Advent of Cyber 2023

Log analysis O Data, All Ye Faithful

After yesterday's resounding success, McHoneyBell walks into AntarctiCrafts' office with a gleaming smile. She takes out her company-issued laptop from her knapsack and decides to check the news. "Traffic on the North-15 Highway? Glad I skied into work today," she boasts. A notification from the Best Festival Company's internal communication tool (HollyChat) pings.

It's another task. It reads, "The B-Team has been tasked with understanding the network of AntarctiCrafts' South Pole site". Taking a minute to think about the task ahead, McHoneyBell realises that AntarctiCrafts has no fancy technology that captures events on the network. "No tech? No problem!" exclaims McHoneyBell.

She decides to open up her Python terminal…

Learning Objectives

In today's task, you will:

- Get an introduction to what data science involves and how it can be applied in Cybersecurity
- Get a gentle (We promise) introduction to Python
- Get to work with some popular Python libraries such as Pandas and Matplotlib to crunch data
- Help McHoneyBell establish an understanding of AntarctiCrafts' network

Solving Day 2

For the second day, we were using Jupiter for data analysis

Answer the questions below

Open the notebook "Workbook" located in the directory "4_Capstone" on the VM. Use what you have learned today to analyse the packet capture.

No answer needed	Question Done

How many packets were captured (looking at the PacketNumber)?

100	Correct Answer	Hint

What IP address sent the most amount of traffic during the packet capture?

10.10.1.4	Correct Answer	Hint

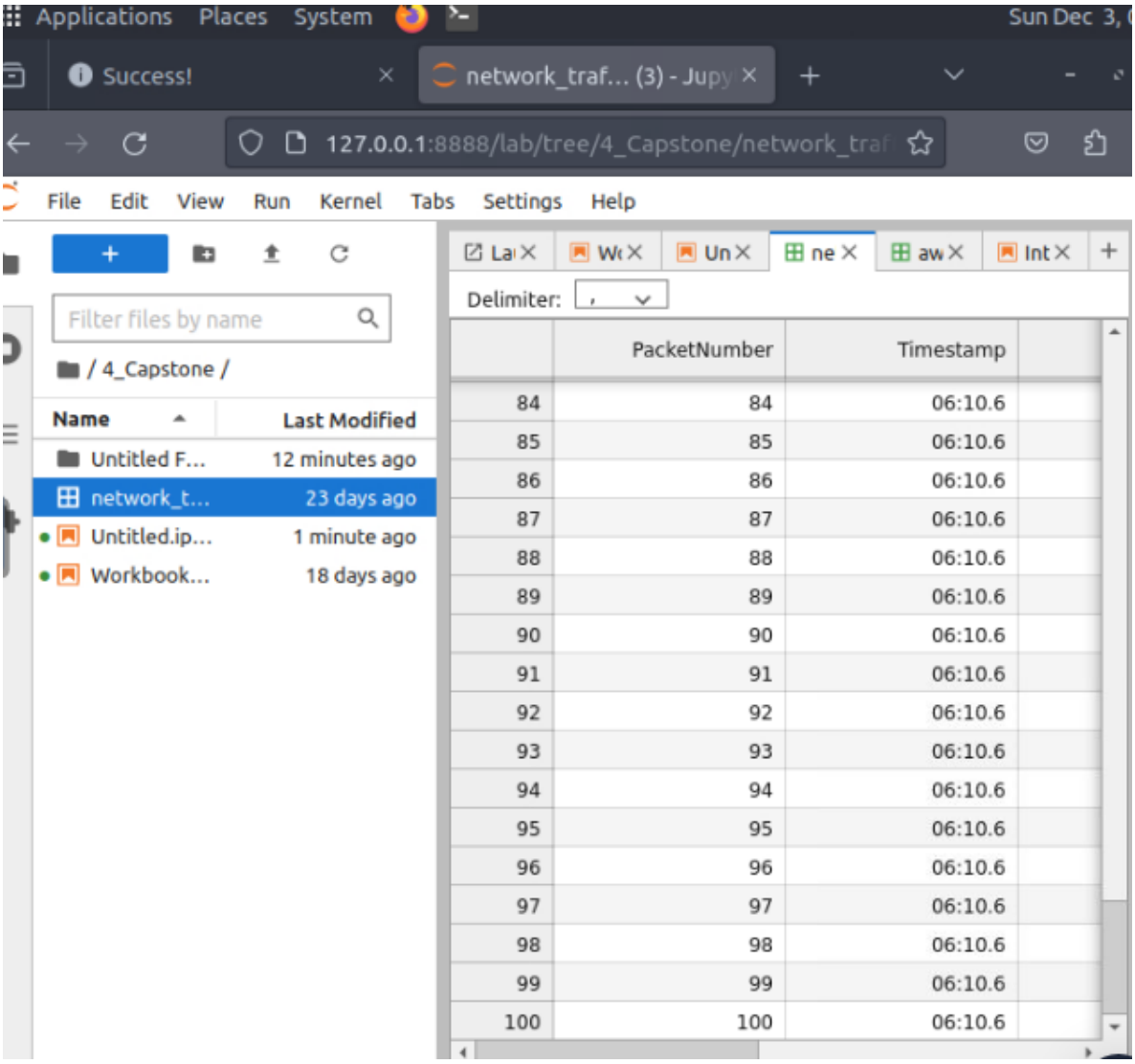
What was the most frequent protocol?

ICMP	Correct Answer	Hint

If you enjoyed today's task, check out the Intro to Log Analysis room.

No answer needed	Question Done

Within the virtual machine's folder in Jupiter, we found a set of data neatly organized into their respective columns.



What I did was create a new workbook, import the pandas and matplotlib.pyplot libraries, specifically using 'pd' and 'plt'. I opened the CSV file and at that moment, I noticed that we had a 'packetnumber,' so I examined the last 5 records.

[1]:

```
import pandas as pd
import matplotlib.pyplot as plt
```

[ ]:

[2]:

```
df = pd.read_csv('network_traffic.csv')
df.tail(5)
```

This revealed that 'packetnumber' 100 was the latest. Here came the initial question.

cketNumber	Timestamp	Source	Destination	Protocol
96	06:10.6	10.10.1.8	10.10.1.3	DNS
97	06:10.6	10.10.1.1	10.10.1.3	ICMP
98	06:10.6	10.10.1.3	10.10.1.3	DNS
99	06:10.6	10.10.1.4	10.10.1.3	TCP
100	06:10.6	10.10.1.5	10.10.1.2	ICMP

Another way to achieve the same result was through 'df.count,' which counted all the packets.

```
df.count()
```

```
[4]:

PacketNumber    100
Timestamp       100
Source          100
Destination     100
Protocol        100
dtype: int64
```

To find the IP with the highest traffic, it was sufficient to group them by their source.

```
df.groupby(['Source']).size()
```

```
[6]:

Source
10.10.1.1      8
10.10.1.10     8
10.10.1.2     12
10.10.1.3     13
10.10.1.4     15
10.10.1.5      5
10.10.1.6     14
10.10.1.7      5
10.10.1.8      9
10.10.1.9     11
dtype: int64

[7]:
```

To identify the most used protocol, we could do exactly the same, but this time grouping them by their protocol instead of their source.

```
df.groupby(['Protocol']).size()
```

[7]:

```
Protocol
DNS     25
HTTP    24
ICMP    27
TCP     24
dtype: int64
```