How to run a disaster day

Overview

- Prepare a scenario
- Introduce the team to the exercise
  - Introduce the team to incident response
  - Introduce the team to post-mortems and remind to create a timeline already during the incident response.
  - Introduce the team to the scenario: expectations, rules, scope, goal
- Trigger the incident
- The team responds and mitigates the incident
- The team collaboratively writes a post-mortem and presents it to the facilitator
- Collect any follow-up action items, collect feedback from the participants

Preparation

In order to conduct a succesfull disaster day, you must gain buy-in from relevant stakeholders.

- Obtain permission to break the environment
  - If the simulated incident involves losing data, have a backup or way to recreate the data and obtain explicit permission for deleting data.
  - Remember to consider all possible stakeholders who will be impacted by the downtime: QA, sales, other teams that may receive alerts etc.
- Agree on a scope for the exercise: what infrastructure is included
- Discuss any special requirements, concerns or ideas
- Schedule the session: dedicate at least half a day for the exercise.

On-site or remote?

Disaster day can be run on-site or remotely. If run on-site:

- Reserve a room or dedicated workspace for the team. Things may get noisy
- Allow other stakeholders to observe, but not interfere with the incident response. Ideally, these stakeholders will act as if they would during a real incident.

If runnign remotely / virtually:

- Establish communication channels, for example a shared video call for the team, beforehand.
- Consider setting up key roles beforehand for easier communication: assign a coordinator and a timeline-keeper.

Preparing a scenario

You need to come up with a simulated incident for the exercise. The team should not know the incident beforehand. The incident should

- Be within the agreed scope
- Not known to the team beforehand
- Not be trivially fixable
    - For example breaking infrastructure provisioned with IaC may not be as interesting as breaking a manually created infrastructure dependency
- Be something you know how to fix
- Ideally be triggered or triggerable via a realistic chain of events. An "oopsie" is a realistic event

Incident ideas

- Delete / remove some infrastructure resources, for example an AKS cluster
- Break some network links or other networking infrastructure like DNS
- Delete or otherwise a cause a container image to go missing
- Application misconfiguration
- Look for ideas and inspiration in existing post-mortems

Schedule the session

- Based on past experience, responsing to a small to medium sized incident takes roughly 1-2h, and writing and analyzing a post-mortem around 1h. The introduction may take 15min-1h, depending on the amount of material covered.
- In total, a good rule of thumb is to budget half a day for a single exercise. Larger exercises, for example if a complex environment needs to be recreated, may take longer.
- It is possible to run multiple scenarios during a single work day if the team wants to double down on incident response. However, also consider the option of instead running two half-day exercixes spread out with at least a srint between them, this allows the team to implement some of the action items from the first exercise before the second incident.

Running the Disaster Day session

Introduce the team to incident response

If the team does not have incident response experience:

- Ask the team how they would respond to an incident
- Introduce the team to the incident response workflow:
  - triage: figure out what is broken and what the impact is, apply any immediate emergency options.
  - examine: find out what exactly is going on
  - diagnose: form a hypothesis of what the problem is
  - test & treat: try to validate the hypothesis by conducting an experiment
  - fix: assess and possible implement a more permanent fix
- Introduce the team to keeping track of the timeline of the incident and writing notes of any actions that were taken.
  - Possibly ask the team to select a single person to be responsible for keeping track of the timeline

Introduce the team to post mortems

If the team does not have an established post-mortem process

- Ask the team about any previous experiences with post-mortems
- Introduce the team to the most important post-mortem concepts. If the organization uses a post-mortem template, make use of that.
- Make sure to explain and emphasize a no-blame post-mortem culture
- Explain to the team why post-mortems are useful: they act as a collective record and a retrospective of the incident.
- Explain to the team that they will collectively write a post-mortem on the incident, with a focus on the lessons learned. The post-mortem is here used as a tool to reflect on the incident and gather learnings.

Introduce the team to the scenario

- Explain to the team the scope of the exercise (which infrastructure is in scope)
- The team is expected to collectively respond to the incident and write a post-mortem.
- The teams immediate goals are:
    i. Understand the impact of the incident
    ii. Restore the service to users
    iii. Find and address the root cause

Trigger & respond

Now it is time to trigger the incident

- Ideally an alerting system will detect the incident, otherwise role-play a call from a customer or otherwise tell the team that something is wrong with the system
- Monitor the teams progress. Allow the team to go off tracks and hunt down red herrings.
    - In case the team steers too far outside of the scope or is about to perform a truly dangerous operation, be ready to step in. Otherwise give the team space.
- Conside role-playing a customer service representative: ask updates on the incident regularly to create a sense of urgency and to validate that the team is working towards understanding the impact of the incident and restoring services

Finding the root cause

- Once the incident has been mitigated, encourage the team to find the root cause
- Each time the team thinks they have found the root cause, ask "why"
- Once the team can no longer go deeper, the root cause is found. The root cause may not be conclusive.

Writing & presenting the post-mortem, wrap up

- Once the team has finished the exercise, ask them to write the post-mortem on a whiteboard or shared document and then present it
- Focus on the lessons learned and facilitate a discussion to identify action items
- Once the action items have been gathered, remember to congratulate the team, they hopefully have had an intense day!