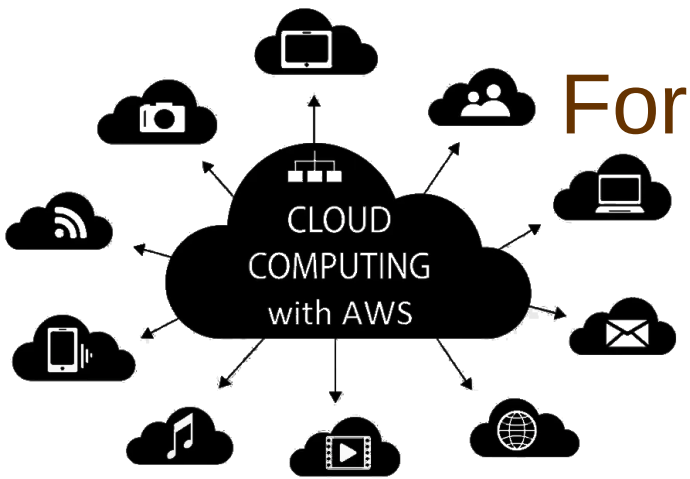
Déployez un modèle dans le cloud



Formation Data scientist
Projet n°8
Bayram DONAT

–

Fruits!

Sommaire

- Problématique
- Jeux de données
- Environnement Big Data dans le cloud
- Chaîne de traitement des images

Problématique

- Contexte
  - Entreprise : "Fruits!"
  - Activité : solutions innovantes pour la récolte des fruits : robots cueilleurs intelligents.
  - Besoins :
    - Une application mobile qui fournit des informations sur le fruit pris en photo
    - Un moteur de classification des images de fruits.
    - Une architecture Big Data nécessaire.
- Mission :
  - dans un environnement Big Data, développer une chaîne de traitement des données
    - le preprocessing
    - une étape de réduction de dimension.
- Contraintes
- Livrables

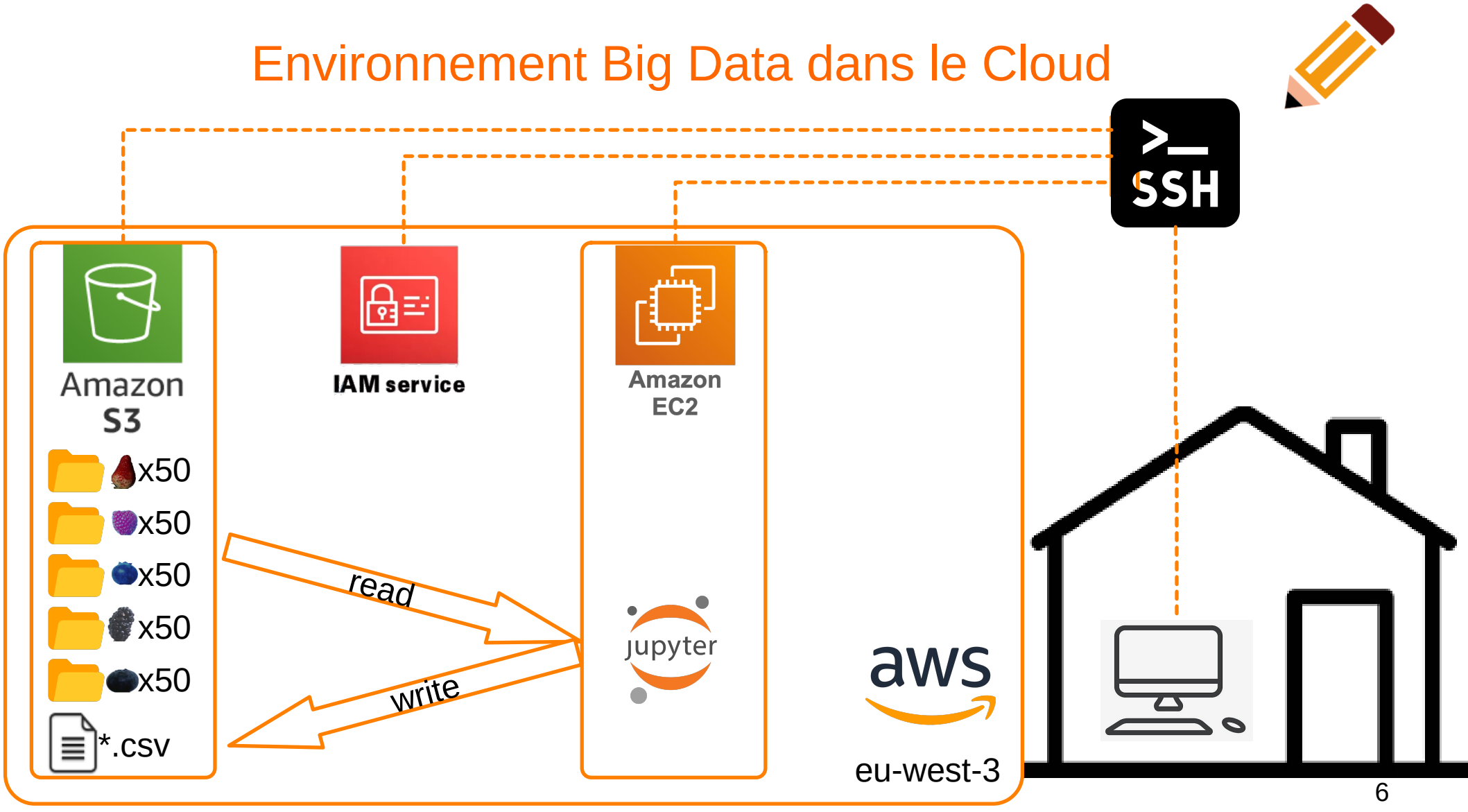
Problématique

- Contraintes
  - le volume de données va augmenter très rapidement
  - Utiliser des scripts en Pyspark
  - Utiliser le cloud AWS avec architecture Big Data (EC2 linux, S3, IAM)
- Livrables
  - Un notebook avec les scripts sur le cloud
  - Les images du jeu de données initial
  - La sortie de la réduction de dimension sur le cloud.

Jeux de données

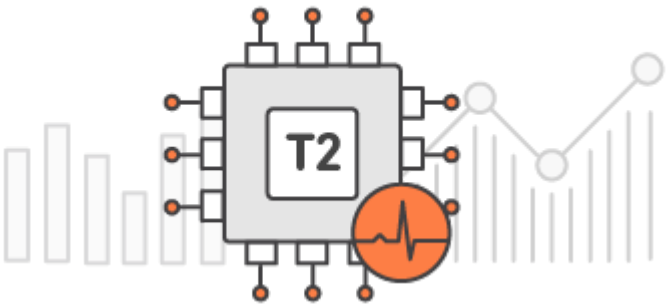
- des images de fruits avec des labels associés
  - 131 Types de fruits avec un repertoire par fruit
  - au minimum 450 photos du fruit de taille 100x100
- Pour des raisons de cout (taille mémoire)
  - 5 fruits rouges (fraise, framboise, huckleberry, mûre, myrtille)
  - 50 photos par fruits

Environnement Big Data dans le Cloud



Environnement Big Data dans le Cloud

- Serveur EC2 :
  - Image logicielle (AMI) : ubuntu 22.04 LTS server 64 bits
  - Type t2.large avec 2 vCPU, 8Go RAM, 20GO SSD, crédits de 36 CPU/heure, prix horaire de 0,0928 $ US
- Bucket S3 : Stockage de fichiers
- Identity and Access Management (IAM) : gestion des droits utilisateurs



Amazon Machine Image (AMI)

ubuntu

Chaine de traitement des images

- Récupération des fichiers jpg
- Mise en tableau avec vecteurs
  - pyspark
    - Ajout de la catégorie du fruit (répertoire)
    - Ajout vecteur redimensionné en 224x224
    - Ajout features de VGG16
    - Ajout features standardisés de VGG16
- PCA sur features standardisés de VGG16 (k=50)
  - Réduction de dimensions
- Sauvegarde sur S3 (*.paquet)

Recommandations

- Que reste-t-il à faire?
  - Une classification kmeans à partir des pca des features de VGG16
  - Projection sur 2 dimensions en PCA ou TSNE
  - Superposition des clusters kmeans et des classes

  - comme le projet 6 : classification des biens de consommation
- Anticiper l'évolution du serveur EC2 en fonction des prévisions des données stockées sur S3

Recommandations

- L’utilisation de spark fait gagner du temps d’exécution
- L’utilisation de spark consomme la RAM notamment en transformant le tableau sql en tableau pandas