Table des matières

Projet : Pipeline de Données sur Azure

Ce projet est une solution de pipeline de données conçue pour répondre à un problème métier fictif. Il a été créé dans le but de renforcer ma compréhension et mon apprentissage des pipelines de données.

Demande métier

Dans ce projet, notre entreprise a identifié un manque de compréhension des données démographiques de ses clients, notamment la répartition par genre et son influence sur les achats de produits. Une grande quantité de données clients étant stockée dans une base de données SQL sur site, les parties prenantes ont demandé la création d'un tableau de bord KPI complet. Ce tableau de bord devra fournir des insights sur les ventes par genre et par catégorie de produit, en affichant le nombre total de produits vendus, le chiffre d'affaires total et une répartition claire des clients selon leur genre.

De plus, il devra inclure des filtres permettant de segmenter les données par catégorie de produit et par genre, ainsi qu'une interface conviviale pour les requêtes basées sur des dates.

Vue d'ensemble de notre solution

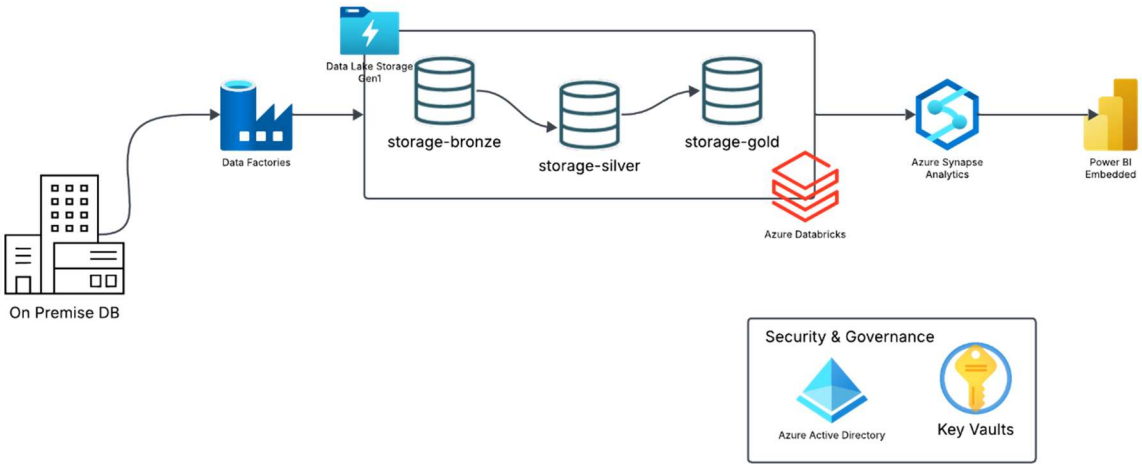
Pour répondre à cette demande, nous allons concevoir une pipeline de données robuste qui extraira les données de la base sur site, les chargera dans Azure et effectuera les transformations nécessaires afin de les rendre plus exploitables. Les données transformées alimenteront ensuite un rapport personnalisé répondant à toutes les exigences spécifiées. Cette pipeline sera planifiée pour s'exécuter automatiquement chaque jour, garantissant ainsi aux parties prenantes un accès permanent à des données précises et actualisées.

Exigences métier

L'entreprise a identifié un manque de compréhension des données démographiques des clients, notamment la répartition des genres et son influence sur les achats. Les principales exigences sont les suivantes :

- **Ventes par genre et catégorie de produit** : Un tableau de bord affichant le nombre total de produits vendus, le chiffre d’affaires total et une répartition des clients selon leur genre.
- **Filtrage des données** : Possibilité de filtrer les données par catégorie de produit, genre et date.
- **Interface conviviale** : Un accès simplifié pour que les parties prenantes puissent interroger facilement les données.

[1]Architecture de la solution



Explication de notre architecture

Pour répondre à ces exigences, la solution est structurée en plusieurs étapes :

1. Ingestion des données

- Extraction des données clients et ventes depuis une base SQL sur site.
- Chargement des données dans **Azure Data Lake Storage (ADLS)** via **Azure Data Factory (ADF)**.

2. Transformation des données

- Nettoyage et transformation des données avec **Azure Databricks**.
- Organisation des données en trois niveaux :
    - **Bronze** : Données brutes
    - **Silver** : Données nettoyées
    - **Gold** : Données agrégées et prêtes pour l'analyse

3. Chargement et Reporting

- Chargement des données transformées dans **Azure Synapse Analytics**.
- Création d'un **tableau de bord Power BI** pour visualiser les insights et permettre une exploration interactive des données.

4. Automatisation

- Planification de l'exécution du pipeline quotidiennement pour garantir des données toujours à jour.

---

[1] By Franklin kana nguedia

Stack Technologique

- **Azure Data Factory (ADF)** : Orchestration du mouvement et de la transformation des données.
- **Azure Data Lake Storage (ADLS)** : Stockage des données brutes et transformées.
- **Azure Databricks** : Traitement et transformation des données.
- **Azure Synapse Analytics** : Entrepôt de données et requêtes SQL analytiques.
- **Power BI** : Visualisation des données et création de rapports.
- **Azure Key Vault** : Gestion sécurisée des identifiants et secrets.
- **SQL Server (On-Premises)** : Source des données clients et ventes.

Instructions d'installation et de configuration

Principaux Etapes dans la construction du pipeline :

Pré-requis

- Un compte Azure avec des crédits suffisants.
- Accès à une base de données **SQL Server sur site**.

Étape 1 : Configuration de l'environnement Azure

1. **Créer un groupe de ressources** sur Azure.
2. **Provisionner les services nécessaires** :
    - Créer une instance **Azure Data Factory**.
    - Configurer **Azure Data Lake Storage** avec les conteneurs Bronze, Silver et Gold.
    - Déployer un **workspace Azure Databricks** et un **workspace Synapse Analytics**.
    - Configurer **Azure Key Vault** pour la gestion des identifiants.

Étape 2 : Ingestion des données

1. Installer **SQL Server** et **SQL Server Management Studio (SSMS)** sur site.
2. Restaurer la base de données **AdventureWorks**.
3. Créer des pipelines dans **ADF** pour copier les données de **SQL Server** vers le conteneur Bronze d'ADLS.

Étape 3 : Transformation des données

1. Configurer Databricks pour accéder à ADLS.

2. Nettoyer et agréger les données dans des notebooks Databricks, en les faisant passer de Bronze à Silver puis à Gold.

Étape 4 : Chargement et Reporting

1. Créer un **pool SQL Synapse** et charger les données Gold pour analyse.
2. Connecter **Power BI** à Synapse et concevoir un tableau de bord basé sur les exigences métier.

Étape 5 : Automatisation et Monitoring

1. Planifier l'exécution quotidienne des pipelines dans **ADF**.
2. Surveiller les exécutions avec les outils de monitoring intégrés d'ADF et Synapse.

Étape 6 : Sécurité et Gouvernance

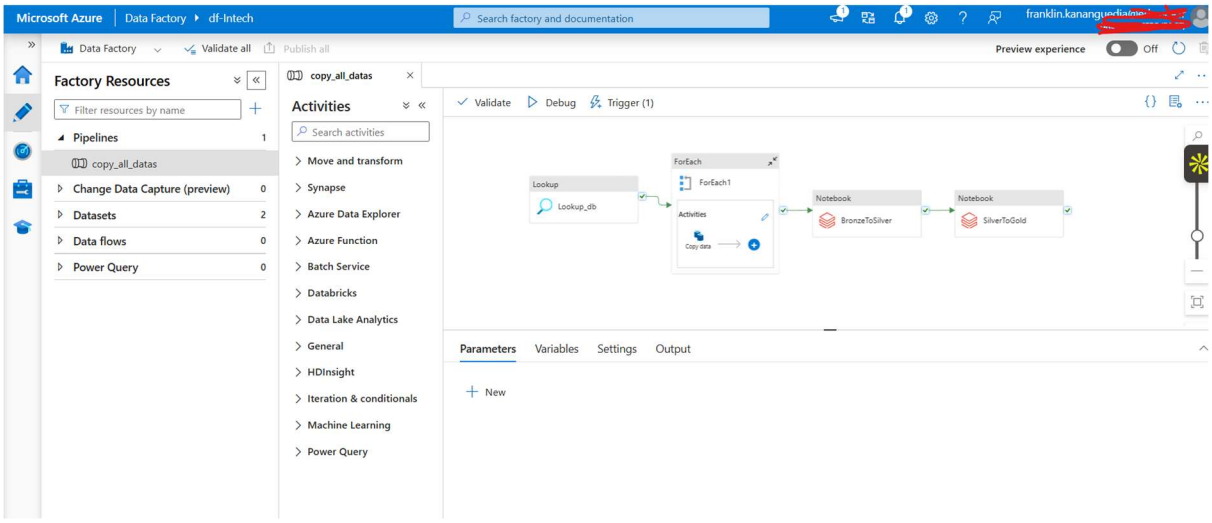
- Mettre en place le **contrôle d'accès basé sur les rôles (RBAC)** avec **Azure Entra ID (ex-Active Directory)**.

Étape 7 : Tests de bout en bout

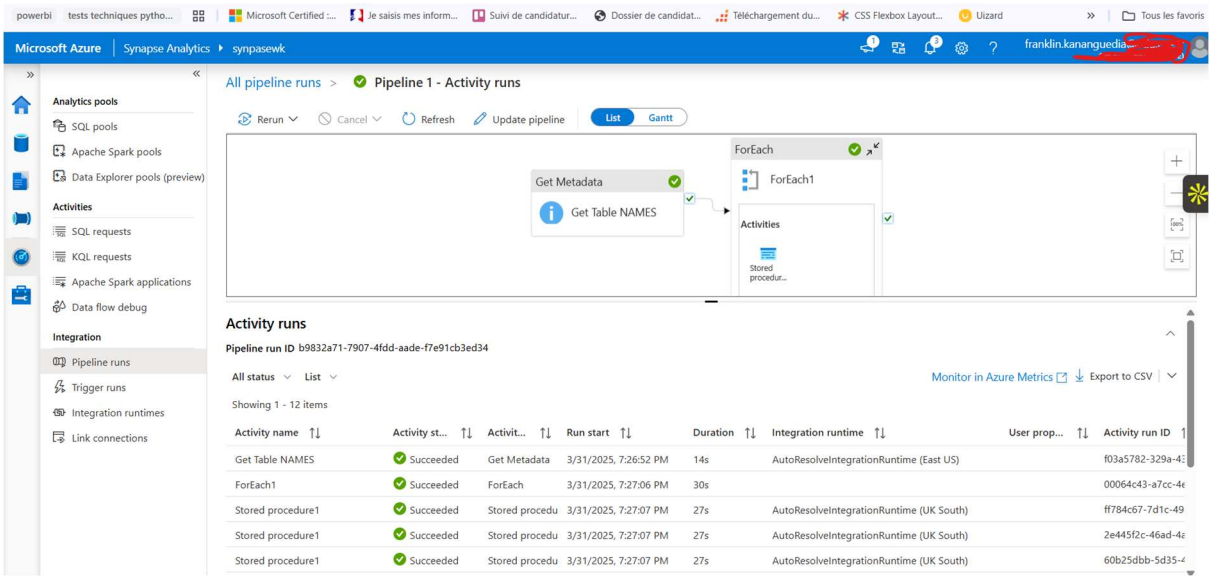
1. Ajouter de nouveaux enregistrements dans la base SQL.
2. Vérifier que tout le pipeline s'exécute correctement et que le tableau de bord Power BI est mis à jour.

Résultats :

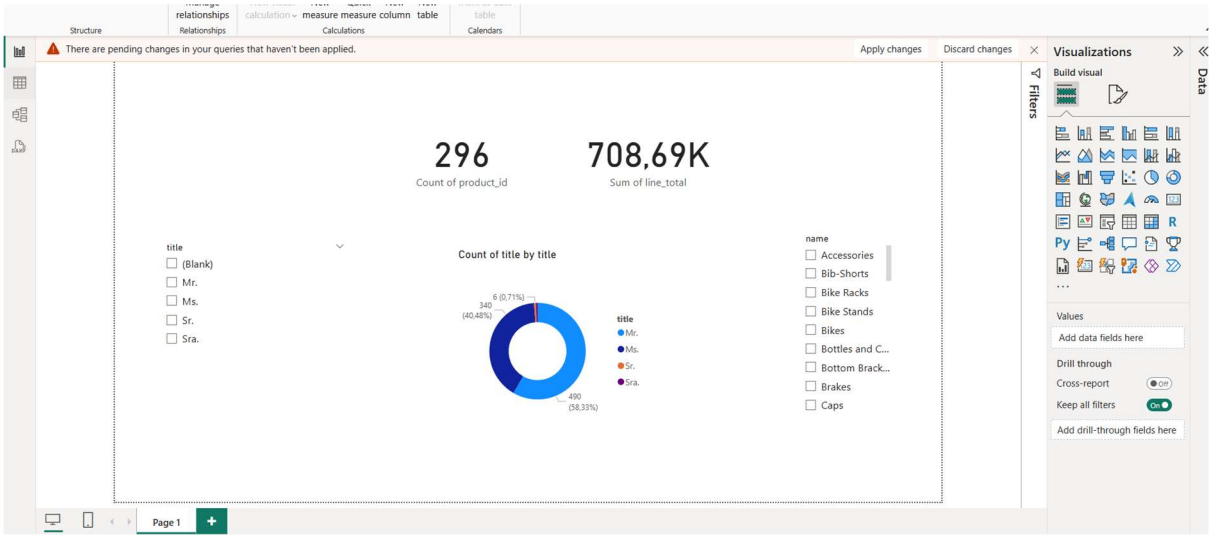
Visuel de notre pipeline azure DataFactory



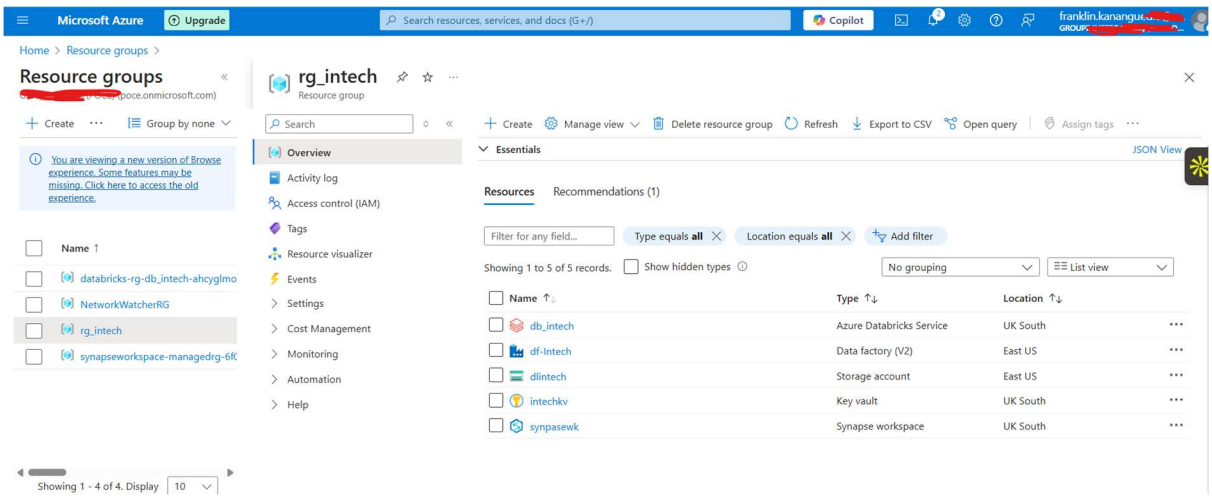
Visuel de notre pipeline Azure Synpase Analytics



Visuel de Power BI (visualisation)



Présentation de notre ressource Group



Conclusion

Ce projet fournit une solution de bout en bout permettant de mieux comprendre la démographie des clients et son impact sur les ventes. Grâce à l'automatisation du pipeline de données, les parties prenantes ont un accès en temps réel aux informations les plus récentes et exploitables.