

Timings of common tasks using the **data.table** package in R

Matthew Dowle

July 21, 2010

* WORK IN PROGRESS *

This document contains a series of tests, followed by a summary table of various timings and comparisons. Please go straight to the summary table first [<here>](#) in which each row has a link back to the test.

This document is reproducible. Simply run the .Rnw file yourself in your environment to confirm the results. Also see `?vignette`, which says that `edit(vignette("datatable-timings"))` will extract the code from this document so you can easily work with it.

The .Rnw included in the package has N=10,000,000. This is a small number so that 'R CMD build' completes in a reasonable time (about 5 minutes). We don't want the nightly builds on R-Forge and CRAN to slow down just to run long timing comparisons. We have increased this to N=100,000,000 ourselves, and included the output on the [datatable homepage \(<link>\)](#).

Contents

1	Timing tests	1
1.1	Extraction	1
1.2	Grouping	2
1.3	Test 3	2
1.4	Test 4	2
1.5	Test 5	2
2	Summary table	2

1 Timing tests

1.1 Extraction

This is a repeat of the test in section 1 of the Introduction vignette. The syntax is explained there. This demonstrates the large difference in speed between vector scans and binary search. Therefore, please avoid using `==` in the `i` expression.

```
> n = ceiling(1e7/26^2) # 10 million rows
> DF = data.frame(x=rep(LETTERS,each=26*n),
+               y=rep(letters,each=n),
+               v=rnorm(n*26^2))
> DT = data.table(DF,key="x,y")
> tables()

      NAME      NROW  MB COLS  KEY
[1,] DT    10,000,068 153  x,y,v x,y
Total: 153MB

> tt=system.time(ans1 <- DF[DF$x=="R" & DF$y=="h",]); tt

      user  system elapsed
1.608    0.644    2.249
```

```
> ss=system.time(ans2 <- DT[J("R","h"),mult="all"]); ss

      user  system elapsed
0.008    0.000    0.008

> mapply(identical,ans1,ans2)

      x      y      v
TRUE TRUE TRUE
```

1.2 Grouping

This is a repeat of the test in section 2 of the Introduction vignette. The syntax is explained there.

```
> ttt=system.time(ans1 <- tapply(DF$v,DF$x,sum)); ttt

      user  system elapsed
5.408    0.648    6.060

> sss=system.time(ans2 <- DT[,sum(v),by=x]); sss

      user  system elapsed
0.304    0.156    0.463

> identical(as.vector(ans1), ans2$V1)

[1] TRUE
```

1.3 Test 3

1.4 Test 4

1.5 Test 5

2 Summary table

```
> ans

      base data.table times faster
==    2.249      0.008      281
tapply 6.060      0.463      13

> toLatex(sessionInfo())

• R version 2.11.1 (2010-05-31), i486-pc-linux-gnu

• Locale: LC_CTYPE=en_GB.utf8, LC_NUMERIC=C, LC_TIME=en_GB.utf8,
  LC_COLLATE=en_GB.utf8, LC_MONETARY=C, LC_MESSAGES=en_GB.utf8,
  LC_PAPER=en_GB.utf8, LC_NAME=C, LC_ADDRESS=C, LC_TELEPHONE=C,
  LC_MEASUREMENT=en_GB.utf8, LC_IDENTIFICATION=C

• Base packages: base, datasets, graphics, grDevices, methods, stats, utils

• Loaded via a namespace (and not attached): tools~2.11.1
```