Rubin Observatory

APDB Status

Fritz Mueller

Previously...

Rubin Observatory

- Jan/Feb: APDB on Cassandra evaluation began, using temporary cluster composed of three borrowed Qserv master nodes, driven by **`ap_proto`** simulator running on verification cluster.

- Feb: preliminary results reported to DMLT: Cassandra looking more feasible to scale than monolithic SQL had done, but many unknowns. Work continues…

- Mar: Andy Salnikov BG3 work was prioritized above APDB; Cassandra eval effort reduced to ~.10 FTE

Recent Progress

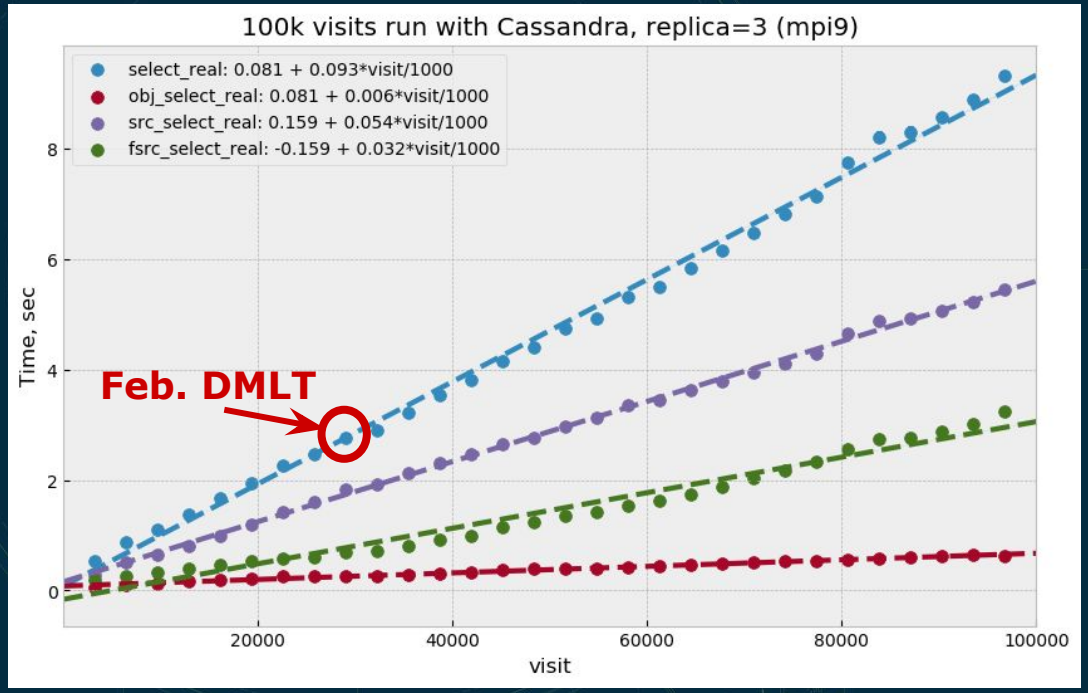
Rubin Observatory

Most work since Feb DMLT captured in ticket DM-23881

- Tuning to use finer Cassandra partitioning
- Do pixelId filtering client-side to reduce query load
- Tuning of JVM memory size to limit swap activity and smooth GC
- Tune frequency of flush/compact
- Tune GC algorithm and parameters.
- Run more instances per node (via Docker) to smooth GC
- Experiments with replication factor (x3 w/ quorum if feasible)

Also started looking at Scylla (C++ Cassandra-alike) DM-24692

Some Recent Results



- Sim 100K visits (~3.3 mos.)
- Same three phy. node cluster
- Blue trend is limiting: average time for all needed db reads per CCD (parallelized)
- Write times (tested concurrently with above) not limiting; ~.5s and roughly constant.
- Improved stability since last report (smoothed GC, disconnects, etc.)
- Details in DM-23881

Late Breaking, from NCSA

Rubin Observatory

Have been experimenting independently, running Andy's `ap_proto` on verification cluster against Postgres. Some findings, via Michelle:

- Achieved Postgres SELECTs faster than the preliminary Cassandra queries reported at previous DMLT **(Caveat: have not heard yet how these tests were configured/conducted!)**
- Report Cassandra/Scylla faster for inserts, but feel Postgres could come likely come close with data layout changes, sharding, etc.
- Overall, testing showed greatest gains to be had from adjusting sharding, partitioning, query design, etc. and lesser from choice of database or hardware.
- Expressed opinion: databases are hard-pressed to meet these requirements right now, but DBAs have seen other projects reaching similar requirements by digging into the details of the SQL and data layout etc.
- Expressed desire: want to meet with arch/developer folks to discuss options and leverage expertise w/ data layout, SQL, and possibilities for improvements to data design

Ongoing Work

- Need to scale beyond three nodes (cloud?)
- Get out to 1yr+ of simulation
- Loop in the NCSA DBAs; see what they have found and what we can learn
- Reminder: only ~.10 FTE allocated in DAX right now due to BG3 contention!