Main page
Recent changes
Server admin log (Prod)
Server admin log (RelEng)
Deployments
SRE/Operations Help
Incident status

Cloud VPS & Toolforge
Cloud VPS documentation
Toolforge documentation
Request Cloud VPS project
Server admin log (Cloud VPS)

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page

Print/export
Create a book
Download as PDF
Printable version

Toolforge webservices are in the final stages of migrating to the toolforge.org domain.
Please help us clean up older documentation referring to tools.wmflabs.org!

Incident documentation/20190417-Jobqueue

< Incident documentation

Contents [hide]

1 Summary
1.1 Impact
1.2 Detection
2 Timeline
3 Conclusions
3.1 What went well?
3.2 What went poorly?
3.3 Where did we get lucky?
4 Links to relevant documentation
5 Actionables

Summary

A change to how MediaWiki encodes jobs in Kafka was added to 1.34.0-wmf1. This caused several different jobs on group1 wikis to fail from 2019-04-17 19:14 up until 2019-04-18 12:58.

Impact

Several different jobs categories completely failed to execute for the duration of the incident.

Further investigation is needed if we want to evaluate the full impact, but for example MassMessage (from meta.wikimedia.org) would silently fail to deliver messages for the duration of the incident.

Detection

The issue was first detected in production by User:Elitre while trying to send a message via MassMessage.

Timeline

All times in UTC.

See also T221368 which has some debugging play-by-play and actions taken.

April 11

- A Fatal error from Special:MassMessage relating to the JobQueue was reported about the Beta Cluster at T220662.
- The error was quickly fixed in Git master, but the user reports that while the fatal error no longer occurred, the actual message still wasn't delivered in Beta Cluster.
- This was assumed to be due to the job runners being offline in Beta Cluster (T220662#5104700).

April 16

- The 1.34.0-wmf1 branch is created from the Git master and deployed to test wikis (group0).

April 17

- 19:14 Wikis in group 1 get updated by the train to use 1.34.0-wmf1. Error rates for jobs start raising immediately.

Apr 18

- 11:42 User:Elitre reports to _joe_ messages sent via the MassMessage extension are not being delivered. This seems like a simple bug, Erica is advised to open a task.
- 12:21 _joe_ starts investigating the reported bug
- 12:31 _joe_ finds out an unusual number of errors are being emitted by the jobrunners - they went from ~ 3

per second to ~ 100 per second. The errors affect multiple types of jobs (not limited to MassMessage) and all report the same error message: "Failed to create job from description / Title Special: is invalid". Investigation of the incident starts.
- 12:51 It is identified that all errors come from wikis on 1.34.0-wmf1, and they started after the train deployment the previous evening.
- 12:54 Reedy prepares a revert for group1 back to the previous version.
- 12:58 deployment of the revert is completed. The rate of jobs failing starts dropping. The impact part of the incident ends here.
- 14:26 Pchelolo identifies the root cause to be mediawiki/core change 500171.
- 17:46 mobrovac deploys 504929 as a temporary measure to resolve the immediate fatal "Invalid title" errors.
- 19:10 mobrovac deploys 504942 to have the JobExecutor behave correctly vis-a-vis the new way of passing the page title to jobs; despite this errors for some job(s) continued (Translate extension).
- 19:20 cdanis stops CP4JQ to minimise impact.
- 20:30 mobrovac restarts CP4JQ back up.
- 20:40 mobrovac deploys 504961 which makes the Translate extension errors cease; outage is over.

Conclusions

We should really consider execution on the jobrunners a first-class citizen.

What went well?

- Not much?

What went poorly?

- The testing infrastructure (including beta) didn't detect the problem.
- The deployment process didn't detect the problem
- Production monitoring did detect the problem, but no alert was set up on the relevant metrics

Where did we get lucky?

- Erica (Elitre) had a deadline to meet for a communication so she reported what seemed as a simple bug with some urgency.

Links to relevant documentation

Kafka_Job_Queue - the documentation of the kafka job queue.

Actionables

- Fix the jobrunner in deployment-prep T215339 ✓ **Done**
- Scap should include jobrunners in its canary process (T172480 - filed in 2017)
- Scap's Logstash checker ("fatal monitor") should include errors from the jobqueue (T172480 - filed in 2017)
- Create an alert on the number of 5xx responses from the jobrunners (TODO: Create task)
- Cookbook for downtiming only a particular service on a set of hosts (TODO: Create task)
- Cookbook procedure for cancelling all downtimes with a particular comment message (TODO: Create task)

Category: Incident documentation

Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. See Terms of Use for details.