Ce projet est cofinancé par l'Union européenne
L'EUROPE S'ENGAGE en région Auvergne-Rhône-Alpes avec le FSE
La Région Auvergne-Rhône-Alpes

OBJECTIFS

Objectifs pédagogiques

En tant que data scientist, certains problèmes nécessitent une intensité de calcul supérieure à ce que proposent Python et Pandas. Ce module a pour objectif de vous initier au Big Data au travers de l'environnement Apache Spark. Vous serez initiés à la répartition des calculs en utilisant Apache Spark, à la création de flux de données et à la répartition des données. Le projet permettra de parcourir les différentes technologies et aspects du big data en utilisant une base de textes de différentes manières.

Compétences développées :

- Installation d'un environnement Spark et Big Data
- Premières manipulations avec les listes et un texte
- Utilisation de DataFrame Spark
- Création de flux de données
- Utilisation de système de fichiers distribué.

Démarche pédagogique (projet, ressources ...)

Pour partir à la découverte du Big Data on va réaliser 3 projets :

- Itération 1 :
  - Installation de Spark
- Itération 2 :
  - Premières manipulations sur des listes puis sur du texte !
- Itération 3 :
  - On ajoute quelques textes.
- Itération 4
  - Manipulation de flux de données.
- Itération 5
  - Utilisation de systèmes de fichiers distribués

Compétences

Itération 1 : installation

- Installation d'un environnement Spark

Itération 2 : petites découvertes

- Premières manipulations sur les listes puis sur un texte

Itération 3 : projet concret sur des textes

- Mise en place et utilisation d'un cluster spark local.

Itération 4 : suite du projet

- Mise en place du traitement de flux

Itération 5 : finalisation

- Utilisation d'un système de fichier distribué
- Création d'un mini Dashboard

MODALITÉS

Durée

5 jours soit 35 heures au total.
Lancement le 24/07/2023 et clotûre le 28/07/2023.

Formateur(s)

Jérémie Suzan, référent module
Marie Thieulin

TRAME

		Planning		Jour	Sujet	Activités
24/07	Big Data	Marie	Marie	1	Sujet 1 Sujet 2	Installation de l'environnement Spark Découverte de Spark à travers un exemple
25/07	Big Data	Marie	Marie	2	Sujet 2	Découverte de Spark sur du traitement de texte
26/07	Big Data	Jérémie	Jérémie	3	Sujet 3	Traitement sur plusieurs livres Mise en place d'un cluster spark
27/07	Big Data	Jérémie	Jérémie	4	Sujet 4	Mise en place du traitement de flux
28/07	Big Data	Jérémie	Jérémie	5	Sujet 5	Utilisation de système de fichiers distribué Mini dashboard

ITÉRATION 0

Un peu de contexte

Modalités

- Travail individuel en autonomie
- $^1/_{42}$ journée en présentiel

0.0 — Encore un lundi matin!

$^1/_{42}$ j — Présentiel

Comme tous les lundis matin, vous venez d'arriver au bureau et de finir le daily avec votre équipe. La semaine s'annonce tranquille, votre backlog comprend les quelques tâches que Stan (le data *titrepompeux* probablement meilleur en blagues qu'en code) n'a pas pu finir, plus deux trois nouveautés.

Comme souvent pendant l'heure post-daily du lundi, vous commencez votre veille data de la semaine sur les différents sites à la mode,... mais en tombant sur un article les techniques de traitement de données en tant réel, pour une organisation au logo comprenant un canari bleu, vous dérivez naturellement sur des vidéos de chats qui font du piano... La semaine s'annonce vraiment tranquille.

Quelque chose se met à faire un bruit ayant une similarité plutôt élevée au "DRING" du téléphone utilisé à une certaine époque. Surpris de trouver un téléphone sur votre bureau, vous répondez et :
c'est Maddie, la responsable marketing/produit du groupe. Vous ne la connaissiez pas, elle a entendu parler de l'Intelligence Artificielle récemment (merci chatgpt!) et a enclenché une dizaine de comités et de réunions afin de convaincre le management et Stan de créer le Shakespeare du "Quotidien temps" , le groupe médiatique pour lequel vous travaillez.

Stan (en utilisant son titre pompeux) a réussi à convaincre Maddie et la hiérarchie qu'avant de recréer Shakespeare, il serait plus judicieux de leurs montrer ce que l'équipe est capable de faire. Il vous a négocié une deadline très large (selon lui) pour prendre en main ces données et montrer

votre talent.

Vous avez une semaine.

Les tâches que Stan vous a donné si gentiment sont donc mises en pause…. L'équipe est inquiète : il y a beaucoup de contraintes sur les données à traiter, il semblerait que la Volumétrie, la Vitesse de traitement nécessaire et la Variété des données soient un peu plus importantes que les habitudes de l'équipe.

De façon nostalgique vous repensez à vos pauses café des derniers mois avec les stagiaires qui ont maintenant terminé et la… PAF ! Un nom donné par Edmund lors d'une des pauses vous apparaît comme par magie "Apache Spark", vous le tapez naïvement sur internet, et… cela vous parait drôlement adapté aux besoins du nouveau projet improvisé. En plus il existe une partie en Python : PySpark. "Mais que ferait-on sans Edmund ?!".

Après une pause café bien méritée, vous décidez de prendre une journée pour prendre en main ce nouveau jouet. La suite de la semaine s'annonce surprenante !

RESSOURCES

- Qu'est ce que le Big Data:
  https://fr.wikipedia.org/wiki/Big_data
  https://upload.wikimedia.org/wikipedia/commons/e/ee/Big_Data.png

COMPÉTENCES ASSOCIÉES

- Compréhension du jargon de Stan
- Veille technique passive en écoutant Edmund