

A brief introduction to R data.table package



If thoR was an R programmer,
his hammer would be data.table

data.table

(i, by)



Miguel P Xochicale

 @mxochicale  @mxochicale

Licence CC-BY-4.0

This presentation is released under the terms of the Creative Commons Attribution-Share Alike license. You are free to reuse it and modify it as much as you want as long as you re-share your presentation under the same terms and you mention Miguel P Xochicale as being the original author.

This presentation was built in Ubuntu 14.04 x64 with Markdown and Pandoc, and is available at:

- <https://github.com/mxochicale/thw-r-datatable>

See README.md for further information

Outline

- What is `data.table`?
- Why bother using `data.table`?
- Basic Examples with `data.table[]`
- Analysing Time Series with `ggplot()` and `data.table()`
- References

What is data.table?

The data.table R package that allows you to do fast data manipulations (for example, 100GB in RAM).

- 678 packages import/depend/suggest data.table (543 CRAN + 135 Bioconductor)
- Github: <https://github.com/Rdatatable/data.table>

Rdatatable / data.table

Watch

177

Star

1,573

Fork

704

R's data.table package extends data.frame: <http://r-datatable.com>

3,492 commits

15 branches

45 releases

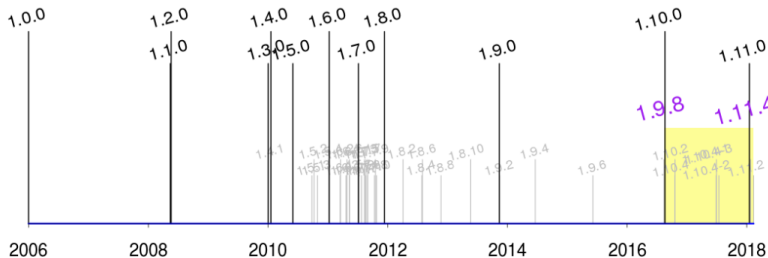
53 contributors

MPL-2.0

The R `data.table` package is 12 years old

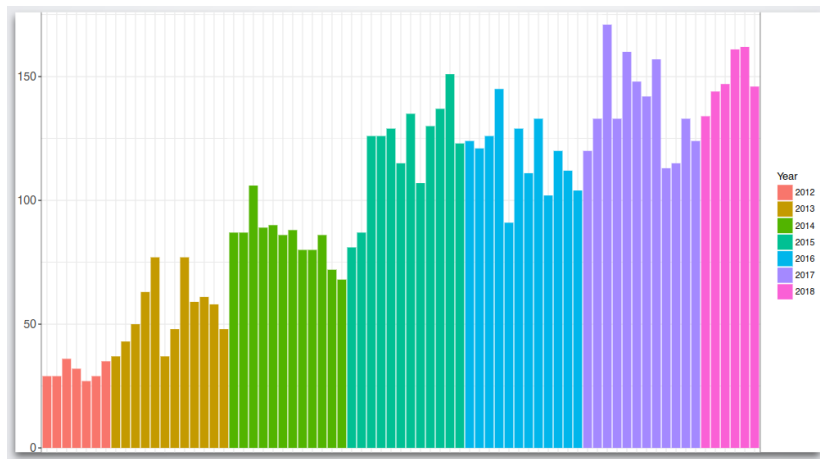
- More than 35 releases of `data.table` since 2006 on CRAN
- 45 releases of `data.table` in Github

`data.table` development timeline:



What's new in `data.table`, (Jan Gorecki, 2018.07)

Stack Overflow Questions from 2012-2018

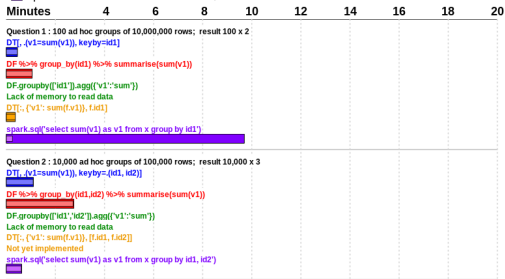


Grouping benchmarks (2018)

Input table: 1,000,000,000 rows x 9 columns (50 GB)

■ data.table 1.11.5 - 2018-09-07 - Total: \$0.19 for 23 minutes
■ dplyr 0.7.99.9000 - 2018-08-27 - Total: \$0.82 for 98 minutes
■ pandas 0.23.4 - 2018-08-04 - Total: \$NA for NA minutes
■ pydatatable 0.6.0 - 2018-09-08 - Total: \$NA for NA minutes
■ spark 2.3.1 - 2018-06-08 - Total: \$0.20 for 24 minutes

■ First time
■ Second time



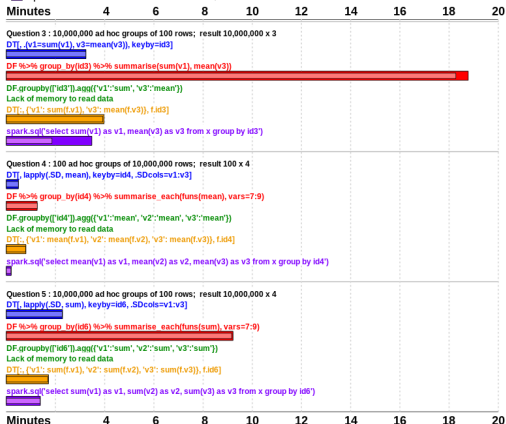
h2oai.github.io/db-benchmark

Grouping benchmarks (2018)

Input table: 1,000,000,000 rows x 9 columns (50 GB)

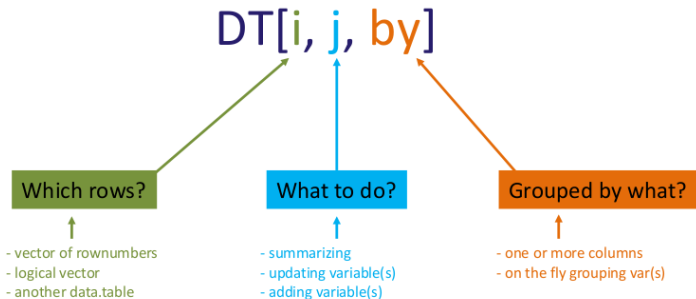
■ data.table 1.11.5 - 2018-09-07 - Total: \$0.19 for 23 minutes
■ dplyr 0.7.99.9000 - 2018-08-27 - Total: \$0.82 for 98 minutes
■ pandas 0.23.4 - 2018-08-04 - Total: \$NA for NA minutes
■ pydatatable 0.6.0 - 2018-09-08 - Total: \$NA for NA minutes
■ spark 2.3.1 - 2018-06-08 - Total: \$0.20 for 24 minutes

■ First time
■ Second time



h2oai.github.io/db-benchmark

data.table Syntax



Code example

```
puts "Hello world."  
def my_awesome_variable  
  puts "My awesome variable"  
end
```

Slide with R Code and Output

```
summary(cars)
```

Slide with text and footnote

Surely this is true.¹

¹Jane Doe, *Says It Here* (New York: Oxford University Press, 2050).

References

- <https://github.com/Rdatatable/data.table/wiki>
- <https://github.com/arunsrinivasan/user2017-data.table-tutorial>
- <https://www.datacamp.com/courses/data-table-data-manipulation-r-tutorial>
- <https://www.datacamp.com/community/tutorials/data-table-cheat-sheet>
- <https://h2oai.github.io/db-benchmark/>