PROJECT 8

QUESTIONS :

* Les **images** du jeu de données initial ainsi que la sortie de la réduction de dimension (une matrice écrite sur un fichier CSV ou autre) disponible dans un espace de stockage sur le cloud.
* K=10 → explained variance ratio. Tu as environ un millier de features donc tu vas avoir de la perte d’information avec ton PCA, c’est normal.

S3 open to all so that the evaluator can go check it, normally restricted to the EMR cluster (special IAM rôle and the allowed users).

**Core node:** A node with software components that run tasks and store data in the Hadoop Distributed File System (HDFS) on your cluster. Multi-node clusters have at least one core node.

We could use task nodes instead of the core ones :

**Task nodes** are optional. You can use them to add power to perform parallel computation tasks on data, such as Hadoop MapReduce tasks and Spark executors.

**Task nodes** don't run the Data Node daemon, nor do they store data in HDFS.

Some Use cases are:

* You can use Task nodes for processing streams from S3. In this case Network IO won't increase as the used data isn't on HDFS.
* Task nodes can be added or removed as no HDFS daemons are running. Hence, no data on task nodes. Core nodes have HDFS daemons running and keep adding and removing new nodes isn't a good practice.

**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.