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

Matthew Dowle

April 12, 2010

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
> DT = data.table(x=rep(LETTERS,each=26*n),
+                 y=rep(letters,each=n),
+                 v=rnorm(n*26^2),
+                 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 <- DT[DT$x=="R" & DT$y=="h",]); tt
   user  system elapsed
  4.376   1.036   5.484
> ss=system.time(ans2 <- DT[J("R","h"),mult="all"]); ss
   user  system elapsed
  0.008   0.000   0.009
> if(!identical(ans1,ans2)) stop("Test 1 not identical")
```

1.2 Grouping

Repeat from Introduction.

2 Summary table

```
> tt[3]

elapsed
  5.484

> ss[3]

elapsed
  0.009

> toLatex(sessionInfo())
```

- R version 2.10.1 (2009-12-14), `i486-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=C`, `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, graphics, grDevices, methods, stats, utils
- Other packages: data.table˜1.3, ref˜0.97
- Loaded via a namespace (and not attached): tools˜2.10.1