Projet Big Data (version 2)

Ingestion, processing, visualisation.

La valeur du big data n'est pas dans la quantité des données, mais dans son utilisation efficace et sa capacité à fournir de la valeur ajoutée, en choisissant l'architecture et les technologies qui vont permettre de l'adapter à un contexte métier particulier.

Le but de ce projet est de développer une pipeline big data qui devra être capable de traiter automatiquement des données provenant de plusieurs sources qui arrivent à des fréquences différentes, voir en temps réel, et les mettre à disposition de votre client de manière visuelle.

Feuille de route :

Vous devrez pour ce travail, imaginer dans un premier temps votre projet dans son ensemble: contexte métier, objectif, sources de données.

Une fois vos sources de données identifiées, vous devrez choisir par quel moyen envoyer quelles données (brutes) dans votre pipeline, et créer les scripts correspondant (producer pour les données qui passent par kafka par exemple)

Ces données devront ensuite être traitées avec pyspark. Pour cela vous pouvez créer des dataset test avec vos données puis faire vos test dans votre machine docker avec le pyspark notebook. Puis une fois vos traitements définis sur notebook, vous pourrez en faire un ou des scripts que l'on pourra soumettre ensuite avec un spark-submit.

Les données une fois traitées devront alimenter une base de données Mongo qui les mettra à disposition pour votre objectif final (webapp, dashboard…)

Automatisation :

Pour lancer vos différents scripts automatiquement à intervalle régulier vous pourrez utiliser crontab (linux) : https://www.linuxtricks.fr/wiki/cron-et-crontab-le-planificateur-de-taches

ou le planificateur de tâche (windows) :
https://www.pcastuces.com/pratique/astuces/5515.htm

Pensez bien dans vos traitements, à gérer vos fichiers entrants : par exemple si votre traitement spark va chercher un fichier dans un répertoire hdfs, ce fichier doit il être déplacé? ou supprimé? afin qu'il ne soit pas retraité à chaque lancement du traitement?

Objectif (visualisation des données):

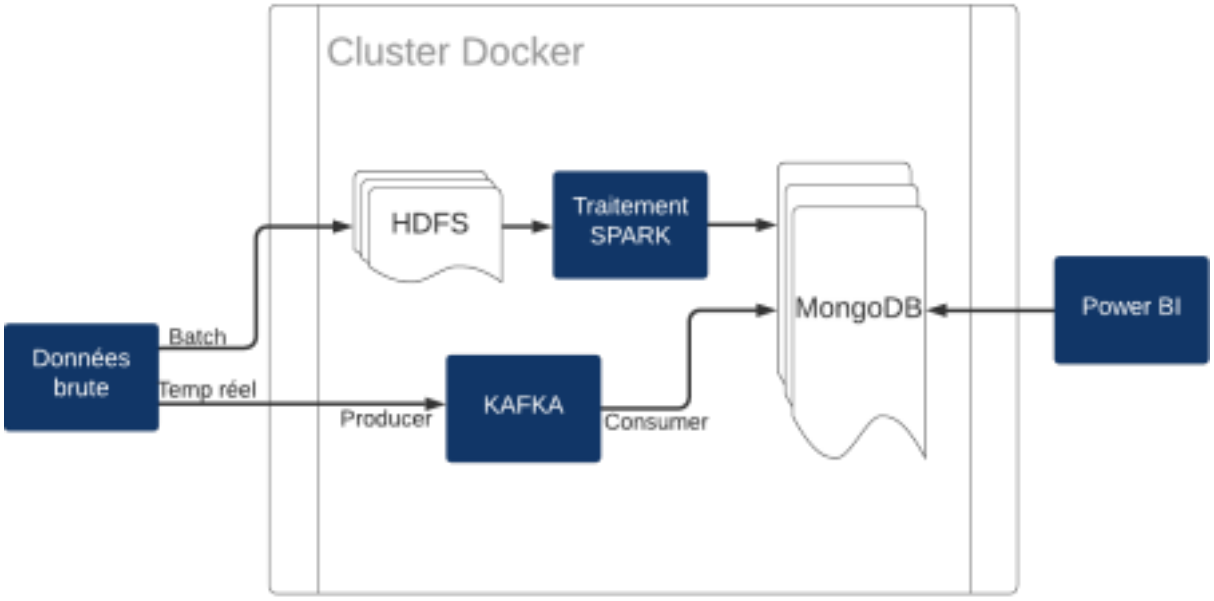
A vous de choisir dans la première étape de la feuille de route le contexte métier et la forme sous laquelle vous voulez exposer vos données. Pour cela vous pouvez vous posez la question suivante:

Qui est l'utilisateur final des données?

Garder la réponse à cette question à l'esprit lorsque vous modélisez votre interface pour donner du sens à l'information, à son utilisation, ainsi qu'à son interprétation, pour permettre une prise de décision rapide et efficace.

Pour réaliser l'interface vous pourrez utiliser les technos de votre choix, dash, flask, streamlit, powerbi, ….

Exemple de schéma de pipeline:



Les rendus et livrables demandés :

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

-méthodologie

-présentation de votre sujet/source de donnée

-présentation de l'architecture (avec un peu de code, montrez vos scripts, expliquez comment ils marchent ensemble)

-justification de vos choix de technos

-difficultés rencontrées, axe d’amélioration

-présentation de votre visuel