Mémoire M1 MIAGE

-

ANDRIN Mathieu

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



Table des matières

[1. Introduction (3 pages) 3](#_Toc173657181)

[1.1. Mise en contexte rapide 3](#_Toc173657182)

[1.2. Définitions de la problématique 4](#_Toc173657183)

[1.3. Explication du plan et de sa logique 5](#_Toc173657184)

[2. Contexte en entreprise (2-3 pages) 5](#_Toc173657185)

[2.1. Présentation d’Electricité De France (EDF) 5](#_Toc173657186)

[2.2. Focus sur l’équipe Portefeuille Client Contrat et du pôle GAZ. 6](#_Toc173657187)

[2.3. Contexte de la problématique 6](#_Toc173657188)

[3. Chaine d’utilisation des données 6](#_Toc173657189)

[3.1. La donnée 7](#_Toc173657190)

[3.2. Qu’est-ce que le cycle de vie d’une donnée. 8](#_Toc173657191)

[3.3. Les étapes les plus sensibles en termes de qualités des données 9](#_Toc173657192)

[4. Etat de l’art (7-10 pages) 9](#_Toc173657193)

[4.1. Introduction 9](#_Toc173657194)

[4.2. Principaux problèmes et caractéristiques de qualité des données 10](#_Toc173657195)

[4.3. Quels sont les mesures de la qualité des données les plus populaires (à passer avant) 10](#_Toc173657196)

[4.4. Les solutions existantes. 12](#_Toc173657197)

[4.5. Les Méthodologies existantes. 14](#_Toc173657198)

[4.6. Limites et implications de l’état actuel 16](#_Toc173657199)

[5. Analyse (7-10 pages) 17](#_Toc173657200)

[5.1. (Introduction) 17](#_Toc173657201)

[5.2. Point sur l’existant dans mon entreprise par rapport à l’état de l’art 17](#_Toc173657202)

[5.3. Mesure de la qualité des données. 17](#_Toc173657203)

[5.4. Evolutions/ améliorations possibles ? 18](#_Toc173657204)

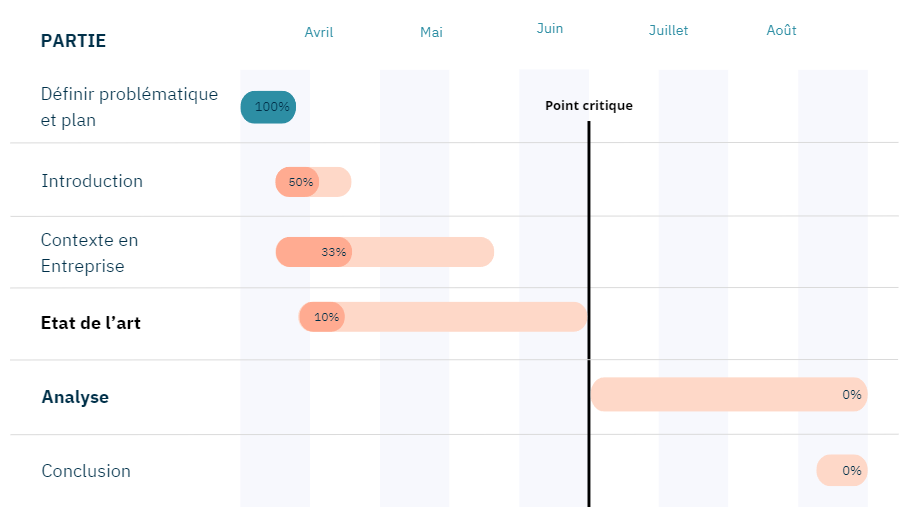
[5.5. (à voir si pertinent)Exemple de projet (en rapport avec la qualité des données) effectué durant mon alternance. 18](#_Toc173657205)

[6. Conclusion (2 pages) 18](#_Toc173657206)

[7. Références générales 18](#_Toc173657207)

[8. Bibliographie 18](#_Toc173657208)

# Introduction (3 pages)



## Mise en contexte rapide

**Contexte et justification :**

Actuellement Alternant en M1 MIAGE à l’université Paris Dauphine PSL, au sein d’EDF[[1]](#footnote-2) en tant que data analyst au sein du portefeuille GAZ de l’entreprise. Dans le cadre de la réalisation d’un mémoire de première année de master, nous allons réfléchir sur la gestion de la qualité des données tout au long de leurs cycles de vie.

Petit passage sur la DATA -> essence de l’économie actuelle, de + en + important (Machine learning, prévision, gestion d’entreprise…) Cette discipline ayant connu un essor durant les dernières années se situe à la croisé des mondes informatiques et mathématiques et de l’informations.

Cependant au sein de la data une nouvelle problématique est apparu : la mauvaise qualité de la données qui peut engendre des client mécontent, collaborateur mécontent, ainsi que des pertes importantes d’argent. Selon le data WareHousing Institute cette problématique aurait par exemple un coût de plus de 600 milliards de dollars sur l’économie Américaine

En tant qu’alternant data-analyst ces problèmes de qualité font parties des principaux enjeux du poste, ces derniers peuvent rendre les analyses obsolètes

Ce mémoire me permettra donc d’appréhender des solutions scientifiques dans le domaine de la data dans lequel je compte continuer d’évoluer à la suite de me études et Egalement pourrait permettre de faire évoluer les méthodes actuelles de l’entreprise.

**Objectif :**

L’objectif étant de couvrir le problème de la manière la plus générale possible.

Pas seulement applicable EDF -> pour toutes instances de ce problème

Avoir une multiplicité de solution

Différence information et une donnée.

## Définitions de la problématique

Ce sujet de réflexion a été trouvé après plusieurs échanges avec Entreprise (Sylvie) et Ecole (Florian), objectif -> réduction de la portée du problème et prendre un sujet déjà traité par (avoir un état de l’art existant pour alimenter la réflexion).

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

Qualité des données

La qualité des données se réfère à la mesure dans laquelle les données sont précises, complètes, fiables, cohérentes et pertinentes pour leur utilisation prévue. Une bonne qualité de données est essentielle pour garantir que les analyses, les décisions et les processus basés sur ces données sont fiables et efficaces. Les dimensions de la qualité des données comprennent souvent l'exactitude, la complétude, la cohérence, la validité, la ponctualité et la fiabilité.

Cycles d’utilisation des données

Les cycles d'utilisation des données font référence aux différentes étapes ou phases par lesquelles les données passent depuis leur collecte jusqu'à leur utilisation pour des analyses, des rapports, des décisions ou d'autres applications. Ces cycles peuvent inclure la collecte des données, leur prétraitement et nettoyage, leur analyse, leur interprétation, leur communication des résultats, leur prise de décision, ainsi que le retour d'information et l'itération sur les processus en fonction des résultats obtenus. Les cycles d'utilisation des données sont souvent itératifs et peuvent varier en fonction des besoins spécifiques et des objectifs des projets ou des organisations. Nous reviendrons sur ces différents éléments dans le mémoire

## Explication du plan et de sa logique

Explication de la logique du plan

Globalement le contexte permet de donner l’existant et du crédit au sujet,

Ensuite nous reviendrons précisément sur les étapes qui compose le cycle de vie de la donnée leurs différents enjeux.

Finalement, l’analyse permettra de croiser la réalité et l’état actuel de cette science. On cherchera à donner des applications à ce qui sera décrit dans l’état de l’art

# Contexte en entreprise (2-3 pages)

## Présentation d’Electricité De France (EDF)

Electricité 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.

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

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

L’objectif du pôle Gaz est d’optimiser l’équilibre économique du sourcing[[2]](#footnote-3) 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 : Equipe dans laquelle j’évolue responsable de toute la partie amont du processus -> gestion des donnés / qualité des données. Nous établissons un bilan sur les variations réelles (inverse de prévisionnelles) du portefeuille, nouveaux clients, sorties de client. L’équipe a également un rôle de comptabilité (Facture d’achat de gaz et de coût d’achemienement). Ainsi que la gestion de la souscription des grosses capacités sur le réseaux (éviter une surcharge).

Schéma récapitulatif du fonctionnement :

Il y a de grands enjeux financiers pour l’entreprise concernant ce

L’équipe PCC effectue également une collaboration interfonctionnelle : Dev, Marketing, commerciaux

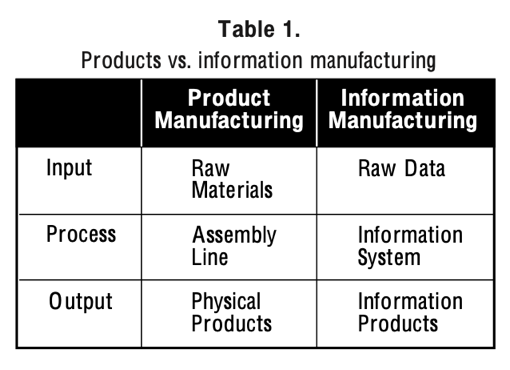
## Contexte de la problématique

Une grande partie du travail opérationnel au sein de l’équipe PCC et d’expliquer d’éventuels écarts de données, et dans certains cas des données de mauvaise qualité peuvent amenées à complexifié ces vérifications.

Aucune analyse ne peut être effectué sereinement s’il n’y a pas une confiance de la part de l’analyste sur la donnée

Nouvel outil -> question de mise en place d’outil de gestion de la qualité avec le SI

# La donnée et



## La donnée et son cycle de vie

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. 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 par leur granularité (détail), leur format (structure), leur nature (quantitative ou qualitative) et leur source (humaine ou machine). Les principaux types de données incluent 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. Méta-données : 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.

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

Description générée automatiquement](https://www.claranet.com/fr/expertises/data-modernisation/big-data/data-et-big-data-comprendre-la-chaine-de-valeur)*

Figure 1 : Graphique de la chaîne de valeur des données

Le schéma ci-dessus présente les différentes étapes de la chaine de valeur des données, Nous nous concentrerons sur les 2 premiers éléments.

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

Le data management existe dans pratiquement toutes les entreprises, de la collecte à la suppression des données se processus se divise en plusieurs étapes.

Donner les différents outils que l’entreprise possède en rapport à la qualité des données.

Les étapes sont les suivantes :

1. Collecte : étapes cruciales une des étapes les plus sensible dans laquelle beaucoup de solutions viennent s’incorporer, cette étape se situant en amont de la chaine si elle respecte la qualité des donnés alors les étapes suivantes peuvent espérer la respectée elles aussi.

Durant cette étape le principe du garbage IN ⬄ Garbage out s’applique, en effet lorsque les données proviennent d’entrée manuelle (saisie par un individu), et si ce-dernier effectue des gestes incorrects par rapport à la procédure SI cela peut poser un problème. Exemple : ne pas saisir sa vraie identité

1. Stockage : Cette étape comme son nom l’indique

Peu sensible -> la qualité de donné à cette étape est le reflet de la qualité des traitements effectués dans la phase de collecte, c’est ici que l’on peut mesurer l’impact de nos règles définis sur les entrées de datas.

En général, il y a un double stockage des données pour limiter les conséquences sur des altérations/suppressions non désirées en base.

Il est nécessaire que les données stockées soit conforme au réglementations en vigueur dans la région d’activité du pays.

Phase intermédiaire : Traitement - Nettoyage des données brut ou encore cryptage

1. Analyse :

Cette partie consiste à créer des données dérivées permettant d’alimenter des processus de décision au sein des entreprises. Également, il peut être possible d’exploiter les données présentes en base pour des études ponctuelles, ou encore pour établir des modèles de prédiction.

1. Sauvegarde/Archivage

Une fois que les données deviennent obsolètes sur la gestion de l’activité, elles sont placées dans des stockages secondaires (ex : datawarehouse), pour être réutilisé en cas d’étude sur le long terme ou éventuellement de litige.

1. Suppression.

La dernière étape du cycle de vie d’une donnée est la suppression la donnée.

Données supprimées -> Fin de période de rétention. Suppression de manière sécurisée pour éviter

\* Modification : Cet élément n’a pas été trouvé à partir de la littérature, mais il a dans notre cas une certaine importance, en effet la gestion de la modification de certaines données peut être importante

## Les étapes les plus sensibles en termes de qualités des données

Collecte : Cette phase est une phase critique, tout d’abord elle est la phase initiale de la vie d’une donnée, si la donnée rentre de mauvaise qualité, elle dégradera la suite du cycle.

Modification

Analyse :

Sensible à l’erreur humaine, erreur d’interprétation, évolution dans le temps.

Évolution de la compréhension d’une donnée, les normes évoluent en internes les manières de calculer peuvent être amené à évoluer avec l’arrivé de nouveauté, plusieurs exemples en entreprise. Cela se rapporte plutôt à un problème de qualité.

# Etat de l’art (7-10 pages)

## Introduction

Cette partie a pour objectif de répondre à différentes questions que nous nous sommes posés durant la réflexion sur le sujet.

Quels sont les éléments pouvant causer des erreurs, comment les détecter ?

*Contrôle de la qualité des données, qualité des prédictions ?*

*Comment gérer des éléments étranges dans les données ?*

<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>

Laure Berti-Equille, Directrice de recherche, Article de décembre 2004

Différentes phases de vie dans les données,

Ouvrage Global qui à l’image de ce mémoire traite le problème de manière global en apportant plusieurs solutions théoriques à différents problèmes

https://theses.hal.science/tel-04011007/

## Principaux problèmes et caractéristiques de qualité des données

On peut classifier la qualité des données dans 4 grandes classes qui regroupe des dimensions :

|  |  |  |  |
| --- | --- | --- | --- |
| Intrinsèque | Accessibilité | Contextuelle | Représentative |

Identification des principaux problèmes en qualité des données, permettra de mieux comprendre l’intérêt des mesures, et les caractéristiques de solution

* Redondance

L’information

* Valeur inexistante
* Valeurs aberrantes
* Incompréhension de la valeur

Il est important de noter des maintenant

## Quels sont les mesures de la qualité des données les plus populaires (à passer avant)

Dans cette partie, nous étudierons les principales mesures de qualité des données. 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.

Il existe 2 grande famille de mesure sur les données :

* Qualitatives
* Quantitatives

Les solutions seront étudiées à la suite des mesures car la plupart des solutions se basent sur les mesures de la qualité

Les 4 plus grandes :

* Précision ( accuracy)

Il est important de préciser que l’adjectif « correct » est subjectif : une donnée peut-être correct si elle n’a pas de valeur null ou bien si toute les valeurs sont justes par rapport à la vérité terrain

De même, à partir de quand considère-t-on une donnée correcte ?

* Complétude (completness)
* Consitance (Consistant)
* Opportunité (timeliness)

<https://www.cairn.info/revue-comptabilite-controle-audit-2005-1-page-23.htm>

Cluster / Groupe expérimental

Completness : <https://www.sciencedirect.com/science/article/pii/S0167923605000023> Vu en ML cette année, reprendre le cours

**AUTEURS METRIQUES VALIDATION**

**THEORIQUE EMPIRIQUE**

|  |  |  |  |
| --- | --- | --- | --- |
| Gray et al. [11] | *n* Complexité structurelle du modèle : *E*=∑*Ei* avec n entités Ei  *i=*1 Complexité pour l'entité i : *Di*=*Ri*∗(*a*∗*FDAi*+*b*∗*NFDAi*) avec 0 < a <= b, Ri = nombre d'associations, FDAi = nombre  d'attributs en dépendances fonctionnelles, NFDAi = nombre d'attributs sans dépendance fonctionnelle | Non | Non |
| Moody et al. [21] | **Complétude** : <nombre d'items du modèle ne correspondant pas aux spécifications des utilisateurs, nombre de spécifications non représentées dans le modèle, nombre d'items du modèle qui correspondent aux spécifications mais qui sont mal définis, nombre d'incohérences dans la modélisation>  **Intégrité** : <nombre de contraintes non satisfaites par les données, nombre de contraintes incluses dans le modèle qui ne correspondent pas exactement à la réalité modélisée>  **Flexibilité : <**nombre d'éléments dans le modèle sujets à modification, coût estimé des modifications, importance stratégique des modifications>  **Compréhensibilité** : <estimation par l'utilisateur du caractère compréhensible et interprétable du modèle> **Correction** : <nombre de violations des conventions du modèle de données, nombre de violations des formes normales,  nombre d'instances redondantes dans le modèle> **Simplicité** : <nombre d'entités et associations ; somme pondérée (aNE + bNR + cNA ), où NE est le nombre d'entités, NR le  nombre d'associations et NA le nombre d'attributs> **Intégration** : <nombre de conflits avec le modèle de données commun, nombre de conflits avec les systèmes existants> **Implémentabilité** : <estimation du risque technique, estimation du risque de planification, , estimation du coût de  développement, nombre d'éléments du niveau physique inclus dans le modèle> | Non | Non |
| Genero et al. in [27] | **NE** : nombre total d'entités dans le modèle ; **NA** : nombre total d'attributs d'entités et d'associations (simples ou composés); **DA** : nombre d'attributs dérivés (i.e., attributs dont la valeur peut être déduite ou calculée) ; **CA** : nombre total d'attributs composés ; **MVA** : nombre total d'attributs multi-valués ; **NR** : nombre total d'associations dans le modèle ; **M:NR** : nombre total d'associations M:N ; **1:NR** : nombre total d'associations 1:N et 1:1 ; **N-AryR** : nombre total d'associations N-aires ; **BinaryR** : nombre total d'associations binaires ; **NIS-AR** : nombre total d'associations IS\_A (généralisation/spécialisation) ; **RefR** : nombre total d'associations cycliques ; **RR** : nombre total d'association redondantes | Oui | Partiellem ent |
| Piattini et al. [28] | **RvsE:**rapportentrelenombred'associationsNRetlenombred'entitésNEdumodèle*RvsE*=(*NR* )2avecNR+NE>0. *NR*+*NE*  **DA:** rapport entre le nombre d'attributs dérivés NDA et le nombre maximal d'attributs dérivés possibles NA : *DA*= *NDA* avec *N A*−1  NA > 1. **CA:** rapport entre le nombre d'attributs composés NCA et le nombre d'attributs total NA : *CA*= *NCA* avec NA > 0.  *NA* **RR:** rapport entre le nombre d'associations redondantes NRR et le nombre d'associations NR : *RR*= *NRR* avec NR > 1.  *NR*−1 **M:NRel:** rapport entre NM:NR, le nombre d'associations M:N sur le nombre d'associations NR : *M*:*NRel*=*NM*:*NR* avec NR >  *NR*  1. **FLeaf :** mesure de la complexité des hiérarchies (IS\_A) : *FLeaf*=*NLeaf* avec NLeaf : nombre d'entités filles dans une  *NE*  hiérarchie généralisation/spécialisation et NE nombre d'entités dans chaque hiérarchie, NE > 0.  **IS\_Arel** : nombre moyen de supertypes directs et indirects par entité non racine ALLSup : *IS*\_*ARel*=*FLeaf*− *RLeaf ALLSup* | Non | Partiellem ent |

Pour les doublons mises en place d’outil sur la

Profiling ->donner un profil aux données

## Les solutions existantes.

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 : ISI, 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)

La plupart des algorithmes utilisés en solution sont basées sur les mesures.

Dans un premier temps nous étudierons les solutions

* **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 présent dans A alors A sinon B.

* **Vérification d’après vérité terrain**

Diagnostique/Correctives

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
* **Suivi des données** Préventive

Cette solution

* **Nettoyage** Correctives :

Partie automatique -> régles techniques (ex : contrôle format numéro tel) ou métier.

Partie « manuelle » contrôle par l’utilisateur

* Implémentation de règle de gestion de la BDD
* Modèle de Markov

## 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.

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

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

Description générée automatiquement

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

[Wang (1998)](https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2022.850611/full#B91)

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

Description générée automatiquement

1. Étude de l’existant :

Remix du PDCA

1. Définir donnée :
2. Mesurer la donnée :
3. Analyser :
4. Améliorer :

* **AIMQ (A Methodology for Information Quality Assessment)**

[Lee et al.](https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2022.850611/full#B57)

* **(ORME) DQ assessment methods**

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

Cette méthode se comporte 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.

Les relations peuvent être replacé dan

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 : Eleve/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, certain seuil 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[[3]](#footnote-4), en cherchant à classifier les éléments dans une classe « NoLoss » ou « Loss »

Pour réaliser ces différentes phases des modules sont utilisés :

Différent Module permette de réaliser ces phases :

1. Module extracteur de connaissance (Phase 1 et 2):
2. Module d’évaluation de la qualité des données (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 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:

1. Module de pilotage:

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

## Limites et implications de l’état actuel

# Analyse (7-10 pages)

Cette partie sera amenée à évoluer en fonction de l’état de l’art les parties sont données à titre indicatif.

Assez compliqué de corréler qualité des données et cout économique absolue, seulement estimation comme dans le ORME-DQ.

Méthodologie de correction + que des méthodologies préventives

## (Introduction)

Analyse

## Point sur l’existant dans mon entreprise par rapport à l’état de l’art

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

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

*CONSOLIDATION : EDF - GRDF*

*Suivi des données*

*Verif d’après vérité terrain*

## Mesure de la qualité des données.

Objectif : à l’image d’un médecin faire des analyses/diagnostiques.

Cette partie a pour objectif de mettre en avant l’état actuel des choses de manière scientifique et d’identifier certains problèmes existants.

*Qualité de la rétroaction, comment les maillons se corrigent les uns et les autres ?*

*Comparaison des résultats en prenant en compte les externalités ? (Covid)*

Est-ce que faire des clusters choisis pourraient-être intéressant ? (Direction, Région, Segment)

## Evolutions/ améliorations possibles ?

Cette partie présentera les améliorations possibles par rapport aux résultats sur la mesure de la qualité des données.

## (à voir si pertinent)Exemple de projet (en rapport avec la qualité des données) effectué durant mon alternance.

*Aspect automatisation qui a un impact sur la fiabilité et sur la qualité -> donner des mesures concrètes -> Projet de comparaison ou power BI pourraient être un bon exemple.*

# Conclusion (2 pages)

# 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

4.1. Identification des sources

4.2. Retranscription des sources

4.2.1. Source 1

4.2.2. Source 2

4.2.3. Source 3

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. EDF -> Electricité De France. [↑](#footnote-ref-2)
2. Sourcing : Achat d’un élément pour répondre à une demande, dans notre cas, cela correspond à l’achat du GAZ pour répondre à la demande cliente. [↑](#footnote-ref-3)
3. Méthodes de classification utilisée dans des disciplines telles que le Machine Learning. [↑](#footnote-ref-4)