جامعة الأخوين
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵏ ⵓⵅⴰⵡⵉⵏ
AL AKHAWAYN UNIVERSITY

CSC 5356 01 - Data Engineering and Visualization

Data pipeline

Supervisor: Dr. Tajjeeddine Rachidi

Definition:

Data pipeline:

Is a set of data processing elements connected in series, where the output of one element is the input of the next one, the operations can be: moving, sorting, reformatting, analyzing and reporting in order to derive value from data. The elements of a pipeline are often executed in parallel or in time-sliced fashion

MAWILab:

Is a database that assists researchers to evaluate their traffic anomaly detection methods

Neo4j:

Transactional database with native graph storage and processing

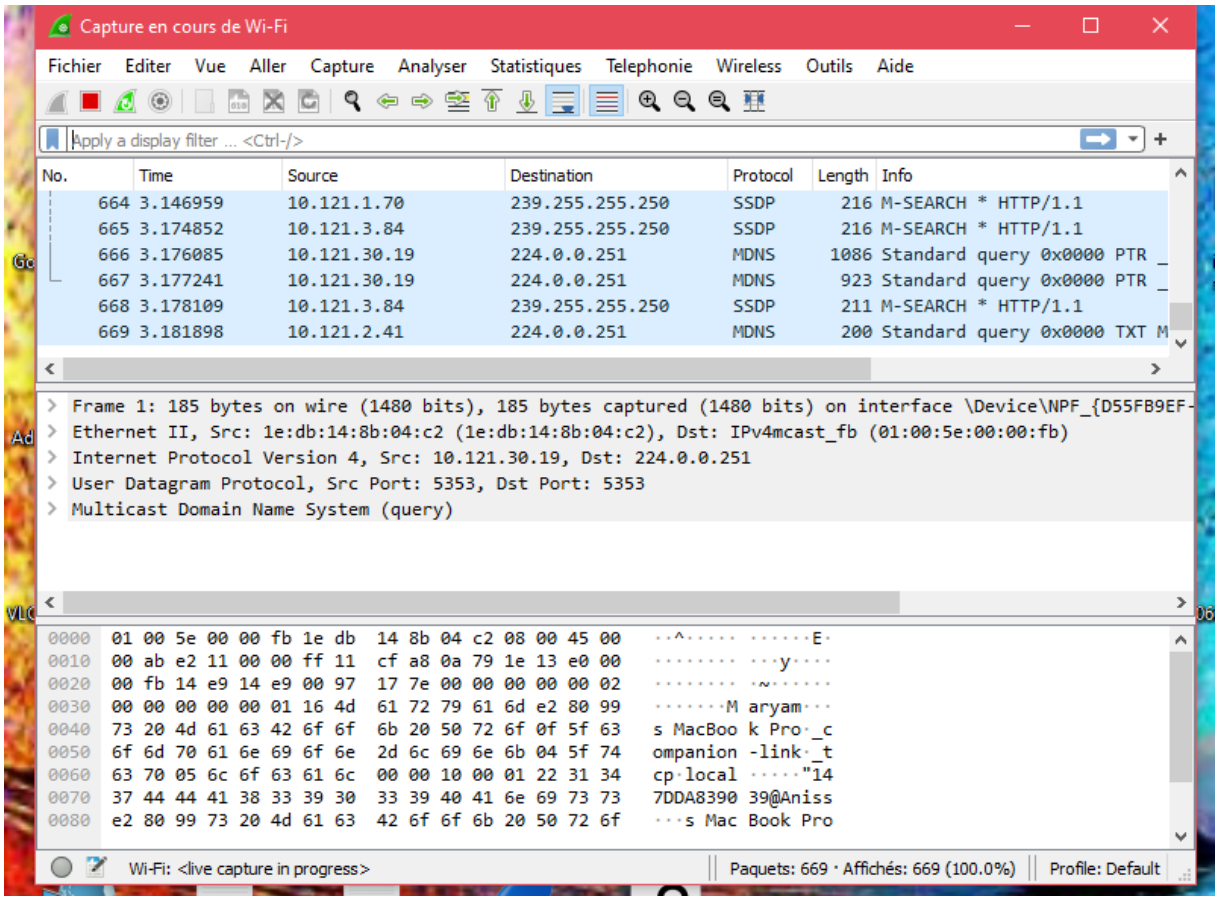
Apache beam:

Programming model to define and execute data processing pipelines, including ETL, batch and stream processing

Question1:

Using python, we will try to integrate network traces to the neo4j Database

First, the data set in MAWI was too large (7g, 5g), so I used the dataset captured from my own computer using wireshark (it is a tool that allows packet traces to be sniffed, captured and analyzed)



- Libraries and packages used:

PYPCAPFILE, SCAPY, PY2NEO

- After imports and installations, I load my pcap file in google drive because I am using google collab

We can read the pcap file by those lines of code:

```
# read pcap file in binary format
pcapfile = open('/content/drive/MyDrive/sample3.pcap', 'rb')
# load the pcapfile into a variable
capfile = savefile.load_savefile(pcapfile, layers=2, verbose=True)
```

- we connect to our Neo4j database after installing neo4j by `pip install neo4j`:

```
from neo4j import GraphDatabase
uri = "bolt://127.0.0.1:7687"
neo4jGraph = GraphDatabase.driver(uri, auth=("neo4j", "123"))
```

- After that we We go through the packets.sessions dictionary, we split the packet (into a list of strings).

Next, we create two nodes of type “Host”, one for the sender, and one for the receiver,the nodes’ name is the IP address.

After that we can see the nodes and the relationships in our neo4j database.

Question2:

Because pcap is not readable in apache beam, I converted my data set to csv, and loaded it in google drive since I used google collab also

- We install and import apache beam

```
! pip   install apache-beam[interactive]
import   apache_beam as beam
```

- We then declare our data pipeline

```
pipeline1 = beam.Pipeline()
```

- Beam has the beam.dataframe.io.read_csv function that emulates pandas.read_csv, but returns a deferred Beam DataFrame.

We are using Interactive Beam to use collect to bring a Beam DataFrame into local memory as a Pandas DataFrame

```
beam_df = pipeline | 'Read CSV' >> beam.dataframe.io.read_csv('solar_ev
ents.csv')

ib.collect(beam_df)
```

- now beam_df is our new dataframe, we will try to integrate it in our neo4j database using the following code :

```
from neo4j import GraphDatabase

transaction_list = beam_df.values.tolist()

transaction_execution_commands = []

for i in transaction_list:
```

```
    neo4j_create_statemenet = "create (t:Transaction {num:" + str(i[0])
 +", time:  " + str(i[1]) +", source: " + str(i[2]) +", destination: '"
 + str(i[3]) + "'})"
    transaction_execution_commands.append(neo4j_create_statemenet)


def execute_transactions(transaction_execution_commands):
    data_base_connection = GraphDatabase.driver(uri = "bolt://localhost
:7687", auth=("neo4j", "123"))
    session = data_base_connection.session()
    for i in transaction_execution_commands:
        session.run(i)


execute_transactions(transaction_execution_commands)
```