myHadoop 0.2a: Hadoop-on-demand on Traditional HPC Resources

Sriram Krishnan, Ph.D.
San Diego Supercomputer Center, University of California at San Diego
sriram@sdsc.edu

1. Introduction

Traditional HPC environments typically support batch job submissions using resource management systems such as the TORQUE Resource Manager (also known as the Portable Batch System – PBS) or the Sun Grid Engine (SGE). On the other hand, Hadoop provides it own scheduling, and manages its own job and task submissions, and tracking. Since both systems are designed to have complete control over the resources that they manage, the challenge is how to enable users to run Hadoop jobs in a typical HPC environment using a scheduler such as PBS or SGE. In this release, we support Hadoop job submissions via PBS and SGE. However, this approach is equally feasible for other schedulers such as Condor, as well.

Our approach is to configure Hadoop clusters "on-demand" by first requesting resources for an N-node Hadoop cluster via PBS. Once the resources are received, the Hadoop configurations and environments are set up based on the set of resources provided by PBS. The Hadoop Distributed File System (HDFS) can be configured in one of two ways – in 1) transient (non-persistent) or 2) persistent modes. In the non-persistent mode, the HDFS is set up to use local storage. In the persistent mode, the HDFS is set to symbolically link to an external location that will be persistent – i.e. data from Hadoop runs will continue to persist even after the Hadoop runs are complete. More details are as follows.

2. Details

The pre-requisite for myHadoop is a valid Hadoop installation – we recommend that you use Hadoop version 0.20.2 since that is the only version of Hadoop that this package has been tested with. Henceforth, we will refer to the location of the Hadoop installation as *HADOOP_HOME*. We will refer to the location of the myHadoop installation (i.e. this package) as *MY_HADOOP_HOME*. The *$MY_HADOOP_HOME/pbs-example.sh* shows an example of how to use myHadoop with PBS. A similar script for SGE can be found in *$MY_HADOOP_HOME/sge-example.sh*.

A step-by-step process for using myHadoop is as follows.

2.1. Initial Configuration

Ensure that the environment variables inside *$MY_HADOOP_HOME/bin/setenv.sh* are set correctly. You can set your *HADOOP_HOME,* and the locations for your HDFS data and log directories using this script. ***You will need to update this script before you can proceed further***.

All the tuning parameters for Hadoop can be found in the *$MY_HADOOP_HOME/etc* directory. There is no need to edit any of the parameters, especially if you are not an expert Hadoop user. If you are familiar with the various Hadoop parameters, you may edit the parameters that fall outside the "*DO NOT EDIT*" sections.

2.2. Request *N* nodes from the Scheduler

Once the environment variables have been set correctly, we are ready to use myHadoop using a regular PBS or SGE submission script. Your PBS script should contain the following lines to initialize PBS as follows:

```
#!/bin/bash

#PBS -q <queue_name>
#PBS -N <job_name>
#PBS -l nodes=4:ppn=1
#PBS -o <output file>
#PBS -e <error_file>
#PBS -A <allocation>
#PBS -V
#PBS -M <user email>
#PBS -m abe
```

In the above case, we are requesting 4 nodes. Note that you must set the processors per node (ppn) to 1.

Your SGE script should contain the following lines to initialize SGE:

```
#!/bin/bash

#$ -V -cwd
#$ -N <job_name>
#$ -pe <queue_name> 4
#$ -o <output file>
#$ -e <error file>
#$ -S /bin/bash
```

For SGE, there is one important rule to remember. The queue name specified above should be pre-configured with an ***allocation_rule*** set to ***1*** (one). This ensures that the

Hadoop cluster is set up such that multiple instances of the Hadoop daemons are not scheduled on the same node.

2.3. Set the myHadoop Environment

Run the *$MY_HADOOP_HOME/bin/setenv.sh* script (that you modified in Section 2.1) to set all the environment variables required by myHadoop.

. ***$MY_HADOOP_HOME/bin/setenv.sh***

Set the HADOOP_CONF_DIR to the directory where Hadoop configs should be generated – all configuration files for the Hadoop run will be picked up from here. Ensure that this directory is accessible to all nodes – and a way to do this is to make sure that this directory is on a shared file system such as NFS or Lustre.

*export HADOOP_CONF_DIR=**<configuration directory>***

2.4. Configure the myHadoop Cluster

You can initialize and configure the Hadoop cluster by using the *$MY_HADOOP_HOME/bin/pbs-configure.sh* (or *sge-configure.sh)* script. You may create a transient or persistent myHadoop cluster by changing the command-line arguments as follows.

For a transient myHadoop cluster, configure it as follows (replace 4 with the total number of nodes requested):

$MY_HADOOP_HOME/bin/pbs-configure.sh -n 4 -c $HADOOP_CONF_DIR

In this mode, you will have to copy all of your data into the myHadoop cluster after it is configured, and copy out the results after the job is complete. All data will be inaccessible from HDFS once the PBS job is complete.

Alternatively, you may set up a persistent myHadoop cluster by using the ***–p*** option, and setting the ***BASE_DIR*** for HDFS as follows:

*$MY_HADOOP_HOME/bin/pbs-configure.sh -n 4 -c $HADOOP_CONF_DIR -p -d **<HDFS BASE_DIR>***

The BASE_DIR should be on a directory accessible to all nodes, to ensure that the data will not be cleaned up after job completion. For instance, the BASE_DIR could be on a Lustre file system. Note that, if ***N-node*** cluster is being created, then the BASE_DIR should have directories named ***1, 2, … , N***. The configuration script sets up symbolic links from node ***I*** to the ***BASE_DIR/I*** directory. When this mode is used, there is no need to copy data back and forth from HDFS to another file system between runs.

2.5. Format HDFS (if need be)

If myHadoop is being used in transient mode, or if it is being used for the first time in persistent mode, then you will have to format the HDFS as follows:

$HADOOP_HOME/bin/hadoop --config $HADOOP_CONF_DIR namenode –format

2.6. Run Hadoop Jobs

You are now all set to start all the Hadoop daemons as follows:

$HADOOP_HOME/bin/start-all.sh

Once the daemons are all started up, you can start using Hadoop as usual. You may also stage data in and out from HDFS, as required.

2.7. Clean up

Although, PBS or SGE may be set up to automatically clean up after your Hadoop job is complete, it is always a good idea to stop all the Hadoop daemons, and use the clean-up script to clean up after yourself.

$HADOOP_HOME/bin/stop-all.sh
$MY_HADOOP_HOME/bin/pbs-cleanup.sh -n 4 OR
$MY_HADOOP_HOME/bin/sge-cleanup.sh -n 4