# La qualité des données en assurance

Amal Dhannoo

le cnam

intd

# La qualité des données en assurance

Mémoire

pour l'obtention du

**Master Sciences humaines et sociales**

**mention humanités numériques**

**Parcours Mégadonnées et analyse sociale** (MEDAS)

Amal DHANNOO

Date et lieu de la soutenance

- 09/09/2020
- Visioconférence Skype

Membres du jury

- Béatrice ARRUABARRENA, Présidente du Jury
- Ghislaine CHARTRON, Membre du jury – CNAM
- Karim KILANI, Directeur de mémoire – CNAM
- Céline DESSONS, Maître d'apprentissage – Generali

Promotion (2018-2020)

« *Et parce qu'une donnée vit, elle naît, grandit, se reproduit, voyage, et prend sa retraite*. »

Jean-Laurent GRANIER, Président Directeur Générale de Generali France

## Résumé

Ce mémoire se concentre sur la façon dont la qualité des données est évaluée dans le milieu de l'assurance. Actuellement, peu d'entreprises disposent d'un service dédié pour mesurer la qualité de leurs données, ce qui peut entraîner une perte de bénéfice pour l'entreprise. En effet, selon une étude de Jack E. OLSEN dans son livre *Data Quality: The Accuracy Dimensions,* les données de mauvaise qualité peuvent conduire à une perte de profit de 15 à 20%, ce qui représente un énorme manque à gagner. La plupart des études qui ont été réalisées jusqu'à présent suggèrent la mise en place d'un service ou d'une équipe qui peut, grâce à ses compétences techniques, mesurer la qualité des données de l'entreprise en utilisant différentes méthodes. Le projet mis en œuvre porte sur l'étude de la qualité des données identitaire et contact des assurés chez Generali. Les méthodes utilisées dans cette étude sont dans un premier temps de calculer le taux de remplissage de chaque champ, et dans un deuxième temps, de relever les anomalies des données.



Mots clés : Algorithme, Business Intelligence, Hadoop, PySpark, Qualité de données, Tableaux de bord



## Abstract

This research focuses on how data quality does help in decision making in the insurance industry. Currently, few companies have a dedicated service to measure the quality of their data, and this can cause the company to lose profit. Indeed, according to a study by Jack E. OLSEN in his book *Data Quality: The Accuracy Dimension*, poor quality data can lead to a loss of profit of 15% to 20%, which will represent a huge shortfall. Most of the studies that have been carried out so far suggest setting up a service or team that can, thanks to its technical skills, measure the quality of the company's data using different methods. This project is about the study of the quality of identity and contact data at Generali. The methods used in this study are first to calculate the filling rate of each field, and secondly, to identify the anomalies of the data.



Keywords: Algorithm, Business Intelligence, Data quality, Hadoop, PySpark, Reporting

# Glossaire

CNIL : Commission Nationale Informatique et Libertés

CNR : Contrats Non Réglés

ETL : Extraction, Transformation, Chargement (*Extract, Transform, Load* en anglais)

HDFS : Hadoop Distributed File System

NPAI : N'habite Pas à l'Adresse Indiquée

RCE : Référentiel Clients Entreprise

RGPD : Règlement Général sur Protection de Données à caractère personnel

RNIPP : Registre National d'Identité des Personnes Physiques

SI : Système d'Information

TA : Technique Assurance

# Remerciements

Ce mémoire représente le fruit du croisement des efforts de plusieurs personnes qui, sans leurs précieuses interventions, m'aurait été difficile de réaliser correctement ce travail. Il m'est donc indispensable d'exprimer ma gratitude envers eux.

Je tiens à remercier mon tuteur Karim KILANI pour son accompagnement et ses conseils pour la réalisation de ce mémoire. Il m'a guidé, aidé et encouragé du tout au long de la réalisation de ce mémoire. Ses suggestions et son expérience ont été un atout pour moi.

Je remercie également tout le corps professoral de l'INTD, en particulier Béatrice ARRUABARRENA et Ghislaine CHARTRON pour leurs cours de Management et Qualité des données qui m'ont beaucoup apporté sur le plan professionnel.

J'adresse mes remerciements au service « Opérations, CNR et pilotage délégataires » de Generali France, et en particulier à Delphine BIGOT, Directrice du service, pour son accueil chaleureux.

Je remercie Didier BOTTAIS, Responsable du service et Sandrine BENITO, Manager de l'équipe des Contrats Non Réglés, pour leur confiance et de m'avoir donné la chance d'intégrer leur équipe afin d'y apprendre beaucoup.

Je remercie Céline DESSONS et Romain METAYER, pour leur encadrement et leur pédagogie dont ils ont fait preuve envers moi. Leur patience, leur disponibilité ainsi que leur expérience m'ont été d'un grand soutient tout au long de mes missions, ainsi qu'à l'ensemble du service auprès de qui j'ai eu la chance de travailler.

# Table des matières

# Introduction

Ce mémoire a été réalisé dans le cadre du Master MégaDonnées et Analyse Sociale, au Conservatoire National des Arts et Métiers (CNAM). Cette formation a été réalisé en alternance au sein de Generali France ou j'ai occupé le poste de Data Analyst pendant deux ans.

Au cours de mes deux années d'alternance, j'ai effectué plusieurs missions et participé à plusieurs projets et cela m'a permis de découvrir une partie de l'analyse de données qui m'étais alors inconnue à ce jour : la qualité des données. En effet, tout au long de mes études en statistique et informatique décisionnel, j'ai été amené à effectuer des analyses statistiques sur des données bien constituées. Ainsi, j'ai acquis des compétences sur l'analyse de la qualité des données en développant des algorithmes de détection d'anomalies, mais également le développement de tableaux de bords afin de restituer les informations essentielles des différents projets.

De nos jours, les assureurs possèdent un grand volume de données issus principalement des contrats des assurés. Ces données représentent un atout concurrentiel car elles vont permettre de mieux comprendre les comportements des assurés et par conséquent de calculer au mieux les différents tarifs ou primes des produits à vendre. Les professionnels de la donnée occupent donc un rôle important dans l'entreprise afin d'exploiter au mieux les données et d'en tirer un plus grand nombre d'informations.

C'est ainsi que la gestion de données, et particulièrement l'analyse de la qualité des données représentent un aspect essentiel du Big Data chez les compagnies d'assurance et offrent donc un panel de paramètres beaucoup plus large dans l'apprentissage et la prédiction des données.

A travers ce mémoire, nous verrons dans un premier temps le contexte ayant amené la problématique de la qualité de données. Nous verrons par la suite les différents aspects de la qualité de données ainsi que les opportunités qu'offrent le Big Data. Enfin, nous expliquerons dans la dernière partie les étapes de la mise en place d'un projet d'évaluation de la qualité de données ainsi que les résultats obtenus par l'étude.

# Chapitre 1 : Contexte et problématique

## I – Contexte de Generali

Generali France appartient au Groupe Generali qui a été créé à Trieste en Italie en 1831 sous le nom d'*Assicurazioni Generali*. La première compagnie d'assurance multirisque est actuellement présente dans plus de 50 pays et emploie près de 72 000 collaborateurs afin de se rendre utile auprès de 60 millions de clients.

Generali propose de nombreux produits d'assurance à ses clients que nous pouvons séparer en deux catégories bien définies : les assurances-vie et l'IARD. Ce dernier type couvre bien et les dommages comme, par exemple, les assurances habitations, automobile et autres. L'assurance-vie en revanche protège les personnes à l'aide de placements. Ces placements peuvent être de trois types différents : de l'épargne, de la retraite individuelle ou collective et de la prévoyance individuelle ou collective.

Generali s'est installé à Bordeaux en 1832 ce qui en fait sa plus ancienne implantation étrangère. Generali France résulte du regroupement de l'ensemble des sociétés du territoire français en 2006.

Le groupe français de place troisième sur le marché derrière l'Italie et l'Allemagne avec 13,3 milliards d'euros de chiffre d'affaires en 2019. Le Président-Directeur Général de Generali France, Jean-Laurent Granier en poste depuis le mois de juin 2017 a restructuré les pôles de l'entreprise afin d'atteindre son objectif d'excellence.

### I.I – Organisation de Generali France

Generali France continue sa transformation dans le cadre du projet stratégique Excellence 2022 pour répondre, d'une part, aux évolutions du contexte économique et financier qui intensifient les conditions de concurrence et, d'autre part, aux exigences des clients renforcées notamment par les nouveaux usages digitaux. Notre objectif est de faire de Generali France un assureur multi-spécialiste de référence, excellent en qualité de service, efficace, agile et innovant.

Les exigences de ce projet nous amènent à adapter notre organisation selon les principes directeurs suivants :

- Focaliser, pour plus de responsabilisation et de fluidité dans notre fonctionnement au quotidien ;
- Simplifier, pour plus d'agilité et d'efficacité dans la gouvernance ;
- Intégrer, pour regrouper par marché les compétences nécessaires (offre, gestion et commercialisation) ;

- Se différencier, par l'excellence du service, en pensant l'expérience client de bout en bout ;

- Collaborer, pour maintenir la transversalité en s'appuyant sur l'intelligence collective.

La nouvelle organisation de Generali France s'articule autour de 5 directions de marché, responsables de l'expérience client de bout en bout (offre, commercialisation, gestion et indemnisation) sur le périmètre considéré :

- Particuliers IARD et prévoyance ;

- Épargne et gestion de patrimoine ;

- Pro-PE et entreprises IARD ;

- Protection sociale des Pro-PE ;

- Protection sociale des entreprises.

La logique organisationnelle poursuivie est donc que chaque direction de marché :

- Intègre les équipes offre, gestion, indemnisation et commercialisation spécifiques à son activité ;

- Intègre le ou les réseaux de distribution dédiés à son activité ;

- S'appuie sur la distribution pour les réseaux propriétaires multi-marchés (réseau salariés Generali et réseau agents) ;

- S'appuie sur les fonctions supports des agents généraux de l'entreprise pour les compétences mutualisées.

Cependant, afin de conserver une taille critique pour certaines activités, les compétences rares sont mutualisées et rattachées à la direction de marché techniquement la plus pertinente. Elles peuvent être ainsi mobilisées pour le compte commun des autres directions. De plus, la Distribution est garante de la stratégie du modèle économique de la distribution pour l'ensemble de l'Entreprise.

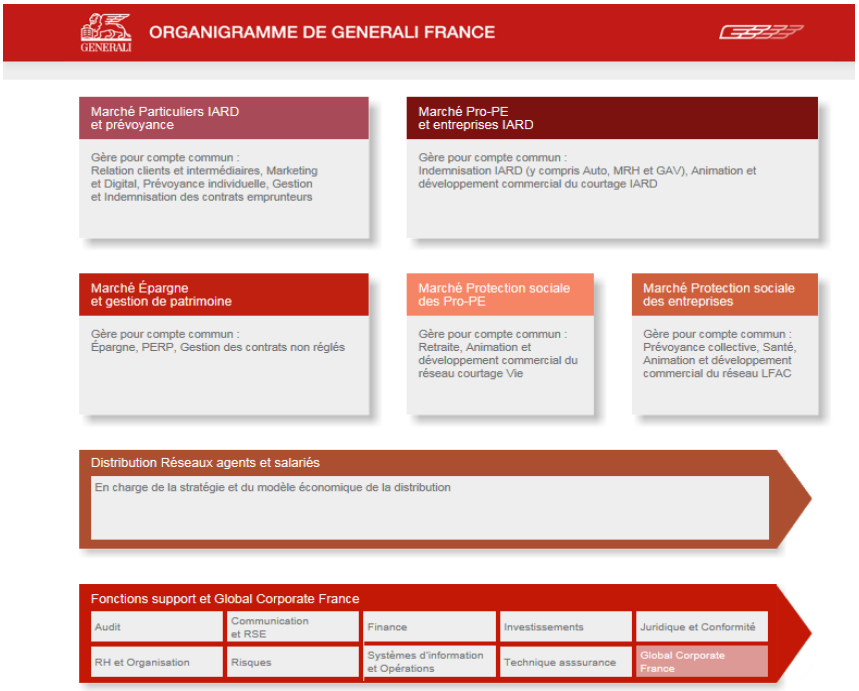



*Figure 1 : Organigramme de Generali France*

## I.II – Programme de transformation Excellence 2022

Excellence 2022 est un programme de transformation sur 5 ans dont l'objectif est de propulser Generali France dans le trio de tête des acteurs français de l'assurance sur les principaux marchés :

- Les particuliers, en épargne et patrimoine, ainsi qu'en assurances de dommages et prévoyance en marque blanche ou en partenariat ;
- Les professionnels, en assurances de dommages, en protection sociale et retraite ;
- Les entreprises, en assurances de dommages et en protection sociales.

Generali veut être la compagnie multi spécialiste de référence par son engagement de service aux clients, l'excellence de son offre, son agilité et sa capacité à innover. Sa distribution, la plus variée du marché français et riche de toutes les formes d'intermédiaires, doit être développée afin d'accroître ses parts de marché, en particulier celle portant la marque Generali.

Le programme Excellence 2022 s'appuie sur trois piliers :

- La conquête de business rentable, chaque direction de marché disposant de moyens étendus pour répondre aux besoins spécifiques de ses clients (Prévention / Commercialisation / Offre / Gestion et Indemnisation) ;

- L’excellence opérationnelle et la digitalisation des process via l’intégration des nouvelles technologies et des usages digitaux pour exceller dans la qualité de service client ;

- L’innovation et la sophistication technique par la captation optimale de la donnée, l’utilisation responsable de l’Intelligence Artificielle et des objets connectés pour délivrer des services et des garanties personnalisées.

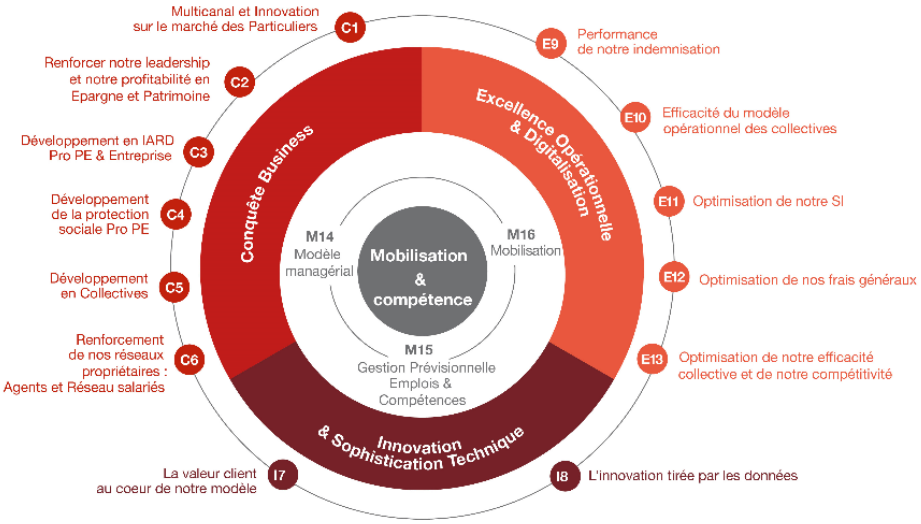



*Figure 2 : Les piliers d'Excellence 2022*

En effet, ces trois piliers sont réalisables avec la mobilisation de tous les services. Les équipes sont encouragées à faire évoluer leurs compétences pour anticiper la transformation des métiers résultants de la révolution digitale et à être animés par un état d’esprit fondé sur l’exigence, l’engagement, l’audace, la coopération et la culture du résultat. L’objectif est de constituer ainsi une entreprise d’assurance d’un type nouveau, intégrant des services novateurs qui permettent de faire converger les intérêts individuels de nos clients et l’intérêt collectif de nos communautés par la valorisation de comportements responsables.

## II – Présentation du service

Le service des Contrats Non Réglés (CNR) est un service qui a été mis en place en mars 2015, afin de renforcer le dispositif de lutte contre la déshérence dans un cadre règlementaire de plus en plus exigeant. La réglementation a régulièrement été renforcer par la promulgation de trois lois : AGIRA 1 en 2005, AGIRA 2 en 2007 et la loi Eckert en 2014 :

**AGIRA 1 - 15/12/2005**

- AGIRA (Association de Gestion des Informations sur le Risque en Assurance) centralise les demandes des bénéficiaires potentiels et les transmets à l'ensemble des organismes d'assurances pour vérifications.
- Permet à toute personne pensant être bénéficaire d'un contrat d'assurance vie de saisir l'AGIRA pour en obtenir la confirmation.

**AGIRA 2 - 17/12/2007**

- Impose aux assureurs de rechercher si les assurés et les bénéficiaires d'un contrat d'assurance vie sont décédés.
- Les entreprises d'assurances avertissent ensuite les bénéficiaires dés lors que le décès d'un assuré est attesté et que les bénéficiaires et leurs coordonnées sont identifiés.

**Loi ECKERT - 14/06/2014**

- Cette loi vient conforter AGIRA 2 (revalorisation post-mortem, élargissement aux contats de capitalisation nominatifs, frais de recherche interdits, délais de 15 jours pour demander les pièces justificatives au bénéficiaire...), reversement CDC.

*Figure 3 : Lois AGIRA 1 et 2 et Eckert*

Dans ce cadre, les assureurs envoient leurs bases de données à l'AGIRA qui les rapprochent du Registre National d'Identité des Personnes Physiques (RNIPP) afin d'identifier d'éventuels décès. Les suspicions de décès sont ensuite renvoyées aux assureurs pour traitement.

L'objectif de ces lois est, entre autres, de palier à la déshérence. Les contrats sont mis en déshérence lorsque suite au décès de l'assuré, du souscripteur ou au terme du contrat, le capital ou la rente n'est pas réglé au(x) bénéficiaire(s) ou ayant droits prévu(s) par le souscripteur, principalement pour les raisons suivantes :

- L'assureur n'a pas connaissance ou ne parvient pas à avoir confirmation du décès du souscripteur ou de l'assuré ;
- Les prestations ne sont pas versées aux bénéficiaires (ni aux ayant droits) parce que l'assureur n'arrive pas à les identifier ou à les localiser après avoir pris connaissance du décès de l'assuré ou du souscripteur (clause bénéficiaire imprécise, coordonnées

absentes, erronées ou obsolètes, non transmission des pièces justificatives pas le(s) bénéficiaire(s), etc.) ;

- La notion est étendue au cas où le souscripteur est en vie au terme du contrat et que ni celui-ci, ni le bénéficiaire, ni les ayants droits ne se sont manifestés auprès de la compagnie pour réclamer le règlement ni pour l'informer d'un changement de coordonnées.

Ainsi, les missions du service CNR sont de :

- Centraliser, animer et sécuriser la gestion des contrats non réglés AGIRA ;
- Coordonner le pilotage et la communication ;
- Traiter les détections AGIRA 2 ;
- Piloter les chantiers du plan d'actions Déshérence ;
- Fiabilisation des données.

Le service CNR est composé de 3 équipes : une équipe « Gestion », une équipe « Pilotage de délégataire » et une équipe « Pilotage et Projet ». C'est au sein de cette dernière équipe que j'ai effectué mon alternance.

## III – Les problématiques *Data* de Generali

La donnée prend de plus en plus d'importance aujourd'hui, et notamment les données à caractère personnel (aussi appelées données personnelles). Une donnée à caractère personnel représente toute information relative à une personne physique qui permet de l'identifier directement ou indirectement (par exemple un nom, un numéro de téléphone, une photographie, une empreinte digitale…). Il existe plusieurs typologies de données :

- Données clients ;
- Données partenaires (agents, courtiers, CGPI) ;
- Données collaborateurs ;
- Données des tiers (fournisseurs, prestataires) ;
- Données de l'entreprise (données financières, données liées à la stratégie, aux offres produits, aux contrats, aux prestations).

Tous ces éléments ont une valeur et constituent le patrimoine de Generali, et un patrimoine grandit, s'entretient et se protège.

L'accès aux informations devient primordial pour les assureurs qui, grâce à l'exploitation des données, ont dés lors une meilleure connaissance des clients et de leurs comportements. Cette connaissance permet par exemple de proposer aux clients des tarifs mieux adaptés et plus justes. Cela permet également de proposer de nouveaux services de prévention et des offres liées à la maison connectée et à l'internet des objets. Les données représentent donc un enjeu *business* pour l'entreprise.

Il y a sept étapes structurantes de la vie d'une donnée qui sont :

- La collecte et/ou la saisie ;
- Le contrôle ;
- La consommation et le traitement ;
- L'évolution ;
- Le transport ;
- Le stockage ;
- L'archivage ou la destruction.

Chacune de ses étapes est importante, et si elles sont réalisées de façon conforme, les données prennent alors de la valeur et deviennent une réelle source de *business* pour Generali.

## IV – Problématique

Il existe de nombreux contrats d'assurance différents, c'est pourquoi chez Generali il existe également de nombreux systèmes d'informations. L'équipe « Pilotage et Projet » travaille sur les multiples bases de données des contrats d'assurance-vie et leurs assurés.

Au sein de cette équipe, l'une des missions capitale est de fiabiliser les données. Au cours de mes deux années d'alternance, j'ai pu analyser le travail déjà mis en place sur l'analyse de la qualité des données des différents systèmes d'informations et par conséquent de bien me familiariser avec les différentes techniques employées pour la détection d'anomalies sur les données de contrats (identitaires et contacts).

Après avoir bien pris connaissances des processus d'analyse et de détection, j'ai contribué à l'élaboration d'un algorithme, permettant de relever ces anomalies sous un nouveau langage de programmation au sein de Generali : PySpark. En effet, jusqu'à présent ces analyses étaient effectuées sous R et Excel. Nous avons donc migré vers les technologies Big Data en faisant également des Tableaux de Bords sous le logiciel Tableau.

Ainsi, la réalisation de ce projet nous a amené à répondre à la question suivante :

**Comment rendre compte de la qualité des données, dans les champs identitaires et contacts, dans le milieu de l'assurance ?**

# Chapitre 2 : Qualité des données et Big data, un état de l'art

## I – Définition, anomalies et algorithmes

### I.I – Définition

Une donnée est de qualité si elle répond parfaitement aux besoins de son utilisateur (Brasseur, 2005). Nous allons distinguer la qualité d'une donnée selon deux niveaux : le premier étant celui de la propreté. Une donnée dite « propre » est une donnée qui n'aura pas besoin d'un retraitement dans le cadre d'un nettoyage de la donnée. Par exemple, dans une base de données contenant des informations sur la date de naissance d'une personne, on va qualifier de propre ce champ si toutes les dates sont cohérentes (pas de date de naissance à 01/01/1900, qui est considérée comme étant le chiffre 1 sous Excel). Ainsi, on pourra admettre, dans un premier temps, que les données relatives aux dates de naissance sont cohérentes et donc de qualité.

Le deuxième niveau est celui de l'usage que l'utilisateur en fera. En effet, un champ de données peut être correctement et proprement rempli, mais s'il ne correspond pas à la demande de l'utilisateur, elle sera inutile à sa problématique. Par exemple, si on souhaite calculer l'ancienneté d'une personne dans un contrat d'assurance, ce ne sera pas la date de naissance qui nous sera utile mais bien la date d'adhésion au contrat. Ainsi, il est nécessaire, pour les utilisateurs des données, de bien comprendre les besoins afin d'admettre la bonne qualité d'une donnée, et afin de satisfaire les utilisateurs, les données doivent respecter certaines spécificités.

Il existe une multitude de spécificités afin d'évaluer la qualité d'une donnée. En effet, dans le livre sur la qualité de donnée de Berti-Equille (2012), l'auteur parle des familles d'indicateurs de qualité à travers cette figure :

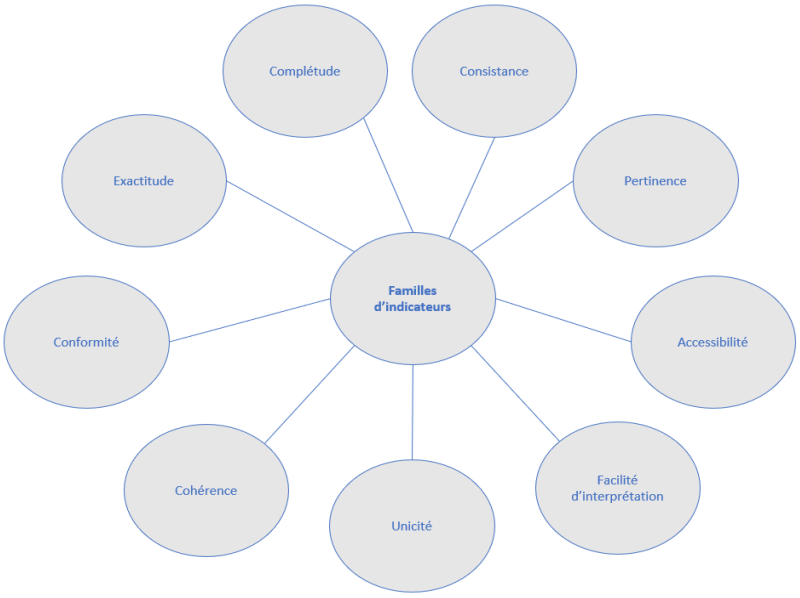



*Figure 4 : Les familles d'indicateurs de qualité de données (Berti-Equille, 2012, p.26)*

Nous distinguons ainsi 9 familles d'indicateurs comme étant les plus significatives par l'auteur. Nous allons les définir ci-dessous :

**Consistance**

La consistance consiste à ce que pour les entités qui sont présentes dans une ou plusieurs base de données, nous retrouvons les mêmes valeurs de champs dans toutes les bases. La consistance est un aspect capital des données car cela va permettre de faire des comparaisons, des rapprochements, ou encore des jointures de tables.

Par exemple, pour un même identifiant client, le nom, le prénom, la date de naissance, ou encore l'adresse doivent être les mêmes dans chaque base de données afin d'éviter toute incohérence dans le rapprochement des données. Si une information est mise à jour dans une base, elle doit l'être également dans toute les autres (Berti-Equille, 2012)

**Pertinence**

La pertinence d'une donnée permet de mesurer la capacité, dans le temps, d'un champs ou d'une donnée, de répondre correctement aux besoins des utilisateurs. En effet, une donnée présente dans une base utilisée sur plusieurs années se doit d'être utilisable à n'importe quel moment ou époque. De ce fait, les bases de données doivent contenir des données utiles et utilisés mais également facilement adaptable avec l'évolution métier (Berti-Equille, 2012).

Par exemple, dans une base de données d'assurance contenant des données de contrats et clients, il est précisé le système d'information (SI) auquel appartient un contrat. Suite à des restructucturations des systèmes, il peut être décidé de regrouper certains systèmes afin de faciliter les traitements statistiques. L'ancienne nomenclature est donc inutilisable alors qu'elle était pertinente jusqu'à présent.

| Ancienne nomenclature des SIs | Nouvelle nomenclature des SIs |
|---|---|
| Epargne salariale | **Epargne** |
| Epargne individuelle | |
| Retraite collective | **Retraite** |
| Retraite individuelle | |
| Assurance vie | **Vie** |
| Assurance santé | |

*Tableau 1 : Exemple de pertinence (fictif)*

**Accessibilité**

L'accessibilité mesure la facilité d'accès aux données par l'utilisateur. Une donnée ou un fichier de données peuvent être utilisés par différents utilisateurs (de l'administrateur de base de données aux lecteurs des données), ainsi les données doivent être facilement accessible avec une ergonomie nécessitant peu de clics de souris. Cela permet d'accéder plus rapidement aux données sans perdre de temps et d'éviter un enchainement d'écran fastidieux (Brasseur, 2005).

Par exemple, des répertoires communs peuvent être créé sur le réseau de l'entreprise pour déposer des fichiers qui seront accessibles par tous les utilisateurs qui se serviront du fichier de donnée en question.

**Facilité d'interprétation**

La facilité d'interprétation permet de savoir si une donnée est facilement compréhensible, dans son analyse et son usage. Il est important que les utilisateurs comprennent bien les données et de ce fait, une documentation avec des descriptions claires des données est essentiel. Pour que cela soit compréhensible par tous, le choix du vocabulaire est important car il faut que les personnes du milieu technique comme les personnes du milieu décisionnel comprennent la même chose concernant une donnée avec le même vocabulaire. Cela permet d'éviter toute confusion et source d'erreur. Ce sont les métadonnées (Berti-Equille, 2012).

Par exemple, pour un fichier de données, il est possible de créer un fichier texte en y inscrivant chaque champ ainsi que sa signification.

**Unicité**

L'unicité permet de garantir un seul enregistrement par entité du monde réel. Cela sert à éviter les doublons. Par exemple, un identifiant client est unique pour chaque individu. On ne le retrouvera pas sur une autre ligne.

**Cohérence**

La cohérence permet d'assurer que les données ne soient pas contradictoires entre elles. Ainsi, cela évite d'avoir des résultats erronés, ou encore de propager des erreurs à tous les niveaux de l'entreprise.

Par exemple, pour un enregistrement de date de naissance d'un fichier client, si l'individu est né le 04/07/1859, on verra que la date ne peut pas être correcte et donc on précisera que la valeur de cette date est incohérente.

**Conformité**

La conformité d'une donnée est liée à un standard, un format ou d'une convention de nommage au sein du service ou de l'entreprise.

Par exemple, pour le champ « sexe », on va le coder de la manière suivante : 0 pour homme et 1 pour femme.

**Exactitude**

L'exactitude d'une donnée permet de mesurer si les données d'une base sont identiques à ceux de la réalité. Cela représente un aspect sensible dans la qualité d'une donnée car, en effet, un bon service rendu au client se doit d'être basé sur des données correctes. Les données non à jour doivent donc être actualisées afin d'avoir la meilleure précision possible (Berti-Equille, 2012).

Par exemple, le milieu de l'assurance rencontre le problème du NPAI (N'habite Pas à l'Adresse Indiquée). En effet, lors des envois de courrier, il est impératif d'avoir la bonne adresse d'un assuré et ce dernier a également une part de responsabilité dans la communication de celle-ci.

**Complétude**

La complétude d'une donnée est principalement liée au modèle de donnée. Il y a quatre critères qui permettent de juger de la complétude d'une donnée : la complétude des entités, la complétude des attributs, la complétude des relations, la complétude des occurrences. Nous allons les détailler avec des exemples. La complétude des entités vérifie que toutes les entités sont bien présente dans le modèle de données (Berti-Equille, 2012). Par exemple, sur une base de données client, il est nécessaire d'avoir l'identifiant du client. Autrement la base de données sera incomplète.

La complétude des attributs permet d'évaluer l'exhaustivité des attributs dans les entités du modèle de données. Par exemple, dans une base de données d'assurance automobile, il est impératif d'avoir les informations sur l'identité de l'assuré (nom, prénom, date de naissance), mais aussi sur le véhicule qui doit être assuré (modèle du véhicule, année de mise en circulation).

La complétude des relations évalue s'il y a suffisamment de relations entre les différentes entités du modèle de donnée. Par exemple, une personne peut acheter un ou plusieurs bien immobilier, notre modèle de donnée doit donc comporter une relation « achète » qui lie les entités « personnes » et les entités « bien_immo ».

La complétude des occurrences évalue l'exhaustivité des occurrences de chaque entité du modèle de donnée. Par exemple, dans la table « détenteur de permis », chaque personne détentrice d'un permis de conduire doit exister sur une ligne spécifique, et personne ne doit être oublié (Berti-Equille, 2012).

Les études sur la donnée de manière général ou sur la qualité d'une donnée sont précieuses car elles permettent de rendre compte les aspects importants de la qualité de donnée. D'un point de vue métier, la qualité des données peuvent aider à la décision et par conséquent de générer des profits. Par exemple, dans le milieu de l'assurance, les entreprises vivent dans un océan de données (Olson, 2003) et le traitement de donnée est l'une des activités la plus répandue dans ce secteur.

La bonne qualité d'une donnée peut permettre d'engendrer d'énormes gains car les informations peuvent être vendues à des sociétés souhaitant faire des campagnes publicitaires ciblées dans certains cas. Il existe trois grandes approches que les entreprises peuvent adopter pour monétiser leurs données (Wixom, Ross, 2017). La première étant de vendre les données directement aux marchés nouveaux et existant en faisant appel à des intermédiaire (*Data Broker*). La deuxième est celle de l'échange d'information. Une donnée peut être échangé contre d'autres données, des biens, ou des services. Les données peuvent également être un point de négociation et d'arrangements avec d'autre entreprises. La troisième consiste à améliorer les performances de l'entreprise en améliorant les procédures et les décisions internes en se basant sur les données. En effet, les données que possèdent l'entreprise peut aider à déterminer les aspects bloquant d'un processus et ainsi trouver des solutions pour y remédier. L'apprentissage par les données peut également permettre de mieux comprendre les besoins des clients, et donc de proposer des solutions plus adaptés mais aussi de créer de nouveaux produits (Bastien L., 2019).

Jour après jour, les données deviennent de plus en plus importantes. La qualité des données est un élément capital du pilotage d'une entreprise (Longet, 2016). Cela représente un atout vis-à-vis des concurrents, mais représente également l'image d'une entreprise. Elles peuvent aider les entreprises et services à prendre les bonnes décisions quant à la stratégie à adopter sur différents sujets (court/moyen/long terme). La données, dans sa globalité, gagne donc en valeur au fur et à mesure du temps et présente donc des enjeux *business* considérable (Olson, 2003).

## I.II – Anomalies

Les anomalies liées aux données peuvent provenir de différents facteurs. En effet, les erreurs apportées aux données peuvent être liés à différents niveaux et étapes de traitements des données. Le tableau suivant présente les problèmes de qualité des données :

| Etapes de traitement des données | Source de problèmes de qualité des données |
|---|---|
| **Création** | • Entrée manuelle: absence de vérification systématiques des formulaires de saisie<br>• Entrée automatique: problèmes de capture OCR, de reconnaissance de la parole, incomplétude, absence de normalisation ou inadéquation de la modélisation conceptuelle de données: attributs peu structurés, absence de |

| | |
|---|---|
| | contraintes d'intégrité pour maintenir la cohérence des données<br>• Entrée des doublons<br>• Approximations<br>• Contraintes matérielles ou logicielles<br>• Erreurs de mesure<br>• Corruption des données: faille de sécurité physique et logique des données |
| **Collecte/import** | • Destruction ou mutilation d'information par des prétraitements inappropriés<br>• Perte des données : *buffer overflows*, problèmes de transmission<br>• Absence de vérification dans les procédures d'import massif<br>• Introduction d'erreurs par les programmes de conversion de données |
| **Stockage** | • Absence de méta-données<br>• Absence de mise à jour et de rafraîchissement des données obsolètes ou répliquées<br>• Modèles et structures de données inappropriés, spécifications incomplètes ou évolution des besoins dans l'analyse et conception du système<br>• Modification *ad hoc*<br>• Contraintes matérielles ou logicielles |
| **Intégration** | • Problème d'intégration de multiples sources de données ayant des niveaux de qualité et d'agrégation divers<br>• Problèmes de synchronisation temporelle<br>• Système de données non conventionnels<br>• Facteurs sociologiques conduisant à des problèmes d'interprétations et d'intégrations des données<br>• Jointures *ad hoc*<br>• Appariements aléatoires<br>• Heuristiques d'appariements des données inappropriées |
| **Recherche et analyse** | • Erreurs humaine<br>• Contraintes liées à la complexité de calcul<br>• Contraintes logicielles, incompatibilité<br>• Problèmes de passage à l'échelle, de performances et de confiance dans les résultats<br>• Approximations dues aux techniques de réduction des grandes dimensions<br>• Utilisation de boîtes noires pour l'analyse<br>• Attachement à une famille de modèles statistiques<br>• Expertise insuffisante d'un domaine<br>• Manque de familiarité avec les données |

*Tableau 2 : Des problèmes de qualité des données (Berti-Equille, 2004)*

Ce tableau présente quelques problèmes que l'on peut rencontrer dans le traitement d'une donnée. Par conséquent, ce ne sont pas toutes des erreurs «classique» qu'un utilisateur peut faire face.

Au cours des différentes analyses qu'un utilisateur peut effectuer, il y a des erreurs typiques qui sont liées à la mauvaise qualité d'une donnée. Il y a, par exemple, les doublons, les coordonnées d'une personne ou encore les erreurs liées aux adresses postales.

Les doublons correspondent à des mêmes enregistrements dans une base de données et qui signifient la même chose. En effet, dans une base de données client, on peut retrouver plusieurs fois la même personne avec les mêmes informations et cela peut porter à confusion car on peut se poser la question de l'homonyme parfait ou encore se demander quel enregistrement choisir pour une opération.

Concernant les coordonnées d'une personne, il n'est pas impossible de trouver dans une base de données des faux numéros ou encore des numéros suspects. Par exemple, si le format du champ n'a pas été spécifié qu'il s'agit d'un numéro de téléphone, il se peut qu'il y ait plus ou moins de chiffre que celui attendu (10 chiffres pour un numéro français). Pour les numéros dit « suspect », il est possible que des numéros « générique » soit renseigné dans les bases de données. Par exemple, on peut retrouver un numéro de contact du type « 01 01 01 02 03 » et par conséquent, c'est une donnée qui porte à confusion car c'est un format de numéro atypique.

On rencontre également des problèmes liés aux adresses postales dans les bases de données client. De manière général, on peut renseigner son adresse sur plusieurs lignes et cela peut porter à confusion pour celui qui rempli cette information. Il y a un risque de retrouver les mauvaises informations sur les mauvaises lignes et donc de ne pas avoir le même type d'adresse pour chaque enregistrement. C'est pour cela qu'une bonne ergonomie est essentiel pour ce type de donnée.

Il existe quatre types d'approches pour évaluer et contrôler la qualité des données (Berti-Équille, 2018) :

Les approches préventives se concentrent sur l'ingénierie des systèmes d'information et le contrôle des processus, et ses techniques sont utilisées pour évaluer la qualité des modèles conceptuels, le développement de logiciels et les processus utilisés pour le traitement des données.

Les approches diagnostiques se concentrent, d'une part, sur les méthodes statistiques, l'analyse et l'exploration de données d'exploration, qui peuvent détecter des anomalies dans l'ensemble de données, et d'autre part, se concentrent sur la vérification de contraintes ou de différents types de données, règles et dépendances fonctionnelles.

Les approches correctives se concentrent sur les techniques de nettoyage et de fusion des données, d'une part à l'aide d'opérateurs d'extraction et de transformation de données (ETL), et d'autre part, des méthodes d'interpolation statistique basées sur les règles, des

heuristiques, de techniques d'apprentissage automatique ou encore des méthodes de remplacements de données.

Les approches adaptatives sont généralement appliquées lors de l'intermédiation ou de l'intégration des données au moment de la requête : elles se concentre sur l'adaptabilité du traitement (requête ou effacement des données), notamment sur la vérification à l'exécution de la qualité des données en prenant également en compte les préférences utilisateurs afin de proposer une personnalisation.

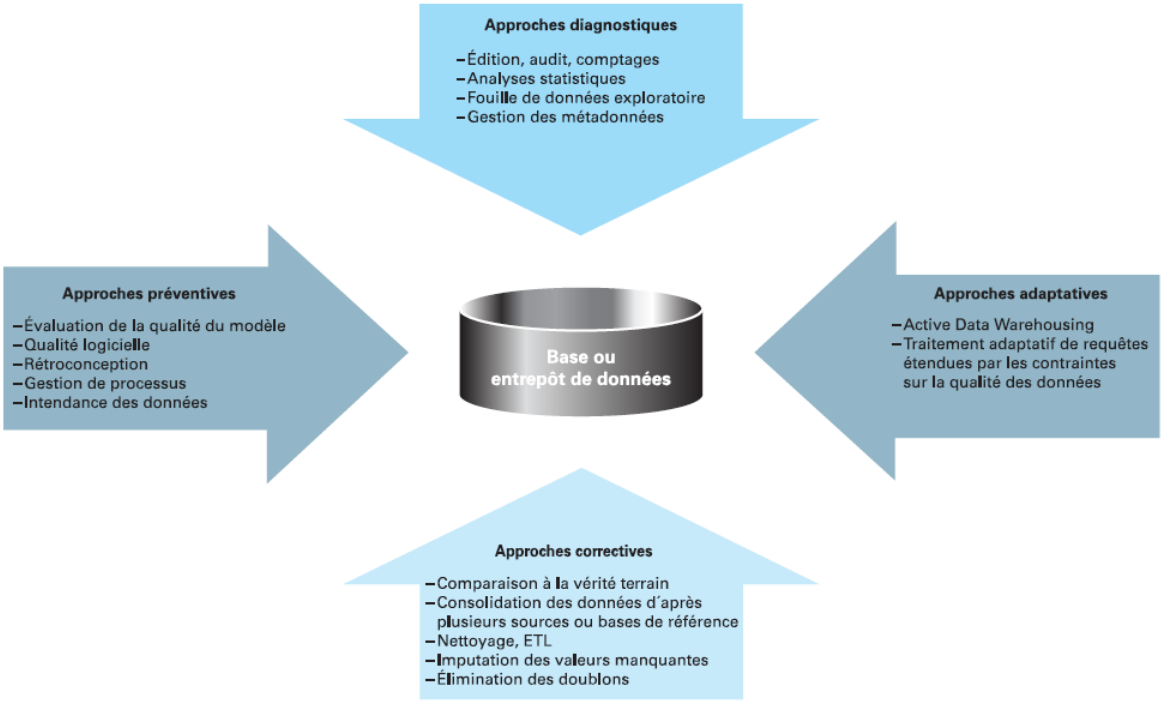



*Figure 5 : Approches d'évaluation et de contrôle de QDD (Berti-Équille, 2018)*

Il existe plusieurs approches pour gérer la qualité des données. En plus de l'approche terrain que nous avons vu précédemment. L'audit de donnée est une méthode pour évaluer la qualité d'une donnée. En effet, l'audit des données consiste à mettre en place un programme qui va vérifier si les données étudiées respectent les spécificités / contraintes qui ont été mise en place par l'utilisateur. Cela peut être le respect des règles de calcul, règles spécifiques à des chaines de caractères, facteurs statistiques par exemple. L'avantage de cette méthode est qu'elle est plus rapide à mettre en place que les approches précédentes mais uniquement dans le but de contrôler la qualité d'une donnée. L'audit de donnée n'aura pas pour but d'améliorer des données, ou de vérifier l'exactitude mais uniquement de voir si les données respectent les règles de données.

Afin d'avoir une donnée de meilleure qualité, on peut appliquer un nettoyage de donnée. Ce processus consiste à uniformiser les données (mettre au même format). Par exemple, si un champ de type date est renseigné comme ci : « 01.10.2000 », elle sera transformée de manière à ce qu'elle soit plus utilisable dans les logiciels de traitement de données et sera sous ce nouveau format : « 01/01/2000 ».

## I.III – Algorithmes de détection d'anomalies

La détection d'anomalies est très répandue dans l'analyse de données. En effet, il est important de savoir quelles données semblent suspectes ou erronées afin d'éviter d'éventuelle perte au niveau du marché sur lequel est une entreprise. Il existe, dans la littérature, des méthodes qui permettent de relever les anomalies liées aux données et le *machine learning* a permis de faire de grandes avancés dans ce domaine d'analyse.

Il n'est pas pertinent de définir une métrique aux anomalies car les cas peuvent être très divers (Talbi, 2019). On admettra donc qu'une anomalie d'une donnée comme étant une observation éloignée du reste des observations. Par exemple, dans une base de données client, on s'attend à lire des données personnes (nom, prénom, date de naissance), par conséquent, il n'est pas possible d'avoir des chiffres dans les champs nom et prénom. Si cela était le cas, on détecterait une anomalie.

Le diagnostic de la qualité des données repose également sur un élément fondamental : les contraintes et spécificités. En effet, il est essentiel de définir les règles de gestion et de mise en œuvre des données et de vérifier la cohérence des données (ne pas avoir de date de naissance dans le champ du nom par exemple). De manière général, le contrôle des données et la détection d'anomalie peut être effectué à différents niveaux. Cela peut aller d'un simple comptage (statistiques descriptives) à des techniques beaucoup plus développées (Berti-Equille, 2004).

Dans le cas d'une base de données issus d'une fusion de plusieurs bases, il est possible d'observer des doublons. Cela peut porter à confusion, il faut donc gérer cet aspect. Pour cela, dans les différents outils de gestion de base de données, il existe des fonctions capable de supprimer les doublons d'un fichier (exemple : l'option *nodupkey* lors d'une *proc sort* en programmation SAS).

Les algorithmes de détection d'anomalies sont nombreux mais pas utilisables dans tous les cas. En effet, les algorithmes sont applicables à des cas particuliers. Il est donc capital de bien choisir sa méthode d'analyse afin de ne pas obtenir des résultats aberrants. Certains outils proposent directement des tableaux de bord applicable à une base de données pour rendre compte du niveau de qualité et par conséquent de détecter les anomalies.

La société IBM propose des produits de qualité des données avec plusieurs options de processus. Il y a une option de nettoyage des données et de surveillance continue de la qualité des données, de gérer ces informations mais également de créer des vues sur les entités d'une base de données. La société propose également des *webinaire* a propos de l'automatisation de l'évaluation de la qualité des données mais également d'utiliser l'apprentissage automatique (*machine learning*) pour générer des donnée fiables et propres à travers des analyses statistiques approfondis. Tous ces outils sont réunis pour être utilisé dans ce qu'IBM appelle « le *DataOps* ».

# II – Le Big Data

## II.I – Les données

Nous sommes aujourd'hui à un niveau ou les entreprises sont très axées sur la collecte et l'analyse de données. Les problématiques *Data* sont de plus en plus important et représentent un sujet très complexe et ambitieux dans un contexte *business*. Il est donc important d'étudier attentivement le type de données que l'on analyse. Les données sont le moteur de la plupart des entreprises et celle-ci les utilisent en grandes quantité, ce qui représente un atout pour obtenir un avantage concurrentiel mais également pour réduire les coûts et maximiser l'efficacité de l'entreprise.

Une donnée est une information de la vie réelle permettant de faire des constats sur différents sujets. Une donnée peut être utilisé pour plusieurs fins comme de la simple collecte, du traitement de donnée dans l'objectif d'apprendre sur un sujet, de l'analyse poussé dans le but de faire des prédictions, ou encore de la diffusion d'information et donc de la communiquer. Dans le monde du *Big Data*, il existe trois types de données : structurées, non structurées, semi-structurées.

Les données structurées sont des données qui respectent un format. Ce sont des données facilement compréhensibles par l'Homme et facilement exploitable par les ordinateurs. Le fait qu'elles soient structurées, les données vont pouvoir être stockées et facilement accessibles. Ce type de données est donc directement exploitable par une machine et n'aura pas besoins de pré-traitement avant usage. Cela peut correspondre à une table client dans une base de données. Les données structurées peuvent être stockées dans des bases de données relationnelle et peut être interrogé en SQL.

Les données non structurées sont des données qui ne sont pas organisées ni formatées comme une donnée structurée. En effet, l'analyse de données non structurées représente un grand défi et cela peut créer des problèmes car ce type de données ne respectent aucun format ou modèle. Ainsi, on retrouvera les données non structurées dans une base de données non relationnelle et pourra être interrogé à l'aide de NoSQL. Une donnée non structurée peut correspondre à une image, un fichier audio ou un commentaire sur un réseau social par exemple.

Les données semi-structurées représentent un format intermédiaire entre les données structurées et semi structurées. De manière générale, il s'agit d'un fichier présentant les deux types de données précédent. Par exemple, un document de type texte est perçu comme un ensemble de données non structurées mais des informations comme des mots-clés permettent de retrouver plus facilement des informations grâce à ces termes. Ce sont alors des données semi-structurées.

## II.II – Big Data : Histoire & concepts

Selon une étude menée par la société IBM, environ 2,5 quintillions de bytes de données sont créés tous les jours, dans le monde. Le terme « *Big Data* » fait référence à la grande quantité de données produite à un âge ou le digital est en plein essor. Cette grande quantité de données, structurées et non structurées, inclue les données issues d'internet généré par l'envoie de messages, la navigation sur des sites, des transactions bancaires, la publication sur les réseaux sociaux et bien d'autre encore. Environ 80% des données mondiale sont de type non structurées sous la forme de texte, photos, images, et ne permettent donc pas d'employer des méthode d'analyse traditionnelle. Les données sont de plus en plus importantes et complexes et vont donc nécessité de nouveaux algorithmes pour en extraire le plus d'informations possible (Holmes, 2017). Le concept du *Big Data* n'est donc pas arrivé d'un jour à l'autre, mais bien par le développement des technologies informatique, la puissance des ordinateurs et la génération importante de données au quotidien.

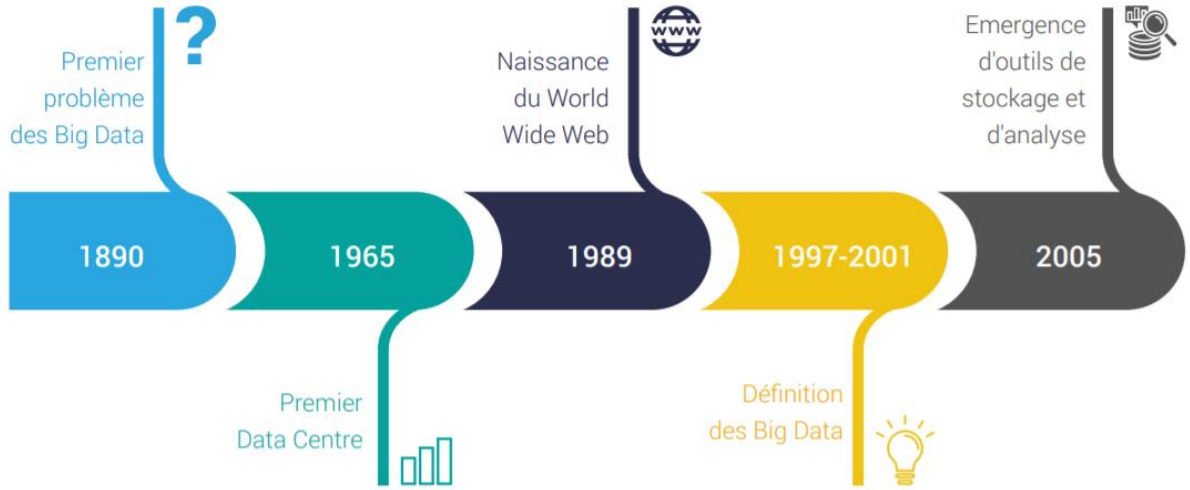



*Figure 6 : Chronologie du Big Data (Rahhal, Mezzour, 2020)*

Au vu de la grande quantité de donnée générées au quotidien, de nouvelles techniques ont du être mises en place pour traiter efficacement les données. On assiste en 2005 à une prise de conscience de la quantité de données que les utilisateurs produisaient sur Facebook, YouTube et d'autres service en ligne, et par conséquent, des infrastructures *open source* (libre accès) ont été crées spécifiquement pour stocker et analyser les jeux de données du *Big Data*. C'est ainsi que Hadoop vit le jour.

Le développement d'infrastructures *open source* comme Hadoop et plus récemment Spark, est essentiel à la croissance du *Big Data* car il facilite son utilisation et réduit les coûts de stockage. Depuis, le volume du *Big Data* a explosé et les utilisateurs continuent de générer énormément de données. On notera qu'à présent, ce ne sont pas seulement les humains qui génèrent des données mais aussi les téléphones mobiles par exemple.

D’après la société Gartner, le terme *Big Data* est défini comme étant des actifs d’information à grand volume, à grande vélocité et/ou à grande variété qui nécessitent des formes de traitement de l’information rentables et innovantes qui permettent d’améliorer la compréhension, la prise de décision et l’automatisation des processus. A travers cette définition, on distingue trois problématiques clés qui composent le trio des 3V :

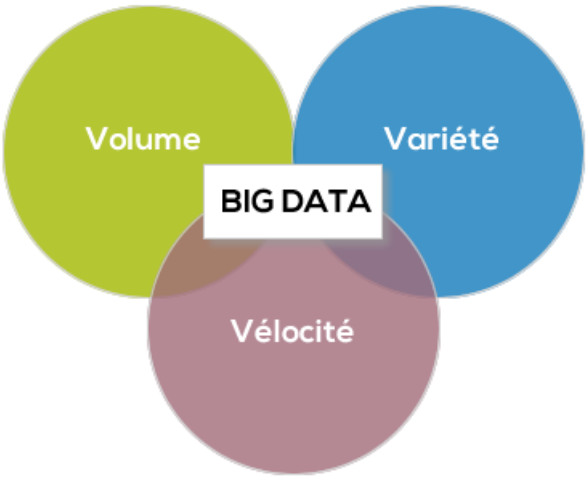



*Figure 7 : Les 3 "V" du Big Data*

**Volume**

Le volume correspond à la quantité de donnée générée par des entreprises ou par des personnes. Les entreprises, quelques soient leur secteur d’activité, devront faire face à cette problématique et donc de trouver des moyens solides pour gérer cette quantité de données. De manière générale, avec le *Big Data*, ce sont de gros volumes de données non structurées.

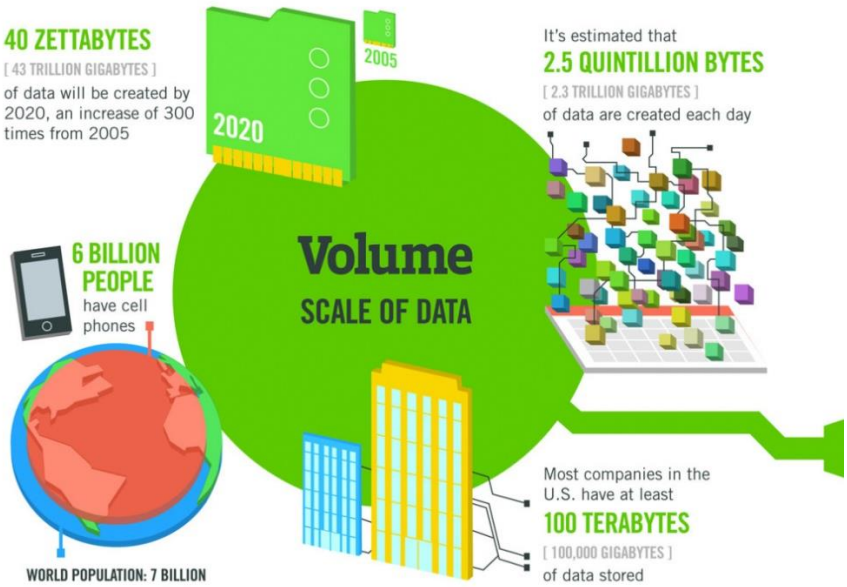



*Figure 8 : Volume, scale of data (Shafakhatullah Khan, Kausar, Nawaz, 2018)*

**Variété**

La variété des données correspond aux différents types de données disponibles. En effet les données traditionnelles que l'on dispose sont de type structurées et se trouvent des les bases de données relationnelles. Avec l'essor du *Big Data*, les données ne sont pas obligatoirement structurées. On trouvera des données de type non structurées et semi-structurées. Par conséquent, ce type de données nécessiteront des pré-traitements afin de pouvoir les exploiter par la suite et d'en dégager toutes les informations disponibles.

Cette spécificité du *Big Data* provient du fait que l'usage d'internet et du numérique dans sa globalité sont diversifiée. En effet, ces données peuvent provenir des réseaux sociaux, des navigateurs automobiles, ou encore des appareils connectés permettant l'analyse de l'état de santé d'une personne. C'est pourquoi la gestion, la protection et la gestion de la qualité des données constituent un nouveau défi pour les systèmes d'information, car les données ne proviennent pas forcément de sources internes et contrôlées, ni conforme aux formats souhaités.

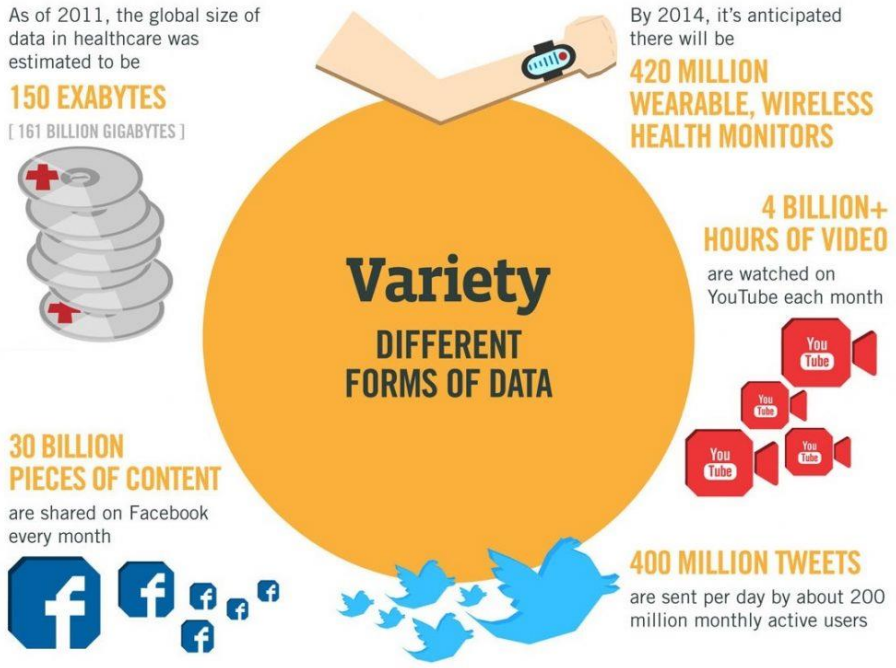



*Figure 9 : Variety, different forms of data (Shafakhatullah Khan, Kausar, Nawaz, 2018)*

**Vélocité**

La vélocité correspond à la fréquence à laquelle les données sont générées, capturées et partagées. Grâce aux avancées technologiques et informatiques, les entreprises et les utilisateurs génèrent beaucoup plus de données en beaucoup moins de temps que les années précédentes. Par exemple, en 2010, il fallait deux jours pour générer 5 exaoctets (5 milliards de Gigaoctet) de données. En 2013, il fallait seulement 10 minutes. Tout cela pour dire qu'il y a quelques années, récolter et traiter des données dans le but d'obtenir des informations exploitables prenait du temps mais qu'aujourd'hui, cela n'en demande que très peu voir exécutable en quasi-temps réel. Par conséquent, plus les outils numériques sont puissants et rapides, plus le nombre de données généré sera important.

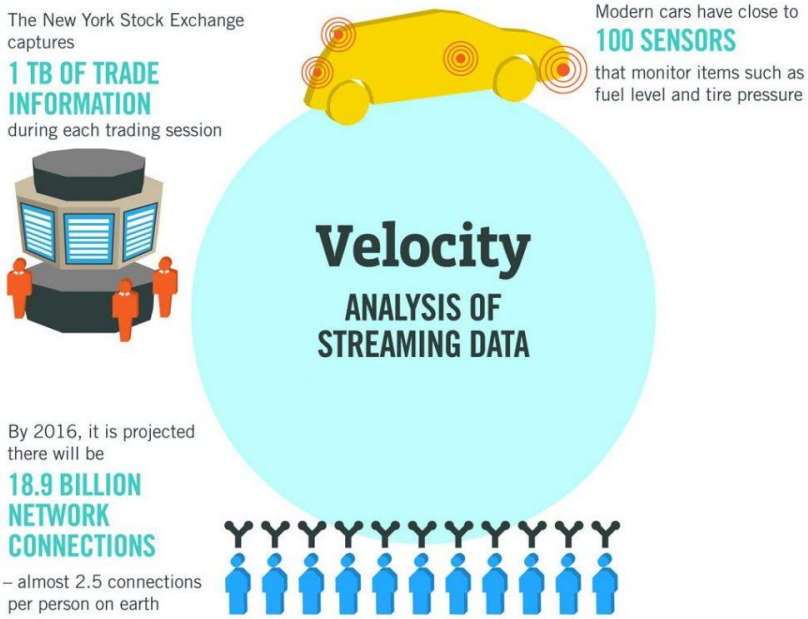



*Figure 10 : Velocity, analysis of streaming data (Shafakhatullah Khan, Kausar, Nawaz, 2018)*

Bien que le *Big Data* repose sur la problématique très connue des 3V que l'on a vu précédemment, il a été admis (Holmes, 2017) que le *Big Data* repose en réalité sur deux autres concepts qui s'ajoutent aux trois précédent : la véracité et la valeur. On parle alors des 5V. Les deux concepts additionnels sont la véracité et la valeur.

**Véracité :**

La véracité correspond à la qualité et la fiabilité des données collectées. Avec une telle quantité de données, la qualité ainsi que la précision sont moins vérifiables. En effet, le manque de qualité et d'exactitude provient, la plus du temps, des gros volumes de données. Par exemple, les données issues des tweets comprenant des *hashtags*, des abréviations, ou encore l'exactitude du contenu, sont tellement nombreuses qu'il est difficile de tous les contrôler avec les méthodes d'avant le *Big Data*. Mais aujourd'hui, il est possible de produire des analyses pertinentes avec ce type de données et d'en tirer des conclusions importantes.

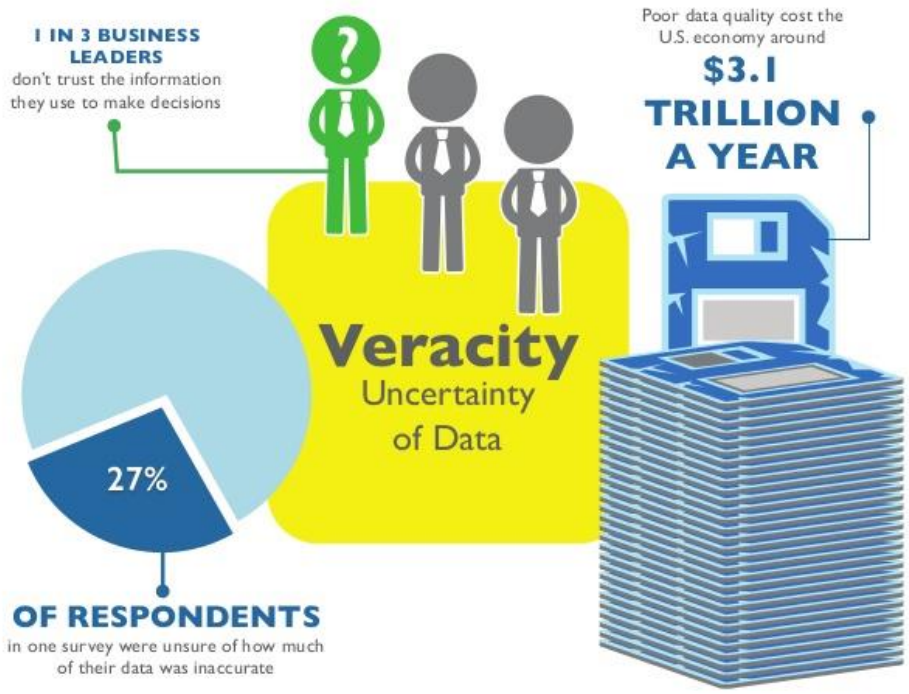



*Figure 11 : Veracity, uncertainty of data (Shafakhatullah Khan, Kausar, Nawaz, 2018)*

**Valeur:**

Enfin, voici le dernier « V » à prendre en compte lorsque que l'on parle du *Big Data*. Le volume correspond au sens que nous allons donner aux données. On l'appelle aussi la « Visualisation ». En effet, il est important de donner de la valeur aux données et par conséquent de les faire comprendre à tous. Ce « V » est considéré comme étant le plus important car il est vrai qu'une donnée a de l'importance uniquement si elle apporte de la valeur ajoutée aux informations. La valeur correspond également à l'aspect financier qu'apporte le *Big Data*. En 2009, l'ensemble des projets liés au *Big Data* représentaient un revenu de 100 millions d'euros. Depuis 2012, grâce aux avancés technologiques et informatiques, ces profits ont prodigieusement augmenté allant jusqu'à près de 42 milliards d'euros de revenu mondial de marché en 2018, et d'après les prédictions émises par le portail Statista, ces revenus pourraient atteindre plus de 100 milliards d'euros d'ici 2030 (Bourany, 2018).

Au fur et à mesure des années, le *Big Data* s'est de plus en plus développer. Les problématiques liées aux données sont de plus en plus importantes pour les entreprises et cela devient un atout concurrentiel majeur. Chaque objet numérique et connecté à internet génère un nombre important de données, il est donc important pour une société de savoir les gérer et d'en tirer un maximum d'information. Tous les jours, chaque personne génère à son échelle des données. Cela peut être des tweets, des publications de vidéos, des transactions bancaires, ou encore le stockage de base de données volumineuse pour une entreprise. Pour une entreprise, les données collectées doivent être la plus fiable possible afin de prendre les meilleures décisions et par conséquent, ces données se doivent être les plus valeureuse possible pour générer un maximum de profit.

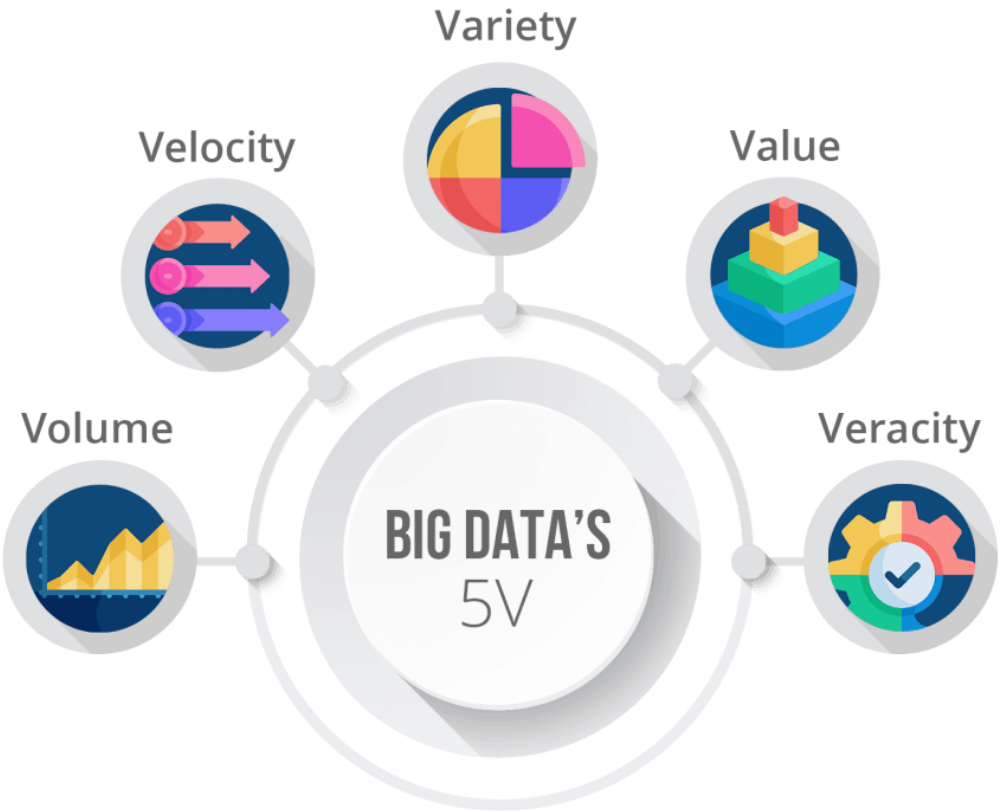



*Figure 12 : Les 5 "V" du Big Data*

## II.III – Hadoop

Avec un nombre conséquent de données produites par jour dans le monde, il est capital pour les entreprises de savoir les gérer et par la suite savoir comment améliorer ses performances. Sur le sujet du *Big Data*, Hadoop en est la principale plateforme. En effet, Hadoop est utilisé pour stocker et traiter une très grande quantité de données (tout types confondus). Ce *Framework* logiciel open source a la possibilité de prendre en charge des nombreuses taches volumineuses (calcul, agrégation…). Hadoop fait parti du projet Apache sponsorisé par *Apache Software Foundation*.

Afin de stocker et de récupérer des fichiers de données en un temps très rapide, il existe un système de fichier distribué d'Hadoop Apache appelé HDFS (Hadoop Distributed File System). Ce système est très adapté au *Big Data* car, en plus de sa vitesse de chargement et de stockage, il permet de stocker un grand volume de donnée (on parle de terabytes ou même petabytes de données). Il est donc idéal pour le traitement du *Big Data*. Avant l'existence de ce type de système, les données devaient être stockées de manière centralisée et cela engendrait des coûts importants et des contraintes de capacité de stockage. Avec ce type de technologies (HDFS), cela n'est plus un problème car il est de nature distribué sur des milliers de serveurs et permet de stocker un fichier sur différent serveurs. Par conséquent, l'entreprise va réduire ses coûts.

HDFS permet de stocker un gros volume de données sur une plateforme/*cloud* de données (nuage de donnée) afin que les données soient disponibles et accessibles par les utilisateurs d'une entreprise. Ainsi, un *cloud* de données doit être mis en place pour obtenir cette facilité d'accès. La société Cloudera propose un *Entreprise Data Cloud* pour les différents besoins d'une société. Ainsi, Cloudera a mis en place un *cloud* de données pour les entreprises afin que celles-ci puissent stocker leurs données. Dans une dimension *Big Data*, Hadoop va en faire distribution de Cloudera ainsi que plusieurs de ses composants.

Bien qu'il soit le socle des projets *Big Data*, Hadoop n'est pas capable de répondre à toutes les problématiques à lui seul. Il faut pour cela utiliser l'ensemble des technologies que présente Hadoop : on parle alors d'écosystème Hadoop.

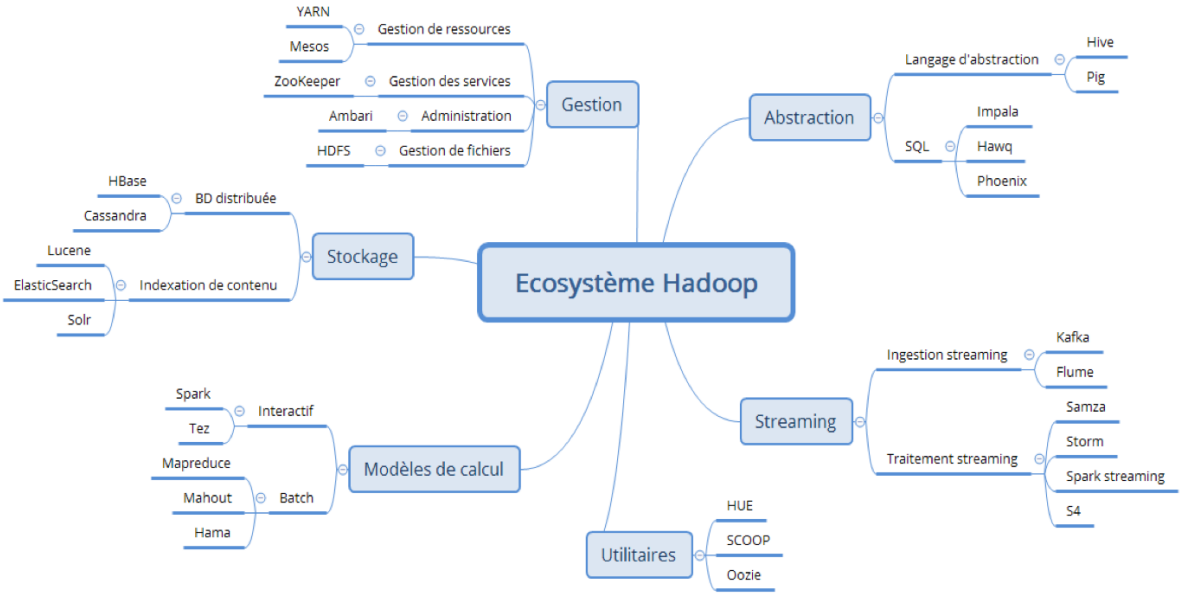



*Figure 13 : Carte heuristique de l'écosystème Hadoop (JVC, 2019)*

La configuration de base de de l'écosystème Hadoop est composé des technologies suivantes : Spark, Hive, PIG, HBase, Sqoop, Storm, ZooKeeper et Oozie. Pour la suite de notre développement, nous verrons plus en détail les technologies Spark et Hive. Mais avant cela, il est important de rappeler un concept : le *MapReduce*.

### MapReduce

*MapReduce* est un modèle de programmation faisant parti intégrante du fonctionnement de l'environnement Hadoop. En effet, ce modèle de programmation permet d'accéder aux données stockées dans HDFS. *MapReduce* va faciliter les traitements de gros volumes de données (allant jusqu'au petaoctects) en les réduisant en de volumes plus petits sur différents serveurs d'Hadoop. Cela étant, le résultat obtenu reste un résultat consolidé de toutes les données.

### Spark

Spark est un système informatique distribué utilisé pour de l'analyse dans le domaine du *Big Data* (Holden Karau, 2015). Il permet d'effectuer des analyses poussées sur des gros volumes de données et de manière très rapide. Grace à cela, les temps de calcul sont divisés par 100 comparé à une utilisation de *MapReduce* sur Hadoop. Il permet de développer des algorithmes alors impossibles à développer en *MapReduce*.

**Hive**

Hive est un système d'entrepôt de donnée interrogeable avec le langage similaire au SQL appelé HiveQL, pour le traitement et l'analyse de données stockées dans HDFS (Guller, 2015). Hive va permettre aux programmeurs de requêter sur des bases de données volumineuses avec un langage similaire au SQL, qui seront ensuite traduit en programme *MapReduce*. En effet, l'avantage de ce système va permettre aux programmeurs de développer des programmes sur un langage connu (SQL) pour la réalisation d'un programme type *MapReduce*.

## III – Réglementation des données

Dans le cadre des traitements des données personnelles, il y a des réglementations à respecter au niveau de la loi et pour cela, un certain nombre de règles ont été mise en place mais aussi des autorités compétentes sur le sujet.

### III.I – RGPD

Pour faire face aux enjeux éthiques, il faut respecter les enjeux juridiques. En effet, le RGPD (Règlement Général sur Protection de Données à caractère personnel) entre en vigueur en 2018 pour instaurer une responsabilisation des données à tous les niveaux. Depuis 2018, les traitements de données doivent respecter les principes fondamentaux cités à l'article 6 du RGPD (Arruabarrena, 2018) :

| Principes | Descriptif |
|---|---|
| **Licéité, loyauté et transparence** | Les données doivent être traitées de manière « licite, loyale et transparente au regard de la personne concernée ». Ce principe renvoie à l'information aux personnes concernées. |
| **Limitation des finalités** | Les finalités doivent être : « déterminées explicites et légitimes ». La finalité sous-tend l'interdiction de collecter des données à caractère personnel pour des finalités non déterminées ou pour d'autres finalités pour lesquelles elles ont été initialement collectées. |
| **Minimisation des données** | La collecte et le traitement des données à caractère personnel doit être « adéquate, pertinentes et limitées », et se limiter aux données qui sont nécessaires pour atteindre ces finalités. |
| **Exactitude** | S'assurer que les données à caractère personnel sont « exactes et tenues à jour » au regard des finalités pour lesquelles elles sont traitées, et les corriger le cas échéant. |
| **Limitation de la conservation** | Les données sont conservées pendant « une durée n'excédant pas celle nécessaire au regard des finalités pour lesquelles elles sont traitées ». S'assurer que les données à caractère personnel ne sont pas conservées plus longtemps que nécessaire pour atteindre les finalités pour lesquelles elles ont été collectées |
| **Intégrité et confidentialité** | Les données doivent être traitées de façon à « garantir une sécurité appropriée des données à caractère personnel », y compris la protection contre la perte, la destruction ou les dégâts d'origine accidentelle, à l'aide de la technologie appropriée |

*Tableau 3 : Principes fondamentaux du RGPD art.6 (Arruabarrena, 2018)*

Bien qu'il existe des lois sur la protection des données à caractère personnel, le RGPD est venu renforcer ces lois.

III.II – CNIL

La commission Nationale Informatique et Libertés (CNIL) est une autorité administrative indépendante française qui contrôle l'application de la réglementation sur la protection des données personnelles stockées en information ou au format papier. Le traitement des données personnelles comprend toutes opérations, y compris manuelles, impliquant la collecte, l'enregistrement, l'utilisation, la modification, la consultation, la communication ou l'effacement de données personnelles. On distingue 3 types de données qui nécessitent un peu plus notre attention : les données à caractère personnel, les données sensibles, les données interdites.

Les données à caractère personnel comprend toute information relative à une personne physique identifiée ou qui peut être identifié, ainsi que des données qui seules ou combinées entre elles, peuvent être rattachées à une personne (exemple : nom, adresse, téléphone, mails, y compris sur internet, numéro de carte bancaire, numéro d'acte décès, vie martiale etc…).

Les données dites « sensibles » peuvent être collectées pour des raisons précises et nécessitent donc un traitement renforcé (exemple : données relatives à la santé).

Les données « interdites » ne peuvent pas faire l'objet d'une collecte sauf cas exceptionnelles (exemple : extrait de casier judiciaire).

Les données personnelles sont protégées par une mise en œuvre de la loi. En effet, ces données disposent d'un cadre réglementaire parmi lequel figure :

- Loi Informatique et libertés du 6 janvier 1978 ;
- Loi pour une République numérique du 7 octobre 2016.

La CNIL a défini pour le secteur de l'assurance un Pack Conformité sur les données collectées pour le cœur de métier d'un assureur. Ce Pack intègre des normes simplifiées et donne des instructions sur l'utilisation des données personnelles, ainsi que des autorisations uniques du traitement de données « sensibles » et « interdites ».

Selon la CNIL, le traitement des données personnelles repose sur cinq principes fondamentaux :

- Finalité : Les données sont collectées pour des objectifs déterminés, explicites et légitimes ;
- Proportionnalité : Les données collectées doivent être adéquates, pertinentes et non excessives au regard des finalités ;
- Sécurité : La confidentialité des données doit être garantie par des mesures adaptées ;

- Consentement : Soit on obtient le consentement des personnes concernées ou bien on dispose d'un fondement légal justifiant le traitement ;

- Droit de rectification : Toutes les personnes ont droit d'accès, de rectification et d'opposition à l'utilisation de leurs données.

Les sociétés devront répondre aux contraintes réglementaires sous peine de sanctions pouvant aller jusqu'à 4% du chiffre d'affaires. Le règlement général sur la protection des données du 27 avril 2016 qui est entré en vigueur le 25 mai 2018 comprenait certaines dispositions déjà présente dans les règlementations applicables avant celle-ci. Sa violation est sanctionnée par les sanctions administratives, financières et pénales (5 ans d'emprisonnement et 1,5 millions d'euros d'amende).

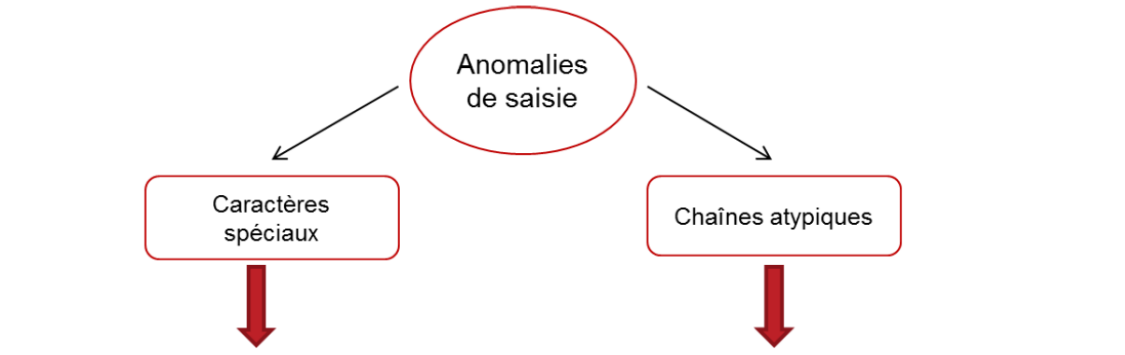
# Chapitre 3 : Screening, projet de détection d'anomalies

## I – Analyse de l'existant chez CNR

J'ai intégré Generali France en Septembre 2018 en tant que Data Analyst, afin d'assurer les missions portants sur les périmètres d'Epargne et de Retraite. A mon arrivé, il m'a été indiqué qu'une procédure existait déjà sur la détection d'anomalies : le Screening, mais qu'il était toujours possible d'améliorer l'algorithme (sous R à cet instant), ou encore d'améliorer les procédures d'analyses.

Le screening est l'action de contrôle de la présence et qualité de données. Il se réalise tous les trimestres sur les données personnes (identitaires) : nom d'usage, prénom, nom de naissance, prénom, date de naissance et sexe. L'objectif est, dans un premier temps, d'automatiser le processus à l'aide de R, afin d'obtenir un champ par données possédant les valeurs « 0 » lorsque la donnée est présente et « 1 » lorsqu'elle est absente. Ensuite, un champ sur la complétude du contrat est alimenté de la façon suivante : « 0 » si le contrat est complet (il ne manque aucune donnée parmi celles étudiées), et les valeurs supérieures à 0 signifient qu'il manque au moins une donnée. Dans un second temps, il est question d'analyser la qualité des données.

Afin de compléter l'analyse, un second niveau d'analyse a été mis en place : la détection d'anomalie. En effet, il est question de relevé les types d'erreur constatés dans les bases de données. On distingue deux grandes catégories d'erreurs : les caractères spéciaux et les chaînes atypiques. Les chaînes atypiques sont séparées en plusieurs sous catégories.




| Types | Exemples | Intitulé anomalie dans le screening |
|---|---|---|
| Caractères numériques | 0 à 9 | Caractères spéciaux |
| Caractères scientifiques | ² / * + = , µ < > ≤ ≥ ° % # | |
| Caractères de ponctuation | « » ( ) [ ] { } _ \ / : , ; ? . ! § … | |
| Caractères monétaires | € $ £ ¥ ¤ | |
| Autres caractères | @ & © ® ™ ← ↑ → ↓ ↔ • \| ¦ | |

| Types | Exemples | Intitulé anomalie dans le screening |
|---|---|---|
| Lettre unique | a, …, z, A, …, Z | Lettre unique |
| Lettre double | aa,…,zz, AA, …., ZZ | Lettre double |
| Lettre triple | aaa, …, zzz, AAA, …, ZZZ | Lettre triple |
| Civilité | Monsieur, M, Madame, MME… | Civilité |
| Mention né/née | Né, Née… | Chaîne atypique |
| Abréviations Saint/Sainte | St, Ste, … | |
| Mention épouse/veuf/ex | Ex, Epouse, veuf, we… | |

*Figure 14 : Type d'anomalies recherchées et intitulés dans le screening*

On appelle ces règles les expressions régulières. Ces règles s'appliquant sur les champs de type caractère, nous avions des autres règles pour les dates de naissance. En effet, il est intéressant de les étudier afin d'évaluer la cohérence des données. Après avoir calculé le taux de remplissage de cette donnée, nous avons mis en place des règles qui indiquerait le type d'erreur que présente une date de naissance :

- Affilié plus vieux que le doyen de France : Cette règle indique que la date de naissance de l'assuré ne peut pas être correcte car elle est antérieure à la date de naissance du doyen de France (né en 1903) ;

- Affilié trop jeune : Cette catégorie est utilisée pour la retraite. En effet, un assuré ne peut pas avoir un contrat de retraite alors qu'il a moins de 16 ans (âge minimum légal en France pour travailler) ;

- Date de naissance erronée : Suite à la fusion de nombreux portefeuilles, on constate des dates de naissance aux «01/01/1900», elles sont donc erronées. Pour rappel, la date «01/01/1900» est considérée comme le chiffre 1 sous Excel ;

- Null : Cette règle correspond aux dates de naissance antérieur à 1900 ou supérieur à 2020 dues à des erreurs de saisies ;

- Date de naissance incohérente : Lorsque la date de naissance est postérieure à la date d'effet (date de début du contrat d'assurance). En effet, il est impossible de créer un contrat d'assurance pour un individu qui n'est pas encore né.

Les procédures étant effectué sur R, nous sommes passés dans une dimension Big Data. Les données étant directement disponible sur un système de fichier distribué (HDFS), la programmation s'est faite en PySpark (utilisation du langage Python sous Spark), qui est le langage de programmation disponible afin de travailler sur les bases de données issu d'HDFS.

## II – Objectifs du projet

Jusqu'à présent, les données étaient livrées par le service informatique de Generali sur plusieurs fichiers. Nous recevions donc un fichier par système (7 au total). Il faut alors lancer un script R par système car les données étant très volumineuses (dépassant parfois 2 millions de lignes), il aurait été difficile de concaténer toutes les bases pour effectuer un seul script, le logiciel R ne l'aurait pas supporté. Les données étant livrées par le service informatique, ils n'étaient pas tous sous le même format ni même la forme : le nom des champs ne sont pas les mêmes. Les extractions étaient fournies tous les trimestres, ce qui aboutit à un total de quatre screening par an.

Avec la dimension Big Data et la mise en place des données disponibles sur une plateforme, l'objectif a été de réaliser le screening de manière mensuelle et de façon automatique. En effet, le service serait beaucoup plus indépendant des services transverses afin de répondre à ses problématiques. Ainsi, le service CNR serait capable de créer sa propre base de données et d'en effectuer les analyses sur une même bases contenant tous les systèmes étudiés.

Ainsi, le projet va se dérouler comme suit :

- Création de la base de données ;
- Premier niveau d'analyse : Taux de remplissage ;
- Deuxième niveau d'analyse : Typologies des anomalies ;
- Restitution des données sous Tableau (*Datavisualisation*).

Nous allons à présent voir plus en détails les différentes étapes du projet après avoir présenter les données étudiées. En plus de celles vues dans la partie précédente, nous ajouterons des nouvelles données à l'étude.

# III – Réalisation du projet

## III.I – Création de la base de données

Les données qui vont constituées notre base de données d'études vont être issues d'un référentiel de données appelé Référentiel Clients d'Entreprise (RCE). Le RCE permet d'avoir toutes les informations de tous les contrats chez Generali. Entre autres, toutes les informations que les clients ont fourni pour la souscription d'un contrat y sont présent. Il y a également les données financières liées aux contrats comme les montants de prime et autres montants calculés.

A travers le RCE, il est possible de retrouver les données d'identification du client (numéro d'assuré, le numéro du sinistres, le numéro de contrats, etc.). Ces informations sont importantes à avoir car le numéro client est un numéro unique et il peut être servi comme une clé lorsqu'il sera question de faire des jointures de tables. Le RCE possède également les données de type « personnes ». Par exemple, le type de client s'il est physique (sphère privée/sphère professionnelle), morale (associations, collectivités, entreprises, établissements, etc.). Le RCE présente aussi les données de types relations. Cela va concerner les données sur l'environnement et le foyer dans lequel vie un assuré, les liens de hiérarchies entre les personnes morales, les interlocuteurs de personnes morales, le rôle des différentes personnes par rapport au contrat, et les moyens de contact et de paiement relatifs au contrat.

En résumé, le RCE contient des données sur les contrats concernant:

- Les sinistres/prestations ;
- Le marketing ;
- Les données financières.

A partir de toutes ces données, il est possible pour Generali d'identifier les clients de différents segments, mais également d'identifier les clients prospects (personne cible à qui l'on souhaite vendre un produit). Ces données peuvent être également mis sous un dispositif de qualité de données :

- Recherche « intelligente » homophonique ;
- Normalisations (dont postale, moyen de paiement) ;
- Détection & déclaration des doublons.

Le RCE n'est pas qu'une seule base de données. En effet toutes les spécificités que présente ce référentiel est scindé en catégories. Il existe 88 tables contenant les données du RCE, mais pour notre étude et la création de notre base de données, nous utiliserons uniquement les données des tables suivantes :

- Personne : contient les données sur l'identité des personnes (nom, prénom, date de naissance, etc.) ;

- Contrat : contient toutes les données relatives au contrat d'un assuré (prime, montant de provision, date d'échéance, système d'information, etc.) ;

- Rôle contrat : indique si la personne est le premier assuré, le deuxième assuré ou le bénéficiaire du contrat ;

- Adresse : contient toutes les données du lieu d'habitation de l'assuré et si l'adresse est toujours valide ;

- Téléphone : contient les numéros de téléphone fixe et portable de l'assuré ;

- Courriel : contient les adresse mail des assurés ;

- Transcodification : les données étant codé de manière numérique, la table de transcodification va permettre d'obtenir les libellés.

Afin de créer notre table d'étude, nous avons besoins des bases de données citées précédemment. En effet, ces bases de données présentent tous les champs que l'on souhaite étudier dans les deux niveaux d'analyses. Voici les champs que l'on va étudier dans notre analyse :

- Nom d'usage ;

- Prénom ;

- Nom de naissance ;

- Lieu de naissance ;

- Date de naissance ;

- Numéro de téléphone fixe ;

- Numéro de téléphone portable ;

- Adresse mail ;

- Top NPAI ;

- Système d'information.

La variable Top NPAI indique si l'adresse de l'assuré est bien à jour. En effet, un assuré peut recevoir des courriers de Generali pour différentes raisons et si le courrier n'a pas pu atteindre le bon destinataire car le la personne recevant le courrier n'est pas la personne concernée, alors l'assuré rechercher aura la valeur 1 dans le champ Top NPAI, et 0 s'il n'y a aucun retour de courrier pour mauvaise adresse. Pour rappel, NPAI signifie N'habite Pas à l'Adresse Indiquée.

Le champ « Système d'information » est indispensable pour avoir le détail des analyses par systèmes. En effet, jusqu'à présent nous recevions un fichier par système (7 au total) mais maintenant que nous sommes capables de créer notre base d'étude autonomement, il est capital d'avoir la précision du système d'information afin de pouvoir effectuer des filtres également.

Les données concernant les partitions et les systèmes d'informations sont aussi essentielles pour les filtres que l'on souhaitera appliquer. Les dates de partitions vont nous permettre d'effectuer un suivi des évolutions mensuelles sur les taux de remplissage mais aussi sur l'évolution des anomalies.

Afin de réaliser la table d'étude, il est nécessaire d'effectuer quelques étapes intermédiaires. En effet, chaque variable que l'on a sélectionné dans chaque base de données sont extraite pour isoler les données que l'on souhaite avoir. Une fois que l'on a fait toutes les tables avec les données à étudier, nous pouvons joindre toutes les tables avec comme clé de jointure le numéro de contrat ainsi que la partition. Il est important de sélectionner la partition en tant que clé de jointure afin d'avoir des données cohérentes. Si nous n'avions pas sélectionné la date de partition avec le numéro de contrat, on aurait eu le risque que les données se mélangent entre les partitions, ce qui rendra fausses les informations de la base d'étude et par conséquent les résultats et analyses.

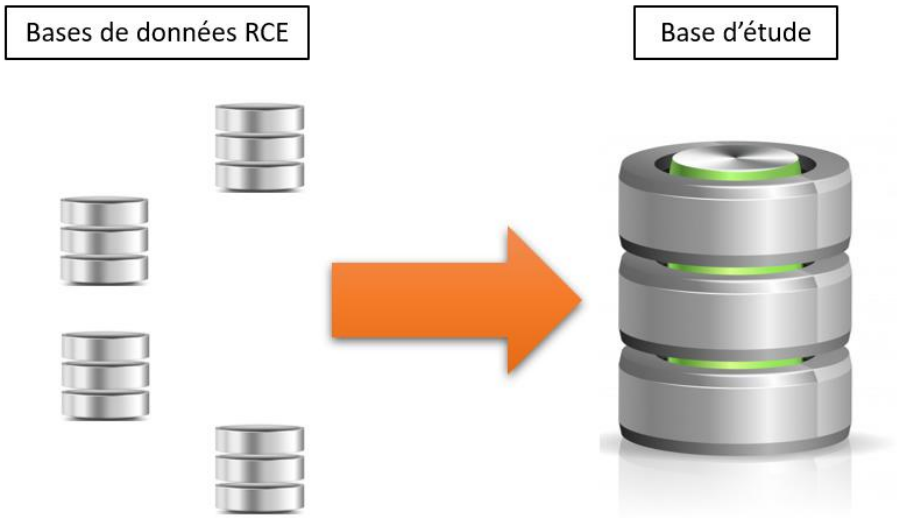



*Figure 15 : Schéma de la création de la base d'étude*

## III.II – Premier niveau d'analyse : Taux de remplissage

A présent que notre base d'étude à été créée, le premier niveau d'analyse peut être réalisé : le taux de remplissage (complétude).

Le taux de remplissage consiste à effectuer un comptage sur la présence/absence de valeur pour chaque champ précisé précédemment (hors « Top NPAI » et « Système d'information »).

La logique de calcul sous PySpark est la suivante : dans un premier temps, une fonction est créée afin de détecter si la donnée est absente dans le champ étudié. Ensuite, cette fonction est associée à une boucle afin de compléter les données vide par un 0 et dans un autre cas (pour les données contenant une valeur) du chiffre 1. Ces données seront stockées dans un vecteur. Une fois notre vecteur complété de 0 et de 1, nous l'associons à la table d'étude, ce qui ajoute un champ calculé à la table :

| Nom | Prénom |
|---|---|
| Dupont | Michel |
| | Catherine |
| Dupré | Aline |

| Nom | Prénom | Pres_Nom |
|---|---|---|
| Dupont | Michel | 0 |
| | Catherine | 1 |
| Dupré | Aline | 0 |

*Figure 16 : Exemple d'ajout du champ de complétude*

Sur les lignes ou la donnée est présente (soit toutes les lignes contenant un 0 dans la colonne d'indicateur), on évaluera la qualité de la donnée. Pour cela, nous faisons un deuxième niveau d'analyse avec des règles qui s'ajoutent à celles que l'on à vu dans la section II de ce chapitre. Ces règles ont été co-écrite avec l'aide du service de la Technique Assurance (TA), qui établisse également des analyses de qualité de données. Leur aide m'a été précieuse car cela nous a permis d'avoir un panel de règle d'anomalie beaucoup plus large et surtout de règle non existante par le passé sur les champs de contact (numéro de téléphone et adresse mail).

## III.III – Deuxième niveau d'analyse : Détection des anomalies

Pour détecter les anomalies, nous avons créés des règles de suspicions : chaque ligne qui aura les critères des règles seront qualifiées de suspect et donc en anomalies. Nous avons créé des plusieurs règles que nous pouvons regrouper en quatre grandes catégories :

- Règles de caractère ;
- Règles d'email ;
- Règles de numéro de téléphone ;
- Règles de date de naissance.

Nous ne détaillerons pas les règles pour les dates de naissance car nous l'avons présenté dans la section I de ce chapitre. Pour les autres grandes catégories de règles, nous feront une liste exhaustive des sous règles.

**Règles de caractère**

Pour cette catégorie de règle, nous avons listé six sous règles. En plus des caractères spéciaux et les chaines atypiques (cf. Chapitre 3, Section I), nous avons ajoutés les règles suivantes :

- La chaine de caractère ne contient pas de voyelle ;
  Exemple : nom d'usage → Dpnt
- La chaine de caractère contient un chiffre ;
  Exemple : prénom → Al4in
- La chaine de caractère contient plus de trois consonnes que ce soit au début, au milieu ou a la fin de la valeur ;
  Exemple : nom d'usage → Duponttt
- La chaine de caractère contient «et» ou «et-» ou commence par la civilité de l'assuré ;
  Exemple : prénom → Alain et Jean
- La chaine de caractère contient plus de trois fois la même lettre à la suite ;
  Exemple : prénom → Fraaancois
- La chaine de caractère contient une consonne isolée de la chaine (toutes les lettres comprises sauf les lettres D, L, H, N, M).
  Exemple : prénom → J Michel

Ces règles ne sont applicables que pour quelques champs analysés : Nom d'usage, Nom de naissance, Prénom et le Lieu de naissance.

**Règles d'email**

Les règles d'email ont été construites en plusieurs temps. Dans un premier temps, nous avons dû lister les mots qui ne peuvent pas constituer un email. Nous avons donc découpé l'email en deux parties : le *username* (chaine de caractère avant l'arobase « @ ») et le domaine (chaine de caractère après l'arobase « @ »). Le service de la Technique Assurance avait déjà établi la liste des *username* et des domaines invalides. Les deux listes étant très grande, nous allons citer que quelques exemples pour montrer la pertinence de la séparation des emails en deux.

| | |
|---|---|
| **Exemples de *username* invalide** | « pasdemail », « 123 », « rien », « sans », « pasdadresse » |
| **Exemples de domaine invalide** | « null », « test », « neant », « noemail » |

*Tableau 4 : Exemples de username et domaine invalide*

Pour cette deuxième grande catégorie de règles, nous avons listé 9 sous règles. Avec ces règles nous avons tenté d'être les plus exhaustif possible car on retrouve une multitude d'email n'étant pas dans un vrai format. Pour analyser pleinement un email, nous avons également pris en compte l'extension afin de voir s'il n'y a pas d'incohérence sur cette partie de l'email. L'extension d'un email correspond à sa dernière partie (exemple : …@gmail.com).

Voici les règles que nous avons appliqués aux emails :

- L'email contient plus d'un arobase (« @ ») ;
  Exemple : usern@me@domaine.fr

- Le *username* et le domaine sont identiques ;
  Exemple : truc@truc.fr

- Le domaine et l'extension sont identiques ;
  Exemple : username@fr.fr

- La longueur de l'email est trop courte (longueur de l'*username* ou du domaine inférieur à 2) ;
  Exemple : a@b.fr

- Répétition de lettres identiques ;
  Exemple : aaa.bbb@domaine.fr

- Répétition de chiffres ;
  Exemple : 12345@domaine.fr

- Ne contient que des chiffres dans le *username* et le domaine ;
  Exemple : 12345@6789.fr

- Contient le caractère « ? » ;
  Exemple : username?@domaine.fr

- Commence par un arobase ;
  Exemple : @username@domaine.fr

**Règles de numéro de téléphone**

Les règles de numéro de téléphone, nous nous sommes basés sur trois sous règles. En effet, les possibilités sont moins importantes que pour un email mais tout de même important. Voici les trois sous règles de cette troisième grande catégorie de règles :

- Le numéro se termine par trois paires égales ;
  Exemple : 01 23 98 98 98

- Le numéro ne contient qu'un chiffre ;

- Le numéro contient d'autres caractères que des chiffres.

Ces règles s'appliquent sur les numéros de téléphone fixe et portable.

Au cours de l'écriture de ces règles, nous avons tenté d'être les plus exhaustifs possible afin de palier tous les cas possibles que nous puissions imaginer. En effet, ces règles peuvent amener à être modifiées, améliorées, augmentées. En effet, nous ne sommes jamais à l'abris d'un email mal écrit mais sous un format que nous ne connaissons pas ou encore un numéro de téléphone qui n'est pas dans les fausses valeurs connues actuellement.

L'objectif de ce deuxième niveau d'analyse est de rendre compte des types d'anomalies que présentent les données. En effet, la bonne qualité des données d'une entreprise montre comment sont gérés les données des assurés et permet de donner une bonne image de la compagnie. Les données représentant un atout concurrentiel dans le milieu de la *data*, il est capital de savoir restituer ces données dans un format plaisant pour les décisionnaires de l'entreprise et c'est en cela que la partie *Reporting : Datavisualisation* est essentiel pour une bonne compréhension des analyses. Dans un milieu ou sont mélangé les demandes métiers, les professionnels de la données et les techniciens, les rapports réalisés doivent être compréhensible par tous dans un langage simple.

## III.IV – Reporting : Data-visualisation sous Tableau Software

Dans cette partie nous allons voir comment nous avons effectuer notre tableau de bord, comment avons-nous pu accéder à nos données exportées dans un entrepôt de données, mais également le dynamisme qu'offre l'outil Tableau.

La production de données chiffrées est une mission importante de CNR. La présentation des résultats doit être claire et surtout lisible par tous. C'est pourquoi, il est toujours intéressant de continuer de développer les *reporting* au rythme des nouveaux indicateurs et des outils disponibles.

**Sauvegarde des données dans le *Cloud***

Afin de disposer des données sous Tableau, nous devons sauvegarder notre table d'étude dans un système de fichier distribué appelé HDFS (cf. Chapitre 2, Section 3, Partie 3). Pour cela, nous procédons de la manière suivante :

```
#On commence par créée une table temporaire à partir de notre dataframe
df_output17.registerTempTable("df_output17")

#on enregistre ensuite la table temporaire dans la table créée via un code sql classique
# toujours stocker les tables à la racine de hive =hdfs://nameservice/met_dsm/TA/hive/MaTable
#    (jamais dans des sous répertoires)
spark.sql("""

CREATE EXTERNAL TABLE IF NOT EXISTS met_ta.CNR_SCREENING_2_aout2020
STORED AS PARQUET
LOCATION 'hdfs://nameservice/met_dsm/TA/hive/CNR_SCREENING_2_aout2020'

AS

SELECT * from df_output17

""")
```

*Figure 17 : Sauvegarde de la table d'étude*

Cette fonction va permettre de sauvegarder nos données dans un format plus adapté pour Tableau , le format *PARQUET*. Cette partie du code va partitionner nos données afin de sauvegarder rapidement sur le système HDFS, mais lors de la connexion avec Tableau, on la retrouvera en une seule partie. Voici comment les données ont été partitionnées dans l'entrepôt de données:

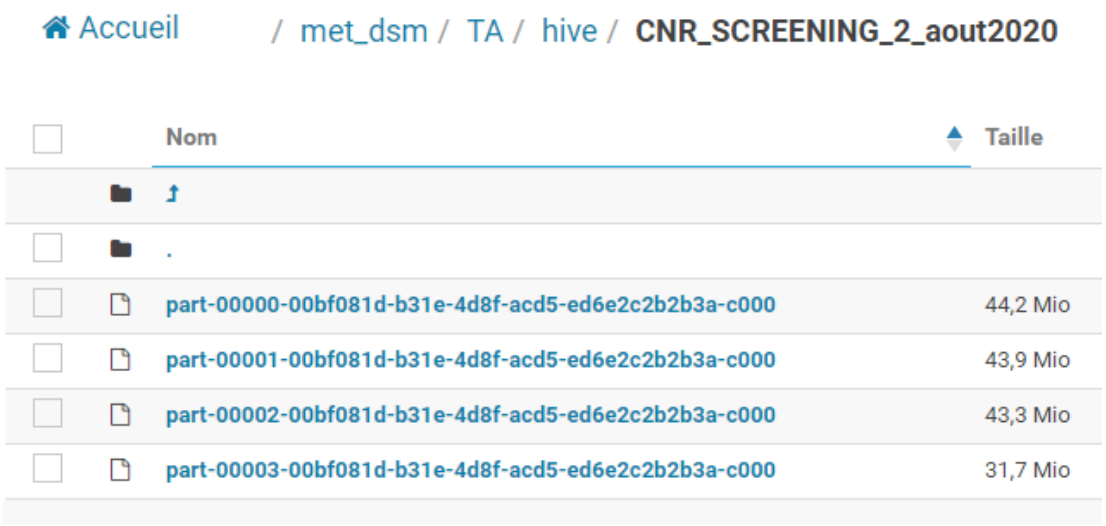



*Figure 18 : Partitionnement des données dans l'entrepôt de données*

**Connexion avec la table sous Tableau**

Maintenant que les données ont été exportées puis partitionnées dans l'entrepôt de données, nous pouvons directement nous connecter avec la table afin d'y effectuer nos tableaux de bords sous Tableau. Voici comment se présente la connexion avec la table :

En ouvrant Tableau, nous devons nous connecter au serveur que Generali utilise afin de pouvoir accéder à tous les schémas et tables qui sont disponible dans l'entrepôt de données. Generali utilise le serveur *Cloudera Hadoop*.

Premièrement, nous devons sélectionner le schéma correspondant (le chemin ou la table a été crée). Dans notre cas, nous devons aller sur «met_ta» :

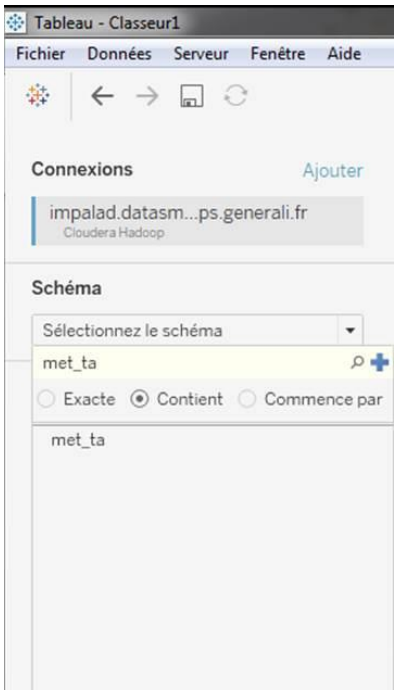



*Figure 19 : Choix du schéma*

Les développements étant assez récents chez Generali, le service CNR n'a pas encore de répertoire dédié pour exporter ses travaux. C'est pour cela que nous utilisons un schéma dédié au service de la Technique Assurance, temporairement en attendant que notre répertoire métier soit créé.

Une fois que nous avons pu sélectionner notre schéma, nous pouvons rechercher notre table d'étude :

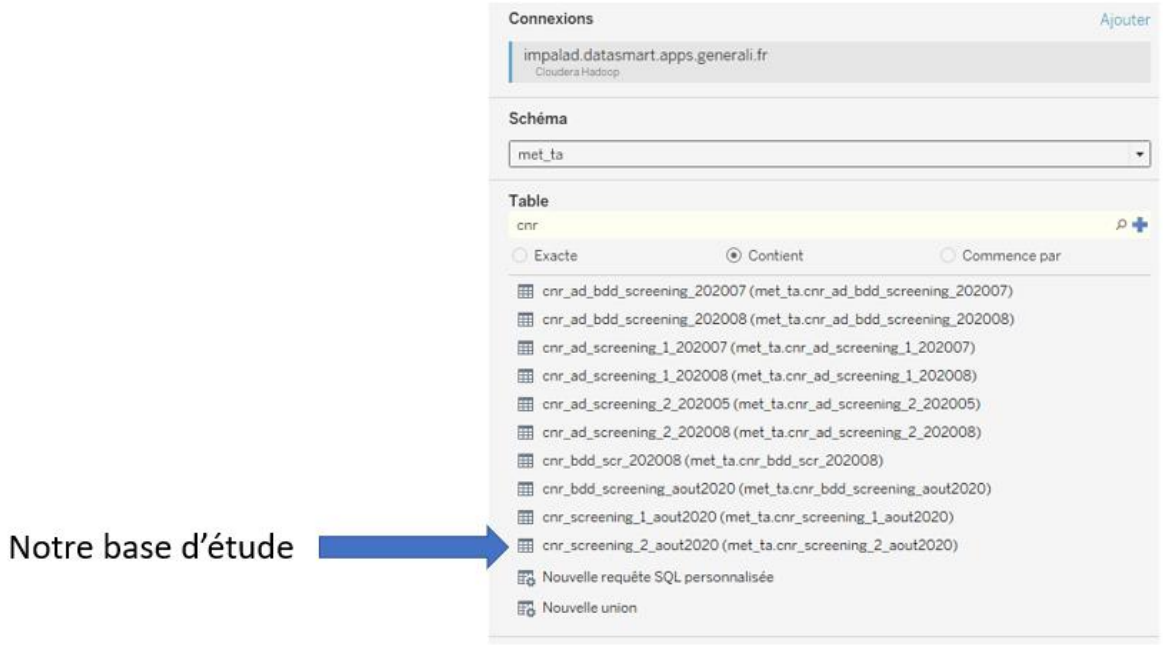



*Figure 20 : Connexion avec la table d'étude*

Une fois notre table sélectionnée, nous pouvons réaliser les tableaux de bords que l'on souhaite.

Les possibilités de conception de tableaux de bords sous Tableau est immensément grande. En effet, l'avantage que représente Tableau pour son lecteur est qu'il est possible de réaliser des tableaux de bords dynamique et interactif. Cela permet d'avoir des rendus visuels de très bonne qualité. En plus de ça, il est possible de se connecter à ses données selon deux façons différentes : soit par la connexion via un entrepôt de données, soit par l'importation des données manuellement. En effet, il est possible d'importer ses propres données (un fichier Excel par exemple) pour en effectuer les tableaux de bords. Tableau est adapté pour les demandes métier dans un environnement de *Business Intelligence*. Il est possible de représenter des données sur un plan géographique ou encore de créer des graphiques qui peuvent être utilisé comme un filtre. Par exemple sur un graphique représentant la répartition hommes/femmes, il est possible d'afficher l'âge moyen de chaque sexe lorsque l'on clique dessus.

Avec Tableau, nous pouvons filtrer sur tout ce que nous pouvons trouver pertinents. En prenant en compte qu'une partie de notre base de données, d'autres graphiques peuvent se mettre à jour automatiquement. Il est également possible d'extraire une table de données à partir d'un graphique. Par exemple, sur un graphique montrant la répartition des systèmes d'informations (nommons les Systèmes A, B et C), il est possible d'extraire les données du système A uniquement lorsque l'on clique dessus. Développer des tableaux de bords avec Tableau est très intéressant car son large panel de possibilité répond encore plus aux besoins métiers du service CNR et de Generali.

Voici quelques exemples de tableaux de bords qui a été réalisé sur le projet d'étude :

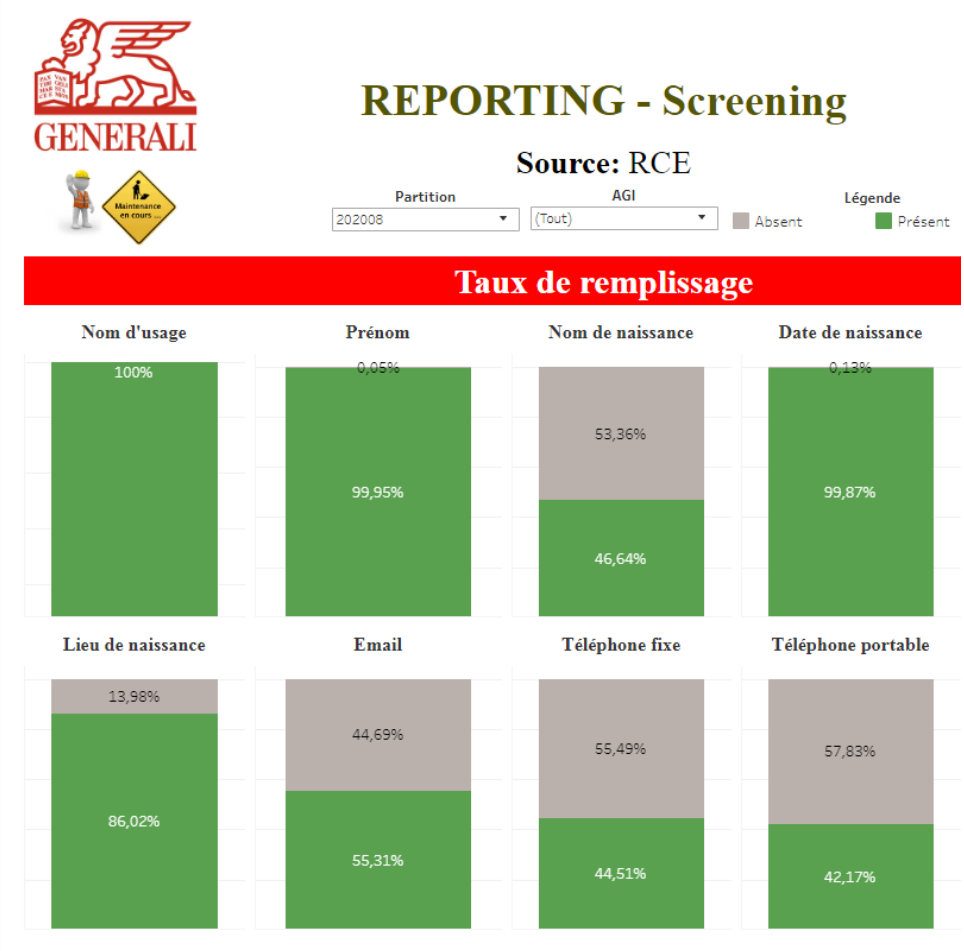



*Figure 21 : Exemple de reporting Tableau (Taux de remplissage)*

Les données ci-dessus représentent le taux de remplissage de chaque donnée analysée. Nous avons en filtre la partition ainsi que l'AGI (Système d'information). La sélection de la partition va nous permettre de voir les évolutions aux cours des mois passé et l'AGI va nous permettre de comparer les systèmes entre eux et de savoir quel système est le mieux rempli parmi les données analysées.

Voici à présent un tableau de bord concernant les anomalies de données avec la possibilité de filtrer en sélectionnant les parties d'un graphique :

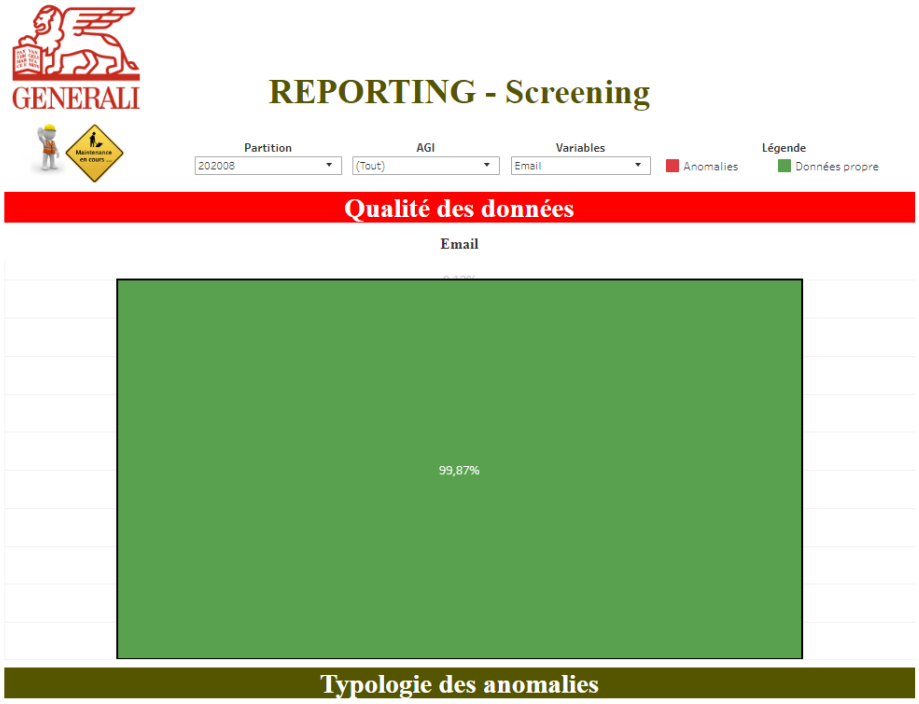



*Figure 22 : Exemple de reporting Tableau (Qualité des données)*

Cet exemple permet de voir la part des données des email qui sont de bonne et de mauvaise qualité, parmi les données présentes. Presque toutes les données sur les emails sont de bonne qualité, et dans ce tableau de bord, nous avons sélectionné les données propres. Par conséquent, aucune typologie d'anomalie n'est affichée.

Voici ce qu'il se passe lorsque l'on clique sur les données en anomalie sur le graphique supérieur :

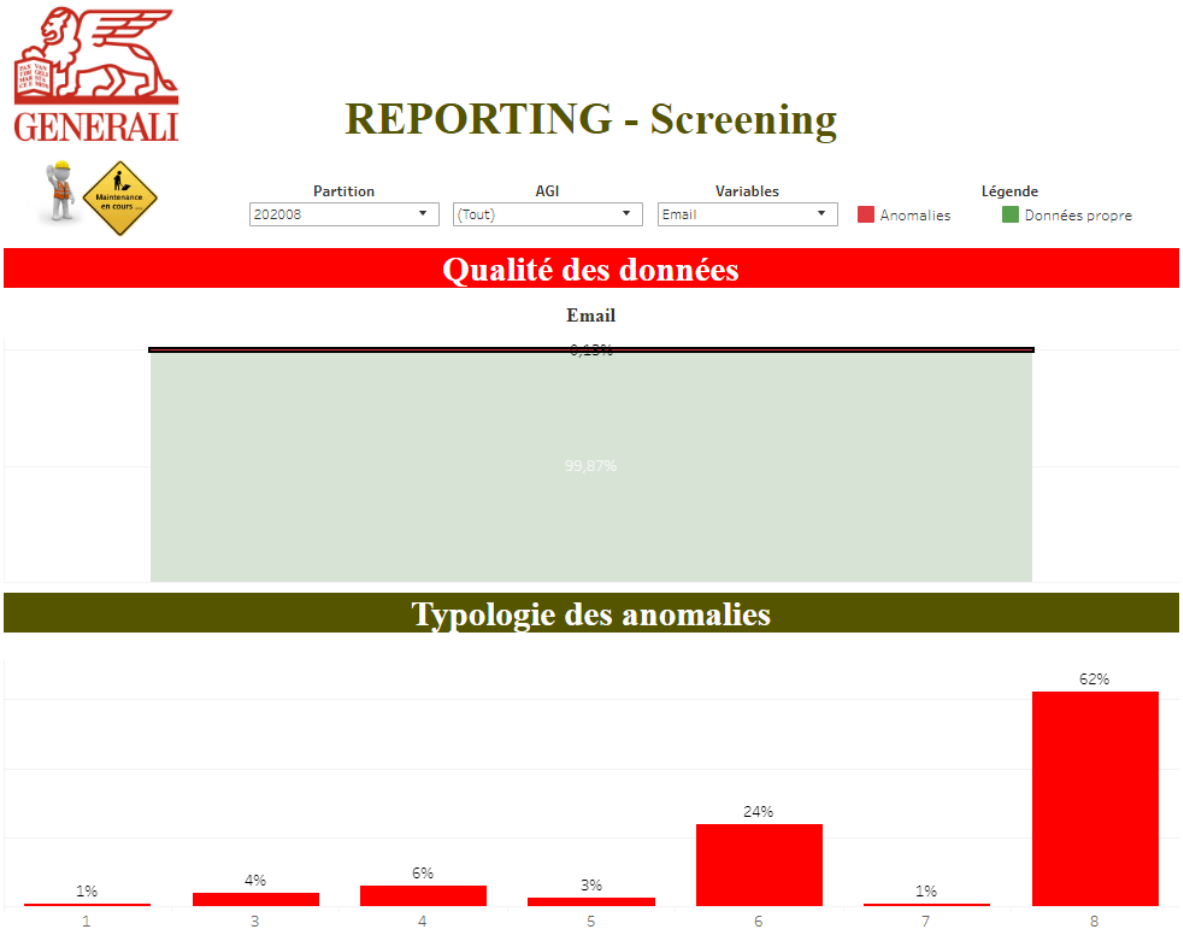



*Figure 23 : Exemple de filtre sur graphique*

En cliquant sur les données en anomalies, nous avons un nouveau graphique qui apparait avec la typologie des anomalies. On peut également changer la variable que l'on souhaite analyser avec la création d'un filtre de données qui prend en compte les variables étudiées dans la table.

# Conclusion

A travers ce mémoire, nous nous sommes intéressés à la qualité des données d'une entreprise dans une dimension Big Data. Nous avons présenté un état de l'art en trois parties en définissant notamment ce qu'est une donnée de bonne qualité, mais aussi définir l'environnement Big Data dans lequel j'ai réalisé ce projet d'analyse de qualité de données.

L'étude réalisé et présenté au cours de ce mémoire a permis de montrer une approche différente de l'analyse de la qualité d'une donnée. Nous avons démontré comment les données sont importantes pour une entreprise d'un point de vue *business* mais aussi sur l'image que donne une entreprise sur la gestion de ses données. En effet, les données représentent un atout concurrentiel car les données nous permettent d'améliorer les procédures mais aussi d'adapter la demande par rapport à l'offre.

A la suite de cela, nous avons appliqué les différentes façons de détecter une anomalie de données dans un exercice de fiabilisation des données identitaires et contactes. Nous avons premièrement calculé le taux de remplissage des champs analysés puis dans un second temps, déterminé des règles de calcul afin de catégoriser les anomalies. Enfin, pour restituer les données, nous sommes passés par l'outil Tableau afin de bénéficier de son aspect visuel, dynamique et interactif.

Conscient que le projet est en partie terminé, il reste tout de même des développements à réaliser afin d'avoir exploiter pleinement les outils et d'améliorer les rendus pour en retirer le plus d'information possible.

La réalisation de ce projet a été très enrichissant pour moi, d'un point de vu personnel et professionnel. En effet, ce projet m'a permis de mettre en pratique les méthodes vues en cours sur des sujets intéressants. Cela m'a permis d'apprendre de nouvelles méthodes de travail mais également de nouveaux outils permettant d'enrichir mes connaissances théoriques et pratiques.

# Bibliographie


ARRUABARRENA, Beatrice. Big Data & Ethique: la qualité des données en débat, 2018.

BARBARA H. WIXOM et JEANNE W. ROSS. How to Monetize Your Data, 2017.

BASTIEN, L. Data Monetization : tout savoir sur la monétisation des données, 2019.

BERTI-EQUILLE, Laure. Qualité des données. Ingénierie des Systèmes d'Information, 2004, pp. 117-143.

BERTI-EQUILLE, Laure. La qualité et la gouvernance des données: au service de la performance des entreprises, 2012.

BERTI-ÉQUILLE, Laure. Qualité des données, 2018, pp. 23.

BOURANY, Thomas. Les 5V du big data, Cairn.info, 2018.

BRASSEUR, Christophe. Data Management : Qualité des données et compétitivité, 2005.

DAWN E. Holmes. Big Data: A Very Short Introduction, 2017.

GULLER, Mohammed. Big Data Analytics with Spark, 2015, pp. 1-15.

KARAU, Holden, KONWINSKI, Andy, WENDELL, Patrick, ZAHARIA, Matei. Learning Spark: Lightning-Fast Big Data Analysis, 2015.

TALBI, Ilyes. Machine learning pour la détection d'anomalies, 2019.

JVC, Juvénal. Hadoop : l'état de l'art des technologies du Big Data, 2019.

OLSON, Jack E. Data Quality: The Accuracy Dimension, 2003.

RAHHAL, Ibrahim, MEZZOUR, Ghita. Etude Big Data, Apports des Big Data pour le suivi de l'activité économique et la prévision, 2020, pp. 55.

SHAFAKHATULLAH KHAN, Mohammed, KAUSAR, Mohammad Abu, NAWAZ, Shaik Shah. BigData Analytics Techniques to Obtain Valuable Knowledge,2018, pp. 1-14.

# Liste des figures

# Liste des tableaux

# Annexes

## Annexe A: Exemple de script PySpark

```
addlibsys_pers_ctr_ = addlibsys_pers_ctr.join(transco_,
                                              on=(transco_["code_vars"] == addrefprd_pers_crt["precisionsystemeorigine"]),
                                              how="left").drop(transco_["code_vars"])
```

```
addlibsys_pers_ctr_ = addlibsys_pers_ctr_.withColumnRenamed("lib_vars", "lib_precisionsystemeorigine")
```

```
addlibsys_pers_ctr_1 = addlibsys_pers_ctr_.join(transco_statut_ctr, on=["systemeorigine",
                                                                        "statut_ctr"],how="left").drop(transco_statut_ctr["sy
```

```
addlibsys_pers_ctr_1 = addlibsys_pers_ctr_1.sort(["idtechcontrat","idtechpersonne","role"],
                                                 ascending=[False]*3).dropDuplicates(["idtechcontrat","idtechpersonne"])
```

```
addlibsys_pers_ctr_1 = addlibsys_pers_ctr_1.select(["societedugroupe", "systemeorigine", "lib_systemeorigine",
                                                    "precisionsystemeorigine", "lib_precisionsystemeorigine",
                                                    "idtechpersonne", "identifiantrce", "nom_usage", "nom_naissance",
                                                    "prenom", "autresprenoms", 'datenaissance', 'lieunaissance',
                                                    'datedeces', 'sexe', 'top_npai', 'tel_poste_perso', 'tel_poste_prof',
                                                    'tel_mobile_perso', 'tel_mobile_prof', 'tel_principal',
                                                    'adressecourriel', "idtechcontrat", "identifiant",
                                                    "statut_ctr", "lib_statut_ctr", "cdprodgauer", 'lib_produit'])
```

```
cdt1 = F.col("systemeorigine").isin(['07','06','13','09','02','12'])
cdt2 = ((F.col("precisionsystemeorigine").isin(['13','09','06'])) & (F.col("systemeorigine") == '11'))
filter_perim = (cdt1) | (cdt2)
```

```
addlibsys_pers_ctr_2 = addlibsys_pers_ctr_1.filter(filter_perim)
```

```
addlibsys_pers_ctr_3 = addlibsys_pers_ctr_2.withColumn("dt_partition", F.lit(str(ReqPartitions)))
```

## Annexe B: Interface de Tableau