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

Matthew Dowle

Revised: November 13, 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

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 
  7.660   0.400   8.085 
```

```
> head(ans1)
```

```
        x y          v
6642058 R h -0.4405890
6642059 R h  1.2837447
6642060 R h -2.5158051
6642061 R h  0.8154541
6642062 R h  1.7378459
6642063 R h  0.8359195
```

```
> dim(ans1)
```

```
[1] 14793     3
```

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

```
   user  system elapsed 
  0.008   0.000   0.006 
```

```
> head(ans2)
```

```
   x y          v
1: R h -0.4405890
2: R h  1.2837447
3: R h -2.5158051
4: R h  0.8154541
5: R h  1.7378459
6: R h  0.8359195
```

```
> 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 
 19.401   1.388  20.850 
```

```
> head(ans1)
```

```
        A         B         C         D         E         F 
-561.4876 1435.8442  616.1335 -465.9479  323.9547 -861.6043 
```

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

```
   user  system elapsed 
  0.652   0.140   0.802 
```

```
> head(ans2)
```

```
   x        V1
1: A -561.4876
2: B 1435.8442
3: C  616.1335
4: D -465.9479
5: E  323.9547
6: F -861.6043
```

```
> 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
==      8.085      0.006         1347
tapply 20.850      0.802           25
```

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

- R version 2.15.2 (2012-10-26), x86_64-pc-linux-gnu
- Locale: LC_CTYPE=en_GB.UTF-8, LC_NUMERIC=C, LC_TIME=en_GB.UTF-8, LC_COLLATE=C, 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, grDevices, graphics, methods, stats, utils
- Other packages: data.table˜1.8.6
- Loaded via a namespace (and not attached): tools˜2.15.2