



DÉPLOYEZ UN
MODÈLE DANS LE
CLOUD

DATA SCIENCE

SOMMAIRE

01

02

03

04

SOMMAIRE

Présentation du projet

01

02

03

04

SOMMAIRE

Présentation du projet

01

Création de
l'environnement cloud

02

03

04

SOMMAIRE

Présentation du projet **01**

Création de
l'environnement cloud **02**

Traitement des images **03**

04

SOMMAIRE

Présentation du projet **01**

Création de
l'environnement cloud **02**

Traitement des images **03**

Conclusion **04**

PRESENTATION DU PROJET



Contexte

- Proposition des solutions innovantes pour la récolte des fruits.
- Mise en place d'une application de classification des images de fruits.
- Application mobile sera la première version de l'architecture Big data.

PRESENTATION DU PROJET

Objectifs

- Développer le traitement des données : le preprocessing et la réduction de dimension..
- Déployer le traitement des données dans un environnement **Big Data**



Contraintes

- Tenir compte de l'augmentation rapidement des données après la livraison de ce projet
- Paramétrer l'installation afin d'utiliser des serveurs situés sur le territoire européen

PRESENTATION DU PROJET

Méthodologie

- Présentation du jeu de données,
- Création de l'environnement Big Data
- Réalisation de la chaîne de traitement des images dans un environnement Big Data
- Démonstration d'exécution du script PySpark sur le Cloud
- Synthèse et conclusion



RAPPEL DE LA PROBLÉMATIQUE ET PRÉSENTATION DU JEU DE DONNÉES

Présentation du jeu de données

Nous avons un jeu de données de milliers photos images :

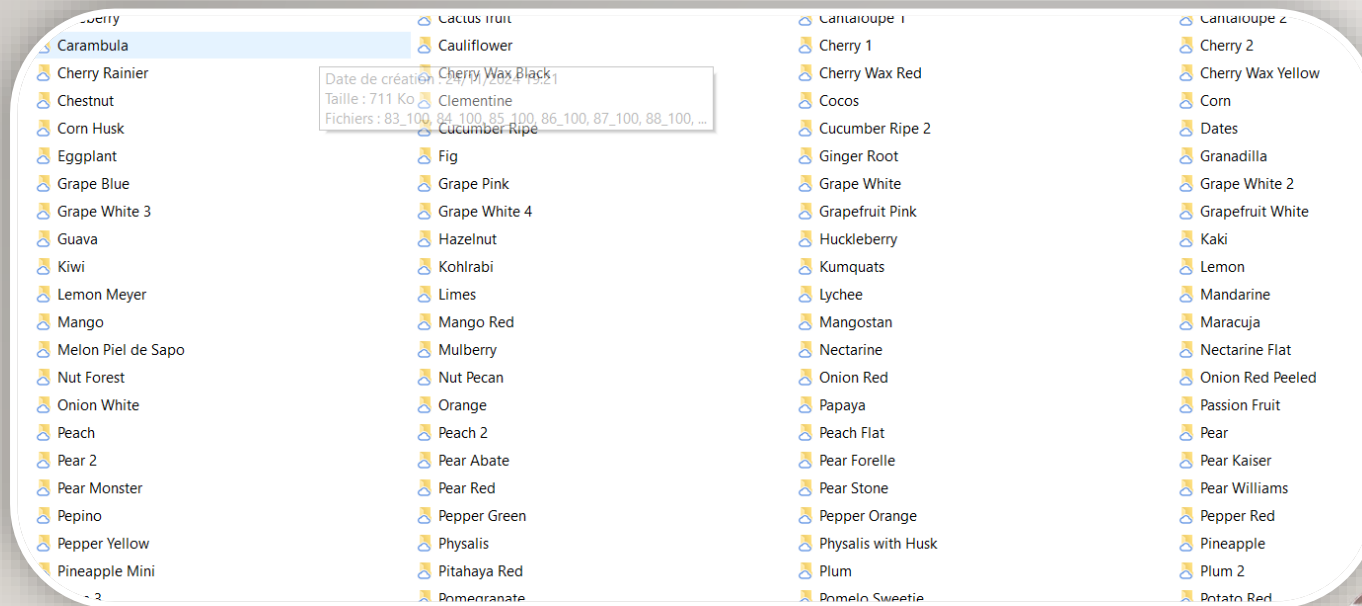
- ❑ Le jeu de données est un répertoire d'espèces de fruit,
- ❑ Chaque espèce de fruit est composé de plusieurs labels (63 labels au total)
- ❑ Chaque label est composé de plus d'une centaine de photos du fruits prises sur plusieurs angles.



RAPPEL DE LA PROBLÉMATIQUE ET PRÉSENTATION DU JEU DE DONNÉES

Présentation du jeu de données

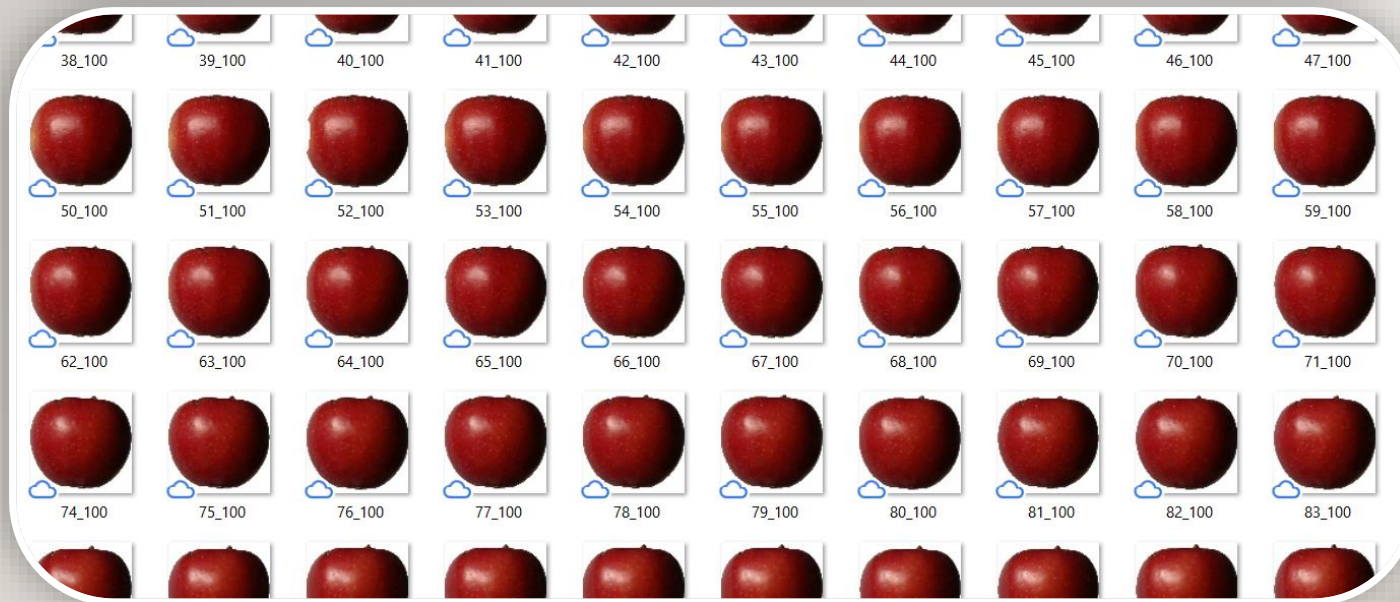
Liste de Labels



RAPPEL DE LA PROBLÉMATIQUE ET PRÉSENTATION DU JEU DE DONNÉES

Présentation du jeu de données

Photos du fruit Apple Brueburn



CRÉATION DE L'ENVIRONNEMENT BIG DATA

Le choix technique

Amazon Web Services

Amazon Web Services (AWS) est le prestataire qui offre à ce jour l'offre la plus large dans le cloud computing.

Certaines de leurs offres sont parfaitement adaptées à notre problématique.

L'objectif premier est de pouvoir, grâce à AWS, louer de la puissance de calcul à la demande :

- ❑ Le **service EMR** permet de louer des **instances EC2** avec des applications préinstallées et configurées,
- ❑ **Amazon S3**, est une solution très efficace pour la gestion du stockage des données.

N.B : Nos services ont été paramétrés sur la région de Paris



CRÉATION DE L'ENVIRONNEMENT BIG DATA

Configuration de l'environnement de travail

Amazon S3

Sur Amazon S3, qui est la Solution de stockage des données, nous allons créer un bucket.

A l'intérieur du bucket :

- ☐ Uploader nos données images à traitées,
- ☐ Uploader notre fichier Bootstrap.sh
- ☐ Uploader le script PySpark exécutable
- ☐ Enregistrer le résultat de notre travail.

Objets (5) Info						
Les objets sont les entités fondamentales stockées dans Amazon S3. Vous pouvez utiliser l' inventaire Amazon S3 pour obtenir une liste de tous les objets de votre compartiment. Pour que d'autres personnes puissent accéder à vos objets, vous devez leur accorder explicitement des autorisations. En savoir plus						
Rechercher des objets en fonction du préfixe						
<input type="checkbox"/>	Nom	▲	Type	▼	Dernière modification	▼
<input type="checkbox"/>	bootstrap.sh		sh		30 Dec 2024 01:40:46 PM GMT	533.0 o
<input type="checkbox"/>	Data/		Dossier		-	-
<input type="checkbox"/>	pca.py		py		30 Dec 2024 04:04:16 PM GMT	10.7 Ko
<input type="checkbox"/>	Preprocessing/		Dossier		-	-
<input type="checkbox"/>	Results/		Dossier		-	-



CRÉATION DE L'ENVIRONNEMENT BIG DATA

Configuration de l'environnement de travail

Configuration du serveur EMR

La configuration va consister à la création d'un cluster; qui s'effectuera en plusieurs étapes :

- ☐ Nom et applications
- ☐ Configuration de cluster
- ☐ Dimensionnement du cluster
- ☐ Actions d'amorçage
- ☐ Sécurité et autorisations
- ☐ Rôle Identity and Access Management (IAM)
- ☐ Créer un cluster



CRÉATION DE L'ENVIRONNEMENT BIG DATA

Configuration du serveur EMR


Nom et applications


▼ Nom et applications - requis [Info](#)
Donnez un nom à votre cluster et choisissez les applications que vous voulez y installer.


Nom


Version Amazon EMR [Info](#)
Une version contient un ensemble d'applications susceptibles d'être installées sur votre cluster.


Offre d'applications


Spark Interactive



Core Hadoop


Flink


HBase


Presto


Trino


Custom


☐ AmazonCloudWatchAgent 1.300032.2
☐ HCatalog 3.1.3
☐ Hue 4.11.0
☐ Livy 0.8.0
☐ Pig 0.17.0
☒ TensorFlow 2.16.1
☒ Zeppelin 0.11.1

☐ Flink 1.19.1
☒ Hadoop 3.4.0
☐ JupyterEnterpriseGateway 2.6.0
☐ Oozie 5.2.1
☐ Presto 0.287
☐ Tez 0.10.2
☐ ZooKeeper 3.9.2

☐ HBase 2.5.10
☐ Hive 3.1.3
☒ JupyterHub 1.5.0
☐ Phoenix 5.2.0
☒ Spark 3.5.2
☐ Trino 446



CRÉATION DE L'ENVIRONNEMENT BIG DATA

Configuration du serveur EMR

Configuration de cluster

▼ Configuration de cluster - *requis* [Info](#)

Choisissez une méthode de configuration pour les groupes de nœuds primaires, principaux et de tâches de votre cluster.

☒ Groupes d'instances uniformes

Choisissez le même type d'instance EC2 et la même option d'achat (à la demande ou Spot) pour tous les nœuds de votre groupe de nœuds. [En savoir plus](#)

☐ Flottes d'instances flexibles

Choisissez parmi la plus grande variété d'options de provisionnement pour les instances EC2 de votre cluster. Diversifiez les types d'instances et les options d'achat, et utilisez une stratégie d'allocation. [En savoir plus](#)

Groupes d'instances uniformes

Primaire

Choisir un type d'instance EC2

m5.xlarge

4 vCore 16 GiB mémoire

EBS uniquement stockage

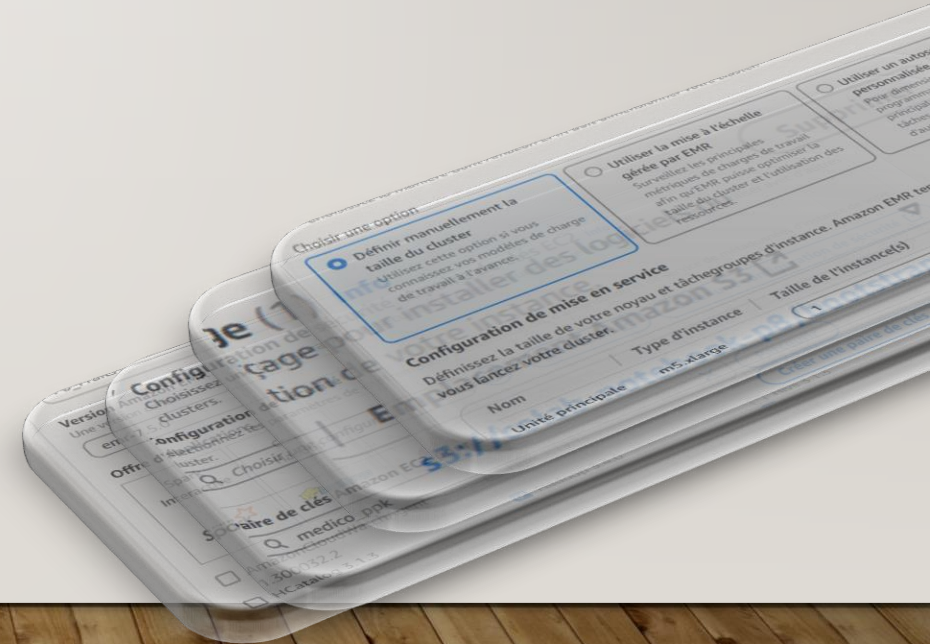
Prix à la demande : 0.224 USD par insta...

Prix Spot le plus bas : 0.081 USD (eu-west...

Actions ▼

☐ Utiliser la haute disponibilité

Lancez des clusters hautement disponibles et plus résilients avec trois nœuds primaires sur des instances à la demande. Cette configuration s'applique pendant toute la durée de vie de votre cluster. [En savoir plus](#)



CRÉATION DE L'ENVIRONNEMENT BIG DATA

Configuration du serveur EMR

Dimensionnement du cluster

▼ Dimensionnement et mise en service du cluster - *requis* [Info](#)

Choisissez la manière dont Amazon EMR doit dimensionner votre cluster.

Choisir une option

☒ Définir manuellement la taille du cluster

Utilisez cette option si vous connaissez vos modèles de charge de travail à l'avance.

☐ Utiliser la mise à l'échelle gérée par EMR

Surveillez les principales métriques de charges de travail afin qu'EMR puisse optimiser la taille du cluster et l'utilisation des ressources.

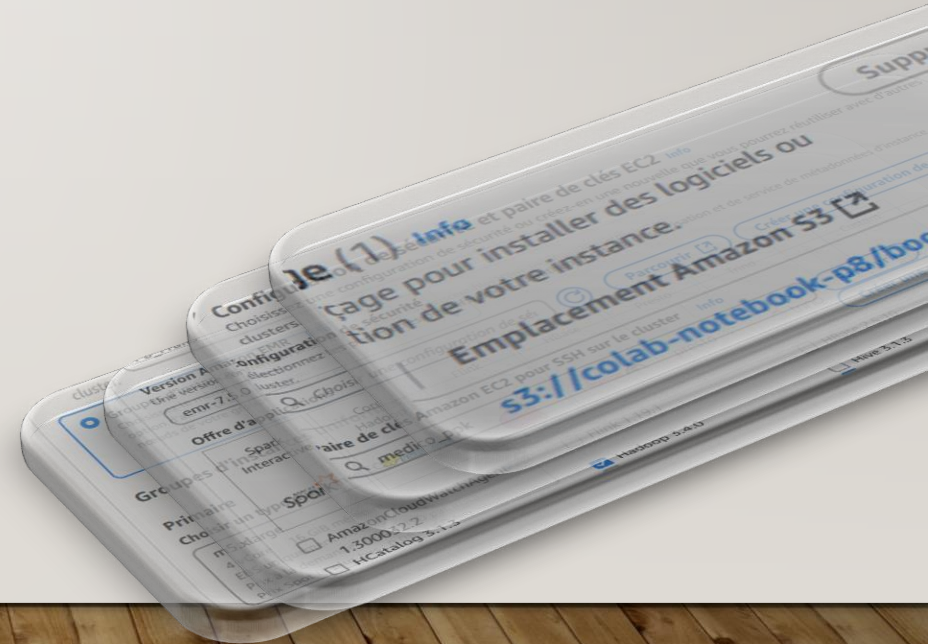
☐ Utiliser un autoscaling personnalisée

Pour dimensionner de manière programmatique les unités principales et les nœuds de tâches, créez des politiques d'autoscaling personnalisées.

Configuration de mise en service

Définissez la taille de votre noyau et tâchesgroupes d'instance. Amazon EMR tente de fournir cette capacité lorsque vous lancez votre cluster.

Nom	Type d'instance	Taille de l'instance(s)	Utiliser l'option d'achat Spot
Unité principale	m5.xlarge	<input type="text" value="1"/>	<input type="checkbox"/>
PCA	m5.xlarge	<input type="text" value="2"/>	<input type="checkbox"/>



CRÉATION DE L'ENVIRONNEMENT BIG DATA

Configuration du serveur EMR

Actions d'amorçage

▼ Actions d'amorçage (1) [Info](#)

Utilisez les actions d'amorçage pour installer des logiciels ou personnaliser la configuration de votre instance.

Supprimer

Modifier

Ajouter

Nom	Emplacement Amazon S3 🔗	Arguments
<input type="radio"/> Librairies	s3://colab-notebook-p8/bootstrap.sh	📄 Installer les librairies manquantes



CRÉATION DE L'ENVIRONNEMENT BIG DATA

Configuration du serveur EMR

Sécurité et autorisations

▼ Configuration de sécurité et paire de clés EC2 [Info](#)

Choisissez une configuration de sécurité ou créez-en une nouvelle que vous pourrez réutiliser avec d'autres clusters.

Configuration de sécurité

Sélectionnez les paramètres de chiffrement, d'authentification, d'autorisation et de service de métadonnées d'instance de votre cluster.



[Parcourir](#)

[Créer une configuration de sécurité](#)

Paire de clés Amazon EC2 pour SSH sur le cluster [Info](#)



[Parcourir](#)

[Créer une paire de clés](#)



CRÉATION DE L'ENVIRONNEMENT BIG DATA

Le Cluster créé

P8_Fruits

Mis à jour il y a 1 minute

Résilier

Cloner dans AWS CLI

Cloner

▼ Récapitulatif

Informations sur le cluster

ID de cluster
j-3436GR9MSGQ57

Configuration de cluster
Groupes d'instances

Capacité
1 primaire(s) | 1 unité(s) principale(s) | 2 tâche(s)

Applications

Version d'Amazon EMR
emr-7.5.0

Applications installées
Hadoop 3.4.0, JupyterHub 1.5.0, Spark 3.5.2, TensorFlow 2.16.1, Zeppelin 0.11.1

Gestion des clusters

Destination des journaux dans Amazon S3
aws-logs-248189923622-eu-west-3/elasticmapreduce

Interfaces utilisateur d'application persistantes
Serveur d'historique Spark
Serveur de chronologie YARN

DNS public du nœud primaire
ec2-15-188-207-185.eu-west-3.compute.amazonaws.com
Connexion au nœud primaire à l'aide de SSH
Connexion au nœud primaire à l'aide de SSM

Statut et heure

Statut
En attente

Heure de création
1 janvier 2025 19:40 (UTC+00:00)

Temps écoulé
10 minutes, 2 secondes

Propriétés

Actions d'amorçage

Instances (Matériel)

Étapes

Applications

Configurations

Surveillance

Événements

Identifications (1)

Système d'exploitation

Version Amazon Linux :
2023.6.20241031.0

Journaux de cluster

Archiver les fichiers journaux dans Amazon S3
Activé

Emplacement Amazon S3
s3://aws-logs-248189923622-eu-west-3/elasticmapreduce/

Chiffrement pour les journaux
Désactivé

Résiliation du cluster et remplacement des nœuds

Option de résiliation
Résilier automatiquement le cluster après le temps d'inactivité

Temps d'inactivité
1 jour, 1 heure

Protection contre la résiliation
Désactivé

Remplacement des nœuds défectueux
Activé

Réseau et sécurité

Réseau
Cloud privé virtuel (VPC)
vpc-0ebc3909dc0f815f7

Sous-réseau(x) et zone(s) de disponibilité
subnet-07c05683bedef6824 | eu-west-3b

Configuration de sécurité

Configuration de sécurité
Aucun

Paire de clés EC2
medico_ppk

Autorisations

Fonction du service pour Amazon EMR
AmazonEMR-ServiceRole-20241210T215905

Profil d'instance EC2
role_EMR_S3

RÉALISATION DE LA CHAÎNE DE TRAITEMENT

Installations, définition PATH et Charger les images

Installations, définition PATH

Dans la première partie de notre script, nous allons procéder aux actions suivantes :

- ❑ Installation des packages (Bootstrapping);
- ❑ Import des librairies;
- ❑ Définition des PATH;
- ❑ Démarrage de la session Spark



RÉALISATION DE LA CHAÎNE DE TRAITEMENT

Traitement des images

Préprocessing

- ❑ Préparation du modèle :
 - Transfert learning : ResNet 50
 - La méthode broadcast des “weights”
- ❑ Fonction Pandas UDF (User Defined Function) :

Cela permet de traiter efficacement nos grandes quantités d'images en parallèle.
- ❑ Application sur nos données images
- ❑ Enregistrement en format Parquet dans bucket



RÉALISATION DE LA CHAÎNE DE TRAITEMENT

Traitement des images

Réduction de dimension (PCA)

- ☐ Vecteurs denses
- ☐ Instance de PCA
- ☐ Enregistrement en format Parquet dans bucket
- ☐ Chargement du résultat :
On charge les données enregistrées dans un **DataFrame Spark**.
- ☐ Affichage de la DataFrame Spark :
 - Les 2 premières lignes;
 - Les dimensions.



DÉMONSTRATION D'EXÉCUTION DU SCRIPT PYSPARK SUR LE CLOUD



23/01/01 20:19:31 INFO
Dimensions: (10605, 4)



CONCLUSION

L'objectif était de pouvoir anticiper une future augmentation de la charge de travail. Notre tâche a donc consisté à créer un réel cluster de calculs.

Amazon Web Services nous a permis de louer à la demande de la puissance de calculs grâce service EMR.

Nous avons également opté pour le service Amazon S3 pour stocker les données de notre projet.



Les données utilisées sont des images de fruit et non est des informations se rapportant à une personne physique identifiée ou identifiable. Nous avons également choisi des serveurs situés sur le territoire européen (Paris)



Nous avons pu exécuter notre script comme si nous étions en local.

Nous avons exécuté le traitement sur l'ensemble des images; ce qui n'aurait pas été possible en local.

Il nous sera également facile de faire face à une montée de la charge de travail en redimensionnant simplement notre cluster de machines.



MERCI

OPENCLASSROOM