Projet#4

Développeur Fullstack Big Data - **L'École Multimédia**

Sommaire

Contexte

Déploiement d'une application web sécurisée pour l'analyse de données météorologiques structurées et non structurées avec l'écosystème Hadoop et ElasticSearch, en utilisant exclusivement les données du site NOAA, dans des conteneurs Docker.

Présentation

Vous travaillez pour une startup spécialisée dans l'analyse des données météorologiques pour prévoir les tendances climatiques et aider les agriculteurs à optimiser leurs récoltes.

Votre mission est de concevoir et déployer une plateforme Big Data permettant de stocker, traiter et analyser de grandes quantités de données météorologiques **structurées et non structurées**, en utilisant exclusivement les données fournies par la **National Oceanic and Atmospheric Administration (NOAA)**.

En raison des ressources limitées, vous utiliserez des conteneurs Docker pour simuler un environnement de production.

Objectifs

A l'issue du projet, vous devrez avoir réalisé les éléments suivant :

1. **Installer et configurer l'écosystème Hadoop** dans des conteneurs Docker pour le stockage et le traitement des données.
2. **Concevoir et déployer un entrepôt de données** pour stocker des données météorologiques structurées et non structurées provenant de la NOAA.
3. **Définir l'architecture des données** en détaillant les flux et le stockage.
4. **Mettre en place ElasticSearch** pour permettre des recherches étendues sur les données.
5. **Développer des requêtes SQL et NoSQL** pour analyser les données volumineuses.
6. **Sécuriser les bases de données** et implémenter des procédures de sauvegarde et de restauration.
7. **Développer une application web sécurisée** pour visualiser et interagir avec les données.

Tâches à réaliser

1. Sélection et récupération des jeux de données météorologiques structurées et non structurées depuis le site de la NOAA
    - Jeux de données structurées attendus :
        - NOAA Global Surface Summary of the Day (GSOD)
            - **Lien :** https://www.ncei.noaa.gov/data/global-summary-of-the-day/access

        - **Description :** Ce jeu de données fournit des relevés quotidiens de stations météorologiques du monde entier, incluant la température, les précipitations, la pression atmosphérique, le vent, etc.
        - **Taille estimée :** Environ **20 Go** pour plusieurs années de données. Pour gérer les ressources, vous pouvez sélectionner une période plus restreinte (par exemple, les 5 dernières années) ou une zone géographique spécifique.
    - NOAA Integrated Surface Database (ISD)
        - **Lien :** https://www.ncei.noaa.gov/products/land-based-station/integrated-surface-database
        - **Description :** L'ISD est une base de données mondiale des observations météorologiques horaires provenant de plus de 35 000 stations.
        - **Taille estimée :** Environ **20 Go** pour plusieurs années de données. Pour gérer les ressources, vous pouvez sélectionner une période plus restreinte (par exemple, les 5 dernières années) ou une zone géographique spécifique.
- Jeux de données non structurées attendus :
    - NOAA Storm Events Database
        - **Lien :** https://www.ncei.noaa.gov/pub/data/swdi/stormevents/csvfiles/
        - **Description :** Cette base de données contient des rapports textuels détaillés sur les événements météorologiques extrêmes aux États-Unis, tels que les tempêtes, les ouragans, les tornades, les inondations, etc.
        - **Taille estimée :** Les fichiers annuels varient en taille. En sélectionnant plusieurs années, vous pouvez obtenir environ **1 à 2 Go** de données.
    - NOAA National Centers for Environmental Information (NCEI) Climate Data Online Text Products
        - **Lien :** https://tgftp.nws.noaa.gov/data/observations/metar/stations/
        - **Description :** Fournit des données textuelles telles que les bulletins climatiques, les rapports de tempêtes, les observations synoptiques, etc.
- Méthode de récupération :
    - **Téléchargement direct :** Téléchargez les données via les liens FTP ou HTTP fournis sur le site de la NOAA.
    - **Scripts d'automatisation :** Écrivez des scripts en Python ou Bash pour automatiser le téléchargement et la préparation des données.
    - **Gestion des volumes de données :** En raison des limitations de ressources, sélectionnez des périodes spécifiques ou des zones géographiques limitées pour maintenir la taille totale des données à environ **10 Go**, vous pouvez ajuster cette taille en fonction de vos conditions particulières de connexion et de matériel, en revanche vous devez absolument avoir des données structurées et non structurées.

2. **Installation et configuration de l'écosystème Hadoop dans des conteneurs Docker** *(Compétence 6.2)*
    - **Déployer Hadoop :** Utilisez Docker pour installer Hadoop 3.x, en configurant HDFS pour le stockage distribué et YARN pour la gestion des ressources.
    - **Configuration du cluster :** Créez un cluster Hadoop avec un NameNode et deux DataNodes, chacun dans un conteneur séparé.
    - **Allocation des ressources :** Configurez les conteneurs Docker pour utiliser une quantité raisonnable de RAM et de CPU (par exemple, 2 Go de RAM et 2 CPU par conteneur).
3. **Conception et déploiement d'un entrepôt de données structurées et non structurées** *(Compétence 6.3)*
    - Utiliser Hive pour les données structurées :
        - **Tables Hive :** Créez des tables pour les données GSOD ou ISD, en définissant les schémas correspondant aux fichiers téléchargés.
        - **Partitionnement :** Optimisez les tables en les partitionnant par année, par mois et par station pour améliorer les performances des requêtes.
    - Utiliser HBase pour les données non structurées :
        - **Stockage des rapports textuels :** Importez les rapports de la Storm Events Database dans HBase.
        - **Organisation des données :** Utilisez des colonnes familiales pour structurer les données (par exemple, "métadonnées", "description", "localisation").
    - Importation des données :
        - **Scripts d'ingestion :** Utilisez des outils comme Apache Pig, Apache Spark ou des scripts MapReduce pour importer et transformer les données dans Hive et HBase.
        - **Vérification de l'intégrité :** Effectuez des requêtes de test pour vous assurer que les données sont correctement importées.
4. **Définition de l'architecture des données** *(Compétence 6.4)*
    - Schéma architectural détaillé :
        - **Diagrammes :** Utilisez des outils comme Draw.io ou Lucidchart pour dessiner des diagrammes représentant les composants du système et les flux de données entre eux.
        - **Flux de données :** Décrivez comment les données sont collectées, stockées, traitées et présentées à l'utilisateur final.
    - Justification des choix technologiques :
        - **Documentation :** Expliquez pourquoi vous avez choisi Hive pour les données structurées, HBase pour les données non structurées, et ElasticSearch pour les fonctionnalités de recherche avancée.
5. **Mise en place d'ElasticSearch pour la recherche étendue** *(Compétence 6.6)*
    - Installer ElasticSearch :
        - **Version :** Utilisez ElasticSearch 7.x pour bénéficier des dernières fonctionnalités.
        - **Docker :** Déployez ElasticSearch dans un conteneur Docker, en allouant suffisamment de mémoire (par exemple, 2 Go de RAM).
    - Indexer les données :

        - **Connecteur Hadoop-Elasticsearch :** Utilisez le connecteur Elasticsearch-Hadoop pour indexer les données de Hive et HBase.
        - **Mapping personnalisé :** Créez des mappings pour définir la structure des index, en spécifiant les types de données et les analyseurs pour le texte.
    - Requêtes de recherche avancées :
        - **Requêtes sur les rapports de tempêtes :** Recherchez des événements spécifiques, des zones géographiques, ou des types de phénomènes météorologiques.
        - **Agrégations :** Utilisez les agrégations d'ElasticSearch pour obtenir des statistiques sur les données (par exemple, nombre d'événements par type et par année).
6. **Développement de requêtes SQL et NoSQL pour traiter des données volumineuses** *(Compétence 6.9)*
    - Requêtes HiveQL :
        - **Analyses statistiques :** Calculer des moyennes, des maxima, des minima de températures, précipitations, etc., par région et par période.
        - **Tendances climatiques :** Identifiez les tendances sur plusieurs années, comme le réchauffement ou le refroidissement dans certaines zones.
    - Requêtes HBase :
        - **Accès aux données non structurées :** Écrivez des scripts pour extraire des rapports contenant certains mots-clés (par exemple, "tornado", "flood").
        - **Jointures avec Hive :** Combinez les données structurées et non structurées pour des analyses plus approfondies.
7. **Sécurisation des bases de données et mise en place de procédures de sauvegarde et de restauration** *(Compétence 6.10)*
    - Authentification et autorisations :
        - **Kerberos :** Intégrez Kerberos pour sécuriser l'accès à Hadoop, Hive et HBase.
        - **Permissions :** Configurez les permissions pour contrôler l'accès aux données sensibles.
    - Chiffrement des données :
        - **Données au repos :** Activez le chiffrement HDFS Transparent Data Encryption (TDE) pour chiffrer les données stockées.
        - **Données en transit :** Configurez SSL/TLS pour chiffrer les communications entre les services (optionnel selon vos possibilités, si pas réalisé à ajouter au dossier).
    - Sauvegarde et restauration :
        - **Scripts de sauvegarde :** Créez des scripts pour sauvegarder régulièrement les données de HDFS, Hive et HBase vers un stockage externe (par exemple, un dossier local simulant un stockage cloud).
        - **Procédures de restauration :** Documentez les étapes pour restaurer les données à partir des sauvegardes en cas de perte ou de corruption.
8. **Développement d'une application web sécurisée** *(Compétences 7.2 à 7.5)*
    - Backend sécurisé :

        - **Framework :** Utilisez Flask ou Fastapi (Python) ou Express.js (Node.js) pour le backend.
        - **Authentification :** Implémentez JWT (JSON Web Tokens) pour l'authentification des utilisateurs.
        - **Endpoints :** Créez des API pour récupérer les données depuis ElasticSearch et Hive/HBase.
    - Frontend interactif :
        - **Framework :** Utilisez React.js ou Vue ou Angular pour le frontend.
        - **Visualisations :** Intégrez des bibliothèques comme D3.js ou Chart.js pour afficher des graphiques interactifs des données météorologiques.
    - Sécurisation du code :
        - **Protection contre les injections :** Validez toutes les entrées utilisateur et utilisez des requêtes préparées.
        - **Protection contre les attaques XSS et CSRF :** Implémentez des mesures de sécurité standard dans le frontend.
    - Optimisation et contre-mesures :
        - **Rate limiting :** Limitez le nombre de requêtes pour prévenir les abus.
        - **Logs et alertes :** Mettez en place un système de logs pour surveiller les activités suspectes et générer des alertes.
    - Étude des menaces et solutions :
        - **Analyse des vulnérabilités :** Utilisez des outils comme OWASP ZAP pour scanner l'application web.
        - **Documentation :** Rédigez un rapport listant les menaces identifiées et les mesures prises pour les contrer.

Contraintes

- Vous travaillerez **seul sur ce projet**
- **Utilisation de Docker :** Tous les composants doivent être déployés dans des conteneurs Docker séparés.
- **Limitation des ressources :** Optimisez la configuration des conteneurs pour fonctionner avec des ressources limitées (par exemple, un total de 8 Go de RAM pour tous les conteneurs).
- **Taille des jeux de données :** Les jeux de données sélectionnés doivent totaliser environ **10 Go** pour être gérables tout en restant représentatifs de données volumineuses.
- **Documentation complète :** Chaque étape doit être clairement documentée avec les commandes utilisées, les scripts, les configurations et les justifications techniques.

Livrables

Votre rendu final prendra la forme **d'une archive Zip dans votre dossier personnel de Drive de l'École** (si vous ne parvenez pas à trouver ce lien merci de contacter la Coordinatrice Pédagogique) et devra comporter les éléments suivants :

1. **Rapport écrit détaillé :**
    - **Introduction :** Présentation du projet, des objectifs et des enjeux.

- **Sélection des jeux de données :** Description des datasets NOAA choisis, leur taille, et la méthode de récupération.
- **Architecture du système :** Schémas et explications détaillées.
- **Installation et configuration :** Étapes pour chaque composant, avec les commandes et les configurations.
- **Requêtes développées :** Présentation et explication des requêtes SQL et NoSQL.
- **Sécurité :** Détails sur les mesures de sécurité mises en place, y compris les procédures de sauvegarde et de restauration.
- **Application web :** Description des fonctionnalités, de l'interface utilisateur et des aspects de sécurité.
- **Analyse des menaces :** Liste des vulnérabilités potentielles et des solutions appliquées.
- **Conclusion :** Bilan du projet, défis rencontrés et perspectives d'amélioration.

2. **Code source et scripts :**
   - **Dockerfiles et fichiers de configuration Docker Compose** pour chaque service.
   - **Scripts de téléchargement et d'ingestion des données NOAA**.
   - **Requêtes HiveQL et scripts HBase**.
   - **Code de l'application web** (backend et frontend), avec les instructions pour l'installation et l'exécution.
   - **Scripts de sauvegarde et de restauration**.
3. **Guide d'utilisation :**
   - **Instructions de déploiement :** Étapes pour lancer l'environnement complet sur une machine locale.
   - **Utilisation de l'application web :** Guide pour naviguer dans l'application, exécuter des analyses et interpréter les résultats.
   - **Procédures de maintenance :** Comment mettre à jour les données, gérer les conteneurs Docker, etc.
4. **Présentation orale :**
   - **Slides :** Résumé visuel du projet, incluant les objectifs, l'architecture, les réalisations, les défis et les solutions.
   - **Démonstration en direct :** Présentation fonctionnelle de l'application web et des principales fonctionnalités du système.

Vous devez livrer une archive de votre livrable avec l’ensemble des éléments.

Cette archive aura comme titre vos nom et prénom suivi de votre classe.

Exemple: projet#3_groupe_A_DATA01.zip

Evaluations

- Sélection et récupération des jeux de données NOAA
- Installation et configuration de Hadoop avec Docker *(compétence 6.2)*
- Conception de l'entrepôt de données *(compétence 6.3)*
- Définition de l'architecture des données *(compétence 6.4)*
- Mise en place d'ElasticSearch *(compétence 6.6)*
- Développement des requêtes SQL et NoSQL *(compétence 6.9)*
- Sécurisation des bases de données et procédures de sauvegarde *(compétence 6.10)*
- Développement de l'application web sécurisée *(compétences 7.2 à 7.5)*
- Qualité de la documentation et de la présentation

- **Modalité d'évaluation** : Préparation et présentation individuelle.
- **Type** : Jury (travaux à déposer la veille du jury). 30 minutes de présentation et 10 minutes de questions des jurés.
- **Date du jury** : mardi 3 et mercredi 4 décembre 2024, à distance sur Teams

Conseils

- Bien prendre le temps d'analyser le brief et comprendre le client
- Organisez-vous et planifiez votre travail : donnez vous des objectifs intermédiaires
- Réalisez la partie conception avant de commencer l'implémentation
- Faites directement des documents de conception présentables
- Ne jamais être trop ambitieux
- Complétez les documents à rendre au fur et à mesure
- Mettez en oeuvre les bonnes pratiques vues en cours
- Refactoriser pour éviter le code redondant
- Soignez la qualité de votre code (commentaires, indentation, nommage)
- Pensez à la qualité du résultat !

Instructions supplémentaires

- **Gestion des données NOAA :**
  - **Téléchargement des données structurées :**
    - **Global Surface Summary of the Day (GSOD) :**
      - Les données sont disponibles au format CSV compressé (.tar.gz).
      - Vous pouvez télécharger les données par année via FTP ou HTTP.

Exemple de lien pour l'année 2020 :
https://www.ncei.noaa.gov/data/global-summary-of-the-day/access/2020/

Script de téléchargement automatisé (exemple) :

```
# Télécharger les données GSOD de 2015 à 2020
for year in {2015..2020}; do
    wget -r -np -nd -A '*.csv'
"https://www.ncei.noaa.gov/data/global-summary-of-the-day/access/$year/" -P ./data/gsod/$year/
done
```

  - **Téléchargement des données non structurées :**
    - **NOAA Storm Events Database :**
      - Les données sont disponibles au format CSV et incluent des descriptions textuelles détaillées.
      - Les fichiers peuvent être téléchargés par année.

Exemple de lien pour l'année 2020 :
https://www1.ncdc.noaa.gov/pub/data/swdi/stormevents/csvfiles/

Script de téléchargement automatisé (exemple) :

```
# Télécharger les rapports de tempêtes détaillés de 2015 à 2020
for year in {2015..2020}; do
    wget
"https://www1.ncdc.noaa.gov/pub/data/swdi/stormevents/csvfiles/StormEvents_details-ftp_v1.0_
${year}.csv.gz" -P ./data/storm_events/
done
```

- **Importation dans HBase :**
  - **Prétraitement des données :** Les données textuelles peuvent nécessiter un prétraitement pour extraire les champs pertinents.
  - **Utilisation de MapReduce ou Apache Spark :** Pour transformer et charger les données dans HBase.

- **Analyse du texte :**
    - **Traitement du langage naturel (NLP) :**
        - Utilisez des bibliothèques comme NLTK ou spaCy pour analyser les descriptions d'événements.
        - Extraire des informations telles que les types d'événements, les localisations, les impacts.
- **Sécurité avancée :**
    - **Gestion des identités et des accès (IAM) :**
        - Créez des utilisateurs et des groupes avec des permissions spécifiques.
        - Simulez différents rôles (administrateur, analyste, utilisateur standard).
    - **Monitoring et alertes :**
        - Configurez des outils de monitoring pour surveiller les performances et la sécurité.
        - Collectez les logs et analysez-les pour détecter des anomalies.
- **Documentation :**
    - **Inclure des diagrammes UML :** Pour représenter l'architecture logicielle et les interactions.
    - **Fournir un glossaire :** Des termes techniques utilisés dans le projet.

Remarques finales :

- **Respect des licences :**
    - Les données de la NOAA sont généralement publiques et libres d'utilisation, mais vérifiez les conditions spécifiques sur le site.
- **Gestion des ressources :**
    - Surveillez l'utilisation des ressources pour éviter de saturer la machine hôte.
    - Nettoyez régulièrement les conteneurs et images Docker inutilisés.
- **Documentation claire :**
    - Commentez votre code et vos scripts.
    - Incluez des captures d'écran pour illustrer les étapes importantes.
- **Collaboration :**
    - Si vous travaillez en groupe, répartissez les tâches et documentez la contribution de chacun.
    - Utilisez un système de contrôle de version comme Git.

Bon courage pour ce projet enrichissant qui vous permettra de développer vos compétences en Big Data, en traitement de données structurées et non structurées, en sécurité informatique et en développement web, en utilisant exclusivement les données du site NOAA !