“THE BIG ONE”

Ajuna Kyaruzi

Marketwatch.com/story/the-big-one-is-a-threat-even-if-you-dont-live-in-california-2018-05-18

Datadog capabilities

Infrastructure Monitoring

- Containers
- Serverless
- Network Performance Monitoring
- Network Device Monitoring
- Cloud Cost Management

Application Performance Monitoring

- Distributed Tracing
- Error Tracking
- Continuous Profiler
- Database Monitoring
- Universal Service Monitoring

Digital Experience Monitoring

- Synthetics
- Real User Monitoring
- Session Replay

Log Management

- Observability Pipelines
- Sensitive Data Scanner
- Audit Trails
- Log Forwarding

Security

- Cloud Security Management
- Application Security Management
- Cloud SIEM

Developer Experience

- CI Visibility
- Continuous Testing

Watchdog AI

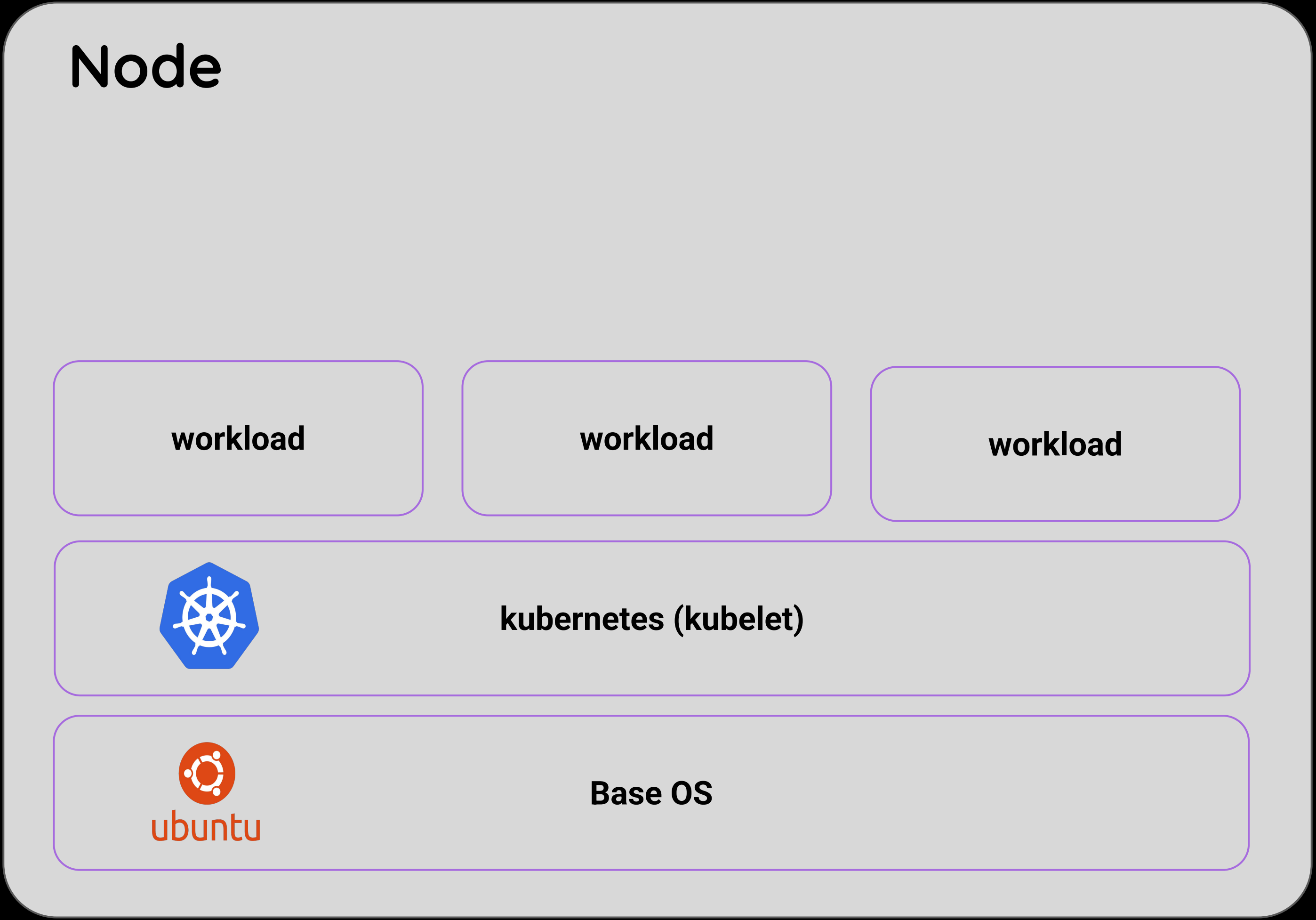
Insights ● Impact Analysis ● Root Cause Analysis ● Anomaly Detection ● Alerts ● Correlation ● Optimizations

Shared Platform Services

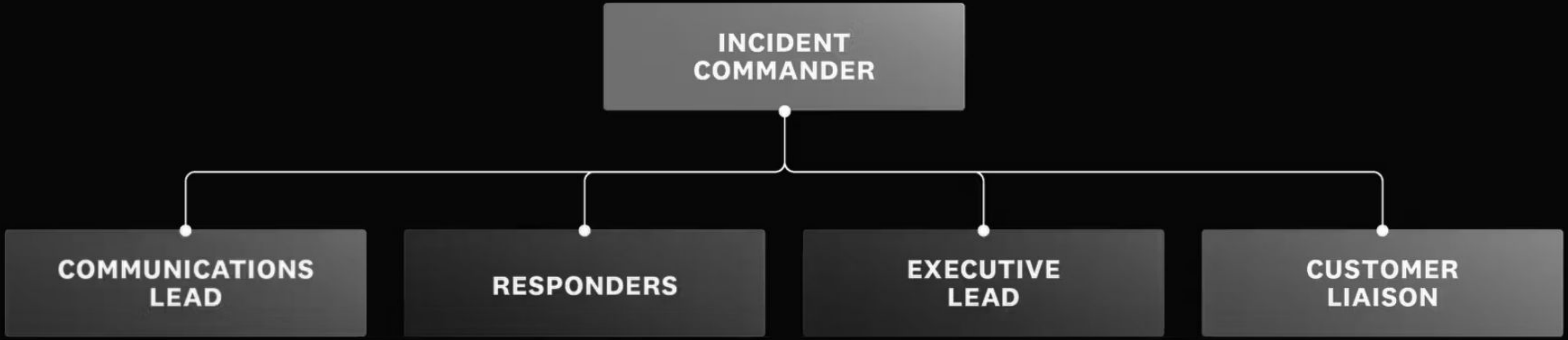
Collaboration ● Dashboards ● Mobile ● Agents ● Notebook ● Workflows ● Open Telemetry ● Service Catalog

UNIFIED METRICS, LOGS, TRACES

- 06:00 UTC: Systemd upgrade starts, triggering the outage
- 06:03 UTC: Monitoring detects a problem with Datadog
- 06:08 UTC: Two engineering teams are paged
- 06:18 UTC: High-severity incident is opened
- 06:23 UTC: Incident commander joins response
- 06:27 UTC: Executive on call joins response
- 06:31 UTC: First status page update is posted
- 06:32 UTC: Global outage is officially diagnosed
- 06:40 UTC: Additional responders join for triage


Node
workload
workload
workload
kubernetes (kubelet)
Base OS
ubuntu

Incident Response



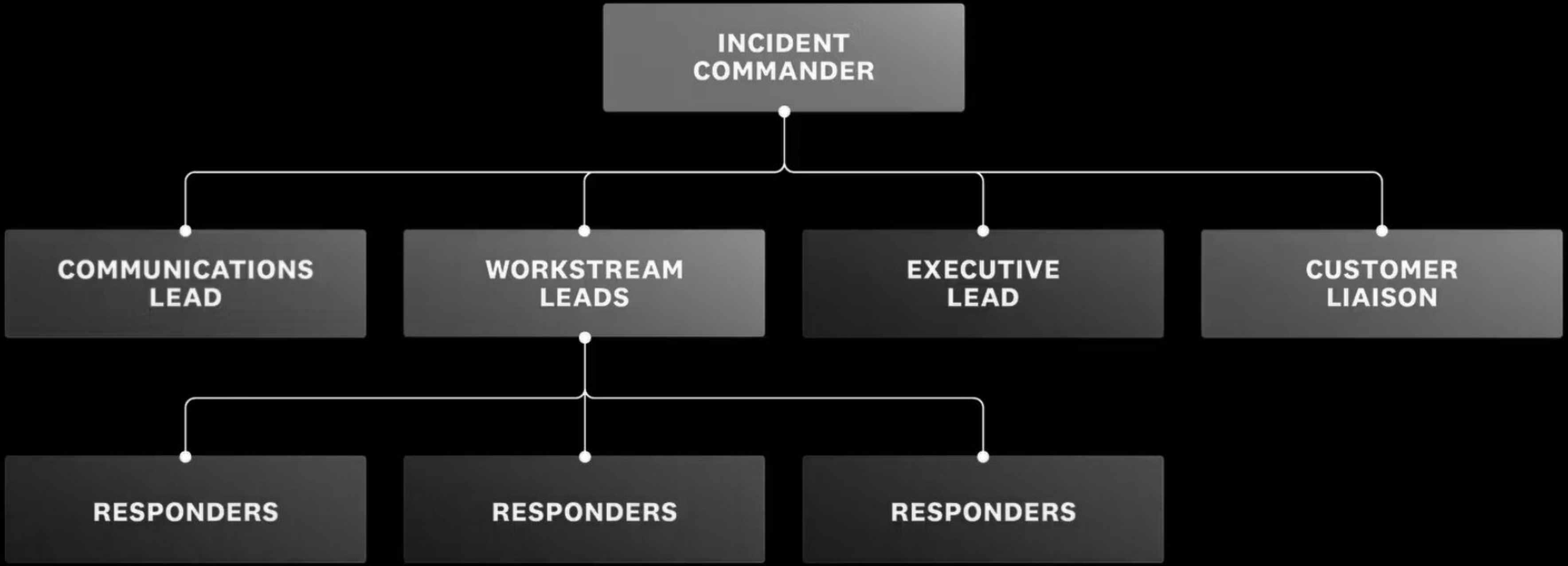
INCIDENT COMMANDER

COMMUNICATIONS LEAD

RESPONDERS

EXECUTIVE LEAD

CUSTOMER LIAISON



INCIDENT COMMANDER

COMMUNICATIONS LEAD

WORKSTREAM LEADS

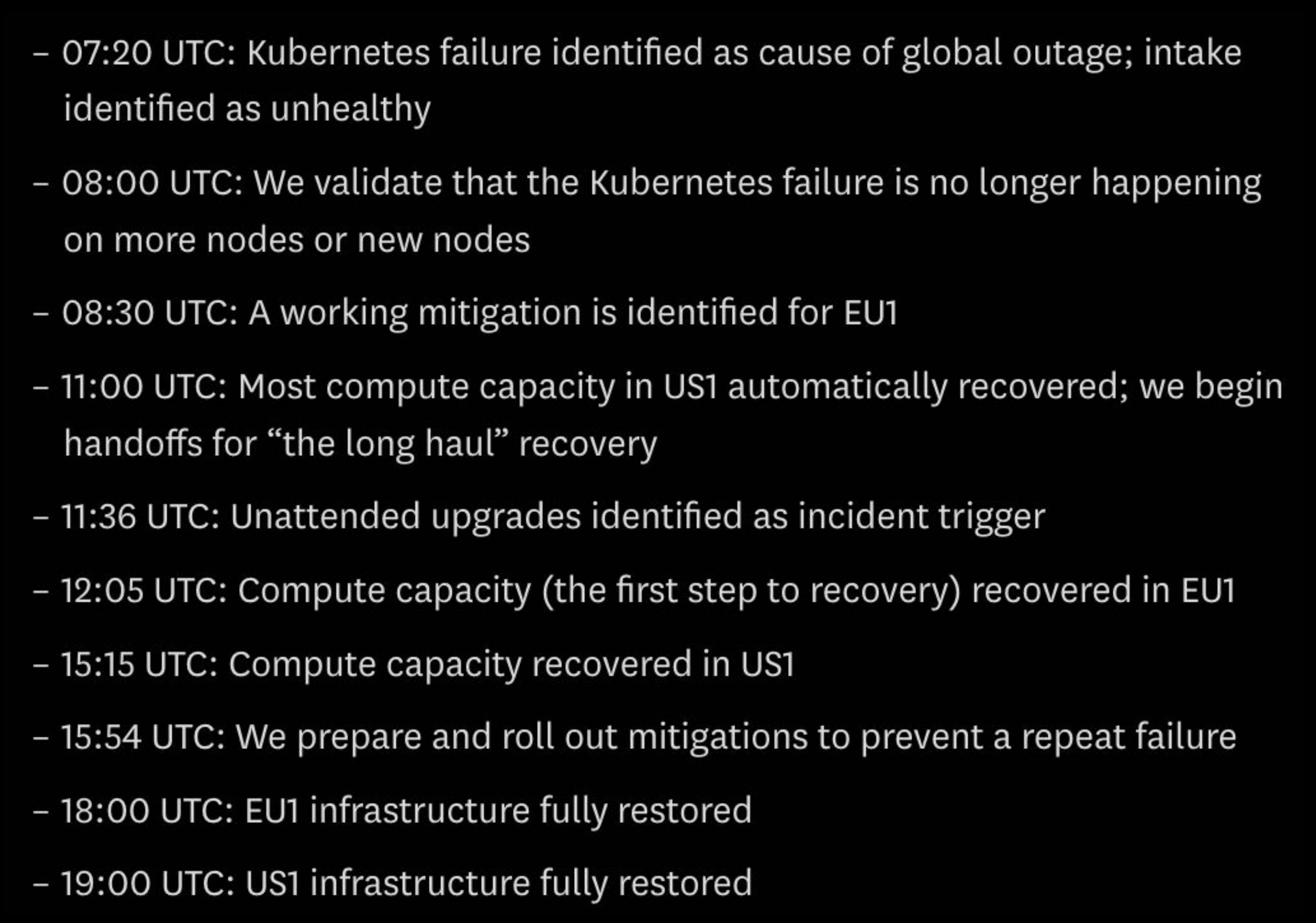
EXECUTIVE LEAD

CUSTOMER LIAISON

RESPONDERS

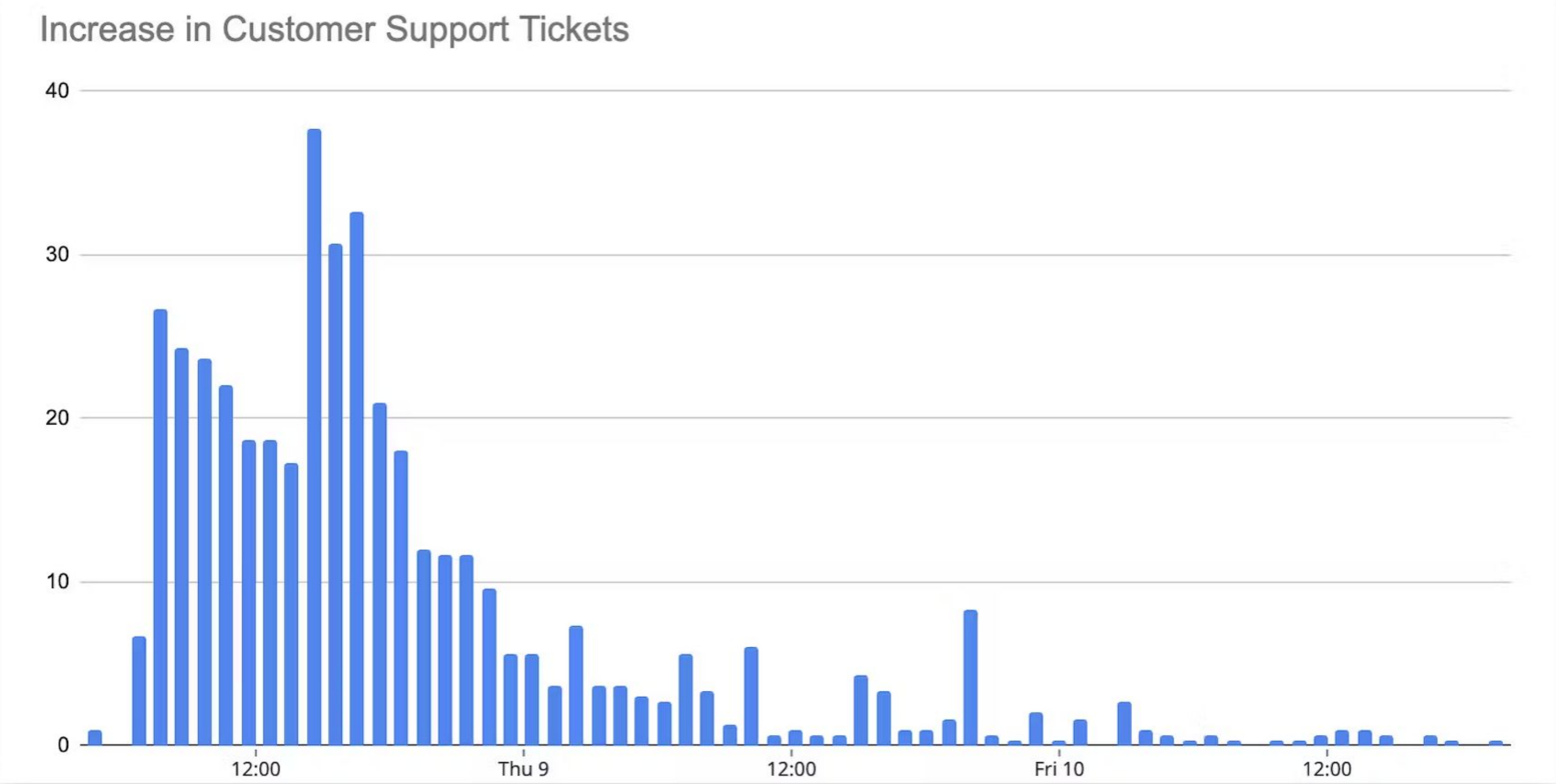
RESPONDERS

RESPONDERS



- 07:20 UTC: Kubernetes failure identified as cause of global outage; intake identified as unhealthy
- 08:00 UTC: We validate that the Kubernetes failure is no longer happening on more nodes or new nodes
- 08:30 UTC: A working mitigation is identified for EU1
- 11:00 UTC: Most compute capacity in US1 automatically recovered; we begin handoffs for "the long haul" recovery
- 11:36 UTC: Unattended upgrades identified as incident trigger
- 12:05 UTC: Compute capacity (the first step to recovery) recovered in EU1
- 15:15 UTC: Compute capacity recovered in US1
- 15:54 UTC: We prepare and roll out mitigations to prevent a repeat failure
- 18:00 UTC: EU1 infrastructure fully restored
- 19:00 UTC: US1 infrastructure fully restored

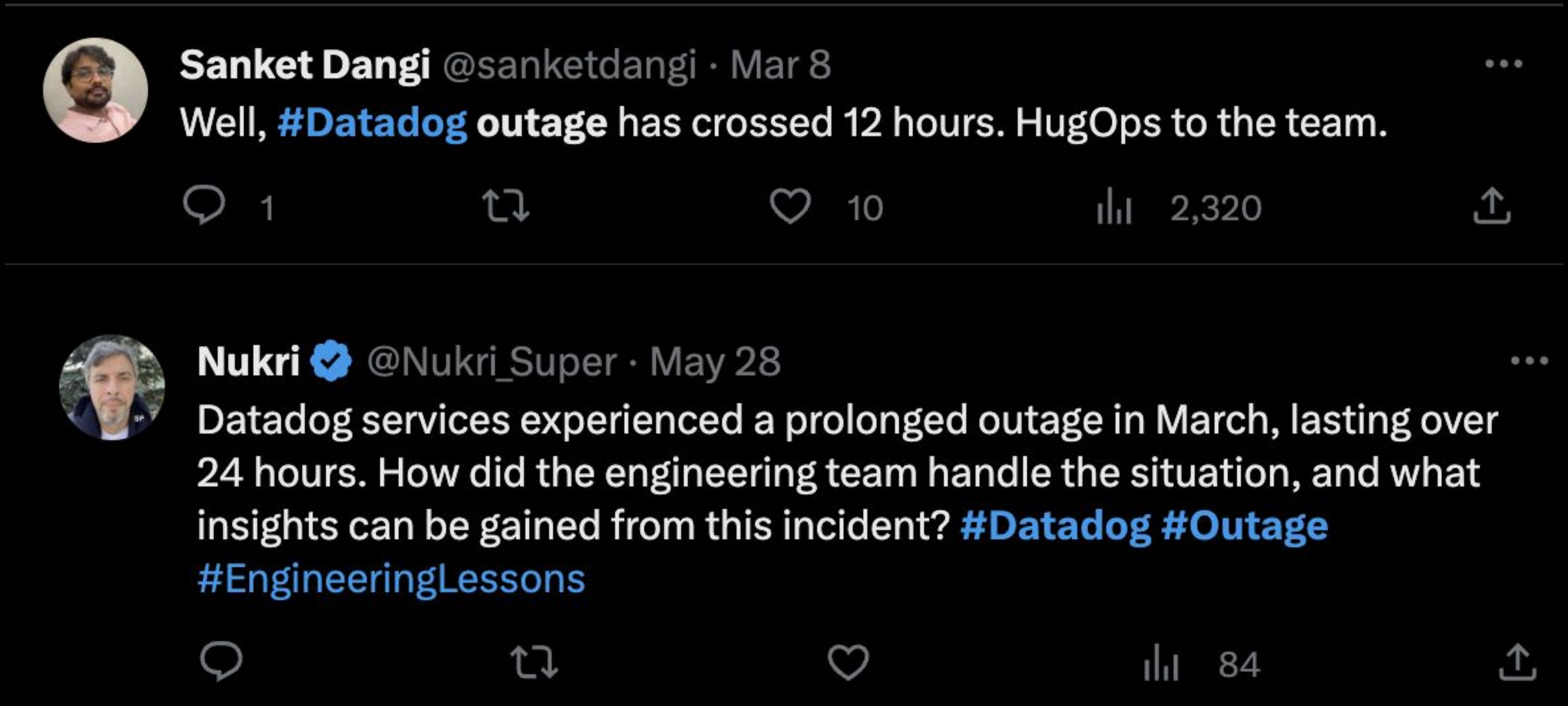
Interesting Challenges



We received about 25 times as many tickets as usual during the first 12 hours of the March 8 incident. The number of tickets received during a typical hour is shown as 1x.

Lessons Learned

Lessons Learned

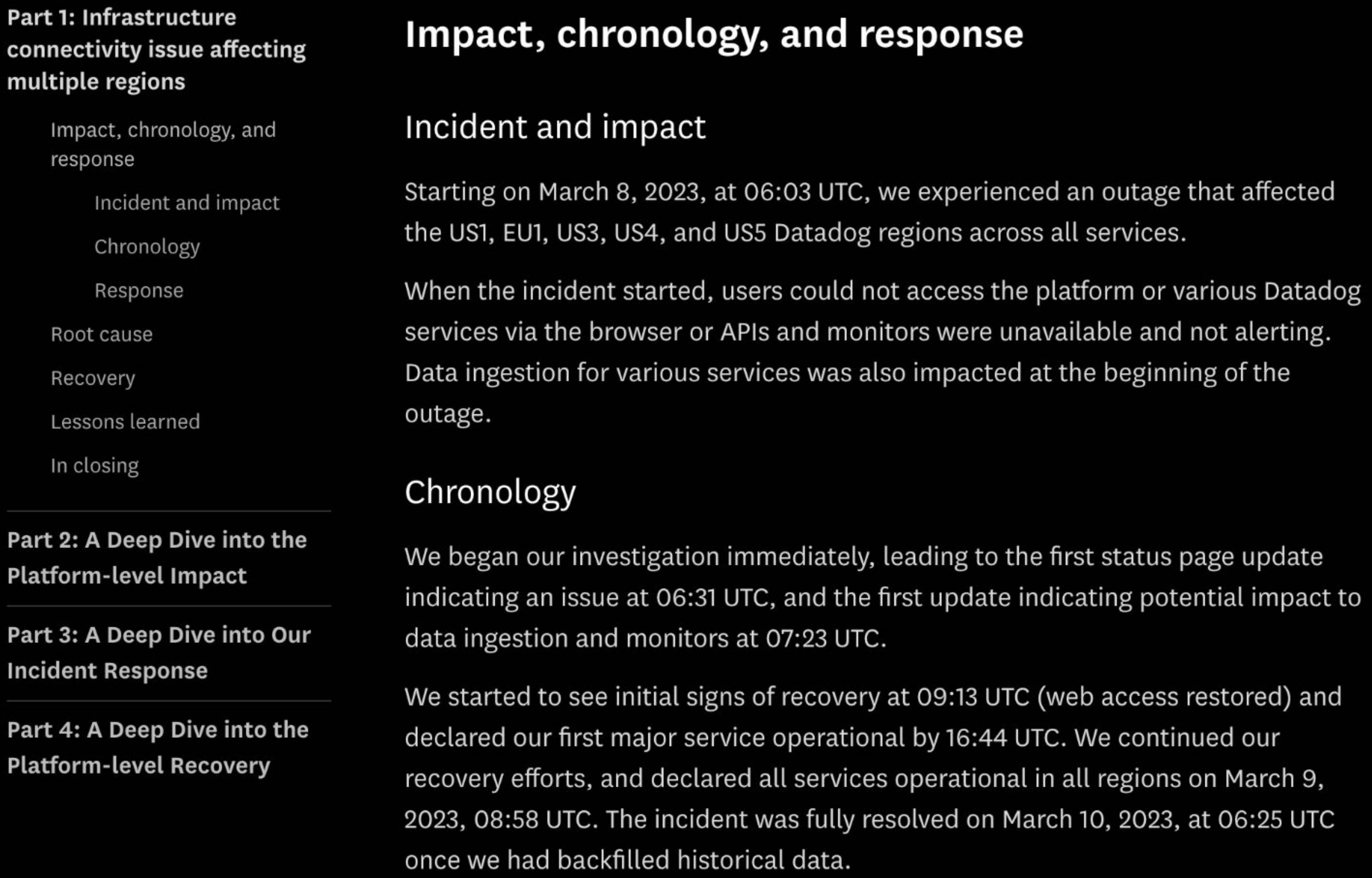


Sanket Dangi @sanketdangi · Mar 8

Well, #Datadog **outage** has crossed 12 hours. HugOps to the team.

Nukri @Nukri_Super · May 28

Datadog services experienced a prolonged outage in March, lasting over 24 hours. How did the engineering team handle the situation, and what insights can be gained from this incident? #Datadog #Outage #EngineeringLessons



Impact, chronology, and response

Incident and impact

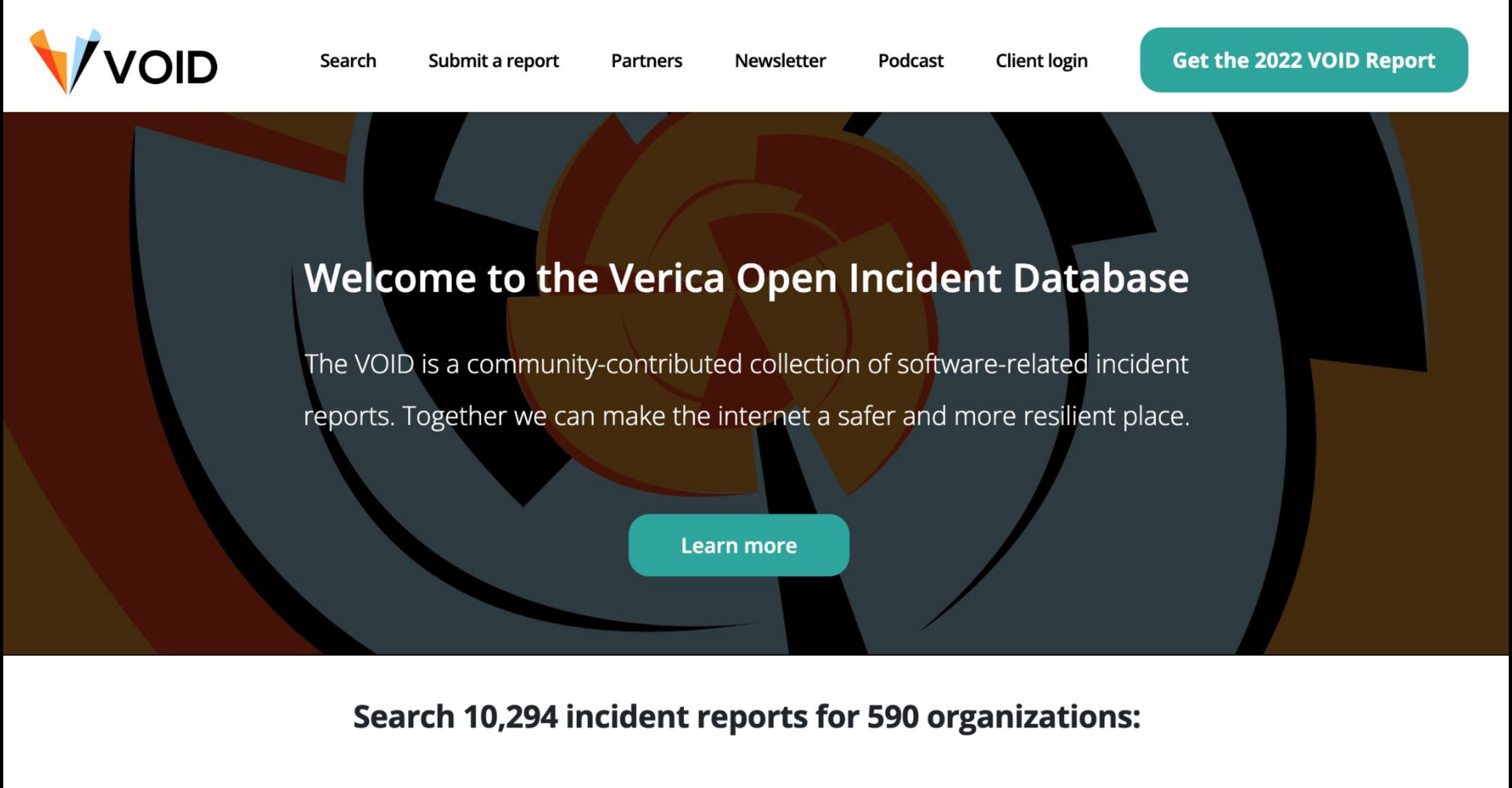
Starting on March 8, 2023, at 06:03 UTC, we experienced an outage that affected the US1, EU1, US3, US4, and US5 Datadog regions across all services.

When the incident started, users could not access the platform or various Datadog services via the browser or APIs and monitors were unavailable and not alerting. Data ingestion for various services was also impacted at the beginning of the outage.

Chronology

We began our investigation immediately, leading to the first status page update indicating an issue at 06:31 UTC, and the first update indicating potential impact to data ingestion and monitors at 07:23 UTC.

We started to see initial signs of recovery at 09:13 UTC (web access restored) and declared our first major service operational by 16:44 UTC. We continued our recovery efforts, and declared all services operational in all regions on March 9, 2023, 08:58 UTC. The incident was fully resolved on March 10, 2023, at 06:25 UTC once we had backfilled historical data.



Welcome to the Verica Open Incident Database

The VOID is a community-contributed collection of software-related incident reports. Together we can make the internet a safer and more resilient place.

Search 10,294 incident reports for 590 organizations:

Ajuna Kyaruzi

SRE & DevOps Advocate
ajuna@datadog.com

DATADOG

ajunaky

“THE BIG ONE”

Learning from incidents

What took so long to get a postmortem out?