myHadoop: Hadoop-on-demand via Traditional Schedulers

Sriram Krishnan
sriram@sdsc.edu

1. Introduction

Traditional HPC environments typically support batch job submissions using environments 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. The Triton and Dash resources at SDSC support PBS – hence, we will limit our discussion to PBS in this document. However, this approach is equally feasible for other schedulers such as SGE, 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 or 2) persistent modes. In the transient mode, the HDFS is set up to use local storage (in the case of Dash, we use Flash storage). In the persistent mode, the HDFS is set to symbolically link to an external location that will be persistent – i.e. will not be cleaned after the Hadoop run is complete. We provide more details on these two modes as follows.

2. Details

Hadoop is installed at */opt/hadoop/hadoop-0.20.2/* on Triton, and at */usr/apps/utilities/hadoop-0.20.2/* on Dash. Henceforth, we will refer to this location as *HADOOP_HOME*. The myHadoop component is at *$HADOOP_HOME/contrib/myHadoop* on both installations, which we will refer to as *MY_HADOOP_HOME*. An example of how to use myHadoop is provided in *$MY_HADOOP_HOME/example.sh*.

For both the modes, the job submission is via a PBS batch script. The individual steps for using myHadoop are as follows:

2.1 Request *N* nodes from PBS

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 – you also have to ensure that the processors per node (ppn) are set to 1.

2.2 Set the Hadoop Configuration Directory

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>***

2.3 Configure the myHadoop Cluster

You can initialize and configure the Hadoop cluster by using the *$MY_HADOOP_HOME/bin/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/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 purged 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/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, such that the data will not be cleaned up after job completion. For instance, the BASE_DIR could be on Data Oasis. Furthermore, 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.

2.3 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

2.3 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.

2.4 Clean up

Although, PBS 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/cleanup.sh -n 4