Mémoire M1 MIAGE APP

-

ANDRIN Mathieu

Comment garantir la qualité des données tout au long de leur cycle d'utilisation ?

SOMMAIRE

1. Introduction. 3
2. Contexte et émergence de la problématique. 4
   2.1. EDF, une entreprise leader dans le monde de l'énergie. 4
   2.2. Focus sur l'activité gazière : Pôle GAZ et Portefeuille Client Contrat 5
   2.3. Illustration de la qualité des données. 6
3. La Donnée : Définition et cycle de vie. 7
   3.1. Qu'est-ce qu'une Donnée ? 7
   3.2. Le cycle de vie d'une Donnée : de la création à la destruction. 8
4. État des lieux actuel de la Qualité des Données. 10
   4.1. Qualité des données : Définition et principales problématiques. 10
   4.2. Les dimensions de la Qualité des Données : un aperçu détaillé. 11
   4.3. Solutions actuelles pour améliorer la Qualité des Données : Typologie et analyse. 14
   4.4. Méthodologies d'amélioration de la Qualité des Données. 16
5. Analyse critique de l'état actuel. 22
   5.1. Diagnostic de la situation actuelle. 22
   5.2. Les limites et implications en qualité des données. 24
   5.3. Perspectives d'amélioration et innovation 26
6. La Qualité des Données en entreprise : cas pratiques et solutions. 28
   6.1. Power BI, une automatisation au profit de la qualité de l'information. 28
   6.2. Mise en place d'un SAS de contrôle à l'entrée des données. 29
   6.3. Des solutions au quotidien : pratiques et outils pour maintenir la qualité. 30
7. Conclusion 31
8. Références générales. 33

1. Introduction.

Actuellement en alternance en tant qu'étudiant en Master 1 MIAGE à l'Université Paris Dauphine-PSL et travaillant chez EDF en tant que data analyst au sein du portefeuille GAZ, je suis amené à réfléchir sur des problématiques essentielles liées à la gestion des données. Dans le cadre de la rédaction de mon mémoire de première année, j'ai choisi de me concentrer sur la gestion de la qualité des données tout au long de leur cycle de vie.

La donnée est aujourd'hui le moteur de l'économie moderne, jouant un rôle de plus en plus central dans les prises de décisions des organisations. Cette discipline, en plein essor au cours des dernières années, se situe à l'intersection des mondes de l'informatique, des mathématiques, et de la gestion de l'information. Cependant, cette expansion rapide s'accompagne de nouveaux défis, notamment celui de la qualité des données.

En effet, une mauvaise qualité des données peut avoir des conséquences graves : des clients et des collaborateurs mécontents, ainsi que des pertes financières significatives. Selon le Data Warehousing Institute, ce problème aurait un coût estimé à plus de 600 milliards de dollars par an pour l'économie américaine, soit environ 2,5% de son PIB de 2022, ce qui souligne ainsi l'ampleur du sujet.

De plus, en tant que data analyst, je suis régulièrement confronté à ces défis liés à la qualité des données, qui figurent parmi les enjeux majeurs de ce métier.

Ce mémoire s'articulera autour de la question suivante :

Comment garantir la qualité des données tout au long de leur cycle d'utilisation ?

Ce sujet de réflexion a émergé à la suite de plusieurs échanges avec mon équipe, notamment avec Mme Brugère, ainsi qu'avec mon école, représentée par M. Yger. L'un des principaux objectifs était de restreindre la portée du problème pour le rendre abordable, tout en sélectionnant un sujet déjà exploré par d'autres auteurs, afin de s'appuyer sur des travaux existants.

Nous commencerons par définir les différents termes de la problématique, ce qui permettra de mieux cerner les enjeux auxquels nous sommes confrontés.

Le plan du mémoire s'articulera comme suit :

1. **Présentation de l'entreprise :** Dans un premier temps, nous reviendrons brièvement sur les principales activités de l'entreprise. Un focus particulier sera mis sur le département dans lequel j'évolue, avec une recontextualisation de la problématique par rapport à mon expérience en alternance et aux projets auxquels j'ai participé.
2. **Définition des concepts clés :** Ensuite, nous approfondirons les éléments constitutifs de la problématique : qu'est-ce qu'une donnée ? Qu'est-ce qu'un cycle de vie des données ? Ces concepts seront analysés pour fournir une base solide à notre réflexion

3. **Étude de l'existant en matière de qualité des données :** Cette partie du mémoire se concentrera sur l'analyse de la qualité des données, en définissant ses principes fondamentaux et en présentant les principales méthodologies et solutions existantes dans le domaine.

4. **Réflexion critique sur l'état de l'art :** Nous examinerons ensuite les différentes méthodes identifiées, en comparant leurs avantages et leurs limites. Nous discuterons des pistes d'amélioration possibles pour ces méthodes.

5. **Mise en œuvre des solutions en entreprise :** Pour conclure, nous mettrons en lumière les principales solutions mises en place au sein de l'entreprise, en les évaluant à travers le prisme des analyses effectuées dans les parties précédentes. Cela permettra de relier la théorie à la pratique, et d'évaluer l'efficacité des approches adoptées.

Enfin, une conclusion viendra synthétiser les apports de ce mémoire, en soulignant les contributions principales et les perspectives pour la suite.

2. Contexte et émergence de la problématique.

2.1. EDF, une entreprise leader dans le monde de l'énergie.

Électricité de France (EDF), fondée en 1946 par le gouvernement français, est un acteur majeur de l'industrie énergétique mondiale. Créée dans le contexte de la reconstruction post-guerre, EDF s'est rapidement distinguée par son expertise en producti on nucléaire, inaugurant sa première centrale à Chinon en 1963. Depuis, l'entreprise a diversifié ses activités vers les énergies renouvelables, incluant l'hydraulique, l'éolien et le solaire.

EDF s'engage également dans le développement durable, avec des filiales comme EDF Renouvelables, Cyclife et IZIVIA, se concentrant respectivement sur les énergies renouvelables, la gestion des déchets nucléaires, et la mobilité électrique. En tant que fournisseur historique d'électricité en France, EDF joue un rôle clé dans la régulation du marché énergétique et collabore étroitement avec les autorités nationales.

Avec l'État français comme actionnaire unique, EDF contribue activement à la formulation des politiques énergétiques et à la transition énergétique. Face aux défis environnementaux mondiaux, EDF continue d'investir dans les énergies propres et les solutions innovantes, soulignant ainsi son engagement à construire un avenir énergétique durable.

2.2. Focus sur l'activité gazière : Pôle GAZ et Portefeuille Client Contrat.

Lors de ma première année d'alternance, j'ai évolué au sein de la Direction Sourcing Economy and Finance, qui dépend du Pôle Client Services et Territoires (CST) et regroupe 30 000 salariés répartis dans différentes directions et filiales.

Cette direction comprend un département gaz dont l'objectif est d'optimiser l'équilibre économique du sourcing de gaz de l'entreprise, c'est-à-dire d'acheter le gaz nécessaire pour répondre à la demande des clients, tout en limitant les aléas liés aux variations des marchés de l'énergie.

Les différents acteurs de ce Pôle sont les suivants :

- **Coût et marché :** Responsable de la création des offres et des prix, ainsi que de la gestion des marges pour couvrir les risques.
- **Optimisation :** Chargé des ordres d'achat de gaz en bout de chaîne.
- **Prévision :** Équipe qui anticipe les fluctuations de consommation du portefeuille.
- **Portefeuille Client et Contrat (PCC) :** L'équipe dans laquelle j'effectue mon alternance.

Le Portefeuille Client et Contrat (PCC) se situe en amont des processus du pôle et, est en partie, responsable des analyses de données. Nous établissons un bilan des variations réelles du portefeuille, telles que les nouveaux clients et les sorties de clients. L'équipe joue également un rôle comptable, et de gestionnaire de la souscription des capacités de consommation journalières des gros consommateurs.

L'équipe PCC, par son utilisation des données de consommation des clients, collabore également de manière interfonctionnelle avec les métiers du marketing, du commercial, et de la comptabilité.

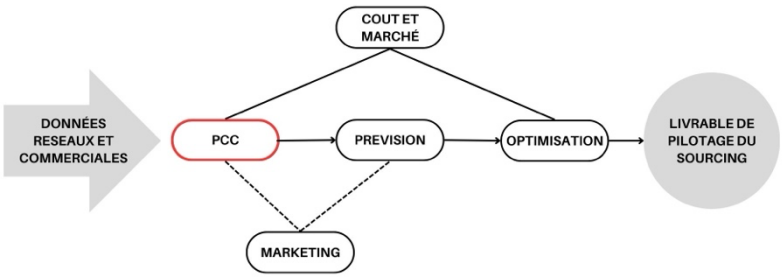

Illustration 1 : Schéma descriptif du pôle GAZ chez EDF

Ci-contre, un graphe représentant la structure simplifiée des différentes entités du pôle gaz. L'objectif est de mettre en avant le flux métier suivant : PCC rapporte les analyses des données réalisées, Prévision effectue des Prévisions courts termes et moyens termes basés sur ces dernières, puis l'équipe Optimisation gère les achats de gaz pour équilibrer le stock sur les quatre prochaines années. L'équipe Coût et marché gère les différentes options de sourcing dans les offres destinées aux clients.

Ainsi, des enjeux financiers importants sont présents pour l'entreprise concernant le gaz, même si cela représente une proportion plus petite comparée à l'électricité.

2.3. Illustration de la qualité des données.

Comme développé dans la partie précédente, une grande partie du travail opérationnel au sein de l'équipe PCC consiste à acquérir et analyser des données. Cela inclut l'explication des variations et des écarts éventuels, mais dans certains cas, des éléments de mauvaise qualité peuvent compliquer ces analyses. Nous reviendrons sur la définition de la qualité dans la suite de ce rapport.

Durant ma première année d'alternance, j'ai participé activement à la vie opérationnelle de l'équipe, notamment en ce qui concerne la qualité des données. De plus, j'ai été impliqué dans trois grands projets transverses qui ont donné du sens à cette problématique. En voici un aperçu :

- **Création de rapports Power BI[1] :** Mon activité principale en tant qu'alternant a été de remplacer des rapports Excel générés à l'aide d'outils par des rapports BI. Lors de cette transition, j'ai dû reprendre toutes les sources de données utilisées, dont j'avais une connaissance limitée. Pendant le développement, je me suis donc interrogé sur ces sources, notamment sur leur qualité.
- **Développement et mise en place d'un nouvel outil :** Un autre élément clé de l'apparition de cette problématique a été l'introduction d'un nouvel outil, destiné à remplacer Le référentiel du portefeuille gaz. Son objectif est de récupérer des flux de données provenant des commerciaux et des gestionnaires de réseaux, puis de les agréger pour faciliter les étapes d'analyse, et finalement permettre le sourcing des offres gaz. Ce nouveau dispositif a été développé par une équipe SI, et lors des points de suivi auxquels j'ai pu participer, les principaux sujets concernaient justement la gestion des données.
- **Développement d'un outil d'analyse :** Durant mes premiers mois dans l'entreprise, j'ai effectué des analyses et développé des outils pour faciliter l'explication de certains écarts dans un rapport. Les principales causes d'écart identifiées étaient liées à des différences temporelles entre des processus, ainsi qu'à la présence, dans certains cas, d'informations incohérentes.

La question de la qualité des données est donc omniprésente dans mon travail de data analyst, que ce soit dans des projets tels que ceux décrits plus haut ou dans des études ponctuelles. Cela m'a conforté dans mon désir de mieux comprendre ce phénomène.

[1] Power BI : Outil développé par Microsoft, permettant d'automatiser l'extraction, l'analyse et la visualisation des données.

3. La Donnée : Définition et cycle de vie.

3.1. Qu’est-ce qu’une Donnée ?

Une donnée est un élément fondamental d'information qui peut être collecté, enregistré et analysé. Elle représente une mesure, une observation ou un fait brut, symbolisant une information ou un concept qui peut être manipulé par des systèmes informatiques pour générer des connaissances. Dans la littérature, le concept d’information correspond à l’état final de la donnée, une fois que cette dernière a subi certains traitements.

Les données peuvent être structurées (dans des bases de données), semi-structurées (comme les fichiers XML ou JSON), ou non structurées (tels que les textes, images, vidéos). Elles se distinguent également par leur granularité (détail), leur format (structure), leur nature (quantitative ou qualitative), et leur source (humaine ou machine). Les données peuvent être classifiées en plusieurs types, en fonction de leur origine et de leur utilisation :

1. **Données de référence :** Il s'agit des informations de base qui définissent les entités avec lesquelles une organisation interagit (clients, produits, fournisseurs, etc.).
2. **Données transactionnelles :** Elles capturent les interactions et les événements opérationnels au sein des systèmes (ventes, achats, etc.).
3. **Données dérivées :** Issues de traitements et d'analyses, elles sont utilisées pour prendre des décisions (statistiques, modèles prédictifs).
4. **Métadonnées :** Des données sur les données, fournissant des informations contextuelles telles que la source, le format, la date de création, etc.

En général, plusieurs rôles peuvent être définis concernant l’utilisation des données :

- **Fournisseur :** Responsable de la création ou de la récupération des données.
- **Fabricant :** Conçoit, développe et maintient les données et l’infrastructure associée.
- **Consommateur :** Donne un sens métier à la donnée.
- **Gestionnaire :** Régit la donnée tout au long de son cycle de vie.

Cruciales pour la prise de décision, l'opérationnalité, et l'innovation, les données sont au cœur des analyses et des décisions stratégiques, ainsi que des opérations quotidiennes et des initiatives innovantes. Comprendre et définir précisément les données est essentiel pour garantir leur qualité tout au long de leur cycle de vie.

3.2. Le cycle de vie d'une Donnée : de la création à la destruction.

Comme vu précédemment, la gestion des données, ou data management, est un processus omniprésent dans presque toutes les entreprises, englobant chaque étape depuis la collecte jusqu'à la suppression des données. Ce processus complexe et structuré se décompose en plusieurs phases distinctes, chacune ayant un impact significatif sur la qualité globale des données.

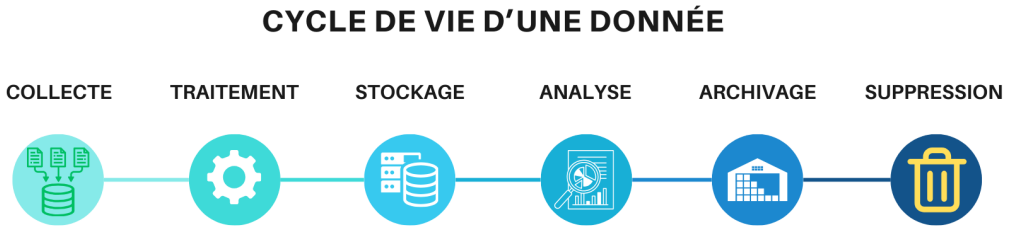

Illustration 2 : Cycle de vie d'une Donnée

1. **Collecte :**

   La collecte des données est une étape cruciale et extrêmement sensible dans le cycle de vie des données. Une collecte rigoureuse et de haute qualité conditionne la fiabilité et la précision des étapes suivantes. Lors de cette phase, le principe du "Garbage In, Garbage Out" (GIGO) s'applique : des données incorrectes ou imprécises saisies au départ entraînent des résultats erronés en aval. Par exemple, des erreurs manuelles lors de la saisie des données, comme l'entrée d'identités incorrectes, peuvent compromettre l'intégrité des données collectées.
   Outils et techniques :

   - Formulaires électroniques.
   - Capteurs IoT (Internet des objets) pour la collecte automatique de données.
   - Systèmes de gestion de l'information (SGI).

2. **Traitement et Nettoyage :**

   Cette phase, bien qu'elle soit placée en deuxième position, peut également réapparaître à d'autres étapes du cycle (avant/pendant le stockage ou avant les analyses). Elle inclut le traitement et le nettoyage des données brutes pour améliorer leur qualité et les rendre exploitables. Les logiciels ETL (Extract, Transform, Load) facilitent cette étape.

3. **Stockage :**

   Le stockage des données, bien que perçu comme une phase moins sensible que la collecte, joue un rôle crucial dans la préservation de la qualité des données initiales. Il reflète la qualité des données telles qu'elles ont été collectées.
   Le stockage doit se conformer aux régulations locales et internationales en matière de protection des données et de vie privée
   Pratiques courantes :

   - Double stockage des données pour éviter les pertes accidentelles.
   - Utilisation de bases de données relationnelles.
   - Conformité aux normes comme le RGPD (Règlement Général sur la Protection des Données).
   - De plus, cette étape doit également appliquer une gestion rigoureuse des modifications. Les changements doivent être contrôlés et tracés pour maintenir l'intégrité des données.

4. **Analyse :**

   L'analyse des données consiste à générer des informations exploitables pour la prise de décision au sein de l'entreprise. Cette étape permet de créer des suivis opérationnels précieux à partir des données stockées, de réaliser des études ponctuelles et de développer des modèles prédictifs.
   Outils et techniques :

   - Logiciels d'analyse statistique (R, Python).
   - Plateformes de business intelligence (BI) comme Tableau ou Power BI.
   - Outils de Machine Learning pour les prévisions et les prédictions.
   - Cette étape est sensible en termes de qualité des données.

5. **Sauvegarde et Archivage :**

   Une fois que les données ne sont plus nécessaires pour les opérations courantes, elles sont déplacées vers des systèmes de stockage secondaires. Ces données archivées peuvent être réutilisées pour des analyses futures ou en cas de litige. Lors de cette étape, les données doivent être conformes aux politiques de rétention des données.

6. **Suppression :**

   La suppression des données est la dernière étape du cycle de vie. À la fin de la période de rétention, les données doivent être supprimées de manière sécurisée pour prévenir tout risque de récupération non autorisée.
   Outils et techniques :

   - Protocoles de suppression sécurisée (par exemple, DOD 5220.22-M).
   - Logiciels de gestion de la suppression des données.
   - Certification de destruction des données.

Ces phases permettent de garantir une gestion optimale de la donnée par rapport à ses caractéristiques. Selon les éléments de littérature, d'autres étapes peuvent être données, ou avec des appellations différentes, mais le concept global reste celui défini plus haut.

4. État des lieux actuel de la Qualité des Données.

4.1. Qualité des données : Définition et principales problématiques.

Comme nous l'avons mentionné au début de ce rapport, l'importance des données ne cesse de croître au sein des organisations à travers le monde. Les disciplines transversales à ce domaine suivent la même tendance, notamment la qualité des données. Dans cette partie, nous allons explorer les principes de ce domaine, ainsi que les problèmes récurrents en termes de qualité.

En premier lieu, il est essentiel de comprendre que la qualité des données ou QD[2] garantit une prise de décision optimale, car elle améliore l'efficacité opérationnelle et la confiance des consommateurs d'informations. Même avec des valeurs économiques importantes, des taux d'erreur relativement faibles peuvent entraîner des coûts considérables.

Maintenant que nous avons décrit les principaux enjeux, il est important de donner un sens à la notion de qualité, qui est un concept subjectif. Dans notre cas, des dimensions de la QD ont été établies, telles que la précision ou la complexité. Ces dimensions fournissent un cadre et des ordres de grandeur pour les différentes caractéristiques des données. Nous les étudierons en détail dans la partie suivante.

La norme ISO 8000 standardise la qualité des données à l'échelle internationale. Son objectif est notamment de garantir la conformité aux réglementations et d'améliorer la collaboration entre les divers partenaires commerciaux.

Une petite illustration simplifiée de l'impact de la qualité des données :

- Consommation annuelle totale de gaz en France : 500 TWh
- Prix du TWh : 1 000 000 * 40[3] = 40 000 000 €
- Coût total : 500 * 40 000 000 = 20 000 000 000 € (20 milliards)

Supposons que la mauvaise qualité des données soit de 2% ; sur la consommation totale, cela pourrait représenter un écart de 400 millions d'euros (2% de 20 milliards), et une variation de 0,1%

[2] QD ⇔ Qualité des Données
[3] Représentatif du prix du MWH des moyennes du marché sur 2023

représenterait, quant à elle, 20 millions d'euros. Évidemment, la majorité des éléments extérieur n'est pris en compte ici, mais cela donne une idée de l'ampleur que peut avoir cette justesse.

Nous allons maintenant étudier les principaux problèmes de qualité des données, ce qui nous permettra de mieux comprendre l'intérêt des mesures ainsi que des solutions. Voici une liste des principaux problèmes en qualité des données :

1. **Données manquantes ou incomplètes :** Absence de valeurs dans certaines colonnes ou enregistrements. Données essentielles non renseignées.
2. **Données dupliquées :** Présence de doublons dans les bases de données. Entrées répétées sous des formes légèrement différentes.
3. **Incohérences des données :** Données contradictoires dans différents systèmes ou à différents moments. Variabilité des formats de données (dates, adresses, etc.).
4. **Erreurs de saisie :** Fautes de frappe ou d'orthographe. Utilisation incorrecte des champs de saisie.
5. **Problèmes de précision et d'exactitude :** Données inexactes ou incorrectes. Estimations ou approximations non fiables.
6. **Problèmes de validation :** Données qui ne respectent pas les règles de validation ou les contraintes définies. Valeurs en dehors des plages autorisées.
7. **Problèmes de format :** Incohérences dans le format des données (par exemple, des dates écrites différemment). Utilisation de formats non standardisés.

Dans la suite de cette étude de l'existant, nous définirons les principales dimensions en lien avec les problématiques évoquées ci-dessus. Ensuite, nous examinerons les types de solutions qui ont été mises en place. Nous conclurons par un aperçu des principales méthodologies de QD.

4.2. Les dimensions de la Qualité des Données : un aperçu détaillé.

Dans cette partie, nous étudierons les principales dimensions de la qualité des données (QD). Ce sont des caractéristiques essentielles pour la suite de l'état de l'art car elles constituent la base des méthodologies que nous examinerons par la suite. En qualité des données, les dimensions sont donc quantifiables par des mesures. Il en existe différentes, plus ou moins complexes pour chaque dimension.

Il est notable que la définition de ces composantes a été influencée par les principales problématiques définies précédemment.

De nombreux chercheurs ont défini diverses dimensions, certaines étant très proches les unes des autres. En voici une liste non exhaustive :

- **Précision (Accuracy) :**

La précision évalue la différence entre l'information contenue dans la base de données et la réalité. En d'autres termes, cela revient à identifier les valeurs incorrectes. Un niveau d'imprécision d'une valeur peut également être mesuré.

- **Complétude (Completeness) :**

Cette caractéristique a pour objectif de donner un ordre de grandeur sur les valeurs manquantes.

- **Cohérence (Consistency) :**

L'objectif ici est de déterminer la proportion de valeurs cohérentes par rapport aux standards et aux règles métier définis dans le système.

- **Conformité (Validity) :**

La conformité vérifie le respect des formats (adresse e-mail, numéro de téléphone, etc.) dans les données. Elle est proche de la dimension de précision, car une donnée qui ne respecte pas le format a de grandes chances d'être incorrecte. Le nombre de valeurs non conformes correspond à la somme des valeurs pour chaque catégorie.

- **Âge (Timeliness) :**

Cette mesure vise à évaluer la fraîcheur des données. L'importance de cette dimension varie selon l'environnement ; certaines données peuvent nécessiter une actualisation journalière, mensuelle, ou annuelle. Il est également nécessaire de disposer d'une date de dernière mise à jour pour chaque donnée en base.

- **Unicité (Uniqueness) :**

La dimension d'unicité quantifie l'importance des données qui se répètent dans la base. Il est essentiel de définir ce qu'est un doublon dans le jeu de données (par exemple, un client sous deux contrats en même temps).

- **Disponibilité (Accessibility) :**

Cette dimension évalue la simplicité d'accès aux informations du point de vue du consommateur de données. Le contrôle peut être réalisé à travers des questionnaires réfléchis soumis aux consommateurs, bien que la subjectivité puisse entrer en jeu.

- **Facilité d'opération (Ease of Operation) :**

La facilité d'opération définit la simplicité d'utilisation des systèmes de gestion des données. Elle peut être quantifiée à l'aide de questionnaires, mais aussi par des métriques telles que la complexité du modèle de données. En effet, plus cette complexité est grande, plus les requêtes se complexifient (jointures, filtres, etc.).

- **Traçabilité (Tracability) :**

Cette évaluation assure que chaque donnée peut être retracée à sa source et que toutes les modifications sont enregistrées, garantissant ainsi l'intégrité et la transparence des données.

- **Compréhension (Understandability) :**

L'objectif de la compréhension est de s'assurer que les données sont présentées de manière claire et compréhensible pour les utilisateurs, ce qui est essentiel pour leur utilisation efficace dans les analyses et la prise de décision. Pour cela, on peut utiliser un taux de compréhension des données pour chaque type d'utilisateur.

Dans chaque dimension on peut définir des éléments propres aux attentes de l'entreprise, par exemple un taux de précision seuil en dessous duquel on considère une erreur.

Dans certains éléments de la littérature on peut retrouver une classification des différentes dimensions en 4 piliers principaux, Les éléments intrinsèques, contextuels, représentatifs et d'accessibilités. Ce regroupement permet de rassembler les dimensions les plus proches et donc d'optimiser les analyses. Voici une répartition des éléments vus plus haut dans ces catégories :

Intrinsèque	Contextuelle	Représentative	Accessibilité
Précision	Age	Compréhension	Disponibilité
Cohérence	Complétude	Conformité	Facilité d'opérations
Unicité			

Les dimensions peuvent comporter différentes mesures pour établir une moyenne globale. Également chacune d'elle peut avoir une importance différente selon le domaine d'une entreprise. Il peut donc être nécessaire de faire varier la pondération de certaines d'entre elle pour établir un bilan global de la qualité.

4.3. Solutions actuelles pour améliorer la Qualité des Données : Typologie et analyse.

Dans cette partie, nous étudierons les différents types de solution pour améliorer la qualité des données.

En premier lieu il est nécessaire de savoir qu'il existe différents types d'approche sur **les solutions** :

1. Diagnostiques : analyse mathématique permettant de faire ressortir les caractéristiques des données et donc intrinsèquement la qualité.
2. Préventives : Réparage des anomalies en amont, ce sont des solutions proches des domaines d'architecture des systèmes informatiques.
3. Adaptatives : S'adapter à ce qui existe en trouvant des parades pour retrouver ce que l'on cherche vraiment.
4. Correctives : Amélioration de la qualité sur le long terme en venant corriger d'éventuelle anomalies.

Il est également possible de caractériser une solution comme active si elle met en place des actions, ou passive dans le cas inverse. Nous allons faire passer en revue les principales solutions en qualité des données

- **Modèle Conceptuel de Données** (MCD) - Préventive :

Le Modèle Conceptuel de Données (MCD) joue un rôle très influent sur la qualité des données. Il définit les différentes relations et entités présentes dans une base. Lors de sa conception, diverses mesures sont prises en compte, telles que la complexité structurelle du modèle ou le nombre d'associations et d'entités. La structure du MCD peut simplifier ou compliquer la mise en place de la qualité des données.

- **Intégrité référentielle -** Préventive :

L'intégrité référentielle est un bon exemple de solution passive car elle implique la mise en place de règles techniques sur les données, telles que des bornes maximales, des contraintes de format, ou le contrôle des clés primaires pour éviter les doublons. Elle est dite passive car, une fois ces contrôles intégrés, elle gère automatiquement les valeurs erronées. Cette intégrité est étroitement liée au MCD défini précédemment.

- **Consolidation –** Adaptative :

La consolidation des données intervient lors de la collecte, lorsque l'on dispose de la même information provenant de deux sources différentes. La consolidation consiste à choisir entre les deux visions. Par exemple, si la source A est jugée plus fiable que B, la règle suivante peut être appliquée : "Si A est présent dans A, alors A, sinon B."

- **Vérification d'après vérité terrain -** Diagnostique/Corrective :

Ce mécanisme est probablement le plus sûr en termes de diagnostic. Il consiste à vérifier la justesse d'une donnée en interrogeant directement la source initiale de l'information, ce qui permet de fournir plus de contexte à une donnée. Bien que cette solution soit très puissante, elle est également très lourde. Dans mon travail en tant que data analyst, cette solution est régulièrement utilisée pour les informations importantes qui n'ont pas pu être traitées automatiquement par les solutions mises en place en amont.

- **Filtre –** Adaptative :

Le filtre consiste à supprimer manuellement certaines informations avant de faire ces analyses, par exemple enlever les valeurs aberrantes ou nulles d'un échantillon. Cette solution est efficace mais n'est pas pérenne.

- **Audit –** Diagnostique :

Un audit est une analyse de l'état actuel des données. Il permet de mettre en avant les forces et les faiblesses du système de gestion des données.

- **Suivi des données** – Préventive :

Cette solution consiste à assurer un pilotage et un contrôle permanent des données. En effet, il ne suffit pas de vérifier les données ponctuellement.

- **Nettoyage** - Correctives :

Le nettoyage est une solution élémentaire qui consiste à retirer les éléments indésirables des données (totalement erronées, doublons, etc.).

- **Règle de gestion métier** - Préventive :

Les règles de gestion sont des éléments métier qui permettent de garantir la cohérence (dimension de la qualité des données). Elles sont implémentées à l'entrée des bases de données. Par exemple, un achat de 100 euros pourrait être enregistré en catégorie B car le prix total est compris entre 50 et 150 euros.

Il existe donc un grand nombre de solutions existantes, des outils informatiques ont été développés pour faciliter la mise en place de ces dernières. Nous pouvons prendre les exemples de Power center – Informatica (minimum 5 000 $ par an) pour le nettoyage par ETL ou encore Avellino – Discovery (80 000 €) pour l'audit.

4.4. Méthodologies d'amélioration de la Qualité des Données.

Nous venons de détailler les différentes solutions que nous pourrions qualifiées de « primitives ». Nous allons à présent passer en revue certaines des méthodologies les plus populaires en gestion de la qualité des données. La plupart de ces solutions comportent plusieurs phases qui reprennent les éléments vus précédemment.

Ces méthodologies suivent approximativement des phases similaires, en premier lieu une cartographie de l'état actuel, suivi d'une mesure de la qualité données, puis une étape d'amélioration/réparation, et dernièrement la création d'un pilotage continu. De toute évidence, toutes les méthodes n'appliquent pas forcément les mêmes étapes.

Voici ci-dessous, un tableau comparatif des principales méthodologies donné par Batini [7] (en rouge les 3 que nous étudierons). Nous pouvons remarquer que chaque système exploite une phase d'analyse de donnée, et de mesure (sauf CIHI). Cependant, ils ne s'appuient pas tous sur une phase de modélisation des processus, ou d'analyse des besoins. De plus, il est possible de voir que toutes les méthodes ne sont pas flexibles quant à l'utilisation de nouvelle mesure.

Phases / Méthodologies	Analyse des données	Analyse des besoins de QD	Identification des éléments critiques	Modélisation des processus	Mesure de la qualité	Possibilité d'ajouter des mesures
TDQM	+	+	+	+	+	Non
DWQ	+	+	+		+	Oui
AIMQ	+		+		+	Non
CIHI	+		+			Non
DQA	+		+		+	Oui
IQM	+				+	Oui
ISTAT	+				+	Non
AMEQ	+		+	+	+	Oui
COLDQ	+	+	+	+	+	Oui
DaQuinCIS	+		+	+	+	Oui
QAFD	+	+			+	Non
CDQ	+	+	+	+	+	Oui
ORME-DQ	+		+	+	+	Oui

Les méthodologies appliquant toutes les étapes devraient-être plus complètes mais également plus complexes (plus chères) à appliquer.

Avant de commencer le tour d'horizon des trois méthodes, il est essentiel de préciser que certains mécanismes de ces méthodes sont complexes et seront survolés dans les explications.

- TDQM [6] (Total Data Quality Management)

Richard Y. Wang

Cette méthodologie, très populaire dans le domaine de la qualité des données, est inspirée du cycle de Deming (Plan-Do-Check-Act), reprenant les quatre phases principales, Définir – Mesurer – Analyser – Améliorer, comme représenté sur le schéma-ci contre.

TDQM est difficilement applicable en complément ou en support, car il est conçu pour être un élément complet. L'un des concepts clés de cette stratégie est de considérer la donnée de la même manière qu'un objet physique.

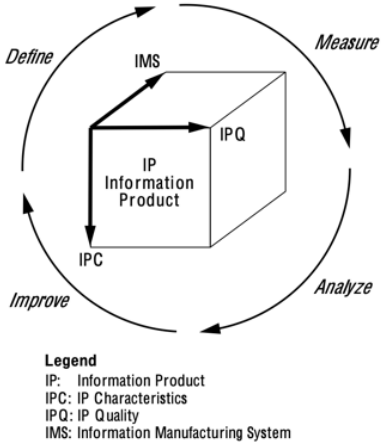

Illustration 3 : Cycle TDQM Définir-Mesurer-Analyser-Améliorer

Voici les principales phases :

1. Définir la donnée :

Cette partie consiste à effectuer un état des lieux de la donnée et de son environnement. Elle se divise en 3 parties principales : Tout d'abord, une étude préliminaire sur la donnée est réalisée, on cherche à définir ces principales applications et caractéristiques. Dans un second temps, on cherche à établir les exigences de la QD, pour cela tous les rôles (Fournisseur, Fabricant, Utilisateur et Gestionnaire) exposent leur point de vue sur chaque dimension de la qualité des données. Puis dernièrement, une analyse du système de production de la donnée permet de mettre en lumière les rôles de chaque acteur ainsi que les différents traitements suivis par la donnée. Cette étape de cartographie est cruciale et permettra de simplifier les mesures et les analyses, grâce à la Matrice d'Analyse de la Fabrication de l'Information.

2. Mesurer la donnée :

Durant cette phase, les différentes mesures et dimensions que nous avons définies précédemment vont être établies. Ces dernières peuvent donc suivre des aspects fondamentaux tels que l'exactitude, ou la cohérence, mais aussi des règles commerciales plus complexes et des indicateurs spécifiques à la fabrication de l'information

3. Analyser la donnée :

À partir des résultats de la phase de mesure, il s'agit de déterminer les causes d'éventuelles anomalies dans les données, dans le but de trouver une solution de régulation. La méthode suggère d'étudier en profondeur les hypothèses et les justifications. Des méthodes d'analyses statistiques ou des diagrammes de Pareto [4]peuvent être effectués pour simplifier ces éléments.

4. Amélioration de la qualité des données :

Sur la base de l'analyse, il convient maintenant de mettre en avant les principaux éléments à améliorer. Une illustration serait une amélioration de la cohérence entre les besoins commerciaux et les données ou encore l'application de nouvelles procédures opérationnelles.

[4] Diagramme de Pareto : Classification des différentes causes d'un phénomène par importance.

- **AIMQ [5] (A Methodology for Information Quality Assessment)**

Yang W. Lee, Diane M. Strong, Beverly K. Kahn, Richard Y. Wang

AIMQ est la méthodologie avec le moins de phases parmi celles que nous allons étudier, car elle ne comporte pas de phase d'état des lieux sur la structure existante. Elle se concentre cependant sur la réalisation d'un benchmark [5]entre différentes organisations.

Trois composantes principales sont mises en avant dans la création de cette méthode :

1. PSP/IQ :

L'objectif de cette approche est de regrouper les dimensions de la qualité des données (QD) en quatre grandes catégories, en distinguant le produit (la donnée) du service (le Système d'Information)

- **Fiabilité** (Sound) : Produit conforme aux spécifications.
- **Fiable et cohérente** (Dependable) : Produit qui répond aux attentes.
- **Utile** (Useful) : Service conforme aux spécifications
- **Utilisable** (Usable) : Service qui répond aux attentes des consommateurs.

	Conforms to specifications	Meets or exceeds consumer expectations
Product Quality	Sound information IQ dimensions Free-of-error Concise representation Completeness Consistent representation	Useful information IQ dimensions Appropriate amount Relevancy Understandability Interpretability Objectivity
Service Quality	Dependable information IQ dimensions Timeliness Security	Usable information IQ dimensions Believability Accessibility Ease of operation Reputation

Illustration 4 : Modèle PSP/IQ

Cette classification permet d'évaluer dans quelle mesure les données et les services associés répondent aux attentes et aux spécifications des consommateurs et des gestionnaires.

2. IQA :

L'élément IQA est chargé d'identifier les mesures les plus pertinentes par rapport au modèle PSP/IQ. Des études sur la fiabilité et l'interdépendance des dimensions ont démontré que, bien que la qualité des données soit multidimensionnelle, elle constitue un élément unique. De plus, un écrémage des mesures associées à chaque dimension a été effectué pour réduire le nombre de mesures et ainsi simplifier l'application de la méthode. Finalement, des notations (moyenne de toutes les dimensions) sont associées aux quatre divisions du modèle PSP/IQ, ce qui permet d'identifier les forces et faiblesses de la qualité des données.

3. IQ :

Enfin, cette composante se concentre sur l'analyse de la situation en prenant en compte plusieurs points :

   - Analyse des résultats précédemment établis, avec l'ajout d'éléments d'analyse tels que des pondérations.

[5] Benchmark : Comparaison qualitative, qui étudie les différentes techniques et résultats obtenus par les organisations concurrentes.

- Benchmark par rapport aux meilleures entreprises du même secteur, afin de visualiser les points d'amélioration possibles.
- Comparaison des différentes perceptions de la qualité des données du point de vue des gestionnaires et des consommateurs, dans le but d'identifier les écarts de vision et de renforcer la synchronisation entre les différentes parties.

- (ORME) [7] DQ assessment methods

Batini Carlo, Barone Daniele, Mastrella Michelle, Maurino Andrea, Ruffini Claudio

DQ est une méthode qui a été dérivée plusieurs fois par différents chercheurs comme Pipino et al. (2002) et Maydanchik (2007) ou encore la version ORME-DQ de Batini et al. (2009), qui est celle que nous allons étudier.

Cette méthode comporte également quatre phases, ces dernières ont été renommé pour faciliter la compréhension :

1- Reconstruction de l'état actuel.

Pour savoir quels sont les éléments à prioriser, on regarde tous les flux de données utilisés et échangés entre les différentes parties d'une organisation. Cette phase permet de mettre en avant les principales données, ainsi que leurs principaux utilisateurs et gestionnaires. A la fin de cette étape, nous avons donc un plan des point critiques (utilisateurs, processus, générateur) des données.

2- Identification des risques économiques liés à la qualité des données

L'objectif de cette étape est de classer les processus en fonction de la valeur des pertes potentielles. Sur cette étape les valeurs économiques peuvent être classifiées en trois grandes familles :

- Absolue : 100 euros
- Pourcentage : 10% de main d'œuvre en plus
- Qualitative : Élevé/Moyen/Faible

3- Mesure de la qualité des données.

Durant cette étape, à partir des mesures définies plutôt (partie 4.3), il va falloir choisir pour chaque donnée quelle est la métrique la plus appropriée. Une fois ce choix effectué, il va falloir déployer des sondes sur les données sélectionnées qui permettront d'effectuer des mesures sur la justesse des informations. Pour avoir une vision sur les coûts économiques, il est possible d'associer à une sonde les coûts économiques directs et indirects liés à l'anomalie. Par exemple, si ma sonde détecte un taux d'erreur de 10 % sur les données de consommation de mes clients, cela peut engendrer un coût économique important lié à une surproduction ou à une sous-production.

4- Pilotage des risques.

En utilisant les sondes définies à l'étape précédente, certains seuils vont déclencher une alerte sur des données. Ce qui est intéressant durant cette phase est que les manières dont on va définir.

Les seuils d'alerte sont multiples. Il est notamment possible d'utiliser la méthode de l'analyse discriminante[6], en cherchant à classifier les éléments dans une classe « NoLoss » ou « Loss »

La méthodologie comporte également un framework (cadre) qui est composé de plusieurs phases que nous allons détailler.

a. Module extracteur de connaissance -knowledge extractor (Phase 1 et 2):

Ce module permet de réaliser les matrices de la phase 1, les matrices sont ensuite stockées dans le knowledge repository (répertoire de connaissance).

b. Module d'évaluation de la qualité des données - Data quality assesment (Phase 3) :

Utilise des méthodes et des algorithmes pour mesurer la qualité des données récupérées par les sondes, en passant par les dimensions étudiées précédemment (exactitude syntaxique, complétude, ...). C'est également dans ce module que l'on définit les spécifications des sondes (ID,mesure,heure), Dans ce modèle, les sondes (i.e. Probes) sont des éléments qui se déclenchent à intervalle bien précis et qui donnent un état des lieux de la qualité à un instant T.

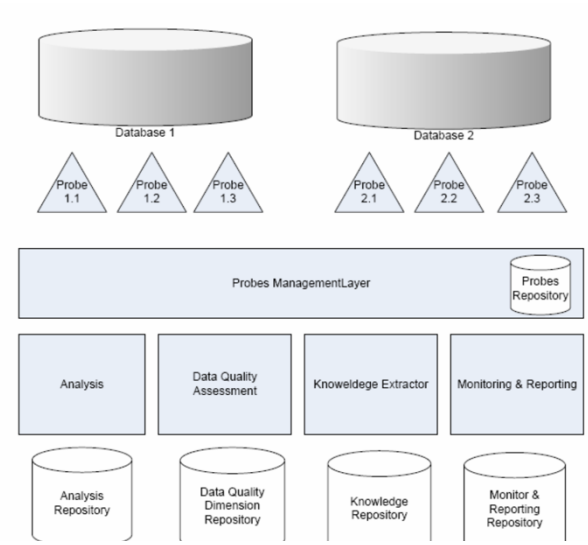

Illustration 5 : Les différents modules de l'ORME-DQ

c. Module de gestion des sondes :

Ce module permet de gérer la mise en place et le pilotage des sondes.

d. Module d'analyse - analysis:

Le module permet de traiter les informations récupérées par les sondes et calculées par l'évaluation de la qualité des données. Il gère également le stockage des informations sur les mesures, notamment le moment où elles sont effectuées.

e. Module de pilotage - monitoring & Reporting :

Ce-dernier est en quelque sorte l'IHM de contrôle, c'est lui qui remonte les alertes sur la qualité, et qui permet de suivre l'évolution des différentes mesures dans le temps.

[6] Méthodes de classification utilisée dans des disciplines telles que le Machine Learning.

5. Analyse critique de l'état actuel.

5.1. Diagnostic de la situation actuelle.

Nous avons étudié certaines solutions primitives, ainsi que leurs familles. Dans le cadre de notre problématique, nous allons replacer ces familles de solutions par rapport à leur cycle d'utilisation.

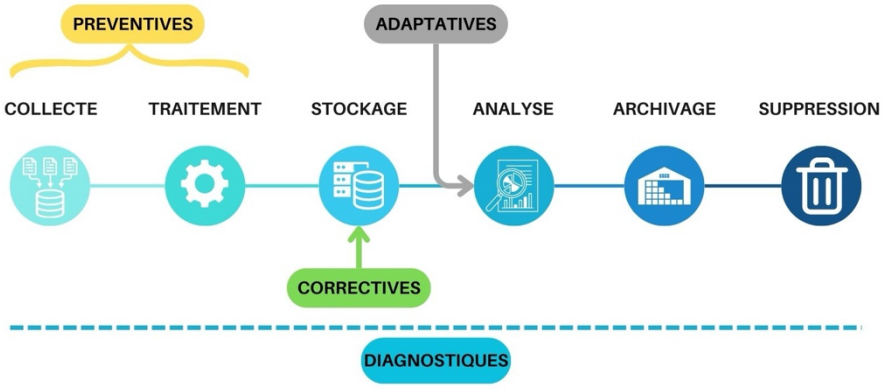

Illustration 6 : Typologies de solution dans le cycle de vie d'une donnée.

Ensuite plusieurs stratégies de gestion de la qualité des données ont été établies, notamment la méthode TDQM, qui est l'un des piliers du domaine de la qualité des données. Cette méthode a été créée par Richard Yang, un auteur influent dans ce domaine. Il a également contribué à l'élaboration de la méthode AIMQ.

Les trois méthodes étudiées effectuent des mesures et des analyses sur les données pour en définir les caractéristiques principales. Cependant, l'utilisation des dimensions et des mesures varie quelque peu. Par exemple TDQM et AIMQ regroupent les dimensions en quatre groupes pour simplifier les analyses.

ORME-DQ et TDQM incluent toutes deux une cartographie, ce qui permet d'identifier indirectement le cycle de vie des données, bien que cela reste un élément d'arrière-plan. En revanche, dans AIMQ il n'y a pas vraiment de phase d'analyse des flux. Les étapes du cycle de vie des données sont gérées de manière implicite dans les méthodologies, qui comprennent également des éléments de pilotage pour garantir une qualité élevée dans le temps. Ces méthodes intègrent donc non seulement des aspects diagnostiques, mais aussi préventifs.

La méthode AIMQ, quant à elle, est constituée uniquement de solutions de type diagnostique (analyse, audit, questionnaire). Elle est donc plus simple et nécessite moins d'expertise que les autres, ce qui la rend plus accessible. Cependant, elle n'est pas aussi complète que les deux autres méthodes. Son avantage réside dans le fait qu'elle compare les résultats par rapport à ceux des concurrents, ce qui est pertinent car chaque domaine a ses spécificités.

ORME-DQ est la seule des trois méthodologies à placer les coûts économiques au centre du raisonnement. Cela peut être un avantage ou un inconvénient, notamment parce qu'associer des failles de qualité des données à des pertes potentielles est relativement complexe. Cependant, cela permet de garantir une priorisation optimale de la gestion de la qualité et de rassurer la direction en quantifiant les gains et les pertes probables. Cette méthode se distingue également par son principe de sondes, qui permet un pilotage de la qualité des données en temps réel.

Un point faible de la méthode ORME-DQ, par rapport aux deux autres, est qu'elle n'utilise pas de questionnaire pour mettre en lumière les différentes perceptions des utilisateurs. Ces questionnaires, bien qu'étant subjectifs, permettent de révéler des incohérences techniques et fonctionnelles, et donc d'améliorer la mise en place des solutions.

Pour clôturer cette analyse, nous allons illustrer les arguments précédents par trois cas différents, en identifiant les solutions optimales pour chacun d'eux :

Cas 1 : PME industrielle avec un chiffre d'affaires d'environ 30 millions d'euros, souhaitant améliorer la qualité de ses données de production.

Dans cette situation, il faut prendre en compte que l'expertise informatique représente un budget important pour l'entreprise. Il serait préférable de contacter un petit cabinet de conseil ou de faire appel à des freelances, avec un coût journalier de 700[7] euros par consultant. L'entreprise pourrait engager trois consultants pour une durée de six mois (environ 400 000 €). Ces consultants ne chercheraient pas à mettre en place une méthodologie complexe, mais pourraient s'appuyer sur les principes de l'analyse AIMQ. La première partie de leur travail consisterait à identifier les dimensions de qualité pertinentes, à les mesurer, puis à les analyser. Ensuite, selon ce que nous avons établi plus haut, il serait nécessaire de mettre en place des solutions correctives, puis préventives, pour éviter la répétition des erreurs identifiées lors de l'analyse. Si le temps le permet, un certain pilotage de la collecte des données pourrait également être mis en place.

Cas 2 : Importante infrastructure de santé gouvernementale, avec 70 millions d'utilisateurs.

L'objectif principal ici n'est pas un gain financier direct, mais une gestion optimale des services de santé. Un bon exemple serait l'approvisionnement de certains médicaments qui ne doivent pas être en rupture de stock.

Dans ce cas, l'infrastructure et les volumes de données sont beaucoup plus importants que dans le cas 1. Il serait donc pertinent d'avoir un service interne dédié à la qualité des données, avec des experts capables de mettre en place des méthodologies plus complexes.

[7] Sur la base des profils freelance présent sur le site malt.

Parmi les méthodes étudiées, le choix se porterait entre TDQM et ORME-DQ. Cette dernière, étant centrée sur l'association de la qualité aux risques financiers, pourrait être plus complexe à mettre en œuvre. L'utilisation de TDQM ou d'une méthode similaire serait donc optimale ici. La présence d'une division composée d'experts sur le sujet permettrait non seulement de mettre en place cette méthodologie, mais aussi de suivre la phase de pilotage pour garantir une qualité des données dans le temps.

Cas 3 : Groupe d'assurance mondial avec un chiffre d'affaires annuel de 150 milliards d'euros.

Ce cas illustre l'importance de la qualité des données pour les assurances qui traitent un grand nombre d'informations dans divers domaines pour établir les risques et proposer des grilles tarifaires. La qualité des données est donc un prérequis pour l'activité de l'entreprise.

L'activité de l'entreprise et sa taille justifient un investissement massif dans la gestion des données. Pour cela, l'entreprise pourrait faire appel à des cabinets d'experts, mais elle devrait également posséder une division interne dédiée à la qualité des données.

Dans ce contexte, la mise en place d'une méthodologie TDQM pourrait être une solution. Cependant, ORME-DQ, orientée sur le gain économique et proche de TDQM, pourrait être une meilleure option. L'application des différents modules et des sondes garantirait la qualité des données.

L'entreprise pourrait très largement envisager l'achat d'outil de nettoyage ou d'audit.

5.2. Les limites et implications en qualité des données.

La qualité des données des données est un sujet complexe, et malgré l'importance économique de ce phénomène aucune solution miracle n'a été trouvé. Cela s'explique notamment par les différentes limites au sujet que nous allons définir ici.

Un monde du travail qui ne facilite pas la QD:

La rotation des postes dans les entreprises ne fait que s'accélérer, la durée moyenne des postes au sein de mon environnement est d'environ 4 ans. Les services informatiques étant en général externalisés, ils sont aussi victimes de changements de gestionnaire fréquents, ce qui complique la gestion des outils. Ces causes amènent des conséquences :

- Manque de connaissances : La gestion des données est comme nous avons pu le voir auparavant est un élément complexe, sur lesquelles un temps d'apprentissage propre à chaque environnement est nécessaire. L'augmentation du turn-over entraine donc une dégradation du savoir, qui réduit indirectement la qualité de la donnée.
- Changement de vision : Lors des rotations de gestionnaires ou de consommateurs, les objectifs et les priorités sont amenés à changer, mais les outils mises en place ne peuvent pas vraiment suivre ces changements.

- Décalage entre métier et techniques : Ces 2 mondes sont très différents, cela s'explique notamment par des profils et des formations assez différentes. De plus les objectifs et les visions des 2 mondes étant mobiles, un accroissement de la frontière technique et métier se créé. Ce point justifie la mise en place de cursus pluridisciplinaires comme la MIAGE.

Une des principales solutions à toutes ces conséquences est la documentation, qui définit chaque élément par du texte pour les rendre intelligibles par tous. Or, ce sujet n'a pas été discuté dans la littérature sur laquelle je me suis appuyé. La documentation est un sujet vaste et complexe qui pourrait représenter un sujet de mémoire. Les principales limites sont notamment l'uniformité de format, la redondance d'informations due à la multiplicité des parties prenantes, ou encore une sous-utilisation.

La qualité logiquement dépendante des moyens économiques mis en place :

La mise en place de solutions pour améliorer la Qualité des Données (QD) représente un investissement significatif pour une entreprise. Cela s'explique par la nécessité de faire appel à des experts, car la gestion de la qualité des données est complexe et requiert une évaluation minutieuse des options disponibles. En plus des coûts liés aux ressources humaines, l'achat de logiciels spécialisés ou le développement de solutions internes représentent également des dépenses importantes.

Il est toutefois difficile d'estimer avec précision les gains économiques que pourrait apporter une meilleure qualité des données. Bien que cela améliore généralement les prises de décision, l'impact financier direct reste difficile à quantifier. Par exemple, il serait nécessaire de se demander :

- Quelle est l'ampleur de la variation des résultats due à la qualité des données ?
- Quel est l'impact de cette variation sur les décisions prises par l'entreprise ?
- Quel est le gain ou la perte associé à ces décisions ?

Ces questions illustrent la complexité d'évaluer les bénéfices économiques de l'amélioration de la qualité des données. Néanmoins, ces investissements, bien que coûteux, sont essentiels pour garantir la compétitivité et l'efficacité à long terme de l'entreprise.

Des volumes tout simplement trop importants ?

Les volumes de données trop importants peuvent poser des défis majeurs. Dans un environnement simple, les anomalies sont faciles à repérer. Cependant, à mesure que le volume de données augmente, la diversité et le nombre d'erreurs croît, tandis que la capacité de contrôle diminue. Pour pallier ce problème, des solutions complexes sont mises en place tout au long du cycle de vie des données, augmentant ainsi la complexité des flux informatiques et les coûts associés.

Cette complexité accrue se manifeste à chaque étape du cycle de vie des données. Par exemple, si des anomalies sont détectées dès la collecte, il devient nécessaire d'ajouter des mesures préventives et des traitements correctifs. Ces interventions, bien que nécessaires, alourdissent les processus et peuvent introduire de nouvelles contraintes, telles que la nécessité de corriger et de documenter les erreurs résiduelles, augmentant encore la charge sur le système. Ce phénomène de complexité croissante peut aussi ralentir les opérations, augmentant les risques de retards et d'inefficacités.

5.3. Perspectives d'amélioration et innovation

Dans cette section, nous argumenterons sur différents points d'évolutions sur ce qui existe en qualité des données en s'appuyant sur les recherches effectuées, ainsi que de mon expérience professionnelle.

Corriger le plus tôt possible :

Dans notre problématique, nous avons mis en avant les étapes du cycle de vie d'une donnée par rapport à la qualité, ce qui n'a pas été fait dans la littérature ou implicitement. Or, visualiser l'impact de la qualité à chaque étape permet de comprendre comment les actions en amont influencent les résultats en aval.

Il est particulièrement crucial de reconnaître que l'amélioration de la qualité des données dès les premières étapes du cycle, telles que la collecte et le traitement, peut avoir des répercussions positives significatives sur les usages ultérieurs. Les dimensions intrinsèques, contextuelles et représentatives de la qualité des données doivent donc être rigoureusement contrôlées dès le début du cycle de vie. En ciblant les efforts de qualité sur la phase de collecte, on réduit non seulement la propagation des erreurs dans les étapes suivantes, mais on simplifie également la gestion de la qualité en concentrant la complexité à un point stratégique. Cette approche proactive de la gestion des données peut réduire les coûts de correction en aval, limiter les risques d'erreurs coûteuses et améliorer globalement l'efficacité des processus.

Dans les méthodes on veut aligner oui mais également Simplifier le plus possible les processus :

Aujourd'hui, la complexité croissante des exigences des clients pousse les entreprises à développer des produits et services de plus en plus sophistiqués, ce qui, à son tour, complexifie les processus métier et les systèmes informatiques associés. Cependant, cette complexité doit être soigneusement contrôlée pour éviter qu'elle ne devienne un frein à l'efficacité et à la qualité des données.

Pour y parvenir, les entreprises doivent chercher à simplifier autant que possible leurs processus tout en tenant compte des besoins de leur environnement. Il est essentiel de maintenir des systèmes informatiques aussi simples que possible, ce qui facilite leur compréhension par les utilisateurs finaux. Une meilleure compréhension des systèmes par les utilisateurs leur permet

de mieux gérer les aspects complexes de leur métier, réduisant ainsi le besoin de corrections coûteuses en aval.

La simplicité des processus contribue également à une plus grande agilité de l'entreprise, lui permettant de s'adapter plus rapidement aux changements du marché tout en maintenant un haut niveau de qualité des données. Cette approche simplifiée favorise également la collaboration entre les équipes métier et informatique, réduisant les silos organisationnels et améliorant la prise de décision.

Mise en place d'indicateurs :

La qualité des données, bien que primordiale, ne sera jamais parfaite. Il est donc pertinent d'introduire des indicateurs permettant d'estimer un taux de confiance des données, qui pourrait être intégré dans les rapports de gestion. Ces indicateurs offriraient plusieurs avantages. D'une part, ils amélioreraient la prise de décision en fournissant aux gestionnaires une compréhension plus nuancée de la fiabilité des informations sur lesquelles ils se basent. D'autre part, ils sensibiliseraient la direction aux enjeux de la qualité des données, en mettant en lumière les domaines nécessitant des améliorations.

Cependant, l'introduction de tels indicateurs doit être faite avec précaution. La direction pourrait hésiter à accepter l'idée que certaines données ne sont pas parfaitement fiables, ce qui pourrait conduire à une résistance initiale. Néanmoins, ces indicateurs doivent être présentés comme des outils stratégiques, permettant de mieux comprendre les risques et d'améliorer les processus de manière continue. Ils jouent un rôle clé dans la gestion des données en offrant une transparence accrue et en facilitant une culture de l'amélioration continue au sein de l'entreprise.

6. La Qualité des Données en entreprise : cas pratiques et solutions.

Durant la réalisation de ce mémoire, j'ai pu retrouver des familiarités avec mon environnement professionnel. Il n'y a pas d'application de méthodologies précise, cependant des solutions primitives sont appliquées à des endroits stratégiques des processus. Nous allons donc passer en revue ces dernières. Pour cela nous reprendrons le sujet du nouvel outil et de power BI mentionné lors de la justification du choix de cette problématique.

6.1. Power BI, une automatisation au profit de la qualité de l'information.

NB : Power BI est un outil de la suite Microsoft qui permet d'analyser les données et de créer des visualisations. Les rapports s'actualisent automatiquement avec de nouvelles données, et les analyses ainsi que les visualisations sont mises à jour lorsque cela est souhaité.

Des outils comme Power BI contribuent indirectement à l'amélioration de la qualité des données, car ils automatisent et normalisent les partages de données entre les différentes parties de l'entreprise. Les dimensions de la qualité qui sont améliorées ici sont principalement la représentativité et l'accessibilité, notamment en termes de facilité d'opération et de compréhension. Aussi, de maniéré générale Power BI fait croitre la qualité des analyses et facilite les décisions issues de ces analyses.

Durant ma première année d'alternance, j'ai eu l'opportunité de migrer certains rapports Excel vers ce nouvel outil. L'objectif de ces livrables était de présenter les variations du portefeuille par catégorie de clients, tout en vérifiant les prévisions. La finalité est d'évaluer chaque mois les hypothèses, c'est-à-dire comparer les prévisions et les réalisations, afin de prendre ou non, avec la direction marketing, les décisions d'ajustements.

Ces rapports sont alimentés par des chiffres de prévision et de réalisation issus de notre département, mais aussi par des données marketing, ce qui permet de valider le bon alignement des flux métiers.

Auparavant, des rapports Excel étaient générés par un outil d'analyse, mais nécessitaient des contrôles ainsi que des copies de données, ce qui représentait un temps de travail considérable et un risque accru d'erreurs. Avec la nouvelle version, notre équipe dispose d'un meilleur contrôle sur les données en entrée grâce à des requêtes SQL.

De plus, les opérations de copier-coller ont été éliminées, les données étant désormais récupérées automatiquement depuis un SharePoint appartenant aux commerciaux, plutôt que par courriel. Les différents calculs sont automatisés, mais restent consultables et modifiables facilement. Tout au long du rapport, les données présentées sont commentées, ce qui facilite la compréhension des utilisateurs.

En cohérence avec ce qui a été décrit comme une limite, nous avons mis en place une documentation autour de ces rapports, pour expliquer la provenance des différentes données, leurs intérêts, et leurs caractéristiques. Cela permet de garantir une bonne compréhension part tous les acteurs durant la présentation, en voici une petite illustration :

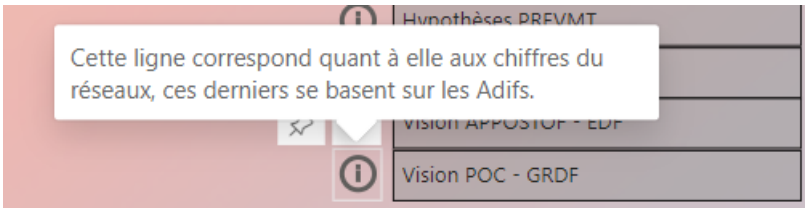

Illustration 7 : Exemple de documentation Power BI mise en place

Dans ce projet, l'objectif initial était de simplifier les tâches, mais l'une des finalités était aussi d'améliorer la qualité des données. Cela montre à quel point la qualité des données dépend de nombreux éléments.

On peut également préciser que Power BI améliore d'autres composantes des données, comme la sécurité et la visualisation des données.

6.2. Mise en place d'un SAS de contrôle à l'entrée des données.

Comme expliqué succinctement dans le contexte, un nouvel outil est actuellement mis en place au sein de mon équipe. Cet outil a pour objectif de traiter différents flux de données reçus par les réseaux de transport et distribution de gaz et du CRM[8], à la suite de la saisie des commerciaux.

Cependant, des anomalies peuvent apparaître dans ces données reçues. Auparavant, ce type d'anomalie pouvait être intégré directement. Avec le nouvel outil, un SAS a été mis en place : lorsque des flux incohérents arrivent, ils sont placés dans une "salle d'attente" où un membre de mon équipe vérifie les informations en se basant sur la vérité terrain.

Évidemment, ce SAS n'est utilisé que pour les données ayant un poids économique important ; il serait compliqué de vérifier les informations de chaque client particulier, étant donné le volume élevé de ces données.

Nous sommes donc ici dans une approche de solution préventive, qui permet d'améliorer les dimensions intrinsèques et représentatives de la qualité des données (QD). C'est une excellente illustration de l'importance de contrôler les informations en début de cycle, comme mentionné

[8] CRM = Gestion de la relation client (Customer Relationship Management)

dans les possibilités d'amélioration. Le temps perdu à vérifier ces données représente un gain de temps sur l'analyse et sur la fiabilité des données.

6.3. Des solutions au quotidien : pratiques et outils pour maintenir la qualité.

Au sein de mon département, des solutions de contrôle de l'information sont mises en place quotidiennement. Nous allons en décrire quelques-unes.

Tout d'abord, une chaîne opérationnelle rétroactive composée de plusieurs maillons qui se contrôlent mutuellement a été mise en place. Le rapport principal est alimenté par toutes les équipes, chacune vérifiant et recroisant les données par rapport à son environnement, afin de s'assurer que tout est conforme. Cela permet de garantir la cohérence des données finales. Cette approche constitue une solution de diagnostic et de correction.

En outre, des vérifications basées sur la réalité du terrain sont régulièrement effectuées pour les clients les plus importants. Par exemple, si un client industriel majeur quitte le portefeuille sans raison apparente, des demandes d'explications peuvent être adressées aux commerciaux. De même, des investigations sont menées si un client sous-consomme ou surconsomme par rapport aux prévisions.

Des points de suivi des variations de portefeuille sont organisés de manière hebdomadaire. Au cours de ces réunions, la cohérence des informations et la bonne évolution des données sont vérifiées, tout comme le bon déroulement des opérations liées au système d'information (SI). Par ailleurs, des réunions régulières entre les équipes SI et les équipes métiers sont tenues pour aligner les flux métier et techniques, un élément essentiel qui est défendu dans les méthodes étudiées.

Des solutions techniques ont également été implémentées dans les systèmes informatiques, telles que l'intégrité référentielle et la consolidation des données entre les réseaux et les équipes commerciales. Toutefois, un manque de documentation avait été constaté sur ces aspects, ce qui a pu entraîner des problèmes de compréhension des données en sortie. Des améliorations ont été demandées à ce sujet, et des requêtes ont été formulées pour que la documentation soit enrichie sur les nouveaux outils. Comme mentionné précédemment, cela permet d'améliorer les dimensions d'accessibilité et de représentativité des données.

D'autres solutions techniques sont développées au sein de l'équipe, qui contribuent également à l'amélioration indirecte de la qualité des données. Il s'agit en grande partie de scripts Python qui automatisent certains processus métier. Par exemple, ces scripts prennent en charge des inputs partagés par d'autres entités, qui peuvent parfois être corrompus. Ils traitent ces données, détectent d'éventuelles anomalies et produisent les outputs souhaités, assurant ainsi une meilleure fiabilité et une cohérence accrue des données.

Comme nous venons de le voir, un grand nombre de solutions de qualité des données (QD) ont été mises en place tout au long des différents processus métier de l'entreprise pour garantir cette qualité. Bien que ces solutions soient placées de manière non structurée, elles permettent, ensemble, de limiter les risques de manière significative.

7. Conclusion

Ce mémoire a permis d'explorer en profondeur les enjeux liés à la qualité des données, notamment sous les angles économiques et décisionnels. Nous avons examiné les caractéristiques des données ainsi que les rôles des différents acteurs impliqués. Nous avons également décrit les étapes du cycle de vie des données, en soulignant leur importance relative en matière de qualité.

Notre analyse s'est ensuite concentrée sur les dimensions de la qualité des données, que nous avons étudiées en détail. Les principales solutions existantes ont été passées en revue, suivies par l'examen de trois des méthodologies les plus couramment utilisées pour assurer la qualité des données. Nous avons enrichi notre étude par une analyse croisée des éléments issus de la littérature, mettant en lumière les limites actuelles et les possibilités d'amélioration. Enfin, nous avons appliqué ces concepts à des cas pratiques, en examinant les initiatives de qualité des données au sein de mon département.

Que conclure du rapport ?

Les données jouent un rôle central dans la prise de décision et le pilotage économique des entreprises, et leur qualité a donc un impact significatif. Cependant, cette dernière est influencée par plusieurs facteurs, notamment au cours des premières phases du cycle de vie des données. Des solutions préventives, correctives, adaptatives, et de diagnostiques ont été développées pour répondre à ces défis. De nombreuses solutions existent, mais leur complexité varie, ce qui permet une adaptation aux besoins spécifiques de chaque environnement. Des logiciels spécialisés ont été conçus pour faciliter l'implémentation de ces solutions, bien qu'ils soient souvent coûteux.

Des méthodologies plus complexes ont également été développées, nécessitant une expertise avancée, et sont généralement réservées aux entreprises gérant de grands volumes de données et disposant de ressources économiques importantes. Ces méthodologies ne sont pas universelles ; chaque entreprise doit composer un ensemble unique de solutions pour répondre à ses besoins en matière de qualité des données.

En théorie, la qualité des données devrait être parfaite, mais en pratique, cela reste inatteignable en raison de la complexité des environnements et de la difficulté à quantifier les impacts financiers d'une mauvaise qualité des données. La croissance exponentielle des volumes de données complexifie encore davantage cette problématique.

Nous avons identifié trois axes principaux d'amélioration : intervenir le plus tôt possible dans la vie des données pour gérer les problèmes de qualité, prendre en compte l'impact de la complexité métier sur la qualité des données, et mettre en place des indicateurs de qualité des données pour les livrables.

Puisque nous avons mis en avant qu'il était complexe de définir les gains économiques potentiels d'une amélioration de la qualité des données, il serait à présent intéressant de se demander comment estimer le coût d'une mauvaise qualité des données pour une entreprise ?

Bilan personnel/professionnel/Apport du mémoire :

Ce mémoire a été pour moi une expérience enrichissante, tant sur le plan personnel que professionnel. Malgré les défis liés à la vie d'alternant, je suis satisfait d'avoir mené à bien ce projet. Cela m'a permis de renforcer mes connaissances en vue de mon Master 2 en Informatique Décisionnelle et d'envisager des perspectives professionnelles, notamment dans le domaine de la qualité des données.

Le sujet s'est avéré plus complexe que je ne l'avais imaginé, révélant qu'il n'existe pas de solution unique et optimale. Chaque situation demande un arbitrage spécifique, ce qui met en évidence la complexité inhérente à la gestion de la qualité des données. Une des conclusions marquantes est que la qualité parfaite des données est une utopie difficilement atteignable, en raison des nombreuses variables à prendre en compte. Cependant, l'exploration de la relation entre la qualité des données et le cycle de vie de celles-ci, a apporté une dimension nouvelle à mon sujet en offrant une perspective différente sur cette problématique.

Sur le plan professionnel, je tiens à exprimer ma gratitude envers les membres de mon équipe. J'ai eu l'opportunité de m'impliquer pleinement dans la vie de l'équipe, contribuant activement à ses projets tout en croisant les résultats de mes recherches théoriques à nos activités opérationnelles. Ce travail m'a permis de faire le lien entre la théorie et la pratique, bien que je regrette de ne pas avoir pu approfondir davantage ce recroisement, notamment en ce qui concerne l'application des méthodologies de qualité des données.

Enfin, ce mémoire a permis de mettre en lumière l'importance de considérer le cycle de vie des données en parallèle avec leur qualité, une approche qui semble peu explorée dans la littérature existante. En confrontant les éléments théoriques à mon expérience pratique au sein d'EDF, j'ai pu formuler certaines idées d'améliorations et dégager des pistes pour de futures recherches, soulignant ainsi l'apport de mon travail à ce domaine en constante évolution.

8. Références générales.

[1] *Laure Berti-Équille* –**La qualité des données–** 2006

[2] *Kumar Rahul, Rohitash Kumar Banyal* – **Data Life Cycle Management in Big Data Analytics –** 2020 - p365 à p371

[3] *Nicolas Boisnic & Zeenea* – **Le guide du data quality management –** 2022

[4] *Lisa Ehrlinger and Wolfram Wöß* – **A Survey of Data Quality Measurement and Monitoring Tools –** 2022 (https://www.frontiersin.org/articles/10.3389/fdata.2022.850611/full)

[5] *Yang W. Lee, Diane M. Strong, Beverly K. Kahn, Richard Y. Wang* - **AIMQ: a methodology for information quality assessment** – p133 à p146 – 2001

[6] *Richard Y. Wang* - **A Product Perspective on Total Data Quality Management** – p58 à p65 – 1998

[7] *Batini Carlo, Barone Daniele, Mastrella Michelle, Maurino Andrea, Ruffini Claudio* – **A frameworrk and a methodology for data quality assessments and monitoring –** 2005

[8] Autres références :

Données et son cycle de vie :

- https://www.talend.com/fr/resources/cycle-vie-donnees/
- https://www.inist.fr/wp-content/uploads/donnees/co/module_Donnees_recherche_7.html
- https://www.ibm.com/fr-fr/topics/data-lifecycle-management
- https://www.claranet.com/fr/expertises/data-modernisation/big-data/data-et-big-data-comprendre-la-chaine-de-valeur

[9] Des outils basés sur de l'intelligence artificielle générative tels que ChatGPT ont été utilisés pour la rédaction de se mémoire. Ils ont permis d'améliorer la rédaction et de corriger d'éventuelles fautes d'orthographes.

[10] Des éléments de mon rapport de stage effectué l'année dernière ont été récupérés notamment sur la présentation de l'entreprise.