Homework Lab

Introduction to Big Data

Jie FAN
Minh Quang HOANG

Sommaire

Introduction

Contexte

Dans le cadre académique, les étudiants sont confrontés à de multiples pressions, notamment des attentes élevées en matière de performance scolaire. Ces exigences, associées à des facteurs comme les troubles de santé mentale, les habitudes de sommeil, et les niveaux de stress, peuvent avoir un impact significatif sur leur bien-être et leur réussite académique.

Pour mieux comprendre ces interrelations, notre projet s'appuie sur des données réelles pour examiner comment le stress, les troubles mentaux, et d'autres facteurs liés à la vie étudiante influencent les performances scolaires. En analysant des aspects tels que le sommeil, les niveaux de stress, et les diagnostics de santé mentale, nous cherchons à proposer des stratégies pour améliorer la santé mentale et les résultats scolaires.

Objectif

L'objectif principal de cette étude est d'analyser et de comprendre l'impact de la santé mentale et des facteurs de stress sur les performances académiques des étudiants. En exploitant les données disponibles, nous visons à identifier les schémas et corrélations clés pour proposer des solutions adaptées.

Les sous-objectifs incluent :

- Analyser les niveaux de stress et leurs relations avec les performances scolaires.
- Évaluer l'impact des troubles mentaux et des habitudes de sommeil sur la concentration et les résultats.
- Fournir des recommandations pour réduire les niveaux de stress et améliorer la santé mentale des étudiants

Sources de données

Description datasets

Voici un aperçu des colonnes présentes dans chaque fichier, avec une explication de leurs significations et des valeurs qu'elles contiennent :

1. StressLevelDataset.csv

Ce fichier contient des informations liées au stress, aux conditions de vie et à d'autres facteurs qui peuvent influencer la santé mentale. Voici les colonnes et leur signification :

- **`anxiety_level`** : Niveau d'anxiété sur une échelle de 1 à 20 (valeurs plus élevées indiquent un niveau d'anxiété plus important).
- **`self_esteem`** : Niveau d'estime de soi sur une échelle de 1 à 30 (valeurs plus élevées indiquent une estime de soi plus élevée).
- **`mental_health_history`** : Historique de troubles mentaux (0 = aucun historique, 1 = historique présent).
- **`depression`** : Niveau de dépression sur une échelle de 1 à 15.
- **`headache`** : Fréquence des maux de tête (1 = rare, 5 = très fréquent).
- **`blood_pressure`** : Niveau de tension artérielle (1 = normal, 5 = très élevé).
- **`sleep_quality`** : Qualité du sommeil sur une échelle de 1 à 5 (1 = très mauvaise, 5 = excellente).
- **`breathing_problem`** : Problèmes respiratoires perçus (1 = aucun, 5 = sévère).
- **`noise_level`** : Niveau de bruit dans l'environnement (1 = très calme, 5 = très bruyant).
- **`living_conditions`** : Conditions de vie générales (1 = très mauvaises, 5 = excellentes).
- **`basic_needs`** : Niveau de satisfaction des besoins de base (1 = non satisfait, 5 = entièrement satisfait).
- **`academic_performance`** : Performance académique (1 = très faible, 5 = très élevée).
- **`study_load`** : Charge d'étude perçue (1 = très légère, 5 = très lourde).
- **`teacher_student_relationship`** : Qualité de la relation enseignant-étudiant (1 = mauvaise, 5 = excellente).
- **`future_career_concerns`** : Niveau de préoccupation concernant la carrière future (1 = aucune, 5 = très préoccupé).
- **`social_support`** : Niveau de soutien social perçu (1 = aucun, 5 = très élevé).
- **`peer_pressure`** : Pression des pairs (1 = très faible, 5 = très forte).
- **`extracurricular_activities`** : Niveau d'implication dans les activités extrascolaires (1 = aucune, 5 = très élevée).
- **`bullying`** : Fréquence du harcèlement (1 = jamais, 5 = très fréquent).

- **`stress_level`** : Niveau global de stress sur une échelle de 1 à 10 (1 = très faible, 10 = très élevé).

2. Student Mental health.csv

Ce fichier documente les troubles mentaux déclarés et quelques informations démographiques. Voici les colonnes et leur signification :

- **`Timestamp`** : Heure et date de la réponse au questionnaire.
- **`Choose your gender`** : Genre de l'étudiant (Male = masculin, Female = féminin, Other = autre).
- **`Age`** : Âge de l'étudiant en années.
- **`What is your course?`** : Programme d'études suivi par l'étudiant.
- **`Your current year of Study`** : Année actuelle d'études (exemple : Year 1, Year 2).
- **`What is your CGPA?`** : Moyenne générale cumulée (exemple : 3.00 - 3.49).
- **`Marital status`** : Statut marital de l'étudiant (Yes = marié, No = non marié).
- **`Do you have Depression?`** : L'étudiant déclare-t-il souffrir de dépression (Yes = oui, No = non).
- **`Do you have Anxiety?`** : L'étudiant déclare-t-il souffrir d'anxiété (Yes = oui, No = non).
- **`Do you have Panic attack?`** : L'étudiant déclare-t-il avoir des crises de panique (Yes = oui, No = non).
- **`Did you seek any specialist for a treatment?`** : L'étudiant a-t-il consulté un spécialiste (Yes = oui, No = non).

3. Student_sleep_patterns.csv

Ce fichier explore les habitudes de sommeil des étudiants et leurs modes de vie. Voici les colonnes et leur signification :

- **`Student_ID`** : Identifiant unique de chaque étudiant.
- **`Age`** : Âge de l'étudiant.
- **`Gender`** : Genre de l'étudiant.
- **`University_Year`** : Année d'université (exemple : 1st Year, 2nd Year).
- **`Sleep_Duration`** : Nombre moyen d'heures de sommeil par jour.
- **`Study_Hours`** : Heures consacrées à l'étude par jour.
- **`Screen_Time`** : Temps moyen passé devant un écran par jour (en heures).
- **`Caffeine_Intake`** : Quantité de consommation de caféine (échelle de 1 à 5, 1 = faible, 5 = élevée).
- **`Physical_Activity`** : Durée ou fréquence d'activité physique (mesurée en minutes ou sessions).

- **Sleep_Quality** : Qualité perçue du sommeil (échelle de 1 à 10).
- **Weekday_Sleep_Start** : Heure moyenne de coucher en semaine (exemple : 22.30 = 22h30).
- **Weekend_Sleep_Start** : Heure moyenne de coucher le week-end.
- **Weekday_Sleep_End** : Heure moyenne de réveil en semaine.
- **Weekend_Sleep_End** : Heure moyenne de réveil le week-end.

4. StudentPerformanceFactors.csv

Ce fichier regroupe des facteurs influençant les performances académiques. Voici les colonnes et leur signification :

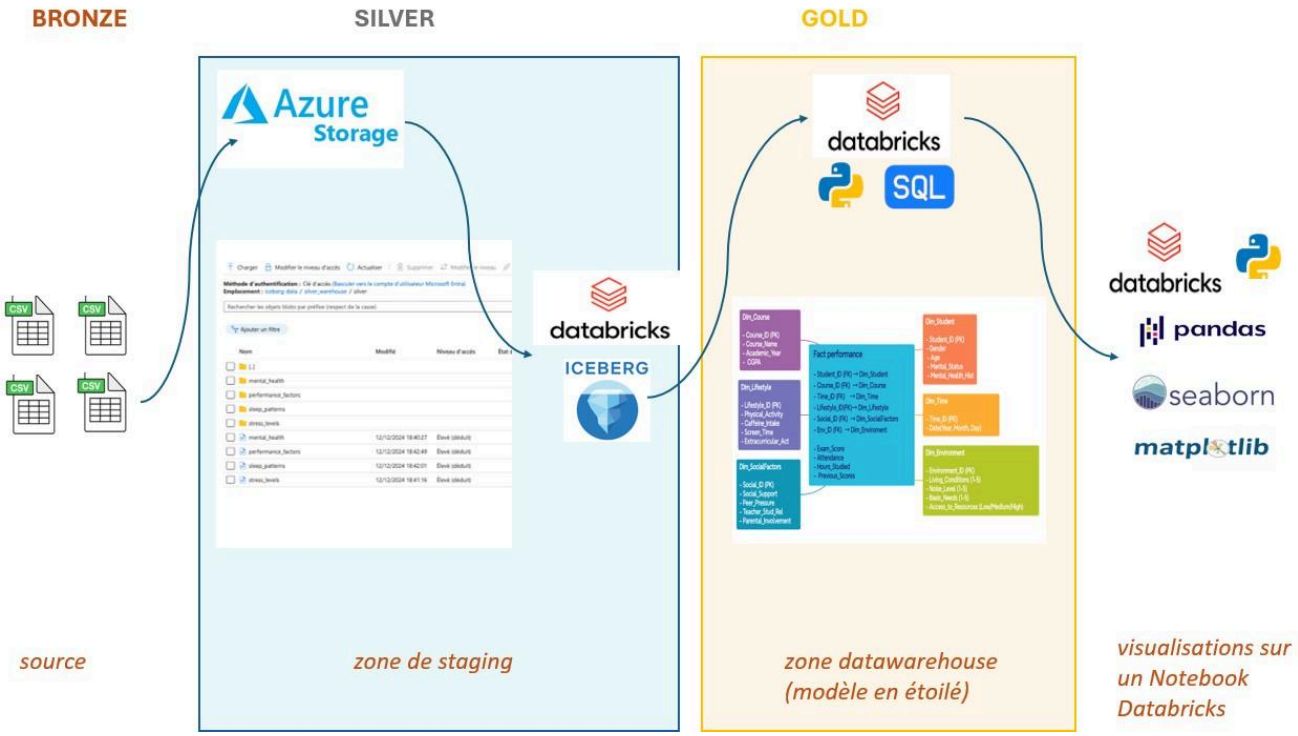
- **Hours_Studied** : Heures consacrées à l'étude chaque semaine.
- **Attendance** : Taux de présence en classe (en pourcentage).
- **Parental_Involvement** : Niveau d'implication des parents (Low = faible, Medium = moyen, High = élevé).
- **Access_to_Resources** : Accès aux ressources éducatives (Low, Medium, High).
- **Extracurricular_Activities** : Participation à des activités extrascolaires (Yes = oui, No = non).
- **Sleep_Hours** : Nombre moyen d'heures de sommeil par nuit.
- **Previous_Scores** : Notes obtenues précédemment (en pourcentage).
- **Motivation_Level** : Niveau de motivation (Low, Medium, High).
- **Internet_Access** : Accès à Internet (Yes = oui, No = non).
- **Tutoring_Sessions** : Nombre de sessions de tutorat suivies.
- **Family_Income** : Niveau de revenu familial (Low = faible, Medium = moyen, High = élevé).
- **Teacher_Quality** : Qualité perçue des enseignants (Low, Medium, High).
- **School_Type** : Type d'établissement scolaire (Public = public, Private = privé).
- **Peer_Influence** : Influence des pairs (Positive, Neutral, Negative).
- **Physical_Activity** : Fréquence ou durée de l'activité physique (en unités).
- **Learning_Disabilities** : Présence de troubles d'apprentissage (Yes = oui, No = non).
- **Parental_Education_Level** : Niveau d'éducation des parents (High School, College, Postgraduate).
- **Distance_from_Home** : Distance entre le domicile et l'établissement scolaire (Near, Moderate, Far).
- **Gender** : Genre de l'étudiant.
- **Exam_Score** : Résultat à un examen (en pourcentage).

Origine données

Les données utilisées dans ce projet proviennent de la plateforme Kaggle, un site web connu pour son large éventail de jeux de données publics, utilisés principalement pour des projets d'analyse de données et de visualisation. Les fichiers sélectionnés ont été choisis en raison de leur pertinence pour l'étude de la santé mentale et des performances académiques des étudiants.

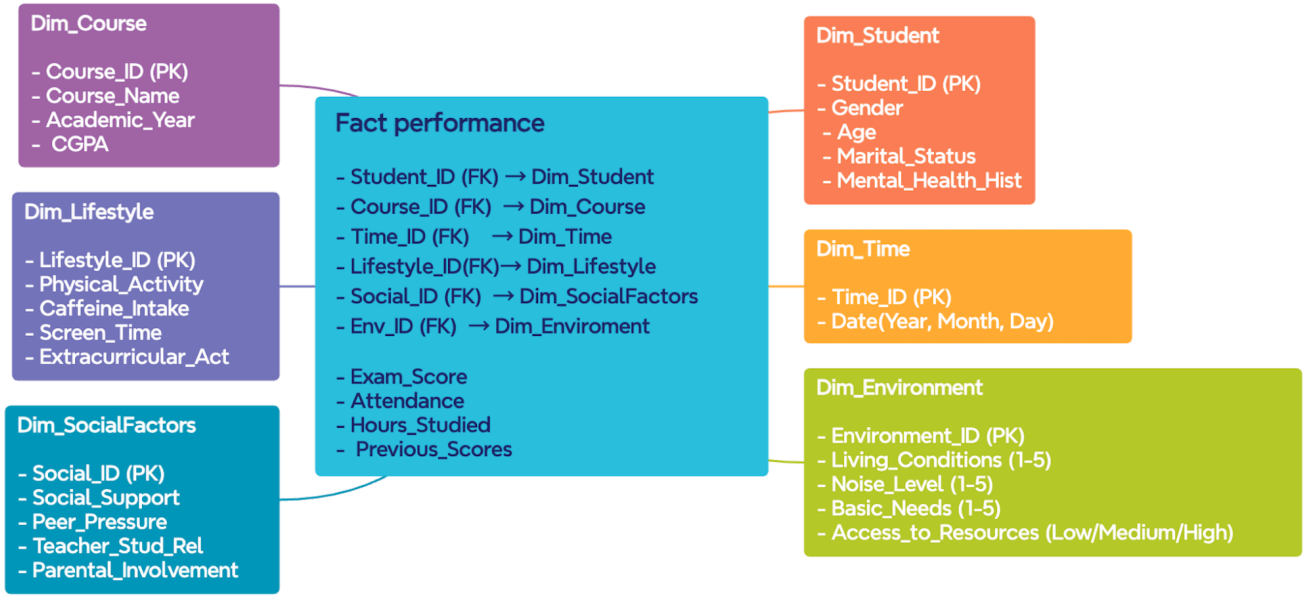
Architecture et gestion des données

Schéma d/architecture



Modèle dimensionnel

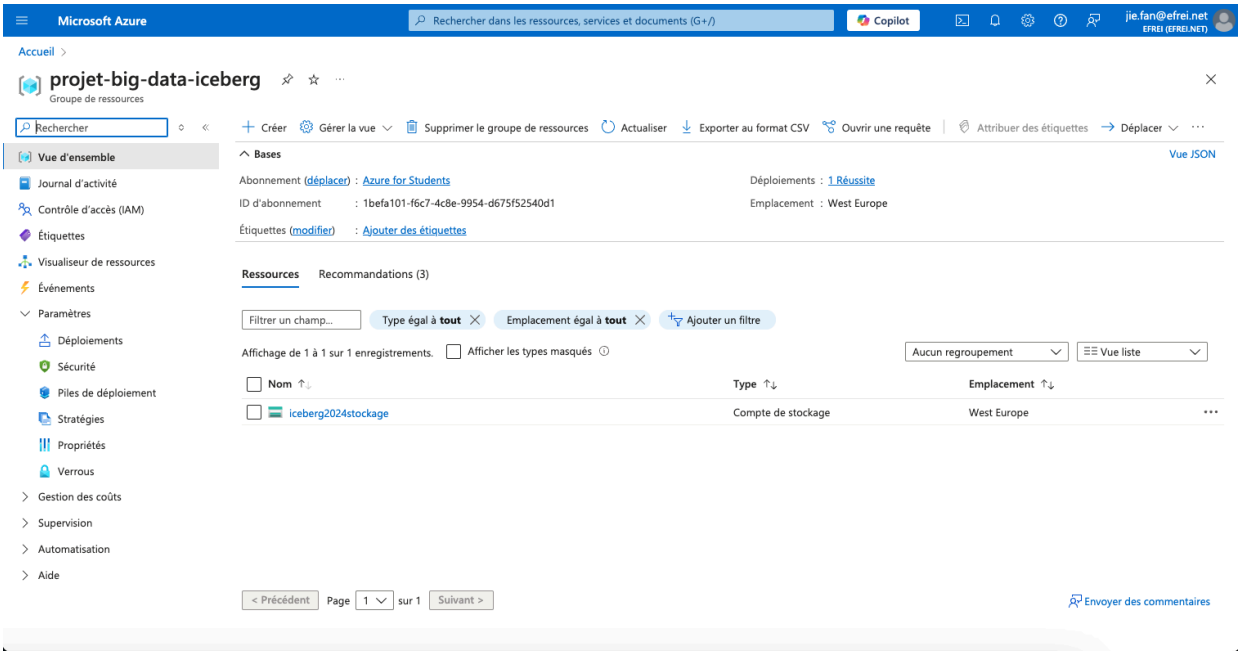
schéma en étoile



Traitement des données

Infrastructures et Ressources

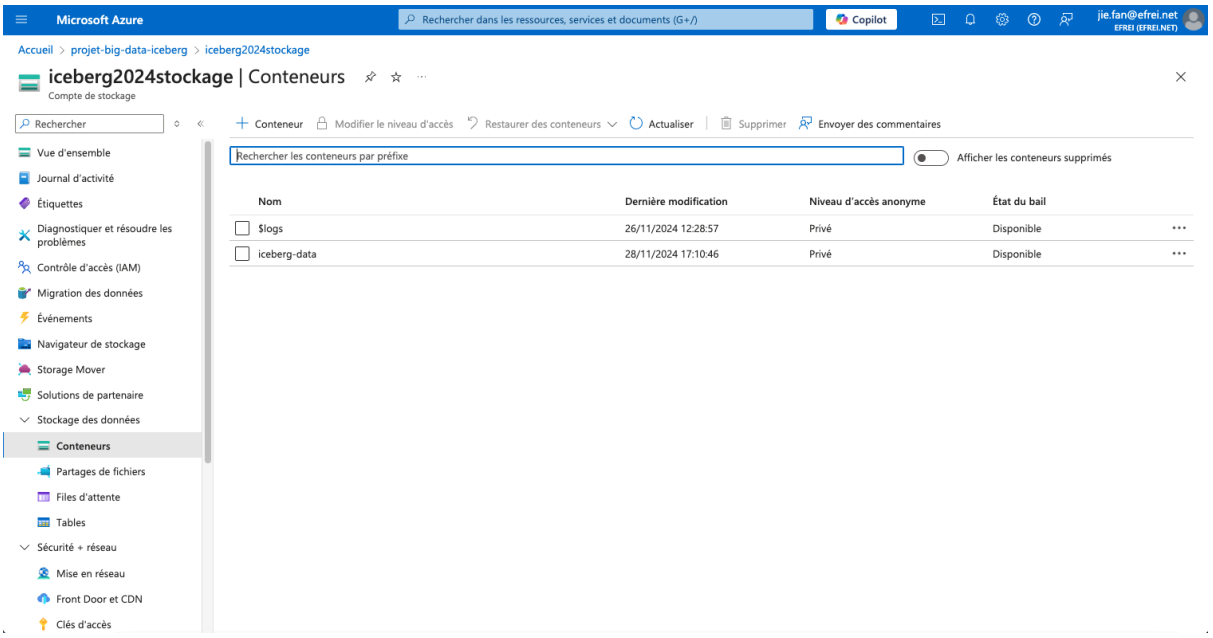
Groupe de Ressources Azure et Compte de Stockage



- **Groupe de ressources** : `projet-big-data-iceberg`
- **Compte de stockage** : `iceberg2024stockage`

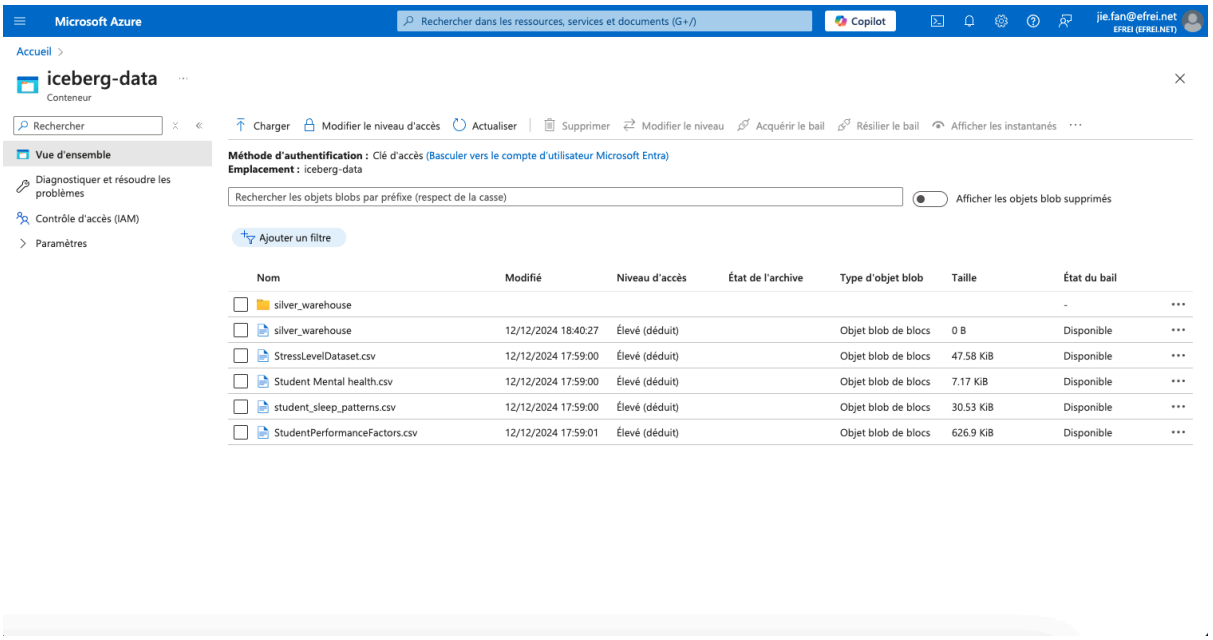
Conteneurs Blob Storage

Deux conteneurs principaux ont été créés :

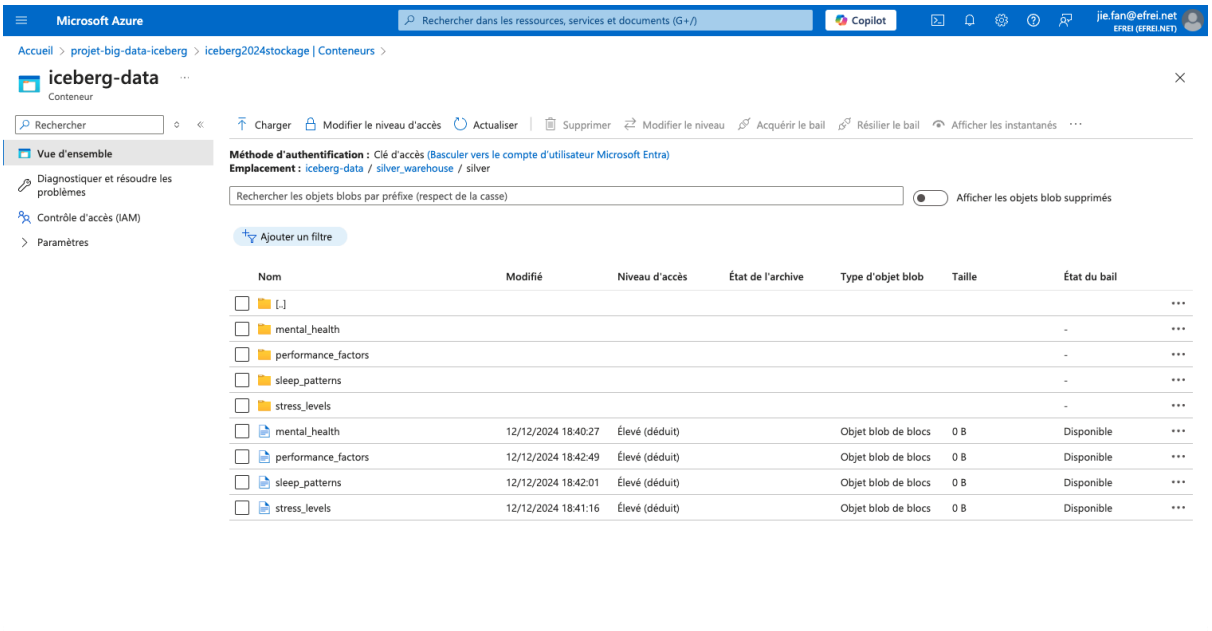


1. **`iceberg-data`** : Contient les fichiers CSV bruts (Bronze) et les tables Iceberg (Silver).
2. **`$logs`** : Contient les journaux de diagnostic.

Contenu du conteneur `iceberg-data` :



- 4 Fichiers CSV originaux (bronze):
  - `Student Mental health.csv`,
  - `StressLevelDataset.csv`,
  - `student_sleep_patterns.csv`,
  - `StudentPerformanceFactors.csv`
- Répertoire `silver_warehouse` : Contient les métadonnées des tables Iceberg (Silver).



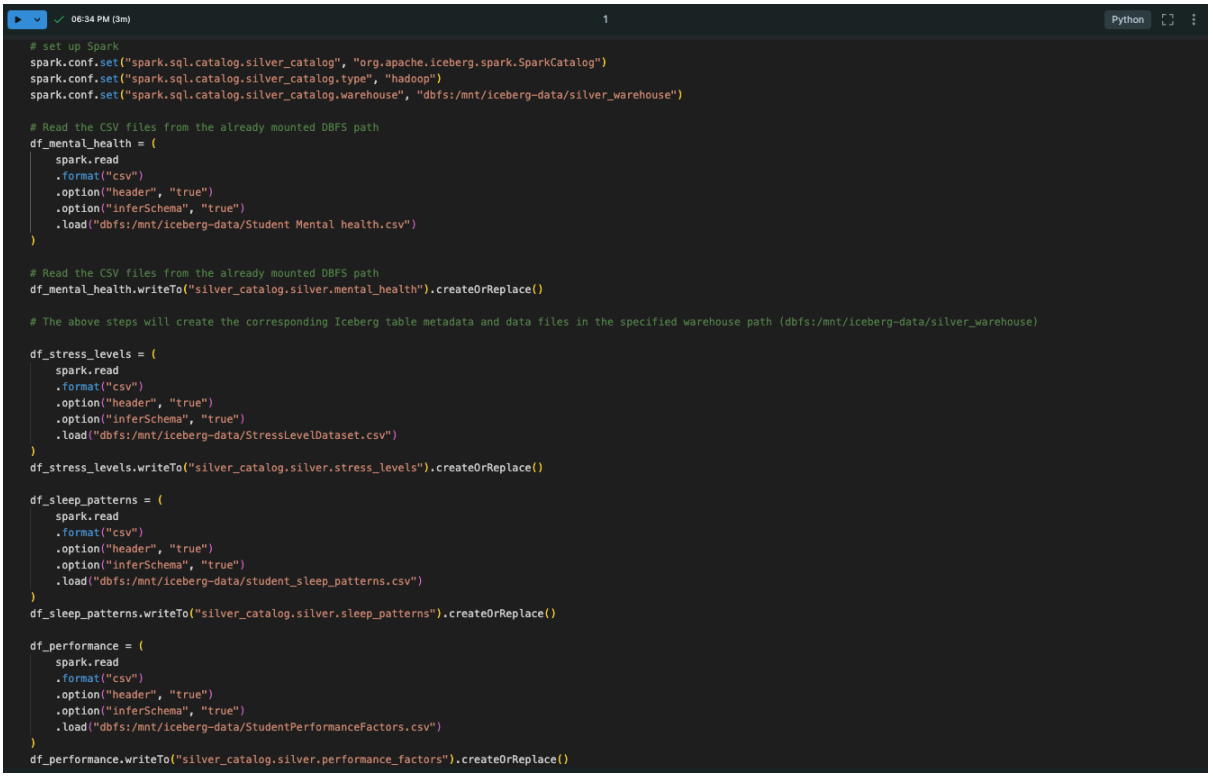
Étapes d'Implémentation

Étape 1 : Montage dans Databricks

- Le conteneur `iceberg-data` a été monté sur le chemin **`dbfs:/mnt/iceberg-data`** dans Databricks.
  Cela permet d'accéder directement aux fichiers dans Databricks sans spécifier de chemin Azure.

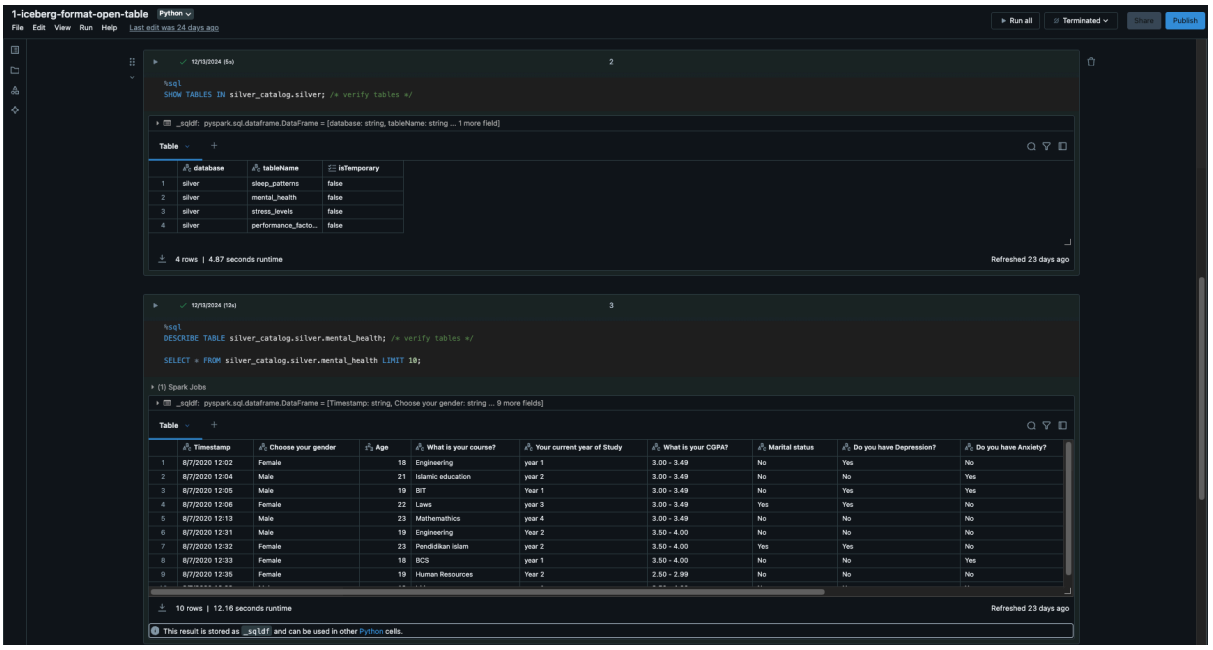
Étape 2 : Transformation en Tables Iceberg (Silver)

- Les fichiers CSV bruts (Bronze) sont lus via Spark, nettoyés, puis stockés sous forme de tables Iceberg dans le répertoire `silver_warehouse`.



```
# set up Spark
spark.conf.set("spark.sql.catalog.silver_catalog", "org.apache.iceberg.spark.SparkCatalog")
spark.conf.set("spark.sql.catalog.silver_catalog.type", "hadoop")
spark.conf.set("spark.sql.catalog.silver_catalog.warehouse", "dbfs:/mnt/iceberg-data/silver_warehouse")

# Read the CSV files from the already mounted DBFS path
df_mental_health = (
    spark.read
    .format("csv")
    .option("header", "true")
    .option("inferSchema", "true")
    .load("dbfs:/mnt/iceberg-data/Student Mental health.csv")
)

# Read the CSV files from the already mounted DBFS path
df_mental_health.writeTo("silver_catalog.silver.mental_health").createOrReplace()

# The above steps will create the corresponding Iceberg table metadata and data files in the specified warehouse path (dbfs:/mnt/iceberg-data/silver_warehouse)

df_stress_levels = (
    spark.read
    .format("csv")
    .option("header", "true")
    .option("inferSchema", "true")
    .load("dbfs:/mnt/iceberg-data/StressLevelDataset.csv")
)
df_stress_levels.writeTo("silver_catalog.silver.stress_levels").createOrReplace()

df_sleep_patterns = (
    spark.read
    .format("csv")
    .option("header", "true")
    .option("inferSchema", "true")
    .load("dbfs:/mnt/iceberg-data/student_sleep_patterns.csv")
)
df_sleep_patterns.writeTo("silver_catalog.silver.sleep_patterns").createOrReplace()

df_performance = (
    spark.read
    .format("csv")
    .option("header", "true")
    .option("inferSchema", "true")
    .load("dbfs:/mnt/iceberg-data/StudentPerformanceFactors.csv")
)
df_performance.writeTo("silver_catalog.silver.performance_factors").createOrReplace()
```

Étape 3 : Vérification des Tables Iceberg



```
%sql
SHOW TABLES IN silver_catalog.silver; /* verify tables */
```

	database	tableName	isTemporary
1	silver	sleep_patterns	false
2	silver	mental_health	false
3	silver	stress_levels	false
4	silver	performance_facto...	false

4 rows | 4.87 seconds runtime — Refreshed 23 days ago

```
%sql
DESCRIBE TABLE silver_catalog.silver.mental_health; /* verify tables */

SELECT * FROM silver_catalog.silver.mental_health LIMIT 10;
```

	Timestamp	Choose your gender	Age	What is your course?	Your current year of Study	What is your CGPA?	Marital status	Do you have Depression?	Do you have Anxiety?
1	8/7/2020 12:02	Female	18	Engineering	year 1	3.00 - 3.49	No	Yes	No
2	8/7/2020 12:04	Male	21	Islamic education	year 2	3.00 - 3.49	No	No	Yes
3	8/7/2020 12:05	Male	19	BIT	Year 1	3.00 - 3.49	No	Yes	Yes
4	8/7/2020 12:06	Female	22	Laws	year 3	3.00 - 3.49	Yes	Yes	No
5	8/7/2020 12:13	Male	23	Mathemathics	year 4	3.00 - 3.49	No	No	No
6	8/7/2020 12:31	Male	19	Engineering	Year 2	3.50 - 4.00	No	No	No
7	8/7/2020 12:32	Female	23	Pendidikan islam	year 2	3.50 - 4.00	Yes	Yes	No
8	8/7/2020 12:33	Female	18	BCS	year 1	3.50 - 4.00	No	No	Yes
9	8/7/2020 12:35	Female	19	Human Resources	Year 2	2.50 - 2.99	No	No	No

10 rows | 12.16 seconds runtime — Refreshed 23 days ago

This result is stored as `_sqldf` and can be used in other Python cells.

Étape nettoyage des données

Dans cette section, nous détaillons les étapes de nettoyage des données effectuées dans le cadre de notre projet. Ce processus est crucial pour garantir la qualité des données en vue de leur exploitation dans la construction des tables de dimensions et de faits.

1. Initialisation de la session Spark

```
spark = SparkSession.builder \
    .appName("Data Cleaning and Dimension/Fact Table Creation") \
    .config("spark.sql.catalog.silver_catalog", "org.apache.iceberg.spark.SparkCatalog") \
    .config("spark.sql.catalog.silver_catalog.type", "hadoop") \
    .config("spark.sql.catalog.silver_catalog.warehouse",
"dbfs:/mnt/iceberg-data/silver_warehouse") \
    .getOrCreate()
```

- **Objectif :** Configurer et initialiser une session Spark.
- **Explication :** Cette étape initialise une session Spark avec des configurations spécifiques pour utiliser le catalogue "Iceberg". Cela nous permet de lire et d'écrire des tables directement dans le *data lake* situé dans le répertoire `dbfs:/mnt/iceberg-data/silver_warehouse`.

2. Chargement des tables sources

```
df_mental_health = spark.table("silver_catalog.silver.mental_health")
df_sleep_patterns = spark.table("silver_catalog.silver.sleep_patterns")
df_stress_levels = spark.table("silver_catalog.silver.stress_levels")
df_performance_factors = spark.table("silver_catalog.silver.performance_factors")
```

- **Objectif :** Charger les données brutes depuis le catalogue "Iceberg".
- **Explication :** Quatre tables (santé mentale, habitudes de sommeil, niveaux de stress, et facteurs de performance) sont chargées dans des *DataFrames* Spark. Ces données constituent les sources pour les étapes ultérieures.

3. Création des tables dimensionnels

Ajout d'un identifiant synthétique

Dans chaque table de dimension, on conserve deux références au même identifiant :

- **`Synthetic_ID`** (le même que dans les données source/fact) pour pouvoir effectuer des jointures (JOIN) précises avec la table de faits.
- **`col("Synthetic_ID").alias("XXX_ID")`** (par ex. `Student_ID`, `Course_ID`) pour définir la clé primaire de la dimension et respecter la logique du modèle en étoile (Star Schema).

Ainsi, `Synthetic_ID` est la référence de correspondance avec la table de faits, tandis que `XXX_ID` est la clé propre à la dimension.

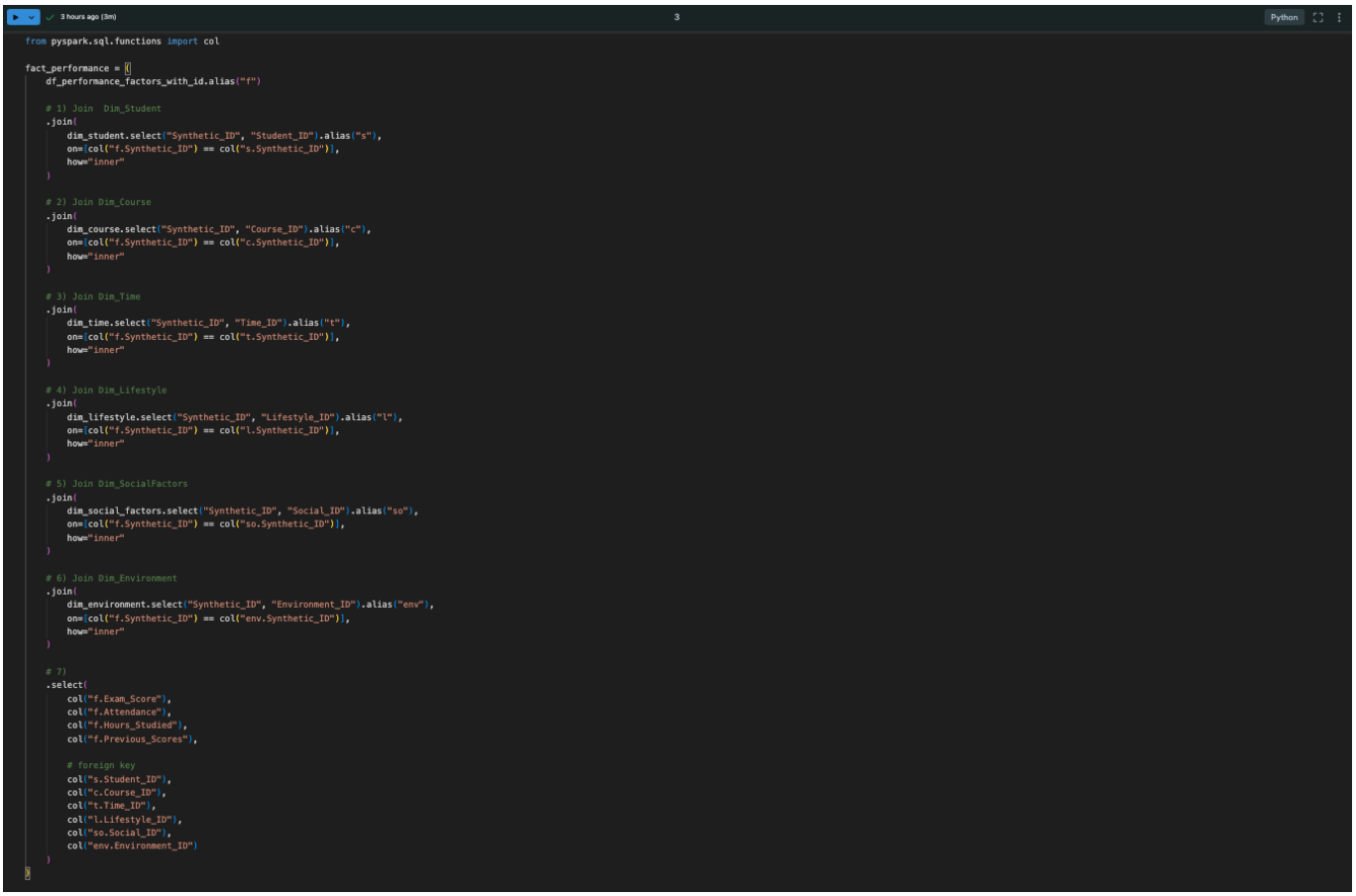
Nettoyage et standardisation des colonnes

Les transformations suivantes ont été appliquées au niveau des tables pour harmoniser les colonnes :

- **Renommage des colonnes :** Par exemple, `Choose your gender` devient `Gender`, `What is your course?` devient `Course_Name`, etc. Cela permet de simplifier la lecture des données et d'uniformiser les noms des colonnes.

- **Conversion des types de données :** Pour certaines colonnes comme les dates (`Timestamp`), une conversion explicite en format `yyyy-MM-dd` a été effectuée pour les rendre compatibles avec les outils d'analyse.

4. Création de la table faits

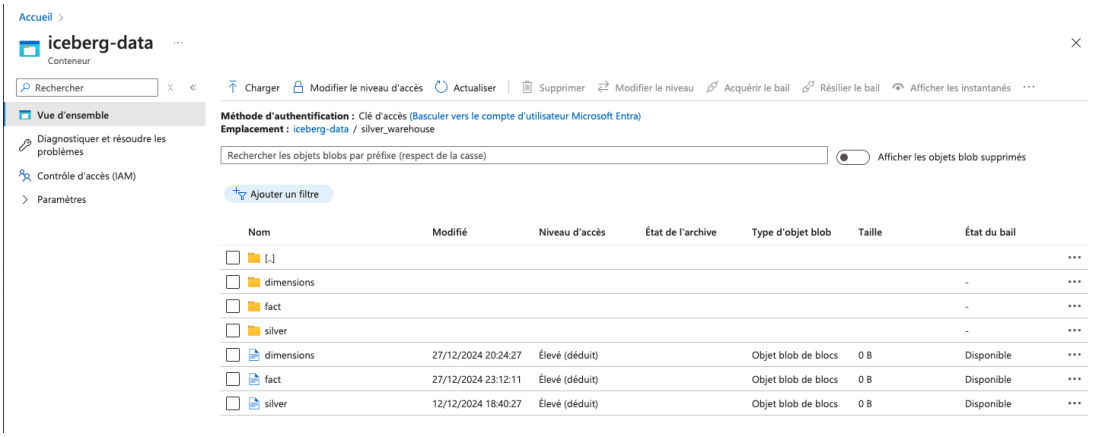


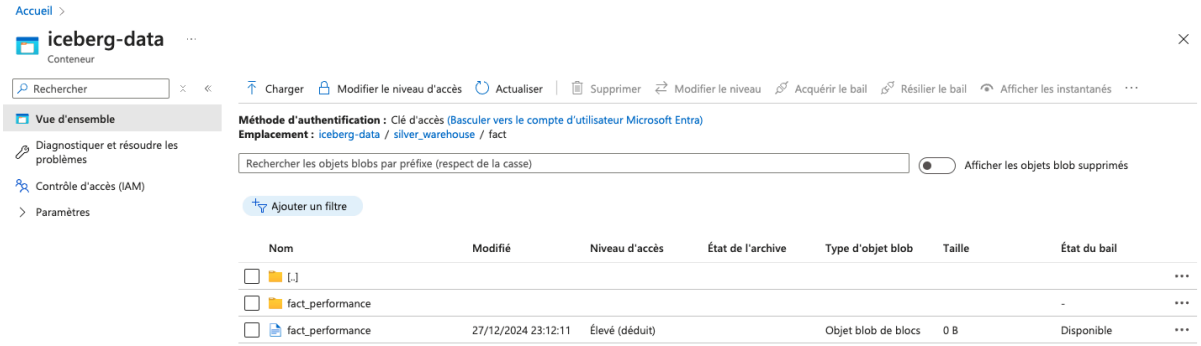
```
from pyspark.sql.functions import col

fact_performance = (
    df_performance_factors_with_id.alias("f")

    # 1) Join  Dim_Student
    .join(
        dim_student.select("Synthetic_ID", "Student_ID").alias("s"),
        on=[col("f.Synthetic_ID") == col("s.Synthetic_ID")],
        how="inner"
    )

    # 2) Join Dim_Course
    .join(
        dim_course.select("Synthetic_ID", "Course_ID").alias("c"),
        on=[col("f.Synthetic_ID") == col("c.Synthetic_ID")],
        how="inner"
    )

    # 3) Join Dim_Time
    .join(
        dim_time.select("Synthetic_ID", "Time_ID").alias("t"),
        on=[col("f.Synthetic_ID") == col("t.Synthetic_ID")],
        how="inner"
    )

    # 4) Join Dim_Lifestyle
    .join(
        dim_lifestyle.select("Synthetic_ID", "Lifestyle_ID").alias("l"),
        on=[col("f.Synthetic_ID") == col("l.Synthetic_ID")],
        how="inner"
    )

    # 5) Join Dim_SocialFactors
    .join(
        dim_social_factors.select("Synthetic_ID", "Social_ID").alias("so"),
        on=[col("f.Synthetic_ID") == col("so.Synthetic_ID")],
        how="inner"
    )

    # 6) Join Dim_Environment
    .join(
        dim_environment.select("Synthetic_ID", "Environment_ID").alias("env"),
        on=[col("f.Synthetic_ID") == col("env.Synthetic_ID")],
        how="inner"
    )

    # 7)
    .select(
        col("f.Exam_Score"),
        col("f.Attendance"),
        col("f.Hours_Studied"),
        col("f.Previous_Scores"),

        # foreign key
        col("s.Student_ID"),
        col("c.Course_ID"),
        col("t.Time_ID"),
        col("l.Lifestyle_ID"),
        col("so.Social_ID"),
        col("env.Environment_ID")
    )
)
```

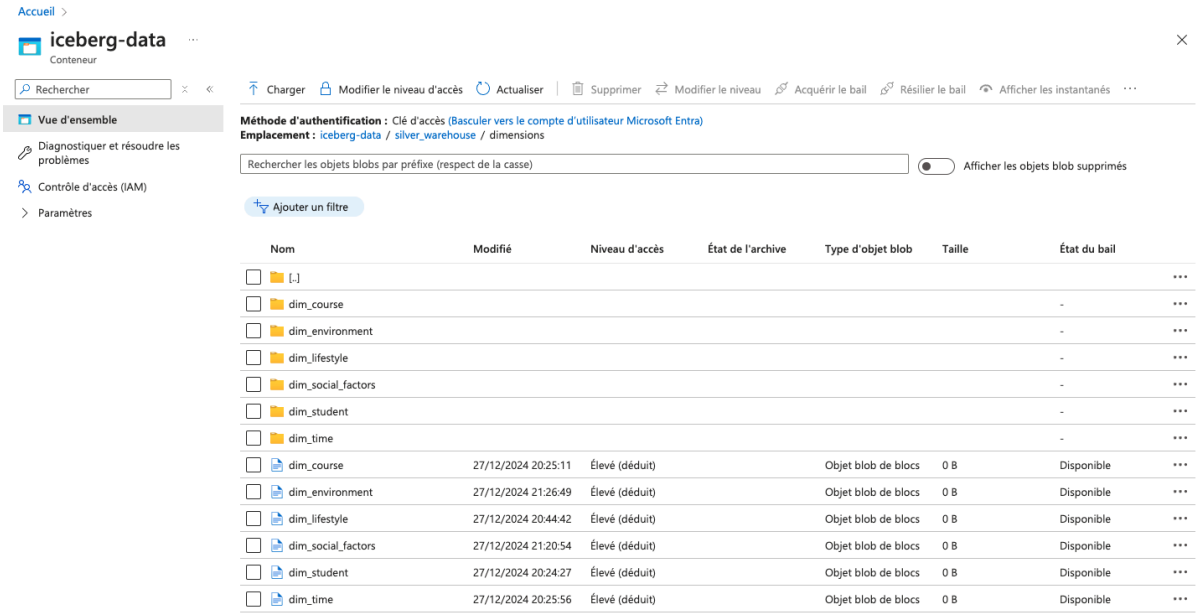
Résultats du nettoyage

Après cette étape, les données sont :

1. **Prêtes pour les jointures :** Grâce à l'ajout des identifiants synthétiques.
2. **Normalisées :** Les noms de colonnes sont cohérents et les formats de données sont standardisés.
3. **Sans doublons :** Les colonnes clés sont dédupliquées afin de réduire les incohérences dans les analyses futures.


Accueil
iceberg-data
Conteneur
Rechercher
Vue d'ensemble
Diagnostiquer et résoudre les problèmes
Contrôle d'accès (IAM)
Paramètres
Charger
Modifier le niveau d'accès
Actualiser
Supprimer
Modifier le niveau
Acquérir le bail
Résilier le bail
Afficher les instantanés
Méthode d'authentification : Clé d'accès (Basculer vers le compte d'utilisateur Microsoft Entra)
Emplacement : iceberg-data / silver_warehouse
Rechercher les objets blobs par préfixe (respect de la casse)
Afficher les objets blob supprimés
Ajouter un filtre
Nom
Modifié
Niveau d'accès
État de l'archive
Type d'objet blob
Taille
État du bail
[..]
dimensions
fact
silver
dimensions 27/12/2024 20:24:27 Élevé (déduit) Objet blob de blocs 0 B Disponible
fact 27/12/2024 23:12:11 Élevé (déduit) Objet blob de blocs 0 B Disponible
silver 12/12/2024 18:40:27 Élevé (déduit) Objet blob de blocs 0 B Disponible


Accueil
iceberg-data
Conteneur
Rechercher
Vue d'ensemble
Diagnostiquer et résoudre les problèmes
Contrôle d'accès (IAM)
Paramètres
Charger
Modifier le niveau d'accès
Actualiser
Supprimer
Modifier le niveau
Acquérir le bail
Résilier le bail
Afficher les instantanés
Méthode d'authentification : Clé d'accès (Basculer vers le compte d'utilisateur Microsoft Entra)
Emplacement : iceberg-data / silver_warehouse / fact
Rechercher les objets blobs par préfixe (respect de la casse)
Afficher les objets blob supprimés
Ajouter un filtre
Nom
Modifié
Niveau d'accès
État de l'archive
Type d'objet blob
Taille
État du bail
[..]
fact_performance
fact_performance 27/12/2024 23:12:11 Élevé (déduit) Objet blob de blocs 0 B Disponible


Accueil
iceberg-data
Conteneur
Rechercher
Vue d'ensemble
Diagnostiquer et résoudre les problèmes
Contrôle d'accès (IAM)
Paramètres
Charger
Modifier le niveau d'accès
Actualiser
Supprimer
Modifier le niveau
Acquérir le bail
Résilier le bail
Afficher les instantanés
Méthode d'authentification : Clé d'accès (Basculer vers le compte d'utilisateur Microsoft Entra)
Emplacement : iceberg-data / silver_warehouse / dimensions
Rechercher les objets blobs par préfixe (respect de la casse)
Afficher les objets blob supprimés
Ajouter un filtre
Nom
Modifié
Niveau d'accès
État de l'archive
Type d'objet blob
Taille
État du bail
[..]
dim_course
dim_environment
dim_lifestyle
dim_social_factors
dim_student
dim_time
dim_course 27/12/2024 20:25:11 Élevé (déduit) Objet blob de blocs 0 B Disponible
dim_environment 27/12/2024 21:26:49 Élevé (déduit) Objet blob de blocs 0 B Disponible
dim_lifestyle 27/12/2024 20:44:42 Élevé (déduit) Objet blob de blocs 0 B Disponible
dim_social_factors 27/12/2024 21:20:54 Élevé (déduit) Objet blob de blocs 0 B Disponible
dim_student 27/12/2024 20:24:27 Élevé (déduit) Objet blob de blocs 0 B Disponible
dim_time 27/12/2024 20:25:56 Élevé (déduit) Objet blob de blocs 0 B Disponible

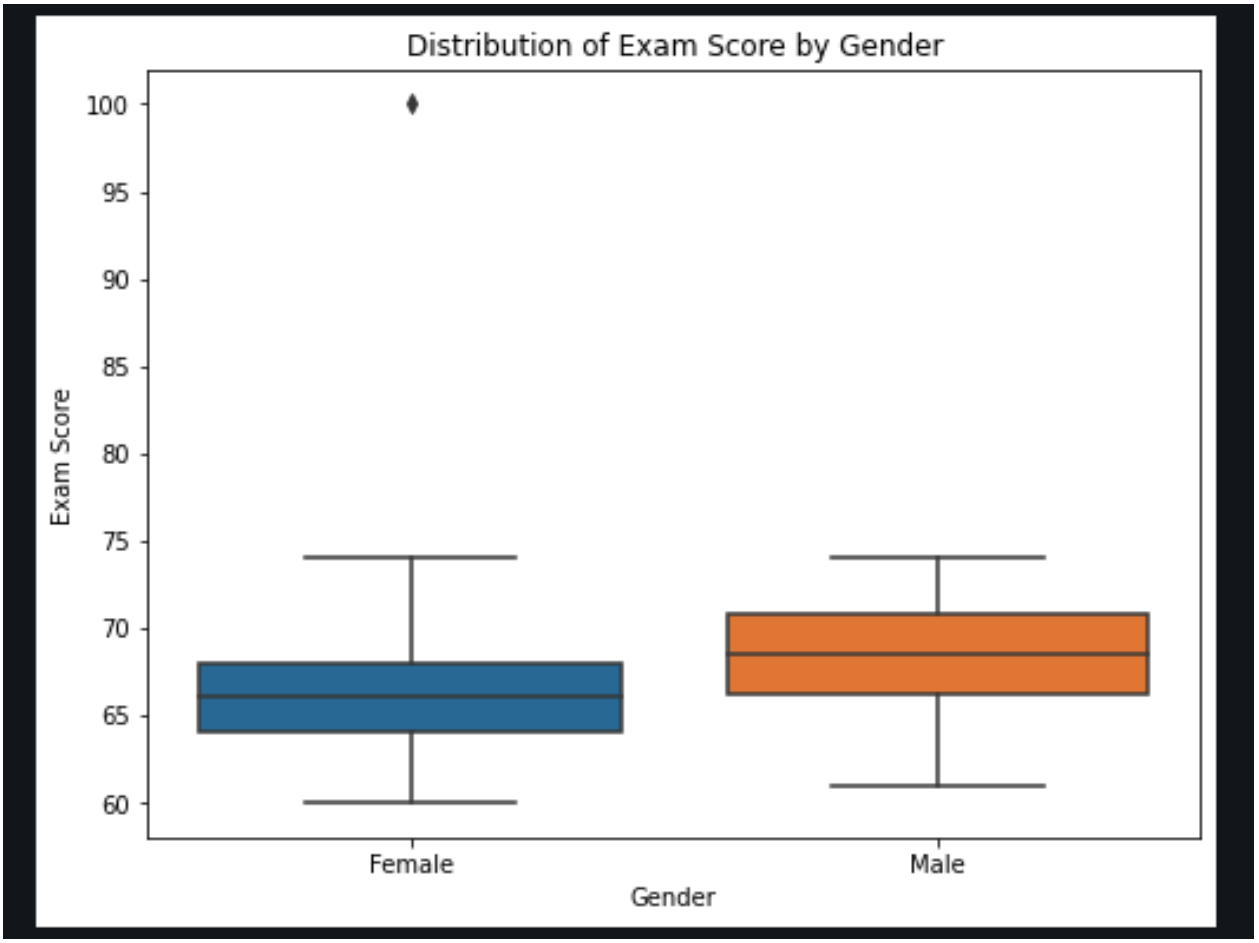
Analyse des résultats

Visualisation des données

Nous avons effectué une jointure entre la table de faits et les différentes tables de dimensions afin de rassembler toutes les informations nécessaires sur un seul DataFrame. Cela nous permet d'analyser conjointement les facteurs de style de vie, l'environnement, le soutien social et la performance académique des étudiants.

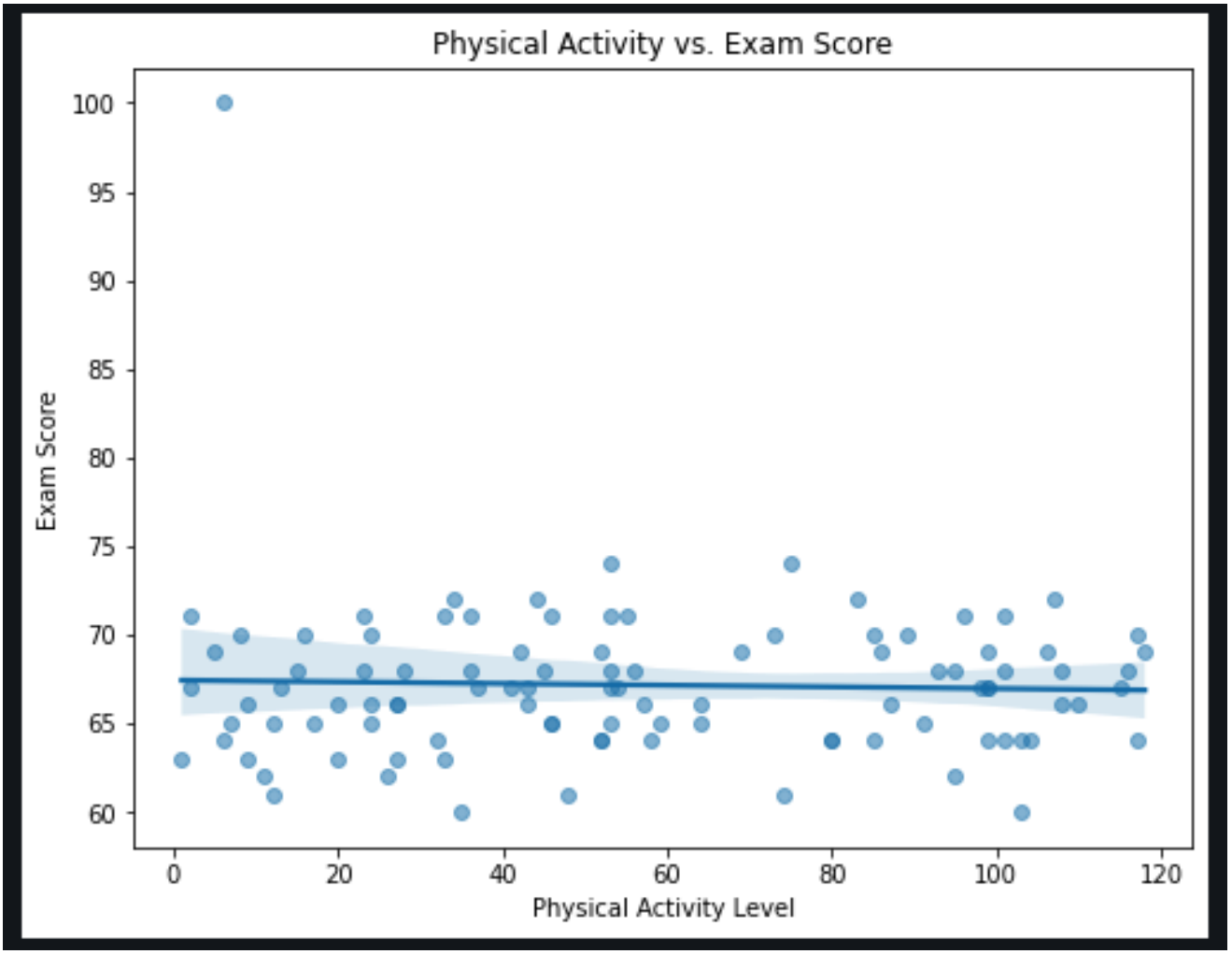
1.Distribution des notes d'examen selon le genre

Interprétation possible : Déterminer si l'un des genres obtient systématiquement des notes plus élevées ou si les distributions sont semblables.



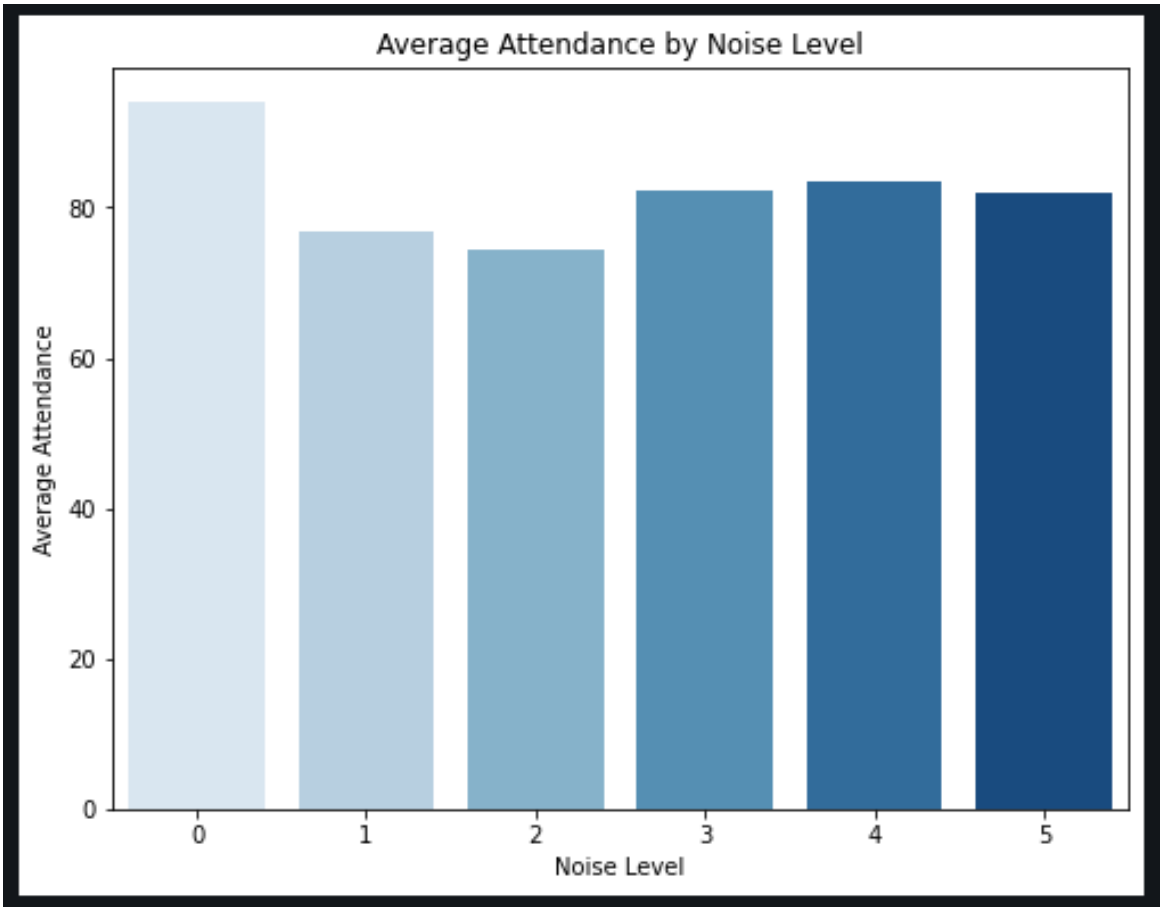
2. Corrélation entre l’activité physique et la note d’examen

Interprétation possible : Une pente positive suggère qu’augmenter l’activité physique coïncide avec de meilleures notes.



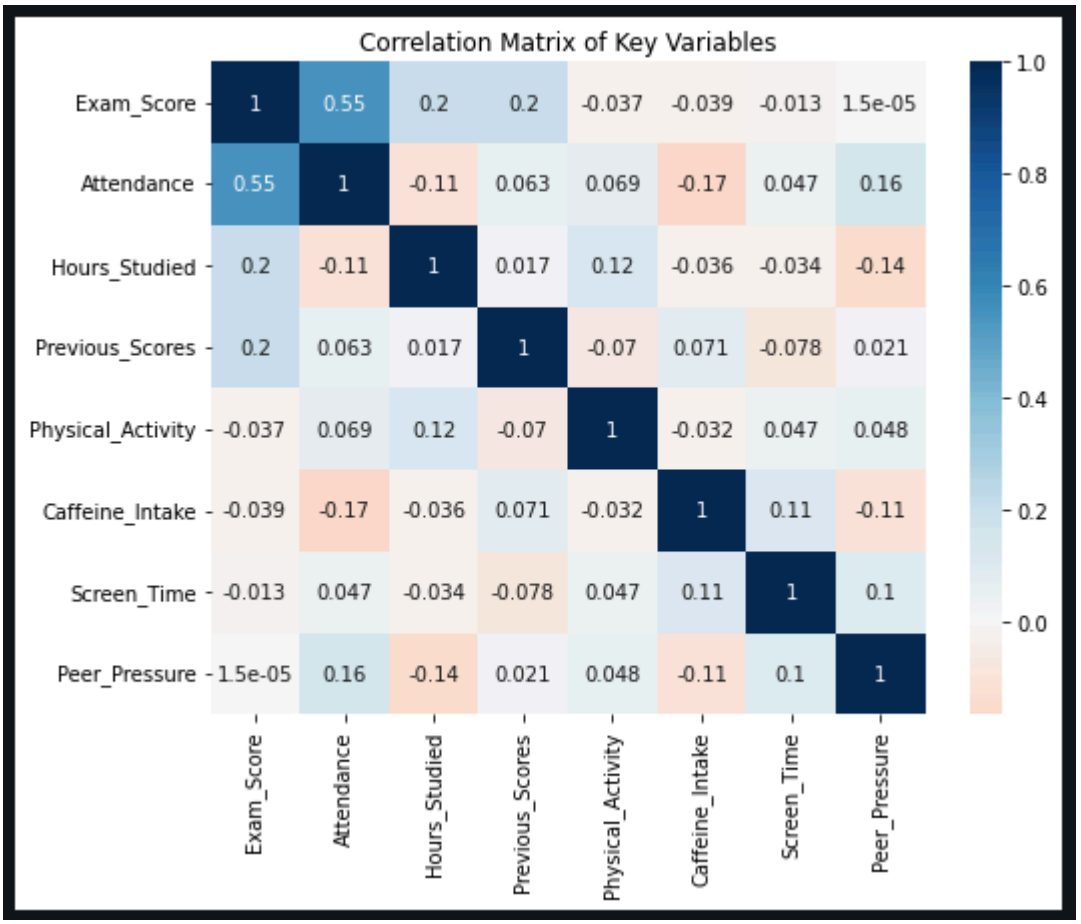
3. Impact du niveau de bruit sur le taux de présence

Interprétation possible : Un environnement plus calme contribuerait-il à un meilleur engagement ?



4. Carte de corrélation pour plusieurs variables

Interprétation possible : Identifier rapidement quelles variables sont corrélées positivement ou négativement.



Conclusion

Ce projet Big Data axé sur la santé mentale et les performances académiques des étudiants nous a démontré l'importance d'une architecture de données pour une meilleure compréhension et un suivi continu. Grâce aux différentes couches (Bronze, Silver et Gold), nous avons pu nettoyer, transformer puis organiser nos données de façon fiable, garantissant leur cohérence et leur intégrité tout au long du pipeline.

L'utilisation d'Iceberg et de modèles en étoile (tables de faits et dimensions) nous a permis de structurer les informations clés (sommeil, stress, environnement, etc.) et de faciliter l'exploration analytique. Les visualisations finales ont mis en évidence certaines corrélations

significatives, par exemple entre le niveau de stress et la réussite académique, ou encore l'impact positif de bonnes conditions de vie sur les notes et la santé mentale.

En définitive, l'architecture mise en place et les analyses menées offrent une vision de l'état de la santé mentale et de la performance des étudiants.

Annexes

Code et scripts : https://github.com/Jie01236/Projet-Big-Data.git