Projet 8
Déployez un modèle dans le cloud

Parcours Data Scientist - Présenté par Kilian ALLIOT - 16/08/2023

Plan

Introduction

Fruits!

Je suis Data Scientist dans une très jeune start-up de l'AgriTech, nommée "Fruits!", qui cherche à proposer des solutions innovantes pour la récolte des fruits.
La volonté de l’entreprise est de préserver la biodiversité des fruits en permettant des traitements spécifiques pour chaque espèce de fruits en développant des robots cueilleurs intelligents.

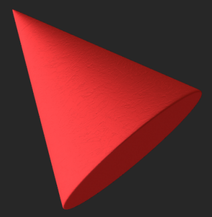
Fruits souhaite dans un premier temps se faire connaître en mettant à disposition du grand public une application mobile qui permettrait aux utilisateurs de prendre en photo un fruit et d'obtenir des informations sur ce fruit.
Pour la start-up, cette application permettrait de sensibiliser le grand public à la biodiversité des fruits et de mettre en place une première version du moteur de classification des images de fruits.
De plus, le développement de l’application mobile permettra de construire une première version de l'architecture Big Data nécessaire.

Objectifs

- Développer une première chaîne de traitement des données qui comprendra le preprocessing et une étape de réduction de dimension.
- Tenir compte du fait que le volume de données va augmenter très rapidement après la livraison de ce projet, ce qui implique de:
  - Déployer le traitement des données dans un environnement Big Data
  - Développer les scripts en pyspark pour effectuer du calcul distribué

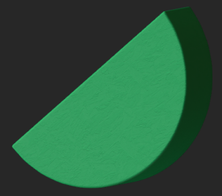
Traitement des données

1



Pyspark

2



MobileNetV2

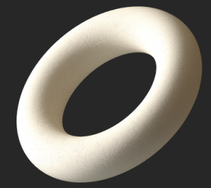
3

Preprocessing

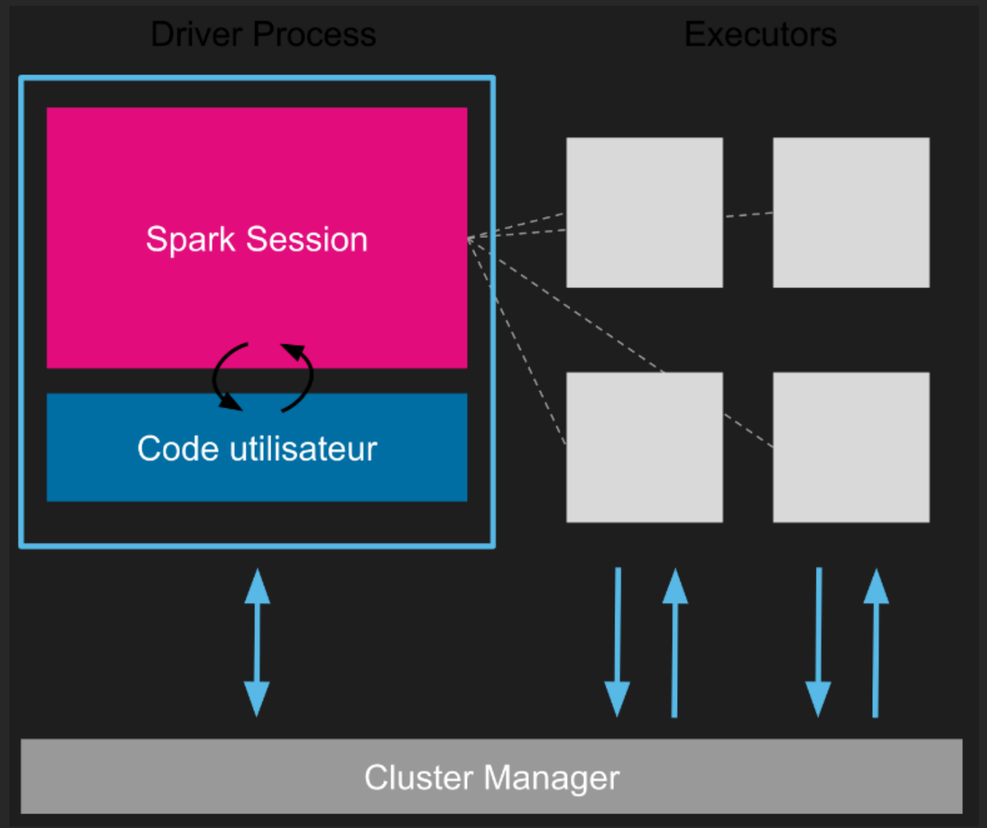
4

ACP

5



Validation



Input	Operator	t	c	n	s
$224^2 \times 3$	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	-	

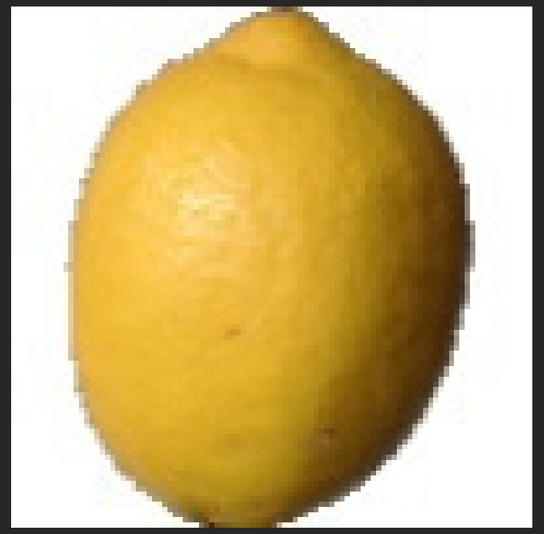
Tous les workeurs doivent pouvoir accéder au modèle ainsi qu'à ses poids.
Une bonne pratique consiste à charger le modèle sur le driver puis à diffuser ensuite les poids aux différents workeurs.
Cette étape correspond au traitement de diffusion des poids du modèle Tensorflow sur les clusters (broadcast des "weights" du modèle) demandé dans l'énoncé du projet.

```
brodcast_weights = sc.broadcast(new_model.get_weights())
```

✓ 0.6s

- Essai local : 64 images
- Cloud : 22688 images

```
+--------------------+-------------------+------+--------------------+
|                path|   modificationTime|length|             content|
+--------------------+-------------------+------+--------------------+
|file:/home/kilian...|2021-09-12 20:26:06|  5735|[FF D8 FF E0 00 1...|
|file:/home/kilian...|2021-09-12 20:26:06|  5729|[FF D8 FF E0 00 1...|
|file:/home/kilian...|2021-09-12 20:26:06|  5720|[FF D8 FF E0 00 1...|
|file:/home/kilian...|2021-09-12 20:26:06|  5715|[FF D8 FF E0 00 1...|
|file:/home/kilian...|2021-09-12 20:26:06|  5678|[FF D8 FF E0 00 1...|
+--------------------+-------------------+------+--------------------+
```

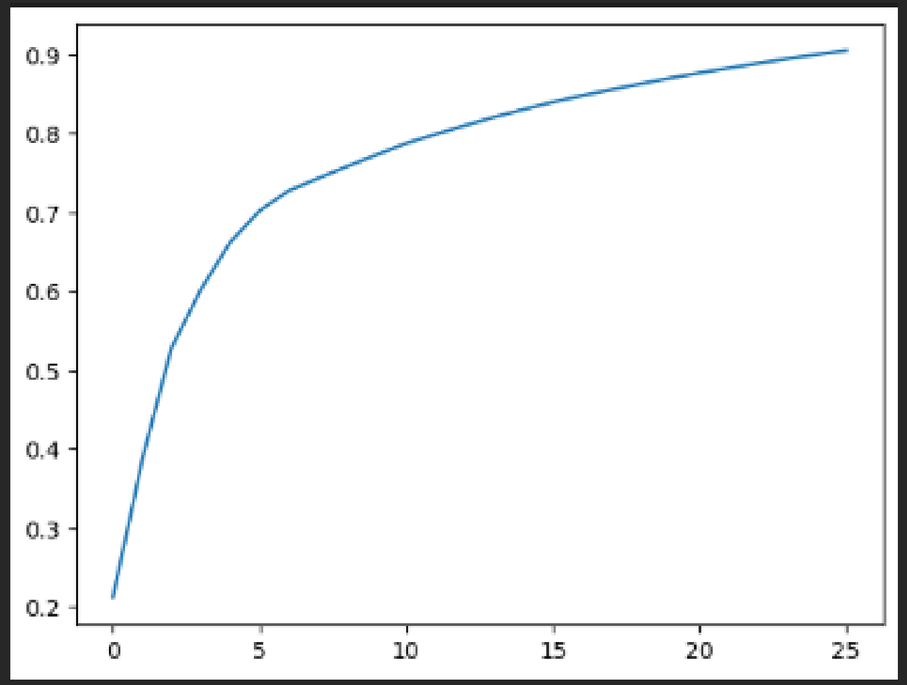


	path	label	features
0	file:/home/kilian/workspaces/workspace-data/OC...	Clementine	[0.15727534890174866, 0.0, 0.0, 0.0, 0.0, 0.20...
1	file:/home/kilian/workspaces/workspace-data/OC...	Strawberry	[1.3659480810165405, 0.09833923727273941, 0.00...
2	file:/home/kilian/workspaces/workspace-data/OC...	Banana	[1.4225854873657227, 0.08237817138433456, 0.0,...
3	file:/home/kilian/workspaces/workspace-data/OC...	Clementine	[0.511864423751831, 0.07806168496608734, 0.0, ...
4	file:/home/kilian/workspaces/workspace-data/OC...	Banana	[1.7645273208618164, 0.0011792182922363281, 0....

```
+--------------------+----------+--------------------+--------------------+--------------------+
|                path|     label|            features|     scaled_features|        pca_features|
+--------------------+----------+--------------------+--------------------+--------------------+
|file:/home/kilian...|Clementine|[0.15727534890174...||[-1.2595558414230...|[5.24053856546943...|
|file:/home/kilian...|Strawberry|[1.36594808101654...||[1.05345249320842...|[7.60194355094782...|
|file:/home/kilian...|    Banana|[1.42258548736572...||[1.16183815415434...|[-25.419240298328...|
|file:/home/kilian...|Clementine|[0.51186442375183...||[-0.5809871406445...|[7.44430266889824...|
|file:/home/kilian...|    Banana|[1.76452732086181...||[1.81620412917869...|[-25.570767635673...|
+--------------------+----------+--------------------+--------------------+--------------------+
```

	path	label	pca_features
0	file:/home/kilian/workspaces/workspace-data/OC...	Clementine	[5.2405386, -1.6420962, 19.637547, 7.7231736, ...
1	file:/home/kilian/workspaces/workspace-data/OC...	Strawberry	[7.6019435, 25.94066, -7.2671494, 0.48385996, ...
2	file:/home/kilian/workspaces/workspace-data/OC...	Banana	[-25.41924, -2.742938, -1.231952, 10.927477, 2...
3	file:/home/kilian/workspaces/workspace-data/OC...	Clementine	[7.4443026, -5.86748, 24.184364, -6.5451164, 1...
4	file:/home/kilian/workspaces/workspace-data/OC...	Banana	[-25.570768, -3.2196324, -2.7823904, -1.531308...

ACP



Déploiement de la solution sur le cloud
Création de l'environnement Big Data

- AWS
- EMR
- EC2
- S3
- IAM

Amazon S3 > Compartiments > kiliandatadev-p8-data

kiliandatadev-p8-data Infos

Objets | Propriétés | Autorisations | Métriques | Gestion | Points d'accès

Objets (6)

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

1

Nom	Type	Dernière modification	Taille	Classe de stockage
bootstrap-emr.sh	sh	14 Aug 2023 06:30:41 PM CEST	406.0 o	Standard
j-3EFKRZFRW6JMF/	Dossier	-	-	-
jupyter/	Dossier	-	-	-
logs/	Dossier	-	-	-
Results/	Dossier	-	-	-
Test/	Dossier	-	-	-


```
$ bootstrap-emr.sh
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
7   sudo python3 -m pip install pyarrow
8   sudo python3 -m pip install boto3
9   sudo python3 -m pip install s3fs
10  sudo python3 -m pip install fsspec
11  sudo python3 -m pip install pyspark
12  sudo python3 -m pip install tensorflow
```

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.

Résilier l'instance | Redimensionner le groupe d'instances | Ajouter un groupe d'instances de tâches

Rechercher des instances par st… | Rechercher des ressources par ID ou type ; ou rechercher le texte dans les résultats chargés

Type et nom	ID	Statut Motif du dernier changemen	Instances	Option d'achat et prix
Primaire	ig-3JZIPZT62ZYT3	En cours d'exécution -	1	-
Principal (Unité principale)	ig-1W25TKARWUA2D	En cours d'exécution -	2	-

Configurations de cluster

Les configurations de cluster sont définies lorsque vous créez un cluster.

Filtrer la classification

Trouver des configurations de cluster | Toute classification

Classification	Propriété	Valeur	Source
jupyter-s3-conf	s3.persistence.bucket	kiliandatadev-p8-data	Configurations de cluster
jupyter-s3-conf	s3.persistence.enabled	true	Configurations de cluster

```
Warning: Permanently added 'ec2-15-188-48-225.eu-west-3.compute.amazonaws.com' (
ED25519) to the list of known hosts.

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/
7 package(s) needed for security, out of 9 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRR
E::::::::::::::::::E M::::::M         M::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M:::::::M       M:::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M::::::::M     M::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E::::::EEEEEEEEEE  M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R::::::::::RR
  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:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-47-19 ~]$
```

Amazon EMR > EMR sur EC2: Clusters > P8_Fruits

P8_Fruits

Mise à jour il y a moins d'une minute | Actions

Récapitulatif

Informations sur le cluster

ID de cluster
j-1MLR1TLAR57OF

Configuration de cluster
Groupes d'instances

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

Applications

Version d'Amazon EMR
emr-6.12.0

Applications installées
Hadoop 3.3.3, JupyterHub 1.4.1, Spark 3.4.0

Gestion des clusters

Destination des journaux dans Amazon S3
kiliandatadev-p8-data/logs

Interfaces utilisateur d'application persistantes
Serveur d'historique Spark
Serveur de chronologie YARN

DNS public du nœud primaire
ec2-15-188-48-225.eu-west-3.compute.amazonaws.com
Connexion au nœud primaire à l'aide de SSH

Statut et heure

Statut
En attente

Heure de création
14 août 2023 19:01 (UTC+02:00)

Temps écoulé
1 heure, 47 minutes

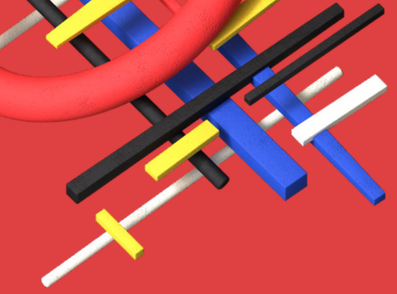
Interfaces utilisateur d'application sur le nœud primaire

Cela nécessite l'activation du tunneling SSH. Suivez les instructions de la section ${managementGuideLink}.

Enable an SSH connection

Application	URL de l'interface utilisateur
Gestionnaire de ressources	http://ec2-15-188-48-225.eu-west-3.compute.amazonaws.com:8088/
JupyterHub	https://ec2-15-188-48-225.eu-west-3.compute.amazonaws.com:9443/
Nom du nœud HDFS	http://ec2-15-188-48-225.eu-west-3.compute.amazonaws.com:9870/
Serveur d'historique Spark	http://ec2-15-188-48-225.eu-west-3.compute.amazonaws.com:18080/

i-0c0875e0af54a9e79	En cours d'exécution		m5.xlarge	2/2 vérifications réussies	Aucune al...	eu-west-3c	ElasticMapReduce-master
i-0e96cc8c96488315f	En cours d'exécution		m5.xlarge	2/2 vérifications réussies	Aucune al...	eu-west-3c	ElasticMapReduce-slave
i-05773d118bf6ba015	En cours d'exécution		m5.xlarge	2/2 vérifications réussies	Aucune al...	eu-west-3c	ElasticMapReduce-slave



AWS IAM

IAM > Rôles

Rôles (9) Infos

Un rôle IAM est une identité que vous pouvez créer et qui dispose d'autorisations spécifiques avec des informations d'identification valides pendant de courtes durées. Les rôles peuvent être endossés par des entités de confiance.

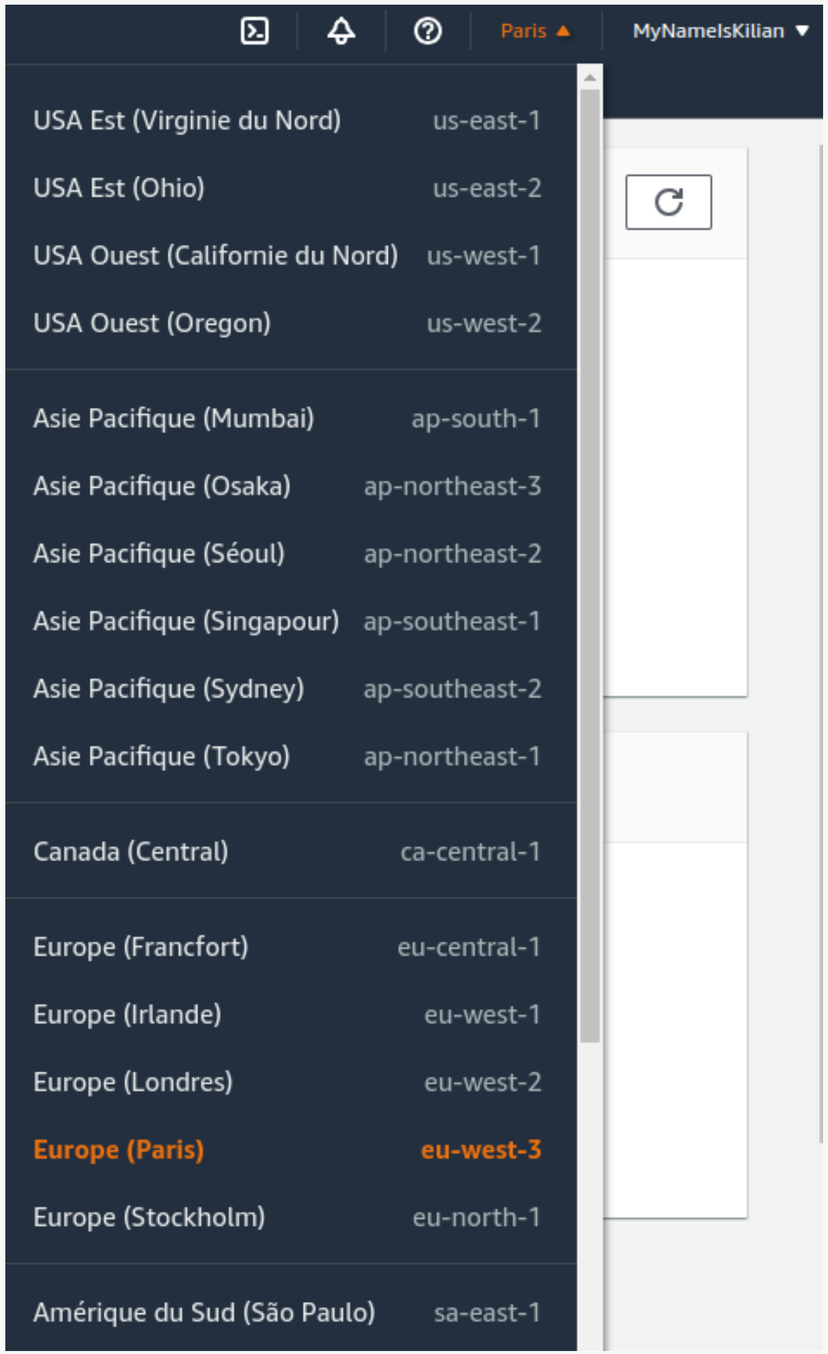
Supprimer | Créer un rôle

Rechercher

1

Nom du rôle	Entités de confiance	Dernière activité
AmazonEMR-InstanceProfile-20230810T193430	Service AWS: ec2	Il y a 11 minutes
AmazonEMR-InstanceProfile-20230814T183833	Service AWS: ec2	-
AmazonEMR-InstanceProfile-20230814T184753	Service AWS: ec2	-
AmazonEMR-ServiceRole-20230810T193446	Service AWS: elasticmapreduce	Il y a 16 minutes
AWSServiceRoleForEMRCleanup	Service AWS: elasticmapreduce (Rôle lié à un service)	Il y a 10 minutes
AWSServiceRoleForRDS	Service AWS: rds (Rôle lié à un service)	Il y a 13 minutes
AWSServiceRoleForSupport	Service AWS: support (Rôle lié à un service)	-
AWSServiceRoleForTrustedAdvisor	Service AWS: trustedadvisor (Rôle lié à un service)	-
rds-monitoring-role	Service AWS: monitoring.rds	-

RGPD


Paris
MyNamelsKilian
USA Est (Virginie du Nord) us-east-1
USA Est (Ohio) us-east-2
USA Ouest (Californie du Nord) us-west-1
USA Ouest (Oregon) us-west-2
Asie Pacifique (Mumbai) ap-south-1
Asie Pacifique (Osaka) ap-northeast-3
Asie Pacifique (Séoul) ap-northeast-2
Asie Pacifique (Singapour) ap-southeast-1
Asie Pacifique (Sydney) ap-southeast-2
Asie Pacifique (Tokyo) ap-northeast-1
Canada (Central) ca-central-1
Europe (Francfort) eu-central-1
Europe (Irlande) eu-west-1
Europe (Londres) eu-west-2
Europe (Paris) eu-west-3
Europe (Stockholm) eu-north-1
Amérique du Sud (São Paulo) sa-east-1

Retour critique

Cluster avec 3 instances EC2 m5.xlarge

Stage Id ▾	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
29	Job group for statement 25 parquet at NativeMethodAccessorImpl.java:0 +details	2023/08/14 18:48:20	7.6 min	24/24		2.5 MiB	95.0 MiB	
27	Job group for statement 25 parquet at NativeMethodAccessorImpl.java:0 +details	2023/08/14 18:37:50	10 min	709/709	98.4 MiB			95.0 MiB
26	Job group for statement 21 showString at NativeMethodAccessorImpl.java:0 +details	2023/08/14 18:30:40	26 s	1/1			4.0 MiB	
24	Job group for statement 21 showString at NativeMethodAccessorImpl.java:0 +details	2023/08/14 18:19:35	11 min	709/709	98.4 MiB			95.0 MiB
23	Job group for statement 19 treeAggregate at RowMatrix.scala:171 +details	2023/08/14 18:15:28	2 s	4/4			150.2 MiB	
22	Job group for statement 19 treeAggregate at RowMatrix.scala:171 +details	2023/08/14 18:07:50	7.6 min	24/24			93.4 MiB	150.2 MiB
20	Job group for statement 19 isEmpty at RowMatrix.scala:441 +details	2023/08/14 18:07:30	20 s	1/1			3.9 MiB	
18	Job group for statement 19 treeAggregate at Statistics.scala:58 +details	2023/08/14 18:07:30	62 ms	4/4			256.4 KiB	
17	Job group for statement 19 treeAggregate at Statistics.scala:58 +details	2023/08/14 17:59:57	7.5 min	24/24			93.4 MiB	256.4 KiB
15	Job group for statement 19 first at RowMatrix.scala:62 +details	2023/08/14 17:59:38	20 s	1/1			3.9 MiB	
13	Job group for statement 19 first at PCA.scala:44 +details	2023/08/14 17:59:13	24 s	1/1			3.9 MiB	
12	Job group for statement 19 rdd at PCA.scala:89 +details	2023/08/14 17:49:12	9.9 min	709/709	98.4 MiB			93.4 MiB
11	Job group for statement 18 first at StandardScaler.scala:113 +details	2023/08/14 17:44:14	0.2 s	1/1			607.0 KiB	
8	Job group for statement 18 first at StandardScaler.scala:113 +details	2023/08/14 17:36:37	7.6 min	24/24			93.4 MiB	607.0 KiB
6	Job group for statement 18 first at StandardScaler.scala:113 +details	2023/08/14 17:26:03	10 min	709/709	98.4 MiB			93.4 MiB
5	Job group for statement 7 showString at NativeMethodAccessorImpl.java:0 +details	2023/08/14 17:24:32	0.2 s	1/1				
4	Job group for statement 6 count at NativeMethodAccessorImpl.java:0 +details	2023/08/14 17:24:09	0.2 s	1/1			34.6 KiB	
2	Job group for statement 6 count at NativeMethodAccessorImpl.java:0 +details	2023/08/14 17:21:12	3.0 min	709/709				34.6 KiB
1	Job group for statement 5 showString at NativeMethodAccessorImpl.java:0 +details	2023/08/14 17:20:55	1 s	1/1	43.0 KiB			
0	Listing leaf files and directories for 131 paths: s3://kiliandatadev-p8-data/Test/Apple Braeburn, ... load at NativeMethodAccessorImpl.java:0 +details	2023/08/14 17:20:34	8 s	131/131				

Cluster avec 3 instances EC2 m5.xlarge

ID ▾	Description	Submitted	Duration	Job IDs
5	Job group for statement 25 +details	2023/08/14 18:37:50	18 min	[15][16]
4	Job group for statement 21 +details	2023/08/14 18:19:34	12 min	[13][14]
3	Job group for statement 18 +details	2023/08/14 17:26:03	18 min	[5][6][7]
2	Job group for statement 7 +details	2023/08/14 17:24:32	0.4 s	[4]
1	Job group for statement 6 +details	2023/08/14 17:21:11	3.0 min	[2][3]
0	Job group for statement 5 +details	2023/08/14 17:20:55	2 s	[1]

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(1)	0	0.0 B / 400 MiB	0.0 B	0	0	0	0	0	3.1 h (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Dead(6)	0	0.0 B / 28.6 GiB	0.0 B	12	0	0	3788	3788	2.5 h (29 s)	393.5 MiB	542.2 MiB	528.1 MiB	0
Total(7)	0	0.0 B / 29 GiB	0.0 B	12	0	0	3788	3788	5.6 h (29 s)	393.5 MiB	542.2 MiB	528.1 MiB	0

Executors

Show 20 entries Search:

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Logs
driver	ip-172-31-39-164.eu-west-3.compute.internal:37781	Active	0	0.0 B / 400 MiB	0.0 B	0	0	0	0	0	3.1 h (0.0 ms)	0.0 B	0.0 B	0.0 B	stdout stderr
1	ip-172-31-41-222.eu-west-3.compute.internal:40729	Dead	0	0.0 B / 4.8 GiB	0.0 B	2	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	stdout stderr
2	ip-172-31-41-222.eu-west-3.compute.internal:44679	Dead	0	0.0 B / 4.8 GiB	0.0 B	2	0	0	843	843	6.2 min (0.7 s)	43 KiB	34.6 KiB	34.6 KiB	stdout stderr
3	ip-172-31-41-222.eu-west-3.compute.internal:39947	Dead	0	0.0 B / 4.8 GiB	0.0 B	2	0	0	734	734	36 min (5 s)	98.4 MiB	94 MiB	94 MiB	stdout stderr
4	ip-172-31-41-222.eu-west-3.compute.internal:38083	Dead	0	0.0 B / 4.8 GiB	0.0 B	2	0	0	768	768	51 min (9 s)	98.4 MiB	349.1 MiB	243.9 MiB	stdout stderr
5	ip-172-31-41-222.eu-west-3.compute.internal:41555	Dead	0	0.0 B / 4.8 GiB	0.0 B	2	0	0	710	710	22 min (6 s)	98.4 MiB	4 MiB	95 MiB	stdout stderr
6	ip-172-31-41-222.eu-west-3.compute.internal:34741	Dead	0	0.0 B / 4.8 GiB	0.0 B	2	0	0	733	733	36 min (9 s)	98.4 MiB	95 MiB	95 MiB	stdout stderr

Cluster avec 3 instances EC2 m5.xlarge

Completed Jobs (17)

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
16 (25)	Job group for statement 25 parquet at NativeMethodAccessorImpl.java:0	2023/08/14 18:48:20	7.6 min	1/1 (1 skipped)	24/24 (709 skipped)
15 (25)	Job group for statement 25 parquet at NativeMethodAccessorImpl.java:0	2023/08/14 18:37:50	11 min	1/1	709/709
14 (21)	Job group for statement 21 showString at NativeMethodAccessorImpl.java:0	2023/08/14 18:30:40	26 s	1/1 (1 skipped)	1/1 (709 skipped)
13 (21)	Job group for statement 21 showString at NativeMethodAccessorImpl.java:0	2023/08/14 18:19:35	11 min	1/1	709/709
12 (19)	Job group for statement 19 treeAggregate at RowMatrix.scala:171	2023/08/14 18:07:50	7.7 min	2/2 (1 skipped)	28/28 (709 skipped)
11 (19)	Job group for statement 19 isEmpty at RowMatrix.scala:441	2023/08/14 18:07:30	20 s	1/1 (1 skipped)	1/1 (709 skipped)
10 (19)	Job group for statement 19 treeAggregate at Statistics.scala:58	2023/08/14 17:59:57	7.5 min	2/2 (1 skipped)	28/28 (709 skipped)
9 (19)	Job group for statement 19 first at RowMatrix.scala:62	2023/08/14 17:59:38	20 s	1/1 (1 skipped)	1/1 (709 skipped)
8 (19)	Job group for statement 19 first at PCA.scala:44	2023/08/14 17:49:12	10 min	2/2	710/710
7 (18)	Job group for statement 18 first at StandardScaler.scala:113	2023/08/14 17:44:14	0.2 s	1/1 (2 skipped)	1/1 (733 skipped)
6 (18)	Job group for statement 18 first at StandardScaler.scala:113	2023/08/14 17:36:37	7.6 min	1/1 (1 skipped)	24/24 (709 skipped)
5 (18)	Job group for statement 18 first at StandardScaler.scala:113	2023/08/14 17:26:03	11 min	1/1	709/709
4 (7)	Job group for statement 7 showString at NativeMethodAccessorImpl.java:0	2023/08/14 17:24:32	0.2 s	1/1	1/1
3 (6)	Job group for statement 6 count at NativeMethodAccessorImpl.java:0	2023/08/14 17:24:09	0.2 s	1/1 (1 skipped)	1/1 (709 skipped)
2 (6)	Job group for statement 6 count at NativeMethodAccessorImpl.java:0	2023/08/14 17:21:12	3.0 min	1/1	709/709
1 (5)	Job group for statement 5 showString at NativeMethodAccessorImpl.java:0	2023/08/14 17:20:55	1 s	1/1	1/1
0 (4)	Listing leaf files and directories for 131 paths: s3://kiliandatadev-p8-data/Test/Apple Braeburn, ... load at NativeMethodAccessorImpl.java:0	2023/08/14 17:20:34	11 s	1/1	131/131

Cluster avec 4 instances EC2 m5.2xlarge

Completed Jobs (13)

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
12 (23)	Job group for statement 23 parquet at NativeMethodAccessorImpl.java:0	2023/08/14 21:49:49	1.7 min	1/1 (1 skipped)	24/24 (709 skipped)
11 (23)	Job group for statement 23 parquet at NativeMethodAccessorImpl.java:0	2023/08/14 21:48:00	1.8 min	1/1	709/709
10 (17)	Job group for statement 17 treeAggregate at RowMatrix.scala:171	2023/08/14 21:46:07	1.7 min	2/2 (1 skipped)	28/28 (709 skipped)
9 (17)	Job group for statement 17 isEmpty at RowMatrix.scala:441	2023/08/14 21:45:54	13 s	1/1 (1 skipped)	1/1 (709 skipped)
8 (17)	Job group for statement 17 treeAggregate at Statistics.scala:58	2023/08/14 21:44:27	1.4 min	2/2 (1 skipped)	28/28 (709 skipped)
7 (17)	Job group for statement 17 first at RowMatrix.scala:62	2023/08/14 21:44:11	16 s	1/1 (1 skipped)	1/1 (709 skipped)
6 (17)	Job group for statement 17 first at PCA.scala:44	2023/08/14 21:42:05	2.1 min	2/2	710/710
5 (16)	Job group for statement 16 first at StandardScaler.scala:113	2023/08/14 21:42:04	0.1 s	1/1 (2 skipped)	1/1 (733 skipped)
4 (16)	Job group for statement 16 first at StandardScaler.scala:113	2023/08/14 21:40:22	1.7 min	1/1 (1 skipped)	24/24 (709 skipped)
3 (16)	Job group for statement 16 first at StandardScaler.scala:113	2023/08/14 21:38:21	2.0 min	1/1	709/709
2 (5)	Job group for statement 5 showString at NativeMethodAccessorImpl.java:0	2023/08/14 21:38:15	0.8 s	1/1	1/1
1 (4)	Job group for statement 4 showString at NativeMethodAccessorImpl.java:0	2023/08/14 21:38:13	2 s	1/1	1/1
0 (3)	Listing leaf files and directories for 131 paths: s3://kiliandatadev-p8-data/Test/Apple Braeburn, ... load at NativeMethodAccessorImpl.java:0	2023/08/14 21:38:04	4 s	1/1	131/131

Cluster avec 4 instances EC2 m5.2xlarge

▾ Completed Stages (16)

Page: 1

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

Stage Id ▾	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
23	Job group for statement 23 parquet at NativeMethodAccessorImpl.java:0 +details	2023/08/14 21:49:49	1.7 min	24/24		2.6 MiB	95.0 MiB	
21	Job group for statement 23 parquet at NativeMethodAccessorImpl.java:0 +details	2023/08/14 21:48:00	1.8 min	709/709	98.4 MiB			95.0 MiB
20	Job group for statement 17 treeAggregate at RowMatrix.scala:171 +details	2023/08/14 21:47:46	2 s	4/4			150.2 MiB	
19	Job group for statement 17 treeAggregate at RowMatrix.scala:171 +details	2023/08/14 21:46:07	1.6 min	24/24			93.4 MiB	150.2 MiB
17	Job group for statement 17 isEmpty at RowMatrix.scala:441 +details	2023/08/14 21:45:54	13 s	1/1			3.9 MiB	
15	Job group for statement 17 treeAggregate at Statistics.scala:58 +details	2023/08/14 21:45:54	43 ms	4/4			256.5 KiB	
14	Job group for statement 17 treeAggregate at Statistics.scala:58 +details	2023/08/14 21:44:27	1.4 min	24/24			93.4 MiB	256.5 KiB
12	Job group for statement 17 first at RowMatrix.scala:62 +details	2023/08/14 21:44:11	16 s	1/1			3.9 MiB	
10	Job group for statement 17 first at PCA.scala:44 +details	2023/08/14 21:43:54	17 s	1/1			3.9 MiB	
9	Job group for statement 17 rdd at PCA.scala:89 +details	2023/08/14 21:42:05	1.8 min	709/709	98.4 MiB			93.4 MiB
8	Job group for statement 16 first at StandardScaler.scala:113 +details	2023/08/14 21:42:04	99 ms	1/1			607.0 KiB	
5	Job group for statement 16 first at StandardScaler.scala:113 +details	2023/08/14 21:40:22	1.7 min	24/24			93.4 MiB	607.0 KiB
3	Job group for statement 16 first at StandardScaler.scala:113 +details	2023/08/14 21:38:21	2.0 min	709/709	98.4 MiB			93.4 MiB
2	Job group for statement 5 showString at NativeMethodAccessorImpl.java:0 +details	2023/08/14 21:38:15	0.8 s	1/1				
1	Job group for statement 4 showString at NativeMethodAccessorImpl.java:0 +details	2023/08/14 21:38:13	1 s	1/1	43.0 KiB			
0	Listing leaf files and directories for 131 paths: s3://kiliandatadev-p8-data/Test/Apple Braeburn, ... load at NativeMethodAccessorImpl.java:0 +details	2023/08/14 21:38:04	3 s	131/131				

Cluster avec 4 instances EC2 m5.2xlarge

Executors

Show Additional Metrics

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(7)	0	0.0 B / 29 GiB	0.0 B	12	5	0	2363	2368	2.4 h (17 s)	295.2 MiB	518.4 MiB	433 MiB	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(7)	0	0.0 B / 29 GiB	0.0 B	12	5	0	2363	2368	2.4 h (17 s)	295.2 MiB	518.4 MiB	433 MiB	0

Executors

Show 20 entries Search:

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Logs
driver	ip-172-31-34-91.eu-west-3.compute.internal:41939	Active	0	0.0 B / 400 MiB	0.0 B	0	0	0	0	0	14 min (0.0 ms)	0.0 B	0.0 B	0.0 B	stdout stderr
1	ip-172-31-36-84.eu-west-3.compute.internal:39299	Active	0	0.0 B / 4.8 GiB	0.0 B	2	2	0	391	393	21 min (3 s)	48.9 MiB	54.8 MiB	71.9 MiB	stdout stderr
2	ip-172-31-36-84.eu-west-3.compute.internal:46467	Active	0	0.0 B / 4.8 GiB	0.0 B	2	1	0	392	393	23 min (3 s)	48.3 MiB	104.1 MiB	71.3 MiB	stdout stderr
3	ip-172-31-32-112.eu-west-3.compute.internal:43155	Active	0	0.0 B / 4.8 GiB	0.0 B	2	0	0	397	397	22 min (3 s)	49 MiB	100.3 MiB	71.9 MiB	stdout stderr
4	ip-172-31-32-112.eu-west-3.compute.internal:39005	Active	0	0.0 B / 4.8 GiB	0.0 B	2	1	0	392	393	22 min (2 s)	48.9 MiB	96.7 MiB	72 MiB	stdout stderr
5	ip-172-31-34-91.eu-west-3.compute.internal:43611	Active	0	0.0 B / 4.8 GiB	0.0 B	2	0	0	399	399	22 min (4 s)	50.4 MiB	62.5 MiB	73.3 MiB	stdout stderr
6	ip-172-31-34-91.eu-west-3.compute.internal:34695	Active	0	0.0 B / 4.8 GiB	0.0 B	2	1	0	392	393	22 min (3 s)	49.6 MiB	100 MiB	72.6 MiB	stdout stderr

Executors

Show Additional Metrics

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(1)	0	0.0 B / 400 MiB	0.0 B	0	0	0	0	0	18 min (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Dead(6)	0	0.0 B / 28.6 GiB	0.0 B	12	0	0	2368	2368	2.3 h (18 s)	295.2 MiB	538.2 MiB	433 MiB	0
Total(7)	0	0.0 B / 29 GiB	0.0 B	12	0	0	2368	2368	2.6 h (18 s)	295.2 MiB	538.2 MiB	433 MiB	0

Executors

Show 20 entries Search:

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Logs
driver	ip-172-31-34-91.eu-west-3.compute.internal:41939	Active	0	0.0 B / 400 MiB	0.0 B	0	0	0	0	0	18 min (0.0 ms)	0.0 B	0.0 B	0.0 B	stdout stderr
1	ip-172-31-36-84.eu-west-3.compute.internal:39299	Dead	0	0.0 B / 4.8 GiB	0.0 B	2	0	0	393	393	23 min (3 s)	48.9 MiB	62.7 MiB	71.9 MiB	stdout stderr
2	ip-172-31-36-84.eu-west-3.compute.internal:46467	Dead	0	0.0 B / 4.8 GiB	0.0 B	2	0	0	393	393	23 min (3 s)	48.3 MiB	108.2 MiB	71.3 MiB	stdout stderr
3	ip-172-31-32-112.eu-west-3.compute.internal:43155	Dead	0	0.0 B / 4.8 GiB	0.0 B	2	0	0	397	397	22 min (3 s)	49 MiB	100.3 MiB	71.9 MiB	stdout stderr
4	ip-172-31-32-112.eu-west-3.compute.internal:39005	Dead	0	0.0 B / 4.8 GiB	0.0 B	2	0	0	393	393	22 min (2 s)	48.9 MiB	100.6 MiB	72 MiB	stdout stderr
5	ip-172-31-34-91.eu-west-3.compute.internal:43611	Dead	0	0.0 B / 4.8 GiB	0.0 B	2	0	0	399	399	22 min (4 s)	50.4 MiB	62.5 MiB	73.3 MiB	stdout stderr
6	ip-172-31-34-91.eu-west-3.compute.internal:34695	Dead	0	0.0 B / 4.8 GiB	0.0 B	2	0	0	393	393	23 min (3 s)	49.6 MiB	104 MiB	72.6 MiB	stdout stderr

Démonstration

Conclusion

Nous avons réalisé ce projet **en deux temps** en tenant compte des contraintes qui nous ont été imposées.

Nous avons **dans un premier temps développé notre solution en local** sur une machine virtuelle dans un environnement Linux Ubuntu.

La première phase a consisté à **installer l'environnement de travail Spark**. **Spark** a un paramètre qui nous permet de travaillé en local et nous permet ainsi de **simuler du calcul partagé** en considérant **chaque cœur d'un processeur comme un worker indépendant**.
Nous avons travaillé sur un plus **petit jeu de donnée**, l'idée était simplement de **valider le bon fonctionnement de la solution**.

Nous avons fait le choix de réaliser du **transfert learning** à partir du model **MobileNetV2**.
Ce modèle a été retenu pour sa **légèreté** et sa **rapidité d'exécution** ainsi que pour la **faible dimension de son vecteur en sortie**.

Les résultats ont été enregistrés sur disque en plusieurs partitions au format "**parquet**".

La solution a parfaitement fonctionné en mode local.

La deuxième phase a consisté à créer un **réel cluster de calculs**.
L'objectif était de pouvoir **anticiper une future augmentation de la charge de travail**.

Le meilleur choix retenu a été l'utilisation du prestataire de services **Amazon Web Services** qui nous permet de **louer à la demande de la puissance de calculs**, pour un **coût tout à fait acceptable**.
Ce service se nomme **EC2** et se classe parmi les offres **Infrastructure As A Service** (IAAS).

Nous sommes allez plus loin en utilisant un service de plus haut niveau (**Plateforme As A Service** PAAS) en utilisant le service **EMR** qui nous permet d'un seul coup d'**instancier plusieurs serveur (un cluster)** sur lesquels nous avons pu demander l'installation et la configuration de plusieurs programmes et librairies nécessaires à notre projet comme **Spark**, **Hadoop**, **JupyterHub** ainsi que la librairie **TensorFlow**.

En plus d'être plus **rapide et efficace à mettre en place**, nous avons la **certitude du bon fonctionnement de la solution**, celle-ci ayant été préalablement validé par les ingénieurs d'Amazon.

Nous avons également pu installer, sans difficulté, **les packages nécessaires sur l'ensembles des machines du cluster**.

Enfin, avec très peu de modification, et plus simplement encore, nous avons pu **exécuter notre notebook comme nous l'avions fait localement**.
Nous avons cette fois-ci exécuté le traitement sur **l'ensemble des images de notre dossier "Test"**.

Nous avons opté pour le service **Amazon S3** pour **stocker les données de notre projet**.
S3 offre, pour un faible coût, toutes les conditions dont nous avons besoin pour stocker et exploiter de manière efficace nos données.
L'espace alloué est potentiellement **illimité**, mais les coûts seront fonction de l'espace utilisé.

Il nous sera **facile de faire face à une monté de la charge de travail** en **redimensionnant** simplement notre cluster de machines (horizontalement et/ou verticalement au besoin), les coûts augmenteront en conséquence mais resteront nettement inférieurs aux coûts engendrés par l'achat de matériels ou par la location de serveurs dédiés.

Merci de votre attention