. #Caution! rmr 1.3 has been released and the wiki documentation is being updated. Thanks for your patience. In particular the Tutorial, Getting data in and out, Efficient rmr techniques and Writing Composable Mapreduce jobs should be considered outdated.

Overview

This R package allows an R programmer to perform statistical analysis via MapReduce on a Hadoop cluster.

Prerequisites and installation

- A Hadoop cluster, CDH3 and higher or Apache 1.0.2 and higher but limited to mr1, not mr2. For more details on Hadoop compatibility see see [[Which Hadoop for rmr]].
- R installed on each node of the cluster (developed and tested on R 2.14.1). Revolution R Community 4.3 or 5.0 can be used, if you upgrade to RJSONIO 0.95 (which must be downloaded from CRAN, as it is not available in the REVO 2.12 repository) and create a symbolic link from /usr/bin/Revoscript to /usr/bin/Rscript.
- Install the following R packages on each node: RJSONIO (0.95-0 or later recommended), itertools and digest
- rmr itself needs to be installed on each node.
- Make sure that the packages are installed in a default location accessible to all users (R will run on the cluster as a different user from the one who has started the R interpreter where the mapreduce calls have been executed)
- Make sure that the environment variables `HADOOP_CMD` and `HADOOP_STREAMING` are properly set. For some distributions, `HADOOP_HOME` is still sufficient for R to find everything that's needed so if that works for you you can keep it that way.

Examples:

```
export HADOOP_CMD=/usr/bin/hadoop
export HADOOP_STREAMING=/usr/lib/hadoop/contrib/streaming/hadoop-streaming-<version>.jar
```

For people who use RPMs for their deployments, courtesy of jseidman, we have RPMs for rmr and its dependencies (digest, iterators, itertool, rjsonio). These RPMs are available in this repo: https://github.com/jseidman/pkgs. Note that currently there's only CentOS 5.5 64bit RPMs, but the source files to create the RPMs are in the same repo, so it should be easy to build for other RH-based distros. jseidman reports using RPMs along with Puppet to deploy all packages, applications, etc. to their (Orbitz) Hadoop clusters.

For people who use EC2 (not EMR), in the source package under the tools directory, a whirr script to fire up an EC2 rmr cluster.

If you use Globus Provision, check out this https://github.com/nbest937/gp-rhadoop (very alpha as of this edit), courtesy nbest.

Contents

- [[Changelog]]
- [[Design Philosophy]]
- [[Tutorial]]
- [[Debugging rmr programs]]
- [[Efficient rmr techniques]]
- [[Writing composable mapreduce jobs]]
- [[Use cases]]
- [[Getting data in and out]]
- [[FAQ]]