Mémoire M1 MIAGE APP

-

ANDRIN Mathieu

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



Table des matières

[1. Introduction (3 pages) 3](#_Toc174197331)

[1.1. Mise en contexte rapide 3](#_Toc174197332)

[1.2. Définitions de la problématique 3](#_Toc174197333)

[2. Contexte en entreprise (2-3 pages) 4](#_Toc174197334)

[2.1. Présentation d’Électricité De France (EDF) 4](#_Toc174197335)

[2.2. Pôle GAZ et Portefeuille Client Contrat. 5](#_Toc174197336)

[2.3. Contexte de la problématique 5](#_Toc174197337)

[3. La donnée et son cycle de vie 6](#_Toc174197338)

[3.1. La donnée 6](#_Toc174197339)

[3.2. Qu’est-ce que le cycle de vie d’une donnée. 7](#_Toc174197340)

[4. État de l’art 9](#_Toc174197341)

[4.1. Qualité des données et principaux problèmes associés 9](#_Toc174197342)

[4.2. Dimensions et mesures de la qualité des données. 10](#_Toc174197343)

[4.3. Les solutions existantes. 12](#_Toc174197344)

[4.4. Les Méthodologies existantes. 14](#_Toc174197345)

[5. Analyse. 19](#_Toc174197346)

[5.1. Analyse de l’état de l’art. 19](#_Toc174197347)

[5.2. Limite et implication de l’état actuel. 21](#_Toc174197348)

[5.3. Évolutions/ améliorations possibles ? 22](#_Toc174197349)

[6. En interne 22](#_Toc174197350)

[7. Conclusion (2 pages) 23](#_Toc174197351)

[8. Références générales 23](#_Toc174197352)

[9. Bibliographie 23](#_Toc174197353)

# Introduction (3 pages)

## Mise en contexte rapide

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 pour l'économie américaine, soulignant ainsi l'ampleur de l'enjeu.

De plus en tant que data analyste, je suis régulièrement confronté à ces défis liés à la qualité des données, qui figurent parmi les enjeux majeurs de ce métier. Ce mémoire me permettra d'explorer des solutions scientifiques à ces problématiques, dans le but de continuer à évoluer dans le domaine de la data après mes études.

## Définitions de la problématique

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 entreprise, 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 fait sur le département dans lequel j'évolue, avec une recontextualisation de la problématique par rapport à mon expérience en alternance et les 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

# Contexte en entreprise (2-3 pages)

## Présentation d’Électricité De France (EDF)

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 production nucléaire, inaugurant sa première centrale à Chinon en 1963. L'entreprise a depuis 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 seul actionnaire, 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 son engagement à construire un avenir énergétique durable.

## 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 de Pôle Client Services et Territoires (CST) et regroupe 30 000 salariés répartis dans différentes directions et filiales.

La direction comprend un département gaz dont l’objectif est d’optimiser l’équilibre économique du sourcing de gaz de l’entreprise, en d’autres termes acheter le GAZ pour répondre à la demande client, 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** : Création des offres et des prix, gestion des marges pour risques,
* **Optimisation** : Responsable des ordres d’achats du gaz en bout de chaine.
* **Prévision** : équipe donnant les anticipations de fluctuations du portefeuille.
* **Portefeuille Client et Contrat** : L’équipe dans laquelle j’effectue mon alternance.

Portefeuille Client et Contrat ou PCC se situe en amont des processus du pôle, elle est en partie responsable des analyses de donné. Nous établissons un bilan sur les variations réelles[[1]](#footnote-2) du portefeuille, nouveaux clients, sorties de client. L’équipe joue également un rôle de comptabilité (Facture d’achat de gaz et de coût d’acheminement). Ainsi que la gestion de la souscription des grosses capacités sur le réseau.

L’équipe PCC, par son utilisation des données de consommations clientes a également une collaboration interfonctionnelle avec des métier marketings, commerciaux ou encore comptables.

Schéma récapitulatif du fonctionnement :

Il y a des enjeux financiers pour l’entreprise concernant le gaz, même si cela représente une petite proportion comparée à l’électricité.

## Contexte de la problématique

Comme explicité dans la partie précédente, une grande partie du travail opérationnel au sein de l’équipe PCC est d’analyser des données. Cela comprend d’expliquer les variations, et d’éventuels écarts de données, cependant dans certains cas des éléments de mauvaise qualité peuvent complexifier ces vérifications. Nous reviendrons sur la définition de qualité dans la suite de ce rapport.

Durant ma 1ére année d’alternance j’ai pu participer à la vie opérationnelle de l’équipe qui est concerné la qualité des données. De plus, j’ai pu être acteur dans 3 grands projets transverses sur cette première année d’alternance qui ont donnée du sens à cette problématique, en voici un aperçu :

* **Création de rapport Power BI**[[2]](#footnote-3) : Mon activité principale en tant qu’alternant a été de remplacer des rapports Excel généré à l’aide d’outil par des Rapports BI. Lors de cette transition j’ai été amené à reprendre toutes les sources de données utilisées, dont j’avais une légère connaissance. Durant le développement je me suis donc posé des questions sur ces dernières, notamment sur leur qualité.
* **Apparition d’un nouvel outil**: Un autre élément clef de l’apparition de cette problématique a été l’arrivé d’un nouvel outil venant en remplacer un ancien. Son objectif est de récupérer des data flows[[3]](#footnote-4) reçus des commerciaux et des gestionnaires de réseaux et les agrégés pour faciliter les étapes d’analyses. Ce nouveau dispositif a été développé par une équipe SI, et lors des points de suivi auxquels j’ai pu participer, les principaux sujet concernés la gestion des données.
* **Développement d’un outil d’analyse** : Durant mes premiers mois dans l’entreprise j’avais effectué des analyses et du développement pour faciliter l’explication de certains écarts sur un rapport. Les principales causes d’écart avaient pu être identifiées, notamment des différences temporelles entre des processus, mais également dans certains cas la présence d’informations incohérentes.

Cette caractéristique de qualité est omniprésente en tant que data analyst que ce soit sur des projets comme décrit plus haut ou encore dans des études ponctuelles. Cela m’a conforté dans le fait de souhaiter mieux comprendre ce phénomène.

# La donnée et son cycle de vie

## La 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 manipulable 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.

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.

Plusieurs rôles sont définis sur l’utilisation des données :

* Fournisseur : Responsable de la création/récupération des données
* Fabricant : Conçoit, développe et maintient les données et l’infrastructures associée.
* Consommateur :
* Gestionnaire : Régisseur de la donnée, tout au long de leur cycle de vie.

## Qu’est-ce que le cycle de vie d’une donnée.

Comme vu précédemment la gestion des données (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.

Une image contenant texte, capture d’écran, Police, conception

Description générée automatiquement

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)

1. **Traitement et Nettoyage**

Cette phase est placée en 2éme position mais elle peut également réapparaitre à d’autres endroits 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) permettent notamment de faciliter cette étape.

1. **Stockage**

Le stockage des données, bien que perçu comme une phase moins sensible que la collecte, il 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 et de data lakes
* Conformité aux normes comme le RGPD (Règlement Général sur la Protection des Données)

De plus cette étape doit également appliquer une gestion de la modification bien définie. Les changements doivent être rigoureusement contrôlés et tracés pour maintenir l'intégrité des données.

1. **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.

1. **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 tels que les datawarehouses. Ces données archivées peuvent être réutilisées pour des analyses futures ou en cas de litige.

Lors de cette étape les données doivent être conforme aux politiques de rétention des données

1. **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

# État de l’art

<http://rdoc.univ-sba.dz/bitstream/123456789/3226/1/D3C_Inf_BENKHALED_Hamid_Naceur.pdf>

<https://www.researchgate.net/profile/Laure-Berti-Equille/publication/220438866_Qualite_des_donnees/links/54f45cdb0cf24eb8794debac/Qualite-des-donnees.pdf?__cf_chl_tk=cqRIs9YghJ17m3pTIwkRnzCsfITN3vSpWJoC_V0shFU-1714932767-0.0.1.1-1919>

Ebook sur les dimensions <https://info.zeenea.com/hubfs/Le-guide-du-data-quality-management.pdf>

## Qualité des données et principaux problèmes associés

Comme nous l’avons mentionnée dans le début de ce rapport, l’importance de la donnée ne fait que croitre au sein des organisation à travers le monde. Les disciplines transverses à ce domaine suivent la même tendance, c’est le cas de la qualité des données. Dans cette partie nous allons voir quels sont les principes de ce domaine, mais également les problèmes récurrents en termes de qualité.

En premier lieu, il faut comprendre que la QD[[4]](#footnote-5) garantit une prise de décision optimale, car elle améliore l’efficacité opérationnel et la confiance des consommateurs d’informations. Sur des valeurs économiques importantes des taux d’erreur assez faible, peuvent couter beaucoup.

Maintenant que nous avons décrit les principaux enjeux, il faut donner un sens à la qualité qui est un concept subjectif. Dans notre cas des dimensions de la QD ont été établit, comme la précision ou la complexité. Elles donnent un cadre et des ordres de grandeur sur les différentes caractéristiques de la donnée. Nous les étudierons en détail dans la partie suivante.

La norme ISO-8000 standardise la qualité des données à l’échelle internationale. Cette dernière à notamment pour objectif de garantir la conformité au réglementation et améliorer la collaboration entre les divers partenaires commerciaux.

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

* Consommation annuel total de gaz en France : 500 TWh
* Prix du TWh : 1 000 000 \*40[[5]](#footnote-6) = 40 000 000
* Coût totale : 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 peut revenir à une perte de 400 millions d’euros (2% de 20 milliards) et une variation de 0,1% représente quant à elle 20 millions. Évidement ici aucune externalité n’est prise en compte, mais cela donne une vision de l’ampleur que cette justesse peut avoir.

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 :

Les 7 problématiques les plus importantes en matière de qualité des données sont maintenant sélectionnées et affichées. Voici le résumé :

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 allons définir les principales dimensions par rapport aux problématiques vu ci-dessus. Puis nous regarderons les types de solutions qui ont été mis en place. Nous conclurons sur un survol des principales méthodologies de DQ.

## Dimensions et mesures de la qualité des données.

Dans cette partie, nous étudierons les principales mesures de QD. Ces éléments seront très importants pour la suite de l’état de l’art, les mesures sont à la base des méthodologies que nous étudierons par la suite.

En qualité des données les mesures sont regroupées dans des dimensions, cela permet de définir le concept qualité en le divisant en plusieurs groupes comme la précision ou encore la complétude.

On peut classifier la qualité des données dans 4 grandes classes.

Un grand nombre de chercheur ont défini des dimensions, certaines sont très proche les unes des autres. Voici donc une liste non-exhaustive des principales dimensions :

* **Précision** (accuracy)

La précision met en avant la différence entre l’information contenu en base et la réalité. En d’autres termes selon revient à considérer les valeurs qui sont incorrect. Éventuellement un niveau d’imprécision d’une valeur peut être utilisé.

* **Complétude** (completness)

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

* **Cohérence** (Consistency)

L’objectif cette fois est d’avoir une idée de la proportion de valeurs cohérentes par rapport aux standards et aux règles métier définis dans le système.

* **Conformité** (Validity)

La conformité permet de vérifier le respect des formats (mail/téléphone/…) dans les données. Elle est proche de la dimension de précision, car une donnée qui ne respectent pas le format à de grande chance d’être fausse. Le nombre de valeur inconforme correspond à la somme des valeurs pour chaque

* **Age** (Timeliness)

On cherche ici à établir une mesure du rafraichissement des données, l’importance de cette dimension varie selon l’environnement, certaines données peuvent nécessiter une actualisation journalière, mensuel ou encore annuel. De plus il est nécessaire d’avoir une date de dernière mise à jour pour chaque donnée en base.

* **Unicité** (Uniqueness)

La dimension d’unicité permet de quantifier l’importance données qui se répètent dans la base. Encore une fois il est nécessaire de définir à quoi correspond un doublon dans le jeu de donnée (exemple : Un client sur 2 contrats en même temps).

* **Disponibilité** (Accessibility)

Ici on se place du côté consommateur de données et on cherche à évaluer la simplicité d’accès aux informations. Le contrôle peut se faire à travers des questionnaires réfléchis soumis au consommateur, une subjectivité peut donc rentrer en compte sur ces éléments.

* **Facilité d’opération** (Ease of operation).

La facilité d’opération défini la facilité d’utilisation des systèmes de gestion des données.

Pour la quantifier, il est possible de se baser sur des questionnaires, mais également sur des métriques tels que la complexité du Modèle de données. En effet plus cette dernière est grande plus les requête se complexifie (jointure, filtre, …).

* **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 et 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.

Il est possible de reclassifier ces différentes dimensions dans de nouvelle catégorie :

|  |  |  |  |
| --- | --- | --- | --- |
| 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 chaque division 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.

## Les solutions existantes.

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, assez proche des domaines d’architecture des systèmes informatiques, évaluer la qualité d’une BDD (relation, processus, traitement,)
3. Adaptatives : Nettoyage et consolidation / utilisation ETL
4. Correctives : Temps réel -> vérification dans la requête (ex : on ne prend pas les valeurs nulls)

Nous allons faire un tour d’horizon des principales solutions en qualité des données

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

Le Modèle Conceptuel de Donnée (MCD) est un élément très influent sur la qualité des données, c’est lui qui définit les différentes relations et entité présent dans une base. Lors de sa conception différentes mesures sont prises comptes, ces dernières seront réutilisées par la suite dans les études de qualité, comme la complexité structurelle du modèle, ou encore le nombre d’association et d’entité. Sa structure permet ou non de simplifier ou non la mise en place de la qualité des données.

* **Consolidation –** Adaptative :

La consolation des données est un processus de choix dans le cycle de vie d’une donnée lors de la collecte : Nous disposons d’une même information provenant de 2 sources différentes, la consolidation est le processus de choix entre les 2 visions, si j’estime que ma source A et plus fiable que B, on peut donner la règle suivante : Si A 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 donner plus du contexte à une donnée.

C’est une solution très puissante mais très lourde, cette solution est régulièrement utilisée dans le cadre de mon travail, en tant que data-analyst si une information importante est erronée et qu’elle n’a pas pu être traité automatiquement par les solutions mise en place en amont cela signifie qu’il est nécessaire de comprendre l’apparition de cette information (Uniquement pour les cas ayant de fort impact sur les analyses)

Un très bon exemple de vérifications d’après la vérité terrain et la confirmation des emails ou des numéros de téléphone par des codes lors de l’inscription à un site.

* **Filtre –** Adaptative :

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

NB : Le clustering est également une solution appartenant au registre du filtre.

* **Audit –** Diagnostique :

Un audit est équivalent à une analyse de l’état actuel. Il permet de mettre en avant les forces et les faiblesses de notre système de gestion des données.

* **Suivi des données** – Préventive :

Cette solution consiste à avoir un pilotage et un contrôle permanent des données, en effet il ne suffit pas seulement de vérifier de temps en temps les données.

* **Nettoyage** - Correctives :

Le nettoyage est une solution élémentaire, cela revient tout simplement à retirer des données les éléments indésirables (totalement erronées, les doublons, ...).

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

Les règles de gestion sont des éléments métiers qui permettent de garantir une 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 5000 $ par an) pour le nettoyage par ETL ou encore Avellino – Discovery (80 000 €) pour l’audit.

## Les Méthodologies existantes.

Nous venons de faire un tour d’horizon des différentes solutions que nous pourrions qualifiées de « primitive », maintenant nous allons 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 (en rouge les 3 que nous étudierons). Nous pouvons remarquer que chaque système exploit une phase d’analyse de donnée, et de mesure (sauf CIHI). Mais 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.

Une image contenant texte, capture d’écran, nombre, Police

Description générée automatiquement

Les méthodologies appliquant toutes les étapes devraient-être plus complète mais également plus complexe (plus cher) à appliquer.

Avant de commencer le tour d’horizon des 3 méthodes, il est essentiel de préciser que certains mécanismes de ces méthodes sont complexes et seront survolés dans les explications.

* **TDQM (Total Data Quality Management)**

*Richard Y. Wang*

Une image contenant texte, diagramme, capture d’écran, cercle

Description générée automatiquement

Cette méthodologie est très populaire en qualité des données et est inspirée du cycle de Deming (Plan-Do-Act-Check), en reprenant les 4 phases principales vues plus haut. Elle est difficilement applicable en complément ou en support, elle est complète. Un des principaux concepts de cette stratégie et de considérer la donnée de la même façon qu’un objet physique.

1. Définir donnée :

Cette partie revient a effectué un état des lieux sur la donnée et son environnement

Tout d’abord une *Évaluation de la qualité de l’information* est effectuée : Tout les rôles (Fournisseur, Fabricant, Utilisateur et Gestionnaire) donne une évaluation de leur point de vue sur chaque dimension de la qualité des données.

Également, une *Identification du système de production de la donnée* permet de mettre en avant les rôles de chaque acteur ainsi que les différents traitements suivis par la donnée. Cette étape est cruciale et permettra de simplifier les mesures et les analyses, à l’aide de la Matrice d'Analyse de la Fabrication de l'Information[[6]](#footnote-7).

1. Mesurer la donnée :

Durant cette phase il y a logiquement une utilisation des différentes mesures et dimensions adaptées sur les données suivies.

Il y a également une prise en compte du respect des différentes règles de gestion métier.

1. Analyser la donnée :

À partir des résultats de la phase de mesure, il faut déterminer les causes d’éventuel anomalie dans les données, dans l’objectif de trouver une solution de régulation.

La méthode suggère d’étudier les hypothèses et les justifications en profondeur.

De plus il est important de prendre en compte les bénéfices associés à une amélioration potentiel.

1. Amélioration et pilotage permanent :

En se basant sur l’analyse, il faut maintenant mettre en avant les principaux éléments à améliorer. En ayant pour objectif :

* + Aligner les flux de données et les flux de travail avec le bon système de fabrication de la donnée.
  + Créer de la cohérence entre les besoins commerciaux et les données.

La matrice d’analyse de la fabrication de l’information défini lors de la première phase permet d’atteindre ces objectifs.

* **AIMQ (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 phase que nous allons étudier, car elle ne possède pas de phase d’état des lieux sur la structure existante. Elle cherche également à faire un benchmark [[7]](#footnote-8)entre les différentes organisations.

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement3 composantes principales de la création de la méthode sont mises en avant :

1. PSP/IQ

Ici l’objectif est de regrouper les dimensions d’analyse de la DQ en 4 grandes parties (Intrinsèque, contexte, représentative, accessibilité). Cela permet d’évaluer la conformité aux attentes et spécifications des consommateurs ainsi que des gestionnaires.

1. IQA

L’élément IQA a la charge de faire ressortir les mesures les plus pertinentes par rapport au modèle PSP/IQ. Des études sur la fiabilité et l’interdépendances des dimensions ont permis de montrer que bien que la qualité des données soit multidimensionnelle, elle représente un élément unique.

De plus un écrémage des mesures lié à chaque dimension a été effectué pour réduire le nombre de mesure et ainsi simplifié l’application de la méthode.

Finalement des notations (moyenne de toutes les dimensions) sont associé pour les 4 division PSP/IQ, ce qui permet d’identifier les forces et faiblesses de la DQ.

1. IQ

Dernièrement, nous somme sur un composant d’analyse de la situation, pour cela plusieurs points sont pris en compte :

* + Analyse des résultats établis précédemment en ajoutant des éléments d’analyses, comme des pondérations.
  + Benchmark par rapport aux meilleures entreprises du même domaine, pour visualiser les éléments d’améliorations.
  + Comparaison des différentes visions de la qualité des données du côté gestionnaire et consommateur, ayant pour but d’établir des décalages de visions et pouvoir améliorer la synchronisation entre les différentes parties.
* **(ORME) DQ assessment methods**

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

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 se comporte également de 4 phases :

1. Reconstruction de l’état actuel.

Pour savoir quels sont les éléments à prioriser on regarde tous les flux de données utilisé et échangé entre les différentes parties d’une organisation.

Cette phase permet de mettre en avant les principales données, ainsi que leurs principaux utilisateurs et gestionnaire.

A la fin de cette étape nous avons donc un plan des point critiques (utilisateurs, processus, générateur) des données.

1. Identification des risques économiques lié 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 3 grandes familles :

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

1. 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 économique, il est possible d’associer à une sonde les couts économiques direct et indirect lié 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.

1. 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[[8]](#footnote-9), 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.

1. Une image contenant texte, capture d’écran, diagramme, conception

   Description générée automatiquementModule 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)

1. 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 définis les spécification des sondes (ID,mesure,heure), Dans ce modèle les sondes (ie. Probes) sont des éléments qui se déclenche à intervalle bien précis et qui donne un état des lieux de la qualité à un instant t. Il existe un module de gestion des sondes, mais nous n’irons pas dans le détail

1. 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 a été effectué cette dernière ou encore les parties de l’organisation jouant un rôle par rapport à la mesure.

1. 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.

# Analyse.

## Analyse de l’état de l’art.

Une image contenant texte, capture d’écran, Police, conception

Description générée automatiquementNous avons étudié certaines solutions primitives, ainsi que leurs familles, dans le cadre de notre problématique nous allons replacer ces familles de solutions par rapport au cycle d’utilisation de ces dernières

Nous avons plusieurs stratégies de gestion de la qualité de la donnée notamment la méthode TDQM qui est un des piliers dans le domaine de la DQ, créée par Richard Yang qui participera à un grand nombre de recherche et qui est un auteur populaire dans le domaine. Il participe notamment à l’élaboration de l’AIMQ.

Logiquement, les 3 méthodes étudiées effectuent des mesures et des analyses sur les données pour en définir les caractéristiques principales. Cependant l’utilisation des dimensions et mesures ne sont pas tout à fait les mêmes, TDQM et AIMQ effectue même un regroupement des dimensions dans 4 groupes.

ORME-DQ et TDQM effectuent toutes 2 une cartographie, sur cette étape on identifie donc indirectement le cycle de vie des données, mais cela reste un élément d’arrière-plan. Alors que dans la 3 éme Méthode nous n’avons pas vraiment de phase d’analyse des flux. Les étapes du cycle de vie de la donnée sont gérées de manière implicite dans les méthodologies. Ces dernières possèdent également éléments de pilotage, ce qui permet de garantir une qualité élevée dans le temps. Dans ces méthodes il y a donc non seulement du diagnostiques mais égalent de la prévention.

En ce qui concerne la méthode AIMQ, elle est quant à elle constituée de solution uniquement de type diagnostique (analyse, audit, questionnaire). Elle donc plus simple et nécessite moins d’expertise que les autres, ce qui la rend plus accessible. Cependant, elle ne constitue pas une méthode aussi complète que les 2 autres. Elle présente cependant un avantage, c’est qu’elle compare ces résultats par rapport à des concurrents ce qui est pertinent car chaque domaine

ORME DQ est la seule méthodologie des 3 ou les coûts économiques sont placé au centre du raisonnement. Ce qui peut être à la fois un avantage ou inconvénient, notamment car associer des failles de DQ à des pertes potentielles est relativement complexe. Mais cela permet de garantir une priorisation optimale de la gestion de la qualité et de mettre en confiance la direction en quantifiant les gains-pertes probables. Elle est également différente par son principe de sonde, qui remonte des informations sur

Un point faible de la méthode de la méthode ORME-DQ par rapport au 2 autres est que cette dernière n’exploite pas de questionnaires pour mettre en avant les différentes visions des utilisateurs. Or ces questionnaires bien qu’étant subjectif, permettent de mettre en lumière des incohérences techniques et fonctionnelles, et donc d’améliorer la mise en place.

|  |  |  |  |
| --- | --- | --- | --- |
| Caractéristiques / Méthode | TDQM | ORME DQ | AIMQ |
| Cartographie | X | X |  |
|  |  |  |  |

Pour clôturer cette analyse, nous donnerons 3 cas différents qui permettront d’illustrer les arguments précédents ce que nous avons vu précédemment et les solutions qui seraient les plus optimales :

**Cas 1** : PME industrielle avec un chiffre d’affaires d’une trentaine de millions d’euro, voulant 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 sera préférable de contacter un petit cabinet de conseil ou des éléments en freelance[[9]](#footnote-10), nous prendrons un cout journalier de 700[[10]](#footnote-11) euros par éléments recrutés. L’entreprise prendra 3 éléments pour une durée de 6 mois (≈ 400 000€). Les éléments ne chercheront pas à mettre en place une méthodologie complexe au sein de l’entreprise, il pourrait cependant s’appuyer sur les principes de l’analyse AIMQ. La première partie de leur travail sera de mettre en avant les dimensions de qualité cohérente, de les mesurer puis de les analyser. Ensuite selon ce que nous avons établi plus haut il faudrait mettre en place des solutions correctives puis préventives pour éviter la répétition des erreurs trouvé lors de l’analyse. Si le temps le permet mettre il faudrait mettre en place un peu de pilotage sur la collecte des données.

Cas 2 : Importante infrastructure de santé gouvernementale non lucrative, avec 70 millions d’utilisateurs.

Cas 3 : Groupe d’assurance mondiale présente un chiffre d’affaires de 150 milliards annuel. L’objectif est sur la partie française d’auditer sa qualité des données pour améliorer sa rentabilité

## Limite et implication de l’état actuel.

Dans les méthodologies les avis des différents utilisateurs jouant un rôle sont pris en compte. Sur la plupart des exemples les visions peuvent varier de manière importante. En générale les utilisateurs finaux ont une vision moins bonne que les gestionnaire, forme d’incertitude. Il y a une frontière entre monde technique et opérationnel qui pose un problème (Témoin de ce phénomène entreprise).

**Impact du turn Over** :

De plus 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 générales externalisés, ils sont aussi victimes de ces changements de gestionnaire fréquent, ce qui complique la gestion des outils. Ces causes amènent des conséquences :

* Manque de connaissances : La gestion données est comme nous avons pu le voir auparavant un élément complexe, sur lesquelles un temps d’apprentissage 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 les priorités sont amenés à changer, mais les outils mises en place ne peuvent pas vraiment suivre ces changements.
* Décalage entre métier et techniques : Les objectifs et les visions des 2 mondes étant mobiles, un accroissement de la frontière se créé.

Une des principales solutions est la documentation, de définir 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.

Économique

Ces méthodes nécessitent un engagement total de la direction de l’entreprise ce qui est compliqué sachant que les coûts de mise en place à l’échelle entreprise sont important et que les gains sont difficilement quantifiables. Assez compliqué de corréler qualité des données et cout économique absolue, seulement estimation comme dans l’ORME-DQ.

Il y a une complexité, les méthodologies se basent sur des éléments assez techniques (AIQM + simple mais complexe malgré tout), nécessité d’avoir des métiers spécialisés.

* Méthodologie de correction + que des méthodologies préventives
* Dans un jeu de donnée il peut y avoir des données avec des lignes plus importantes que d’autre, si une ligne est fausse pour un client à 10 euros ok mais si 100 000 euros moins cool

## Évolutions/ améliorations possibles ?

Mieux vaut prévenir que guérir

Gain de simplicité tant dans les flux métier et

Pris en compte dans les formations

# En interne

Pas particulièrement de méthodologie précise comme TDQM qui s’appliquent mais plutôt l’application de solution simple à des endroits stratégiques des processus.

Des outils comme Power BI conduisent indirectement à une amélioration de la qualité des données, car il automatise et normalise les échanges de données entre les différentes parties de l’entreprise. Il limite les opérations possibles des consommateurs sur les données.

La documentation est également importante

Il y a également des contrôle métier qui sont effectué tout le long des processus

Chaine rétroactive Plusieurs maillons qui se contrôle les uns les autres.

*Avoir un aller-retour entre les différents partis pour avoir leur vision de la qualité*

*Avoir une vision globale des choses et des résultats.*

*Directement dans les bases :*

*CONSOLIDATION : EDF - GRDF*

*Suivi des données*

*Verif d’après vérité terrain*

# Conclusion (2 pages)

Impact écologique

# Références générales

# Bibliographie

1. Introduction (3 pages)

1.1. Mise en contexte rapide

1.2. Revenir sur la problématique

1.3. Explication du plan et de sa logique

2. Contexte en entreprise (3-5 pages)

2.1. Présentation d’Electricité De France (EDF)

2.2. Focus sur l’équipe Portefeuille Client Contrat et du pôle GAZ.

2.3. Contexte de la problématique

3. Chaine d’utilisation des données

4. Etat de l’art (7-10 pages)

<https://www.claranet.com/fr/expertises/data-modernisation/big-data/data-et-big-data-comprendre-la-chaine-de-valeur>

<https://www.talend.com/fr/resources/cycle-vie-donnees/#:~:text=Le%20Data%20lifecycle%20management%20(DLM,collecte%2Fcr%C3%A9ation%20%C3%A0%20sa%20suppression>

<https://www.inist.fr/wp-content/uploads/donnees/co/module_Donnees_recherche_27.html>

AIMQ - http://mitiq.mit.edu/Documents/Publications/TDQMpub/2002/AIMQ.pdf

<https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2022.850611/full>

4.1. Identification des sources

4.2. Retranscription des sources

4.3. Principaux problèmes de qualité des données

4.4. Limites et implications de l’état actuel

5. Analyse (7-10 pages)

5.1. Comparaison entre état de l’art et réalité en entreprise

5.2. Illustration par un projet (en rapport avec la qualité des données) effectué durant mon alternance.

5.3. Mesure de la qualité des données

5.4. Evolutions possibles ?

6. Conclusion (2 pages)

7. Références générales

8. Bibliographie

1. Variations réelles : Inverse de prévisionnelle, c’est ce qui a été réalisé. [↑](#footnote-ref-2)
2. Power BI : Outil développé par Microsoft, permettant d’automatiser l’extraction, l’analyse et la visualisation des données. [↑](#footnote-ref-3)
3. Data flows ⬄ Flux de données. [↑](#footnote-ref-4)
4. QD ó Qualité des Données [↑](#footnote-ref-5)
5. Représentatif du prix du MWH des moyennes du marché sur 2023 [↑](#footnote-ref-6)
6. Cette matrice se réfère à l’ouvrage suivant : *Ballou, D.P., Wang, R.Y., Pazer, H., and Tayi, G.K. Modeling informa- tion manufacturing systems to determine information product quality. Management Science (1997).*  [↑](#footnote-ref-7)
7. Benchmark : Comparaison qualitative, qui étudie les différentes techniques et résultats obtenus par les organisations concurrentes. [↑](#footnote-ref-8)
8. Méthodes de classification utilisée dans des disciplines telles que le Machine Learning. [↑](#footnote-ref-9)
9. Travailleur indépendant. [↑](#footnote-ref-10)
10. Sur la base des profils freelance présent sur le site [malt](https://www.malt.fr/s?q=data+quality&sourceComponent=navbar_search_expanded). [↑](#footnote-ref-11)