Atelier 4 : Pig

Exercice 1 :

Le jeu de données proposé est une base de films. Nous disposons de deux fichiers : La liste des films et la liste des artistes.

1. Copier les deux fichiers `'artists.json'` et `'movies.json'` sous HDFS.
2. Charger les deux fichiers en utilisant **JsonLoader**

```
movies = LOAD 'atelier4/movies.json'
  USING JsonLoader('id:chararray, title:chararray, year:chararray,
                    genre:chararray,country:chararray,
                    director: (id:chararray,lastName:chararray,
                              firstName:chararray,birthDate:chararray),
                    actors: {(id:chararray, role:chararray)}');
```

3. Afficher la liste des films américains par année. (Afficher l'année et les noms des films)
4. Afficher pour chaque metteur en scène la liste de ses films américains.
5. Afficher les triplets (idFilm, idActeur, role).
6. Afficher pour chaque film les descriptions complètes de ses acteurs.
7. Afficher pour chaque film américain le nombre de ses acteurs.

Exercice 2 :

1. Copier le fichier '`combined_access_log.txt`' sous HDFS.
2. Ecrire un script **log_script.pig** qui permet de :
   a. Charger le fichier en utilisant **org.apache.pig.piggybank.storage.apachelog.CombinedLogLoader**

```
REGISTER /usr/lib/pig/piggybank.jar;

logs = LOAD 'atelier4/combined_access_log.txt' USING
org.apache.pig.piggybank.storage.apachelog.CombinedLogLoader()
AS (addr: chararray, logname: chararray, user: chararray, time: chararray,
    method: chararray, uri: chararray, proto: chararray,
    status: int, bytes: int, referer: chararray, userAgent: chararray);
```

   b. Déterminer le nombre d'apparition de chaque adresse
   c. Afficher les sites les plus référencés et le nombre de références
   d. Déterminer les adresses dont il y a un problème d'autorisation d'accès (status =304)

Tous les résultats doivent être stockés sous le répertoire '/user/cloudera/out'