Mémoire M1 MIAGE

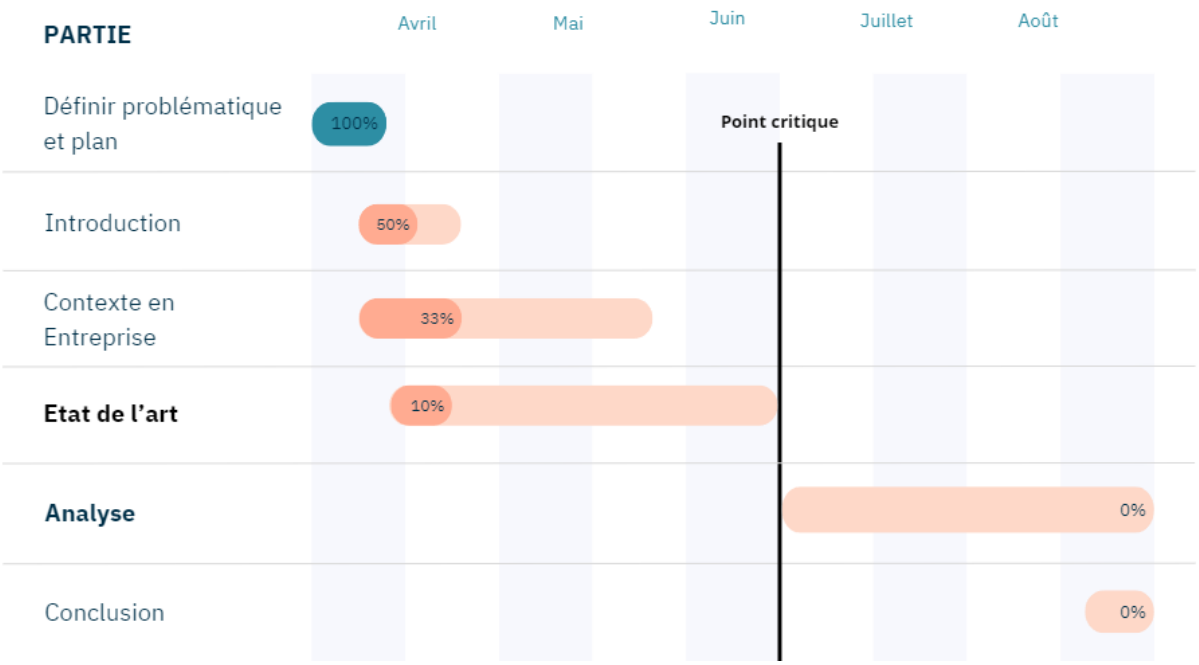
-

ANDRIN Mathieu

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

Table des matières

1. Introduction (3 pages)



1.1. Mise en contexte rapide

Contexte et justification :

Actuellement Alternant en M1 MIAGE à l'université Paris Dauphine PSL, au sein d'EDF[1] en tant que data analyste 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 (IA, prévision, ;....) Croisé des mondes informatiques et mathématiques.

[1] EDF -> Electricité De France.

Une mauvaise qualité -> client mécontent, collaborateur mécontent, pertes importantes d'argent

Les problèmes de la qualités existent

Dans mon rôle d'Alternant dans une équipe de data-analyst -> enjeux principaux contient analyse, visualisation et la qualité des données, Récurences de problèmes de qualité des données

Ce sujet me permettra d'appréhender des solutions scientifiques dans le domaine de la data dans lequel je compte continuer d'évoluer à la suite de me études. Permettra éventuellement de faire évoluer les méthodes actuelles de l'entreprise.

Objectif :

L'objectif étant de couvrir le problème de la manière la plus générale possible. Nous allons éffé

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

Avoir une multiplicité de solution

Différence information et une donnée.

1.2. Définitions de la problématique

Ce sujet de réflexion a été trouvé après plusieurs échanges avec Entreprise (Sylvie) et Ecole (Florian), objectif -> reduction de la portée du problème et prend un sujet commun (avoir un état de l'art existant pour alimeneter 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

1.3. 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, ainsi que la

Puis l'état de l'art donnera une expertise sur le sujet.

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

2. Contexte en entreprise (2-3 pages)

2.1. 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 actionnaire principal, 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.

2.2. 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 pôles Gaz est d'optimiser l'équilibre économique du sourcing[2] 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 : Reponsable des ordres d'achats du gaz en bout de chaine
- Prevision : 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 /etc

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

2.3. Contexte de la problématique

3. Chaine d'utilisation des données

3.1. Qu'est-ce que le cycle de vie d'une données.

Big Data :

[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.

Tout le monde partage des données pour la plupart des échanges en réseaux

Il y a intérêts commerciaux pour les entreprises -> Génération de modèles prévisionnels. *Illustration par les applications en "Ordonnancement et Gestion de production" -> Les prévisions sont à la base de la plupart des modèles de gestion, hors ces dernières s'alimentent des données passés.*

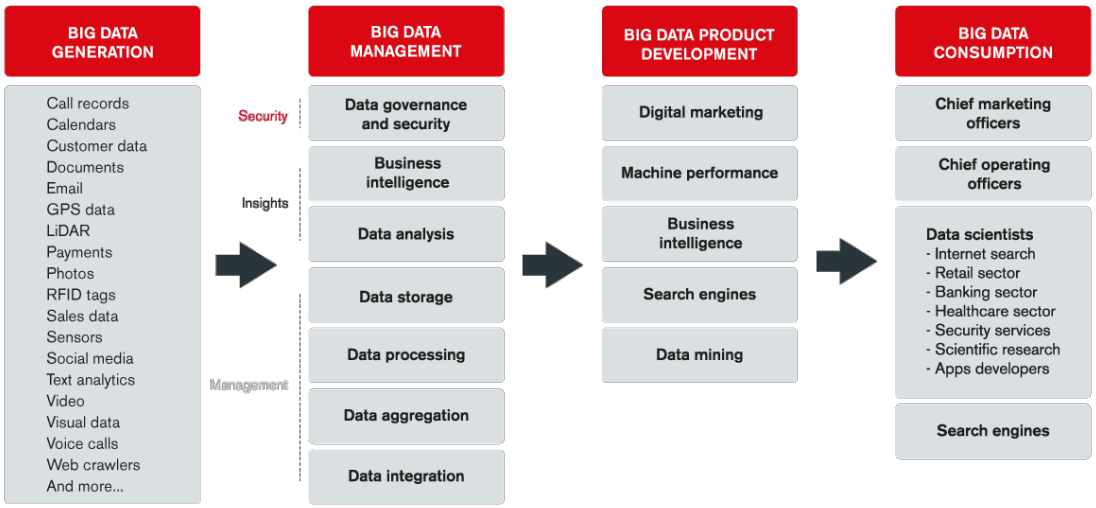


Figure 1 : Graphique de la chaîne de valeur des données

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

Collecte de données ->

Big Data management

Impact financier difficilement quantifiable mais très important exemple EDF GAZ

Définir une BDD/Entrepôt de données et son intérêt (emplacement physique, « maison des données »)

3.2. Les différentes étapes.

Le data management existe dans pratiquement toutes les entreprises, de là de 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 :

2. Stockage :
3. Traitement :

Changement de format, sécurisation(cryptage), nettoyage

4. Analyse
5. Sauvegarde
   Double copie des données permettant de sécuriser ces derniéres
6. Réutilisation
7. Suppression

3.3. Les étapes les plus sensibles en termes de qualités des données

Collecte : Données érronés

Traitement :

Analyse :

4. Etat de l'art (7-10 pages)

4.1. 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 ?

La logique actuelle de mon état de l'art est la suivante : Quels sont les principaux problèmes et leurs solutions ? Et comment reconnaitre la présence de problèmes et leurs sources (Mesure mathématique et informatique)

4.2. Retranscription des sources (à enlever, je placerais ces éléments dans les parties suivantes

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

4.2.1. Article 1

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

Exemple de couple **Problèmes/Solution** :

- Incertain dans les BDD / La théorie des probabilités, la théorie de Dempster-Schafer et la théorie des ensembles flous.
- la complétude du résultat d'une requête sur des relations partiellement complètes [20][17]/
- /l'extension des langages de requêtes pour permettre le nettoyage des données a également suscité bon nombre de travaux de recherche dans le domaine [29][10][30][33][5]
- Informations incomplètes. /

Détection et correction des données :

Différents types d'approche sur **les solutions** :

- Diagnostiques : analyse mathématique permettant de faire ressortir les données
- Préventives : ISI, évaluer la qualité d'une BDD (relation, processus, traitement,..)
- Adaptatives : Nettoyage et consolidation / utilisation ETL
- Correctives : Temps réel -> vérification dans la requête (ex : on ne prend pas les valeurs nulls)

Exemple **de solution** :

- Verif d'après la vérité terrain :
  CONSOLIDATION -> comparé plusieurs basent entre elles, si identiques alors correcte sinon difficulté à connaitre la valeur juste (Imputation). Probléme : identique pas forcément juste, et pas de contrôle à l'entrée. Idée de la matrice de confusion
- Audit : programme qui contrôle certaines caractéristiques, simple à mettre en place.
- Suivi de données :
- Nettoyage des données
-

4.2.2. Article 2

4.2.3. Article 3

4.3. Principaux problèmes de qualité des données

4.4. Les solutions existantes.

4.5. Quels sont les mesures de la qualité des données les plus populaires

4.6. Limites et implications de l'état actuel

5. 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.

5.1. (Introduction)

5.2. 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

5.3. Mesure de la qualité des données.

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)

5.4. Evolutions/ améliorations possibles ?

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

5.5. (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.

6. Conclusion (2 pages)

7. Références générales

8. 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

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