Lecture 31:
- Quiz 5
- Overview of R-5 (neural nets, resources, tensorflow, pytorch)
- SQL on MapReduce Example: Apache Hive (cont)

Announcements:
- R-4 due
- HW-5 out
- R-5 out soon
- Exam 2: Next Friday (4/19)
- Last Assignment: Out next Wednesday-ish

Hive Compiler

Q: How can we perform a join using MapReduce? ... group-by (shuffle) merge!

**Optimizer:** multiple passes over logical plan to build physical plan ...
- combine multiple joins sharing join key into single map-reduce job
- add repartition operations called ReduceSinkOperators
- each ReduceSinkOperator marks boundary between map and reduce phase
- added for each join, group-by, custom map-reduce script
- prune columns early to minimize data transfer
- push predicates to table scans to minimize data transfer
- prune partitions from input early (based on existing partitioning)
- prune buckets from input early (based on existing buckets)
Example (†): Generate tables using HiveSQL’s multi-table insert

- Facebook status updates and profile information
- To identify trends in status changes for a day by gender and school

```
FROM (  
    SELECT s.status, p.school, p.gender  
    FROM status_updates s JOIN profiles p ON (s.user_id = p.user_id)  
    WHERE s.ds = '2009-03-20'  
) s1

-- load school summary table
INSERT OVERWRITE TABLE school_summary PARTITION(ds='2009-03-20')
SELECT s1.school, COUNT(*)
GROUP BY s1.school

-- load gender summary table
INSERT OVERWRITE TABLE gender_summary PARTITION(ds='2009-03-20')
SELECT s1.gender, COUNT(*)
GROUP BY s1.gender
```

(†) From: Thusoo et al. “Hive – A Warehousing Solution Over a Map-Reduce Framework”, VLDB’09

© S. Bowers CPSC 324, Spring 2024 3