R integrations with Hadoop

http://barbra-coco.dyndns.org/student/hadoop/Big-Data-Analytics-with-R-and-Hadoop.pdf

RR package (more integrated with plots)

RHIPE is specially designed for providing a lower-level interface over Hadoop.

three categories: Initialization, HDFS, and MapReduce operations.

The RHIPE package uses the Divide and Recombine technique to perform

data analytics over Big Data. In this technique, data is divided into subsets, computation is performed over those subsets by specific R analytics operations, and the output is combined. RHIPE has mainly been designed to accomplish two goals that are as follows:

* Allowing you to perform in-depth analysis of large as well as small data.
* Allowing users to perform the analytics operations within R using a lower- level language. RHIPE is designed with several functions that help perform Hadoop Distribute File System (HDFS) as well as MapReduce operations using a simple R console.

* RHIPE is a lower-level interface as compared to HDFS and MapReduce operation.

Rhadoop:

rhdfs, rmr, rhbase

Hadoop Streaming - slow

Hive

Pig

snow

R benchmarks :

http://www.revolutionanalytics.com/revolution-revor-enterprise-benchmark-details

Matrix multiply

Cholesky Factorization

Singular value decompositions

Principal Components Analysis

Linear Discriminant analysis

Benchmarks for Hadoop:

<http://www.michael-noll.com/blog/2011/04/09/benchmarking-and-stress-testing-an-hadoop-cluster-with-terasort-testdfsio-nnbench-mrbench/>

Matrix multiplication for big ones

http://www.cse.buffalo.edu/faculty/miller/Courses/CSE633/Ortega-Fall-2012-CSE633.pdf