Conceptos y Aplicaciones en Big Data

2do semestre 2021

Práctica 6 - Spark

1) Indique como queda el DAG y qué se ejecuta (y cuántas veces lo hace) en el siguiente script:

```
rdd1 = sc.parallelize(list)
for i in range(6):
      rdd2 = rdd1.map(fmap2)
      r1 = rdd2.reduce(fReduce)
      rdd3 = rdd2.union(rdd).map(fmap3) \
                  .reduceByKey(fReduceByKey)
      if(i > 3):
            rdd4 = rdd3.filter(fFilter1).persist()
            print (rdd4.collect())
print(rdd2.collect())
print(rdd3.collect())
```

2) Se desea calcular el promedio de las potencias de 2 a 5 de los primeros cinco números naturales.
   - $1^2+2^2+3^2+4^2+5^2 = 55 \Rightarrow 55 / 5 = 11$
   - $1^3+2^3+3^3+4^3+5^3 = 225 \Rightarrow 225 / 5 = 45$
   - $1^4+2^4+3^4+4^4+5^4 = 979 \Rightarrow 979 / 5 = 195.8$
   - $1^5+2^5+3^5+4^5+5^5 = 4425 \Rightarrow 4425 / 5 = 885$

   ¿Cuál es el error del siguiente script?

```
rdd = sc.parallelize([1,2,3,4,5])
for i in range(2,6):
      acc = sc.broadcast(i)
      rdd = rdd.map(lambda v: v ** acc.value)
      r = rdd.reduce(lambda x,y : x+y)
      r = r / 5
      print (r)
```

3) Plantee un algoritmo que permita aproximarse a la mediana de manera iterativa, que resulte más eficiente que el método visto en la teoría.

4) Plantee un algoritmo iterativo que permita imprimir el nombre y apellido de los clientes del banco que tienen un número primo de cajas de ahorro.

5) Plantee un algoritmo iterativo que permita resolver el método de Jacobi como el planteado en el ejercicio 7 de la práctica 2.

6) Dado el dataset Genealogía el cual está formado por:
<nombre_individuo, dni_individuo, dni_mamá>
realice distintas funciones que:
    a) Dado los dni de dos individuos indicar si son primos (dos individuos son primos si tienen la misma abuela)
    b) Dado los dni de dos individuos i_1 y i_2 indicar si i_1 es ancestro de i_2.
    c) El nombre de la "abuela" que tiene más descendientes
    d) Los nombres de los hermanos de la familia más numerosa (la cantidad de integrantes de una familia solo se calcula con la cantidad de hermanos más la mamá). Podría existir más de una familia más numerosa, en cuyo caso se deben imprimir todos los nombres de los hermanos integrantes de cada familia.