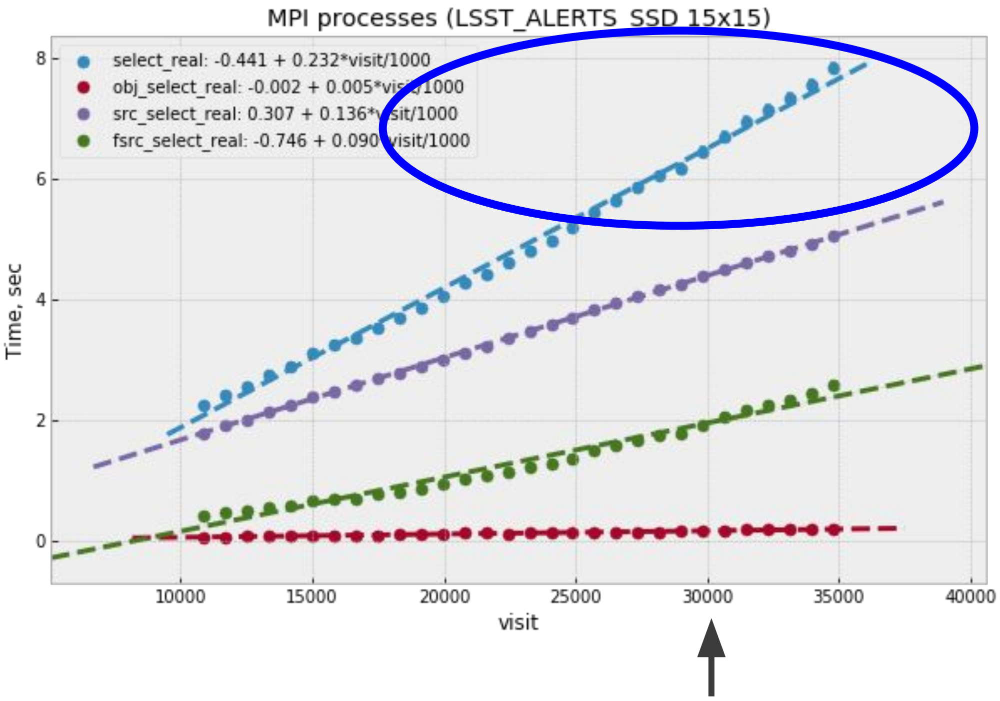
PPDB Status

Large Synoptic Survey Telescope

DMTN on relational DB testing

- Summary at https://dmtn-113.lsst.io/
  - Using Postgres and extrapolating the very-linear-looking behavior, runtime is off by factor of ~few
  - Oracle gave more confusing behaviors, but eventually became linear with similar performance miss
-
- Performed with a variety of borrowed/rented hardware, at IN2P3 and Google

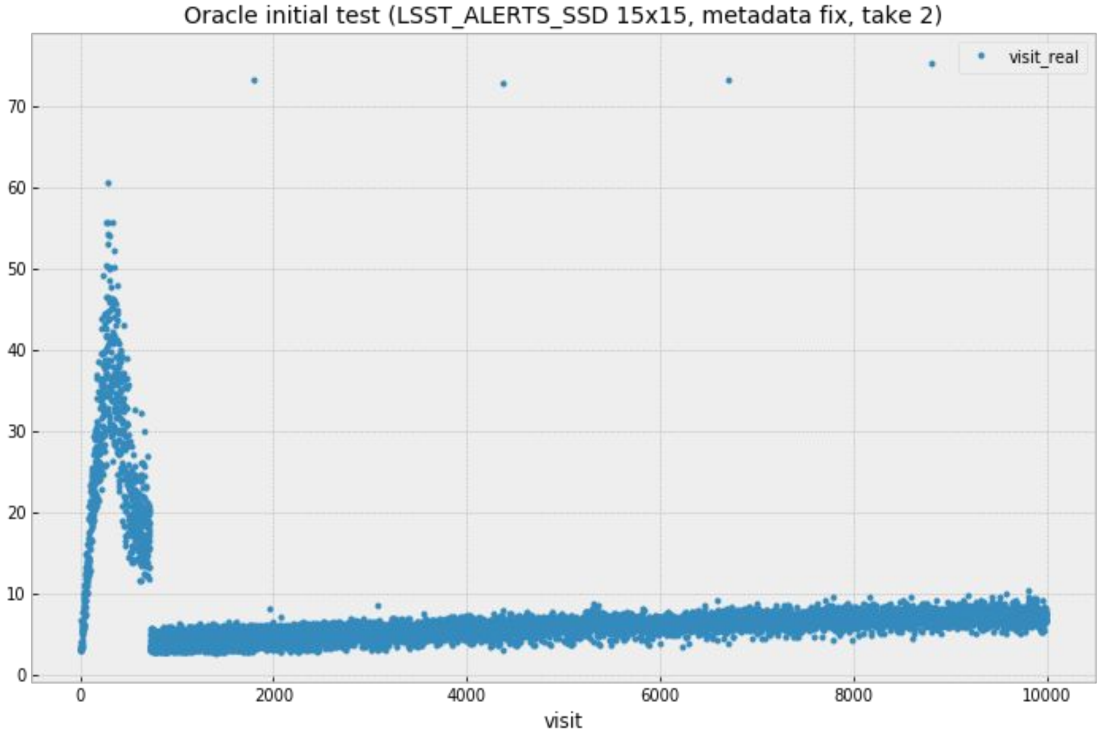
DiaSource SELECT time dominates



One Month

- Using SSD, parallel “image processing” nodes
- ~6 seconds after 1 month -> 72 seconds for 12 month history
- Query time is proportional to both data size on disk in the DB and returned result size, don’t currently have data to disambiguate

Oracle



- Oracle gives similar asymptotic performance, but with weird startup transients that are hard to control.
- No clear benefit to further studies down this path

Cassandra

- Cassandra testing: DM-20580
- This is the most reasonable-ish looking off-the-shelf technology for a distributed time-domain DB system (log-structured merge tree).
- Some skew between its design and our use case, but gives us multi-node capabilities.
- Upcoming procurement includes new nodes configured as qserv Czars, will initially use these for PPDB testing (+ an existing Czar node)
- Expected to be procured Nov/Dec by NCSA, available to DAX in January.

Beyond Cassandra

- Experiment with custom solutions
  - Can we put together a system from smaller stock parts, write some of our own code?
  - E.g. use an object store for static “blobs” of records from past nights + combine with DB results for tonight’s latest updates.
  - Goal would be to better exploit the structure of the problem
- Push back on requirements
  - Most significant is probably alert time-series as currently conceived. Perhaps less history, or simplify “sliding window” design?
  - What could be gained by relaxing 60-second alert constraint?

Public PPDB Releases

- Got general agreement with Bob Blum to make the PPDB contents world-public (though not necessarily accessible to those without data rights)
- Appropriate wording was added to the data rights doc; haven’t seen the result post-NSF review.
- This potentially makes life easier for certain brokers, even absent any technical changes.
- Potentially a larger (scientific & usability) gain to be had from enabling “mirroring” of the PPDB; technical implementation not started due to overall PPDB uncertainty.

Naming

Proposed naming tweak, to reduce ambiguity:

- “APDB” (constrained, used during AP, maybe custom)
- “PPDB” (conventional RDBMs, released product)

This is already de-facto. Shall we make it official? If so, what needs to be done?