PROJECT 8

QUESTIONS :

* Je dois faire quoi en plus dans mon script à part le PCA? d’un traitement de diffusion des poids du modèle Tensorflow sur les clusters (broadcast des “weights” du modèle) qui avait été oublié par l’alternant. Vous pourrez vous appuyer sur l’article “Model inference using TensorFlow Keras API (en anglais)” disponible dans les ressources. C’est pas déjà fait ?
* K=10 → explained variance ratio: [0.10140967559306546,0.08004581517922967,0.06344774721367424,0.050141178300393364,0.035335157853023255,0.029163995040856393,0.027726724002401414,0.02285213787857822,0.019861126212198935,0.019093213598319357]. What k should I use ?
* Difference entre PCA et IncrementalPCA variances ok ? What does that mean ?

**Sélectionner les outils du Cloud permettant de traiter et stocker les données d'un projet Big Data conforme aux normes RGPD en vigueur afin de concevoir une application de qualité supportant le traitement de données massives.**

CE1 Le candidat a identifié les différentes briques d'architecture nécessaires pour la mise en place d'un environnement Big Data.

CE2 Le candidat a identifié les outils du cloud permettant de mettre en place l'environnement Big Data conforme aux normes RGPD en vigueur.

**Prétraiter, analyser et modéliser des données (en veillant à leur conformité RGPD) dans un environnement Big Data et en utilisant les outils du Cloud afin de concevoir une application sécurisée de qualité supportant le traitement de données massives.**

CE1 Le candidat a chargé les fichiers de départ et ceux après transformation dans un espace de stockage cloud conforme à la réglementation RGPD.

CE2 Le candidat a exécuté les scripts en utilisant des machines dans le cloud.

CE3 Le candidat a réalisé un script qui permet d'écrire les sorties du programme directement dans l'espace de stockage cloud.

**Réaliser des calculs distribués sur des données massives en utilisant les outils adaptés et en prenant en compte le RGPD afin de permettre la mise en œuvre d’applications à l’échelle.**

CE1 Le candidat a identifié les traitements critiques lors d'un passage à l'échelle en termes de volume de données.

CE2 Le candidat veille à ce que l’exploitation des données soit conforme au RGPD. Dans le cadre de ce projet :

* les données sont stockées, et les traitements sont réalisés, sur des serveurs situés sur le territoire européen.

CE3 Le candidat a développé les scripts s’appuyant sur Spark.

CE4 Le candidat s'est assuré que toute la chaîne de traitement est exécutée dans le cloud.