VE572 — Methods and tools for big data

Lab 3

Jing and Manuel — UM-JI (Summer 2018)

Goals of the lab

- Install Hadoop
- Setup a Hadoop cluster
- Run a simple test program

Ex. 1 — *Hadoop installation*

Download, install, and set up Hadoop.

Ex. 2 — *Simple Hadoop streaming*

The goal is now to test the Hadoop installation. This exercise can be completed using any programming language.

1. Write a short program that uses the lists of first-names and last-names to create a `csv` file where the first columns contains a list of students, the second a ten digit random student ID, and the third one a random grade in the range 0 to 100. Each students should appear a random number of times all along the file with different grades.
2. Write a short program which extracts the grades from the previous file and for each line outputs on the standard output a pair of values constructed as follows: studentID`<TAB>`grade, e.g. 1234567890 34. Name this program `mapper`.
3. Write a short program which reads pairs from the standard input. Each tab-separated pair is composed of a studendID and a list of grades. Return the max grade for each student on the standard output. Name this program `reducer`.
4. Copy the `cvs` file on HDFS and use Hadoop streaming to process it using the previous mapper and reducer programs.