

Outline for “Computing with R and Hadoop”

This section will last approximately 40 minutes. The first 20-25 minutes will introduce Hadoop and how it is used with R, and the remaining 15-20 minutes will walk participants through a demonstration lab.

Learning goals

1. What Hadoop is
2. Current R/Hadoop integrations
3. *When* to use R with Hadoop (guidelines)
4. *How* to use R with Hadoop (lab)

Extra materials (helpful, but not required)

1. Slides for this section’s presentation: <http://goo.gl/wmX6Xd>
2. Virtual machine used to run R/Hadoop for the lab (requires ~4GB RAM): <http://goo.gl/R5Okcr>
3. R script for lab: <http://goo.gl/dBaxrq>
4. Lab instructions: <http://goo.gl/jPSqOK>
5. PuTTY (SSH client for Windows): <http://goo.gl/vMv6ra>

Key definitions

Hadoop: Open source software for enabling distributed storage and computing capabilities on networked servers.

MapReduce: Hadoop’s

R: Open source statistical computing software designed with strong built-in support for statistical needs like fitting models, making predictions, and drawing inferences.

R/Hadoop integration: Extra R packages that let practitioners run R code in Hadoop MapReduce jobs.

Key ideas

1. R and Hadoop integrate best when using the strengths of both technologies
2. R and Hadoop do not integrate well for all projects. Simple data processing and iterative algorithms may be best implemented with other technologies.
3. Hadoop facilitates distributed computing with the MapReduce programming paradigm