ITÉRATION 1

Installation de Spark

Modalités

- Travail individuel en autonomie
- ½ journée en présentiel

Livrables

L'environnement Spark est bien installé sur votre machine.

Objectifs

Installer Spark sur votre machine. Référez-vous aux ressources fournies ci-dessous.

Compétences

- Je sais installer l'environnement Spark sur ma machine

1.0 — Un peu de motivation

2m — Présentiel

Vous venez de croiser Henri dans les couloirs d'un étage environ égal au deuxième sous-sol. Henri est l'administrateur système de votre service. Après vous avoir annoncé à votre grand étonnement qu'il part en vacances ce matin, vous avez réussi à lui soutirer une information utile, il n'y a pas encore d'environnement Spark sur les machines de votre service et vous allez devoir l'installer.

COMPÉTENCES ASSOCIÉES

- Discussion entre gorilles

N'oubliez pas de noter dans un fichier les commandes que vous avez utilisées dans les étapes suivantes.
Ces notes vous feront gagner un temps précieux dans la suite du module.

1.1 — Pré Requis avant l'installation de PySpark sur Ubuntu

0.3h — Présentiel

Le framework Apache Spark est écrit dans le langage Scala, un langage de programmation fonctionnel et orienté objet permettant entre autres de distribuer des calculs. Le langage Scala est basé sur Java. Pour utiliser PySpark, il va donc être nécessaire d'installer le *"Java Development Kit"* ou **jdk**.

- Inspirez-vous de la ressource proposée ci-dessous pour installer le jdk

RESSOURCES

- Tutoriel bien plus avancé que ce dont vous aurez besoin (**focalisez vous sur la première option: "*Option 1 — Installing the Default JRE/JDK*" pour l'installation de JRE et JDK**): https://www.digitalocean.com/community/tutorials/how-to-install-java-with-apt-on-ubuntu-22-04

COMPÉTENCES ASSOCIÉES

- Installer un paquet dans ubuntu avec apt.

1.2 — Environnement virtuel et PySpark

0.2h — Présentiel

Maintenant que java est installé,

- Avec votre gestionnaire de Python préféré (ex: conda), créez un environnement virtuel (vous pourrez lui donner un nom évocateur tel que "bigdata"). Vous pourrez essayer de fixer la version de Python que vous voulez utiliser dans ce projet lors de la création de l'environnement (choisissez par exemple la Python 3.10).
- Activez cet environnement virtuel
- Pour finir, installez le paquet `pyspark` dans ce nouvel environnement virtuel

RESSOURCES

- Environnements virtuels avec conda (recommandé si vous avez déjà Anaconda ou Miniconda d'installé) :
  - installation de miniconda: miniconda documentation
  - Conda - managing environments
- Environnements virtuels avec venv dans python:
  - venv — Creation of virtual environments — Python 3.10 documentation
- Différences entre Spark et PySpark:
  - https://www.ksolves.com/blog/big-data/spark/pyspark-vs-spark-lets-unravel-the-bond
  - https://stackoverflow.com/questions/51728177/can-pyspark-work-without-spark

COMPÉTENCES ASSOCIÉES

- Créer un environnement virtuel Python
- Installer un paquet dans un environnement virtuel

1.3 — Test de l'installation de PySpark

0.5h — Présentiel

Maintenant que PySpark est installé, le moment est venu de voir s'il est fonctionnel

Pour cela vous avez deux possibilités :

- Lancer un interpréteur python puis :
  - depuis `pyspark` importer `SparkContext`
  - créer l'objet `SparkContext` avec comme paramètres `"local"` et `"test"`, stockez cet objet dans une variable nommée `sc`.
  - vérifier que l'interface web de spark est bien démarrée.
    (elle devrait se situer là : http://localhost:4040)
  - utiliser la fonction `stop` de l'objet `sc`.

Ou(/Et):

- Dans un terminal Linux :
  - lancer directement pyspark dans un terminal linux
  - vérifier que l'interface web de spark est bien démarrée (elle devrait se situer là : http://localhost:4040)
  - stopper le shell pyspark

Essayez de tester les deux possibilités! Lorsque vous avez fini cette partie, informez vous sur l'objet SparkContext à l'aide des ressources proposées ci-dessous.

RESSOURCES

- Spark Web UI: https://spark.apache.org/docs/latest/web-ui.html

- Documentation officielle du SparkContext : https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.SparkContext.html

COMPÉTENCES ASSOCIÉES

- Première prise en main de pyspark