UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

LES PME ET LE TRAITEMENT DE L'INFORMATION À L'ÈRE DES DONNÉES MASSIVES

MÉMOIRE PRÉSENTÉ
COMME EXIGENCE PARTIELLE DE LA

MAÎTRISE EN SCIENCES DE LA GESTION

PAR
NADA FEJJAR

Octobre 2021

Université du Québec à Trois-Rivières

Service de la bibliothèque

Avertissement

L'auteur de ce mémoire ou de cette thèse a autorisé l'Université du Québec à Trois-Rivières à diffuser, à des fins non lucratives, une copie de son mémoire ou de sa thèse.

Cette diffusion n'entraîne pas une renonciation de la part de l'auteur à ses droits de propriété intellectuelle, incluant le droit d'auteur, sur ce mémoire ou cette thèse. Notamment, la reproduction ou la publication de la totalité ou d'une partie importante de ce mémoire ou de cette thèse requiert son autorisation.

SOMMAIRE

Adoptant une approche configurationnelle, cette recherche étudie le traitement de l'information réalisé dans les petites et moyennes entreprises (PME) à l'ère des données massives. Il en résulte différentes configurations d'activités et d'utilisation d'outils de gestion qui assurent l'obtention d'une haute ou d'une moyenne qualité de l'information, mesurée par l'intermédiaire des six différents critères, tels que : l'exactitude, la cohérence, la conformité, la disponibilité, l'actualité et l'unicité. Plus précisément, la première activité du traitement effectué est la *génération des données* qui se réalise lors de la consultation des données provenant des systèmes internes et externes de l'entreprise. Vient ensuite *l'acquisition et l'analyse des données*, en une seule et même étape, et ce, par l'intermédiaire de sous-activités relatives aux pratiques de gestion des données, à la préparation, partage et traitement des données, aux outils d'intelligence d'affaires, ainsi qu'à l'utilisation d'outils non informatisés. Suit *le stockage des données* à l'aide des dispositifs de stockage internes et externes et, enfin, la *diffusion / visualisation de l'information* lors du partage des données et des informations destinées au personnel de l'entreprise, ainsi qu'à toute personne en lien avec l'entreprise, mais n'appartenant pas à son personnel. Suivant la collecte et l'analyse de nos données de recherche, le présent mémoire permet ainsi non seulement d'identifier des pratiques et des outils de gestion plus précis, mais aussi de mieux comprendre comment ceux-ci s'articulent lors du traitement de l'information au sein des PME à l'ère des données massives. L'étude a été menée par une enquête en ligne auprès de 40 PME québécoises appartenant aux secteurs d'activités du manufacturier, du transport / logistique, du commerce de gros, du commerce de détail, des services aux entreprises (industriels/scientifiques), ainsi que de la finance et assurances. Pour procéder à l'analyse des données, l'étude a mobilisé une approche par l'analyse qualitative comparée par les ensembles flous (fsQCA)[1] qui utilise l'algèbre

[1] Tout au long de ce document, nous utiliserons toutefois l'acronyme anglophone qui est mieux connu/reconnu, soit fsQCA (pour *fuzzy-set Qualitative Comparative Analysis).*

booléenne pour déterminer les combinaisons de caractéristiques organisationnelles permettant d'aboutir à un résultat donné (Boswell et Brown, 1999; Fiss, 2007; Ragin, 2000, 2014), qui est, dans notre cas, la qualité de l'information. Plus spécifiquement, la combinaison configurationnelle des activités et outils de gestion obtenue par la fsQCA a résulté en deux configurations pour la présence d'une haute qualité de l'information, huit configurations pour la présence d'une qualité moyenne de l'information et neuf configurations pour une absence de qualité de l'information. De manière plus précise, afin d'assurer l'obtention d'une haute qualité de l'information, les principaux résultats suggèrent que les PME québécoises devraient aller vers une relation de partenariat et de collaboration avec les partenaires d'affaires (fournisseurs, distributeurs et clients) en partageant leurs données avec eux. Elles devraient également favoriser le stockage de leurs données en externe, par exemple via l'utilisation de l'informatique en nuage, et ce, tout en continuant d'utiliser le papier, mais avec une moindre importance par rapport aux outils informatisés. En contrepartie, il y a plusieurs choses à éviter, puisque celles-ci mènent à l'absence de qualité de l'information, par exemple l'utilisation exclusive des outils non informatisés ou l'absence de dispositifs de stockage internes et externes, et ce, en même temps. Même chose pour l'utilisation des outils d'intelligence d'affaires, comme le système d'information d'aide à la décision (SIAD) et les outils non informatisés, comme les rapports papier, ainsi que pour la diffusion des données destinée au personnel de l'entreprise et celles destinées à toute personne n'appartenant pas au personnel de l'entreprise. Notre étude contribue aux connaissances de multiples manières. D'abord, les contributions théoriques consistent en la fusion de deux activités distinctes au départ, soit *l'acquisition des données* et *l'analyse des données*, qui deviennent une seule activité appelée *acquisition et analyse des données*, mieux adaptée au contexte des PME. Une autre contribution est de démontrer qu'il existe plusieurs façons d'atteindre une haute ou une qualité moyenne de l'information. Quant aux contributions pratiques, elles résident dans des recommandations aux PME et leurs dirigeants concernant les configurations d'activités et d'outils assurant une haute qualité ou une qualité moyenne de l'information à utiliser ou à adopter. Par la même occasion, cette recherche prodigue donc des conseils

sur ce qu'il faut éviter de faire afin de préserver la qualité de l'information tout au long de son traitement. Cette recherche comporte des limites qui ne permettent pas la généralisation des résultats. D'abord, il y a le nombre d'entreprises interrogées qui est insuffisant. De plus, notre étude a été restreinte aux PME québécoises. Par ailleurs, il y a également certaines limites spécifiques qui sont liées à la méthode d'analyse des données utilisée (fsQCA) (de Guinea et Raymond, 2020; Glaesser et Cooper, 2014). Enfin, les études existantes sur les données massives sont plus faites sur les grandes entreprises, en plus d'avoir été réalisées hors du Canada. Donc, il y a un manque d'exemples d'utilisation des données massives dans les PME au Canada. Des recherches futures devraient permettre d'approfondir l'analyse de la situation actuelle concernant l'utilisation des données massives dans les PME canadiennes.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS

API	Interface applicative de programmation (ou *Application Programming Interface*)
B2B	*Business to Business* (ou service/commerce interentreprises)
B2C	*Business to Customer* (ou service/commerce aux particuliers)
BD	Base de données
BPM	*Business process management* (ou gestion des performances commerciales)
C2B	*Consumer to business* (ou le commerce électronique consommateur-entreprise)
C2C	*Consumer to consumer* (ou le commerce électronique interconsommateurs)
CRM	*Customer Relationship Management* (ou la gestion de la relation client - GRC)
CTI	Classification type des industries (ou *Standard Industrial Classification*)
EDI	Échange de données informatisées
ETC	Extraction, Transformation, Chargement (ou *Extract, Tranformation, Loading)*
FsQCA	Analyse qualitative comparée par les ensembles flous (pour *fuzzy-set Qualitative Comparative Analysis)*.
GPS	Géospatiales ou système de positionnement global
HDD	Disque dur (ou *Hard Disk Drive)*
IaaS	*Infrastructure as a Service* (ou infrastructure en tant que service)
NoSQL	Bases de données relationnelles non traditionnelles (ou *non traditional relational databases)*
OLAP	Traitement analytique en ligne
PaaS	*Platform as a Service* (ou plateforme en tant que service)
PB	Pétaoctets
PGI	Progiciel de gestion intégré (ou *Enterprise Resource Planning*)
PME	Petite et Moyenne Entreprise
QD	Qualité des données
QI	Qualité de l'information
RAM	Mémoire à accès aléatoire (ou *Random Access Memory*)
SaaS	*Software as a Service* (ou logiciel en tant que service)
SGBD	Système de gestion de base de données
SI	Système d'information (ou *Information System*)

SIAD	Système interactif d'aide à la décision
SID	Système d'information pour dirigeants
SIG	Systèmes d'information de gestion
SOA	*Service Oriented Architecture (*Architecture orientée services)
SQL	Langage de requête structurée (ou *Structured Query Language)*
SSD	Supports de stockage non mécaniques (ou *Solid State Drive)*
STT	Système de traitement des transactions
TB	Téraoctet
TI	Technologies de l'information
TIC	Technologies de l'information et de la communication
ZB	Zettaoctets

REMERCIEMENTS

La réalisation de ce mémoire a été possible grâce au concours de plusieurs personnes à qui je voudrais témoigner toute ma gratitude. Tout d'abord, je voudrais exprimer toute ma reconnaissance à ma directrice de recherche, Mme Claudia PELLETIER, professeure en systèmes d'information et membre de l'Institut de recherche sur les PME (INRPME), pour son encadrement et son soutien tout au long de la rédaction de mon mémoire.

Je désire aussi adresser mes sincères remerciements à Mme Rahma CHOUCHANE, chercheuse postdoctorale à l'Institut de recherche sur les PME (INRPME), pour son mentorat et ses guidances concernant la rédaction de ce mémoire et plus spécifiquement sur la phase méthodologique.

Un grand merci également à mes très chers parents pour m'avoir donné l'occasion de venir au Canada et étudier à l'Université du Québec à Trois-Rivières (UQTR) ainsi que mes sœurs qui m'ont toujours soutenue mentalement.

Pour finir, je tiens à témoigner toute ma gratitude à toute personne qui a contribué de près ou de loin à l'élaboration et la réalisation de ce mémoire de recherche

Nada FEJJAR

INTRODUCTION

La mondialisation a fait que le monde est devenu plus petit grâce aux échanges économiques de biens et de services au niveau international qui sont en augmentation continue (Julien et Morin, 2011). Dans ce contexte, les technologies de l'information (TI) ont permis d'établir le lien entre les clients et les fournisseurs tout en facilitant les échanges d'informations et un gain de temps considérable. Par ailleurs, les affaires électroniques interentreprises (B2B) ou des entreprises aux particuliers (B2C) permettent désormais aux clients d'effectuer leurs achats en ligne sur le site Web de l'entreprise. Les publications sur les médias sociaux ainsi que les commentaires peuvent aussi être des données utiles à l'entreprise. Tout cela contribue à une explosion de données à analyser provenant de différentes sources qui doivent être traitées dans le but d'en extraire une information de qualité qui sera exploitée par les différents services ou départements de l'entreprise (service qualité, marketing, logistique, finance, etc.). Les entreprises sont donc submergées par les données provenant de plusieurs sources de façon structurées (fichiers Excel, rapports internes, etc.) ou non structurées (vidéos, messageries, etc.). Cette abondance de données de façon continue est appelée « données massives » (ou *Big Data*) (Karoui et al., 2014). À titre illustratif, le volume des données stockées est passé de 800 000 pétaoctets[2] en 2000 à une estimation de 35 zettaoctets[3] en 2020 (Taleb et al., 2018; Zikopoulos et Eaton, 2011).

Les données massives sont caractérisées par : le *volume* qui représente la quantité des données créées chaque jour, *la variété* de ces données émanant de différentes sources; interne ou externe; structurées (documents, images, etc.) ou non structurées (données GPS, capteurs, tweets, etc.) et leur *vélocité* c'est-à-dire la rapidité avec laquelle nous produisons les données et qui est en croissance permanente (Karoui et al., 2014). Ainsi, à

[2] 1 PB = 1 pétaoctet = 10^{15} octets
[3] 1 ZB = 1 zettaoctet = 10^{21} octets

l'ère de l'abondance des données, les flux d'information en entreprise qui sont considérés comme son capital intellectuel ont gagné de l'importance au fil du temps et sont devenus indispensables pour la continuité de l'entreprise. Plus spécifiquement pour les Petites et Moyennes Entreprises (PME) qui manquent généralement de ressources financières, humaines et technologiques (Filion et Allali, 2007; Wang Shouhong, 2020). Qui plus est ces données doivent être traitées le plus rapidement possible pour que l'entreprise soit compétitive sur le marché et pour répondre aux besoins des clients (Filion et Allali, 2007), d'où l'importance d'un traitement efficace et d'une bonne gestion des données massives auxquelles les PME font face au quotidien.

La PME, dans le contexte des pays nord-américains où le Québec est situé, représente les entreprises autonomes (ne faisant pas partie d'un grand groupe) qui ont un nombre d'employés inférieur ou égal à 500, ainsi qu'un chiffre d'affaires annuel ne dépassant pas 50 millions de dollars (Filion et Allali, 2007). En décembre 2019, les PME représentaient 99,8% du total des entreprises canadiennes avec employés pour un nombre total de 1,22 million de PME, contre 2 978 grandes entreprises, alors qu'au Québec elles représentaient 254 645 entreprises (Gouvernement du Canada, 2020). Cela justifie le rôle important que jouent ces PME dans le marché d'emploi canadien.

Même si le concept des données massives est un sujet d'actualité et qu'il existe déjà des articles abordant ce phénomène, il y a toujours des lacunes. D'abord, les études qui ont été menées abordent généralement le phénomène des données massives dans les grandes entreprises. De plus, ces études s'intéressent plus au côté analytique de ces données massives à l'aide d'outils statistiques sophistiqués qui nécessitent des ressources spécifiques (la compétence et le financement) et négligent le côté managérial de gestion de l'information. Par ailleurs, ces études ne traitent pas la relation de ces outils analytiques avec la finalité ou l'objectif principal de l'entreprise qui est d'assurer une qualité de l'information. Ce faisant, la plupart des articles ne parlent que de la performance et du côté révolutionnaire de ces outils sans faire de lien avec leur valeur ajoutée sur la qualité

de l'information de l'entreprise dans le cadre de ses activités courantes et de ses opérations. Ainsi, malgré les quelques études et articles existants sur le sujet des données massives, très peu ont essayé de comprendre la manière dont sont gérées ces données massives dans les PME afin d'avoir une qualité de l'information, plus précisément celles faisant des affaires dans la province du Québec.

Le but de cette étude est de mieux comprendre comment les PME acquièrent, évaluent, traitent et stockent les données massives tout en assurant la qualité de leurs informations. À cet effet, l'objectif général de cette étude est d'explorer les pratiques permettant d'assurer une haute qualité de l'information au sein des PME, et ce, à travers l'articulation des différentes configurations d'activités et d'outils de gestion que l'on retrouve au sein des PME à l'ère des données massives. Par la même occasion, nous explorerons aussi celles qui mènent à une absence et une qualité moyenne de l'information. Cette recherche permettra de répondre à la question de recherche suivante : Quels sont les activités et les outils de traitement des données qui permettent d'assurer la qualité de l'information au sein des PME et comment ceux-ci s'articulent-ils à l'ère des données massives ?

Pour ce faire, la recherche mobilise une approche configurationnelle fondée sur l'analyse qualitative comparée par les ensembles flous (aussi appelée fsQCA pour *fuzzy-set Qualitative Comparative Analysis*). Cela à travers une enquête en ligne via le site de Surveymonkey.com et la diffusion de l'hyperlien du questionnaire sur les réseaux sociaux professionnels de la directrice de recherche. La collecte des données porte sur un échantillon de 40 PME québécoises ayant un effectif inférieur ou égal à 500 employés. Ces PME appartiennent aux secteurs d'activité suivants : le manufacturier, transport / logistique, commerce de gros, commerce de détail, services aux entreprises (industriels/scientifiques), finance et assurances. Une fois les données collectées et extraites, elles sont analysées en utilisant la méthode fsQCA qui fait partie des approches de la théorie des ensembles. Elle utilise l'algèbre booléenne pour déterminer les

combinaisons de caractéristiques organisationnelles permettant d'aboutir au résultat en question (Boswell et Brown, 1999; Fiss, 2007; Ragin, 2000, 2014).

La combinaison configurationnelle des activités et outils de gestion obtenue par la fsQCA a résulté en deux configurations pour la présence d'une haute qualité de l'information, huit configurations pour la présence d'une qualité moyenne de l'information et neuf configurations pour une absence de qualité de l'information. Ces différents agencements d'activités et d'outils composant le cycle de traitement de l'information par les PME et leurs dirigeants font également l'objet d'une discussion plus approfondie.

Ce mémoire de recherche est composé de cinq chapitres. Le premier chapitre présente la formulation de la problématique managériale. Le deuxième chapitre est la définition du cadre conceptuel. Le troisième chapitre porte sur la méthodologie de recherche. Le quatrième chapitre présente sur les résultats de l'étude réalisée ; tandis que le cinquième chapitre élabore sur la discussion. Une conclusion, incluant les limites et pistes futures, clôture le tout.

CHAPITRE 1 – LA PROBLÉMATIQUE MANAGÉRIALE

Dans cette section, nous allons présenter le thème de la recherche qui est « le traitement de l'information à l'ère des données massives » et définir les caractéristiques et le cycle de vie des données massives. Ensuite, nous allons présenter les données massives dans les PME, leurs applications, leurs atouts et leurs défis. Puis, nous allons définir la qualité des données ou la qualité de l'information. Enfin, nous présenterons le problème de recherche avec la question générale et les objectifs de recherche.

1.1 TRAITEMENT DE L'INFORMATION À L'ÈRE DES DONNÉES MASSIVES

Depuis quelques années, on ne peut parler des PME et leur développement sans aborder la mondialisation économique (Julien et Morin, 2011). Cette mondialisation est une tendance clé qui a radicalement changé le commerce mondial, de la fabrication au service client ainsi que la variété et les formats des données (Krishnan, 2013). Elle a permis une augmentation des échanges de biens et services et un abaissement des coûts de transport et de communications par la réduction des barrières douanières comme le libre-échange ou les marchés communs et par la création de nouveaux concurrents internationaux basés en Asie (Julien et Morin, 2011). De plus, l'innovation technologique a transformé la façon dont nous nous engageons dans les affaires, fournissons des services et la mesure associée de la valeur et de la rentabilité (Krishnan, 2013). Cela par le développement de nouvelles technologies d'informations et de communication, ainsi que de nouveaux modèles d'affaires. D'un côté, la technologie a évolué au cours des 20 dernières années (Krishnan, 2013), et ce, depuis le début des années 2000 avec l'apparition d'Internet et des systèmes Web 1.0 qui sont caractérisés par les moteurs de recherche comme Google et Yahoo et des entreprises de commerce électronique telles qu'Amazon et eBay, qui permettent aux organisations de présenter leurs activités en ligne

et d'interagir directement avec leurs clients (Hsinchun et al., 2012). Après 2004, le monde des affaires a connu l'apparition de nombreuses applications Web 2.0 qui ont également créé une abondance de contenu généré par les utilisateurs à partir de divers médias sociaux en ligne. C'est ainsi que se sont multipliés les forums, les groupes de discussion en ligne, les blogues, les sites de réseautage social, les sites multimédias sociaux (pour les photos et vidéos) de même que les mondes virtuels, ainsi que des jeux accessibles en ligne (Hsinchun et al., 2012; O'reilly, 2007). De plus, il y a la croissance de nouvelles technologies de gestion, d'analyse et de stockage des données comme l'informatique en nuage (ou le *cloud computing*) qui créent davantage de données. Dépassant ainsi en quantité et en relations mutuelles les capacités des architectures et infrastructures informatiques des entreprises existantes (Chen et al., 2014).

D'un autre côté, la transformation du modèle d'affaires fait que les entreprises sont passées de l'orientation produit à l'orientation service, où la valeur de l'organisation selon ses clients est mesurée par l'efficacité du service et non par l'utilité du produit (Krishnan, 2013). Cela leur impose la nécessité de produire plus de données en termes de produits et services pour répondre à chaque segment et canal de clients, par exemple sur les médias sociaux, les enquêtes et les forums rétroaction directe. Cette transformation représente les affaires électroniques qui contiennent plusieurs modèles d'affaires tels que le commerce électronique interentreprises (B2B) ou celui destiné aux particuliers (B2C) (Baltzan, 2018; Krishnan, 2013) et où les transactions entre clients et fournisseurs ainsi que le partage d'informations se font en ligne. Cela finit par générer de plus en plus de données où la quantité de données produites et consommées par chaque organisation dépasse aujourd'hui ce que la même organisation produisait avant la transformation de l'entreprise (Krishnan, 2013). Cette énorme quantité de données hétérogènes de plus en plus nombreuses pose un problème de stockage et de gestion avec des exigences modérées en matière d'infrastructure matérielle et logicielle (Chen et al., 2014).

Les données massives sont très importantes dans l'entreprise moderne (Bellavance et Labrie, 2017) et cette importance réside dans le processus d'analyse de ces données qui a pour but de les « faire parler », c'est-à-dire de transformer ces données massives aux données intelligentes (*smart data*) permettant de faciliter le processus de prise de décision de l'entreprise (Bellavance et Labrie, 2017). La capacité de traiter et d'utiliser les données massives est reconnue comme le principal moteur de développement de l'organisation, ainsi que la base de la survie du marché, du succès de l'innovation, de l'amélioration de la compétitivité et une prise de décision plus efficace (Hagen et al., 2013; Olszak et Zurada, 2019; Rising et al., 2014).

Toutes les entreprises détiennent des données provenant de différentes sources qui sont essentielles à leur fonctionnement quotidien, mais l'accumulation de ces données qui n'est pas souvent structurée rend ces entreprises impuissantes et incapables de les traiter efficacement. Souvent, elles négligent les informations qu'elles peuvent dégager de ces données et leurs utilités à l'entreprise. Toutes les données ne sont pas forcément utiles et exploitables par l'entreprise, mais dans ces données massives, il y a forcément une information qui permettra à l'entreprise de s'améliorer et de dégager plus de profits. Par exemple, pour le processus de gestion des relations clients, l'entreprise peut avoir des milliers de commentaires des clients exprimant leurs avis sur l'entreprise et son service allant jusqu'à donner des suggestions d'amélioration de la qualité de service à cette entreprise (Bellavance et Labrie, 2017). Mais, si l'entreprise ne voit jamais ces commentaires parce qu'elle ne sait pas comment les traiter, elle ne saura pas satisfaire et fidéliser ses clients. L'article de Bellavance et Labrie (2017) évoque l'exemple du précieux deux dollars qui est le coût de lecture et de compilation d'un commentaire de l'un des clients de l'entreprise. Ce coût peut atteindre des dizaines de milliers de dollars pour les grandes entreprises, mais la plupart des entreprises ne prennent pas le temps de le faire même si ces commentaires peuvent cacher de précieuses informations utiles à l'entreprise. Souvent, l'information dont l'entreprise a besoin est devant ses yeux se

cachant dans ces données massives, mais pour y accéder il faut savoir chercher cette information en analysant ces données.

1.1.1 Caractéristiques et cycle de vie des données massives

Nous sommes à l'ère de l'abondance des données provenant de différentes sources qui se multiplient grâce au développement technologique. Cette abondance de données appelées données massives, méga données ou *Big Data* change la gestion des entreprises actuelles et leurs relations à leurs clients (Bellavance et Labrie, 2017). Les données massives représentent, à la fois, l'abondance de données de façon continue et la capacité technologique d'analyser et de traiter ces données pour en tirer de la valeur qui est de l'information de qualité permettant une bonne prise de décision et en temps réel (Karoui et al., 2014; Ohlhorst, 2012). Dégager de la valeur à partir de ces données massives n'est pas une chose simple à faire. C'est pour cela qu'il faut voir les données massives comme un concept multidimensionnel avec les quatre caractéristiques suivantes[4] : un *volume* énorme de données (V), créées avec une grande *vélocité* (V), de grande *variété* (V) et de *complexité* élevée (C). Le volume caractérise la quantité de données créées chaque jour qui est en augmentation continue et qui a été estimée mondialement pour l'année 2012 à 2,5 x 10^6 téraoctets[5] ou 2,5 exaoctets (10^{18} octets) par jour et on s'attend qu'elle double chaque 40 mois (Coleman et al., 2016; McAfee et al., 2012). L'estimation des données collectées par Walmart, à lui seul, à partir des transactions de ses clients est de plus de 2,5 pétaoctets (10^{15} octets) de données par heure (McAfee et al., 2012). Selon Taleb et al. (2018) et Zikopoulos et Eaton (2011), le volume des données stockées en 2000 était de 800 000 pétaoctets (1 PB = 10^{15} octets) alors que le volume estimé pour l'année 2020 est de 35 zettaoctets (1 ZB = 10^{21} octets). Cette

[4] Ces caractéristiques sont souvent présentées avec l'acronyme VVVC (Coleman et al., 2016; Ohlhorst, 2012).
[5] 1 téraoctet = 10^{12} octets

augmentation exponentielle du stockage des données provient des moteurs de recherche sur le Web comme Google et Yahoo qui permettent l'interrogation et l'agrégation de grandes quantités de données peu structurées, ainsi que les applications telles que Facebook, Amazon, Twitter, YouTube, les capteurs d'Internet des objets (ou *Internet of Things*) et les téléphones intelligents mobiles qui sont les principaux acteurs et générateurs de données (Taleb et al., 2018). La vélocité c'est la rapidité avec laquelle on produit les données qui est en croissance permanente, et cela grâce aux nouvelles technologies de l'information qui permettent d'obtenir l'information à temps réel (Coleman et al., 2016). La variété des données émane de différentes sources, de l'interne ou de l'externe de l'entreprise, et qui sont de différents formats, structurées (documents, image, etc.) ou non structurées (données GPS, capteurs, tweets, etc.) (Coleman et al., 2016). La complexité de ces données réside dans le fait qu'elles sont à plusieurs variables, de différents formats et débits ainsi que de résolutions multiples (Coleman et al., 2016). On obtient ainsi une énorme quantité de données qui covarient entre elles pour une entité cible, par exemple le service des achats, tout en provenant de différentes sources, dans différents formats et de diverses granularités, c'est-à-dire des niveaux différents de précision (Coleman et al., 2016).

Les données sont classifiées selon différentes catégories : structurées, semi-structurées et non structurées. Les données structurées se trouvent dans les bases de données traditionnelles (SGBD[6], SQL, Excel, etc.) où les données sont structurées, déjà définies et indexées sous forme de tables ce qui facilite la filtration et le tri de ces données (Ohlhorst, 2012). Contrairement aux données structurées, les données non structurées ne sont pas organisées sous forme de tables et ne peuvent pas être interprétées par des bases de données (Ohlhorst, 2012). Il peut s'agir de messages ou de vidéos circulant par l'intermédiaire de divers dispositifs (courriels, messagerie instantanée, réseaux sociaux, etc.). Les données semi-structurées sont les données se trouvent entre les deux, c'est-à-dire qu'elles n'ont pas de structure formelle comme c'est le cas d'une base de données

[6] Une liste des abréviations de nature technique est fournie au début du présent document.

contenant des tables et des relations. Elles peuvent néanmoins avoir des étiquettes et d'autres marqueurs pour séparer les éléments et fournir une hiérarchie des enregistrements et des champs qui définissent les données (Ohlhorst, 2012).

L'écosystème des données massives est organisé comme un cycle de vie de la chaîne de valeur, qui va de la création à la visualisation des données (Taleb et al., 2018). Au cours de leur cycle de vie, ces données massives passent par différentes étapes qui sont : la génération, la collecte, l'acquisition, la transmission, le stockage, le prétraitement, l'analyse et la visualisation des données (Taleb et al., 2018). Cela dans le but de transformer ces données massives en information exploitable pour l'entreprise et ses dirigeants.

1.1.2 Données massives dans les PME : applications, atouts et défis

Par définition, au Canada, la PME possède un nombre d'employés inférieur ou égal à 500 et un chiffre d'affaires annuel ne dépassant pas 50 millions de dollars (Filion et Allali, 2007). Ces entreprises constituent aujourd'hui la majeure partie du système productif et son principal facteur de renouvellement (Salles, 2006). En décembre 2019, les PME représentaient 99,8% du total des entreprises canadiennes avec employés pour un nombre total de 1,22 million de PME contre 2 978 grandes entreprises et au Québec ces PME représentaient 254 645 (Gouvernement du Canada, 2020). Cela justifie le rôle important que jouent ces PME dans le marché d'emploi canadien.

Le concept des données massives a déjà fait ses preuves dans différentes entreprises de différents secteurs d'activités. C'est le cas des entreprises Harry's et UPS qui ont utilisé le principe analytique des données massives dans leurs processus de prise de décision. Cette stratégie leur a permis de dégager des bénéfices et de dépasser leurs concurrents. Par exemple, Harry's, une entreprise New-Yorkaise vendant des rasoirs, a

proposé à ses clients de leur livrer des rasoirs au besoin en se basant sur l'analyse des habitudes de navigation de ses clients potentiels sur le Web pour cibler sa clientèle et leur proposer un service adéquat (Bellavance et Labrie, 2017). Quant à UPS (*United Parcel Service*) qui est une entreprise postale mondiale, elle s'est servie de la technologie analytique des données massives sur sa flotte de camions dans le but de détecter et de prévenir la défaillance liée à la surchauffe et la vibration et ainsi réduire les coûts (Karoui et al., 2014). Les PME peuvent aussi bénéficier de l'utilisation des données massive qui leur permettent de croître. Prenons l'exemple d'un vendeur de bois à la façon traditionnelle qui met des pancartes sur le bord de la route pour indiquer le prix de ses produits. S'il recueille plus de données sur ses clients et leur tendance de consommation de bois de chauffage ainsi que la saisonnalité de cette consommation, il pourra répondre à leur besoin au bon moment. Cela, afin de les fidéliser en leur proposant des services de livraison ou en leur faisant découvrir d'autres produits qui peuvent les intéresser. Ainsi, il pourra augmenter ses ventes et faire croître son entreprise (Bellavance et Labrie, 2017).

L'enquête d'*Oxford Economics* en 2013 a déclaré que la priorité stratégique de la croissance des PME est la technologie et l'innovation. Les données massives sont considérées comme l'un des principaux moteurs de cette croissance permettant ainsi d'accroître la flexibilité, la productivité, la réactivité, l'anticipation et la capacité à répondre aux besoins des clients, et de meilleures prises de décisions (Sen et al., 2016). C'est pourquoi les PME devraient s'intéresser aux données massives pour obtenir un avantage concurrentiel et une croissance, et ce, en analysant leurs performances passées et en les combinant à des données externes hétérogènes leur permettant de comprendre le comportement du marché et de faire des prédictions basées sur ces informations (Karoui et al., 2014; MacInnes, 2013; Preez, 2014; Sen et al., 2016). C'est cette capacité de croiser des données hétérogènes en temps réel et d'imaginer des combinaisons et corrélations possibles qui constitue la véritable richesse d'un projet de *Big data* (Karoui et al., 2014). Les PME créent et stockent de plus en plus de données transactionnelles sous format numérique, et collectent des informations de performance plus précises et détaillées sur

tout, allant des inventaires aux jours de maladie de leur personnel, et ainsi exposer la variabilité et améliorer les performances (Brown et al., 2011; Sen et al., 2016). De ce fait, les données massives permettent d'utiliser ces données collectées pour une segmentation plus étroite des clients et donc des produits ou services plus précis. Cela aide les organisations à améliorer l'efficacité de la formulation de la demande et de la planification de l'offre (Brown et al., 2011; Sen et al., 2016). Puisque les PME manquent généralement de ressources humaines, matérielles et financières (Filion et Allali, 2007), il existe des solutions infonuagiques et des systèmes libres (*open sources*) rentables et abordables que les PME peuvent utiliser pour traiter et analyser ses données sans devoir investir en matériel et /ou en logiciel (Sen et al., 2016). Plusieurs services d'hébergement Web peuvent offrir la puissance de calcul, le stockage et les plateformes pour l'analyse. Transformant ainsi le marché des données massives en un système de paiement basé sur la consommation et l'utilisation réelles de ces outils technologiques (Ohlhorst, 2012).

Il existe des défis et obstacles à l'exploitation des PME des données massives. Coleman et al. (2016) a identifié les principaux obstacles. Par exemple, le manque de compréhension des principes de l'analyse des données massives par les représentants et dirigeants des PME. La dominance des PME spécialistes du domaine qui ne cherchent pas à s'ouvrir à d'autres filières et à l'actualité des affaires. Cela crée une barrière de culture et un conservatisme chez ces PME. La pénurie d'expertise interne en matière d'analyse des données par le manque de cas d'affaires existants sur lesquelles ces PMEs peuvent se baser et avoir un retour d'expériences. S'ajoutent la pénurie de consultants utiles et abordables offrant des services d'analyse d'affaires, ainsi que les préoccupations concernant la sécurité et la confidentialité des données. Enfin, les barrières financières restent l'obstacle principal à la croissance des PME du fait que les PME ont moins accès au financement par emprunt par rapport aux grandes entreprises.

1.2 QUALITÉ DES DONNÉES OU QUALITÉ DE L'INFORMATION

La croissance des entrepôts de données et l'accès direct à l'information provenant de diverses sources, par les gestionnaires et les utilisateurs de l'information, ont accru le besoin et la sensibilisation à l'information de haute qualité dans les organisations. Cette recherche de qualité est ainsi devenue une préoccupation essentielle et un domaine actif de la recherche sur les systèmes d'information de gestion (SIG) (Lee et al., 2002).

Dans le monde actuel des affaires, on suppose que le fait d'avoir une donnée en sa possession veut dire forcément que nous avons de la connaissance. Par conséquent, détenir une énorme quantité de données (ou données massives) veut dire par déduction que nous avons une grande connaissance. Cela n'est pas vrai, car on risque d'être submergé par ces données, en plus de ne pas savoir comment les traiter. C'est le cas de la plupart des entreprises et plus particulièrement dans les PME. Selon Labrinidis et Jagadish (2012), nous avons l'habitude de penser que les données massives nous disent toujours la vérité, mais, ce n'est pas toujours le cas, donc il faut toujours être prêt à faire face à des données erronées. D'où la nécessité de l'évaluation de la qualité de ces données tout au long du processus de transformation (ou cycle de vie) de ces données massives qui passent par plusieurs étapes: la génération, la collecte, l'acquisition, la transmission, le stockage, le prétraitement, l'analyse et la visualisation des données (Taleb et al., 2018).

Il y a toujours eu des problèmes de qualité des données, et ce, avant même l'avènement des données massives, mais celles-ci augmentent considérablement les besoins en nettoyage qui devient plus compliqué, notamment à cause des caractéristiques des données massives (volume, vélocité, variété, complexité) et des sources multiples de celles-ci (Oliveira et al., 2005; Rahm et Do, 2000; Taleb et al., 2015). Des problèmes de qualité des données surviennent lorsque les exigences de qualité ne sont pas satisfaites sur les valeurs des données (Fürber et Hepp, 2011; Taleb et al., 2018). Ils peuvent être dus à plusieurs facteurs ou intervenus à différentes étapes du cycle de transformation de ces données, par exemple au niveau : des sources de données (manque de fiabilité), de la

génération des données (erreur de saisie) et de *l'acquisition des données* (collecte, prétraitement, transmission), etc. (Taleb et al., 2018).

1.3 PROBLÈME DE RECHERCHE

Le domaine des données massives est actuellement considéré comme l'un des domaines de recherche les plus dynamiques au monde (Chen et al., 2012; Davenport et al., 2012; LaValle et al., 2011; Olszak et Zurada, 2019). Mais, bien que ce soit un sujet d'actualité et qu'il existe déjà des articles abordant ce phénomène, il y a des lacunes dans la revue de littérature. Premièrement, les études qui ont été faites se focalisent plus sur les données massives dans les grandes entreprises que dans les PME. Deuxièmement, ces recherches s'intéressent au côté analytique de ces données massives ainsi qu'aux compétences reliées à cela en oubliant le côté managérial de gestion de l'information. Elles associent les données massives à des outils statistiques sophistiqués de gestion de ces données qui nécessitent des ressources spécifiques (la compétence et le financement). Cela en faisant l'inventaire de nouvelles technologies existantes sur les données massives qui sont plus destinées à des spécialistes et qui sont incompréhensibles et difficiles à mettre en place dans le contexte des PME. Troisièmement, ces études ne s'intéressent pas à la finalité et l'objectif principal pour lequel ces outils analytiques de données massives ont été créés qui est d'assurer une qualité de l'information. Ainsi, on parle plus souvent de la performance de ces outils et de leur côté révolutionnaire, mais sans relier cela à la valeur ajoutée concernant la qualité de l'information de l'entreprise. Toutefois, il existe quelques études et articles sur le sujet des données massives. Mais, très peu ont essayé de comprendre la manière de gérer et traiter ces données massives dans les PME, plus particulièrement dans le but d'obtenir une meilleure qualité de l'information. Ce constat s'applique aussi aux PME faisant des affaires dans la province du Québec, au Canada.

Plus précisément, le but de cette étude est de mieux comprendre comment les PME acquièrent, évaluent, traitent et stockent les données massives tout en assurant la

qualité de leurs informations. À cet effet, l'objectif général est d'explorer les pratiques permettant d'assurer une haute qualité de l'information au sein des PME, et ce, à travers l'articulation des différentes configurations d'activités et d'outils de gestion que l'on retrouve au sein des PME à l'ère des données massives. Par la même occasion, nous explorerons aussi ce qui mène à une absence et une qualité moyenne de l'information. Cette recherche permettra de répondre à la question de recherche suivante : Quels sont les activités et les outils de traitement des données qui permettent d'assurer la qualité de l'information au sein des PME et comment ceux-ci s'articulent-ils à l'ère des données massives ?

Plus spécifiquement cette étude répondra aux sous-questions suivantes :

- Quels sont les différentes configurations d'activités et d'outils de traitement des données qui permettent d'assurer une haute qualité de l'information au sein des PME et comment ceux-ci s'articulent-ils à l'ère des données massives ?
- Quels sont les différentes configurations d'activités et d'outils de traitement des données qui permettent d'assurer une qualité moyenne de l'information au sein des PME et comment ceux-ci s'articulent-ils à l'ère des données massives ?
- Quels sont les différentes configurations d'activités et d'outils de traitement des données qui mènent à une faible, voire à une absence de qualité de l'information au sein des PME et comment ceux-ci s'articulent-ils à l'ère des données massives ?

CHAPITRE 2 – CONTEXTE THÉORIQUE

Cette section définit l'approche configurationnelle. Ensuite, elle présente le cycle de vie des données massives et ses concepts. Puis, elle aborde la qualité de l'information. Enfin, elle présente les composantes du cadre conceptuel et sa figure, ainsi que la représentation du modèle configurationnel.

2.1 DÉFINITION DE L'APPROCHE CONFIGURATIONNELLE POUR LE TRAITEMENT DE L'INFORMATION DANS LES PME

Les théories configurationnelles émanent de la compréhension des modèles et des combinaisons d'éléments et de la façon dont ils, en tant que configurations, provoquent certains résultats (de Guinea et Raymond, 2020). Une configuration est une combinaison spécifique de variables causales (appelées éléments ou conditions) qui génèrent un résultat d'intérêt (El Sawy et al., 2010; Rihoux et Ragin, 2008). L'approche configurationnelle repose sur la prémisse fondamentale selon laquelle les modèles d'attributs présenteront des caractéristiques différentes et conduiront à des résultats différents selon la manière dont ils sont organisés (Fiss, 2007). Elle suggère que les organisations sont mieux comprises comme des grappes de structures et de pratiques interconnectées, plutôt que comme des entités modulaires ou faiblement couplées dont les composantes peuvent être comprises isolément. Les partisans de cette approche adoptent donc une vision systémique et holistique des organisations, où des modèles ou des profils plutôt que des variables indépendantes individuelles sont liés à un résultat (Delery et Doty, 1996) qui est, dans la présente étude, la qualité de l'information. Il s'agit d'une rupture nette avec le paradigme linéaire prédominant (Meyer et al., 1993). Plutôt que d'impliquer une causalité singulière et des relations linéaires, l'approche configurationnelle suppose une causalité complexe et des relations non linéaires où « les variables dont la relation

causale dans une configuration peut être sans relation ou même inversement liée dans une autre » (Fiss, 2007, p. 3).

Dans les recherches stratégiques ou managériales, on utilise souvent les approches basées sur la variance (telle que la régression et la modélisation d'équations structurelles) comme techniques d'analyse des données afin d'expliquer un phénomène précis (de Guinea et Raymond, 2020). Mais, ces méthodes ont des limites et ne s'adaptent pas à notre recherche, notamment à cause de leurs deux caractéristiques qui sont : l'unifinalité et la symétrie causale. Plus précisément, les études sur la variance sont basées sur l'unifinalité qui présume qu'un facteur ou plusieurs facteurs spécifiques conduisent à un résultat donné. De plus, ces études utilisent des techniques fondées sur la corrélation qui supposent une symétrie causale parce que les corrélations sont par nature symétriques (Fiss, 2011; Ragin, 2009). Dans la même direction, certains chercheurs affirment qu'un résultat d'intérêt organisationnel résulte rarement de facteurs causaux uniques (Woodside, 2013). Par conséquent, l'adoption de l'approche configurationnelle est plus compatible avec notre point d'intérêt qui est la qualité de l'information.

L'approche configurationnelle, contrairement à l'approche de la variance, permet l'équifinalité et l'asymétrie causale (de Guinea et Raymond, 2020). Plus précisément, l'analyse configurationnelle met l'accent sur le concept d'équifinalité, qui se réfère à une situation où « un système peut atteindre le même état final, à partir des conditions initiales différentes et par une variété de chemins différents » (Fiss, 2007, p. 3). Alors que l'unifinalité suppose qu'il existe une configuration optimale, l'équifinalité suppose que deux ou plusieurs configurations organisationnelles peuvent être également efficaces pour atteindre une qualité de l'information, et ce, même si elles sont confrontées aux mêmes contingences (Galunic et Eisenhardt, 1994; Gresov et Drazin, 1997). Autrement dit, les éléments du système, soit les activités et les outils de gestion du cycle de traitement de l'information, qui peuvent être combinés d'une façon multiple et complexe pour produire un même résultat concernant la qualité de l'information désirée

(Meyer et al., 1993). Ce qui veut dire que la qualité de l'information peut être atteinte de manière égale grâce à différentes configurations d'activités et d'outils de gestion lors du cycle de transformation de l'information (Ragin, 2000). L'asymétrie causale signifie aussi que les causes menant à la présence d'un résultat d'intérêt peuvent être très différentes de celles conduisant à l'absence de ce résultat (Rihoux et Ragin, 2008). Si différentes configurations d'activités et d'outils de traitement des données sont équifinales pour conduire à une haute qualité de l'information, cela signifie également qu'un élément particulier dans une configuration doit être présent pour que la haute qualité de l'information se produise, tandis que dans une autre configuration, ce même élément sera absent. Autrement dit, le même élément pourrait permettre dans une configuration la réalisation d'une haute qualité de l'information; tandis qu'il l'inhiberait dans une autre (de Guinea et Raymond, 2020). Ainsi, dans le contexte de cette étude, le concept d'asymétrie causale permet aux configurations équifinales de varier en fonction des différents niveaux de qualité de l'information souhaités par les PME et leurs dirigeants (Fiss, 2011). En résumé, un ensemble spécifique de configurations peut être associé à une haute qualité de l'information, tandis qu'un ensemble différent peut conduire à une qualité moyenne de l'information, et encore un autre ensemble de configurations peut être associé à une faible, voire à une absence de qualité de l'information.

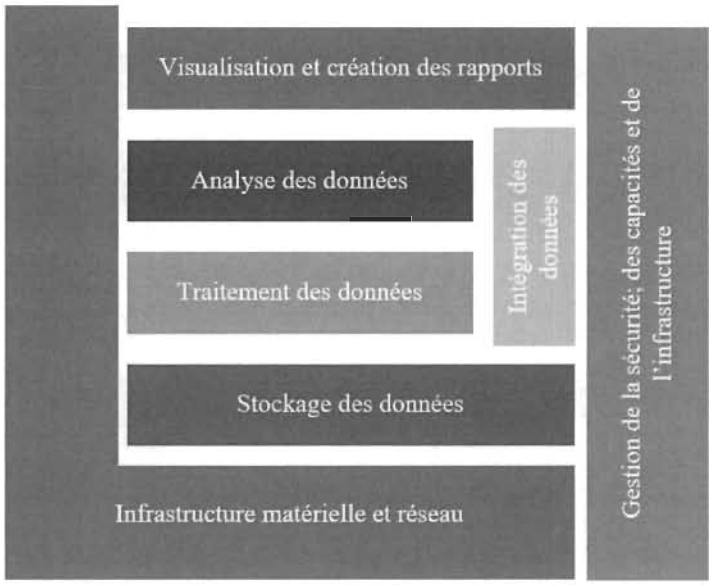
2.2 CYCLE DE VIE DES DONNÉES MASSIVES

Selon Kiron et al. (2012) dans une étude publiée dans la revue de la direction du MIT Sloan, il existe une forte corrélation entre générer un avantage concurrentiel et l'innovation concernant l'analyse et l'efficacité de l'entreprise lors de la gestion du cycle de transformation de l'information. Ce cycle consiste à capturer les données, analyser l'information, agréger et intégrer les données, utiliser les informations pour guider la stratégie future et, enfin, diffuser les informations et les idées (Coleman et al., 2016). En particulier, le processus de production des informations précieuses (ou de valeur) pour l'utilisateur final nécessite des infrastructures pour la collecte, le traitement et la gestion

d'énormes volumes de données. Selon Labrinidis et Jagadish (2012), que les données soient massives ou non, le *pipeline* d'analyse des données comporte plusieurs étapes. Dans le contexte des données massives, à chaque étape, il y a du travail à faire et des défis à relever. Ces étapes sont le cycle de vie de la gestion des données qui est composé de : l'acquisition, l'organisation, l'analyse et la décision (Coleman et al., 2016). Elles sont illustrées à la Figure 1 qui représente une architecture de référence standard pour l'analyse des données massives, qui se caractérise par une structure à plusieurs niveaux.

Figure 1

Infrastructure informatique de référence pour l'analyse des données massives[7]



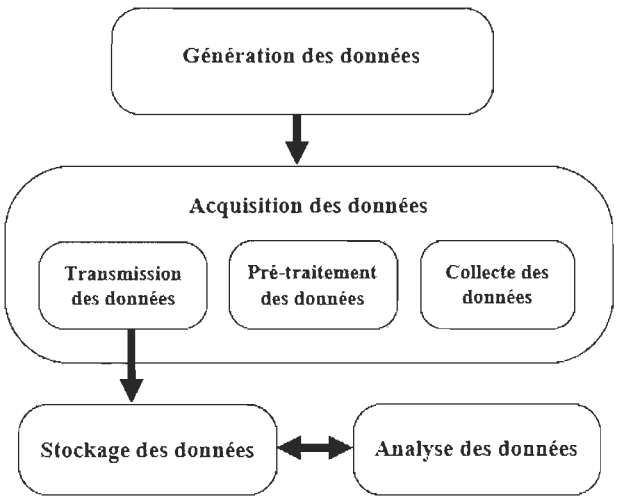
Dans cette étude, nous nous intéresserons plus particulièrement aux activités de : visualisation et création des rapports, analyse des données, traitement des données, stockage des données et intégration des données. Conséquemment, nous ne traiterons pas de l'infrastructure matérielle et réseau, de la gestion de la sécurité, des capacités et de l'infrastructure.

[7] Tiré de Coleman et al. (2016, p. 5).

Les articles de Taleb et al. (2015) et Hu et al. (2014), portant sur le prétraitement des données massives, considèrent que dans les systèmes des données massives, les données sont une ultime source de connaissances. Dans son cycle de vie, les données voyagent à travers quatre différentes étapes, comme le montre la Figure 2 ci-dessous, qui sont : *la génération des données*, *l'acquisition des données*, *le stockage des données* et *l'analyse des données* (Taleb et al., 2015).

Figure 2

Cycle de vie des données massives (2015)[8]

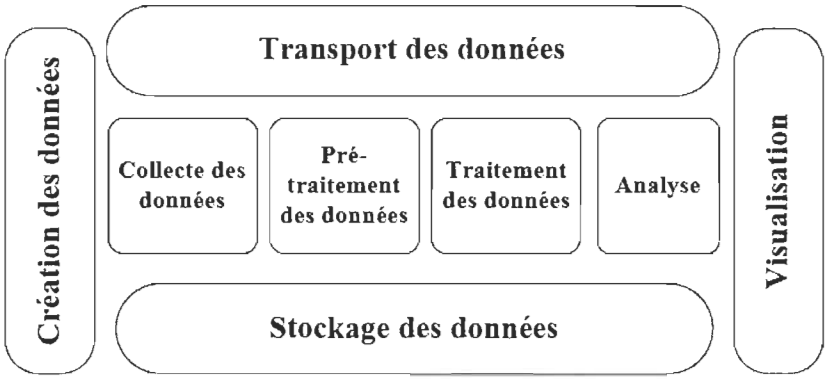


Pour sa part et tel qu'illustré par Taleb et al. (2018) concernant la qualité des données massives, l'écosystème des données massives est organisé comme un cycle de vie de la chaîne de valeur, qui va de la création des données à la visualisation. Il décrit les étapes principales du cycle de vie des données massives par la Figure 3 suivante qui débute par la création des données et se termine par la visualisation des données, en passant par des systèmes de collecte, de traitement, d'analyse, de transport et de stockage des données.

[8] Tiré de Taleb et al. (2015, p. 2).

Figure 3

Cycle de vie des données massives (2018)[9]



On remarque dans ces deux figures sur le cycle de vie des données massives que ce processus commence par la phase de création ou *génération des données* et finit par l'étape de *visualisation des données*. Donc, les activités principales de la gestion du cycle de vie des données massives sont : *la génération des données*, *l'acquisition des données* ; qui comporte la collecte, le prétraitement et la transmission des données ; *le stockage des données*, *l'analyse des données* et pour finir *la visualisation des données*. Dans cette partie, nous allons voir de près la signification de chacune de ces étapes de transformation des données massives en information exploitable par l'utilisateur final.

2.2.1 Génération des données

C'est la phase où les données sont créées (Hu et al., 2014; Taleb et al., 2018). Dans ce cas, le terme données massives désigne l'ensemble des données volumineuses, diversifiées et complexes qui sont générées à partir de diverses sources de données responsables de la création de ces données. Par exemple, les capteurs utilisés pour recueillir les informations sur le climat, les appareils de surveillance, les publications sur

[9] Tiré de Taleb et al. (2018, p. 2).

les réseaux sociaux, les vidéos et images fixes, les enregistrements des transactions, les indices boursiers, la localisation GPS des téléphones portables, les flux de clics et autres sources numériques disponibles.

Il existe d'énormes quantités de données provenant d'Internet en termes de recherche d'entrées, de messages de forums, d'enregistrements de discussions et des messages de microblogues qui sont générées tous les jours (Chen et al., 2014). Ces données peuvent être sans valeur si utilisées individuellement, mais si elles sont accumulées et exploitées d'une manière efficace, elles génèreront des informations de la vie quotidienne des gens qui est utile à l'entreprise. Puisque les données changent constamment, une analyse rapide de ces données est nécessaire et les résultats analytiques doivent être renvoyés dans un délai très court. Cela grâce à l'utilisation de l'analyse en temps réel qui est principalement utilisée dans le commerce électronique et la finance (Chen et al., 2014). Par exemple, en donnant des indications sur le comportement, les habitudes, les loisirs et les humeurs émotionnelles des utilisateurs ou consommateurs potentiels. Les principales sources de données massives sont les informations opérationnelles et commerciales dans les entreprises, les informations logistiques et de détection dans l'Internet des objets, les informations d'interaction humaine et les informations de position, de même que les données générées dans la recherche scientifique.

Cette étude se basera sur les données provenant des systèmes internes de l'entreprise (p. ex. les informations opérationnelles et commerciales, etc.) et celles provenant des systèmes externes de l'entreprise (p. ex. Internet, médias sociaux, contrats et échanges avec des partenaires d'affaires, etc.).

2.2.1.1 Consultation des données provenant des systèmes internes de l'entreprise

Le rapport d'analyse « les applications des données massives au monde réel » de IBM, publié en 2013, indique que les données internes des entreprises sont les principales

sources des données massives (Chen et al., 2014). Ces données internes se composent principalement des données du commerce en ligne (ou *e-commerce*) et des données analytiques en ligne, dont la plupart sont des données historiquement statiques et sont gérées par des systèmes de gestion de base de données (SGBD) de manière structurée. En outre, les données de production, les données d'inventaire, les données de vente et les données financières, etc., constituent également des données internes à l'entreprise. Elles visent à capturer les informations et les activités axées sur les données dans les entreprises, de manière à enregistrer toutes les activités sous forme de données internes.

Pour la gestion de leurs données internes, la plupart des entreprises utilisent des progiciels de gestion intégré (PGI) (ou *Enterprise Resource Planning - ERP)* qui, comme son nom l'indique, « intègre tous les services et toutes les fonctions d'une organisation en un seul système d'information (ou en un ensemble intégré de systèmes d'information), de sorte que les employés peuvent prendre des décisions éclairées après avoir examiné les données concernant toutes les activités d'affaires de l'entreprise dans son ensemble » (Baltzan, 2018, p. 278). De plus, les entreprises peuvent utiliser les systèmes d'information géographique et données géospatiales ou système de positionnement global (GPS) qui utilise les données géospatiales pour guider n'importe qui du point A au point B (Krishnan, 2013).

2.2.1.2 Consultation des données provenant des systèmes externes de l'entreprise

Avant, c'était les employés qui généraient les données, mais, aujourd'hui, pour une organisation ce sont tous les acteurs de la chaîne logistique qui génèrent les données, par exemple les clients, les partenaires, les compétiteurs de l'organisation (Krishnan, 2013). Identifiées par (Krishnan, 2013), ce qui suit est une énumération de quelques exemples de sources de génération des données, , qui peuvent être considérées comme sources externes en entreprise. Les données de la machine (ou appareil) que nous utilisons au quotidien allant des appareils industriels aux appareils personnels. Le journal des applications qui est une autre forme de génération des données par la machine qui génère

des journaux à des rythmes et des formats différents, par exemple les machines à rayons X, les scanners corporels dans les aéroports, les données générées par les nouvelles technologies comme les téléphones portables et les tablettes. Les journaux de flux des clics (ou *Clickstream*) sont les statistiques d'utilisation de la page Web qui sont capturées dans les données de parcours de clics. Ces documents, numériques ou papier, fournissent un aperçu du comportement de l'utilisateur sur la page Web et peuvent indiquer des données très utiles pour l'analyse du comportement et de l'utilisabilité, le marketing et la recherche générale. Les données externes ou tierces sont plusieurs ensembles de données que les organisations achètent ou obtiennent sous forme de flux provenant de sources externes. Celles-ci sont souvent en gros volumes, de différents formats et non structurées, par exemple les données météorologiques. Les courriels qui sont générés quotidiennement par ses employés, ses clients et ses dirigeants. Les contrats qui sont liés aux différentes activités de l'entreprise comme les ressources humaines, les services juridiques, les fournisseurs, les clients, etc. Enfin, les médias sociaux, par exemple Facebook et Twitter qui permettent de partager des données dans plusieurs formats (texte, image ou vidéo).

2.2.2 Acquisition des données

Le processus d'acquisition des données comprend trois sous-étapes : la collecte, le prétraitement et la transmission des données (Hu et al., 2014). Après avoir collecté les données brutes des données massives pendant la phase de l'acquisition, elles seront transmises vers un système de stockage approprié pour une analyse ultérieure grâce aux différentes applications analytiques (Chen et al., 2014). Les données collectées contiennent souvent beaucoup de données redondantes ou n'ayant pas de valeur ajoutée, ce qui prend plus d'espace de stockage et risque de fausser les résultats de l'analyse des données. Pour réduire ce problème, la sous-activité prétraitement des données est nécessaire pour le tri et le nettoyage de ces données grâce à des technologies et pratiques d'intégration, d'enrichissement, de transformation, de réduction et de nettoyage des

données qui seront effectuées avant les phases de stockage et d'analyse des données (Taleb et al., 2018).

Dans cette étude, les deux sous-étapes « collecte des données » et « prétraitement des données » seront fusionnées en une seule sous-étape appelée « préparation des données » et la sous-étape « transmission des données » sera conservée.

2.2.2.1 Préparation des données

Cette première sous-activité de l'étape de l'acquisition des données comprend, à son tour, deux sous-activités ou dimensions, soit la collecte des données qui consiste à utiliser des techniques spéciales de collecte des données pour acquérir des données brutes à partir d'un environnement de génération de données spécifique (Chen et al., 2014). Ensuite, le prétraitement des données qui est une étape obligatoire est indispensable pour affiner et valoriser les données (Taleb et al., 2015).

Le processus de collecte des données doit être bien conçu sinon il y aura des conséquences et des répercussions sur les étapes suivantes comme l'analyse des données (Hu et al., 2014). Donc, si les données collectées en amont sont erronées elles engendreront des résultats invalides. Dans le même temps, les méthodes de collecte de données dépendent non seulement des caractéristiques physiques des sources de données, mais également des objectifs de l'analyse des données (Hu et al., 2014). Il existe de nombreuses méthodes de collecte des données, mais dans le cadre d'une PME des exemples des plus utilisés sont donnés comme suit. L'intégration des données qui consiste à combiner les données provenant de sources multiples pour une vue unifiée des données pour l'utilisateur (Taleb et al., 2015). Par exemple l'utilisation d'un progiciel de gestion intégré (PGI). Cette intégration des données permet la collecte des données provenant des différents acteurs de la chaîne de valeur grâce aux saisies des données et des opérations transactionnelles quotidiennes. L'information transactionnelle « comprend toutes les données contenues dans un même processus d'affaires ou une même unité de travail »

(Baltzan, 2018, p. 44). Elle est utilisée par les gestionnaires pour effectuer ou exécuter les tâches opérationnelles quotidiennes leur permettant la prise des décisions structurées comme la définition du stock minimum. Pour la gestion et la collecte de cette information transactionnelle, on utilise des systèmes de traitement des transactions (STT) qui sont destinés au suivi des opérations utilisé au niveau opérationnel comme le système d'enregistrement des commandes (Baltzan, 2018). Au niveau tactique et stratégique, il y a l'informatique analytique qui « inclut toutes les données transactionnelles compilées ou agrégées ainsi que les données externes comme celles qu'on obtient auprès des sources extérieures au marché et à l'industrie » (Baltzan, 2018, p. 46). Par exemple, les montants des ventes groupées par région, les statistiques sur les produits et les prévisions de croissance. Elle est utilisée pour la prise de décisions semi-structurées, par exemple la construction d'une nouvelle usine de fabrication.

En raison de la grande variété des sources de données, les ensembles des données collectées varient en ce qui concerne le bruit, la redondance et la cohérence. Le stockage et le transfert de ces données brutes seront sans doute un gaspillage et entraîneraient des coûts inutiles. De plus, certaines méthodes analytiques ont de sérieuses exigences en matière de la qualité des données (Chen et al., 2014; Hu et al., 2014; Taleb et al., 2015). Par conséquent, afin de permettre une analyse efficace des données, il est nécessaire de passer par une phase de prétraitement des données dans laquelle des activités telles que le nettoyage des données, la déduplication, la compression, le filtrage et la conversion de format ont lieu (Taleb et al., 2015). D'autres tâches de prétraitement sont nécessaires telles que l'intégration des données (Bansal, 2014; Dong et Srivastava, 2013; Taleb et al., 2015) et la fusion (Taleb et al., 2015; Yang et al., 2014) de plusieurs sources hétérogènes. Ces tâches ont aussi un effet considérable sur les données transformées qui en résultent et sur l'analyse globale des résultats (Taleb et al., 2015). L'étape du prétraitement est obligatoire et indispensable pour affiner et valoriser les données et améliorer la précision de l'analyse tout en réduisant les frais de stockage (Chen et al., 2014; Hu et al., 2014; Taleb et al., 2015). Dans cette partie nous allons définir les différentes activités du prétraitement des

données qui sont : l'intégration, le nettoyage, la transformation, la réduction et l'amélioration / enrichissement des données.

Les techniques d'intégration des données visent à combiner des données résidant dans différentes sources et à fournir aux utilisateurs une vue unifiée des données (Hu et al., 2014; Lenzerini, 2002). Auparavant, deux approches prévalaient, la méthode de l'entrepôt de données également appelée ETC (extraction, transformation et chargement) et la méthode de fédération des données (Hu et al., 2014; Lenzerini, 2002). Cette dernière consiste à créer une base de données virtuelle, contenant des métadonnées sur les données réelles et leur emplacement, pour interroger et agréger des données provenant de sources disparates (Hu et al., 2014). Aujourd'hui, il y a le progiciel de gestion intégré (PGI). Ce PGI permet d'intégrer toutes les fonctions et activités de l'entreprise (comptabilité, finance, production, approvisionnement, gestion des ressources humaines, etc.), appelées modules, en un seul système d'information (Baltzan, 2018). Son objectif est de donner une vision unifiée des données de l'entreprise à tous les membres du personnel tout en leur permettant de partager les données en temps réel sur cette base de données centralisée. Il permet d'avoir des données structurées en éliminant les redondances et les incohérences tout en diminuant les pertes de temps. Les fournisseurs du PGI les plus connus pour les grandes entreprises sont : SAP, Oracle et Microsoft (Baltzan, 2018). Pour les moyennes entreprises sont : Infor, Lawson, Epicor et Sage. Et pour les petites entreprises, il y a : Exact Globe, Syspro, NetSuite et Consona.

La technique de nettoyage des données fait référence au processus visant à déterminer les données inexactes, incomplètes ou déraisonnables, puis à remplacer, modifier ou supprimer ces données pour améliorer leur qualité (Gao et al., 2016; Hu et al., 2014). Il est considéré comme essentiel pour maintenir la cohérence, la mise à jour et la précision des données (Taleb et al., 2015). Un cadre général (Hu et al., 2014; Maletic et Marcus, 2000) pour le nettoyage des données comprend cinq étapes complémentaires. Ces étapes consistent à définir et déterminer les types d'erreurs; rechercher et identifier

les instances d'erreur; corriger les erreurs; documenter les instances d'erreurs et les types d'erreurs; modifier les procédures de saisie des données pour réduire les erreurs futures.

La transformation des données convertit un ensemble de valeurs de données du format de données d'un système source au format de données d'un système de destination (Gao et al., 2016; Osborne, 2002) ou une migration de données d'un système à un autre (Taleb et al., 2015). Par exemple, lors de la conversion d'un document .docx en document .pdf, des tableurs et chiffriers .xls convertis en format .csv ou des fichiers numériques (avec audio) transformés en fichier .mp4, etc.

La redondance des données est la répétition ou la superfluité des données, qui est un problème courant pour divers ensembles de données (Hu et al., 2014). Cela parce qu'elle augmente inutilement la surcharge de transmission et de stockage des données avec un espace de stockage gaspillé, une incohérence des données, une fiabilité réduite et une corruption des données. D'où la nécessité d'effectuer une réduction de ces données. Cette réduction consiste à générer des vues réduites des données sans impact sur les résultats d'analyse en utilisant des pratiques et des méthodes diverses telles que la détection de redondance (Zhang et al., 2002) et la compression des données (Hu et al., 2014; Salomon, 2004; Taleb et al., 2015). Par exemple, les techniques de compression vidéo qui sont largement utilisées pour réduire la redondance des données vidéo en utilisant des normes importantes (comme MPEG-2, MPEG-4, H.263, H.264 / AVC) (Hu et al., 2014; Symes, 2004).

L'amélioration et l'enrichissement des données consistent à définir les processus d'augmentation de la valeur d'un ensemble d'informations en combinant des données provenant de sources multiples et des collections de données telles que l'intégration et la fusion de données (Holley et al., 2014; Loshin, 2001; Taleb et al., 2015).

2.2.2.2 Transmission des données

Une fois que nous avons rassemblé les données brutes, nous devons les transférer dans une infrastructure de stockage des données, généralement dans un centre de données, pour un traitement ultérieur (Hu et al., 2014). Pour notre étude, la transmission des données représente les moyens de partage et de transfert des données entre les différents services en interne et avec les partenaires en externe. Quelques exemples d'outils et de systèmes d'information les plus utilisés dans les PME pour la transmission, le transfert et le partage des données sont présentés comme suit. Le système de gestion de base de données (SGBD) qui permet de gérer les données d'une base de données en créant, lisant, mettant à jour et supprimant les données de cette base de données, tout en contrôlant l'accès et la sécurité (Baltzan, 2018). L'entrepôt de données « permet de compiler les données provenant de bases de données relationnelles internes qui contiennent les transactions de l'entreprise de même que celles issues de jeux de données externes de fournisseurs par extraction, transformation et chargement (ETC) » (Baltzan, 2018, p. 217 - 218). Cela consiste en l'extraction des données situées dans des bases de données internes, à les transformer et les stocker dans un entrepôt de données (Baltzan, 2018). Le progiciel de gestion intégré (PGI) qui permet d'intégrer toutes les fonctions et activités de l'entreprise en un seul système d'information (Baltzan, 2018). L'informatique en nuage qui « représente un modèle d'accès au réseau qui est omniprésent, pratique et sur demande » (Baltzan, 2018, p. 183). Elle se caractérise par un ensemble partagé de ressources informatiques configurables (réseaux, de serveurs, d'entreposage, d'applications et de services) qui peut rapidement être activé et désactivé (Baltzan, 2018). Les outils collaboratifs et messageries permettent à plusieurs personnes de contribuer et de partager des données et des documents stockés tels que des documents textes, des films, des images, des calendriers, etc. (Belanger et Van Slyke, 2011). Ils contrôlent l'accès des utilisateurs et la gestion des versions, évitant à deux utilisateurs de modifier un document simultanément. Les courriels qui sont générés quotidiennement par les employés, les clients et les dirigeants (Krishnan, 2013). Ces courriels permettent le partage des données informationnelles et transactionnelles entre les membres de la chaîne logistique. Les

papiers et documents physiques qui représentent la gestion traditionnelle de l'entreprise par le partage des documents papier (comme la facture) entre les différents départements de l'entreprise en interne et entre l'entreprise et ses partenaires en externe. Ce type de gestion est toujours utilisé par les entreprises et plus spécifiquement par les PME, et ce, malgré l'utilisation croissante des nouvelles technologies de l'information et de la communication (TIC).

2.2.3 Stockage des données

L'activité stockage des données représente les composants logiciels pour le stockage et la gestion permanente de grands référentiels de données (Coleman et al., 2016; Hu et al., 2014). Traditionnellement, en tant qu'équipement auxiliaire du serveur, le périphérique de stockage des données est utilisé pour stocker, gérer, rechercher et analyser des données avec des SGBD[10] structurés (Chen et al., 2014). Avec la forte croissance des données, le périphérique de stockage des données devient de plus en plus important et de nombreuses sociétés Internet cherchent une grande capacité de stockage pour être compétitives. Un système de stockage des données peut être divisé en deux parties : l'infrastructure matérielle et la gestion des données (Hu et al., 2014). L'infrastructure matérielle est composée d'un ensemble de ressources TI partagées et organisées de manière élastique pour diverses tâches en réponse à leur demande instantanée. Cette infrastructure doit évoluer et être reconfigurée dynamiquement pour répondre à différents types d'environnements d'application. Le logiciel de gestion des données est déployé au-dessus de l'infrastructure matérielle pour maintenir des ensembles de données à grande échelle. De plus, pour analyser ou interagir avec les données stockées, les systèmes de stockage doivent fournir plusieurs fonctions d'interface, des requêtes rapides et d'autres modèles de programmation. Le cadre de gestion des données concerne la manière d'organiser les informations de manière pratique pour un traitement efficace. Par exemple,

[10] Une liste des abréviations de nature technique est fournie au début du présent document.

les bases de données NoSQL (c'est-à-dire les bases de données relationnelles non traditionnelles) deviennent de plus en plus populaires pour le stockage des données massives. Elles proposent des modes flexibles, une prise en charge d'une copie simple et facile, une API (interface applicative de programmation) simple, une cohérence à terme et une prise en charge de données volumineuses (Chen et al., 2014).

Cette étude se basera sur les dispositifs de stockage internes (p. ex. les disques durs, les entrepôts de données, etc.) et les dispositifs de stockage externes (p. ex. l'informatique en nuage, l'informatique en grille, etc.).

2.2.3.1 Dispositifs de stockage internes

Les dispositifs de stockage internes représentent les entrepôts de données (ou *data warehouse*) qui se trouvent sur les lieux physiques de l'entreprise. Les technologies typiques de stockage interne (ou archives numériques) sont présentées dans ce qui suit (Hu et al., 2014). La mémoire à accès aléatoire (RAM) est une forme de stockage des données informatiques associée à des types de mémoire volatile, qui perd ses informations lorsqu'elle est mise hors tension. Les disques magnétiques et matrices de disques, tels que les disques durs (HDD), sont les principales composantes des systèmes de stockage modernes. Un disque dur consiste en un ou plusieurs disques rigides à rotation rapide avec des têtes magnétiques disposées sur un bras d'actionnement mobile pour lire et écrire des données sur les surfaces. Contrairement à la RAM, un disque dur conserve ses données même lorsqu'il est éteint. La mémoire de classe de stockage fait référence à des supports de stockage non mécaniques, tels que la mémoire flash. En général, la mémoire flash est utilisée pour construire des disques SSD (*Solid State Drive*). Contrairement aux disques durs, les disques SSD n'ont pas de composants mécaniques, fonctionnent plus silencieusement et ont des temps d'accès plus courts et moins de latence que les disques durs. Cependant, les disques SSD restent plus chers par unité de stockage que les disques durs. Enfin, il y a la gestion traditionnelle de l'entreprise par l'archivage des documents

de gestion dans les locaux de l'entreprise pour une durée déterminée définie par la politique interne.

2.2.3.2 Dispositifs de stockage externes

Il existe plusieurs types de dispositifs de stockage externes, par exemple les systèmes de fichiers distribués et l'informatique en nuage. Les systèmes de fichiers distribués (ou *file systems*) sont la base du stockage des données massives offrant un stockage, une tolérance aux pannes, une évolutivité, une fiabilité et une disponibilité (Coleman et al., 2016; Hu et al., 2014). Les technologies de base de cette couche sont, par exemple, le système de fichiers Google GFS (ou *Google File System*) et le système de fichiers distribué Hadoop (Coleman et al., 2016). Un système de fichiers distribué permet d'accéder aux fichiers de partage de plusieurs hôtes via un réseau informatique, ce qui permet à plusieurs utilisateurs sur plusieurs machines de partager des fichiers et des ressources de stockage (Bi et Cochran, 2014). Lorsque ces systèmes distribués sont utilisés, le traitement est distribué. Les données peuvent également être réparties sur deux ou plusieurs nœuds, comme sur des serveurs de fichiers, des serveurs SQL[11], des postes de travail, etc. (Stamper, 1988). Les données distribuées et les transactions distribuées peuvent avoir un impact significatif sur l'utilisation des ressources du réseau. La mise en œuvre du traitement parallèle permet à différentes méthodes d'améliorer les performances globales des processus d'extraction, de transformation et de chargement (ETC) lorsqu'il s'agit de gros volumes de données (Bi et Cochran, 2014; Stonebraker et al., 2010). L'informatique en nuage est définie comme un modèle d'accès réseau à la demande à un ensemble partagé de ressources informatiques configurables (Coleman et al., 2016; Mell et Grance, 2011). Son avantage est qu'elle permet de stocker de grands volumes de données (Karoui et al., 2014). Les solutions proposées peuvent être classées en trois grandes catégories, selon qu'elles proposent des services logiciels, des plateformes ou des

[11] Langage SQL : ou *Structured Query Language*, (SQL) est un « Langage informatique dédié à l'interrogation et à la gestion des données dans un système de gestion de base de données » (Baltzan et al., 2018, p. 205).

infrastructures de base spécifiques (Coleman et al., 2016; Zhang et al., 2010). Ces plateformes représentent un moyen économique de créer l'infrastructure matérielle des données grâce au marché actuel qui propose plusieurs solutions à cet égard comme Amazon Web Services, Microsoft Azure et Google Cloud DataLab (Coleman et al., 2016). Cela constitue une solution abordable et économique pour les PME grâce à un modèle flexible leur garantissant l'indépendance de la plateforme, la portabilité et la flexibilité, ainsi que la possibilité d'arrêt à n'importe quel moment pour ainsi réduire les coûts fixes de détention d'infrastructures de stockage (Coleman et al., 2016).

Il existe aussi la gestion traditionnelle de l'entreprise par l'archivage des documents de gestion dans les locaux de l'entreprise pour une durée déterminée définie par la politique intérieure de l'entreprise. Dans ce cas, les archives papier sont autorisées de quitter les locaux de l'entreprise en cas de besoin. On peut dire la même chose pour les dispositifs de stockage internes où il y a différentes technologies typiques de stockage des données, et ce, telles que présentées précédemment par Hu et al. (2014), mais dans ce cas, les archives numériques peuvent quitter les lieux physiques de l'entreprise.

2.2.4 Analyse des données

Une fois les informations mises à disposition, des méthodes d'analyse des données sont utilisées pour rechercher les informations utiles (Bi et Cochran, 2014). L'analyse des données est un processus d'inspection, de nettoyage, de transformation et de modélisation des données dans le but de découvrir des informations utiles, de suggérer des conclusions et d'appuyer la prise de décision (Gao et al., 2016; Hu et al., 2014). Le but de l'analyse des données est d'extraire autant d'informations pertinentes que possible pour le sujet considéré (Hu et al., 2014). La nature du sujet et les objectifs peuvent varier considérablement. Par exemple, pour extrapoler et interpréter les données et ainsi déterminer comment les utiliser. Afin de vérifier la légitimité des données, ou donner des

conseils et aider à la prise de décision. Pour diagnostiquer et déduire les raisons d'un problème (ou un défaut). Enfin, pour le but de prédire ce qui se passera dans le futur.

L'analyse des données utilise des méthodes ou des outils analytiques pour inspecter, transformer et modéliser les données afin d'en extraire de la valeur (Hu et al., 2014). Il existe six types d'applications des données massives, organisées par type de données, qui sont : l'analyse des données structurées, l'analyse de texte, l'analyse Web, l'analyse multimédia, l'analyse de réseau et l'analyse mobile (Hu et al., 2014). Parmi ces types d'analyses, deux correspondent le plus au contexte des PME qui sont : l'analyse des données structurées et l'analyse Web. Dans ce dernier cas, l'exploration des données à ce stade est un processus d'extraction d'informations et de connaissances cachées, inconnues mais potentiellement utiles à partir des données massives, incomplètes, bruyantes, floues et aléatoires (Chen et al., 2014).

Cette étude se basera sur les applications et outils d'analyse appartenant ou qui sont développés par l'entreprise (p. ex. les systèmes de traitement de facturation ou de transactions, les PGI, etc.) et les applications et outils d'analyse développés par des tiers (p. ex. par l'intermédiaire de l'infrastructure Web ou des logiciels-service, etc.).

2.2.4.1 Applications et outils d'analyse appartenant ou développés par l'entreprise

Les outils d'analyse les plus utilisés par les PME constituent les fonctions de base d'analyse des données sont ceux qui permettent une analyse des données structurées. Une grande quantité de données structurées est générée dans les domaines des affaires et de la recherche scientifique. La gestion de ces données structurées repose sur la maturité du système de gestion de bases de données (SGBD), l'entreposage des données, le traitement analytique en ligne (OLAP) et la gestion des performances commerciales (BPM) (Chaudhuri et al., 2011; Hu et al., 2014). Comme défini précédemment, le SGBD est un système permettant la création, la lecture, la mise à jour et la suppression des données d'une base de données structurée, ainsi que le contrôle d'accès et la sécurité

(Baltzan, 2018). Les SGBD les plus connu sont : « MySQL, Microsoft Access, SQL Server, FileMaker, Oracle et FoxPro » (Baltzan, 2018, p. 205). La gestion et l'entreposage des données sont considérés comme la base de l'analyse des données structurées. La conception des comptoirs des données (ou *data marts*) et d'outils d'extraction, de transformation et de chargement (ETC) sont essentiels pour convertir et intégrer des données spécifiques à l'entreprise (Hsinchun et al., 2012). Les requêtes de base de données tel que le langage SQL (ou *Structured Query Langage*) qui est « un langage informatique dédié à l'interrogation et la gestion des données dans un système de gestion de base de données » (Baltzan, 2018, p. 205), le traitement analytique en ligne (OLAP) et les outils de création de rapports basés sur des graphiques intuitifs, mais simples sont utilisés pour explorer les caractéristiques importantes des données (Hsinchun et al., 2012). L'OLAP représente l'analyse de l'information interne de l'entreprise provenant des systèmes de traitement de transactions d'un côté et celles provenant de l'extérieur comme les partenaires d'un autre côté ayant comme objectif la prise de décisions stratégiques de l'entreprise (Baltzan, 2018). Le progiciel de gestion intégré (PGI) permet aussi la gestion des données structurées par l'entreprise. Il est divisé en deux composantes : les composantes de base et les composantes élargies (Baltzan, 2018). Les composantes de base du PGI représentent les activités centrées sur les activités internes de l'entreprise qui sont : la composante en comptabilité et finance, la gestion de la production et celle des ressources humaines. Les composantes élargies du PGI représentent les activités externes de l'entreprise qui ne sont pas couvertes dans les composantes de base et qui nécessitent souvent l'interaction avec les partenaires d'affaires via l'utilisation d'Internet. Les plus courantes sont : l'intelligence d'affaires, la gestion de la relation client, la gestion de la chaîne d'approvisionnement et les affaires électroniques. De plus, il y a le système de traitement des transactions (STT) qui est un système de suivi des opérations utilisé au niveau opérationnel comme le système d'enregistrement des commandes. Par ailleurs, il y a les systèmes interactifs d'aide à la décision (SIAD) qui « modélisent l'information au moyen du traitement analytique en ligne, ce qui contribue à évaluer plusieurs possibilités et à choisir parmi ces dernières » (Baltzan, 2018, p. 46 - 47). Enfin, la gestion des

performances commerciales (BPM) à l'aide de cartes de performance et de tableaux de bord permettant d'analyser et de visualiser une variété de mesures de performance (Hsinchun et al., 2012).

2.2.4.2 Applications et outils d'analyse développés par des tiers

Les outils d'analyse développés par des tiers sont souvent des outils d'analyse et de calcul infonuagiques via le Web permettant de bénéficier des plateformes et logiciels en ligne qui sont facturés à l'utilisation. Permettant ainsi d'économiser sur les coûts d'investissement pour les PME. La croissance explosive des pages Web au cours des dernières années avec les outils du Web 2.0 a engendré un nouveau domaine actif qui est l'analyse Web qui vise à récupérer, extraire et évaluer automatiquement les informations pour la découverte de connaissances à partir de documents et de services Web (Hu et al., 2014).

Une composante émergente majeure de la recherche en analyse Web est le développement de plateformes et de services de l'informatique en nuage qui comprennent des applications, des logiciels système et du matériel fournis sous forme de services sur Internet (Hsinchun et al., 2012). Basée sur l'architecture orientée services (SOA), la virtualisation des serveurs et l'informatique utilitaire, l'informatique en nuage peut être proposée sous forme de logiciel en tant que service (ou *Software as a Service* - SaaS), d'infrastructure en tant que service (ou *Infrastructure as a Service* - IaaS) ou de plateforme en tant que service (ou *Platform as a Service* - PaaS). Le logiciel-service est un « service offrant des applications payables à l'utilisation par l'entremise de l'informatique en nuage » (Baltzan, 2018, p. 186). Par exemple, l'utilisation de Salesforce qui propose une application d'automatisation de la force de vente. L'infrastructure-service est un « service offrant du matériel de mise en réseau, notamment des serveurs, des ressources réseau et de l'espace de stockage, par l'entremise de l'informatique en nuage, le tout étant payable à l'utilisation » (Baltzan, 2018, p. 185). Par exemple, l'utilisation d'*Amazon Elastic Compute Cloud* (EC2) qui est une interface Web où les clients peuvent charger leurs

applications et gérer eux-mêmes leur environnement d'exploitation en créant, exécutant et interrompant des services selon leurs besoins. Il existe d'autres exemples pour les fournisseurs qui proposent le stockage en ligne des données tels que le service de stockage simple (S3) et le Google Bigtable (Hsinchun et al., 2012). Enfin, la plateforme-service est un « service proposant le déploiement de systèmes en entier qui comprend le matériel informatique, les ressources réseau ainsi que les applications. Le tout est offert par l'entremise de l'informatique en nuage et est également payable à l'utilisation » (Baltzan, 2018, p. 186). Par exemple, le moteur d'application de Google qui « conçoit et déploie des applications Web à l'intention des entreprises » (Baltzan, 2018, p. 187).

Le texte est l'une des formes les plus courantes d'informations stockées et comprend la communication par courrier électronique, les documents d'entreprise, les pages Web et le contenu des médias sociaux. L'analyse de texte, connue sous le nom de *text mining* (ou extraction de texte), fait référence au processus d'extraction d'informations et de connaissances utiles à partir d'un texte non structuré (Hu et al., 2014).

L'augmentation des données multimédias (l'image, l'audio et la vidéo) à un rythme phénoménal a engendré un nouveau type d'analyse qui est l'analyse de contenu multimédia. Elle fait référence à l'extraction de connaissances intéressantes et à la compréhension de la sémantique capturée dans les données multimédias. Ces données qui sont plus riches en informations que les simples données structurées et les données textuelles nécessite de passer par un processus d'extraction d'informations (Hu et al., 2014).

Le contenu généré par les utilisateurs explose sur les réseaux sociaux à cause de l'évolution de la technologie Web 2.0 où le terme médias sociaux est utilisé pour nommer ce contenu généré par l'utilisateur. Par exemple les blogues, les microblogues, le partage de photos et de vidéos, le marketing numérique, les sites de réseautage social, les actualités sociales et les wikis. En règle générale, les réseaux sociaux contiennent une

énorme quantité de données de liaison et de contenu. Les données de liaison sont essentiellement la structure graphique, représentant les communications entre les entités. Les données de contenu contiennent du texte, des données multimédias (images, vidéos), des emplacements et des commentaires dans les réseaux. Presque tous les sujets de recherche sur l'analyse de données structurées, l'analyse de texte et l'analyse multimédia peuvent être traduits en analyse des médias sociaux (Hu et al., 2014). Néanmoins, l'analyse des médias sociaux est confrontée à certains défis comme une augmentation constante des données à analyser dans un délai raisonnable contenant des données bruyantes à nettoyer. Par ailleurs, les réseaux sociaux sont dynamiques, en constante évolution et mis à jour rapidement. Les applications de l'analyse de texte dans les réseaux sociaux incluent les recherches de mots clés, les classifications, le regroupement et l'apprentissage par transfert dans des réseaux hétérogènes (Hu et al., 2014).

2.2.5 Diffusion / visualisation de l'information

De nos jours, les analystes doivent présenter les résultats dans des visualisations puissantes facilitant l'interprétation et la prise en charge de la collaboration des utilisateurs. Cela en proposant des visualisations se basant sur des sources interactives permettant aux utilisateurs de définir les critères et les scénarios hypothétiques telles que des plages de dates, des emplacements géographiques ou des requêtes statistiques. De plus, en quelques clics, l'utilisateur devrait être en mesure d'approfondir chaque élément de données et de comprendre sa provenance. Cela est une caractéristique clé pour comprendre les données voire de faire une simulation en proposant d'autres scénarios possibles afin d'avoir pour chaque scénario le résultat à obtenir avant la prise de la décision finale (Ohlhorst, 2012).

Puisque la visualisation est la dernière étape du cycle de vie des données massives, ce processus comportera plus d'informations structurées que des données, mais

dans cette partie il n'y aura pas de différence entre les données et l'information. L'activité *diffusion / visualisation des données (ou de l'information)* permet de « produire des représentations graphiques des tendances et des relations complexes dans de grandes quantités de données » (Baltzan, 2018, p. 48). Le processus de visualisation consiste à visualiser et à valider les données issues de la chaîne de valeur des données massives afin de soutenir la formulation des décisions, de rendre compte des mises à jour continues sur l'état de ces données massives (Serhani et al., 2016). Les données peuvent être présentées à l'aide de différentes vues, un résumé des résultats de la surveillance, des graphiques, des schémas de lecture, et même un rapport sur les écarts de mesures, puis générer des actions préventives automatiques. Donc, la visualisation consiste à présenter les résultats analytiques aux décideurs sous la forme d'un rapport statique ou d'une application interactive (Miller et Mork, 2013). Elle prend en charge l'exploration et le raffinement des résultats pour objectif de fournir aux principales parties prenantes des informations significatives dans un format qu'elles peuvent facilement utiliser pour prendre des décisions critiques.

Cette étude se basera sur les données destinées au personnel de l'entreprise (p. ex. le système d'information pour dirigeants (SID), intranet, etc.) et les données destinées à toute personne n'appartenant pas au personnel de l'entreprise (p. ex. extranet, affaires électroniques, etc.).

2.2.5.1 Données destinées au personnel de l'entreprise

Les outils de diffusion / visualisation de l'information en interne les plus utilisés en PME sont cités dans ce qui suit. Pour les dirigeants, il y a le système d'information pour dirigeants (SID) qui est « un système d'aide à la décision spécialisé conçu pour les cadres supérieurs de l'organisation » (Baltzan, 2018, p. 48) afin de prendre des décisions non structurées nécessitant d'avoir une vue d'ensemble. Les SID utilisent de la visualisation à travers les tableaux de bord numériques comportant des indicateurs de performance. Les systèmes interactifs d'aide à la décision (SIAD) sont des systèmes

informatiques destinés aux cadres supérieurs (Baltzan, 2018). Ils leur permettent de manipuler de grandes quantités de données internes et externes et les modéliser pour leur proposer plusieurs possibilités à évaluer et leur laisser le choix de la ou les solutions optimales. Il y a aussi le tableau de bord numérique qui est un outil permettant de suivre en temps réel les indicateurs de performance et les facteurs clés de succès (Baltzan, 2018). Il permet de compiler l'information issue de différentes sources afin de prendre des décisions et résoudre les problèmes le plus tôt possible, et ce, d'une manière quotidienne. Le progiciel de gestion intégré (PGI) permet également la diffusion des données en interne à travers l'intégration de toutes les fonctions et activités de l'entreprise en un seul système d'information pour donner une vision unifiée des données de l'entreprise à tous les membres du personnel (Baltzan, 2018). Enfin, il y a l'intranet qui est une « partie internalisée d'Internet, à l'abri de tout accès de l'extérieur, qui permet à une organisation d'offrir uniquement à son personnel un accès à l'information et aux logiciels d'application » (Baltzan, 2018, p. 244). Il peut contenir tout type d'information lié à l'entreprise et son fonctionnement en interne, qui est destinée à son personnel, grâce à des publications sur le Web comme les bulletins de paie, les catalogues de produits, etc. Il peut aussi être utilisé comme un moyen de communication en interne à travers l'envoi et la réception des messages et courriels ainsi que les documents numériques entre les membres de l'organisation. De plus, il y a une possibilité de l'utiliser pour l'exploitation et la gestion de l'entreprise en concevant des applications personnalisées pour effectuer les différentes activités de l'entreprise, telles que le traitement des commandes, la gestion des stocks et la gestion des ventes. Cela à travers l'utilisation du navigateur Web qui peut être utilisé de n'importe quel emplacement du réseau.

2.2.5.2 Données destinées aux personnes n'appartenant pas au personnel de l'entreprise

Les outils de diffusion / visualisation de l'information en externe les plus utilisés en PME sont cités dans ce qui suit. L'extranet est un « service d'intranet qui est mis à la disposition d'alliés stratégiques (comme des clients, des fournisseurs ou des partenaires) » (Baltzan, 2018, p. 244). Donc, il propose les mêmes fonctionnalités d'intranet pour les

partenaires d'affaires. De plus, la diffusion de l'information peut se faire via les affaires électroniques dont l'expression est dérivée du commerce en ligne qui représente « l'achat des biens et services par Internet » (Baltzan, 2018, p. 86). Les affaires électroniques vont au-delà de la vente en ligne via Internet vers l'échange de l'information en ligne avec les collaborateurs et partenaires d'affaires. Il existe différents types de modèles d'affaires des affaires électroniques: le commerce électronique interentreprises (B2B), le commerce électronique entreprise-consommateur (B2C), le commerce électronique consommateur-entreprise (C2B), ou le commerce électronique interconsommateurs (C2C). Dans le B2B qui « s'applique aux activités des entreprises effectuant des achats et des ventes entre elles par Internet » (p. 94), le client a l'accès en ligne aux données liées à la transaction (comme les dates d'expédition prévues et l'état des envois) dans son espace client qui sont fournies par le vendeur (Baltzan, 2018).

Enfin, dans le contexte de cette étude, nous considérons évidemment que les PME utilisent toujours la diffusion traditionnelle de l'information en interne, entre les différents départements de l'entreprise, et externe, entre l'entreprise et ses partenaires, via les rapports et documents papier.

2.3 QUALITÉ DE L'INFORMATION

Concernant la qualité de l'information, la littérature utilise les termes Qualité de l'information (QI) et Qualité des données (QD) de manière souvent interchangeable (Rafique et al., 2012). Pourtant, il existe une différence de signification entre les données et les informations. Selon ISO/IEC 25012:2008(E)[12], les données sont une représentation réinterprétable d'informations d'une manière formalisée adaptée à la communication, à l'interprétation ou au traitement; tandis que l'information est une connaissance concernant des objets, tels que des faits, des événements, des choses, des processus ou des idées, y

[12] ISO/IEC 25012:2008(E) : Exigences de qualité et évaluation des produits logiciels (SQuaRE) - Modèle de qualité des données.

compris des concepts, qui dans certains contextes ont une signification particulière (Rafique et al., 2012). Lorsque les données sont mises dans un contexte et combinées dans une structure, des informations émergent. Par conséquent, lorsque nous discutons de la qualité de l'information (QI), nous devons tenir compte de sa structure et de son contexte. Sinon, elle sera considérée comme une donnée.

Définir la qualité des données et des informations n'est pas une chose facile parce qu'elle dépend de plusieurs facteurs, ainsi que du fait qu'elle est sensible aux facteurs comme le contexte, le domaine des données, la zone ou les domaines dans lesquels ces données sont utilisées. De plus, la qualité des données est parfois comprise différemment dans le milieu universitaire comparativement à l'industrie (Oliveira et al., 2005; Taleb et al., 2018). La qualité des données (QD) est un concept bien connu dans la communauté de recherche en gestion de bases de données et a été un domaine actif de recherche pendant de nombreuses années (Chiang et Miller, 2008; Yeh et Puri, 2010). En général, il existe un consensus sur le fait que la qualité des données dépend toujours de la qualité de la source des données (Maier et al., 2013). C'est pour cela qu'il est important d'associer la qualité aux données dès leur création, car en transitant par plusieurs étapes du processus, la qualité de ces données sera affectée positivement ou négativement (Fürber et Hepp, 2011). Donc, si les données en intrants sont de mauvaise qualité, cela engendrera des données sortantes qui manquent également de fiabilité. Selon Sidi et al. (2012) et Taleb et al. (2018), la qualité des données est définie comme la réponse aux exigences des utilisateurs ou la satisfaction des besoins des utilisateurs en matière d'information. Donc, des problèmes de qualité des données surviennent lorsque les exigences de qualité ne sont pas satisfaites sur les valeurs des données (Fürber et Hepp, 2011).

Dans tout système de qualité des données, celles-ci sont soumises à un audit, à un profilage et à l'application des règles de qualité dans le but de maintenir et/ou d'améliorer leur qualité (Taleb et al., 2015). Mais, l'application des concepts de qualité aux données massives rencontre des problèmes en termes du temps et des coûts du

prétraitement de ces données. D'autant plus que ces techniques ont été développées en contexte des données structurées alors que les données massives contiennent aussi des données non structurées, par exemple celles provenant des réseaux sociaux où il n'existe aucune référence de qualité. D'où la nécessité d'évaluer la qualité des données dès la phase de démarrage (*génération des données*) et tout au long du cycle de vie de ces données. Dans le cadre des données massives, toute application de qualité des données doit être sélectionnée en fonction de l'origine, du domaine, de la nature, du format et du type de données sur lesquelles elle est appliquée. Une bonne gestion de ces systèmes de qualité des données est essentielle pour résoudre les nombreux problèmes qui se posent lorsqu'il s'agit de grands ensembles de données.

Il existe des référentiels internationaux, qui permettent l'évaluation de la qualité des données. Le référentiel ISO/IEC 25012:2008 définit un modèle général de la qualité des données pour les données conservées dans un format structuré au sein d'un système informatique et vise à soutenir la mise en œuvre des processus du cycle de vie du système (Rafique et al., 2012). Il rassemble les principales caractéristiques de la qualité des données souhaitables pour tout ensemble de données en catégorisant les attributs de la qualité des données en quinze caractéristiques divisées en deux catégories : inhérent (intrinsèque) et dépendant du système (contextuelle) (Merino et al., 2016; Serhani et al., 2016; Taleb et al., 2015). La qualité inhérente des données fait référence au degré auquel les caractéristiques de la qualité des données ont un potentiel intrinsèque pour satisfaire les besoins explicites et implicites quand les données sont utilisées sous certaines conditions (Rafique et al., 2012). Les caractéristiques liées à la qualité inhérente des données sont : l'exactitude, la complétude, la cohérence, la crédibilité, l'actualité, l'accessibilité, la conformité, la confidentialité, l'efficacité, la précision, la traçabilité et la compréhensibilité. La qualité des données dépendante du système fait référence au degré auquel la qualité des données est atteinte et préservée au sein d'un système informatique quand les données sont utilisées sous des conditions spécifiques. De ce point de vue, la qualité des données dépend du domaine technologique dans lequel les données

sont utilisées. Cela est atteint par les capacités des composants du système informatique. Les caractéristiques liées à la qualité des données dépendante du système sont : l'accessibilité, la conformité, la confidentialité, l'efficacité, la précision, la traçabilité, la compréhensibilité, la disponibilité, la portabilité et la récupérabilité. Le Tableau 1 ci-dessous résume toutes ces qualités qu'il est possible d'attribuer aux données.

Tableau 1

ISO 25012 Qualité des données[13]

Caractéristique	Inhérent	Dépendant du système
Exactitude	x	
Complétude	x	
Cohérence	x	
Crédibilité	x	
Actualité	x	
Accessibilité	x	x
Conformité	x	x
Confidentialité	x	x
Efficacité	x	x
Précision	x	x
Traçabilité	x	x
Compréhensibilité	x	x
Disponibilité		x
Portabilité		x
Récupérabilité		x

Suivant le Tableau 1, les caractéristiques de la qualité des données qui sont à la fois inhérentes et dépendantes du système sont : l'accessibilité, la conformité, la confidentialité, l'efficacité, la précision, la traçabilité et la compréhensibilité. Les caractéristiques d'évaluation de la qualité des données sont au nombre de quinze, mais dans notre étude nous allons réduire ce nombre à huit caractéristiques principales, extraites du Tableau 1, qui sont : l'exactitude, la crédibilité, la compréhensibilité, la cohérence, la conformité, la disponibilité, la complétude et l'actualité. Nous allons ajouter une autre caractéristique très importante pour l'évaluation de la qualité de l'information

[13] Tiré de Merino et al. (2016, p. 3) et Rafique et al. (2012, p. 3).

(ou des données) qui est : « l'unicité » pour indiquer qu'il n'y a pas de redondance dans les données (Baltzan, 2018). Finalement, nous aurons neuf caractéristiques d'évaluation de la qualité des données.

L'objectif de notre recherche est de trouver les configurations d'activités et d'outils de traitement des données permettant d'assurer une haute qualité de l'information et celles qui permettent une qualité moyenne, de même que celles menant à une faible, voire à une absence de la qualité de l'information. Ces neuf caractéristiques de l'évaluation de la qualité de l'information sont : l'exactitude, la crédibilité, la compréhensibilité, la cohérence, la conformité, la disponibilité, la complétude, l'actualité et l'unicité. Le Tableau 2 résume les définitions que nous associons à chacune de ces caractéristiques.

Tableau 2

Résumé des définitions des caractéristiques de qualité de l'information

Caractéristique	Définition	Sources
Exactitude	Mesure si les données ont été enregistrées correctement et reflètent des valeurs réalistes.	Taleb et al. (2015); Gao et al. (2016)
Crédibilité	Mesure si les informations sont réputées objectives et fiables.	Rafique et al. (2012)
Compréhensibilité	Mesure si l'information est facile à comprendre par l'utilisateur.	Rafique et al. (2012)
Cohérence	Mesure si l'information fournie est présentée dans des formats qui coïncident entre eux.	Taleb et al. (2015)
Conformité	Mesure si l'information est applicable et utile pour la tâche à accomplir	Belanger et Van Slyke (2011)
Disponibilité	Mesure si l'information est récupérable, disponible et prête à être utilisée au bon moment par l'utilisateur	Rafique et al. (2012)
Complétude	Mesure si toutes les données pertinentes sont enregistrées sans entrées ou valeurs manquantes	Taleb et al. (2015); Rafique et al. (2012)
Actualité	Mesure si les données sont à jour	Taleb et al. (2015); Merino et al. (2016)
Unicité	Mesure s'il n'y a pas de redondance inutile dans l'information	Baltzan (2018)

L'exactitude des données mesure si les données ont été enregistrées correctement et reflètent des valeurs réalistes (Taleb et al., 2015). Pour Gao et al. (2016), elle fait

référence à la proximité des résultats des observations avec les vraies valeurs ou les valeurs acceptées comme vraies. Donc, pour considérer l'information fournie comme exacte, elle doit être correcte, sans erreurs et fiable. La crédibilité est la mesure dans laquelle les informations sont réputées objectives (impartiales) et fiables (vraies et crédibles) (Rafique et al., 2012). Dans son article, il a considéré qu'elle est une sous-caractéristique de l'exactitude, mais dans notre étude nous allons la considérer comme une des neuf caractéristiques choisies. Donc, pour que l'information fournie soit crédible, elle doit être objective (non biaisée) et fiable (vraie, selon la connaissance de l'utilisateur). La compréhensibilité est la mesure dans laquelle les informations ont des attributs qui leur permettent d'être lues et interprétées par les utilisateurs, et sont exprimées dans des langues, symboles et unités appropriés dans un contexte d'utilisation spécifique (Rafique et al., 2012). Donc, pour que l'information soit compréhensible elle doit être facile à comprendre par l'utilisateur. La cohérence est la mesure dans laquelle l'information est présentée dans des formats qui coïncident entre eux (Taleb et al., 2015). Une information conforme doit être applicable et utile pour la tâche à accomplir (Belanger et Van Slyke, 2011). La disponibilité est le degré auquel les informations ont des attributs qui leur permettent d'être récupérées par des utilisateurs et/ou des applications autorisées dans un contexte d'utilisation spécifique (Rafique et al., 2012). Donc, pour que l'information fournie soit disponible elle doit être récupérable, disponible et prête à être utilisée au bon moment par l'utilisateur. La complétude mesure si toutes les données pertinentes sont enregistrées sans entrées ou valeurs manquantes (Taleb et al., 2015). Selon Rafique et al. (2012) la complétude est la mesure dans laquelle les informations fournies par une application Web sont d'une ampleur, d'une profondeur et d'une portée suffisantes pour la tâche à accomplir. Donc, pour que l'information fournie soit complète elle doit contenir toutes les valeurs nécessaires pour répondre aux besoins de l'utilisateur ou pour remplir la tâche à accomplir. L'actualité mesure si les données sont à jour (Taleb et al., 2015). Selon Merino et al. (2016), les données doivent être correctement mises à jour pour la tâche à accomplir, afin qu'elles aient un âge convenable pour l'analyse, du fait que dans certains cas, la fusion des données ayant différents niveaux d'actualité peut ne pas

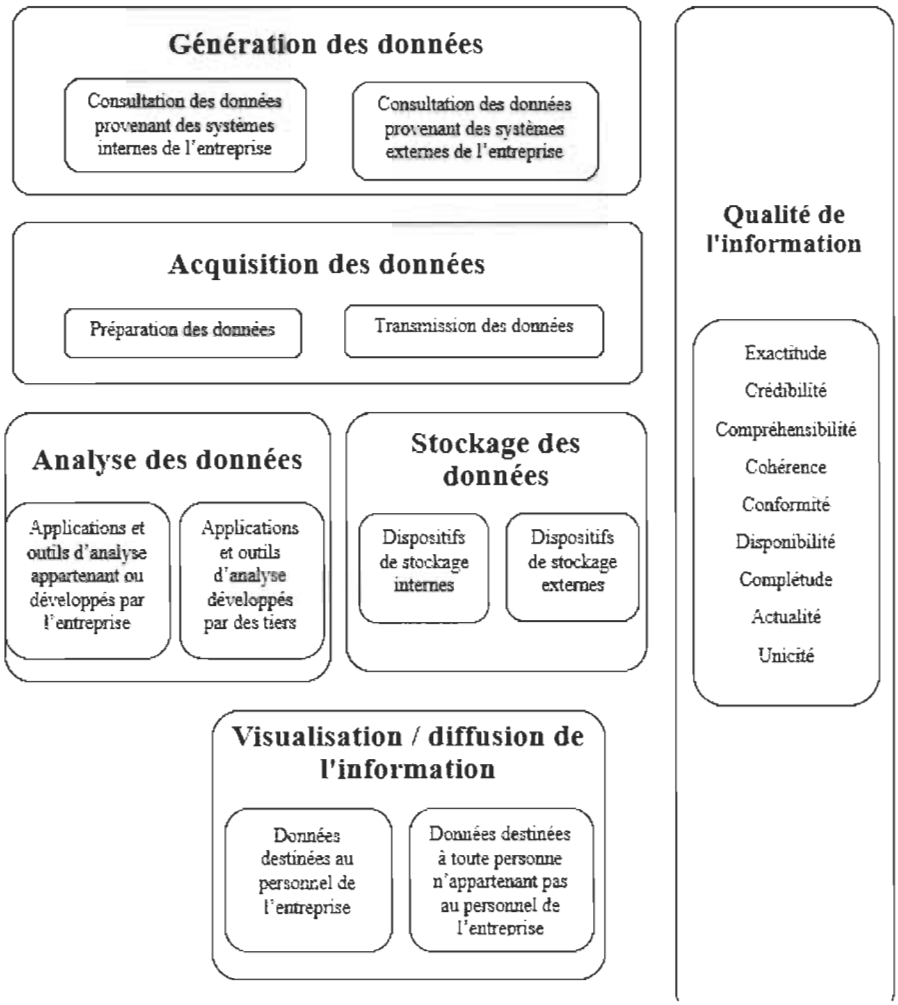
conduire à une analyse solide. Donc, pour que l'information fournie soit actuelle elle doit être est suffisamment à jour pour le travail à accomplir. L'unicité c'est quand chaque élément (transaction, entité ou événement) est présenté une seule fois sans redondance de données (Baltzan, 2018). Donc, pour que l'information fournie soit unique, il ne doit y avoir de redondance inutile entre les différentes sources de l'information.

2.4 CADRE CONCEPTUEL ET MODÈLE CONFIGURATIONNEL

Résumant les éléments précédemment présentés, la Figure 4 présente le modèle conceptuel de transformation des données en information dans les PME à l'ère des données massives. Ce modèle inclut les cinq activités de transformation des données massives en information qui sont : *la génération, l'acquisition*, *le stockage*, *l'analyse des données* et *la visualisation / diffusion de l'information*, auxquelles est ajouté le concept de *la qualité de l'information* à travers neuf caractéristiques qui sont : l'exactitude, la crédibilité, la compréhensibilité, la cohérence, la conformité, la disponibilité, la complétude, l'actualité et l'unicité.

Figure 4

Cadre conceptuel du traitement de l'information dans les PME à l'ère des données massives[14]



La première activité du traitement de l'information, tel que nous le concevons, est la *génération des données*. L'objectif de celle-ci est de dresser un inventaire de toutes les sources de données auxquelles l'entreprise peut accéder ou consulter. La *génération des données* comporte deux sous-activités sur les types des sources de données (internes ou externes). La première est la consultation des données provenant des systèmes internes

[14] Adapté de Coleman et al. (2016), Gao et al. (2016), Hu et al. (2014), Merino et al. (2016), Rafique et al. (2012), Taleb et al. (2015) et Taleb et al. (2018).

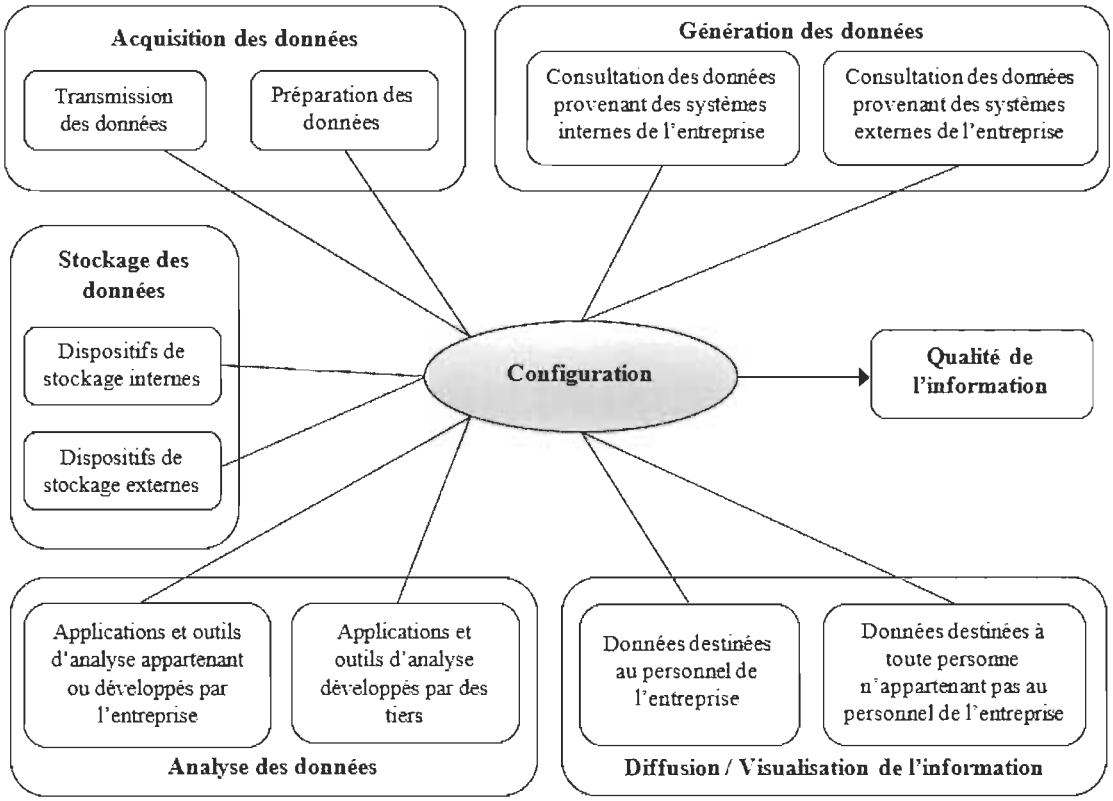
de l'entreprise qui représente les sources internes de consultation des données, informatisées ou non informatisées. La deuxième sous-activité est la consultation des données provenant des systèmes externes de l'entreprise qui représente les sources de données provenant des systèmes externes, informatisés ou non informatisés, et produites par une tierce personne n'appartenant pas au personnel de l'entreprise. La deuxième activité du traitement de l'information est *l'acquisition des données* qui est composée de deux sous-activités, soit la préparation des données et la transmission des données. La troisième activité est le *stockage des données*. Cette activité est composée de deux sous-activités concernant les types d'outils de stockage des données. La première sous-activité concerne l'utilisation des dispositifs de stockage internes qui représentent les outils de stockage appartenant à l'entreprise et toutes les données stockées en interne et ne quittant pas les lieux physiques de l'entreprise. La deuxième sous-activité concerne les dispositifs de stockage externes qui représentent les outils ou les espaces de stockage n'appartenant pas à l'entreprise et qui peuvent être loués par l'entreprise pour l'utilisation pendant une durée déterminée ou tout autre dispositif permettant de stocker les données en dehors des locaux de l'entreprise. La quatrième activité du traitement de l'information est l'*analyse des données*. Cette activité est composée de deux sous-activités concernant les moyens utilisés pour procéder à l'analyse des données. La première sous-activité concerne les applications et outils d'analyse appartenant ou développés par l'entreprise qui représentent les outils d'analyse appartenant ou développés par l'entreprise. La deuxième sous-activité concerne les applications et outils d'analyse développés par des tiers qui représentent les outils d'analyse n'appartenant pas à l'entreprise (p. ex. outils développés par des agences spécialisées en marketing numérique). Ces moyens peuvent aussi être offerts par l'intermédiaire de logiciels-services dont l'utilisation est payée en fonction de l'usage, pendant une durée déterminée (p. ex. Salesforce). La cinquième activité est la *diffusion / visualisation de l'information*. Cette activité est composée de deux sous-activités portant sur les types des données. La première sous-activité concerne les données destinées au personnel de l'entreprise qui représentent les données destinées à une diffusion interne pour le personnel de l'entreprise. La deuxième sous-activité concerne

les données destinées à toute personne n'appartenant pas au personnel de l'entreprise qui représentent toutes les données destinées à une diffusion externe pour toute personne agissant en dehors de l'entreprise, qui ne fait pas partie du personnel salarié, ou de l'actionnariat. Enfin, le résultat à expliquer de notre cadre conceptuel est la *qualité de l'information* qui représente la variable dépendante ou la variable expliquée. Ce concept est composé de neuf caractéristiques de l'évaluation de la qualité de l'information sélectionnées parmi la liste présentée précédemment et qui sont : l'exactitude, la crédibilité, la compréhensibilité, la cohérence, la conformité, la disponibilité, la complétude, la disponibilité et l'unicité (Baltzan, 2018; Merino et al., 2016; Rafique et al., 2012; Taleb et al., 2015).

Fondé sur le cadre conceptuel préalablement présenté, la Figure 5 ci-dessous présente le modèle configurationnel de transformation des données en information dans les PME à l'ère des données massives. De manière à soutenir l'investigation du thème et des questions de recherche formulées, ce modèle représente ce que nous retenons concernant les activités et les outils de traitement des données *(génération, acquisition, stockage, analyse des données, diffusion / visualisation de l'information)* qui constituent les variables indépendantes, incluant leurs sous-activités (ou dimensions), dans le but d'expliquer la variable dépendante qui est le niveau de *la qualité de l'information* à atteindre (haute, moyenne, absence).

Figure 5

Modèle configurationnel du traitement de l'information dans les PME à l'ère des données massives[15]



[15] Adapté de Coleman et al. (2016), Gao et al. (2016), Hu et al. (2014), Merino et al. (2016), Rafique et al. (2012), Taleb et al. (2015) et Taleb et al. (2018).

CHAPITRE 3 – MÉTHODOLOGIE

Notre étude cherche à décrire et comprendre le phénomène du traitement de l'information au sein des PME à l'ère des données massives à travers l'utilisation d'un devis quantitatif[16]. Pour ce faire, la recherche mobilise une approche configurationnelle fondée sur l'analyse qualitative comparée par les ensembles flous (aussi appelée fsQCA pour *fuzzy-set Qualitative Comparative Analysis*). Cette approche est aussi adaptée au constat que les connaissances existantes sur le phénomène du traitement de l'information au sein des PME à l'ère des données massives sont rares (Coleman et al., 2016).

Un certificat d'éthique du comité d'éthique de la recherche (CER) de l'Université du Québec à Trois-Rivières nous permettant d'effectuer la recherche avec des êtres humains a été émis. Ce certificat porte le numéro CER-19-263-07.16 et sa période de validité s'étend du 15 janvier 2021 au 15 janvier 2022 (voir ANNEXE A).

3.1 CONSTITUTION DE L'ÉCHANTILLON

Selon Fortin et Gagnon (2016, p. 275) « le recrutement des participants fait référence au processus de sélection des participants ». Pour la constitution de notre échantillon, nous avons utilisé la base de données DataDb.ca dans laquelle nous avons choisi les PME québécoises respectant les critères d'inclusion. C'est-à-dire celles qui ont un effectif supérieur à 10 et inférieur ou égal à 500 employés et appartenant aux secteurs d'activité du manufacturier, transport / logistique, commerce de gros, commerce de détail, services aux entreprises (industriels/scientifiques), ainsi que finance et assurances. Ensuite, nous avons fait l'extraction de cette liste sur un tableur (Excel) selon l'effectif de la PME et le code de Classification Type des Industries (CTI) qui constitue notre liste

[16] Le devis choisi pour répondre à notre question de recherche était initialement un devis mixte. La pandémie du COVID19 n'a toutefois pas permis d'exploiter le volet qualitatif de notre recherche.

d'échantillonnage. Un courriel contenant l'hyperlien menant vers le questionnaire en ligne sur le site de Surveymonkey.com a été envoyé aux contacts de cette liste d'échantillonnage. Afin d'assurer un plus grand nombre de répondants, le réseau professionnel de la directrice de recherche a aussi été mis à contribution, notamment par la diffusion de l'hyperlien menant au questionnaire sur les réseaux sociaux Facebook et LinkedIn. Pour cette raison, l'échantillon final pouvait inclure des PME dont le nombre d'employés variait de l'objectif initial, soit ayant un effectif inférieur ou égal à 10.

Au départ, la recherche s'appuyait sur un échantillon initial de 61 PME québécoises appartenant aux secteurs d'activité suivants : le manufacturier, transport / logistique, commerce de gros, commerce de détail, services aux entreprises (industriels / scientifiques), finance et assurances. Ce nombre a été réduit à 40 PME, notamment à cause de questionnaires qui ont été seulement ouverts, mais qui sont restés vides de réponses exploitables. L'échantillon final était composé de 29 questionnaires complets et de 11 autres contenant des réponses partielles. Afin de pallier aux réponses contenant des valeurs vides, les données ont fait l'objet d'une analyse préliminaire pour le remplacement des données manquantes. La manière de procéder sera précisée plus loin.

De ces 40 répondants, 35% étaient des femmes contre 65% d'hommes. L'âge moyen de ces répondants se situe entre 40 et 60 ans. La majorité détient un diplôme universitaire de 1er cycle (40%) ou un diplôme collégial (17,5%). Les postes les plus occupés sont les postes de président-directeur général (30%). Ces répondants déclarent majoritairement siéger au comité de direction de leur entreprise (57,5%) et ils sont très impliqués dans les projets TI de leurs entreprises (30%). Un investissement inférieur à 50 000$ (45%) est consacré aux projets TI. La plupart de ces PME n'ont aucune personne dédiée à la gestion des TI (37,5%). La majorité de ces PME œuvrent dans le secteur manufacturier (30%) et la prestation de service (25%). Ces PME sont plus localisées dans les régions Capitale-Nationale (15%) et Montréal (12,5%). De plus, celles-ci ont un effectif majoritairement de 1 à 9 employés (30%) ou de 20 à 49 employés (22,5%). Leurs

chiffres d'affaires se situent majoritairement entre 500 000 et 4 999 999$ (30%) ou il est supérieur à 10 000 000$ (17,5%).

3.2 MÉTHODE DE COLLECTE DES DONNÉES

Le questionnaire est l'instrument de mesure le plus utilisé par les chercheurs.(Fortin et Gagnon, 2016, p. 326). Il permet de recueillir une information factuelle sur différents phénomènes ou sujets d'étude comme des événements, des croyances, ou des connaissances. Les questions, auxquelles les participants doivent répondre, suivent un ordre logique et peuvent être ouvertes ou fermées.

3.2.1 Outils de collecte des données

Présenté à l'ANNEXE B, l'instrument de mesure utilisé est un questionnaire en ligne intitulé « les PME et le traitement de l'information à l'ère des données massives ». Il est composé de 23 questions divisées en sept parties : *la génération des données* (2 questions), *l'acquisition des données* (2 questions), *le stockage des données* (2 questions), *l'analyse des données* (2 questions), *la visualisation / diffusion de l'information* (2 questions), *la qualité de l'information* (1 question) et une partie d'informations sociodémographiques (12 questions).

Ce questionnaire en ligne publié sur Surveymonkey.com débute par une lettre d'information-consentement, fournie à l'ANNEXE C, afin de fournir aux participants toutes les informations nécessaires sur ce projet, par exemple les objectifs, les composantes principales du questionnaire, les bénéfices, les risques, la confidentialité, la compensation, le cas échéant, ainsi que les informations pour contacter la chercheuse et sa directrice de recherche, en cas de besoin. La durée de complétion du questionnaire est estimée à 20 minutes. L'envoi du questionnaire aux répondants est fait à l'aide des

fonctionnalités du site Surveymonkey, incluant un courriel d'invitation. Plusieurs courriels de rappel ont été envoyés aux non-répondants, ainsi qu'aux courriels associés à des questionnaires indiquant des réponses partielles. La collecte s'est complétée par la diffusion de l'hyperlien du questionnaire sur les réseaux sociaux de la directrice de recherche, permettant ainsi d'ajouter une dizaine de répondants à ceux provenant des listes initiales.

3.2.2 Préparation des données

Après la collecte des données, il faut vérifier que le questionnaire est bien rempli. La révision porte alors sur l'admissibilité des informations recueillies, l'incomplétude des données et l'attribution des données aux bons répondants (Fortin et Gagnon, 2016, p. 350).

Concernant les données manquantes évoquées précédemment, la méthode d'imputation simple a été utilisée et « consiste à remplacer chaque donnée manquante par une valeur plausible » (Glasson-Cicognani et Berchtold, 2010, p. 2).

Les données recueillies sont conservées sur le site de Surveymonkey qui permet d'extraire le résultat de ces données sur différents outils d'analyse de données tels que SPSS (*Statistical Package for the Social Sciences*), sous format .csv ou sous format de tableur .xls (Excel).

Selon Fortin et Gagnon (2016, p. 350)« le codage est un processus par lequel les données originales sont transformées en symboles ou codes compatibles avec l'analyse assistée par ordinateur ou d'autres types d'analyses ». Concernant nos variables d'études, nous avons opté pour un codage de type échelle (ou variables continues). Afin de simplifier l'analyse des données, nous avons utilisé des abréviations pour coder ces variables d'études, par exemple *la génération des données* sera codée *GD* et *la diffusion*

/ visualisation de l'information sera codée *VI*. Pour les données sociodémographiques, le codage numérique ordinal a été utilisé pour les variables scolarité, âge, effectif, chiffre d'affaires, investissement TI et implication gestion TI. Tandis que pour les variables genre, secteur d'activité, région, comité de direction et gestion TI, nous utilisions un codage nominal.

3.3 MESURES

3.3.1 Génération des données (GD)

Nous avons mesuré la *génération des données* par deux dimensions : la consultation des données provenant des systèmes internes de l'entreprise (10 items; p. ex. pour générer les données des systèmes internes, je consulte les données provenant du progiciel de gestion intégrée comme SAP, Oracle, etc.) et la consultation des données provenant des systèmes externes de l'entreprise (10 items; p. ex. pour générer les données des systèmes externes, je consulte les publications sur les réseaux sociaux produites par d'autres organisations comme Facebook, Instagram, etc.). Pour ce faire, nous avons utilisé et adapté le matériel et les outils de Baltzan (2018), Bélanger et Van Slyke (2011), Chen et al. (2014), Hu et al. (2014) et Taleb et al. (2018). Pour cette mesure, les participants évaluaient la fréquence de consultation de différentes sources de *génération des données* selon une échelle de type Likert (1932) allant de 1 (jamais) à 5 (très souvent). Cette mesure nous permettra de déterminer les données provenant des systèmes internes informatisés ou non informatisés ainsi que celles provenant des systèmes externes, informatisés ou non informatisés, ou produits par une tierce personne n'appartenant pas au personnel de l'entreprise en définissant la fréquence de leur consultation par ces PME.

3.3.2 Acquisition des données (ACD)

Nous avons mesuré *l'acquisition des données* par deux dimensions : la préparation des données et la transmission des données. La préparation des données est composée de deux sous-activités; la collecte des données (4 items; p. ex. pour collecter les données, j'utilise les outils analytiques en ligne, les outils informatisés de gestion d'activités de l'entreprise, etc.) et le prétraitement (4 items; p. ex. pour le prétraitement des données collectées, j'effectue le nettoyage des données, la réduction des données, la transformation des données chiffriers .xls convertis en .csv, etc.). Pour ce faire, nous avons utilisé et adapté le matériel et les outils de Baltzan (2018), Taleb et al. (2015), Gao et al. (2016) et Osborne (2002), Hu et al. (2014), Salomon (2004), et Zhang et al. (2002). Pour cette mesure, les participants évaluaient la fréquence d'utilisation des activités de préparation des données (collecte et prétraitement des données) pendant la phase de *l'acquisition des données* selon une échelle de type Likert (1932) allant de 1 (jamais) à 5 (très souvent). Cette mesure nous permettra de déterminer les activités de collecte et prétraitement des données de l'entreprise en définissant la fréquence de leur utilisation par ces PME. La transmission des données est mesurée par 7 items (p. ex. pour la transmission des données, j'utilise un système de gestion de bases de données interne, un progiciel de gestion intégré, des serveurs internes entrepôts ou dépôts de données, etc.). Pour ce faire, nous avons utilisé et adapté le matériel et les outils de Baltzan (2018), Krishnan (2013) et Belanger et Van Slyke (2011). Pour cette mesure, les participants évaluaient la fréquence d'utilisation des moyens de transmission des données pendant la phase de *l'acquisition des données* selon une échelle de type Likert (1932) allant de 1 (jamais) à 5 (très souvent). Cette mesure nous permettra de déterminer les moyens de transmission, de transfert et de partage des données vers différents services de l'entreprise en définissant la fréquence de leur utilisation par ces PME.

3.3.3 Stockage des données (SD)

Nous avons mesuré *le stockage des données* par deux dimensions : les dispositifs de stockage internes (3 items; p. ex. pour conserver les données avec des dispositifs de stockage internes de l'entreprise, j'utilise des serveurs internes se trouvant sur les lieux physiques de l'entreprise, des archives numériques conservées ailleurs que sur les serveurs comme les disques durs externes, etc.) et les dispositifs de stockage externes (4 items; p. ex. pour conserver les données avec des dispositifs de stockage externes de l'entreprise, j'utilise des serveurs et bases de données externes en nuage Google drive, l'informatique en grille, etc.). Pour ce faire, nous avons utilisé et adapté le matériel et les outils de Baltzan (2018), Bi et Cochran (2014), Hu et al. (2014), Coleman et al. (2016) et Mell et Grance (2011). Pour cette mesure, les participants évaluaient la fréquence d'utilisation des dispositifs de *stockage des données* selon une échelle de type Likert (1932) allant de 1 (jamais) à 5 (très souvent). Cette mesure nous permettra de déterminer les dispositifs de stockage internes et externes des données pour la conservation des données de l'entreprise en définissant la fréquence de leur utilisation par ces PME.

3.3.4 Analyse des données (AD)

Nous avons mesuré l'analyse des données par deux dimensions : applications et outils d'analyse appartenant ou développés par l'entreprise (8 items; p. ex. pour analyser les données avec des moyens internes, j'utilise des chiffriers électroniques pour manipuler les données et en extraire des informations utiles, des rapports provenant des systèmes de traitement de transactions, etc.) et applications et outils d'analyse développés par des tiers (3 items; p. ex. pour analyser les données avec des moyens externes, j'utilise des outils d'analyse d'intelligence d'affaires sur des données externes comme les affaires électroniques de l'entreprise, l'infrastructure ou logiciels-service (IaaS) pouvant être loués ou facturés à l'utilisation, etc.). Pour ce faire, nous avons utilisé et adapté le matériel

de Baltzan (2018). Pour cette mesure, les participants évaluaient la fréquence d'utilisation des moyens d'analyse des données selon une échelle de type Likert (1932) allant de 1 (jamais) à 5 (très souvent). Cette mesure nous permettra de déterminer les moyens internes et externes d'analyse des données de l'entreprise en définissant la fréquence de leur utilisation par ces PME.

3.3.5 Diffusion / visualisation de l'information (VI)

Nous avons mesuré *la diffusion / visualisation de l'information* par deux dimensions : données destinées au personnel de l'entreprise (5 items; p. ex. j'effectue la diffusion des données destinées au personnel de l'entreprise en utilisant des formulaires ou documents papier de gestion destinés à une utilisation interne, des rapports et extraits de tableaux de bord numériques (indicateurs de performance) destinés aux dirigeants de l'entreprise, etc.) et données destinées à toute personne n'appartenant pas au personnel de l'entreprise (7 items; p. ex. j'effectue la diffusion des données destinées aux personnes n'appartenant pas au personnel de l'entreprise en utilisant le suivi des comptes-clients/comptes-fournisseurs en temps réel dans les affaires électroniques, les données soutenant les activités du marketing numérique, etc.). Pour ce faire, nous avons utilisé et adapté le matériel de Baltzan (2018). Pour cette mesure, les participants évaluaient la fréquence d'utilisation des outils de *diffusion / visualisation de l'information* dans les PME selon une échelle de type Likert (1932) allant de 1 (jamais) à 5 (très souvent). Cette mesure nous permettra de déterminer les outils de diffusion internes, qui sont destinés au personnel de l'entreprise, et les outils de diffusion en externe, qui sont destinés aux personnes n'appartenant pas au personnel de l'entreprise, en définissant la fréquence de leur utilisation par ces PME.

3.3.6 Qualité de l'information (QI)

La qualité de l'information est composée de neuf caractéristiques qui sont : l'exactitude, la crédibilité, la compréhensibilité, la cohérence, la conformité, la disponibilité, la complétude, l'actualité et l'unicité. Chaque caractéristique est mesurée par un seul item. L'item qui mesure l'exactitude de l'information est adapté de Taleb et al. (2015) et Gao et al. (2016). Les items qui mesurent la crédibilité de l'information, la compréhensibilité de l'information et la disponibilité de l'information sont adaptés de Rafique et al. (2012). L'item qui mesure la cohérence de l'information est adapté de Taleb et al. (2015). L'item qui mesure la conformité de l'information est adapté de Belanger et Van Slyke (2011). L'item qui mesure la complétude de l'information est adapté de Taleb et al. (2015) et Rafique et al. (2012). L'item qui mesure l'actualité de l'information est adapté de Taleb et al. (2015) et Merino et al. (2016). L'item qui mesure l'unicité de l'information est adapté de Baltzan (2018). Pour cette mesure, les participants évaluaient la fréquence des caractéristiques de la qualité de l'information en entreprise selon une échelle de type Likert (1932) allant de 1 (pas du tout) à 5 (très souvent). Cette mesure nous permettra de faire l'évaluation de la qualité de l'information qui circule en entreprise.

3.4 PROCESSUS D'ANALYSE DES DONNÉES

Le processus d'analyse des données collectées passe par deux étapes. La première étape est une analyse descriptive des données qui consiste à faire des analyses préliminaires pour la préparation des données, notamment en remplaçant les valeurs manquantes avec la méthode d'imputation simple (Glasson-Cicognani et Berchtold, 2010). Puis une analyse factorielle sera faite sur la base des données imputées pour le but de structuration des données sous forme de facteurs. Cela permettra de passer à la vérification de la fiabilité et de la validité des construits et des variables utilisées. C'est aussi à cette étape que nous établirons le portrait de notre échantillon.

La deuxième étape consiste à faire une analyse fsQCA sur les moyennes obtenues à l'étape précédente.

3.4.1 Analyses descriptives

Les résultats obtenus nous permettront d'avoir les informations nécessaires sur les statistiques descriptives de ces variables (moyennes, fréquences, etc.) et de faire une évaluation de leurs fiabilités en utilisant un indice de cohérence interne tel que le coefficient alpha de Cronbach et la fiabilité composite avec le calcul de Rhô de Fornell et Larcker (1981). De plus, nous allons évaluer les validités de ces construits.

3.4.1.1 Analyse et imputation des valeurs manquantes

L'analyse des valeurs manquantes est une analyse qui permet de définir les valeurs manquantes pour chaque variable étudiée. Des valeurs sont considérées manquantes lorsqu'elles sont absentes parce que les répondants ont oublié de répondre à certaines questions. Pour diminuer ce risque lors de la collecte des données, nous avons défini les questions importantes de notre enquête comme obligatoires. Ce qui veut dire que lorsque le répondant oublie de répondre à une question importante, il y a un message qui s'affiche lui indiquant que la question est obligatoire. Il existe plusieurs types de données manquantes selon la classification de Rubin (1987). Les trois types qui sont les plus répandus sont : les valeurs manquantes non aléatoires (ou *Missing not at random* - MNAR), les valeurs manquantes aléatoires (ou *Missing at random*- MAR), et les valeurs manquantes complètement aléatoires (ou *Missing completely at random*- MCAR). Puisque cette dernière correspond à notre recherche, nous la définissons. Les données manquantes sont complètement aléatoires (MCAR) lorsque la probabilité qu'une valeur soit manquante dépend uniquement des paramètres extérieurs et pas de la variable étudiée ni des autres variables de la base de données (Rubin, 1987). Ce type de données manquantes est très rare.

Une fois que nous avons défini les valeurs manquantes de notre base de données codée, et ce, par le logiciel SPSS V. 27 qui effectuera cette analyse, la prochaine étape est de faire une imputation simple de ces données manquantes. La méthode d'imputation simple est une des méthodes de traitement des données manquantes qui consiste à « remplacer chaque donnée manquante par une valeur plausible » (Glasson-Cicognani et Berchtold, 2010, p. 2). Il existe plusieurs méthodes d'imputation simple, mais celle qui sera utilisée dans cette étude est la méthode d'imputation par régression qui permet de « remplacer les données manquantes par des valeurs prédites selon un modèle de régression » (Glasson-Cicognani et Berchtold, 2010, p. 2). Cette méthode d'imputation est intégrée au logiciel SPSS V. 27 qui fera tous les calculs pour nous.

3.4.1.2 Analyse factorielle exploratoire

La validité par analyse factorielle est une technique statistique qui est « utilisée pour examiner la structure d'un grand ensemble de variables et pour déterminer l'existence des dimensions sous-jacentes à cet ensemble » (Fortin et Gagnon, 2016, p. 302). Cette analyse nécessite un grand échantillon qui est représentatif de la population à l'étude (Waltz et al., 2010). L'analyse factorielle utilise les procédures corrélationnelles pour examiner les relations entre les différents items (ou énoncés) de l'instrument de mesure, qui est dans notre cas le questionnaire en ligne, pour former un facteur qui est composé des items (ou énoncés) qui sont fortement corrélés entre eux. Selon Fortin et Gagnon (2016, p. 302) et (Nunnally, 1994) « les énoncés qui mesurent la même dimension d'un construit appartiennent au même facteur ; les énoncés qui mesurent différentes dimensions devraient relever de facteurs différents » et ceux qui n'appartiennent à aucun facteur par l'absence de corrélation entre les énoncés doivent être enlevés du test (DeVon et al., 2007).

Il existe deux types d'extraction des facteurs ayant deux buts différents qui sont : l'analyse en facteurs communs et spécifiques (AFCS) et l'analyse en composantes

principales (ACP) (Conway et Huffcutt, 2003). Pour notre étude nous utiliserons cette dernière qui « sert à réduire le nombre de variables requises pour calculer un score total pour le trait mesuré, tout en conservant le plus d'informations possible de l'ensemble original » (Bourque et al., 2006, p. 3; Bryant et Yarnold, 1995; Fabrigar et al., 1999). En effet, nous avons choisi d'effectuer cette ACP sur les items et dimensions de chacune des cinq variables étudiées : *génération des données (GD)*, *acquisition des données (ACD)*, *stockage des données (SD)*, *analyse des données (AD)*, *diffusion / visualisation de l'information (VI)* et *la qualité de l'information (QI)*.

Pour l'extraction des facteurs, selon Pett et al. (2003) il est recommandé d'utiliser la matrice des corrélations au lieu de celle des covariances par la facilité d'interprétation de la matrice de structure et la minimisation des risques d'incohérence des solutions obtenues. De plus, il faut vérifier si les corrélations entre les items ne sont pas très fortes ($r > 0,80$) ou très faibles ($r < 0,30$). Cela peut être testé avec le test de sphéricité de Bartlett qui consiste à tester « l'hypothèse nulle selon laquelle la matrice des corrélations serait une matrice d'identité et qu'il n'existerait donc aucune relation entre les items. Si le résultat est significatif, l'hypothèse nulle est rejetée, signalant ainsi qu'il existe des corrélations inter-items » (Bourque et al., 2006, p. 5). Un autre test permet de vérifier les corrélations inter- items qui est le test Kaiser-Meyer-Olkin (KMO) qui permet de « vérifier qu'une fois l'effet linéaire des autres items contrôlé, les corrélations partielles de chaque paire d'items sont faibles, ce qui confirmerait la présence de facteurs latents liant les items entre eux » (Bourque et al., 2006, p. 5). Une valeur de KMO qui est inférieure à 0,60 signifie que l'ajustement des items aux facteurs latents est insuffisant, donc, l'utilisation de l'analyse factorielle n'est pas recommandée (Pett et al., 2003). Un KMO de 0,60 à 0,70 est considéré faible ; de 0,70 à 0,80 est un ajustement moyen ; de 0,80 à 0,90 l'ajustement est bon et au-delà de 0,90 il est excellent. Pour la détermination du nombre des facteurs à extraire des données, il existe plusieurs règles, la plus populaire est le critère Kaiser-Guttman (Conway et Huffcutt, 2003; Fabrigar et al., 1999; Henson et al., 2001; Park et al., 2002; Pohlmann, 2004; Russell, 2002) qui consiste à ne conserver

que les facteurs ayant des valeurs propres supérieures à 1 qui est fixé par défaut dans la majorité des logiciels tel que SPSS d'où sa popularité. Une autre méthode souvent utilisée consiste à fixer au départ le nombre de facteurs voulus en se basant sur les critères théoriques définis dans la partie théorique pour vérifier si la répartition des items en facteurs correspond aux attentes théoriques (Bourque et al., 2006).

Afin de faciliter l'interprétation de la solution factorielle on utilise souvent ce qu'on appelle la rotation des axes (Kieffer, 1998) qui consiste à « faire pivoter les axes autour de l'origine afin d'obtenir un ajustement optimal à la distribution empirique des données » (Bourque et al., 2006, p. 9). Il existe deux types de rotations : la rotation orthogonale (p. ex. Varimax, Quartimax et Equamax) et la rotation oblique (p. ex. Oblimin). Les rotations orthogonales produisent des facteurs indépendants non corrélés contrairement aux rotations obliques qui permettent des corrélations entre les facteurs (Conway et Huffcutt, 2003). La rotation orthogonale de type Varimax proposée par Kaiser est la rotation la plus utilisée (Fabrigar et al., 1999; Henson et al., 2001; Park et al., 2002; Pohlmann, 2004). L'objectif de ce type de rotation est de « minimiser le nombre d'items attribués à un facteur donné en maximisant la variance intra-facteurs, c'est-à-dire l'écart entre les coefficients de saturation factorielle élevés et faibles » (Bourque et al., 2006, p. 9). Donc, pour notre étude nous utiliserons le type de rotation orthogonale de type Varimax.

3.4.1.3 Fidélité des construits

Selon Fortin et Gagnon (2016, p. 293) « la fidélité renvoie à la précision et à la constance des mesures obtenues à l'aide d'un instrument de mesure ». Dans le cadre de notre étude, nous avons choisi de vérifier la fidélité de notre instrument de mesure (questionnaire en ligne) en évaluant le coefficient alpha de Cronbach (α) ainsi que l'évaluation de la fiabilité composite avec le calcul de Rhô de Fornell et Larcker (1981). Selon Fortin et Gagnon (2016, p. 296) « le coefficient alpha de Cronbach (α) est une statistique servant à estimer la cohérence interne des énoncés d'une échelle de mesure

(test). Il est utilisé lorsqu'il existe plusieurs choix d'établissement des scores, comme le propose l'échelle de Likert ». Il se situe entre 0 et 1, plus il est proche de 1 plus la cohérence interne des items de l'échelle est grande (Gliem et Gliem, 2003). George et Mallery (2016, p. 231) fournissent les règles empiriques suivantes pour le coefficient alpha de Cronbach: $\alpha > 0{,}90$ est considéré excellent, $\alpha > 0{,}80$ est bon, $\alpha > 0{,}70$ est acceptable, $\alpha > 0{,}60$ est discutable, $\alpha > 0{,}50$ est médiocre et $\alpha < 0{,}50$ est inacceptable. Nous allons également utiliser un autre indicateur de vérification de la fiabilité d'un construit qui consiste à calculer le Rhô de Fornell et Larcker (1981) permettant de calculer la fiabilité composite de nos variables et dimensions en suivant la formule suivante (p. 45) :

Équation 1. Rhô de la fiabilité composite de Fornell et Larcker (1981) :

$$\rho_c = \frac{\left(\sum_{i=1}^{p} \lambda_{yi}\right)^2}{\left(\sum_{i=1}^{p} \lambda_{yi}\right)^2 + \sum_{i=1}^{p} Var(\varepsilon_i)}$$

Avec $\boldsymbol{\lambda_{yi}}$ contribution factorielle et ε_i variance de l'erreur de l'item i qui est équivalent à

$$\mathbf{1-\lambda_{yi}^2}$$

3.4.1.4 Validité des construits

Selon Fortin et Gagnon (2016, p. 299) « la validité désigne le degré selon lequel un instrument de mesure reflète bien ce qu'il est censé mesurer ». La validité des construits permet « d'évaluer dans quelle mesure les relations entre les énoncés de l'instrument sont cohérentes avec la théorie et les concepts définis de façon opérationnelle » (p. 301). Dans notre étude nous allons nous intéresser à la validité par convergence, la validité par divergence (ou discriminante) et la validité par analyse factorielle (expliquée dans la partie sur l'analyse factorielle exploratoire) qui sont des composantes de la validité des construits. La validité par convergence « consiste à déterminer si différentes méthodes servant à mesurer deux tests d'un même construit donne des résultats similaires. Il y a validité convergente lorsque les mesures sont corrélées positivement entre elles » (p. 303). La validité discriminante ou la validité par

divergence « sert à déterminer la capacité de différencier deux tests mesurant des construits différents par la même méthode ou par des méthodes différentes » (p. 303). C'est-à-dire qu'il devrait y avoir des résultats différents entre les échelles valides et les échelles ne mesurant pas les mêmes construits. En d'autres termes, « on établit qu'il y a validité divergente si les mesures sont corrélées négativement » (p. 303). Ces deux validités seront testées en se basant sur les critères de Fornell et Larcker (1981) qui suggèrent d'utiliser la variance moyenne extraite (VME) qui est calculée selon la formule suivante :

Équation 2. Rhô de la validité convergente de Fornell et Larcker (1981):

$$\rho_{vc} = \frac{\sum_{i=1}^{p} \lambda_{yi}^2}{\sum_{i=1}^{p} \lambda_{yi}^2 + \sum_{i=1}^{p} Var(\varepsilon_i)}$$

Avec $\boldsymbol{\lambda_{yi}}$ contribution factorielle et ε_i variance de l'erreur de l'item i qui est équivalent à $\mathbf{1-\lambda_{yi}^2}$.

Selon ces auteurs, pour s'assurer de la validité convergente, la VME (ρ_{vc}) doit être supérieure à 0,5. De plus, Fornell et Larcker (1981) mesurent la validité discriminante en vérifiant que la racine carrée des VME est plus grande que les différentes corrélations entre les construits.

3.4.2 Analyse fSQCA

Après les analyses descriptives, la prochaine étape est de déterminer les configurations en utilisant une approche configurationnelle à l'aide du logiciel fsQCA permettant de réaliser l'analyse qualitative comparée par les ensembles flous (de Guinea et Raymond, 2020). Nous allons présenter la méthode fsQCA et expliquer les étapes à suivre concernant les données collectées et prétraitées, à savoir : le calibrage des données, l'analyse de nécessité, la table de vérité, les conditions suffisantes, les différents types de

solutions proposées par la méthode (complexe, intermédiaire et parcimonieuse), ainsi que la signification des éléments principaux et périphériques.

La fsQCA est la méthode choisie pour analyser les données collectées pendant la phase empirique. Elle représente une technique d'analyse configurationnelle de deuxième génération (Ragin, 2000, 2009). Elle a été développée à l'origine en science politique par le professeur Charles C. Ragin dans le domaine des sciences sociales (Mendel et Korjani, 2012) pour traiter des échantillons de petite taille qui sont inférieurs à 15 (Cooper et Glaesser, 2016). Cependant, elle peut également être appliquée à des échantillons de taille intermédiaire (12 à 50 cas) qui sont considérés comme des échantillons de petite taille ou small-N par Greckhamer et al. (2013) et à des échantillons de grande taille (50 cas ou plus) ou large-N. De plus, la QCA a parfois été appliquée pour analyser des échantillons de 12 cas ou même moins. Concernant les PME, la fsQCA peut être utilisée pour les entreprises de taille intermédiaire (12 à 499 cas). Tel est notre cas ici avec un échantillon de 40 PME. Elle s'utilise aussi pour les grandes entreprises (500 cas et plus) et parfois aux petites entreprises (inférieure à 12 cas).

Au lieu d'utiliser des effets d'interaction, des algorithmes de *clustering* ou des scores d'écart, une approche de la théorie des ensembles, utilise l'algèbre booléenne pour déterminer quelles combinaisons de caractéristiques organisationnelles se combinent pour aboutir au résultat en question (Boswell et Brown, 1999; Fiss, 2007; Ragin, 2000, 2014). Au centre des approches de la théorie des ensembles se trouve l'idée que les relations entre les différentes variables sont souvent mieux comprises en termes d'appartenance à l'ensemble. Les développements récents utilisent des ensembles flous (ou *fuzzy-set*) où l'appartenance à l'ensemble n'est pas limitée aux valeurs binaires de 0 et 1 (ensembles nets ou *crisp set*), mais peut à la place être définie à l'aide de scores d'appartenance allant de valeurs ordinales à des valeurs continues (Ragin, 2000, 2005). La valeur de 0 est définie comme étant totalement hors de l'ensemble pertinent (ou la non-appartenance à l'ensemble complet) et la valeur de 1 comme l'appartenance à l'ensemble complet. Un

ensemble flou simple et gradué peut contenir les six valeurs suivantes : 1 = appartenance totale; 0,80 = principalement dedans; 0,60 = plus dedans que dehors; 0,40 = plus dehors que dedans; 0,20 = principalement dehors; 0 = exclusion totale.

3.4.2.1 Calibrage des données

La première étape de l'analyse fsQCA consiste à calibrer les conditions et la variable dépendante en ensembles flous. Le calibrage repose sur les connaissances de fond des chercheurs (sur le domaine du problème, le modèle de mesure et/ou les cas) et peut être effectué de deux manières différentes (de Guinea et Raymond, 2020; Liu et al., 2017; Ragin, 2009). La première méthode est la méthode directe en identifiant trois points d'appartenance à un ensemble : l'appartenance totale (1), le point de croisement (0,5) et l'exclusion totale (0). La deuxième méthode est la méthode indirecte en évaluant qualitativement les cas et en remettant à l'échelle les mesures d'origine. Lorsque des échelles de Likert et d'indices sont utilisées pour la collecte des données quantitatives, il est recommandé d'utiliser la procédure d'étalonnage (ou calibrage) direct. Ainsi, compte tenu des échelles de mesure utilisées dans notre recherche (échelles de Likert), un calibrage direct sera utilisé. Donc, le point de croisement sera utilisé comme point d'ancrage à partir duquel les scores d'écart sont calculés, en prenant les valeurs de l'appartenance complète et de la non-appartenance complète comme limites supérieure et inférieure (Fiss, 2011). Dans la version actuelle du progiciel fsQCA (3.0), la transformation est automatisée dans la commande « compute » et peut être facilement exécutée une fois les trois seuils définis (Fiss, 2011).

3.4.2.2 Analyse de nécessité

L'étude des conditions nécessaires (ou des éléments / variables de configuration) est généralement la deuxième étape de l'analyse fsQCA (de Guinea et Raymond, 2020; Ragin, 2000). Une condition causale est nécessaire (mais pas suffisante) (Cn) pour un résultat lorsque les instances d'un résultat constituent un sous-ensemble d'instances d'une cause (Ragin, 2006). Selon Depeyre et Vergne (2018, p. 13) « une condition Cn est

nécessaire au phénomène P si et seulement si à chaque fois que P est observé, alors Cn est présente ». Dans son livre sur la science sociale des ensembles flous, Ragin (2000) a introduit un indicateur appelé « cohérence nécessaire », qui est valable à la fois pour la csQCA[17] et la fsQCA. La mesure de la « cohérence nécessaire » donne le degré de nécessité d'une seule condition ou d'une configuration de conditions pour le résultat, c'est-à-dire la proportion de cas révélant, à la fois, la condition et le résultat parmi les cas révélant le résultat (Rihoux et Marx, 2013). En d'autres termes, c'est la somme des valeurs minimales d'une condition et d'un résultat dans tous les cas divisée par la somme des valeurs de ce résultat dans tous les cas (Ragin, 2000). En supposant que Y est le résultat et Xi est la condition, cela peut être écrit comme suit : Cohérence nécessaire = $\sum$ (min (Xi, Y)) / $\sum$ (Y). Une condition est nécessaire lorsque son score de cohérence est supérieur à 0,9 (de Guinea et Raymond, 2020; Schneider et Wagemann, 2012; Schneider et al., 2010), plus la mesure est proche de 1, plus la condition doit être cohérente. Par exemple une cohérence nécessaire de 0,98 signifierait que les relations nécessaires entre la condition ou la configuration des conditions sont supportées par la table de vérité.

3.4.2.3 Table de vérité

Après avoir calibré les conditions et la variable expliquée (dépendante) ainsi que la définition des conditions nécessaires, la troisième étape consiste à utiliser ces mesures d'ensemble pour construire une matrice des données connue sous le nom de table de vérité avec 2^k lignes, où k est le nombre de conditions causales utilisées dans l'analyse (Fiss, 2011). Chaque ligne de ce tableau est associée à une combinaison spécifique d'attributs, et le tableau complet répertorie ainsi toutes les combinaisons possibles. La table de vérité est ensuite triée en fonction de la fréquence et de la cohérence (de Guinea et Raymond, 2020; Pappas et al., 2017; Ragin, 2009). La fréquence représente le nombre d'observations pour chaque configuration possible. Pour les échantillons de plus de 150 cas, il est recommandé de définir le seuil de fréquence à 3, tandis que pour les échantillons plus

[17] **csQCA :** *crisp set Qualitative Comparative Analysis* ou l'analyse qualitative comparée des ensembles nets.

petits, le seuil recommandé est de 2 (Ragin, 2006). Puisque nous avons un échantillon de 40 PME qui est petit, nous avons choisi un seuil de fréquence de 1. La cohérence se réfère au degré auquel les cas correspondent aux relations de la théorie des ensembles exprimées dans une solution (Fiss, 2011). Un moyen simple d'estimer la cohérence lors de l'utilisation d'ensembles flous est la proportion d'observations cohérente avec le résultat. Puisque le seuil minimum recommandé pour la cohérence est de 0,75 (Ragin, 2006; Rihoux et Ragin, 2008), nous choisissons une cohérence minimale de 0,80 pour les solutions. Pour les configurations au-dessus du seuil de cohérence, le résultat de la variable sera fixé à 1 (car ces configurations sont celles qui expliquent complètement le résultat) et le reste sera fixé à 0 (de Guinea et Raymond, 2020; Pappas et al., 2017).

La table de vérité utilise un algorithme basé sur l'algèbre booléenne ou l'algèbre de Boole, qui est un système de notation permettant la manipulation algébrique d'énoncés logiques (Fiss, 2007). Cet algorithme permet la réduction logique des lignes obtenues en combinaisons simplifiées pour avoir l'expression la plus courte possible appelée la formule minimale listant les principaux impliquants (Fiss, 2011; Ragin, 2005; Rihoux et Ragin, 2008). Il est basé sur une analyse contrefactuelle des conditions causales, qui a l'avantage de permettre une catégorisation des conditions causales en conditions centrales et périphériques qu'on va expliquer dans la partie consacrée à cet effet. Il appartient ensuite au chercheur d'interpréter la formule minimale obtenue, éventuellement en termes de causalité (Rihoux et Marx, 2013). Selon Fiss (2007), cette procédure de réduction (ou minimisation) utilise l'algorithme QuineMcCluskey, un algorithme commun pour simplifier les énoncés de la théorie des ensembles qui est implémentée dans des progiciels tels que le QCA et le fs / QCA (Drass et Ragin, 1999).

3.4.2.4 Conditions suffisantes

Une condition est suffisante lorsqu'un résultat sera toujours obtenu si l'attribut en question est présent (Fiss, 2007). Selon Rihoux et Marx (2013), le degré de suffisance d'une condition indique dans quelle mesure une condition peut être liée à l'explication d'un

résultat. Après avoir défini les conditions causales qui mènent au résultat souhaité, le logiciel de fsQCA utilise l'algèbre booléenne pour avoir au final la formule minimale composée des conditions les plus pertinentes (Depeyre et Vergne, 2018). Cette formule minimale « liste les différentes configurations de conditions suffisantes, ainsi que les scores de cohérence et de couverture qui évaluent la validité de la solution » (p. 15). Dans notre étude, il y a trois configurations de conditions suffisantes. La première configuration est sur la présence d'une haute qualité de l'information des données massives dans les PME. La deuxième est sur la présence d'une qualité moyenne de l'information des données massives dans les PME. La dernière est sur l'absence de qualité de l'information des données massives dans les PME.

3.4.2.5 Solutions complexes, parcimonieuses et intermédiaires

Pour traiter le problème de la diversité limitée à l'aide de l'analyse contrefactuelle, l'algorithme de la table de vérité distingue les solutions parcimonieuses et intermédiaires sur la base de contrefactuels « faciles » et « difficiles » (Fiss, 2011; Ragin, 2009). Les contrefactuels « faciles » font référence à des situations dans lesquelles une condition causale redondante est ajoutée à un ensemble de conditions causales qui, par elles-mêmes, conduisent déjà au résultat en question. Dans ce cas-là, le chercheur se pose la question suivante : est-ce que l'ajout d'une autre condition causale ferait une différence? Si la réponse est non, on peut procéder avec l'expression simplifiée. En revanche, les contrefactuels « difficiles » font référence à des situations dans lesquelles une condition est supprimée d'un ensemble de conditions causales menant à un résultat en supposant que cette condition est redondante. Dans une analyse contrefactuelle difficile, un chercheur pose la question suivante : est-ce que la suppression d'une condition causale ferait une différence? Cette question est plus difficile à répondre. La distinction des contrefactuels faciles et difficiles permet d'établir deux types de solutions. La première est une solution parcimonieuse qui inclut toutes les hypothèses simplificatrices, qu'elles soient basées sur des contrefactuels faciles ou difficiles. Elle produit les conditions les plus importantes, appelées conditions ou éléments de base ou principaux, qui ne peuvent être omis d'aucune

configuration (de Guinea et Raymond, 2020; Pappas et al., 2019). La seconde est une solution intermédiaire qui n'inclut que des hypothèses simplificatrices basées sur des contrefactuels faciles (Fiss, 2011). Elle comprend la solution parcimonieuse et fait partie de la solution complexe (Liu et al., 2017; Rihoux et Ragin, 2008). La troisième solution, est la plus complexe qui n'inclut ni les contrefactuels faciles ni difficiles. Cependant, une telle solution est généralement inutilement complexe et fournit assez peu d'informations sur les configurations causales.

La notion de conditions causales, appartenant à des configurations principales ou périphériques, est basée sur les solutions parcimonieuses et intermédiaires : les conditions principales sont celles qui font partie à la fois aux solutions parcimonieuses et intermédiaires, et les conditions périphériques sont celles qui sont éliminées de la solution parcimonieuse et n'apparaissent donc que dans la solution intermédiaire (Fiss, 2011). La recommandation est d'utiliser une combinaison de solutions parcimonieuses et intermédiaires pour interpréter les résultats de la fsQCA (de Guinea et Raymond, 2020). Plus précisément, le chercheur doit identifier les conditions de la solution parcimonieuse dans la solution intermédiaire afin de pouvoir créer un tableau qui comprend à la fois les éléments principaux et périphériques. Il en résulte une solution combinée qui présente les éléments principaux et périphériques qui aident l'interprétation des configurations résultantes.

3.4.2.6 Éléments principaux et périphériques

Selon Fiss (2011) et Hannan et al. (1996), les éléments principaux sont essentiels et les éléments périphériques sont moins importants et peuvent être échangeables. Les éléments principaux sont définis comme des conditions causales pour lesquelles la preuve indique une forte relation causale avec le résultat d'intérêt, et les éléments périphériques comme celles pour lesquelles la preuve d'une relation causale avec le résultat est plus faible. Les impliquants principaux qui sont les éléments principaux d'une configuration sont marqués par de grands cercles (Mikalef et Pateli, 2017; Mikalef et al., 2015). S'il y

a une présence d'une condition principale, elle sera représentée par un grand cercle noir (●) indiquant la présence de la condition (Rihoux et Ragin, 2008), dans le cas contraire elle sera représentée par un grand cercle vide avec une croix à l'intérieur (⊗) indiquant l'absence de la condition principale. Les éléments périphériques sont marqués par des petits cercles. En cas de présence d'une condition périphérique, elle sera représentée par un petit cercle noir (•), sinon, on utilisera un petit cercle vide avec une croix à l'intérieur (⊗) indiquant l'absence de la condition périphérique. Les espaces vides sont une indication d'une situation ou une condition immatérielle (ou « *don't care* » en anglais) dans laquelle la condition causale peut être présente ou absente sans modifier la relation causale entre la configuration et le résultat (de Guinea et Raymond, 2020; Ragin, 2009).

CHAPITRE 4 – RÉSULTATS

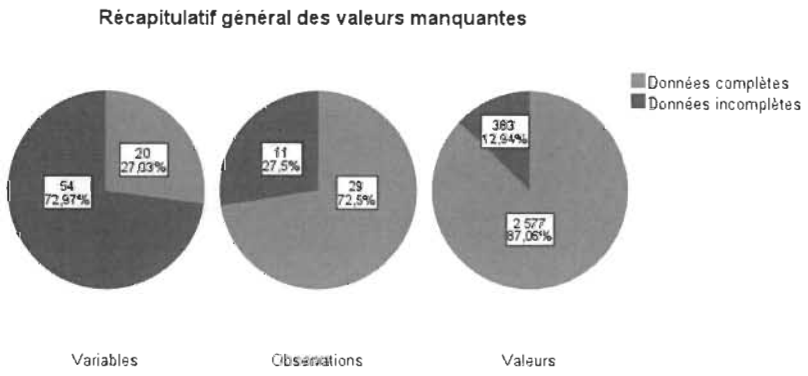
Cette section présente les résultats de l'analyse des données. Elle débute par une analyse préliminaire des données collectées, incluant la procédure de remplacement des valeurs manquantes avec la méthode d'imputation simple. Puis, une étude de vérification de la fiabilité et de la validité des concepts sera effectuée. Nous terminons avec les résultats de l'analyse fsQCA.

4.1 ANALYSES PRÉLIMINAIRES

Les statistiques univariées obtenues après l'analyse des données manquantes nous permettent d'avoir une vision globale de toutes les variables[18]. Fournissant ainsi des informations sur le nombre de valeurs complètes (N) et les valeurs manquantes, ainsi que le pourcentage de ces dernières. La Figure 6 ci-dessous présente un récapitulatif général des valeurs manquantes avec trois diagrammes circulaires qui illustrent des aspects différents des valeurs manquantes dans les données.

Figure 6

Récapitulatif général des valeurs manquantes du questionnaire



[18] Les analyses complètes des valeurs manquantes et du processus d'imputation pourront être fournies sur demande du lecteur.

Le diagramme *Variables* indique que 54 variables d'analyse (72.97%) contiennent au moins une valeur manquante pour une observation. Le diagramme *Observations* indique que 11 des 40 observations contiennent au moins une valeur manquante pour une variable. Le diagramme *Valeurs* indique que 383 des 2960 valeurs (observations × variables) sont manquantes. Cela signifie que l'élimination des observations incomplètes supprimerait quelques informations utiles dans l'ensemble des données.

Avant d'imputer les données manquantes, il faut déterminer si ces valeurs sont aléatoires et dues au hasard. Les données manquantes sont considérées complètement aléatoires (MCAR) quand la valeur de $p > 0.05$ (Tabachnick et al., 2007). En utilisant la méthode EM (*expectation-maximisation*) sur SPSS pour l'analyse des données manquantes, nous obtenons plusieurs tableaux, dont le Tableau « Moyennes EM » qui représente le test MCAR pour chaque variable. La moyenne minimale alors observée est de 1,36 pour la variable ADI8 (voir ANNEXE D) qui est supérieure à 0,05. On conclut ainsi que les données manquantes dans notre base de données sont considérées complètement aléatoires (MCAR).

Les statistiques descriptives au Tableau 3 présentent les récapitulatifs de l'imputation des variables contribuant au modèle étudié.

Tableau 3

Statistiques descriptives de la variable TD1

TD1						
Données	**Imputation**	**N**	**Moy**	**Écart type**	**Min**	**Max**
Données d'origine		35	2,7143	1,67282	1,0000	5,0000
Valeurs imputées	1	5	3,2501	1,35427	1,8543	5,0000
Données complètes après imputation	1	40	2,7813	1,63091	1,0000	5,0000

Suivant le Tableau 3, les données qui ont été imputées de l'item TD1 (voir ANNEXE D) sont au nombre de 5 et les données complètes sont de 35 valeurs. La moyenne et l'écart-type des données d'origine étaient respectivement de 2,7143 et 1,67282 avec un minimum de 1 et un maximum de 5 et après imputation des données ils sont devenus 2,7813 et 1,63091 qui est proche des valeurs d'origine en gardant les mêmes valeurs minimales et maximales. Cette analyse d'imputation simple des données génère une autre base de données avec des valeurs imputées où les données manquantes ont été remplacées. C'est-à-dire que notre nouvelle base de données imputées ne contient pas de valeurs manquantes. Dorénavant nous utiliserons cette base de données pour toutes nos prochaines analyses.

4.2 FIABILITÉ ET VALIDITÉ DES CONCEPTS

Pour vérifier la fiabilité et la validité de chaque concept de cette recherche, il faut passer par plusieurs étapes : l'analyse descriptive, la validation de la structuration factorielle, le test de la cohérence interne (la fiabilité) et les validités convergente et discriminante[19].

4.2.1 Analyse factorielle exploratoire et fiabilité du concept Génération des données *(GD)*

4.2.1.1 GD : variable bidimensionnelle

Après trois itérations de l'analyse factorielle (ACP) où nous avons retiré les items CDI1, CDI2, CDI5, CDI8, CDI9, CDE6 et CDE7 ayant une valeur d'extraction inférieure à 0,400, nous obtenons une valeur de test KMO 0,733 qui est moyen (Bourque et al., 2006) et un test de Bartlett significatif et une variance totale expliquée de 57,283 % avec

[19] Les analyses complètes sur l'analyse factorielle exploratoire des données pourront être fournies sur demande du lecteur.

deux facteurs. Comme indiqué à l'ANNEXE E, les qualités des représentations sont bonnes allant de 0,402 à 0,784. Dans l'ANNEXE F le premier facteur « consultation des données provenant des systèmes internes de l'entreprise » contient les sept items suivants : CDI3, CDI6, CDI7, CDI10, CDE8, CDE9 et CDE10. Le deuxième facteur « consultation des données provenant des systèmes externes de l'entreprise » contient les six items suivants : CDE1, CDE2, CDE3, CDE4, CDE5, CDI4.

4.2.1.2 Fiabilité de l'instrument de mesure de GD

Comme indiqué à l'ANNEXE E, la cohérence interne des mesures est évaluée par le coefficient alpha de Cronbach obtenu par le logiciel SPSS. Pour la variable *GD* qui contient 13 items, nous obtenons un alpha égal à 0,856 qui ne laisse pas envisager une mauvaise qualité psychométrique de notre échelle, mais une bien bonne qualité. Pour les dimensions CDI (7 items) et CDE (6 items), nous obtenons des valeurs respectives de 0,841 et 0,838 qui sont bonnes. Nous avons également calculé le Rhô de Fornell et Larcker (1981) et nous avons obtenu une valeur de la fiabilité composite égale à 0,930, ce qui est excellent. En moyenne, les répondants font occasionnellement (M=2,64) *la génération des données*. Plus particulièrement, ils consultent rarement (M=2,45) les données provenant des systèmes internes de l'entreprise et occasionnellement (M= 2,85) les données provenant des systèmes externes de l'entreprise.

4.2.2 Analyse factorielle exploratoire et fiabilité du concept Acquisition des données (ACD)

4.2.2.1 ACD : variable bidimensionnelle

Après trois itérations de l'ACP où nous avons retiré les items PD_CD1, TD3 et TD5 ayant une valeur d'extraction inférieure à 0,400, nous obtenons un KMO de 0,681 qui est faible, mais acceptable (Bourque et al., 2006) et un test de Bartlett significatif et une variance totale expliquée de 58,374 % avec deux facteurs. Comme indiqué à

l'ANNEXE E, les qualités des représentations sont bonnes allant de 0,398 (proche de 0,400) à 0,891. Dans l'ANNEXE F le premier facteur « préparation des données » contient les dix items suivants : PD_CD2, PD_CD3, PD_PT1, PD_PT2, PD_PT3, PD_PT4, TD1, TD2, TD6, TD7. Le deuxième facteur « transmission des données » contient les deux items suivants : TD4, PD_CD4.

4.2.2.2 Fiabilité de l'instrument de mesure de ACD

Comme indiqué à l'ANNEXE E, pour la variable *ACD* qui contient 12 items, nous obtenons un alpha égal à 0,831 qui ne laisse pas envisager une mauvaise qualité psychométrique de notre échelle, mais une bonne qualité. Pour les dimensions PD (10 items) et TD (2 items), nous obtenons des valeurs respectives de 0,880 et 0,878 qui sont bonnes. Nous avons également calculé le Rhô de Fornell et Larcker (1981) et nous avons obtenu une valeur de 0,932, ce qui est bon. En moyenne, les répondants font occasionnellement (M=2,76) *l'acquisition des données*. Plus particulièrement, ils réalisent occasionnellement (M=2,78) les activités de préparation des données et utilisent occasionnellement (M=2,65) les moyens de transmission des données.

4.2.3 Analyse factorielle exploratoire et fiabilité du concept Stockage des données (SD)

4.2.3.1 SD : variable bidimensionnelle

Après trois itérations de l'ACP où nous avons retiré l'item DSE1 ayant une valeur d'extraction inférieure à 0,400, nous obtenons un KMO de 0,532 qui est insuffisant (Bourque et al., 2006) et un test de Bartlett significatif et une variance totale expliquée de 58,120 % avec deux facteurs. Comme indiqué à l'ANNEXE E, les qualités des représentations sont bonnes allant de 0,473 à 0,708. Dans l'ANNEXE F le premier facteur « dispositifs de stockage externes » contient les trois items suivants : DSE2, DSE3, DSI3.

Le deuxième facteur « dispositifs de stockage internes » contient les trois items suivants : DSI1, DSI2, DSE4.

4.2.3.2 La fiabilité de l'instrument de mesure du SD

Comme indiqué à l'ANNEXE E, pour la variable *SD* qui contient 6 items, nous obtenons un alpha égal à 0,627 qui est discutable. Pour les dimensions DSI (3 items) et DSE (3 items), nous obtenons des valeurs respectives de 0,562 et 0,642 qui est médiocre pour la première et discutable pour la deuxième. Cela peut être expliqué par le nombre faible d'items (3) pour chacune de ces deux dimensions sachant que le calcul du coefficient de fiabilité α prend en considération le nombre d'items dans sa formule qui est la suivante (Gliem et Gliem, 2003) : $rk / [1 + (k-1) r]$ où k est le nombre d'éléments considérés et r est la moyenne des corrélations inter-éléments. Nous avons également calculé le Rhô de Fornell et Larcker (1981) et nous avons obtenu une valeur de 0,872 qui est bonne. En moyenne, les répondants font occasionnellement (M=2,92) *le stockage des données*. Plus particulièrement, ils utilisent occasionnellement (M=3,15) les dispositifs de stockage internes et occasionnellement (M=2,68) les dispositifs de stockage externes.

4.2.4 Analyse factorielle exploratoire et fiabilité du concept Analyse des données (AD)

4.2.4.1 AD : variable bidimensionnelle

Après trois itérations de l'ACP où nous avons retiré l'item ADI2 ayant une valeur d'extraction inférieure à 0,400, nous obtenons un KMO de 0,844 qui est bon (Bourque et al., 2006) et un test de Bartlett significatif et une variance totale expliquée de 62,111% avec deux facteurs. Comme indiqué à l'ANNEXE E, les qualités des représentations sont bonnes allant de 0,518 à 0,749. Dans l'ANNEXE F le premier facteur « applications et outils d'analyse appartenant ou développés par l'entreprise » contient les six items suivants : ADI1, ADI3, ADI4, ADI5, ADI6, ADI7. Le deuxième facteur « applications et

outils d'analyse développés par des tiers » contient les quatre items suivants : ADE1, ADE2, ADE3, ADI8.

4.2.4.2 La fiabilité de l'instrument de mesure du AD

Comme indiqué à l'ANNEXE E, pour la variable *AD* qui contient 10 items, nous obtenons un alpha égal à 0,866 qui ne laisse pas envisager une mauvaise qualité psychométrique de notre échelle, mais une bonne qualité. Pour les dimensions ADI (6 items) et ADE (4 items), nous obtenons des valeurs respectives de 0,848 et 0,788 qui sont bonnes. Nous avons également calculé le Rhô de Fornell et Larcker (1981) et nous avons obtenu une valeur de la fiabilité composite égale à 0,917 qui confirme la consistance interne. En moyenne, les répondants font rarement (M=2,02) *l'analyse des données*. Plus particulièrement, ils utilisent rarement (M=2,29) les applications et outils d'analyse appartenant ou développés par l'entreprise et rarement (M=1,61) les applications et outils d'analyse développés par des tiers.

4.2.5 Analyse factorielle exploratoire et fiabilité du concept Diffusion / visualisation de l'information (VI)

4.2.5.1 VI : variable bidimensionnelle

Après deux itérations de l'ACP où nous avons retiré l'item VIE6 ayant une valeur d'extraction inférieure à 0,400, nous obtenons un KMO de 0,630 qui est faible, mais acceptable (Bourque et al., 2006) et un test de Bartlett significatif et une variance totale expliquée de 56,673 % avec deux facteurs. Comme indiqué à l'ANNEXE E, les qualités des représentations sont bonnes allant de 0,388 (proche de 0,400) à 0,776. Dans l'ANNEXE F le premier facteur « données destinées à toute personne n'appartenant pas au personnel de l'entreprise » contient les huit items suivants : VIE1, VIE2, VIE3, VIE4, VIE5, VIE7, VII4, VII5. Le deuxième facteur « données destinées au personnel de l'entreprise » contient les trois items suivants : VII1, VII2, VII3.

4.2.5.2 La fiabilité de l'instrument de mesure de VI

Comme indiqué à l'ANNEXE E, pour la variable *VI* qui contient 11 items, nous obtenons un alpha égal à 0,817 qui ne laisse pas envisager une mauvaise qualité psychométrique de notre échelle, mais une bonne qualité. Pour les dimensions VII (3 items) et VIE (8 items), nous obtenons des valeurs respectives de 0,790 et 0,843 qui sont bonnes. Nous avons également calculé le Rhô de Fornell et Larcker (1981) et nous avons obtenu une valeur de la fiabilité composite égale à 0,923 qui confirme la consistance interne. En moyenne, les répondants font occasionnellement (M=2,68) *la diffusion / visualisation de l'information.* Plus particulièrement, ils effectuent occasionnellement (M=3,09) la diffusion des données destinées au personnel de l'entreprise, mais ils effectuent rarement (M=2,53) la diffusion des données destinées à toute personne n'appartenant pas au personnel de l'entreprise.

4.2.6 Analyse factorielle exploratoire et fiabilité du concept Qualité de l'information (QI)

4.2.6.1 QI : variable unidimensionnelle

Après quatre itérations de l'ACP où nous avons retiré les items QI2, QI3 et QI7 ayant une valeur d'extraction inférieure à 0,400, nous obtenons un KMO de 0,819 qui est bon (Bourque et al., 2006) et un test de Bartlett qui est significatif. Comme indiqué à l'ANNEXE E, les qualités des représentations sont bonnes allant de 0,410 à 0,728. La variance totale expliquée est de 63,591% de l'échantillon avec un seul facteur. Dans l'ANNEXE F ce facteur est composé des six items suivants : QI1 (exactitude), QI4 (cohérence), QI5 (conformité), QI6 (disponibilité), QI8 (actualité) et QI9 (unicité).

4.2.6.2 La fiabilité de l'instrument de mesure de QI

Comme indiqué à l'ANNEXE E, pour la variable *QI*, nous obtenons un alpha égal à 0,883 qui ne laisse pas envisager une mauvaise qualité psychométrique de notre échelle, mais une bonne qualité voire une excellente qualité. Nous avons également calculé le Rhô de Fornell et Larcker (1981) et nous avons obtenu une valeur égale à 0,912 qui confirme la consistance interne. En moyenne, les répondants considèrent souvent (M=4,09) que l'information qui circule est exacte, cohérente, conforme, disponible, actuelle et unique.

En résumé, l'analyse factorielle exploratoire des construits nous a permis de purifier les variables étudiées avant de mesurer la fiabilité de chaque variable et dimension en se basant sur deux méthodes qui sont : le coefficient alpha de Cronbach (α) et le Rhô de Fornell et Larcker (1981) ou la fiabilité composite (FC). Cette analyse est synthétisée dans les Tableaux de l'ANNEXE E et F. De plus, elle nous a permis de réduire le nombre de critères de la qualité de l'information de neuf critères, définis dans le modèle conceptuel initial, à six critères qui sont : l'exactitude, la cohérence, la conformité, la disponibilité, l'actualité et l'unicité.

4.2.7 Validité convergente / discriminante

Selon Fornell et Larcker (1981) pour s'assurer de la validité convergente, la VME (ρ_{vc}) doit être supérieure à 0,5. De ce fait, nous avons calculé la VME de toutes les variables selon la formule de Fornell et Larcker (1981) qui sont présentées dans le Tableau 4 ci-dessous. Nous remarquons que toutes les VME obtenues sont supérieures ou égales à 0,5 avec des valeurs comprises entre 0,512 et 0,636. Dans l'ensemble, nous pouvons admettre que la validité convergente est confirmée.

Tableau 4

Variance moyenne extraite et fiabilité composite des variables

Variables	GD	ACD	SD	AD	VI	QI
VME	0,512	0,538	0,537	0,527	0,526	0,636
FC	0,930	0.932	0,873	0.917	0,923	0,912

Pour la validité discriminante, Fornell et Larcker (1981) la mesurent en vérifiant que la racine carrée des VME est plus grande que les différentes corrélations entre les construits. Dans le Tableau 5 ci-dessous, nous remarquons que la racine carrée des VME de chaque variable est largement supérieure aux corrélations entre celles-ci sauf pour les variables *acquisition des données* et *analyse des données* où la valeur de la racine carrée de la VME (0,73) de ces deux variables est à peu près égale à la valeur de la corrélation entre celles-ci (0,727). Cela veut dire que les répondants ont considéré que les activités *d'acquisition des données* et *analyse des données* appartiennent à une seule phase ou processus. Dans ce cas-là nous ne pouvons pas affirmer que la validité discriminante des construits est confirmée.

Tableau 5

Analyse de la validité discriminante des construits

	1	2	3	4	5	6	VME
1. Génération des données	**0,72**						**0,512**
2. Acquisition des données	,401*	**0,73**					**0,538**
3. Stockage des données	,437**	,365*	**0,73**				**0,537**
4. Analyse des données	,388*	,727**	0,211	**0,73**			**0,527**
5. Diffusion / visualisation de l'information	,313*	,602**	,536**	,405**	**0,73**		**0,526**
6. Qualité de l'information	0,030	-,336*	-0,082	-0,121	-0,140	**0,80**	**0,636**
VME	**0,512**	**0,538**	**0,537**	**0,527**	**0,526**	**0,636**	

**. La corrélation est significative au niveau 0.01; En diagonale : racine carrée de VME

Selon Farrell (2010) si les problèmes de validité discriminante persistent, il n'existe peut-être pas d'autre choix que de combiner les construits en une seule mesure globale. Donc, nous fusionnons les activités *acquisition des données* et *analyse des données* en une seule activité qu'on appellera *acquisition et analyse des données* et qui sera codée *AAD*. La prochaine étape consiste à repasser par toutes les étapes précédentes de calcul de l'ACP, de la cohérence interne, de la validité convergente et la validité discriminante de notre nouvelle variable *acquisition et analyse des données* ou *AAD*.

Après deux itérations de l'ACP de la nouvelle variable *AAD*, après factorisation (sans les items PD_CD1, TD3, TD5 et ADI2), et en forçant le nombre de facteurs voulus à quatre facteurs, nous obtenons un KMO de 0,631 qui est faible, mais acceptable (Bourque et al., 2006). Comme indiqué à l'ANNEXE E, le test de Bartlett est significatif et la variance totale expliquée est de 65,173 % avec quatre facteurs. Les qualités des représentations sont bonnes allant de 0,382 à 0,803. Dans l'ANNEXE F le premier facteur, qui sera appelé *outils et pratiques de gestion des données* et codé *PGD*, contient les sept items suivants : PD_CD2, PD_CD3, PD_PT3, TD1, TD2, ADI1, ADI6. Le deuxième facteur, qui sera appelé *préparation, partage et traitement des données* et codé *PTD*, contient les huit items suivants : PD_PT1, PD_PT2, PD_PT4, TD6, TD7, ADI3, ADI5, ADI7. Le troisième facteur, qui sera appelé *outils d'intelligence d'affaires* et codé *OIA*, contient les cinq items suivants: ADI4, ADI8, ADE1, ADE2, ADE3. Le quatrième facteur, qui sera appelé *outils non informatisés* et codé *ONI*, contient les deux items suivants : PD_CD4, TD4.

Comme indiqué à l'ANNEXE E, la fiabilité de la variable *AAD* a été évaluée par le coefficient de Cronbach avec une valeur de 0,906 qui est excellente, et par la fiabilité composite en calculant le Rhô de Fornell et Larcker (1981) qui nous a donné une valeur de 0,950 qui est excellente. Les valeurs respectives de la fiabilité des dimensions *PGD*, *PTD*, *OIA* et *ONI* selon le coefficient alpha de Cronbach sont de : 0,889; 0,852; 0,819 et 0,878 qui sont bonnes. En moyenne, les répondants font rarement (M=2,42) *l'acquisition*

et l'analyse des données. Plus particulièrement, ils utilisent occasionnellement : les outils et pratiques de gestion des données(M=2,62); la préparation, le partage et le traitement des données (M= 2,68) et les outils non informatisés (M=2,65). Mais, ils utilisent rarement (M= 1,62) les outils d'intelligence d'affaires.

Pour conclure, nous obtenons cinq variables qui sont : *GD*, *AAD*, *SD*, *VI* et *QI* pour lesquelles nous allons vérifier la validité convergente et la validité discriminante. Pour la validité convergente, le Tableau 6 ci-dessous montre que les valeurs des VME varient entre 0,476 (proche de 0,50) et 0,636 qui sont considérées supérieures ou égales à 0,5. Donc, dans l'ensemble, nous pouvons admettre que la validité convergente des construits est confirmée.

Tableau 6

Variance moyenne extraite et fiabilité composite des variables

après fusion des variables ACD et AD en AAD

Variables	GD	AAD	SD	VI	QI
VME	0,512	0,476	0,537	0,526	0,636
FC	0,930	0.950	0,873	0,923	0,912

Dans le Tableau 7, nous remarquons que la racine carrée des VME de chacune de ces variables est largement supérieure aux corrélations entre celles-ci. Donc, dans l'ensemble, nous pouvons admettre que la validité discriminante des construits est confirmée.

Tableau 7

Analyse de la validité discriminante des construits après fusion des variables ACD et AD en AAD

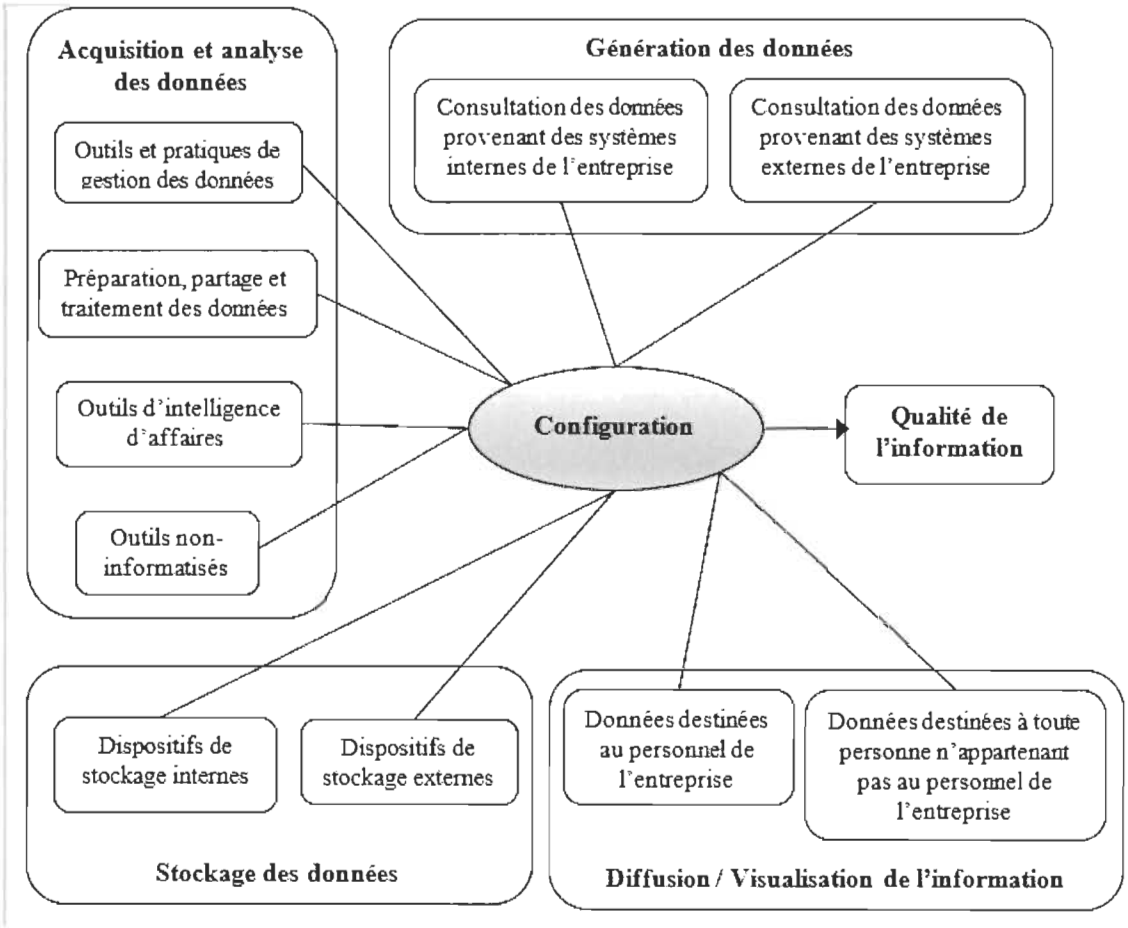
	1	2	3	4	5	VME
1. Génération des données	**0,72**					**0,512**
2. Acquisition et analyse des données	,425**	**0,69**				**0,476**
3. Stockage des données	,437**	,314*	**0,73**			**0,537**
4. Diffusion / visualisation de l'information	,313*	,547**	,536**	**0,73**		**0,526**
5. Qualité de l'information	0,030	-0,252	-0,082	-0,140	**0,80**	**0,636**
VME	**0,512**	**0,476**	**0,537**	**0,526**	**0,636**	

**. La corrélation est significative au niveau 0.01; En diagonale : racine carrée de VME

Maintenant prêt à procéder à l'analyse fsQCA, la Figure 7 présente le modèle configurationnel modifié prenant en compte les résultats des analyses préliminaires réalisées. Cette Figure constitue le modèle configurationnel final du traitement de l'information dans les PME à l'ère des données massives, et ce, par l'intermédiaire des activités et outils de gestion (*génération*, *acquisition et analyse*, *stockage des données*, *diffusion / visualisation de l'information*) avec leurs sous-activités (ou dimensions) respectives. Le but étant d'expliquer la variable dépendante qui est représentée par la qualité de l'information à atteindre en aval (haute, moyenne, absence).

Figure 7

Modèle configurationnel final du traitement de l'information dans les PME à l'ère des données massives[20]



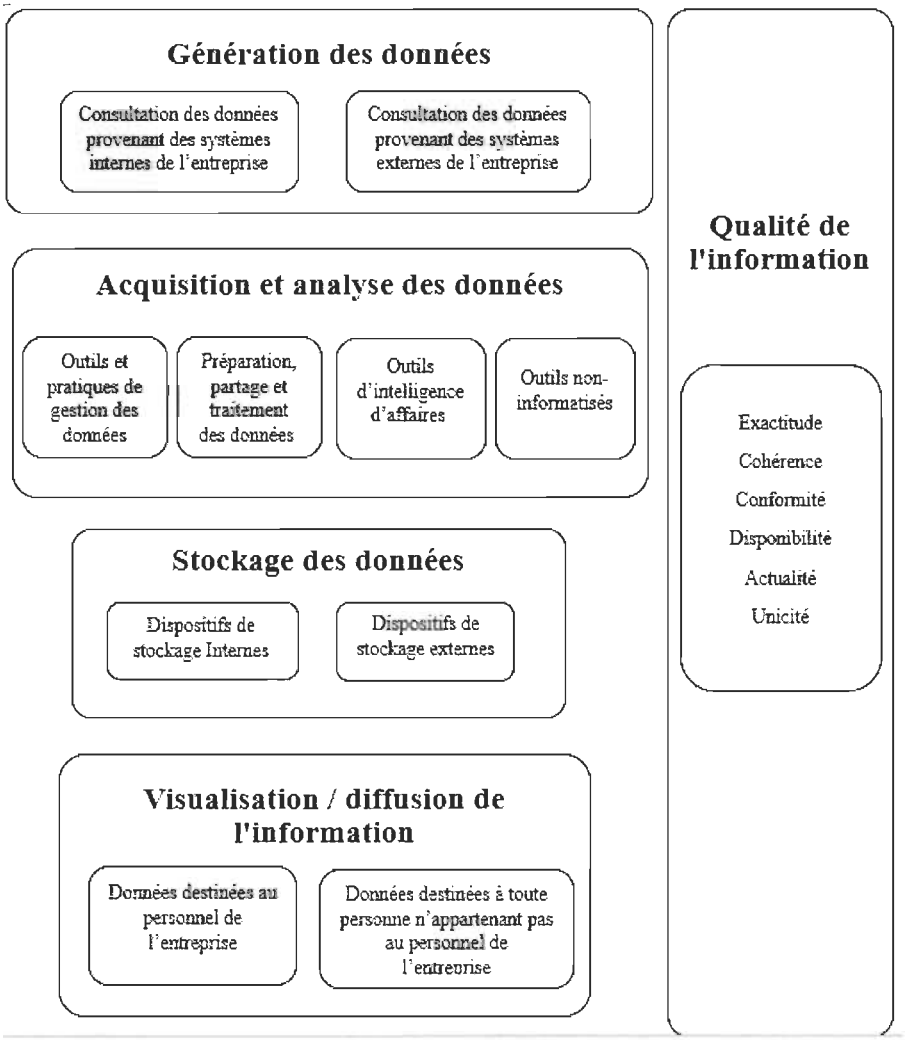
Quant à la Figure 8 ci-dessous, elle reprend le cadre conceptuel précédemment présenté, mais une fois finalisé. Elle illustre ainsi comment s'organise et s'enchaîne le contenu des quatre activités du traitement de l'information qui sont désormais : *la génération, l'acquisition et l'analyse des données, le stockage des données* et *la visualisation / diffusion de l'information.* Elle rappelle, en outre, les six caractéristiques (ou critères) retenues, après l'analyse de factorisation, pour qualifier *la qualité de*

[20] Adapté de Coleman et al. (2016), Gao et al. (2016), Hu et al. (2014), Merino et al. (2016), Rafique et al. (2012), Taleb et al. (2015) et Taleb et al. (2018).

l'information dans cette étude et qui sont : l'exactitude, la cohérence, la conformité, la disponibilité, l'actualité et l'unicité.

Figure 8

Cadre conceptuel final du traitement de l'information dans les PME à l'ère des données massives[21]



[21] Adapté de Coleman et al. (2016), Gao et al. (2016), Hu et al. (2014), Merino et al. (2016), Rafique et al. (2012), Taleb et al. (2015) et Taleb et al. (2018).

4.3 ANALYSE CONFIGURATIONNELLE FSQCA

4.3.1 Calibrage et analyse des conditions nécessaires

Pour le calibrage de nos variables, nous avons défini les points d'appartenance de chacune des variables en utilisant des percentiles prenant comme point d'appartenance totale le 80^{e} percentile des valeurs obtenues. Ici, le point de croisement est le 50^{e} percentile et la non-appartenance totale est le 20^{e} percentile. Les mêmes seuils sont utilisés pour les résultats souhaités (présence d'une haute qualité de l'information, présence d'une qualité moyenne de l'information et absence de la qualité de l'information). Le Tableau 8 présente les statistiques descriptives (moyenne, médiane, écart-type, minimum, maximum) et le calibrage de chaque variable. Plus précisément, les valeurs des trois premières colonnes représentent les seuils d'appartenance ayant été définis. Une fois que les seuils sont définis, nous utilisons le logiciel fsQCA v3.0 pour convertir les résultats de chaque variable en ensembles flous avec des valeurs entre 0 et 1 (de Guinea et Raymond, 2020; Liu et al., 2017; Ragin et Davey, 2014; Ragin, 2009; Thiem, 2014).

Tableau 8

Statistiques descriptives et calibrage des variables

	Calibrage			Moy.	Méd.	É-T	Min	Max
	Non-appartenance totale	Point de croisement	Appartenance totale					
Consultation des données provenant des systèmes internes de l'entreprise (CDI)	1,60	2,29	3,26	2,45	2,29	0,81	1,14	4,43
Consultation des données provenant des systèmes externes de l'entreprise (CDE)	2,00	2,83	3,63	2,85	2,83	0,94	1,00	4,67
Outils et pratiques de gestion des données (PGD)	1,43	2,67	3,71	2,62	2,67	1,08	1,00	4,86
Préparation, partage et traitement des données (PTD)	1,78	2,75	3,50	2,68	2,75	0,92	1,00	4,88
Outils d'intelligence d'affaires (OIA)	1,00	1,20	2,07	1,62	1,20	0,81	1,00	4,00
Outils non informatisés (ONI)	1,50	2,50	3,50	2,65	2,50	1,14	1,00	5,00
Dispositifs de stockage internes (DSI)	2,33	3,33	4,07	3,15	3,33	0,98	1,00	5,00
Dispositifs de stockage externes (DES)	1,67	2,67	3,67	2,68	2,67	0,95	1,00	4,33
Données destinées au personnel de l'entreprise (VII)	1,67	3,13	4,40	3,09	3,13	1,24	1,00	5,00
Données destinées à toute personne n'appartenant pas au personnel de l'entreprise (VIE)	1,84	2,38	3,20	2,53	2,38	0,90	1,00	5,00
Qualité de l'information (QI)	3,67	4,00	4,61	4,09	4,00	0,54	2,83	5,00

N=40. Points de coupure du calibrage : appartenance totale = 0.8; point de croisement = médiane (0.5) ; non-appartenance totale = 0.2.

Ensuite, selon l'analyse des conditions nécessaires représentée dans le Tableau 9 qui suit, le score de cohérence de chacune des conditions de notre étude varie entre 0,46 et 0,62 pour la présence de la qualité de l'information (QI) et entre 0,52 et 0,71 pour l'absence de la qualité de l'information (~QI).

Tableau 9

Analyse des conditions nécessaires

	Qualité de l'information		~ Qualité de l'information	
	Cohérence	Couverture	Cohérence	Couverture
Consultation des données provenant des systèmes internes de l'entreprise (CDI)	0,62	0,65	0,52	0,46
Consultation des données provenant des systèmes externes de l'entreprise (CDE)	0,54	0,58	0,59	0,54
Outils et pratiques de gestion des données (PGD)	0,51	0,55	0,71	0,64
Préparation, partage et traitement des données (PTD)	0,61	0,67	0,56	0,52
Outils d'intelligence d'affaires (OIA)	0,58	0,61	0,59	0,52
Outils non informatisés (ONI)	0,57	0,58	0,64	0,54
Dispositifs de stockage internes (DSI)	0,46	0,53	0,65	0,63
Dispositifs de stockage externes (DES)	0,59	0,63	0,56	0,50
Données destinées au personnel de l'entreprise (VII)	0,53	0,56	0,70	0,62
Données destinées à toute personne n'appartenant pas au personnel de l'entreprise (VIE)	0,53	0,58	0,61	0,57

Ces résultats révèlent que toutes les conditions ont une cohérence de moins de 0,90 (Ragin, 2009) pour les deux résultats (QI et ~QI). Cela signifie qu'aucune de ces conditions n'est nécessaire pour assurer la qualité de l'information ou son absence.

4.3.2 Analyses des conditions suffisantes

Pour effectuer les analyses des conditions suffisantes dans le logiciel fsQCA v3.0, nous définissons comme critères une fréquence de 1 cas et une cohérence de 0,80 pour les deux résultats : la présence d'une qualité moyenne de l'information et l'absence de la qualité de l'information. Pour le résultat d'une haute qualité de l'information, la fréquence est de 1 cas et la cohérence est de 0,90.

4.3.2.1 Présence d'une haute QI des données massives dans les PME

La combinaison configurationnelle des conditions de notre recherche indique deux configurations de la présence d'une haute qualité de l'information. Celles-ci sont présentées dans le Tableau 10 ci-dessous. La solution a une cohérence élevée de 0,89 et une couverture de 0,20. Cela indique que les résultats sont robustes. La couverture globale indique que 20% de nos PME participantes sont couvertes par ces deux configurations.

Tableau 10

Configurations suffisantes pour obtenir une haute qualité de l'information

Configurations / Conditions	Haute qualité de l'information	
	HQI1	**HQI2**
Consultation des données provenant des systèmes internes de l'entreprise (CDI)		
Consultation des données provenant des systèmes externes de l'entreprise (CDE)		
Outils et pratiques de gestion des données (PGD)		⊗
Préparation, partage et traitement des données (PTD)		
Outils d'intelligence d'affaires (OIA)		
Outils non informatisés (ONI)	•	•
Dispositifs de stockage internes (DSI)	⊗	
Dispositifs de stockage externes (DSE)	●	●
Données destinées au personnel de l'entreprise (VII)	⊗	⊗
Données destinées à toute personne n'appartenant pas au personnel de l'entreprise (VIE)	●	●
Couverture brute	0,18	0,18
Couverture unique	0,02	0,02
Cohérence	0,93	0,89
Couverture globale	0,20	
Cohérence globale	0,89	

Note: ● = condition causale principale (présente); ⊗ = condition causale principale (absente); • = condition causale périphérique (présente); ⊗ = condition causale périphérique (absente). Les espaces vides indiquent que la condition peut être présente ou absente sans avoir d'impact sur le résultat.

La première configuration (HQI1 : cohérence=0,93; couverture=0,18) se caractérise par la présence d'une utilisation importante des dispositifs de stockage externes, la présence d'une forte diffusion des données destinées à toute personne

n'appartenant pas au personnel de l'entreprise, mais une absence d'utilisation des dispositifs de stockage internes et de diffusion des données destinées au personnel de l'entreprise qui sont des conditions centrales ainsi que l'utilisation des outils non informatisés qui est une condition périphérique. La deuxième configuration (HQI2 : cohérence=0,89; couverture=0,18) se caractérise par la présence d'une utilisation importante des dispositifs de stockage externes, la présence d'une forte diffusion des données destinées à toute personne n'appartenant pas au personnel de l'entreprise, mais une absence d'utilisation des outils et pratiques de gestion des données, ainsi que de diffusion des données destinées au personnel de l'entreprise, qui sont des conditions centrales ainsi que la présence de l'utilisation des outils non informatisés qui est une condition périphérique.

4.3.2.2 Présence de la QI des données massives dans les PME

La combinaison configurationnelle des conditions de notre recherche a résulté sur huit configurations concernant la présence d'une qualité moyenne de l'information. Celles-ci sont présentées au Tableau 11. La solution possède une cohérence globale de 0,79 et une couverture globale de 0,50. Cela indique que les résultats sont robustes. La couverture globale indique que 50% de nos PME participantes sont couvertes par ces huit configurations.

Tableau 11

Configurations suffisantes pour une qualité moyenne de l'information

Configurations / Conditions	Qualité de l'information							
	QI1	QI2	QI3	QI4	QI5	QI6	QI7	QI8
Consultation des données provenant des systèmes internes de l'entreprise (CDI)	●	●	•	•				●
Consultation des données provenant des systèmes externes de l'entreprise (CDE)								●
Outils et pratiques de gestion des données (PGD)		⊗		⊗			⊗	
Préparation, partage et traitement des données (PTD)								●
Outils d'intelligence d'affaires (OIA)	•	•						●
Outils non informatisés (ONI)			•	•	•	•	•	●
Dispositifs de stockage internes (DSI)	•	•	•	•		⊗		
Dispositifs de stockage externes (DSE)					•	●	●	●
Données destinées au personnel de l'entreprise (VII)	⊗		⊗		⊗			
Données destinées à toute personne n'appartenant pas au personnel de l'entreprise (VIE)					●	•	●	●
Couverture brute	0,21	0,16	0,22	0,21	0,22	0,21	0,21	0,21
Couverture unique	0,05	0,01	0,00	0,00	0,00	0,00	0,00	0,05
Cohérence	0,78	0,88	0,83	0,88	0,82	0,85	0,82	0,79
Couverture globale	0,50							
Cohérence globale	0,79							

Note: ● = condition causale principale (présente); ⊗ = condition causale principale (absente); • = condition causale périphérique (présente); ⊗ = condition causale périphérique (absente). Les espaces vides indiquent que la condition peut être présente ou absente sans avoir d'impact sur le résultat.

La première configuration (QI1 : cohérence=0,78; couverture=0,21) se caractérise par la présence d'une consultation importante des données provenant des systèmes internes de l'entreprise, mais une absence de diffusion des données destinées au personnel de l'entreprise qui sont des conditions centrales. De plus, la présence de l'utilisation des outils d'intelligence d'affaires avec la présence de l'utilisation des dispositifs de stockage internes sont des conditions périphériques. La deuxième configuration (QI2 : cohérence=0,88; couverture=0,16) se caractérise par la présence d'une consultation importante des données provenant des systèmes internes de

l'entreprise avec une absence des outils et pratiques de gestion des données qui sont des conditions centrales. De plus, la présence de l'utilisation des outils d'intelligence d'affaires avec la présence de l'utilisation des dispositifs de stockage internes sont des conditions périphériques. La troisième configuration (QI3 : cohérence=0,83; couverture=0,22) se caractérise par une présence de la consultation des données provenant des systèmes internes de l'entreprise, la présence de l'utilisation des outils non informatisés, la présence de l'utilisation des dispositifs de stockage internes, mais une absence de diffusion des données destinées au personnel de l'entreprise qui sont des conditions périphériques. Cette configuration ne contient aucune condition centrale. La quatrième configuration (QI4 : cohérence=0,88; couverture=0,21) se caractérise par la présence d'une consultation des données provenant des systèmes internes de l'entreprise, la présence de l'utilisation des outils non informatisés, la présence de l'utilisation des dispositifs de stockage internes, mais une absence de l'utilisation des outils et pratiques de gestion des données qui sont des conditions périphériques. Cette configuration ne contient aucune condition centrale. La cinquième configuration (QI5 : cohérence=0,82; couverture=0,22) se caractérise par la présence d'une forte diffusion des données destinées à toute personne n'appartenant pas au personnel de l'entreprise, mais une absence de diffusion des données destinées au personnel de l'entreprise qui sont des conditions centrales. De plus, la présence de l'utilisation des outils non informatisés avec la présence de l'utilisation des dispositifs de stockage externes sont des conditions périphériques. La sixième configuration (QI6 : cohérence=0,85; couverture=0,21) se caractérise par la présence d'une forte utilisation des dispositifs de stockage externes et une absence d'utilisation des dispositifs de stockage internes qui sont des conditions centrales. De plus, la présence de l'utilisation des outils non informatisés avec la présence de diffusion des données destinées à toute personne n'appartenant pas au personnel de l'entreprise sont des conditions périphériques. La septième configuration (QI7 : cohérence=0,82; couverture=0,21) se caractérise par la présence d'une utilisation importante des dispositifs de stockage externes avec une présence d'une forte diffusion des données destinées à toute personne n'appartenant pas au personnel de l'entreprise,

mais une absence d'outils et pratiques de gestion des données qui sont des conditions centrales ainsi que la présence des outils non informatisés qui est une condition périphérique. Dans la huitième configuration (QI8 : cohérence=0,79; couverture=0,21), toutes les conditions sont centrales et présentes, sauf les outils et pratiques de gestion des données, les dispositifs de stockage internes et la diffusion des données destinées au personnel de l'entreprise qui sont des conditions immatérielles, ce qui signifie qu'elles n'ont pas d'impact sur la présence, de même que sur l'absence de qualité de l'information. Les conditions centrales sont : une consultation importante des données provenant des systèmes internes de l'entreprise, une consultation importante des données provenant des systèmes externes de l'entreprise, une forte préparation, partage et traitement des données, une forte utilisation des outils d'intelligence d'affaires, une utilisation importante des outils non informatisés, une utilisation importante des dispositifs de stockage externes et une forte diffusion des données destinées à toute personne n'appartenant pas au personnel de l'entreprise.

4.3.2.3 Absence de QI des données massives dans les PME

Enfin, la combinaison configurationnelle des conditions de notre recherche permet d'identifier neuf configurations illustrant l'absence de la qualité de l'information. Elles sont présentées dans le Tableau 12 ci-dessous. La solution a une cohérence de 0,82 et une couverture de 0,67. Cela indique que les résultats sont robustes. La couverture globale indique que 67% de nos PME participantes sont couvertes par ces neuf configurations.

Tableau 12

Configurations suffisantes pour l'absence de la qualité de l'information

Configurations / Conditions	Absence de qualité de l'information								
	~ QI1	~ QI2	~ QI3	~ QI4	~ QI5	~ QI6	~ QI7	~ QI8	~ QI9
Consultation des données provenant des systèmes internes de l'entreprise (CDI)		⊗				•		•	•
Consultation des données provenant des systèmes externes de l'entreprise (CDE)			•	•			•	•	•
Outils et pratiques de gestion des données (PGD)			●	●	•		●	●	•
Préparation, partage et traitement des données (PTD)			•	•	⊗	⊗			⊗
Outils d'intelligence d'affaires (OIA)	⊗	⊗					•	•	•
Outils non informatisés (ONI)	•	•		⊗	•	•	•	⊗	
Dispositifs de stockage internes (DSI)	⊗						•	•	•
Dispositifs de stockage externes (DSE)	⊗	⊗				•	•	•	•
Données destinées au personnel de l'entreprise (VII)			⊗		●	●	⊗	•	•
Données destinées à toute personne n'appartenant pas au personnel de l'entreprise (VIE)					●	●		•	•
Couverture brute	0,24	0,27	0,22	0,28	0,23	0,19	0,17	0,14	0,14
Couverture unique	0,02	0,04	0,00	0,08	0,00	0,02	0,01	0,00	0,00
Cohérence	0,87	0,84	0,82	0,80	0,98	0,92	0,93	0,92	1,00
Couverture globale	0,67								
Cohérence globale	0,82								

Note: ● = condition causale principale (présente); ⊗ = condition causale principale (absente); • = condition causale périphérique (présente); ⊗ = condition causale périphérique (absente). Les espaces vides indiquent que la condition peut être présente ou absente sans avoir d'impact sur le résultat.

La première configuration (~QI1 : cohérence=0,87; couverture=0,24) se caractérise par une absence d'utilisation des outils d'intelligence d'affaires, des dispositifs de stockage internes, ainsi que des dispositifs de stockage externes qui sont des conditions centrales; tandis que la présence de l'utilisation des outils non informatisés est une condition périphérique. La deuxième configuration (~QI2 : cohérence=0,84; couverture=0,27) se caractérise par une absence de consultation des données provenant des systèmes internes de l'entreprise, des outils d'intelligence d'affaires et des dispositifs de stockage externes qui sont des conditions centrales; tandis que la présence de l'utilisation des outils non informatisés est une condition périphérique. La troisième

configuration (~QI3 : cohérence=0,82; couverture=0,22) se caractérise par la présence d'une utilisation importante des outils et pratiques de gestion des données, mais une absence de diffusion des données destinées au personnel de l'entreprise qui sont des conditions centrales. De plus, la présence de la consultation des données provenant des systèmes externes de l'entreprise avec la présence de la préparation, partage et traitement des données sont des conditions périphériques. La quatrième configuration (~QI4 : cohérence=0,80; couverture=0,28) se caractérise par la présence d'une utilisation importante des outils et pratiques de gestion des données, mais une absence d'outils non informatisés qui sont des conditions centrales. De plus, la présence de la consultation des données provenant des systèmes externes de l'entreprise avec la présence de la préparation, partage et traitement des données sont des conditions périphériques. La cinquième configuration (~QI5 : cohérence=0,98; couverture=0,23) se caractérise par la présence d'une diffusion importante des données destinées au personnel de l'entreprise, une diffusion importante des données destinées à toute personne n'appartenant pas au personnel de l'entreprise, mais une absence de préparation, partage et traitement des données qui sont des conditions centrales. Par ailleurs, la présence de l'utilisation des outils et pratiques de gestion des données avec la présence des outils non informatisés sont des conditions périphériques. La sixième configuration (~QI6 : cohérence=0,92; couverture=0,19) se caractérise par la présence d'une diffusion importante des données destinées au personnel de l'entreprise, des données destinées à toute personne n'appartenant pas au personnel de l'entreprise, mais une absence de préparation, partage et traitement des données qui sont des conditions centrales. Par ailleurs, la présence de la consultation des données provenant des systèmes internes de l'entreprise avec la présence des outils non informatisés et la présence des dispositifs de stockage externes sont des conditions périphériques. La septième configuration (~QI7 : cohérence=0,93; couverture=0,17) se caractérise par la présence d'une forte utilisation des outils et pratiques de gestion des données, mais une absence de diffusion des données destinées au personnel de l'entreprise qui sont des conditions centrales. Par ailleurs, la présence de la consultation des données provenant des systèmes externes de l'entreprise, des outils

d'intelligence d'affaires, des outils non informatisés, des dispositifs de stockage internes et des dispositifs de stockage externes sont des conditions périphériques. La huitième configuration (~QI8 : cohérence=0,92; couverture=0,14) se caractérise par la présence d'une forte utilisation des outils et pratiques de gestion des données, mais une absence d'outils non informatisés qui sont des conditions centrales. De plus, la présence de la consultation des données provenant des systèmes internes de l'entreprise, des données provenant des systèmes externes de l'entreprise, des outils d'intelligence d'affaires, des dispositifs de stockage internes, des dispositifs de stockage externes, de la diffusion des données destinées au personnel de l'entreprise et de la diffusion des données destinées à toute personne n'appartenant pas au personnel de l'entreprise sont des conditions périphériques. La neuvième configuration (~QI9 : cohérence=1,00; couverture=0,14) ne contient que des conditions périphériques. Ces conditions consistent en la présence de la consultation des données provenant des systèmes internes et externes de l'entreprise, des outils et pratiques de gestion des données, des outils d'intelligence d'affaires, des dispositifs de stockage internes et externes, de la diffusion des données destinées au personnel de l'entreprise et de la diffusion des données destinées à toute personne n'appartenant pas au personnel de l'entreprise. Il y a toutefois absence de préparation, partage et traitement des données.

CHAPITRE 5 – DISCUSSION

Rappelons la question de recherche qui est : Quels sont les activités et les outils de traitement des données qui permettent d'assurer la qualité de l'information au sein des PME et comment ceux-ci s'articulent-ils à l'ère des données massives ? Afin d'y répondre, nous avons identifié les configurations des activités et outils de gestion assurant la présence d'une haute qualité de l'information (deux configurations), la présence d'une qualité moyenne de l'information qui est acceptable (huit configurations), et celles menant à une absence de qualité de l'information (neuf configurations) à travers une analyse fsQCA.

Avant de présenter une interprétation des résultats obtenus par la fsQCA, il importe d'aborder le jumelage des activités *acquisition des données* et *analyse des données* en une seule activité appelée *acquisition et analyse des données*, puisqu'il s'agit en soi d'un premier résultat spécifiquement applicable au contexte des PME.

5.1 RÉSULTAT DE FUSION : ACQUISITION ET ANALYSE DES DONNÉES

L'analyse traditionnelle des données signifie utiliser des méthodes statistiques appropriées pour analyser des données massives. L'objectif est d'extraire les données utiles cachées dans un lot d'ensembles de données chaotiques afin de maximiser la valeur de ces données. L'analyse des données massives peut être considérée comme la technique d'analyse d'un type particulier de données (Chen et al., 2014). Néanmoins, notre recherche a montré qu'en contexte de PME les activités *acquisition des données* et *analyse des données* se font en même temps et qu'elles doivent être fusionnées en une seule activité que nous avons appelée *acquisition et analyse des données*. Cette activité a conséquemment résulté en quatre sous-activités. La première est l'utilisation des *outils et pratiques de gestion des données* relatifs à l'intégration des données par le recours aux

PGI et autres systèmes intégrés de gestion, les outils de traitement des données structurées et les techniques de réduction des données. La deuxième est la sous-activité *préparation, partage et traitement des données*, par exemple par l'utilisation des pratiques de nettoyage, d'amélioration et de transformation des données. De plus, il y a l'utilisation des outils collaboratifs, le partage des fichiers en ligne, l'utilisation du système de traitement des transactions ainsi que les composantes de base du PGI. La troisième sous-activité est l'utilisation des *outils d'intelligence d'affaires* sur les données internes et externes, ou l'utilisation de l'infrastructure ou logiciels-service (IaaS / SaaS). Ces activités incluent également le recours aux scénarios et prévisions des systèmes interactifs d'aide à la décision (SIAD)[22]. Enfin, la quatrième sous-activité est l'utilisation traditionnelle des *outils non informatisés* sous forme de documents papier. En d'autres mots, cela signifie que plusieurs méthodes traditionnelles d'analyse des données peuvent encore être utilisées pour l'analyse de ces données massives dont beaucoup proviennent des statistiques et de l'informatique (Chen et al., 2014).

La fusion de ces deux activités constitue en soi un résultat intéressant indiquant que les PME, contrairement aux grandes entreprises, ne font pas de distinction entre les deux activités *d'acquisition des données* et *d'analyse des données*. Cela a été confirmé par notre étude. Bien qu'il soit plausible d'attribuer cette situation à un manque possible de ressources internes, cela permet tout de même de confirmer une certaine spécificité des PME concernant leurs manières d'acquérir et d'analyser les données à l'ère des données massives. Comme l'a mentionné Coleman et al. (2016) sur les obstacles à l'exploitation des données massives, les barrières financières restent l'obstacle principal à la croissance des PME par la difficulté d'accès au financement. De plus, cette étude est allée plus loin en définissant les sous-activités de cette nouvelle activité *d'acquisition et analyse des données* dans le contexte des PME.

[22] Une liste des abréviations de nature technique est fournie au début du présent document.

5.2 QUE FAIRE ?

Rappelons la définition des conditions centrales (principales) et périphériques. Une condition est centrale (ou principale) lorsqu'il y a une forte relation causale avec le résultat d'intérêt contrairement à la condition périphérique pour laquelle la preuve d'une relation causale avec le résultat est plus faible (Fiss, 2011; Hannan et al., 1996). La condition immatérielle est une condition qui peut être présente ou absente sans modifier la relation causale entre la configuration et le résultat (de Guinea et Raymond, 2020; Ragin, 2009). Le Tableau 13 présente l'ensemble de toutes les configurations qui ont été identifiées lors des analyses effectuées.

Tableau 13

Tableau récapitulatif des configurations

	Haute qualité de l'information		Qualité de l'information								Absence de qualité de l'information								
Configurations / **Conditions**	HQI 1	HQI 2	QI1	QI2	QI3	QI4	QI5	QI6	QI7	QI8	~ QI1	~ QI2	~ QI3	~ QI4	~ QI5	~ QI6	~ QI7	~ QI8	~ QI9
Consultation des données provenant des systèmes internes de l'entreprise (CDI)			●	●	•	•				●		⊗				•		•	•
Consultation des données provenant des systèmes externes de l'entreprise (CDE)										●			•	•			•	•	•
Outils et pratiques de gestion des données (PGD)		⊗		⊗		⊘			⊗				●	●	•		●	●	•
Préparation, partage et traitement des données (PTD)										●			•	•	⊗	⊗			⊘
Outils d'intelligence d'affaires (OIA)			•	•						●	⊗	⊗					•	•	•
Outils non informatisés (ONI)	•	•			•	•	•	•	•	●	•	•		⊗	•	•	•	⊗	
Dispositifs de stockage internes (DSI)	⊗		•	•	•	•		⊗			⊗						•	•	•
Dispositifs de stockage externes (DSE)	●	●					•	●	●	●	⊗	⊗				•	•	•	•
Données destinées au personnel de l'entreprise (VII)	⊗	⊗	⊗		⊘		⊗						⊗		●	●	⊗	•	•
Données destinées à toute personne n'appartenant pas au personnel de l'entreprise (VIE)	●	●					●	•	●	●					●	●		•	•
Couverture brute	0,18	0,18	0,21	0,16	0,22	0,21	0,22	0,21	0,21	0,21	0,24	0,27	0,22	0,28	0,23	0,19	0,17	0,14	0,14
Couverture unique	0,02	0,02	0,05	0,01	0,00	0,00	0,00	0,00	0,00	0,05	0,02	0,04	0,00	0,08	0,00	0,02	0,01	0,00	0,00
Cohérence	0,93	0,89	0,78	0,88	0,83	0,88	0,82	0,85	0,82	0,79	0,87	0,84	0,82	0,80	0,98	0,92	0,93	0,92	1,00
Couverture globale	0,20		0,50								0,62								
Cohérence globale	0,89		0,79								0,87								

Note: ● = condition causale principale (présente); ⊗ = condition causale principale (absente); • = condition causale périphérique (présente); ⊘ = condition causale périphérique (absente). Les espaces vides indiquent que la condition peut être présente ou absente sans avoir d'impact sur le résultat.

Suivant les résultats de la fsQCA réalisée, nous remarquons d'abord que pour avoir une haute qualité de l'information, les PME devraient utiliser les dispositifs de stockage externes, tels que l'informatique en grille et les serveurs et bases de données externes en nuage, avec la diffusion des données destinées à toute personne n'appartenant pas au personnel de l'entreprise, par exemple via l'extranet et les affaires électroniques, qui sont des conditions centrales. D'un autre côté, ces PME ne devraient pas utiliser les outils de diffusion des données destinées au personnel de l'entreprise (absence d'une condition centrale) tels que des rapports de gestion internes informatisés qui se trouvent sur le système interne de la PME, ou des formulaires numériques (ou papier) de gestion destinée à une utilisation interne qui constitue la gestion traditionnelle de l'information. Cela peut être expliqué par le fait que l'utilisation des outils de diffusion des données destinées à toute personne n'appartenant pas au personnel de l'entreprise remplace l'utilisation de ceux destinés au personnel de l'entreprise. Ces outils favorisent la collaboration et le partenariat entre les partenaires d'affaires par exemple, via l'utilisation de l'Extranet pour le partage de l'information avec les partenaires d'affaires (Baltzan, 2018). Mais, ils intègrent aussi les activités de gestion et de partage des données en interne, et ce, via le Web avec l'Intranet (Baltzan, 2018) et les de plates-formes et services de l'informatique en nuage (Hsinchun et al., 2012). Nous remarquons aussi que les PME pourront continuer à utiliser les outils non informatisés tels que les papiers et documents physiques dans leurs activités quotidiennes, mais avec une moindre importance (condition périphérique). Privilégiant, par la même occasion, le recours plus fréquent aux outils informatisés. Les autres conditions sont immatérielles, c'est-à-dire que leur présence ou absence n'affecte pas le résultat (de Guinea et Raymond, 2020; Ragin, 2009). Dans notre cas, ce résultat est représenté par la qualité de l'information. Ces conditions immatérielles sont : la consultation des données provenant des systèmes internes et externes de l'entreprise; la préparation, partage et traitement des données; les outils d'intelligence d'affaires, les dispositifs de stockage internes (pour HQI2) et les outils et pratiques de gestion des données (pour HQI1).

En résumé, ce qui précède signifie que pour que l'entreprise ait une haute qualité de l'information, elle doit favoriser une relation de partenariat et de collaboration avec tous les acteurs de la chaîne logistique allant du fournisseur du fournisseur jusqu'au client final (en amont, en interne, et en aval) en partageant ses données avec eux. Par exemple, par l'utilisation d'un système de collaboration qui est « un ensemble d'outils fondés sur les technologies de l'information qui soutient le travail d'individus et d'équipes en facilitant le partage et la circulation de l'information » (Baltzan, 2018, p. 104). Souvent ces systèmes utilisent les plateformes Web tels que l'Intranet ou l'Extranet pour le partage des données. Plus précisément, l'intranet peut contenir tout type d'information lié à l'entreprise et son fonctionnement destinée à son personnel, comme les bulletins de paie ou les catalogues de produits. Il peut aussi être utilisé comme un moyen de communication à travers l'envoi et la réception des messages et courriels ainsi que les documents numériques entre les membres de l'organisation par la simple utilisation d'un navigateur Web. L'extranet utilise le même principe d'intranet mais destiné à des alliés stratégiques et partenaires d'affaires (les clients, les fournisseurs, etc.) (Baltzan, 2018). De plus, il y a les affaires électroniques et les plateformes et services de l'informatique en nuage. À ce propos, l'utilisation des affaires électroniques permet, à la fois, le commerce en ligne et l'échange de l'information en ligne avec les collaborateurs et partenaires d'affaires selon différents modèles d'affaires, tels que le commerce électronique interentreprises (B2B), le commerce électronique entreprise-consommateur (B2C), le commerce électronique consommateur-entreprise (C2B) et le commerce électronique inter-consommateurs (C2C). Un dernier conseil est de favoriser le stockage de ses données en externes par exemple via l'utilisation de l'informatique en nuage ou l'utilisation de l'informatique en grille. Celle-ci permet la distribution du stockage et de l'effort de calcul sur plusieurs ordinateurs interconnectés pour améliorer la performance du système d'information (Baltzan, 2018). Enfin, l'utilisation du papier en gestion reste importante, mais il reste évidemment préférable de lui donner une moindre importance par rapport aux outils informatisés.

Pour avoir une qualité moyenne de l'information, mais qui reste néanmoins acceptable, les résultats obtenus des huit autres configurations montrent plusieurs pistes pour les PME concernant le traitement de l'information à l'ère des données massives.

Premièrement, nous remarquons que dans les configurations QI1, QI2, QI3 et QI4, la consultation des données internes est toujours présente avec la présence des dispositifs de stockage internes (condition périphérique). Cette condition est centrale avec la présence des outils d'intelligence d'affaires (condition périphérique), mais elle est périphérique avec la présence des outils non informatisés (condition périphérique). Ce qui veut dire que l'entreprise consulte plus ses données internes lorsqu'elle détient des outils d'intelligence d'affaires que lorsqu'elle effectue une gestion plus traditionnelle avec des outils non informatisés tels que le papier et les documents physiques. Cela s'explique par la facilité d'accès à l'information et la praticité des outils d'intelligence d'affaires permettant d'avoir un historique de l'activité de l'entreprise, ainsi qu'une vision globale de la performance de l'entreprise plus rapidement qu'avec les outils non informatisés. Avec ces deux méthodes, l'entreprise utilise des dispositifs de stockage internes. Par exemple, par archivage de ses documents papier dans les locaux de l'entreprise pour celles qui utilisent les outils non informatisés ou l'utilisation des serveurs internes ou l'archivage numérique (disques durs, clefs USB, etc.) pour celles qui utilisent les outils d'intelligence d'affaires.

Deuxièmement, dans les configurations QI5, QI6 et QI7, la présence des dispositifs de stockage externes (condition centrale dans QI6 et QI7) est accompagnée de la diffusion des données destinées à toute personne n'appartenant pas au personnel de l'entreprise (condition centrale dans QI5 et QI7) et des outils non informatisés (conditions périphériques). De plus, les dispositifs de stockage internes sont soit absents (QI6) ou immatériels. Même chose pour la diffusion des données destinées au personnel de l'entreprise. Cela veut dire qu'il devrait y avoir une combinaison des dispositifs de stockage externes, par exemple via l'informatique en nuage ou des serveurs externes, avec

la diffusion des données destinées à toute personne n'appartenant pas au personnel de l'entreprise par exemple sur le Web via l'Extranet. Évidemment, l'utilisation du papier dans la gestion reste toujours utile dans les PME, et ce, même à l'ère des données massives.

Troisièmement, dans la configuration QI8 toutes les conditions sont présentes et centrales sauf pour les deux conditions : les dispositifs de stockage internes et la diffusion des données destinées au personnel de l'entreprise qui sont des conditions immatérielles, donc sans effet sur la qualité de l'information obtenue.

Quatrièmement, nous remarquons aussi que lorsque les dispositifs de stockage internes sont présents (QI1, QI2, QI3 et QI4), les dispositifs de stockage externes sont des conditions immatérielles n'affectant pas la qualité de l'information. Alors que lors de la présence des dispositifs de stockage externes (QI5, QI6, QI7 et QI8), les dispositifs de stockage internes sont absents (QI6) ou immatériels. Cela veut dire que les configurations sont mutuellement exclusives sur le stockage, c'est-à-dire que les entreprises utilisent, soit des dispositifs internes, soit des dispositifs externes, mais pas les deux.

Pour conclure, il existe plusieurs manières d'atteindre une qualité moyenne de l'information. D'abord, la combinaison d'une consultation importante des données internes, par exemple les données provenant des outils bureautiques, avec les outils d'intelligence d'affaires, comme le SIAD, et les dispositifs de stockage internes tels que les serveurs internes. Puis, la combinaison d'une consultation moyenne des données internes avec les outils non informatisés, comme le papier, ainsi que les dispositifs de stockage internes. Ensuite, donner une moindre importance à la diffusion des données destinées au personnel de l'entreprise comme les rapports internes de gestion informatisés diffusés sur les systèmes internes. Puis, l'utilisation des dispositifs de stockage externes, par exemple sur des serveurs externes, est accompagnée de la diffusion des données destinées à toute personne n'appartenant pas au personnel de l'entreprise, tel que sur le

Web via l'Extranet ou l'informatique en nuage. Ajouté à cela une moindre importance, voire une absence de l'utilisation de la diffusion des données destinées au personnel de l'entreprise. Par exemple, concernant les systèmes internes de la PME et l'utilisation des outils non informatisés, tel que le papier, qui reste toujours utilisé dans le cadre actuel des affaires dans les PME.

5.3 QUE DOIT-ON ÉVITER ?

Également utiles à des fins d'amélioration des pratiques de gestion, les résultats des neuf configurations menant à l'absence de qualité de l'information nous montrent ce qu'il faut éviter de faire. L'objectif étant évidemment de préserver la qualité de l'information lors des différentes activités de son traitement. À cet effet, rappelons que la qualité de l'information était définie par six critères, tels que l'exactitude, la cohérence, la conformité, la disponibilité, l'actualité et l'unicité.

D'abord, nous remarquons dans les configurations ~ QI1 et ~ QI2 qu'il y a une utilisation exclusive des outils non informatisés (condition périphérique), mais une absence des outils d'intelligence d'affaires et des dispositifs de stockage externes (conditions centrales).

Donc, la première chose à éviter est de n'utiliser que les outils non informatisés, tels que le papier et les documents physiques dans la gestion de la PME. Cela en évitant l'absence des outils d'intelligence d'affaires, par exemple le SID ou le SIAD, avec l'absence des dispositifs de stockage internes (~ QI1), comme les serveurs internes, ou les dispositifs de stockage externes, telle que l'informatique en nuage.

Dans les configurations ~ QI3 et ~ QI4, on observe la présence importante des outils et pratiques de gestion des données (condition centrale) qui sont relatifs à l'intégration des données par le recours aux PGI et autres systèmes intégrés de gestion.

Cette condition est absente ou immatérielle dans les autres configurations menant à une qualité moyenne de l'information. De plus, il y a une présence moyenne des outils de consultation des données externes (condition périphérique), tels que les outils de recherche, de présentation et d'agrégation de contenus. Par ailleurs, il y a la présence de la préparation, partage et traitement des données (condition périphérique), par exemple lors de l'utilisation des pratiques de nettoyage, d'amélioration et de transformation des données. En comparaison, ces deux dernières conditions sont immatérielles dans les configurations menant à une qualité moyenne de l'information. De plus, les outils d'intelligence d'affaires représentent une condition immatérielle (n'affectant pas la qualité de l'information). C'est également le cas concernant les outils non informatisés, tel que le papier, qui est également une condition immatérielle dans ~ QI3, mais absente dans ~ QI4 (centrale). Cela veut dire que la PME, dans ce cas précis, n'utilise ni les outils d'intelligence d'affaires ni les outils non informatisés dans la gestion de ses ressources. Par ailleurs, nous remarquons que les dispositifs de stockage internes et externes sont des conditions immatérielles. La même chose se produit concernant la diffusion des données destinées à toute personne n'appartenant pas au personnel de l'entreprise, qui est une condition immatérielle, alors que la diffusion des données destinées au personnel de l'entreprise est absente dans ~ QI3 (centrale) et immatérielle dans ~ QI4.

Donc la deuxième recommandation est, pour la gestion des données en entreprise, d'avoir des outils d'intelligence d'affaires comme le SIAD ou des outils non informatisés tel que le papier et ne pas avoir une absence des deux. En plus de cela, il faut avoir des dispositifs de stockage internes comme les serveurs internes ou les dispositifs de stockage externes comme l'informatique en nuage. Même chose pour la diffusion des données, il faut avoir une diffusion des données destinée au personnel de l'entreprise, par exemple des rapports de gestion internes informatisés qui se trouvent sur le système interne de la PME, ou une diffusion des données destinées à toute personne n'appartenant pas au personnel de l'entreprise par exemple via le Web avec l'extranet et les affaires

électroniques. Certes, ces recommandations peuvent sembler intuitives, voire basiques, mais nous croyons tout de même important de les rappeler.

Quant aux configurations ~ QI7 et ~ QI8, on observe une présence importante des outils et pratiques de gestion des données (condition centrale), qui sont relatifs à l'intégration des données par le recours aux PGI et autres systèmes intégrés de gestion, avec une présence moyenne des outils de consultation des données externes (condition périphérique), tels que les outils de recherche, de présentation et d'agrégation de contenus. De plus, il y a la présence des outils d'intelligence d'affaires (condition périphérique) comme le SIAD avec la présence des dispositifs de stockage internes, comme les serveurs internes et les dispositifs de stockage externes (condition périphérique), comme les serveurs externes ou l'informatique en nuage. Par ailleurs, il y a la présence de la diffusion des données destinées au personnel de l'entreprise (condition périphérique), par exemple des rapports de gestion internes informatisés, avec la présence de la diffusion des données destinées à toute personne n'appartenant pas au personnel de l'entreprise (condition périphérique), par exemple via le Web avec l'extranet et les affaires électroniques, dans la configuration ~ QI8. Alors que dans la configuration ~ QI7, il y a une absence de diffusion des données destinées au personnel de l'entreprise (condition centrale), alors que la diffusion des données destinées à toute personne n'appartenant pas au personnel de l'entreprise est une condition immatérielle.

Donc la troisième recommandation est de donner plus d'importance aux outils d'intelligence d'affaires, tel que le SID ou le SIAD, par rapport aux outils non informatisés, tel que le papier, et vice versa, mais ne pas donner d'importance aux deux en même temps. Même chose pour les dispositifs de stockage internes, par exemple les serveurs internes, par rapport aux dispositifs de stockage externes, tel que les serveurs externes ou l'informatique en nuage, et vice versa. Également pour les types de diffusion des données, donner plus d'importance à la diffusion des données destinées à toute personne n'appartenant pas au personnel de l'entreprise, par exemple via le Web avec

l'extranet et les affaires électroniques, par rapport à celle destinée au personnel de l'entreprise tel que les rapports de gestion internes informatisés qui se trouvent sur le système interne de la PME.

Ensuite, nous remarquons dans les configurations ~ QI5 et ~ QI6 qu'il y a la présence de la diffusion des données destinées au personnel de l'entreprise (condition centrale), par exemple des rapports de gestion internes informatisés qui se trouvent sur le système interne de la PME, avec la présence de la diffusion des données destinées à toute personne n'appartenant pas au personnel de l'entreprise (condition centrale), par exemple via le Web avec l'extranet et les affaires électroniques. De plus, il a une absence de la préparation, partage et traitement des données (condition centrale) par exemple par l'utilisation des pratiques de nettoyage, d'amélioration et de transformation des données. Par ailleurs, il y a la présence des outils non informatisés (condition périphérique) tel que le papier.

Donc, la quatrième recommandation est que la préparation, le partage et le traitement des données, par exemple par l'utilisation des pratiques de nettoyage, d'amélioration et de transformation des données, peut être une condition immatérielle sans influence sur la qualité de l'information, plutôt qu'absente au sens strict. De plus, les configurations menant à la qualité de l'information recommandent pour la diffusion des données, de donner plus d'importance à la diffusion des données destinées à toute personne n'appartenant pas au personnel de l'entreprise, par exemple via le Web avec l'extranet et les affaires électroniques, par rapport à celle sur la diffusion des données destinées au personnel de l'entreprise, par exemple des rapports de gestion internes informatisés qui se trouvent sur le système interne de la PME, qui peut être une condition immatérielle. Cela signifie qu'au sens strict, ce n'est pas la diffusion d'une information qui assure sa qualité (en amont), mais que des partenaires d'affaires sont peut-être plus susceptibles de faire des commentaires qui pourraient améliorer (éventuellement) la qualité de l'information qui est diffusée, en interne comme en externe. Par exemple, par

les commentaires des clients sur le site de l'entreprise dans les affaires électroniques ou une communication efficace et agile avec les fournisseurs dans l'Extranet. Cela permet une meilleure compréhension du besoin du client et une meilleure collaboration avec le fournisseur et ainsi un meilleur fonctionnement interne en évitant les tâches inutiles dues au manque de compréhension et de communication avec les partenaires d'affaires.

Pour conclure, la première chose à éviter est d'utiliser exclusivement les outils non informatisés tels que le papier et les documents physiques dans la gestion de la PME. La deuxièmement recommandation est d'avoir une méthode de gestion de ses données, c'est-à-dire soit par l'utilisation des outils d'intelligence d'affaires comme le système d'information d'aide à la décision (SIAD), ou par l'utilisation des outils non informatisés tel que le papier. Même chose pour les dispositifs de stockage où il faut avoir, soit les dispositifs de stockage internes comme les serveurs internes, ou les dispositifs de stockage externes comme l'informatique en nuage et pas les deux en même temps. Également pour la diffusion des données, soit par l'utilisation de la diffusion des données destinée au personnel de l'entreprise, par exemple des rapports de gestion internes informatisés qui se trouvent sur le système interne de la PME, ou l'utilisation de la diffusion des données destinées à toute personne n'appartenant pas au personnel de l'entreprise par exemple via le Web avec l'extranet et les affaires électroniques. La troisième et la quatrième recommandation est de donner de l'importance aux outils d'intelligence d'affaires par rapport aux outils non informatisés, et aux dispositifs de stockage externes par rapport aux dispositifs de stockage internes. Même chose pour la diffusion des données, favoriser la diffusion des données destinées à toute personne n'appartenant pas au personnel de l'entreprise, par exemple via le Web avec l'extranet et les affaires électroniques, à celles destinées au personnel de l'entreprise tels que les rapports de gestion internes informatisés qui se trouvent sur le système interne de la PME. Cela s'explique par le fait que les partenaires d'affaires sont plus susceptibles de faire des remarques et/ou des suggestions d'amélioration de la qualité de l'information qui est diffusée, en interne comme en externe.

CONCLUSION

L'objectif de notre étude était de décrire et comprendre le phénomène du traitement de l'information au sein des PME à l'ère des données massives. Cette recherche cherchait à répondre à la question de recherche suivante :
Quels sont les activités et les outils de traitement des données qui permettent d'assurer la qualité de l'information au sein des PME et comment ceux-ci s'articulent-ils à l'ère des données massives ?

Pour ce faire, notre étude a adopté une approche configurationnelle fondée sur l'analyse qualitative comparée par les ensembles flous (dont l'acronyme connu est fsQCA). Elle a été réalisée par l'intermédiaire d'une enquête en ligne auprès de 40 PME québécoises appartenant à différents secteurs d'activités. L'analyse fsQCA qui a été réalisée a permis d'identifier les configurations des activités et outils de gestion assurant la présence d'une haute qualité de l'information (deux configurations), la présence d'une qualité moyenne de l'information qui est acceptable (huit configurations), et celles menant à une absence de qualité de l'information (neuf configurations). Ces configurations reposaient d'abord sur *la génération des données* qui permet la consultation des données provenant des systèmes internes et externes de l'entreprise. Ensuite, *l'acquisition et l'analyse des données* à travers quatre sous-activités relatives aux outils et pratiques de gestion des données, à la préparation, le partage et traitement des données, aux outils d'intelligence d'affaires, ainsi qu'aux outils non informatisés. Puis, il y a *le stockage des données* avec les dispositifs de stockage internes et externes. Enfin, il y a *la diffusion / visualisation de l'information* à travers le partage des données et des informations destinées au personnel de l'entreprise, ainsi qu'à toute personne en lien avec l'entreprise, mais n'appartenant pas à son personnel comme les partenaires d'affaires (fournisseurs, clients, distributeurs, etc.).

En résumé, afin d'assurer une haute qualité de l'information, nos résultats suggèrent que les PME québécoises devraient aller vers une relation de partenariat et de collaboration avec les partenaires d'affaires en partageant leurs données avec eux. De plus, elles devraient stocker leurs données en externes par exemple via l'utilisation de l'informatique en nuage. L'utilisation du papier restera comme une pratique de gestion, mais en lui donnant une moindre importance par rapport aux outils informatisés. Toutefois, pour préserver la qualité de l'information, il y a des choses à éviter. Par exemple, la première chose à éviter est d'utiliser exclusivement les outils non informatisés tels que le papier et les documents physiques dans la gestion de la PME. La deuxième chose est de détenir au moins un moyen pour la gestion de ses données, soit les outils d'intelligence d'affaires ou les outils non informatisés (papier). Même chose pour les dispositifs de stockage, avoir soit les dispositifs de stockage internes comme les serveurs internes, ou les dispositifs de stockage externes comme l'informatique en nuage. De plus, favoriser l'utilisation de la diffusion des données destinées à toute personne n'appartenant pas au personnel de l'entreprise par exemple via le Web avec l'extranet et les affaires électroniques à celles destinées au personnel de l'entreprise, par exemple des rapports de gestion internes informatisés qui se trouvent sur le système interne de la PME.

Certes cette recherche a des avantages, mais elle a aussi des limites. D'abord, il y a la taille de l'échantillon qui est faible (n = 40). De plus, cette étude se concentrait que sur les PME québécoises. Donc, les résultats obtenus ne peuvent pas être généralisables à d'autres PME hors du Québec. Par ailleurs, la revue de littérature ne permettait pas d'avoir le retour d'expériences des PME québécoises, voire canadiennes, sur leur utilisation des données massives par manque de cas. En effet, la plupart des articles abordaient la relation des grandes entreprises aux données massives ou encore il s'agissait d'articles rédigés hors Canada. Enfin, il existe des limites par rapport à la méthode d'analyse des données utilisée (fsQCA) elle-même. Par exemple, les décisions concernant le calibrage de nos mesures pourraient affecter nos résultats (de Guinea et Raymond, 2020; Glaesser et Cooper, 2014). Considérant ce qui précède, comme pistes futures de recherche, il serait

alors intéressant d'étendre cette recherche sur d'autres PME se situant hors du Québec pour que les résultats soient plus robustes et ainsi obtenir une analyse de la situation plus précise et plus actuelle de l'utilisation des données massives dans les PME en général.

Notre étude permet la contribution aux connaissances de diverses manières. D'abord, pour les contributions théoriques, il y a l'adaptation du modèle conceptuel de Taleb et al. (2015, p. 2) sur le cycle de vie des données massives au contexte des PME. Cette adaptation concerne plus précisément les figures conceptuelles et configurationnelles finales du traitement de l'information dans les PME à l'ère des données massives qui ont été dressés avant les analyses fsQCA (voir Figures 7 et 8). Ces figures contiennent les activités et sous-activités par lesquelles les données massives passent pour les convertir en information exploitable et où nous proposons la fusion de l'activité *acquisition des données* et de l'activité *analyse des données* en une nouvelle activité nommée *acquisition et analyse des données*. Ce faisant, nous indiquons que les PME ne font pas de distinction entre ces deux activités, et ce, contrairement aux grandes entreprises. Conséquemment, il y a également la définition des sous-activités de chacune de ces activités dans le contexte des PME qui constitue une autre contribution théorique. Plus spécifiquement, la définition de ces activités et sous-activités permet de répondre au problème de recherche en s'intéressant au côté managérial des données massives dans les PME, ayant pour le but d'assurer la qualité de l'information. Une autre contribution théorique est qu'il existe plusieurs manières d'atteindre une haute ou une qualité moyenne de l'information. Cela grâce à l'équifinalité de l'approche configurationnelle de la fsQCA.

Quant aux contributions pratiques, nos résultats permettent d'offrir quelques recommandations aux PME et leurs dirigeants concernant les configurations d'activités et d'outils de traitement des données qui assurent une haute ou une qualité moyenne de l'information. Enfin, une dernière contribution pratique est d'offrir quelques conseils concernant les pratiques de gestion à éviter (ou améliorer), puisqu'elles mènent vers une faible qualité de l'information.

RÉFÉRENCES

Baltzan, P. (2018). *Gestion des Technologies d'Affaires* (Chenelière éd.). Montréal (Québec): Chenelière Éducation.

Bansal, S. K. (2014). *Towards a semantic extract-transform-load (ETL) framework for big data integration.* Communication présentée 2014 IEEE International Congress on Big Data.

Belanger, F., et Van Slyke, C. (2011). *Information systems for business: an experiential approach*: John Wiley & Sons.

Bellavance, F., et Labrie, F. (2017). Bienvenue à l'ère du gestionnaire décodeur. *Gestion, 42*(1), 38-46. http://dx.doi.org/10.3917/riges.421.0038

Bi, Z., et Cochran, D. (2014). Big data analytics with applications. *Journal of Management Analytics, 1*(4), 249-265. http://dx.doi.org/10.1080/23270012.2014.992985

Boswell, T., et Brown, C. (1999). The scope of general theory: Methods for linking deductive and inductive comparative history. *Sociological Methods & Research, 28*(2), 154-185.

Bourque, J., et al. (2006). Évaluation de l'utilisation et de la présentation des résultats d'analyses factorielles et d'analyses en composantes principales en éducation. *Revue des sciences de l'éducation, 32*(2), 325-344.

Brown, B., et al. (2011). Are you ready for the era of 'big data'. *McKinsey Quarterly, 4*(1), 24-35.

Bryant, F. B., et Yarnold, P. R. (1995). Principal-components analysis and exploratory and confirmatory factor analysis.

Chaudhuri, S., et al. (2011). An overview of business intelligence technology. *Commun. ACM, 54*(8), 88-98. http://dx.doi.org/10.1145/1978542.1978562

Chen, H., et al. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly, 36*(4), 1165-1188. http://dx.doi.org/10.2307/41703503

Chen, M., et al. (2014). Big data: A survey. *Mobile networks and applications, 19*(2), 171-209.

Chiang, F., et Miller, R. J. (2008). Discovering data quality rules. *Proceedings of the VLDB Endowment, 1*(1), 1166-1177.

Coleman, S., et al. (2016). How Can SMEs Benefit from Big Data? Challenges and a Path Forward. *Quality and Reliability Engineering International, 32*(6), 2151-2164. http://dx.doi.org/10.1002/qre.2008

Conway, J. M., et Huffcutt, A. I. (2003). A review and evaluation of exploratory factor analysis practices in organizational research. *Organizational research methods, 6*(2), 147-168.

Cooper, B., et Glaesser, J. (2016). Exploring the robustness of set theoretic findings from a large n fsQCA: an illustration from the sociology of education. *International Journal of Social Research Methodology, 19*(4), 445-459.

Davenport, T. H., et al. (2012). How'big data'is different.

de Guinea, A. O., et Raymond, L. (2020). Enabling innovation in the face of uncertainty through IT ambidexterity: A fuzzy set qualitative comparative analysis of industrial service SMEs. *International Journal of Information Management, 50*, 244-260.

Delery, J. E., et Doty, D. H. (1996). Modes of theorizing in strategic human resource management: Tests of universalistic, contingency, and configurational performance predictions. *Academy of Management Journal, 39*(4), 802-835.

Depeyre, C., et Vergne, J.-P. (2018). L'analyse qualitative comparative (QCA).

DeVon, H. A., et al. (2007). A psychometric toolbox for testing validity and reliability. *Journal of Nursing scholarship, 39*(2), 155-164.

Dong, X. L., et Srivastava, D. (2013). *Big data integration.* Communication présentée 2013 IEEE 29th international conference on data engineering (ICDE).

Drass, K., et Ragin, C. (1999). fs/QCA: Fuzzy set/qualitative comparative analysis. *Evanston, IL: Institute for Policy Research, Northwestern University*.

El Sawy, O. A., et al. (2010). Research commentary—seeking the configurations of digital ecodynamics: It takes three to tango. *Information systems research, 21*(4), 835-848.

Fabrigar, L. R., et al. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods, 4*(3), 272.

Farrell, A. M. (2010). Insufficient discriminant validity: A comment on Bove, Pervan, Beatty, and Shiu (2009). *Journal of Business Research, 63*(3), 324-327.

Filion, L. J., et Allali, B. (2007). *Management des PME : de la création à la croissance.* Saint-Laurent, Québec: Éditions du Renouveau pédagogique.

Fiss, P. C. (2007). A set-theoretic approach to organizational configurations. *Academy of management review, 32*(4), 1180-1198.

Fiss, P. C. (2011). Building better causal theories: A fuzzy set approach to typologies in organization research. *Academy of Management Journal, 54*(2), 393-420.

Fornell, C., et Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of marketing research, 18*(1), 39-50.

Fortin, M.-F., et Gagnon, J. (2016). *Fondements et étapes du processus de recherche: méthodes quantitatives et qualitatives*. Montréal (Québec) Canada: Chenelière éducation.

Fürber, C., et Hepp, M. (2011). *Towards a vocabulary for data quality management in semantic web architectures.* Communication présentée Proceedings of the 1st International Workshop on Linked Web Data Management.

Galunic, D. C., et Eisenhardt, K. M. (1994). Renewing the strategy-structure-performance paradigm. *Research in organizational behavior, 16*, 215-215.

Gao, J., et al. (2016). *Big Data Validation and Quality Assurance--Issuses, Challenges, and Needs.* Communication présentée 2016 IEEE symposium on service-oriented system engineering (SOSE).

George, D., et Mallery, P. (2016). SPSS for Windows Step by Step: A Simple Guide and Reference. 11.0 update , 2003: Boston: Allyn & Bacon.

Glaesser, J., et Cooper, B. (2014). Exploring the consequences of a recalibration of causal conditions when assessing sufficiency with fuzzy set QCA. *International Journal of Social Research Methodology, 17*(4), 387-401. http://dx.doi.org/10.1080/13645579.2013.769782

Glasson-Cicognani, M., et Berchtold, A. (2010). *Imputation des données manquantes: Comparaison de différentes approches.* Communication présentée 42èmes Journées de Statistique.

Gliem, J. A., et Gliem, R. R. (2003). *Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales.* Communication présentée.

Gouvernement du Canada. (2020). Principales statistiques relatives aux petites entreprises. Repéré à https://www.ic.gc.ca/eic/site/061.nsf/fra/h_03126.html

Greckhamer, T., et al. (2013). Chapter 3 The two QCAs: From a small-N to a large-N set theoretic approach. *Configurational theory and methods in organizational research, 38*, 49-75.

Gresov, C., et Drazin, R. (1997). Equifinality: Functional equivalence in organization design. *Academy of management review, 22*(2), 403-428.

Hagen, C., et al. (2013). Big data and the creative destruction of today's business models. *Retrieved January, 5*, 2015.

Hannan, M. T., et al. (1996). Inertia and change in the early years: Employment relations in young, high technology firms. *Industrial and Corporate Change, 5*(2), 503-536.

Henson, R. K., et al. (2001). Reporting Practice and Use of Exploratory Factor Analysis in Educational Research Journals.

Holley, K., et al. (2014). *Enrichment patterns for big data.* Communication présentée 2014 IEEE International Congress on Big Data.

Hsinchun, C., et al. (2012). BUSINESS INTELLIGENCE AND ANALYTICS: FROM BIG DATA TO BIG IMPACT. *MIS Quarterly, 36*(4), 1165-1188.

Hu, H., et al. (2014). Toward scalable systems for big data analytics: A technology tutorial. *IEEE access, 2*, 652-687.

Julien, P.-A., et Morin, M. (2011). *Mondialisation de l'économie et PME québécoises*: PUQ.

Karoui, M., et al. (2014). *Karoui M., Devauchelle G., Dudezert A. (2014), « Big Data : Mise en perspective et enjeux pour les entreprises », N° Spécial "Big Data", Revue Ingénierie des Systèmes d'Information, 19 (3), 73-92 Hermès* (Vol. 19).

Kieffer, K. M. (1998). Orthogonal versus Oblique Factor Rotation: A Review of the Literature regarding the Pros and Cons.

Kiron, D., et al. (2012). Innovating with analytics. *MIT Sloan Management Review, 54*(1), 47.

Krishnan, K. (2013). *Data Warehousing in the Age of Big Data.* Amsterdam: Morgan Kaufmann.

Labrinidis, A., et Jagadish, H. V. (2012). Challenges and opportunities with big data. *Proc. VLDB Endow., 5*(12), 2032-2033. http://dx.doi.org/10.14778/2367502.2367572

LaValle, S., et al. (2011). Big data, analytics and the path from insights to value. *MIT Sloan Management Review, 52*(2), 21-32.

Lee, Y. W., et al. (2002). AIMQ: a methodology for information quality assessment. *Information & Management, 40*(2), 133-146. http://dx.doi.org/https://doi.org/10.1016/S0378-7206(02)00043-5

Lenzerini, M. (2002). *Data integration: A theoretical perspective.* Communication présentée Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology.*

Liu, Y., et al. (2017). Applying configurational analysis to IS behavioural research: a methodological alternative for modelling combinatorial complexities. *Information Systems Journal, 27*(1), 59-89.

Loshin, D. (2001). *Enterprise knowledge management: The data quality approach*: Morgan Kaufmann.

MacInnes, B. (2013). Help SMEs benefit from big data on a small scale. *MicroScope.*

Maier, M., et al. (2013). Towards a big data reference architecture. *University of Eindhoven.*

Maletic, J. I., et Marcus, A. (2000). *Data Cleansing: Beyond Integrity Analysis.* Communication présentée Iq.

McAfee, A., et al. (2012). Big data: the management revolution. *Harvard Business Review, 90*(10), 60-68.

Mell, P., et Grance, T. (2011). The NIST definition of cloud computing.

Mendel, J. M., et Korjani, M. M. (2012). Charles Ragin's Fuzzy Set Qualitative Comparative Analysis (fsQCA) used for linguistic summarizations. *Information Sciences, 202*, 1-23. http://dx.doi.org/https://doi.org/10.1016/j.ins.2012.02.039

Merino, J., et al. (2016). A Data Quality in Use model for Big Data. *Future Generation Computer Systems, 63*, 123-130. http://dx.doi.org/https://doi.org/10.1016/j.future.2015.11.024

Meyer, A. D., et al. (1993). Configurational approaches to organizational analysis. *Academy of Management Journal, 36*(6), 1175-1195.

Mikalef, P., et Pateli, A. (2017). Information technology-enabled dynamic capabilities and their indirect effect on competitive performance: Findings from PLS-SEM and fsQCA. *Journal of Business Research, 70*, 1-16.

Mikalef, P., et al. (2015). Purchasing alignment under multiple contingencies: a configuration theory approach. *Industrial Management & Data Systems.*

Miller, H. G., et Mork, P. (2013). From data to decisions: a value chain for big data. *It Professional*(1), 57-59.

Nunnally, J. C. (1994). *Psychometric theory 3E*: Tata McGraw-hill education.

O'reilly, T. (2007). What is Web 2.0: Design patterns and business models for the next generation of software. *Communications & strategies*(1), 17.

Ohlhorst, F. (2012). Big Data Analytics: Turning Big Data into Big Money. Dans F. Ohlhorst (Éd.), *Big Data Analytics: Turning Big Data into Big Money*.

Oliveira, P., et al. (2005). *A formal definition of data quality problems.* Communication présentée ICIQ.

Olszak, C., et Zurada, J. (2019). *Big Data-driven Value Creation for Organizations.* Communication présentée Proceedings of the 52nd Hawaii International Conference on System Sciences.

Osborne, J. (2002). Notes on the use of data transformations. *Practical assessment, research, and evaluation, 8*(1), 6.

Pappas, I., et al. (2017). Value co-creation and trust in social commerce: An fsQCA approach.

Pappas, I. O., et al. (2019). Fuzzy set analysis as a means to understand users of 21st-century learning systems: The case of mobile learning and reflections on learning analytics research. *Computers in Human Behavior, 92*, 646-659. http://dx.doi.org/https://doi.org/10.1016/j.chb.2017.10.010

Park, H. S., et al. (2002). The use of exploratory factor analysis and principal components analysis in communication research. *Human Communication Research, 28*(4), 562-577.

Pett, M. A., et al. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*: sage.

Pohlmann, J. T. (2004). Use and interpretation of factor analysis in The Journal of Educational Research: 1992-2002. *the Journal of Educational research, 98*(1), 14-23.

Preez, D. D. (2014). Big data for small business. *Raconteur.[Online] Retrieved from: http://raconteur. net/technoloev/bie-data-for-small-business [Accessed on June 25, 2017]*.

Rafique, I., et al. (2012). Information quality evaluation framework: Extending ISO 25012 data quality model. *World Academy of Science, Engineering and Technology, 65*, 523-528.

Ragin, C., et Davey, S. (2014). fs/QCA [Computer Programme]. *Version [2.5/3.0]. Irvine: University of California*.

Ragin, C. C. (2000). *Fuzzy-set social science*: University of Chicago Press.

Ragin, C. C. (2005). From fuzzy sets to crisp truth tables. *Comparative Methods for the Advancement of Systematic cross-case analysis and Small-N studies (COMPASS)*.

Ragin, C. C. (2006). Set relations in social research: Evaluating their consistency and coverage. *Political analysis*, 291-310.

Ragin, C. C. (2009). *Redesigning social inquiry: Fuzzy sets and beyond*: University of Chicago Press.

Ragin, C. C. (2014). *The comparative method: Moving beyond qualitative and quantitative strategies*: Univ of California Press.

Rahm, E., et Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull., 23*(4), 3-13.

Rihoux, B., et Marx, A. (2013). QCA, 25 years after "The comparative method" mapping, challenges, and innovations—Mini-Symposium. *Political Research Quarterly, 66*(1), 167-235.

Rihoux, B., et Ragin, C. C. (2008). *Configurational comparative methods: Qualitative comparative analysis (QCA) and related techniques* (Vol. 51): Sage Publications.

Rising, C., et al. (2014). Is Big data too big for SMEs? Leading trends in information technology. *Stanford University (Online). https://web. stanford. edu/class/msande238/projects/2014/GainIT. pdf. Access April, 4*, 2019.

Rubin, D. B. (1987). *Statistical analysis with missing data*: Wiley.

Russell, D. W. (2002). In search of underlying dimensions: The use (and abuse) of factor analysis in Personality and Social Psychology Bulletin. *Personality and social psychology bulletin, 28*(12), 1629-1646.

Salles, M. (2006). Decision making in SMEs and information requirements for competitive intelligence. *Production Planning & Control, 17*(3), 229-237. http://dx.doi.org/10.1080/09537280500285367

Salomon, D. (2004). *Data compression: the complete reference*: Springer Science & Business Media.

Schneider, C. Q., et Wagemann, C. (2012). *Set-theoretic methods for the social sciences: A guide to qualitative comparative analysis*: Cambridge University Press.

Schneider, M. R., et al. (2010). Mapping the institutional capital of high-tech firms: A fuzzy-set analysis of capitalist variety and export performance. *Journal of International Business Studies, 41*(2), 246-266.

Sen, D., et al. (2016). An Overview of Big Data for Growth in SMEs. *Procedia - Social and Behavioral Sciences, 235*, 159-167. http://dx.doi.org/https://doi.org/10.1016/j.sbspro.2016.11.011

Serhani, M. A., et al. (2016). *An hybrid approach to quality evaluation across big data value chain.* Communication présentée 2016 IEEE International Congress on Big Data (BigData Congress).

Sidi, F., et al. (2012). *Data quality: A survey of data quality dimensions.* Communication présentée 2012 International Conference on Information Retrieval & Knowledge Management.

Stamper, D. A. (1988). *Business data communications*: Benjamin-Cummings Publishing Co., Inc.

Stonebraker, M., et al. (2010). MapReduce and parallel DBMSs: friends or foes? *Communications of the ACM, 53*(1), 64-71.

Symes, P. D. (2004). *Digital video compression*: McGraw Hill Professional.

Tabachnick, B. G., et al. (2007). *Using multivariate statistics* (Vol. 5): Pearson Boston, MA.

Taleb, I., et al. (2015). *Big data pre-processing: A quality framework.* Communication présentée 2015 IEEE international congress on big data.

Taleb, I., et al. (2018). *Big data quality: A survey.* Communication présentée 2018 IEEE International Congress on Big Data (BigData Congress).

Thiem, A. (2014). Membership function sensitivity of descriptive statistics in fuzzy-set relations. *International Journal of Social Research Methodology, 17*(6), 625-642.

Waltz, C. F., et al. (2010). *Measurement in nursing and health research*: Springer publishing company.

Wang Shouhong, S. (2020). Big data for small and medium-sized enterprises (SME): a knowledge management model. *Journal of Knowledge Management, 24*(4), 881.

Woodside, A. G. (2013). Moving beyond multiple regression analysis to algorithms: Calling for adoption of a paradigm shift from symmetric to asymmetric thinking in data analysis and crafting theory: Elsevier.

Yang, G.-Z., et al. (2014). Multi-sensor fusion *Body sensor networks* (pp. 301-354): Springer.

Yeh, P. Z., et Puri, C. A. (2010). *An efficient and robust approach for discovering data quality rules.* Communication présentée 2010 22nd IEEE International Conference on Tools with Artificial Intelligence.

Zhang, Y., et al. (2002). *Novelty and redundancy detection in adaptive filtering.* Communication présentée Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval.

Zikopoulos, P., et Eaton, C. (2011). *Understanding big data: Analytics for enterprise class hadoop and streaming data*: McGraw-Hill Osborne Media.

ANNEXE A
CERTIFICAT D'ÉTHIQUE

3250

UQTR
Savoir.
Surprendre.

CERTIFICAT D'ÉTHIQUE DE LA RECHERCHE AVEC DES ÊTRES HUMAINS

En vertu du mandat qui lui a été confié par l'Université, le Comité d'éthique de la recherche avec des êtres humains a analysé et approuvé pour certification éthique le protocole de recherche suivant :

Titre : **La PME à l'ère des données massives (Big Data)**

Chercheur(s) : Nada Fejjar
Département de marketing et systèmes d'information

Organisme(s) : Aucun financement

N° DU CERTIFICAT : **CER-19-263-07.16**

PÉRIODE DE VALIDITÉ : **Du 15 janvier 2020** **au 15 janvier 2021**

En acceptant le certificat éthique, le chercheur s'engage à :

- Aviser le CER par écrit des changements apportés à son protocole de recherche avant leur entrée en vigueur;
- Procéder au renouvellement annuel du certificat tant et aussi longtemps que la recherche ne sera pas terminée;
- Aviser par écrit le CER de l'abandon ou de l'interruption prématurée de la recherche;
- Faire parvenir par écrit au CER un rapport final dans le mois suivant la fin de la recherche.

Bruce Maxwell
Président du comité

Fanny Longpré
Secrétaire du comité

Décanat de la recherche et de la création **Date d'émission :** 15 janvier 2020

3250

UQTR
Savoir. Surprendre.

CERTIFICAT D'ÉTHIQUE DE LA RECHERCHE AVEC DES ÊTRES HUMAINS

En vertu du mandat qui lui a été confié par l'Université, le Comité d'éthique de la recherche avec des êtres humains a analysé et approuvé pour certification éthique le protocole de recherche suivant :

Titre : **La PME à l'ère des données massives (Big Data)**

Chercheur(s) : Nada Fejjar
Département de marketing et systèmes d'information

Organisme(s) : Aucun financement

N° DU CERTIFICAT : **CER-19-263-07.16**

PÉRIODE DE VALIDITÉ : **Du 15 janvier 2021** **au 15 janvier 2022**

En acceptant le certificat éthique, le chercheur s'engage à :

- Aviser le CER par écrit des changements apportés à son protocole de recherche avant leur entrée en vigueur;
- Procéder au renouvellement annuel du certificat tant et aussi longtemps que la recherche ne sera pas terminée;
- Aviser par écrit le CER de l'abandon ou de l'interruption prématurée de la recherche;
- Faire parvenir par écrit au CER un rapport final dans le mois suivant la fin de la recherche.

Me Richard LeBlanc
Président du comité

Fanny Longpré
Secrétaire du comité

Décanat de la recherche et de la création

Date d'émission : 15 décembre 2020

ANNEXE B

QUESTIONNAIRE DE RECHERCHE

Les PME et le traitement de l'information à l'ère des données massives (Big Data)

INSTRUCTIONS

- Ce questionnaire s'adresse au dirigeant principal et à ses collaborateurs capables de répondre à des questions qui portent sur le traitement de l'information et sa qualité au sein des petites ou moyennes entreprises (PME) québécoises. Une section d'informations d'ordre socio-démographique complète le questionnaire.
- Cette étude est destinée à des gestionnaires et non des spécialistes du domaine informatique. Veuillez répondre à chaque question au meilleur de vos connaissances en cliquant sur la case correspondant à la réponse choisie sur l'échelle proposée ou en saisissant votre réponse sur la ligne prévue à cet effet.
- Toutes les données collectées seront rendues anonymes pour les fins d'analyse. Ces données seront conservées par l'étudiante, sous format numérique, pour la durée de ses études et l'exploitation de ses données, par exemple à des fins de publications scientifiques et professionnelles. Elles seront, ensuite, détruites.
- Les seules personnes qui auront accès aux données seront par exemple la directrice de recherche, la professeure Claudia PELLETIER, les évaluateurs du mémoire de recherche et les autres personnes impliquées pour l'analyse des données, le cas échéant. Elles signeront toutes un engagement à la confidentialité.
- Les résultats de la recherche seront diffusés sous forme d'un mémoire de recherche et ne permettront pas d'identifier les participants.
- La présente recherche respecte les règles d'éthique et de déontologie de la recherche à l'Université du Québec à Trois-Rivières, tel qu'en témoigne le certificat émis sous le numéro CER-19-263-07.16.

Merci pour votre précieuse collaboration à cette recherche!

Pour toute information relative à cette recherche, vous pouvez joindre :

NADA FEJJAR
Étudiante en Maîtrise Sciences de la Gestion
Université du Québec à Trois-Rivières (UQTR)
nada.fejjar@uqtr.ca
(+1) 819-384-7368

CLAUDIA PELLETIER, DBA
Directrice de recherche
Professeure en systèmes d'information
Membre de l'Institut de recherche sur les PME (INRPME)
Université du Québec à Trois-Rivières (UQTR)
Claudia.Pelletier@uqtr.ca
(+1) 819-376-5011, poste 4271

GÉNÉRATION DES DONNÉES

C'est la phase où les données sont créées. Différentes sources de données sont responsables de création de ces données comme par exemple les capteurs utilisés pour recueillir les informations sur le climat, les appareils de surveillance, les publications sur les réseaux sociaux, etc. (traduction libre) (Taleb et al., 2018, p. 2).
L'objectif de cette section est de faire un inventaire de toutes les sources de données auxquelles l'entreprise peut accéder ou consulter.

On distingue deux types de sources, soit internes et externes :

- Les sources internes représentent toutes les données provenant des systèmes de l'entreprise qu'ils soient informatisés ou non informatisés.
- Les sources externes représentent toutes les données provenant des systèmes externes de l'entreprise, informatisés ou non informatisés, ou produites par une tierce personne n'appartenant pas au personnel de l'entreprise.

➢ Consultation des données provenant des systèmes internes de l'entreprise (sources internes):

- **Progiciel de gestion intégré (PGI) ou Enterprise Resource Planning (ERP)**: « Le progiciel de gestion intégré (PGI), comme son nom l'indique, intègre tous les services et toutes les fonctions d'une organisation en un seul système d'information » (Baltzan, 2018, p. 278).

Selon le barème suivant, indiquez, en cliquant sur la case représentant le chiffre qui correspond le mieux à votre opinion sur l'échelle de 1 à 5, **(1) indiquant « jamais », (2) « rarement », (3) « occasionnellement », (4) « souvent » et (5) « très souvent »**. Veuillez répondre à tous les énoncés.

1. Dans quelle mesure consultez-vous les sources internes des données et d'informations suivantes :

Énoncés	Jamais	Rarement	Occasionnellement	Souvent	Très souvent
Pour générer les données des sources internes, je consulte ...					
a) ... les données provenant d'un progiciel de gestion intégrée/ERP *(p. ex.: SAP, Oracle, etc.)*	1	2	3	4	5
b) ... les données provenant des outils bureautiques *(p. ex. : tableurs, traitement de texte, etc.)*	1	2	3	4	5
c) ... les données provenant des activités de l'entreprise sur le Web *(p. ex. : Affaires électroniques, marketing numérique, etc.)*	1	2	3	4	5
d) ...les données provenant de publications de l'entreprise sur le Web *(p. ex. : Réseaux sociaux, blogues, sites, etc.)*	1	2	3	4	5

e) ...les formulaires ou documents numériques de gestion *(p. ex.: facturation, inventaire, gestion de la production, ventes, etc.)* destinés à une utilisation interne	**1**	**2**	**3**	**4**	**5**
f) ...les formulaires ou documents papier de gestion *(p. ex.: facturation, inventaire, gestion de la production, ventes, etc.)* destinés à une utilisation interne	**1**	**2**	**3**	**4**	**5**
g) ...les données de localisation GPS *(p. ex.: transporteurs, etc.)*	**1**	**2**	**3**	**4**	**5**
h) ...les enregistrements des appareils de surveillance *(p. ex.: bandes vidéos)*	**1**	**2**	**3**	**4**	**5**
i)...les images / photos prises au sein de l'entreprise *(p. ex.: des produits, communication, conférences, présentations, etc.)*	**1**	**2**	**3**	**4**	**5**
j)...les vidéos pour usage interne *(p. ex.: pour formation, communication, etc.)*	**1**	**2**	**3**	**4**	**5**
k) Autres :	**1**	**2**	**3**	**4**	**5**

- **Consultation des données provenant des sources extérieures (sources externes) :**

Selon le barème suivant, indiquez, en cliquant sur la case représentant le chiffre qui correspond le mieux à votre opinion sur l'échelle de 1 à 5, **(1) indiquant « jamais », (2) « rarement », (3) « occasionnellement », (4) « souvent » et (5) « très souvent »**. Veuillez répondre à tous les énoncés.

2. Dans quelle mesure consultez-vous les sources externes suivantes des données et d'informations :

Énoncés	**Jamais**	**Rarement**	**Occasionnellement**	**Souvent**	**Très souvent**
Pour générer les données des sources externes, je consulte...					
a) ...la messagerie instantanée *(p. ex.: www.aim.com, talk.google.com, Teams, etc.)*	**1**	**2**	**3**	**4**	**5**
b) ...le contenu collaboratif *(p. ex.: Wikipédia, Documents modifiables via Google Docs, Onedrive, etc.)*	**1**	**2**	**3**	**4**	**5**
c) ...la visioconférence et leurs enregistrements *(p. ex.: ConnectWebConference, Zoom, Skype, Teams, etc.)*	**1**	**2**	**3**	**4**	**5**
d) ...les publications sur les réseaux sociaux produites par d'autres organisations *(p. ex.: Facebook, LinkedIn, Instagram, etc.)*	**1**	**2**	**3**	**4**	**5**

e)	...les blogues et sites spécialisés rédigés par des personnes en dehors de l'entreprise *(p. ex.: consultants, organismes socioéconomiques, ministères, etc.)*	**1**	**2**	**3**	**4**	**5**
f)	...les outils de recherche, de présentation et d'agrégation de contenus *(p. ex.: Google, Flipboard, Feedly, etc.)*	**1**	**2**	**3**	**4**	**5**
g)	...les informations produites dans le cadre d'affaires électroniques où je suis un client *(p. ex.: transactions en ligne avec une autre entreprise, accès à une zone privée sur le site d'un tiers, etc.)*	**1**	**2**	**3**	**4**	**5**
h)	...les données relatives à l'industrie, mon secteur d'activités ou l'économie en générale *(p. ex.: gouvernement, revues économiques, marchés boursiers, etc.)*	**1**	**2**	**3**	**4**	**5**
i)	...autres documents informatisés de nature transactionnelle concernant des partenaires d'affaires *(p. ex.: fournisseurs, clients, distributeurs, etc.)*	**1**	**2**	**3**	**4**	**5**
j)	... autres documents papier de nature transactionnelle concernant des partenaires d'affaires *(p. ex.: fournisseurs, clients, distributeurs, etc.)*	**1**	**2**	**3**	**4**	**5**
k)	Autres:	**1**	**2**	**3**	**4**	**5**

ACQUISITION DES DONNÉES

L'acquisition des données consiste à agréger les informations sous format numérique pour un stockage et une analyse ultérieure. Le processus comprend deux sous-étapes, la préparation et la transmission des données (traduction libre) (Hu et al., 2014, p. 10).

➢ Préparation des données :

Cette première étape de l'acquisition comprend, à son tour, deux activités, soit la **collecte des données** qui consiste à utiliser des techniques spéciales de collecte des données pour acquérir des données brutes à partir d'un environnement de génération de données spécifique (traduction libre) (Chen et al., 2014, p. 11). Ensuite, le **prétraitement des données** est la phase au cours de laquelle des activités telles que le nettoyage des données, la déduplication, la compression, le filtrage et la conversion de format ont lieu. Cette étape obligatoire est indispensable pour affiner et valoriser les données. (traduction libre) (Taleb et al., 2015, p. 2).

Selon le barème suivant, indiquez, en cliquant sur la case représentant le chiffre qui correspond le mieux à votre opinion sur l'échelle de 1 à 5, **(1) indiquant « jamais », (2) « rarement », (3) « occasionnellement », (4) « souvent » et (5) « très souvent »**. Veuillez répondre à tous les énoncés.

3. Dans quelle mesure réalisez-vous les activités de préparation des données suivantes :

Énoncés	Jamais	Rarement	Occasionnellement	Souvent	Très souvent
Pour la collecte des données, j'utilise…					
a) … les outils analytiques en ligne *(p. ex. : Google Analytics, Hubspot, etc.)*	1	2	3	4	5
b) …les techniques d'intégration et d'unification des données *(p. ex. : requêtes dans les bases de données, etc.)*	1	2	3	4	5
c) … les outils informatisés de gestion d'activités de l'entreprise *(p. ex. : enregistrements d'appels, rapports numériques provenant des services et départements de l'entreprise, répertoires et fichiers informatisés concernant les clients/fournisseurs, etc.)*	1	2	3	4	5
d) …les outils non informatisés de gestion d'activités de l'entreprise *(p. ex. : rapports papier d'appels de services, documents papier divers provenant des services et départements de l'entreprise, filières papier concernant les clients/fournisseurs, etc.).*	1	2	3	4	5
Pour le prétraitement des données collectées, j'effectue …					
e) … l'amélioration et l'enrichissement des données en combinant des données provenant de sources multiples et des collections des données telles que l'intégration et la fusion des données.	1	2	3	4	5
f) …la transformation des données (normalisation et agrégation) en convertissant les données d'un format à un autre *(p. ex.: documents .docx mis en .pdf, tableurs et chiffriers .xls convertis en .csv, fichiers numériques (avec audio) transformés en .mp4, etc.)*	1	2	3	4	5
g) …la réduction des données en générant des vues réduites des données sans impact sur les résultats d'analyse *(p. ex. : la compression des données, le regroupement, la réduction de dimension, etc.)*	1	2	3	4	5
h) …le nettoyage des données pour identifier les données inexactes, incomplètes ou déraisonnables, puis les mettre à jour, les réparer ou les supprimer pour améliorer la qualité de mes données *(p. ex.: vérification et effacement des*	1	2	3	4	5

données de création concernant les fichiers, anti-virus, correction des erreurs typographiques avec Antidote, élimination des redondances dans des listes informatisées, etc.)					
i) Autres :	**1**	**2**	**3**	**4**	**5**

➢ Transmission des données :

Pour notre étude, la transmission des données représente les moyens de partage et de transfert des données entre les différents services en interne et avec les partenaires en externe.

Selon le barème suivant, indiquez, en cliquant sur la case représentant le chiffre qui correspond le mieux à votre opinion sur l'échelle de 1 à 5, **(1) indiquant « jamais », (2) « rarement », (3) « occasionnellement », (4) « souvent » et (5) « très souvent »**. Veuillez répondre à tous les énoncés.

4. Dans quelle mesure utilisez-vous les moyens ci-dessous pour effectuer la transmission, le transfert ou le partage des données vers les différents services de l'entreprise :

Énoncés	Jamais	Rarement	Occasionnellement	Souvent	Très souvent
Pour la transmission des données, j'utilise…					
a) …un Système de Gestion de Bases de Données interne *(p. ex.: MS Access, etc.)*	**1**	**2**	**3**	**4**	**5**
b) …des serveurs internes (entrepôts ou dépôts de données)	**1**	**2**	**3**	**4**	**5**
c) …un Progiciel de Gestion Intégré / ERP *(p. ex.: Oracle, SAP, etc.)*	**1**	**2**	**3**	**4**	**5**
d) …des papiers et documents physiques	**1**	**2**	**3**	**4**	**5**
e) …des courriels	**1**	**2**	**3**	**4**	**5**
f) … des outils collaboratifs et messageries *(p. ex. : Teams, Slack, etc.)*	**1**	**2**	**3**	**4**	**5**
g) …le partage des fichiers en ligne (infonuagique/*cloud*) *(p. ex. : Google Drive, Onedrive dans MS Office 365, DropBox, WeTransfer, etc.)*	**1**	**2**	**3**	**4**	**5**
h) Autres :	**1**	**2**	**3**	**4**	**5**

STOCKAGE DES DONNÉES

L'activité stockage de données représente les composants logiciels pour le stockage et la gestion de grands référentiels de données (traduction libre) (Coleman et al., 2016, p. 9). C'est-à-dire la façon et les outils permettant l'entreposage des données de l'entreprise.

On distingue deux types d'outils de stockage / entreposage des données:

- Les outils internes (ou dispositifs internes) représentent tous les outils de stockage appartenant à l'entreprise et toutes les données stockées en interne et ne quittant pas les lieux de l'entreprise.
- Les outils externes (ou dispositifs externes) représentent tous les outils ou l'espaces de stockage n'appartenant pas à l'entreprise et qui peuvent être loués par l'entreprise pour l'utilisation pendant une durée déterminée ou toute donnée stockée en dehors des locaux de l'entreprise.

➢ Dispositifs de stockage internes :

Selon le barème suivant, indiquez, en cliquant sur la case représentant le chiffre qui correspond le mieux à votre opinion sur l'échelle de 1 à 5, **(1) indiquant « jamais », (2) « rarement », (3) « occasionnellement », (4) « souvent » et (5) « très souvent »**. Veuillez répondre à tous les énoncés.

5. Dans quelle mesure utilisez-vous les dispositifs de stockage internes ci-dessous pour la conservation des données de l'entreprise :

Énoncés	Jamais	Rarement	Occasionnellement	Souvent	Très souvent
Pour conserver les données avec des dispositifs de stockage internes de l'entreprise, j'utilise...					
a) ...des serveurs internes se trouvant sur les lieux physiques de l'entreprise (entrepôts ou dépôts de données)	1	2	3	4	5
b) ...des archives et documents papier ne quittant pas les lieux physiques de l'entreprise	1	2	3	4	5
c) ...des archives numériques conservées ailleurs que sur les serveurs *(p. ex.: disques durs externes, clefs USB, etc.)* ne quittant pas les lieux physiques de l'entreprise	1	2	3	4	5
d) Autres :	1	2	3	4	5

➢ Dispositifs de stockage externes :

Selon le barème suivant, indiquez, en cliquant sur la case représentant le chiffre qui correspond le mieux à votre opinion sur l'échelle de 1 à 5, **(1) indiquant « jamais », (2) « rarement », (3) « occasionnellement », (4) « souvent » et (5) « très souvent »**. Veuillez répondre à tous les énoncés.

6. Dans quelle mesure utilisez-vous les dispositifs de stockage externes ci-dessous pour la conservation des données de l'entreprise :

Énoncés	Jamais	Rarement	Occasionnellement	Souvent	Très souvent
Pour conserver les données avec des dispositifs de stockage externes de l'entreprise, j'utilise...					
a) ...des serveurs et bases de données externes en nuage (*p. ex.: Google Drive, DropBox, WeTransfer, Microsoft Azure, AWS d'Amazon, etc.*)	1	2	3	4	5
b) ...des archives sur papier qui peuvent quitter les lieux physiques de l'entreprise	1	2	3	4	5
c) ...des archives numériques (*p. ex.: disques durs externes, clefs USB, etc.*) qui peuvent quitter les lieux physiques de l'entreprise	1	2	3	4	5
d) ...l'informatique en grille / stockage partitionné (stockage sur plusieurs ordinateurs)	1	2	3	4	5
e) Autres :	1	2	3	4	5

ANALYSE DES DONNÉES

L'analyse des données est un processus d'inspection et de modélisation des données dans le but de découvrir des informations utiles, de suggérer des conclusions et d'appuyer la prise de décision. L'analyse des données a de multiples facettes et approches, englobant diverses techniques sous divers noms, dans différents domaines des affaires, des sciences et des sciences sociales (traduction libre) (Gao et al., 2016, p. 5).

On distingue ici, deux types de moyens d'analyse de données : les moyens internes et externe :

- Les moyens internes représentent tous les outils d'analyse appartenant ou développés par l'entreprise.
- Les moyens externes représentent tous les outils d'analyse n'appartenant pas à l'entreprise ou qui sont développés par des tiers (p. ex. agences spécialisées). Ces moyens peuvent toutefois être offerts par l'intermédiaire de logiciels-services dont l'utilisation est payée en fonction de l'usage, pendant une durée déterminée (*p. ex. :* Salesforce).

➢ Applications et outils d'analyse appartenant ou développés par l'entreprise (moyens internes) :

Selon le barème suivant, indiquez, en cliquant sur la case représentant le chiffre qui correspond le mieux à votre opinion sur l'échelle de 1 à 5, **(1) indiquant « jamais », (2) « rarement », (3) « occasionnellement », (4) « souvent » et (5) « très souvent »**. Veuillez répondre à tous les énoncés.

7. Dans quelle mesure utilisez-vous les moyens internes ci-dessous pour l'analyse de vos données :

Énoncés	Jamais	Rarement	Occasionnellement	Souvent	Très souvent
Pour analyser les données avec des moyens internes, j'utilise…					
a) …des requêtes dans les bases de données internes *(p. ex.: SGBD, SQL, etc.)*	1	2	3	4	5
b) …des chiffriers électroniques pour manipuler les données et en extraire des informations utiles *(p. ex.: Excel, Visual basic (VBA), tableaux croisés dynamiques, etc.)*	1	2	3	4	5
c) …des rapports provenant des systèmes de traitement de transactions (STT) *(p. ex. : système comptable de l'entreprise, système de gestion des ressources humaines pour la paie, système informatisé pour les ventes, etc.)*	1	2	3	4	5
d) …des scénarios et prévisions à l'aide des systèmes interactifs d'aide à la décision (SIAD) *(p. ex. : prévisions pour la planification de la production, simulations permettant de prendre des décisions financières, etc.)*	1	2	3	4	5
e) …des composantes de base du PGI / ERP *(p. ex.: module comptable / finance, gestion de production, gestion des ressources humaines, ventes, etc.)*	1	2	3	4	5
f) …des composantes élargies du PGI / ERP *(p. ex.: module gestion de relation client, gestion d'approvisionnement / logistique, affaires électroniques, etc.)*	1	2	3	4	5
g) …des composantes internes de gestion ne provenant pas d'un PGI / ERP *(p. ex.: SGBD, système comptable informatisé, système de gestion de la production / inventaires, système de gestion des ressources humaines, système de gestion des ventes / marketing, etc.)*	1	2	3	4	5
h) …des outils d'intelligence d'affaires (Business intelligence – BI) sur des données internes *(p. ex.: Power BI, Tableau, Cognos, etc.)*	1	2	3	4	5
i) Autres :	1	2	3	4	5

➢ Applications et outils d'analyse développés par des tiers (moyens externes) :

- **Infrastructure-service / Infrastructure as a Service (IaaS) :** « Service offrant du matériel de mise en réseau, notamment des serveurs, des ressources réseau et de l'espace de stockage, par l'entremise de l'informatique en nuage, le tout étant payable à l'utilisation » (Baltzan, 2018, p. 185).

Selon le barème suivant, indiquez, en cliquant sur la case représentant le chiffre qui correspond le mieux à votre opinion sur l'échelle de 1 à 5, **(1) indiquant « jamais », (2) « rarement », (3) « occasionnellement », (4) « souvent » et (5) « très souvent »**. Veuillez répondre à tous les énoncés.

8. Dans quelle mesure utilisez-vous les moyens externes ci-dessous pour le traitement de vos données ?

Énoncés	Jamais	Rarement	Occasionnellement	Souvent	Très souvent
Pour analyser les données avec des moyens externes, j'utilise…					
a) …des requêtes sur bases de données externes *(SQL server, etc.)*	1	2	3	4	5
b) …des outils d'analyse d'intelligence d'affaires (Business intelligence – BI) sur des données externes comme les affaires électroniques de l'entreprise *(p. ex. : statistiques de fréquentation du site web de l'entreprise, Google Analytics, Facebook Analytics, Power BI, Tableau, etc.).*	1	2	3	4	5
c) …l'Infrastructure ou logiciels-service (IaaS) pouvant être loués ou facturés à l'utilisation. *(p. ex. : AWS d'Amazon, Microsoft Azure, etc.)*	1	2	3	4	5
d) Autres :	1	2	3	4	5

DIFFUSION / VISUALISATION DE L'INFORMATION

L'activité diffusion /visualisation des données permet de « produire des représentations graphiques des tendances et des relations complexes dans de grandes quantités de données » (Baltzan, 2018, p. 48).
C'est la forme sous laquelle les données sont diffusées dans l'entreprise après leur traitement par / pour chaque service ou département.

On distingue ici deux types de données : les données internes et externes :

- Les données internes sont destinées au personnel de l'entreprise.
- Les données externes sont destinées à toute personne agissant en dehors de l'entreprise, qui ne fait pas partie du personnel salarié, ou de l'actionnariat, le cas échéant.

➢ Données destinées au personnel de l'entreprise (données internes) :

Selon le barème suivant, indiquez, en cliquant sur la case représentant le chiffre qui correspond le mieux à votre opinion sur l'échelle de 1 à 5, **(1) indiquant « jamais », (2) « rarement », (3) « occasionnellement », (4) « souvent » et (5) « très souvent »**. Veuillez répondre à tous les énoncés.

9. Dans quelle mesure utilisez-vous les outils de diffusion internes suivants :

Énoncés	Jamais	Rarement	Occasionnellement	Souvent	Très souvent
J'effectue la diffusion des données destinées au personnel de l'entreprise en utilisant...					
a) ...des rapports de gestion informatisés internes *(p. ex.: graphiques, figures, tableaux, etc.)*	**1**	**2**	**3**	**4**	**5**
b) ...des rapports et extraits de tableaux de bord numériques (indicateurs de performance) destinés à l'ensemble de l'entreprise *(p. ex.: MS Excel, PGI, Power BI, etc.)*	**1**	**2**	**3**	**4**	**5**
c) ...des rapports et extraits de tableaux de bord numériques (indicateurs de performance) destinés aux dirigeants de l'entreprise	**1**	**2**	**3**	**4**	**5**
d) ...des formulaires ou documents papier de gestion *(p. ex.: facturation, inventaire, comptabilité, etc.)* destinés à une utilisation interne	**1**	**2**	**3**	**4**	**5**
e) ...des formulaires ou documents numériques de gestion *(p. ex.: facturation, inventaire, comptabilité, etc.)* destinés à une utilisation interne	**1**	**2**	**3**	**4**	**5**
f) Autres :	**1**	**2**	**3**	**4**	**5**

- **Données destinées à toute personne n'appartenant pas au personnel de l'entreprise (données externes):**

Selon le barème suivant, indiquez, en cliquant sur la case représentant le chiffre qui correspond le mieux à votre opinion sur l'échelle de 1 à 5, **(1) indiquant « jamais », (2) « rarement », (3) « occasionnellement », (4) « souvent » et (5) « très souvent »**. Veuillez répondre à tous les énoncés.

10. Dans quelle mesure utilisez-vous les outils de diffusion en externe suivants :

Énoncés	Jamais	Rarement	Occasionnellement	Souvent	Très souvent
J'effectue la diffusion des données destinées aux personnes n'appartenant pas au personnel de l'entreprise en utilisant…					
a) …des rapports de gestion publics à des fins informationnelles ou liées à des obligations légales *(p. ex.: rapports annuels ou périodiques, états financiers de société publique, etc.)*	1	2	3	4	5
b) …des formulaires ou documents numériques de gestion de nature transactionnelle avec d'autres entreprises *(p. ex.: factures, bons de commandes, etc.)*	1	2	3	4	5
c) …des formulaires ou documents papier de gestion de nature transactionnelle *(p. ex.: facturation, bons de commande/livraison, etc.)* destinés à une diffusion externe à l'entreprise *(p. ex. : partenaires, distributeurs, gouvernements, etc.)*	1	2	3	4	5
d) …des formulaires ou documents numériques de gestion de nature informationnelle *(p. ex.: états financiers, documentation promotionnelle, etc.)* destinés à une diffusion externe à l'entreprise *(p. ex.: partenaires, collaborateurs, distributeurs, gouvernement, etc.)*	1	2	3	4	5
e) …des formulaires ou documents papier de gestion de nature informationnelle *(p. ex.: états financiers, plan stratégique, documentation promotionnelle, etc.)* destinés à une diffusion externe à l'entreprise *(p. ex.: partenaires, collaborateurs, distributeurs, gouvernement, etc.)*	1	2	3	4	5
f) …le suivi des comptes-clients/comptes-fournisseurs en temps réel *(p. ex.: zone privée pour les comptes clients, historique des commandes, etc.)* dans les affaires électroniques	1	2	3	4	5
g) …les données soutenant les activités du marketing numérique *(p. ex.: l'automatisation des envois, notifications automatisées, courriels de suivi, etc.)*	1	2	3	4	5
h) Autres :	1	2	3	4	5

QUALITÉ DE L'INFORMATION

La qualité des données en gestion est définie par leur conformité aux besoins de l'utilisateur (Traduction libre) (Taleb et al., 2015, p. 2). Une information de qualité doit avoir les neuf caractéristiques suivantes : l'exactitude, la crédibilité, la compréhensibilité, la cohérence, la conformité, la disponibilité, la complétude, la disponibilité et l'unicité (traduction libre) (Merino et al., 2016; Rafique et al., 2012; Taleb et al., 2015). Selon le barème suivant, indiquez, en cliquant sur la case représentant le chiffre qui correspond le mieux à votre opinion sur l'échelle de 1 à 5, **(1) indiquant « pas du tout », (2) « rarement », (3) « occasionnellement », (4) « souvent » et (5) « très souvent »**. Veuillez répondre à tous les énoncés.

11. Je considère que l'information qui circule dans l'entreprise est :

Énoncés	Pas du tout	Rarement	Occasionnellement	Souvent	Très souvent
a) Exacte : L'information fournie est correcte, sans erreurs et fiable.	1	2	3	4	5
b) Crédible : L'information fournie est objective (non biaisée) et fiable (vraie, selon ce que j'en sais).	1	2	3	4	5
c) Compréhensible : L'information est facile à comprendre.	1	2	3	4	5
d) Cohérente : L'information fournie est présentée dans des formats qui coïncident entre eux.	1	2	3	4	5
e) Conforme : L'information est applicable et utile pour la tâche à accomplir.	1	2	3	4	5
f) Disponible : L'information est récupérable, disponible et prête à être utilisée au bon moment par l'utilisateur	1	2	3	4	5
g) Complète : L'information fournie contient toutes les valeurs nécessaires. L'information est généralement complète pour répondre à nos besoins ou pour remplir la tâche à accomplir.	1	2	3	4	5
h) Actuelle : L'information fournie est suffisamment à jour pour le travail à accomplir.	1	2	3	4	5
i) Unique : Il n'y a pas de redondance inutile entre les différentes sources d'information.	1	2	3	4	5

INFORMATIONS SOCIO-DÉMOGRAPHIQUES (entreprise et répondant.e)

12. Indiquez l'année de fondation de l'entreprise (format aaaa) : _____

13. Indiquez le secteur qui correspond le mieux aux *activités principales* de l'entreprise (**un seul choix possible**).

Division	Code SIC	Classification type des industries (CTI) / Standard Industrial Classification (SIC)	
D	20-39	Fabrication	O
E	40-49	Transport, communications, électricité, gaz et services sanitaires	O
F	50-51	Commerce de gros	O
G	52-59	Commerce de détail	O
*H	60-67	Finance, assurance et immobilier	O
	60	Institutions de dépôt	O
	61	Établissements de crédit non-dépositaire	O
	62	Courtiers, concessionnaires, bourses et services en matière de sécurité et de matières premières	O
	63	Transporteurs d'assurance	O
	64	Agents, courtiers et services d'assurance	O
I	70-89	Prestations de service	O
	73	Services aux entreprises	O

14. Indiquez la région où se déroule la majeure partie des activités de l'entreprise (**un seul choix possible**) :

Code	Régions administratives du Québec (nom)		Code	Régions administratives du Québec (nom)	
01	Bas St-Laurent	O	*10*	Nord-du-Québec	O
02	Saguenay-Lac-St-Jean	O	*11*	Gaspésie-Îles-de-la-Madeleine	O
03	Capitale-Nationale	O	*12*	Chaudière-Appalaches	O
04	Mauricie	O	*13*	Laval	O
05	Estrie	O	*14*	Lanaudière	O
06	Montréal	O	*15*	Laurentides	O
07	Outaouais	O	*16*	Montérégie	O
08	Abitibi-Témiscamingue	O	*17*	Centre-du-Québec	O
09	Côte-Nord	O			

15. Indiquez dans quelle tranche se situe la moyenne du nombre d'**employés salariés à temps plein** pour l'année 2019.

Tranches	2019
De 1 à 9 employé.e.s	O
De 10 à 19 employé.e.s	O
De 20 à 49 employé.e.s	O
De 50 à 99 employé.e.s	O
De 100 à 249 employé.e.s	O
De 250 à 500 employé.e.s	O
Ne sais pas	O

16. Indiquez dans quelle tranche se situe le chiffre d'affaires de l'entreprise[23] pour l'année financière 2019.

Tranches	2019
Moins de 250 000 $	O
De 250 000 $ à 499 999 $	O
De 500 000 à 999 999 $	O
De 1 000 000 à 4 999 999 $	O
De 5 000 000 à 9 999 999 $	O
Plus de 10 000 000 $	O
Ne sais pas	O

17. Indiquez dans quelle tranche se situe l'investissement global réalisé concernant les ressources technologiques de l'entreprise (p. ex. équipement informatique, logiciels, réseaux) pour l'année financière 2019.

Tranches	2019
Aucun investissement réalisé	O
Moins de 10 000 $	O
De 10 000 à 49 999 $	O
De 50 000 $ à 99 999 $	O
De 100 000 à 249 999 $	O
Plus de 250 000 $	O
Ne sais pas	O

18. Vous êtes : O Femme O Homme

19. Indiquez votre tranche d'âge : O 20-29 ans O 30-39 ans O 40-49 ans
O 50-59 ans O 60 ans et +

20. Indiquez le plus haut degré de scolarité complété :
O Secondaire O Collégial O Universitaire (1er cycle)
O Universitaire (2e-3e cycle)

21. Indiquez votre poste/titre actuel dans l'entreprise : _____

22. Est-ce que l'entreprise a une personne dédiée à la gestion des technologies de l'information (TI) ?
O Oui O Non
➲ *Si oui, siégez-vous au comité de direction de votre entreprise?* O Oui O Non
➲ *Si non, indiquez dans quelle mesure vous êtes impliqué.e dans les projets TI de l'entreprise :*

Aucune implication	Je suis un peu impliqué(e)	Je suis moyennement impliqué(e)	Je suis souvent impliqué(e)	Je suis toujours impliqué(e)
O	O	O	O	O

23. Souhaitez-vous obtenir un résumé des résultats de cette recherche : O Oui O Non
➲ *Si oui*, indiquez une adresse courriel : _____@_____

Merci pour votre précieuse collaboration à cette recherche!

[23] Nous vous rappelons que **TOUTE l'information fournie demeurera strictement confidentielle**.

ANNEXE C

LETTRES D'INFORMATION-CONSENTEMENT

LETTRE D'INFORMATION – QUESTIONNAIRE (EN LIGNE)

LES PME ET LE TRAITEMENT DE L'INFORMATION À L'ÈRE DES DONNÉES MASSIVES

Votre participation à la recherche, qui vise à comprendre comment les PME acquièrent, évaluent, traitent et stockent les données massives tout en assurant la qualité de leurs informations, serait grandement appréciée. Le but de cette lettre d'information est de vous aider à comprendre exactement ce qu'implique votre éventuelle participation à la recherche de sorte que vous puissiez prendre une décision éclairée à ce sujet. Prenez donc le temps de la lire attentivement et n'hésitez pas à poser toute question que vous jugerez utile.

Objectifs

L'objectif général de cette étude est d'explorer les activités et les outils de gestion permettant d'assurer la qualité de l'information au sein des PME à l'ère des données massives. Cela en conduisant une étude exploratoire sur quelques PME québécoises et leurs manières de traiter les données provenant de différentes sources pour les transformer en information de qualité.
Plus précisément, la recherche vise à déterminer et comprendre les différentes configurations d'activités et d'outils de gestion qui assurent une haute qualité de l'information et celles qui mènent à une faible qualité de l'information au sein des PME à l'ère des données massives.

Tâches

Votre participation à ce projet de recherche consiste à remplir un questionnaire en ligne. La durée estimée pour remplir ce questionnaire est d'environ 20 minutes. Le questionnaire est composé des huit sections suivantes concernant la gestion des données dans les PME :

1. Instructions
2. Génération des données
3. Acquisition des données
4. Stockage des données
5. Analyse des données
6. Visualisation / diffusion de l'information
7. Qualité de l'information
8. Informations sociodémographiques

Risques

Aucun risque n'est associé à votre participation. Le temps consacré à remplir ce questionnaire par les personnes-ressources demeure le seul inconvénient.

Bénéfices

La participation à cette recherche pour une entreprise comporte l'avantage d'inciter les personnes-ressources et les dirigeants à se questionner face au processus de transformation des données, à

revoir et évaluer ce processus en allant vers la définition des bonnes pratiques, les outils de gestion ainsi que les activités permettant l'amélioration du traitement de leurs données. Les résultats de la recherche seront communiqués aux participants sous forme d'un rapport électronique envoyé par courriel sur demande aux personnes concernées.

Compensation ou incitatif
La participation à cette recherche n'implique aucune compensation d'ordre monétaire.

Confidentialité
La présente recherche respecte les règles d'éthique et de déontologie de la recherche à l'Université du Québec à Trois-Rivières, tel qu'en témoigne le certificat émis sous le numéro CER-19-263-07.16.
Les données recueillies par cette étude sont entièrement confidentielles et ne pourront en aucun cas mener à votre identification. Toutes les données collectées seront rendues anonymes pour les fins d'analyse. Ces données seront conservées par l'étudiante, sous format numérique, pour la durée de ses études. Les résultats de la recherche seront diffusés sous forme de mémoire de recherche et leur exploitation sera destinée à des fins de publications scientifiques et professionnelles. Elles seront, ensuite, détruites.
Le cas échéant, les seules personnes qui auront accès aux données feront partie de l'équipe de recherche par exemple la directrice de recherche Mme Claudia PELLETIER, les évaluateurs du mémoire de recherche et les autres personnes impliquées pour l'analyse des données, le cas échéant. Elles signeront toutes un engagement à la confidentialité.

Participation volontaire
Votre participation à cette étude se fait sur une base volontaire. Vous êtes entièrement libre de participer ou non, de refuser de répondre à certaines questions ou de vous retirer en tout temps sans préjudice et sans avoir à fournir d'explications.

Responsable de la recherche
Pour obtenir de plus amples renseignements ou pour toute question concernant ce projet de recherche, vous pouvez communiquer avec l'étudiante ou sa directrice de recherche aux coordonnées suivantes :

Nada FEJJAR, étudiante	**CLAUDIA PELLETIER, DBA**
École de gestion, UQTR	Professeure en systèmes d'information
nada.fejjar@uqtr.ca	École de gestion, UQTR
819-384-7368	Claudia.Pelletier@uqtr.ca
	819-376-5011, poste 4271.

Question ou plainte concernant l'éthique de la recherche

Pour toute question ou plainte d'ordre éthique concernant cette recherche, vous devez communiquer avec la secrétaire du comité d'éthique de la recherche de l'Université du Québec à Trois-Rivières, par téléphone (819) 376-5011, poste 2129 ou par courrier électronique CEREH@uqtr.ca.

Votre collaboration est précieuse.
Nous l'apprécions et vous en remercions.

FORMULAIRE DE CONSENTEMENT – QUESTIONNAIRE (EN LIGNE)

LES PME ET LE TRAITEMENT DE L'INFORMATION À L'ÈRE DES DONNÉES MASSIVES

Engagement de la chercheuse

Moi, **Nada FEJJAR** m'engage à procéder à cette étude conformément à toutes les normes éthiques qui s'appliquent aux projets comportant la participation de sujets humains.

Consentement du participant

Je, _____, confirme avoir lu et compris la lettre d'information au sujet du projet « **LES PME ET LE TRAITEMENT DE L'INFORMATION À L'ÈRE DES DONNÉES MASSIVES** ». J'ai bien saisi les conditions, les risques et les bienfaits éventuels de ma participation. On a répondu à toutes mes questions à mon entière satisfaction. J'ai disposé de suffisamment de temps pour réfléchir à ma décision de participer ou non à cette recherche. Je comprends que ma participation est entièrement volontaire et que je peux décider de me retirer en tout temps, sans aucun préjudice.

J'accepte donc librement de participer à ce projet de recherche

En cliquant sur le bouton de participation, vous indiquez

- avoir lu l'information
- être d'accord pour participer

Oui, j'accepte de participer

ANNEXE D

CODAGE DES VARIABLES D'ÉTUDE APRÈS FACTORISATION

Tableau D

Codage des variables d'études après factorisation et fusion des variables *ACD* et *AD* en *AAD*

	Items	Type	Code
Génération des données		**Variable**	**GD**
Consultation des données provenant des systèmes internes de l'entreprise	Pour générer les données des sources internes, je consulte …	Dimension 1	CDI
	… les données provenant des activités de l'entreprise sur le Web (p. ex. : Affaires électroniques, marketing numérique, etc.)	Item 1	CDI3
	…les formulaires ou documents papier de gestion (p. ex. facturation, inventaire, gestion de la production, ventes, etc.) destinés à une utilisation interne	Item 2	CDI6
	…les données de localisation GPS (p. ex. transporteurs, etc.)	Item 3	CDI7
	…les vidéos pour usage interne (p. ex.: pour formation, communication, etc.)	Item 4	CDI10
	…les données relatives à l'industrie, mon secteur d'activités ou l'économie en générale (p. ex. gouvernement, revues économiques, marchés boursiers, etc.)	Item 5	CDE8
	…autres documents informatisés de nature transactionnelle concernant des partenaires d'affaires (p. ex. fournisseurs, clients, distributeurs, etc.)	Item 6	CDE9
	… autres documents papier de nature transactionnelle concernant des partenaires d'affaires (p. ex. fournisseurs, clients, distributeurs, etc.)	Item 7	CDE10
Consultation des données provenant des systèmes externes de l'entreprise	Pour générer les données des sources externes, je consulte…	Dimension 2	CDE
	…la messagerie instantanée (p. ex. www.aim.com, talk.google.com, Teams, etc.)	Item 1	CDE1
	…le contenu collaboratif (p. ex. Wikipédia, Documents modifiables via Google Docs, Onedrive, etc.)	Item 2	CDE2
	…la visioconférence et leurs enregistrements (p. ex. ConnectWebConference, Zoom, Skype, Teams, etc.)	Item 3	CDE3
	…les publications sur les réseaux sociaux produites par d'autres organisations (p. ex. Facebook, LinkedIn, Instagram, etc.)	Item 4	CDE4
	…les blogs et sites spécialisés rédigés par des personnes en dehors de l'entreprise (p. ex. consultants, organismes socioéconomiques, ministères, etc.)	Item 5	CDE5
	…les données provenant de publications de l'entreprise sur le Web (p. ex. : Réseaux sociaux, blogs, sites, etc.)	Item 6	CDI4
Acquisition et analyse des données		**Variable**	**AAD**
Outils et pratiques de gestion des données		Dimension 1	PGD
	…les techniques d'intégration et d'unification des données (p. ex. : requêtes dans les bases de données, etc.)	Item 1	PD_CD2

	Items	Type	Code
	... les outils informatisés de gestion d'activités de l'entreprise (p. ex. : enregistrements d'appels, rapports numériques provenant des services et départements de l'entreprise, répertoires et fichiers informatisés concernant les clients/fournisseurs, etc.)	Item 2	PD_CD3
	...la réduction des données en générant des vues réduites des données sans impact sur les résultats d'analyse (p. ex. : la compression des données, le regroupement, la réduction de dimension, etc.)	Item 3	PD_PT3
	...un Système de Gestion de Bases de Données interne (p. ex. MS Access, etc.)	Item 4	TD1
	...des serveurs internes (entrepôts ou dépôts de données)	Item 5	TD2
	...des requêtes dans les bases de données internes (p. ex. SGBD, SQL, etc.)	Item 6	ADI1
	...des composantes élargies du PGI / ERP (p. ex. module gestion de relation client, gestion d'approvisionnement / logistique, affaires électroniques, etc.)	Item 7	ADI6
Préparation, partage et traitement des données		Dimension 2	PTD
	... l'amélioration et l'enrichissement des données en combinant des données provenant de sources multiples et des collections des données telles que l'intégration et la fusion des données.	Item 1	PD_PT1
	...la transformation des données (normalisation et agrégation) en convertissant les données d'un format à un autre (p. ex. documents .docx mis en .pdf, tableurs et chiffriers .xls convertis en .csv, fichiers numériques (avec audio) transformés en .mp4, etc.)	Item 2	PD_PT2
	...le nettoyage des données pour identifier les données inexactes, incomplètes ou déraisonnables, puis les mettre à jour, les réparer ou les supprimer pour améliorer la qualité de mes données (p. ex. vérification et effacement des données de création concernant les fichiers, anti-virus, correction des erreurs typographiques avec Antidote, élimination des redondances dans des listes informatisées, etc.)	Item 3	PD_PT4
	... des outils collaboratifs et messageries (p. ex. : Teams, Slack, etc.)	Item 4	TD6
	...le partage des fichiers en ligne (infonuagique/cloud) (p. ex. : Google Drive, Onedrive dans MS Office 365, DropBox, WeTransfer, etc.)	Item 5	TD7
	...des rapports provenant des systèmes de traitement de transactions (STT) (p. ex. : système comptable de l'entreprise, système de gestion des ressources humaines pour la paie, système informatisé pour les ventes, etc.)	Item 6	ADI3
	...des composantes de base du PGI / ERP (p. ex. module comptable / finance, gestion de production, gestion des ressources humaines, ventes, etc.)	Item 7	ADI5

	Items	Type	Code
	…des composantes internes de gestion ne provenant pas d'un PGI / ERP (p. ex. SGBD, système comptable informatisé, système de gestion de la production / inventaires, système de gestion des ressources humaines, système de gestion des ventes / marketing, etc.)	Item 8	ADI7
Outils d'intelligence d'affaires		Dimension 3	OIA
	…des scénarios et prévisions à l'aide des systèmes interactifs d'aide à la décision (SIAD) (p. ex. : prévisions pour la planification de la production, simulations permettant de prendre des décisions financières, etc.)	Item 1	ADI4
	…des outils d'intelligence d'affaires (Business intelligence – BI) sur des données internes (p. ex. Power BI, Tableau, Cognos, etc.)	Item 2	ADI8
	…des requêtes sur bases de données externes (SQL server, etc.)	Item 3	ADE1
	…des outils d'analyse d'intelligence d'affaires (Business intelligence – BI) sur des données externes comme les affaires électroniques de l'entreprise (p. ex. : statistiques de fréquentation du site web de l'entreprise, Google Analytics, Facebook Analytics, Power BI, Tableau, etc.).	Item 4	ADE2
	…l'Infrastructure ou logiciels-service (IaaS) pouvant être loués ou facturés à l'utilisation. (p. ex. : AWS d'Amazon, Microsoft Azure, etc.)	Item 5	ADE3
Outils non informatisés		Dimension 4	ONI
	…les outils non informatisés de gestion d'activités de l'entreprise (p. ex. : rapports papier d'appels de services, documents papier divers provenant des services et départements de l'entreprise, filières papier concernant les clients/fournisseurs, etc.).	Item 1	PD_CD4
	…des papiers et documents physiques	Item 2	TD4
Stockage des données		**Variable**	**SD**
Dispositifs de stockage internes	Pour conserver les données avec des dispositifs de stockage internes de l'entreprise, j'utilise…	Dimension 1	DSI
	…des serveurs internes se trouvant sur les lieux physiques de l'entreprise (entrepôts ou dépôts de données)	Item 1	DSI1
	…des archives et documents papier ne quittant pas les lieux physiques de l'entreprise	Item 2	DSI2
	…l'informatique en grille / stockage partitionné (stockage sur plusieurs ordinateurs)	Item 3	DSE4
Dispositifs de stockage externes	Pour conserver les données avec des dispositifs de stockage externes de l'entreprise, j'utilise…	Dimension 2	DSE
	…des archives sur papier qui peuvent quitter les lieux physiques de l'entreprise	Item 1	DSE2
	…des archives numériques (p. ex. disques durs externes, clefs USB, etc.) qui peuvent quitter les lieux physiques de l'entreprise	Item 2	DSE3
	…des archives numériques conservées ailleurs que sur les serveurs (p. ex.: disques durs externes, clefs USB, etc.) ne quittant pas les lieux physiques de l'entreprise	Item 3	DSI3

	Items	Type	Code
Diffusion / visualisation de l'information		**Variable**	**VI**
Données destinées au personnel de l'entreprise	J'effectue la diffusion des données destinées au personnel de l'entreprise en utilisant…	Dimension 1	VII
	…des rapports de gestion informatisés internes (p. ex. graphiques, figures, tableaux, etc.)	Item 1	VII1
	…des rapports et extraits de tableaux de bord numériques (indicateurs de performance) destinés à l'ensemble de l'entreprise (p. ex. MS Excel, PGI, Power BI, etc.)	Item 2	VII2
	…des rapports et extraits de tableaux de bord numériques (indicateurs de performance) destinés aux dirigeants de l'entreprise	Item 3	VII3
Données destinées à toute personne n'appartenant pas au personnel de l'entreprise	J'effectue la diffusion des données destinées aux personnes n'appartenant pas au personnel de l'entreprise en utilisant…	Dimension 2	VIE
	…des rapports de gestion publics à des fins informationnelles ou liées à des obligations légales (p. ex. rapports annuels ou périodiques, états financiers de société publique, etc.)	Item 1	VIE1
	…des formulaires ou documents numériques de gestion de nature transactionnelle avec d'autres entreprises (p. ex. factures, bons de commandes, etc.)	Item 2	VIE2
	…des formulaires ou documents papier de gestion de nature transactionnelle (p. ex. facturation, bons de commande/livraison, etc.) destinés à une diffusion externe à l'entreprise (p. ex. : partenaires, distributeurs, gouvernements, etc.)	Item 3	VIE3
	…des formulaires ou documents numériques de gestion de nature informationnelle (p. ex. états financiers, documentation promotionnelle, etc.) destinés à une diffusion externe à l'entreprise (p. ex.: partenaires, collaborateurs, distributeurs, gouvernement, etc.)	Item 4	VIE4
	…des formulaires ou documents papier de gestion de nature informationnelle (p. ex. états financiers, plan stratégique, documentation promotionnelle, etc.) destinés à une diffusion externe à l'entreprise (p. ex.: partenaires, collaborateurs, distributeurs, gouvernement, etc.)	Item 5	VIE5
	…les données soutenant les activités du marketing numérique (p. ex. l'automatisation des envois, notifications automatisées, courriels de suivi, etc.)	Item 6	VIE7
	…des formulaires ou documents papier de gestion (p. ex. facturation, inventaire, comptabilité, etc.) destinés à une utilisation interne	Item 7	VII4
	…des formulaires ou documents numériques de gestion (p. ex. facturation, inventaire, comptabilité, etc.) destinés à une utilisation interne	Item 8	VII5
Qualité de l'information	**Je considère que l'information qui circule dans l'entreprise est …**	**Variable**	**QI**
	Exacte : L'information fournie est correcte, sans erreurs et fiable.	Item 1	QI1
	Cohérente : L'information fournie est présentée dans des formats qui coïncident entre eux.	Item 2	QI4

	Items	Type	Code
	Conforme : L'information est applicable et utile pour la tâche à accomplir.	Item 3	QI5
	Disponible : L'information est récupérable, disponible et prête à être utilisée au bon moment pour l'utilisateur	Item 4	QI6
	Actuelle : L'information fournie est suffisamment à jour pour le travail à accomplir.	Item 5	QI8
	Unique : Il n'y a pas de redondance inutile entre les différentes sources d'information.	Item 6	QI9

ANNEXE E

SYNTHÈSE DES RÉSULTATS DES ACP ET FIABILITÉ DES ÉCHELLES DE MESURE

Tableau E1

Résultats de l'épuration des instruments de mesure après plusieurs itérations d'ACP après fusion des variables *ACD* et *AD* en *AAD*

ANALYSES EN COMPOSANTES PRINCIPALES				**ITEMS RETIRÉS DE L'ANALYSE**
Variables	**KMO**	**Significativité de Bartlett**	**Qualité de représentation**	**Qualité de représentation ≤ 0,4**
GD	0,733	0,000	0,402 → 0,784	CDI1, CDI2, CDI5, CDI8, CDI9, CDE6, CDE7
AAD	0,631	< 0,001	0,382 → 0,803	PD_CD1, TD3, TD5, ADI2
SD	0,532	0,000	0,473 → 0,708	DSE1
VI	0,630	0,000	0,388 → 0,776	VIE6
QI	0,819	0,000	0,410 → 0,728	QI2, QI3, QI7

Tableau E2

Synthèse des résultats des ACP et fiabilité des échelles de mesure après la fusion des variables *ACD* et *AD* en *AAD*

ANALYSES EN COMPOSANTES PRINCIPALES					FIABILITÉ DE L'ÉCHELLE	
Variables	**Nbr d'items**	**Qualité de représentation**	**Contributions factorielles**	**Variances expliquées**	**Alpha de Cronbach (1951)**	**Rhô de Fornell et Karell (1981)**
GD	**13**	0,402 → 0,784	F1 : *(CDI)* 0,509→0,886 F2 : *(CDE)* 0,633→0,786	57,283 %	**0,856**	**0,930**
CDI	7				0,841	0,868
CDE	6				0,838	0,871
AAD	**22**	0,382 → 0,803	F1 : -0,215→0,828 F2 : -0,026→0,869 F3 : -0,096→0,824 F4 : -0,396→0,882	65,173 %	**0,906**	**0,950**
PGD	7				0,889	0,864
PTD	8				0,852	0,827
OIA	5				0,819	0,836
ONI	2				0,878	0,842
SD	**6**	0,473 → 0,708	F1 : *(DSE)* 0,686→0,833 F2 : *(DSI)* 0,576→0,839	58,120 %	**0,627**	**0,872**
DSI	3				0,562	0,753
DSE	3				0,642	0,793
VI	**11**	0,388 → 0,776	F1 : *(VIE)* 0,527→0,800 F2 : *(VII)* 0,733→0,878	56,673 %	**0,817**	**0,923**
VII	3				0,790	0,855
VIE	8				0,843	0,876
QI	**6**	0,410 → 0,728	0,641→0,853	63,591 %	**0,883**	**0,912**

ANNEXE F

RÉCAPITULATIF DES VARIABLES D'ÉTUDE APRÈS FACTORISATION

Tableau F

Récapitulatif des variables d'études après factorisation et fusion des variables *ACD* et *AD* en *AAD*

Variable	Dimension	Item
Génération des données (GD)		
	Consultation des données provenant des systèmes internes de l'entreprise (CDI)	CDI3, CDI6, CDI7, CDI10, CDE8, CDE9, CDE10
	Consultation des données provenant des systèmes externes de l'entreprise (CDE)	CDE1, CDE2, CDE3, CDE4, CDE5, CDI4
Acquisition et analyse des données (AAD)		
	Outils et pratiques de gestion des données (PGD)	PD_CD2, PD_CD3, PD_PT3, TD1, TD2, ADI1, ADI6
	Préparation, partage et traitement des données (PTD)	PD_PT1, PD_PT2, PD_PT4, TD6, TD7, ADI3, ADI5, ADI7
	Outils d'intelligence d'affaires (OIA)	ADI4, ADI8, ADE1, ADE2, ADE3
	Outils non informatisés (ONI)	PD_CD4, TD4
Stockage des données (SD)		
	Dispositifs de stockage internes (DSI)	DSI1, DSI2, DSE4
	Dispositifs de stockage externes (DSE)	DSE2, DSE3, DSI3
Diffusion / Visualisation de l'information (VI)		
	Données destinées au personnel de l'entreprise (VII)	VII1, VII2, VII3
	Données destinées à toute personne n'appartenant pas au personnel de l'entreprise (VIE)	VIE1, VIE2, VIE3, VIE4, VIE5, VIE7, VII4, VII5
Qualité de l'information (QI)		QI1, QI4, QI5, QI6, QI8, QI9