Create account Log in

Page Discussion Read View source View history Search Wikitech

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/20190616-AQS

< Incident documentation

Contents [hide]

Summary

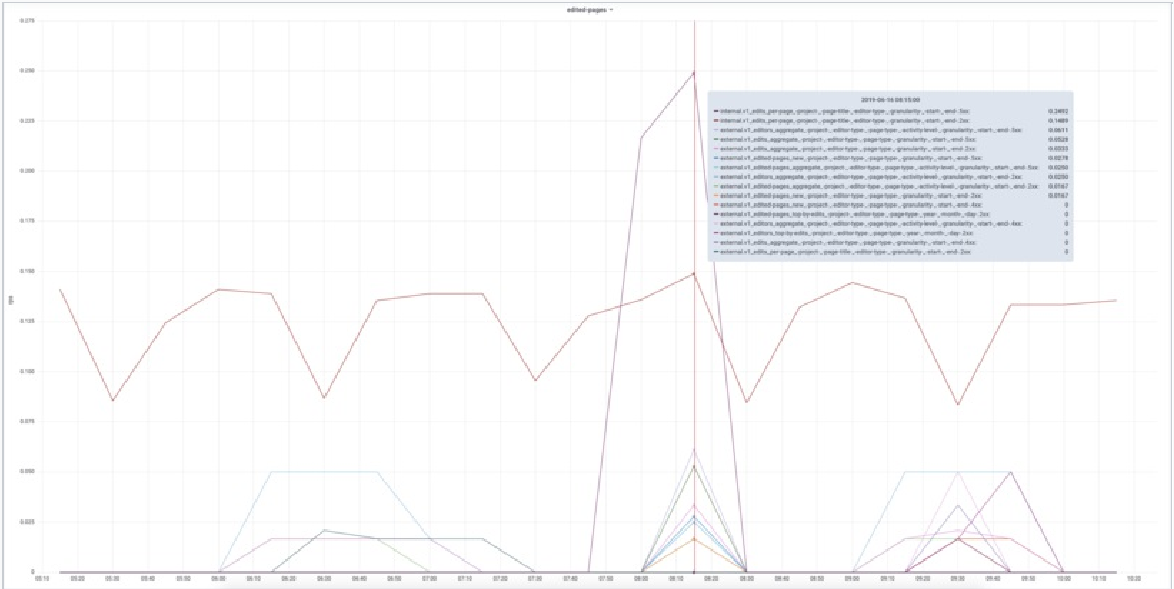
STATUS: final

The Analytics AQS service is the backend for Analytics APIs like Pageviews, Unique Devices and more recently all the data to support the new Wikistats 2 website. In order to implement the edit API for Wikistat 2, AQS needs to query another service, Druid. Recurrent systemd timers on the Analytics Coordinator node (an-coord1001) drop old data from Druid to increase its efficiency and disk/memory space used on the hosts in which it runs (druid100[4-6]).

A systemd timer meant to clean up old MediaWiki history snapshots on Druid (the data needed by Wikistats 2) failed to run due to a change in the script that needed to be ran, since an argument was changed and puppet had not been updated accordingly. Luca merged https://gerrit.wikimedia.org/r/#/c/operations/puppet/+/517203/ to fix the systemd timer, and he manually restarted its service unit to force the timer to run in order to verify that everything was running fine. The old snapshots on Druid were deleted, but sadly this triggered an old Druid bug that we thought to have resolved, namely all Druid brokers locked up in some state without being able to respond to any query (https://phabricator.wikimedia.org/T220111). This in turn caused the AQS API to start responding 50x HTTP errors for all the API requests for Edit data, including the ones coming from the Wikistats 2 website.

Impact

Every HTTP call to the AQS service for edit-related APIs got a 503 during the timeframe of the outage. The Wikstats V2 website (still in beta) ran in degraded mode for the timeframe of the outage.



HTTP 50X impact during the outage. Since the AQS MediaWiki Edit history part of the API is low volume, the impact was minimal.

Detection

The AQS application level health checks for the edit API fired first, followed by the Pybal ones indicating that the LVS backend hosts behind druid-public-broker.svc.eqiad.wmnet were all down.

Timeline

All times in UTC.

- 2019-06-15 7:10 - The analytics team receives an alarm from Icinga about a failure registered by the last execution of the refinery-druid-drop-public-snapshots systemd timer on an-coord1001 (non-zero return code from the service unit).
- 2019-06-16 8:09 - Luca notices the problem, creates/merges https://gerrit.wikimedia.org/r/#/c/operations/puppet/+/517203/ and forces a puppet run on an-coord1001 to fix the systemd timer. Then Luca forces the execution of its service unit via a manual restart of the refinery-druid-drop-public-snapshots service unit.
- 2019-06-16 8:09 - The druid coordinator leader starts the procedure to drop the two mediawiki history snapshots from the historical nodes, following what (correctly) indicated by the refinery-druid-drop-public-snapshots unit.
- 2019-06-16 8:10 - All the druid brokers on druid100[4-6] stopped serving HTTP queries (stuck in a weird/buggy state). **OUTAGE BEGINS**
- 2019-06-16 8:13 - First alarm fired for aqs1005's AQS endpoint (health check failure).
- 2019-06-16 8:21 - Luca restarts all the druid brokers on druid100[4-6] and full recovery was reached few seconds after it. **OUTAGE ENDS**

Conclusions

What went well?

The monitoring of the AQS endpoint triggered the first alert few minutes after the outage started, so a manual intervention was quickly able to limit the damage and restore the functionality of the service.

What went poorly?

This issue already happened in the past, but it was meant to be fixed by all the work done in https://phabricator.wikimedia.org/T220111. We thought that the issue was related to an aggressive data drop procedure, causing lag and timeouts in Druid and leading to this failure scenario. Our tests suggested that the issue was fixed, but it is clearly not the case.

Luca didn't follow any procedure (since there is none at the moment) to attempt to dump the status of a Druid broker before restarting it (for example via jstack etc..). This would have been extremely useful to identify a bug and report it to upstream.

Where did we get lucky?

The refinery-druid-drop-public-snapshots runs once every month, and its was meant to run on a Saturday during early EU morning, when usually people from the Analytics team are not around. It luckily broke, so when Luca restarted it he was able to fix the issue very quickly since he already knew how to solve it from his past experience. It would have probably caused a longer outage if the script executed correctly on Saturday (few SREs around, without a lot of knowledge of this particular failure scenario).

Links to relevant documentation

The AQS alarms were carrying the following wiki page: Services/Monitoring/aqs. It is difficult from a first read, if you are not familiar with the relationship between AQS and Druid, to figure out that part of the API relies on Druid to function properly. This needs to be addressed as action item.

Actionables

- Work again on T220111 to figure out what is the problem with Druid and how to safely drop data without causing outages (T226035)
- Improve documentation about AQS with details about what its relationship with Druid is. (TODO as part of T226035)
- Allow the execution of the druid datasource drop script only during workdays, to avoid causing outages during the weekend (Done in https://gerrit.wikimedia.org/r/#/c/operations/puppet/+/519361/).
- Add documentation about how to use jstack when Java daemons are in trouble. (Done in T226035#5288322)

Category: Incident documentation