# CSE 444: Homework 6

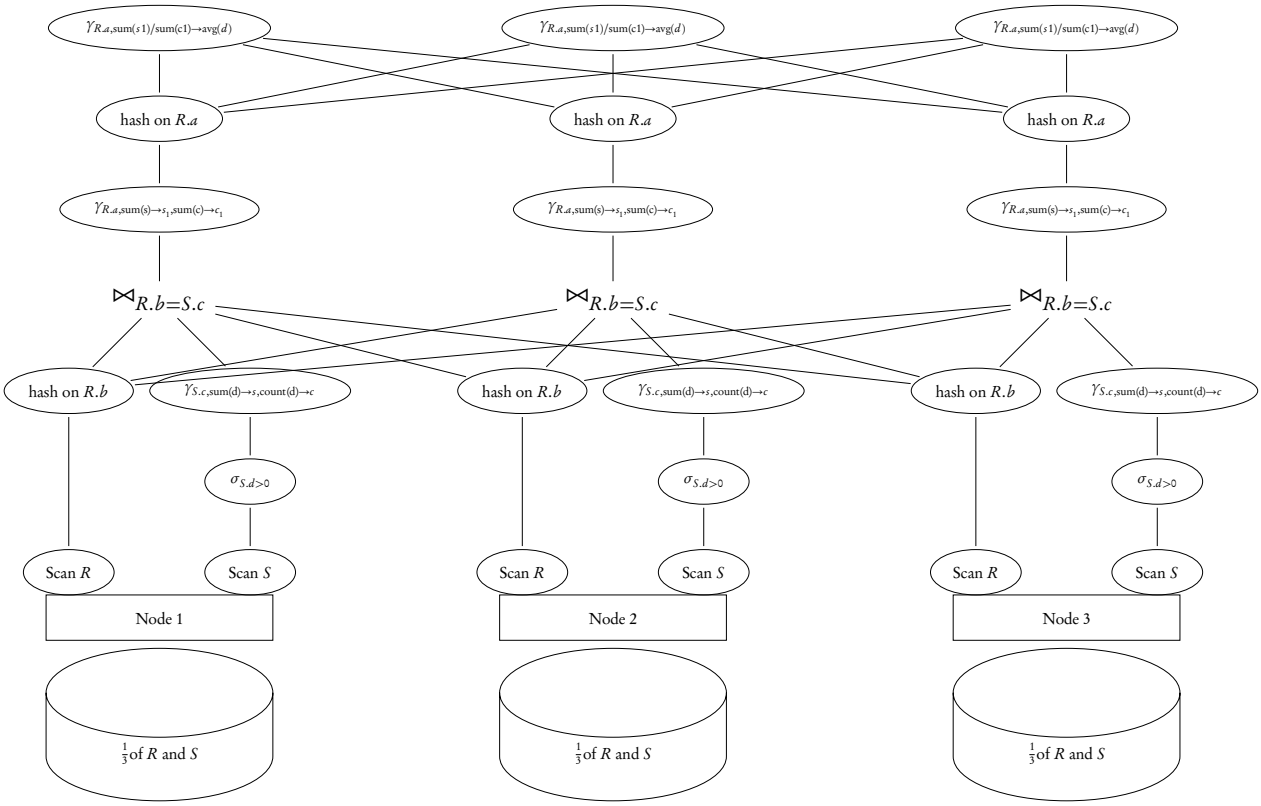
Linxing Preston Jiang Winter 2018

March 7, 2018

---

## 1 Parallel Data Processing

**1.** See below




**2.** We will MapReduce twice:

- First round:
  - Map: Map function on $R$: Use $R.b$ as key, write value as $('R', R.a)$; map function on $S$: First apply filter on $S.d > 0$, use $S.c$ as key, write value as $('S', S.d)$
  - Reduce: The same value of $R.b$ and $S.c$ as the key, perform a local join and output value is $(R.a, S.d)$
- Second round (MapReduce on $(R.a, S.d)$):
  - Map: Use $R.a$ as the key, output value as $(S.d)$
  - Reduce: Input is now $S.d$ with the same value of $R.a$. Output $R.a$ and $\frac{sum(S.d)}{count(S.d)}$ as the average.

## 2 Distribution and Replication

1. Subordinate will scan the log file and find there is PREPARE by no COMMIT/ABORT. It will keep asking the coordinator for a final decision. If it's commit, we redo the transaction. If it's abort, we undo the transaction

2.
- Single master
  - Asynchronus approach has better availability because it does not need to update all replicas together at the same time
  - Asynchronus approach has worse consistency because when the master of asynchronous approach fails, it might lose some recent write updates which haven't be sent to replicas
- Multi masters
  - Has the same better availability as an asynchronus approach, also allows more than one transactions to run together.
  - Asynchronus approach may introduce conflicts when different transactions write to different values of the same object on multiple replicas. Need conflicts detection and resolution.