

CHAPITRE

1

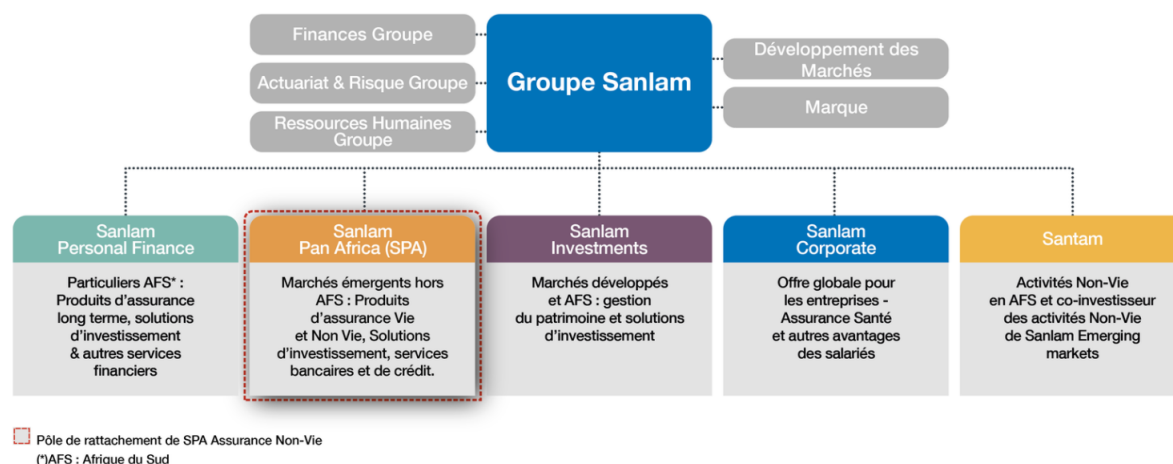
CADRE INSTITUTIONNEL DE L'ÉTUDE

1.1 CADRE CONTEXTUEL

1.1.1 PRÉSENTATION DE L'ORGANISME D'ACCUEIL

Sanlam Group

Fondée en 1918 en tant que compagnie d'assurance vie, Sanlam (South African National Life Assurance Company Limited) est aujourd'hui un groupe leader des services financiers diversifiés, basé en Afrique du Sud. Il déploie ses activités à travers l'ensemble du continent africain ainsi qu'en Malaisie, aux Etats-Unis, en France, en Suisse, en Inde et en Australie. Groupe financier de référence coté à la bourse de Johannesburg, Sanlam propose une offre de solutions financières complètes et personnalisées dans tous les segments du marché, à travers ses 5 pôles d'activités : Sanlam Personal Finance, Sanlam Pan Africa, Sanlam Investments, Sanlam Corporate et Santam (South African National Trust and Assurance Company Limited).



En Afrique, Sanlam est le plus grand groupe d'assurance en termes de capitalisation boursière. C'est également l'un des plus grands groupes d'assurance ayant une couverture internationale, dans le monde, en termes de présence, avec une présence directe et indirecte dans 44 pays, à l'exception de l'Afrique du Sud. Elle est de ce fait l'entreprise d'assurance la plus présente en Afrique avec notamment une présence directe dans 32 pays. Sanlam Group, fournit des solutions et produits financiers aux clients particuliers et aux entreprises à savoir :

- > planification financière et conseil ;
- > assurance (Incendies Accidents et Risques Divers, Vie, spécialiste et réassurance) ;
- > gestion de patrimoine / intermédiaire boursier ;
- > retraite/ gestions de fonds / santé ;
- > asset management.

À Sanlam Group, on développe une vision portée sur la création de valeur pour le client, placé au cœur de toute stratégie de développement. Au fil des années, Sanlam s'est érigé en tant qu'acteur de référence des services financiers dans les marchés émergents, en Afrique et en Asie. Par ailleurs, il ambitionne de consolider ses positions sur le segment des solutions d'investissement sur les marchés développés. Sa vision stratégique est de créer de la valeur durable pour toutes les parties prenantes. Elle se décline sur 3 axes principaux :

- > pour l'Afrique du Sud : être leader dans la gestion de patrimoine, le management et la protection ;
- > pour l'Afrique, l'Inde, la Malaisie et le Liban : être un Groupe de services financiers panafricain de premier plan avec une présence significative en Inde et en Malaisie ;
- > pour les marchés développés : se positionner sur la niche de gestion du patrimoine et de placements sur des marchés développés spécifiques.



Sanlam Pan-Africa

Sanlam Pan-Africa (SPA) est le pôle d'activités de Sanlam Group chargé de la gestion des services financiers sur les marchés émergents (hors Afrique du Sud). SPA assure ainsi le développement et le déploiement d'une gamme diversifiée de produits d'assurance Vie et Non-Vie, ainsi que des solutions d'investissement, des services bancaires et de crédit à la consommation, la bancassurance, la gestion d'actifs et les produits d'assurances Non-Vie spécialisées. L'Afrique est aujourd'hui une composante fondamentale de la vision de Sanlam Group, d'où la mission stratégique fondamentale dévolue à SPA : ériger un groupe de services financiers panafricain de premier plan. Fidèle à la vision et aux orientations du groupe, SPA ambitionne de se hisser au rang de leader sur chaque territoire d'implantation. Ceci en capitalisant sur les synergies-groupe, son expertise-métier ainsi que la connaissance de ses marchés respectifs. S'y ajoute son offre complète de produits dédiés aussi bien aux entreprises qu'aux particuliers & professionnels. Disposant de la première empreinte panafricaine qui lui garantit un rayonnement continental unique, SPA se positionne notamment en tant que partenaire privilégié des multinationales et autres réseaux de distribution.

Digital Factory (Not Yet Done)

1.1.2 QUALITÉ DES DONNÉES DANS LE SECTEUR DE L'ASSURANCE

Grande consommatrice de données, le secteur de l'assurance est un secteur qui a pour principale mission de fournir une prestation lors de la survenance d'un événement incertain et aléatoire souvent appelé «*risque*» : il s'agit d'une assurance. La prestation, généralement financière, peut être destinée à un individu, une association ou une entreprise, en échange de la perception d'une cotisation ou prime. L'assurance se définit également comme étant une opération par laquelle l'assuré transfère ses risques à l'assureur en contrepartie du paiement d'une prime.

Bien qu'étant un secteur majeur dans les activités économiques d'un pays, l'assurance se distingue de la majorité de ces dernières par l'inversion du cycle de production. En effet, dans le cas de l'assurance, il est impossible aux compagnies de savoir avec certitude, combien la prestation qu'elles vendent leur coûtera, la prime étant payée par le client avant que la prestation (indemnisation en cas de sinistre) n'ait été fournie par l'assureur. Ainsi, pour fixer le montant de sa prime, l'assureur ne peut se baser que sur des études statistiques lui permettant de se faire une idée de combien lui coûtera sa prestation (indemnisation du sinistre) en analysant par exemple le taux de sinistralité et le montant moyen des sinistres des années passées. Cela ne donne pas pour autant la certitude qu'il n'aura pas à faire face à de plus gros coûts (en cas d'augmentation du taux de sinistralité dans l'année en cours, par exemple, ou en cas de survenance d'une catastrophe naturelle imprévisible – tremblement de terre, tempête, épidémie mondiale etc.). Le secteur de l'assurance se trouve ainsi donc dans un environnement hautement

incertain et malgré tout concurrentiel. La fiabilité des systèmes d'informations et la qualité des données doivent alors constituer un objectif permanent.

Ce sujet, est au cœur du modèle d'affaires des assureurs. En effet, si la parfaite maîtrise des systèmes d'information et des dispositifs de sécurité associés sont souvent des objectifs identifiés comme stratégiques ; cette maîtrise est vitale pour une bonne gestion et surtout une bonne performance dans ce domaine. Elle permet notamment, de dégager des avantages compétitifs :

- > en terme de tarification, grâce à une segmentation plus efficiente avec la possibilité de créer de nouvelles offres et de mieux comprendre le marché ;
- > en terme de gestion des risques, par une optimisation des couvertures et des provisions,
- > sans oublier la création de meilleurs systèmes de détection des fraudes.

L'amélioration de la qualité des données constitue de ce fait, un enjeu majeur pour les organismes assureurs. L'enjeu est crucial à tout niveau : que ce soit pour une bonne appréhension des risques, pour mener les études actuarielles, pour réaliser les tarifications, pour évaluer les provisions ou fiabiliser les modèles, etc.

Les organismes assureurs sont donc, sensibles aux gains de productivité espérés qui pourront se traduire dans la compétition avec les autres acteurs du marché. Et, les défauts de qualité des données sont autant de freins dans cette compétition, ces défauts étant coûteux pour plusieurs raisons. Tout d'abord, ils rendent plus difficiles l'ensemble des travaux de production puisqu'ils complexifient les traitements. De plus, des données de mauvaises qualités sont susceptibles de conduire à une dégradation ou à l'allongement des travaux et des analyses qui en résultent. En effet, selon une étude du Massachusetts Institute of Technology (MIT) [1], la non-qualité occasionne une perte d'argent estimée entre 15% et 25% du chiffre d'affaire total d'une entreprise. De même, dans [2], l'institut Gartner estime que plus de 25% des données critiques des plus grandes entreprises mondiales sont erronées et précise également qu'un tel problème n'a pas une conséquence informatique mais plutôt une conséquence commerciale, se chiffrant en millions de devise monétaire.

Cette perte représentant environ un quart des revenus, s'explique par les mauvais choix stratégiques opérés à partir d'informations erronées, mais aussi par le temps perdu par les services informatiques à traiter ces données inexactes. Ramener au secteur de l'assurance, une non-qualité des données peut nuire aux décisions prises s'agissant aussi bien des exigences réglementaires que des choix de l'entreprise (mauvaise interprétation de la situation actuelle par exemple) [1]. Il urge alors de se pencher plus sérieusement sur cette problématique tout en prenant en compte le contexte technologique. Nous touchons là au cœur de l'activité des compagnies d'assurances. Dans un environnement économique-financier sinistré, il est impensable que les données de l'assureur ne soient pas

à la hauteur des attentes. Une défaillance sur ce volet-là pourrait avoir des conséquences très importantes.


1.2 MANAGEMENT DU PROJET

1.2.1 CAHIER DE CHARGES

Le présent projet s'inscrit dans un projet plus global de gouvernance des données (*data governance*). La Data Gouvernance, ou gouvernance des données, désigne l'ensemble des conventions et process qui régissent la collecte, la documentation, et l'usage des données dans une organisation. Les données prennent une place de plus en plus importante dans les prises de décisions des entreprises. Mais comment réunir les conditions nécessaires à une exploitation saine de ces données ? La Gouvernance des données a pour rôle de s'assurer de la qualité et de la sécurité des données au sein d'une organisation. Pour cela, elle détermine un ensemble de processus, rôles, règles, normes et métriques permettant d'assurer une utilisation efficace et efficiente des informations, dans le but d'aider les entreprises à atteindre leurs objectifs. Elle définit les procédures et les responsabilités garantissant la qualité et la sécurité des données au sein d'une entreprise ou d'une organisation. Elle définit également qui peut effectuer quelle action, sur quelles données, dans quelle situation et selon quelle méthode. Une stratégie de gouvernance des données claire est fondamentale pour toute organisation traitant les *big data*, et explique comment la société peut bénéficier de procédures et de responsabilités communes et cohérentes. La qualité des données est à l'origine de la plupart des activités de gouvernance des données. La précision, l'exhaustivité et l'homogénéité des sources de données sont les piliers essentiels d'initiatives réussies. Également appelée nettoyage des données, l'épuration des données concourt également à assurer la qualité des données.

À la Digital Factory de Saham Maroc, le besoin s'est fait ressentir de mettre en place dans le cadre de la gouvernance des données un outil permettant d'industrialiser la gestion de la qualité des données au sein de Saham Assurance. Cet outil viendra remplacer les test manuels (SQL) et les explorations effectués sans industrialisation.

Objectif général

L'objectif générale du présent projet est de détecter et de corriger les problèmes de qualité dans le cadre du projet socle de données de la Digital Factory. Il est de ce fait subdiviser en deux volets : le volet outil et le volet qualité de donnée. 

> **Objectifs spécifiques : Volet outil** Il s'agit ici de :

1. prendre en main l'outil d'audit de la qualité des données Apache Griffin ;

2. faire une analyse comparative entre Apache Griffin et Great Expectations ;
3. établir la connexion avec les différentes bases de données de l'équipe socle de données.

> **Objectifs spécifiques : Volet qualité de données** Il s'agira ici de se servir de l'outil pour identifier les différents problèmes de qualités de données et le cas échéant proposer des mesures de redressement. Plus précisément, il faudra faire :

1. une revue des différentes données du socle de données et détecter des incohérences par rapport aux différentes sources de données / extractions utilisées par le métier ;
2. une priorisation des champs à mettre en qualité en urgence ;
3. une implémentation d'algorithmes de redressement des données quand cela est possible ;
4. éventuellement une injection d'open data pour une mise en qualité des données prioritaires.

1.2.2 CHOIX DE LA MÉTHODE AGILE

La méthode Agile est une méthodologie de gestion de projet. Il s'agit d'une organisation de travail en cycles courts, permettant aux équipes de développement de gérer un produit de manière souple, adaptative et itérative. Son but est d'améliorer leur process et réduire leur taux d'échec. Pour cela, elle place le client au cœur du projet et s'adapte tout le long du fil du projet. De plus, au lieu de planifier le projet de A à Z dès le départ, ce qui laisse peu de place aux imprévus, des objectifs courts sont fixés, par exemple à deux ou trois semaines. Le projet est divisé en sous-projets et l'on ne passe au suivant que lorsque le précédent est réglé.

Le principal avantage est la flexibilité, la possibilité de s'adapter en fonction des nouvelles exigences du client ou des évolutions du marché. Cela permet aussi un meilleur contrôle des coûts, puisqu'un point budget peut être fait à chaque étape. Les effets positifs se font enfin ressentir sur la motivation : les collaborateurs voient les tâches avancer, au lieu de passer plusieurs mois d'affilé sur un gros dossier qui semble stagner et de redouter la date limite finale.

Il existe en réalité plusieurs méthodes qui ont toutes un point commun : elles découlent toutes du Manifeste Agile. Scrum est aujourd'hui l'approche Agile la plus répandue, il s'agit plus précisément d'un cadre méthodologique plutôt que d'une méthode. Elle est d'ailleurs celle implémentée par la Digital Factory. Son objectif est d'améliorer la productivité des équipes, tout en permettant une optimisation du produit grâce à des retours d'expériences réguliers avec les utilisateurs finaux. De plus, Scrum est une pratique Agile élémentaire qui permet également une mise à l'échelle, autrement dit le

déploiement progressif de l'agilité à l'échelle de l'entreprise. Scrum est constitué d'une définition des rôles, de réunions et d'artefacts [3] [4].

Rôles

1. **Le «*Product Owner*» ou «*PO*»** : il porte la vision du produit à réaliser et il s'agit donc généralement d'un expert métier. Il travaille en collaboration directe avec l'équipe de développement et a notamment la charge de remplir le «*Product Backlog*» et de déterminer la priorité des «*user stories*» (phrase simple, rédigée dans un langage courant, qui permet de décrire avec suffisamment de précision le contenu d'une fonctionnalité à développer) à réaliser. Il peut être interne ou externe, même s'il s'agit généralement du client.
2. **Le «*Scrum Master*» ou «*SM*»** : Il s'agit d'un membre à part entière de l'équipe de projet, et il doit maîtriser Scrum car il est chargé de s'assurer que la méthodologie est correctement appliquée. Il ne faut surtout pas le confondre avec un chef de projet. Son rôle n'est pas de diriger, mais de faciliter le dialogue et le travail entre les différents intervenants, de façon à ce que l'équipe soit pleinement productive. Il agit donc plutôt comme un coach, aussi bien auprès du «*Product Owner*» que de l'équipe de développement. Il doit être un bon communicant et faire preuve de pédagogie, afin de pouvoir résoudre les éventuels conflits qui pourraient apparaître durant le projet. Il anime généralement les différentes réunions, qu'il s'agisse du «*scrum daily meeting*», de la revue de sprint ou encore de la rétrospective. Généralement, le rôle de «*Scrum Master*» change régulièrement au sein de l'équipe projet, chacun pouvant le devenir à tour de rôle.
3. **L'équipe de développement** : généralement composée de 4 à 6 personnes, elle est chargée de transformer les besoins exprimés par le «*Product Owner*» sous la forme de «*user stories*» en fonctionnalités réelles, opérationnelles et utilisables. L'équipe est généralement composée de plusieurs profils, ne se limitant pas à des développeurs, et peut intégrer des architectes, un administrateur de base de données, un graphiste, un ergonomiste ou encore un ingénieur système ou réseau, en fonction des besoins.

Réunions

1. **Planification du *Sprint*** (*Sprint* = itération) : au cours de cette réunion, l'équipe de développement sélectionne les éléments prioritaires du «*Product Backlog*» (liste ordonnancée des exigences fonctionnelles et non fonctionnelles du projet) qu'elle pense pouvoir réaliser au cours du *Sprint* (en accord avec le «*Product Owner*»).
2. **Revue de *Sprint*** : au cours de cette réunion qui a lieu à la fin du *Sprint*, l'équipe de développement présente les fonctionnalités terminées au cours du

sprint et recueille les feedbacks du «*Product Owner*» et des utilisateurs finaux. C'est également le moment d'anticiper le périmètre des prochains *Sprints* et d'ajuster au besoin la planification de *release* (nombre de *Sprints* restants).

3. **Rétrospective de *Sprint*** : la rétrospective qui a généralement lieu après la revue de sprint est l'occasion de s'améliorer (productivité, qualité, efficacité, conditions de travail, etc) à la lueur du "vécu" sur le *Sprint* écoulé (principe d'amélioration continue).
4. **Mêlée quotidienne** : il s'agit d'une réunion de synchronisation de l'équipe de développement qui se fait debout (elle est aussi appelée «*stand up meeting*») en 15 minutes maximum au cours de laquelle chacun répond principalement à 3 questions : «*Qu'est ce que j'ai terminé depuis la dernière mêlée ? Qu'est ce que j'aurai terminé d'ici la prochaine mêlée ? Quels obstacles me retardent ?*».

Artefacts

1. **Le *Sprint*** : il s'agit d'une période pendant laquelle un travail spécifique doit être mené à bien avant de faire l'objet d'une révision.
2. **Le *Product Backlog*** : Il s'agit d'une liste hiérarchisée des exigences initiales du client concernant le produit à réaliser.
3. **Le *Sprint Backlog*** : c'est le plan détaillé de la réalisation de l'objectif du *Sprint*, défini lors de la réunion de planification du *Sprint*.
4. **Le *Task Board*** : outil central du *Sprint* scrum, ce tableau de bord du projet permet de suivre en temps réel la progression de la réalisation des différentes tâches. Il est composé de trois colonnes comportant les tâches à faire, les tâches en cours et les tâches terminées. Il est généralement situé dans un endroit visible de l'ensemble de l'équipe (un tableau accroché au mur ou affiché sur un grand écran) et est actualisé directement par les développeurs dès que l'activité sur laquelle ils travaillent évolue.
5. **Le *Burndown Chart*** : il s'agit d'un graphique simple permettant de visualiser le degré d'avancement de chacune des tâches. Il permet de fournir à l'ensemble de l'équipe une vision claire et actualisée de l'état d'avancement des travaux et de la quantité de travail restante. Il est généralement mis à jour lors de la réunion quotidienne.

1.2.3 ORGANISATION ET PLANNING DU PROJET (DÉCRIRE L'IMPLÉMENTATION DU SCRUM DANS LE CADRE DU PROJET ET LES SPRINTS –NOT YET DONE)

CHAPITRE

2

CADRE THÉORIQUE ET MÉTHODOLOGIQUE DE L'ÉTUDE

2.1 CLARIFICATION CONCEPTUELLE

2.1.1 BIG DATA

Le *big data* (données massives), fait désormais partie du quotidien de toutes les entreprises. Bien qu'omniprésentes dans notre actualité, les données massives constituent des phénomènes nouveaux et parfois difficiles à définir. Le terme *big data* a été popularisé par John Mashey, informaticien chez Silicon Graphics dans les années 1990 [5]. Ce dernier, faisait référence, aux bases de données trop grandes et complexes pour être étudiées avec les méthodes statistiques traditionnelles – et, par extension, à tous les nouveaux outils d'analyse de ces données. En 2001, Douglas Laney a analysé cette nouvelle tendance à travers une liste très simple de trois « V », ensuite élargie à cinq « V » [5] [6] :

- > le volume : pour désigner la grande quantité de données ou d'informations contenue dans ces bases de données ;
- > la vélocité : également appelée vitesse correspondant à la rapidité à laquelle les données sont générées, collectées et circulent pour transmission et analyse ;
- > la variété : pour désigner la multiplicité des types de données disponibles, autrement dit les différences de natures, formats et structures ;

- > la valeur : pour désigner la capacité de ces données à générer du profit ; chaque donnée devant apporter une valeur ajoutée à l'entreprise ;
- > la véracité : qui désigne la fiabilité des données et qui est essentielle pour pouvoir en tirer profit et les transformer en information utilisable dans l'entreprise.

Ces cinq « V » permettent donc de décrire et de caractériser les *big data* . Dans un contexte où, le volume de données augmente de manière exponentielle, ce sont généralement les entreprises qui commencent à tirer des avantages incroyables de leurs *big data*. Selon les gestionnaires et les économistes, les entreprises qui ne s'intéressent pas sérieusement au *big data* risquent d'être pénalisées et écartées [7]. Mais, pour pouvoir en tirer pleinement profit, des données de haute qualité sont la condition préalable à l'analyse et à l'utilisation du *big data* ainsi qu'à la garantie de leur valeur. On fait ainsi référence à leur véracité. Au vu de la masse des données, il y a beaucoup de chose à affiner. D'où la nécessité de prendre des mesures de précaution pour minimiser les biais liés au manque de fiabilité des données. Les méthodes permettant d'améliorer et de garantir la qualité des *big data* sont essentielles pour prendre des décisions commerciales précises, efficaces et fiables. Mais qu'est-ce qu'une donnée de qualité ?

2.1.2 QUALITÉ DES DONNÉES

Définir la qualité des données n'est pas une opération aisée, il est souvent plus simple de tenter de définir la non-qualité [8]. En effet, il est plus facile dans ce cas d'identifier ce qu'on ne souhaite pas avoir dans nos données. Toutefois, il conviendrait de faire un point conceptuel et de définir ce qu'on entend par qualité, information et donnée avant de revenir sur la définition de la qualité des données en elle même.

À cet effet, d'après [9], la qualité pourrait se définir comme le degré auquel un ensemble de caractéristiques inhérentes à un objet répond aux exigences. On parle également de conformité aux exigences. Toujours selon [9], l'exigence se définit comme un besoin ou une attente énoncée ; généralement implicite ou obligatoire.

Dans [8], on retiendra que les données « *sont des faits et des statistiques qui peuvent être quantifiées, mesurées, comptées, et stockées* » et que l'information quant à elle « *est un ensemble de données organisées selon une ontologie qui définit les relations entre certains sujets* ». La qualité des données pourrait donc se définir simplement comme le degré auquel un ensemble de caractéristiques inhérentes aux données (des faits ou des statistiques) répond aux exigences [9]. Plus précisément, Wang et Strong(1996) cité dans [10], définissent la qualité comme l'aptitude à l'emploi et proposent que le jugement de la qualité des données dépende des consommateurs de données. Il s'agit bien souvent, d'avoir à disposition des données exemptes d'erreurs, d'incohérences, de redondances, de formatage médiocre et d'autres problèmes susceptibles d'empêcher une utilisation aisée [11].

Bien que la littérature diffère sur la définition exacte de la qualité des données, tous les auteurs s'accordent sur le fait que : la qualité des données dépend non seulement de ses caractéristiques propres, mais aussi de l'environnement dans lequel ces données sont utilisées [10]. Le fait qu'un ensemble de données contienne ou non des informations de qualité est déterminé en dernier ressort par l'objectif à atteindre. Cela révèle ainsi le caractère hautement subjectif de cette notion. Aussi peut-on lire dans [12] que :

- > pour le consommateur, des données de qualité sont des données :
 - qui sont aptes à être utilisées ;
 - qui répondent à ses attentes ou les dépassent ;
 - qui satisfont aux exigences de leur utilisation prévue ;

- > pour l'entreprise, des données de qualité sont des données :
 - qui sont aptes à être utilisées dans leurs rôles opérationnels, décisionnels et autres prévus ou qui présentent une conformité aux normes qui ont été fixées ;
 - qui sont adaptées aux utilisations prévues dans le cadre des opérations, de la prise de décision et de la planification ;
 - capable de satisfaire les exigences commerciales, systémiques et techniques déclarées de l'entreprise ;

- > et du point de vue des normes, la qualité des données est mesurée par :
 - le degré auquel un ensemble de caractéristiques inhérentes (dimensions de qualité) d'un objet (données) répond aux exigences ;
 - l'utilité, la précision et l'exactitude des données pour leur utilisation.

Cette multiplicité de point de vue, témoigne aisément de la nature subjective de cette notion et se justifie par le fait qu'avec l'avènement du *big data* , contrairement au passé, les utilisateurs de données ne sont pas nécessairement les producteurs de ces données. Abondant dans le même sens, sur la définition de la qualité des données, le NISS (National Institute of Statistical Sciences) [13] identifie sept(7) principes clés permettant de saisir sa quitéssence, nous en citerons 5. Ils s'énoncent comme suit :

1. les données peuvent être considéré comme un produit, avec des clients pour lesquels elles ont à la fois un coût et une valeur ;
2. étant un produit, les données ont une qualité résultant du processus par lequel elles ont été générées ;
3. cette qualité dépend de multiples facteurs dont (au moins) l'objectif pour lequel les données sont utilisées, l'utilisateur, le moment, ... ;
4. la qualité des données est multidimensionnelle et comprend des attributs de qualité intrinsèque, d'accessibilité, de contexte et de représentation ;
5. en principe, la qualité des données peut être mesurée et améliorée.

Somme toute, cette clarification conceptuelle montre sans équivoque que la qualité des

données fait référence à la capacité d'un ensemble de données à remplir une fonction, ou des attentes prévues. En allant plus loin, ces attentes, spécifications et exigences sont énoncées en termes de caractéristiques ou de dimensions, ce qui la rends mesurable. Décrite comme un concept à dimensions multiples, chaque dimension de la qualité des données fait alors référence à un aspect spécifique de celle-ci. Une dimension de qualité des données (QD) est un terme utilisé par les professionnels de la gestion des données pour décrire une caractéristique des données qui peut être mesurée ou évaluée par rapport à des attentes prédéfinies [14].

En conjonction avec la discussion sur les dimensions de la qualité des données, la définition de mesures spécifiques est nécessaire pour leur quantification dans la pratique, on parle alors de métrique. Une métrique de qualité des données selon [15], est une fonction qui associe une dimension de qualité à une valeur numérique, ce qui permet d'interpréter la réalisation d'une dimension. Une telle métrique peut être mesurée à différents niveaux d'agrégation : au niveau des valeurs, des colonnes ou des attributs, des tuples ou des enregistrements, des tables ou des relations, ainsi qu'au niveau de la base de données. À la faveur de cette clarification conceptuelle, on pourrait alors se demander comment mesurer la qualité de nos données ?

2.2 REVUE DE LITTÉRATURE

Être apte à satisfaire l'utilisation prévu : c'est ainsi que peut se résumer la notion de qualité des données. Elle est bien souvent à tort réduite à une mesure d'exactitude : par exemple, le nom de la ville "Abidjan" mal orthographié en "Abdjan" serait le seul type de problème rencontré. En effet, on considère qu'une donnée est de mauvaise qualité si des fautes de frappe sont présentes ou si des valeurs erronées s'y trouvent. Mais cela ne se résume pas qu'à cela : la qualité des données ne se limite pas qu'à l'exactitude. D'autres dimensions non moins importantes sont nécessaires pour pleinement la caractériser. Une étude de la qualité des données ne saurait alors se faire sans avoir clairement identifier des dimensions cibles. Il s'agira ici de faire d'entrée, une revue des différents événements qui peuvent entraver la qualité des données, pour ensuite déboucher sur quelques dimensions présentent dans la littérature et enfin les pistes de correctifs proposées.

2.2.1 DÉFIS DE LA QUALITÉ DES *BIG DATA* ET LES SOURCES DE NON QUALITÉ

Le *big data*, loin de ce qu'il pourrait laisser imaginé, n'est évidemment pas qu'une simple question de taille. L'extraction et le traitement de données de haute qualité, massives, variables et compliquées devient une question urgente. En effet, l'ère de l'information moderne produit des tonnes de données à chaque seconde. Avec l'avènement

des téléphones intelligents et de l'internet, des quantités phénoménales de données sont créés. Avec les données volumineuses viennent de plus grandes responsabilités et de plus grands défis. À mesure que le volume et la variété des données augmentent, il devient plus difficile de contrôler chaque entrée de données afin de s'assurer de leur bonne qualité. Cai et Zhu [10], citent les défis auxquels le *big data* est confronté de nos jours :

- > la diversité des sources de données, entraîne une abondance de types de données et de structures de données complexes, augmentant ainsi la difficulté de leur intégration ;
- > un volume énorme de données, rend difficile l'appréciation de la qualité des données dans un délai raisonnable ;
- > avec les données qui arrivent en temps réel, comme les données en continu (*streaming*) ou l'internet des objets, il devient plus difficile d'évaluer la qualité. Lorsqu'elle ne peut être évaluée, les données ne sont pas fiables, ce qui entraîne une prise de décision imprécise ;
- > l'inexistence de normes unifiées et approuvées de qualité des données ;
- > de plus, la recherche sur la qualité des données du *big data* est relativement récente.

Ce sont ainsi, les défis qu'une entreprise à la recherche d'un cadre méthodologique d'analyse de la qualité de ses *big data* doit relever. En effet, aucune entreprise ne saurait aspirer à une réelle expansion sans intégrer le *big data* et une stratégie de gouvernance des données. C'est pour cela qu'il est important d'analyser les dimensions de qualité existantes, largement utilisées pour évaluer la qualité des données, afin de déterminer dans quelles mesures elles sont applicables au *big data* . Les problèmes de qualité des données surviennent lorsque les exigences qualité ne sont pas satisfaites. Ces problèmes sont dûs à plusieurs facteurs ou processus. Il s'agit principalement comme on peut le lire dans [16], [17],[8], [18] de :

- > l'entrée des données par l'homme ; en effet, un champ de saisie laissé libre induit un risque d'erreur élevé rendant difficile la mesure de la qualité de la donnée ;
- > la dégradation de la donnée dans les chaînes et processus de traitement : troncatures, caractères mal interprétés, erreurs lors des conversions, rejets non traités, absence de contrôles sur le format, passage d'un système d'encodage à un autre, ... ;
- > la corruption volontaire ou intentionnelle des données à des fins malhonnêtes ;
- > des données non actualisées qui deviennent une source d'inexactitude avec le temps ;
- > des défauts de conception qui en laissant subsister une ambiguïté sémantique peuvent amener à des erreurs de valorisation et également d'interprétation de la donnée par la suite ;
- > des définitions d'attributs pas suffisamment bien structurées ou normalisées lors de la modélisation conceptuelle des données ainsi qu'un schéma non valide et/ou un manque de contraintes d'intégrité et de procédure pour maintenir la cohérence

- des données ;
- > l'intégration de données provenant de sources externes et faisant l'objet de contradiction ou d'incohérences avec celles locales.

La complexité de l'organisation d'une entreprise et la multiplicité des chaînes de traitement augmentent le risque de tous ces facteurs. Combinée avec les caractéristiques majeurs du *big data*, ils laissent ainsi surgir de nouveaux défis. Plus l'anomalie ou la non-qualité est détectée tôt, suivie et corrigée à la source, plus l'ensemble du patrimoine global de données sera fiable. D'où la nécessité d'un outil d'audit de la qualité des données. Mais avant de penser à un outil il serait judicieux de se pencher sur les aspects de la qualité qu'on souhaite abordé.

2.2.2 DIMENSIONS DE LA QUALITÉ DES DONNÉES

La qualité des données est un concept aux multiples facettes, et différentes dimensions concourent à la définir. Différents travaux dans la littérature, ont mis en exergue un ensemble de dimensions dans le but de mesurer la qualité des données. Afin de garantir des données d'une certaine qualité aux utilisateurs, et des décisions pertinentes, chaque organisation se doit de définir les dimensions qu'elle utilisera dans son processus. Ben Salem [17], va plus loin en disant que chaque organisme doit créer ses propres définitions opérationnelles en fonction des objectifs et priorités, afin de définir des indicateurs pour chacune des dimensions choisies, et vérifier par des mesures régulières leur évolution dans le temps. Les données doivent avoir la qualité nécessaire pour supporter le type d'utilisation prévue.

Pour les données structurées, la littérature offre plusieurs contributions qui proposent différentes dimensions. Mais, les *big data* posent de nouveaux défis liés à leurs principales caractéristiques : volume, vitesse et variété. En particulier, afin d'aborder les questions de volume et de vitesse, il est nécessaire de repenser les méthodes d'évaluation pour exploiter les scénarios de calcul parallèle et pour réduire l'espace de calcul. La littérature sur les dimensions de la qualité des données est plutôt féconde bien que récente. Plusieurs auteurs se sont prêtés à cet exercice de revue des différentes dimensions utilisées pour attester de la qualité d'un ensemble de données. Notre étude porte principalement sur l'utilisation d'Apache Griffin, comme outil d'audit de la qualité des données. Apache Griffin, s'inspire de la définition que donne la Data Administration Management Association United Kingdom (DAMA UK) [14], qui identifie six principales dimensions pour l'évaluation de la qualité des données [14], [15] :

1. l'exhaustivité ou complétude (*completeness*) : qui mesure l'absence de valeurs manquantes (chaînes de caractères nulles ou vides, données numériques manquantes). Il est à noter que le nombre de valeurs manquantes peut être calculé de différentes manières, soit en ne prenant en compte que les vraies valeurs manquantes (i.e. null), soit les valeurs par défaut ou une entrée textuelle mentionnant

"NaN" (c'est-à-dire, *Not a Number*);

2. l'unicité (*uniqueness*) : cette dimension permet l'analyse des valeurs uniques. L'unicité est l'inverse d'une évaluation des doublons. Ces deux aspects représentent deux faces d'une même pièce;
3. l'actualité (*timeliness*) : décrit le degré de fraîcheur des données pour la tâche à accomplir et est étroitement liée aux notions de fréquence de mise à jour des données et de volatilité (vitesse à laquelle les données deviennent non pertinentes). Une autre définition indique que l'actualité peut être interprétée comme la probabilité qu'un attribut soit toujours à jour [15]. Elle se mesure en détectant le taux de valeurs obsolètes dans la base de données par rapport à une date prédéfinie;
4. la validité (*validity*) : une donnée est jugée valide si elle est conforme aux exigences (format, type, plage) de sa définition. Il s'agit d'une comparaison entre les données et les métadonnées ou la documentation;
5. l'exactitude (*accuracy*) : bien que parfois décrite comme la dimension la plus importante, présente un certain nombre de définitions différentes. Cette dimension évalue la mesure dans laquelle un système d'information décrit correctement ou se rapproche du monde réel qu'il est censé modéliser. Elle se mesure en détectant le taux de valeurs correctes dans la base de données au regard d'une source de données définie comme référence;
6. la cohérence (*consistency*) : il existe également différentes définitions de la dimension cohérence. Selon Batini et Scannapieco cité par [15], la cohérence capture la violation des règles sémantiques définies sur les données, où les éléments peuvent être des tuples de tables relationnelles ou des enregistrements dans un fichier. Les contraintes d'intégrité de la théorie relationnelle sont un exemple de telles règles. Elle se mesure donc par rapport à l'ensemble des contraintes en détectant les données qui ne les satisfont pas.

Dans une tentative de synthèse, [19] fait remarquer en 2004 que quel que soit le domaine d'application, les mesures de qualité de données les plus fréquemment mentionnées dans la littérature, sont l'exactitude, la complétude, l'actualité, la cohérence. Mais à l'ère du *big data*, il revient de se demander si ces dimensions sont toujours d'actualité ou nécessitent des ajustements. Ainsi, pour évaluer la qualité des *big data*, Cai et Zhu [10], proposent une hiérarchie de dimensions à deux niveaux :

- > la disponibilité de la donnée caractérisée par :
 - l'accessibilité : existe-t-il des facilités d'accès aux données ?;
 - l'actualité;
 - l'autorisation : a-t-on le droit d'utiliser ces données ? a-t-on les accès nécessaires ?;
- > l'utilisabilité de la donnée caractérisée par :

- l'existence d'une documentation ;
 - la crédibilité : concerne la fiabilité de la source de données, la normalisation des données et le moment où les données sont produites ;
 - l'existence de méta-données ;
- > la fiabilité de la donnée caractérisée par quatre des six dimensions identifiées par [14]
- l'exactitude ;
 - l'intégrité = validité [14] ;
 - la cohérence ;
 - la complétude ;
 - l'auditabilité : l'auditabilité signifie que les auditeurs peuvent évaluer équitablement l'exactitude et l'intégrité des données dans des limites rationnelles de temps et de main-d'œuvre au cours de la phase d'utilisation des données ;
- > la pertinence de la donnée caractérisée par :
- l'aptitude de la donnée à satisfaire son utilisateur ;
- > la qualité de la présentation caractérisée par :
- la lisibilité : est ce que la donnée est bien définie suivant une terminologie connus et courante ? ;
 - la structuration : la structure fait référence à la difficulté à transformer les données semi-structurées ou non structurées en données structurées.

Plusieurs autres auteurs, ont également proposé des cadres méthodologiques d'analyse de la qualité des *big data* . Ramasamy et Chowdhury [20], en font un résumé. En plus de l'accessibilité [10], la lisibilité [10], la crédibilité et la confiance [20] (la crédibilité [10]), la cohésion [20] (la cohérence [14], [10]) ainsi que la confidentialité [20] (l'autorisation [10]), ils font ressortir quatre autres dimensions jugées importantes :

- > le pédigrée ou la lignée : cette dimension permet de connaître la source des données afin de pouvoir corriger les éventuelles incohérences directement là bas ;
- > la capacité d'analyse en temps réel ;
- > la redondance : elle fait référence à la capacité à représenter le monde réel sans répétition des informations ;
- > le volume : cette dimension permet d'analyser le volume des données extrait.

Au regard de ces dimensions sur la qualité des *big data* , on constate que plusieurs dimensions prennent en compte le fait que les *big data* ne sont pas forcément produite en entreprise mais pour la plupart proviennent de sources externes et que cela à un impact sur la qualité. Toutefois, les dimensions retenues pour les données structurées demeurent applicable dans le contexte des *big data* .