

Webalizer

[grafička analiza logova]

priređio: Dinko Korunić
verzija 1.0, siječanj 2002.

Tijekom prezentacije

- ako što **nije jasno** - pitajte!
- ako što **nije točno** - ispravite!
- diskusija je **poželjna** i **produktivna**
- ako je **prebrzo** - tražite da se uspori!
- ako je pak **presporo** i **uspavljuje** vas - lako se ubrza sa sadržajem
- podijelimo zajedno vlastita stručna iskustva

Ciljevi prezentacije

- osnovni prikaz paketa **Webalizer**
 - instalacija paketa na sistem (Linux, Unix)
 - osnovno korištenje programa
- kako koristiti i iskoristiti **Webalizer**
 - tipovi logova
 - razumijevanje rezultata
 - što se može poboljšati
 - napredna instalacija i konfiguracija

Potrebno predznanje

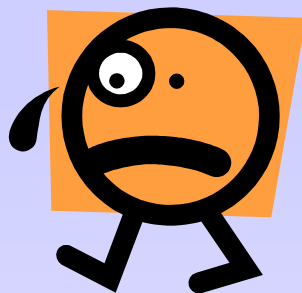
- **osnovna** računalna pismenost
 - datoteke, direktoriji, hijerarhija
 - korišćenje tipkovnice, miša
- **poželjno** poznavanje
 - osnovnog rada u nekom tekst editoru ili uređivaču teksta
 - rad sa httpd-om, npr. Apache
 - formati Apache logova

Ideja paketa

- **grafičko predočenje** posjećenosti:
 - kratki pregled
 - detaljno - prema sadržajima, zemljama, itd.
- **ulazno/izlazna propusnost** koju zahtijeva Web site
- povijest posjećenosti prema mjesecima
- podaci na atraktivan način iz često "**nečitljivih**" logova

Značajke

- automatski resolve IP adresa (asyncDNS)
- generiran html
 - **grafički i tekstualni** prikaz 12 **mjesečne** analize
- **cronjob** ili ručno pokretanje
- automatska **dekompresija** logova
- semi-inteligentno parsiranje
 - različiti tipovi logova
- **inkrementalno** procesiranje



Webalizer (1)

izvedba,
osnovno konfiguriranje,
pokretanje i korištenje

Priprema paketa

- `./configure --with-language=croatian && make all install`
- `--with-gdlib=<staza-do-gd-librarya>`
- `--with-gd=<staza-do-gd-includeova>`
- `--enable-dns`

- **libgd:** <http://www.boutell.com/gd/>
- **zlib:** <http://www.libpng.org/pub/png/>
- **jpeg6b:** <http://www.ijg.org/>

Izvedba paketa

- izvršne datoteke:
 - `webalizer`, `webazolver` (simbolički link!)
- man stranice:
 - `webalizer.1`
- txt dokumentacija:
 - `Readme`, `Install`, `DNS.README`, `README.FIRST`
- konfiguracijska datoteka:
 - `sample.conf`

Konfiguriranje (općenito)

- komplicirana, centralizirana konfiguracija:
 - `webalizer.conf`
 - bez "-c" traži `./webalizer.conf` i `/etc/webalizer.conf`
- može i kao parametar:
 - `webalizer -c imedatoteke`
 - `imedatoteke` može biti i `/staza/imedatoteke`
- konfiguracijskih opcija mnogo, utiču na **performanse i izlazne rezultate**

Konfiguriranje (quickstart)

- podesiti **webalizer.conf**:

- LogFile /staza/ulazna_log_datoteka
 - OutputDir /staza_za_generirani_web
 - HostName ime_poslužitelja

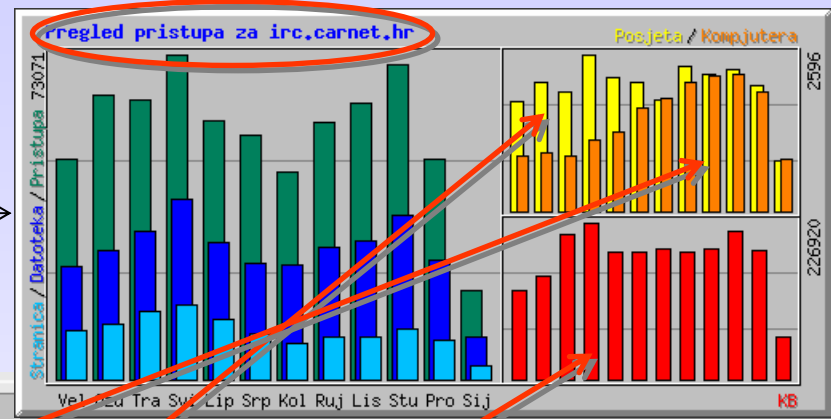
- **crontab**:

- 0 * * * * /usr/local/bin/webalizer

- neprecizna konfiguracija, nepraktična za **virtualne hostove**, ne radi resolve IP

Primjer statistike

grafičko reprezentiranje
tekstualnih rezultata

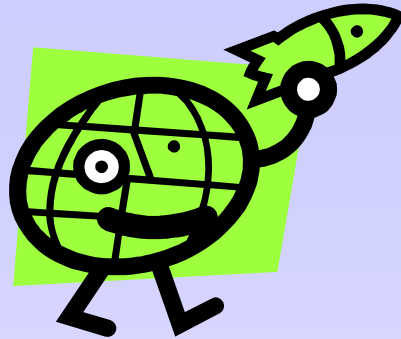


Pregled po mjesecima

Mjesec	Dnevni prosjek				Mjesečni brojevi						
	Pristupa	Datoteka	Stranica	Posjeta	Kompjuteru	KP	Posjeta	Stranica	Datoteka	Pristupa	
Sij 2002	2204	1068	359	93	898	61744	844	3233	9612	19839	
Pro 2001	1598	804	279	66	1964	187347	2070	8660	26721	49568	
Stu 2001	2363	1229	373	77	2266	214952	2338	11192	36879	70907	
Lis 2001	2601	1009	313	72	2230	189781	2600	9708	31304	62041	
Ruj 2001	1920	994	321	79	2141	184024	2396	9659	29804	57614	
Kol 2001	1504	826	265	69	1864	188231	1845	8219	25630	46642	
Srp 2001	1769	842	333	69	1710	184634	2142	10351	26115	54844	
Lip 2001	1939	1020	445	73	1312	184240	2200	13359	30608	58187	
Svi 2001	2357	1300	542	83	1180	226920	2596	16841	40330	73071	
Tra 2001	2088	1107	513	66	921	210521	1983	15401	33218	62643	
Ožu 2001	2056	939	397	68	972	149330	2121	12313	29111	63746	
Vel 2001	1770	902	387	64	920	128084	1812	10840	25273	49564	
Ukupno za sve						2109938	24607	129776	344642	668666	

Napomene u radu

- ne koristiti na **živim** logovima
- koristiti uvijek najnoviju inačicu:
 - redirect **buffer overflow!**
- Webalizer rasipa **memoriju**
 - veći logovi = veći load, itd.
- **hitovi** na Webalizer statistiku utiču na samu statistiku?
- osnovna konfiguracija često nedovoljna!



Webalizer (2)

razumijevanje ulaznih i
izlaznih podataka,
problemi u radu i naprednije
podešavanje

Ulazni logovi (1)

- podržani tipovi logova:
 - CLF logovi
 - Combined log format (NCSA)
 - wu-ftpd xferlog
 - squid proxy logovi
 - komprimirani (.gz) logovi
- način procesiranja:
 - iz datoteke (LogFile)
 - iz stdina (webalizer -opcije -)

Ulazni logovi (2)

- Apache logovi:
 - zasebna linija za svaki zahtjev
 - u liniji su zapisi odvojeni razmacima
 - Apache dokumentacija: [mod_log_config.html](#)
 - **CLF**: "%h %l %u %t \"%r\" %>s %b"
 - **CLF sa Virtual Host**: "%v %h %l %u %t \"%r\" %>s %b"
 - **NCSA extended/combined log format**:
"%h %l %u %t \"%r\" %>s %b \"%{Referer}i\"
\"%{User-agent}i\""

Ulazni logovi (3)

- Apache 1.2:
 - `LogFormat "%h %l %u %t \"%r\" %s %b \"%{Referer}i\" \"%{User-agent}i\""`
- Apache 1.3:
 - `CustomLog /var/lib/httpd/logs/access_log combined`
- potrebno za ispravan prikaz referiranja (Referers) i **UA** (User Agents)

Problemi sa logovima

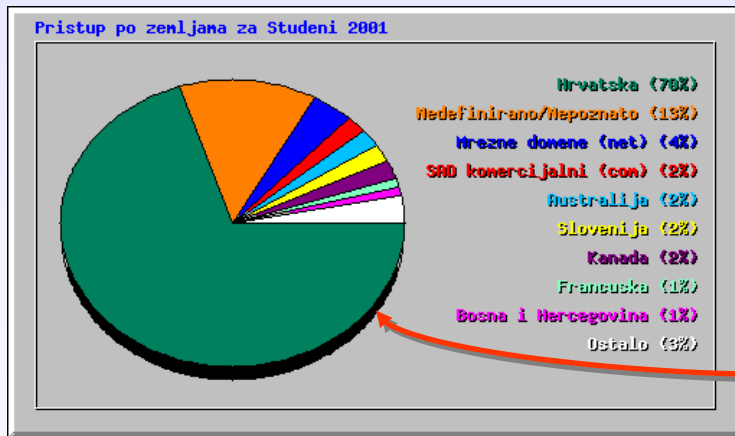
- procesiranje "živih" logova:
 - opasno - istovremeno pisanje i čitanje iz toka
 - netočno - da li privremeno ugasi Apache?
- **rotiranje logova** - prebrzo?
- namjerni prekid rada Webalizera
- **opetovano** pokretanje:
 - dupliciranje podataka
 - netočni rezultati
- rješenje: **timestamping!**

DNS resolve (1)

- Apache ne mora raditi resolve IP adresa:
 - libc DNS pozivi = spori, **blokirajući**/sinkroni, UDP
 - rješenje: async DNS, multithread, fork, zasebni resolve izvan Apachea
- webalizer = webazolver:
 - oprez: **ne** raditi resolve na **tekućim** logovima!
 - DNSCache `dns_cache.db`
 - DNSChildren `20`

DNS resolve (2)

- dva načina rada webalizer **resolviranja**:
 - tijekom normalnog procesiranja logova
 - zasebno od procesiranja logova
- oprez: **preveliki** broj DNS klijenata = DoS!



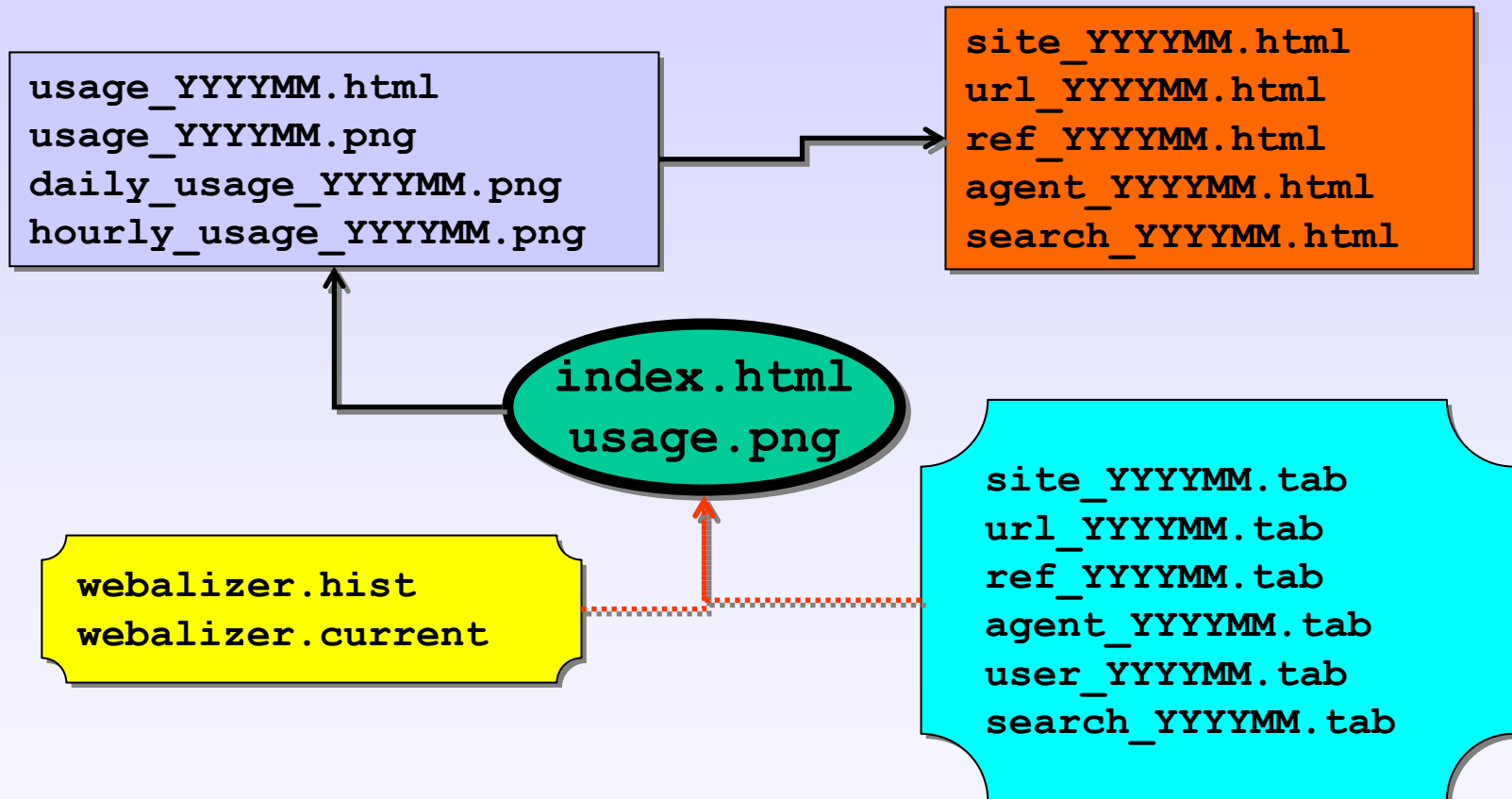
Prvih 30 od 47 zemalja

#	Pristupa	Datoteka	KB	Zemlja			
1	49476	69.78%	24178	65.56%	130558	60.74%	Hrvatska
2	9342	13.18%	5412	14.68%	30016	13.96%	Nedefinirano/Nepoznato
3	2921	4.12%	1921	5.21%	13158	6.12%	Mrezne domene (net)
4	1670	2.36%	993	2.69%	12457	5.80%	SAD komercijalni (com)
5	1211	1.71%	235	0.64%	1842	0.86%	Australija
6	1206	1.70%	1129	3.06%	5235	2.44%	Slovenija
7	1196	1.69%	545	1.48%	3144	1.46%	Kanada
8	909	1.28%	263	0.71%	2131	0.99%	Francuska
9	578	0.82%	204	0.55%	1115	0.50%	Bosna i Hercegovina
10	336	0.47%	300	0.82%	2127	0.99%	Jugoslavija
11	278	0.39%	266	0.72%	1701	0.79%	Njemačka
12	238	0.34%	196	0.53%	1196	0.56%	Austrija
13	212	0.30%	190	0.52%	1104	0.51%	Belgija

Inkrementalno procesiranje

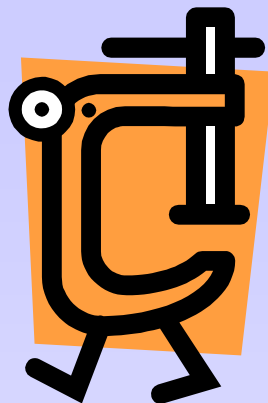
- veliki logovi = dugotrajno procesiranje = **CPU/memory hog**:
 - **razdijeliti** na više manjih
 - procesirati **odvojeno**
 - snimiti **interno stanje** Webalizera
 - nema **dupliciranja** podataka - drži se TS zadnjeg procesiranog zapisa - moguće ponovno procesiranje istog (npr. rotiranog) loga!
 - oprez: **ne** mijenjati konfiguraciju!

Izlazni podaci (struktura)



Izlazni podaci - kategorije

- **Pristupa (Hits)**
 - zahtjev upućen poslužitelju
 - cgi, grafika, html, multimedija
 - ulazni promet
- **Datoteka (Files)**
 - datoteke koje poslužitelj proslijedi
 - izlazni promet
- **Stranica (Pages)**
 - čisti html dokumenti posluženi klijentu
- **Kompjuteri (Sites)**
 - svaki jedinstveni IP sa kojeg je došao zahtjev
- **Posjeta (Visits)**
 - "novi" dolazak na stranice (visit timeout)
- **KB (KBytes)**
 - ...



Webalizer (3)

detaljno konfiguriranje,
napredne opcije i parametri,
preuređivanje izlaza

Konfiguracija detaljno (1)

- **LogFile** - apsolutno ili relativno ime, ako nije definirano = **stdin**
- **LogType** = clf, ftp, squid; oprez: "invalid record"
- **OutputDir** = apsolutna staza ili radni direktorij
- **HistoryName** - **relativno** prema OutputDir
- **ReportTitle** = "Usage Statistics for"
- **HostName** = ime poslužitelja, inače **uname** ili **localhost**

Konfiguracija detaljno (2)

- `UseHTTPS` = "https://", rijetko se koristi
- `Quiet` - ugasi informacije, greške i dalje na **stderr**
- `TimeMe` - vremenske statistike (Quiet!)
- `GMTTime` - GMT (UTC) umjesto lokalnog vremena
- `Debug`
- `IgnoreHist` - ignoriraj prošle mjesece, **oprez!**

Konfiguracija detaljno (3)

- `FoldSeqErr` - ignorira vremenske greške u logovima
- `VisitTimeout` - vrijeme da se **visit** pretvori u **hit**, standardno = 30 minuta
- `PageType` = **htm***, **cgi** (web), **txt** (ftp)
- `GraphLegend`, `GraphLines`, `CountryGraph`, `DailyGraph`, `DailyStats`, `HourlyGraph`, `HourlyStats` - grafički rezultati

Konfiguracija detaljno (4)

- **IndexAlias** - pretvaranje /staza/index.html u /staza i za druge dokumente (npr. home.htm)
- **MangleAgents** - 6 nivoa baratanja sa UA stringom, 0 je nepromijenjeno
- **SearchEngine** - identifikacija pretraživača, itd.
- **Incremental** - inkrem. procesiranje, oprez!
- **IncrementalName** - inače webalizer.current u OutputDir

Konfiguracija detaljno (5)

- **DNSCache** - dns_cache.db, SleepyCat DB
- **DNSChildren** - 20ak fork(), blocking resolve
- dodatne kontrolne riječi:
 - TopAgents, AllAgents, TopCountries, TopReferrers, AllReferrers, TopSites, TopKSites, AllSites, TopURLs, TopKURLs, AllURLs, TopEntry, TopExit, TopSearch, TopUsers, AllUsers, AllSearchStr

Parametri pri izvršavanju (1)

- većina parametara ima ekvivalente u konfiguracijskoj datoteci
- **-d** = Debug
- **-F** = LogType ftp (wu-ftpd)
- **-f** = FoldSeqErr
- **-i** = IgnoreHist (oprez!)
- **-p** = Incremental
- **-q** ili **-Q** = Quiet/ReallyQuiet

Parametri pri izvršavanju (2)

- -T = TimeMe
- -c datoteka
- -n imehosta = HostName
- -o direktorij = OutputDir
- -x ime_ekstenzije = HTMLExtension
- -P ekstenzije = PageType
- itd.

Upravljanje izlaznim podacima

- četiri tipa modificiranja sadržaja:
 - **skrivanje** određenih sadržaja - ne vidi se u Top, svejedno utiču na prosjeke
 - **grupiranje** sadržaja - vide se kao grupe, pojedinačni utiču na prosjeke
 - **ignoriranje** sadržaja - nema ih nigdje
 - **forsiranje** sadržaja - ne poštuje se ignoriranje
- previše Include*, Ignore*, Group* = veliko usporenje

Sakrivanje sadržaja

- skriva se objekt (UA, referer, ulazno računalo, URL, korisnici) od prikaza u "Top" tablicama:
 - [HideAgent](#) - beskorisno!
 - [HideReferrer](#) - obično vlastiti web
 - [HideSite](#) - vlastiti/lokalna računala, itd.
 - [HideAllSites](#) - sakriva sve negrupirane rezultate
 - [HideURL](#) - obično za sakrivanje svih ne-HTML sadržaja
 - [HideUser](#) - samo uz http autentifikaciju

Grupiranje sadržaja

- grupiraju se objekti, najčešće se koristi uz Hide* parametre:
 - GroupReferrer - obično za pretraživače
 - GroupURL - za stabla direktorija
 - GroupSite - TLD, dialupovi, itd.
 - GroupAgent
 - GroupDomains - xxx.xxx.xxx = 2, xxx.xxx = 1
 - GroupUser - za grupe korisnika
 - GroupShading, GroupHighlight - drukčiji prikaz

Izbacivanje sadržaja (1)

- za kompletno **izbacivanje** sadržaja/objekata iz **svih** statistika:
 - IgnoreSite - npr. lokalni
 - IgnoreURL - korišteno za "privremene" direktorije
 - IgnoreReferrer
 - IgnoreAgent
 - IgnoreUser
 - IncludeSite - forsira se procesiranje (bez obzira ako paše u IgnoreSite)

Izbacivanje sadržaja (2)

- nadalje analogno:
 - IncludeURL
 - IncludeReferrer
 - IncludeAgent
 - IncludeUser

Primjeri (Hide, Include, Group)

- "prvi/drugi/treći": "prv*", "drugi" ili "*treći"
- limit od **80** znakova!

- GroupURL /help/*
HideURL /help/*
- IgnoreAgent MSIE*
- IgnoreSite *.srk.fer.hr
IncludeSite gnjilux.srk.fer.hr

Spremanje u bazu podataka

- eksportiranje u baze podataka - polja u čistom tekstu, tab delimiter:
 - `DumpPath` - apsolutna ili relativna staza za db
 - `DumpExtension` - sufiks db
 - `DumpHeader` - dodati zaglavlje u db
 - `DumpSites`, `DumpURLs`, `DumpReferrers`,
`DumpAgents`, `DumpUsers`, `DumpSearchStr`
- lak unos u bilo koju stvarnu bazu

HTML oblikovanje izlaza

- dodavanje HTML koda u dijelove weba:
 - **HTMLExtension** - obično "html" (bez točke!)
 - **HTMLPre** - na vrh datoteke (HTML 3.2 DOCTYPE)
 - **HTMLHead** - kod između <HEAD></HEAD> (Javascript, php)
 - **HTMLBody** - kod nakon <HEAD> (logo, boje)
 - **HTMLPost** - prije <HR> (logo)
 - **HTMLTail** - dno desne strane, u <TABLE>
 - **HTMLEnd** - obično </BODY></HTML>

Primjer procesiranja

```
#!/bin/sh
kill `cat /var/lib/httpd/logs/httpd.pid`
ACCESS_LOG=/var/lib/httpd/logs/old/access_log.`date +%y%m%d-%H%M%S`
ERROR_LOG=/var/lib/httpd/logs/old/error_log.`date +%y%m%d-%H%M%S`
cp /var/lib/httpd/logs/access_log
  /var/lib/httpd/logs/access_log.backup
mv /var/lib/httpd/logs/access_log $ACCESS_LOG
mv /var/lib/httpd/logs/error_log $ERROR_LOG
/usr/sbin/httpd
/bin/gzip $OLD_ACCESS_LOG
/bin/gzip $OLD_ERROR_LOG
/usr/bin/webalizer -Q /var/lib/httpd/logs/access_log.backup
```

Literatura

- **SleepyCat DB** - <http://www.sleepycat.com/>
- **Bind** - <http://www.isc.org/>
- **Apache** - <http://www.apache.org/>
- **Squid** - <http://www.squid-cache.org/>
- **Wu-Ftpd** - <http://www.wu-ftp.org/>
- **Webalizer** - www.mrunix.net/webalizer/
- **Webalizer Win32** - <http://www.medasys-lille.com/webalizer/>