simpleCache: R caching for reproducible, distributed, large-scale projects

Nathan Sheffield¹, VP Nagraj¹, and Vince Reuter¹

¹ University of Virginia

Summary

simpleCache is an R (R Core Team 2016) package that provides functions for caching R objects. Its purpose is to encourage writing reusable, restartable, and reproducible analysis for projects with large data and computational requirements. Like its name indicates, simpleCache is intended to be simple. Users specify a location to store caches, and then provide nothing more than a cache name and instructions (R code) for how to produce an R object. simpleCache either creates and saves or simply loads the result as necessary with just a single function call.

In addition to this basic functionality, simpleCache has advanced options for assigning objects to specific environments, recreating caches, reloading caches, and even distributing caching operations to cluster computing resources via the batchools (Lang, Bischl, and Surmann 2017) interface. These features make the package particularly useful for large-scale data analysis and research projects. simpleCache is most helpful for caching objects that are computationally expensive to create, but used in multiple scripts or by multiple users.

simpleCache is also useful to enhance performance in a package that relies on large databases. For example, simpleCache has been incorporated with the LOLA R package (Sheffield and Bock 2016) to more efficiently cache and retrieve genomic region databases. Similarly, simpleCache has been used to store cached baseline statistical tables for faster lookup to determine statistical differences on tables with hundreds of millions of data points (Sheffield et al. 2017).

In summary, simpleCache provides a user-friendly interface to help the R programmer manage computationally intensive, repeated data analysis.

References


