



Formation Data Scientist

OpenClassrooms

Projet 8: Déployez un modèle dans le cloud

Etudiant : Monine Chan

Evaluateur : Ugo Aubri

Mardi 14 Mars 2023



PLAN

1. Rappel la problématique et du jeu de données (3 min)
2. Présentation du processus de création de l'environnement Big Data, S3 et EMR (6 min)
3. Présentation de la réalisation de la chaîne de traitement des images dans un environnement Big Data dans le cloud (6 min)
4. Démonstration d'exécution du script PYSpark sur le Cloud (2 min)
5. Conclusion



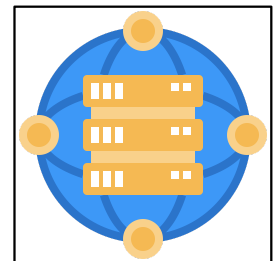
PLAN

1. Rappel la problématique et du jeu de données (3 min)
2. Présentation du processus de création de l'environnement Big Data, S3 et EMR (6 min)
3. Présentation de la réalisation de la chaîne de traitement des images dans un environnement Big Data dans le cloud (6 min)
4. Démonstration d'exécution du script PYSpark sur le Cloud (2 min)
5. Conclusion

1. RAPPEL DE LA PROBLÉMATIQUE ET DU JEU DE DONNÉES

PROBLÉMATIQUE

- La start-up de l'Agri-Tech « Fruits! » souhaite développer une application mobile qui reconnaît un fruit pris en photo et en fournit des informations (moteur de classification d'images de fruits).
- Cette startup souhaite créer une première version l'architecture Big Data nécessaire avec la contrainte d'anticiper une augmentation très rapide des données.
- Le but de ce projet est:
 - de créer un prototype de la chaîne de traitement des données par la création d'une instance EMR opérationnelle dans le cloud AWS.
 - d'extraire les features des images de fruit et réaliser une réduction de dimension PCA en PySpark.



1. RAPPEL DE LA PROBLÉMATIQUE ET DU JEU DE DONNÉES

JEU DE DONNÉES



- Le jeu de données (<https://www.kaggle.com/moltean/fruits>) contient environ 90 000 images au format .jpg de 131 fruits et légumes différents, en taille 100x100 pixels et avec différentes orientations.

- Pour notre projet:

- On utilise 10 types de fruits

Apple Golden 1	Avocado	Banana	Blueberry	Mangostan	Orange	Pear	Strawberry	Tomato 1	Watermelon
									

- Pour chaque type de fruits, on utilise 10 images (pour limiter les coûts).



PLAN

1. Rappel la problématique et du jeu de données (3 min)
2. Présentation du processus de création de l'environnement Big Data, S3 et EMR (6 min)
3. Présentation de la réalisation de la chaîne de traitement des images dans un environnement Big Data dans le cloud (6 min)
4. Démonstration d'exécution du script PYSpark sur le Cloud (2 min)
5. Conclusion

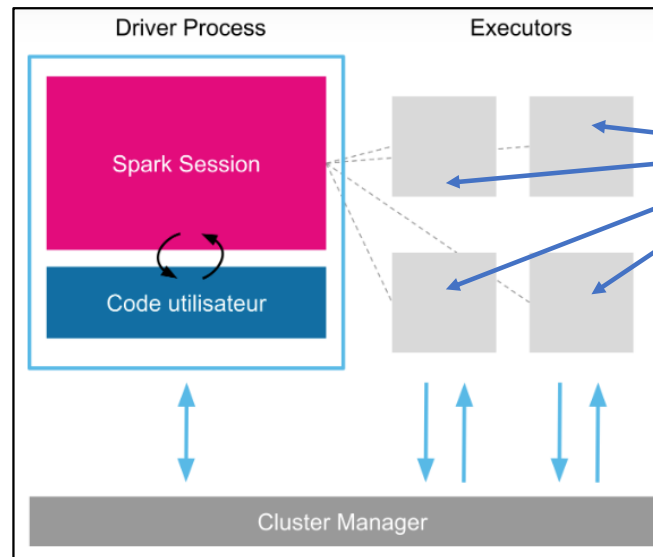
2. PROCESSUS DE LA CRÉATION DE L'ENVIRONNEMENT BIG DATA, S3 ET EMR



➤ On utilise **Spark** : qu'est ce que c'est?



Spark est un framework de calcul distribué (i.e. réparti sur plusieurs machines comme un cluster de calcul) pour le traitement et l'analyse de données massives.



Processus Java Virtual Machine
→ nombre de CPU et quantité
de mémoire configurables

➤ **Pourquoi utiliser Spark?**

Car Spark écrit les données en RAM et non sur le disque dur ce qui le rend plus rapide pour le traitement des données par rapport à un framework comme Hadoop MapReduce

2. PROCESSUS DE LA CRÉATION DE L'ENVIRONNEMENT BIG DATA, S3 ET EMR

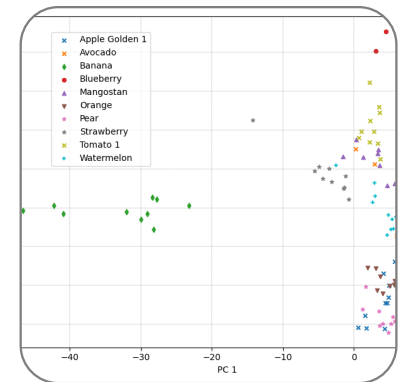
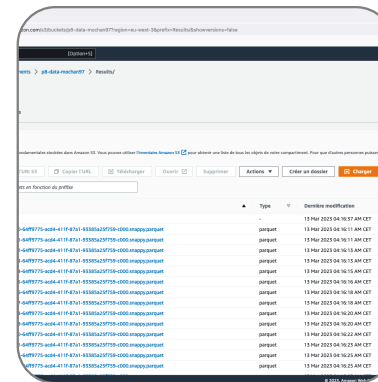
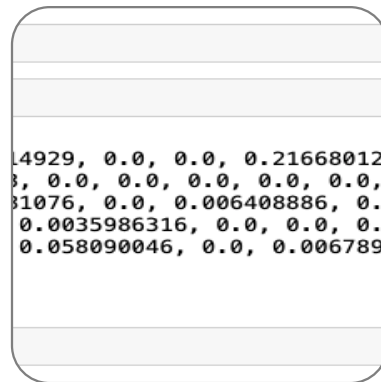
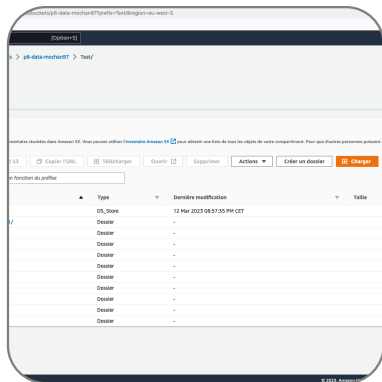


- **Briques d'une architecture Big Data en utilisant Amazon Web Services (AWS):**
 - **Service de Calcul distribué** (Cluster AWS Elastic Map Reduce = AWS EMR) : exécuter un calcul sur des ensembles de données plus petit et agréger les résultats intermédiaires obtenus pour construire le résultat final.
 - **Service de Stockage** (Bucket AWS Simple Storage Solution = AWS S3) : stocker des fichiers avec des droits d'accès sécurisés sans limite de place
 - **Résilience aux pannes** : duplication des données et partitionnement
 - **RGPD** : stockage des données dans la région / le pays pertinent

2. PROCESSUS DE LA CRÉATION DE L'ENVIRONNEMENT BIG DATA, S3 ET EMR



CHAINE DE TRAITEMENT



Stockage des données d'entrée

- Bucket AWS S3
- Images au format .jpg

Calcul distribué

- Cluster AWS EMR
- Extraction des features avec MobileNetV2

Stockage des résultats

- Bucket AWS S3
- Features au Format parquet

Réduction de dimension

- PCA avec Pyspark



PLAN

1. Rappel la problématique et du jeu de données (3 min)
2. Présentation du processus de création de l'environnement Big Data, S3 et EMR (6 min)
3. **Présentation de la réalisation de la chaîne de traitement des images dans un environnement Big Data dans le cloud (6 min)**
4. Démonstration d'exécution du script PYSpark sur le Cloud (2 min)
5. Conclusion

3. RÉALISATION DE LA CHAÎNE DE TRAITEMENT DES IMAGES DANS UN ENVIRONNEMENT BIG DATA DANS LE CLOUD



➤ Création du bucket S3 dans la région eu-west-3 (Paris):

The screenshot shows the AWS S3 console interface. The browser address bar displays the URL: `https://s3.console.aws.amazon.com/s3/buckets/p8-data-mochan97?region=eu-west-3&tab=properties`. The console header includes a search bar, a notification bell, a help icon, and a user profile dropdown labeled 'moninechan'. The breadcrumb navigation shows 'Amazon S3 > Compartiments > p8-data-mochan97'. The bucket name 'p8-data-mochan97' is prominently displayed with an 'Info' link. Below this, a tabbed interface shows 'Objets', 'Propriétés' (selected), 'Autorisations', 'Métriques', 'Gestion', and 'Points d'accès'. The 'Présentation des compartiments' section contains a table with the following data:

Région AWS	Amazon Resource Name (ARN)	Date de création
EU (Paris) eu-west-3	arn:aws:s3::p8-data-mochan97	03 Mar 2023 11:25:58 AM CET

3. RÉALISATION DE LA CHAÎNE DE TRAITEMENT DES IMAGES DANS UN ENVIRONNEMENT BIG DATA DANS LE CLOUD



➤ Chargement des fichiers de départ dans le bucket S3 :

The screenshot shows the AWS S3 console interface. The left sidebar displays the 'Amazon S3' menu with options like 'Compartiments', 'Points d'accès', and 'Storage Lens'. The main content area shows the 'Test/' bucket. The 'Objets (11)' section lists the objects in the bucket. A green box highlights the 'Apple Golden 1/' folder in the left sidebar, and another green box highlights the list of objects in the 'Apple Golden 1/' folder. A green arrow points from the folder name to the object list.

Nom	Type	Dernière modification	Taille	Classe de stockage
.DS_Store	DS_Store	04 Mar 2023 06:03:50 PM CET	6.0 Ko	Standard
0_100.jpg	jpg	04 Mar 2023 06:03:50 PM CET	4.8 Ko	Standard
233_100.jpg	jpg	04 Mar 2023 06:03:50 PM CET	5.5 Ko	Standard
311_100.jpg	jpg	04 Mar 2023 06:03:50 PM CET	4.9 Ko	Standard
45_100.jpg	jpg	04 Mar 2023 06:03:50 PM CET	5.4 Ko	Standard
r_0_100.jpg	jpg	04 Mar 2023 06:03:50 PM CET	5.0 Ko	Standard
r_10_100.jpg	jpg	04 Mar 2023 06:03:50 PM CET	5.0 Ko	Standard
r_100_100.jpg	jpg	04 Mar 2023 06:03:50 PM CET	5.4 Ko	Standard
r_201_100.jpg	jpg	04 Mar 2023 06:03:50 PM CET	5.7 Ko	Standard
r_24_100.jpg	jpg	04 Mar 2023 06:03:50 PM CET	5.0 Ko	Standard
r_306_100.jpg	jpg	04 Mar 2023 06:03:50 PM CET	5.1 Ko	Standard

3. RÉALISATION DE LA CHAÎNE DE TRAITEMENT DES IMAGES DANS UN ENVIRONNEMENT BIG DATA DANS LE CLOUD



➤ Création d'une instance de cluster AWS EMR avec JupyterHub activé :

The screenshot displays the AWS Management Console for an Amazon EMR cluster. The cluster name is 'P8_Fruits_EMR_6_10_0_V5_03' and it is currently in the 'En attente' (Pending) state. The console shows various tabs for cluster management, including 'Résumé' (Summary), 'Historique de l'application' (Application history), 'Surveillance' (Monitoring), 'Matériel' (Hardware), 'Configurations', 'Événements' (Events), 'Étapes' (Steps), and 'Actions d'amorçage' (Startup actions). The 'Résumé' tab is selected, showing the cluster's ID, creation date, and other key information. The 'Détails de configuration' (Configuration details) tab is also visible, showing the application and distribution settings. The 'Réseau et matériel' (Network and hardware) tab shows the VPC, subnets, and security groups. The 'Sécurité et accès' (Security and access) tab shows the IAM role and security groups. The 'Application user interfaces' (Application user interfaces) tab shows the JupyterHub and Spark history server connections.

Section	Détails
Résumé	<ul style="list-style-type: none">ID : j-2966IDXEE6ZBMDate de création : 13-03-2023 03:54 (UTC+1)Temps écoulé : 1 heure, 25 minutesRésiliation automatique : Cluster waitsProtection de la résiliation : DésactivéBalises : --DNS public principal : ec2-13-38-44-18.eu-west-3.compute.amazonaws.com
Détails de configuration	<ul style="list-style-type: none">Étiquette de version : emr-6.10.0Distribution Hadoop : Amazon 3.3.3Application : JupyterHub 1.5.0, TensorFlow 2.11.0, Spark 3.3.1URI de connexion : s3://aws-logs-067745069111-eu-west-3/elasticmapreduce/Vue cohérente EMRFS : DésactivéID d'AMI personnalisée : --Version d'Amazon Linux : 2.0.20230207.0
Réseau et matériel	<ul style="list-style-type: none">Zone de disponibilité : eu-west-3aID de sous-réseau (subnet) : subnet-0a52da2399d12d6a1Maître : En cours d'exécution 1 m5.xlargePrincipal : En cours d'exécution 2 m5.xlargeTâche : --Cluster scaling : Not enabledRésiliation automatique : Arrêter en cas d'inactivité pour 1 heure
Sécurité et accès	<ul style="list-style-type: none">Nom de clé : aws_cours_openclassroomsProfil d'instance EC2 : EMR_EC2_DefaultRoleRôle EMR : EMR_DefaultRoleVisible pour tous les utilisateurs : TousGroupes de sécurité pour le principal : sg-08393b8e674f238f1 (ElasticMapReduce-principal)Groupes de sécurité pour la base et les tâches : sg-06c2d874ce0a98f4f (ElasticMapReduce-slave)
Application user interfaces	<ul style="list-style-type: none">Service d'historique : Spark history server, YARN timeline serverConnexions : Nom du nœud HDFS, Serveur d'historique Spark, JupyterHub, Gestionnaire de ressources

3. RÉALISATION DE LA CHAÎNE DE TRAITEMENT DES IMAGES DANS UN ENVIRONNEMENT BIG DATA DANS LE CLOUD



➤ Configuration du cluster :

https://eu-west-3.console.aws.amazon.com/elasticmapreduce/home?region=eu-west-3#cluster-details:j-2966IDXE6ZBM

Chercher [Option+S]

The new EMR console will become the default console soon. Switch to the new console. If you want, you can still switch back. Learn more

Cloner Réinitialiser Exporter AWS CLI

Cluster : P8_Fruits_EMR_6_10_0_V5_03 **En attente** Cluster ready to run steps.

Récapitulatif Historique de l'application Surveillance Matériel Configurations Événements Étapes Actions d'amorçage

Ajouter un groupe d'instances de tâches

Groupes d'instances

Filtre : Filtrer les groupes d'instances... 2 groupes d'instances (tous chargés)

ID	Status	Nom et type de nœud	Type d'instance	Nombre d'instances	Option d'achat
ig-1FS8LP7514O5L	En cours d'exécution	CORE Groupe d'instances principal - 2	m5.xlarge 4 Cœurs virtuels, 16 Gio de mémoire, stockage EBS uniquement Stockage sur EBS : 64 Gio	2 Instances Redimensionner	A la demande ⓘ
ig-2O62WVCYS8BIE	En cours d'exécution	MASTER Groupe d'instances maître - 1	m5.xlarge 4 Cœurs virtuels, 16 Gio de mémoire, stockage EBS uniquement Stockage sur EBS : 64 Gio	1 Instances	A la demande ⓘ

Cluster Scaling Policy

No scaling enabled [Edit](#)

Résiliation automatique

Sélectionnez une heure pour que le cluster soit arrêté une fois que le cluster est inactif. Choisissez un minimum de 1 minute ou un maximum de 24 heures. [En savoir plus](#)

Résiliation automatique : Arrêter en cas d'inactivité pour 1 heure

3. RÉALISATION DE LA CHAÎNE DE TRAITEMENT DES IMAGES DANS UN ENVIRONNEMENT BIG DATA DANS LE CLOUD



➤ Amorçage: installation des packages

The screenshot displays the AWS EMR console interface. The browser address bar shows the URL: `https://eu-west-3.console.aws.amazon.com/elasticmapreduce/home?region=eu-west-3#cluster-details:j-2966IDXEE6ZBM`. A notification banner at the top states: "The new EMR console will become the default console soon. Switch to the new console. If you want, you can still switch back. Learn more". Below the notification are buttons for "Cloner", "Résilier", and "Exporter AWS CLI".

The cluster details for "Cluster : P8_Fruits_EMR_6_10_0_V5_03" are shown, with the status "En attente" (Waiting) and the note "Cluster ready to run steps." The "Actions d'amorçage" (Bootstrap Actions) tab is selected, displaying a table with one action:

Nom	Emplacement	Arguments facultatifs
1 Action personnalisée	s3://p8-data-mochan97/bootstrap-emr-v5.sh	

A green arrow points from the script location in the table to a terminal window titled "bootstrap-emr-v5.sh". The terminal displays the following commands:

```
1 #!/bin/bash
2 sudo python3 -m pip install -U setuptools
3 sudo python3 -m pip install -U pip
4 sudo python3 -m pip install wheel
5 sudo python3 -m pip install pillow
6 sudo python3 -m pip install pandas
7 sudo python3 -m pip install pyarrow
8 sudo python3 -m pip install s3fs
9 sudo python3 -m pip install fsspec
10 sudo python3 -m pip install keras==2.11.0
11 sudo python3 -m pip install matplotlib
12 sudo python3 -m pip install plotly
13
```



PLAN

1. Rappel la problématique et du jeu de données (3 min)
2. Présentation du processus de création de l'environnement Big Data, S3 et EMR (6 min)
3. Présentation de la réalisation de la chaîne de traitement des images dans un environnement Big Data dans le cloud (6 min)
4. **Démonstration d'exécution du script PYSpark sur le Cloud (2 min)**
5. Conclusion

3. DÉMONSTRATION DE L'EXÉCUTION DU SCRIPT PYSPARK DANS LE CLOUD



➤ Exécution du script PySpark dans une instance de cluster AWS EMR, accès via tunnel SSH:

```
In [1]: %info

Current session configs: {'driverMemory': '1000M', 'executorCores': 2, 'proxyUser': 'jovyan', 'kind': 'pyspark'}

No active sessions.

In [2]: import pandas as pd
import numpy as np
import io
import os
import tensorflow as tf
from PIL import Image
import matplotlib.pyplot as plt
from pyspark.sql.functions import col, pandas_udf, PandasUDFType, element_at, split

Starting Spark application

ID      YARN Application ID  Kind  State  Spark UI  Driver log  User  Current session?
1  application_1678676471236_0002  pyspark  idle  Link  Link  None  ✓

SparkSession available as 'spark'.

In [3]: PATH = 's3://p8-data-mochan97'
PATH_Data = PATH+'/Test'
PATH_Result = PATH+'/Results'
print('PATH: '
      PATH+'\nPATH_Data: '
      PATH_Data+'\nPATH_Result: '+PATH_Result)

PATH:      s3://p8-data-mochan97
PATH_Data: s3://p8-data-mochan97/Test
PATH_Result: s3://p8-data-mochan97/Results
```

```
Proj_8 -- hadoop@ip-172-31-2-41:~$ ssh -i ~/.ssh/aws_cours_openclassrooms.pem -D 8157 hadoop@ec2-13-38-44-18.eu-west-3.compute.amazonaws.com
(base) wonrechanpc2 Proj_8 % ssh -i ~/.ssh/aws_cours_openclassrooms.pem -D 8157 hadoop@ec2-13-38-44-18.eu-west-3.compute.amazonaws.com
The authenticity of host 'ec2-13-38-44-18.eu-west-3.compute.amazonaws.com (13.38.44.18)' can't be established.
ECDSA key fingerprint is SHA256:vguass2S05U02akn+V0V0R3jYnR0Y0B8N0W0B8.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-13-38-44-18.eu-west-3.compute.amazonaws.com,13.38.44.18' (ECDSA) to the list of known hosts.

hadoop@ip-172-31-2-41:~$ sudo yum update
Amazon Linux 2 AMI
https://aws.amazon.com/amazon-linux-2/
5 package(s) needed for security, out of 13 available
Run "sudo yum update" to apply all updates.

hadoop@ip-172-31-2-41:~$
```

3. DÉMONSTRATION DE L'EXÉCUTION DU SCRIPT PYSPARK DANS LE CLOUD



➤ Ecritures des sorties du notebook dans l'espace de stockage cloud AWS S3 :

1.4. Exécutions des actions d'extractions de features

```
In [16]: features_df = images.repartition(24).select(col("path"),
                                                    col("label"),
                                                    featurize_udf("content").alias("features")
                                                    )

In [17]: print(PATH_Result)
s3://p8-data-mochan97/Results

In [18]: features_df.write.mode("overwrite").parquet(PATH_Result)
```

Amazon S3 > Compartiments > p8-data-mochan97 > Results/

Objets (25)

Nom	Type	Dernière modification	Taille	Classe de stockage
part-00000-64f99775-acd4-411f-87a1-93385a25f759-c000.snappy.parquet	parquet	13 Mar 2023 04:16:37 AM CET	0 o	Standard
part-00001-64f99775-acd4-411f-87a1-93385a25f759-c000.snappy.parquet	parquet	13 Mar 2023 04:16:11 AM CET	25.2 Ko	Standard
part-00002-64f99775-acd4-411f-87a1-93385a25f759-c000.snappy.parquet	parquet	13 Mar 2023 04:16:11 AM CET	21.1 Ko	Standard
part-00003-64f99775-acd4-411f-87a1-93385a25f759-c000.snappy.parquet	parquet	13 Mar 2023 04:16:13 AM CET	24.7 Ko	Standard
part-00004-64f99775-acd4-411f-87a1-93385a25f759-c000.snappy.parquet	parquet	13 Mar 2023 04:16:15 AM CET	16.9 Ko	Standard
part-00005-64f99775-acd4-411f-87a1-93385a25f759-c000.snappy.parquet	parquet	13 Mar 2023 04:16:16 AM CET	17.6 Ko	Standard
part-00006-64f99775-acd4-411f-87a1-93385a25f759-c000.snappy.parquet	parquet	13 Mar 2023 04:16:18 AM CET	20.7 Ko	Standard
part-00007-64f99775-acd4-411f-87a1-93385a25f759-c000.snappy.parquet	parquet	13 Mar 2023 04:16:18 AM CET	17.3 Ko	Standard
part-00008-64f99775-acd4-411f-87a1-93385a25f759-c000.snappy.parquet	parquet	13 Mar 2023 04:16:20 AM CET	17.1 Ko	Standard
part-00009-64f99775-acd4-411f-87a1-93385a25f759-c000.snappy.parquet	parquet	13 Mar 2023 04:16:20 AM CET	20.4 Ko	Standard
part-00010-64f99775-acd4-411f-87a1-93385a25f759-c000.snappy.parquet	parquet	13 Mar 2023 04:16:22 AM CET	20.9 Ko	Standard
part-00011-64f99775-acd4-411f-87a1-93385a25f759-c000.snappy.parquet	parquet	13 Mar 2023 04:16:23 AM CET	20.8 Ko	Standard
part-00012-64f99775-acd4-411f-87a1-93385a25f759-c000.snappy.parquet	parquet	13 Mar 2023 04:16:25 AM CET	17.1 Ko	Standard
part-00013-64f99775-acd4-411f-87a1-93385a25f759-c000.snappy.parquet	parquet	13 Mar 2023 04:16:25 AM CET	15.2 Ko	Standard

3. DÉMONSTRATION DE L'ÉXÉCUTION DU SCRIPT PYSPARK DANS LE CLOUD



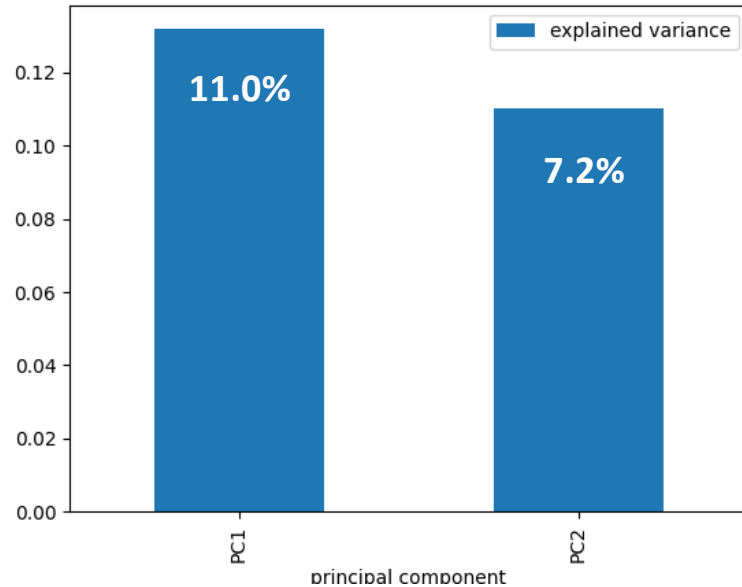
➤ Ecritures des sorties du notebook dans l'espace de stockage cloud AWS S3 :

```
In [24]: df.to_csv(PATH + '/df_avec_features.csv')
```

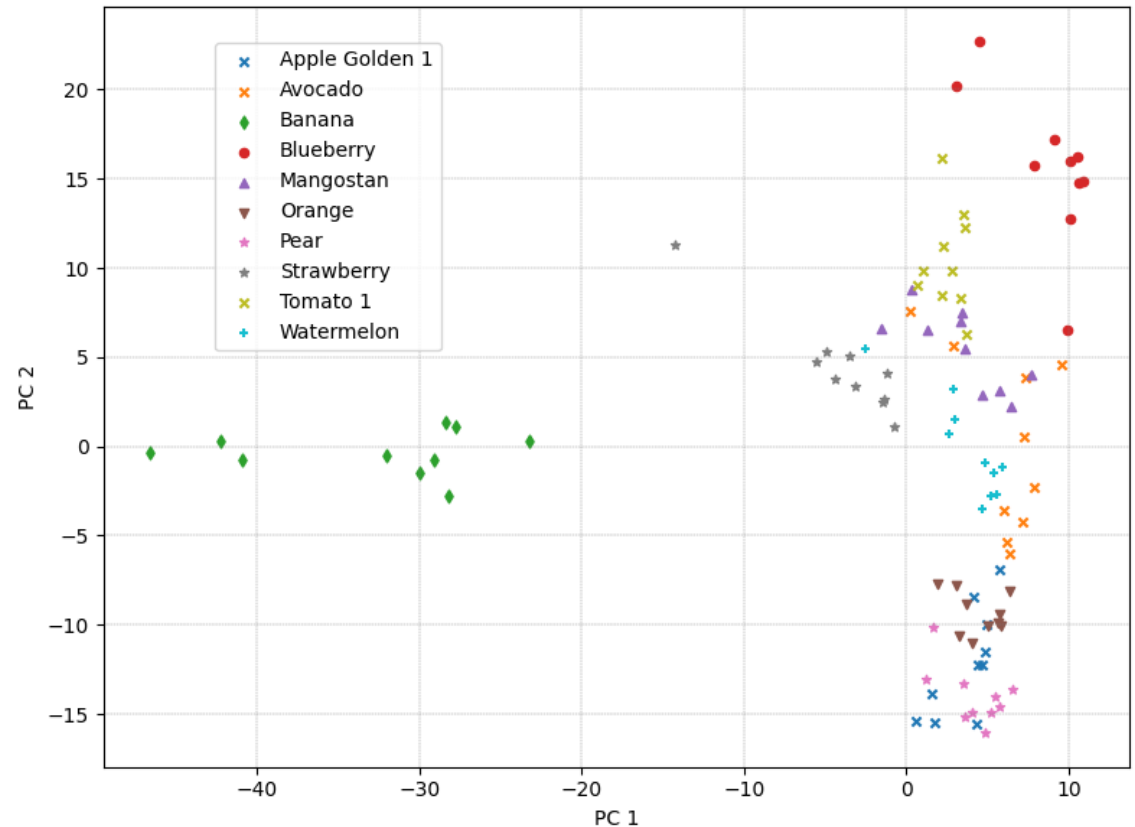
The screenshot shows the AWS S3 console interface for the bucket 'p8-data-mochan97'. The 'Objets' tab is selected, displaying a list of 13 objects. The object 'df_avec_features.csv' is highlighted with a green box, and a green arrow points from the code cell above to it.

	Nom	Type	Dernière modification	Taille	Classe de stockage
<input type="checkbox"/>	.DS_Store	DS_Store	12 Mar 2023 08:57:35 PM CET	6.0 Ko	Standard
<input type="checkbox"/>	bootstrap-emr-v2.sh	sh	06 Mar 2023 06:01:41 AM CET	339.0 o	Standard
<input type="checkbox"/>	bootstrap-emr-v3.sh	sh	09 Mar 2023 04:42:51 PM CET	371.0 o	Standard
<input type="checkbox"/>	bootstrap-emr-v4.sh	sh	09 Mar 2023 06:32:28 PM CET	413.0 o	Standard
<input type="checkbox"/>	bootstrap-emr-v5-iris.sh	sh	12 Mar 2023 06:11:00 PM CET	499.0 o	Standard
<input type="checkbox"/>	bootstrap-emr-v5.sh	sh	09 Mar 2023 06:58:31 PM CET	413.0 o	Standard
<input type="checkbox"/>	bootstrap-emr-v6.sh	sh	12 Mar 2023 08:57:35 PM CET	452.0 o	Standard
<input type="checkbox"/>	bootstrap-emr.sh	sh	04 Mar 2023 12:19:01 AM CET	331.0 o	Standard
<input type="checkbox"/>	df_avec_features.csv	csv	13 Mar 2023 04:16:38 AM CET	13.1 Ko	Standard
<input type="checkbox"/>	Iris/	Dossier	-	-	-
<input type="checkbox"/>	Jupyter/	Dossier	-	-	-
<input type="checkbox"/>	Results/	Dossier	-	-	-
<input type="checkbox"/>	Test/	Dossier	-	-	-

4. DÉMONSTRATION DE L'ÉXÉCUTION DU SCRIPT PYSPARK DANS LE CLOUD



- On réalise une PCA à partir des features extraites des images.
- On choisit un nombre de composants = 2 et on représente le résultats de la PCA en fonction des deux composantes principales PC1 et PC2.





PLAN

1. Rappel la problématique et du jeu de données (3 min)
2. Présentation du processus de création de l'environnement Big Data, S3 et EMR (6 min)
3. Présentation de la réalisation de la chaîne de traitement des images dans un environnement Big Data dans le cloud (6 min)
4. Démonstration d'exécution du script PYSpark sur le Cloud (2 min)
5. Conclusion

4. CONCLUSION



- Une chaîne de traitement des images a été mise en place dans le cloud AWS avec un cluster EMR.
- Le framework PySpark a été utilisé pour traiter une centaine d'images.
- L'algorithme d'extraction des features utilisé est MobileNetV2.
- Acquisition de nouvelles compétences : cloud AWS (EMR en particulier).

Perspectives:

- Réaliser une étape de transfer learning.
- Tester sur un plus grand nombre d'images (limitation coût).



Merci de votre attention!