

The background of the slide is a blurred image of a person in a light blue shirt and tie, holding a pen and pointing at a tablet. Overlaid on this are various financial data visualizations, including a candlestick chart on the left with numerical values like 928,545, 28,545, 128,150, 548,125, 215,810, 9,007, 337,296, 124,545, and 289,000. A line graph with blue and red dots is also visible. The overall theme is big data and analytics.

Big Data Analytics WS 20/21

Final Presentation

AGENDA

- 01** TASK
- 02** COLLABORATION
- 03** DATA
- 04** IT LANDSCAPE
- 05** DATA PREPARATION
- 06** DATA ANALYSIS



TASK

The client **Frachtwerk GmbH** asked us to **perform pattern recognition** on a dataset of approximately 15 Mio data points of internet traffic retrieved by their **firewall** between Sept-Dec 2020. The deliverables were the analysis of the data, the identification of security threat patterns and a visual interface to view the most relevant results.

COLLABORATION | Trello



Backlog

Hand-in-deliverable_2: Technical documentation/Source Code

Jan 28

1

Hand-in-deliverable_1: Conceptual documentation

Jan 28

Hand-in-deliverable_3: Presentation

Jan 28

Software recommendations for Conclusion in technical doc

Add column internal IP address or external IP address Origin and Target

+ Add another card

ToDo

Data Analysis

J

IT Landscape

LW

Data Preparation

LW

Business Understanding

LW

Evaluation / Lessons Learned

J

Setup Technical Documentation

J

+ Add another card

In Progress

+ Add a card

Done 22.01

Setup Grafana with citus datasource

LW

acc Spalte analyse?

Setup technical documentation

Describe IT Landscape in concept. doc.

J

IP: Write algorithm to add country to IP address

1

3

J

Describe Data Preparation in concept. doc.

LW

Integrate classes into main script & run full test

LW

DATA | pfsense.log

```
7 11:45:47 monitoring rsyslogd: [origin software="rsyslogd" swVersion="8.1901.0"
rsyslog.com"] start
7 11:45:50 gw01.extranet.frachtwerk.de filterlog: 200,,,1599469328,vtnet3,match,b
172.23.0.2,23.54.112.189,51636,443,0,R,3551626076,,,0,,
```

← Debug log entry

← Normal log entry

Most common fields*

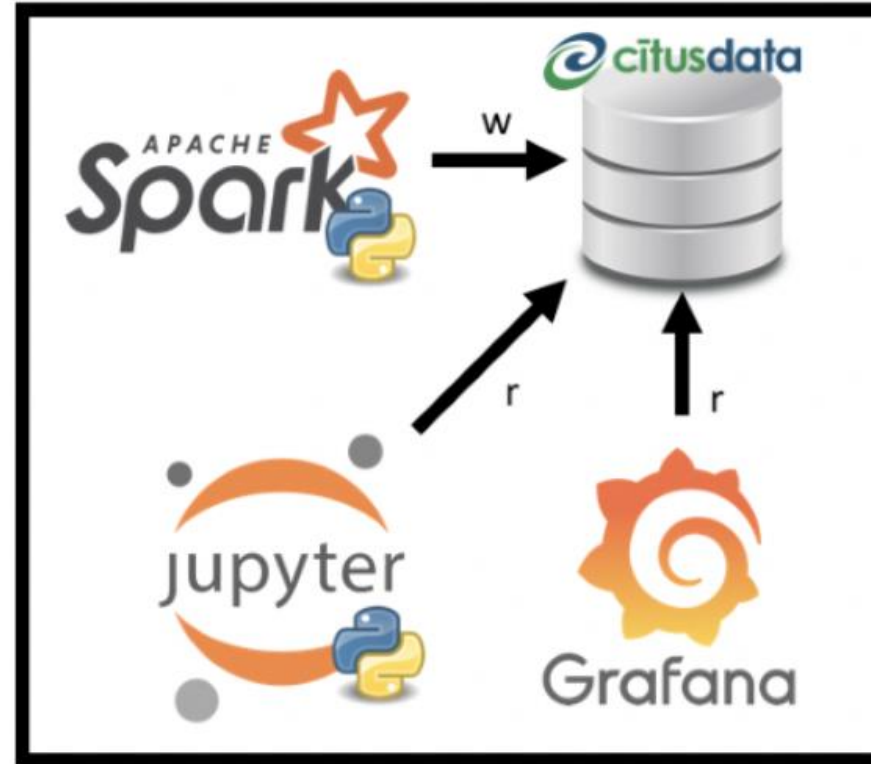
Name	Definition
Action	Shows the rule that generated the log entry (e.g. pass or block)
Time	The time that the packet arrived.
Interface	Where the packet entered the firewall (e.g. em0).
Rule Number	The firewall rule ID number which generated the log entry, if available.
Source	The source IP address and port.
Destination	The destination IP address and port.
Protocol	The protocol of the packet, e.g. ICMP, TCP, UDP, etc.

Field name	Approx. unique entries	Data type
Source IP	309240	string
Source port	65551	int
Destination IP	633	string
Destination port	65514	int
Protocol	9	string

*Full list of fields: <https://docs.netgate.com/pfsense/en/latest/monitoring/logs/raw-filter-format.html>
(Accessed 28. Jan 2021)

IT LANDSCAPE

Source Code
Management



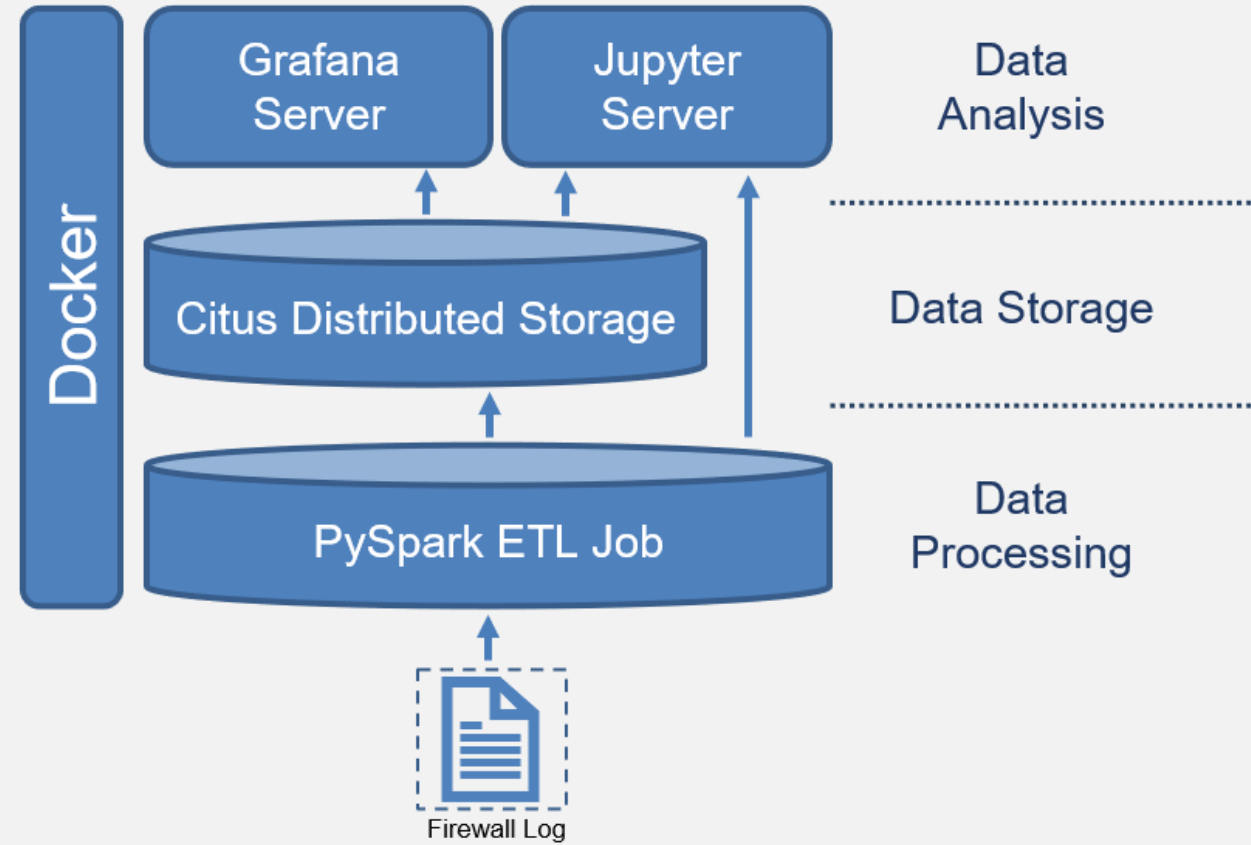
*Image Sources: <https://www.python.org>, <https://gitlab.com>, <https://www.docker.com>,
<https://spark.apache.org>, <https://grafana.com>, <https://www.citusdata.com> (Accessed 28.01.2021)

DATA PIPELINE

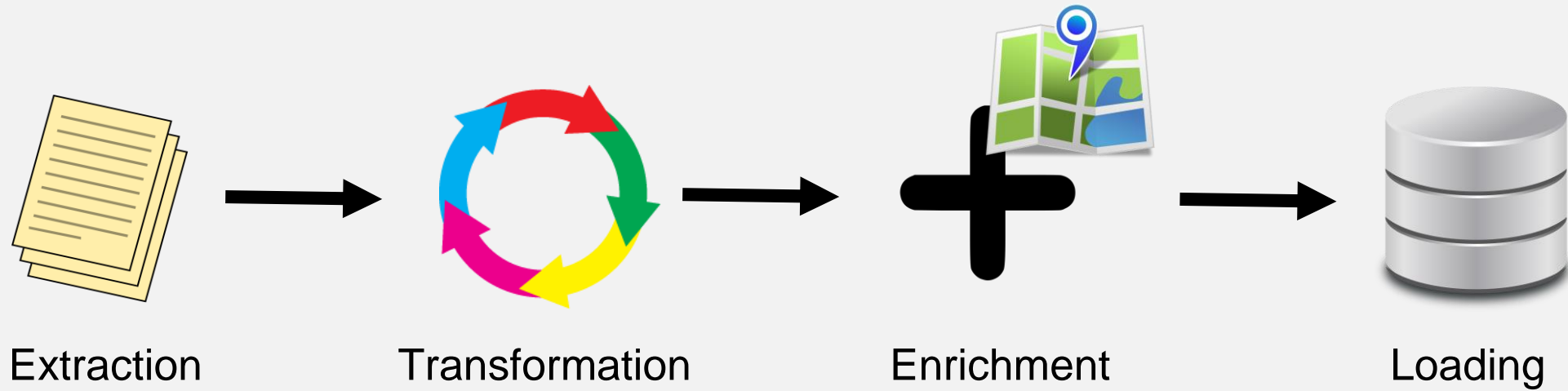
Reliability

Scalability

Maintainability



DATA PREPARATION



*Image Sources: Retrieved from bing online in PowerPoint. All images are under the Creative Commons license

DATA ANALYSIS | Jupyter Notebook → Timestamp

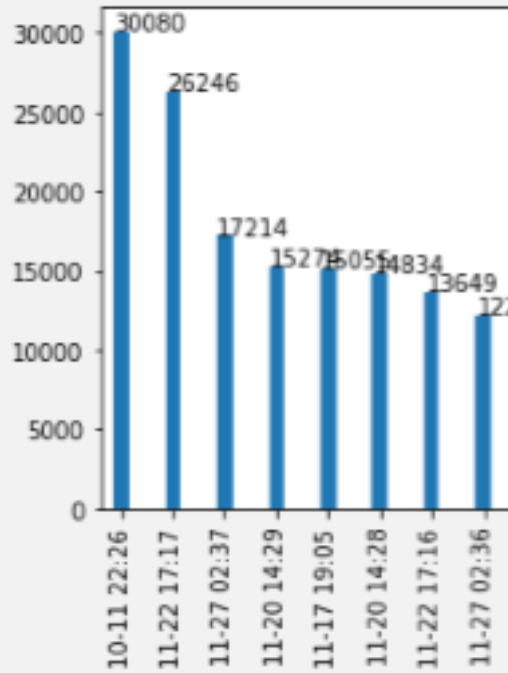
```
0    Sep  7 11:45:50 gw01.extranet.frachtwerk.de
1    Sep  7 11:45:52 gw01.extranet.frachtwerk.de
2    Sep  7 11:46:00 gw01.extranet.frachtwerk.de
3    Sep  7 11:46:00 gw01.extranet.frachtwerk.de
4    Sep  7 11:46:12 gw01.extranet.frachtwerk.de
```



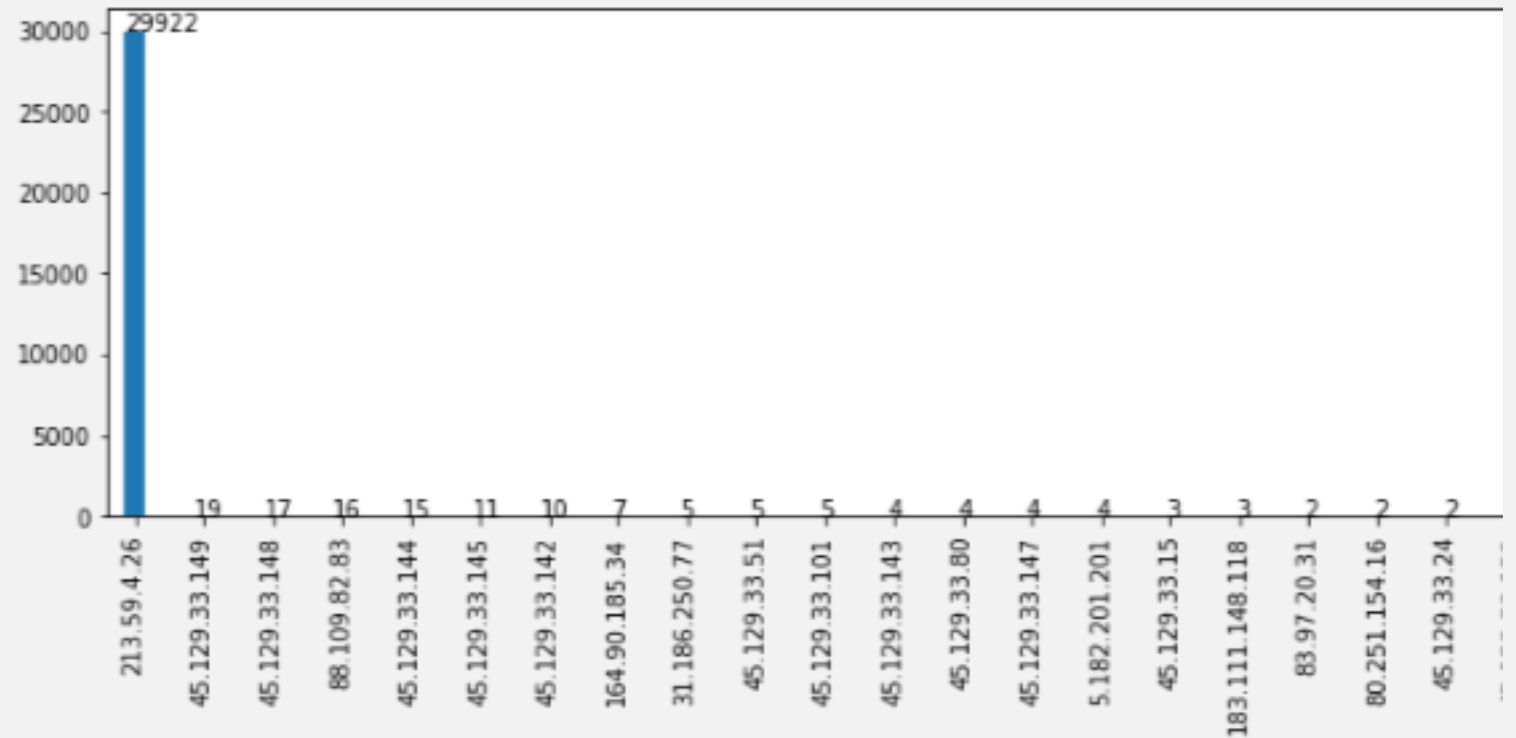
log aggregation per minute

```
count    130593.000000
mean      117.871310
std       256.359653
```

DATA ANALYSIS | Jupyter Notebook → Timestamp



Logs per minute desc



10-11 22:26 source IP frequency desc

DATA ANALYSIS | Jupyter Notebook → Timestamp

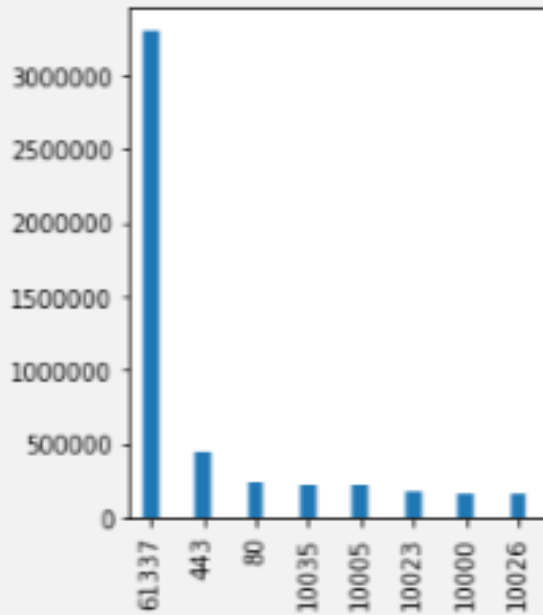
Assumption:
Server
performs
DoS attacks

 **213.59.4.26**

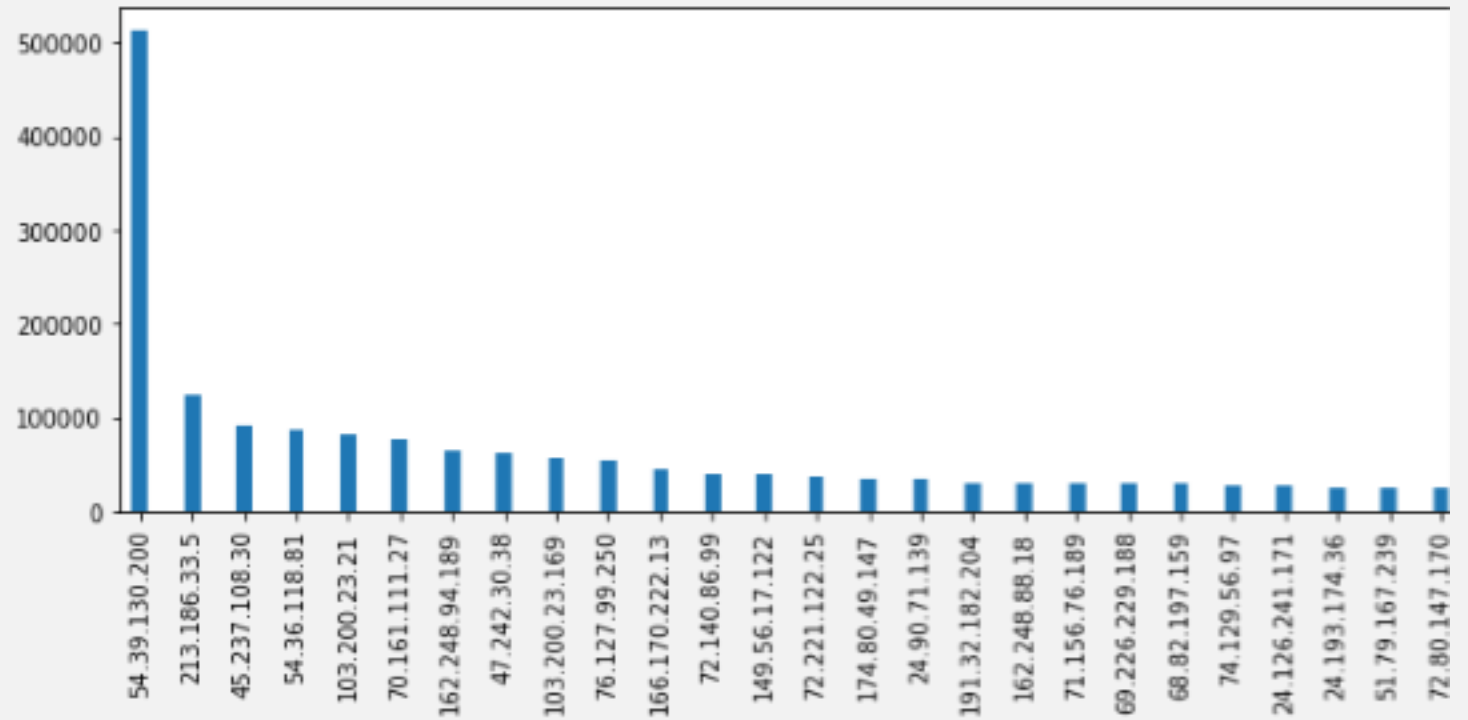
Country	Russia
Organization	Rostelecom
ISP	Rostelecom
Last Update	2021-01-19T09:50:17.354054
ASN	AS8342

Shodan.io (Accessed 28.01.21)

DATA ANALYSIS | Jupyter Notebook → Ports




Ports frequency desc



Port 61337 frequency desc

DATA ANALYSIS | Jupyter Notebook → Ports

Assumption:
Server
infected by
trojan and
part of botnet

 54.39.130.200 firewallroozservers.bairesrp.net	
self-signed	
<hr/>	
City	Victoria
Country	Canada
Organization	OVH SAS
ISP	OVH SAS
Last Update	2021-01-28T13:48:39.643447
Hostnames	firewallroozservers.bairesrp.net
ASN	AS16276

Shodan.io (Accessed 28.01.21)

DATA ANALYSIS | Jupyter Notebook → Bandwidth

```
0    Sep  7 11:45:50 gw01.extranet.frachtwerk.de
1    Sep  7 11:45:52 gw01.extranet.frachtwerk.de
2    Sep  7 11:46:00 gw01.extranet.frachtwerk.de
3    Sep  7 11:46:00 gw01.extranet.frachtwerk.de
4    Sep  7 11:46:12 gw01.extranet.frachtwerk.de
```

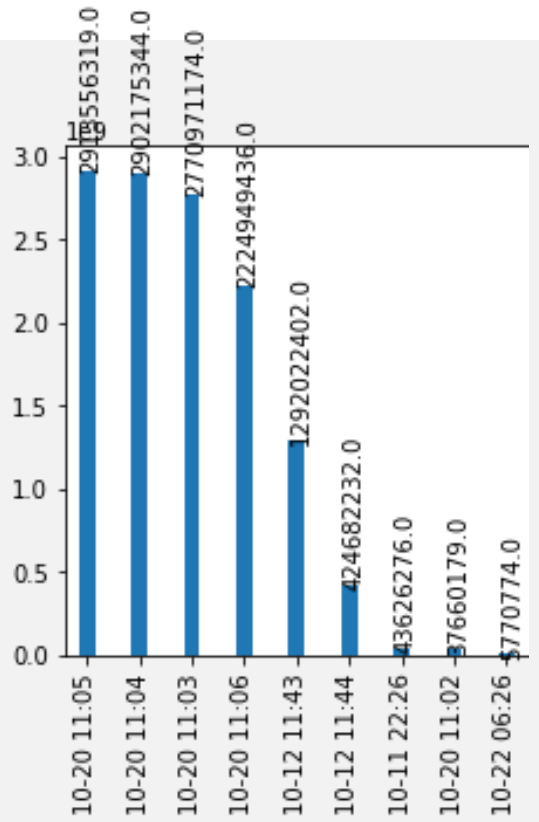


log aggregation per minute and
data_length

```
count    1.305930e+05
mean     9.897071e+04
std      1.550425e+07
```

98971

DATA ANALYSIS | Jupyter Notebook → Bandwidth

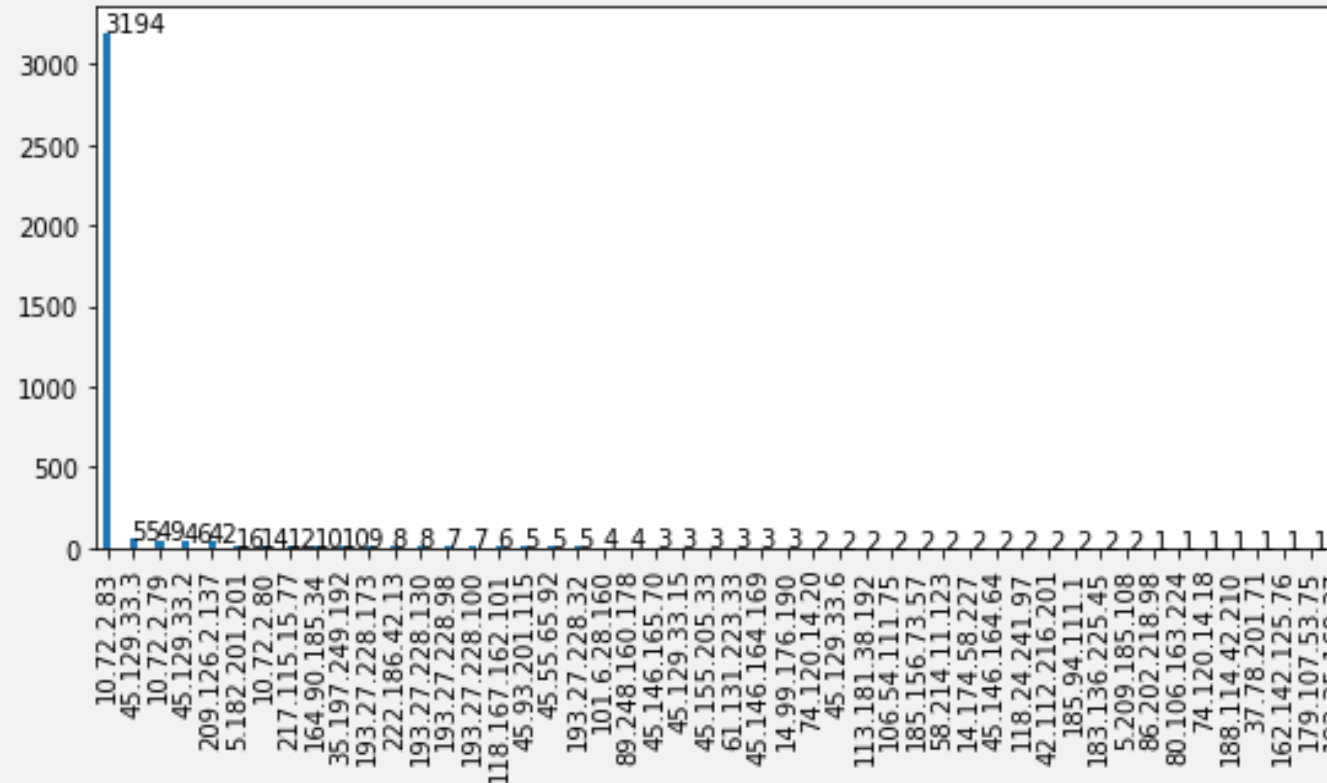


Nr	Time	Pattern
1	10-20 11:05	1 & 2 & 3 & 4 & 8 Possible denial of service attack
2	10-20 11:04	1 & 2 & 3 & 4 & 8 Possible denial of service attack
3	10-20 11:03	1 & 2 & 3 & 4 & 8 Possible denial of service attack
4	10-20 11:06	1 & 2 & 3 & 4 & 8 Possible denial of service attack
5	10-12 11:43	5 & 6 Possible denial of service attack
6	10-12 11:44	5 & 6 Possible denial of service attack
7	10-11 22:26	Possible denial of service attack
8	10-20 11:02	1 & 2 & 3 & 4 & 8 Possible denial of service attack

Data_length per minute desc

DATA ANALYSIS | Jupyter Notebook → Bandwidth

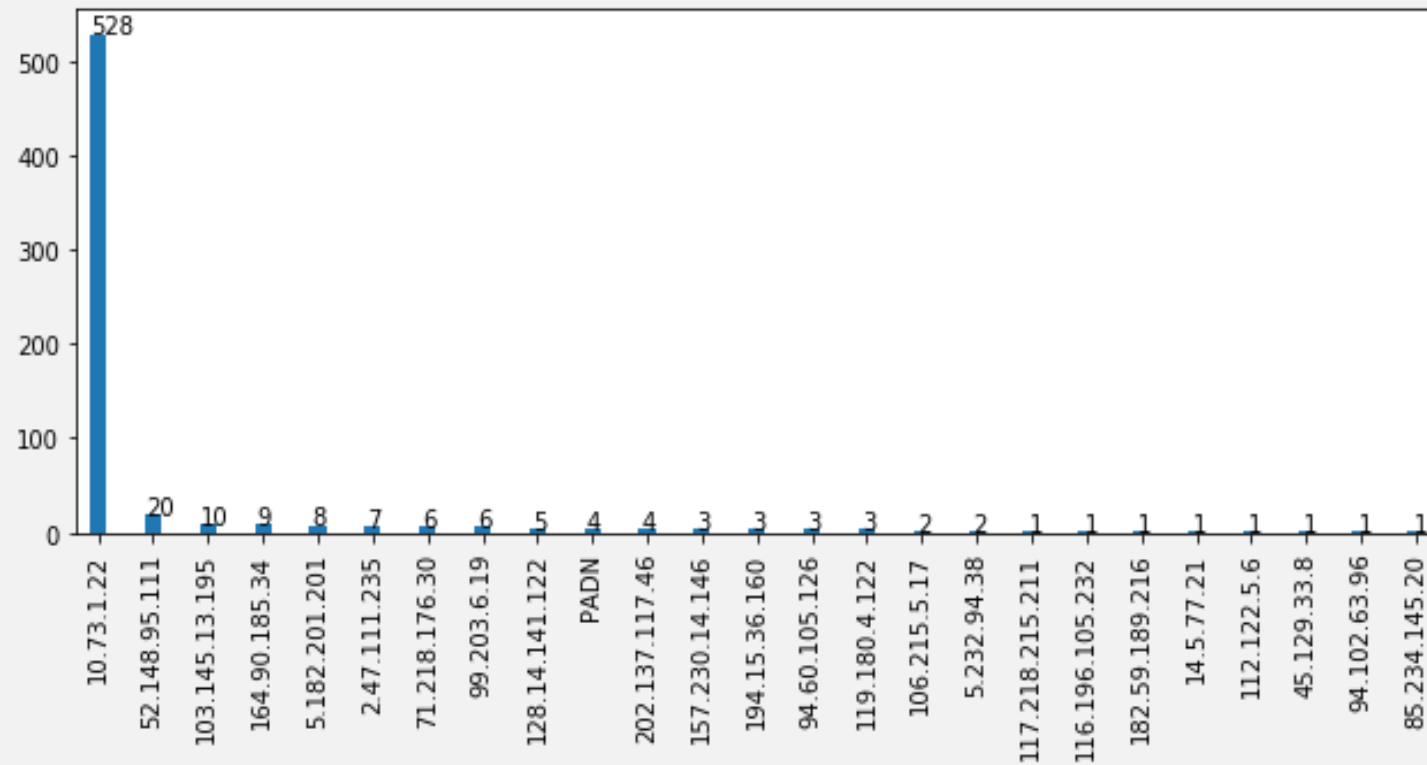
Assumption:
Internal
Server
infected by
trojan



Source IPs causing data_length spike
1 & 2 & 3 & 4 & 8

DATA ANALYSIS | Jupyter Notebook → Bandwidth

Assumption:
Internal
Server
infected by
trojan



Source IPs causing data_length spike
5 & 6

DATA ANALYSIS | Grafana



VIELEN DANK!

LORENZ WACKENHUT
JONAS FÄHRMANN
DENA KARINI

