Database Meeting 2022-10-26

Date

26 Oct 2022

Attendees

- Igor Gaponenko Fabrice Jammes Fritz Mueller Andy Salnikov John Gates Kian-Tat Lim Colin Slater Joanne Bogart

Notes from the previous meeting:

- Database Meeting 2022-10-12

Discussion items

Discussed	Item	Who	Notes
✓	Project news	Fritz Mueller Colin Slater	Fritz Mueller: • any news from DM Leadership Team Virtual Face-to-Face Meeting - 2022-10-18? ○ (operationally) Qserv is being moved from SLAC Infrastructure (Richard) to LSST Data Services (reported to Frossie) ○ presented the road map for supporting the user-defined features in Qserv ○ (easier problem) ingesting single-use transient tables has been approved. The tables could be ingested w/o getting TAP involved ○ (harder problem) of ingesting persistent tables into Qserv using the existing Ingest API and possibly new extensions still further discussions ▪ there are a few complications here, including decisions on the partitioning parameters, which are difficult to make at the TAP level • on plans for the next 6 months: ○ main goal: adding support for the user-generated data products in Qserv ○ DP02 testing using Google Big Query to compare with Qserv ○ new 15 nodes Qserv hardware arriving in November; the new Qserv will be based on Kubernetes; the existing 6 nodes cluster may still be retained for some time while the new one becomes stable and useful ○ hardware (12 nodes) for APDB (Andy S) is also arriving soon ▪ Andy Salnikov was wandering about the hardware specs of the cluster. ▪ Fritz Mueller will look for the specs ▪ Kian-Tat Lim the machines are going to join the Kubernetes cluster. Although they will be locked for the specific use case. ○ Fritz Mueller realistically speaking, the new nodes will be available in January 2023 • any word from the Google colleagues on ingesting & testing DP02 into the BigTable? ○ nothing yet ○ they seem to be unblocked Continued discussion on the office spaces at ROB to relocate the DAX team from B50. This is still in progress.
✓	User-generated data products	team	Context: • started discussing the topic at the previous meeting Database Meeting 2022-10-12 • Fritz Mueller made two presentations mentioning the topic at the last VF2F DMLT on Oct 18th, 2002: ○ DM Leadership Team Virtual Face-to-Face Meeting - 2022-10-18 ○ https://confluence.lsstcorp.org/display/DM/DM+Leadership+Team+Virtual+Face-to-Face+Meeting++-+2022-10-18?preview=/197235494/197245341/User-Generated%20Data%20Products.pdf Next steps? • it seems that we've been given the "green light" to proceed with implementing the single-use query • Fritz Mueller and Igor Gaponenko will continue discussing practical steps toward implementing this • We will start with developing the single-short ingest (REST) API for ingesting the tables • Initially, we will only support CSV as the input data format. Support for ingesting the VOtables will be added later. • We need a schema for ingesting user data. CSV only supports the names of the columns. • Kian-Tat Lim eCSV supported by AstroPy allows schema specs. It's available in Python only. • Igor Gaponenko (as an option) we might end up building another REST service in front of the core Ingest API to perform data transformation and schema extraction before interacting with the core API. • Fritz Mueller an architectural decision on where to put this operation (closer to the TAP services or inside Qser) is yet to be made. This needs to be discussed with Frossie. • Fritz Mueller we need to support both ingest options: 1) by reference, and 2) by value (the "push" mode) • Igor Gaponenko for the "push" mode we need to improve `qhttp` to support multi-part attachments in the request body • Fritz Mueller we might look at Boost Beast to see if we could use it (as a whole, or just the relevant tools) for that purpose

✓	(Possible) bug in Qserv `czar` when handling failed chunk queries	team	Context: • the problematic query involves 3 tables: Object, TruthMatch, and MAtchesTruth (RefMatch) • It's a large result query (a few **GB**, 10k, or 100k chunks of each chunk, **58** chunks involved) • Qserv `czar` leaves the failed queries in the pending state if the failures were triggered by the worker restarts. Workers were restarted in `k8s` due to OOM (memory pressure) • It's been seen with the `replication_level=1` after recovery (query retry) attempts made by `czar` • we might not see this problem in the past because the replication level was higher ... or, perhaps, the specific query time might trigger the issue. • In some cases, the queries are staying in the `EXECUTING` state. In other cases (USDF tests below), `czar` gets into a strange state by refusing to process any further queries • The problem seems to be reproducible (it's been reproduced using Qserv `slac6` at USD, though, differently Eventually, Qserv `czar` got crashed. • Igor Gaponenko "worker restarts" are different in `k8s` and the host environment. In the former case, the IP addresses associated with the workers would disappear from `k8s` DNS. In the case of the host-based Qserv deployments (so-called "iGor" mode), the hosts are still staying in DNS. Only the XROOTD servers would disappear. This *may* affect the outcome of the problem (how it's handled by `czar`). How do we investigate this problem? • ✓ Igor Gaponenko will work with Yee to allow Fritz Mueller and John Gates to log into the Qserv cluster at USDF. • ? John Gates will investigate the bug
✓	Status of `qserv-ingest` and `qserv-operator`	Fabrice Jammes	Fabrice Jammes There is the following proposal from FrDF to implement the fast Parquet-to-CSV translator in C++. Possible options include a separate application or integration with an existing partitioning tool: https://lsstc.slack.com/archives/C996604NR/p1666811747284709 Fritz Mueller is in favor of the latter option. We should also support (eventually) VO tables. Igor Gaponenko we need to use Parquet "row groups" to allow parallel translation of the files. The columns could be efficiently compressed if they have repeating data patterns (all zeroes Kian-Tat Lim: it's not done yet by the Pipeline. No JIRA ticket exists yet for this improvement. Though, a need in having the row groups has been recognized by the developers. Colin Slater: column-oriented format provided by the Parquet data format is essential for the data analysis based on these files. Qserv is not the only (or the main) user of these files. Igor Gaponenko mentioned that the source files of the partitioner are now a part of the Qserv source tree. The partitioner's binaries are now built as a part of the Qserv binary container. There are concerns about bringing extra dependencies into the Qserv container. Fritz Mueller thinks we could introduce refined binary containers to separate Qserv itself from the partitioning tools. This may lead to better control over the dependencies. Fritz Mueller on the practical steps in this direction: • Fabrice Jammes and the team will begin working on a prototype of the idea using the Qserv development container • The rest of the DAX team will provide support if needed
✓	Status of the `ObsCore` table	Andy Salnikov	Any news? Slow progress so far. An implementation of the "live ObsCore manager for Butler" (PostgreSQL) has finished. It works. Still requires more testing. A few PostgreSQL extensions are required by ObsCore. One set has been installed at USDF. More will be needed. More details on the status of the project can be found at Live ObsTAP service deployment

Action items

□