QPP Incident Response Plan

Quick Links

How to report an incident:

Starting the Incident Process

Incident Response Teams (this has the email addresses for each PagerDuty alert)

Sample Incident Response Template

Help! What do I do?

- I'm starting the Incident Process
- I'm noticing an issue and I think it might be an incident
- I'm the Incident Commander and I'm starting the Incident Process
- I'm the Incident Commander for team A and I want to hand off my duties to another team B
- I'm responding to an alert via PagerDuty
- I'm a member of the SI responding to an incident report from the service center
- I'm the Incident Commander and I'm closing an incident.
- I've been e-mailed PII or someone posted it in the Google Group

Table Of Contents

Child Pages

- Incident Response Teams
- PagerDuty - Escalation Policies
- Sample Incident Response Template
- What to do when someone emails you PII or posts it to the Google Group

Overview

This document outlines the process all QPP teams must follow during an incident.

Summary

- Anyone may open an incident
- "Incident" is defined as an issue that requires **immediate** corrective action
- PagerDuty is the preferred mechanism for opening an incident
- Response teams must designate a single point of contact (called the Incident Commander) for all external communication, which happens primarily via an incident response document
- The Incident Commander must announce all incidents in the QPP Incidents HipChat room, and all real-time response coordination chat must take place in that room.
- Response teams will attempt to mitigate the issue as fast as possible before attempting to determine the root cause
- Response teams work on resolving the incident until they and the affected business owners are satisfied and declare the incident closed. The Incident Commander is ultimately responsible for making this determination.
- After every incident, teams must complete a root cause analysis and take steps to prevent a recurrence.

What is an “incident”?

Software breaks all the time. Sometimes it breaks and is still usable; other times it breaks and systems grind to a halt. For the purpose of this document, an incident is defined as the deviation of a system from its normal operation such that immediate corrective action is required.

Examples of incidents include: a production web service cannot connect to its database and returns an error on every request; a software license expires causing a system to stop responding; a new release suffers from a remotely-exploitable security vulnerability; private user information is available to unauthenticated users on the Internet.

Examples of non-incidents include: errors in development or implementation environments; a UI bug that does not affect a large number of users (escalate to Product Owner for prioritization) ; intermittent performance degradation.

In general terms, an incident:

- requires an immediate response

- may require cross-team collaboration
- disrupts the activities or privacy of end users
- has a significant impact on normal business operations

When in doubt - treat every issue as an active incident until proven otherwise.

How to report an incident

If you have encountered an incident, send an e-mail to PagerDuty as described here: Starting the Incident Process

Anyone may report an incident. Early warning is crucial for mitigating the harm an incident may cause, and for quickly returning to normal operations.

The person reporting an incident must use PagerDuty to report the incident. PagerDuty is used by all QPP teams to manage on-call response teams, and using it to alert the responsible team will ensure the fastest, most reliable response. PagerDuty allows for tracking the response and escalation, and integrates well with other tools QPP teams use (i.e. Slack, HipChat, and New Relic).

If it is during business hours and you want to ask a question about whether or not something is an incident, follow this process: I've noticed an issue and I think it might be an incident.

If you don't get a prompt response to your question or it's not during business hours, treat the issue as an incident and send an e-mail to PagerDuty as described here.

Don't be hesitant to report something you are unsure about. Reporting an incident when you are unsure is always a good idea and the program will thank you for it, even if it turns out to be a false positive. Conversely, failing to correctly report an issue because you were unsure if it was an incident could cause serious harm to our users and the product.

If the incident involves PII, it is considered a security incident and must be reported to the CMS IT Service Help Desk within one hour of identification.

Positive example of reporting an incident

```
To: example@cms-qpp.pagerduty.com

From: reporter@cms.hhs.gov

Subject: many errors on Auth service

Starting at 23:50pm, we’ve seen a massive increase in errors logged in the Auth service logs in Splunk.

Here’s a link to a search that shows the errors: https://splunk.example.com/...

New Relic is also showing errors in the same timeframe: https://rpm.newrelic.com/example-url/...

Please acknowledge and advise on next steps
```

Notes:

- The email is to PagerDuty, so it will immediately alert the current on-call engineer.
- The body of the email has enough substance, including evidence, for the responder to understand the scope and severity of the issue.

Negative example of reporting an incident

```
To: developer@qppcontractor.com, other-developer@qppcontractor.com, unrelated-developer@qppcontractor.com

From: reporter@cms.hhs.gov

Subject: Auth is not forwarding Splunk logs

John Sysadmin just called me and said that Auth Production systems are not forwarding Splunk
logs.  Please contact him as soon as possible to resolved (sic) this issue.  His phone number is 419-555-1234.

Thanks,

Jeff Sysadmin
```

Notes:

- This is an actual incident report from WDS Team during Open Enrollment, December 2016. The email was shunted to “Low Priority” by Google Inbox, delaying a response.
- The email was sent to 3 potential responders, who may assume someone else will respond and therefore ignore the email.
- The email doesn’t offer critical details.

Teams are advised to not report incidents by:

- sending an email to a distribution list
- calling, texting, or emailing an individual
- demanding a person or team join a call or webconference

Using these channels will delay the initial response, and in the worst case, may cause the initial incident report to be ignored.

Each QPP team is responsible for maintaining an up-to-date on-call schedule in PagerDuty.

Responding to an alert

Not all PagerDuty alerts are necessarily incidents. **If you are currently responding to a PagerDuty alert, follow this script: I'm responding to an alert via PagerDuty**

The goals of the alert responder are to, in order:

1. Determine if the issue is an incident, and if so, start the incident response process
2. If the issue is not an incident, document the issue in JIRA and if applicable, the resolution as well

In general, if an alert responder is not sure how to handle a situation, they should escalate the alert in PagerDuty

The response time goal from initial notification of a standard automated alert is 15 minutes. For alerts that are likely to indicate an incident, including reports directly to incident@qpp-cms.pagerduty.com, the response time goal is 5 minutes. PagerDuty schedules should automatically escalate unacknowledged issues after the response time goal period has elapsed.

System-wide alerts

If an alert is not specific to a particular team, for instance, if it originates from the service center or from a generic site-wide alert, an SI on-call responder will receive the notification.

If you are an SI on-call responder responding to an alert, follow this script: I'm a member of the SI responding to an incident report from the service center

If the situation is clearly an incident (e.g. the site is down), they should immediately start the incident response process and alert two engineers to the incident. Otherwise, they should alert the appropriate team for triage.

Responding to an incident

When an alert or report has been confirmed as an incident, the primary goal is to minimize the disruption caused by the incident, then the restoration of normal business operations.

If you have just discovered an incident, follow this script: Starting the Incident Process

If you are the Incident Commander for a new incident, follow this Script: I'm the Incident Commander and I'm starting the Incident Process

The first steps of the incident process will determine a single **Incident Commander**, responsible for managing the incident response and communicating the current state of the response. This person is also the single point of contact for all inbound inquiries from other entities, such as business owners. The engineers who are responding to the incident (typically gathered by the Incident Commander), and are working on mitigating or triaging the incident, are referred to as the **Incident Team**.

Each development team on QPP should maintain a rotation of people who are trained in the process of being an Incident Commander and are comfortable executing the incident process. These are the people who will receive the alerts from their team's incident response PagerDuty.

A good candidate for an Incident Commander is someone who:

- Can communicate in a clear, concise manner
- Understands the basic technical principles involved in the incident and response
- Has access to communication channels, including CMS HipChat and email
- Can devote their entire time and attention to managing the response and managing external entities

The Incident Commander is also responsible for handling communication with any necessary external sources (contacts are listed here: Production Monitoring and Contacts). They should also ensure that there are no blockers preventing the Incident Team from mitigating the issue as quickly as possible, and should take immediate action to remove any blockers that should arise.

If the Incident Commander encounters an issue, dispute, or blocker that they cannot resolve, they should immediately escalate to their Product Owner and the QPP CTO. This is true even if the incident occurs in the middle of the night or on the weekend. The Incident Commander should already have the appropriate contact information to reach their Product Owner and the QPP CTO.

During the response

Effective incident response is predicated on:

- Allowing engineers time and space to investigate the issue and implement a fix
- Real-time communication of status

The engineers on the Incident Team must be focused on diagnosing the issue and developing a solution. The initial solution may be a code change, adjusting operational processes (i.e. toggling a feature flag, throttling a service), or reverting to a known good release. Response teams are encouraged to provide incremental resolution, rather than wait for a “big bang” fix to be deployed.

We should follow these rules for any incident response:

1. Do no harm. When responding to an incident, don't do anything that is going make the problem worse or create new problems.
2. Fix as much as possible as quickly as possible. If you can get 60% back in 20 minutes, then you have bought yourself time to get the other 40% back.
3. Communicate as much as possible as often as possible to those who are impacted, without breaking rules 1 and 2.

It is imperative that the engineers not be subject to external interruptions. The Incident Commander must act as a proxy between the response team and all external entities.

If an incident spans many hours, it is necessary to rotate teams in order to avoid errors caused by fatigue or stress. If a response team rotation occurs, the Incident Commander is responsible for briefing the new team on status, and for broadcasting the changeover to external entities.

Engineers responding to an issue should prioritize work in this order:

1. Triage and mitigate the issue. In short, make the system work as well as it can under the circumstances, as fast as possible. (see a longer note on triage and mitigation below)
2. Determine the extent of the damage. This is needed for root cause analysis but also for later auditing/postmortems and potentially other damage control activity (e.g. notifying affected parties after a PII leak). It is also a double check that there are no other parts of the system that were affected that were missed in the triage and mitigation step
3. Assess and stabilize the state of the system, and determine a long term plan. Basically, evaluate any short-term fixes in place and any system degradation that is still happening, and ensure that it is sufficient for the period of time needed to put the long term fix into place. If not, find a more stable short-term solution that will enable us to get to the long term. It is possible but uncommon (for a system of this size) that you will need to do this several times on a longer timescale.

A quote on the first step above, triage and mitigation, from the Google SRE book:

> Your first response in a major outage may be to start troubleshooting and try to find a root cause as quickly as possible. Ignore that instinct!
>
> Instead, your course of action should be to make the system work as well as it can under the circumstances. This may entail emergency options, such as diverting traffic from a broken cluster to others that are still working, dropping traffic wholesale to prevent a cascading failure, or disabling subsystems to lighten the load. Stopping the bleeding should be your first priority; you aren’t helping your users if the system dies while you’re root-causing. Of course, an emphasis on rapid triage doesn’t preclude taking steps to preserve evidence of what’s going wrong, such as logs, to help with subsequent root-cause analysis.
>
> Novice pilots are taught that their first responsibility in an emergency is to fly the airplane [Gaw09]; troubleshooting is secondary to getting the plane and everyone on it safely onto the ground. This approach is also applicable to computer systems: for example, if a bug is leading to possibly unrecoverable data corruption, freezing the system to prevent further failure may be better than letting this behavior continue.
>
> This realization is often quite unsettling and counterintuitive for new SREs, particularly those whose prior experience was in product development organizations.

Communication methods

The Incident Response Document is the primary channel for communicating incident status. By maintaining a live document with a running log of actions the team has taken, external parties can quickly learn the current status without interrupting the response team. The Incident Response Document must be accessible to all parties that are affected by the incident; Confluence is the preferred method for publishing the document.

During the incident response, the Incident Commander must keep the Incident Response Document up to date with all activities. The Incident Commander should update the document on a regular interval, at least once every 2 hours, even if only to note that no new information is available. The Incident Commander may also make regular updates via email or HipChat to external entities on a mutually agreed upon schedule.

The Incident Commander will use the “QPP Incidents” HipChat room to host real-time chat between the various teams involved in the response. In general, HipChat should be preferred to conference calls, as it’s faster and simpler to see the chat history, and to see who is involved in the response.

Closing an incident

An incident is deemed to be resolved when the system has returned to normal operations. In general, both the response team and the affected business owners should reach consensus on the correctness of the resolution before declaring the incident to be over. The Incident Commander is responsible for gathering the affected parties to reach this consensus, and finding a path to resolution for any disputes. An effective method for determining if an incident is closed is to announce: “Is anyone opposed to me closing this incident?” As long as anyone has a valid reason to keep the incident open, it remains open.

The response team must remain available and dedicated to resolving the incident as long as the incident has not been declared closed. In general terms, this means that the Incident Commander is available and responsive to all external entities, typically responding to inquiries within 15 minutes. The response team should be dedicated entirely to resolving the incident, and should not be performing any work other than incident response.

Likewise, business owners must also be available for the duration of the incident to approve actions necessary for the response, such as emergency code deployments and configuration changes, and for assisting with managing the external impact of the incident. Generally, this is accomplished by business owners temporary delegating their authority to a single person in the form of the Federal Liason, but business owners are also expected to be available for specific questions during the incident.

After an incident

After the incident is declared closed, the response team must write a post mortem document, using the template here, then spend time filing bugs to track all necessary follow up work. Teams may consider adding additional alerting, tests, or larger changes as necessary.

The response team must also complete a root cause analysis (RCA) that may be shared with other teams within CMS. The RCA should clearly state what the problem was, how the team responded, and what steps they are taking to prevent the issue from recurring. The RCA should be completed within one business day of the incident.

How to write a good postmortem

- Be precise. Use precise, clear language. Refer to specific items and actions taken, not vague categories. Don't assume the reader knows to what you are referring. Don't add superfluous adjectives, which often carry judgment and can obscure analysis.
- Use simple language. Write at a 6th-grade reading level or lower. (Use the Hemingway app as a writing guide.) Substitute simpler, more direct words for complicated ones.
- Don't assign blame or use judgmental language. Incidents occur, outages happen to everyone—the important thing is how we respond to them and learn from them. The point of the document is not to fall on your sword, or drive it into others. Don't use blaming language, on yourself or others.
- Write in the 3rd-person. The document should appear to be authored by the team, not by a particular person. Don't use "I", "my", "me", "we", etc.
- Don't use the passive-voice, except to avoid assigning blame. It's important to know who was performing the action. Don't write "the logs were copied to a remote machine", instead, "Sally copied the logs to a remote machine". The exception is in lessons-learned-type sections, were the active voice could be read as assigning blame.
- Be comprehensive. Error on the side of including more information about the incident rather than less. It may not be apparent at the time of writing what information is relevant to future analyses. Add start and end times to the outage and incident (an outage typically starts before the incident starts--an incident is when the team acknowledges it as such--and ends before the incident ends--after any cleanup or resolving notifications are sent) in the timeline, and mark them clearly for ease of scanning.
- Don't use casual language. Avoid informal words and joking.
- Connect the dots. Draw connections between actions and outcomes, observations and analysis. Explain how and why the team came to conclusions. Don't leave it to the reader to piece it all together, they won't have enough context.

Credit

This document is deeply influenced by the Google Site Reliability Engineering book, especially the chapter “Managing Incidents”: https://landing.google.com/sre/book/chapters/managing-incidents.html.

Appendix

Sample Incident Response Document: see 2018-01-29 AWS Credentials Logged to Splunk

Quick Response Scripts

I'm starting the Incident Process

#	Steps
1	Do you know which team can handle this incident? If not, email incident-report@cms-qpp.pagerduty.com
2	Given that you know which team should handle the incident, email <team name>-incident-report@cms-qpp.pagerduty.com Full list here: Incident Response Teams
3	The incident commander receives the PagerDuty, clicks 'acknowledge' on the alert and runs 'I'm the Incident Commander and I'm starting the Incident Process' *Note:* For the generic 'incident-all' list, the incident commanders should discuss in the 'QPP Incidents' room in HipChat and determine who should take command before acknowledging the alert.

I'm noticing an issue and I think it might be an incident

#	Steps
1	First, do you feel fairly confident this is an incident? If so, go to **Starting the Incident Process**
2	If you are not sure but want to find out if something is an incident, post your question with an @channel in the "QPP Incidents" HipChat room.
3	In addition to step 2, if your team has designated incident commanders, seek them out and ask them specifically
4	If no one responds to your question in "QPP Incidents" HipChat within 15 minutes, you need to escalate the process by going to Starting the Incident Process
5	If the incident commander believes the issue is an incident, they will start the incident process (see Starting the Incident Process)

I'm the Incident Commander and I'm starting the Incident Process

#	Steps
1	Gather anyone already working on the incident (the Incident Team) and move all real-time coordination to the "QPP Incidents" HipChat chat room. If this is happening off hours, phone call anyone who needs to be involved in this incident. *Note:* Each team should furnish their Incident Commanders with a list of phone numbers for all team members so that they can pull in whoever is needed
2	Ensure that the Incident Team has prioritized mitigation of the issue above any other tasks and work on mitigation has begun

3	Create an incident tracking document in confluence (cloning this template: Sample Incident Response Template) as a child page of the list of incidents here: Incidents List At this point make sure to fill in the "PagerDuty Tracking Incident" in the template. Making sure the tracking incident was actually created and acknowledged will help ensure our incident response time metrics are correct. See Measurement: Incident Recovery Times. Remember to publish this page frequently throughout the Incident so that other people can see your updates.

4	Send an e-mail to inform the stakeholders listed here that there is an active incident. Include the link to the incident document.
	
	- Product Owners
	- SI team
	- Service Center
	- CTO & CPO
	
	Subject line: **"[Active Incident] {description}"**. Such as, "[Active Incident] TIN PII in C2Q Logs".
	
	Specifically,
	
	`Stan.Ostrow@cms.hhs.gov;`
	`Christopher.Reinartz@cms.hhs.gov;`
	`Mindy.Riley@cms.hhs.gov;`
	`Lauren.Erickson@cms.hhs.gov;`
	`Cate.Corradini@cms.hhs.gov;`
	`betina.fletcher@cms.hhs.gov;`
	`Ashley.Hain@cms.hhs.gov;`
	`kenneth.howard@cms.hhs.gov;`
	`matthew.leipold@cms.hhs.gov;`
	`kati.moore@cms.hhs.gov;`
	`Dach.Mou@cms.hhs.gov;`
	`Manoj.Nagelia@cms.hhs.gov;`
	`Justyna.Sardin@cms.hhs.gov;`
	`James.Woodey@cms.hhs.gov;`
	`anil.chillakuru2@cms.hhs.gov;`
	`natalie.medler@us.pwc.com;`
	`shwoods@qssinc.com;`
	`ckahler@qssinc.com;`
	`Amanda.Slay@ventech.hcqis.org;`
	`sfradkin@flexion.us;`
	`shasse@flexion.us;`
	`dknesek@flexion.us;`
	`pkendall@flexion.us;`
	`ctedrick@flexion.us;`
	`bruth@flexion.us;`
	`wsadler@flexion.us;`

5	Set up a video chat/phone bridge for coordination and make the call-in information for that phone bridge the topic of the QPP Incidents HipChat room: `Active Incident. Conference Line: [link] Incident Document: [link]`

6	**Security Incidents Only:** Assign someone who is not working on remediating the incident to fill out the attached CMS Incident form as much as possible. This form should be e-mailed to the CMS IT Service Help Desk (CMS_IT_Service_Desk@cms.hhs.gov) **within one hour of incident identification with available information.** The form may be resubmitted periodically as additional information is obtained. **Note:** any incident involving PII, whether internal or external, must be reported to the CMS IT Service Help Desk. Even if the incident is closed before the hour is complete, CMS must be notified. In addition, you can report an incident by calling the CMS IT Service Desk: 410-786-2580 or 1-800-562-1963 CMS Incident R..._Comments.docx Refer to the CMS Privacy Data Breach page for more information about responding to a security incident.
7	Continue to track ongoing work in the confluence document, and ensure that the Incident Team has no blockers
8	If you need help getting in touch with another team or external resource, the System Integrator team should be on-hand to help pull in those resources. You can also trigger that team's Incident Response PagerDuty in order to escalate the request.

I'm the Incident Commander for team A and I want to hand off my duties to another team B

#	Steps
1	If they are not already aware, get team B (and other teams, e.g. team C, team D, if necessary) aware of the incident onto the QPP Incidents channel and phone bridge (the SI can help with that)
2	All teams involved should come to a consensus on who the new incident commander will be (say from team B). If you can't come to a consensus, notify the product owners from all teams and the CTO
3	The active Incident Commander from team A should explicitly say in the QPP Incidents channel that they are handing off command of the incident to the Incident Commander from team B
4	The Incident Commander from team B should explicitly say in the QPP Incidents that they now have command

I'm responding to an alert via PagerDuty

#	Steps
1	Do you have the necessary knowledge to triage this alert? If you're not sure, escalate the alert to the next level in PagerDuty
2	Are there signs that functionality of the site for users, or users' privacy might be seriously adversely affected? If so, go to **Starting the Incident Process**
3	Do you think something might be wrong, but aren't sure if you can determine the extent of the issue within a short period of time (15 minutes)? If so, escalate the alert and continue working on the issue
4	Is this alert clearly a false positive? If so, create a JIRA ticket on your team's JIRA board to fix the alert later, and label the ticket 'qpp-incident-response'
5	Is the issue we were alerted about not an incident? • Can you fix it quickly? If so, put in a quick fix, and write a JIRA ticket quickly describing the issue and fix on your team's JIRA board, and label the ticket 'qpp-incident-response' • Otherwise, write a JIRA ticket quickly describing the issue on your team's JIRA board, and label the ticket 'qpp-incident-response'
6	Review the ticket with your product owner as soon as they are free (likely the start of the next business day if the alert is triggered off hours)

I'm a member of the SI responding to an incident report from the service center

#	Steps
1	Is this clearly an incident (is it affecting functionality of the site for a lot of users, or affecting users' privacy)? If so, go to **Starting the Incident Process** If you do not believe this is an incident, move to step 2
2	Forward the report the appropriate DevOps team via PagerDuty

I'm the Incident Commander and I'm closing an incident.

#	Steps
1	If you believe the incident is ready to be closed, confirm that you have the consensus of the response team and impacted business owners by asking “Is anyone opposed to me closing this incident?” If anyone is opposed, keep the incident open. If no one is opposed, move to step 2.
2	Shut down the video conference/phone bridge, and reset the QPP Incidents room topic to: `No current incident - Incident Response Teams: https://confluence.cms.gov/display/QPPGUIDE/Incident+Response+Teams` Mark the "PagerDuty Tracking Incident" in the Incident Document as resolved. See Measurement: Incident Recovery Times.
3	Make sure that any outstanding action items are clearly documented in the Incident Document and schedule a retrospective with everyone involved in the incident.

I've been e-mailed PII or someone posted it in the Google Group

Follow these instructions

What to do when someone emails you PII or posts it to the Google Group

---

Incident Response Flow Diagram

Incident Response Strategy V7.2.pdf