Cybersecurity Analysis: Suspicious Web Threat Interactions

A Data Analytics Project Using MySQL

---

1. Introduction

With the rapid growth of cloud-based applications, cybersecurity threats have become more frequent and complex. Organizations rely heavily on real-time traffic monitoring systems to detect suspicious activities and protect sensitive data.

This project focuses on analyzing **web traffic logs collected via AWS CloudWatch** to identify **suspicious web interactions** using **SQL-based data analytics techniques**. Instead of relying on machine learning models, this project emphasizes **structured query analysis**, exploratory data analysis (EDA), and logical reasoning — which are core responsibilities of a **Data Analytics Intern**.

The entire project is implemented using **MySQL Workbench**, ensuring simplicity, transparency, and real-world applicability.

---

2. Project Objectives

The primary objectives of this project are:

- To understand patterns in web traffic data
- To identify indicators of suspicious or abnormal behavior
- To analyze traffic based on geography, ports, and session duration
- To perform data cleaning and exploratory data analysis using SQL
- To derive meaningful cybersecurity insights from structured data

---

3. Dataset Description

The dataset used in this project consists of **282 web traffic records** captured from a production web server and monitored using AWS CloudWatch.

Each record represents a web interaction session that was flagged or observed during traffic monitoring.

Key Attributes in the Dataset

Column NameDescription	
bytes_in	Amount of data received by the server
bytes_out	Amount of data sent from the server
creation_time	Start time of the web session
end_time	End time of the web session
src_ip	Source IP address
src_ip_country_code	Country of origin
protocol	Network protocol used
response_code	HTTP response status
dst_port	Destination port
detection_types	Type of detection triggered

The dataset is well-structured and does not contain missing values, making it suitable for analytical exploration.

---

4. Tools and Technologies Used

- **Database Tool:** MySQL Workbench
- **Query Language:** SQL
- **Data Source:** AWS CloudWatch Web Traffic Logs
- **Domain:** Cybersecurity Analytics

---

5. Database Design and Data Import

A dedicated database was created in MySQL Workbench to store and analyze the data.
The dataset was imported using the **Table Data Import Wizard**, ensuring correct data types and avoiding manual loading errors.

A structured table design was used to store all relevant fields such as traffic volume, timestamps, IP information, and detection details.

---

6. Data Cleaning and Preparation

Data cleaning was performed directly using SQL queries to ensure analytical accuracy.

6.1 Duplicate Removal

Duplicate records were checked and removed to avoid skewed analysis results.

6.2 Data Standardization

- Country codes were converted to uppercase for consistency.
- Timestamp fields were stored in proper datetime format.

6.3 Feature Engineering

A new feature called **session_duration** was created by calculating the time difference between session start and end times.
This helped in identifying unusually long web sessions, which may indicate suspicious activity.

---

7. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted using SQL queries to uncover hidden patterns and trends.

7.1 Traffic Volume Analysis

The average incoming and outgoing traffic was analyzed to understand normal behavior versus anomalies.

Key Observation:
Some sessions had unusually high incoming traffic with low outgoing responses, which may indicate probing or reconnaissance attempts.

7.2 Country-wise Traffic Analysis

Traffic was grouped based on the source country code to identify regions contributing most to suspicious traffic.

Insight:
A small number of countries accounted for a disproportionately large share of flagged interactions, suggesting targeted or automated attacks.

7.3 Port Usage Analysis

Destination port usage was analyzed to detect misuse of standard or non-standard ports.

Insight:
Most interactions occurred over port **443 (HTTPS)**, indicating that encrypted channels are often used for malicious traffic, making detection more challenging.

7.4 Detection Type Distribution

Different detection types were analyzed to understand how traffic was being flagged.

Insight:
A majority of the records were associated with web application firewall (WAF) rules, confirming that the dataset focuses on suspicious or abnormal activity rather than normal traffic.

---

8. Advanced Analysis

8.1 High-Risk Traffic Identification

Sessions with extremely high incoming data and low outgoing responses were identified as potential security risks.

8.2 Long Session Detection

Sessions with unusually long durations were analyzed, as persistent connections may indicate unauthorized access attempts or data scraping activities.

---

9. Key Findings and Insights

- Suspicious web traffic is **pattern-based**, not random
- Certain geographic regions repeatedly appear in threat logs
- HTTPS traffic does not guarantee safety
- Session behavior (duration and volume) is a strong indicator of risk
- SQL alone is sufficient to extract meaningful cybersecurity insights

---

10. Conclusion

This project demonstrates how **structured SQL analysis** can be effectively used to analyze cybersecurity data and detect suspicious web interactions.

By focusing on **data cleaning**, **exploratory analysis**, and **logical reasoning**, the project provides transparent and explainable insights that are valuable for security teams and decision-makers.

For a Data Analytics Intern, this project showcases:

- Strong SQL fundamentals
- Real-world analytical thinking
- Ability to derive insights from raw operational data

The project avoids unnecessary complexity and presents a clean, professional, and industry-relevant analytical workflow