dplyr: a grammar of data manipulation

Hadley Wickham[1,*]

1. RStudio

*Contact author: hadley@rstudio.com

Keywords: data manipulation, databases

dplyr is a new package which provides a set of tools for efficiently manipulating datasets in R. dplyr is the next iteration of plyr, focussing on only data frames. dplyr is faster, has a more consistent API and should be easier to use. There are three key ideas that underlie dplyr:

1. Your time is important, so Romain Francois has written the key pieces in Rcpp to provide blazing fast performance. Performance will only get better over time, especially once we figure out the best way to make the most of multiple processors. For some cases, dplyr is 10,000x faster than dplyr.

2. Tabular data is tabular data regardless of where it lives, so you should use the same functions to work with it. With dplyr, anything you can do to a local data frame you can also do to a remote database table. PostgreSQL, MySQL, SQLite and Google bigquery support is built-in; adding a new backend is a matter of implementing a handful of S3 methods.

3. The bottleneck in most data analyses is the time it takes for you to figure out what to do with your data, and dplyr makes this easier by having individual functions that correspond to the most common operations (group_by(), summarise(), mutate(), filter(), select() and arrange()). Each function does one only thing, but does it well.