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 en début de Kit.

Compétences

- Je sais installer l'environnement Spark sur ma machine

1.0 — Un peu de motivation

2m — Présentiel

Vous venez de croiser dans les couloirs d'un étage supérieur à -2, Henri (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

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.

- Installer le paquet `openjdk-8-jdk` ce paquet contient la version 8 de java (si il n'est pas présent, essayez avec la version 11 ou la version 19)

RESSOURCES

- Tutoriel bien plus avancé que ce dont vous avez besoin: 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 (conda), créez un environnement virtuel (vous pourrez lui donner un gros nom comme "big_data",...)
- Activez cet environnement virtuel

- Installez le paquet `pyspark` dans ce nouvel environnement virtuel

RESSOURCES

- Environnements virtuels avec conda (recommandé) : https://docs.conda.io/projects/conda/en/latest/user-guide/tasks/manage-environments.html
- Environnements virtuels avec venv dans python (un peu plus complexe) : https://docs.python.org/3.10/library/venv.html

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

- 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

(Il est recommandé de tester les deux possibilités)

Lorsque vous avez fini cette partie, informez vous sur l'objet SparkContext.

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

- Installation de pyspark

1.4 — (Bonus) Ajout de Jupyter dans la configuration pyspark 0.5h — Présentiel

Il est aussi possible de configurer pyspark pour lancer automatiquement un notebook jupyter contenant le SparkContext déjà instancié.

Pour cela vous devrez:

- Configurer deux variables d'environnements
- Vérifier que pyspark lance directement un notebook
- Vérifier que pyspark fonctionne dans le notebook
- S'il ne fonctionne pas, installer `findspark` et réessayer les étapes précédentes.

RESSOURCES

- Installation de pyspark dans jupyter: https://sparkbyexamples.com/pyspark/install-pyspark-in-anaconda-jupyter-notebook/
- Documentation de findspark (lire le readme): https://github.com/minrk/findspark

COMPÉTENCES ASSOCIÉES

- Configuration avancée de pyspark et jupyter
