Capacity Scheduler Guide

Table of contents

1. Purpose

This document describes the Capacity Scheduler, a pluggable Map/Reduce scheduler for Hadoop which provides a way to share large clusters.

2. Features

The Capacity Scheduler supports the following features:

- Support for multiple queues, where a job is submitted to a queue.
- Queues are guaranteed a fraction of the capacity of the grid (their 'guaranteed capacity') in the sense that a certain capacity of resources will be at their disposal. All jobs submitted to a queue will have access to the capacity guaranteed to the queue.
- Free resources can be allocated to any queue beyond its guaranteed capacity. These excess allocated resources can be reclaimed and made available to another queue in order to meet its capacity guarantee.
- The scheduler guarantees that excess resources taken from a queue will be restored to it within N minutes of its need for them.
- Queues optionally support job priorities (disabled by default).
- Within a queue, jobs with higher priority will have access to the queue's resources before jobs with lower priority. However, once a job is running, it will not be preempted for a higher priority job.
- In order to prevent one or more users from monopolizing its resources, each queue enforces a limit on the percentage of resources allocated to a user at any given time, if there is competition for them.
- Support for memory-intensive jobs, wherein a job can optionally specify higher memory-requirements than the default, and the tasks of the job will only be run on TaskTrackers that have enough memory to spare.

3. Picking a task to run

Note that many of these steps can be, and will be, enhanced over time to provide better algorithms.

Whenever a TaskTracker is free, the Capacity Scheduler first picks a queue that needs to reclaim any resources the earliest (this is a queue whose resources were temporarily being used by some other queue and now needs access to those resources). If no such queue is found, it then picks a queue which has most free space (whose ratio of # of running slots to guaranteed capacity is the lowest).

Once a queue is selected, the scheduler picks a job in the queue. Jobs are sorted based on

when they're submitted and their priorities (if the queue supports priorities). Jobs are considered in order, and a job is selected if its user is within the user-quota for the queue, i.e., the user is not already using queue resources above his/her limit. The scheduler also makes sure that there is enough free memory in the TaskTracker to tun the job's task, in case the job has special memory requirements.

Once a job is selected, the scheduler picks a task to run. This logic to pick a task remains unchanged from earlier versions.

4. Reclaiming capacity

Periodically, the scheduler determines:

- if a queue needs to reclaim capacity. This happens when a queue has at least one task pending and part of its guaranteed capacity is being used by some other queue. If this happens, the scheduler notes the amount of resources it needs to reclaim for this queue within a specified period of time (the reclaim time).
- if a queue has not received all the resources it needed to reclaim, and its reclaim time is about to expire. In this case, the scheduler needs to kill tasks from queues running over capacity. This it does by killing the tasks that started the latest.

5. Installation

The capacity scheduler is available as a JAR file in the Hadoop tarball under the *contrib/capacity-scheduler* directory. The name of the JAR file would be on the lines of hadoop-*-capacity-scheduler.jar.

You can also build the scheduler from source by executing *ant package*, in which case it would be available under *build/contrib/capacity-scheduler*.

To run the capacity scheduler in your Hadoop installation, you need to put it on the *CLASSPATH*. The easiest way is to copy the `hadoop-*-capacity-scheduler.jar` from to `HADOOP_HOME/lib`. Alternatively, you can modify *HADOOP_CLASSPATH* to include this jar, in `conf/hadoop-env.sh`.

6. Configuration

6.1. Using the capacity scheduler

To make the Hadoop framework use the capacity scheduler, set up the following property in the site configuration:

Property	Value
mapred.jobtracker.taskScheduler	org.apache.hadoop.mapred.CapacityTaskScheduler

6.2. Setting up queues

You can define multiple queues to which users can submit jobs with the capacity scheduler. To define multiple queues, you should edit the site configuration for Hadoop and modify the *mapred.queue.names* property.

You can also configure ACLs for controlling which users or groups have access to the queues.

For more details, refer to Cluster Setup documentation.

6.3. Configuring properties for queues

The capacity scheduler can be configured with several properties for each queue that control the behavior of the scheduler. This configuration is in the *conf/capacity-scheduler.xml*. By default, the configuration is set up for one queue, named *default*.

To specify a property for a queue that is defined in the site configuration, you should use the property name as *mapred.capacity-scheduler.queue.<queue-name>.<property-name>*.

For example, to define the property *guaranteed-capacity* for queue named *research*, you should specify the property name as *mapred.capacity-scheduler.queue.research.guaranteed-capacity*.

The properties defined for queues and their descriptions are listed in the table below:

Name	Description
mapred.capacity-scheduler.queue.<queue-name>	Percentage of the number of slots in the cluster that are guaranteed to be available for jobs in this queue. The sum of guaranteed capacities for all queues should be less than or equal 100.
mapred.capacity-scheduler.queue.<queue-name>	The amount of time, in seconds, before which resources distributed to other queues will be reclaimed.
mapred.capacity-scheduler.queue.<queue-name>	If true, priorities of jobs will be taken into account in scheduling decisions.
mapred.capacity-scheduler.queue.<queue-name>	Each queue enforces a limit on the percentage of resources allocated to a user at any given time, if there is competition for them. This user

	limit can vary between a minimum and maximum value. The former depends on the number of users who have submitted jobs, and the latter is set to this property value. For example, suppose the value of this property is 25. If two users have submitted jobs to a queue, no single user can use more than 50% of the queue resources. If a third user submits a job, no single user can use more than 33% of the queue resources. With 4 or more users, no user can use more than 25% of the queue's resources. A value of 100 implies no user limits are imposed.

6.4. Configuring the capacity scheduler

The capacity scheduler's behavior can be controlled through the following properties.

Name	Description
mapred.capacity-scheduler.reclaimCapacity.interv	The time interval, in seconds, between which the scheduler periodically determines whether capacity needs to be reclaimed for any queue. The default value is 5 seconds.

6.5. Reviewing the configuration of the capacity scheduler

Once the installation and configuration is completed, you can review it after starting the Map/Reduce cluster from the admin UI.

- Start the Map/Reduce cluster as usual.
- Open the JobTracker web UI.
- The queues you have configured should be listed under the *Scheduling Information* section of the page.
- The properties for the queues should be visible in the *Scheduling Information* column against each queue.