Tuesday Quiz

1. [1 point] What is an eager operation in Spark?

   An eager operation (action) is something that gets executed immediately.

2. [2 points] Where does the MapReduce program store 1. Input data, 2. Intermediate files, and 3. Output data? Why do we care about where they store the data?

   Input data - DFS [0.5 point]

   Intermediate files - Local FS [0.5 point]

   Output data - DFS [0.5 point]

   We care about where they store data as we can look into failures and make sure that minimum tasks have to be rescheduled to carry out the successful completion of the map-reduce application. [0.5 point]

3. [1 point] Which of the following computations on a large set of integers (that may contain duplicates) require(s) a two-stage Map Reduce program?
   a. Compute the integers divisible by 7 in the set
   b. Compute the count of integers in the set
   c. Compute the largest integer in the set
   d. Compute the count on distinct integers in the set
4. [1 point] Consider multiplying two matrices A (3X3) and B (3X2). Consider the **one-stage** approach to matrix multiplication (AXB) as discussed in class.

   A = [1 1 1]
   [2 1 2]
   [1 2 1]

   B= [1 0]
   [0 1]
   [1 2]

   If the Mapper takes as the input the element A[2,2], which of the following key-value pairs will be in its output?
   a. ((2,1),(A,2,21,A[2,2]))
   b. ((2,2),(A,2,1,A[2,2]))
   c. ((1,1),(A,1,1,A[2,2]))
   d. ((1,2),(A,1,2,A[2,2]))

   Answer - emit ((i,k), (‘A’, i, j, A[i,j]) for k in 1..2

5. [1 point] Consider multiplying two matrices A (3X3) and B (3X2). Consider the **two-stage** approach to matrix multiplication(AXB) as discussed in class.

A = [1 1 1]
[2 1 2]
[1 2 1]

B= [1 0]
[0 1]
[1 2]

If the Mapper in **stage 1** takes as the input the element B[3,2], which of the following key-value pairs will be in its output?
   a. (3,(B,2,B[3,2]))
   b. (2,(B,2,B[3,2]))
   c. (2,(B,3,B[3,2]))
   d. (3,(B,3,B[3,2]))

Answer - emit(j , (B, k, B[j,k])

6. [2 points] Write the map reduce solution for Distributed Sort. We would like to sort a very large list of (firstName, lastName) pairs by lastName followed by firstName
Examples of outputs:
Smith Anne
Smith John
Smith Ken
- Map Task: [0.75 points]
emit(lastName, firstName)
- Group By Keys [0.5 points]
- Reduce Task: [0.75 points]
For each lastName key, if there are multiple firstName values, emit(lastName, firstName) in alphabetical order. Merge output from all reduce tasks.

7. [2 points] Write the Map and Reduce tasks and their output for joining these two tables:

Order(orderId, account, date)
1, aaa, d1
2, aaa, d2
3, bbb, d3

LineItem(orderId, itemId, quantity)
1, 10, 1
1, 20, 3
2, 10, 5

2, 50, 100
3, 20, 1

```
Map(key, value) [0.75 points]
  If table is "Order":
      emit(orderId, ["Order", (orderId, account, date)])
  If table is "LineItem":
      emit(orderId, ["LineItem", (orderId, itemId, quantity)])
```

```
Reduce(key, value) [0.75 points]
  emit(orderId, Order JOINS LineItem for each key-value pair(exclude table name from results))
```

The output of the map step for the example will be:

```
(1, ["Order", (1, aaa, d1)])
(2, ["Order", (2, aaa, d2)])
(3, ["Order", (3, bbb, d3)])
(1, ["LineItem", (1, 10, 1)])
(1, ["LineItem", (1, 20, 3)])
(2, ["LineItem", (2, 10, 5)])
(2, ["LineItem", (2, 50, 100)])
(3, ["LineItem", (3, 20, 1)])
```

The output of the reduce step for the example will be:

```
(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)
```

Final output [0.5 points]