DBA for KCIDB

Request for Proposals

Summary

The KernelCI's Maestro project runs thousands of tests on dozens of Linux kernel trees everyday generating a great amount of data. The KernelCI's KCIDB project receives that data in addition to similar data from seven other CI systems (so far), and stores and aggregates it to provide long-term archive, notifications and dashboards to the community.

This Request For Proposals aims to attract contract offers for making the KCIDB's backend database more scalable and maintainable long-term, in the face of ever increasing amounts of arriving data.

Last edited	2025-01-22
Version	0.1
Authors	

The KernelCI Project

KernelCI was created several years ago by Linux kernel maintainers in order to provide an automated test system for the upstream kernel. It then became a Linux Foundation project in

2019 and the founding members are still part of the project. More information can be found about it on the kernelci.org website. An annual budget is collected from the members' contributions which makes it possible to fund work packages, this being one of them.

KCIDB challenges

At the moment, KCIDB receives up to half a million of test and related data rows per day (330K average). We keep six months worth of data in PostgreSQL (117GB in tables and 16GB in indexes) for displaying on dashboards, analyzing, and generating notifications. The complete archive is stored in BigQuery.

While our dashboards and notifications generally function, their performance has been degrading recently, due to increasing ingestion, no longer leaving us space and time to experiment and develop new and exciting features. Additionally we would like to keep more historical data in the operational PostgreSQL database, so we can make queries over a larger timespan.

Deliverables

- A design of a scalable database architecture, and an assessment of capacity required to hold 10x the current load/scale.
- Documented production deployment code integrated with KCIDB CI (existing or modified), deploying the following:
  - A PostgreSQL server tuned for serving the main KCIDB database.
  - An empty KCIDB database tuned and ready for accepting data.
- Automated server/schema upgrades NOT included in the automatic deployment, but available separately.
- Google Cloud and/or Docker/Podman containers as supported deployment targets.
- All code licensed under GPLv2 or LGPL-2.1

The deliverable process needs to be iterative, gathering extensive feedback multiple times from the Sysadmin working group and the community at large.

Vendor Considerations

Proposed Timeline

The deadline for responding to this RFP is six weeks after it has been made public. Then the KernelCI Advisory Board of Members will vote and respond within two weeks, so a decision should have been reached within eight weeks. By starting this timeline on `12 Feb 2025`, the deadline to respond is `26 Mar 2025` with a board vote on the `2 Apr 2025`. Exact dates

might be subject to change in case of a major practical issue or unavailability of voting members.

It is expected that the (part-time) work in this work package will last two to four months. The envisioned dates are from `1 May 2025` to `31 Aug 2025`. Alternative dates may be used in proposals and any improvements should be completed within their own defined timeframe.

Contact information

Please send inquiries, clarification questions and responses to this RFP via email to Nikolai Kondrashov <spbnick@gmail.com> and the KernelCI governing board <kernelci-members@groups.io>.

Current KCIDB schema

public

checkouts

_timestamp	***timestamp with time zone***	
id	***text***	« pk »
origin	***text***	« nn »
tree_name	***text***	
git_repository_url	***text***	
git_commit_hash	***text***	
git_commit_name	***text***	
git_repository_branch	***text***	
patchset_files	***jsonb***	
patchset_hash	***text***	
message_id	***text***	
comment	***text***	
start_time	***timestamp with time zone***	
contacts	***jsonb***	
log_url	***text***	
log_excerpt	***character varying(16384)***	
valid	***boolean***	
misc	***jsonb***	
git_commit_tags	***text[]***	
git_commit_message	***text***	
git_repository_branch_tip	***boolean***	

checkouts_pkey	***constraint***	« pk »
checkouts__timestamp	***index***	
checkouts_git_commit_hash_patchset_hash	***index***	
checkouts_origin	***index***	
checkouts_start_time	***index***	
checkouts_git_repository_url	***index***	
checkouts_git_repository_branch	***index***	
checkouts_valid	***index***	
checkouts_tree_name	***index***	
checkouts_git_commit_tags	***index***	
checkouts_git_repository_branch_tip	***index***	

builds

_timestamp	***timestamp with time zone***	
checkout_id	***text***	« nn »
id	***text***	« pk »
origin	***text***	« nn »
comment	***text***	
start_time	***timestamp with time zone***	
duration	***double precision***	
architecture	***text***	
command	***text***	
compiler	***text***	
input_files	***jsonb***	
output_files	***jsonb***	
config_name	***text***	
config_url	***text***	
log_url	***text***	
log_excerpt	***character varying(16384)***	
valid	***boolean***	
misc	***jsonb***	

builds_pkey	***constraint***	« pk »
builds__timestamp	***index***	
builds_checkout_id	***index***	
builds_origin	***index***	
builds_start_time	***index***	
builds_architecture	***index***	
builds_config_name	***index***	
builds_compiler	***index***	
builds_valid	***index***	

tests

_timestamp	***timestamp with time zone***	
build_id	***text***	« nn »
id	***text***	« pk »
origin	***text***	« nn »
environment_comment	***text***	
environment_misc	***jsonb***	
path	***text***	
comment	***text***	
log_url	***text***	
log_excerpt	***character varying(16384)***	
status	***public.status***	
waived	***boolean***	
start_time	***timestamp with time zone***	
duration	***double precision***	
output_files	***jsonb***	
misc	***jsonb***	
environment_compatible	***text[]***	
number_value	***double precision***	
number_unit	***text***	
number_prefix	***public.unit_prefix***	

tests_pkey	***constraint***	« pk »
tests__timestamp	***index***	
tests_build_id	***index***	
tests_origin	***index***	
tests_start_time	***index***	
tests_path	***index***	
tests_status	***index***	
tests_waived	***index***	
tests_environment_compatible	***index***	
tests_number_value	***index***	
tests_number_unit	***index***	

issues

_timestamp	***timestamp with time zone***	
id	***text***	« pk »
version	***integer***	« pk »
origin	***text***	« nn »
report_url	***text***	
report_subject	***text***	
culprit_code	***boolean***	
culprit_tool	***boolean***	
culprit_harness	***boolean***	
build_valid	***boolean***	
test_status	***text***	
comment	***text***	
misc	***jsonb***	

issues_pkey	***constraint***	« pk »
issues__timestamp	***index***	
issues_origin	***index***	
issues_report_url	***index***	
issues_culprit_code	***index***	
issues_culprit_tool	***index***	
issues_culprit_harness	***index***	

incidents

_timestamp	***timestamp with time zone***	
id	***text***	« pk »
origin	***text***	« nn »
issue_id	***text***	« nn »
issue_version	***integer***	« nn »
build_id	***text***	
test_id	***text***	
present	***boolean***	
comment	***text***	
misc	***jsonb***	

incidents_pkey	***constraint***	« pk »
incidents__timestamp	***index***	
incidents_build_id	***index***	
incidents_test_id	***index***	
incidents_issue_id	***index***	
incidents_origin	***index***	
incidents_issue_version	***index***	
incidents_present	***index***	