PROJET À COMPLÉTER

Réalisez un traitement dans un environnement Big Data sur le Cloud

40 heures

Mis à jour le mercredi 29 mars 2023

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.

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.
Votre collègue Paul vous indique l'existence d'un document, formalisé par un alternant qui vient de quitter l'entreprise. Il a testé une première approche dans un environnement Big Data AWS EMR, à partir d'un jeu de données constitué des images de fruits et des labels associés (en

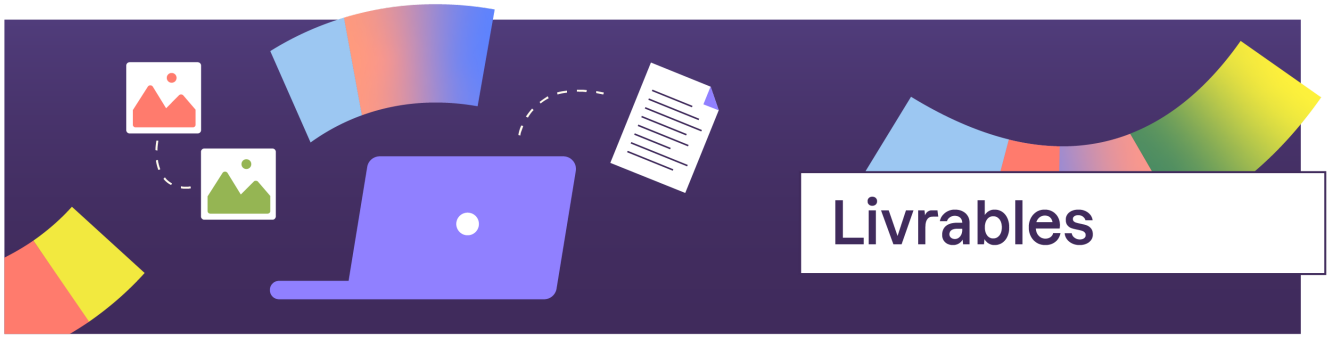
téléchargement direct à ce lien). Le notebook réalisé par l'alternant servira de point de départ pour construire une partie de la chaîne de traitement des données.

> Vous êtes donc chargé de vous approprier les travaux réalisés par l'alternant et de compléter la chaîne de traitement. 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 !

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 continuerez donc à développer des scripts en Pyspark et à utiliser le cloud AWS pour profiter d'une architecture Big Data (EMR, S3, IAM). Si vous préférez, vous pourrez transférer les traitements dans un environnement Databricks
- Vous devez faire une démonstration de la mise en place d'une instance EMR opérationnelle, ainsi qu' expliquer pas à pas le script PySpark, que vous aurez complété :
    - 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 "Distributed model inference using TensorFlow Keras" disponible dans les ressources
    - d'une étape de réduction de dimension de type PCA en PySpark
- Vous respecterez les contraintes du RGPD : dans notre contexte, vous veillerez à paramétrer votre installation afin d'utiliser des serveurs situés sur le territoire européen
- Votre retour critique de cette solution sera également précieuse, avant de décider de la généraliser
- La mise en œuvre d'une architecture Big Data de type EMR engendrera des coûts. Vous veillerez donc à ne maintenir l'instance EMR opérationnelle que pour les tests et les démos.

> 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 mise à jour du Script PySpark, en limitant l'utilisation du serveur EMR à l'implémentation et aux tests, permet de réduire sensiblement ce coût.



- Un **notebook** sur le cloud contenant les scripts en Pyspark exécutables (le preprocessing et une étape de réduction de dimension de type PCA).
- 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 ;
    - la démarche de mise en oeuvre de l'environnement Big Data (EMR ou Databricks)
    - les étapes de la chaîne de traitement PySpark.

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_012022.*

Pendant la soutenance, l'évaluateur jouera le rôle de Paul. Vous lui présenterez l'ensemble de votre travail.

- **Présentation (20 minutes)**
  - Rappel de la problématique et présentation du jeu de données (3 minutes),
  - Présentation du processus de création de l'environnement Big Data, S3 et EMR ou Databricks (6 minutes),
  - 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 (6 minutes),
  - Démonstration d'exécution du script PYSpark sur le Cloud (2 minutes),
  - Synthèse et conclusion (3 minutes).
- **Discussion (5 minutes)**
  - L'évaluateur vous challengera sur votre compréhension de concepts et techniques mis en oeuvre.
- **Débriefing (5 minutes)**
  - À la fin de la soutenance, vous pourrez débriefer ensemble.

> 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.

Référentiel d'évaluation

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 Vous avez identifié les différentes briques d'architecture nécessaires pour la mise en place d'un environnement Big Data.
CE2 Vous avez 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 Vous avez chargé les fichiers de départ et ceux après transformation dans un espace de stockage cloud conforme à la réglementation RGPD.
CE2 Vous avez exécuté les scripts en utilisant des machines dans le cloud.
CE3 Vous avez 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 Vous avez identifié les traitements critiques lors d'un passage à l'échelle en termes de volume de données.
CE2 Vous avez veillé 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 Vous avez développé les scripts s'appuyant sur Spark.
CE4 Vous vous êtes assurés que toute la chaîne de traitement est exécutée dans le cloud.

Compétences évaluées

- Modéliser des données dans un environnement Big Data et en utilisant les outils du Cloud
- Réaliser des calculs distribués sur des données massives en utilisant les outils adaptés
- Sélectionner les outils du Cloud permettant de traiter et stocker des données Big Data


Télécharger dans
l'App Store