Technical Report 01−2003−09−08

A shell script to update statistics for web servers using http−analyze

If you run **http−analyze** for one server only, than a simple *cron(8)* job will do that. But if you are in the hosting business where several virtual hosts have to be processed, a shell script comes handy to automate the process. This report explains the helper utility **run−ha**, a shell script to run **http−analyze** on the logfiles of all web servers. **run−ha** uses **ipresolve** to resolve IP numbers into hostnames and **http−analyze** to create a statistics report for the server.

The first version of **run−ha** uses a hard−coded list of server names defined in **SERVERLIST**. It assumes that the server root is **/www/***sitename*, that the statistics directory is **/www/***sitename***/stats** and that logfiles are below **/www/***sitename***/httpd/logs/**. The name of the current logfile is **access** and archived logfiles are named **log***YYYY***/access.***MM***.gz** (gzip’ed format), where *YYYY* is the year and *MM* is the month. The configuration file is **/www/***sitename***/httpd/analyze.conf**. The options **−d**, **−m** and **−v** are passed directly to **http−analyze**. If a month and optionally a year is given, the logfile for this period is used. If no month is specified, the logfile for the current month will be processed.

```
#!/bin/ksh
#
# Run http−analyze for all servers
#
USAGE="$(basename $0) [−mdv] [MM [YYYY]]"

HA_PATHNAME=/usr/local/bin/http−analyze         # pathname of http−analyze
HA_CONFNAME=httpd/analyze.conf                  # name of configuration file

SERVERLIST="clientA clientB"                    # list of servers to analyze
SERVERROOT=/www                                 # name of the server root

ECHO=": "

while [ $# −gt 0 ]; do
        case $1 in
         −[md]) DEFMODE="$DEFMODE $1" ;;
         −v)    DEFMODE="$DEFMODE $1" ; ECHO=echo ;;
         [01][0−9])             MONTH="$1" ;;
         [012][0−9][0−9][0−9])  YEAR="$1" ;;
         *)     echo "Invalid parameter: $1\nUsage: $USAGE" 1>&2; exit 1 ;;
        esac
        shift
done

if [ −n "$MONTH" ]; then
        : ${DEFMODE:=−m}
        : ${YEAR:=$(date +%Y)}
        LOGFILE="httpd/logs/log$YEAR/access.$MONTH"
else
        : ${DEFMODE:=−d}
fi

cd $SERVERROOT || { echo "panic: can’t change into $SERVERROOT" 1>&2; exit 1; }

for server in $SERVERLIST; do
        $ECHO "Generating new statistics for $server"
        if [ −z "$LOGFILE" ]; then           # use default logfile from configuration file
            $HA_PATHNAME $DEFMODE −3f −c $server/$HA_CONFNAME
        elif [ −f "$server/$LOGFILE" ]; then
            $ECHO "processing $SERVERROOT/$server/$LOGFILE"
            $HA_PATHNAME $DEFMODE −3f −c $server/$HA_CONFNAME $server/$LOGFILE
        elif [ −f "$server/${LOGFILE}.gz" ]; then
            $ECHO "processing $SERVERROOT/$server/${LOGFILE}.gz"
            gzcat $server/${LOGFILE}.gz |
                $HA_PATHNAME $DEFMODE −3f −c $server/$HA_CONFNAME −
        else
            echo "no logfile for $MONTH $YEAR of $server found" 1>&2
        fi
        $ECHO "\n\c"
done
exit 0
```

Continued on page 2

The second version of **run−ha** is a **bash**−script, which creates a list of servers to process automatically from the configuration file of an Apache web server. The server root is **/www/vhosts/***sitename*, where *sitename* is the domain name of the virtual server (either **www.site.domain** or just **site.domain**). The document root is under **/www/vhosts/***sitename***/htdocs/**. The statistics directory is located below the document root (**htdocs/stats/**). The logfiles are under **/www/vhosts/***sitename***/logs/**. A missing **logs** directory means that **run−ha** should skip this server. In addition to **http−analyze**, the program **ipresolve** is called to resolve IP numbers into hostnames.

```
#
# Run http−analyze for all our customers
#
USAGE="$(basename $0) [−v]"

APACHE_CFG=/usr/local/etc/httpd/conf/httpd.conf

IR_PATHNAME=/usr/local/bin/ipresolve
IR_DATABASE=/var/tmp/DNS
HA_PATHNAME=/usr/local/bin/http−analyze
HA_CONFNAME=http−analyze.conf
DEFMODE="−m3f −F elf −b 102400"

SERVERLIST=‘sed −n ’s/^ServerName[      ][      ]*\(.*\)/\1/p’ $APACHE_CFG‘
SERVERROOT=/www/vhosts

ECHO=": "

while [ $# −gt 0 ]; do
        case $1 in
        −h)     echo "Usage: $USAGE"; exit 0 ;;
        −v)     DEFMODE="$DEFMODE $1" ; ECHO=echo ;;
        −e)     ECHO=echo ;;
        [01][0−9])              MONTH="$1" ;;
        [012][0−9][0−9][0−9])   YEAR="$1" ;;
        *)      echo "Invalid parameter: $1\nUsage: $USAGE" 1>&2; exit 1 ;;
        esac
        shift
done

if [ −n "$MONTH" ]; then
        : ${YEAR:=$(date +%Y)}
        LOGTEMPLATE="log$YEAR/access.$MONTH"
fi

# change into the root directory of all virtual hosts
cd $SERVERROOT || { echo "Couldn’t change into directory $SERVERROOT" 1>&2; exit 1; }

for server in $SERVERLIST; do
        #
        # First, check for the document root with and w/o the prefix "www."
        #
        if [ −d www.$server/ ]; then server=www.$server; fi
        if [ ! −d $server ]; then
                echo "No document root found for $server in $SERVERROOT" 1>&2
                continue
        fi

        # absence of logs directory means skip this vhost
        if [ ! −d $server/logs ]; then continue; fi

        # select proper logfile
        if [ −n "$MONTH" ]; then
                LOGFILE="$server/logs/log$YEAR/access_log.$MONTH"
        else    LOGFILE="$server/logs/access_log"
        fi

        if [ ! −s "$LOGFILE" ]; then
                echo "Can’t open the logfile \‘$SERVERROOT/$LOGFILE’ for vhost $server" 1>&2
                continue
        fi

        # Create the statistics directory
        STATSDIR=$server/stats
        if [ ! −d "$STATSDIR" ]; then mkdir $STATSDIR 2>/dev/null; fi
        if [ ! −d "$STATSDIR" ]; then
                echo "Can’t create directory \‘$SERVERROOT/$STATSDIR’ for statistics" 1>&2
                continue
        fi

        #
        # Generate statistics
        #
        $ECHO "Generating new statistics for $server"
        if [ −f $server/$HA_CONFNAME ]; then
                $IR_PATHNAME −d $IR_DATABASE $LOGFILE |
                    $HA_PATHNAME $DEFMODE −c $server/$HA_CONFNAME −o $STATSDIR −
        else
                $IR_PATHNAME −d $IR_DATABASE $LOGFILE |
                    $HA_PATHNAME $DEFMODE −S $server −o $STATSDIR −
        fi
done
exit 0
```

Continued on page 3

To have the **run−ha** script executed automatically, add an entry like the following in the crontab of the server user (do **NOT** use **root**'s crontab!):

```
# crontab file for http−analyze statistics
#
# Format of lines:
# min hour daymo month daywk cmd
#
17  1,13 * * * /usr/local/bin/run−ha −m
17  8−22 * * * /usr/local/bin/run−ha −d
```

This causes the script to be called at 01:17 and 13:17 each day to produce a full statistics and each hour between 08:17 and 22:17 to produce a short (**−d**) statistics report. On the 1st of a month, the **rotate−httpd** script, which rotates (saves) the logfile can also call **run−ha** explicitely, in which case you should avoid to run it via *cron(8)*. See *TR−02−2003−09−08* for more information.

The script can be customized to almost all web server installations. Some things to keep in mind are:

❑ Always have **http−analyze** process all logfiles for the current month in full statistics mode, otherwise the statistics can get zeroed.

❑ If you analyze older periods than the current month, anything between this period and the current month will be lost except when you specify the **−n** option to not update the history! **run−ha** passes this option to **http−analyze**.

❑ If you use **http−analyze** 2.5 or **ipresolve** 2.0, they can both read gzip'ed logfiles directly.

❑ The **logs** subdirectory is always under the site's server root. The **stats** subdirectory can reside inside or outside the document root. Access to certain parts of the statistics or the whole report might require authentication. The configuration file for **http−analyze** should always be outside the document root.

run−ha is included in **http−analyze** since version 2.4.
rotate−httpd is included in **http−analyze** since version 2.4.
ipresolve 2.0 is available through our Customer Support site and will shortly become available to everyone.

Please send comments, enhancements, tips and tricks to: **office@rent−a−guru.de**.