1. Úloha IOS (2019)

Popis úlohy

Cílem úlohy je vytvořit skript pro analýzu záznamů webového serveru. Skript bude filtrovat záznamy a poskytovat statistiky podle zadání úživatele.

Specifikace chování skriptu

JMÉNO

- `wana` - analyzátor webových logů

POUŽITÍ

- `wana [FILTR] [PŘÍKAZ] [LOG [LOG2 [...]]`

VOLBY

- `PŘÍKAZ` může být jeden z:
  - `list-ip` – výpis seznamu zdrojových IP adres.
  - `list-hosts` – výpis seznamu zdrojových doménových jmen.
  - `list-uri` – výpis seznamu cílových zdrojů (URI).
  - `hist-ip` – výpis histogramu četností dotazů podle zdrojových IP adres.
  - `hist-load` – výpis histogramu zátěže (tj. počtu dotazů ve jednotlivých časových intervalech).
- `FILTR` může být kombinace následujících:
  - `-a DATETIME` – after = jsou uvažovány pouze záznamy PO tomto datu (bez tohoto data). `DATETIME` je formátu `YYYY-MM-DD HH:MM:SS`.
  - `-b DATETIME` – before, jsou uvažovány pouze záznamy PŘED tímto datem (bez tohodo data).
  - `-ip IPADDR` – jsou uvažovány pouze záznamy odpovídající požadavkům ze zdrojové adresy `IPADDR`. Formát `IPADDR` odpovídá IPv4 nebo IPv6.
  - `-uri URI` – jsou uvažovány pouze záznamy týkající se dotazů na konkrétní webovou stránku. `URI` je základní regulární výraz.

POPIS

1. Skript filtruje záznamy z webového serveru. Pokud je skriptu zadán také příkaz, nad filtrovanými záznamy daný příkaz provede.
2. Pokud skript nedostane ani filtr ani příkaz, opisuje záznamy na standardní výstup.
3. Skript umí zpracovat záznamy webového serveru komprimované pomocí nástroje gzip (v případě, že název souboru končí `.gz`).

4. V případě, že skript na příkazové řádce nedostane soubory se záznamy webového serveru (`LOG`, `LOG2` ...), očekává záznamy na standardním vstupu.
5. Pokud má skript vypsat seznam, každá položka je vypsána na jeden řádek a pouze jednou. Na pořadí nezáleží. Položky se nesmí opakovat.
6. Víceřádkový histogram je vykreslen pomocí ASCII a je otočený doprava. Každý řádek histogramu udává kategorii (např. IP adresu nebo časový interval). Četnost dané kategorie je vyobrazena posloupností znaku mřížky `#`. Formát je `"%s (%d): %s"`, kde první argument identifikuje kategorii, druhý je četnost v číselné podobě a třetí je četnost vykreslená pomocí mřížek.
7. Histogram podle IP adres (`hist-ip`) je seřazen od nejčetnějších po nejméně četné dotazy.
8. Histogram zátěže (`hist-load`) má jednotlivé časové intervaly po celých hodinách. Do dané kategorie spadají všechny záznamy počínající danou hodinou. V histogramu budou pouze časové údaje s nenulovým výskytem záznamů. Formát kategorie je `YYYY-MM-DD HH:00`. Celkový časový rozsah je dán časovým rozsahem vstupních nebo filtrovaných záznamů.
9. URI je identifikátor, který se v záznamu nachází za identifikátorem metody protokolu HTTP, viz RFC2616, Sec. 9.

PODROBNÉ POŽADAVKY

1. Skript analyzuje záznamy (logy) pouze ze zadaných souborů.
2. Skript žádný soubor nemodifikuje. Skript nepoužívá dočasné soubory.
3. IP adresa může být IPv4 (např. `147.229.176.19`), IPv6 standardního (např. `2001:67c:1220:8b0:0:0:93e5:b013`) nebo IPv6 komprimovaného (např. `2001:67c:1220:8b0::93e5:b013`) formátu (viz RFC 1884, sekce 2.2).
4. Skript nebere ohled na význam IP adres. IP adresy rozlišuje podle jejich textové reprezentace.
5. Skript neuvažuje časové zóny. Předpokládá se, že všechny záznamy i filtry podle data mají časovou značku ve stejné časové zóně.
6. Doménové jméno podle IP adresy zjistěte pomocí příkazu `host`. Pokud nelze doménové jméno získat, bude místo něj použita IP adresa.

NÁVRATOVÁ HODNOTA

- Skript vrací úspěch v případě úspěšné operace. Interní chyba skriptu nebo chybné argumenty budou doprovázeny chybovým hlášením a neúspěšným návratovým kódem.

Implementační detaily

- Skript by měl mít v celém běhu nastaveno `POSIXLY_CORRECT=yes`.
- Skript by měl běžet na všech běžných shellech (dash, ksh, bash). Můžete použít GNU rozšíření pro `sed` či `awk`. Jazyk Perl nebo Python povolen není.

- Skript musí běžet na běžně dostupných OS GNU/Linux, BSD a MacOS. Studentům je k dispozici virtuální stroj s obrazem ke stažení zde: http://www.fit.vutbr.cz/~lengal/public/trusty.ova (pro VirtualBox, login: trusty / heslo: trusty), na kterém lze ověřit správnou funkčnost projektu.
- Skript nesmí používat dočasné soubory. Povoleny jsou dočasné soubory nepřímo tvořené příkazem `sed` (např. argument `sed -i`).

Příklady použití

- Ukázky záznamů webového serveru jsou dostupné zde: https://pajda.fit.vutbr.cz/ios/ios-19-1-logs

Příklady:

```
$ ./wana list-ip ios-example.com.access.log
147.229.13.201
198.27.69.191
2001:67c:1220:808::93e5:8ad
2001:67c:1220:80c:d4:985a:df2c:d717
40.77.167.115
66.249.66.45
66.249.66.49
82.202.69.253

$ ./wana list-hosts ios-example.com.access.log
hele.fit.vutbr.cz.
ns504614.ip-198-27-69.net.
perchta.fit.vutbr.cz.
2001:67c:1220:80c:d4:985a:df2c:d717
msnbot-40-77-167-115.search.msn.com.
crawl-66-249-66-45.googlebot.com.
crawl-66-249-66-49.googlebot.com.
82.202.69.253

$ ./wana -a "2019-02-22 9:00" -b "2019-02-22 9:44:54" ios-example.com.access.log

147.229.13.201 - - [22/Feb/2019:09:24:33 +0100] "-" 408 3275 "-" "-"
147.229.13.201 - - [22/Feb/2019:09:24:33 +0100] "-" 408 3275 "-" "-"
198.27.69.191 - - [22/Feb/2019:09:43:13 +0100] "GET / HTTP/1.1" 200 22311 "-" "Mozilla/5.0 (
198.27.69.191 - - [22/Feb/2019:09:43:24 +0100] "GET / HTTP/1.1" 200 22313 "-" "Mozilla/5.0 (
198.27.69.191 - - [22/Feb/2019:09:43:42 +0100] "GET /?gf_page=upload HTTP/1.1" 200 22304 "-"
198.27.69.191 - - [22/Feb/2019:09:44:07 +0100] "GET / HTTP/1.1" 200 22313 "-" "Mozilla/5.0 (
198.27.69.191 - - [22/Feb/2019:09:44:37 +0100] "GET /?up_auto_log=true HTTP/1.1" 200 22315 "
198.27.69.191 - - [22/Feb/2019:09:44:54 +0100] "GET /wp-admin/ HTTP/1.1" 302 3711 "-" "Mozil

$ ./wana -a "2019-02-22 9:00" -b "2019-02-22 9:44:54" list-uri ios-example.com.access.log
```

```
/
/?gf_page=upload
/?up_auto_log=true
/wp-admin/

$ ./wana hist-ip ios-example.com.access.log
198.27.69.191 (8): ########
2001:67c:1220:80c:d4:985a:df2c:d717 (4): ####
82.202.69.253 (2): ##
2001:67c:1220:808::93e5:8ad (2): ##
147.229.13.201 (2): ##
66.249.66.49 (1): #
66.249.66.45 (1): #
40.77.167.115 (1): #

$ ./wana -ip 2001:67c:1220:808::93e5:8ad hist-load ios-example.com.access.log.1
2019-02-21 08:00 (1): #
2019-02-21 10:00 (1): #
2019-02-21 14:00 (1): #
2019-02-21 16:00 (1): #
2019-02-21 19:00 (1): #
2019-02-21 20:00 (1): #
2019-02-21 22:00 (1): #
2019-02-21 23:00 (1): #
2019-02-22 02:00 (1): #
2019-02-22 03:00 (2): ##
2019-02-22 05:00 (1): #
2019-02-22 07:00 (1): #

$ ./wana -uri "/robots\.txt" list-hosts *log*
msnbot-157-55-39-17.search.msn.com.
msnbot-157-55-39-35.search.msn.com.
crawl7.bl.semrush.com.
crawl22.bl.semrush.com.
crawl-66-249-66-45.googlebot.com.
crawl-66-249-66-49.googlebot.com.

$ ./wana -uri "/robots\.txt" hist-load *log*
2019-02-20 13:00 (2): ##
2019-02-20 18:00 (1): #
2019-02-21 11:00 (1): #
2019-02-21 23:00 (1): #
2019-02-22 07:00 (2): ##
2019-02-22 10:00 (1): #
```