Scalable Data Mining
Quiz-1
Hadoop

1. Consider the dataset where each record is of the form $(1, studentRollNo, StudentName)$ or $(2, studentRollNo, SubjectName, Marks)$. A map reduce program which prints average marks for each student with his/her name. What will be mapper output key and mapper output value?

   (a) $studentRollNo$, $(type, StudentName/Marks)$ where $type$ is 1 or 2 depending on the record.

   (b) $studentRollNo$, $(SubjectName, StudentName/Marks)$

   (c) $StudentName$, $(studentRollNo, Marks)$

   (d) $(type, studentRollNo)$, $(type, StudentName/Marks)$

   Correct answer a . The key is studentRollNo, hence it's the mapper output key. Value will be StudentName or Marks depending on the type of record.

2. Complete the reducer logic for the above problem assuming that mapper output key is $studentRollNo$:

```
sum = 0
count = 0
for each record(r):
  if type = 1:
    (1) _______________
  else:
    (2) _______________
    count = count + 1
print name , (3) ________
```

   (a) (1) sum = Marks (2) name = SubjectName (3) sum/count

   (b) (1) name = SubjectName (2) sum = sum + Marks (3) sum/count

   (c) (1) name = StudentName (2) sum = sum + Marks (3) sum/count

   Correct answer c

3. Where is the intermediate data passed between the map and reduce worker, stored?

(a) External File System

(b) Local disk of map worker

(c) Local disk of reduce worker

(d) Not stored

Correct answer b

4. Suppose we have four mappers with outputs being as follows:

   Mapper 1: (a,1) (b,2)

   Mapper 2: (c,3) (c,6)

   Mapper 3: (a,5) (c,2)

   Mapper 4: (b,7) (c,8)

   What will be the key value pairs that will be fed as input to the reducer (i) with Combiner (ii) without Combiner ?

   (a) (i) (a,1) (a,5) (b,2) (b,7)(c,3) (c,6) (c,2) (c,8)
   (ii) (a,1) (b,2) (c,3) (c,6) (a,5) (c,2) (b,7) (c,8)

   (b) (i) (a,[1,5]) (b,[2,7]) (c,[3,6,2,8])
   (ii) (a,1) (a,5) (b,2) (b,7)(c,3) (c,6) (c,2) (c,8)

   (c) (i) (a,[1,5]) (b,[2,7]) (c,[2,9,8]
   (ii) (a,[1,5]) (b,[2,7]) (c,[3,6,2,8])

   (d) (i) (a,6) (b,9) (c,18)
   (ii) (a,[1,5]) (b,[2,7]) (c,[3,6,2,8])

   Correct answer c

5. Consider a map-reduce program which takes two matrices as input and computes the product matrix. The input format is:

   `<Matrix id> <Row Index> <Column Index> <Value>`

   Here `<Matrix id>` is 1 for the first matrix and 2 for the second matrix. Also, assume that the dimensions of the matrices are $m \times n$ and $n \times k$. Complete the following mapper code:

   Mapper Code:

```
for line in sys.stdin:
    matrixid,row,col,value = line.split("\t")
    if matrixid == 1:
        for i in range(1,k+1):
             print((---,i),[matrixid,col,value],sep='\t')
    else:
         for i in range(1,m+1):
         print((i,---),[matrixid,row,value],sep='\t')
```

   (a) `col` and `row`

(b) `row` and `col`

(c) `1` and `col`

(d) `row` and `1`

Correct answer b : This is because the mapper generates copies of rows for the first matrix with column ids of the second and vice versa.
Consider matrix 1 : $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ and matrix 2: $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$. Running the algorithm will fetch us 2 keys: (1,1) and (2,1) and their respective values from each matrix.

6. Following the question above, and assuming that mapper aggregates the correct records in the reducers, complete the reducer code:

```
Reducer Code:

for line in sys.stdin:
    key,value = line.split('\t')
    matrixid, index, mult = value.split(' ')
    if matrixid == ---:
        mat1[index]=mult
    else:
        mat2[index]=mult

sum=0
for i in range(1,---):
    sum=sum+mat1[i]*mat2[i]

print(key,sum)
```

(a) `1` and `n+1`

(b) `2` and `index`

(c) `1` and `index`

(d) None of the above

Correct answer a : This is because all the corresponding entries given by index should be multiplied and added to give the result. For each key obtained from mapper, we multiply and add values. Intuitively, # of cells of a product matrix is equal to number of keys. The reducer code will work for each key.