Building the High Speed Cyber Security Data Pipeline Using Apache NiFi

Praveen Kanumarlapudi

Cyber Security

60% of Small Businesses Fold Within 6 Months of a Cyber Attack.

How to make it success ?

Global Security Key Stake Holders

Security Operations Center

An information security operations center ("ISOC" or "SOC") is a facility where enterprise information systems (websites, applications, databases, data centers and servers, networks, desktops and other endpoints) are monitored, assessed, and defended.

Technology : SIEM

Data Scientists

Security data scientists have the skills to understand complex algorithms and build advanced models for threat and anomaly detection and applying these concepts to real security data sets in single or clustered environments.

Technology : Python, R, Big Data, Spark/Scala or MATLAB…

Data Analysts

Map and trace the data from system to system for solving a given business or incident problem.
Design and create data reports using various reporting tools that help business executive to make better decisions.
Implements new metrics for business (KPIs)

Technology : SQL, SIEM, Big Data, Reporting tools

Executives

CSO's, CISO's

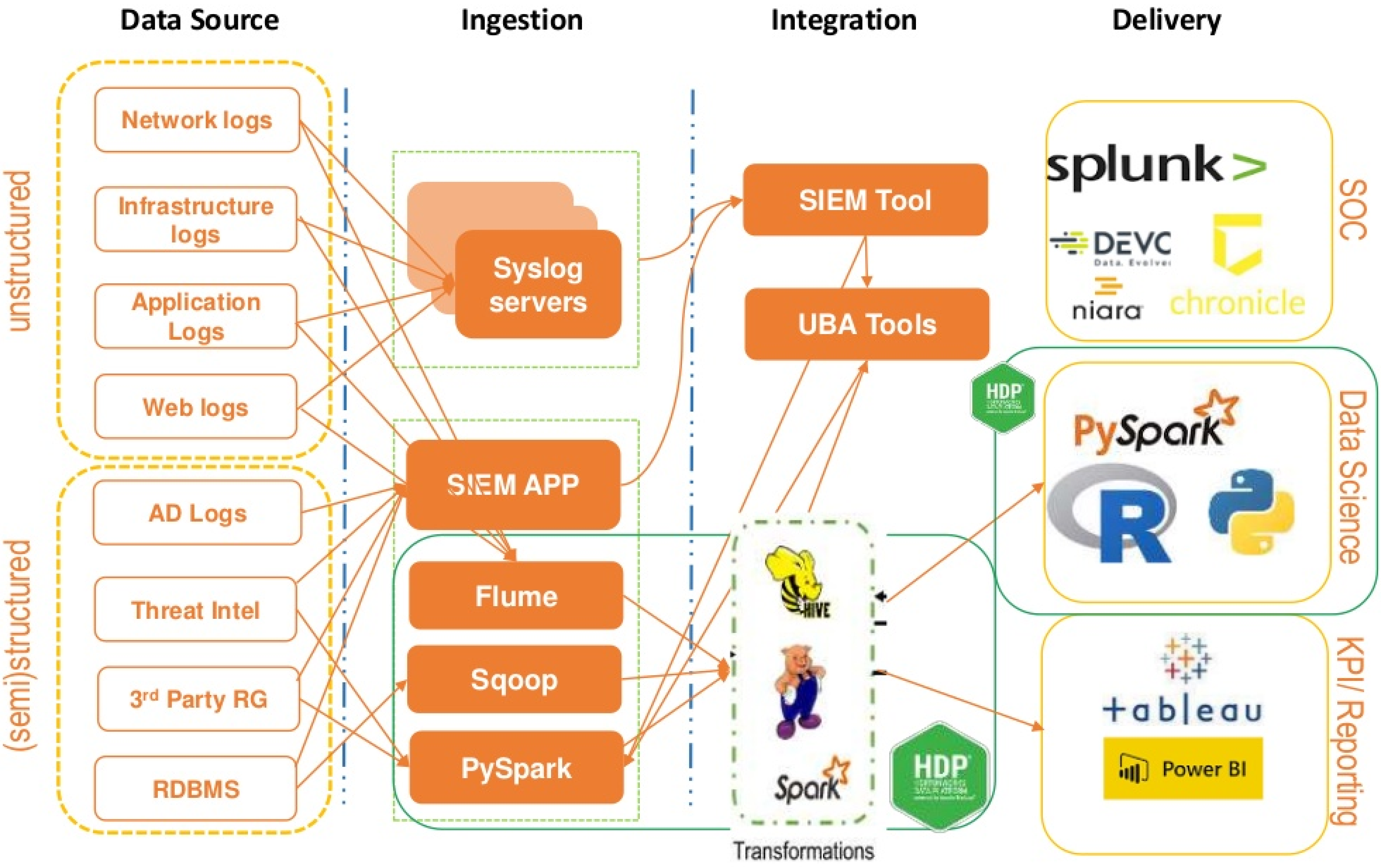
Cyber Security ‘BIG data’ challenges

- Speed , Volume and Variety
    - Data Ingestion
    - Cleansing
    - Transformation

- data reliance
    - Executives – KPI Metrics
    - Data scientists
    - SOC
    - Data Analysts

- Real-Time context

A couple of years Ago !



Data Source

Ingestion

Integration

Delivery

unstructured

Network logs

Infrastructure logs

Application Logs

Web logs

(semi)structured

AD Logs

Threat Intel

3rd Party RG

RDBMS

Syslog servers

SIEM APP

Flume

Sqoop

PySpark

SIEM Tool

UBA Tools

Transformations

SOC

Data Science

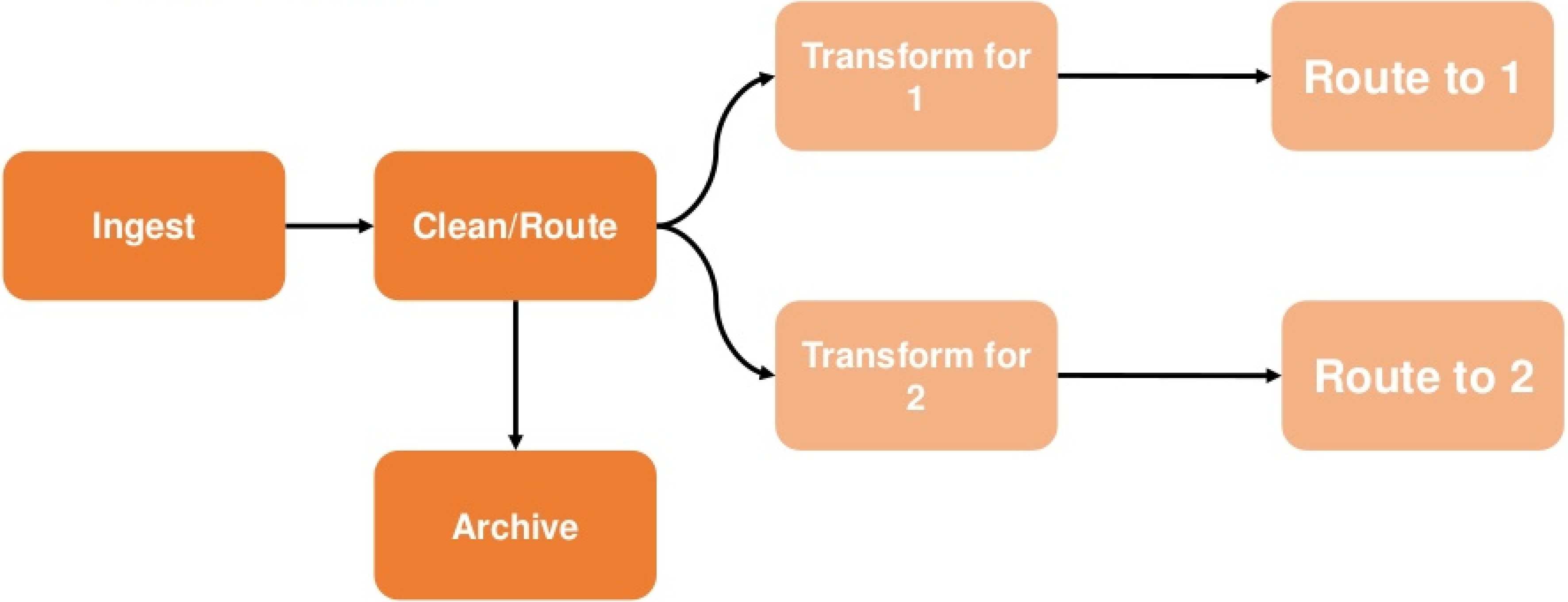
KPI/ Reporting

Challenges

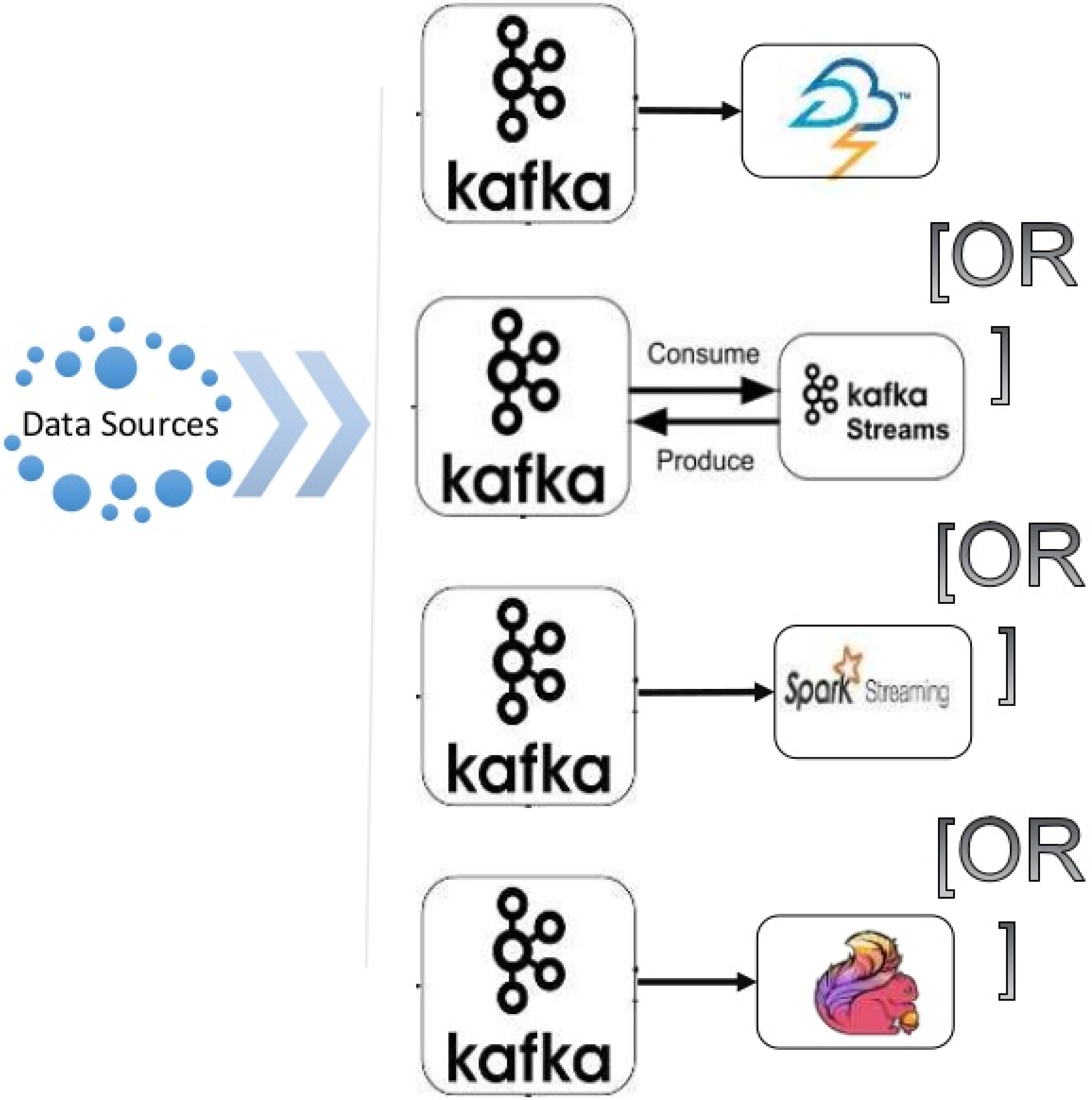
- Complexity of Architecture
- Debugging
- Data Source Dependencies
- Lack of Centralized logging
- Multiple Data Copies
- Stress on Network
- Transformations with respect to destination

Solution Framework

- ✓ Single Data entry point – avoids network traffic and duplicate data flowing around
- ✓ Transformations according destination – reduces the reliance on source
- ✓ Should be capable of handling different formats and different sources

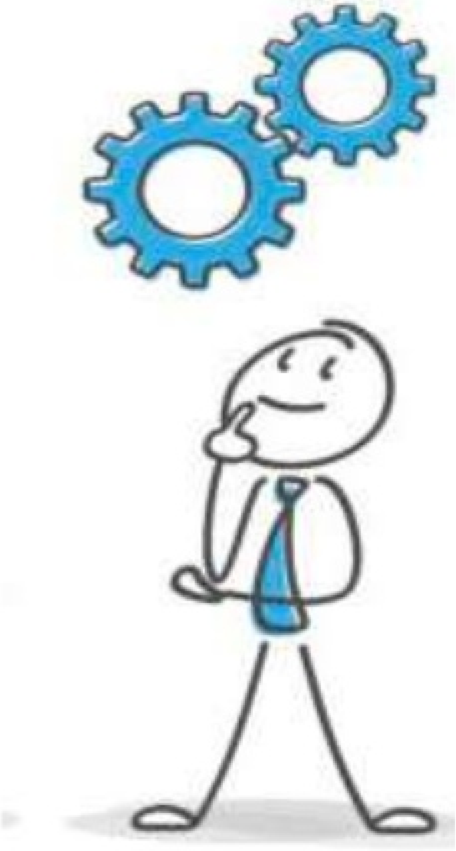


Deployment Models



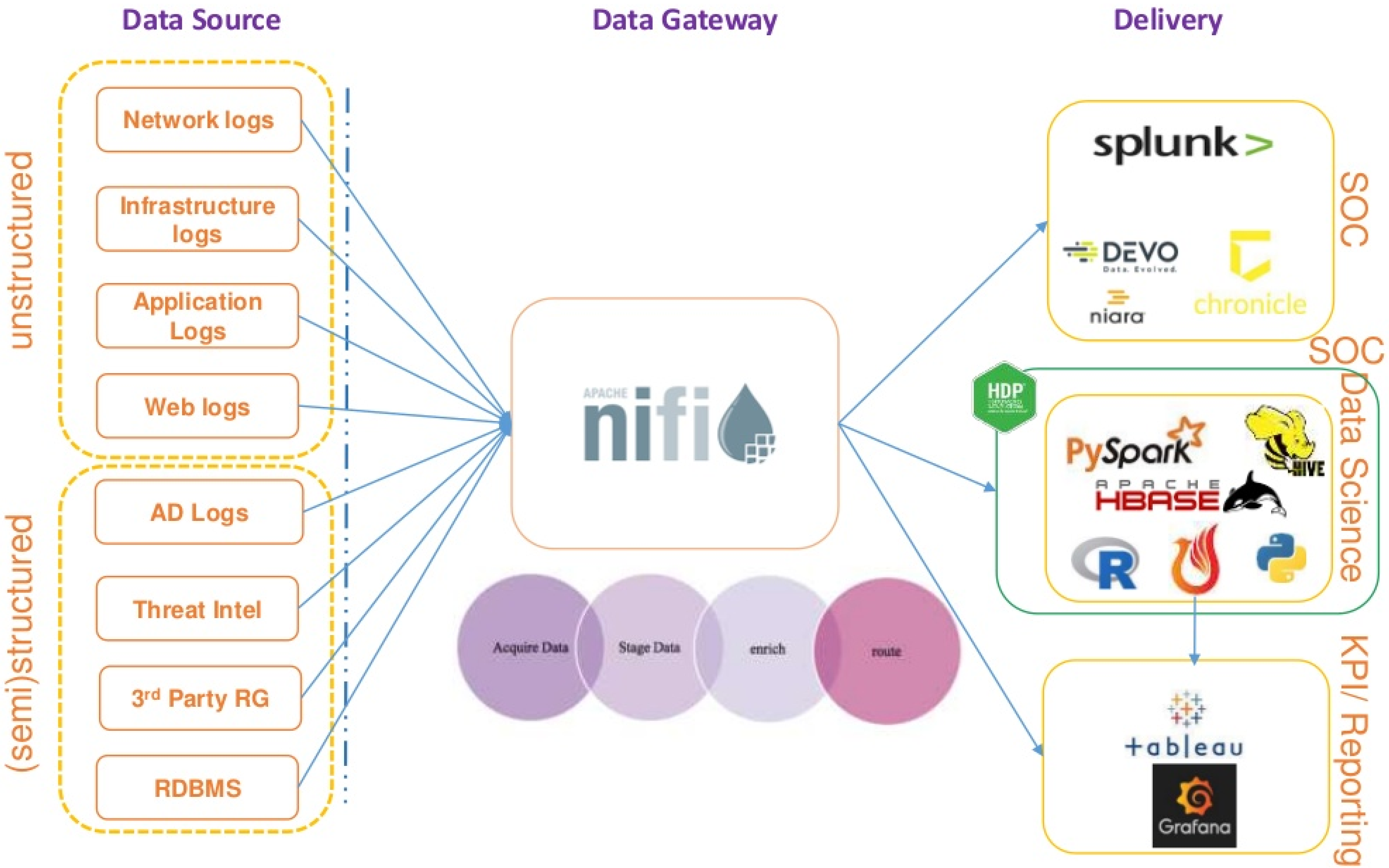
Challenges

- ✓ Good architectural understanding of all systems
- ✓ Good amount of coding effort
- ✓ Long development hours
- ✓ Maintenance overheads
- ✓ Maintain the sync between the systems
- ✓ Provenance



- Guaranteed delivery
- Processors that supports multiple formats
- Ease to develop the flows and deploy in minutes
- Open Source and rich community

The Data Gateway



Data Source

Data Gateway

Delivery

unstructured

Network logs

Infrastructure logs

Application Logs

Web logs

(semi)structured

AD Logs

Threat Intel

3rd Party RG

RDBMS

Acquire Data

Stage Data

enrich

route

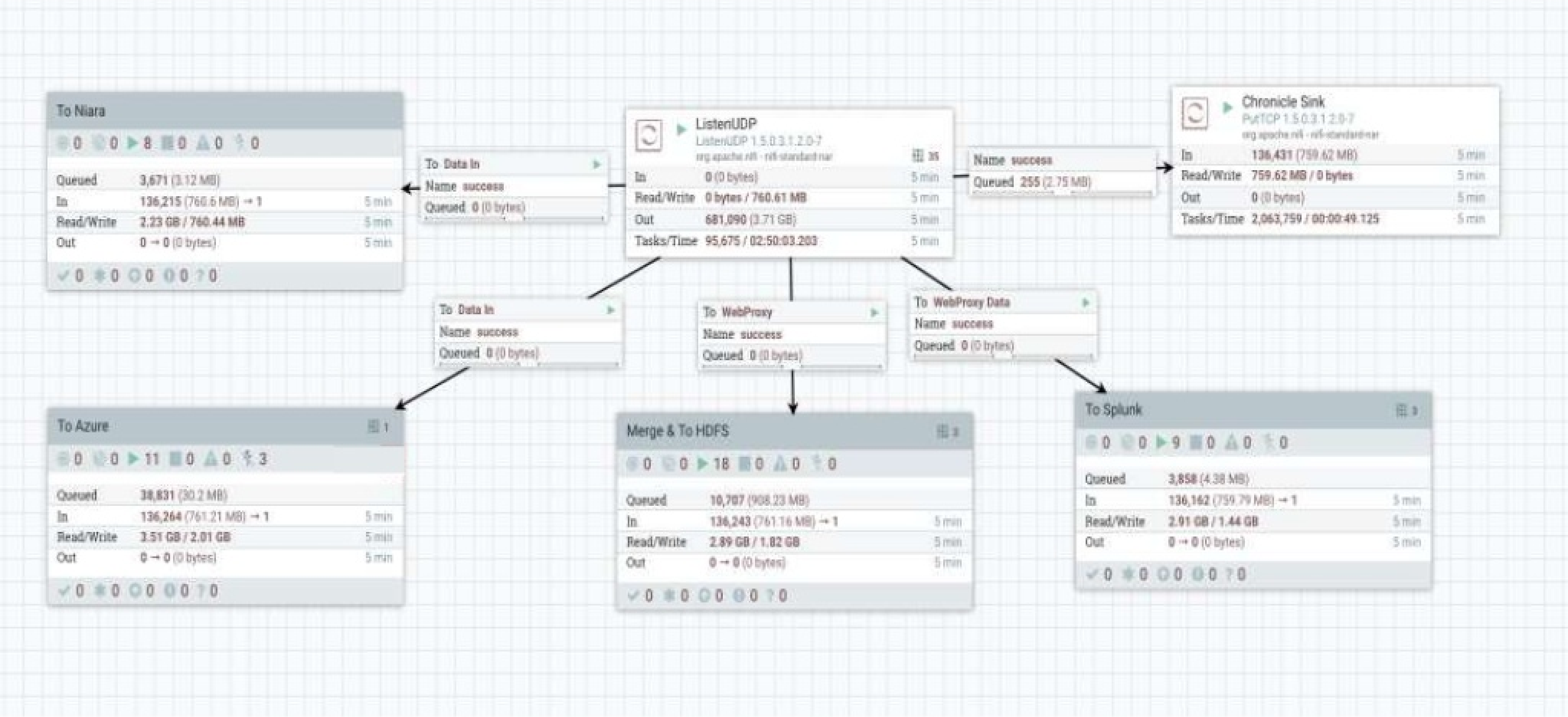
SOC

SOC

Data Science

KPI/ Reporting

Sample Flow



To Niara

0 0 ▶ 8 ■ 0 ⚠ 0 0

Queued 3,671 (3.12 MB)
In 136,215 (760.6 MB) → 1 5 min
Read/Write 2.23 GB / 760.44 MB 5 min
Out 0 → 0 (0 bytes) 5 min

✓ 0 0 0 0 ? 0

To Data In
Name success
Queued 0 (0 bytes)

ListenUDP
ListenUDP 1.5.0.3.1.2.0-7
org.apache.nifi - nifi-standard-nar
35

In 0 (0 bytes) 5 min
Read/Write 0 bytes / 760.61 MB 5 min
Out 681,090 (3.71 GB) 5 min
Tasks/Time 95,675 / 02:50:03.203 5 min

Name success
Queued 255 (2.75 MB)

Chronicle Sink
PutTCP 1.5.0.3.1.2.0-7
org.apache.nifi - nifi-standard-nar

In 136,431 (759.62 MB) 5 min
Read/Write 759.62 MB / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 2,063,759 / 00:00:49.125 5 min

To Data In
Name success
Queued 0 (0 bytes)

To WebProxy
Name success
Queued 0 (0 bytes)

To WebProxy Data
Name success
Queued 0 (0 bytes)

To Azure
1

0 0 ▶ 11 ■ 0 ⚠ 0 3

Queued 38,831 (30.2 MB)
In 136,264 (761.21 MB) → 1 5 min
Read/Write 3.51 GB / 2.01 GB 5 min
Out 0 → 0 (0 bytes) 5 min

✓ 0 0 0 0 ? 0

Merge & To HDFS
3

0 0 ▶ 18 ■ 0 ⚠ 0 0

Queued 10,707 (908.23 MB)
In 136,243 (761.16 MB) → 1 5 min
Read/Write 2.89 GB / 1.82 GB 5 min
Out 0 → 0 (0 bytes) 5 min

✓ 0 0 0 0 ? 0

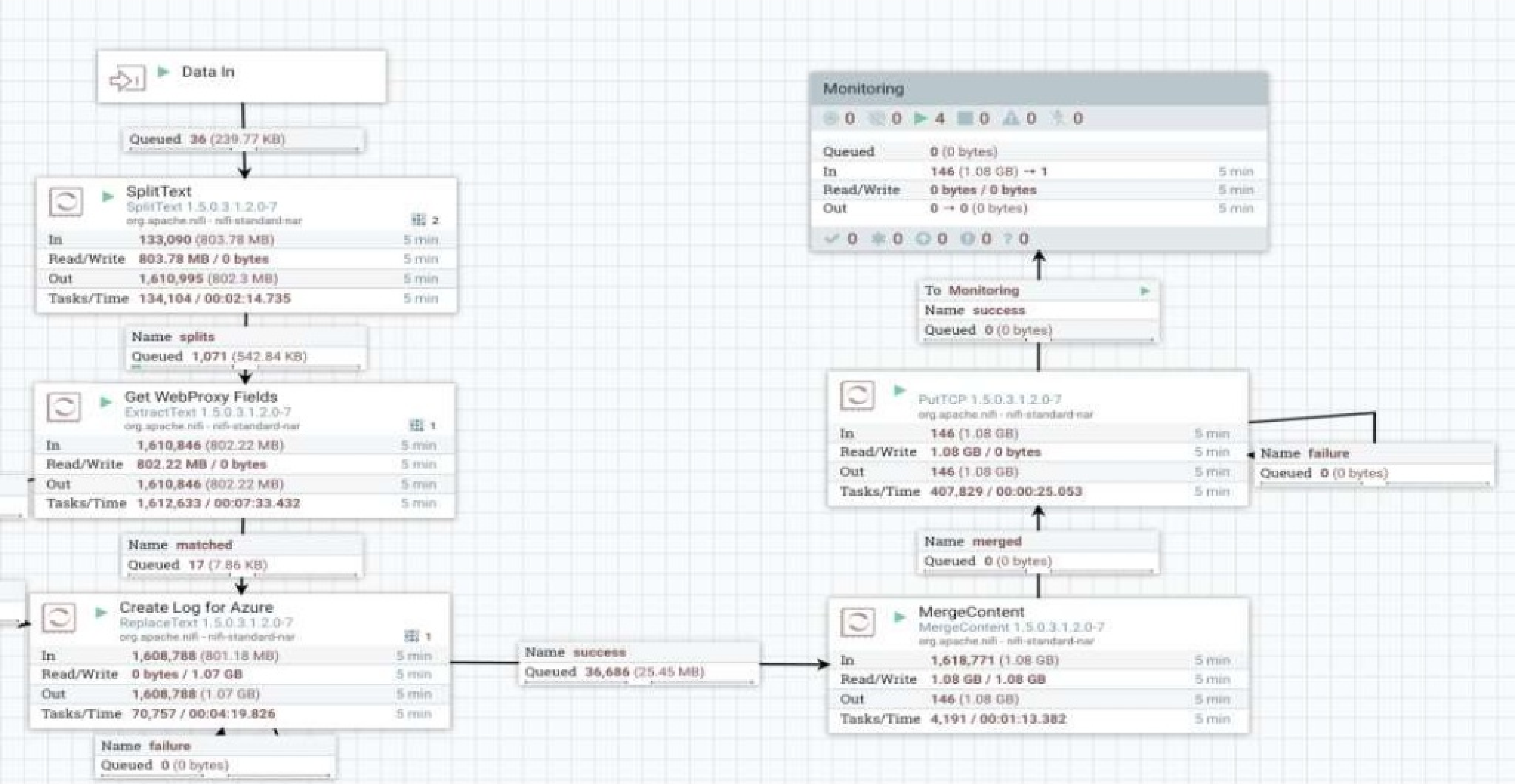
To Splunk
3

0 0 ▶ 9 ■ 0 ⚠ 0 0

Queued 3,858 (4.38 MB)
In 136,162 (759.79 MB) → 1 5 min
Read/Write 2.91 GB / 1.44 GB 5 min
Out 0 → 0 (0 bytes) 5 min

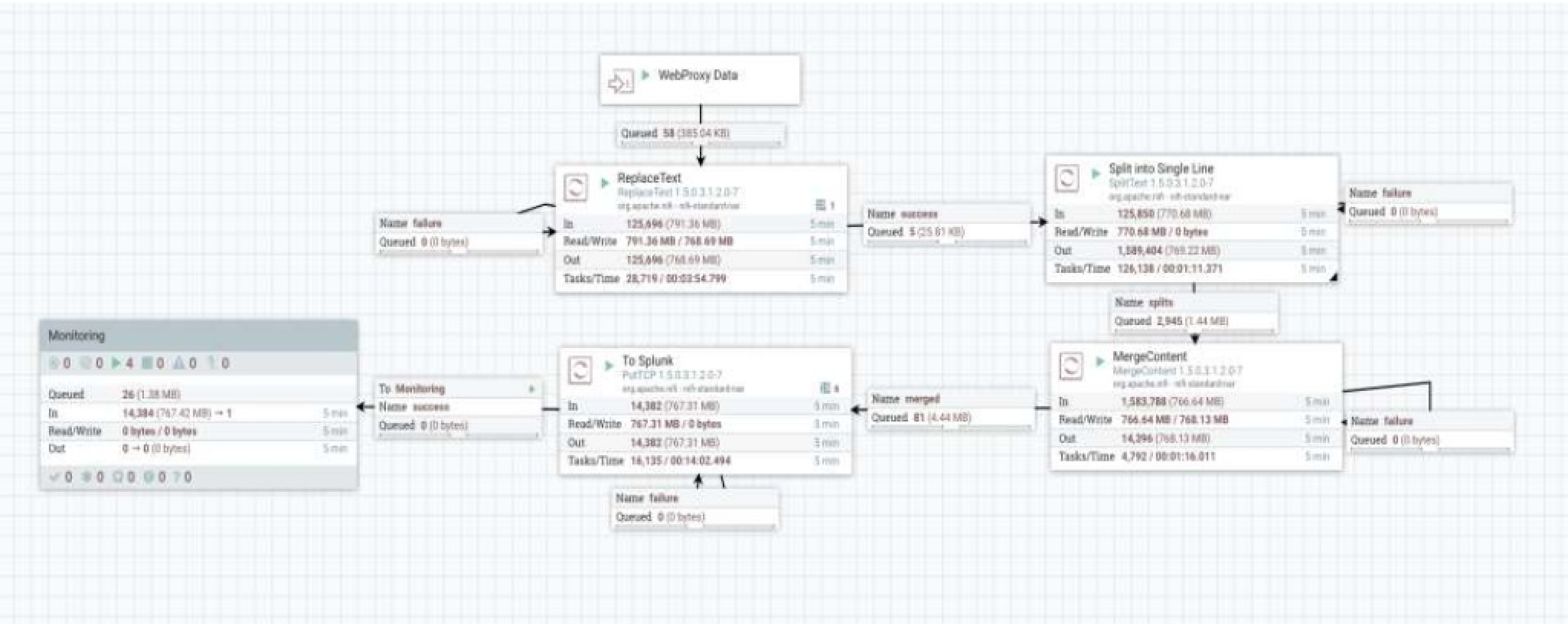
✓ 0 0 0 0 ? 0

To Azure



```
Data In
Queued  36 (239.77 KB)

SplitText
SplitText 1.5.0.3.1.2.0-7
org.apache.nifi - nifi-standard-nar                2
In           133,090 (803.78 MB)                   5 min
Read/Write   803.78 MB / 0 bytes                   5 min
Out          1,610,995 (802.3 MB)                  5 min
Tasks/Time   134,104 / 00:02:14.735                5 min

Name  splits
Queued  1,071 (542.84 KB)

Get WebProxy Fields
ExtractText 1.5.0.3.1.2.0-7
org.apache.nifi - nifi-standard-nar                1
In           1,610,846 (802.22 MB)                 5 min
Read/Write   802.22 MB / 0 bytes                   5 min
Out          1,610,846 (802.22 MB)                 5 min
Tasks/Time   1,612,633 / 00:07:33.432              5 min

Name  matched
Queued  17 (7.86 KB)

Create Log for Azure
ReplaceText 1.5.0.3.1.2.0-7
org.apache.nifi - nifi-standard-nar                1
In           1,608,788 (801.18 MB)                 5 min
Read/Write   0 bytes / 1.07 GB                     5 min
Out          1,608,788 (1.07 GB)                   5 min
Tasks/Time   70,757 / 00:04:19.826                 5 min

Name  failure
Queued  0 (0 bytes)

Name  success
Queued  36,686 (25.45 MB)

MergeContent
MergeContent 1.5.0.3.1.2.0-7
org.apache.nifi - nifi-standard-nar
In           1,618,771 (1.08 GB)                   5 min
Read/Write   1.08 GB / 1.08 GB                     5 min
Out          146 (1.08 GB)                         5 min
Tasks/Time   4,191 / 00:01:13.382                  5 min

Name  merged
Queued  0 (0 bytes)

PutTCP 1.5.0.3.1.2.0-7
org.apache.nifi - nifi-standard-nar
In           146 (1.08 GB)                         5 min
Read/Write   1.08 GB / 0 bytes                     5 min
Out          146 (1.08 GB)                         5 min
Tasks/Time   407,829 / 00:00:25.053                5 min

Name  failure
Queued  0 (0 bytes)

To  Monitoring
Name  success
Queued  0 (0 bytes)

Monitoring
0   0   4   0   0   0
Queued       0 (0 bytes)
In           146 (1.08 GB) → 1                     5 min
Read/Write   0 bytes / 0 bytes                     5 min
Out          0 → 0 (0 bytes)                       5 min
✓ 0   0   0   0   ? 0
```

To Splunk



WebProxy Data

Queued 58 (385.04 KB)

ReplaceText
ReplaceText 1.5.0.3.1.2.0-7
org.apache.nifi - nifi-standard-nar

In	125,696 (791.36 MB)	5 min
Read/Write	791.36 MB / 768.69 MB	5 min
Out	125,696 (768.69 MB)	5 min
Tasks/Time	28,719 / 00:03:54.799	5 min

Name failure
Queued 0 (0 bytes)

Name success
Queued 5 (25.81 KB)

Split into Single Line
SplitText 1.5.0.3.1.2.0-7
org.apache.nifi - nifi-standard-nar

In	125,850 (770.68 MB)	5 min
Read/Write	770.68 MB / 0 bytes	5 min
Out	1,589,404 (769.22 MB)	5 min
Tasks/Time	126,138 / 00:01:11.371	5 min

Name failure
Queued 0 (0 bytes)

Name splits
Queued 2,945 (1.44 MB)

MergeContent
MergeContent 1.5.0.3.1.2.0-7
org.apache.nifi - nifi-standard-nar

In	1,583,788 (766.64 MB)	5 min
Read/Write	766.64 MB / 768.13 MB	5 min
Out	14,396 (768.13 MB)	5 min
Tasks/Time	4,792 / 00:01:16.011	5 min

Name failure
Queued 0 (0 bytes)

Name merged
Queued 81 (4.44 MB)

To Splunk
PutTCP 1.5.0.3.1.2.0-7
org.apache.nifi - nifi-standard-nar

In	14,382 (767.31 MB)	5 min
Read/Write	767.31 MB / 0 bytes	5 min
Out	14,382 (767.31 MB)	5 min
Tasks/Time	16,135 / 00:14:02.494	5 min

Name failure
Queued 0 (0 bytes)

To Monitoring
Name success
Queued 0 (0 bytes)

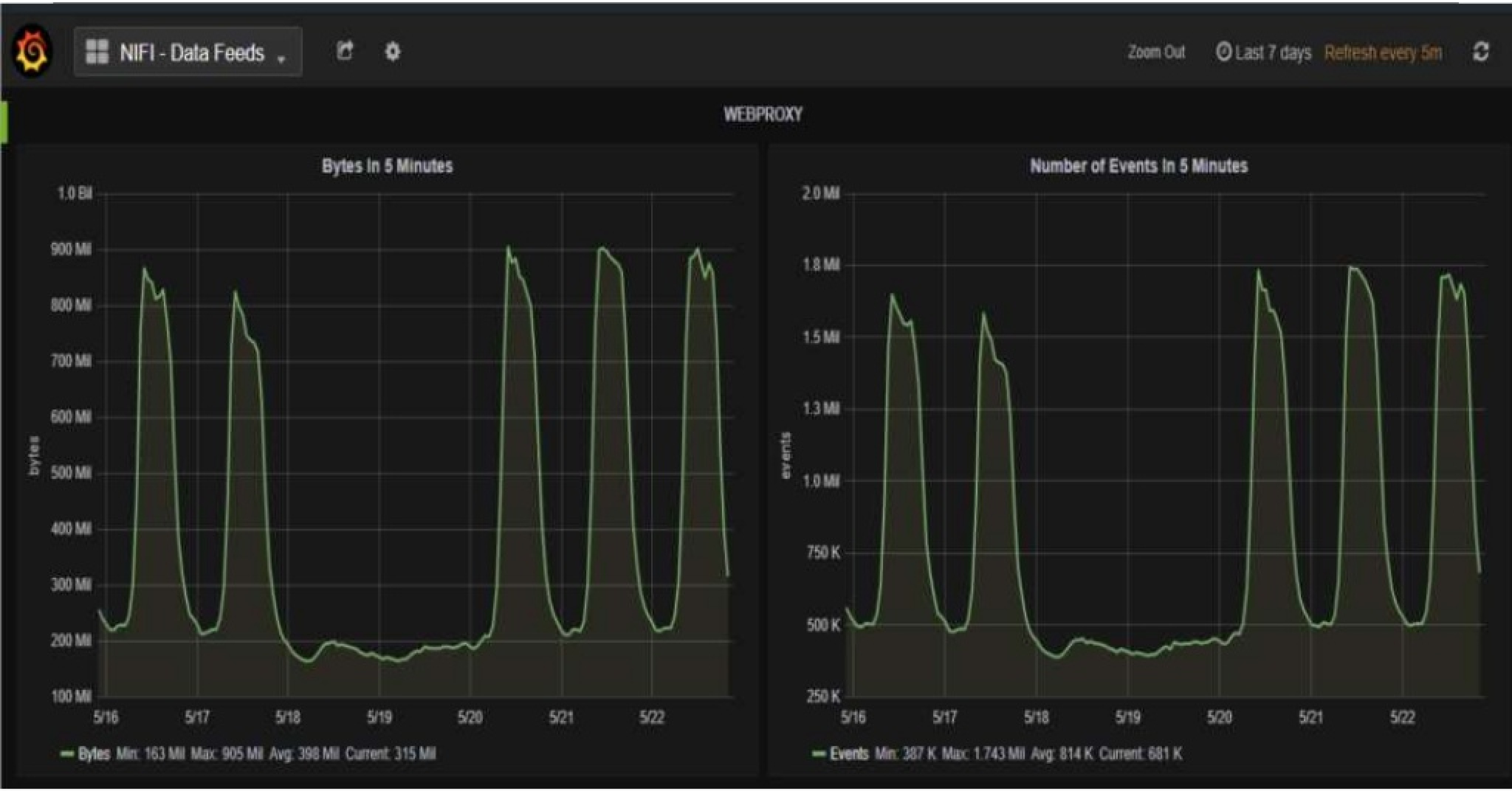
Monitoring

0 0 ▶ 4 ■ 0 ⚠ 0 0

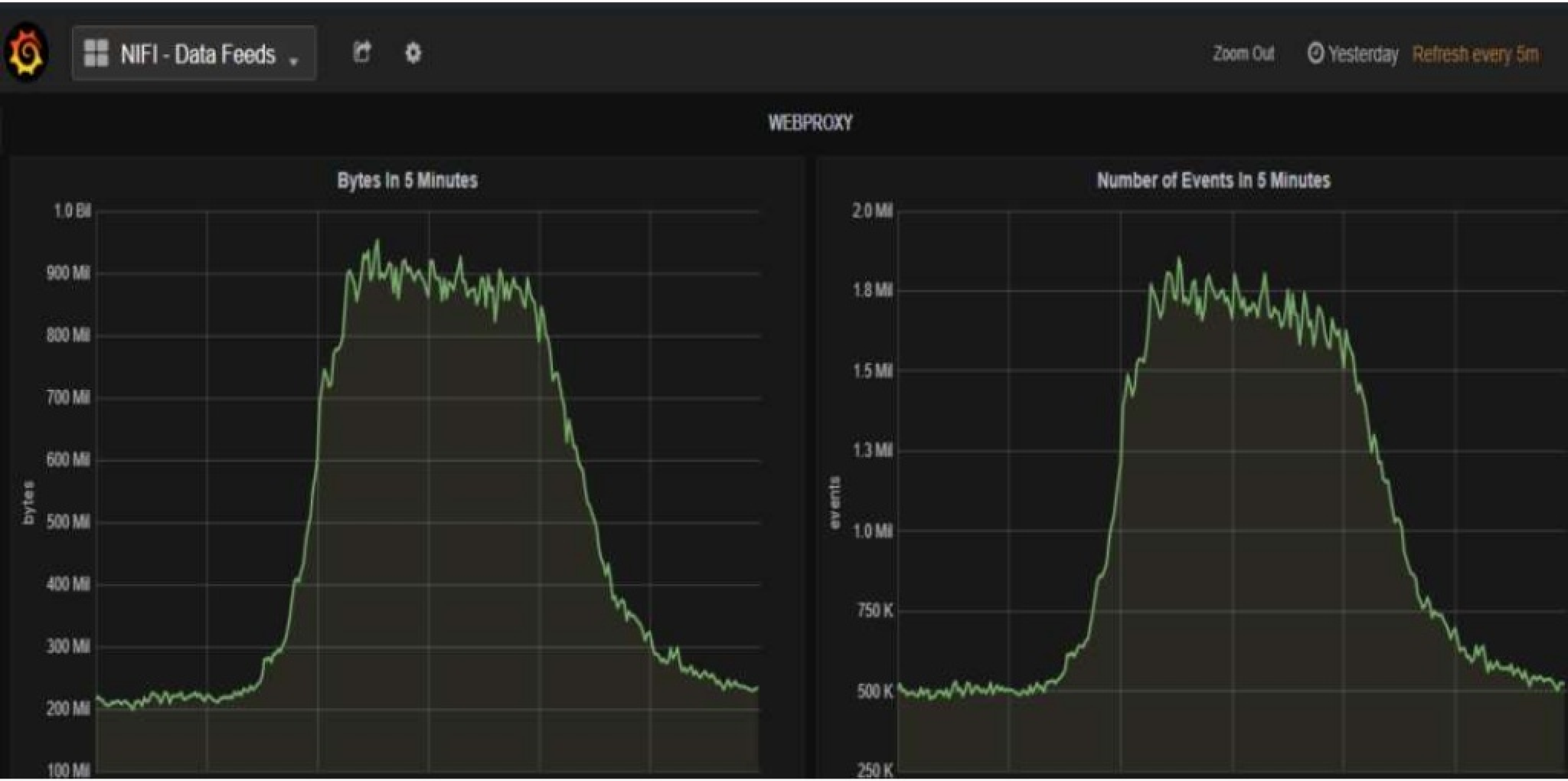
Queued	26 (1.38 MB)	
In	14,384 (767.42 MB) → 1	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 → 0 (0 bytes)	5 min

✓ 0 0 0 0 0

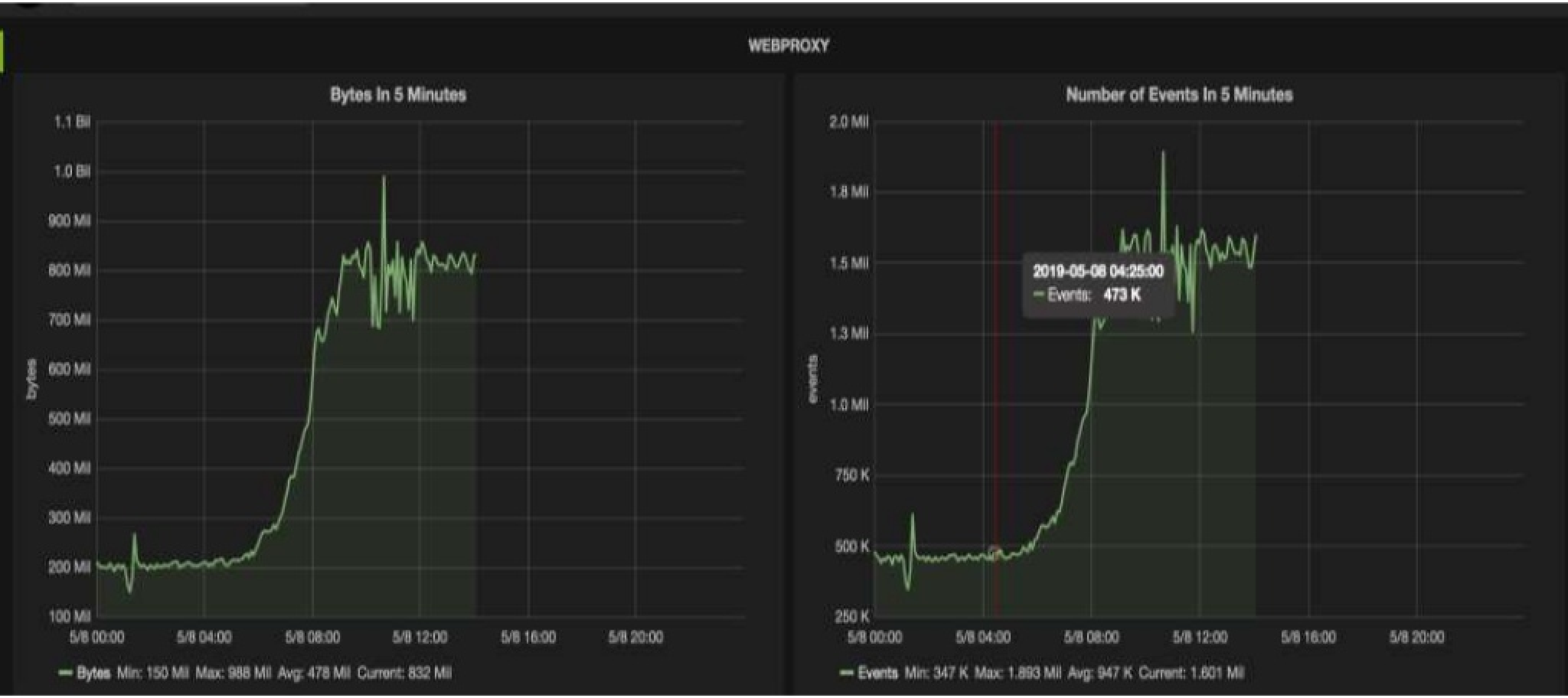
Grafana dashboard – Last 7 days



Yesterday


NIFI - Data Feeds
Zoom Out
Yesterday
Refresh every 5m
WEBPROXY
Bytes In 5 Minutes
1.0 Bil
900 Mil
800 Mil
700 Mil
600 Mil
500 Mil
400 Mil
300 Mil
200 Mil
100 Mil
bytes
Number of Events In 5 Minutes
2.0 Mil
1.8 Mil
1.5 Mil
1.3 Mil
1.0 Mil
750 K
500 K
250 K
events

In the middle of the day



Metrics

- ✓ 100+ production flows
- ✓ ~ 20 Billion events
- ✓ 1000+ Transformations

Next ?

- ✓ MiNiFi
- ✓ Stateless NiFi
- ✓ Registry
- ✓ SAM
- ✓ Real-Time Model training
- ✓ CI/CD, NiFi API’s

thank
you