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:
2. Virtual machine used to run R/Hadoop for the lab (requires ~4GB RAM): http://goo.gl/8ZjXUZ
3. R script for lab:

Key definitions

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

MapReduce: Hadoop's

R: Popular 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 technolgoies.
3. Hadoop facilitates distributed computing with the MapReduce programming paradigm

Lab outline

GOAL: Make a data-driven business decision for an insurance company to start a new program.

1. Fit a logistic regression model to a small amount of data obtained from a pilot study of customer feedback data.
2. In a virtual cloud, use the `RHadoop` package to run a MapReduce job that summarizes which other customers the model predicts will provide positive feedback.
3. (Optional) Modify the MapReduce job to refine the analysis.

Monitor Hadoop:

- Open “The Hadoop UI” (HUE): http://192.168.1.105:8888/
- Username and password: `cloudera`
- View status of MapReduce jobs: http://192.168.1.105:8888/jobbrowser/
- View contents of **hdfs**: http://192.168.1.105:8888/filebrowser/#/