CSC 590, CSC 690 – Data Analytics
Exam #1, Fall 2024

First/Given Name: ____________________________

Last/Family Name: ____________________________

This exam contains 7 pages (including this cover page) and 6 questions.

- Clearly identify your answer for each problem, and try to organize your work in a reasonably coherent way, in the space provided. If you decided to use the back of a paper, note this clearly so the instructor can find your answer.
- Partial credit will be given for incorrect or incomplete answers that show a partial understanding of the relevant concepts. Irrelevant and meaningless answers will not receive partial credit.
- No electronic devices, including calculators, are allowed.
- Each student is allowed to use only a cheat sheet of size 8.5″×5.75″, which is equivalent to a half of a standard letter-sized paper. The cheat sheet can be used on both sides. Only handwritten cheat sheets are allowed, and each student is required to write their name on their cheat sheet. The cheat sheet must be submitted along with the exam upon completion.

Question	Points	Score
1	1.00	
2	1.00	
3	1.50	
4	2.00	
5	2.00	
6	1.50	
Total:	9.00	

I acknowledge that it is the responsibility of every student at Missouri State University to adhere to the university's policies on Student Academic Integrity. I confirm that I have neither given nor received any unauthorized assistance during this exam.

Signature: ____________________________

1. Assume we have 36 TB of data.

   (a) (0.50 points) How many days is required to explore 36TB of data by a single node with a speed of 50MB/sec? (1TB $\approx$ 1000GB, and 1GB $\approx$ 1000 MB)

   (b) (0.50 points) If we want to explore this amount of data using a cluster in 1 hour, how many nodes should be in that cluster?

2. (1.00 points) Indicate whether the following statements are true or false:

   (a) Hadoop follows a scale-up strategy, so it cannot add new computing nodes.

   (b) The Hadoop Distributed File System (HDFS) is good at reading big files, but not so good at making/writing random changes in the data.

   (c) There is a cluster of five computing nodes, in which each node has a drive of 1 TB dedicated to the HDFS. It is not possible to store a file that takes 5 TB.

   (d) The Node Manage lives in the master node.

Points earned for this question: ______

3. Assume we have the following list of key-value pairs:

$$[(4,2),(7,1),(3,3),(5,1),(3,5)]. \tag{1}$$

(a) (0.50 points) We design a map function map(k, v) that returns $(k + v, k)$ for each key-pair (k, v). What is the output of applying this function on the list in (1).

(b) (0.50 points) Using the output from the previous part, what would be the output of the shuffle phase?

(c) (0.50 points) For the output of the previous shuffle, we define a reduce function as lambda $x, y : x + y - 1$. What is the output of this phase?

Points earned for this question: ______

4. (2.00 points) Design map and reduce functions to obtain the histogram of the areas of a given collection of rectangles. Each rectangle is given by a key-value pair $(k, (x, y))$ in which k is the key, x is the width, and y is the length of the rectangle. A sample input for this problem looks like this:

$$(a, (3, 4)) \quad (b, (5, 4)) \quad (c, (6, 2)) \quad (d, (3, 5)) \quad (e, (10, 2)). \tag{2}$$

Accordingly, the areas of each of the rectangles with keys a and c is 12, the areas of each of the rectangles with keys b and e is 20, and the area of the rectangle with key d is 15. Hence, the area 12 is appeared 2 times, the area 20 is appeared 2 times, and the area 15 is appeared only 1 times.

The output for this example is:

$$(12, 2) \quad (20, 2) \quad (15, 1). \tag{3}$$

Use the following diagram to illustrate your solution with this example. The diagram should show the map function and its result on the input, the result of the shuffle phase, and the reduce function along with the result

Map: $(k, (x, y)) \rightarrow$ ____________________

Input: key, width, and length

Output: _______

(a, (3, 4))

(b, (5, 4))

(c, (6, 2))

(d, (3, 5))

(e, (10, 2))

Shuffle (network)

Reduce: ____________________

Points earned for this question: _____

5. (2.00 points) Design map and reduce functions to find for each pair of friends, the list of all persons who are not friends with either one of those two friends The input is a file in which each line is in the form of

$$\textit{person} : \textit{list of friends}.$$

Illustrate your solution using the following example. Draw a diagram similar to the one on the previous page.

Sample input

$a : b, c, f$

$b : a, c$

$c : a, b$

$d : e$

$e : d$

$f : a$

Output for the sample input

$((a, b), [d, e])$

$((a, c), [d, e])$

$((b, c), [d, e, f])$

$((d, e), [a, b, c, f])$

$((a, f), [b, c])$

Map: $(\text{person}, \mathit{list}(\text{friend})) \rightarrow$ ____________________

Shuffle (network)

Reduce: ______________________

Input: key, width, and length **Output:** ________

$a: b, c, f$

$b: a, c$

$c: a, b$

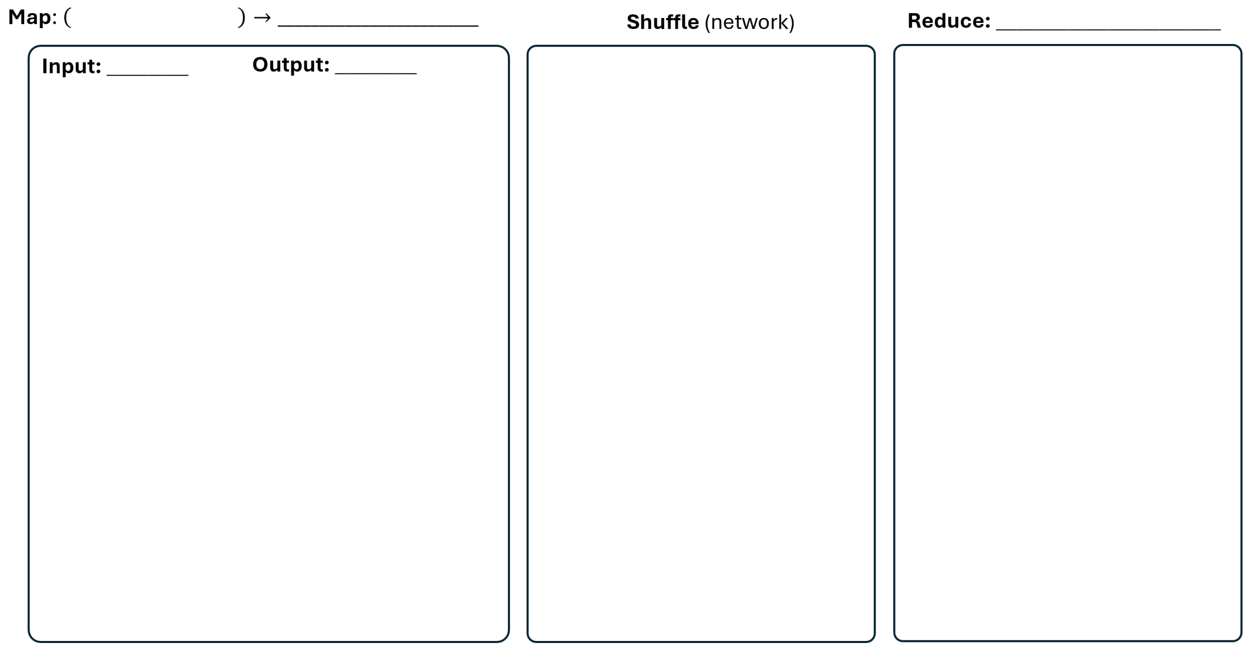
$d: e$

$e: d$

$f: a$

Points earned for this question: ______

We can use the following diagram in case you need it.



Points earned for this question: ______

6. (1.50 points) Let $P = [(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)]$ be a list of points in the plane, and $o = (o_1, o_2)$ and $r = (r_1, r_2)$ be the locations of two post offices. Complete the following Python program by implementing **only one map function** and **only one reduce function** so it computes and prints a tuple (d_1, d_2) in which d_1 is the minimum distance of the points in P to o, and d_2 is the maximum distance of the points in P to r. Formally,

$$d_1 = \min_{p \in P} \sqrt{(p[0] - o[0])^2 + (p[1] - o[1])^2},$$

and

$$d_2 = \max_{p \in P} \sqrt{(p[0] - r[0])^2 + (p[1] - r[1])^2}.$$

Note that we have already written the code to compute the distance between two points. You only need to implement the map and reduce functions.

Important: To receive full credit, your implementation must include only one map function and one reduce function. Partial credit will be awarded if you use more than one map function or more than one reduce function.

```python
import math
from functools import reduce

P = [(1, 2), (3, 4), (7, -2), (-5, 2), (0, -7)]
o = (3, 3)
r = (-2, -2)

# This function returns the distance between points p and q
def dist(p, q):
    return math.sqrt((p[0]-q[0])**2 + (p[1]-q[1])**2)
```

Points earned for this question: ______