DSCI 632: Applied Cloud Computing
Midterm Assessment

Name: ________ Score: ______

Q1 (20 Points): When do we need to write functions in a *map-reduce* way? What is the syntax difference between Python *map-reduce* and Pyspark *map-reduce* way?

Q2 (20 Points): Obtain the maximum element of a very large list (Do not use max built-in function). Implement it in Python reduce way.

Q3 (20 Points): Assume a very large list is given. Write a function that removes duplicates in the list using map-reduce in Python (returns a list/set that there are no duplicates in it).

Hint 1 (for reduce step):

```
A = {1, 2}
B = {2, 3, 4}
print(A.union(B))
{1, 2, 3, 4}
```

Hint 2 (for map step):

```
a = 3
print(set([a]))
{3}
```

Q4 (20 Points): We have setup a cluster on Dataproc at Google Cloud Platform (GCP) with the following master and workers specifications:

Item	Machine Type	Virtual CPUs	Attached persistent disk	Number in cluster
Master Node	n1-standard-4	4	500 GB	1
Worker Nodes	n1-standard-4	4	500 GB	5

How much should we pay if we use this cluster for a Big Data Processing with Pyspark for 2 hours?

Hint:
Dataproc charge = # of vCPUs * hours * Dataproc price
Dataproc price = $0.01

Q5 (20 Points): Mention 8 Pyspark DataFrame methods that its name or its functionality is different from Pandas.

Q6 (30 Points): The following dataset (in.txt has two values at each row, we say it has two columns) and code in Pyspark is given. How can we obtain the average of each column? Write down your code with Pyspark to obtain the average value for each column.

```
from pyspark.sql import SparkSession

spark = SparkSession.builder.appName("Pyspark average
example").config("spark.some.config.option", "some-value").getOrCreate()

lines = spark.read.text("in.txt").rdd
row_rdd = lines.map(lambda x: x.value.split(','))
row_rdd.take(5)
```

Hint 1:
a ='0.727832' -> float(a) is equal to 0.727832
b = '0.427831'-> float(b) is equal to 0.427831

Hint 2: Above, was RDD based mean computation. You can do Pyspark DF based mean computation.

Optional Bonus Question:

Q7 (30 Points): for Titanic data set, create a new column as ‘Age_Normalize’. Each value in this column would be Age - Mean(‘Age’) .

Hint: We also need udf here