Projet 11

Réalisez un traitement dans un environnement Big Data sur le Cloud

Zaccaria Amillou

Sommaire

- Contexte
- Données
- EDA
- Modélisation
- RGPD
- Cloud
- Conclusions

Contexte

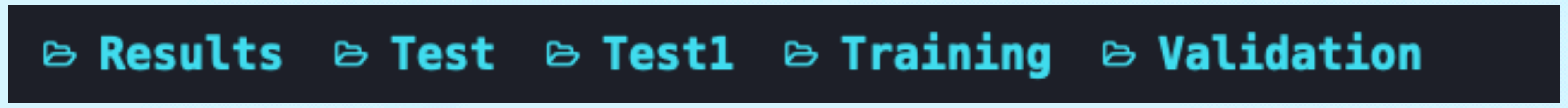
Entreprise

Objectif

Mise en place d’une traitement des données sur le cloud

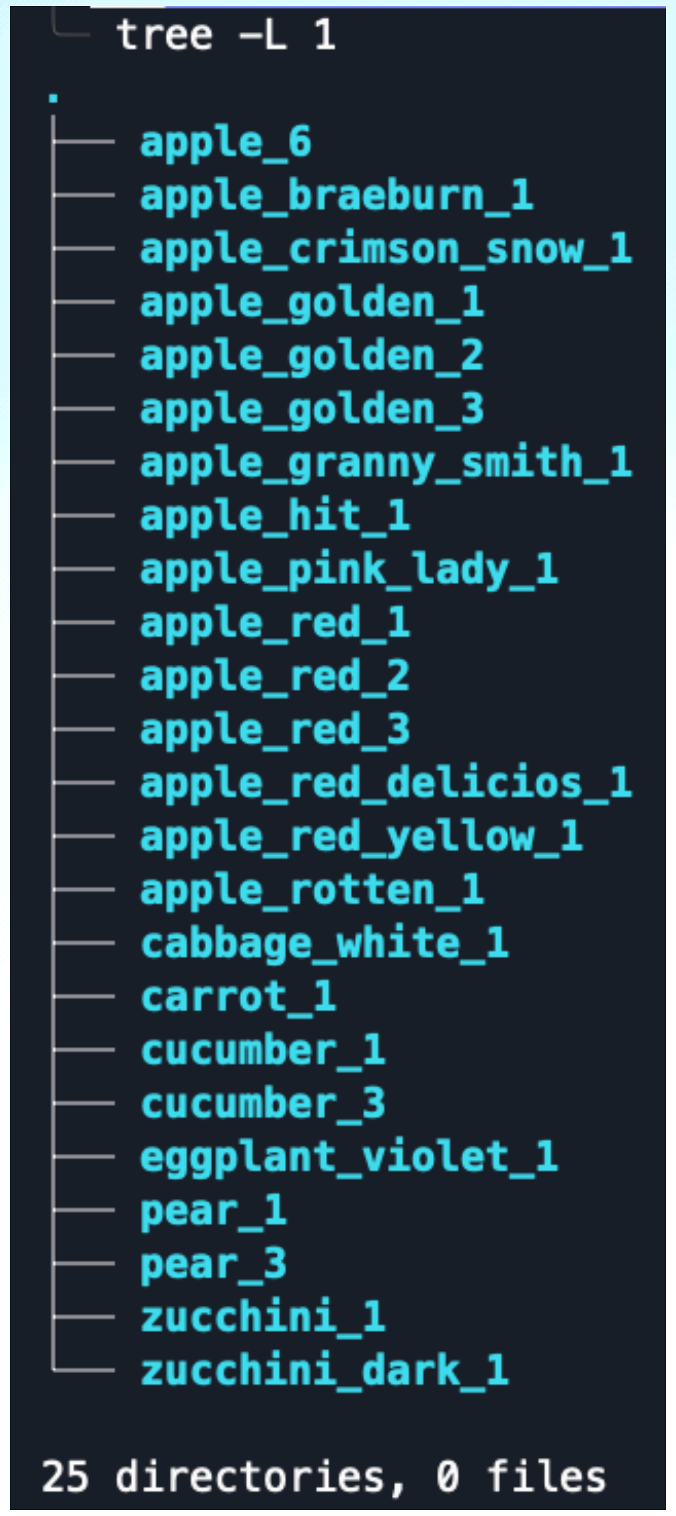
Données

Fichier Data



```
🗁 Results  🗁 Test  🗁 Test1  🗁 Training  🗁 Validation
```

Structure Test

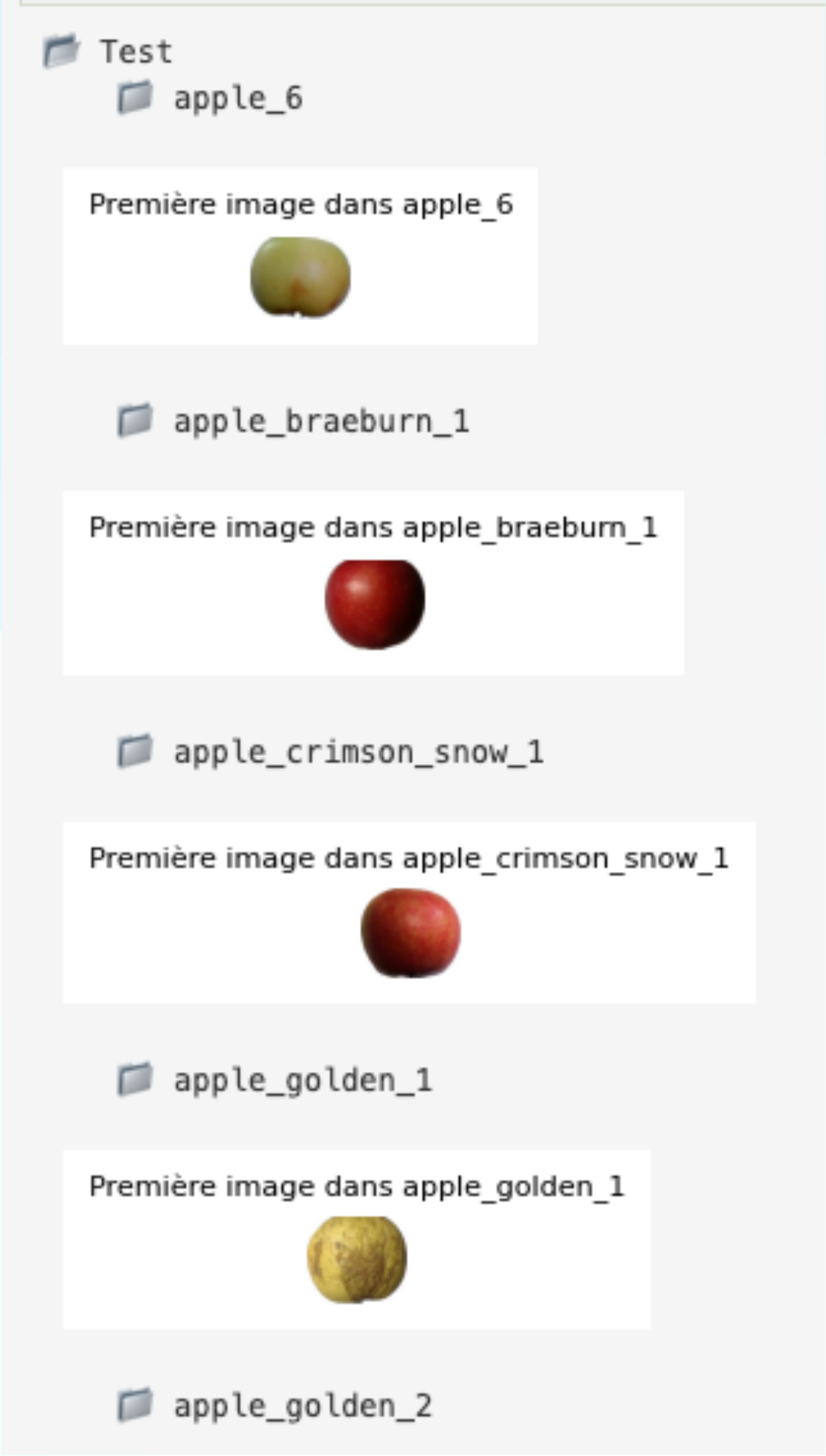


```
 tree -L 1
.
├── apple_6
├── apple_braeburn_1
├── apple_crimson_snow_1
├── apple_golden_1
├── apple_golden_2
├── apple_golden_3
├── apple_granny_smith_1
├── apple_hit_1
├── apple_pink_lady_1
├── apple_red_1
├── apple_red_2
├── apple_red_3
├── apple_red_delicios_1
├── apple_red_yellow_1
├── apple_rotten_1
├── cabbage_white_1
├── carrot_1
├── cucumber_1
├── cucumber_3
├── eggplant_violet_1
├── pear_1
├── pear_3
├── zucchini_1
└── zucchini_dark_1

25 directories, 0 files
```

Exemple Images

EDA

Affichage données



```
Test
    apple_6
```

Première image dans apple_6

```
    apple_braeburn_1
```

Première image dans apple_braeburn_1

```
    apple_crimson_snow_1
```

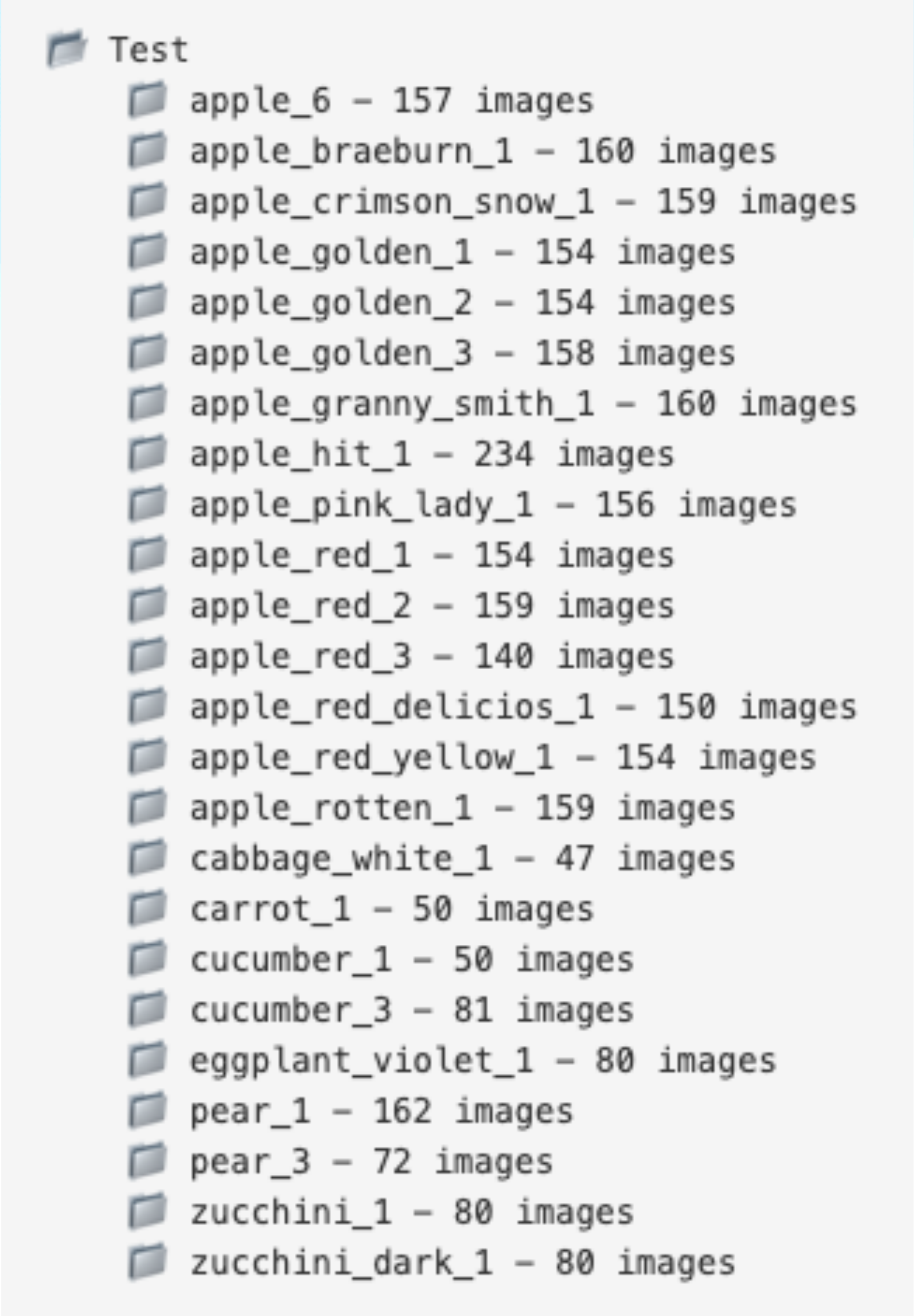
Première image dans apple_crimson_snow_1

```
    apple_golden_1
```

Première image dans apple_golden_1

```
    apple_golden_2
```

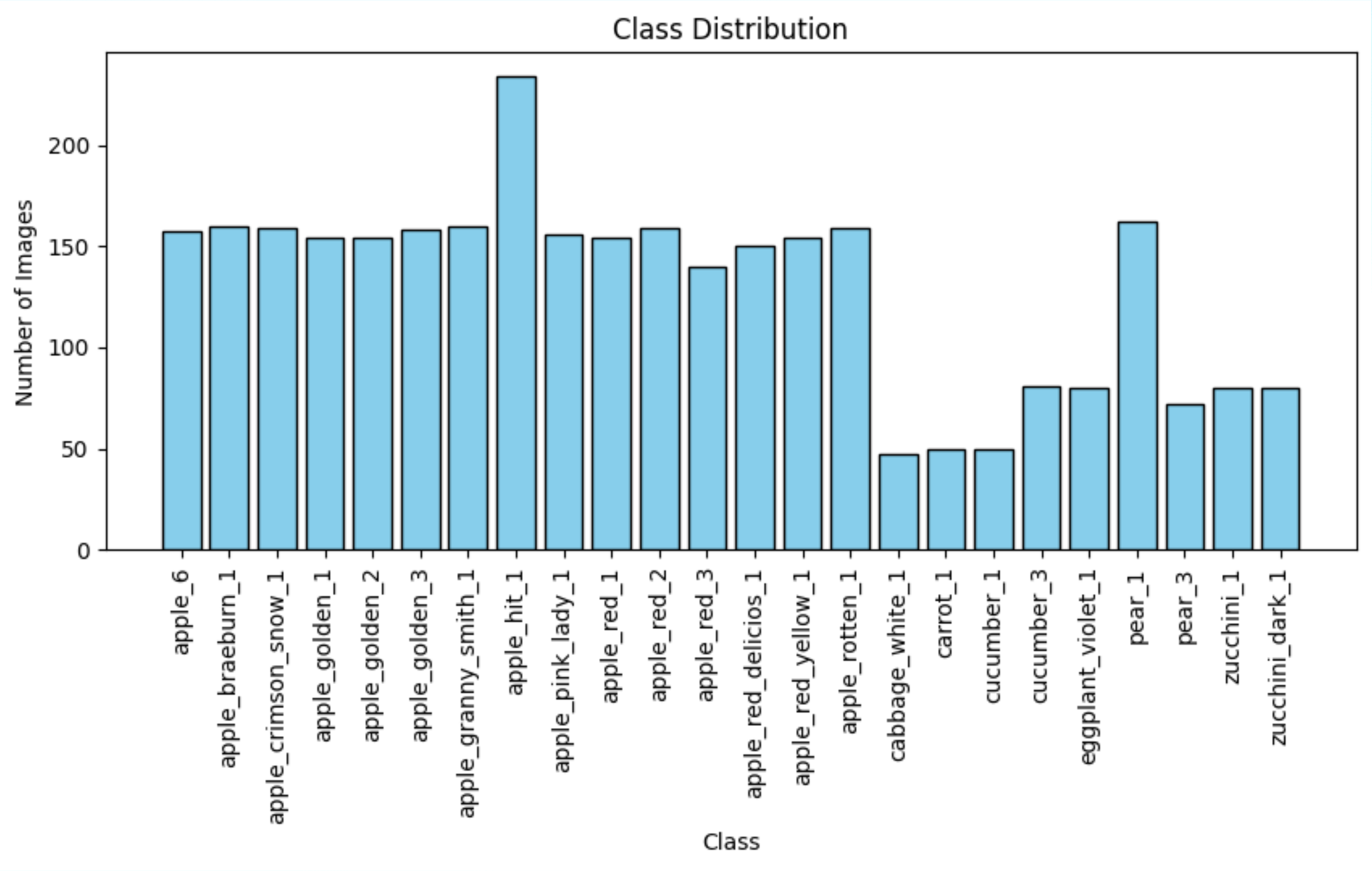
Comptage nombre images

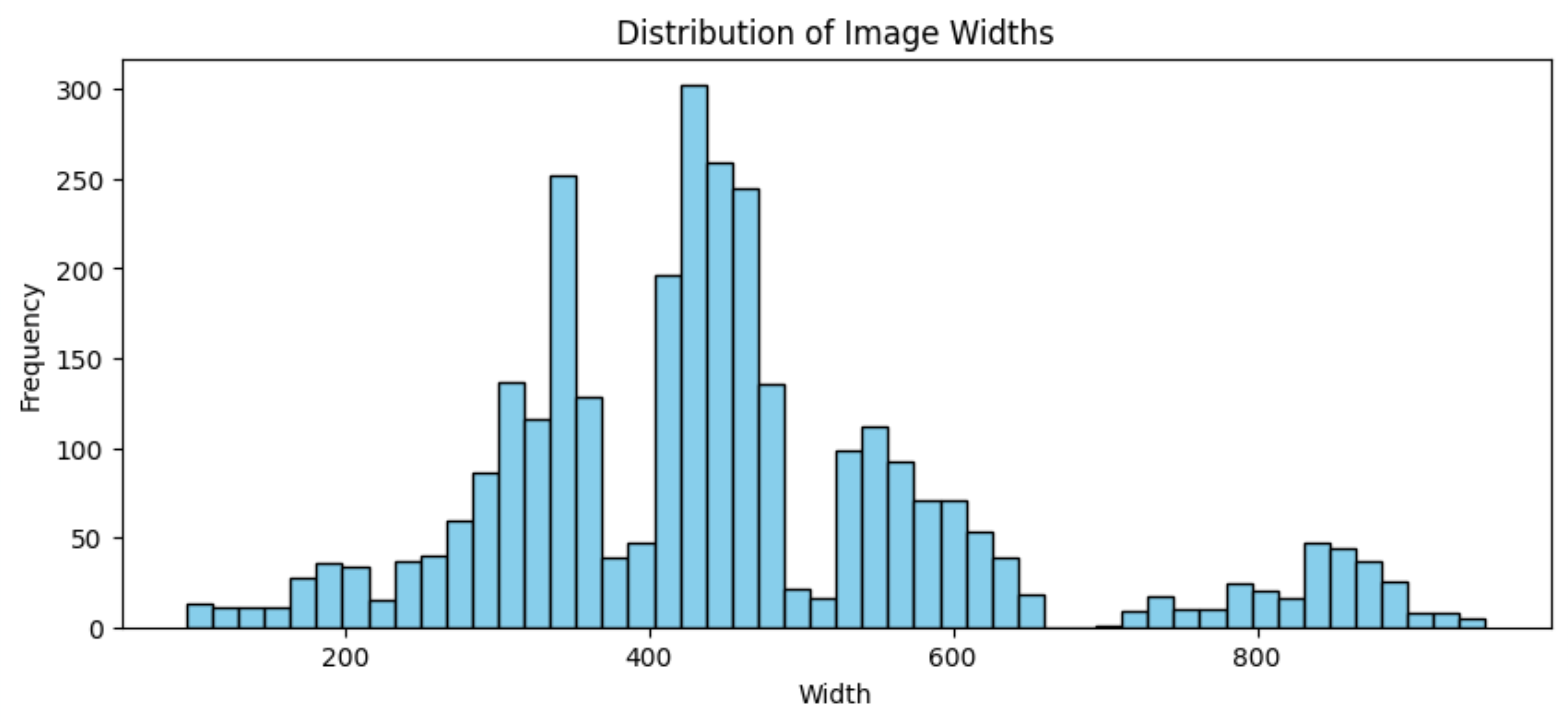


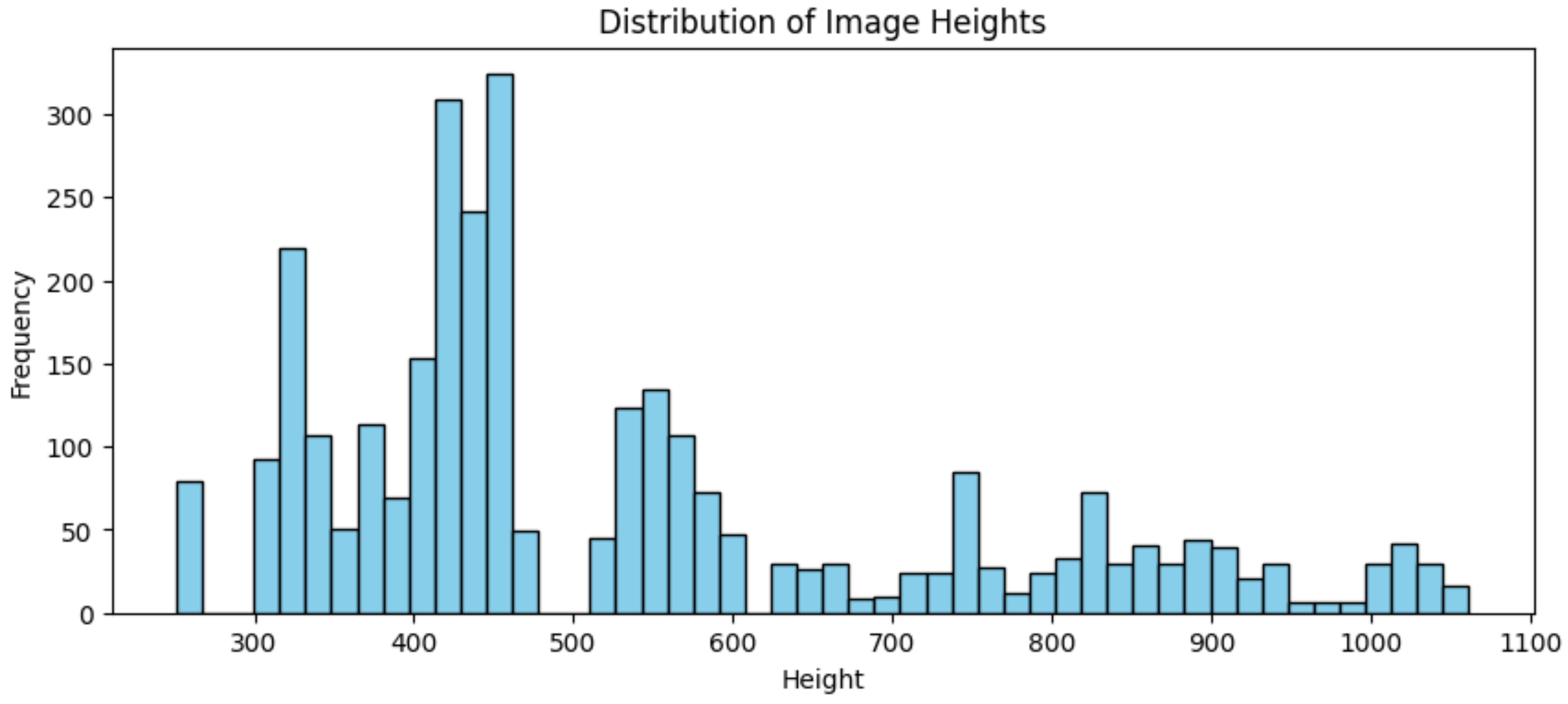
```
Test
    apple_6 - 157 images
    apple_braeburn_1 - 160 images
    apple_crimson_snow_1 - 159 images
    apple_golden_1 - 154 images
    apple_golden_2 - 154 images
    apple_golden_3 - 158 images
    apple_granny_smith_1 - 160 images
    apple_hit_1 - 234 images
    apple_pink_lady_1 - 156 images
    apple_red_1 - 154 images
    apple_red_2 - 159 images
    apple_red_3 - 140 images
    apple_red_delicios_1 - 150 images
    apple_red_yellow_1 - 154 images
    apple_rotten_1 - 159 images
    cabbage_white_1 - 47 images
    carrot_1 - 50 images
    cucumber_1 - 50 images
    cucumber_3 - 81 images
    eggplant_violet_1 - 80 images
    pear_1 - 162 images
    pear_3 - 72 images
    zucchini_1 - 80 images
    zucchini_dark_1 - 80 images
```

EDA

Graphiques

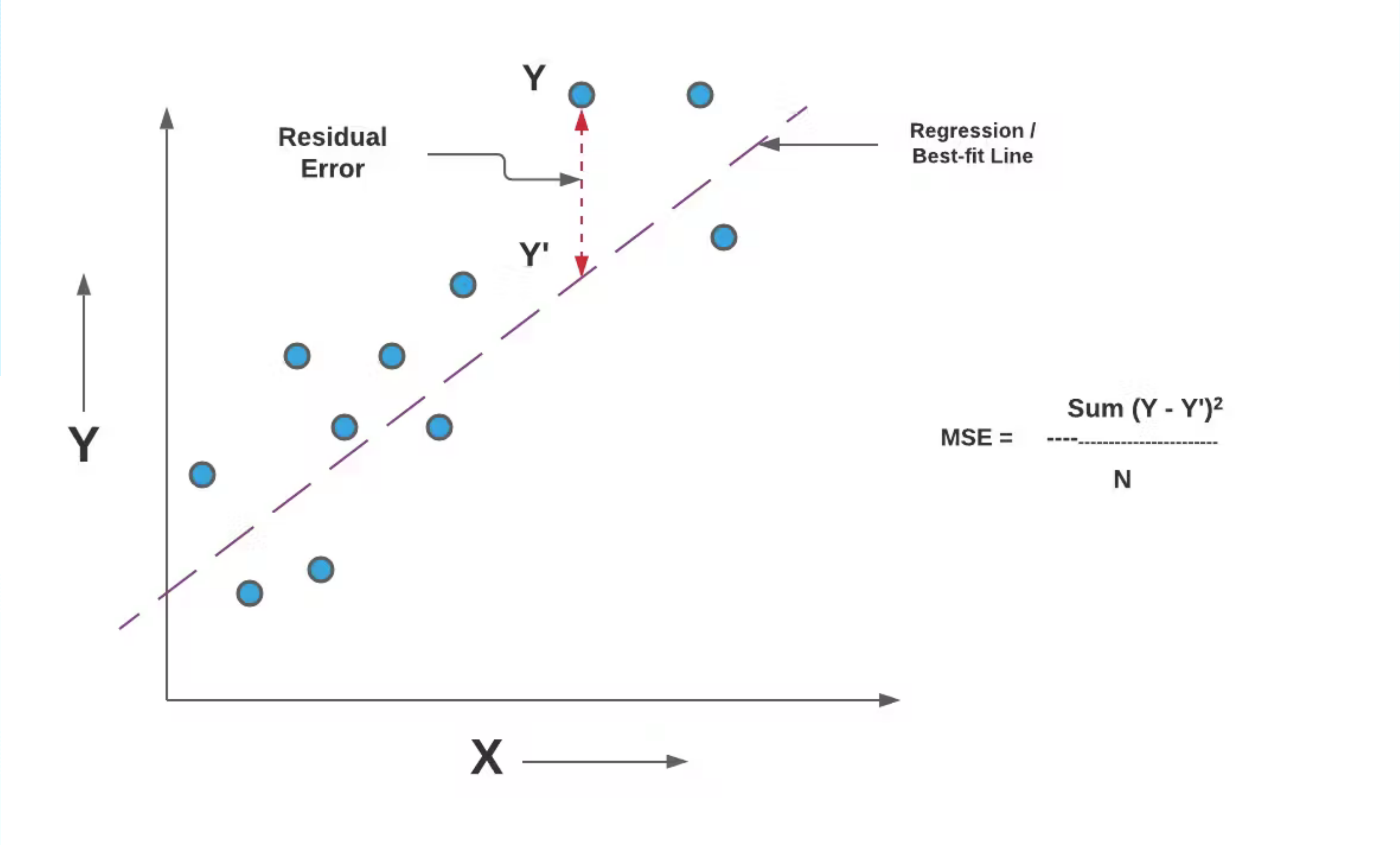


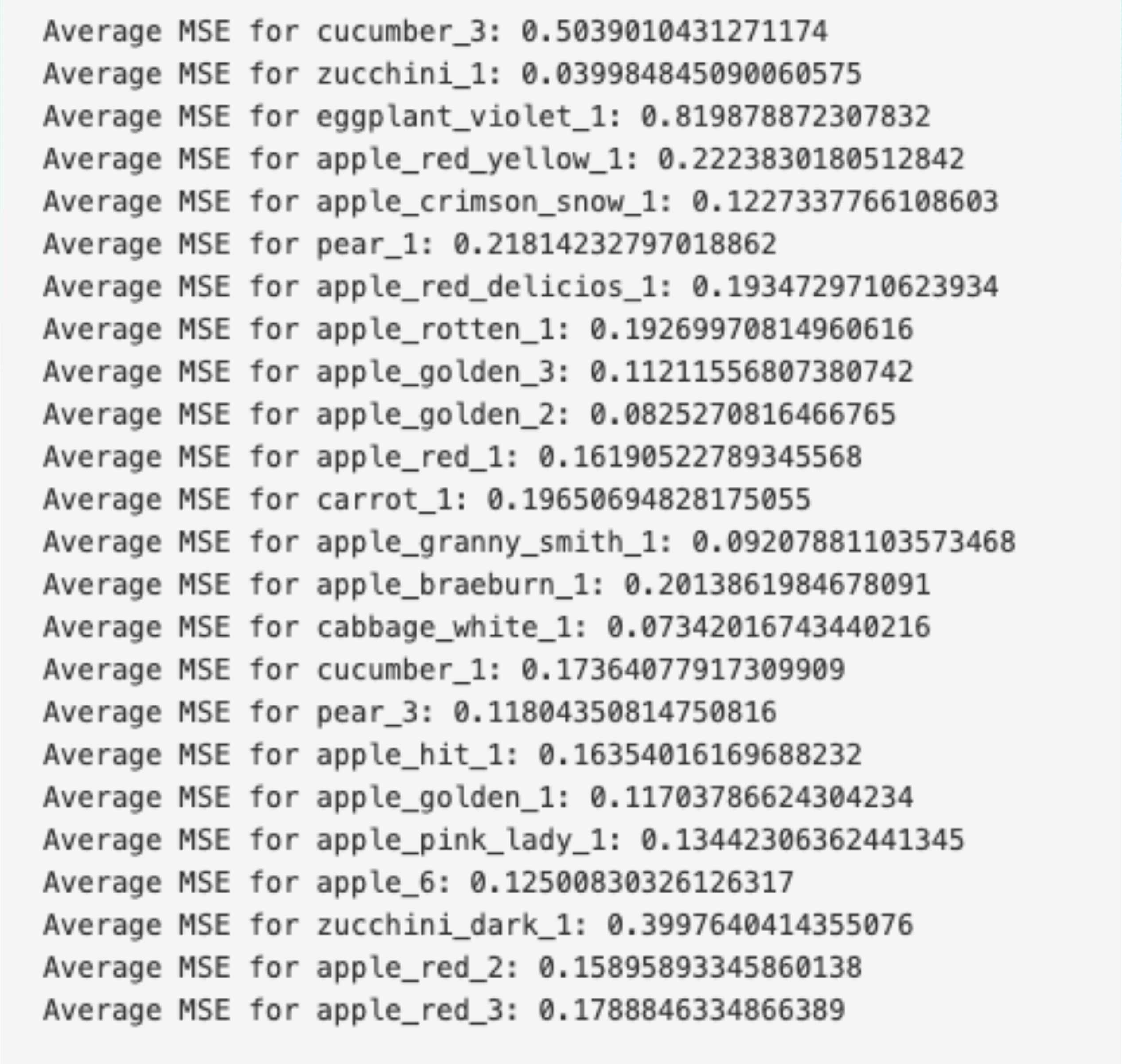




EDA

Calcul difference entre images





```
Average MSE for cucumber_3: 0.5039010431271174
Average MSE for zucchini_1: 0.039984845090060575
Average MSE for eggplant_violet_1: 0.819878872307832
Average MSE for apple_red_yellow_1: 0.2223830180512842
Average MSE for apple_crimson_snow_1: 0.1227337766108603
Average MSE for pear_1: 0.21814232797018862
Average MSE for apple_red_delicios_1: 0.1934729710623934
Average MSE for apple_rotten_1: 0.19269970814960616
Average MSE for apple_golden_3: 0.11211556807380742
Average MSE for apple_golden_2: 0.0825270816466765
Average MSE for apple_red_1: 0.16190522789345568
Average MSE for carrot_1: 0.19650694828175055
Average MSE for apple_granny_smith_1: 0.09207881103573468
Average MSE for apple_braeburn_1: 0.2013861984678091
Average MSE for cabbage_white_1: 0.07342016743440216
Average MSE for cucumber_1: 0.17364077917309909
Average MSE for pear_3: 0.11804350814750816
Average MSE for apple_hit_1: 0.16354016169688232
Average MSE for apple_golden_1: 0.11703786624304234
Average MSE for apple_pink_lady_1: 0.13442306362441345
Average MSE for apple_6: 0.12500830326126317
Average MSE for zucchini_dark_1: 0.3997640414355076
Average MSE for apple_red_2: 0.15895893345860138
Average MSE for apple_red_3: 0.1788846334866389
```

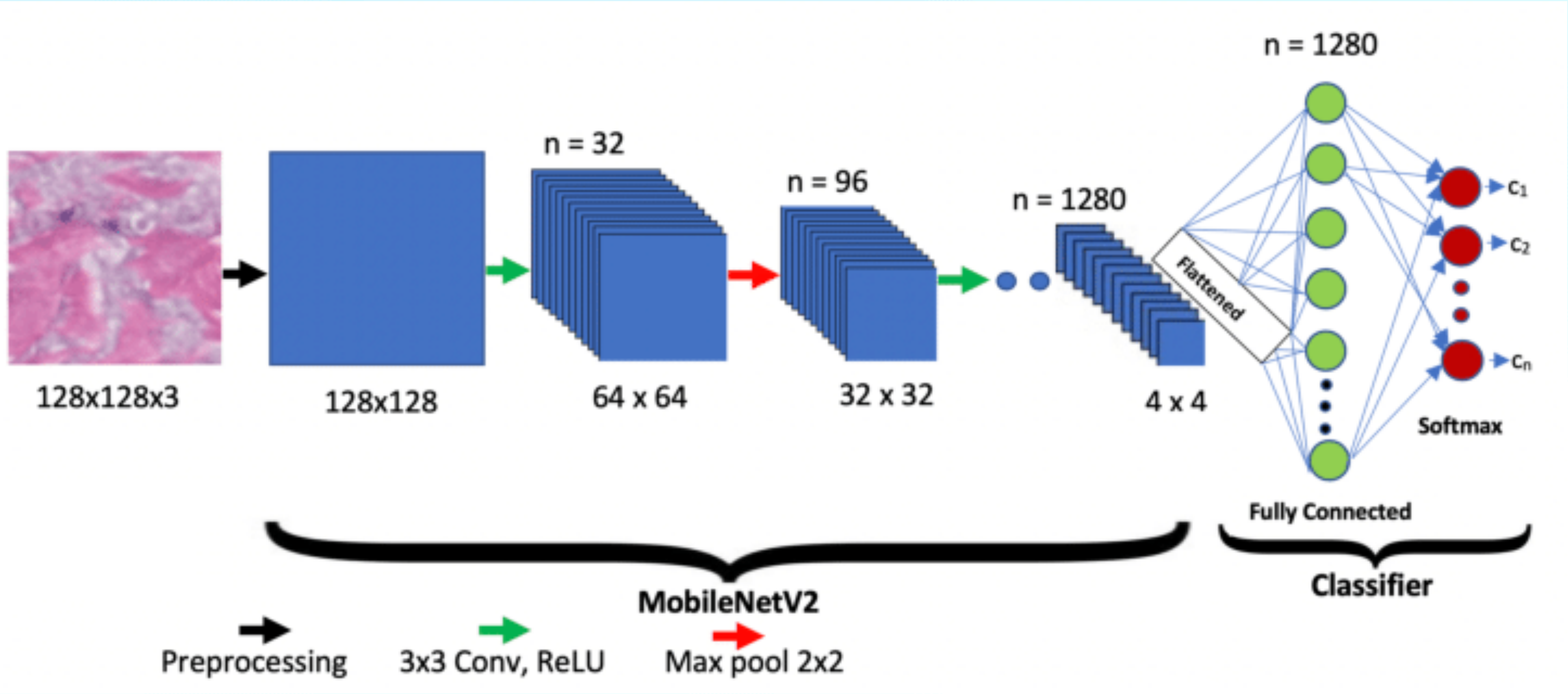
EDA

Fruits choisi pour l’entrainement

- **apple_granny_smith_1** (nombre élevé et relativement constant d'images : 160)
- **pear_1** (nombre similaire d’images: 162)
- **cucumber_3** (81 images, pour équilibrer l’étude)
- **zucchini_1** (80 images, pour équilibrer l’étude).

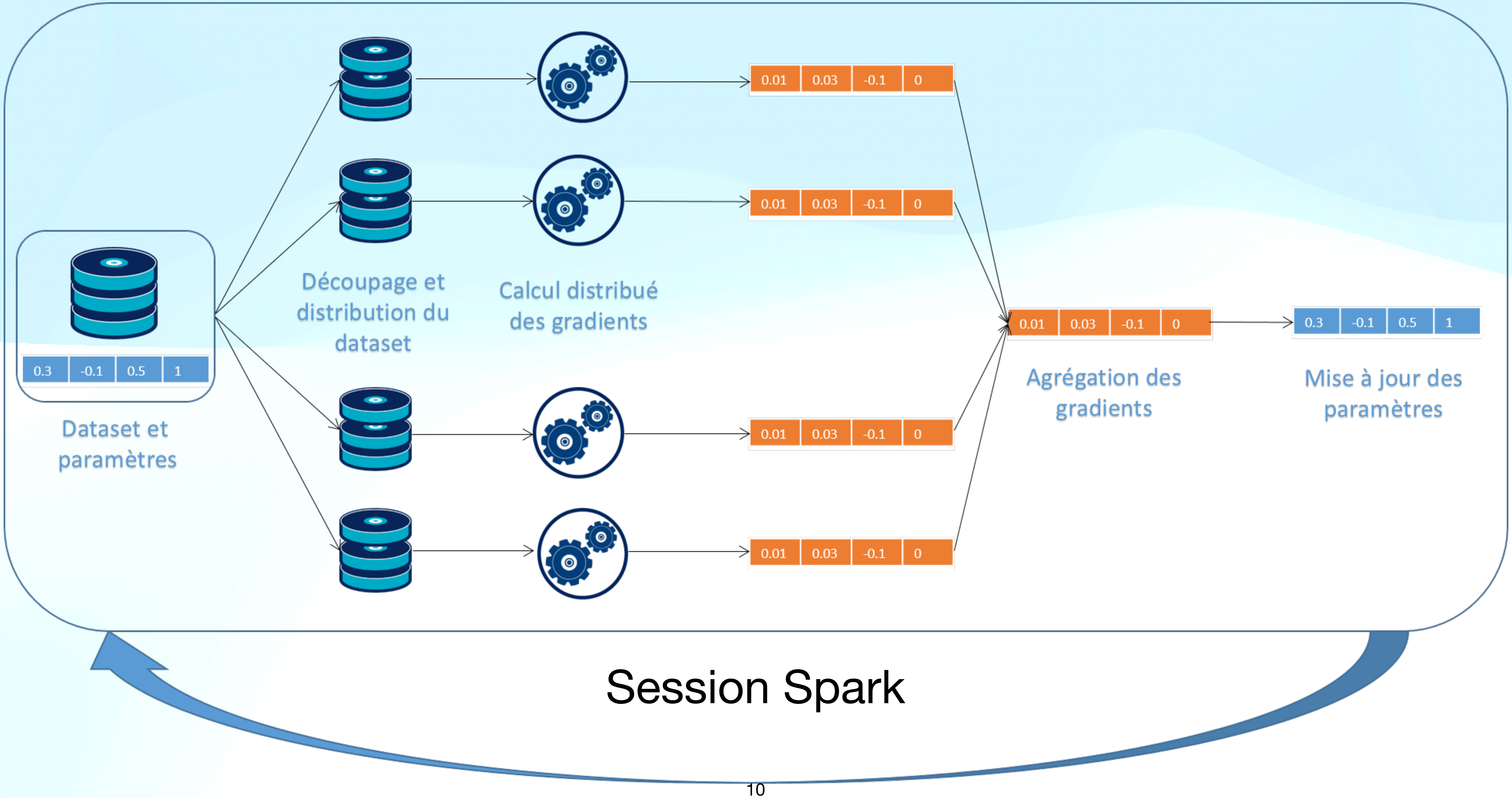
Modélisation

Architecture utilisé



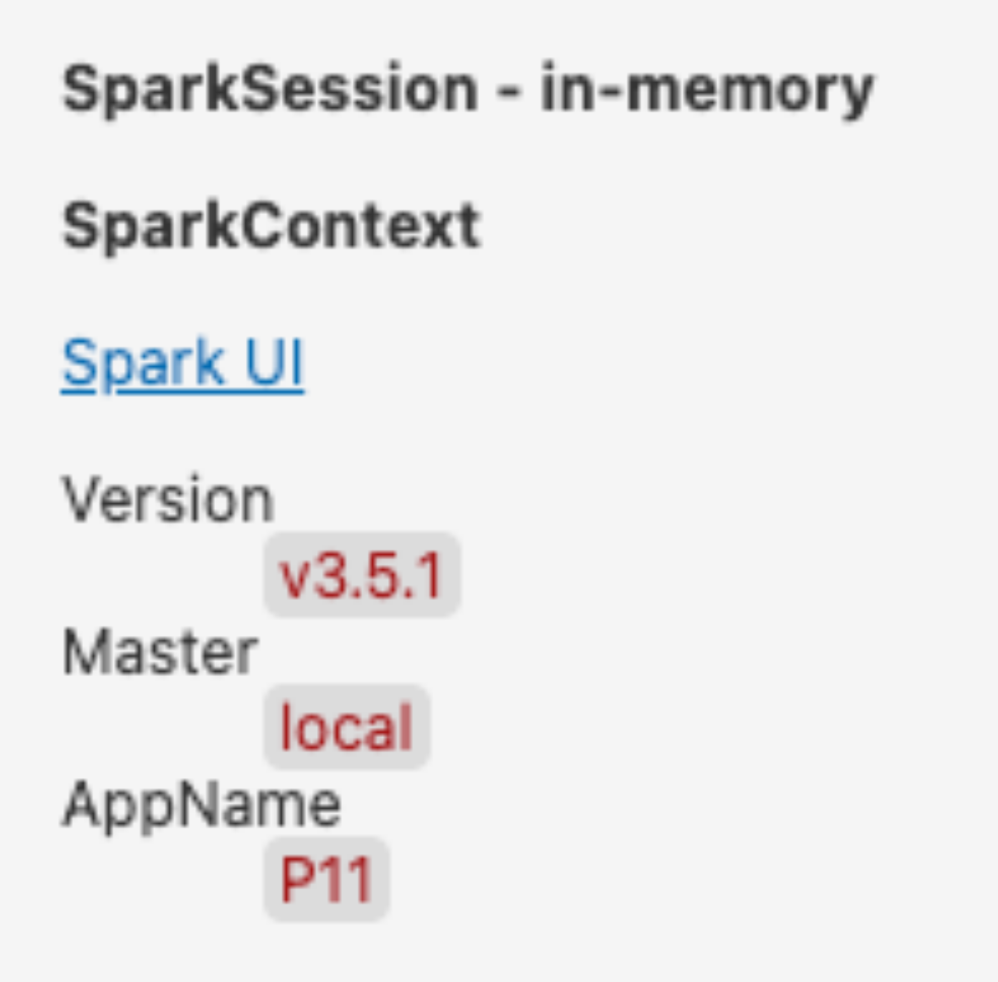
Traitement de données distribuées

Modélisation


0.3
-0.1
0.5
1
Dataset et paramètres
Découpage et distribution du dataset
Calcul distribué des gradients
0.01
0.03
-0.1
0
Agrégation des gradients
Mise à jour des paramètres
Session Spark

Modélisation

Session Spark



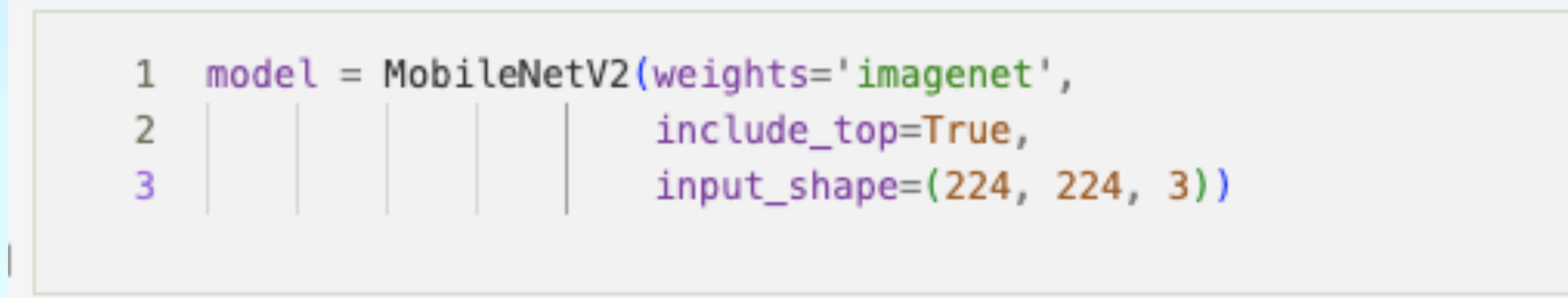
SparkSession - in-memory

SparkContext

Spark UI

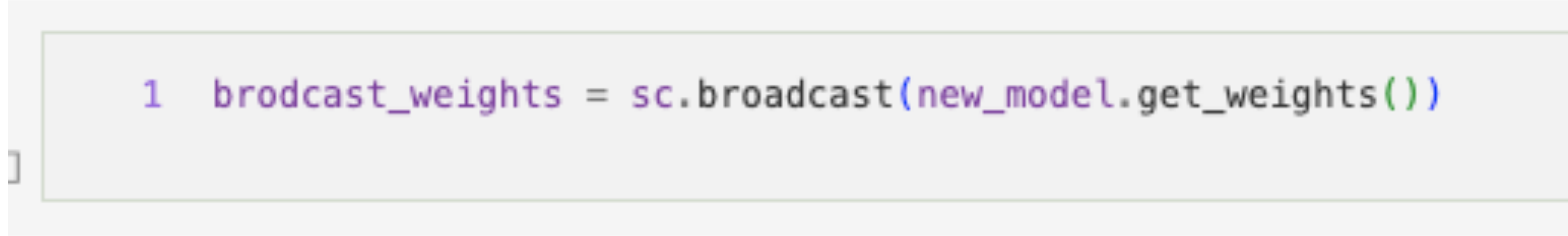
Version
v3.5.1
Master
local
AppName
P11

Initialisation modèle



```python
model = MobileNetV2(weights='imagenet',
                    include_top=True,
                    input_shape=(224, 224, 3))
```

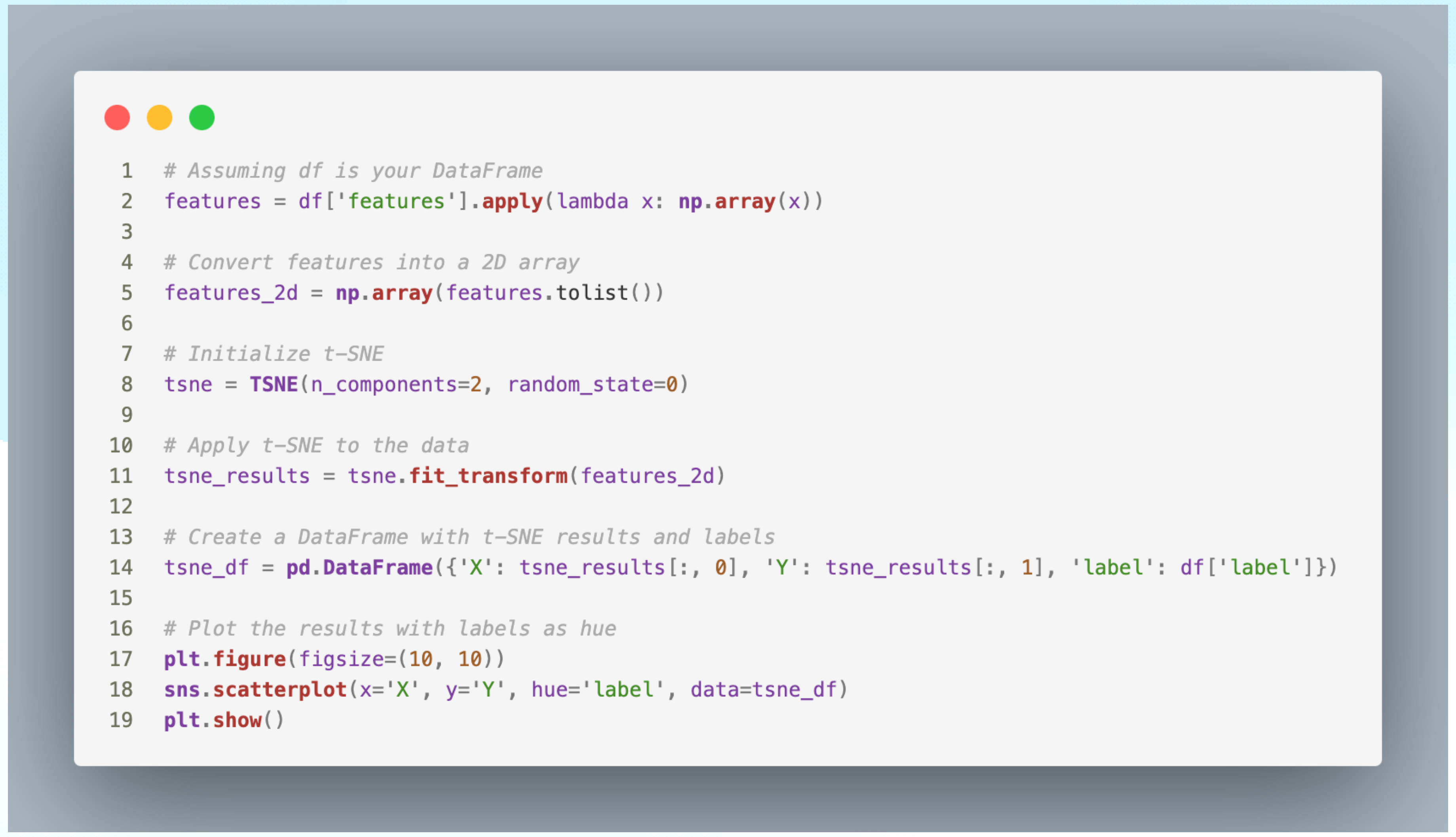
Diffusion poids du modèle



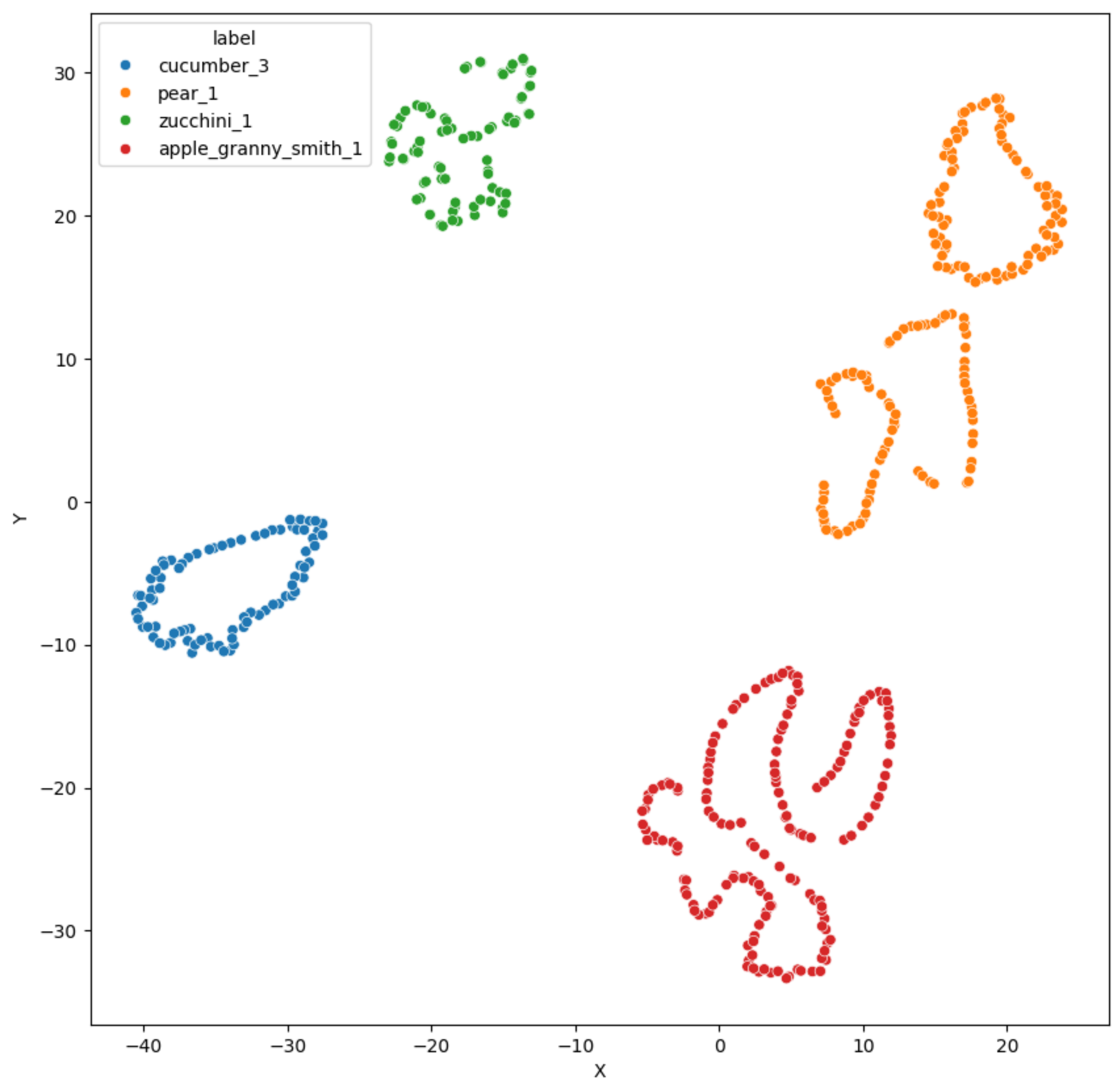
```python
brodcast_weights = sc.broadcast(new_model.get_weights())
```

Modélisation

T-Sne

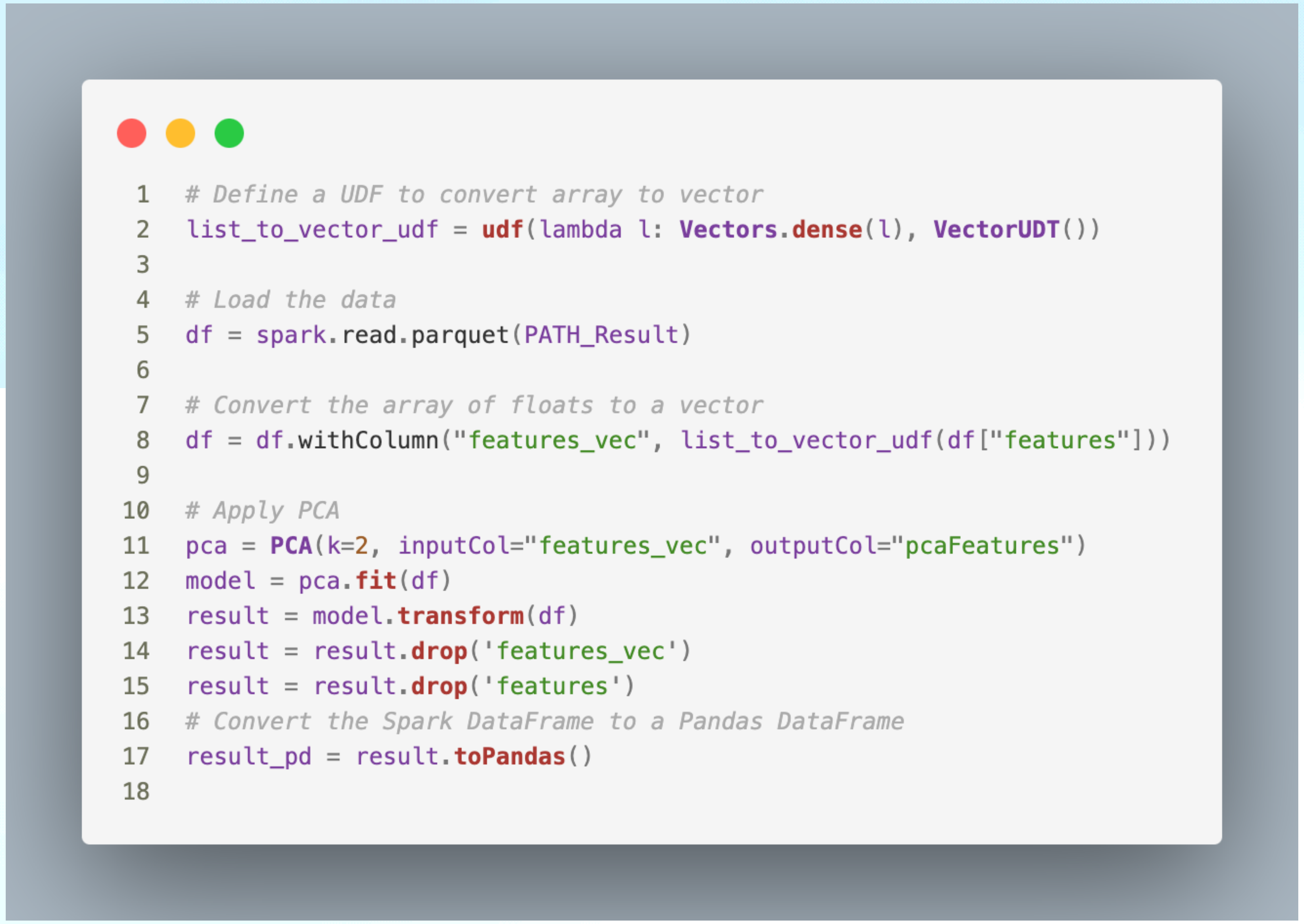


```python
# Assuming df is your DataFrame
features = df['features'].apply(lambda x: np.array(x))

# Convert features into a 2D array
features_2d = np.array(features.tolist())

# Initialize t-SNE
tsne = TSNE(n_components=2, random_state=0)

# Apply t-SNE to the data
tsne_results = tsne.fit_transform(features_2d)

# Create a DataFrame with t-SNE results and labels
tsne_df = pd.DataFrame({'X': tsne_results[:, 0], 'Y': tsne_results[:, 1], 'label': df['label']})

# Plot the results with labels as hue
plt.figure(figsize=(10, 10))
sns.scatterplot(x='X', y='Y', hue='label', data=tsne_df)
plt.show()
```



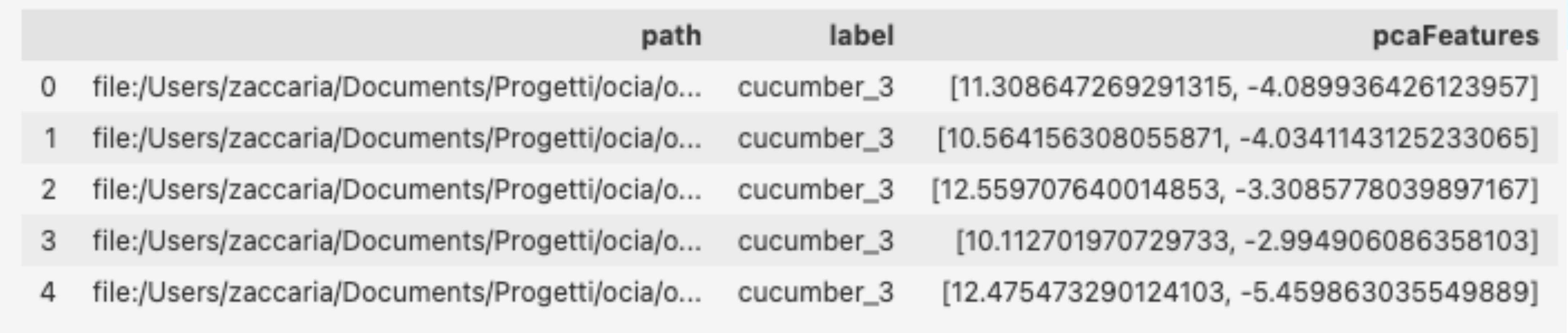
Modélisation

PCA



```python
# Define a UDF to convert array to vector
list_to_vector_udf = udf(lambda l: Vectors.dense(l), VectorUDT())

# Load the data
df = spark.read.parquet(PATH_Result)

# Convert the array of floats to a vector
df = df.withColumn("features_vec", list_to_vector_udf(df["features"]))

# Apply PCA
pca = PCA(k=2, inputCol="features_vec", outputCol="pcaFeatures")
model = pca.fit(df)
result = model.transform(df)
result = result.drop('features_vec')
result = result.drop('features')
# Convert the Spark DataFrame to a Pandas DataFrame
result_pd = result.toPandas()

```

Résultats

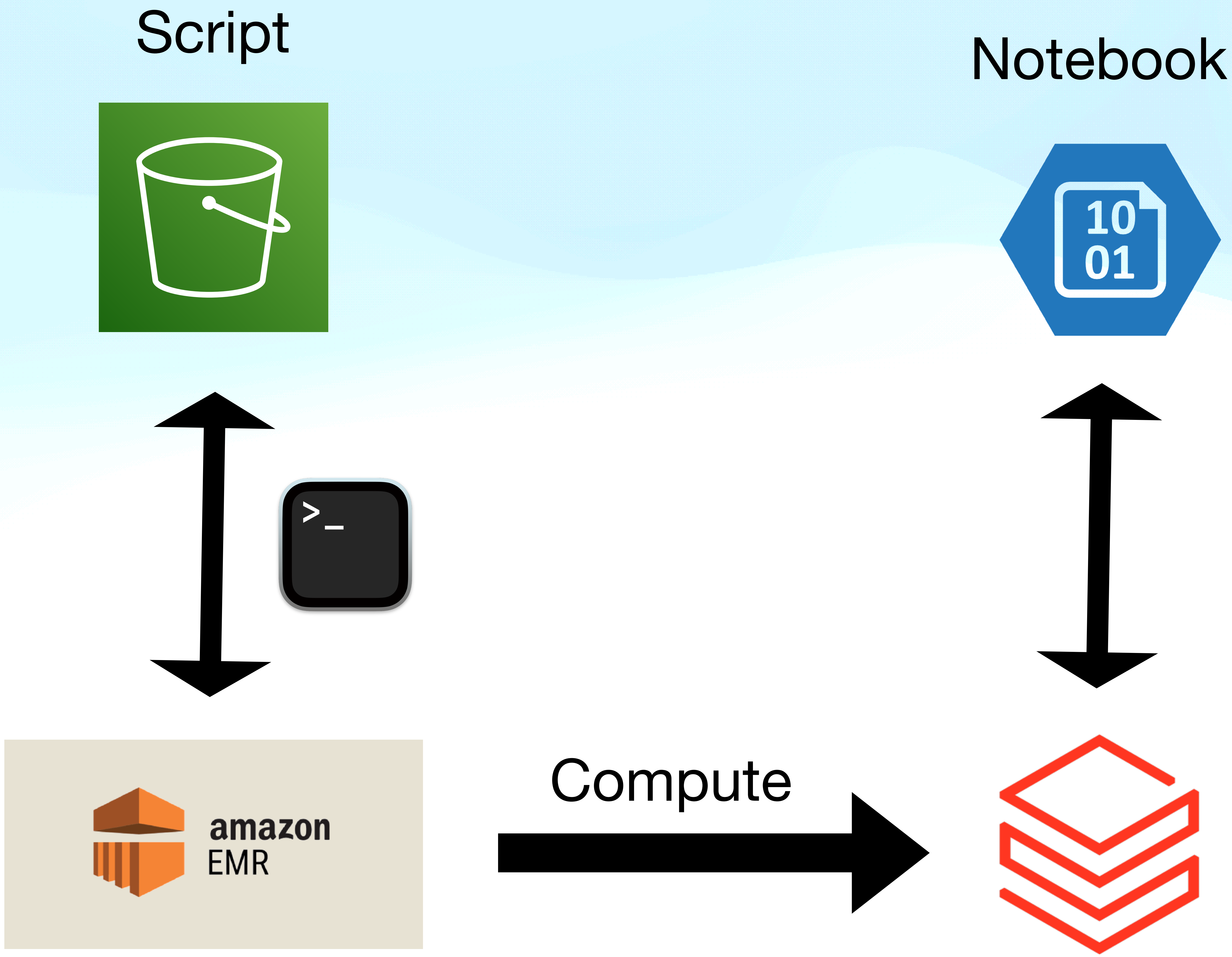


	path	label	pcaFeatures
0	file:/Users/zaccaria/Documents/Progetti/ocia/o...	cucumber_3	[11.308647269291315, -4.089936426123957]
1	file:/Users/zaccaria/Documents/Progetti/ocia/o...	cucumber_3	[10.564156308055871, -4.0341143125233065]
2	file:/Users/zaccaria/Documents/Progetti/ocia/o...	cucumber_3	[12.559707640014853, -3.3085778039897167]
3	file:/Users/zaccaria/Documents/Progetti/ocia/o...	cucumber_3	[10.112701970729733, -2.994906086358103]
4	file:/Users/zaccaria/Documents/Progetti/ocia/o...	cucumber_3	[12.475473290124103, -5.459863035549889]

Cloud

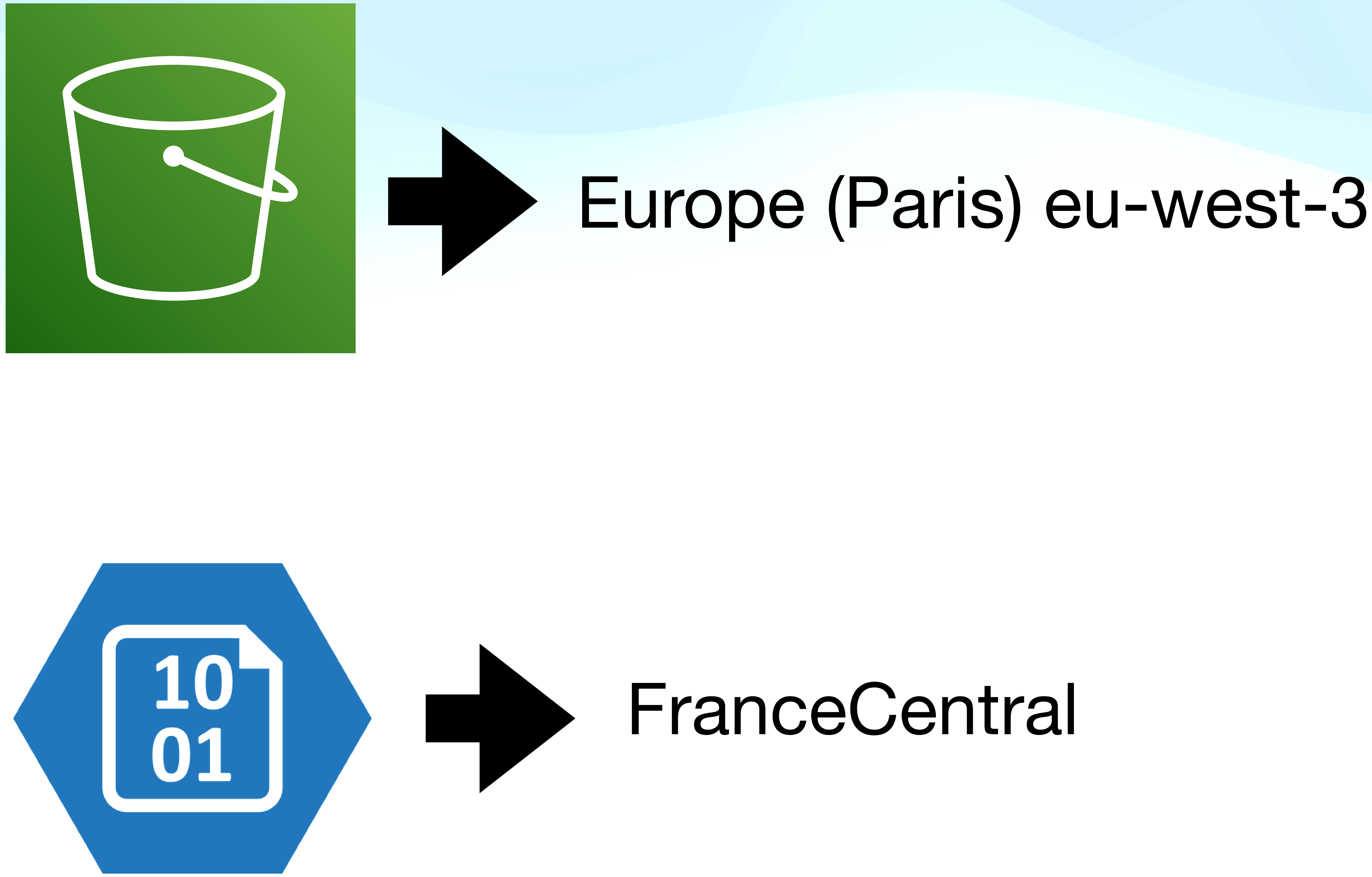
3 Approches

- Local
- Script
- Notebook



RGPD

Article 3 – Territorial Scope

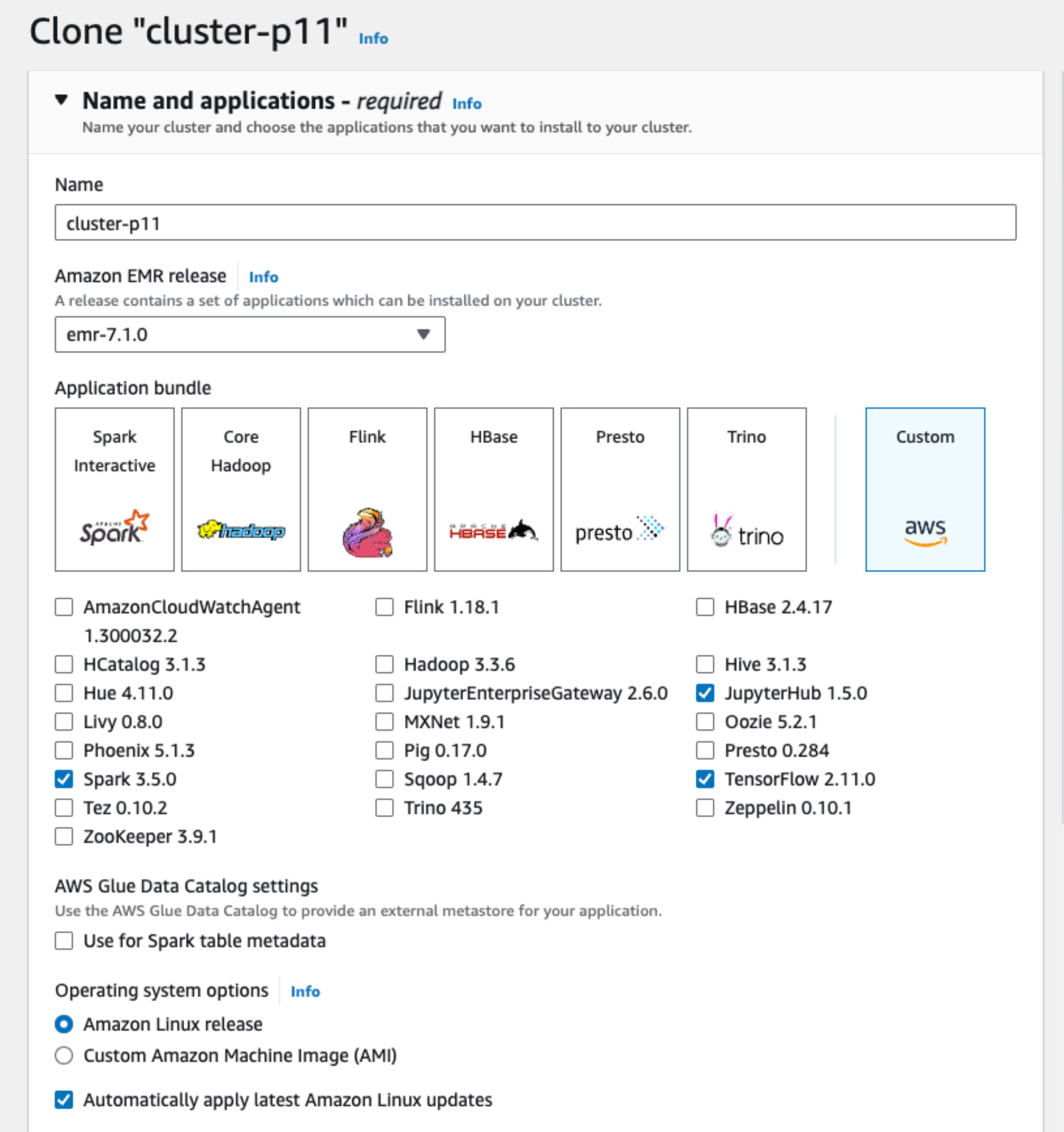


Europe (Paris) eu-west-3

FranceCentral

Cloud

Cluster EMR



Clone "cluster-p11" Info

▼ **Name and applications** - *required* Info

Name your cluster and choose the applications that you want to install to your cluster.

Name

cluster-p11

Amazon EMR release Info

A release contains a set of applications which can be installed on your cluster.

emr-7.1.0 ▼

Application bundle

- ☐ AmazonCloudWatchAgent 1.300032.2
- ☐ HCatalog 3.1.3
- ☐ Hue 4.11.0
- ☐ Livy 0.8.0
- ☐ Phoenix 5.1.3
- ☑ Spark 3.5.0
- ☐ Tez 0.10.2
- ☐ ZooKeeper 3.9.1
- ☐ Flink 1.18.1
- ☐ Hadoop 3.3.6
- ☐ JupyterEnterpriseGateway 2.6.0
- ☐ MXNet 1.9.1
- ☐ Pig 0.17.0
- ☐ Sqoop 1.4.7
- ☐ Trino 435
- ☐ HBase 2.4.17
- ☐ Hive 3.1.3
- ☑ JupyterHub 1.5.0
- ☐ Oozie 5.2.1
- ☐ Presto 0.284
- ☑ TensorFlow 2.11.0
- ☐ Zeppelin 0.10.1

AWS Glue Data Catalog settings

Use the AWS Glue Data Catalog to provide an external metastore for your application.

☐ Use for Spark table metadata

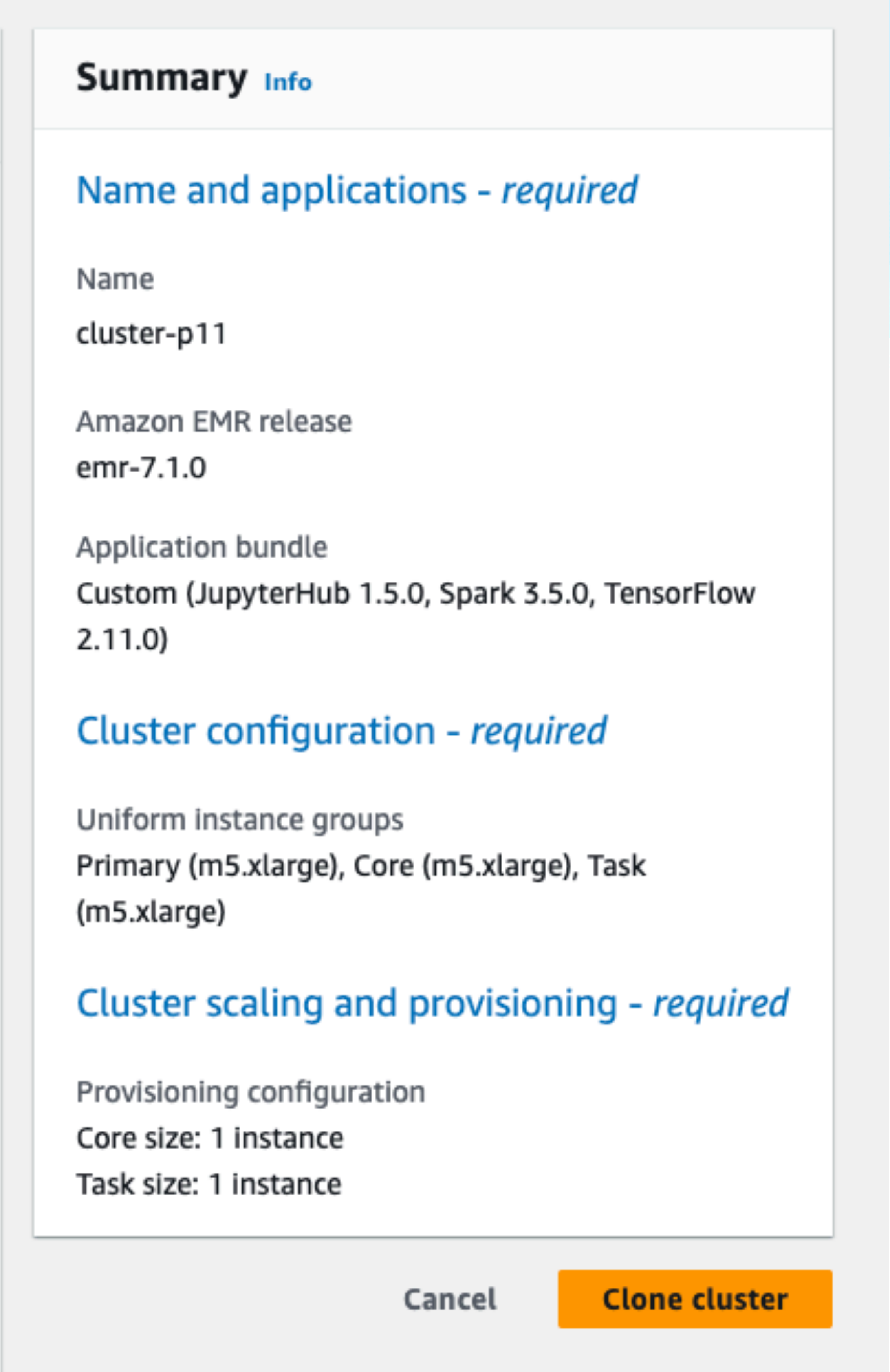
Operating system options Info

● Amazon Linux release

○ Custom Amazon Machine Image (AMI)

☑ Automatically apply latest Amazon Linux updates

Sommaire



Summary Info

Name and applications - *required*

Name

cluster-p11

Amazon EMR release

emr-7.1.0

Application bundle

Custom (JupyterHub 1.5.0, Spark 3.5.0, TensorFlow 2.11.0)

Cluster configuration - *required*

Uniform instance groups

Primary (m5.xlarge), Core (m5.xlarge), Task (m5.xlarge)

Cluster scaling and provisioning - *required*

Provisioning configuration

Core size: 1 instance

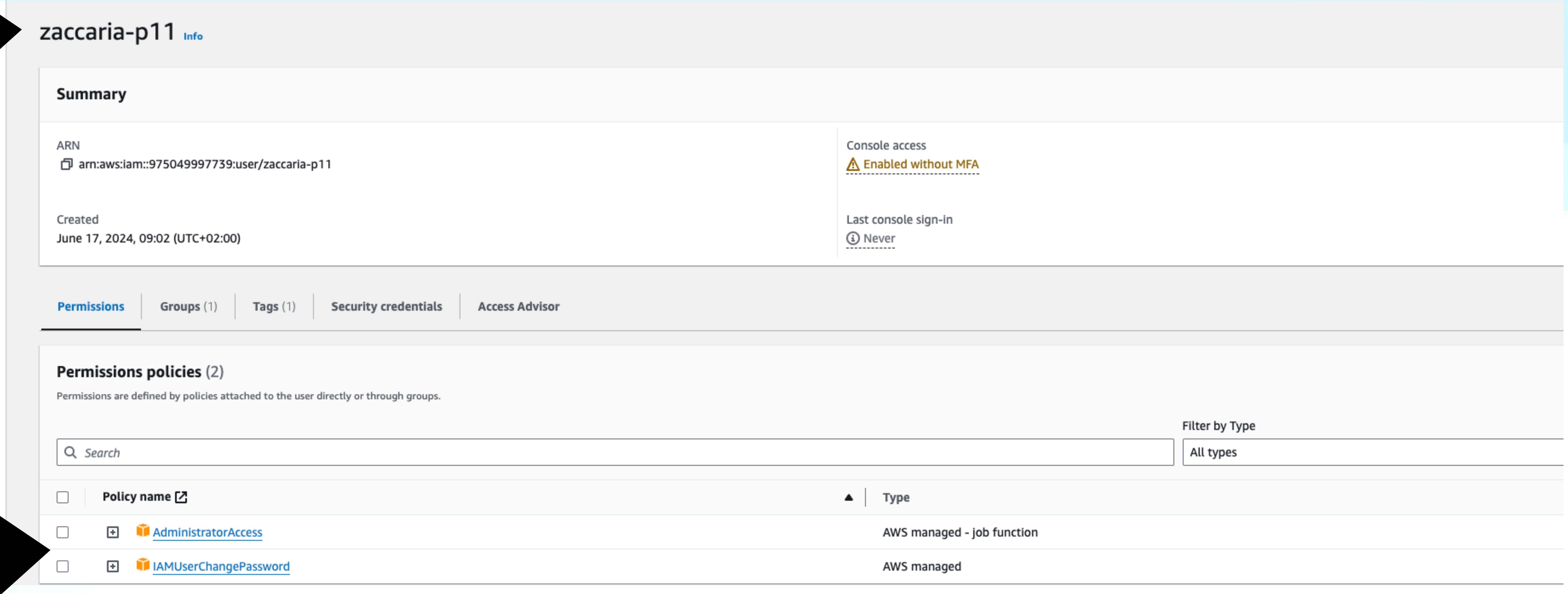
Task size: 1 instance

Cloud

IAM

Utilisateur

Policies



zaccaria-p11 Info

Summary

ARN
arn:aws:iam::975049997739:user/zaccaria-p11

Console access
⚠ Enabled without MFA

Created
June 17, 2024, 09:02 (UTC+02:00)

Last console sign-in
ⓘ Never

Permissions policies (2)

Permissions are defined by policies attached to the user directly or through groups.

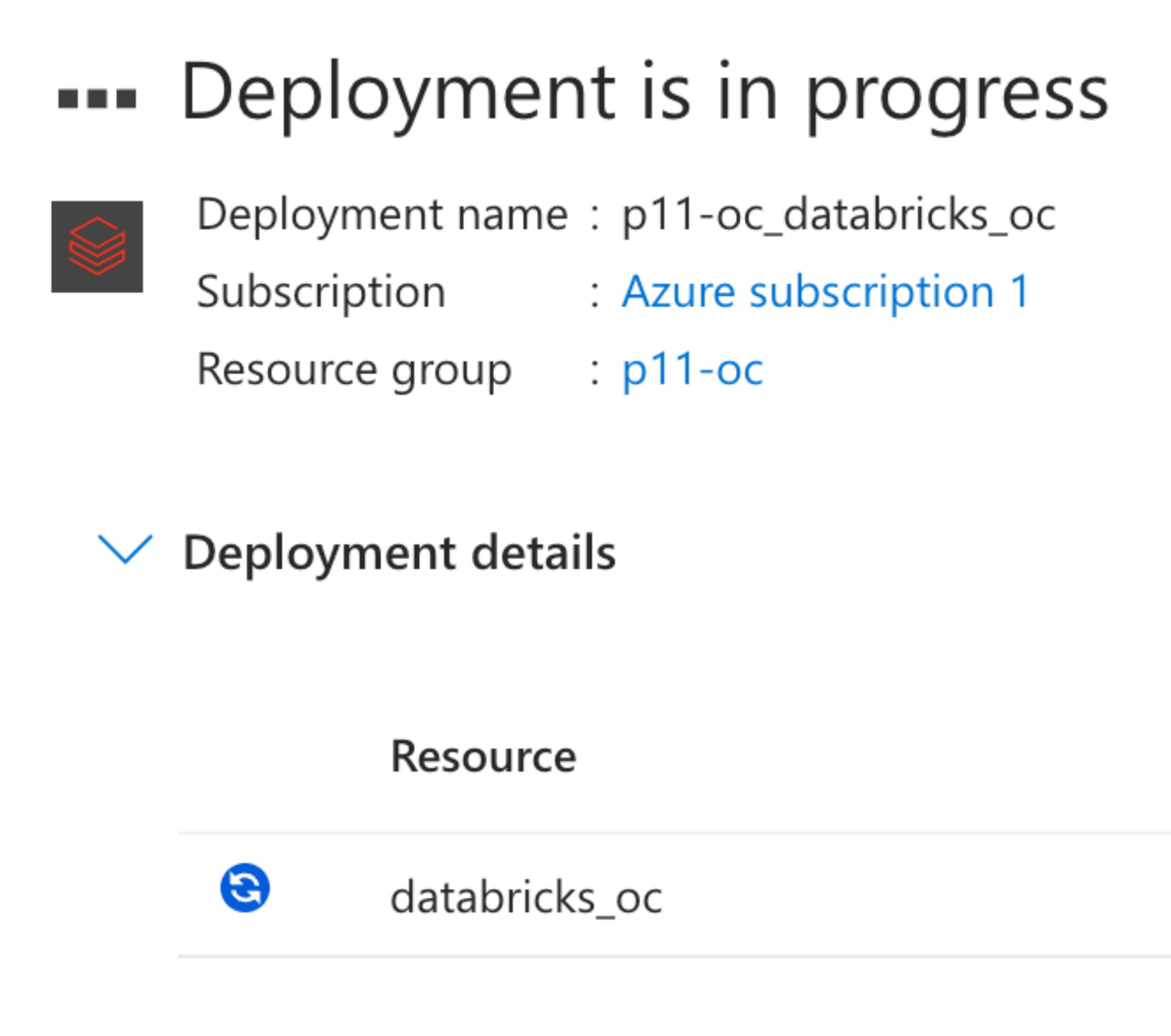
All types

□	Policy name	Type
□	AdministratorAccess	AWS managed - job function
□	IAMUserChangePassword	AWS managed

Cloud

Databricks

Deployment



■■■ Deployment is in progress

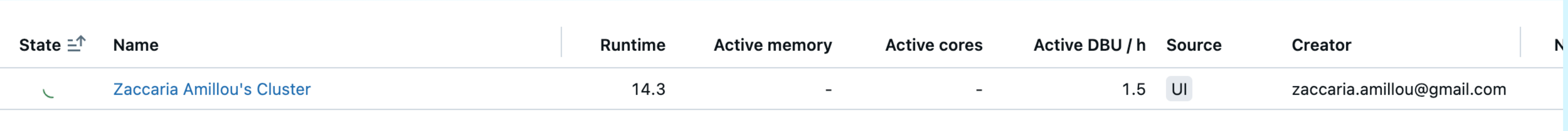
Deployment name : p11-oc_databricks_oc
Subscription : Azure subscription 1
Resource group : p11-oc

Deployment details

Resource

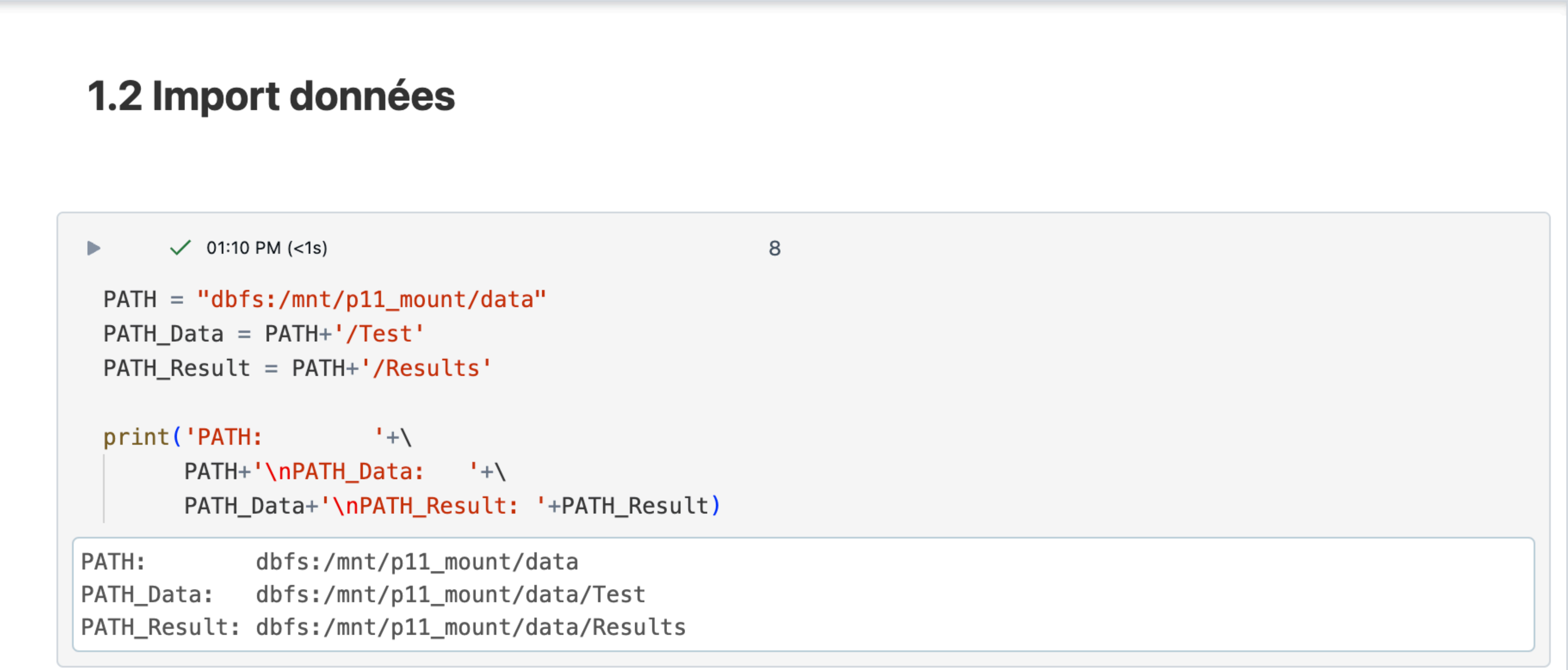
databricks_oc

Initialisation cluster



State	Name	Runtime	Active memory	Active cores	Active DBU / h	Source	Creator	N
	Zaccaria Amillou's Cluster	14.3	-	-	1.5	UI	zaccaria.amillou@gmail.com	

Import des données



1.2 Import données

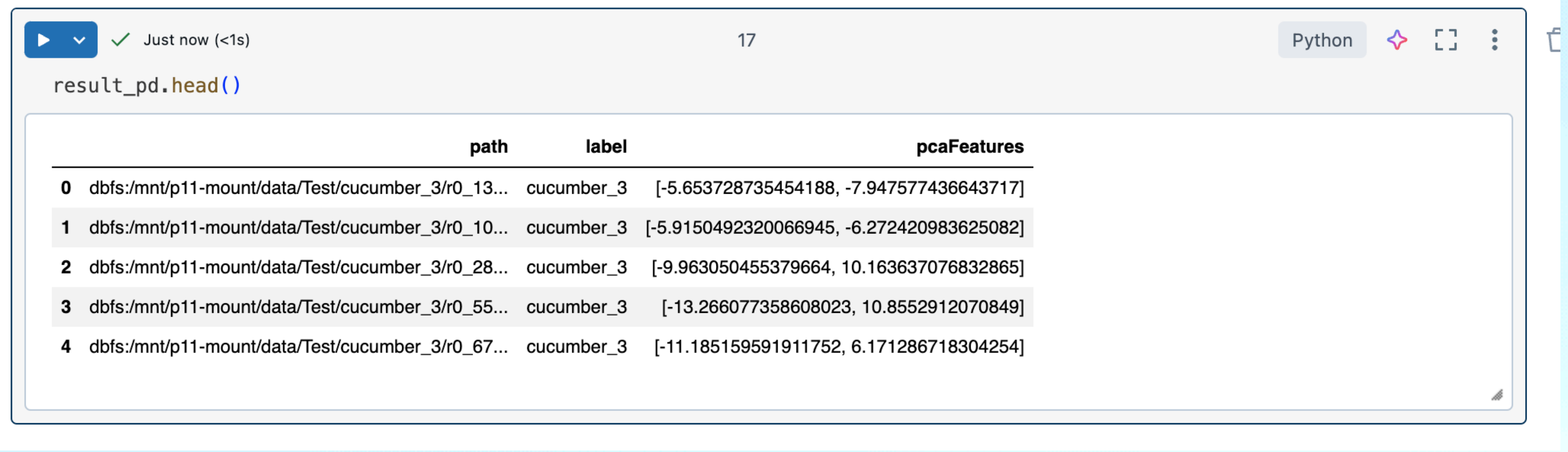
✓ 01:10 PM (<1s) 8

```
PATH = "dbfs:/mnt/p11_mount/data"
PATH_Data = PATH+'/Test'
PATH_Result = PATH+'/Results'

print('PATH:        '+\
      PATH+'\nPATH_Data:   '+\
      PATH_Data+'\nPATH_Result: '+PATH_Result)
```

```
PATH:        dbfs:/mnt/p11_mount/data
PATH_Data:   dbfs:/mnt/p11_mount/data/Test
PATH_Result: dbfs:/mnt/p11_mount/data/Results
```

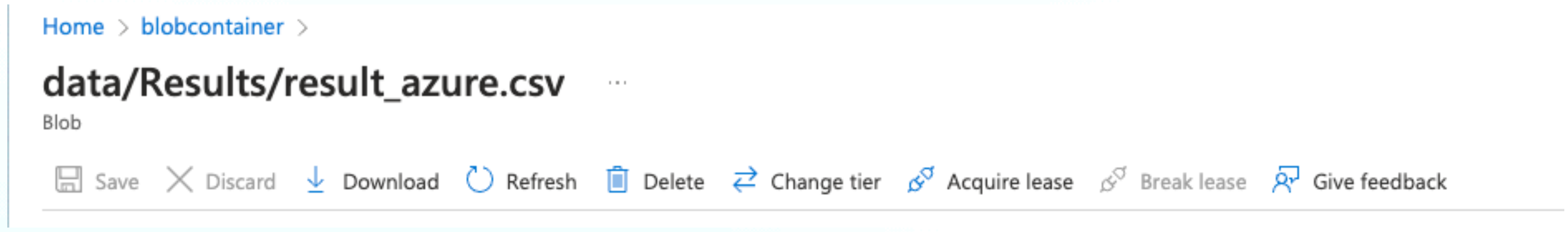
Cloud

Résultats



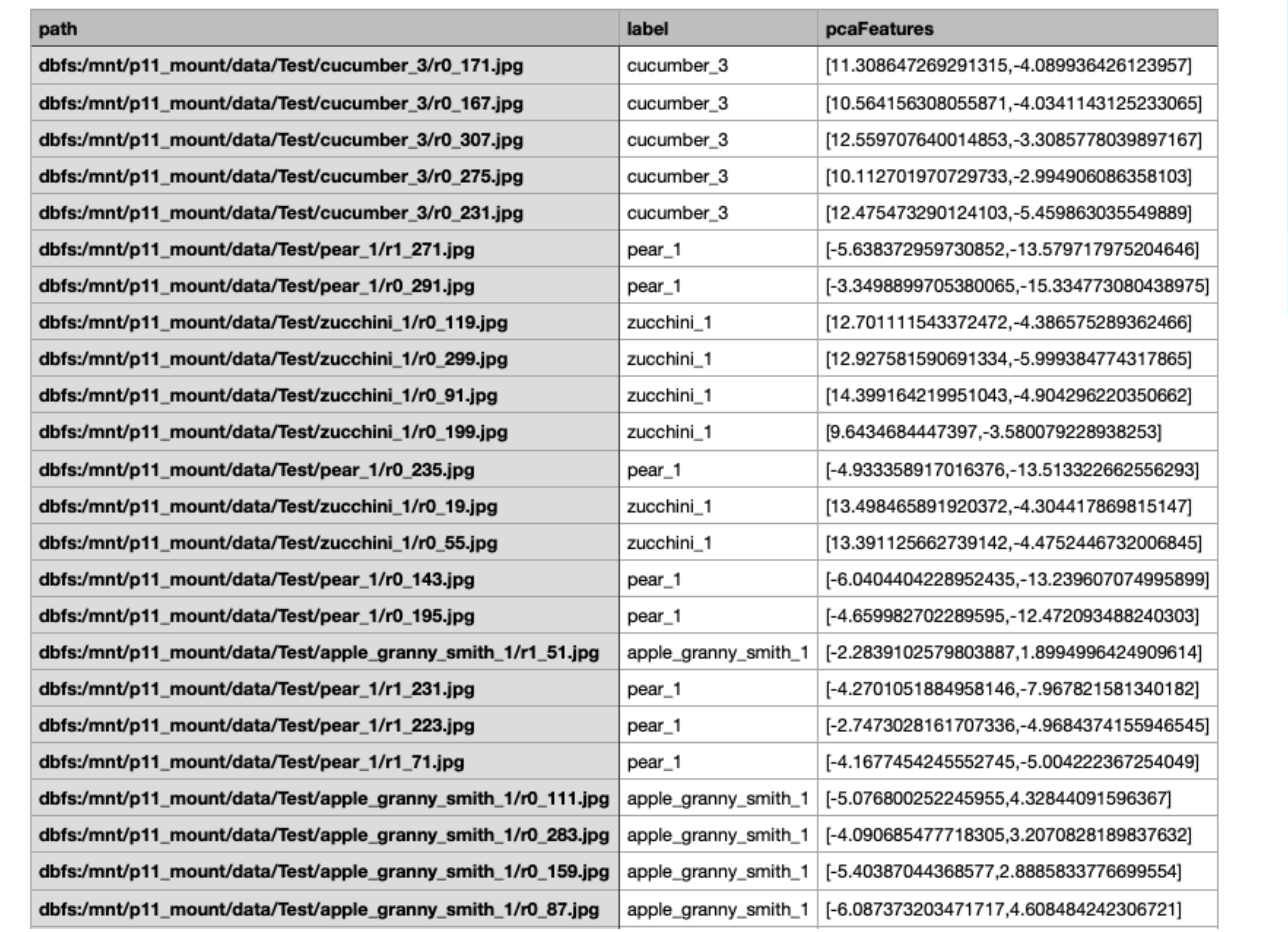
```python
result_pd.head()
```

	path	label	pcaFeatures
0	dbfs:/mnt/p11-mount/data/Test/cucumber_3/r0_13...	cucumber_3	[-5.653728735454188, -7.947577436643717]
1	dbfs:/mnt/p11-mount/data/Test/cucumber_3/r0_10...	cucumber_3	[-5.9150492320066945, -6.272420983625082]
2	dbfs:/mnt/p11-mount/data/Test/cucumber_3/r0_28...	cucumber_3	[-9.963050455379664, 10.163637076832865]
3	dbfs:/mnt/p11-mount/data/Test/cucumber_3/r0_55...	cucumber_3	[-13.266077358608023, 10.8552912070849]
4	dbfs:/mnt/p11-mount/data/Test/cucumber_3/r0_67...	cucumber_3	[-11.185159591911752, 6.171286718304254]



data/Results/result_azure.csv

Blob



path	label	pcaFeatures
dbfs:/mnt/p11_mount/data/Test/cucumber_3/r0_171.jpg	cucumber_3	[11.308647269291315,-4.089936426123957]
dbfs:/mnt/p11_mount/data/Test/cucumber_3/r0_167.jpg	cucumber_3	[10.564156308055871,-4.0341143125233065]
dbfs:/mnt/p11_mount/data/Test/cucumber_3/r0_307.jpg	cucumber_3	[12.559707640014853,-3.3085778039897167]
dbfs:/mnt/p11_mount/data/Test/cucumber_3/r0_275.jpg	cucumber_3	[10.112701970729733,-2.994906086358103]
dbfs:/mnt/p11_mount/data/Test/cucumber_3/r0_231.jpg	cucumber_3	[12.475473290124103,-5.459863035549889]
dbfs:/mnt/p11_mount/data/Test/pear_1/r1_271.jpg	pear_1	[-5.638372959730852,-13.579717975204646]
dbfs:/mnt/p11_mount/data/Test/pear_1/r0_291.jpg	pear_1	[-3.3498899705380065,-15.334773080438975]
dbfs:/mnt/p11_mount/data/Test/zucchini_1/r0_119.jpg	zucchini_1	[12.701111543372472,-4.386575289362466]
dbfs:/mnt/p11_mount/data/Test/zucchini_1/r0_299.jpg	zucchini_1	[12.927581590691334,-5.999384774317865]
dbfs:/mnt/p11_mount/data/Test/zucchini_1/r0_91.jpg	zucchini_1	[14.399164219951043,-4.904296220350662]
dbfs:/mnt/p11_mount/data/Test/zucchini_1/r0_199.jpg	zucchini_1	[9.6434684447397,-3.580079228938253]
dbfs:/mnt/p11_mount/data/Test/pear_1/r0_235.jpg	pear_1	[-4.933358917016376,-13.513322662556293]
dbfs:/mnt/p11_mount/data/Test/zucchini_1/r0_19.jpg	zucchini_1	[13.498465891920372,-4.304417869815147]
dbfs:/mnt/p11_mount/data/Test/zucchini_1/r0_55.jpg	zucchini_1	[13.391125662739142,-4.4752446732006845]
dbfs:/mnt/p11_mount/data/Test/pear_1/r0_143.jpg	pear_1	[-6.0404404228952435,-13.239607074995899]
dbfs:/mnt/p11_mount/data/Test/pear_1/r0_195.jpg	pear_1	[-4.659982702289595,-12.472093488240303]
dbfs:/mnt/p11_mount/data/Test/apple_granny_smith_1/r1_51.jpg	apple_granny_smith_1	[-2.2839102579803887,1.8994996424909614]
dbfs:/mnt/p11_mount/data/Test/pear_1/r1_231.jpg	pear_1	[-4.2701051884958146,-7.967821581340182]
dbfs:/mnt/p11_mount/data/Test/pear_1/r1_223.jpg	pear_1	[-2.7473028161707336,-4.9684374155946545]
dbfs:/mnt/p11_mount/data/Test/pear_1/r1_71.jpg	pear_1	[-4.1677454245552745,-5.004222367254049]
dbfs:/mnt/p11_mount/data/Test/apple_granny_smith_1/r0_111.jpg	apple_granny_smith_1	[-5.076800252245955,4.32844091596367]
dbfs:/mnt/p11_mount/data/Test/apple_granny_smith_1/r0_283.jpg	apple_granny_smith_1	[-4.090685477718305,3.2070828189837632]
dbfs:/mnt/p11_mount/data/Test/apple_granny_smith_1/r0_159.jpg	apple_granny_smith_1	[-5.40387044368577,2.8885833776699554]
dbfs:/mnt/p11_mount/data/Test/apple_granny_smith_1/r0_87.jpg	apple_granny_smith_1	[-6.087373203471717,4.608484242306721]

Conclusions

Analyse des données

Mise en place du modèle

Traitement des données sur le cloud

Respect des normes RGPD

Merci pour votre attention