Analysing and Visualising ALB Log data using AWS Athena and Quick-Sight

Amogh A Sura
Department of Computer Science andEngineering
Faculty of Engineering and Technology
– Jain (Deemed-to-be University)
Bengaluru, India 19btrct003@jainuniversity.ac.in

Vedanth S Tammewar
Department of Computer Science andEngineering
Faculty of Engineering and Technology
– Jain (Deemed-to-be University)
Bengaluru, India 19btrct045@jainuniversity.ac.in

Krishna Prasad M L
Department of Computer Science andEngineering
Faculty of Engineering and Technology
– Jain (Deemed-to-be University)
Bengaluru, India 19btrct019@jainuniversity.ac.in

Vishesh Singh
Department of Computer Science andEngineering
Faculty of Engineering and Technology
– Jain (Deemed-to-be University)
Bengaluru, India 19btrct061@jainuniversity.ac.in

Prateek Bharadwaj
Department of Computer Science andEngineering
Faculty of Engineering and Technology
– Jain (Deemed-to-be University)
Bengaluru, India 19btrct033@jainuniversity.ac.in

Dr. A. Vijay Kumar
Department of Computer Science andEngineering
Faculty of Engineering and Technology
– Jain (Deemed-to-be University)
Bengaluru, India ak.vijay@jainuniversity.ac.in

Abstract

In the modern world, servers produce many log entries. Most of the data are not being used to their full potential; we have developed solution to generate valuable information.

An EC2 instance running the NGINX server has beenused to deploy a highly available website. To achieve high availability, we have developed an application load balancer to share the traffic between the instances. The log data from the Load Balancer is transformed into a database using Amazon Athena and is saved in S3 Buckets. Amazon Quick Sight is used to visualize this logdata. Using various AWS Services, we are displaying thelog data to provide greater insights. The firm can use this to make an informed decision because they can get personalized insights that will support them based on their requirements.

I. INTRODUCTION

Logging is the process of writing down details aboutimportant events on a permanent medium so that otherscan read them later. A huge file called a log is used to keep track of all activities in software programmes. Inorder to address the issues with Towards network blockage and overloading of the servers which guarantee todeliver continuous and well ended services, load balancingtechnology and Web caching technology are given toppriority [1].

A workflow's log events may include redundant data, such as recurring log events and groups of log events that always occur together [2]. Interpreting andextrapolating from the ALB logs is a challenge for most enterprises. The company must go through alaborious process to find the necessary data. Each log is derived from a unique source. As a result, the formatin which each source generates and reports logs to thelog monitoring tools for analysis varies. Several eventlog monitoring solutions include a standard log formatto address this issue, however it is impossible for all logs to adhere to the same format. There is therefore no

assurance that all of your incoming log data will have the format that is similar to what is being currently analysed, even with the installation of a standard log format.

How to efficiently store, query, analyse, and use massive datasets in the cloud is the fundamental challenge [3]. By analysing and visualising the ALB logs, our project hopes to gain insights from them. There are numerous uses, some of which include the ability to count how many times a user accesses a specific URL and to estimate request access time using Amazon Athena andAWS Quick-Sight. This benefits the business in anumber of ways, including by assisting with marketing and sales strategies and business expansion.

Log analysis can improve logging's capacity as a trouble-shooter. Fast searches and excellent visualisation features are frequently included with log analysis tools, which could speed up problem diagnosis and resolution. Nevertheless, log analysis can help you take it a step further and stop problems intheir tracks. You can develop your ability to spot trendsthat portend impending issues by studying prior events.

Log analysis can help you uncover security breaches, traffic surges on your web servers, and otherissues.

II. PROBLEM STATEMENT

1. In the modern world, servers produce many log entries. Most of the data are not being used to their full potential.
2. Every action in the software applications is recorded in a huge file called a log. The Application Load Balancer automatically logs the message, reports oferror, file access requests, file transfers, and sign in and sign out requests (ALB).
3. The majority of businesses find it challenging to extrapolate from and analyze the ALB logs. The organization must go through a laborious process tofind the needed data.
4. This makes it harder for businesses to make right decisions even when analyzing the problem using visualization techniques.
5. Data quality is an additional issue with log data. It might be difficult to use log data for analysis since it may be inconsistent, erroneous, or incomplete. This may result in false positives or false negatives, which may affect the analysis's effectiveness.
6. Log data may include private information like IP addresses or user passwords. In order to prevent illegal access, this information must be stored securely. Everyone is aware that the log is full withuseful information. Data about user activity, system operation status, and other topics can be gathered by log analysis [4]. In order to avoid unwanted access,which could jeopardize the security of the system, *C. AWS S3:*

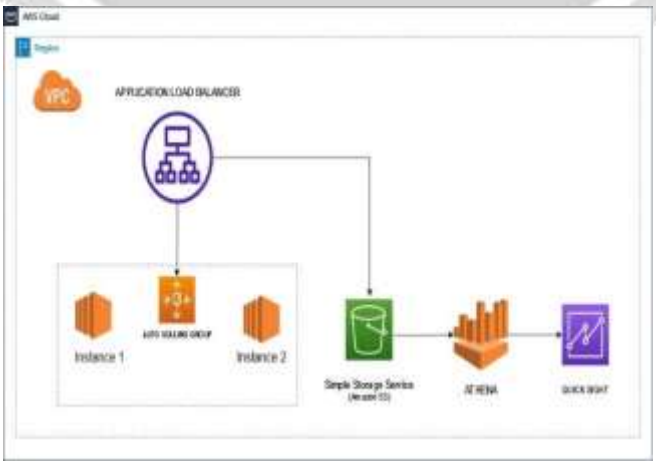

Fig.1. Architecture

access to log data also needs to be properly regulated.

7. Depending on the legal, regulatory, or business requirements, organizations must decide how long to keep log data.

Managing this can be difficult, especially when working with big data sets. Also, it might be difficult to handle log data properly because retention policies are subject to change overtime.

III. TOOLS USED

A. AWS ALB:

A technique called load balancing is used to divide the requests among the servers which are running so that requests responses can be supplied more quickly[5]. An application load balancer works on applications layer which falls on the seventh layers of the OSI model , when the request is recieved by the load balancer , it goes through the rules provided by the listener in an order based on the priority which determines what rule that has to be applied and then it selects target from different groups for rule action

B. AWS EC2:

EC2 provides us multiple types of instances which can also be optimized to fit our conditions and other cases ,its also useful to combine multiple Memory , CPU , storage , and other networking capacities and also provied you the flexibility to choose and apply for different resources for your application. Each instances includes multiple instances and multiple sizes which allows you to scale your requirements based on your necessity of your workload.

Performance, security, and scalability of an object storage service called Amazon Simple Storage Service are unequalled in the industry. Cloud storage is a novel technology that allows information to be hoarded on servers (data centres), accessed online, and maintained by the cloud provider [6]. Any quantity of data can be hoarded and secured by customers of variety of sizes and parts for practically any use case, comprising datalakes, cloud-native applications, and mobile applications. You can cut down expenses, sort data, and set-up limited access restrictions to comply with certain business, and organisational requirements with the help of practical storage classes and simple administration tools.

D. AWS QuickSight:

Quicksight allows you to understand your data by unsing natural languages in any organization and brings out interactive dashboards by looking through the patters which are promoted/powered by machine learning.

E. AWS Athena:

AWS Amazon provides Athena, a serverless interactive analytics solution that may be used to quickly acquire understanding of data stored in S3. The SQL queries are executed by Athena using a distributed SQL engine named Presto. Hive, a well-known open-source technology, forms the foundation of Presto, which uses it to store structured, semi-structured, and unstructured data.

IV. WORKING METHODOLOGY

1. Initially, we will create 2 Linux EC2 instances (Instance 1 & Instance 2) in different availability zones to make our website highly available. SSH to the Linux Instances using Putty.
2. Low memory usage, high concurrency, great extensibility, a variety of third-party modules, and open source codes are all characteristics of Nginx [7]. Install the NGINX server in both instances and replace the default website at
   /usr/sbin/nginx/index.html with our custom website.
3. Create a target group (ProjectTG), and include both instances (Instance 1 and Instance 2) into the target group.
4. Select an Application Load Balancer by using the create new Load Balancer button among different types of Balancers and provide a name (ProjectLB) using the target groups which were created earlier.
5. Create the S3 bucket (ctisprojectalblogs) to store all the log details of the Application Load Balancer in the S3.
6. Go to Application Load Balancer attributes and specify the S3 bucket URL (s3://ctisprojectalblogs) in the Access Log section.
7. Go to AWS Athena, and create a new Database.(Log_db)
8. Write a query to create a table (log_table) for storing the log details.
9. Now, we have all the log details in tabular format.
10. In the AWS Quick Sight dashboard, create a Dataset and map it to the log table created in AWS Athena.

11. Using AWS QuickSight, we can visualize all thelog parameters for better insights.

V. FUTURE WORK

- Organizations may store and analyze log data moreaffordably and flexibly with the aid of cloud-based technologies.
- Additionally, cloud-based solutions enablebusinesses to access log data from any location,which is beneficial for remote teams in particular Another trend that will influence the future of logdata is data visualization. Organizations can quickly and efficiently make sense of massive amounts of log data with the aid ofdata visualization tools.
  With data visualization tools, firms may rapidly andeffectively uncover issues by identifying patterns and trends in log date. We will monitor log data in real-time to spot issues as real-time technologies become more prevalent.
- Organizations may respond to security events and performance problems in real-time, before they have a chance to do any impact.
- Advanced analytics is one of the major themes thatwill influence how log data is can be the future.
- Organizations will need cutting-edge analytics toolsto make sense of the data as computer systems create more log data.
- Sophisticated analytics technologies, we plan toassist organizations in finding patterns and trends in
  log data, enabling them to rapidly and effectively identify security events, performance difficulties, and other concerns.
  - We plan to add artificial intelligence is anotherdevelopment that will influence how log data is used in the future (AI)
  -

VI. CONCLUSION

The analysis's findings will show them how to raise awareness of low priority services, advertise them, and promote them, as well as how to modify the website as needed[8]. For debugging, monitoring, and auditing computer systems, applications, and networks, log data is a useful source of information. Yet, log data has a unique set ofproblems that must be handled well. Organizations can take action to handle log data properly by being aware of these issues, which can aid in the quick and efficient identificationof issues.

By considering log data as a substantial component of company data supply and exploiting it with Big Data techniques, organisations may now access a wealth of knowledge that is contained in unstructured log files that are already being gathered but underused due to their great diversity and volume.[9]. The majority of organizations don'tcare about log data, so we created a project that enables the organization to analyze and visualize the log information from their application load balancer, which is then used by the organization to make important decisions and track the application's active users. We have built a highly accessible website using two EC2 instances.

Using various AWS Services, we are displaying the log datato provide greater insights. The firm can use this to make an informed decision because they can get personalized insightsthat will support them based on their requirements.

VII. REFERENCES

[1] Chi, Xiaoni; Liu, Bichuan; Niu, Qi; Wu, Qiuxuan (2012). [IEEE 2012 Third International Conference on Digital Manufacturing and Automation (ICDMA) - Guilin, China (2012.07.31-2012.08.2)] 2012 Third International Conference on Digital Manufacturing & Automation - Web Load Balance and Cache Optimization Design Based Nginx under High-Concurrency Environment. , (), 1029– 1032. doi:10.1109/ICDMA.2012.241.

[2] S. Locke, H. Li, T. -H. P. Chen, W. Shang and W. Liu, "LogAssist: Assisting Log Analysis Through Log Summarization," in IEEE Transactions on Software Engineering, vol. 48, no. 9, pp. 3227-3241, 1 Sept. 2022, doi: 10.1109/TSE.2021.3083715.

[3] R. Buyya, J. Broberg, A. Goscinski, "Cloud Computing: Principles and Pradigms", Wiley, 201l.

[4] X. Lin, P. Wang and B. Wu, "Log analysis in cloud computing environment with Hadoop and Spark," 2013 5th IEEE International Conference on Broadband Network & Multimedia Technology, Guilin, China, 2013, pp. 273-276, doi: 10.1109/ICBNMT.2013.6823956.

[5] M. Rahman, S. Iqbal and J. Gao, "Load Balancer as a Service in Cloud Computing," *2014 IEEE 8th International Symposium on Service Oriented System Engineering*, Oxford,
UK, 2014, pp. 204-211, doi: 10.1109/SOSE.2014.31
[6] A. Gupta, A. Mehta, L. Daver and P. Banga, "Implementation of Storage in Virtual Private Cloud using Simple Storage Service on AWS," 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bangalore, India, 2020, pp. 213-217,doi: 10.1109/ICIMIA48430.2020.9074899.
[7] Z. Wen, G. Li and G. Yang, "Research and Realization of Nginx-based Dynamic Feedback Load Balancing Algorithm," 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, 2018, pp. 2541-2546, doi: 10.1109/IAEAC.2018.8577911.
[8] S. Narkhede, T. Baraskar and D. Mukhopadhyay, "Analyzing web application log files to find hit count through the utilization of Hadoop MapReduce
in cloud computing environment," *2014 Conference on IT inBusiness, Industry and Government (CSIBIG)*, Indore, India,2014, pp. 1-7, doi: 10.1109/CSIBIG.2014.7056950.
[9] M. Lemoudden and B. E. Ouahidi, "Managing cloud- generated logs using big data technologies," 2015 International Conference on Wireless Networks and Mobile Communications (WINCOM), Marrakech, Morocco, 2015, pp. 1-7, doi: 10.1109/WINCOM.2015.7381334.