Expedition To The LLM & AI Frontier

Sanghamitra Goswami

Senior Director, PagerDuty

Data Science & Machine Learning

October 17th, LeadDev West Coast 2023

Introduction to PagerDuty

- Pagerduty Operations Cloud
- PagerDuty - AI

AI to Generative AI

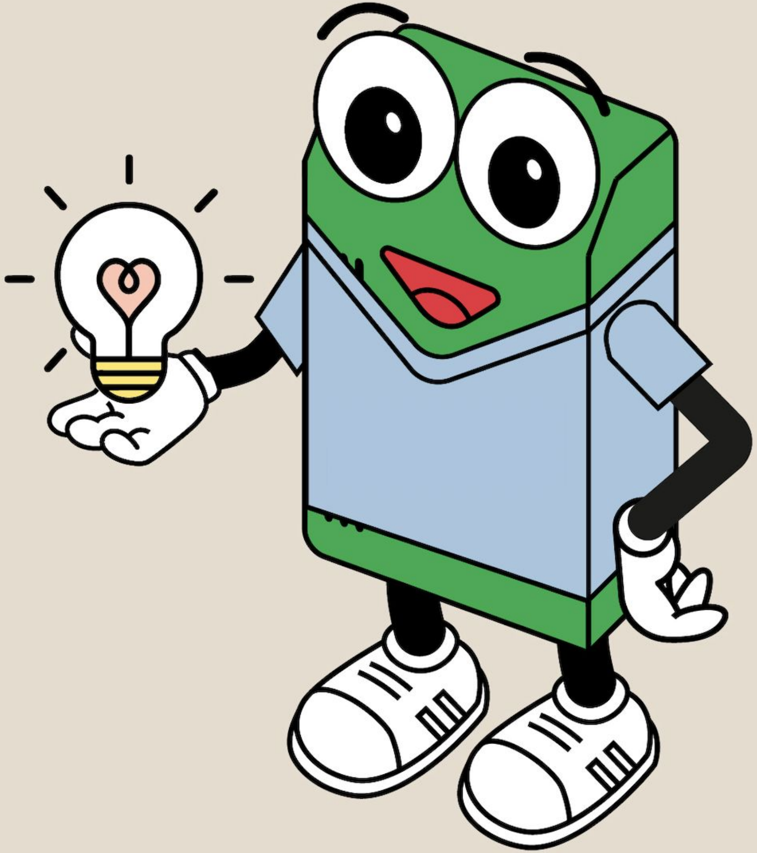
- Challenges
- Multi-modal Approach

New Generative AI Features

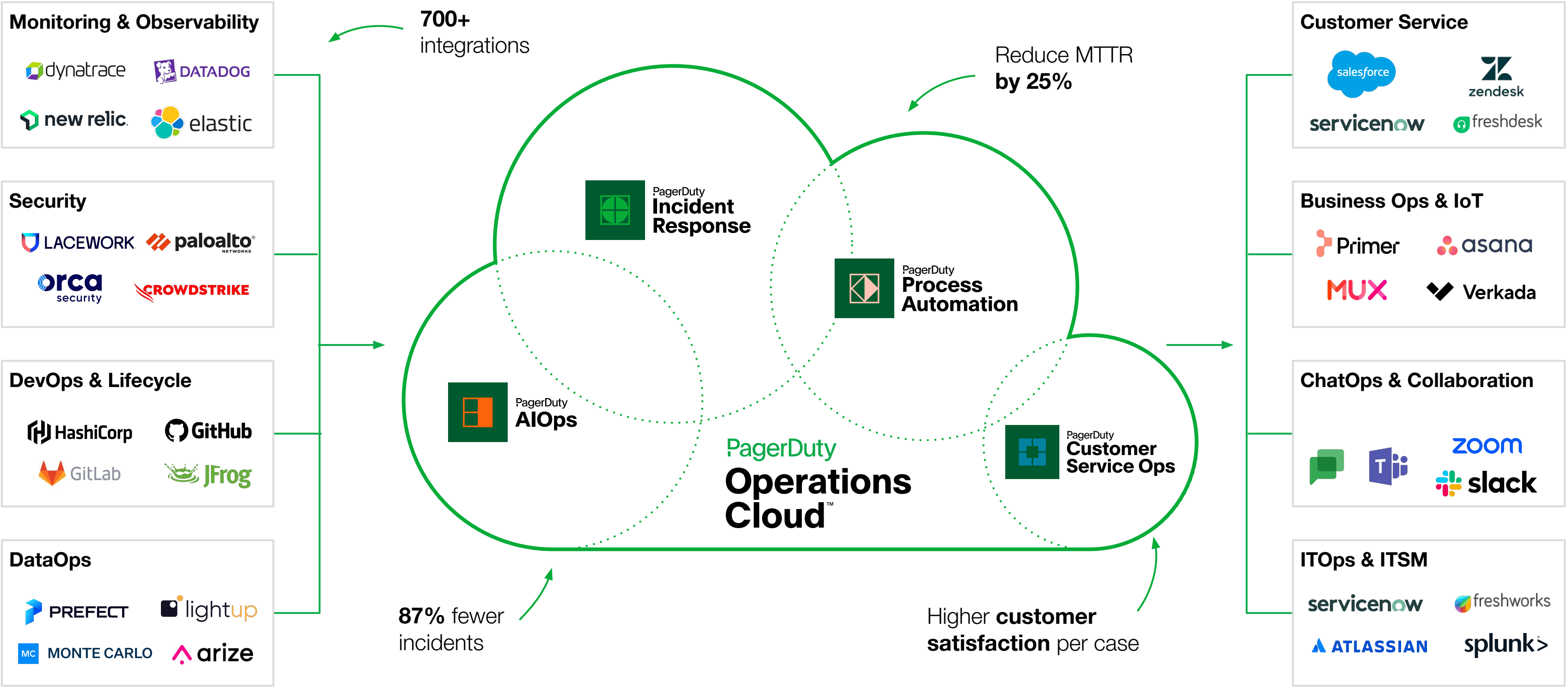
- SMART STUNT
- AI Generated Postmortems
- AI Generated Runbooks

PagerDuty

Through its SaaS-based platform, PagerDuty empowers developers, DevOps, IT operations and business leaders to prevent and resolve business-impacting incidents



PagerDuty Operations Cloud Ecosystem



Monitoring & Observability

dynatrace · DATADOG · new relic. · elastic

Security

LACEWORK · paloalto NETWORKS · orca security · CROWDSTRIKE

DevOps & Lifecycle

HashiCorp · GitHub · GitLab · JFrog

DataOps

PREFECT · lightup · MC MONTE CARLO · arize

700+ integrations

Reduce MTTR **by 25%**

PagerDuty **Incident Response**

PagerDuty **Process Automation**

PagerDuty **AIOps**

PagerDuty **Customer Service Ops**

PagerDuty **Operations Cloud™**

87% fewer incidents

Higher **customer satisfaction** per case

Customer Service

salesforce · zendesk · servicenow · freshdesk

Business Ops & IoT

Primer · asana · MUX · Verkada

ChatOps & Collaboration

zoom · slack

ITOps & ITSM

servicenow · freshworks · ATLASSIAN · splunk>

PagerDuty

AI is part of our DNA

AI has been part of our platform

AIOps

Incident Response

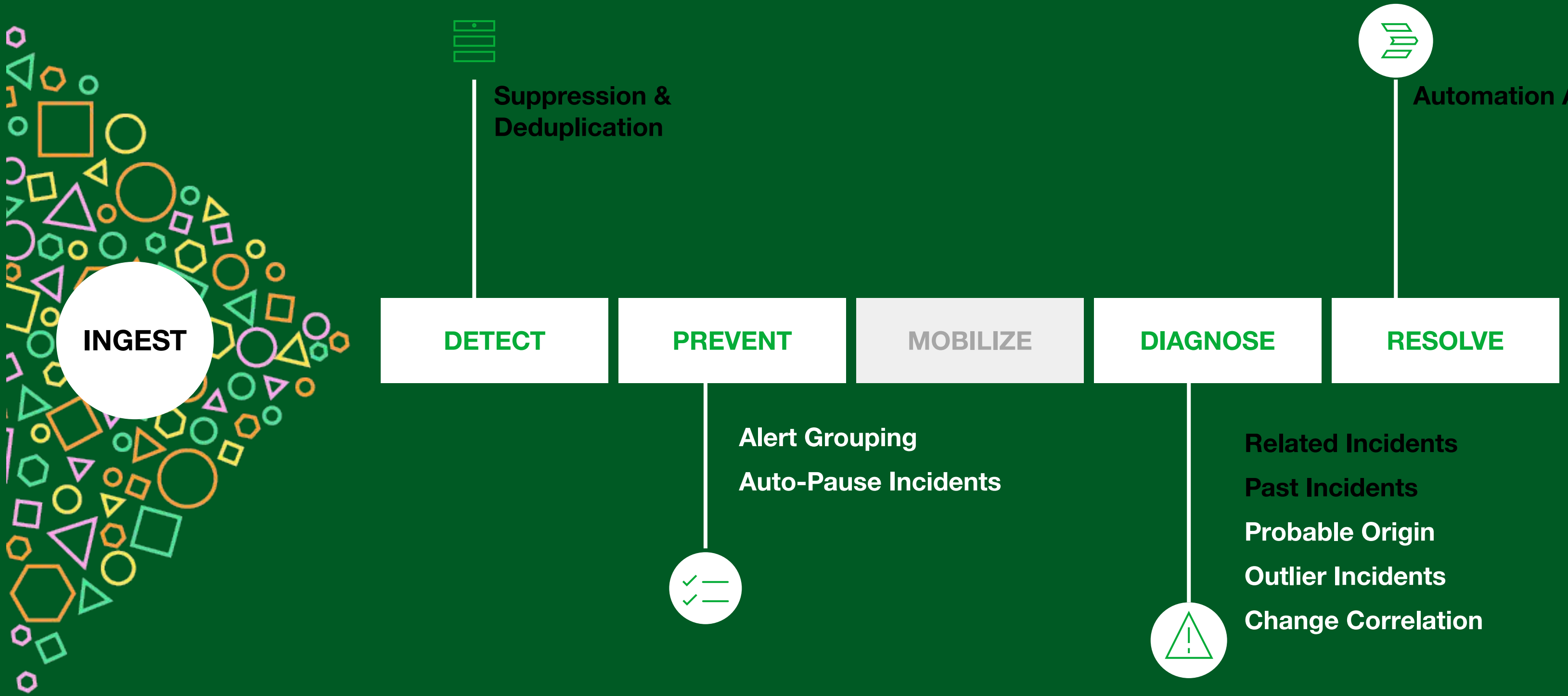
Process Automation

Customer Service Ops

We plan to incorporate AI throughout our product portfolio

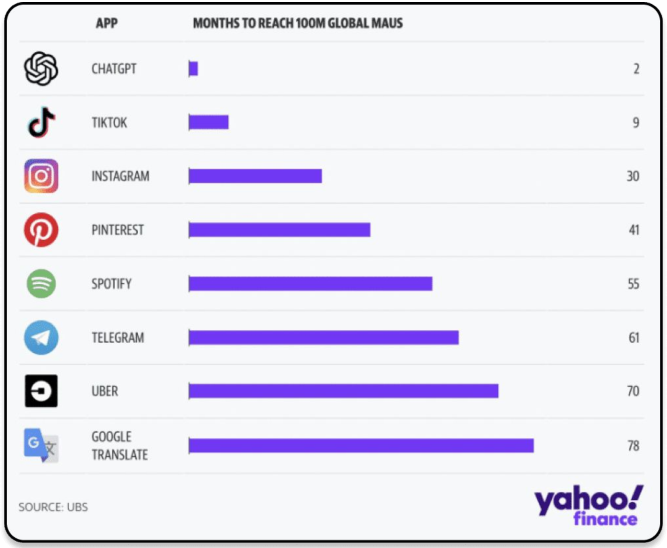
PagerDuty AI

Reduce Noise, Create Context, Automate Toil



AI To Generative AI - Our Expedition

Generative AI Landscape a Few Months Back



SOURCE: UBS

10 months ago OpenAI released ChatGPT

- Crazy adoption rate as people were amazed by its capabilities.
- Sparked a lot of businesses to integrate it into their products
- We wanted to innovate and add to our Product

Our Goal:

Let’s adopt Generative AI and develop Product features

Key Challenges

- New planning
- Ratio of data scientists to software engineers low
- New technology
- Concurrent workflows
- Roadmap realignment

PagerDuty

Our Multi-Modal Approach

- Empower & Educate
- Establish Rules of Engagement
- Discover Your Use Cases
- Establish Data Privacy & Legal guidelines
- Determine the LLM of Your Choice
- Set Operating Principles To Release

Empower & Educate

Hack Week → Work in a team, learn from each other

Several Gen AI focussed slack channels → Share the learning, curated courses

Data Science/AI Office hours → Ask questions to your Data Science Team

Form a core lead group→ A multidisciplinary group with AI, legal, security and engineering

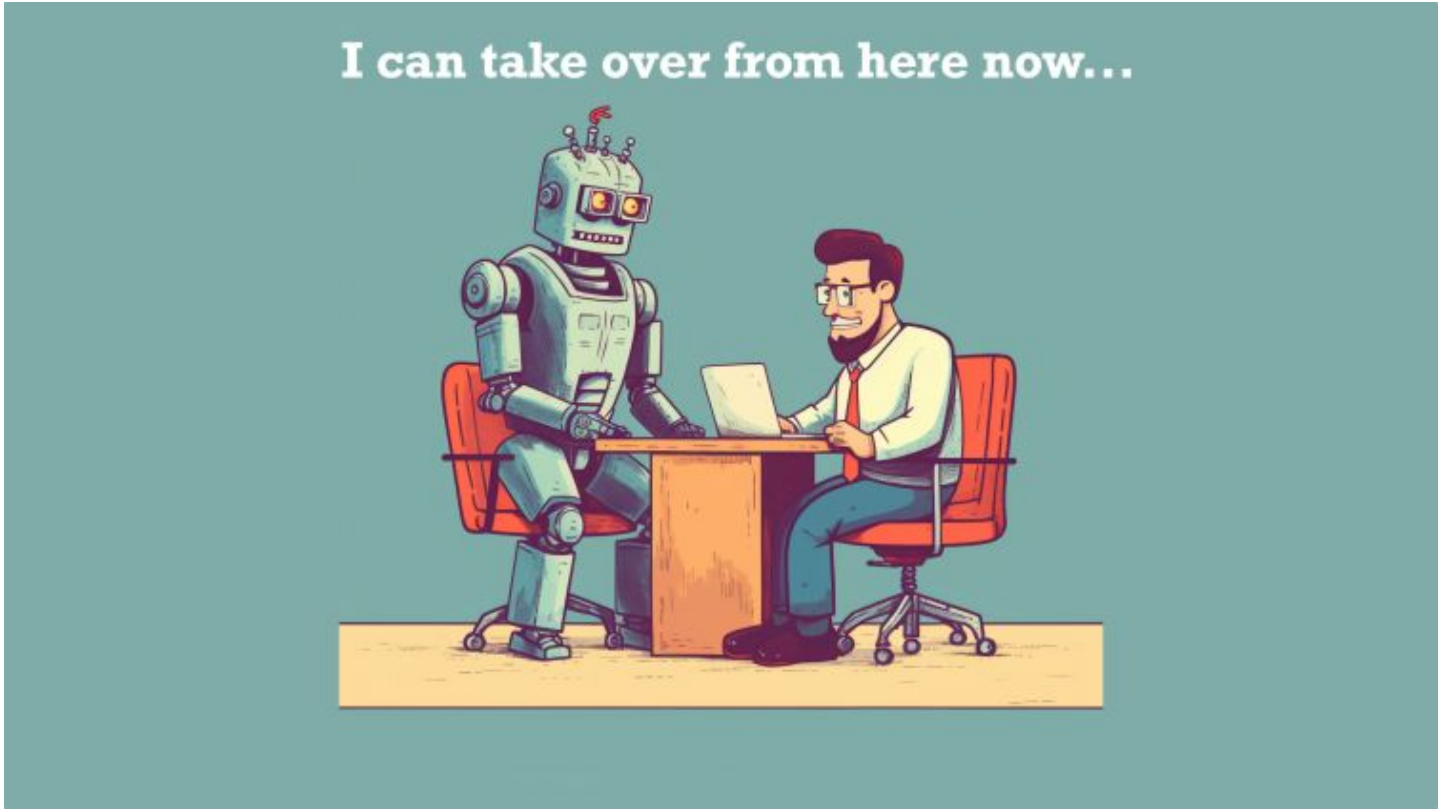
Dedicate x% of resources from all engineering teams to learn → AI champions in every team

Establish Rules on Engagement

- Involve a Data Scientist in early conversations
- Start a design document early
- Involve Legal and Security early

PagerDuty

You Need Data Scientists



Role Evolving…

Explore data feasibility	Garbage in garbage out
Building prompt Vs. Fine tuning	Costly interactions
Build within a Generative AI feature	Make Product unique
Evaluate feature performance	Measure performance & quality
Assist Product in selecting right customers	Product customer fit

Discover your Use Cases

Code Generation

- We pivoted to one of our values **#champinionthecustomer**
- Where can Gen AI help?
- What are some of the use cases for PagerDuty?
- Limitations?

Chat Bot

Summarization

Determine the LLM of your Choice

Accuracy → Score outputs from LLMs

Cost → Queries, especially those requiring long prompts can be expensive

Latency → It is not uncommon for a query to take several seconds

Context Window → Several of our use cases require large context windows

Data Governance → Control over our data

Ease of setup & SLA → Effort for PagerDuty to access, manage and maintain

Operating Principles To Release

- Internal Preview
- External Preview
- Early Access
- General Access
- Human-in-the-loop

PagerDuty

Key Takeaways

- Preserve Optionality
  - Build for evolution, not perfection
- Manage Risks
  - Customer Trust is the #1 priority
- Focus on customer adoption
  - Show new value against their jobs-to-be-done

GenAI features just announced

1. AI-Generated
Status Updates

2. AI-Generated
Incident Postmortems

3. AI-Generated
Process Automation

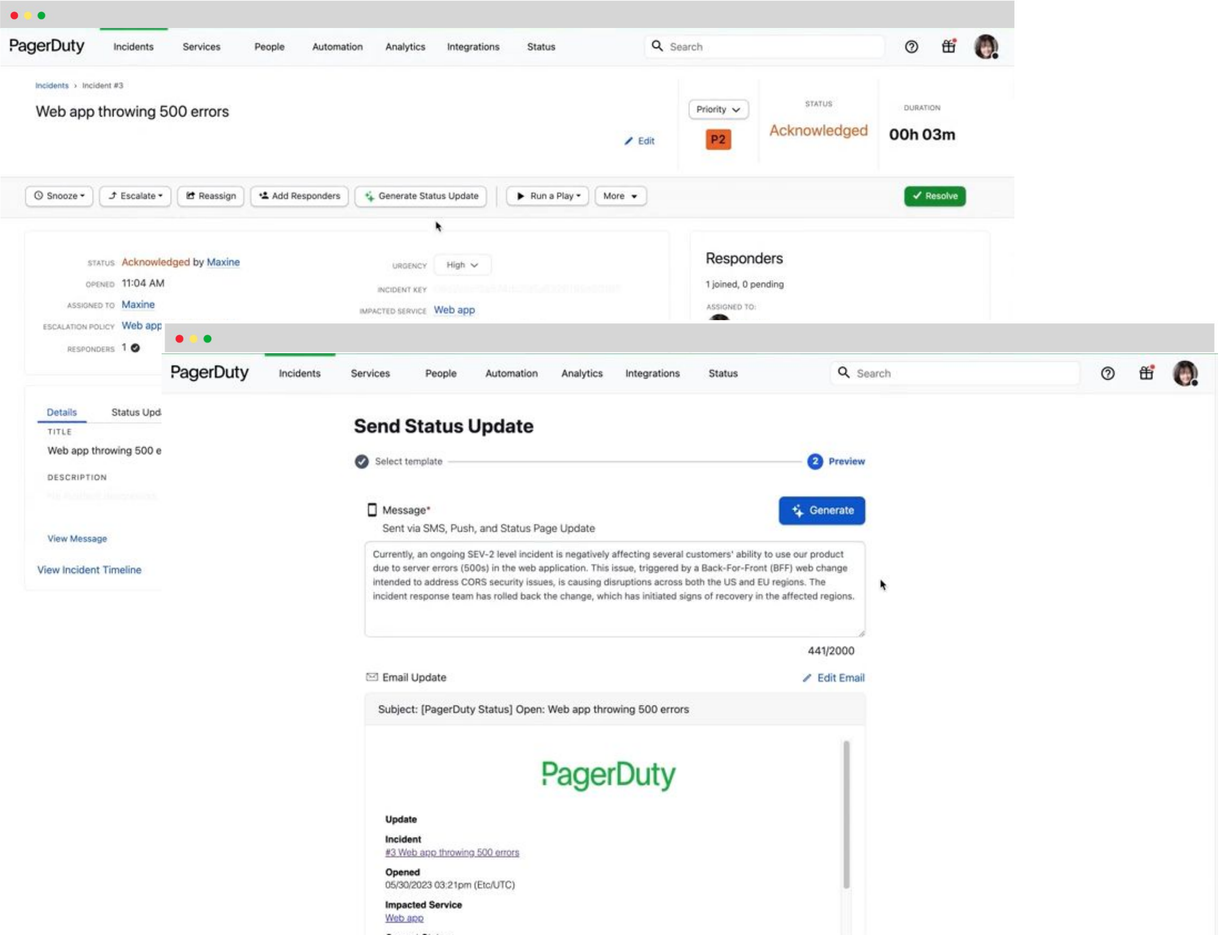
AI-Generated Status Updates

One-click status updates, built for any audience

- Generate persona-specific status updates with just a few clicks
- Make it easy for responders to keep internal stakeholders and executives in the loop

Demo

Waitlist available



Incidents › Incident #3

Web app throwing 500 errors

P2

STATUS Acknowledged

DURATION 00h 03m

STATUS Acknowledged by Maxine
OPENED 11:04 AM
ASSIGNED TO Maxine
ESCALATION POLICY Web app
RESPONDERS 1

URGENCY High
INCIDENT KEY
IMPACTED SERVICE Web app

Responders
1 joined, 0 pending
ASSIGNED TO:

TITLE
Web app throwing 500 e

DESCRIPTION

Send Status Update

Select template — 2 Preview

Message*
Sent via SMS, Push, and Status Page Update

Currently, an ongoing SEV-2 level incident is negatively affecting several customers' ability to use our product due to server errors (500s) in the web application. This issue, triggered by a Back-For-Front (BFF) web change intended to address CORS security issues, is causing disruptions across both the US and EU regions. The incident response team has rolled back the change, which has initiated signs of recovery in the affected regions.

441/2000

Email Update

Subject: [PagerDuty Status] Open: Web app throwing 500 errors

PagerDuty

Update

Incident
#3 Web app throwing 500 errors

Opened
05/30/2023 03:21pm (Etc/UTC)

Impacted Service
Web app

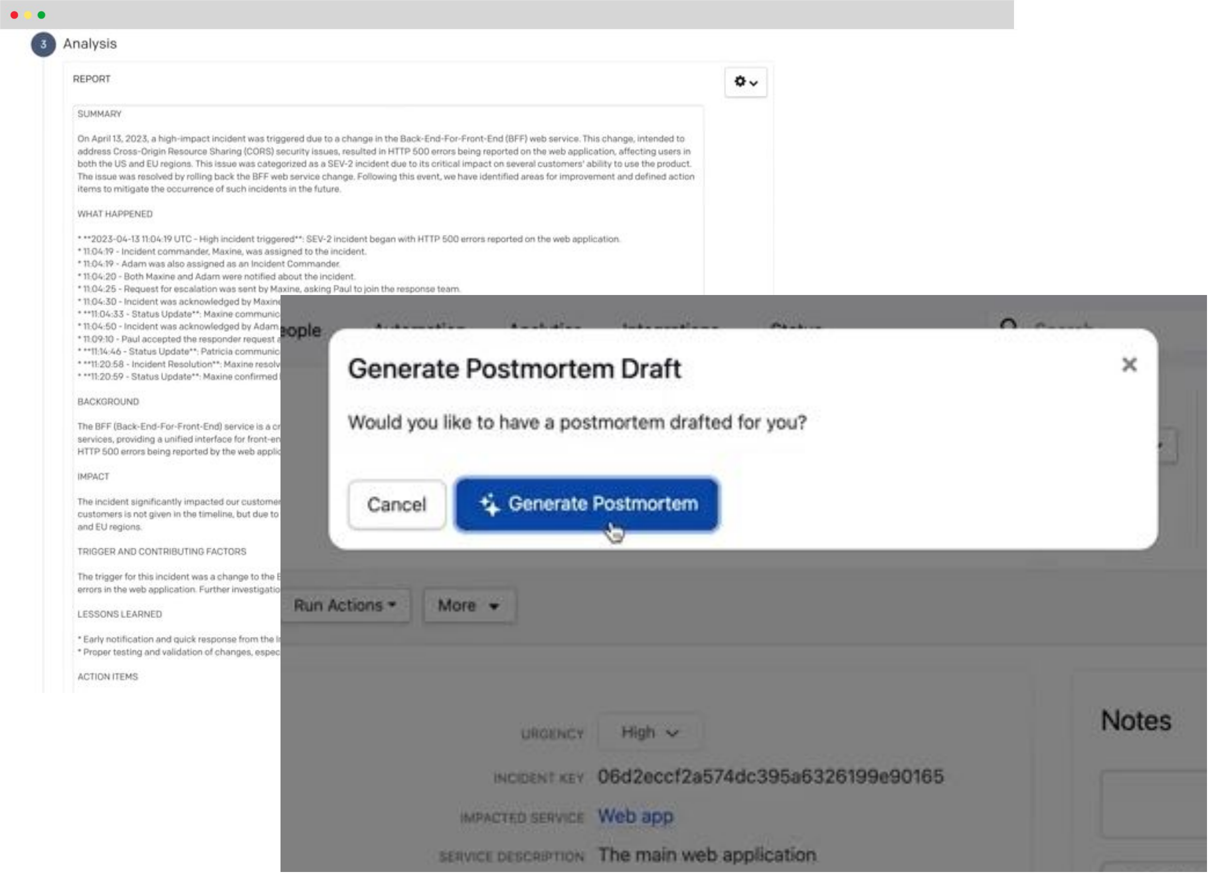
Save time and remove toil from postmortems

- Use AI to create a comprehensive summary of what happened, when, how it was resolved, and key actions for next time
- Automatically collect and collate incident data (including logs, metrics, and relevant Slack or Microsoft Teams conversations

Demo

Waitlist available

AI-Generated Incident Postmortems



3 Analysis

REPORT

SUMMARY

On April 13, 2023, a high-impact incident was triggered due to a change in the Back-End-For-Front-End (BFF) web service. This change, intended to address Cross-Origin Resource Sharing (CORS) security issues, resulted in HTTP 500 errors being reported on the web application, affecting users in both the US and EU regions. This issue was categorized as a SEV-2 incident due to its critical impact on several customers' ability to use the product. The issue was resolved by rolling back the BFF web service change. Following this event, we have identified areas for improvement and defined action items to mitigate the occurrence of such incidents in the future.

WHAT HAPPENED

```
* **2023-04-13 11:04:19 UTC - High incident triggered**: SEV-2 incident began with HTTP 500 errors reported on the web application.
* 11:04:19 - Incident commander, Maxine, was assigned to the incident.
* 11:04:19 - Adam was also assigned as an Incident Commander.
* 11:04:20 - Both Maxine and Adam were notified about the incident.
* 11:04:25 - Request for escalation was sent by Maxine, asking Paul to join the response team.
* 11:04:30 - Incident was acknowledged by Maxine
* **11:04:33 - Status Update**: Maxine communic
* 11:04:50 - Incident was acknowledged by Adam
* 11:09:10 - Paul accepted the responder request a
* **11:14:46 - Status Update**: Patricia communic
* **11:20:58 - Incident Resolution**: Maxine resolv
* **11:20:59 - Status Update**: Maxine confirmed
```

BACKGROUND

The BFF (Back-End-For-Front-End) service is a cr
services, providing a unified interface for front-en
HTTP 500 errors being reported by the web applic

IMPACT

The incident significantly impacted our customer
customers is not given in the timeline, but due to
and EU regions.

TRIGGER AND CONTRIBUTING FACTORS

The trigger for this incident was a change to the E
errors in the web application. Further investigatio

LESSONS LEARNED

* Early notification and quick response from the I
* Proper testing and validation of changes, espec

ACTION ITEMS

Generate Postmortem Draft

Would you like to have a postmortem drafted for you?

URGENCY: High

INCIDENT KEY: 06d2eccf2a574dc395a6326199e90165

IMPACTED SERVICE: Web app

SERVICE DESCRIPTION: The main web application

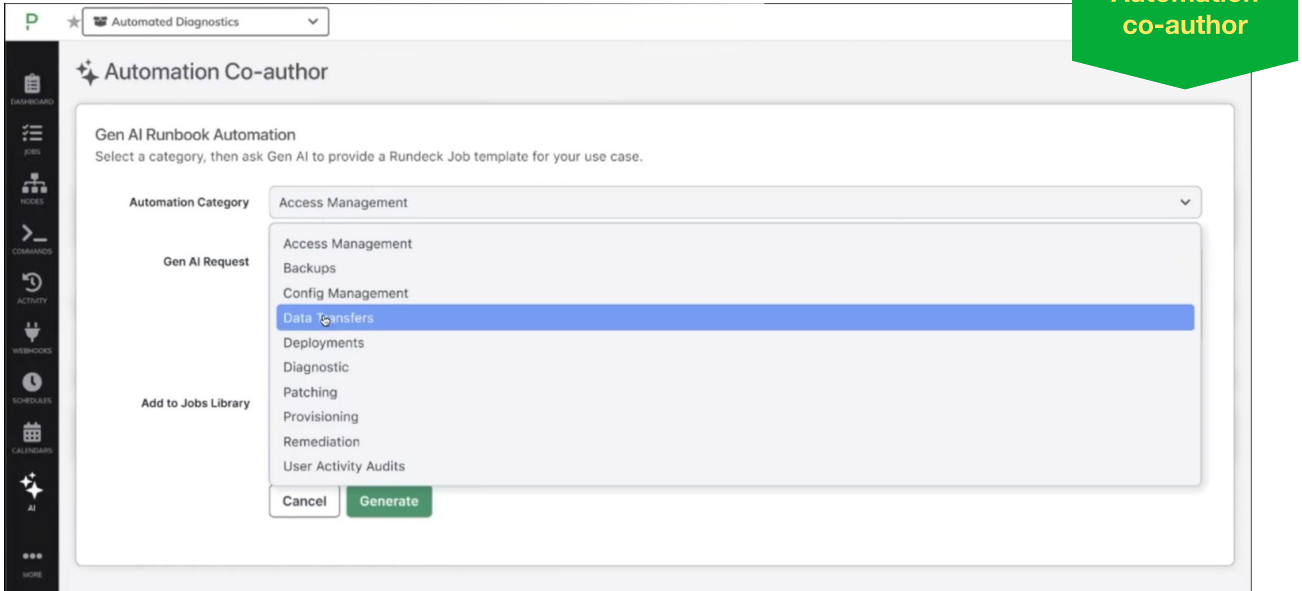
Generate Process Automation job definitions via AI

- Get a fast start in developing new workflows
- Speed up learning and problem solving
- Generate results for any scripting language

Demo

Waitlist available

Automation co-author



Automation Co-author

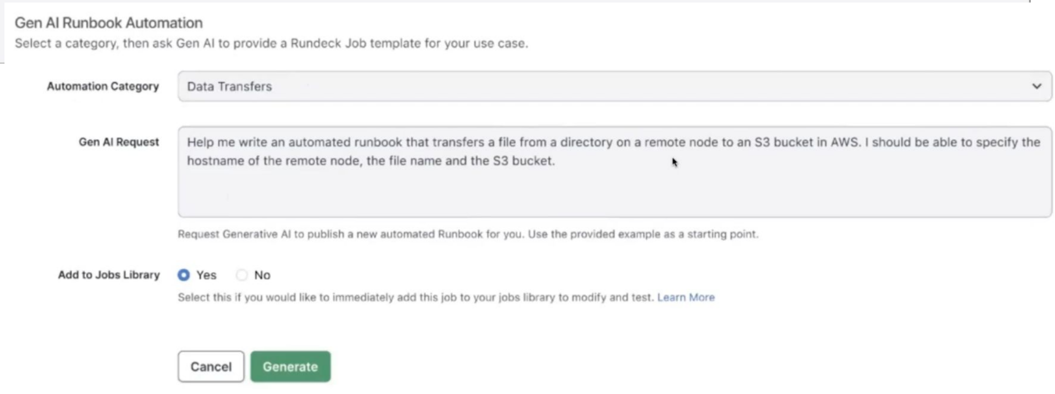
Gen AI Runbook Automation

Select a category, then ask Gen AI to provide a Rundeck Job template for your use case.

Automation Category: Access Management

Gen AI Request

Add to Jobs Library



Gen AI Runbook Automation

Select a category, then ask Gen AI to provide a Rundeck Job template for your use case.

Automation Category: Data Transfers

Gen AI Request: Help me write an automated runbook that transfers a file from a directory on a remote node to an S3 bucket in AWS. I should be able to specify the hostname of the remote node, the file name and the S3 bucket.

Request Generative AI to publish a new automated Runbook for you. Use the provided example as a starting point.

Add to Jobs Library: ● Yes ○ No

Select this if you would like to immediately add this job to your jobs library to modify and test. Learn More

What’s Next

- Sign up for PagerDuty's waitlist
- Watch an extended demo that shows some of these features in action https://youtu.be/OgBS2cvkiCs
- Hear more from the team on learnings https://youtu.be/2mEVY6rmX3M

Thank You

Establish Rules on Engagement

Start a design document early

- This tightens everything up and gives a basis for discussion
- Include the API contract when working with external vendors

Involve Legal and Security Early

- We partnered with legal from the beginning, involving them in the earliest stages of the process. This allowed us to get early advice on how to think about the short, medium and long-term release phases

Key Challenges

New planning → We wanted to move fast while the industry landscape is continuously evolving, changing the requirements

Ratio of data scientists to software engineers low → Not enough headcount with the required skill set

New technology → Lots of learning on every front (AI, Security, Legal, etc.,)

Concurrent workflows, everyone is excited → Work duplication (code and skill development)

Roadmap realignment → Introducing a completely new item in the roadmap

Establish Rules on Engagement

You have a potential Use Case for LLM → follow this process

Involve a Data Scientist in early conversations

- Feasibility validation
- Suggesting alternative solutions (do you need an LLM?)
- What support you need
    - Prompt engineering
    - Testing plan
    - Use case - model efficacy
    - API consolidation