Hive on Spark: Now and Future

Xuefu Zhang
Cloudera
Apache Hive PMC
Short History

- Incepted in Summer 2014
- Tracked by HIVE-7292
- First release in Apache Hive 1.1 in March 2015
- CDH Beta in 5.3
- Support provided for selected customers since CDH 5.4
- GA will be in CDH5.7 (coming soon)
Current Status

• Functional parity with MR
• A few times better performance than MR
• Extensive testing: scale, performance, stress
• Bug fixes, usability improvement, integration with other components
• Production ready
Use Hive on Spark

• Minimum configurations
  – hive.execution.engine=spark
  – Spark.master=yarn-cluster

• Other modes work, though yarn-cluster is highly recommended and supported in CDH
Spark Configurations

• Defines how Spark utilizes YARN resources (core, memory)
  - spark.executor.cores
  - spark.executor.memory and spark.yarn.executor.memoryOverhead
  - spark.driver.memory and spark.yarn.driver.memoryOverhead

• Executor core and memory allocation directly impacts performance
Executor Allocation and Warming

- Static allocation
- Dynamic allocation
- Prewarm Spark executors
Fine Performance Tuning

• Number of executors mostly determines performance
• Share most of existing performance related configurations
• One important exception
  – hive.auto.convert.join.noconditionaltask.size
• Parallelism (number of reducers)
  hive.exec.reducers.bytes.per.reducer
Trouble-Shooting

• Try MapReduce
• Console output
• HiveServer2 log
• YARN application log
• YARN container log
• Spark driver webui and history server
  – Especially useful to see task statistics
  – Find slow nodes or long-trailing tasks
Future

• A lot of optimization work ahead
• Incorporate new Spark features (such as Tungsten)
• Integrate with new features in Hive: LLAP, CBO, Hbase HMS
• Better integration with other Hadoop components: Oozie, Hue
• Usability, scalability, performance
• Hive community is committed to make Hive on Spark better and faster
Summary

- Bearing solid foundations in design and architecture
- Well tested
- Great initial Performance
- Minimum configurations
- Production ready
- Grow with Hive
- Start using it and replacing your MR engine
- Please provide your feedback