1. Introduction

Contexte et Objectifs

L’entreprise de commerce en ligne exploite diverses sources de données (transactions clients, logs serveurs, données des réseaux sociaux, campagnes publicitaires) pour améliorer ses décisions stratégiques. Le Data Lake a été conçu pour centraliser ces données, automatiser leur transformation, et fournir un accès simplifié aux analyses et tableaux de bord.

Portée

Le projet intègre les quatre types de données mentionnés dans un Data Lake hébergé sur **MongoDB Atlas**, avec ingestion automatisée via Kafka, transformation via PySpark, et exposition via une API dédiée.

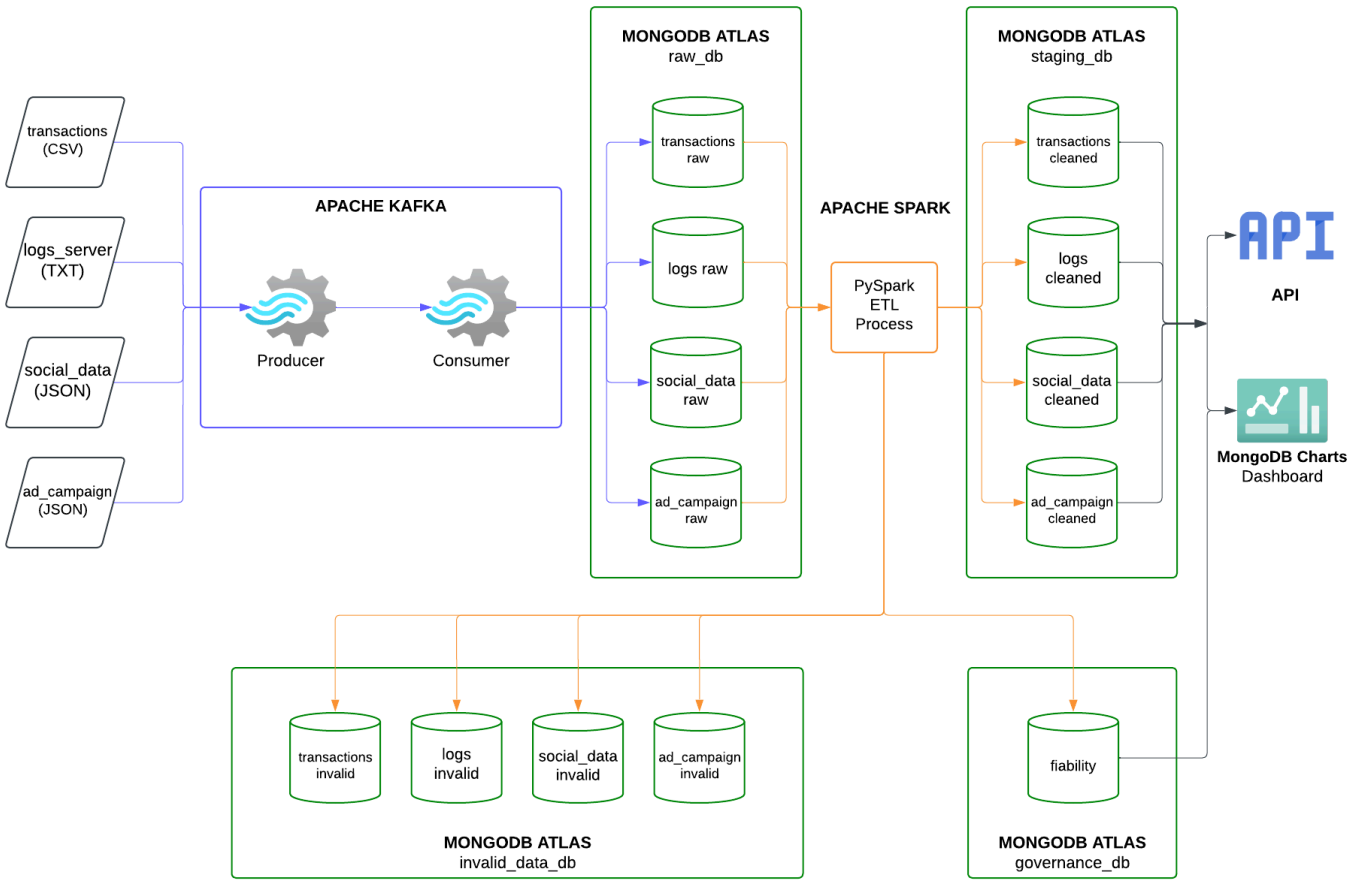
2. Architecture du Data Lake

2.1 Vue d'ensemble

L'architecture se compose de plusieurs couches fonctionnelles :

1. **Ingestion :** Collecte continue des données locales via Kafka.
2. **Raw Layer :** Stockage brut dans la base de données `raw_db` sur MongoDB Atlas.
3. **Transformation :** Nettoyage et enrichissement des données brutes via PySpark.
4. **Staging Layer :** Stockage des données nettoyées dans `staging_db` sur MongoDB Atlas.
5. **Accès aux données :** Exposition des données nettoyées via une API REST et visualisation sur un tableau de bord MongoXDB Atlas.

2.2 Diagramme des flux de données



3. Source et Stockage des Données

3.1 Sources de données locales

Les données proviennent de quatre fichiers locaux avec les formats suivants :

Fichier	Format	Contenu	Exemple
`transactions.csv`	CSV	Transactions des clients.	ID, montant, date, etc.
`logs_server.txt`	Fichier texte	Journaux des serveurs web.	Méthodes HTTP, URL, IP, etc.
`social_data.json`	JSON	Publications et interactions sociales.	Hashtags, réactions, texte, etc.
`ad_campaign.json`	JSON	Données de campagnes publicitaires.	Impressions, clics, conversions, etc.

3.2 Conversion des fichiers en JSON

- Les fichiers locaux sont surveillés par un **Producer Kafka**.
- **Transformation des formats :**
    1. Les fichiers CSV (`transactions.csv`) sont convertis en JSON.
    2. Les fichiers texte (`logs_server.txt`) sont analysés et convertis en JSON.
    3. Les fichiers JSON (`social_data.json`, `ad_campaign.json`) sont directement ingérés.

3.3 Stockage des données

Les données sont stockées dans deux bases de données sur MongoDB Atlas :

1. **raw_db** (données brutes) :
    - Collections : `transactions_raw`, `logs_raw`, `social_data_raw`, `ad_campaign_raw`.
    - Structure des fichiers après ingestion, sans transformation.
2. **staging_db** (données nettoyées) :
    - Collections : `transactions_cleaned`, `logs_cleaned`, `social_data_cleaned`, `ad_campaign_cleaned`.
    - Données nettoyées et enrichies via PySpark.

4. Flux de Données

4.1 Ingestion via Kafka

- **Technologie utilisée :** Kafka (avec Zookeeper).
- **Étapes :**
    1. Les fichiers locaux sont surveillés par le Producer Kafka.
    2. Le Producer Kafka effectue les conversions nécessaires en JSON.
    3. Les fichiers JSON sont publiés dans des **topics Kafka** spécifiques :
        - `transactions_topic`
        - `logs_topic`
        - `social_data_topic`
        - `ad_campaign_topic`
    4. Le Consumer Kafka consomme les messages et insèrent les données dans MongoDB Atlas (`raw_db`).

4.2 Transformation via PySpark

- **Technologie utilisée :** PySpark.
- **Étapes :**
    1. Lecture des données brutes dans `raw_db`.
    2. Nettoyage et transformation des données :
        - Suppression des doublons.
        - Conversion des dates dans un format unique (`ISO 8601`).
        - Correction des anomalies (ex. gestion des valeurs manquantes).
    3. Enregistrement des données nettoyées dans `staging_db`.

4.3 Exposition des données

- **API REST :**
    - Une API expose les données de `staging_db`.
    - Fonctionnalités :
        - Interroger les transactions (`customer_id`).
        - Analyser les clics et conversions par campagne (`ad_id`).
        - Accéder aux moyens de paiements (payment_method)

- **Tableaux de bord MongoDB Atlas :**

- Visualisent les données nettoyées pour des cas d’utilisation spécifiques :
    - Performances des campagnes publicitaires.
    - Comportements des utilisateurs sur le site web.

5. Sécurité et Gouvernance

5.1 Sécurité

- **Contrôle d'accès :**
    - RBAC (Role-Based Access Control) sur MongoDB Atlas.
    - Kafka protégé par des ACLs pour sécuriser les topics.

- **Chiffrement des données :**
    - Chiffrement au repos : AES-256 sur MongoDB.
    - Chiffrement en transit : TLS/SSL entre Kafka, PySpark et MongoDB.

5.2 Gouvernance

- **Documentation des données :**
    - Utilisation d'un catalogue des données

- **Politique de rétention :**
    - Logs serveur : Conserver pendant 1 an.
    - Données transactionnelles et sociales : Rétention illimitée.

- **Rôles des différents Data Owners :**

    DSI (Direction des Systèmes d'Information)

    - La DSI est responsable de la **sécurité des données**, en mettant en place les outils, les politiques, et les protocoles nécessaires pour protéger les données (ex. chiffrement, gestion des accès, sauvegardes).
    - Elle garantit également la **fiabilité des infrastructures** techniques (ex. bases de données MongoDB, Kafka) et veille à la conformité des systèmes avec les réglementations (ex. RGPD).
    - Enfin, la DSI supervise les processus d'ingestion, de transformation, et de stockage des données pour assurer leur disponibilité et leur intégrité.

Marketing

- Le département Marketing est le **propriétaire des données des campagnes publicitaires**. Il est responsable de la définition des règles de collecte et d'utilisation des données liées aux campagnes (ad campaigns), comme les impressions, clics, et conversions.
- Le rôle inclut également l'analyse des performances des campagnes et la garantie que les données sont conformes aux besoins opérationnels et stratégiques.

Marketing Digital

- Ce département est chargé des **données sociales** issues des réseaux sociaux (social_data).
- Il définit les politiques de gouvernance pour l'analyse des interactions (hashtags, réactions) et s'assure que les données sont utilisées dans le respect des politiques de confidentialité des plateformes (ex. utilisation des API des réseaux sociaux).
- Le marketing digital joue également un rôle clé dans la valorisation des données pour améliorer les stratégies numériques.

Finance

- Le département Finance est responsable des **données transactionnelles**. Il garantit leur exactitude et leur exhaustivité, car ces données sont essentielles pour la gestion des revenus et la comptabilité.
- Il supervise les règles de rétention et de classification des données sensibles, tout en assurant leur conformité avec les réglementations financières et fiscales.
- En outre, Finance collabore avec la DSI pour s'assurer que les données de paiement et de transactions sont sécurisées.

6. Outils et Technologies

Fonctionnalité	Outil/Technologie	Description
Ingestion	Kafka	Surveillance des fichiers locaux et ingestion en continu.
Stockage	MongoDB Atlas	Base de données NoSQL pour les données brutes et nettoyées.
Transformation	PySpark	Nettoyage, enrichissement et uniformisation des données.
Accès aux données	API REST	Exposition des données nettoyées.
Visualisation	Tableau de bord MongoDB	Analyse des données nettoyées.