

Incident Management For Data People



Kyle Kirwan
Founder & CEO

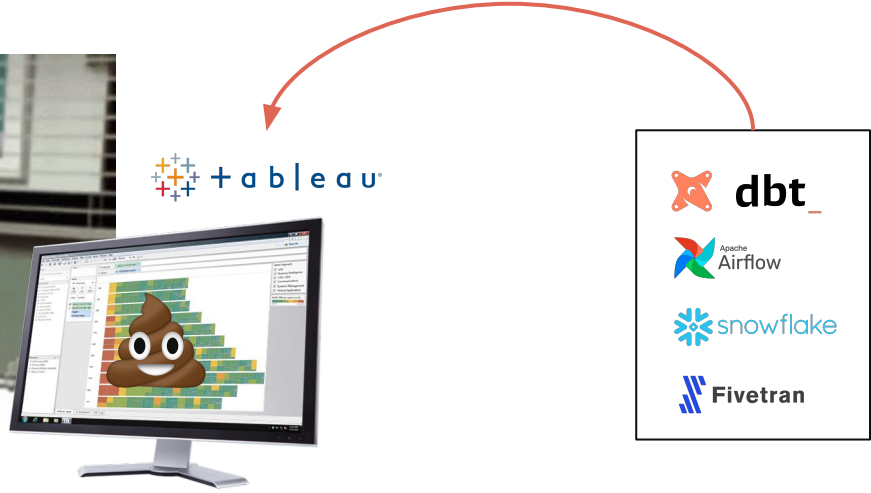


**An unplanned interruption
to a service or a reduction
in quality of a service.**

ITIL 4, 2019



Your users



You

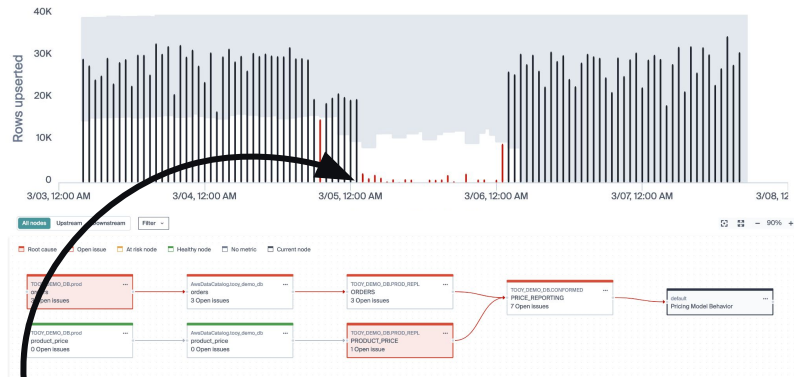


Identification

Monitoring system sends alerts + pages.







API endpoint latency up



Snowflake table inserts down

Mobilization

Incident management team forms.

		Decide	Document	Communicate	Resolve
	Commander Staff Data Eng	✓	✗	✗	✗
	Scribe Data Scientist	✗	✓	✗	✗
	Liaison Data Eng. Mgr.	✗	✗	✓	✗
	SME Analytics Eng. Data Eng.	✗	✗	✗	✓

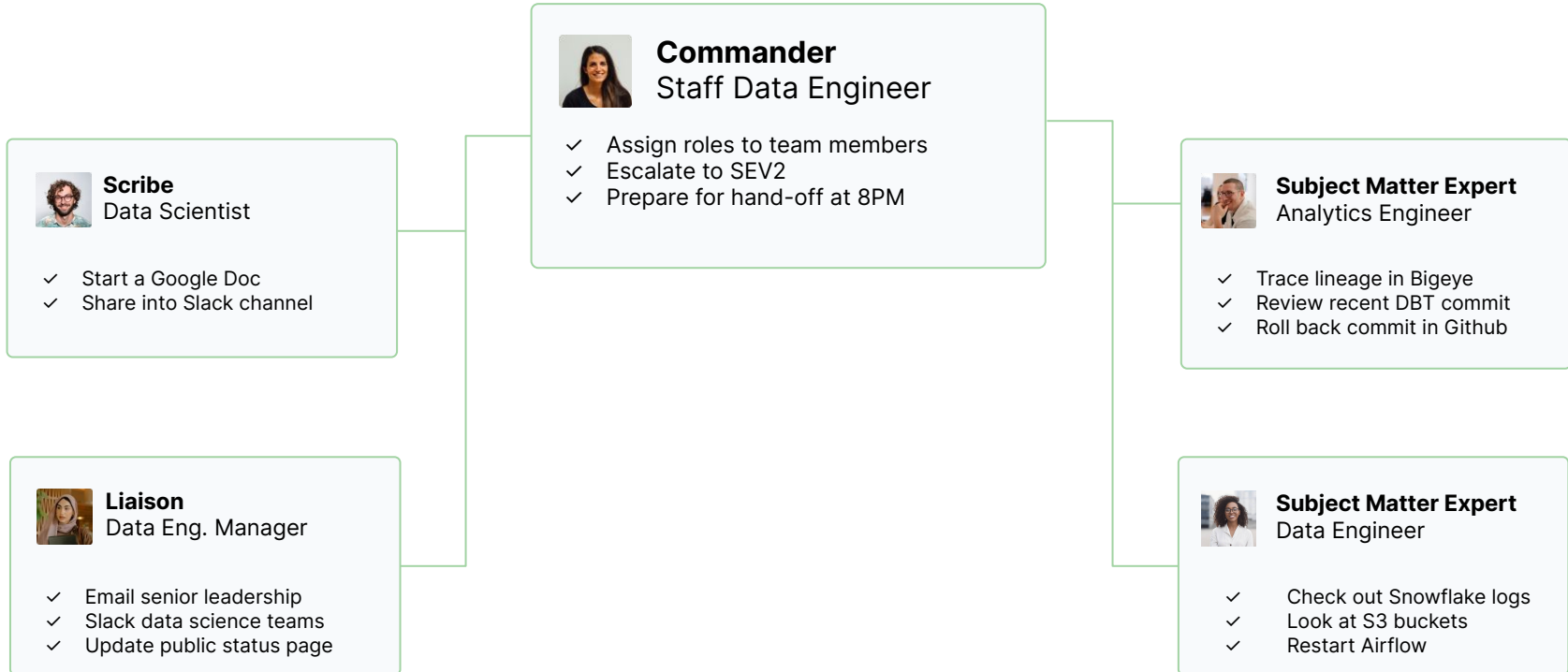
Diagnosis

Assess impact. Escalate. Start root cause analysis.

	Service degradation	Impact to users
SEV 1	Unavailable	Most users affected
SEV 2	Significant problems	Many users affected
SEV 3	Performance problems	Some users affected
SEV 4	Performance problems	None
SEV 5	Low level annoyances	None

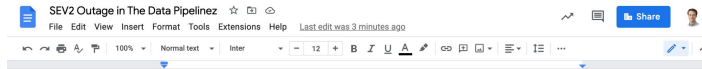
Resolution

Take actions to resolve. Confirm recovery.



Closure

Review and identify preventive measures.



EXAMPLE:

No specific items in the backlog that could have improved this service. There is a note about improvements to flow typing, and these were ongoing tasks with workflows in place. There have been tickets submitted for improving integration tests but so far they haven't been successful.

Recurrence

Now that you know the root cause, can you look back and see any other incidents that could have the same root cause? If yes, note what mitigation was attempted in those incidents and ask why this incident occurred again.

EXAMPLE:

This same root cause resulted in incidents HOT-13432, HOT-14932 and HOT-19452.

Lessons learned

Discuss what went well in the incident response, what could have been improved, and where there are opportunities for improvement.

EXAMPLE:

- Need a unit test to verify the rate-limiter for work has been properly maintained
- Bulk operation workloads which are atypical of normal operation should be reviewed
- Bulk ops should start slowly and monitored, increasing when service metrics appear nominal

Corrective actions

Describe the corrective action ordered to prevent this class of incident in the future. Note who is responsible and when they have to complete the work and where that work is being tracked.

EXAMPLE:

1. Manual auto-scaling rate limit put in place temporarily to limit failures
2. Unit test and re-introduction of job rate limiting
3. Introduction of a secondary mechanism to collect distributed rate information across cluster to guide scaling effects



The Hidden Costs of Poor Incident Management

FireHydrant Solutions Product Resources Integrations Pricing Log In Start for free

INCIDENT MANAGEMENT BLOG Product releases How to Engage

LISTEN February 16, 2023

The hidden costs of poor incident management

We're all looking to maximize every dollar spent, every line made, every hour logged. But there's one cost center you might not be thinking about — incident management. This post explores the explicit and implicit costs associated with incidents.

[Read article](#) By Robert Ross



Incident Management Guide

PagerDuty Incident Response

Incident Response Home

Home

This documentation covers parts of the PagerDuty Incident Response process. It is a cut-down version of our internal documentation used at PagerDuty for any major incidents and to prepare new employees for on-call responsibilities. It provides information not only on preparing for an incident, but also what to do during and after the incident. It is intended to be used by on-call practitioners and those involved in an operational incident response process for those wanting to enact a formal incident response process. See the [about page](#) for more information on what this documentation is and why it exists.

👉 Don't know where to start?

[How to handle incident response \(and don't get a formal process in your organization, we recommend\)](#)

Additional links:



SRE Handbook

Chapter 3 - Incident Response

Incident Response

By Jennifer Mace, Jelena Oertel, Stephen Thome, and Anup Chakrabarti (PagerDuty) with Jian Ma and Jessie Yang

Everyone wants their services to run smoothly all the time, but we live in an imperfect world in which outages do occur. What happens when a not-so-ordinary, urgent problem requires multiple individuals or teams to resolve it? You are suddenly faced with simultaneously managing the incident response and resolving the problem.

Incident Handbooks

ATLASSIAN

Incident management for high-velocity teams

Atlassian Incident Handbook

- Incident Management home
- Incident response
- Responsibilities

Teams using tech services today are expected to maintain 24/7 availability. When something goes wrong, whether it's an outage in a broken feature, team members need to respond immediately and restore service. The process is called **incident management**, and it's an ongoing, complex challenge for engineering products.

Thanks!



Kyle Kirwan
kyle@bigeye.com

