

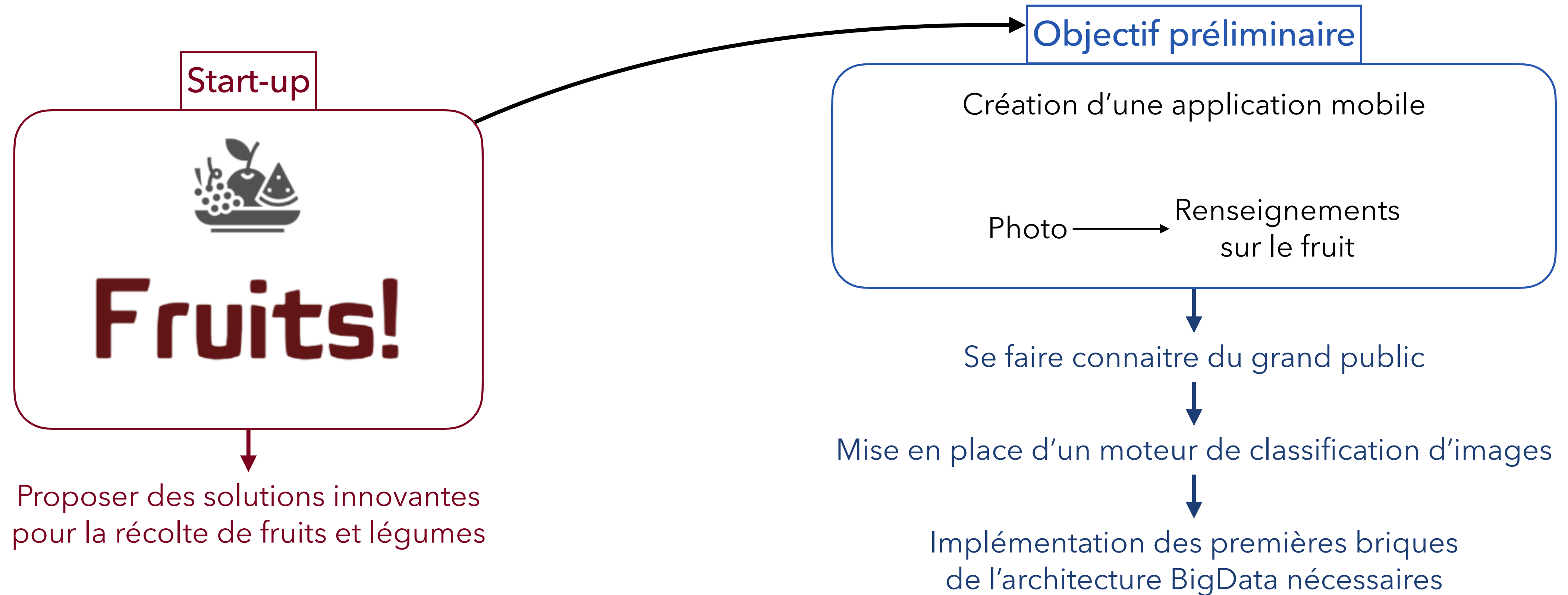
Projet 8

Déployez un modèle dans le cloud

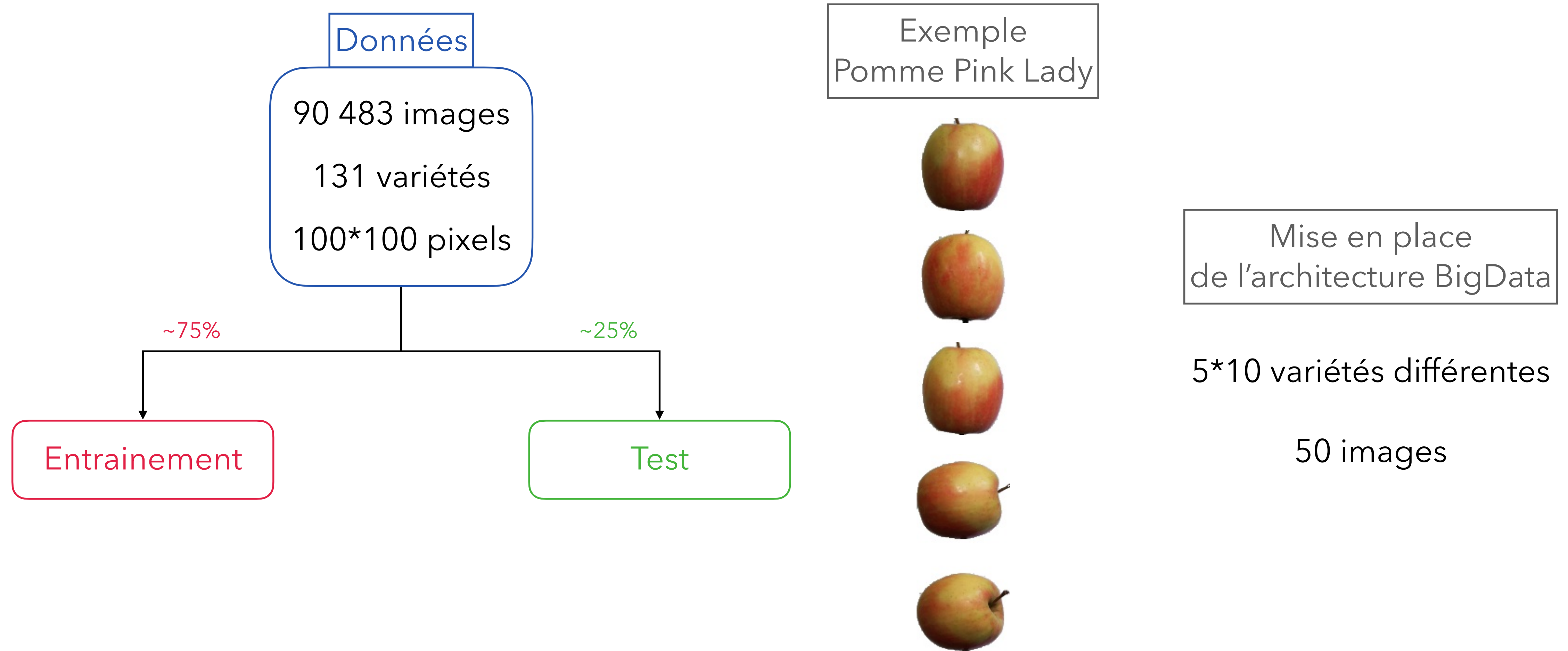
Plan de la présentation

- 1 - Contexte et objectifs du projet
- 2 - Présentation du jeu de données
- 3 - Infrastructure et outils
- 4 - Processus de création de l'environnement BigData
- 5 - Chaine de traitement des images
- 6- Conclusion

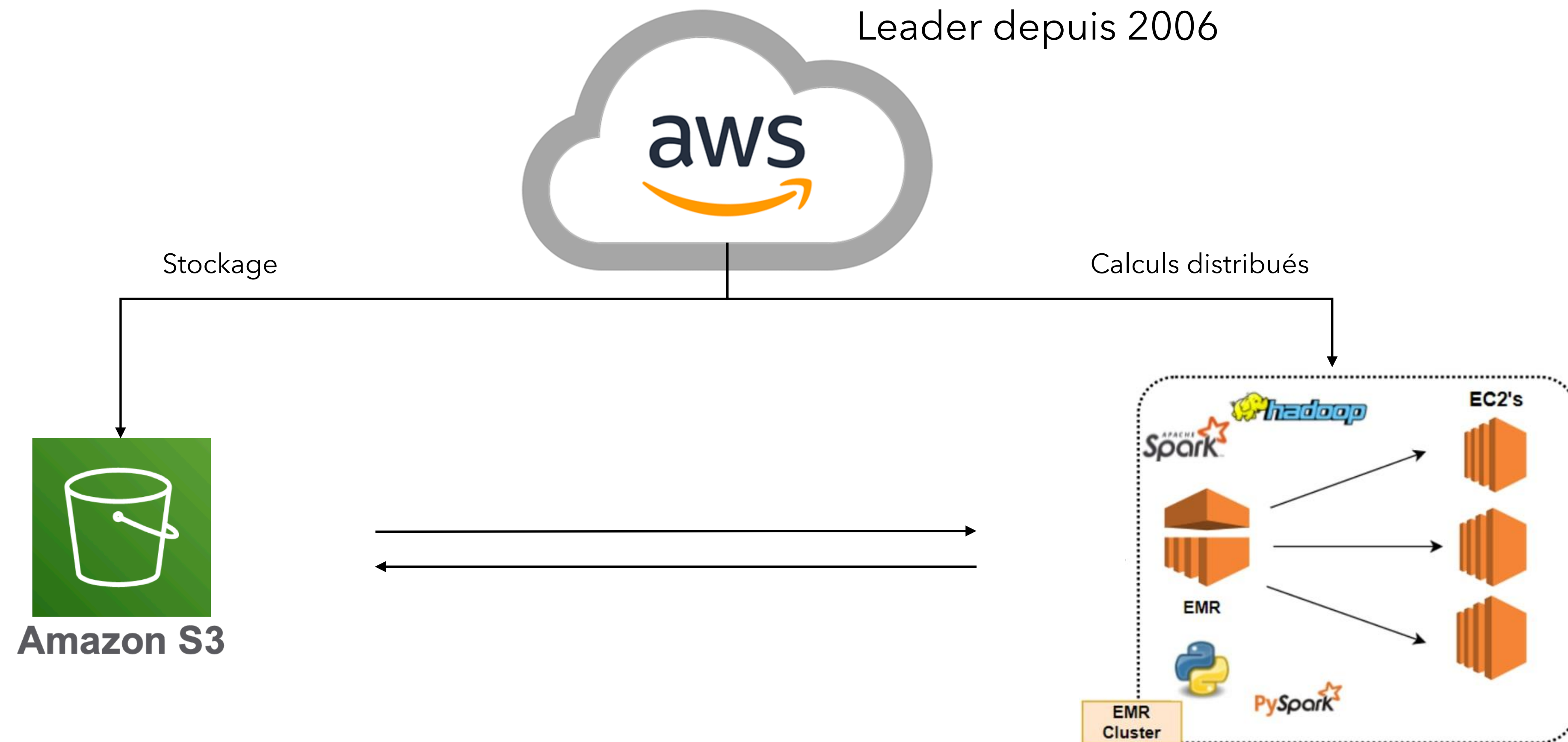
1- Contexte et objectifs



2 - Présentation du jeu de données



3 - Infrastructure et outils



PySpark, outils de communication avec Spark via le langage Python

Spark, framework permettant la distribution et la planification de tâches entre différents exécutants

4 - Processus de création de l'environnement BigData

1. Installation de **AWS Cli** (Interface en ligne de commande) pour l'interaction avec les services

2. Configuration des accès à AWS depuis notre machine locale

```
(projet_8_ocr-hiVB9eyE-py3.10) maxime@MacBook-Pro-14 AWS_keys % aws configure
AWS Access Key ID [None]: 
AWS Secret Access Key [None]: 
Default region name [None]: eu-west-3
Default output format [None]:
```

Paris

Identifiants de sécurité de compte

3. Création du Bucket s3 et dépôt des images

Amazon S3 > Compartiments > p8ocrbucket > img_cloud/		
<input type="checkbox"/>	Nom	Type
<input type="checkbox"/>	Apple Golden 3/	Dossier
<input type="checkbox"/>	Apple Pink Lady/	Dossier
<input type="checkbox"/>	Banana Red/	Dossier
<input type="checkbox"/>	Beetroot/	Dossier
<input type="checkbox"/>	Clementine/	Dossier
<input type="checkbox"/>	Kumquats/	Dossier
<input type="checkbox"/>	Lemon/	Dossier
<input type="checkbox"/>	Peach/	Dossier
<input type="checkbox"/>	Pineapple/	Dossier
<input type="checkbox"/>	Tomato 1/	Dossier

Chargement de 10 dossiers de 5 images

4 - Processus de création de l'environnement BigData

4. Création du cluster EMR

Nom et applications

Nom

p8ocrcluster

Version Amazon EMR

Une version contient un ensemble d'applications susceptibles d'être installées sur votre cluster.

emr-6.15.0

Offre d'applications

Spark Interactive

Core Hadoop

Flink

HBase

Presto

Trino

Custom

☐ Flink 1.17.1

☐ HCatalog 3.1.3

☐ Hue 4.11.0

☐ Livy 0.7.1

☐ Phoenix 5.1.3

☒ Spark 3.4.1

☐ Tez 0.10.2

☐ ZooKeeper 3.5.10

☐ Ganglia 3.7.2

☒ Hadoop 3.3.6

☐ JupyterEnterpriseGateway 2.6.0

☐ MXNet 1.9.1

☐ Pig 0.17.0

☐ Sqoop 1.4.7

☐ Trino 426

☐ HBase 2.4.17

☐ Hive 3.1.3

☒ JupyterHub 1.5.0

☐ Oozie 5.2.1

☐ Presto 0.283

☒ TensorFlow 2.11.0

☐ Zeppelin 0.10.1

Paramètres du catalogue de données AWS Glue

Utilisez le catalogue de données AWS Glue pour fournir un metastore externe à votre application.

☐ Utiliser pour les métadonnées de table Spark

Options du système d'exploitation

☒ Version Amazon Linux :

☐ Amazon Machine Image (AMI) personnalisée

☒ Appliquez automatiquement les dernières mises à jour Amazon Linux

5. Location d'instances EC2 (Elastic Compte Cloud)

Dimensionnement et mise en service du cluster

Configurez des configurations de dimensionnement et de provisionnement pour les groupes de nœuds principaux et de tâches de votre cluster.

Choisir une option

☒ Définir manuellement la taille du cluster

Utilisez cette option si vous connaissez vos modèles de charge de travail à l'avance.

☐ Utiliser la mise à l'échelle gérée par EMR

Surveillez les principales métriques de charges de travail afin qu'EMR puisse optimiser la taille du cluster et l'utilisation des ressources.

☐ Utiliser un autoscaling personnalisée

Pour dimensionner de manière programmatique les unités principales et les nœuds de tâches, créez des politiques d'autoscaling personnalisées.

Configuration de mise en service

Définissez la taille de votre noyau et tâchegroupes 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
Workers	m5.xlarge	2	<input type="checkbox"/>
Unité principale	m5.xlarge	1	<input type="checkbox"/>

- Performances équilibrées, puissance de calcul / mémoire de stockage
- Bon compromis qualité/prix

6. Localisation des instances EC2 - accord avec la réglementation RGPD

Serveurs localisés en Europe : Paris, Londres, Francfort, Stockholm et Irlande

Sélection de Paris : réduction de latence

4 - Processus de création de l'environnement BigData

5. Paramétrages de l'EMR

5.1 Ajout des librairies nécessaires

5.2 Paramètres logiciel

5.3 Configuration de la sécurité d'accès

Fichier bootstrap comme actions d'amorçage

```
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 boto3
9  sudo python3 -m pip install s3fs
10 sudo python3 -m pip install fsspec
```

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

Paire de clés publique / privée permettant une connexion sécurisée via un tunnel SSH

▼ Paramètres logiciels - *facultatif* [Info](#)

☒ Entrer la configuration

☐ Charger JSON à partir d'Amazon S3

```
1  [
2  {
3    "Classification": "jupyter-s3-conf",
4    "Properties": {
5      "s3.persistence.bucket": "p8ocrbucket",
6      "s3.persistence.enabled": "true"
7    }
8  }
9  ]
```

Persistances des données utilisées ou générées par jupyter

4 - Processus de création de l'environnement BigData

6. Création de rôles spécifiques IAM (Identity and Access Management) pour notre cluster et nos instances

Fonction du service

cluster_P8

Profil d'instance

EC2_S3_P8

Ajout d'un niveau de sécurité → limitation des interactions avec les autres services AWS



Environ 10 minutes

Statut et heure	Statut et heure	Statut et heure
<div>Statut</div> <div>⋮ Démarrage en cours</div>	<div>Statut</div> <div>⋮ Action d'amorçage</div>	<div>Statut</div> <div>✔ En attente</div>
<div>Heure de création</div> <div>9 janvier 2024 10:23 (UTC+01:00)</div>	<div>Heure de création</div> <div>9 janvier 2024 10:23 (UTC+01:00)</div>	<div>Heure de création</div> <div>9 janvier 2024 10:23 (UTC+01:00)</div>
<div>Temps écoulé</div> <div>1 seconde</div>	<div>Temps écoulé</div> <div>2 minutes, 31 secondes</div>	<div>Temps écoulé</div> <div>10 minutes, 25 secondes</div>

4 - Processus de création de l'environnement BigData

7. Autorisation d'écoute des tunnels SSH

EC2 > Groupes de sécurité > sg-03892863226fb3809

sg-03892863226fb3809 - ElasticMapReduce-master

-	sgr-015683a6df8d199...	IPv4	SSH	TCP	22
-	sgr-0e21a5f6a3e2f11b5	IPv6	SSH	TCP	22

8. Connexion à l'EMR ssh -i '1. Données/AWS_keys/ec2_p8.pem' -D 5555 hadoop@ec2-15-237-117-33.eu-west-3.compute.amazonaws.com

Last login: Tue Jan 9 09:24:46 2024



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/

Chemin local de stockage de la clé privée

Définition du port proxy pour accéder aux applications

EEEEEEEEEEEEEEEEEEEE MMMMMMM RRRRRRRRRRRRRR
E:::E M:::M M:::M R:::R
EE:::EEEEEEEE:::E M:::M M:::M R:::RRRRR:::R
E:::E EEEEE M:::M M:::M RR:::R R:::R
E:::E M:::M M:::M M:::M R:::R R:::R
E:::EEEEEEEE M:::M M:::M M:::M R:::RRRRR:::R
E:::E M:::M M:::M M:::M R:::RRRRR:::R
E:::E EEEEE M:::M M:::M M:::M R:::R R:::R
EE:::EEEEEEEE:::E M:::M M:::M R:::R R:::R
E:::E M:::M M:::M RR:::R R:::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM RRRRRRR RRRRRR




[hadoop@ip-172-31-8-189 ~]\$

9. Installation et paramétrage de FoxyProxy



4 - Processus de création de l'environnement BigData

7. Accès au Notebook

Application	URL de l'interface utilisateur 
Gestionnaire de ressources	 http://ec2-15-237-117-33.eu-west-3.compute.amazonaws.com:8088/
JupyterHub	 https://ec2-15-237-117-33.eu-west-3.compute.amazonaws.com:9443/
Nom du nœud HDFS	 http://ec2-15-237-117-33.eu-west-3.compute.amazonaws.com:9870/
Serveur d'historique Spark	 http://ec2-15-237-117-33.eu-west-3.compute.amazonaws.com:18080/

Sign in

Username:

jovyan

Password:

.....

Sign in

jupyterhub Notebook_cloud (modifié)

LogoutControl Panel

FileEditViewInsertCellKernelWidgetsHelp

Non fiablePySpark

Exécuter

Markdown

Projet 8 - Déployez un modèle dans le cloud : Notebook cloud

Table des matières

- 1. [Introduction](#)
 - 1.1. [Contexte](#)
 - 1.2. [Mission](#)
 - 1.3. [Contraintes](#)
 - 1.4. [NOTE](#)
- 2. [Démarrage de la session Spark et importation des librairies](#)
- 3. [Démarrage de la session Spark](#)
 - 3.1. [Importation des librairies](#)
- 4. [Définition des PATH pour le chargement des images et l'enregistrement des résultats](#)

5- Chaîne de traitement des images

1. Démarrage de la session Spark

```
1 # L'exécution de cette cellule démarre l'application Spark
```

[1]

... Starting Spark application

ID	YARN Application ID	Kind	State	Spark UI	Driver log	User	Current session?
0	application_1704291990624_0001	pyspark	idle	Link	Link	None	✓

... SparkSession available as 'spark'.
...

2. Importation des librairies nécessaires

```
1 import pandas as pd
2 from PIL import Image
3 import numpy as np
4 import io
5 from typing import Iterator
6
7 import tensorflow as tf
8 from tensorflow.keras.applications.mobilenet_v2 import MobileNetV2, preprocess_input
9 from tensorflow.keras.preprocessing.image import img_to_array
10 from tensorflow.keras import Model
11 from pyspark.sql.functions import col, pandas_udf, PandasUDFType, element_at, split, udf
12 from pyspark.ml.linalg import Vectors, VectorUDT
13 from pyspark.sql.types import ArrayType, FloatType
14 from pyspark.ml.feature import PCA
```

3. Définition des chemins d'accès aux données

```
1 PATH = 's3://p8ocrbucket'
2 PATH_Data = PATH+'/img_cloud'
3 PATH_Result = PATH+'/Results'
```

PATH : s3://p8ocrbucket
PATH_Data : s3://p8ocrbucket/img_cloud
PATH_Result : s3://p8ocrbucket/Results

Définition simple du fait de l'autorisation de lecture / écriture des EC2 sur le s3

5- Chaîne de traitement des images

4. Chargement des images dans un DataFrame Spark

s3

PATH_Data

path	modificationTime	length	content
s3://p8ocrbucket/...	2024-01-02 10:58:55	6555	[FF D8 FF E0 00 1...
s3://p8ocrbucket/...	2024-01-02 10:58:56	6533	[FF D8 FF E0 00 1...
s3://p8ocrbucket/...	2024-01-02 10:58:56	6473	[FF D8 FF E0 00 1...
s3://p8ocrbucket/...	2024-01-02 10:58:57	5750	[FF D8 FF E0 00 1...
s3://p8ocrbucket/...	2024-01-02 10:58:55	5576	[FF D8 FF E0 00 1...

only showing top 5 rows

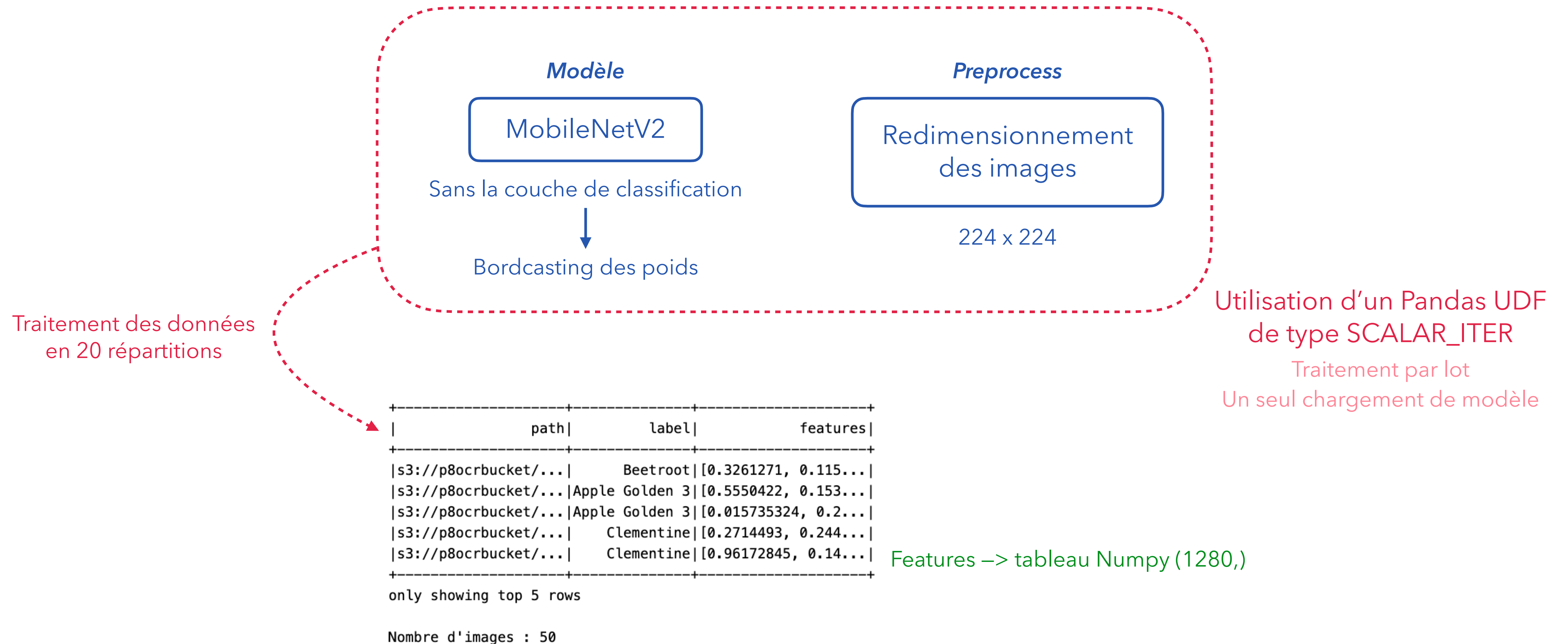
5. Récupération des paths et création des labels

path	label
s3://p8ocrbucket/img_cloud/Pineapple/3_100.jpg	Pineapple
s3://p8ocrbucket/img_cloud/Pineapple/232_100.jpg	Pineapple
s3://p8ocrbucket/img_cloud/Pineapple/217_100.jpg	Pineapple
s3://p8ocrbucket/img_cloud/Pineapple/r_75_100.jpg	Pineapple
s3://p8ocrbucket/img_cloud/Pineapple/r_208_100.jpg	Pineapple

only showing top 5 rows

5- Chaîne de traitement des images

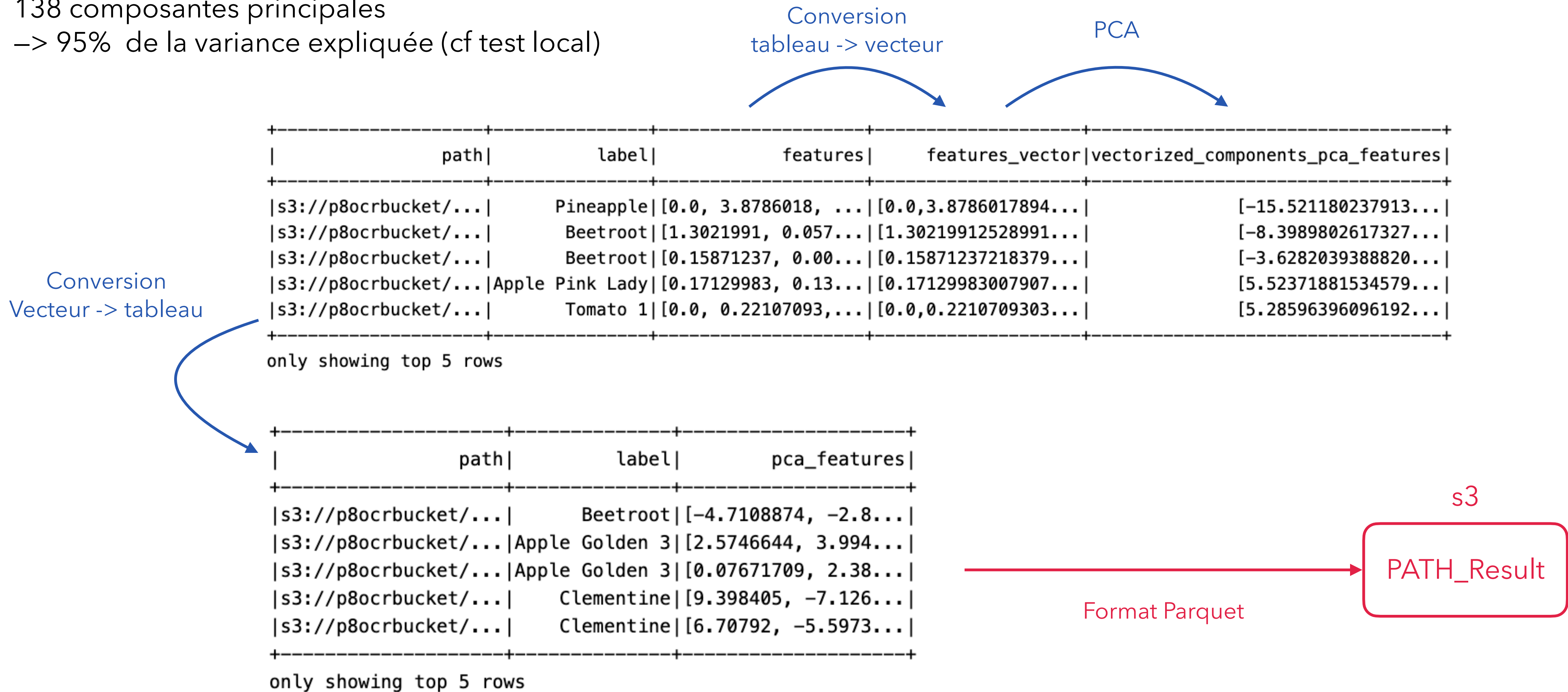
5. Extraction des features



5- Chaîne de traitement des images

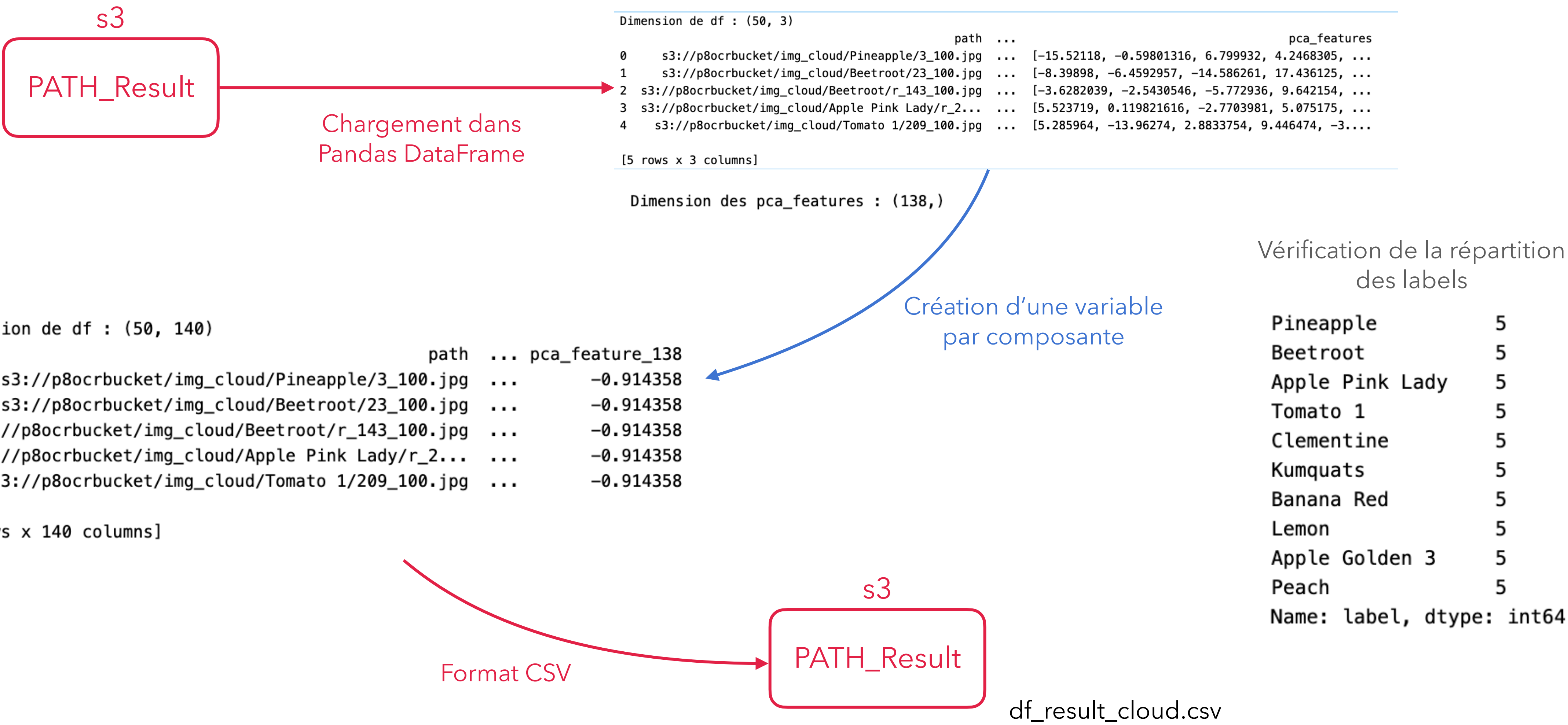
6. Réduction dimensionnelle (ACP)

138 composantes principales
→ 95% de la variance expliquée (cf test local)



5- Chaîne de traitement des images

7. Validation du traitement



CONCLUSION

Cloud AWS particulièrement adapté à notre problématique

- ▶ Simplicité d'utilisation
- ▶ Stockage possible d'un grand volume de données
- ▶ Adaptabilité des ressources en fonction des besoins

Bémols à noter :

- ▶ Coût financier non négligeable pour une utilisation en continue
- ▶ Apprentissage non négligeable pour l'optimisation de ces derniers

Cas de l'étude : location de 3 instances m5.xlarge sur Paris (0,224\$/instance/h)

Autres sites européens possibles: Stockholm (0,204\$/instance), Irlande (0,214\$/instance), Londres(0,222\$/instance) et Francfort (0,230\$/instance).

Question ?