TP n3 en Big Data Analytics

Exercice 1.

1. Charger le fichier **breweries.csv** en utilisant l'API spark.read.csv

```
scala> val df = spark.read.csv("hdfs://namenode:9000/user/root/input/breweries.csv")
```

2. Editer le schéma du ficher : scala> df.show()
3. Lister les 10 premières lignes dans le fichier
4. Lister le contenu de la colonne city
5. Compter le nombre de données par city : scala> df.groupBy("cityr").count().show()

Exercice 2 :

1. Créer un fichier JSON intitulé employe.json dans le répertoire /root/input de votre spark-master, avec le contenu suivant:

```
{"id" : "1201", "name" : "ahmed", "age" : "25"}
{"id" : "1202", "name" : "salma", "age" : "58"}
{"id" : "1203", "name" : "amina", "age" : "39"}
{"id" : "1204", "name" : "ali", "age" : "23"}
{"id" : "1205", "name" : "mourad", "age" : "23"}
```

2. Charger le fichier json

```
val df = sqlcontext.read.json("/root/input/employe.json")
//val df=spark.read.format("csv").option("header","true").load("d:/tpspark/flights.csv")
```

3. Editer le schéma du ficher
4. Lister les 3 premières lignes dans le fichier
5. Lister les noms des employés
6. Compter le nombre des employés par age
7. Filter les données par âge (age>23)
8. Grouper les données par âge et compter le nombre de personnes pour chaque âge

Exercice 3 : Une entreprise d'activité e-commerce souhaite une routine qui doit compter tous les utilisateurs uniques qui ont visité la boutique en ligne, un jour donné. Elle veut calculer également la moyenne du nombre de produits ajoutés au panier de chaque client, tout en recherchant les dix premiers produits communs ajoutés à tous les paniers de la journée.

Il s'agit de créer une série de tâches pour lire les données d'événement relatives à l'activité de l'utilisateur sur un site de commerce électronique fictif, dans le but de proposer ce qui suit trois rapports quotidiens communs :

• Trouver les utilisateurs actifs quotidiens (ou utilisateurs uniques quotidiens).
• Calculer le nombre moyen quotidien d'articles dans tous les paniers d'utilisateurs.
• Générer les dix articles les plus ajoutés dans tous les paniers d'utilisateurs.

1. Créer un fichier csv intitulé activiteUser.csv dans le répertoire /root/input de votre spark-master, avec le contenu suivant:

```
userId, catId, itemId
"u1", "c1", "i1",
"u1", "c1", "i2",
"u2", "c2", "i1",
"u3", "c3", "i3",
"u4", "c4", "i3"
```

2. Charger le fichier

```
scala> val df=spark.read.format("csv").option("header","true").load("d:/tpspark/activiteUser.csv")
```

3. Lister les lignes dans le fichier
4. Trouver les utilisateurs actifs quotidiens pour un jour donné
5. Calculer le nombre moyen quotidien d'articles Dans tous les paniers d'utilisateurs
6. Générer les dix éléments les plus ajoutés dans tous les paniers d'utilisateurs.