[Pipeline de Traitement des Données pour l'Analyse de la Consommation Énergétique des Entreprises]

[JUILLET 2024]

Auteur : [Bya Amine]
[Mounir Menguelti]

efrei PARIS
École d'ingénieurs du numérique

Table des matières

Introduction :

Dans le cadre de ce projet, nous avons développé une solution complète pour analyser la consommation énergétique des entreprises en utilisant les données ouvertes fournies par Enedis. L'objectif principal de ce projet était de créer un pipeline de traitement des données permettant de charger, transformer, et visualiser ces données en utilisant les services Azure.

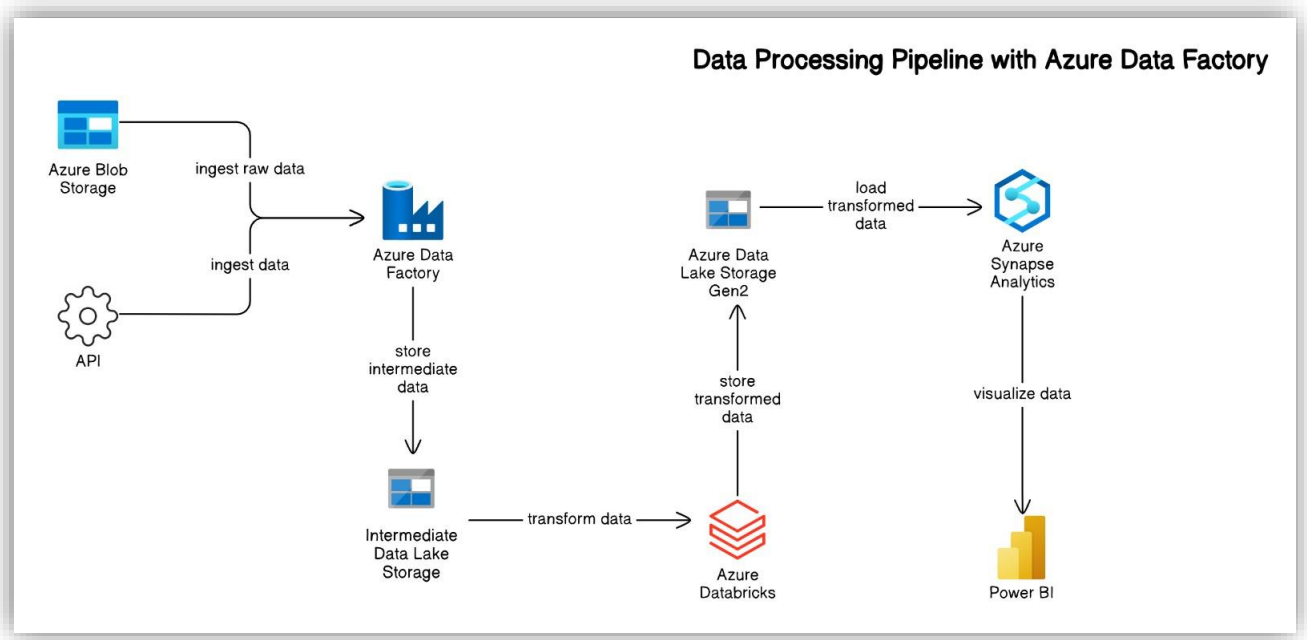
Enedis, principal gestionnaire du réseau de distribution d'électricité en France, met à disposition des données ouvertes sur la consommation énergétique à travers son portail Open Data. Ces données sont une source précieuse d'informations pour comprendre les habitudes de consommation des entreprises et identifier des opportunités d'optimisation énergétique.

Pour mener à bien ce projet, nous avons utilisé plusieurs services Azure, notamment Azure Data Factory pour l'ingestion des données, Azure Databricks pour la transformation des données, Azure Data Lake Storage Gen2 pour le stockage des données transformées, et Azure Synapse Analytics pour l'analyse des données. Enfin, Power BI a été utilisé pour la visualisation des résultats, permettant ainsi de présenter les insights de manière claire et interactive.

Ce rapport détaillera les différentes étapes de notre solution, de l'ingestion des données à la visualisation des résultats, en passant par les transformations effectuées. Il présentera également les résultats obtenus et les conclusions tirées de cette analyse, ainsi que les bénéfices potentiels pour les entreprises en termes d'optimisation de leur consommation énergétique.

Architecture de la Solution

L'architecture de notre solution s'articule autour de plusieurs services Azure, chacun jouant un rôle clé dans le processus de traitement des données, de l'ingestion à la visualisation. Le schéma ci-dessous illustre les différentes étapes de notre pipeline de traitement des données.



Ingestion des Données

Les données initiales sont ingérées à partir de deux sources principales : Azure Blob Storage et une API. Azure Data Factory (ADF) est utilisé pour orchestrer et automatiser ce processus. ADF permet de connecter, extraire et charger les données de manière efficace et sécurisée. Les types de données ingérées incluent des fichiers CSV et JSON contenant des informations détaillées sur la consommation énergétique des entreprises.

Stockage des Données Intermédiaires

Pour gérer les étapes de transformation successives, les données ingérées sont stockées dans un espace de stockage intermédiaire, Azure Data Lake Storage. Ce stockage permet de gérer efficacement les volumes de données et de garantir la disponibilité des données pour les transformations ultérieures.

Transformation des Données

Les données sont ensuite transformées à l'aide d'Azure Databricks. Les transformations incluent le nettoyage des données, la normalisation des formats, l'agrégation des informations pertinentes. Azure Databricks offre un environnement de traitement de données puissant et scalable, adapté aux besoins de ce projet.

Stockage des Données Transformées

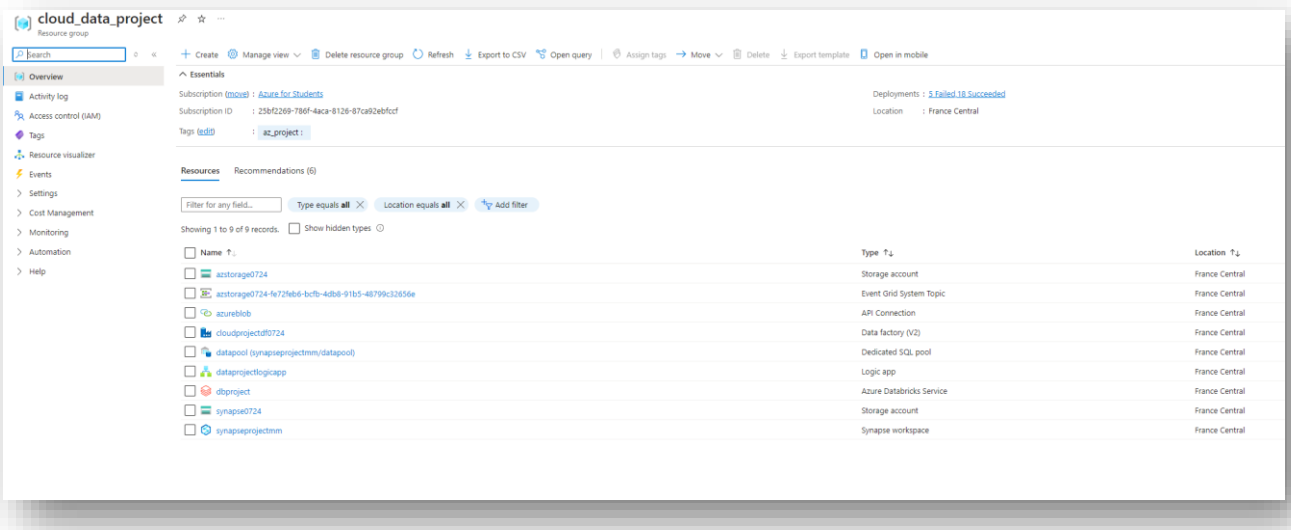
Les données transformées sont ensuite stockées dans Azure Data Lake Storage Gen2. Ce service de stockage est hautement performant et scalable, offrant une solution sécurisée pour conserver les données prêtes à être analysées.

Chargement et Analyse des Données

Les données transformées sont chargées dans Azure Synapse Analytics pour des analyses plus approfondies. Azure Synapse Analytics permet l'exécution de requêtes SQL complexes, l'analyse en temps réel et l'intégration de données provenant de diverses sources.

Visualisation des Données

Enfin, les résultats de l'analyse sont visualisés à l'aide de Power BI. Power BI permet de créer des tableaux de bord interactifs et dynamiques, facilitant la visualisation des tendances de consommation, des pics de demande, et des comparaisons entre différentes entreprises. Ces visualisations offrent des insights précieux pour optimiser l'utilisation énergétique et identifier des opportunités d'économie.
Cette architecture modulaire et scalable assure une gestion efficace du pipeline de traitement des données, de l'ingestion à la visualisation, en passant par les transformations et l'analyse. Chaque composant de la solution joue un rôle crucial pour garantir l'intégrité, la disponibilité et la pertinence des données tout au long du processus.



Ingestion des Données

L'ingestion des données est la première étape cruciale de notre pipeline de traitement. Cette étape consiste à collecter les données brutes provenant de différentes sources et à les intégrer dans notre système pour un traitement ultérieur. Pour ce projet, nous avons utilisé Azure Data Factory (ADF) pour orchestrer ce processus de manière efficace et automatisée.

Sources de Données

Azure Blob Storage :

Ce service de stockage d'objets est utilisé pour stocker des fichiers contenant des données sur la consommation énergétique. Ces fichiers sont généralement au format CSV ou JSON et sont régulièrement mis à jour avec de nouvelles informations.

API Enedis Open Data :

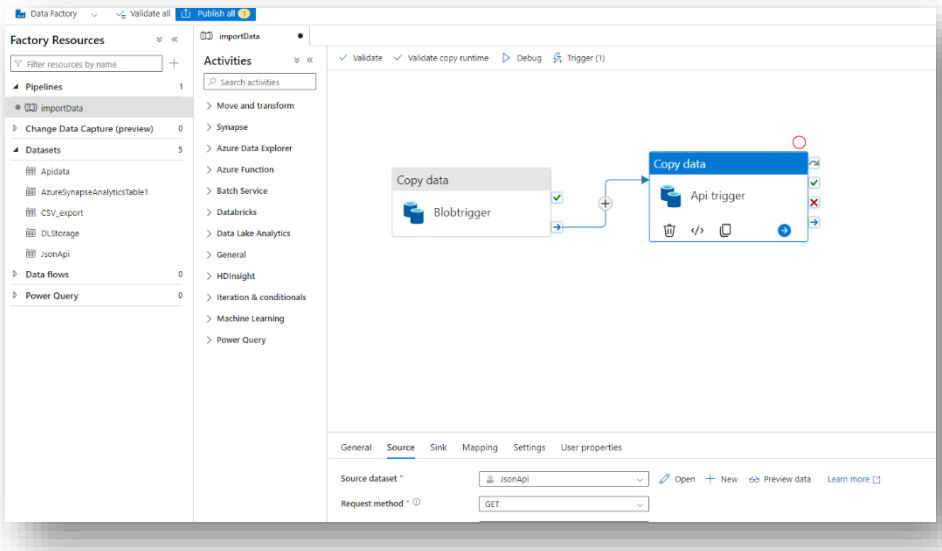
Enedis fournit une API permettant d'accéder aux données ouvertes sur la consommation énergétique. Cette API est utilisée pour récupérer des données en temps réel ou selon un calendrier défini, assurant ainsi que notre pipeline dispose toujours des informations les plus récentes.

Utilisation d'Azure Data Factory

Azure Data Factory est un service de gestion de flux de travail de données qui permet de créer, planifier et orchestrer des pipelines de données. Voici comment ADF a été utilisé dans notre projet :

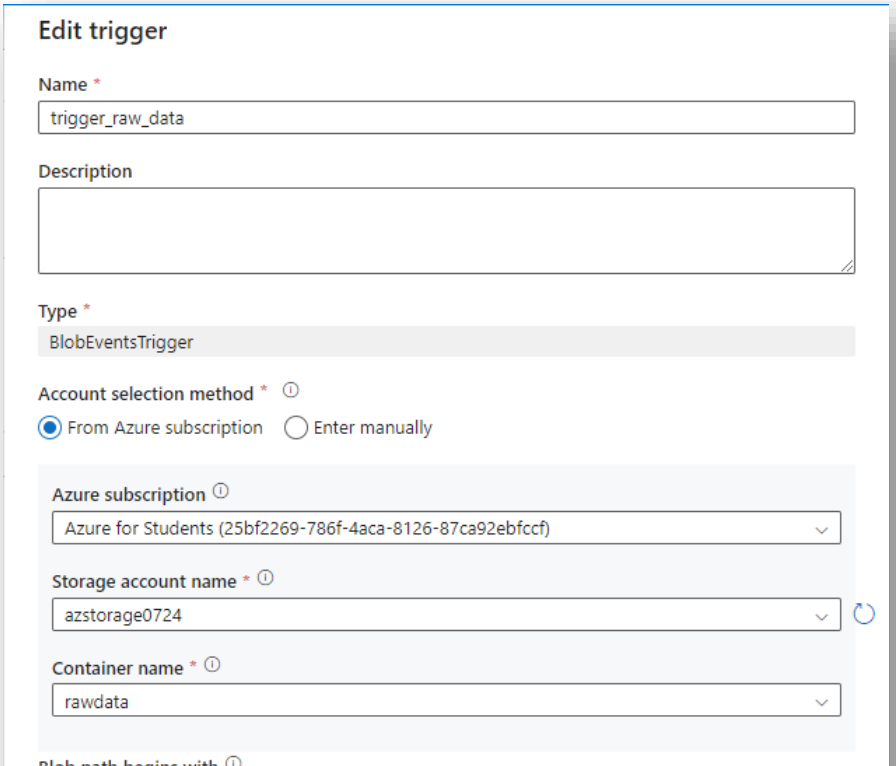
Configuration des Pipelines d'Ingestion :

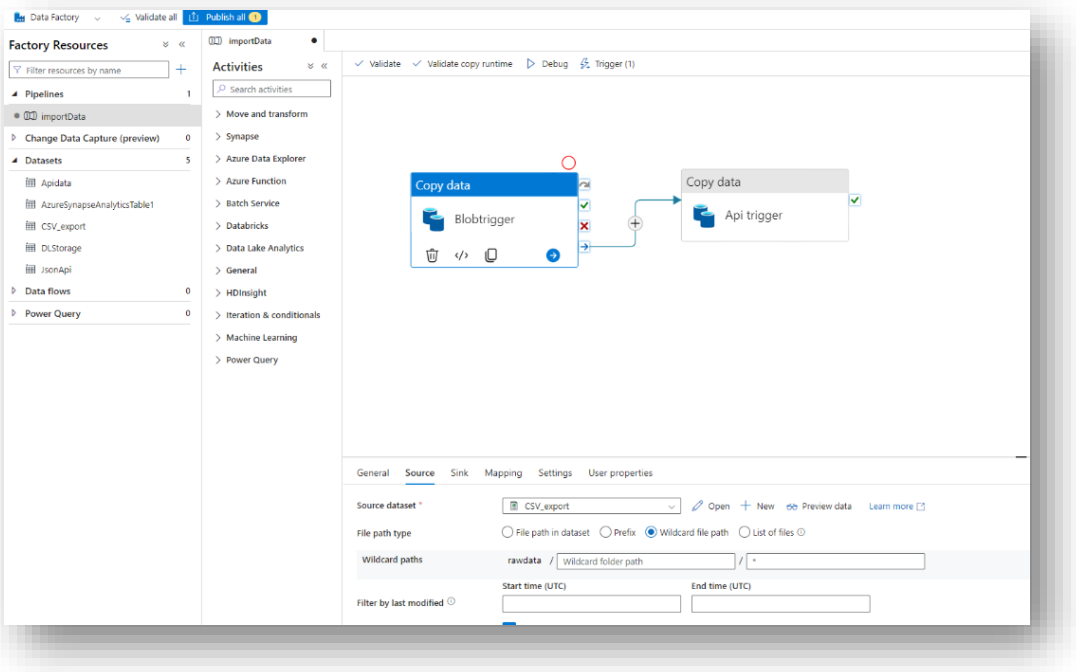
Nous avons configuré plusieurs pipelines d'ingestion dans ADF pour automatiser le processus de collecte des données. Chaque pipeline est responsable de l'ingestion de données à partir d'une source spécifique, comme Azure Blob Storage ou l'API Enedis.



Automatisation et Planification :

ADF permet de planifier l'exécution des pipelines à intervalles réguliers, garantissant que les données sont ingérées de manière continue et en temps opportun. Nous avons configuré des déclencheurs basés sur le dépôt de fichier CSV dans le conteneur.





Stockage des Données Intermédiaires

Le stockage des données intermédiaires est une étape cruciale dans notre pipeline de traitement des données, permettant de gérer les données en transit entre les différentes phases de traitement. Pour ce projet, nous avons utilisé Azure Data Lake Storage pour stocker ces données intermédiaires, en les formatant en Parquet pour des raisons de performance et de compacité.

Utilisation d'Azure Data Lake Storage

Azure Data Lake Storage est une solution de stockage évolutive et sécurisée, idéale pour gérer de grands volumes de données. Nous avons organisé notre Data Lake en différentes couches pour les données brutes, nettoyées et transformées. Après l'ingestion initiale via Azure Data Factory, les données sont stockées au format Parquet, ce qui permet une compression efficace et un accès rapide lors des étapes de transformation et d'analyse.

Avantages du Stockage Intermédiaire

1. **Fiabilité et Disponibilité** : Azure Data Lake Storage assure que les données sont toujours disponibles pour les étapes suivantes du pipeline, réduisant ainsi les risques de perte de données.

2. **Facilitation des Transformations** : Le stockage intermédiaire permet de découpler les étapes d'ingestion et de transformation, facilitant le traitement asynchrone des données.

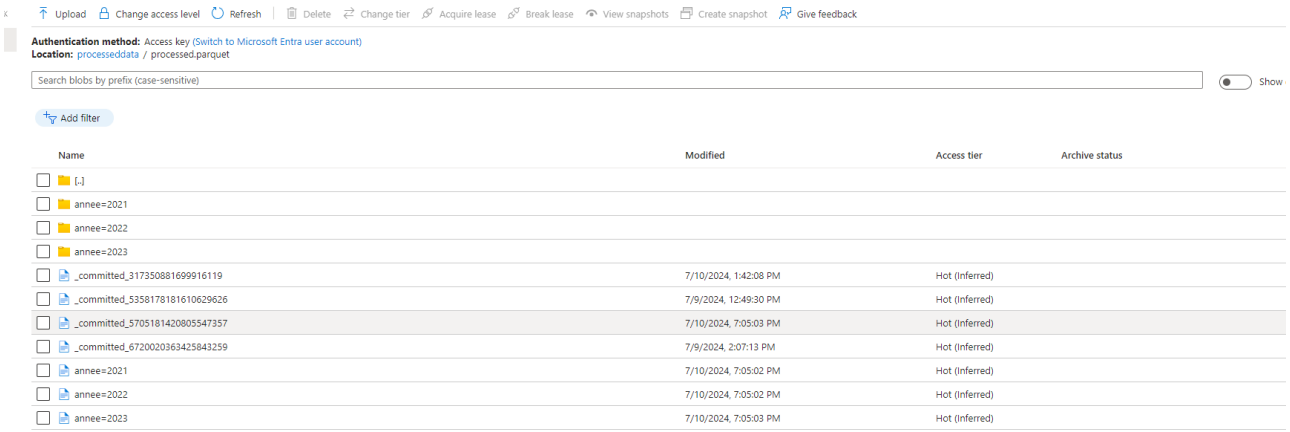
Transformation des Données

Azure Databricks a été utilisé pour transformer les données brutes en informations exploitables pour l'analyse de la consommation énergétique des entreprises. Vous pouvez retrouver le notebook utilisé pour ces transformations sur notre GitHub à ce lien.

https://github.com/mounirm96/Azure-Data-ingestion-

Stockage des Données Transformées

Les données transformées sont stockées de manière sécurisée et efficace dans Azure Data Lake Storage Gen2 pour faciliter l'analyse ultérieure et assurer la disponibilité des données.



Utilisation d'Azure Data Lake Storage Gen2

Azure Data Lake Storage Gen2 offre une capacité de stockage évolutive et des performances élevées pour répondre aux besoins de gestion des données à grande échelle. Voici comment nous avons utilisé ce service pour le stockage des données transformées :

Format Parquet : Les données transformées sont stockées au format Parquet, optimisé pour la compression et la consultation rapide des données. Ce format est idéal pour les analyses complexes et l'extraction d'insights.

Avantages du Stockage des Données Transformées

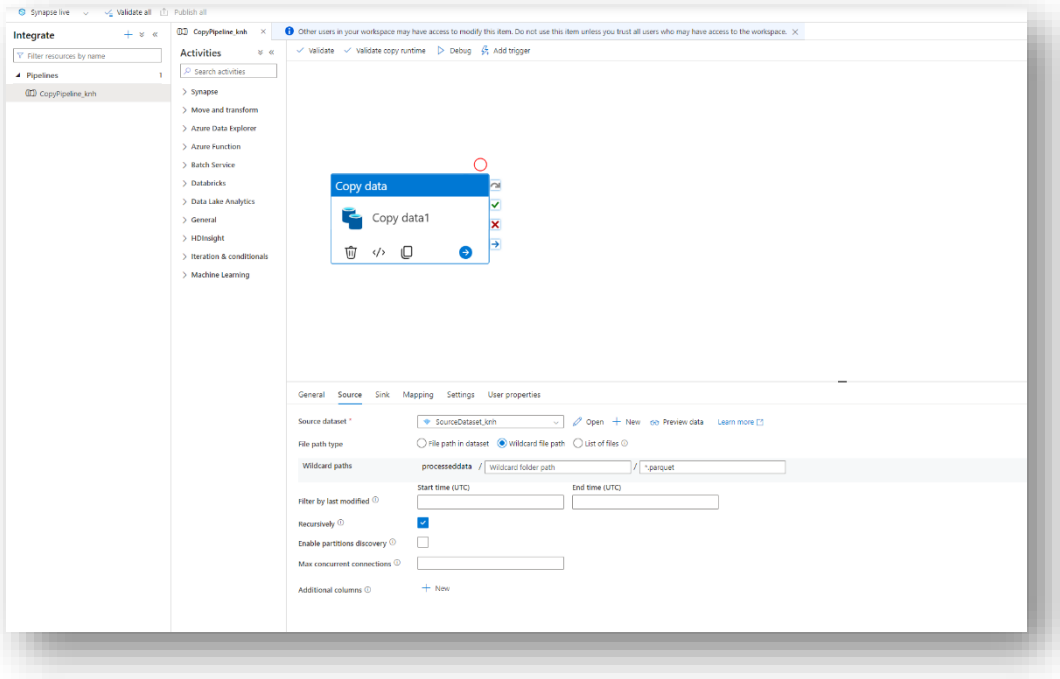
Scalabilité : La capacité de stockage évolutive permet de gérer efficacement les volumes croissants de données transformées.

Performance : Le format Parquet optimise les performances des requêtes, ce qui est crucial pour les analyses complexes et les visualisations interactives.

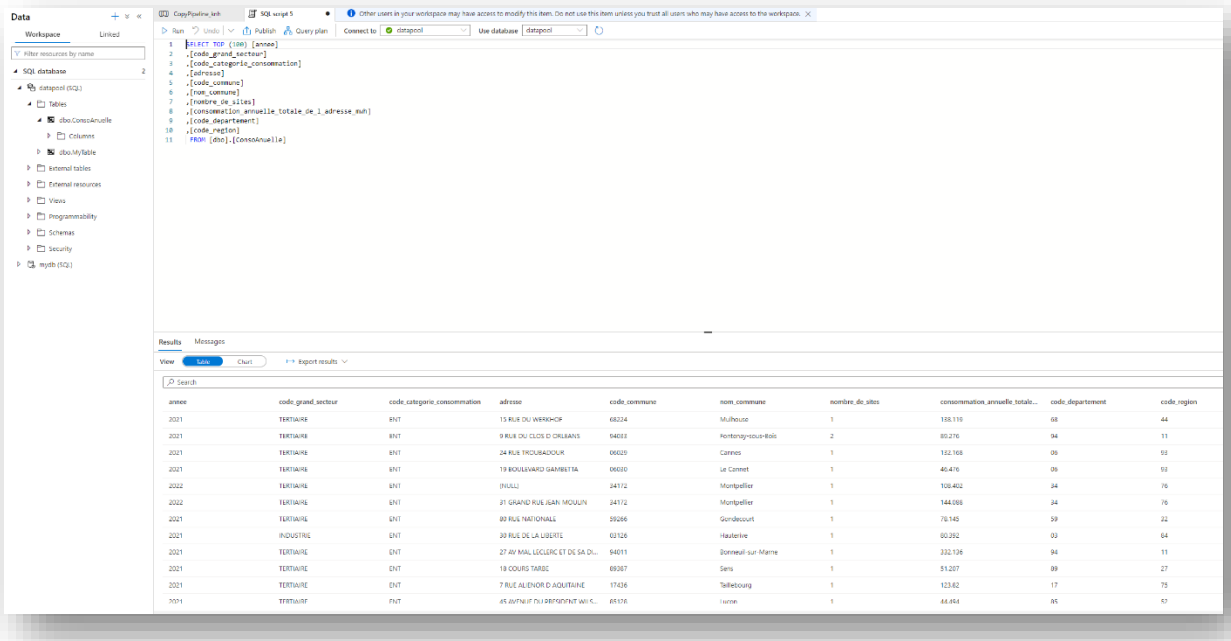
Utilisation d'Azure Synapse Analytics

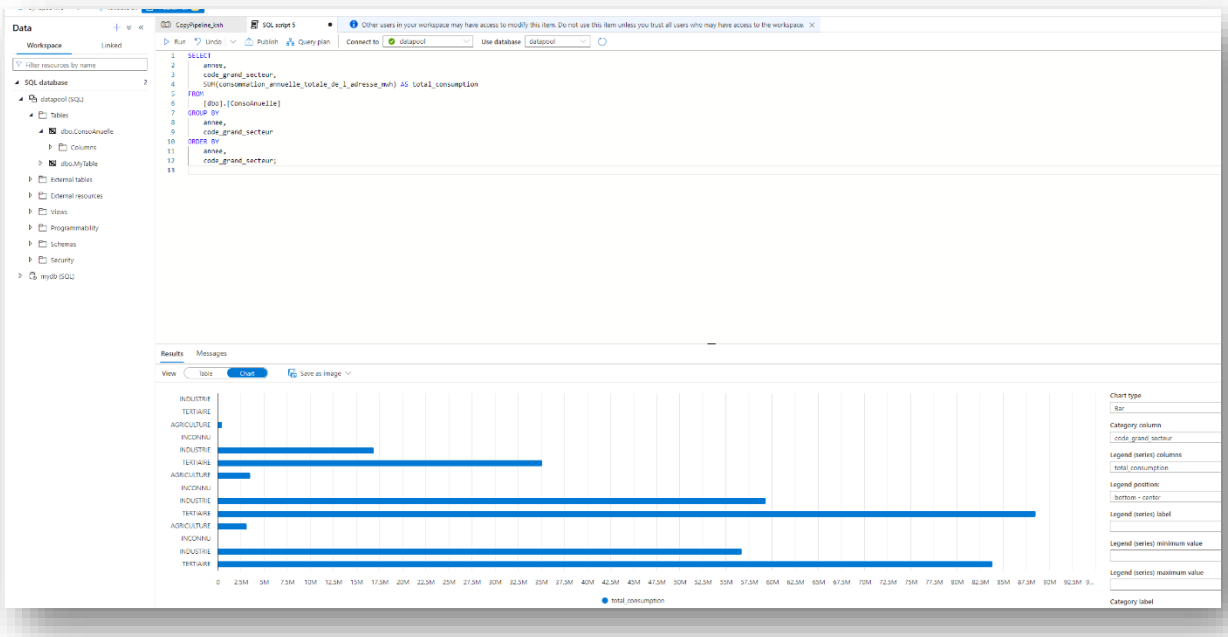
Azure Synapse Analytics, anciennement connu sous le nom de Azure SQL Data Warehouse, combine des fonctionnalités de data warehousing traditionnelles avec des capacités de big data. Voici comment nous l'avons utilisé dans notre projet :

Chargement des Données Transformées : Les données transformées, stockées dans Azure Data Lake Storage Gen2 au format Parquet, sont chargées dans Azure Synapse Analytics. Ce processus est optimisé pour la vitesse et la scalabilité, assurant que les données sont disponibles rapidement pour l'analyse.



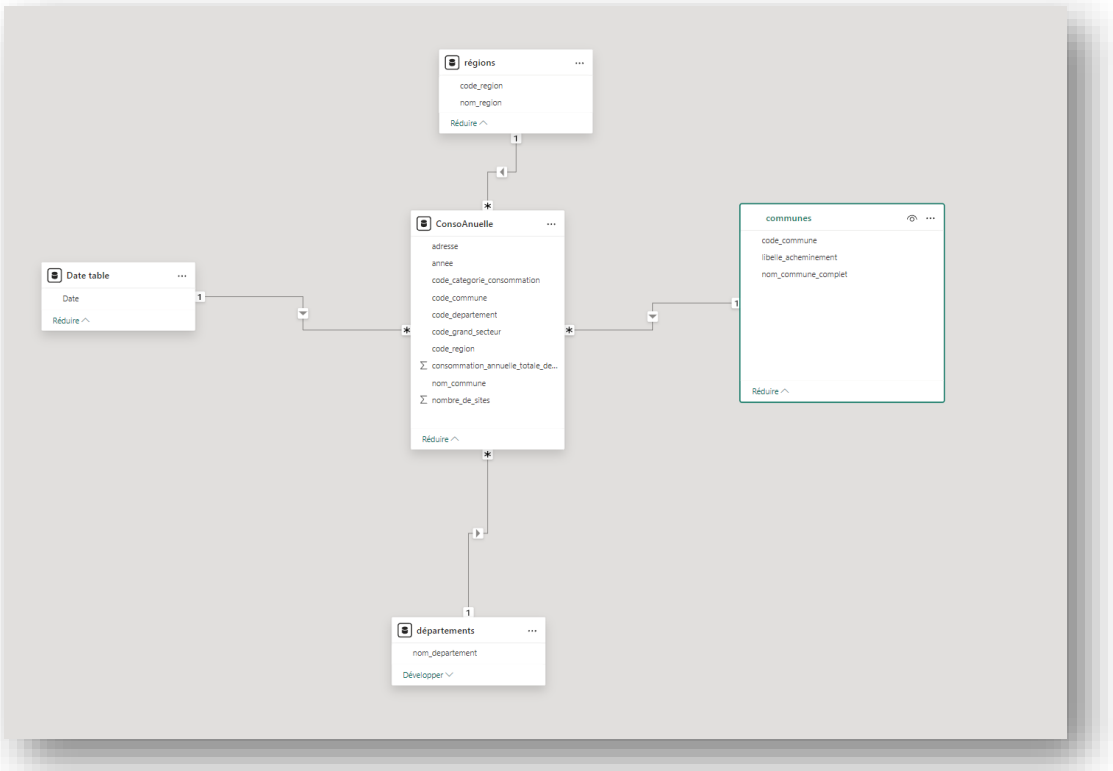
Exécution de Requêtes SQL : Azure Synapse Analytics permet l'exécution de requêtes SQL complexes sur de grands ensembles de données. Nous avons utilisé ces capacités pour agréger, filtrer et analyser les données de consommation énergétique des entreprises.





Visualisation des Données

Power BI a été utilisé pour la visualisation interactive des insights tirés de l'analyse de la consommation énergétique des entreprises, offrant des tableaux de bord dynamiques et informatifs.



Utilisation de Power BI

Power BI est une plateforme de business intelligence qui permet de créer des visualisations interactives et des tableaux de bord à partir de données. Voici comment nous l'avons utilisé dans notre projet :

1. **Connexion aux Données** : Power BI se connecte directement à Azure Synapse Analytics pour accéder aux données analysées et transformées. Cela permet une intégration transparente des insights dans les visualisations.
2. **Création de Visualisations** : Nous avons créé plusieurs types de visualisations, telles que des graphiques, des tableaux croisés dynamiques et des cartes géographiques, pour représenter différents aspects de la consommation énergétique des entreprises.
3. **Tableaux de Bord Interactifs** : Les tableaux de bord interactifs de Power BI permettent aux utilisateurs d'explorer les données, de filtrer les informations selon différents critères et d'obtenir des insights en temps réel.