# ITÉRATION 4

## Utilisation de système de fichiers distribués

### Modalités

- Travail individuel en autonomie
- 1 journée en présentiel

### Livrables

Le système de fichier installé
Le job adapté sur le cluster spark avec les données dans le système de fichiers

### Objectifs

Installer HDFS sur une ou plusieurs machines pour stocker des quantités de données importantes

### Compétences

- Je sais installer l'environnement Hadoop et Spark sur ma machine
- Je sais écrire un job qui utilise un système distribué (HDFS + Spark) pour traiter et stocker des données

## 4.0 — Le fameux stagiaire

2m — Présentiel

Au détour d'une pause café vous remarquez une tête qui vous est familière, c'est Edmund, votre stagiaire sauveur, qui est passé faire coucou une dernière fois avant de retourner travailler dans l'entreprise très connue dont il ne faut pas prononcer le nom.

Vous décidez de vous confier car il semble tout de même avoir mûri professionnellement depuis qu'il est parti de l'entreprise. Vous lui racontez vos péripéties de la semaine entre Stan et Spark, et les différents petits problèmes que vous avez pu rencontrer au détour de votre montée en compétence expresse.

Au moment où vous lui avez parlé de quelques Giga de données sur le système de fichier "bateau" de votre machine, vous avez cru qu'il s'évanouissait! Mais il a tenu le choc et vous a laissé terminer. Au cours de la discussion où vous cherchiez à trouver une nouvelle blague à faire sur l'ordinateur de Stan (qui oublie de verrouiller son ordinateur aussi souvent que d'amener les croissants), Edmund vous demande pourquoi vous n'utilisez pas un système de fichiers distribué comme HDFS de Hadoop, qui permettrait notamment de stocker les fichiers d'entrée et de sortie de façon distribuée et redondante.

Ceci vous paraît être une excellente idée, vous interrompez votre pause café quotidienne pour vous renseigner sur cette pépite que Edmund vous a encore dégotée. Vous vous dites "Décidément on aurait dû l'embaucher cet Edmund!"

### RESSOURCES

- Une vidéo d'introduction à HDFS: https://yewtu.be/watch?v=jFsYao4C3cw
- Votre requête google: https://gprivate.com/65whp

### COMPÉTENCES ASSOCIÉES

- Pause café
- Écoute active de stagiaire
- Recherche Google de termes obscures

---

## 4.1 — Récupération des briques logicielles

3h — Présentiel

Pour essayer cette technologie, une idée va être de mettre en place un premier système de fichier hdfs. Après une lecture rapide du manuel et des différentes ressources, vous décidez de mettre en place un “Single Node Cluster”.

Pour cela deux possibilités s’offrent à vous:

- Installer HDFS directement sur votre ordinateur
  HDFS étant aussi basé sur java, il n’y aura pas besoin de réinstaller java pour l’installer.
- Utiliser docker pour lancer HDFS

Choisissez des deux possibilités

- Installez une Instance de HDFS
- Vérifiez sa configuration
- si vous pouvez accéder aux différents outils graphique (WebUI) et au système de fichier en ligne de commande passez à la suite.

### RESSOURCES

- Documentation officielle de hdfs:
  https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html
- Tutoriel de la documentation officielle:
  https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html
- Tutoriels de mise en place de hdfs sur ubuntu:
  https://www.edureka.co/blog/install-hadoop-single-node-hadoop-cluster
  https://tecadmin.net/install-hadoop-on-ubuntu-20-04/
- Lien vers un github qui contient un docker compose contenant hdfs (et un peu plus):
  https://github.com/big-data-europe/docker-hadoop
- Un dossier avec un docker compose tout configurer pour hdfs:
  https://github.com/gchq/gaffer-docker/tree/develop/docker/hdfs
- Hadoop page de téléchargement :
  https://hadoop.apache.org/releases.html

### COMPÉTENCES ASSOCIÉES

- Installation de HDFS

---

## 4.2 — Les fichiers dans HDFS

1h — Présentiel

En utilisant l'outil en  ligne de commande fourni par hdfs:

- affichez les fichiers de la racine du système hdfs
- créez un dossier nommé data_test
- regarder les manuels des commandes put et get de hdfs.
- ajoutez le csv sur les arbres de paris dans le dossier data_test
- afficher le contenu du dossier data_test, faites de même pour le contenu d'arbre de paris
- que constatez vous?
- ajoutez le fichier tar.gz des livres dans hdfs, que constatez vous?
- supprimez ce fichier
- utiliser les modes récursif des commandes d'ajout pour ajouter tous les livres.

**RESSOURCES**

- Doc officielle de la cli hadoop/hdfs : https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HDFSCommands.html
- CheatSheet de la cli HDFS: https://dzone.com/articles/hdfs-cheat-sheet

**COMPÉTENCES ASSOCIÉES**

- Utilisation de l'utilitaire en ligne de commande de HDFS

---

## 4.3 — Le Fameux job sur les textes

3h — Présentiel

- Adaptez le job pour prendre en entrées les répertoires voulus de HDFS
- Stockez des résultats intermédiaires dans HDFS
- Stockez les résultats finaux dans HDFS

**RESSOURCES**

- Doc officielle de la cli hadoop/hdfs : https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HDFSCommands.html
- Doc spark sur la lecture de fichier (marche sur HDFS)
  https://spark.apache.org/docs/3.4.1/api/python/reference/api/pyspark.SparkContext.wholeTextFiles.html
  https://spark.apache.org/docs/3.4.1/api/python/reference/api/pyspark.SparkContext.textFile.html
  https://spark.apache.org/docs/3.4.1/api/python/reference/pyspark.mllib.html#clustering

**COMPÉTENCES ASSOCIÉES**

- Combiner HDFS et Spark

---

## 4.4 — Un peu de clustering…

∞h — Présentiel

Nous allons maintenant chercher à faire du clustering sur les textes:

- Effectuer votre pré processing sur les textes afin de pouvoir stocker un résultat par texte dans hdfs
- Utilisez une méthode de clustering de spark mllib pour créer des cluster parmis les livres prétraités
- Stockez le résultat du clustering dans hdfs

**RESSOURCES**

- https://spark.apache.org/docs/3.4.1/api/python/reference/pyspark.mllib.html#clustering

**COMPÉTENCES ASSOCIÉES**

- Combiner HDFS et Spark
- Spark MLLib on RDD.

### Livrables

À la fin de l'itération vous devez avoir:

➔ Le système de fichier HDFS configuré et opérationnel (en single node)
➔ Un job spark qui utilise HDFS