Projet **Manipulations de données avec OMOP**

Ce projet a pour but de simuler le cycle de vie des données de santé et de vous confronter aux problématiques inhérentes à leur collecte, leur traitement et leur utilisation. Il est composé de trois parties distinctes portant respectivement sur la planification d’un ETL à partir d’une source de données, le stockage de données dans un format standardisé, et l’utilisation des données dans le cadre d’une étude clinique.

Il vous est recommandé de lire la totalité du sujet avant de commencer le projet. De même, il vous est vivement conseillé de prendre le temps de réfléchir aux problématiques et difficultés que vous pourriez rencontrer après la prise en compte du sujet, avant de vous lancer dans l’implémentation technique du projet.

Ce projet est à réaliser en binôme. Un rapport de projet est attendu, ainsi qu’un support de présentation pour la soutenance.

Date limite d’envoi du rapport : **vendredi 27 mai, 18h00.**

## Partie I : design d’ETL

L’objectif de cette partie est de produire des spécifications d’[ETL](https://en.wikipedia.org/wiki/Extract,_transform,_load). A partir de données sources et à l’aide d’outils de la communautés OHDSI, vous produirez un document technique permettant d’expliciter les transformations nécessaires pour adapter les données sources au format OMOP CDM v5.4. Ce document doit permettre à tout intégrateur d’implémenter un ETL de ces données sources sans ambigüité ni interprétations personnelles.

## Récupération de données synthétiques

Récupérez le jeu de données **1K Sample Synthetic Patient Records** au format **CSV** sur le site : <https://synthea.mitre.org/downloads>

## Génération d’un rapport sur les données sources

Téléchargez l’outil **WhiteRabbit** depuis le dépôt d’OHDSI. La *release* disponible contient également l’outil RabbitInAHat, qui sera également utile plus tard.

Dans WhiteRabbit, chargez l’ensemble des fichiers CSV du jeu de données précédemment téléchargé.

Générez un rapport sur les données en cliquant sur « Scan Tables ». Assurez-vous d’avoir sélectionné l’ensemble des lignes du jeu de données et demandé la génération de statistiques numériques avant de générer le rapport.

Le rapport généré est un fichier Excel.

* + **Commentez son contenu.**
  + **Interprétez les différentes feuilles qu’il contient et explicitez quelles informations vous en retenez.**
  + **Documentez ce que vous apprenez du jeu de données source et de ses tables.**

## Réalisation d’un mapping de données

Générez un nouveau rapport avec WhiteRabbit, cette fois-ci uniquement sur les tables Patients, Encounters, Conditions, Medications, Observations, Procedures, Allergies et Immunizations.

Ouvrez l’outil **RabbitInAHat**. Chargez le rapport précédemment généré par WhiteRabbit.

Réalisez un mapping initial des fichiers sources par table OMOPv5, sur la base de votre analyse précédente. Attention, un fichier peut alimenter plusieurs tables, et une table peut être alimentée par plusieurs fichiers.



*Exemple de mapping multiple*

Pour chaque mapping de table, double-cliquez sur la flèche pour accéder au mapping des champs.

La sélection des flèches des mappings (de table comme de champ) permet de définir la logique de mapping à implémenter ainsi que d’éventuels commentaires.

Une fois vos mappings réalisés, exportez-le en tant que document Word. Il sera à verser en annexe de votre rapport.

Documentez dans votre rapport les étapes réalisées dans cette partie I.

## Partie II : Implémentation de données au modèle OMOP

## Récupération de données synthétiques

Téléchargez le jeu de données covid19, modèle **OMOPv5** au format **CSV** sur le site : <https://dev.azure.com/shambergerm/Covid19/_git/covid19Storage>

## Récupération des terminologies

Sur la plateforme **Athena** d’OHDSI, téléchargez les terminologies utilisées dans le jeu de données. Il s’agit de LOINC, CVX, SNOMED et RxNorm. Le format à télécharger est le 5.

Le téléchargement des terminologies est soumis à inscription et prend quelques dizaines de minutes à être disponible au téléchargement entre la demande et la génération du lien.

## Création d’une base OMOP

Chargez l’ensemble des données (patients et terminologiques dans une base de données). Postgres est conseillé mais l’implémentation est libre et possible sur toute base de données SQL.

Les spécifications et contraintes du CDM sont décrites dans [la documentation fournie par OHDSI](https://ohdsi.github.io/CommonDataModel/cdm60.html). En complément de cette documentation, OHDSI fournit également des implémentations formelles en SQL, pour faciliter la tâche des développeurs. Ces implémentations pour une dizaine de types de serveur SQL sont disponibles sur [un dépôt dédié](https://github.com/OHDSI/CommonDataModel).

Toutefois, gardez à l’esprit que les données de santé ne sont jamais des données parfaites. Intégrer des données dans un modèle requiert toujours des prises de décisions pour définir les ETL. Vous devrez probablement altérer le modèle ou transformer les données pour arriver à un résultat final.

Documentez dans votre rapport les choix faits et les étapes réalisées dans cette partie II.

## Partie III : Exploration de données

A partir de la base générée en partie II, vous allez réaliser des analyses portant sur les codes utilisés dans les données.

Dans le cadre de la pandémie mondiale de la maladie émergente COVID-19 due au nouveau coronavirus SRAS-CoV-2, la caractérisation clinique et biologique de la maladie est primordiale ainsi que l’évaluation de l’impact de la crise sanitaire sur les patients pris en charge dans les établissements de santé.

Sur la base du jeu de données collecté dans la partie II, vous simulez votre participation à une étude clinique en lien avec cette pandémie.

### Etude : Participation à l’étude 4CE

|  |
| --- |
| En mars 2020, le Pr. Isaac Kohane, directeur du i2b2 National Center for Biomedical Computing (Harvard Medical School), a lancé une initiative académique internationale de constitution de jeux de données ouvertes en lien avec l’épidémie COVID19 (https://covidclinical.net). Le consortium 4C (Consortium for Clinical Characterization of COVID-19 by EHR) est un réseau international de recherche associant des acteurs majeurs de l'informatique médicale issus de centres hospitalo-universitaires prestigieux dans le monde (Etats-Unis, Europe et Asie) dont l’objectif est de conduire des recherches épidémiologiques collaboratives sur des données partagées concernant la Covid19. Les centres participant au réseau 4CE ont l’opportunité de décrire l’évolution clinique des patients atteint de Covid ou impactés par la crise sanitaire et identifier les nombreuses et quelques variations spécifiques de l’état de santé des patients.  À cette fin, les centres qui souhaitent participer doivent produire des fichiers de données agrégées :   * **DailyCounts.csv** :   + Une ligne par jour   + 4 colonnes     - Date     - Nombre de nouveau patient diagnostiqué covid (PCR +)     - Nombre de patient covid en réanimation     - Nombre de décès lié au covid   + Note : Tout décès durant le séjour ou un patient est covid 19 confirmé est considéré lié au covid, tout décès en dehors d’un séjour de ce type n’est pas lié au covid * **Demographics.csv :**    + Une ligne par sexe   + Une colonne par tranche d’âge et la colonne du sexe :     - Sexe     - Nombre de patient de 0−2 ans,     - Nombre de patient de 3−5 ans,     - Nombre de patient de 6-11 ans,     - Nombre de patient de 12-17 ans,     - Nombre de patient de 18-25 ans,     - Nombre de patient de 26-49 ans,     - Nombre de patient de 50-69 ans,     - Nombre de patient de 70-79 ans,     - Nombre de patient de 80 ans et plus   + Note : Age à la date de la PCR * **Labs.csv :**    + 1 ligne par type d’analyse biologique et par nombre de jour depuis la PCR positive   + 5 colonnes par ligne :     - L’analyse biologique (LOINC code)     - Le nombre de jour depuis la PCR positive     - Le nombre de patient     - La valeur moyenne de l’analyse biologique     - L’écart type   + Notes : Analyses biologiques concernées :     1. Créatininémie (deux loinc groups possibles : LG50019-5 ou LG50024-5)     2. C-reactive protein (loinc : 1988-5)     3. D-Dimer (loinc : 48065-7)     4. Bilirubine totale (loinc : 1975-2)     5. Nombre de globule blanc (loinc group : LG32857-1) * **Diagnoses.csv :**    + 1 ligne par code CIM9 ou CIM10   + 2 colonnes :     - Code CIM9 ou CIM10     - Le nombre de patient avec ce diagnostic entre :       * Date de la PCR positive moins 1 semaine       * Aujourd’hui   + Note : Ne pas doublonner l’information entre CIM9 et CIM10   Produisez les 4 fichiers demandés à partir de votre jeu de données.  L'article de l'étude est disponible pour plus d’informations. |

### Bonus : Data Science sur le projet

|  |
| --- |
| Sur la base des fichiers générés dans l'étape précédentes, pour présenterez un notebook d'analyse et d'exploration des données exportées.  Le notebook décrira les tendances des données, leur tendances exploitables et proposera des visualisations graphiques de ces éléments.  Suggestions de contenu:   * Description de la cohorte (proportions, intervalles de confiance à 95%) * Comparaisons en fonction de variables discriminantes (diagnostiques par exemple). * Réalisez une analyse statistique des résultats observés   Justifiez les méthodes employées et discutez les résultats avancés. |

S’agissant de données synthétiques, il n’y a pas de nécessité à masquer les effectifs limités.

Gardez à l’esprit que les données utilisées simulent des données réelles, données dont la qualité est toujours variable. Notamment, la qualité du codage et la disponibilité des codes peut être partielle et variable. Toutefois, une grande partie d’entre eux peuvent être récupérés par le biais d’alignements ; la plateforme Athéna peut vous aider à cette fin.

Documentez dans votre rapport les étapes réalisées dans cette partie III et pour consignez pour chacune d’entre-elles l’approche adoptée. Justifiez vos choix et vos décisions.

# Modalités

Le projet est à réaliser en binôme (groupes de deux étudiants).

Le rapport de projet est à envoyer en amont de la soutenance, avant le 27 mai, 18h.

Soutenance le lundi 30 mai.

* 10 minutes de présentation
* Diapositives de support au format libre

# Historique des questions posées

## Partie I

* *Est-ce qu'il faut faire attention au type des données ? (faire une conversion de type en SQL)*

L’objectif de cette partie est de produire des spécifications d’ETL. Il s'agit d'un document technique permettant d’expliciter les flux et transformations nécessaires pour alimenter le format OMOP avec des données sources. Le typage des données est un élément majeur du travail de l'ETL. Pour un champ source et champ destination donnés, soit les types d'entrée et de destination sont identiques et l'ETL réalise une copie simple du champ ; soit ils diffèrent et l'ETL réalise une opération de conversion ; cette opération est alors expliquée (pour le mapping concerné).

Dans le cas de RabbitInAHat, cela consiste à documenter la logique liée à une flèche en particulier. Vous réalisez cela en remplissant le champ logique après sélection de la flèche de mapping précédemment créée entre un champ source et un champ destination.

* *Est-ce qu'il faut qu'on relie les champs qui finissent par \_id à leur instance de l'objet ?*

Attention, il n'est ici à aucun moment question d'instance d'objet. Uniquement de champs de valeurs et d'opération de remplissage et éventuellement de transformation.

Gardez également à l'esprit qu'il est fortement improbable qu'une source de données corresponde parfaitement à un modèle : si une information prévue par le modèle n'est pas présente dans le format des données sources, alors il ne saurait y avoir de mapping vers les champs correspondant. Inversement, les données sources peuvent contenir des éléments sans réalité au sein du modèle de destination ; elles ne seront alors tout simplement pas mappés. Le graphe de mapping ne saurait alors être un graphe complet.

* *Comment peut-on vérifier nos résultats ? (Tout est relié au moins une fois? etc ...)*

Vous pouvez vérifier visuellement la couverture de vos mappings en navigant dans l'outil, ou en une fois en consultant l'export Word de votre travail. Vous pouvez alors confronter votre travail avec votre objectif. Gardez toutefois en tête la réponse à la question précédente, l'objectif n'est pas nécessairement une affectation de tous les champs sources et tous les champs destination.