Performance Testing and Comparative Benchmarking for Creating a Self-Sustaining Ecosystem for data.table

Doris Afriyie Amoakohene, Toby Hocking

School of Informatics, Computing & Cyber Systems — NAU

NAU NORTHERN ARIZONA UNIVERSITY

Office of Undergraduate Research and Creative Activity

Introduction

- data.table is an extension of R's data.frame, designed to handle large datasets efficiently. It provides a syntax that is both concise and expressive, allowing users to perform complex data manipulations with ease
- This is a project funded by the NSF POSE program, the project aims to establish a new governance model and promote a sustainable open-source ecosystem around the data.table package

Methods

- The atime package in R is used for benchmarking the performance of R packages (data.table) , by comparing it with similar functions in other R packages and benchmarking different versions of R packages (data.table).

```
atime::atime(
N=10^seq(1,20),
setup={
    ...
},
"data.table::fwrite" = {
    data.table::fwrite()
  },
"pandas::to_csv" = {
    reticulate::py_run_string()
    }
  )
```

```
atime::atime_versions(
pkg.path = "~/data.table",
pkg.edit.fun = pkg.edit.fun,
N = 10^seq(1,20)
setup = {
...
},
expr=data.table:::`[.data.table`(...),
  "slow"="15f0598b9828d3af2eb8ddc9b38e0356f42afe4f",
  "fast"="6f360be0b2a6cf425f6df751ca9a99ec5d35ed93"
  )
```

Path to git clone of repo containing R package(data.table).

function called to edit package before installation

Conclusion

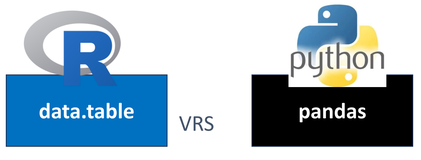
- data.table is an efficient package for data manipulation.
- data.table is a modern, community-driven open-source project that prioritizes sustainability and collaboration.
- atime package proves to be exceptionally useful for conducting comparative benchmarking and performance testing.

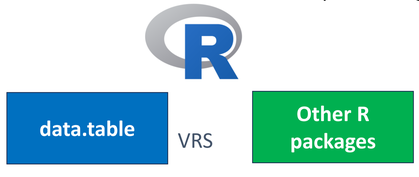
References

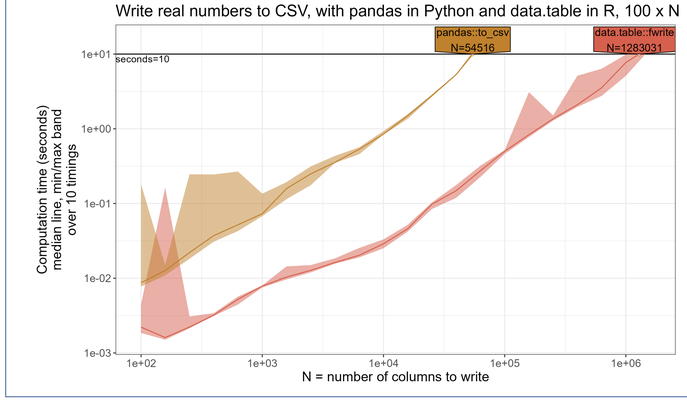
- atime: Asymptotic Time and Memory Complexity, https://github.com/tdhock/atime

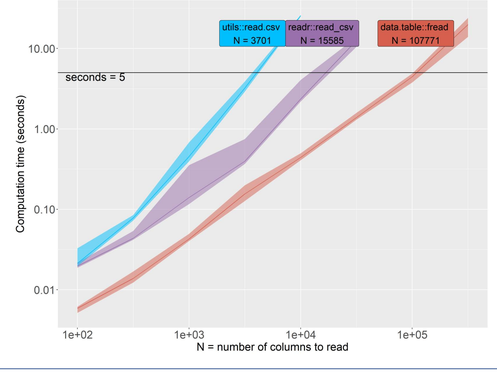
Comparative Benchmarking

- **Comparative Benchmarking**: Comparing data.table to other packages in R and python that perform same tasks
- The following graphs provide a comparative benchmarking analysis between writing CSV files using pandas in Python and data.table. Additionally, the other graphs also showcase a comparison between data.table and other functions in R for performing similar tasks.









Performance & Continuous Performance Testing

- Performance Testing: We evaluate the performance of different versions of the data.table repository by benchmarking their memory and time usage, focusing mainly on time.
- GitHub Action: To monitor data.table's high-performance standards, this initiative aimed to implement automated monitoring for performance regressions and run for every pull request
- Slow : This refers to a release that a caused slowness or late execution of a particular function
- Fast: Commit where the performance has been restored or improved beyond the point of regression
- CRAN: Latest version on the CRAN platform
- base: PR target
- HEAD: PR source
- Merge=base: The common ancestor between base and HEAD