Data Processing with AWK

A command-line tool to slice & dice textual data

Krishnakumar Gopalakrishnan

krishnak [at] vt [dot] edu

Computational Methods Hub
Imperial College London

April 16 2019

Introduction

What is Awk?

Why the *awkward* name?

- Aho, Al

What is Awk?

Why the *awkward* name?

- Aho, Al
- Weinberger, Peter

What is Awk?

Why the *awkward* name?

- Aho, Al
- Weinberger, Peter
- Kernighan, Brian

What is Awk?

Why the *awkward* name?

- Aho, Al
- Weinberger, Peter
- Kernighan, Brian

What is Awk?

Why the *awkward* name?

- Aho, Al
- Weinberger, Peter
- Kernighan, Brian

What can we do with Awk?

- Text manipulation in powerful ways

What is Awk?

Why the *awkward* name?

- Aho, Al
- Weinberger, Peter
- Kernighan, Brian

What can we do with Awk?

- Text manipulation in powerful ways
- Data filtering, cleaning & other pre-processing tasks

What is Awk?

Why the *awkward* name?

- Aho, Al
- Weinberger, Peter
- Kernighan, Brian

What can we do with Awk?

- Text manipulation in powerful ways
- Data filtering, cleaning & other pre-processing tasks
- Read & write data in a variety of formats

What is Awk?

Why the *awkward* name?

- **A**ho, Al
- **W**einberger, Peter
- **K**ernighan, Brian

What can we do with Awk?

- Text manipulation in powerful ways
- Data filtering, cleaning & other pre-processing tasks
- Read & write data in a variety of formats
- Produce tabular reports for the web (advanced)

What is Awk?

Why the *awkward* name?

- **A**ho, Al
- **W**einberger, Peter
- **K**ernighan, Brian

What can we do with Awk?

- Text manipulation in powerful ways
- Data filtering, cleaning & other pre-processing tasks
- Read & write data in a variety of formats
- Produce tabular reports for the web (advanced)
- Task automation (advanced)

Where is Awk available?

Summary of Awk Variants

- Lots of variants (awk, bwk, nawk, mawk, gawk, goawk)
- Most popular version of awk is GNU Awk (gawk)
- Version 5.0 released Apr 12, 2019!

Where is Awk available?

*nix Machines

- Pre-installed on most Unix-like OSes
  - MacOS
  - Linux

Where is Awk available?

*nix Machines

- Pre-installed on most Unix-like OSes
  - MacOS
  - Linux
  - All BSD variants

Where is Awk available?

*nix Machines

- Pre-installed on most Unix-like OSes
  - MacOS
  - Linux
  - All BSD variants
  - Solaris, Illumos & many others

Where is Awk available?

*nix Machines

- Pre-installed on most Unix-like OSes
  - MacOS
  - Linux
  - All BSD variants
  - Solaris, Illumos & many others

Where is Awk available?

*nix Machines

- Pre-installed on most Unix-like OSes
  - MacOS
  - Linux
  - All BSD variants
  - Solaris, Illumos & many others

Windows Machines

- Plethora of options available
  - Cygwin: complete unix-like environment (heavy)

Where is Awk available?

*nix Machines

- Pre-installed on most Unix-like OSes
  - MacOS
  - Linux
  - All BSD variants
  - Solaris, Illumos & many others

Windows Machines

- Plethora of options available
  - Cygwin: complete unix-like environment (heavy)
  - Windows Subsystem for Linux (WSL): Windows 10 1703 & above (requires admin privileges)

Where is Awk available?

*nix Machines

- Pre-installed on most Unix-like OSes
  - MacOS
  - Linux
  - All BSD variants
  - Solaris, Illumos & many others

Windows Machines

- Plethora of options available
  - Cygwin: complete unix-like environment (heavy)
  - Windows Subsystem for Linux (WSL): Windows 10 1703 & above (requires admin privileges)
  - ezwinports (`https://sourceforge.net/projects/ezwinports/`)

Where is Awk available?

*nix Machines

- Pre-installed on most Unix-like OSes
  - MacOS
  - Linux
  - All BSD variants
  - Solaris, Illumos & many others

Windows Machines

- Plethora of options available
  - Cygwin: complete unix-like environment (heavy)
  - Windows Subsystem for Linux (WSL): Windows 10 1703 & above (requires admin privileges)
  - ezwinports (`https://sourceforge.net/projects/ezwinports/`)
  - Git for Windows (`https://git-scm.com/download/win`)

Where is Awk available?

*nix Machines

- Pre-installed on most Unix-like OSes
  - MacOS
  - Linux
  - All BSD variants
  - Solaris, Illumos & many others

Windows Machines

- Plethora of options available
  - Cygwin: complete unix-like environment (heavy)
  - Windows Subsystem for Linux (WSL): Windows 10 1703 & above (requires admin privileges)
  - ezwinports (`https://sourceforge.net/projects/ezwinports/`)
  - Git for Windows (`https://git-scm.com/download/win`)
  - Cmder (`https://github.com/cmderdev/cmder/releases/download/v1.3.11/cmder.zip`)

Where is Awk available?

*nix Machines

- Pre-installed on most Unix-like OSes
  - MacOS
  - Linux
  - All BSD variants
  - Solaris, Illumos & many others

Windows Machines

- Plethora of options available
  - Cygwin: complete unix-like environment (heavy)
  - Windows Subsystem for Linux (WSL): Windows 10 1703 & above (requires admin privileges)
  - ezwinports (`https://sourceforge.net/projects/ezwinports/`)
  - Git for Windows (`https://git-scm.com/download/win`)
  - Cmder (`https://github.com/cmderdev/cmder/releases/download/v1.3.11/cmder.zip`)
  - Log on to a *nix remote server

Applicability of Awk

- Great for:

Applicability of Awk

- Great for:
  - Manipulating text files divided into lines and columns

Applicability of Awk

- Great for:
  - Manipulating text files divided into lines and columns
  - All lines are not required to be in the same format

Applicability of Awk

- Great for:
  - Manipulating text files divided into lines and columns
  - All lines are not required to be in the same format
  - Performs best on a structured file (eg tabular data)

Applicability of Awk

- Great for:
  - Manipulating text files divided into lines and columns
  - All lines are not required to be in the same format
  - Performs best on a structured file (eg tabular data)
- Small one-line Awk programs can:

Applicability of Awk

- Great for:
  - Manipulating text files divided into lines and columns
  - All lines are not required to be in the same format
  - Performs best on a structured file (eg tabular data)
- Small one-line Awk programs can:
  - Find interesting lines in a data file

Applicability of Awk

- Great for:
  - Manipulating text files divided into lines and columns
  - All lines are not required to be in the same format
  - Performs best on a structured file (eg tabular data)
- Small one-line Awk programs can:
  - Find interesting lines in a data file
  - Output only columns of data matching some criterion

Applicability of Awk

- Great for:
  - Manipulating text files divided into lines and columns
  - All lines are not required to be in the same format
  - Performs best on a structured file (eg tabular data)
- Small one-line Awk programs can:
  - Find interesting lines in a data file
  - Output only columns of data matching some criterion
  - Swapping the order of columns

Applicability of Awk

- Great for:
  - Manipulating text files divided into lines and columns
  - All lines are not required to be in the same format
  - Performs best on a structured file (eg tabular data)
- Small one-line Awk programs can:
  - Find interesting lines in a data file
  - Output only columns of data matching some criterion
  - Swapping the order of columns
  - Combine multiple columns into one

Applicability of Awk

- Great for:
  - Manipulating text files divided into lines and columns
  - All lines are not required to be in the same format
  - Performs best on a structured file (eg tabular data)
- Small one-line Awk programs can:
  - Find interesting lines in a data file
  - Output only columns of data matching some criterion
  - Swapping the order of columns
  - Combine multiple columns into one
  - Split single column into multiple

Applicability of Awk

- Great for:
  - Manipulating text files divided into lines and columns
  - All lines are not required to be in the same format
  - Performs best on a structured file (eg tabular data)
- Small one-line Awk programs can:
  - Find interesting lines in a data file
  - Output only columns of data matching some criterion
  - Swapping the order of columns
  - Combine multiple columns into one
  - Split single column into multiple
  - Sophisticated data operations (joins, merges etc)

Applicability of Awk

- Great for:
  - Manipulating text files divided into lines and columns
  - All lines are not required to be in the same format
  - Performs best on a structured file (eg tabular data)
- Small one-line Awk programs can:
  - Find interesting lines in a data file
  - Output only columns of data matching some criterion
  - Swapping the order of columns
  - Combine multiple columns into one
  - Split single column into multiple
  - Sophisticated data operations (joins, merges etc)
- Not good for:

Applicability of Awk

- Great for:
  - Manipulating text files divided into lines and columns
  - All lines are not required to be in the same format
  - Performs best on a structured file (eg tabular data)
- Small one-line Awk programs can:
  - Find interesting lines in a data file
  - Output only columns of data matching some criterion
  - Swapping the order of columns
  - Combine multiple columns into one
  - Split single column into multiple
  - Sophisticated data operations (joins, merges etc)
- Not good for:
  - Manipulating binary files (Excel/Word)

Applicability of Awk

- Great for:
  - Manipulating text files divided into lines and columns
  - All lines are not required to be in the same format
  - Performs best on a structured file (eg tabular data)
- Small one-line Awk programs can:
  - Find interesting lines in a data file
  - Output only columns of data matching some criterion
  - Swapping the order of columns
  - Combine multiple columns into one
  - Split single column into multiple
  - Sophisticated data operations (joins, merges etc)
- Not good for:
  - Manipulating binary files (Excel/Word)
  - Not a web programming language (eg parsing HTML)

Hands-on Exercises

Exercise 1

- Reverse names in `names.txt`

Questions?