



Name: _____

MapReduce

Question 1

Consider two relations $R(a, b)$ and $S(b, c)$.

```
SELECT R.b, max(S.c) as cmax
FROM R, S
WHERE R.b = S.b
      AND R.a <= 100
GROUP BY R.b;
```

For the **Map** function, what are the computations performed, and what will be its outputs? Assume that the Map function reads a block of R or S relation as input.

Question 2

Consider two relations R(a, b) and S(b, c).

```
SELECT R.b, max(S.c) as cmax
FROM R, S
WHERE R.b = S.b
      AND R.a <= 100
GROUP BY R.b;
```

For the **Reduce** function, what will be its inputs, what are the computations performed, and what will be its outputs?

Parallel Processing

Given the following relations D(A, B) and E(A, C)

Suppose that D and E are partitioned across 3 different machines using random block partitioning, and no indexes are available on any of the machines. If we use a hash-join (aka. shuffle-join) in the relational algebra plan to execute the queries below, determine whether the following conditionals can be computed before or whether they **must** occur after the shuffle.

Use for questions 3 - 5.

Question 3

```
SELECT D.A  
FROM D, E  
WHERE D.A = E.A  
AND E.C > 10;
```

Do we need to shuffle before 'E.C > 10' can be determined?

Question 4

```
SELECT D.A
FROM D, E
WHERE D.A = E.A
AND E.C - D.B > 20;
```

Do we need to shuffle before 'E.C - D.B > 20' can be determined?

Question 5

```
SELECT D.A
FROM D, E
WHERE D.A = E.A
GROUP BY D.A
HAVING MAX(E.C) < 100;
```

Do we need to shuffle before 'GROUP BY D.A' can be determined?

Do we need to shuffle before 'HAVING MAX(E.C) < 100' can be determined?

Question 6

Describe in a few sentences how you would partition the data between machines if your goal is to maximize performance of the above query.

LetterID	SenderAddr	RecipientAddr	Status	ContentType	Content
12345	1600 Pennsylvania Ave	185 E Stevens Way	Delivered	Text	"Dear Hannah, I would like to request a regrade of the midterm ..."
67890	3800 E Stevens Way	185 E Stevens Way	InTransit	DVD	(lots and lots of 0s and 1s)

All of the attributes have uniformly distributed data except for recipients (eg, the President of the United States gets an unusually large or unusually small amount of letters). Consider the partitioning strategies we know:

- Block (Horizontal)
- Range (Horizontal) - please specify attribute set
- Hash (Horizontal) - please specify attribute set
- Vertical - please specify attribute set

Which one would you choose under the following circumstances? If you choose Range, Hash, or Vertical partitioning, please specify the attributes that you would partition on.
