Observing *from* Incidents

Cory Watson
SREcon20 Americas

Surprises, incidents, and mistakes are a fact of life.

Cory Watson
Reliability at Stripe

Past: **SRE** @ Twitter, **Observability** @ Twitter & Stripe, **Director** @ SignalFx, **Advocacy and Product** @ Splunk

Things We **Do** From Incidents

Add an alert for that

- Page someone
- Email someone
- Slack someone

Add a chart for that

Add a dashboard for that

Add a metric for that

Add that to the runbook

Filter those shapes out

Catch this in code review

Add tests for that

Test this in (QA, Staging, Pre-Prod, Blue/Green)

Fix that one thing nobody wants to touch

Org wide “someone needs to fix QA”

Rewrite it all

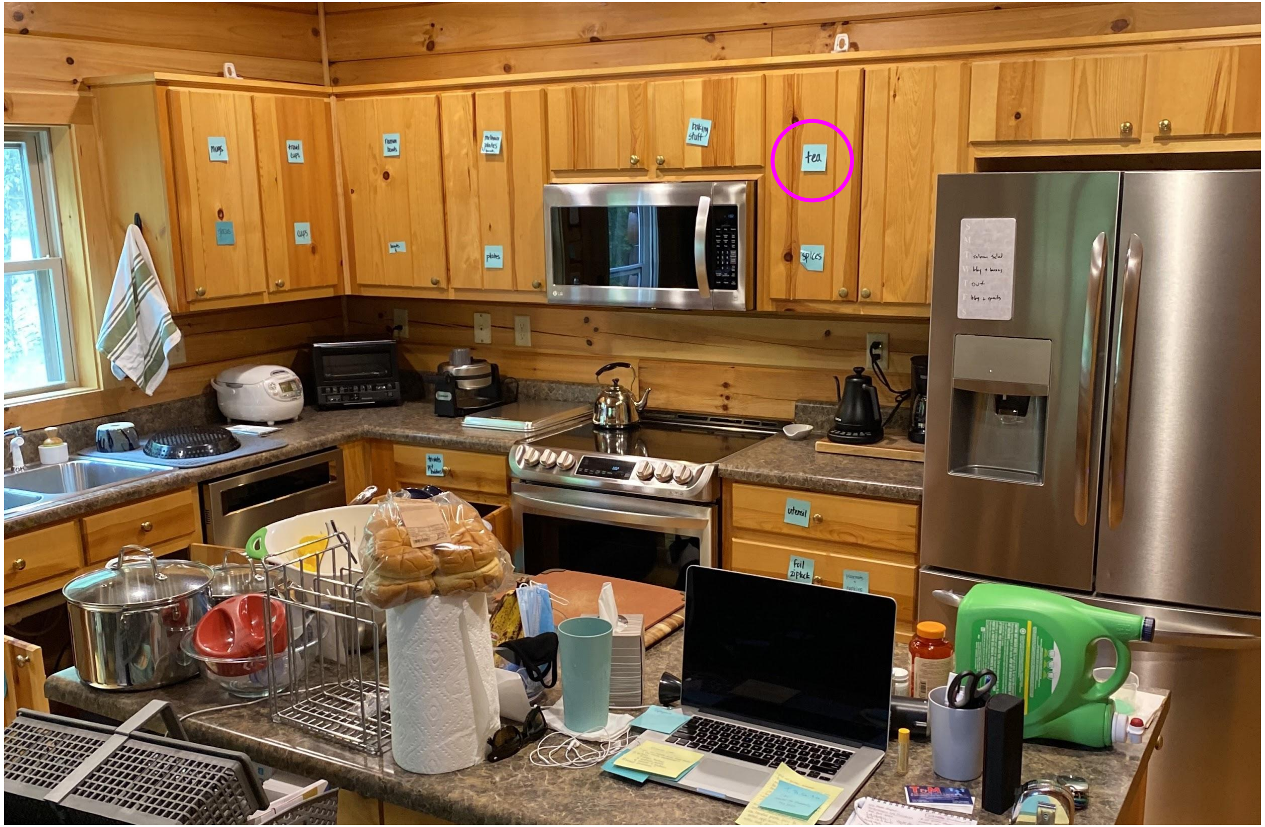
Learning from Incidents

Photo by Ivo Rainha from Pexels

Information Overload

“Perhaps a better way is to make memory unnecessary: **put the required information into the world**”

Don Norman, *The Design of Everyday Things*



Example: Labels

How We Interact With Our Systems

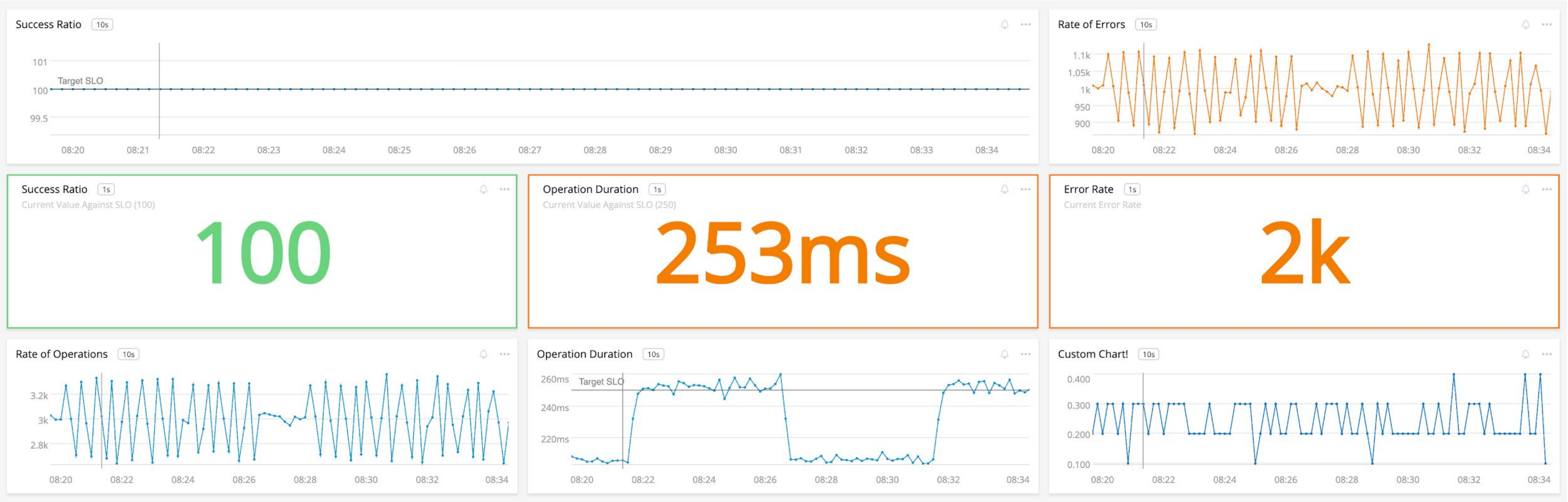
Observing *from* Incidents

Putting what we learn
into the world.

Goal: Add What We Learn To The Tools

- Dashboards, mostly
- Minimizing the need for memory / recall
- Quick, “good enough”
- Awareness of boundaries
- Feedback

Dashboards



Models of Our Systems

Problems we’re looking for

Surprise

Confusion

Dead ends

Frustration

Edges

Misunderstanding

Boundaries

Inefficiency

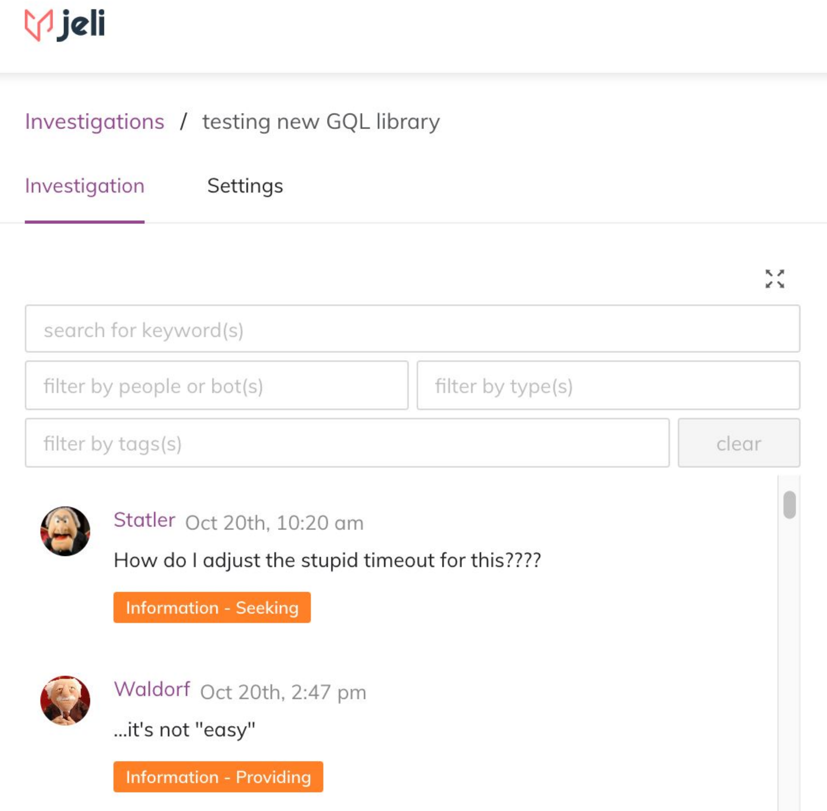
Timeliness

Staleness

Lack of confidence

General gripes

Incidents can inform how we need to improve our tools



Investigations / testing new GQL library

Statler Oct 20th, 10:20 am

How do I adjust the stupid timeout for this????

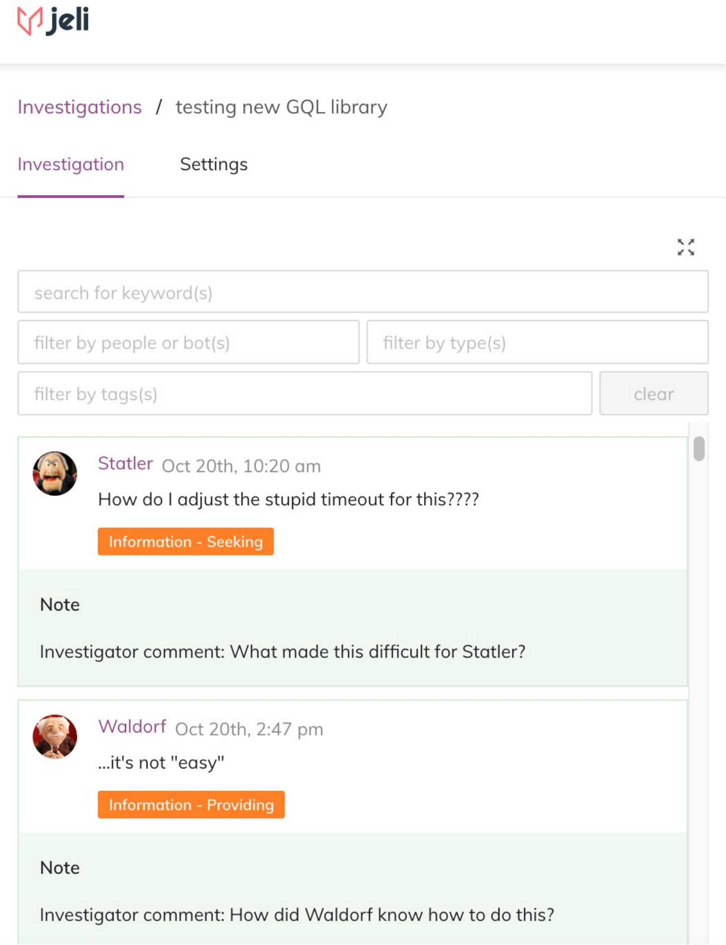
Information - Seeking

Waldorf Oct 20th, 2:47 pm

...it's not "easy"

Information - Providing

When someone asks a question in an incident, take note of it



Statler Oct 20th, 10:20 am

How do I adjust the stupid timeout for this????

Information - Seeking

Note

Investigator comment: What made this difficult for Statler?

Waldorf Oct 20th, 2:47 pm

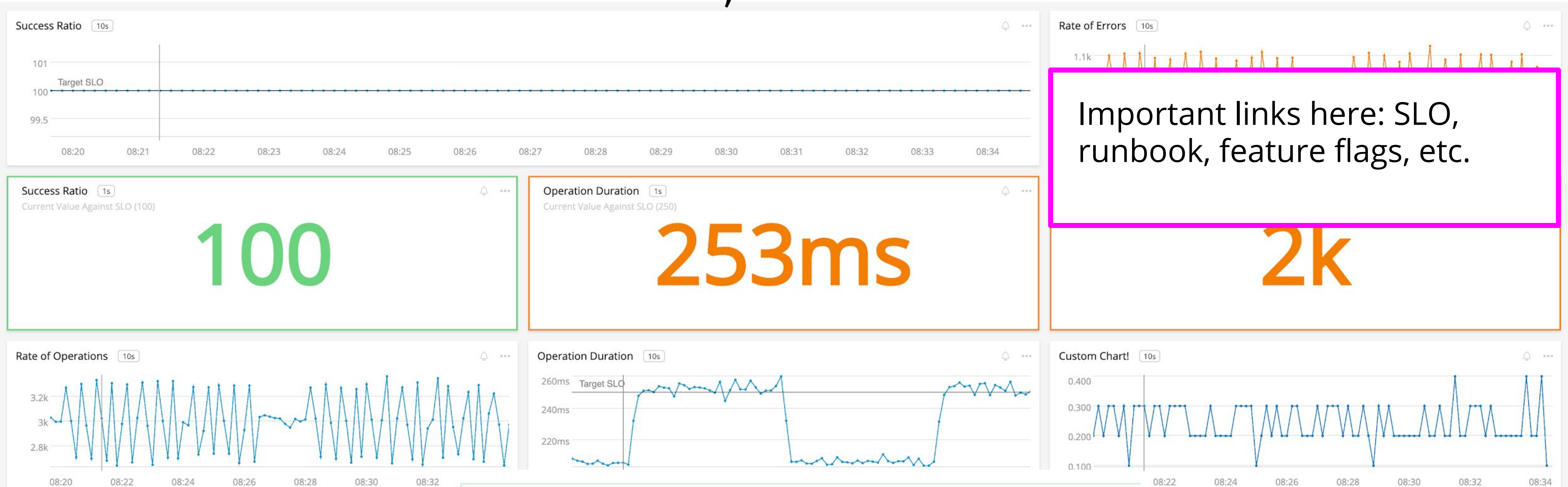
...it's not "easy"

Information - Providing

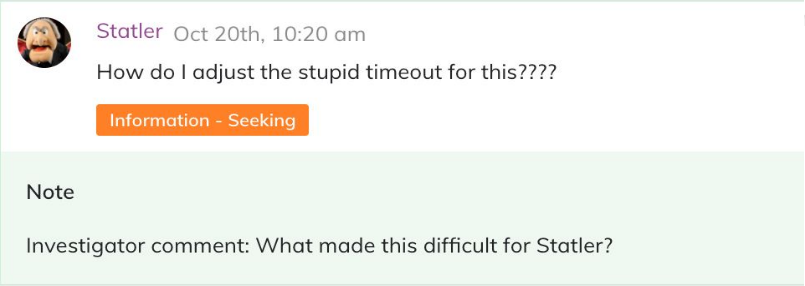
Note

Investigator comment: How did Waldorf know how to do this?

Link, Describe



Important links here: SLO, runbook, feature flags, etc.



Statler Oct 20th, 10:20 am

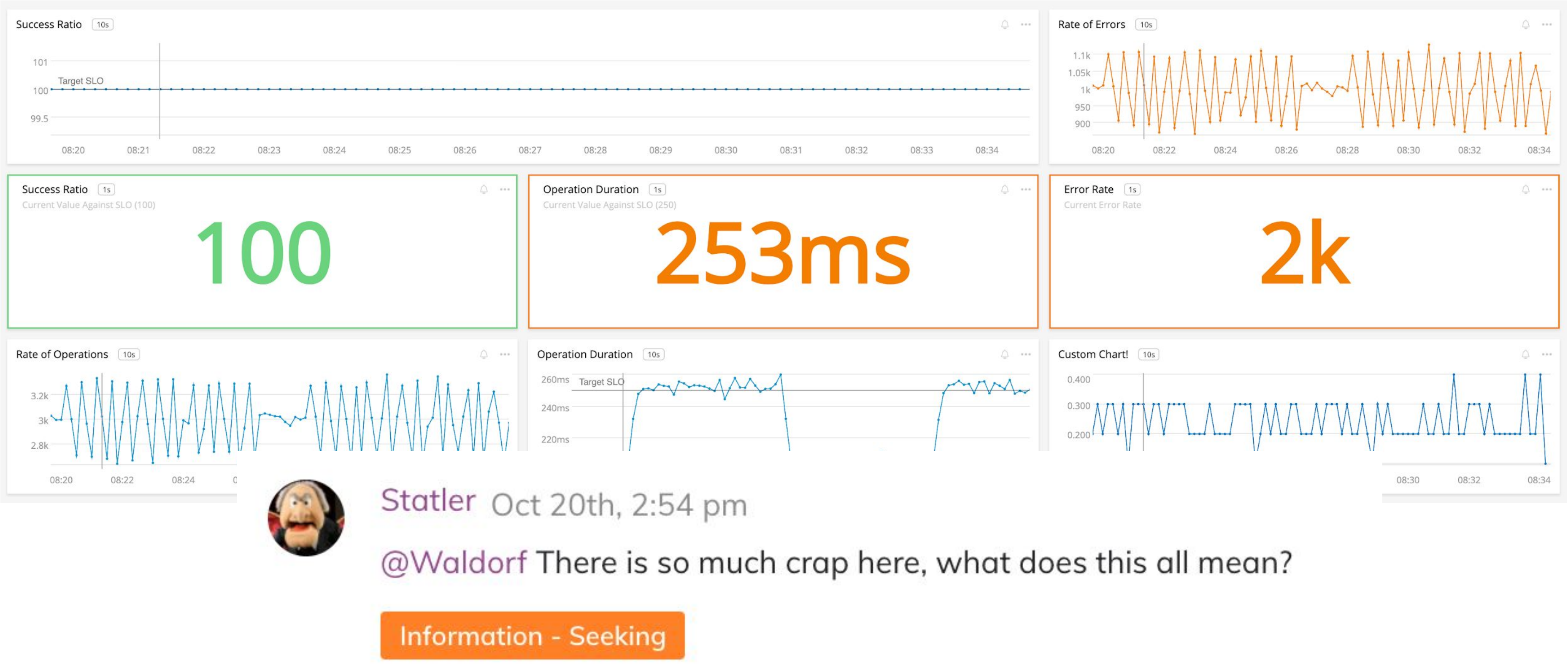
How do I adjust the stupid timeout for this????

Information - Seeking

Note

Investigator comment: What made this difficult for Statler?

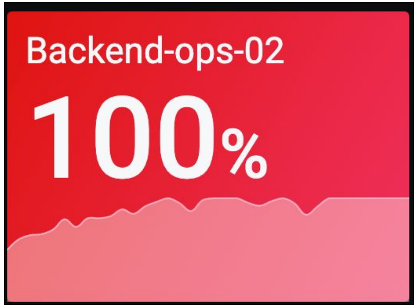
Declutter



Statler Oct 20th, 2:54 pm

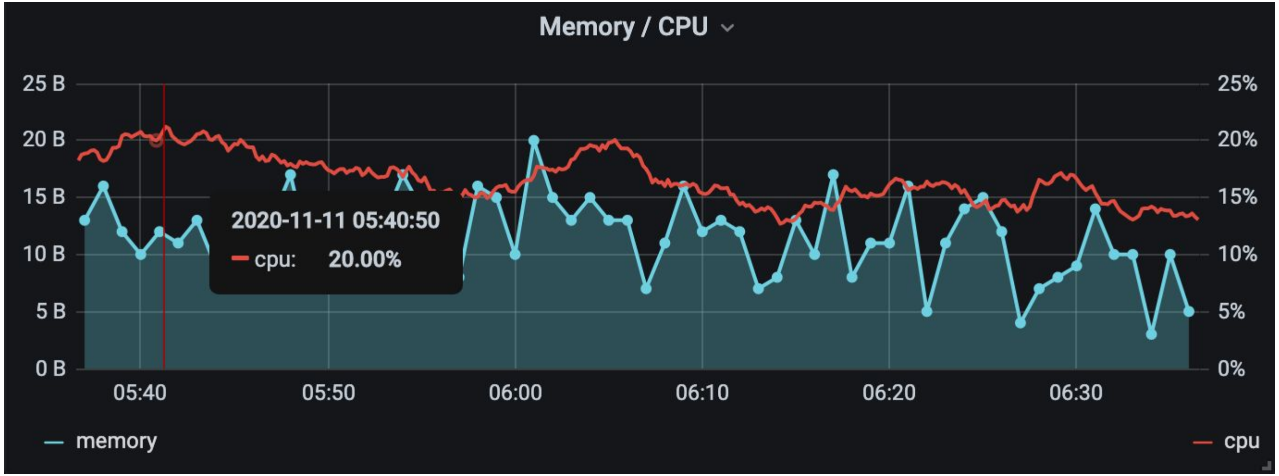
@Waldorf There is so much crap here, what does this all mean?

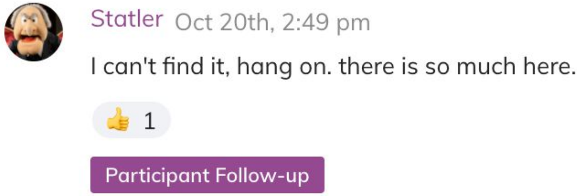
Combine



Backend-ops-02

100%





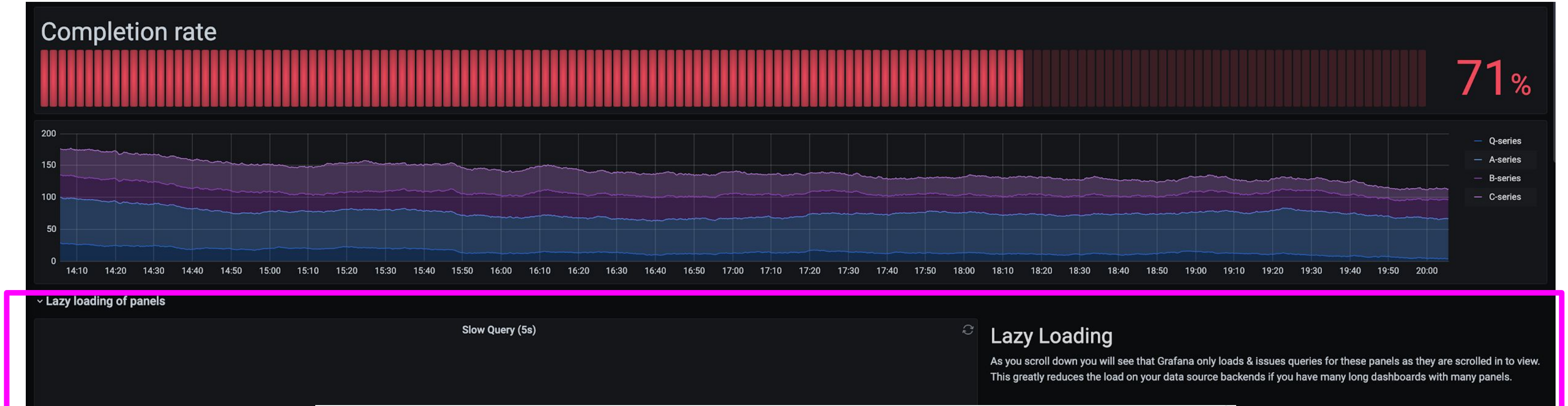
Statler Oct 20th, 2:49 pm

I can't find it, hang on. there is so much here.

👍 1

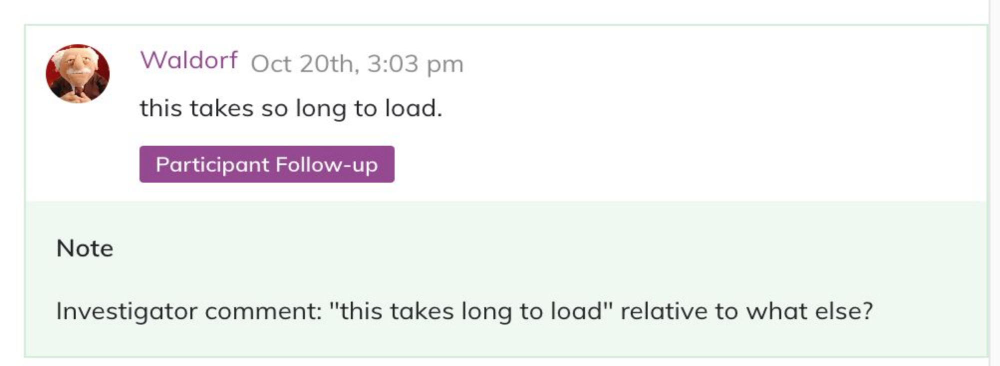
Participant Follow-up

Divide



Lazy Loading

As you scroll down you will see that Grafana only loads & issues queries for these panels as they are scrolled in to view. This greatly reduces the load on your data source backends if you have many long dashboards with many panels.



Waldorf Oct 20th, 3:03 pm

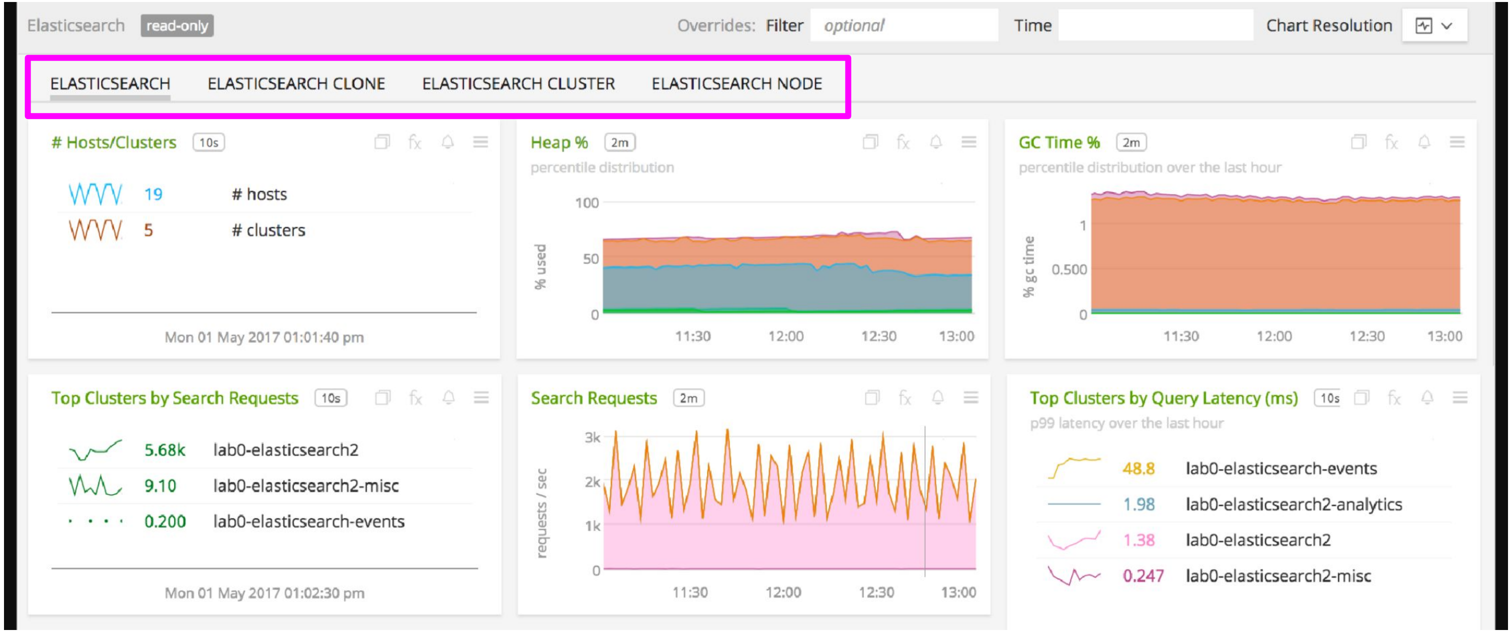
this takes so long to load.

Participant Follow-up

Note

Investigator comment: "this takes long to load" relative to what else?

Divide (cont)



# Hosts/Clusters 10s

19	# hosts
5	# clusters

Mon 01 May 2017 01:01:40 pm

Heap % 2m

percentile distribution

GC Time % 2m

percentile distribution over the last hour

Top Clusters by Search Requests 10s

5.68k	lab0-elasticsearch2
9.10	lab0-elasticsearch2-misc
0.200	lab0-elasticsearch-events

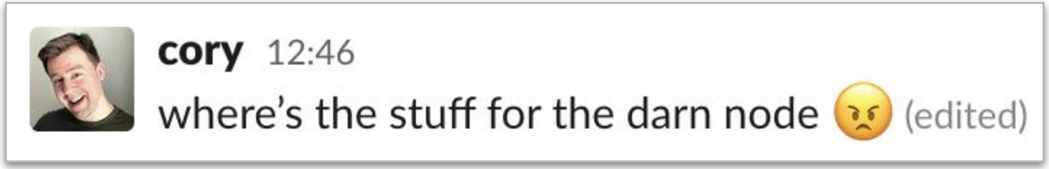
Mon 01 May 2017 01:02:30 pm

Search Requests 2m

Top Clusters by Query Latency (ms) 10s

p99 latency over the last hour

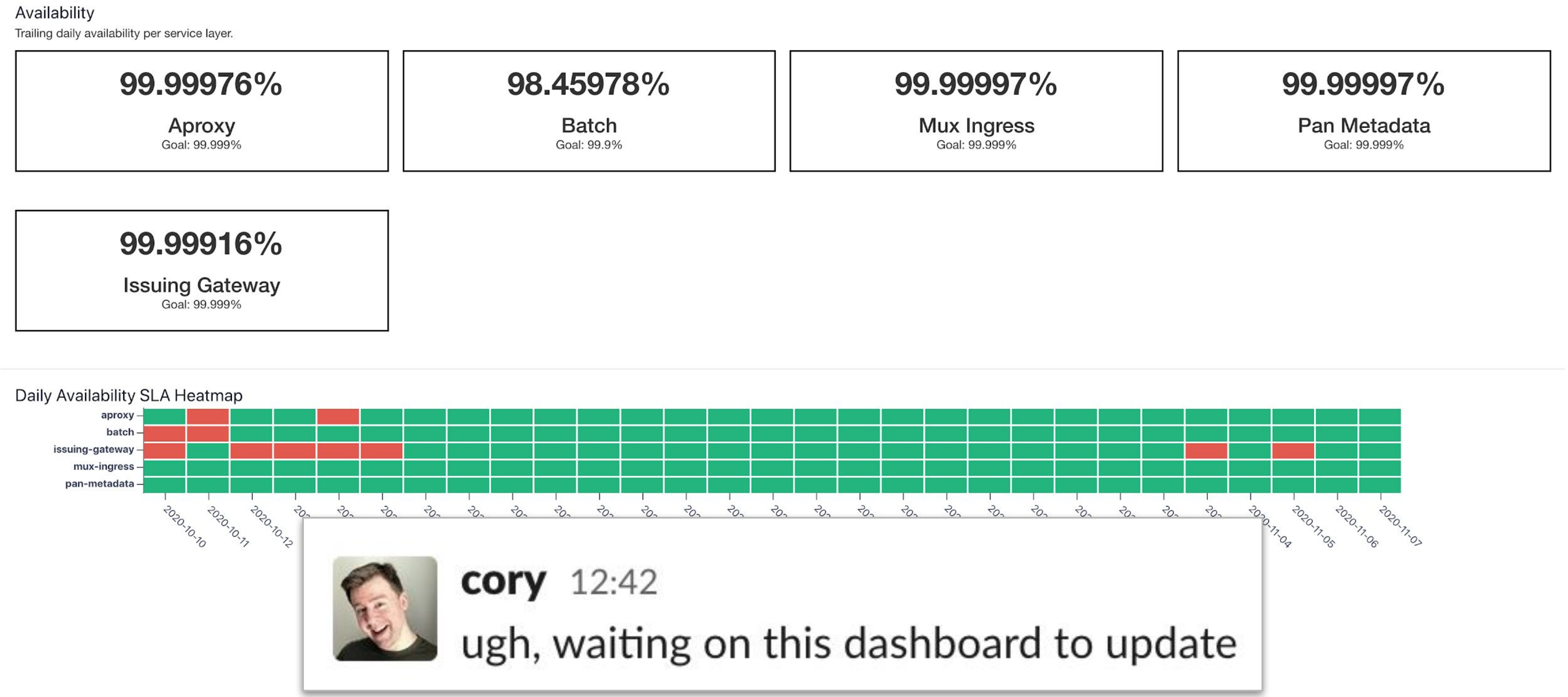
48.8	lab0-elasticsearch-events
1.98	lab0-elasticsearch2-analytics
1.38	lab0-elasticsearch2
0.247	lab0-elasticsearch2-misc



cory 12:46

where’s the stuff for the darn node 😠 (edited)

Timeliness & Rate of Change



Availability

Trailing daily availability per service layer.

99.99976%
Aproxy
Goal: 99.999%

98.45978%
Batch
Goal: 99.9%

99.99997%
Mux Ingress
Goal: 99.999%

99.99997%
Pan Metadata
Goal: 99.999%

99.99916%
Issuing Gateway
Goal: 99.999%

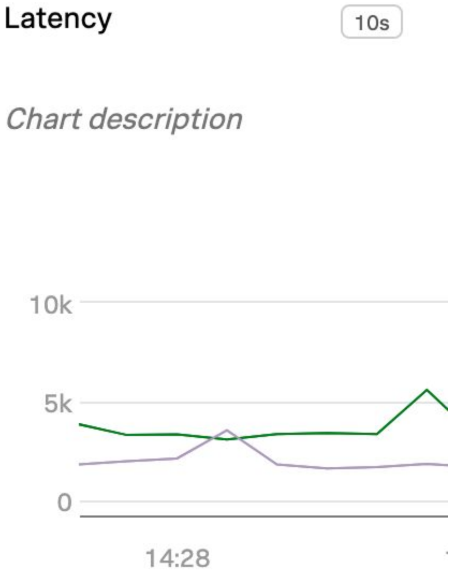
Daily Availability SLA Heatmap

	2020-10-10	2020-10-11	2020-10-12	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	2020-11-04	2020-11-05	2020-11-06	2020-11-07
aproxy	🟩	🟥	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
batch	🟥	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
issuing-gateway	🟥	🟩	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟥	🟩	🟩
mux-ingress	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
pan-metadata	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩

cory 12:42

ugh, waiting on this dashboard to update

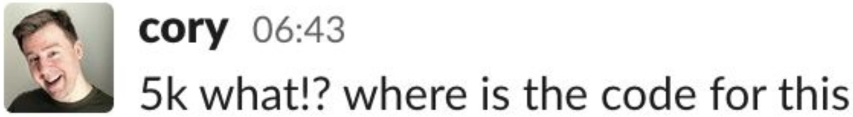
Units



Latency

10s

Chart description



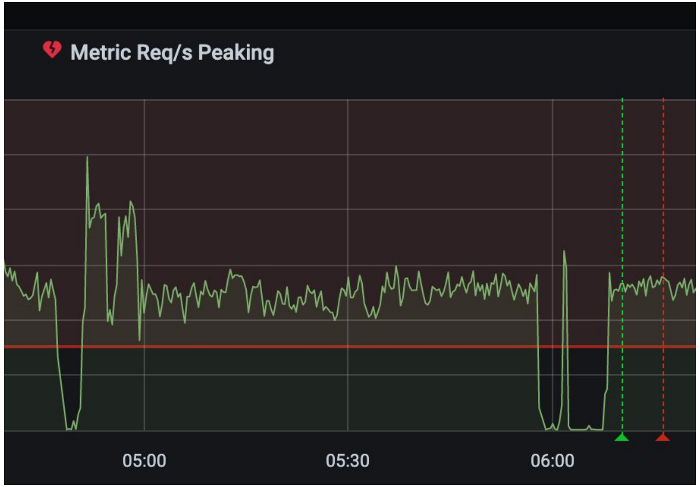
cory 06:43

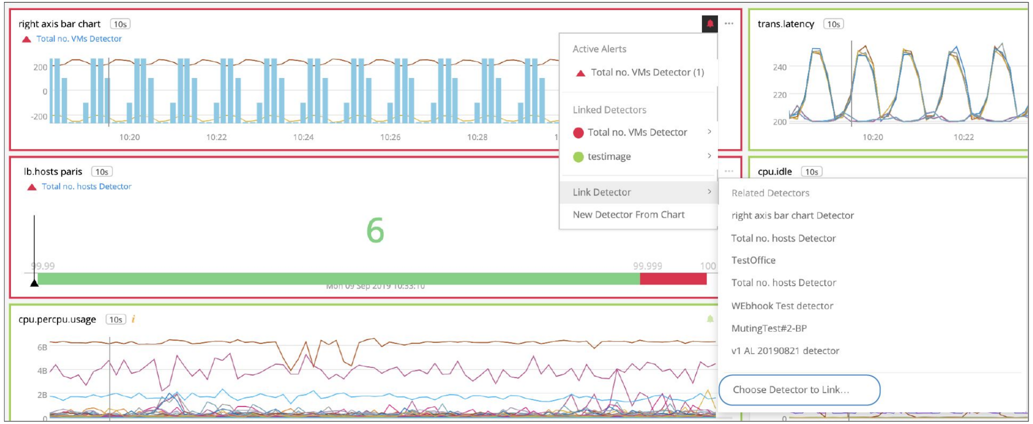
5k what!? where is the code for this

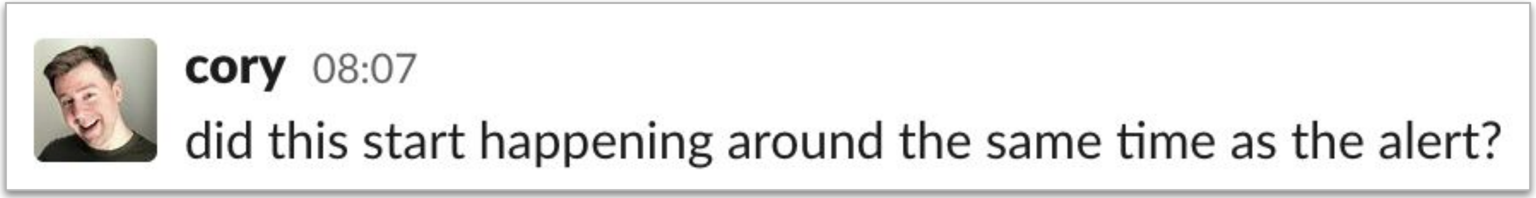
"…the focus should be on control of behavior by **making the boundaries explicit and known**…"

Rasmussen, *Risk management in a dynamic society*

Colocate Signals



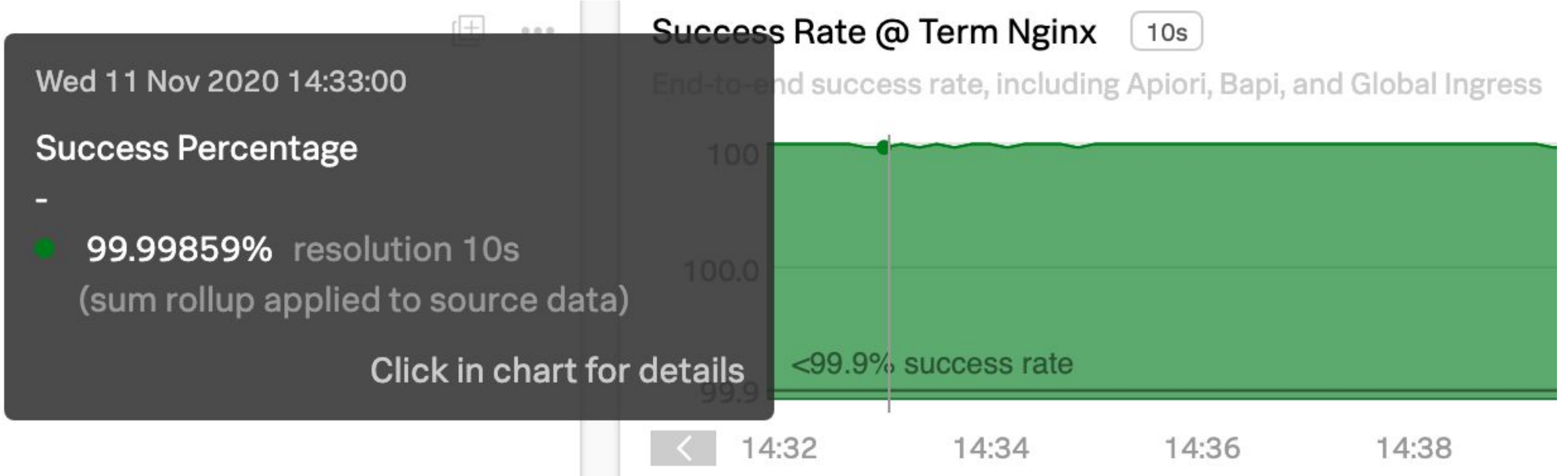




cory 08:07

did this start happening around the same time as the alert?

Communicate Goals



Success Rate @ Term Nginx 10s

End-to-end success rate, including Apiori, Bapi, and Global Ingress

Wed 11 Nov 2020 14:33:00

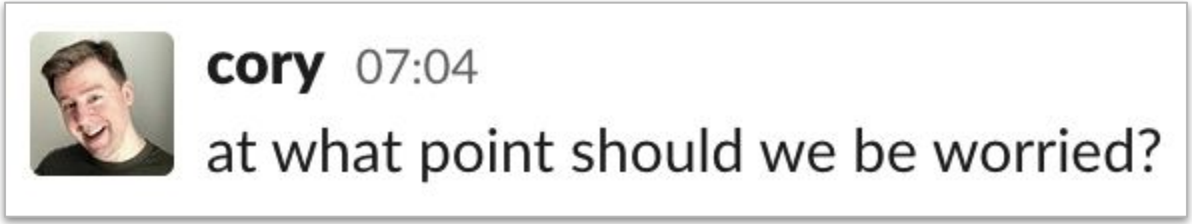
Success Percentage

-

● 99.99859% resolution 10s

(sum rollup applied to source data)

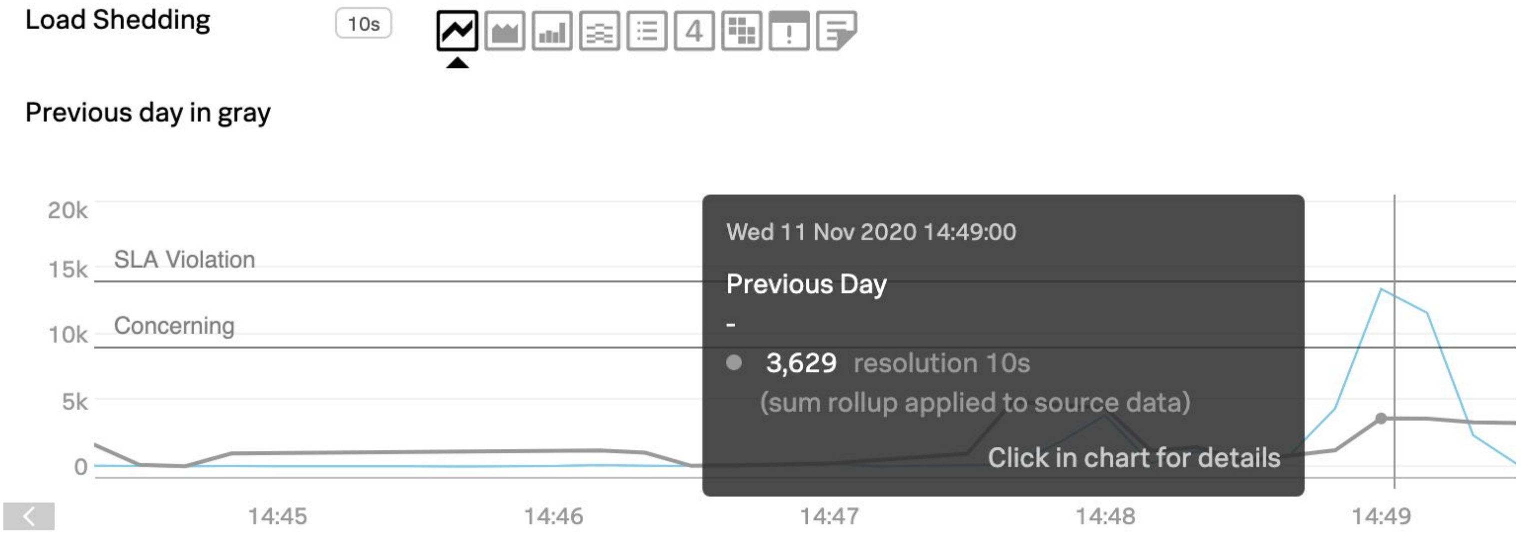
Click in chart for details



cory 07:04

at what point should we be worried?

Give Context



Load Shedding

10s

Previous day in gray

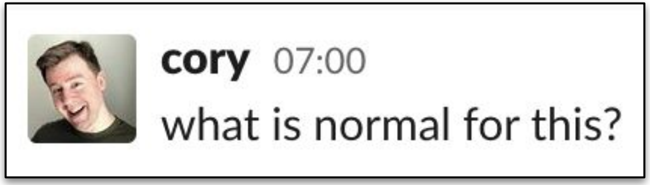
Wed 11 Nov 2020 14:49:00

Previous Day

-

● 3,629 resolution 10s
(sum rollup applied to source data)

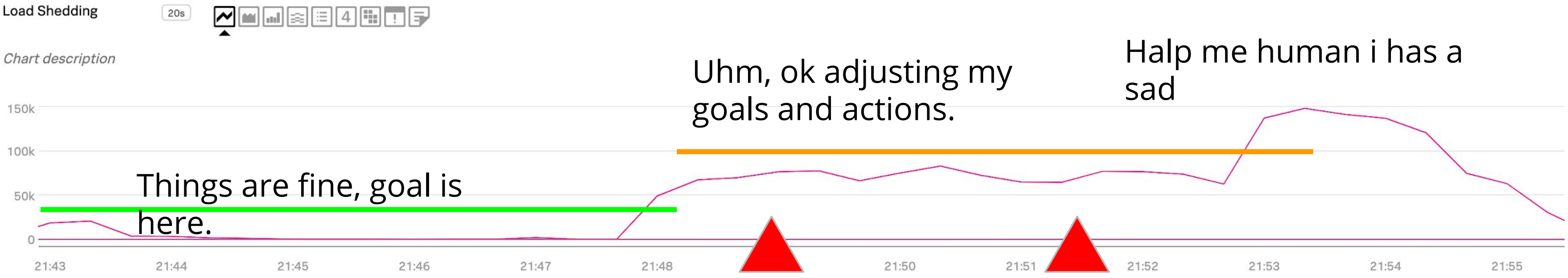
Click in chart for details



cory 07:00

what is normal for this?

Work With Automation



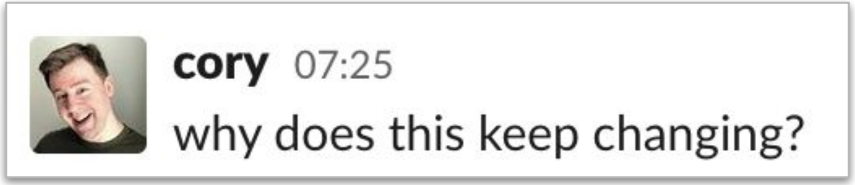
Things are fine, goal is here.

Uhm, ok adjusting my goals and actions.

Halp me human i has a sad

Pew pew mitigating

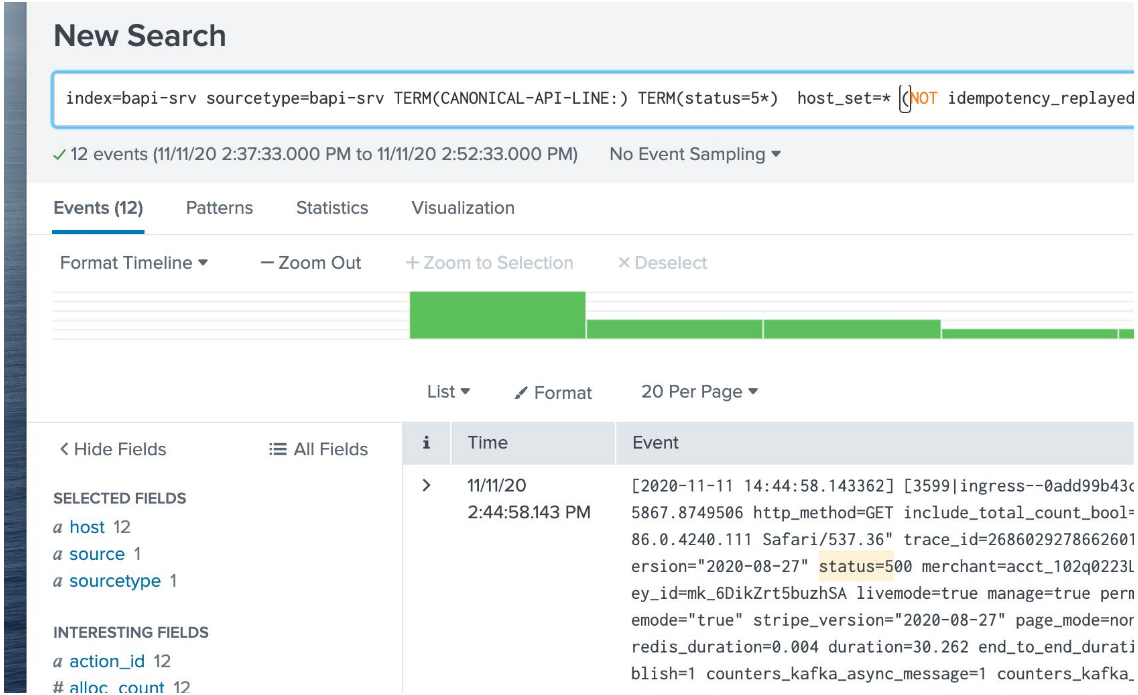
Im givin' her all shes got captain



cory 07:25

why does this keep changing?

Go Analytic



```
index=bapi-srv sourcetype=bapi-srv TERM(CANONICAL-API-LINE:) TERM(status=5*)  host_set=* (NOT idempotency_replayed
```

✓ 12 events (11/11/20 2:37:33.000 PM to 11/11/20 2:52:33.000 PM)   No Event Sampling ▾

SELECTED FIELDS

a host 12
a source 1
a sourcetype 1

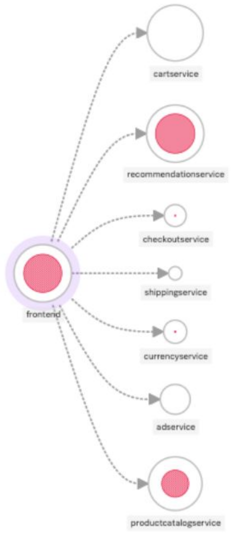
INTERESTING FIELDS

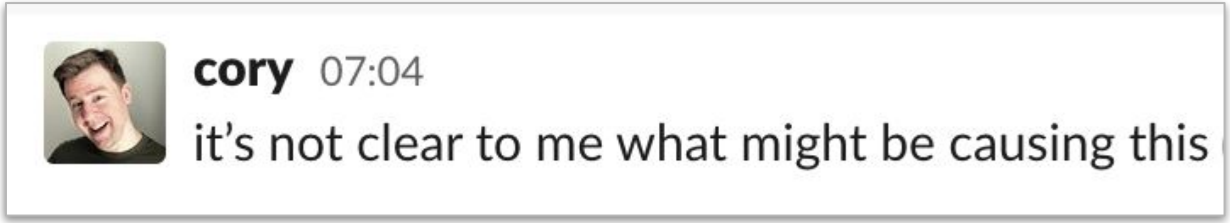
a action_id 12
alloc_count 12

› 11/11/20 2:44:58.143 PM

```
[2020-11-11 14:44:58.143362] [3599|ingress--0add99b43c
5867.8749506 http_method=GET include_total_count_bool=
86.0.4240.111 Safari/537.36" trace_id=268602927866260
ersion="2020-08-27" status=500 merchant=acct_102q0223L
ey_id=mk_6DikZrt5buzhSA livemode=true manage=true perm
emode="true" stripe_version="2020-08-27" page_mode=nor
redis_duration=0.004 duration=30.262 end_to_end_durati
blish=1 counters_kafka_async_message=1 counters_kafka_
```

Stay Perceptive / Mental Models

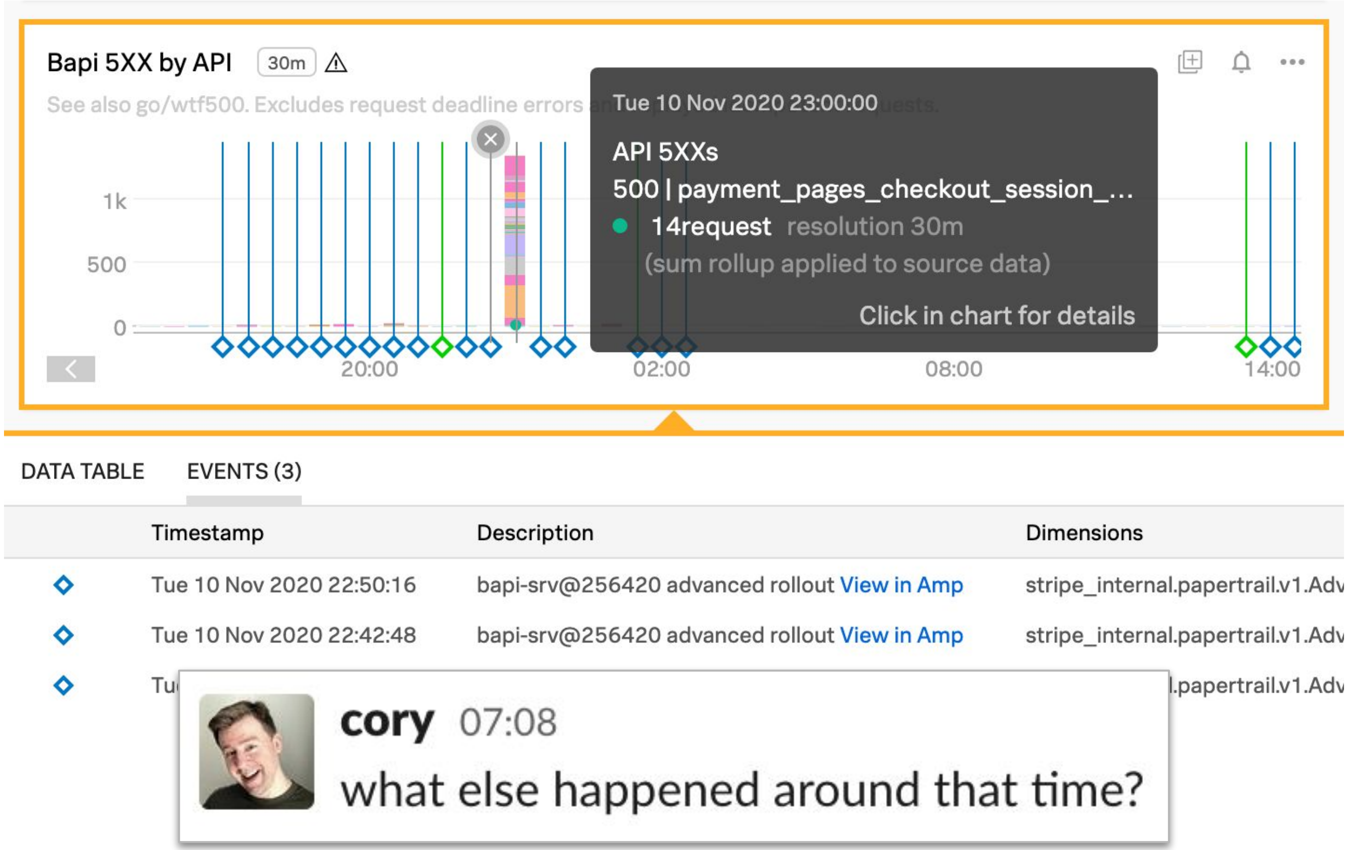




cory 07:04

it's not clear to me what might be causing this

Spot Contributing Factors



Tue 10 Nov 2020 23:00:00

API 5XXs

500 | payment_pages_checkout_session_…

● 14request resolution 30m

(sum rollup applied to source data)

Click in chart for details

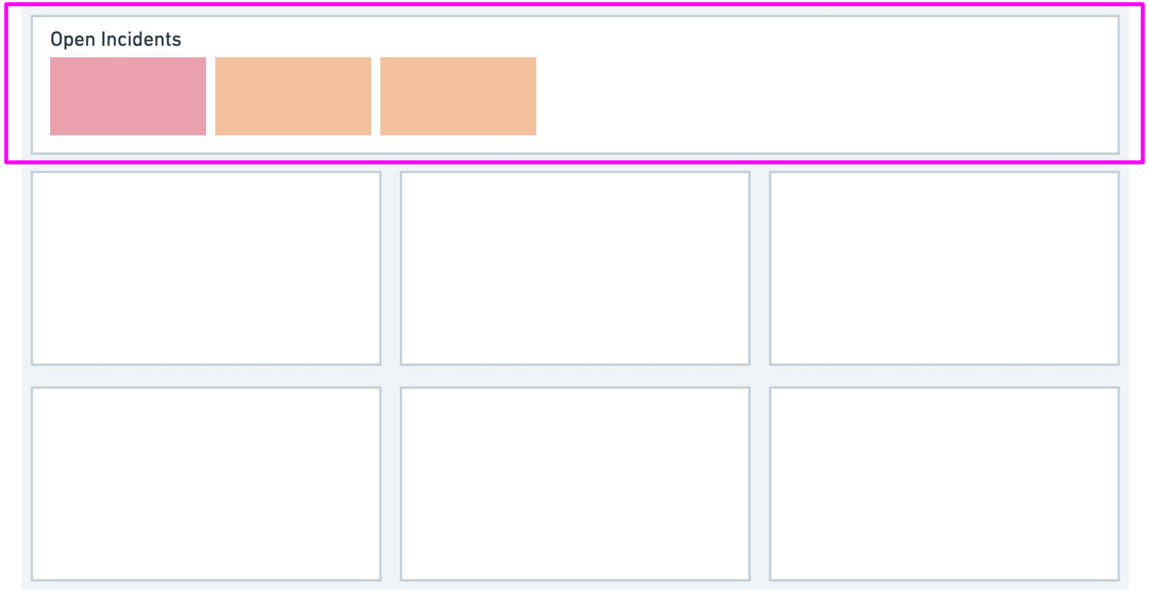
DATA TABLE EVENTS (3)

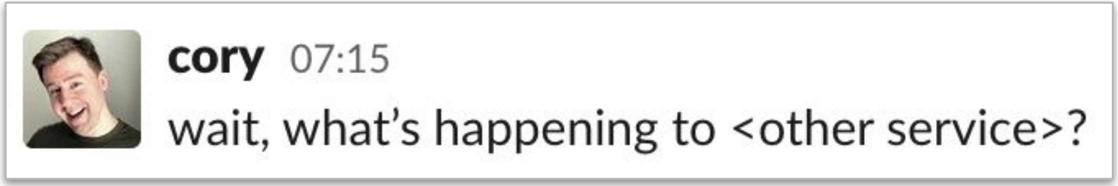
	Timestamp	Description	Dimensions
◇	Tue 10 Nov 2020 22:50:16	bapi-srv@256420 advanced rollout View in Amp	stripe_internal.papertrail.v1.Adv
◇	Tue 10 Nov 2020 22:42:48	bapi-srv@256420 advanced rollout View in Amp	stripe_internal.papertrail.v1.Adv
◇	Tu		l.papertrail.v1.Adv

cory 07:08

what else happened around that time?

Help with Situation Awareness

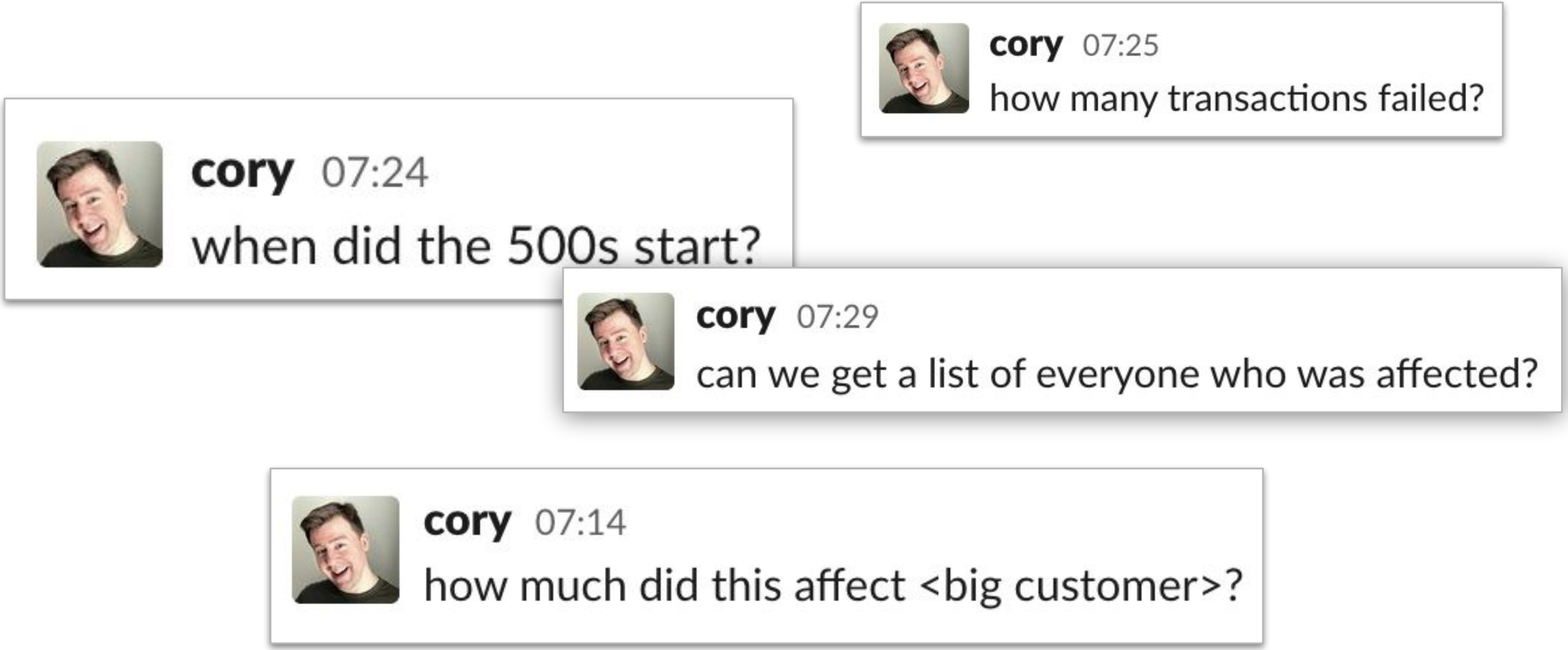




cory 07:15

wait, what’s happening to <other service>?

Encode Common Workflows



cory 07:25
how many transactions failed?

cory 07:24
when did the 500s start?

cory 07:29
can we get a list of everyone who was affected?

cory 07:14
how much did this affect <big customer>?

Summary

Learn

From Your

~~Incidents~~ Opportunities

No neutral in design

Easy to fall into

Organic growth

Examine usage, collect signals

Sit with users

Share your work

Not (just) your vendor / tool problem

What's your model?

Collecting more feedback

Use what you have

Spot the signs

- Dashboards as a tool: need a goal, get state, way to affect state, model
- Mental models, avoid dead ends, allow side to side, analytic
- Events to aid awareness (stuff happening, alert state, using alerts as an "info" channel)
- Bounds, limits, norms
- Automation actions, decisions, and goals

Would you like to know more?

- *The Design of Everyday Things*, Don Norman
- *Above the Line, Below the Line*, Dr. Richard Cook
- *Risk management in a dynamic society*, Rasmussen
- Ecological Interface Design

Thanks!

Twitter: @gphat
Email: cory@onemogin.com