Running Apache Kafka at scale on Kubernetes with Strimzi

May 2024

MAERSK

Speakers

Steffen Karlsson
Lead Software Engineer
Streaming Services, Maersk

steffen.karlsson@maersk.com

Maciej Tatarski
Software Engineer
RETinA Team, Maersk

maciej.tatarski@maersk.com

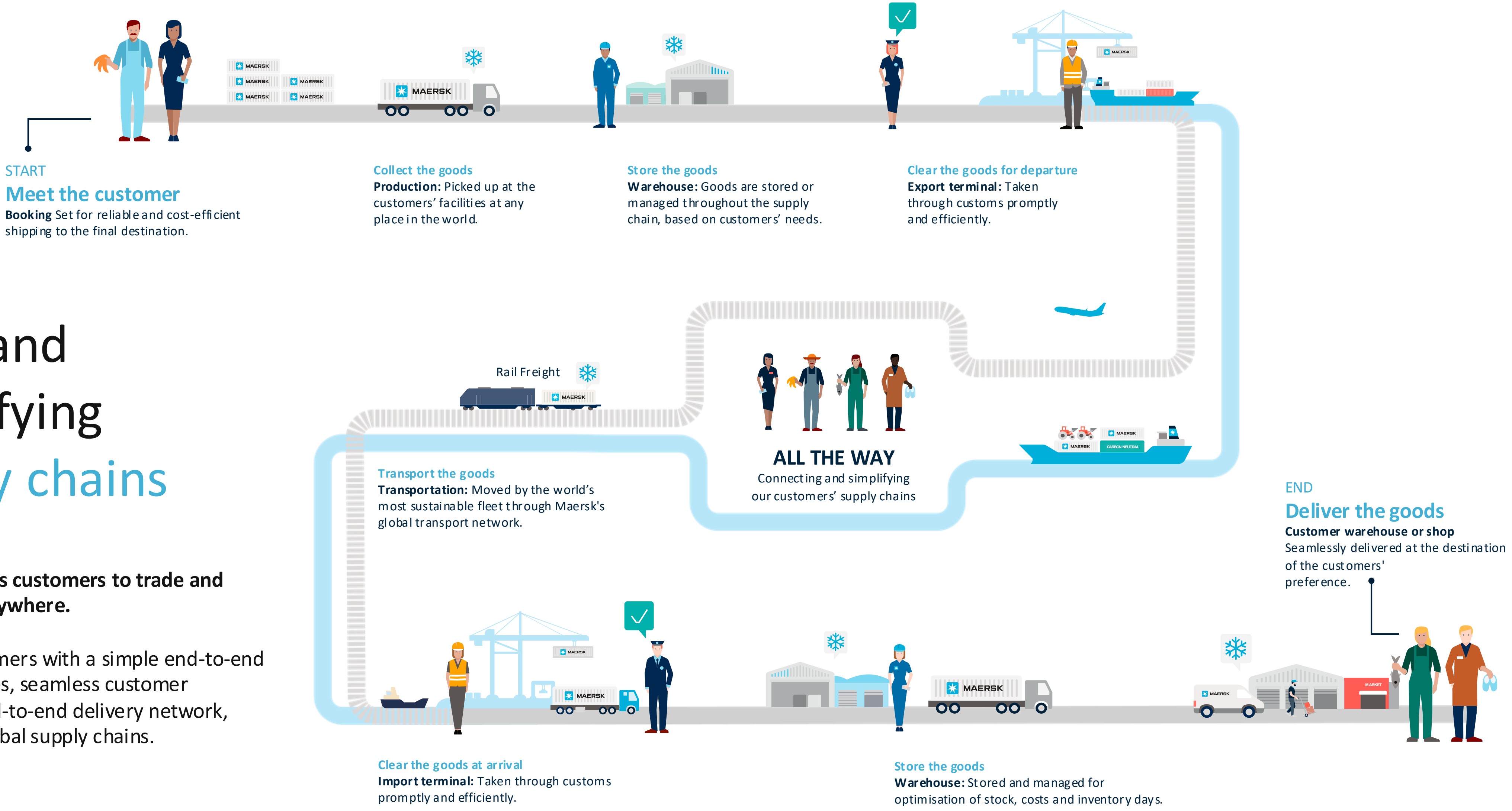
Agenda

- What is Maersk?
- Backstory
- DevX Vision
- Technology backbone of Maersk
- Running Kafka on Kubernetes?
- Architecture
- On-Prem / Edge
- Tools
- Observability & Monitoring
- Alerting
- Stream Processing
- Reflections

Connecting and simplifying global supply chains

A.P. Moller - Maersk enables its customers to trade and grow by transporting goods anywhere.

Maersk works to provide customers with a simple end-to-end offering of products and services, seamless customer engagement and a superior end-to-end delivery network, taking the complexity out of global supply chains.



START

Meet the customer

Booking Set for reliable and cost-efficient shipping to the final destination.

Collect the goods

Production: Picked up at the customers' facilities at any place in the world.

Store the goods

Warehouse: Goods are stored or managed throughout the supply chain, based on customers' needs.

Clear the goods for departure

Export terminal: Taken through customs promptly and efficiently.

Rail Freight

Transport the goods

Transportation: Moved by the world's most sustainable fleet through Maersk's global transport network.

ALL THE WAY

Connecting and simplifying our customers' supply chains

END

Deliver the goods

Customer warehouse or shop

Seamlessly delivered at the destination of the customers' preference.

Clear the goods at arrival

Import terminal: Taken through customs promptly and efficiently.

Store the goods

Warehouse: Stored and managed for optimisation of stock, costs and inventory days.

Backstory

Strategy to move from batch to event driven enterprise architecture

Fast and easy bootstrap with PaaS and SaaS

Gain trust of the business

Cost-benefit analysis: Pay-as-you-go

Transition

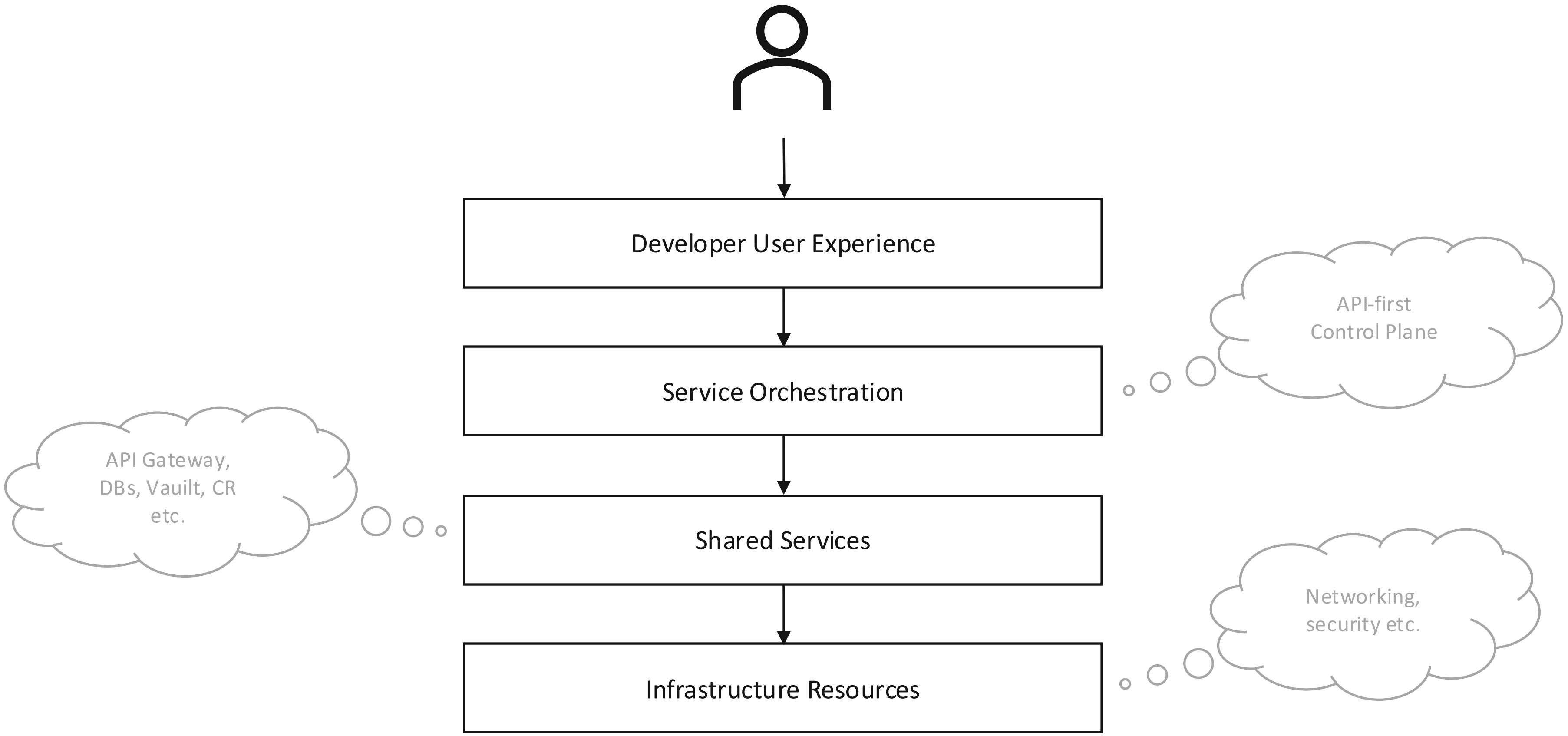
Open-source adoption strategy, benefitting from the economy of scale

Building up internal capabilities and inner-sourcing for a more tailored solution

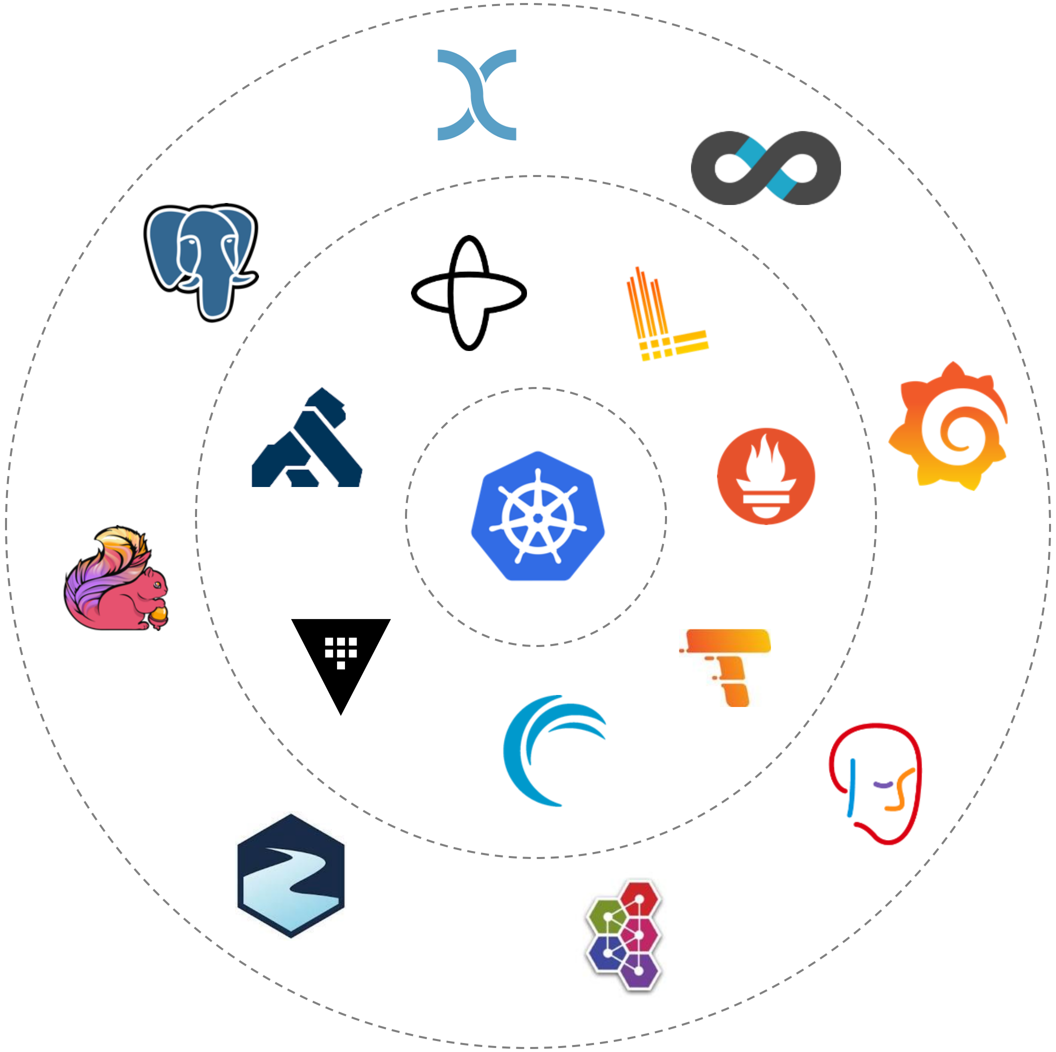
Internal knowledge on how to manage and administer a Kafka ecosystem

Rapidly increasing adoption of the platform, point of no return

DevX Vision

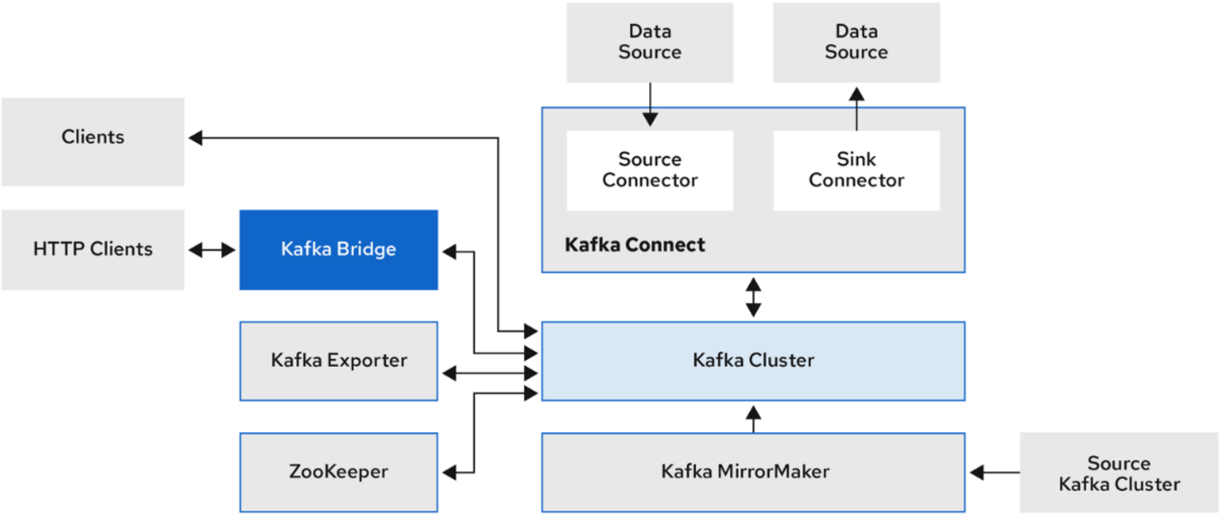

Developer User Experience
Service Orchestration
Shared Services
Infrastructure Resources
API-first
Control Plane
API Gateway,
DBs, Vauilt, CR
etc.
Networking,
security etc.

Technology backbone of Maersk

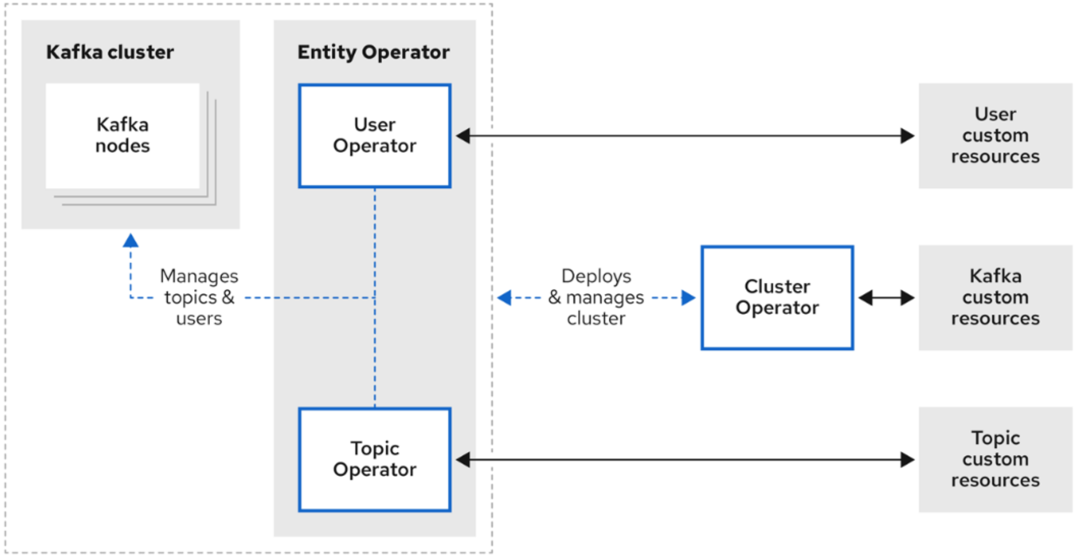


Running Kafka on Kubernetes

- Strimzi provides container images and operators for running Kafka on Kubernetes
- Strimzi operators are purpose-built with specialist operational knowledge to effectively manage Kafka on Kubernetes
- Apache Kafka components are provided for deployment to Kubernetes with the Strimzi distribution



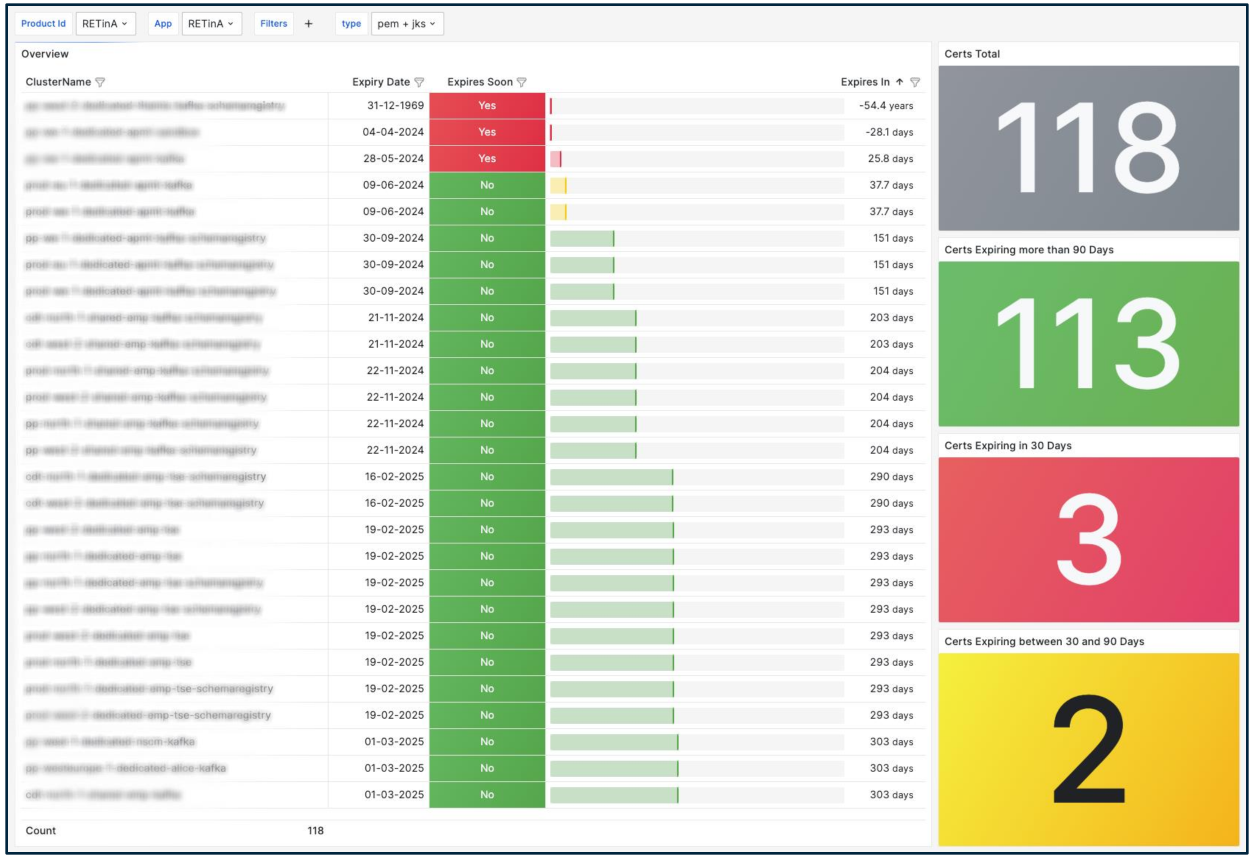
https://strimzi.io/docs/operators/latest/overview#overview-components_str



- Operators are a method of packaging, deploying, and managing Kubernetes applications
- An operator provide a way to extend the Kubernetes API and simplify the administration tasks associated with specific applications, i.e. Kafka Topic

Strimzi Contributions

- #3761 Provide metrics to monitor certificates expiration
- #2779 CrdGenerator validate @JsonPropertyOrder
- #8732 Enhance KafkaBridge resource with consumer inactivity timeout and HTTP consumer/producer parts enablement
- #9537 Kafka Exporter Grafana dashboard too long URL error
- #7374 Improve the Kafka brokers Grafana dashboard



Product Id: RETinA · App: RETinA · Filters + · type: pem + jks

Overview

ClusterName	Expiry Date	Expires Soon	Expires In
[redacted]	31-12-1969	Yes	-54.4 years
[redacted]	04-04-2024	Yes	-28.1 days
[redacted]	28-05-2024	Yes	25.8 days
[redacted]	09-06-2024	No	37.7 days
[redacted]	09-06-2024	No	37.7 days
[redacted]	30-09-2024	No	151 days
[redacted]	30-09-2024	No	151 days
[redacted]	30-09-2024	No	151 days
[redacted]	21-11-2024	No	203 days
[redacted]	21-11-2024	No	203 days
[redacted]	22-11-2024	No	204 days
[redacted]	22-11-2024	No	204 days
[redacted]	22-11-2024	No	204 days
[redacted]	22-11-2024	No	204 days
[redacted]	16-02-2025	No	290 days
[redacted]	16-02-2025	No	290 days
[redacted]	19-02-2025	No	293 days
[redacted]	19-02-2025	No	293 days
[redacted]	19-02-2025	No	293 days
[redacted]	19-02-2025	No	293 days
[redacted]	19-02-2025	No	293 days
[redacted]	19-02-2025	No	293 days
[redacted]	19-02-2025	No	293 days
[redacted]	19-02-2025	No	293 days
[redacted]	01-03-2025	No	303 days
[redacted]	01-03-2025	No	303 days
[redacted]	01-03-2025	No	303 days
Count	118		

Certs Total: 118

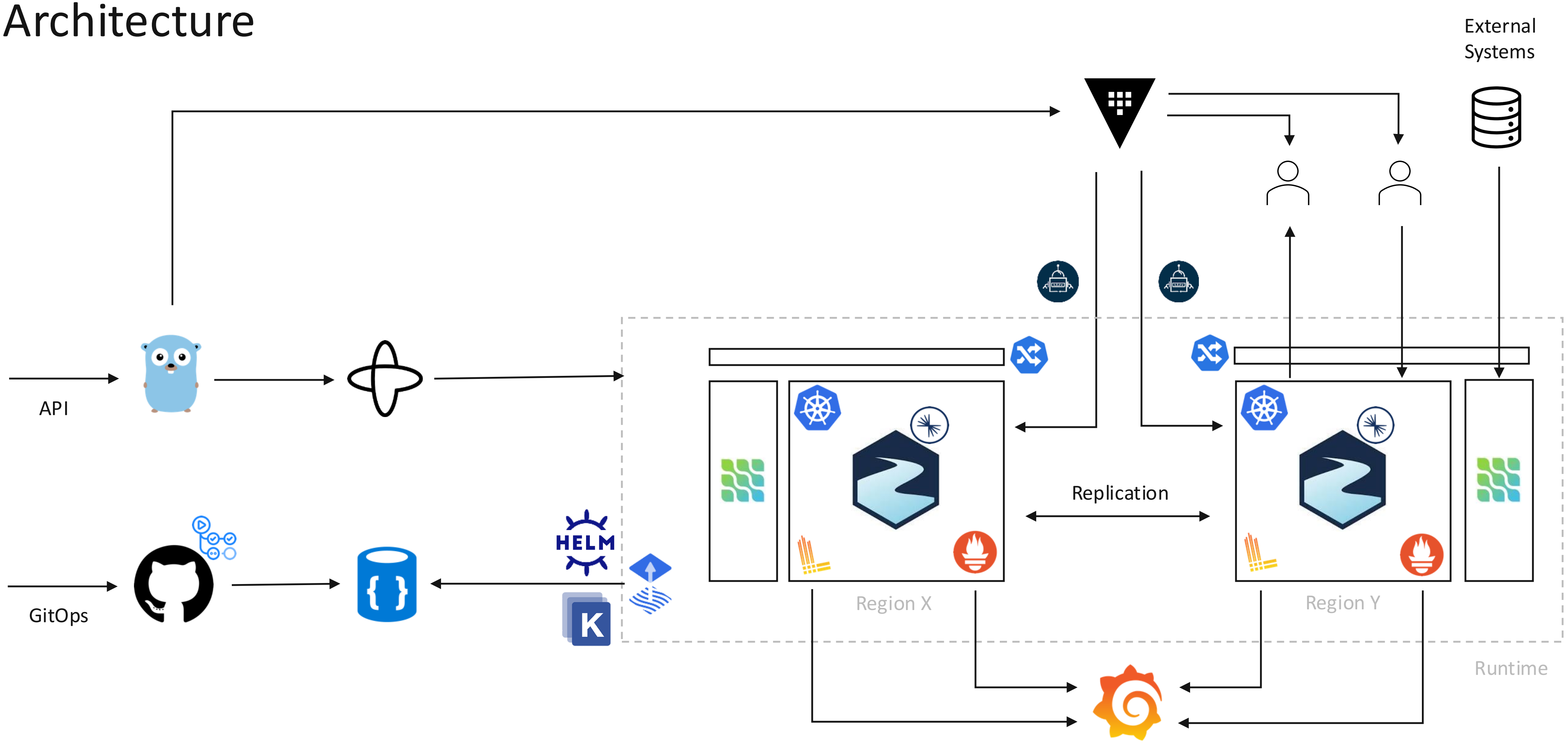
Certs Expiring more than 90 Days: 113

Certs Expiring in 30 Days: 3

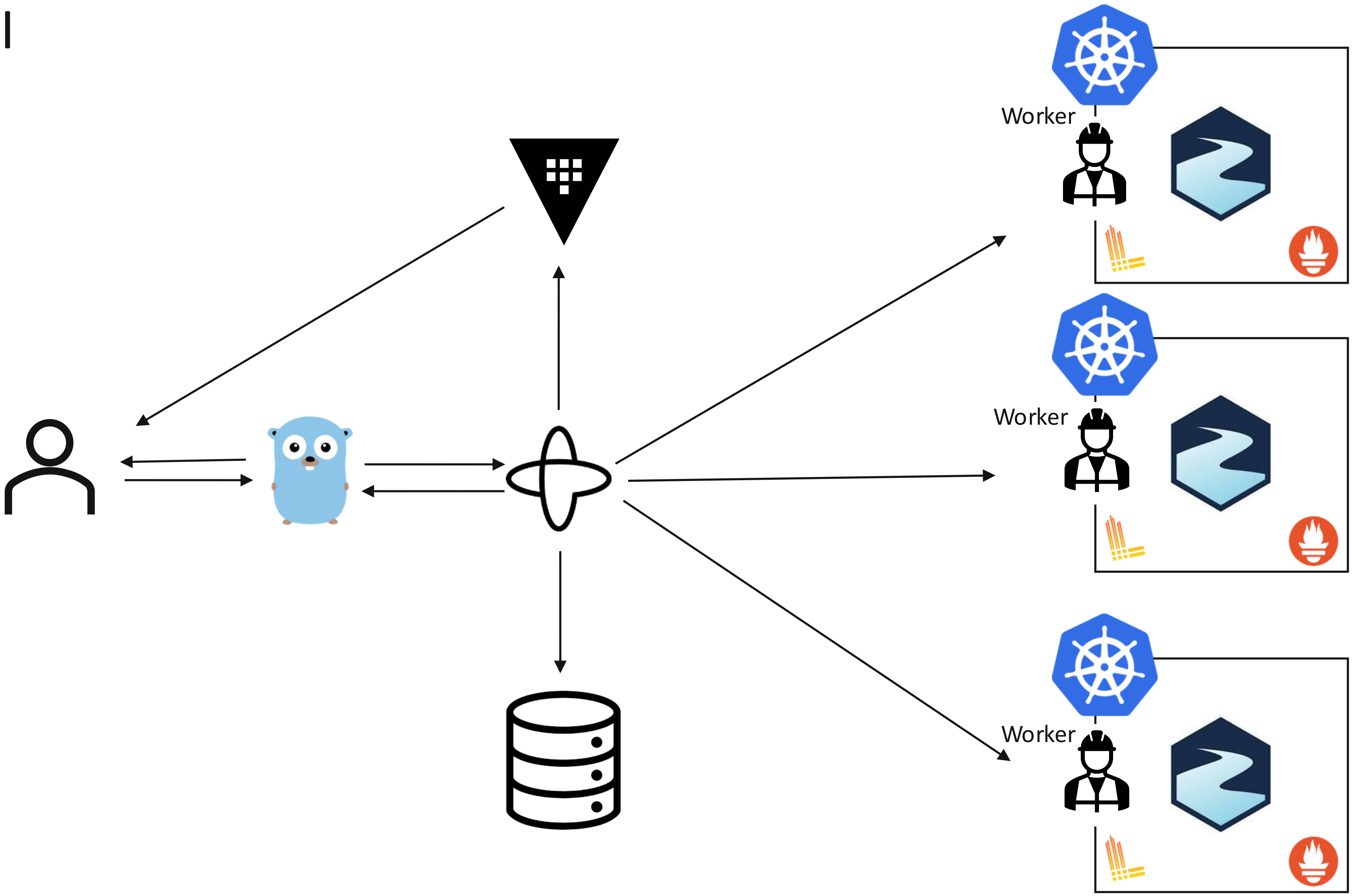
Certs Expiring between 30 and 90 Days: 2

MAERSK

Architecture



New API



OUR PURPOSE

Improving life for all by integrating the world

— The integration illustrated by five years of Automatic Identification System (AIS) transponder data from A.P. Moller - Maersk vessels registered in the company's scheduling system GSIS

● Gateway and hub terminals

Tenants

Number of teams using our Strimzi-based solution	+300
Number of Strimzi-based clusters actively running	81
Number of brokers in all our clusters	+450

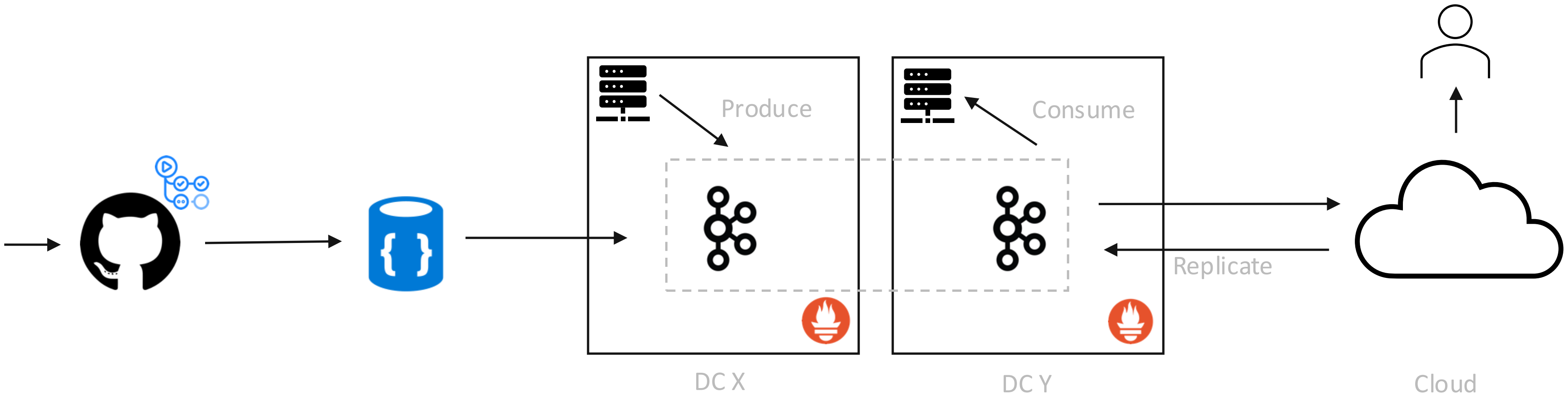
Topics

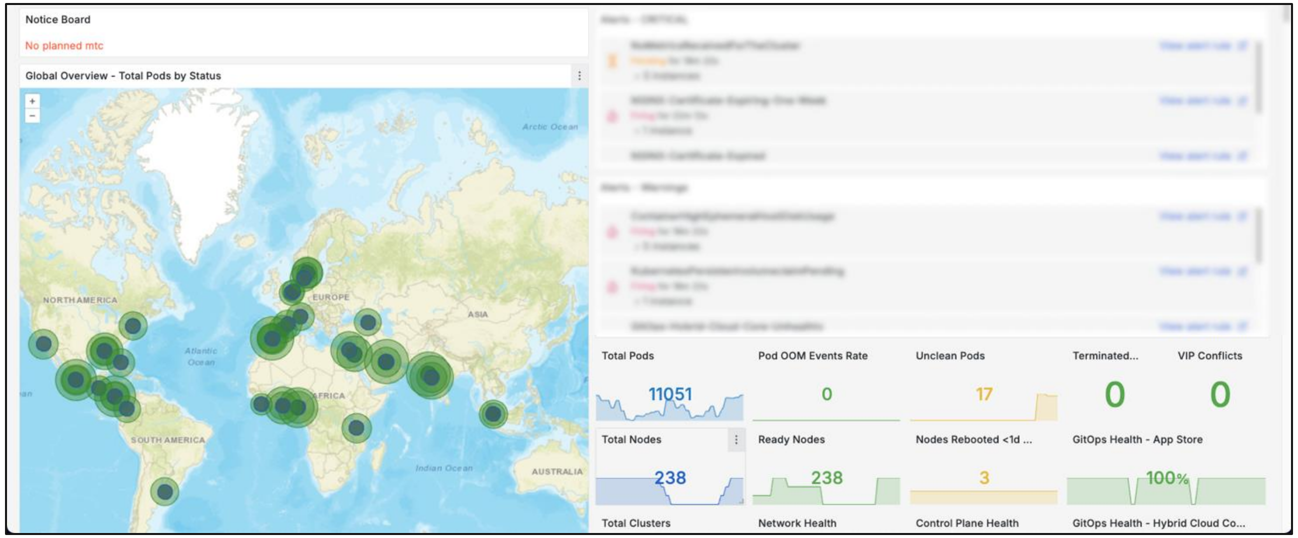
Total number of topics on all our clusters	+36K
Total number of partitions on all our topics	+410K
Number of AVRO and JSON schema versions in Schema Registry	+42K

Events

Total number of unique messages produced per day	+6Bil
Bytes in per second, avg	~140Mb
Bytes out per second, avg	~260Mb

On-Prem and Edge




Notice Board

No planned mtc

Global Overview - Total Pods by Status

Total Pods 11051

Pod OOM Events Rate 0

Unclean Pods 17

Terminated... 0

VIP Conflicts 0

Total Nodes 238

Ready Nodes 238

Nodes Rebooted <1d ... 3

GitOps Health - App Store 100%

Total Clusters

Network Health

Control Plane Health

GitOps Health - Hybrid Cloud Co...

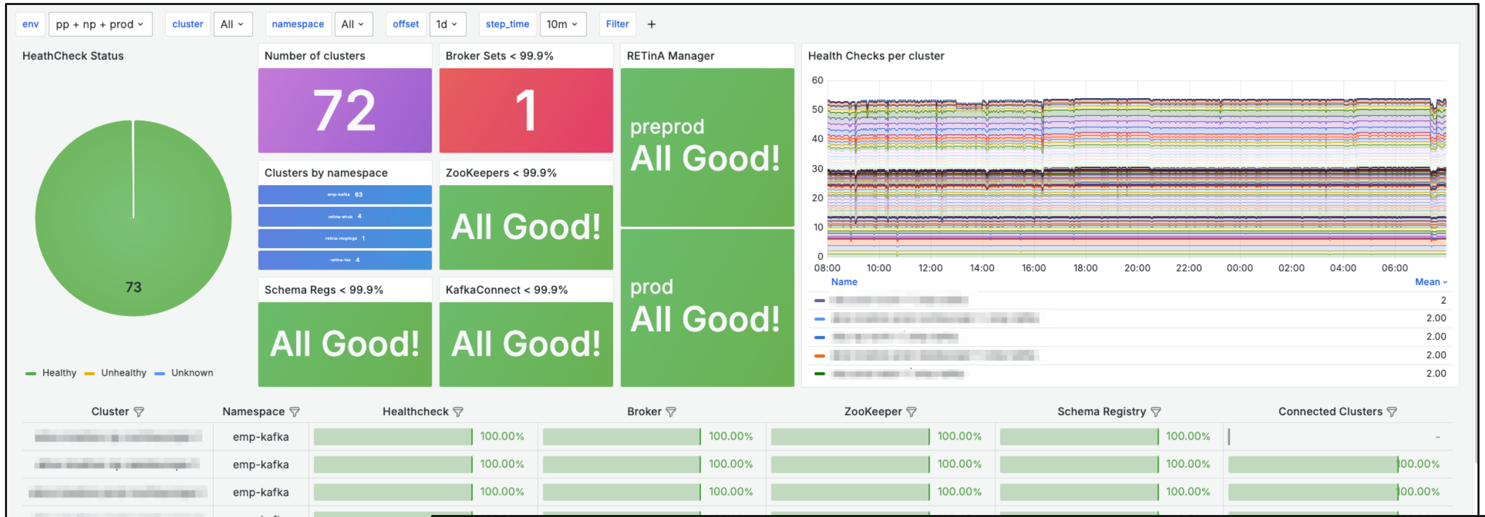
Edge

Number of topics in the terminals	+3.5K
Number of Strimzi-based clusters actively running in the terminals	37
Unique geographical locations	+17

On-prem

Number of topics on-prem	+1.6K
Total number of unique messages produced per day	+100M
Total number of partitions on all our topics	+20K

Observability & Monitoring



env | pp + np + prod | cluster | All | namespace | All | offset | 1d | step_time | 10m | Filter | +

HeathCheck Status

Number of clusters: 72

Broker Sets < 99.9%: 1

RETinA Manager: preprod All Good! / prod All Good!

Clusters by namespace: emp-kafka 63 · retina-ehub 4 · retina-moplogs 1 · retina-tse 4

ZooKeepers < 99.9%: All Good!

Schema Regs < 99.9%: All Good!

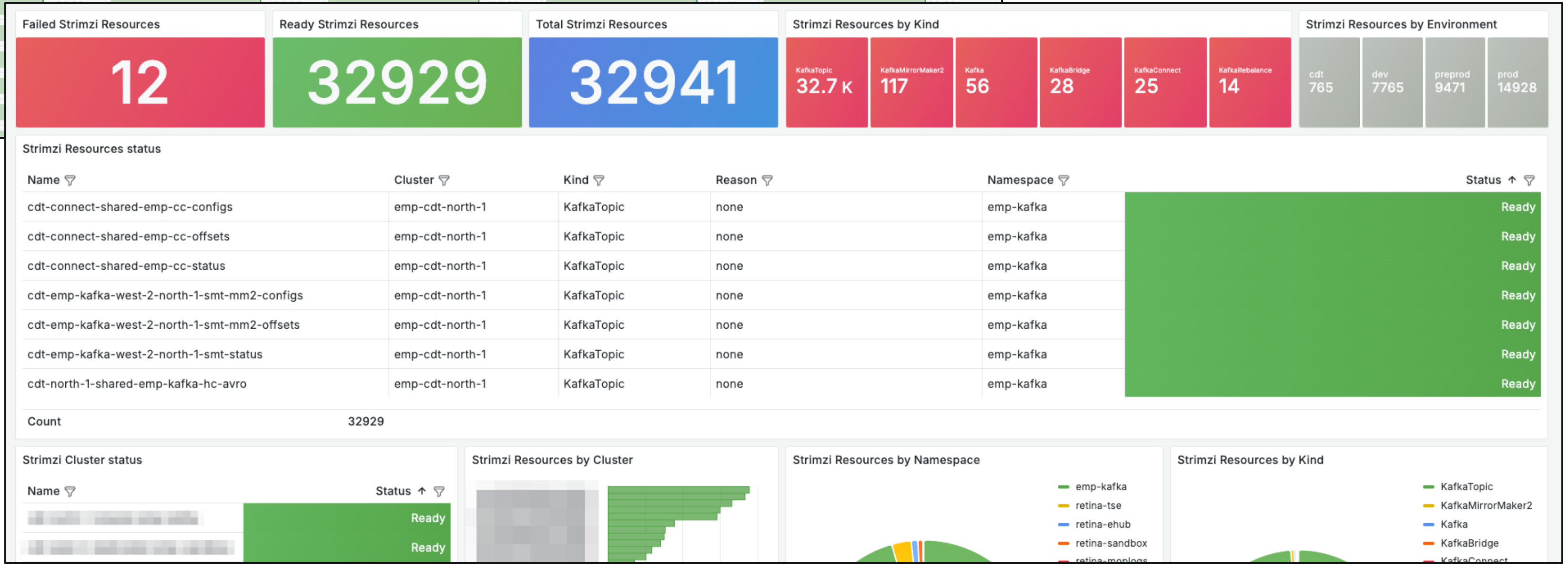
KafkaConnect < 99.9%: All Good!

Health Checks per cluster

Name	Mean
[redacted]	2
[redacted]	2.00
[redacted]	2.00
[redacted]	2.00
[redacted]	2.00

Cluster	Namespace	Healthcheck	Broker	ZooKeeper	Schema Registry	Connected Clusters
[redacted]	emp-kafka	100.00%	100.00%	100.00%	100.00%	-
[redacted]	emp-kafka	100.00%	100.00%	100.00%	100.00%	100.00%
[redacted]	emp-kafka	100.00%	100.00%	100.00%	100.00%	100.00%
[redacted]	emp-kafka					
[redacted]	emp-kafka					
[redacted]	emp-kafka					
[redacted]	emp-kafka					
[redacted]	emp-kafka					

- Fully internal and open-source based observability platform
- Based on the LGTM and OpenTelemetry
- Healthchecks
- Real User Monitoring
- Synthetic Monitoring



Failed Strimzi Resources: 12

Ready Strimzi Resources: 32929

Total Strimzi Resources: 32941

Strimzi Resources by Kind: KafkaTopic 32.7 K · KafkaMirrorMaker2 117 · Kafka 56 · KafkaBridge 28 · KafkaConnect 25 · KafkaRebalance 14

Strimzi Resources by Environment: cdt 765 · dev 7765 · preprod 9471 · prod 14928

Strimzi Resources status

Name	Cluster	Kind	Reason	Namespace	Status ↑
cdt-connect-shared-emp-cc-configs	emp-cdt-north-1	KafkaTopic	none	emp-kafka	Ready
cdt-connect-shared-emp-cc-offsets	emp-cdt-north-1	KafkaTopic	none	emp-kafka	Ready
cdt-connect-shared-emp-cc-status	emp-cdt-north-1	KafkaTopic	none	emp-kafka	Ready
cdt-emp-kafka-west-2-north-1-smt-mm2-configs	emp-cdt-north-1	KafkaTopic	none	emp-kafka	Ready
cdt-emp-kafka-west-2-north-1-smt-mm2-offsets	emp-cdt-north-1	KafkaTopic	none	emp-kafka	Ready
cdt-emp-kafka-west-2-north-1-smt-status	emp-cdt-north-1	KafkaTopic	none	emp-kafka	Ready
cdt-north-1-shared-emp-kafka-hc-avro	emp-cdt-north-1	KafkaTopic	none	emp-kafka	Ready

Count 32929

Strimzi Cluster status

Name	Status ↑
[redacted]	Ready
[redacted]	Ready

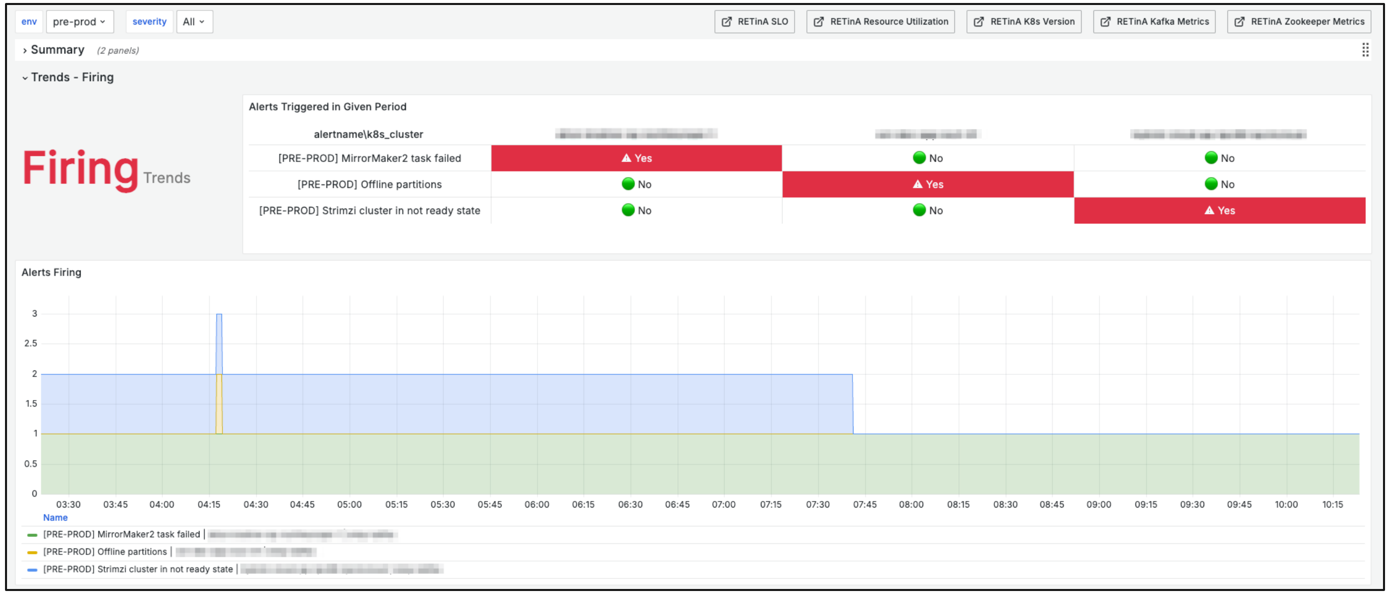
Strimzi Resources by Cluster

Strimzi Resources by Namespace

Strimzi Resources by Kind

Alerting

- Aggregated alerting by cluster and namespace
- Alerts and Dashboards as code
- Automated deployment and synchronization using GitHub Actions
- SLA/SLO 99.99%

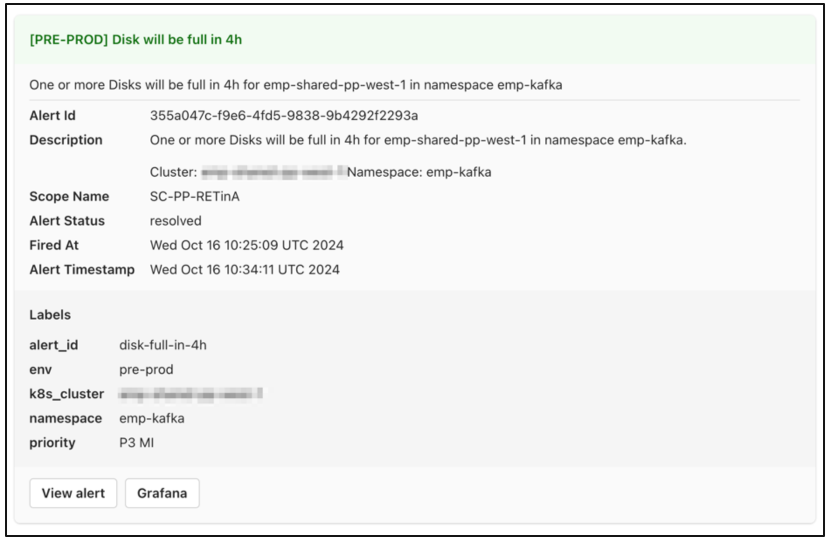


Alerts Triggered in Given Period

alertname\k8s_cluster	[redacted]	[redacted]	[redacted]
[PRE-PROD] MirrorMaker2 task failed	⚠ Yes	● No	● No
[PRE-PROD] Offline partitions	● No	⚠ Yes	● No
[PRE-PROD] Strimzi cluster in not ready state	● No	● No	⚠ Yes

Alerts Firing

- [PRE-PROD] MirrorMaker2 task failed | [redacted]
- [PRE-PROD] Offline partitions | [redacted]
- [PRE-PROD] Strimzi cluster in not ready state | [redacted]



[PRE-PROD] Disk will be full in 4h

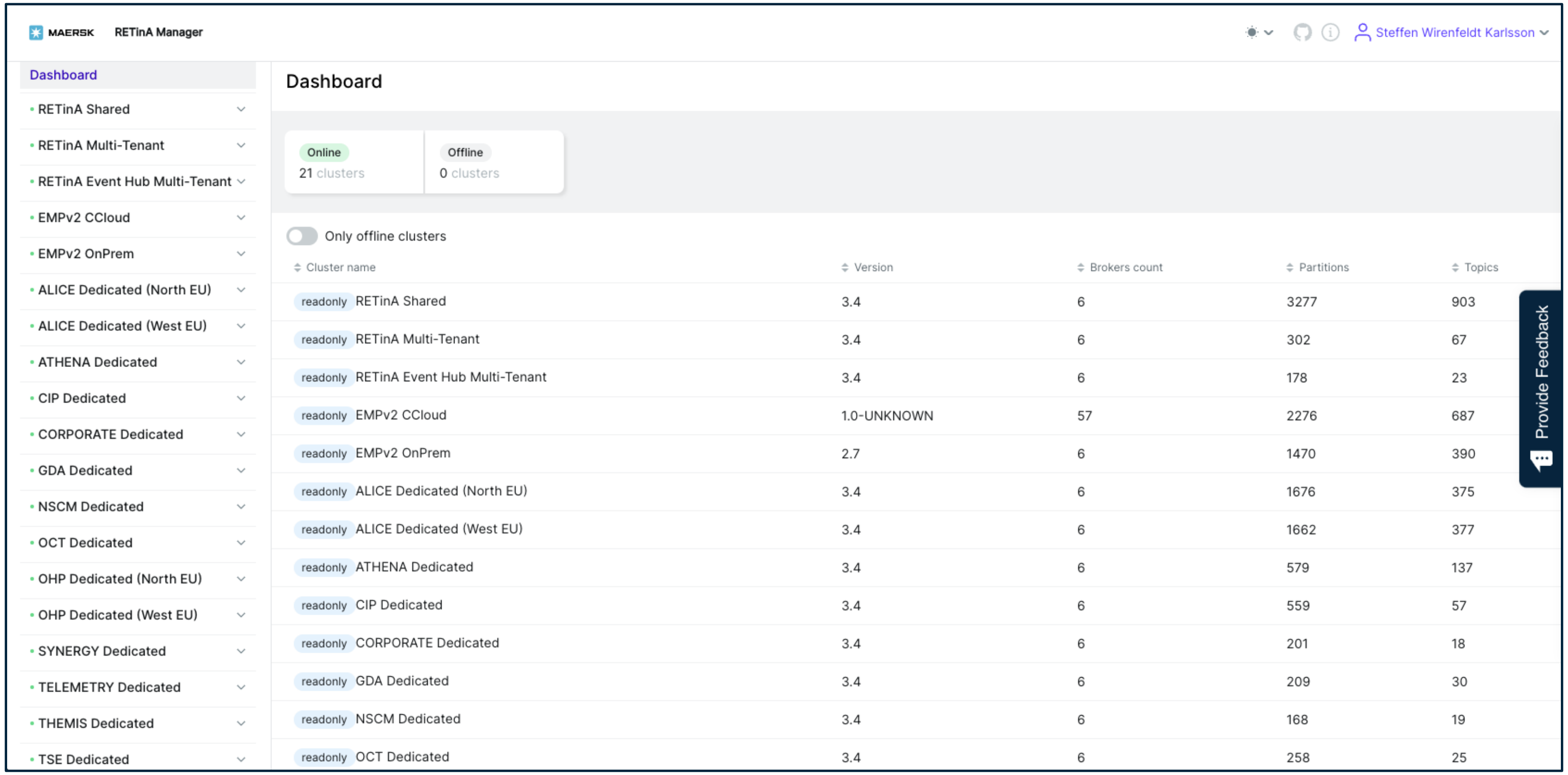
One or more Disks will be full in 4h for emp-shared-pp-west-1 in namespace emp-kafka

Alert Id	355a047c-f9e6-4fd5-9838-9b4292f2293a
Description	One or more Disks will be full in 4h for emp-shared-pp-west-1 in namespace emp-kafka. Cluster: [redacted] Namespace: emp-kafka
Scope Name	SC-PP-RETinA
Alert Status	resolved
Fired At	Wed Oct 16 10:25:09 UTC 2024
Alert Timestamp	Wed Oct 16 10:34:11 UTC 2024

Labels

alert_id	disk-full-in-4h
env	pre-prod
k8s_cluster	[redacted]
namespace	emp-kafka
priority	P3 MI

RETinA Manager



- Dashboard
- RETinA Shared
- RETinA Multi-Tenant
- RETinA Event Hub Multi-Tenant
- EMPv2 CCloud
- EMPv2 OnPrem
- ALICE Dedicated (North EU)
- ALICE Dedicated (West EU)
- ATHENA Dedicated
- CIP Dedicated
- CORPORATE Dedicated
- GDA Dedicated
- NSCM Dedicated
- OCT Dedicated
- OHP Dedicated (North EU)
- OHP Dedicated (West EU)
- SYNERGY Dedicated
- TELEMETRY Dedicated
- THEMIS Dedicated
- TSE Dedicated

Dashboard

Online
21 clusters

Offline
0 clusters

Only offline clusters

Cluster name	Version	Brokers count	Partitions	Topics
readonly RETinA Shared	3.4	6	3277	903
readonly RETinA Multi-Tenant	3.4	6	302	67
readonly RETinA Event Hub Multi-Tenant	3.4	6	178	23
readonly EMPv2 CCloud	1.0-UNKNOWN	57	2276	687
readonly EMPv2 OnPrem	2.7	6	1470	390
readonly ALICE Dedicated (North EU)	3.4	6	1676	375
readonly ALICE Dedicated (West EU)	3.4	6	1662	377
readonly ATHENA Dedicated	3.4	6	579	137
readonly CIP Dedicated	3.4	6	559	57
readonly CORPORATE Dedicated	3.4	6	201	18
readonly GDA Dedicated	3.4	6	209	30
readonly NSCM Dedicated	3.4	6	168	19
readonly OCT Dedicated	3.4	6	258	25

https://github.com/provectus/kafka-ui

Unique users per day

+300

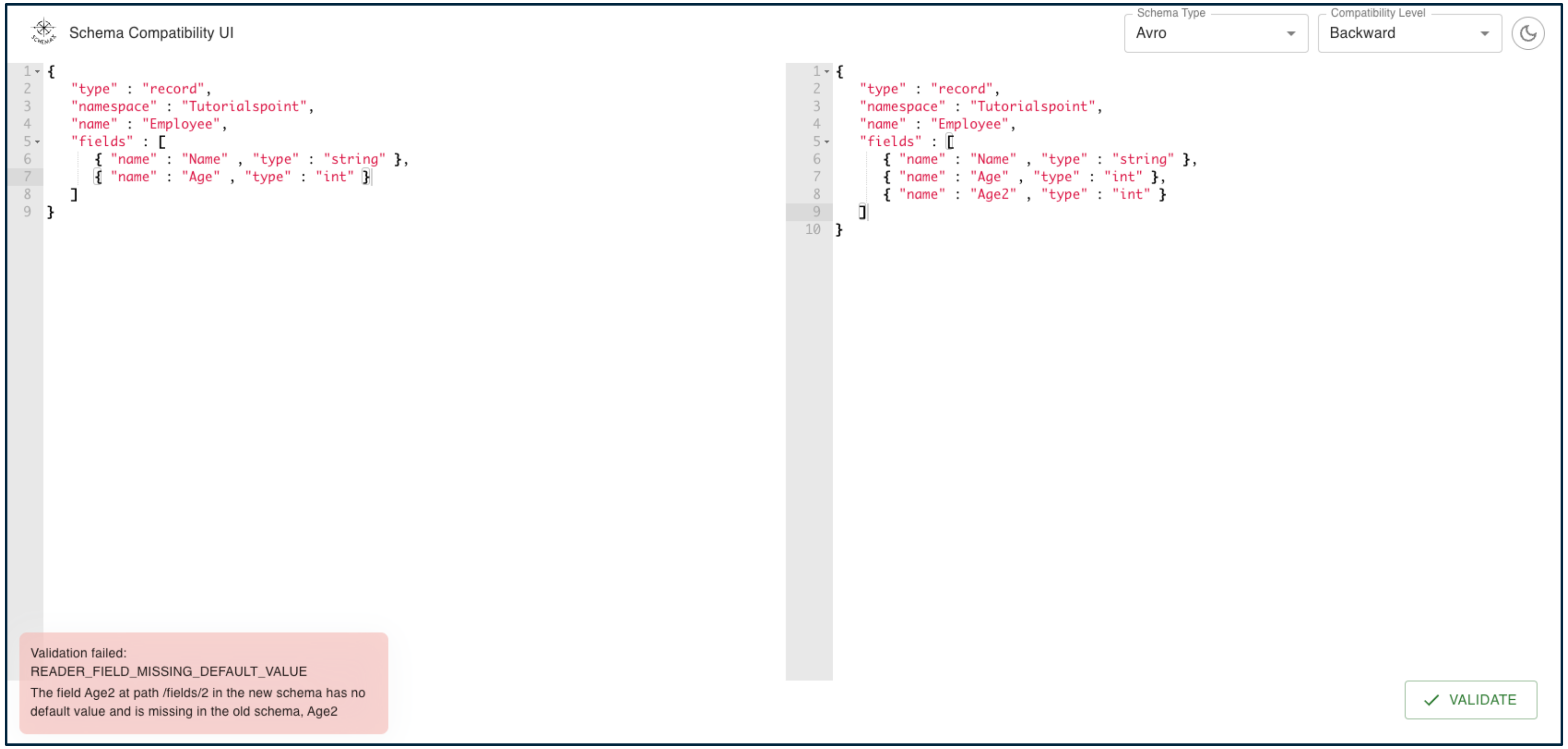
Requests per day

+7.5K

Schema Compatibility UI

- Standalone and open-source Apache-2.0 license
- Fail fast and enable tenants to be more independent to decrease time-to-market
- Full self-service and transparency in schema compatibility and comparison

- **Schema Types:**
  Avro, JSON and Protobuf
- **Compatibility levels:**
  Backwards, Forwards, Full and None



Schema Type: Avro

Compatibility Level: Backward

```
{
  "type" : "record",
  "namespace" : "Tutorialspoint",
  "name" : "Employee",
  "fields" : [
    { "name" : "Name" , "type" : "string" },
    { "name" : "Age" , "type" : "int" }
  ]
}
```

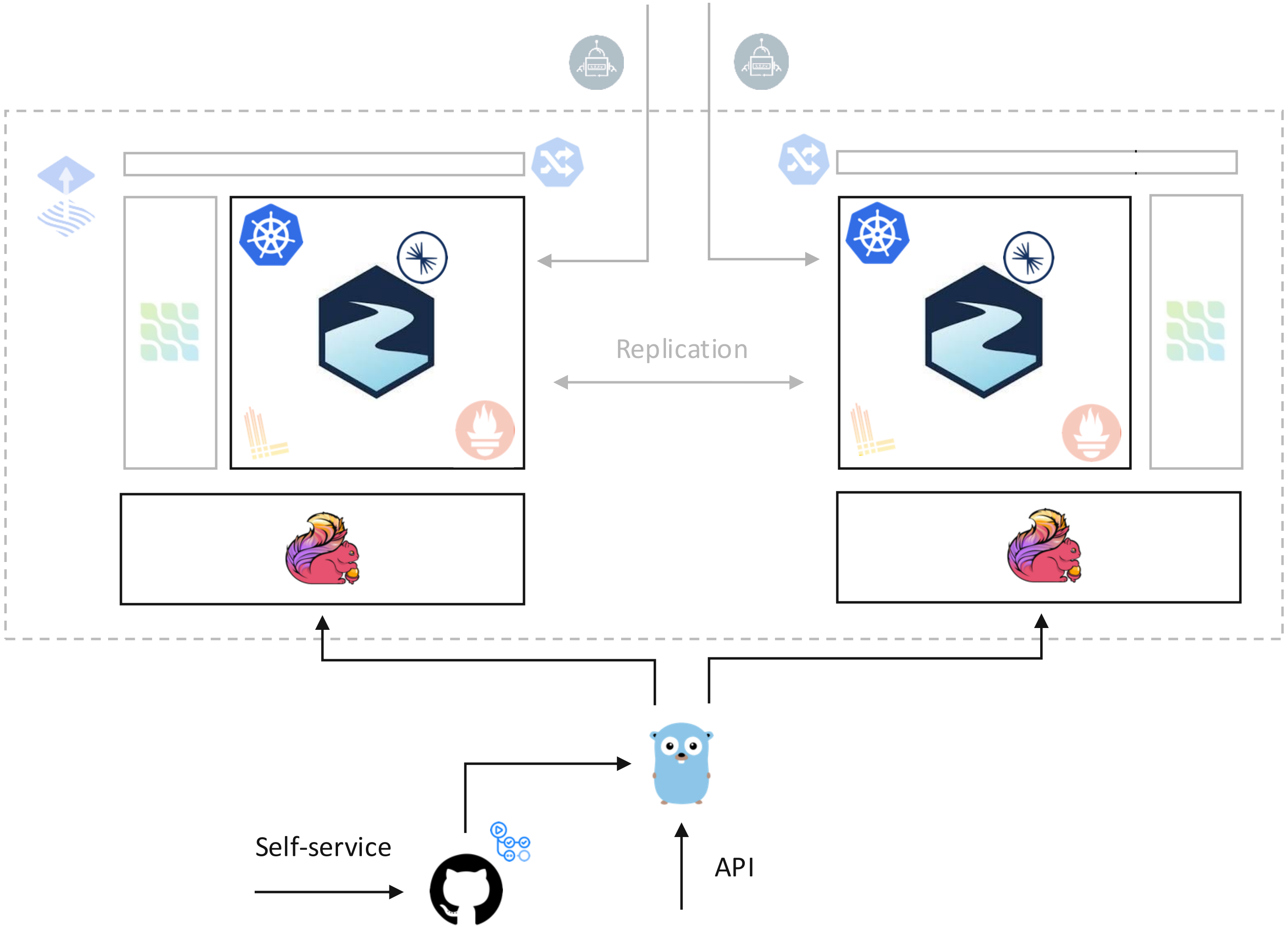
```
{
  "type" : "record",
  "namespace" : "Tutorialspoint",
  "name" : "Employee",
  "fields" : [
    { "name" : "Name" , "type" : "string" },
    { "name" : "Age" , "type" : "int" },
    { "name" : "Age2" , "type" : "int" }
  ]
}
```

Validation failed:
READER_FIELD_MISSING_DEFAULT_VALUE
The field Age2 at path /fields/2 in the new schema has no default value and is missing in the old schema, Age2

MAERSK

Stream Processing

- Automated self-serviceable stream processing on-top of Strimzi with Flink and GitHub Actions
- API-first architecture for better system-to-system integrations
- Predefined template jobs for better overall performance on the cluster
- Deployed using open-source community operator for better stability and configuration of Flink
- Fully monitored and observed using the Grafana and Prometheus stack



Reflections

- Migrations are hard in kafka world
- More clusters = more problems
- Supporting multiple onboarding scenarios ain't easy
- Good observability ease support pain
- In our case migrating from confluent cloud cause significant savings even with more adoption of current platform
- Operational stability is a key
- Supporting multiple deployment environments (edge, onprem, cloud) requires wide knowladge spread across team

MAERSK