**Health Data Project — Rapport Final**

DATA ENGINEERING ON CLOUD



Sekari Inès - Nkuida Malaïka

**Table des matières**

1. **Présentation du sujet**
   * + - **Contexte**
       - **Objectifs du projet**
       - **Les sources de données**
2. **Outils et technologies utilisés**
3. **Architecture**
   * + - **Détails**
4. **Modélisation dimensionnelle/MCD**
   * + - **Schéma en étoile**
5. **Implémentation technique**
   * + - **Traitements et transformations**
       - **ETL**
6. **Construction du modèle sémantique/Power BI**
7. **Visualisation et analyses**
8. **Conclusion et perspectives**
9. **Sources**
10. **Présentation du sujet**

## **1. Introduction du projet**

Ce projet s’inscrit dans le cadre d’une démarche de valorisation des données de santé publique à l’échelle mondiale. À partir d’un dataset issu de **Kaggle - Global Health Statistics**, notre objectif a été de mettre en place un pipeline complet de traitement de données et de machine learning permettant à la fois :

* Une **analyse exploratoire et visuelle des indicateurs de santé globaux** via Power BI ;
* La **prédiction de certaines variables sanitaires critiques** telles que les taux de mortalité ou de récupération ;
* La **mise en forme des données selon une architecture modulaire** inspirée des bonnes pratiques du Data Engineering : Bronze, Silver et Gold.

## **2. Objectif du projet**

Notre projet "Global Health" poursuit un double objectif :

1. **Analyser et visualiser les données mondiales de santé publique** afin d’identifier les pays et régions à risque, suivre l’évolution des maladies, et produire des indicateurs de performance clés (KPIs).
2. **Mettre en place un modèle prédictif** capable d’estimer la mortalité ou la récupération selon les pays, en fonction de plusieurs variables socio-économiques et médicales (espérance de vie, dépenses de santé, accès aux soins, etc.).

## **3. Description du dataset**

Le dataset provient de Kaggle, intitulé **" Global Health Statistics"**. Ce dernier fournit des statistiques détaillées sur la santé mondiale, en mettant l’accent sur divers types de maladies, leurs traitements et les résultats associés. Il couvre plusieurs pays et années, offrant ainsi une base riche pour des analyses en épidémiologie, des recherches en santé publique, avec des **variables quantitatives et catégorielles**:

Une image contenant symbole, Police, logo, Rectangle

Le contenu généré par l’IA peut être incorrect.

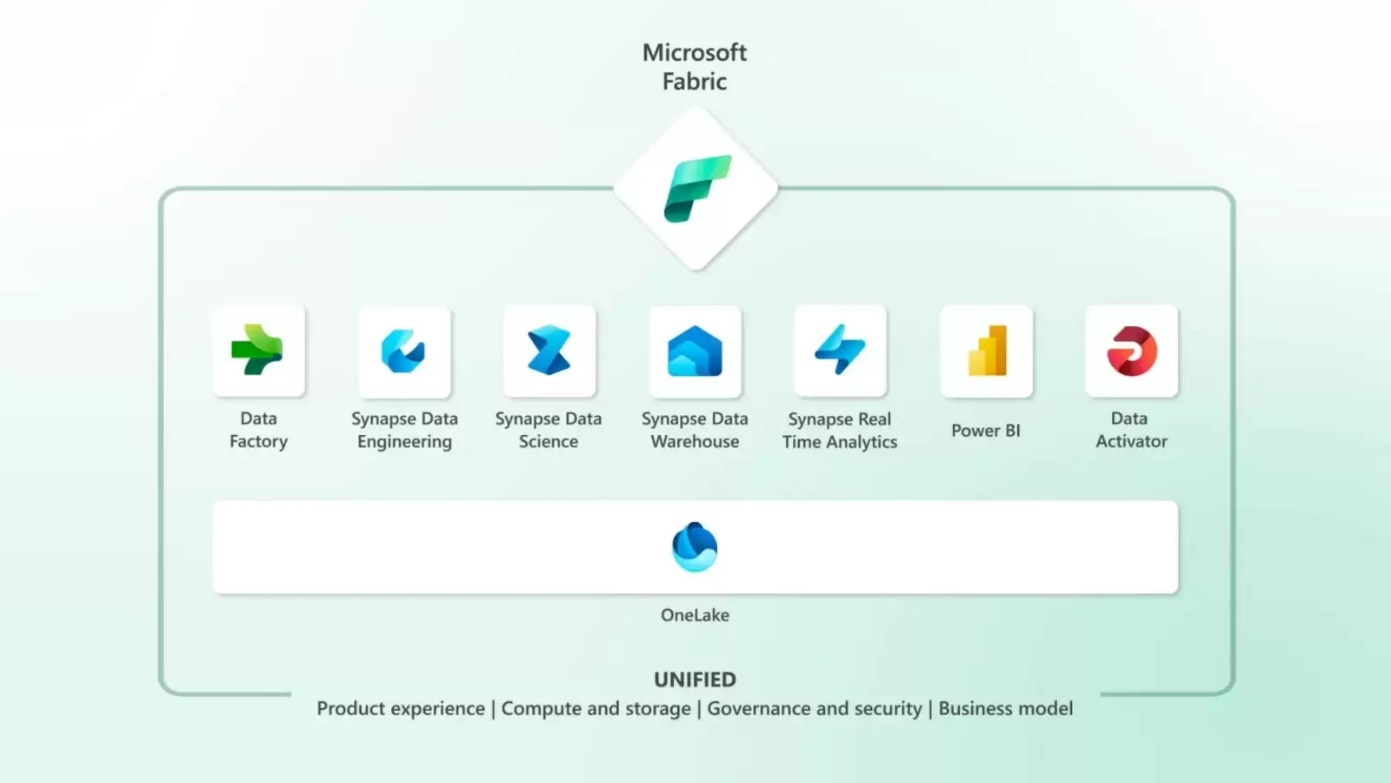
**global\_health.csv**

Ce fichier contient des données anonymisées sur les visites aux urgences, symptômes, diagnostics. Les données issues de ce fichier sont très riches et couvrent plusieurs dimensions du séjour patient, structurées autour de quatre grands ensembles.

### Description des colonnes :

* **Country** : Nom du pays dans lequel les données de santé ont été enregistrées.
* **Year** : Année de collecte des données.
* **Disease Name** : Nom de la maladie ou du trouble de santé suivi.
* **Disease Category** : Catégorie de la maladie (ex : Infectieuse, Non transmissible).
* **Prevalence Rate (%)** : Pourcentage de la population touchée par la maladie.
* **Incidence Rate (%)** : Taux de nouveaux cas diagnostiqués dans la population.
* **Mortality Rate (%)** : Taux de décès parmi les personnes atteintes de la maladie.
* **Age Group** : Tranche d’âge la plus concernée par la maladie.
* **Gender** : Genre(s) concerné(s) par la maladie (Homme, Femme, Les deux).
* **Population Affected** : Nombre total d’individus affectés par la maladie.
* **Healthcare Access (%)** : Pourcentage de la population ayant accès à des soins de santé.
* **Doctors per 1000** : Nombre de médecins pour 1 000 habitants.
* **Hospital Beds per 1000** : Nombre de lits d’hôpital disponibles pour 1 000 habitants.
* **Treatment Type** : Type principal de traitement utilisé (ex : Médicaments, Chirurgie).
* **Average Treatment Cost (USD)** : Coût moyen du traitement de la maladie (en dollars américains).
* **Availability of Vaccines/Treatment** : Indique si des vaccins ou traitements sont disponibles.
* **Recovery Rate (%)** : Taux de guérison observé chez les personnes atteintes.
* **DALYs** (Disability-Adjusted Life Years) : Mesure de la charge de morbidité (années de vie ajustées sur l’incapacité).
* **Improvement in 5 Years (%)** : Amélioration des indicateurs de santé liés à la maladie sur les cinq dernières années.
* **Per Capita Income (USD)** : Revenu moyen par habitant dans le pays concerné.
* **Education Index** : Indice moyen de niveau d’éducation du pays.
* **Urbanization Rate (%)** : Pourcentage de la population vivant en zone urbaine.

**2. Outils et technologies utilisés**



Pour répondre aux enjeux d’intégration, d’analyse et de valorisation des données de santé hétérogènes, notre projet s’appuie sur un écosystème moderne, performant et évolutif.  
Voici les principaux outils et plateformes exploités :

**Microsoft Fabric (Lakehouse)**

Microsoft Fabric constitue le socle central de notre architecture data. Il s’agit d’une plateforme de données et d’analytique complète, dédiée à l’unification du stockage, du traitement et de la valorisation des données.

Parmi ses atouts majeurs :

* **Solution Tout-en-Un** : Fabric centralise l’ensemble des opérations (ingestion, traitement, stockage, restitution) en une seule plateforme, réduisant la complexité et évitant l’intégration manuelle de briques multiples.
* **Interface conviviale** : la prise en main est facilitée pour les data engineers/scientists comme pour les métiers, optimisant la gestion quotidienne de la donnée.
* **Modèle SaaS** : la plateforme est entièrement gérée par Microsoft. Les mises à jour, la scalabilité et la maintenance sont prises en charge, ce qui nous permet de nous concentrer sur la valeur métier.
* **Sécurité et conformité** : Fabric garantit un haut niveau de sécurité, de traçabilité et de conformité réglementaire, essentiels pour les données sensibles du secteur santé.

**PySpark**

Le traitement, le nettoyage et l’exploitation avancée des volumes importants de données sont réalisés via **PySpark**, la version Python de l’API Spark.

* Permet de manipuler efficacement de grands jeux de données (distribution, parallélisation).
* Adapté aux opérations complexes de croisement et d’enrichissement.

**Power BI**

Pour la restitution et la visualisation des analyses, **Power BI** s’impose comme l’outil privilégié :

* Puissant, interactif, connecté en natif à Fabric.
* Permet la création de tableaux de bord dynamiques, favorisant l’exploration et la prise de décision pour les utilisateurs métiers.
* Assure la sécurité au niveau de la restitution des données sensibles.

**GitHub**

La gestion du code, des scripts de transformation et de l’ensemble des artefacts du projet s’effectue via **Git/GitHub** :

* Facilite la collaboration entre les membres de l’équipe.
* Permet l’historisation, la gestion des versions, et une traçabilité complète des évolutions techniques.
* Sécurise les développements et facilite le passage à l’échelle ou la reproductibilité du projet.

**Ce socle technologique garantit flexibilité, robustesse, sécurité et efficacité pour traiter et valoriser les données de santé, du brut à l’insight décisionnel.**

**3. Architecture**

Une image contenant capture d’écran, diagramme, texte, Plan

Description générée automatiquement

**Détails de l’architecture :**

* **Stockage initial (Zone Lakehouse – Raw Data) :**  
  Tous les fichiers de données sources (hospital-triage.csv, meds\_cat.csv, icd\_code.csv, motif\_categorie.csv) sont déposés dans un espace de stockage du Lakehouse spécialement dédié.  
  Cette zone permet de centraliser et d’archiver toutes les données brutes, sans transformation.

**Objectif :** disposer d’un référentiel de données original, assurant l’auditabilité et la traçabilité.

* **Bronze Layer (données brutes intégrées) :**  
  À l’aide de notebooks PySpark/Spark, chaque fichier source est importé **tel quel** dans la couche dite *Bronze* du Lakehouse avec un léger pré-nettoyage.  
  Les jeux de données sont chargés dans des tables distinctes dont la source principale éclatée en 4 tables :
  + Table patient provenant d’hospital-triage (caractéristiques patients)
  + Table maladie provenant d’hospital-triage (pathologies antérieures et diagnostic)
  + Table motif\_admission provenant d’hospital-triage (motifs de prise en charge)
  + Table médicament provenant d’hospital-triage (traitements consommés lors de la prise en charge)
  + référentiels : icd\_code, meds\_code, motifs\_code  
    Aucune transformation ni nettoyage significatif n’est appliqué à cette étape : le focus est sur la conservation de l’historique et la fidélité des données d’origine.
* **Silver Layer – Zone de Données Nettoyées et Transformées**  
  Les données bronzes sont nettoyées, standardisées, dédoublonnées et enrichies.
  + Correction des valeurs invalides ou manquantes
  + Standardisation des variables métier (âge, genre…)
  + Extraction de features analytiques (diagnostics, médicament, motifs d’admission, timestamps…)
  + Préparation des données pour la modélisation en étoile (tables de faits et dimensions).
* **Gold Layer - Zone Analytique**

Résultat final, structuré pour l’exploitation métier ; création d’un modèle dimensionnel pour l’analyse (tables de faits/dimensions adaptées). Dans cette couche, les données sont nettoyées, normalisées, auditées ; elles sont croisées pour créer une **modélisation analytique** adaptée à la visualisation et à l’analyse.

* + Construction d’un **modèle en étoile** :
    - Table de faits : Consultation patient, centralisant les mesures d’intérêt
    - Tables de dimensions :
      * dim\_country
      * dim\_disease
      * dim\_demographics
* **Modèle sémantique :**  
  Un **modèle sémantique** est construit dans Fabric pour synthétiser les indicateurs clés et faciliter la consommation des données par les utilisateurs finaux.  
  Il centralise les mesures (par exemple : taux de passage pour motif respiratoire, score de gravité moyen), les dimensions d’analyse (âge, période, motif…), et les calculs métiers, assurant ainsi l’homogénéité des analyses.
* **Restitution & DataViz – Power BI :**  
  Le modèle sémantique est exposé à Power BI pour permettre une analyse visuelle interactive.  
  Les utilisateurs finaux peuvent explorer dynamiquement les parcours patients, les tendances de pathologies/motifs ou encore l’usage des classes thérapeutiques.

**Cette architecture modulaire et robuste est conçue pour évoluer facilement : elle garantit rapidité d’analyse, intégrité des données sources, traçabilité totale, et simplicité de restitution métier.**

**4. Modélisation dimensionnelle/MCD**

**Schéma en étoile**

Pourquoi un schéma en étoile ?

Nous avons opté pour un modèle en étoile car :

* Il est facile à comprendre et facilite l’analyse.
* La structure centrée sur une table de faits entourée de tables de dimensions permet d’accéder rapidement aux attributs utiles sans jointures complexes.
* C’est le modèle de référence pour la BI sur des données transactionnelles/sanitaires complexes.

**Une image contenant texte, capture d’écran, diagramme, conception

Description générée automatiquement**

Dans ce schéma :

Fact\_consultation est la table de faits centrale, représentant les consultations patients et centralisant les mesures d’intérêt.

Les tables de dimensions entourent la table de faits, fournissant des contextes supplémentaires pour l'analyse.

**5. Implémentation technique**

**Traitements et transformations**

**Lien Github :** https://github.com/inessekari/data-cloud-project1

Dans notre démarche, nous utilisons deux services principaux de Microsoft Fabric : l’un pour la partie ETL (traitement et transformation des données), l’autre pour la visualisation et la restitution analytique. Tous les scripts de transformation, de nettoyage et de modélisation sont versionnés sur GitHub via des notebooks pour une meilleure traçabilité et collaboration.

**ETL – Création du workspace**

La phase ETL (Extract, Transform, Load) est réalisée à travers les notebooks fournis par Microsoft Fabric, ce qui nous permet d'automatiser chaque étape du processus, de l’import brut jusqu’au modèle prêt à l’analyse.

Nous avons tout d’abord créé un **Workspace** dédié afin de centraliser le projet, d’organiser le travail en binôme et de structurer proprement l’ensemble des ressources.

Une image contenant texte, capture d’écran, logiciel, Police

Le contenu généré par l’IA peut être incorrect. Une image contenant texte, capture d’écran, logiciel, Icône d’ordinateur

Le contenu généré par l’IA peut être incorrect.

**ETL – Création du lakehouse**

Dans ce workspace, nous avons mis en place un **Lakehouse** qui nous sert de socle unique pour stocker, transformer et requêter l’ensemble des données du projet.

Une image contenant capture d’écran, texte, ligne

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, capture d’écran, Police, nombre

Le contenu généré par l’IA peut être incorrect.Une image contenant texte, capture d’écran, logiciel, Icône d’ordinateur

Le contenu généré par l’IA peut être incorrect.

**ETL – Chargement des données**

Pour organiser les sources de données, nous avons créé dans l’espace “Files” un sous-dossier « Bronze » dans lequel nous déposons les différents fichiers CSV nous servant de sources. Cela garantit une gestion claire des sources ainsi qu’une reprise aisée en cas de mise à jour des fichiers.

Une image contenant texte, Police, nombre, capture d’écran

Description générée automatiquement

Nous avons utilisé des notebooks pour ingérer notre source brute en table **Bronze** (raw data). Ces notebooks documentent les opérations de typage, de renommage et de contrôle de qualité de la donnée lors de l’import, offrant visibilité et facilité de maintenance. Chaque notebook est renommé pour plus de précision et de clarté.

Une image contenant texte, capture d’écran, nombre, Police

Description générée automatiquement

Comme précisé dans les détails de notre Architecture, nous avons quatre sources de données et dont une source principale et des référentiels. Notre premier objectif est de scinder la source principale en quatre sources : patient, maladie, médicament et motif d’admission.

**📓 Notebook 1 – Ingestion et nettoyage (Bronze → Silver)**

* Chargement des données brutes issues de Kaggle.
* Uniformisation des formats (dates, pays, codes, valeurs manquantes).
* Gestion des doublons, renommage des colonnes pour standardiser les noms.
* Filtrage des variables pertinentes (nous avons écarté certaines colonnes non exploitables).
* Résultat : un **dataset Silver**, propre et exploitable pour l’analyse.

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

Une image contenant texte, Police, capture d’écran

Description générée automatiquement

Une image contenant texte, capture d’écran, reçu

Description générée automatiquement

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

Une image contenant texte, reçu, capture d’écran

Description générée automatiquement

Une image contenant texte, capture d’écran

Description générée automatiquement

Une image contenant texte, capture d’écran, Police, reçu

Description générée automatiquement

### 📓 **Notebook 2 – Construction du modèle en étoile (Silver → Gold)**

* Structuration des données selon un **modèle en étoile** pour optimiser l’analyse dans Power BI :
  + **Table des faits** : données mesurables (ex : taux de mortalité, dépenses de santé).
  + **Dimensions** : pays, années, types de maladies, sexe, groupe d'âge.
* Création de la **table Gold** destinée à Power BI.
* Jointures et agrégations nécessaires pour assurer une cohérence entre les tables.
* Résultat : données enrichies, prêtes pour visualisation et modélisation.

Une image contenant texte, capture d’écran, Police, reçu

Description générée automatiquement

Une image contenant texte, capture d’écran, conception

Description générée automatiquement

Une image contenant texte, capture d’écran, nombre

Description générée automatiquement

Une image contenant texte, capture d’écran

Description générée automatiquement

Une image contenant texte, capture d’écran, Police, lettre

Description générée automatiquement

Une image contenant texte, capture d’écran

Description générée automatiquement

Une image contenant texte, capture d’écran

Description générée automatiquement

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

Une image contenant texte, capture d’écran, diagramme, Tracé

Description générée automatiquement

**📓 Notebook 3 – Machine Learning (Prédiction)**

* Objectif : **Prédire la mortalité ou le taux de récupération** en fonction d’autres variables.
* Préparation des features : normalisation, encodage, split train/test.
* Tests de plusieurs modèles :
  + **Régression linéaire**
  + **Random Forest**
  + **Gradient Boosting**
* Sélection du modèle le plus performant selon le contexte et les métriques (RMSE, R²).
* Visualisation des performances et interprétation des résultats.

Une image contenant texte, capture d’écran, Police, algèbre

Description générée automatiquement

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

Une image contenant texte, capture d’écran

Description générée automatiquement

Une image contenant texte, Police, capture d’écran

Description générée automatiquement

Une image contenant texte, capture d’écran

Description générée automatiquement

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

Une image contenant texte, capture d’écran

Description générée automatiquement

**Une image contenant texte, capture d’écran

Description générée automatiquement**

**Une image contenant texte, capture d’écran, logiciel, Page web

Description générée automatiquement**

**Une image contenant texte, ligne, nombre, capture d’écran

Description générée automatiquement**

**Une image contenant texte, ligne, nombre, capture d’écran

Description générée automatiquement**

**Une image contenant texte, Police, capture d’écran

Description générée automatiquement**

**ETL – Workflow du projet *Global Health***

Dans le cadre du projet *Global Health*, nous avons structuré le pipeline de traitement de données autour d’un workflow cohérent, rigoureux et reproductible, basé sur une architecture multi-zones (Bronze, Silver, Gold). Ce processus est orchestré dans *Microsoft Fabric* à l’aide de notebooks PySpark, avec une gestion de version centralisée sur GitHub.

Les principales étapes sont les suivantes :

**• Étape 1 : Ingestion des données brutes et contrôle initial *(Zone Bronze – Notebook 01)***

Les données issues du dataset *Global Health Statistics* (source : Kaggle) sont tout d’abord chargées en zone Bronze sans modification. Un premier diagnostic est effectué : typage automatique, structure du schéma, comptage des lignes, exploration statistique et détection des valeurs nulles. Cette phase permet d’évaluer la qualité initiale des données. Un log technique est généré pour documenter cette ingestion.

**• Étape 2 : Nettoyage, normalisation et structuration *(Zone Silver – Notebook 01)***

Les données sont ensuite nettoyées : suppression des valeurs nulles, harmonisation des noms de colonnes, suppression des doublons et formatage conforme aux bonnes pratiques de structuration analytique. Le dataset ainsi traité est sauvegardé en format Delta dans la zone Silver, prêt pour des traitements analytiques ou de modélisation.

**• Étape 3 : Modèle en étoile et préparation analytique *(Zone Gold – Notebook 02)***

À partir des données Silver, nous avons construit un **modèle en étoile** en définissant des tables de dimensions (pays, maladie, démographie) et une table de faits centralisant les indicateurs de santé (taux de mortalité, DALYs, coûts, etc.). Cette structuration vise à faciliter l’intégration dans Power BI et les analyses multidimensionnelles. Des étapes de contrôle de qualité et des statistiques descriptives sont également intégrées.

**• Étape 4 : Modélisation prédictive et visualisation *(Notebook 03 + Power BI)***

En parallèle de l’analyse descriptive, des modèles de machine learning ont été entraînés pour prédire :

* le **taux de mortalité** (régression),
* la **disponibilité d’un vaccin ou traitement** (classification),
* et les **DALYs** (régression).

Les résultats et insights sont mis en valeur dans Power BI à travers des tableaux de bord interactifs, des visualisations statistiques, et des KPIs dynamiques facilitant la prise de décision.

**6. Construction du modèle sémantique/Power BI**

Le modèle sémantique permet aux utilisateurs finaux d’explorer et d’analyser les données de façon intuitive, sans s’encombrer de la complexité technique des sources brutes ou des étapes ETL. Il repose sur une organisation logique avec une table de faits, des dimensions, des colonnes pré-nettoyées, ainsi que d’éventuelles hiérarchies.

Dans notre projet, nous incluons toutes les tables transformées (uniquement Gold).

Une image contenant texte, capture d’écran, Police, document

Description générée automatiquement

Une image contenant texte, logiciel, diagramme, Icône d’ordinateur

Description générée automatiquement

**7. Visualisation et analyses**

**8. Conclusion et perspectives**

Ce projet s’est inscrit dans une démarche méthodologique rigoureuse de traitement et d’analyse des données de santé publique à l’échelle mondiale. En mobilisant une architecture en trois zones (Bronze, Silver, Gold), nous avons mis en œuvre un pipeline de transformation de données assurant la qualité, la traçabilité et l’exploitabilité des informations issues d’un jeu de données complexe.

Dans la première phase, les données brutes ont été ingérées, nettoyées et structurées afin d’éliminer les incohérences, valeurs manquantes et doublons, garantissant ainsi une base fiable pour l’analyse. La seconde phase a permis la modélisation des données selon un **modèle en étoile**, facilitant leur intégration dans des outils de visualisation comme Power BI et leur interrogation selon une logique dimensionnelle.

La troisième phase a consisté en la mise en œuvre de plusieurs modèles d’apprentissage automatique : régression (Gradient Boosted Trees) pour la prédiction du **taux de mortalité** et des **DALYs**, et classification pour la prédiction de la **disponibilité d’un traitement ou d’un vaccin**. Les performances obtenues confirment la pertinence des variables explicatives choisies et montrent l’intérêt des approches prédictives dans le domaine de la santé publique.

En somme, ce projet illustre la capacité à combiner une architecture de données robuste avec des techniques de data science pour extraire des connaissances à forte valeur ajoutée, susceptibles de contribuer à une meilleure compréhension des déterminants de la santé à l’échelle internationale.

Ce projet démontre **l’efficacité du cloud analytique** pour transformer des données de santé complexes en **insights exploitables** pour améliorer les soins.

**9. Sources**

**https://learn.microsoft.com/fr-fr/fabric/fundamentals/microsoft-fabric-overview**