Projet Big Data – Étude de cas finale

Sujet : Conception et mise en œuvre d'une plateforme Big Data de surveillance et d'analyse du trafic urbain en temps réel

1. Contexte général

Dans le cadre du développement des Smart Cities, les collectivités territoriales exploitent aujourd'hui des volumes massifs de données issues de multiples sources (capteurs routiers, systèmes de comptage, données météorologiques, événements urbains, etc.). Ces données, caractérisées par leur volume, leur vélocité et leur variété, nécessitent la mise en place d'architectures Big Data robustes, capables de traiter à la fois des données historiques et des flux temps réel.

Une métropole souhaite concevoir une plateforme Big Data permettant :

- L'analyse historique du trafic urbain,
- La surveillance en temps réel des conditions de circulation,
- La détection d'événements anormaux (congestion, ralentissements inhabituels).

Vous intervenez en tant qu'équipe Big Data chargée de concevoir, déployer et documenter une solution technique complète, en respectant les principes de l'architecture Big Data, du data streaming et de la gestion de pipelines de données.

2. Objectifs du projet

Ce projet a pour objectif de mettre en application les notions vues dans le module *Déploiement de solution – Architecture Big Data*, notamment :

- Concevoir une architecture Big Data cohérente et scalable
- Mettre en œuvre un traitement batch sur des données volumineuses
- Intégrer un système de streaming de données
- Développer un pipeline de données de bout en bout
- Documenter une solution dans une logique de maintenabilité et de déploiement

3. Organisation du travail

- **Travail en groupe** de 3-4 personnes
- **Rendu :** rapport technique

4. Données

4.1 Données batch (données historiques)

Vous utiliserez un jeu de données massif représentant le trafic routier urbain :

Par exemples US Traffic Dataset

- Données de trafic et données météorologiques associées

Lien : Traffic Prediction Dataset

4.2 Données streaming (temps réel)

Vous avez la possibilité de simuler un flux de données temps réel représentant des capteurs de trafic urbain.

Exemples de données :

- Identifiant du capteur
- Horodatage
- Vitesse moyenne
- Densité du trafic

Les messages devront être envoyés sous forme JSON vers un système de streaming.

5. Architecture attendue

L'architecture globale devra inclure, a minima :

- Une source de données batch
- Une source de données streaming
- Un système de messagerie pour le streaming
- Un moteur de traitement Big Data
- Une zone de stockage
- Un pipeline de données documenté

Un schéma d'architecture clair et lisible est **obligatoire** dans le rapport.

6. Technologies autorisées

Stack recommandée :

- **Apache Kafka** : ingestion des flux temps réel
- **Apache Spark** :

    - Spark SQL pour le batch
    - Spark Streaming pour le streaming
- **Apache NiFi** (optionnel) : gestion des flux et pipelines
- **Apache Airflow** (optionnel) : orchestration
- **HDFS** ou système de fichiers local : stockage
- **Docker / Docker Compose** : déploiement et simulation d'infrastructure

7. Travaux demandés

7.1 Conception

- Définir l'architecture Big Data
- Justifier les choix techniques

7.2 Traitement batch

- Ingestion des données historiques
- Traitements analytiques avec Spark SQL
- Stockage optimisé (partitionnement, formats)

7.3 Streaming

- Mise en place d'Apache Kafka
- Développement d'un producteur de données
- Consommation et traitement en temps réel avec Spark Streaming

7.4 Pipeline de données

- Description complète du pipeline
- Gestion des erreurs et surveillance (conceptuelle ou implémentée)

8. Livrable attendu

Rapport technique

Le rapport devra contenir :

1. Présentation du contexte et des enjeux
2. Description de l'architecture globale
3. Détail des outils et environnements utilisés
4. Explication des pipelines de données

5. Captures d’écran commentées
6. Difficultés rencontrées et solutions apportées
7. Limites et axes d’amélioration

9. Critères d’évaluation

- Pertinence de l’architecture Big Data
- Mise en œuvre du streaming
- Qualité du pipeline de données
- Justification des choix techniques
- Qualité et clarté du rapport