Projet Data Lake

Rapport sur le Projet Data Lake

BOUCHIDA YOUSSEF

HAJAR AKHERRAZ

MERIEM ACHAQ

---

21 janvier 2024

1 Introduction

Au coeur de l'ère numérique actuelle, la gestion et l'analyse des données massives constituent le pilier central des avancées technologiques et analytiques en entreprise. Dans ce contexte, notre formation en Master 2 Business Intelligence et Analytics, nous confère la responsabilité et le défi d'appréhender et de maîtriser les architectures de données de nouvelle génération, parmi lesquelles le Data Lake s'impose comme une solution incontournable.

Ce projet, s'insérant dans un cadre académique rigoureux et pragmatique, vise à concrétiser nos acquis théoriques par la mise en oeuvre d'un Data Lake opérationnel, capable de stocker, de traiter et de valoriser des volumes conséquents de données hétérogènes. Cette initiative nous permet de simuler une situation réelle où la polyvalence des formats et la diversité des sources de données mettent à l'épreuve notre agilité et notre capacité à innover.

Dans cette optique, la construction de notre Data Lake suit une stratégie structurée en trois zones distinctes, chacune correspondant à un niveau de traitement et de raffinement des données. La Landing Zone, notre point d'entrée des données brutes, la Curated Zone, où les données sont transformées et qualifiées, et la Refined Zone, qui prépare les données pour l'exploration et les analyses avancées.

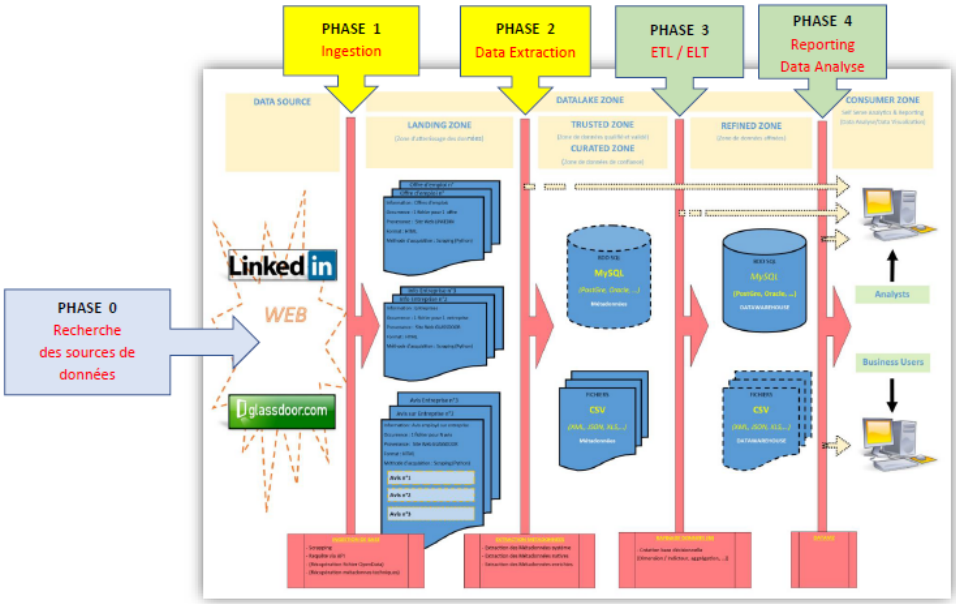

FIGURE 1 – Schèma représentant les différentes phases

L'ambition est de doter le Data Lake d'une flexibilité et d'une évolutivité telles, qu'il puisse répondre aux besoins analytiques et opérationnels d'une entreprise fictive. Cela, en exploitant des jeux de données réalistes, notamment des offres d'emploi extraites depuis les sites web LinkedIn et Glassdoor, reflétant ainsi les tendances actuelles du marché du travail dans le secteur de l'informatique.

Ce document, faisant office d'introduction à notre démarche, détaille les fondements, les motivations et les objectifs qui sous-tendent la réalisation de ce Data Lake. Il sert de prélude aux explications techniques et aux analyses qui seront développées par la suite, illustrant notre approche méthodique et notre contribution à la science des données.

2 Phase 1 : Ingestion

La première phase de notre projet consistait à ingérer des données, une étape cruciale dans la construction d'un Data Lake fiable. Le dossier "0_SOURCE_WEB" était notre point de départ initial, contenant une abondance de fichiers HTML issus de scrapping méticuleux des sites LinkedIn et Glassdoor. Cette hétérogénéité de données constituait notre matière première à traiter.

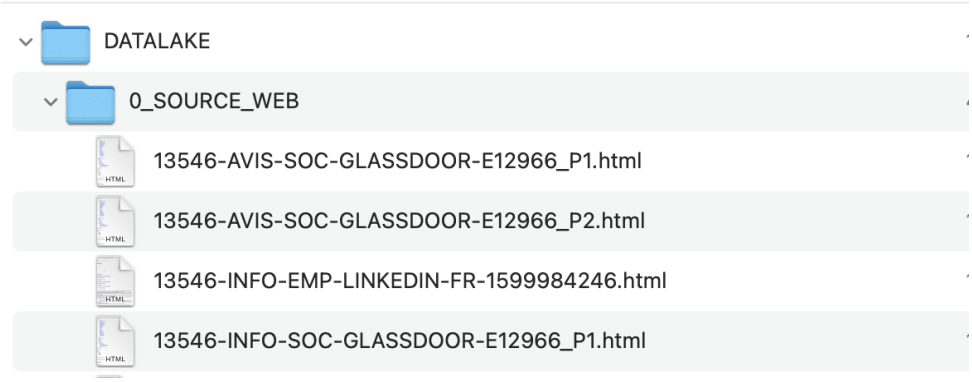

FIGURE 2 – Le dossier "0_SOURCE_WEB"

Afin de catégoriser efficacement cette masse d'informations, nous avons élaboré un script Python dédié à la séparation des données en fonction de leur provenance. Ce processus de ségrégation a abouti à la création de deux dossiers distincts, "LinkedIn" et "Glassdoor", tous deux logés dans le répertoire "1_LANDING_ZONE". Cette organisation initiale était essentielle pour préparer le terrain aux étapes d'extraction et de transformation qui allaient suivre.

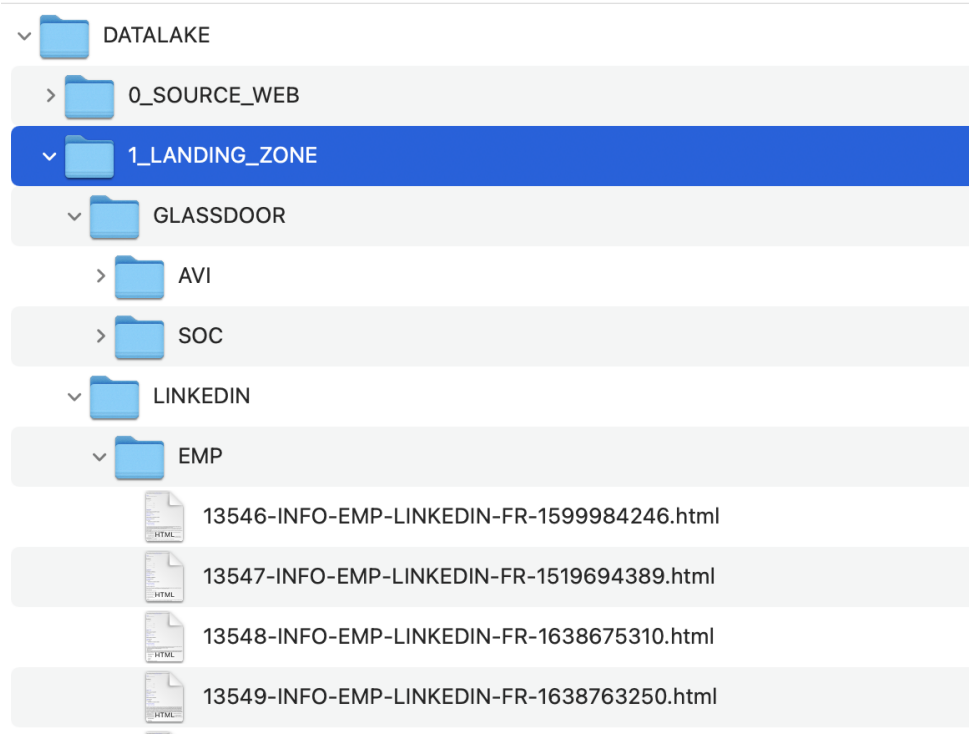

FIGURE 3 – Le dossier "1_LANDING_ZONE"

3 Phase 2 : Extraction des Données

L'exploration des données a commencé par une observation attentive du contenu des pages HTML via Notepad++, nous permettant de déterminer les balises significatives. Cette analyse préliminaire était nécessaire pour cibler les informations pertinentes et structurer notre approche d'extraction. Avec cette compréhension approfondie, le script Python que nous avons conçu a pu cibler et extraire les données avec précision, en les convertissant en un format CSV exploitable.

FIGURE 4 – Repération des balises avec Notepad++

Les fichiers CSV générés, représentant des tables de données, ont été méticuleusement stockés dans "2_CURATED_ZONE". Ce répertoire intermédiaire fonctionnait comme une bibliothèque de données nettoyées, prêtes pour l'intégration et l'analyse. Par la suite, ces tables ont été transférées vers "3_PRODUCTION_ZONE" et le dossier "BDD", signifiant la finalisation de la préparation des données et leur aptitude à être utilisées dans un contexte de production. La création d'une table de faits, en tirant parti des divers fichiers CSV, a permis d'établir un socle solide pour les visualisations et les insights à venir.

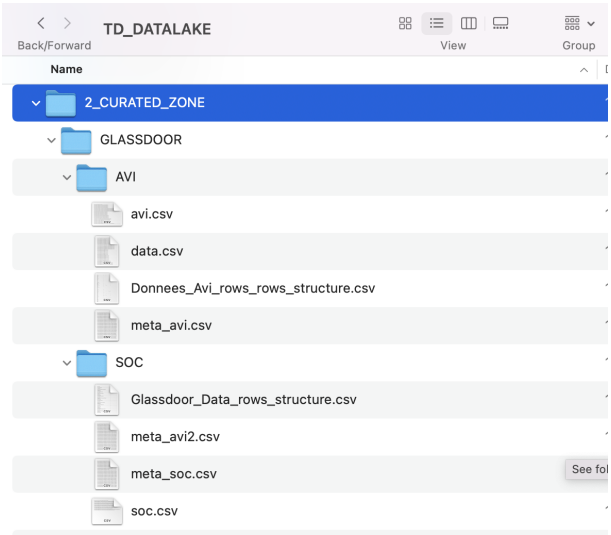

FIGURE 5 – Le dossier "2_CURATED_ZONE"

4 Phase 3 : ETL

Le nettoyage des données représente souvent une tâche ardue mais indispensable pour garantir la qualité et la fiabilité des analyses de données. En utilisant des scripts Python, nous avons implémenté des procédures ETL (Extract, Transform, Load) pour purifier et structurer nos données CSV. Chaque script a été soigneusement conçu pour traiter des aspects spécifiques de la qualité des données : suppression des doublons, correction des erreurs de formatage, et validation des types de données.

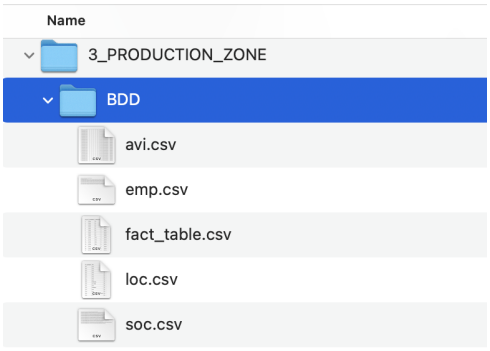

FIGURE 6 – Le dossier "3_PRODUCTION_ZONE"

Grâce à ces efforts méticuleux, nous avons obtenu des fichiers de données propres et organisés, prêts pour une utilisation analytique poussée. Ces fichiers structurés, désormais localisés dans "3_PRODUCTION_ZONE", formaient la base de notre modèle relationnel. La structuration cohérente des données nous a permis de visualiser un modèle relationnel cohérent et optimisé pour le reporting et l'analyse avancée, comme illustré ci-dessous.

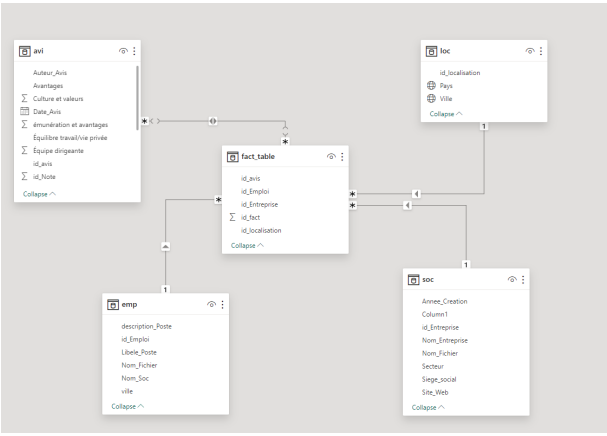

FIGURE 7 – Modèle relationnel de la base de données.

5 Phase 4 : Reporting

Au terme de notre projet Data Lake, nous avons atteint l'étape cruciale du reporting. Cette phase finale se concrétise par la création d'un ensemble de visualisations qui transforment nos données structurées en insights actionnables. Nous avons élaboré divers tableaux de bord qui non seulement reflètent la richesse des données extraites et traitées mais offrent également une interface intuitive pour l'exploration et l'analyse.

Le tableau de bord "Vue d'ensemble de l'entreprise" a été le point de départ de notre reporting. Nous avons conçu des KPIs tels que la moyenne des notes, la moyenne des salaires et le nombre total d'avis, qui sont présentés sous forme de cartes numériques, offrant une lecture immédiate des métriques clés. Un graphique en ligne illustre la tendance des notes moyennes au fil des années, révélant l'évolution de la satisfaction au sein des entreprises représentées. Un diagramme à barres et un graphique en camembert coloré affiche la répartition des entreprises par secteur, démontrant la diversité du paysage industriel dans notre ensemble de données.

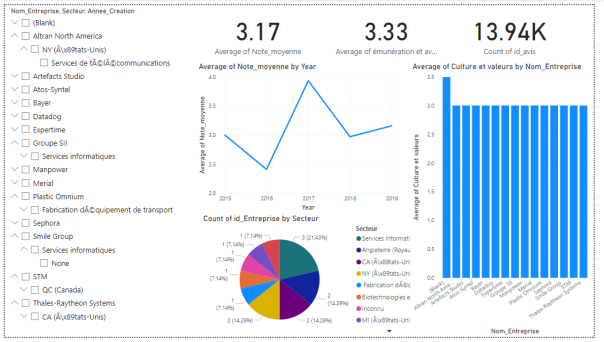

FIGURE 8 – Modèle relationnel de la base de données.

L'exploration se poursuit avec le tableau de bord "Sentiment des employés", qui met l'accent sur les perceptions et opinions des employés. À travers des visualisations telles que des nuages de mots pour les avantages et inconvénients, nous avons pu identifier les termes les plus fréquemment cités, ce qui donne un aperçu des points forts et des domaines à améliorer selon les employés.

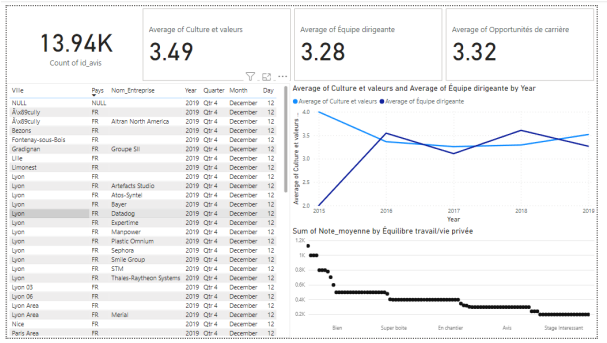

FIGURE 9 – Modèle relationnel de la base de données.

Le tableau de bord "Aperçu détaillé de l'entreprise" fournit une liste exhaustive des entreprises avec des informations pertinentes telles que le nom, le secteur, l'année de création, et les avantages et inconvénients mentionnés dans les avis. Cette table interactive

permet aux utilisateurs de filtrer les données selon différents critères pour affiner leur recherche et leur analyse.

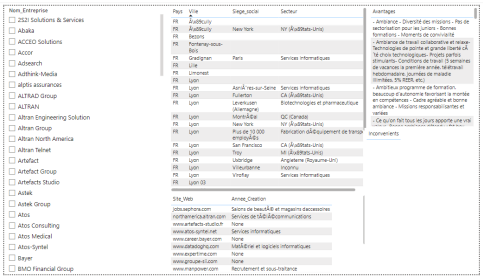

FIGURE 10 – Modèle relationnel de la base de données.

Enfin, le tableau de bord "Aperçu des insights de localisation" présente une carte géographique qui montre la distribution des emplois et des entreprises à travers différentes régions. Cette visualisation spatiale est accompagnée de barres horizontales qui comparent la moyenne des notes par ville, soulignant ainsi les variations géographiques de la satisfaction au travail. Chacun de ces tableaux de bord constitue un outil d'analyse

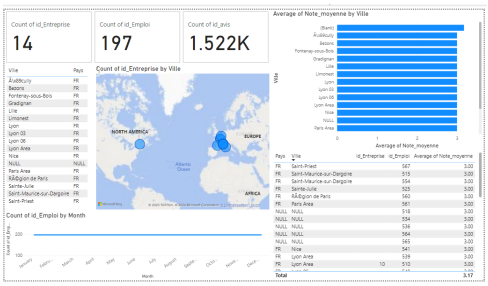

FIGURE 11 – Modèle relationnel de la base de données.

décisionnelle puissant, capable d'orienter les stratégies d'entreprise grâce à une meilleure compréhension des tendances du marché et des sentiments des employés. Ils sont le fruit d'un processus méticuleux d'ingestion, d'extraction, de traitement ETL, et de modélisation des données, témoignant de l'efficacité de notre pipeline de données du début à la fin.

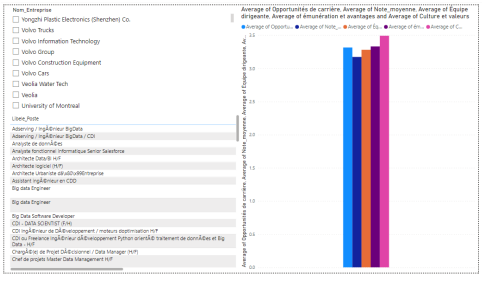

FIGURE 12 – Modèle relationnel de la base de données.

6 Conclusion

En finalisant ce projet ambitieux de Data Lake pour l'analyse de données massives, nous avons traversé un parcours exhaustif qui a illustré la puissance transformatrice de l'informatique décisionnelle dans le traitement et l'analyse des données. De l'ingestion initiale des données à la création de visualisations significatives, chaque étape a renforcé notre compréhension des défis et des opportunités qu'offre le Big Data.

La Phase 1, dédiée à l'ingestion, a posé les fondations de notre architecture de données. La séparation des fichiers HTML dans une structure organisée a permis de garantir que notre Data Lake démarre sur des bases solides, avec une distinction claire entre les données brutes issues de LinkedIn et Glassdoor. La création des répertoires "LinkedIn" et "Glassdoor" au sein de la "1_LANDING_ZONE" a non seulement reflété notre souci d'organisation mais a aussi préparé le terrain pour une extraction de données efficace.

La Phase 2 nous a confrontés directement avec le contenu des données. L'exploration des pages HTML via Notepad++ a révélé la complexité des données web et l'importance de cibler des informations pertinentes pour l'extraction. Les scripts Python développés pour cette tâche ont fait preuve d'une précision remarquable, transformant les données en fichiers CSV structurés pour le stockage dans "2_CURATED_ZONE". Cette transformation a facilité la transition des données vers "3_PRODUCTION_ZONE", où les données ont été finalement préparées pour l'analyse.

La Phase 3 a représenté l'épicentre de notre projet, où les données ont été nettoyées, transformées et chargées à l'aide de processus ETL méticuleux. Les fichiers CSV ont été purifiés de toute incohérence, rendant les données prêtes pour des analyses avancées. L'application d'un modèle relationnel correct et la création d'une table de faits ont concrétisé notre objectif de fournir une vue unifiée et intégrée des données.

Enfin, la Phase 4, le reporting, a été le point culminant de notre projet. Les visualisations créées, résultant d'un processus méticuleux d'ingestion, d'extraction, de traitement ETL et de modélisation des données, témoignent de l'efficacité de notre pipeline de données. Les tableaux de bord ont non seulement fourni une clarté sans précédent sur les données mais ont également ouvert la voie à des découvertes approfondies et à des décisions éclairées pour toutes les parties prenantes. Les visualisations, allant des cartes numériques aux graphiques interactifs, ont illustré de manière vivante les tendances du marché du travail et les sentiments des employés, soulignant l'importance des données dans la stratégie d'entreprise.

Ce projet a renforcé notre conviction que la maîtrise des Data Lakes et des compétences en visualisation sont essentielles pour tout professionnel de la donnée souhaitant apporter de la valeur dans un monde de plus en plus axé sur les données. Il a également confirmé l'importance de la flexibilité et de l'évolutivité dans la conception des systèmes de données pour répondre aux besoins changeants des analyses d'entreprise.

En regardant vers l'avenir, nous sommes convaincus que les compétences et les connaissances acquises grâce à ce projet nous positionnent favorablement pour relever les défis futurs dans le domaine de l'analyse de données massives. Nous sommes impatients de poursuivre notre voyage dans le vaste et dynamique paysage du Big Data, armés des outils et des stratégies développés au cours de ce projet captivant.