Data Analysis in Geophysics
ESCI 7205

Bob Smalley
Room 103 in 3892 (long building), x-4929

Tu/Th - 13:00-14:30
CERI MAC (or STUDENT) LAB

Lab – 21, 11/07/13

Computers make it easier to do a lot of things, but most of the things they make it easier to do don't need to be done.

Andy Rooney

Computers may save time but they sure waste a lot of paper. About 98 percent of everything printed out by a computer is garbage that no one ever reads.

Andy Rooney

Introduction

AWK programming language

awk Programming Language

standard UNIX language that is geared for text processing and creating formatted reports

But is very valuable to seismologists because it uses floating point math, and is designed to work with columnar data

syntax similar to C and bash

one of the most useful UNIX tools at your command

(Softpanorama says "AWK is a simple and elegant pattern scanning and processing language.")

`awk` considers text files as having records (lines), which have fields (columns)

Performs floating & integer arithmetic and string operations

Has loops and conditionals

Can define your own functions (subroutines)

Can execute UNIX commands within the scripts and process the results

Basic structure of `awk` use

The essential organization of an `awk` program follows the form:

```
pattern { action }
```

The pattern specifies when the action is performed.

Like most UNIX utilities, awk is line oriented.

It may include an explicit test pattern that is performed with each line read as input.

If there is no explicit test pattern, or the condition in the test pattern is true, then the action specified is taken.

The default test pattern (no explicit test) is something that matches every line, i.e. is always true.

(This is the blank or null pattern.)

Versions/Implementations

awk: original awk

nawk: new awk, dates to 1987

gawk: GNU awk has more powerful string functionality

- NOTE -

We are going to use **awk** as the generic program name (like Kleenex for facial tissue)

Wherever you see **awk**, you can use **nawk** (or **gawk** if you are using that on a LINUX box).

Every CERI UNIX system has at least **awk**. The SUNs also have nawk.

Rumor has it that in OS X **awk** is actually **nawk**, although I've not been able to establish this.

If this is the case no changes to **nawk** codes are necessary (except the name).

If you have lots of scripts with **nawk** and it is not found/installed, put the following in your .cshrc or .bashrc file

```
alias nawk='awk' (csh) or alias nawk=awk (bash)
```

Command line functionality

you can call **awk** from the command line two ways, we have seen/used the first – put the **awk** commands on the command line in the construct '{...}' or read **awk** commands from a script file.

```
awk [options] 'pattern { commands }' variables infile(s)
awk –f scriptfile variables infile(s)
```

or you can create an executable **awk** script

```
%cat << EOF > test.awk
#!/usr/bin/awk
some set of commands
EOF

%chmod 755 test.awk
%./test.awk
```

How awk treats text

awk commands are applied to every record or line of a file that passes the test.

it is designed to separate the data in each line into a number of fields and can processes what is in each of these fields.

essentially, each field becomes a member of an array with the first field identified as `$1`, second field `$2` and so on.

`$0` refers to the entire record (all fields).

Field Separator

How does awk break the line into fields.
It needs a field separator.
The default field separator is one or more white spaces

```
$1   $2     $3   $4    $5   $6    $7      $8         $9        $10    $11
1    1918    9    22    9    54   49.29  -1.698    98.298     15.1   ehb
```

So `$1 = 1, $2=1918, …, $10=15.1, $11=ehb`

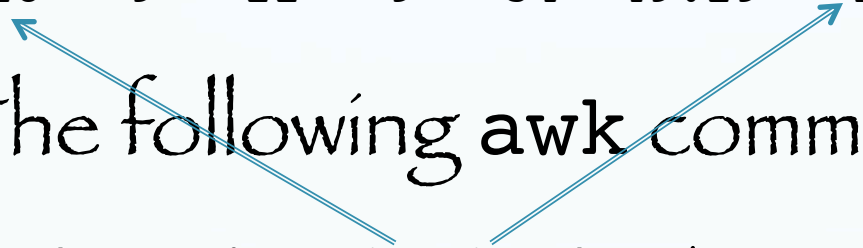
Notice that the fields may be integer, floating point (have a decimal point) or strings.

awk is generally smart enough to figure out how to use them.

print

print is one of the most common awk commands (e.x. for an input line)

```
1 1918    9    22    9    54   49.29   -1.698    98.298    15.1   ehb
```



The following awk command '{...}' will produce

```
%awk '{ print $2 $8}'   /somewhere/inputfile
1918-1.698
```

The two output fields (1918 and -1.698) are run together - this is probably not what you want.

This is because awk is insensitive to white space in the inside the command '{...}'

print

```
%awk '{ print $2 $8}'   /somewhere/inputfile
1918-1.698
```

The two output fields (1918 and -1.698) are run together – two solutions if this is not what you want.

```
%awk '{ print $1 "   " $8}'   /somewhere/inputfile
1918 -1.698
%awk '{ print $1, $8}'   /somewhere/inputfile
1918 -1.698
```

The awk print command different from the UNIX printf commad (more similar to echo).

any string (almost – we will see the caveat in a minute) or numeric text can be explicitly output using double quotes "..."

Assume our input file looks like this

```
1 1 1918 9 22 9 54 49.29  -1.698 98.298 15.0 0.0  0.0 ehb
```

Specify the character strings you want to put out with the "...".

```
1 1 1918 9 22 9 54 49.29  -1.698  98.298 15.0 0.0  0.0 ehb FEQ  x
1 1 1918 9 22 9 54 49.29   9.599 -92.802 30.0 0.0  0.0 ehb FEQ  x
1 1 1918 9 22 9 54 49.29   4.003  94.545 20.0 0.0  0.0 ehb FEQ  x
```

```
%awk '{print "latitude:",$9,"longitude:",$10,"depth:",$11}' SUMA.Loc
latitude: -1.698 longitude: 98.298 depth: 15.0
latitude: 9.599 longitude: -92.802 depth: 30.0
latitude: 4.003 longitude: 94.545 depth: 20.0
```

Does it for each line.

Notice that the output does not come out in nice columnar output (similar to the input).

If you wanted to put each piece of information on a different line, you can specify a newline several ways

```
%awk '{print "latitude:",$9; print "longitude:",$10}' SUMA.Loc
%awk '{print "latitude:",$9}{print "longitude:",$10}' SUMA.Loc
%awk '{print "latitude:",$9"\n"longitude:",$10}'  SUMA. Loc
latitude: -1.698
longitude: 98.298
```

Stop printing with “;” or } (the } marks the end of statement/block) and then do another print statement (you need to start a new block with another { if you closed the previous block),

or put out a new line character (\n) (it is a character so it has to be in double quotes "...").

awk variables

You can create awk variables inside your awk blocks in a manner similar to sh/bash. These variables can be character strings or numeric - integer, or floating point.

awk treats a number as a character string or various types of number based on context.

awk variables
In Shell we can do this

```
796 $ a=text
797 $ b='test $TEXT'
798 $ c="check $0"
799 $ d=10.7
800 $ echo $a $b, $c, $d
text test $TEXT, check –bash, 10.7
```

No comma between $a and $b, so no comma in output (spaces count and are output as spaces, comma produces comma in output)

In awk we would do the same thing like this (spaces don't count here, comma produces spaces in output)

```
809 $ echo text | nawk '{b="test $TEXT";a=b;c="'$0'";d=10.7;print $1, a, b, c, d}'
text test $TEXT test $TEXT -bash 10.7
810 $ echo text | nawk '{b="test $TEXT";a=b;c="'$0'";d=10.7;print $1, a b, c, d}'
text test $TEXTtest $TEXT -bash 10.7
```

Aside - Several ways to enter awk command(s) (some more readable than others

separate commands on the command line by ";"

```
809 $ echo text | nawk '{b="test $TEXT";a=b;c="'$0'";d=10.7;print $1, a, b, c, d}'
text test $TEXT test $TEXT -bash 10.7
```

Or you can put each command on its own line (newlines replace the ";"s)

```
506 $ echo text | nawk '{
b="test $TEXT"
a=b
c="'$0'"
d=10.7
print $1, a, b, c, d
}'
text test $TEXT test $TEXT -bash 10.7
507 $
```

Aside - Several ways to enter awk command(s) (some more readable than others)

Or you can make an executable shell file – tst.awk – the file has what you would type on the terminal (plus a #!/bin/sh at the beginning).

```
$ vi tst.awk
i#!/bin/sh
nawk '{
b="test $TEXT"
a=b
c="'$0'"
d=10.7
print $1, a, b, c, d
}'esc
:wq
$ x tst.awk
$ echo text | tst.awk
text test $TEXT test $TEXT ./tst.awk 10.7
```

Aside - Several ways to enter awk command(s)

Make a file, `tst.awk.in`, containing an awk "program" (the commands you would type in). Notice that here we do not need the single quotes (stuff not going through shell) or the { } around the block of commands (outside most set of braces optional as file is used to define the block). Use –f to id file with commands.

```
$ vi tst.awk.in
i#!/bin/sh
b="test $TEXT"
a=b
c="'$0'"
d=10.7
print $1, a, b, c, d
esc
:wq
$ echo text | tst.awk –f tst.awk.in
text test $TEXT test $TEXT ./tst.awk 10.7
```

```
$ cat tst.awk.in
b="test $TEXT"
a=b
c="'$0'"
d=10.7
print $1, a, b, c, d
```

Back to awk variables

```
#!/bin/sh
column=$1
nawk '{print $'$column'}'

822 $ ls -l *tst
-rwx------@ 1 robertsmalley  staff  2250 Aug 16  2004 az_map_tst
-rwx------@ 1 robertsmalley  staff   348 Aug 16  2004 tst
823 $ ls -l *tst | Column2.sh 9
az_map_tst
tst
824 $
```

Here **column** is a shell variable that is set to the first command line argument, **$1**.
'$column' is then expanded to **9** , the value of the first command line argument above, creating the **awk** variable **$9**

```
nawk '{print $9}'
```

And another tangent - this example also demonstrates a very powerful (***and very dangerous***) idea and capability.

The field to be printed is determined in real-time while the program is being executed.

It is not "hard coded".

So the program is effectively writing itself.
This is a very, very simple form of self-modifying-code.
(self-modifying-code is very hard to debug because you don't know what is actually being executed! You better hope it is the guilty party.)

You will find that it is very convenient to write scripts to write scripts!

You can write shell scripts (or C or Fortran for that matter) to write new shell scripts.

You can write shell scripts (or C or Fortran for that matter) to write SAC macros, etc. (Vocabulary/jargon - SAC macros are like shell scripts, but are in the SAC command "language".)

Built-In Variables

FS: Field Separator

The default field separator is the space, what if we want to use some other character.

The password file looks like this

```
root:x:0:1:Super-User:/:/sbin/sh
```

The field separator seems to be (is) "**:**"
We can reset the field separator using the –F command line switch (the lower case switch, -f, is for specifying scriptfiles as we saw before).

```
% awk –F":"   '{print $1, $3}'   /etc/passwd
root 0
```

Built-In Variables

FS: Field Separator

There is another way to reset the FS variable that is more general (in that it does not depend on having a command line switch to do it – so it works with other built-in variables).

```
root:x:0:1:Super-User:/:/sbin/sh

% awk 'BEGIN {FS=":"} {print $1, $3}'  /etc/passwd
root 0
```

More awk program syntax

`BEGIN {…}` : the begin block contains everything you want done before awk procedures are implemented (before it starts processing the file)

`{…}[{…}…]` (list of procedures to be carried out on all lines)

`END {…}` : the end block contains everything you want done after the whole file has been processed.

BEGIN and **END** specify actions to be taken before any lines are read, and after the last line is read.

The **awk** program:

```
BEGIN { print "START" }
    { print }
END { print "STOP" }
```

adds one line with "**START**" before printing out the file and one line "**STOP**" at the end.

Field Separator

Can use multiple characters for Field Separators simultaneously

```
FS = "[:, -]+"
```

Built-In Variables

NR: record number is another useful built-in awk variable
it takes on the current line number, starting from 1

```
root:x:0:1:Super-User:/:/sbin/sh

% awk –F":"  '{print NR, $1, $3}'  /etc/passwd
1 root 0
```

RS : record separator specifies when the current record ends and the next begins
default is "\n" (newline)
useful option is "" (blank line)

OFS : output field separator
default is " " (whitespace)

ORS : output record separator
default is a "\n" (newline)

NF : number of fields in the current record

think of this as awk looking ahead to the next RS
to count the number of fields in advance

```
$ echo 1 1 1918 9 22 9 54 49.29  -1.698 98.298 15.0 0.0  0.0 ehb
FEQ  x | nawk '{print NF}'
16
```

FILENAME : stores the current filename

OFMT : output format for numbers
example OFMT="%.6f" would make all numbers
output as floating points

Accessing shell variables in `awk`

3 methods to access shell variables inside a `awk` script ...

Method 1 - Say I have a script with command arguments and I want to include them in the output of `awk` processing of some input data file:

Now we have a little problem

The Shell takes `$0, $1, etc.` as (the value of) variables for the command that is running, its first argument, etc.

While `awk` takes `$0, $1, etc.` as (the value of) variables that refer to the whole line, first field, second field, etc. of the input file.

So what does `'{print $1}'` refer to?

So what does `'{print $1}'` refer to?

It refers to the `awk` definition (the single quotes protect it from the shell) – the first field on the line coming from the input file.

The problem is getting stuff into `awk` other than what is coming from the input stream (a file, a pipe, stndin).

In addition to the shell command line parameter/field position variable name problem, say I have some variable in my shell script that I want to include in the awk processing (using the shell variables `$0, $1`, is really the same as this problem with the addition of the variable name confusion).

What happens if I try

```
$ a=1
$ awk '{print $1, $2, $a}'
```

awk will be very disappointed in you!

Unlike the shell awk does not evaluate variables within strings.

If I try putting the shell variables into quotes to make them part of a character string to be output

```
'{print "$0\t$a" }'
```

awk would print out

```
$0      $a
```

Inside quotes in awk, the $ is not a metacharacter (unlike inside double quotes in the shell where variables are expanded). Outside quotes in awk, the $ corresponds to a field (so far), not evaluate and return the value of a variable (you could think of the field as a variable, but you are limited to variables with integer names).

Aside - The \t is a tab

```
{print "$0\t$a" }
```

Another difference between awk and a shell processing the characters within double quotes.

AWK understands special characters follow the "\" character like "t".

The Bourne and C UNIX shells do not.

To make a long story short, what we have to do is stop protecting the $ from the shell by judicious use of the single quotes.

For number valued variables, just use single quotes, for text valued variables you need to tell **awk** it is a character string with double quotes.

```
$ a=A;b=1;c=C
$ echo $a $c | nawk '{print $1 '$b' $2, "'$0'"}'
A1C –bash
```

If you don't use double quotes for character variables, it may not work

```
$ echo $a $b | nawk '{print $1 '$b' $2, '$0'}'
A1C -0
```

The single quotes really group like this.

```
$ echo $a $c | awk '{print $1 '$b' $2, "'$0'"}'
A1C –bash
```

The first single quote turns off shell interpretation of everything after it until the next single quote. So the single quote before the `$b` turns off the quote before the `{`. The `$b` gets passed to the shell for interpretation. It is a number so `awk` can handle it without further ado. The single quote after the `$b` turns off shell interpretation again, until the single quote before the `$0`.

The single quotes really group like this.

```
$ echo $a $c | awk '{print $1 '$b' $2, "'$0'"}'
A1C –bash
```

The $0 returns the name of the program you are running, in this case the shell –bash. This is a character string so it needs to be in double quotes, thus the "'$0'". The single quote after the $0 turns "quoting" back on and it continues to the end of the awk block of code , dignified by the }.

The quotes are switches that turn shell interpretation off (first one) and back on (second one).

The single quotes really group like this.

```
$ echo $a $c | awk '{print $1 '$b' $2, "'$0'"}'
A1C –bash
```

Practically, since you always have the first and last, you can think about the ones about the `'$b'` and `'$0'` as pairs – but they really match up operationally as discussed.

Same for variables you define - if it is a text string you have to say it is a text string with the double quotes.

```
$ b=B
$ echo $a $b | nawk '{print $1 "'$b'" $2}'
ABC
```

If the variable was a number, you can still print it out as a text string (awk treats numbers as text strings or numbers as necessary in context, while text strings are stuck as text strings.)

```
$ b = 1
$ echo $a $b | nawk '{print $1 "'$b'" '$b' $2}'
A11C
```

Aside

How to print out quotes
(this is a very long line, no \ for continuation – wraps on its own).

```
620 $ echo $a $c | nawk '{print $1, '$b', $2, "'$0'", "\"", "\"ab
\"", "*", "'"'''"'", "a'"'"b", "/" }'
A 1 C -bash " "ab" * ''' a'b /
```

Aside
How to print out quotes

```
581:> nawk 'BEGIN { print "Dont Panic!" }'
Dont Panic!
582:> nawk 'BEGIN { print "Don'\''t Panic!" }'
Don't Panic!
583:> nawk 'BEGIN { print "Don'"'"'t Panic!" }'
Don't Panic!
586:> echo Don"'"t Panic! | nawk "{print}"
Don't Panic!
584:> echo Don\'t Panic! | nawk '{print}'
Don't Panic!
585:> echo Don\'t Panic! | nawk "{print}"
Don't Panic!
```

Look carefully at the 2 lines above – you can (sometimes) use either quote (‘ or “) to protect the nawk program (depends on what you are trying to also protect from the shell).

Aside
How to print out quotes

```
alpaca.587:> nawk 'BEGIN { print "\"Don’\’’t Panic!\"" }'
"Don’t Panic!"
```

Method 2. Assign the shell variables to awk variables after the body of the script, but before you specify the input file

```
VAR1=3
VAR2="Hi"

awk '{print v1, v2}' v1=$VAR1 v2=$VAR2 input_file

3 Hi
```

Also note: awk variables are referred to using just their name (no $ in front)

There are a couple of constraints with this method

Shell variables assigned using this method are not available in the **BEGIN** section (will see this, and **END** section, soon).

If variables are assigned after a filename, they will not be available when processing that filename

```
awk '{print v1, v2}' v1=$VAR1 file1 v2=$VAR2 file2
```

In this case, **v2** is not available to **awk** when processing file1.

Method 3. Use the -v switch to assign the shell variables to awk variables.

This works with awk, but not all flavors.

```
awk -v v1=$VAR1 -v v2=$VAR2 '{print v1, v2}' input_file
```

Aside - why use variables?

Say I'm doing some calculation that uses the number π.

I can put `3.1416` in whenever I need to use it.

But say later I decide that I need more precision and want to change the value of π to `3.1415926`.

It is a pain to have to change this and as we have seen global edits sometimes have unexpected (mostly because we were not really paying attention) side effects.

Aside - Why use variables?

Using variables (the first step to avoid hard-coding) – if you use variables you don't have to modify the code in a thousand places where you used `3.1416` for π.

If you had set a variable

```
pi=3.1416
```

And use `$pi`, it becomes trivial to change its value everywhere in the script by just editing the single line

```
pi=3.1415926
```

you don't have to look for it & change it everywhere

Examples:

Say we want to print out the owner of every file

Record/Field/column separator (`RS=" "`)

The output of `ls -l` is

```
-rwxrwxrwx  1 rsmalley user    7237 Jun 12  2006 setup_exp1.sh
```

So we need fields 3 and 9.

Do using an executable shell script

Create the file **owner.nawk** and make it executable.

```
$ vi owner.nawk
i#!/bin/awk -f
BEGIN { print "File\tOwner" }
{ print $9, "\t", $3}
END { print " - DONE -" }esc
:wq
$ x owner.nawk
```

Now we have to get the input into the program.

Pipe the long directory listing into our shell script.

```
507:> ls -l | owner.nawk
File    Owner
*CHARGE-2002-107*        rsmalley
022285A.cmt      rsmalley
190-00384-07.pdf         rsmalley
. . .
zreal2.f         rsmalley
zreal2.o         rsmalley
 - DONE -
508:>
```

So far we have just been selecting and rearranging fields. The output is a simple copy of the input field.

What if you wanted to change the format of the output with respect to the input

Considering that **awk** was written by some of UINX's developers, it might seem reasonable to guess that they "reused" some useful UNIX tools.

If you guessed that you would be correct.

So if you wanted to change the format of the output with respect to the input – you just use the UNIX `printf` command.

We already saw this command, so we don't need to discuss it any further (another UNIX philosophy based attitude).

```
$ echo text | awk '{
b="test $TEXT"
a=b
c="'$0'"
d=10.7
printf("%s, %s, %s, %s, %6.3f\n", $1, a, b, c, d)
}'
text, test $TEXT, test $TEXT, -bash, 10.700
```

Notice a few differences with the UNIX `printf` command

You need parens (...) around the arguments to the `printf` command.

You need commas between the items in the variable list.

```
$ echo text | awk '{
b="test $TEXT"
a=b
c="'$0'"
d=10.7
printf("%s, %s, %s, %s, %6.3f\n", $1, a, b, c, d)
}'
text, test $TEXT, test $TEXT, -bash, 10.700
```

The output of `printf` goes to stndout.

`sprintf`

Same as `printf` but sends formatted print output to a string variable rather to stndout

```
n=sprintf ("%d plus %d is %d", a, b, a+b);
```

"Simple" awk example:

Say I have some sac files with the horrid IRIS DMC format file names

```
1999.289.10.05.26.0000.IU.KMBO.00.LHZ.SAC
```

and it would rename it to something more "user friendly" like `KMBO.LHZ` to save on typing while doing one of Chuck's homeworks.

```
alpaca.540:> more rename.sh
#!/bin/sh

#to rename horrid iris dmc file names

#call with rename.sh A x y
#where A is the char string to match, x and y are the field
#numbers in the original file name you want to use in the
#final name, and using the period/dot for the field seperator

#eg if the file names look like
#1999.289.10.05.26.0000.IU.KMBO.00.LHZ.SAC
#and you would ;ike to rename it KMBO.LHZ
#the 8th field is the station name, KMBO
#and the 10th field is the component name, LHZ
#so you would call rename.sh SAC 8 10
#(it will do it for all file names in your directory
#containing the string "SAC")

for file in `ls -1 *$1*`
do
mv $file `echo $file | nawk -F. '{print $'$2'"."$'$3'}'`
done
alpaca.541:>
```

Loop is in Shell, not awk.