How to SRE When Everything is Already on Fire

Alex Hidalgo (@ahidalgosre)
Alex Lee (@ahl91)
Squarespace Site Reliability Engineering
Tuesday, March 5th, 2019
20:00 ET

It is always darkest before the dawn
A PHENOMENAL EVENING

<table>
<thead>
<tr>
<th>Team</th>
<th>1st Half</th>
<th>2nd Half</th>
<th>Final</th>
</tr>
</thead>
<tbody>
<tr>
<td>VCU Rams</td>
<td>28</td>
<td>43</td>
<td>71</td>
</tr>
<tr>
<td>GMU Patriots</td>
<td>21</td>
<td>15</td>
<td>36</td>
</tr>
</tbody>
</table>
The same ignored problem that had been cropping up for weeks had returned.
Who are we?
Alex Hidalgo
@ahidalgosre

Squarespace Observability SRE

Alex Lee
@alee_sre
Service Reliability

85% ➞ 99.9%
None of this is new
NONE OF THIS IS NEW

Alarm Fatigue is studied
- Healthcare
- Mining
- Construction
- NUCLEAR POWER INDUSTRY
“A liar will not be believed, even when they speak the truth.” - Aesop

(Like 2600 years ago)
NONE OF THIS IS NEW (REdux)

Service reliability has been studied, too
- Alert on what actually matters
- Develop SLIs and SLOs
- Increase your Observability
- Improve tooling and automation
- Trust proven paradigms
- Conduct meaningful postmortems
CONTEXT
How we got here
Spring 2015

The introduction of ELK
ELK @ SQUARESPACE

- Open-source log aggregation
- Scale observability platforms with growing Squarespace infrastructure
- Highest-trafficked service at Squarespace
ELK @ SQUARESPACE

- Log Shippers
- KAFKA
- LOGSTASH
- ELASTICSEARCH
- KIBANA

Diagram shows the flow of data from Log Shippers to KAFKA, then to LOGSTASH, and finally to ELASTICSEARCH, with KIBANA providing a dashboard view.
August 2018

An unhealthy stack
1. ALERT ON WHAT MATTERS
   Put your users first
   DEVELOP MEANINGFUL SLOs
   INCREASE YOUR OBSERVABILITY
   IMPROVE YOUR ENVIRONMENT
   TRUST PROVEN PARADIGMS
   CONDUCT MEANINGFUL POSTMORTEMS
OLD ALERTS

- Logstash process
- Logstash-to-Kafka connection
- Logstash-to-Elasticsearch connection
- Logstash-to-Elasticsearch throughput
- Elasticsearch process
- Elasticsearch “cluster block”
Noisy

Only “known-unknowns”

Not user-focused
Alert on what matters. Put your users first.
“My logs are delayed. Is ELK having issues?”
ELK @ SQUARESPACE

Log Shippers → KAFKA → LOGSTASH → ELASTICSEARCH → KIBANA

START

END
NEW ALERT, SINGULAR

\[
\frac{\text{Kafka lag (msg)}}{\text{Logstash rate (msg/s)}} = \text{End-to-end Latency (s)}
\]
“My logs are delayed. Is ELK having issues?”
“Yes, logs are delayed by ~5 minutes.”
ALERT ON WHAT MATTERS

DEVELOP MEANINGFUL SLOs
Don’t try to be perfect

INCREASE YOUR OBSERVABILITY

IMPROVE YOUR ENVIRONMENT

TRUST PROVEN PARADIGMS

CONDUCT MEANINGFUL POSTMORTEMS
SERVICE RELIABILITY PRINCIPLES

1. Reliability is the most important feature of your service.
2. Your users determine what reliable means.
3. Nothing works all the time, so don’t aim for it.
SERVICE LEVEL INDICATORS

- A metric used to define how a service is operating
- Most often a ratio of good events over total events
- Measures how your service is doing from user’s perspective
THE RELIABILITY STACK
SERVICE LEVEL OBJECTIVES

- A target percentage informed by an SLI
- Often with a threshold involved
- Nothing is ever 100% reliable, so an SLO lets you pick a reasonable number
THE RELIABILITY STACK

Error Budget

SLO

SLI
ERROR BUDGETS

- Calculating how well your SLO has performed over a period of time
- An SLO implies acceptable levels of errors or problems
- For example, 99.9% available also means “we’re gonna have 43 bad minutes every 30 days.”
ERROR BUDGETS ARE AWESOME

Surplus Error Budget? → Do what you want!

Out of Error Budget? → Focus on reliability.
Tuesday, March 5th, 2019
20:00 ET

Yet another incident begins
Tuesday, March 5th, 2019
20:36 ET

Error budget exhaustion declared
Monday, March 4th, 2019
16:29 ET

SLO was defined
“A logline will be processed on average within 5 minutes 99% of the time.”
99% target = 
1% = 7h 18m 17s 
bad time/30 days
Jon Thornton  
Mar 5, 2019

This sounds like a great v1 SLO

Tanya Reilly  
Mar 5, 2019

+1!
With no remaining error budget, we gave ourselves permission to go all-in
3

ALERT ON WHAT MATTERS

DEVELOP MEANINGFUL SLOs

INCREASE YOUR OBSERVABILITY
You need to know what is happening to fix it

IMPROVE YOUR ENVIRONMENT

TRUST PROVEN PARADIGMS

CONDUCT MEANINGFUL POSTMORTEMS
ELK @ SQUARESPACE

Log Shippers → KAFKA → LOGSTASH → ELASTICSEARCH → KIBANA
ALERT ON WHAT MATTERS

DEVELOP MEANINGFUL SLOs

INCREASE YOUR OBSERVABILITY

IMPROVE YOUR ENVIRONMENT
Tooling and automation are your friends

TRUST PROVEN PARADIGMS

CONDUCT MEANINGFUL POSTMORTEMs
ELK @ SQUARESPACE

KAFKA → LOGSTASH → ELASTICSEARCH

Photo of a black hole
NASA, April 2019
ELK @ SQUARESPACE

Diagram showing the flow of Log Shippers to Kafka and then to Kibana.
ELK @ SQUARESPACE

- Log Shippers
- Log Shippers
- Log Shippers

KAFKA → LOGSTASH → ELASTICSEARCH

KIBANA
ALERT ON WHAT MATTERS

DEVELOP MEANINGFUL SLOs

INCREASE YOUR OBSERVABILITY

IMPROVE YOUR ENVIRONMENT

TRUST PROVEN PARADIGMS
We can learn from those that came before us

CONDUCT MEANINGFUL POSTMORTEMs
The tech industry is hurtling towards adopting known processes instead of continuing to invent our own.
THIS RELIABILITY STUFF ISN’T NEW

- Engineers have been working on reliability for as long as humans have been building stuff
- Statisticians have been analyzing data for centuries
- Emergency responders have been focused on response for just as long
THE INCIDENT COMMAND SYSTEM

- Formalized in 1968 by Fire Chiefs in Phoenix, Arizona
- They had resolved to streamline and improve response
- Based upon serious research and data
PROBLEMS THE ICS ADDRESSES

- Lack of accountability
- Poor communications
- No established hierarchy
- Too much freelancing
Delegation of Duties
Handing Off
INCIDENT
COMMANDER 2
A ticking time bomb was getting ready to explode...
Thursday, March 21st, 2019
20:00 ET

We’re better working together
TIMELINE OF A 37-HOUR INCIDENT

- **2019-03-21 20:00 ET** - Consumer lag starts increasing
- **20:05 ET** - Node with too many new shards identified
- **20:10 ET** - Ran cancel_future command to move shards
- **2019-03-22 00:30 ET** - Disengaged while the cluster recovers
- **01:00 ET** - precreate_indices script runs
- **01:00-06:30 ET** - Indexing slump as shards are moved
- **08:05 ET** - Allocation set to primaries
- **08:10 ET** - An apparently stuck node is restarted
- **08:48 ET** - Translog heap changed from 512MB to 2GB
- **10:00 ET** - All logstash nodes are restarted
- **13:02 ET** - Identified new erroring logs to filter out
- **14:18 ET** - Incident Commander hands off and goes to bed
- Probably tried changing something about the indexers
- Maybe tried to move more shards around
- Let’s try restarting the load balancers, or something?

2019-03-23 - 09:23 ET - ALL CLEAR
When your back is against the wall your perspective changes.
Sunday, March 24th, 2019
Afternoon
A beautiful day
SEE THE FOREST FOR THE TREES

- The problem was clearly shard related.
- But, what if it wasn’t the new shards...
- What if it was the **total** number of shards?
Google, “How many shards should I have in my Elasticsearch cluster?”
How many shards should I have in my Elasticsearch cluster?

By Christian Dahlqvist
“A node with a 30GB heap should therefore have a maximum of 600 shards, but the further below this limit you can keep it the better.”
600 < 2200
THE UNSHARDENING
ALERT ON WHAT MATTERS

DEVELOP MEANINGFUL SLOs

INCREASE YOUR OBSERVABILITY

IMPROVE YOUR ENVIRONMENT

TRUST PROVEN PARADIGMS

CONDUCT MEANINGFUL POSTMORTEMs

Learn from your own past
“A postmortem is an argument for change.”
- Nida Farrukh, Monitorama 2019
KEY COMPONENTS

Data Collection
1. Impact Assessment
   ▫ User-focused

2. Timeline
   ▫ Started, Detected, Engaged, Mitigated

3. Contributing Factors
   ▫ Root cause fallacy
KEY COMPONENTS

Data Collection

Lessons Learned
LESSONS LEARNED

- What went well?
- What went poorly?
- Where did we get lucky?
KEY COMPONENTS

Data Collection

Lessons Learned

Repair Items
REPAIR ITEMS

Incident Response
- Timeline analysis
- TT Detect
- TT Engage
- TT Mitigate

System
- Preventative / Mitigative
- “Why do we even X?”

Diversity of backgrounds and expertise is key!
April 23, 2019

Recording this chapter of ELK
Postmortem Report: 2019 State of ELK
Squarespace Engineering | Operational Excellence

Incident Date: 2019-01 to 2019-04-04
COE Jira Ticket: COE-730, COE-753, COE-774, COE-790
Authors: Alex Lee
Contributors: Mike Du Russel, Hannan Butt, Alex Conway, Weitao Jiang, Alex Hidalgo

Customer Impact
Over a period of several months, Squarespace engineers could not reliably depend on ELK.
April 29, 2019

The End

... Or is it?
ITERATE OVER EVERYTHING

Again and again and again...
You can make good be great.
PROGRESS IS INCREMENTAL

Things we continued to do for a while:
   ◦ Everything!

Things we continue to do until this day:
   ◦ Everything!
Friday, May 10th, 2019
15:30 ET

SLO target improved
"A logline will be processed on average within 5 minutes 99% of the time."
NEW:

“A logline will be processed on average within 2 minutes 99.9% of the time.”
Alex Hidalgo
3:30 PM May 10

We are updating this to 99.9% processed within an average of 2 minutes.

Jon Thornton
5:19 PM May 10

:dogeintesifies:
CONCLUSIONS
ALERT ON WHAT MATTERS
Put your users first

DEVELOP MEANINGFUL SLOs
Don’t try to be perfect

INCREASE YOUR OBSERVABILITY
You need to know what is happening to fix it

IMPROVE YOUR ENVIRONMENT
Tooling and automation are your friends

TRUST PROVEN PARADIGMS
We can learn from those that came before us

CONDUCT MEANINGFUL POSTMORTEMs
Learn from your own past, too

ITERATE OVER EVERYTHING
Again and again and again...
It’s always darkest before the dawn
“You’re going to be amazing.”
Thank you!

Alex Hidalgo - @ahidalgosre
Alex Lee - @ahl91

Shout Outs:
Squarespace Engineering
https://engineering.squarespace.com
Slidesgala.com