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

Matthew Dowle

Revised: April 11, 2014
(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 = as.data.table(DF)
> system.time(setkey(DT,x,y))  # one-off cost, usually

   user  system elapsed
  0.184   0.060   0.244

> 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
  3.096   0.400   3.495
```

```
> head(ans1)
```

```
        x y          v
6642058 R h -0.6654744
6642059 R h -0.9309739
6642060 R h -0.3062619
6642061 R h -0.1371070
6642062 R h -0.7355361
6642063 R h  0.3557500
```

```
> dim(ans1)
```

```
[1] 14793     3
```

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

```
   user  system elapsed
  0.004   0.000   0.002
```

```
> head(ans2)
```

```
   x y          v
1: R h -0.6654744
2: R h -0.9309739
3: R h -0.3062619
4: R h -0.1371070
5: R h -0.7355361
6: R h  0.3557500
```

```
> 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
  5.988   1.560   7.551
```

```
> head(ans1)
```

```
         A          B          C          D          E          F
-270.29919 1082.74046 -772.47111 -524.97521  -53.17756 -186.90401
```

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

```
   user  system elapsed
  0.212   0.168   0.383
```

```
> head(ans2)
```

```
   x         V1
1: A -270.29919
2: B 1082.74046
3: C -772.47111
4: D -524.97521
5: E  -53.17756
6: F -186.90401
```

```
> 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
==     3.495      0.002         1747
tapply 7.551      0.383           19
```

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

- R version 3.0.3 Patched (2014-03-06 r65386), x86_64-unknown-linux-gnu
- Locale: LC_CTYPE=en_US.UTF-8, LC_NUMERIC=C, LC_TIME=en_US.UTF-8, LC_COLLATE=C, LC_MONETARY=en_US.UTF-8, LC_MESSAGES=en_US.UTF-8, LC_PAPER=en_US.UTF-8, LC_NAME=C, LC_ADDRESS=C, LC_TELEPHONE=C, LC_MEASUREMENT=en_US.UTF-8, LC_IDENTIFICATION=C
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: data.table˜1.9.3
- Loaded via a namespace (and not attached): Rcpp˜0.11.1, plyr˜1.8.1, reshape2˜1.2.2, stringr˜0.6.2, tools˜3.0.3