Supersize Your Analysis: Cluster Computing with R and Condor

Shukry Zablah

July 06, 2018

Introduction

R is a powerful language that provides multiple functions for every step of the data analysis process. Students have the ability to stretch their skills as far as the R libraries go and to create their own functions to analyze their data. However, in every big analysis the overarching limitation is computing power.

Enter computer clusters into the picture – an idea that has been around for more than three decades. You can think of computer clusters as big machines made up of smaller units. The smaller units are the machines we usually are familiar with, like a laptop. We are able to treat the big machines as a single entity and make use of their extensive computing power.

At Amherst College we have a couple of computer clusters, including one with a scheduling software called Condor and another one running Spark. We will try to leverage these computer clusters to supersize our analysis.

Learning Goals

1. Deploy a simple R script to computer cluster
2. Setup and run a job in Condor

Our R script

```
#!/usr/bin/env Rscript
args = commandArgs(trailingOnly=TRUE)
if(length(args) != 1) {
  stop("Usage: Rscript --vanilla square.R <number>")
} else {
  num <- as.numeric(args[1])
  print(paste0("Number: ", num))
  print(paste0("Square: ", num^2))
}
```

Steps

1. Log in to Condor
2. Change directories to cluster-scratch
3. Use mkdir to create new folder for your project.
4. Use scp to copy file from your computer to the project folder.
5. Use chmod to make the R script executable.
6. Use the Python command file maker to make your project.cmd file
7. Use condor_submit to queue your job.