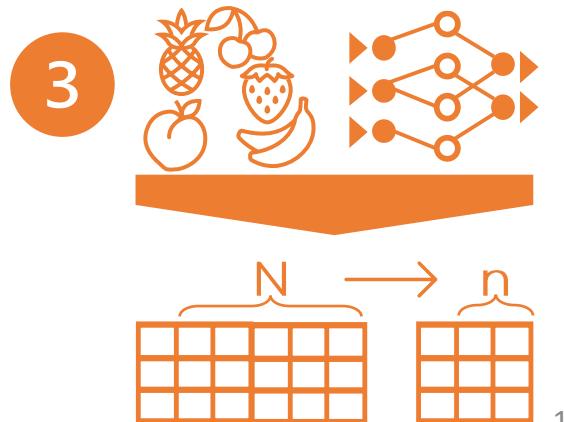
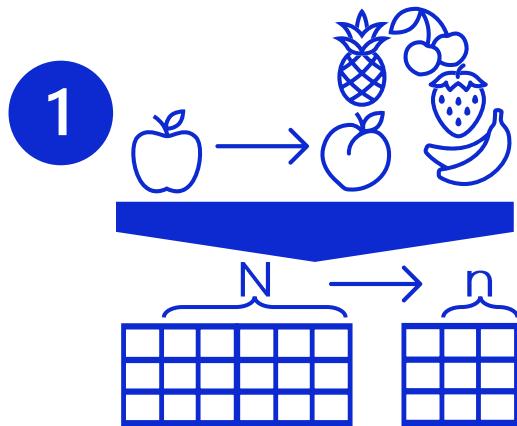
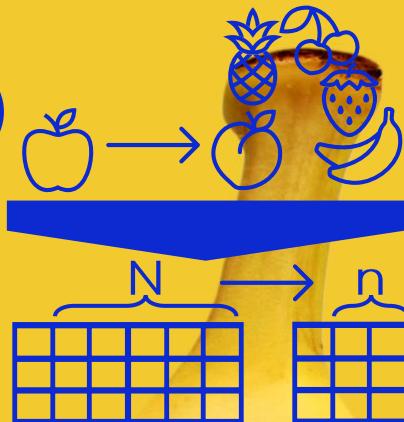


Déployez un modèle dans le cloud

Sébastien De Rosa | OpenClassrooms | Data Scientist | Projet 8



1



Analyse du besoin

Acquisition des données
Visualisation des images
Besoin réduction dimension
Solution envisagée



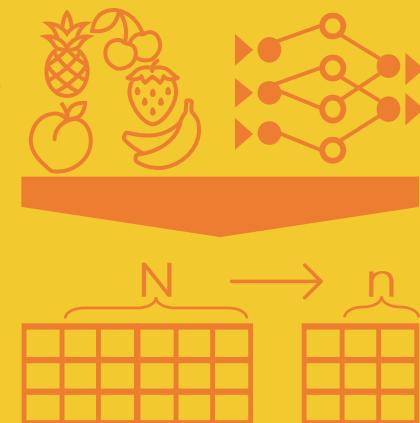
2



Big Data | Rôle et création

Schéma général
Solution stockage AWS S3
Instance AWS EC2 Linux
Hadoop vs Spark
Session PySpark

3



Chaine de Traitement

Modèle VGG16
Preprocessing
Fonctions UDF Pandas
Sauvegarde dans s3 / Check
Remarques & préconisations

1

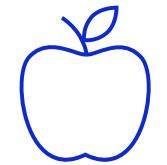
Analyse du besoin



Images cible :

<https://www.kaggle.com/datasets/moltean/fruits>

Mihai Oltean, Fruits 360 dataset: new research directions, Technical report, 2021



Acquisition des données :

12 455 fichiers jpg

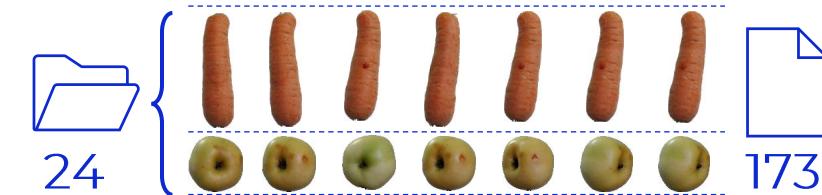
558 Mo total



3 répertoires :

- Test / 3 110
- Training / 6 231
- Validation / 3 114

fruits-360-original-size / Test :



24

173

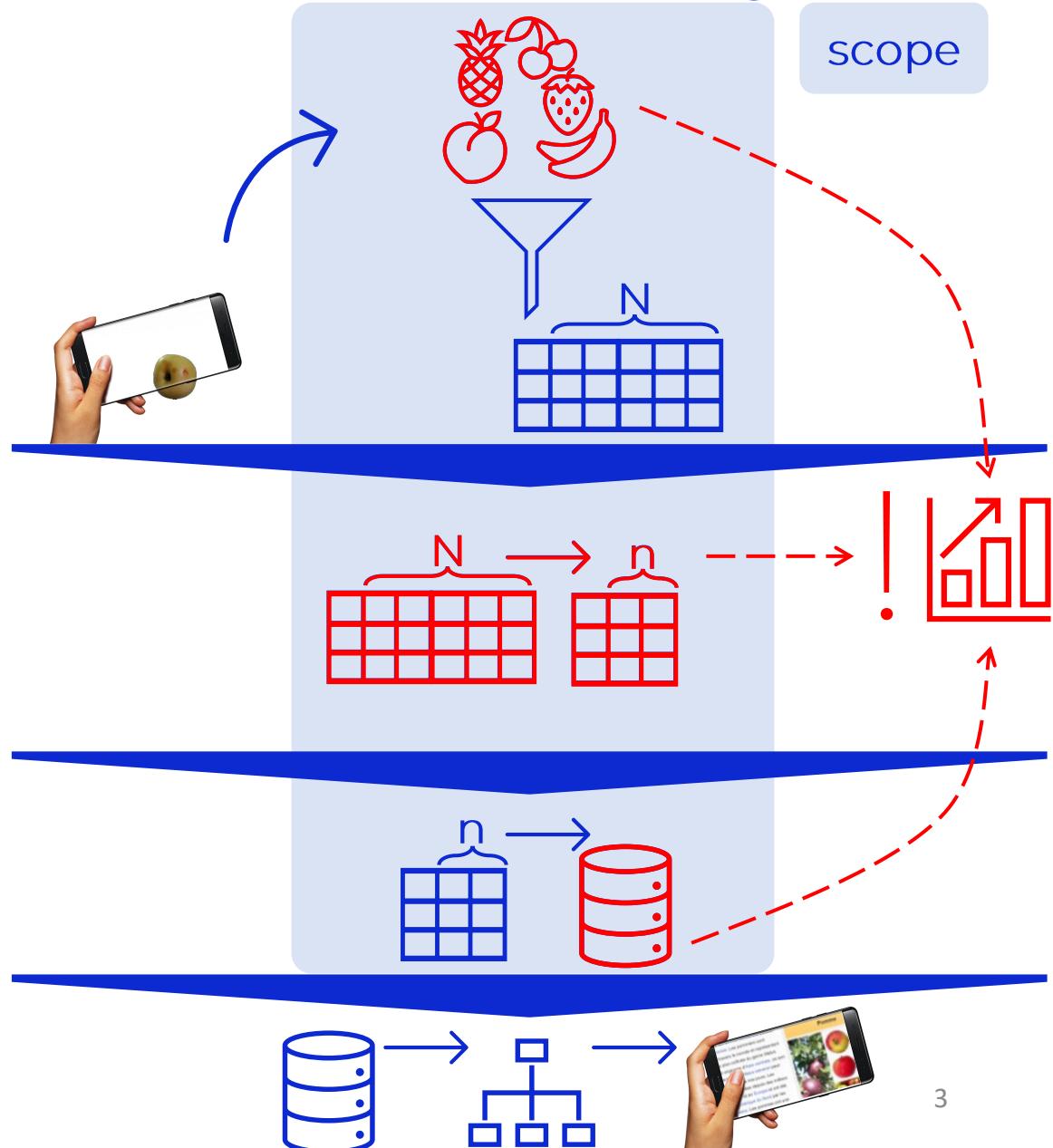


Nom	Dimensions	Date
r0_30.jpg	217 x 649	12/09/2021 20:24



217 x 649 x 3
422 499 features

Solution envisagée



Volume



Vitesse



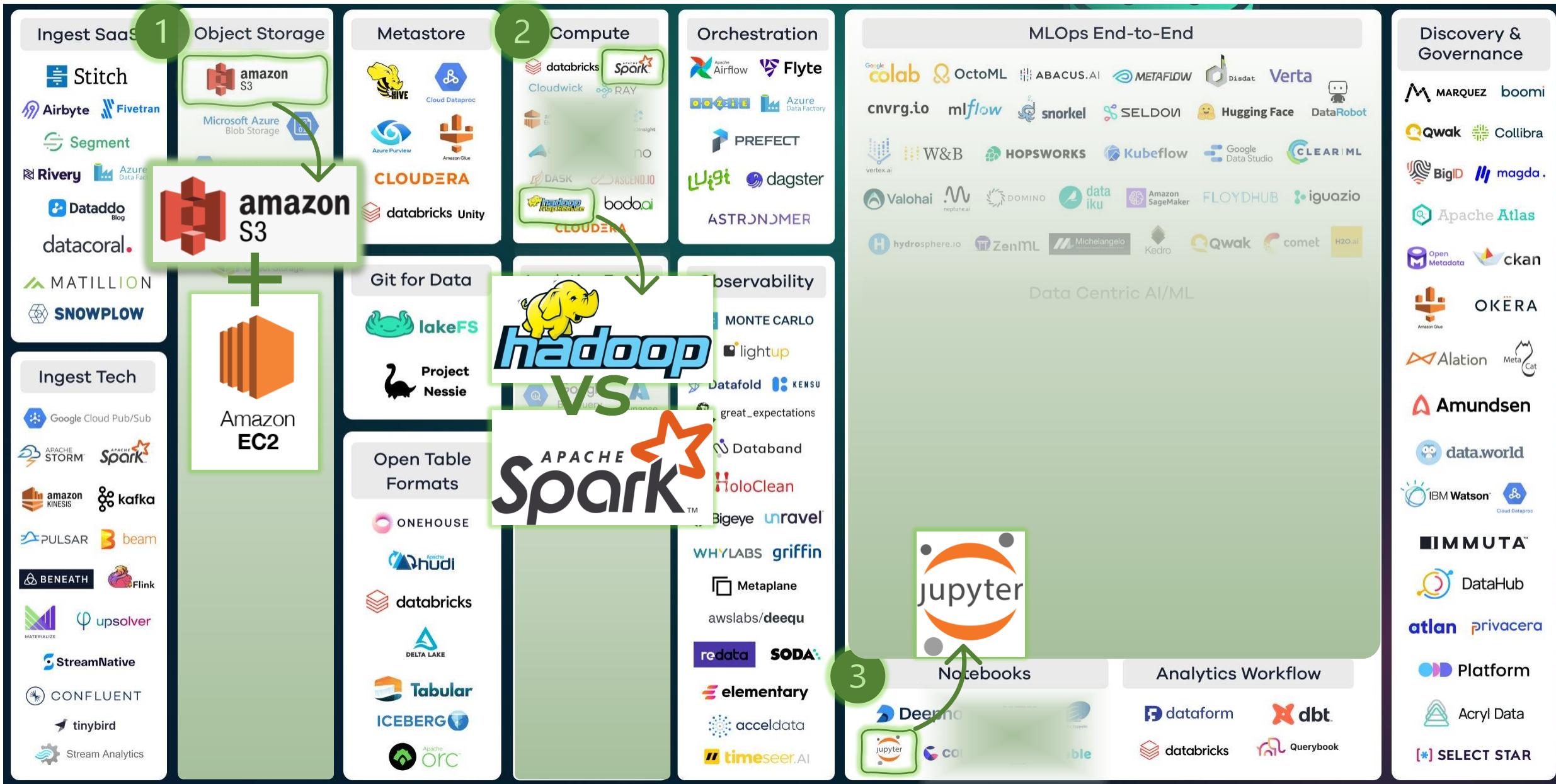
Variété



2

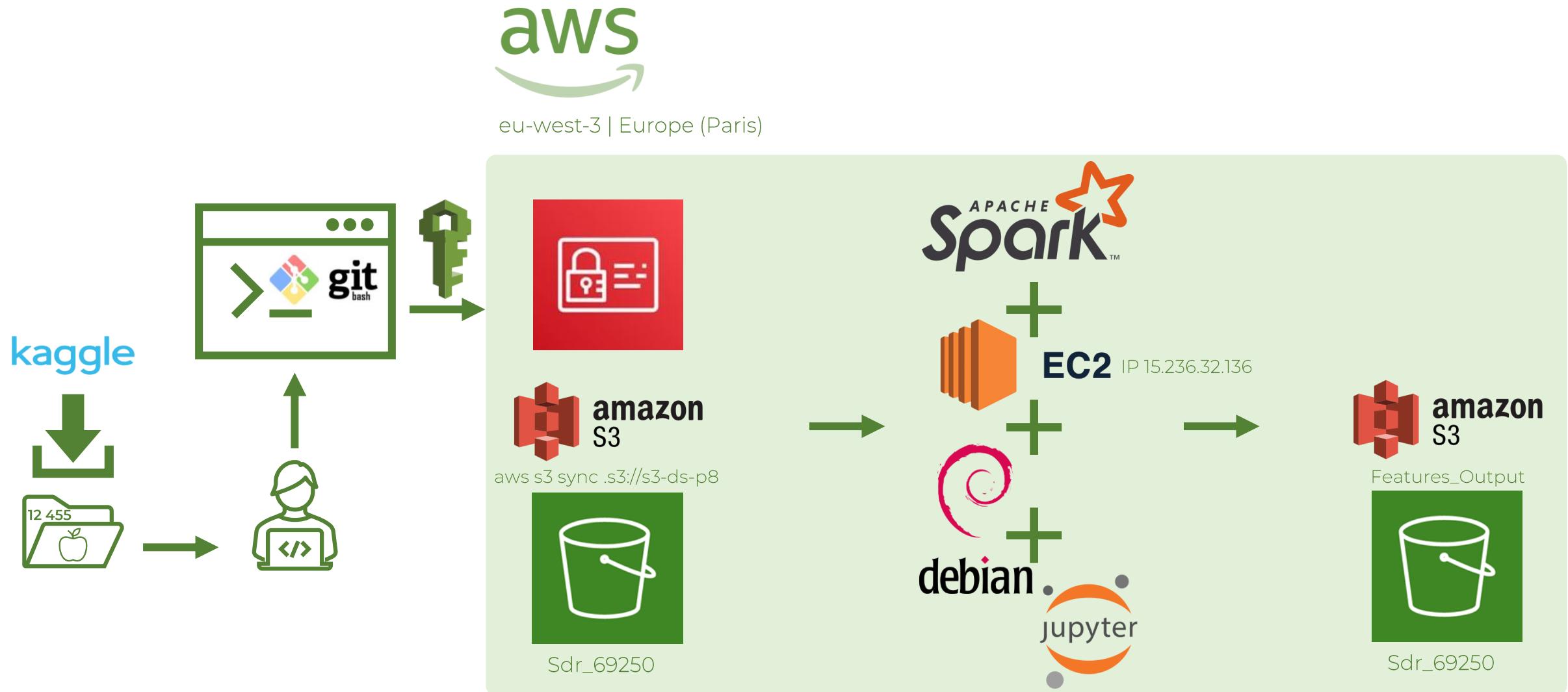
Big Data | Rôle et création

Schéma général



2 Big Data | Rôle et création

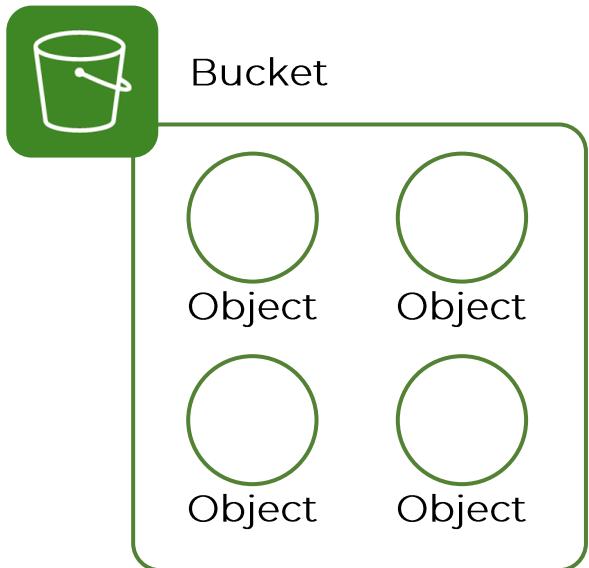
Schéma général



2

Big Data | Rôle et création

Solution stockage AWS S3



Calculer la taille totale

Résumé

Source	Nombre total d'objets	Taille totale
s3://n3-ds-p8	12 473	2.2 Go

Objets spécifiés

Nom	Type	Dernière modification	Taille	Nombre total d'objets	Erreur
Features_Output/	Dossier	-	1.6 Go	18	-
Test/	Dossier	-	139.6 Mo	3110	-
Training/	Dossier	-	279.0 Mo	6231	-
Validated/	Dossier	-	139.2 Mo	3114	-

Classes de stockage

S3 Standard	Hiérarchisation intelligente S3*	S3 Standard - IA	S3 One Zone-IA†	S3 Glacier Instant Retrieval	S3 Glacier Flexible Retrieval	S3 Glacier Deep Archive
Conçu pour la durabilité	99,999999999 % (11 9s)	99,999999999 % (11 9s)	99,999999999 % (11 9s)	99,999999999 % (11 9s)	99,999999999 % (11 9s)	99,999999999 % (11 9s)
Conçu pour la disponibilité	99,99 %	99,9 %	99,9 %	99,5 %	99,9 %	99,99 %
Disponibilité SLA	99,9 %	99 %	99 %	99 %	99 %	99,9 %
Zones de disponibilité	≥3	≥3	≥3	1	≥3	≥3
Frais de capacité minimale par objet	N/A	N/A	128 Ko	128 Ko	128 Ko	40 Ko
Frais minimum de durée de stockage	N/A	N/A	30 jours	30 jours	90 jours	90 jours
Frais d'extraction	N/A	N/A	par Go extrait	par Go extrait	par Go extrait	par Go extrait
Latence du premier octet	millisecondes	millisecondes	millisecondes	millisecondes	minutes ou heures	heures
Type de stockage	Objet	Objet	Objet	Objet	Objet	Objet
Transitions du cycle de vie	Oui	Oui	Oui	Oui	Oui	Oui

Fréquence consultation

Coût

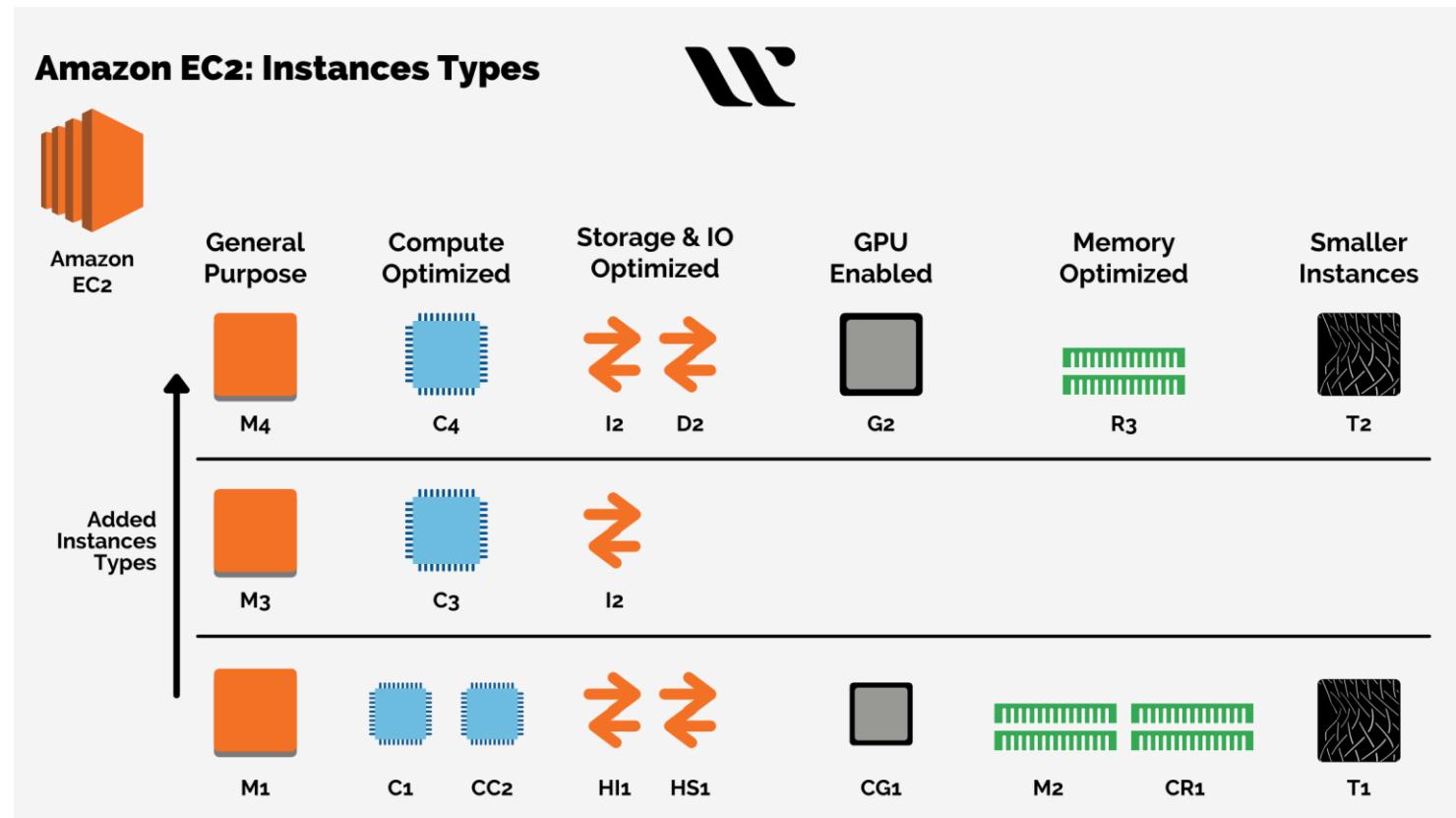
Accès sécurisé à s3 :

```
# identifiants nécessaires pour que PySpark accède à AWS  
AWS_CREDENTIALS = json.loads(os.popen("aws sts get-session-token").read())['Credentials']
```

Python

```
# Necessary ressources SparkConf to run PySPark on AWS  
# PySPark was coded in Scala itself coded in Java  
# To run on AWS, we need to install AWS packages  
# For the access we ask for AWS Credential  
conf.set('spark.jars.packages', 'org.apache.hadoop:hadoop-aws:3.2.2')  
conf.set('spark.hadoop.fs.s3a.aws.credentials.provider', 'org.apache.hadoop.fs.s3a.  
TemporaryAWSCredentialsProvider')
```

```
# Credentials passed to the SparkConf to get access to S3 Bucket  
conf.set('spark.hadoop.fs.s3a.access.key', AWS_CREDENTIALS['AccessKeyId'])  
conf.set('spark.hadoop.fs.s3a.secret.key', AWS_CREDENTIALS['SecretAccessKey'])  
conf.set('spark.hadoop.fs.s3a.session.token', AWS_CREDENTIALS['SessionToken'])
```

**Offres tarifaires**

- Offre gratuite
- Instances Spot
- Saving Plans
- Instances Réservées
- A la demande
- Hôtes dédiés



Facturation à la seconde



Microsoft

ubuntu

debian

SUSE

aws
Amazon Linux

Résumé de l'instance pour i-049925baedcaa13f4 (SDREC2_P8) Informations

ID d'instance: i-049925baedcaa13f4 (SDREC2_P8)

Adresse IPv4 publique: 15.236.32.136 | adresse ouverte

État de l'instance: Arrêté(e)

Nom DNS de l'IP privé (IPv4 uniquement): ip-172-31-36-181.eu-west-3.compute.internal

Type d'instance: t2.large

ID de VPC: vpc-00bce038c22dbc96a

Adresses IPv4 privées: 172.31.36.181

DNS IPv4 public: ec2-15-236-32-136.eu-west-3.compute.amazonaws.com | adresse ouverte

Adresses IP élastiques: 15.236.32.136 [IP publique]

Rôle IAM: -

ID de sous-réseau: subnet-0cf2ea59b1cbe3e50

Nom du groupe Auto Scaling: -

Détails | Sécurité | Mise en réseau | Stockage | Vérifications de statut | Surveillance | Balises

Détails de l'instance Informations

Plateforme: Debian (déduit)

Informations sur la plateforme: Linux/UNIX

Protection contre l'arrêt: Désactivé

Récupération automatique de l'instance: Par défaut

Index de lancement de l'AMI:

ID AMI: ami-002ff2c881c910aa8

Nom de l'AMI: debian-11-amd64-20220503-998

Heure de lancement: Mon Nov 28 2022 11:54:29 GMT+0100 (heure normale d'Europe centrale) (1 day)

Cycle de vie: normal

Nom de la paire de clés:

Surveillance: désactivé

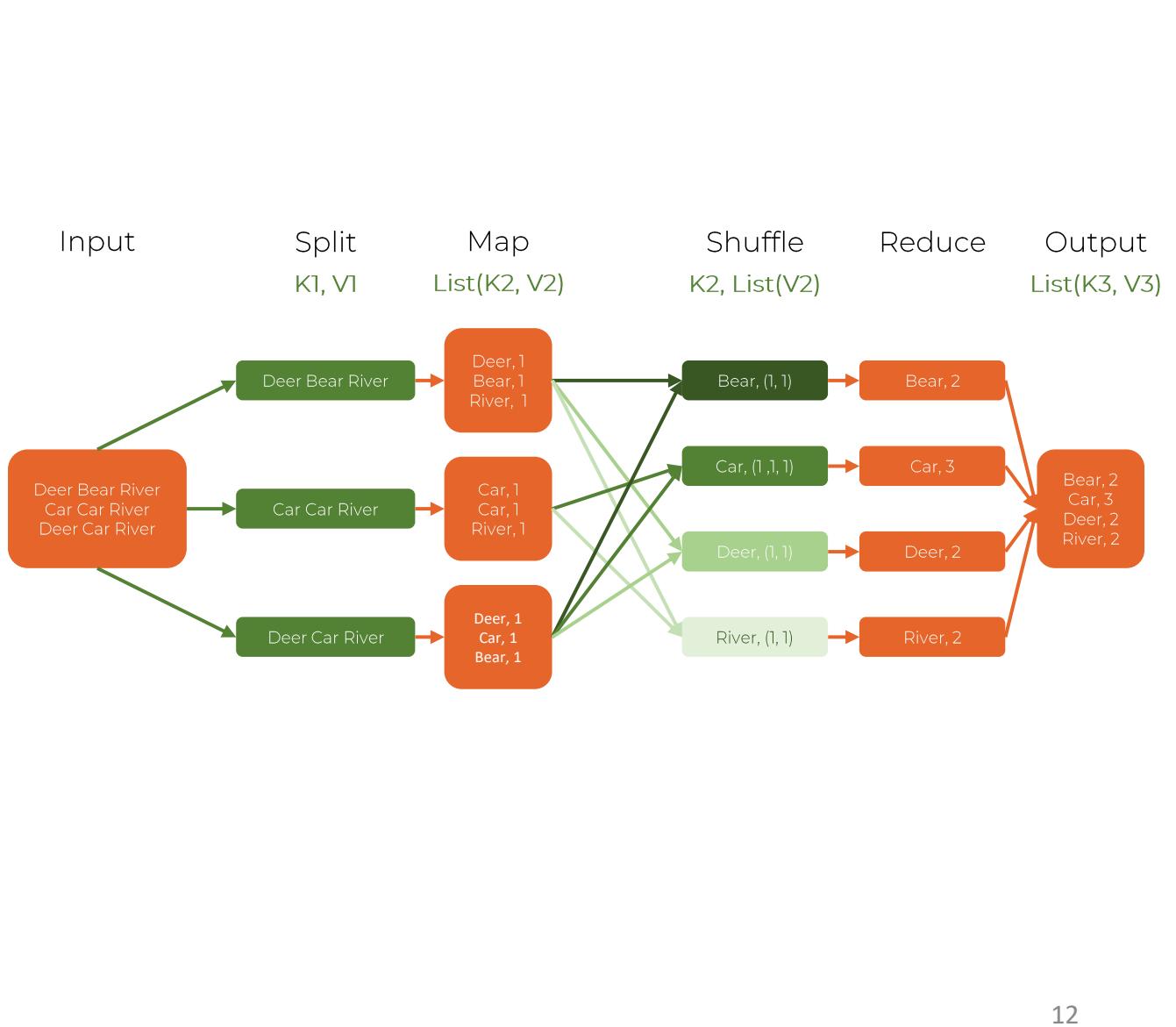
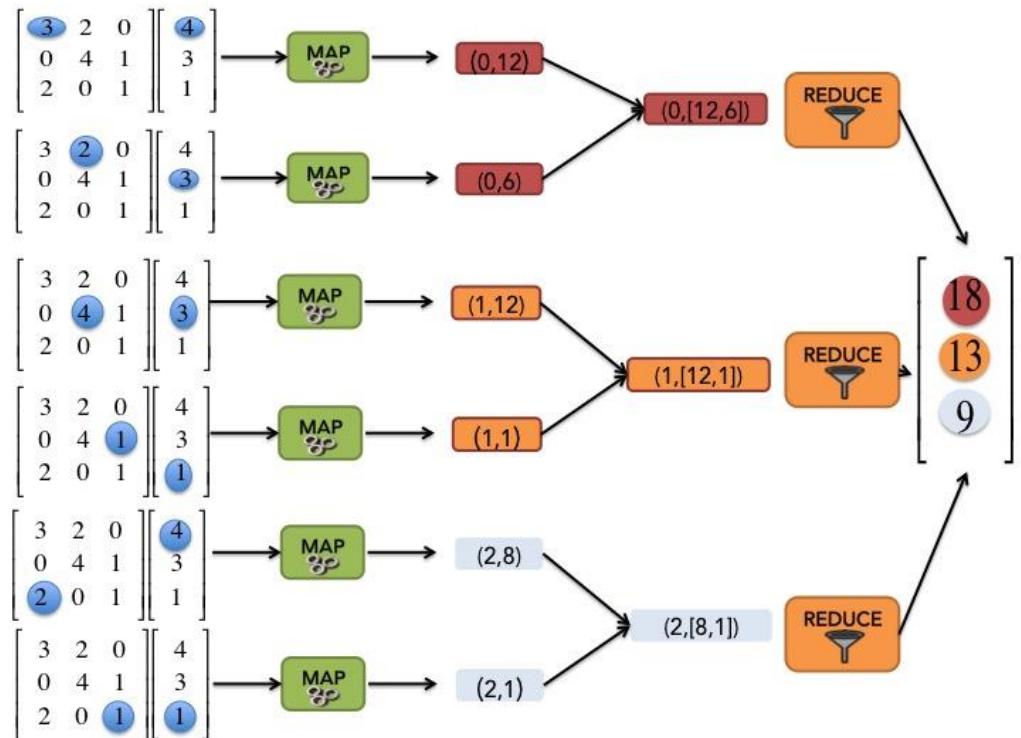
Protection de la résiliation: Désactivé

Emplacement de l'AMI: amazon/debian-11-amd64-20220503-998

Comportement Arrêt - Mise en veille prolongée: désactivé

Motif de transition de l'état:

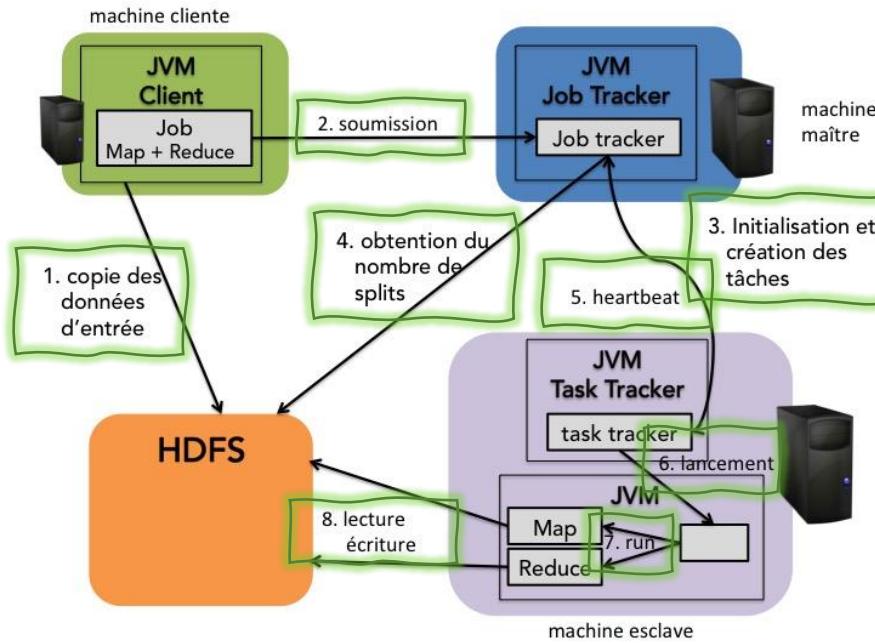
Map Reduce



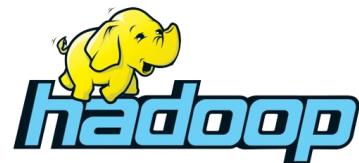
2 Big Data | Rôle et création

Hadoop vs Spark

Map Reduce Hadoop 1.x



Map Reduce + HDFS

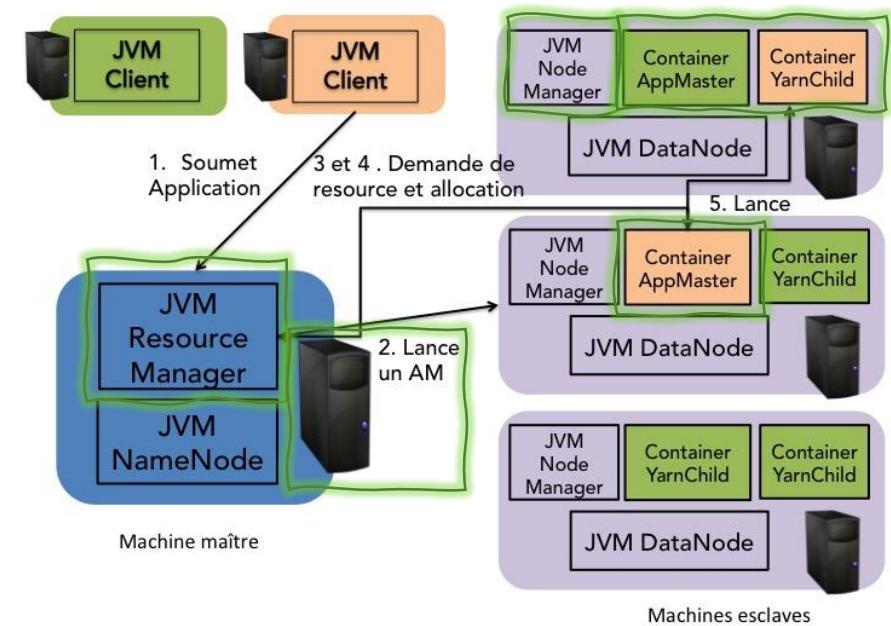


Transformer
algorithme

Problèmes
complexes
nécessitent
d'enchaîner
Map/Reduce

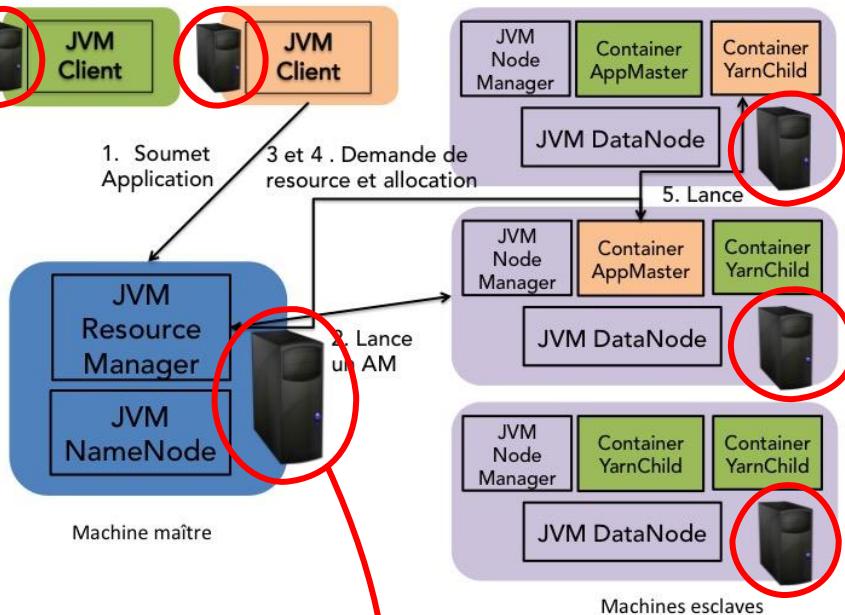
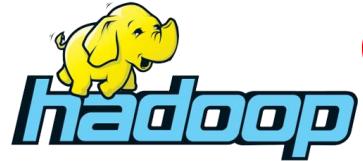
- Job Tracker:
- ressources cluster
 - ordonner jobs

Map Reduce Hadoop 2.x



Map Reduce + HDFS + YARN

2 Big Data | Rôle et création

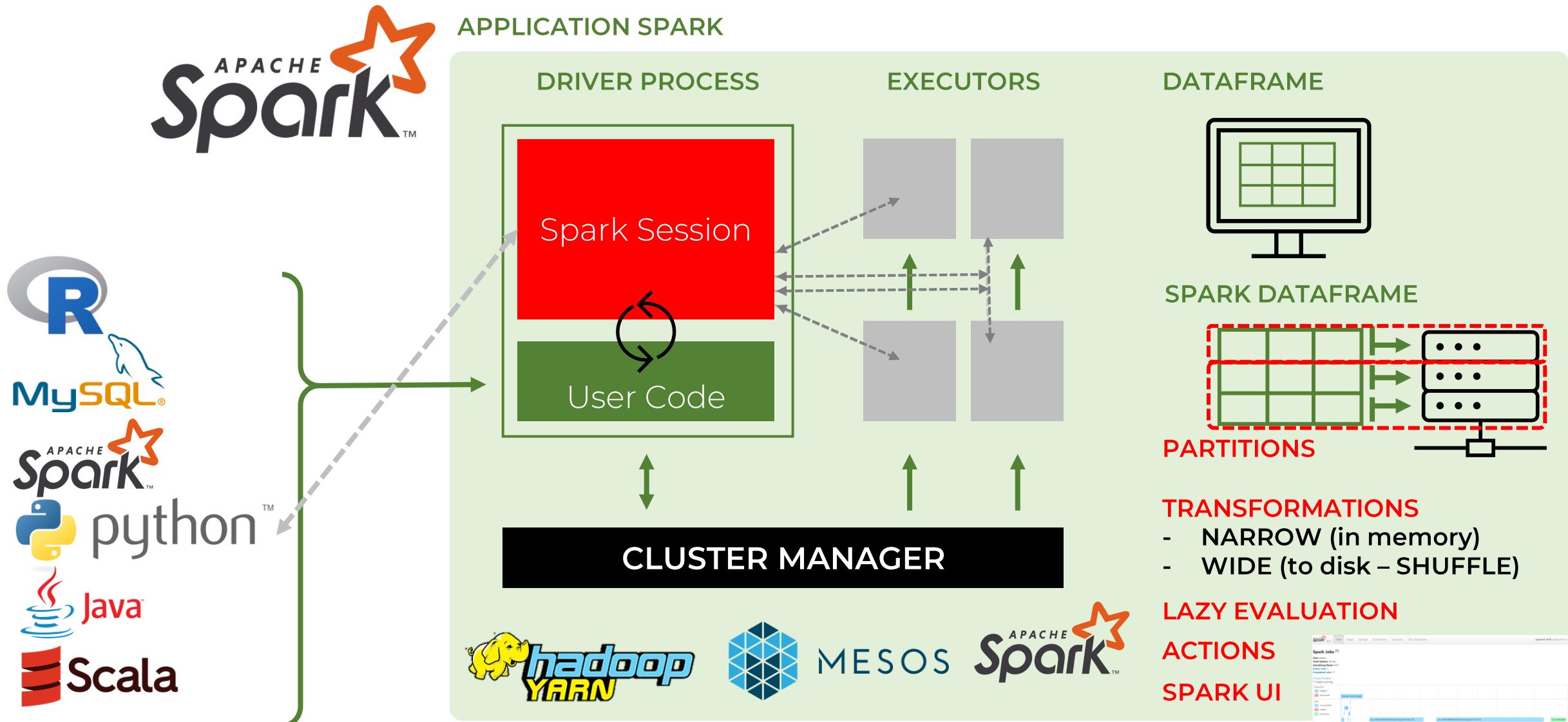


Technologie	Latence (s)	Taux de transfert (Go/s)
Disque dur	10^{-2}	0.15

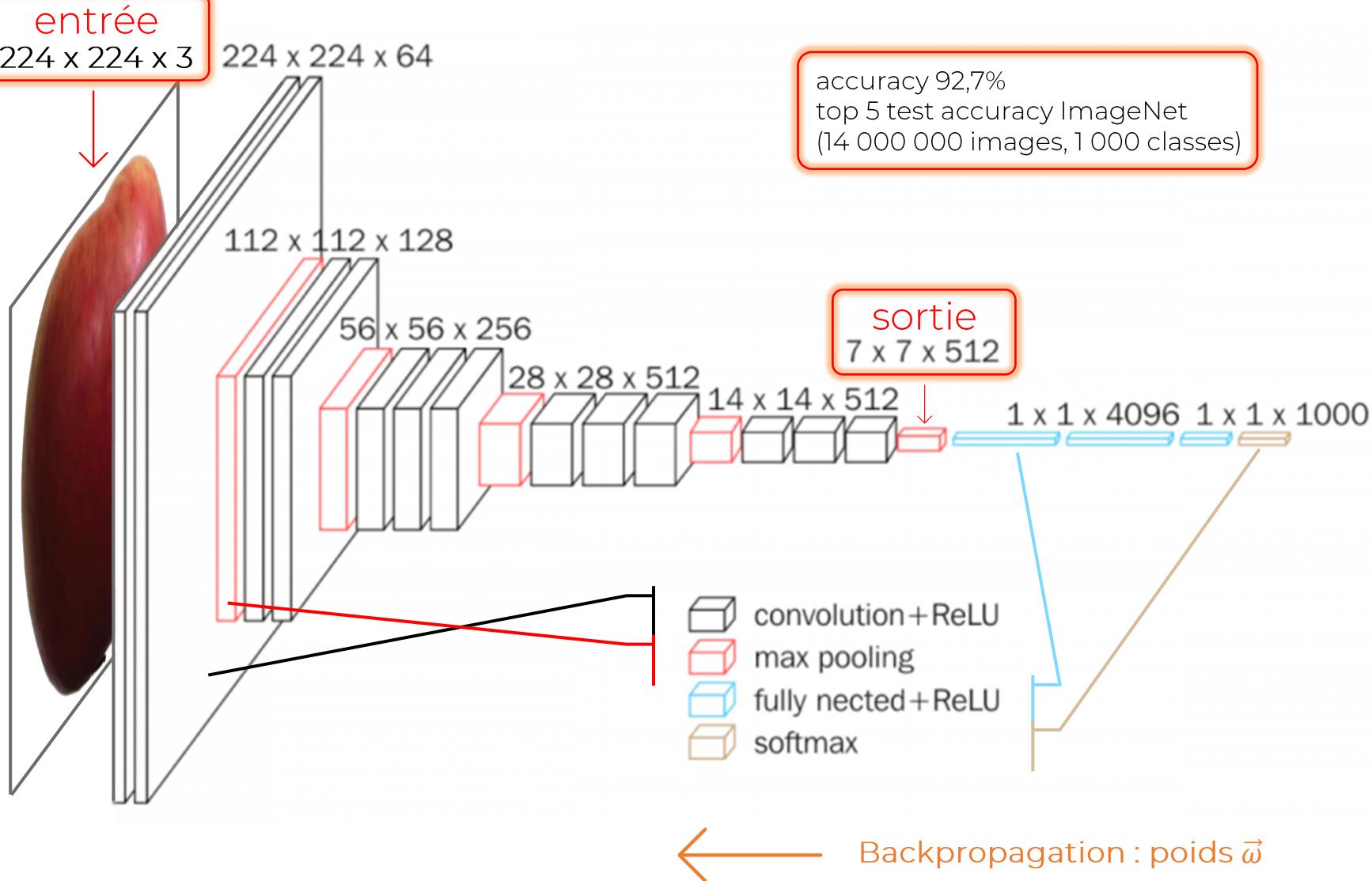


High Level Operators





3 Chaine de Traitement



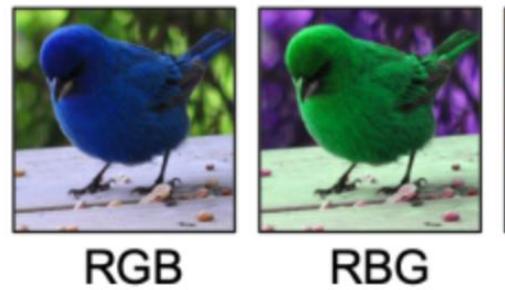
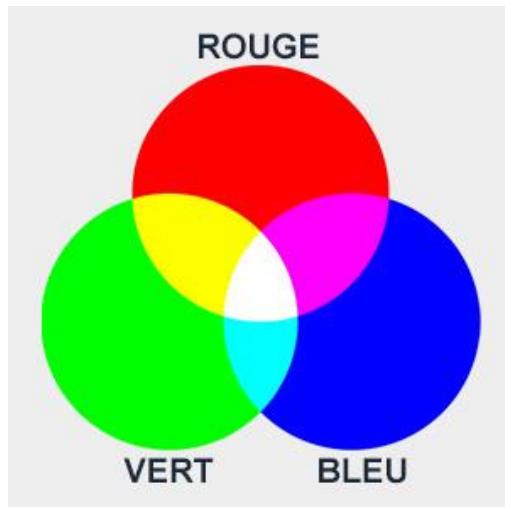
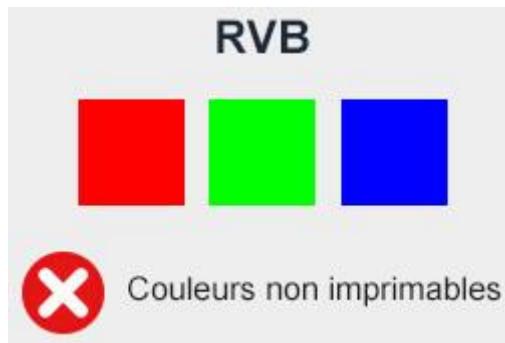
Modèle VGG16

```
# Change input shape dimensions for fine-tuning with Keras
# remove last layer (include_top=False) to get a Tensor
model = VGG16(weights="imagenet", include_top=False)
model.trainable = False
# model = Model(inputs = model.input, outputs = model.output)
model.summary() # verify that the top layer is removed
```

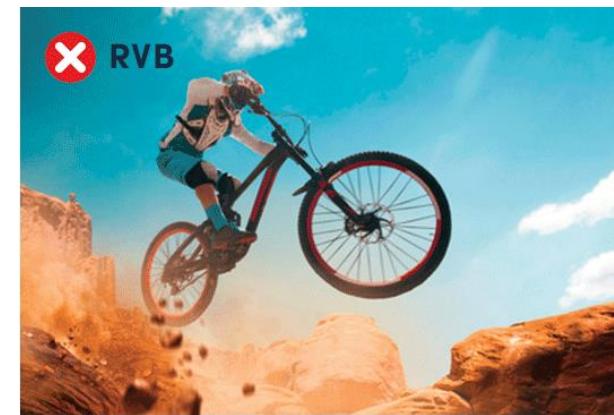
Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[None, None, None, 3]	0
block1_conv1 (Conv2D)	(None, None, None, 64)	1792
block1_conv2 (Conv2D)	(None, None, None, 64)	36928
block1_pool (MaxPooling2D)	(None, None, None, 64)	0
block2_conv1 (Conv2D)	(None, None, None, 128)	73856
block2_conv2 (Conv2D)	(None, None, None, 128)	147584
block2_pool (MaxPooling2D)	(None, None, None, 128)	0
block3_conv1 (Conv2D)	(None, None, None, 256)	295168
block3_conv2 (Conv2D)	(None, None, None, 256)	590080
block3_conv3 (Conv2D)	(None, None, None, 256)	590080
block3_pool (MaxPooling2D)	(None, None, None, 256)	0
block4_conv1 (Conv2D)	(None, None, None, 512)	1180160
block4_conv2 (Conv2D)	(None, None, None, 512)	2359808
block4_conv3 (Conv2D)	(None, None, None, 512)	2359808
block4_pool (MaxPooling2D)	(None, None, None, 512)	0
block5_conv1 (Conv2D)	(None, None, None, 512)	2359808
block5_conv2 (Conv2D)	(None, None, None, 512)	2359808
block5_conv3 (Conv2D)	(None, None, None, 512)	2359808
block5_pool (MaxPooling2D)	(None, None, None, 512)	0

```
Total params: 14,714,688
Trainable params: 0
Non-trainable params: 14,714,688
```

3 Chaine de Traitement



sources : <https://www.easyflyer.fr/l/explanations-rvb-rgb/>, <https://datascience.stackexchange.com/questions/65260/how-important-is-the-channel-order-in-deep-learning-computer-vision-tasks>



$$3! = 6$$

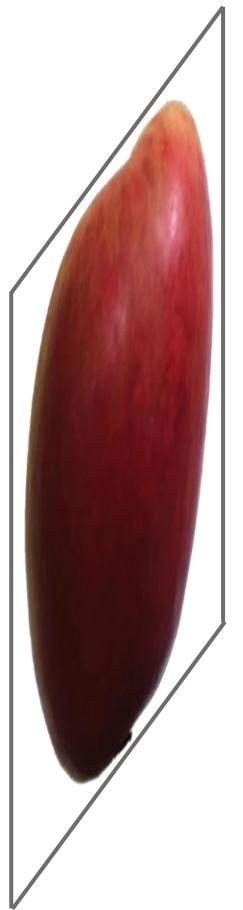
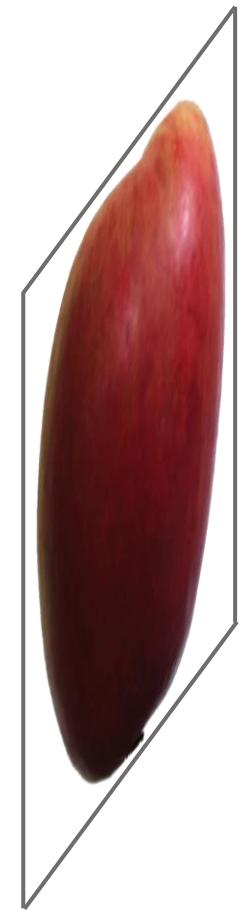


Preprocessing

- ✓ `resize([224, 224])`
- ✓ `vgg16.preprocess_input`

$YYY \times ZZZ \times 3$ | **RGB**

$224 \times 224 \times 3$ | **BGR**



3 Chaine de Traitement

```
@pandas_udf('array<float>', PandasUDFType.SCALAR_ITER)
def featurize_udf(content_series_iter):
```

```
@pandas_udf('array<float>', PandasUDFType.SCALAR_ITER)
def featurize_udf(content_series_iter):
```

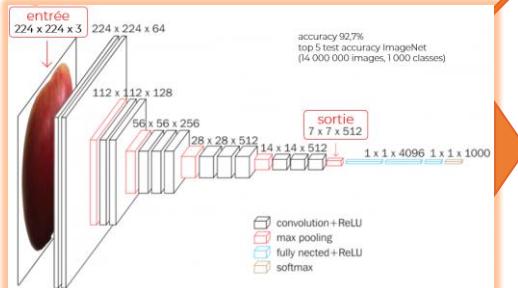
This method is a Scalar Iterator pandas UDF wrapping our featurization function.
The decorator specifies that this returns a Spark DataFrame column of type ArrayType(FloatType).

:param content_series_iter: This argument is an iterator over batches of data, where each batch is a pandas Series of image data.

'''
With Scalar Iterator pandas UDFs, we can load the model once and then re-use it
for multiple data batches. This amortizes the overhead of loading big models.

for content_series in content_series_iter:
 yield featurize_series(model, content_series)

Python



```
def featurize_series(model, content_series):
```

'''
Featurize a pd.Series of raw images using the input model.
:return: a pd.Series of image features

```
"""
input_ = np.stack(content_series.map(preprocess))
feats = model.predict(input_)
# For some layers, output features will be multi-dimensional tensors.
# We flatten the feature tensors to vectors for easier storage in Spark DataFrames.
output = [p.flatten() for p in feats]
return pd.Series(output)
```

Fonctions UDF Pandas

```
# We can now run featurization on our entire Spark DataFrame.
```

```
# NOTE: This can take a long time (about 10 minutes)
# since it applies a large model to the full dataset.
```

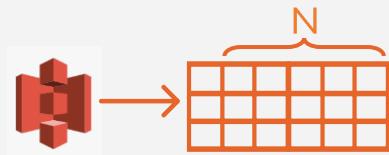
```
# number of workers in parallel = 5
features_test = images_test.repartition(5).select(col("path"), featurize_udf("content").alias("features"))
```

Python



```
# creation of dataframe for Test files
# filter on file extension, recursive file lookup, bucket path
images_test = spark.read.format("binaryFile") \
.option("pathGlobFilter", "*.jpg") \
.option("recursiveFileLookup", "true") \
.load("s3a://s3-ds-p8/Test/*")
```

Python

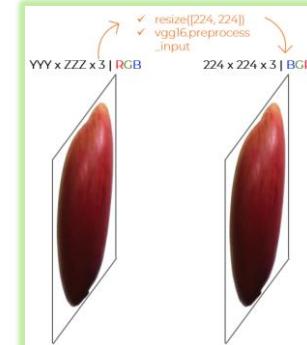


```
# kept the resize command, as preprocess_input() did not take care of it
def preprocess(content):
```

Preprocesses raw image bytes for prediction.

```
"""
img = Image.open(io.BytesIO(content)).resize([224, 224])
arr = img_to_array(img)
return preprocess_input(arr)
```

Python



18

3 Chaine de Traitement

Sauvegarde dans s3 / Check

```
features_test.withColumn("features", col("features").cast("string"))\  
.write.mode('overwrite')\  
.option("header",True)\  
.csv("s3a://s3-ds-p8/Features_Output/Test/")
```

Test



15min 24s

```
features_training.withColumn("features", col("features").cast("string"))\  
.write.mode('overwrite')\  
.option("header",True)\  
.csv("s3a://s3-ds-p8/Features_Output/Training/")
```

Training

30min 28s

```
features_validation.withColumn("features", col("features").cast("string"))\  
.write.mode('overwrite')\  
.option("header",True)\  
.csv("s3a://s3-ds-p8/Features_Output/Validation/")
```

Validation

23min 30s

3

Chaine de Traitement

Sauvegarde dans s3 / Check

Amazon S3 > Compartiments > s3-ds-p8

s3-ds-p8 Info

Objets **Propriétés** **Autorisations** **Métriques** **Gestion**

Objets (4)

Les objets sont les entités fondamentales stockées dans Amazon S3. Vous pouvez utiliser l'invite pour accéder explicitement des autorisations. [En savoir plus](#)

Copier l'URI S3 Copier l'URL Télécharger

Rechercher des objets en fonction du préfixe

<input type="checkbox"/>	Nom	Type
<input type="checkbox"/>	Features_Output/	Dossier
<input type="checkbox"/>	Test/	Dossier
<input type="checkbox"/>	Training/	Dossier
<input type="checkbox"/>	Validation/	Dossier

Amazon S3 > Compartiments > s3-ds-p8 > Features_Output/ > Test/

Test/

Objets **Propriétés**

Objets (6)

Les objets sont les entités fondamentales stockées dans Amazon S3. Vous pouvez utiliser l'invite 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

<input type="checkbox"/>	Nom	Type	Dernière modification	Taille	Classe de stockage
<input type="checkbox"/>	_SUCCESS	-	23 Nov 2022 09:45:53 AM CET	0 o	Standard
<input type="checkbox"/>	part-00001-39ca36d0-bd1a-44f6-a2f8-39b56794527-c000.csv	csv	23 Nov 2022 09:45:46 AM CET	84.4 Mo	Standard
<input type="checkbox"/>	part-00001-39ca36d0-bd1a-44f6-a2f8-39b56794527-r000.csv	csv	23 Nov 2022 09:45:49 AM CET	84.0 Mo	Standard
<input type="checkbox"/>	part-00001-39ca36d0-bd1a-44f6-a2f8-39b56794527-r0000.csv	csv	23 Nov 2022 09:45:48 AM CET	82.6 Mo	Standard
<input type="checkbox"/>	part-00001-39ca36d0-bd1a-44f6-a2f8-39b56794527-r0001.csv	csv	23 Nov 2022 09:45:52 AM CET	81.9 Mo	Standard
<input type="checkbox"/>	part-00004-39ca36d0-bd1a-44f6-a2f8-39b56794527-c000.csv	csv	23 Nov 2022 09:45:51 AM CET	84.1 Mo	Standard

Amazon S3 > Compartiments > s3-ds-p8 > Features_Output/ > Training/

Training/

Objets **Propriétés**

Objets (6)

Les objets sont les entités fondamentales stockées dans Amazon S3. Vous pouvez utiliser l'invite 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

<input type="checkbox"/>	Type	Dernière modification	Taille	Classe de stockage	
<input type="checkbox"/>	_SUCCESS	23 Nov 2022 10:16:22 AM CET	0 o	Standard	
<input type="checkbox"/>	part-00000-9fb0096-fd01-4eeb-b18a-99bf2287f6e-c000.csv	csv	23 Nov 2022 10:16:17 AM CET	168.6 Mo	Standard
<input type="checkbox"/>	part-00001-9fb0096-fd01-4eeb-b18a-99bf2287f6e-c000.csv	csv	23 Nov 2022 10:16:13 AM CET	167.0 Mo	Standard
<input type="checkbox"/>	part-00002-9fb0096-fd01-4eeb-b18a-99bf2287f6e-c000.csv	csv	23 Nov 2022 10:16:15 AM CET	166.5 Mo	Standard
<input type="checkbox"/>	part-00003-9fb0096-fd01-4eeb-b18a-99bf2287f6e-c000.csv	csv	23 Nov 2022 10:16:19 AM CET	165.1 Mo	Standard
<input type="checkbox"/>	part-00004-9fb0096-fd01-4eeb-b18a-99bf2287f6e-c000.csv	csv	23 Nov 2022 10:16:20 AM CET	168.2 Mo	Standard

Amazon S3 > Compartiments > s3-ds-p8 > Features_Output/ > Validation/

Validation/

Objets **Propriétés**

Objets (6)

Les objets sont les entités fondamentales stockées dans Amazon S3. Vous pouvez utiliser l'invite 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

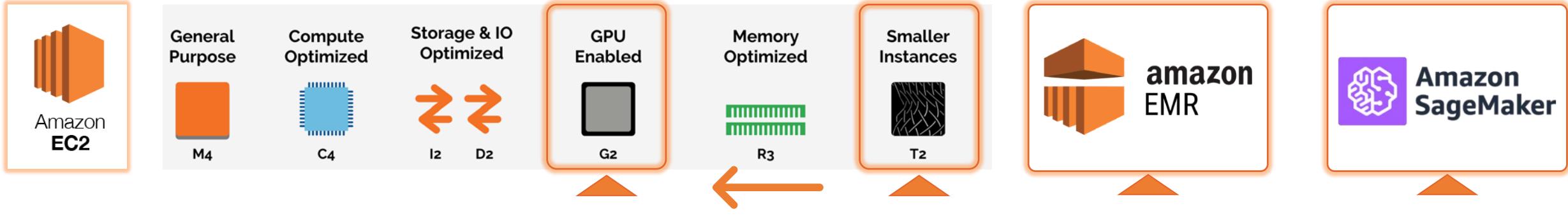
Rechercher des objets en fonction du préfixe

<input type="checkbox"/>	Nom	Type	Dernière modification	Taille	Classe de stockage
<input type="checkbox"/>	_SUCCESS	-	23 Nov 2022 10:39:52 AM CET	0 o	Standard
<input type="checkbox"/>	part-00000-7c979eca-8ffd-494a-abcd-651172b0773b-c000.csv	csv	23 Nov 2022 10:39:51 AM CET	84.6 Mo	Standard
<input type="checkbox"/>	part-00001-7c979eca-8ffd-494a-abcd-651172b0773b-c000.csv	csv	23 Nov 2022 10:39:49 AM CET	84.0 Mo	Standard
<input type="checkbox"/>	part-00002-7c979eca-8ffd-494a-abcd-651172b0773b-c000.csv	csv	23 Nov 2022 10:39:44 AM CET	82.6 Mo	Standard
<input type="checkbox"/>	part-00003-7c979eca-8ffd-494a-abcd-651172b0773b-c000.csv	csv	23 Nov 2022 10:39:47 AM CET	82.1 Mo	Standard
<input type="checkbox"/>	part-00004-7c979eca-8ffd-494a-abcd-651172b0773b-c000.csv	csv	23 Nov 2022 10:39:46 AM CET	84.1 Mo	Standard

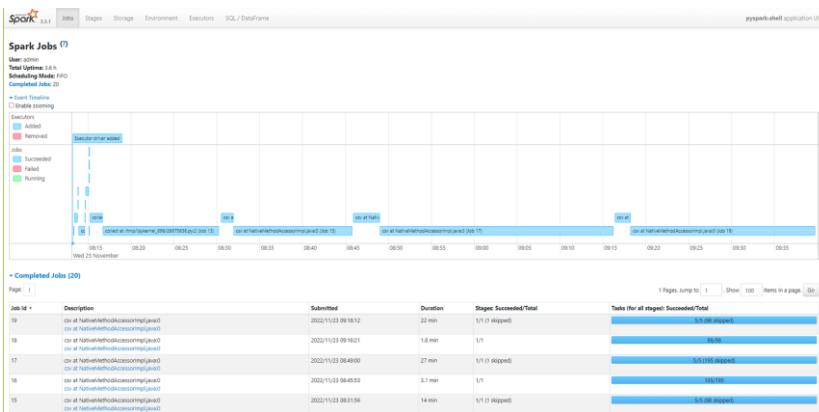


3 Chaine de Traitement

Remarques & préconisations



NoSQL



VS





Questions ?

Merci !