Atelier 2 - Ingestion des données

Dans la continuité du premier atelier, vous collectez les données nécessaires à la constitution de l'entrepôt modélisé précédemment.

Étape 1 - Outils et données pour la collecte

GitHub propose différents outils pour collecter des données :

- GitHub CLI : Ligne de commande produisant des données en JSON
- API REST : https://docs.github.com/fr/rest?apiVersion=2022-11-28
- GraphQL : https://docs.github.com/fr/graphql

Selon vos préférences et/ou votre spécialité, choisir un outils.

A partir de la modélisation du précédent atelier, parcourir la documentation pour identifier 5 à 10 tables de données brutes qui serviront, **après traitement**, à alimenter ce modèle.

Créer un dépôt git public et, dans une rubrique du readme, lister l'outil et les sources de données identifiées.

Étape 2 - Script-s de collecte

Automatiser la récupération de vos données au format `parquet` avec la combinaison d'outils/modules de votre choix, comme par exemple :

- Un ou plusieurs scripts shell utilisant curl ou GitHub CLI couplé à `json2parquet` (https://github.com/domoritz/arrow-tools)
- Un ou plusieurs scripts python appelant l'API REST ou GraphQL couplé à pyarrow ou polars pour produire des fichiers `parquet`
- ...

Historiser les scripts et les fichiers obtenus dans le dépôt git.

Étape 3 - "Lac" DuckDB

Écrire un script (shell utilisant DuckDB CLI ou python) qui charge ces données brutes sans transformation dans une base DuckDB en préfixant chaque table de `raw.`.

Historiser le script.

Remise

Par mail, le lien vers le dépôt GitHub public à sylvain.labasse@mail-formateur.net avant le 27/01.