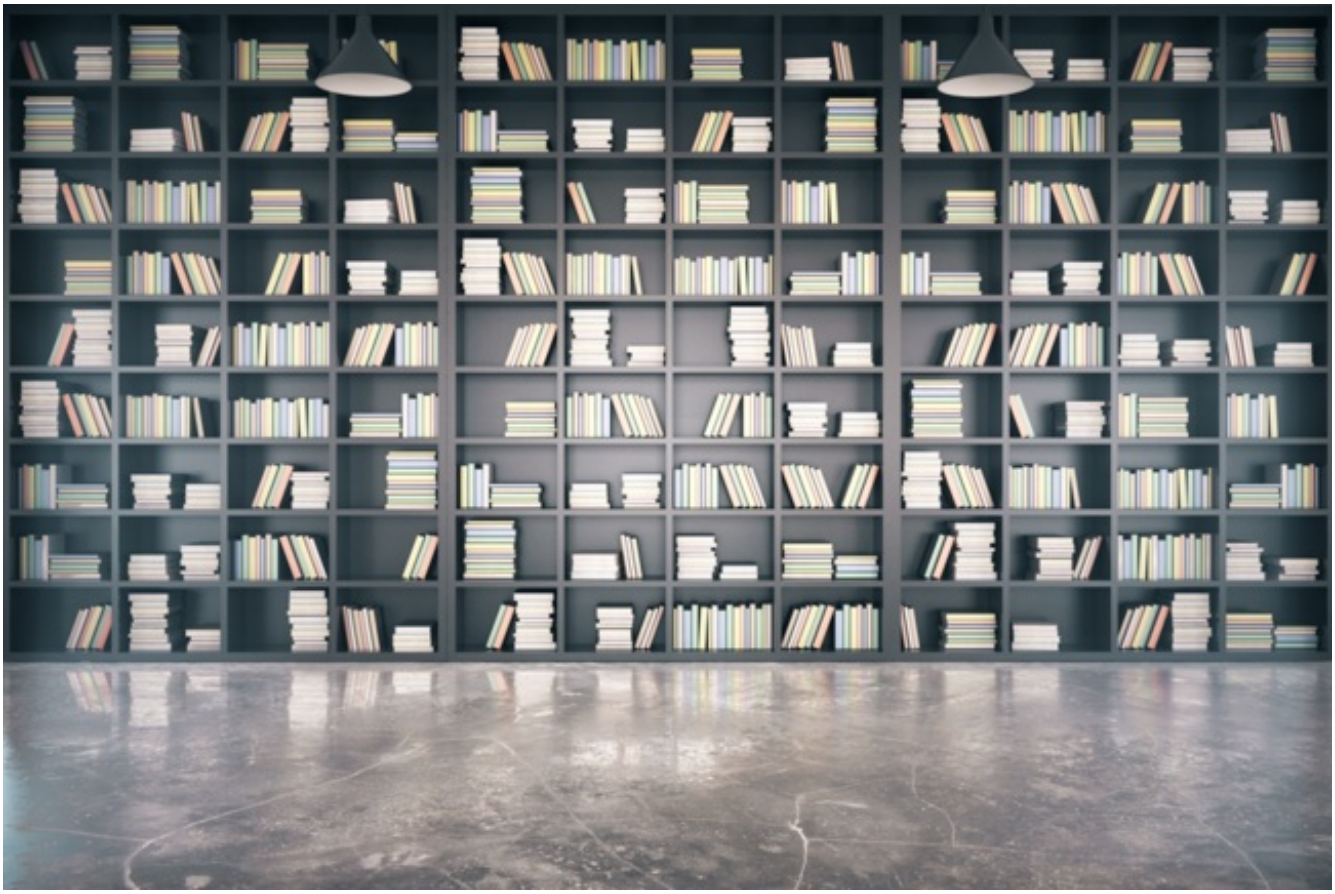
Rhadoop libraries `rhdfs`, `rmr2` and `plyrmr`

futurelearn.com /courses/big-data-r-hadoop/2/steps/216590



© Shutterstock

There are several `RHadoop` packages that we have to use in order to be able to connect `R` and `Hadoop`. Here we give a dense presentation of `rhdfs`, `rmr2` and `plyrmr` together with available commands. Note that within this MOOC we demonstrate only few of those commands.

`rhdfs`

The library package rhdfs provides commands for file manipulation in terms of reading, writing and moving files. Namely, `R` is a programming language that offers data processing. The data themselves are stored in files in the Hadoop filesystem. The following commands are part of the `rhdfs` package:

- File Manipulations

`hdfs.copy`, `hdfs.move`,
`hdfs.rename`, `hdfs.delete`, `hdfs.rm`, `hdfs.del`, `hdfs.chown`, `hdfs.put`,
`hdfs.get`

- File Read/Write

`hdfs.file`, `hdfs.write`, `hdfs.close`, `hdfs.flush`, `hdfs.read`, `hdfs.seek`, `hdfs.tell`,
`hdfs.line.reader`, `hdfs.read.text.file`

- Directory

`hdfs.dircreate`, `hdfs.mkdir`

- Utility

`hdfs.ls, hdfs.list.files, hdfs.file.info, hdfs.exists`

- Initialization

`hdfs.init, hdfs.defaults`

rmr2

The `rmr2` package allows the `Hadoop` MapReduce facility to be used inside the `R` environment. The package documentation lists the following commands:

- The big-data object

`big.data.object`

- Backend-independent file manipulation

`dfs.empty`

- Equijoins using map-reduce

`equijoin`

- Read or write `R` objects from or to the file system

`from.dfs`

- Important Hadoop settings in relation to rmr2

`hadoop.settings`

- Create, project or concatenate key-value pairs

`keyval`

- Create combinations of settings for flexible IO

`make.input.format`

- MapReduce using Hadoop Streaming

`mapreduce`

- Function to set and get package options

`rmr.options`

- Sample large data sets

`rmr.sample`

- Print a variable's content

`rmr.str`

- Functions to split a file over several parts or to merge multiple parts into one

`scatter`

- Set the status and define and increment counters for a Hadoop job

`status`

- Create map-and-reduce functions from other functions

`to.map`

plyrmr

This package aims to provide a wide palette of predefined operations to cover the basic data-manipulation needs. The documentation of the `plyrmr` package provides the following list of available commands:

- data manipulation

`bind.cols` (add new columns), `where`(select rows), `select`(select columns), `rbind`, `transmute` (all of the above plus summaries)

- summaries

`transmute`, `sample`, `count.cols`, `quantile.cols`, `top.k`, `bottom.k`

- set operations

`union`, `intersect`, `unique`, `merge`

`transmute` appears twice because it is a generalization over `transform` and `summarize` that allows us to increase or decrease the number of columns or rows, covering the need for multi-row summaries, flattening of data structures, etc.

© PRACE and Faculty of information studies in Novo mesto