Lab instructions for "Computing with R and Hadoop"

This lab uses a simplified scenario to illustrate how R and Hadoop can be applied to practical problems.

Instructions

1. Read the problem/scenario description below.
2. Read the solution sketch below.
3. Download the R script with the code to run the lab: http://goo.gl/dBaxrq
4. Connect to `TheCantina` wireless network.
5. Use a shell/terminal window (Mac or *nix) or PuTTY (Windows) to connect to Hadoop and open R.
   - Log in to a server that can submit MapReduce jobs to Hadoop:
     `ssh cloudera@192.168.1.105` (password: `cloudera`)
   - Start R:
     `R`
6. Open the Hadoop monitoring pages (links at end of document).
7. Run the R script to execute the code.
8. Review code and output.
9. (Optional) Modify pieces of the code to change the analysis.

Problem/Scenario

A car insurance company launched a **small pilot study** to evaluate a new program they are considering offering to all of their customers. At the end of the study the participants were asked whether or not they would like to stay enrolled in the offering.

- The company would like to use the participants' **demographic information** and their feedback to help **predict** whether the program can be **profitable** if offered to all customers.
- For marketing purposes, they are additionally interested in knowing if the program is very popular with specific subsets of their customers.

Solution sketch

1. Analyze study data
   - Fit a logistic regression model to the study data; regress the participants' feedback (whether they would like to stay enrolled in the program) on their demographic information.
2. Make predictions **(in parallel)**
   - Use the "'RHadoop"' package to run a MapReduce job that first uses the logistic regression model to predict whether each of the company's other customers would like the new offering, and then summarizes the predictions by gender and region.
3. Evaluate/interpret predictions
   - Use the "'RHadoop"' package to retrieve the summarized predictions
4. (Optional) Modify the MapReduce job to refine the analysis.

Note: The R script for this lab is broken into sections with code to implement each of these general steps.

Ideas for modifications or variations

- Use a different regression/classification model
- Compute more detailed summaries of customer predictions
  - E.g., Group by gender, age, and region
  - E.g., Work with raw estimates and compute averages or variances instead of percentages
- Identify model weaknesses
  - What factor levels are present in customer records but not in pilot records?
  - Which or how many customers have these types of levels?
- Explore functions and objects in `rmr2`
  - Look at `mapred.result`
  - Write data to **hdfs** in different formats

Lab materials and resources

R script for lab (on the internet):

- http://goo.gl/dBaxrq

Webpages to monitor Hadoop system and parallel computation status (on `TheCantina` local network):

- 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/#/