What We Discovered From the Log File

From analyzing the Apache access log file, we discovered the following:

-Total number of HTTP requests recorded in the log file was **10,000**

-The **distribution of request methods** was:

- `GET`: 9952 requests
- `POST`: 5 requests
- `HEAD`: 43 requests

-The number of **unique IP addresses** that accessed the server was **1,753**

-The **IP address that made the most GET requests** was `66.249.73.135` with **482 requests**

-The **IP address that made the most POST requests** was `78.173.140.106` with **3 requests**.

-There were **220 failed requests** (client-side or server-side errors)

-Breakdown of those **error status codes**:

- 404 Not Found: 213 times
- 500 Internal Server Error: 3 times
- 416 Range Not Satisfiable: 2 times
- 403 Forbidden: 2 times

- The **distribution of requests by hour** showed traffic across all hours, with a peak around hours 14 and 15.

-The **distribution of failed requests by day** was:

  - 18/May/2015: 66 errors
  - 19/May/2015: 66 errors
  - 20/May/2015: 58 errors
  - 21/May/2015: 30 errors

- The average number of requests made by each IP address is approximately **5.7**

- Most visitors make a small number of requests, indicating normal browsing behavior

- **Finding:** 6 IP addresses made more than 100 requests

357 130.237.218.86

102 209.85.238.199

364 46.105.14.53

113 50.16.19.13

482 66.249.73.135

273 75.97.9.59

- HTTP Status Code Distribution

Finding: Most responses were successful (200), but there were also some 404 (Not Found) and 500 (Server Error) responses

9126 200

445 304

213 404

164 301

45 206

3 500

2 416

2 403

- Top 10 Most Active Ips

Finding: The top IP made 482 requests

482 66.249.73.135

364 46.105.14.53

357 130.237.218.86

273 75.97.9.59

113 50.16.19.13

102 209.85.238.199

99 68.180.224.225

84 100.43.83.137

83 208.115.111.72

82 198.46.149.143

- Hourly Request Distribution

Finding: Peak activity occurred between 10:00 AM and 8:00 PM

443 10

459 11

462 12

475 13

498 14

496 15

473 16

484 17

478 18

493 19

- Most Requested Resources

Finding: Many requests were for static files like `/favicon.ico`, CSS, and images

807 /favicon.ico

546 /style2.css

538 /reset.css

533 /images/jordan-80.png

516 /images/web/2009/banner.png

488 /blog/tags/puppet?flav=rss20

224 /projects/xdotool/

217 /?flav=rss20

197 /

180 /robots.txt

- IPs Causing the Most 404 Errors

Finding: IP `208.91.156.11` alone caused 60 "Not Found" errors.

60 208.91.156.11

14 144.76.95.39

8 91.236.75.25

8 66.249.73.135

6 75.97.9.59

5 176.92.75.62

4 84.137.208.44

4 130.237.218.86

3 95.78.54.93

3 78.173.140.106

-Top User Agents (Browsers and Bots)

Finding: Most traffic came from Chrome and Firefox browsers

1044 Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/32.0.1700.107 Safari/537.36

369 Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/33.0.1750.91 Safari/537.36

364 UniversalFeedParser/4.2-pre-314-svn +http://feedparser.org/

296 Mozilla/5.0 (Windows NT 6.1; WOW64; rv:27.0) Gecko/20100101 Firefox/27.0

...

– Most Requested Pages (excluding images, CSS, JS, etc.)

Finding: RSS feeds and project pages are popular among users and bots.

488 /blog/tags/puppet?flav=rss20

224 /projects/xdotool/

217 /?flav=rss20

197 /

180 /robots.txt

...

- Request Distribution by Day

Finding: Peak traffic occurred on **May 19th, 2015**.

1632 17/May/2015

2893 18/May/2015

2896 19/May/2015

2579 20/May/2015

- Request Methods Used

Finding: Over 99% of requests are `GET`, showing typical browsing behavior

9952 GET

42 HEAD

5 POST

1 OPTIONS

- Top Referrers (External and Internal Sources)

Finding: Most users accessed the site directly or via internal links.

4073 -

689 http://semicomplete.com/presentations/logstash-puppetconf-2012/

656 http://www.semicomplete.com/projects/xdotool/

406 http://semicomplete.com/presentations/logstash-scale11x/

...

Commands Used and What Each Produced

Here are the exact Bash commands we used and the results they gave us:

1. **Total number of requests:**
   awk 'END {print NR}' apache_logs

Result: 10000 requests

- **Meaning**: The log contains 10,000 lines, each representing one HTTP request.

2. Request method counts:

awk '{print $6}' apache_logs | sort | uniq -c

Result:

```
9952 "GET
 42 "HEAD
  1 "OPTIONS
  5 "POST
```

Meaning: Most of the traffic involved GET requests; very few POSTs and HEADs were recorded.

3. Count of unique IP addresses:

awk '{print $1}' apache_logs | sort | uniq | wc -l

Result: 1753

Meaning: 1,753 distinct IP addresses accessed the server

4. Most active IP (GET requests):

grep '"GET' apache_logs | awk '{print $1}' | sort | uniq -c | sort -nr | head -1

Result:

482 66.249.73.135

Meaning: This IP issued the highest number of GET requests

5. Most active IP (POST requests):

grep '"POST' apache_logs | awk '{print $1}' | sort | uniq -c | sort -nr | head -1

Result:

3 78.173.140.106

Meaning: This IP made the most POST requests (still a very small number).

6. Total number of failed requests:

awk '$9 ~ /^4|^5/ {print $9}' apache_logs | wc -l

Result: 220

Meaning: There were 220 client/server errors (status codes starting with 4 or 5).

7. Breakdown of error status codes:

awk '$9 ~ /^4|^5/ {print $9}' apache_logs | sort | uniq -c | sort -nr

Result:

213 404

3 500

2 416

2 403

Meaning: 404s dominated the errors, meaning many requests were made to non-existent pages

8. Requests per hour:

awk -F: '{print $2}' apache_logs | uniq -c

Sample Result:

361 [00
383 [01
...
498 [14
496 [15

Meaning: Helps visualize hourly traffic volume — useful for spotting peak usage times

9. **Failed requests per day:**

awk '$9 ~ /^4|^5/ {print $4}' apache_logs | cut -d: -f1 | sed 's/\[//' | uniq -c

Result:

66 18/May/2015
66 19/May/2015
58 20/May/2015
30 21/May/2015

Meaning: Error activity was higher earlier in the week, possibly indicating resolved issues or changing usage patterns

10. **Average Number of Requests per IP**

awk '{print $1}' apache_logs | sort | uniq -c | awk '{total+=$1; count++} END {print total/count}'

Result:

5.70451

İnterpretation: Most visitors make a small number of requests, indicating normal browsing behavior.

11. **IPs with More Than 100 Requests**

awk '{print $1}' apache_logs | sort | uniq -c | awk '$1 > 100'

Result:

357 130.237.218.86
102 209.85.238.199
364 46.105.14.53
113 50.16.19.13
482 66.249.73.135
273 75.97.9.59

Interpretation: These could be bots or highly active users and might require further monitoring

12.HTTP Status Code Distribution

```
awk '{print $9}' apache_logs | sort | uniq -c | sort -nr
```

Result:

```
9126 200
 445 304
 213 404
 164 301
  45 206
   3 500
   2 416
   2 403
```

Interpretation: The server is generally healthy, but some requests target missing or problematic resources.

13- Top 10 Most Active Ips

```
awk '{print $1}' apache_logs | sort | uniq -c | sort -nr | head -n 10
```

Result:

```
482 66.249.73.135
364 46.105.14.53
357 130.237.218.86
273 75.97.9.59
113 50.16.19.13
102 209.85.238.199
```

99 68.180.224.225
84 100.43.83.137
83 208.115.111.72
82 198.46.149.143

Interpretation: These may be bots or automated scanners and should be reviewed.

14- Hourly Request Distribution

awk -F: '{print $2}' apache_logs | cut -d[-f2 | cut -d] -f1 | cut -d: -f1 | sort | uniq -c

Result:

443 10
459 11
462 12
475 13
498 14
496 15
473 16
484 17
478 18
493 19

15- Most Requested Resources

awk '{print $7}' apache_logs | sort | uniq -c | sort -nr | head -10

Finding: Many requests were for static files like `/favicon.ico`, CSS, and images.

Result:

807 /favicon.ico

546 /style2.css

538 /reset.css

533 /images/jordan-80.png

516 /images/web/2009/banner.png

488 /blog/tags/puppet?flav=rss20

224 /projects/xdotool/

217 /?flav=rss20

197 /

180 /robots.txt

Interpretation: Could be a mix of real users and bots scanning all linked resources

16- IPs Causing the Most 404 Errors

```
awk '$9 == 404 {print $1}' apache_logs | sort | uniq -c | sort -nr | head -10
```

Finding: IP `208.91.156.11` alone caused 60 "Not Found" errors.

60 208.91.156.11

14 144.76.95.39

8 91.236.75.25

8 66.249.73.135

6 75.97.9.59

5 176.92.75.62

4 84.137.208.44

4 130.237.218.86

3 95.78.54.93

3 78.173.140.106

Interpretation: Likely a bot probing for vulnerable or outdated URLs. Should be monitored or blocked.

17- Top User Agents (Browsers and Bots)

```
awk -F\" '{print $6}' apache_logs | sort | uniq -c | sort -nr | head -10
```

Finding: Most traffic came from Chrome and Firefox browsers.

Result: 1044 Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/32.0.1700.107 Safari/537.36

369 Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/33.0.1750.91 Safari/537.36

364 UniversalFeedParser/4.2-pre-314-svn +http://feedparser.org/

296 Mozilla/5.0 (Windows NT 6.1; WOW64; rv:27.0) Gecko/20100101 Firefox/27.0

...

Interpretation: There's significant bot traffic, especially from **Googlebot** and **RSS readers** like UniversalFeedParser and Tiny Tiny RSS.

18- Most Requested Pages (excluding images, CSS, JS, etc.)

```
awk '{print $7}' apache_logs | grep -vE "\.(jpg|png|gif|css|js|ico)$" | sort | uniq -c | sort -nr | head -10
```

Finding: RSS feeds and project pages are popular among users and bots

Result:

488 /blog/tags/puppet?flav=rss20

224 /projects/xdotool/

217 /?flav=rss20

197 /

180 /robots.txt

...

Interpretation: The presence of `/robots.txt` indicates web crawlers activity.

19- Request Distribution by Day

awk -F['{print $2}' apache_logs | cut -d: -f1 | uniq -c

Finding: Peak traffic occurred on **May 19th, 2015**

Result:

1632 17/May/2015

2893 18/May/2015

2896 19/May/2015

2579 20/May/2015

Interpretation: Traffic was relatively stable throughout the observed days.

20- Request Methods Used

awk '{print $6}' apache_logs | cut -d'"' -f2 | sort | uniq -c | sort -nr

Finding: Over 99% of requests are `GET`, showing typical browsing behavior

Result:

9952 GET

42 HEAD

5 POST

1 OPTIONS

Interpretation: Few `POST` requests suggest minimal form submissions or API calls

21- Top Referrers (External and Internal Sources)

```
awk -F\" '{print $4}' apache_logs | sort | uniq -c | sort -nr | head -10
```

Finding: Most users accessed the site directly or via internal links.

Result:

```
4073 -

 689 http://semicomplete.com/presentations/logstash-puppetconf-2012/

 656 http://www.semicomplete.com/projects/xdotool/

 406 http://semicomplete.com/presentations/logstash-scale11x/

 ...
```

Interpretation: Significant traffic came from presentations, technical blogs, and articles hosted on the same domain

Final Insights

- The website attracts **both human users and a lot of bots**, including crawlers and feed readers.
- The site's **technical content** is the most popular (projects and blogs).
- There are **minor backend issues (500 errors)** and **broken/missing links** that should be addressed.
- Bot traffic, especially requesting `favicon.ico`, is substantial and could be optimized/cached.
- **Peak hours and dates** are useful for performance tuning and capacity planning

Comprehensive Apache Log Analysis Summary

1. Visitor Activity & Traffic Distribution

- The **average number of requests per unique IP** is approximately **5.7**, suggesting a **low engagement per user** or the presence of many one-time/bot visits.
- Some IPs had **very high activity** (e.g., `66.249.73.135` with 482 requests), which is often a sign of bots like Googlebot or scraping tools.

2. Top IPs and Potential Bots

- The top 10 most active IPs are heavily skewed toward known bot behavior or automated tools. IPs like `66.249.73.135` and `130.237.218.86` likely represent web crawlers.

3. HTTP Response Codes

- Majority of responses were `200 OK` (9126), which is normal.
- But we also saw:
  - `213` times: **404 Not Found** → broken links or bots probing for missing resources.
  - `3` times: **500 Internal Server Error** → potential backend/server issue.
  - `445` times: **304 Not Modified** → caching works correctly.

4. Most Frequent 404 IPs

- IP `208.91.156.11` caused the highest number of 404 errors (60 times), indicating possible scanning or misconfigured bot behavior.

5. Hourly Traffic Trends

- Traffic increases gradually during the day and peaks between **11:00 to 19:00**, which reflects **typical working hours**.

. Most Requested Resources

- `/favicon.ico` was requested 807 times → often by browsers/bots automatically.
- Pages like `/blog/tags/puppet?flav=rss20` and `/projects/xdotool/` had high engagement, which indicates user interest in **technical articles and projects**.

7. Traffic by Date

- The **19th of May 2015** had the highest traffic (2896 requests), possibly due to promotion, publication, or bot activity spike.

8. HTTP Methods Used

- `GET` was overwhelmingly used (9952 times), which is expected.
- A few `POST`, `HEAD`, and `OPTIONS` methods appeared, likely from bots, scanners, or limited dynamic interactions.

9. User Agent Analysis

- Chrome and Firefox dominate browser access.
- A lot of hits from known bots:
  - `Googlebot`
  - `UniversalFeedParser`
  - `Tiny Tiny RSS`

  This confirms high bot traffic and RSS feed usage

10. Top Referrers

- Most traffic came **directly (-)**, followed by internal links.
- External referrers came from:
  - Technical articles and project pages
  - Presentations hosted on the same domain

  This implies **content-driven inbound traffic**.

Evidence-Based Security Insights

1. **Unusual Number of Requests from Certain IPs**
   `66.249.73.135` made **482 requests**,
   `46.105.14.53` made **364 requests**

Insight: Such high request counts from single IPs may indicate **brute-force attempts**, **web scraping**, or **DoS probing**.

Recommendation: Consider rate-limiting or temporarily blocking such IPs via a firewall or IDS system

2.High Failure Rates (4xx and 5xx Errors)

`404 errors = 213, 500 errors = 3`

Insight: Repeated failed access attempts often mean users or bots are probing for **non-existent** or **restricted resources**.

Recommendation: Monitor IPs generating frequent failures, and use WAF rules to block or redirect

Suspicious Behavior from IPs Causing Most 404s
IP `208.91.156.11` caused **60** of the 404 errors.

Insight: This IP may be scanning for vulnerabilities or old paths (e.g., WordPress plugins, admin panels).

Recommendation: Block IPs with high 4xx activity or alert security team for deeper inspection.

User-Agent Analysis Revealing Bots and Crawlers

- `Googlebot` User-Agent made **237 requests**
- `UniversalFeedParser` made **364 requests**

Insight: These agents are either legit crawlers or impersonating bots trying to scrape data.

Recommendation: Respect robots.txt, use CAPTCHA for forms, and verify bot authenticity via reverse DNS

Suspicious Referrers

4073 requests had referrer `"-"` (empty), and others came from internal presentation links.

Insight: Empty referrers might suggest **direct/scripted requests** or **automated attacks**.

Recommendation: Add CSRF tokens, log referrers more deeply, and watch for unusual spikes.

Requests to /robots.txt and Other Metadata Files

`/robots.txt` accessed **180 times**

Insight: Indicates bots checking what is "disallowed" – could mean **malicious scanning**.

Recommendation: Avoid putting sensitive paths in `robots.txt` and monitor its access.

Security Recommendations

- Block or rate-limit suspicious IPs.
- Monitor failed requests for patterns.
- Restrict or validate access to sensitive URLs.
- Regularly review and update server security configurations
