

RHadoop: MapReduce Jobs in R

Jamie F Olson^{1,*}

1. Microsoft Algorithms & Data Science

*Contact author: jamieo@microsoft.com

Keywords: Hadoop, R

The RHadoop projects aims to enable *R* developers and statisticians to scale up their workflow with the Hadoop ecosystem by writing MapReduce jobs in the *R* language. RHadoop's `rmr2` package integrates the R environment with the Hadoop infrastructure, providing a seamless experience when transitioning between local execution and massively parallel distributed workloads across a cluster. This tutorial will provide an overview of the RHadoop APIs and architecture while focusing primarily on hands-on experience writing MapReduce jobs using RHadoop. We will walk through data import and export for a variety of data formats as well as some simple but instructive use cases highlighting key aspects of RHadoop functionality. Participants will be provided access to a training environment via a web browser and need not attempt any RHadoop installation/configuration.