Log File Analysis Report

Apache Web Server Log Analysis Using Bash Script

Submitted by: Youssef Mohammed Ahmed
ID: 2205168
Submission for: Log File Analysis Task

Contents

1 Introduction

This report presents a comprehensive analysis of an Apache web server log file (`apache_logs.txt`) using a Bash script, as per the requirements of the Log File Analysis Task. The log spans from 17 May 2015 to 20 May 2015, capturing 145 requests. The analysis includes request counts, unique IP addresses, failure requests, most active users, daily averages, hourly patterns, status code distributions, and failure trends, followed by actionable suggestions to improve system performance and security.

2 Request Counts

The total number of requests, along with the breakdown of GET and POST requests, is as follows:

- Total Requests: 145
- GET Requests: 145
- POST Requests: 0

The absence of POST requests suggests the server primarily handles read-only operations or serves static content.

3 Unique IP Analysis

The log file contains requests from 43 unique IP addresses. The breakdown of GET and POST requests per IP is saved in the file `Uniq_IPs`. Key entries include:

```
38.99.236.50 GET: 25 POST: 0
83.149.9.216 GET: 23 POST: 0
66.249.73.135 GET: 13 POST: 0
46.105.14.53 GET: 8 POST: 0
```

The full breakdown (43 IPs) is available in `Uniq_IPs`. The complete list is:

```
38.99.236.50 GET: 25 POST: 0
83.149.9.216 GET: 23 POST: 0
66.249.73.135 GET: 13 POST: 0
46.105.14.53 GET: 8 POST: 0
208.115.111.72 GET: 7 POST: 0
218.30.103.62 GET: 7 POST: 0
110.136.166.128 GET: 6 POST: 0
63.140.98.80 GET: 6 POST: 0
91.177.205.119 GET: 6 POST: 0
93.114.45.13 GET: 6 POST: 0
212.180.233.101 GET: 6 POST: 0
```

```
108.28.155.98 GET: 6 POST: 0
71.212.224.97 GET: 5 POST: 0
86.1.76.62 GET: 5 POST: 0
81.220.24.207 GET: 5 POST: 0
200.49.190.101 GET: 3 POST: 0
66.249.73.185 GET: 3 POST: 0
207.241.237.220 GET: 2 POST: 0
207.241.237.227 GET: 2 POST: 0
198.46.149.143 GET: 2 POST: 0
92.115.179.247 GET: 2 POST: 0
50.16.19.13 GET: 2 POST: 0
91.151.182.109 GET: 2 POST: 0
82.165.139.53 GET: 2 POST: 0
207.241.237.225 GET: 1 POST: 0
207.241.237.228 GET: 1 POST: 0
207.241.237.101 GET: 1 POST: 0
209.85.238.199 GET: 1 POST: 0
123.125.71.35 GET: 1 POST: 0
50.150.204.184 GET: 1 POST: 0
67.214.178.190 GET: 1 POST: 0
87.169.99.232 GET: 1 POST: 0
24.236.252.67 GET: 1 POST: 0
200.49.190.100 GET: 1 POST: 0
108.174.55.234 GET: 1 POST: 0
121.107.188.202 GET: 1 POST: 0
107.170.41.69 GET: 1 POST: 0
180.76.6.130 GET: 1 POST: 0
173.192.238.41 GET: 1 POST: 0
109.163.234.2 GET: 1 POST: 0
68.180.224.225 GET: 1 POST: 0
5.10.83.82 GET: 1 POST: 0
74.125.40.20 GET: 1 POST: 0
```

4 Failure Requests

Requests with status codes in the 4xx or 5xx range are considered failures:

- Failed Requests (4xx/5xx): 3
- Failure Percentage: 2.07%

The low failure rate indicates robust server performance, though specific 404 errors warrant investigation.

5 Most Active IP

The IP address with the most requests is:

- Most Active IP: 38.99.236.50 (Requests: 25)

This IP accounts for 17.24% of all requests, raising potential concerns about its activity.

6 Daily Request Averages

The log file spans 4 unique days, with an average request rate of:

- Average Requests per Day: 36.25

7 Days with Highest Failures

The days with the highest number of failure requests are:

```
2 [20/May/2015
1 [17/May/2015
```

Failures are relatively evenly distributed, with a slight increase on 20 May 2015.

8 Requests by Hour

The distribution of requests by hour of the day is:

```
92 10
50 21
 3 11
 0 20
```

Significant spikes occur at 10:00 (92 requests) and 21:00 (50 requests), indicating peak usage periods.

9 Status Codes Breakdown

The frequency of each status code is:

```
200 136
304 6
404 3
```

The majority of requests (93.79%) are successful (200), with a small number of cached responses (304) and errors (404).

10 Most Active Users by Method

The IP addresses with the most GET and POST requests are:

- Top GET IP: 38.99.236.50 (Requests: 25)
- Top POST IP: None (No POST requests)

11 Failure Patterns by Hour

The hours with the most failure requests are:

```
2 21
1 10
```

Failures primarily occur at 21:00, aligning with a peak request period.

12 Request Trends

The data reveals clear temporal trends:

- **Hourly Spikes**: The majority of requests occur at 10:00 (63.45% of total) and 21:00 (34.48% of total), suggesting concentrated user or bot activity during these hours.
- **Daily Consistency**: Requests are relatively evenly distributed across the 4 days (average 36.25 per day), with no extreme daily fluctuations.
- **No POST Activity**: The absence of POST requests indicates the server is not handling form submissions or API write operations in this log sample.

13 Patterns in Failure Requests

Failures are minimal (3 total), but patterns include:

- **Temporal Concentration**: Two of the three 404 errors occur at 21:00 on 20 May 2015, during a high-traffic period.
- **Specific Resources**: The 404 errors are for:
  - `/doc/index.html?org/elasticsearch/action/search/SearchResponse.ht` (17/May/2015, 10:05:22)
  - `/files/logstash/logstash-1.3.2-monolithic.jar` (20/May/2015, 21:05:11)

- `/presentations/logstash-puppetconf-2012/images/office-space-print` (20/May/2015, 21:05:36)

These suggest missing files or incorrect URLs.

14 Analysis Suggestions

Based on the analysis, the following suggestions are proposed:

- **Investigate 404 Errors**: The three 404 errors indicate missing resources. Verify if these files (`logstash-1.3.2-monolithic.jar`, `office-space-printer-bea` and the Elasticsearch documentation page) are intentionally absent or need to be restored. Implement redirects or error pages to handle such requests gracefully.
- **Monitor High-Activity IP**: The IP 38.99.236.50 accounts for 25 requests (17.24% of total). This could indicate legitimate user activity (e.g., viewing a presentation) or automated scraping. Consider rate limiting or monitoring this IP to prevent potential abuse.
- **Optimize for Peak Hours**: The request spikes at 10:00 (92 requests) and 21:00 (50 requests) suggest high user or bot activity. Scale server resources (e.g., increase CPU/memory or enable caching) during these hours to maintain performance.
- **Address Bot Traffic**: IPs like 66.249.73.135 (Googlebot) and 208.115.111.72 (Ezooms) indicate crawler activity. Ensure the site is optimized for SEO (e.g., proper `robots.txt`, sitemaps) and monitor crawler impact on server load. The multiple requests for `robots.txt` (e.g., from 218.30.103.62, 180.76.6.56) suggest bots are checking crawling permissions.
- **Investigate Lack of POST Requests**: If the server is expected to handle POST requests (e.g., for form submissions or API writes), investigate why none are present. This could indicate a configuration issue or that the log sample only captures read-only operations.
- **Enhance Logging**: To better diagnose failures, enhance logging to capture additional details (e.g., error stack traces for 404s). This can help identify whether 404s result from user errors or server misconfigurations.
- **Security Monitoring**: While no immediate security threats are evident, the high activity from 38.99.236.50 and bot traffic warrant ongoing monitoring. Implement intrusion detection or anomaly detection to catch unusual patterns (e.g., rapid requests from a single IP).

15 Conclusion

The analysis of the Apache web server log file reveals a stable system with a low failure rate (2.07%) and consistent daily request patterns. The server primarily handles GET requests, with significant activity from a single IP (38.99.236.50) and during peak hours (10:00 and 21:00). The minimal failures (three 404s) suggest robust performance, but missing resources and bot activity require attention. By implementing the suggested improvements—such as investigating 404 errors, optimizing for peak hours, and monitoring high-activity IPs—the system's performance, reliability, and security can be further enhanced.