Lab 12

SQLi Pattern Analysis

Lab Objectives

- Manually parse through SQLi Logs via Opensearch
- Query the LLM to parse through the SQLi Index
- Prompt Engineer different queries for different results

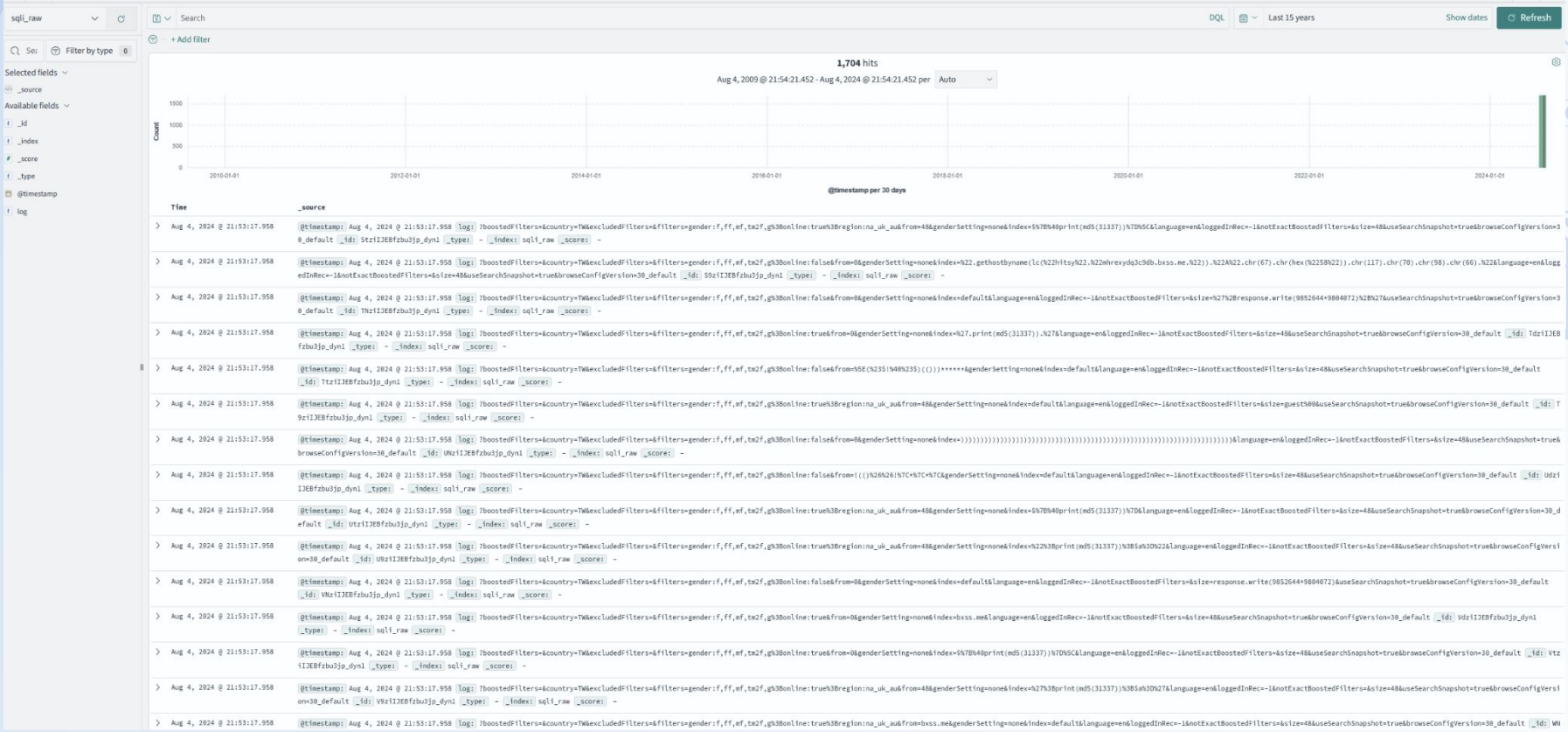
Different Indexes for... the same data?

As mentioned previously, the data for each lab is duplicated into two indexes:

- $ITEM_knn
- $ITEM_raw

This is done for primarily due to the embeddings generated with the knn index and the embedding model. If we were to search through the, for example, sqli_knn index, we would not only see the logs, but a bunch of integers alongside the logs in the embeddings field. There is no easy way to get around this. As such, the index patterns will always be focused on the “_raw” index. Lets go to the Discover page and look at our SQLi data.

Discover Page



What a mess

What a messy screen. Most, if not all of these logs, contain some form of SQL Injection. See if you can use the Search bar to find SQL Injection. The next slide will have a query you can use to get an event that has SQLi should you be unable to formulate the query.

SQLi Query

This Query Identifies a few SQLi attempts from the Netherlands:

```
log:SELECT AND log:nl
```

Identifying Countries of interest

Depending on if you were able to find SQLi yourself, or with the help of the provided query, you may have a variety of countries that have popped up. Without even writing a line of DQL, we can ask our LLM what countries are present in the dataset!

```
POST /_plugins/_ml/agents/yoLKXpgBlaNTCsEI6cTR/_execute
{
  "parameters": {
    "question": "What countries are present in the dataset."
  }
}
```

Two countries

If the LLM was able to receive results from the embedding model- you will only have two countries that pop up, the Netherlands, and Taiwan. From this limited dataset, your gut reaction may be to block these locations at the edge.. But we don't have time to write firewall rules...


```
{
  "inference_results": [
    {
      "output": [
        {
          "name": "MLModelTool",
          "result": """From the given context, we can see that the dataset contains log entries from different countries. The country code is a part of the URL
            parameter 'country'.

Let's analyze the data to find out which country codes are present in the dataset.

From the provided log entries, we can observe two distinct country codes: TW and NL. These are the country codes for Taiwan (Taiwan) and Netherlands (Netherlands
  ), respectively.

Therefore, based on our analysis of the data, the country codes present in the dataset are TW and NL."""
        }
      ]
    }
  ]
}
```

Firewall rules

For this- try to ask your LLM to write iptables(or other preferred firewall) rules to block inbound traffic the Netherlands and Taiwan. Try doing this with and without the RAG agent, and see how the results differ.

Lab Challenge - Prompt Engineering

Within this dataset- there are some attempts at fuzzing to gain access. You can view those events in Discover with this query:

```
log:nslookup AND log:nl
```

Try and use the LLM to find these same results. No worries if you do not succeed- this will be a part of the upcoming conversation. A query that (should) give you the wanted output will be provided at the end of the lab for you to test.

Lab End