Computing with R and Hadoop

November 19, 2014

Learning goals

This section should teach participants:

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)

Brief introduction to Hadoop

Hadoop

Key ideas: enables distributed computing; open source; widely used

> *The Apache Hadoop project develops open-source software for reliable, scalable, distributed computing. The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage... Source: Apache Software Foundation - What is Apache Hadoop?*

Hadoop is one of the older, more mature modern “cloud computing” technologies in use today.

Key features

Two of Hadoop's main features are particularly relevant to this talk:

R/Hadoop integrations

Key integration projects

Project	Sponsors/Maintainers
RHadoop	RevolutionAnalytics
RHIPE	tesseradata

Integration purposes

- Let people use Hadoop to execute R code
- Let people use R to access data stored in Hadoop

Consider integrating R and Hadoop when...

Your computing needs align with natural strengths of R and Hadoop

Evaluate alignment with the following factors:

Lab - Use R/Hadoop

Lab goals

1. Present an example of a problem to integrate
2. Connect to Hadoop via R
3. Work through a basic integration
4. Modify the analysis on your own

Lab problem (simplified)

- 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.

Summary

Stay efficient, stay practical

- Practical computing requires balancing computing speed, programming efficiency, and personal comfort. R/Hadoop integrations offer more ways to achieve balance.
- R/Hadoop integrations give practitioners opportunities to use strengths of both technologies.
- R/Hadoop integrations give R programmers access to “non-R” technologies.

Topics for further reading

R/Hadoop integrations

- RHadoop documentation and examples on github
- RHIPE’s example analysis of airplane dataset
- Any online tutorial for logistic regression or k-means via R/Hadoop