



## **Data Engineering on Cloud**

**Sujet:** L'Impact de l'Indice de Position Sociale des Élèves sur la Répartition des Effectifs en France

**Présenter Par :** Fatat TARRAF, Mohamad DIAG, Nikola ZIVKOVIC, Adam EL OUNISSI

**Présenter a :** M. Alexandre BERGÈRE

Mastère Dev Manager Full Stack

Groupe 2

## **Table des Matières**

Introduction.....	3
Étapes du projet et outils utilisés .....	3
Sources et Collecte des Données .....	4
Types de données et leur format.....	6
Modèle d'Architecture du Système .....	8
Unité de ressources Azure .....	9
Opérations réalisées.....	10
Configuration du service lié .....	11
Création du pipeline de copie .....	11
Exécution et validation .....	12
Stockage des Données dans Azure Data Lake Storage (ADLS).....	14
Databricks : Transformation des Données.....	14
Transformation des Données avec Apache Spark.....	15
Préparation des Tables de Dimensions .....	16
Modèle Conceptuel de Données (MCD).....	18
Notebook.....	21
Conclusion.....	23

# Introduction

## Présentation du projet

- Dans le cadre de notre formation en *Big Data*, nous avons réalisé un projet visant à démontrer nos compétences dans le traitement et l'analyse de données à grande échelle. Ce projet nous a permis de manipuler des données complexes en utilisant des outils et services cloud pour répondre à une problématique éducative.
- Nous avons choisi ***d'étudier l'impact de l'indice de position sociale des élèves sur la répartition des effectifs par niveau et sur la gestion des classes dans les écoles en FRANCE***. Cette thématique nous a permis d'explorer comment des facteurs sociaux influencent l'organisation des établissements scolaires.
- Pour mener à bien ce projet, nous avons exploité les services cloud de Microsoft Azure, une plateforme offrant des outils performants pour la collecte, le stockage, le traitement et l'analyse des données. La structure de notre travail et les outils utilisés sont détaillés ci-après.

## Étapes du projet et outils utilisés :

### 1. Collecte et ingestion des données

- *Azure Data Factory* : Nous avons utilisé Azure Data Factory pour collecter et automatiser l'ingestion des données issues de différentes sources. Cet outil a permis de gérer efficacement les flux de données brutes tout en préparant ces dernières pour les étapes de traitement ultérieures.

### 2. Stockage des données

- *Azure Data Lake Storage Gen2* : Les données brutes et transformées ont été stockées dans Azure Data Lake Storage Gen2. Ce service de stockage sécurisé et évolutif a été essentiel pour gérer de grands volumes de données tout en garantissant leur accessibilité.

### 3. Traitement et transformation des données

- *Databricks* : Les données collectées ont été analysées et transformées à l'aide de Databricks. Cet environnement basé sur Apache Spark nous a permis d'effectuer des calculs avancés, tels que la mise en corrélation des indices sociaux avec la répartition des effectifs scolaires.

### 4. Visualisation des résultats

- *Databricks visualization* : Pour rendre nos résultats accessibles et compréhensibles, nous avons conçu des visualisations interactives. Ces graphiques et tableaux de bord ont permis d'illustrer les disparités dans la répartition des effectifs et d'identifier les facteurs sociaux ayant un impact significatif.

## Sources et Collecte des Données

Dans le cadre de ce projet, nous avons utilisé quatre ensembles de données provenant de la plateforme **data.gouv.fr**. Ces données couvrent différents aspects de la répartition des effectifs et de la gestion des classes dans les écoles françaises, en lien avec l'indice de position sociale des élèves.

### 1. Effectifs des élèves et classes par école

- Cette source fournit des données sur le nombre d'élèves et de classes par école, collectées chaque année au début du mois d'octobre. Elle permet d'analyser la répartition des élèves dans les écoles françaises, par niveau scolaire (primaire, collège, lycée) et par région. Ces informations sont essentielles pour comprendre comment les effectifs sont répartis et comment cela affecte la gestion des classes dans les écoles.

### 2. Indices de position sociale des élèves

- Ces données, disponibles à partir de 2022, communiquent des informations sur les indices de position sociale dans les écoles françaises. Ces indices sont des indicateurs clés pour évaluer l'impact de la situation socio-économique des élèves sur leur scolarité. En croisant ces indices avec les données sur les effectifs scolaires, il est possible d'analyser l'influence de l'indice de position sociale sur la gestion des classes et la répartition des élèves par niveau.

3. **Effectifs d'élèves par niveau, sexe, et langues vivantes 1 et 2 les plus fréquentes, par collège**

- Ces indicateurs visent à évaluer l'action de chaque collège pour assurer la réussite de ses élèves.

4. **Indices de position sociale dans les collèges de France métropolitaine et des DROM (à partir de 2022)**

- L'indice de position sociale permet d'appréhender le statut social des élèves à partir des professions et catégories sociales de leurs parents.

**Types de données et leur format**

1. **Effectifs des élèves et classes par école**

○ Colonnes :

- Rentrée scolaire
- Région académique
- Académie
- Département
- Commune
- Numéro de l'école
- Dénomination principale
- Patronyme
- Secteur
- REP
- REP +
- Nombre total de classes
- Nombre total d'élèves
- Nombre d'élèves en pré-élémentaire hors ULIS
- Nombre d'élèves en élémentaire hors ULIS
- Nombre d'élèves en ULIS
- Nombre d'élèves en CP hors ULIS
- Nombre d'élèves en CE1 hors ULIS
- Nombre d'élèves en CE2 hors ULIS
- Nombre d'élèves en CM1 hors ULIS

- Nombre d'élèves en CM2 hors ULIS
- Tri
- Code Postal

## 2. **Indices de position sociale des élèves**

- Colonnes :
  - Rentrée scolaire
  - Académie
  - Code du département
  - Département
  - UAI
  - Nom de l'établissement
  - Code INSEE de la commune
  - Nom de la commune
  - Secteur
  - Effectifs
  - IPS

## 3. **Effectifs des élèves et classes par collège**

- Colonnes :
  - num\_ligne
  - Rentrée scolaire
  - Code région académique
  - Région académique
  - Code académie
  - Académie
  - Code département
  - Département
  - Code postal
  - Commune
  - UAI
  - Dénomination principale
  - Patronyme
  - Secteur
  - REP
  - REP +
  - Nombre d'élèves total

- Nombre d'élèves total hors Segpa hors ULIS
- Nombre d'élèves total Segpa
- Nombre d'élèves total ULIS
- 6èmes total
- 6èmes hors Segpa hors ULIS
- 6èmes Segpa
- 6èmes ULIS
- 6èmes filles
- 6èmes garçons

#### 4. **Indices de position sociale des élèves**

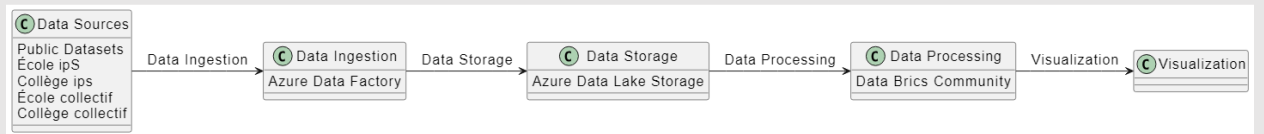
##### ○ Colonnes :

- Rentrée scolaire
- Académie
- Code du département
- Département
- UAI
- Nom de l'établissement
- Code INSEE de la commune
- Nom de la commune
- Secteur
- Effectifs
- IPS
- Ecart-type de l'IPS

Ces sources de données fournissent des informations essentielles pour analyser l'impact de l'indice de position sociale des élèves sur la répartition des effectifs et la gestion des classes dans les écoles. Grâce à ces données, il est possible d'explorer les liens entre les variables socio-économiques et l'organisation scolaire.

# Modèle d'Architecture du Système

## 1. Diagramme de l'Architecture





## Unité de ressources Azure : GPResource-cloud

The screenshot shows the Microsoft Azure portal interface. At the top, there's a search bar and a user profile. Below the header, the 'Services Azure' section is visible, with 'Microsoft Fabric' highlighted. Underneath, the 'Ressources' section displays a table of recent resources.











Nom	Type	Dernier affichage
bigdatadlgcn2	Compte de stockage	il y a 2 jours
ADFBigDataProject7	Fabrique de données (V2)	il y a 2 semaines
BigDataProject	Groupe de ressources	il y a 2 semaines
dlkefrei7	Compte de stockage	il y a 2 semaines
rgData	Groupe de ressources	il y a 3 semaines

Figure 1: Succès de l'activité de copie

## 2. Opérations réalisées

ADF : Extraction et collecte des données

Collecte des fichiers CSV depuis GitHub

 <b>DiagM</b> add gold	4f57aa4 · 2 hours ago	 <b>14 Commits</b>
 bronze	Delete bronze/uykkjh	2 hours ago
 gold	add gold	2 hours ago
 silver	Delete silver/t	2 hours ago
 README.md	Create README.md	2 hours ago
 ecole effectif	Create ecole effectif	2 hours ago
 fr-en-college-effectifs-niveau-sexe-lv.csv	College effectifs	2 hours ago
 fr-en-ips-colleges-ap2022.csv	ips college	2 hours ago
 fr-en-ips-ecoles-ap2022.csv	add file ips ecole	2 hours ago

## 1. Création d'un pipeline de copie dans Azure Data Factory (ADF)

### Objectif :

- Déplacer les fichiers CSV hébergés sur GitHub vers Azure Data Lake Storage Gen2 (ADLS Gen2) en utilisant un pipeline de copie dans ADF.

## A. Configuration du service lié (Linked Service)

### 1. Source - GitHub (HTTP)

- Un service lié de type HTTP a été configuré pour accéder aux fichiers CSV stockés sur GitHub.
- Dans Linked Services d'ADF, un service lié de type HTTP a été créé, avec l'URL de base pointant vers le dépôt GitHub contenant les fichiers de données (ex. : [DiagM/IPS eleve BigData: projet Big Data M1 Semestre1:](#)).
- Les paramètres nécessaires pour l'accès aux fichiers ont été définis, garantissant une connexion sécurisée et stable.

### 2. Destination - Azure Data Lake Storage Gen2 (ADLS Gen2)

- Un service lié de type Azure Data Lake Storage Gen2 a été configuré comme destination.
- Dans Linked Services, un service lié de type ADLS Gen2 a été configuré pour pointer vers notre compte de stockage, avec les informations d'authentification nécessaires (par exemple, clé d'accès ou identités managées).

## B. Création du pipeline de copie

### 1. Activité de copie (Copy Activity)

- Une activité de copie a été ajoutée dans le pipeline afin de transférer les fichiers CSV de GitHub vers ADLS Gen2.
- Cette activité assure une copie exacte des fichiers sans modification de format, permettant une ingestion rapide et fiable des données.

### 2. Configuration de la source

- Dans les paramètres de la source, le format de fichier a été défini comme "DelimitedText" pour gérer correctement les fichiers CSV.
- Le chemin exact des fichiers CSV sur GitHub a été fourni, y compris le nom complet des fichiers

### 3. Configuration de la destination (Sink)

- La destination de l'activité de copie a été configurée pour écrire les fichiers dans Azure Data Lake Storage Gen2.
- Le service lié à ADLS Gen2 a été sélectionné, et le chemin de destination a été défini pour les fichiers CSV.
- Le format "DelimitedText" a été choisi pour maintenir les fichiers au format CSV afin de faciliter leur traitement ultérieur.

## C. Exécution et validation

- Une fois la configuration terminée, le pipeline a été exécuté pour transférer les fichiers CSV depuis GitHub vers ADLS Gen2.
- Après l'exécution, une validation a été effectuée pour vérifier que les fichiers avaient bien été transférés et stockés au bon emplacement dans le Data Lake, tout en conservant leur intégrité.

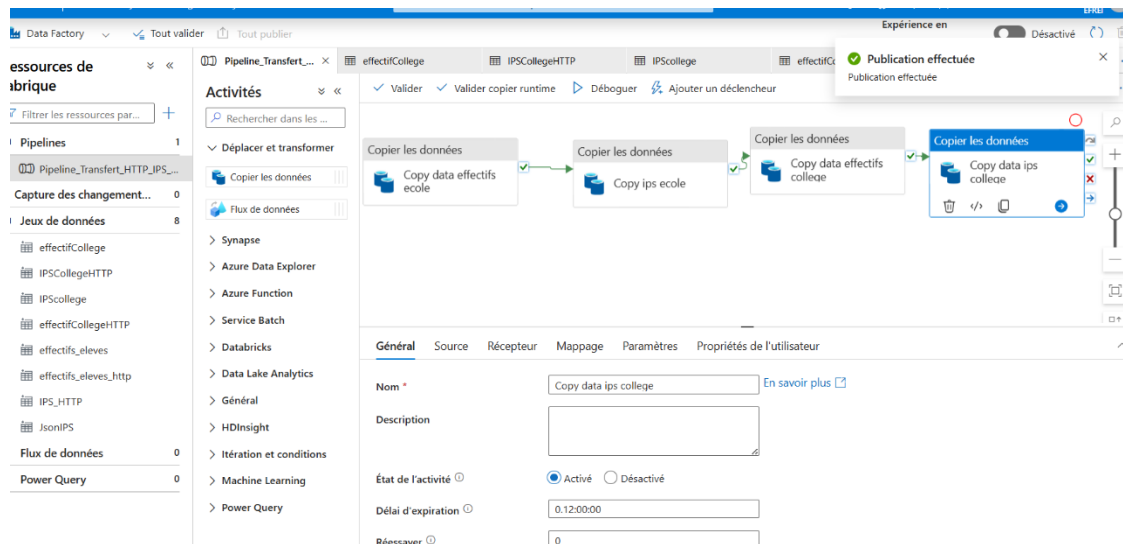


Figure 2: Configuration du service lié de destination ADLS Gen2

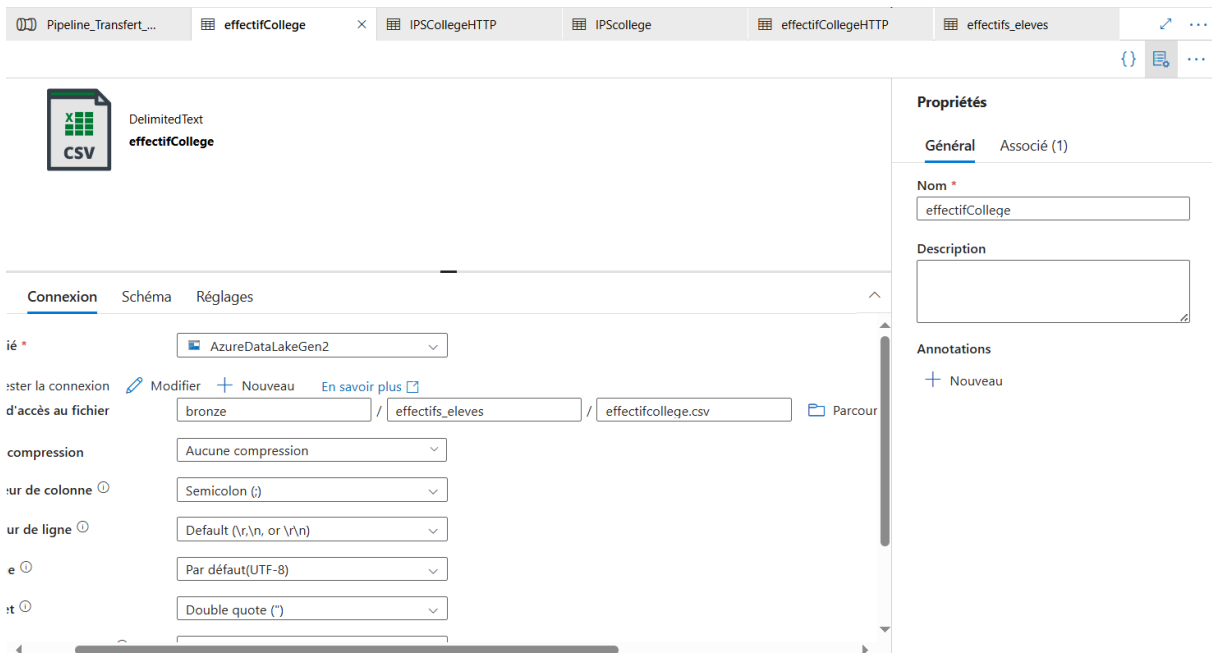


Figure 3: Configuration du service lié Source http

## **A. Stockage des Données dans Azure Data Lake Storage (ADLS)**

Une fois les données ingérées depuis les fichiers CSV, nous avons validé le succès de l'opération en vérifiant que les fichiers étaient correctement transférés et stockés dans le conteneur raw-data d'Azure Data Lake Storage (ADLS). Cette étape a permis de garantir que les données étaient complètes, bien formatées et accessibles pour les traitements suivants.

### **1. Validation de l'Ingestion**

- ❖ Nous avons confirmé que les fichiers CSV, contenant des informations cruciales sur les effectifs scolaires, les indices de position sociale, etc., étaient présents dans le conteneur raw-data d'ADLS.
- ❖ Les fichiers ont été examinés pour s'assurer qu'ils étaient intègres, que les données étaient complètes et correctement structurées, sans perte ni corruption des informations.

### **2. Stockage des Données Brutes**

- ❖ Les données ont été stockées dans leur état brut, tel qu'elles ont été ingérées depuis les différentes sources, pour préserver leur intégrité. Ce stockage permet de garder une trace fidèle des données initiales avant toute transformation ou analyse.
- ❖ Cette approche est essentielle pour garantir que les modifications futures sur les données, telles que l'analyse des tendances ou la préparation des rapports, reposent sur des données non altérées et fidèles à la source.

### **3. Organisation et Sécurisation des Données**

- ❖ Les fichiers CSV ont été organisés dans un conteneur structuré pour faciliter l'accès et l'exploitation des données. Chaque type de données est stocké dans des dossiers clairement identifiés pour permettre une navigation efficace et une gestion cohérente des fichiers.
- ❖ Des mesures de sécurité ont été mises en place pour protéger les données, notamment l'application de règles d'accès basées sur des rôles (RBAC) et un chiffrement des données, assurant ainsi que seules les personnes autorisées peuvent accéder ou modifier les informations sensibles.

### **B.Databricks : Transformation des Données**

#### **1. Connexion d'Azure Databricks au Compte de Stockage Azure**

Pour établir la connexion entre Azure Databricks et notre compte de stockage Azure Data Lake Storage (ADLS), nous avons exploré deux solutions pour gérer l'accès sécurisé aux ressources Azure.

##### **1.1 App Registration**

###### **Principe :**

- Cette solution consiste à créer une application dans Azure Active Directory (AAD), puis à configurer les permissions nécessaires pour accéder aux ressources du compte de stockage.

###### **Configuration :**

- Les identifiants nécessaires (ID client, ID locataire, clé secrète) sont utilisés dans Databricks pour l'authentification OAuth, ce qui permet de sécuriser l'accès aux données dans ADLS Gen2.

##### **1.2 Azure Key Vault**

###### **Principe :**

- Utiliser Azure Key Vault pour stocker les identifiants sensibles permet une gestion sécurisée des secrets, essentielle pour les environnements de production.

###### **Configuration :**

- **Databricks est configuré pour accéder aux secrets stockés dans Azure Key Vault afin de se connecter en toute sécurité aux ressources de stockage.**

## **C. Transformation des Données avec Apache Spark**

Une fois la connexion établie, nous avons utilisé Apache Spark dans Azure Databricks pour appliquer des transformations sur les données stockées dans ADLS Gen2, permettant ainsi de préparer les données brutes pour des analyses plus poussées.

### **1.Montage des Conteneurs dans Databricks**

Nous avons monté les conteneurs d'ADLS Gen2 dans Databricks pour les rendre accessibles directement dans l'environnement de transformation.

```
▶ ✓ 12/9/2024 (<1s) 3

storage_name="bigdatadlgen2"
container_name="bronze"
access_key="Gpxe5iI7R82B8ceiOpMBKwhE4H/NdZb5gd4LeiaZS1q5oB4Z3PcPgLrIK/IZjzQ0smkjKZowakQk+AStNP+ebQ=="
mount_point_name="/mnt/bronze"
```

```
1 dbutils.fs.mount(
2     source = f"wasbs://{container_name}@{storage_name}.blob.core.windows.net/",
3     mount_point = mount_point_name,
4     extra_configs = {
5         f"fs.azure.account.key.{storage_name}.blob.core.windows.net": access_key
6     }
7 )
```

## **2. Préparation des Tables de Dimensions et des Tables de Faits :**

- Pour répondre aux besoins analytiques, nous avons transformé les données brutes en plusieurs tables de dimensions et tables de faits, en utilisant les modèles existants.

## **Principales Transformations :**

### **Table de Dimension : Établissements Éducatifs**

**Contient des informations détaillées sur les écoles, telles que :**

- Nom
- Emplacement géographique
- Niveaux d'enseignement
- Effectifs
- Table de Dimension : Dates

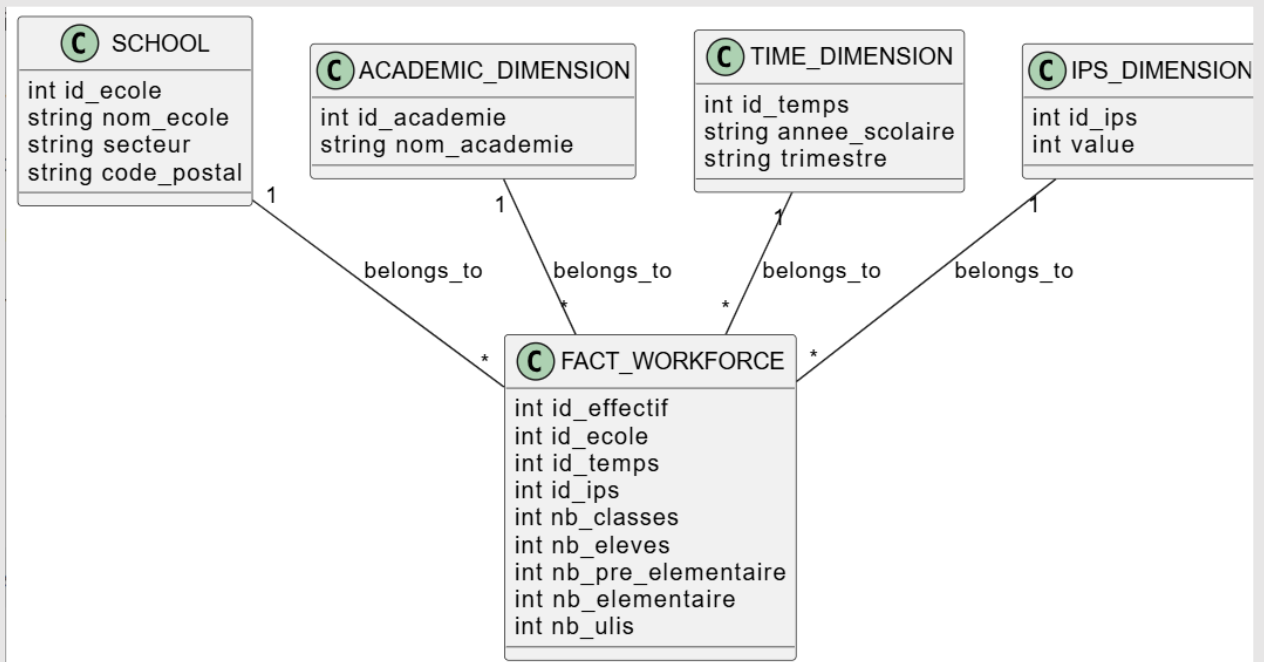
Inclut toutes les dates comprises entre le 1<sup>er</sup> janvier 2020 et le 30 décembre 2022, avec un accent particulier sur les périodes scolaires et les événements clés.

### **Tables de Faits : Données Éducatives**

**Ces tables contiennent des indicateurs et mesures essentiels, notamment :**

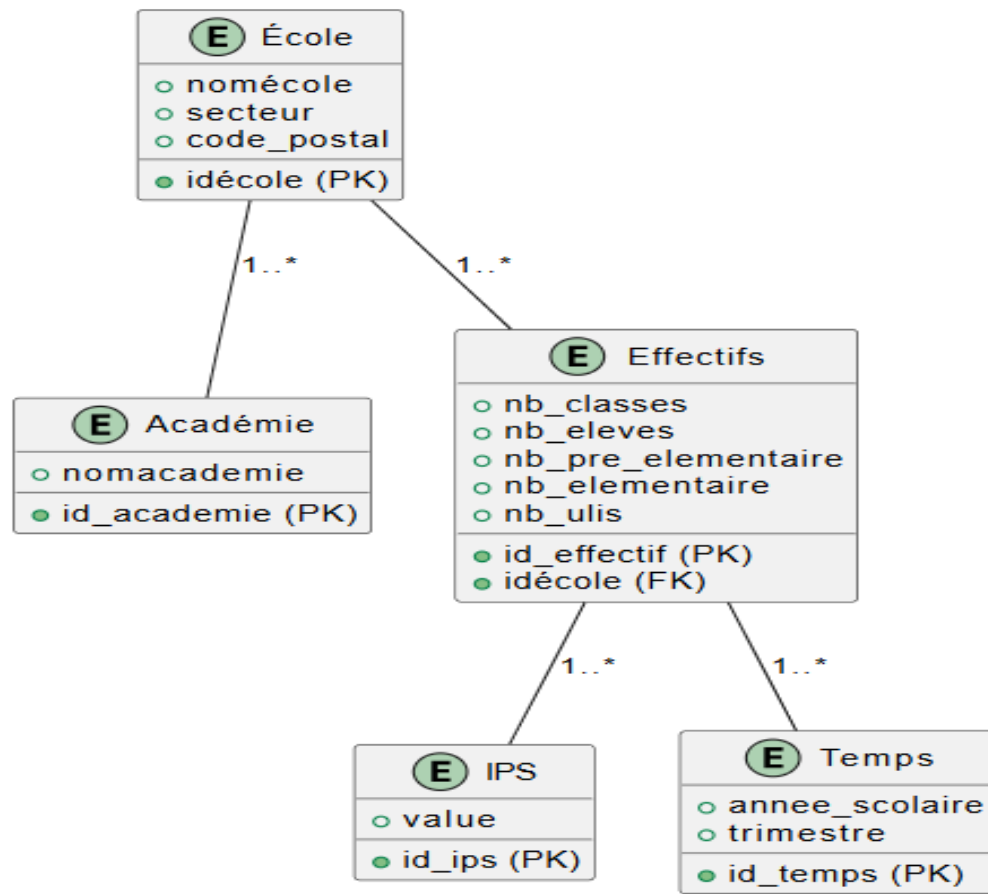
- Résultats scolaires (moyennes, taux de réussite)
- Effectifs d'élèves par niveau et par année
- Statistiques de performance des établissements
- Les tables de faits et dimensions existantes ont été utilisées pour structurer et enrichir ces transformations, garantissant une intégration cohérente avec les modèles analytiques.
- Les données transformées sont ensuite stockées dans le conteneur projet-education d'ADLS Gen2. Elles sont désormais prêtes pour des analyses avancées avec des outils comme Power BI, offrant une visibilité accrue sur les performances et les tendances éducatives.





## Modèle Conceptuel de Données (MCD)

- ✓ Le Modèle Conceptuel de Données (MCD) décrit les relations entre les entités principales qui permettent l'analyse et la gestion des données éducatives dans notre projet. Ce modèle inclut des entités telles que les étudiants, les enseignants, les cours, les inscriptions et les évaluations, qui sont essentielles à la gestion de l'information scolaire.



## **Entités principales**

- **École: Représente un établissement scolaire individuel.**
  - Attributs:
    - id\_école (clé primaire) : Identifiant unique de l'école
    - nom\_école : Nom de l'école
    - secteur : Secteur géographique
    - code\_postal : Code postal de l'école
- **Académie: Représente une académie, une circonscription administrative regroupant plusieurs écoles.**
  - Attributs:
    - id\_académie (clé primaire) : Identifiant unique de l'académie
    - nom\_académie : Nom de l'académie
- **Effectifs: Contient les informations quantitatives sur les élèves d'une école à un moment donné.**
  - Attributs:
    - id\_effectif (clé primaire) : Identifiant unique de l'effectif
    - nb\_classes : Nombre de classes
    - nb\_élèves : Nombre total d'élèves
    - nb\_pré\_élémentaire, nb\_élémentaire, nb\_ulis : Effectifs par niveau scolaire
    - id\_école (clé étrangère) : Identifie l'école à laquelle est rattaché l'effectif
    - IPS : Indice de position sociale de l'effectif
- **IPS: Représente l'indice de position sociale, un indicateur qui permet de classer les élèves en fonction du milieu social duquel ils sont issus.**
  - Attributs:
    - id\_ips (clé primaire) : Identifiant unique de l'indice
    - value : Valeur numérique de l'indice

- **Temps: Représente une période scolaire (année scolaire, trimestre).**

- Attributs:

- id\_temps (clé primaire) : Identifiant unique de la période
    - année\_scolaire : Année scolaire
    - trimestre : Trimestre

**Les relations entre les entités :**

- Une école est rattachée à une seule académie. (relation un-à-un entre École et Académie)
- Une école peut avoir plusieurs effectifs au cours du temps. (relation un-à-plusieurs entre École et Effectifs)
- Un effectif est associé à une école et à une période spécifique. (relation un-à-un entre Effectifs et École, et un-à-un entre Effectifs et Temps)
- Un effectif possède un indice de position sociale. (relation un-à-un entre Effectifs et IPS)

**Ce schéma permet de :**

- Suivre les effectifs des écoles au fil du temps.
- Analyser les inégalités sociales entre les établissements.
- Évaluer l'impact des politiques éducatives sur la réduction des inégalités.

## Notebook : Visualisation

- Nous avons utilisé le Notebook Dashboard de Databricks pour visualiser les données issues de nos analyses et mieux comprendre les relations entre les effectifs, les catégories d'IPS, les départements, et les niveaux scolaires. Cet outil nous a permis de créer des visualisations interactives, facilitant l'exploration des données et la communication des résultats.

### Visualisations Réalisées :

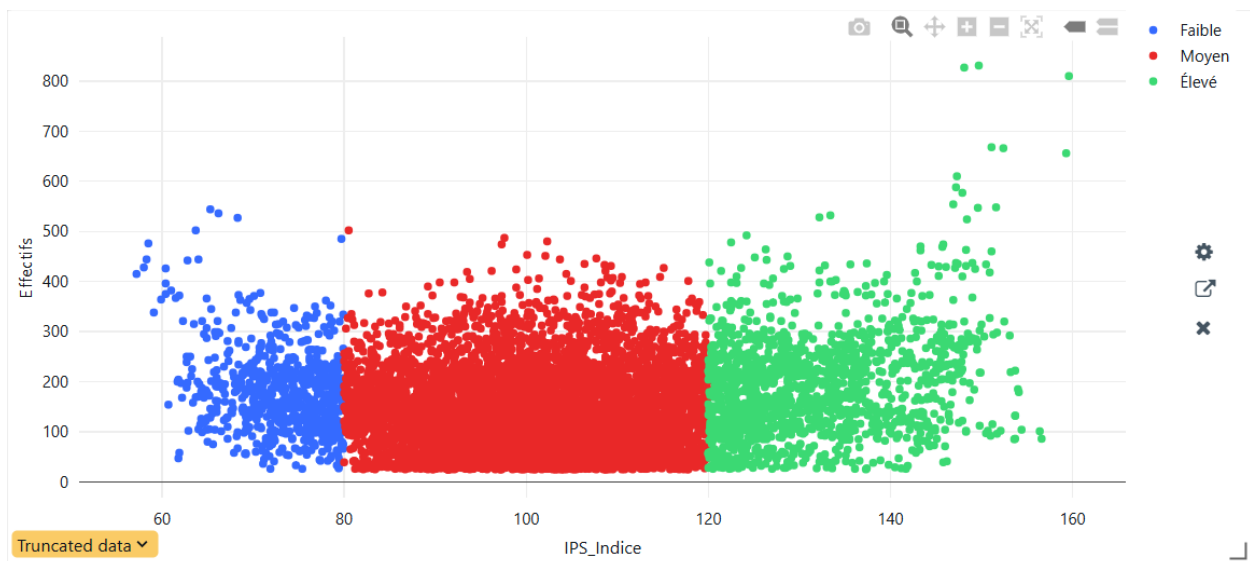
#### Répartition des Effectifs par Catégorie IPS

#### Représentation : Scatter Plot (diagramme de dispersion)

**Objectif : Identifier la distribution des effectifs en fonction des différentes catégories d'IPS.**

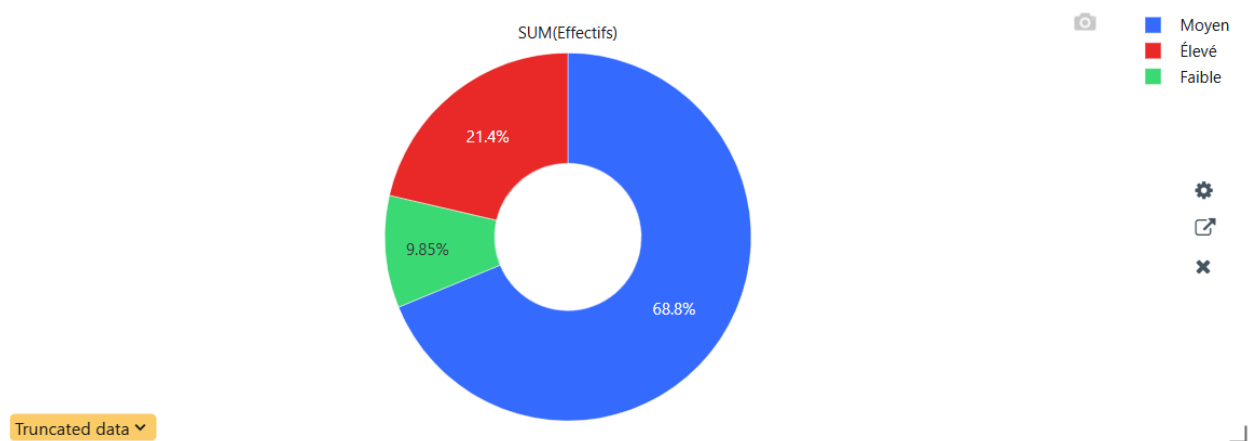
#### Résultat :

Cette visualisation met en évidence les écarts et les regroupements spécifiques au sein des données.



- Répartition des Effectifs par Niveau Scolaire et Catégorie IPS
- Représentation : Pie Chart (diagramme circulaire)
- **Objectif :** Illustrer la répartition des effectifs par niveau scolaire (école, collège) et leur correspondance avec les indices IPS.

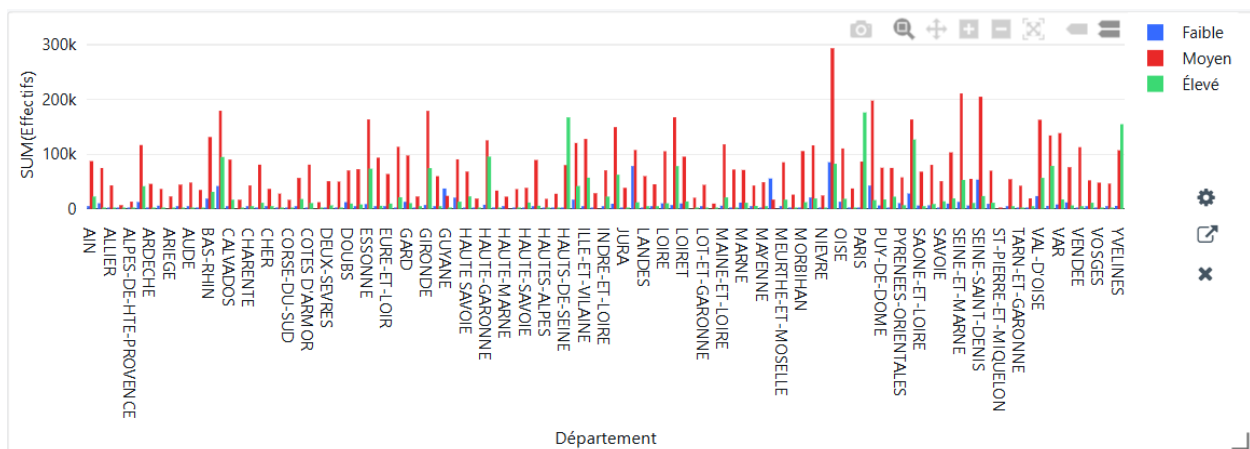
- **Résultat :** Ce graphique permet de visualiser les proportions relatives des effectifs selon les catégories



- Répartition des Effectifs par Département et Catégorie IPS
- Représentation : Bar Chart (diagramme en barres)

**Objectif :** Comparer les effectifs par département tout en tenant compte des indices IPS.

**Résultat :** Cette représentation facilite l'analyse des disparités géographiques présentes dans les données.



- Grâce à ces visualisations interactives, nous avons pu extraire des insights pertinents et mieux comprendre les dynamiques de répartition des effectifs en fonction de plusieurs dimensions clés.

## **Conclusion**

- ✓ Ce projet a permis de mettre en lumière l'impact significatif de l'indice de position sociale des élèves sur la répartition des effectifs dans les établissements scolaires en France. En utilisant des outils avancés de traitement de données sur la plateforme Microsoft Azure, nous avons pu collecter, stocker, transformer et analyser des données complexes liées à la situation socio-économique des élèves.
- ✓ Nos analyses ont révélé des disparités notables dans la répartition des élèves selon leur indice de position sociale, mettant en évidence comment des facteurs socio-éducatifs peuvent influencer les pratiques de gestion des classes et l'organisation scolaire.
- ✓ En créant des visualisations interactives dans Databricks, nous avons pu mieux comprendre et communiquer nos résultats, tout en soulignant l'importance d'une approche centrée sur les données pour l'évaluation et l'amélioration des politiques éducatives.
- ✓ Les résultats de cette étude peuvent servir de base pour des recommandations d'actions ciblées visant à réduire les inégalités dans l'éducation. Par exemple, les décideurs pourraient considérer des mesures d'accompagnement supplémentaires dans les écoles à forte concentration d'élèves venant de milieux défavorisés.
- ✓ En conclusion, ce projet démontre non seulement notre capacité à utiliser des technologies de Big Data, mais aussi l'importance de l'analyse de données dans la prise de décisions éclairées pour l'avenir du système éducatif français.
- ✓ Les résultats obtenus pourraient également être enrichis par des études futures, intégrant d'autres variables et facteurs socioculturels, afin de fournir une vue encore plus holistique sur les dynamiques en jeu dans l'éducation.