

Déployez un modèle dans le cloud

 openclassrooms.com/fr/projects/deployez-un-modele-dans-le-cloud/assignment

Vous êtes Data Scientist dans une très jeune start-up de l'AgriTech, nommée "**Fruits!**", qui cherche à proposer des solutions innovantes pour la récolte des fruits.

La volonté de l'entreprise est de préserver la biodiversité des fruits en permettant des traitements spécifiques pour chaque espèce de fruits en développant des robots cueilleurs intelligents.



Fruits!

Votre start-up souhaite dans un premier temps se faire connaître en mettant à disposition du grand public une application mobile qui permettrait aux utilisateurs de prendre en photo un fruit et d'obtenir des informations sur ce fruit.

Pour la start-up, cette application permettrait de sensibiliser le grand public à la biodiversité des fruits et de mettre en place une première version du moteur de classification des images de fruits.

De plus, le développement de l'application mobile permettra de construire une première version de l'architecture Big Data nécessaire.

Les données

Votre collègue Paul vous indique l'existence d'un jeu de données constitué des images de fruits et des labels associés, qui pourra servir de point de départ pour construire une partie de la chaîne de traitement des données.

Votre mission

Vous êtes donc chargé de développer dans un environnement Big Data une première chaîne de traitement des données qui comprendra le preprocessing et une étape de réduction de dimension.

Il n'est pas nécessaire d'entraîner un modèle pour le moment.

L'important est de mettre en place les premières briques de traitement qui serviront lorsqu'il faudra passer à l'échelle en termes de volume de données !

Contraintes

Lors de son brief initial, Paul vous a averti des points suivants :

- Vous devrez tenir compte dans vos développements du fait que le volume de données va augmenter très rapidement après la livraison de ce projet. Vous développerez donc des scripts en Pyspark et utiliserez par exemple le cloud AWS pour profiter d'une architecture Big Data (EC2, S3, IAM), basée sur un serveur EC2 Linux.
- La mise en œuvre d'une architecture Big Data sous (par exemple) AWS peut nécessiter une configuration serveur plus puissante que celle proposée gratuitement (EC2 = t2.micro, 1 Go RAM, 8 Go disque serveur).

Ce coût, qui devrait rester inférieur à 10 euros pour une utilisation raisonnée, reste à votre charge. L'utilisation d'un serveur local pour la conception, en limitant l'utilisation du serveur EC2 à l'implémentation et aux tests, permet de réduire sensiblement ce coût.

Livrables attendus

- Un **notebook** sur le cloud contenant les scripts en Pyspark exécutables (le preprocessing et une étape de réduction de dimension).
- 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.
- Un support de **présentation** pour la soutenance, présentant :
 - les différentes briques d'architecture choisies sur le cloud ;
 - leur rôle dans l'architecture Big Data ;
 - les étapes de la chaîne de traitement.

Pour faciliter votre passage devant le jury, déposez sur la plateforme, dans un dossier zip nommé "**Titre_du_projet_nom_prénom**", votre livrable nommé comme suit : **Nom_Prénom_n° du livrable_nom du livrable_date de démarrage du projet**. Cela donnera :

- *Nom_Prénom_1_notebook_mmaaaa*
- *Nom_Prénom_2_images_mmaaaa*
- *Nom_Prénom_3_presentation_mmaaaa*

Par exemple, votre premier livrable peut être nommé comme suit :

Dupont_Jean_1_notebook_12022.

Modalités de la soutenance

5 min - Rappel de la problématique et présentation du jeu de données




15 min - Présentation de la réalisation de la chaîne de traitement des images dans un environnement Big Data dans le cloud (à l'aide de votre support de présentation)

5 min - Conclusion et recommandations

5 à 10 minutes de questions-réponses

Votre présentation devrait durer 20 minutes (+/- 5 minutes). Puisque le respect des durées des présentations est important en milieu professionnel, les présentations en dessous de 15 minutes ou au-dessus de 25 minutes peuvent être refusées.

Compétences évaluées

-  Paralléliser des opérations de calcul avec Pyspark
-  Utiliser les outils du cloud pour manipuler des données dans un environnement Big Data
-  Identifier les outils du cloud permettant de mettre en place un environnement Big Data