Awk — A Pattern Scanning and Processing Language
(Second Edition)

Alfred V. Aho

Brian W. Kernighan

Peter J. Weinberger

Bell Laboratories
Murray Hill, New Jersey 07974

ABSTRACT

Awk is a programming language whose basic operation is to search a set of files for patterns, and to perform specified actions upon lines or fields of lines which contain instances of those patterns. *Awk* makes certain data selection and transformation operations easy to express; for example, the *awk* program

length > 72

prints all input lines whose length exceeds 72 characters; the program

NF % 2 == 0

prints all lines with an even number of fields; and the program

{ $1 = log($1); print }

replaces the first field of each line by its logarithm.

Awk patterns may include arbitrary boolean combinations of regular expressions and of relational operators on strings, numbers, fields, variables, and array elements. Actions may include the same pattern-matching constructions as in patterns, as well as arithmetic and string expressions and assignments, **if-else**, **while**, **for** statements, and multiple output streams.

This report contains a user's guide, a discussion of the design and implementation of *awk*, and some timing statistics.

September 1, 1978

Awk — A Pattern Scanning and Processing Language
(Second Edition)

Alfred V. Aho

Brian W. Kernighan

Peter J. Weinberger

Bell Laboratories
Murray Hill, New Jersey 07974

1. Introduction

Awk is a programming language designed to make many common information retrieval and text manipulation tasks easy to state and to perform.

The basic operation of *awk* is to scan a set of input lines in order, searching for lines which match any of a set of patterns which the user has specified. For each pattern, an action can be specified; this action will be performed on each line that matches the pattern.

program *grep* unix program manual will recognize the approach, although in *awk* the patterns may be more general than in *grep*, and the actions allowed are more involved than merely printing the matching line. For example, the *awk* program

```
{print $3, $2}
```

prints the third and second columns of a table in that order. The program

```
$2 ~ /A|B|C/
```

prints all input lines with an A, B, or C in the second field. The program

```
$1 != prev { print; prev = $1 }
```

prints all lines in which the first field is different from the previous first field.

1.1. Usage

The command

```
awk program [files]
```

executes the *awk* commands in the string **program** on the set of named files, or on the standard input if there are no files. The statements can also be placed in a file **pfile**, and executed by the command

```
awk −f pfile [files]
```

1.2. Program Structure

An *awk* program is a sequence of statements of the form:

pattern { *action* }
pattern { *action* }
...

Each line of input is matched against each of the patterns in turn. For each pattern that matches, the associated action is executed. When all the patterns have been tested, the next line is fetched and the matching starts over.

Readers familiar with the UNIX†

†UNIX is a Trademark of Bell Laboratories.

Either the pattern or the action may be left out, but not both. If there is no action
to the output. (Thus a line which matches several patterns can be printed several
the action is performed for every input line. A line which matches no pattern is ig
Since patterns and actions are both optional, actions must be enclosed in braces t

1.3. Records and Fields

Awk input is divided into ‘‘records’’ terminated by a record separator. The def
awk processes its input a line at a time. The number of the current record is avail

Each input record is considered to be divided into ‘‘fields.’’ Fields are normally
but the input field separator may be changed, as described below. Fields are ref
first field, and **$0** is the whole input record itself. Fields may be assigned to. Th
able in a variable named **NF**.

The variables **FS** and **RS** refer to the input field and record separators; they may
The optional command-line argument **−F***c* may also be used to set **FS** to the chara

If the record separator is empty, an empty input line is taken as the record separa
field separators.

The variable **FILENAME** contains the name of the current input file.

1.4. Printing

An action may have no pattern, in which case the action is executed for all lines
record; this is accomplished by the *awk* command **print**. The *awk* program

```
{ print }
```

prints each record, thus copying the input to the output intact. More useful is to
stance,

```
print $2, $1
```

prints the first two fields in reverse order. Items separated by a comma in the prin
put field separator when output. Items not separated by commas will be concaten

```
print $1 $2
```

runs the first and second fields together.

The predefined variables **NF** and **NR** can be used; for example

```
{ print NR, NF, $0 }
```

prints each record preceded by the record number and the number of fields.

Output may be diverted to multiple files; the program

```
{ print $1 >"foo1"; print $2 >"foo2" }
```

writes the first field, **$1**, on the file **foo1**, and the second field on file **foo2**. The >>

```
print $1 >>"foo"
```

appends the output to the file **foo**. (In each case, the output files are created if necessary.) The file name can be a variable or a field as well as a constant; for example,

print $1 >$2

uses the contents of field 2 as a file name.

Naturally there is a limit on the number of output files; currently it is 10.

Similarly, output can be piped into another process (on UNIX only); for instance,

print | "mail bwk"

mails the output to **bwk**.

The variables **OFS** and **ORS** may be used to change the current output field separator and output record separator. The output record separator is appended to the output of the **print** statement.

Awk also provides the **printf** statement for output formatting:

printf format expr, expr, ...

formats the expressions in the list according to the specification in **format** and prints them. For example,

printf "%8.2f %10ld\n", $1, $2

prints **$1** as a floating point number 8 digits wide, with two after the decimal point, and **$2** as a 10-digit long decimal number, followed by a newline. No output separators are produced automatically; you must add them yourself, as in this example. The version of **printf** is identical to that used with C. C programm language prentice hall 1978

2. Patterns

A pattern in front of an action acts as a selector that determines whether the action is to be executed. A variety of expressions may be used as patterns: regular expressions, arithmetic relational expressions, string-valued expressions, and arbitrary boolean combinations of these.

2.1. BEGIN and END

The special pattern **BEGIN** matches the beginning of the input, before the first record is read. The pattern **END** matches the end of the input, after the last record has been processed. **BEGIN** and **END** thus provide a way to gain control before and after processing, for initialization and wrapup.

As an example, the field separator can be set to a colon by

BEGIN { FS = ":" }
... rest of program ...

Or the input lines may be counted by

END { print NR }

If **BEGIN** is present, it must be the first pattern; **END** must be the last if used.

2.2. Regular Expressions

The simplest regular expression is a literal string of characters enclosed in slashes, like

/smith/

This is actually a complete *awk* program which will print all lines which contain any occurrence of the name ‘‘smith’’. If a line contains ‘‘smith’’ as part of a larger word, it will also be printed, as in

blacksmithing

Awk regular expressions include the regular expression forms found in the UNIX text editor *ed* and *grep* (without back-referencing). In addition, *awk* allows parentheses for grouping, | for alternatives, **+** for ‘‘one or more’’, and **?** for ‘‘zero or one’’, all as in *lex*. Character classes may be abbreviated: **[a−zA−Z0−9]** is the set of all letters and digits. As an example, the *awk* program

/[Aa]ho|[Ww]einberger|[Kk]ernighan/

will print all lines which contain any of the names ‘‘Aho,’’ ‘‘Weinberger’’ or ‘‘Kernighan,’’ whether capitalized or not.

Regular expressions (with the extensions listed above) must be enclosed in slashes, just as in *ed* and *sed*. Within a regular expression, blanks and the regular expression metacharacters are significant. To turn off the magic meaning of one of the regular expression characters, precede it with a backslash. An example is the pattern

/\/.∗\//

which matches any string of characters enclosed in slashes.

One can also specify that any field or variable matches a regular expression (or

The program

$1 ~ /[jJ]ohn/

prints all lines where the first field matches ‘‘john’’ or ‘‘John.’’ Notice that this w

... To restrict it to exactly **[jJ]ohn**, use

$1 ~ /^[jJ]ohn$/

The caret ^ refers to the beginning of a line or field; the dollar sign **$** refers to the

2.3. Relational Expressions

An *awk* pattern can be a relational expression involving the usual relational oper

$2 > $1 + 100

which selects lines where the second field is at least 100 greater than the first field

NF % 2 == 0

prints lines with an even number of fields.

In relational tests, if neither operand is numeric, a string comparison is made; oth

$1 >= "s"

selects lines that begin with an **s**, **t**, **u**, etc. In the absence of any other informatio

$1 > $2

will perform a string comparison.

2.4. Combinations of Patterns

A pattern can be any boolean combination of patterns, using the operators || (or)

$1 >= "s" && $1 < "t" && $1 != "smith"

selects lines where the first field begins with ‘‘s’’, but is not ‘‘smith’’. **&&** and |

from left to right; evaluation stops as soon as the truth or falsehood is determined

2.5. Pattern Ranges

The ‘‘pattern’’ that selects an action may also consist of two patterns separated b

pat1, pat2 { ... }

In this case, the action is performed for each line between an occurrence of **pat1**

example,

/start/, /stop/

prints all lines between **start** and **stop**, while

NR == 100, NR == 200 { ... }

does the action for lines 100 through 200 of the input.

3. Actions

An *awk* action is a sequence of action statements terminated by newlines or sem

do a variety of bookkeeping and string manipulating tasks.

3.1. Built-in Functions

Awk provides a ‘‘length’’ function to compute the length of a string of character

its length:

{print length, $0}

length by itself is a ‘‘pseudo-variable’’ which yields the length of the current

yields the length of its argument, as in the equivalent

{print length($0), $0}

The argument may be any expression.

Awk also provides the arithmetic functions **sqrt**, **log**, **exp**, and **int**, for square root, base *e* logarithm, exponential, and integer part of their respective arguments.

The name of one of these built-in functions, without argument or parentheses, stands for the value of the function on the whole record. The program

```
length < 10 || length > 20
```

prints lines whose length is less than 10 or greater than 20.

The function **substr(s, m, n)** produces the substring of **s** that begins at position **m** (origin 1) and is at most **n** characters long. If **n** is omitted, the substring goes to the end of **s**. The function **index(s1, s2)** returns the position where the string **s2** occurs in **s1**, or zero if it does not.

The function **sprintf(f, e1, e2, ...)** produces the value of the expressions **e1**, **e2**, etc., in the **printf** format specified by **f**. Thus, for example,

```
x = sprintf("%8.2f %10ld", $1, $2)
```

sets **x** to the string produced by formatting the values of **$1** and **$2**.

3.2. Variables, Expressions, and Assignments

Awk variables take on numeric (floating point) or string values according to context. For example, in

```
x = 1
```

x is clearly a number, while in

```
x = "smith"
```

it is clearly a string. Strings are converted to numbers and vice versa whenever context demands it. For instance,

```
x = "3" + "4"
```

assigns 7 to **x**. Strings which cannot be interpreted as numbers in a numerical context will generally have numeric value zero, but it is unwise to count on this behavior.

By default, variables (other than built-ins) are initialized to the null string, which has numerical value zero; this eliminates the need for most **BEGIN** sections. For example, the sums of the first two fields can be computed by

```
    { s1 += $1; s2 += $2 }
END { print s1, s2 }
```

Arithmetic is done internally in floating point. The arithmetic operators are +, −, ∗, /, and **%** (mod). The C increment **++** and decrement **−−** operators are also available, and so are the assignment operators **+=**, **−=**, **∗=**, **/=**, and **%=**. These operators may all be used in expressions.

3.3. Field Variables

Fields in *awk* share essentially all of the properties of variables — they may be used in arithmetic or string operations, and may be assigned to. Thus one can replace the first field with a sequence number like this:

```
{ $1 = NR; print }
```

or accumulate two fields into a third, like this:

```
{ $1 = $2 + $3; print $0 }
```

or assign a string to a field:

```
{ if ($3 > 1000)
      $3 = "too big"
  print
}
```

which replaces the third field by ‘‘too big’’ when it is, and in any case prints the record.

Field references may be numerical expressions, as in

```
{ print $i, $(i+1), $(i+n) }
```

Whether a field is deemed numeric or string depends on context; in ambiguous cases like

```
if ($1 == $2) ...
```

fields are treated as strings.

Each input line is split into fields automatically as necessary. It is also possible t

```
n = split(s, array, sep)
```

splits the the string **s** into **array[1]**, ..., **array[n]**. The number of elements found used as the field separator; otherwise **FS** is used as the separator.

3.4. String Concatenation

Strings may be concatenated. For example

```
length($1 $2 $3)
```

returns the length of the first three fields. Or in a **print** statement,

```
print $1 " is " $2
```

prints the two fields separated by ‘‘ is ’’. Variables and numeric expressions may

3.5. Arrays

Array elements are not declared; they spring into existence by being mentione example of a conventional numeric subscript

```
x[NR] = $0
```

assigns the current input record to the **NR**-th element of the array **x**. In fact, it i process the entire input in a random order with the *awk* program

```
    { x[NR] = $0 }
END { ... program ... }
```

The first action merely records each input line in the array **x**.

Array elements may be named by non-numeric values, which gives *awk* a c Snobol tables. Suppose the input contains fields with values like **apple**, **orange**, e

```
/apple/     { x["apple"]++ }
/orange/    { x["orange"]++ }
END         { print x["apple"], x["orange"] }
```

increments counts for the named array elements, and prints them at the end of the

3.6. Flow-of-Control Statements

Awk provides the basic flow-of-control statements **if-else**, **while**, **for**, and stater the **if** statement in section 3.3 without describing it. The condition in parenthese ing the **if** is done. The **else** part is optional.

The **while** statement is exactly like that of C. For example, to print all input fielc

```
i = 1
while (i <= NF) {
      print $i
      ++i
}
```

The **for** statement is also exactly that of C:

```
for (i = 1; i <= NF; i++)
      print $i
```

does the same job as the **while** statement above.

There is an alternate form of the **for** statement which is suited for accessing the e

```
for (i in array)
      statement
```

does *statement* with **i** set in turn to each element of **array**. The elements are acce ensue if **i** is altered, or if any new elements are accessed during the loop.

The expression in the condition part of an **if**, **while** or **for** can include relational and **!=** (‘‘not equal to’’); regular expression matches with the match operators ~

-- --

of course parentheses for grouping.

The **break** statement causes an immediate exit from an enclosing **while** or **for**; the **continue** statement causes the next iteration to begin.

The statement **next** causes *awk* to skip immediately to the next record and begin scanning the patterns from the top. The statement **exit** causes the program to behave as if the end of the input had occurred.

Comments may be placed in *awk* programs: they begin with the character **#** and end with the end of the line, as in

print x, y # this is a comment

As might be expected, *awk* is not as fast as the specialized tools *wc* , *sed* , or th the general tool *lex*. In all cases, the tasks were about as easy to express a guages; tasks involving fields were considerably easier to express as *awk* prog *awk*, *wc*, *sed* and *lex*. $LISTS$

4. Design

The UNIX system already provides several programs that operate by passing input through a selection mechanism. *Grep* , the first and simplest, merely prints all lines which match a single specified pattern. *Egrep* provides more general patterns, i.e., regular expressions in full generality; *fgrep* searches for a set of keywords with a particularly fast algorithm. *Sed* unix programm manual provides most of the editing facilities of the editor *ed* , applied to a stream of input. None of these programs provides numeric capabilities, logical relations, or variables.

Lex lesk lexical analyzer cstr provides general regular expression recognition capabilities, and, by serving as a C program generator, is essentially open-ended in its capabilities. The use of *lex* , however, requires a knowledge of C programming, and a *lex* program must be compiled and loaded before use, which discourages its use for one-shot applications.

Awk is an attempt to fill in another part of the matrix of possibilities. It provides general regular expression capabilities and an implicit input/output loop. But it also provides convenient numeric processing, variables, more general selection, and control flow in the actions. It does not require compilation or a knowledge of C. Finally, *awk* provides a convenient way to access fields within lines; it is unique in this respect.

Awk also tries to integrate strings and numbers completely, by treating all quantities as both string and numeric, deciding which representation is appropriate as late as possible. In most cases the user can simply ignore the differences.

Most of the effort in developing *awk* went into deciding what *awk* should or should not do (for instance, it doesn't do string substitution) and what the syntax should be (no explicit operator for concatenation) rather than on writing or debugging the code. We have tried to make the syntax powerful but easy to use and well adapted to scanning files. For example, the absence of declarations and implicit initializations, while probably a bad idea for a general-purpose programming language, is desirable in a language that is meant to be used for tiny programs that may even be composed on the command line.

In practice, *awk* usage seems to fall into two broad categories. One is what might be called ‘‘report generation’’ — processing an input to extract counts, sums, sub-totals, etc. This also includes the writing of trivial data validation programs, such as verifying that a field contains only numeric information or that certain delimiters are properly balanced. The combination of textual and numeric processing is invaluable here.

A second area of use is as a data transformer, converting data from the form produced by one program into that expected by another. The simplest examples merely select fields, perhaps with rearrangements.

5. Implementation

The actual implementation of *awk* uses the language development tools available on the UNIX operating system. The grammar is specified with *yacc* ; yacc johnson cstr the lexical analysis is done by *lex* ; the regular expression recognizers are deterministic finite automata constructed directly from the expressions. An *awk* program is translated into a parse tree which is then directly executed by a simple interpreter.

Awk was designed for ease of use rather than processing speed; the delayed evaluation of variable types and the necessity to break input into fields makes high speed difficult to achieve in any case. Nonetheless, the program has not proven to be unworkably slow.

Table I below shows the execution (user + system) time on a PDP-11/70 of the UNIX programs *wc* , *grep* , *egrep* , *fgrep* , *sed* , *lex* , and *awk* on the following simple tasks:

1. count the number of lines.
2. print all lines containing ‘‘doug’’.
3. print all lines containing ‘‘doug’’, ‘‘ken’’ or ‘‘dmr’’.
4. print the third field of each line.
5. print the third and second fields of each line, in that order.
6. append all lines containing ‘‘doug’’, ‘‘ken’’, and ‘‘dmr’’ to files ‘‘jdoug’’, ‘‘jken’’, and ‘‘jdmr’’, respectively.
7. print each line prefixed by ‘‘line-number : ’’.
8. sum the fourth column of a table.

The program *wc* merely counts words, lines and characters in its input; we have already mentioned the others. In all cases the input was a file containing 10,000 lines as created by the command *ls −l* ; each line has the form

−rw−rw−rw− 1 ava 123 Oct 15 17:05 xxx

The total length of this input is 452,960 characters. Times for *lex* do not include compile or load.

Program	Task 1	2	3	4	5	6	7	8
wc	8.6							
grep	11.7	13.1						
egrep	6.2	11.5	11.6					
fgrep	7.7	13.8	16.1					
sed	10.2	11.6	15.8	29.0	30.5	16.1		
lex	65.1	150.1	144.2	67.7	70.3	104.0	81.7	92.8
awk	15.0	25.6	29.9	33.3	38.9	46.4	71.4	31.1

Table I. Execution Times of Programs. (Times are in sec.)

The programs for some of these jobs are shown below. The *lex* programs are generally too long to show.

AWK:

1.
```
END {print NR}
```

2.
```
/doug/
```

3.
```
/ken|doug|dmr/
```

4.
```
{print $3}
```

5.
```
{print $3, $2}
```

6.
```
/ken/ {print >"jken"}
/doug/      {print >"jdoug"}
/dmr/{print >"jdmr"}
```

7.
```
{print NR ": " $0}
```

8.
```
        {sum = sum + $4}
END {print sum}
```

SED:

1.
```
$=
```

2.
```
/doug/p
```

3.
```
/doug/p
/doug/d
/ken/p
/ken/d
/dmr/p
/dmr/d
```

4.
```
/[^ ]* [ ]*[^ ]* [ ]*\([^ ]*\) .*/s//\1/p
```

5.
```
/[^ ]* [ ]*\([^ ]*\) [ ]*\([^ ]*\) .*/s//\2 \1/p
```

6.
```
/ken/w jken
/doug/w jdoug
/dmr/w jdmr
```

LEX:

1.
```
%{
int i;
%}
%%
\n      i++;
.       ;
%%
yywrap() {
        printf("%d\n", i);
}
```

2.
```
%%
^.*doug.*$ printf("%s\n", yytext);
.       ;
\n      ;
```