



Réalisez un traitement dans un environnement Big Data sur le Cloud

« Une application pour mieux connaître les fruits, un projet pour mieux maîtriser les données »

Projet: 9

Formation : Data Scientist

Oumou Faye

Mentor : Medina Hadjem

Plan de présentation



Introduction
-
Mission



Processus
de gestion
du projet



Architecture
Big Data sur
le Cloud



Chaîne de
traitement
PySpark



Conclusion
et
Perspectives



Entreprise



Start-up AgriTech
« FRUITS »



Propose des solutions innovantes pour la récolte des fruits



Mission



Développer une application mobile de reconnaissance de fruits

Processus

1. Le consommateur prend un fruit en photo



2. Le consommateur obtient instantanément des informations pertinentes sur ce fruit

Problématique

La quantité de données est volumineuse



La quantité de données va augmenter au fur et à mesure



Comment gérer le traitement d'un grand volume de données ?

Objectif

Mettre en place une première version du **moteur de reconnaissance des images de fruits**



Profiter d'une **architecture** sur le **Cloud** et réaliser les étapes de **traitement sur les images**



Sensibiliser le grand public à la biodiversité des fruits

Introduction et Mission



Dataset

Nom du jeu de données

Fruits-360 dataset

Nombre d'image

138.704

Nombre de fruit

206

- Le jeu de données est disponible sur le site Kaggle
- Le jeu de données va rapidement augmenter

Processus de gestion du projet

1
↓

Traitement local initial des images avec le framework PySpark pour valider la chaîne



2
↓

Choix du fournisseur cloud : sélection d'AWS (écosystème Big Data performant)

Identification des services AWS nécessaires (S3, EMR, EC2, IAM, EMR Studio)



4
↓

Mise en place de l'architecture cloud (configuration des services)



3
↑

Migration de la chaîne PySpark vers le cloud et exécution du traitement sur EMR



5
↑



6
↓

Enregistrement des résultats (au format parquet) sur S3



Architecture Big Data sur le Cloud



Simple Storage Service

S3

Stockage distribué
pour les images et les
résultats



Amazon
S3



PaaS : Platform as a service

IAM

« Identity Access
Management »

Gestion des **droits**
d'accès et des rôles
utilisateurs dans
l'environnement AWS.

EMR

Elastic MapReduce

Service pour exécuter
Spark/Hadoop sur des
clusters dans le cloud

EC2

Elastic Compute Cloud

Machines virtuelles
(workers) à la demande
pour exécuter des
traitements

EMR STUDIO






Elastic MapReduce Studio

Interface web pour **coder**
et **analyser** les données
via des notebook
connecté à des clusters
EMR

Architecture Big Data sur le Cloud



Le rôle des briques dans Amazon Web Services

<u>S3</u>  Amazon S3	<u>EMR</u>  amazon EMR	<u>EC2</u>  amazon web services™ EC2	<u>EMR STUDIO</u> 	<u>IAM</u>  aws IAM
Stockage centralisé pour les images brutes à traiter et les résultats du traitement.	Exécute les traitements Big Data distribués (Spark, Hadoop) sur un cluster.	Machine virtuelle à la demande, pouvant supporter EMR ou d'autres traitement spécifique.	Interface web Jupyterlab qui est reliée au cluster EMR.	Gère les permissions, rôles et accès aux services AWS. Garantissant la conformité RGPD .
↓	↓	↓	↓	↓
Le support de stockage distribué et durable.	L'environnement d'exécution du code PySpark (ex. transformation, PCA).	Ressource de calcul modulable selon les besoins (scalabilité).	Développement, test et visualisation du code Spark dans un environnement interactif.	Sécurité et contrôle d'accès aux données et traitements.

Architecture Big Data sur le Cloud



Configuration S3 – Création du compartiment

Amazon S3 > Compartiments > Créer un compartiment

Créer un compartiment [Infos](#)

Les compartiments sont des conteneurs pour les données stockées dans S3.

Configuration générale

Région AWS
Europe (Paris) eu-west-3

Nom du compartiment [Infos](#)

Le nom de compartiment doit comporter de 3 à 63 caractères et être unique dans l'espace de noms global. Les noms des compartiments doivent également commencer et se terminer par une lettre ou un chiffre. Les caractères valides sont les suivants : a-z, 0-9, points (.) et tirets (-).
[En savoir plus](#)

Copier les paramètres depuis un compartiment existant - facultatif
Sélectionnez un compartiment existant dans la configuration actuelle sans copier.

[Sélectionner un compartiment](#)

Format : <[1-5].<[a-z0-9]+>.[a-z0-9]+

Propriété d'objets [Infos](#)

Contrôlez la propriété des objets écrits dans ce compartiment à partir d'autres comptes AWS et l'utilisation des listes de contrôle d'accès (ACL). La propriété des objets détermine qui peut spécifier l'accès aux objets.

☒ **Listes ACL désactivées (recommandé)**
Tous les objets de ce compartiment sont gérés par ce compte. L'accès à ce compartiment et à ses objets est spécifié en utilisant uniquement des politiques.

☐ **Listes ACL actives**
Les objets de ce compartiment peuvent être gérés par d'autres comptes AWS. L'accès à ce compartiment et à ses objets peut être spécifié à l'aide des listes ACL.

Propriété d'objets
Propriétaire du compartiment appliqué

Paramètres de blocage de l'accès public pour ce compartiment

L'accès public aux compartiments et aux objets est accordé via des listes de contrôle d'accès (ACL), des stratégies de compartiment, de point d'accès ou tous ces éléments à la fois. Pour bloquer l'accès public à votre compartiment et aux objets qu'il contient, activez le paramètre Bloquer tous les accès publics. Il s'applique uniquement à ce compartiment et à ses points d'accès. AWS recommande de bloquer tous les accès publics, mais avant d'appliquer ces paramètres, vérifiez que vos applications fonctionnent correctement sans accès public. Si vous souhaitez autoriser un certain niveau d'accès public pour votre compartiment ou ses objets, vous pouvez personnaliser les paramètres individuels ci-dessous en fonction de vos besoins en stockage. [En savoir plus](#)

☐ **Bloquer tous les accès publics**
L'activation de ce paramètre empêche à la fois les quatre paramètres ci-dessous. Chacun des paramètres suivants est indépendant l'un de l'autre.

☐ **Bloquer l'accès public aux compartiments et aux objets, accordé via de nouvelles listes de contrôle d'accès (ACL)**
S3 bloque les autorisations d'accès public accordées aux compartiments ou objets récemment ajoutés et empêche la création de listes ACL d'accès public pour les compartiments et objets existants. Ce paramètre ne modifie pas les autorisations existantes qui permettent l'accès public aux ressources S3 qui utilisent les listes ACL.

☐ **Bloquer l'accès public aux compartiments et aux objets, accordé via d'importer quelles listes de contrôle d'accès (ACL)**
S3 ignore toutes les listes ACL aux objets existants publics, aux compartiments et aux objets.

☐ **Bloquer l'accès public aux compartiments et aux objets, accordé via de nouvelles stratégies de compartiment ou de point d'accès public**
S3 bloque les nouvelles stratégies de compartiment et de point d'accès qui accordent un accès public aux compartiments et objets. Ce paramètre ne modifie pas les stratégies existantes qui accordent l'accès public aux ressources S3.

☐ **Bloquer l'accès public et entre comptes aux compartiments et objets via d'importe quelles stratégies de compartiment ou de point d'accès public**
S3 ignore l'accès public et entre comptes pour les compartiments ou points d'accès avec des stratégies qui accordent l'accès public aux compartiments et aux objets.

Si le paramètre « Bloquer l'accès public » est désactivé, ce compartiment et les objets qu'il contient peuvent devenir publics.
AWS vous recommande de bloquer tous les accès publics, sauf si celui-ci est requis dans des cas d'utilisation spécifiques et vérifiés, tels que l'hébergement de site Web statique.

☒ Je suis conscient, qu'avec les paramètres actuels, ce compartiment et les objets qu'il contient peuvent devenir publics.

Bonnes pratiques - facultatif [Infos](#)

Ajouter des bonnes pratiques à ce compartiment.

[Ajouter une bonne pratique](#)

Chiffrement par défaut [Infos](#)

Le chiffrement client s'applique automatiquement à tous les nouveaux objets stockés dans ce compartiment.

Type de chiffrement [Infos](#)

☒ Chiffrement client serveur avec des clés gérées par Amazon S3 (SSE-S3)

☐ Chiffrement client serveur avec des clés AWS Key Management Service (SSE-KMS)

☐ Chiffrement client serveur avec des clés d'Amazon S3 (SSE-S3) et des clés AWS Key Management Service (SSE-KMS)

☐ Chiffrement client serveur avec des clés d'Amazon S3 (SSE-S3) et des clés d'Amazon S3 (SSE-S3) et des clés d'Amazon S3 (SSE-S3)

Côté de chiffrement

☒ Désactiver

☐ Activer

Paramètres avancés

[Après avoir créé le compartiment, vous pouvez y charger des fichiers et des dossiers et configurer des paramètres de compartiment supplémentaires.](#)

Nom de compartiment unique sur AWS

ACL (Access Control List) désactivées : accès géré uniquement par les politiques IAM

Blocage de l'accès public non activé

Chiffrement SEE-S3 géré par AWS



Configuration S3 – Création du compartiment

Nom du bucket S3

Amazon S3

aws-emr-studio-065693560228-eu-west-3

Objets (5)

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

Objet	Type	Dernière modification	Taille	Classe de stockage
e-E82R12H9J1MWBECBC25H8RRR/	Dossier	-	-	-
elasticmapreduce/	Dossier	-	-	-
Results_folder\$	-	24 Jun 2025 04:14:01 PM CEST	-	0 o Standard
Results/	Dossier	-	-	-
Test1/	Dossier	-	-	-

Résultat des images

Images à traiter

- Le bucket S3 a été créé avec succès.
- Deux dossiers y ont été ajoutés manuellement :
 - un pour les images à traiter,
 - un pour les résultats du traitement.

Architecture Big Data sur le Cloud

Configuration EMR - Le cluster



aws [Alt+S] Recherche Europe (Paris) AdministratorAccess/p9-user-2

Amazon EMR > EMR sur EC2: Clusters > Projet-Reconnaissance-Fruits

Mise à jour il y a moins d'une minute [Résilier] [Cloner dans AWS CLI] [Cloner]

▼ Récapitulatif

Informations sur le cluster	Applications	Gestion des clusters	Statut et heure
ID de cluster j-11UE2ZXP63IP1 ARN du cluster arn:aws:elasticmapreduce:eu-west-3:065693560228:cluster/j-11UE2ZXP63IP1 Configuration de cluster Groupes d'instances Capacité 1 primaire(s) 0 unité(s) principale(s) 0 tâche(s)	Version d'Amazon EMR emr-7.9.0 Applications installées Hadoop 3.4.1, JupyterEnterpriseGateway 2.6.0, JupyterHub 1.5.0, Livy 0.8.0, Spark 3.5.5, TensorFlow 2.16.1	Destination des journaux dans Amazon S3 aws-emr-studio-065693560228-eu-west-3/elasticmapreduce Interfaces utilisateur d'application persistantes Serveur d'historique Spark YARN Timeline Server DNS public du nœud primaire ec2-15-237-118-246.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 ✓ En attente Heure de création 25 juin 2025 15:28 (UTC+02:00) Temps écoulé 9 minutes, 5 secondes

Identifie le cluster de façon unique.

Contient les outils nécessaires au traitement Big Data.

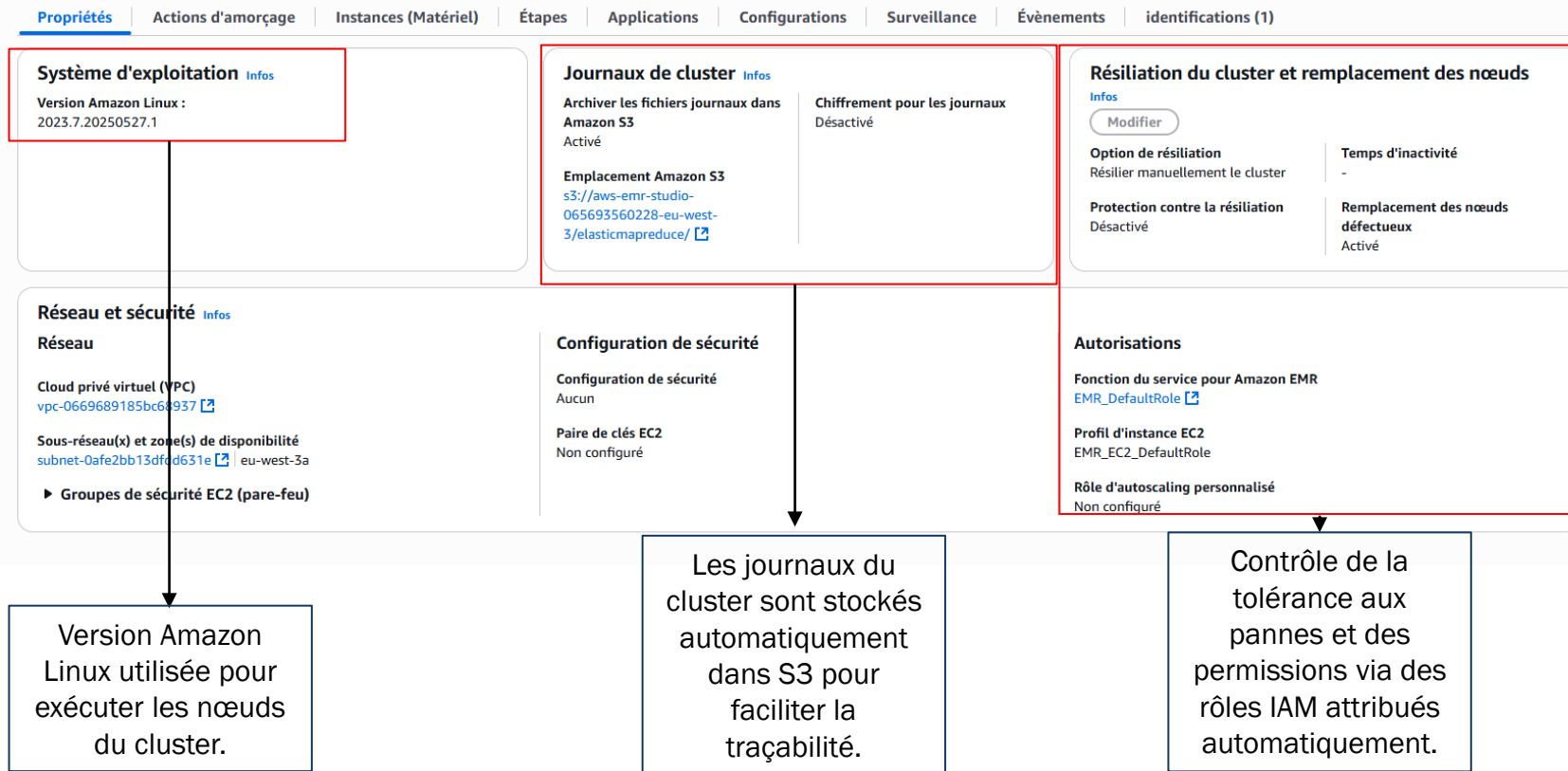
Permet la connexion, la journalisation et le suivi via S3.

Affiche le statut, la durée d'exécution et la date de création du cluster.

Architecture Big Data sur le Cloud



Configuration EMR - Le cluster





Primaire

Choisir un type d'instance EC2

m5.xlarge

4 vCore 16 GiB mémoire

EBS uniquement stockage

Prix à la demande : -

Prix Spot le plus bas : 0.075 USD (eu-west-...)

Actions ▼

- Nous avons utilisé l'instance EC2 : **m5.xlarge**.
- Elle est suffisamment **puissante** pour le **traitement parallèle d'images**.
- Elle est **équilibrée** entre **performance et coût**.
- Elle est adaptée à l'exécution de « User Defined Functions » de deep learning léger (ex: MobileNetV2).

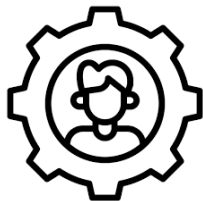


Configuration IAM



AWS IAM

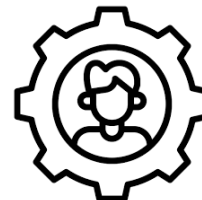
EMR_DefaultRole



☐ | Nom de la politique [?](#)

- ☐ [+ AmazonEC2FullAccess](#)
- ☐ [+ AmazonEMRFullAccessPolicy_v2](#)
- ☐ [+ AmazonSSMManagedInstanceCore](#)
- ☐ [+ EMR_Bucket_Access_Policy](#)
- ☐ [+ EMR-EC2-Describe-Permissions](#)
- ☐ [+ PermissionEC2](#)

EMR_EC2_DefaultRole



☐ | Nom de la politique [?](#)

- ☐ [+ AmazonElasticMapReduceforEC2Role](#)
- ☐ [+ AmazonEMRServicePolicy_v2](#)
- ☐ [+ AmazonS3FullAccess](#)
- ☐ [+ AmazonSSMManagedInstanceCore](#)
- ☐ [+ EMR_Bucket_Access_Policy](#)

Architecture Big Data sur le Cloud

Configuration EMR Studio



aws [Rechercher] [Alt+S] Europe (Paris) AdministratorAccess/p9-user-2

Amazon EMR > EMR Studio: Studios > studio-cluster-p9

Amazon EMR <

EMR Serverless

▼ EMR sur EC2

Clusters

Blocs-notes et référentiels Git

Événements

Bloquer l'accès public

Configurations de sécurité

▼ EMR sur EKS

Clusters virtuels

▼ EMR Studio

Mise en route

Studios

Espaces de travail (Blocs-notes)

Nouveautés

Visite guidée en vidéo

Mode compact

studio-cluster-p9

Paramètres Studio

ID de studio es-AF1NWWOZ2UCTKRKMBN4ON24YO	VPC vpc-0669689185bc68937	Groupe de sécurité du moteur sg-0b247740bf6a7312d	Authentifié par IAM
Description -	Sous-réseaux subnet-0afe2bb13dfdd631e	Groupe de sécurité de l'instance WorkSpace sg-027502636ddadd152	Fonction du service arn:aws:iam::065693560228:role/p9-user-2
identifications -	Stockage de l'instance WorkSpace s3://aws-emr-studio-065693560228-eu-west-3/e-EBZR12H9J1MW8IECBC25H8RRR		
URL https://es-AF1NWWOZ2UCTKRKMBN4...	Stockage de l'instance Clé KMS -		

Accès à EMR Studio via JupyterLab pour créer des notebooks connectés au cluster EMR avec PySpark.

Bucket S3 associé pour stocker les images d'entrée et les résultats produits.

Architecture Big Data sur le Cloud

Configuration EMR Studio



EMR Studio X

Dashboard
Workspaces
Query editor New
Powered by Athena

▼ Serverless
Applications

▼ Clusters
EMR on EC2

▼ Service Integrations
SageMaker Data Wrangler New
Low-code data prep and ML

What's New
Submit feedback
Logout

EMR Studio > Workspaces

Workspaces

Studio: studio-cluster-p9

Workspaces (1) Info

Find Workspaces by name, status, or last modified by

Actions Launch Workspace Create Workspace

	Workspace name	Status	Cluster ID	Creation time (UTC+02:00)	Last modified by	Last modified (UTC+02:00)
<input type="radio"/>	cluster-emr-studio-final3	<input type="radio"/> Idle	-	June 19, 2025, 10:29	p9-user-2	June 24, 2025, 23:22

- L'espace de travail **EMR Studio** a bien été crée.
- Il est prêt à être lancé pour accéder à **JupyterLab** et déployer la chaîne de traitement sur Amazon Web Services.



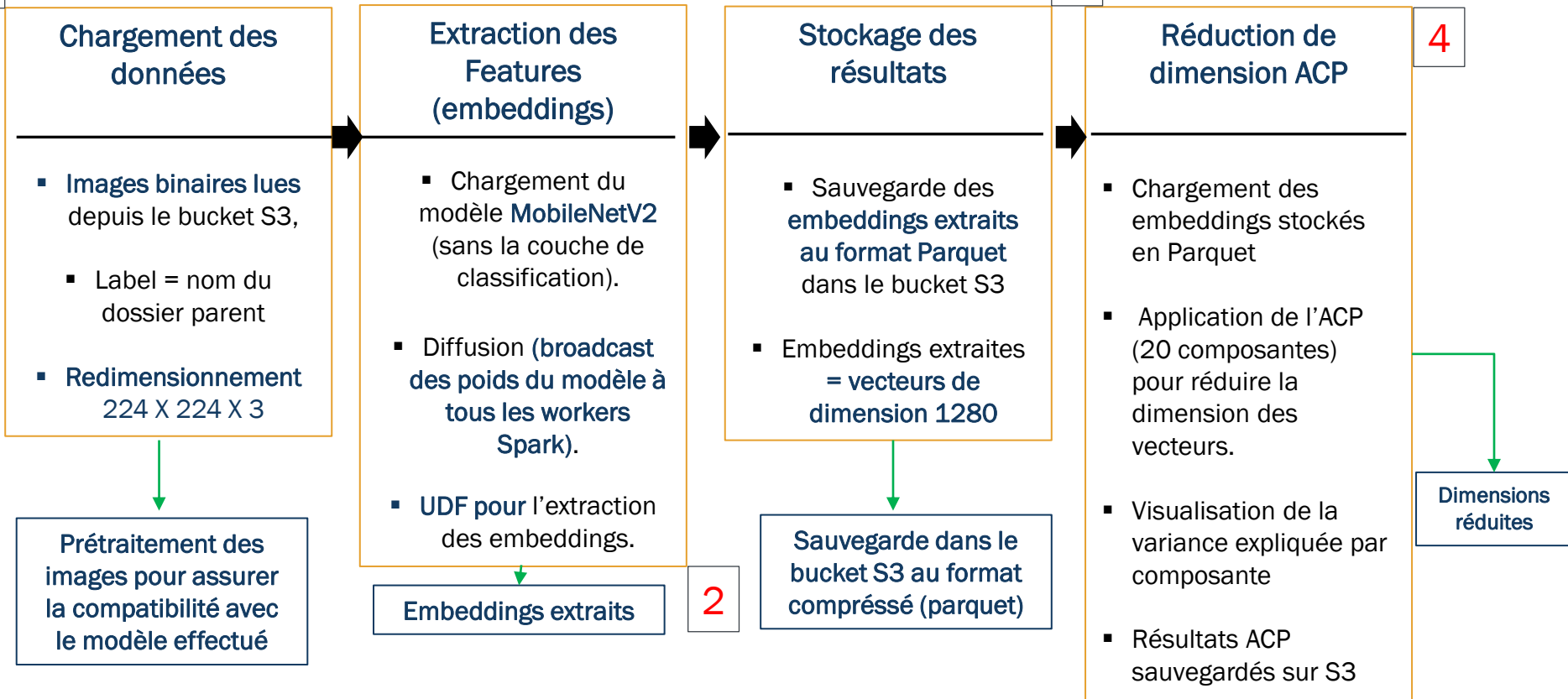
Résumer général des configurations AWS

- ✓ Le **bucket S3** a été créé avec succès.
- ✓ Le **cluster EMR** est au statut « **en attente** » prêt à être connecté à un notebook via le Kernel PySpark.
- ✓ Les rôles **EMR_DefaultRole** et **EMR_EC2_Default** possèdent les **politiques nécessaires** pour interagir avec le bucket contenant les images et avec le cluster EMR.
- ✓ L'environnement **EMR Studio** a bien été créé, et nous avons accès à l'interface **JupyterLab** via l'espace de travail.
- ✓ Bien que nos images ne contiennent **aucune donnée personnelle identifiable**, toutes les configurations sont déployées dans la **région européenne (eu-west-3)**, assurant la conformité avec le **RGPD** (**R**èglement **G**énérale sur la **P**rotection des **D**onnées).

Chaîne de traitement PySpark



Pipeline PySpark de traitement des images

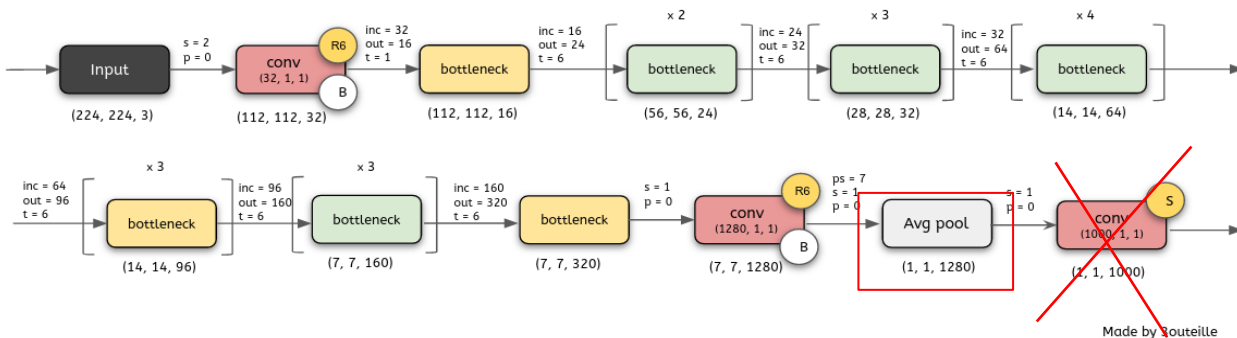


Chaîne de traitement PySpark

Particularité du modèle MobileNetV2



floor = True (for Pytorch Conv2d)



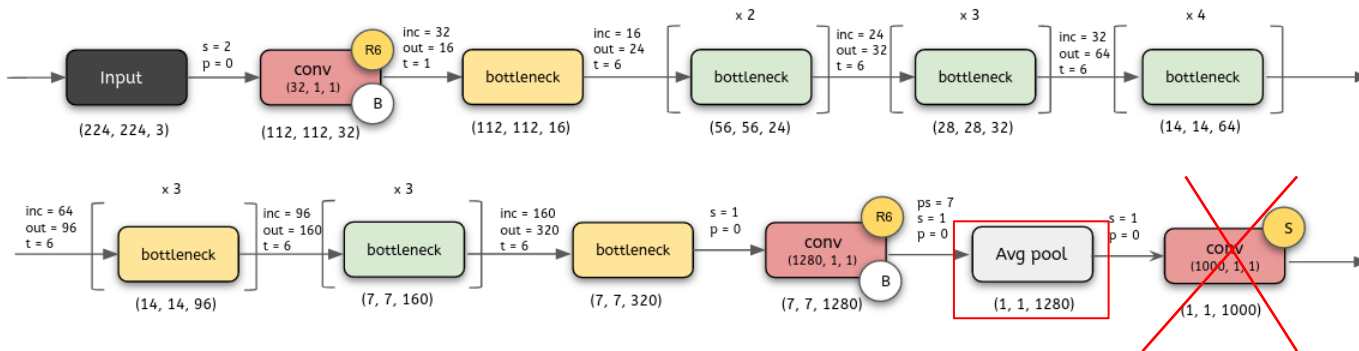
- Modèle **léger** et **rapide**, idéal pour **des environnements distribués** (comme **Spark**).
- Pré-entraîné sur **ImageNet**, ce modèle possède une bonne capacité à **extraire des caractéristiques visuelles**.
- Utilise une architecture optimisée avec des blocs « **bottleneck** » et des connexions **résiduelles**, facilitant l'apprentissage.
- Compatible avec des entrées d'image **224x224x3** (**couramment utilisées**).
- Permet d'extraire efficacement des **embeddings visuels**, en utilisant les couches **avant la classification finale**.

Chaîne de traitement PySpark

Particularité du modèle MobileNetV2



floor = True (for Pytorch Conv2d)



- ✓ Le modèle **MobileNetV2** a été utilisé pour **extraire des embeddings visuels** à partir des images.
- ✓ **MobileNetV2** est utilisé pour de nombreuses tâches de vision par ordinateur (computer vision) : classification d'images, détection d'objets, segmentation ou la reconnaissance faciale.
- ✓ **Seule l'avant-dernière couche** (juste avant la classification finale) a été conservée. Elle produit un **vecteur de 1280 dimensions**, servant de **représentation visuelle** de chaque image.

Chaîne de traitement PySpark

Utilité de PySpark dans le pipeline



- Traitement distribué des images en parallèle
- Utilisation des UDF pour appliquer MobileNetV2
- Optimisation du temps d'exécution



- Stockage dans EMR instance EC2
- Adaptation des ressources selon la charge
- Optimisation des coûts grâce à EC2

- Diffusion automatique des paramètres du modèle (broadcast)
- Tous les workers Spark appliquent localement le modèle
- Evite les rechargements inutiles => gain de performance



Réduction de dimension avec l'analyse en composante principale

L'objectif était de réduire dimensionnalité des caractéristiques (1280 dimensions) tout en conservant l'essentiel de l'information.

Composante | Var. expliquée (%) | Var. cumulée (%)

1	66.15	66.15
2	19.35	85.50
3	5.26	90.76
4	3.17	93.94
5	2.22	96.15
6	2.09	98.24
7	1.13	99.37
8	0.31	99.68
9	0.20	99.88
10	0.10	99.98
11	0.02	100.00
12	0.00	100.00
13	0.00	100.00
14	0.00	100.00
15	0.00	100.00
16	0.00	100.00
17	0.00	100.00
18	0.00	100.00
19	0.00	100.00
20	0.00	100.00

Avantages :

1. Réduction du temps de calcul
2. Moins de mémoire utilisé
3. Moins de bruit et redondance dans les données
4. Préparation efficace pour effectuer la prédiction avec le modèle

- Les 2 premières composantes expliquent déjà 85,5 % de la variance.
- Avec seulement 6 composantes, on conserve 98 % de l'information.

Chaîne de traitement PySpark

Surveillance via les logs



▼ Completed Jobs (62)

Page: 1 1 Pages. Jump to: 1 Show 100 Items in a page. Go

Job Id (Job Group) *	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
61 (65)	Job group for statement 65 parquet at NativeMethodAccessorImpl.java:0	2025/06/24 07:01:24	38 s	1/1	2/2
60 (64)	Job group for statement 64 showString at NativeMethodAccessorImpl.java:0	2025/06/24 07:00:08	0.5 s	1/1	1/1
59 (64)	Job group for statement 64 treeAggregate at RowMatrix.scala:139	2025/06/24 07:00:06	1 s	1/1	2/2
58 (64)	Job group for statement 64 isEmpty at RowMatrix.scala:441	2025/06/24 07:00:06	0.2 s	1/1	1/1
57 (64)	Job group for statement 64 isEmpty at RowMatrix.scala:441	2025/06/24 07:00:06	0.3 s	1/1	1/1
56 (64)	Job group for statement 64 treeAggregate at Statistics.scala:58	2025/06/24 07:00:05	0.7 s	1/1	2/2
55 (64)	Job group for statement 64 first at RowMatrix.scala:62	2025/06/24 07:00:05	0.5 s	1/1	1/1
54 (64)	Job group for statement 64 first at PCA.scala:44	2025/06/24 06:59:29	36 s	1/1	1/1
53 (62)	Job group for statement 62 isEmpty at RowMatrix.scala:441	2025/06/24 06:56:17	0.7 s	1/1	1/1
52 (62)	Job group for statement 62 isEmpty at RowMatrix.scala:441	2025/06/24 06:56:16	1 s	1/1	1/1
51 (62)	Job group for statement 62 treeAggregate at Statistics.scala:58	2025/06/24 06:56:15	1 s	1/1	2/2
50 (62)	Job group for statement 62 first at RowMatrix.scala:62	2025/06/24 06:56:14	0.8 s	1/1	1/1

- Le traitement des tâches est vérifié dans l'onglet « Jobs » de l'interface Spark.
- La durée d'exécution de chaque tâche est accessible.
- Vérification que tous les « stages » et « tasks » sont bien en statut « succeeded ».



Visualisation de l'activation des « executors » (workers) dans le cluster.

- Observation du lancement parallèle des jobs.
- Confirmation que les poids du modèle ont été diffusés correctement sur l'ensemble des workers.



Nous avons effectué :

- Le traitement **distribué** d'images grâce à PySpark + AWS EC2.
- Utilisation de **workers** (machines virtuelles) connectés à AWS EC2 pour **accélérer les traitements** à grande échelle.
- Réduction de la dimension des images grâce à l'analyse en composantes principales (PCA) :
 - Pour réduire l'utilisation mémoire et accélérer les calculs.
- Le pipeline est prêt à intégrer un **modèle de prédiction supervisé** (fit/predict).

Perspectives :

- Test d'autres modèles d'extraction tels que **ResNet (Residual Network)** ou **EfficientNet (Efficient Convolutional Network)**.
- Entraînement d'un modèle de **prédiction** (ex. MLP – **Multi-Layer Perceptron**, SVM – **Support Vector Machine**) à partir des embeddings extraits.
- Déploiement de l'interface utilisateur « front-end » de l'application mobile.
- Pipeline prêt à traiter un volume massif d'images.



Merci de votre attention

Question(s) ? / Réponse(s)

