NBA stats 🏀

Introduction

Vous allez travailler sur des données de la nba depuis les années 2000.

Les données sont brutes, i.e. elles ont été récupérées via l'api stats nba et elles n'ont quasiment pas été retouchées.

Le modèle de données n'est pas fourni ! Ce sera à vous d'explorer vous même la base de données.

> Tip
>
> Dans votre vie professionnelle, dans 95 % des cas vous allez hériter de missions où la documentation est obsolète voire inexistante.
>
> Il faut donc apprendre à vous débrouiller dès maintenant. Et au passage, vous convaincre que faire et maintenir la doc ça aide et c'est bien !

1. Lancement des services

- ☐ Connectez-vous au SSP Cloud
- ☐ Lancez un service **PostgreSQL**
- ☐ Lancez un service **cloudBeaver**
  - ouvrez ce service et ouvrez la connexion PostgreSQL

En cas de besoin, vous trouverez plus d'informations dans l'onglet Datalab.

À la fin du TP

N'oubliez pas de sauvegarder votre travail et libérer les ressources :

- ☐ Sauvegardez votre travail
  - Copiez-collez vos scripts SQL sur votre machine
- ☐ Mettez en pause ou Arrêter les services du Datalab

2. Importer la base de données

Aujourd'hui, pas de script sql pour créer les tables et importer les données, vous allez utiliser un **dump**.

> **i** Définition
>
> Un dump de base de données est un fichier qui contient une copie complète d'une base de données.
>
> Il sert à :
>
> - Sauvegarder une base de données
> - Transférer une base vers un autre serveur
> - Reconstituer la base à l'identique (tables, données, index, etc.)
>
> En bref : c'est une photographie de la base de données à un instant donné, utilisable pour la recréer ailleurs.

Le fichier *nba.dump* est disponible sur le stockage S3 de l'enseignant.

Pour récupérer ce fichier et l'utiliser pour votre base de données :

- ☐ Suivez les instructions du README de ce répo : https://github.com/ludo2ne/SQL-NBA-POSTGRESQL/
  - ⚠️ Si vous êtes sur le datalab du GENES
  - ↪️ Avant de lancer le script `5_import_for_students`
  - ↪️ Dans le FILE_PATH, remplacez *ludo2ne* par *ldeneuville-ensai*
- ☐ Une fois les données récupérées, arrêtez le service *Vscode-python*

> **i** Explications
>
> Le script *5_import_for_students.py* :
>
> - Se connecte au S3 de l'enseignant pour récupérer le fichier *nba.dump*
>   - Vous remarquerez que le fichier est dans un dossier nommé *diffusion*
>   - Ce dossier est accessible en lecture aux autres utilisateurs
> - Utilise les variables d'envionnement pour se connecter à votre base de données
>   - Celles que vous avez renseigné dans le fichier *.env*
> - Utilise l'utilitaire *pg_restore* pour créer la base de données à partir du dump

3. Rendez-vous en base inconnue

Je n'ai pas réussi à inviter Laury T. pour vous accompagner dans cette aventure mais voici quelques requêtes utiles pour explorer une base de données PostgreSQL.

```sql
-- Liste des tables du schema
SELECT table_name
  FROM information_schema.tables
 WHERE table_schema = '<schema_name>'
   AND table_type = 'BASE TABLE'
 ORDER BY table_name;

-- Colonnes d une table
SELECT column_name,
```

```
       data_type,
       is_nullable,
       column_default
  FROM information_schema.columns
 WHERE table_schema = '<...>'
   AND table_name   = '<...>'
 ORDER BY ordinal_position;

-- Colonnes commencant ou terminant par id
SELECT table_name,
       column_name,
       data_type
  FROM information_schema.columns
 WHERE table_schema = '<...>'
   AND (column_name ILIKE 'id%' OR column_name ILIKE '%id')
ORDER BY table_name, column_name;

-- Generer les CREATE TABLE basiques
SELECT 'CREATE TABLE ' || table_schema || '.' || table_name || ' ( ' || E'\n' ||
       STRING_AGG(
           '  ' || column_name || ' ' || UPPER(data_type),
           ',' || E'\n'
           ORDER BY ordinal_position
       )
       || E'\n);' AS create_table_script
  FROM information_schema.columns
 WHERE table_schema = '<...>'
 GROUP BY table_schema, table_name
 ORDER BY table_name;
```

Requêtes plus complètes

Liste des colonnes des tables d'un schéma

```
WITH
params AS (
    SELECT
        '<...>' AS schemaname
),
cols AS (
    SELECT
        c.table_schema,
        c.table_name,
        c.column_name,
        c.ordinal_position,
        c.is_nullable,
        c.udt_name,
        c.character_maximum_length,
        c.column_default
    FROM information_schema.columns c
    JOIN params p ON c.table_schema = p.schemaname
),
pks AS (
    SELECT DISTINCT
        kcu.table_schema,
        kcu.table_name,
        kcu.column_name
    FROM information_schema.table_constraints tc
    JOIN information_schema.key_column_usage kcu
         ON tc.constraint_name = kcu.constraint_name
        AND tc.table_schema = kcu.table_schema
    JOIN params p ON tc.table_schema = p.schemaname
    WHERE tc.constraint_type = 'PRIMARY KEY'
),
uniques AS (
    SELECT DISTINCT
        kcu.table_schema,
        kcu.table_name,
        kcu.column_name
    FROM information_schema.table_constraints tc
    JOIN information_schema.key_column_usage kcu
         ON tc.constraint_name = kcu.constraint_name
        AND tc.table_schema = kcu.table_schema
    JOIN params p ON tc.table_schema = p.schemaname
    WHERE tc.constraint_type = 'UNIQUE'
),
fks AS (
    SELECT DISTINCT ON (kcu.table_schema, kcu.table_name, kcu.column_name)
```



```
WITH     kcu.column_name,
params AS cu.table_name AS ref_table,
```

i Glossaire

Colonne	Signification
matchup	Description du match : ex. `LAL @ BOS` (à l'extérieur) ou `LAL vs BOS` (à domicile).
wl	Résultat pour l'équipe : `W` (win), `L` (loss).
min	Minutes totales jouées par l'équipe (toujours `240` sauf prolongation).
pts	Total des points de l'équipe.
fgm	Field goals made — tirs réussis (2 pts + 3 pts).
fga	Field goals attempted — tirs tentés.
fg_pct	Taux de réussite : `fgm / fga`.
fg3m	3-points réussis.
fg3a	3-points tentés.
fg3_pct	Pourcentage à 3 points.
ftm	Free throws made — lancers francs réussis.
fta	Free throws attempted — lancers francs tentés.
ft_pct	Pourcentage aux lancers.
oreb	Rebonds offensifs.
dreb	Rebonds défensifs.
reb	Rebonds totaux (oreb + dreb).
ast	Passes décisives.
stl	Interceptions.
blk	Contres.
tov	Ballons perdus (turnovers).
pf	Fautes personnelles.

Si un modèle de données est nécessaire pour vous, vous pouvez par exemple :

- Générer les CREATE TABLE
- Demander à une IAgen de vous générer le modèle
- Par exemple au format utilsé par ce site : https://dbdiagram.io/

4. Explorons avec les Spurs

Tip

Vous allez avoir parfois des incompatibilités de types.

Vous pouvez utiliser par exemple `<colname>::int` ou `<colname>::str` pour caster.

- ☐ Quel est l'id de l'équipe des Spurs de San Antonio ?
- ☐ Listez les joueurs des Spurs

- ☐ Listez tous les matchs des Spurs de la saison régulière *2024-2025*
  - classez ces matchs par date
- ☐ Combien de lignes correspondent à chaque *game_id* dans la table *game* ?
  - Choisissez des exemples de *game_id* et vérifiez
  - Ou mieux écrivez une requête qui vérifie l'hypothèse pour tous les *game_id*

Ce choix de conception a des avantages et des inconvénients. Nous y reviendrons par la suite lorsque nous afficherons les scores de chaque match.

- ☐ Proposez une clé primaire pour la table *game*
- ☐ Ecrivez une requête pour vérifier si votre proposition est correcte

Regardons maintenant les stats des joueurs.

- ☐ Affichez les stats des joueurs des Spurs pour la saison régulière *2024-2025*
- ☐ Affichez les stats des joueurs des Spurs par match
- ☐ Est-ce qu'il n'y aurait pas des doublons dans cette table ?
- ☐ Est-ce que cette table contient les stats des joueurs par match de toutes les saisons ?

5. Afficher les matchs

Dans toute cette partie nous nous interesserons à la saison *2020-2021*.

- ☐ Listez les matchs de la saison
- ☐ Combien y-a-t-il eu de match cette saison
- ☐ Distinguez saison régulière et Playoffs

Interessons-nous la saison régulière.

- ☐ Listez les matchs de *Utah Jazz*
- ☐ Comptez leur nombre de victoires et de défaites
- ☐ Utilisez COUNT et FILTER pour avoir le résultat sur une seule ligne
  - ajoutez le nombre de matchs joués
- ☐ Affichez les mêmes colonnes pour chaque équipe de la conférence *Ouest*
  - classez par nombre de victoires décroissant

Passons maintenant aux Playoffs.

- ☐ Listez les matchs de *Utah Jazz*
- ☐ Pour chacun de leurs matchs, affichez le score, le nom de leurs adversaires et le résultat
- ☐ Affichez les scores de matchs

Une dernière question pour la route :

- ☐ Si vous deviez afficher toutes les finales et leur score, comment feriez-vous ?

Aide

- ☐ Quelles étapes sont réalisées ci-dessous ?

Tout simplement :

- Trouvez le dernier match des playoffs i.e. celui qui a le plus grand id
- Déduisez l'équipe ayant gagné le dernier match et la finaliste
- Enfin listez tous les matchs de la finales pour avoir le score

```
WITH
last_playoff_games AS (
SELECT season_id, MAX(game_id) AS last_game_id
  FROM nba.game
 WHERE season_type = 'Playoffs'
 GROUP BY season_id
),
final_teams AS (
SELECT g.season_id,
       g.team_id AS team_id_win,
       g2.team_id AS team_id_los
  FROM nba.game g
   JOIN last_playoff_games lpg ON g.season_id = lpg.season_id AND g.game_id =
lpg.last_game_id
  JOIN nba.game g2 ON g2.game_id = g.game_id AND g2.team_id <> g.team_id
 WHERE g.wl = 'W'
)
SELECT gwin.season_id,
       twin.full_name AS equipe,
        COUNT(*) FILTER (WHERE gwin.wl = 'W') || ' - ' || COUNT(*) FILTER (WHERE
gwin.wl = 'L') AS score,
       tlos.full_name AS adversaire
  FROM nba.game gwin
    JOIN final_teams ft ON ft.team_id_win = gwin.team_id AND ft.season_id =
gwin.season_id
  JOIN nba.team twin ON twin.id = gwin.team_id
  JOIN nba.game glos ON glos.game_id = gwin.game_id AND glos.team_id = ft.team_id_los
  JOIN nba.team tlos ON tlos.id = glos.team_id
 WHERE gwin.season_type = 'Playoffs'
 GROUP BY gwin.season_id, twin.full_name, tlos.full_name
 ORDER BY gwin.season_id DESC;
```

6. Triple zéro

Lors d'un match, les principales statistiques utilisées pour mesurer les performances des joueurs sont : les points, les rebonds, les passes décisives, les interceptions et les contres.

Un joueur réalise alors un :

- **double-double** : s'il atteint au moins 10 dans deux de ces statistiques
- **triple-double** : au moins 10 dans trois de ces statistiques (grosse perf)
- **quadruple-double** : au moins 10 dans quatre de ces statistiques (très rare)

Il existe également d'autres performances très rares :

- **double triple-double** : s'il atteint au moins 20 dans trois de ces statistiques
- **Five-by-five** : au moins 5 dans les cinq catégories

Vous aussi en avez assez d'entendre toujours les mêmes noms réaliser ce genre de performance : Wemby, Curry, Gilgeous-Alexander, Doncic, Jokic…

L'heure de la revanche a sonné. Mettons aujourd'hui en lumière les vrais champions, ceux qui réalisent des 0 dans ces stats, font des fautes, perdent des balles.

- ☐ Affichez les statistiques par match de chaque joueur
- ☐ Conservez uniquement ceux ayant joué plus de 10 minutes par match

> Tip
>
> - Commencez par retirer les lignes ou la colonne *minutes* est vide
> - Puis utilisez `CAST(SPLIT_PART(minutes, ':', 1) AS integer)`

- ☐ En plus du nom du joueur et du nombre de minutes jouées, affichez les cinq statistiques demandées
- ☐ Est-ce que nous avons un champion avec un *quintuple zéro* ?
- ☐ ⭐ ⭐ ⭐ Un peu plus compliqué, est-ce qu'il y a des quadruples zéros ?
  - *quadruple zéro* : 0 points et 3 autres stats à 0
  - Le `CASE WHEN` est votre ami ici

Quelques autres titres à décerner (toujours avec au moins 10min de jeu):

- ☐ Celui qui a le plus souvent marqué 0 points
- ☐ Qui est le pire shooter d'un match ?
  - i.e. qui n'a marqué aucun point, et a tenté le plus grand nombre de tirs, lancers francs inclus
- ☐ Liste des joueurs qui lors d'un match ont eu plus de pertes de balle (turnover) que de points marqués
- ☐ Top 10 de ceux qui ne terminent pas souvent les matchs (ils sont expulsés après 6 fautes personnelles)

La colonne *plusminuspoints* est la différence entre le nombre de points marqués par l'équipe et le nombre de points encaissés lorsque le joueur est sur le terrain.

Cela donne une idée de l'influence et l'impact du joueur lorsqu'il est sur le terrain.

- ☐ Quel joueur a eu la pire influence sur un seul match
- ☐ Quels sont les joueurs qui ont le plus grand nombre de matchs avec une influence négative sur l'équipe

Ils ont essayé de faire perdre leur équipe mais ils n'ont pas réussi :

- ☐ Quel joueur a eu la pire influence, et pourtant son équipe a gagné
- ☐ Ajoutez les deux équipes et le score du match

7. Classement par saison

Pour occuper les plus rapides 🚀

- ☐ Affichez le classement de la conférence OUEST de la saison régulière 2008-2009
  - Triez par pourcentage de victoires
  - En cas d'égalité, vous pouvez si vous le souhaitez trouver les autres critères de déparatage mais cela complique grandement la requête
- ☐ Pour chaque saison, affichez l'équipe gagnante, ainsi que son classement dans sa conférence lors de la saison régulière
  - Procédez par étapes
- ☐ Affichez combien d'équipes classées 3e lors de la saison régulière ont gagné
- ☐ Quelle équipe a réussi le meilleur bilan sur une saison (meilleur pourcentage de victoire)
- ☐ Classement des équipes ayant le plus gagné / perdu des matchs par moins de 5 points d'écarts

Pour chaque saison régulière donnez le match :

- ☐ Avec le plus grand écart de points
- ☐ Avec le plus de points marqués (ne pas inclure les éventuelles prolongations)
- ☐ Avec le moins de points marqués

Solutions

- Correction TP5