Tasks

1. Clone the scala-spark-training project in:
   https://github.com/cpliyanage/scala-spark-training
   (java version 1.8, gradle latest)
2. Build it using 'gradle build'
3. Open in IntelliJ IDEA and make sure you enable creation of content roots when you import the project.

Scala Exercise 1

1. Create a feature branch
2. Using scala create a program to read a text file of clickstream data in a given path (in/clickstream.csv) and calculate some metrics (one line of input corresponds to one click event). The format of one line of the text file is;
   <userId>,<productCategory>,<productId>,<channel>

Ex:
user1,Clothing,product1,Webstore
user2,Kitchen,product2,Mobile
user3,Clothing,product3,Webstore
user1,Bathroom,product4,Tablet
user4,Kitchen,product2,Webstore

Your task is to calculate the number of clicks

- Per User
- Per Product
- Per Product Category
- Per Channel
- Per Product and Channel

Sample output format for number of clicks (Same format should be followed for other aggregations) :

"User1" : 2
"User2" : 1
"User3" : 1
"User4" : 1

3. Generate the Top 5 products according to the number of clicks

4. Implement test cases to test the above logic. You may use the ScalaTest package. Look at http://www.scalatest.org/user_guide/writing_your_first_test

The maven dependency for ScalaTest:
https://mvnrepository.com/artifact/org.scalatest/scalatest_2.11/3.0.4

5. Commit the changes and push

Spark Exercise 1

1. Repeat the above Scala Exercise 1 using the Spark RDD API
    - Read the input data as a Spark RDD
    - Perform the calculations using Spark RDD API

2. Repeat the above Scala Exercise 1 using the Spark Dataframe API
    - Read the input data as a Spark Dataframe
    - Perform the calculations using Spark Dataframe API