Quiz 1 – Thursday - Rubrics

1. [1 point] State all the stages of the Map-Reduce paradigm.
   Map ; Group-by-key ; Reduce

2. [1 point] What 2 properties should a reduce function have for it to be used as a combiner?
   The reduce function should be commutative and associative.

3. [1 point] Briefly explain the advantages of a distributed file system.
   High Availability - stores data across multiple disks for availability and persistence
   Moves the computation closer to the data to minimize data movement.

4. [1 point] What is the typical size of Chunk server?
a. 16 MB
b. 64 MB
c. 128 MB
d. 512 MB

5. [1 point] Which of the following statement(s) about Combiner in MapReduce is (are) correct? (Select all that apply)
. Combiner is often called a mini-reducer.
a. Combiner takes the output from the Mapper as the input
b. The results of combiners will be sent to the Reducers.
c. Combiner may reduce the amount of data transferred from Mappers to Reducers
   Answer - All the above/a,b,d - It does go through sorter/shuffler but that is an in-build task, eventually it goes to the reducer.

6. [2 points] Design a MapReduce algorithm that takes a very large file of integers and produces as output all unique integers from the original file that are evenly divisible by 3. Just indicate the logic needed using pseudocode.
   Map (key, value list):[1 point]
   for v in value list:
      if (v % 3) == 0:
      emit (v, 1)

   Reduce (key, values)[1 point]
   : Eliminate duplicates
   emit (key, 1)
   [1 point] Where did you check divisibility by 3, Map or Reduce? Why?
   In Map Task so as to reduce communication i.e. send less data over network.

7. [2 points] Recall the “evil-doer” example when we talked about Bonferroni's principle. Suppose that we make the following assumptions. We track 2 million people for 100 days. Each person stays in a hotel 1% of the time. Each hotel holds 100 people and there are 100 hotels.What is the expected number of “suspicious” pairs of people (i.e., they went to the same hotel on some two days)?
   Expected number of suspicious people is 2500 * 4
   Let p be the probability of person p being in the hotel

p = 1/100
Let q be the probability of person q being in the hotel
q = 1/100
A = probability that p and q are in the same hotel on day d
A = 1/100 * 1/100 * 1/100 = 10^ -6

B = probability that p and q are in the same hotel on day d1 and d2
B = A * A = 10^-6 * 10^-6 = 10^(-12) [0.5 point]

Choose 2 days = 100C2
C = probability that p and q are in the same hotel on day d1 and d2 with two days selected
C = 10^(-12) * 100*100/2 = 10^-8/2 [0.5 point]

Choose pairs of people = 10^6C2 = 9*10^12/2 [0.5 point]
D = the expected number of "suspicious" pairs of people
D = (10^-8)/2 * (10^12)/2 = 10^4/4 = 2500*4 [0.5 point]

the expected number of "suspicious" pairs of people = 10000