Cahier des charges - Projet Fil Rouge : Data Engineering

Par : Yassine Harrati

Objectif du projet

Le projet vise à démontrer les compétences en data engineering acquises lors de la formation de 8 mois à Youcode. Il consistera en la conception et la réalisation d'un pipeline de données pour le traitement des données provenant d'une plateforme d'e-commerce Alkasba Store. L'objectif final est de fournir une solution fonctionnelle, opérationnelle et sécurisée pour améliorer la présence sur le marché des entreprises du secteur du commerce en ligne.

Contexte

Le projet se concentrera sur la collecte, le traitement, le stockage, l'analyse en temps réel et batch, ainsi que la visualisation des données relatives au commerce en ligne à partir de la plateforme d'ecommerce Alkasba Store. Il intégrera également la mise en œuvre d'un pipeline avec deux workflows distincts : l'un pour la prise de décision en temps réel et l'autre pour le traitement des Big Data.

Étapes clés

1. Organisation et modélisation

Utilisation d'outils de gestion de projet avec Jira pour organiser et suivre l'avancement du projet.

2. Analyse des exigences

Compréhension des objectifs commerciaux spécifiques liés au secteur du commerce en ligne (augmentation des ventes, analyse de la concurrence).

Définition de la portée du projet en se concentrant sur les besoins d'analyse de données pour améliorer la performance et la compétitivité.

3. Collecte et ingestion des données

Utilisation de bibliothèques de Python pour collecter des données pertinentes à partir de l'API de la plateforme d'e-commerce Alkasba Store.

Mettre en place des mécanismes d'ingestion de données en temps réel pour obtenir des données de manière continue.

Gestion des erreurs, des tentatives et validation des données lors de l'ingestion.

4. Traitement initial des données

Nettoyage, prétraitement et transformation des données collectées en un format cohérent et utilisable.

Enrichissement des données en les combinant avec d'autres sources pour une analyse plus approfondie.

5. Stockage des données

Choix de solutions de stockage appropriées en fonction du flux de travail (bases de données NoSQL (ElastickSearch) pour le temps réel, datalakes pour le stockage à grande échelle).

6. Traitement et analyse des données

Mise en œuvre d'une logique d'analyse en temps réel pour permettre des prises de décision rapides concernant les données du commerce en ligne.

Mise en place de mécanismes d'alerte pour des événements spécifiques en temps réel.

Mise en place de jobs de traitement par lots pour une analyse plus approfondie des données en batch.

7. Chargement des données

Chargement des données traitées dans les systèmes appropriés (par exemple, tableaux de bord pour l'analyse en temps réel, entrepôts de données pour l'analyse BI).

8. Visualisation et rapports

Conception et mise en œuvre de tableaux de bord interactifs pour visualiser les données du commerce en ligne.

Création de rapports BI pour des analyses approfondies.

9. Data Modeling

Utilisation de modèles de machine learning pour prédire les tendances du commerce en ligne et optimiser les performances.

10. Sécurité et conformité

Assurer la confidentialité et la sécurité des données, en particulier pour les données sensibles du commerce en ligne.

Mettre en place des contrôles d'accès et le chiffrement des données conformément aux réglementations.

Livrables attendus

- Documentation détaillée du processus de collecte, de traitement, de stockage, d'analyse et de visualisation des données.

- Code source documenté.

- Tableaux de bord interactifs et rapports BI.

Contraintes

- Utilisation des technologies spécifiées, notamment Python, SQL, Talend, Power BI, NoSQL databases (Cassandra, MongoDB), PostgreSQL, etc.

- Respect des compétences du référentiel de formation.