A Linux and/or VM Solution for Analyzing any Web Server Log

SHARE 99, Session 9108, August 2002

Gordon W. Wolfe, Ph.D.

Senior VM Systems Programmer

VM Technical Services, Shared Systems Group, The Boeing Company

e-mail gordon.w.wolfe@boeing.com

Who's Looking At Your Webserver?

- Is your website effective?
- Does anyone look at it?
- WHO's looking at it? Where are they?
- Do they come direct or are they referred by another site?
- WHAT do they look at?
- How many hits are you getting?

The Problem at Boeing

- CA VM:Webserver sites have over 1 million hits per month
- Unix and NT servers have logs too big to analyze locally, want to run on mainframe.
- How do we find out what's the most effective use of hardware and software?
- Tried writing our own analysis program,but must be customized for every server.

Stephen Turner's ANALOG Program

- Someone's already written this program!
- Stephen Turner, University of Cambridge (UK) Statistical Laboratory
- The program is FREE!
- *The* standard analysis program - runs on 90% of analysis sites.
- Extremely flexible - will analyze virtually all webserver log formats, with APACHE and NT built in.
- Written in C, runs on Unix, Windows, Macintosh, z/OS, DEC, AS/400, OS/2, BUT NOT VM or Linux!

Porting ANALOG 4.0 to VM/OpenExtension

- This author took Unix version and ported to VM OpenExtension. Simpler than rewriting whole thing to use CMS file structures.
- Used Neale Ferguson's "Porting UNIX Applications to OpenEdition for VM/ESA" redbook SG24-4747-00
- Port was not simple, even though program already had provision for EBCDIC.
  - Rewrote Makefile
  - Migrate LE370 H files into /usr/lib/include. Resolved conflicts with TCPIP H files of same name.
  - Rewrite 3 of Turner's H files and one subroutine for VM/CMS
  - Mr. Turner had one small bug in his code that only affected VM/CMS.
  - Took 4 months to do the port!
  - Complete compile and link-edit takes 40 minutes on R85!
- Currently working on port for Analog 5.0

Porting ANALOG 5.0 to Linux/390

- Linux/390 port was simple.
- Took UNIX version of ANALOG, copied to Linux. Compliled, linked and ran clean the first time!

Running from OE

- Put ANALOG program in /usr/local/bin
- Put run-time files in /usr/local/Analog4.0/
  - Usdom.tab
  - Us.lng
- Put work files in /tmp or on CMS
  - Dns.file
- Put analog.cfg in user directory /home/user
  - Config file is the key to running Analog
  - Tells where the data is
  - Tells where the run-time files are
  - Tells how you want the data analyzed.
- Run by entering analog +g /home/user/analog.cfg

Running from CMS Directly

- Takes advantage of the fact that you can run OE directly from CMS with OPENVM EXEC
- Create an ANALOG EXEC to run it, place on 19E disk (See Appendix)
- OE programs can call CMS files by simply using the structure //filename.filetype.filemode
- Run using EXEC ANALOG configfn configft configfm
- 350,000 records analyzed in 9 seconds!

Remote Jobs

- Use CA's VM:Batch's RJE facility
- ANALOG EXEC on 19E disk
- For accounting, set up dummy userids
  - USER REMOTE1 XXX G - perhaps same as node name of remote server
  - ACCOUNT CHARGE1
  - No IPL card or MDISK cards - charge only for CPU time.
- Then send job deck to VMBATCH in RJE format
  - lst card: *RJECMD USER REMOTE1 PASS XXX NODE REMOTE1
  - 2nd card: EXEC ANALOG ANALOG CONFIG A
  - 3rd card: *ENDCMD
  - 4th card :READ ANALOG CONFIG A
  - Add cards for above file
  - 5th card :READ ANALOG DATA A
  - Add cards for above file
- Send whole deck in NETDATA format.

Running ANALOG on Linux

- Binary executable /usr/bin/analog
- Place files in /usr/src: /usr/src/analog-5.01/lang/usdom.tab and /usr/src/analog-5.01/lang/us.lng. Also analogo.gif to /analog subdir of /html subdirectory of APACHE.
- Dummy control file /usr/bin/analog.cfg just has one entry: CONFIGFILE /etc/analog.cfg to point to real config file in r/w subdirectory /etc.
- Point /etc/analog.cfg entries to where you have (or want) files:
  - LOGFILE /var/log/httpd/access_log <==the log file to be analyzed
  - OUTFILE /home/httpd/html/analog.html <==where you want the output (for APACHE?)
  - IMAGEDIR analog/ <== output relative to above directory, e.g. /home/httpd/html/analog
  - DOMAINSFILE /usr/src/analog-5.01/lang/usdom.tab
  - LANGFILE /usr/src/analog-5.01/lang/us.lng
  - DNSFILE /tmp/dns.file <== Temporary DNS lookup file created by analog
  - DNSLOCKFILE /tmp/dnslock <== Temp lock file created by analog
- Then to run just enter /usr/bin/analog

The Config File - Key to ANALOG processing

- LOGFORMAT … to inform analog of format of logs
- LOGFILE tells analog name of log file to analyze
- OUTFILE tells analog where to put results. Output is report or direct to HTML for web display.
- DOMAINSFILE - where the table of domains is
- DNSFILE tells analog where to put the dns cache file
- All above are minimum requirements to run Analog.
- Other entries to specify what reports you want, what data to include and exclude, what certain file extensions mean, and so on.

LOGFORMAT values

- Analog's default is CERN/NCSA common format, which is used by NCSA server and APACHE.
- Has built-in values for Windows (95,98,NT), WEBSTAR and NETSCAPE formats
- Anything else requires custom format
  - Analog is really picky about the format
  - Must match the format of the log exactly
- For VM:Webserver, use the VIWLOGEX utility to convert log files to CERN format.

Sample Results 1

General Summary

(Figures in parentheses refer to the 7-day period ending Jul 09 2001 at 4:03 PM).
Successful requests: 17,253 (487)
Average successful requests per day: 811 (69)
Successful requests for pages: 14,573 (284)
Average successful requests for pages per day: 685 (40)
Failed requests: 4,387 (10)
Redirected requests: 1 (0)
Distinct files requested: 16,001 (280)
Distinct hosts served: 33 (13)
Data transferred: 126.132 Mbytes (7.734 Mbytes)
Average data transferred per day: 5.936 Mbytes (1.104 Mbytes)

BOEING

Sample Results 2

Hourly Summary

Each unit (▬) represents 80 requests for pages or part thereof.

hour:	#reqs:	#pages:
0:	0:	0:
1:	0:	0:
2:	1:	1:
3:	0:	0:
4:	1341:	1195:
5:	1622:	1609:
6:	1642:	1362:
7:	95:	8:
8:	73:	37:
9:	56:	8:
10:	88:	16:
11:	61:	14:
12:	42:	6:
13:	65:	7:
14:	119:	12:
15:	52:	6:
16:	113:	108:
17:	113:	113:
18:	95:	93:
19:	2740:	1808:
20:	3385:	3285:
21:	3644:	3013:
22:	1314:	1280:

Sample Results 3

Listing domains, sorted by the amount of traffic.

```
#reqs: %bytes: domain
-----: ------: ------
17232: 99.86%: .com (Commercial)
17232: 99.86%:   boeing.com
  274:  1.37%:     ca.boeing.com
   21:  0.14%: [unresolved numerical addresses]
```

BOEING

Sample Results 4

Listing extensions with at least 0.1% of the traffic, sorted by the amount of traffic.

```
#reqs: %bytes: extension
-----: ------: ---------
13905: 58.28%: .html [Hypertext Markup Language]
  934:  9.16%: [no extension]
  370:  7.24%: .txt  [Plain text]
   12:  6.73%: .pdf  [Adobe Portable Document Format]
  637:  3.29%: [directories]
  454:  2.24%: .gif  [GIF graphics]
   16:  1.82%: .dvi
    1:  1.81%: .rtf  [Rich Text Format]
   31:  1.33%: .htm  [Hypertext Markup Language]
    2:  0.92%: .src
   22:  0.92%: .1
  119:  0.62%: .jpg  [JPEG graphics]
   21:  0.59%: .doc  [Microsoft Word document]
   75:  0.45%: .c
   48:  0.35%: .yo
   17:  0.32%: .tex
```

BOEING

Sample Results 5

Request Report

Listing files with at least 20 requests, sorted by the number of requests.

#reqs:	%bytes:	last time:	file
73:	0.49%:	Jul/ 9/01 1:17 PM:	/
61:	0.22%:	Jul/ 9/01 1:17 PM:	/bkgrnd.jpg
60:	1.76%:	Jul/ 9/01 1:17 PM:	/penguin.gif
58:	0.06%:	Jul/ 9/01 1:17 PM:	/g4700.gif
58:	0.40%:	Jul/ 9/01 1:17 PM:	/linux6.jpg
57:	0.08%:	Jul/ 9/01 1:17 PM:	/linux390.gif
57:	0.09%:	Jul/ 9/01 1:17 PM:	/apache.gif
57:	0.07%:	Jul/ 9/01 1:17 PM:	/suse.gif
57:	0.06%:	Jul/ 9/01 1:17 PM:	/s390.gif
57:	0.12%:	Jul/ 9/01 1:17 PM:	/ssg.gif
16658:	96.65%:	Jul/ 9/01 1:18 PM:	[not listed: 15,991 files]

BOEING

Where Can I Get Analog?

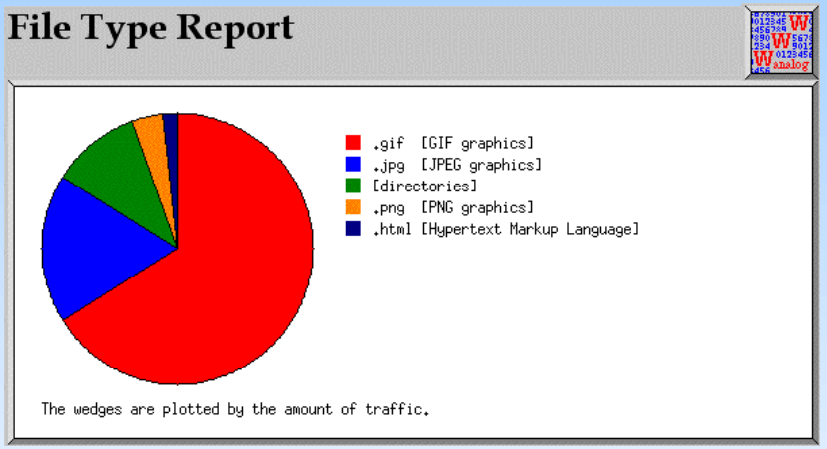
- CMS version 4.0 from this author or at Neale Ferguson's OE site: http://pucc.princeton.edu/~neale/vmoe03.html#ANALOG or from this author.
- Linux/390 version 5.0 from http://linuxvm.org/Info/analg501.tgz or this author only at this time.
- Source code at Stephen Turner's site: http://www.analog.cx/
- Complete on-line manual for ANALOG also at Turner's site

ANALOG Listserver

- Have questions or problems with ANALOG?
- Try the ANALOG listserver.
- Questions answered by people who've had the same problems
- Frequently will be answered by Stephen Turner himself!
- http://www.analog.cx/docs/mailing.html

Analog 5.0

- Minor new features, commands and bug fixes
- Big news is pie charts and JPEGs!



BOEING

Appendix: Sample Configfile for VM:Webserver

```
  LOGFORMAT COMBINED
  LOGFILE //SCM1.CERN0106.C
  HOSTNAME "Boeing VM:Webserver"
  LANGFILE /usr/local/analog4.0/us.lng
  OUTFILE //SCM10106.HTML.F
  DOMAINSFILE /usr/local/analog4.0/ukdom.tab
  OUTPUT HTML
  DNSFILE //dns.file
  DNS WRITE
  DNSGOODHOURS 800
  IMAGEDIR images/
GOTOS OFF
REQINCLUDE pages
LINKINCLUDE pages
REFLINKINCLUDE pages
UNCOMPRESS *.gz,*.Z "gzip -cd"
BROWOUTPUTALIAS Mozilla Netscape
BROWOUTPUTALIAS "Mozilla (compatible)" "Netscape (compatible)"
BROWOUTPUTALIAS IWENG AOL
SUBTYPE *.gz,*.Z
TYPEOUTPUTALIAS .html    ".html [Hypertext Markup Language]"
TYPEOUTPUTALIAS .htm     ".htm  [Hypertext Markup Language]"
TYPEOUTPUTALIAS .ps      ".ps   [PostScript]"
TYPEOUTPUTALIAS .gz      ".gz   [Gzip compressed files]"
...
```

Appendix: Sample Configfile for Linux Apache

```
LOGFILE /var/log/httpd/access_log
OUTFILE /home/httpd/html/analog.html
DOMAINSFILE /usr/src/analog-5.01/lang/usdom.tab
LANGFILE /usr/src/analog-5.01/lang/us.lng
HOSTNAME "Linux/390 Apache Webserver"
OUTPUT HTML
DNSFILE /tmp/dns.file
DNSLOCKFILE /tmp/dnslock
DNS WRITE
DNSGOODHOURS 800
# following directory is relative to /home/httpd/html/
IMAGEDIR analog/
GOTOS OFF
LINKINCLUDE pages
REQLINKINCLUDE pages
REFLINKINCLUDE *
REDIRREFLINKINCLUDE *
FAILREFLINKINCLUDE *
OSREP ON
UNCOMPRESS *.gz,*.Z "gzip -cd"
BROWOUTPUTALIAS Mozilla Netscape
BROWOUTPUTALIAS "Mozilla (compatible)" "Netscape (compatible)"
BROWOUTPUTALIAS IWENG AOL
SUBBROW */*
SUBTYPE *.gz,*.Z
# Add whichever of these types of pages you have on your server, or others.
PAGEINCLUDE *.shtml
PAGEINCLUDE *.asp
PAGEINCLUDE *.jsp
```

Appendix: Sample EXEC for CMS

```
/* EXEC to run Analog 4.0 from CMS                                    */
/* 'EXEC ANALOG' <config file>                                        */
/*          the config file should reside on your a-disk.             */
/* by Gordon Wolfe, VM Technical Services                  03/16/00 */
address command
arg cfgfn cfgft cfgfm .

'EXEC OPENVM MOUNT /../VMBFS:VMSYS:ROOT/ /'
'EXEC OPENVM SET DIRECTORY /home/webstat'

r = rc
if r <> 0 then do
  say 'Cannot mount BFS'
  exit r
end

'EXEC OPENVM RUN /usr/local/bin/analog +g//'cfgfn'.'cfgft'.'cfgfm
exit rc
```