Planification du Projet Big Data – EcomData

1. Contexte et Enjeux

EcomData est une entreprise spécialisée dans la vente de produits électroniques via une plateforme e-commerce. Chaque jour, des milliers de clics, de consultations et de commandes génèrent un volume massif de données. Actuellement, ces données sont dispersées sur plusieurs systèmes, stockées sous des formats variés comme CSV et JSON, et gérées manuellement. Cela cause des erreurs, des retards et limite l'exploitation analytique. L'objectif est donc de construire un pipeline Big Data robuste, automatisé, et conforme aux exigences réglementaires.

---

2. Objectifs du Projet

Le projet vise à automatiser l'ingestion des données issues des commandes et des interactions utilisateurs. Les données doivent être stockées de manière organisée sur Amazon S3 pour les fichiers CSV et envoyées vers Kafka pour les fichiers JSON. L'ensemble du processus sera orchestré par Apache Airflow, qui assurera l'automatisation et le bon enchaînement des étapes.

---

3. Phases du Projet

Phase 1 : Analyse et Conception

Cette phase consiste à comprendre les besoins métiers et techniques, définir la structure des données à traiter (commandes et logs), et concevoir une architecture cible intégrant les outils nécessaires : Python, NiFi, Airflow, Kafka, S3. Un schéma clair des données à générer est établi ici.

Phase 2 : Génération des Données

Un script Python sera développé pour simuler des commandes au format CSV et des interactions utilisateurs au format JSON. Les fichiers générés seront enregistrés dans un dossier local nommé `./Staging`, qui servira de source pour les étapes suivantes.

Phase 3 : Ingestion avec Apache NiFi

Deux flux de données seront créés dans NiFi. Le premier permettra de détecter les fichiers CSV dans le dossier `./Staging` et de les envoyer automatiquement vers Amazon S3. Le second flux détectera les fichiers JSON et les publiera dans un topic Kafka. Chaque flux repose sur des processeurs NiFi tels que `GetFile`, `PutS3Object`, et `PublishKafkaRecord`.

Phase 4 : Orchestration avec Apache Airflow

Un DAG Airflow sera conçu pour automatiser tout le pipeline. Il exécutera le script Python toutes les heures, puis déclenchera les flux NiFi pour ingérer les fichiers. Enfin, une vérification automatique confirmera que les données ont bien été transférées vers S3 et Kafka.

Phase 5 : Conformité RGPD et Gouvernance

Afin d'assurer la conformité avec le RGPD, les données sensibles devront être anonymisées si nécessaire. Des règles de gouvernance seront mises en place pour garantir la qualité, la sécurité, et la traçabilité des données. Un catalogue de données détaillé sera développé pour documenter les sources, formats et définitions des données traitées.

Phase 6 : Tests et Livraison

Cette dernière phase comprend les tests de bout en bout du pipeline (génération → ingestion → stockage). Une fois validé, le pipeline sera déployé sur l'environnement de production. Toute la documentation technique et fonctionnelle sera également livrée.

---

4. Ressources et Responsabilités

Le projet impliquera plusieurs profils : un data engineer pour développer le script Python, les flows NiFi et le DAG Airflow ; un architecte data pour concevoir l'architecture technique ; un expert sécurité pour s'assurer de la conformité RGPD ; et un chef de projet pour coordonner l'ensemble.

---

5. Outils et Technologies Utilisés

Le script de génération sera écrit en Python. Pour l'ingestion, Apache NiFi sera utilisé afin de simplifier le transfert vers S3 et Kafka. Apache Airflow orchestrera le pipeline, avec des tâches planifiées toutes les heures. Les données CSV seront stockées sur Amazon S3, tandis que les JSON seront envoyées vers Kafka avec l'option d'intégration à Iceberg. Enfin, des outils de gouvernance de données (comme un Data Catalog en YAML ou open-source) seront utilisés pour documenter l'ensemble du pipeline.

---

6. Suivi et Indicateurs Clés

La réussite du projet sera mesurée à l'aide d'indicateurs comme le taux d'automatisation du pipeline, le temps moyen d'ingestion, le taux d'erreur des tâches Airflow, le niveau de couverture RGPD (anonymisation, gestion des droits), et le nombre de jeux de données documentés dans le catalogue.