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
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
When someone asks a question in an incident, take note of it.
Important links here: SLO, runbook, feature flags, etc.

100

253ms

2k

**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

Success Ratio: 100%
Operation Duration: 253ms
Error Rate: 2k

Statler  Oct 20th, 2:54 pm
@Waldorf  There is so much crap here, what does this all mean?

Information - Seeking
I can't find it, hang on. there is so much here.
Divide (cont)

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: 99.99%

99.99997%
Pan Metadata
Goal: 99.99%

99.99916%
Issuing Gateway
Goal: 99.999%

Daily Availability SLA Heatmap

Cory 12:42
ugh, waiting on this dashboard to update
Units

Latency

Chart description

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

Wed 11 Nov 2020 14:33:00

Success Percentage

- 99.99859% resolution 10s
  (sum rollup applied to source data)

Click in chart for details

<99.9% success rate

Cory 07:04
at what point should we be worried?
Give Context

Load Shedding

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?
Things are fine, goal is here.

Uhm, ok adjusting my goals and actions.

Pew pew mitigating

Im givin’ her all shes got captain

Halp me human i has a sad

why does this keep changing?
## New Search

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

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

### Events (12)

<table>
<thead>
<tr>
<th>i</th>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;</td>
<td>11/11/20 2:44:58.143 PM</td>
<td>[2020-11-11 14:44:58.143362] [3599]ingress--0add99b43+x586.8749506 http_method=GET include_total_count_bool:86.0.4208.111 Safari/537.36 trace_id=268682927866260 version=&quot;2020-08-27&quot; status=500 merchant=acct_102q0223 ey_id=mk_601k2zt5buzhSA livemode=true manage=true persist=true stripe_version=&quot;2020-08-27&quot; page_mode=non redis_duration=0.004 duration=30.262 end_to_end_duration=68.316 birsch=1 counters_kafka_async_message=1 counters_kafka_</td>
</tr>
</tbody>
</table>
Stay Perceptive / Mental Models

cory  07:04
it’s not clear to me what might be causing this
Spot Contributing Factors

DATA TABLE

<table>
<thead>
<tr>
<th>Timestamp</th>
<th>Description</th>
<th>Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tue 10 Nov 2020 22:50:16</td>
<td>bapi-srv@256420 advanced rollout View in Amp</td>
<td>stripe_internal.papertrail.v1.Adv</td>
</tr>
<tr>
<td>Tue 10 Nov 2020 22:42:48</td>
<td>bapi-srv@256420 advanced rollout View in Amp</td>
<td>stripe_internal.papertrail.v1.Adv</td>
</tr>
<tr>
<td>Tue 10 Nov 2020 22:40:48</td>
<td>bapi-srv@256420 advanced rollout View in Amp</td>
<td>stripe_internal.papertrail.v1.Adv</td>
</tr>
</tbody>
</table>

 Cory 07:08
 what else happened around that time?
Help with Situation Awareness

Open Incidents

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

cory 07:24
when did the 500s start?

cory 07:25
how many transactions failed?

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