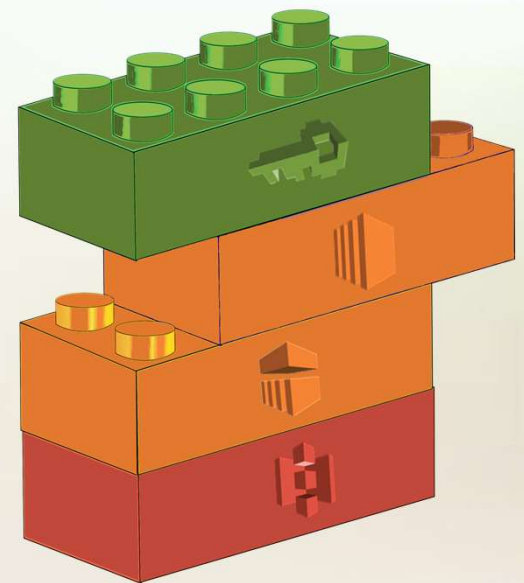


Déployez un modèle dans le

cloud



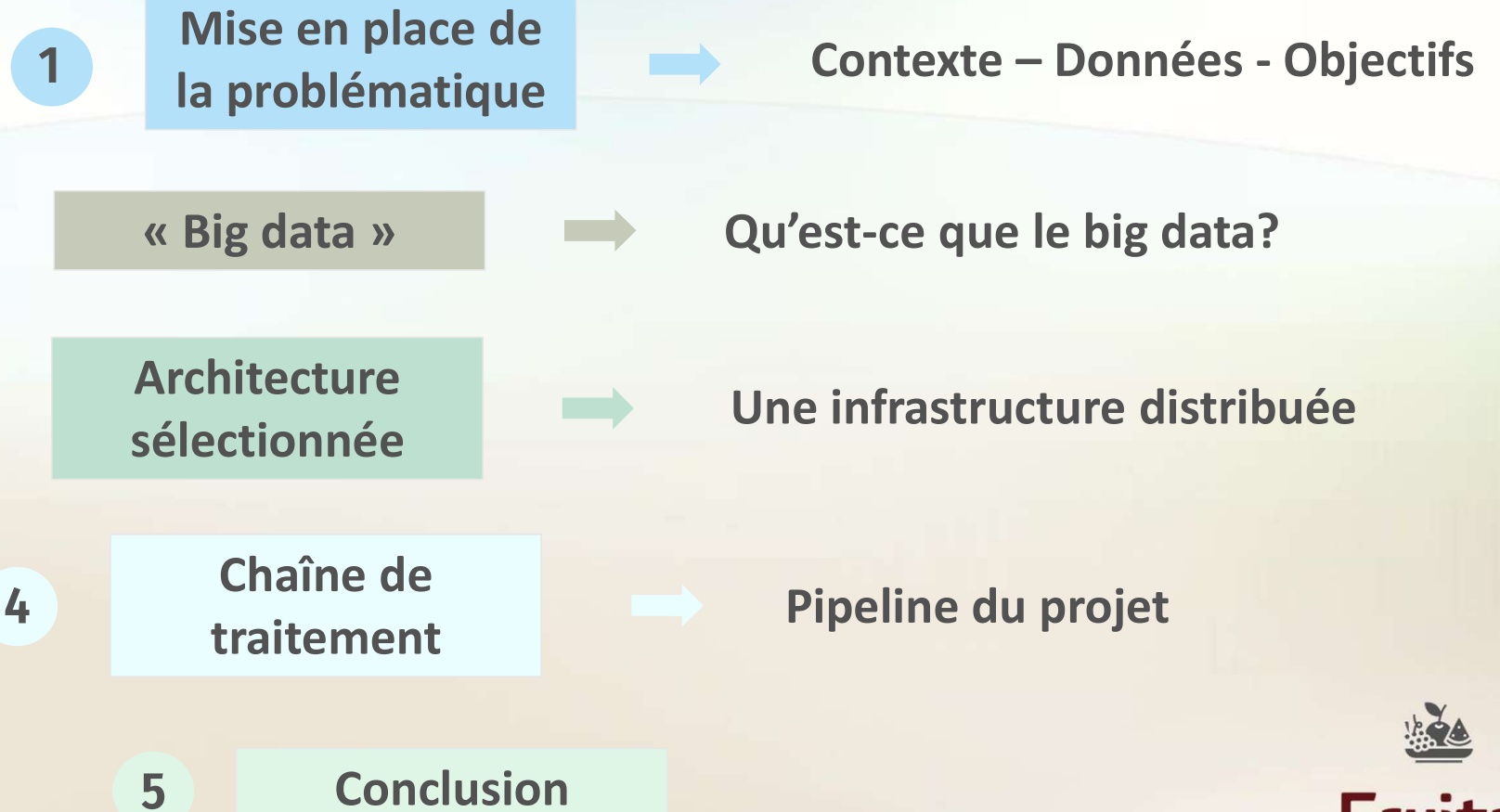
Fruits!



MAI 2021

Christelle Troussard

Sommaire



Fruits!

Problématique

1

Contexte – Données - Objectifs

Le contexte


Fruits!
Solutions innovantes
pour la récolte des fruits



AgriTech : l'IA au service de l'agriculture

Fruits! : Startup de l'AgriTech



Les données initiales

Jeux avec étiquettes

- ✓ 90 380 images
- ✓ 131 classes

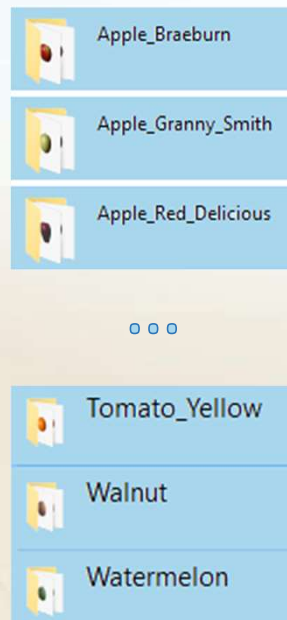
Jeu entraînement

- ✓ 67 692 images
- ✓ 131 classes

Jeu test

- ✓ 22 688 images
- ✓ 131 classes

131 dossiers



Photos 360° de fruits et légumes



Fond blanc, 100x100 pixels



Un seul fruit/légume par image

Jeux sans étiquette

- ✓ multi fruits
- ✓ 103 images

Quelles solutions pour répondre aux enjeux?

Big data

2

Qu'est-ce que le big data?

Quelles solutions pour répondre aux enjeux?

Le big data

Données massives



Volume de données considérables à traiter

Volume



Le Big Data c'est quoi ?

Explosion de la quantité des données
Le partage des données
La recherche des données
Le stockage des données
Le traitement des flux de données

3 V

Big data



Variété

Variété d'informations

diverses sources, non-structurées



Vélocité



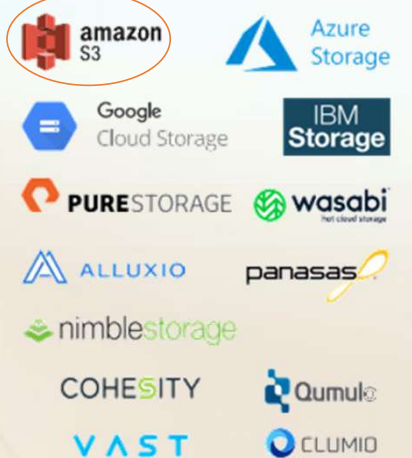
Vitesse de création,
fréquence de création
collecte et partage des données

Une infrastructure distribuée

Stocker , traiter et diffuser des mégadonnées

Solutions existantes

STORAGE



Notre solution

Stockage distribué



évolutivité

Pas de limite de place
Ressources à l'échelle
Scalabilité
Disponibilité



résilience

Redondance : copie des objets
sur plusieurs systèmes
Tolérance aux pannes

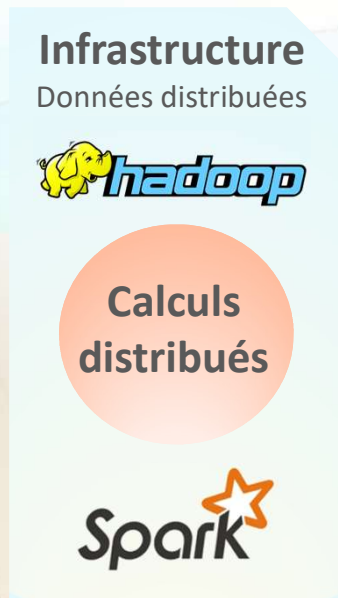
performance

Durabilité
Bonne compression

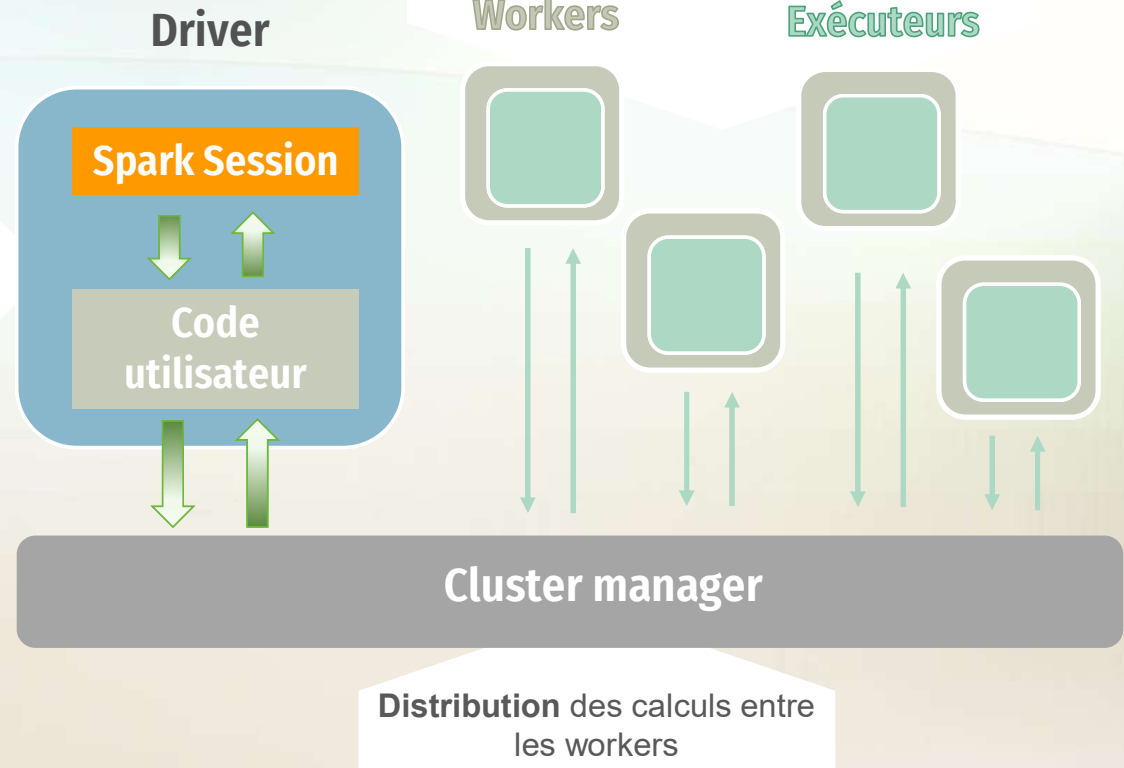
Une infrastructure distribuée

traiter et diffuser des mégadonnées

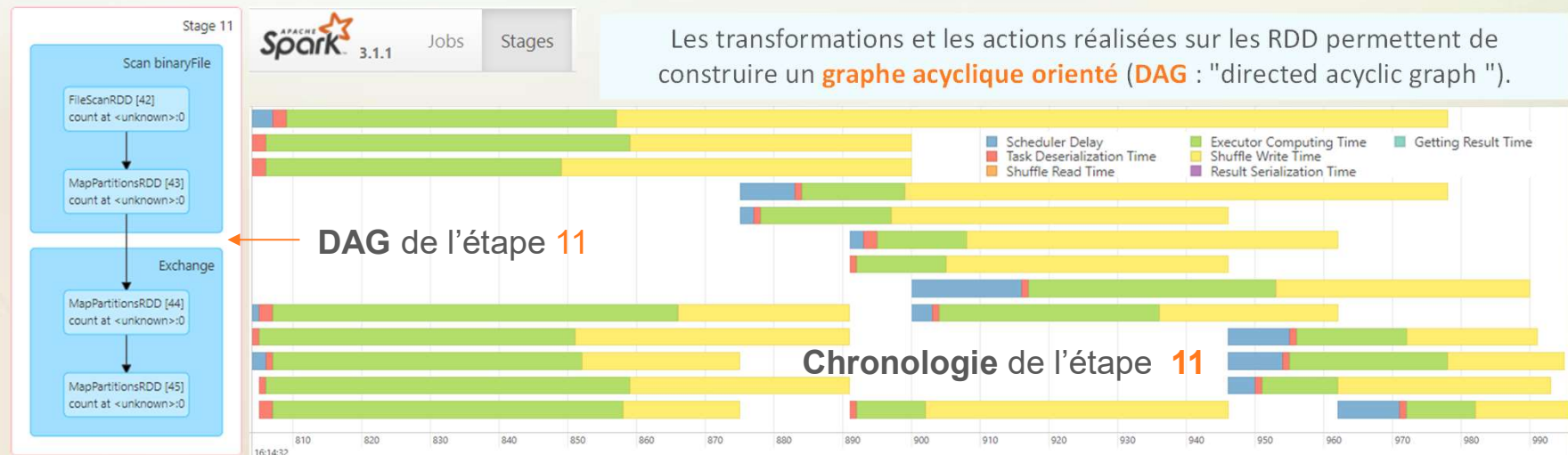
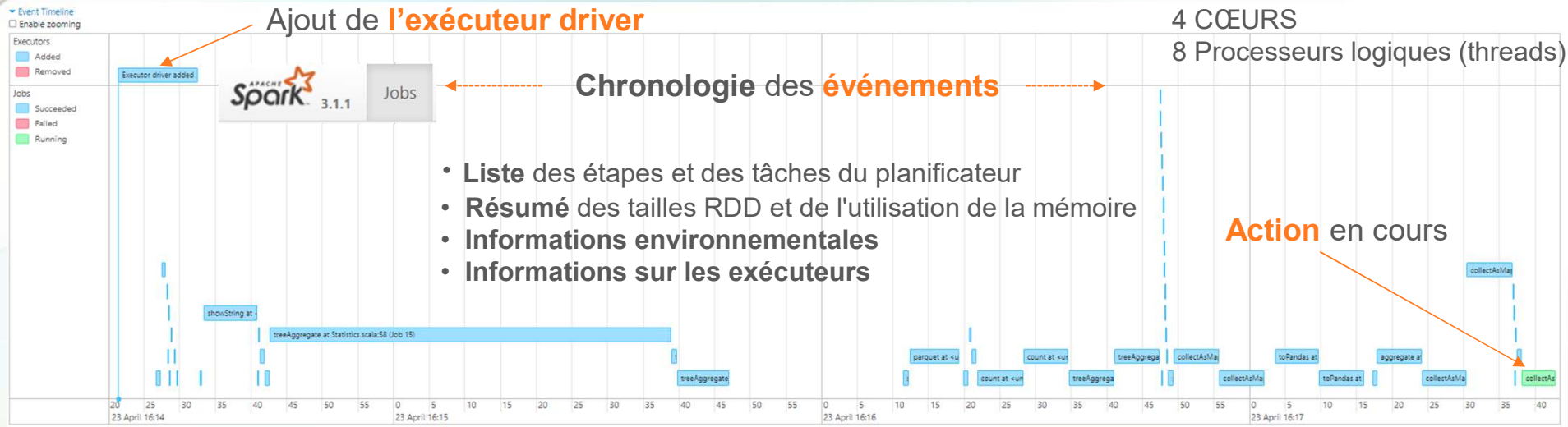
Notre solution




Configuration
Initialisation
Agrégation des calculs



MapReduce : Map (transformer) Reduce (agréger) **Divisez pour distribuer pour régner**



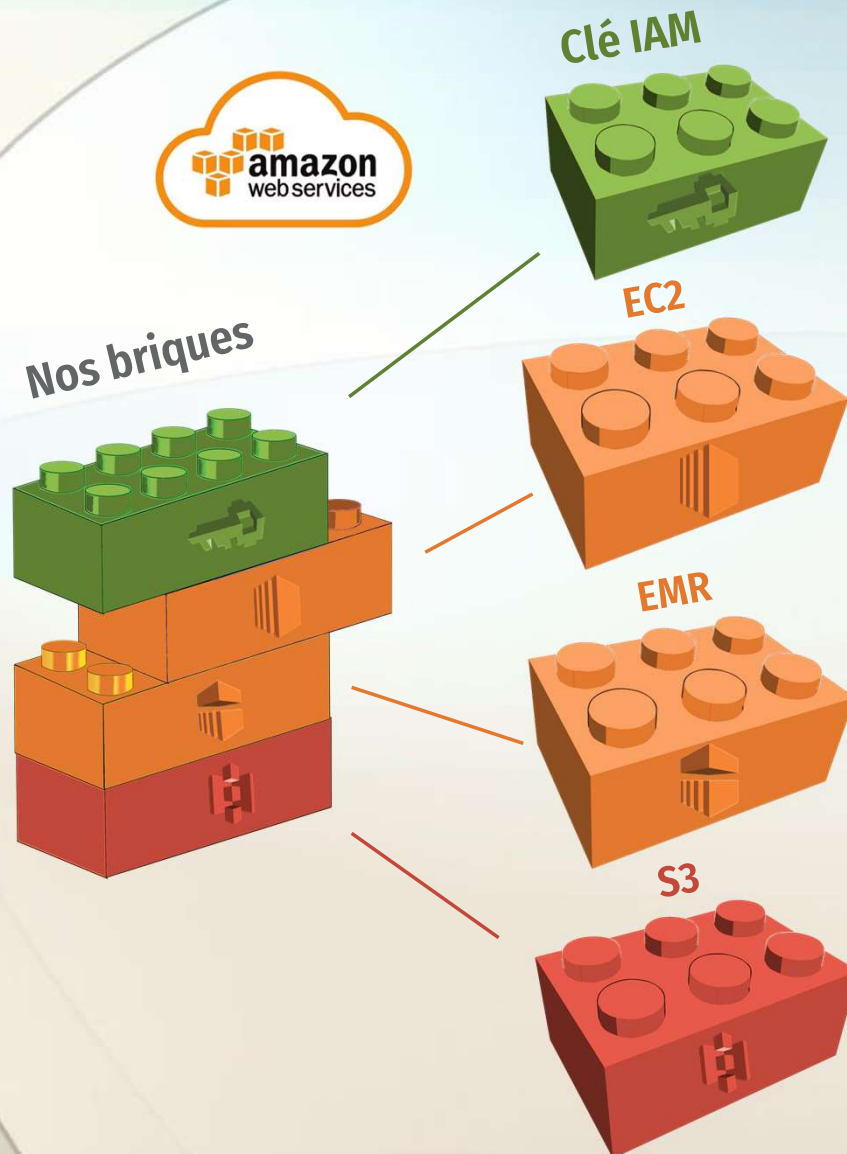
Architecture



3

Une infrastructure distribuée

Architecture



IAM Sécurité renforcée

Identity and Access Management

Contrôlez de façon sécurisée l'accès aux services et ressources AWS.

EC2 Serveurs virtuels dans le cloud

Elastic Compute Cloud

Capacité de calcul sécurisée et redimensionnable pouvant prendre en charge quasiment tout type de charge de travail

☑ Une infrastructure à la demande fiable et évolutive

EMR Analyse

Elastic Map Reduce

Amazon EMR est un service qui utilise Apache Spark pour traiter et analyser de grandes quantités de données.

☑ Exécutez et mettez à l'échelle facilement les cadres Apache Spark, Hive, Presto et d'autres cadres de Big Data.

S3 Stockage scalable dans le cloud

Simple Storage Service

Stockage d'objets conçu pour stocker et récupérer n'importe quelle quantité de données, n'importe où

☑ Performances, scalabilité, disponibilité et durabilité de pointe

Un « ordinateur virtuel »



Mais bien plus encore ...

Local versus cloud

	Local	Cloud
Stockage	Disque dur : limité Panne possible : perte des données Données disponibles localement	Illimité Redondance : tolérance aux pannes Données disponibles partout
Puissance de calcul	Dépendante du matériel informatique à disposition	Evolutive en fonction de la charge de travail
Ethique	Les informations restent au sein de l'entreprise Protection en interne des données à caractère confidentiel	Contraintes juridiques liées à l'hébergement des données (si le centre d'hébergement se trouve à l'étranger). Confidentialité des données
Sécurité	Choix de l'utilisateur	

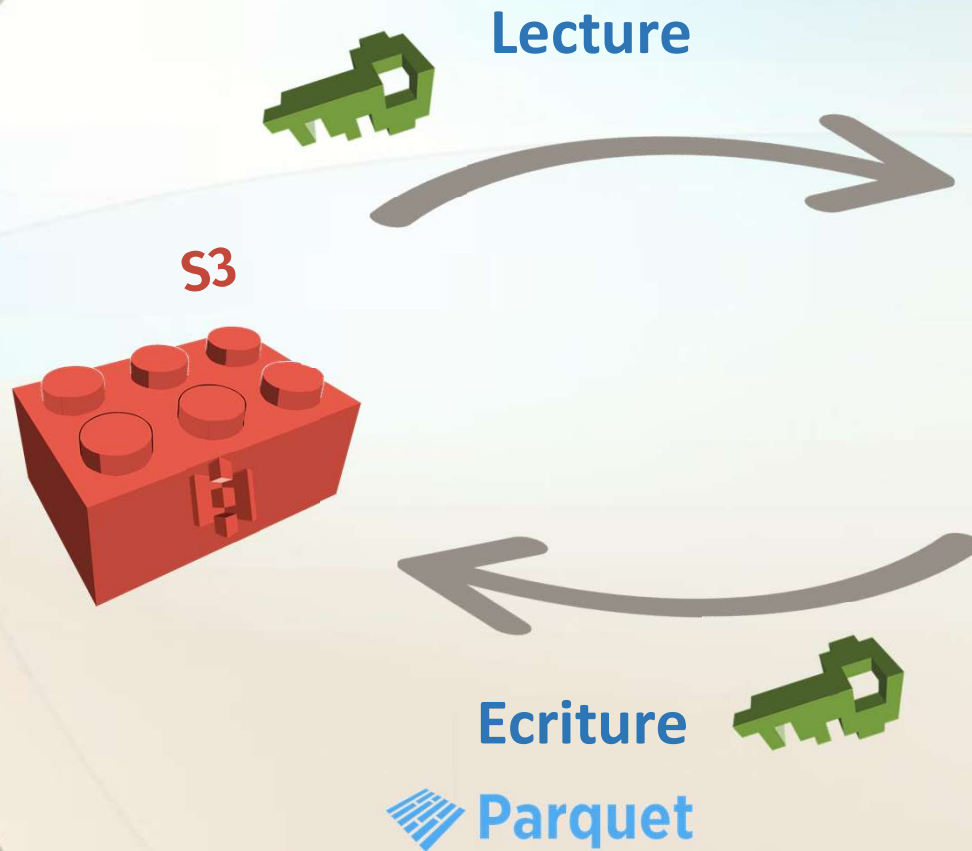
Il s'agit de découvrir de nouveaux ordres de grandeur concernant : capture, recherche, partage, stockage, analyse et présentation des données.

Chaîne de traitement

4

Pipeline du projet

Chaîne de traitement



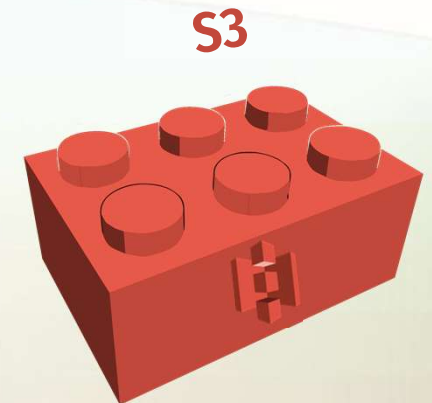
Traitement



- **Traiter** de très grandes quantités de données
- **Créer** des pipelines d'apprentissage automatique

PySpark

Stockage S3



s3://p8bucket

Compartiment S3

Nom	
app_P8.py	
configuration.json	
emr_bootstrap.sh	
images/	
logs/	
resultat_parquet/	

Données chargées

Données enregistrées

Données initiales

Code python

Configuration

Librairies à installer

Fichiers en format parquet

```
# Création d'une SparkSession
spark = SparkSession.builder\
    .appName('P8_preprocess_images')\
    .getOrCreate()

# Chargement des données
# En format "binaryFile"
df_binary = spark.read.format("binaryFile") \
    .option("pathGlobFilter", "*.jpg") \
    .option("recursiveFileLookup", "true") \
    .load(data_source)
```

```
[
  {
    "Classification": "spark-env",
    "Configurations":
      [
        {
          "Classification": "export",
          "Properties":
            {
              "PYSPARK_PYTHON": "/usr/bin/python3"
            }
        }
      ]
  }
]
```

```
#!/bin/bash
sudo pip install numpy
sudo pip install pandas
sudo pip install Pillow
sudo pip install findspark
sudo pip install pyarrow==0.15.1
sudo python3 -m pip install numpy
sudo python3 -m pip install Pillow
sudo python3 -m pip install findspark
sudo python3 -m pip install pandas
sudo python3 -m pip install pyarrow==0.15.1
```

._SUCCESS.crc	23/04/2021 16:16	Fichier CRC	1 Ko
.part-00000-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00001-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00002-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00003-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00004-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00005-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00006-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00007-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00008-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00009-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00010-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00011-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00012-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00013-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00014-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00015-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko

Création d'un cluster EMR



Configuration

Configuration des logiciels

Libérer : emr-6.1.0

<input checked="" type="checkbox"/> Hadoop 3.2.1	<input type="checkbox"/> Zeppelin 0.9.0	<input checked="" type="checkbox"/> Livy 0.7.0
<input type="checkbox"/> JupyterHub 1.1.0	<input type="checkbox"/> Tez 0.9.2	<input type="checkbox"/> Flink 1.11.0
<input type="checkbox"/> Ganglia 3.7.2	<input type="checkbox"/> HBase 2.2.5	<input type="checkbox"/> Pig 0.17.0
<input type="checkbox"/> Hive 3.1.2	<input type="checkbox"/> Presto 0.232	<input type="checkbox"/> PrestoSQL 338
<input type="checkbox"/> ZooKeeper 3.4.14	<input type="checkbox"/> MXNet 1.6.0	<input type="checkbox"/> Sqoop 1.4.7
<input type="checkbox"/> Hue 4.7.1	<input type="checkbox"/> Phoenix 5.0.0	<input type="checkbox"/> Oozie 5.2.0
<input checked="" type="checkbox"/> Spark 3.0.0	<input type="checkbox"/> HCatalog 3.1.2	<input checked="" type="checkbox"/> TensorFlow 2.1.0

Modifier les paramètres du logiciel

☐ Entrer la configuration ☒ Charger JSON à partir de S3

s3://p8fruitsbucket/configuration.json

configuration.json

m5.xlarge

EC2

EMR

Ajout d'étape

Type d'étape : Application Spark				Ajouter une étape
Nom	Action sur échec	Emplacement JAR	Arguments	
Application Spark	Continuer	command-runner.jar	spark-submit --deploy-mode client s3://p8fruitsbucket/app_P8.py -- data_source s3://p8fruitsbucket/data/* -- output_uri s3://p8fruitsbucket/resultats_par quets	

Arguments

```
spark-submit --deploy-mode
client
s3://p8fruitsbucket/app_P8.py --
data_source
s3://p8fruitsbucket/data/* --
output_uri
s3://p8fruitsbucket/resultats_par
quets
```

Amorçage

Actions d'amorçage			
Les actions d'amorçage sont des scripts exécutés lors de la configuration avant le démarrage de Hadoop sur chaque nœud de cluster. Vous pouvez les utiliser pour installer des logiciels supplémentaires et personnaliser vos applications. En savoir plus			
Type d'action d'amorçage	Nom	Emplacement JAR	Arguments facultatifs
Action personnalisée	Action personnalisée	s3://p8fruitsbucket/emr_bootstrap.sh	

Ajouter une action d'amorçage : Action personnalisée

emr_bootstrap.sh

Clés

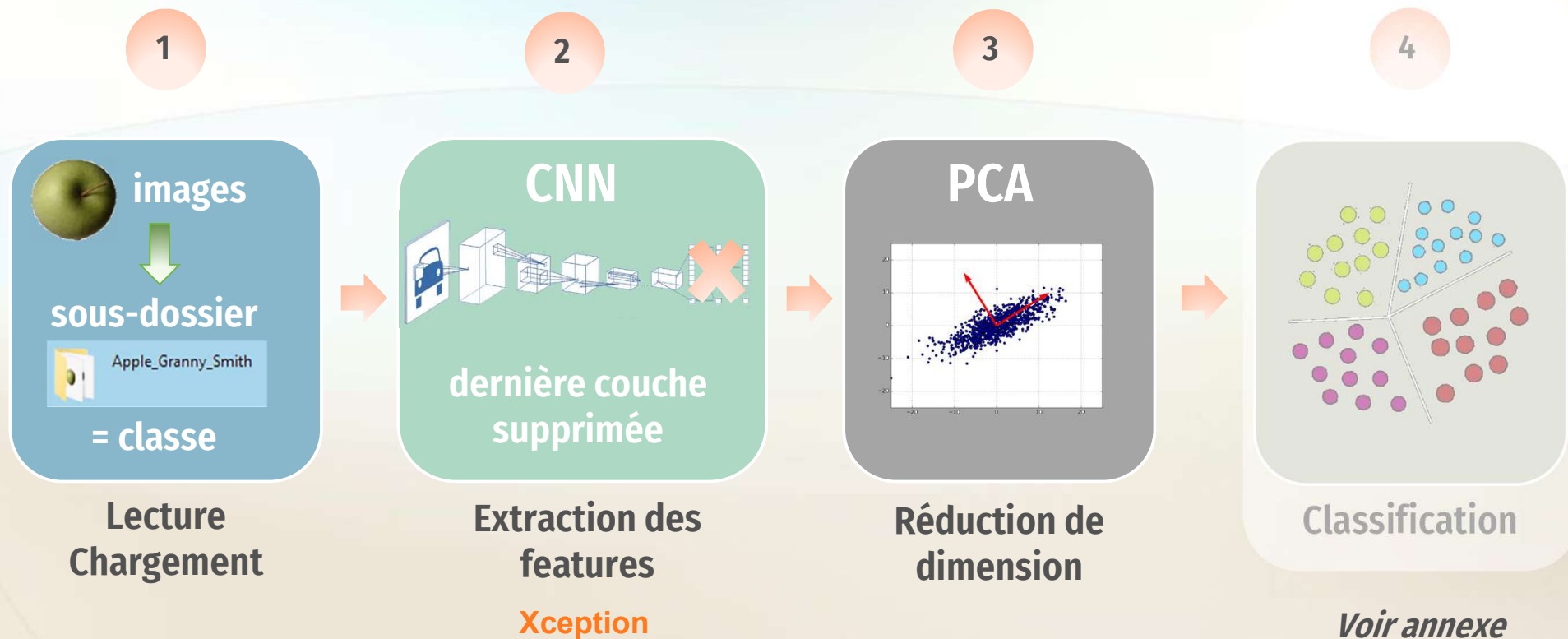
Options de sécurité

Paire de clés EC2 : key_london

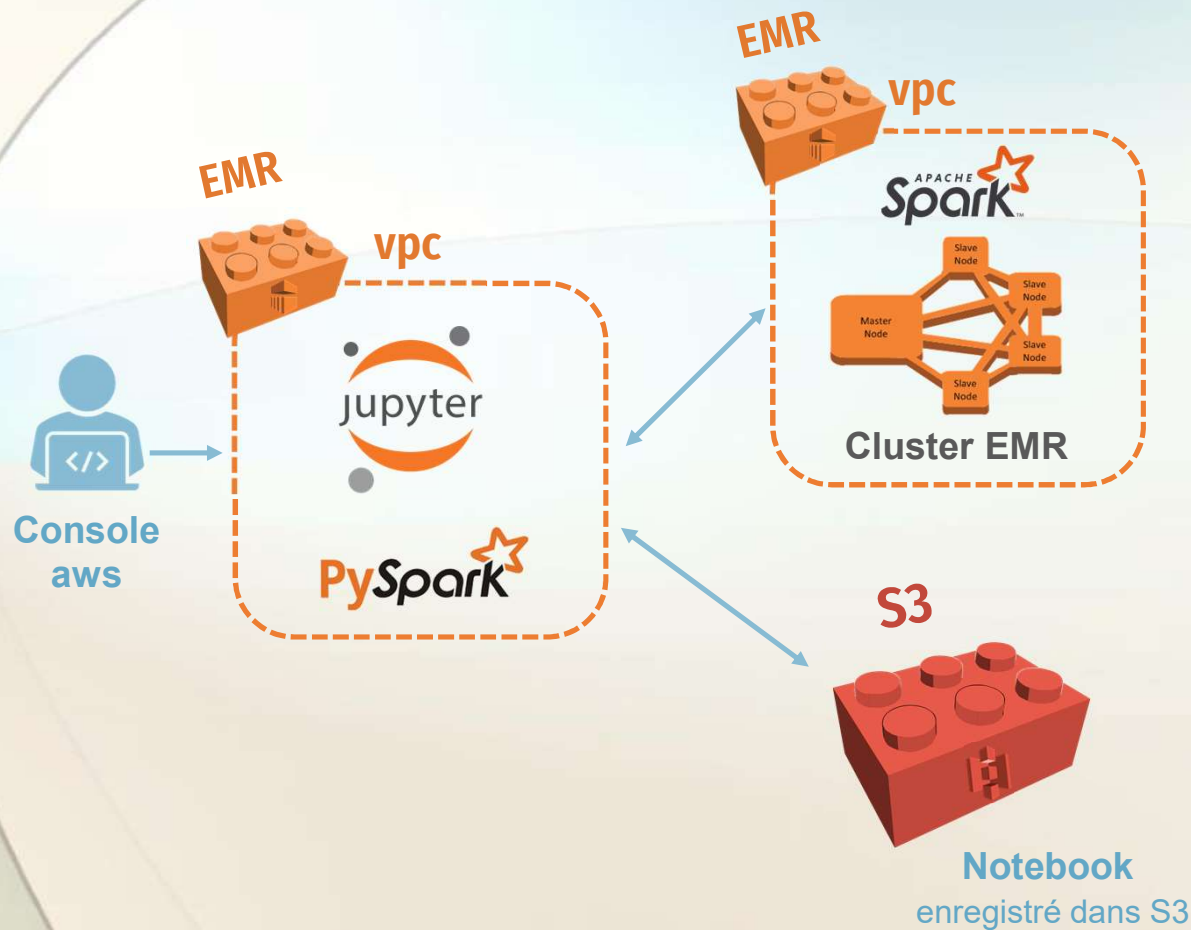
☒ Cluster visible pour tous les utilisateurs IAM du compte



Vcpu : 4
Mémoire : 16 Gio



Amazon EMR notebooks



Starting Spark application

ID	YARN Application ID	Kind	State	Spark UI	Driver log	Current session?
1	application_1619165088340_0003	pyspark	idle	Link	Link	✓

SparkSession available as 'spark'.

```
Entrée [7]: # Création d'une SparkSession
spark = SparkSession.builder()
    .appName('P8_preprocess_images') \
    .getOrCreate()
```

```
Entrée [9]: data_source = 's3://p8bucket/images/**'
```

```
Entrée [10]: # Chargement des données
# En format "binaryFile"
df_binary = spark.read.format("binaryFile") \
    .option("pathGlobFilter", "*.jpg") \
    .option("recursiveFileLookup", "true") \
    .load(data_source)
```

```
Entrée [11]: # On extrait la classe de chaque image de fruit
df_binary = df_binary.withColumn('classe', element_at(split(df_binary['path'], "/"), -2))
```

```
Entrée [12]: df_binary.show(5)
```

Spark Job Progress

path	modificationTime	length	content	classe
s3://p8bucket/ima...	2021-04-23 06:33:02	6989	[FF D8 FF E0 00 1...	Watermelon
s3://p8bucket/ima...	2021-04-23 06:34:21	6987	[FF D8 FF E0 00 1...	Watermelon
s3://p8bucket/ima...	2021-04-23 06:33:14	6984	[FF D8 FF E0 00 1...	Watermelon
s3://p8bucket/ima...	2021-04-23 06:33:00	6982	[FF D8 FF E0 00 1...	Watermelon
s3://p8bucket/ima...	2021-04-23 06:33:01	6973	[FF D8 FF E0 00 1...	Watermelon

only showing top 5 rows

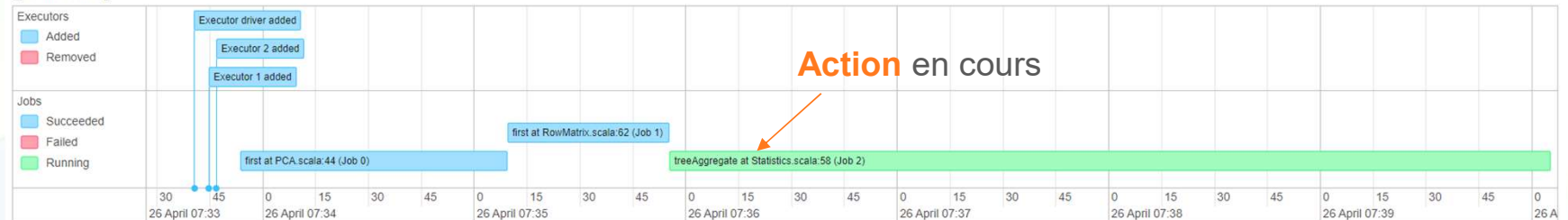
```
Entrée [13]: ## Extraction des features
# Modèle
model = Xception(
    include_top=False, # top layer supprimé
    weights=None,
    input_shape=(100,100,3),
    pooling='max'
)
```

Spark UI

Spark Jobs (?)

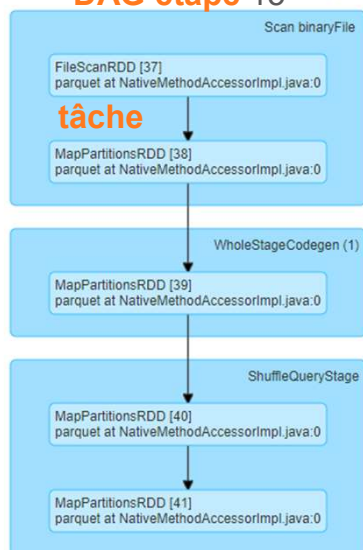
User: hadoop
Total Uptime:
Scheduling Mode: FIFO
Active Jobs: 1
Completed Jobs: 2

Event Timeline
☐ Enable zooming



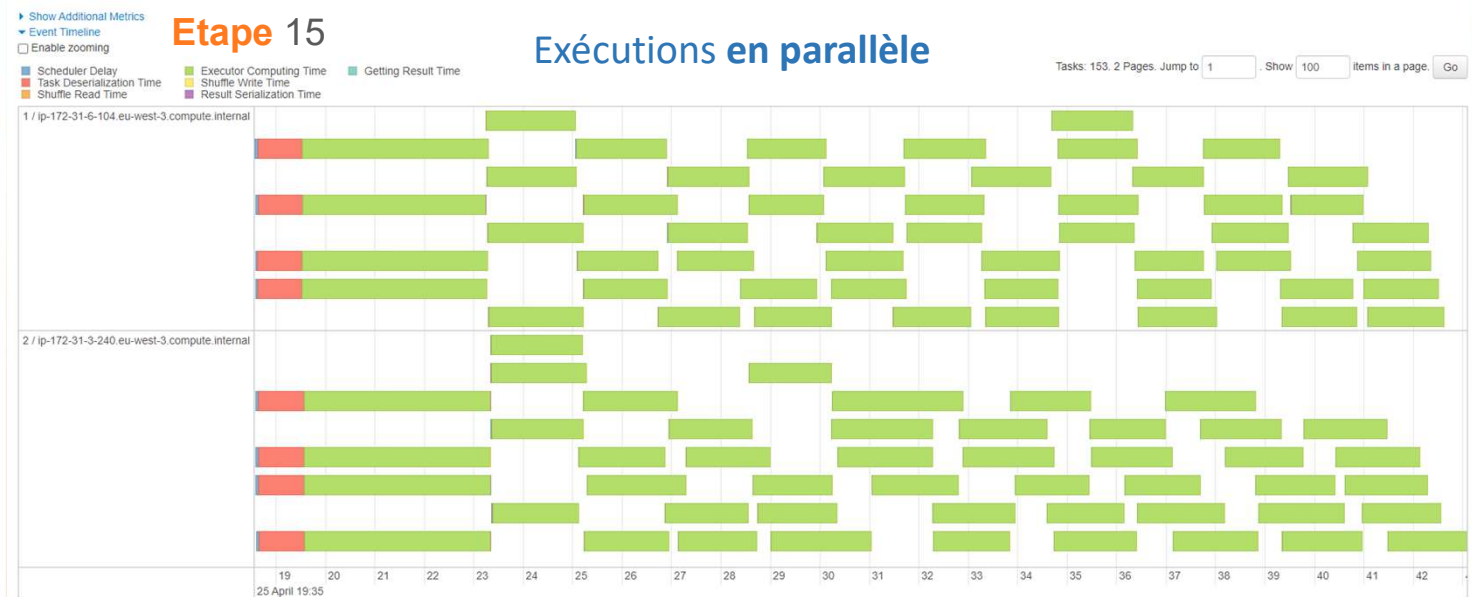
DAG étape 15

Stage 15



Etape 15

Exécutions en parallèle



Enregistrements des données

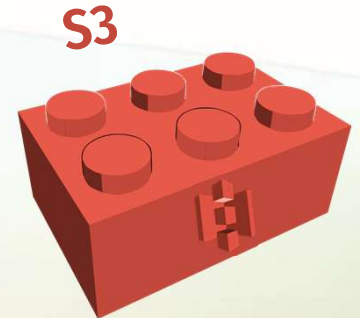
Images Classe Features array Features vectors Vecteurs PCA

s3://p8bucket/resultats_parquet

path	classe	X_features	X_vectors	X_vectors_pca
s3://p8bucket/ima...	Strawberry	[7.528077E-5, 5.7...	[7.52807682147249...	[-0.0024979452002...
s3://p8bucket/ima...	Peach	[7.164011E-5, 4.6...	[7.16401118552312...	[-0.0022347604461...
s3://p8bucket/ima...	Peach	[7.0007634E-5, 4...	[7.00076343491673...	[-0.0022332804300...
s3://p8bucket/ima...	Orange	[5.2437896E-5, 5...	[5.24378956470172...	[-0.0022457391958...
s3://p8bucket/ima...	Apple_Red_1	[6.39236E-5, 5.58...	[6.39236022834666...	[-0.0022016862054...
s3://p8bucket/ima...	Peach	[6.562176E-5, 4.3...	[6.56217598589137...	[-0.0018781298588...
s3://p8bucket/ima...	Apple_Red_1	[6.592149E-5, 6.1...	[6.592149293283E-...	[-0.002221958511...
s3://p8bucket/ima...	Strawberry	[8.3148494E-5, 5...	[8.31484940135851...	[-0.0025633449203...
s3://p8bucket/ima...	Peach	[5.6137997E-5, 4...	[5.61379965802188...	[-0.0023045842887...
s3://p8bucket/ima...	Peach	[4.3400338E-5, 5...	[4.34003377449698...	[-0.0020065563145...
s3://p8bucket/ima...	Strawberry	[5.3844553E-5, 4...	[5.38445528945885...	[-0.0024053836598...
s3://p8bucket/ima...	Peach	[7.0101734E-5, 5...	[7.01017343089915...	[-0.0020346894213...
s3://p8bucket/ima...	Peach	[7.544263E-5, 5.2...	[7.54426291678100...	[-0.0019785420552...
s3://p8bucket/ima...	Strawberry	[6.331178E-5, 5.0...	[6.33117815596051...	[-0.0023900177974...
s3://p8bucket/ima...	Peach	[4.662218E-5, 6.2...	[4.66221790702547...	[-0.0023086338846...
s3://p8bucket/ima...	Strawberry	[8.312277E-5, 6.4...	[8.31227735034190...	[-0.0025797946202...
s3://p8bucket/ima...	Pear	[6.2984975E-5, 4...	[6.29849746474064...	[-0.0018198891807...
s3://p8bucket/ima...	Strawberry	[8.780114E-5, 6.1...	[8.78011414897628...	[-0.0025548053255...
s3://p8bucket/ima...	Peach	[6.663117E-5, 5.0...	[6.66311680106446...	[-0.0023358706523...
s3://p8bucket/ima...	Peach	[5.9448852E-5, 5...	[5.94488519709557...	[-0.0023381211192...

enregistré au format distribué


Parquet



._SUCCESS.crc	23/04/2021 16:16	Fichier CRC	1 Ko
.part-00000-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00001-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00002-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00003-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00004-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00005-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00006-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00007-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00008-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00009-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00010-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00011-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00012-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00013-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00014-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00015-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko

Nom
app_P8.py
configuration.json
emr_bootstrap.sh
images/
logs/
resultat_parquet/

Bonne compression , conçu pour les données massives



Conclusion

Architecture retenue - Passage à l'échelle



Montée en compétence – Difficultés rencontrées

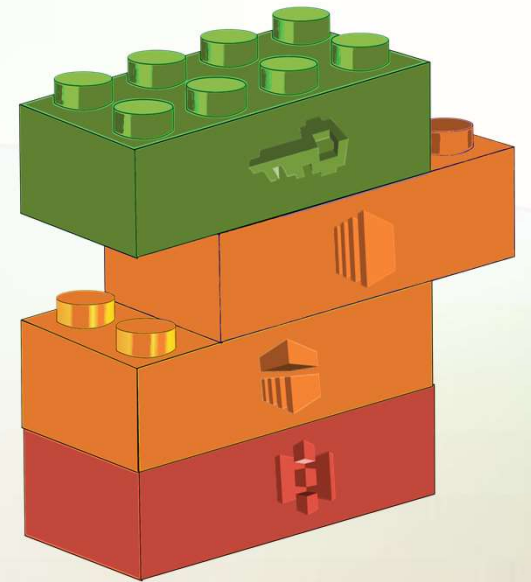
- ✓ Découverte de l'écosystème **Hadoop**, du moteur de traitement de données massives **Spark**, prise en main de l'**API Pyspark**, et du système d'exploitation **Linux** (Ubuntu 20.04 LTS via **wsl**)
- Découverte de l'écosystème **AWS**.
- ✗ Nombreuses **erreurs, peu explicites** pour le profane.
- Possibilités techniques nombreuses : difficile avec peu d'expérience, d'être assuré d'avoir fait **le bon choix**!

Perspectives – Améliorations possibles

- Scripts en **scala**
- **Gpu versus Cpu** ! Réflexions à mener : *Spark 3 Demo: Comparing Performance of GPUs vs. CPUs*
<https://www.youtube.com/watch?v=tGqEZYUgexY> (video nvidia...)



Fruits!

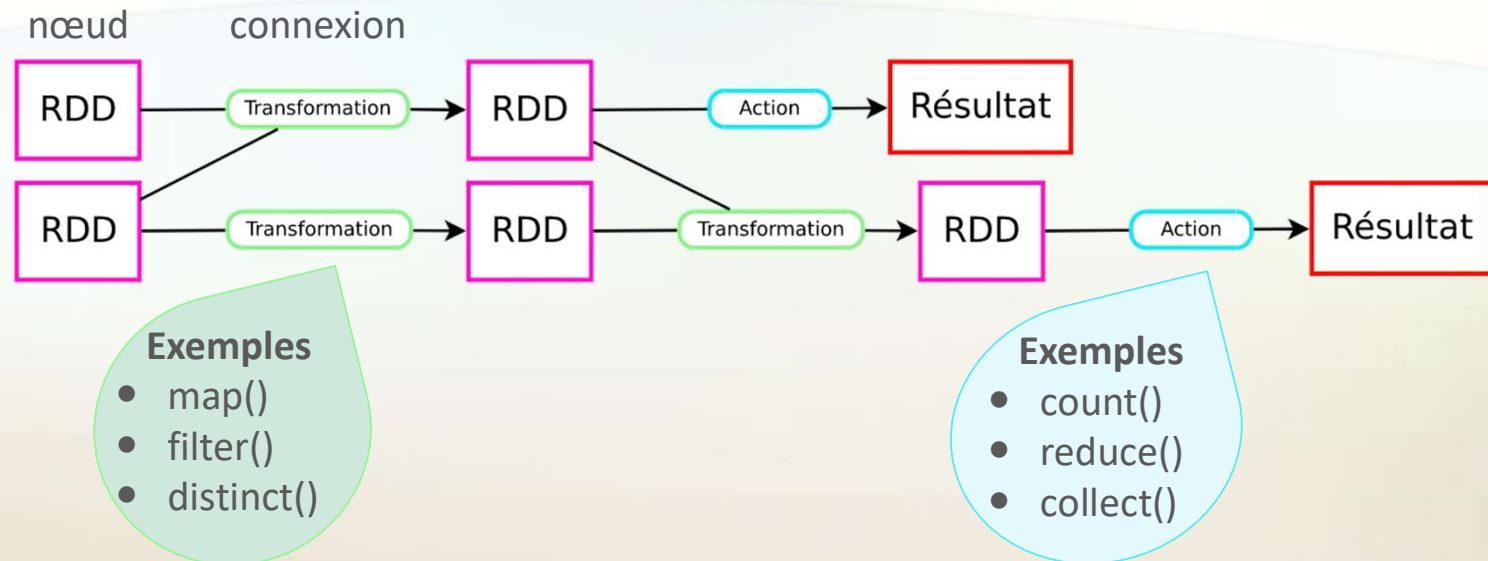


Une infrastructure distribuée

traiter et diffuser des mégadonnées

Dans une application Spark, les **transformations** et les **actions** réalisées sur les RDD permettent de construire un **graphe acyclique orienté (DAG : "directed acyclic graph")**

Les
**Resilient
Distributed
Dataset
(RDD)**



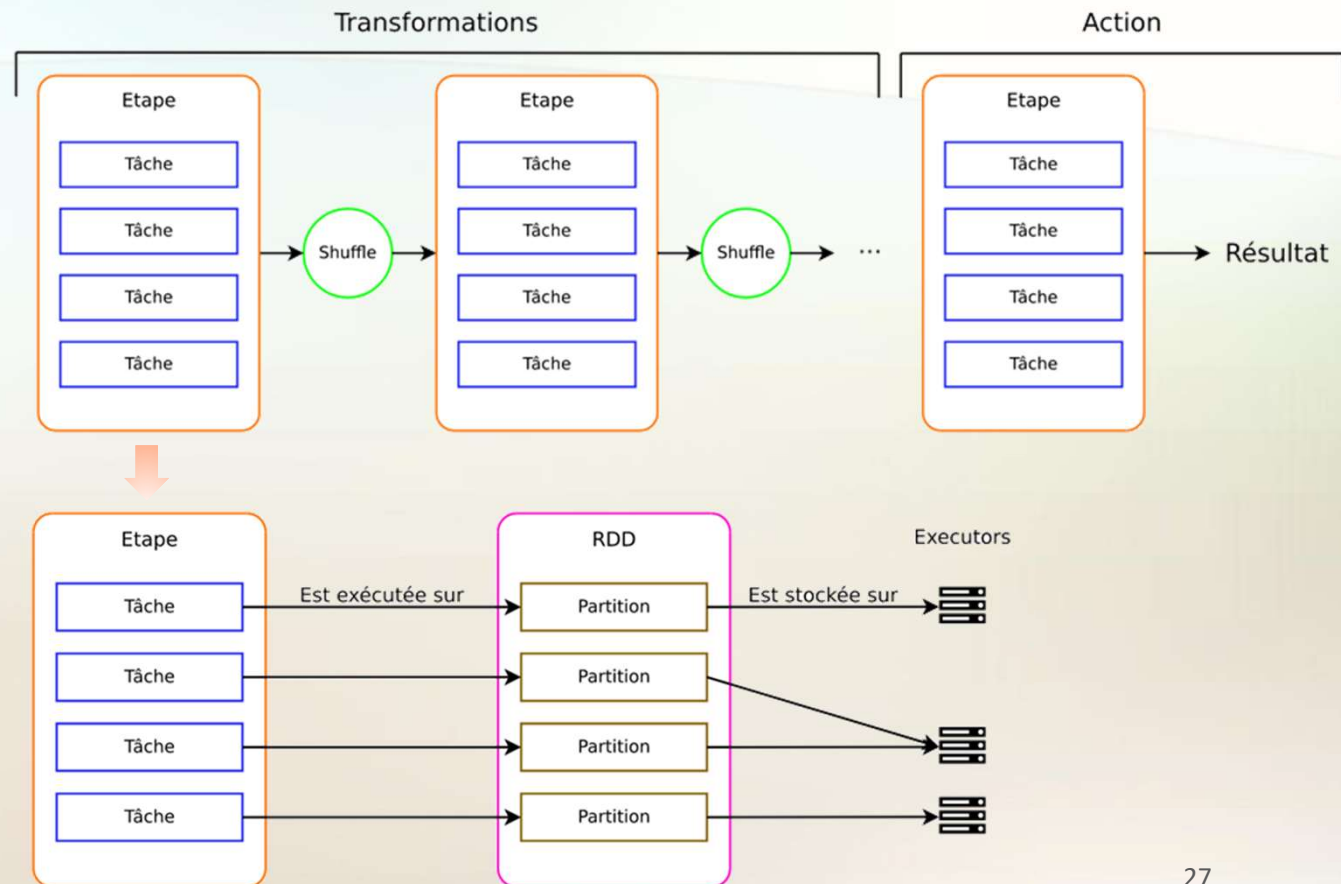
Lorsqu'un nœud devient indisponible, il peut être régénéré à partir de ses nœuds parents. C'est précisément ce qui permet la **tolérance aux pannes** des applications Spark. Spark utilise une **évaluation paresseuse**, ce qui signifie qu'il ne fait aucun travail jusqu'à ce que vous demandiez un résultat.

Une infrastructure distribuée

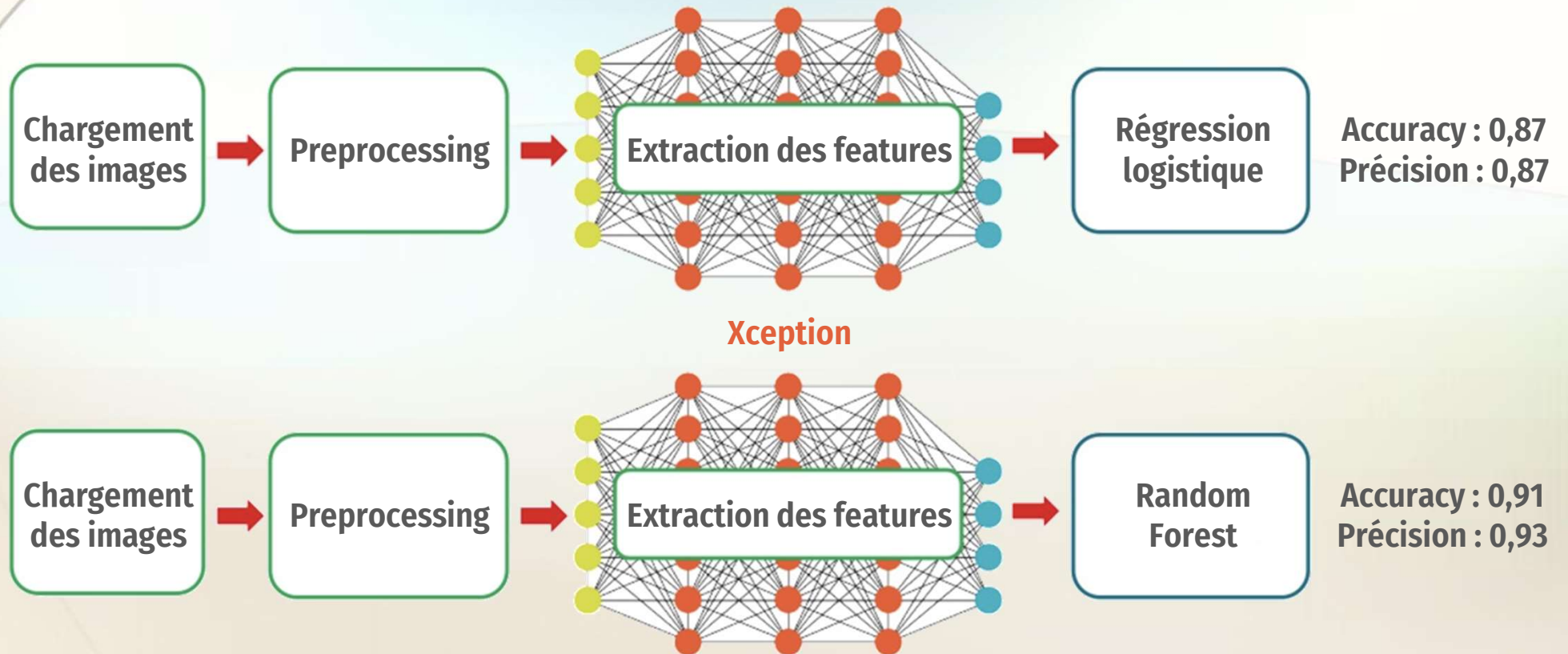
traiter et diffuser des mégadonnées

Un **job** Spark correspond à une action sur un RDD et est composé de plusieurs **étapes** séparées par des **shuffles**.

- **Partitions** : découpage des données
- **Tâche** : traitement d'une partition
- **Étape** : ensemble de tâches réalisées en parallèle
- **Shuffle** : redistribution des données entre les nœuds



Pipelines classification



Logout de l'application enregistré sur S3

path	classe	X_vectors	X_vectors_pca
s3://p8bucket/ima...	Watermelon	[4.30195686931256...	[-0.0014724724050...
s3://p8bucket/ima...	Watermelon	[3.85823259421158...	[-0.0014240677179...
s3://p8bucket/ima...	Watermelon	[2.91877313429722...	[-0.0013982374124...
s3://p8bucket/ima...	Watermelon	[4.27223603765014...	[-0.0013877593994...
s3://p8bucket/ima...	Watermelon	[3.48064459103625...	[-0.0013936819953...
s3://p8bucket/ima...	Clementine	[4.13163252233061...	[-0.0023987243574...
s3://p8bucket/ima...	Watermelon	[2.83777862932765...	[-0.0015022272896...
s3://p8bucket/ima...	Watermelon	[3.01670806948095...	[-0.0014527697217...
s3://p8bucket/ima...	Clementine	[3.88991320505738...	[-0.0024666961772...
s3://p8bucket/ima...	Clementine	[4.27966697316151...	[-0.0025012090477...
s3://p8bucket/ima...	Clementine	[4.84576812596060...	[-0.0024858020994...
s3://p8bucket/ima...	Clementine	[5.10514291818253...	[-0.0025488732644...
s3://p8bucket/ima...	Clementine	[4.07514417020138...	[-0.0023361143739...
s3://p8bucket/ima...	Clementine	[5.23504895681981...	[-0.0024441455624...
s3://p8bucket/ima...	Strawberry	[3.92002475564368...	[-0.0016503117639...
s3://p8bucket/ima...	Clementine	[4.92607614432927...	[-0.0023981603203...
s3://p8bucket/ima...	Orange	[3.83260485250502...	[-0.0020354908982...
s3://p8bucket/ima...	Orange	[3.81613535864744...	[-0.0020261696806...
s3://p8bucket/ima...	Apple_Red_1	[4.32711349276360...	[-0.0014783001823...
s3://p8bucket/ima...	Strawberry	[3.16732912324368...	[-0.0016672352046...

only showing top 20 rows

Mini-classification

classe	X_vectors_pca
Apricot	[-0.0017985031932...
Lemon	[-0.0023381406432...
Peach	[-0.0017537706410...

only showing top 3 rows

classe	X_vectors_pca	labelIndex
Watermelon	[-0.0013619884910...	9.0
Watermelon	[-0.0014813502034...	9.0
Watermelon	[-0.0014462156936...	9.0

only showing top 3 rows

Training Dataset Count: 3396
Test Dataset Count: 1490

Regression logistique

prediction	labelIndex
0.0	0.0
0.0	0.0
0.0	0.0
0.0	0.0
0.0	0.0

only showing top 5 rows

Test Error = 0.126846
Accuracy = 0.873154
Test Error = 0.12471
Precision = 0.87529

Random forest

prediction	labelIndex
0.0	0.0
0.0	0.0
0.0	0.0
0.0	0.0
0.0	0.0

only showing top 5 rows

Test Error = 0.0885906
Accuracy = 0.911409
Test Error = 0.0741965
Precision = 0.925803