Conceptos y aplicaciones en Big Data

2do semestre 2021

Práctica 1 – Paradigma MapReduce

1) Dado el siguiente dataset:

Split 1		**Split 2**		**Split 3**		**Split 4**	
Key	Value	Key	Value	Key	Value	Key	Value
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 para cada job: ¿Cuántas veces (invocaciones) se ejecuta la función map? ¿Cuántas veces (invocaciones) se ejecuta la función reduce? ¿Cuántos *mappers* se ejecutan? ¿Cuántos *reducers* se ejecutan? ¿Qué datos recibe cada función reduce? ¿Cuál es la salida de cada job?

a) Job A

```
def map(k1, v1, context):
      context.write(1, v1)

def reduce(k2, v2, context):
      n = 0
      for v in v2:
            n = n + 1
      context.write(k2, n)
```

b) Job B

```
def map(k1, v1, context):
      context.write(1, v1)

def reduce(k2, v2, context):
      n = 0
      for v in v2:
            n = n + v
      context.write(k2, n)
```

c) Job C

```
def map(k1, v1, context):
     if (v1 < 30):
          context.write(1, k1)
     else:
          context.write(2, k1)

def reduce(k2, v2, context):
     max = -1
     for v in v2:
          if(v > max):
               max = v
     context.write(k2, max)
```

d) Job D

```
def map(k1, v1, context):
     for v in range(v1):
          context.write(k1, v1)

def reduce(k2, v2, context):
     n = 0
     for v in v2:
          n = n + 1
     context.write(k2, n)
```

e) Job E

```
def map(k1, v1, context):
     context.write(v1, k1)

def reduce(k2, v2, context):
     n = 0
     for v in v2:
          n = n + 1
          context.write(v, n)
```

2) El dataset Libros provisto por la cátedra almacena libros cada uno en un archivo separado. Dentro de cada archivo, la primera línea tiene el título del libro y luego en las líneas siguientes un párrafo por línea. Ejecute el proyecto WordCount dado por la cátedra para saber cuántas veces es utilizada cada palabra.

3) En el ejercicio anterior ¿Cómo haría para obtener el top 20 de las palabras más usadas?

4) Modifique el proyecto WordCount para contar cuántas vocales, consonantes, dígitos, espacios y otros caracteres posee el data set Libros.

5) Indique si utilizando el dataset Libros es posible resolver los siguientes problemas:
    a. Obtener los títulos de todos los libros
    b. Obtener la cantidad de palabras promedio por párrafo
    c. Obtener la cantidad de párrafos promedio por libro
    d. Obtener la cantidad de caracteres del párrafo más extenso
    e. Cantidad total de párrafos con diálogos (se entiende por párrafo con diálogo aquel que empieza con un guión)
    f. El diálogo más largo (se entiende por diálogo a una secuencia de párrafos con diálogo que aparecen de manera consecutiva)
    g. El top 20 de las palabras más usadas por cada libro

6) Una empresa proveedora de internet realizó una encuesta para conocer el grado de satisfacción de sus clientes, en un formulario web los clientes debían completar un campo con los textos "Muy satisfecho", "Algo satisfecho", "Poco satisfecho", “Disconforme” o "Muy disconforme". Utilice el dataset Encuesta para saber cuántos clientes están en cada una de las cinco categorías.

7) El dataset Inversionistas posee los nombres, dni, fecha de nacimiento (día, mes y año como campos separados) e importe invertido por diferentes personas en la apertura de un nuevo negocio en la ciudad. Se desea saber:
    a. El nombre del inversionista más joven
    b. El total del importe invertido por todos los inversionistas
    c. El promedio de edad

    Implemente una solución en MapReduce. ¿Se puede resolver los tres problemas en un único job?

8) Si contáramos con un cluster donde podemos configurar 100 nodos para la tarea de reduce ¿De qué manera se podrían usar esos 100 nodos en el ejemplo de los eventos POSITIVO, NEGATIVO y NEUTRO visto en la teoría?