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   [[
[[[[[[[[[

Subject: Re: Reading a very large ascii data file
Posted by Paul van Delst on Fri, 24 Aug 2001 18:11:44 GMT

View Forum Message <> Reply to Message

Mirko Vukovic wrote:
>
> 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).

Can you provide more information about your data files? E.g. are the number of columns fixed? Are the number of lines fixed? If not, is there a maximum number of lines which the files won't exceed?

Try the DDREAD.PRO and associated IDL code. Have a look at

  http://www.dfanning.com/tips/unknown_rows.html

for some issues and a link to the source code.

paulv

--
Paul van Delst A little learning is a dangerous thing;
CIMSS @ NOAA/NCEP Drink deep, or taste not the Pierian spring;
Ph: (301)763-8000 x7274 There shallow draughts intoxicate the brain,
Fax:(301)763-8545 And drinking largely sobers us again.
Alexander Pope.

Subject: Re: Reading a very large ascii data file
Posted by mvukovic on Fri, 24 Aug 2001 22:22:55 GMT
View Forum Message <> Reply to Message

Paul van Delst <paul.vandelst@noaa.gov> wrote in message
news:<3B8698E0.B3F13251@noaa.gov>...
> Mirko Vukovic wrote:
>>
>> 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).
>
> Can you provide more information about your data files? E.g. are the number of columns
> fixed? Are the number of lines fixed? If not, is there a maximum number of lines which the
> files won't exceed?
>
> Try the DDREAD.PRO and associated IDL code. Have a look at
>
> http://www.dfanning.com/tips/unknown_rows.html
>
> for some issues and a link to the source code.
>
> paulv

Thanks for the comments,

The file format is variable. The file contains a log of data of a
variable number of channels, and of arbitrary duration. It is
generated by the TrendLink software from Fluke.

The file consists of a header, which has as many lines as diagnostics.
Next comes the data, with one column for the time and date, and a
column each for each channel.

I therefore use a two-pass system. In the first, I read all the lines, and count their number, and from the last line also extract the number of channels.

With this info, I then initialize the header and data structures, and then go again through the file, and store the stuff.

In that sense, I am not using the very slow procedure noted by martin (appending a line to the matrix). However, I am going explicitly through a very long loop, twice.

One methode may be to open the file in binary mode, get info about the number of bytes, initialize a byte vector to appropriate size, and then read the file into it. Now, with the file stored in memory (although it can be megabytes in size), go through it, ``reading'' line by line.

This actually looks to be a quite generic procedure. Any idea whether it has been implemented already?

Any more suggestions?

Thanks,

Mirko

Subject: Re: Reading a very large ascii data file
Posted by R.Bauer on Sat, 25 Aug 2001 11:32:09 GMT
View Forum Message <> Reply to Message

Mirko Vukovic wrote:
>
> Paul van Delst <paul.vandelst@noaa.gov> wrote in message news:<3B8698E0.B3F13251@noaa.gov>...
>> Mirko Vukovic wrote:
>>>
>>> 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).
>>

>> Can you provide more information about your data files? E.g. are the number of columns
>> fixed? Are the number of lines fixed? If not, is there a maximum number of lines which the
>> files won't exceed?
>>
>> Try the DDREAD.PRO and associated IDL code. Have a look at
>>
>> http://www.dfanning.com/tips/unknown_rows.html
>>
>> for some issues and a link to the source code.
>>
>> paulv
>
> Thanks for the comments,
>
> The file format is variable. The file contains a log of data of a
> variable number of channels, and of arbitrary duration. It is
> generated by the TrendLink software from Fluke.
>
> The file consists of a header, which has as many lines as diagnostics.
> Next comes the data, with one column for the time and date, and a
> column each for each channel.
>
> I therefore use a two-pass system. In the first, I read all the
> lines, and count their number, and from the last line also extract the
> number of channels.
>
> With this info, I then initialize the header and data structures, and
> then go again through the file, and store the stuff.
>
> In that sense, I am not using the very slow procedure noted by martin
> (appending a line to the matrix). However, I am going explicitly
> through a very long loop, twice.
>
> One methode may be to open the file in binary mode, get info about the
> number of bytes, initialize a byte vector to appropriate size, and
> then read the file into it. Now, with the file stored in memory
> (although it can be megabytes in size), go through it, ``reading''
> line by line.
>
> This actually looks to be a quite generic procedure. Any idea whether
> it has been implemented already?
>
> Any more suggestions?
>
> Thanks,
>
> Mirko

Dear Mirko,

you should use our read_data_file.

This routine itselfs separates header, datablock and trailer.
The datablock must be a tabular of numbers.
You got returned a structure .header, .separator, .data
because you haven't a trailer.

data is a tabular of n columns and m lines

This routine is very fast.

http://www.fz-juelich.de/icg/icg1/idl_icglib/idl_source/idl_
html/dbase/download/read_data_file.tar.gz

regards

Reimar

--
Reimar Bauer

Institut fuer Stratosphaerische Chemie (ICG-1)
Forschungszentrum Juelich
email: R.Bauer@fz-juelich.de
http://www.fz-juelich.de/icg/icg1/
========== ======
a IDL library at ForschungsZentrum Juelich
http://www.fz-juelich.de/icg/icg1/idl_icglib/idl_lib_intro.h tml

http://www.fz-juelich.de/zb/text/publikation/juel3786.html
========== ======

read something about linux / windows
http://www.suse.de/de/news/hotnews/MS.html

Subject: Re: Reading a very large ascii data file
Posted by R.Bauer on Sat, 25 Aug 2001 11:34:02 GMT
View Forum Message <> Reply to Message

Our server is down and will be up on Monday (I hope)

Reimar

--

Reimar Bauer

Institut fuer Stratosphaerische Chemie (ICG-1)
Forschungszentrum Juelich
email: R.Bauer@fz-juelich.de
http://www.fz-juelich.de/icg/icg1/
 ========== ======
a IDL library at ForschungsZentrum Juelich
 http://www.fz-juelich.de/icg/icg1/idl_icglib/idl_lib_intro.h tml

http://www.fz-juelich.de/zb/text/publikation/juel3786.html
 ========== ======

read something about linux / windows
http://www.suse.de/de/news/hotnews/MS.html

Subject: Re: Reading a very large ascii data file
Posted by mvukovic on Tue, 28 Aug 2001 15:07:24 GMT
View Forum Message <> Reply to Message

Reimar Bauer <r.bauer@fz-juelich.de> wrote in message
news:<3B878CB9.EF89AFC6@fz-juelich.de>...
> Mirko Vukovic wrote:
>>
>> Paul van Delst <paul.vandelst@noaa.gov> wrote in message
news:<3B8698E0.B3F13251@noaa.gov>...
>>> Mirko Vukovic wrote:
>>>>
...lots of stuff deleted

I tried read_data_file. It does not quite work, as in my case the
first column consists of date and time (non-numeric characters).
However, I studied the routines, and picked up some salient points and
subroutines. The code is now much faster. Thank you very much.

Mirko
