



HEALTH  
DATA HUB

# Mise en situation

Poste de stagiaire au sein du pôle  
Data Innovation  
2025

Au HDH, nous proposons dans notre mode de recrutement, une mise en situation pour vous projeter dans le poste et échanger ensemble sur la manière dont vous vous y prendriez pour adresser une problématique propre au HDH.

# Mise en situation - contexte et scénario

**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:0000000000000000001X), 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.

# Mise en situation - Données fournies

---

## Données fournies :

- **ir\_ben\_r** : Table contenant les informations des assurés.
- **er\_prs\_f** : Table contenant les informations sur les prestations remboursées.
- **T\_mcoaaE** : Table contenant des informations sur les établissements de santé.
- **ir\_act\_v** et **ir\_spe\_v** : Tables contenant des informations sur les professionnels de santé.

# Mise en situation - 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*.

# Mise en situation - Critères d'évaluation, ressources

## 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 :

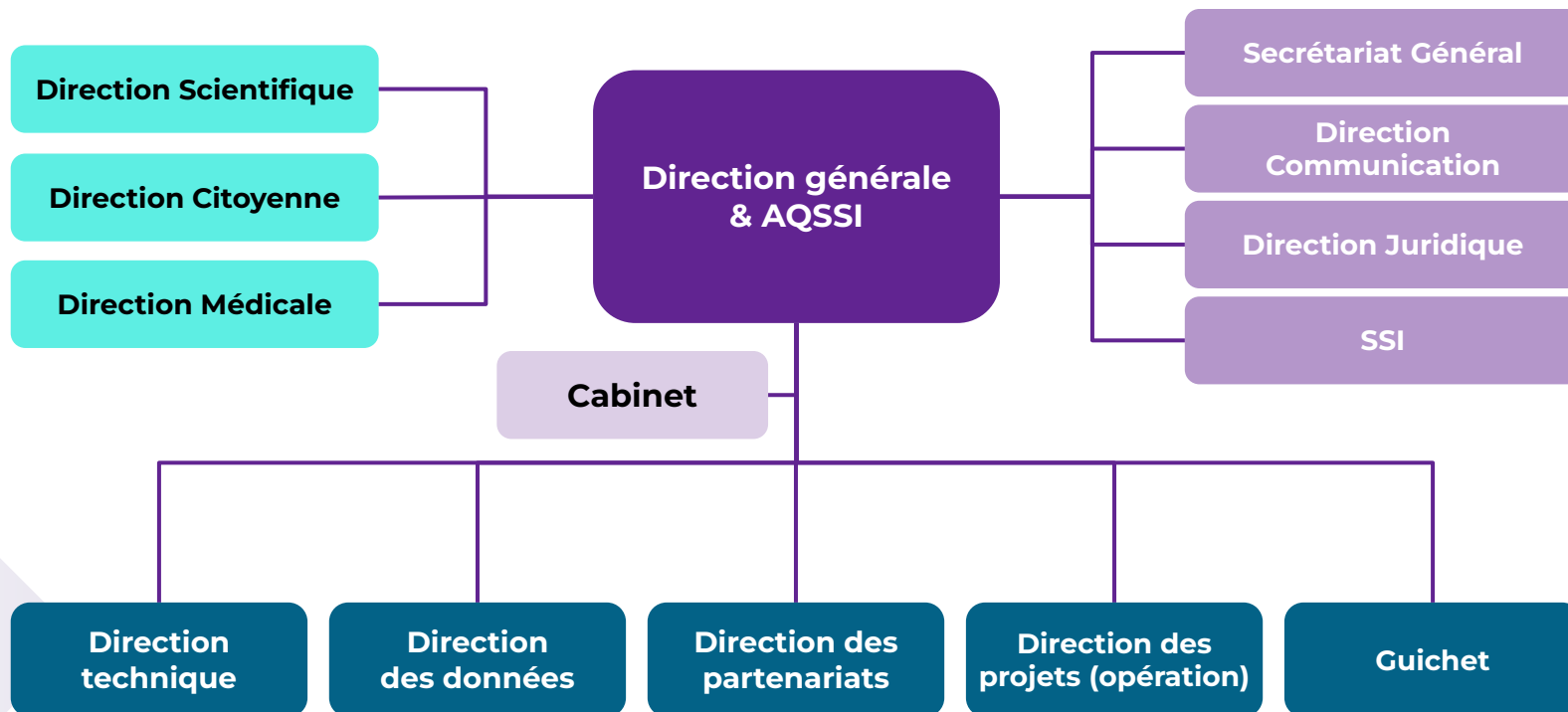
- **Documentation OMOP** : Vous pouvez trouver la documentation sur les identifiants de concept spécifiques aux divers types d'informations et bien d'autres détails sur le site officiel de [l'OHDSI](#) et [ATHENA](#).
- **Documentation SNDS** : Pour en savoir plus sur le SNDS et comment l'utiliser, nous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

# Mise en situation - 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 ?

# Organigramme





# Annexes 1/2

**La Direction scientifique du Health Data Hub** est chargée de l'animation académique et industrielle nationale et internationale, de la coordination et la contribution aux initiatives européennes et internationales phares en lien avec la donnée de santé. Elle définit la stratégie scientifique du Hub en se basant notamment sur les avis du Conseil Scientifique Consultatif et collabore étroitement avec les autres équipes sur les sujets scientifiques. Elle anime et fédère un vaste réseau national et international de chercheurs, d'industriels et d'institutionnels et crée des liens internes et externes.

**La Direction citoyenne du Health Data Hub** met en œuvre l'obligation légale d'informer les citoyens, de promouvoir et faciliter leurs droits. Elle vise aussi à produire une culture de la donnée de santé. Elle développe ainsi des actions qui visent à écouter de la société civile, mais aussi à l'informer, la former et l'impliquer.

**La Direction médicale du Health Data Hub** intervient dans la stratégie de développement en apportant une expertise médicale, en contribuant à la communication avec le monde des professionnels de santé et en fédérant cet écosystème.

**Le Secrétariat général du Health Data Hub** assure le bon fonctionnement administratif de l'organisation, sa politique et gestion RH, son pilotage budgétaire, ainsi que le suivi des marchés publics.

**La Direction Communication du Health Data Hub** intervient en support des autres directions, pour contribuer à développer la notoriété du HDH, veiller à son image, communiquer sur les actions menées, les projets accompagnés et mettre en valeur la position du HDH dans l'éco-système national, européen et international.

**La Direction juridique du Health Data Hub** intervient sur tous les aspects juridiques, en support des autres métiers, pour concourir au bon accomplissement de l'ensemble des missions du Health Data Hub et garantir la sécurité juridique de son fonctionnement et de ses actions. La supervision de la protection des données à caractère personnel est aussi assurée par l'équipe juridique.

**La Direction des Partenariats du Health Data Hub** intervient sur l'accompagnement des projets pilotes, des responsables de données, des partenariats et des offres de services

# Annexes 2/2

**L'équipe SSI du Health Data Hub** intervient sur tous les aspects sécurité , en support des autres métiers, pour garantir la conformité sécurité des SI du Hub aux référentiels et recommandations de sécurité (SNDS, ANSSI, CNIL etc.). Elle définit la politique de sécurité des systèmes d'information (PSSI) et veille à son application. L'équipe SSI est en charge de l'organisation d'exercice de gestion de crise d'origine cyber une fois par an. La supervision et la veille sécurité ainsi que la sensibilisation des agents et des utilisateurs aux risques cyber est également assurée par l'équipe SSI.

**La Direction des données du Health Data Hub** a pour mission de : (1) définir des stratégies novatrices sur la gestion, l'exploitation et le partage de données de santé ; (2) Partager et mutualiser en open source les outils et les connaissances nécessaires à l'analyse des données de santé ; (3) Gérer et mettre à disposition les données au sein de la plateforme ; (4) Fournir un soutien technique aux projets accompagnés par le HDH via la mise à disposition d'experts du SNDS, de data scientists et de data engineers.

**La Direction projets et services utilisateurs** accompagne l'ensemble des projets et bases utilisateurs exploitant la plateforme du HDH (partenaires ou non), en coordination avec les autres directions ressources (Guichet, Data, Technique, Partenariat). Elle assure un suivi tout au long du cycle de vie du projet, participe aux opérations projets et constitue un point de contact privilégié pour les utilisateurs (hors partenaires). A ce titre, elle a également la charge de partager au sein du HDH la vision consolidée du portefeuille de projets et bases (exploitant la plateforme) et d'aider à l'amélioration du produit et du processus projet, en lien avec les retours des utilisateurs.

**La Direction Accès aux données du Health Data Hub** assure le secrétariat des demandes d'accès aux données de santé (b) accompagne dans ce cadre les porteurs de projets dans leur démarche de soumission (appui réglementaire et sur les questions relatives au SNDS) (c) assure le secrétariat du CESREES (d) est force de proposition dans le déploiement des mesures de simplification de l'accès aux données

**La Direction Produit du Health Data Hub** a la responsabilité de définir, porter et mettre en œuvre la vision produit du Health Data Hub, par 1) l'identification des besoins des utilisateurs cibles du HDH 2) la priorisation et exécution de la roadmap produit en collaboration étroite avec les directions Plateforme et Données 3) la veille internationale sur les cas d'usage identifiés d'utilisation des données de santé

**La Direction technique du Health Data Hub** est chargée (1) de la conception, de l'implémentation et de la maintenance de la plateforme technologique des données de santé, (2) de la conduite des projets de développement des différents applicatifs, (3) de fournir une expertise technique sur les projets stratégiques du Health Data Hub et (4) d'opérer les SI Interne et métiers



Suivez-nous sur les réseaux sociaux !

