Transforming Production Readiness

Panagiotis (Panos) Moustafellos

SREcon24 EMEA, Dublin - Oct 31st, 2024

Hello

Panagiotis Moustafellos

Distinguished Engineer - SRE

Systems engineer with 20 years of experience in diverse tech environments.

Main areas of expertise around systems architecture, observability, and security.
Scaling software systems and infrastructure.
Mentoring and growing newer generations of technical leaders.

Until recently the overall technical lead of the SRE organisation at Elastic.

Currently, a Distinguished Engineer in the Observability space at Elastic. Building products for the SRE practitioner & driving the production readiness transformation for Elastic Cloud.

elastic

What are we covering?

Operational environment and business context

Strategic planning for change

Service offering boundaries & limitations

Insights on SLOs at scale

Phased rollout approaches, readiness criteria

Empowering engineers to carry the pager for their services

Takeaways - Lessons learned

elastic

Business context

$1.26B

Revenue

Total revenue FY24

32%

%YoY Growth

Growth of Elastic Cloud business

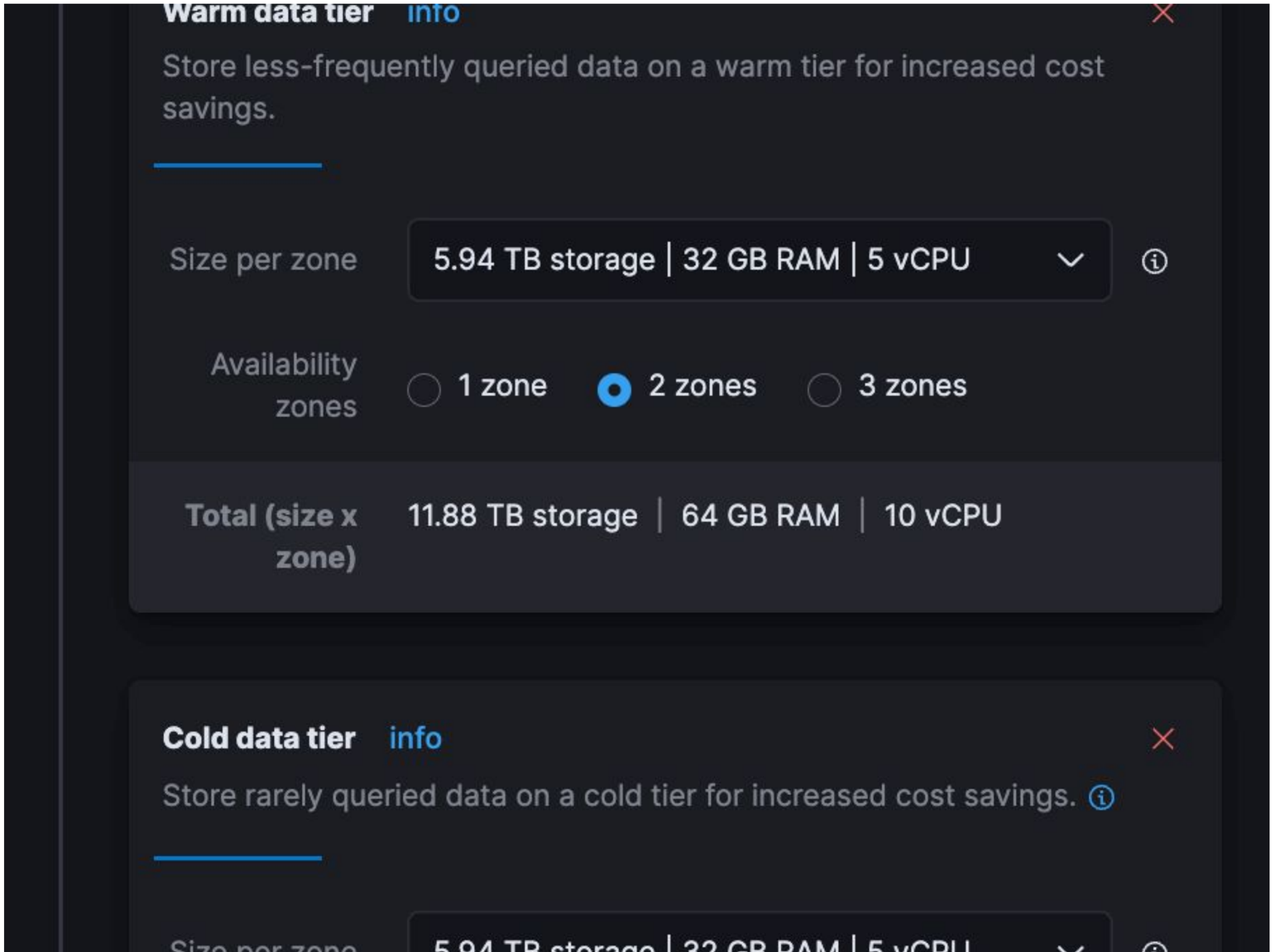
$1.43B

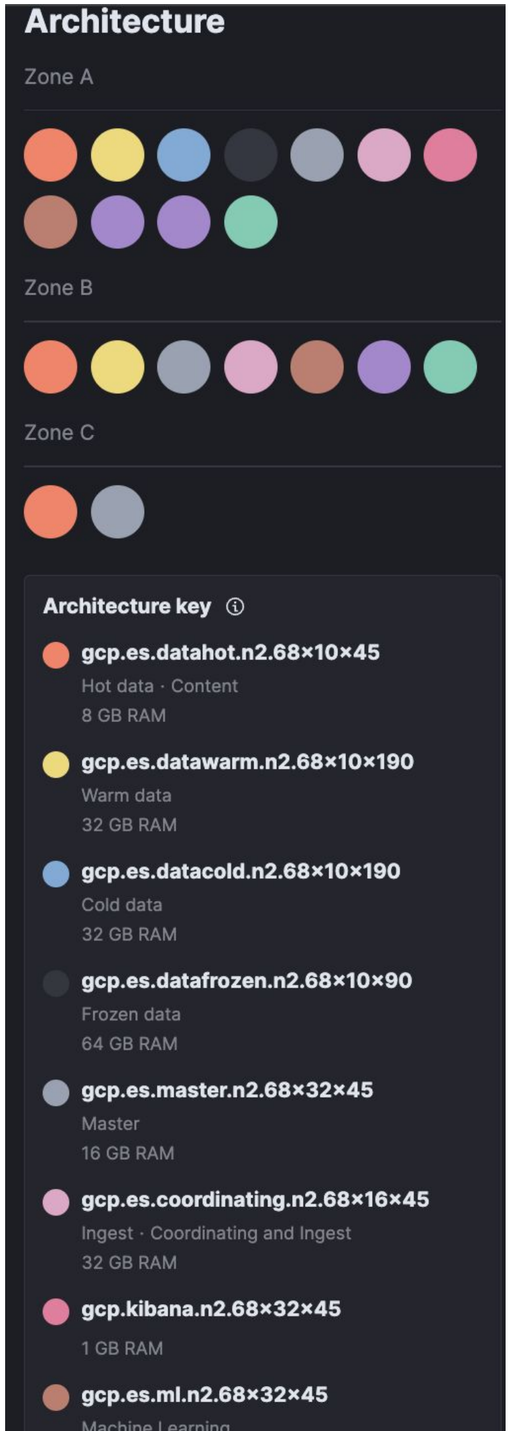
Projected Rev

Revenue guidance for FY25

elastic

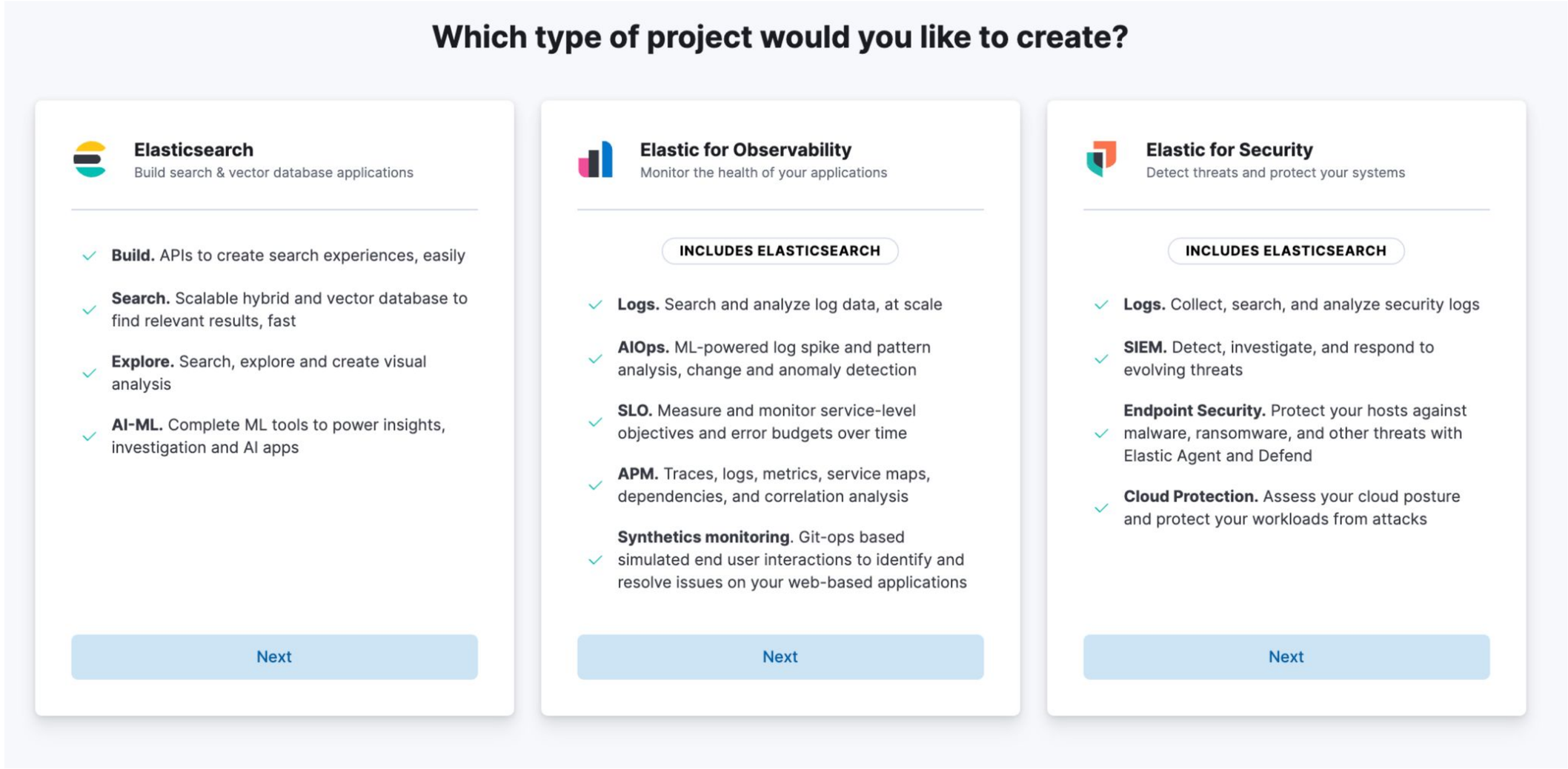
Elastic Cloud - Hosted service





elastic

Elastic Cloud - Serverless



elastic

What does that change mean operationally?

Responsibility & business model shift

- Customer must not care at all about the operational aspect
  - Physicality is abstracted away from the customer
  - They are no longer co-responsible for the operation of their service
  - Scaling and **reliability** are system properties
- Consumption based pricing
  - Business and customer incentives align

What does that change mean operationally?

Architectural **changes** to support the new model

- Cell-based architecture for infrastructure and services
- Multi-tenancy shift
- Robust infrastructure and private networking substrate
- Replatforming to Kubernetes
- Separating Compute from Storage, separating Ingestion from Search
- Autoscaling in many dimensions (and future scale to zero)
- Regional failure domain, System for global configuration w/ RPO = 0
- System for inferring customer happiness E2E, on a per individual customer basis
- Improvements in production readiness, change management and safety

elastic

At what scale

60+

Geo regions

Spanning all major geographies, multiple points of presence

10sK

Customers

Dozens of thousands of customers

58

SRE

An organisation of 7 SRE teams totalling ~60 engineers

elastic

The SRE team could never* scale to support that operating model.

And they shouldn’t*.

The engineering teams building the systems must* carry the pager.

So what’s next?

Preparing for organizational change

Navigating complexity, dispelling FUD

- Senior engineering leadership seeks alignment
- Once aligned, determine the initiative's leader
- Provide consulting services from SRE teams around on-call structure models
- Working with HR, Legal, and Finance
  - Navigating local laws, compliance, compensation for on-call
- Coming up with a **rollout plan**
- Testing the waters and iterating

Drawing your “box”

Your product’s or service’s capabilities live in the “box”

- **Purpose of known Limitations, Quotas**
  - Aligning engineering, support, and field teams on expectations
  - Provide reasonable SLAs
- **Impact on customer onboarding**
  - Preventing mismatches with high-demand use cases; directing some customers to other options
- **Continuous improvement strategy**
  - Expanding the “box” as automation and self-healing capabilities grow

elastic

Production Readiness Review (PRR) checklist

Provide actionable guidance to service owning teams

- It’s simple and efficient to provide a PRR checklist in the form of conformance criteria, universally examined for each service
- Sections can include:
  - Key documents and reports
  - Service overview and architecture
    - Durability review, threat modelling, failure domains
  - Operational mechanisms and review venues
  - Monitoring (including SLOs) and Incident response
  - Change management safety
  - Scalability
  - Development & test

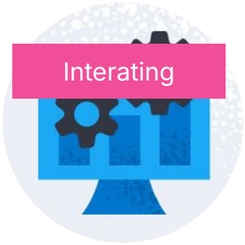
elastic

Phased product launches

Private beta

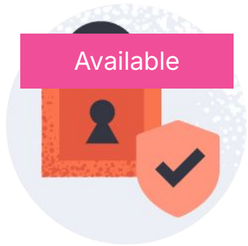
Onboarding sets of “friendly” customers.

Close and iterate any open PRR items, establish procedures to respond on high SLO burn rates. Exercise on-call and incident response



Public beta

Gated and then ungated public beta launch.
Iterating on operational improvements with any findings. Adjusting the “box”



General availability

SLAs apply.
Service is ready to accept production workload.
Operational workload (Toil / KTLO) from previous phase is at reasonable levels

elastic

Drawing clear escalation lines

Clear incident and alert management flows, with automated routing to owning team or simple decision matrix to triage

- **Challenges:** Who owns what? Who is on call? When and who do I escalate to?
- **Improvements:**
  - Service inventory
  - All alerts with attached runbooks and routing
  - Proactive and reactive flows for alert, incident and case management
  - On-call management tools available to all engineering leaders
- **Continuous improvement:** Data-driven decisions given alert and escalation rate analysis

Revamping incident management processes

Senior SREs are still Incident Commanders for high Sevs, but every engineering team should manage their incidents

- **Challenges:** Need for clarity, role alignment, and rapid incident resolution
- **Improvements:** Simplified severity levels, RCA format update, on-call training for development teams
- **Incident management analytics:** Leveraging data for incident trends and service enhancements
- **Automation and tooling:** Leveraging a vendor solution to simplify the flows of incident management, making them accessible to engineers on call

Empowering engineering teams

IDP, CI/CD, and Workflow execution systems

- **Key systems:**
  - IDP, service inventory & integrations with other systems
  - Managed CI/CD pipeline
    - Progressive rollouts (QA → Stg → Prod canarying → Prod)
  - GitOps-controlled software delivery, and defaults overrides
  - Quality gates w/ automated SLO burn rate checks
  - Workflow execution
    - Controlled debugging in prod, ad hoc runbook automation
  - Vulnerability management (deps scanning, supply chain checks, SAST)
- **Outcome:** Ensuring reliable, low-risk deployment paths for development teams

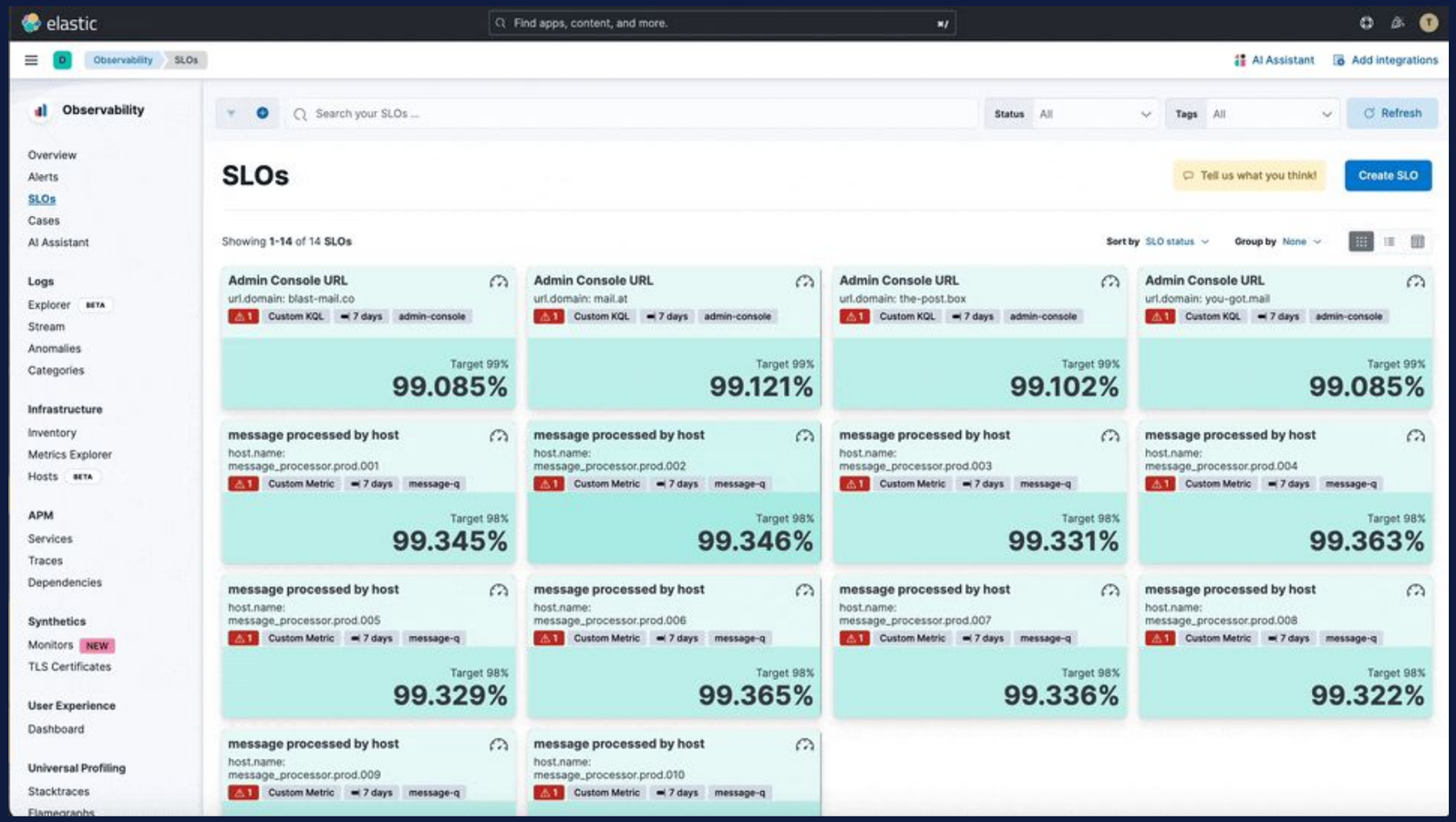
Empowering engineering teams

Observability “as a Service”

- **Self-service model**
  - All services in conformance, have their o11y signals automagically pushed in our Observability solution
- **Capabilities for teams:** Defining SLOs, dashboards, alerting, and monitoring for proactive and reactive incident management
- **GitOps management:** Everything as-code
- **Single pane of glass:** O11y signals and data sources are spread across the globe, however **everything is accessible through one Kibana interface** through Cross-Cluster search

SLOs as a common language

- Proactive and strategic approach to monitor and maintain Service Level Objectives (SLOs).
- Customer focused. Ensures that your service meets predefined performance standards and user expectations.
- Baseline metrics and highlight how changes affect those measurements
- **Act as a communication layer with common language to align across teams and business goals. At high level they should translate to customer happiness**



elastic

Elastic's SRE team maintains a single pane of glass cluster for the Observability needs of Elastic Cloud.

Connecting over 180 Elasticsearch clusters across the globe, through Cross-Cluster Search.

Holding PBs of data, Ingesting over 500 TB/day

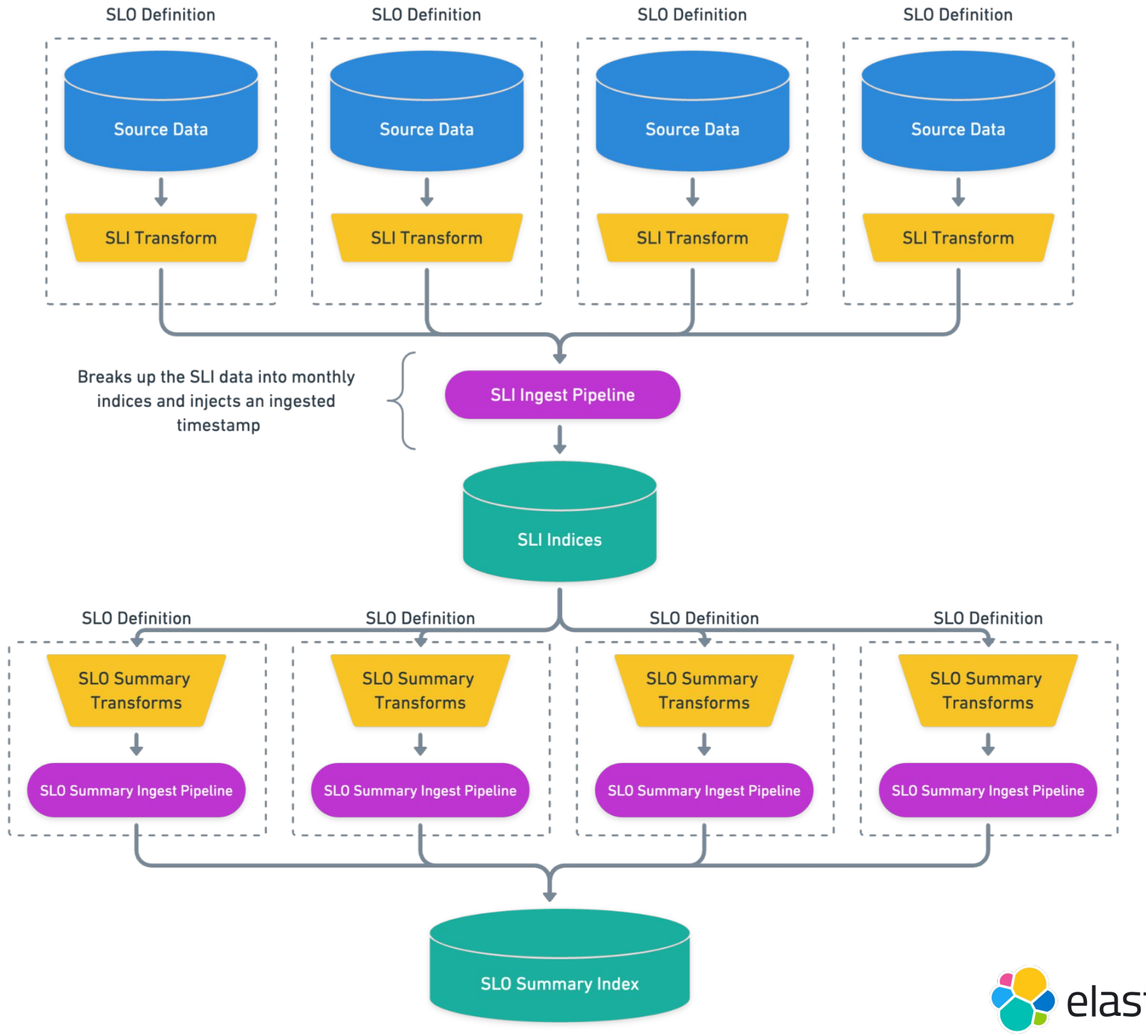
As of 2hrs ago, over 900k SLOs are tracked

elastic

Basis of Architecture

SLOs is powered by the transform service which built on search and indexing primitives. It's scale is dependent on the scalability of Elasticsearch.

- The transform service is a GA Elasticsearch feature
- To fulfill requirements, we need to transform data into:
  - Number of good events
  - Number of total events
  - Timeslices
- Doesn't require an additional service to build or manage.
- Transforms naturally allowed us to implement the "group by" to enable SREs to manage large amounts of SLOs



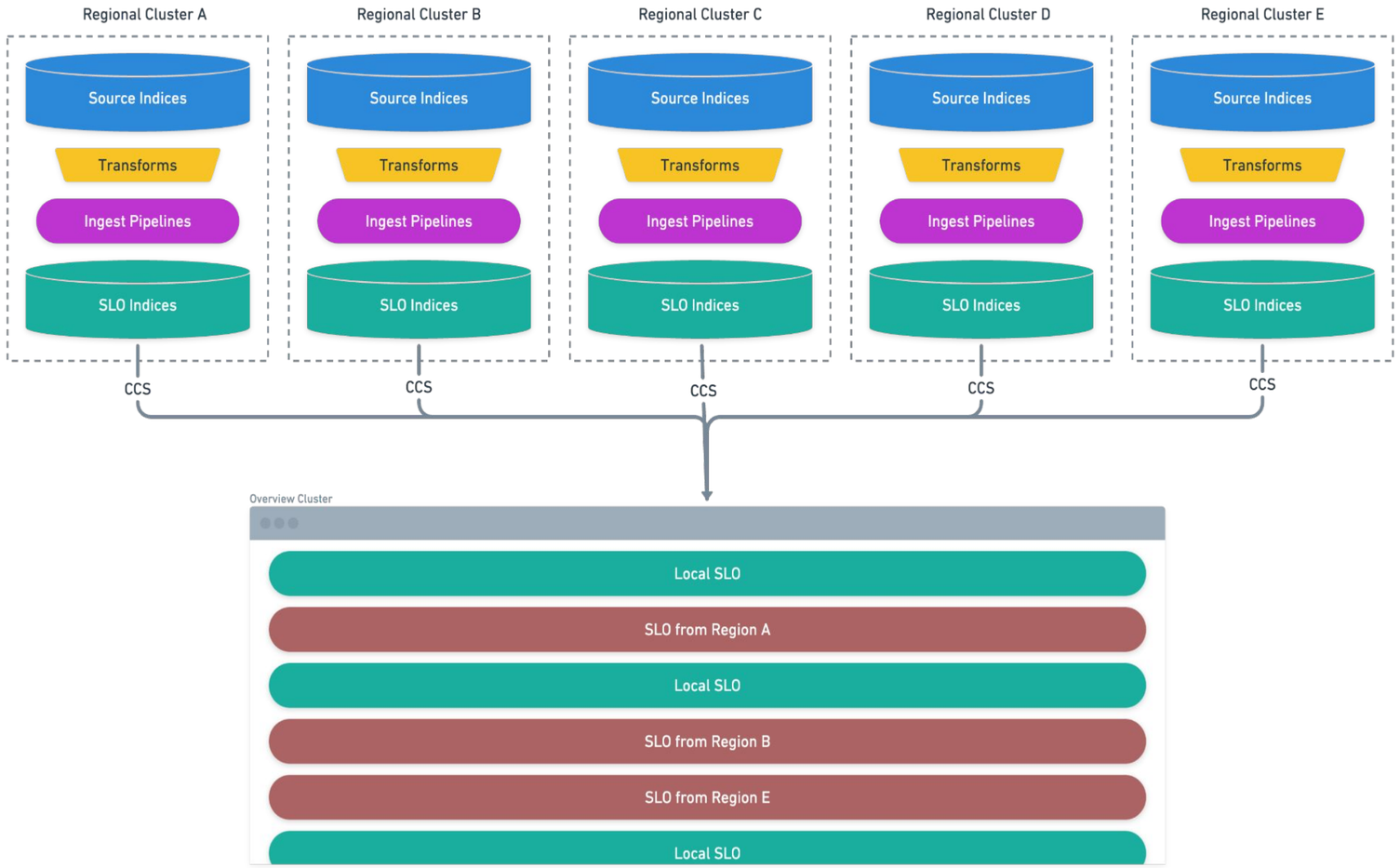
elastic

Federated SLOs to work in a CCS environment

Reducing the amount of query traffic between an overview cluster via CCS and the source data. Only the summarized data is queried via CCS.

This is accomplished by running the transform as close to the source data as possible while still giving SREs and all engineering teams alike, the same “single pane of glass” experience.

We annotate the details of the remote cluster so it’s clear to the user where the original SLO and source data is stored.

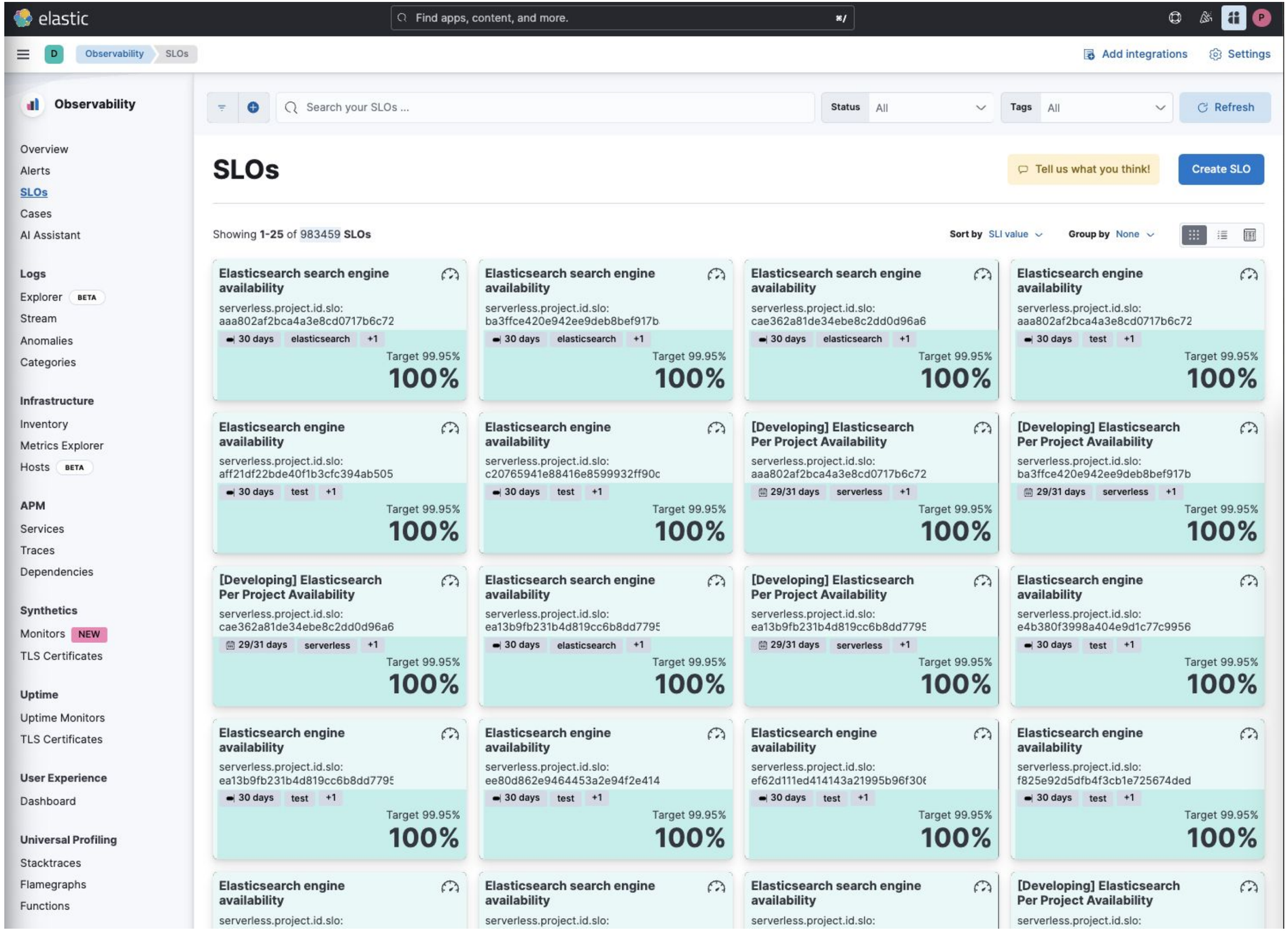


The Elastic Cloud “Overview” cluster is one of the largest Cross-Cluster Search environment in the world.

Using transforms extensively there was a real battle-test for the SLO product

This was testing both SLI transforms and CCS _search primitives.

elastic


elastic
Find apps, content, and more.
Observability
SLOs
Add integrations
Settings
Observability
Overview
Alerts
SLOs
Cases
AI Assistant
Logs
Explorer BETA
Stream
Anomalies
Categories
Infrastructure
Inventory
Metrics Explorer
Hosts BETA
APM
Services
Traces
Dependencies
Synthetics
Monitors NEW
TLS Certificates
Uptime
Uptime Monitors
TLS Certificates
User Experience
Dashboard
Universal Profiling
Stacktraces
Flamegraphs
Functions
Search your SLOs ...
Status All
Tags All
Refresh
SLOs
Tell us what you think!
Create SLO
Showing 1-25 of 983459 SLOs
Sort by SLI value
Group by None
Elasticsearch search engine availability
serverless.project.id.slo: aaa802af2bca4a3e8cd0717b6c72
30 days elasticsearch +1
Target 99.95%
100%
Elasticsearch search engine availability
serverless.project.id.slo: ba3ffce420e942ee9deb8bef917b
30 days elasticsearch +1
Target 99.95%
100%
Elasticsearch search engine availability
serverless.project.id.slo: cae362a81de34ebe8c2dd0d96a6
30 days elasticsearch +1
Target 99.95%
100%
Elasticsearch engine availability
serverless.project.id.slo: aaa802af2bca4a3e8cd0717b6c72
30 days test +1
Target 99.95%
100%
Elasticsearch engine availability
serverless.project.id.slo: aff21df22bde40f1b3cfc394ab505
30 days test +1
Target 99.95%
100%
Elasticsearch engine availability
serverless.project.id.slo: c20765941e88416e8599932ff90c
30 days test +1
Target 99.95%
100%
[Developing] Elasticsearch Per Project Availability
serverless.project.id.slo: aaa802af2bca4a3e8cd0717b6c72
29/31 days serverless +1
Target 99.95%
100%
[Developing] Elasticsearch Per Project Availability
serverless.project.id.slo: ba3ffce420e942ee9deb8bef917b
29/31 days serverless +1
Target 99.95%
100%
[Developing] Elasticsearch Per Project Availability
serverless.project.id.slo: cae362a81de34ebe8c2dd0d96a6
29/31 days serverless +1
Target 99.95%
100%
Elasticsearch search engine availability
serverless.project.id.slo: ea13b9fb231b4d819cc6b8dd7795
30 days elasticsearch +1
Target 99.95%
100%
[Developing] Elasticsearch Per Project Availability
serverless.project.id.slo: ea13b9fb231b4d819cc6b8dd7795
29/31 days serverless +1
Target 99.95%
100%
Elasticsearch engine availability
serverless.project.id.slo: e4b380f3998a404e9d1c77c9956
30 days test +1
Target 99.95%
100%
Elasticsearch engine availability
serverless.project.id.slo: ea13b9fb231b4d819cc6b8dd7795
30 days test +1
Target 99.95%
100%
Elasticsearch engine availability
serverless.project.id.slo: ee80d862e9464453a2e94f2e414
30 days test +1
Target 99.95%
100%
Elasticsearch engine availability
serverless.project.id.slo: ef62d111ed414143a21995b96f30
30 days test +1
Target 99.95%
100%
Elasticsearch engine availability
serverless.project.id.slo: f825e92d5dfb4f3cb1e725674ded
30 days test +1
Target 99.95%
100%
Elasticsearch engine availability
serverless.project.id.slo:
Elasticsearch search engine availability
serverless.project.id.slo:
Elasticsearch search engine availability
serverless.project.id.slo:
[Developing] Elasticsearch Per Project Availability
serverless.project.id.slo:

Operational KPI reviews

Keeping the pressure on operational excellence

- **Review mechanisms:**
  - Creating venues for operational insights and continuous learning
    - High level one - CxOs, VPs, Directors, PMs, Tech Leads, Engineers
    - Lower level ones - Organizational, or even team level
- **Customer-Facing SLOs and KTLO / Toil review:** Examining incidents, and SLO compliance, KTLO trends
- **Post-mortem reviews:** Examining the Action Items and takeaways from recent post-mortems
- **Leveraging analytics:** Incident trend analysis and KPI tracking to drive improvement

Takeaways

Engage leaders early to prepare for the change. Give them the steering wheel

Simplify & communicate strategy

Know your first team

Reliability is a system property

Have supporting systems and processes to enable ownership

Give Things for Free(™)

Empowered eng teams love* on call

Thank you

@pmoust

elastic

Transforming Production Readiness

Panagiotis (Panos) Moustafellos

SREcon24 EMEA, Dublin - Oct 31st, 2024