Hewlett Packard
Enterprise

DATA STAGING PLATFORM
HOW WE AGGREGATE AND ENRICH FORTINET SYSLOG EVENTS

Gabriela Aumayr, Dr. Hugo Hromic

July 14, 2021

AGENDA

- Introduction
- Background
- Data Staging Platform
  - architecture
  - feed onboarding methodology
  - how we Aggregate and Enrich Fortinet Syslog Events
- Demo
- Conclusion and future steps
- Q&A

MOTIVATION

- Scalable processing of cyber security data

- Challenges found in Security Operations Centers (SOC)
  - **Scalability**: high data volume, limited storage for historical data
  - **High Availability**: ingestion, processing and storage must be kept operating
  - **Data Quality**: accuracy of de-duplication, flow stitching, enrichment, field sanitization
  - **Low-latency Output**: short processing time and fast query time
  - **Security:** observability data is sensitive and should be kept protected

- Big Data technologies
  - Scalability - distributed computing
  - High Availability - replication
  - Low-latency output - massive parallel processing
  - Quick integration and deployment - container orchestration

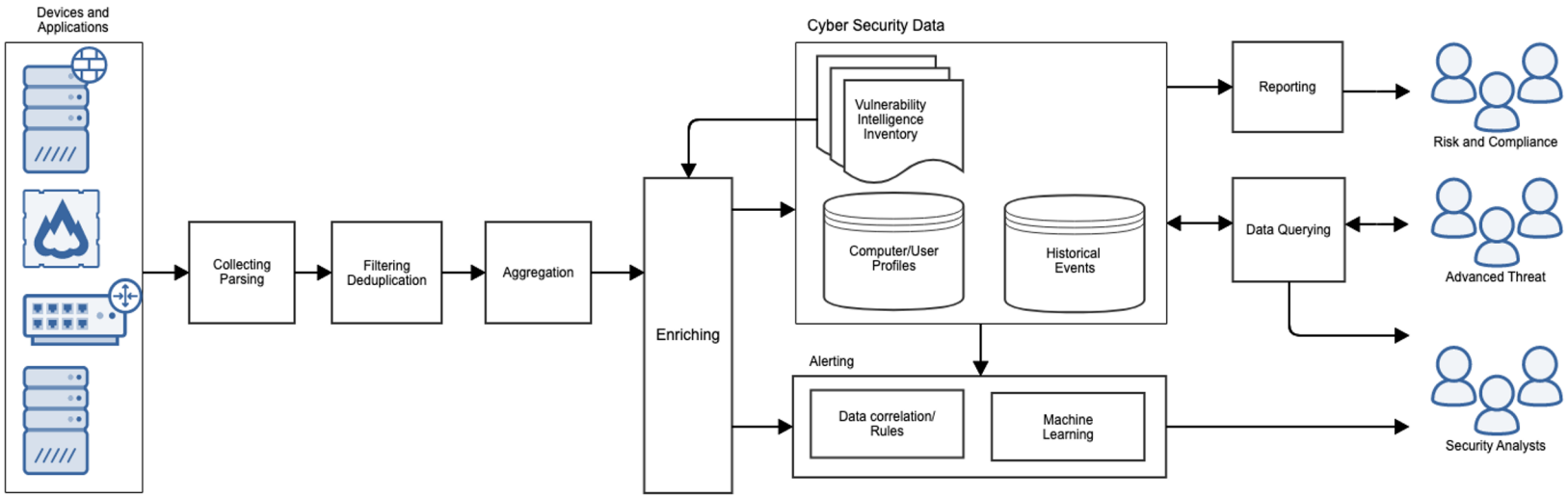
TOOLS STACK

- Big Data
  - **Hadoop**: distributed file system and map-reduce framework
  - **Spark Streaming:** distributed streaming framework
  - **Hazelcast, Hazelcast Jet**: distributed in-memory data grid and streaming framework
  - **Vector**: ultra-fast and reliable observability data pipeline
  - **Kafka**: distributed message broker
  - **Vertica**: distributed columnar store

TOOLS STACK IN DSP

- Big Data
  - **Hadoop**: distributed file system and map-reduce framework
  - **Spark Streaming:** distributed streaming framework
  - **Hazelcast, Hazelcast Jet**: distributed in-memory data grid and streaming framework
  - **Vector**: ultra-fast and reliable observability data pipeline
  - **Kafka**: distributed message broker
  - **Vertica**: distributed columnar store
- Containerization
  - **Docker in Swarm mode:** containerisation platform
  - **Portainer:** container management tool
  - **HPE Trusted Registry:** on-premise registry to store and manage Docker images
- Monitoring
  - **Prometheus:** application metrics
  - **Loki:** distributed log aggregation system
  - **Grafana:** visualisation and alerting

BACKGROUND



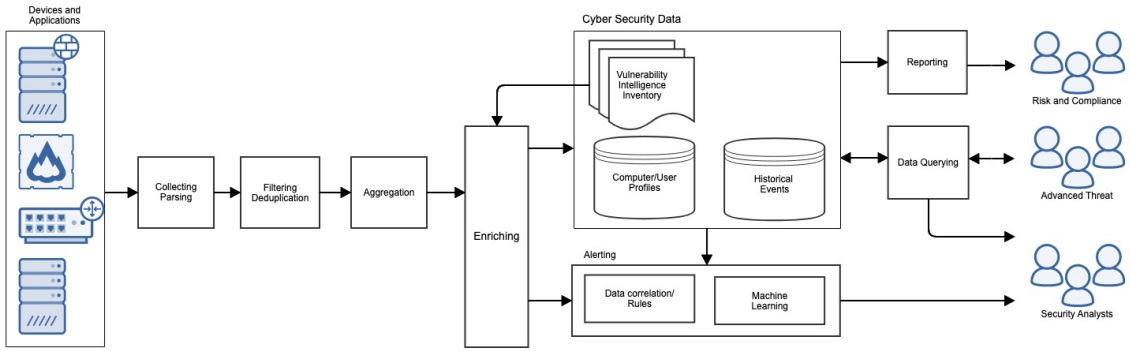
- SAPPAN, EU-funded project
  - methodology for scalable cyber security data processing using traditional big data technologies
  - modular architecture

BACKGROUND

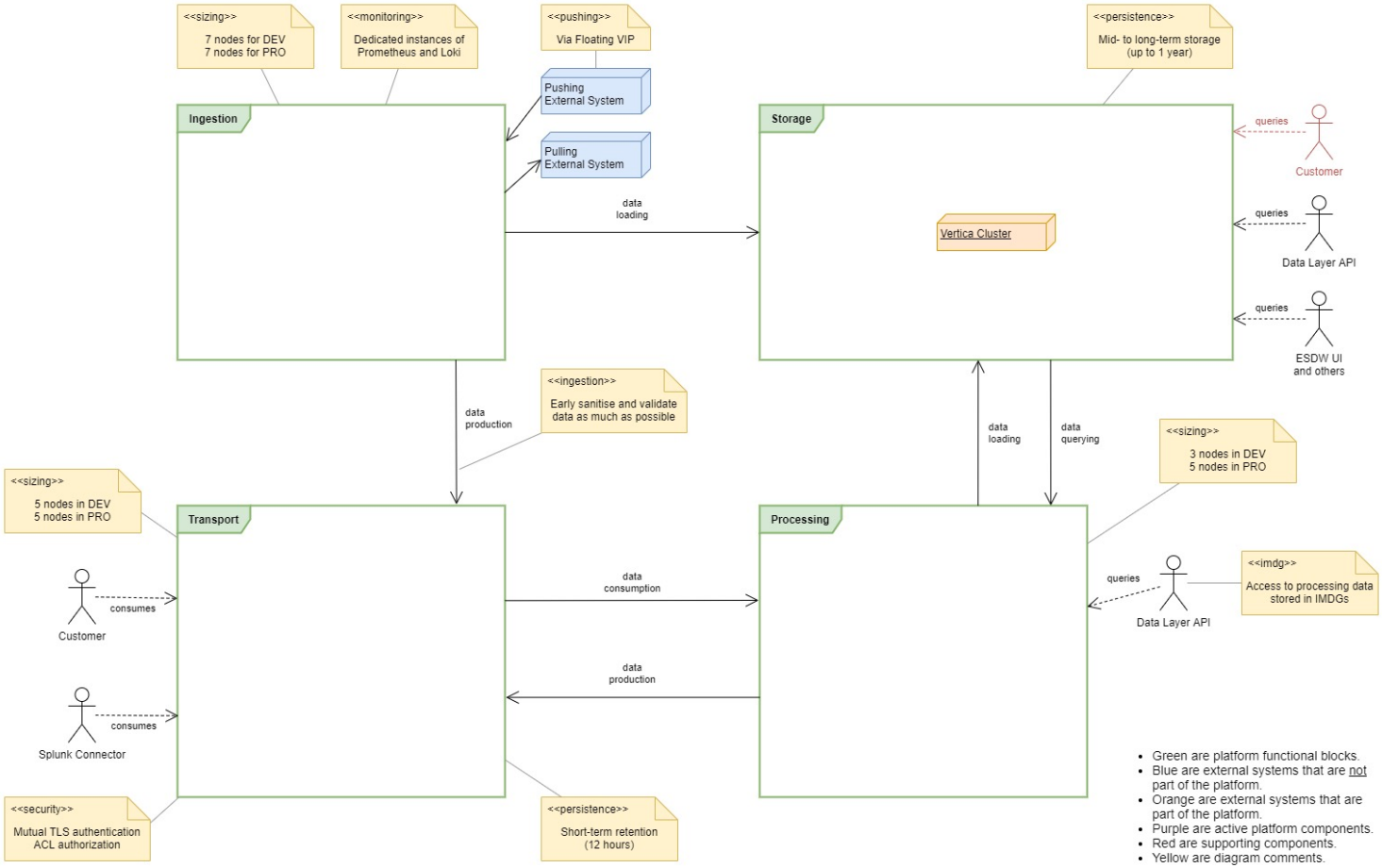
Iterative process

- First iteration – Hortonworks, Spark/Scala
  - PROS
    - Scalable, performant
  - CONS
    - Shared platform, unstable environment
    - At the limit of available resources

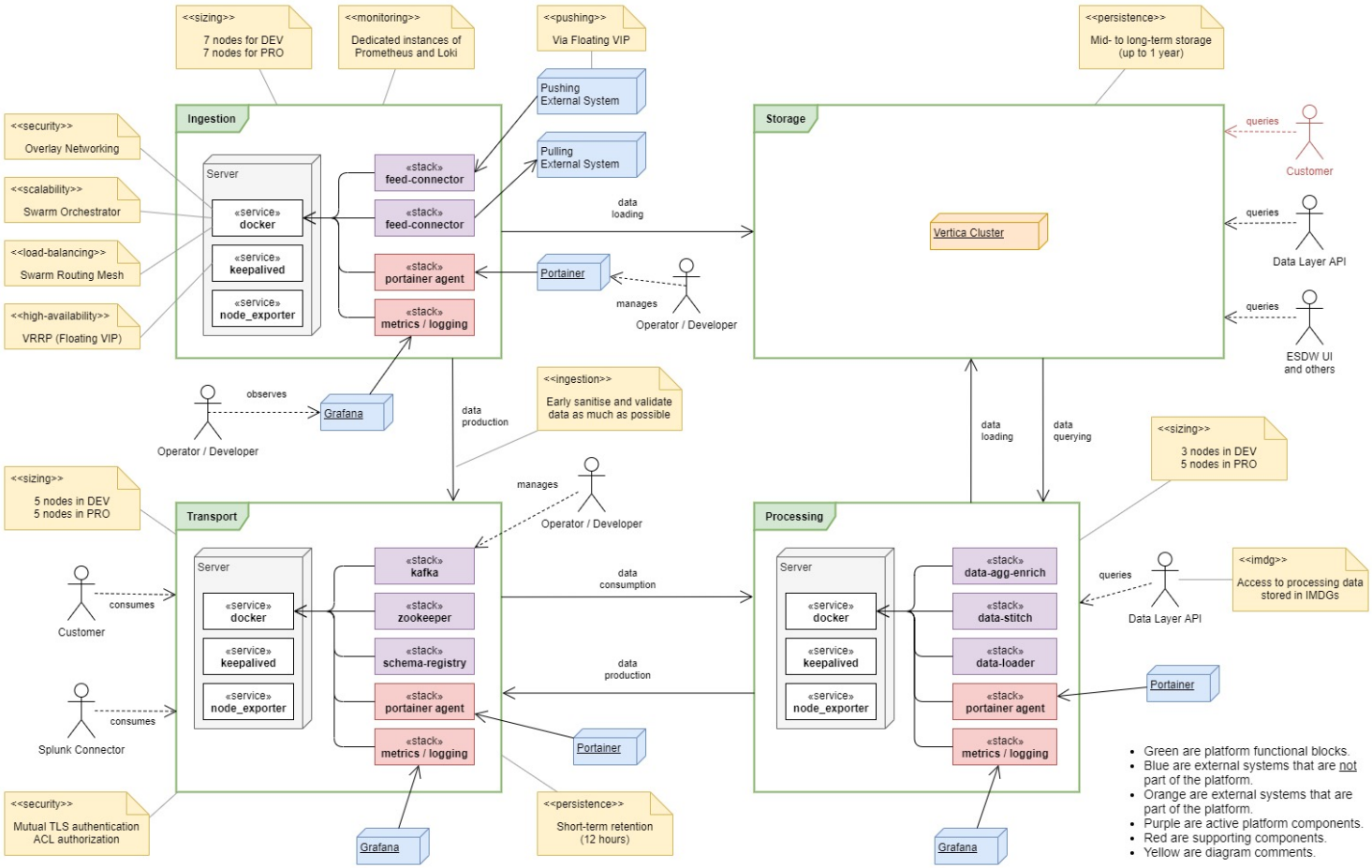
- Second iteration – MapR, Spark/Scala
  - PROS
    - our own platform
  - CONS
    - non-transferable technology
    - different Kafka libraries in MapR
    - difficult to maintain



DATA STAGING PLATFORM ARCHITECTURE



DATA STAGING PLATFORM ARCHITECTURE



FEED ONBOARDING METHODOLOGY IN DSP

- Interaction with stakeholders:
  - data feed prioritization
  - get sample data
  - identify operational characteristics
  - user requirements gathering (data fields, processing steps: aggregation, enrichment, etc.)

- Get infrastructure ready:
  - resource allocation
  - connectivity: NCS, NCR requests etc.

- Data feed documentation
  - established feed template
  - all feeds documentation is standardized

Summary

General

- **Name:** Fortinet Firewall
- **Source:** Fortinet Analyzer
- **HPE Owner(s):** Carnell Tolbert, Patrick MacRoberts
- **Portfolio/ApplicationID:** N/A (Firewalls do not fall under a single EPRID)
- **Admin Contact(s):** John Carnicle (Admin), Barry Or (Network team manager)
- **Support Channel(s):** hpsm-assignment-group: W-INCLV3-ITIO-GT-IPSFW, PDL: gt.network.security@dxc.com, Sample HPSM ticket: IM30204758
- **NCS Request(s):** 20910 , 21122, NCS Excel Files

Operational

- **Modes:** `stream`
- **Strategies:** `incremental`
- **Methods:** `listen`
- **Protocols:** `udp`
- **Authentication:** `n/a`
- **Storage:** `vertica`
- **Vol/Freq:** `70K-80K eps` , `every 60 seconds`
- **Monitoring type:** `internal-dsp`

Known Consumers

- CDC team (cybersecurity-cdc@hpe.com)
- SIEM team (gcs-cdc-eng@hpe.com)
- ATR&I team (cs-atri@hpe.com, Guarav Shah)

Description

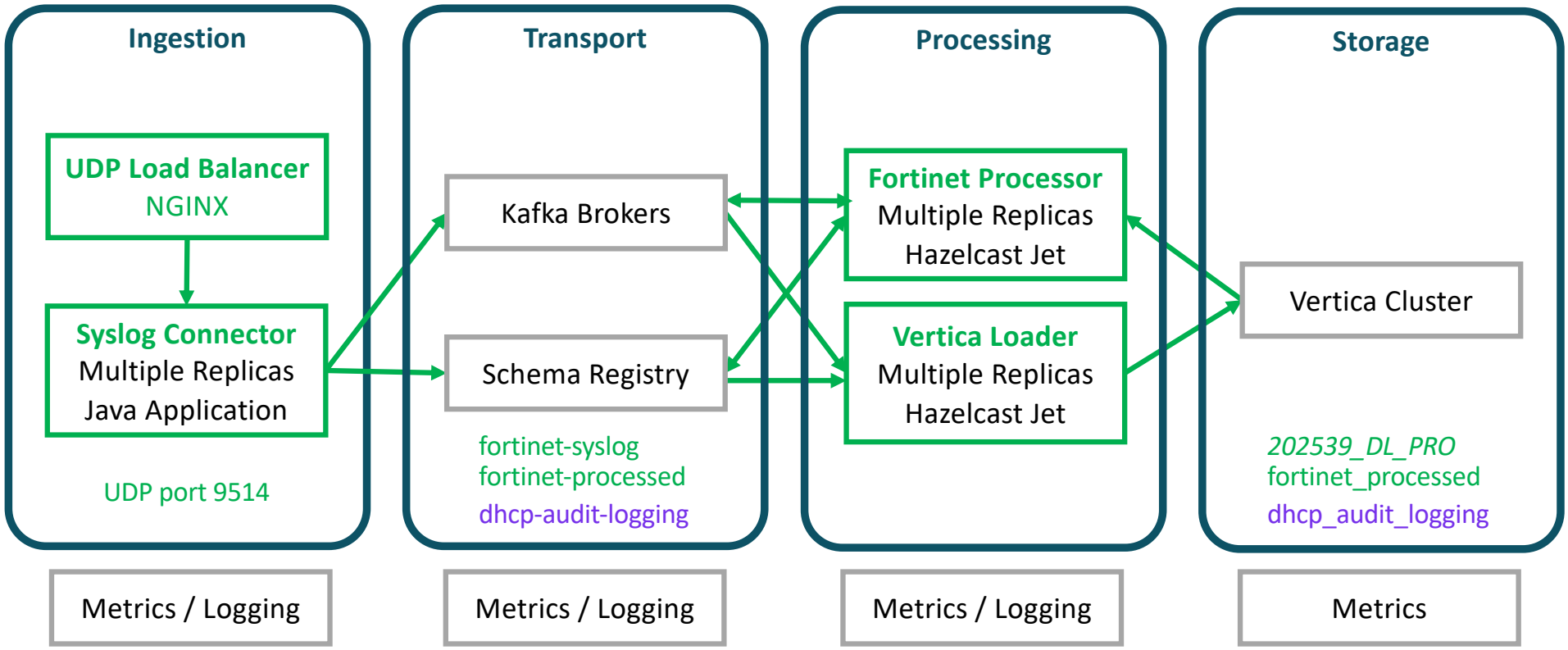
Data: Fortinet Firewall data contains information about which devices are allowed or denied access to which network resources by the company Fortinet firewall devices.
Purpose: The purpose of the data is to provide firewall traffic information for cyber security investigations.
Source: The data is provided by the FAZ (Fortinet management) devices over UDP to port 8514.

Ingestion

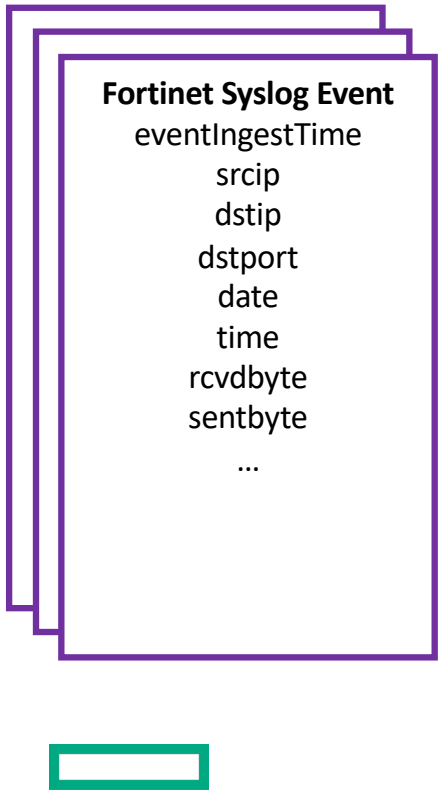
Components

- https://github.hpe.com/OpsEngineering/dsp-syslog-connector

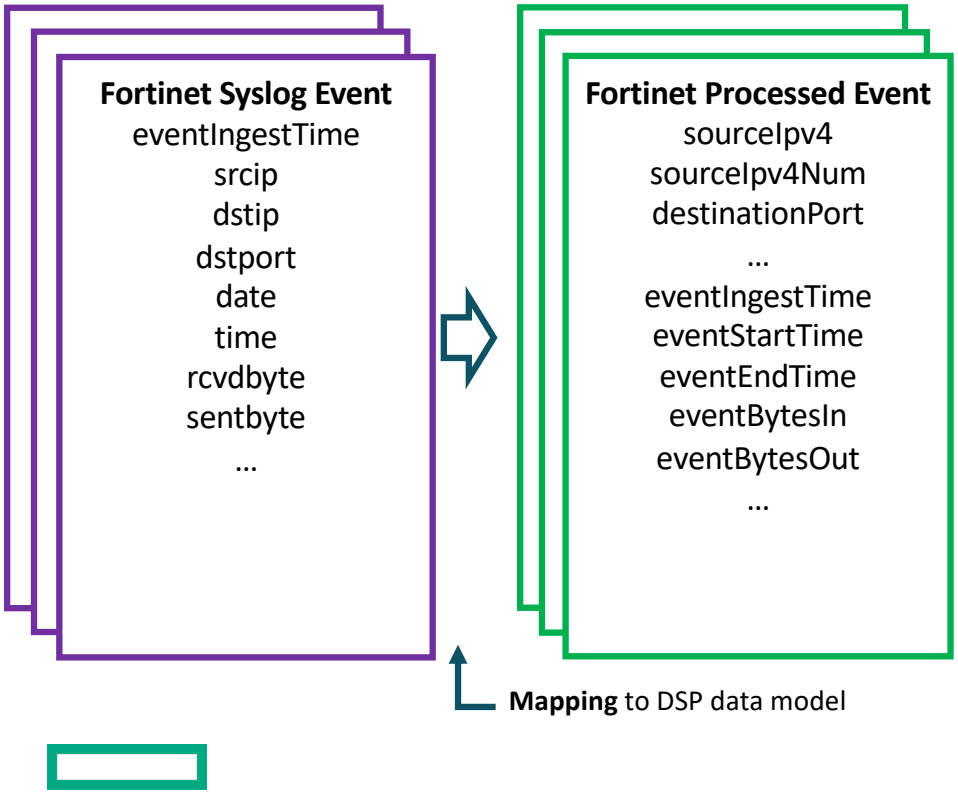
HOW WE INGEST, PROCESS AND STORE FORTINET SYSLOG EVENTS



FORTINET PROCESSING: AGGREGATION AND ENRICHMENT

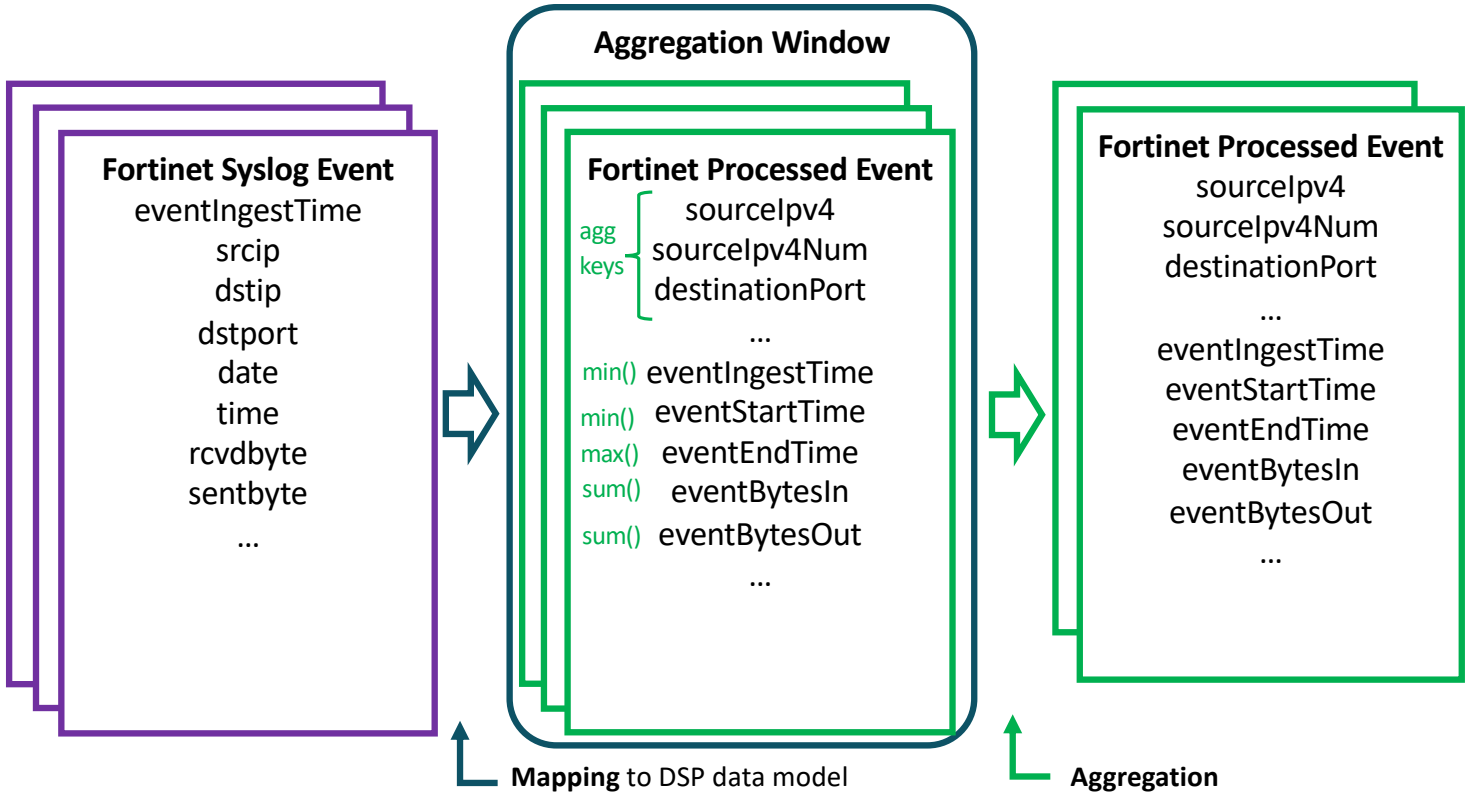


FORTINET PROCESSING: AGGREGATION AND ENRICHMENT



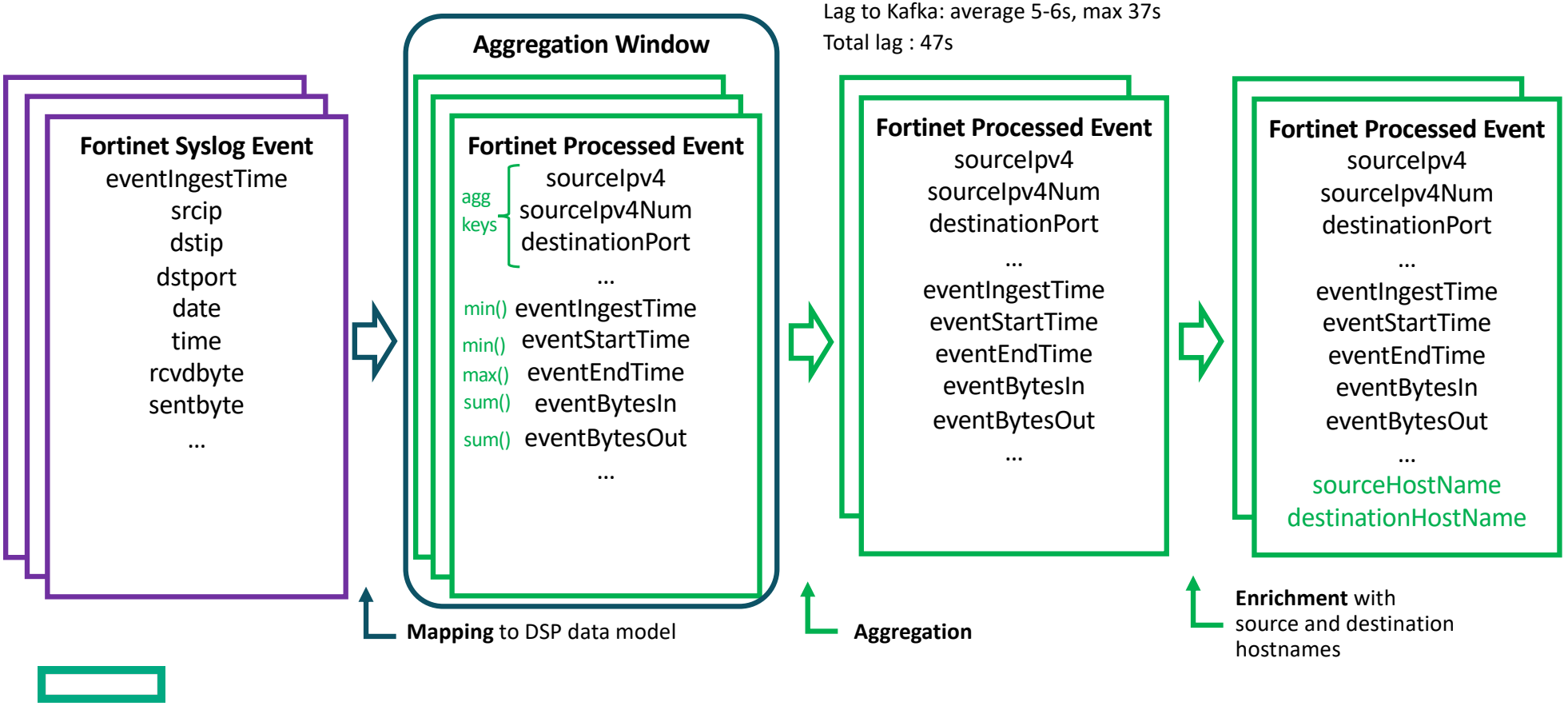
FORTINET PROCESSING: AGGREGATION AND ENRICHMENT

In average, for a 60s window the aggregation ratio is **~67%**



FORTINET PROCESSING: AGGREGATION AND ENRICHMENT

In average, for a 60s window, the aggregation ratio is **~67%**
Lag to Kafka: average 5-6s, max 37s
Total lag : 47s



DEMO

CONCLUSIONS

Solid foundation for the DSP

- Distributed, performant
  - –low latency, good aggregation rate
- Modular architecture
- Cloud oriented
  - –easy to deploy/maintain
- Secure
  - –encryption and access control
- Observable
  - –metrics, logging
  - –specialized dashboards (per feed, per process)
- Unified data model
- Well documented and following best practices
  - –git, code review, release cycles

Future steps

- More feeds onboarding
  - –Palo Alto, Zscaler, Checkpoint
- Handover to OEIS/OE
  - –Finalize operations playbooks
- Actionable Intelligence
  - –**Antonio Neri, Discover 2021:**
    - –“We are entering the age of insight”
    - –“(the future) is not about simply capturing data, but about how fast we can extract value from it”
    - –“Collecting AND connecting data and applying ML at the enterprise scale”

Q&A