Rapport Projet

–

Données massives

MASTER 2 BUSINESS INTELLIGENCE & ANALYTICS

BROUTIER CHARLENE
TARHAMI KAWTAR
LAVERIE CAMILLE

PROFESSEUR RÉFÉRENT : ERIC KLOECKLE

Université Lumière Lyon 2
Institut de la communication

Table des matières

Glossaire

Mots techniques	Description
API	Interface de Programmation Applicative, utilisée dans le projet pour effectuer des requêtes au module Gemini et récupérer des informations sur les offres d'emploi.
Curated Zone	Deuxième zone de stockage des données. C'est l'étape où les données brutes de la Landing Zone sont extraites, transformées, nettoyées et enrichies (notamment via l'utilisation de Gemini) pour créer une base de données de qualité, contenant uniquement les informations pertinentes pour l'analyse. Dans ce projet, le dossier est vide puisque ce sont des données descriptives qui sont donc entreposées dans le fichier 'METADATA_CURATED_ZONE.csv' dans le dossier 00_METADATA.
Datavisualisation	Représentation graphique des données, réalisée dans le projet via Power BI pour afficher les résultats de façon claire et interactive.
Dictionnaire de données	Document listant les données jugées utiles pour l'analyse, en précisant leur nom, leur type et leur description.
ETL	Processus d'Extraction, Transformation et Chargement des données (Extract, Transform, Load) réalisé dans le projet grâce à des scripts Python.
Gemini	Intelligence Artificielle (IA) générative de Google utilisée pour enrichir les données en extrayant des informations clés (niveau d'étude, technologies, soft skills, etc.) à partir des descriptions d'offres d'emploi.
IA	Intelligence Artificielle, terme général pour désigner la technologie de Gemini dans le contexte du projet.
Landing Zone	Première zone de stockage temporaire des données brutes copiées depuis la source web (étape 1 de l'ETL).

Maquette	Représentation visuelle des écrans du rapport Power BI, réalisée en amont pour déterminer quelles données sont utiles à récupérer pour l'analyse.
Modèle de données	Schéma d'organisation des données en tables de dimensions (DIM) et tables de faits (FAIT), pour structurer les données avant la datavisualisation.
Production Zone	Dernière zone de stockage des données. C'est l'étape de chargement (Load) où les données raffinées de la Curated Zone sont pivotées et structurées sous forme de tables (ou fichiers "type BDD") pour être prêtes à la datavisualisation (rapport Power BI). Cela génère les fichiers finaux : `table_AVIS_SOC.csv`, `table_EMP.csv` et `table_INFO_SOC.csv`.
Python	Langage de programmation utilisé pour développer l'intégralité des scripts de la partie ETL.
Tables de dimensions (DIM)	Tables contenant des données descriptives (ex: entreprises, avis, emplois) utilisées comme critères de filtrage dans le rapport.
Table de fait (FAIT)	Table contenant les mesures (quantités) à analyser (ex: nombre d'avis, nombre d'emplois).
Tokens	Unités de consommation de l'IA (Gemini), mentionnées comme une contrainte d'utilisation dans le cadre du projet étudiant.
Workflow	Enchaînement d'opérations et de processus. Dans ce projet, le fichier workflow_datalake.bat permet d'exécuter automatiquement les différents scripts d'extraction, transformation et de chargement des données.

Introduction

Dans le cadre de l'optimisation des processus analytiques, nous avons réalisé un projet pour collecter et exploiter des données issues de sites internet et les analyser dans un rapport Power BI. Ce projet consiste à analyser des données sur les entreprises ainsi que leurs offres d'emploi et permettre la recherche d'informations sur ces sujets. Il permet de mieux comprendre les secteurs, entreprises et emplois grâce à des visualisations simples.

Dans un premier temps, nous allons présenter les prémices du projet avec les maquettes et le dictionnaire de données. Dans un second temps, nous allons détailler la phase d'ETL qui consiste à récupérer les données issues des sources web et à les préparer. Dans un dernier temps, nous allons restituer ces données sous forme de graphiques afin de réaliser des statistiques descriptives et faciliter la recherche d'entreprises et d'emplois.

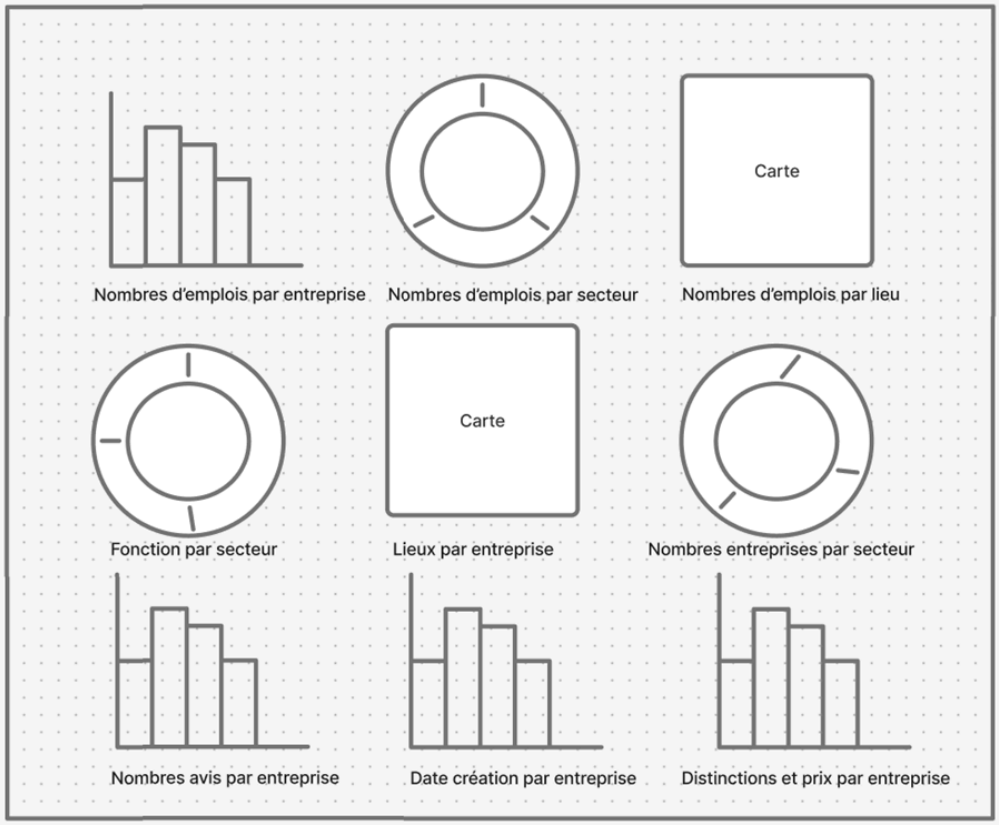
Informations complémentaires
Projet Github : https://github.com/CamilleLV/Lyon2_Data_lake_Project Lien du projet Google Drive : https://drive.google.com/drive/folders/1KrB7Kzc1aZHHAXFTqZHCsfLKSyCHk_ci?usp=sharing

1. Prémices du projet

1.1. Maquettes

Nous avons réalisé des maquettes dans le but de faciliter l'étape d'ETL qui consiste à récupérer les données. En effet, il est plus facile de représenter ce que l'on souhaite analyser pour ensuite récupérer seulement les données nécessaires.

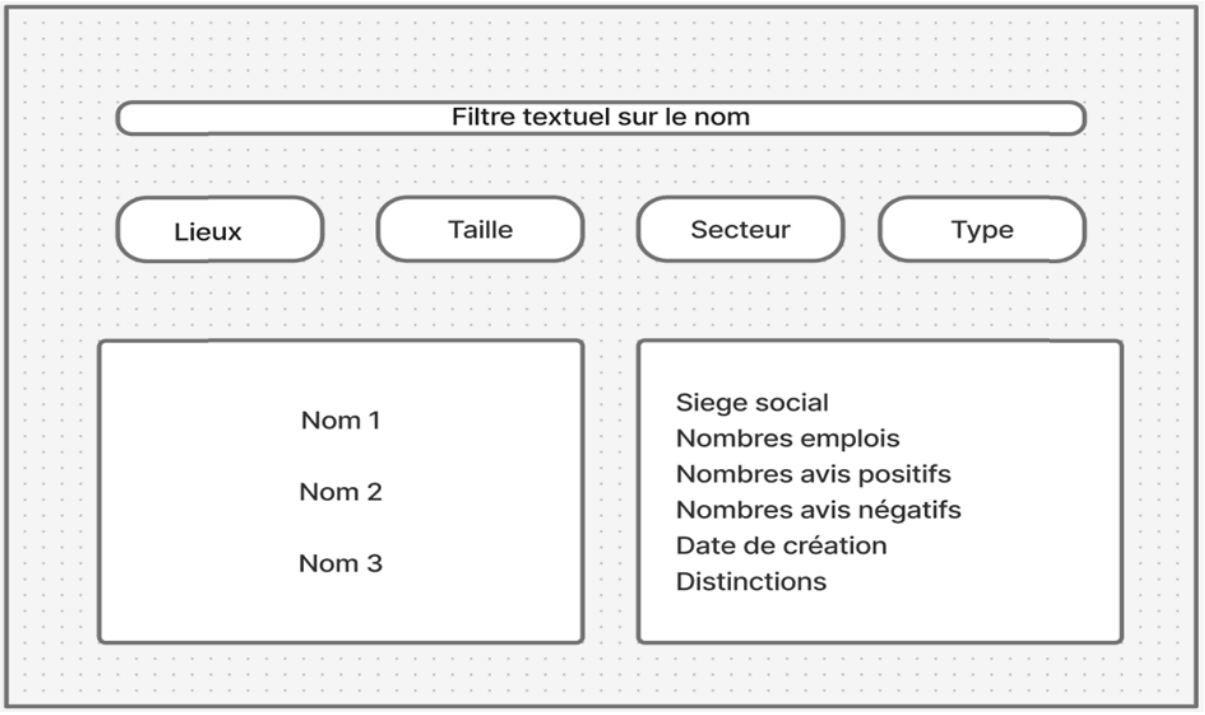
Tout d'abord, nous avons décidé de commencer notre rapport Power BI par une page de statistiques descriptives. Nous avons utilisé des histogrammes, des donuts et des cartes pour analyser des critères qui nous semblent importants lors de la recherche d'une entreprise ou d'un emploi c'est-à-dire les lieux, les secteurs, les avis, …



Capture 1: Maquette - Statistiques descriptives

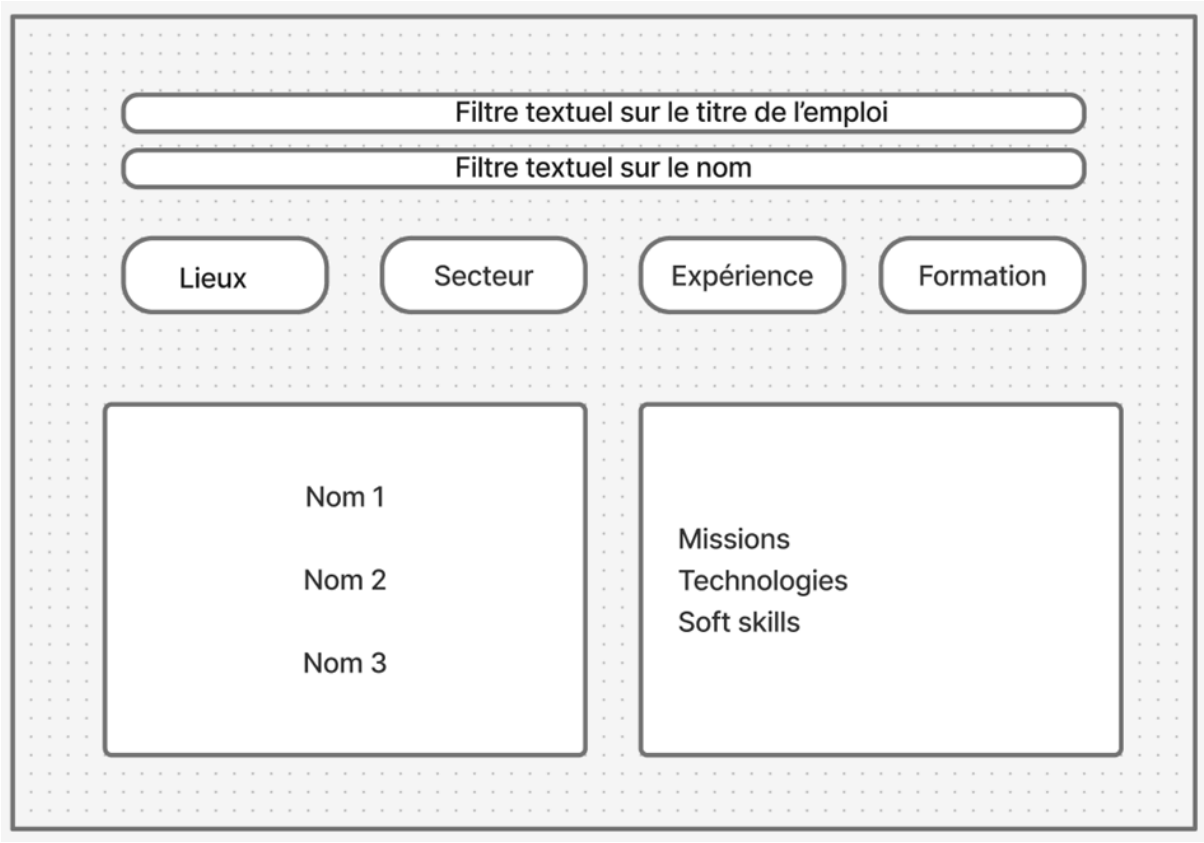
Ensuite, nous sommes partis sur une page "entreprises". C'est une page qui permet de filtrer une entreprise directement depuis son nom ou en utilisant les filtres:

lieux, taille, secteur et type. Une fois l'entreprise filtrée et sélectionnée, d'autres informations complémentaires, qui ne sont pas les filtres, sont affichées sur la droite.



Capture 2: Maquette - Informations entreprises

Enfin, nous avons réalisé une page "emplois" qui, similairement à la page "entreprises", permet de filtrer une offre. Elle comporte deux filtres directs: un filtre sur le titre de l'offre et un sur le nom de l'entreprise. Sinon, nous pouvons filtrer sur les lieux, le secteur, l'expérience et la formation. Une fois l'offre sélectionnée, une description plus précise de l'offre apparaît sur la droite.



Capture 3: Maquette - Informations emplois

Ces maquettes ont pour objectif de préciser quelles données récupérer lors de la phase d'ETL et ne reflètent pas le rapport de datavisualisation final.

1.2. Dictionnaire des données

Les maquettes nous ont permis de déterminer les données que nous jugeons utiles pour notre analyse. Nous avons donc décidé de construire un dictionnaire de données répertoriant, dans des tables, nos données en précisant leur nom, leur type et leur description.

Par exemple, nous avons créé une table "DIM_Soc" qui contient les données des entreprises. Nous y avons répertorié, entre autres, le nom de la société "Nom_soc, sous forme de chaîne de caractères et la taille de la société "Taille_soc" également sous forme de chaîne de caractères.

DIM_Soc

Nom de l'attribut	Type de l'attribut	Description de l'attribut
ID_soc	INT	Clé primaire de la dimension soc
Nom_soc	VARCHAR	Nom de la société
Taille_soc	VARCHAR	Taille de la société
Date_creation_soc	DATE	Date de création de la société
Siege_social_soc	VARCHAR	Siège social de la société
Type_soc	VARCHAR	Type de la société
Secteur_soc	VARCHAR	Secteur de la société
Lieu_soc	VARCHAR	Lieu de la société
Distinctions_prix_soc	VARCHAR	Distinctions et prix de la société
ID_departement	INT	Clé étrangère de la hiérarchie Departement

Capture 4: Extrait du dictionnaire de données

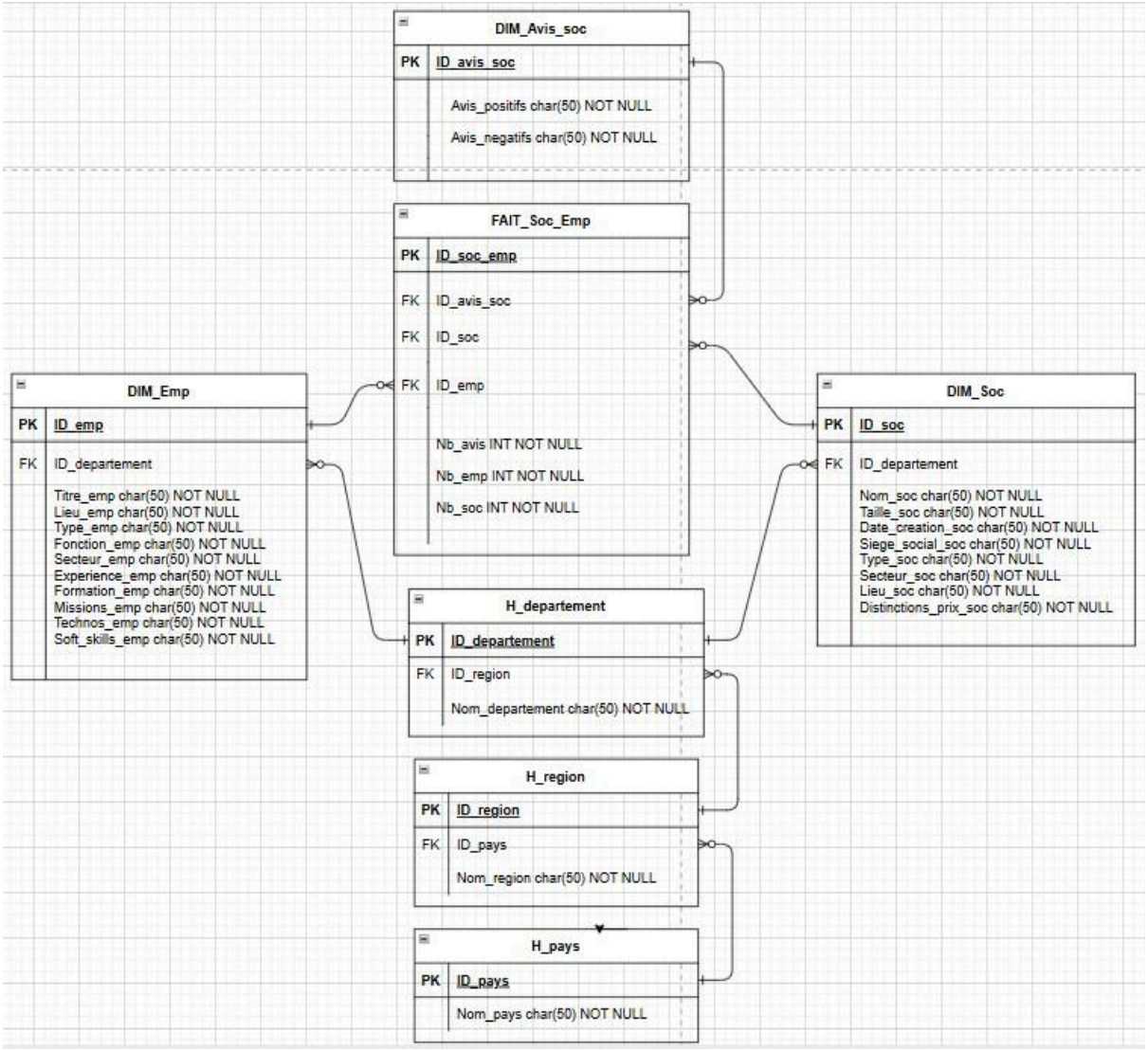
Nous avons précisé, dans le dictionnaire de données, l'intégralité des données que nous souhaitons récupérer. Le reste du dictionnaire se trouve dans le Drive. Ce dictionnaire nous permet ensuite de réfléchir à un modèle de données pour la partie datavisualisation.

1.3. Modèle de données

À la suite du dictionnaire de données, nous avons procédé à la réalisation du modèle de données. Il nous sert à organiser nos données en tables de dimensions, tables de faits et hiérarchies.

Tout d'abord, nous avons créé les tables de dimensions DIM_Avis_soc, DIM_Emp et DIM_Soc qui contiennent respectivement les données des avis des entreprises, des entreprises et des emplois. Ensuite, nous avons relié ces dimensions à une table de fait FAIT_Soc_Emp qui contient les mesures à analyser

soit le nombre d'avis, le nombre d'emplois et le nombre d'entreprises. Enfin, nous avons rajouté des hiérarchies H_departement, H_region et H_pays pour les tables DIM_Soc et DIM_Emp. Ces hiérarchies nous permettent de faire des analyses plus précises concernant les lieux des emplois et des entreprises.

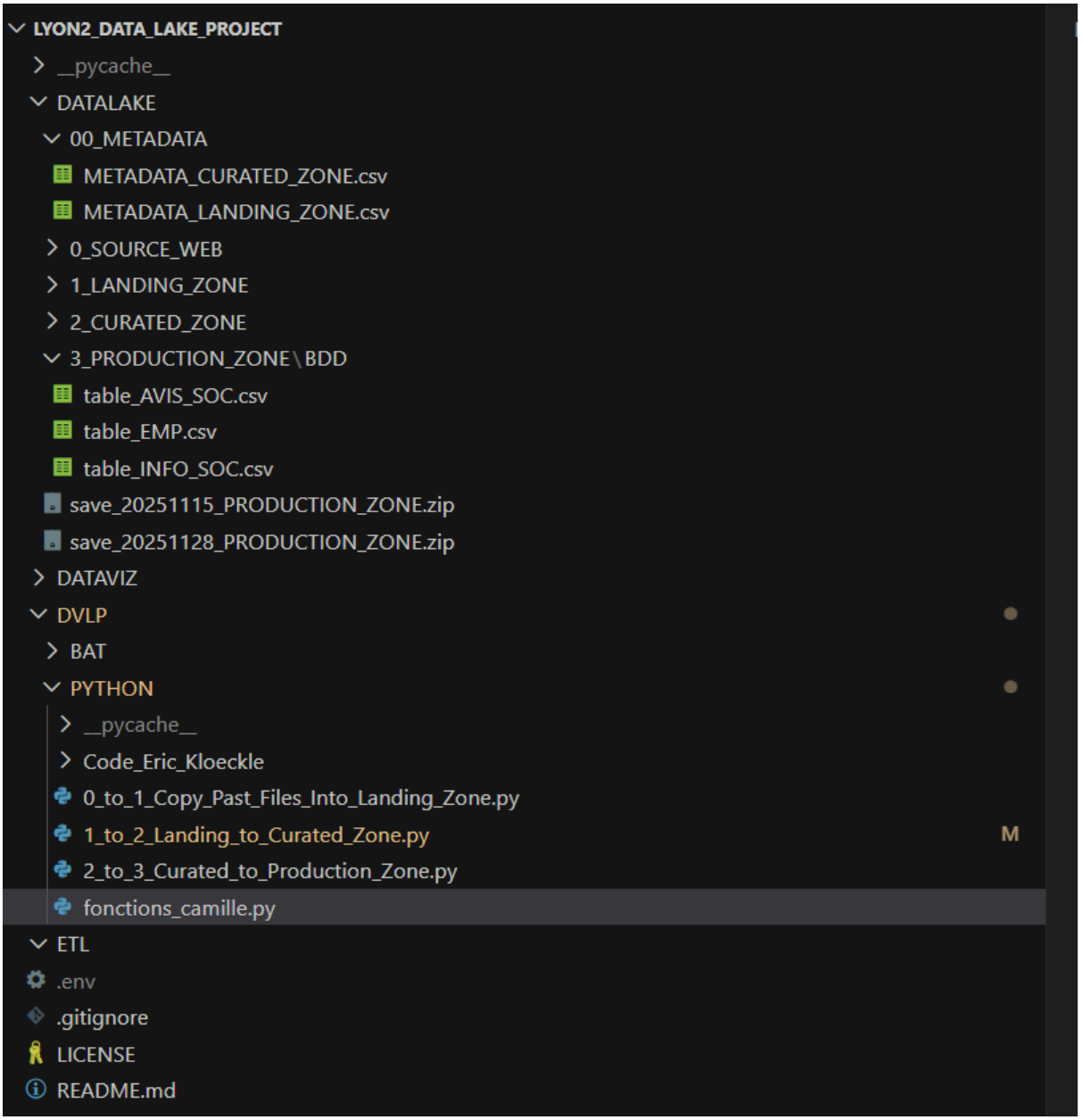


Capture 5: Modèle de données

Pour plus de clarté et de lisibilité, le modèle est disponible sur le drive sous un format .drawio. Ce modèle est à titre indicatif et peut donc être différent du modèle final.

2. ETL

Pour commencer sur cette partie, il faut savoir que toute la partie ETL s'est faite grâce à Python. Nous nous sommes basés sur l'arborescence indiquée par M Eric Kloeckle, en y apportant quelques modifications:



Capture 6: Arborescence du projet pendant le développement

De bas en haut, nous avons d'abord des fichiers de configuration pour le repository github: https://github.com/CamilleLV/Lyon2_Data_lake_Project

Nous avons fait le choix de ce site pour historiser notre projet. Le fichier “.env” contient notamment les codes API pour le développement de l’algorithme avec Gemini, ces codes étant des données sensibles, ce fichier n’est jamais envoyé sur le site de github.

Le dossier ETL est vide, puisque tout a été fait dans la partie DVLP, nous avons gardé ce dossier pour la forme, afin que l’arborescence corresponde toujours à celle donnée par notre professeur.

Dans le dossier DVLP, nous avons un dossier BAT dans lequel se situe le fichier workflow_datalake.bat, qui permet d'exécuter automatiquement nos différents scripts d’extraction, transformation et de chargement des données.

```
@echo off
ECHO Lancement du script 1...
"C:/Users/camil/AppData/Local/Programs/Python/Python310/python.exe" "C:/Users/camil/Cours/Lyon 2/Données massives/Lyon2_Data_lake_Project/0_to_1_Copy_Past_Files_Into_Landing_Zone.py"
    ECHO Script 1 termine avec succes. Lancement du script 2...
    "C:/Users/camil/AppData/Local/Programs/Python/Python310/python.exe" "C:/Users/camil/Cours/Lyon 2/Données massives/Lyon2_Data_lake_Project/1_to_2_Landing_to_Curated_Zone.py" && (
        ECHO Script 2 termine avec succes. Lancement du script 3...
        "C:/Users/camil/AppData/Local/Programs/Python/Python310/python.exe" "C:/Users/camil/Cours/Lyon 2/Données massives/Lyon2_Data_lake_Project/2_to_3_Curated_to_Production_Zone.py
            ECHO Script 3 termine avec succes. Tous les scripts sont termines.
        )
    )
)

REM Cette ligne capture le code d'erreur final (0 si tout a reussi)
exit /b %errorlevel%
```

Capture 7: Texte du fichier workflow_datalake.bat

Dans le dossier PYTHON, on retrouve le dossier avec les scripts donnés par notre professeur, le fichier fonction_camille.py contenant les fonctions déjà créées par Camille dans de précédents fichiers qui pourront être importées dans nos différents scripts, ainsi que 3 fichiers qui représentent les 3 grandes étapes de la partie ETL:

- 0_to_1_Copy_Past_Files_Into_Landing_Zone.py
- 1_to_2_Landing_to_Curated_Zone.py
- 2_to_3_Curated_to_Production_Zone.py.

2.1. Copy Past Files Into Landing Zone

Ce fichier contient l’algorithme permettant de copier les fichiers (et de vérifier qu’ils ont bien été copiés) du dossier 0_SOURCE_WEB vers 1_LANDING_ZONE, tout en catégorisant les fichiers pour les placer dans le bon sous-dossier, et en générant un fichier de métadonnées permettant de suivre les actions de déplacement de nos données.

2.2. Landing to Curated Zone

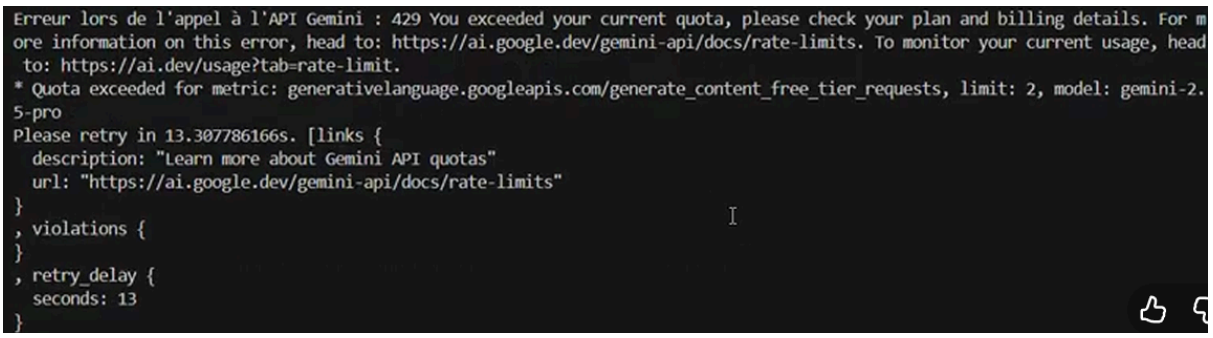
C'est ce fichier le corps du projet. Depuis la landing zone, chaque document présent dans les sous-dossier possède des informations que l'on va extraire afin de créer notre base de données, avec uniquement des données qui nous intéressent et de qualité. Plusieurs fonctions de récupération d'information dans les documents nous ont été fournies par notre professeur, mais nous en avons rajoutées quelques-unes qui nous semblent pertinentes.

2.3. Utilisation de Gemini pour renforcer nos données

Une des problématiques majeures que nous avons rencontrées était l'impossibilité de récupérer les informations clés de la description d'un emploi. De plus, certaines entreprises font appel à un cabinet de recrutement donc sur certaines offres d'emploi, c'est le nom du cabinet qui est à la place de celui de l'entreprise, cela peut contrarier la recherche d'emploi lors de la visualisation des données à posteriori.

Les descriptions contiennent généralement des informations très pertinentes comme le niveau d'étude demandé, les technologies et soft skills requis, etc. C'est pour cela que l'on a implémenté un module permettant d'effectuer une requête API à Gemini, l'IA générative de Google. On récupère la description et on lui écrit un prompt indiquant de récupérer plusieurs informations, ou ne rien retourner si l'IA ne trouve rien.

Cela nous a permis de renforcer nos données initiales, ajoutant des informations supplémentaires lors de la recherche d'emploi dans le rapport Power BI. Malheureusement dans le contexte de ce projet étudiant, nous n'avons pas beaucoup de tokens, ce sont des points d'utilisation de l'IA, donc certains emplois n'ont pas pu être traités. Nous souhaitions quand même mettre en avant cette solution, qui fonctionne, mais pas dans un modèle gratuit tel que ce projet étudiant.



Capture 8: Message d'erreur de Gemini

Après récupération de toutes ces données dans les documents, nous avons créé les métadonnées descriptives dans le fichier METADATA_CURATED_ZONE.csv. Dans le contexte de notre projet, nous n'utilisons pas le dossier 2_CURATED_ZONE, puisque nous avons décidé de centraliser nos métadonnées dans un dossier fraîchement créé : 00_METADATA. Nous avons laissé le dossier 2_CURATED_ZONE pour la forme, mais nous ne l'utilisons pas, on a créé un petit fichier dedans pour que le dossier soit visible dans Github.

Pour ce projet, puisque nous utilisons des fichiers et non des tables, nous effectuons une suppression des lignes doublons après avoir ajouté toutes nos lignes. Dans un contexte professionnel, une mise à jour correctement exécutée serait plus adéquate.

2.4. Curated to Production Zone

Enfin, nous souhaitons, pour la visualisation des données, posséder plusieurs tables pour les différents types de fichiers (INFO_SOC, AVIS_SOC, EMPLOIS). Nous avons donc créé des ID lors de la création des métadonnées pour que l'on puisse récupérer plusieurs lignes et les transformer en plusieurs champs d'une ligne "type BDD". Ce dernier script Python pivote donc les données et crée les trois fichiers finaux : table_AVIS_SOC.csv, table_EMP.csv, table_INFO_SOC.csv

```
PS C:\Users\camil\Cours\Lyon 2\Données massives\Lyon2_Data_lake_Project> & C:/Users/camil/AppData
/Local/Programs/Python/Python310/python.exe "c:/Users/camil/Cours/Lyon 2/Données massives/Lyon2_D
ata_lake_Project/DVLP/PYTHON/2_to_3_Curated_to_Production_Zone.py"
Lecture et pivotement du fichier : ./DATALAKE/00_METADATA/METADATA_CURATED_ZONE.csv...
Lecture et pivotement terminés.
Écriture des fichiers de sortie...
-> Écriture de ./DATALAKE/3_PRODUCTION_ZONE/BDD\table_AVIS_SOC.csv (209 lignes)
-> Écriture de ./DATALAKE/3_PRODUCTION_ZONE/BDD\table_INFO_SOC.csv (140 lignes)
-> Écriture de ./DATALAKE/3_PRODUCTION_ZONE/BDD\table_EMP.csv (254 lignes)
Traitement terminé !
```

Capture 9: Vue du terminal lors de l'exécution du script

Cette partie du développement python récupère donc nos fichiers depuis une source web (ici simplement un dossier), les copies dans un dossier "Staging", récupère les informations majeures et génère des tables pour la visualisation des données, renforcées par l'IA générative de Google : Gemini.

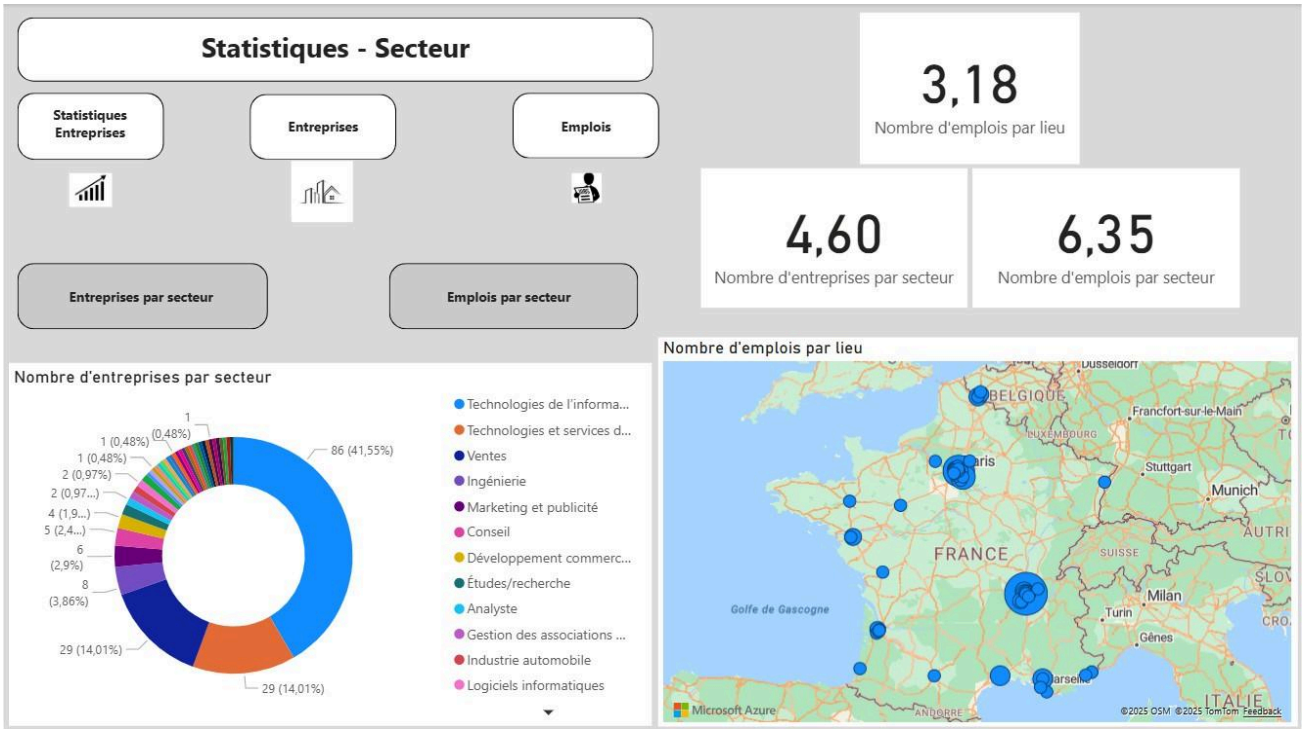
3. Datavisualisation

Une fois les données récupérées via l'ETL, nous pouvons passer à la phase de datavisualisation sur Power BI.

Ce premier écran a pour objectif d'afficher des statistiques avec une précision sur les secteurs. Dans un premier temps, nous avons réalisé deux donuts à propos du nombre d'emplois par secteur et du nombre d'entreprises par secteur. Chaque graphique sera affiché en appuyant sur le bouton correspondant.

Dans un second temps, nous avons créé une carte "map" qui représente le nombre d'emplois par ville.

Dans un dernier temps, nous avons produit des cartes représentant la moyenne des mesures utilisées dans les graphiques précédents, c'est-à-dire le nombre d'emplois par lieu, le nombre d'entreprises par secteur et le nombre d'emplois par secteur.



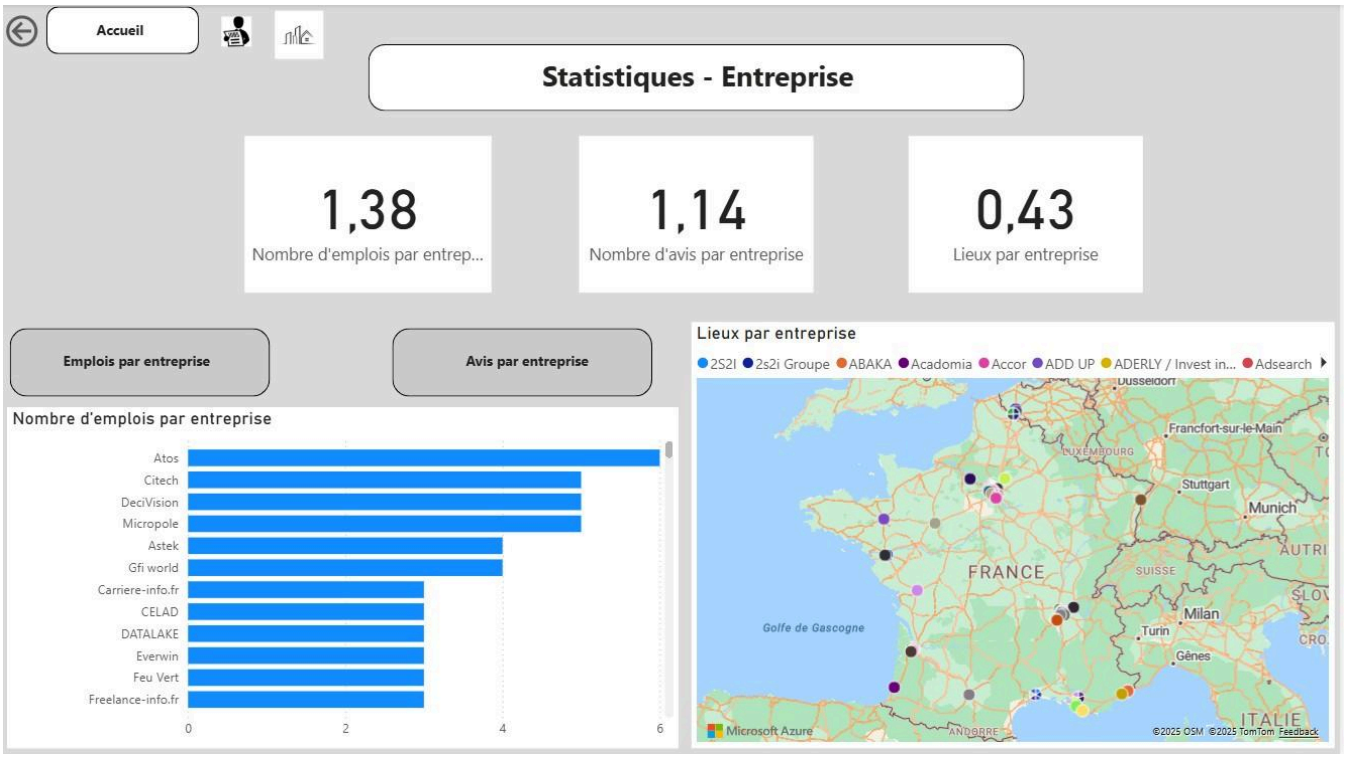
Capture 10: Écran 1 - Statistiques - secteur

Ce deuxième écran a pour but d'afficher des statistiques avec une précision sur les entreprises.

Premièrement, nous avons utilisé deux diagrammes à barres horizontales à propos du nombre d'emplois par entreprise et du nombre d'avis par entreprise. Chaque graphique sera affiché en appuyant sur le bouton correspondant.

Deuxièmement, nous avons créé une carte "map" qui représente les lieux par entreprise.

Dernièrement, nous avons produit des cartes représentant la moyenne des mesures utilisées dans les graphiques de cet écran, c'est-à-dire le nombre d'emplois par entreprise, le nombre d'avis par entreprise et les lieux par entreprise.

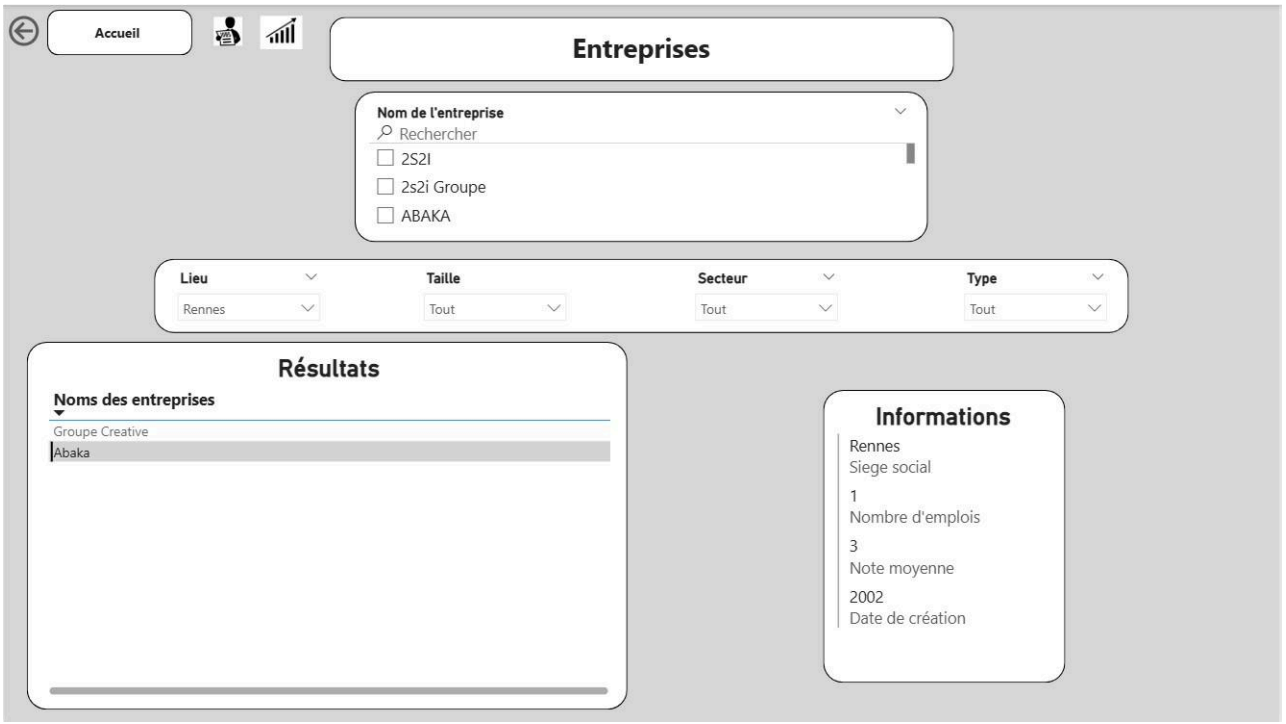


Capture 11: Écran 2 - Statistiques - entreprise

Ce troisième écran est nécessaire pour chercher des informations sur une entreprise. Tout d'abord, nous avons un filtre direct sur le nom de l'entreprise. Ensuite, si l'on souhaite chercher autrement qu'avec le nom, il est possible de filtrer les entreprises grâce aux filtres lieu, taille, secteur et type.

Les résultats sont ensuite affichés dans le tableau “résultats”. Nous pouvons ensuite regarder les informations complémentaires spécifiques à une entreprise dans le tableau “informations”.

Dans notre cas, nous souhaitons rechercher une entreprise à Rennes. Nous avons donc filtré le lieu sur Rennes et deux résultats sont sortis. Afin d’avoir des informations plus précises, nous avons sélectionné Abaka.



Capture 12: Écran 3 - Entreprises

Ce quatrième écran est nécessaire pour chercher des informations sur une offre. Tout d’abord, nous avons un filtre direct sur le nom de l’offre et un autre sur le nom de l’entreprise. Ensuite, si l’on souhaite chercher autrement qu’avec le nom, il est possible de filtrer les offres grâce aux filtres lieu, secteur, expérience et formation.

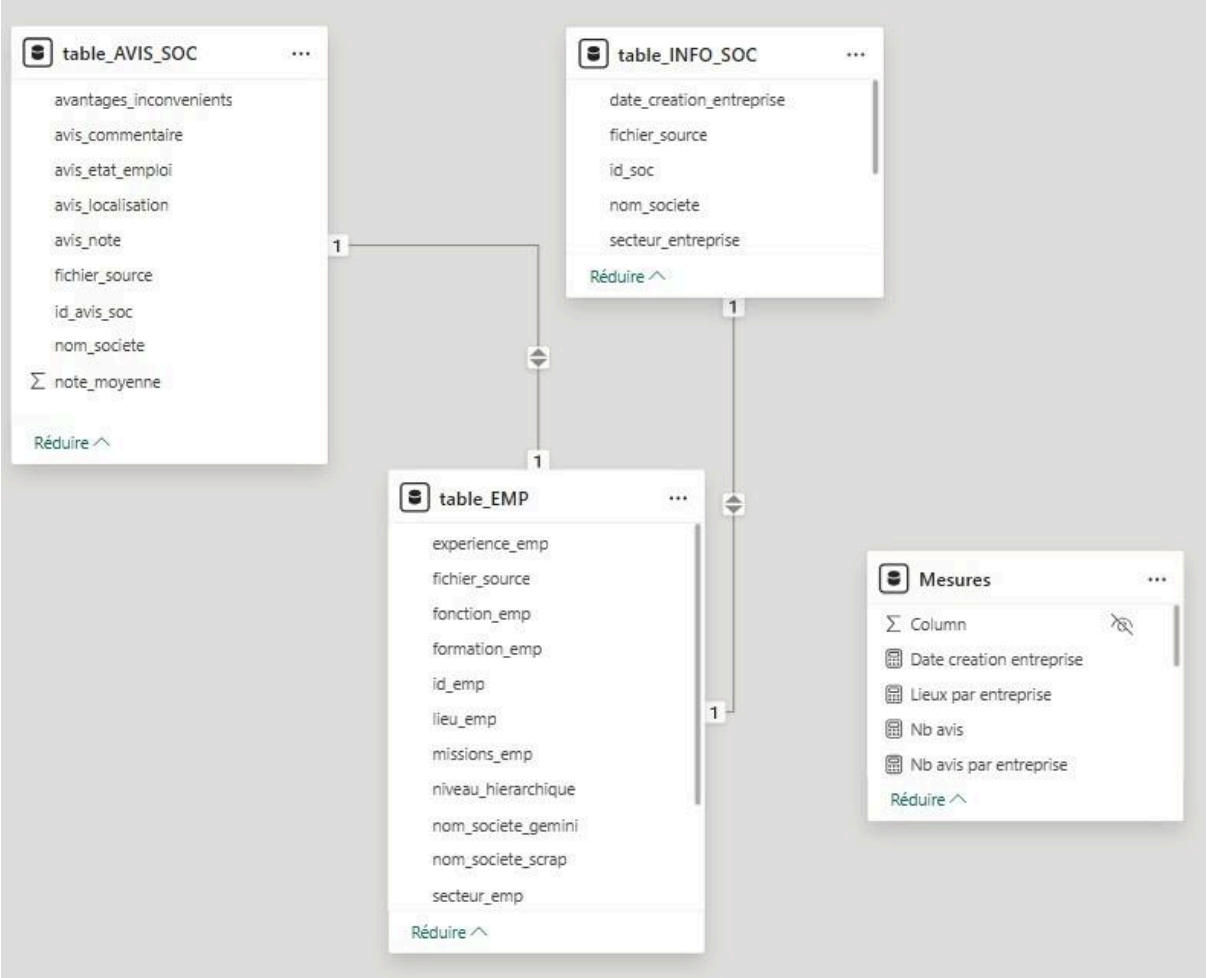
Les résultats sont ensuite affichés dans le tableau “résultats”. Nous pouvons ensuite regarder les informations complémentaires spécifiques à une offre dans le tableau “informations”.

Dans notre cas, nous souhaitons rechercher une offre dans le secteur des assurances. Nous avons donc filtré le secteur sur assurances et un résultat est sorti. Afin d'avoir des informations plus précises, nous avons sélectionné l'offre.

Capture 13: Écran 4 - Emplois

À l'origine, nous avions choisi de réaliser trois écrans différents: un pour les statistiques, un pour les entreprises et un pour les emplois. Finalement, nous en avons créé quatre. Ceux des entreprises et des emplois sont identiques à la maquette. Cependant, celui des statistiques est différent. En effet, nous avions prévu de nombreux graphiques et, pour des raisons de lisibilité et de dynamisme, les séparer sur deux écrans distincts était la meilleure solution.

Concernant le modèle, il est aussi différent de celui de départ. Nous avons pris le parti de ne pas créer de table de fait mais de créer une table de mesures directement dans Power BI. Enfin, compte tenu de la disparité des données, nous n'avons pas créé les hiérarchies qui nous semblaient obsolètes.



Capture 14: Modèle dans Power BI

4. Conclusions personnelles

Charlène Broutier :
Pour conclure personnellement, j'ai beaucoup appris sur le principe des zones qui permettent de séparer les données en prenant en compte leur phase de traitement. J'ai également appris sur Power BI. En effet, j'ai pour habitude d'utiliser le logiciel et les graphiques pour faire des analyses et des statistiques comme les deux premiers écrans. Je n'avais jamais pensé à les utiliser comme filtres et éléments de recherche.

Kawtar Tarhami :
J'ai appris qu'on peut concevoir une maquette à partir des données que l'on prévoit de récupérer, en imaginant les statistiques et analyses souhaitées. Cependant, une fois les données réellement collectées, on se rend souvent compte que d'autres indicateurs, parfois plus pertinents et mieux adaptés à la réalité des données disponibles, peuvent être mis en place.

Camille Laverie :
Ce projet m'a permis de renforcer mes acquis en Python, on a pu aussi faire de la gestion de projet puisque celui-ci était tout de même imposant et nous a demandé une certaine rigueur dans les échanges de travail entre camarades. J'ai aussi appris à requêter une IA générative, je suis content du travail rendu, la datavisualisation est vraiment pertinente. Le projet était très complet et le sujet très clair.

Conclusion

Ce projet a été une expérience complète et enrichissante, menant à la création d'un rapport Power BI interactif pour l'analyse des données d'entreprises et d'offres d'emploi. Le cycle de vie du projet s'est articulé autour de trois phases majeures.

La première phase contenait la définition des besoins via des maquettes de datavisualisation, la structuration des données avec un Dictionnaire de données, et l'élaboration d'un modèle de données.

La seconde phase de ce projet contenait l'ETL (Extraction, Transformation, Chargement), entièrement réalisé en Python. Il a permis d'acheminer les données brutes à travers les différentes zones (Landing Zone, Curated Zone, Production Zone). Un point fort a été l'intégration de Gemini, l'IA générative de Google, pour enrichir les descriptions d'emplois avec des informations clés (technologies, soft skills), prouvant la pertinence des outils d'IA pour l'amélioration de la qualité des données, malgré la contrainte de tokens rencontrée.

Enfin, la dernière phase a abouti à la création d'un rapport Power BI dynamique, composé de pages dédiées aux entreprises, aux statistiques sectorielles et à la recherche ciblée d'informations. Ce tableau de bord offre une lecture claire et exploitable du marché de l'emploi, tout en apportant des informations contextuelles essentielles sur les entreprises.

Au cours de ce projet, nous avons rencontré plusieurs difficultés. D'abord pour l'ajout du renforcement des données par Gemini, une contrainte due au nombre de tokens utilisateurs nous a empêché d'utiliser Gemini sur l'ensemble des données. Nous avons une version gratuite du modèle de développeur donc nous avons simplement accepté, pour ce projet, de ne renforcer qu'une partie des données. Ensuite, le modèle de données final ne correspond pas exactement au modèle imaginé et souhaité lors des prémices du projet. Ce n'est pas une réelle problématique, mais cela nous a beaucoup questionné tout au long du projet. Enfin, le projet nous a demandé de l'investissement, prenant une majorité du peu de temps hors travail/école où l'on pouvait travailler à plusieurs, cela nous a demandé une vraie rigueur dans le travail.

En conclusion, malgré les défis techniques, le projet a atteint son objectif. Il a permis de transformer des documents en une source d'information structurée et analysable, offrant une meilleure compréhension des tendances du marché de l'emploi et des entreprises grâce à des visualisations claires et efficaces.

Table des illustrations