VE572 — Methods and tools for big data

Project (part 2)

Jing and Manuel — UM-JI (Summer 2018)

Goals of the project

- Work inside/with Hadoop
- Write Drill storage plugins
- Visualise and present data
- Optionally explore the power of Spark

Preparation

This part of the project splits into a mandatory and an optional task. The first focuses on Drill and the second on Spark.

Note: any group investigating the Spark task should contact the teaching team as a larger cluster might be necessary. In such a case several groups could join their efforts, but *only* for this specific task. The Drill part must be completed per group.

Retrieve the task archive as well as the full *Million Song Dataset* from ve572 server at ji.csproject.org, port 2572, using login `motivatedstudent` and password `Iwanttoworkhardinve572`.

Mandatory tasks: Drill

The Drill queries to run are described in the task file.

Task 1: write a program which takes advantage of Drill storage plugin to read the `.btf` file and extracts the tasks. Translate them into SQL queries and runs them on the Million Song Dataset using Drill.

Task 2: draw a map showing the location of all the artists.

Task 3: generate a graph showing how similar artists relate to each others.

Optional tasks: Spark

Task 4: write a Spark program to perform task 3 from project part 1, over the whole dataset, i.e. perform k-mean for k=1,2,3,4,5 over the whole Million Song dataset.

Task 5: create a subset of about 1000 songs that are, **without any question**, all of type Jazz. Build a statistical model which determines the specifics of Jazz music based on the segment analysis. Based on this model, list all the songs that are misclassified.