Name: _____ USC ID: _____

INF 551 – Fall 2018 (Afternoon section)

Quiz 9: SQL via Hadoop MapReduce (10 points), 10 minutes

Consider a sales data stored as a csv file (fields are category, state, year, and amount):

Cell,CA,2017,30
Cell,CA,2016,30
Cell,NY,2018,80
Laptop,CA,2017,15
Laptop,CA,2018,50
Laptop,CA,2015,35
Tablet,NY,2015,50

Consider computing the answer to the following SQL query using Hadoop MapReduce (assume that NO combiner is used).

```
Select category, state, avg(amount)
From sales
Where amount > 20
Group by category, state
Having count(*) > 1
```

1. [5 points] Write the logic of **map** function in pseudocode. What will the map function output for the data above?

```
def map(key, value):
#key: offset of each line of the file, value: the string content of each line
    data = value.strip().split(',')
    if data[3] > 20:
        output ((data[0], data[1]), data[3])
```

output: (('Cell','CA'),30), (('Cell','CA'),30), (('Cell','NY'),80), (('Laptop','CA'),50), (('Laptop','CA'),35), (('Tablet','NY'),50)

2. [5 points] Write the logic of **reduce** function in pseudocode. What will the reduce function output for the data above?

```
def reduce(key, values):
    sum = 0
    count = len(values)
    if count:
        for value in values:
            sum += value
        output (key, float(sum)/count)
```

output: (('Cell', 'CA'), 30), (('Laptop', 'CA'), 42.5)