Document de Projet : Architecture de Big Data Lake

Pour l'Analyse Climatique et des Risques Naturels

Introduction

Ce document présente le projet d'architecture de Big Data Lake spécialement conçu pour l'analyse des données climatiques et des risques naturels. L'objectif principal était de développer une plateforme scalable et fiable, capable d'ingérer, stocker, traiter et visualiser des volumes massifs de données hétérogènes provenant de sources scientifiques et opérationnelles clés telles que la NOAA (National Oceanic and Atmospheric Administration) et l'USGS (United States Geological Survey). Ce projet répond aux enjeux croissants de centralisation des données environnementales, de prévision des risques et d'aide à la décision dans un contexte de changement climatique.

Architecture du Projet

Architecture globale

Notre solution repose sur une architecture en couches conçue pour supporter des flux de données variés (batch, streaming, API) et des cas d'usage scientifiques et métiers :

1. **Couche d'ingestion :** Collecte de données depuis NOAA, USGS, et autres sources (satellites, capteurs IoT, données ouvertes).
2. **Couche de stockage :** Data Lake scalable basé sur un système de stockage distribué.
3. **Couche de traitement :** Moteurs de traitement batch et streaming pour l'analyse et l'enrichissement.
4. **Couche de service :** APIs d'accès aux données et services métier.
5. **Couche de visualisation** : Tableaux de bord interactifs dédiés aux indicateurs climatiques et de risques.

Composants techniques

- **Stockage :** Hadoop HDFS / AWS S3 pour les données NOAA (images satellitaires, données météo) et USGS (sismiques, hydrologiques).
- **Orchestration :** Apache Airflow pour l'automatisation des pipelines d'ingestion réguliers.
- **Traitement batch :** Apache Spark pour le nettoyage, l'agrégation et l'analyse des séries temporelles.
- **Traitement streaming :** Apache Kafka + Spark Streaming pour la surveillance en temps réel des flux USGS (sismicité) et NOAA (alertes météo).

- **Catalog :** Apache Hive Metastore pour le catalogage des jeux de données climatiques.
- **Sécurité** : Kerberos + Apache Ranger pour la gestion des accès aux données sensibles.

Ce que nous avons réalisé

Phase 1 : Conception et planification

- Analyse des besoins métier : prévision des risques naturels, analyse climatique historique, recherche scientifique.
- Identification des sources principales : NOAA (données atmosphériques, océaniques) et USGS (données géologiques, hydrologiques).
- Design des zones de données : Raw (données brutes NOAA/USGS), Curated (données nettoyées), Analytics (indicateurs climatiques), Sandbox (expérimentation ML).

Phase 2 : Infrastructure de base

- Configuration du stockage distribué avec une structure organisée par source (dossiers /noaa/, /usgs/).
- Mise en place des pipelines d'ingestion automatisés pour les datasets NOAA (ex : GSOD, NEXRAD) et USGS (ex : flux sismiques en temps réel).
- Implémentation des mécanismes de sécurité et de gouvernance.

Phase 3 : Traitement des données

- Développement de pipelines ETL/ELT pour :
    - Nettoyer et formater les données NOAA (températures, précipitations, vents).
    - Enrichir les données USGS avec des référentiels géographiques.
    - Calculer des indicateurs agrégés : moyennes mensuelles, écarts à la normale, indices de sécheresse.
- Mise en place de la traçabilité des données (data lineage) pour garantir l'auditabilité des analyses.

Les Agrégations

Types d'agrégations implémentées

- Agrégations temporelles : Données NOAA regroupées par heure, jour, mois, saison.
- Agrégations spatiales : Données USGS agrégées par région géographique, bassin versant.

- Agrégations métier : Calcul d'indices climatiques (ex : indice de danger incendie, risque inondation).
- Agrégations en temps réel : Surveillance des flux sismiques USGS pour détection d'événements anormaux.

Technologies utilisées

- Spark SQL pour les agrégations complexes sur les séries temporelles NOAA.
- Presto/Trino pour les requêtes ad-hoc croisant données USGS et référentiels externes.
- Matérielisation des vues pour accélérer les requêtes sur les jeux de données fréquemment interrogés.

Corrélation et Analytics

Analyse des corrélations

- Calcul de corrélations entre variables climatiques NOAA (température, humidité) et activité sismique USGS.
- Analyse de séries temporelles pour détecter des patterns saisonniers ou des tendances longues.
- Utilisation d'algorithmes de clustering pour regrouper des régions aux profils climatiques similaires.

Outils déployés

- MLlib pour l'apprentissage automatique distribué (prédiction de précipitations à partir des données NOAA).
- Python avec Pandas/NumPy pour l'analyse exploratoire dans des notebooks Jupyter.
- Apache Superset pour la visualisation avancée des corrélations.

Backend

Services principaux

- API Gateway : Point d'entrée unique pour interroger les données NOAA/USGS.
- Service d'accès aux données : APIs RESTful permettant de récupérer des séries temporelles, des métadonnées, des indicateurs calculés.
- Service de métadonnées : Gestion du catalogage des jeux de données (source, date, qualité).
- Service de monitoring : Surveillance de l'état des pipelines et de la fraîcheur des données.

Technologies backend

- Langages : Python (FastAPI), Scala (Spark).
- Base de données : PostgreSQL pour stocker les métadonnées et les résultats d'agrégation.
- Authentification : OAuth2/JWT pour sécuriser l'accès aux données sensibles.

Frontend

Application de visualisation

- Dashboards interactifs :
    - Carte des risques inondation (données USGS hydrologiques + précipitations NOAA).
    - Graphiques d'évolution des températures (NOAA) par région.
    - Visualisation des séismes récents (flux USGS) superposée aux infrastructures critiques.
- Recherche de datasets : Interface permettant de découvrir les jeux de données NOAA/USGS disponibles.
- Éditeur de requêtes : Interface SQL avec autocomplétion pour interroger le Data Lake.

Stack technique

- Framework : React.js avec TypeScript.
- Visualisation : D3.js pour les cartes choroplèthes, Apache ECharts pour les séries temporelles.
- État : Redux pour la gestion des filtres métier (période, région, type de risque).

Gestion des données

Zones de données

- Zone Raw : Données brutes NOAA (fichiers CSV, NetCDF) et USGS (flux JSON, XML).
- Zone Curated : Données nettoyées, standardisées, prêtes pour l'analyse.
- Zone Analytics : Indicateurs calculés, agrégats, modèles ML déployés.
- Zone Sandbox : Espace d'expérimentation pour la recherche et le développement de nouveaux indicateurs.

Gouvernance

- Lineage : Traçabilité complète depuis la source NOAA/USGS jusqu'au dashboard.

- Qualité : Métriques sur la complétude, la fraîcheur et la cohérence des données.
- Accès : RBAC (Role-Based Access Control) pour différencier accès recherche, opérationnel, public.
- Conformité : Gestion des données sensibles (localisations précises, données critiques) conformément aux réglementations.

Conclusion

Le projet d'architecture de Big Data Lake que nous avons conçu est spécialement adapté aux enjeux climatiques et des risques naturels. En intégrant des sources scientifiques majeures comme NOAA et USGS, nous avons construit une plateforme qui permet :

- ✅ Centralisation : Une source unique de vérité pour les données environnementales.
- ✅ Agilité : Analyses exploratoires rapides et calculs d'indicateurs en quasi-temps réel.
- ✅ Innovation : Support des cas d'usage avancés (ML, streaming) pour la prévision et la gestion de crise.
- ✅ Gouvernance : Traçabilité, sécurité et conformité intégrées.

Bénéfices principaux :

- Réduction des délais d'accès aux données climatiques et sismiques.
- Amélioration de la précision des modèles de prévision des risques.
- Création d'une base solide pour la recherche scientifique et l'aide à la décision publique.
- Évolutivité garantie face à l'augmentation du volume et de la variété des données environnementales.

Cette architecture constitue une fondation robuste et adaptable pour tout organisme œuvrant dans les domaines du climat, de l'environnement ou de la gestion des risques naturels.