

AWS re:Invent

NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

ARC308

Improving resiliency with the correction of error process

Johnny Hanley (he/him)

AWS Well-Architected Solutions Architect
Amazon Web Services

Juan Ossa (he/him)

Enterprise Support Lead
Amazon Web Services



Agenda

1. Why is the correction of error (COE) process important?
2. What is a COE?
3. COE components
4. Example scenario
5. Demo – COE
6. Demo – Incident Manager, a capability of AWS Systems Manager
7. Cultivating a COE culture

“Everything fails, all the time.”

Dr. Werner Vogels

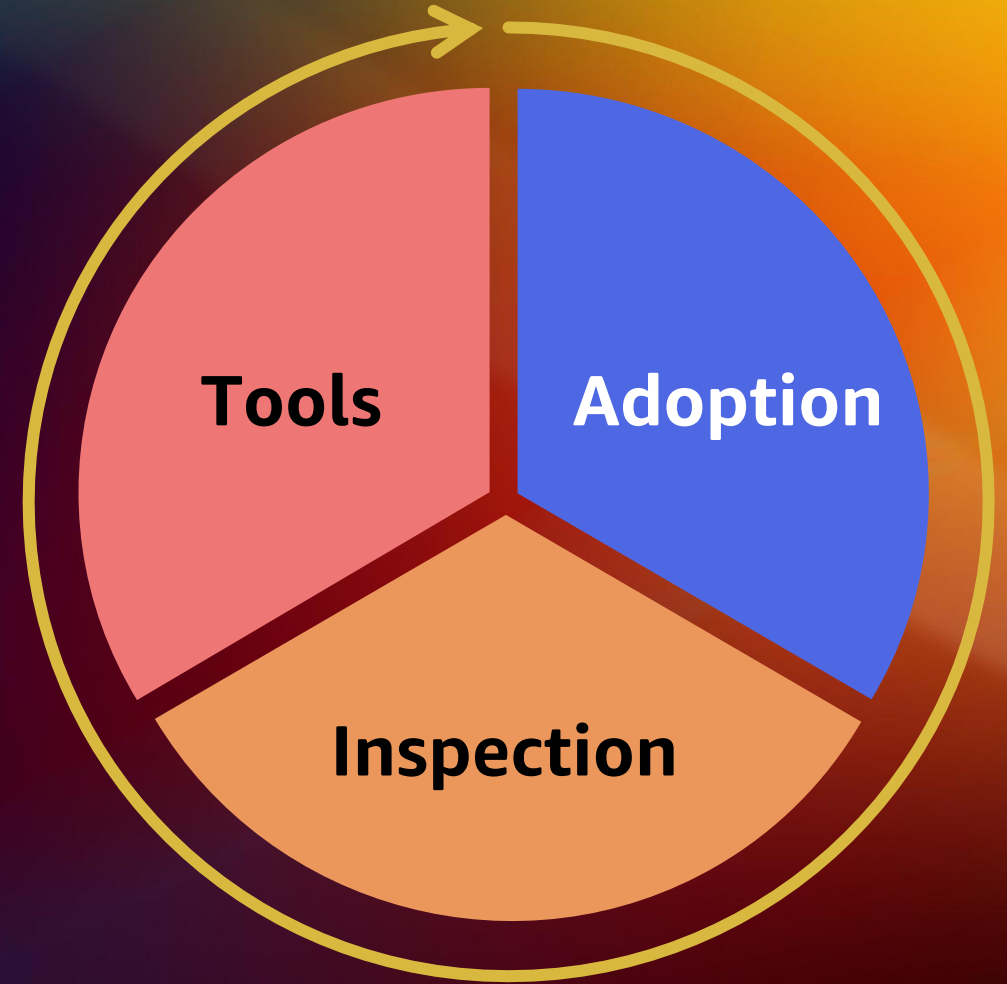
VP and CTO at Amazon.com



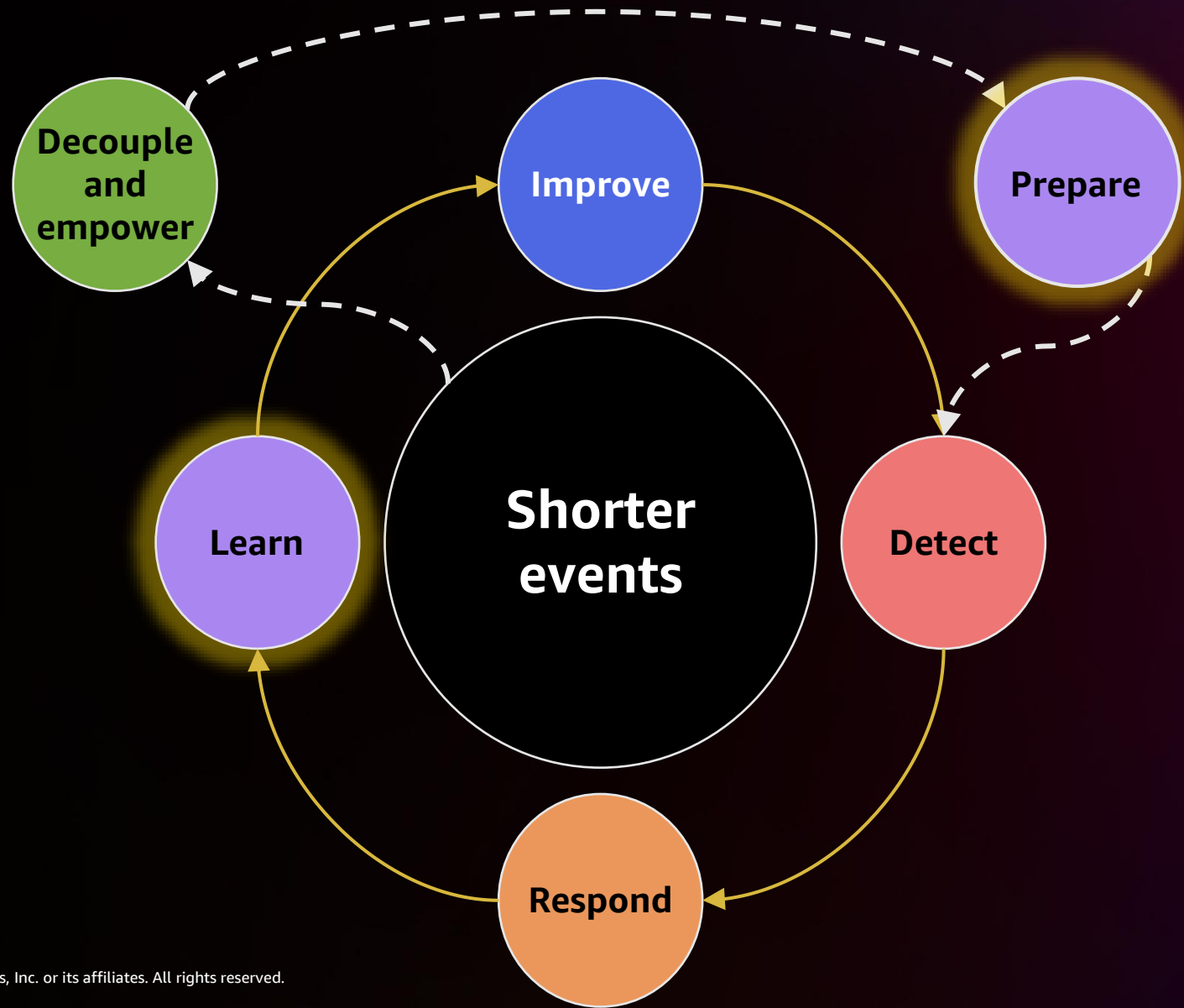
“Good intentions never work, you need **good mechanisms to make anything happen.”**

Jeff Bezos

Founder and Executive Chair of Amazon



The uptime flywheel



Why is the COE process important?

Benefits of the COE process

- Identify the root cause and remediation path
- Document and share knowledge
- Implement and measure improvements

What is a COE?



The COE document

- A **mechanism** to:
 - Identify and fix problems
 - Drive ownership of action items
 - Maximize lessons learned
 - Prevent recurrence of the problem
- **Not** to blame
- **Not** to punish

When is a COE necessary?

- Customer-impacting events
- Procedural miss (missed test case)
- Process miss (missed use case)
- Any event that reveals an opportunity for improvement

Who owns a COE?

- Beneficiaries of lessons learned
- Team where root cause belongs
- Single or multiple COEs
- Cross-team COE – area owners

General tips for writing a COE


- Gather data while it lasts
- Reference:
 - Root cause
 - Action items
 - Lessons learned
- Standardize timestamps (single time zone)
- Include links

COE components



Anatomy of a COE

- Summary
- Impact
- Timeline
- Metrics
- Event questions
- The 5 whys
- Action items
- Related items



Supporting information
(what happened)

Corrections
(learning and action)

Summary

- Write it last
- What, where, and why
- Reference the root cause
- Outline the details
 - Introduce the systems involved
 - Spell out acronyms at first mention
- Concise executive summary

Impact

A concise paragraph on the customer impact of the event

- Who was impacted?
- What was the customer experience?
- How long did it last?
- Quantify everything
- Cast a wide net

Timeline

Do

- Bullet list of essential moments
- Links where relevant
- Use consistent date/time format
 - Include time zone
 - Include year
- Highlight critical times

Don't

- Apply blame
 - Don't use names
- Expose PII
- Be verbose
- Leave major, unaccounted-for gaps

Metrics

- Show the impact and recovery
 - Lack of metrics is a valid action item
- Quantify your information
- If you use graphs:
 - Use a consistent timescale
 - Include explanations
 - Identify critical events

Event questions

Start asking questions to analyze the event and start identifying key aspects of the issue

Detection

- When did you learn there was customer impact?
- How did you learn there was customer impact?
- How can we cut the time to detection in half?

Diagnosis

- What was the underlying cause of the customer impact?
- Was an internal activity happening during the event (for example, a maintenance window)?
- How can we cut the time to diagnosis in half?

Mitigation

- When did customer impact return to pre-event levels?
- How does the system owner know that the system is properly restored?
- How did you determine where and how to mitigate the problem?
- How can we cut the time to mitigation in half?

The 5 (or more) whys

Getting at the root cause of the problem

- Identify root causes
- Build a causal chain
 - Reference action items and lessons learned
 - Identify the process failure
- Be prepared for multiple root causes or contributing causes
- Be concise
- Don't justify

Action items

Who's going to implement what fixes and when?

- Prevent recurrence
- Coordinate with owners
- Maintain a sense of urgency
- Deliver results quickly
- Owner and due date are nonnegotiable

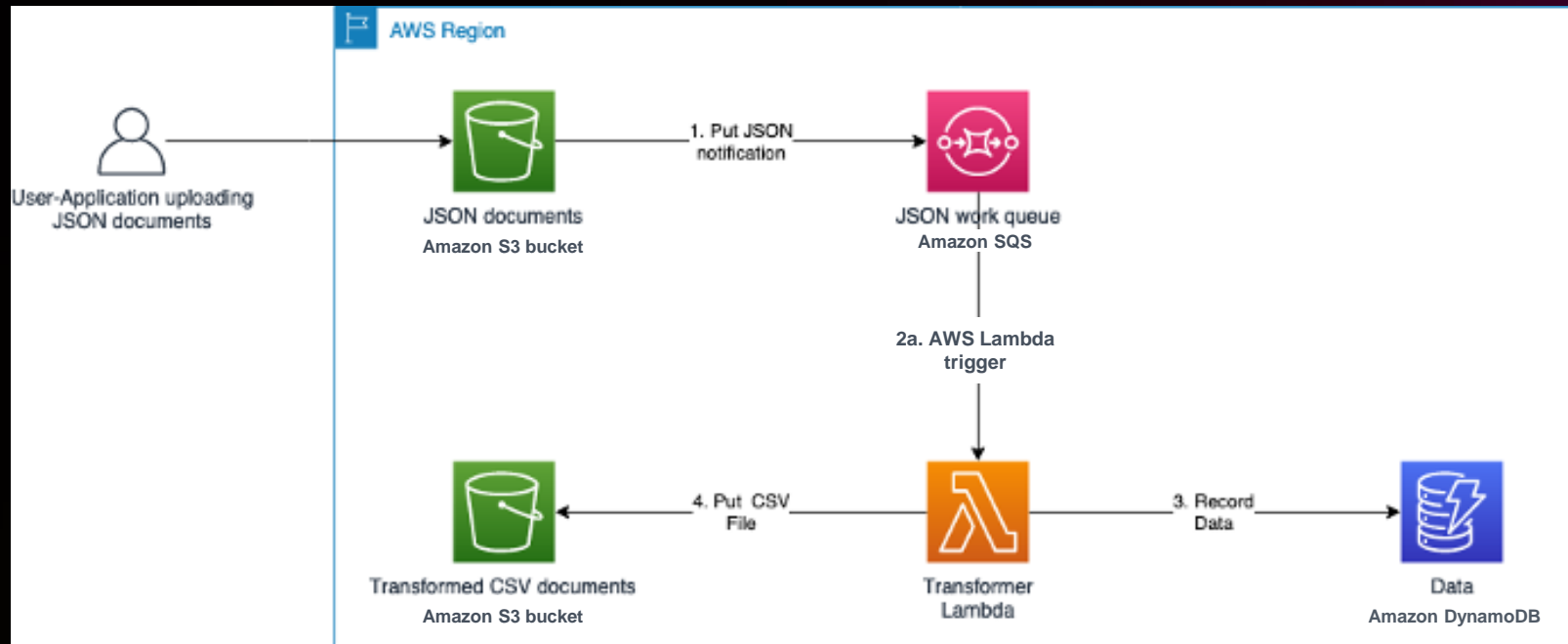
Related items

Is this case related to another COE?

- Did something similar happened before?
- Is this a series of events?
- Were there pending actions?

Example scenario

Example architecture before



Example failure scenario

Multiple customers calling the contact center indicating problems with the application

The team checked the dashboard again and realized that everything seemed to be working well; the trend was fine and the percentages seemed to be okay

Deeper investigation is needed

Example summary

Write it last



Example impact

Over 10,000 files that were processed were not successfully transformed

The customers received a message that their uploads were successful, but the data was never reflected in the application

The event started at 9:38:18 am (GMT-5) and was resolved at 11:38:24 am (GMT-5)

Example timeline

- 9:00:00 am (GMT-5) 4/1/2022 – The application is pushed to production
- 9:25:00 am (GMT-5) 4/1/2022 – Engineers verify that the dashboard metrics are as expected
- 9:38:18 am (GMT-5) 4/1/2022 – Transformer Lambda errors increase
- 9:40:00 am (GMT-5) 4/1/2022 – Call center customer complaints surge
- 9:45:00 am (GMT-5) 4/1/2022 – Call center notifies service team of customer complaints
- 9:47:00 am (GMT-5) 4/1/2022 – Engineers review dashboards, and metrics are acceptable
- 9:53:00 am (GMT-5) 4/1/2022 – Engineers broaden search to all logs
- 10:25:00 am (GMT-5) 4/1/2022 – Engineers notice increased error rate in transformer Lambda logs
- 10:45:00 am (GMT-5) 4/1/2022 – Engineers deploy a patch to the test environment
- 10:55:00 am (GMT-5) 4/1/2022 – The test environment completes successful acceptance testing
- 10:59:00 am (GMT-5) 4/1/2022 – Engineers deploy a patch to production
- 11:25:00 am (GMT-5) 4/1/2022 – The system starts to show recovery
- 11:38:24 am (GMT-5)- 4/1/2022 – The system recovery is complete

Example metrics

- Original dashboard metrics
 - Queue depth of documents to be processed
 - Percentage of documents to be processed
- Proposed additional metrics
 - Transformer Lambda error rate
 - Number of DynamoDB updates per period

Example event questions

Start asking questions to analyze the event and start identifying key aspects of the issue

Detection

- When did you learn there was customer impact?
- How did you learn there was customer impact?
- How can we cut the time to detection in half?

Diagnosis

- What was the underlying cause of the customer impact?
- Was an internal activity (for example, a maintenance window) happening during the event?
- How can we cut the time to diagnosis in half?

Mitigation

- When did customer impact return to pre-event levels?
- How does the system owner know that the system is properly restored?
- How did you determine where and how to mitigate the problem?
- How can we cut the time to mitigation in half?

Example 5 (or more) whys

The application crashed!

1. Why did the application crash?
2. Why didn't uploaded files show up in the application?
3. Why didn't the transformer Lambda update the database?
4. Why did the transformer Lambda return an execution error?
5. Why didn't the application gracefully handle invalid data types?
6. Why didn't the application know there was an error?
7. (More if necessary)

Example action items

- Update the runbook for this event – service team, 5/31/2022
- Propose new metrics – service team, 5/31/2022
 - Transformer Lambda error rate
 - Number of DynamoDB updates per period
- Update application awareness – service team, 6/31/2022
 - Add error handling into the application
 - Add input validation into the application

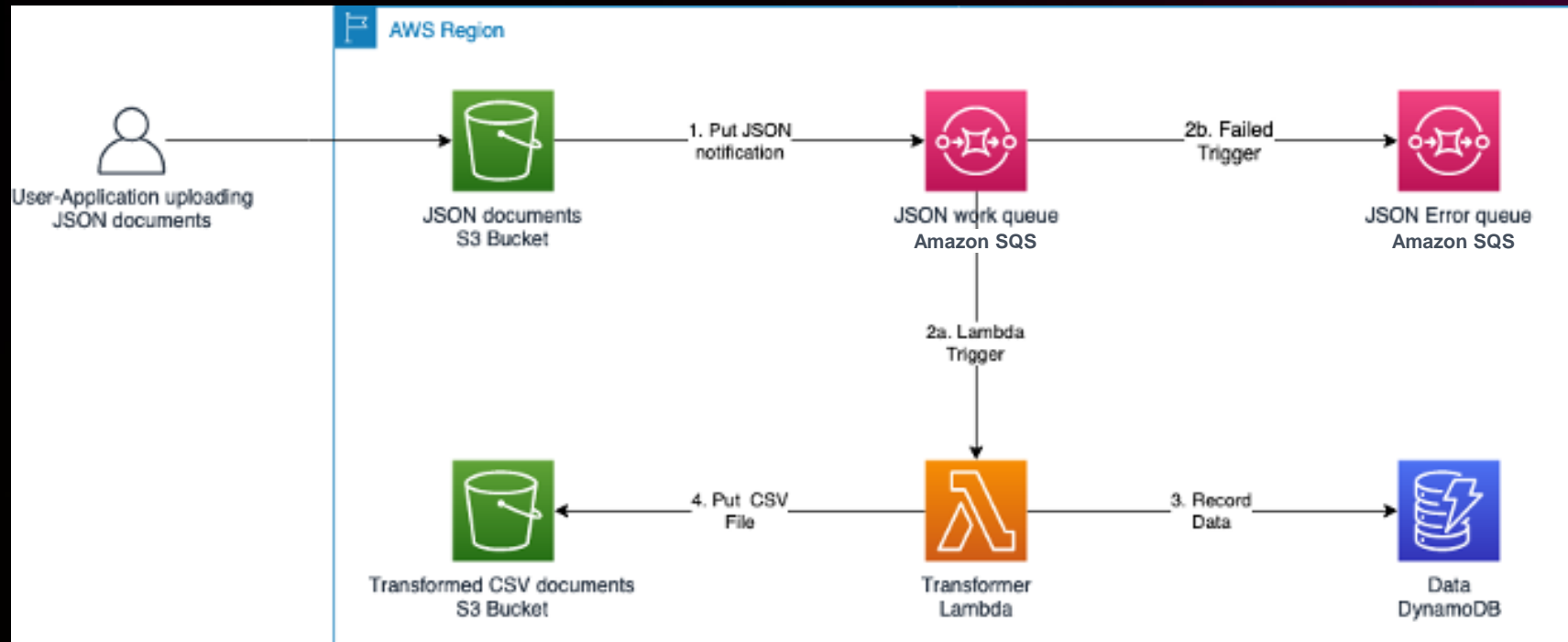
Example related items

This is the first occurrence of this issue

Example summary

The application was launched on 4/1/2022. The service team validated that all metrics were as expected. The service team received notification that customers were reporting errors. Over 10,000 files that were processed were not successfully transformed. The customers received a message that their uploads were successful, but the data was never reflected in the application. The service team began broader investigations into all the application logs to determine the cause. It was determined that the transformer Lambda was erroring due to an invalid data type. The service team developed and applied a patch to resolve the issue. Full system recovery occurred.

Application after improvements



Demo – Incident Manager, a capability of AWS Systems Manager



Cultivating a COE culture



Create a community of practice

- Establish a community for early adopters
- Foster informal sharing of lessons learned
- Standardize and document best practices
- Identify champions throughout the organization
- Train the trainer (with champions)
- Improve the process

What have you learned?



Thank you!

Johnny Hanley

<https://linkedin.com/in/johnnyhanley>
@johnnyhanley

Juan Ossa

<https://linkedin.com/in/jossaq/>



Please complete the session survey in the **mobile app**



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.