Page   **Discussion**

Read   **View source**   **View history**   Search Wikitech

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/20181024-WDQS

< Incident documentation

Summary

WDQS updater have stopped updating on the 24th Oct 2018 due to inability to parse Kafka stream. Icinga also reported SSH failure for all WDQS servers. During this period, WDQS is accessible from (https://query.wikidata.org) and queries worked correctly. Restarting/power cycling some of the servers (wdqs2005 and wdqs1009) solved the SSH issue but not the Updater issue. WDQS updater spilled `java.time.format.DateTimeParseException` which is related to date format change in `dt` field of Kafka events, coming from Mediawiki change. The cause is most likely this patch: https://gerrit.wikimedia.org/r/c/mediawiki/extensions/EventBus/+/468482

See here for further details: https://phabricator.wikimedia.org/T207817

Timeline

This is a step by step outline of what happened to cause the incident and how it was remedied.

- Oct 24 00:28:30 (UTC) Mediawiki update, which includes https://gerrit.wikimedia.org/r/c/mediawiki/extensions/EventBus/+/468482
- Oct 24 00:28:37 (UTC) First error in Kafka poller for WDQS Updater, after change from mediawiki.org using new date format
- Oct 24 1:11 (UTC) Icinga reported the first set of errors:
    - PROBLEM - SSH on wdqs1010 is CRITICAL: CRITICAL - Socket timeout after 10 seconds 2:11 AM PROBLEM - SSH on wdqs1008 is CRITICAL: CRITICAL - Socket timeout after 10 seconds 2:11 AM PROBLEM - SSH on wdqs2001 is CRITICAL: CRITICAL - Socket timeout after 10 seconds 2:11 AM PROBLEM - SSH on wdqs2006 is CRITICAL: CRITICAL - Socket timeout after 10 seconds
- Oct 24 1:15 (UTC): More of these errors touching other servers were reported by icinga
- Oct 24 1:17 (UTC): It was confirmed that these errors are related to: `java.time.format.DateTimeParseException` as wdqs-updater threw this errors continuously on all servers.
- Oct 24 1:24 (UTC): wdqs2005 was powercycled to solve the SSH problem which worked (we could login to the server but wdqs-updater issues persisted.
- Oct 24 1:30 (UTC): We discovered new update touching the updater was made earlier and we started making attempt to revert update back on one of the server(wdqs1003)
- Oct 24 2:11 (UTC): Attempts to revert updates back on wdqs1003 failed due to `git rebase` issues
- Oct 24 2:15 (UTC): Attempt were made to login into wdqs1009 (autodeploy/test server). This failed and needed to be powercycled.
- Oct 24 2:34 (UTC): wdqs1009 was powercycled.
- Oct 24 2:43 (UTC): Revert of Updater was done successfully on wdqs1009 but wdqs-updater issues persisted.
- Oct 24 2:46 (UTC): It was concluded that since the servers were still serving queries during this period, we could wait till some more people are awake.
- Oct 24 5:59 (UTC): patch to switch Updater to recentchanges API
- Oct 24 6:33 (UTC): SSH issues persists and wdqs1004 was power cycled.

- Oct 24 6:35 (UTC): WDQS servers at codfw were powercycled one after another.
- Oct 24 6:38 (UTC): powerclycing and the patch to fix wdqs-updater worked and this was propagated on other servers.
- Oct 24 7:11 (UTC): It was confirmed that mediawiki update train impacted wdqs-updater
- Oct 24 08:52 (UTC): https://gerrit.wikimedia.org/r/c/mediawiki/extensions/EventBus/+/468482 is reverted from 1.33.0-wmf.1 branch

Further Investigation is ongoing as to why SSH failed totally.

Conclusions

What weakness did we learn about and how can we address them?

- Communication of changes that are likely to impact other platforms depending on mediawiki should be sent out before deployment.
- Our setup (at least on WDQS servers) is way too fragile - Updater problem should not cause log congestion, and log congestion should not bring down a system to the point where reboot is required to even log in
- We need better monitoring for Updater failures - 45 minutes between Updater failing completely and us starting to research is too long.
- We need graceful failure setup - if one component fails, it should not bring down the whole system in futile attempts to restart, but rather alert somebody ASAP.
- It took us > 5h to address the problem. SSH failure to all 14 wdqs nodes should have been handled with higher priority.
- We may also have coverage gap in evening PDT/night CET - 01:00 UTC to 07:00 UTC may be late enough in PDT so Stas can be offline, and early enough in CET so Gehel is not online yet. We should figure out who could help if something happens in that time window.

Links to relevant documentation

Wikidata query service/Runbook

Actionables

Explicit next steps to prevent this from happening again as much as possible, with Phabricator tasks linked for every step.

NOTE: Please add the #wikimedia-incident Phabricator project to these follow-up tasks and move them to the "follow-up/actionable" column.

- Add how to recover from this type of issue into the WDQS runbook - Wikidata query service/Runbook
- Status: ▌ **Done** Make Updater more tolerant to date formats phab:T207873
- Status: ▌ **Done** Refactor WDQS logging so that it does not use journald for routine logging phab:T207834
- Status: ▌ **Done** Change Updater systemd setup so it does not try to restart every 3 seconds if it's not working phab:T207843
- Specify Kafka data formats (esp. datetimes) more precisely, and provide notifications when they change
- Status: ▌ **Done** Switch WDQS cluster back to Kafka updating when things calm down

Category : Incident documentation