09/11/2024

# PROJET 5CLOUD

**Master of engineering II**


**Etudiant(e)s :**

Elisabeth NOKAM DASSI TAGUEMNE

Alexandre MOINDROT

Ibrar HAMOUDA

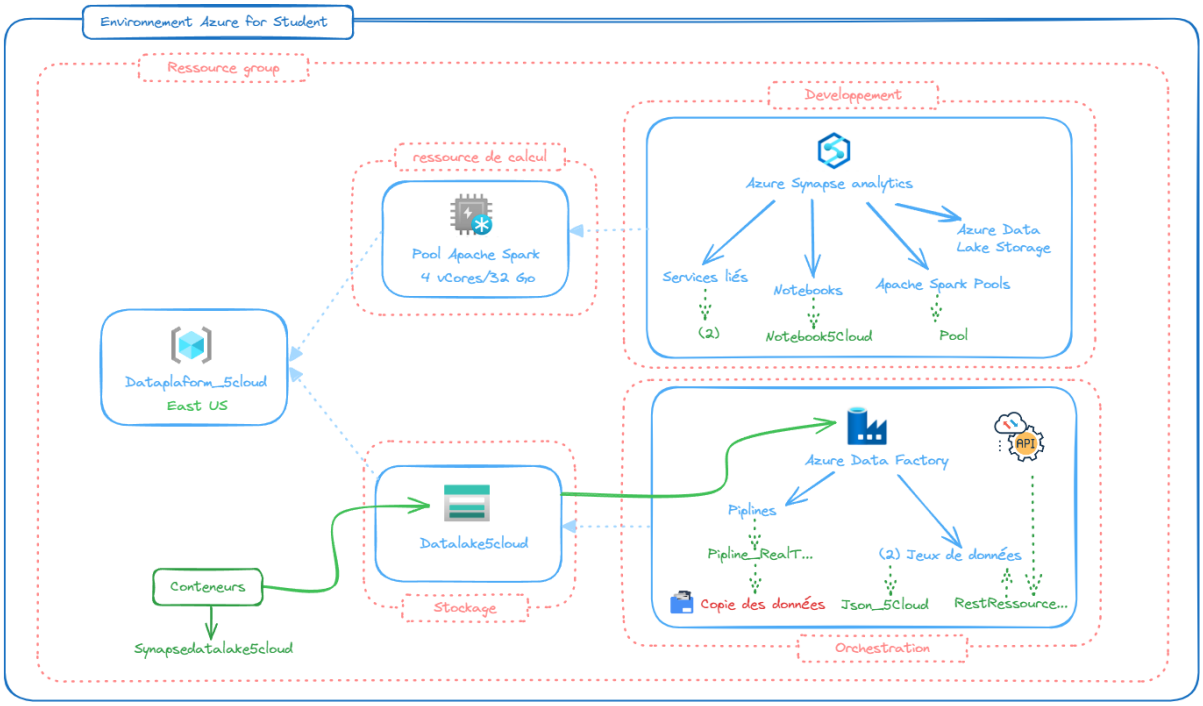


Année Académique 2024 - 2025

## Présentation du Projet :

**Objectif :** Créons une plateforme de données unifiée pour l'analyse d'événements en temps réel à l'aide de Microsoft Azure (ou d'outils locaux alternatifs).

### Architecture Azure :




Accès au repo Github : https://github.com/deszr/5CLOU.git

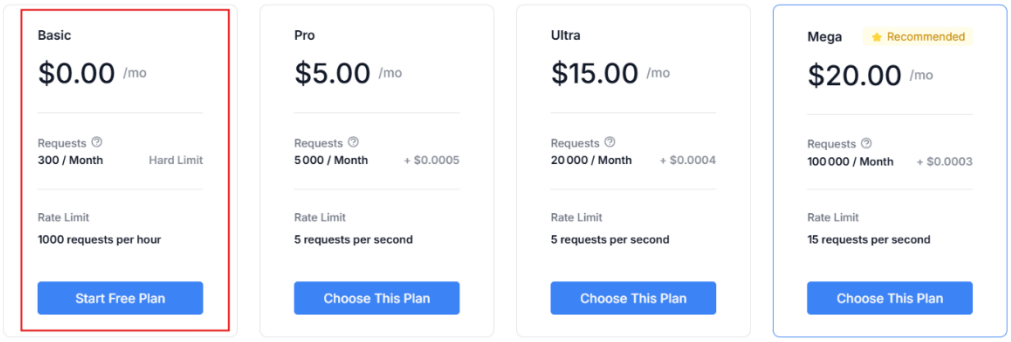
### Étape 1 : Choix de la source de données

Comme source de données en temps réel, nous avons opté pour **Yahoo Finance** qui est un flux boursier.

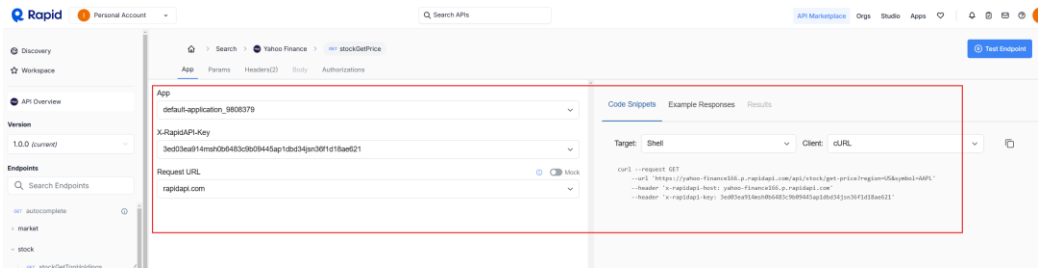
Pour accéder aux données de Yahoo Finance, nous le ferons via une API **RapidAPI** :

**Étapes pour Accéder à l'API Yahoo Finance via RapidAPI**

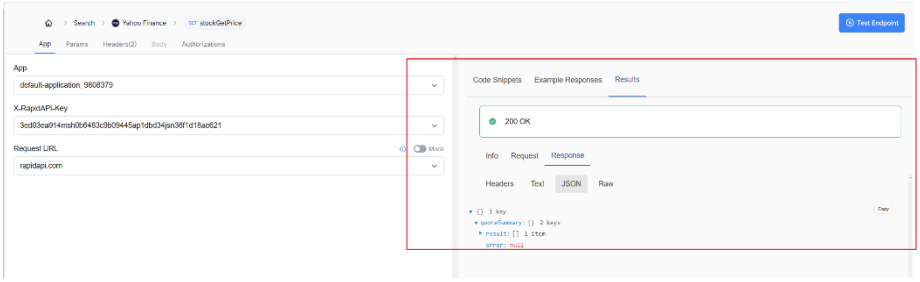
- Création d'un compte sur RapidAPI
- Sélectionner l'API Yahoo Finance sur RapidAPI
- S'abonner à un Plan de l'API



- Obtention de la clé donnant accès à l'API




- Testons un Endpoint de l'API Yahoo Finance avec RapidAPI

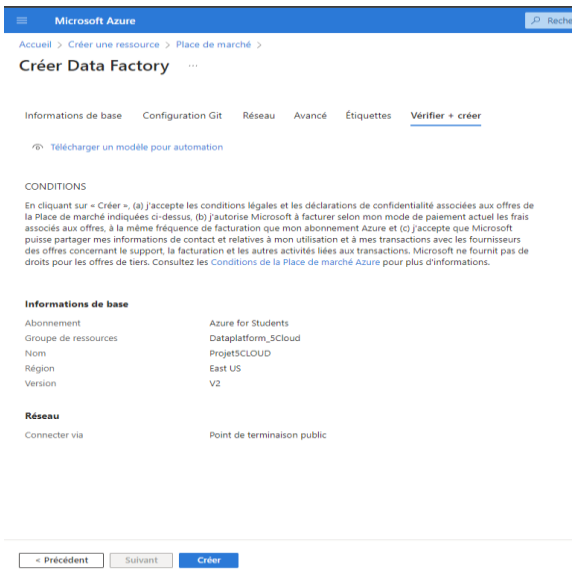



## Étape 2 : Configuration du Pipeline d'Ingestion de Données

**Objectif :** Configurons un pipeline d'ingestion capable de récupérer les données en continu et de les stocker pour effectuer traitement.

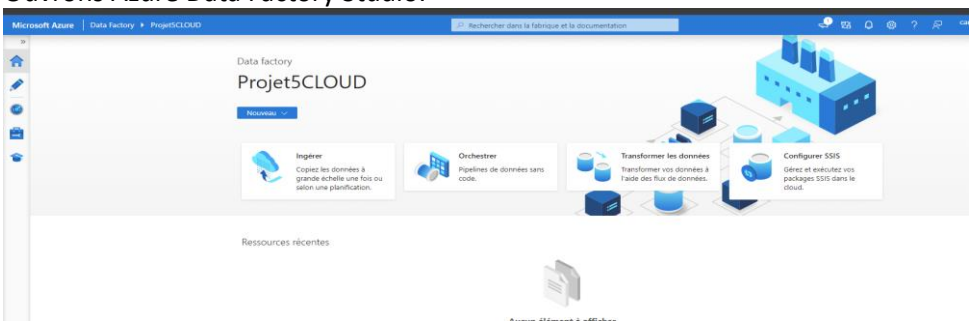
Créons un service **Data Factory** dans le portail Azure :

- Recherchons et créons une instance d'**Azure Data Factory**.
- Assignons un nom et un groupe de ressources au service ensuite cliquons sur « **Créer** »

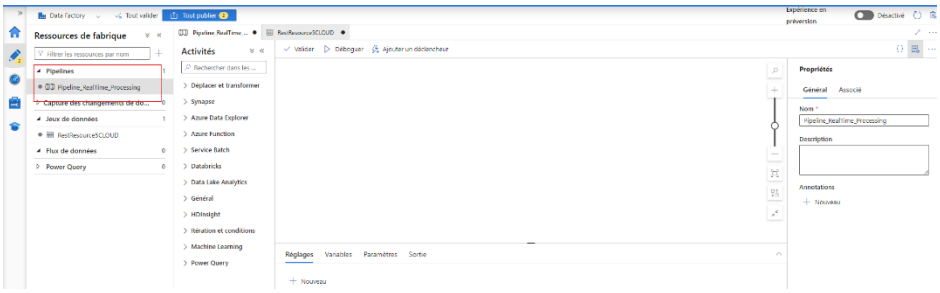



**Configurons le pipeline d'ingestion**

Ouvrons Azure Data Factory Studio.

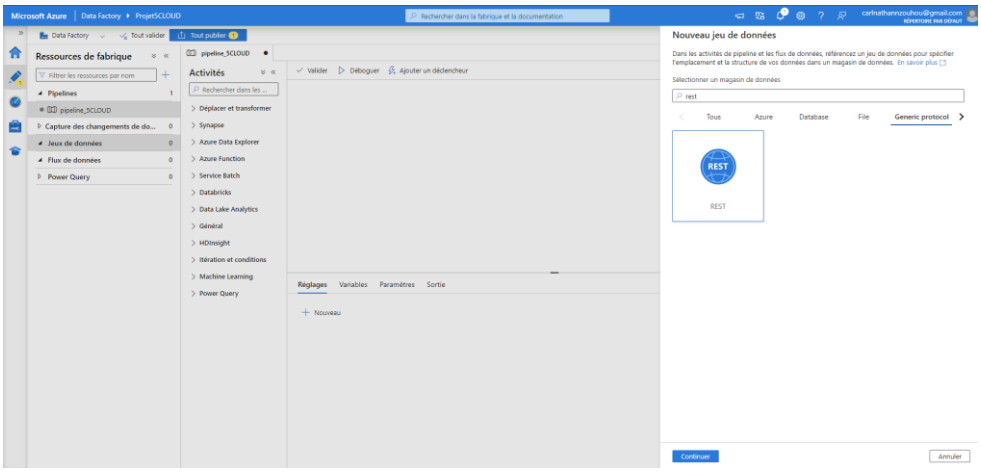



Créons un nouveau pipeline

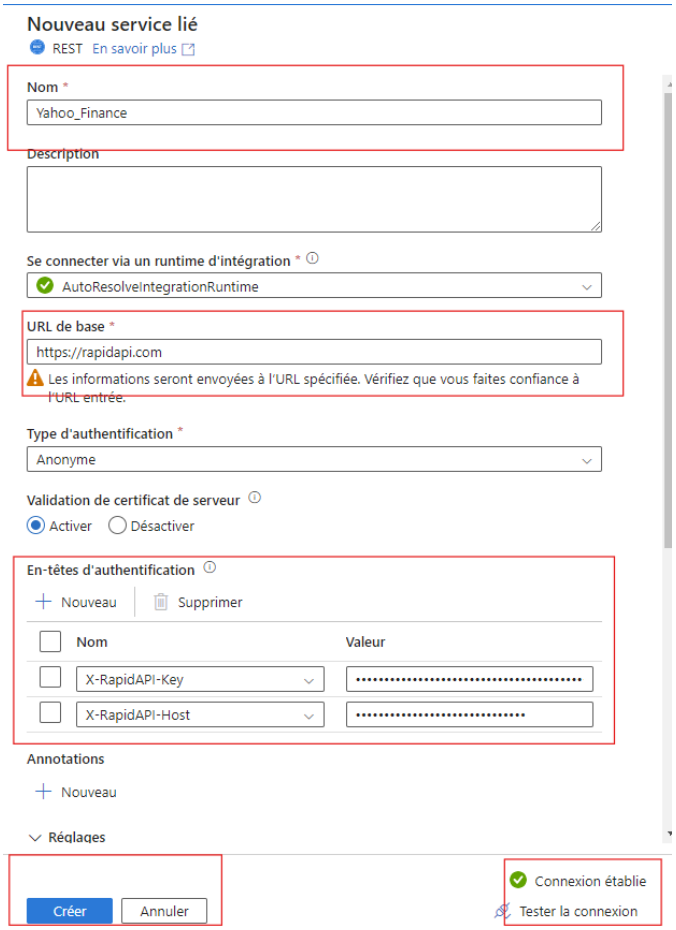


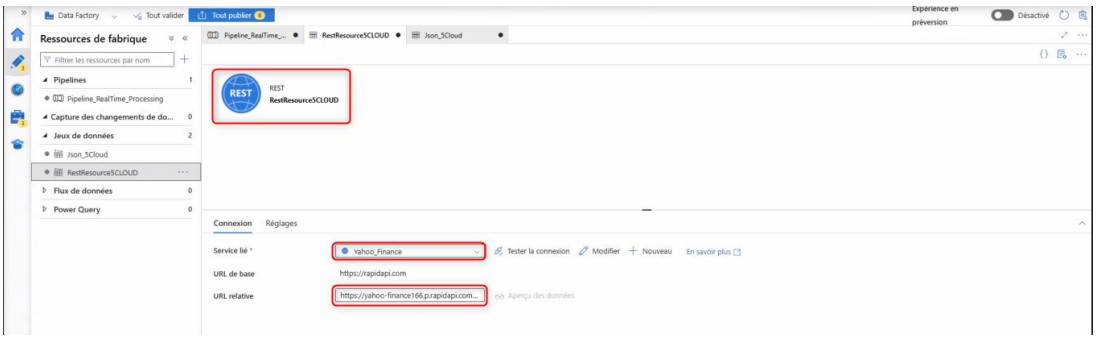

Ajoutons notre jeu de données

Pour cela, nous avons configuré la source en sélectionnant “REST”(**generic protocol**) :




Ensuite nous renseignons L’URL de base (celle de l’API), ainsi que les paramètres d’authentification. Une fois cela fait, on se rassure que la connexion a bien été établie. Ainsi on peut créer notre connexion.



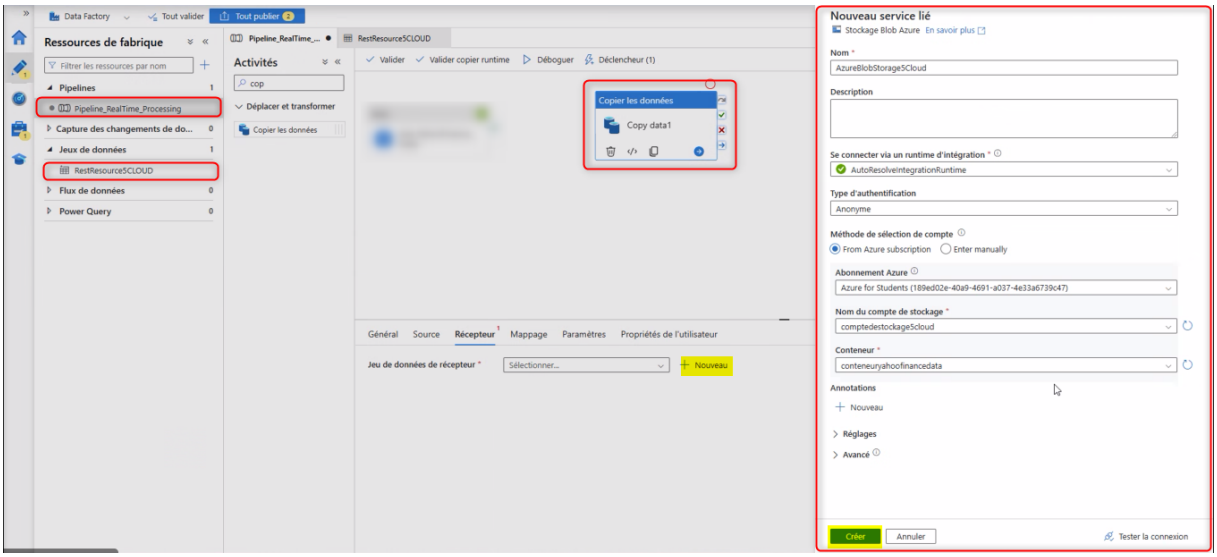




Revenons à notre Pipeline « **Pipeline_RealTime_Processing** » créé au départ :

**Configurons l'Activité Copy Data** :

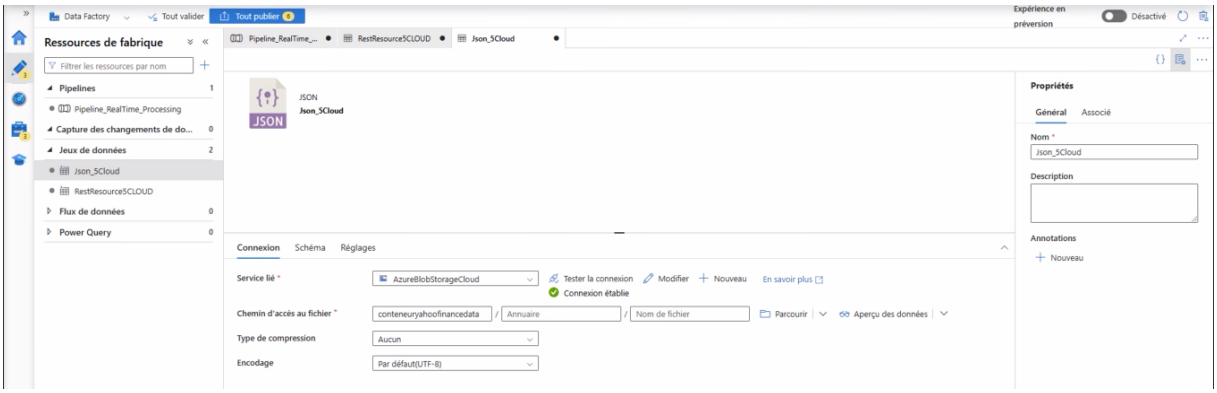
A partir du « **Pipeline_RealTime_Processing** » Dans l'onglet **Source** de l'activité "Copy Data", on configure la connexion vers la source de données depuis un Blob Storage avec du « CSV » comme format de données.

Dans les paramètres de configuration du « copy data » on configure notre jeu de données de récepteur en cliquant sur « Nouveau »

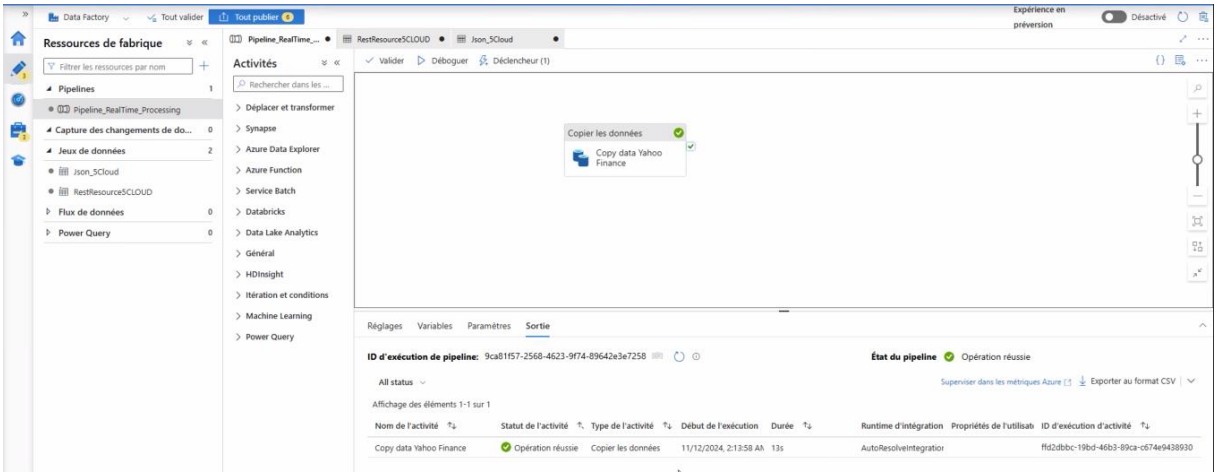



**Testons l'Activité de notre jeu de données :**

A partir du jeu de données « Json_5cloud » on clique sur tester la connexion :




Ensuite on effectue un test de debug :




On peut constater que l'opération a réussie.

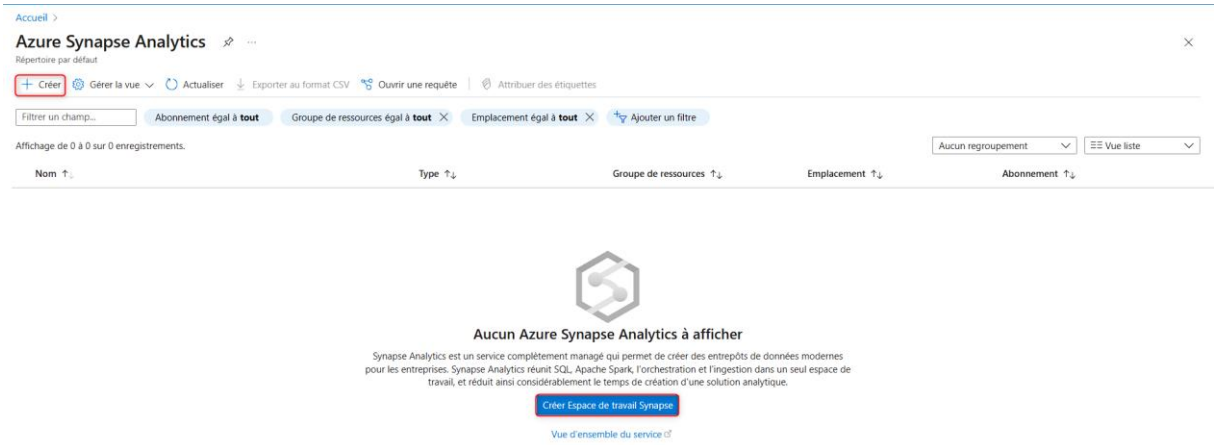
Vérification des données dans Azure Blob Storage :

**Etape 3 : Traitement des données en temps réel avec Synapse Analytics**

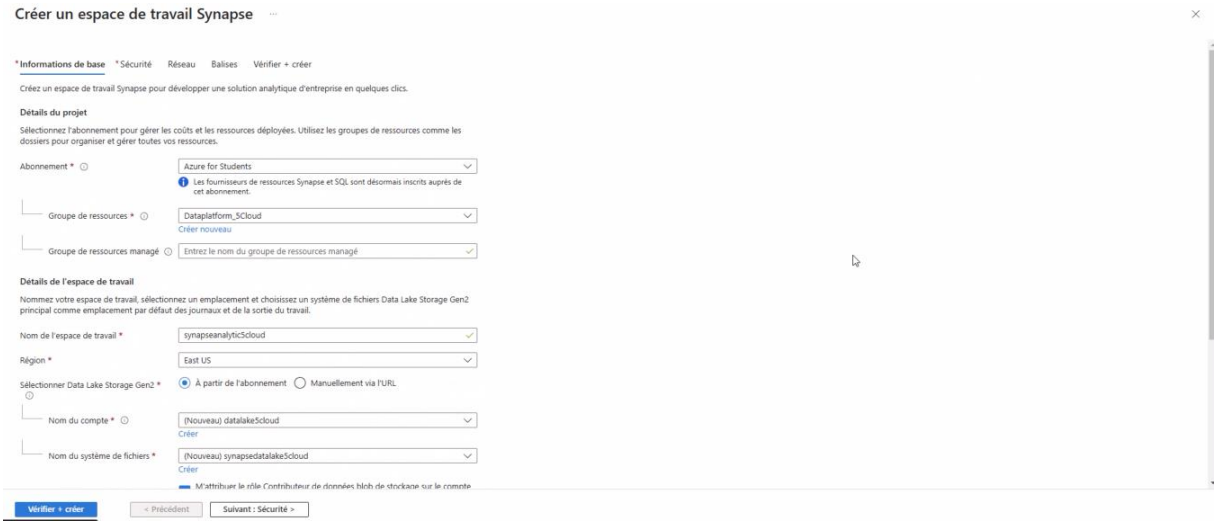
**3.1 : Configuration Synapse Workspace**

On créait un espace de travail dans Synapse :

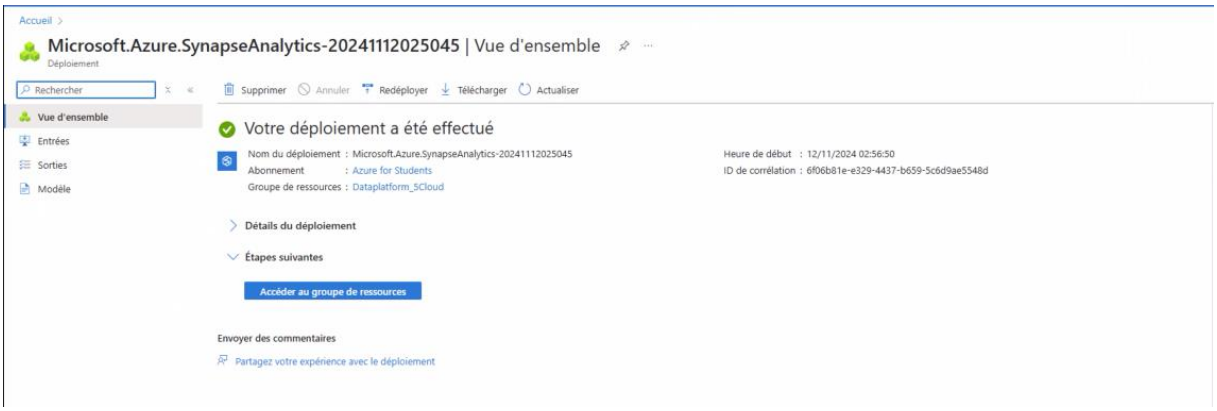
Dans Azure Sélectionne Azure Synapse Analytics puis on clique sur créer :



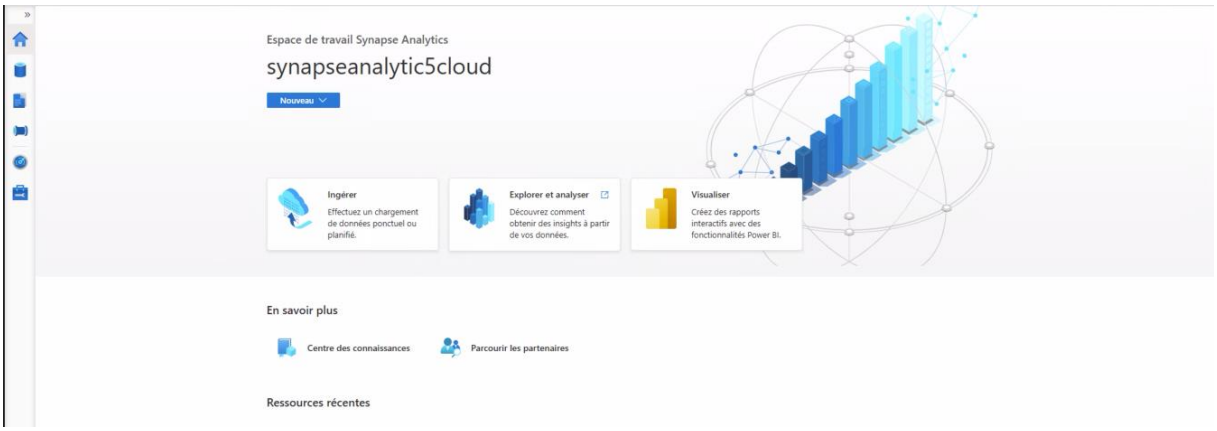

On renseigne ensuite les informations de la synapse :



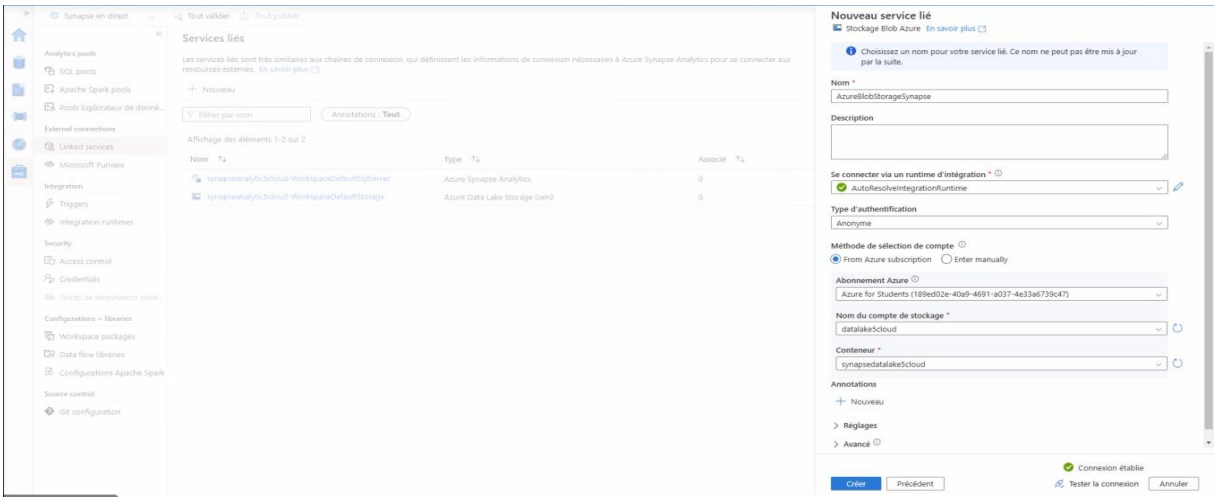

Déploiement de la ressource :



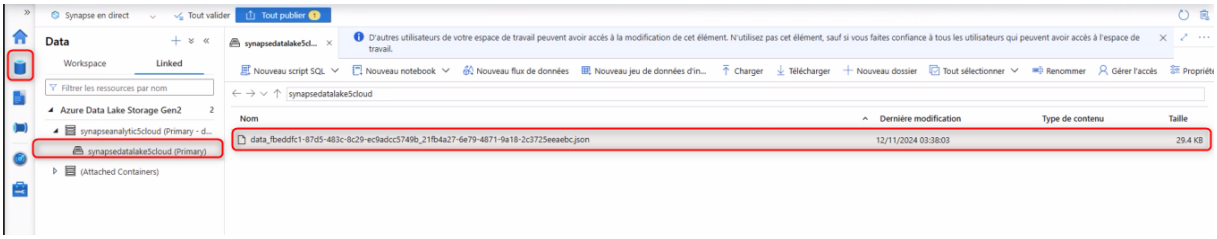

Visualisation de l'environnement Azure Synapse Studio :



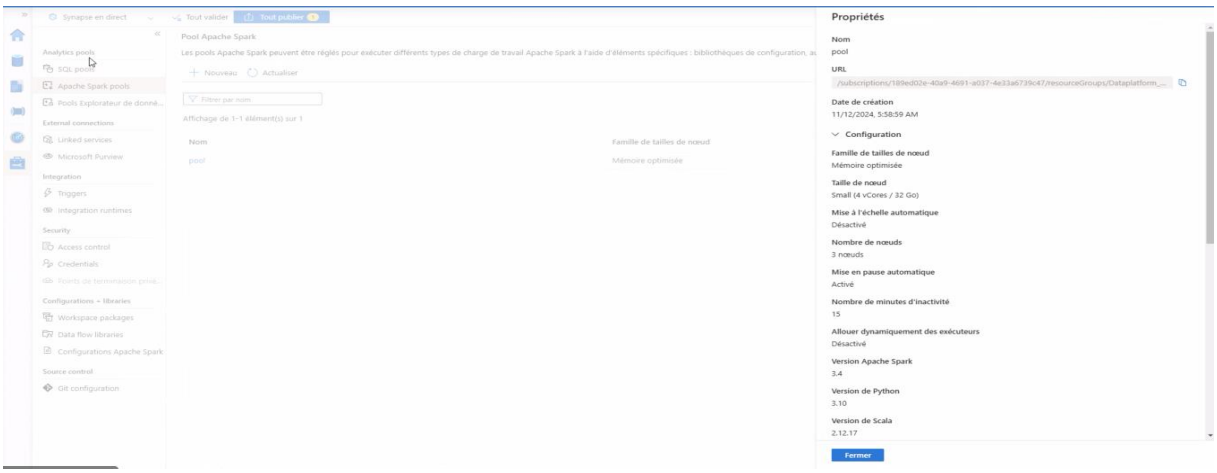

Connection de Synapse Workspace à notre Blob Storage :



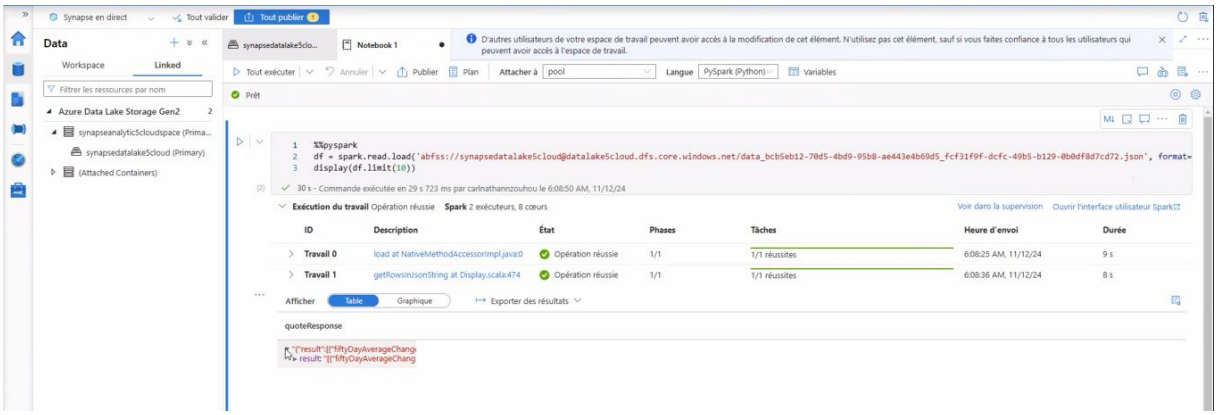

Vérification des données :



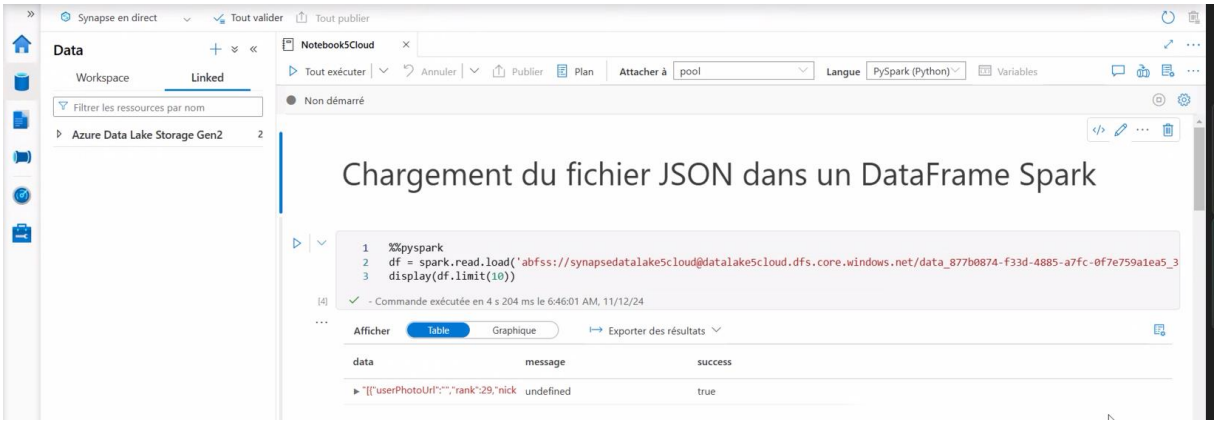

### 3.2 : Configuration d'un pool Spark pour le traitement des données avec python



**Chargement du notebook :**




**Chargement du data Frame Spark :**

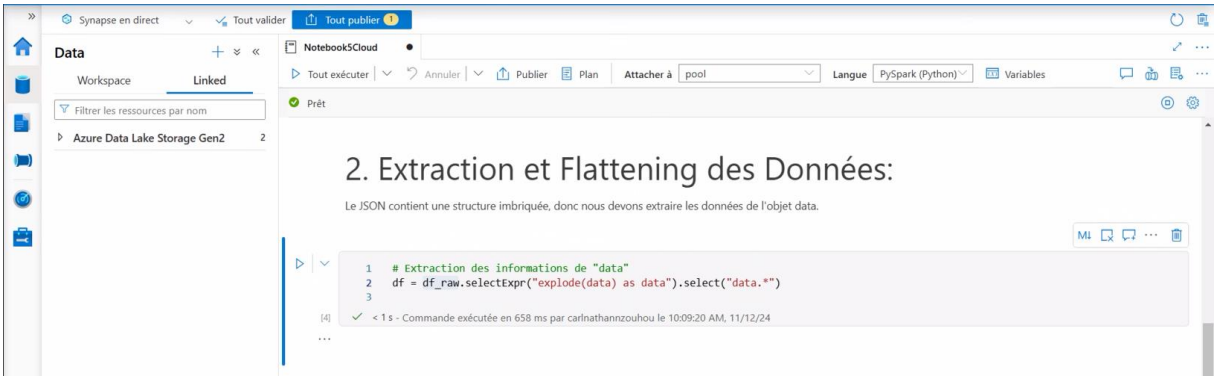



Rédaction des requêtes de transformation qui nettoient et formatent les données pour la visualisation :

### 3.3 : Traitement des données avec python

### Extraction et Flattening des Données

Le JSON contient une structure imbriquée, donc nous devons extraire les données de l'objet data

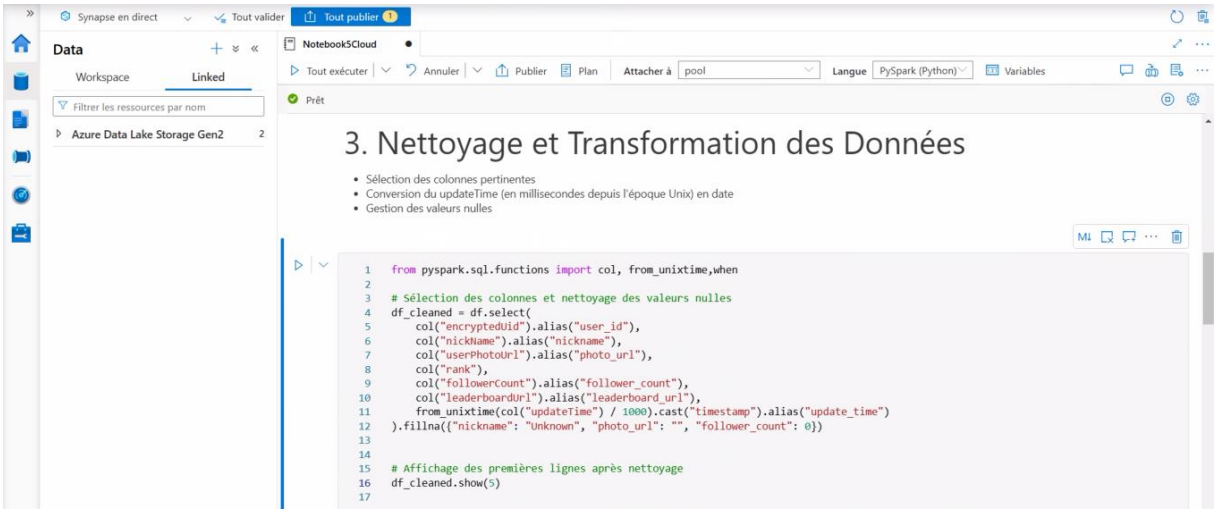



### Nettoyage et Transformation des Données

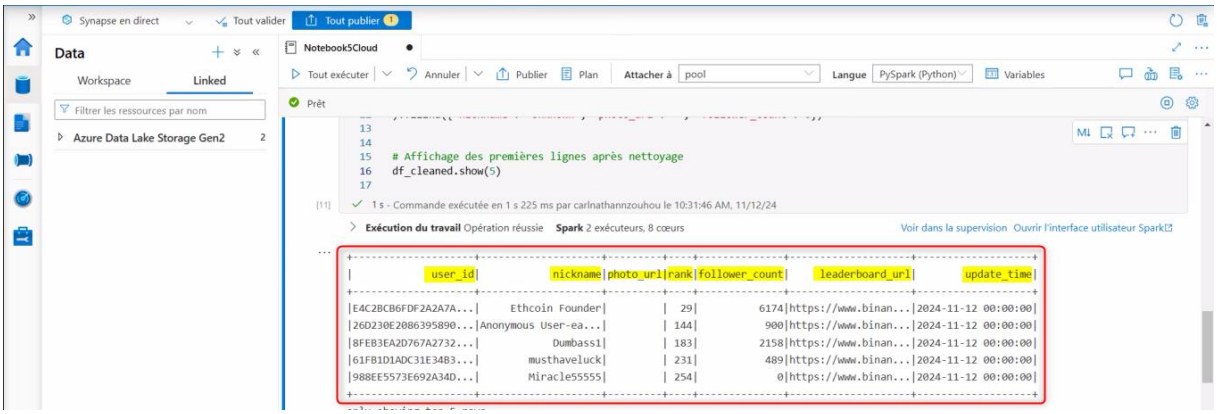
### Cette partie se fera en trois étapes :

- Sélection des colonnes pertinentes

- Conversion de l’update Time (en millisecondes depuis l'époque Unix) en date
- Gestion des valeurs nulles

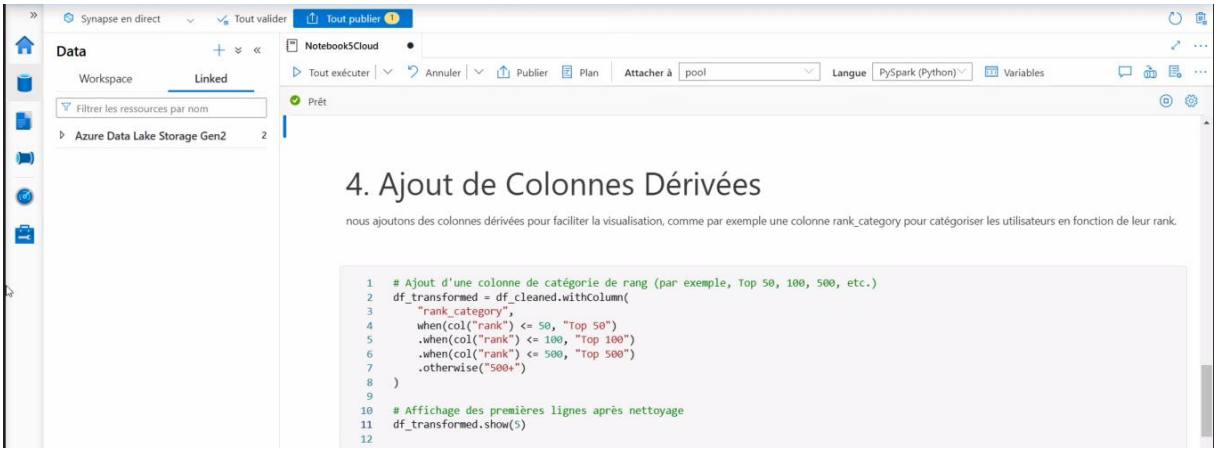



Aperçu des données des 5 premières lignes uniquement

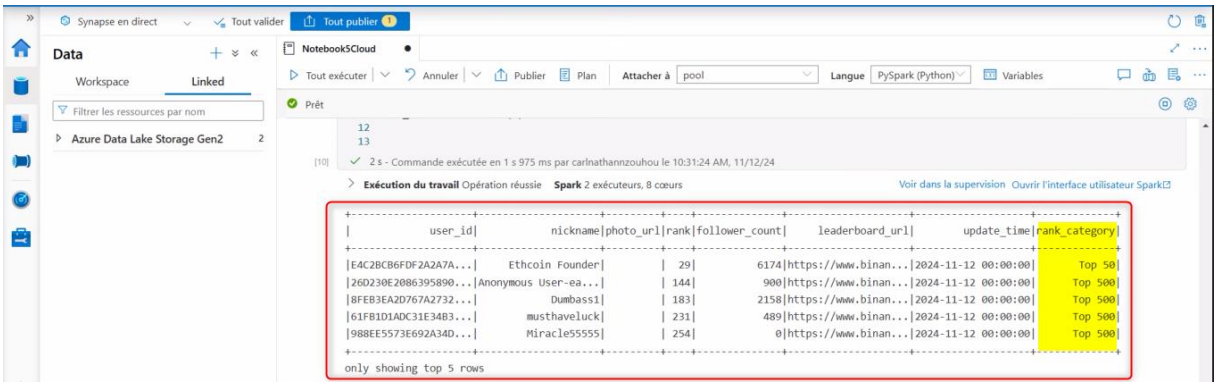



**Ajout de Colonnes Dérivées**

Pour cette partie il nous a fallu ajouter des colonnes dérivées pour faciliter la visualisation, comme une colonne **rank_category** pour catégoriser les utilisateurs en fonction de leur **rank**

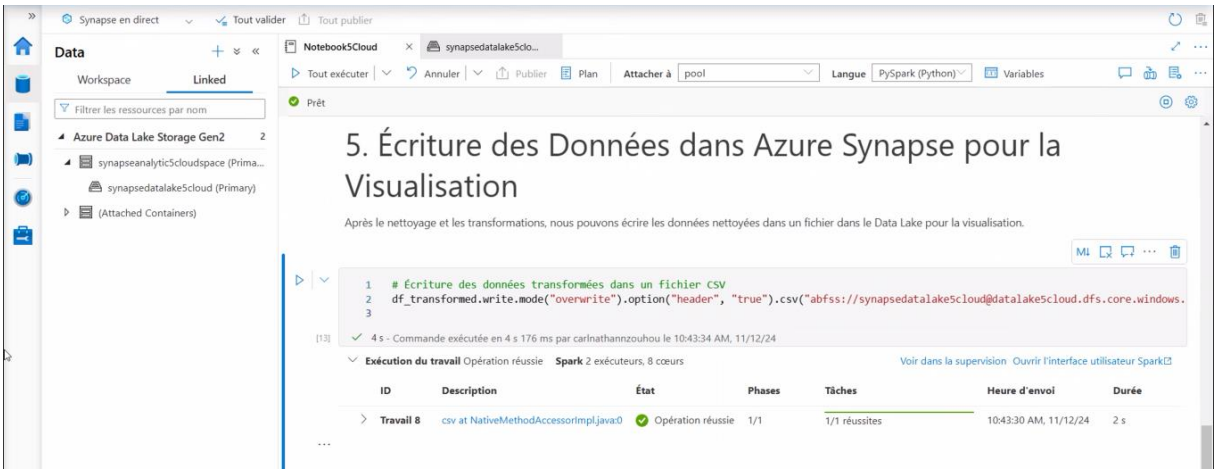



Aperçu des données des 5 premières uniquement

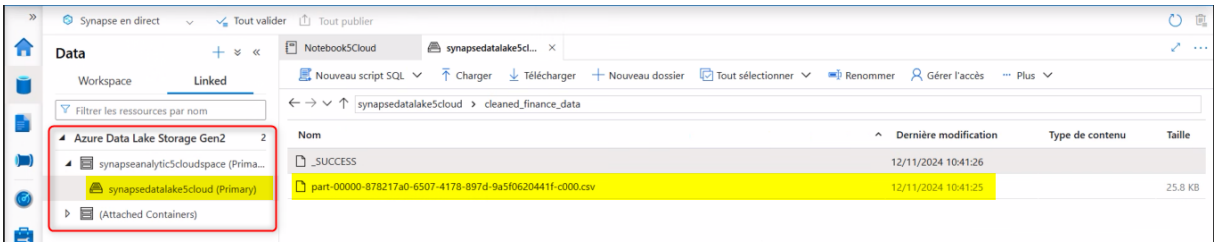



**Écriture des Données dans Azure Synapse pour la Visualisation :**

Après nettoyage et transformation des données, nous écrivons les données nettoyées dans une table **Synapse**




Le résultat peut être visible dans notre espace de stockage « **Azure Data Lake Storage Gen2** »



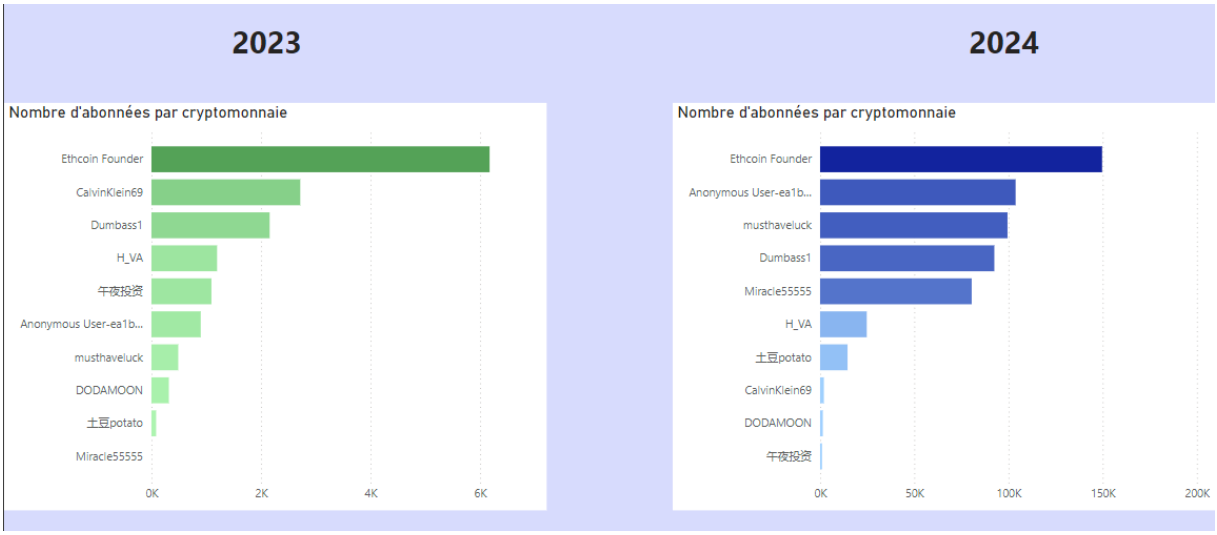

**Étape 4 : Visualisation des données avec Power BI**

Pour cette partie visualisation des données, nous allons utiliser Power BI.

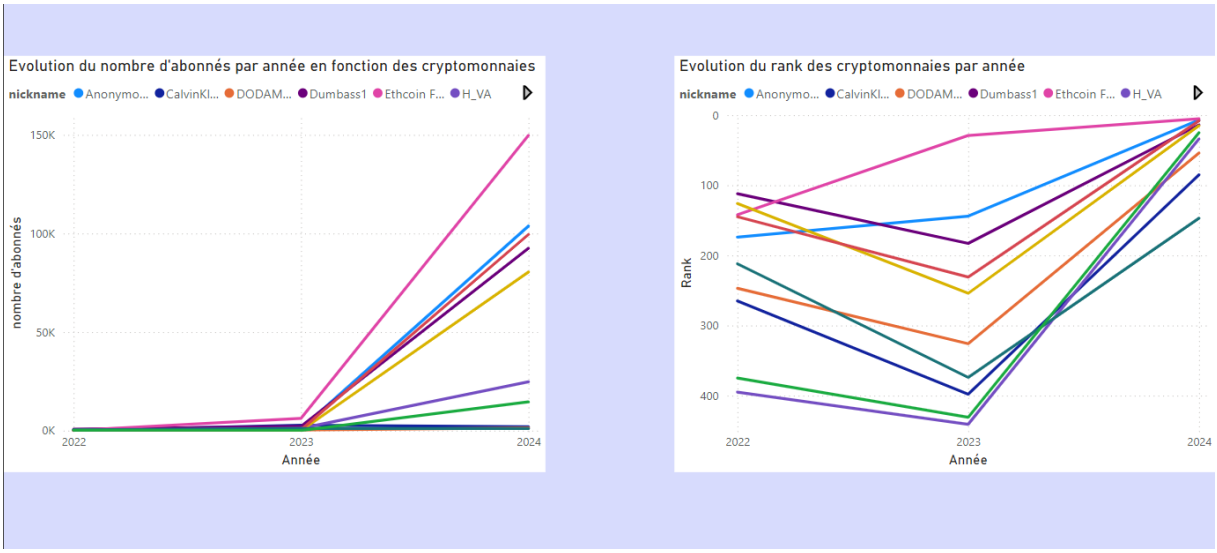
Voici notre rapport :

Nous avons créé des TOP pour identifier les cryptomonnaies les mieux classé.

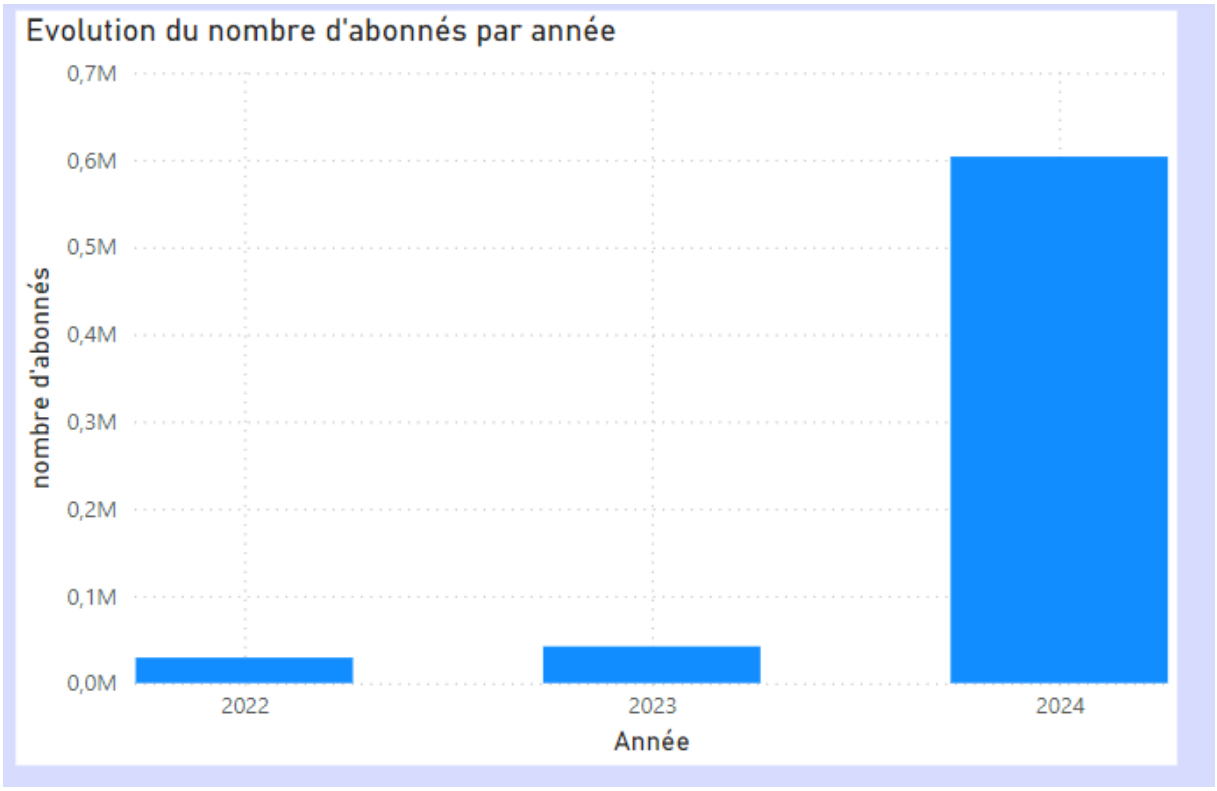

Nous avons fait la même chose pour le nombre d'abonnés.



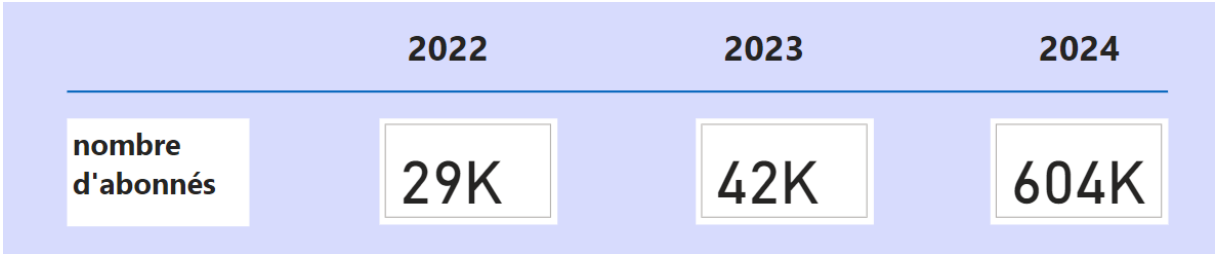

Nous avons réalisé des graphiques sur l'évolution des cryptomonnaies en fonction du rank et du nombre d'abonnés par années.



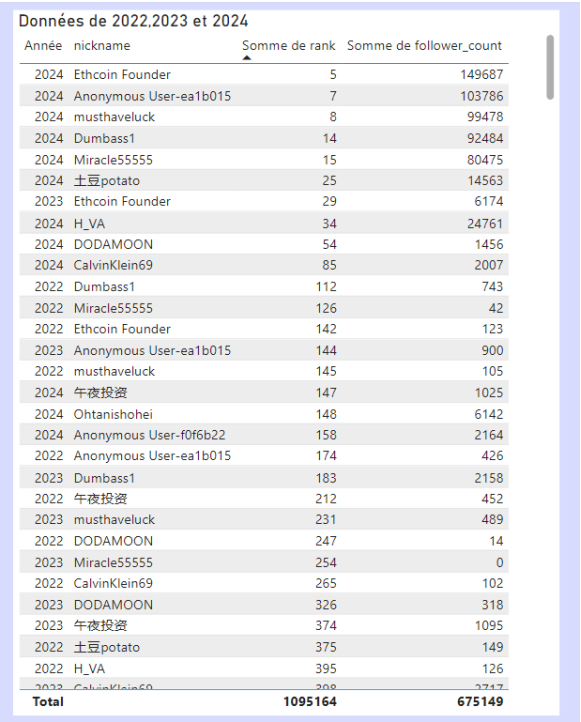

Nous avons créé un histogramme pour représenter l'évolution du nombre d'abonnés total par années.




C'est le même graphique que le précédent mais sous forme différente avec des indicateurs.



Enfin nous avons les données sous forme tabulaire.



Données de 2022,2023 et 2024

| Année | nickname | Somme de rank | Somme de follower_count |
|---|---|---|---|
| 2024 | Ethcoin Founder | 5 | 149687 |
| 2024 | Anonymous User-ea1b015 | 7 | 103786 |
| 2024 | musthaveluck | 8 | 99478 |
| 2024 | Dumbass1 | 14 | 92484 |
| 2024 | Miracle55555 | 15 | 80475 |
| 2024 | 土豆potato | 25 | 14563 |
| 2023 | Ethcoin Founder | 29 | 6174 |
| 2024 | H_VA | 34 | 24761 |
| 2024 | DODAMOON | 54 | 1456 |
| 2024 | CalvinKlein69 | 85 | 2007 |
| 2022 | Dumbass1 | 112 | 743 |
| 2022 | Miracle55555 | 126 | 42 |
| 2022 | Ethcoin Founder | 142 | 123 |
| 2023 | Anonymous User-ea1b015 | 144 | 900 |
| 2022 | musthaveluck | 145 | 105 |
| 2024 | 午夜投资 | 147 | 1025 |
| 2024 | Ohtanishohei | 148 | 6142 |
| 2024 | Anonymous User-f0f6b22 | 158 | 2164 |
| 2022 | Anonymous User-ea1b015 | 174 | 426 |
| 2023 | Dumbass1 | 183 | 2158 |
| 2022 | 午夜投资 | 212 | 452 |
| 2023 | musthaveluck | 231 | 489 |
| 2022 | DODAMOON | 247 | 14 |
| 2023 | Miracle55555 | 254 | 0 |
| 2022 | CalvinKlein69 | 265 | 102 |
| 2023 | DODAMOON | 326 | 318 |
| 2023 | 午夜投资 | 374 | 1095 |
| 2022 | 土豆potato | 375 | 149 |
| 2022 | H_VA | 395 | 126 |
| **Total** | | **1095164** | **675149** |

Après analyse des données nous pouvons conclure qu'il y a une énorme augmentation du nombre d'abonnés sur les cryptomonnaie que nous avons étudié.

En 2023 on remarque une tendance générale en diminution pour les rank des cryptomonnaies sélectionné puis une augmentation en 2024.