Hive Background: What is it?

An ETL/Data Warehouse system for Hadoop:

• SQL->MR Compiler and Execution Engine

• SerDes: Pluggable Data Format Handlers

• MetaStore: Persistent Metadata Storage
Hive Evolution

• Original Vision:
  – Let users express their queries in a high-level language without having to write MR programs

• Now more and more:
  – A parallel SQL DBMS that happens to use Hadoop for its storage and execution layer.
What do users expect from a DBMS?

- **Sessions/Concurrency**
  - Persistent client state on the server-side
  - Ability to run multiple client concurrently
- **ODBC/JDBC**
  - SQL IDEs, BI, ETL, …
- **Authentication/Authorization**
- **Auditing/Logging**
What’s Missing?

• Sessions/Concurrency
  – Current Thrift API can’t support concurrency

• ODBC/JDBC
  – Thrift API doesn’t support common ODBC/JDBC

• Authentication/Authorization
  – Incomplete implementations

• Auditing/Logging
  – Multiple plugin interfaces in need of consolidation
What’s Missing

Concurrency/Sessions

• Current Thrift API can’t support multiple connections or client sessions.
• User/Global Configuration and Session Info
• Query compiler memory leaks
ODBC/JDBC

- Thrift API can’t support common ODBC/JDBC calls:
  - SQLGetInfo
  - SQLGetTypeInfo
  - SQLCancel
  - SQLGetFunctions
What’s Missing

Authentication/Authorization

• SASL Authentication for HiveServer
• Hive supports GRANT/ROLE based authorization, but implementation is incomplete.
• Code injection vectors: ADD JAR, TRANSFORM, SET x, …
Project Milestones

• HiveServer2 Thrift API Spec
• JDBC/ODBC HiveServer2 Drivers
• Concurrent Thrift clients
  – Fix query compiler memory leaks
  – User/Global session/configuration information
• Authentication (Kerberos)
• Authorization
  – Extend to configuration, ADD x, TRANSFORM, …
Who’s working on it?

• Carl Steinbach
  – carl@cloudera

• Prasad Mujumdar
  – prasadm@cloudera
Resources

• HIVE-2935: Implement HiveServer2

• HiveServer API Proposal:
  – https://cwiki.apache.org/confluence/display/Hive/HiveServer2+Thrift+API
Questions?

- Questions?
- Questions?
- Questions?
- Questions?