Project

Daniel Hagimont - Daniel.Hagimont@enseeiht.fr

USTH – Teaching Unit MI3.02

December 2018

The objective of this project is to implement an application scenario which illustrates the use of the following techniques :
- docker : it can be used to deploy a cluster of (virtual) machines on your laptop, but can also be used in a distributed setting.
- spark : a spark infrastructure, including hdfs, a master and several slaves must be deployed in the docker infrastructure.

You should select and implement an application scenario in the Spark environment. This application scenario should be realistic.
You should evaluate the benefits from parallelizing such an application. You can run it locally (sharing the file system, benefiting from CPU parallelism) or distributed on several laptops (for the final evaluations, benefiting from IO parallelism). Your evaluations should demonstrate where the improvements (due ti parallelism) or non-improvements, come from. For instance, evaluations can be performed locally or distributed, with more or less computations.

The results should include:
- the source code and data (Java source code, scripts, Dockerfiles, samples, …)
- a performance analysis of parallelism (CPU and IO)
- a short report describing your achievements and allowing to reproduce your experiments on my laptop (with as less manipulations as possible, detail the pre-requisites in the report).

You should be careful, not to plagiarize resources on the internet, and not to have a project too similar to other projects from the same class.

The projet has to be realized as a pair.

Your achievements should be returned by email (hagimont@enseeiht.fr) before January 6th, 2019.