VE472 — Methods and tools for big data

Lab 5

Manuel — UM-JI (Summer 2022)

Goals of the lab

- Scale up a cluster
- Run a heavy computation over the cluster
- Observe the benefits of using Hadoop

Before moving on to the second part of the course where algorithms to efficiently analyse big data will be studied, we want to fully observe the benefits of Hadoop, Drill, and Spark and evaluate all the power they brought in term of data analysis.

Therefore the tasks in this lab are very simply explained:

- Connect all the clusters together to run a single large cluster over all the devices.
- Revisit homework 1, exercise 2.
  - Briefly describe the "large cluster" used in this lab;
  - Use Drill to perform all the database requests;
  - Take advantage of Spark to test a statistical model over the whole flight data set.
  - Using your understanding of the course, explain the benefits of a distributed system in the cases of I/O and compute bound jobs.