Projet "Entrepôt de données"

Rahma GARGOURI

2023-2024

1 Objectif

A partir de fichiers sources, nous allons mettre en place un entrepôt de données qui nous permettra de répondre aux questions suivantes :

- Quel est le pourcentage de passages aux urgences pour suspicion de covid-19 par rapport au nombre de passages total, par mois et par région ?
- Quel est le pourcentage de passages aux urgences pour suspicion de covid-19 par rapport au nombre de passages total, par tranche d'âge en 2022 ?
- Quel est le pourcentage de passages aux urgences pour suspicion de covid-19 par rapport au nombre de passages total, pour les personnes agées de plus de 65 ans, en 2023 ?
- Quel est le pourcentage de passages aux urgences pour suspicion de covid-19 par rapport au nombre de passages total pour les femmes par an et par département ?
- Quel est le pourcentage de passages aux urgences pour suspicion de covid-19 par rapport au nombre de passages total pour les hommes par an et par département ?
- Quel est le rapport entre le nombre des hospitalisations des hommes et celui des femmes par jour et par région ?

2 Sources de données

Les sources de données sont les suivantes :

- *donnees-urgences-SOS-medecins.csv* représente nombre de passages aux urgences pour suspicion de COVID-19, nombre total de passages aux urgences avec un diagnostic médical renseigné, nombre d'hospitalisations parmi les passages aux urgences pour suspicion de COVID-19, nombre total d'actes médicaux SOS Médecins pour suspicion de COVID-19, nombre total d'actes médicaux SOS Médecins avec un diagnostic médical renseigné.
- *code-tranches-dage-donnees-urgences.csv* représente le mapping des tranches d'âge dans le fichier précédent.
- *departements-region.json* contient les départements français et les régions correspondantes.

Le fichier *metadonnee-urgenceshos-sosmedecin-covid19-quot.csv* contient l'explication des différents champs du fichier *donnees-urgences-SOS-medecins.csv*.

3 Travail demandé

A l'aide des fichiers sources, modélisez un entrepôt de données en utilisant un schéma en étoile, qui permet, en utilisant des requêtes simples, de répondre aux questions de la section *Objectif*.

NB : Dans l'entrepôt de données, nous n'allons pas calculer ces pourcentages et ces ratios, mais nous allons mettre en place des tables permettant, avec un calcul simple, d'obtenir une réponse à ces questions.

Y a-t-il un nettoyage à faire ?
Y a-t-il des colonnes à supprimer des fichiers sources ?
Quelle est la table des faits ? Quelles sont les tables de dimension ?
Faut-il renommer les colonnes pour être plus explicite ?

Une fois le schéma de l'entrepôt de données fixé, il faut mettre en place un DAG Airflow qui :

— extrait les données sources

— fait les transformations nécessaires

— crée les tables de l'entrepôt de données

— alimente l'entrepôt de données

D'autres fonctionnalités peuvent être ajoutées à ce DAG. Un repo Git est à partager avec un fichier *readme.md* détaillant la procédure à suivre pour lancer le DAG.

Peut-on ajouter des tests de qualité sur les données sources ?

Peut-on ajouter un *callback* qui envoie le mail en cas d'échec ou de réussite du DAG ?