Expanding the Open Source Ecosystem around Data.table in R

Doris Amoakohene, MSc.
Toby Hocking, Professor

Funded by National Science Foundation(NSP), under the program for open source software for science and Engineering (POSE) project

Analyzing Efficiency and Performance Regressions in Packages.

Who am I

Personal

I am Doris Amoakohene Afriyie

Experience

Graduate Research Assistant– Northern Arizona University(Aug 2023 - present

Project

Expanding the Open source ecosystem around data.table in R

Education

Msc. Informatics

Interests

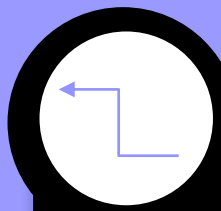
I have been using R since 2018
Using data.table since 2021, but very efficiently from 2023

The Goals For This Project

- Modify atime compare-data.table-tidyverse vignette, and analyze efficiency of packages such as polars, arrow, collapse, spark.
- Examine the history of data.table repository, including issues about performance regressions, to create two relevant performance tests, and use atime to analyze three different code branches (before regression, regression, fix regression).
- Fork data.table, and create github action for performance testing, which is run for every Pull request.
- Create slides and present results in machine learning group meeting

Overview of Presentation

- What is Data.table? History of data.table Repository
- atime and its importance in performing this analysis
- Overview of Packages to be examined(Polars, arrow, collapse, spark in R)
- Conclusion



What is Data.table?
History of data.table
Repository

What is Data.table? History of data.table Repository

Data.table is a powerful R package that provides an enhanced version of the traditional data.frame object. It is designed for efficient data manipulation, particularly for large datasets.

- Data.table has more powerful R code syntax, and C code implementation
- R package on CRAN since 2006
- Created by Matt Dowle, co-author Arun Srinivasan since 2013, 50+ contributors
- 1463 other CRAN packages require data.table (in most popular 0.05% of all CRAN packages, rank 11/19932 as of 1 Oct 2023)

(source:TobyPresentation slide)

DT

	Id	val
1	d	3
2	c	5
3	f	6
4	s	7
5	s	1
6	a	4

Comparing Data.table and Data Frame

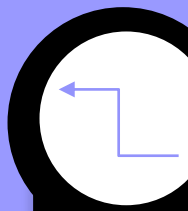
Similarities	data.table	data.frame
Tabular Data Structure	Yes	Yes
Data Manipulation	Yes	Yes
Column Naming	Yes	Yes
Column Types	Numeric, Character, Logical, Factors, and more	Numeric, Character, Logical, Factors, and more
Data Analysis	Yes	Yes
Integration	Can be used together within the same code or project	Can be used together within the same code

Differences	data.table	data.frame
Performance	Efficient for large datasets	May be slower for large datasets
Syntax	Concise and expressive	Traditional R syntax
Memory usage	Uses less memory	May use more memory
Compatibility	Compatible with most R functions and packages	Widely used default data structure in R
Additional features	Advanced join operations, fast grouping and aggregating functions, modify data by reference	Standard functionality without specialized features

Compare Data.table and Tidyverse

Similarities	data.table	tidyverse
Data Manipulation	Yes	Yes
Tabular Data Structure	Yes	Yes
Column Naming	Yes	Yes
Data Analysis	Yes	Yes
Integration	Can be used together within the same code or project	Can be used together

Differences	data.table	tidyverse
Syntax	Specific syntax unique to "data.table"	Consistent syntax across "dplyr," "tidyr," and other packages
Performance	Optimized for large datasets	May be slower for large datasets
Memory usage	Uses less memory	May use more memory
Additional features	Advanced join operations, modify data by reference	Wide range of packages for data manipulation and analysis



atime and its importance in performing this analysis

atime for Performance Efficiency

What is atime

- The atime package in R is designed to facilitate the computation and visualization of comparative asymptotic timings of different algorithms and code versions.
- It offers functionality for comparing empirical timings with expected references based on asymptotic computational complexity.

What is atime use for?

- It has features used for measuring asymptotic memory usage and other quantities.
- Measuring and comparing the execution times of various expressions or functions for different data sizes. Users can gain insights into the scalability and efficiency of their code.

Key Function

- The key function provided by the 'atime' package is called atime().
- This function allows users to specify a numeric vector of data sizes (N) and an expression or a list of expressions (expr.list) to be timed

atime importance in performing this analysis

- Compare the performance of these packages in terms of their speed, memory usage using the atime

- When assessing the efficiency and performance of these packages, it is important to consider that they can vary depending on factors such as the nature of the dataset, available hardware resources, and specific use cases. In this context, utilizing the atime metric will greatly assist us in accurately determining these variations.

Overview of Packages to be examined

Packages

Polars

- Polars is a fast and efficient data manipulation and analysis library for R.
- It provides a DataFrame API similar to pandas in Python and allows you to perform operations like filtering, aggregating, joining, and transforming data.
- Polars is designed for working with large datasets and utilizes parallel processing to achieve high performance.

spark

- The sparklyr package provides an R interface to Apache Spark, a distributed computing system.
- Spark is known for its ability to process large-scale data sets in a distributed and parallel manner.

Collapse

- The collapse package provides various functions for aggregating and collapsing data in R.
- It includes functions for collapsing data by group, creating contingency tables, calculating summary statistics, and performing other data aggregation tasks.
- The package aims to provide efficient and convenient methods for summarizing data.

Packages

Dplyr

- dplyr is a widely used package for data manipulation in R.
- It provides a set of functions that allow you to easily filter, arrange, mutate, and summarize data. dplyr follows a "grammar of data manipulation" approach, which means that it provides a consistent and intuitive syntax for performing common data manipulation tasks..

tidyr

- tidyr is another popular package for data manipulation in R.
- It focuses on transforming data between "wide" and "long" formats, also known as data reshaping or tidying. tidyr provides functions like pivot-longer and pivot wider that help you reshape your data to a more suitable format for analysis

Reshape2

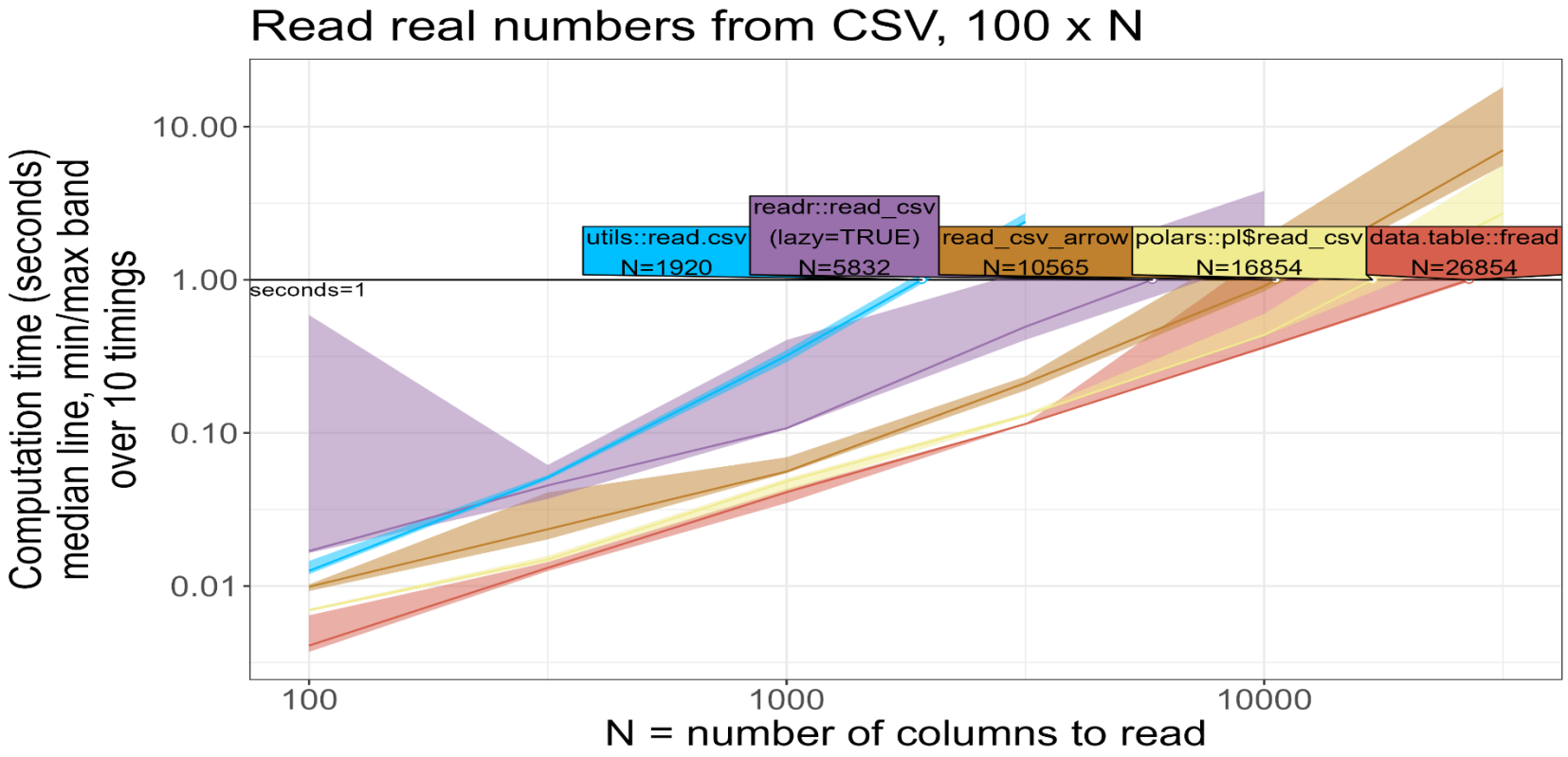
- reshape2 is an older package in R that provides functions for reshaping and transforming data.
- It offers tools for converting data between different formats, such as converting data from wide to long format or vice versa.

Plot showing analyzes of efficiency of data.table and packages

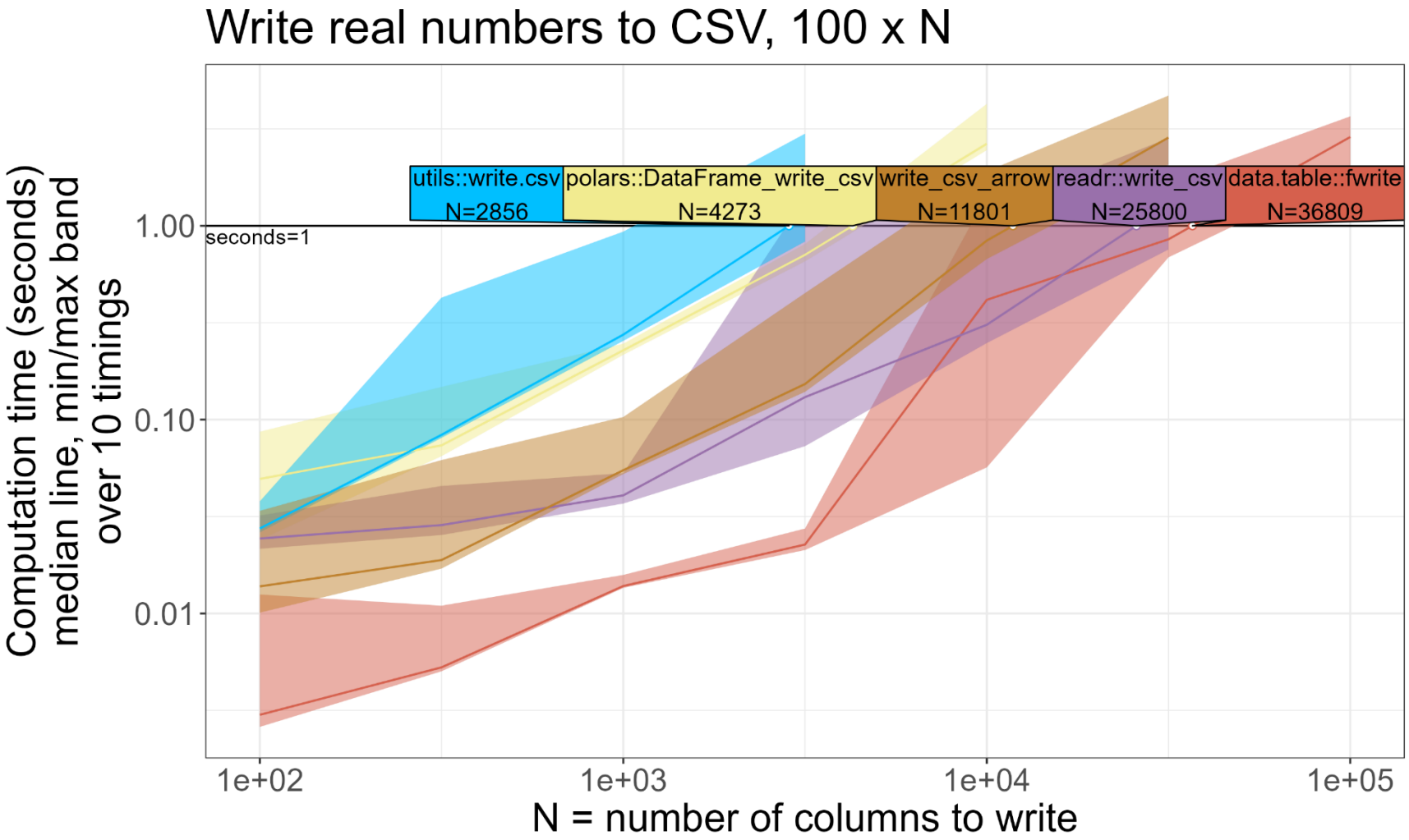
Analyzing Performance Regressions(Before, Regression and Fix)

- Identify Performance Issues
- Creating relevant performance tests allows you to set benchmarks and measure the performance of the system under various conditions.
- Validate Fixes
- Maintain User Satisfaction
- Future Performance Monitoring

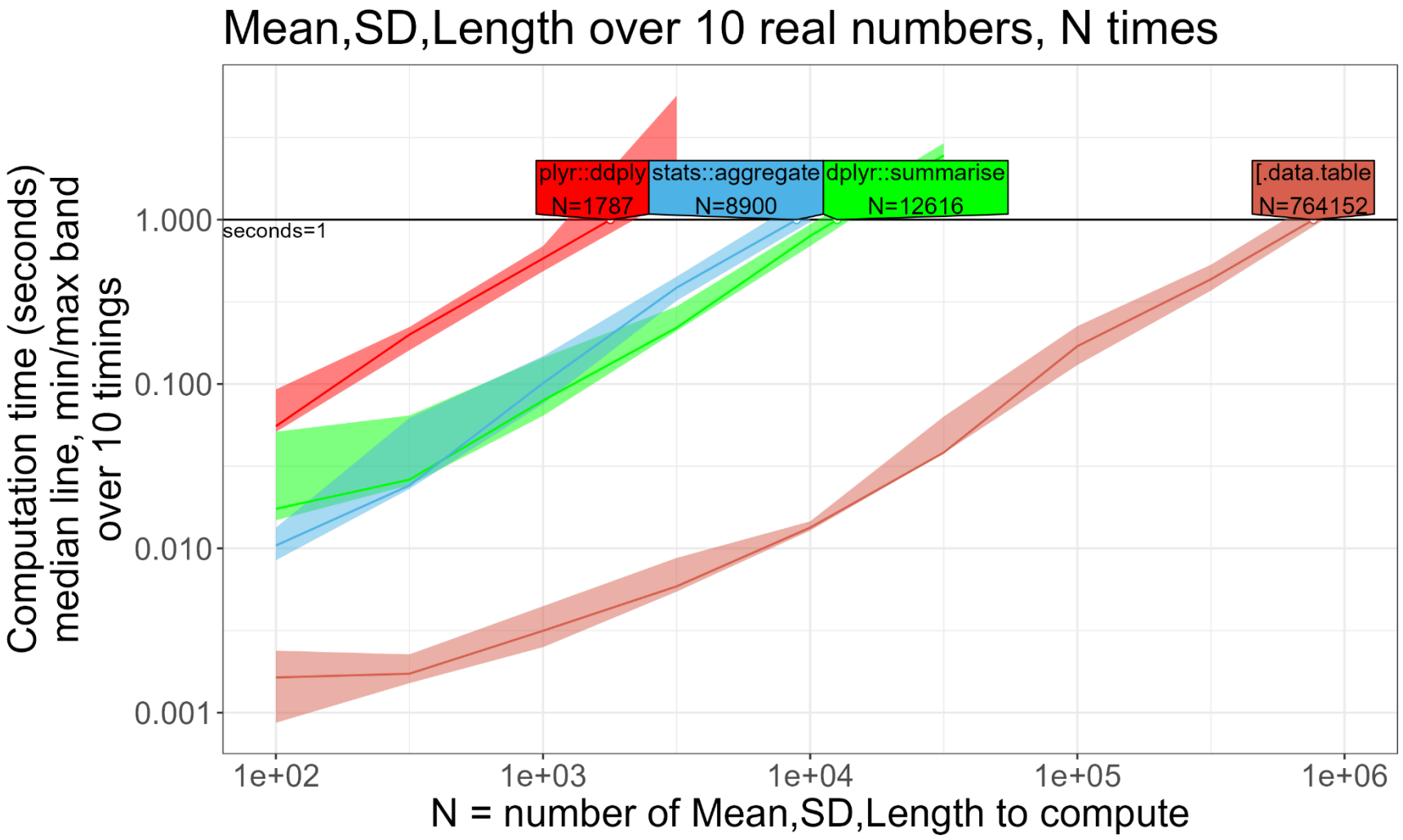
Reading CSV Files in R



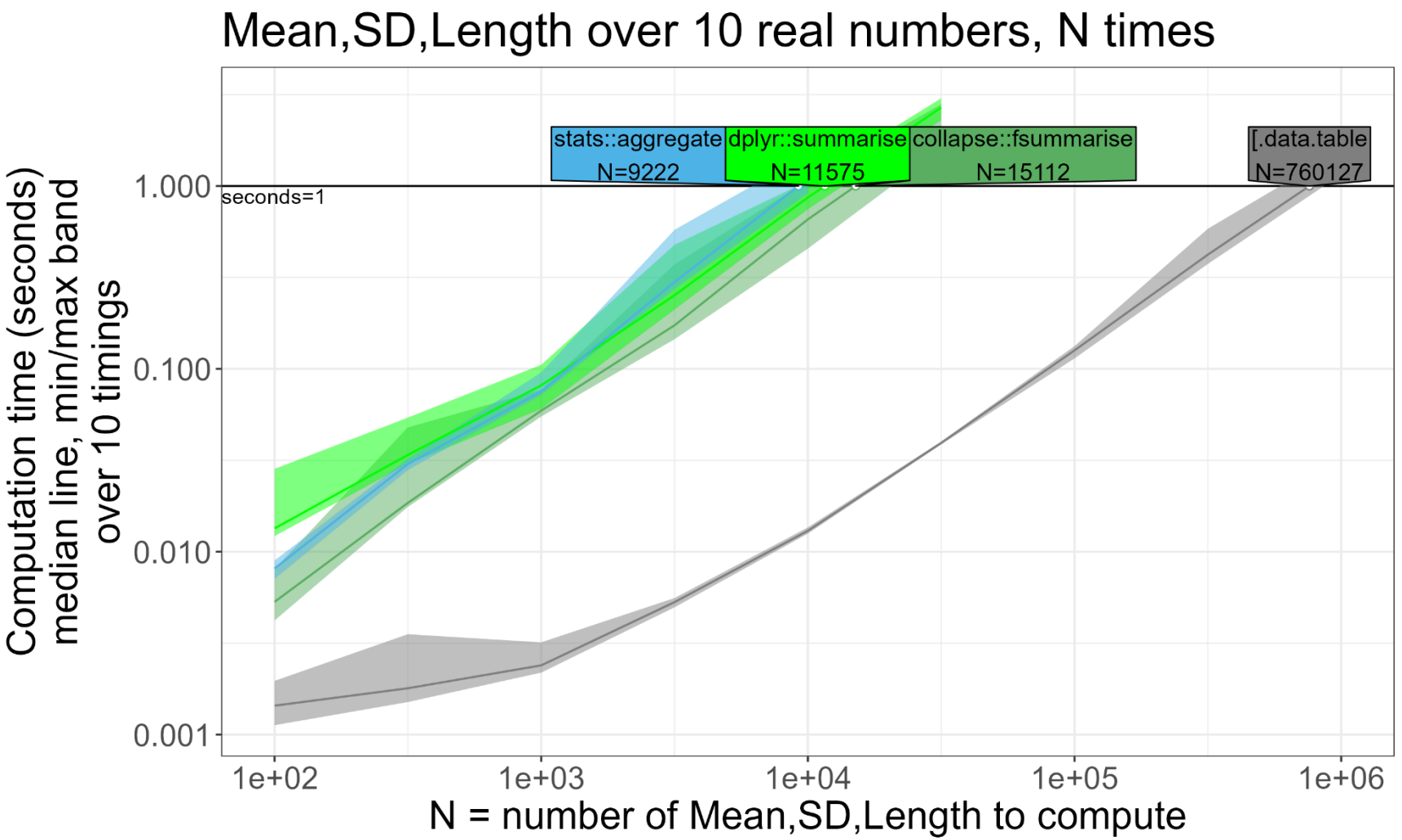
Writing CSV Files in R



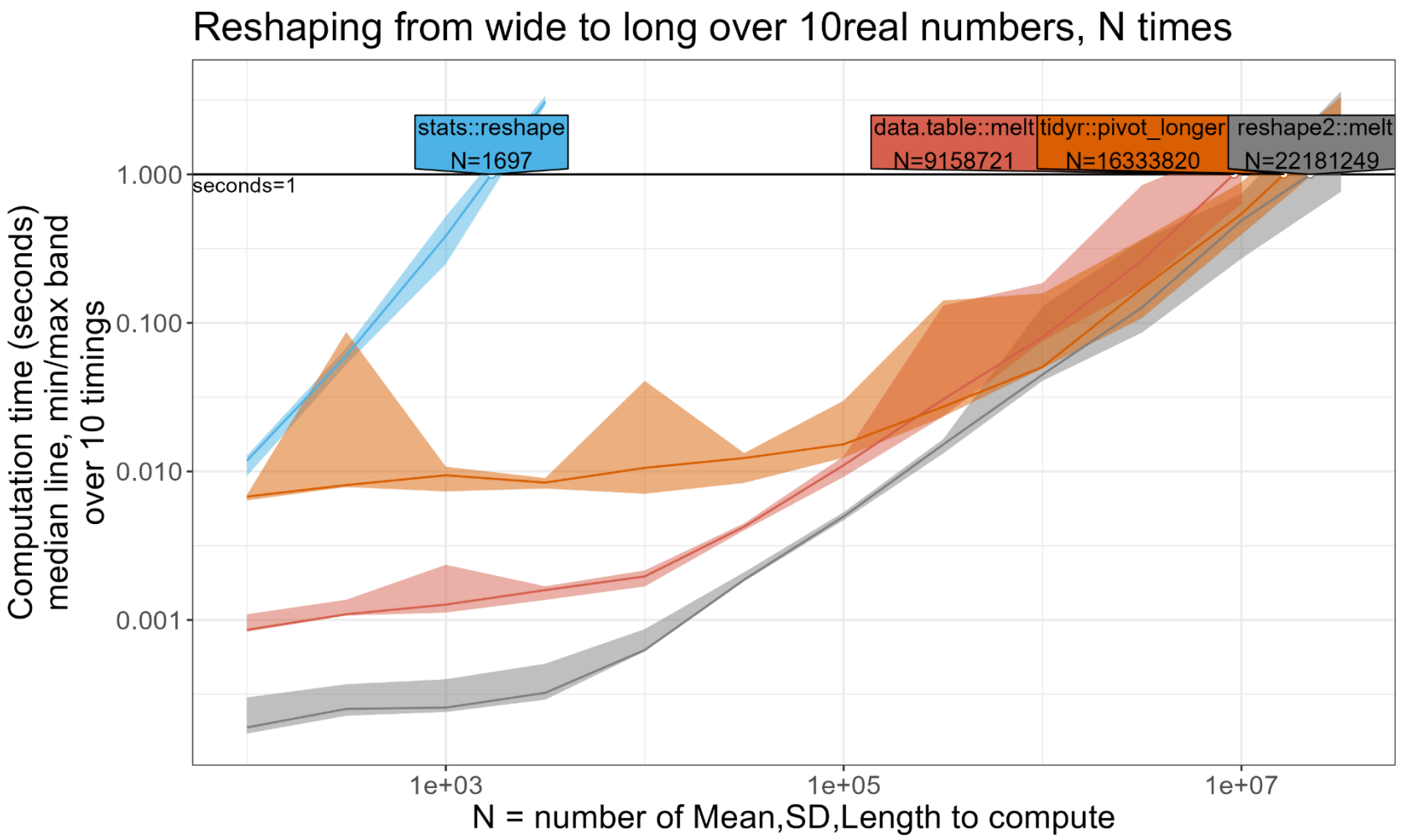
Creating Data Summary in R



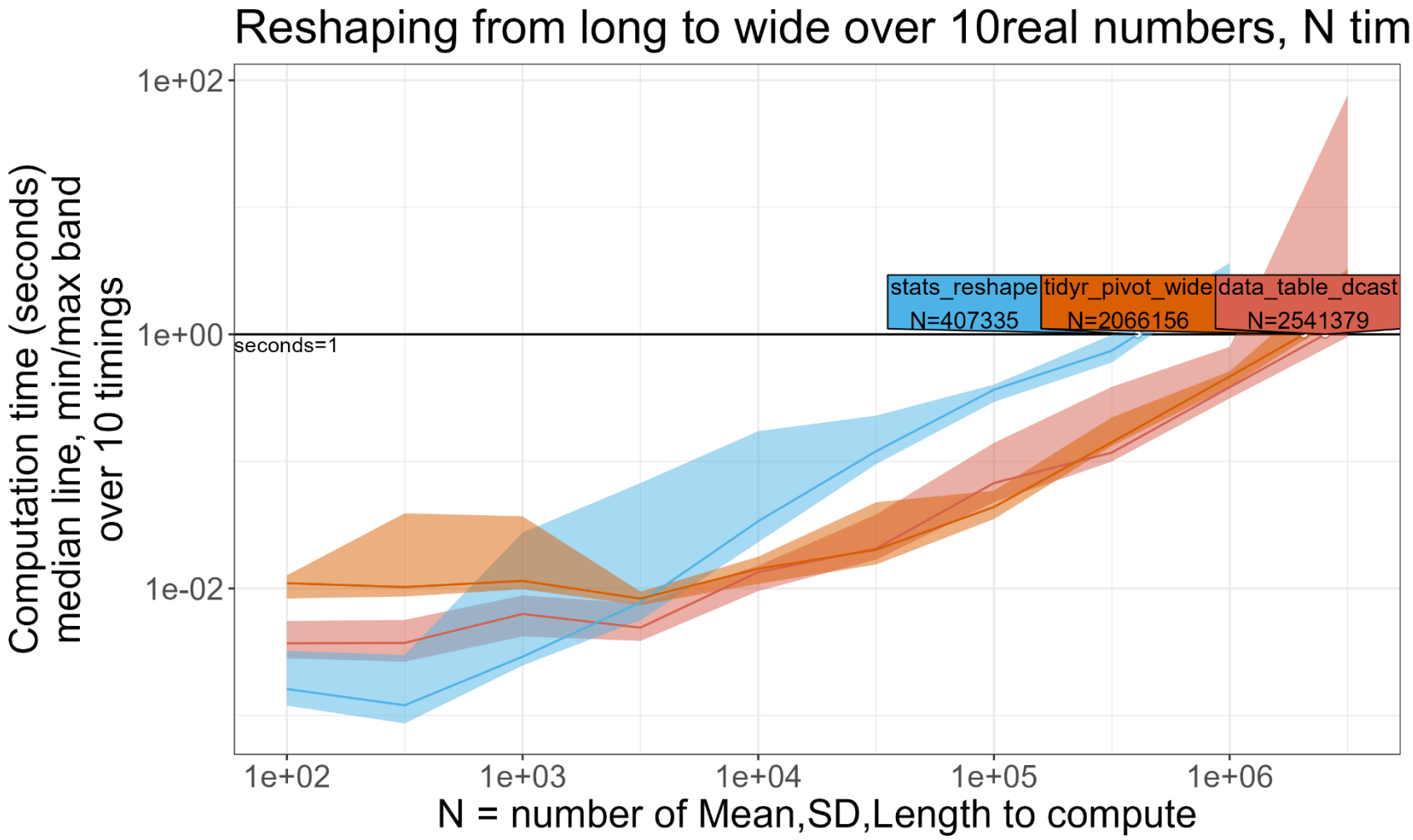
Creating an Expanded Data summary in R

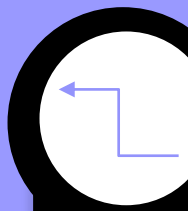


Reshaping Data from wide to long format



Reshaping Data from Long to wide Format in R





Performance Regression on Github Issues

Comparing the efficiency of the data.table package with other packages is important for several reasons:

1. Performance Optimization: Efficiency comparisons help identify the most efficient package for handling large datasets and performing complex operations.

2. Decision-making: By comparing the efficiency of data.table with other packages, you can make informed decisions about which package to use for specific tasks

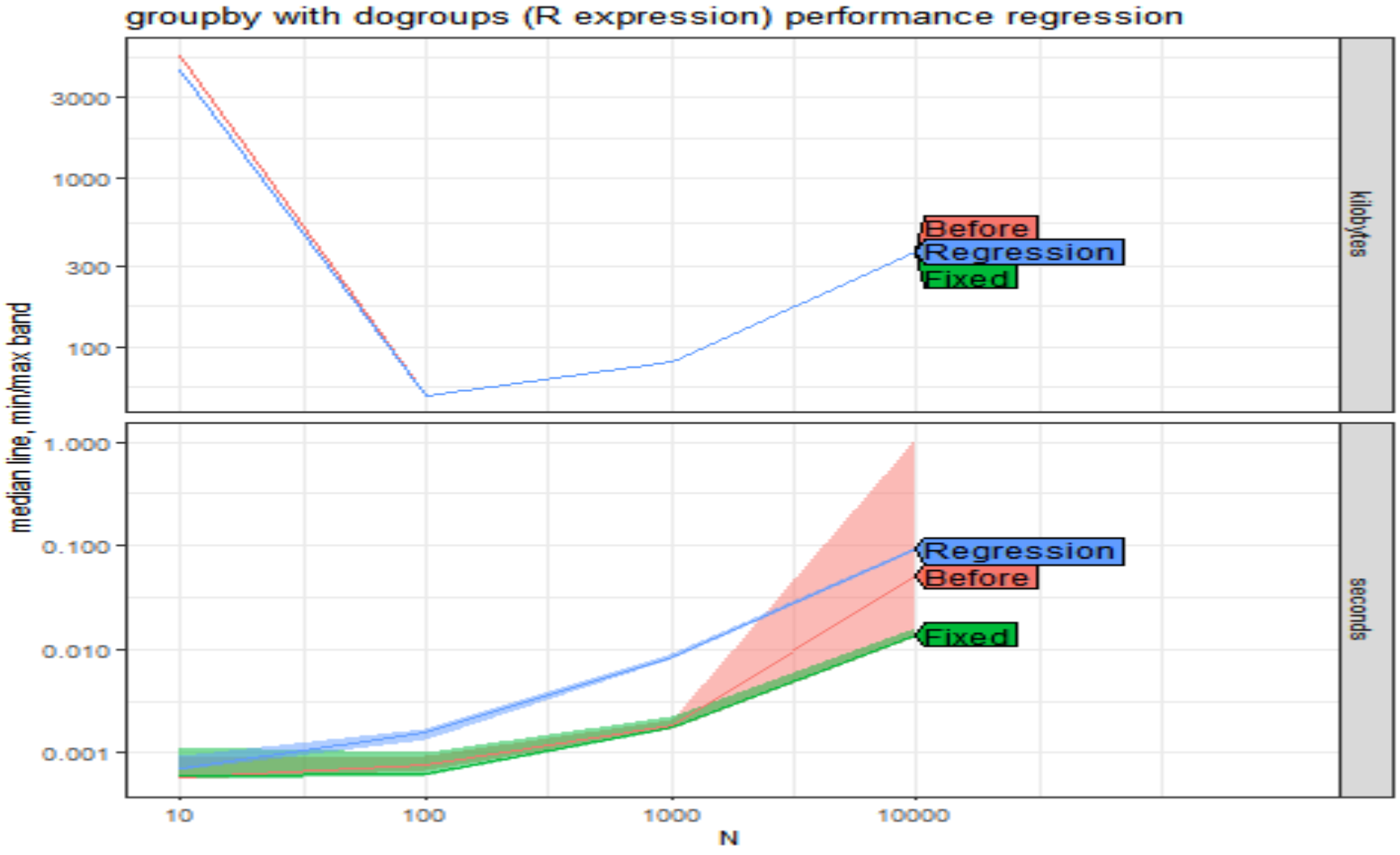
3. Compatibility and Integration: Understanding the efficiency of different packages helps you assess their compatibility with your existing codebase, libraries, and dependencies

4. Benchmarking and Reproducibility: Efficiency comparisons enable benchmarking, allowing you to measure the performance of different packages objectively. This helps researchers, developers, and data scientists reproduce experiments, validate results, and ensure consistency across different analyses.

groupby with dogroups (R expression) performance regression #4200

- This issue reported a performance regression when performing group computations, specifically when running R's C eval on each group (q7 and q8) in the db-benchmark, indicating a slowness in the implementation of the code.
- Regression: The regression was specifically related to the evaluation of C code within each group of data, specifically q7 and q8 in the "db-benchmark“
- Fixes Regression: The Regression was fixed : The regression was fixed Regression by the addition of const int nth = getDTthreads

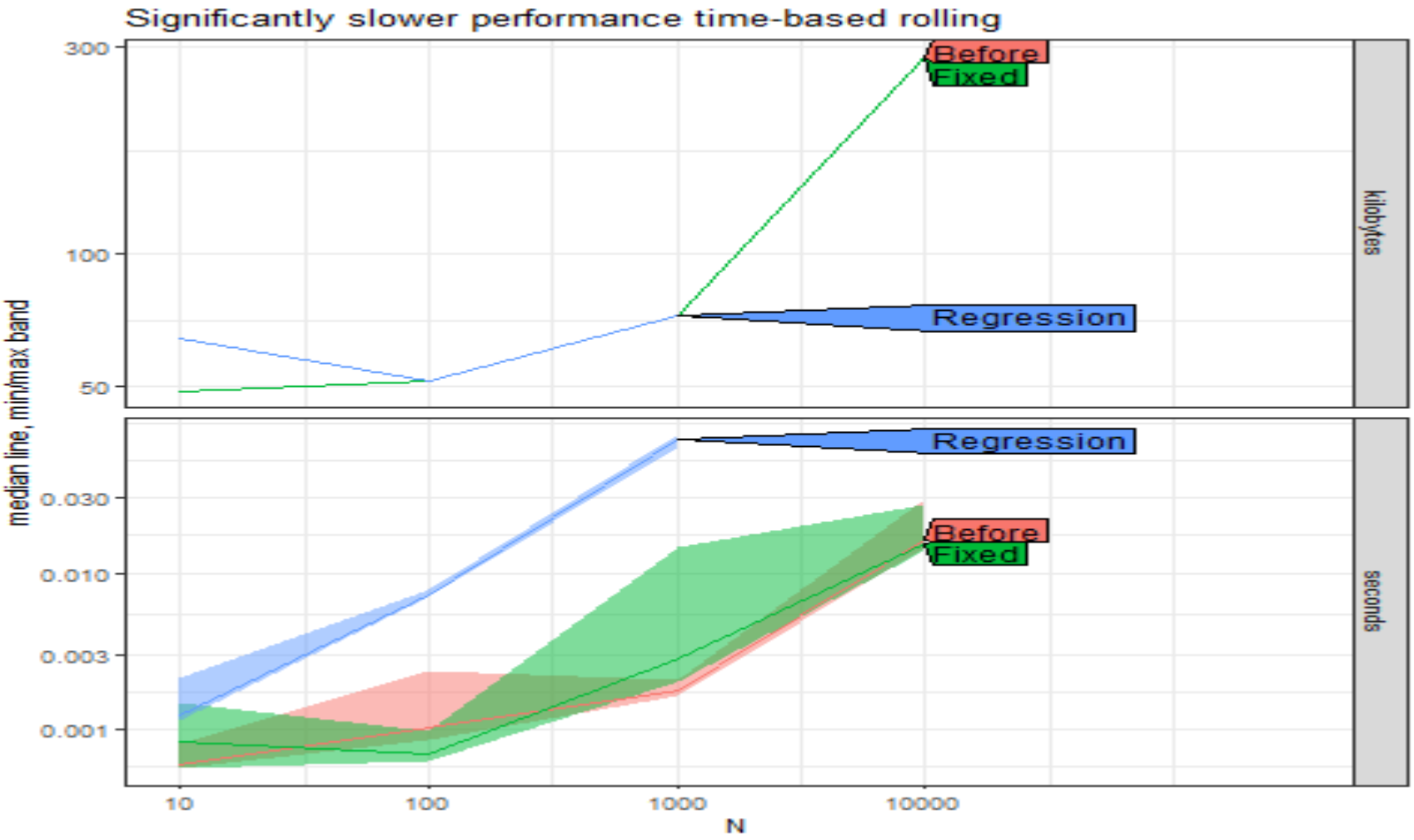
groupby with dogroups (R expression) performance regression #4200



Significantly slower performance time-based rolling and issue5371

- The cause of the regression is related to the addition of the snprintf function in the assign.c.
- Regression: The regression was specifically related to significantly slower performance time-based rolling
- Fixes Regression: The Regression was fixed by creating targetDesc function and adding snprintf in assign.c

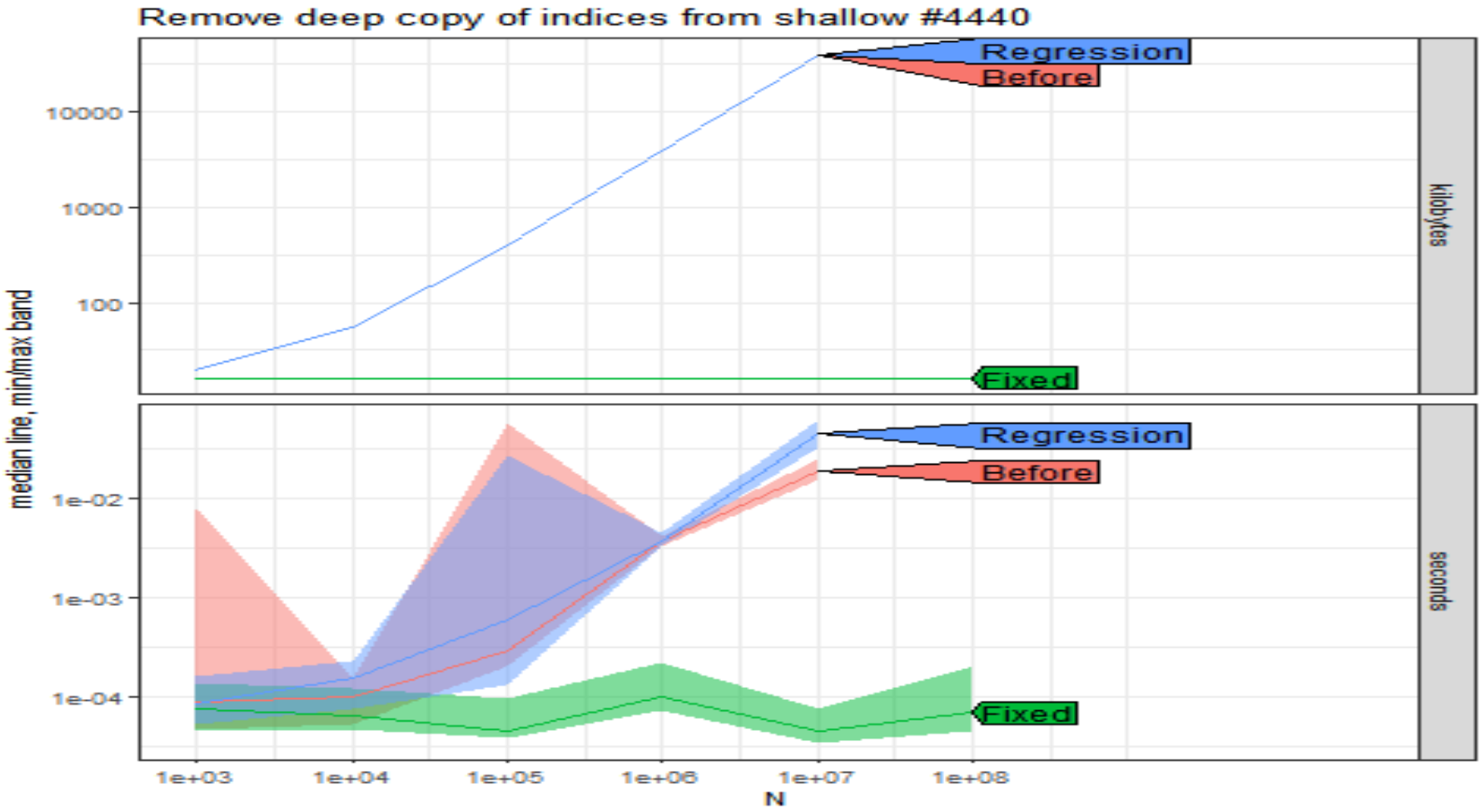
Plot


Significantly slower performance time-based rolling
median line, min/max band
kilobytes
seconds
Before
Fixed
Regression
Regression
Before
Fixed
300
100
50
0.030
0.010
0.003
0.001
10
100
1000
10000
N

Remove deep copy of indices from shallow

- This issue reported that when using the dt[selector, foo := bar] syntax with an index defined, the performance of that operation can be significantly slower compared to when there is no index.
- Regression: The regression was Remove deep copy of indices from shallow slow
- Fixes Regression: The Regression was fixed by passing shallow(dt.s4) to the isS4() function

Plot


Remove deep copy of indices from shallow #4440
Regression
Before
Fixed
10000
1000
100
kilobytes
median line, min/max band
1e-02
1e-03
1e-04
seconds
1e+03
1e+04
1e+05
1e+06
1e+07
1e+08
N

CONCLUSION

- data.table has a strong track record as a high-performance package for data manipulation.
- It excels in tasks such as reading and writing CSV files, grouping data, and reshaping data between wide and long formats.
- data.table demonstrates responsiveness when it comes to addressing and resolving performance-related issues, actively engaging in discussions and providing timely fixes.

Thank You!!!