14 October 2022

NSF POSE : R’s data.table : collaboration

Name: Matt Dowle, creator and maintainer of data.table
Affiliation: Principal Software Engineer, H2O.ai
Relationship to proposing team: PI Toby Dylan Hocking is a contributor to data.table

data.table is a package for R and already benefits from R’s mature open source ecosystem. I created and published v1.0 in 2006 as a faster and more memory efficient alternative to data.frame. Since then there have been 56 updates. There are 1,304 CRAN and 298 Bioconductor packages using data.table directly, and 3,241 packages which use those packages, so depend on data.table indirectly. There are 860,000 downloads per month.

In 2015 my O1 visa application was approved and I moved with my family from the UK to the US to work full time for H2O.ai in Mountain View, California. I collaborated with Mchael Lawrence at Genentech (an R-core member with write access to R) and in 2016 R’s core sort algorithm was replaced with data.table’s code that I wrote together with Arun Srinivasan. This has probably been the largest impact of data.table due to now being used by all R users (despite most R users not knowing they are in fact using data.table’s sort). In recent years I have implemented parallelism using OpenMP in several key areas of data.table, and others in the R community have followed my lead.

The package is MPL2 licensed. This encourages contribution because the code contributed belongs to the contributor. The license of the package cannot be changed without asking permission from all contributors. The MPL, unlike GPL, does permit the library to be used in closed-source software which I believe strikes a happy balance and encourages usage and contribution from for-profit companies too.

I wholeheartedly support this POSE application and will collaborate with Toby’s team. The data.table project has difficulties in terms of time to review and accept pull requests which is dependent on me currently. Toby has many good proposals on this. Benchmarking is another area where the project is in need of resources. In terms of impact to science, I do feel that the biggest impact would be to extend data.table to support out of memory operations. This is often requested. The operations data.table provides are inherently ordered which is unlike SQL. This is why users request data.table to be out of memory rather than using a database. R does now support vectors with more than 4 billion items, and memory mapping to disk too, but that support is at a low level. Work is needed to extend data.table to utilize those new capabilities. No other package, including R’s data.frame itself, has done this yet. However this needs skilled developers to implement and I see that POSE does not fund such effort directly. Perhaps by expanding the community of data.table users, and documentation about the internals of data.table at C level, that effort will increase developer skills sufficiently to achieve out-of-memory data.table.

Sincerely,

Matt Dowle 10.14.2022

Matt Dowle