Conceptos y aplicaciones en Big Data

2do semestre 2021

Práctica 4 - Spark

1) Dado el siguiente RDD almacenado en la variable *rdd*:

Partition 1		Partition 2		Partition 3		Partition 4	
34	21	23	45	3	21	30	91
21	34	12	12	15	10	31	32
10	18	36	18	14	18	32	53
32	45	4	97	3	15	19	35

responda ¿Qué imprime cada uno de los siguientes scripts (sin ejecutarlo)?

a.
```
res = rdd.map(lambda t: t[0] + t[1] * 2)
print(res.first())
```

b.
```
res = rdd.filter(lambda t: t[0] >= t[1])
print(res.take(3))
```

c.
```
res = rdd.map(lambda t: (t[0], t[1], t[0] / t[1]))
res = res.filter(lambda t: t[2] < 0.5)
res = res.reduce(lamdda t1, t2:
                    t1 if t1[2] < t2[2] else t2))
print(res)
```

d.
```
r1 = rdd.map(lambda t: t[0])
r2 = rdd.map(lambda t: t[1])
r1 = r1.distinct()
r2 = r2.distinct()
res = r2.union(r1)
print(res.collect())
```

2) Dados los siguientes scripts en Spark, dibuje el DAG correspondiente ¿Qué se termina ejecutando?

a.

```
A = sc.textFile("Caso A")
B = A.map(fMap1)
C = B.filter(fFilter1)
final = B.reduce(fReduce1)
```

b.

```
A = sc.textFile("Caso B")
B = A.map(fMap1)
C = B.filter(fFilter1)
C = B.map(fmap2)
D = C.filter(fFilter2)
final = D.reduce(fReduce1)
```

c.

```
A = sc.textFile("Caso C")
B = A.map(fMap1)
C = B.filter(fFilter1)
D = B.filter(fFilter1)
E = B.filter(fFilter1)
C = C.union(D)
D = C.intersection(E)
E = C.substract(E)
final = E.reduce(fReduce1)
```

d.

```
A = sc.textFile("Caso D.1")
B = sc.textFile("Caso D.2")
C = sc.textFile("Caso D.3")
A = A.map(fMap1)
B = B.filter(fFilter1)
D = A.filter(fFilter1)
E = D.map(fFilter1)
F = D.union(A).union(E)
final = D.count()
```

3) Usando el dataset Banco, escriba un script en Python usando Spark para responder a las siguientes preguntas:

a. Nombre y apellidos de los clientes capricornianos.
b. Nombre y apellido de los clientes de nacionalidad argentina.
c. Del resultado de a) cuántos nacieron en verano.
d. Del resultado de b) quién es el cliente más joven y quién el más viejo.
e. El ID de la caja que tiene asociado el préstamo con mayor cantidad de cuotas y entre las que tienen la misma cantidad, el de mayor monto.
f. Los ID de clientes (únicos) con al menos una caja de ahorro (en positivo) cuyo saldo es mayor a 300 U$S.
g. Del dataset Movimientos, el monto del mayor movimiento y el id de caja del último movimiento.

4) Es posible resolver los siguientes problemas (por separado) utilizando una única función *reduce*:
    a. El promedio de edades de los clientes
    b. Determinar la cantidad de cuentas con saldo positivo y la cantidad de cuentas con saldo negativo.

5) El dataset EstacionesMeteorológicas posee información sobre registros de datos climáticos tomados por sus estaciones. Este dataset tiene tuplas con la siguiente información:
   <ID_Estación, fecha_registro, temperatura, humedad, precipitación>
   Y además está conformado por dos archivos:
    a. estacionNorte.txt almacena la información en grados centígrados, porcentaje de humedad, y milímetros de lluvia.
    b. estacionSur.txt almacena la información en grados Fahrenheit, porcentaje de humedad y centímetros de lluvia.

   Implemente una solución en Spark que permita obtener:
    - el promedio de temperatura, de humedad y precipitación total entre todas las estaciones.
    - el ID de la estación y la fecha que registró
        - la temperatura más fría
        - la temperatura más calurosa
        - la de mayor humedad
        - la de menor humedad
        - la de más precipitación
        - la de menor precipitación

   NOTA: En caso de dos estaciones con igual máximo o mínimo, devolver cualquiera de las dos.