Big Data & Hadoop
- The Future of the Information Economy

Sirish Shrestha[1,⋆]

1. Department of Statistics, West Virginia University
⋆Contact author: sshrestha@mix.wvu.edu

Keywords: Big Data, Hadoop, Map-Reduce, RHIPE, R

Since the advent of the computer, there has been a growing need for a method to analyze heaps of data efficiently. As technology advanced, access and availability of data grew exponentially. From this, the term Big Data arose, and along with it, the concern of processing and analyzing it. Traditionally, the data was processed and stored in a single computer/server. This required the hardware to be either exceptionally powerful in order to obtain the result in a timely manner, or the analyst had to wait an extended period of time—both of which are infeasible. *Hadoop* provides a software framework solution for storing and processing massive amounts of data in a distributed fashion on large clusters of commodity hardware while minimizing the wait time. It allows massive data storage, improves processing speed, and has data redundancy to promote fault tolerance. An already powerful framework for processing big data coupled with powerful tools in its ecosystem has made *Hadoop* a desirable tool in the world of Big Data. Although *Hadoop* is built using *Java*, many programming languages have created packages that allows the integration of *Hadoop* functions into their own environment providing evidence of wide acceptance of the framework—one of which is *R*, a free open source functional language for statistical computing and graphics. Due to the availability of various libraries and a rich ecosystem with a vast collection of contributed packages, which are reproducible, transparent, reusable, and reconfigurable, it perhaps has become a de facto tool for expert statisticians and data scientists. *R* has a few packages for analyzing big data despite its limitation with memory, but it also has packages that integrates *Hadoop* in its environment for distributed processing of Big Data.

In this article, an implementation of *Hadoop* MapReduce framework is shown using **RHIPE** (*R* and *Hadoop* Integrated Programming Environment, an *R* package) on Twitter feeds of Occupy Wall Street of November 15, 2011.