Projet Hadoop et IA

Objectif Global

Le but global de ce projet est de développer un projet complet de gestion de grandes bases de données et d'utilisation de modèles d'intelligence artificielle. Le projet se distingue en 2 grandes parties :

1. **Projet Hadoop** (Infrastructure et Traitement des données)
2. **Projet IA** (Développement de l'API pour un LLM et un modèle d'images)

Projet Hadoop

Le but du **Projet Hadoop** est de déployer une infrastructure **Big Data** capable de stocker, traiter et exploiter de grandes quantités de données en texte et en image. Le projet se compose des étapes suivantes :

- **Création** d'une infrastructure Hadoop et de bases de données associées.
- **Récolte** de deux bases de données existantes (Textes & Images).
- **Création** d'un flux de données entrant via scraping web.
- **Utilisation** d'une API d'IA externe (Projet IA) pour envoyer les données prétraitées et récupérer les résultats d'analyse.

Projet IA

Le but du **Projet IA** est de développer une **API centralisée** qui expose deux modèles d'IA :

- Un **LLM (Large Language Model)** pour analyser et traiter des textes.
- Un **modèle de Vision (CNN, Transformer)** pour classifier des images.

Projet Hadoop : Infrastructure et Traitement des Données

1. **Déploiement Hadoop & Stockage**
    - **Création** d'une architecture Hadoop avec **Docker compose** permettant de simuler votre environnement. Cette architecture doit être modulable en nombre de nœuds DataNode (i.e. le nombre de nœuds du cluster doit être un paramètre).
    - Le **choix des bases de données** utilisées pour le cluster multi-noeuds de Hadoop est libre (**Hive, HBase, etc.**). Vous pouvez utiliser plusieurs systèmes de bases de données selon vos besoins.
    - **Automatisation** de l'installation avec **Ansible**. Le nombre de nœuds du cluster peut être fixé manuellement : il n'est pas obligatoire de rendre cette partie également modulable.

2. **Ingestion et streaming des données**
    - **Scraping web** en temps réel (choix libre de l'outil, si possible **Kafka + Spark Streaming**). Le but est de récolter de nouvelles données pour compléter vos bases de données existantes ou pour créer une nouvelle base de données.
    - **Transformation et stockage** dans la base de données choisie dans l'environnement Hadoop.

3. **Prétraitement des données (avant envoi à l'API IA)**
    - **Nettoyage** des données pour les 2 bases de données. Ce traitement peut être prévu au moment de la création des bases de données ou du scraping.
    - **Tokenisation** des textes pour envoyer vers l'API. Cette partie peut aussi être faite côté API selon vos choix architecturaux.
    - **Conversion** des images dans un format optimisé pour envoi sur une API. Le format adapté pour l'image correspond à un tableau d'octet.

4. **Interaction avec l'API IA**
    - **Envoi** des données textes et images via l'API IA (Projet IA). Mettre en forme les données réceptionnées pour qu'elles soient exploitables.
    - **Stockage** des résultats IA dans la base de données de votre choix pour exploitation. Ce système de base de données peut être différent de vos autres systèmes mis en place.

5. **DevOps & Automatisation**
    - **Mise en place** d'un pipeline CI/CD pour les traitements Hadoop avec GitLab CI/GitHub Actions. Le choix du pipeline et des éléments mises en place est libre.
    - **Ajouter** tout autre élément qui vous semble utile (monitoring, logging, testing).
    - **Optionnel** : Mise en place d'un déploiement automatique de votre solution. Par exemple, génération automatique d'une image Docker à chaque push sur la branche principale.

6. **Visualisation & Exploitation**
    - **Développement** de tableaux de bord avec des outils au choix. Le but est de visualiser les analyses faites par vos IAs sur les données de votre base de données.
    - **Optionnel** : Ajout d'une visualisation générale pour votre base de données (nombre de données, informations diverses sur la base de données, etc.).

Projet IA : API pour LLM & Vision

1. **Développement de l'API IA**
    - **Mettre en place** une API REST commune pour les deux modèles. L'objectif est d'avoir une URL unique avec laquelle communiquer, un point d'entrée unique pour envoyer des textes et images.
    - **Déploiement** avec FastAPI ou Flask selon votre choix.
    - Retourner les données pour un format optimisé pour la base de données que vous avez choisie dans la phase d'Hadoop.

2. **Modèle IA**
    - **Adapter** un modèle avec le fine-tuning d'un **LLM type GPT / BERT** pour effectuer une tâche au choix :
        - Classification des sujets.
        - Résumé automatique.
        - Analyse de sentiment.
        - Etc..
    - **Adapter** le modèle d'image Yolo pour faire de la détection d'images.

3. **Ré-entraînement du modèle YOLO avec les images scrappées**
    - **Récupération** des images depuis Hadoop (HDFS/HBase). Cette fonctionnalité n'a pas besoin de passer par votre API. Elle peut figurer autant dans la partie Hadoop que la partie IA selon votre architecture.
    - **Prétraitement** des images (redimensionnement, conversion format, nettoyage des données).
    - **Génération** des annotations (manuelle ou semi-automatique). Il est possible de récupérer des annotations déjà existantes lors de la phase de scraping.
    - **Fine-tuning** de YOLO avec les images scrappées. Vous pouvez vous baser sur ce tutoriel pour faire un entrainement : https://docs.ultralytics.com/fr/modes/train/

4. **DevOps & Industrialisation**
    - **Mise en place** d'un pipeline CI/CD pour les traitements Hadoop avec GitLab CI/GitHub Actions. Le choix du pipeline et des éléments mises en place est libre.
    - **Ajouter** tout autre élément qui vous semble utile (monitoring, logging, testing).
    - **Mise en place** d'un déploiement automatique de votre solution. Cela implique la génération automatique d'une image Docker.

Rédaction du rapport

Le rapport doit contenir :

- De maniere generale :
  - Présentation du projet et de l'architecture globale.
- Pour la partie Hadoop :
  - Description de l'architecture Hadoop et explication des choix de bases de données.
  - Description des bases de données récoltées et des traitements prévus.
  - Description de la stratégie mise en place pour le scraping de données.
  - Analyse des performances selon la taille du cluster.
- Pour la partie IA :
  - Description de l'architecture de l'API IA et explication des choix de framework utilisés.
  - Présentation du LLM et du modèle Vision.
  - Analyse des performances des modèles (temps d'exécution, ressources utilisées, etc.).
- Pour chaque partie :
  - Présentation de la stratégie DevOps et CI/CD.

Tout autre élément qui pourrait étayer votre présentation sera le bienvenu.

Livrables attendus

Le projet devra se composer :

- D'un rapport sous forme de document Word ou PDF contenant toutes vos explications et descriptions du projet. Le document peut être en français ou en anglais.
- 2 repositories GitHub contenant les codes pour la partie Hadoop d'un côté (à mettre dans l'organisation https://github.com/orgs/hadoop-89) et ceux pour la partie IA de l'autre (à mettre dans l'organisation https://github.com/orgs/data-mining-ia-89). Tout le contenu devra être en anglais (commentaires et README inclus).

Deadline

Rapport pour le 3 Juin 2025, à 23h59.

Code pour le 12 Juin 2025, à 23h59.

Soutenance orale le 13 Juin 2025, de 9h à 12h30 et de 14h à 17h30.