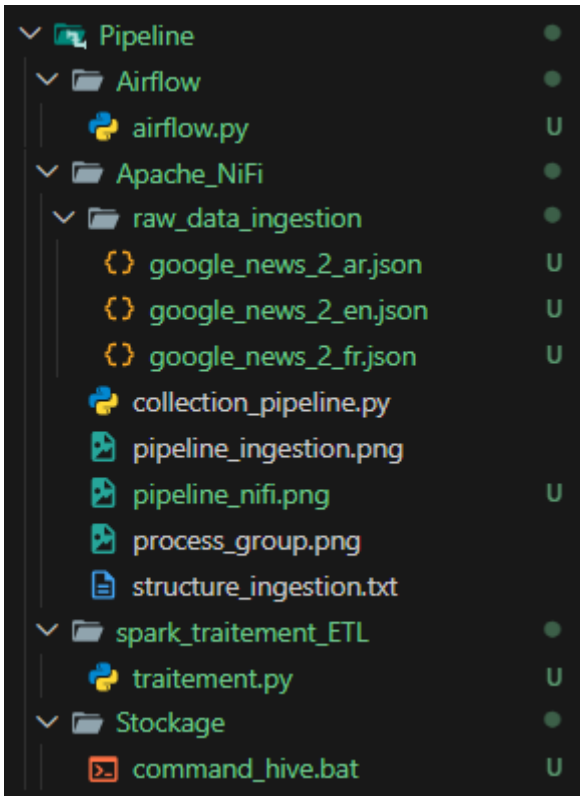
Etape 7 : Pipeline de Données Multilingues avec NiFi, Hive, Spark et Airflow

1. Introduction

Ce projet vise à orchestrer un pipeline de traitement des données multilingues à travers une série d'étapes allant de l'ingestion des données jusqu'à leur transformation et leur traitement en utilisant plusieurs outils de l'écosystème Apache. Le pipeline inclut les étapes suivantes :

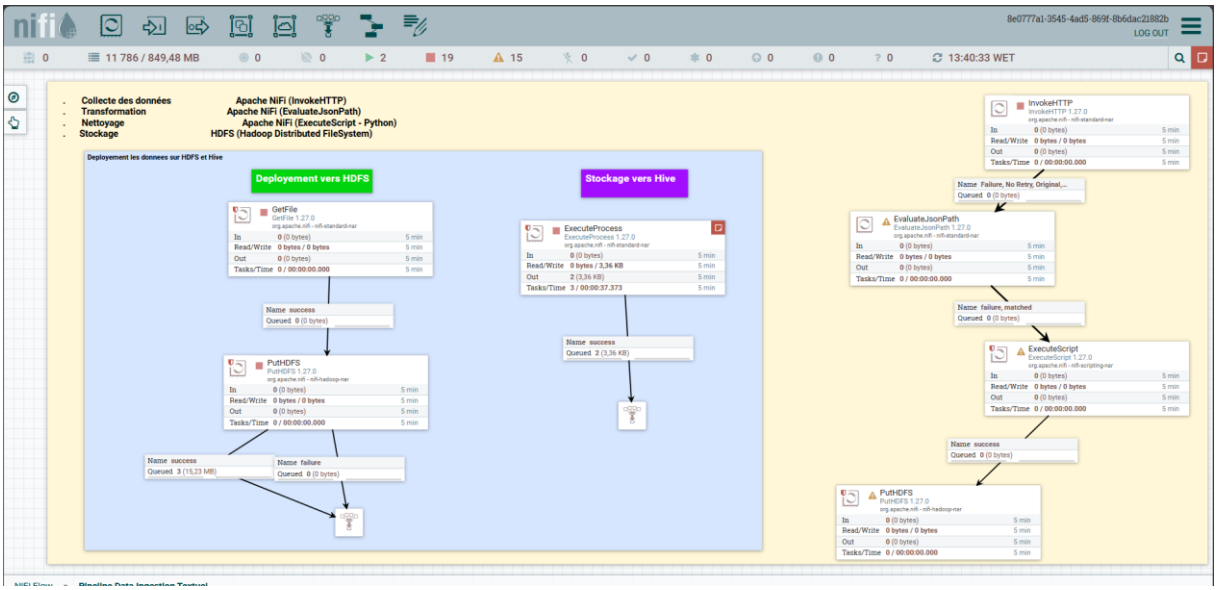
1. **Ingestion des données avec Apache NiFi** : Les données JSON sont collectées depuis une source externe et transférées dans HDFS (Hadoop Distributed File System).
2. **Stockage et gestion des données dans Apache Hive** : Les données sont ensuite transférées dans une table Hive pour leur traitement ultérieur.
3. **Nettoyage et transformation des données avec Apache Spark** : Le script Spark effectue des transformations nécessaires sur les données stockées dans Hive.
4. **Orchestration et automatisation avec Apache Airflow** : Un DAG Apache Airflow est utilisé pour orchestrer le pipeline, automatiser le processus et assurer un traitement parallèle des données multilingues.



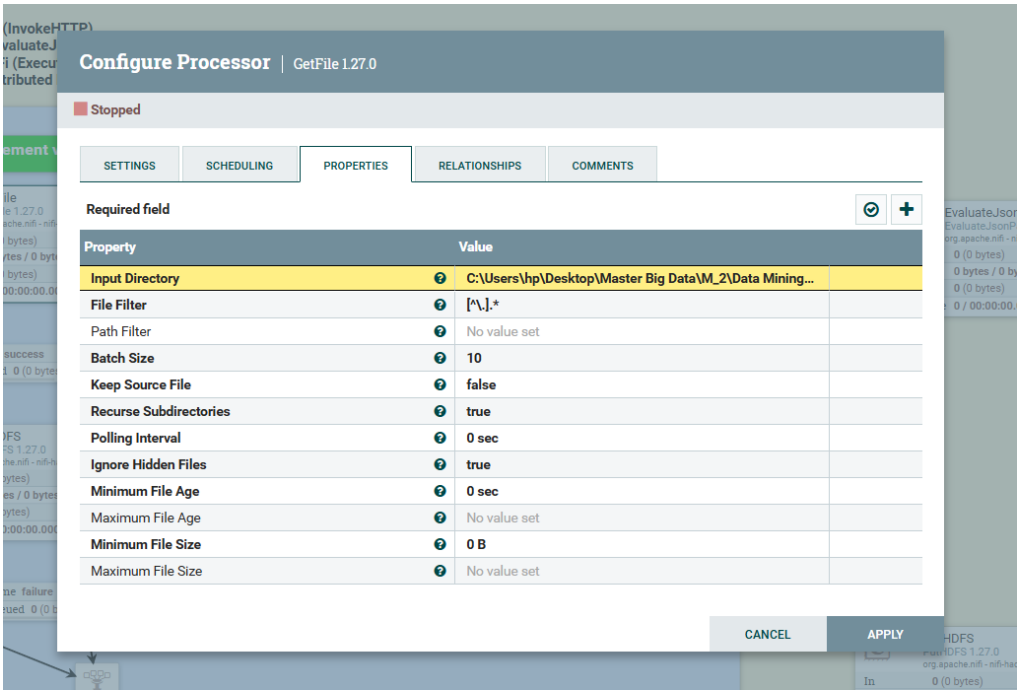
2. Architecture du Pipeline de Données

2.1. Apache NiFi

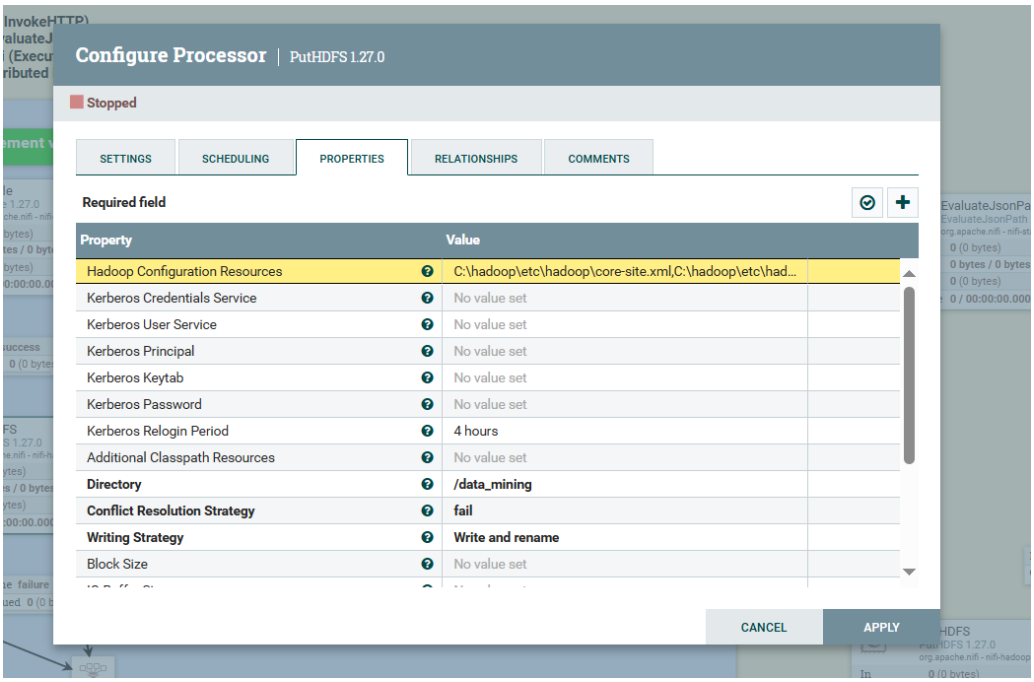
Apache NiFi est utilisé pour l'ingestion des données JSON. Les données sont collectées via un processeur InvokeHTTP qui récupère les fichiers JSON depuis une source externe et les place dans un répertoire d'entrée pour un traitement ultérieur. Ensuite, les processeurs comme PutHDFS et PutHiveQL sont utilisés pour déplacer les données dans HDFS et Hive respectivement.



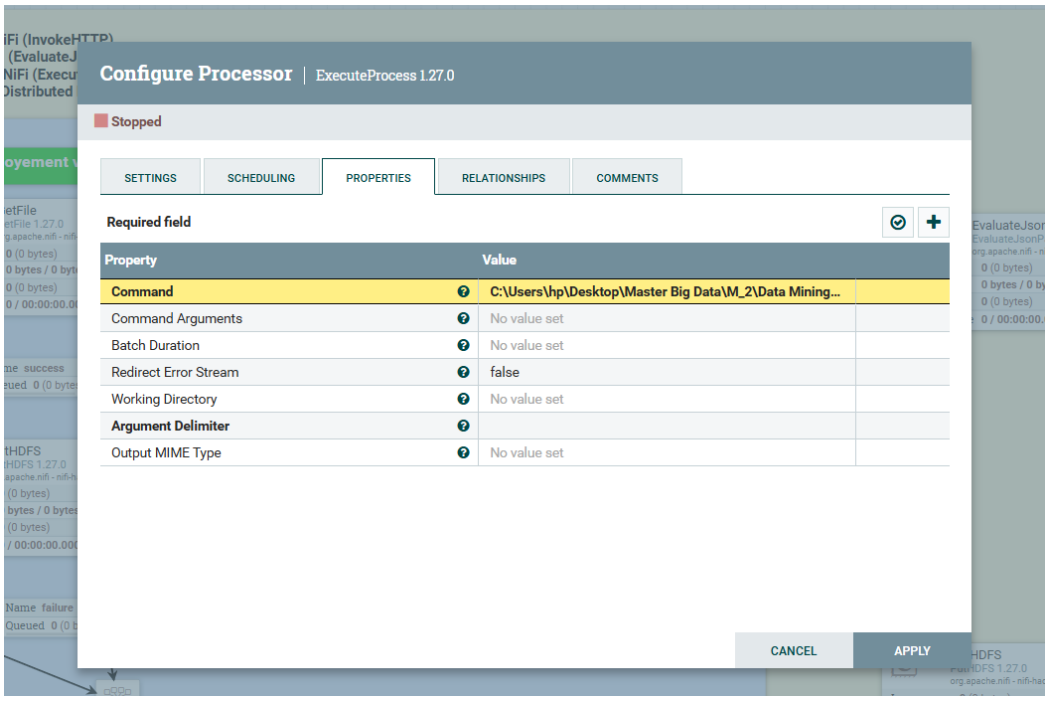
Processus GetFile



Processus PutHDFS



Processus Exécute Process



2.2. Apache Hive

Apache Hive est utilisé pour stocker et interroger les données après leur transfert depuis NiFi. Hive permet de stocker les données dans des tables adaptées aux formats de données complexes comme JSON. La table data_mining.news_fr, news_ar, et news_en contient les titres des actualités pour chaque langue respective.

Structure de la Table Hive

Les tables dans Hive ont les colonnes suivantes :

- **id** (entier, auto-incrémenté) : Identifiant unique pour chaque enregistrement.
- **title** (chaîne de caractères) : Titre de l'actualité.

2.3. Apache Spark

Apache Spark est utilisé pour le traitement des données et leur nettoyage. Le script cleaning_script.py prend en entrée les données stockées dans Hive, les nettoie et les transforme (par exemple, gestion des valeurs manquantes, mise en forme des dates, etc.).

Le script est déclenché via spark-submit avec un argument de langue (--lang) pour traiter les données des tables news_fr, news_ar, et news_en.

2.4. Apache Airflow

Apache Airflow est utilisé pour orchestrer l'ensemble du pipeline de données. Le DAG Airflow comprend plusieurs tâches :

- **Validation des données dans Hive** : Avant de lancer le nettoyage avec Spark, Airflow vérifie que les données existent et sont accessibles dans les tables Hive pour chaque langue (news_fr, news_ar, news_en).
- **Exécution du script Spark** : Pour chaque langue, le script cleaning_script.py est exécuté avec le paramètre --lang, ce qui permet de traiter les données de manière spécifique pour chaque table.
- **Orchestration des dépendances** : Le DAG est conçu pour exécuter les tâches en parallèle pour chaque langue tout en respectant les dépendances nécessaires (validation avant exécution du script).

2.5. Workflow de Traitement

1. **NiFi** récupère les fichiers JSON de la source et les charge dans HDFS.
2. **Hive** reçoit les données depuis NiFi et les stocke dans les tables correspondantes (ex. news_fr, news_ar, news_en).
3. **Airflow** valide les données dans Hive, puis déclenche le script Spark pour chaque langue.
4. **Spark** nettoie et transforme les données selon les règles définies (comme le nettoyage des colonnes title).
5. **Les résultats finaux** sont stockés de nouveau dans Hive ou dans une autre destination (par exemple, HDFS, ou une table différente dans Hive).

4. Conclusion

Ce pipeline de données multilingues repose sur l'intégration de plusieurs outils de l'écosystème Apache pour garantir une ingestion, une transformation et un stockage efficaces des données. Apache NiFi permet d'ingérer les données JSON, Apache Hive gère le stockage et les requêtes, Apache Spark nettoie et transforme les données, et Apache Airflow orchestre le tout, assurant une automatisation complète.

Ce système est flexible et peut être étendu pour intégrer d'autres langues ou ajouter d'autres étapes de traitement, ce qui en fait une solution robuste pour la gestion des données multilingues à grande échelle.