



A brief introduction to R data.table package

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



Miguel P Xochicale

 @mxochicale  @mxochicale

Licence

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`
- Time Series Analysis with `data.table` and `ggplot`
- References

What is `data.table`?

*The R `data.table` package extends `data.frame`.
`data.table` allows you to do fast data
manipulations (for example, 100GB in RAM).
`data.table` goals are reduce both programming
time and compute time.*

- 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

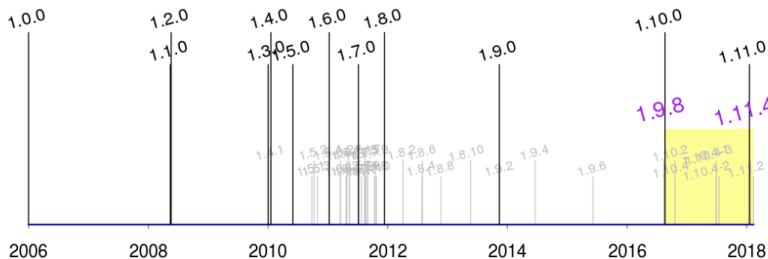
MPL-2.0

Why bother using `data.table`?

The R data.table package is 12 years old

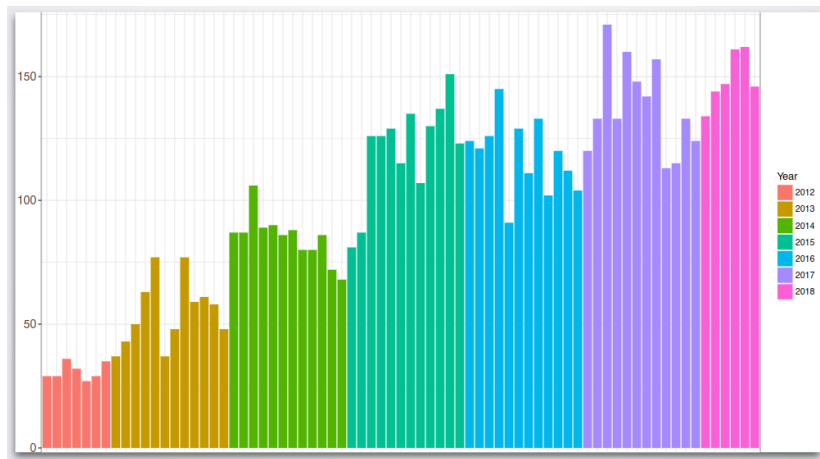
- More than 35 releases of data.table since 2006 on CRAN
- 45 releases of data.table in Github

data.table development timeline:



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

Stack Overflow Questions from 2012-2018



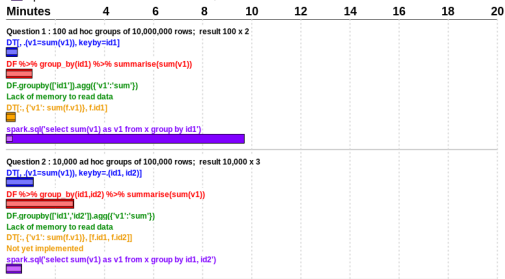
12 years of data.table, (Arun Srinivasan, 2018.07)

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



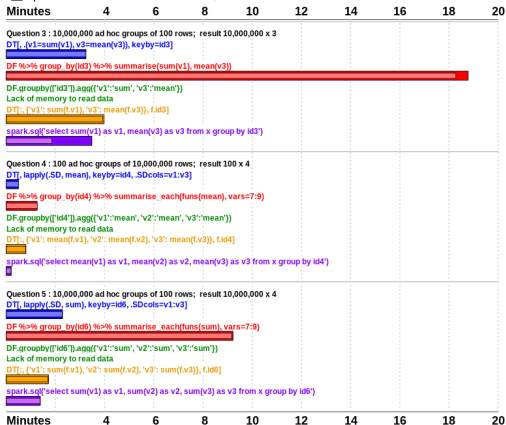
h2oai.github.io/db-benchmark

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

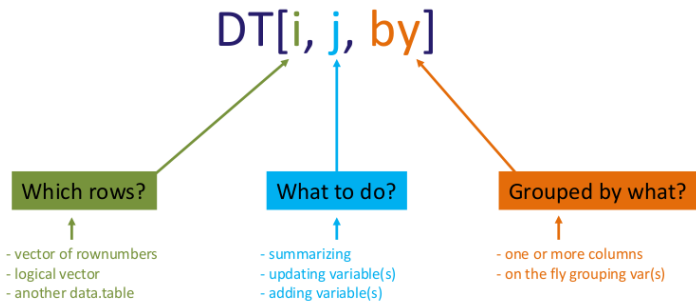
■ First time
■ Second time



h2oai.github.io/db-benchmark

Basic examples with `data.table`

General form



tutorial uRos (walhouti, 2018)

Examples

00-lib-dependencies.R

01-basics.R

02-counts.R

03-aggregating.R

04-group-by.R

05-group-by-SD.R

06-updating-variables.R

07-adding-variables.R

08-deleting-variables.R

09-joining-datasets.R

Examples with `data.table` and `ggplot`

Examples

01-scatterplot.R

02-boxplot.R

03-histogram.R

04-densitycurve.R

05-addingsmoothers.R

06-faceting.R

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/>