

DÉPLOIEMENT D'UN MODÈLE DANS LE CLOUD



Fruits!

[HTTPS://GITHUB.COM/PIMOUSS75/PROJET_8](https://github.com/pimouss75/projet_8)

MARC SELLAM

PROJET 8

DATA SCIENTIST

AVRIL 2023

Mission	P1
Les informations fournies	P2
Les contraintes	P3
Mise en œuvre	P4
Concepts clés	P5
Traitement des images en local en utilisant PySpark	P13
Déploiement de la solution au niveau du Cloud	P17
Démonstration	P28
Conclusion	P38

SOMMAIRE



Je suis Data Scientist dans une start-up nommée Fruits! .

Elle souhaite une application mobile qui permettrait aux utilisateurs de prendre en photo un fruit et d'en obtenir les informations de mettre en place une première version d'un moteur de classification des images de fruits.

MISSION

I

- Un notebook testé dans un environnement Big Data AWS EMR.
- Un dossier compressé d'images de fruits au format jpg dont le lien en téléchargement direct a été fourni, contenant :
 - 131 dossiers de Fruits différents
 - Un total de 22688 images au format 100 x 100



- En raison de l'augmentation du volume de données après la livraison de ce projet, on souhaite mettre en place une solution Big Data.
- Démonstration d'une instance dans le Cloud et expliquer script PySpark, qui aura été complété :
 - d'un traitement de diffusion des poids.
 - d'une étape de réduction de dimension de type PCA en PySpark.
 - d'une sortie d'une matrice écrite sur un fichier au format csv.
- Respect des contraintes du RGPD .
- Un retour critique de cette solution.

- Complétion du script PySpark et test en local avec un petit volume de données (3 Fruits).

Puis :

- Mise en place de l'environnement Big Data dans le cloud.
- Déploiement du traitement des données dans l'environnement Big Data.
- Test du script avec un volume de données plus importants (7 Fruits) dans le cloud en mode de calcul distribué.
- Surveiller les couts d'utilisation du serveur dans le cloud.

Qu'est-ce que le Big Data ?

Les trois « V » du Big Data

Volume

La quantité de données a son importance. Avec le Big Data, vous devrez traiter de gros volumes de données non structurées et à faible densité. Il peut s'agir de données de valeur inconnue, comme des flux de données Twitter, des flux de clics sur une page Internet ou une application mobile ou d'un appareil équipé d'un capteur. Pour certaines entreprises, cela peut correspondre à des dizaines de téraoctets de données. Pour d'autres, il peut s'agir de centaines de pétaoctets.

Vitesse

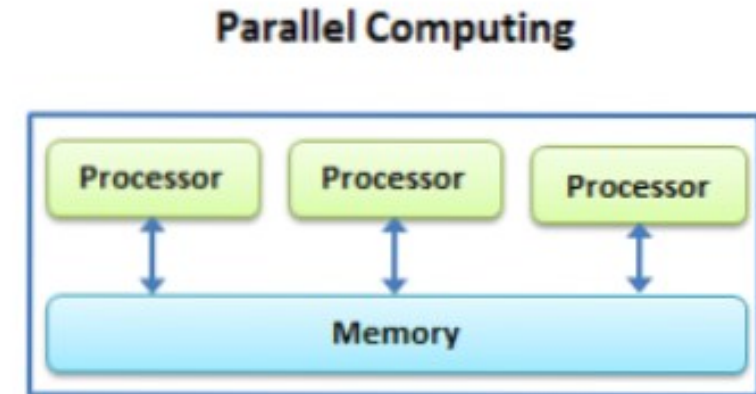
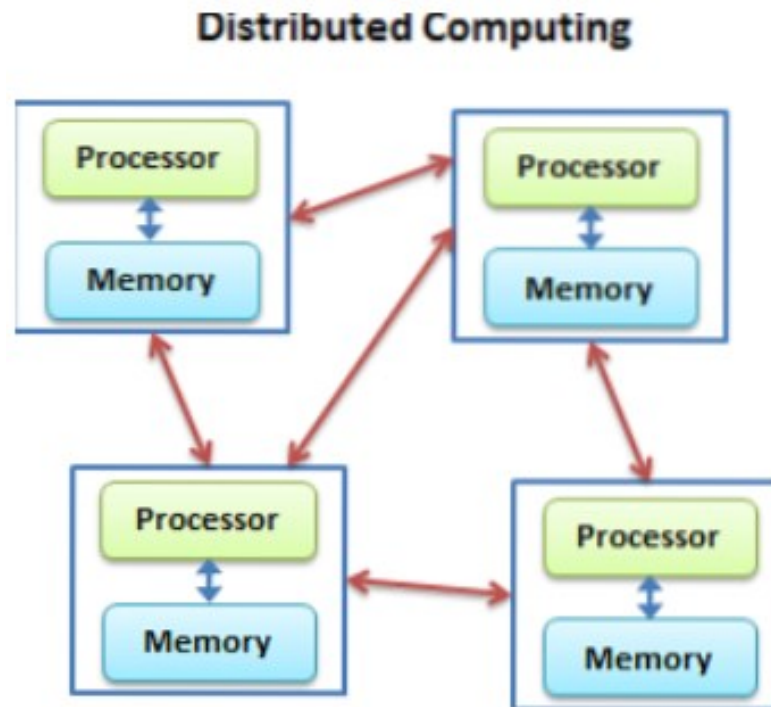
La vitesse à laquelle les données sont reçues et éventuellement traitées. Normalement, les données haute vitesse sont transmises directement à la mémoire, plutôt que d'être écrites sur le disque. Certains produits intelligents accessibles via Internet opèrent en temps réel ou quasi réel et nécessitent une évaluation et une action en temps réel.

Variété

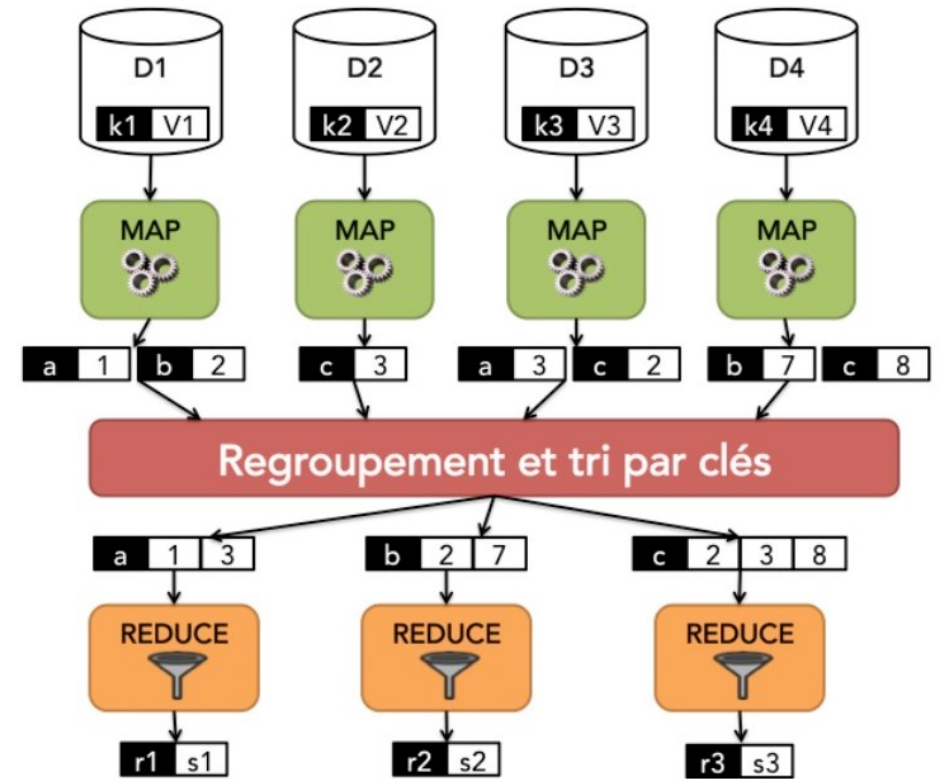
La variété fait allusion aux nombreux types de données disponibles. Les types de données traditionnels ont été structurés et trouvent naturellement leur place dans une [base de données relationnelle](#). Avec l'augmentation du Big Data, les données ne sont pas nécessairement structurées. Les types de données non structurés et semi-structurés, tels que le texte, l'audio et la vidéo, nécessitent un prétraitement supplémentaire pour en déduire le sens et prendre en charge les métadonnées.

Le Calcul distribué

Le modèle distribué permet une plus grande tolérance aux pannes : lorsqu'un nœud du cluster subit une panne, il suffit d'affecter la tâche qu'il était en train de traiter à un autre nœud, alors que dans le modèle parallèle la machine sur laquelle le calcul est exécuté constitue un point unique de défaillance.



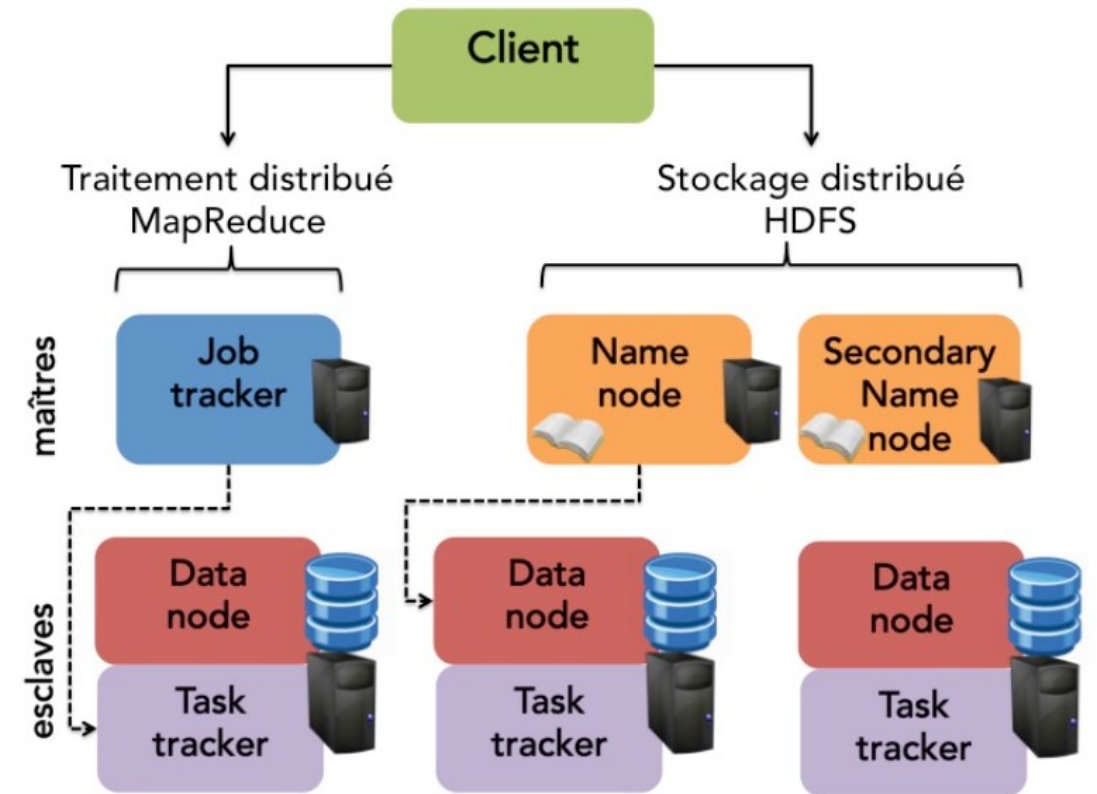
MapReduce

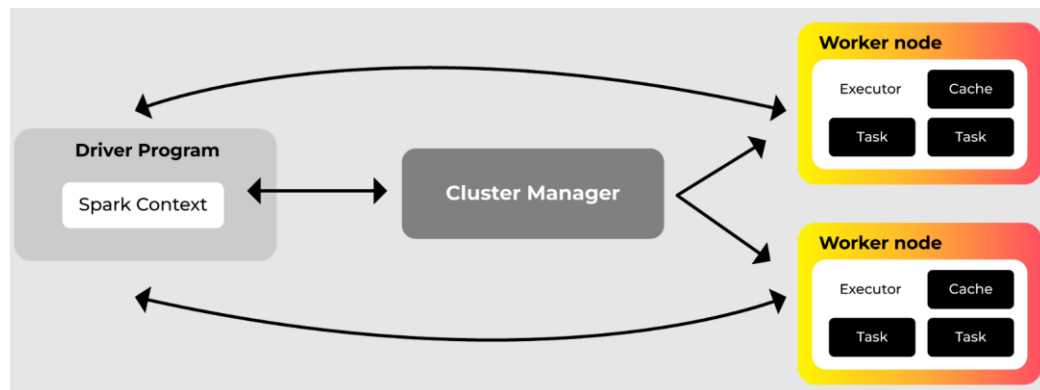


CONCEPTS CLÉS

Socle technique hadoop

- Toute l'architecture support nécessaire pour l'orchestration de MapReduce
- Un système de fichiers HDFS





3 composants principaux :

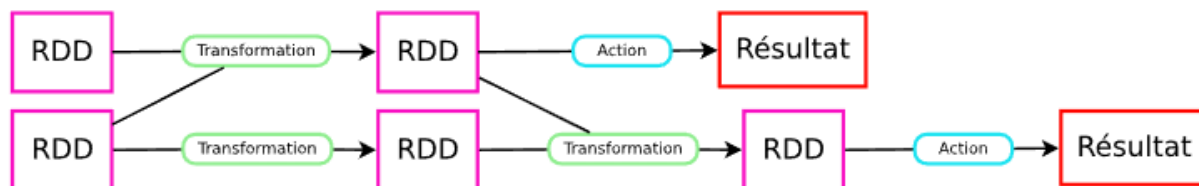
- Un **driver**
- Un **cluster manager**
- Un ou plusieurs **workers**

Apache Spark est une alternative à Hadoop MapReduce pour le **calcul distribué** qui vise à **résoudre** ces deux défauts:

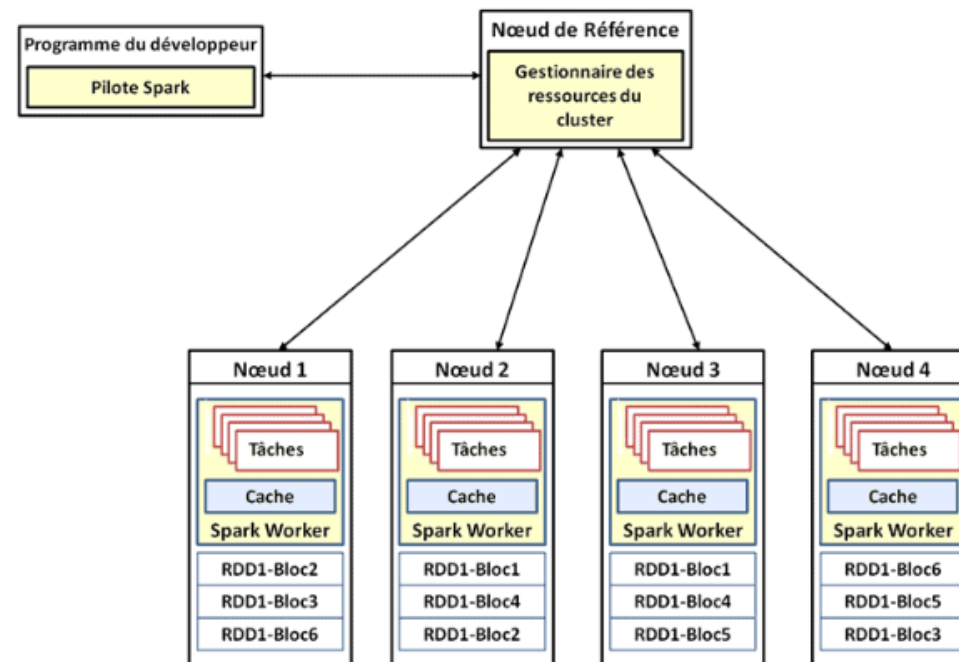
- Des écritures et lectures sont coûteuses en temps.
- Le jeu d'expressions composé exclusivement d'opérations map et reduce est très limité et peu expressif.

La différence fondamentale entre **Hadoop MapReduce** et **Spark** est que **Spark** écrit les données en **RAM**, et non sur disque ainsi que sur l'architecture globale de **Spark**.

Calculs distribués sous forme de graphe avec les DAG



- Un **job** Spark correspond à une action sur un RDD et est composé de plusieurs **étapes** séparées par des **shuffles**.
- Chaque **étape** est composée de **tâches**.
- Chaque tâche s'exécute sur une **partition** différente des données.
- Les partitions sont réparties sur les différents **executors**.
- Les partitions sont créées par les **Resilient Distributed Datasets** (RDD).



PySpark

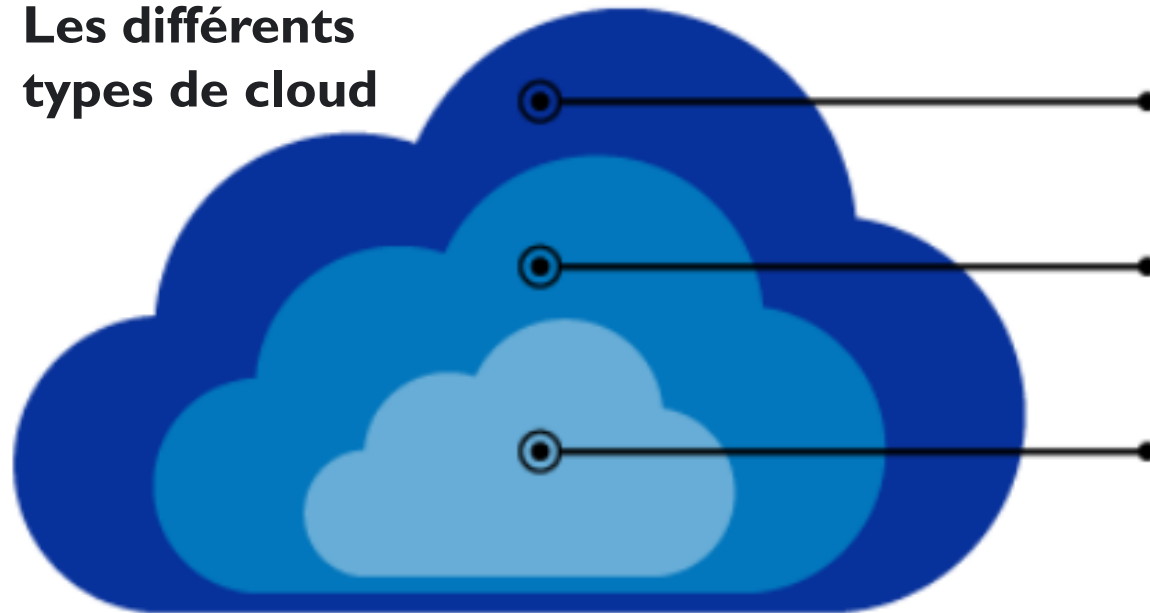
PySpark fournit une API Python pour utiliser les fonctionnalités de Spark, ce qui facilite l'écriture de programmes Spark en Python.



Le Cloud

Le cloud est l'accès et le stockage grâce à Internet. Il s'oppose ainsi à l'utilisation d'un stockage physique ou local par l'intermédiaire d'un disque dur d'ordinateur.

Les différents types de cloud



SaaS
Utiliser

(Software as a Service) : on vous fournit l'accès à un logiciel sous forme de service sous la forme d'une application web (ex. : Microsoft Office)



PaaS
Déployer

(Platform as a Service) : non seulement on vous fournit un accès à l'infrastructure, mais on s'occupe aussi de gérer le nombre de machines nécessaires pour que votre application fonctionne bien.



IaaS
Héberger

(Infrastructure as a Service) : un prestataire vous fournit un accès à tout ou partie de son infrastructure technique, c'est-à-dire à ses serveurs.

Diffusion des poids du modèle de transfert learning

On charge les poids du model pour
etre exécuté sur tous les nœuds

```
In [16]: broadcast_weights = sc.broadcast(new_model.get_weights())
```

Mettons cela sous forme de fonction :

```
In [17]: def model_fn():
    """
    Returns a MobileNetV2 model with top layer removed
    and broadcasted pretrained weights.
    """
    model = MobileNetV2(weights='imagenet',
                        include_top=True,
                        input_shape=(224, 224, 3))
    for layer in model.layers:
        layer.trainable = False
    new_model = Model(inputs=model.input,
                     outputs=model.layers[-2].output)
    new_model.set_weights(broadcast_weights.value)
    return new_model
```

Ajout d'une étape de réduction de dimension

```
n_components = 100
pca = PCA(
    k = n_components,
    inputCol = 'scaledFeatures',
    outputCol = 'pcaFeatures'
).fit(df_scaled)

df_pca = pca.transform(df_scaled)
df_pca.show(6)
```

path	label	scaledFeatures	pcaFeatures
file:/C:/Users/mu...	Apple Red 1	[-1.1977509841651...	[-6.2871393865860...
file:/C:/Users/mu...	Apple Red 1	[-1.1977509841651...	[-5.9051367560626...
file:/C:/Users/mu...	Apple Red 1	[-0.9021714007865...	[-6.7050055947995...
file:/C:/Users/mu...	Apple Red 1	[-0.7579063437929...	[-7.3259522917575...
file:/C:/Users/mu...	Apple Red 1	[-0.8929905795362...	[-6.6761411351968...
file:/C:/Users/mu...	Apple Red 1	[0.60797174798140...	[-10.620241979613...

only showing top 6 rows

```
print('Explained Variance Ratio', pca.explainedVariance.sum())
```

Explained Variance Ratio 0.9240176084006985

Sortie de la réduction de dimension (fichier au format csv)

```
+-----+-----+
|      label|count|
+-----+-----+
|      Dates|   166|
|      Kaki |   166|
|Apple Red 1|   164|
+-----+-----+
```

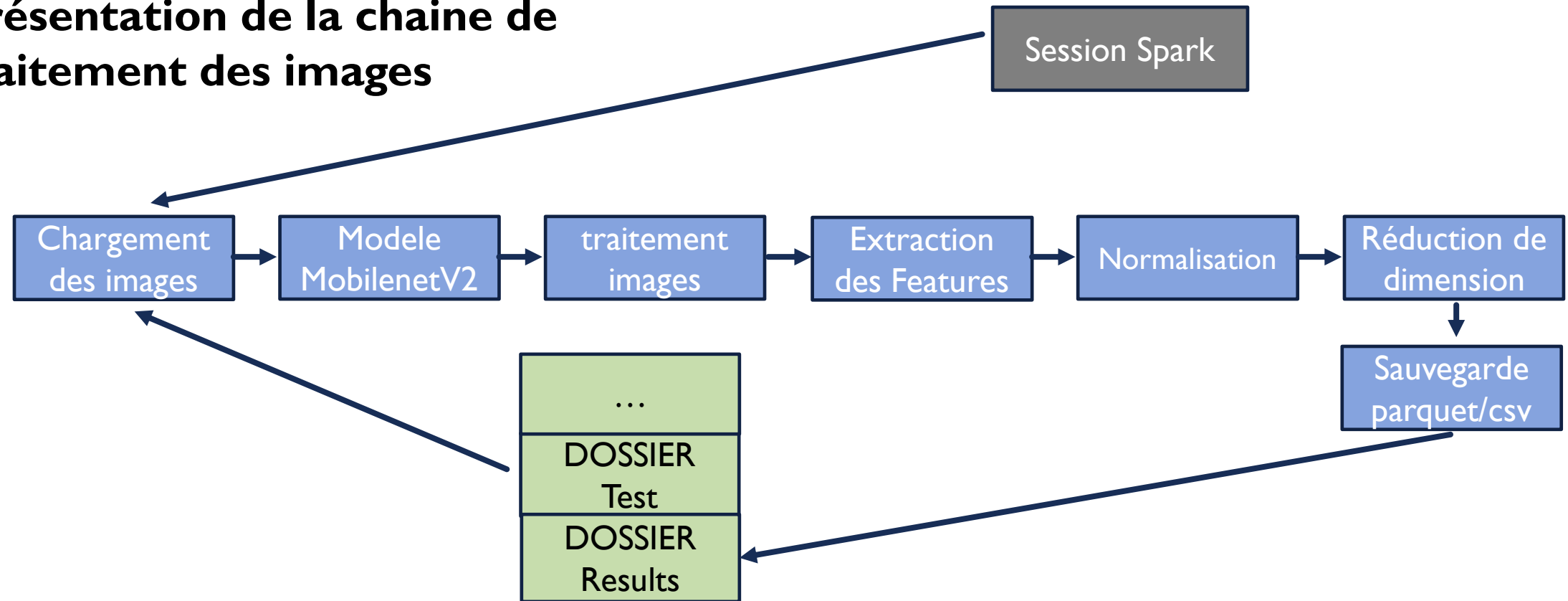
```
df = df_pca.select("*").toPandas()
```

```
df.head()
```

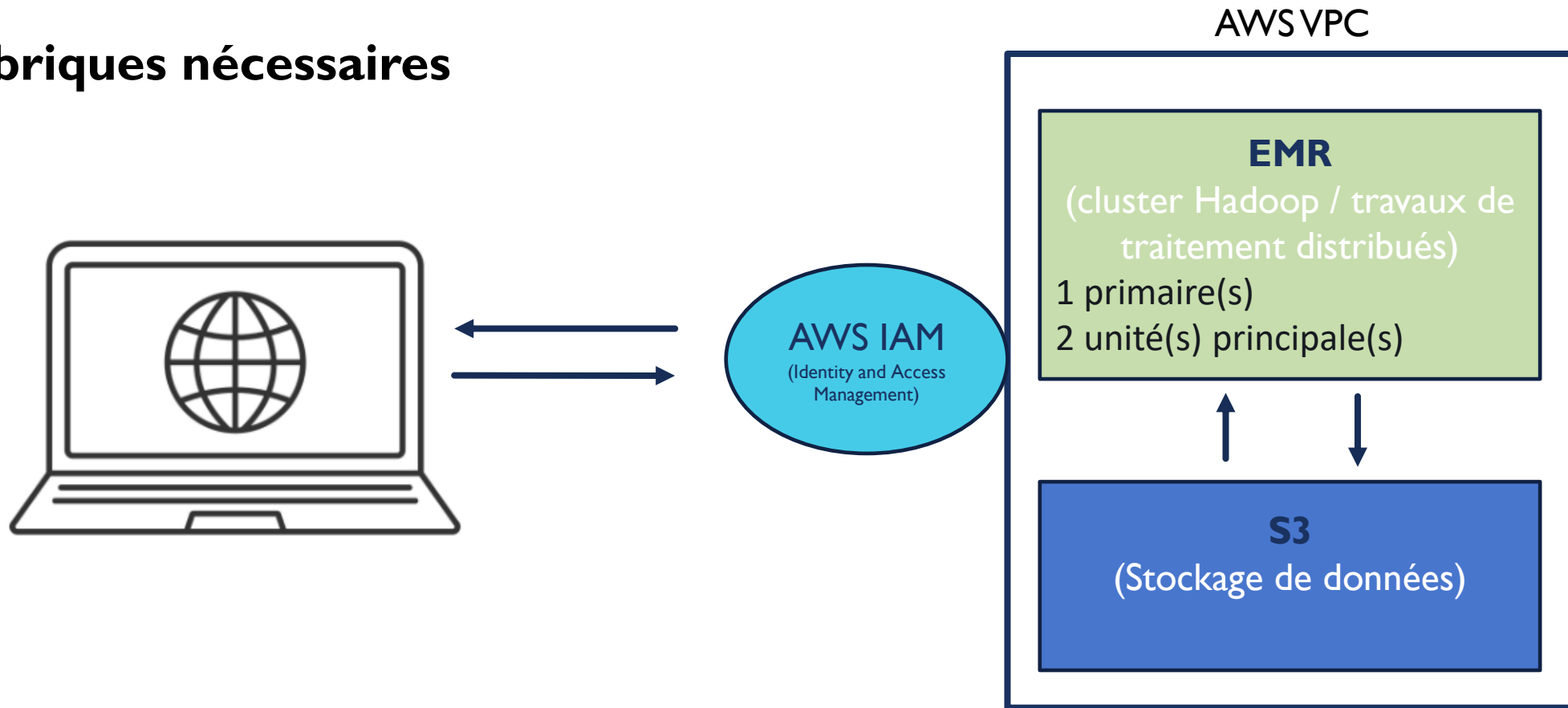
	path	label	pcaFeatures
0	file:/C:/Users/multi/Desktop/Projets_DS/P8_ds/...	Apple Red 1	[-6.287139386586003, -17.056464787267654, -2.4...
1	file:/C:/Users/multi/Desktop/Projets_DS/P8_ds/...	Apple Red 1	[-5.905136756062631, -17.296257849603094, -2.5...
2	file:/C:/Users/multi/Desktop/Projets_DS/P8_ds/...	Apple Red 1	[-6.705005594799506, -12.167035846753548, -3.6...
3	file:/C:/Users/multi/Desktop/Projets_DS/P8_ds/...	Apple Red 1	[-7.3259522917575355, -16.293423930727133, -5....
4	file:/C:/Users/multi/Desktop/Projets_DS/P8_ds/...	Apple Red 1	[-6.676141135196839, -16.836243784648985, -6.7...

```
df.to_csv("{}results.csv".format(PATH_Result), index=False)
```

Présentation de la chaine de traitement des images



Les briques nécessaires



DÉPLOIEMENT DE LA SOLUTION AU NIVEAU DU CLOUD

Configuration AWS

- Création d'une clé d'accès AWS (AWS IAM)
- Installation et configuration de AWS CLI

```
C:\Users\multi\Desktop>aws configure
AWS Access Key ID [None]: XXXXXXXXXXXXPJMM6A67NM
AWS Secret Access Key [None]: XXXXXXXXXXXcZbApzy5mCDUGgm2OnLz9FKEvf/sSD
Default region name [None]: eu-west-3
Default output format [None]: json

aws swf list-domains --registration-status REGISTERED --output json|
```

Création d'un bucket S3 et upload des images dans le dossier Test

```
C:\Users\multi\Desktop>aws s3 mb s3://p8-data  
  
C:\Users\multi\Desktop>aws s3 ls  
2023-03-28 14:57:27 p8-ms  
  
aws s3 sync Test/ s3://p8-ms/Test
```

Amazon S3 > Compartiments > p8-ms > Test/

Test/

Objets | Propriétés

Objets (7)

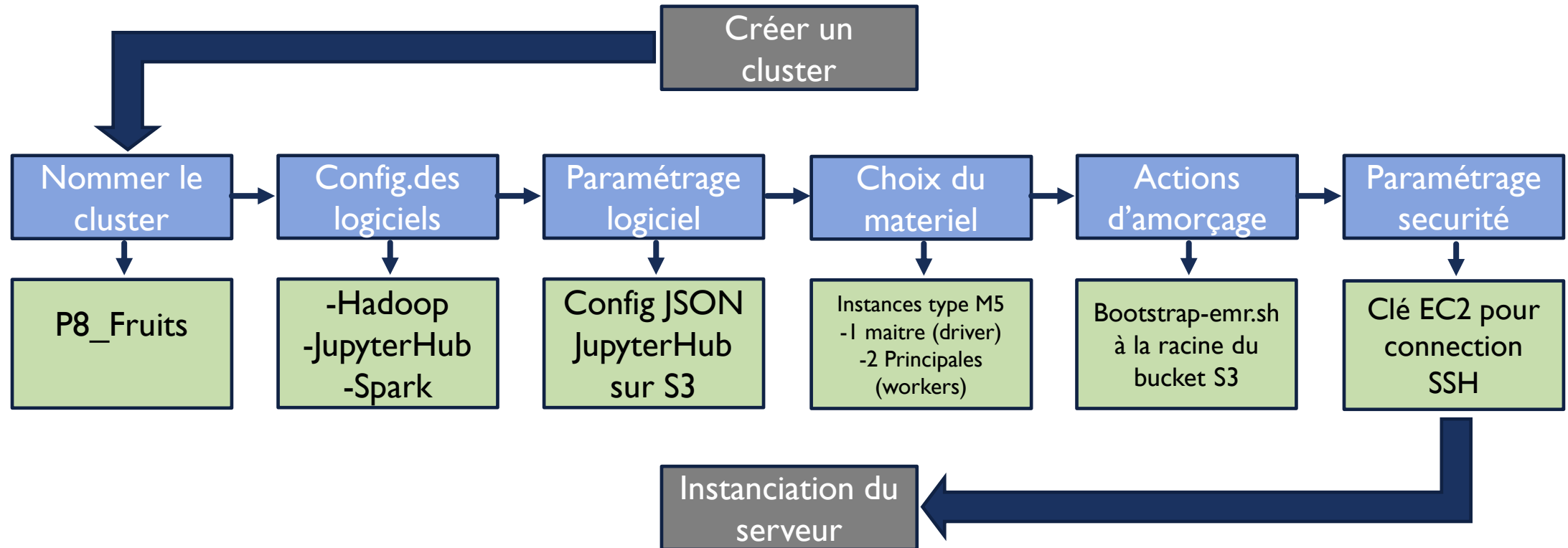
Les objets sont les entités fondamentales stockées dans Amazon S3. Vous pouvez utiliser l'inventaire Amazon S3 pour obtenir leur accord explicite des autorisations. [En savoir plus](#)

Copier l'URI S3 Copier l'URL Télécharger Ouvrir S

Rechercher des objets en fonction du préfixe

<input type="checkbox"/>	Nom	Type	Dernière modification
<input type="checkbox"/>	Apple Golden 1/	Dossier	-
<input type="checkbox"/>	Banana/	Dossier	-
<input type="checkbox"/>	Corn/	Dossier	-
<input type="checkbox"/>	Kaki/	Dossier	-
<input type="checkbox"/>	Lemon/	Dossier	-
<input type="checkbox"/>	Orange/	Dossier	-
<input type="checkbox"/>	Pineapple/	Dossier	-

Configuration de l'EMR



Paramétrage logiciel

Propriétés	Actions d'amorçage	Instances	Étapes	Applications	Configurations	Surveillance	Évènements	identifications (0)
Configurations de cluster Les configurations de cluster sont définies lorsque vous créez un cluster.								
<input type="text" value="Rechercher des configurations"/> Toute classification ▼ < 1 >								
Classification ▼	Propriété ▼	Valeur ▼	Source ▼					
jupyter-s3-conf	s3.persistence.bucket	p8-ms	Configurations de cluster					
jupyter-s3-conf	s3.persistence.enabled	true	Configurations de cluster					

Afficher JSON

Afficher JSON pour les configurations de cluster

```
1 [
2   {
3     "Classification": "jupyter-s3-conf",
4     "Properties": {
5       "s3.persistence.bucket": "p8-ms",
6       "s3.persistence.enabled": "true"
7     }
8   }
9 ]
```

Actions d'amorçage

```
C:\Users\multi\Desktop>aws s3 cp bootstrap-emr.sh s3://p8-ms/  
upload: .\bootstrap-emr.sh to s3://p8-ms/bootstrap-emr.sh
```

Propriétés	Actions d'amorçage	Instances	Étapes	Applications	Configurations
Actions d'amorçage (1) Info					
Nom		Emplacement Amazon S3 🔗			
Action personnalisée		s3://p8-ms/bootstrap-emr.sh			

Contenu du fichier bootstrap-emr.sh

```
#!/bin/bash  
sudo python3 -m pip install -U setuptools  
sudo python3 -m pip install -U pip  
sudo python3 -m pip install wheel  
sudo python3 -m pip install pillow  
sudo python3 -m pip install pandas==1.2.5  
sudo python3 -m pip install pyarrow  
sudo python3 -m pip install boto3  
sudo python3 -m pip install s3fs  
sudo python3 -m pip install fsspec  
sudo python3 -m pip install tensorflow
```


Paramétrage de sécurité

The screenshot displays the AWS Management Console interface for configuring an Amazon EMR cluster. The top navigation bar includes the 'Services' menu, a search bar with the placeholder 'Rechercher', and a keyboard shortcut '[Alt+S]'. On the right side of the header, there are icons for help, notifications, and a user profile dropdown showing 'Paris' and 'Pimouss75'.

The main content area is divided into three columns:

- Réseau (Network):**
 - Cloud privé virtuel (VPC): [vpc-08a4be2488bb609c0](#)
 - Sous-réseau(x) et zone(s) de disponibilité: [subnet-017647de2df3b044f](#) | eu-west-3b
 - ▼ Groupes de sécurité EC2 (pare-feu):
 - Nœud primaire**
Groupe de sécurité géré EMR
[sg-0f5564a70684f0dfb](#)
 - Additional security groups: -
 - Unités principales et nœuds de tâches**
Groupe de sécurité géré EMR
[sg-0efe45d54449faf51](#)
 - Additional security groups: -
- Configuration de sécurité (Security Configuration):**
 - Configuration de sécurité: Aucun
 - Paire de clés EC2: [my_key](#) (indicated by a blue arrow)
- Autorisations (Permissions):**
 - Fonction du service pour Amazon EMR: [EMR_DefaultRole](#)
 - Rôle IAM pour le profil d'instance: [EMR_EC2_DefaultRole](#)
 - Rôle de dimensionnement automatique: Non configuré

Instanciación du serveur

P8_Fruits

Mise à jour il y a moins d'une minute

Actions

Récapitulatif

Informations sur le cluster

ID de cluster
j-37KV6WG5ZS1I5

Configuration de cluster
Groupes d'instances

Capacité
1 primaire(s) | 2 unité(s) principale(s) | 0 tâche(s)

Applications

Version d'Amazon EMR
emr-6.9.0

Applications installées
Hadoop 3.3.3, JupyterHub 1.4.1, Spark 3.3.0

Gestion des clusters

URI du journal Amazon S3
[s3://p8-ms/](#)

DNS public du nœud primaire
[ec2-13-38-114-127.eu-west-3.compute.amazonaws.com](#)

[Connexion au nœud primaire à l'aide de SSH](#)

Statut et heure

Statut
[Action d'amorçage](#)

Heure de création
9 avril 2023 13:46 (UTC+02:00)

Temps écoulé
4 minutes, 30 secondes

Propriétés

Actions d'amorçage

Instances

Étapes

Applications

Configurations

Surveillance

Évènements

identifications (0)

Paramètres du groupe d'instances

Info

Modifier l'option de mise à l'échelle du cluster

Option de mise à l'échelle du cluster
Définir manuellement la taille du cluster

Unité principale

Nom et nombre maximal de nœuds principaux dans le cluster
Groupe d'instances principal - 2 | 2 instances

Groupes d'instances (2)

Info

Avec la configuration des groupes d'instances, chaque type de nœud est constitué du même type d'instance et de la même option d'achat d'instances : à la demande ou Spot.

Redimensionner le groupe d'instances

Ajouter un groupe d'instances de tâches

Rechercher une ressource

< 1 >

Type de ... ▲	Nom ▼	ID ▼	Statut ▼	Instances ▼	Type d'instance ▼	Option d'achat ▼	Prix actuel	Heure de création (UTC+02... ▼	Heure de préparation (UTC... ▼
● Primaire	Groupe d'inst...	ig-1WIC2STJJUBJU	⊖ Action d'amorç	0 (1 demandées)	m5.xlarge	À la demande	0.224 USD/h	9 avril 2023 13:46	-
● Principal	Groupe d'inst...	ig-2J5M3MDLTR3OS	⊖ Action d'amorç	0 (2 demandées)	m5.xlarge	À la demande	0.224 USD/h	9 avril 2023 13:46	-

Création des autorisations sur les connexions entrantes

EC2 > Groupes de sécurité > sg-0f5564a70684f0dfb - ElasticMapReduce-master

sg-0f5564a70684f0dfb - ElasticMapReduce-master

Actions

Détails

Nom du groupe de sécurité

ElasticMapReduce-master

ID du groupe de sécurité

sg-0f5564a70684f0dfb

Description

Master group for Elastic MapReduce created on 2023-03-17T03:37:21.909Z

ID de VPC

vpc-08a4be2488bb609c0

Propriétaire

122912766930

Nombre de règles entrantes

9 Entrées d'autorisation

Nombre de règles sortantes

1 Entrée d'autorisation

Règles entrantes

Règles sortantes

Balises

Vous pouvez désormais vérifier la connectivité réseau avec Reachability Analyzer

Exécuter Reachability Analyzer

Règles entrantes (9)

↻

Gérer les balises

Modifier les règles entrantes

🔍 Filtrer les règles des groupes de sécurité

< 1 > ⓘ

<input type="checkbox"/>	Name	ID de règle de grou...	Version IP	Type	Protocole	Plage de ports	Source	Description
<input type="checkbox"/>	-	sgr-0b2372b9c356c40...	-	Tous les ICMP - IPv4	ICMP	Tous	sg-0efe45d54449faf5...	-
<input type="checkbox"/>	-	sgr-05451761dd6626...	IPv4	SSH	TCP	22	0.0.0.0/0	-
<input type="checkbox"/>	-	sgr-0e682d2f353b3c577	-	Tous les UDP	UDP	0 - 65535	sg-0f5564a70684f0df...	-
<input type="checkbox"/>	-	sgr-0671da68c4f40358d	-	Tous les ICMP - IPv4	ICMP	Tous	sg-0f5564a70684f0df...	-
<input type="checkbox"/>	-	sgr-0142441b9e5338...	-	Tous les UDP	UDP	0 - 65535	sg-0efe45d54449faf5...	-
<input type="checkbox"/>	-	sgr-0550a27e3397e6abf	-	Tous les TCP	TCP	0 - 65535	sg-0f5564a70684f0df...	-
<input type="checkbox"/>	-	sgr-06a80465d76d36...	IPv4	TCP personnalisé	TCP	8443	52.46.82.0/23	-
<input type="checkbox"/>	-	sgr-0d544a75702910...	-	Tous les TCP	TCP	0 - 65535	sg-0efe45d54449faf5...	-
<input type="checkbox"/>	-	sgr-02d2945e0d68c3e...	IPv6	SSH	TCP	22	::/0	-

Création du tunnel SSH vers le driver et configuration de foxyproxy

```
C:\Users\multi\Desktop>ssh -i my_key.pem -D 8157 hadoop@ec2-52-47-35-61.eu-west-3.compute.amazonaws.com
The authenticity of host 'ec2-52-47-35-61.eu-west-3.compute.amazonaws.com (52.47.35.61)' can't be established.
ED25519 key fingerprint is SHA256:Rq7cLhUrQsdjBowXecRHqWMut/GIMk5ActJ/54IRXqY.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-52-47-35-61.eu-west-3.compute.amazonaws.com' (ED25519) to the list of known hosts.
```

```
--|  --|  )
_| (  /   Amazon Linux 2 AMI
---| \---|---
```

```
https://aws.amazon.com/amazon-linux-2/
22 package(s) needed for security, out of 25 available
Run "sudo yum update" to apply all updates.
```

```
EEEEEEEEEEEEEEEEEEEE MMMMMMMM MMMMMMMM RRRRRRRRRRRRRRRR
E:EEEEEEEEEEEEEEEE M:EEEEEEEE M:EEEEEEEE R:EEEEEEEEEEEE
EE:EEEEEEEEEEEEEEEE M:EEEEEEEE M:EEEEEEEE R:EEEEEEEEEEEE
E:EE EEEEE M:EEEEEEEE M:EEEEEEEE RR:EEEE R:EEEE
E:EE M:EEEEEEEE M:EEEEEEEE M:EEEEEEEE R:EEEE R:EEEE
E:EEEEEEEEEEEE M:EEEEEEEE M:EEEEEEEE M:EEEEEEEE R:EEEEEEEEEEEE
E:EEEEEEEEEEEE M:EEEEEEEE M:EEEEEEEE M:EEEEEEEE R:EEEEEEEEEEEE
E:EE M:EEEEEEEE M:EEEEEEEE M:EEEEEEEE R:EEEE R:EEEE
E:EE EEEEE M:EEEEEEEE MMM M:EEEEEEEE R:EEEE R:EEEE
EE:EEEEEEEEEEEE M:EEEEEEEE M:EEEEEEEE R:EEEE R:EEEE
E:EEEEEEEEEEEE M:EEEEEEEE M:EEEEEEEE RR:EEEE R:EEEE
EEEEEEEEEEEEEEEEEEEE MMMMMMMM MMMMMMMM RRRRRRRR RRRRRR
```

```
[hadoop@ip-172-31-16-206 ~]$
```



Proxy mode:

Proxies

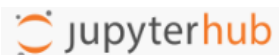
Enabled	Color	Proxy Name	Proxy Notes	Host or IP Address	Port	SOCKS proxy?	SOCKS Version
✓	Blue	localhost:8157		localhost	8157	✓	5
✓	Blue	Default	These are the settings that are used when no patterns match an URL				5

[Import your proxies from FoxyProxy on Mozilla Firefox or from another computer.](#)

[Please Donate](#)

[Buy Proxy Service](#)

Connexion au notebook JupyterHub



Interfaces utilisateur d'application sur le nœud primaire

Cela nécessite l'activation du tunneling SSH. Suivez les instructions de la section [Afficher les interfaces web hébergées sur des clusters Amazon EMR](#).

Application	URL de l'interface utilisateur
Gestionnaire de ressources	http://ec2-13-38-114-127.eu-west-3.compute.amazonaws.com:8088/
Interface utilisateur Spark	http://ec2-13-38-114-127.eu-west-3.compute.amazonaws.com:18080/
JupyterHub	https://ec2-13-38-114-127.eu-west-3.compute.amazonaws.com:9443/
Nom du nœud HDFS	http://ec2-13-38-114-127.eu-west-3.compute.amazonaws.com:9870/

Sign in

Username:

Password:

Sign in

Session Spark

In [1]: `# L'exécution de cette cellule démarre l'application Spark`

Starting Spark application

ID	YARN Application ID	Kind	State	Spark UI	Driver log	User	Current session?
0	application_1680809092392_0001	pyspark	idle	Link	Link	None	✓

SparkSession available as 'spark'.

Définition des PATH :
PATH: s3://p8-ms
PATH_Data: s3://p8-ms/Test
PATH_Result: s3://p8-ms/Results

Chargement des images

Les images sont
chargées au
format binaire

path	modificationTime	length	content
s3://p8-ms/Test/P...	2023-04-04 09:03:12	6574	[FF D8 FF E0 00 1...
s3://p8-ms/Test/P...	2023-04-04 09:03:09	6573	[FF D8 FF E0 00 1...
s3://p8-ms/Test/P...	2023-04-04 09:03:11	6573	[FF D8 FF E0 00 1...
s3://p8-ms/Test/P...	2023-04-04 09:02:57	6572	[FF D8 FF E0 00 1...
s3://p8-ms/Test/P...	2023-04-04 09:03:06	6572	[FF D8 FF E0 00 1...

Chargement du modèle MobilenetV2 en retirant l'avant dernière couche

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

```
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),...
```

```
Downloading data from https://storage.googleapis.com/tensorflow/keras-applications/mobilenet\_v2/mobilenet\_v2\_weights\_tf\_dim\_ordering\_tf\_kernels\_1.0\_224.h5
```

```
14536120/14536120 [=====] - 1s 0us/step
```

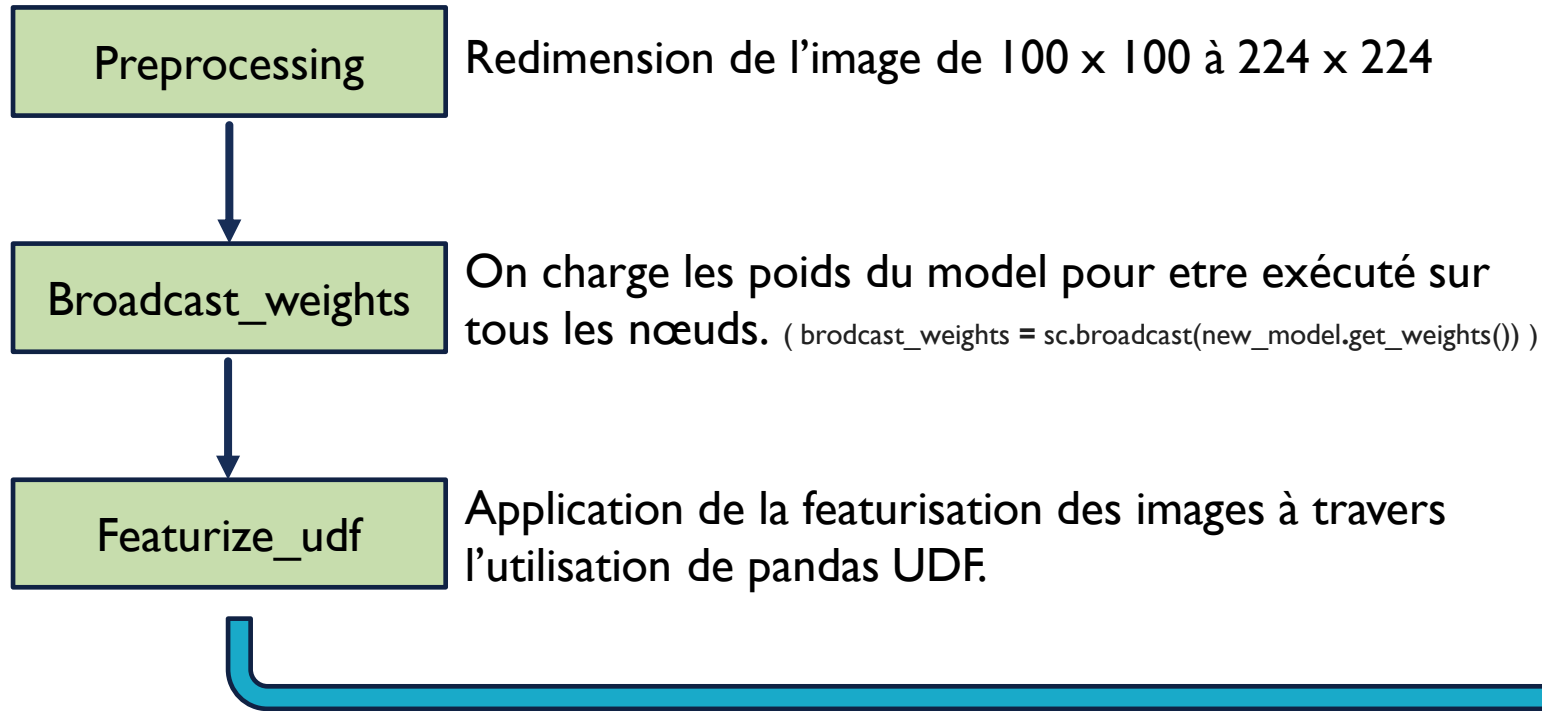
```
new_model = Model(inputs=model.input,  
                  outputs=model.layers[-2].output)
```

```
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),...
```

en entrée : l'entrée du modèle MobileNetV2

en sortie : l'avant dernière couche du modèle MobileNetV2

Traitement des images et extraction des Features



label	count
Orange	160
Pineapple	166
Apple Golden 1	160
Kaki	166
Banana	166
Lemon	164
Corn	150

path	label	features
s3://p8-ms/Test/P...	Pineapple	[0.0, 3.632566, 0...
s3://p8-ms/Test/P...	Pineapple	[0.0, 4.1071906, ...
s3://p8-ms/Test/P...	Pineapple	[0.0, 3.6676247, ...
s3://p8-ms/Test/A...	Apple Golden 1	[0.0, 0.053870715...
s3://p8-ms/Test/P...	Pineapple	[0.0039413804, 2....

Vectorisation des features

```
# Apply the 'array_to_vector' function to the 'features_vectorized' column  
features_df = features_df.withColumn("features_vectorized", array_to_vector("features"))
```

Normalisation des features

```
scaler = StandardScaler(  
    inputCol = 'features_vectorized',  
    outputCol = 'scaledFeatures',  
    withMean = True,  
    withStd = True  
).fit(features_df)  
  
# when we transform the dataframe, the old  
# feature will still remain in it  
df_scaled = scaler.transform(features_df)  
df_scaled.show(6)
```

Application de l'ACP

```
n_components = 100
pca = PCA(
    k = n_components,
    inputCol = 'scaledFeatures',
    outputCol = 'pcaFeatures'
).fit(df_scaled)

df_pca = pca.transform(df_scaled)
df_pca.show(6)
```

path	label	pcaFeatures
s3://p8-ms/Test/P...	Pineapple	[4.24952868426703...
s3://p8-ms/Test/P...	Pineapple	[4.78663711046702...
s3://p8-ms/Test/P...	Pineapple	[3.92177294440557...
s3://p8-ms/Test/A...	Apple Golden 1	[7.16397370645045...
s3://p8-ms/Test/P...	Pineapple	[6.37464657896214...

```
print('Variance expliquée:', pca.explainedVariance.sum())
```

FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),...

Variance expliquée: 0.906408857244218

Enregistrement des Resultats

```
# au format parquet
df_pca.write.mode("overwrite").parquet(PATH_Result)
```

```
df = df_pca.select("*").toPandas()
```

```
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),...
```

```
df.head()
```

```
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),...
```

	path	...	pcaFeatures
0	s3://p8-ms/Test/Pineapple/209_100.jpg	...	[4.249528684267031, 24.29155204063384, -16.101...
1	s3://p8-ms/Test/Pineapple/24_100.jpg	...	[4.786637110467027, 23.88517594245971, -15.595...
2	s3://p8-ms/Test/Pineapple/197_100.jpg	...	[3.92177294444055713, 23.539003012275177, -16.1...
3	s3://p8-ms/Test/Apple Golden 1/130_100.jpg	...	[7.163973706450456, 4.234909962497389, 10.7138...
4	s3://p8-ms/Test/Pineapple/r_81_100.jpg	...	[6.374646578962141, 18.568245346159973, -11.51...

```
[5 rows x 3 columns]
```

```
# au format csv
df.to_csv("{}results.csv".format(PATH_Result), index=False)
```

Résultats sur le bucket S3

Amazon S3

Compartiments

Points d'accès

Points d'accès de l'objet Lambda

Points d'accès multi-région

Opérations par lot

IAM Access Analyzer pour S3

Paramètres de blocage de l'accès public pour ce compte

Storage Lens

Tableaux de bord

Paramètres AWS Organizations

Fonctionnalité spot 3

Les objets sont les entités fondamentales stockées dans Amazon S3. Vous pouvez utiliser l'inventaire Amazon S3 pour obtenir une liste de tous les objets de votre compartiment. Pour que d'autres personnes puissent accéder à vos objets, vous devez leur accorder explicitement des autorisations. En savoir plus

Copier l'URI S3

Copier l'URL

Télécharger

Ouvrir

Supprimer

Actions

Créer un dossier

Charger

Rechercher des objets en fonction du préfixe

Nom

Type

Dernière modification

Taille

Classe de stockage

part-00018-d92582ba-a289-4144-9ea7-103ba385385c-c000.snappy.parquet

parquet

06 Apr 2023 09:47:16 PM CEST

37.0 Ko

Standard

part-00019-d92582ba-a289-4144-9ea7-103ba385385c-c000.snappy.parquet

parquet

06 Apr 2023 09:47:18 PM CEST

37.0 Ko

Standard

part-00020-d92582ba-a289-4144-9ea7-103ba385385c-c000.snappy.parquet

parquet

06 Apr 2023 09:47:20 PM CEST

40.1 Ko

Standard

part-00021-d92582ba-a289-4144-9ea7-103ba385385c-c000.snappy.parquet

parquet

06 Apr 2023 09:47:21 PM CEST

40.1 Ko

Standard

part-00022-d92582ba-a289-4144-9ea7-103ba385385c-c000.snappy.parquet

parquet

06 Apr 2023 09:47:23 PM CEST

39.3 Ko

Standard

part-00023-d92582ba-a289-4144-9ea7-103ba385385c-c000.snappy.parquet

parquet

06 Apr 2023 09:47:25 PM CEST

36.9 Ko

Standard

results.csv

csv

06 Apr 2023 09:48:47 PM CEST

2.3 Mo

Standard

Historique Spark

sur le serveur

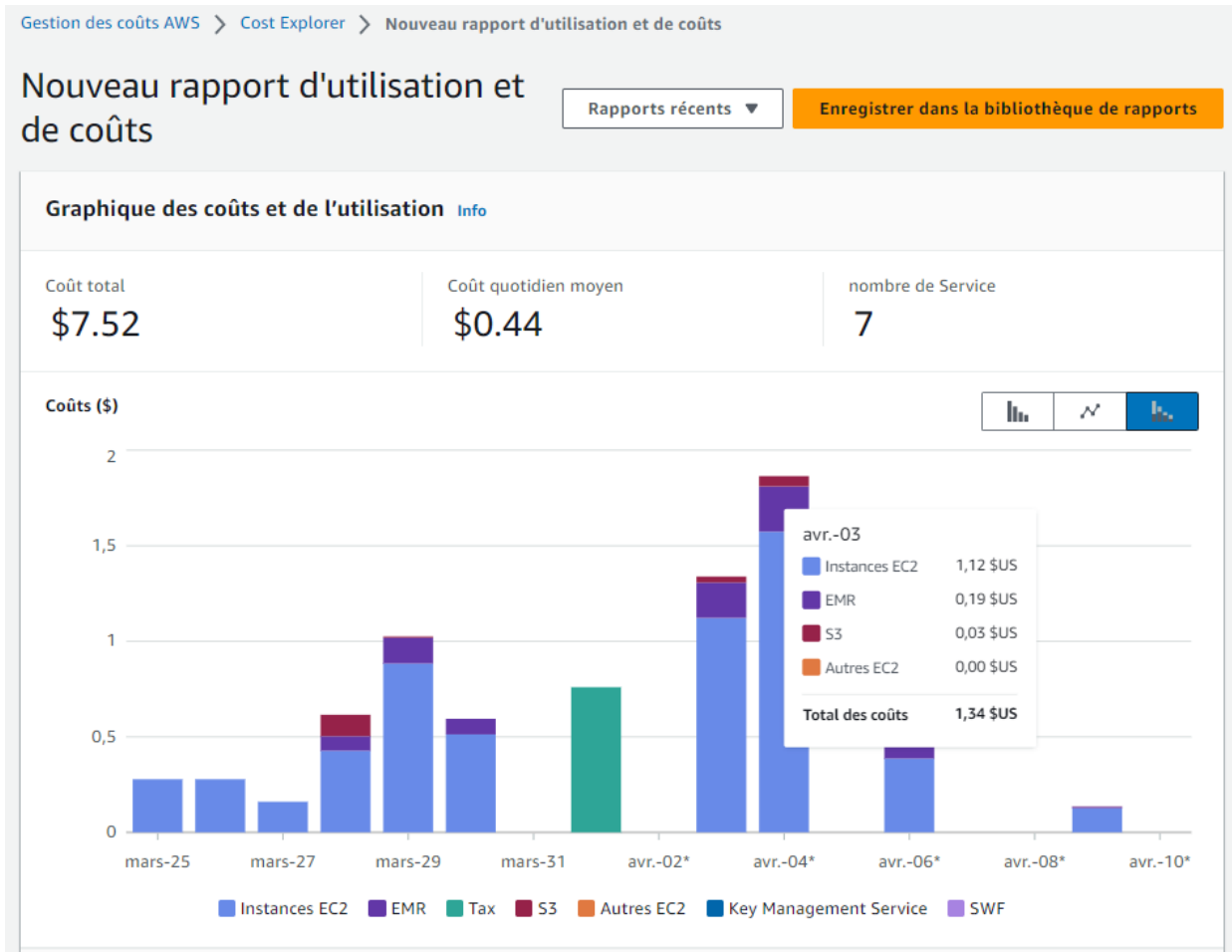
Completed Jobs (38)

Page: 1

1 Pages. Jump to 1. Show 100 items in a page. Go

Job Id (Job Group) ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
37 (40)	Job group for statement 40 toPandas at <stdin>:1	2023/04/06 19:47:56	45 s	1/1 (1 skipped)	24/24 (36 skipped)
36 (40)	Job group for statement 40 toPandas at <stdin>:1	2023/04/06 19:47:28	28 s	1/1	36/36
35 (38)	Job group for statement 38 parquet at NativeMethodAccessorImpl.java:0	2023/04/06 19:46:35	49 s	1/1 (1 skipped)	24/24 (36 skipped)
34 (38)	Job group for statement 38 parquet at NativeMethodAccessorImpl.java:0	2023/04/06 19:46:06	29 s	1/1	36/36
33 (35)	Job group for statement 35 showString at NativeMethodAccessorImpl.java:0	2023/04/06 19:45:47	3 s	1/1 (1 skipped)	1/1 (36 skipped)
32 (35)	Job group for statement 35 showString at NativeMethodAccessorImpl.java:0	2023/04/06 19:45:18	30 s	1/1	36/36
31 (31)	Job group for statement 31 collect at <stdin>:1	2023/04/06 19:44:12	43 s	1/1 (1 skipped)	24/24 (36 skipped)
30 (30)	Job group for statement 30 collect at <stdin>:1	2023/04/06 19:43:22	46 s	1/1 (1 skipped)	24/24 (36 skipped)
29 (30)	Job group for statement 30 javaToPython at NativeMethodAccessorImpl.java:0	2023/04/06 19:42:58	24 s	1/1	36/36

Evolution du cout





En résumé, nous avons déployé sur le cloud une solution qui permet avec notre script PySpark d'extraire les caractéristiques de chaque image avec un algorithme de transfert learning, d'effectuer une réduction de dimension (ACP) sur ces caractéristiques et leur sauvegarde pour une utilisation ultérieure.