CS 686: Special Topics in Big Data

Discussion: MapReduce

Lecture 18

In the next few weeks…

- P2 will be posted
- We'll look at:
  - Storm
  - Spark
  - RDDs
- And we'll keep moving towards more dynamic programming models / computation techniques
- Analysis approaches, machine learning

Project 1 Grading

- Lots of interesting approaches for implementing the DFS!
- If you are missing a small piece of functionality or have a bug and want to fix it, talk to me first
  - Common: no list/disk space function
- Come on in at your appointment time:
  - Set up your cluster (10 Storage Nodes)
  - Prepare test files
- Note: we’re not limited to the exact ordering / test cases listed online!

Project 2

- I will post Project 2 this weekend
- This is a smaller project (15%), and less focused on development
  - We'll use Hadoop to analyze a large dataset
  - You'll write small MapReduce applications to learn more about the data, produce visualizations, etc.

MapReduce

- **Map**: filtering data, picking out the records you want
- **Reduce**: combining and summarizing the results

Tweets

- “Big Data problems solved in two steps – Map & Reduce”
- “Convention over configuration for distributed data processing”
- “Simple yet powerful programming interface which enables parallelization and distribution of large-scale computations”
- “If coding for fault tolerance and parallelization is too eerie put them under the MapReduce library”
- “Move computation! Don’t move data!”

Today's Discussion

- Groups of up to **4 people**
- One point raised in many of the evaluations was the suitability of the MapReduce model for certain problems…
- Come up with three problems that are well-aligned with MapReduce, and thee problems that are not
  - WordCount should not be included here ☺
- For problems that work well, you will demonstrate the workflow your group designs
- For problems that don't work well, you'll explain why

- Present @ 12:25
- Pick your favorite “good” case **OR** ”bad” case to present