Unique records and counting

```
$ head /var/log/access.log
175.101.18.115 - - [23/Jan/2016:16:04:58 +0200] "GET /doc/dbcp-full.png HTTP/1.1" 200 60918 "https://www.google.co.in/" "Mozilla/5.0 (Win
68.67.87.25 - - [23/Jan/2016:16:05:35 +0200] "GET /ver.html HTTP/1.1" 200 7242 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.3
134.109.104.224 - - [23/Jan/2016:16:06:31 +0200] "GET /doc/dbcp-full.png HTTP/1.1" 200 60918 "https://www.google.de/" "Mozilla/5.0 (Macin
68.67.87.25 - - [23/Jan/2016:16:06:37 +0200] "GET /ver.html HTTP/1.1" 200 7242 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.3
109.42.1.60 - - [23/Jan/2016:16:07:02 +0200] "GET /doc.html HTTP/1.1" 200 1881 "http://www.umlgraph.org/download.html" "Mozilla/5.0 (Maci
109.42.1.60 - - [23/Jan/2016:16:07:06 +0200] "GET /UMLGraph-5.7_2.3-SNAPSHOT.zip HTTP/1.1" 200 3974489 "http://www.umlgraph.org/download.
68.67.87.25 - - [23/Jan/2016:16:07:46 +0200] "GET /ver.html HTTP/1.1" 200 7242 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.3
207.46.13.54 - - [23/Jan/2016:16:08:05 +0200] "GET /robots.txt HTTP/1.1" 404 208 "-" "Mozilla/5.0 (compatible; bingbot/2.0; +http://www.b
207.46.13.54 - - [23/Jan/2016:16:08:05 +0200] "GET /robots.txt HTTP/1.1" 404 208 "-" "Mozilla/5.0 (compatible; bingbot/2.0; +http://www.b
66.249.65.41 - - [23/Jan/2016:16:08:46 +0200] "GET /robots.txt HTTP/1.1" 404 208 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www
$ logresolve /var/log/access.log >resolved
$ head resolved
175.101.18.115 - - [23/Jan/2016:16:04:58 +0200] "GET /doc/dbcp-full.png HTTP/1.1" 200 60918 "https://www.google.co.in/" "Mozilla/5.0 (Win
68-67-87-25.wavecable.com - - [23/Jan/2016:16:05:35 +0200] "GET /ver.html HTTP/1.1" 200 7242 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64) App
224-104-109-134.csn.tu-chemnitz.de - - [23/Jan/2016:16:06:31 +0200] "GET /doc/dbcp-full.png HTTP/1.1" 200 60918 "https://www.google.de/"
68-67-87-25.wavecable.com - - [23/Jan/2016:16:06:37 +0200] "GET /ver.html HTTP/1.1" 200 7242 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64) App
ip-109-42-1-60.web.vodafone.de - - [23/Jan/2016:16:07:02 +0200] "GET /doc.html HTTP/1.1" 200 1881 "http://www.umlgraph.org/download.html"
ip-109-42-1-60.web.vodafone.de - - [23/Jan/2016:16:07:06 +0200] "GET /UMLGraph-5.7_2.3-SNAPSHOT.zip HTTP/1.1" 200 3974489 "http://www.uml
68-67-87-25.wavecable.com - - [23/Jan/2016:16:07:46 +0200] "GET /ver.html HTTP/1.1" 200 7242 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64) App
msnbot-207-46-13-54.search.msn.com - - [23/Jan/2016:16:08:05 +0200] "GET /robots.txt HTTP/1.1" 404 208 "-" "Mozilla/5.0 (compatible; bingl
msnbot-207-46-13-54.search.msn.com - - [23/Jan/2016:16:08:05 +0200] "GET /robots.txt HTTP/1.1" 404 208 "-" "Mozilla/5.0 (compatible; bingl
crawl-66-249-65-41.googlebot.com - - [23/Jan/2016:16:08:46 +0200] "GET /robots.txt HTTP/1.1" 404 208 "-" "Mozilla/5.0 (compatible; Google
$ cut -d ' ' -f 1 resolved | # Obtain domain name
> head
175.101.18.115
68-67-87-25.wavecable.com
224-104-109-134.csn.tu-chemnitz.de
68-67-87-25.wavecable.com
ip-109-42-1-60.web.vodafone.de
ip-109-42-1-60.web.vodafone.de
68-67-87-25.wavecable.com
msnbot-207-46-13-54.search.msn.com
msnbot-207-46-13-54.search.msn.com
crawl-66-249-65-41.googlebot.com
$ cut -d ' ' -f 1 resolved | # Obtain domain name
> awk -F. '{print $NF}' | # Obtain top-level domain
> head
115
com
de
com
de
de
com
com
com
com
$ cut -d ' ' -f 1 resolved | # Obtain domain name
> awk -F. '{print $NF}' | # Obtain top-level domain
> grep -v '[0-9]' | # Remove numeric IP addresses
> head
com
de
com
de
de
com
com
com
com
com
$ cut -d ' ' -f 1 resolved | # Obtain domain name
> awk -F. '{print $NF}' | # Obtain top-level domain
> grep -v '[0-9]' | # Remove numeric IP addresses
> sort | # Order by TLD
> head
ar
ar
ar
ar
ar
ar
ar
ar
ar
be
$ cut -d ' ' -f 1 resolved | # Obtain domain name
> awk -F. '{print $NF}' | # Obtain top-level domain
> grep -v '[0-9]' | # Remove numeric IP addresses
> sort | # Order by TLD
> uniq -c | # Count duplicates
> head
      9 ar
      1 be
      1 bg
    101 br
      1 cn
    643 com
      2 cz
    161 de
     35 dk
      5 edu
$ cut -d ' ' -f 1 resolved | # Obtain domain name
> awk -F. '{print $NF}' | # Obtain top-level domain
> grep -v '[0-9]' | # Remove numeric IP addresses
> sort | # Order by TLD
```

```
> uniq -c | # Count duplicates
> sort -rn | # Order by number, descending
> head
    643 com
    382 net
    161 de
    101 br
     36 fi
     35 dk
     26 ru
     21 se
     15 nl
     11 it
$
```

Common records between files

```
$ ls /bin | head
bash
bunzip2
busybox
bzcat
bzcmp
bzdiff
bzegrep
bzexe
bzfgrep
bzgrep
$ ls /bin >linux.bin
$ ssh freefall.freebsd.org ls /bin >freebsd.bin
$ comm linux.bin freebsd.bin | head -20
        [
bash
bunzip2
busybox
bzcat
bzcmp
bzdiff
bzegrep
bzexe
bzfgrep
bzgrep
bzip2
bzip2recover
bzless
bzmore
                cat
chacl
        chflags
chgrp
        chio
$ comm -23 freebsd.bin linux.bin
[
chflags
chio
expr
kenv
link
pax
pgrep
pkill
pwait
rcp
realpath
rmail
test
unlink
uuidgen
$ comm -13 freebsd.bin linux.bin | wc -l
109
$ comm -12 freebsd.bin linux.bin | head
cat
chmod
cp
csh
date
dd
df
domainname
echo
ed
$
```

Relational joins

```
$ md5sum README.txt
632d2138da54c6e6c094d7f3e6b43907 *README.txt
$ find . -type f -print0 | # Output all files
> xargs -0 md5sum >md5-sum.out # Run md5-sum on each of them
$ head md5-sum.out # See resulting file
16238df89c4288520d484ac4781aecb0 *./ace/acconfig.h
c579006acd813c1c7fd38d4e2a76368e *./ace/ace/Acceptor.cpp
2899d6d59d7c338f29bff2daff7dfbcf *./ace/ace/Acceptor.h
ef3794b6ed6bb2f29419107ddf38e459 *./ace/ace/ace-dll.icc
04e28ed17630fd508480a9b6f7470882 *./ace/ace/ace-lib.icc
8b7c1d60f8d3534a4b55c7d356fe2aa2 *./ace/ace/ACE.cpp
6d9cb60083d9ffbcee1d3659ef983bb1 *./ace/ace/ace.dsw
733cd34139d60ee26afbc4a86df4a67c *./ace/ace/ACE.h
5acadd604cadf0902fd79a50381a454b *./ace/ace/ACE.i
b73fc4d2a6614eb28ce95ef478f0a05c *./ace/ace/ace.icc
```

```
$ cut -d ' ' -f 1 md5-sum.out | # Obtain first field
> head
16238df89c4288520d484ac4781aecb0
c579006acd813c1c7fd38d4e2a76368e
2899d6d59d7c338f29bff2daff7dfbcf
ef3794b6ed6bb2f29419107ddf38e459
04e28ed17630fd508480a9b6f7470882
8b7c1d60f8d3534a4b55c7d356fe2aa2
6d9cb60083d9ffbcee1d3659ef983bb1
733cd34139d60ee26afbc4a86df4a67c
5acadd604cadf0902fd79a50381a454b
b73fc4d2a6614eb28ce95ef478f0a05c
$ cut -d ' ' -f 1 md5-sum.out | # Obtain first field (MD5 sum)
> sort | # Sort
> uniq -d >duplicates # Obtain only duplicate lines
$ head duplicates # See duplicate MD5 sums
001c24c26aec0b2325c8eb698457e574
00c119acd6a63150cbabd1e3b782dc45
01a66a413bfdbc294ae15defad3e56e8
01cf48c3b81297fc57b5e6426ef20159
024144e6b5c751d98467b04ab0918306
0297c14b15cb015314034c78baf17b0f
03985ddf4bab7590ea4c53a5df655023
03e3a32c845b1fa0eb12a986542751df
04f80355c73293385e08bc02d1aa5490
053927c17147ac125efbe30c34d86a34
$ sort md5-sum.out | # Sort list of all files
> join - duplicates | # Join with duplicates on first field (MD5 sum)
> head
001c24c26aec0b2325c8eb698457e574 *./vcf/tests/Menus/bitmap1.bmp
001c24c26aec0b2325c8eb698457e574 *./vcf/tests/Toolbars/bitmap1.bmp
00c119acd6a63150cbabd1e3b782dc45 *./netbsdsrc/sys/arch/mvme68k/stand/bootst/dev_tape.h
00c119acd6a63150cbabd1e3b782dc45 *./netbsdsrc/sys/arch/sun3/stand/tapeboot/dev_tape.h
00c119acd6a63150cbabd1e3b782dc45 *./netbsdsrc/sys/arch/sun3x/stand/tapeboot/dev_tape.h
01a66a413bfdbc294ae15defad3e56e8 *./apache/src/ap/.indent.pro
01a66a413bfdbc294ae15defad3e56e8 *./apache/src/include/.indent.pro
01a66a413bfdbc294ae15defad3e56e8 *./apache/src/main/.indent.pro
01a66a413bfdbc294ae15defad3e56e8 *./apache/src/modules/example/.indent.pro
01a66a413bfdbc294ae15defad3e56e8 *./apache/src/modules/experimental/.indent.pro
$
```

Processes as input arguments

```
$ comm -23 <( # Specify first input
> tar tvf dds.20160512.tar | # List first archive contents
> cut -c 49- | # Isolate file name
> sort) <( # Sort and specify second input
> tar tvf dds.20160518.tar | # List second archive contents
> cut -c 49- | # Isolate file name
> sort) | # Sort
> head -5
/home/dds/src/master/.git/objects/fc
/home/dds/src/master/.git/objects/fc/0c98d54d098b8923a427a9fcfd30f9455200e8
/home/dds/src/master/.git/objects/fc/21cf3199146431e8305b20f60e0d9ca58fb143
/home/dds/src/master/.git/objects/ff
/home/dds/src/master/.git/objects/ff/63a6b8ebc396d9caeac11c51e2ace21235191a
/home/dds/src/master/.git/objects/ff/a24395c548c2643340df1d6890800364e3fd80
$
```

Processes as output arguments

```
$ ls |
> tee >(wc -l) # Count output's lines
if-down.d
if-post-down.d
if-pre-up.d
if-up.d
interfaces
interfaces.d
run
7
$
```