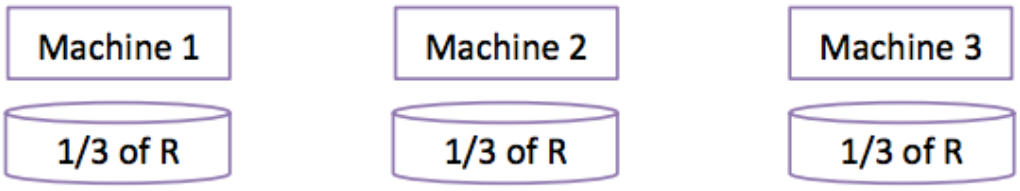
Section 10 – Big Data

CSE 344

Parallel Data Processing

Given the following query, show a (parallel) relational algebra plan for this query. There are 3 machines and the data is block-partitioned evenly across each machine.

```
SELECT a, max(b) as topb
FROM R
WHERE a > 0
GROUP BY a;
```



MapReduce

Suppose you have two relations: R(a,b) and S(b,c)
MapReduce needs (key, value) pairs as input, so we parse the above relations into such pairs.
R.a will be the key for each tuple in R. Imagine it as a map: { R.a → (R.a, R.b, tag="R") }
S.b will be the key for each tuple in S. Imagine it as a map: { S.b → (S.b, S.c, tag="S") }
For each relational plan below, write pseudocode for the Map and Reduce functions.

Select tuples from R: $\sigma_{a<10}R$

Eliminate duplicates from R: $\delta(R)$

Natural join of R and S: R |><|$_{R.b=S.b}$ S