Benchmarking Performance 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.
- Its efficiency is particularly evident when dealing with tasks like, Extracting, Transforming, Loading, filtering, grouping, aggregating, and joining data

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

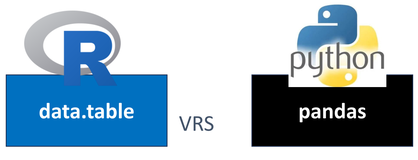
- data.table is a high-performance R package that enables efficient handling of large datasets, making it valuable for machine learning practitioners.
- Its optimized data structure and syntax streamline both pre-processing data sets, prior to running ML algorithm, and for post-processing, when you want to visualize the results of a machine learning analysis, also enhancing the speed and simplicity of data manipulation in the ML workflow.
- atime package proves to be exceptionally useful for conducting comparative benchmarking and performance testing.

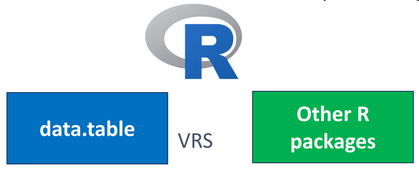
References

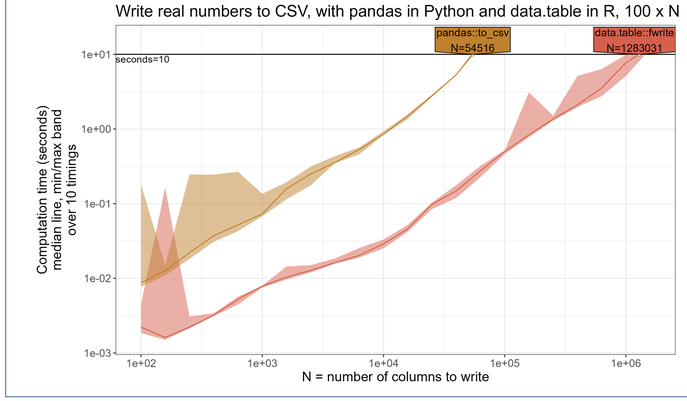
- atime: Asymptotic Time and Memory Complexity, https://github.com/tdhock/atime

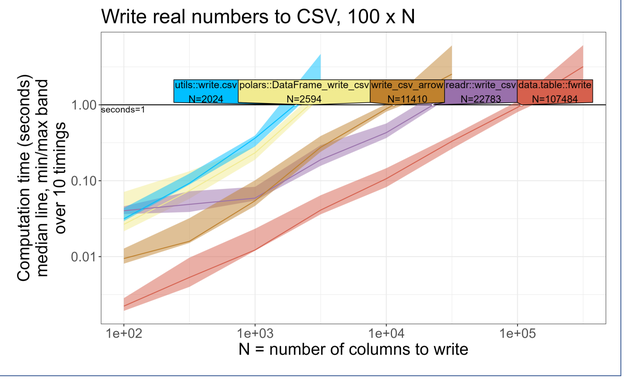
Comparative Benchmarking

- **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 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.
- Before : A version that existed prior to a specific regression used to differentiate between the states before and after that fix
- Slow : This refers to a release that a caused slowness or late execution of a particular function or version.
- Regression : A release that introduces new issues not present in previous versions, leading to a decline in performance.
- Fast/fixed: A version that fixed the regression/slowness on enhancing performance and speed