Incident Analysis

Your Organization's Secret Weapon

jeli

jeli.io

jeli

We've all had incidents.

The good news—it means your organization is important enough to have incidents

jeli

“What can we do to prevent this from ever happening again?”

jeli

“What caused this?”

jeli

“Why did this take so long to fix?”

jeli

“Why do you do incident reviews or postmortems?”

- “I’m honestly not sure”
- “Management wants us to”
- “It gives the engineers space to vent”
- “I think people would be mad if we didn’t”
- “Obligations to customers”
- “Tracking purposes”

jeli

Post-incident reviews are important

jeli

Activity: Cultivating curiosity & leveling up learning.

Rally Software
ca technologies A Broadcom Company

Root Cause Analysis:
Outage due to auto-patching

The following is a detailed accounting of the service outage that Rally users experienced on October 4th, 2019.

Root Cause Analysis Summary

Event Date	10/4/2019
Event Start	3:12am MDT
Time Detected	3:19am MDT
Time Resolved	4:39am MDT
Event End Time	4:57am MDT
Root Cause	Our DNS hosts all scheduled a normal reboot for security patches, resulting in a simultaneous outage of all DNS servers in our environment. As a result, all connections between all hosts in our environment failed due to DNS lookup issues. We were down for long enough that all reconnects between services and data stores broke, and our services were hard down. We investigated the confusing outage symptoms, and brought up all services.
Customer Impact	• Site was down for approximately one hour • Authentication service was down for a longer period • Most recent analytics data was not ingested in a timely fashion. 17 workspaces had stale data until resolved later in the morning • 2 support cases logged

Future Preventative Measures

Actions that should be taken to prevent this Event in the future.

Actions	Description
Quick restart playbook	Document quick ALM roll script and provide examples of when it should be used
Remove automated reboots for Windows hosts	Stop the servers from rebooting automatically, and add them back into the monthly patching stories.
Additional redundancy	Have one off DC AD as a redundancy for DNS. Add a domain controller in opposite datacenter to the resolvers on our linux hosts.
Documentation update	Update windows configuration documentation to include disabling automatic reboots

jeli.io

jeli

Rally Software
ca A Broadcom Company technologies

Root Cause Analysis:
Outage due to auto-patching

The following is a detailed accounting of the service outage that Rally users experienced on October 4th, 2019.

Root Cause Analysis Summary

Event Date	10/4/2019
Event Start	3:12am MDT
Time Detected	3:19am MDT
Time Resolved	4:39am MDT
Event End Time	4:57am MDT
Root Cause	Our DNS hosts all scheduled a normal reboot for security patches, resulting in a simultaneous outage of all DNS servers in our environment. As a result, all connections between all hosts in our environment failed due to DNS lookup issues. We were down for long enough that all reconnects between services and data stores broke, and our services were hard down. We investigated the confusing outage symptoms, and brought up all services.
Customer Impact	• Site was down for approximately one hour • Authentication service was down for a longer period • Most recent analytics data was not ingested in a timely fashion. 17 workspaces had stale data until resolved later in the morning • 2 support cases logged

Future Preventative Measures

Actions that should be taken to prevent this Event in the future.

Actions	Description
Quick restart playbook	Document quick ALM roll script and provide examples of when it should be used
Remove automated reboots for Windows hosts	Stop the servers from rebooting automatically, and add them back into the monthly patching stories.
Additional redundancy	Have one off DC AD as a redundancy for DNS. Add a domain controller in opposite datacenter to the resolvers on our linux hosts.
Documentation update	Update windows configuration documentation to include disabling automatic reboots

Did this report answer all your questions? What other questions do you have about the event?

jeli

Organizational approaches to post-incident activity

- Make Sure This Will Never Happen Again
- Fix The Broken Things
- Don’t shoot the messenger
- Learn From Incidents

Organizational approaches to post-incident activity

- Make Sure This Will Never Happen Again
- Fix The Broken Things
- Don’t shoot the messenger
- Learn From Incidents

How would you characterize your approach?

jeli

Post-incident reviews are important

jeli

Post-incident reviews are important

...but they're not "good"

jeli

Quantifying incident reviews

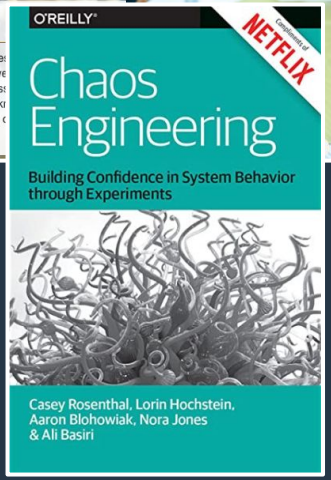
"Where are the people in this tracking?"

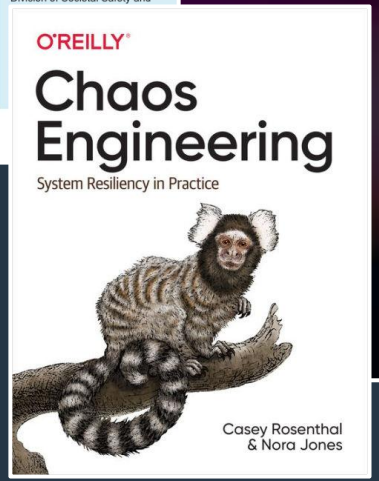
"Where are **you**?"

jeli

Gathering useful data about incidents does not come for free.

You need time and space to determine it.

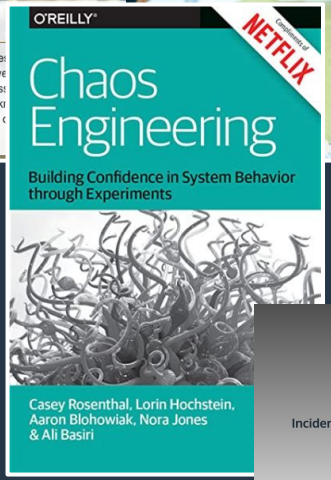


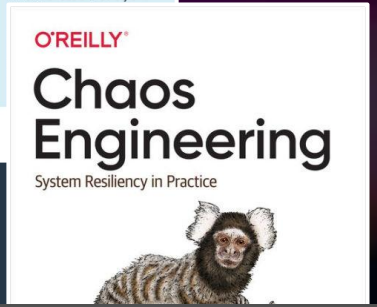


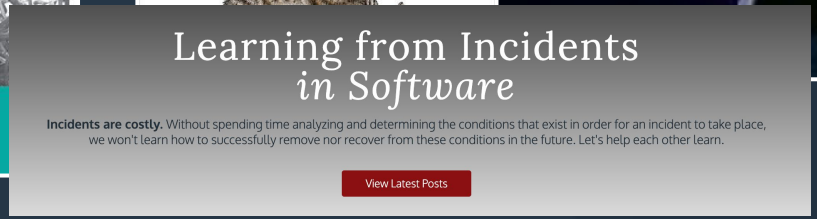
slack NETFLIX jet

jeli

jeli.io







slack NETFLIX jet

jeli

jeli.io

Performance Improvement = Errors ⬇ + Insights ⬆

Seeing What Others Don't - Gary Klein

Cases and Stories

Note: these are based on true events that I witnessed, but names and details have been changed

STORY ONE

We needed to make our Chaos Engineering tooling work better

jeli

Here’s the secret:
Incident analysis
is not actually about
the incident.

jeli

The incident is a catalyst to understanding how your org is structured in theory vs. how it is structured in practice.

jeli

The incident is a catalyst to understanding where you need to improve your sociotechnical system.

jeli

The incident is a catalyst to showing you what your organization is good at, and what needs improvement.

jeli

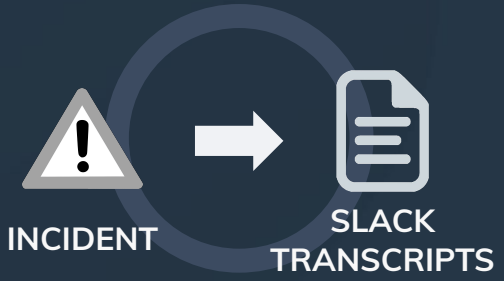
STORY TWO

It wasn’t just “human error”

Post-incident reviews are important

...but they're not "good"

jeli



Initial analysis of incident channel(s) to identify opportunities and initial interviewees.

Individual Interviews to determine:

- What their understanding of the event was
- What stood out for them as important
- What stood out for them as confusing/ambiguous/unclear
- What they believe they know about the event and how things actually work that they believe others don't

jeli

Cognitive Interviews

Knowledge and perspective gleaned in early interviews can point to important new topics to continue exploring:

- Relevant ongoing projects
- Past incidents
- Past experiences

jeli

SLACK transcripts
#incd-xxx, #dev-yyyy,
#zzzz, etc.

DOCS architecture diagrams,
previous related postmortems,
feature/product descriptions, etc.

GITHUB code, issues,
comments, etc.

XJIRA related tickets, previous
incident “action items,”
discussions, etc.

...other sources of data

jeli

STORY THREE

Promotion packets were due

jeli

A good incident analysis should tell you where to look

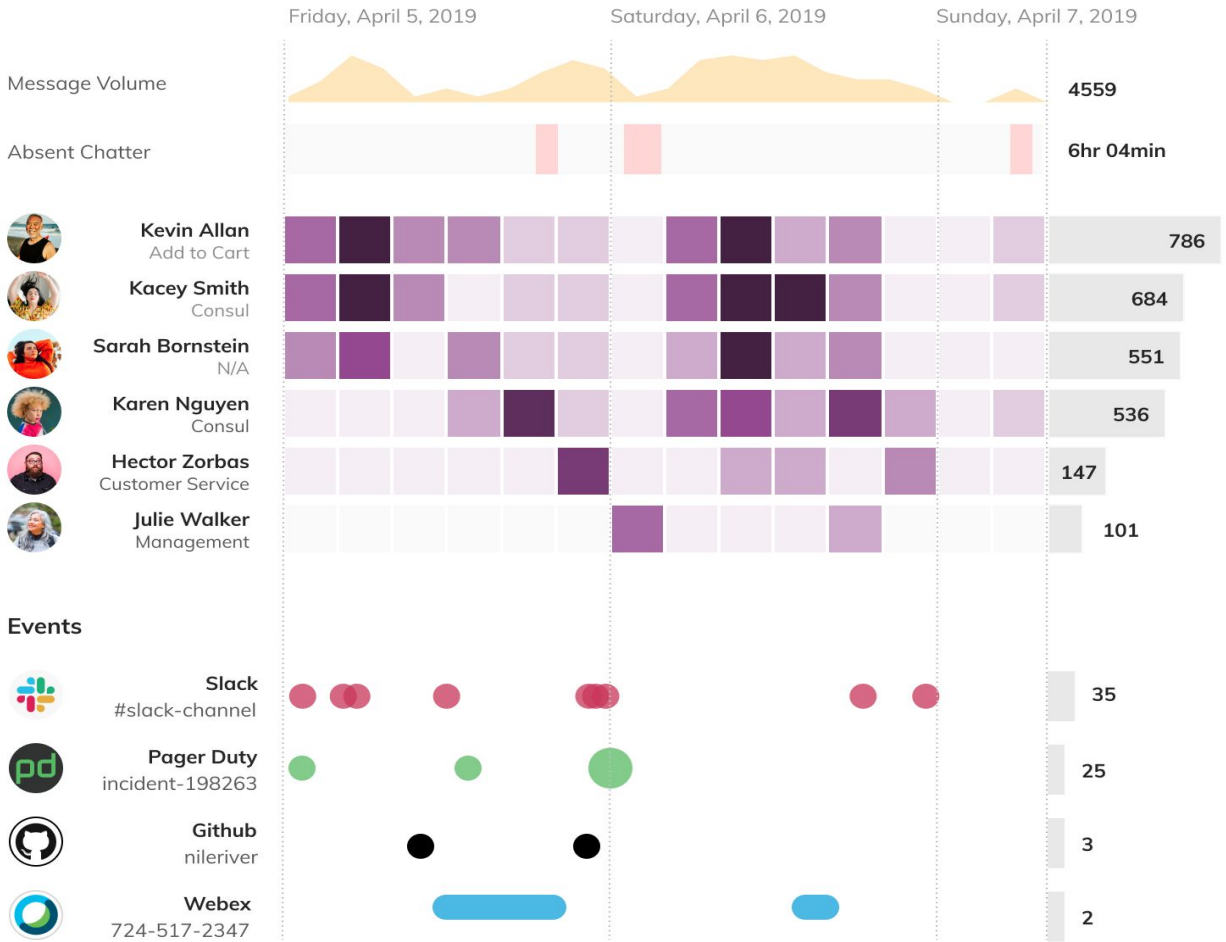
jeli

Messages

Group together: No grouping

Showing: All participants

Sorted by: Years of Tenure



A good incident analysis can help you with:

- Headcount
- Training
- Planning promotion cycles
- Quarterly planning
- Unlocking tribal knowledge
  Who came into the incident that wasn't supposed to
- How much are coordination efforts during incidents actually costing you
- Understanding bottlenecks...*in people*

jeli

Which kinds of incidents should be given more “time and space” to analyze?

jeli

There were more than two teams involved—especially if they hadn't worked together before

jeli

There were more than two teams involved—especially if they hadn't worked together before

It involved a misuse of something that seemed trivial (*cough*—like expired certs)

jeli

There were more than two teams involved—especially if they hadn't worked together before

It involved a misuse of something that seemed trivial (**cough**—like expired certs)

The "incident" or event, was almost *really* bad

jeli

There were more than two teams involved—especially if they hadn't worked together before

It involved a misuse of something that seemed trivial (**cough**—like expired certs)

The "incident" or event, was almost *really* bad

It took place during a big event (like an earnings call)

jeli

There were more than two teams involved—especially if they hadn't worked together before

It involved a misuse of something that seemed trivial (**cough**—like expired certs)

The "incident" or event, was almost *really* bad

It took place during a big event (like an earnings call)

A new service or interaction between services was involved

jeli

There were more than two teams involved—especially if they hadn't worked together before

It involved a misuse of something that seemed trivial (**cough**—like expired certs)

The "incident" or event, was almost *really* bad

It took place during a big event (like an earnings call)

A new service or interaction between services was involved

More people joined the channel than usual

jeli

When are we ready for “incident analysis”?

jeli

Having customers means you are ready to benefit from incident analysis.

jeli

What can you do today to improve incident analysis?

- Give folks time and space to get better at analysis—this can be trained
- Come up with different metrics—look at the people
- Investigator on-call rotations
- Allow time for investigation of the “big ones”
- Jeli workshop - Move Fast and Learn from Incidents
- Jeli tooling

jeli

How do you know if it’s working?

jeli

How do you know if it's working?

- More folks are reading the incident review
- More folks are attending the incident review
- You're not seeing the same folks pop into every incident
- Folks feel more confident
- Teams are collaborating more
- Better shared understanding of the definition of an incident

jeli

“I just changed the way I was proposing to use $X in a design as a result of reading this document”

jeli

“Never have I ever seen such an in-depth analysis of any software system that I’ve had the pleasure to work with. Anyone who will read this document should come out more informed and even have a better understanding of the services that started off as having 1 or 2 persons understanding. This is a beautiful educational piece that anyone who plans on using $x should read.”

jeli

“Hearing from both devs and platform on terraform-related work was valuable”

-- attendee from Articulate review

jeli

“Hearing from others in the room about how roles are created was really valuable.”

-- attendee from Articulate review

What are the components of a strong post-incident process?

1. Incident occurs
2. Investigation assigned
3. Investigation accepted
4. Initial analysis by investigator to identify interviewees and opportunities
5. Investigator analysis of disparate sources: *Slack*, *PRs*, *Tickets*, *PagerDuty*, etc.
6. Individual interviews
7. Calibration Document (align with participants on the event)
8. Facilitated Meeting
9. Report
10. Action Items

jeli

Further resources on incident analysis

www.learningfromincidents.io

The Error of Counting Errors by Robert L. Wears

jeli