Towards Terrabytes of TAQ

John W. Emerson (**Jay**) and Michael J. Kane (**Mike**)

Yale University

john.emerson@yale.edu, michael.kane@yale.edu

`http://www.stat.yale.edu/~jay/RinFinance2012/`

R in Finance 2012

Motivation

A Yale College undergraduate did his senior essay with me on asset price & volume sensitivity to Reuters news releases.

He hit the “big data wall” with **a single day** of TAQ data.

He had all of 2010 to analyze and needed a little help.

His Reuters data scrape, example 1

Turning Text into Data

FBR raises price target on Google shares

Recommend | Be the first of your friends to recommend this.

Tweet 0
Share
Share this
+1 0
Email
Print

Mon Jan 4, 2010 11:25am EST

Jan 4 (Reuters) - FBR Capital Markets raised its price target on the stock of Google Inc (GOOG.O) 19 percent to $810, and said the world's No. 1 Internet search firm is the best positioned of the large-cap internet companies.

Related News

Wall Street cool on rumored Google phone launch
Mon, Jan 4 2010

Baird raises Intel to outperform
Mon, Jan 4 2010

ANALYSIS-Amazon could pay for Kindle sales coyness
Thu, Dec 31 2009

```
    Date                                  Title        Time Comments Comp1 Comp2 Comp3 Comp4 Comp5
20100104 FBR raises price target on Google shares 11:25am EST        0  GOOG  <NA>  <NA>  <NA>  <NA>
```

His Reuters data scrape, example 2

More Text Into Data

CES-Freescale takes aim at tablet computer market

Recommend | Be the first of your friends to recommend this.

Mon Jan 4, 2010 12:00am EST

* Sub-$200 devices would run Google's Android
* Tablet products could hit stores this summer
* Freescale prototypes at Consumer Electronics Show

By Gabriel Madway

SAN FRANCISCO, Jan 4 (Reuters) - Chipmaker Freescale Semiconductor Inc [FSLSM.UL] is staking its claim on the tablet computer market, an emerging product category that will generate plenty of interest in 2010.

Although next-generation tablet PCs are scarcely evident on the market, the technology world is abuzz about their potential, as Apple Inc (AAPL.O) is expected to unveil its offering in 2010.

Freescale's announcement comes ahead of this week's Consumer Electronics Show in Las Vegas, where rival chipmakers are expected to show off new so-called smartbooks, which aim to bridge the gap between laptops and smartphones.

Privately held Freescale unveiled its reference design for a 7-inch, touchscreen tablet running on the company's low-power ARM-based processor and priced at less than $200.

The company said such a device will be able to run either Google Inc's (GOOG.O) Android mobile software or Linux, with Wi-Fi and 3G capability.

Tweet 0
Share
Share this
+1 0
Email
Print

Related News

Lenovo unveils ThinkPad design with AMD chips
Sun, Jan 3 2010

Phones, PCs to drive tech rally into 2010
Thu, Dec 31 2009

Amazon could pay for Kindle sales coyness
Thu, Dec 31 2009

Apple shares hit new high on tablet excitement
Thu, Dec 24 2009

Apple to host product event in January: report
Thu, Dec 24 2009

Date	Title	Time	Comments	Comp1	Comp2	Comp3	Comp4	Comp5
20100104	CES-Freescale takes aim at tablet computer market	12:00am EST	0	AAPL	GOOG	AMZN	QCOM	NVDA

Raw data: Daily 2010 TAQ files from Wharton

Each day of TAQ data: lines like

```
  SYMBOL      DATE     TIME PRICE SIZE CORR COND EX
1      A 20100104 9:30:02 31.32   98    0    Q  T
2      A 20100104 9:30:50 31.39  100    0    F  T
3      A 20100104 9:30:50 31.40  300    0    F  T
```

what	typical day	worst day
compressed size (gz)	90 MB	210 MB
uncompressed size (CSV)	1 GB	2.5 GB
size in a data frameafter `read.csv()`	1.3 GB	3.4 GB
rows	24 million	65 million
peak memory usage during `read.csv()`	2.7 GB	6.5 GB
waiting during the `read.csv()`	90 seconds	a few minutes

Subsequent **basic** data manipulation and exploration on **a single day**: dangerous, caused swapping (turning a 4-minute job into a 24-hour bomb)

Hardware

- Newish quad-core laptop (nothing special)
- 8 GB RAM
- 2 TB external eSATA hard drive
- Not Windows

Abstract (revised)

I will give away a trivial bit code for creating a $\sim$ ~~500~~ **250** GB big.matrix **of integer values** ($\sim$ ~~6~~ **6.38** billion rows, 10 columns) containing a year of TAQ data associated with $\sim$ 6500 ticker symbols featured in Reuters news releases. It took $\sim$ ~~24~~ **15** hours to create ~~the~~ **two** big.matrix **objects** from the original compressed TAQ files, working on a stock Dell laptop with 8 GB RAM and a 2 TB eSATA external hard drive. Note that R itself is incapable of storing a single column of this data set in a single vector (even if 64+ GB of RAM were available). I'll demonstrate basic functionality of packages bigmemory, bigtabulate, and biganalytics for working directly with the data. All examples are scalable, limited only by available disk space.

`http://www.stat.yale.edu/~jay/RinFinance2012/`

www.stat.yale.edu/~jay/RinFinance2012/

```
for (i in 1:length(files)) {
  x <- read.csv(gzfile(paste(taqdir, files[i], sep='')),
                header=TRUE, as.is=TRUE)
  ...
  tickernum <- match(x$SYMBOL, tickers$ticker)
  these <- x$SYMBOL %in% tickers$ticker
  these <- these & dayminute >= as.integer(570) &
                   dayminute <= as.integer(960)
  ...
  b <- big.matrix(ncol=length(thenames), nrow=sum(these),
         type='integer', dimnames=list(NULL, thenames),
         backingfile='sensible binary file name',
         descriptorfile='sensible descriptor file name',
         backingpath=dailybackingdir)
  ...
  b[,4] <- tickernum[these]; rm(tickernum); gc()
  b[,2] <- as.integer(100*x[these,'PRICE']); gc()
  ...
}
```

www.stat.yale.edu/~jay/RinFinance2012/

The following runs instantly, technically giving you access to each of the daily `big.matrix` objects in turn; we only make use of the dimension for calculating the total number of rows we have (about 6.38 billion).

```
numrows <- 0
for (i in 1:length(files)) {
  b <- attach.big.matrix(
         paste('taq_', datestamps[i], '.desc', sep=''),
         path=dailybackingdir)
  numrows <- numrows + nrow(b)
}
```

www.stat.yale.edu/~jay/RinFinance2012/

Now create the two `big.matrix` objects to hold the full year. The first is just 10 columns of TAQ data, essentially the result of a giant `rbind()`. The second has a row for each ticker symbol, and a column for each minute of the year (with NAs for minutes with no data).

```
all <- big.matrix(ncol=ncol(b), nrow=numrows,
                  type='integer',
                  dimnames=list(NULL, colnames(b)),
                  backingfile='alltaq.bin',
                  descriptorfile='alltaq.desc',
                  backingpath=bigbackingdir)

agg <- big.matrix(ncol=length(files)*(960-570+1),
                  nrow=nrow(tickers), init=NA,
                  dimnames=list(tickers$ticker, NULL),
                  backingfile='aggtaq.bin',
                  descriptorfile='aggtaq.desc',
                  backingpath=bigbackingdir)
```

www.stat.yale.edu/~jay/RinFinance2012/

Create the matrix of naive minute price averages for each ticker symbol.

```
library(foreach); library(doMC); registerDoMC(2)
startrow <- 1
for (i in 1:length(files)) {
  b <- attach.big.matrix( <omitted>)
  minuteinds <- bigsplit(b, ccols='dayminute')
  ans <- foreach(j=names(minuteinds)) %dopar% {
    foo <- tapply(b[minuteinds[[j]],'PRICE'],
                  b[minuteinds[[j]],'tickernum'],
                  mean, simplify=TRUE)
    agg[as.numeric(names(foo)),
        (i-1)*(960-570+1)+as.numeric(j)-570+1] <- foo
    return(TRUE)
  }
  startrow <- startrow + nrow(b)
}
```

www.stat.yale.edu/~jay/RinFinance2012/

Plot the minute-by-minute naive average for Google:

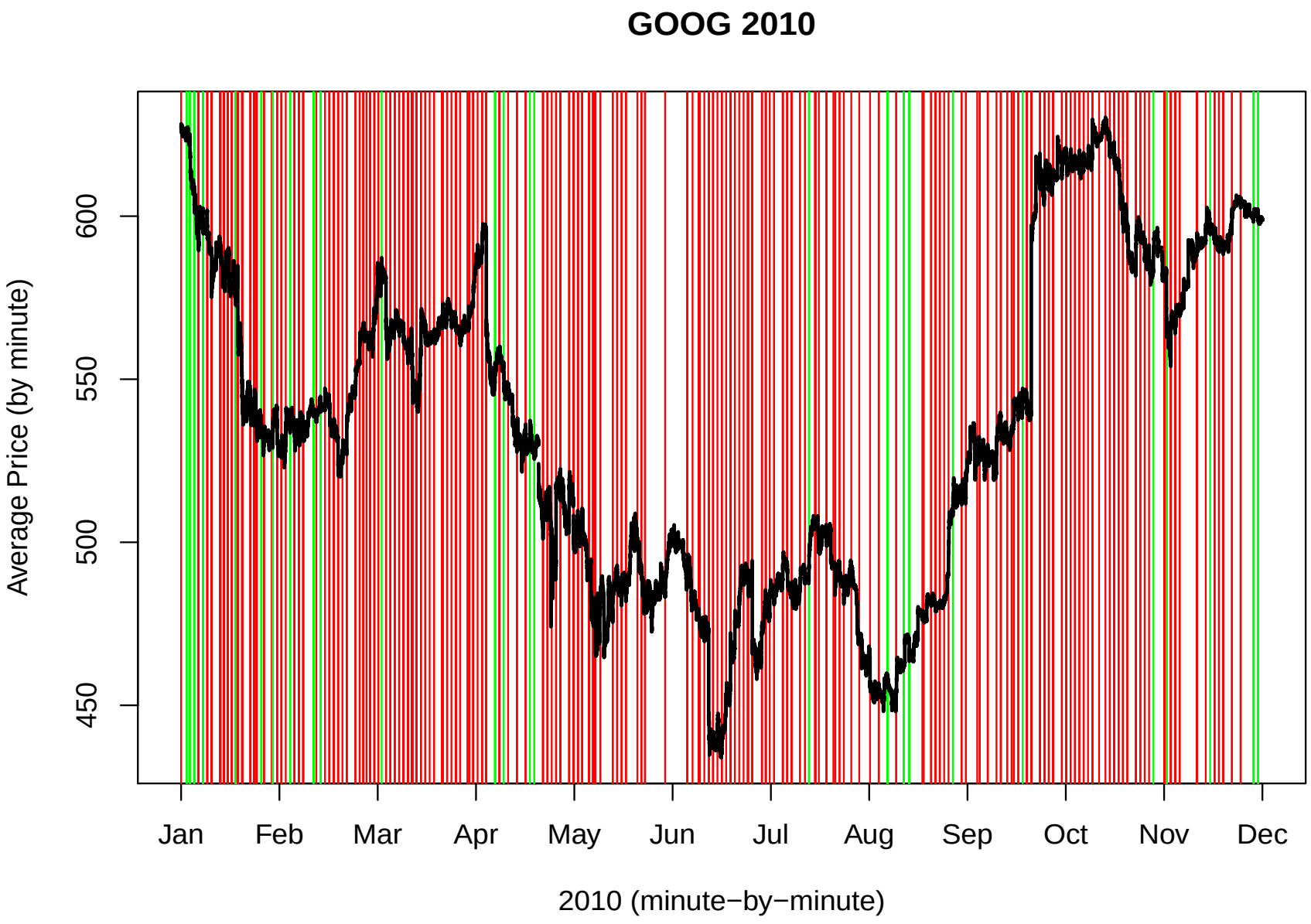
```
ticker <- 'GOOG'

agg <- attach.big.matrix('aggtaq.desc',
                         path=bigbackingdir)

...

plot(agg[ticker,]/100, type='l', <details omitted>,
     xlab='2010 (minute-by-minute)',
     ylab='Average Price (by minute)',
     main=paste(ticker,'2010')))

...
```


GOOG 2010
Average Price (by minute)
450
500
550
600
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
2010 (minute−by−minute)

Thanks!

- Dirk Eddelbuettel, Bryan Lewis, Steve Weston, and Martin Schultz, for their feedback and advice over the last few years
- Bell Laboratories (Rick Becker, John Chambers and Allan Wilks), for development of the S language
- Ross Ihaka and Robert Gentleman, for their work and unselfish vision for R
- The R Core team

`http://www.stat.yale.edu/~jay/RinFinance2012/`