Projet
DATA ENGINEERING ON CLOUD

Présenté par :

Rayane Besrour
Noémie Ktourza
Arthur Manceau
Florian Germain

Table des matières

I.Présentation du sujet

1. Contexte

Dans un monde professionnel en constante évolution, la gestion des ressources humaines est devenue un levier stratégique essentiel pour les entreprises. Les dynamiques du marché du travail, influencées par des facteurs économiques, sociaux et technologiques, poussent les organisations à adapter leurs pratiques pour répondre aux attentes des salariés tout en restant compétitives.

Analyser la satisfaction au travail et les opportunités professionnelles des employés est devenu un enjeu majeur. Ces deux dimensions sont étroitement liées à des variables clés telles que le niveau d'étude, l'expérience professionnelle, la mobilité géographique et l'influence familiale. Ces facteurs jouent un rôle déterminant dans les trajectoires professionnelles, le bien-être des salariés et leur capacité à saisir des opportunités dans un environnement en perpétuel changement.

Dans ce contexte, l'exploitation des informations permettra de mieux comprendre les dynamiques de l'entreprise et de proposer des solutions aux équipes RH visant à

2. Objectif

L'objectif est d'étudier l'impact du niveau d'étude, de l'expérience, de la mobilité géographique et de l'influence familiale sur la satisfaction au travail et les opportunités professionnelles, en identifiant les leviers d'action prioritaires.

3. Dataset

Dans notre solution nous allons utiliser 3 fichiers csv qui seront nos sources de données :

- **Carrer_change_dynamique.csv**

Ce fichier contient toutes les informations liées à la carrière professionnelle des employés, et leurs évolutions de carrière.

- **Demography_Education.csv**

Ce fichier contient les informations démographiques et educatives des employés.

- **Job_satisfaction.csv**

Ce fichier contient la satisfaction des employés au sein de leur entreprise

Types de données :

1. **carrer_change_dynamique.csv**

o Type de données : carrière et évolution des employés
o Format : CSV
o Colonnes :
- Career Change Interest : Intérêt pour un changement de carrière.
- Skills Gap : Niveau de lacunes en compétences.
- Mentorship Available : Disponibilité du mentorat.
- Certifications : Nombre de certifications obtenues.
- Geographic Mobility : Mobilité géographique.
- Professional Networks : Accès aux réseaux professionnels.
- Career Change Events : Nombre d'événements favorisant un changement de carrière.
- Technology Adoption : Adoption des technologies modernes.
- Likely to Change Occupation : Probabilité de changement d'emploi.

2. **Demography_Education.csv**

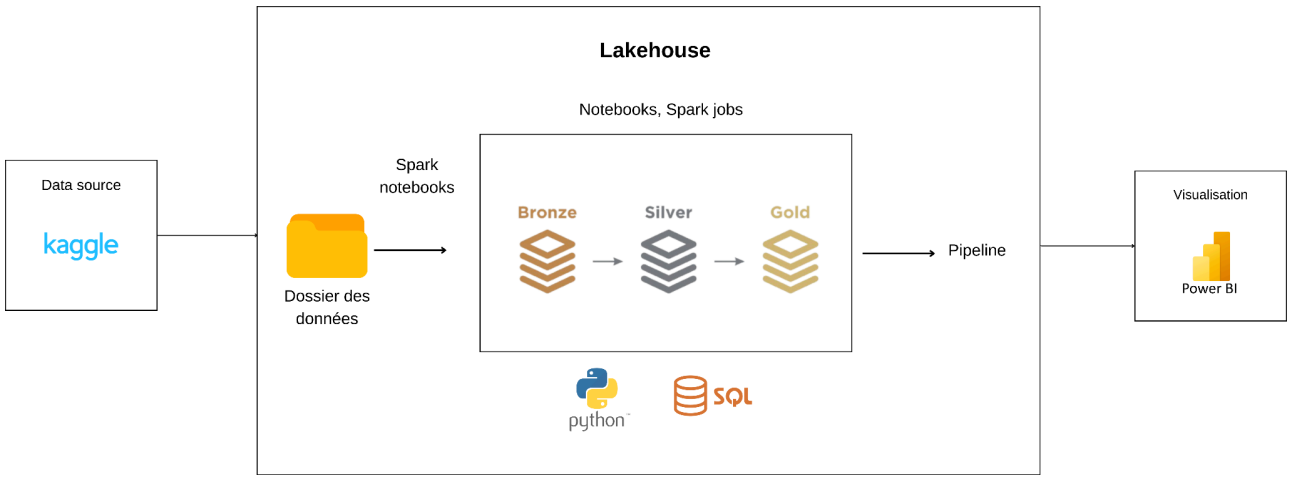
o Type de données : données démographiques des employés
o Format : CSV
o Colonnes :
- Field of Study : Domaine d'étude (ex. : médecine, éducation, arts).
- Age : Âge de l'employé.
- Gender : genre de l'employé.
- Education Level : Niveau d'éducation (lycée, licence, master, doctorat).
- Years of Experience : Nombre d'années d'expérience professionnelle.
- Family Influence : Influence familiale sur la carrière (faible, moyenne, élevée).

3. **Job_satisfaction.csv**

o Type de données : satisfaction des employés
o Format : CSV
o Colonnes :
- Current Occupation : Poste occupé actuellement.
- Industry Growth Rate : Taux de croissance de l'industrie (faible, moyen, élevé).
- Job Satisfaction : Niveau de satisfaction au travail (échelle de 1 à 10).

- Work-Life Balance : Équilibre entre vie professionnelle et personnelle (échelle de 1 à 10).
- Job Opportunities : Nombre d'opportunités disponibles (en pourcentage).
- Salary : Salaire annuel (en dollars).
- Job Security : Sécurité de l'emploi (échelle de 1 à 10).
- Freelancing Experience : Expérience de travail en tant qu'indépendant

II. Architecture du système



Les fichiers sont d'abord importés dans un espace dédié au sein du Lakehouse que nous configurons. Ainsi, les données brutes seront stockées dans le "Bronze Layer", une zone où les données sont simplement importées telles qu'elles sont reçues sans avoir effectué de transformation ou de nettoyage au préalable.
En effet, cette étape est très importante pour assurer une traçabilité complète des données d'origines afin d'avoir la possibilité de revenir à l'état brute des données en cas de problème.

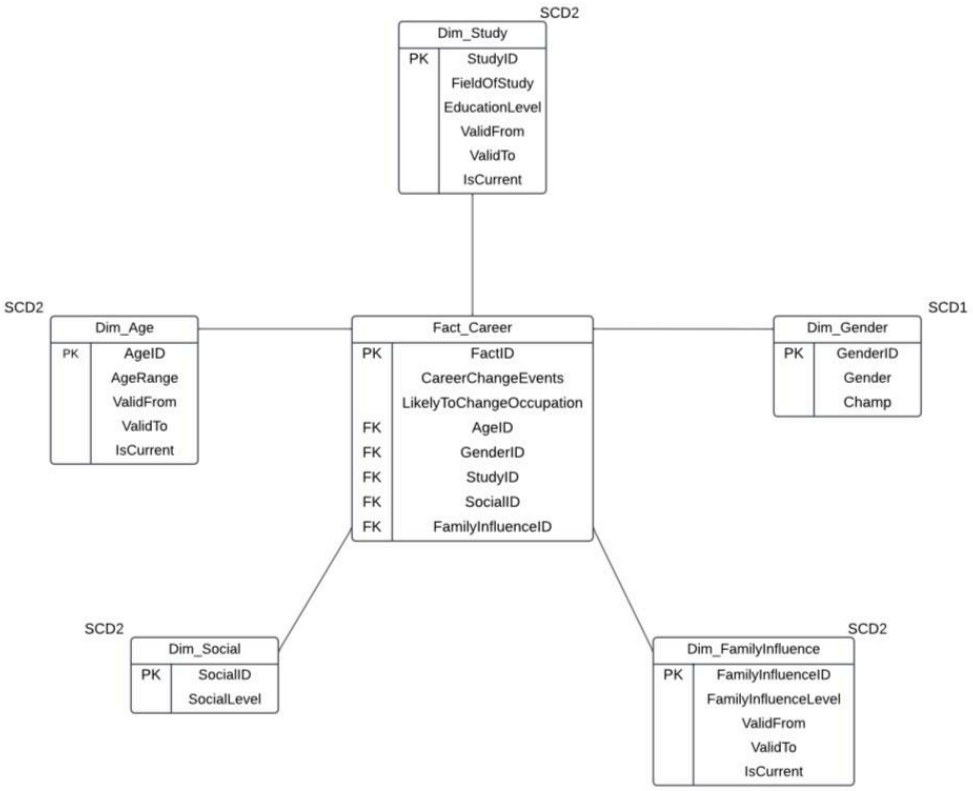
Une fois les données brutes collectées, elles passeront au "Silver Layer" où elles seront nettoyées et partiellement transformées comme supprimer des doublons.
Cette étape nous permet d'avoir des données prêtes à l'analyse tout en ayant une structure intermédiaire entre les données brutes et la visualisation finale.

Enfin, nous allons les structurer dans le "Gold Layer" de manière à simplifier l'analyse. Pour cela, nous utilisons un modèle dimensionnel, un modèle étoile, comprenant une table de fait et plusieurs tables de dimensions.

Ces données seront alors transformées pour les uniformiser et rendre les données exploitables.

Par la suite, les données seront connectées à Power BI pour offrir une vue d'ensemble sur nos données

III.MCD



IV.Traitements effectués

Lien GitHub : https://github.com/armanceau/RH-Data

Récupération des fichiers depuis github

La gestion des données brutes

1. Le passage de Bronze à Silver

Le fichier est chargé dans le DataFrame Pandas, une structure de données pour le traitement. Une fois que les données sont chargées, la première étape de nettoyage consiste à supprimer toutes les lignes avec des valeurs manquantes. Cette opération est importante pour garantir que les données qui seront utilisées sont complètes et ne risquent pas de fausser les résultats.

```
bronze_files = os.listdir(BRONZE_PATH)

for file in bronze_files:
    if file.endswith(".csv"):
        print(f"Traitement du fichier : {file}")

        df = pd.read_csv(f"{BRONZE_PATH}/{file}")

        # On supprime les lignes avec des valeurs manquantes
        df_cleaned = df.dropna()
```

Nous convertissons par la suite la colonne ‘Age’ en valeurs numériques.
Cette conversion permet de garantir que seules les valeurs valides sont prises en compte.
Après le nettoyage, les fichiers sont sauvegardés dans le répertoire Silver

```
if 'Age' in df_cleaned.columns:
    df_cleaned['Age'] = pd.to_numeric(df_cleaned['Age'], errors='coerce')

# On sauvegarde dans ds-silver
df_cleaned.to_csv(f"{SILVER_PATH}/{file}", index=False)
print(f"Fichier nettoyé sauvegardé : {file}")
```

Les trois fichiers sont alors transformés et sauvegardés dans leurs répertoires **Silver** respectifs, avec un passage en **Delta Lake** :

Passage des fichiers Silver en format Delta

Dans cette étape, nous avons converti les fichiers Silver au format Delta afin de bénéficier des avantages de ce dernier, notamment en termes de performance, de fiabilité et de gestion des métadonnées. Voici les détails pour chaque fichier traité :

1. Carrer_change_dynamique.csv

Le fichier **Carrer_change_dynamique.csv** a été transformé en format Delta. Il contient des informations liées aux changements de carrière, aux compétences, au mentorat et à d'autres aspects professionnels.

- **Schéma des données** :
  Career Change Interest; Skills Gap; Mentorship Available; Certifications; Geographic Mobility; Professional Networks; Career Change Events; Technology Adoption; Likely to Change Occupation.
- **Statistiques** :
  - Nombre total de lignes : **38 444**.
  - Valeurs minimales : 0;10;0;0;0;10;0;10;0.
  - Valeurs maximales : 1;9;1;1;1;9;2;2;1.
- **Format initial** : Parquet.
- **Format final** : Delta.

2. Demography_Education.csv

Le fichier *demography.csv* a également été converti en Delta. Il contient des données démographiques telles que l'âge, le genre, le niveau d'éducation et l'expérience professionnelle.

- **Schéma des données** :
  Field of Study; Age; Gender; Education Level; Years of Experience; Family Influence.
- **Statistiques** :
  - Nombre total de lignes : **38 444**.
  - Valeurs minimales : Arts;20;Female;Bachelor's;11;Low.
  - Valeurs maximales : Psychology;59;Male;PhD;38;None.
- **Format initial** : Parquet.
- **Format final** : Delta.

3. Job_satisfaction.csv

Enfin, le fichier *job_satisfaction.csv* a été transformé en Delta. Ce fichier inclut des informations sur la satisfaction au travail, l'équilibre entre vie professionnelle et personnelle, les opportunités d'emploi, et d'autres indicateurs clés.

- **Schéma des données** :
  Current Occupation; Industry Growth Rate; Job Satisfaction; Work-Life Balance; Job Opportunities; Salary; Job Security; Freelancing Experience.
- **Statistiques** :
  - Nombre total de lignes : **38 444**.
  - Valeurs minimales : Artist;High;10;10;1;52108;3;0.
  - Valeurs maximales : Teacher;Medium;9;9;96;88977;4;0.
- **Format initial** : Parquet.
- **Format final** : Delta.

Cette conversion permet d'assurer une meilleure intégration des données dans l'étape suivante, où les informations seront dénormalisées et consolidées dans une table de faits.

2. Le passage de Silver à Gold

Tout d’abord, nous commençons par catégoriser l’âge afin que l’analyse des données soit plus simple et plus compréhensible. Les différentes catégories sont alors : "Jeune", "Adulte", "Senior", et "Très Senior"

```python
# 1. Catégorie age
if 'Age' in df.columns:
    df['Age Group'] = pd.cut(
        df['Age'],
        bins=[0, 25, 35, 50, np.inf],
        labels=['Jeune', 'Adulte', 'Senior', 'Très Senior'],
        right=False
    )
```

Ensuite, le script va créer une nouvelle colonne “Combined Satisfaction” qui combine les 2 colonnes : Job Satisfaction et Work-Life pour obtenir une mesure unique de la satisfaction générale des employés.

```python
# 2. Indice de satisfaction
if 'Job Satisfaction' in df.columns and 'Work-Life Balance' in df.columns:
    df['Combined Satisfaction'] = (
        df['Job Satisfaction'] * 0.6 + df['Work-Life Balance'] * 0.4
    )
```

Nous avons aussi regrouper les niveaux d’éducations avec les catégories "Basique", "Intermédiaire", "Avancé", "Très Avancé"

```python
# 3. Niveau éducation
if 'Education Level' in df.columns:
    df['Education Category'] = df['Education Level'].map({
        'High School': 'Basique',
        'Bachelor': 'Intermédiaire',
        'Master': 'Avancé',
        'PhD': 'Très Avancé'
    }).fillna('Inconnu')
```

Enfin, nous calculons la probabilité moyenne de changement de carrière en prenant la moyenne des valeurs de la colonne “Career Change”

```
# 4. Probabilités de changement carrière
if 'Career Change' in df.columns:
    prob_career_change = df['Career Change'].mean()
    df['P(Career Change)'] = prob_career_change

df.to_csv(f"{GOLD_PATH}/{file}", index=False)
print(f"Fichier enrichi sauvegardé : {file}")
```

Passage de gold à delta : transformation en table de fait

Les données Gold sont dénormalisées et intégrées dans une **table de fait** appelée **fact_career**. Cette table est optimisée pour faciliter les analyses multidimensionnelles. Voici sa structure :

1. **Fait principal** : Changement de carrière (provenant du fichier **career**).
2. **Dimensions ajoutées** :
   - **Dimensions d’âge** : Catégories d’âge ("Jeune", "Adulte", "Senior", "Très Senior").
   - **Dimensions de genre** : Données issues de Gender dans **demography**.
   - **Dimensions d’étude** : Catégories d’éducation regroupées ("Basique", "Intermédiaire", etc.).
   - **Dimensions sociales** : Équilibre travail-vie personnelle et satisfaction professionnelle (issues de **job_satisfaction**).
   - **Dimension d’influence de la famille** : Données provenant de Family Influence dans **demography**.

La table **fact_career** offre une vision consolidée des facteurs influençant les décisions de **changement de carrière**.

Les 3 fichiers csv sont alors transformé au format delta dans des dossiers dédiés

1. **Carrer_change_dynamique.csv**

Le fichier **Carrer_change_dynamique.csv** a été converti avec les spécifications suivantes :

- **Schéma des données** :
- **Statistiques** :
    - **Nombre d'enregistrements** : 38 444.
    - **Valeurs minimales** : 0;10;0;0;0;10;0;10;0.
    - **Valeurs maximales** : 1;9;1;1;1;9;2;2;1.
- **Format Delta** : Parquet compressé (Snappy).
- **Fichier de sortie** : part-00001-dc82573f...snappy.parquet.

2. Demography_Education.csv

Le fichier **Demography_Education.csv** a été converti avec les spécifications suivantes :

- **Schéma des données** :
- **Statistiques** :
    - **Nombre d'enregistrements** : 38 444.
    - **Valeurs minimales** : Arts;20;Female;Bachelor's;11;Low.
    - **Valeurs maximales** : Psychology;59;Male;PhD;38;None.
- **Format Delta** : Parquet compressé (Snappy).
- **Fichier de sortie** : part-00001-c2c78d53...snappy.parquet.

3. Job_satisfaction.csv

Le fichier **Job_satisfaction.csv** a été transformé avec les spécifications suivantes :

- **Schéma des données** :
- **Statistiques** :
    - **Nombre d'enregistrements** : 38 444.
    - **Valeurs minimales** : Artist;High;10;10;1;52108;3;0.
    - **Valeurs maximales** : Teacher;Medium;9;9;96;88977;4;0.
- **Format Delta** : Parquet compressé (Snappy).
- **Fichier de sortie** : part-00001-efdd9c9e...snappy.parquet.

Dénormalisation et Consolidation :

Une fois les fichiers convertis en **format Delta**, l'étape suivante consiste à dénormaliser les données et les consolider dans une **table de faits unique**. Voici le processus :

1. **Jointure des trois tables Delta** sur une clé commune (exemple : **ID employé** si disponible ou une clé arbitraire).

2. **Agrégation des colonnes** pertinentes provenant des trois tables pour avoir une vue complète.

3. **Création de la table de faits** finale qui inclura :

   - **Informations de carrière** (skills gap, changement d'occupation, etc.)

   - **Données démographiques** (âge, niveau d'éducation, années d'expérience, etc.)

   - **Satisfaction au travail** (équilibre vie pro/vie perso, salaire, satisfaction, etc.)

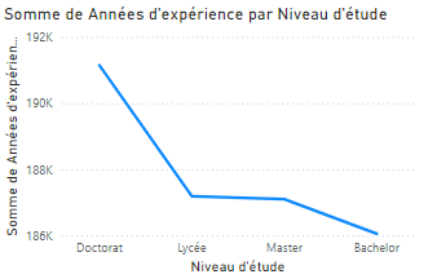
La table consolidée dans la couche **Gold** fournira une **vue 360° des employés** avec les métriques clés issues des trois fichiers initiaux. Cette table de faits sera prête pour les analyses ultérieures, incluant le reporting ou le machine learning.

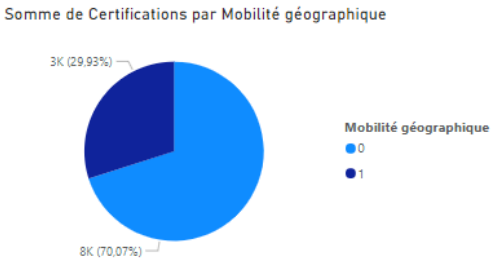
IV. Visualisation des données : rapport PowerBI

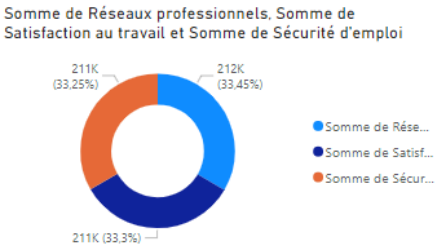
A partir des données stockées dans le LakeHouse, nous avons créé un rapport PowerBI qui vise à analyser les résultats que nous avions obtenus afin de fournir une vue d'ensemble sur ces données.
En utilisant nos visualisations, ce rapport va permettre de faciliter l'interprétation des données et de tirer des conclusions sur notre objectif, d'identifier les principaux facteurs contribuant aux changements de carrière.

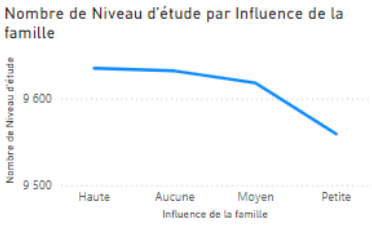
Les visualisations sont les suivantes :

- Courbe sur la somme des années d'expérience par niveau d'étude
- Diagramme circulaire pour l'impact de la certification sur les mobilités géographiques.
- Diagramme en anneaux pour l'impact de la sécurité d'emploi de la satisfaction avec le réseau professionnels.
- Courbe pour le nombre de niveau d'étude par influence de la famille
- Tableau : satisfaction au travail/somme équilibre travail-vie personnelle

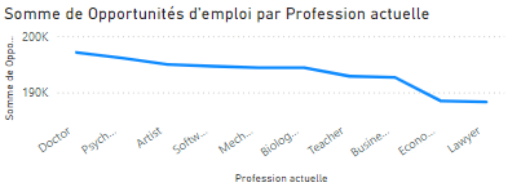








Équilibre travail-vie personnelle	Somme de Satisfaction au travail
1	21089
2	20997
3	21608
4	20539
5	20975
6	21377
7	20683
8	21265
9	21079
10	21433
Total	**211045**



L'analyse met en évidence que les niveaux d'étude élevés, comme le doctorat, favorisent une expérience professionnelle plus longue, tandis que la faible mobilité géographique (70 % des salariés non mobiles) limite souvent les opportunités. L'influence familiale peut freiner les choix éducatifs et professionnels, et l'équilibre entre satisfaction au travail, réseaux professionnels et sécurité d'emploi est essentiel pour les employés.

De plus, certaines professions offrent plus d'opportunités que d'autres, et un bon équilibre travail-vie personnelle joue un rôle clé dans la satisfaction. Les entreprises doivent donc encourager la mobilité, soutenir l'éducation continue, renforcer les réseaux professionnels et promouvoir des politiques favorisant l'équilibre entre vie professionnelle et personnelle.

Conclusion :

Pour conclure, l'éducation continue, le renforcement des réseaux professionnels et l'amélioration de l'équilibre travail-vie personnelle sont essentielles pour améliorer l'épanouissement des salariés tout en renforçant la compétitivité des entreprises.