

Clos

Closed

RCA for 2018-12-21 Gitaly Outage

RCA for 2018-12-21 Gitaly Outage

Please note: if the incident relates to sensitive data, or is security related consider labeling this issue with [security](#) and mark it confidential.

Summary

A brief summary of what happened. Try to make it as executive-friendly as possible.

1. Service(s) affected : Gitaly Storage Nodes
2. Team attribution : Infrastructure
3. Minutes downtime or degradation : 35 Minutes

Impact & Metrics

Start with the following:

- What was the impact of the incident? All customer and internal requests to Git data nodes were unable to be serviced for 34 minutes.
- Who was impacted by this incident? External Customers, CI jobs
- How did the incident impact customers? See impact above
- How many attempts were made to access the impacted service/feature?
- How many customers were affected? All
- How many customers tried to access the impacted service/feature?

Include any additional metrics that are of relevance.

Provide any relevant graphs that could help understand the impact of the incident and its dynamics.

Detection & Response

Start with the following:

- How was the incident detected? PagerDuty/Slack Alerts
- Did alarming work as expected? Yes
- How long did it take from the start of the incident to its detection? Approximately 10 minutes
- How long did it take from detection to remediation? 30 minutes
- Were there any issues with the response to the incident? (i.e. bastion host used to access the service was not available, relevant team memeber wasn't page-able, ...) Yes - terraform module repos that were only on .com delayed us in running terraform to stand the compute node back up until a local copy was used.

Timeline

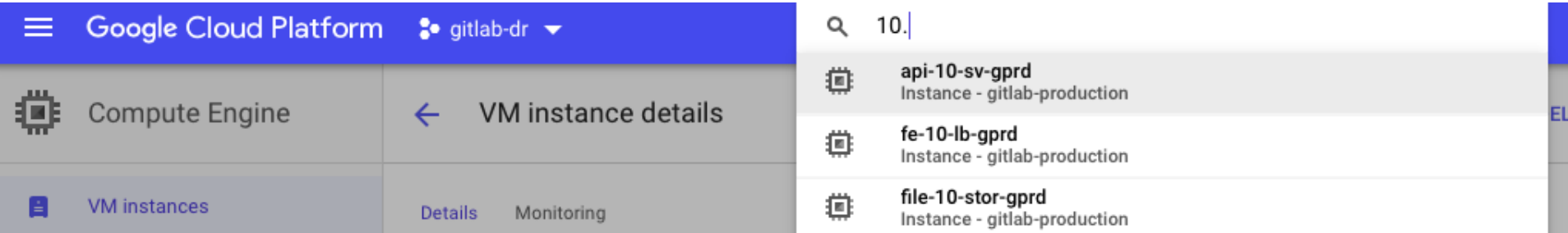
On the production issue.

Root Cause Analysis

While working on our Disaster Recovery project and region, an SRE on our team was unable to use Terraform to properly remove some nodes in the the DR project and region. They chose to go to the GCP console to perform the deletes to get the Terraform state back to good. While doing this, they searched for some nodes and per the illustration below had their project switched from gitlab-dr to gitlab-production. It was not clear that the project switch had been made and they proceeded to remove the gitaly compute instances (file-[1-24]) in the gitlab-production project. At that point, monitoring started to alert us to the problem and the team started to restore the deleted compute nodes.

Illustration

When attempting to search for something in the GCP search bar which has a partial match, one would expect that pressing enter here would execute a search for a partial IP address, which in this case returns no results.



Pressing enter results in this:

Closed

RCA for 2018-12-21 Gitaly Outage

Compute Engine

←

VM instance details

EDIT

RESET

CREATE SIMILAR

VM instances

Details

Monitoring

Which appears to have searched and found nothing. However, it did not search - the first line in the dropdown above was highlighted - so the project changed. At this point, going and deleting the DR file nodes resulted in the gitlab-production nodes being deleted rather than the gitlab-dr nodes.

What went well

Start with the following:

- Identify the things that worked well or as expected.
 - Any additional call-outs for what went particularly well.
- Quickness of team to jump on a zoom and start to mitigate the issue.
 - We were able to restore the affected infrastructure with no data loss

What can be improved

Start with the following:

- Using the root cause analysis, explain what can be improved to prevent this from happening again.
 - Is there anything that could have been done to improve the detection or time to detection?
 - Is there anything that could have been done to improve the response or time to response?
 - Is there an existing issue that would have either prevented this incident or reduced the impact?
 - Did we have any indication or beforehand knowledge that this incident might take place?
- Look at different node names for disaster recovery compute and storage node names vs production.
 - Mirror all Terraform repos (environments and modules) on ops.gitlab.net to prevent issues with access when GitLab.com is down.
 - Look at further enhancing procedures for any deletes in production requiring two sets of eyes and ways to prevent needing to do any interaction with the cloud console. Further automation to ask and double check before performing the delete.

Corrective actions

- [#5815 \(closed\)](#) : Add deletion protection for gitaly servers
- [#5816 \(closed\)](#) : Move terraform modules to the ops instance
- [#5868](#) : Start practicing incident response
- [#5867](#) : Create list of incident response scenarios
- [#5869](#) : Setup Atlantis for Terraform deployments on ops instance
- [#5945 \(closed\)](#) : Change to TF process to prevent conflicts that sent us to Cloud console

Guidelines

- [Blameless RCA Guideline](#)
- [5 whys](#)

Edited 1 year ago by [David Smith](#)

Linked issues ⓘ

8

Relates to

Gitaly service down

production#632

Move terraform modules to the ops instance

#5816

AS Team 201...

2

Enable deletion protection flag for gitaly,,pages and share

#5815

Completed 2...

2

Color code GCP projects in the console

#5848

Completed 2...

Create Cadence to practice Incident response

https://gitlab.com/gitlab-com/gl-infra/infrastructure/-/issues/5813

2/5

#5868



Closed RCA for 2018-12-21 Gitaly Outage

Ensure integrity of terraform changes before apply/merge

#5869

5

Share learnings about Terraform changes from Dec 21 RCA

#5945

DS Team 201... Jan 18, 2019 2



David Smith @dawsmith marked this issue as related to [production#632 \(closed\)](#), 1 year ago



David Smith @dawsmith mentioned in issue [production#632 \(closed\)](#), 1 year ago



Devin Sylva @devin changed the description 1 year ago



David Smith @dawsmith changed the description 1 year ago



Alex Hanselka @ahanselka · 1 year ago

Owner

The RCA text says:

It was not clear that the project switch had been made and they proceeded to remove the Git compute nodes in the gitlab-production project.

However, the issue title says "gitaly". Was it gitaly (storage) nodes, git compute nodes, or both?



David Smith @dawsmith · 1 year ago

Owner

The gitaly file-[1-24] compute instances - will update text



David Smith @dawsmith changed the description 1 year ago



John Jarvis @jarv changed the description 1 year ago



John Jarvis @jarv · 1 year ago

Owner

When attempting to search for something in the GCP search bar which has a partial match, one would expect that pressing enter here would execute a search for a partial IP address, which in this case returns no results.

I think the issue here is that the search is global across projects, not sure if this is something we can control or not :(The first item highlighted was a server in the gitlab-production project so the project was updated.



John Jarvis @jarv · 1 year ago

Owner

Added a corrective action [#5815 \(closed\)](#) to enable deletion protection for file servers which is something we probably should have had in place previously.



John Jarvis @jarv mentioned in issue [#5816 \(closed\)](#), 1 year ago



John Jarvis @jarv · 1 year ago

Owner

Added [#5816 \(closed\)](#) to move terraform modules to the ops instance



John Jarvis @jarv changed the description 1 year ago



Amarbayer Amarsanaa @aamarsanaa mentioned in issue [#5848 \(closed\)](#), 1 year ago



Amarbayer Amarsanaa @aamarsanaa · 1 year ago

Maintainer


It was not clear that the project switch had been made...

Po Closed RCA for 2018-12-21 Gitaly Outage

projects in GCP console. (Currently it color-codes production, staging and internal since these are the ones I have access to. We can add DR to it and color-code it too).

Issue: #5848 (closed). Script: https://ops.gitlab.net/gitlab-com/gl-infra/gcp-project-color-coder

I hope this becomes useful for us.




John Jarvis @jarv · 1 year ago

Owner


Nice @aamarsanaa ! Maybe we should move this to https://gitlab.com/gitlab-com/gl-infra/infrastructure/tree/master/onboarding/browser-scripts and add it to onboarding?


Also I notice there were a bunch of alert() messages in the script, left for debugging?


Also it looks like in this case when you are searching and choose the auto-completed result the color doesn't change, I think we need to adjust it a bit.





Edited by John Jarvis 1 year ago


- 

David Smith 🌴 @dawsmith added 1 deleted label 1 year ago
- 


David Smith 🌴 @dawsmith changed milestone to %Completed 2019-01-02 1 year ago
- 

David Smith 🌴 @dawsmith added s1 label 1 year ago
- 

David Smith 🌴 @dawsmith marked this issue as related to #5816 (closed). 1 year ago
- 

David Smith 🌴 @dawsmith marked this issue as related to #5815 (closed). 1 year ago
- 

David Smith 🌴 @dawsmith marked this issue as related to #5848 (closed). 1 year ago




David Smith 🌴 @dawsmith · 1 year ago

Owner

Looking to close this out soon. One further discussion point @dsylva and I had was around what we can do to prevent issues like the kind he ran into with Terraform which forced him to go to the console. It sounded like we got ourselves a little stuck by trying to do 2 big things at once with TF - 1. The module break out 2. the Kernel updates. That left us with a less stable master.

Question for @gitlab-com/gl-infra was - should we look to any team agreements / changes to how bring in change to TF? Maybe be thoughtful on how break up bigger batches into smaller changes? I only want create/change something here if we feel it would have helped prevent the issues we had.




Amarbayer Amarsanaa @aamarsanaa · 1 year ago

Maintainer

@jarv - Nice that we already have a place where we collect these scripts! I was exactly thinking the same thing, searched for a few keywords but couldn't find the one you provided above. I have moved my script there, removed the alerts (yep - they were for debugging), and will create an issue to address the color-coding issue when project switch is made via the autocomplete search box.


- 

David Smith 🌴 @dawsmith marked this issue as related to #5868 1 year ago
- 

David Smith 🌴 @dawsmith marked this issue as related to #5867 1 year ago
- 

David Smith 🌴 @dawsmith changed the description 1 year ago
- 

Craig Barrett 🌴 @craig mentioned in issue #5869 1 year ago
- 

Craig Barrett 🌴 @craig changed the description 1 year ago
- 

Craig Barrett 🌴 @craig marked this issue as related to #5869 1 year ago






Craig Barrett 🌴 @craig · 1 year ago

Owner

[@dawsmith](#) when I had initially evaluated [Atlantis](#), we discounted it because of it's [security issues](#) with public repos. Since the pipelines are running on a private instance (ops), that no longer applies, so I've added an [RCA](#) to setup Atlantis for distributed locking.







Closed RCA for 2018-12-21 Gitaly Outage


- **Craig Barrett** 🌴 @craig changed the description 1 year ago
- **Craig Barrett** 🌴 @craig changed the description 1 year ago
- **David Smith** 🌴 @dawsmith changed milestone to [%DS Team 2019 Week 3](#) 1 year ago

**David Smith** 🌴 @dawsmith · 1 year ago

Owner

Also noting that we have since added badges in <https://ops.gitlab.net/gitlab-com/gitlab-com-infrastructure/pipelines> to the repo so we can see if the pipeline has been passing.

- **David Smith** 🌴 @dawsmith mentioned in issue [#5945 \(closed\)](#) 1 year ago
- **David Smith** 🌴 @dawsmith changed the description 1 year ago
- **David Smith** 🌴 @dawsmith marked this issue as related to [#5945 \(closed\)](#) 1 year ago
- **David Smith** 🌴 @dawsmith assigned to [@dawsmith](#) 1 year ago
- **David Smith** 🌴 @dawsmith changed milestone to [%DS team 2019 Week 5](#) 1 year ago
- **David Smith** 🌴 @dawsmith added [moved 1](#) label 1 year ago

**GitLab Bot** 🤖 @gitlab-bot · 1 year ago

Maintainer




Hi [@dawsmith](#),

This issue does not appear to have an issue weight set.

As a general guidelines use a weight of 1 for an access request issue or a simple configuration update. Use this as a multiplier for setting the weight. If you are unsure about what weight to set it is better to add a generous estimate and change it later. If the weight on this issue is 8 or larger then it might be a good idea to consider splitting this issue up into smaller pieces.

Thanks for your help! ❤️

You are welcome to help [improve this comment](#).

- **David Smith** 🌴 @dawsmith changed weight to 3 1 year ago
- **David Smith** 🌴 @dawsmith closed 1 year ago
- **ops-gitlab-net** 💬 @ops-gitlab-net mentioned in issue [#6130 \(closed\)](#) 1 year ago

Please [register](#) or [sign in](#) to reply