CSE 444 – Homework 6
Parallelism and Distribution

Name: _____

Question	Points	Score
1	20	
2	20	
Total:	40	

1 Parallel Data Processing

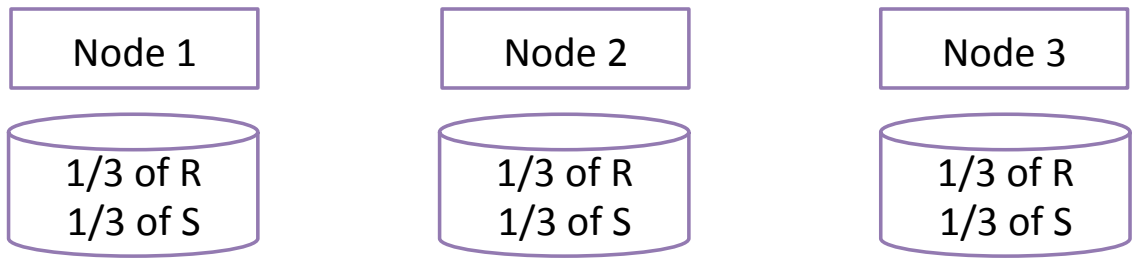
1. (20 points)

   (a) (10 points) Consider two relations `R(a,b)` and `S(c,d)` that are both horizontally partitioned across $N = 3$ nodes as shown in the diagram below. Each node locally stores approximately $\frac{1}{N}$ of the tuples in `R` and $\frac{1}{N}$ of the tuples in `S`. The tuples of `R` are *randomly* organized across machines (i.e., `R` is block partitioned across machines) while the tuples of `S` are *hash-partitioned* on `S.c`.

   Show a relational algebra plan for the following query and how it will be executed across the $N = 3$ machines. Pick an *efficient* plan that leverages the parallelism as much as possible. Include operators that need to re-shuffle data and add a note explaining how these operators will re-shuffle that data. For example, if you need to re-hash the data, add a "hash" operator into your query plan.

   Draw the parallel query plan. Indicate the edges that re-shuffle data across machines by drawing them as dashed lines:

```
SELECT a, avg(d) as avg
FROM R, S
WHERE R.b = S.c
AND S.d > 0
GROUP BY a
```



Answer:

(b) (10 points) Explain how the query would be executed in MapReduce (**not Pig**). Make sure to specify the computation performed in the map and the reduce functions. You do not need to show pseudocode. An English-language description will suffice.

2 Distribution and Replication

2. (20 points)

   (a) (10 points) In the two-phase commit protocol, describe what happens if a subordinate receives a PREPARE message, replies with a YES vote, crashes, and restarts.

(b) (10 points) Explain the benefits and challenges of asynchronous replication (also called lazy replication) in contrast to synchronous replication. Discuss both the configuration that uses a single master and one that uses multiple masters.