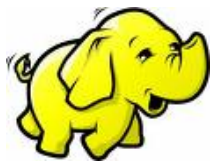
# Hadoop Capacity Scheduler

Rong-En Fan (rafan)
Yahoo! Search Engineering

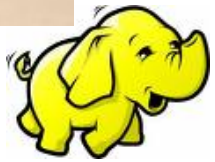
Hadoop Taiwan User Group meeting 2009



# A Large Hadoop Cluster

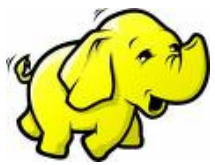
- Lots of computing power
- But... Provision? Resource share? Utilization?

*Yahoo! Hadoop Cluster*



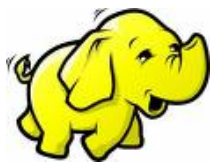
# Default Job Scheduler

- First-In-First-Output (FIFO) with priority support
- One job at a time
- Low utilization, monopoly

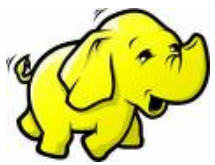
# Hadoop On Daemon (HOD)

- Divides cluster into many sub-clusters
- Nodes are dedicated to the requester
- Uses Torque/Maui for resources management, not easy to setup
- Not-so-good overall cluster utilization



# Pluggable Job Scheduler

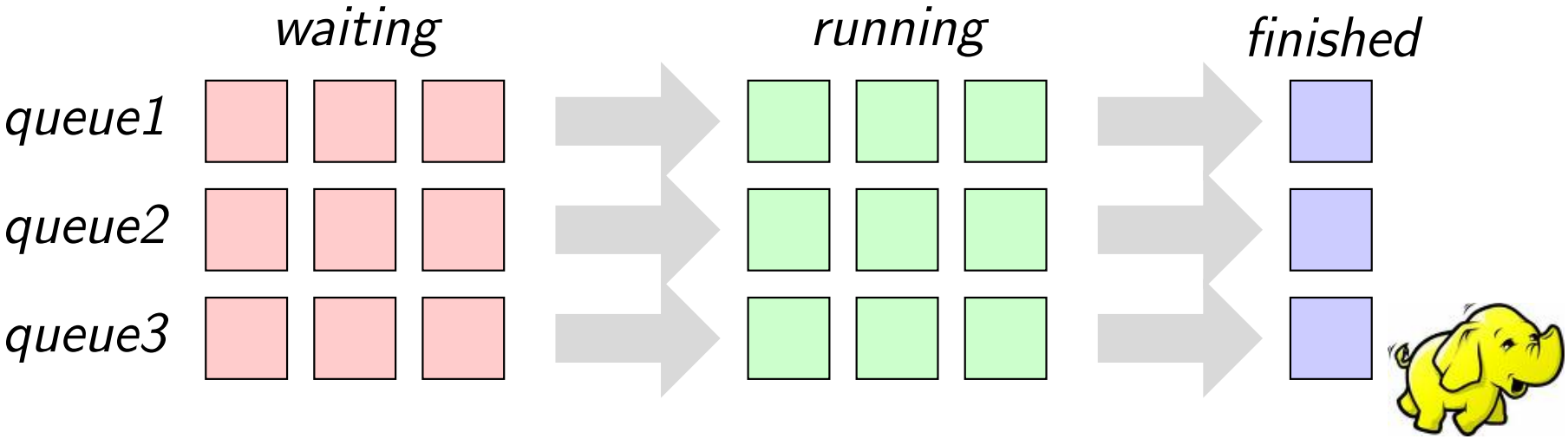
- A pluggable framework for job scheduling algorithm available since Hadoop 0.19
- Two new schedulers are born
  Capacity Scheduler by Yahoo!
  Fair Scheduler by Facebook



# Capacity Scheduler

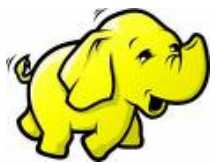
- Organizes jobs into queues
- Queue shares as %'s of cluster

  Optionally, can limit maximum resources per queue
- FIFO scheduling within each queue
- Contributed by Yahoo!



# Capacity Scheduler (cont'd)

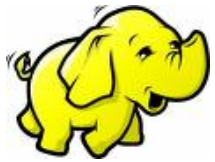
- Free resources are given to queues beyond its capacity
- Supports preemption[1]
- Supports memory-intensive jobs if job specifies memory requirement
- Supports job priorities; disabled by default
- Can enforce maximum resources used by a user per queue if there is competition



[1] Removed in 0.20.1 due to HADOOP-5726

# Job Scheduler Security

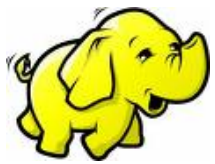
- Not necessary tied to Capacity Scheduler
- Each queue can have its own ACL control
- Tasks can be executed on the behalf of the users via `LinuxTaskController`[2]



[2] Available in 0.21 or Yahoo! Hadoop Distribution

# Installation

- Put `hadoop-*-capacity-scheduler.jar` to classpath
  - Modify `HADOOP_CLASSPATH` in `conf/hadoop-env.sh` *or*
  - Copy it to `lib/`



# Configuration Files

- Hadoop Config (`mapred-site.xml`)

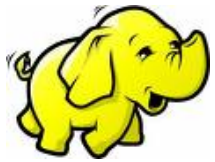
  Set

  `mapred.jobtracker.taskScheduler`

  to

  `org.apache.hadoop.mapred.CapacityTaskScheduler`

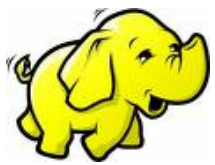
  Also define queues, ACL, etc.
- Queues' resources (`capacity-scheduler.xml`)



# Job Submission

- Passes queue name via `mapred.job.queue.name` property

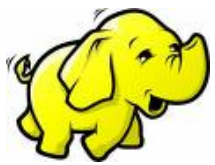
  `hadoop -Dmapred.job.queue.name=<name> ...`



# Define Queue Capacity

- Define two queues: *production* and *research*; each shares 75% and 25% of cluster, respectively

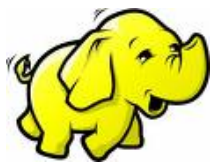
```
<?xml version="1.0"?>
<configuration>
  <property>
    <name>mapred.capacity-scheduler.queue.production.capacity</name>
    <value>75</value>
  </property>
  <property>
    <name>mapred.capacity-scheduler.queue.research.capacity</name>
    <value>25</value>
  </property>
</configuration>
```



# Limit Resources per User

- Queue *production*: each user has guaranteed 50% resource, at most 2 users running jobs at a time
- Queue *research*: each user has guaranteed 20% resource, at most 5 users running jobs at a time

```
<?xml version="1.0"?>
<configuration>
  <property>
    <name>
      mapred.capacity-scheduler.queue.production.minimum-user-limit-percent
    </name>
    <value>50</value>
  </property>
  <property>
    <name>
      mapred.capacity-scheduler.queue.research.minimum-user-limit-percent
    </name>
    <value>20</value>
  </property>
</configuration>
```

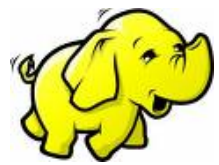


# Queue Information in Job Tracker UI

- The *Scheduling Information* section in Job Tracker UI

| | |
|---|---|
| [redacted] | Queue configuration<br>Capacity Percentage: 9.0%<br>User Limit: 20%<br>Priority Supported: NO<br>----------<br>Map tasks<br>Capacity: 332 slots<br>Used capacity: 7 (2.1% of Capacity)<br>Running tasks: 7<br>Active users:<br>User 'hadoopqa': 7 (100.0% of used capacity)<br>----------<br>Reduce tasks<br>Capacity: 166 slots<br>Used capacity: 1 (0.6% of Capacity)<br>Running tasks: 1<br>Active users:<br>User 'hadoopqa': 1 (100.0% of used capacity)<br>----------<br>Job info<br>Number of Waiting Jobs: 0<br>Number of users who have submitted jobs: 1 |

*Queue Configuration*



# Running Jobs in a Queue

- Click *queue name* in the Job Tracker UI, *Scheduling Information* section

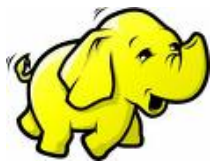
**Job Summary for the Queue :: [redacted]**

*(In the order maintained by the scheduler)*

| Jobid | Priority | User | Name | Map % Complete | Map Total | Maps Completed | Reduce % Complete | Reduce Total | Reduces Completed | Job Scheduling Information |
|---|---|---|---|---|---|---|---|---|---|---|
| job_200911040612_19339 | NORMAL | hadoopqa | | 5.00% | 20 | 1 | 0.00% | 1 | 0 | 19 running map tasks using 19 map slots. 0 additional slots reserved. 1 running reduce tasks using 1 reduce slots. 0 additional slots reserved. |

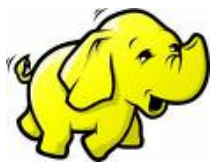
This release is based on the Yahoo! Distribution of Hadoop, powering the largest Hadoop clusters in the Universe!

*Running Jobs*



# Fair Scheduler

- Different philosophy, but similar in functionalty
- A little bit more flexible on configurating resource limit
- Contributed by Facebook

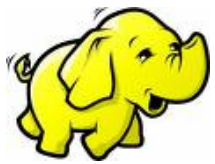


# Reference

- Hadoop Documentations

  Cluster Setup

  Capacity Scheduler Guide
- Job Scheduling with the Fair and Capacity Schedulers by Matei Zaharia, Hadoop Summit 2009
- Cloudera Blog: Job Scheduling in Hadoop



# Questions?