neo4J (Neo4J, 2021) constitue un bon exemple de *Graph Store Database*. Enfin, la *Key-Value Store* reste le type de base de données NoSQL le plus simple d'utilisation (Raghav *et al.*, 2018). Les informations peuvent être non structurées ou structurées et sont accessibles par une clé unique. La valeur de cette clé doit être connue si l'on veut obtenir une information. Voldemort (Project Voldemort, 2021) en constitue un bon exemple.

⁸ XML signifie *Extensible Markup Language*. L'objectif de ce langage est de faciliter le traitement automatisé des informations, l'idée étant de structurer les informations à l'aide de balises (Université libre de Bruxelles, 2021).

⁹ JSON signifie *JavaScript Object Notation*. Il s'agit d'un langage d'échange de données qui permet de représenter de l'information structurée (JSON, 2021).

2.4.2. Focus sur les datacenters : cas de Google

Pour terminer, intéressons-nous au concept de centres de données, ou *datacenters*. Ces sites représentent des infrastructures technologiques impressionnantes équipées de serveurs, baies de stockage et autres composants réseaux, le tout situé dans un environnement intégralement contrôlé : alimentation électrique, sécurité, climatisation, dispositifs anti-incendie, etc. (Daydé & Veynante, 2017). A titre d'exemple en Belgique, la célèbre entreprise Google détient un gigantesque *datacenter* implanté dans la région de Mons, opérationnel depuis 2010 (Google, 2021a). Néanmoins, ces centres de données couvrent des enjeux de taille. Quels sont-ils ? Nous identifions trois points importants : les ressources, l'énergie et la sécurité.

i. Économie de ressources

Dans cette sous-section, nous distinguerons deux types de ressources : les ressources financières d'une part, les ressources humaines d'autre part.

Tout d'abord, en ce qui concerne les ressources financières, un *datacenter* requiert des investissements importants nécessaires au lancement de l'activité (Daydé & Veynante, 2017). Autrement dit, les coûts liés à la construction du site, qui concernent donc le bâtiment, le matériel de stockage (entre autres serveurs et baies de stockage), les dispositifs de sécurité, les mécanismes de refroidissements du matériel de stockage des données, etc. Une fois le site mis en place et opérationnel, un centre de données implique des coûts liés à l'activité de stockage elle-même (i.e. coûts d'exploitation) (Daydé & Veynante, 2017). Citons les coûts de maintenance informatique, frais de personnel (*cf. infra*), frais d'électricité et d'alimentation du site, etc.

Ainsi, un centre de données coûte cher. En effet, Google évalue l'investissement total en Belgique lié à ses centres de données à près de 2,2 milliards d'euros, dont 300 millions d'euros dernièrement investis dans son *datacenter* de la région de Mons (Samain, 2020). C'est la raison pour laquelle, en Belgique, les centres de données sont dans la plupart des cas mis à disposition des entreprises qui désirent stocker des données marketing relatives à leurs clients, en leur évitant ainsi ces coûts très élevés, ceux-ci étant alors mutualisés (Witsenburg, 2016).

Par ailleurs, en termes de ressources humaines, celles-ci s'avèrent essentielles dans la gestion des centres de données. Des techniciens aux directeurs de site, en passant par les installateurs,

ingénieurs et responsables de la sécurité du site (Google, 2021c), les entreprises doivent s'entourer de profils variés pour assurer le bon fonctionnement du site de stockage.

ii. Économie d'énergie

Le second élément souligné est l'énergie. En 2020, l'organisme allemand Statista (2021a) évalue la quantité de données créées dans le monde à 50 zettaoctets. En comparaison, il faudrait 470 millions des plus gros disques durs actuels pour pouvoir supporter cette quantité immense de données, dont l'impact énergétique est considérable. Toutefois, cet impact n'est aujourd'hui que trop peu considéré et interrogé (Cappy, 2017). En effet, le stockage de ces informations dans les *datacenters* représente 48 % de la consommation énergétique associée au Big Data (Cappy, 2017). Que représente cette énergie ? L'ensemble du matériel fonctionne principalement grâce à l'énergie électrique. Afin de diminuer les coûts d'alimentation du site, des organisations (à l'instar de Google) investissent dans des infrastructures éoliennes et solaires afin d'en récupérer l'énergie pour alimenter leurs sites de stockage (Cappy, 2017). Par ailleurs, l'énergie est également utilisée dans les *datacenters* pour refroidir les environnements de stockage, ceux-ci dégageant une forte chaleur étant donné leur fonctionnement continu. Des systèmes de climatisation sont ainsi mis en place pour pallier ces surchauffes (Google, 2021a). Dans le cas de Google (2021a) en Belgique, l'entreprise a quelque peu innové en matière de refroidissement du matériel : ce centre de données utilise l'eau recyclée d'un canal à proximité du site pour refroidir ses serveurs, maximisant ainsi l'efficacité de son matériel.

iii. Garantie de sécurité

Le dernier point à considérer n'est pas des moindres : il s'agit de la sécurité. Dans cette sous-section, nous développerons surtout l'enjeu de la sécurité du site de stockage de données, la protection des données elles-mêmes ayant déjà été couverte dans une section précédente (*cf.* section 2.2.4).

Les données personnelles relatives aux consommateurs étant d'une grande sensibilité, Google ne plaisante pas avec la sécurité de ses sites de stockage de données (Google, 2021b). Chacun de ses centres de données est protégé par une série de niveaux de sécurité imaginés pour empêcher les accès non autorisés. A l'extérieur, le site de stockage est protégé par des caméras de type standard et thermique à haut niveau de portée, des barrières et portails sensibles au

contact humain. Ainsi, une alerte est émise au service de sécurité si un individu approche le portail frontalier ou le touche. A l'intérieur du complexe, les autorisations d'accès de chaque personne (qu'elle soit collaborateur ou non) sont vérifiées par des lecteurs de badges que l'on retrouve à chaque point d'accès du centre de données. Une autre règle empêche une personne qui en suit une autre de près de pénétrer dans une zone à accès restreint sans passer un badge. Si le système détecte une porte ouverte trop longtemps, il alerte immédiatement le personnel de sécurité : toute porte ou portail extérieur doit se fermer avant que la personne suivante puisse s'identifier grâce à son badge et accéder à une zone. Par ailleurs, à certains endroits des centres de données, chaque collaborateur doit faire glisser son badge, puis entrer dans un sas circulaire ou une porte en tube (similaire au dispositif de sécurité de certains aéroports), qui vérifie l'authenticité du badge et effectue un scanner rétinien. Si l'individu est reconnu, il accède au niveau suivant du *datacenter* (Google, 2021b).

En ce qui concerne les données elles-mêmes, celles-ci sont stockées dans des disques durs (eux-mêmes repris dans des baies de stockage), répartis à plusieurs endroits différents, et ce de manière cryptée (Google, 2021b). Ces disques durs sont conçus pour être traçables en quelques secondes, avec un degré élevé de précision, ce afin d'éviter que les données n'arrivent entre de les mains d'individus dotés de mauvaises intentions. Des contrôles de la performance du matériel sont effectués régulièrement, en vue d'identifier les dysfonctionnements informatiques et d'évaluer la qualité du matériel de stockage.

2.5. Revue de la littérature : récapitulatif

Ainsi, nous nous sommes intéressés brièvement au concept de Big Data, en se focalisant ensuite sur le rôle qu'il joue dans le marketing. Puis, nous nous sommes concentrés sur les données marketing et leurs différents types, pour s'intéresser ensuite de manière détaillée à la qualité des données stockées : d'une part, en identifiant les attributs qui caractérisent une donnée de qualité ; d'autre part en relevant les types d'anomalies rencontrées dans les bases de données, dont les répercussions impactent négativement la qualité des données stockées. Par ailleurs, nous avons mis en avant un schéma de travail de contrôle de la qualité des données proposé par la littérature, ainsi que quelques outils commerciaux spécialisés dans la gestion de la qualité de la donnée. Au terme de ce dernier chapitre, nous sommes ainsi parvenus à construire une liste qui rassemble une série de critères liés au stockage de données marketing de qualité, répartis en deux dimensions : une première dimension liée à la qualité des données elles-mêmes, une seconde qui concerne la qualité de leur processus de stockage. Ces deux dimensions sont représentées dans le tableau de synthèse 2.5.1 ci-dessous.

Qualité des données stockées	Qualité du processus de stockage
▸ Complétude ▸ Pertinence ▸ Exactitude et Précision (<i>Accuracy</i>) ▸ Conformité ▸ Consistance et Cohérence (<i>Consistency</i>) ▸ Récence ▸ Unicité ▸ Accessibilité et Disponibilité ▸ Intégrité	▸ Économie de ressources ▸ Économie d'énergie ▸ Garantie de sécurité

Tableau 2.5.1. Les critères d'un stockage de données marketing de qualité identifiés depuis la littérature.

Chapitre 3 : Méthodologie de l'étude qualitative

La revue de la littérature aura ainsi permis de constituer une liste de critères qui font d'une donnée marketing une donnée de qualité, répartis en deux dimensions : une première dimension qui concerne la qualité des données elles-mêmes, une seconde liée à la qualité du processus de stockage de celles-ci. Pour rappel, la principale finalité du mémoire revient à proposer une liste fiable et pertinente de critères qui déterminent le degré de qualité d'une donnée marketing. La liste établie au travers de la revue de la littérature restant toutefois très théorique, il est dès lors intéressant de confronter cette liste à des expériences pratiques que rencontrent des professionnels de la donnée.

Nous opterons ainsi pour une approche empirique inductive. Par opposition à l'approche déductive, cette démarche qualitative permet d'établir une théorie sur base d'observations effectuées sur le terrain (Lerum, 2014). Plus concrètement, nous tenterons d'affiner la liste de critères initialement construite au travers de la revue de la littérature en la soumettant à une série d'experts du métier. Nous prévoyons une entrevue auprès de cinq experts dans le stockage et la qualité des données (*cf.* Tableau 3.1), que nous présenterons brièvement ci-après. Ces entrevues seront de type semi-structuré. Cette méthode est utilisée pour approfondir un sujet de recherche en récoltant des données ciblées et qualitatives auprès d'intervenants experts, offrant ainsi une meilleure compréhension du domaine spécifique de notre recherche (Saunders, Lewis, & Thornhill, 2016). Aussi, nous fixerons la durée des entretiens à une heure afin de rassembler un suffisamment d'informations, que nous exploiterons dans le cadre de notre analyse empirique. Par ailleurs, les propos de ces cinq individus seront recueillis au travers d'un court guide d'entretien (*cf.* Annexe 6, p. 81). Ce guide servira uniquement de fil rouge durant la discussion, l'objectif étant de laisser l'interlocuteur parler librement. D'ailleurs, nous supposerons que lorsque le critère n'est pas abordé par l'intervenant, celui-ci le considère comme de moindre importance, l'idée étant de garantir un contenu totalement original, sans biais.

Pour des raisons pratiques, ces entrevues seront enregistrées moyennant l'accord de chaque interlocuteur. Notons que sitôt les entretiens remis par écrit sous la forme de compte-rendu (*cf.* Annexe 7, p. 82), ces enregistrements seront définitivement détruits. Une fois cela, nous confronterons l'opinion de chaque interlocuteur à notre grille de critères établie dans le chapitre

précédent (cf. Section 4.3.1, Tableau 4.3.1.1). Notre analyse empirique aura pour finalité d'établir une liste de critères d'un stockage de données marketing de qualité, classés en fonction de l'importance apportée par les intervenants à chacun d'entre eux. En effet, nous aurons ainsi eu l'occasion de confronter la théorie ainsi que des opinions et expériences d'individus experts, qui auront permis de solidifier notre liste de critères construite préalablement.

Parmi les interlocuteurs sollicités dans le cadre de la démarche, Jean-Christophe Dewitte est *Business Analyst Policy* dans l'entreprise spécialiste de la téléphonie, Orange. Tout au long de son parcours, Jean-Christophe aura eu l'occasion de se s'intéresser aux environnements des données et du *campaigning* (traitement et stockage de données marketing), ce qui en fait un intervenant tout à fait légitime dans le cadre de la démarche qualitative envisagée.

Ensuite, Marguerite Paris est actuellement *Data Marketing Scientist* chez Trade Republic, néo-courtier en bourse allemand. Son avis sur la problématique de la qualité des données marketing est davantage crédibilisé lorsqu'on observe son parcours. Elle fait ses débuts chez HelloFresh (fournisseur de kits repas) en tant qu'analyste marketing, puis intègre l'entreprise allemande N26 (banque en ligne) pour devenir *Data Marketing Product Owner*.

Par ailleurs, Camille Laimay est directrice commerciale chez Weborama France, organisme spécialisé dans la collecte de données marketing qui permettent d'établir des profils de consommateurs cibles, de calculer des paniers moyens, etc. L'entreprise met ces informations à disposition de clients durant un certain temps pour différents cas d'usage, en échange d'un forfait. Weborama ne vend donc pas les données qu'elle parvient à obtenir, elle en reste ainsi propriétaire.

Éric de Bellefroid est *Senior Marketing Consultant* chez Freedelity. Cette entreprise propose une solution aux retailers qui permet de leur apprendre à mieux connaître leurs consommateurs. Elle les aide ainsi à collecter des données personnelles sur le lieu de vente, en vue de se constituer une base de données qui soit apte à représenter au mieux la relation qu'a le distributeur avec ses consommateurs.

Enfin, Ali El Maghraoui est directeur exécutif chez EY dans le département *Data Analytics*. Ce département est en charge de l'étude des stratégies *data* des entreprises (vision, *business case*

data, roadmap, gouvernance de la donnée, architecture et modèles de données, etc.). Son expertise est plus fonctionnelle que technique dans le domaine. Le point de vue de cet intervenant s'avère d'un grand intérêt dans notre démarche qualitative. En effet, la consultance l'amène à collaborer avec plusieurs entreprises simultanément, élargissant ainsi son opinion sur le sujet étudié.

Interlocuteur	Organisation	Occupation	Données
Jean-Christophe Dewitte	Orange	<i>Business Analyst Policy</i>	Marketing (<i>First-party data</i>)
Marguerite Paris	Trade Republic	<i>Marketing Data Scientist</i>	Marketing (<i>Third-party data</i>)
Camille Laimay	Weborama	<i>Head of Business</i>	
Éric de Bellefroid	Freedelity	<i>Senior Marketing Consultant</i>	Marketing (<i>First-party data</i>)
Ali El Maghraoui	EY	<i>Executive Director</i>	Tous types de données, dont marketing

Tableau 3.1. Intervenants sollicités dans l'approche qualitative. Nom et prénom, organisation, occupation actuelle, type de données qui concerne l'intervenant.

Chapitre 4 : Résultats empiriques

Nous avons ainsi interagi avec plusieurs spécialistes en stockage de données. Pour rappel, ces entrevues se sont déroulées de manière semi-structurée, laissant à chaque interlocuteur la liberté de s'exprimer de manière la plus spontanée possible. Au terme de ce chapitre, nous serons en mesure de confronter notre liste théorique aux expériences de chacun des intervenants, en se concentrant à la fois sur les points communs aux analyses théorique et empirique, mais aussi sur les divergences survenues au cours des discussions. Ce dernier point est intéressant dans la mesure où il nous permettra de compléter notre liste à la hausse et/ou à la baisse.

4.2. Observations

Dans cette section, nous nous intéresserons en détail aux propos de chacun des intervenants en restituant de manière claire leurs commentaires pour chaque critère identifié dans la littérature. Finalement, nous construirons un tableau de synthèse qui confrontera les points de vue de l'ensemble des interlocuteurs, qui nous permettra alors d'attribuer un score à chacun des critères théoriques sélectionnés en fonction du nombre de fois où ce dernier a été relevé.

4.2.1. Qualité des données stockées

i. Observations relatives aux critères identifiés dans la littérature

Complétude

Intervenant	Commentaire
J-C. Dewitte	-
M. Paris	-
C. Laimay	L'intervenante ne commente pas le critère mais marque son accord. Elle ajoute tout de même que ce critère est d'une grande importance.
E. de Bellefroid	▸ L'utilisation de la carte d'identité comme moyen de collecte des données personnelles assure l'exhaustivité des différents champs de données attendus (ex. nom et prénom, date de naissance, adresse postale, etc.). Cela permet d'éviter les doublons dans les bases de données. ▸ Si dans 80 % des enregistrements stockés dans la base de données aucune adresse postale n'est renseignée, cela ne sert à rien.

<i>Complétude – Suite</i>	
Intervenant	Commentaire
A. El Maghraoui	La complétude des données se définit bien par le fait que toutes les données soient bien présentes dans l’environnement de données. Le degré de complétude ne concerne que la manière dont on présente les résultats. Il faut partir du principe que pour mesurer une qualité des données, on doit d’abord définir comment on doit mesurer (i.e. complétude = s’assurer que le champ n’est pas vide). Ensuite, on détermine un seuil qui nous dit que la qualité des données est bonne lorsque 90 % des données sont complètes.

Pertinence

Intervenant	Commentaire
J-C. Dewitte	Stocker une information n’est pertinent que si l’on est certain d’obtenir un retour important sur l’utilisation de cette donnée.
M. Paris	La pertinence est d’une grande importance. Toutefois, Trade Republic est une entreprise qui travaille beaucoup avec des <i>third-party data</i> , que celle-ci va récolter directement via des API ¹⁰ (cf. Annexe 8, p. 91) de Google, Facebook ou de plus petits fournisseurs de données. On suppose donc que les données avec lesquelles Trade Republic travaille sont pertinentes.
C. Laimay	L’intervenante ne commente pas le critère mais marque son accord. Elle ajoute tout de même que ce critère est d’une grande importance.
E. de Bellefroid	La pertinence est importante. En effet, les décisions prises sur la base de toutes les modélisations qui seront effectuées vont dépendre des données stockées elles-mêmes.
A. El Maghraoui	En ce qui concerne la pertinence des données, celle-ci doit pouvoir se différencier de la dimension précédente. On ne devrait pas pouvoir considérer cela comme un critère de stockage, car cela dépendra toujours de l’utilisation qu’on fera de cette donnée. Nous avons parlé de pertinence de stockage : doit-on stocker cette valeur ou non ? Néanmoins, à partir du moment où on se demande si on doit stocker cette donnée ou non, cela concerne surtout le design fonctionnel plutôt qu’une vraie mesure de qualité. Car si on ne stocke pas l’information, on ne sait pas déterminer le niveau de qualité de cette donnée étant donné qu’elle n’est pas stockée. Il y a une nuance. A partir du besoin, on détermine quelles seront les données à stocker dans les systèmes d’informations. La pertinence n’est donc pas un indicateur de qualité de donnée.

¹⁰ *Application Programming Interface* ou API désigne un ensemble de protocoles qui permettent à deux applications de communiquer entre elles sans connaître les informations de chacune de leur mise en œuvre (RedHat, 2021).

Exactitude et précision (*Accuracy*)

Intervenant	Commentaire
J-C. Dewitte	▸ Efficacité et propreté des processus garantissent des données stockées qui soient exactes et précises. ▸ Les métiers d'insertion d'informations propres dans des base de données sont souvent négligés, toutefois d'une grande importance. ▸ Une donnée inexacte peut entraîner une perte de la connaissance du client concerné par l'anomalie.
M. Paris	-
C. Laimay	L'intervenante ne commente pas le critère mais marque son accord. Elle ajoute tout de même que ce critère est d'une grande importance.
E. de Bellefroid	L'utilisation de la carte d'identité comme moyen de collecte des données assure l'exactitude et la précision des données. En effet, le risque d'anomalies diminue largement étant donné que les données sont déjà standardisées, pas de risque d'erreur d'encodage. La qualité des données stockées par Freedelity dépend ainsi de la manière dont ces informations sont renseignées par nul autre que les services de population, dépendant de l'État belge lui-même.
A. El Maghraoui	Dans la pratique, l'exactitude et la pertinence (<i>accuracy</i>) représentent la dimension de la qualité des données la plus compliquée à atteindre et à définir. Dire qu'une donnée doit correspondre à la réalité de ce qu'elle représente c'est très bien, mais il faut également déterminer ce que représente cette réalité. L'implémenter dans un code SQL ou autre est très complexe car toute la logique qui gravite autour de cette donnée doit être refaite, ce afin de vérifier que tout est correct.

Conformité

Intervenant	Commentaire
J-C. Dewitte	▸ Il n'y a pas assez de travail effectué sur la standardisation des informations. Par facilité, ce travail de standardisation des données doit être effectué en amont par des experts-métiers qui comprennent tous les impacts, car des données de mauvaise qualité font diminuer l'efficacité globale d'une entreprise. ▸ Plus on accorde de libertés dans une base de données, plus la probabilité d'y rencontrer des anomalies (ex. doublons) augmente. Il faut contrôler la validité des données qui entrent dans la base de données.
M. Paris	La documentation doit être claire sur les données (métadonnées), qu'il s'agisse de <i>first-</i> ou <i>third-party data</i> . Sans information sur l'intitulé des colonnes d'une table par exemple, ou encore la nature de la donnée (ex. un chiffre, une chaîne de caractères, etc.), il est difficile de comprendre la donnée, et donc de l'utiliser : quelle est la source des données, quelle est la fréquence de rafraîchissement des données ?

<i>Conformité – Suite</i>	
Intervenant	Commentaire
C. Laimay	L'intervenante ne commente pas le critère mais marque son accord. Elle ajoute tout de même que ce critère est d'une grande importance.
E. de Bellefroid	- La documentation est importante. Mais il est très compliqué d'avoir un taux de documentation homogène de la donnée en règle générale. - La standardisation des données permet d'éviter les doublons. Freedelity utilise des logiciels algorithmiques qui standardisent entre autres les adresses postales. En effet, l'entreprise travaille avec des référentiels de rue qui permettent d'adjoindre à chaque adresse un identifiant unique d'adresse. C'est cette clé unique qui est utilisée pour différencier les consommateurs.
A. El Maghraoui	La conformité des données renvoie à la validité des données. Deux concepts surviennent souvent de manière conjointe : la conformité et la validité des données. La conformité (plutôt fonctionnelle) renvoie au respect du format qui a été défini pour la donnée. La validité (plutôt technique) quant à elle permet de contrôler la donnée stockée (par exemple un âge ne peut pas être inférieur à 0).

Consistance et cohérence (*Consistency*)

Intervenant	Commentaire
J-C. Dewitte	La réplication des base de données due à la saturation des espaces de stockage induit des incohérences de données : dans deux endroits de l'environnement de donnée, nous ferons face à deux données temporellement différentes.
M. Paris	Si une colonne devait changer d'intitulé dans l'environnement de données d'un fournisseur de données, il faudra impérativement contrôler toutes les pipelines de données. Ce critère est important et requiert une série de processus de conduite du changement à ne pas négliger.
C. Laimay	L'intervenante ne commente pas le critère mais marque son accord. Elle ajoute tout de même que ce critère est d'une grande importance.
E. de Bellefroid	-
A. El Maghraoui	Le critère suivant qui concerne la consistance et la cohérence des données représente l'absence de contradictions dans l'environnement de donnée, ainsi que l'homogénéité des représentations dans ce même environnement. A nouveau, ce critère doit être mesurable. Si on compare deux enregistrements de date de naissance pour Maxime, 1998 d'un côté et 1999 de l'autre, il y a clairement un problème de consistance. En revanche, une date stockée dans tel format dans une base de données, ou un autre format, c'est un choix de design, pas de rapport avec la donnée elle-même. Souvent, si on compare des bases de données Oracle avec SQL Server, certains stockeront la donnée comme année, mois, jour, heure, minute, seconde ; tandis que d'autres proposeront de stocker en année, mois, jour ou heure, minute, seconde. Cela dépend des contraintes techniques du système.

Récence

Intervenant	Commentaire
J-C. Dewitte	▸ La réplication des base de données due à la saturation des espaces de stockage induit des incidents de fraîcheur des données : dans deux endroits de l'environnement de donnée, nous ferons face à deux données temporellement différentes. ▸ Il arrive très souvent que des mécontentements surviennent en raison de problèmes liés à la récence des données. La gestion des déménagements constitue un exemple de cas concret : environ 10 % de la population déménage chaque année. Si l'on ne s'attarde pas sur la mise à jour des données, en quelques années les données de localité ne valent plus rien.
M. Paris	▸ Il est clair que les données doivent être rafraichies si l'on veut capter leur valeur la plus actualisée possible. Toutefois, celles-ci ne doivent pas constamment changer. En tous les cas, si elles doivent changer, elles le peuvent, mais changements continus ne doivent en rien détruire le travail effectué en amont de ce rafraîchissement. ▸ Dans une table de données, l'idéal est d'avoir une colonne qui renseigne la dernière date de mise à jour de la donnée, de sorte que l'utilisateur de cette donnée puisse s'assurer de la qualité de la donnée qu'il s'apprête à exploiter.
C. Laimay	▸ L'intervenante ne commente pas le critère mais marque son accord. Elle ajoute tout de même que ce critère est d'une grande importance. ▸ L'entreprise recommande d'ailleurs vivement une fraîcheur de données inférieure ou égale à trente jours.
E. de Bellefroid	▸ Avec l'unicité, c'est là sans doute la caractéristique la plus importante. ▸ Freedelity met à disposition des retailers un service de mutualisation de la tenue à jour des données personnelles des consommateurs entre retailers. Collecter de la donnée consommateur est une chose, toutefois la tenir à jour dans le temps constitue un défi de taille pour les entreprises. ▸ La collecte des données via lecture de la carte d'identité belge vient directement mettre à jour les données dans le <i>CRM</i> des entreprises.
A. El Maghraoui	La récence est primordiale. Elle renvoie à l'actualisation des données, mais aussi à la fréquence de mise à jour des données : à quelle fréquence ces données sont-elles mises à jour ? Tous les jours ? Toutes les minutes ? Il faut s'assurer qu'il s'agisse de la dernière version de la donnée.

Unicité

Intervenant	Commentaire
J-C. Dewitte	▸ L'idéal est d'avoir une et une seule donnée avec un faible taux de réplcation. ▸ Plus on accorde de libertés dans une base de données, plus la probabilité d'y rencontrer des anomalies (ex. doublons) augmente. La gestion des doublons est extrêmement difficile à gérer. Souvent, ces anomalies ne sont pas traitées. Il faut contrôler la validité des données qui entrent dans la base de données.
M. Paris	A nouveau, éviter à tout prix les doublons dans les base de données.
C. Laimay	L'intervenante ne commente pas le critère mais marque son accord. Elle ajoute tout de même que ce critère est d'une grande importance.
E. de Bellefroid	▸ Avec la récence, c'est là sans doute la caractéristique la plus importante. ▸ L'utilisation de la carte d'identité comme moyen de collecte des données personnelles (ex. nom et prénom, date de naissance, adresse postale, etc.) permet d'éviter toute forme de doublons dans les bases de données, principal ennemi des entreprises en termes de qualité des données. ▸ Dans son référentiel de consommateurs, Freedelity permet de fusionner des données en présence de doublons. Ainsi par exemple, si pour une même carte de fidélité on retrouve deux enregistrements différents de nom, l'utilisateur de la donnée a la possibilité de faire en sorte que les deux lignes soient fusionnées, en prenant pour repère la ligne la plus récente. Cette fusion n'est effectuée qu'après un calcul de probabilité de ressemblance entre deux lignes de consommateurs.
A. El Maghraoui	L'unicité de la donnée renvoie à la représentation unique des données, ainsi qu'à l'absence de répétitions dans l'environnement de stockage de données. La plupart du temps, surtout dans les systèmes de stockage actuels, cette caractéristique est vérifiée par le design. On perd souvent du temps à construire un indicateur de qualité pour l'unicité car la plupart du temps elle est vérifiée par le design, au travers des clés primaires et uniques de contrôle.

Accessibilité et disponibilité

Intervenant	Commentaire
J-C. Dewitte	▸ Il y a un besoin très important mettre la main sur les informations stockées le plus rapidement possible. ▸ L'environnement de stockage est tellement lent qu'il en devient parfois impossible d'y accéder durant un certain laps de temps. ▸ La flexibilité des accès aux plateformes de données doit être prise en compte.

<i>Accessibilité et disponibilité – Suite</i>	
Intervenant	Commentaire
M. Paris	▸ Les données doivent être accessibles facilement et rapidement, mais aussi compréhensibles et utilisables. ▸ Les données sont collectées dans les API (<i>cf.</i> Annexe 8, p. 91) par des ingénieurs de la donnée pour les implémenter dans le système de travail interne de l'entreprise. C'est seulement dès cet instant que les équipes de traitement de données peuvent commencer à travailler. Ces interventions préliminaires ralentissent le processus de stockage de données de par leur difficulté et le temps nécessaire pour les appliquer.
C. Laimay	L'intervenante ne commente pas le critère mais marque son accord. Elle ajoute tout de même que ce critère est d'une grande importance.
E. de Bellefroid	▸ Les données sont collectées à une vitesse remarquable étant donné que la majorité d'entre elles proviennent de la lecture de la puce de la carte d'identité belge. ▸ Freedelity est une plateforme à haut degré de transactions, avec des degrés de disponibilité importants.
A. El Maghraoui	L'accessibilité et la disponibilité des données reprend la facilité d'accès, de compréhension et d'utilisation des données. Ce critère est souvent mentionné, cependant une question se pose : comment mesurer cela ? S'il s'agit d'un indicateur, il faut pouvoir en mesurer la performance. Or ici, c'est très compliqué. Par définition, un indicateur doit être mesurable. Cela étant, cette caractéristique peut être mesurée au travers des différents processus opérationnels. La seule manière de mesurer cet indicateur est de vérifier si le processus informatique qui a permis à une donnée d'être chargée et stockée a été exécuté ou non. A ce moment-là, cela devient un check opérationnel, et non plus un indicateur de qualité.

Intégrité

Intervenant	Commentaire
J-C. Dewitte	-
M. Paris	Si une colonne devait changer d'intitulé dans l'environnement de données d'un fournisseur de données, il faudra impérativement contrôler toutes les pipelines de données. Ce critère est important et requiert une série de processus de conduite du changement à ne pas négliger.
C. Laimay	L'intervenante ne commente pas le critère mais marque son accord. Elle ajoute tout de même que ce critère est d'une moins grande importance, toutefois à ne pas négliger.

<i>Intégrité – Suite</i>	
Intervenant	Commentaire
E. de Bellefroid	Le processus de collecte des données par Freedelity se fait en deux temps : scanning de la carte de fidélité de l'enseigne, puis lecture de la carte d'identité du consommateur. Ce processus très rapide en deux temps permet à Freedelity de créer des liens entre un consommateur et une à n cartes de fidélité, construisant ainsi un référentiel de consommateurs muni d'une interface qui permet, en présence de doublons, de fusionner les données.
A. El Maghraoui	L'intégrité revient à s'assurer que dans un système un client Maxime soit renseigné, et que dans un second système, ce même client soit retrouvé. C'est aussi simple que cela. Cela peut se traduire par les clés primaires et clés étrangères, ou encore amener cela au niveau fonctionnel.

ii. Nouveaux critères identifiés au terme des discussions

Automatisation

Intervenant	Commentaire
J-C. Dewitte	-
M. Paris	Une donnée sera de meilleure qualité si elle ne nécessite aucunement une intervention humaine. En effet, ce type d'intervention augmente considérablement le risque d'erreur et impacte directement la qualité des données concernées.
C. Laimay	-
E. de Bellefroid	Freedelity ne fonctionne qu'avec des scripts. Aucune intervention humaine ne sera effectuée sur les données stockées. Ces erreurs influenceront défavorablement la qualité et la robustesse des données et des modèles de données qui en découlent.
A. El Maghraoui	-

Source de la donnée

Intervenant	Commentaire
J-C. Dewitte	-
M. Paris	La source de la donnée est importante et est garante d'un niveau de qualité de donnée. Trade Republic travaille le plus souvent avec des API qui vont puiser des informations dans les bases de données de fournisseurs (notamment Google et Facebook). Ce sont là de grands acteurs <i>third-party</i> garants d'un certain degré de qualité de données, en comparaison aux plus petits fournisseurs de données, dont la qualité est d'un niveau plus faible.

<i>Source de la donnée – Suite</i>	
Intervenant	Commentaire
C. Laimay	La méthode de collecte permet déjà de se faire une première idée de la qualité des données concernées, et d'en déterminer la fiabilité. Une source sera par exemple fiable si les protocoles de collecte sont respectés par le fournisseur de données. En d'autres termes, si la méthode de collecte est bonne, la qualité des informations qui ont été collectées est un peu assurée.
E. de Bellefroid	-
A. El Maghraoui	-

Stabilité

Intervenant	Commentaire
J-C. Dewitte	-
M. Paris	-
C. Laimay	-
E. de Bellefroid	-
A. El Maghraoui	La stabilité est une dimension de plus en plus soulignée par des clients. Cette caractéristique renvoie à s'assurer qu'à travers le temps, il n'y ait pas de variations majeures constatées dans la base de données. Ainsi par exemple, si chaque mois on enregistre un million de clients, et qu'un mois on en enregistre 10 millions, il y a de quoi se poser des questions.

4.2.2. Qualité du processus de stockage

i. Observations relatives aux critères identifiés dans la littérature

Économie de ressources

Intervenant	Commentaire
J.-C. Dewitte	▸ L'espace nécessaire au stockage de fichiers de données coûte de moins en moins cher à l'heure actuelle, en comparaison de l'époque. Toutefois, les incidences qui concernent la saturation de l'espace de stockage sont très coûteuses à corriger. ▸ On a conservé des bases de données solides mais peu flexibles à la fois sur l'augmentation de la puissance de calcul (moindre, difficile à augmenter, coûteuse), mais aussi sur l'augmentation de stockage. ▸ Les entreprises ont pris pour habitude de stocker toujours et encore plus d'informations. Au plus de l'information est stockée, au plus l'entreprise considère qu'elle se porte bien. Cependant, l'accumulation de ces données et le manque de ressources font que la globalité des données stockées d'une entreprise empire et s'agrandit de manière exponentielle. Aujourd'hui, il est en effet difficile de demander à un département informatique (responsable du matériel de stockage) de mettre en cause la nécessité de conserver les informations du côté marketing. Par conséquent, le stockage arrive souvent à son niveau maximum de capacité, et ce trop rapidement. ▸ Il n'existe pas de nomenclatures pour les fichiers de données, ni de conventions d'archivage : on ne surveille pas la taille des données que l'on utilise. Il faut un cadre qui précise les limitations des outils, ainsi que la temporalité des purges de données. ▸ La flexibilité de l'accès des utilisateurs aux plateformes de données (tant les collaborateurs métiers – ex. marketing – que les collaborateurs techniques – ex. informatique) doit être prise en compte.
M. Paris	Beaucoup d'outils sont mobilisés pour évaluer et améliorer la qualité des données, comme Snowflake ou Redshift.
C. Laimay	Weborama (à l'échelle internationale) détient un et un seul centre de données.

<i>Économie de ressources – Suite</i>	
Intervenant	Commentaire
E. de Bellefroid	▸ Freedelity héberge ses propres serveurs auprès d'un prestataire externe dans un centre de données. Ces serveurs sont configurés par l'entreprise elle-même, et situés dans des racks informatiques. Ils sont hébergés dans deux endroits spécifiques. En cas d'incident, ce sont les équipes de Freedelity qui se mobilisent pour intervenir sur le matériel. ▸ Les données sont stockées dans des silos de données propres à chaque enseigne, tous extrêmement sécurisés. ▸ Le contrôle de qualité des données se fait automatiquement via des algorithmes de contrôle, notamment dans le cas où le consommateur ne dispose pas de sa carte d'identité ou n'est pas un citoyen belge. ▸ Freedelity met à disposition des services de contrôle de qualité des données, de manière directe ou indirecte (via deux entreprises spécialisées). Ces prestations incluent par exemple la comparaison du référentiel consommateurs de Freedelity (qui pour rappel concerne près de 6,5 millions de consommateurs belges) à d'autres bases de données pour en enrichir le contenu. ▸ Dès que des anomalies de qualité des données sont identifiées, les modifications ne sont pas effectuées manuellement : tout est corrigé au travers de scripts informatiques.
A. El Maghraoui	Le problème avec ce critère, c'est qu'on ne peut pas l'associer uniquement au stockage, cela va au-delà de ça. Il faut considérer le cycle de vie de la donnée en général en général.

Économie d'énergie

Intervenant	Commentaire
J-C. Dewitte	-
M. Paris	-
C. Laimay	-
E. de Bellefroid	-

<i>Économie d'énergie – Suite</i>	
Intervenant	Commentaire
A. El Maghraoui	A nouveau, comment mesurer cet indicateur ? Il ne faut pas oublier que ces critères, il va falloir les implémenter en entreprise. A un moment donné, ils devront apparaître dans un dashboard avec des couleurs vertes, rouges et orange, en disant si bon ou pas bon. Il faut alors être en mesure de trouver une méthode d'évaluation et de facturation de ce critère. Si on n'y parvient pas, ce critère n'est pas à prendre en compte. Actuellement, beaucoup d'entreprises tendent à imaginer des indices (vs. Indicateurs) qui rendent compte de l'impact sociétal de leurs activités Big Data. Le stockage en tant que tel n'a pas vraiment d'impact énergétique. Prenons l'exemple de la cryptomonnaie. Ce n'est pas le fait de la stocker qui fait qu'elle est énergivore, c'est plutôt tout le processus de miner qui coûte de l'énergie. Ce critère énergétique dépendra de l'utilisation que l'on fait de la donnée, et de comment on la calcule, plutôt que de comment on stocke cette information. Si on veut considérer l'énergie comme un critère de qualité de données, il faudra penser plus large, ne pas se concentrer sur le stockage des informations, qui est tout à fait négligeable comparé à toute la puissance de calcul utilisée pour l'extraire, la transformer et la mettre quelque part.

Garantie de sécurité

Intervenant	Commentaire
J-C. Dewitte	-
M. Paris	▸ En des termes de sécurité des données, les systèmes sont parfois tellement sécurisés qu'il est difficile d'y pénétrer. ▸ Le secteur bancaire exige que les données sensibles des utilisateurs soient anonymisées avant utilisation. Ce processus entraîne à la fois une lourde perte de temps, mais aussi une perte partielle de données. ▸ Les entreprises doivent se soumettre aux réglementations imposées par le RGPD. Outre cela, les entreprises qui font aux données d'autres entreprises (c'est le cas de Trade Republic, qui utilise des données mises à disposition par Apple) doivent se soumettre à la politique de données de leurs fournisseurs. Ainsi par exemple, Apple a récemment mis à jour sa politique de confidentialité (iOS 14.5), qui permet à chaque utilisateur d'un appareil Apple de refuser l'accès à leur identifiant publicitaire (Apple, 2021). En d'autres termes, l'utilisateur pourra refuser d'être pisté.

<i>Garantie de sécurité – Suite</i>	
Intervenant	Commentaire
C. Laimay	▸ Ce critère est l'un des plus importants. Camille s'étonne d'ailleurs que ce critère arrive en dernière position. Toutefois, notons que l'ordre dans lequel les critères sont classés est purement aléatoire. ▸ La collecte et l'utilisation des données sont soumises aux réglementations prescrites par le RGPD. Ce document s'intéresse entre autres au consentement des utilisateurs, à la collecte et gestion des cookies. Aussi, le RGPD remet en cause la manière dont les entreprises gèrent l'abandon d'un utilisateur (ex. désinscription).
E. de Bellefroid	▸ Les données sont stockées dans des silos de données propres à chaque enseigne, tous extrêmement sécurisés. Effectivement, il ne s'agit pas que d'autres acteurs s'approprient la connaissance stratégique du client appartenant aux entreprises qui utilisent les services de Freedelity. ▸ Les données de Freedelity sont stockées dans un datacenter sécurisé. Seuls trois techniciens nommés de l'entreprise ont accès aux serveurs de stockage. Ces serveurs sont d'ailleurs hébergés dans deux endroits spécifiques ▸ Des audits de sécurité des données sont effectués occasionnellement pour garantir que ces données sont bel et bien sous protection, en l'échange d'une certification ISO, nécessaire pour prétendre à un stockage de données sécurisé. ▸ Freedelity fait appel aux services d'un <i>Data Protection Officer</i> indépendant, spécialiste en sécurité informatique. Dans le cadre de sa mission, cet individu a audité l'ensemble des systèmes de stockage de Freedelity. D'autres personnes ont également procédé à l'audit de ces infrastructures, notamment les prestataires internes des entreprises clientes chez Freedelity. ▸ Tous les flux de données vers les États-Unis sont entièrement bloqués par un dispositif de sécurité informatique avancé.
A. El Maghraoui	La garantie de sécurité est primordiale : RGPD, sécurité des bases de données, sécurité des centres de données, sécurité des infrastructures Cloud.

4.3. Confrontation des sous-items associés à chaque critère avec Ali El Maghraoui

Cette section vise à comparer les sous-items du tableau de synthèse 2.3.1.1 (*cf.* Section 2.3.1) avec l'avis d'expertise de l'un de nos intervenants, Ali El Maghraoui, consultant chez EY. Durant la démarche qualitative, nous avons fourni à l'intervenant les différents sous-items relevés dans la revue de la littérature, qui en retour commentait chacun d'entre eux. Le tableau 4.2.1 ci-dessous fait état des différentes modifications survenues à la suite de l'échange conduit avec l'intervenant.

Critère	Sous-critères associés Source : Batini & Scannapieco, 2016 ; Di Ruocco <i>et al.</i> , 2012 ; Eckerson, 2002 ; Loshin, 2011 ; Mahanti, 2018 ; Zhang <i>et al.</i> , 2019	Commentaires Source : Propos de l'expert Ali El Maghraoui, consultant chez EY (<i>cf.</i> Annexe 7.5, p. 88).
<i>Complétude</i>	▸ Présence de tous les éléments essentiels ▸ Degré de complétude requis	Le premier item est conservé. Insistons sur l'importance du terme "essentiels". Le second item renvoie surtout à la manière dont les résultats seront représentés. Il renvoie au seuil au-delà duquel on considèrera que la base de données marketing est complète.
<i>Pertinence</i>	▸ Capacité de répondre à l'objectif visé par l'utilisation des données ▸ Profondeur et étendue de données au regard de l'objectif visé	Les deux items sont validés par l'expert, toutefois ce dernier estime (personnellement) que ce critère n'est pas à considérer comme un critère de qualité de stockage de données : pourquoi le stocker si la donnée n'est pas pertinente pour l'utilisation qui en sera faite ?
<i>Exactitude & Précision (Accuracy)</i>	▸ Représentation précise de la réalité ▸ Niveau de détail requis ▸ Cohérence de l'évolution des relations entre entités de données	Les deux premiers items sont validés. Un troisième item est ajouté. En réalité, celui-ci est déplacé du macro-critère <i>Intégrité</i> vers celui-ci.
<i>Conformité</i>	▸ Formatage de données ▸ Standardisation de données ▸ Correspondance avec les métadonnées ▸ Validité des données (aspect technique)	Les deux premiers items sont validés. Un troisième item est ajouté à la liste. La validité des données renvoie au contrôle technique du stockage des données. Il s'agit là de la dimension technique du macro-critère <i>Conformité</i> .
<i>Consistance & Cohérence (Consistency)</i>	▸ Homogénéité des représentations dans l'environnement de données ▸ Absence de contradictions	Le premier item n'est pas validé. En effet, les représentations des données (c'est-à-dire la manière dont elles apparaissent dans la base de données) concernent uniquement l'aspect design du stockage des données. Peu importe qu'une date de naissance soit renseignée comme 08/11/1998 ou 1998/11/08. Néanmoins, ce qui compte sera surtout de s'assurer de l'absence de contradictions, d'où l'importance du second item, qui lui est validé.
<i>Récence</i>	▸ Actualisation de données ▸ Fréquence de mise à jour des données	Tous les items sont validés par l'expert.
<i>Unicité</i>	▸ Représentation unique des données ▸ Absence de répétitions	Tous les items sont validés par l'expert.

<i>Suite</i>			
Critère		Sous-critères associés	Commentaires
		Source : Batini & Scannapieco, 2016 ; Di Ruocco <i>et al.</i> , 2012 ; Eckerson, 2002 ; Loshin, 2011 ; Mahanti, 2018 ; Zhang <i>et al.</i> , 2019	Source : Propos de l'expert Ali El Maghraoui, consultant chez EY (<i>cf.</i> Annexe 7.5, p. 88).
<i>Accessibilité & Disponibilité</i>		▸ Facilité d'accès ▸ Facilité de compréhension ▸ Facilité d'utilisation ▸ Disponibilité de la documentation sur les données	Tous les items sont validés par l'expert.
<i>Intégrité</i>		▸ Présence de tous les liens relationnels essentiels ▸ Validité des liens relationnels dans l'environnement de données ▸ Cohérence de l'évolution des relations entre entités de données	Les deux premiers items sont validés par l'expert. Néanmoins, le dernier item est supprimé de la liste. En réalité, il trouve plus sa place dans le macro-critère <i>Exactitude & Précision (Accuracy)</i> .
Légende			
texte	Le sous-critère (ou macro-critère) n'est pas validé par l'intervenant		texte Le sous-critère est ajouté ou déplacé d'un macro-critère à un autre

Tableau 4.2.1. Qualité des données : revue des sous-dimensions par critère de qualité de stockage avec Ali El Maghraoui (consultant chez EY) sur la base du tableau de synthèse 2.3.1.1. p. 18.

4.4. Conclusions relatives à l'étude qualitative

Avant de se plonger dans la synthèse des observations développées ci-avant, nous nous attarderons brièvement sur les trois critères nouvellement identifiés grâce à cette étude qualitative. Ces nouveaux critères s'insèrent tous dans la première dimension, qui concerne la qualité des données à proprement parler.

La stabilité constitue le premier critère. Selon Ali El Maghraoui, intervenant directeur exécutif chez EY, cette caractéristique permet de s'assurer qu'à travers le temps, il n'y ait pas de variations majeures constatées dans la base de données. Elle est également soulignée par Marguerite Paris, qui estime qu'une donnée de qualité est une donnée qui ne change pas trop rapidement ni brusquement.

Par ailleurs, les deux autres concernent l'automatisation du processus de collecte, ainsi que la source dont provient la donnée. En réalité, ces deux caractéristiques concernent surtout les métadonnées, qui pour rappel renvoient aux données sur les données (*cf. supra*). Effectivement, l'automatisation rend simplement compte de la manière dont la donnée a été collectée

(intervention humaine vs. processus de collecte automatisé), et constitue ainsi un élément d'information concernant la donnée elle-même. Il en va de même pour la source dont provient la donnée stockée. Elle reste certes garante de la qualité des données, mais fait également office d'information qui concerne la donnée elle-même. Dès lors, nous regrouperons ces éléments en un seul critère, que nous appellerons *Documentation*. En effet, pour reprendre les propos de Marguerite Paris, plus une donnée sera documentée, plus celle-ci sera de qualité. Camille Laimay confirme : accéder aux informations qui indiquent par exemple la source dont proviennent les données permet de se faire une idée de la qualité de celles-ci, ou encore des protocoles de collecte qui ont été appliqués.

La liste de critère est ainsi revue : les critères *Automatisation* et *Source de la donnée* disparaissent de la première dimension de la liste, tous deux couverts sous le nouveau critère *Documentation*. Ces changements sont visibles dans le tableau 4.3.1.1 de la prochaine section.

4.4.1. Synthèse des observations

	Intervenant Critère	Jean-Christophe Dewitte <i>Orange</i>	Marguerite Paris <i>Trade Republic</i>	Camille Laimay <i>Weborama</i>	Éric de Bellefroid <i>Freedelity</i>	Ali El Maghraoui <i>EY</i>	Occurrences	
Qualité des données stockées	Complétude			✓	✓	✓	3/5	60 %
	Pertinence	✓	✓	✓	✓		4/5	80 %
	Exactitude et précision (<i>Accuracy</i>)	✓		✓	✓	✓	4/5	80 %
	Conformité	✓	✓	✓	✓	✓	5/5	100 %
	Consistance et cohérence (<i>Consistency</i>)	✓	✓	✓		✓	4/5	80 %
	Récence	✓	✓	✓	✓	✓	5/5	100 %
	Unicité	✓	✓	✓	✓	✓	5/5	100 %
	Accessibilité et disponibilité	✓	✓	✓	✓	✓	5/5	100 %
	Intégrité		✓	✓	✓	✓	4/5	80 %
	NEW Stabilité		✓			✓	2/5	40 %
	NEW Documentation		✓		✓	✓	3/5	60 %
Qualité du processus de stockage	Économie de ressources	✓	✓	✓	✓		4/5	80 %
	Économie d'énergie						0/5	0 %
	Garantie de sécurité		✓	✓	✓	✓	4/5	80 %
Légende								
✓	Le critère est relevé par l'intervenant				Le critère n'est pas relevé par l'intervenant			
	Le critère est soulevé à l'unanimité				Le critère est soulevé par une minorité d'intervenants			
	Le critère est soulevé par la majorité des intervenants				Le critère n'est soulevé par aucun des intervenants			

Tableau 4.3.1.1. Table de comparaison des intervenants par critère relevé dans la revue de la littérature, mais aussi ceux identifiés au cours des entretiens dans la démarche qualitative.

Le tableau 4.3.1.1 ci-avant confronte ainsi les points de vue des différents intervenants. Pour chaque critère, pour chaque interlocuteur, le tableau indique si le critère a été relevé lors de l'échange (✓). Ce tableau reprend également les critères non identifiés dans la revue de la littérature. Désormais au nombre de deux, ces critères sont signalés à l'aide de la mention NEW.

Par ailleurs, les deux dernières colonnes attribuent un score calculé en fonction du nombre de fois que le critère concerné a été débattu. La sous-colonne de gauche exprime ce résultat en termes d'occurrences (ex. 3 fois sur 5), la sous-colonne de droite transforme cette valeur en un pourcentage (ex. 60 %). Ce score permettra de déterminer si oui ou non nous décidons de conserver le critère dans la liste fournie au terme de ce mémoire. Pour cela, nous conviendrons de quelques règles de gestion. Ces règles sont consultables dans le tableau 4.3.1.2 ci-dessous, et prennent pour fondement le nombre de fois qu'un critère a été relevé.

	Score	Commentaire	Décision
1	Score équivalent à 100 %	Le critère est soulevé à l'unanimité	Critère conservé
2	Score équivalent à 60 ou 80 %	Le critère est soulevé par la majorité des intervenants	Critère conservé
3	Score équivalent à 40 %	Le critère est soulevé par une minorité d'intervenants	Critère conservé
4	Score équivalent à 0 %	Le critère n'est soulevé par aucun des intervenants	Critère rejeté

Tableau 4.3.1.2. Règles de gestion.

Ainsi, les trois premières règles impliquent la validation du critère dans notre liste finale, puisqu'au moins deux intervenants sur les cinq sollicités soulignent l'importance du critère concerné. Quant à la dernière règle, celle-ci induit le rejet du critère dont il est question, étant donné qu'aucun interlocuteur ne l'aura relevé lors des échanges.

4.4.2. Résultats et classification des critères par ordre d'importance

Au vu des résultats représentés dans le tableau 4.3.1.1, et en se tenant aux règles établies dans le tableau 4.3.1.2, il semble que tous les critères trouvent leur place dans notre liste finale, à l'exception de l'un d'entre eux. Il s'agit du critère *Économie d'énergie*. En effet, ce critère n'est relevé par aucun des intervenants, et obtient ainsi un score nul. D'après la règle de gestion n°4

(cf. Section 4.3.1, Tableau 4.3.1.2), un critère non relevé dans la démarche qualitative, ayant ainsi obtenu un score de 0 %, se voit retiré du panel initial de critères. En se fondant sur les résultats de la démarche qualitative, il semble que ce critère énergétique n'impacte en rien la qualité du processus de stockage des données. En effet, d'après Ali El Maghraoui, stocker la donnée est négligeable d'un point de vue énergétique. En revanche, l'impact énergétique se ressent surtout au travers de la puissance de calcul utilisée dans les processus d'extraction, de transformation et d'importation/de chargement de ces données. L'intervenant cite d'ailleurs l'exemple de la cryptomonnaie. Ce n'est pas le fait de la stocker qui fait qu'elle est énergivore, c'est plutôt tout le processus de miner qui coûte de l'énergie. Ce critère énergétique dépendra de l'utilisation que l'on fait de la donnée, et de comment celle-ci est calculée, plutôt que la manière dont cette information est stockée. En outre, ce score nul pourrait également s'expliquer entre autres par l'avènement récent des technologies *Cloud*, qui proposent des infrastructures de stockage intégralement virtualisées, agiles et flexibles (O'Reilly, 2017). Cet aspect sera couvert au travers du cinquième chapitre.

Dès lors, nous sommes en mesure de pouvoir finalement fournir un classement de ces différents critères en fonction de leur importance. Ce niveau d'importance est déterminé au travers du score attribué à chacune des caractéristiques. Pour rappel, ce classement (visible dans le tableau 4.3.2.1 ci-dessous) prend pour critère de choix le nombre de fois que le critère a été relevé durant les entrevues de la démarche qualitative. La prudence reste toutefois de mise dans nos conclusions étant donné le faible échantillon d'experts sollicités.

Ainsi, les critères qui concernent la conformité, la récence, l'unicité, l'accessibilité et la disponibilité des données sont considérés comme de haute importance étant donné leur score à hauteur de 100 % des occurrences. Par ailleurs, complétude, pertinence, exactitude et précision (*Accuracy*), consistance et cohérence (*Consistency*), intégrité, documentation, économie de ressources et garantie de sécurité des données ayant obtenu un score entre 60 et 80 %, ceux-ci seront considérés comme d'importance moyenne. Pour finir, la stabilité des données reste de faible importance, avec un score équivalent à 40 %.

Importance	Qualité des données	Qualité du processus de stockage
Haute	▸ Conformité ▸ Récence ▸ Unicité ▸ Accessibilité et disponibilité	-
Moyenne	▸ Complétude ▸ Pertinence ▸ Exactitude et précision (<i>Accuracy</i>) ▸ Consistance et cohérence (<i>Consistency</i>) ▸ Intégrité ▸ Documentation	▸ Économie de ressources ▸ Garantie de sécurité
Faible	▸ Stabilité	-

Tableau 4.3.2.1. Les critères d'un stockage de données marketing de qualité sur base (i) d'une revue de la littérature, (ii) d'une étude qualitative menée auprès de cinq experts intervenants.

4.5. Limites dans la démarche qualitative

Il est important de garder à l'esprit que les conclusions tirées au terme de la démarche qualitative se fondent sur l'intervention de cinq experts du domaine. Il est évident que dans le cas où des intervenants supplémentaires (ou différents de ceux ayant accepté de participer à l'étude) avaient été sollicités, les résultats (*cf.* Tableau 4.3.1.1) et le classement des macro-critères (*cf.* Tableau 4.3.2.1) eurent pu s'avérer radicalement différentes. En effet, les décisions de rejeter un critère ou non ont été fondées sur des règles de gestion (*cf.* Tableau 4.3.1.2), elles-mêmes construites sur la base des scores attribués aux différents macro-critères (*cf.* Tableau 4.3.1.1).

Par ailleurs, nous aurions pu nous intéresser plus en profondeur à la qualité du processus de stockage de données en prenant part à quelques visites de centres de données privés afin d'étayer nos conclusions liées à la qualité du processus de stockage en masse de données. Aussi, une recherche de type quantitatif sous la forme d'une enquête à questions fermées menée auprès d'autres experts eut été judicieuse. Ce type de recherche permet de couvrir un plus large échantillon d'enquêtés (par exemple, contacter des experts à l'échelle internationale), mais fournit des réponses rigides, parfois peu profondes, et difficiles à interpréter (Pleyers, 2020).

Chapitre 5 : Réflexion sur l'aspect sociétal du stockage de données

Nous avons ainsi couvert au travers des parties précédentes les apports scientifique et managérial de la thématique de ce mémoire. Ainsi scientifiquement, nous avons eu l'occasion de compléter et corriger une liste de plusieurs critères de qualité de stockage de données grâce à des entretiens qualitatifs conduits auprès de cinq intervenants experts. En des termes managériaux, nous avons mis en avant une série de points d'attention et proposé des recommandations à destination des entreprises, comme par exemple l'attention constante à la mise à jour des données ou la lutte constante contre les doublons.

Si le critère de l'économie d'énergie n'a pas été soulevé par nos interlocuteurs (score nul), c'est certainement que tous sont des utilisateurs des données, sans intervention aucune dans le processus du stockage de celles-ci. Ainsi, dans ce chapitre, nous développerons une brève réflexion sur l'aspect sociétal (entre autres énergétique) du stockage en masse de données. Nous insistons sur le caractère tout à fait **personnel** de cette réflexion.

Actuellement, il existe deux solutions de stockage de données en masse : les centres de données (dont nous avons parlé dans le second chapitre) et la très récente technologie du *Cloud*. Cette dernière solution propose des infrastructures de stockage entièrement virtualisées, agiles et flexibles (O'Reilly, 2017). En d'autres termes, les données sont stockées et accessibles via des serveurs à distance sur Internet.

5.1. Impact environnemental mitigé du stockage des données

5.1.1. Augmentation de la consommation en énergie des *datacenters*

En 2019, Data Center Map recensait près de 4.800 centres de données dans le monde (Data Center Map, 2021). A l'heure actuelle, il en existe près de 8.000 à l'échelle globale (Statista, 2021b), soit une croissance d'environ 67 % des installations en deux ans. Cette statistique est surprenante, voire inquiétante. Ces dispositifs de stockage en masse de données consomment une très large quantité d'énergie, constituant ainsi une grave menace pour l'environnement. En effet, selon une étude menée par l'association Négawatt (2017), la consommation énergétique des centres de données situés en France s'élevait, en 2015, à près de 10 TWh (i.e. 10.000.000.000.000 Wh), soit un peu plus de 2 % de la consommation électrique nationale

française. A l'échelle globale, les datacenters nécessitent aujourd'hui 3 % de la production électrique mondiale (McCormick, 2020). Pour quelle raison ? Outre l'énergie électrique nécessaire au bon fonctionnement des centres et de leurs milliers de serveurs (alimentation, éclairage, sécurité), ceux-ci dégagent une quantité immense de chaleur qui doit impérativement être contrée et évacuée, ce le plus rapidement possible. Sans nul doute, ces dégagements de chaleur expliquent partiellement le réchauffement climatique. Ainsi, une grande partie du montant de la facture d'électricité d'un *datacenter* se justifiera par les systèmes de refroidissement (climatisation). C'est d'ailleurs la raison pour laquelle des grandes entreprises localisent stratégiquement leurs centres de données dans des endroits peu chauds (ex. Facebook dans les pays nordiques) ou proches de sources naturelles de refroidissement (ex. Google en Belgique – cf. *supra*). Enfin, comme tout matériau, les serveurs et autres outils de stockage tendent à s'user, jusqu'à devenir un déchet au terme de leur vie, déchet qui ne disparaîtra jamais réellement...

5.1.2. Amélioration de l'efficacité énergétique du *Cloud* au regard des *datacenters*

En revanche, les technologies du *Cloud* ont de meilleures répercussions sur l'environnement. De fait, cette solution en ligne nécessiterait moins d'électricité étant donné l'absence de site. Pas d'éclairage ou de refroidissement des infrastructures de stockage, ce qui représente une quantité considérable d'énergie (pas de climatisation, facteur de première ligne). Solution de stockage dématérialisée dans son entièreté, le *Cloud* permet ainsi de réduire les émissions de gaz à effet de serre (Carbon Disclosure Project, 2013), les déchets (entre autres économie de papier grâce aux solutions à l'instar de Dropbox, Google Drive ou encore OneDrive) ainsi que la consommation d'énergie. L'empreinte environnementale du *Cloud* semble ainsi inférieure à celle d'un centre de données ordinaire sur site. Attention, ce dernier paragraphe ne tend pas à faire croire que le *Cloud* est énergétiquement neutre, mais bien énergétiquement **plus efficace** qu'un datacenter.

5.2. Impact social du stockage des données

Par sociétal, nous n'entendons pas uniquement environnement. Il est également question de la société, de la population, dans notre cas les propriétaires des données marketing stockées et utilisées par les entreprises, qu'il s'agisse de données primaires (*first-party*) ou tierces (*third-party*) (cf. *supra*).

5.2.1. Remise en question du fonctionnement des entreprises en matière de protection de la vie privée de leurs utilisateurs

Nous en avons parlé dans le chapitre deux, la vie privée des individus est aujourd'hui protégée par le Règlement général sur la protection des données (RGPD, *cf. supra*). Ce document législatif européen contraint entre autres les entreprises à accorder aux individus consommateurs (ou prospects) la possibilité d'accepter ou de refuser la collecte, le stockage et l'utilisation de leurs données personnelles à n'importe quel moment. Autrement dit, comment gérer l'abandon soudain d'un individu dont une entreprise dispose des données personnelles ? C'est là tout l'enjeu. Or, à l'heure actuelle, une étude conduite par le réseau de cabinets d'audit RSM (2019) montre que près d'un tiers (30 %) des entreprises européennes avouent ne pas être conformes aux directives du règlement. Par ailleurs, seuls 57 % des entreprises estiment être en règle, tandis que 13 % doutent de leur conformité¹¹. Ces statistiques ne sont pas surprenantes, dans la mesure où ces remises en question peuvent entraîner des coûts parfois trop importants.

5.2.2. Méfiance des utilisateurs envers les entreprises

Ces derniers chiffres ne rassurent en rien. Certains restent en effet méfiants quant à la divulgation de leurs données personnelles. Rien d'étonnant au regard des scandales liés à la divulgation de données supposément privées, tels celui qui concerne la tristement célèbre affaire Cambridge Analytica. Pour rappel, cette entreprise britannique s'était intéressée aux données de millions d'individus à leur insu pour nul autre que l'ex-président américain Donald Trump dans le cadre de sa campagne électorale de 2016 (Audureau, 2018). Cette méfiance se répercute par ailleurs sur la performance marketing des entreprises. Il est effectivement des personnes qui renseignent volontairement une fausse adresse de courrier électronique ou encore un faux numéro de téléphone dans un formulaire d'inscription en ligne (par exemple) pour se prémunir de l'utilisation qui en sera faite ultérieurement. 36 % des Allemands et 34 % des Britanniques déclarent avoir donné de faux détails en ligne afin d'éviter de divulguer leurs informations personnelles (Symantec, 2015). Cependant, une entreprise ne pourra jamais se satisfaire de données incorrectes. La confiance envers les entreprises n'est ainsi pas encore au rendez-vous... D'ailleurs, c'est à se demander si elle le sera un jour.

¹¹ *Remarque.* La majorité des entreprises interrogées dans le cadre de cette étude sont des entreprises européennes provenant de 34 pays, de taille moyenne, et dont le chiffre d'affaires annuel n'excède pas 100 millions d'euros.

Chapitre 6 : Conclusion

Nous nous sommes ainsi intéressés à la qualité du stockage des données marketing. Pour rappel, notre problématique était la suivante : “*Quels sont les critères qui font d’un stockage de données marketing un stockage de qualité ?*” Afin d’y répondre, nous avons préalablement restitué le contenu repris dans la littérature, ce afin d’établir une première liste de critères de qualité de stockage de données. Ensuite, nous avons critiqué cette liste au travers d’une étude qualitative menée auprès de cinq spécialistes du domaine, en s’intéressant à leurs expériences pratiques. A terme, la réponse à notre problématique aura pris la forme d’un tableau à double entrée dans lequel nous avons classé par ordre d’importance les critères relevés à la fois dans la littérature mais aussi au travers de la démarche empirique.

Par ailleurs, ce mémoire couvre une intention à **trois niveaux**. Le premier niveau renvoie à l’apport **scientifique** de notre rédaction. En effet, nous aurons eu l’occasion d’aborder le sujet d’étude sous tous les angles au travers du chapitre consacré à la revue de la littérature. Pour mener à bien notre recherche, nous avons ainsi couvert les concepts de Big Data, données, qualité des données et ses différentes dimensions. Parmi celles-ci, les critères qui concernent la conformité, la récence, l’unicité, l’accessibilité et la disponibilité des données sont considérés comme de première importance. Nous recommandons dès lors de prêter attention à ces critères en priorité. Bien entendu, il ne s’agit pas là des seuls points d’attention. En effet, exactitude et précision (*Accuracy*), consistance et cohérence (*Consistency*), complétude et sécurité des données sont – entre autres – à compter parmi les critères d’importance moyenne.

En outre, la seconde intention est **managériale**. Notre recherche a en effet pour finalité de proposer aux entreprises et acteurs du stockage des données une liste relativement complète et fiable de critères de qualité de données à considérer impérativement en vue d’en extraire un maximum de valeur. Le non-respect de ces caractéristiques fait effectivement survenir des anomalies au sein des bases de données. Citons par exemple les doublons. Dans le cas d’une entreprise, ces incidences induiront entre autres perte de temps, d’argent, et baisse de performance. Afin de se prémunir face à de telles anomalies, de nombreux outils existent à l’heure actuelle, à l’instar de logiciels de standardisation de données. Par ailleurs, des entreprises spécialisées dans l’audit de base de données (*data auditing*) sont disponibles pour proposer des actions correctives une fois l’environnement de stockage examiné de près. En

somme, ce ne sont pas les solutions qui manquent. L'essentiel reste surtout d'être conscient de l'importance de la qualité des données et de leur processus de stockage, de l'entretenir (il s'agit d'un travail continu), de même que des conséquences que cela peut avoir sur le succès d'une entreprise dans le cas où cette qualité n'est pas suffisamment considérée. Car oui, la majorité des actions marketing se fondent aujourd'hui sur le contenu des bases de données. S'il s'avérait qu'un consommateur soit mal enregistré, c'est avec des données erronées ou incomplètes qu'une entreprise devra travailler.

Enfin, ce mémoire tend à se positionner **partiellement** sur le plan **sociétal**. La réflexion personnelle non exhaustive proposée dans le dernier chapitre a effectivement pour ambition de montrer de manière limitée l'importance des répercussions environnementales et sociales que provoquent le stockage en masse données et plus généralement le Big Data. Les problématiques liées à la préservation de l'environnement et de ses occupants restent non négligeables, et prennent en effet une place de plus en plus prépondérante. Dès lors, nous retenons une augmentation de la consommation énergétique des centres de données (malgré une amélioration de l'efficacité énergétique des solutions *Cloud*). Aussi, le Règlement général sur la protection des données personnelles contraint depuis 2018 les entreprises à repenser leur mode de fonctionnement en matière de protection de la vie privée de leurs consommateurs. Sachant qu'une entreprise sur trois seulement déclare être conforme aux directives établies dans le texte de loi (*cf. supra*), ces chiffres relativement faibles sont source de méfiance de la part des utilisateurs, ceux-ci allant parfois jusqu'à renseigner de fausses informations personnelles pour se protéger. Cela étant, ces comportements peuvent-ils simplement se justifier par un souci de protection individuelle ? Ou pourrions-nous parler de sabotage volontaire dans un but de propagande idéologique ?

En ce qui concerne les limites de ce mémoire, rappelons que les conclusions tirées au terme de l'étude qualitative se fondent sur les expériences et propos de cinq spécialistes de la donnée. En effet, des témoignages supplémentaires ou différents auraient sans nul doute amené à d'autres scores et issues. Aussi, notons que nous nous sommes avant tout intéressés à l'importance des critères de qualité. Il eut été également bien pensé d'étudier la performance de chacun d'entre eux, au travers d'un audit de base de données par exemple. Cela étant, cette démarche aurait selon nous conduit à plusieurs difficultés en termes de faisabilité. De plus, une recherche

quantitative sous la forme d'une enquête menée auprès d'autres experts (par exemple) eut été intéressante en vue d'en arriver à des conclusions moins subjectives.

En bref, si une donnée est de qualité lorsqu'elle répond aux besoins des acteurs qui en font usage (Brasseur, 2005), nous ne devrions retenir qu'une chose : il est inutile de stocker une donnée et de la conserver dans un environnement de stockage si, à terme, aucun usage n'en sera fait. Une pareille décision est dénuée de tout sens : capacité de stockage occupée inutilement, énergie, argent et temps dépensés sans succès... D'ici 2035, le volume global de données créées devrait atteindre 2.000 zetaoctets, soit 40 fois plus que le volume enregistré l'an dernier (Statista, 2021a). Au vu de cette croissance exponentielle de données produites, les entreprises n'auront nul autre choix que d'éviter ces démarches vaines à l'avenir.

Bibliographie

- ACCURITY. (2021). *Data quality monitoring*. Retrieved June 6th, 2021 from <https://www.accuracy.ai/solutions/data-quality-monitoring/>
- ADSQUARE. (2021). *Adsquare is helping brands connect people and places in a privacy-first world*. Retrieved June 4th, 2021 from <https://www.adsquare.com>
- ALWAN, H. B., & KU-MAHAMUD, K. R. (2019, Sep. 25-27th). *Big Data: definition, characteristics, life cycle, applications and challenges*. The 6th International Conference on Software Engineering & Computer Systems, Pahang, Malaysia. doi: 10.1088/1757-899X/769/1/012007
- APPLE. (2021). *Politique de confidentialité d'Apple*. Retrieved June 22nd, 2021 from <https://www.apple.com/ca/fr/legal/privacy/fr-ca/>
- AUDUREAU, W. (2018). Ce qu'il faut savoir sur Cambridge Analytica, la société au cœur du scandale Facebook. Retrieved August 2nd, 2021 from https://www.lemonde.fr/pixels/article/2018/03/22/ce-qu-il-faut-savoir-sur-cambridge-analytica-la-societe-au-c-ur-du-scandale-facebook_5274804_4408996.html
- AUTORITÉ DE PROTECTION DES DONNÉES. (2021). *Organisation*. Retrieved May 13th, 2021 from <https://autoriteprotectiondonnees.be/citoyen/l-autorite/organisation>
- BATHELOT, B. (2015). Conversion. In *L'Encyclopédie illustrée du marketing*. Retrieved May 13th, 2021 from <https://www.definitions-marketing.com/definition/conversion/>
- BATHELOT, B. (2016). Base de données marketing. In *L'Encyclopédie illustrée du marketing*. Retrieved May 19th, 2021 from <https://www.definitions-marketing.com/definition/base-de-donnees-marketing/>
- BATHELOT, B. (2018a). Données clients. In *L'Encyclopédie illustrée du marketing*. Retrieved May 18th, 2021 from <https://www.definitions-marketing.com/definition/donnees-client/>
- BATHELOT, B. (2018b). Persona en marketing. In *L'Encyclopédie illustrée du marketing*. Retrieved June 16th, 2021 from <https://www.definitions-marketing.com/definition/persona-en-marketing/>

- BATINI, C., & SCANNAPIECO, M. (2016). *Data and Information Quality: Caractéristiques, Principles and Techniques*. Switzerland: Springer.
- BERTI-EQUILLE, L. (2004). Qualité des données. *Ingénierie des Systèmes d'Information*, 9(5-6), pp. 117-143. doi: 10.3166/isi.9.5-6.117-143
- BERTI-EQUILLE, L. (2017). La qualité des données : champs de recherche et enjeux. In M. Bouzeghoub & R. Mosseri (Eds.), *Les Big Data à découvert* (pp. 64-65). Paris, France: CNRS Éditions.
- BIBARD, O. (2003). *Avellino Discovery dresse l'état de santé des SGBDR*. Retrieved June 6th, 2021 from <https://bfmbusiness.bfmtv.com/01-business-forum/avellino-discovery-dresse-letat-de-sante-des-sgbdr-209057.html>
- BIDOIT-TOLLU, N., & DOUCET, A. (2017). La diversité des données. In M. Bouzeghoub & R. Mosseri (Eds.), *Les Big Data à découvert* (pp. 24-25). Paris, France: CNRS Éditions.
- BOCQUET, M., & MULLER, C. (2017). Stockage de l'information : de la peinture rupestre aux data centers. In M. Bouzeghoub & R. Mosseri (Eds.), *Les Big Data à découvert* (pp. 76-77). Paris, France: CNRS Éditions.
- BRASSEUR, C. (2005). *Data Management: Qualité des données et compétitivité*. Cachan, France: Lavoisier.
- BROSCHKE, K., & KUMAR, A. (2014). Realtime Data Accelerates Online Marketing. In O. Busch (Ed.), *Programmatic Advertising, The Successful Transformation to Automated, Data-Driven Marketing in Real-Time* (pp. 221-232). Switzerland: Springer.
- CAPPY, A. (2017). Impact énergétique des Big Data. In M. Bouzeghoub & R. Mosseri (Eds.), *Les Big Data à découvert* (pp. 82-83). Paris, France: CNRS Éditions.
- CARBON DISCLOSURE PROJECT. (2013). *How Environmentally Sustainable is Cloud Computing and Storage?* Retrieved July 29th, 2021 from <https://globalwarmingisreal.com/2013/09/12/sustainable-cloud-computing/>
- CASSANDRA. (2021). *What is Apache Cassandra?* Retrieved May 26th, 2021 from <https://cassandra.apache.org/cassandra-basics/>

- CHEN, H., CHIANG, R., & STOREY, V. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, 36(4), pp. 1165-1188. doi: 10.2307/41703503
- CHERUVU, R. (2015). *Big Data Applications in Self-Driving Cars* (final semester paper). Harvard University, Cambridge. doi: 10.13140/RG.2.2.12302.05445
- CHEVALIER, L. (2020). *Big Data Applied to Marketing, LLSMS2004*. Course material, Louvain School of Management, Louvain-la-Neuve.
- CHINTAGUNTA, P., HANSSENS, D. M., & HAUSER, J. R. (2016). Marketing Science and Big Data. *Marketing Science*, 35(3), pp. 341-342. doi: 10.1287/mksc.2016.0996
- COURTHEOUX, R. J. (2003). Marketing data analysis and data quality management. *Journal of Targeting, Measurement and Analysis for Marketing*, 11(4), pp. 299-313. doi: 10.1057/palgrave.jt.5740086
- COUSSEMENT, K., VAN DEN BOSSCHE, F. A. M., & DE BOCK, K. W. (2014). Data accuracy's impact on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees. *Journal of Business Research*, 67(1), pp. 2751-2758. doi: 10.1016/j.jbusres.2012.09.024
- DATA CENTER MAP. (2021). *Colocation Data Centers*. Retrieved August 8th, 2021 from <https://www.datacentermap.com>
- DATARADE. (2021). *Demographic Data*. Retrieved May 12th, 2021 from <https://datarade.ai/data-categories/demographic-data>
- DATARADE. (2021). *Psychographic Data*. Retrieved May 12th, 2021 from <https://datarade.ai/data-categories/psychographic-data>
- DAVENPORT, T. (2013). *Enterprise Analytics: Optimize Performance, Process and Decisions Through Big Data*. Upper Saddle River, New Jersey: Pearson Education.
- DAVENPORT, T. H., & PATIL, D. J. (2012). Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review*, 90(10), pp. 70-76. Retrieved from <https://www.hbs.edu/faculty/Pages/item.aspx?num=43110>

- DAVENPORT, T. H., BARTH, P., & BEAN, R. (2012). How 'Big Data' Is Different. *MIT Sloan Management Review* 54(1), pp. 22-24. Retrieved from <https://www.hbs.edu/faculty/Pages/item.aspx?num=43026>
- DAVOUST, A., & VOLLE, P. (2014). *La gestion des « Big data » en marketing : défis, opportunités et voies de recherche*. 30^e congrès de l'Association française de Marketing, Montpellier, France.
- DAY, G. S. (2011). Closing the Marketing Capabilities Gap. *Journal of Marketing*, 75(4), pp. 183-195. doi: 10.1509/jmkg.75.4.183
- DAYDÉ, M., & VEYNANTE, D. (2017). Les grands sites de calcul et de stockage. In M. Bouzeghoub & R. Mosseri (Eds.), *Les Big Data à découvert* (pp. 80-81). Paris, France: CNRS Éditions.
- DI RUOCCO, N., SCHEIWILER, J-M., & SOTNYKOVA, A. (2012). La qualité des données: concepts de base et techniques d'amélioration. In L. Berti-Equille (Ed.), *La qualité et la gouvernance des données au service de la performance des entreprises* (pp. 25-54). Cachan, France: Lavoisier.
- DIACONU, C., & MASSOL, M. (2017). Préservation et archivage. In M. Bouzeghoub & R. Mosseri (Eds.), *Les Big Data à découvert* (pp. 78-79). Paris, France: CNRS Éditions.
- ECKERSON, W. W. (2002). *Data Quality and The Bottom Line: Achieving Business Success through a Commitment to High Quality Data*. Report, The Data Warehousing Institute, Renton.
- ELMASRI, R., & NAVATHE, S.B. (2016). *Fundamentals of Database Systems* (7th Edition). Hoboken, New jersey: Pearson Higher Education.
- ENCYCLOPÉDIE LAROUSSE EN LIGNE. (2021). *Octet*. Retrieved May 11th, 2021 from <https://www.larousse.fr/encyclopedie/divers/octet/187782>
- ENGLISH, P. L. (2009). *Information Quality Applied: Best Practices for Improving Business Information, Processes and Systems*. New York, United States: Wiley.

- EUROPA. (2021). *Directive 95/46/CE du Parlement européen et du Conseil, du 24 octobre 1995, relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données*. Retrieved May 13th, 2021 from <https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=celex%3A31995L0046>
- EUROPA. (2021). *Règlements, directives et autres actes législatifs*. Retrieved May 13th, 2021 from https://europa.eu/european-union/law/legal-acts_fr
- EVEN, A., & SHANKARANARAYANAN, G. (2009). Dual Assessment of Data Quality in Customer Databases. *ACM Journal of Data and Information Quality*, 1(3), pp 1-29. doi: 10.1145/1659225.1659228
- EVEN, A., SHANKARANARAYANAN, G., & BERGER, P. D. (2007). Assessing data quality: A value-driven approach. *Data Base for Advances in Information Systems*, 38(2), pp. 76-93.
- EVEN, A., SHANKARANARAYANAN, G., & BERGER, P. D. (2010). Managing the Quality of Marketing Data: Cost/benefit Tradeoffs and Optimal Configuration. *Journal of Interactive Marketing*, 24(3), pp. 209-221. doi: 10.1016/j.intmar.2010.01.001
- FURNER, J. (2020). Definitions of “Metadata”: A Brief Survey of International Standards. *Journal of the Association for Information Science and Technology*, 71(6), pp. 33-42. doi: 10.1002/asi.24295
- GANDOMI, A., & HAIDER, M. (2014). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), pp. 137-144. doi: 10.1016/j.ijinfomgt.2014.10.007
- GERTZ, O., & MCGLASHAN, D. (2014). Consumer-Centric Programmatic Advertising. In O. Busch (Ed.), *Programmatic Advertising, The Successful Transformation to Automated, Data-Driven Marketing in Real-Time* (pp. 55-74). Switzerland: Springer.
- GIANELLONI, J-L., & LE NAGARD, E. (2016). Big Data et marketing prédictif: Vers un “consommateur nu”? *Décisions Marketing*, 82, pp. 5-15. doi: 10.7193/DM.082.05.15
- GOOGLE. (2021a). *Centres de données: Saint-Ghislain, Belgique*. Retrieved May 26th, 2021 from <https://www.google.com/intl/fr/about/datacenters/locations/st-ghislain/>

- GOOGLE. (2021b). *One percent of Googlers get to visit a data center, but I did*. Retrieved May 26th, 2021 from <https://blog.google/inside-google/infrastructure/how-data-center-security-works/>
- GOOGLE. (2021c). *La vie dans les centres de données Google*. Retrieved May 26th, 2021 from <https://www.google.com/intl/fr/about/datacenters/life/>
- GROTH, A., & ZAWADZKI, V. (2014). Granularity Creates Added Value for Every Objective. In O. Busch (Ed.), *Programmatic Advertising, The Successful Transformation to Automated, Data-Driven Marketing in Real-Time* (pp. 87-102). Switzerland: Springer.
- GUDIVADA, V. N., APON, A., & DING, J. (2017). Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations. *International Journal on Advances in Software*, 10(1), pp. 1-20.
- HEWLETT PACKARD. (2021). *Qu'est-ce que le stockage de données ?* Retrieved August 1st, 2021 from <https://www.hpe.com/be/fr/what-is/data-storage.html>
- HUBSPOT. (2021). *The What, Why, and How of Behavioral Marketing [+ Examples]*. Retrieved May 12th, 2021 from <https://blog.hubspot.com/marketing/what-is-behavioral-marketing>
- IBM. (2003). *Ascential Enterprise Integration Suite: A complete data integration solution designed for the right-time enterprise*. Software Announcement, United States.
- IBM. (2016). *Data profiling process*. Retrieved May 25th, 2021 from <https://www.ibm.com/docs/en/iis/11.5?topic=columns-data-profiling-process>
- IBM. (2019). *Introduction to data cleansing*. Retrieved June 6th, 2021 from <https://www.ibm.com/docs/en/iis/11.7?topic=jobs-introduction-data-cleansing>
- INFORMATICA. (2021). *PowerCenter*. Retrieved June 6th, 2021 from <https://www.informatica.com/fr/products/data-integration/powercenter.html>
- JOURNAL OFFICIEL DE L'UNION EUROPÉENNE. (2016). Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE. *L119/33, Article 4*.

- JSON. (2021). *Introducing JSON*. Retrieved May 27th, 2021 from <https://www.json.org/json-en.html>
- KAROUI, M., DAVAUCHELLE, G., & DUDEZERT, A. (2014). Big Data: Mise en perspective et enjeux pour les entreprises. *Ingénierie des Systèmes d'Information*, 19(3), pp. 73-92. doi: 10.3166/isi.19.3.73-92
- KHAN, S., LIU, X., ALI, S. A., & ALAM, M. (2019). *Storage Solutions for Big Data Systems: A Qualitative Study and Comparison*. <https://arxiv.org/abs/1904.11498>
- LAMBIN, J-J., & DE MOERLOOSE, C. (2021). *Marketing stratégique et opérationnel : La démarche marketing dans une perspective responsable* (10^e éd.). Malakoff, France : Dunod.
- LERUM, K. (2014). *Deductive & Inductive Research* [Video]. YouTube. Retrieved June 2nd, 2021 from <https://www.youtube.com/watch?v=gctAuZGrdbQ>
- LOSHIN, D. (2011). *The Practitioner's Guide to Data Quality Improvement*. Burlington, United States: Elsevier.
- LOTAME. (2019). *1st Party Data, 2nd Party Data, 3rd Party Data: What Does It All Mean?* Retrieved May 13th, 2021 from <https://www.lotame.com/1st-party-2nd-party-3rd-party-data-what-does-it-all-mean/#what-is-2nd-party-data>
- MAHANTI, R. (2018). *Data Quality: Dimensions, Measurement, Strategy, Management, and Governance*. Milwaukee, Wisconsin, United States: ASQ Quality Press.
- MARKARIAN, R., PARK, A-N., & GREYER, M. (2014). Borderless Media Management. In O. Busch (Ed.), *Programmatic Advertising, The Successful Transformation to Automated, Data-Driven Marketing in Real-Time* (pp. 17-24). Switzerland: Springer.
- MARR, B. (2015). *Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance*. Chichester, West Sussex, United Kingdom: Wiley
- MAYDANCHIK, A. (2007). *Data Quality Assessment*. New Jersey, United States: Technics Publications.

- MCCORMICK, P. (2020). *Le Cloud est-il plus écoresponsable ?* Retrieved July 29th, 2021 from <https://valiantys.com/fr/blog/itsm-fr/le-cloud-est-il-plus-eco-responsable/>
- MOHANTY, S., JAGADEESH, M., & SRIVATSA, H. (2013). *Big Data Imperatives: Enterprise Big Data Warehouse, BI Implementations and Analytics*. New York, United States: Springer.
- MONGODB. (2021). *The database for modern applications*. Retrieved May 26th, 2021 from <https://www.mongodb.com>
- NAIR, R., & NARAYANAN, A. (2012). *Benefitting from Big Data: Leveraging Unstructured Data Capabilities for Competitive Advantage*. Report, Booz & Company, United States.
- NEGAWATT. (2017). *La révolution numérique fera-t-elle exploser nos consommations d'énergie ?* Retrieved July 29th, 2021 from <https://decrypterlenergie.org/la-revolution-numerique-fera-t-elle-exploser-nos-consommations-denergie>
- NEO4J. (2021). *We Help the World Make Sense of Data*. Retrieved May 26th, 2021 from <https://neo4j.com/company/>
- NIELSEN. (2021). *About Nielsen*. Retrieved June 4th, 2021 from <https://www.nielsen.com/us/en/about-us/>
- NORMAD1. (2021). *Restructuration, normalisation et validation postale*. Retrieved June 6th, 2021 from <https://www.normad1.fr/fr/solutions/accueil-rnvp-restructuration-normalisation-validation-postale>
- O'REILLY, J. (2017). *Network Storage: Tools and Technologies for Storing Your Company's Data*. Cambridge, United States: Elsevier.
- ORACLE. (2021a). *Qu'est-ce qu'une base de données relationnelle?* Retrieved May 26th, 2021 from <https://www.oracle.com/be-fr/database/what-is-a-relational-database/>
- ORACLE. (2021b). *Oracle Advertising (Oracle Data Cloud)*. Retrieved June 4th, 2021 from <https://www.oracle.com/cx/advertising/>

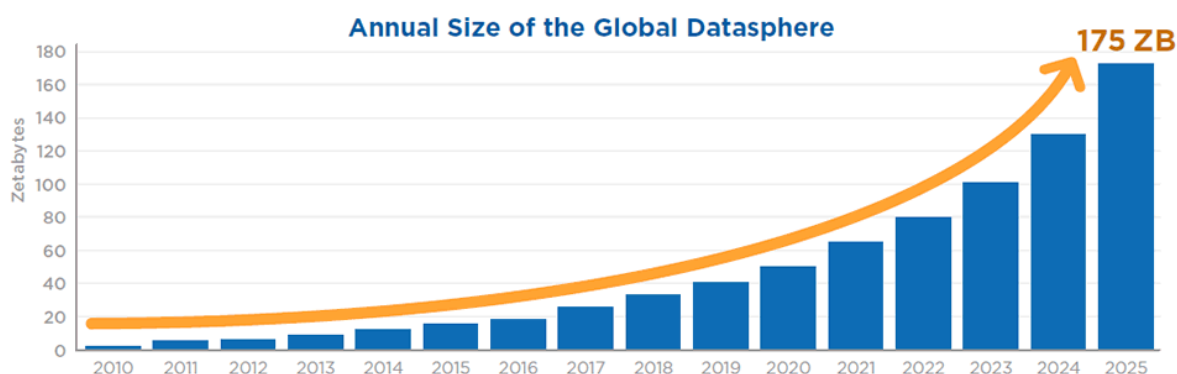
- PÊCHEUX, F. (n.d.). Big Data. In *Encyclopædia Universalis*. Retrieved May 10th, 2021 from <https://www.universalis.fr/encyclopedie/big-data/>
- PLEYERS, G. (2020). *Marketing Research, LLSMS2000*. Course material, Louvain School of Management, Louvain-la-Neuve.
- PROJECT VOLDEMORT. (2021). Project Voldemort: A distributed database. Retrieved May 26th, 2021 from <https://www.project-voldemort.com/voldemort/>
- RAGHAV, R. S., AMUDHAVAL, J., & DHAVACHELVAN, P. (2018). A Survey of NoSQL Database For Analyzing Large Volume Of Data In Big Data Platform. *International Journal of Engineering & Technology*, 7(2.32), pp. 181-186. doi: 10.14419/ijet.v7i2.32.15563
- REDHAT. (2021). *Qu'est-ce qu'une API ?* Retrieved June 22nd, 2021 from <https://www.redhat.com/fr/topics/api/what-are-application-programming-interfaces>
- RSM. (2019). *30% of European businesses are still not compliant with GDPR*. Retrieved August 1st, 2021 from <https://www.rsm.global/news/30-european-businesses-are-still-not-compliant-gdpr>
- RUSSOM, P. (2011). *Big Data Analytics: Best Practices Report, Fourth Quarter*. Report, The Data Warehousing Institute, Renton, United States.
- SALESFORCE. (2021a). *CRM 101: What is CRM?* Retrieved May 11th, 2021 from <https://www.salesforce.com/crm/what-is-crm/>
- SALESFORCE. (2021b). *Predictive Marketing & Why You Should Look Into It*. Retrieved May 19th, 2021 from <https://www.salesforce.com/products/marketing-cloud/best-practices/predictive-marketing/>
- SAMAIN, M. (2020). *Google va construire un 5e data center en Belgique*. Retrieved May 26th, 2021 from <https://www.lecho.be/entreprises/technologie/google-va-construire-un-5e-data-center-en-belgique/10267250.html>

- SAMITSCH, C. (2015). *Data Quality and its Impacts on Decision-Making: How Managers can benefit from Good Data* (Master's Thesis). Management Center Innsbruck, Innsbruck, Austria. doi: 10.1007/978-3-658-08200-0
- SAS. (2021). *About SAS and DataFlux*. Retrieved June 6th, 2021 from https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/whatsdiff/n0k0t20kfebo9n1ktvgchplt1yk.htm
- SATHI, A. (2014). *Engaging Customers Using Big Data: How Marketing Analysis Are Transforming Business*. New York: Palgrave Macmillan.
- SAUNDERS, M., LEWIS, P., & THORNHILL, A. (2016). *Research methods for business students* (7th Ed). Edinburgh, England: Pearson.
- SCHWAB, P-N. (2016). *Obsolescence des données : quelle est votre stratégie ?* Retrieved May 19th, 2021 from <https://www.intotheminds.com/blog/obsolescence-des-donnees-quelle-est-votre-strategie/>
- SIDDIQA, A., KARIM, A., & GANI, A. (2017). Big data storage technologies: a survey. *Frontiers of Information Technology & Electronic Engineering*, 18(8), pp. 1040–1070. doi: 10.1631/FITEE.1500441
- STATISTA. (2021a). *Le big bang du big data*. Retrieved May 26th, 2021 from <https://fr.statista.com/infographie/17800/big-data-evolution-quantite-donnees-numeriques-creees-dans-le-monde/>
- STATISTA. (2021b). *Data centers : les pays les mieux équipés*. Retrieved July 29th, 2021 from <https://fr.statista.com/infographie/24147/pays-avec-le-plus-de-data-centers-centres-de-donnees/>
- STROHBACH, M., DAUBERT, J., RAVKIN, H., & LISCHKA, M. (2016). Big Data Storage. In J. M. Cavanillas, E. Curry, W. Wahlster (Eds.), *New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe* (pp. 119-142). Switzerland: Springer.

- SYMANTEC. (2015). *State of Privacy Report*. Retrieved from <https://www.nortonlifelock.com/content/dam/nortonlifelock/pdfs/reports/state-of-privacy-report-en-2015.pdf>
- UNIVERSITÉ LIBRE DE BRUXELLES. (2021). *Introduction au langage xml*. Retrieved May 27th, 2021 from <https://www2.ulb.ac.be/demo/introduction-xml/index.html>
- WEBORAMA. (2020). *A propos de Weborama*. Retrieved June 4th, 2021 from <https://weborama.com>
- WITSENBURG, P. (2016). *Centre de données en Belgique: L'accès numérique au cloud*. Retrieved May 26th, 2021 from <https://cloudmakelaar.be/wp-content/uploads/2017/01/De-Digitale-hub-Datacenters-in-België-deel-1-3-FR-e-paper.pdf>
- ZHANG, R., INDULSKA, M., & SADIQ, S. (2019). Discovering Data Quality Problems. *Business & Information Systems Engineering*, 61(5), pp. 575–593. doi: 10.1007/s12599-019-00608-0

Annexes

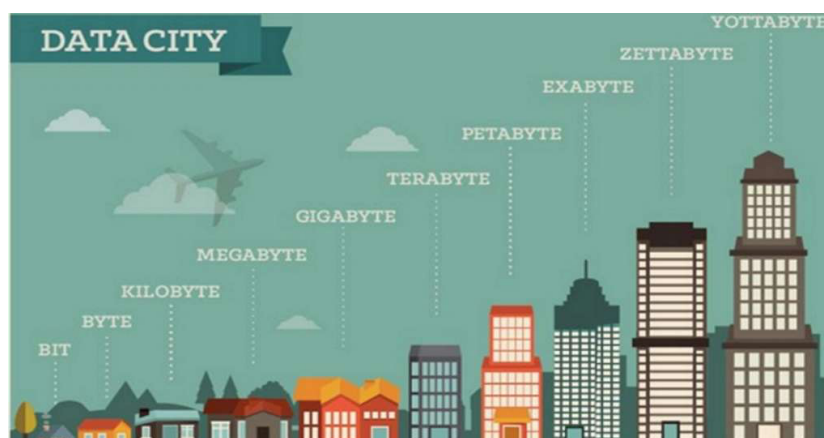
Annexe 1. Datasphère à l'échelle mondiale par année



Source : https://www.researchgate.net/figure/Annual-Size-of-the-Global-Datasphere_fig1_346973885

Remarque : 1 ZB (Zettabyte) = 10^{21} bytes (cf. Annexe 2)

Annexe 2. Mesures courantes de stockage de données

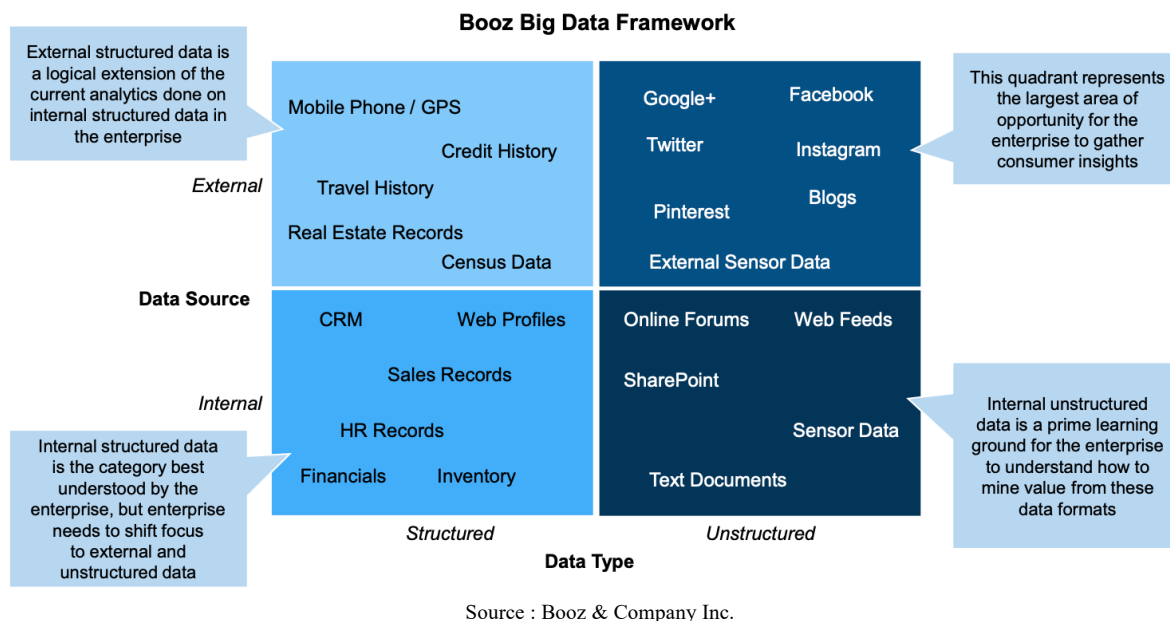


Source : <https://www.futura-sciences.com/tech/dossiers/informatique-stockage-donnees-informatiques-105/page/2/>

Unit	Shortened	Storage Capacity
Bit	Binary digit: 1 or 0	1 or 0
Byte	Byte	8 bits
Kilobyte	KB	1024 bytes / 10^3 bytes
Megabyte	MB	1024 KB / 10^6 bytes
Gigabyte	GB	1024 MB / 10^9 bytes
Terabyte	TB	1024 GB / 10^{12} bytes
Petabyte	PB	1024 TB / 10^{15} bytes
Exabyte	EB	1024 PB / 10^{18} bytes
Zettabyte	ZB	1024 EB / 10^{21} bytes
Yottabyte	YB	1024 ZB / 10^{24} bytes

Source : <https://solutionsreview.com/data-storage/data-storage-units-of-measurement-chart-from-smallest-to-largest/>

Annexe 3. Typologies de données



Annexe 4. Trois exemples de droits fondamentaux individuels défendus par le RGPD

Droit d'accès de la personne concernée

La personne concernée a le droit d'obtenir du responsable du traitement la confirmation que des données à caractère personnel la concernant sont ou ne sont pas traitées et, lorsqu'elles le sont, l'accès auxdites données à caractère personnel [...]. (RGPD, 2016, Article 15, p. 43)

Droit à l'effacement

La personne concernée a le droit d'obtenir du responsable du traitement l'effacement, dans les meilleurs délais, de données à caractère personnel la concernant et le responsable du traitement a l'obligation d'effacer ces données à caractère personnel dans les meilleurs délais [...]. (RGPD, 2016, Article 17, p. 43)

Décision individuelle automatisée, y compris le profilage

La personne concernée a le droit de ne pas faire l'objet d'une décision fondée exclusivement sur un traitement automatisé, y compris le profilage [...]. (RGPD, 2016, Article 22, p. 46)

Annexe 5. Quelques fournisseurs de données tierces (*Third-party data*)

Oracle Data Cloud

Oracle Data Cloud est une plateforme *cloud* de gestion de données marketing qui accompagne les entreprises (notamment Disneyland Paris et Panasonic) dans la personnalisation de leurs campagnes de marketing en ligne, hors ligne et mobile. Elle aide ainsi les plus grands spécialistes du marketing à obtenir de meilleurs résultats en touchant les cibles appropriées, en mesurant l'impact de leurs campagnes et en améliorant leur stratégie numérique. Oracle Data Cloud utilise la plus grande *data marketplace* au monde, alimentée par plus de 15 millions de sites web. AddThis, BlueKai, ou encore Datalogix sont à compter parmi son portefeuille de marques.

(Oracle, 2021b)

Weborama

Créée en 1998, Weborama est une entreprise française qui propose des solutions précises de connaissance du consommateur pour permettre aux entreprises de générer de la croissance tout en rationalisant les coûts marketing. De cette manière, elle permet d'optimiser les stratégies médias et marketing. Weborama offre une combinaison de technologies, de data et d'expertises conformes au RGPD (*cf.* section 2.2.4) qui tentent de comprendre et cibler au mieux les populations cibles de ces entreprises. L'entreprise est aujourd'hui active dans une trentaine de pays, et accompagne de nombreuses marques dont Danone et Intermarché sur le territoire français.

(Weborama, 2020)

Adsquare

Adsquare soutient les spécialistes du marketing dans la compréhension de leurs consommateurs en exploitant les informations d'audience et de localisation pour mettre au jour leur comportement. A nouveau, ces données permettent aux entreprises de cibler et de s'engager auprès de leurs consommateurs, améliorant ainsi leur efficacité marketing. Parmi ces entreprises, nous retrouvons Coca Cola, IKEA ou encore McDonald's.

(Adsquare, 2021)

Nielsen

Nielsen met tout en œuvre pour comprendre précisément ce que les individus écoutent et regardent. Pour cela, l'entreprise analyse sans cesse une grande quantité de chaînes et plateformes, sans oublier tout le contenu présent sur les réseaux sociaux en passant par la télévision. De cette manière, Nielsen est en mesure de découvrir des opportunités

considérables et contribuer à l'accélération de la croissance des entreprises qui font appel à ses services.

(Nielsen, 2021)

Annexe 6. Guide d'entretien soumis aux interlocuteurs dans la démarche qualitative

0. Rappel méthodologique

Les entrevues sont de type semi-structuré. Celles-ci dureront en moyenne 30 à 45min, et seront enregistrées pour des raisons pratiques. L'idée est de laisser parler les intervenants, pour ensuite pouvoir décortiquer chacun de leurs propos dans le cadre de la démarche qualitative exposée dans le quatrième chapitre de ce mémoire. Un compte-rendu écrit de chacune de ces entrevues est disponible en annexe (cf. Annexe 7).

1. Introduction

- Présentation des objectifs de l'entrevue et du mémoire,
- Demander l'accord de l'interlocuteur concernant l'enregistrement de la session,
- Présentation de l'interlocuteur en quelques mots (parcours et occupation actuelle), ainsi que l'entreprise et/ou département dans lequel il travaille.

2. Objectifs de la discussion

Dans les propos de l'intervenant, repérer les différents critères dont il est question. L'objectif n'étant pas de poser des questions trop précises, mais bien de créer un échange, une discussion.

- Identifier les types de données auxquels l'interlocuteur est confronté dans l'exercice de son métier,
- Demander une définition personnelle de qualité des données,
- Tenter d'identifier le ou les critères les plus importants pour l'interlocuteur (ex. demander sur quoi mettre la priorité avec un petit budget alloué à la gestion de la qualité des données, ou s'intéresser aux différentes anomalies rencontrées dans l'exercice de leur travail),
- Si possible, identifier les outils utilisés dans le cadre de l'amélioration de la qualité des données.

3. Conclusion

- Rappel des objectifs de l'entrevue,
- Remercier l'interlocuteur pour son temps de parole et sa disponibilité,
- Couper l'enregistrement de l'entretien.
- Répondre aux questions éventuelles de l'intervenant.

Annexe 7. Comptes-rendus des entrevues menées dans la démarche qualitative

Annexe 7.1. Rencontre avec Jean-Christophe Dewitte

Jean-Christophe Dewitte travaille pour l'entreprise de téléphonie Orange depuis 10 ans. Il occupe aujourd'hui la position de Business Analyst Policy. Il travaille depuis le début de sa carrière dans l'environnement de la data (traitements de fichiers d'adresses postales, gestion de bases de données, création de modèles, segments, etc.) et du campaigning. Jean-Christophe expose tout d'abord l'environnement de stockage de fichiers dans lequel il a travaillé il y a 20 ans maintenant. Ces fichiers à la structure identique étaient utilisés pour différents traitements : enrichissement, sélection et déduplication d'informations. Cette expérience lui aura permis de se rendre compte de la rigueur nécessaire dans le cadre de la gestion des données marketing : pas de nomenclatures de fichiers, la taille des données stockées n'est pas vérifiée, etc. L'interlocuteur insiste sur un premier point important : avant de parler de structure et de stockage de données, il faut s'attarder sur la rigueur et imposer un cadre de travail, que très peu d'entreprises ont actuellement. Ce cadre permettrait de se fixer des limites de travail, notamment dans la gestion du stockage de la donnée. A l'époque, l'espace de stockage coûte très cher. Le peu d'espace de stockage disponible était trop rapidement atteint. Pourquoi ? La volonté de stocker toutes les informations collectées et produites. Jean-Christophe insiste également sur le monde de l'entreprise qui coûte plus cher que le milieu privé : un gigaoctet pour un usage privé coûte moins cher qu'un gigaoctet utilisé à des fins professionnelles. Toujours dans un environnement data, dans un contexte bancaire cette fois, Jean-Christophe évoque les connexions sur les bases de données (Oracle) pour traiter les informations marketing via SAAS (sélection de champs, regroupements, filtres) en vue de construire des modèles marketing prédictifs : qui va partir, qui va acheter, qui va acheter plus ou moins. A nouveau, absence de cadre. Chez Orange, la manière de travailler est similaire. L'entreprise utilise les produits IBM. Jean-Christophe identifie la faible flexibilité de ces machines. D'une part l'augmentation de la puissance de calcul (moindre, difficile à augmenter, coûteuse) et l'augmentation du stockage. Un gros piège relève de cette réflexion : IBM fournit du matériel rigoureux et coûteux, qui garantit un niveau de stabilité et d'accès à l'information exceptionnel. L'utilisation exponentielle des données (surtout dans un contexte professionnel non cadré ni challengé) fait qu'actuellement, les entreprises ont ce besoin de stocker un maximum d'informations (plus j'ai de l'information, mieux je me porte) qui sont pour la plupart inutiles. Ces données sont historisées, archivées. Pourquoi la rendre disponible ? Le manque de cadre agrandit considérablement les infrastructures de stockage des données en entreprise (de manière exponentielle). C'est là une difficulté pour le département IT, qui ne parle pas le même langage que le marketing et estime alors que plus de données il y a, mieux ils se portent. Finalement, plus de stockage disponible : on constate que le calcul de dimensionnement des machines était beaucoup trop optimiste. Que font les sociétés pour pallier cela ? Ils ont tendance à répliquer les données pour les implémenter dans un nouvel environnement de stockage agrandi, action à faible coût. A proscrire ! L'information stockée est en décalage avec l'environnement initial. L'idéal est d'avoir une seule et unique source d'information avec un très faible taux de

réplication. En ce qui concerne le monde des campagnes, ils ont leur propre environnement séparé, une grande quantité de données utilisées, pas beaucoup de stockage. Ce sont souvent les contraintes techniques qui font qu'on fait un pas en avant dans la technologie. En somme, puissance de calcul phénoménale, espace de stockage pas cher, mais manque de cadre (dépend des ressources des entreprises) : efficacité, propreté. Malgré le faible coût, le monde du stockage de la donnée évolue tellement vite, le monde du cloud s'ouvre à nous (Amazon, Google, Microsoft Azure), qui rendent beaucoup plus flexibles les contraintes actuelles liées au stockage de la donnée. Nous avons donc déjà abordé plusieurs critères qui concernent la qualité des données : récence, accessibilité, unicité, consistance des données. En ce qui concerne la qualité des informations dans le monde du marketing, la gestion des doublons pose énormément de souci. Une grande majorité de cas ne sont pas traités. Par exemple, des adresses mails en majuscules et minuscules. Jean-Christophe souligne le manque de standardisation des données. Plus on laisse de liberté, plus on laisse des problèmes rentrer dans la base de données. Un autre exemple : pas de check de validation des données (présence de « ; » dans les adresses mail. Les entreprises perdent énormément de temps. Manque d'apport de qualité à l'insertion des informations en amont. Toutes ces anomalies (accents mal transformés, problèmes de conversion, etc.) entraînent un manque de sérieux dans le monde du campaigning. Cela amène plein de problèmes, notamment de duplication (un client qui reçoit un mail ou un courrier en double). Plus généralement, quand on rentre dans une agence ou un magasin et qu'on ne nous retrouve pas dans le système de fidélité car une information est mal renseignée ou mal associée à un code postal. Pire encore, les entreprises ne sont pas prêtes au métier d'ingérer des données proprement. Très peu de gens n'ont pas conscience de l'importance que cela a. Passer du temps à traiter une colonne (ex. une bonne date de naissance), que faire avec les informations mal renseignées ? Le temps passé ici va faire gagner un temps dingue pour les gens en fin de ligne. Il serait intéressant de mesurer, mais très difficile. Jean-Christophe parle d'expérience personnelle. Gros point d'attention : prendre suffisamment en compte les doublons et les impacts de bout en bout qu'ils génèrent. Le mot d'ordre est qu'il faut mettre à ces places d'ingestion de données des individus qui comprennent les impacts. Cela demande beaucoup de soin. C'est un métier à proprement parler. En termes de logiciels d'amélioration de la qualité des données, Jean-Christophe a utilisé des logiciels de standardisation d'adresses postales (ex. Road65). Cet outil est utilisé dans plein d'endroits différents. Dans le cadre des audits de bases de données avec un outils, un diagnostic des données stockées est effectué. Le problème est que pallier tous ces problèmes devient très vite très onéreux, souvent on ne sait pas par quoi ni où commencer. Avec la digitalisation, l'endroit où vit un consommateur est de plus en plus difficile à savoir. L'adresse postale est cruciale pour déterminer de nouvelles informations sur les individus, une très grande valeur de connaissance client : on peut déterminer le milieu social, la gamme de revenus en fonction de l'âge, dès qu'il y a des soucis là-dedans, on ne connaît plus le client. Jean-Christophe est plus que jamais convaincu que l'adresse postale a une grande valeur à l'heure actuelle. En observant le coût d'amélioration de la qualité des données : mettre en évidence les données qui ne peuvent pas être utilisées au moyen d'un flag d'enrichissement. Ensuite, Jean-Christophe engage une personne qui s'intéresse à ses clients pour améliorer le contenu de la qualité des informations et ainsi couvrir l'essentiel de la clientèle. Par ailleurs, en sachant que 10% de la population déménagent par an, il faut s'intéresser à l'actualité des

adresses postales. Mais pas uniquement, beaucoup de changements dans tous les coins de la base de données. Il faut également standardiser les informations, cela ne doit pas coûter cher mais judicieux de mettre une équipe qui standardise les données. Enfin, s'intéresser à la déduplication des informations (plus cher). Attention aux liens qui disparaissent suite à la suppression des données. Globalement Jean-Christophe insiste sur agir tout de suite, ne pas attendre une fois le problème identifié, agir directement. Monitorer toutes les sources de données. Pour terminer, Jean-Christophe définit une donnée marketing de qualité comme une donnée qui lorsqu'elle est utilisée apporte quelque chose d'important en retour. C'est bien beau d'avoir une adresse de qualité, mais cela ne donne pas l'assurance que c'est celle-là qu'il faut sélectionner. Un client est intéressant lorsqu'on sait qu'il interagit avec nous. Plus on cadre les actions enregistrées qu'il a faites, plus c'est intéressant pour l'entreprise. Il faut utiliser l'information déjà disponible : les campagnes marketing ne sont pas plus sophistiquées aujourd'hui qu'elles ne l'étaient il y a 10 ou 15 ans. Les entreprises qui y arrivent le plus sont les entreprises qui connaissent le mieux leurs clients. Il faut que ça soit clean partout, ne pas se focaliser sur une partie de la base de données. En somme, nous pouvons retenir que dans le monde du campaigning, il faut utiliser les informations quand on les a. bien souvent, les entreprises ne le font pas par commodité, par manque d'envie. Souvent, elles restent focalisées sur leur manière de faire (« on a toujours fait comme ça »). Avec un peu d'huile de coude, il est possible de faire des miracles avec peu d'informations. Faire cela intelligemment.

Annexe 7.2. Rencontre avec Marguerite Paris

Marguerite Paris est une ancienne étudiante de la *Louvain School of Management*. Elle travaille dans l'environnement data depuis quelques années maintenant, milieu très quantitatif. Elle commence sa carrière chez HelloFresh en tant qu'analyste dans la team *CRM*, beaucoup de traitement de données, beaucoup de *AB testing* pour optimiser les campagnes. Ensuite, elle a travaillé chez N26, une banque en ligne allemande, principalement pour de la coordination de projet, en quelque sorte *product owner* entre l'équipe data et l'équipe marketing. Elle devient ensuite *data scientist* chez Trade Republic (banque), pour un rôle très technique de codage. Dans les trois cas, elle travaillait surtout avec des *third-party data* (ex. données de Facebook). La difficulté est de combiner les données extérieures avec des données propres au système de l'entreprise. En s'intéressant à la qualité des données, en marketing il y a beaucoup de marketing channels (ex. Google et Facebook). Elle ne rencontre pas énormément d'incidents liés à la qualité des données étant donné qu'elle travaille surtout avec des API, notamment Google et Facebook qui ont des API très solides en termes de qualité des données. La difficulté survient surtout lorsqu'elle doit traiter avec de plus petits prestataires, ou lorsque les processus ne sont pas entièrement automatisés. La qualité des données sera plus facilement assurée lorsque c'est automatisé. Une autre difficulté liée à la qualité des données vient du fait qu'en marketing il y a beaucoup de sources de données différentes. Recréer le *customer journey* est très difficile. Une donnée de qualité est une donnée dans la documentation (métadonnée) est la plus claire possible, sinon on ne sait pas ce qu'on regarde, on ne comprend pas la donnée, on ne sait pas l'interpréter ni l'utiliser ! La fréquence de rafraichissement doit être clairement

renseignée. La nature de la donnée également. L'absence de doublons est primordiale aussi. Une donnée de qualité est aussi une donnée qui ne change pas tout le temps, si cela doit changer, cela ne doit pas casser le travail qui a été fait au préalable. Une donnée sera qualitative si elle est disponible rapidement, mais cela dépendra du processeur, pas de la donnée elle-même. Pas d'intervention humaine ! Le *change management* n'est pas à négliger, surtout auprès des petits providers. Enfin, la sécurité (RGPD) assure la qualité d'une donnée, elle demande beaucoup de temps (surtout dans le secteur bancaire), car toutes les transactions doivent être anonymisées : pas de nom de famille, pas de prénom, pas d'adresse. Trade Republic ne conserve que l'année de naissance et non le jour. Chez HelloFresh, ils allaient trop à l'essentiel, ne collectaient que des données essentielles (nom et adresse), mais pas de genre par exemple. Or, difficile de faire de l'analytics derrière.

Annexe 7.3. Rencontre avec Camille Laimay

Camille Laimay est responsable du pôle *Business* chez Weborama France, entreprise spécialisée dans la collecte de données marketing (ex. cookies), données qui sont temporairement mises à disposition de clients en échange d'un forfait, et ce pour différents types d'utilisations (ex. campagne de communication). Camille insiste : les données ne sont pas vendues, Weborama en reste propriétaire. Grâce aux données collectées, Weborama retravaille également les données pour en tirer quelque chose de valorisable et utilisable. Par exemple, établir un *persona*¹² ou calculer un panier moyen. Camille rejoint fortement la notion de qualité des données telle qu'elle est décrite au travers du chapitre consacré à la revue de la littérature. Plus précisément, l'intervenante marque son accord pour chacun des critères relevés : complétude, pertinence, exactitude et précision (*accuracy*), conformité, consistance et cohérence (*consistency*), récence (l'entreprise recommande une fraîcheur de données inférieure ou égale à 30 jours), unicité, accessibilité et disponibilité. Ceux-ci sont primordiaux si l'on veut maximiser la performance de l'utilisation des données. Le critère qui concerne l'intégrité des données reste de moindre importance selon Camille, cependant à ne pas négliger.

En revanche, le critère lié à la sécurité des données apparaît trop tard. Ce critère renvoie surtout au Rgpd, qui légifère les politiques de consentement de collecte, stockage et d'utilisation des données personnelles. En réalité, ce règlement contraint une entreprise à gérer rapidement l'abandon d'un client dont on possède les données personnelles.

Enfin, au cours de la discussion, nous sommes parvenus à mettre en avant deux critères : la source de la donnée, ainsi que la nature de la donnée (ex. *CRM*). D'une part, la collecte de la donnée permet effectivement de juger de la fiabilité de celle-ci. En effet, la source de cette information en dit long sur la qualité de la donnée, par exemple sur le suivi des protocoles de

¹² En marketing, un *persona* désigne un individu imaginaire représentant la cible d'une activité marketing, qu'il s'agisse d'un produit, d'un service ou d'une campagne de communication par exemple (Bathelot, 2018).

collecte. Camille estime que si la méthode de collecte est bonne, la qualité des informations qui ont été collectées est un peu assurée. D'autre part, la nature de la donnée.

Annexe 7.4. Rencontre avec Éric de Bellefroid

Éric de Bellefroid est directeur marketing et consultant marketing Senior chez Freedelity. Cette entreprise propose une solution aux distributeurs pour apprendre à mieux connaître leurs consommateurs. Elle aide les retailers à collecter des données personnelles sur le lieu de vente, et de se constituer une base de données qui soit apte à représenter au mieux la relation qu'à le distributeur avec ses consommateurs. Freedelity propose une *consumer data platform* aux retailers qui a la particularité de se mutualiser avec les données d'autres retailers qui travaillent avec le même outil. Car collecter de la donnée n'est pas difficile, la tenir à jour est une autre affaire. Ainsi en Belgique, Freedelity utilise la puce électronique de la carte d'identité des individus citoyens belges. La collecte s'effectue sur le lieu de vente lui-même via des bornes de lecture intégrées avec le système de caisse du retailer. L'entreprise collecte ainsi les données personnelles après insertion de la carte d'identité dans un lecteur (ex. adresse postale) mais aussi une adresse mail, un numéro de mobile, ainsi que la traçabilité des consentements quant au traitement de ces données (car règlementé par le RGPD). L'usage de la carte d'identité réduit le risque d'incidences liées à la qualité des données. L'offre de Freedelity est ainsi axée autour de plusieurs piliers : la mutualisation des données entre retailers, bornes/kiosques intégrés avec la plateforme Custocentrix (application hébergée en Belgique par Freedelity qui permet à chaque enseigne de réconcilier des données d'identité avec des données de dépense via les systèmes de caisse ou des API), le portail MyFreedelity à destination des consommateurs qui leur permet de garder le contrôle de l'usage des données qui sont faits après la collecte. Custocentrix comprend plusieurs modules : le *consumer datahub* (gère la problématique de collecte des données personnelles et de consentement et fait bénéficier aux retailers de la mise à jour des données), la *consumer data platform* (permet aux enseignes qui le souhaitent de stocker sur les serveurs sécurisés de Freedelity des données comportementales et transactionnelles), la gestion du programme de fidélité (modalités à points), le module de géomarketing (enrichit chaque adresse de domicile avec des positions XY sur une carte – attitude et longitude – et enrichit la donnée avec un code secteur statistique). Les codes-secteur statistiques sont utiles pour effectuer des recensements de population de manière informatisée. Enfin, le module complémentaire *data source* permet de retourner à une enseigne l'ensemble des données collectées sous forme de fichiers dénormalisés utilisables via des outils de statistiques avancés. En ce qui concerne la qualité des données, il faut un traitement algorithmique qui permet de s'assurer que les enregistrements ne vont pas créer de doublons. Par ailleurs, Freedelity met doucement sur le marché des services de gestion de la qualité de données pour des enseignes qui n'utilisent pas l'application Custocentrix, pour ainsi bénéficier du fait qu'on a un consentement de 6,5 millions de consommateurs belges quant au traitement de leurs données personnelles. Le fichier Freedelity au cumul de l'activité de l'ensemble des enseignes qui travaillent avec l'entreprise concerne 6,5 millions de consommateurs belges, soit le plus gros fichier de données basé sur le consentement en Belgique. Ce consentement permet

de faire plusieurs prestations, notamment la mise à jour des données comme évoqué ci-avant, mais aussi de prêter des services de *data quality* de manière directe (via des *marketing service providers*) ou indirecte, qui permettent à des annonceurs de valider l'existence d'un consommateur en comparant ce qu'ils détiennent dans leurs fichiers avec celui de Freedelity (ex. M. de Bellefroid réside bien à telle adresse), de les prévenir si une autre adresse postale a été détectée plus récemment, si nom et prénom sont corrects ou non, de valider le code de langue et de genre dans leur fichier également, ou encore valider une date de naissance. Ce service ne concerne pas de l'enrichissement de données. Pour faire cela, plusieurs algorithmes. Le premier renvoie à la capacité à standardiser les adresses postales (identifiant unique de voie) via géocodage. Pour cela, Freedelity a créé un référentiel d'immeubles qui connaît tous les immeubles en Belgique avec informations XY pour cartographier cela, pour adjoindre à chaque adresse un identifiant d'immeuble. Utile car permet de faire des regroupements de personnes qui résident à une même adresse (ex. envoyer un seul courrier dans un logement qui couvre trois personnes ayant chacune une carte de fidélité). Un second algorithme permet, une fois l'adresse identifiée, d'utiliser des logiques qui calculent des proximités entre profils pour ainsi définir un seuil au-delà duquel on estime qu'il y a ressemblance entre deux profils pour les considérer comme un et un seul profil client. Tout le travail effectué autour de la qualité des données est automatisé via des scripts. Dès qu'il y a souci de *data quality*, on le résout par script. Aujourd'hui, aucun acteur du secteur du *marketing automation* a une réponse suffisamment robuste à la gestion de la qualité des données, ni même à l'intérieur des CRM. D'ailleurs, l'échec des CRM s'explique avant tout par des problèmes liés à la qualité de données. D'autres processus de qualité des données sont effectués tels que la gestion de la réputation IP de l'adresse mail utilisée dans le cadre de campagnes de communication. Unicité, récence, processus de tenue à jour, l'exhaustivité de la donnée également (complétude). Compliqué d'avoir un taux de documentation important sur les données collectées en règle générale (conformité). Pertinence de la donnée ultra importante car la modélisation de données découlera des données elles-mêmes. Si l'on désire connaître ses clients, c'est sur l'unicité des enregistrements de données qu'il faut mettre la priorité, avec la tenue à jour de ces données également. Éviter les *garbages*... En termes de stockage de données, celles-ci sont stockées dans des serveurs qui sont propres à Freedelity, ultra-sécurisés (flux de données impossibles vers les EU). Ces serveurs sont configurés par Freedelity, qui a la possibilité d'intervenir en tout temps à distance. En cas de problème technique, seuls trois techniciens ont les droits d'accéder à la zone de stockage. Le stockage des données fait beaucoup l'objet de certifications ISO qui sont coûteuses à obtenir d'une part mais aussi à conserver d'autre part. comme dans toute entreprise, on a besoin d'un DPO (*Data Protection Officer*). Freedelity a choisi de faire affaire avec un DPO indépendant, spécialiste en sécurité informatique qui vient de l'environnement bancaire. Des tiers (ex. prestataires propres aux enseignes) ont des droits d'auditer les données stockées sur les serveurs Freedelity. Il y a donc toute une série de normes et bonnes pratiques de protection qui sont mises en place. Seulement un très petit nombre de personnes nommées ont le droit d'accès aux données, qui elles sont stockées dans des serveurs différents. Le degré de disponibilité des données stockées est hautement important étant donné que Freedelity est une plateforme hautement transactionnelle.

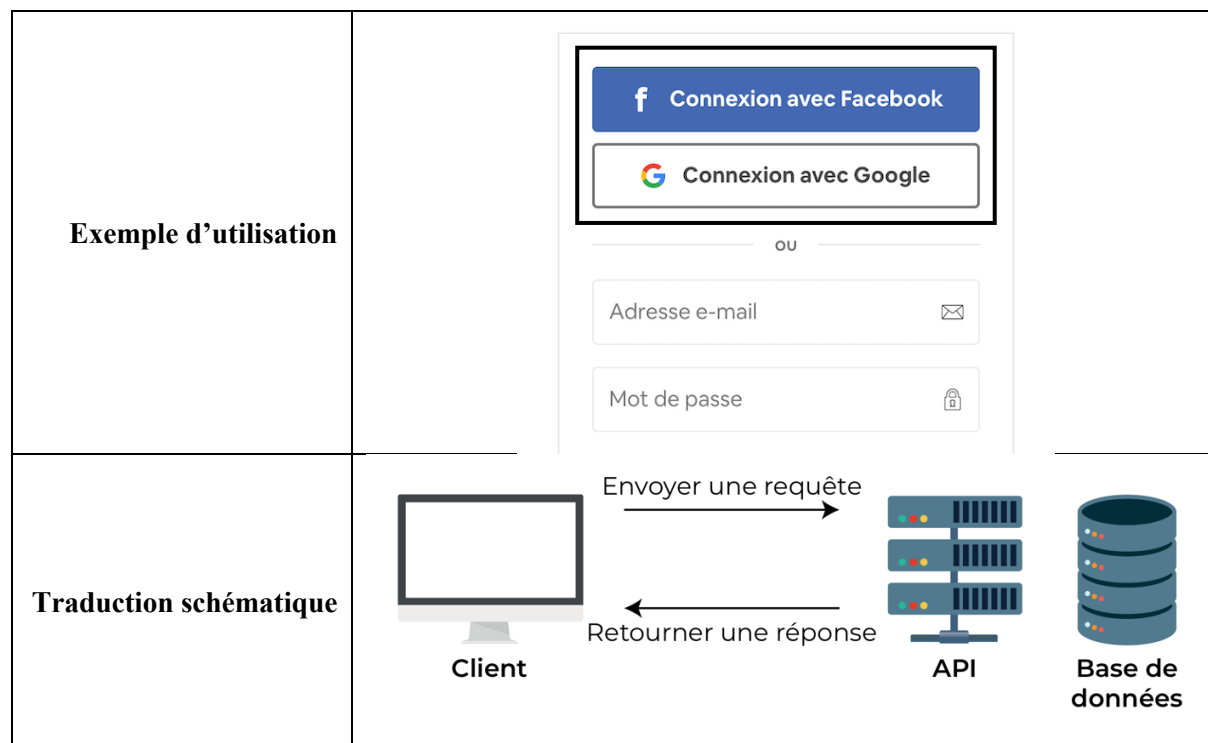
Annexe 7.5. Rencontre avec Ali El Maghraoui

Ali El Maghraoui est directeur au sein de EY plus spécifiquement dans le département Data Analytics, responsable de l'étude des stratégies data des entreprises (vision, business case data, roadmaps, gouvernance de la donnée, architecture et modèles de données, etc.). Sa spécialité est plus fonctionnelle que technique dans le domaine de la donnée. En ce qui concerne son parcours, l'intervenant travaille depuis 14 ans dans le domaine de la consultance. Il commence dans une petite boîte de consultance en Wallonie, puis Accenture dans le département *Information Management*. Ensuite, il dirige le département *Data Analytics* chez Avanade. Enfin, il rejoint l'équipe de EY. En ce qui concerne les types de données qui le concernent, son parcours dans des entreprises de télécommunications, énergétiques et bancaires fait qu'il aura fait face à tous types de données : données clients, marketing, crédits, finance, utilisation de réseaux, etc. Le concept reste le même mais l'aspect fonctionnel diffère. Pour Ali, une donnée est créée pour répondre à un besoin spécifique. Cette donnée restera de qualité à partir du moment où elle sera toujours en mesure de répondre à ce besoin-là. De manière factuelle, comment on s'assure qu'elle réponde toujours à ce besoin, c'est notamment à travers des dimensions de données telles que complétude, récence, etc. La complétude des données se définit bien par le fait que toutes les données soient bien présentes dans l'environnement de données. Le degré de complétude ne concerne que la manière dont on présente les résultats. Il faut partir du principe que pour mesurer une qualité des données, on doit d'abord définir comment on doit mesurer (i.e. complétude = s'assurer que le champ n'est pas vide). Ensuite, on détermine un seuil qui nous dit que la qualité des données est bonne lorsque 90 % des données sont complètes. La récence est primordiale. Elle renvoie à l'actualisation des données, mais aussi à la fréquence de mise à jour des données : à quelle fréquence ces données sont-elles mises à jour ? Tous les jours ? Toutes les minutes ? Il faut s'assurer qu'il s'agisse de la dernière version de la donnée. La conformité des données renvoie à la validité des données. Deux concepts surviennent souvent de manière conjointe : la conformité et la validité des données. La conformité (plutôt fonctionnelle) renvoie au respect du format qui a été défini pour la donnée. La validité (plutôt technique) quant à elle permet de contrôler la donnée stockée (par exemple un âge ne peut pas être inférieur à 0). Dans la pratique, l'exactitude et la pertinence (*accuracy*) représentent la dimension de la qualité des données la plus compliquée à atteindre et à définir. Dire qu'une donnée doit correspondre à la réalité de ce qu'elle représente c'est très bien, mais il faut également déterminer ce que représente cette réalité. L'implémenter dans un code SQL ou autre est très complexe car toute la logique qui gravite autour de cette donnée doit être refaite, ce afin de vérifier que tout est correct. En ce qui concerne la pertinence des données, celle-ci doit pouvoir se différencier de la dimension précédente. On ne devrait pas pouvoir considérer cela comme un critère de stockage, car cela dépendra toujours de l'utilisation qu'on fera de cette donnée. Nous avons parlé de pertinence de stockage : doit-on stocker cette valeur ou non ? Néanmoins, à partir du moment où on se demande si on doit stocker cette donnée ou non, cela concerne surtout le design fonctionnel plutôt qu'une vraie mesure de qualité. Car si on ne stocke pas l'information, on ne sait pas déterminer le niveau de qualité de cette donnée étant donné qu'elle n'est pas stockée. Il y a une nuance. A partir du besoin, on détermine quelles

seront les données à stocker dans les systèmes d'informations. La pertinence n'est donc pas un indicateur de qualité de donnée. L'unicité de la donnée renvoie à la représentation unique des données, ainsi qu'à l'absence de répétitions dans l'environnement de stockage de données. La plupart du temps, surtout dans les systèmes de stockage actuels, cette caractéristique est vérifiée par le design. On perd souvent du temps à construire un indicateur de qualité pour l'unicité car la plupart du temps elle est vérifiée par le design, au travers des clés primaires et uniques de contrôle. L'accessibilité et la disponibilité des données reprend la facilité d'accès, de compréhension et d'utilisation des données. Ce critère est souvent mentionné, cependant une question se pose : comment mesurer cela ? S'il s'agit d'un indicateur, il faut pouvoir en mesurer la performance. Or ici, c'est très compliqué. Par définition, un indicateur doit être mesurable. Cela étant, cette caractéristique peut être mesurée au travers des différents processus opérationnels. La seule manière de mesurer cet indicateur est de vérifier si le processus informatique qui a permis à une donnée d'être chargée et stockée a été exécuté ou non. A ce moment-là, cela devient un check opérationnel, et non plus un indicateur de qualité. Le critère suivant qui concerne la consistance et la cohérence des données représente l'absence de contradictions dans l'environnement de donnée, ainsi que l'homogénéité des représentations dans ce même environnement. Seul le premier item est correct, on laisse tomber le second. A nouveau, ce critère doit être mesurable. Si on compare deux enregistrements de date de naissance pour Maxime, 1998 d'un côté et 1999 de l'autre, il y a clairement un problème de consistance. En revanche, une date stockée dans tel format dans une base de données, ou un autre format, c'est un choix de design, pas de rapport avec la donnée elle-même. Souvent, si on compare des bases de données Oracle avec SQL Server, certains stockeront la donnée comme année, mois, jour, heure, minute, seconde ; tandis que d'autres proposeront de stocker en année, mois, jour ou heure, minute, seconde. Cela dépend des contraintes techniques du système. On garde donc le premier item. Présence de tous les liens relationnels essentiels, validité des liens relationnels dans l'environnement de données et cohérence de l'évolution des relations entre entités de données définissent l'intégrité des données. L'intégrité revient à s'assurer que dans un système un client Maxime soit renseigné, et que dans un second système, ce même client soit retrouvé. C'est aussi simple que cela. Cela peut se traduire par les clés primaires et clés étrangères, ou encore amener cela au niveau fonctionnel. En ce qui concerne le processus de stockage, la garantie de sécurité est primordiale : RGPD, sécurité des bases de données, sécurité des centres de données, sécurité des infrastructures Cloud. En outre, le critère énergétique est plutôt complexe. A nouveau, comment mesurer cet indicateur ? émissions de CO2, consommation électrique, etc. Il ne faut pas oublier que ces critères, il va falloir les implémenter en entreprise. A un moment donné, ils devront apparaître dans un dashboard avec des couleurs vertes, rouges et orange, en disant si bon ou pas bon. Il faut alors être en mesure de trouver une méthode d'évaluation et de facturation de ce critère. Si on n'y parvient pas, ce critère n'est pas à prendre en compte. Actuellement, beaucoup d'entreprises tendent à imaginer des indices (vs. Indicateurs) qui rendent compte de l'impact sociétal de leurs activités Big Data. Le stockage en tant que tel n'a pas vraiment d'impact énergétique. Prenons l'exemple de la cryptomonnaie. Ce n'est pas le fait de la stocker qui fait qu'elle est énergivore, c'est plutôt tout le processus de miner qui coûte de l'énergie. Ce critère énergétique dépendra de l'utilisation que l'on fait de la donnée, et de comment on la calcule, plutôt que de comment on stocke cette information. Si on

veut considérer l'énergie comme un critère de qualité de données, il faudra penser plus large, ne pas se concentrer sur le stockage des informations, qui est tout à fait négligeable comparé à toute la puissance de calcul utilisée pour l'extraire, la transformer et la mettre quelque part. Enfin, le critère d'économie de ressources renvoie à toutes les ressources nécessaires au bon fonctionnement du stockage des données (ressources financières, ressources humaines). Peu importe ce qu'on définit, on doit pouvoir à la fin essayer de représenter d'une manière ou d'une autre. Le problème avec les critères liés à la qualité du processus de stockage (énergie et ressources notamment), c'est qu'on ne peut pas à associer ces critères uniquement au stockage, cela va au-delà de ça. Il faut considérer le data life cycle en général : à partir du moment où on crée la donnée, on l'envoie, etc. Dans la vie pratique, si l'objectif du mémoire est de proposer une liste de critères, il faut pouvoir mesurer chacun d'entre eux et se positionner d'année en année (ex. est-ce qu'on est rouge, orange ou vert). Deux critères supplémentaires : la méthode de collecte de la donnée (automatisation), qui concerne plutôt les métadonnées (la donnée est collectée de telle ou telle manière).

Annexe 8. APIs – Application Programming Interfaces



Source : <https://openclassrooms.com/fr/courses/6573181-adoptez-les-api-rest-pour-vos-projets-web/6816951-initiez-vous-au-fonctionnement-des-api>

Annexe 9. Donnée, information, connaissance

Source : Codet, B. (2009). *Donnée, information, connaissance*. Retrieved August 8th, 2021 from <https://brunochaudet.wordpress.com/2009/03/30/donnee-information-connaissance/>

Par définition, une donnée est un élément brut, qui n'a pas encore été interprétée, mis en contexte.

Et c'est là toute la différence entre une information et une donnée. En effet, une information est par définition une donnée interprétée. En d'autres termes, la mise en contexte d'une donnée crée de la valeur ajoutée pour constituer une information.

On peut considérer la connaissance comme une information comprise, c'est-à-dire assimilée et utilisée, qui permet d'aboutir à une action. « Nonaka et Takeuchi (deux experts du knowledge management) différencient deux formes de connaissance : une connaissance tacite et une connaissance explicite.

La connaissance tacite : c'est la connaissance que possèdent les individus. Elle n'est pas formalisée et difficilement transmissible. Ce sont les compétences, les expériences, l'intuition, les secrets de métiers, les tours de main qu'un individu a acquis et échangés lors d'échanges internes et externes à l'entreprise. La connaissance tacite se transmet par imitation et imprégnation. On le sait sans le savoir. On met en œuvre des pratiques sans vraiment s'en rendre compte.

La connaissance explicite : C'est la connaissance formalisée et transmissible sous forme de documents réutilisables. Ce sont les informations concernant les processus, les projets, les clients, les fournisseurs, etc. La connaissance explicite se transmet par des documents formalisés et normalisés.

L'utilisation des connaissances dans le contexte d'activité s'appelle du savoir-faire ou des compétences.

La conceptualisation de ces savoir-faire et compétences et l'expérience de ces situations de travail forment ce que l'on appelle l'expertise.

La numérisation croissante constitue l'un des plus grands défis actuels rencontrés par les entreprises. En effet, il serait difficile d'imaginer notre société vivre sans smartphone, ordinateur et connexion Internet. Cette tendance démultiplie toujours plus la quantité de données produites dans le monde. Le marketing figure parmi les branches du management les plus impactées par cette révolution numérique. La variété et la quantité des données disponibles rendent effectivement la tâche de leurs utilisateurs particulièrement difficile et impactent la qualité des bases de données concernées. Aussi, la production constante de données encourage à tort ces mêmes utilisateurs à en stocker un maximum dans la mesure du possible. C'est dès lors à se demander quelles sont les données qu'une entreprise doit conserver mais aussi entretenir, car toutes les données ne sont pas bonnes à stocker...