Projet Big Data

UCAD – M2BI
décembre 2022

Ce projet vise à présenter des problématiques réelles et diversifiées afin de vous permettre d'approfondir et consolider les connaissances acquises durant le parcours Big Data.

Le travail est individuel.

A rendre :

- le script .scala ou notebook (.dbc ou .ipynb)
- le lien git de votre projet
- + démonstration (code, tables hive, fichiers export)

Dates à retenir :
– travaux à rendre au plus tard le dimanche 08 janvier à 23h59.
– Premier groupe de présentation samedi 14 janvier à partir de 9h.
– Second groupe de présentation samedi 14 janvier à partir de 13h.

Pour faciliter la planification des présentations, chaque étudiant disposera de 20 minutes pour exposer son travail (installations, codes et résultats uniquement, les slides ne seront pas nécessaires).

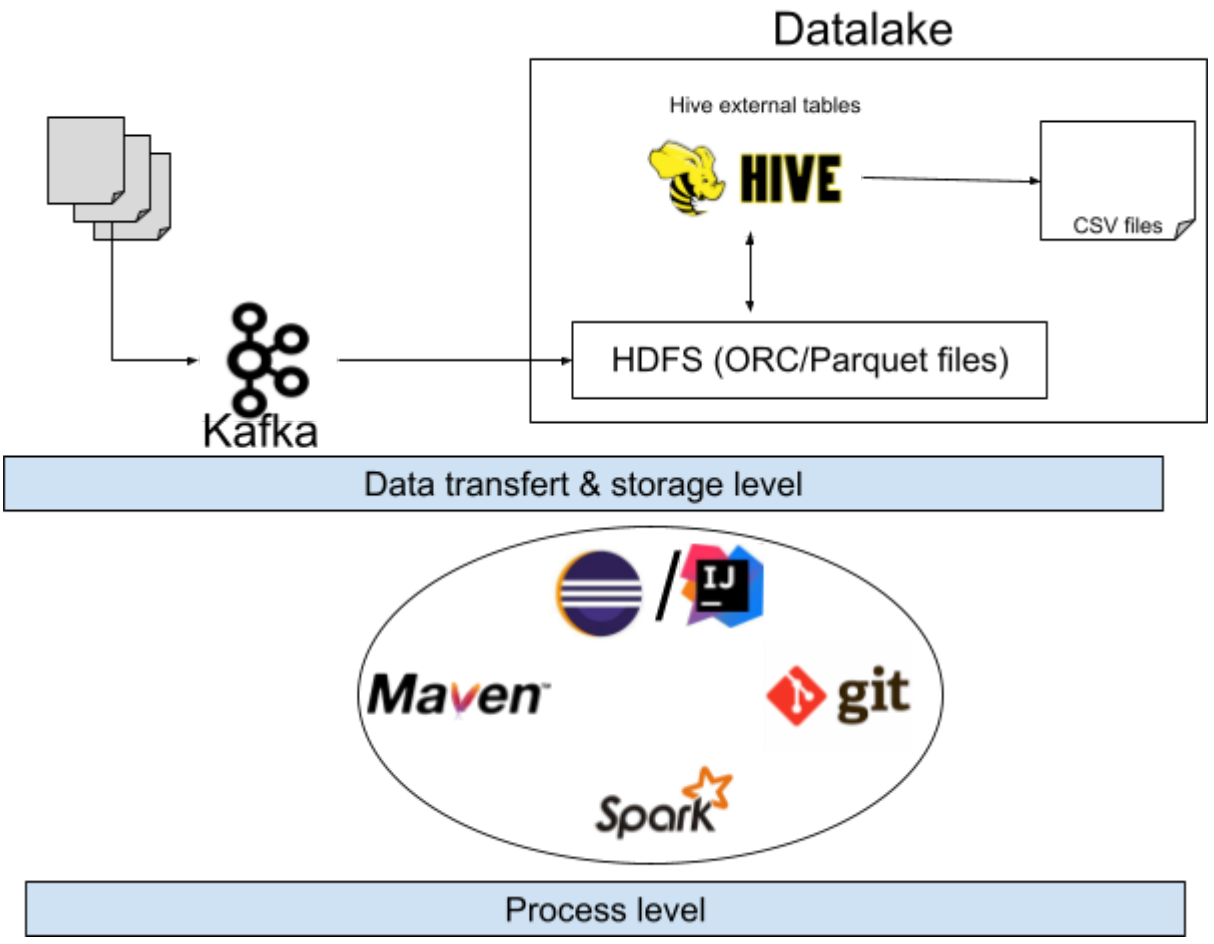
Le libellé ci-dessous est scindé en quatre (4) parties:

1. Sujet
2. Architecture
3. Cas d'usage
4. Livrable

1. Sujet

Vous travaillez pour une société de la place à dimension mondiale dénommée **GaindeBi** qui vient d'implanter un département (service) appelé Stratégie. Ce département représente la vision futuriste de **GaindeBI** et renferme en son sein une filière Big Data à laquelle vous êtes rattaché.
Pour cela, vous êtes chargé de déblayer le terrain pour la création et l'implémentation des processus et concepts pour la filière Big Data. Il vous est demandé de concevoir une plateforme générique pour la collecte, l'ingestion dans le Datalake, le traitement et la restitution (architecture ci-dessous).

2. Architecture à mettre en place



- Utiliser Intellij/Eclipse
- Utiliser Maven pour les librairies
- Utiliser Git pour le versioning de votre projet
- Fichiers chargés dans HDFS
- Données chargées dans des tables Hive qui vont pointer sur les fichiers hdfs (table interne ou externe)
- Resultat de requêtes exportés dans des fichiers Export.csv avec header et footer.

3. **Cas d'usage**

   La société **LightInvest** souhaite investir dans d'autres entreprises établies dans différents pays. **LightInvest** sous-traite avec votre société **GaindeBI** afin d'analyser les différents risques en utilisant les technologies Big Data. Les problématiques et les données seront fournies par **LightInvest** et **GaindeBI** (c'est-à-dire vous) proposera la plateforme de traitement et les requêtes.

   Problématiques et données

   Pour la compréhension des affaires et des données, la société **LightInvest** a deux contraintes mineures pour ses investissements: elle veut investir entre 4 et 16 millions d'euros par tour d'investissement et souhaite investir uniquement dans les pays anglophones en raison de la facilité de communication avec les entreprises dans lesquelles elle investirait.

   L'objectif est d'identifier les meilleurs secteurs et pays appropriés pour effectuer des investissements. La stratégie globale consiste à investir là où d'autres sociétés investissent (on suppose que les meilleurs secteurs et pays sont ceux où la plupart des investisseurs autres que LightInvest investissent).

   Pour votre analyse, considérez qu'un pays est anglophone si et seulement si l'anglais est l'une des langues officielles de ce pays. Vous pouvez utiliser la liste des pays anglophones dans le fichier `pays.csv`.

Partie 1: exploration, compréhension et nettoyage de la donnée

- **Question 0**: charger respectivement les fichiers societe.txt, cartographie.csv, tours.csv et pays.csv dans les dataframes societeDf, cartoDf, toursDf et paysDf et respectivement dans les tables Hive t_societe, t_carto, t_tours et param_pays.
- **Question 1**: quel est le nombre de sociétés différentes présentes dans le dataframe toursDf?
- **Question 2**: quel est le nombre de sociétés différentes présentes dans le dataframe societesDf?
- **Question 3**: dans le dataframe societesDf, quelle colonne peut être utilisé comme clé unique pour chaque société. Donnez le nom de la colonne.
- **Question 4**: y a-t-il des sociétés dans le dataframe toursDf qui ne sont pas présentes dans le dataframe societesDf ?

- **Question 5**: fusionner les deux dataframes afin que toutes les colonnes du dataframe societesDf soient ajoutées au dataframe toursDf. Nommez le nouveau dataframe obtenu mergedDf. Combien d'observations sont présentes dans mergedDf ?

Partie 2: analyse des pays

- **Question 1**: quels sont les sept premiers pays qui ont reçu l'investissement total le plus élevé (dans tous les secteurs pour le type d'investissement choisi).
- **Question 2**: pour le type d'investissement choisi, créez un dataframe nommé top7CountriesDf avec les sept premiers pays (en fonction du montant total d'investissement reçu par chaque pays).
- **Question 3**: identifiez les trois premiers pays anglophones dans le dataframe top7CountriesDf.

Partie 3: analyse des secteurs

- **Question 1**: extraire le secteur primaire de chaque liste de catégories de la colonne liste_categorie.
- **Question 2**: Utilisez le dataframe cartoDf (fichier cartographie.csv) pour mapper chaque secteur primaire à l'un des huit secteurs principaux (« autres » peut également être considéré comme étant un secteur principal). – **Question 3**: créez trois dataframe différents df1, df2 et df3 pour chacun des trois pays contenant les observations de type de financement FT comprises dans la fourchette 4-16 millions d'euros.
- **Question 4**: Les trois dataframe devraient contenir toutes les colonnes de mergedDf ainsi que le secteur principal et le secteur primaire.
- **Question 5**: À l'aide des trois dataframe, calculer le nombre total d'investissements et le montant total des investissements dans chaque secteur principal pour chacun des trois pays.
    - Pour le premier secteur en termes de comptage, quelle entreprise a reçu l'investissement le plus élevé ?
    - Pour le deuxième meilleur secteur en termes de comptage, quelle entreprise a reçu l'investissement le plus élevé ?

Partie 2: analyse du type d'investissement
calculez la valeur la plus représentative du montant de l'investissement pour chaque type d'investissement (colonne type_tour_investissement du datafarme toursDf).

4. **Livrables attendus**
    a. Architecture (ci-dessus à respecter).
    b. Code et requêtes
    c. Tables et fichiers d'export
    d. "Démo"