ECE472 — Methods and tools for big data

Lab 4

Manuel — JI (Summer 2024)

Goals of the lab

- Install Drill and Spark
- Run Drill and Spark on a Hadoop cluster
- Optionally connect R with Drill and Spark

Ex. 1 — *Drill and Spark installation*

Download, install, and set up Drill and Spark on the Hadoop cluster from the previous lab. Optionally install drillR or Sergeant and SparklyR to use Drill and Spark from inside R.
Note: before installing Spark ensure Scala is properly working.

Ex. 2 — *Simple Drill queries*

The goal is now to test the Drill installation. The following questions can be completed directly in the Drill shell, running SQL queries, in R using DrillR or Sergeant dplyr interface, or using pydrill.

1. Using exercise 2 from lab 4, generate a file of size at least 5 GB, and copy it onto the Hadoop cluster.
2. Determine the name of the student who had the
   a) Lowest grade;
   b) Highest average score;
3. Calculate the median over all the scores.

Ex. 3 — *Simple Spark*

The goal is now to test the Spark installation. The following questions can be completed directly in the Spark shell, in R using SparkR or SparklyR interfaces, or using pyspark.

1. Create an RDD from the grade file generated in the previous exercise.
2. Apply a simple `flatMap()` transformation on the RDD to create pairs composed of the student ID and a grade.
3. Use the `reduce()` action to retrieve the maximum grade of each student.
4. Present some graphs and/or tables showing how the speed evolves as the size of the file increases. Compare the results between MapReduce (lab 2 and homework 3 exercise 1) and SparkR.