mmm, mmm, mmap good

Unlimited size, zero memory, blazingly fast data for R (and more!)

Jeff Ryan CRUG - June 16, 2021

I am …

R user since 2002

Creator of quantmod, xts, and other R packages

Creator/Co-founder R/Finance 2009 (12th happening this Oct!)

Instigator of (C)RUG - 2008, arguably first RUG anywhere! (yes Joe, before SF)

Recently (ex) Citadel quant

Building multiple companies now. Data (quantkiosk) and new database (shhhh).

Consulting new quant/systematic shops and independents (quantatlarge)

Founder of rpeat.io (devops for devs and researchers)

What is mmap ?

a.k.a. memory mapped pages

OS level system to move bytes from disk to memory

a.k.a. “virtual memory”

Create mapping	Access bytes	Remove mapping
mmap( )	m[0] … m[1002007700]	munmap( )

What is mmap ?

a.k.a. memory mapped pages

Used extensively by your operating system

Nearly every database relies on this to move data from disk to process

Binary data is cross language and platform!

You can build your own database with 2 commands !!!!!

(Yes, seriously… well sort of… mostly)

Let’s try it!

Three Billion Doubles

```
> r <- rnorm(3e9)
>
> gc()
            used    (Mb) gc trigger    (Mb)   max used    (Mb)
Ncells    270699    14.5     655322    35.0     447568    24.0
Vcells 3000542769 22892.4 4323372718 32984.8 3000543773 22892.4
> writeBin(r, 'r.bin')
```

Create data, write file

```
> system.time( r <- mmap('r.bin', double()) )
   user  system elapsed
      0       0       0
> r
<mmap:r.bin>  (double) num [1:3e+09] 0.08358927 0.8656447 0.6805671 ...
> r[3e9]
[1] -0.5641944
> r[2328888212]
[1] 1.20722
>
>
> gc()
         used (Mb) gc trigger (Mb) max used (Mb)
Ncells 341043 18.3     655322   35   486087 26.0
Vcells 697328  5.4    8388608   64  1776211 13.6
>
```

mmap and index

Let’s try it!

Three Billion Doubles (23GB)

```
> r <- rnorm(3e9)
>
> gc()
            used    (Mb) gc trigger    (Mb)   max used    (Mb)
Ncells     270699    14.5     655322    35.0     447568    24.0
Vcells 3000542769 22892.4 4323372...
> writeBin(r, 'r.bin')
```

Create data, write file

mmap package mimics system call

magically converts bytes to R types (16 types)

macOS + Windows + Linux/Unix

stable since 2010

supports C-style structs, big/little endianness, 2^{53} elements

```
> system.time( r <- mmap('r.bin', double()) )
   user  system elapsed
      0       0       0
> r
<mmap:r.bin> (double) num [1:3e+09] 0.08358927 0.8656447 0.6805671 ...
> r[1]
[1] -0.5641944
> r[...]
[1] 1.20722
>
>
> gc()
            used  (Mb) gc trigger  (Mb)  max used  (Mb)
Ncells 341043 18.3     655322   35   486087 26.0
Vcells 697328  5.4    8388608   64  1776211 13.6
```

mmap and index

Designed to be infrastructure - build upon it!

Searching EDGAR 13F

indexing Package Example

13F-HR's contain quarterly holdings for HF
XML - converted to tables (thanks QuantKiosk!)
~57 million holdings since 2013

indexing package

built with mmap
data.frame semantics
database performance
easy to use

Searching EDGAR 13F

indexing Package Example

```
> db[issuerTicker=='IBM' & reportPeriod == 20201231, data.frame(impliedPrice=value*1000 / shrsOrPrnAmt, filerCik), limit=5]
  impliedPrice filerCik
1     125.8945  1000275
2     125.9275  1000275
3     125.9048  1000275
4     125.8797  1000275
5     125.8801  1000275
> system.time( db[issuerTicker=='IBM' & reportPeriod == 20201231, data.frame(impliedPrice=value*1000 / shrsOrPrnAmt, filerCik), limit=5] )
   user  system elapsed
  0.064   0.008   0.073
> gc()
         used (Mb) gc trigger (Mb) max used  (Mb)
Ncells 384104 20.6     655322   35   655322  35.0
Vcells 795034  6.1   19125234  146 42807331 326.6
> db
Indexed Environment: 57925096 rows by 8 columns
```

query

transform/select

limit, order

Insane fast

No RAM harmed

14GB (disk)

NoDB DB powered by mmap and base R

Searching The Stars

RCosmos Package Example

rcosmo: Handle and Analyse Spherical, HEALPix and Cosmic Microwave Background data on a HEALPix grid.

build error | CRAN 1.1.2 | CRAN 1.1.2 – a year ago | downloads 14K | downloads 89/week | lifecycle maturing

Use this R package as an advanced toolkit for performing Cosmic Microwave Background (CMB) data analytics. The CMB is remnant electromagnetic radiation from the epoch of recombination. As an ancient source of data on the early universe, the CMB is helping us unlock the mysteries of the Big Bang and the structure of time and space. With increasingly high resolution satellite data, intensive investigations in the past few years have resulted in many physical and mathematical results to characterize CMB radiation.

Publication

Daniel Fryer, Ming Li and Andriy Olenko, rcosmo: R Package for Analysis of Spherical, HEALPix and Cosmological Data, The R Journal (2020) 12:1, pages 206-225.

```
> filename1 <- "CMB_map_smica2048.fits"
> downloadCMBMap(foreground = "smica", nside = 2048, filename1)
> system.time(sky <- CMBDataFrame(filename1))
   user  system elapsed
   1.36    0.29    1.73

> system.time(fits <- FITSio::readFITS(filename1))
   user  system elapsed
 822.28   90.05  942.14
```

Uses mmap to access FITS data format

talk |> more

www.github.com/jaryan

www.linkedin.com/in/jeffreyaryan/