RHIPE: R and Hadoop Integrated Programming Environment; Tutorial

Jeremiah Rounds[*]

Purdue University
[*]Contact author: jrounds@stat.purdue.edu

Keywords: Rhipe, Hadoop, R, MapReduce, Divide & Recombine

Use *R* to do intricate analysis of large data sets via Hadoop. Large complex data sets that can fill up several large hard drives (or more) are becoming commonplace, and many *R* using data analyst will have to confront that reality in the coming years. Data parallel distributed computing paradigms such as MapReduce have emerged as able tools to deal with large data, but until now very little has been done to put those paradigms into the hands of *R* users. **Rhipe** is a software package that allows the *R* user to create MapReduce jobs that work entirely within the *R* environment using *R* expressions. This integration with *R* is a trans-formative change to MapReduce; it allows an analyst to quickly specify Maps and Reduces using the full power, flexibility, and expressiveness of the *R* interpreted language.

In this half-day tutorial, the audience will be introduced to distributed computing, Hadoop, MapReduce, **Rhipe**, and common statistical paradigms that can appear in data parallel algorithms. No prior experience will be assumed, but the ideal participant has an interest in distributed computing, Hadoop, MapReduce, and *R*. For those interested in following along with hands on material, a virtual machine with Hadoop, *R* and **Rhipe** preinstalled will be available for download. More information about **Rhipe** is available at `www.rhipe.org`.

With data analysis examples, we will introduce Divide and Recombine (D&R) for the analysis of large complex data. In D&R data are divided into subsets in one or more ways. Numeric and visualization methods are applied to each of the subsets separately. Then the results of each method are recombined across subsets. By introducing and exploiting parallelization of data, D&R using **Rhipe** succeeds in making it possible to apply to large complex data almost any existing analysis method already available in *R*.