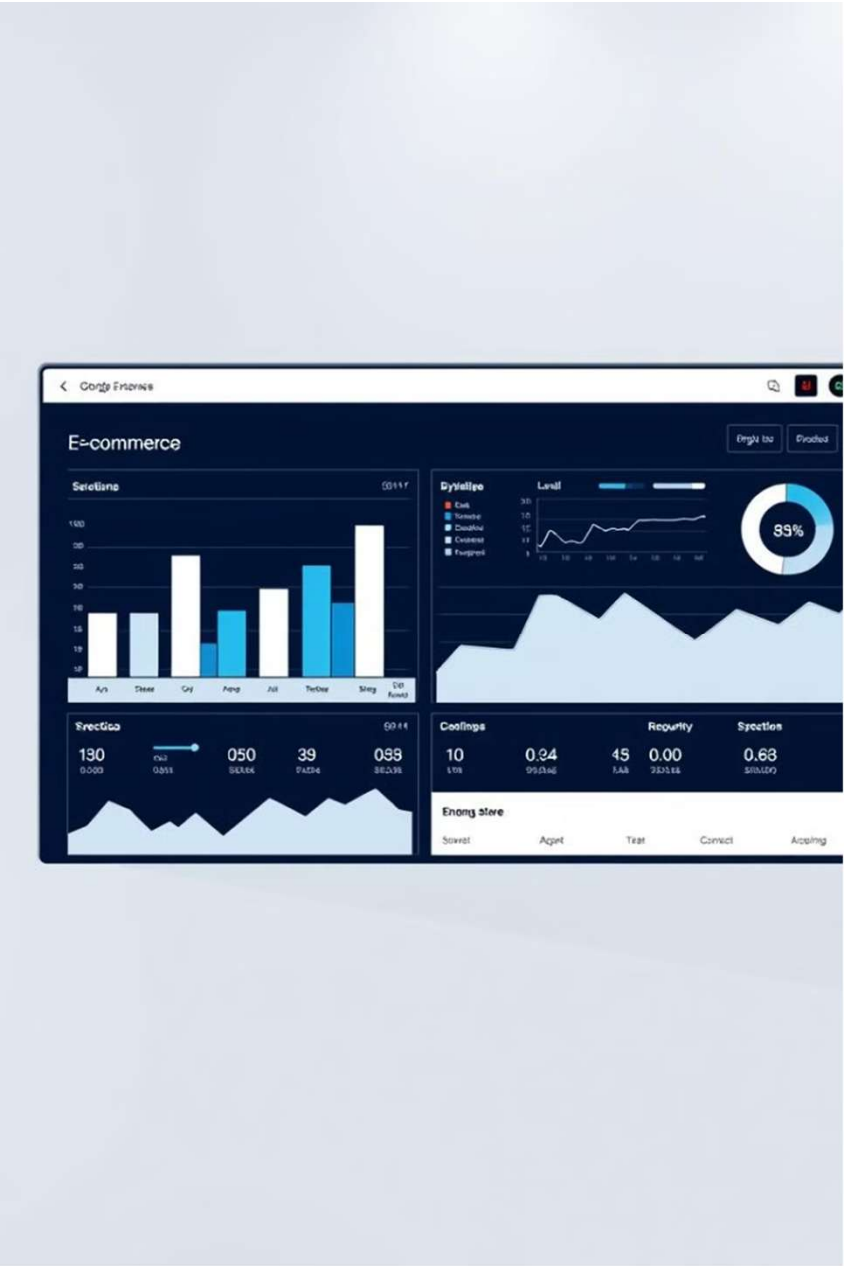


Projet d'Analyse de Données E-commerce

Ce document offre une vue d'ensemble complète du projet d'analyse de données e-commerce, en détaillant les outils, les technologies utilisées, l'architecture du pipeline et le flux de travail. De l'ingestion des données brutes à la création d'insights exploitables, le projet vise à construire un pipeline de données de bout en bout pour extraire, transformer et visualiser les données e-commerce. L'objectif est de créer un système scalable et facile à maintenir, fournissant des insights pertinents via des tableaux de bord interactifs.

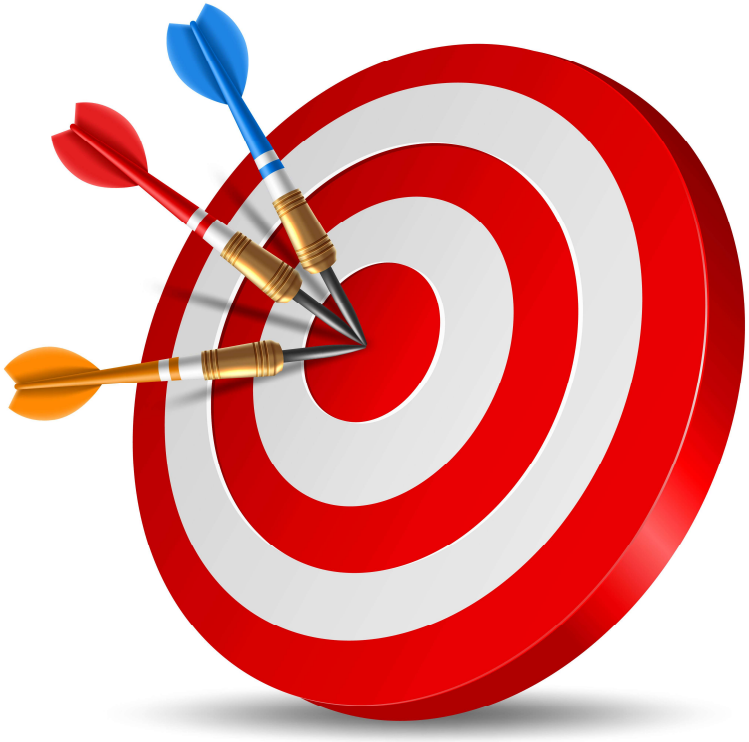
Objectifs Clés du Projet

1 Pipeline de Données Scalable

Construire un pipeline de données scalable et facile à maintenir est primordial. Un tel pipeline assure une gestion efficace des volumes croissants de données e-commerce, tout en simplifiant les mises à jour et les extensions futures.

2 Schéma en Étoile

Créer un schéma en étoile avec des tables de faits et de dimensions est essentiel pour répondre aux besoins analytiques. Cette structure facilite l'interrogation des données et l'analyse des performances e-commerce sous différents angles.



3 Data Marts pour le Reporting

Développer des data marts pour un reporting ciblé permet de fournir des informations spécifiques et pertinentes aux différents départements. Cela améliore la prise de décision et l'efficacité opérationnelle.

Outils et Technologies Utilisés

Google BigQuery

Utilisé comme entrepôt de données principal pour stocker les données brutes, transformées et agrégées. BigQuery offre une scalabilité et une performance élevées pour l'analyse de grands ensembles de données.

dbt (Data Build Tool)

dbt permet des transformations modulaires et scalables pour créer des tables de faits, des tables de dimensions et des data marts. Il facilite la gestion et la maintenance des transformations de données.

Python

Utilisé pour extraire les données de BigQuery et les charger dans PostgreSQL. Les bibliothèques google-cloud-bigquery, pandas et sqlalchemy sont employées pour cette tâche.

Power BI

Power BI est utilisé pour créer des tableaux de bord interactifs et des KPIs, directement connecté à BigQuery pour des insights en temps réel. Il offre une visualisation claire et intuitive des données.

Architecture du Pipeline de Données

1

Sources de Données

Les données sources sont des fichiers CSV contenant des transactions e-commerce. Ces fichiers incluent des détails sur les produits, les clients, les ventes et les marges de profit.

2

Transformation

Le pipeline de transformation comprend l'ingestion des données brutes, la création d'une couche de staging pour le nettoyage et la standardisation, ainsi que la modélisation dimensionnelle pour structurer les données.

3

Tables de Faits

Les tables de faits capturent les détails transactionnels et les agrégations de ventes. Deux tables principales sont utilisées : fait_ventes et fait_tendances_locales.

4

Data Marts

Les data marts sont créés pour agréger les données de ventes par ville et par mois (dm_tendances_locales), ainsi que les données de ventes et de profit par produit, catégorie et temps (dm_ventes_produits).

Aperçu du Flux de Travail

Ingestion des Données

1

Charger le fichier CSV brut dans BigQuery et assurer une structure de table appropriée avec des types de données corrects.

2

Transformation des Données

Utiliser dbt pour créer des tables de staging, des dimensions et des tables de faits, en appliquant des transformations pour obtenir des données propres et cohérentes.

Création des Data Marts

3

Agréger les données pour créer des data marts ciblés pour le reporting, en fonction des besoins spécifiques.

4

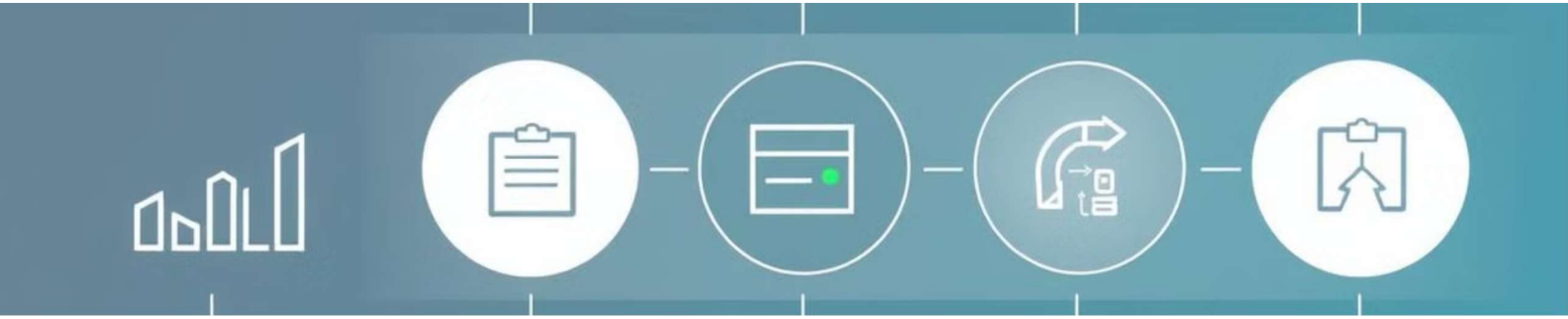
Extraction et Chargement

Utiliser des scripts Python pour exporter les data marts de BigQuery vers des fichiers CSV, puis charger les données extraites dans PostgreSQL.

Reporting

5

Connecter Power BI à BigQuery pour créer des tableaux de bord interactifs, affichant des KPIs tels que les ventes totales, le profit et les produits les plus performants.



Fonctionnalités Clés du Projet

Scalabilité

L'infrastructure cloud-native de BigQuery garantit une scalabilité pour de grands volumes de données, permettant de gérer efficacement la croissance des données e-commerce.

Automatisation

Le pipeline peut être automatisé à l'aide d'outils comme Apache Airflow, réduisant ainsi les interventions manuelles et assurant un flux de données continu.

Insights en Temps Réel

La connexion directe entre Power BI et BigQuery garantit des mises à jour quasi instantanées dans les tableaux de bord, offrant une vision en temps réel des performances e-commerce.

Analyse des Ventes – Tableau de Bord Power BI

1 Synthèse des Résultats

• Revenu Total & Profit Total : Suivi global des ventes et de la rentabilité.

• Marge de Profit : Indicateur clé de performance financière.

• Filtre temporel : Sélection d’une période spécifique pour l’analyse.

2 Analyse des Performances

• Classement des meilleures villes en termes de chiffre d’affaires.

• Analyse comparative des bénéfices et des revenus.

• Identification des produits les plus populaires.

3 Détails et Tendances

Analyse approfondie des produits les plus performants et de la répartition du profit par catégorie. Suivi de l'évolution des ventes au fil du temps.

Défis et Solutions

Gestion des Doublons

Le défi consiste à gérer les doublons d'identifiants produits avec des noms incohérents, ce qui peut fausser les analyses. La solution est d'utiliser ROW_NUMBER() en SQL pour sélectionner le nom de produit le plus fréquent pour les ID en doublon.

Intégrité des Données

Assurer l'intégrité des données lors des transformations est crucial. Cela est résolu par une validation approfondie des données avec des tests dbt pour détecter les anomalies et garantir la qualité des données.

Conclusion et Prochaines Étapes

Ce projet démontre l'intégration d'outils modernes de gestion des données pour créer un pipeline analytique scalable, maintenable et riche en insights. En utilisant BigQuery, dbt, Python et Power BI, la solution offre des insights exploitables tout en garantissant la qualité et la fiabilité des données. Les prochaines étapes pourraient inclure l'automatisation complète du pipeline avec Apache Airflow et l'exploration de modèles d'apprentissage automatique pour des analyses prédictives.