Projet Big Data

Ingestion, processing, visualisation

Faire du Spark sur des données pédagogiques c’est bien. Se confronter à des problématiques réellement Big Data, c’est mieux. Le but de ce projet est de développer un pipeline big data qui devra être capable de traiter une forte volumétrie de données à intervalles réguliers (batch) OU/ET des données en temps réel (streaming).

# Contexte :

Votre “client” (source de donnée que vous aurez choisie) génère des données qu’il va vous envoyer sur votre infrastructure. Vous devez les traiter avant le prochain batch ou en temps réel et les mettre à disposition de votre client de manière visuelle.

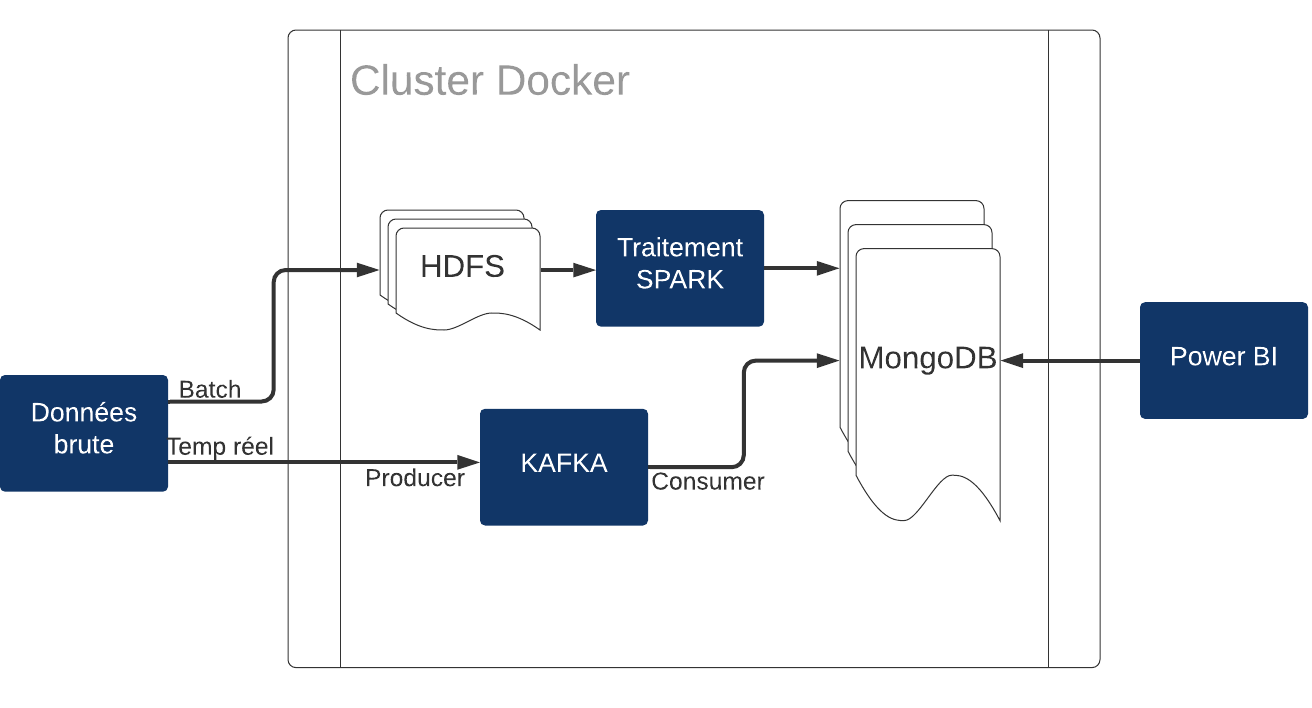
Vous aurez à créer l’architecture de votre pipeline qui devra faire des statistiques descriptives simples.

Les contraintes de vos clients sont les suivantes :

* Vous aurez accès a un serveur mongodb et un cluster Hadoop HDFS pour stocker les données.
* Vous devez effectuer le processing des données avec Apache Spark ou/et Apache

Kafka

Exemple de pipeline:



# Livrables :

1) Présentation powerpoint (ou autre outil de présentation) avec:

* Méthodologie
* Présentation de votre sujet/source de donnée
* Présentation de l’architecture
* Justification de vos choix de technos
* Difficultés rencontrées, axe d’amélioration

BONUS:

* Connecter un Power BI qui sera chargé de faire la visualisation des données processées qui se mettra à jour automatiquement (en temps réel ou à chaque nouveau batch).
* Ordonnancer vos traitements spark avec des crontab
* Présenter votre dashboard