

Arrow and beyond: Collaborating on next generation tools for open source data science

JJ Allaire

2018-04-19

Categories: [News Arrow](#) Tags: [News Arrow](#) [Ursa Labs](#)

Two years ago, [Wes McKinney](#) and [Hadley Wickham](#) got together to discuss some of the systems challenges facing the Python and R communities. Data science teams inevitably work with multiple languages and systems, so it's critical that data flow seamlessly and efficiently between these environments. Wes and Hadley wanted to explore opportunities to collaborate on tools for improving interoperability between Python, R, and external compute and storage systems. This discussion led to the creation of the [feather](#) file format, a very fast on-disk format for storing data frames that can be read and written to by multiple languages.

Feather was a successful project, and has made it easier for thousands of data scientists and data engineers to collaborate across language boundaries. In this post, we want to update you on how we think about cross-language collaboration, and share some exciting new plans.

Beyond file-based interoperability

File-based interoperability is a great first step, but is fundamentally clunky: to communicate between R and Python running on the same computer, you have to save out from one and load into the other. What if there were some way to share data in memory without having to copy objects or round trip to disk?

You may have experienced a taste of this if you've tried the [reticulate](#) package. It makes it possible to use Python objects and functions from R. But reticulate is focused on solving only one part of the problem, for R and Python. It doesn't help pass data from R to Julia, or Julia to Python, or Python to [Apache Spark](#). What if there were some way to share data between multiple languages without having to write a translation layer between every pair of languages? That challenge is the inspiration for the [Apache Arrow](#) project, which defines a standardized, language independent, columnar memory format for analytics and data science.

A new data science runtime

The Apache Arrow project has been making great progress, so we can now start to think about what could be built on top of that foundation. Modern hardware platforms provide huge opportunities for optimization (cache pipelining, CPU parallelism, GPUs, etc.), which should allow us to use a laptop to interactively analyze 100GB datasets. We should also be getting dramatically better performance when building models and visualizing data on smaller datasets.

We think that the time has come to build a modern data science runtime environment that takes advantage of the computational advances of the last 20 years, and can be used from many languages (in the same way that [Project Jupyter](#) has built an interactive data science environment that supports many languages). We don't think that it makes sense to build this type of infrastructure for a single language, as there are too many difficult problems, and we need diverse viewpoints to solve them. Wes has been thinking and talking publicly about [shared infrastructure for data science](#) for some time, and recently RStudio and Wes have been talking about what we could do to begin making this a reality.

These discussions have culminated in a plan to work closely together on building a new data science runtime powered by Apache Arrow. What might this new runtime look like? Here are some of the things currently envisioned:

- A core set of C++ shared libraries with bindings for each host language
- Runtime in-memory format based on the Arrow columnar format, with auxiliary data structures that can be described by composing Arrow data structures
- Reusable operator “kernel” containing functions utilizing Arrow format as input and output. This includes pandas-style array functions, as well as SQL-style relational operations (joins, aggregations, etc.)
- Multithreaded graph dataflow-based execution engine for efficient evaluation of lazy data frame expressions created in the host language
- Subgraph compilation using LLVM; optimization of common operator patterns
- Support for user-defined operators and function kernels
- Comprehensive interoperability with existing data representations (e.g., data frames in R, pandas / NumPy in Python)
- New front-end interfaces for host languages (e.g., dplyr and other “tidy” front ends for R, evolution of pandas for Python)

When you consider the scope and potential impact of the project, it's hopefully easy to see why language communities need to come together around making it happen rather than work in their own silos.

Ursa Labs

Today, Wes has [announced Ursa Labs](#), an independent open-source development lab that will serve as the focal point for the development of a new cross-language data science runtime powered by Apache Arrow. [Ursa Labs](#) isn't a startup company and won't have its own employees. Instead, a variety of individuals and organizations will contribute to the effort.

RStudio will serve as a host organization for Ursa Labs, providing operational support and infrastructure (e.g., help with hiring, DevOps, QA, etc.) which will enable Wes and others to dedicate 100% of their time and energy to creating great open-source software.

Hadley will be a key technical advisor to Ursa, and collaborate with Wes on the design and implementation of the data science runtime. Hadley and his team will also build a dplyr back end, as well as other tidy interfaces to the new system.


It might sound strange to hear that Wes, who is so closely associated with Python, will be working with RStudio. It might also sound strange that RStudio will be investing in tools that are useful for R and Python users alike. Aren't these languages and tools out to succeed at each other's expense? That's not how we see it. Rather, we are inspired to work together by the common desire to make the languages our users love more successful. Languages are vocabularies for interacting with computation, and like human vocabularies, are rich and varied. We succeed as tool builders by understanding the users that embrace our languages, and by building tools perfectly suited to their needs. That's what Wes, Hadley, and RStudio have been doing for many years, and we think everyone will be better off if we do it together!

We are tremendously excited to see the fruits of this work, and to continue R's tradition of providing fluent and powerful interfaces to state-of-the-art computational environments. Check out the [Ursa Labs](#) website for additional details and to find out how you can get involved with the project!

✚ [Summer Interns](#) →

2 Comments

RStudio Blog

 Login ▾ Recommend 8 Share

Sort by Best ▾



Join the discussion...

LOG IN WITH

OR SIGN UP WITH DISQUS 

Avatar

Thierry • 6 days ago

Thanks for this news

I don't know if it's related or not

This week oracle has announced a new virtual machine <https://blogs.oracle.com/de...>

With this VM it's possible to exchange in memory (data structure between different languages)

I do some small benchmarks between GraalVM R and cran R and I have some performance increase by a *5 factor.

^ | ▾ • Reply • Share ›



Avatar

Tal Galili • 6 days ago

Wonderful update/news.

Thank you jj, and the rest, for the amazing and inspiring work you are all doing!

^ | ▾ • Reply • Share ›

ALSO ON RSTUDIO BLOG

Extend the tidyverse workshop

1 comment • 9 months ago

essay writing websites — At least, this blog shows on how to use that kind of tool. A person could even get more information if they attend**Summer interns | RStudio Blog**





5 comments • 2 months ago

Bella Feng — wow, amazing opportunity! wish I was 20 years younger. :)**RStudio v1.1 Preview - Object Explorer**

11 comments • 8 months ago

Duy Thọ Nguyễn — How about copy/paste to terminal windows? I use RStudio in Ubuntu and could not use Ctrl + Shift + C / Ctrl + Shift**RStudio Connect v1.5.4 - Now Supporting Plumber!**

1 comment • 9 months ago

Ian Fellows — Very cool. This was the missing killer feature for me. Subscribe  Add Disqus to your site  Add Disqus  Privacy**Search**

Type and press Enter

You may subscribe by Email or the [RSS feed](#).

Email

Read our [privacy policy](#).

Subscribe

News & Events

Effective Application of the R Language — November 1-3 — Boston, MA [↗](#)

rstudio::conf — Jan 31-Feb 1 — San Diego, CA [↗](#)

© RStudio, Inc. 2011 - 2017