Projet Big Data Infrastructure

Contact : Daniel Hagimont

hagimont@enseeiht.fr

http://hagimont.perso.enseeiht.fr/resources-N7/certificat/certificat.html

Projet Big Data Infrastructure

- Objectif
  - Installer et utiliser Spark/HDFS en cluster
  - Evaluer sa scalabilité
    - On ajoute des machines, on baisse le temps d’exécution
- Condition du projet
  - En trinomes
  - Support en ligne : slack / discord
  - Un CR (email / 5 lignes) par semaine
  - Deadline : 31/01/2021

2 options au choix

- Option Docker
  - Plutot pour les non informaticiens
  - Le cluster est simulé sous la forme de containers Docker sur une machine (Linux)
  - Chaque container travaille avec un seul coeur
  - On étudie la scalabilité en augmentant le nombre de containers
  - Peu d’installation, les containers Docker incluant Spark/HDFS sont deja installé
  - Plutot en mode observation

2 options au choix

- Option AWS
  - Plutot pour les informaticiens
  - Vous utilisez un compte AWS perso
    - Vous avez droit à 750h de calcul gratuit
  - Vous avez accès à un TP décrivant le déploiement de Spark/HDFS en mode cluster
  - Vous devez tout installer dans des machines virtuelles AWS
  - On étudie la scalabilité en augmentant le nombre de machines virtuelles