Hadoop External JARs

When Hadoop jobs get run on a cluster, all of the dependencies must be available on each node so the code can actually run. In the simplest case, that is just your code (which is automatically distributed by yarn jar) and the core Hadoop classes (which are automatically on every node).

If your application depends on JARs outside of those, you have to tell YARN to actually distribute them to the nodes doing work.

In your shell, you need to set the HADOOP\_CLASSPATH variable with a **colon** separated list of JARs:

export HADOOP\_CLASSPATH=/path/to/jar1.jar:/path/to/jar2.jar

Then when submitting the job, you give a **comma** separated list of JARs as the -libjars argument:

${HADOOP\_HOME}/bin/yarn jar project.jar MainClass \

-libjars /path/to/jar1.jar,/path/to/jar2.jar inputfiles outputfiles