Project Overview

CS 5531

(Rao)

Fall 2012

Apache Hadoop

- Project site: http://hadoop.apache.org
- Open-source, scalable, reliable distributed computing
- Distributed processing of large data sets (Big Data) on a large cluster of machines using a simple programming model
- Written in Java

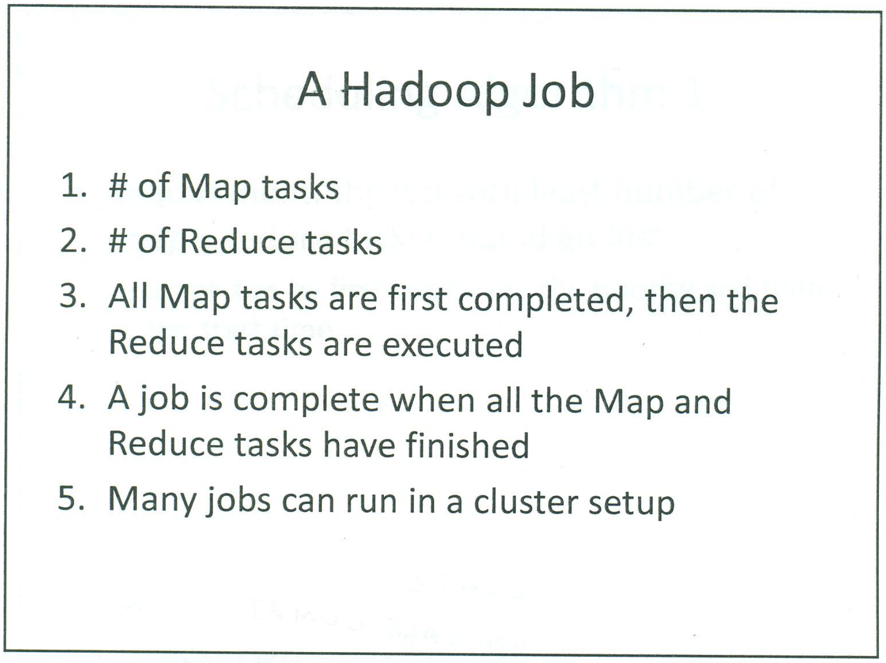
Project

- Phase 1 - setup and testing of Hadoop on IBM Cloud
- Phase 2 - modify the source code of Hadoop to add a new functionality; testing
- Phase 3 - performance evaluation on IBM Cloud
- Phase 4 - in-class presentation

Design a New Scheduling Algorithm for Apache Hadoop

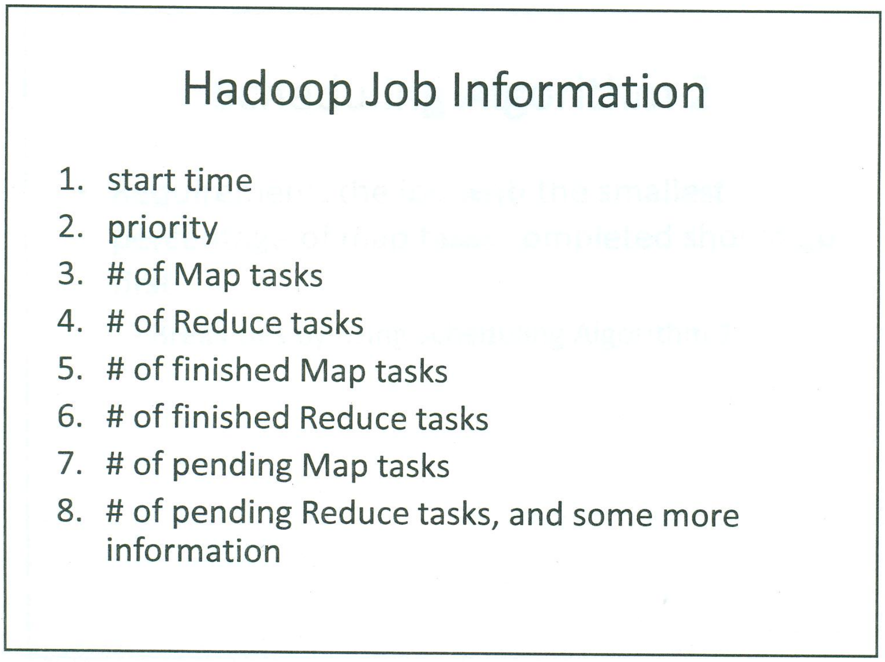
- Existing scheduling algorithms

1. First-come first-served (FIFO)
2. Shortest job first
3. Round robin
4. Priority scheduling
5. Fair-share scheduling



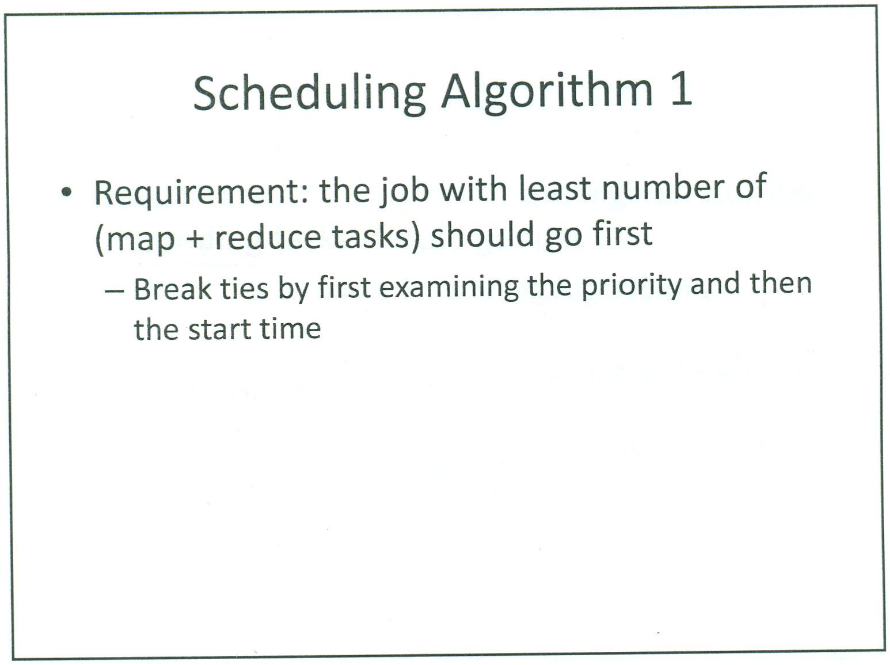
A Hadoop Job

1. # of Map tasks
2. # of Reduce tasks
3. All Map tasks are first completed, then the Reduce tasks are executed
4. A job is complete when all the Map and Reduce tasks have finished
5. Many jobs can run in a cluster setup



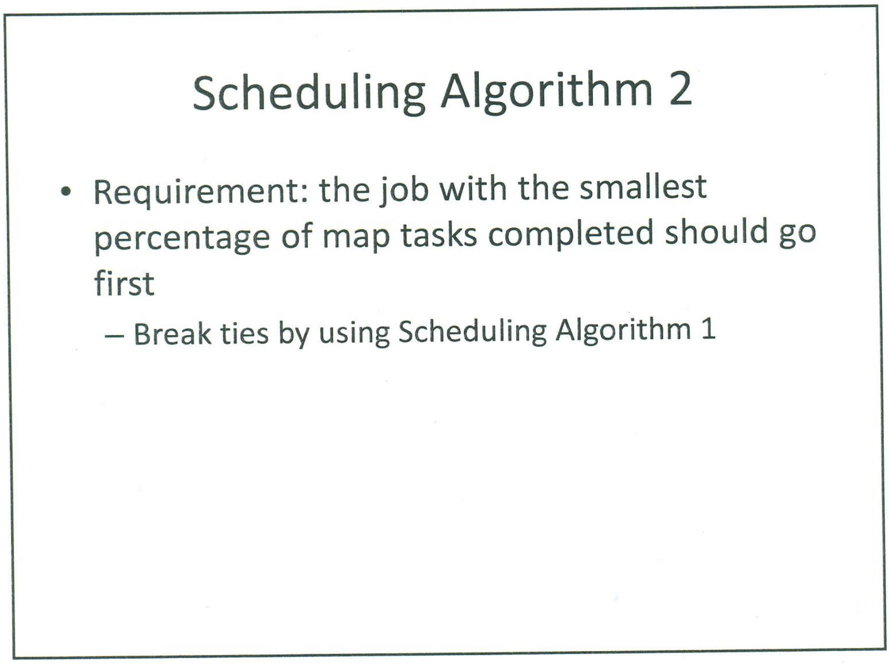
Hadoop Job Information

1. start time
2. priority
3. # of Map tasks
4. # of Reduce tasks
5. # of finished Map tasks
6. # of finished Reduce tasks
7. # of pending Map tasks
8. # of pending Reduce tasks, and some more information



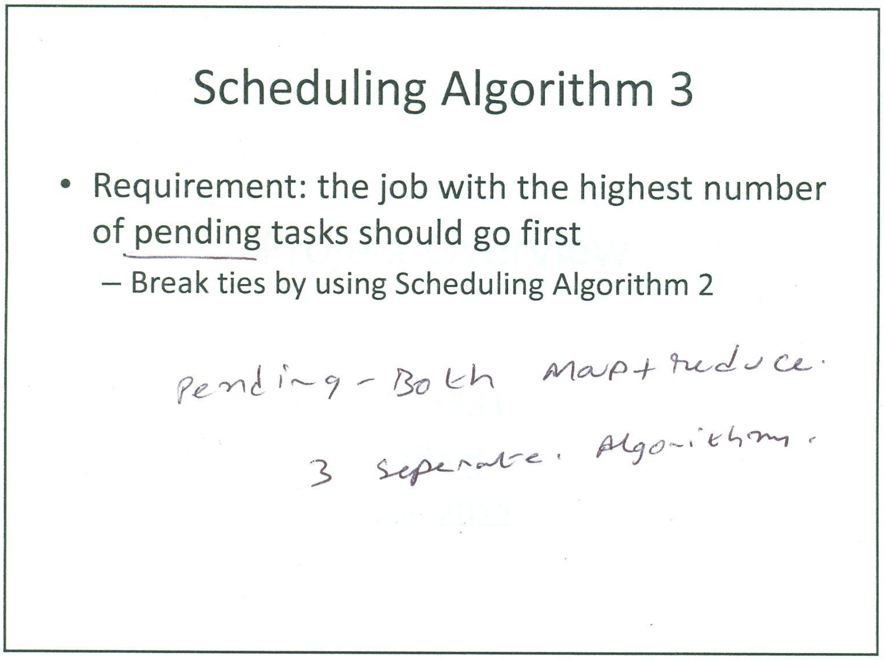
Scheduling Algorithm 1

- Requirement: the job with least number of (map + reduce tasks) should go first
  - Break ties by first examining the priority and then the start time



Scheduling Algorithm 2

- Requirement: the job with the smallest percentage of map tasks completed should go first
  - Break ties by using Scheduling Algorithm 1



Scheduling Algorithm 3

- Requirement: the job with the highest number of pending tasks should go first
  - Break ties by using Scheduling Algorithm 2

Pending - Both Map+reduce.

3 seperate. Algorithms.