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

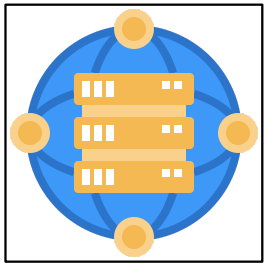
1. Rappel la problématique et du jeu de données (3 min)
2. Présentation du processus de création de l'environnement Big Data, S3 et EMR (6 min)
3. Présentation de la réalisation de la chaîne de traitement des images dans un environnement Big Data dans le cloud (6 min)
4. Démonstration d'exécution du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



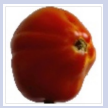
- Le but de ce projet est:
  - de créer un prototype de la chaîne de traitement des données par la création d'une instance EMR opérationnelle dans le cloud AWS.
  - d'extraire les features des images de fruit et réaliser une réduction de dimension PCA en PySpark.

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).

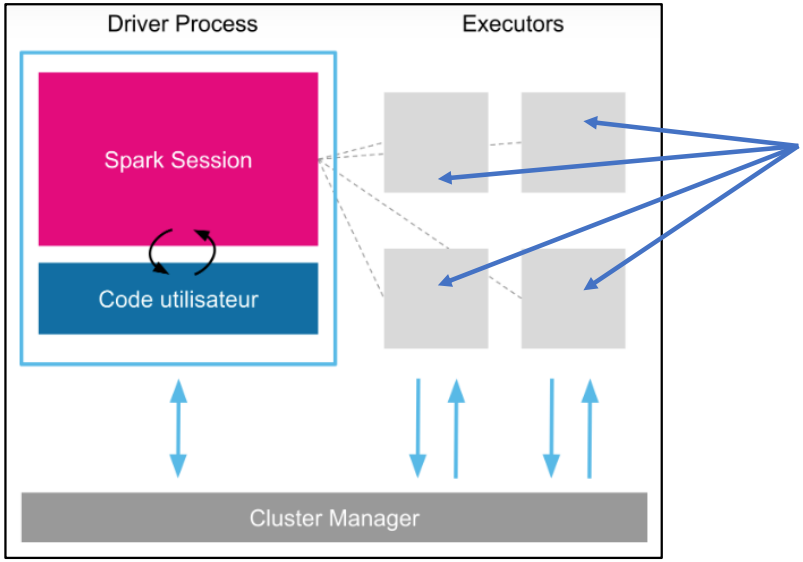
PLAN

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.

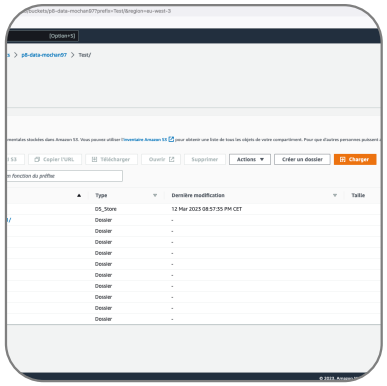
- **Service de Stockage** (Bucket AWS Simple Storage Solution = AWS S3) : stocker des fichiers avec des droits d'accès sécurisés sans limite de place

- **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

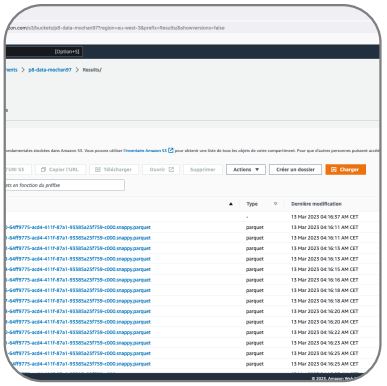


Stockage des données d'entrée

- Bucket AWS S3
- Images au format .jpg

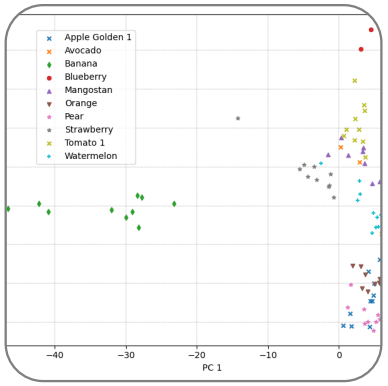
Calcul distribué

- Cluster AWS EMR
- Extraction des features avec MobileNetV2



Stockage des résultats

- Bucket AWS S3
- Features au Format parquet



Réduction de dimension

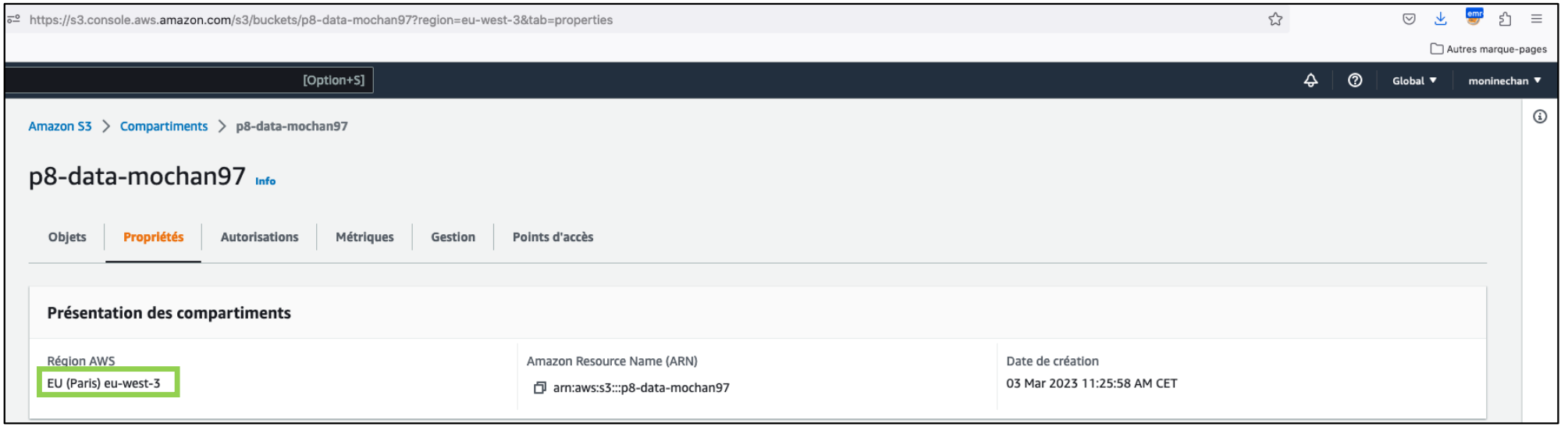
- PCA avec Pyspark

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

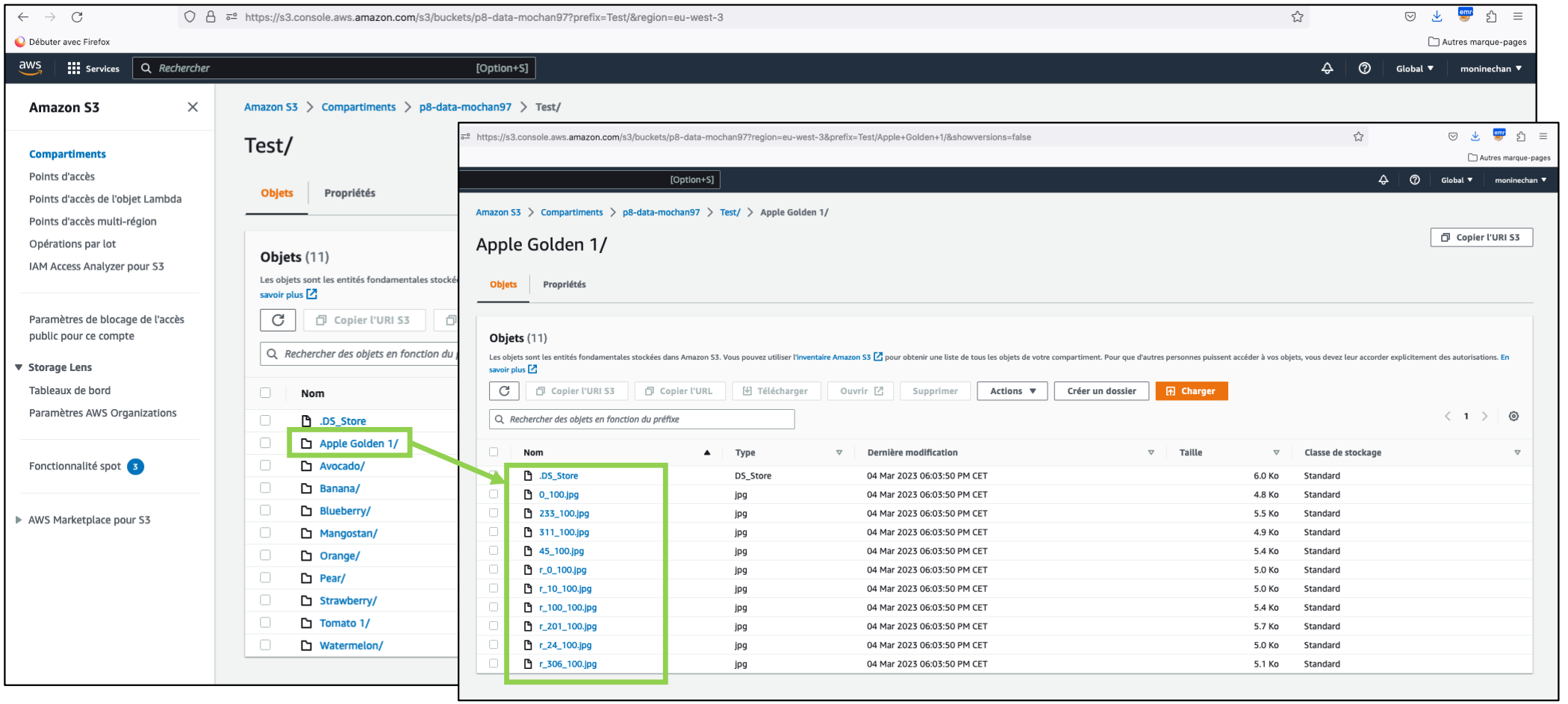
3. RÉALISATION DE LA CHAÎNE DE TRAITEMENT DES IMAGES DANS UN ENVIRONNEMENT BIG DATA DANS LE CLOUD

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



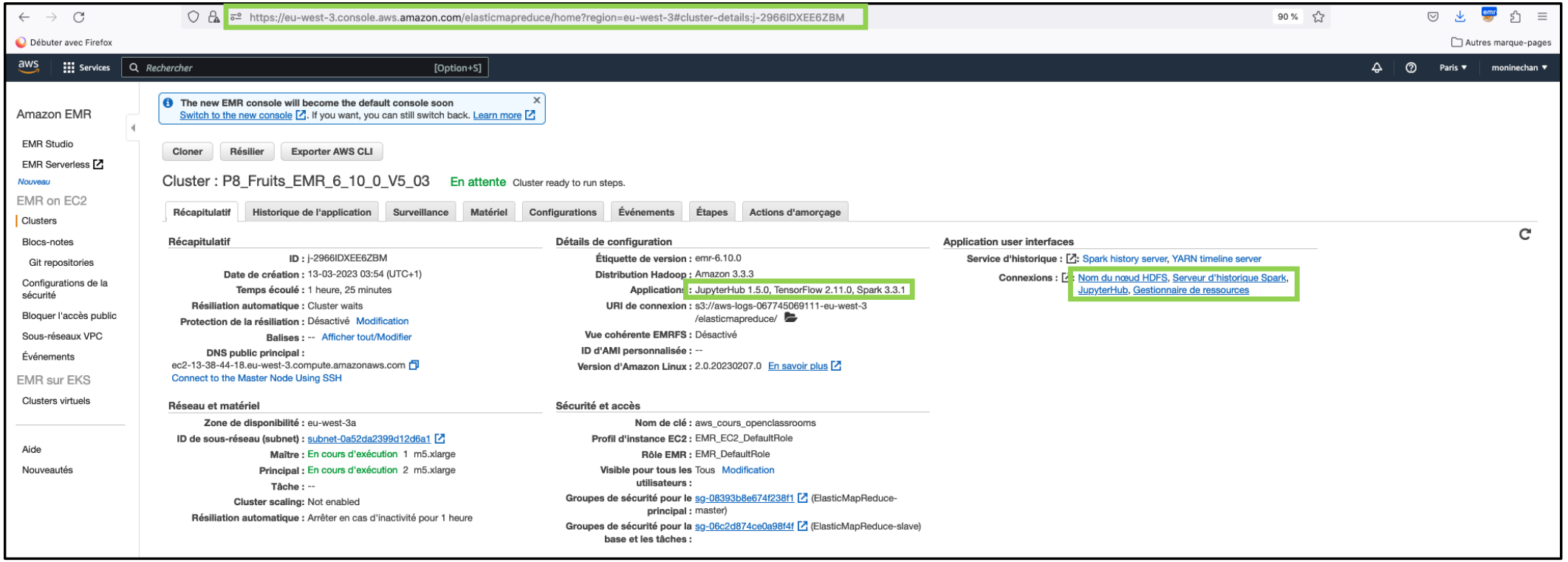
3. RÉALISATION DE LA CHAÎNE DE TRAITEMENT DES IMAGES DANS UN ENVIRONNEMENT BIG DATA DANS LE CLOUD

➢ Chargement des fichiers de départ dans le bucket S3 :



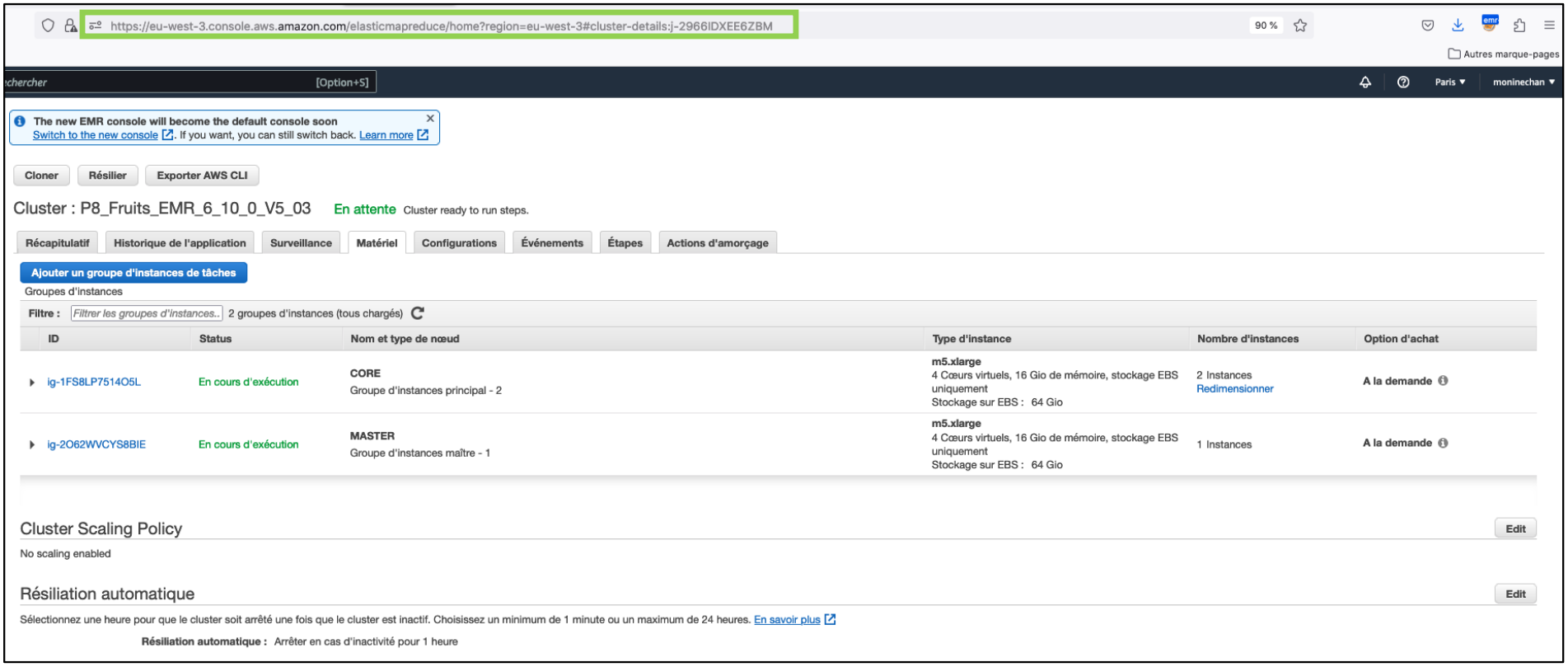
3. RÉALISATION DE LA CHAÎNE DE TRAITEMENT DES IMAGES DANS UN ENVIRONNEMENT BIG DATA DANS LE CLOUD

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



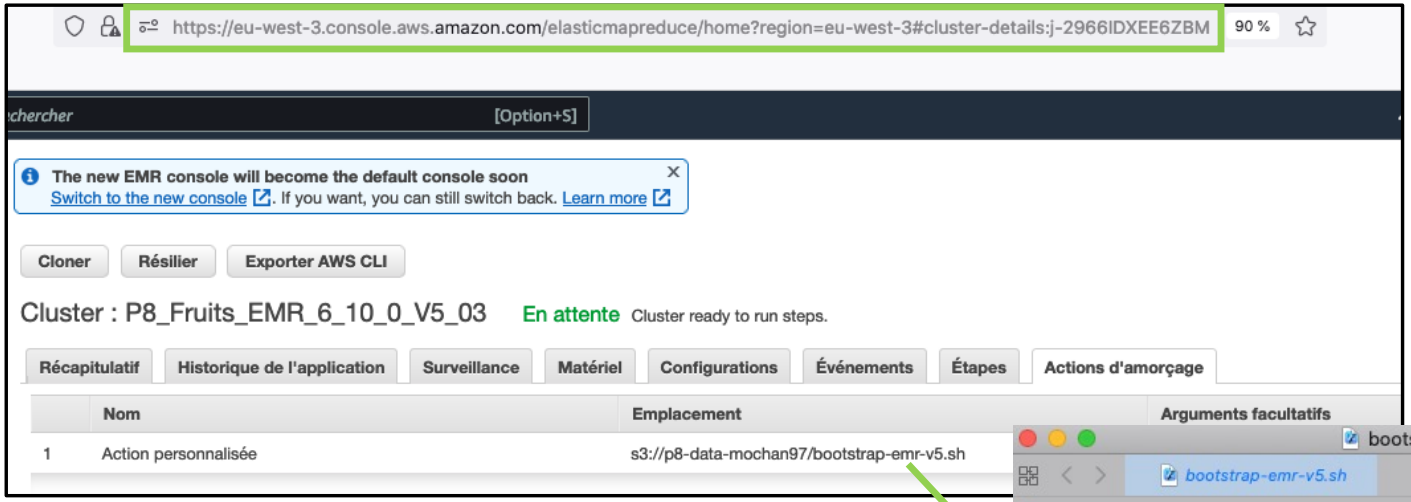
3. RÉALISATION DE LA CHAÎNE DE TRAITEMENT DES IMAGES DANS UN ENVIRONNEMENT BIG DATA DANS LE CLOUD

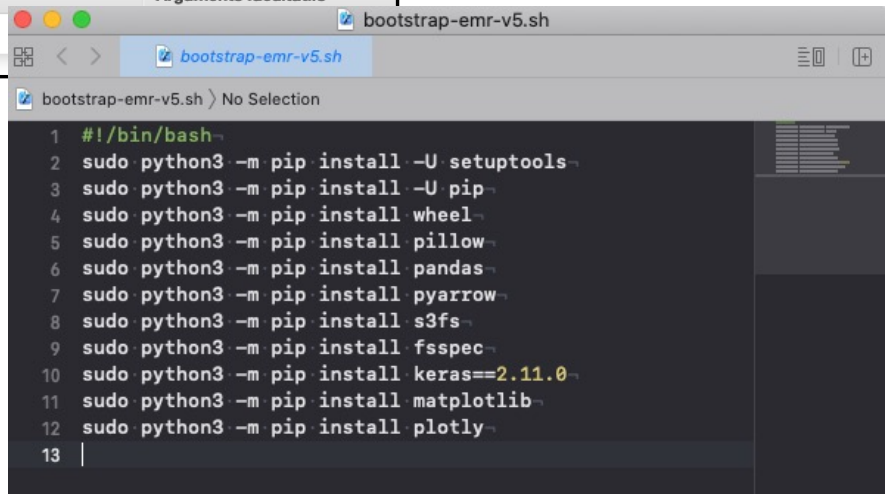
➢ Configuration du cluster :



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

- Amorçage: installation des packages



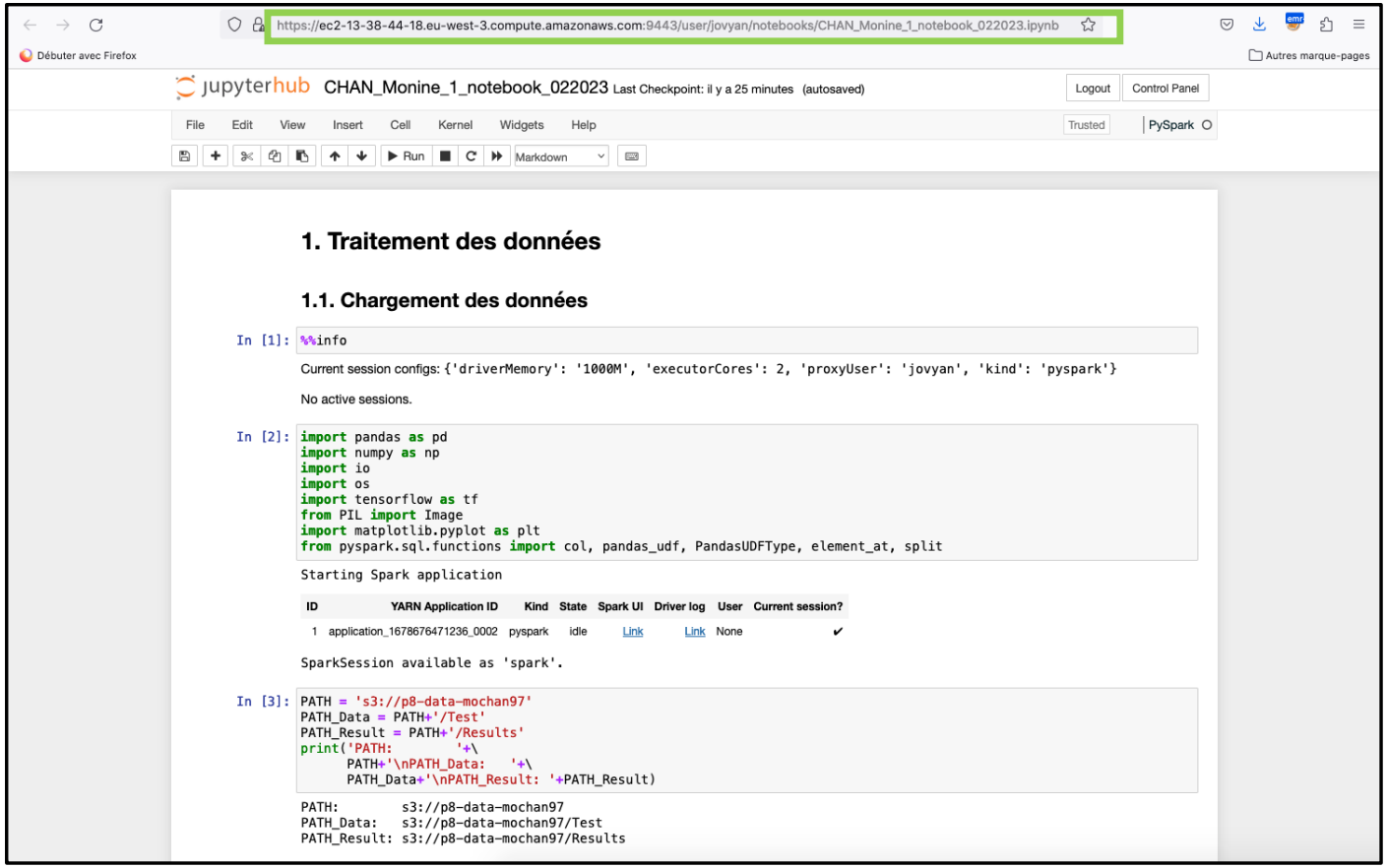


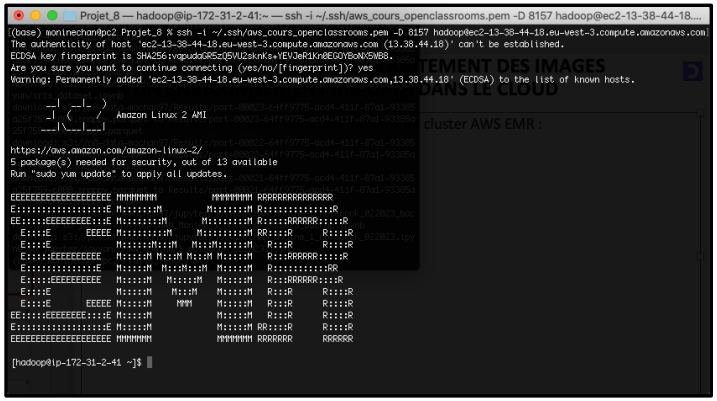
PLAN

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'ÉXÉCUTION DU SCRIPT PYSPARK DANS LE CLOUD

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



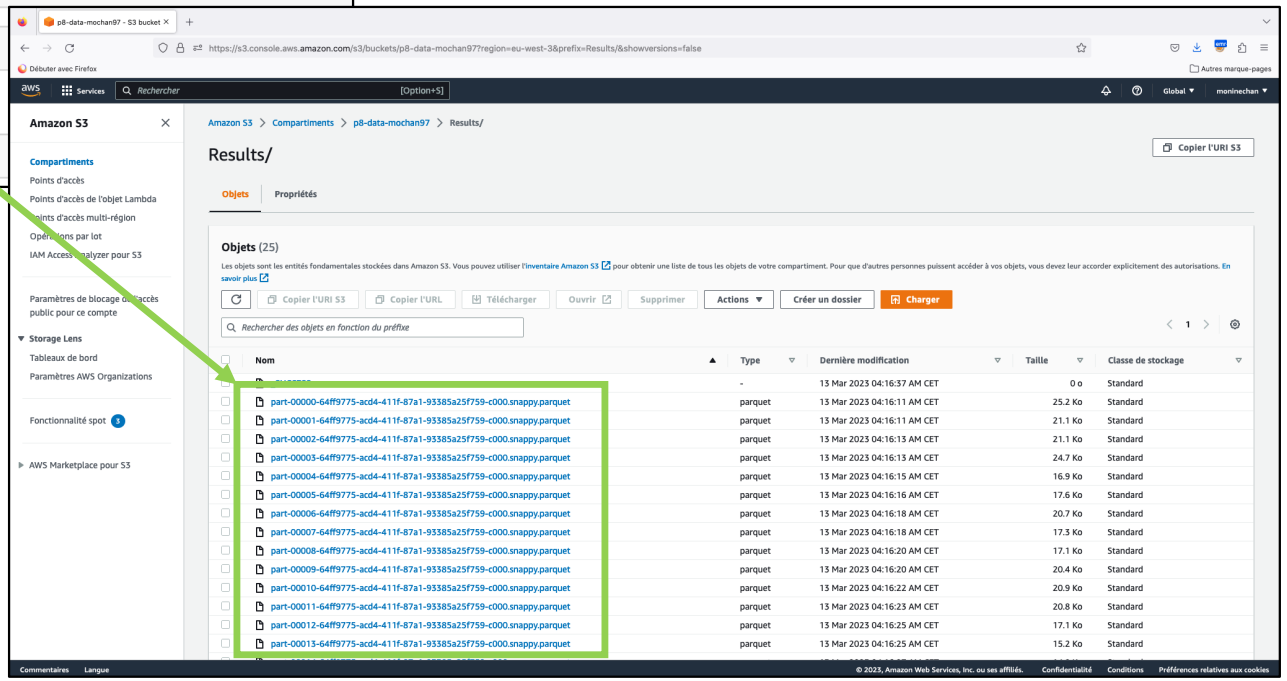


3. DÉMONSTRATION DE L'ÉXÉCUTION DU SCRIPT PYSPARK DANS LE CLOUD

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

1.4. Exécutions des actions d'extractions de features

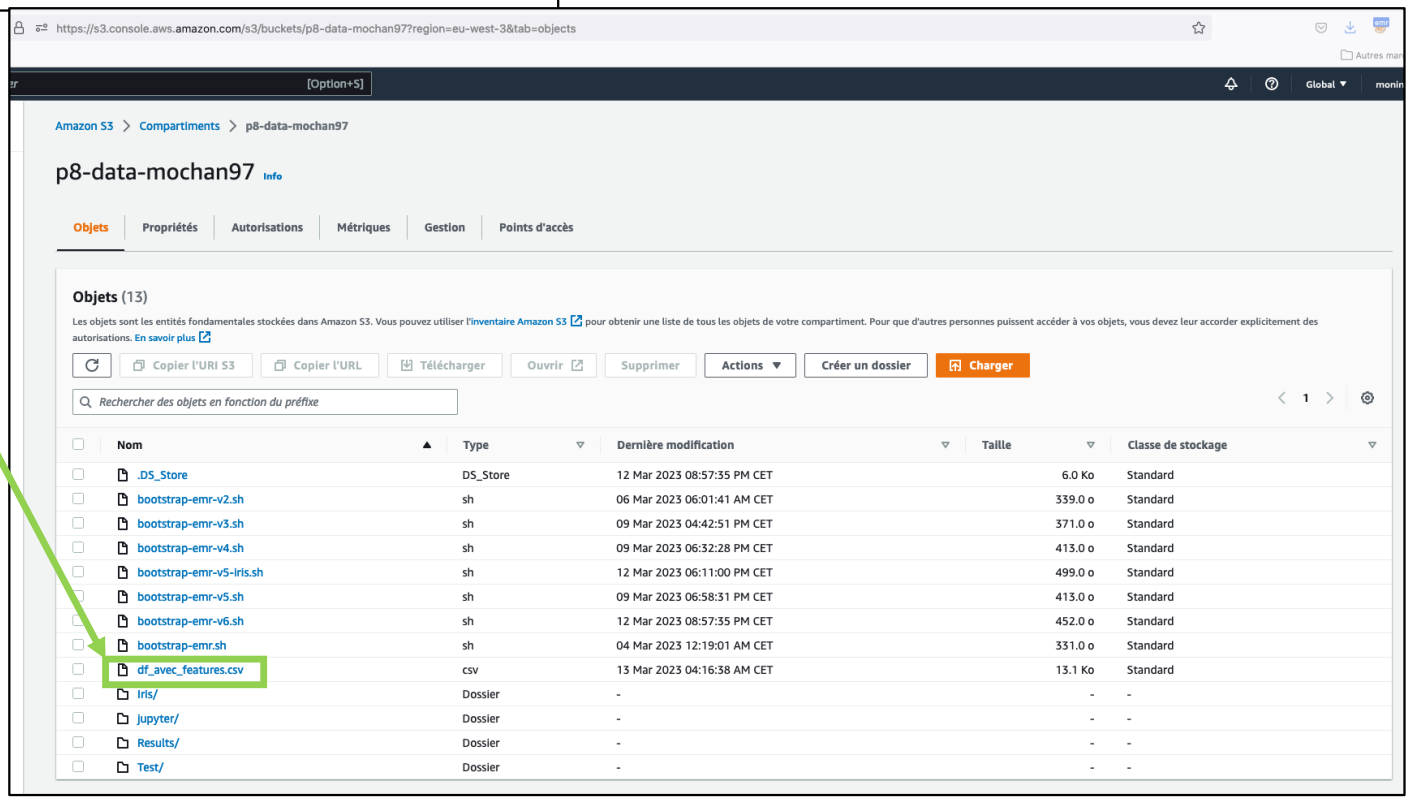
```
In [16]: features_df = images.repartition(24).select(col("path"),
                                                     col("label"),
                                                     featurize_udf("content").alias("features")
                                                    )

In [17]: print(PATH_Result)

         s3://p8-data-mochan97/Results

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



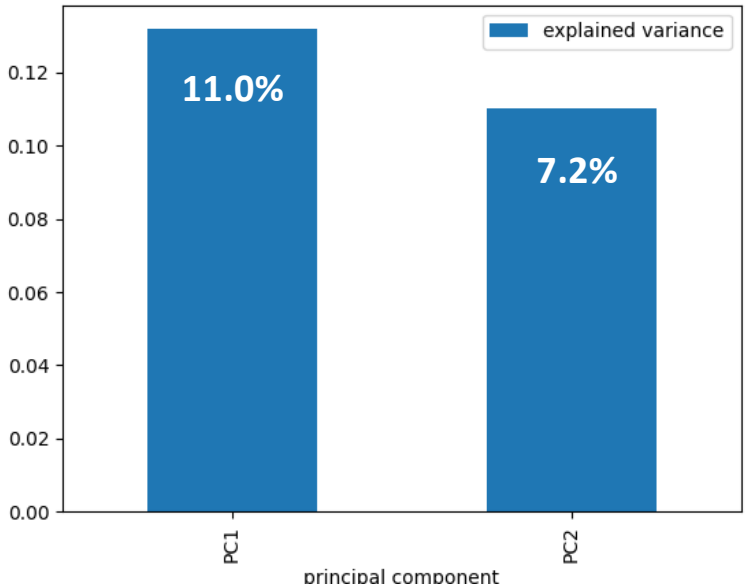
3. DÉMONSTRATION DE L'ÉXÉCUTION DU SCRIPT PYSPARK DANS LE CLOUD

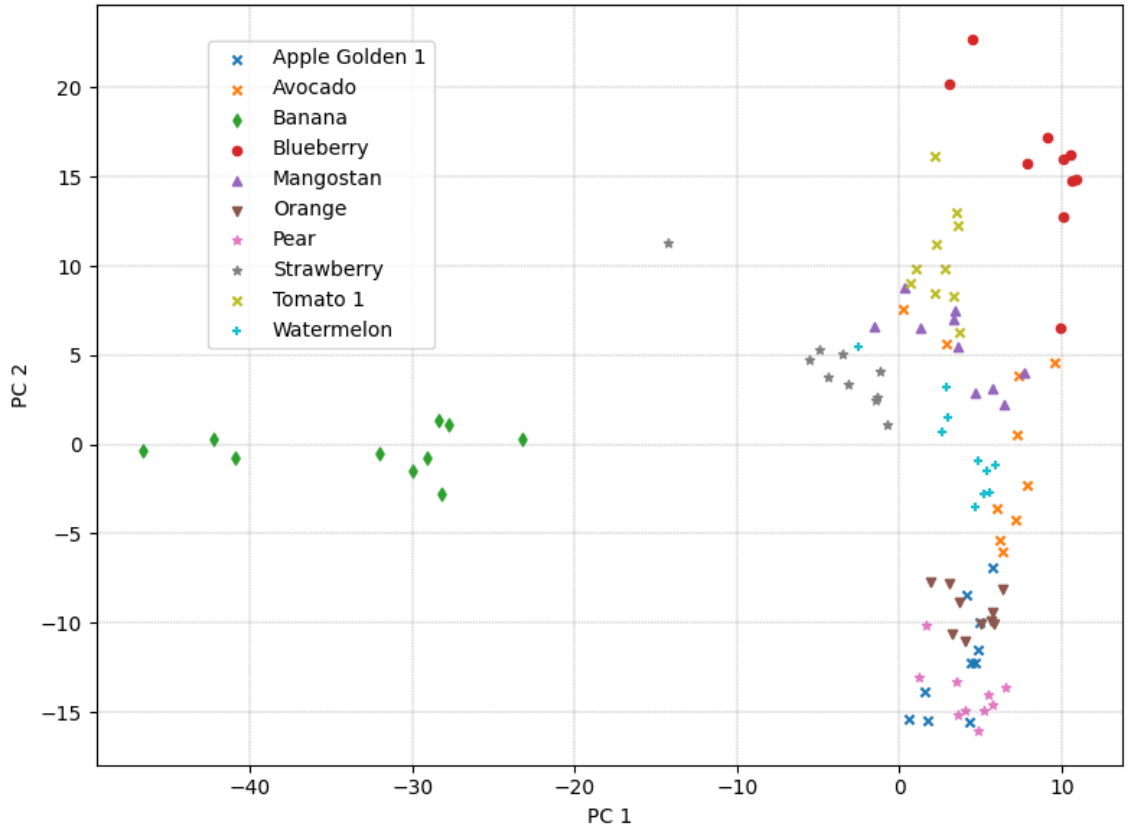
➢Ecritures des sorties du notebook dans l'espace de stockage cloud AWS S3 :

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



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!