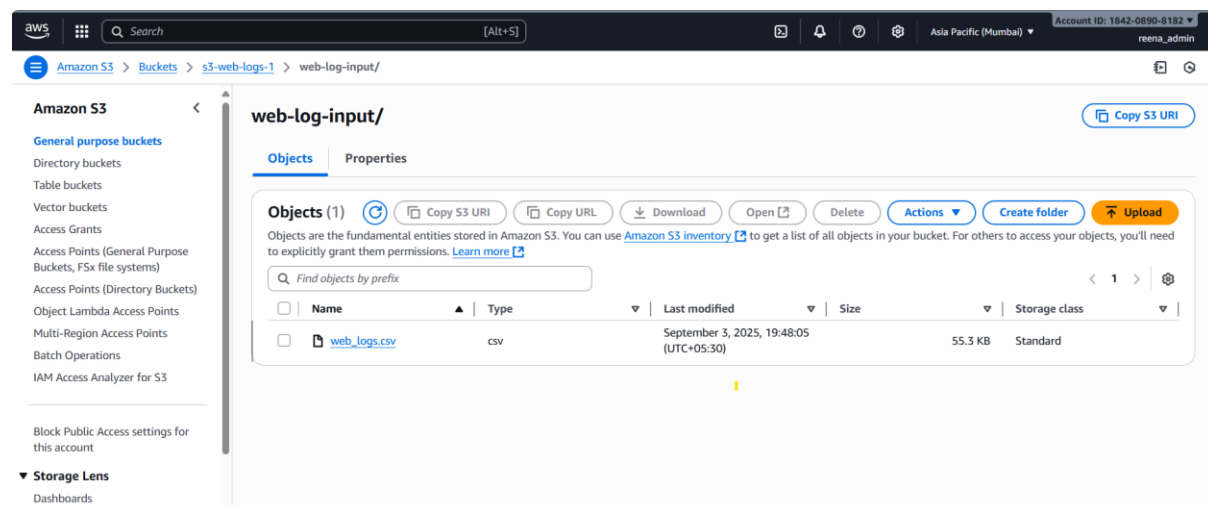**Scenario:**

You are a Data Engineer at an e-commerce company. The website stores **web server logs** in **Amazon S3** in .csv format. You want to analyze the logs to answer business questions such as:

- How many users visited the site per day?

- What are the top 5 most visited product pages?

- Which IPs generated the most traffic (possible bots)?

Solution:

Step 1: Upload Data to S3

Step 2: Create Database in Athena

CREATE DATABASE IF NOT EXISTS ecommerce_logs;

Step3: Create Table for Logs

CREATE EXTERNAL TABLE IF NOT EXISTS `ecommerce_logs`.`web_logs` (

  `ip_address` string,

  `timestamp` timestamp,

  `url` string,

  `status_code` bigint,

  `user_agent` string

)

ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'

WITH SERDEPROPERTIES ('separatorChar' = ',', 'quoteChar' = '\"')

STORED AS INPUTFORMAT 'org.apache.hadoop.mapred.TextInputFormat' OUTPUTFORMAT 'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'

LOCATION 's3://s3-web-logs-1/web-log-input/'

TBLPROPERTIES (

 'classification' = 'csv',

 'skip.header.line.count' = '1',

 'has_encrypted_data' = 'false'

);

Preview table query
The preview table query will be populated in the query editor. Creating tables allows you to be ready for real-time querying in the query editor.

```
1  CREATE EXTERNAL TABLE IF NOT EXISTS `ecommerce_logs`.`web_logs` (
2    `ip_address` string,
3    `timestamp` timestamp,
4    `url` string,
5    `status_code` bigint,
6    `user_agent` string
7  )
8  ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
9  WITH SERDEPROPERTIES ('separatorChar' = ',', 'quoteChar' = '\"')
10 STORED AS INPUTFORMAT 'org.apache.hadoop.mapred.TextInputFormat' OUTPUTFORMAT 'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'
11 LOCATION 's3://s3-web-logs-1/web-log-input/'
12 TBLPROPERTIES (
13   'classification' = 'csv'
```

4. Analysis Queries and outputs screenshots

1. Daily Visitor's

SELECT

date(parse_datetime(timestamp, 'yyyy-MM-dd HH:mm:ss')) AS visit_date,

COUNT(DISTINCT ip_address) AS unique_visitors

FROM ecommerce_logs.web_logs

GROUP BY 1

ORDER BY 1;

Query results   **Query stats**

✓ Completed | Time in queue: 61 ms | Run time: 522 ms | Data scanned: 55.25 KB

Results (7) 🗐 Copy   Download results CSV

🔍 Search rows   ‹ 1 › ⚙

#	visit_date	unique_visitors
1	2025-08-01	126
2	2025-08-02	159
3	2025-08-03	143
4	2025-08-04	144
5	2025-08-05	149
6	2025-08-06	148
7	2025-08-07	131

2. Top 5 pages

SELECT url, COUNT(*) AS hits

FROM ecommerce_logs.web_logs

GROUP BY url

ORDER BY hits DESC

LIMIT 5;

Query results

Completed   Time in queue: 58 ms   Run time: 445 ms   Data scanned: 55.25 KB

Results (5)

#	url	hits
1	/home	141
2	/search?q=shoes	133
3	/search?q=shirts	131
4	/checkout	127
5	/product/123	126

3. Top IP's

SELECT ip_address, COUNT(*) AS requests

FROM ecommerce_logs.web_logs

GROUP BY ip_address

ORDER BY requests DESC

LIMIT 10;

Query results

Completed   Time in queue: 55 ms   Run time: 447 ms   Data scanned: 55.25 KB

Results (10)

#	ip_address	requests
1	71.81.120.114	1
2	132.120.246.48	1
3	157.90.246.113	1
4	89.172.130.215	1
5	192.127.137.181	1
6	28.149.1.189	1
7	118.14.151.255	1
8	43.195.39.197	1
9	57.3.191.33	1
10	216.83.124.247	1