

Rapport Projet Données Réparties

- **Etude de scalabilité:**

- On utilisera 1 NameNode (situé dans la machine dragon), 3 serveurs (minotaure, succube et hippogriffe) et 3 workers (situés dans les mêmes machines que les serveurs).
- On étudiera l'influence de la taille du fichier sur la vitesse de l'exécution.
- On utilisera le script generate.sh (./generate.sh n) pour générer un fichier contenant 2^n fois le fichier filesample.txt
- On obtient le résultat suivant

n	5	10	15	20
taille	41,8 Ko	1,3 Mo	42,8 Mo	1,4 Go
Séquentielle	36ms	111ms	830ms	21000ms
Réparties	520ms	562ms	765ms	7605ms

- **Corrections apportées pour obtenir la version opérationnelle:**

- Les fichiers sont stockés dans le dossier /tmp pour éviter des dépassements de quota lors de l'utilisation sur les machines de l'enseeiht.
- Correction dans HdfsClient : lors de la commande write, les lignes du fichier ne sont plus stockées dans une variable pour éviter une erreur de Java (OutOfMemoryError). On les envoie directement après la lecture.

- **Outils pour le déploiement:**

- **data/generate.sh n:** générer un fichier contenant 2^n fois le fichier filesample.txt pour obtenir des fichiers de grandes tailles.
- **src/lancerNameNode.sh utilisateur:** permet de se connecter en ssh à dragon.enseeiht.fr et lancer le NameNode sur cette machine. Pour changer de machine, on doit changer le fichier de configuration (config/nameNode.xml) et changer ce script pour se connecter à la même machine. On suppose que le dossier hidoop se trouve dans le répertoire courant.

- **src/lancerServeur1.Sh utilisateur:** se connecte à la machine correspondante au serveur et lance HdfsServer. Pour changer la machine on doit changer le fichier de configuration et la machine sur laquelle on se connecte dans le script. On suppose que le dossier hidoop se trouve dans le répertoire courant.
- **src/lancerWorker1.Sh utilisateur:** se connecte à la machine correspondante au worker et lance HdfsServer. Pour changer la machine on doit changer le fichier de configuration et la machine sur laquelle on se connecte dans le script. On suppose que le dossier hidoop se trouve dans le répertoire courant.
- Pour ajouter un serveur ou un worker, on doit créer le fichier de configuration et le script correspondants.

- **Améliorations envisagées:**
 - Envoyer les données directement et ne plus créer de fichier intermédiaire.
 - Permettre plusieurs reducers.

- **Application choisie pour évaluer la version améliorée:**
 - *QuasiMonteCarlo.java* : calcul de π par une méthode de Monte-Carlo.