Vous êtes un ingénieur de données qui travaille dans un cabinet de consultance en Big Data pour le ministère de la santé.

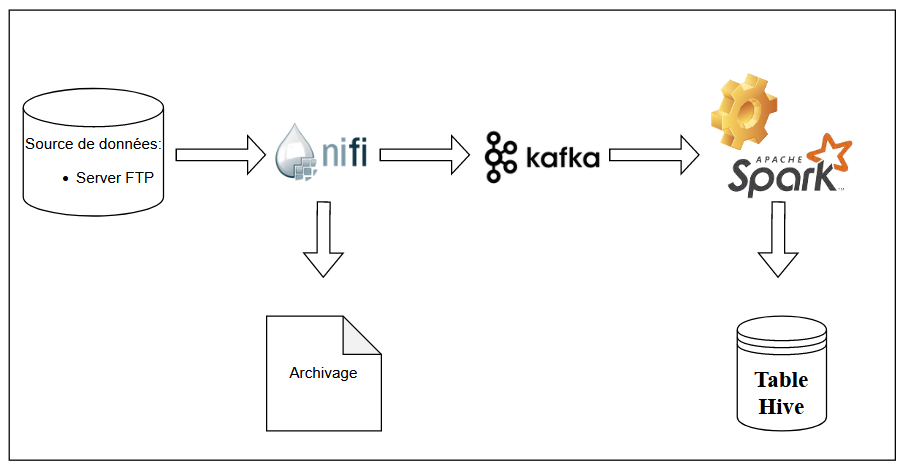
On vous confie la mission de gérer l’ingestion des données du Covid jusqu’à la mise en disposition de données prétraitées pour les agents du ministère à travers leur tableau de bord.

Le scénario est le suivant :

Les données du covid sont déposées momentanément dans un répertoire sur un serveur distant, sous forme de fichiers json contenant comme information les éléments contenus dans le tableau suivant

|  |  |  |
| --- | --- | --- |
| **champs** | **Description** | **Commentaire** |
| date | La date d’enregistrement | Date au format yyyy-MM-dd |
| variant | Le type de variant |  |
| nb\_cas | Le nombre de cas enregistré par rapport au variant |  |
| id\_pays | L’id du pays concerné |  |

Votre Techlead vous donne l’architecture ci-dessous avec les différentes étapes à suivre :



1. Étape 0 :
   1. Créer un topic kafka dénommé **ism2022\_m2\_votreNom\_covid**
   2. Créer dans HDFS un répertoire **/user/student\_ismX/m2\_bd\_2022/votreNom/examen/raw\_covid**
2. Étape 1 : (data orchestration avec Nifi)
   1. En utilisant vos identifiants, récupérer les fichiers se trouvant dans le répertoire **/home/ism\_teacher/ifall/examen\_m2/data/jsons\_covid** depuis le serveur **m10223.contaboserver.net** en utilisant les connecteurs **ListSftp**, **FetchSftp**
   2. Avec le connecteur RouteOnAttributes, vous pouvez rediriger un flow de  
      donnée dans le processeur suivant, en fonction de conditions
      1. Définir la condition sur la taille du fichier : pour ne récupérer que les fichiers dont la taille  
         dépasse 400Bytes
      2. Définir la condition sur le nom du fichier : pour ne récupérer que les fichiers ayant  
         l’extension ‘.json’
   3. Depuis le connecteur RouteOnAttributes, utiliser les connecteurs **MergeContent**, **UpdateAtttribute** et **putHDFS**, compresser les fichiers et mettre dans le répertoire Raw (données archivées pour un gain d’espace).
      1. **MergeContent :** choirir le format **zip**
      2. **UpdateAtttribute :** changer le nom du fichier compressé en utilisant l’expression ‘votreNom\_${now():format("yyyyMMddHH")}.zip’
      3. **putHDFS :** stocker les données zippées dans **/user/student\_ismX/m2\_bd\_2022/votreNom/examen/raw\_covid**
   4. Depuis le connecteur RouteOnAttributes (question b), utilisez le connecteur publishKafka pour publier le contenu des données dans le topic **ism2022\_m2\_votreNom\_covid**
3. Étape 2 : (data processing/storing avec Spark et Hive)
   1. En utilisant Spark Streaming
      1. Lisez les données contenues dans le topic **ism2022\_m2\_votreNom\_covid** tous les 15 minutes
      2. Faites la jointure avec le fichier **/user/ism\_teacher/ifall/examen\_m2/data/pays\_covid.csv** pour récupérer le libellé des pays.
      3. Filtrez en enlevant les variants de type **non\_who**
      4. Calculez le nombre de cas covid par année et par pays
      5. Écrivez les données de la question iv en mode append dans une table hive avec les caractéristiques suivantes :
         * Nom database : ism\_m2\_2022\_examspark
         * Nom table  : spark\_stream\_Votrenom
         * Chemin HDFS : /user/student\_ismX/m2\_bd\_2022/votreNom/examen/output/
         * Format : parquet