Subject: Re: Reading a very large ascii data file
Posted by Martin Schultz on Fri, 24 Aug 2001 18:01:03 GMT
View Forum Message <> Reply to Message

mvukovic@taz.telusa.com (Mirko Vukovic) writes:

> I am reading some large ascii data files in csv (comma separated
> fields) format, and would like to speed the process up.
>
> I recall someone discussing reading such files as binaries and then
> converting to ascii after finding line breaks, but was un-able to find
> the discussion on the group.
>
> Can anyone offer pointers, code, or suggestions on who might have
> discussed it (so that I can look again on the newsgroup).
>
> Thanks,
>
> Mirko

Well, the most important speed-up is probably gained from "blocking" the input. At least, if you read the file in that "classical" way as:

```
readf, lun, line
text = [ text, line ]
```

This is very unefficient, and shoul dbe replaced with something like:

```
count = 0L
text = StrArr(10000L)
WHILE NOT Eof(lun) DO BEGIN
  Readf, lun, line
  text = line
  count = count + 1
  IF count MOD 10000L EQ 0 THEN text = [ text, StrArr(10000) ]
ENDWHILE
text = text[0:count-1]
```

In principle, you can use a similar technique to read the file in binary format as well (not tested):

```
LEN = 1000000L
text = BytArr(LEN)
WHILE NOT Eof(lun) DO BEGIN
  ReadU, lun, text, count=count     ;; wasn't this something lately?
  IF count EQ LEN THEN text = [ text, BytArr(LEN) ]
ENDWHILE
```

```
;; The following is system dependent
cr = String(13B)
lf = String(10B)
crlf = Where(text EQ lf, cnt)   ;; these are your line breaks in Unix
   ;; on a Mac it's simply cr, I believe, and in Windows it's cr+lf
```

Hope this helps somewhat,

Martin

--
[[[[[[[[[
[[Dr. Martin Schultz   Max-Planck-Institut fuer Meteorologie   [[
[[Bundesstr. 55, 20146 Hamburg   [[
[[phone: +49 40 41173-308   [[
[[fax: +49 40 41173-298   [[
[[martin.schultz@dkrz.de   [[
[[[[[[[[[