Quiz 2-Solutions(Spring 2020)

1.
Input and Output : are stored on a distributed file system (DFS) → **0.5 mark**
Intermediate Results : on local file system of Map workers →**0.5 mark**

We care about where the data is stored because we are interested in decreasing communication costs and also helps to deal with node failures →**1 mark**

2.
Map Task: Emit (key, value) pair →**0.5 marks if key and value is mentioned correctly**
Key is **orderid** (key used for join)
Value is tuple of **(Order,orderid,account,date)/(LineItem,orderid,itemid,qty)**

Eg: key=1 value=(Order,1,aaa,d1)
Key=1 value=(LineItem,1,10,1)

Reduce Task: emit joined values (without table names)→ **0.5 marks**

Eg:
input :
Key=1, values=[(Order,1,aaa,d1), (LineItem,1,10,1)]

Output:
1,aaa,d1,1,10,1

Final Output: → **1 mark**

1,aaa,d1,1,10,1
1,aaa,d1,1,20,3
2,aaa,d2,2,10,5
2,aaa,d2,2,50,100
3,bbb,d3,3,20,1

3.
1st Map Function: →**0.75 marks**
For each matrix element A[ij] :
emit(j , (A, i, A[i,j])
For each matrix element B[jk] :
emit(j , (B, k, B[j,k])

1st Reduce Function: →**0.75 marks**
For each value of (i,k) which comes from A and B, i.e. (A, i, A[ij]) and (B, k, B[jk]):
emit((i,k), (A[ij] * B[jk]))

2nd Map Function: →0.75 marks

map(key,value):
#Let the pair of (((i,k), (A[ij] * B[jk])) pass through

2nd Reduce Function: →0.75 marks

reduce(key,values):
For each (i,k) we will add up the values, thus emitting:
emit((i,k),Sum(values))

*** If you write in two stages correctly, your will get 3 points**

4.

Reducer Size: is the upper bound on the number of values that are allowed to appear in the list associated with a single key. **→1mark**

Replication Rate:
number of key-value pairs produced by all the Map tasks on all the inputs, divided by the number of inputs. That is, the replication rate is the average communication from Map tasks to Reduce tasks per input. **→1mark**

We are concerned about it as the running time will be greatly reduced if we can avoid having to move data repeatedly between main memory and disk. **→1mark**