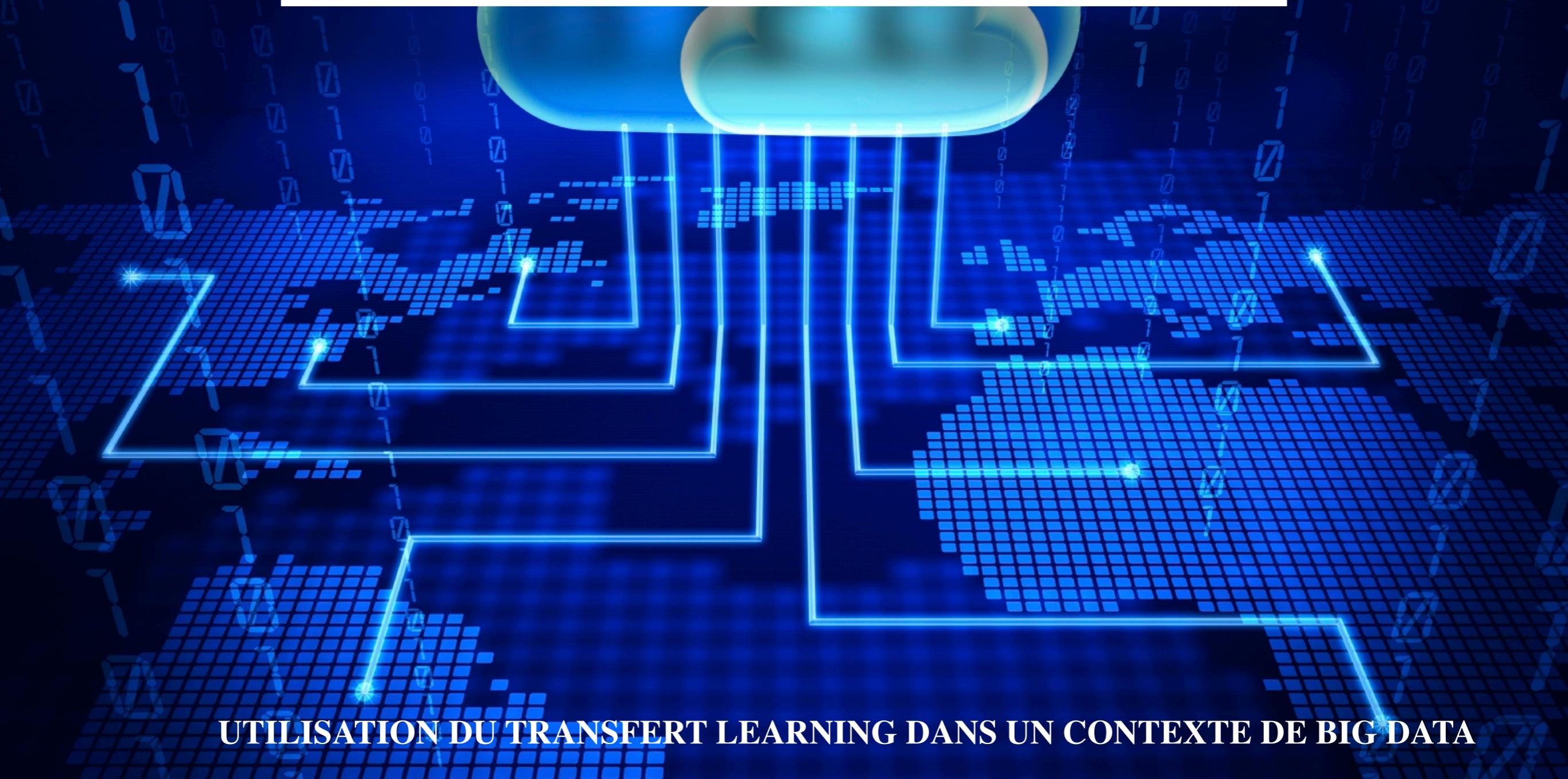


Déployez un modèle dans le cloud



UTILISATION DU TRANSFERT LEARNING DANS UN CONTEXTE DE BIG DATA

SOMMAIRE

PROBLÉMATIQUE

PRÉSENTATION
DES DONNÉES

PREPROCESSING

EXTRACTION
DE FEATURES
ET
RÉDUCTION
DE DIMENSION

EXÉCUTION
DU CODE :
LOCAL ET CLOUD



Fruits!

CONTEXTE GÉNÉRAL

Contexte :

La start-up **AgriTech** tente de développer des robots cueilleurs capables de reconnaître les fruits. En première approche, l'entreprise souhaite **mettre en ligne une application de reconnaissance de fruit**.

Le volume de données peut devenir très important, requérant une architecture spécifique du Big Data.

Un alternant a réalisé un premier script (notebook) posant les bases d'une classification dans un contexte Big Data.

Mission :

1. Vérifier le code de l'alternant en local.
2. Ajouter une étape de standardisation et réduction de dimension (PCA) dans un contexte de calculs distribués (SPARK).
3. Création d'un environnement Cloud (AWS) pour déployer le code.

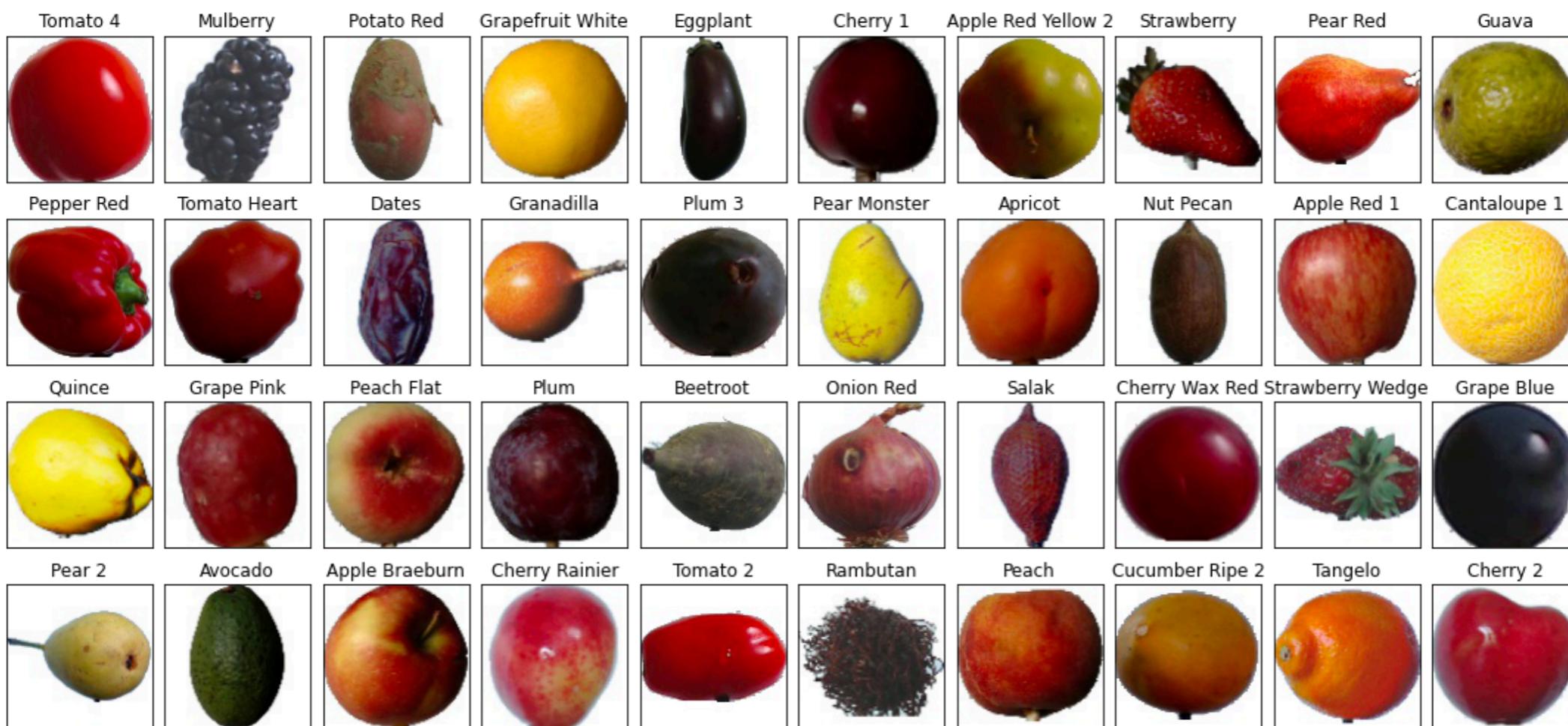


Fruits!

- Jeu de données **Fruits360** provenant de Kaggle: <https://www.kaggle.com/datasets/moltean/fruits/data>
- Données d'entraînement : **67692 images**.
- Données supplémentaires (test) : **22688 images**.
- **131 variétés** de fruits et légumes.
- Image size: **100x100 pixels**, le format **.jpg**

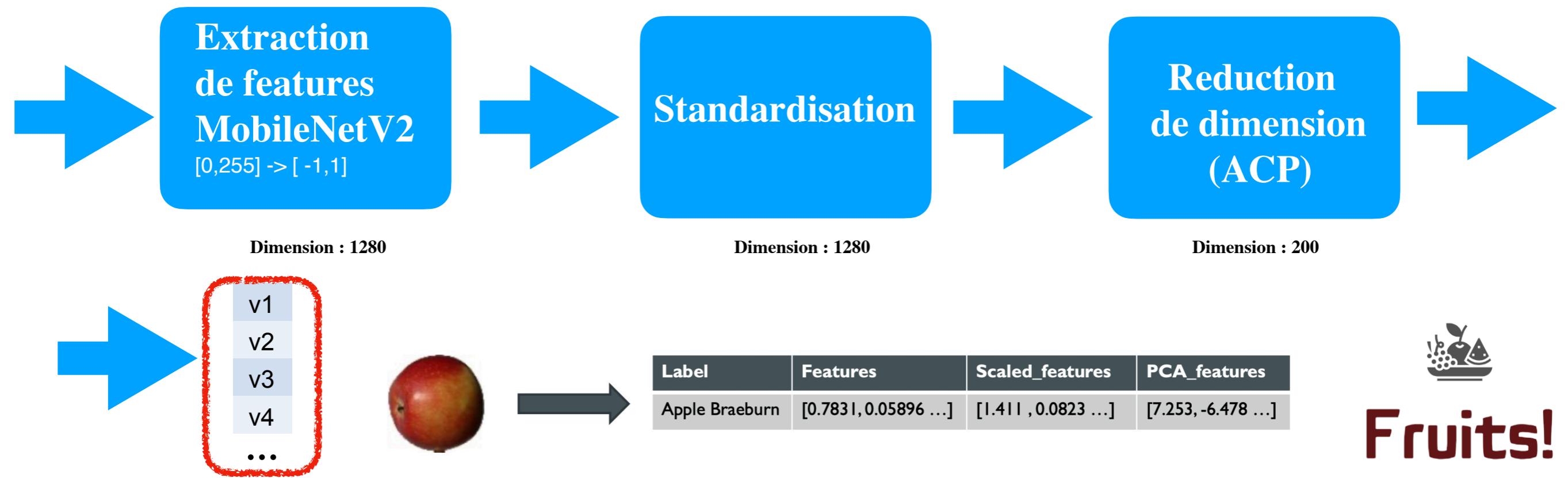
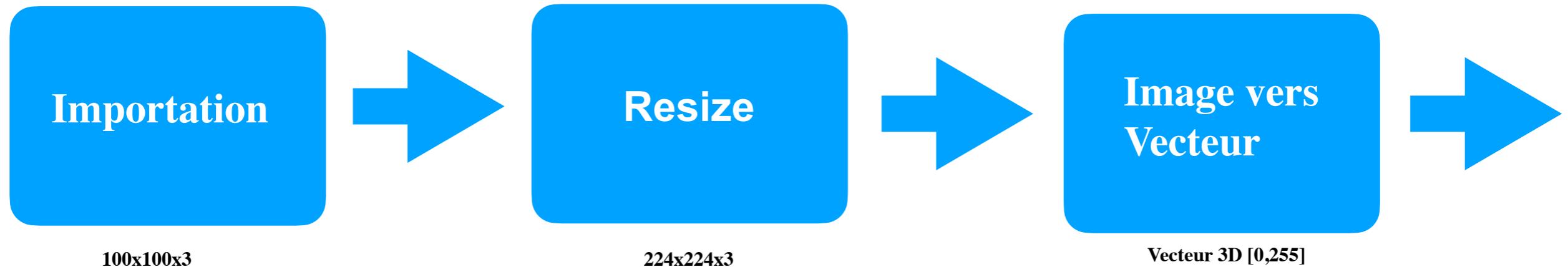
Objectif :

Utiliser la méthode de Transfert Learning afin d'extraire des « features » de ces images (*prémices d'une classification*).



Fruits!

Transformation des données



EXTRACTION DE FEATURES ET RÉDUCTION DE DIMENSION

MobileNET V2

MobileNetV2 utilise une architecture spécialisée optimisée pour les appareils mobiles. Cela inclut l'utilisation de ce que l'on appelle les « inverted residuals » et les « linear bottlenecks», qui permettent au modèle de rester léger et rapide sans perte significative de précision.

Spécificités : Bottleneck Residual Blocks - Réduction du nombres de paramètres - Temps de calculs - Poids du modèle

Modèle Tensorflow

Transfer Learning

Dernière couche retirée (classification)

Images en 224x224

1280 features

Input	Operator	<i>t</i>	<i>c</i>	<i>n</i>	<i>s</i>
$224^2 \times 3$	conv2d	-	32	1	2
$112^2 \times 32$	bottleneck	1	16	1	1
$112^2 \times 16$	bottleneck	6	24	2	2
$56^2 \times 24$	bottleneck	6	32	3	2
$28^2 \times 32$	bottleneck	6	64	4	2
$14^2 \times 64$	bottleneck	6	96	3	1
$14^2 \times 96$	bottleneck	6	160	3	2
$7^2 \times 160$	bottleneck	6	320	1	1
$7^2 \times 320$	conv2d 1x1	-	1280	1	1
$7^2 \times 1280$	avgpool 7x7	-	-	1	-
$1 \times 1 \times 1280$	conv2d 1x1	-	k	-	-

MobileNetV 2,257,984 1280



Le Big Data ?

Problématiques

- Grand nombre d'images à traiter.
- Opérations longues.
- Ressources matérielles locales limitées au niveau mémoire et capacités de calcul.
- Risques en cas de panne.

5V : Volume * Vitesse * Variété * Valeur * Véracité

Le calcul distribué

Problème : Comment traiter simultanément des millions d'images rapidement ?

1. Environnement de calculs distribués;
2. Architecture Cloud pour le passage à l'échelle.

• LA MÉTHODE MAPREDUCE

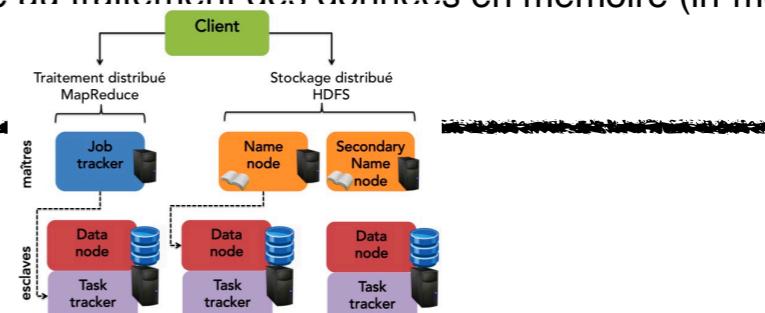
- **Traitement Distribué** : MapReduce est un modèle de programmation pour le traitement et la génération de grands ensembles de données en utilisant des algorithmes parallèles sur des clusters.
- **Deux Phases Principales** : Comprend deux phases principales - "Map" (mise en carte) et "Reduce" (réduction), chacune effectuant certaines opérations sur les données.
- **Scalabilité et Fiabilité** : Conçu pour une mise à l'échelle efficace d'un seul serveur à des milliers de serveurs.

• L'ENVIRONNEMENT HADOOP

- **Écosystème de Big Data** : Hadoop est un framework pour le stockage et le traitement de grandes quantités de données, comprenant divers composants tels que HDFS, YARN et MapReduce.
- **Stockage Distribué Fiable** : Particulièrement fort dans le stockage de grandes quantités de données (HDFS) avec une haute résilience aux pannes.
- **Fonctionnalités Étendues** : Prend en charge divers outils et langages (par exemple, Pig, Hive) pour le traitement des données.

• L'ENVIRONNEMENT SPARK

- **Traitement Haute Performance** : Spark est un moteur rapide et polyvalent pour le traitement de grandes quantités de données, supportant des tâches nécessitant une grande puissance de calcul.
- **Capacités Avancées** : Fournit des fonctionnalités supplémentaires (par exemple, Spark SQL, MLlib pour l'apprentissage automatique, GraphX).
- **Traitement Efficace en Mémoire** : Optimisé pour des calculs rapides, notamment grâce au traitement des données en mémoire (in-memory processing).



Fruits!

Frameworks BigData

Comparing Hadoop MapReduce and Spark



Fast

Batch Processing

Stores Data on Disk

Written in java

Low Cost

100x faster than MapReduce
in-memory processing

Real-time Processing

Stores Data Memory

Written in Scala

High Cost

Machine learning

Hadoop

Spark

Fonctionnalités ML
via l'intégration
de bibliothèques
externes.

Fonctionnalités ML
intégrées
(SparkML).

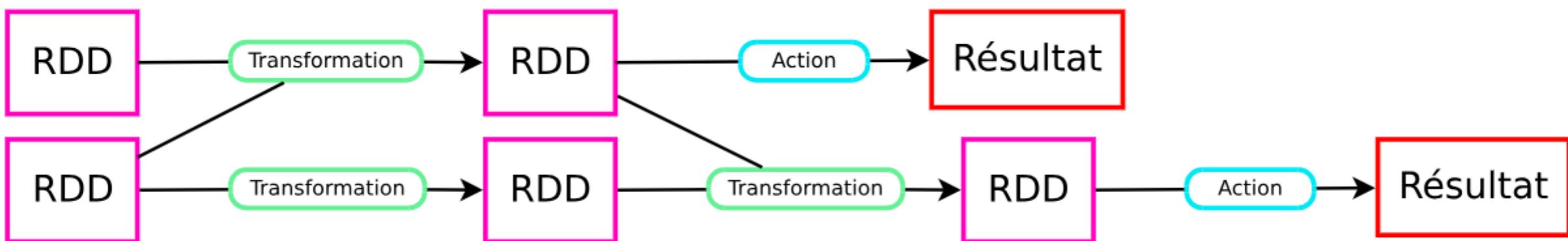
Possibilité d'utiliser Spark en
conjonction avec Hadoop pour
tirer parti du meilleur des 2
technologies



Fruits!

Solution retenue : SPARK

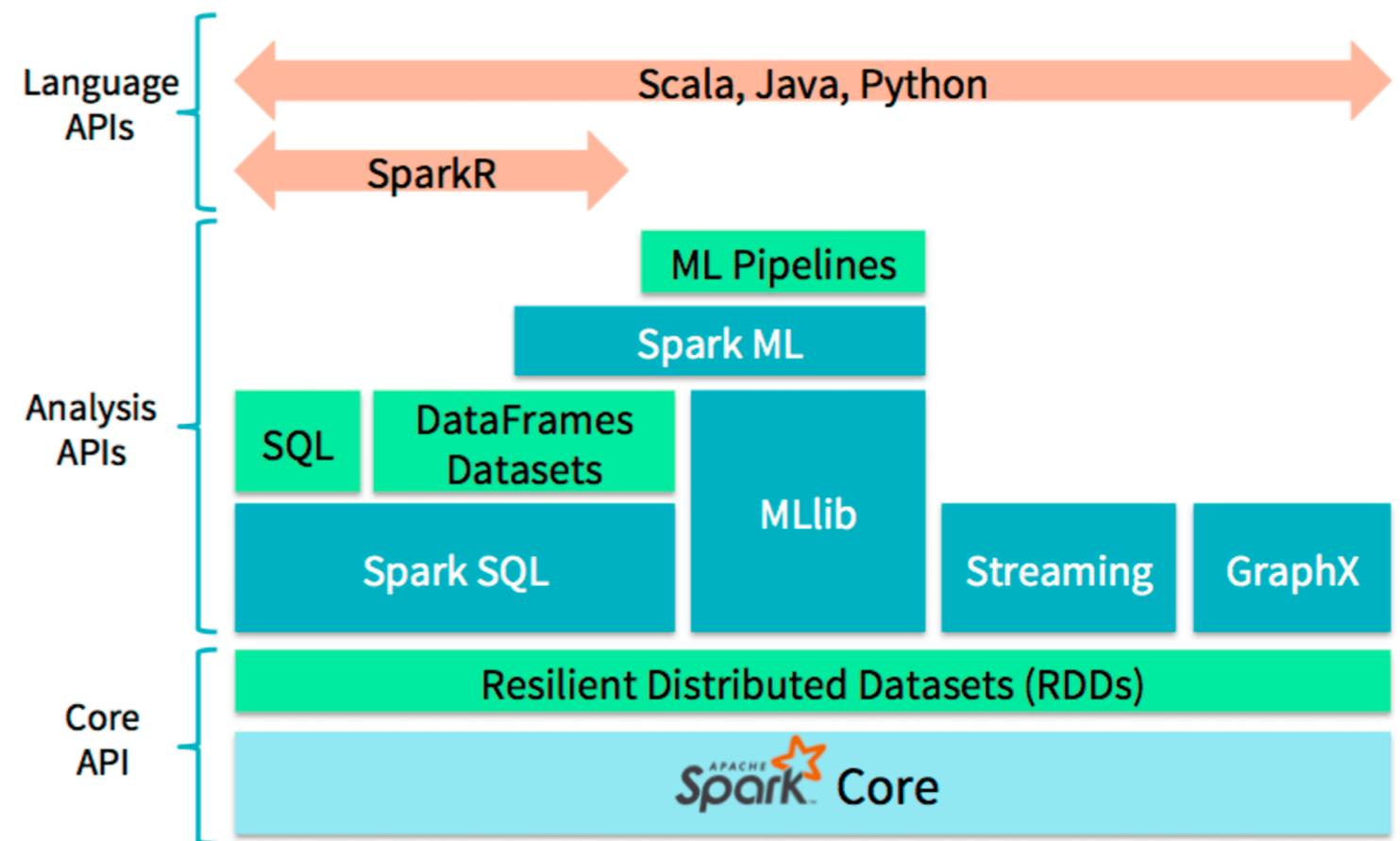
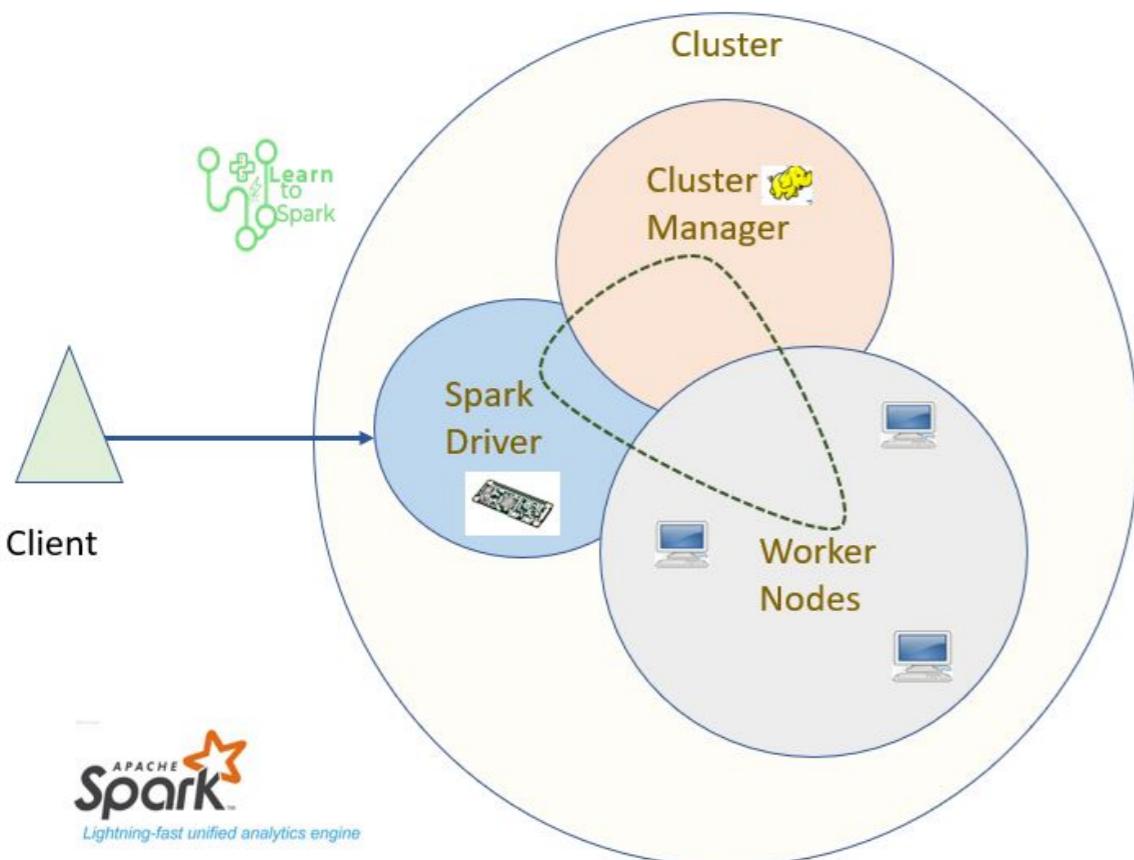
- **Rapidité via l'utilisation de la RAM**
- **Outils ML intégrés.**
- **Fonctionnement par :**
 - **Transformations** : permettent de construire un plan de transformation logique (**lazy évaluation**)
Les opérations sur les données ne sont effectuées qu'avant l'utilisation directe des résultats de ces opérations. Grâce à cela, la puissance de calcul n'est pas gaspillée pour des calculs qui seront nécessaires dans le futur.
 - **Actions** : pour déclencher le calcul, il faut utiliser une **action**. (visualisation, count, collect,...)
- **Traitement parallèle et opérations de combinaison** - distribue les données et les calculs sur plusieurs nœuds du cluster, effectuant différentes opérations de traitement en parallèle et en temps réel. Cela le distingue de MapReduce, dans lequel chaque étape ultérieure de travail avec un ensemble de données nécessite l'achèvement de la précédente.
- **Une résilience aux pannes** - le stockage des ensembles de données et des informations sur les transformations effectuées à la fois sur plusieurs nœuds du réseau de cluster. RDD permet à Spark de restaurer les données en cas de panne et d'optimiser les calculs.
- **Flexibilité** - prise en charge d'une variété de sources et de formats de données.



Graphe Orienté Acyclique utilisé dans l'ordonnancement des tâches



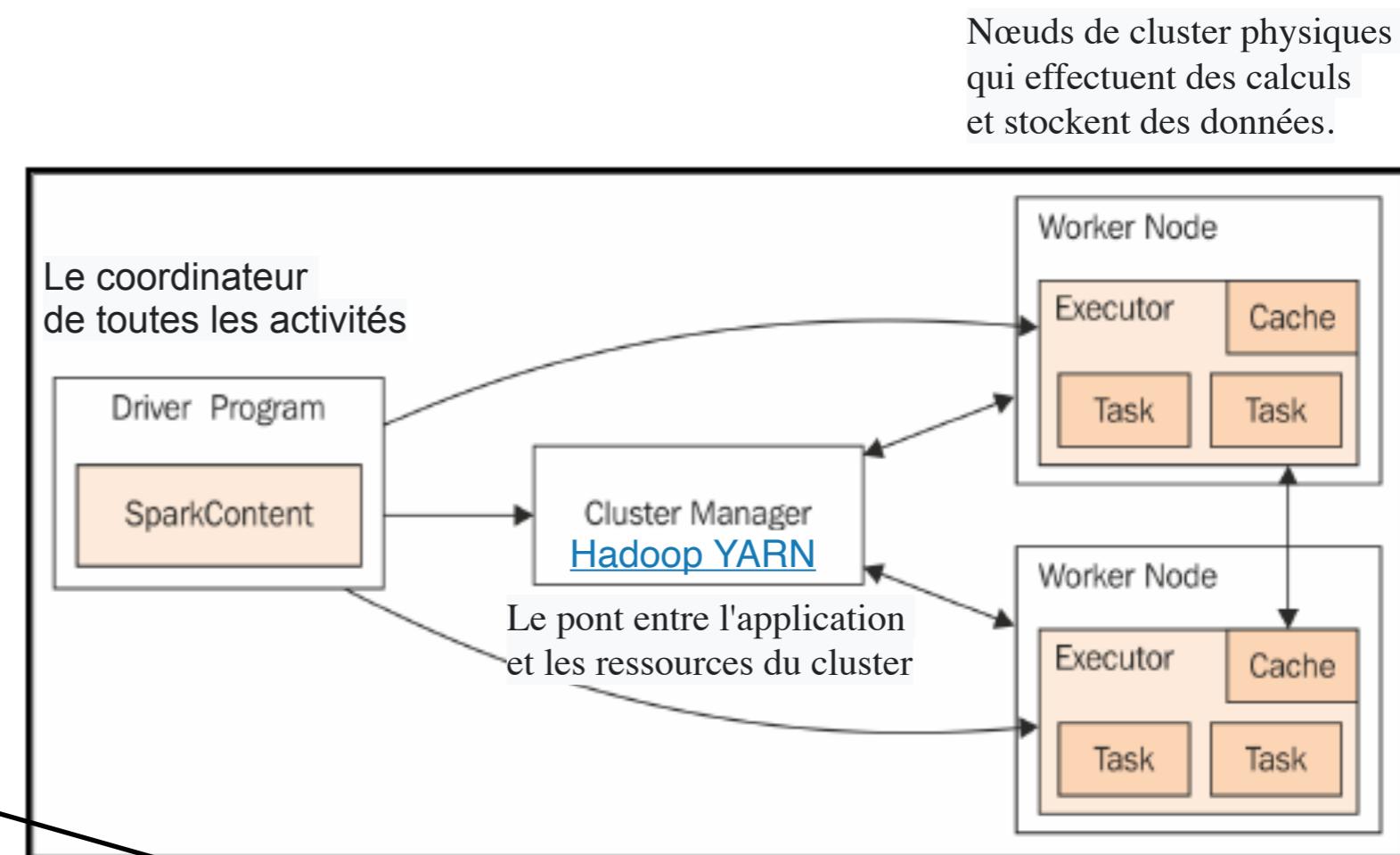
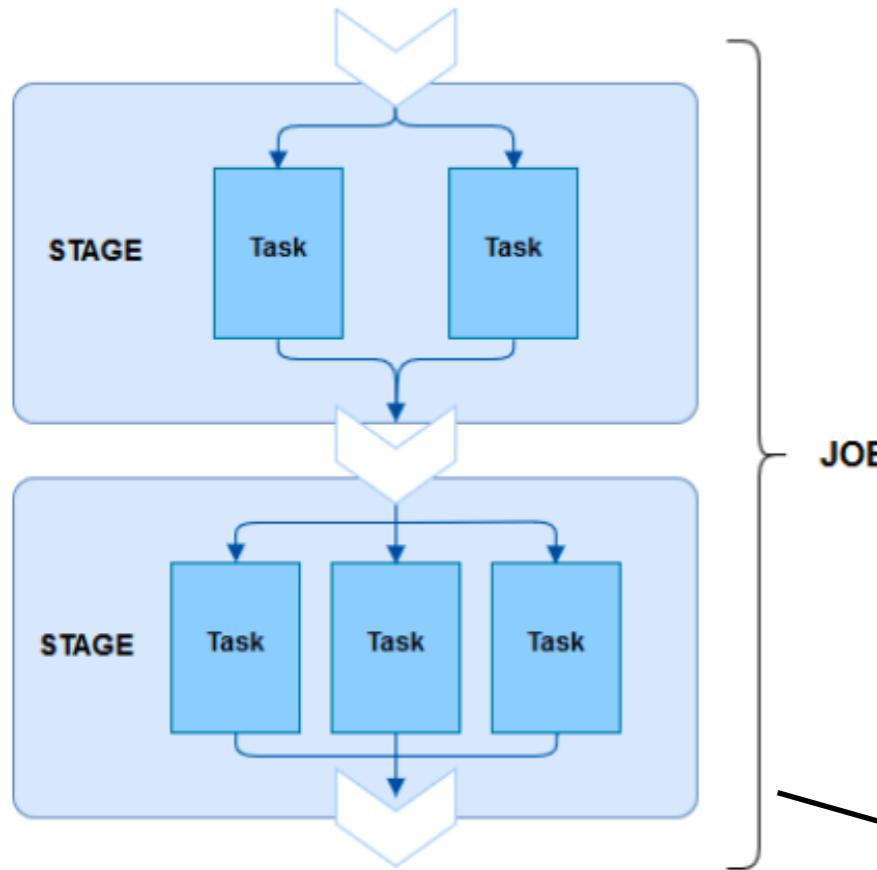
Caractéristiques Spark



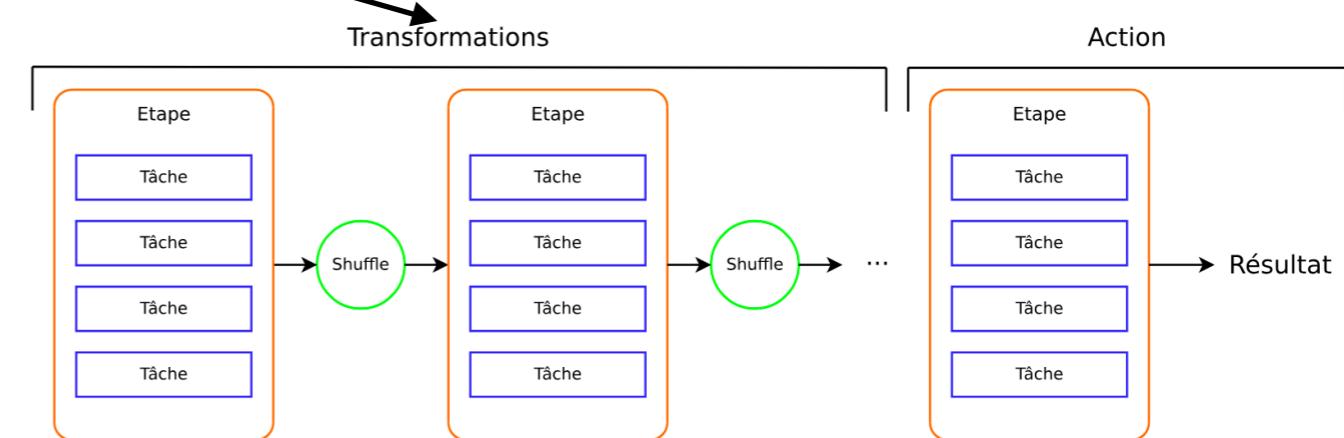
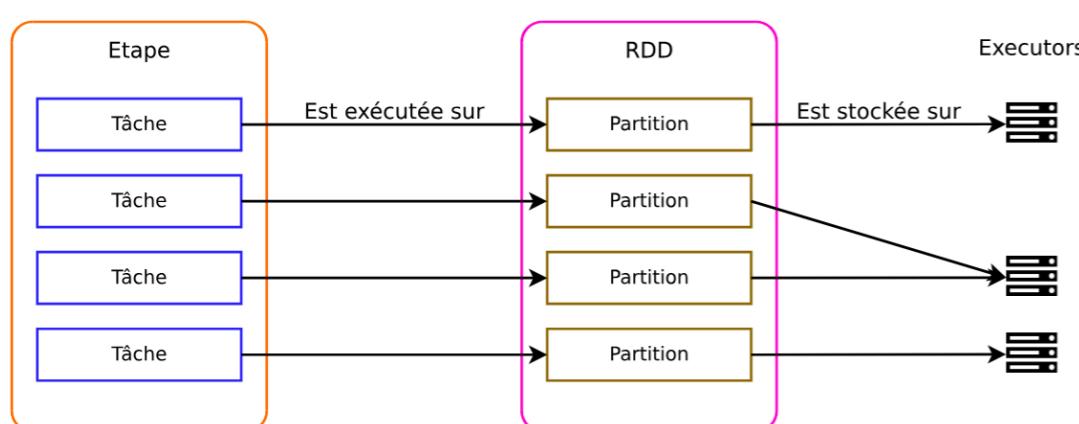
Fruits!

Architecture dans Spark

Chaque **étape** est composée de **tâche**



- Chaque tâche s'exécute sur une **partition** différente des données.

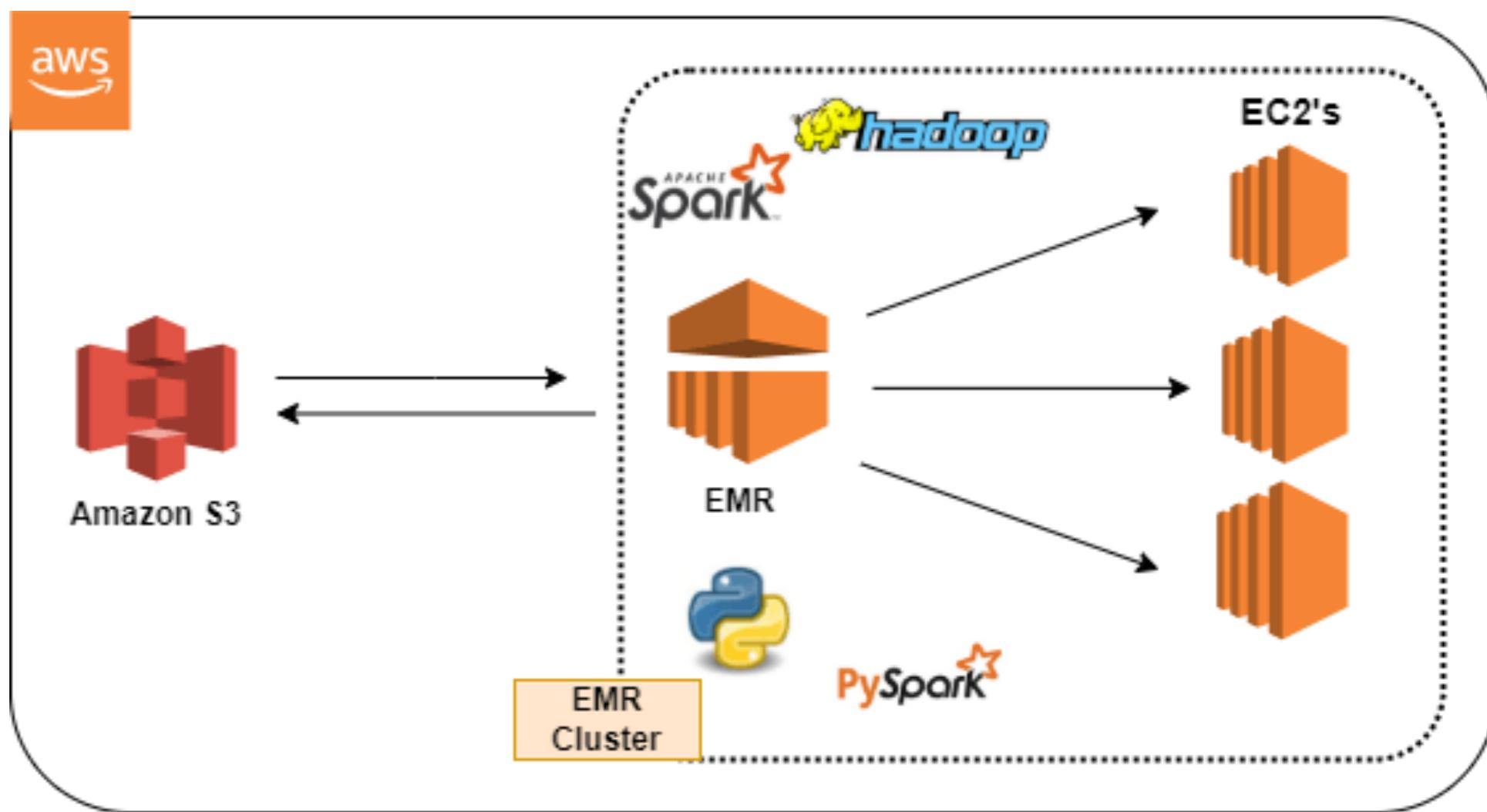


- Les partitions sont réparties sur les différents **executors**.
- Les partitions sont créées par les **Resilient Distributed Datasets (RDD)**.
- Un **job** Spark correspond à une action sur un RDD et est composé de plusieurs **étapes** séparées par des **shuffles**.

Fruits!

Environnement AWS EMR

- S3 : service de stockage persistant et sécurisé
- EMR : service de création de clusters Spark(PAAS)
- EC2 : service de machines virtuelles (IAAS)

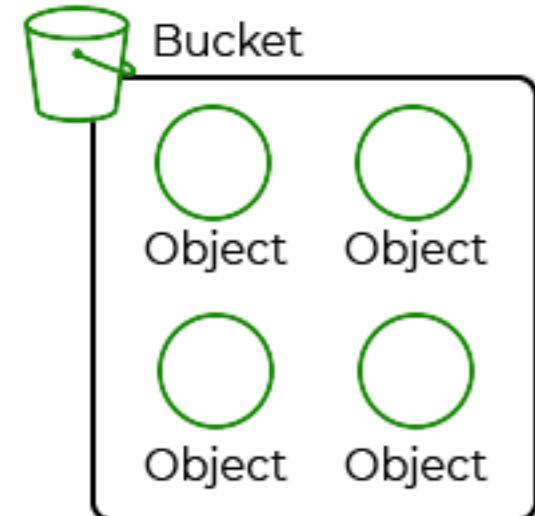
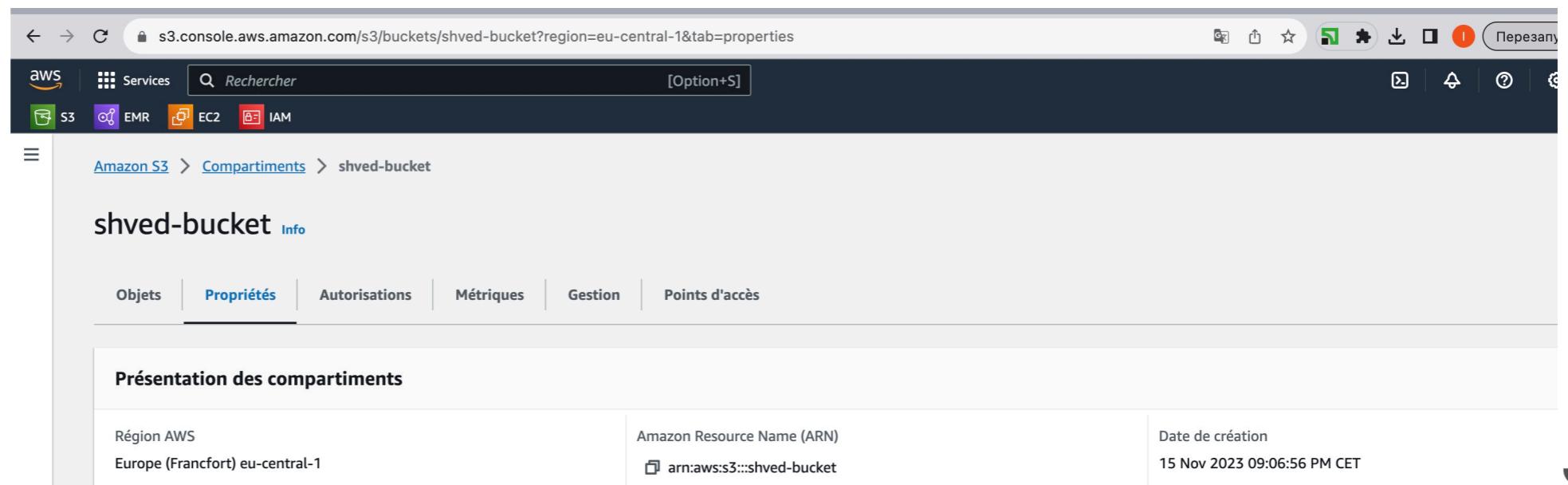


Fruits!

L'ESPACE DE STOCKAGE CLOUD (bucket) S3

- Création du Bucket S3 via la console AWS (interface web)
- Création d'un utilisateur IAM (administrator)
- Configuration de l'Interface en Lignes de Commande (CLI) d'AWS
- Chargement des fichiers par CLI sur le bucket S3

Compartiment (bucket) situé en Allemagne



Fruits!

Création du cluster EMR

Instances EC2 situées en Allemagne

Cluster basé sur Spark

Choisissez une méthode de configuration pour les groupes de nœuds primaires, principaux et de tâches de votre cluster.

Groupes d'instances

Choisir un type d'instance par groupe de nœuds

Flottes d'instances

Choisir une combinaison de types d'instance au sein de chaque groupe de nœuds

Groupes d'instances

Primaire

Choisir un type d'instance EC2

m5.xlarge

4 vCore 16 GiB mémoire EBS uniquement stockage
Prix à la demande : 0.230 USD par instance/heure
Prix Spot le plus bas : \$0.075 (eu-central-1b)

Actions ▾

Utiliser plusieurs nœuds primaires

Pour améliorer la disponibilité du cluster, utilisez trois nœuds primaires avec les mêmes actions de configuration et d'amorçage.
Vous ne pouvez pas utiliser plusieurs nœuds primaires avec des flottes d'instances.

▶ Configuration de nœud - facultatif

Unité principale

Choisir un type d'instance EC2

m5.xlarge

4 vCore 16 GiB mémoire EBS uniquement stockage
Prix à la demande : 0.230 USD par instance/heure
Prix Spot le plus bas : \$0.075 (eu-central-1b)

Actions ▾

▶ Configuration de nœud - facultatif

Tâche 1 sur 1

Nom

Tâche 1

Retirer le groupe d'instances

Créer un cluster Info

Nom et applications Info

Nom

shved-cluster

Version Amazon EMR Info

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

emr-6.14.0

Offre d'applications



- Flink 1.17.1
- HCatalog 3.1.3
- Hue 4.11.0
- Livy 0.7.1
- Phoenix 5.1.3
- Spark 3.4.1
- Tez 0.10.2
- ZooKeeper 3.5.10

- Ganglia 3.7.2
- Hadoop 3.3.3
- JupyterEnterpriseGateway 2.6.0
- MXNet 1.9.1
- Pig 0.17.0
- Swoop 1.4.7
- Trino 422

- HBase 2.4.17
- Hive 3.1.3
- JupyterHub 1.5.0
- Oozie 5.2.1
- Presto 0.281
- TensorFlow 2.11.0
- Zeppelin 0.10.1

Paramètres du catalogue de données AWS Glue

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

- Utiliser pour les métadonnées de table Spark

Options du système d'exploitation Info

- Version Amazon Linux :
- Amazon Machine Image (AMI) personnalisée



Fruits!

Amorçage du cluster

Chargement des logiciels réalisé sur tous les membres du cluster via un fichier d'amorçage

C eu-central-1.console.aws.amazon.com/emr/home?region=eu-central-1#/createCluster

Services Rechercher [Option+S]

EMR EC2 IAM

Résilier automatiquement le cluster après le temps d'inactivité (Recommandé)

Temps d'inactivité
Saisissez la durée avant la résiliation de votre cluster.

1 jour ▾ 01:00:00
Choisissez une durée supérieure à 1 minute (00:01:00) et inférieure à 7 jours. L'heure est au format hh:mm:ss (24 heures).

Utiliser la protection contre la résiliation
Protégez vos instances EC2 de la résiliation accidentelle.

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

Actions d'amorçage (1)

Supprimer Modifier Ajouter

Nom	Emplacement Amazon S3	Arguments
amorçage	s3://shved-bucket/bootstrap-emr.sh	-

Journaux de clusters – facultatif Info

(i) Nous archivons automatiquement vos fichiers journaux sur Amazon S3. Vous pouvez spécifier votre propre emplacement S3 ou utiliser l'emplacement S3 par défaut pour Amazon EMR. L'emplacement des journaux par défaut est prérenseigné dans le champ Emplacement Amazon S3.

Publier les journaux spécifiques au cluster sur Amazon S3

Emplacement Amazon S3

s3://aws-logs-429430768867-eu-central-1/elasticmapreduc

Format : utiliser s3://bucket/prefix

Chiffrer les journaux spécifiques au cluster

jupyter bootstrap-emr.sh

Минулої неділі о 14:23

File Edit View Language

```
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=2.0.0
7 sudo python3 -m pip install pyarrow
8 sudo python3 -m pip install boto3
9 sudo python3 -m pip install tensorflow
10
```



EMR et EC2

zon EMR > EMR sur EC2: Clusters

Clusters (1) Info

Filtrer les clusters par statut ▾ Rechercher des clusters Filtrer les clusters par date et heure de création < 1 > ⚙️

ID de cluster	Nom du cluster	Statut	Heure de création (UTC+01:00)	Temps écoulé	He no
j-32FL2XRAUMGGR	shved-cluster	En attente Étapes prêtes à exécuter	16 novembre 2023 02:06	10 minutes, 23 secondes	0

eu-central-1.console.aws.amazon.com/ec2/home?region=eu-central-1#Instances:

Services Rechercher [Option+S] Francfort InShed

EMR EC2 IAM

Ordre EC2 Instances (1/3) Informations Lancer des instances

Rechercher Instance par attribut ou identification (case-sensitive) < 1 > ⚙️

Name	ID d'instance	État de l'instance	Type d'insta...	Contrôle d...	Statut d'alar...	Zone de dispon...	DNS IPv4 public	Adresse IPv...
<input type="checkbox"/>	i-01e55e9792bdccaca0	En cours d'exéc...	m5.xlarge	2/2 vérificati	View alarms +	eu-central-1a	ec2-3-71-3-72.eu-centr...	3.71.3.72
<input type="checkbox"/>	i-0316a3c97c4ee529c	En cours d'exéc...	m5.xlarge	2/2 vérificati	View alarms +	eu-central-1a	ec2-18-184-220-21.eu-...	18.184.220.
<input checked="" type="checkbox"/>	i-000038396b107fed6	En cours d'exéc...	m5.xlarge	2/2 vérificati	View alarms +	eu-central-1a	ec2-3-70-177-235.eu-c...	3.70.177.23

eu-central-1.console.aws.amazon.com/emr/home?region=eu-central-1#/clusterDetails/j-H8GCCXF3VLW8

Services Rechercher [Option+S] Francfort InShed

EMR EC2 IAM AWS CloudFormation VPC Resource Groups & Tags Editor

Mise à jour il y a moins d'une minute		Résilier	Cloner dans AWS CLI	Cloner
<p>tion des clusters</p> <p>ination des journaux dans Amazon S3 -logs-429430768867-eu-central-1 asticmapreduce</p> <p>surfaces utilisateur d'application persistantes eur d'historique Spark </p>	<p>Statut et heure</p> <p>Statut</p> <p> Résilié</p> <p>Heure de création</p> <p>16 novembre 2023 16:44 (UTC+01:00)</p>			

Accès aux applications du serveur EMR via le tunnel ssh

```
Відновлено 16 лист. 2023 р., 22:06:22
Last login: Thu Nov 16 21:53:55 on console
Restored session: чт 16 лис 2023 21:08:07 CET
(base) innakonar@Innas-MacBook-Pro ~ % ssh -i /Users/innakonar/Desktop/Stas11/my_keys.pem -l
Last login: Thu Nov 16 20:46:31 2023
      _#
 ~\_ #####      Amazon Linux 2
 ~~ \#####\_
 ~~ \|##|      AL2 End of Life is 2025-06-30.
 ~~ \|#/ ___
 ~~ \~' '-->
 ~~~ /      A newer version of Amazon Linux is available!
 ~~-. /_
 _/_/      Amazon Linux 2023, GA and supported until 2028-03-15.
 _/m/      https://aws.amazon.com/linux/amazon-linux-2023/
16 package(s) needed for security, out of 23 available
Run "sudo yum update" to apply all updates.

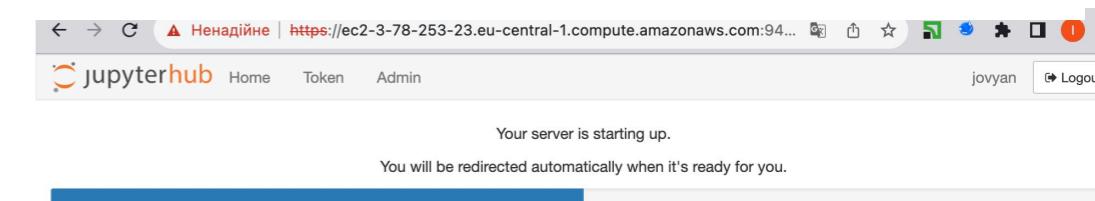
EEEEEEEEEEEEEEEEE MMMMMMM RRRRRRRRRRRRRRR
E:::::::::::E M:::::M M:::::M R:::::R:::::R
EE:::::E EEEEEEE M:::::M M:::::M R:::::RRRRRR:::::R
E:::::E M:::::M M:::::M M:::::M R:::R R:::::R
E:::::E EEEEEEEEEE M:::::M M:::::M M:::::M R:::RRRRRR:::::R
E:::::E EEEEEEEEEE M:::::M M:::::M M:::::M R:::::RRRRRR:::::R
E:::::E M:::::M M:::::M M:::::M R:::R R:::::R
E:::::E EEEEE M:::::M MMM M:::::M R:::R R:::::R
EE:::::E EEEEEEEEEE M:::::M M:::::M R:::R R:::::R
E:::::E M:::::M RR:::::R R:::::R
EEEEEEEEEEEEEEEEE MMMMMMM RRRRRRR
[.hadoop@ip-10-0-20-116 ~]$ [hadoop@ip-10-0-20-116 ~]$
```



Sign in

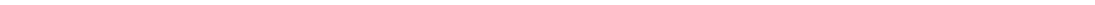
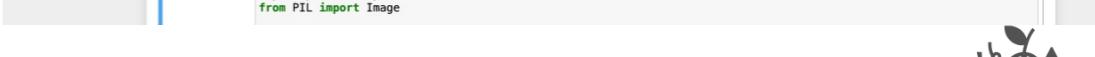
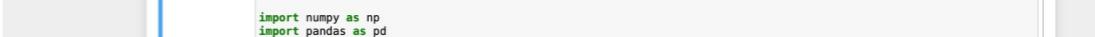
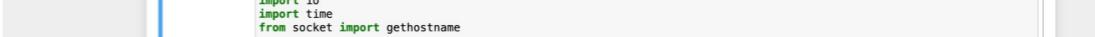
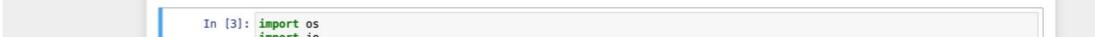
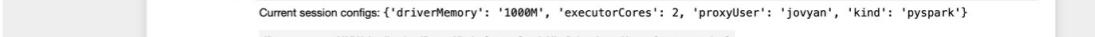
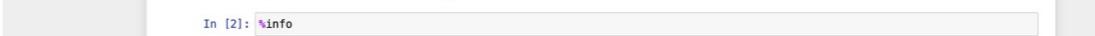
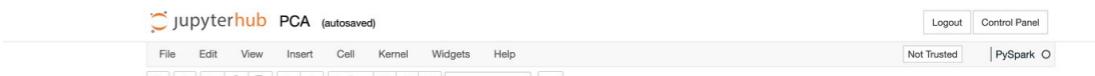
Username:

Password:



Event log

ид the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.



Implementation

Utilisation d'un pipeline SparkML avec des transformers:

- Récupération du label à partir du chemin du fichier (PathToLabelTransformer);
- Transformation de l'image en features (ImageFeatureTransformer);
- Transformation du tableau de features en vecteur de features (ArrayToVectorTransformer);
- Transformation du label en index numérique;
- MinMaxScaler;
- PCA.



Exécution locale

APACHE Spark 3.5.0 Jobs Stages Storage Environment Executors SQL / DataFrame FruitScout application U

Spark Jobs (?)

User: innakonar
Total Uptime: 2,2 h
Scheduling Mode: FIFO
Completed Jobs: 20

Event Timeline Enable zooming

Executors: Added (blue), Removed (red). Jobs: Succeeded (blue), Failed (red), Running (green).

Timeline events:

- Executor driver added at 00:10
- first at MinMaxScaler.scala:120 (Job 10) at 00:25
- treeAggregate at Statistics.scala:58 (Job 14) at 00:45
- treeAggregate at RowMatrix.scala:171 (Job 1) at 01:10
- parquet at NativeMethodAccessorImpl.java:0 (Job 19) at 01:35

00:10 00:20 00:30 00:40 00:50 01:00 01:10 01:20 01:30 01:40 01:50 02:00 02:10 02:20

Thu 16 November

Completed Jobs (20)

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

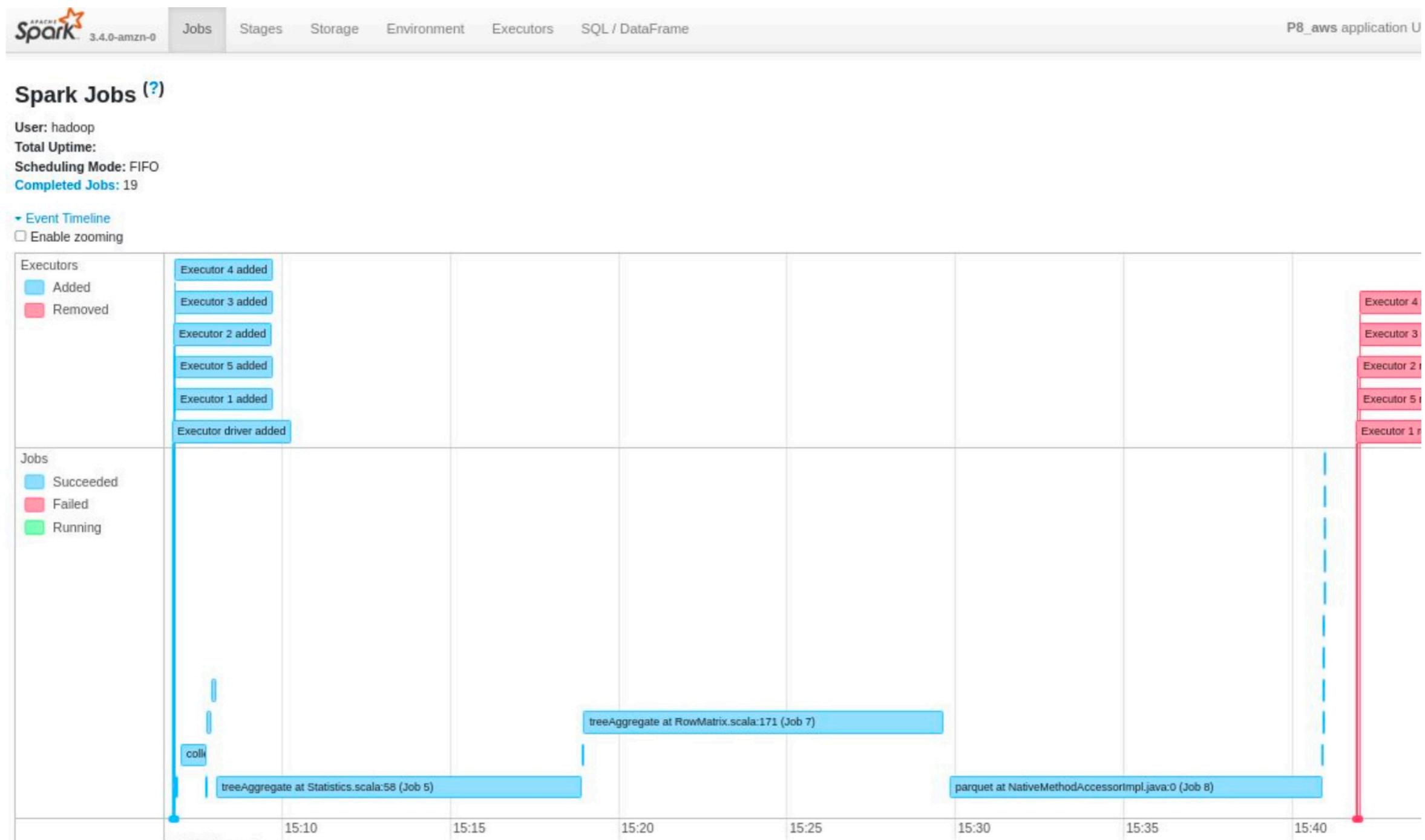
Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
19	parquet at NativeMethodAccessorImpl.java:0 parquet at NativeMethodAccessorImpl.java:0	2023/11/16 01:34:13	47 min	1/1	709/709

- 1 nœud 16Gb
- 4 cœurs utilisés
- 2,2h



Fruits!

Exécution sur AWS



-1 noeud primaire

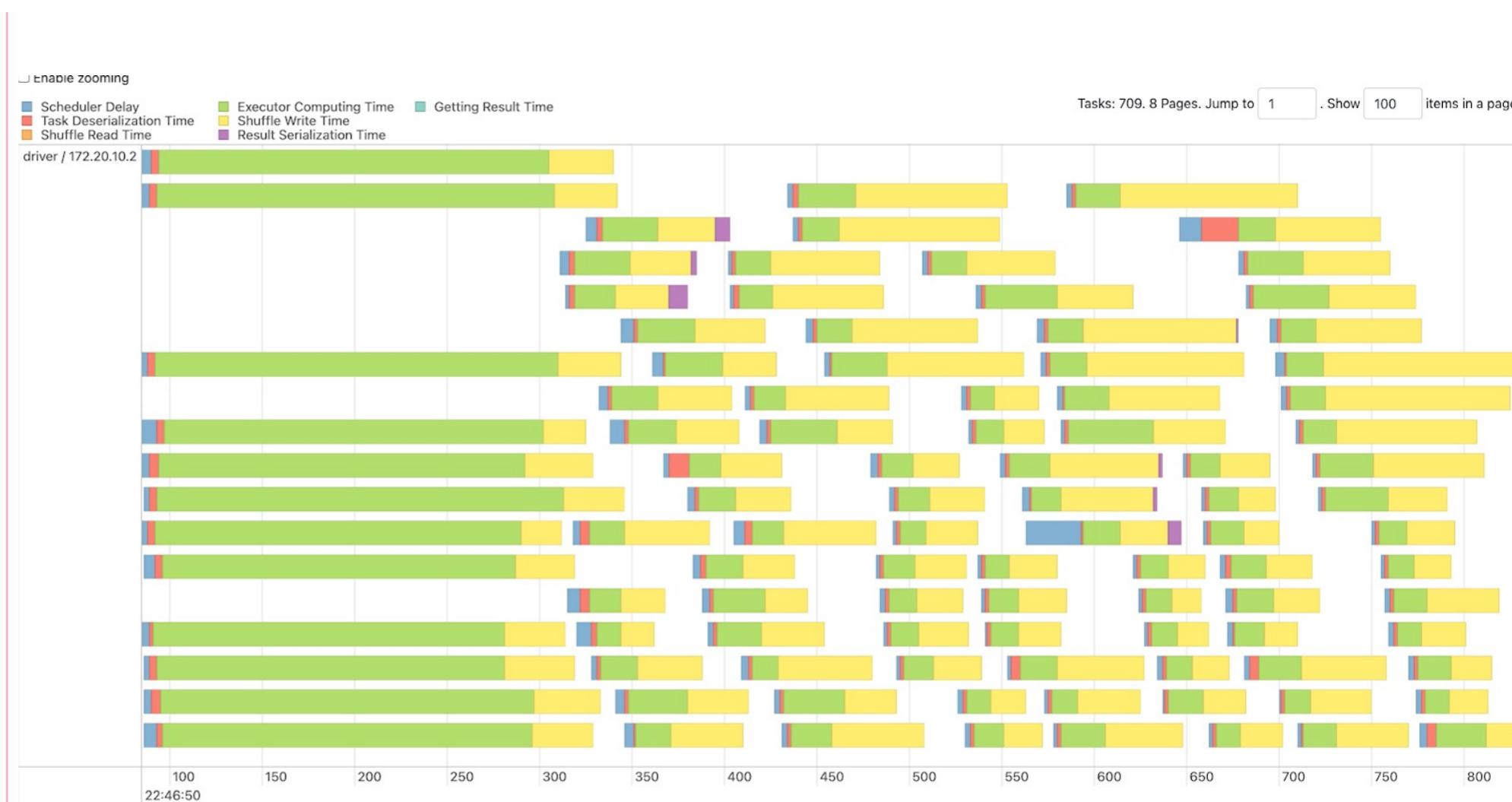
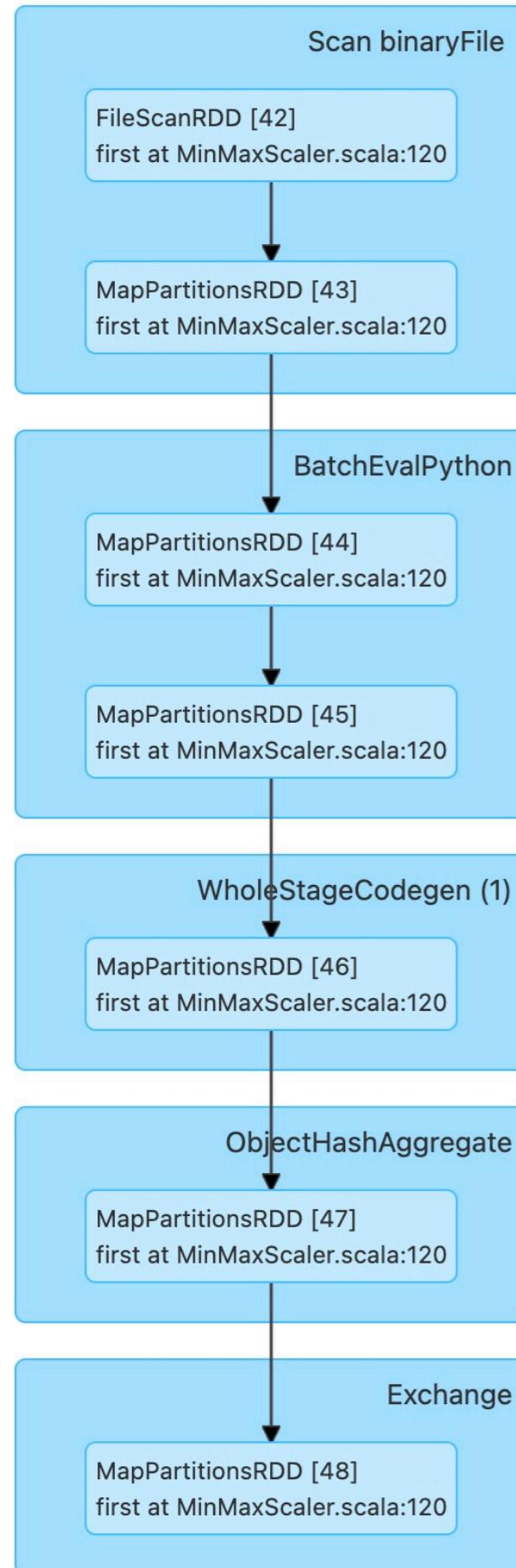
-3 instances de 4 coeurs + 16 Gb pour les workers

-41 mn



Fruits!

Stage 13



Fruits!

Stages for All Jobs

Completed Stages: 22

Skipped Stages: 5

Completed Stages (22)

Page: 1

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

Stage Id ▾	Description		Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
26	parquet at NativeMethodAccessImpl.java:0	+details	2023/11/16 01:34:13	47 min	709/709	3.1 MiB	665.4 MiB		
25	transform at /var/folders/0g/rj_1t2x52bz2m5m552wspnkc0000gn/T/ipykernel_1189/4212335374.py:2+details	+details	2023/11/16 01:34:07	0,5 s	1/1			716.4 KiB	
23	javaToPython at NativeMethodAccessImpl.java:0	+details	2023/11/16 01:34:02	6 s	709/709				716.4 KiB
22	treeAggregate at RowMatrix.scala:171	+details	2023/11/16 01:33:47	5 s	26/26			4.3 GiB	
21	treeAggregate at RowMatrix.scala:171	+details	2023/11/16 01:11:01	23 min	709/709	3.1 MiB			4.3 GiB
20	isEmpty at RowMatrix.scala:441	+details	2023/11/16 01:10:52	9 s	1/1	7.2 KiB			
19	treeAggregate at Statistics.scala:58	+details	2023/11/16 01:10:52	0,2 s	26/26			9.9 MiB	
18	treeAggregate at Statistics.scala:58	+details	2023/11/16 00:48:37	22 min	709/709	3.1 MiB			9.9 MiB
17	first at RowMatrix.scala:62	+details	2023/11/16 00:48:31	6 s	1/1	7.2 KiB			
16	first at PCA.scala:44	+details	2023/11/16 00:48:25	6 s	1/1	7.2 KiB			
15	first at MinMaxScaler.scala:120	+details	2023/11/16 00:48:24	0,7 s	1/1			12.8 MiB	
13	first at MinMaxScaler.scala:120	+details	2023/11/16 00:25:03	23 min	709/709	3.1 MiB			12.8 MiB
12	fit at /var/folders/0g/rj_1t2x52bz2m5m552wspnkc0000gn/T/ipykernel_1189/4212335374.py:1	+details	2023/11/16 00:25:01	0,1 s	1/1			716.4 KiB	
10	javaToPython at NativeMethodAccessImpl.java:0	+details	2023/11/16 00:24:58	4 s	709/709				716.4 KiB
9	showString at NativeMethodAccessImpl.java:0	+details	2023/11/16 00:24:19	0,2 s	1/1	78.5 KiB			
8	Listing leaf files and directories for 131 paths: file:/Users/innakonar/Desktop/Projet8/Test/Tomato 4, ... load at NativeMethodAccessImpl.java:0	+details	2023/11/16 00:24:15	0,7 s	131/131				
7	fit at /var/folders/0g/rj_1t2x52bz2m5m552wspnkc0000gn/T/ipykernel_1189/4212335374.py:1	+details	2023/11/16 00:23:42	0,1 s	1/1			716.4 KiB	



Fruits!

Fichiers résultats

Amazon S3 > Compartiments > shved-bucket

shved-bucket [Info](#)

[Objets](#) [Propriétés](#) [Autorisations](#) [Métriques](#) [Gestion](#) [Points d'accès](#)

Objets (9)

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](#)

<input type="checkbox"/>	Nom	Type	Dernière modification	Taille	Classe de stockage
<input type="checkbox"/>	bootstrap-emr.sh	sh	17 Nov 2023 04:19:03 PM CET	426.0 o	Standard
<input type="checkbox"/>	emr_config.json	json	17 Nov 2023 01:38:19 PM CET	165.0 o	Standard
<input type="checkbox"/>	jupyter/	Dossier	-	-	-
<input type="checkbox"/>	PCA.ipynb	ipynb	17 Nov 2023 05:32:49 PM CET	21.1 Ko	Standard
<input type="checkbox"/>	pipeline_model/	Dossier	-	-	-
<input type="checkbox"/>	Result_PCA.csv	csv	18 Nov 2023 01:28:48 AM CET	1.3 Go	Standard
<input type="checkbox"/>	Results_PCA/	Dossier	-	-	-
<input type="checkbox"/>	Results/	Dossier	-	-	-
<input type="checkbox"/>	Test/	Dossier	-	-	-

[Copier l'URI S3](#) [Copier l'URL](#) [Télécharger](#) [Ouvrir](#) [Supprimer](#) [Actions](#) [Créer un dossier](#) [Charger](#)

Rechercher des objets en fonction du préfixe

Amazon S3 > Compartiments > shved-bucket > pipeline_model/

pipeline_model/

[Objets](#) [Propriétés](#)

Objets (2)

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](#)

<input type="checkbox"/>	Nom	Type	Dernière modification	Taille	Classe de stockage
<input type="checkbox"/>	metadata/	Dossier	-	-	-
<input type="checkbox"/>	stages/	Dossier	-	-	-

[Copier l'URI S3](#)

Rechercher des objets en fonction du préfixe

Amazon S3 > Compartiments

► Instantané de compte

Storage Lens offre une visibilité sur l'utilisation du stockage et les tendances d'activité. [En savoir plus](#)

[Afficher le tableau de bord de Storage Lens](#)

Compartiments (3) [Info](#)

Les compartiments sont des conteneurs pour les données stockées dans S3. [En savoir plus](#)

[Rechercher des compartiments par nom](#)

<input type="checkbox"/>	Nom	Région AWS	Accéder	Date de création
<input type="radio"/>	aws-emr-studio-429430768867-eu-central-1	Europe (Francfort) eu-central-1	Compartiment et objets non publics	16 Nov 2023 06:20:44 PM CET
<input type="radio"/>	aws-logs-429430768867-eu-central-1	Europe (Francfort) eu-central-1	Compartiment et objets non publics	16 Nov 2023 02:06:03 AM CET
<input type="radio"/>	shved-bucket	Europe (Francfort) eu-central-1	Compartiment et objets non publics	15 Nov 2023 09:06:56 PM CET

[Copier l'ARN](#) [Vider](#) [Supprimer](#) [Créer un compartiment](#)

Amazon S3 > Compartiments > shved-bucket > Results_PCA/

Results_PCA/

[Objets](#) [Propriétés](#)

Objets (710)

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](#)

<input type="checkbox"/>	Nom	Type	Dernière modification	Taille	Classe de stockage
<input type="checkbox"/>	_SUCCESS	-	18 Nov 2023 12:17:31 AM CET	0 o	Standard
<input type="checkbox"/>	part-00000-f32e340c-6dc9-4cd-b68c-7ed23aea4ec-c000.snappy.parquet	parquet	18 Nov 2023 12:32:05 AM CET	251.6 Ko	Standard
<input type="checkbox"/>	part-00001-f32e340c-6dc9-4cd-b68c-7ed23aea4ec-c000.snappy.parquet	parquet	18 Nov 2023 12:39:06 AM CET	251.6 Ko	Standard
<input type="checkbox"/>	part-00002-f32e340c-6dc9-4cd-b68c-7ed23aea4ec-c000.snappy.parquet	parquet	18 Nov 2023 12:39:07 AM CET	251.6 Ko	Standard
<input type="checkbox"/>	part-00003-f32e340c-6dc9-4cd-b68c-7ed23aea4ec-c000.snappy.parquet	parquet	18 Nov 2023 12:39:08 AM CET	251.6 Ko	Standard
<input type="checkbox"/>	part-00004-f32e340c-6dc9-4cd-b68c-7ed23aea4ec-c000.snappy.parquet	parquet	18 Nov 2023 12:39:09 AM CET	251.6 Ko	Standard

[Copier l'URI S3](#) [Copier l'URL](#) [Télécharger](#) [Ouvrir](#) [Supprimer](#) [Actions](#) [Créer un dossier](#) [Charger](#)

Rechercher des objets en fonction du préfixe

Amazon S3 > Compartiments > shved-bucket > Results/

Results/

[Objets](#) [Propriétés](#)

Objets (710)

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](#)

<input type="checkbox"/>	Nom	Type	Dernière modification	Taille	Classe de stockage
<input type="checkbox"/>	_SUCCESS	-	18 Nov 2023 12:06:19 AM CET	0 o	Standard
<input type="checkbox"/>	part-00000-4217ee94-d0f2-497f-90f-0a4c0d4704d1-c000.snappy.parquet	parquet	18 Nov 2023 12:06:19 AM CET	1.1 Mo	Standard
<input type="checkbox"/>	part-00001-4217ee94-d0f2-497f-90f-0a4c0d4704d1-c000.snappy.parquet	parquet	18 Nov 2023 12:06:19 AM CET	1.1 Mo	Standard
<input type="checkbox"/>	part-00002-4217ee94-d0f2-497f-90f-0a4c0d4704d1-c000.snappy.parquet	parquet	18 Nov 2023 12:06:19 AM CET	1001.5 Ko	Standard
<input type="checkbox"/>	part-00003-4217ee94-d0f2-497f-90f-0a4c0d4704d1-c000.snappy.parquet	parquet	18 Nov 2023 12:06:19 AM CET	1000.0 Ko	Standard
<input type="checkbox"/>	part-00004-4217ee94-d0f2-497f-90f-0a4c0d4704d1-c000.snappy.parquet	parquet	18 Nov 2023 12:06:20 AM CET	1006.4 Ko	Standard

[Copier l'URI S3](#) [Copier l'URL](#) [Télécharger](#) [Ouvrir](#) [Supprimer](#) [Actions](#) [Créer un dossier](#) [Charger](#)

Rechercher des objets en fonction du préfixe

Amazon S3 > Compartiments > aws-emr-studio-429430768867-eu-central-1

aws-emr-studio-429430768867-eu-central-1 [Info](#)

[Objets](#) [Propriétés](#) [Autorisations](#) [Métriques](#) [Gestion](#) [Points d'accès](#)

Objets (2)

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](#)

<input type="checkbox"/>	Nom	Type	Dernière modification	Taille	Classe de stockage
<input type="checkbox"/>	j-1YRSQSPOSATRY/	Dossier	-	-	-
<input type="checkbox"/>	j-300SYIEBZVA/	Dossier	-	-	-

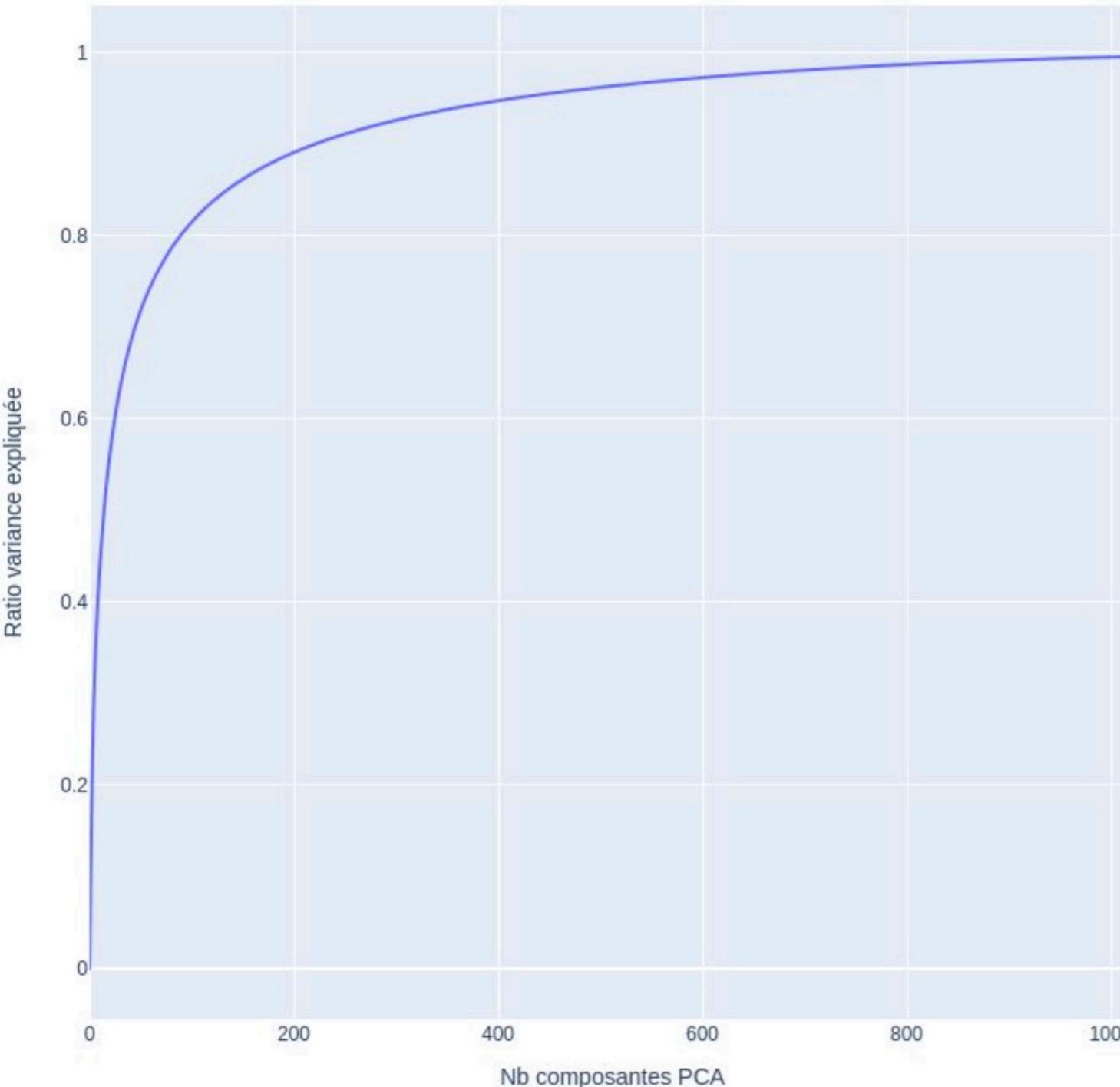
[Copier l'URI S3](#) [Copier l'URL](#) [Télécharger](#) [Ouvrir](#) [Supprimer](#) [Actions](#) [Créer un dossier](#) [Charger](#)

Rechercher des objets en fonction du préfixe



Feature reduction

Résultats de la PCA



PCA

Variances expliquée

97 %

Features

600



Fruits!

Conclusion

Enseignements:

- Prise en main Pyspark.
- Découverte du format distribué parquet.
- Découverte de l'écosystème AWS.
- Comparaison entre Spark en local et Spark déployé sur AWS

Déploiement sur une instance EMR d'AWS facilitant le passage à l'échelle (*EMR permet d'augmenter ou de diminuer facilement le nombre de ressources (telles que des serveurs), permettant au système d'évoluer en fonction des besoins des utilisateurs*).

Prochaines étapes:

- Entraînement d'un classifier (+1 Couche Fully-Connected ou apprentissage supervisé).
- Déploiement du modèle entrainé.
- Créer un chat-bot dans le réseau social Telegram qui reconnaîtrait les légumes et les fruits.
- Tester les solutions existantes sur le marché : API Pl@ntnet.
- Pré-traitement pour cas réels (recadrage, plusieurs fruits, arrière plan, etc.).
- Identifier la maturité des fruits pour les cueillir au bon moment.
- Identifier les pathologies ou les fruits abîmés.



MERCI DE
VOTRE
ATTENTION

