Project Infrastructure

Daniel Hagimont / Boris Teabe

The goal of this project is to build a service for the automatic deployment of a Big Data application in a virtualized environment. This service will rely on KVM with libvirt for providing virtual machines, Terraform for provisioning virtual machines, Ansible for deploying the Spark infrastructure and application.

You should first set up the physical environment. It will be composed of (at least) one server (a laptop) which runs a natively installed Ubuntu system. KVM is used for virtualization and libvirt for enabling the provisionning of virtual machines.

Based on this physical environment, you should use Terraform for provisionning VMs (at least 3 VMs) and Ansible for the deployment of a Spark/HDFS infrastructure which will run the Spark application.

The final result can be a script which takes as parameters the application jar to execute, the datafile to process and the number of slaves VMs you want. Running this script from a client machine will provision VMs, deploy the Spark infrastructure on these VMs, upload data, run the application, download the result, and deprovision VMs.

There are different technologies and tasks (KVM/libvirt, Terraform, Ansible, Spark) that you can distribute between the members of your group. Try to be efficient and not to rely on one student.

You must provide at the end :

- a report (PDF) describing your achievements
- a video presenting a demonstration of the services

A download link must be sent to both of us (including the report and the video). The deadline is october 21st 2024 (11pm Hanoi time).