Introducing RHIPE: R and Hadoop Integrated Processing Environment

Saptarshi Guha[1*]

1. Department of Statistics, Purdue University

* Contact author: sguha@purdue.edu

Keywords: High Performance Computing, Hadoop, MapReduce

With the ready availability of inexpensive yet powerful computer hardware together with breakthroughs in software for distributed computing, it has recently become feasible to analyze unprecedentedly large datasets using very popular interactive languages like R that historically could handle only small data sets. RHIPE is an open source software system that combines R and Hadoop to enable the analysis of massive datasets distributed across a cluster of computers. Hadoop, together with the Hadoop distributed file system, is an open source implementation of Google's MapReduce distributed compute engine. Using RHIPE the R user can implement MapReduce algorithms using code using the R language. The integration of R and Hadoop is accomplished via a set of components written in R and Java. The components handle the passing of information between R and Hadoop. RHIPE has worked successfully on several projects with hundreds of gigabytes of data. Currently, it is in a proof of concept stage, and a version ready for public use will be released soon.

References

Jeffrey Dean and Sanjay Ghemawat (2004). MapReduce: Simplified Data Processing on Large Clusters
`http://labs.google.com/papers/mapreduce.html`

Saptarshi Guha (2009) RHIPE: R and Hadoop Integrated Processing Environment
`http://www.stat.purdue.edu/~sguha/rhipe`

Apache Foundation, Hadoop
`http://hadoop.apache.org/core/`