



# Web Günlük Analizi

Mehmet ULUER  
muluer (at) ford (dot) com (dot) tr  
06.10.2004

# İçerik

- Motivasyon
- Analiz Öncesi
  - Günlük Kütüğü Nedir?
  - Neden Log Analizi?
  - Veri Kaynakları
  - Bilgi Keşif Süreci
- Analiz
  - Önışleme
  - Analiz Yazılımları
    - Analog
    - AWStats
    - Webalizer
  - WUM
- Analiz Sonrası

# Motivasyon I

- Çok hızlı gelişen teknoloji!!!
- Her geçen gün pek çok kurum ve kuruluş, gerçek hayatta yaptıkları işlerini sanal dünyaya taşıyarak, güvenli(!) bir şekilde işlemeye başlıyor.
- Olabildiğince büyük ölçüde, günlük hatta anlık hareket verilerini toplamaya ve saklamaya çalışıyorlar.

# Motivasyon II

- Bu veriler genelde web sunucuları tarafından otomatik olarak toplanıyor ve sunucuya erişim kütüklerinde biriktiriliyor.
- Üretilen bu logların analizini yapmak, bize çok değerli bilgiler kazandırabilir.
  - ✓ örn. reklamın hedef kitleye ulaşması.

# Log(Günlük) Kütüğü Nedir?

- Meydana gelen ~tüm hareketlerin tutulduğu dosyalardır.
- Web sunucusuna gelen her “hit” (.html, .png, .php vb.) log kütüğünde bir kayıt oluşturur.

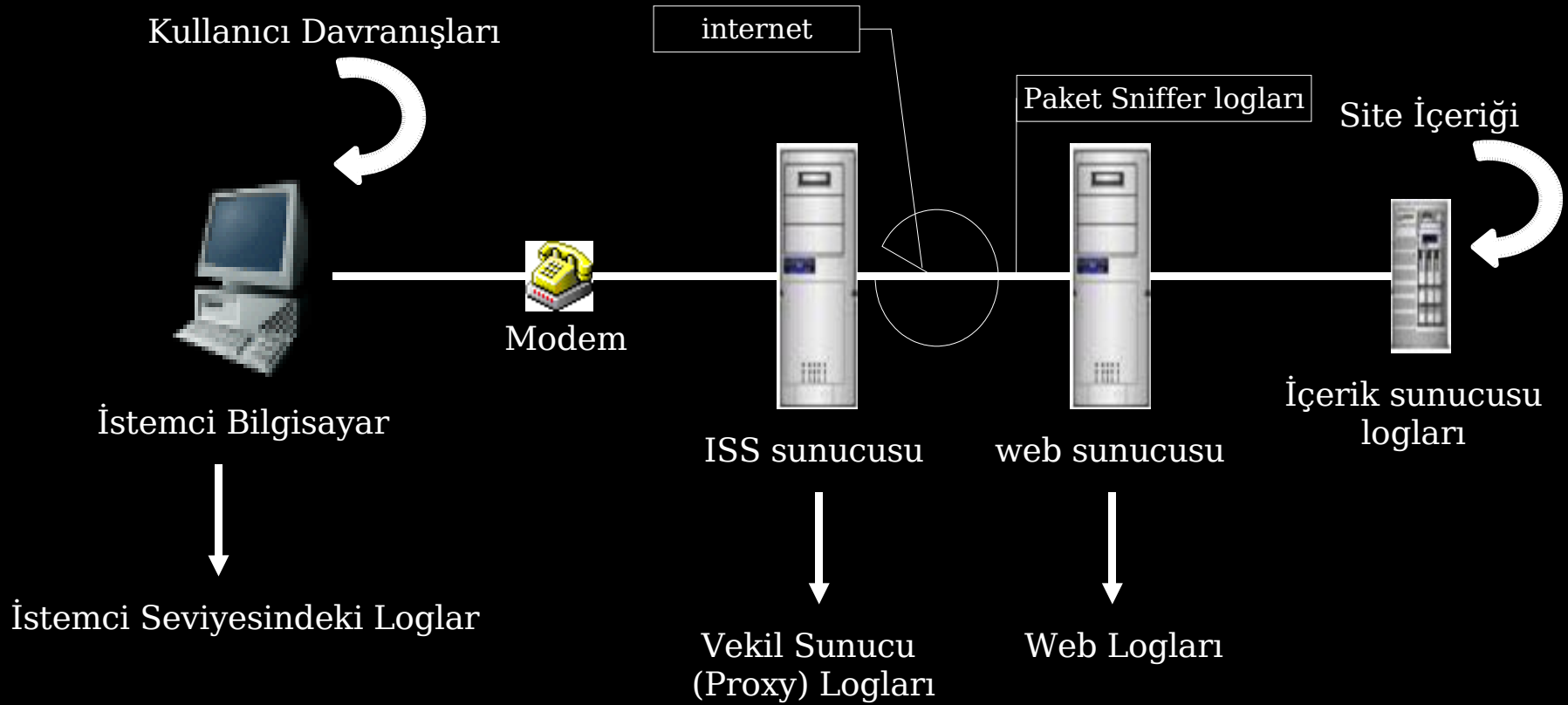
# Neden Log Analizi

- Siteyi bugün kaç kişi ziyaret etti?
- Ziyaretçiler siteyi ne kadar sevdiler?
- En çok ziyaret edilen sayfa hangisi?
- Ziyaretçi siteden neden ayrıldı?
- Siteye kimler link vermiş?
- En çok hangi sayfadan sonra site terk edilmiş?
- Siteyi Internette bulabilmek için hangi anahtar kelimeler kullanılmış?

# WEB'deki Veri Sınıfları

- x İçerik : Gerçek hayattaki veriyi tanımlayan veriler. Html, resim, ses dosyaları vb.
- x Yapı : İçerigin organizasyonu ( linkler )
- x Kullanıcı Profili : Üyelik aşamasında çıkarılan demografik veriler.
- ✓ Kullanım : Web sayfalarının kullanım örüntülerini tanımlayan veriler. IP adresleri, referans sayfaları vb.

# Veri Kaynakları



# Günlük Tipi

- Erişim (Access)
- \_Referans (Referrer)
- Hata (Error)
- CLF
  - common log format. "%h %l %u %t \"%r\" %>s %b", Hostname, Identity, Userid, Time, Request, Status\_code, Bytes.
- ECLF
  - extended/combined common log format. clf + "%Referer \ %Agent\".

# Sunucu Erişim Logları

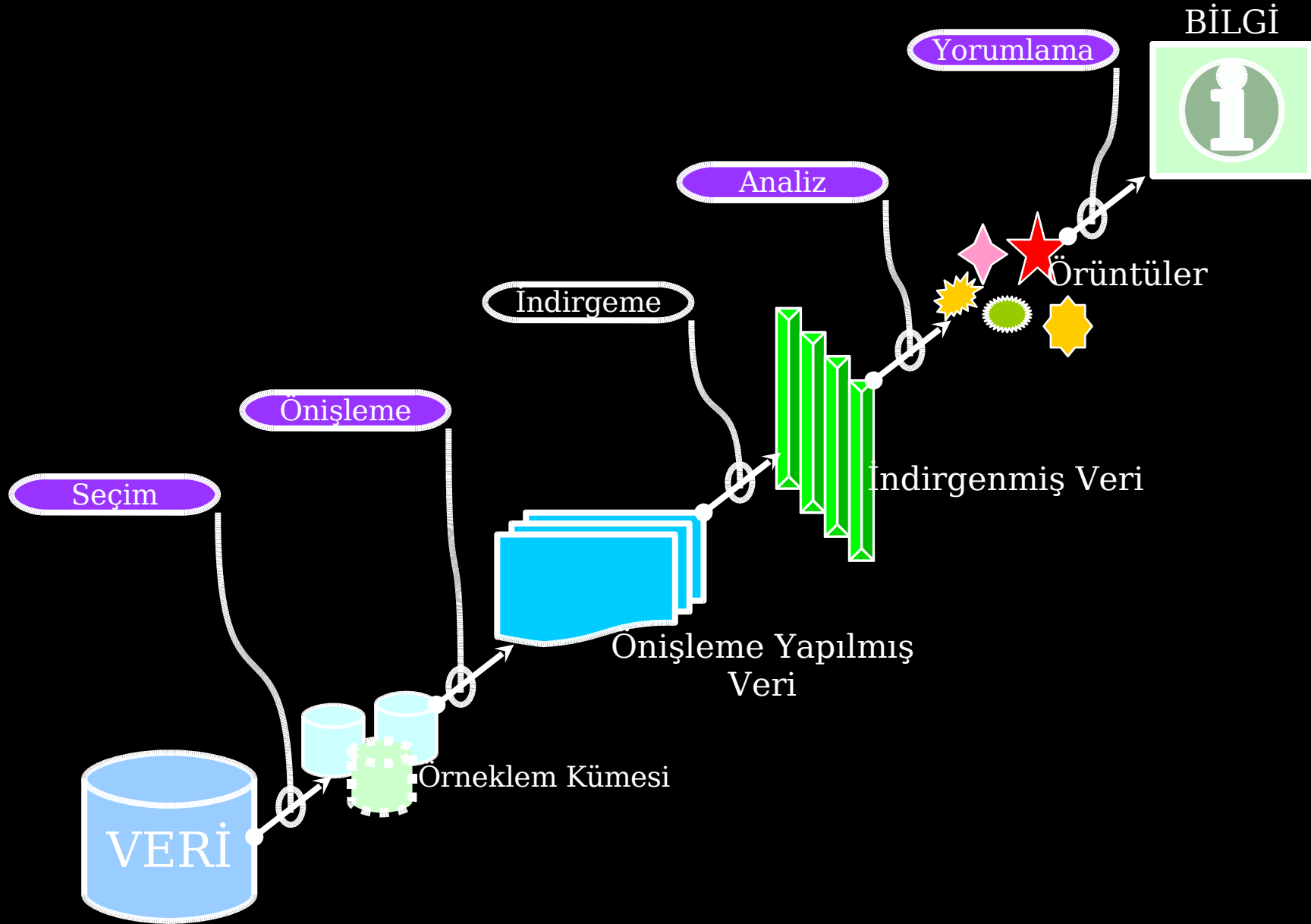
193.140.162.154	www.uluer.com -	[28/Feb/2002:16:26:50 -0500]	GET /home1.html HTTP/1.1	200	6693
213.243.30.5	www.uluer.com -	[08/Mar/2002:20:45:19 -0500]	GET /interests.html HTTP/1.1	200	7647
195.175.137.169	www.uluer.com -	[12/Mar/2002:20:35:15 -0500]	GET /time_table.html HTTP/1.1	200	178
193.140.164.197	www.uluer.com -	[13/Mar/2002:09:45:37 -0500]	GET /courses.html HTTP/1.1	200	5436

193.140.161.67	-	[07/Apr/2003:17:08:44 +0300]	"GET /~muluer/educational/101/index.html HTTP/1.1"	200	14429	http://www.cs.baskent.edu.tr/~muluer/	Mozilla/4.0(compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.0.2914)
193.140.161.67	-	[07/Apr/2003:17:09:23 +0300]	"GET /~muluer/educational/101/101-lab1.txt HTTP/1.1"	304	0	http://www.cs.baskent.edu.tr/~muluer/educational/101/index.html	Mozilla/4.0(compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.0.2914)
193.140.161.67	-	[07/Apr/2003:17:09:27 +0300]	"GET /~muluer/educational/101/101-lab2.txt HTTP/1.1"	304	0	http://www.cs.baskent.edu.tr/~muluer/educational/101/index.html	Mozilla/4.0(compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.0.2914)
193.140.161.67	-	[07/Apr/2003:17:09:31 +0300]	"GET /~muluer/educational/101/101-lab3.txt HTTP/1.1"	304	0	http://www.cs.baskent.edu.tr/~muluer/educational/101/index.html	Mozilla/4.0(compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.0.2914)
193.140.161.67	-	[07/Apr/2003:17:09:38 +0300]	"GET /~muluer/educational/101/101-lab4.txt HTTP/1.1"	304	0	http://www.cs.baskent.edu.tr/~muluer/educational/101/index.html	Mozilla/4.0(compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.0.2914)
193.140.161.67	-	[07/Apr/2003:17:09:45 +0300]	"GET /~muluer/educational/101/101-lab5.txt HTTP/1.1"	200	461	http://www.cs.baskent.edu.tr/~muluer/educational/101/index.html	Mozilla/4.0(compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.0.2914)
193.140.161.67	-	[07/Apr/2003:17:09:54 +0300]	"GET /~muluer/educational/101/not.html HTTP/1.1"	200	35355	http://www.cs.baskent.edu.tr/~muluer/educational/101/index.html	Mozilla/4.0(compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.0.2914)
193.140.161.67	-	[07/Apr/2003:17:10:14 +0300]	"GET /~muluer/educational/101/blist.html HTTP/1.1"	200	7818	http://www.cs.baskent.edu.tr/~muluer/educational/101/index.html	Mozilla/4.0(compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.0.2914)
193.140.161.67	-	[07/Apr/2003:17:10:52 +0300]	"GET /~muluer/educational/101/sy.html HTTP/1.1"	404	1287	http://www.cs.baskent.edu.tr/~muluer/educational/101/index.html	Mozilla/4.0(compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.0.2914)
193.140.161.67	-	[07/Apr/2003:17:10:54 +0300]	"GET /~muluer/educational/101/index.html HTTP/1.1"	304	0	http://www.cs.baskent.edu.tr/~muluer/	Mozilla/4.0(compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.0.2914)

# Sunucudaki tipik bir kayıt

<u>Ip Adresi</u>	→	• 64.12.51.xxx
<u>Sunucu Adı</u>	→	• www.uluer.com
<u>Kullanıcı Adı</u>	→	• -
<u>Zaman</u>	→	• [05/Nov/2003:14:53:48 +0300]
<u>İstek tipi</u>	→	• GET /
<u>Url</u>	→	• http://www.uluer.com
<u>Protokol</u>	→	• HTTP/1.1"
<u>Statüsü</u>	→	• 200-Başarılı, 300-Yönlendir 400-Başarısız, 500-İç Hata
<u>Boyut</u>	→	• 2270
<u>Referansı</u>	→	• http://www.cs.baskent.edu.tr/ ~muluer
<u>Araçları</u>	→	• Mozilla/5.0 (Linux 2.4.18- 27.8.0 i686) Opera 6.12 [en]

# Bilgi Keşif Süreci



# Önişleme

- Veri temizleme: Hata, tutarsızlık, tekrar ve eksik verilerin ayıklanması/düzeltilmesi
- Hareketlerin (transactions) teşhis edilmesi; kullanıcı oturumlarının bulunması
  - İçerik verisi (bağıntı keşfi) zaman
  - Gezinti-İçerik verisi (yol analizi)

# Ön işlemedeki zorluklar I

+ Ham verinin soyut veri tiplerine dönüştürülmesi ( ham veri çok büyük!!! )

+ 1 IP adresi, >1 kullanıcı

Aynı zaman diliminde pek çok kullanıcı bir proksi sunucusu üzerinden işlem yapabilir

+ 1 Kullanıcı, >1 IP adresi

Aynı kullanıcı farklı adreslerden işlem yapabilir

# Ön işlemedeki zorluklar II

+ 1 Sunucu oturumu, >1 IP adresi

Bazı ISS ler veya güvenlik sunucuları istek tipine göre aynı oturum içinde farklı IP ler atayabilirler

+ 1 Kullanıcı, >1 Araç

Aynı kullanıcı aynı anda birden fazla tarayıcı kullanabilir

# Çözüm Önerileri

- + Çerezler: İstemci tarafında tutulan küçük kod parçaları
  - ++ Aynı kullanıcıyı birden fazla oturum için takip edebilir
  - Güvenlik gerekçesi ile kabul edilmeyebilir veya silinebilir
- + Login / Logout : Kullanıcı adı ve parola kullanarak siteye erişimin sağlanması
  - ++ Makineden ve tarayıcıdan ziyade her kullanıcının farklı bir kimlik bilgisinin olması
  - Her kullanıcı kayıt yaptırmak istemeyebilir
- + Zamana göre oturum sonlandırma...

# Oturum Bilgisinin Tamamlanması

	IP	Zaman	URL	Referans	Araç
1	www.uluer.com	08:30:00	A	#	Mozilla
2	www.uluer.com	08:30:01	B	E	Mozilla
3	www.uluer.com	08:30:02	C	B	Mozilla
4	www.uluer.com	08:30:01	B	#	Mozilla
5	www.uluer.com	08:30:03	C	B	Mozilla
6	www.uluer.com	08:30:04	F	#	Mozilla
7	www.uluer.com	08:30:04	B	A	Mozilla
8	www.uluer.com	08:30:05	G	B	Mozilla

Tanımlanan Oturumlar:

$S_1: \# \implies A \implies B \implies G$  1, 7, 8

$S_2: E \implies B \implies C$  2, 3

$S_3: \# \implies B \implies C$  4, 5

$S_4: \# \implies F$  6

# Analiz Yazılımları

- × Ticari yazılım kullan
- × Kendi kodunu kendin yaz
- ✓ Açık kaynak kodlu yazılım kullan

# httpd.conf

- HostnameLookups Off
- ErrorLog /usr/local/apache/logs/error\_log
- LogLevel warn
  - debug, info, notice, warn, error, crit, alert, emerg
- LogFormat "%h %l %u %t \"%r\" %>s %b \"%{Referer}i\" \"%{User-Agent}i\"" combined
- LogFormat "%h %l %u %t \"%r\" %>s %b" common
- LogFormat "%{Referer}i -> %U" referer
- LogFormat "%{User-agent}i" agent
- CustomLog /usr/local/apache/logs/access\_log common
- #CustomLog /usr/local/apache/logs/referer\_log referer
- #CustomLog /usr/local/apache/logs/agent\_log agent
- #CustomLog /usr/local/apache/logs/access\_log combined

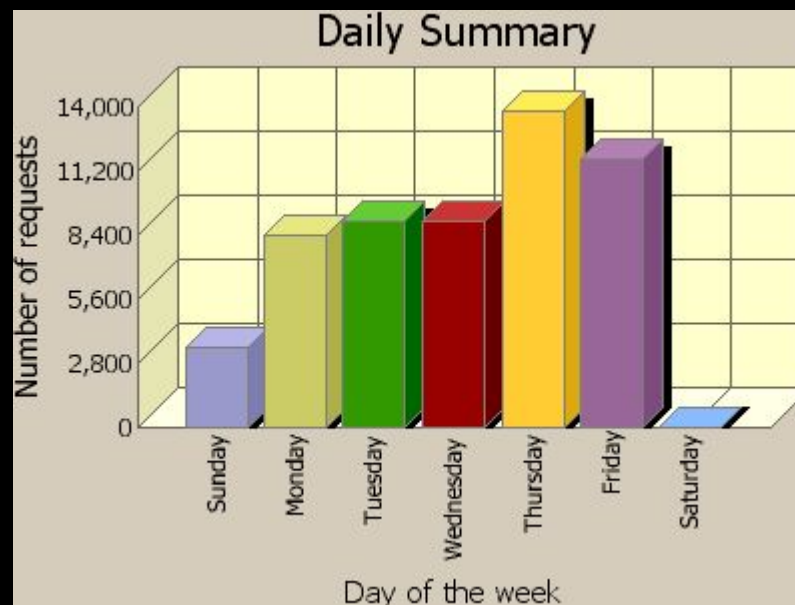
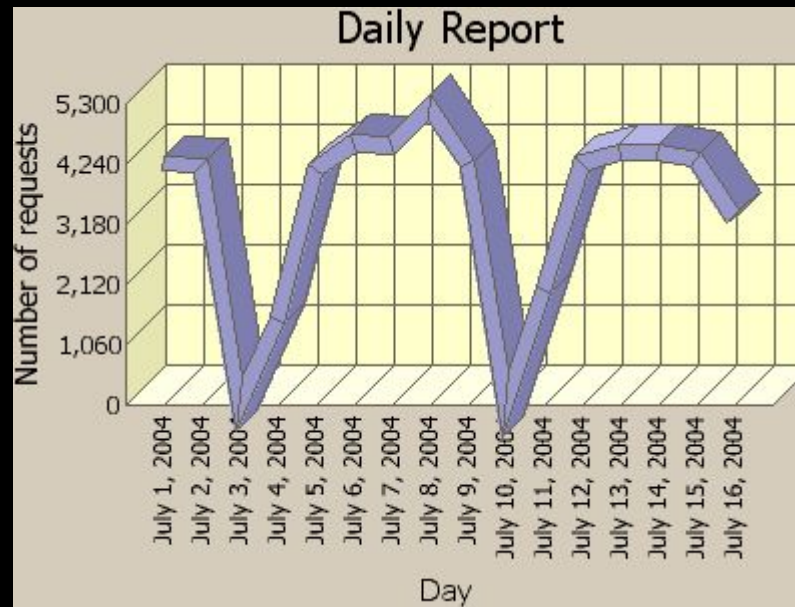
# Analog

- 5.32
- 23 Mart 2003
- 5dk -> 1GB (266 Mhz)
  - 56 milyon log kaydi(satir) 35 dk.
- 100GB
- 33 farkli dil desteđi (Türkçe)
- Ayar vermek kolay
- AWStats X %15 daha hızlı
- <http://www.analog.cx/>

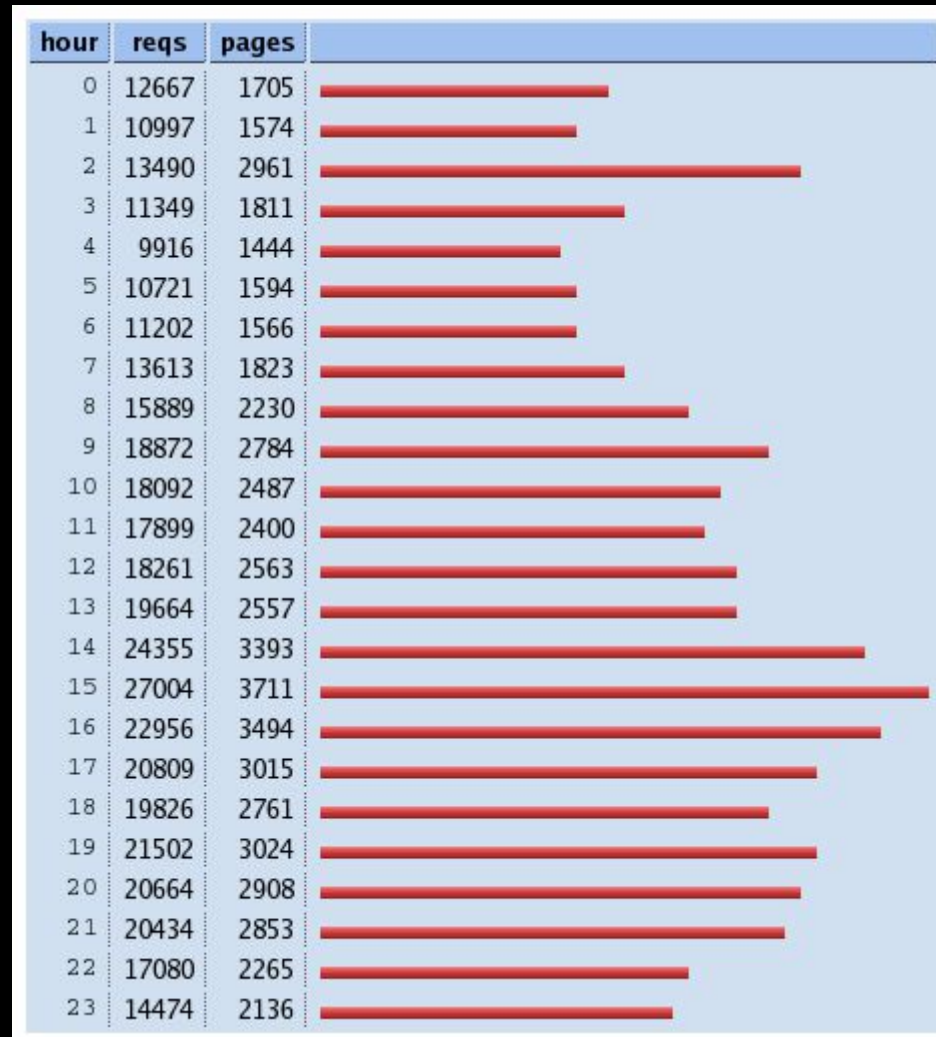
# analog.cfg

- LOGFILE logfilename
- OUTFILE outputfile.html
- HOSTNAME "LKD"
- HOSTURL http://www.lkd.org.tr/
- SEARCHQUERY ON
- LANGUAGE TURKISH
- LogFormat "%h %l %u %t %v \"%r\" %>s %b" myformat
- CustomLog /var/log/apache/access.log myformat
- CASE SENSITIVE
- FILEINCLUDE /cgi-bin/script.pl\*
- FILEALIAS /football.html /soccer.html
- HOSTALIAS tymith tymith.uluer.com
- REQLINKINCLUDE pages, \*.pdf
- SEARCHENGINE http://\*google.\*/\* q

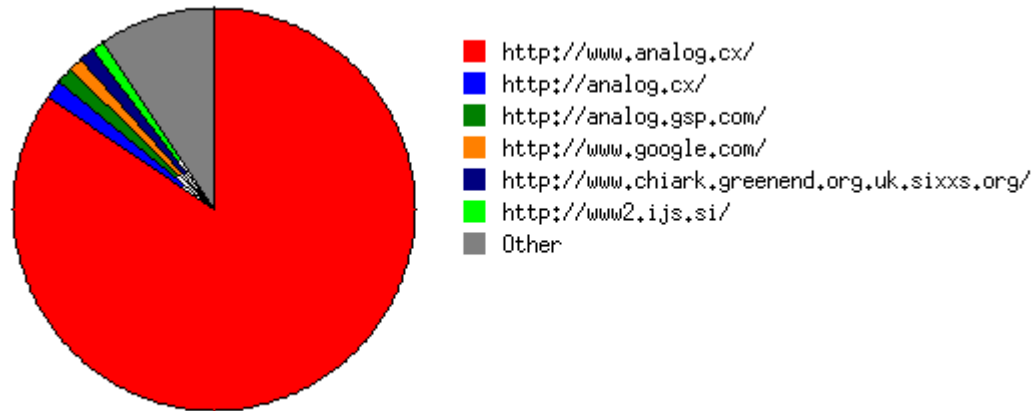
# Analog Raporları (gün)



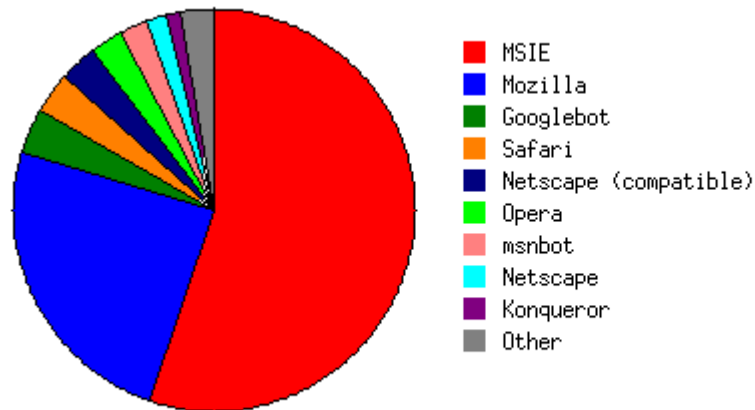
# Analog Raporları (saat)



# Analog Raporları (ref. & araç)

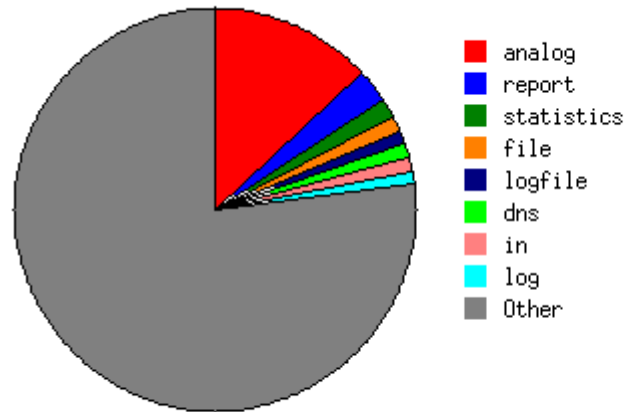


The wedges are plotted by the number of requests.



The wedges are plotted by the number of requests for pages.

# Analog Raporları (kelime & os)



The wedges are plotted by the number of requests.

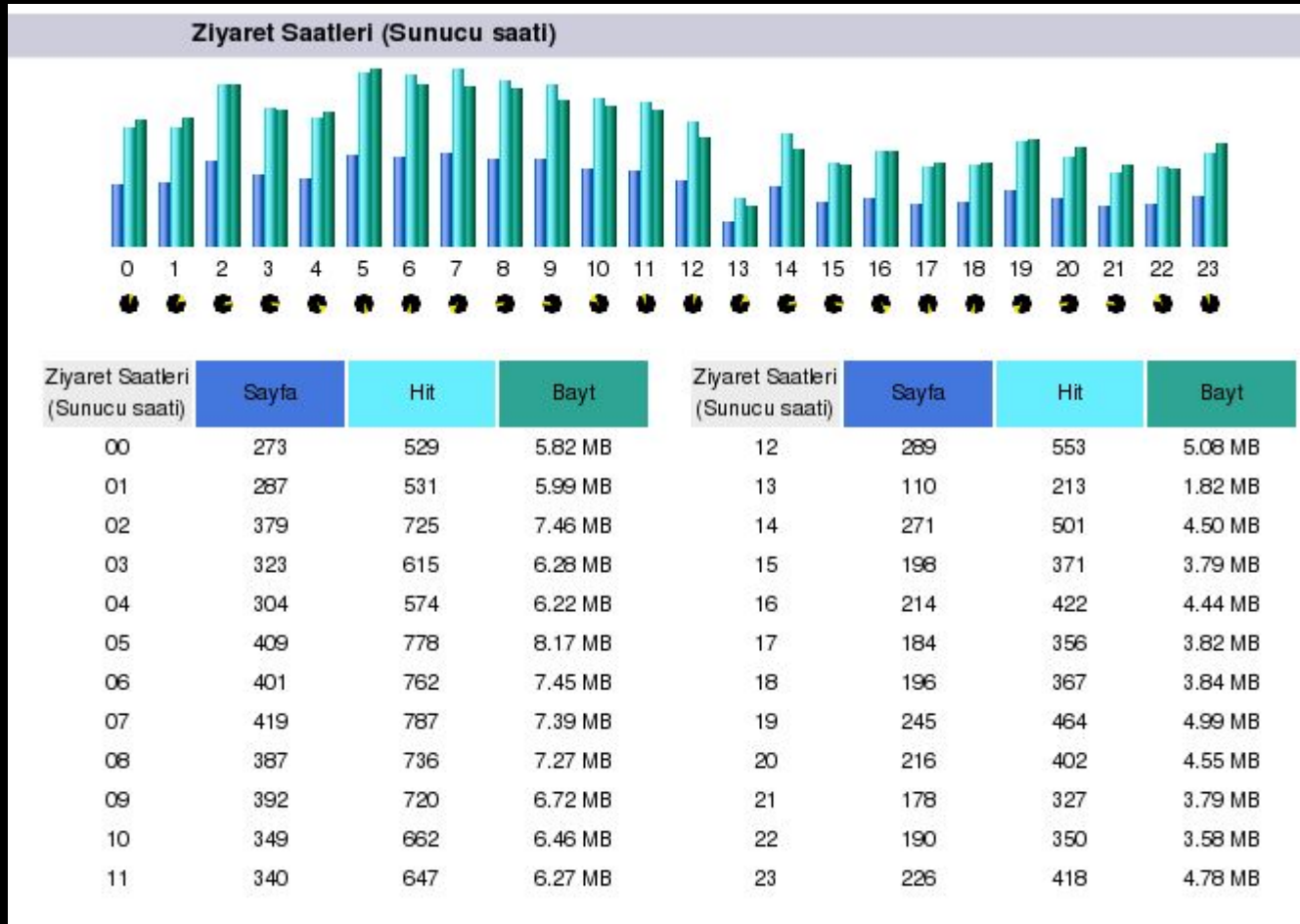


The wedges are plotted by the number of requests for pages.





















# AWStats

- analog++
- 6.2
- Haziran 2004
- Perl
- Web/FTP/e-posta
- Unique (benzersiz) ziyaretçi
- Session (oturum) uzunluğu
- Giriş/Çıkış sayfaları
- CGI sayfalarını tanıyabilme
- <http://awstats.sourceforge.net/>

# AWStats (saat)



# AWStats (ülkeler & unique ziyaretçi)

Countries (En sık kullanılan 10) - Tüm liste					
Countries		Sayfa	Hit	Bayt	
 United States	us	2276	4332	40.00 MB	
 European Union	eu	523	1002	8.88 MB	
 Germany	de	393	759	5.78 MB	
 Australia	au	377	698	8.82 MB	
 Japan	jp	259	449	5.41 MB	
 Netherlands	nl	205	384	4.07 MB	
 France	fr	204	405	3.40 MB	
 Canada	ca	184	349	3.35 MB	
 Great Britain	gb	182	344	3.60 MB	
 China	cn	178	322	5.23 MB	
Diğerleri		1999	3766	41.95 MB	

Hosts (En sık kullanılan 10) - Tüm liste - Son ziyaret - Çözümeyen IP Adresleri				
Hosts : 0 Known, 4715 Bilinmeyen (çözümeyen ip) - 4641 Ayrı Ziyaretçi	Sayfa	Hit	Bayt	Son ziyaret
82.151.203.129	14	26	0	04 Eki 2004 - 11:01
219.93.174.104	13	27	314.89 KB	04 Eki 2004 - 05:47
80.105.101.54	11	11	0	03 Eki 2004 - 14:10
217.57.199.13	11	16	346.22 KB	04 Eki 2004 - 04:13
80.58.1.174	11	21	165.21 KB	04 Eki 2004 - 10:36
203.173.27.244	11	20	353.13 KB	03 Eki 2004 - 18:34
198.60.99.240	10	17	259.63 KB	02 Eki 2004 - 20:42
68.4.254.88	10	18	314.74 KB	03 Eki 2004 - 16:21
81.176.1.140	10	13	0	04 Eki 2004 - 07:20
82.43.209.2	10	22	0	03 Eki 2004 - 18:03
Diğerleri	6669	12619	128.78 MB	

# AWStats (ekran & diđerleri)

Miscellaneous		
Miscellaneous		
Add to favorites (estimated)	0 / 4641 Ziyaretçiler	0 %
Javascript disabled	-	0 %
Browsers with Java support	-	89.4 %
Browsers with Macromedia Director Support	-	36.5 %
Browsers with Flash Support	-	91.3 %
Browsers with Real audio playing support	-	41.1 %
Browsers with Quictime audio playing support	-	48.5 %
Browsers with Windows Media audio playing support	-	85.6 %
Browsers with PDF support	-	74.9 %

Screen sizes (En sık kullanılan 5)	
Screen sizes	Yüzde
1024x768	48.7 %
1280x1024	22.1 %
800x600	10.5 %
1152x864	5.1 %
1600x1200	3.8 %
Diđerleri	9.5 %

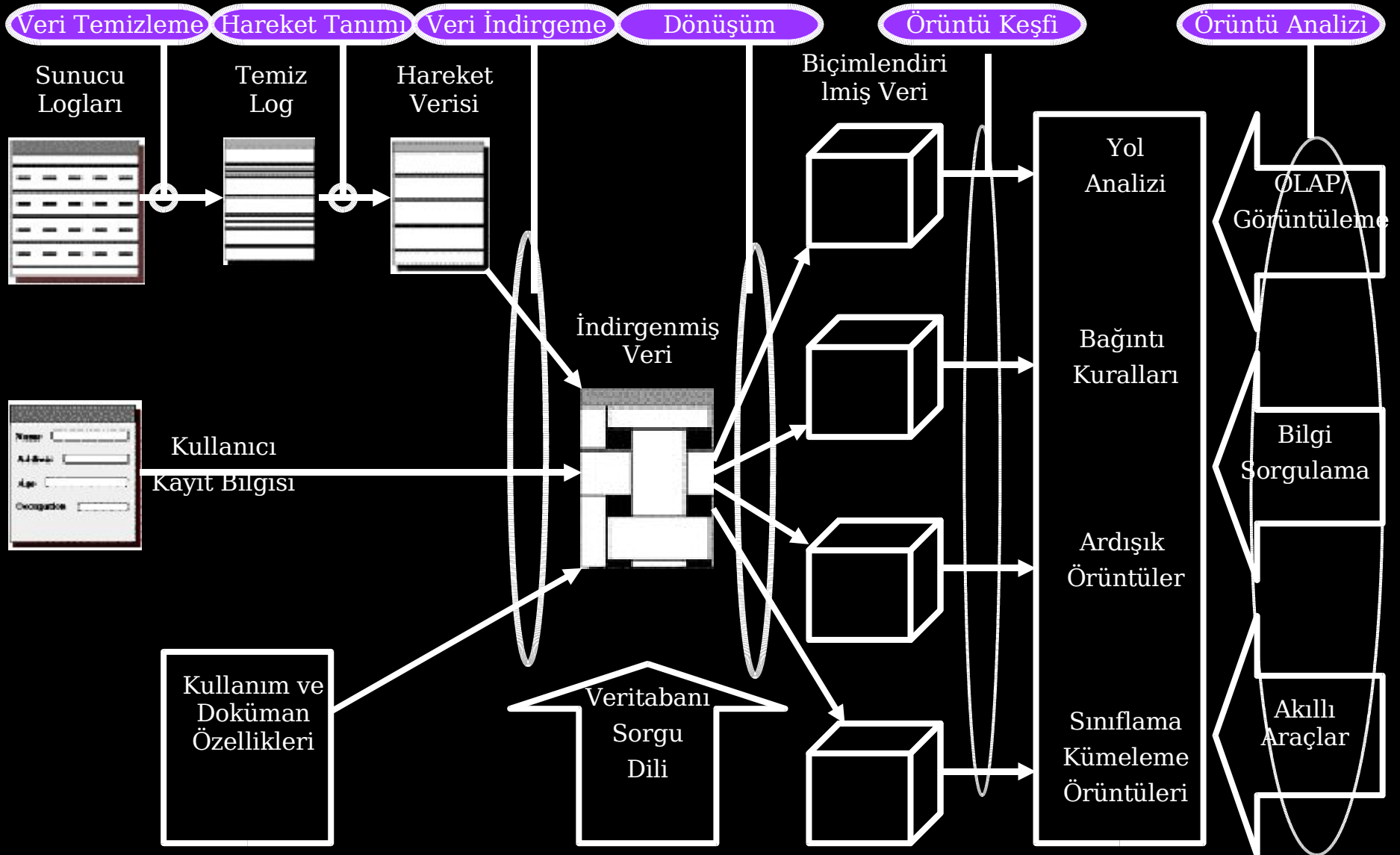
# AWStats (referanslar)

Siteye bağlantı yapanlar				
Köken	Sayfa	Yüzde	Hit	Yüzde
<b>Doğrudan adres / Yer imi</b>	1621	25.9 %	7651	62.3 %
<b>Links from a NewsGroup</b>				
<b>İnternet arama motorundan bağlantı - Tüm liste</b>	1238	19.8 %	1238	10 %
- Google	1123	1123		
- Yahoo	55	55		
- Baidu	23	23		
- MSN	12	12		
- Alexa	5	5		
- Lycos	4	4		
- Netscape	3	3		
- Unknown search engines	2	2		
- AllTheWeb	2	2		
- Excite	2	2		
- Diğerleri	7	7		
<b>Dış sayfalardan bağlantılar (arama motorları hariç diğer web siteleri) - Tüm liste</b>	3260	52.2 %	3260	26.5 %
- <a href="http://sourceforge.net/projects/awstats/">http://sourceforge.net/projects/awstats/</a>	117	117		
- <a href="http://www.9sky.com/aws/www/awstats.www.keyphrases.html">http://www.9sky.com/aws/www/awstats.www.keyphrases.html</a>	98	98		
- <a href="http://sarikata.com/cgi-bin/awstats.pl">http://sarikata.com/cgi-bin/awstats.pl</a>	61	61		
- <a href="http://www.postfix.org/addon.html">http://www.postfix.org/addon.html</a>	39	39		
- <a href="http://www.thefreecountry.com/webmaster/loganalyzers.shtml">http://www.thefreecountry.com/webmaster/loganalyzers.shtml</a>	38	38		
- <a href="http://www.9sky.com/aws/www/awstats.www.keywords.html">http://www.9sky.com/aws/www/awstats.www.keywords.html</a>	36	36		
- <a href="http://taiwan.cnet.com/computer/network/features/0,2000068598,20...">http://taiwan.cnet.com/computer/network/features/0,2000068598,20...</a>	36	36		
- <a href="http://jir.ns2go.com/cgi-bin/awstats/awstats.pl">http://jir.ns2go.com/cgi-bin/awstats/awstats.pl</a>	36	36		
- <a href="http://awstats.org">http://awstats.org</a>	32	32		
- <a href="http://sourceforge.net/forum/forum.php">http://sourceforge.net/forum/forum.php</a>	28	28		
- Diğerleri	2739	2739		
<b>Kökene bilinmeyen</b>	116	1.8 %	116	0.9 %

# Webalizer

- 2.01
- Nisan 2002
- C
- [http://awstats.sourceforge.net/docs/awstats\\_compare.html](http://awstats.sourceforge.net/docs/awstats_compare.html)
  - Athlon 1GHz
  - AWStats X 3 daha hızlı
- Temiz kaynak kodu
- <http://www.mrunix.net/webalizer/>

# Web Günlük Madenciliği Sistem Mimarisi



# WUM

- Web Utilization Miner
  - Java
  - Web Madenciliği
  - Sıralı Örüntü Analizi
  - MINT
  - WUMprep
    - Perl
    - Veri Önışleme
    - Oturum tanıma
    - Robot

# WUM (Yol Analizi)

- + /firma/ürünler/ürün2.html sayfasını ziyaret eden kullanıcıların %70'i, ilk önce /firma arkasından /firma/duyurular, sonra /firma/ürünler daha sonrada /firma /ürünler /ürün1.html yi ziyaret etmişler (zahmet oldu size de!)
- + Site ziyaretçilerinin 80%'i /firma/ürünler/ bağlantısından giriş yapmış!
- + Site ziyaretçilerinin 65%'i 4 veya daha az sayfa ziyaret etmişler!

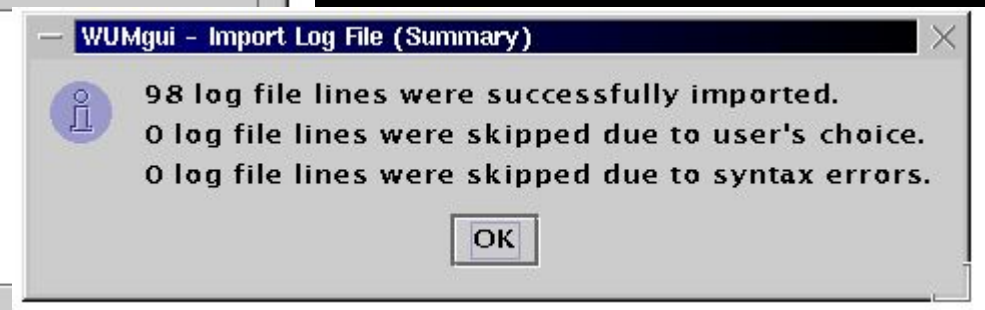
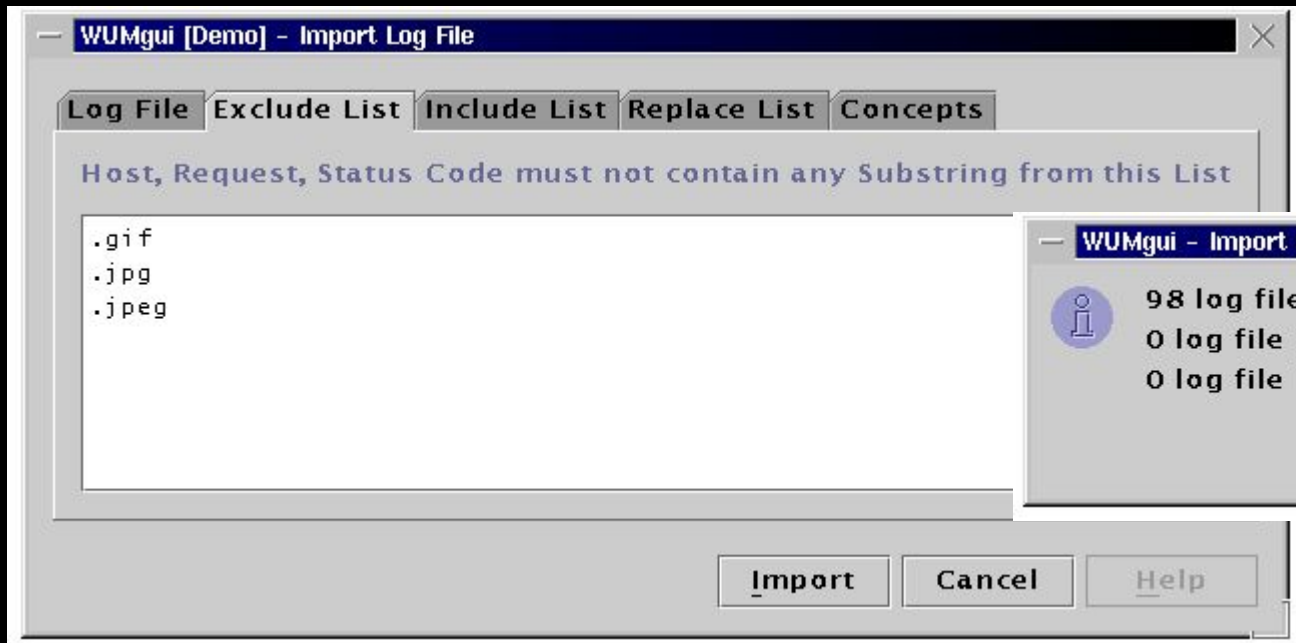
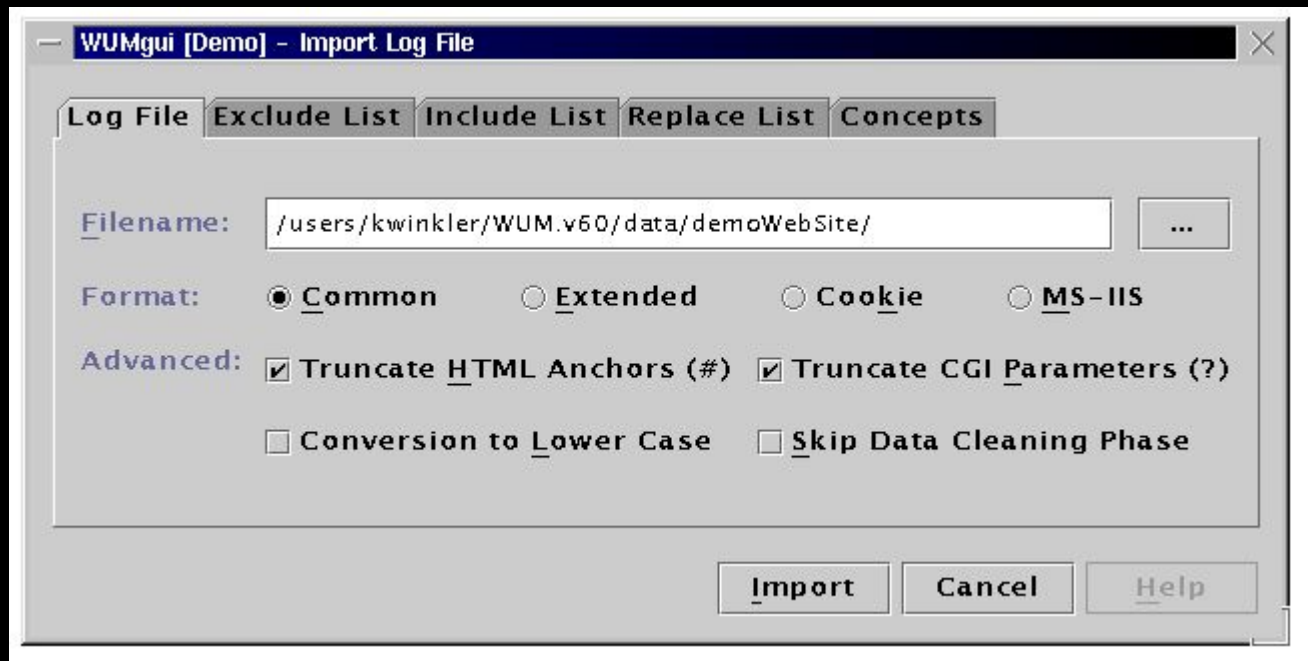
# WUM

The screenshot shows the WUMgui application window with a menu bar (File, Analyzer, Reports, Queries, Visualizer, Tools, Help) and a main display area. The display area contains a log file with several entries. A red box highlights a specific log entry: `141.20.140.224 - - [20/Apr/1998:09:58:44 +0200] "GET /iwi/le" [P.html;1,3] b = [X.html;1,2]`. Below this entry, a SQL query is displayed: `SELECT t FROM NODE AS a, b, TEMP WHERE a.url2 = "S.html" AND b.support > 3`. The application also displays a welcome message: "Welcome to WUM 6.0 Beta: A Web Utilization Miner" and the affiliation: "Humboldt-University Berlin, Germany Institute of Information Systems Leipzig Graduate School of Business".

The screenshot shows the "WUMgui - Create Mining Base" dialog box. It contains the following fields and buttons:

- Name:** Demo
- Web Server:** http://www.demo.org
- Directory:** /users/kwinkler/WUM.v60/data/demoWebSite (with a browse button "...")
- Log Files:** /users/kwinkler/WUM.v60/data/demoWebSite (with a browse button "...")
- Remarks:** small demo mining base |
- Buttons:** Create, Cancel, Help

# WUM



# WUM

WUMgui [Demo] - Create Visitors' Sessions

Criterion:  Maximal Session Duration  
 Maximal Page View Time

Threshold:

WUMgui - Create Visitors' Sessions

 22 sessions were successfully created.

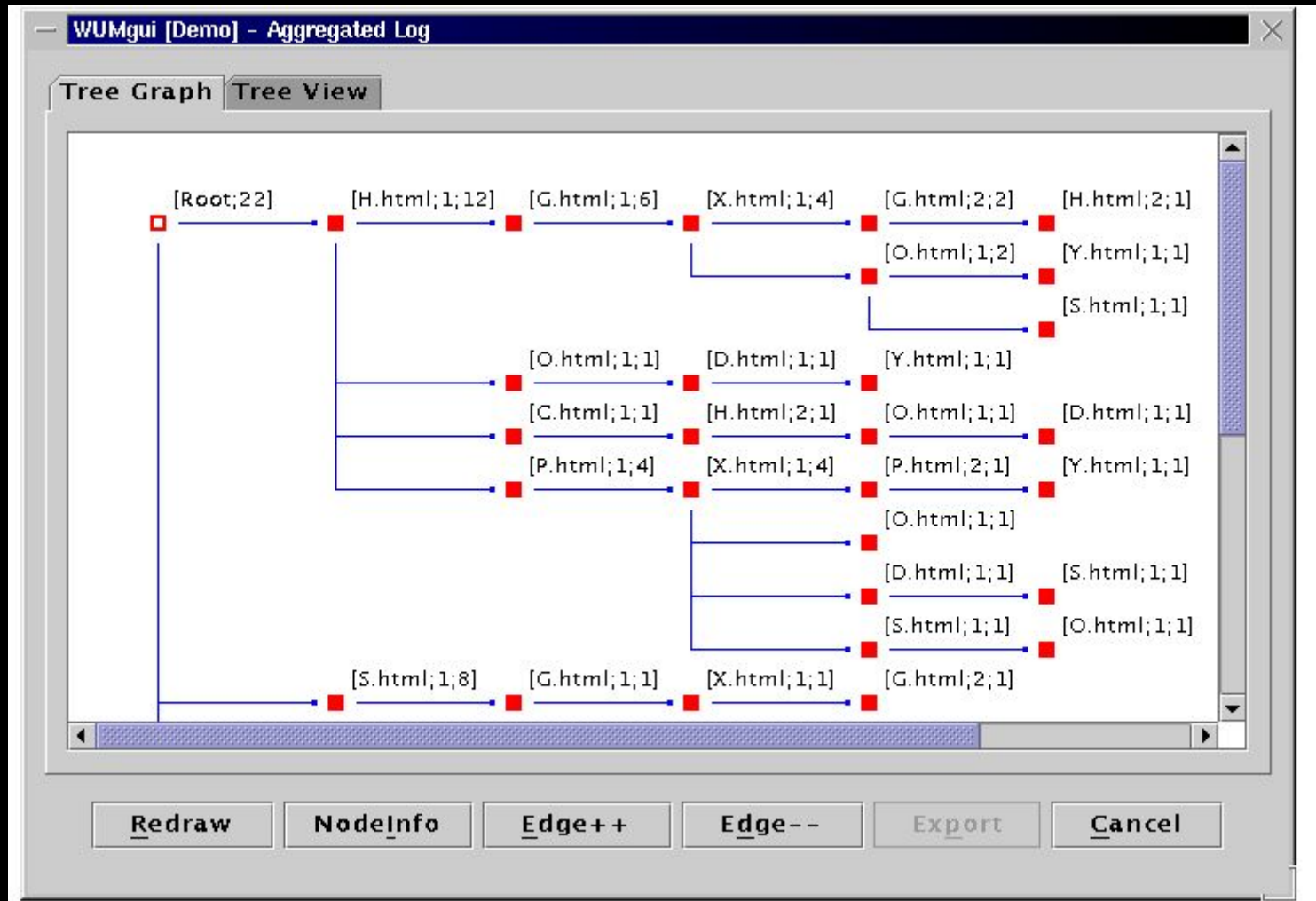
WUMgui [Demo] - Comprehensive Summary

Filename:

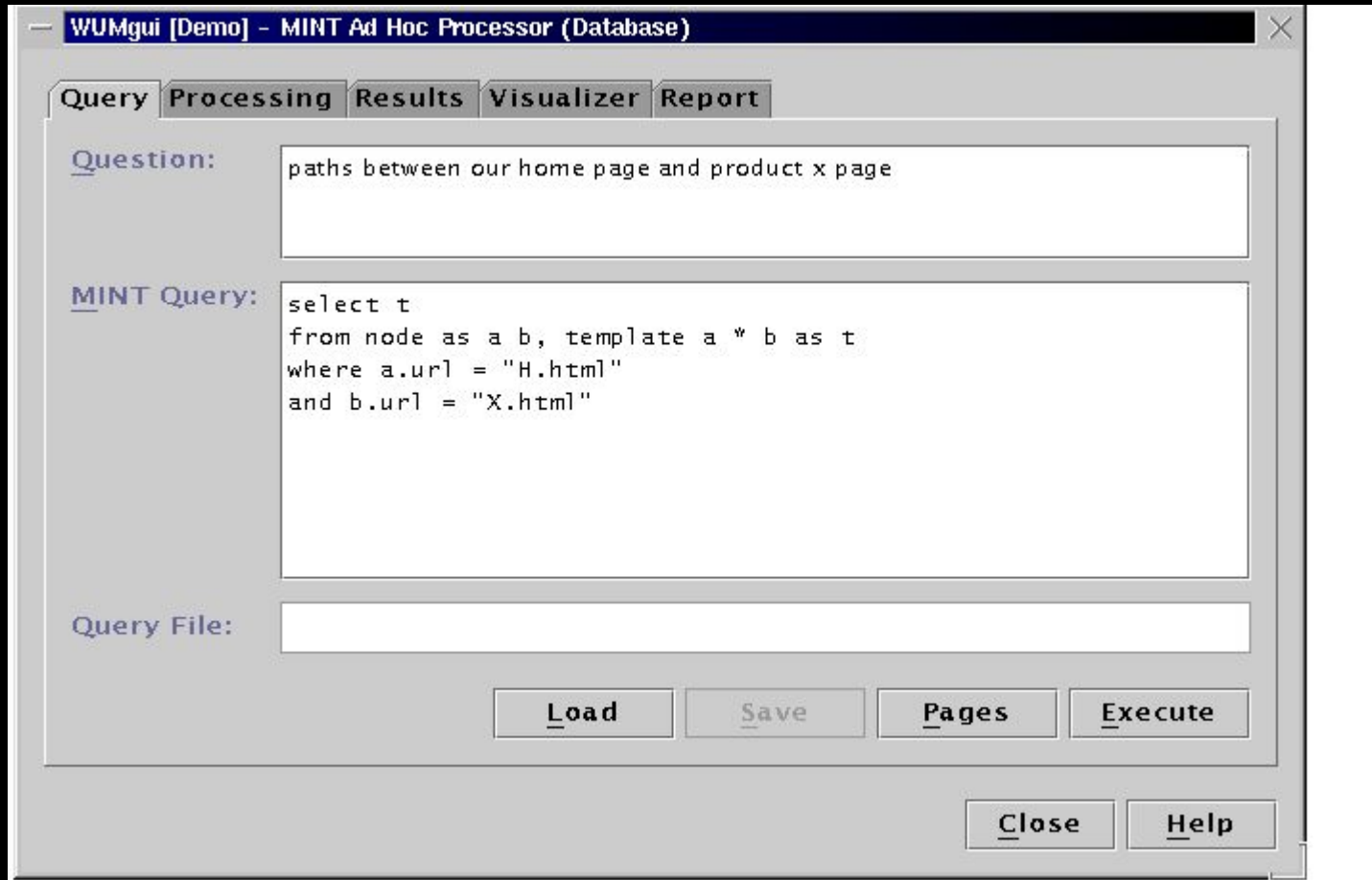
Details:

Report Section	Size
Most Requested Pages	20
Most Requested Directories	10
Least Requested Pages	20
Least Requested Directories	10
Top Entry Pages	20
Top Exit Pages	20
Single Access Pages	20
Most Active Top-Level Domains	20
Most Active 2nd-Level Domains	20
Top Visitors	20
Top Referrer Sites	20
Top Referrer Pages	20
Most Used Browsers	10

# WUM



# WUM



# WUM

WUMgui [Demo] - MINT Ad Hoc Processor (Database)

Query Processing Results Visualizer Report

Please wait: Your MINT query is being executed.

Using standard algorithm and logging in file /users/kwinkler/WUM.v60/data/demoWebSite/WUM.MiningLog.txt...

```
START select t
from node as a b, template a * b as t
where a.url = "H.html"
and b.url = "X.html"
```

-----+-----#-----  
:-----:-----#-----;-----  
-----+-----:-----

STOP 0:00:00.815 |

Stop

Close Help

WUMgui [Demo] - Result Aggregated Tree 1/2

Tree Graph Tree View

```
graph LR
  a["a = [H.html;1;14]"] --> p["[P.html;1;]"]
  a --> g["[G.html;1;]"]
  p --> b["b = [X.html;1;4]"]
  g --> b["b = [X.html;1;4]"]
```

Redraw NodeInfo Edge++ Edge-- Export Cancel

# LOGML

- Log Markup(işaret) Language (Dili)
- LOGML DTD
- LOGML Schema
- <http://www.cs.rpi.edu/~puninj/LOGML/>  
<!ELEMENT logml (graph, hosts?, domains?, directories?, userAgents?, hostReferers?, referers?, keywords\*, summary?, userSessions?)>  
<!ATTLIST logml  
    %global-atts;  
    %xml-atts;  
    start\_date     %date.type;         #REQUIRED  
    end\_date       %date.type;         #REQUIRED  
>

# ROBOTlarrr

- Crawler, Spiders, Arama motoru botu
  - Dizinleme
  - HTML / link doğrulama
  - Yeni ne var?
  - Mirroring (Yansı)
- <http://www.google.com/bot.html>
  - Sürekli dolaşıyor
  - `<META NAME="googlebot" CONTENT="noarchive,nofollow">`

# Robots.txt

- rapid-fire (Bir bot gördüm sanki!)  
# /robots.txt file for http://webcrawler.com/  
# mail webmaster@webcrawler.com for constructive criticism  
User-agent: webcrawler  
Disallow:  
User-agent: lycra  
Disallow: /  
User-agent: \*  
Disallow: /tmp  
Disallow: /logs
- Tüm siteyi kapat
  - User-agent: \*
  - Disallow: /
- <META NAME="robots" CONTENT="noindex">
- <META NAME="robots" CONTENT="nofollow">
- <META NAME="robots" CONTENT="noarchive">
- <http://www.robotstxt.org/>

# Analiz Sonrası

- Ziyaretçi
  - {Kim, Neyi, Ne zaman}, istemiş
  - Profil
- Sunucuyu eniyileştirme
  - Bandwidth
  - Güvenlik
  - Kaynakların doğru kullanımı
- Sayfalalar
  - Prefetching (Önyükleme)
  - Dinamik Link(Bağ) Tavsiyesi

Teşekkürler..

?