data.table: fast and flexible data manipulation

Matt Dowle

Keywords: Large data, ordered joins, update by reference, aggregation

The data.table package inherits from and extends data.frame aiming to reduce two types of time:

- 1. programming time (fewer function calls, less variable name repetition)
- 2. compute time on large data (e.g. 64bit with 8GB+ RAM)

The package offers fast aggregation of large datasets, fast ordered joins, fast add/modify/delete of columns by group using no copies at all, list columns where each cell can itself be a vector/object and a fast file reader: fread(). Although the speed benefits are greatest on large datasets (1GB-100GB), many also use it on small datasets for its brief and flexible syntax.

The general form, including chaining is:

Currently there are 5 active contributors to the project, mainly from Genomics and Finance.

The presentation covers the essential syntax illustrated with examples.

Creating a data.table
Fast and friendly file reading with fread
Basic query syntax
Keys (setkey)
Update by reference (:= and set*)
Ordered joins forwards, backwards, limited and nearest
List columns (each cell can itself be a vector)
Why R?
Recent new features
Future directions
Quality assurance (1,000 tests, release procedures)
A review of online help (1,200 Q&A on Stack Overflow's data.table tag)

References

M Dowle, T Short, S Lianoglou, A Srinivasan, R Saporta, E Antonyan (2008-2014). data.table: Extension of data.frame. http://datatable.r-forge.r-project.org/.