المدرسة الوطنية للعلوم التطبيقية - بني ملال
ⵜⵉⵎⵔⵙⵉⵜ ⵜⵏⴰⵎⵓⵔⵜ ⵏ ⵜⵙⵙⵏⵜⵉⵡⵉⵏ ⵜⵜⵓⵏⵉⵖⵜ - ⴰⵣⵜ ⵅⵏⵎⴰ
Ecole Nationale des Sciences Appliquées - Béni Mellal

Pilotage de la performance et tableaux

Rapport de Mini-Projet : Mise en Œuvre d'un Pipeline de Données avec Azure : Exploration et Analyse de AdventureWorks

Réalisé Par :
HOUBAOUI Mimoune
SARHIR MOHAMED

Encadré Par :
Pr. BE.ELBAGHAZAOUI

Le 24/12/2024

Sommaire

Introduction

Dans le cadre de l'exploration des technologies modernes de gestion et d'analyse de données, ce projet s'inscrit dans une démarche pratique visant à exploiter et transformer les données d'AdventureWorks, une base de données fictive développée par Microsoft. AdventureWorks simule les opérations d'une entreprise multinationale spécialisée dans la fabrication et la vente de vélos, offrant ainsi un cadre réaliste pour expérimenter des scénarios d'ingénierie des données.

L'objectif principal de ce projet est de mettre en place un pipeline de données complet en utilisant les outils de Microsoft Azure et d'autres technologies pertinentes. À travers des étapes allant de l'extraction des données brutes à leur transformation, jusqu'à leur visualisation via Power BI, ce projet illustre les concepts clés de l'ingénierie des données. Les participants doivent ainsi apprendre à nettoyer, organiser, et exploiter les données pour en tirer des insights pertinents pour la prise de décision.
Ce travail intègre des solutions modernes telles qu'Azure Data Factory, Azure Data Lake, et Azure Synapse Analytics pour construire un pipeline de données efficace et adaptable. En parallèle, les méthodologies de gestion de tâches agiles et d'intégration collaborative, via des outils comme Taiga.io et GitHub, permettent d'assurer une coordination optimale des équipes et une qualité élevée des livrables.

Ce rapport détaille les différentes étapes du projet, les outils utilisés, et les résultats obtenus, tout en mettant en lumière les défis rencontrés et les enseignements tirés de cette expérience.

II - Contexte et Description du Projet

II - 1 - Description des données.

AdventureWorks est une base de données d'exemple fournie par Microsoft pour simuler les opérations d'une entreprise fictive appelée AdventureWorks Cycles. Cette entreprise se spécialise dans la fabrication et la vente de vélos, de pièces et d'accessoires dans un contexte international. La base de données contient plusieurs schémas représentant les différents aspects de l'entreprise, allant des ventes et des ressources humaines à la production et aux achats.
Dans le cadre de ce projet, nous avons principalement travaillé avec le schéma Sales d'AdventureWorks 2019. Ce schéma comprend plusieurs tables essentielles qui permettent de suivre les transactions commerciales de l'entreprise, ainsi que l'historique des commandes, des produits, des clients et des territoires de vente. Les tables utilisées dans le schéma Sales sont cruciales pour analyser la performance des ventes et la distribution des produits à travers différents territoires géographiques.

II - 2 - Besoins et problématiques à résoudre

L'objectif de ce projet est de préparer, nettoyer et analyser les données des ventes d'AdventureWorks. Les principales problématiques à résoudre sont :

- **Nettoyage des données** : Avant toute analyse, il est nécessaire d'identifier et de traiter les anomalies dans les données, telles que les valeurs manquantes, les doublons ou les formats incorrects.
- **Intégration et transformation des données** : L'objectif est de concevoir un pipeline qui permette d'intégrer les données des différentes tables du schéma Sales, de les transformer et de les charger dans un format analysable.
- **Visualisation des données** : Une fois les données nettoyées et transformées, il est essentiel de créer des visualisations dans Power BI pour faciliter la prise de décision et obtenir des insights sur les performances des ventes et des territoires.

II - 3 - Présentation des outils utilisés

Pour accomplir ces objectifs, plusieurs outils et technologies ont été utilisés tout au long du projet :

- **AdventureWorks 2019 (SQL Server)** : Cette base de données d'exemple a été installée dans SQL Server Management Studio (SSMS) pour accéder aux tables et explorer les données. Nous avons utilisé les tables du schéma Sales pour réaliser les différentes étapes du projet.

- **Azure Data Factory** : Azure Data Factory a été utilisé pour orchestrer le pipeline de données, facilitant l'extraction des données, leur transformation et leur chargement vers des destinations appropriées comme Azure Data Lake ou un entrepôt de données.

- **Azure Data Lake** : Les données brutes ont été stockées dans Azure Data Lake pour permettre une gestion centralisée et évolutive des données avant transformation.

- **Azure Synapse Analytics** : Azure Synapse Analytics a été utilisé pour effectuer des transformations complexes sur les données, notamment l'agrégation, l'enrichissement et l'analyse à grande échelle. Il permet de préparer les données pour une analyse approfondie à l'aide de requêtes SQL et d'intégrer des workflows analytiques.

- **Azure Databricks** : Power BI a été utilisé pour créer des visualisations interactives permettant de suivre la performance des ventes par produit, par territoire, et par période.

- **Power BI** : Power BI a été utilisé pour créer des visualisations interactives permettant de suivre la performance des ventes par produit, par territoire, et par période.

- **Azure Key Vault** : Power BI a été utilisé pour créer des visualisations interactives permettant de suivre la performance des ventes par produit, par territoire, et par période.

- Azure Entra ID (Active Directory) : gère les identités et les autorisations d'accès pour garantir une sécurité optimale.

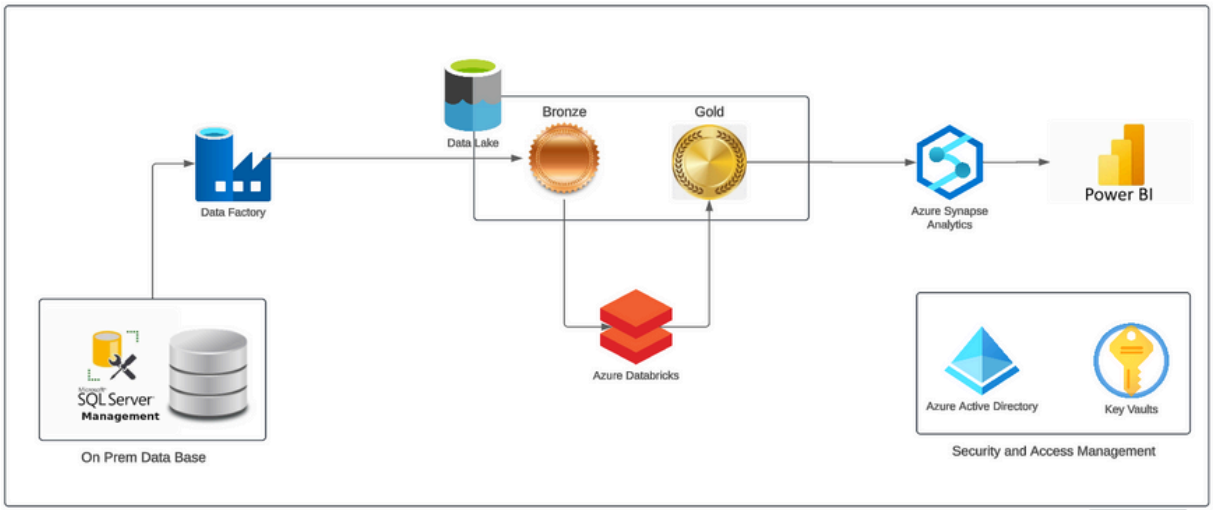
- GitHub : GitHub a servi à gérer les versions des scripts SQL et des rapports créés, ainsi qu'à faciliter la collaboration avec les autres membres de l'équipe.

- Taiga.io : Taiga.io a été utilisé pour la gestion de projet, permettant de suivre l'avancement des tâches à travers des sprints et de s'assurer de la bonne organisation du travail au sein de l'équipe.

III - Architecture et Pipeline de Données

III - 1 - Diagramme d'architecture globale

L'architecture globale de ce projet repose sur une série de technologies Azure qui interagissent pour orchestrer, transformer et visualiser les données dans un flux continu. Le pipeline de données commence par l'extraction des données de sources multiples, puis les données sont stockées, transformées et enfin visualisées à l'aide de Power BI.

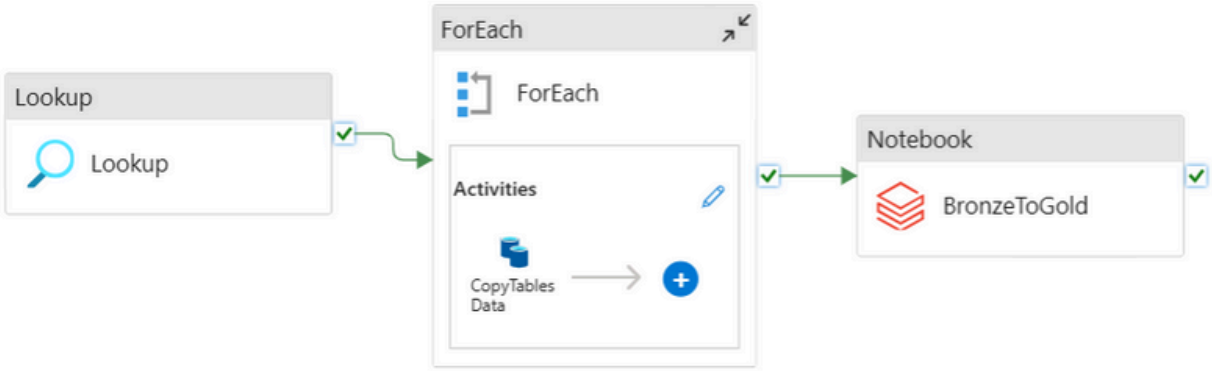


IV - Étapes de Réalisation

Le pipeline de données a été conçu pour orchestrer l'ensemble du processus, depuis l'extraction des données jusqu'à leur visualisation dans Power BI. Voici les étapes principales du pipeline :

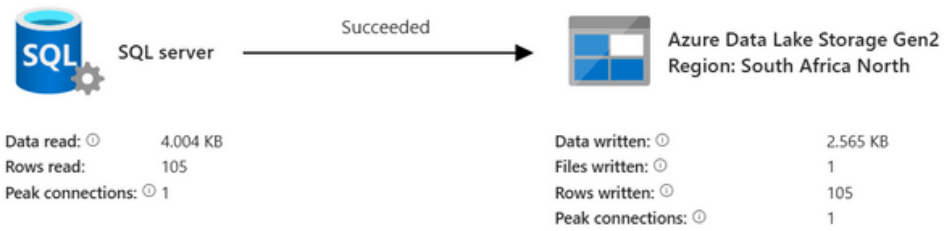
IV - 1 - Extraction des données - ETL

La première étape consiste à extraire les données de différentes sources (par exemple, bases de données SQL, fichiers CSV, API). Azure Data Factory a été utilisé pour orchestrer cette extraction et charger les données brutes dans Azure Data Lake.



IV - 2 - Le Stockage dans Azure Data Lake

Une fois extraites, les données sont stockées dans Azure Data Lake sous leur forme brute. Ce stockage centralisé permet de gérer de grandes quantités de données non structurées ou semi-structurées, en les rendant disponibles pour des transformations ultérieures.



IV - 3 - Transformation des données

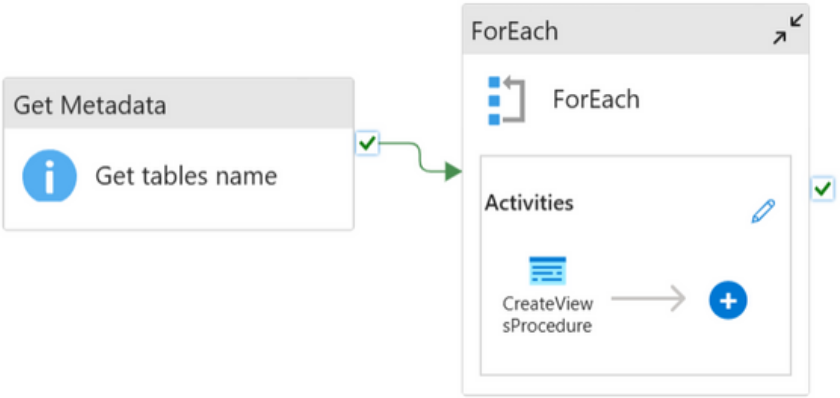
Les données sont extraites des fichiers Parquet stockés dans le répertoire bronze (données brutes) et chargées dans des DataFrames Spark. Les doublons sont supprimés et la colonne ModifiedDate est renommée en Date.

Transformations spécifiques sur chaque table :

- **Customer** : La colonne StoreID est supprimée. Ensuite, les lignes avec des valeurs nulles dans la colonne PersonID sont éliminées et cette colonne est convertie en entier.
- **SalesOrderDetail** : Les valeurs nulles de la colonne CarrierTrackingNumber sont remplacées par la valeur 'UNKNOWN'.
- **SalesOrderHeader** : Plusieurs colonnes inutiles telles que PurchaseOrderNumber, SalesPersonID, Comment et CurrencyRateID sont supprimées. Ensuite, les lignes avec des valeurs nulles dans la colonne CreditCardID sont supprimées et cette colonne est convertie en entier.
- **SalesPerson** : Les lignes avec des valeurs nulles dans la colonne TerritoryID sont supprimées, et les colonnes TerritoryID et SalesQuota sont converties respectivement en entier et en double.
- **SalesTerritoryHistory** : La colonne EndDate est supprimée.
- SpecialOffer : La colonne MaxQty est supprimée.

IV - 4 - Chargement dans l'entrepôt de données

Après la transformation, les données sont chargées dans un entrepôt de données ou une autre destination appropriée, comme Azure Synapse Analytics.



IV - 5 - Visualisation des données

Les données transformées et nettoyées dans le stockage Gold ont été utilisées pour créer des visualisations interactives avec Power BI, offrant des insights sur les performances des ventes, les tendances et les KPI.

Connexion aux données Gold et gestion des relations :
Power BI a été configuré pour se connecter au stockage Gold. Les relations entre les tables (Customer, SalesOrderHeader, SalesPerson, etc.) ont été définies pour garantir un modèle de données cohérent et faciliter les analyses.

Construction des tableaux de bord :
Les tableaux de bord interactifs ont été conçus pour permettre une exploration intuitive des données transformées.

IV - 6 - Sécurisation et gestion des accès

Tout au long du pipeline, la gestion des identités et des accès est assurée par Azure Entra ID (Active Directory), qui garantit que seules les personnes autorisées peuvent accéder aux données et aux outils de transformation. Azure Key Vault est utilisé pour gérer de manière sécurisée les secrets et les clés nécessaires à l'accès aux différentes ressources Azure.

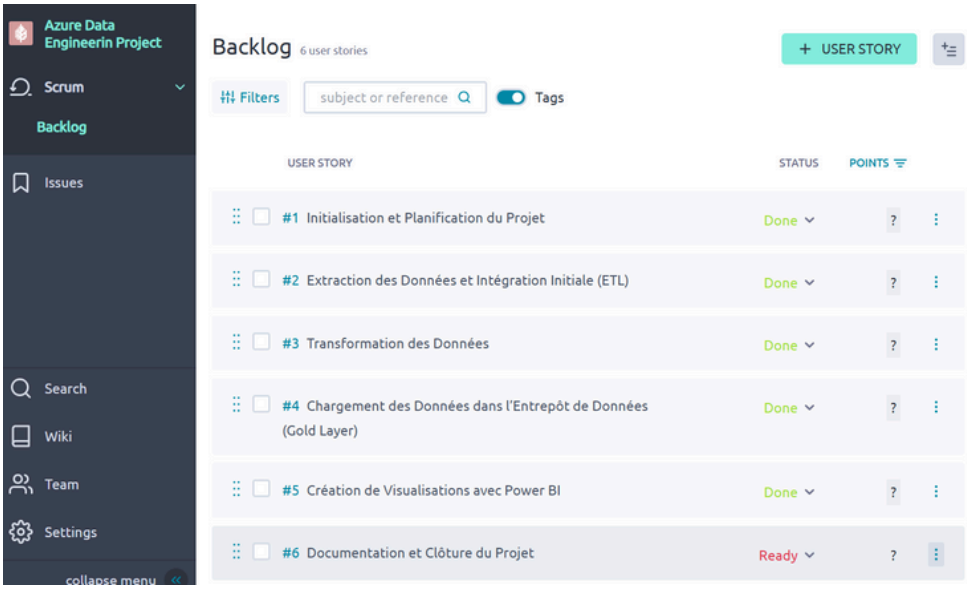
V - Gestion des Tâches et Collaboration

V - 1 - Utilisation de Taiga.io pour la gestion des tâches et des sprints

testPour assurer une organisation optimale du projet, Taiga.io a été utilisé comme outil de gestion des tâches. Les principales activités réalisées sur cette plateforme incluent :

- Planification des sprints pour structurer les étapes clés du projet.
- Attribution des tâches en fonction des rôles et compétences des membres.
- Suivi de l'avancement des tâches via un tableau Kanban, permettant une gestion visuelle et intuitive du progrès.

Cette gestion structurée a permis de respecter les délais et de répartir les responsabilités de manière efficace.



V - 2 - Méthodologie Agile adoptée

La méthodologie Agile a été adoptée pour une exécution flexible et collaborative du projet. Ses points forts incluent :

- Réunions régulières pour faire le point sur l'avancement et ajuster les priorités.
- Une approche itérative permettant de livrer des versions intermédiaires et de recueillir des retours en continu.
- Une capacité d'adaptation rapide face aux changements ou aux imprévus.

Grâce à Agile, l'équipe a pu maintenir une communication fluide et livrer un produit final de qualité.

V - 3 - Collaboration via GitHub

GitHub a joué un rôle central dans la collaboration technique, en assurant une gestion efficace des fichiers et du code. Voici les principaux usages :

- Gestion des versions pour suivre les modifications et revenir à des états antérieurs en cas de besoin.
- Création de branches pour développer et tester des fonctionnalités sans affecter la version principale.
- Intégration des modifications via des pull requests avec validation collective pour garantir la qualité du code.

GitHub a également servi à centraliser les livrables, incluant les scripts SQL, les notebooks Databricks, et les rapports Power BI, favorisant ainsi un environnement collaboratif efficace.

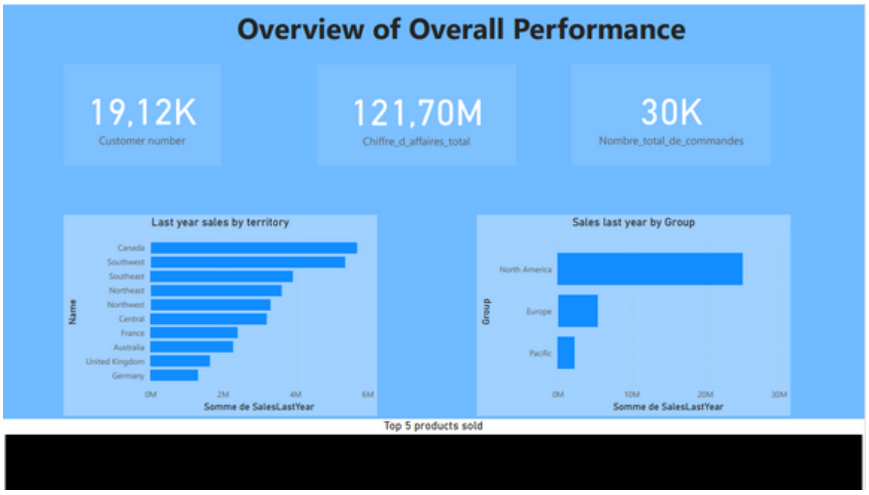
VI - Résultats

VI - 1 - Présentation des visualisations principales

Page 1 : Analyse de la performance globale

Objectif : Cette visualisation fournit un aperçu global de la performance de l'entreprise, en se concentrant sur les indicateurs clés de performance (KPI) tels que les ventes totales, la rentabilité et l'évolution des performances sur plusieurs périodes. L'objectif est de donner une vue d'ensemble de la santé financière et opérationnelle d'AdventureWorks.

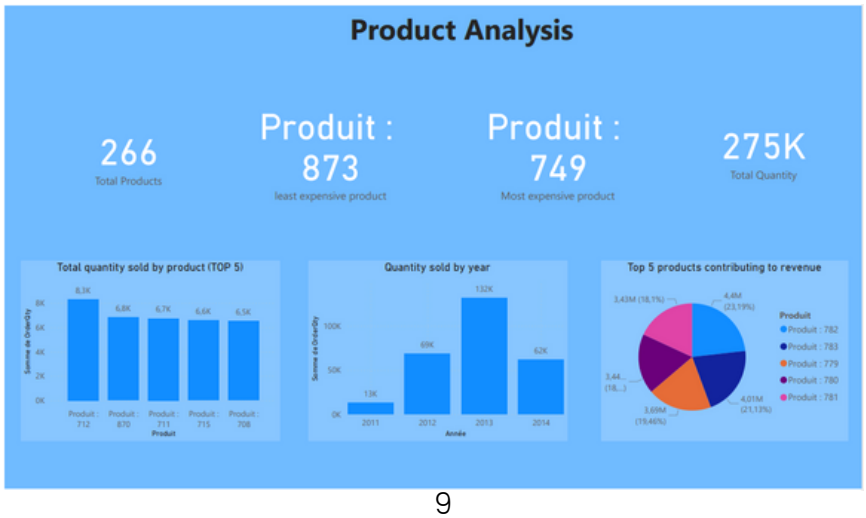
Visualisations utilisées :



Page 2 : Analyse des produits

Objectif : Cette page se concentre sur l'analyse des produits vendus par AdventureWorks. Elle permet d'identifier les produits les plus rentables, les meilleures performances de vente par catégorie, ainsi que les tendances de vente sur des périodes données.

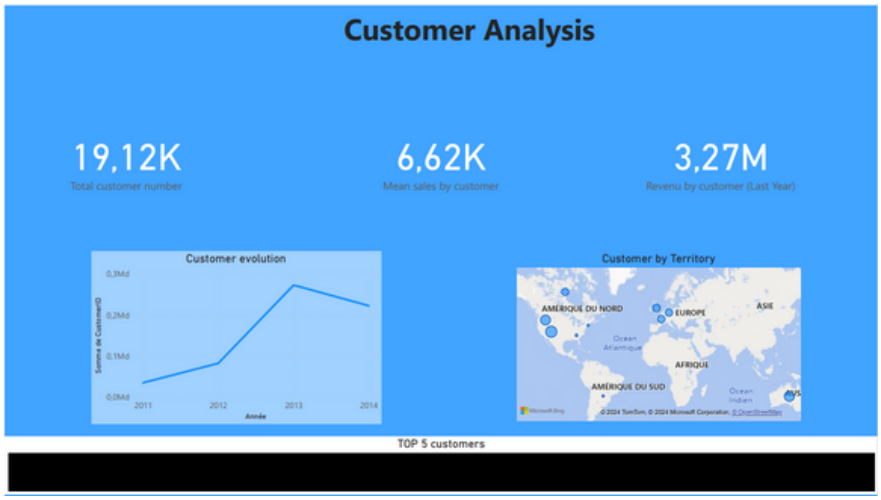
Visualisations utilisées :



Page 3 : Analyse des clients

Objectif : Cette page analyse les données des clients, avec un focus sur leur comportement d'achat et leur valeur à long terme pour l'entreprise. L'objectif est d'identifier les clients les plus précieux et de comprendre les tendances dans leurs achats.

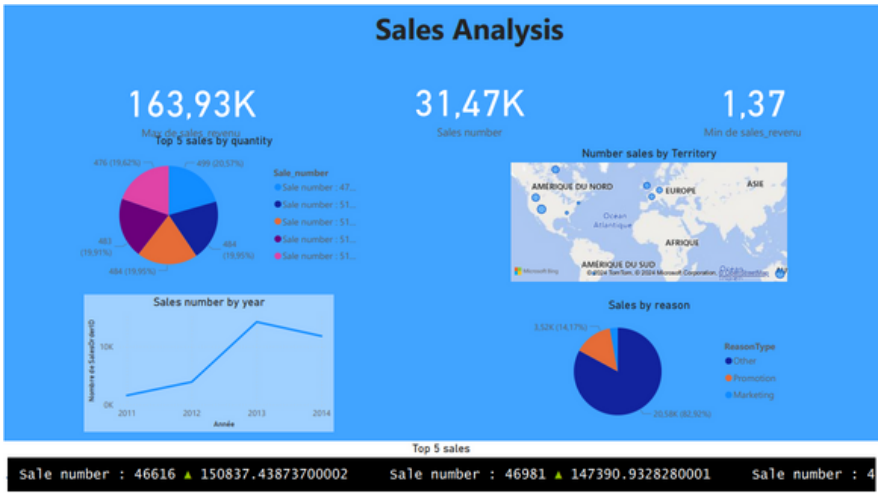
Visualisations utilisées :



Page 4 : Analyse des ventes

Objectif : L'objectif de cette page est de se concentrer sur l'analyse des ventes en fonction des territoires, des employés de vente et des promotions spéciales. Cette visualisation aide à comprendre où se trouvent les meilleures opportunités de vente et où des améliorations peuvent être apportées.

Visualisations utilisées :



VII - Leçons Apprises et Améliorations

Ce projet a permis d'acquérir une meilleure compréhension des enjeux liés à la qualité des données, à l'orchestration des pipelines et à la collaboration en équipe. La phase de nettoyage et de transformation des données a souligné l'importance de données fiables pour garantir des analyses précises, tandis que l'utilisation d'outils comme Azure Data Factory, Databricks, et Synapse a renforcé notre expertise dans la gestion des données à grande échelle. La méthodologie Agile, associée à des outils comme Taiga.io et GitHub, a facilité une gestion efficace des tâches et permis une meilleure coordination entre les membres de l'équipe.

Pour les projets futurs, plusieurs pistes d'amélioration ont été identifiées, notamment l'intégration d'une plus grande automatisation pour réduire les interventions manuelles, une optimisation des performances des outils utilisés, et un renforcement de la documentation pour faciliter la collaboration et la transmission des connaissances. Enfin, l'ajout d'alertes dans Power BI et l'adoption de pipelines CI/CD pourraient améliorer la surveillance des KPI et sécuriser les déploiements.

Conclusion

Ce projet a permis de démontrer la puissance des solutions cloud et des outils modernes de gestion des données dans le contexte d'une architecture de données bout en bout. En intégrant des technologies telles qu'Azure Data Factory, Data Lake, Synapse, Databricks, et Power BI, nous avons construit un pipeline de données robuste et scalable, allant de l'extraction des données brutes jusqu'à leur analyse interactive.

Les visualisations développées dans Power BI ont offert des insights précieux, permettant d'évaluer les performances des ventes et d'identifier des opportunités d'amélioration. La collaboration efficace, facilitée par des outils tels que GitHub et Taiga.io, ainsi que l'adoption d'une méthodologie Agile, ont permis de garantir un suivi rigoureux des tâches et une livraison progressive des résultats.

Ce projet constitue une étape importante dans la maîtrise des outils et techniques modernes de gestion de données. Les leçons apprises et les pistes d'amélioration identifiées seront précieuses pour des projets futurs, notamment dans l'optimisation des performances, l'automatisation, et l'intégration avancée des visualisations. Ce travail témoigne de l'importance croissante des solutions cloud pour répondre aux défis des entreprises en matière de transformation digitale et de gestion des données.

Bibliographie

- Microsoft Azure Documentation : https://learn.microsoft.com/en-us/azure
- Azure Data Factory Documentation : https://learn.microsoft.com/en-us/azure/data-factory/
- Azure Synapse Analytics Documentation : https://learn.microsoft.com/en-us/azure/synapse-analytics/
- Azure Data Lake Documentation : https://learn.microsoft.com/en-us/azure/data-lake-store/
- Power BI Documentation : https://learn.microsoft.com/en-us/power-bi
- GitHub Documentation : https://docs.github.com/
- Taiga.io Documentation : https://taiga.io/support/
- Blog Databricks : "Getting Started with PySpark for Data Transformation," https://databricks.com/blog
- Cours "Base de données et Cloud Computing," ENSA-Beni Mellal.
- Stack Overflow : https://stackoverflow.com/
- Azure Community Blog : https://techcommunity.microsoft.com/t5/azure/ct-p/Azure