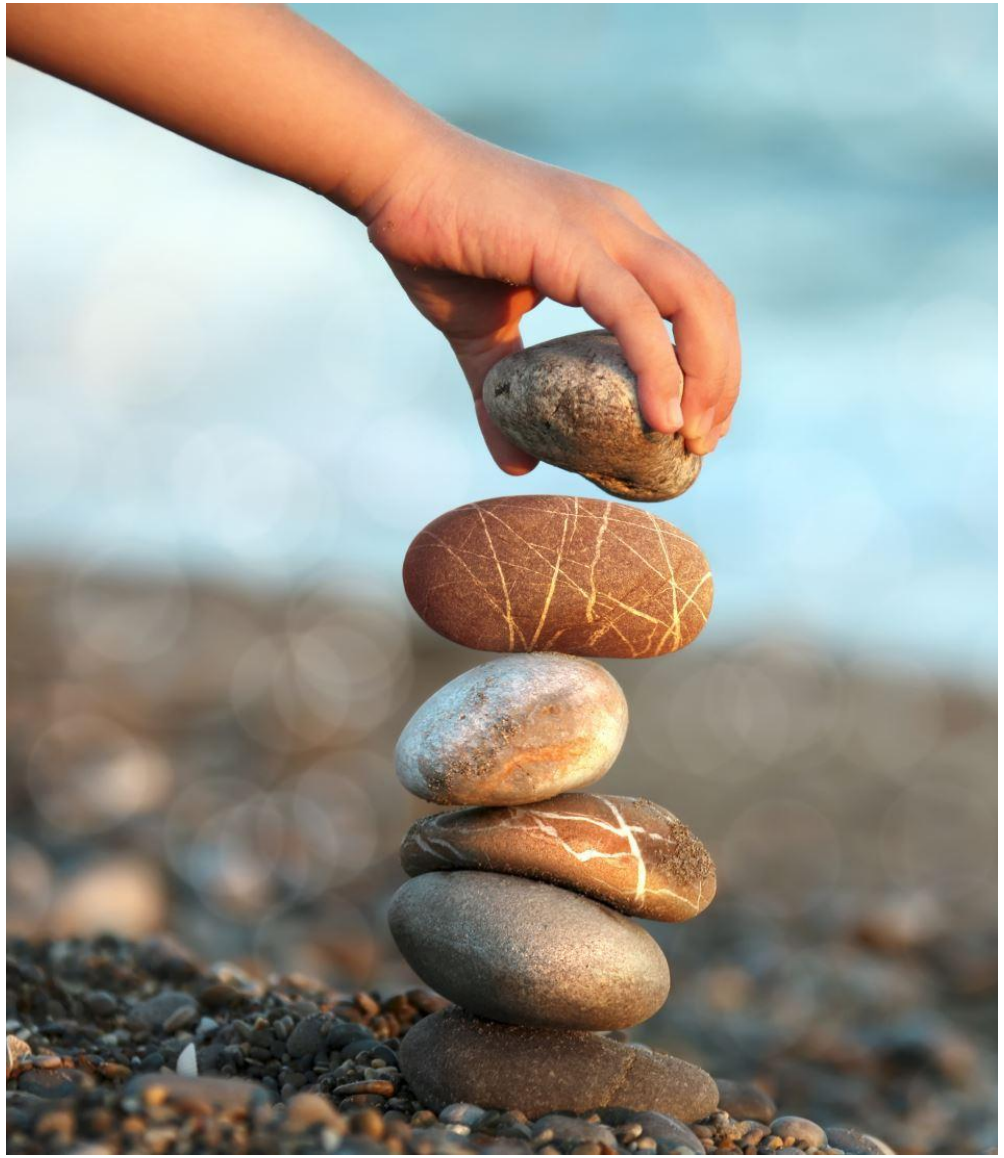
INCIDENT MANAGEMENT

Nishant Roy, Pinterest

Sep 2022



ABOUT ME

Engineering Manager @
Pinterest Ads Serving Platform

Ads + Company Incident
Manager On Call (IMOC)

Worked on 150+ incidents

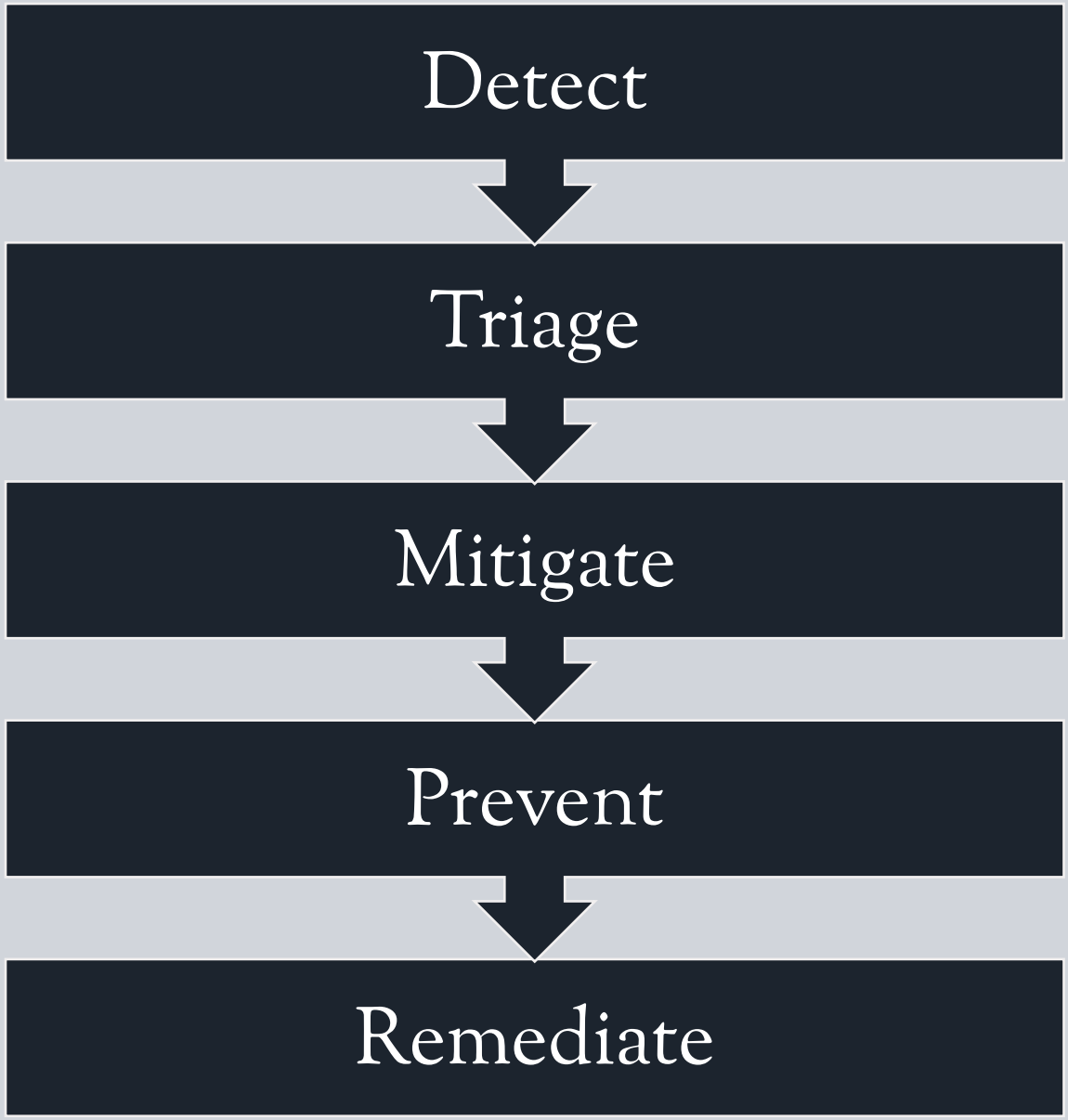
WHAT IS AN INCIDENT?

An incident is any event that is not part of the standard operation of a service and that causes, or may cause, an interruption in service or a reduction in its quality

WHAT IS AN INCIDENT (REALLY)?

An incident is when we need one or more people to drop everything else and fix the bleeding immediately

WHAT IS INCIDENT MANAGEMENT?



DECLARING INCIDENTS

WHEN?

WHO?

HOW?

DEFINING SEVERITY LEVELS

All incidents are not the same

Every component needs custom criteria

General framework: how many people need to respond or need to know?

INCIDENT RESPONSE

What to do when an incident is declared?

COMMUNICATION CHANNELS

Chat room / channel

Video conference and/or physical room

Investigation document *(optional)*

DEFINE ROLES

Incident Runner: Responsible for the outcome of incident response

Incident Manager: Responsible for coordinating incident response

IMPACT AND SEVERITY ASSESSMENT

Escalation to other teams

Communication (internal/external)

Resolution process (hotfix/rollback)

PRIORITY #1: STOP THE BLEEDING!

Root cause can wait, first treat the symptom

- Rollback suspicious changes
- Patch to alleviate symptoms

Update incident status to mitigated

NEXT: RESOLVE ROOT CAUSE

Ensure it won’t happen again for at least a few weeks

Restore all systems to regular operations

Final update to internal / external parties

INCIDENT POSTMORTEM

What to do after an incident is resolved?

DEFINE A POSTMORTEM PROCESS

Goal: Make our systems more resilient

Postmortem document template

In-person review

SLAs for postmortem process and remediation items

POSTMORTEM DOCUMENT

Impact estimate

Root cause

Detailed timeline

Time to detect, mitigate, resolve

Remediation items to improve the above metrics

BLAMELESS POSTMORTEMS

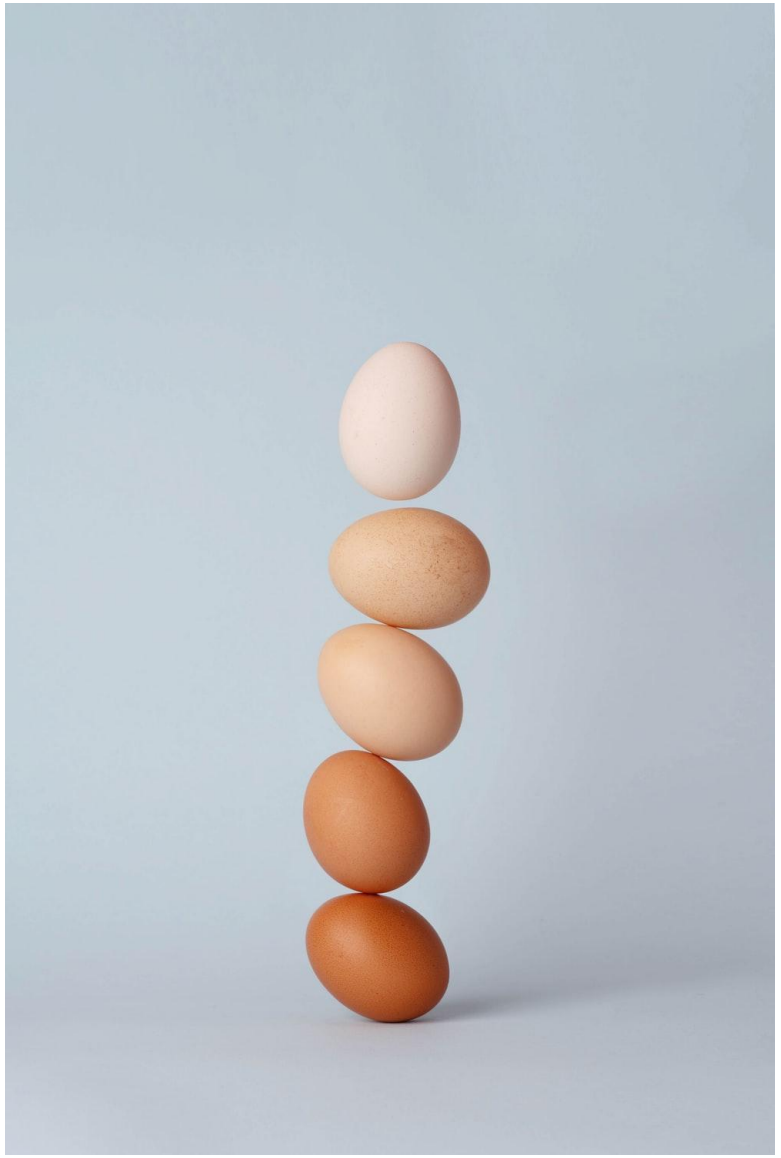
Focus on **WHAT** and **WHY** not who

Humans make mistakes. Systems need to handle them

5 WHY'S

- What? User can't login to their accounts
- Why? The API is rejecting the login request
- Why? The API cannot talk to the AuthN service
- Why? The API does not have the right SSL cert
- Why? Someone copied the wrong config to prod
- Why? There's no validation for the API config

NISHANT'S LESSONS LEARNED



DESTIGMATIZE INCIDENTS

Incidents are a learning opportunity

Celebrate incident responders

False positive >>> False negative

INCIDENT MANAGER:

BE CONFIDENT

Guiding principle: Minimize risk and drive resolution

Example 1: Should we turn off ads entirely or show irrelevant ads for a period?

Example 2: Rollback, fix forward, or hotfix?

Example 3: Do we need a new incident runner?

Example 4: Should we pause investigation till business hours?

INCIDENT MANAGER:

ASK FOR HELP

Rely on your incident runner and/or the subject matter expert

Loop in other managers to help with multiple incidents

MAKE A LIST OF KEY CONTACTS

Who can make the hard decisions?

Who to reach out to for external support?

MEASURE
ITERATE
MEASURE

Goal: Reduce downtime

MTTR: Mean Time To Recovery

MTBF: Mean Time Between Failures

MTTD: Mean Time To Detection

USE ERROR BUDGETS

Max amount of time that a technical system can fail without breaking SLA

(SLA = 99.9%, Error Budget = 8h46m12s)

Used to trade off innovation vs KTLO work

CONCLUSION

What is an incident?

How do we respond to one?

How do we learn from them?

How to encourage our teams to value and prioritize this work?

THANK YOU!

nroy@pinterest.com