

Re: How to filter a .csv file based on the integer value in one specific field (per record)

## Re: How to filter a .csv file based on the integer value in one specific field (per record)

---

*Source:* <http://unix.derkeiler.com/Newsgroups/comp.unix.shell/2008-01/msg00390.html>

---

- *From:* Brian Greaney <[brian@xxxxxxxxxxxxxxxxxxxxx](mailto:brian@xxxxxxxxxxxxxxxxxxxxx)>
  - *Date:* Mon, 14 Jan 2008 22:00:39 GMT
- 

On Mon, 14 Jan 2008 15:39:47 -0600, Ed Morton wrote:

On 1/14/2008 3:23 PM, Brian Greaney wrote:

On Mon, 14 Jan 2008 14:59:17 -0600, Ed Morton wrote:

On 1/14/2008 2:40 PM, Brian Greaney wrote:

Hi, hope someone can help me (again!)  
I have a large .csv file full of text &  
numbers.  
I would like to 'filter' this file based on  
several 'keys', 2 text strings  
(use grep I think) and the integer value of a  
specific field being greater  
than a certain value (e.g. 100000).  
I can see a way of using grep to filter on the  
two strings, but can't see  
an elegant way of getting the numerical test  
on a specific field, bearing  
in mind numbers occur in other fields.  
The text fields I filter on are of the form  
ABC12 the integer from 2 to 6  
digits  
Hope this is clear (and not too  
dumb/newbie/rtfm a question) :)

There are many different ways a ".csv" file could be  
structured as there's no  
specific standard for one (though there are some attempts at  
such on the

Re: How to filter a .csv file based on the integer value in one specific field (per record)

internet) so you need to post a small set of sample input and expected output.

Make sure you tell us which specific fields you're interested in. If the 2 text fields are field 3 and field 7, and the integer's in field 12, the solution MAY be (but probably isn't) as simple as:

```
awk 'BEGIN{FS=OFS=","}$3 == "ABC12" && $7 == "ABC12" && ($12 >=2) && ($12 <=6)' file
```

Ed.

Actual .csv file is 109 columns by >3,000 rows (and yes it should be a database, it just grew and grew) so I didn't really want a post filled with junk.

Filter would typically be Anode (field 45)=LAC01 & Bnode (field 56)=CLH01  
field 3> 100000

A sample row is below although I have masked some text out with xxxx  
The fields 45 = LAC01, 56=LHR01 & 3 = 4842, so I would filter out this record. Note there are a large number of blanks and I've had to chop the lines up to keep the mail program happy!:

```
LAC0102 3 9 LHR0102 2 10,LAC01 3 2 LHR01 3 2,4842,,xxxxxxx Old  
Tower,Green,,,,,,,,LXXX SXXXXXXXXX,Green,,,,,,,,64000,64k  
Voice,VXXXXX,TPX 4969 – VXXXXXX XXX XXXXXXXXX,4w  
Analogue,,,,LAC0102,3,10chEM,,3,30227,  
,28,27,9,3,2,LAC01,10,10,38,38,40,10,10,1,2,40,LHR01,LHR01,3,2,LHR0102,2,10chEM,  
2,6,5,10,,30205,  
,,,,,CXXX810330,CXXX810330,FXXX342947,FXXS342947,CXXX612010,CXXX12010,FXXX34  
FXXX43412,CXXX805295,CXXX805295,0.484,1141,Bundled  
T/S,,,,,,,,4.1,18-Dec-07,,,,0,2 3 3,2 0 3,
```

Thanks for your help!

Based on these requirements/samples for filtering:

- 1) The text fields I filter on are of the form ABC12 the integer from 2 to 6 digits
- 2) Filter would typically be Anode (field 45)=LAC01 & Bnode (field 56)=CLH01  
field 3> 100000
- 3) The fields 45 = LAC01, 56=LHR01 & 3 = 4842, so I would filter out this record.

the text filtering requirement is pretty clear (3 all-upper-case letters followed by 2 digits), but the integer still isn't since "field 3>100000"

Re: How to filter a .csv file based on the integer value in one specific field (per record)

matches neither "from 2 to 6 digits" nor "3 = 4842".

If I assume you actually just want to find integer values between 10 and 100000 then as long as there's no quoted or escaped commas WITHIN the fields and the 3rd field is guaranteed to be an integer, this should do it:

```
awk 'BEGIN{FS=OFS=","}
$45 ~ /^[[:upper:]][[:upper:]][[:upper:]][[:digit:]][[:digit:]]$/ &&
$56 ~ /^[[:upper:]][[:upper:]][[:upper:]][[:digit:]][[:digit:]]$/ &&
($3 >= 10) && ($3 <= 100000)' file
```

Regards,

Ed.

Ed

With your great help I'm getting there.. and I've just found a hiccup  
In the above my field 45 will always start LAC but could be O1 or O2, I  
think therefore I can just loose the 2 [[:digit:]] parts? – BUT I need it  
NOT EQUALS. Can I also pass the match as a string?? (pattern="^CLH01\\$")  
? I think I need to take your other suggestion about the book.

PS the 3rd field is always an integer with no commas

.