Lessons Learned from EnterPrise Users

Xuefu Zhang
Cloudera
Hive PMC

Background

- Seeing an uptrend of Hive usage by enterprise users
- Not just a few gurus, but also those who don't necessarily know much about Hive
- Multiple request sources: Beeline, JDBC, Hue...
- More workload: ETL, DW, Analytics, Reporting, AI
- Ever-increasing data size

Problems

- HS2/HMS OOM
- HS2/HMS pauses, big or small
- Backend DB overheat
- Hue UI unusable
- Large query latency

Causes

- One HS2 instance serves up to tens of concurrent user sessions.
- Small heap causes OOM. Large heap causes long GC pauses
- Compile lock serializes all incoming commands, causing large latency
- Fetching large number of partitions is time consuming for HS2, HMS, and backend DB server.
- Hue brings in more traffic (asynchronous requests)
- Hue brings more opportunities for mistakes: schema browsing, data sampling, etc.
- Unbalanced load to HMS

Lessons learned

- Enterprise user may not have much knowledge of Hive, or Hadoop
- Hive is perceived just as another database
- Capacity planning is usually missing
- It's hard to push thru changes in a deployment
- If used, it's mission-critical, and you get P1s
- Scalability/usability/stability means more than performance
- Problems on the horizon if not seen today

What we can do better?

- Better quality, scalability, supportability, security, HA
- Solve the fundamental partition problem
- Provide help on data lifecycle management
- Eventual enterprise readiness
- Better performance will make the above issues more dormnant.

Thank you!