Conceptos y Aplicaciones en Big Data

2do semestre 2021

Práctica 5 - Spark

1) Indique (pensando en lo que hace cada operación) si las siguientes transformaciones son *narrow* o *wide*.
   a) cartesian
   b) coalesce
   c) distinct
   d) flatMap
   e) flatMapValues
   f) intersection
   g) repartition
   h) substract
   i) union

2) Usando el dataset EstacionesMeteorologicas imprima el id de la estación que tiene el máximo registro de humedad, el id de la estación con máximo registro en temperatura y el id de la estación con el máximo registro de precipitación usando solo seis transformaciones, incluyendo la transformación textFile.

3) Indique en cuántas etapas se ejecutan los siguientes scripts
   a)
```
A = sc.textFile("Caso B1")
B = A.map(fMap1)
C = B.filter(fFilter1)
C = C.map(fmap2)
A = sc.textFile("Caso B2")
B = A.map(fmap3)
D = C.join(B)
D = D.filter(fFilter2)
final = D.reduce(fReduce1)
```
   b)
```
A = sc.textFile("Caso C")
B = A.map(fMap1)
C = B.filter(fFilter1)
D = C.groupByKey()
C = B.filter(fFilter2)
E = C.groupByKey()
E = D.cogroup(E)
final = E.reduce(fReduce1)
```

c)

```
A = sc.textFile("Caso D.1")
B = sc.textFile("Caso D.2")
C = sc.textFile("Caso D.3")
A = A.map(fMap1)
A = A.distinct()
B = B.filter(fFilter1)
C = C.filter(fFilter1)
E = C.join(B)
B = A.map(fMap2)
C = E.map(fMap3)
D = B.union(C)
F = E.map(fMap4)
E = F.filter(fFilter2)
D = D.union(E).union(B)
B = D.substract(E)
final = B.count()
```

4) Utilizando el dataset Banco escriba un script que permita determinar si las siguientes afirmaciones son verdaderas
    a) El banco tiene más clientes europeos que americanos
    b) El promedio de edad de los clientes americanos es menor que el de los europeos
    c) Los americanos deben más plata que los europeos (un cliente debe plata si la suma de montos de todas sus cajas de ahorro es negativa)
    d) Los clientes americanos suelen sacar, en promedio, préstamos con mayor cantidad de cuotas que los europeos

5) Utilizando el dataset Banco escriba un script que permita calcular el factor de riesgo de todos sus clientes. El factor de riesgo de un cliente se calcula de la siguiente manera:

$$factorRiego = \frac{\left(\frac{D}{E} + 0.001\right)^{F}}{\left(\frac{A}{B}\right)^{\frac{1}{b-c+1}}}$$

donde:
A: saldo total entre todas las cajas de ahorro
B: cantidad de cajas de ahorro
C: cantidad de cajas de ahorro con saldo negativo
D: monto total de todos los préstamos
E: promedio de cuotas entre todos los préstamos
F: cantidad de préstamos

6) Realice un script que permita imprimir, por país, los nombres de los clientes cuyo factor de riesgo es menor a 2.

7) Del data set movimientos, el banco desea saber para cada uno de los meses y para cada cliente si sus movimientos tienen tendencia positiva o negativa.
    a) Calcule para cada cliente y para cada mes de cada año la pendiente de la recta determinada por regresión lineal de los movimientos de dicho periodo.
    b) Si en un periodo de tiempo un cliente tuvo tendencia positiva (pendiente positiva), el valor de ese período para ese cliente vale 1 (balance positivo). Si tuvo tendencia negativa (pendiente negativa) el valor del período vale -1 (balance negativo).
    c) Calcule para cada cliente su balance histórico, haciendo una regresión lineal entre todos sus balances.
    d) Imprima un ranking por país calculando
    núm_clientes_balance_historico_pos - núm_clientes_balance_historico_neg
    e) Un país tiene balance positivo en un determinado periodo si la mayoría de los clientes de ese país tienen balance positivo (en el mismo periodo). El país tiene balance negativo en ese periodo en caso contrario
    f) Calcule para cada país su balance histórico, haciendo una regresión lineal entre todos sus balances.
    g) Imprima un ranking de países ordenados por la pendiente de la recta reultado de la regresión.

Cálculo de la pendiente de la recta por regresión lineal:

$$\beta_1 = \frac{\sum x \sum y - n \sum xy}{(\sum x)^2 - n \sum x^2} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$