Exercice de standardisation des données médicales au format OMOP

Contexte : Vous êtes data scientist dans le secteur de la santé et votre tâche consiste à transformer des données de santé hétérogènes en un format standardisé pour en faciliter l'analyse. Vous utiliserez le modèle OMOP (Observational Medical Outcomes Partnership) pour cette standardisation.

Scénario : Camille Honette (code NUM_ENQ = DPXX:0000000000000001X), née le 28/12/1963 et résidant à Paris dans le 14ème arrondissement, a suivi un parcours de soins pour une lombalgie. Elle a consulté son médecin généraliste (code pse_spe_cod = 1) le 4 mars 2013, qui lui a prescrit une radiographie du rachis et de l'ibuprofène 200 mg à prendre trois fois par jour pendant 5 jours. La radiographie a été réalisée à l'Hôpital Privé des Peupliers (eta_num = 75010016), et l'ibuprofène a été acheté à la Pharmacie Plaisance (eta_num = 750023772) le lendemain. Enfin, on lui a prescrite pour 10 séances de kinésithérapie (code pse_act_nat = 26), la première ayant eu lieu le 7 mars 2013.

Toutes ces prestations de santé ont généré des données qui se retrouvent dans les tables du Système National des Données de Santé (SNDS) fournies en pièce jointe.

Données fournies :

- **ir_ben_r** : Table contenant les informations des assurés.
- **er_prs_f** : Table contenant les informations sur les prestations remboursées.
- **T_mcoaaE** : Tables contenant des informations sur les établissements de santé.
- **ir_act_v et ir_spe_v** : Tables contenant des informations sur les professionnels de santé.

Tâches :

- **Table Person (Python)** :
    - Remplir les colonnes suivantes : *person_id, gender_concept_id, year_of_birth, month_of_birth, person_source_value, location_id, gender_source_value.*
    - À partir des informations fournies, générer la table Person en Python en utilisant pandas. Sauvegardez cette table sous forme de fichier CSV.

- **Table Care Site (SQL avec SQLite)** :
    - Remplir les colonnes suivantes : *cc_site_id, care_site_name, location_id, care_site_source_value.*
    - Créer et remplir la table Care Site dans une base de données SQLite.

- **Table Provider (Spark)** :

    - Remplir les colonnes suivantes : *provider_id, specialty_source_value, specialty_concept_id, provider_source_value.*
    - Utilisez Apache Spark pour générer la table Provider. Enregistrez le résultat sous forme de fichier Parquet.

Utilité de la documentation ATHENA : Il est utile d'utiliser la documentation ATHENA pour pouvoir remplir les colonnes `Person.gender_concept_id` et `Provider.specialty_concept_id`.

Critères d'évaluation :

- Qualité du code : Clarté, utilisation efficace des ressources, respect des conventions de nommage, industrialisation.
- Exactitude de la transformation des données selon le modèle OMOP.
- Une attention sera portée sur la clarté et l'explicabilité des transformations.

Ressources supplémentaires :

- **Documentation OMOP :** Vous pouvez trouver la documentation sur les identifiants de concept spécifiques aux divers types d'informations et bien d'autres détails sur le site officiel de l'OHDSI et ATHENA.
- **Documentation SNDS** : Pour en savoir plus sur le SNDS et comment l'utiliser, nous vous invitons à vous référer au **dictionnaire interactif du SNDS** et à la **documentation collaborative**.

Questions ouvertes :

- Comment gérer des transformations de données pour de la grande volumétrie ?
- Quelles sont les différentes étapes d'un projet data ?
- Quels outils utilisez-vous en plus pour que cet exercice devienne un vrai cas d'usage en entreprise ?
- Quelle méthodologie de travail serait adaptée à un projet data ?