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

Matthew Dowle

Revised: September 3, 2013
(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
 11.225   0.316  11.567
```

```
> head(ans1)
```

```
        x y           v
6642058 R h -1.11672722
6642059 R h  0.07458134
6642060 R h  0.42563928
6642061 R h  0.62443876
6642062 R h -0.27554045
6642063 R h -0.47028993
```

```
> 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 -1.11672722
2: R h  0.07458134
3: R h  0.42563928
4: R h  0.62443876
5: R h -0.27554045
6: R h -0.47028993
```

```
> 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
 21.913   1.216  23.200
```

```
> head(ans1)
```

```
         A          B          C          D          E          F
 429.60884 -652.54668  242.62894 -599.44781 1402.31285  -98.82486
```

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

```
   user  system elapsed
  0.572   0.048   0.624
```

```
> head(ans2)
```

```
   x         V1
1: A  429.60884
2: B -652.54668
3: C  242.62894
4: D -599.44781
5: E 1402.31285
6: F  -98.82486
```

```
> 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
==     11.567      0.006         1927
tapply 23.200      0.624           37
```

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

- R Under development (unstable) (2013-08-30 r63776), x86_64-unknown-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=en_GB.UTF-8, 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˜1.8.10
- Loaded via a namespace (and not attached): tools˜3.1.0