



Cloud Operations at Scale

Managing 250 Kubernetes Clusters at Zalando

Heinrich Hartmann & Mikkel Larsen



About us

Heinrich Hartmann

- Senior Principal SRE at Zalando
- 10 years of Reliability Engineering
- Led SRE Department for 2.5 years
- Chief Data Scientist @ Circonus
- PhD in Mathematics

Mikkel Larsen

- Principal Engineer at Zalando
- 9 years of Cloud Infra (AWS/Kubernetes)
- Open Source Maintainer

Women > Shoes > High heels

red High Heels

Shoes

Sneakers

Sandals

Pumps

High heels

Pumps

Peep toes

Sandals

Flat shoes

Mules

Ankle boots

Ballerinas

Boots

Sports shoes

Beach shoes

Bridal shoes

House shoes

Outdoor shoes

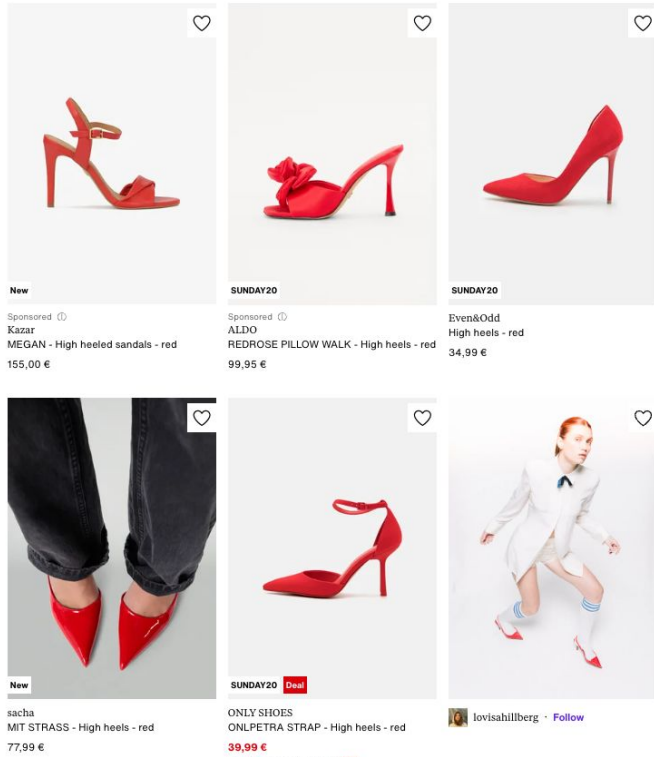
Shoe accessories

Sale

Sort by ▾
 Colour **1** ▾
 Size ▾
 Heel height ▾
 Price ▾
 Type of heel ▾
 Pattern ▾

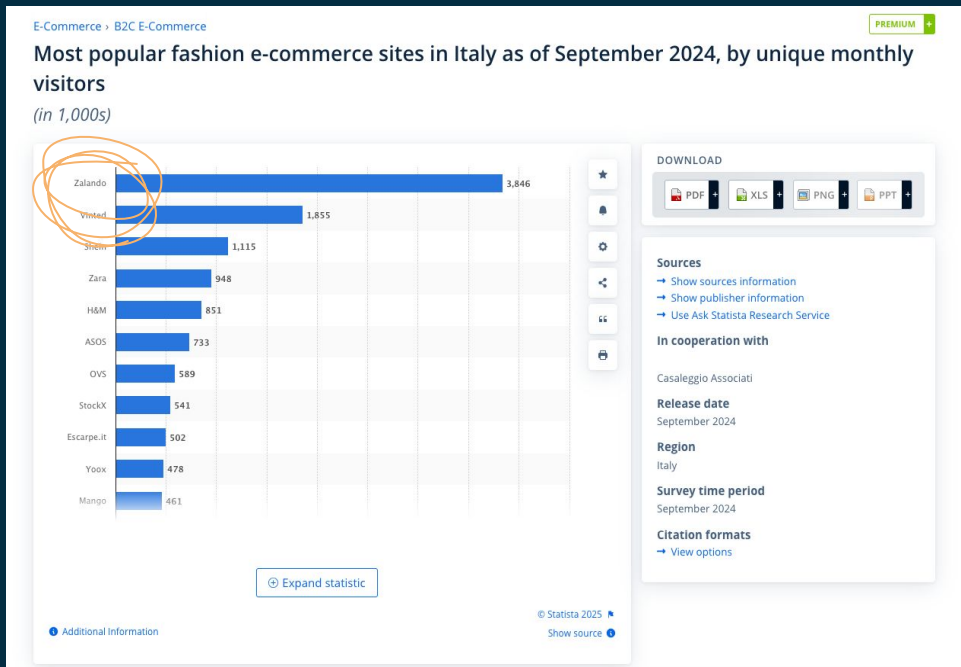
Product standard ▾
 Collection ▾
 Brand ▾
 Toe ▾
 [Show all filters](#)

75 items ①



- Leading Fashion Platform in Europe
- Founded 2008
- 50M+ Customers
- 10B+ Revenue
- 3K+ Software Engineers

Grazie mille!



source: <https://www.statista.com/statistics/1265105/most-popular-fashion-e-commerce-sites-italy/>

Agenda

1. 📡 Cloud Strategy
2. 🚀 Cloud Operations
3. ⚡ Cloud Incidents

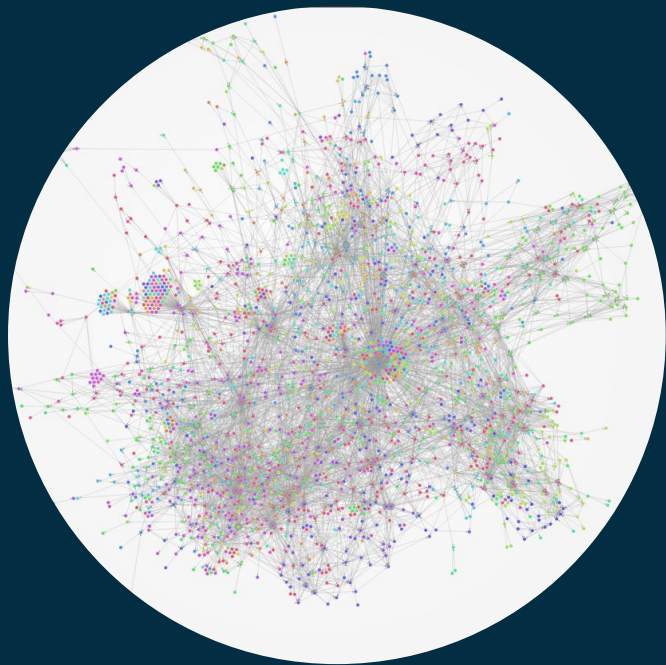




Cloud Strategy



Zalando Scale



Zalando Service Graph (2019)

Business

- 200K+ HTTP requests per second
- Thousands of orders per minute peak

Technology

- 3000+ Microservices
- 100K+ Containers
- 10K+ Nodes

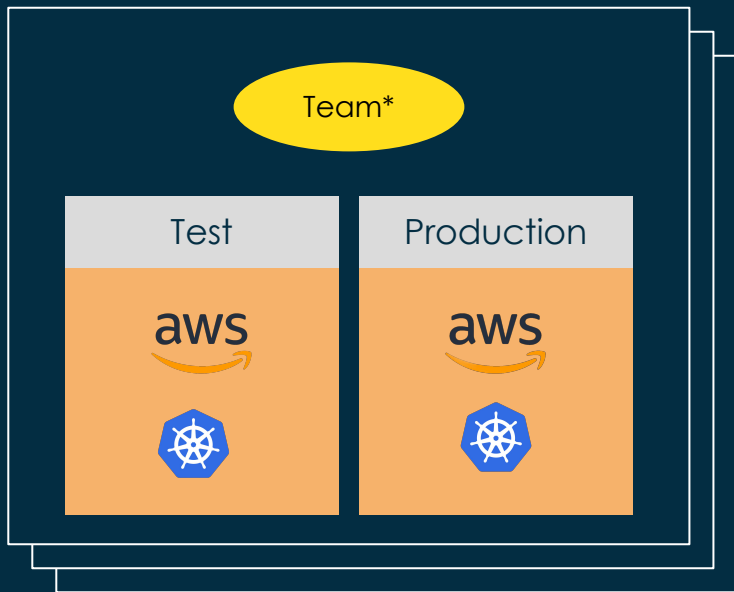
Zalando Cloud Strategy

- No Data Centers*
- Single Cloud (AWS)*
- Single Region (Frankfurt)*
- Everything in Kubernetes*
- Leverage AWS Managed Services**

*) almost **) responsibly



One Team – One Account – One Cluster



300 AWS Accounts, 250 K8S Clusters

Benefits

- Team Autonomy (Conway's Law)
- Reliability & Security
- Cost attribution
- Access management
- Avoid K8S scaling limits (2K nodes)

Drawbacks

- Higher Cost
- Lower Efficiency
- Requires Automation
- ...

Cloud Platform Services

The screenshot displays the Sunrise Cloud Infrastructure Developer Portal. The left sidebar contains navigation links: Search, Ask Sunrise, Create, Applications, APIs, Documentation, Teams, Tooling, Tech Radar, CyberWeek, Reliability, Compliance & Security, rkeep - Route Manager, CI/CD Pipelines, Infrastructure (selected), Cost Insights, Tech Insights, ML Platform, and Support. The main content area is titled 'Cloud Infrastructure' and 'AWS Accounts'. It features a table with columns: Account, Environment, Criticality, Cost Center, Account Group, and Tag Compliance. The table lists several AWS accounts, including 'acid', 'cloud-cost-efficiency', 'data', 'data-services', and 'stups-build', each with its respective environment (production), criticality (Tier 3), cost center, and tag compliance percentage. A search bar and filters for Environment (production), Criticality (Tier 3), and Cost Center are visible on the left.

Account	Environment	Criticality	Cost Center	Account Group	Tag Compliance
acid 263820974194	production	Tier 3	0001104321	acid	100%
cloud-cost-efficiency 299812085149	production	Tier 3	0001104321	cloud-cost-efficiency	100%
data 343040485429	production	Tier 3	0001104715	data	97%
data-services 710180679010	production	Tier 3	0001104321	data-services	50%
stups-build 861068367966	production	Tier 3	0001104715	stups-build	46%

Developer Portal
(Backstage.io)

```
apiVersion: v1
kind: Pod
metadata:
  name: testapp # name of the Pod
  labels:
    application: testapp # name of the application this Pod belongs to
    component: backend # name of the component of the application
spec:
  containers:
    # our Pod has just one container
    - name: testapp # name of the container
      image: container-registry-test.zalando.net/teapot/training-example:master-44
      ports:
        - containerPort: 8080
```

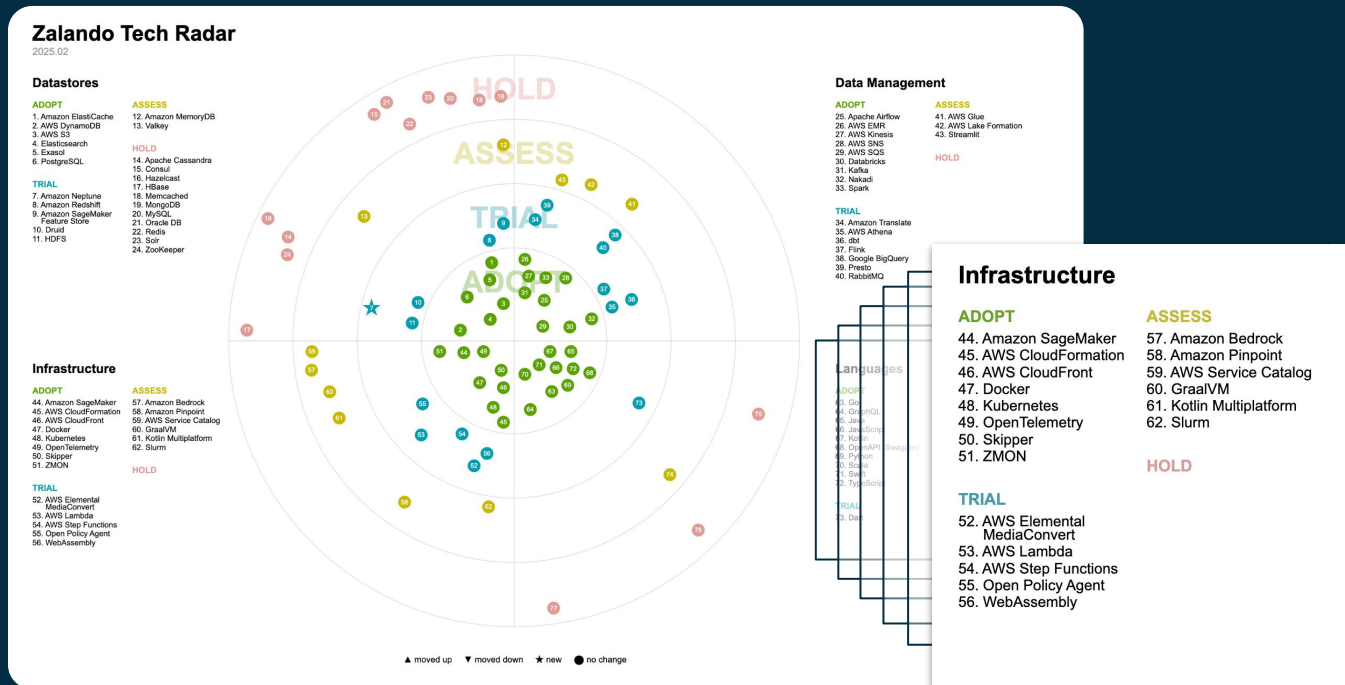
*.yaml on GitHub

```
pipx install zalando-cli-bundle --include-deps
installed package zalando-cli-bundle 24.17.0, installed using Python 3.12.8
These apps are now globally available
- docker-credential-pierone
- jp.py
- kio
- markdown-it
- natsort
- normalizer
- pierone
- piv
- pygmentize
- zalando-cli-bundle
- zaws
- zign
- zkubectl
- ztoken
done! 🌟🌟🌟
```

CLI Tools

Cloud Governance w/ TechRadar

<https://opensource.zalando.com/tech-radar/>





Cloud Operations



Cloud Operations - Scale



1600+
Postgres Databases

2300 Applications



250 Kubernetes Clusters

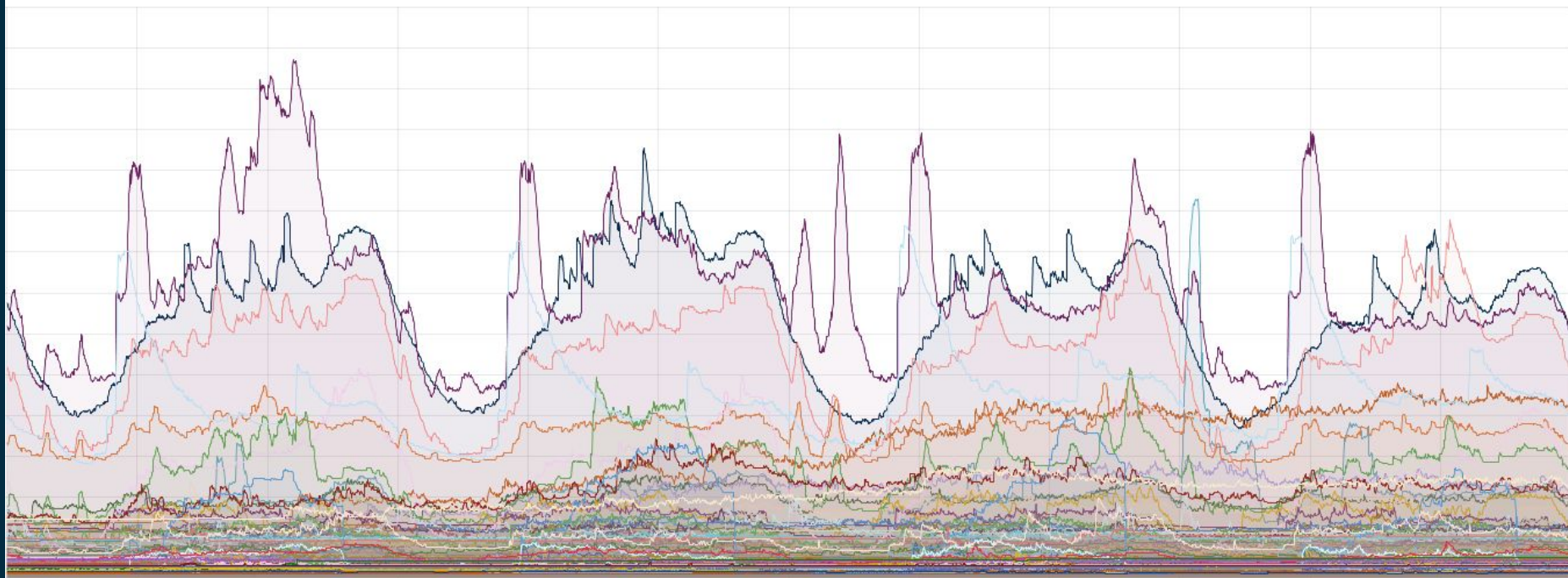
**Dedicated
workloads**



300+ AWS Accounts

Cloud Operations - Dynamic Scale

Number of nodes by Cluster



Cloud Operations - Original “Philosophy”

- **No pet Infrastructure**

Cluster and Account configuration is managed as code in central repositories.

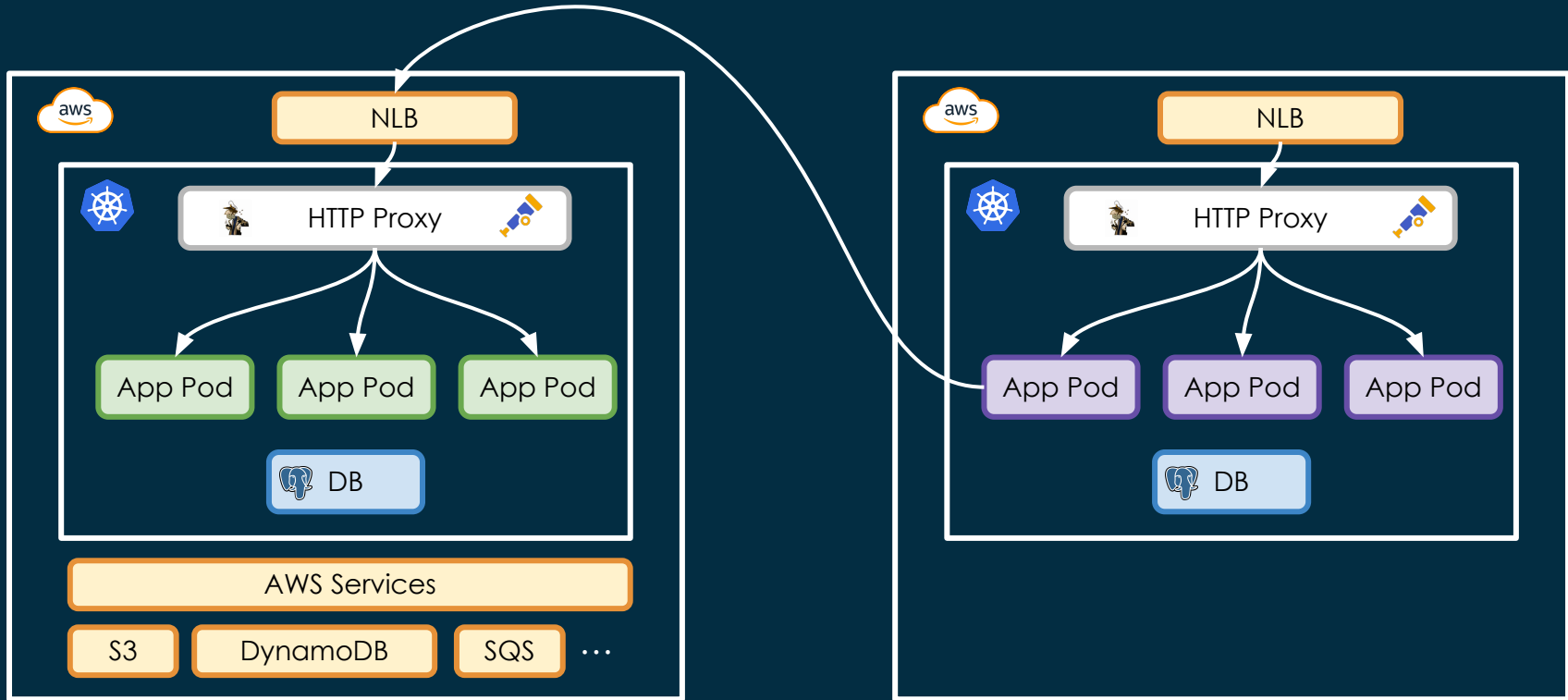
- **Always provide the latest stable Kubernetes version**

Latest features and quick to address possible vulnerabilities.

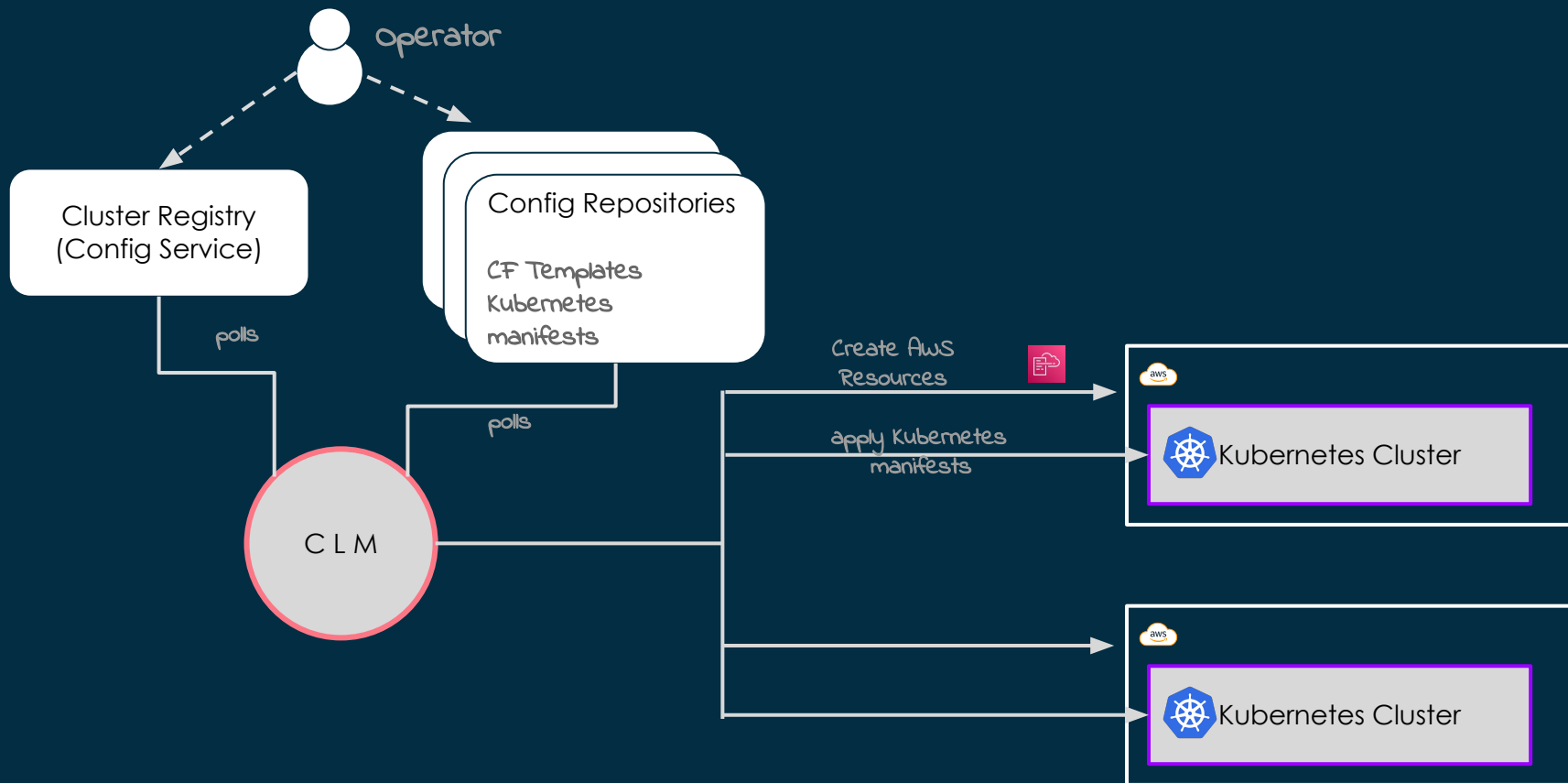
- **Continuous and non-disruptive cluster updates**

No maintenance windows.

Cloud Operations - Architecture



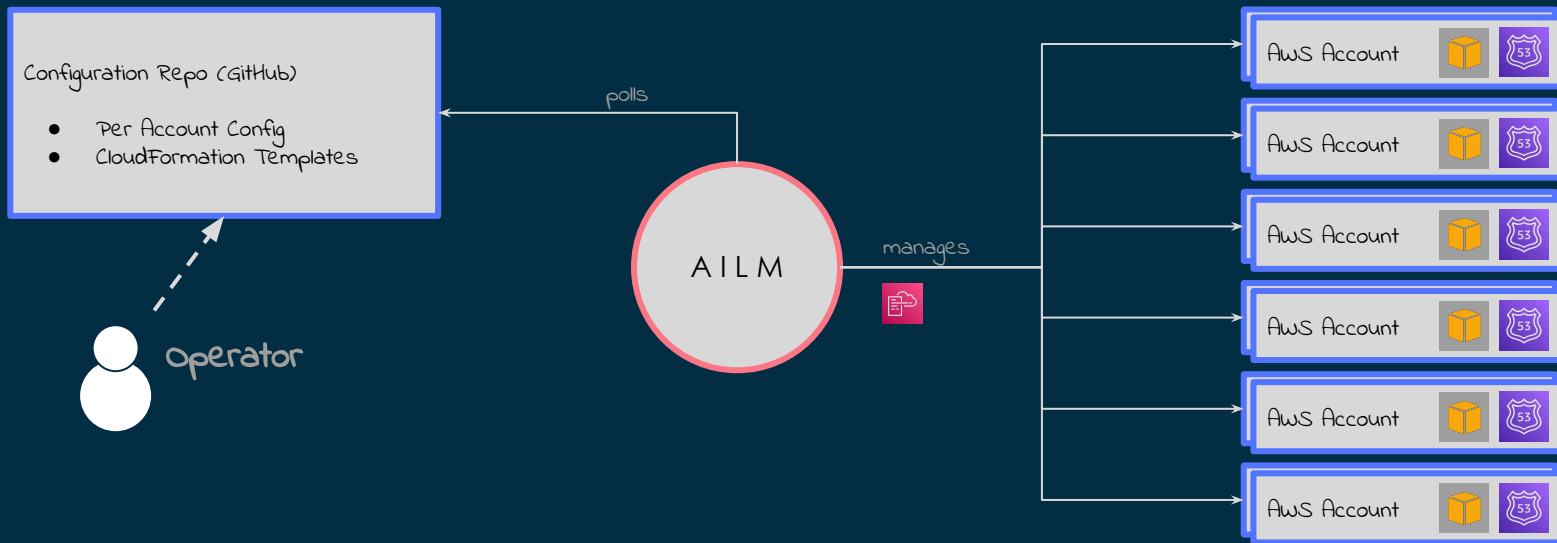
Cloud Operations - Infrastructure as Code (Kubernetes)



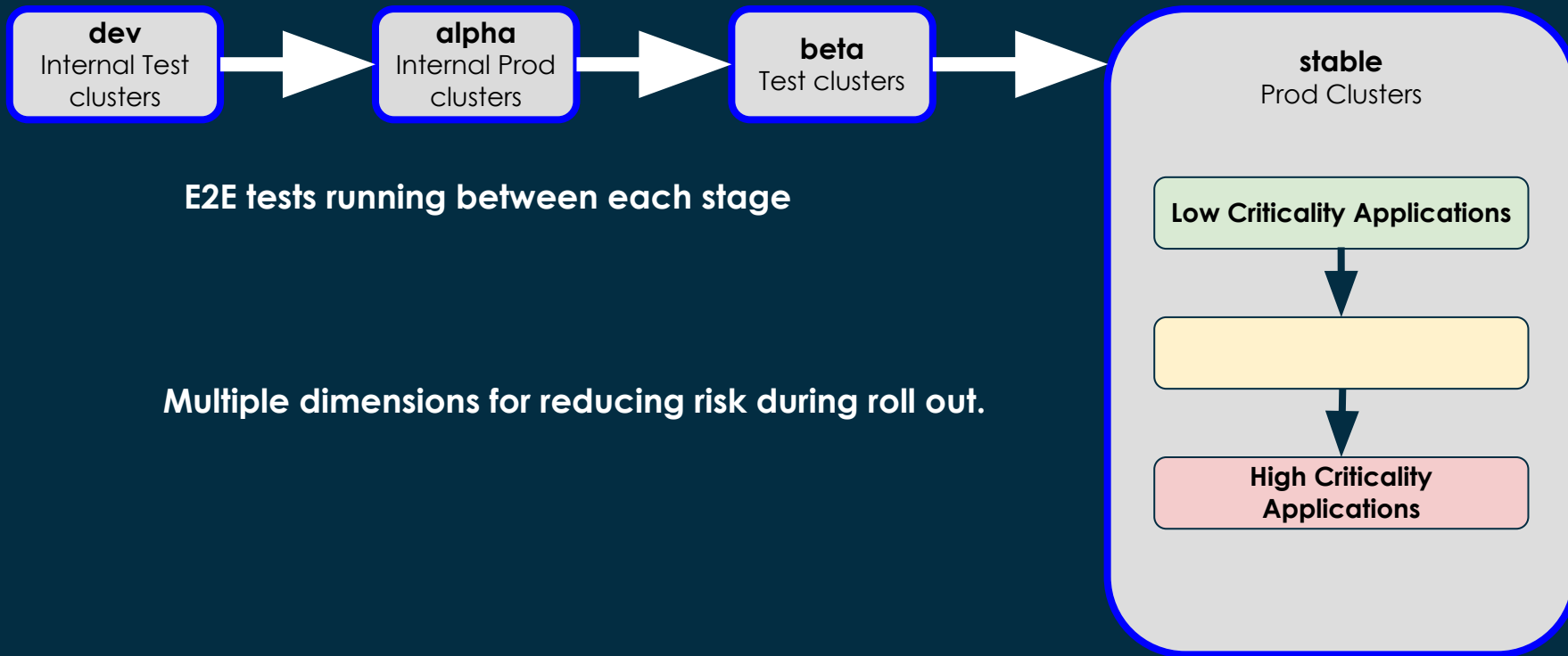
Cloud Operations - Infrastructure as Code (AWS)

Manages all Base Infrastructure

- VPC
- Hosted Zones
- Certificates
- IAM Policies

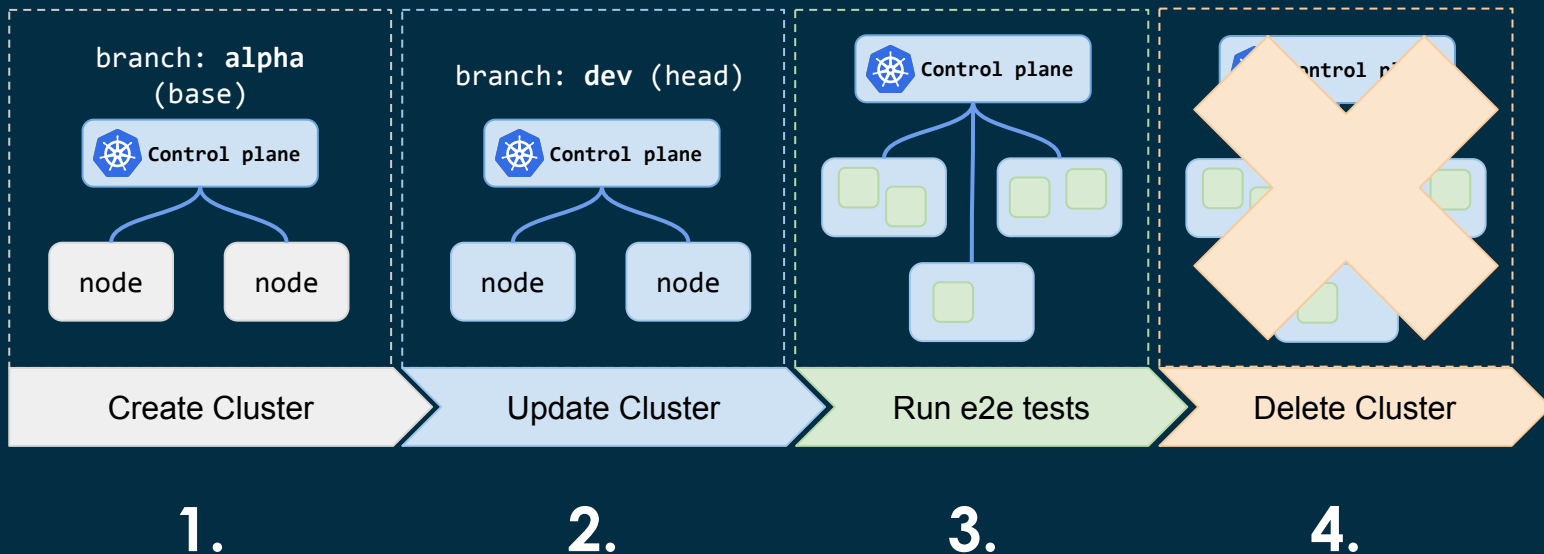


Cloud Operations - Staged roll out (Kubernetes)



Cloud Operations - e2e test on every change

Testing dev to alpha upgrade



⌚ 1 hour for a full e2e test run

Cloud Operations - “Philosophy” vs. Practice

- No pet Infrastructure
 - Almost everything in code.
 - Many per account/cluster configurations.
- Always provide the latest stable Kubernetes version
 - Oldest clusters were upgraded from Kubernetes v1.4 through to v1.31. (8.5 years)
- Continuous and non-disruptive cluster updates
 - Limiting monthly node rotations to reduce disruption of stateful workloads.

Cloud Incidents



The Setting

It's Tuesday, November 17th 2022

...

The Trigger

12:56 - 🚢 PR #716 gets merged

Remove metrics collector trust #716

Merged merged 1 commit into `master` from `fix_role` on Nov 17, 2022

Conversation 2 Commits 1 Checks 0 Files changed 5 +1 -13

Changes from all commits File filter Conversations 0 / 5 files viewed **Review changes**

Filter changed files

- Root
 - AWS-Org-Management
 - iam-apps-roles.cf.yaml
 - Core
 - iam-apps-roles.cf.yaml
 - iam-apps-roles.cf.yaml
 - route53-hostedzones-acm-c...
 - validate.sh

Root/AWS-Org-Management/iam-apps-roles.cf.yaml > 4 +1 -13 Viewed

Root/Core/iam-apps-roles.cf.yaml > 4 +1 -13 Viewed

Root/iam-apps-roles.cf.yaml > 4 +1 -13 Viewed

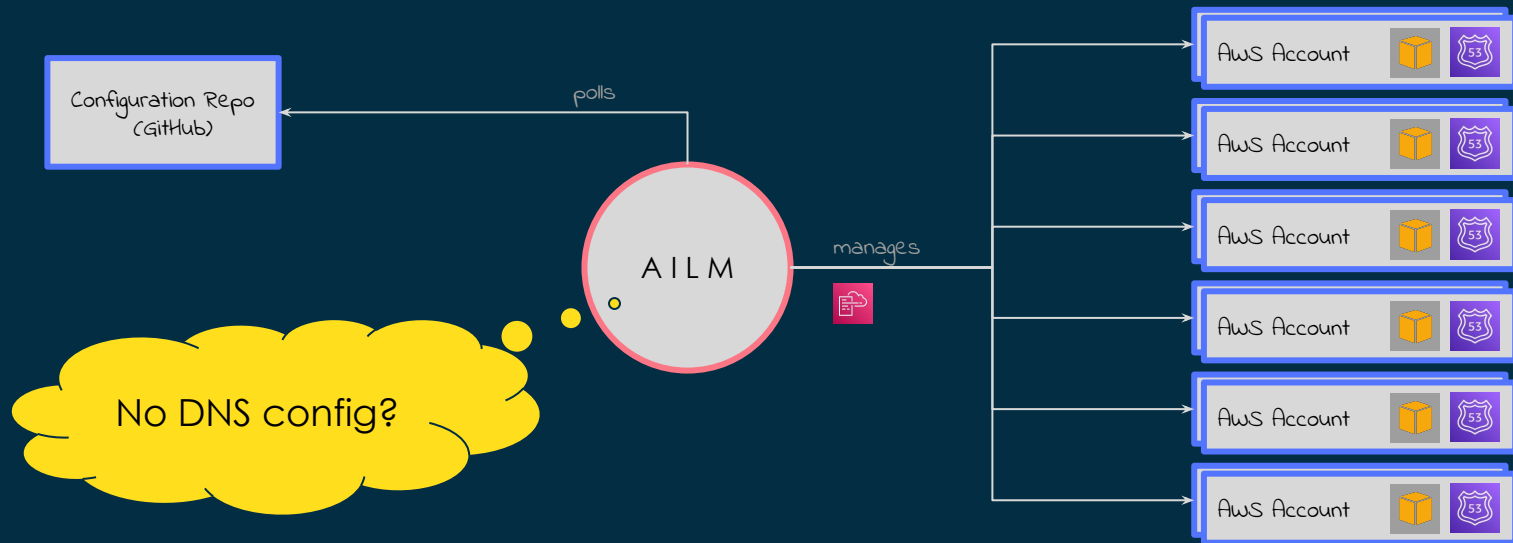
Root/route53-hostedzones-acm-certificates.cf.yaml > 2 +1 -13 Viewed

```
... @@ -1,4 +1,4 @@
1 - Metadata:
  1 + Metadata:
2   2 StackName: route53-hostedzones-acm-certificates
3   3 Scope:
4   4 Type: region
```

validate.sh > 0 +1 -13 Viewed

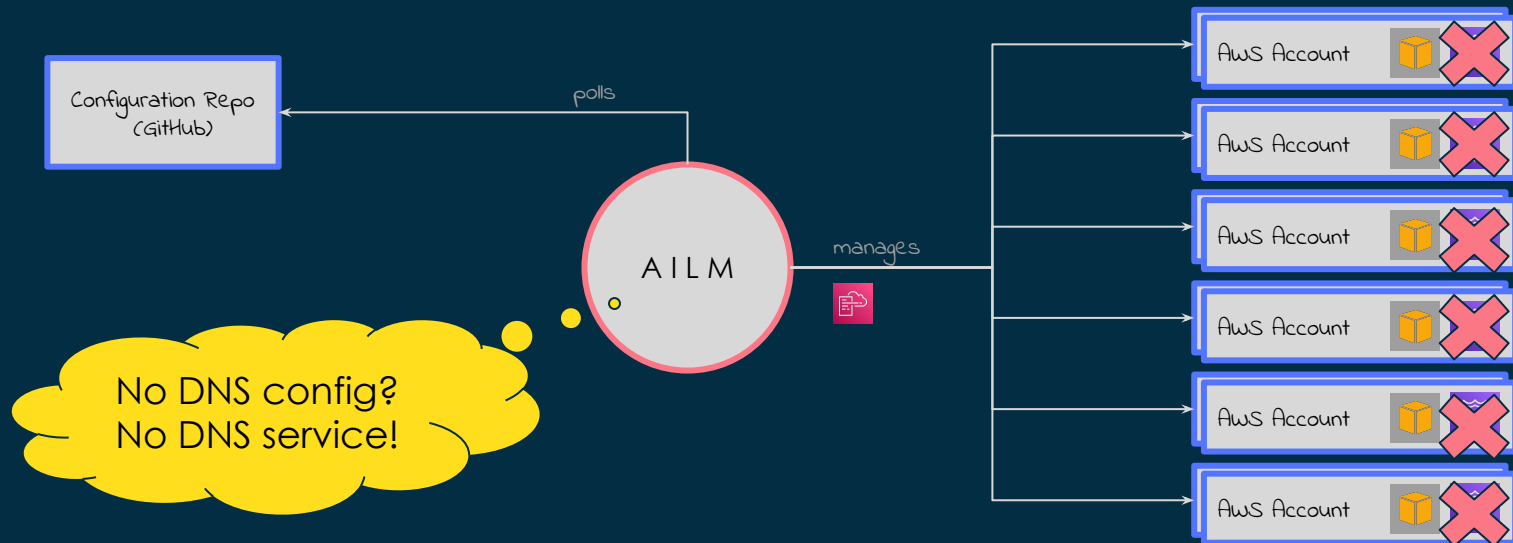
Automation takes over

12:58 - 🤖 AWS Lifecycle Manager processes change



Automation takes over

12:58 - 🤖 AWS Lifecycle Manager processes change



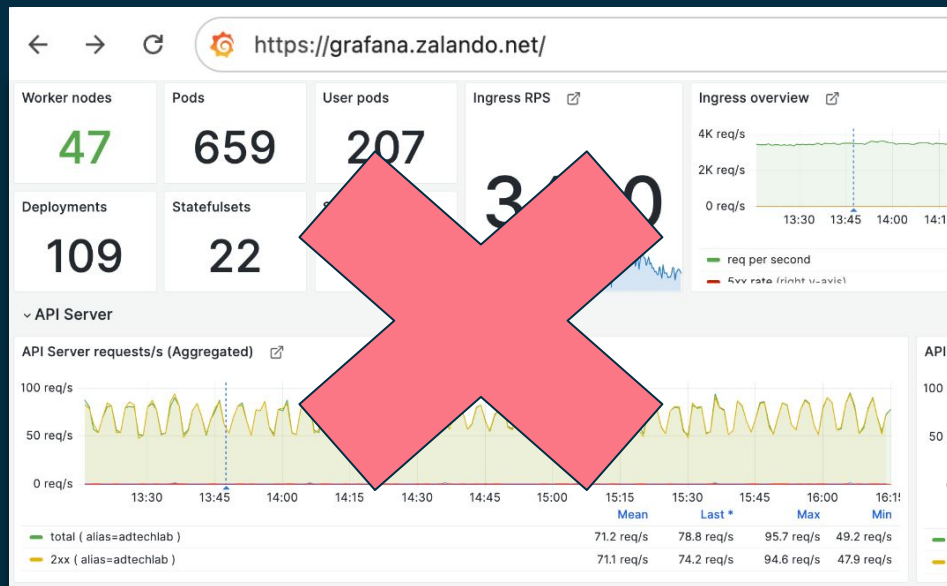
The Fallout

13:10 - 🔥 Zalando goes down



The Fallout

13:10 - 🔥 Monitoring goes also down



The Cleanup

- 13:23 - PR #716 is identified as culprit
- 13:38 - PR #716 is reverted - without success
- 13:40 - 🧑 Manual restoring of DNS entries starts
~ 200 Engineers in Incident Video Meeting

aws-chooser.zalando.net	18.156.81.28	18.198.223.174	52.57.226.107
emergency-access-service.stups.zalan.do	18.156.81.28	18.198.223.174	52.57.226.107
kube-2.stups.zalan.do			
legacy-teams-api.corporate-iam.zalan.do			
kube-web-view.zalando.net			
zmon.zalando.net			
registry.opensource.zalan.do			
pierone.stups.zalan.do			
container-registry.zalando.net			
info.services.auth.zalando.com			
platform-iam-tokeninfo.corporate-iam.zalan.do			
sunrise.platform-infrastructure.zalan.do			

prio		cname- Status		External DNS fixed?
product-availability	Rodrigo	Done	External DNS Logs	Yes
fulfillment	Katyanna	Done	External DNS Logs	Yes
tx-core	Rodrigo	Done	External DNS Logs	Yes
fashion-store	Katyanna	Done	External DNS Logs	Yes
information-experience	Noor	Done	External DNS Logs	Yes
customer-data-platform	Mahmoud	Done	External DNS Logs	Yes
dcis	Noor	Done	External DNS Logs	Yes
search	Katyanna	Done	External DNS Logs	Yes
inbox	Martin	Done	External DNS Logs	Yes
post-purchase	Martin	Done	External DNS Logs	Yes
checkout	Noor	Done	External DNS Logs	Yes
cart	Zak	Done	External DNS Logs	Yes
db	Zak	Done	External DNS Logs	Yes

Restoring Service

- 14:30 - 🛠 Monitoring & Ops tools start to recover
- 14:53 - 👠 Catalog starts to recover
- 15:41 - 💰 Order processing start to recover
- 17:00 - 📦 Warehouse services are restored
- 20:02 - ✅ All systems fully recovered



The Learnings

- **Safeguard Infrastructure Changes**

Validation, Preview, Staged-rollout.

- **Tighten Deployment Policy**

Applies to configuration (not just code) and tooling that *can* impact production.

- **Segregate Infrastructure for Monitoring & Ops Tooling**

Separate rollout stages for internal services.

- **Blameless Post Mortem Culture**

Human error is never the root cause. Fix process to catch mistakes.

Infrastructure Incident Patterns

- **Sharp Edges with Running K8S**

⇒ Fail, Fix, Repeat

⇒ Upstream code & share learnings <https://k8s.af/>.

- **Change Failures with "Supertools"**

⇒ Validation, Preview & Staged rollout

- **Overload of Kubernetes Backplane**

⇒ Overscale. Move to AWS EKS.

- **Upstream AWS Incidents** (Networking Problems, AZ Outages, etc.)

TL;DR

- Single Cloud, Single Region, K8S
- No pet Infrastructure!
- One Cluster per Team can work!
- Cloud automation needs safeguards!

THANK YOU!

