

Data Scientist Projet : 8



# Déployez un modèle dans le cloud

Mayves NKONDO

Mentor : Hamza Tajmouati

# >>> Compétences évaluées

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



# Sommaire

- Problématique
- Présentation des données
- Big Data ?
- Traitement des Images
- Conclusions



# Fruits!

# Problématique





## Fruits !

Souhaite proposer des solutions innovantes pour la récolte des fruits.

Développer des robots cueilleurs intelligents à l'aide d'une application qui permettra aux utilisateurs de prendre en photo un fruit et d'obtenir des informations sur ce fruit.

## Mission

Développer un environnement Big Data

Réaliser une première chaîne de traitement des données avec le preprocessing et une étape de réduction de dimension.

# Présentation des données

- Ensemble de données contenant des images de haute qualité de fruit avec les labels associés.
- 22700 images au format JPG 100 x 100 pixels.
- 131 Variétés différentes.
- Chaque fruit est photographié sous différents angles.



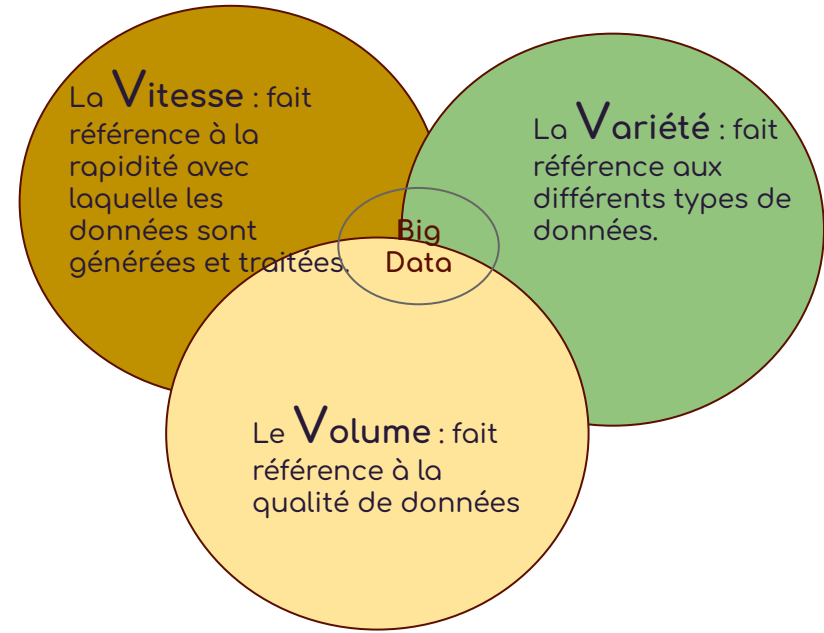
# Pourquoi un environnement Big Data ?

## 1 - Qu'est ce que le Big Data :

Le terme "Big Data" est employé pour décrire des ensembles de données volumineux, complexes et hétérogènes collectés à partir de diverses sources, lesquels ne peuvent pas être traités efficacement par les méthodes traditionnelles de gestion des données.

Les données du Big Data sont caractérisées par les **3V** :

Le Volume, La Variété et La Vitesse.





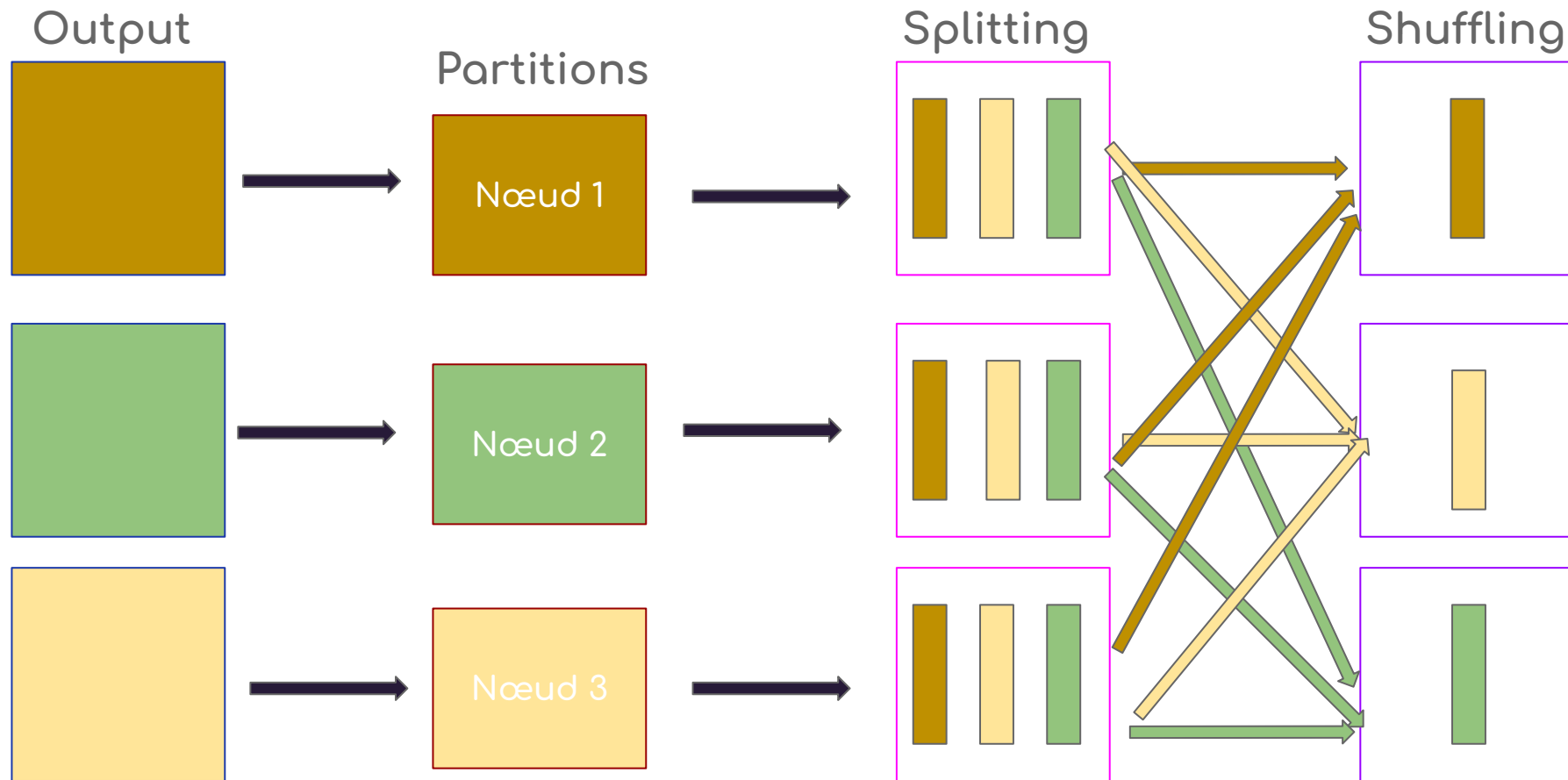


## 2 - Big Data Frameworks :

	Hadoop	Spark
Architecture	Stock et traite les données sur un stockage externe	Stock et traite les données dans la mémoire interne
Traitement de données	Par lots	En temps réel
	Moins rapides	Plus rapides
Performances	Bibliothèques externes	Bibliothèques intégrées
Machine Learning	Kerberos	Authentification avec un mot de passe secret
Sécurité	HDFS	HDFS,...
Bases de données		



### 3 - Comment Spark traite les données :



#### 4 - Amazon Web Service:








Le Cloud AWS est une plateforme de services cloud développée par le géant américain Amazon. AWS regroupe plus de 200 services répartis en diverses catégories telles que le stockage cloud, la puissance de calcul, l'analyse de données, l'intelligence artificielle ou même le développement de jeux vidéo.

# Big Data - Suite :

## Configuration de la console AWS

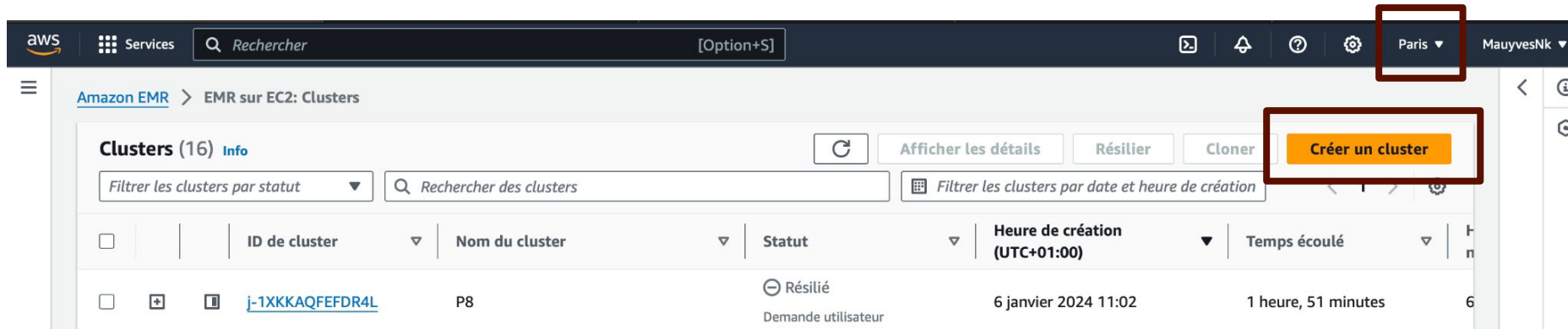
### Etape 1:

Création d'un bucket sur s3 dans lequel, je télécharge le contenu des images dans les deux dossiers (Apple Pink Lady/ et Cherry Wax Yellow/), le fichier d'amorçage (bootstrap-emr.sh) et la création d'un dossier(dataResults/) pour télécharger les parquets

<input type="radio"/>	p8mauyves		Europe (Paris) eu-west-3		Compartiment et objets non publics		02 Jan 2024 05:14:20 PM CET	
<input type="checkbox"/>	Nom ▲	Type ▼	Dernière modification ▼	Taille ▼	Classe de stockage ▼			
<input type="checkbox"/>	 Apple Pink Lady/	Dossier	-	-	-			
<input type="checkbox"/>	 bootstrap-emr.sh	sh	06 Jan 2024 12:04:41 AM CET	516.0 o	Standard			
<input type="checkbox"/>	 Cherry Wax Yellow/	Dossier	-	-	-			
<input type="checkbox"/>	 dataResults/	Dossier	-	-	-			
<input type="checkbox"/>	 jupyter/	Dossier	-	-	-			

## Étape 2:

Création du cluster EMR dans Instances EC2 situées en France sur référence (eu-west-3)



The screenshot shows the AWS Management Console interface for Amazon EMR. The top navigation bar includes the AWS logo, 'Services', a search bar, and the 'Paris' region dropdown, which is highlighted with a red box. The main content area is titled 'Amazon EMR > EMR sur EC2: Clusters'. Below this, there's a section for 'Clusters (16)' with an 'Info' link. A search bar and a filter dropdown are present. A table lists existing clusters, with one cluster 'j-1XKKAQFEFDR4L' in the 'Résilié' (Resilié) state. The 'Créer un cluster' button is highlighted with a red box.

ID de cluster	Nom du cluster	Statut	Heure de création (UTC+01:00)	Temps écoulé
j-1XKKAQFEFDR4L	P8	Résilié Demande utilisateur	6 janvier 2024 11:02	1 heure, 51 minutes


### Nom et applications [Info](#)

Nom


P8


# Big Data - Suite :


## Offre d'applications


Spark  
Interactive  



Core  
Hadoop  


Flink  


HBase  


Presto  


Trino  


Custom  


- |   |  |   |
|---|--|---|
| <input type="checkbox"/> Flink 1.17.1           | <input type="checkbox"/> Ganglia 3.7.2                             | <input type="checkbox"/> HBase 2.4.17                 |
| <input type="checkbox"/> HCatalog 3.1.3         | <input checked="" type="checkbox"/> Hadoop 3.3.6                   | <input type="checkbox"/> Hive 3.1.3                   |
| <input type="checkbox"/> Hue 4.11.0             | <input checked="" type="checkbox"/> JupyterEnterpriseGateway 2.6.0 | <input checked="" type="checkbox"/> JupyterHub 1.5.0  |
| <input type="checkbox"/> Livy 0.7.1             | <input type="checkbox"/> MXNet 1.9.1                               | <input type="checkbox"/> Oozie 5.2.1                  |
| <input type="checkbox"/> Phoenix 5.1.3          | <input type="checkbox"/> Pig 0.17.0                                | <input type="checkbox"/> Presto 0.283                 |
| <input checked="" type="checkbox"/> Spark 3.4.1 | <input type="checkbox"/> Sqoop 1.4.7                               | <input checked="" type="checkbox"/> TensorFlow 2.11.0 |
| <input type="checkbox"/> Tez 0.10.2             | <input type="checkbox"/> Trino 426                                 | <input type="checkbox"/> Zeppelin 0.10.1              |
| <input type="checkbox"/> ZooKeeper 3.5.10       |  |   |

J'ai choisi Spark,  
TensorFlow, JupyterHub  
1.5.0 et Hadoop 3.3.6

## Configuration de mise en service

Définissez la taille de votre noyau et tâche groupes 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="2"/>	<input type="checkbox"/>
Tâche - 1	m5.xlarge	<input type="text" value="1"/>	<input type="checkbox"/>

Je sélectionne 2  
instances principales et  
1 instance maître => 3  
instances EC2

# Big Data - Suite :

## ▼ Actions d'amorçage - facultatif (1) [Info](#)

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

Supprimer

Modifier

Ajouter

```
#!/bin/bash
sudo python3 -m pip install -U setuptools
sudo python3 -m pip install -U pip
sudo python3 -m pip install -U wheel
sudo python3 -m pip install -U pillow
sudo python3 -m pip install -U pyarrow
sudo python3 -m pip install -U boto3
sudo python3 -m pip install -U s3fs
sudo python3 -m pip install -U fsspec
sudo python3 -m pip install -U pandas
sudo python3 -m pip install -U numpy
sudo python3 -m pip install -U scikit-learn
sudo python3 -m pip install -U matplotlib
sudo python3 -m pip install -U tensorflow
```

bootstrap-emr.sh,  
est un fichier pour  
installer les  
bibliothèques  
manquantes

Nom

Emplacement Amazon S3 [🔗](#)

Arguments



p8mauyves

<s3://p8mauyves/bootstrap-emr.sh>

-

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

### Configuration de sécurité

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

Parcourir [🔗](#)
🔗"/>

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



Parcourir

🔗"/>

À cette étape, nous  
sélectionnons les clés EC2  
créées précédemment.  
Celles-ci nous permettront  
de nous connecter en SSH  
à nos instances EC2 sans  
avoir à entrer nos  
identifiants et mots de  
passe.

# Big Data - Suite :

## ▼ Récapitulatif

### Informations sur le cluster

ID de cluster  
j-2F0252GFFAXLL

Configuration de cluster  
Groupes d'instances

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

### Applications

Version d'Amazon EMR  
emr-6.15.0

Applications installées  
Hadoop 3.3.6, JupyterEnterpriseGateway  
2.6.0, JupyterHub 1.5.0, Spark 3.4.1,  
TensorFlow 2.11.0

### Gestion des clusters

Destination des journaux dans Amazon S3  
[aws-logs-340767347534-eu-west-3/elasticmapreduce](#)

Interfaces utilisateur d'application  
persistantes  
[Serveur d'historique Spark](#)  
[Serveur de chronologie YARN](#)

DNS public du nœud primaire  
[ec2-51-44-23-40.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  
8 janvier 2024 06:16 (UTC+01:00)

Temps écoulé  
10 minutes, 31 secondes

Il ne nous reste plus qu'à attendre que le serveur soit prêt. Cette étape peut prendre entre 10 et 20 minutes.

```
Last login: Mon Jan  8 05:37:37 on tty000
(base) admin@Admins-MBP ~ % cd //Users/admin/Downloads
(base) admin@Admins-MBP Downloads % chmod 400 mauvyves-ec2.pem
(base) admin@Admins-MBP Downloads % ssh -i mauvyves-ec2.pem -D 5555 hadoop@ec2-51-44-23-40.eu-west-3.compute.amazonaws.com
Warning: Permanently added 'ec2-51-44-23-40.eu-west-3.compute.amazonaws.com' (ED25519) to the list of known hosts.
Amazon Linux 2
AL2 End of Life is 2025-06-30.

A newer version of Amazon Linux is available!
Amazon Linux 2023, GA and supported until 2028-03-15.
https://aws.amazon.com/linux/amazon-linux-2023/

EEEEEEEEEEEEEEEEEEEE MMMMMMM RRRRRRRRRRRRR
E:::EEEEEEEEEEEEEEEE M:::M M:::M R:::R
EE:::EEEEEEEEEEEEEEEE M:::M M:::M R:::RRRRRR:::R
E:::E EEEEE M:::M M:::M RR:::R R:::R
E:::E M:::M M:::M M:::M M:::M R:::R R:::R
E:::EEEEEEEEEEEE M:::M M:::M M:::M R:::RRRRR:::R
E:::EEEEEEEEEEEE M:::M M:::M M:::M R:::RRRRR:::R
E:::E EEEEE M:::M M:::M M:::M R:::R R:::R
E:::EEEEEEEEEEEE M:::M M:::M M:::M R:::R R:::R
E:::E M:::M M:::M M:::M R:::R R:::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM RRRRRRRRRRRRR
[ hadoop@ip-172-31-34-26 ~]$
```

Nous avons correctement établi le tunnel ssh avec le driver sur le port “5555”.



# Big Data - Suite :

## Étape 3:

Sign in

Username:

jovyan

Password:

\*\*\*\*\*

Sign in

### Objets (2) [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](#)

Copier l'URI S3

Copier l'URL

Télécharger

Ouvrir

Supprimer

Actions ▼

Créer un dossier

Charger

<
1
>

<input type="checkbox"/>	Nom ▲	Type ▼	Dernière modification ▼	Taille ▼	Classe de stockage ▼
<input type="checkbox"/>	<a href="#">.s3keep</a>	s3keep	06 Jan 2024 11:19:16 AM CET	0 o	Standard
<input type="checkbox"/>	<a href="#">P8_emr_cloud.ipynb</a>	ipynb	06 Jan 2024 12:51:15 PM CET	72.5 Ko	Standard

# Traitement des images



Data Sources



S3

Télécharger les données dans S3



EMR

Utiliser Amazon EMR pour le traitement des images +  
Extraction des features +  
Réduction de Dimension (ACP)



S3

Charger des données au format parquet dans S3

<https://www.kaggle.com/datasets/moltean/fruits>

# Modèle MobileNetV2

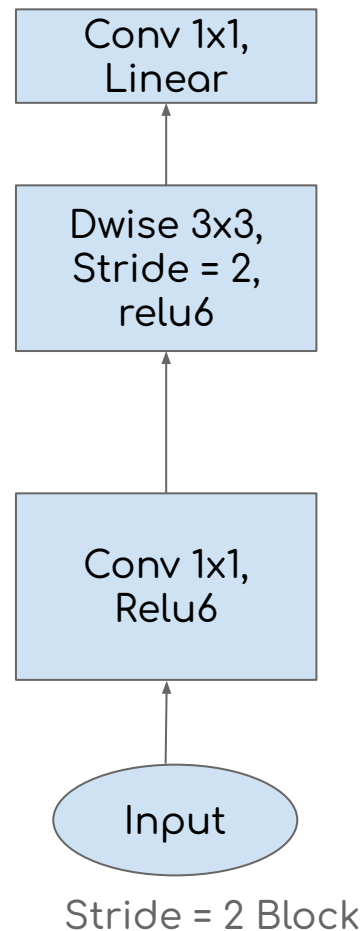
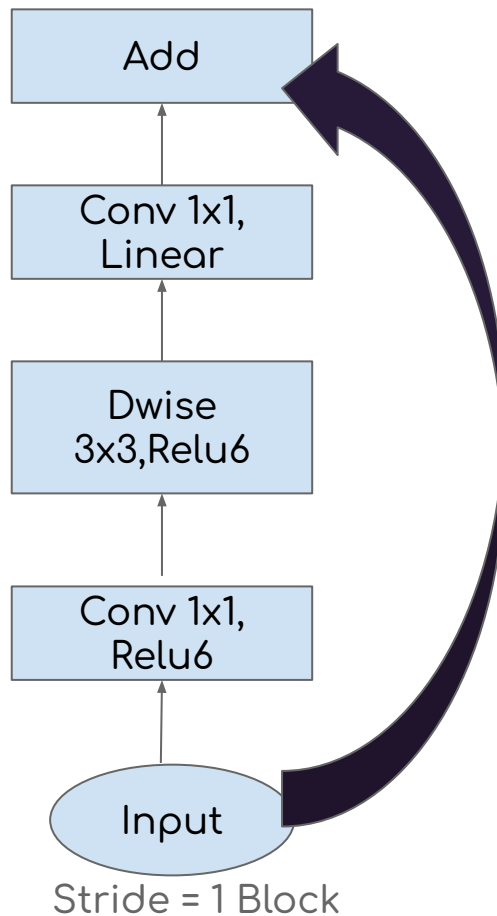
Ce sont les deux composants différents dans le modèle MobileNet V2 :

Le modèle MobileNet V2 comporte 53 couches de convolution et 1 AvgPool avec près de 350 GFLOP. Il se compose de deux composants principaux :

- Bloc résiduel inversé
- Bloc résiduel en entonnoir

Il existe deux types de couches de convolution dans l'architecture MobileNet V2 :

- Convolution 1x1
- Convolution en profondeur 3x3



# Traitement des images - Suite :

- 1280 descripteurs par image
- Conversion au format vecteur dense
- Standardisation

path	label	features	features_vectorized	scaledFeatures
s3://p8mauyves/Ap...	Apple Pink Lady	[0.1960293, 0.113...	[0.19602930545806...	[-0.2075054344109...
s3://p8mauyves/Ap...	Apple Pink Lady	[0.17898463, 0.04...	[0.17898462712764...	[-0.2840561488596...
s3://p8mauyves/Ap...	Apple Pink Lady	[0.98299426, 0.01...	[0.98299425840377...	[3.32689563941691...
s3://p8mauyves/Ch...	Cherry Wax Yellow	[0.1632067, 0.490...	[0.16320669651031...	[-0.3549176716097...
s3://p8mauyves/Ch...	Cherry Wax Yellow	[0.0, 0.672891, 0...	[0.0, 0.6728910207...	[-1.0879082847784...
s3://p8mauyves/Ch...	Cherry Wax Yellow	[0.15166348, 0.17...	[0.15166348218917...	[-0.4067603222123...

only showing top 6 rows

```
# Calculer le nombre de composantes nécessaires pour expliquer 95% de la variance cumulative
cumsum = 0
for i in pca.explainedVariance.cumsum():
    cumsum += 1
    if(i > 0.8):
        print(
            '{} composantes expliquent 80% de la variance'.format(cumsum))
        break
```

FloatProgress(value=0.0, bar\_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),...

40 composantes expliquent 80% de la variance

## PCA finale

label	features_pca
Apple Pink Lady	[-18.099312063480...
Apple Pink Lady	[-18.364402838348...
Apple Pink Lady	[-16.991673369214...
Cherry Wax Yellow	[15.0310022481309...
Cherry Wax Yellow	[19.3920419732819...
Cherry Wax Yellow	[17.0163878334519...

only showing top 6 rows

```
# Appliquer l'analyse en composantes principales (PCA) pour réduire la dimensionnalité des données
# Paramètres :
# - k : Nombre de composantes principales à conserver ou proportion cumulée à conserver (cumsum)
# - inputCol : Colonne d'entrée contenant les caractéristiques mises à l'échelle
# - outputCol : Colonne de sortie pour les caractéristiques réduites mises à l'échelle

pca = PCA(
    k=cumsum,
    inputCol='scaledFeatures',
    outputCol='features_pca')

# Appliquer le modèle PCA aux données mises à l'échelle
model_pca = pca.fit(df_scaled)
df_final = model_pca.transform(df_scaled)

# Supprimer la colonne "scaledFeatures" des données finales
df_final = df_final.drop('path', 'scaledFeatures', 'features', 'features_vectorized')
```

FloatProgress(value=0.0, bar\_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),...

# Traitement des images - Suite :

## Historique de session Spark

Job Id (Job Group) ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
29 (28)	Job group for statement 28 <a href="#">parquet at NativeMethodAccessorImpl.java:0</a>	2024/01/08 06:19:37	17 s	1/1 (1 skipped)	24/24 (10 skipped)
28 (28)	Job group for statement 28 <a href="#">parquet at NativeMethodAccessorImpl.java:0</a>	2024/01/08 06:19:34	3 s	1/1	10/10
27 (27)	Job group for statement 27 <a href="#">showString at NativeMethodAccessorImpl.java:0</a>	2024/01/08 06:19:30	2 s	1/1 (1 skipped)	1/1 (10 skipped)
26 (27)	Job group for statement 27 <a href="#">showString at NativeMethodAccessorImpl.java:0</a>	2024/01/08 06:19:26	3 s	1/1	10/10
25 (26)	Job group for statement 26 <a href="#">treeAggregate at RowMatrix.scala:171</a>	2024/01/08 06:19:00	16 s	2/2 (1 skipped)	28/28 (10 skipped)
24 (26)	Job group for statement 26 <a href="#">isEmpty at RowMatrix.scala:441</a>	2024/01/08 06:18:58	2 s	1/1 (1 skipped)	1/1 (10 skipped)
23 (26)	Job group for statement 26 <a href="#">treeAggregate at Statistics.scala:58</a>	2024/01/08 06:18:41	17 s	2/2 (1 skipped)	28/28 (10 skipped)
22 (26)	Job group for statement 26 <a href="#">first at RowMatrix.scala:62</a>	2024/01/08 06:18:39	2 s	1/1 (1 skipped)	1/1 (10 skipped)
21 (26)	Job group for statement 26 <a href="#">first at PCA.scala:44</a>	2024/01/08 06:18:35	4 s	2/2	11/11
20 (22)	Job group for statement 22 <a href="#">treeAggregate at RowMatrix.scala:171</a>	2024/01/08 06:18:09	16 s	2/2 (1 skipped)	28/28 (10 skipped)
19 (22)	Job group for statement 22 <a href="#">isEmpty at RowMatrix.scala:441</a>	2024/01/08 06:18:07	2 s	1/1 (1 skipped)	1/1 (10 skipped)
18 (22)	Job group for statement 22 <a href="#">treeAggregate at Statistics.scala:58</a>	2024/01/08 06:17:50	17 s	2/2 (1 skipped)	28/28 (10 skipped)
17 (22)	Job group for statement 22 <a href="#">first at RowMatrix.scala:62</a>	2024/01/08 06:17:48	2 s	1/1 (1 skipped)	1/1 (10 skipped)
16 (22)	Job group for statement 22 <a href="#">first at PCA.scala:44</a>	2024/01/08 06:17:42	6 s	2/2	11/11
15 (21)	Job group for statement 21 <a href="#">treeAggregate at RowMatrix.scala:171</a>	2024/01/08 06:17:11	22 s	2/2 (1 skipped)	28/28 (10 skipped)

# Traitement des images - Suite :

## Sauvegarde sur S3 sur le format parquet

**Objets (25)** [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](#)

Copier l'URI S3

Copier l'URL

Télécharger

Ouvrir

Supprimer

Actions ▼

Créer un dossier

Charger

< 1 >



# Conclusion

- Le Projet a permis de déployer un développement de Machine Learning sur le cloud Big Data, en utilisant :
  - Apache spark et Pyspark pour les traitements distribués.
  - AWS, IAM pour la gestion des utilisateurs et des autorisations.
  - AWS S3 pour le stockage des données.
  
- Limites :
  - Outil payant







MERCI DE VOTRE ATTENTION !

Mauyves NKONDO

Mentor : Hamza Tajmouati