Recap

- Hadoop installation
- Running JAR file on cluster
- HDFS API to delete HDFS file

Agenda for today

- MapReduce counters
- Performance tuning in MapReduce jobs
- Ad-hoc analysis with Impala
- Impala as Query processing tool
- MapReduce job chaining

Performance tuning

- Cluster configuration
- Use compression technique
- Tuning # mappers and reducers
- Use combiner
- Appropriate data type
- Reuse objects
- Profiling

https://blog.cloudera.com/blog/2009/12/7-tips-for-improving-mapreduce-performance/

MapReduce Job chaining

- Two separate jobs
- Multiple mappers/reducers within same job

MapReduce Job chaining

- Two separate jobs

1. Configure first job object and run it.
2. Configure second job object and run it

MapReduce Job chaining

- Multiple mappers/reducers within same job

https://mapr.com/blog/how-to-launching-mapreduce-jobs/

Impala SQL

- Create/Drop/Update table
- Insert/Update/Delete data into table
- Partitioning
- Modifying data directly in HDFS
- REFRESH table
- External vs Internal table
- Profiling and Optimization