# Premières requêtes SQL 🦆

## Introduction

Vous allez réaliser ce TP sur un Datalab.

- ☐ Connectez-vous au SSP Cloud

Pour ce premier TP, nous n'allons pas utiliser de base de données PostgreSQL.

Vous allez utiliser les logiciels CloudBeaver et DuckDB pour lire des fichiers de données.

### CloudBeaver

CloudBeaver est une application web légère dédiée à la gestion de bases de données. Elle permet de se connecter à divers types de bases de données, qu'elles soient SQL, NoSQL ou hébergées dans le cloud, à partir d'un seul point d'accès via un navigateur.

CloudBeaver facilite l'exploration, la modification et la visualisation des données sans nécessiter l'installation de logiciels locaux.

### DuckDB

DuckDB est un moteur de base de données relationnelle conçu pour des analyses rapides et efficaces.

DuckDB est idéal pour les charges de travail analytiques en raison de sa simplicité, de sa rapidité et de son extensibilité, surtout pour traiter des fichiers volumineux localement.

| Caractéristique | Description |
|---|---|
| **Simplicité** | Fonctionne sans serveur, intégré dans le processus hôte. Aucune dépendance externe pour la compilation ou l'exécution, simplifiant son déploiement. |
| **Rapidité** | Optimisé pour l'analytique (OLAP) avec un moteur d'exécution vectorisé en colonnes, réduisant le temps de traitement pour les requêtes complexes. |
| **Richesse fonctionnelle** | Supporte des requêtes SQL complexes, fonctions de fenêtre, index secondaires, et assure des garanties ACID grâce au contrôle de concurrence (MVCC). |
| **Extensibilité** | Permet l'ajout de types de données, fonctions et formats de fichiers via des extensions (supporte Parquet, JSON, S3, HTTP(S)). |
| **Gratuit et open-source** | Licence MIT, code source disponible et contributions ouvertes à tous. |

> Tip
>
> Vous allez créer une base de données stockée en mémoire vive. L'avantage est que c'est très rapide. L'inconvénient, c'est que rien ne sera sauvegardé.

# 1. Lancement du service

- ☐ Lancez le service CloudBeaver
  - Allez dans **Catalogue de services** ➡️ Onglet **Databases**
- ☐ Ouvrez le service
  - Renseignez username et password

Une fois dans CloudBeaver, vous allez créer une connexion DuckDB :

- ☐ Créez une nouvelle connexion **DuckDB**
  - En haut à gauche, cliquez sur le +, puis New connection
  - Sélectionner **DuckDB**
  - Database : `:memory:`
  - Cliquez sur le bouton CREATE
- ☐ Ouvrez un éditeur SQL
  - Dans l'explorer à gauche, clic droit sur votre connexion *DuckDB* ➡️ SQL Editor
  - vous pouvez maintenant saisir du code *SQL*

> 🔥 Caution
>
> Votre connexion doit être ouverte 🟢.
>
> Si ce n'est pas le cas : clic droit > Open.
>
> Puis réouvrez une nouvelle fenêtre SQL

# 2. Les prénoms

Le fichier des prénoms contient des données sur les prénoms attribués aux enfants nés en France depuis 1900. Ces données sont disponibles au niveau France et par département.

Dans ce TP, nous allons utiliser la version au format Parquet.

> **Fichier parquet**
>
> Le format **Parquet** est un format de fichier de stockage de données optimisé pour les systèmes de traitement analytique de grande échelle. Voici ses principales caractéristiques :
>
> 1. **Stockage en colonnes** : Parquet stocke les données par colonnes plutôt que par lignes, ce qui améliore l'efficacité de l'accès aux données dans les charges de travail analytiques.
> 2. **Compression efficace** : La compression par colonne permet un taux de compression moyen de 5 à 10 fois par rapport aux formats CSV, voire plus pour de gros fichiers. Cela réduit significativement la taille des données stockées et le coût de stockage.
> 3. **Optimisé pour l'analytique** : Parquet est conçu pour les requêtes en lecture intensive, car il permet de charger uniquement les colonnes nécessaires pour une analyse, ce qui améliore les performances en particulier sur les données volumineuses.
> 4. **Métadonnées Riches et Auto-descriptives** : Parquet inclut des métadonnées détaillées (schéma, types de données, statistiques min/max), permettant une lecture rapide et sans risque d'erreur. Ces métadonnées facilitent également le traitement des données par différents outils.
> 5. **Langage Indépendant et Open Source** : Le format Parquet est indépendant du langage et peut être utilisé avec divers langages de programmation (Python, R, C++, Java). Il est aussi open source et compatible avec la plupart des frameworks de big data.

Pour en savoir plus sur le format Parquet :

- Parquet, qu'est-ce que c'est ?, Databricks
- What is the Parquet File Format?, Upsolver blog

Ici le fichier est relativement modeste (10 Mo), nous pouvons nous permettre de le stocker intégralement en mémoire.

Pour des requêtes plus rapides, nous allons forcer le téléchargement en cache du fichier :

- ☐ Exécutez `SET force_download=true;`

> **Éxécuter une requête**
>
> Cliquez sur la requête pour y positionner le pointeur de la souris, puis au choix :
>
> - Cliquez sur le petit triangle orange ▶
> - CTRL + ENTREE

- ☐ Créez une table `prenoms` à partir du contenu du fichier

```
CREATE OR REPLACE TABLE prenoms AS
FROM 'https://www.insee.fr/fr/statistiques/fichier/8595130/prenoms-2025.parquet';
```

> **Note**
>
> Cette requête est un peu longue et peut prendre parfois jusque 30s.
>
> Si le lien n'est plus valide, allez chercher la dernière version sur le site de l'INSEE.

Vous pourrez ensuite requêter sur cette vue comme si c'était une table.

## 2.1. Premières requêtes

! Prenez le temps et l'habitude d'écrire de jolies requêtes bien alignées !!!

```sql
SELECT c.nom AS club,
       AVG(j.elo) AS moyenne_elo
  FROM joueuse j
 INNER JOIN club c USING (id_club)
 WHERE j.mail IS NOT NULL
 GROUP BY c.nom
 ORDER BY 2 DESC;
```

```sql
SELECT c.nom AS club, AVG(j.elo) AS moyenne_elo FROM joueuse j INNER JOIN club c
USING (id_club) WHERE mail IS NOT NULL GROUP BY c.nom ORDER BY 2 DESC;
```

Dictionnaire des variables

Avant de vous lancer tête baissée, ce qui serait pas mal c'est d'aller jeter un oeil au dictionnaire des variables.

Vous le trouverez sur le site de l'INSEE.

- ☐ Listez tous les éléments de *prenoms*
- ☐ Filtrez pour n'afficher que le niveau géographique *FRANCE* et le prénom *GINETTE*

i Secret statistique

Vous remarquerez que les valeurs sont des multiples de 5.

> Pour chaque prénom, il est indiqué, pour chaque année de naissance et chaque sexe, le nombre de personnes inscrites à l'état civil sous ce prénom, arrondi au multiple de 5 le plus proche.

Vous remarquerez également que certaines années sont manquantes.

> Un prénom est diffusé s'il a été attribué au moins 3 fois à des personnes de sexe féminin ou à des personnes de sexe masculin.

En statistique publique, on exige au moins 3 individus pour garantir le secret statistique. Avec 1 ou 2 personnes, on peut identifier ou déduire leurs données.

Pour **la suite du TP**, nous allons uniquement nous interesser au niveau géographique *FRANCE* :

- ☐ À partir de la table *prenoms*, créez une table *prenoms_fr* conservant uniquement :
  - les lignes où le niveau géographique est égal à *FRANCE*
  - les colonnes suivantes : *sexe*, *prenom*, *periode*, *valeur*

- ☐ Combien de lignes contient cette table ?

## 2.2. Année 2022

Filtrez sur l'année *2022*

- ☐ Listez les prénoms de l'année 2022
  - Classez-les en décroissant par le nombre de fois où ils ont été donnés
- ☐ Affichez uniquement les colonnes *sexe*, *prenom* et *valeur*
  - Affichez également ces colonnes pour les questions suivantes
- ☐ Listez les prénoms de l'année 2022 donnés 2000 fois ou plus
- ☐ Classez-les d'abord par sexe, puis par nombre décroissant d'occurrences
  - Quels sont les prénoms masculins et féminins les plus donnés ? 🏅
- ☐ Listez les prénoms féminins commençant par la lettre *Q*
- ☐ Listez les prénoms contenant exactement deux fois la lettre *U*
- ☐ Comptez le nombre de prénoms commençant par chaque lettre

> **Tip**
>
> Pour extraire la première lettre d'un prénom, vous pouvez utiliser :
>
> - la fonction SUBSTRING
> - le slicing comme en R ou Python

- ☐ Bonus : Même question en ignorant les accents
- ☐ Bonus 2 : Comptez le nombre de prénoms commençant par chaque lettre en différenciant filles et garçons

## 2.3. Statistiques descriptives

- ☐ Lancez la commande `SUMMARIZE prenoms;` pour avoir quelques stats sur la table
- ☐ Affichez pour l'année 2003, les prénoms et leurs nombres de caractères
  - Trouvez une fonction adéquate
- ☐ Affichez le nombre de caractères minimum, maximum et moyen parmi les prénoms de 2003
- ☐ Listez les 3 prénoms de 2003 ayant le nombre de caractères maximum
  - Commencez par écrire une requête donnant le nombre de caractères maximum
  - Utilisez-là comme sous-requête pour afficher les 3 prénoms
- ☐ Modifiez cette dernière requête pour afficher pour chaque année entre 2015 et 2022 le ou les prénoms avec le plus de caractères
  - classez par année
  - Difficulty : ⭐ ⭐ ⭐
- ☐ Affichez pour chaque année la taille moyenne des prénoms
  - classez par année
  - utilisez la fonction `ROUND(<value>, 2)` pour arrondir les moyennes
  - Que remarquez-vous ?

## 2.4. Jérôme, c'est moi

- ☐ Listez les différentes années où le prénom `JÉRÔME` a été donné
  - Classez par année décroissante
  - Que remarquez-vous ?
- ☐ Quelles sont les différentes manières d'écrire ce prénom (avec des accents différents) ?
- ☐ Listez également les `JEROME` avec d'autres accents / sans accent
  - Recherchez une fonction DuckDB qui permet de répondre à ce besoin
- ☐ Affichez pour chaque année le nombre de fois où le prénom JEROME a été donné, quelle que soit l'accentuation

> Aide
>
> - voir Fonction DuckDB strip_accents()
> - Regroupez par strip_accents(prenom) et periode
> - Affichez strip_accents(prenom), periode et la somme de la colonne *valeur*

## 2.5. Vérifions le classement

Avec cette histoire d'accents, posons-nous la question si cela peut modifier le classement des top prénoms. 🤔

- ☐ Affichez le top 10 des prénoms féminins en 2023 ?
- ☐ Quel prénom se classe onzième ?
  - Y-a-t-il d'autres orthographes proches pour écrire ce prénom ?
- ☐ Refaites le classement en ignorant les accents

## 2.6. Prénoms composés

Nous allons maintenant nous interesser aux prénoms composés entre 2000 et 2009

- ☐ Listez les différents prénoms composés donnés entre les années 2000 et 2009 incluses
- ☐ Regroupez-les par *prenom*, puis affichez également le nombre de fois où ils ont été donnés
  - Triez en décroissant par ce nombre
- ☐ Filtrez pour ne garder que les prénoms composés contenant le prénom `JEAN`
  - Que remarquez-vous ? N'y a-t-il pas des intrus ?
- ☐ Trouvez une solution pour remédier à ce souci

## 2.7. Cette année-là

*Difficulty* : ⭐ ⭐ ⭐

Filtrez sur l'année 1962.

- ☐ Comptez le nombre de prénoms distincts donnés
- ☐ Complétez en comptant le nombre de prénoms distincts donnés par sexe
- ☐ Complétez en calculant le nombre total des prénoms distincts donnés, en prenant en compte la distinction entre les sexes
  - exemple : ici *Dominique (H)* et *Dominique (F)* comptent pour deux prénoms
- ☐ Donnez la liste des prénoms donnés à la fois à des filles et des garçons

    - affichez le nombre de fois où ils ont été donnés à chaque sexe
- ☐ Ajoutez à la dernière requête une colonne contenant un booléen qui affiche
    - True si le prénom a été donné à plus de filles que de garçons
    - False sinon
- ☐ Au lieu d'afficher ce booléen, ajoutez une condition pour n'afficher que lorsque le prénom a été plus donné à des filles

## 3. Fichier des individus

Les requêtes peuvent être un peu lentes.

Dans cet exercice, vous allez utiliser le fichier détail des individus du recensement de la population 2020.

- ☐ Créez une vue qui pointe vers le fichier

```
CREATE OR REPLACE VIEW individus AS
FROM    'https://static.data.gouv.fr/resources/recensement-de-la-population-fichiers-detail-individus-localises-au-canton-ou-ville-2020-1/20231023-122841/fd-indcvi-2020.parquet';
```

- ☐ Comptez le nombre d'individus
    - Pourquoi ce nombre d'individus vivant en France paraît faible ?

> Poids de l'individu
>
> Dans de nombreuses bases de données de l'INSEE, la variable `ipondi` représente le poids de l'individu.
>
> i.e. Si une ligne a un IPONDI égal à 2, cela signifie qu'elle représente 2 individus.

- ☐ Au lieu de compter le nombre de lignes, sommez la variable `ipondi`
    - Vous pouvez suffixer votre somme par `::INT` pour convertir ce nombre en entier
- ☐ Affichez 10 lignes
    - Que remarquez-vous ?

> Codification
>
> Pourquoi utiliser une codification pour le fichier national individus localisés au canton-ou-ville du recensement de la population 2020 ?
>
> - **Espace de stockage réduit** : Les codes courts remplacent des chaînes de texte plus longues, ce qui diminue la taille du fichier et rend le stockage et le transfert de données plus efficaces. Même après codification, le fichier pèse plus de 500 Mo
> - **Standardisation et uniformité** : Utiliser des codes standardisés permet de garantir une structure homogène et facilite l'intégration et la comparaison des données entre différents fichiers ou études.
> - **Performances de traitement** : Les codes sont plus rapides à traiter que des textes longs, notamment pour des opérations de recherche, de tri, et de filtrage, ce qui est crucial pour les grands ensembles de données.

- ☐ Pour s'y retrouver, créez une vue vers le dictionnaire des variables

```
CREATE OR REPLACE VIEW variables_individus AS
FROM    'https://static.data.gouv.fr/resources/recensement-de-la-population-fichiers-detail-individus-localises-au-canton-ou-ville-2020-1/20231025-082910/dictionnaire-variables-indcvi-2020.csv'
```

- ☐ Affichez tout le contenu de cette vue
- ☐ Trouvez la modalité représentant le département de résidence de l'individu
- ☐ Affichez le nombre d'habitants par départements
  - ordonnez par numéro de département
- ☐ Complétez la requête en :
  - restreignant sur les individus entre 25 et 29 ans
  - différenciant les hommes et les femmes

Solutions

- Correction TP1

## Arrêtez votre service

C'est la fin du TP, vous pouvez maintenant sauvegarder votre travail et libérer les ressources réservées :

- ☐ Copiez votre code et collez le sur votre machine dans un fichier *.sql*
  - par exemple dans `P:/Cours1A/UE3-Bases-de-donnees-relationnelles/TP1/tp1.sql`
- ☐ Retournez sur la page d'accueil du Datalab
- ☐ Allez dans *Mes Services* et supprimez votre service

## Pour aller plus loin

- 3 explorations bluffantes avec DuckDB
- Avec DuckDB, gavez-vous d'open data!
- DuckDB, un aperçu du futur !