

Filter Data from Differing Text File Formats



Andrew Mallett

Author and Trainer

@theurbanpenguin | www.theurbanpenguin.com



Overview



AWK and File Formats

- The Output Field Separator
- Converting to YAML
- Reading XML data





awk vs tr



**We used tr before to
convert field separators
BUT awk has this built-in**



```
$ awk 'BEGIN {FS=":" ; OFS=","} { print $1,$3,$4,$6,$7 } ' /etc/passwd
```

Convert /etc/passwd to CSV

The default OFS, Output Field Separator is a space. To see this, we must pass each field as their own argument to print. To set to a CSV file we change the OFS variable in the BEGIN block



Demo



Working with the OFS

- Using print arguments
- Converting to CSV



YAML Data

Keys / Value

key: value

Dictionary

Indented pairs

Lists

**Indented and
hyphens**



Sample YAML

```
users:  
  - user: bob  
    passwd: x  
    uid: 1000  
    gid: 1000  
    comment:  
    home: /home/bob  
    shell: /bin/bash  
  - user: fred
```



passwd2yaml.awk -- Extract

```
BEGIN {  
    FS=":";  
    print "users: "  
}  
{  
    for (i=1; i<=NF; i++) {  
        switch(i) {  
            case 1:  
                printf "    - user: %s\n", $i;  
                break;  
            case 2:  
                printf "        password: %s\n", $i;  
                break;  
        }  
    }  
}
```



Demo



Using AWK Script

- Convert to YAML



Demo



Using AWK Script

- Pass variables to allow search





Tagged and XML Data



```
<VirtualHost *:80>  
DocumentRoot /www/example  
ServerName www.example.org  
# Other directives here  
</VirtualHost>
```

XML and Tagged Data

An Apache Configuration file is not exactly an XML file but uses tags in the same way as XML



```
<VirtualHost *:80>
DocumentRoot /www/example

ServerName www.example.org
# Other directives here
</VirtualHost>
<VirtualHost *:80>
DocumentRoot /www/theurbanpenguin
ServerName www.theurbanpenguin.com
# Other directives here
</VirtualHost>
$ sed -i '/^\s*$/d;/^<\/Virt/a \ ' vh.conf
```

Multiple Virtual Hosts

We may have more than virtualhost, some may be separated with a line, other, like this file may not. We can use sed, to delete all extra lines and add them only after a closing virtual host tag



Demo



Formatting a Virtual Host File

- Using sed to ensure no blank lines
- Add blank line between hosts



```
$ vim virtualhost.awk
BEGIN { RS="\n\s*\n" }
$0 ~ pattern { print }
$ awk -v pattern=example -f virtualhost.awk vh.conf
```

Working with the RS Variable

The variable RS is the record separator. This is normally a new line, but in our case we want to be the extra new line between each virtual host record.



Demo



Searching Virtual Hosts

- Implementing the RS variable



| And Breathe....



Summary



AWK and File Formats

- CSV from /etc/passwd
 - OFS=","
 - print \$1,\$3
 - Defaults to space
- YAML
 - White space sensitive
 - Indents sets blocks
 - for, switch blocks
- Tagged Data
 - RS="\n\s*\n"
 - Clean files first with sed



Up Next:

Filtering Data from Log Files

