Quiz 1 – Tuesday - Rubrics

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

2. [1 point] What is the advantage of using combiners?
   Reduce network traffic for data sent to the reduce tasks.

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] Consider a system with “m” mappers and “r” reducers. In the grouping stage, the master controller applies a hash function of the form “(a+b*x) % y”, where “a” and “b” are constants, “x” is the key that comes from the mapper. What should the ideal value of “y” be?
A. m - 1
B. m
C. r
D. r - 1
The master controller typically applies a hash function to keys and produces a bucket number from 0 to r-1. For this to happen, the value of “y” must be “r”

5. [2 points] You are given a list containing the marks of students for different subjects (studentID, subject, score). Write a map-reduce pseudo-code to calculate the total marks of each student. (Same as word-count taken in class)
   map(key, value): [1point]
   For each element in the chunk:
      emit(studentID, score)
   reduce(key, values):[1 point]
   emit(key,sum(values))

6. [1 point] Which of the following mathematical computations are both commutative and associative?
a. Addition
b. Multiplication
c. Mean
d. Median

Answer - a, b

7. [1 point] In data mining, the discovered patterns and models are required to be (Multiple answer correct):
a. Useful
b. Valid
c. Understandable
d. Unexpected
   Answer - All of the above

8. [2 points] Recall the “evil-doer” example when we talked about Bonferroni's principle. Suppose that we make the following assumptions. We track 1 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

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 marks]

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 = 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 [0.5 point]

the expected number of "suspicious" pairs of people = 2500