

## *Projet 8*

# Déployez un modèle dans le cloud

Problématique / Big Data / Stratégie retenue / Conclusions

**Camille BRODIN**



# Fruits!

**Startup AgriTech**

## **Objectifs à court terme :**

- Création d'une application mobile de reconnaissance des variétés de sur photo.

## **La mission :**

- Compléter les travaux d'un alternant et mettre en place une architecture BigData sur l'intégralité des données images.

# Problématique : données images

## Base de données :

**Training = 67 692** images jpeg (100x100p) , **131 Dossiers** classes (labels fruits)

**Test = 22 688** images jpeg (100x100p) , **131 Dossiers** classes (labels fruits)

Abricot (164)



Ananas (166)



- Ce volume de données va augmenter très rapidement après la livraison de ce projet.
- Définir l'architecture Big Data requise (avec ici un échantillon d'images pour limiter les couts)

**Sample d'étude = 779 images jpeg (100x100p), 5 Dossiers classes (labels fruits)**

# Architecture Big Data : Contexte

## Introduction BIG DATA:

### BIG DATA

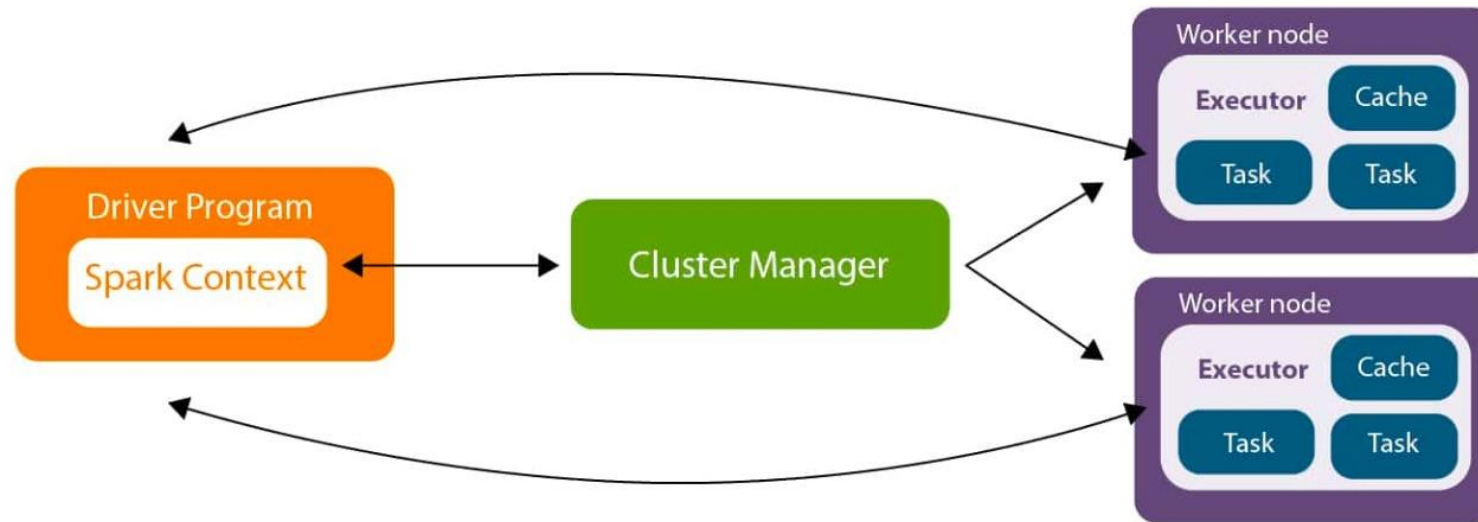


*En Français : les données massives*

- Volume : **trop important pour être stocké et/ou traité sur une seule machine avec des performances acceptables.**
- On référence souvent le Big Data sous les 3 V : **Volume, Vitesse et Variété**. On ajoute parfois Valeur ou Véracité
- Les technologies utilisées pour faire du Big Data sont très nombreuses et très diverses.

# Architecture Big Data : Contexte

**Stratégie BIG DATA:** Spark (ou Apache Spark)

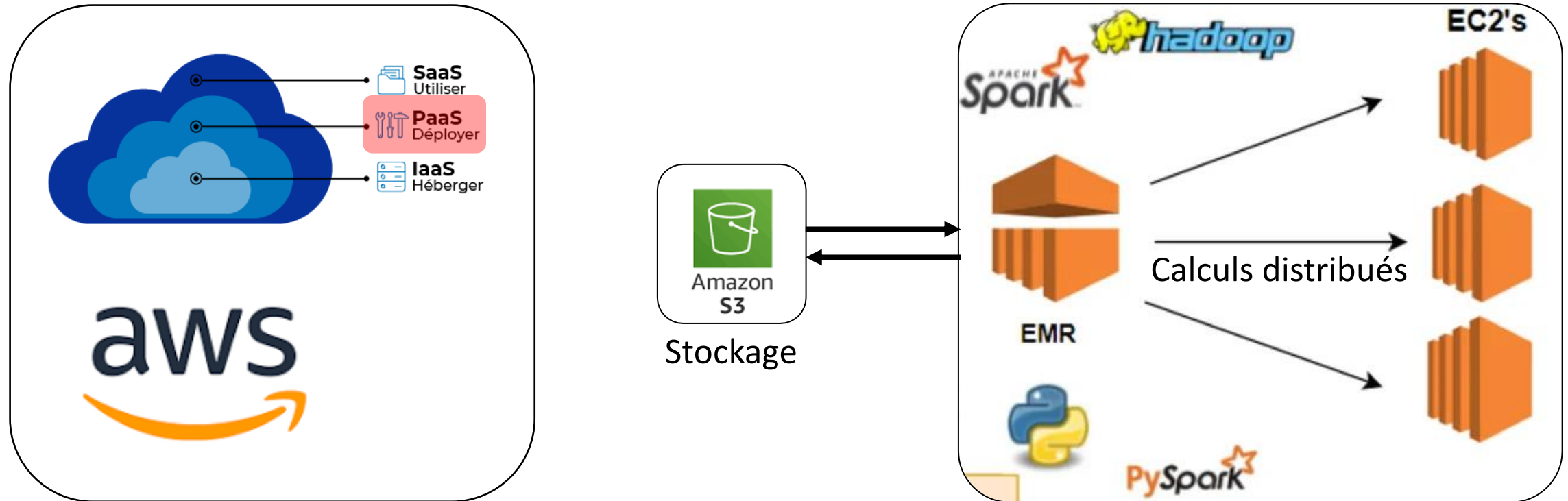


**Spark est un framework open source de calcul distribué pour le traitement et l'analyse de données massives. Il s'agit d'un ensemble d'outils structurés selon une architecture définie.**

- Le processus de pilotage (driver process) est responsable de l'exécution le programme à travers les exécuteurs pour accomplir une tâche donnée.
- Spark emploie un gestionnaire de groupe (cluster manager) qui assure le suivi des ressources disponibles.

# Architecture Big Data : Création de l'environnement

## Zoom sur l'infrastructure AWS



# Architecture Big Data : Création de l'environnement

## 1. Installation de AWS CLI

*L'interface de la ligne de commande AWS (AWS CLI) est un outil qui permet de gérer ses services AWS*

## 2. Configuration des accès à AWS en local

```
PS C:\Users\camil\Documents\Projet8> aws configure
AWS Access Key ID [*****IOG6]: 
AWS Secret Access Key [*****Fckr]: 
Default region name [eu-west-3]: 
Default output format [None]: 
PS C:\Users\camil\Documents\Projet8> aws s3 ls
2024-01-22 12:11:38 aws-logs-405687705265-eu-west-3
2024-01-15 10:51:24 camilleb-projet8
```

Paris

## 3. Création du bucket S3 et import des images

<s3://camilleb-projet8/donneesimages/>

# Architecture Big Data : Création de l'environnement

## 4. Lancement du cluster EMR

Nom

P8\_6

Version Amazon EMR [Info](#)

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

emr-6.3.0

Offre d'applications

Spark

Core Hadoop

HBase

Presto

PrestoSQL

Custom

☐ Flink 1.12.1

☐ HCatalog 3.1.2

☐ Hue 4.9.0

☐ Livy 0.7.0

☐ Phoenix 5.0.0

☐ PrestoSQL 350

☒ TensorFlow 2.4.1

☐ ZooKeeper 3.4.14

☐ Ganglia 3.7.2

☐ Hadoop 3.2.1

☐ JupyterEnterpriseGateway 2.1.0

☐ MXNet 1.7.0

☐ Pig 0.17.0

☒ Spark 3.1.1

☐ Tez 0.9.2

☐ HBase 2.2.6

☐ Hive 3.1.2

☒ JupyterHub 1.2.0

☐ Oozie 5.2.1

☐ Presto 0.245.1

☐ Sqoop 1.4.7

☐ Zeppelin 0.9.0

## 5. Sélection et location d'instances EC2

**Dimensionnement et mise en service du cluster** [Info](#)

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
Unité principale	m5.xlarge	2	<input type="checkbox"/>
Tâche - 1	m5.xlarge	1	<input type="checkbox"/>



# Architecture Big Data : Création de l'environnement

## 6. Paramétrage de l'EMR

### a) Ajout des librairies nécessaires

*Fichier bootstrap comme action d'amorçage*

▼ **Actions d'amorçage – facultatif (1)** Info

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

Nom	Emplacement Amazon S3	Arguments
bootstrap	<a href="s3://camilleb-projet8/bootstrap-emr.sh">s3://camilleb-projet8/bootstrap-emr.sh</a>	-

```
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==1.2.5
7 sudo python3 -m pip install pyarrow
8 sudo python3 -m pip install boto3>=1.9.91
9 sudo python3 -m pip install s3fs==0.4
10 sudo python3 -m pip install fsspec>=0.6.0
11 sudo python3 -m pip install botocore>=1.12.91
12 sudo python3 -m pip install jmespath>=0.7.1
13 sudo python3 -m pip install s3transfer>=0.10.0
14 sudo python3 -m pip install python-dateutil>=2.1
15 sudo python3 -m pip install urllib3>=1.25.4
16 sudo python3 -m pip install six>=1.5
```

### b) Paramètre logiciel

*Persistances des données utilisées ou générées par jupyter*

Afficher JSON pour les configurations de cluster

```
1 {
2   "Classification": "jupyter-s3-conf",
3   "Properties": {
4     "s3.persistance.bucket": "camilleb-projet8",
5     "s3.persistance.enabled": "true"
6   }
7 }
8 }
9 }
```

### c) Configuration de sécurité

*Paire de clés privée/publique pour une connexion sécurisée via tunnel SSH*

Configuration de sécurité

Sélectionnez les paramètres de chiffrement, d'authentification, d'autorisation et de service de métadonnées d'instance de vo

Choisir une configuration de sécu

Parcourir

Créer une configuration de sécurité

---

Paire de clés Amazon EC2 pour SSH sur le cluster

P8\_6\_cle

Parcourir

Créer une paire de clés

# Architecture Big Data : Création de l'environnement

## 7. Création de rôles spécifiques IAM (Identity and Access Management)

**Fonction du service Amazon EMR** Info

La fonction du service est un rôle IAM assumé par Amazon EMR pour mettre en service des ressources et effectuer des actions au niveau du service avec d'autres services AWS.

☒ Choisir une fonction du service existant  
Sélectionnez une fonction du service par défaut ou un rôle personnalisé avec des stratégies IAM attachées afin que votre cluster puisse interagir avec d'autres services AWS.

☐ Créez une fonction du service  
Laissez Amazon EMR créer une nouvelle fonction du service afin que vous puissiez accorder et restreindre l'accès aux ressources d'autres services AWS.

Fonction du service

P8

**Profil d'instance EC2 pour Amazon EMR**

Le profil d'instance attribue un rôle à chaque instance EC2 d'un cluster. Le profil d'instance doit spécifier un rôle qui peut accéder aux ressources pour vos étapes et actions d'amorçage.

☒ Choisir un profil d'instance existant  
Sélectionnez un rôle par défaut ou un profil d'instance personnalisé avec des stratégies IAM attachées afin que votre cluster puisse interagir avec vos ressources dans Amazon S3.

☐ Choisir un profil d'instance  
Laissez Amazon EMR créer un profil d'instance afin de pouvoir spécifier un ensemble personnalisé de ressources auquel il peut accéder dans Amazon S3.

Profil d'instance

ec2\_s3

✓

Votre cluster « P8\_7 » a été créé.

✕

Amazon EMR > EMR sur EC2: Clusters > P8\_7

P8\_7

Mise à jour il y a moins d'une minute

Résilier

Cloner dans AWS CLI

Cloner

▼ Récapitulatif

Informations sur le cluster

ID de cluster  
j-LNMBB8LOWDXL

Configuration de cluster  
Groupes d'instances

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

Applications

Version d'Amazon EMR  
emr-6.3.0

Applications installées  
JupyterHub 1.2.0, Spark 3.1.1, TensorFlow 2.4.1

Gestion des clusters

Destination des journaux dans Amazon S3  
[aws-logs-405687705265-eu-west-3/elasticmapreduce](#)

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

DNS public du nœud primaire  
[ec2-13-38-18-252.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  
22 janvier 2024 16:35 (UTC+01:00)

Temps écoulé  
12 minutes, 19 secondes

10

# Architecture Big Data : Création de l'environnement

## 8. Autorisation d'écoute des tunnels SSH

Groupe de sécurité ElasticMapReduce-master

Règles entrantes (9)									
<input type="text" value="Search"/>									
<div><span>&lt;</span> 1 <span>&gt;</span> <span>⚙</span></div>									
<input type="checkbox"/>	Name	ID de règle de grou...	Version IP	Type	Protocole	Plage de ports	Source		
<input type="checkbox"/>	-	sgr-0b5e997ac3f11d92b	IPv6	SSH	TCP	22	::/0		
<input type="checkbox"/>	-	sgr-0594d4e0129b27acf	IPv4	SSH	TCP	22	0.0.0.0/0		

## 9. Configuration d'un tunnel SSH à l'aide du réacheminement de port dynamique avec l'interface AWS CLI

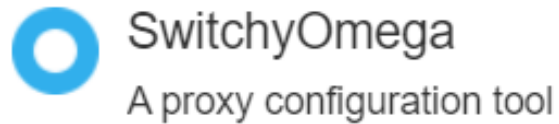
aws emr list-clusters

aws emr socks --cluster-id j-2AL4XXXXXX5T9 --key-pair-file ~/mykeypair.pem

```
PS C:\Users\camil\Documents\Projet8> aws emr socks --cluster-id j-2D8AMHMC58AYF --key-pair-file ~/P8_7_clone1_cle.pem  
ssh -o StrictHostKeyChecking=no -o ServerAliveInterval=10 -ND 8157 -i ~/P8_7_clone1_cle.pem hadoop@ec2-15-236-205-132.eu-west-3.compute.amazonaws.com
```

# Architecture Big Data : Création de l'environnement

## 10. Configurer un proxy SOCKS pour votre navigateur.



### SETTINGS

- Interface
- General
- Import/Export

### PROFILES

- proxy
- emr-socks-proxy**
- auto switch
- + New profile...

### ACTIONS

- Apply changes**
- Discard changes

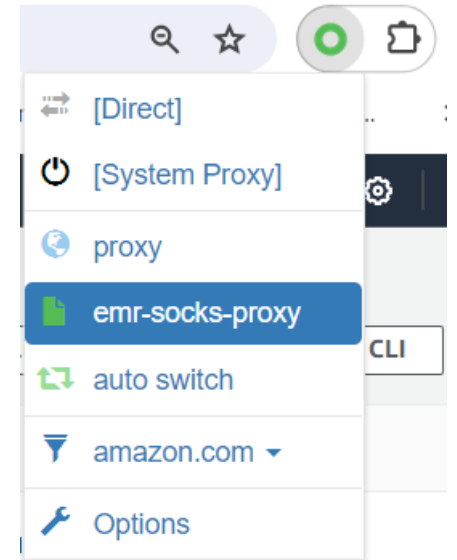
Profile :: emr-socks-proxy

### PAC URL

The PAC script will be updated from this URL. If it is left blank, the following script will be used directly instead.

### PAC Script


```
function FindProxyForURL(url, host) {  
  if (shExpMatch(url, "*ec2*.compute*.amazonaws.com*")) return 'SOCKS5 localhost:8157';  
  if (shExpMatch(url, "*ec2*.compute*")) return 'SOCKS5 localhost:8157';  
  if (shExpMatch(url, "http://10.*")) return 'SOCKS5 localhost:8157';  
  if (shExpMatch(url, "*10*.compute*")) return 'SOCKS5 localhost:8157';  
  if (shExpMatch(url, "*10*.amazonaws.com*")) return 'SOCKS5 localhost:8157';  
  if (shExpMatch(url, "*.compute.internal*")) return 'SOCKS5 localhost:8157';  
  if (shExpMatch(url, "*ec2.internal*")) return 'SOCKS5 localhost:8157';  
  return 'DIRECT';  
}
```

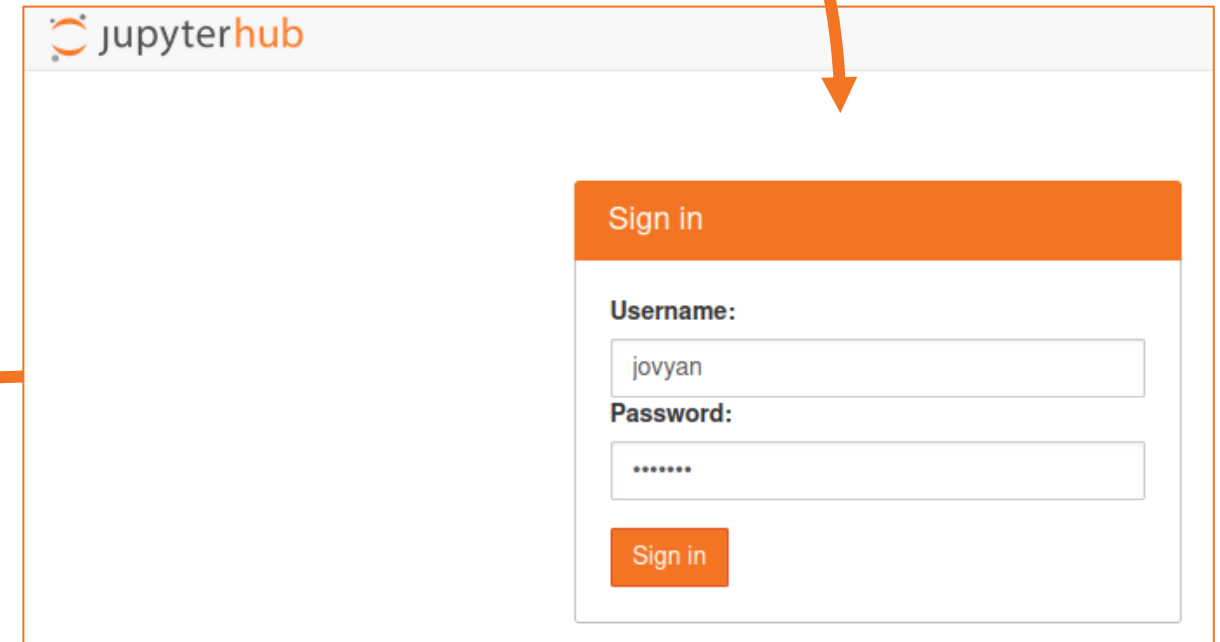
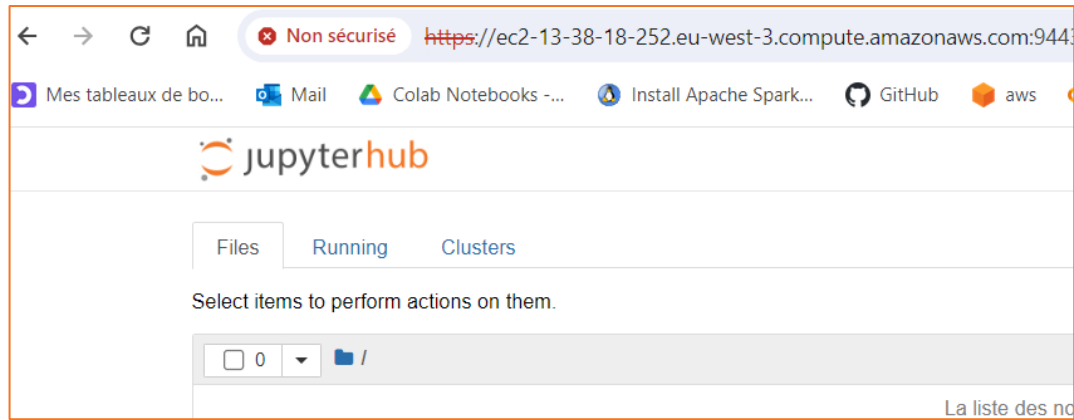


# Architecture Big Data : Création de l'environnement

## Interfaces utilisateur d'application sur le nœud primaire

Celles-ci nécessitent l'activation du tunneling SSH.

Application	URL de l'interface utilisateur 
Gestionnaire de ressources	<a href="http://ec2-15-236-205-132.eu-west-3.compute.amazonaws.com:8088/">http://ec2-15-236-205-132.eu-west-3.compute.amazonaws.com:8088/</a>
JupyterHub	<a href="https://ec2-15-236-205-132.eu-west-3.compute.amazonaws.com:9443/">https://ec2-15-236-205-132.eu-west-3.compute.amazonaws.com:9443/</a>
Nom du nœud HDFS	<a href="http://ec2-15-236-205-132.eu-west-3.compute.amazonaws.com:9870/">http://ec2-15-236-205-132.eu-west-3.compute.amazonaws.com:9870/</a>
Serveur d'historique Spark	<a href="http://ec2-15-236-205-132.eu-west-3.compute.amazonaws.com:18080/">http://ec2-15-236-205-132.eu-west-3.compute.amazonaws.com:18080/</a>



# Architecture Big Data : Chaîne de traitement des images

## 1. Démarrage de la session Spark

Choix du Kernel Pyspark

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

Starting Spark application

ID	YARN Application ID	Kind	State	Spark UI	Driver log	Current session?
0	application_1706185631653_0001	pyspark	idle	<a href="#">Link</a>	<a href="#">Link</a>	✓

SparkSession available as 'spark'.

```
: %%info
```

Current session configs: {'driverMemory': '1000M', 'executorCores': 2, 'proxyU:

ID	YARN Application ID	Kind	State	Spark UI	Driver log	Current session?
0	application_1706185631653_0001	pyspark	idle	<a href="#">Link</a>	<a href="#">Link</a>	✓

## 2. Importation des librairies de la session Spark

### 3. Ajustement des PATHs

```
PATH = 's3://camilleb-projet8'
PATH_Data = PATH+'/donneesimages'
PATH_Result = PATH+'/Results'
print('PATH:          '+\
      PATH+'\nPATH_Data:    '+\
      PATH_Data+'\nPATH_Result: '+PATH_Result)
```

```
PATH:          s3://camilleb-projet8
PATH_Data:     s3://camilleb-projet8/donneesimages
PATH_Result:   s3://camilleb-projet8/Results
```

# Architecture Big Data : Chaîne de traitement des images

## 4. Chargement des images dans un DataFrame Spark



## 5. Récupération des paths et création des labels

```
] : images = spark.read.format("binaryFile") \  
    .option("pathGlobFilter", "*.jpg") \  
    .option("recursiveFileLookup", "true") \  
    .load(PATH_Data)
```

```
] : # Affichage de 5 images  
images.show(5)
```

path	modificationTime	length	content
s3://camilleb-pro...	2024-01-23 10:37:48	6328	[FF D8 FF E0 00 1...
s3://camilleb-pro...	2024-01-23 10:37:39	6322	[FF D8 FF E0 00 1...
s3://camilleb-pro...	2024-01-23 10:37:38	6308	[FF D8 FF E0 00 1...
s3://camilleb-pro...	2024-01-23 10:38:36	6304	[FF D8 FF E0 00 1...
s3://camilleb-pro...	2024-01-23 10:37:35	6300	[FF D8 FF E0 00 1...

only showing top 5 rows

```
[9]: # Ajout d'une nouvelle colonne 'label' au dataframe images  
images = images.withColumn('label', element_at(split(images['path'], '/'), -2))
```

```
# Impression des résultats  
images.select('path', 'label').show(5, False)
```

```
# Impression du schéma du dataframe  
print(images.printSchema())
```

path	label
s3://camilleb-projet8/donneesimages/Apple Golden 1/114_100.jpg	Apple Golden 1
s3://camilleb-projet8/donneesimages/Apple Golden 1/103_100.jpg	Apple Golden 1
s3://camilleb-projet8/donneesimages/Apple Golden 1/101_100.jpg	Apple Golden 1
s3://camilleb-projet8/donneesimages/Apple Golden 1/96_100.jpg	Apple Golden 1
s3://camilleb-projet8/donneesimages/Apple Golden 1/100_100.jpg	Apple Golden 1

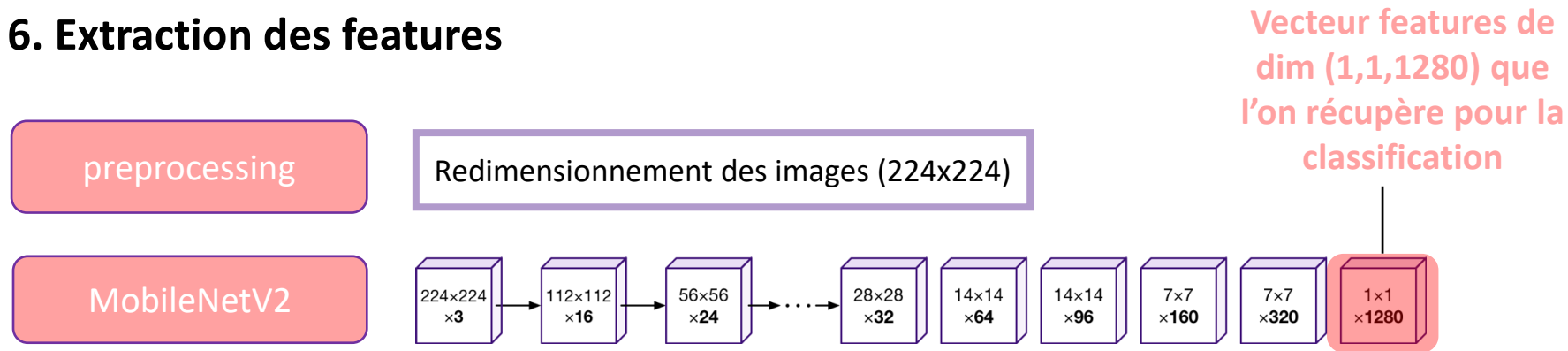
only showing top 5 rows

```
root  
|-- path: string (nullable = true)  
|-- modificationTime: timestamp (nullable = true)  
|-- length: long (nullable = true)  
|-- content: binary (nullable = true)  
|-- label: string (nullable = true)
```

None

# Architecture Big Data : Chaîne de traitement des images

## 6. Extraction des features



`broadcast_weights = sc.broadcast(new_model.get_weights())` -> Diffuser les poids aux unités principales (workers)

path	label	features
s3://camilleb-pro...	Apple Golden 1	[0.0, 0.026634023...
s3://camilleb-pro...	Apple Crimson Snow	[0.0, 0.0, 0.0, 0...
s3://camilleb-pro...	Apple Golden 1	[0.0, 0.010257924...
s3://camilleb-pro...	Apple Golden 1	[0.041660447, 0.0...
s3://camilleb-pro...	Apple Braeburn	[0.5408363, 0.207...

only showing top 5 rows

Nombre d'images : 779

Utilisation d'un Pandas UDF de type `SCALAR_ITER`  
Traitement par lot + modèle chargé une seule fois

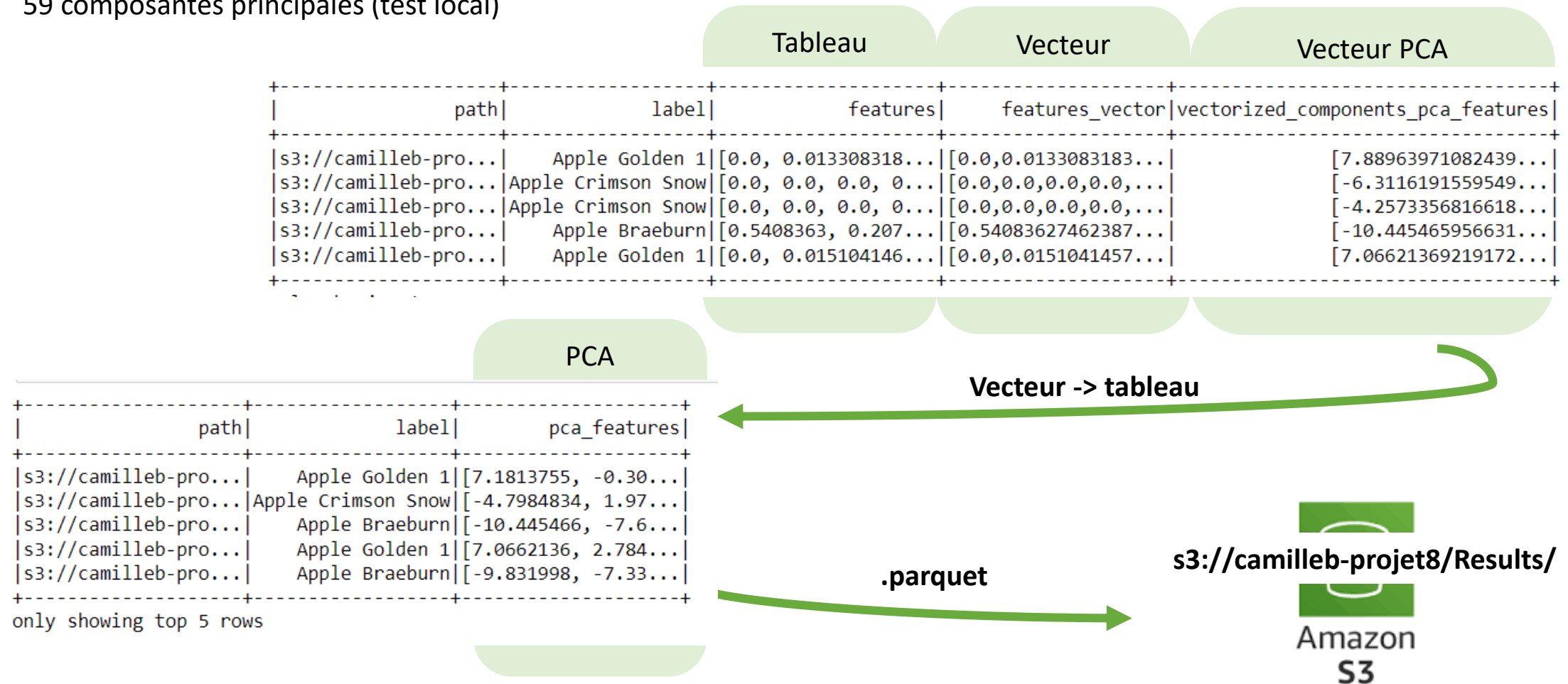
Features -> tableau Numpy



# Architecture Big Data : Chaîne de traitement des images

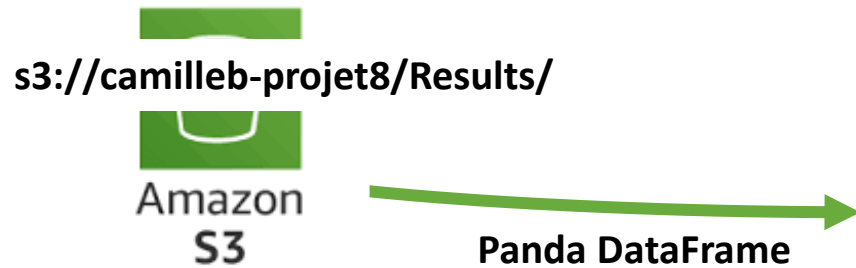
## 7. Réduction dimensionnelle (ACP)

59 composantes principales (test local)



# Architecture Big Data : Chaîne de traitement des images

## 8. Validation du traitement



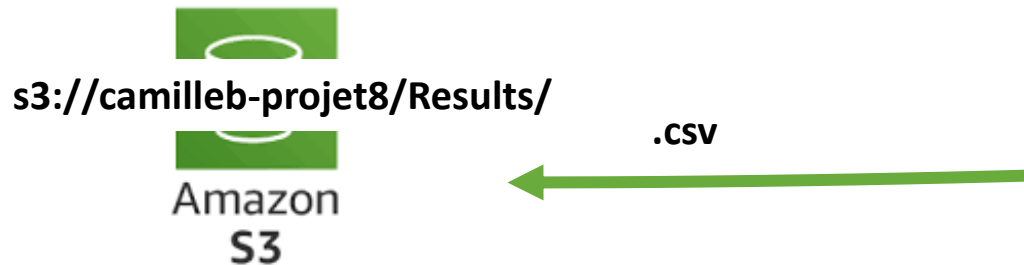
```
Dimension de df : (779, 3)
```

	path	...	pca_features
0	s3://camilleb-projet8/donneesimages/Apple Gold...	...	[7.1813755, -0.30588418, -0.22571632, 0.335645...
1	s3://camilleb-projet8/donneesimages/Apple Crim...	...	[-4.7984834, 1.9792905, -0.1853492, 0.3675855,...
2	s3://camilleb-projet8/donneesimages/Apple Crim...	...	[-9.877423, -6.9510164, -0.16194622, 0.3720356...
3	s3://camilleb-projet8/donneesimages/Apple Crim...	...	[-7.8800607, -5.7731857, -0.26345143, 0.313858...
4	s3://camilleb-projet8/donneesimages/Apple Crim...	...	[-9.340089, -5.0119357, -0.1463543, 0.30558503...

[5 rows x 3 columns]

```
27]: # Validation de la dimension des pca_features :  
print(f"Dimension des pca_features : {df.loc[0, 'pca_features'].shape}")  
Dimension des pca_features : (59,)
```

1 variable par composante



```
Dimension de df : (779, 61)
```

	path	...	pca_feature_59
0	s3://camilleb-projet8/donneesimages/Apple Gold...	...	-0.201702
1	s3://camilleb-projet8/donneesimages/Apple Crim...	...	-0.206420
2	s3://camilleb-projet8/donneesimages/Apple Crim...	...	-0.173428
3	s3://camilleb-projet8/donneesimages/Apple Crim...	...	-0.154983
4	s3://camilleb-projet8/donneesimages/Apple Crim...	...	-0.202378

[5 rows x 61 columns]



## Fruits!

### Cloud AWS adapté à notre problématique



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



Coût financier non négligeable pour une utilisation en continue (location de 3 instances m5.xlarge sur Paris (0,224\$/instance/h))

### Compétences acquises:

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

### Aller plus loin :

- Identifier la maturité des fruits pour les cueillir au bon moment.
- Identifier les pathologies ou les fruits abîmés

**Merci pour votre attention**  
**Des questions ?**