Document de Conception : Architecture du Data Lake

1. Introduction

L'objectif de ce document est de présenter l'architecture du Data Lake conçu pour une entreprise de commerce en ligne souhaitant centraliser et exploiter efficacement ses données provenant de différentes sources. Ce Data Lake permettra de collecter, transformer, analyser et gouverner les données pour faciliter des décisions commerciales.

2. Analyse des Besoins

Sources de Données

Source	Type de Données	Format	Fréquence
Transactions clients	Structurées	SQL	Par lots
Logs des serveurs web	Non structurées	Texte brut (log)	Par lots
Données des médias sociaux	Semi-structurées	JSON	Temps réel
Campagnes publicitaires	Semi-structurées	JSON, Avro	Temps réel

Exigences Fonctionnelles

- Centralisation : Collecter et centraliser les données brutes.
- Flexibilité : Permettre de traiter différents types de données (structurées, non structurées, semi-structurées).
- Analyse : Fournir des données nettoyées pour des outils analytiques.
- Temps réel : Intégrer les flux de données en temps réel (publicités, médias sociaux).

Exigences Techniques

- Stockage scalable : Gestion de grandes quantités de données.
- Compatibilité : Intégration avec outils d'analyse (ex. PySpark, Tableau).
- Coût : Solution économique (par exemple, S3 pour le cloud, HDFS pour local).

3. Architecture Logique

Le Data Lake est divisé en trois zones principales pour organiser les données et faciliter leur traitement :

1. Raw Zone (Zone Brute)

- Stocke les données dans leur format d'origine sans transformation.
- Structure des dossiers basée sur les sources et la date de réception.

  - Exemple :
    - `/data/raw/transactions/yyyy/mm/dd/`
    - `/data/raw/logs/yyyy/mm/dd/`

2. Cleansed Zone (Zone Nettoyée)

- Stocke les données après nettoyage et transformation (normalisation, suppression des doublons).

  - Exemple :
    - `/data/cleansed/transactions/yyyy/mm/dd/`
    - `/data/cleansed/logs/yyyy/mm/dd/`

3. Business Zone (Zone Analytique)

- Contient des données prêtes pour l'analyse, agrégées ou organisées en datasets spécifiques.

  - Exemple :
    - `/data/business/ads_performance/`

4. Architecture Physique

Exemples de Technologies

Composant	Technologie	Justification
Stockage	HDFS	Scalable, performant, gros volumes
Collecte des Données	Kafka	Collecte des données par lots et en temps réel
Gouvernance	Apache Atlas	Suivi des métadonnées

Organisation des Données

Les fichiers sont organisés de manière hiérarchique pour une gestion efficace :

```
/data/
  raw/
    transactions/
    logs/
    social_media/
    ads/
  cleansed/
    transactions/
    logs/
    social_media/
    ads/
  business/
    ads_performance/
    customer_segments/
```

Flux de Données

Exemple de collecte

- Transactions clients : Extraction via Sqoop ou ETL.
- Logs web : Collecte avec Fluentd.
- Médias sociaux : APIs des plateformes (ex. Facebook).
- Publicité en temps réel : Kafka pour les flux.

Exemple de traitement

- Raw Zone → Cleansed Zone : Nettoyage avec PySpark.
- Cleansed Zone → Business Zone : Agrégation et structuration.

Exemple d'analyse

Les données de la Business Zone sont consommées par des outils analytiques (Tableau, Power BI) ou via des notebooks (Jupyter, Databricks).