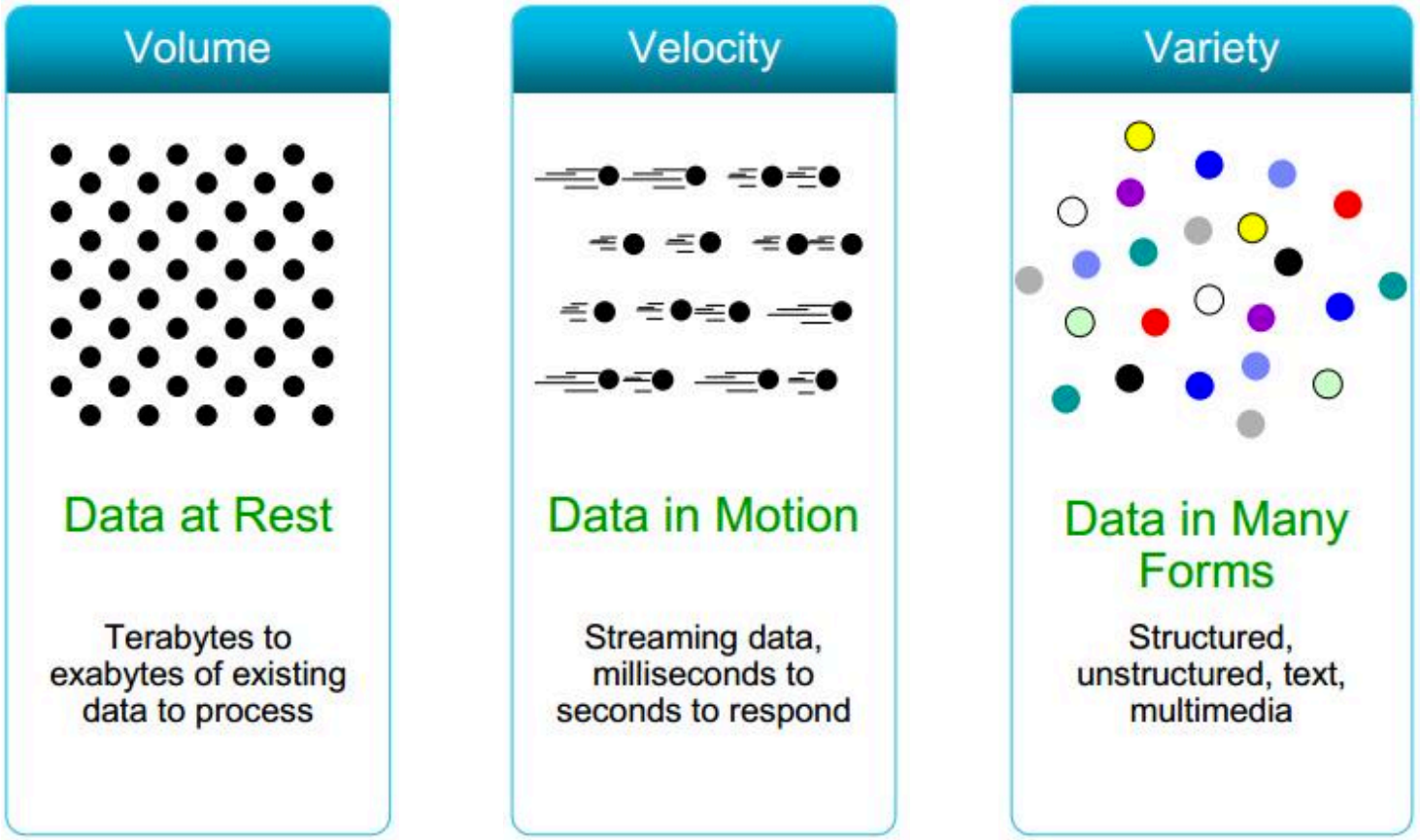
Parallel DBs & MapReduce

CSE 414 – SECTION 10

Big Data



The Three V's of Big Data

A Brief Story…

TECH 2/16/2012 @ 11:02AM | 2,691,904 views

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

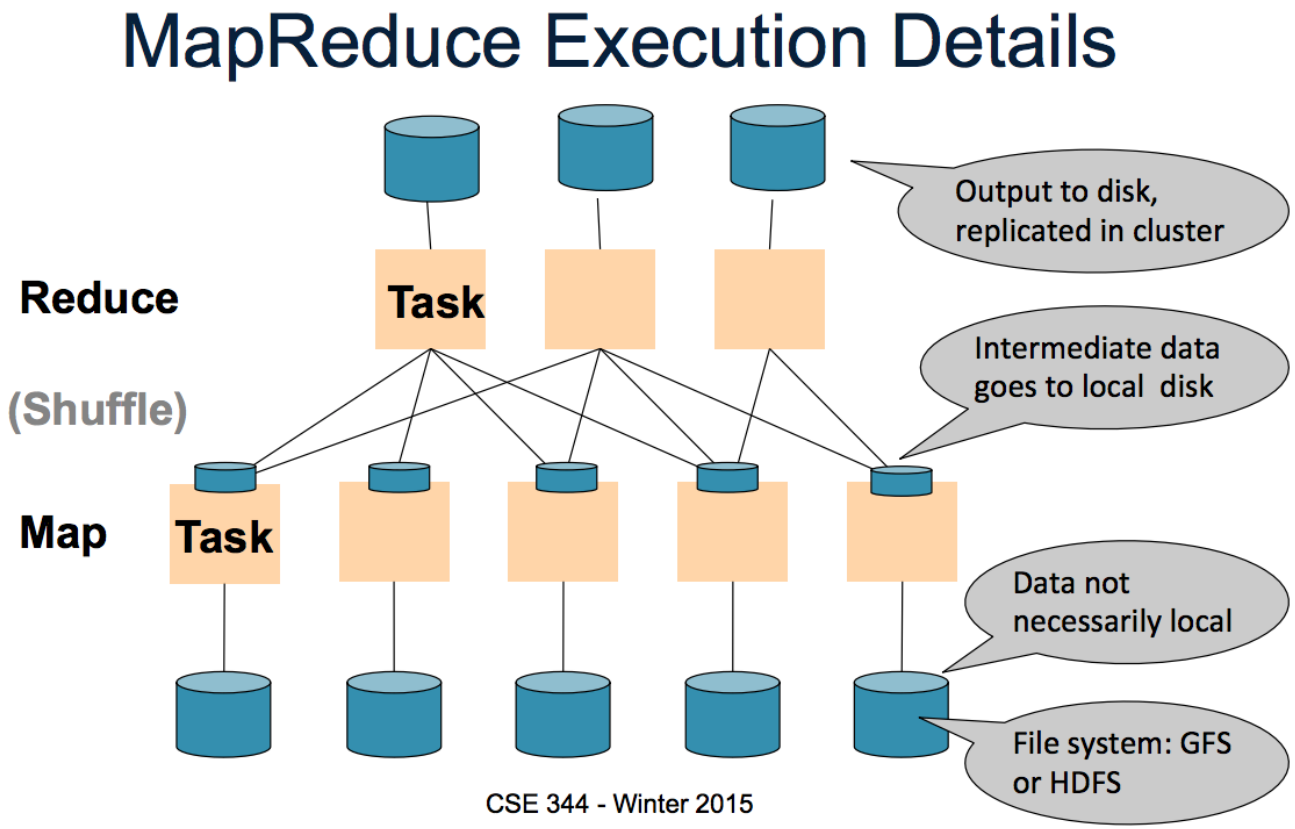
Predicting the future…

> As Pole's computers crawled through the data, he was able to identify about 25 products that, when analyzed together, allowed him to assign each shopper a "pregnancy prediction" score. More important, he could also estimate her due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy.

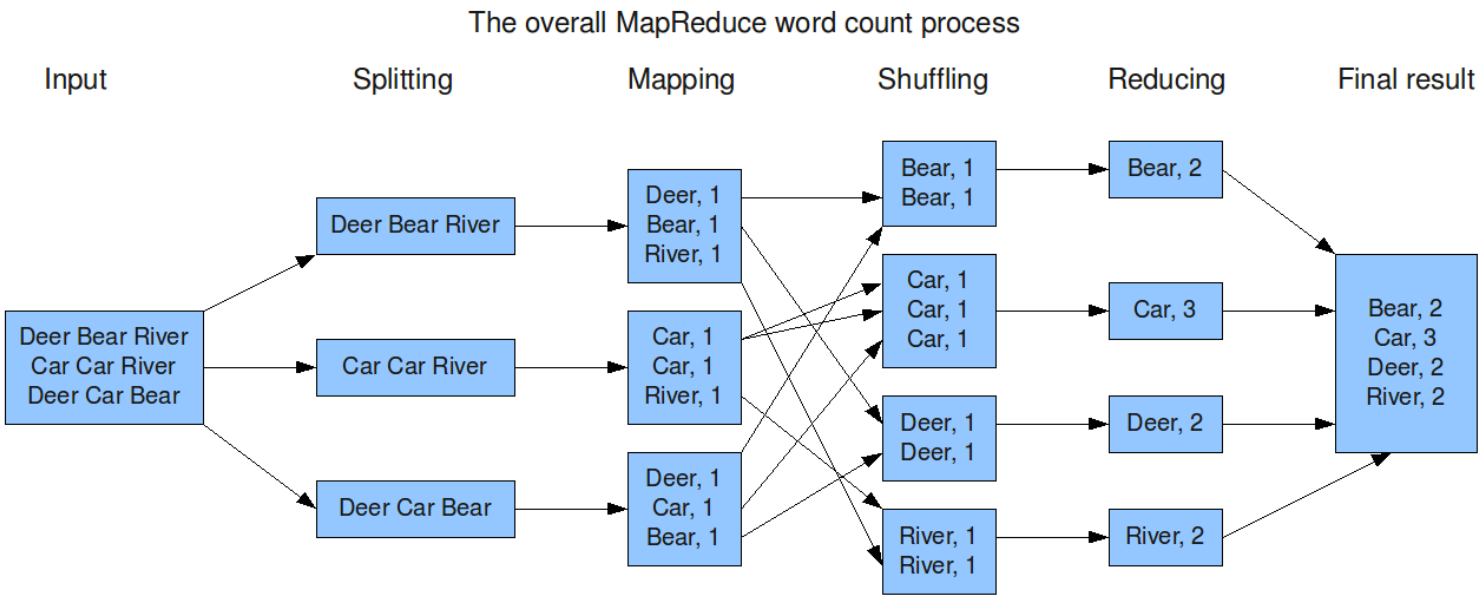
> "My daughter got this in the mail!" he said. "She's still in high school, and you're sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?"

> On the phone, though, the father was somewhat abashed. "I had a talk with my daughter," he said. "It turns out there's been some activities in my house I haven't been completely aware of. She's due in August. I owe you an apology."

And… back to CSE 414

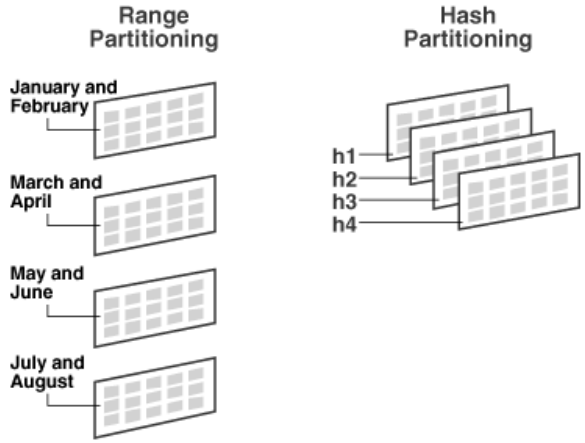


MapReduce Phases



Review of Parallel DBMSs

Block Partition vs. Range Partition vs. Hash Partition



Shuffling & Broadcasting

WordCount Example

```
Map(int id, String[] val)
        for word in val:
                emitIntermediate(word, 1)                    ←shuffling

 Reduce(String word, Iterator vals)
        cnt = 0
        for i in vals:
                cnt ++
        emit(cnt)
        // emit(word + ':' + cnt)
```