Plyrmr: a data manipulation DSL for big data

Antonio Piccolboni^{1,*}

1. Revolution Analytics
*Contact author: rhadoop@revolutionanalytics.com

Keywords: data manipulation, hadoop, big data

plyrmr is the latest package to spawn from RHadoop, an open source project aimed at making *R* work seamlessly with the Hadoop system for the storage and processing of big data on commodity clusters. Like another RHadoop package, **rmr2**, it is specifically targeted to work with Mapreduce, Hadoop's batch computing subsystem. The goal for **plyrmr** was to strike a different compromise of power and ease of use biased toward the latter, thus helping to expand the circle of people who can access and process big data directly. The main compromise we accepted is that **plyrmr** is focused on structured data, specifically data organized in columns, like a data.frame, as feedback from our users indicated this was the most important use case. With this in mind, we set five design guidelines:

- Reduce the need to define functions even for the simplest tasks: to this end we have adopted a programming jargon popularized by the package **plyr** [2] (to which **plyrmr** also owes half of its name). Simple calculations such as the ratio of two columns or the average of another one can be described with expressions thanks to a non-standard but well understood evaluation method.
- Whenever users need to define functions to access advanced functionality, make them simpler and less specialized to their use in a mapreduce context, thus promoting reuse. In fact, all user defined function in the **plyrmr** API accept a data frame as their first argument and return a data frame, enabling the reuse, for instance, of functions from **plyr**, **dplyr** and **reshape2** for processing big data.
- Replace the potentially unfamiliar concept of a key with an SQL-like function group and related.
- Whereas **rmr2** was more a foundational package with a *minimalist* API, **plyrmr** includes mapreduce equivalents of many popular and useful functions, to show Hadoop conversion can be accomplished and to provide a useful set of tools even without any programming by the user. According to [1] **plyrmr** is more in the camp of a *humane interface*, whereby common use cases are worth defining and implementing, no matter how trivial their implementation. Converting more functions for Hadoop use can require as little as a call to the function magic.wand.
- using a technique known as delayed evaluation, reduce the cost of abstraction eliminating redundant I/O when possible.

A touch of syntactic sugar is a Unix-like %|% operator to make nested expressions more readable. The result are programs like input ("path/to/data-set") %|% where (var1/var2 > x) %|% group (id) %|% select (mean (var1)) that can run on the largest commodity clusters in use today, and process the largest data sets.

References

- [1] Fowler, M. (2005, December). HumaneInterface. http://martinfowler.com/bliki/HumaneInterface.html. Accessed 2014-3-20.
- [2] Wickham, H. (2009). plyr: Tools for splitting, applying and combining data. *R package version 0.1 9*, 651.