Incidents & Incident Management

Incidents impact the business of an organization

An Incident Management Procedure keeps the Incident Impact at accepted levels

Incident Definition

Incidents, causing outages or downtime, impact the business of the organization:

- Service Level Agreements might come under pressure.
- Business reputation goes down, new business becomes more difficult
- Customer lose confidence, renewals become harder

Incident Examples

- A customer is unable to login and cannot start new actions or retrieve reports
- Report results are out of sync with operational data
- The API call processing time slows down and impacts customer’s systems
- E-mail deliveries to *@gmail.com slows down down by 30%
- A subset of appliances does not receive new updates
- A disk subsystem fails and causes a database to go down
- A customer’s account is breached and used for spamming
- A customer finds a Cross Site Scripting vulnerability

Historic Incident Links

- Cloudflare
- Wikimedia
- Domain at Marketo - http://bit.ly/wk-inc-1
- Disk at Cheetah Digital - http://bit.ly/wk-inc-2

Incident Management

We need a process to manage Incidents.

Incident Process

- Tracking Mechanism
- Conference Call
- Incident Commander
  - Overall command, communications to stakeholders
  - Uses the Map of Architecture and Applications and associated SMEs
- SMEs
  - SRE, Application engineers, Networking, Security, etc - on demand
  - Aware that quick, priority response is vital
- Incident Communication

Incident Communication

- Incident Commander manages
- Technical Channel for Incident - high volume, log like
- Communication Channel - every 30 minutes status
  - Application, Geography, Customer impact, Severity in Description
  - Receives periodic updates from Incident Commander - every 30 minutes
  - Tip: include “root cause: unknown” in each update, as the question for a root cause is quite natural and unlikely to be known
- Meta Channel - low volume, notifies interested parties of new incidents and their dedicated channel

- Channel: could be Slack, IRC, Teams, etc

Incident Priorities - Example

1	Critical - high impact	A customer-facing service is down for all customers Confidentiality or privacy is breached Customer data loss
2	Major	A customer-facing service is unavailable for a subset of customers Core functionality is significantly impacted
3	Minor	A minor inconvenience to customers, workaround available Usable performance degradation

- Severity vs Priority = Impact vs Urgency
- Often aligned, start with **Priority** for Business Clarity

Mock SME Application Map

- SRE
  - Joana: UTC-8 (+1…), Hans: UTC+1 (+49…) , Joao: UTC+6:30 (+853…)
- SRE Network
  - Jean: UTC-5 (+1…), Jack: UTC (+44…)
- SRE Database
- SWE Mailing Application
- SWE Membership Application

Scorecard Examples

- Determine root cause and possible improvements
  - Might involve Dev, QA
- 5 Whys
- Did the incident process work?
- Wikimedia example
- Quarterly Analysis
  - Events vs Incidents
  - Severities
  - Mean Time to assemble
  - Single Points of Failure
  - Service Level Objectives/Agreements violated?

OnCall Payment

Documents the seriousness of Incidents. Example:

- Level 1 - daytime 500 USD/week
- Level 2 - escalation Level 1 and nighttime 500 USD/week
- Level 3 - escalation Level 2/
- IC - 500/week

Examples: Intercom, Uber, Google

OnCall Payment - Tiers/Payment @Uber

- Tier 0: critical service powering most key services. Oncall needs to acknowledge in less than 5 minutes and includes dedicated oncall with short rotation, all paid.
- Tier 1: key service powering a core flow directly or indirectly. Oncall needs to acknowledge in 10 minutes and is paid.
- Tier 2: a service powering a user-facing experience. Oncall needs to acknowledge in 30 minutes and is not paid.
- Tier 3: a service powering a non-user facing experience. Oncall is “best effort” and not paid.

Post-mortem Examples

- Disk Array Crash, Application Database downtime, Disk Array Reboot, lengthy consistency check
- Validity: yes
- Detection: flood of alerts, confusing, customer complaints, reproducible
- TTA: 30 minutes
- Analysis: slow, troubleshooting tools impacted by disk array crash, 3rd party involved
- Why: disk overloaded, application traffic growth, disk older, no capacity planning, no budget for disk renewal
- Results: 3rd party process included, Improved 3rd party Support, Capacity planning, Independent Disk Usage monitoring, Data Retention policies
- Did the incident process work? Yes

Glossary

- IC - Incident Commander
- IMS - Incident Management System
- SME - Subject Matter Expert
- SLA/SLO/SLI - Service Level Agreement, Service Level Objective, Service Level Indicator