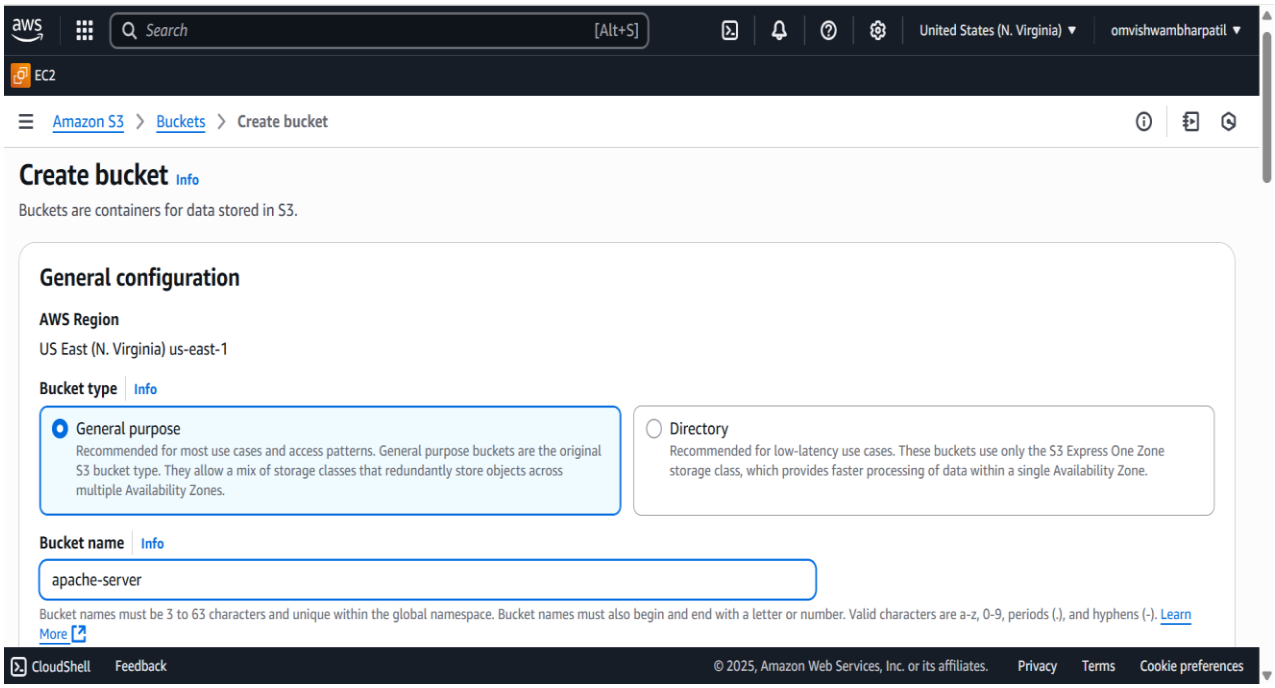
Web Server Log Analyzer

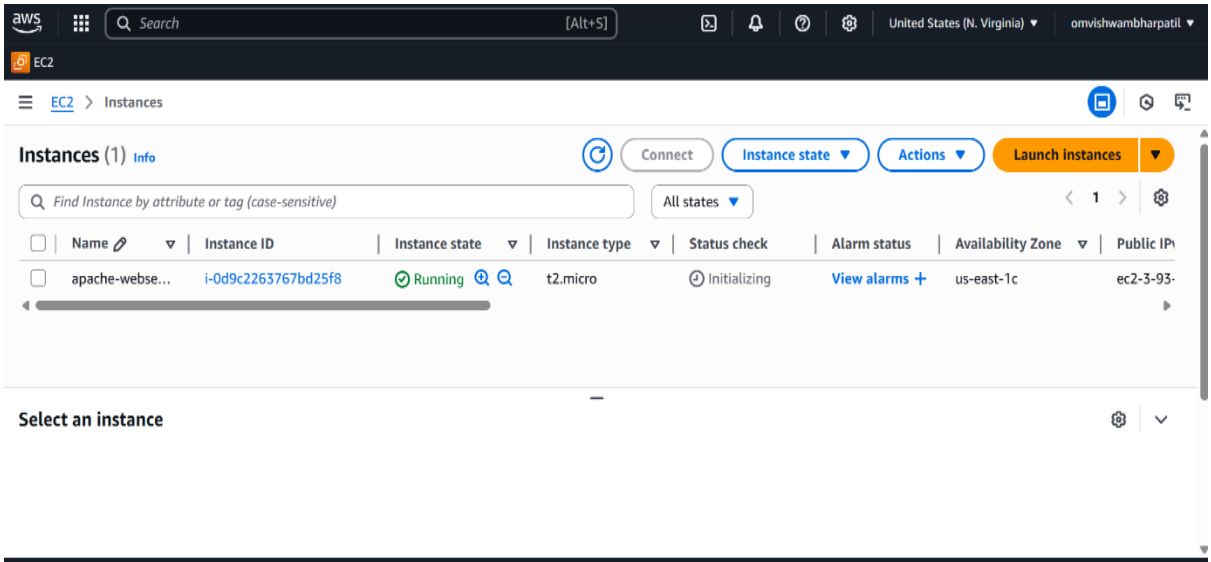
The **Web Server Log Analyzer** project is designed to process and analyze raw server log data stored in **Amazon S3**. Web servers generate log files that contain valuable information about user activity, error messages, IP addresses, request paths, timestamps, and more. These logs, when analyzed properly, can provide insights into website performance, security issues, user behavior, and server health.

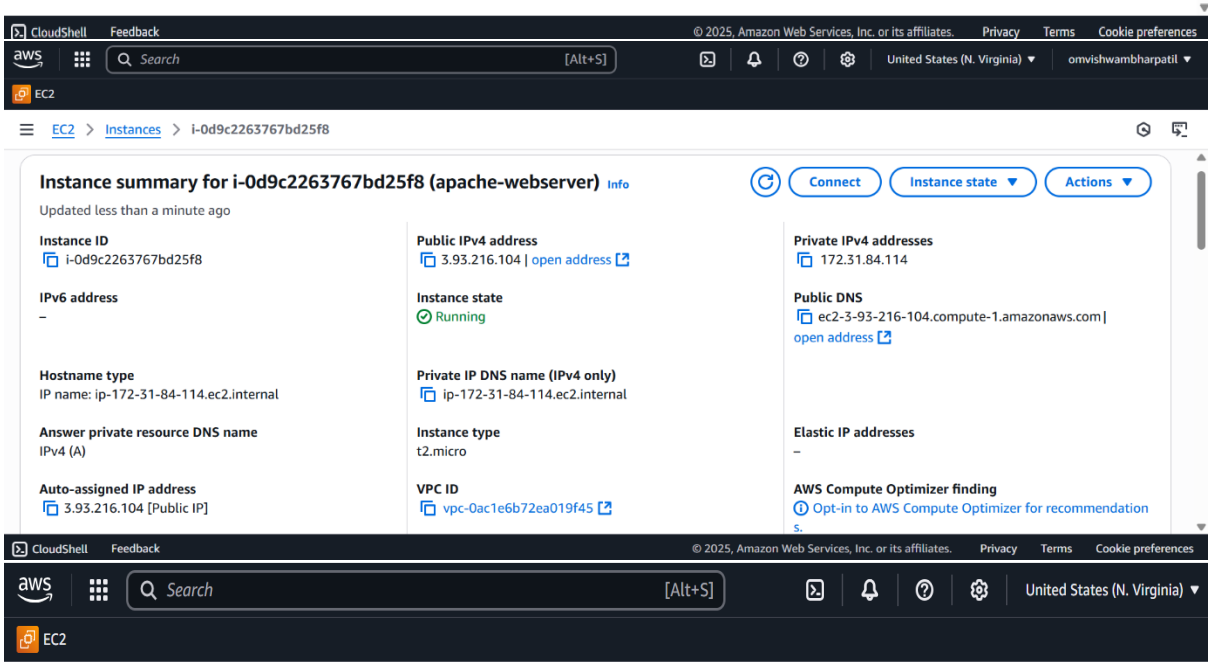
- Apache or Nginx web logs.
- Glue Transformations:
- Parse log lines into structured fields
- Extract metrics (top IPs, most visited URLs, error rate)
- Output (S3): Structured, queryable logs or analytics-ready data.

Apache or Nginx web logs:-

Apache or Nginx web server logs using a cloud-native architecture built on **Amazon Web Services (AWS)**. Both Apache and Nginx generate structured log files in formats such as the **Common Log Format (CLF)** or **Combined Log Format**, which contain valuable information about incoming HTTP requests—such as IP addresses, timestamps, URLs, request methods, status codes, user-agents, and referrers. These logs are typically generated on EC2 instances or web servers running in the cloud, and are then uploaded periodically to an **Amazon S3** bucket for centralized storage.







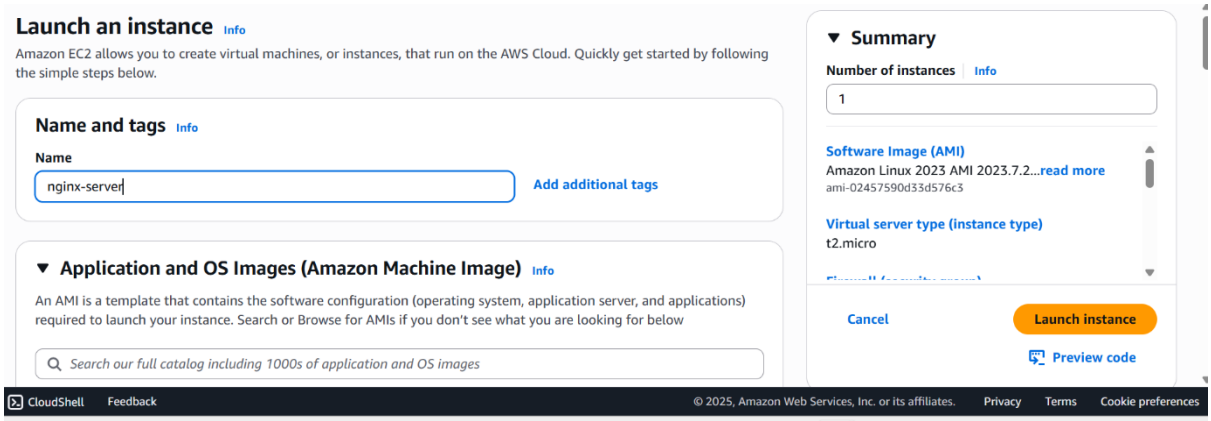

```
  ,     #_
  ~\_  ####_        Amazon Linux 2023
 ~~  \_#####\
 ~~     \###|
 ~~       \#/ ___   https://aws.amazon.com/linux/amazon-linux-2023
  ~~       V~' '->
   ~~~         /
     ~~._.   _/
        _/ _/
      _/m/'
Last login: Tue Jun 10 06:29:50 2025 from 18.206.107.29
[ec2-user@ip-172-31-84-114 ~]$ sudo yum install httpd
Last metadata expiration check: 0:38:22 ago on Tue Jun 10 06:33:47 2025.
Package httpd-2.4.62-1.amzn2023.x86_64 is already installed.
Dependencies resolved.
Nothing to do.
Complete!
[ec2-user@ip-172-31-84-114 ~]$ rpm -qa | grep httpd
httpd-tools-2.4.62-1.amzn2023.x86_64
```

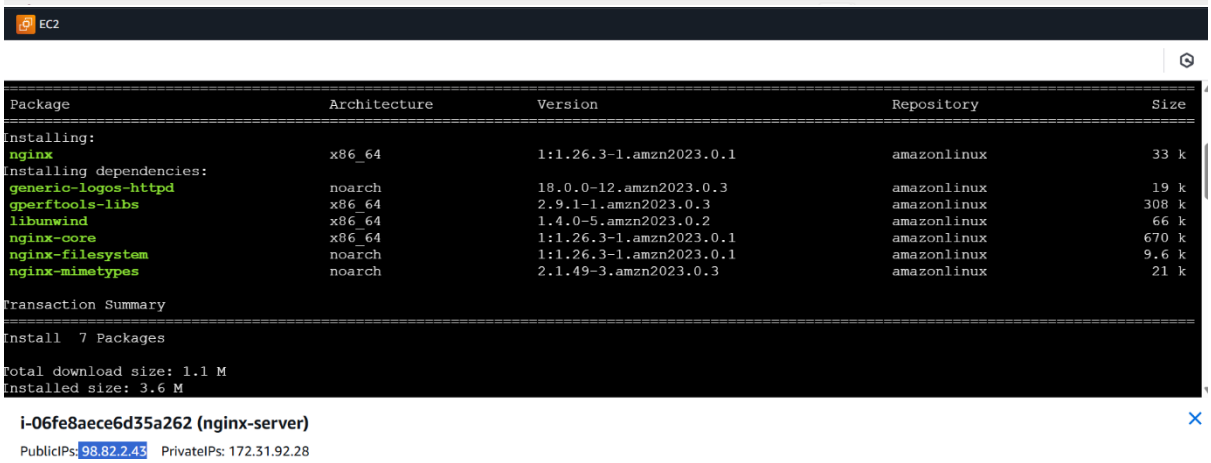

```
ec2-user@ip-172-31-84-114 ~]$ rpm -qa | grep httpd
httpd-tools-2.4.62-1.amzn2023.x86_64
httpd-filesystem-2.4.62-1.amzn2023.noarch
httpd-core-2.4.62-1.amzn2023.x86_64
generic-logos-httpd-18.0.0-12.amzn2023.0.3.noarch
httpd-2.4.62-1.amzn2023.x86_64
ec2-user@ip-172-31-84-114 ~]$ sudo systemctl status httpd
● httpd.service - The Apache HTTP Server
     Loaded: loaded (/usr/lib/systemd/system/httpd.service; disabled; preset: disabled)
     Active: active (running) since Tue 2025-06-10 06:56:49 UTC; 17min ago
       Docs: man:httpd.service(8)
   Main PID: 27175 (httpd)
     Status: "Total requests: 0; Idle/Busy workers 100/0;Requests/sec: 0; Bytes served/sec:   0 B/sec"
      Tasks: 230 (limit: 1111)
     Memory: 16.0M
        CPU: 749ms
     CGroup: /system.slice/httpd.service
             ├─27175 /usr/sbin/httpd -DFOREGROUND
             └─27184 /usr/sbin/httpd -DFOREGROUND
```

Nginx web logs

Nginx web logs provide critical visibility into web server activity and user interactions. Nginx generates logs in formats such as the **Access Log** and **Error Log**, typically following the **Combined Log Format**

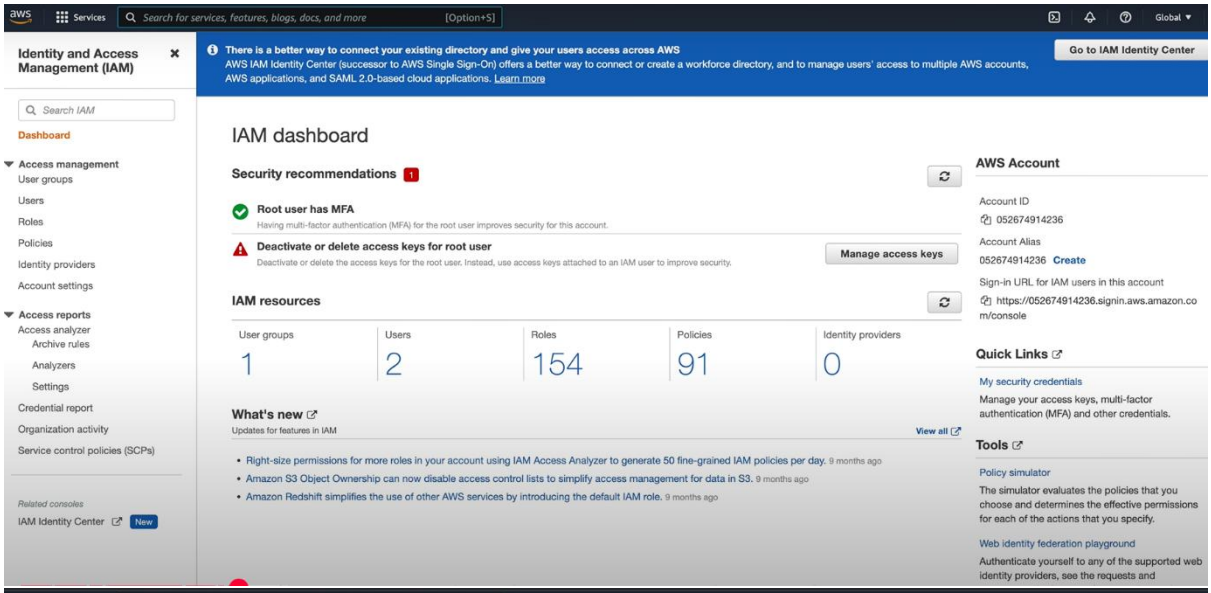
Nginx web logs are continuously pushed to **Amazon S3**, acting as a scalable and durable log repository. This enables decoupling of log generation and analysis processes.

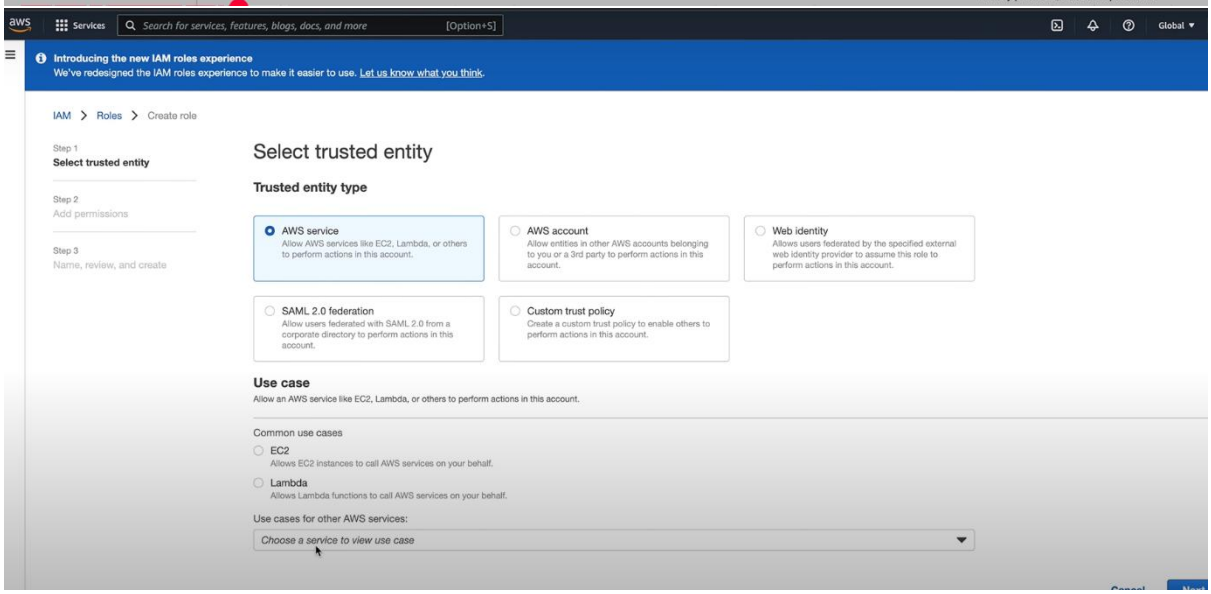




Glue Transformations:

AWS Glue is a fully managed extract, transform, and load (**ETL**) service that enables automatic data discovery, cleaning, enrichment, and preparation for analytics. After web logs are ingested into**Amazon**





aws | Services | Search for services, features, blogs, docs, and more [Option+S] | Global

We've redesigned the IAM roles experience to make it easier to use. Let us know what you think.

IAM > Roles > Create role

Step 1
Select trusted entity

Step 2
Add permissions

Step 3
Name, review, and create

Name, review, and create

Role details

Role name
Enter a meaningful name to identify this role.

aws-glue-service-role-demo

Maximum 64 characters. Use alphanumeric and '+=,.@-_' characters.

Description
Add a short explanation for this role.

Allows Glue to call AWS services on your behalf.

Maximum 1000 characters. Use alphanumeric and '+=,.@-_' characters.

Step 1: Select trusted entities

Ed

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Principal": {
                "Service": "glue.amazonaws.com"
            },
            "Action": "sts:AssumeRole"
        }
```

aws | Services | s | Global

Identity and Access Management (IAM)

Search IAM

Dashboard

Access management
User groups
Users
Roles
Policies
Identity providers
Account settings

Access reports
Access analyzer
Archive rules
Analyzers
Settings
Credential report
Organization activity
Service control policies (SCPs)

Related consoles
IAM Identity Center New

New! Securely access AWS services from your data center with IAM Roles Anywhere. Learn more

Role aws-glue-service-role-demo created | View role

IAM > Roles

Roles (155) Info
An IAM role is an identity you can create that has specific permissions with credentials that are valid for short durations. Roles can be assumed by entities that you trust.

Delete | Create role

Search

1 2 3 4 5 6 7 8

Role name	Trusted entities	Last activity
A2ISageMaker-ExecutionRole-20211024T130530	AWS Service: sagemaker	250 days ago
A2ISageMaker-ExecutionRole-20211219T140706	AWS Service: sagemaker	251 days ago
A2ISageMaker-ExecutionRole-20211219T145014	AWS Service: sagemaker	251 days ago
A2ISageMaker-ExecutionRole-20211220T194051	AWS Service: sagemaker	51 days ago
accessTokenAuth-role-y59klzz6	AWS Service: lambda	80 days ago
AmazonAppStreamServiceAccess	AWS Service: appstream	1 hour ago
AmazonComprehendServiceRole-compre-train	AWS Service: comprehend	-
AmazonComprehendServiceRole-comprehend-custom	AWS Service: comprehend	-
AmazonComprehendServiceRole-entity-recognizer	AWS Service: comprehend	-
AmazonSSMRoleForAutomationAssumeQuickSetup	AWS Service: ssm	-

aws | Services | Search for services, features, blogs, docs, and more [Option+S] | Global

Query with S3 Select Info

Use Amazon S3 Select to retrieve a subset of data from an object using standard SQL queries. Pricing is based on the size of the input, query results, and data transferred. Learn more or see Amazon S3 pricing

Input settings

Path
s3://aws-glue-demo-tutorial-series/data-store/annual_reports/csv_reports/annual-enterprise-survey-2021-financial-year-provisional-csv.csv

Size
6.3 MB (6617174.0 B)

Format
- (●) CSV
- ( ) JSON
- ( ) Apache Parquet

CSV delimiter
- (●) Comma
- ( ) Tab
- ( ) Custom

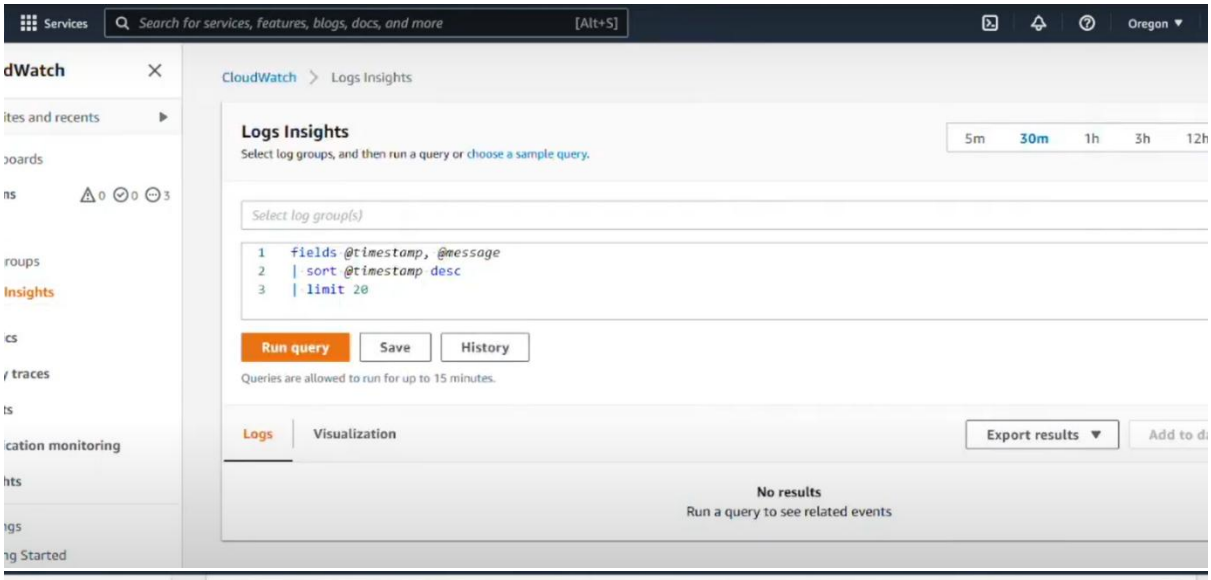
☐ Exclude the first line of CSV data
Enable this setting if CSV contains a header row.

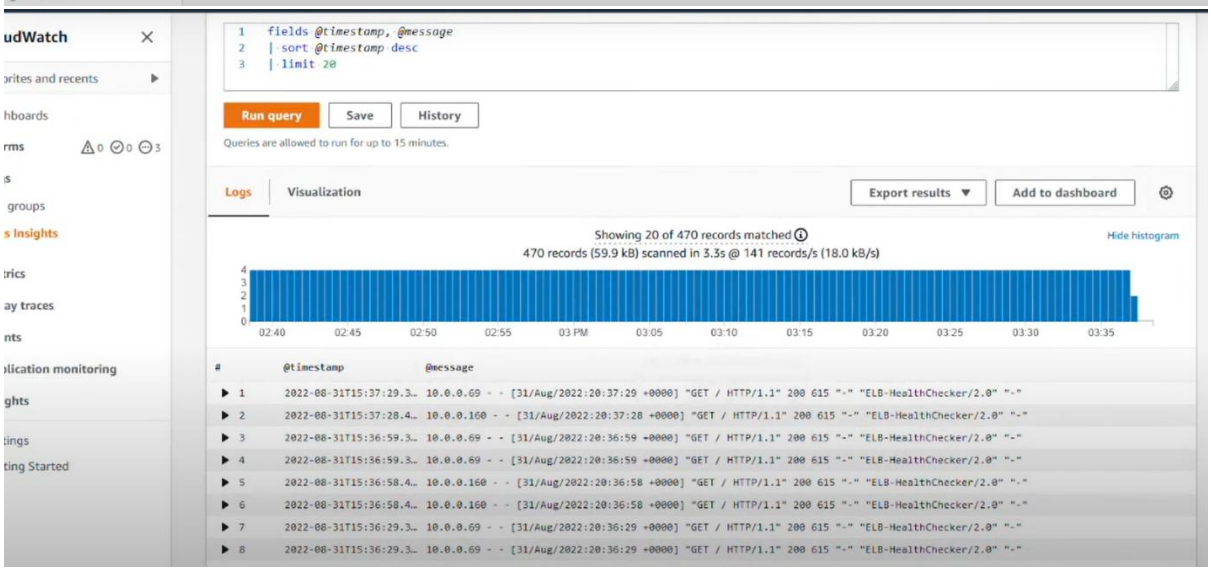
Compression
- (●) None
- ( ) GZIP
- ( ) BZIP2

Output settings

Parse log lines into structured fields:-

parse raw log lines into structured fields. Web logs from servers like Nginx or Apache are typically stored as unstructured text, where each line represents a request in a specific log format such as the Common Log Format (CLF) or the Combined Log Format. These lines contain information like the client IP address, request timestamp, HTTP method, requested URL, HTTP status code, user-agent, and referrer—all separated by spaces or enclosed in quotes.





Extract metrics (top IPs, most visited URLs, error rate)

After parsing the raw Nginx or Apache log data into structured fields, the next crucial step is to extract key metrics that provide insights into server usage and performance. Among the most

valuable metrics are the top client IP addresses, which help identify the most active users or potential malicious actors (e.g., in cases of denial-of-servic

AWS
aws-integration-zfyy9hcv

Overview | Configuration | Uninstall

Logs
1,313 Events/day
View logs

Metrics
108 Points/day
View metrics

Resources
126 Updates/day
View resources

Metrics [API polling] (87)

Selected AWS service metrics and data points collected in the last hour for each service.

AWS services

AWS service	Metrics selected	Points/Hour
AWS/EC2	23 of 34 metrics	108
AWS/RDS	64 of 128 metrics	0

Edit AWS/EC2 metrics

Select metrics

AWS EC2 metrics

All | Selected | Unselected

Metrics (23/34) Select all | Unselect all | Reset to defaults

- [x] CPUCreditBalance
- [x] CPUCreditUsage
- [x] CPUSurplusCreditBalance
- [x] CPUSurplusCreditsCharged
- [x] CPUUtilization
- [x] DedicatedHostCPUUtilization

Tags

CloudWatch Metrics can be filtered by tags. A tag is a key-value pair that you have applied to the resources of this AWS service.

Key | Value

+ Add tag

Dimensions

Cancel | Apply settings