Référent des données (data engineer) H/F
Direction des données du HDH

Contrat – CDI
Rémunération – compétitive
Diplômes – Bac+5/Master
Date de début – dès que possible
Localisation – Paris 15ème (télétravail partiel possible)

LE HEALTH DATA HUB

Comment **améliorer les dépistages** et faire en sorte que les patients soient pris en charge le plus tôt possible ?
Comment leur **proposer les meilleurs traitements** sur le long cours ?
Comment **appuyer les professionnels de santé** dans un contexte clinique qui se complexifie ou en cas de crise sanitaire ?

L'Intelligence Artificielle et les **données de santé** font partie de la réponse. Elles sont incontournables pour la **recherche** et **l'innovation en santé**, par exemple, pour prévenir des insuffisances cardiaques à partir de données issues d'appareils connectés, pour accélérer le dépistage du cancer du sein à partir d'analyses automatiques des examens de mammographies ou même pour réunir assez d'informations afin d'améliorer la prise en charge des maladies rares.

Et pour ça, la France a la chance de disposer de **bases de données extrêmement riches** !
Mais ces données sont souvent sous exploitées car dispersées. Grâce à des **solutions innovantes** telles que l'IA, l'objectif du Health Data Hub est justement de permettre d'accéder de manière **facilitée**, **unifiée**, **transparente** et **sécurisée** à un catalogue de bases de données de santé françaises.

Comment ?
Le Health Data Hub a mis en place une **plateforme technologique** qui met à disposition des porteurs de **projets d'intérêt public**, dans un environnement technologique sécurisé et à l'état de l'art, les données de santé **pseudonymisées** des français. Ces porteurs de projets vont mobiliser des sources de données très **volumineuses**, les croiser entre elles, et utiliser une puissance de calcul pour faire tourner des algorithmes de recherche complexes. Il s'agit par exemple de projets de start-up pour améliorer des logiciels d'aide au professionnel de santé, de projets permettant **d'améliorer la prise en charge des patient**s en comparant l'efficacité de prise en charge, de projets portés par les administrations pour éclairer les politiques publiques.
Notre offre technologique, en constante évolution, peut être consultée ici.
Les défis sont de taille pour traiter ces données de santé sensibles, volumineuses de natures et formats variables. La plateforme doit être un **levier d'innovation** dans l'écosystème de la donnée de santé français.

En résumé, avec le Hub, nous **accompagnons des porteurs de projets innovants** qui contribuent à trouver les solutions de demain pour améliorer la santé de tous les citoyens.

DESCRIPTIF DU POSTE

Direction des données :

Pour mener à bien les missions qui lui ont été confiées, le Health Data Hub a formé la direction des données dont les objectifs principaux sont de :

1) **Définir des stratégies novatrices** sur la gestion, l'exploitation et le partage de données de santé, permettant de réaliser la vision du HDH ;
2) **Partager et mutualiser les outils et les connaissances** nécessaires à l'analyse des données de santé, dans le cadre d'une démarche open source.
3) **Gérer et mettre à disposition les données** qui lui sont confiées aux porteurs de projet au sein de la plateforme technologique du Health Data Hub ;
4) **Soutenir les projets d'intérêt public** que le HDH accompagne, aussi bien sur la compréhension des données de santé que sur leur exploitation via des experts des données de santé, des data scientists et des data engineers.

Pôle "Gestion des données" :

Pour répondre à la troisième mission qui lui a été conférée et définir une approche claire pour l'écosystème de la santé, la direction des données s'est dotée d'un pôle "Gestion des données". Ce pôle est responsable de l'intégralité du cycle de vie des données, et se structure autour des chantiers suivants :

1) **Traitement** des données de santé massives et diverses transmises par les porteurs de projet à la plateforme technologique du Health Data Hub ;
2) **Gestion et mise en qualité** des données de santé stockées dans la plateforme technologique du Health Data Hub ;
3) **Application** de librairies en Python ou R pour faciliter, automatiser et systématiser les traitements des données cités précédemment ;
4) **Analyses exploratoires** de nouvelles fonctionnalités et applications (e.g., cluster spark, lecteur d'images spécifiques au secteur de la santé) à intégrer à la plateforme technologique du Health Data Hub.

Ces missions sont essentielles pour garantir la fiabilité des recherches menées sur la plateforme technologique et présentent d'importants défis au regard du caractère hétérogène des données manipulées (e.g., données médico-administratives, imagerie médicale, compte-rendus médicaux) et des efforts nécessaires pour les rendre utilisables.

Activités du poste :

Au sein du pôle "Gestion des données", vous réaliserez en particulier les traitements nécessaires (1) pour l'ingestion des données dans la plateforme et (2) pour la bonne gestion et la mise en qualité des données présentes sur la plateforme technologique du Health Data Hub. A ce titre, les principales missions seront les suivantes :

- Collaborer conjointement avec la Direction Projets et Services utilisateurs et prendre connaissance du protocole scientifique et des buts premiers de chacun des projets accompagnés. Cette phase s'accompagne d'une découverte du ou des jeu(x) de données complet(s) tant au niveau fonctionnel qu'au niveau technique, ainsi que de la rédaction d'une documentation de ce(s) dernier(s) afin (1) de s'assurer en amont du bon respect des règles de pseudonymisation et (2) effectuer des opérations de vérifications des données en aval de façon automatique ;
- Vérifier le caractère anonymisé des demandes d'import et d'export des données ou codes sur la plateforme, en relation avec des acteurs externes (producteurs de données, porteurs de projets de recherche) ;
- Travailler dans la plateforme technologique du Health Data Hub ;
    - Utiliser les librairies existantes et les compléter pour développer des scripts Python et PySpark permettant de manipuler des grande quantité de données (~To) sous différents formats (e.g., tabulaires, texte libre, images, JSON) reçues sur la plateforme technologique.
    - Vérifier l'intégrité, confidentialité et conformité à certains critères de qualité définis en amont ainsi que de de les préparer pour leur mise à disposition (e.g., reformatage, jointure, transformation parquet, etc.) ;
    - Remonter les besoins permettant l'évolution des librairies et outils existants afin d'améliorer la qualité et la rapidité des opérations,
    - Contribuer à la documentation des opérations,
- Assurer un premier niveau de support technique aux utilisateurs externes pour l'utilisation de leurs projets par exemple avec des exemples de code utilisant leurs données.

La tech stack (pile de technologies) utilisée pour ces missions sera principalement :

- Python comme langage de programmation généraliste :
    - notebooks Jupyter pour accéder à la plateforme et organiser la documentation d'utilisation (tutoriels),
    - pandas pour l'analyse des données CSV de petite taille et Spark / pyspark pour les données volumineuses,
- Gitlab et gitea pour l'utilisation et la synchronisation avec les librairies existantes,
- Microsoft Azure pour le stockage et le requêtage de données volumineuses,
- Suite Google pour la bureautique (Google Docs, Google Sheets, etc.)

Cette liste est non exhaustive; le collaborateur pourra appuyer sa (ses) direction(s) dans d'autres missions.

PROFIL RECHERCHÉ

Compétences indispensables

- Bonne maîtrise du langage Python
- Bonne maîtrise de SQL et de gestion de bases de données
- Bonne maîtrise des librairies de traitement de données (e.g., pandas, dplyr)
-
- Connaissance des outils en ligne de travail collaboratif type Git (GitHub ou GitLab)
- Capacités rédactionnelles
- Bon relationnel : capacité à interagir avec les partenaires externes du HDH (startups, institutions publiques, etc.)

Compétences additionnelles recherchées

- Maîtrise des frameworks de calcul distribué (Spark, Dask)
- Maîtrise de R
- Maîtrise d'environnements cloud (notamment Azure Blob Storage pour le stockage de données)
- Expérience avec des formats de données complexes (par exemple : images DICOM, SVS, JSON complexes, CSV de très grande taille etc.)

POURQUOI CHOISIR LE HEALTH DATA HUB ?

Vous êtes motivé(e) à rejoindre une équipe impliquée dans un projet ambitieux, qui a du sens et une finalité d'intérêt public ? Rejoignez-nous !
Notre récente structure, d'une centaine de collaborateurs/trices, a besoin de talents créatifs, autonomes et proactifs pour continuer de grandir ! Ensemble, nous nous sommes engagés à :

- Accompagner les porteurs de projet visant à analyser les données de santé pour le bien commun.
- Construire et opérer une plateforme technologique pour leur offrir les meilleurs outils avec un très haut niveau de sécurité à respecter.
- Réunir et mettre en forme les données au plus grand potentiel pour la recherche et l'innovation.
- Promouvoir le partage des connaissances, des expertises et du savoir et diffuser une culture de la donnée de santé auprès de tous.

Bon à savoir:

💪 Rejoindre le HDH c'est surtout participer à un projet enrichissant humainement qui a du sens, avec un fort impact sociétal
🏆 Au HDH on favorise la prise d'initiative, dans une ambiance de challenge perpétuel
😎 Ici la bonne humeur et l'esprit d'équipe règnent

PROCÉDURE DE RECRUTEMENT

Après avoir postulé, voilà comment se déroulera le recrutement:

- Un premier entretien avec le directeur de l'équipe Data

- Un test technique à réaliser chez soi
- Un entretien technique, basé sur le test, avec des membres de l'équipe Data
- Un entretien final avec la directrice du Health Data Hub

Pour postuler

RDV sur Welcome to the jungle **ici** (CV et Lettre de motivation obligatoire)