analiza

June 6, 2024

[1]:
```
!pip install pyspark
!pip install kaggle
```

```
Requirement already satisfied: pyspark in /opt/conda/lib/python3.10/site-
packages (3.3.2)
Requirement already satisfied: py4j==0.10.9.5 in /opt/conda/lib/python3.10/site-
packages (from pyspark) (0.10.9.5)
WARNING: Error parsing requirements for jinja2: [Errno 2] No such file or
directory: '/opt/conda/lib/python3.10/site-packages/Jinja2-3.1.2.dist-
info/METADATA'
WARNING: Error parsing requirements for jsonschema: [Errno 2] No such
file or directory: '/opt/conda/lib/python3.10/site-
packages/jsonschema-4.17.3.dist-info/METADATA'
WARNING: Error parsing requirements for platformdirs: [Errno 2] No such
file or directory: '/opt/conda/lib/python3.10/site-
packages/platformdirs-3.5.0.dist-info/METADATA'
WARNING: Error parsing requirements for websocket-client: [Errno 2] No
such file or directory: '/opt/conda/lib/python3.10/site-
packages/websocket_client-1.5.1.dist-info/METADATA'
WARNING: Running pip as the 'root' user can result in broken
permissions and conflicting behaviour with the system package manager. It is
recommended to use a virtual environment instead:
https://pip.pypa.io/warnings/venv
Requirement already satisfied: kaggle in /opt/conda/lib/python3.10/site-
packages (1.6.14)
Requirement already satisfied: six>=1.10 in /opt/conda/lib/python3.10/site-
packages (from kaggle) (1.16.0)
Requirement already satisfied: certifi>=2023.7.22 in
/opt/conda/lib/python3.10/site-packages (from kaggle) (2024.6.2)
Requirement already satisfied: python-dateutil in
```

```
/opt/conda/lib/python3.10/site-packages (from kaggle) (2.8.2)
Requirement already satisfied: requests in /opt/conda/lib/python3.10/site-
packages (from kaggle) (2.28.2)
Requirement already satisfied: tqdm in /opt/conda/lib/python3.10/site-packages
(from kaggle) (4.64.1)
Requirement already satisfied: python-slugify in /opt/conda/lib/python3.10/site-
packages (from kaggle) (8.0.4)
Requirement already satisfied: urllib3 in /opt/conda/lib/python3.10/site-
packages (from kaggle) (1.26.15)
Requirement already satisfied: bleach in /opt/conda/lib/python3.10/site-packages
(from kaggle) (6.0.0)
Requirement already satisfied: webencodings in /opt/conda/lib/python3.10/site-
packages (from bleach->kaggle) (0.5.1)
Requirement already satisfied: text-unidecode>=1.3 in
/opt/conda/lib/python3.10/site-packages (from python-slugify->kaggle) (1.3)
Requirement already satisfied: charset-normalizer<4,>=2 in
/opt/conda/lib/python3.10/site-packages (from requests->kaggle) (3.1.0)
Requirement already satisfied: idna<4,>=2.5 in /opt/conda/lib/python3.10/site-
packages (from requests->kaggle) (3.4)
WARNING: Error parsing requirements for jinja2: [Errno 2] No such file or
directory: '/opt/conda/lib/python3.10/site-packages/Jinja2-3.1.2.dist-
info/METADATA'
WARNING: Error parsing requirements for jsonschema: [Errno 2] No such
file or directory: '/opt/conda/lib/python3.10/site-
packages/jsonschema-4.17.3.dist-info/METADATA'
WARNING: Error parsing requirements for platformdirs: [Errno 2] No such
file or directory: '/opt/conda/lib/python3.10/site-
packages/platformdirs-3.5.0.dist-info/METADATA'
WARNING: Error parsing requirements for websocket-client: [Errno 2] No
such file or directory: '/opt/conda/lib/python3.10/site-
packages/websocket_client-1.5.1.dist-info/METADATA'
WARNING: Running pip as the 'root' user can result in broken
permissions and conflicting behaviour with the system package manager. It is
recommended to use a virtual environment instead:
https://pip.pypa.io/warnings/venv
```

[2]:

```python
import os
os.environ['KAGGLE_CONFIG_DIR'] = "kaggle.json"
```

[3]:
```
# !kaggle datasets download -d teamincribo/cyber-security-attacks
# !unzip cyber-security-attacks.zip
```

[4]:
```
# !gsutil cp cybersecurity_attacks.csv gs://workshop-3
```

[24]:
```
from pyspark.sql import SparkSession

spark = SparkSession.builder \
    .appName('VertexAI-Dataproc') \
    .master('yarn') \
    .config('spark.yarn.access.hadoopFileSystems', 'gs://workshop-3') \
    .getOrCreate()
```

[25]:
```
df = spark.read.csv('gs://workshop-3/cybersecurity_attacks.csv', header=True,␣
 ↪inferSchema=True)
df.describe()
```

[25]:
```
DataFrame[summary: string, Timestamp: string, Source IP Address: string,
Destination IP Address: string, Source Port: string, Destination Port: string,
Protocol: string, Packet Length: string, Packet Type: string, Traffic Type:
string, Payload Data: string, Malware Indicators: string, Anomaly Scores:
string, Alerts/Warnings: string, Attack Type: string, Attack Signature: string,
Action Taken: string, Severity Level: string, User Information: string, Device
Information: string, Network Segment: string, Geo-location Data: string, Proxy
Information: string, Firewall Logs: string, IDS/IPS Alerts: string, Log Source:
string]
```

[26]:
```
df.printSchema()
```

```
root
 |-- Timestamp: string (nullable = true)
 |-- Source IP Address: string (nullable = true)
 |-- Destination IP Address: string (nullable = true)
 |-- Source Port: string (nullable = true)
 |-- Destination Port: string (nullable = true)
 |-- Protocol: string (nullable = true)
 |-- Packet Length: string (nullable = true)
 |-- Packet Type: string (nullable = true)
 |-- Traffic Type: string (nullable = true)
 |-- Payload Data: string (nullable = true)
 |-- Malware Indicators: string (nullable = true)
 |-- Anomaly Scores: string (nullable = true)
 |-- Alerts/Warnings: string (nullable = true)
 |-- Attack Type: string (nullable = true)
 |-- Attack Signature: string (nullable = true)
 |-- Action Taken: string (nullable = true)
 |-- Severity Level: string (nullable = true)
 |-- User Information: string (nullable = true)
 |-- Device Information: string (nullable = true)
 |-- Network Segment: string (nullable = true)
 |-- Geo-location Data: string (nullable = true)
 |-- Proxy Information: string (nullable = true)
 |-- Firewall Logs: string (nullable = true)
 |-- IDS/IPS Alerts: string (nullable = true)
 |-- Log Source: string (nullable = true)
```

[47]:

```
table_name = "workshop3"
```

[48]:

```
gs_path = 'gs://workshop-3/cybersecurity_attacks.csv'
```

[49]:

```
spark.sql(f'CREATE TABLE IF NOT EXISTS {table_name} \
          USING csv \
          OPTIONS (HEADER true, INFERSCHEMA true, NULLVALUE "NA") \
          LOCATION "{gs_path}"')
```

[49]:

```
DataFrame[]
```

[50]:

```
selected_column_df_sql = spark.sql(f"SELECT Protocol FROM {table_name}")
```

[52]:

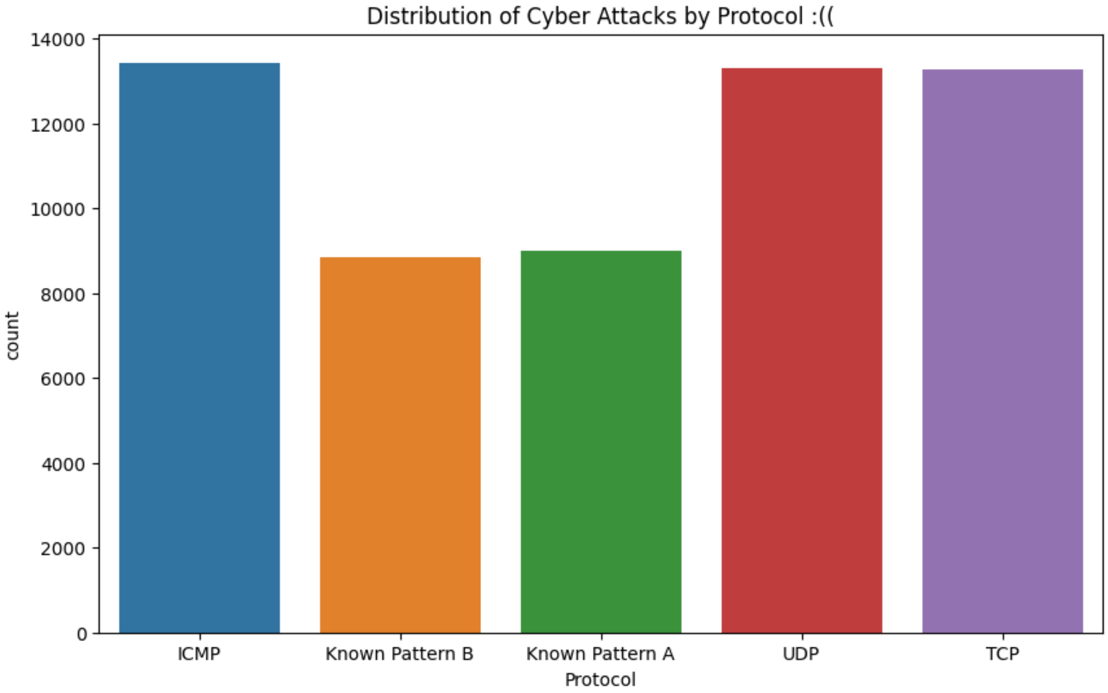
```
selected_column_df_sql.describe().show()
```

```
[Stage 43:===============================================>           (4 + 1) / 5]

+-------+--------+
|summary|Protocol|
+-------+--------+
|  count|   57826|
|   mean|    null|
| stddev|    null|
|    min|    ICMP|
|    max|     UDP|
+-------+--------+
```

[53]:

```
pdf = selected_column_df_sql.toPandas()
```

[56]:
```
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(10, 6))
sns.countplot(data=pdf, x='Protocol')
plt.title('Distribution of Cyber Attacks by Protocol :((')
plt.show()
```


Distribution of Cyber Attacks by Protocol :((
count
ICMP
Known Pattern B
Known Pattern A
UDP
TCP
Protocol