

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

Matthew Dowle

Revised: July 12, 2012

(A later revision may be available on the [homepage](#))

\* 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

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

## 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),
+                 stringsAsFactors=FALSE)
> DT = data.table(DF,key="x,y")
> tables()
```

```
      NAME      NROW  MB COLS  KEY
[1,] DT    10,000,068 229 x,y,v x,y
Total: 229MB
```

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

```

      user  system elapsed
8.392    0.360    8.820

> head(ans1)

      x y          v
6642058 R h  0.1378294
6642059 R h  0.6649357
6642060 R h  0.4235033
6642061 R h  1.2628483
6642062 R h -0.8678718
6642063 R h -2.0255757

> dim(ans1)

[1] 14793      3

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

      user  system elapsed
0.008    0.004    0.013

> head(ans2)

      x y          v
1: R h  0.1378294
2: R h  0.6649357
3: R h  0.4235033
4: R h  1.2628483
5: R h -0.8678718
6: R h -2.0255757

> dim(ans2)

[1] 14793      3

> identical(ans1$v,ans2$v)

[1] 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
21.026    1.540    22.633

> head(ans1)

      A          B          C          D          E          F
-247.8614 -1144.4888  350.0649  264.6673  223.3804 -453.1123

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

      user  system elapsed
1.016    0.160    1.182

> head(ans2)

```

```

      x      V1
1: A  -247.8614
2: B -1144.4888
3: C   350.0649
4: D   264.6673
5: E   223.3804
6: F  -453.1123

```

```
> 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
==	8.820	0.013		678
tapply	22.633	1.182		19

```
> toLatex(sessionInfo())
```

- R version 2.15.1 (2012-06-22), x86\_64-pc-linux-gnu
- Locale: LC\_CTYPE=en\_GB.UTF-8, LC\_NUMERIC=C, LC\_TIME=en\_GB.UTF-8, LC\_COLLATE=en\_GB.UTF-8, LC\_MONETARY=en\_GB.UTF-8, LC\_MESSAGES=en\_GB.UTF-8, LC\_PAPER=C, LC\_NAME=C, LC\_ADDRESS=C, LC\_TELEPHONE=C, LC\_MEASUREMENT=en\_GB.UTF-8, LC\_IDENTIFICATION=C
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Loaded via a namespace (and not attached): tools~2.15.1