Name:Ahmed nabil nour ahmed

ID: 2205245

Log File Analysis Report

Abstract

This report analyzes a web server log file containing 1,569,898 requests over 30 days, processed using a Bash script. The analysis covers request counts, unique IP addresses, failure rates, daily averages, hourly trends, status codes, and failure patterns. Key findings include a low failure rate (0.67%), peak request times between 12:00–15:00, and high failure days at the end of August 1995. Recommendations address failure reduction, peak load management, and potential security concerns.

Introduction

The objective of this analysis is to extract insights from a web server log file to understand request patterns, identify issues, and suggest improvements. The log file, sourced from [placeholder:

https://www.kaggle.com/datasets/adchatakora/nasa-http-access-logs?resource=download , was analyzed using a Bash script to generate statistics on request counts, IP activity, failures, and trends. This report presents the findings, answers the required questions, and provides actionable suggestions.

Analysis Results

The following sections address the requirements outlined in the task.

1. **Request Counts**
   - **Total Requests**: 1,569,898
   - **GET Requests**: 1,565,812 (99.74% of total)
   - **POST Requests**: 111 (0.007% of total)
     Observation: The overwhelming majority of requests are GET, indicating a read-heavy workload, typical for web servers serving static or informational content.
2. **Unique IP Addresses**
   - **Total Unique IPs**: 75,060
   - **Top 5 IPs by GET Requests**:
     - edams.ksc.nasa.gov: 6,528
     - piweba4y.prodigy.com: 4,846
     - 163.206.89.4: 4,791
     - piweba5y.prodigy.com: 4,607
     - piweba3y.prodigy.com: 4,416
   - **Top 5 IPs by POST Requests**:
     - seabrk.mindspring.com: 8
     - 155.33.77.108: 6

pc0139.metrolink.net: 5
n868370.ksc.nasa.gov: 5
163.205.1.19: 4

Observation: The top GET IPs are significantly more active than POST IPs, suggesting heavy browsing activity from specific domains (e.g., NASA and Prodigy).

3. **Failure Requests**

   Total Failures: 10,489 (0.67% of total requests)

   Failure Percentage:

   10,4891,569,898×100=0.67%\frac{10,489}{1,569,898} \times 100 = 0.67\%1,569,89810,489×100=0.67%
   Observation: The low failure rate indicates a generally stable system, though specific failure patterns require attention.

4. **Top User**

   Most Active IP: edams.ksc.nasa.gov (6,530 requests)

   Observation: This IP's high activity suggests it may belong to a frequent user or automated system (e.g., a crawler or monitoring tool).

5. **Daily Request Averages**

   Total Days: 30

   Average Requests per Day: 52,329.93

   (1,569,89830\frac{1,569,898}{30}301,569,898)
   Observation: The consistent daily average suggests stable traffic, with potential peaks on specific days.

6. **Failure Analysis**

   Days with Highest Failures:

   30/Aug/1995: 601 failures

31/Aug/1995: 541 failures
07/Aug/1995: 538 failures
29/Aug/1995: 453 failures
25/Aug/1995: 444 failures

Observation: Failures are concentrated toward the end of August, possibly due to increased traffic or system issues.

7. **Requests by Hour**

Peak Hours:

Hour 15: 109,465 requests
Hour 12: 105,143 requests
Hour 13: 104,536 requests

Low Hours:

Hour 04: 26,756 requests
Hour 05: 27,587 requests
Hour 03: 29,995 requests

Observation: Requests peak during midday (12:00–15:00), likely corresponding to business hours, and drop significantly overnight.

8. **Request Trends**

Requests increase steadily from 06:00, peak at 15:00, and decline after 17:00.

The highest activity occurs during typical working hours (08:00–17:00), suggesting a user base active during daytime.

Low activity overnight (00:00–05:00) indicates reduced global usage during these hours.

9. **Status Codes Breakdown**

Success/Redirect Codes:

    - 200 (OK): 1,396,473 (88.96%)
    - 304 (Not Modified): 134,138 (8.54%)
    - 302 (Found): 26,422 (1.68%)
  - **Failure Codes (4xx/5xx)**:
    - 404 (Not Found): 9,978 (0.64%)
    - 403 (Forbidden): 171 (0.01%)
    - 501 (Not Implemented): 27 (<0.01%)
    - 500 (Internal Server Error): 3 (<0.01%)
    - Others (e.g., 509, 527): Minor occurrences

      Observation: Most failures are 404 errors, suggesting missing resources or broken links. Server errors (5xx) are minimal.

10. **Most Active User by Method**

    GET: edams.ksc.nasa.gov (6,528 requests)

    POST: seabrk.mindspring.com (8 requests)

    Observation: POST requests are rare, and their activity is distributed across fewer IPs.

11. **Patterns in Failure Requests**

    Top Failure Hours:
    - Hour 12: 688 failures
    - Hour 13: 631 failures
    - Hour 02: 622 failures

    Low Failure Hours:
    - Hour 06: 139 failures
    - Hour 05: 174 failures
    - Hour 04: 182 failures

    Observation: Failures peak during high-traffic hours (12:00–13:00), suggesting load-related issues, but

Hour 02’s high failures are anomalous, possibly due to automated processes or errors.

Trends and Patterns

- **Hourly Trends**: Traffic peaks during midday (12:00–15:00), with Hour 15 being the busiest (109,465 requests). Overnight hours (00:00–05:00) see the lowest activity, indicating a user base primarily active during daytime.
- **Failure Patterns**: Failures correlate with peak traffic hours (12:00–13:00), except for Hour 02, which has high failures (622) despite low traffic (32,508 requests). This suggests potential issues with automated scripts or maintenance tasks.
- **Daily Failure Trends**: High failure days (e.g., 30–31 August) align with the end of the month, possibly due to increased user activity or system updates.
- **IP Activity**: The top IP (edams.ksc.nasa.gov) accounts for a significant portion of GET requests, indicating heavy usage from a single source, possibly a crawler or institutional user.

Suggestions

Based on the analysis, the following recommendations address failures, performance, and security:

1. **Reducing Failures**

404 Errors (9,978 occurrences): Audit the website for broken links and missing resources. Implement redirects for deprecated URLs and ensure content is properly maintained.

Peak Hour Failures (Hours 12–13): Scale server capacity during peak hours (12:00–15:00) using load balancing or cloud-based resources to handle high traffic.

Hour 02 Anomalies: Investigate high failures during low-traffic Hour 02. This could indicate misconfigured scripts, bots, or maintenance tasks causing errors.

2. **Days/Times Needing Attention**

End-of-Month Failures (30–31 August): Monitor system performance at month-end, as increased failures suggest higher traffic or system strain. Schedule maintenance outside these periods.

Peak Hours (12:00–15:00): Optimize server performance during these hours by caching static content and prioritizing critical requests.

3. **Security Concerns and Anomalies**

High Activity from Single IP (edams.ksc.nasa.gov, 6,530 requests): Investigate this IP’s behavior to determine if it’s a legitimate user (e.g., NASA crawler) or a potential bot. Implement rate-limiting for IPs exceeding a request threshold.

POST Request IPs: Monitor IPs making POST requests (e.g., seabrk.mindspring.com), as these are rare and could indicate form submissions or API interactions. Ensure POST endpoints are secure against abuse.

- **Unusual Status Codes (e.g., 786, 669)**: Investigate non-standard status codes to confirm they are intentional or identify misconfigurations.

4. **System Improvements**
   - **Content Delivery Network (CDN)**: Deploy a CDN to reduce server load during peak hours and improve response times for global users.
   - **Logging Enhancements**: Add more granular logging (e.g., request paths, user agents) to better diagnose 404 errors and anomalous failures.
   - **Automated Monitoring**: Implement real-time monitoring for failure spikes and alert administrators during high-failure periods (e.g., Hour 02 or 30–31 August).

Conclusion

The log file analysis reveals a stable web server with a low failure rate (0.67%) and predictable traffic patterns, peaking during midday hours. However, concentrated failures at month-end and during specific hours (e.g., Hour 02) indicate areas for improvement. By addressing 404 errors, scaling capacity during peak times, and investigating high-activity IPs, the system can achieve better reliability and security. The log file has been uploaded to GitHub for reference.

The log_file source

https://www.kaggle.com/datasets/adchatakora/nasa-http-access-logs?resource=download