Conceptos y aplicaciones en Big Data

2do semestre 2021

Práctica 3 - Hadoop MapReduce

1) El siguiente job en MapReduce permite contabilizar cuantas palabras comienzan con cada una de las letras del abecedario.

```
def map(key, values, context):
      words = values.split()
      for w in words:
            context.write(w[0], 1)
def reduce(key, values, context):
      c=0
      for v in values:
            c=c+1
      context.write(key, c)
```

   a. Solucione el problema del "case sensitive" usando comparadores.
   b. ¿Cuántos reducers se ejecutan en este problema? Sabiendo que se cuenta con el doble de nodos para la tarea de reduce ¿Cómo podría usar los comparadores para aprovechar todos los nodos?

2) ¿Qué operaciones de resumen realizadas por los reducers se ven beneficiados por la definición de funciones de comparación personalizadas? ¿Qué consideraciones habría que tener en cuenta relacionado con la tarea de los *mappers*?

3) ¿La operación de inner join mejora su performance con la función combiner?

4) Muchos cálculos aritméticos necesitan ordenar una serie de números para obtener su resultado, como por ejemplo la mediana.

   La mediana es el "número en el medio" de una lista ordenada de números.

   3, 5, 7, 12, 13, 14, 21, ***23****, 23, 23, 23, 29, 39, 40, 56*

   Implemente una solución MapReduce que permita calcular la mediana de una serie de valores. Use como prueba el dataset website para calcular la mediana del tiempo de permanencia.

5) Implemente una solución MapReduce para el método de Jacobi utilizando el dataset jacobi2 y los valores iniciales están en un dataset con el formato:

incognita_i	valor
var1	1
var2	2
var3	3

Nota: En esta solución los valores de las incógnitas NO pueden ser pasados por parámetros a los *mappers* y *reducers*. Deben ser recibidos como una entrada del job.

6) Implemente una solución MapReduce que permita crear el dataset con valores iniciales arbitrarios para Jacobi. Suponiendo que hay 1000000 de variables ¿Cómo optimizaría la creación de todos los valores aprovechando la capacidad de paralelismo del paradigma?
Nota: Piense como entrada del job un único archivo con un único valor: el número de variables que hay que crear.

7) Dado el dataset Banco (visto en teoría) el cual está compuesto por tres datasets:
Cliente: <ID_Cliente, nombre, apellido, DNI, fecha de nacimiento, nacionalidad>
CajaDeAhorro: <ID_Caja, ID_Cliente, saldo>
Prestamos: <ID_Caja, cuotas, monto>
Implemente una solución MapReduce para resolver las siguientes consultas SQL describiendo el DAG correspondiente.

a.

```
SELECT nacionalidad, Count(*) AS cuantos
FROM Cliente AS C INNER JOIN CajaDeAhorro AS CA
     ON C.ID_Cliente = CA.ID_Cliente
          INNER JOIN Prestamos AS P
               ON CA.ID_Caja = P.ID_Caja
GROUP BY nacionalidad
ORDER BY cuantos DESC
LIMIT 1
```

b.

```
SELECT (Year(fecha_nacimiento) % 100) AS decada,
          Avg(saldo)
FROM Cliente AS C INNER JOIN CajaDeAhorro AS CA
     ON C.ID_Cliente = CA.ID_Cliente
WHERE nacionalidad = "ITA"
GROUP BY decada
```

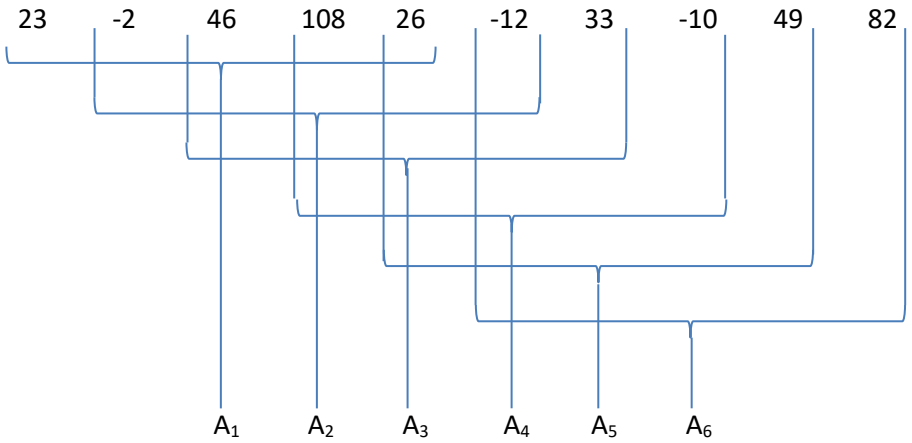
c.

```
SELECT nombre, apellido, DNI
FROM Cliente AS C INNER JOIN CajaDeAhorro AS CA
     ON C.ID_Cliente = CA.ID_Cliente
WHERE C.ID_Cliente IN
      ( SELECT ID_Cliente
        FROM CajaDeAhorro AS CA INNER JOIN
            Prestamos AS P ON CA.ID_Caja = P.ID_Caja
        GROUP BY ID_Cliente
        HAVING Count(*) > 2 )
AND Year(fecha_nacimiento) < 2000
GROUP BY C.ID_Cliente
HAVING Max(saldo) > 500000
```

8) El banco almacena en un cuarto archivo (llamado Movimientos) todas las transacciones realizadas y las almacena con este formato:
<ID_Caja, monto, timestamp>
Donde para cada caja de ahorro almacena en un determinado timestamp (en formato AAAA-MM-DD HH:mm:SS) el monto del movimiento (puede ser negativo o positivo según si se hizo una extracción o un depósito).
Implemente una solución en MapReduce que permita calcular para cada cliente del banco su media móvil (moving average) en un mes y año determinado. Tanto el año, como el mes y el ancho de la ventana deben ser parámetros de la consulta.

<u>Media móvil (moving average)</u>

Dada una serie numérica ordenada en el tiempo, la media móvil de ancho 5 se calcula como:



donde A_1 .. A_6 es el promedio aritmético de cada sub-serie de cinco valores. La media móvil es el promedio de todos los A_i.