Réalisez un traitement dans un environnement Big Data sur le Cloud

Réaliser par:Kaoutar EL Mardi

jury:Emmanuel Goudot

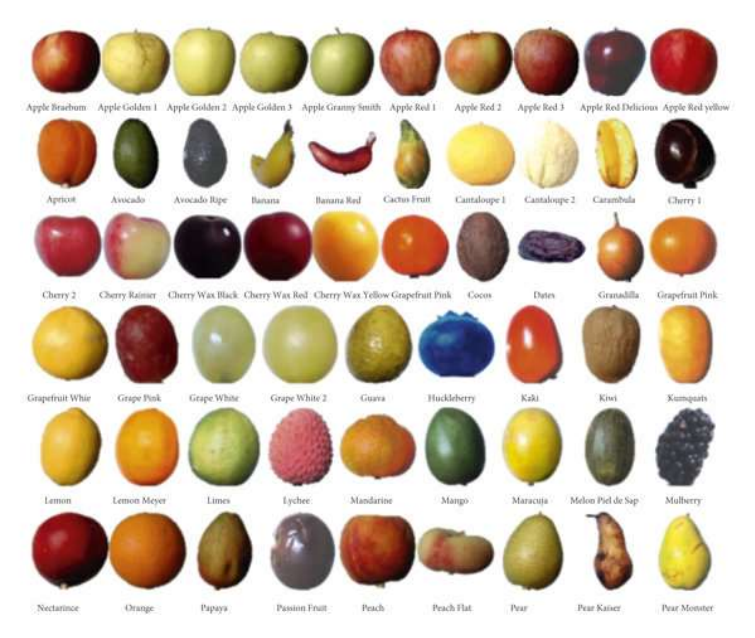
Table of contents

01
Problématique et données

Problématique

Fruits!" développe des robots cueilleurs intelligents pour préserver la biodiversité. Pour sensibiliser le public, l'entreprise lance une **application mobile d'identification des fruits.** Cette application servira à construire **une architecture Big Data** évolutive, respectant le RGPD et optimisant les coûts.

Jeu de données



Caractéristiques principales :

- **Images :** Plus de 70 000 images de fruits.
- **Catégories :** Représente plus de 100 catégories de fruits et légumes.
- **Diversité :** Images prises sous différents angles et conditions d'éclairage.
- **Étiquetage :** Chaque image est soigneusement étiquetée, facilitant la reconnaissance précise des fruits.

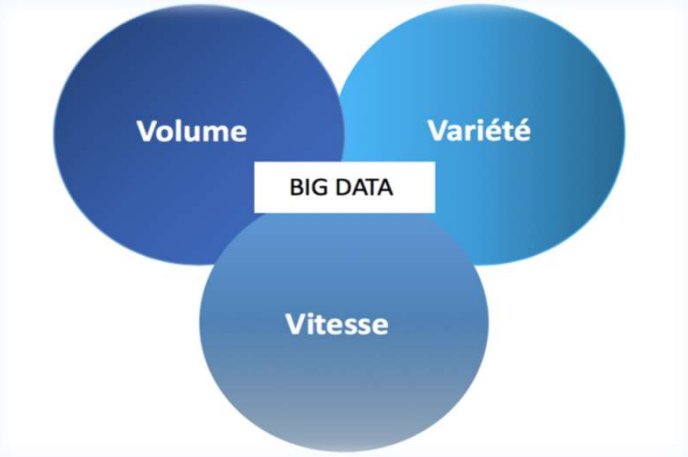
https://www.kaggle.com/datasets/moltean/fruits/data

02

Processus de Création de de l’environnement Big Data

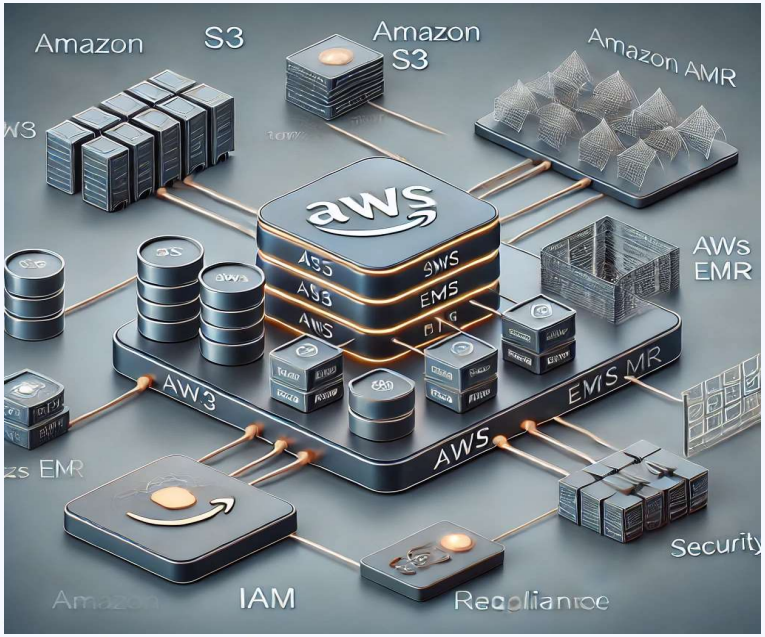
Pourquoi Envirenement Big Data

L'environnement Big Data est indispensable pour gérer la croissance rapide des données de l'application "Fruits!". Les trois dimensions du Big Data — Volume, Variété, et Vitesse — nécessitent une architecture robuste

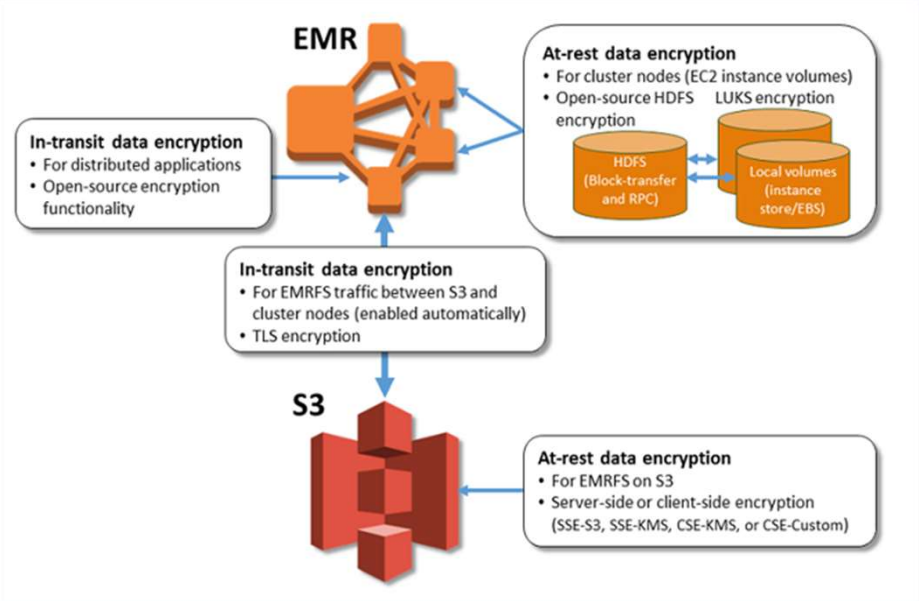


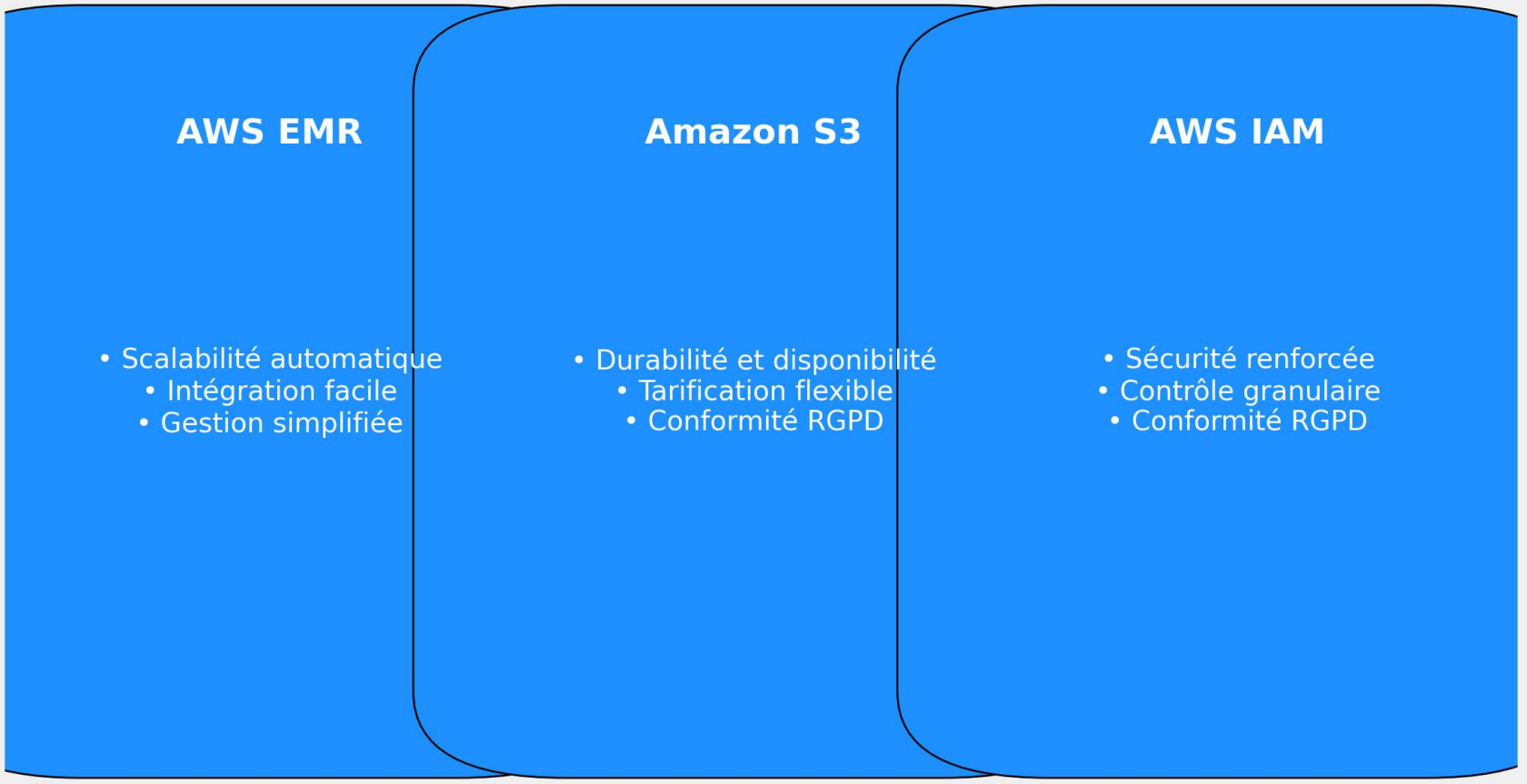
Outils et déploiement

Amazon AWS offre les outils nécessaires pour traiter ces données efficacement tout en assurant la conformité avec le RGPD.

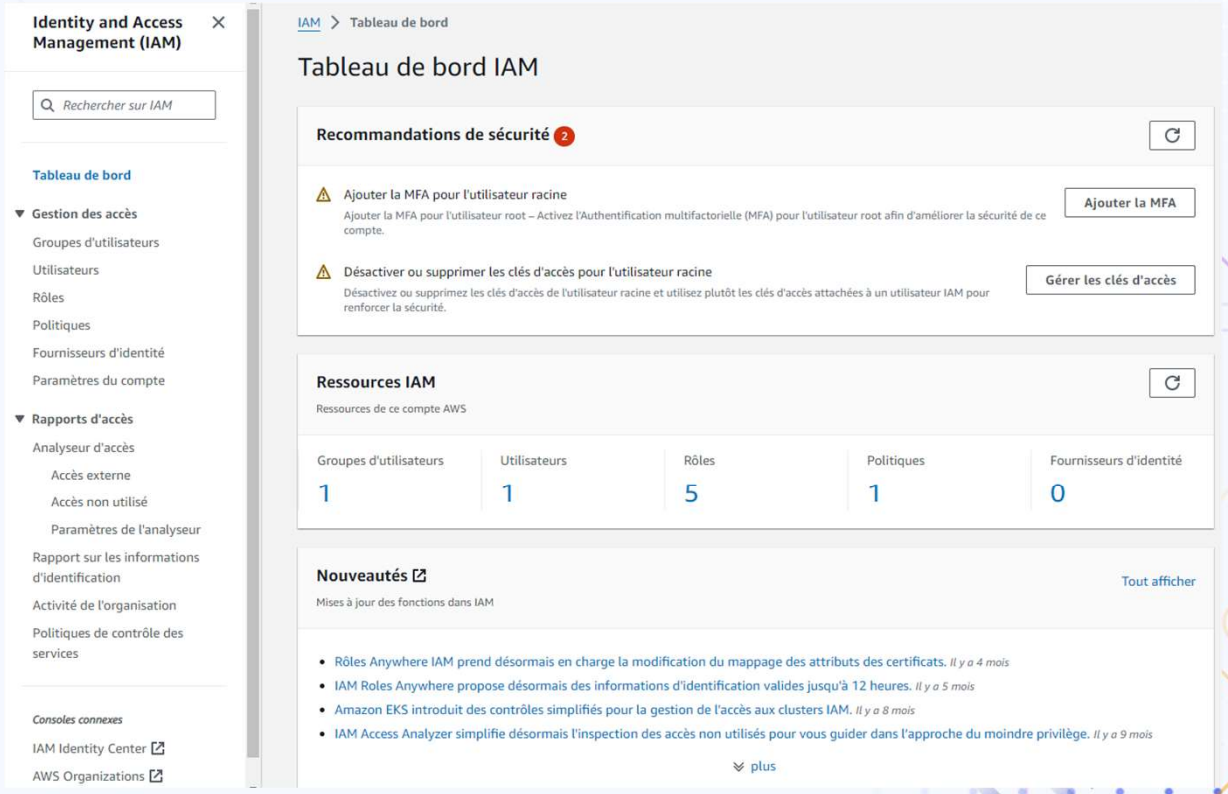


processus de création de l'environnement Big Data

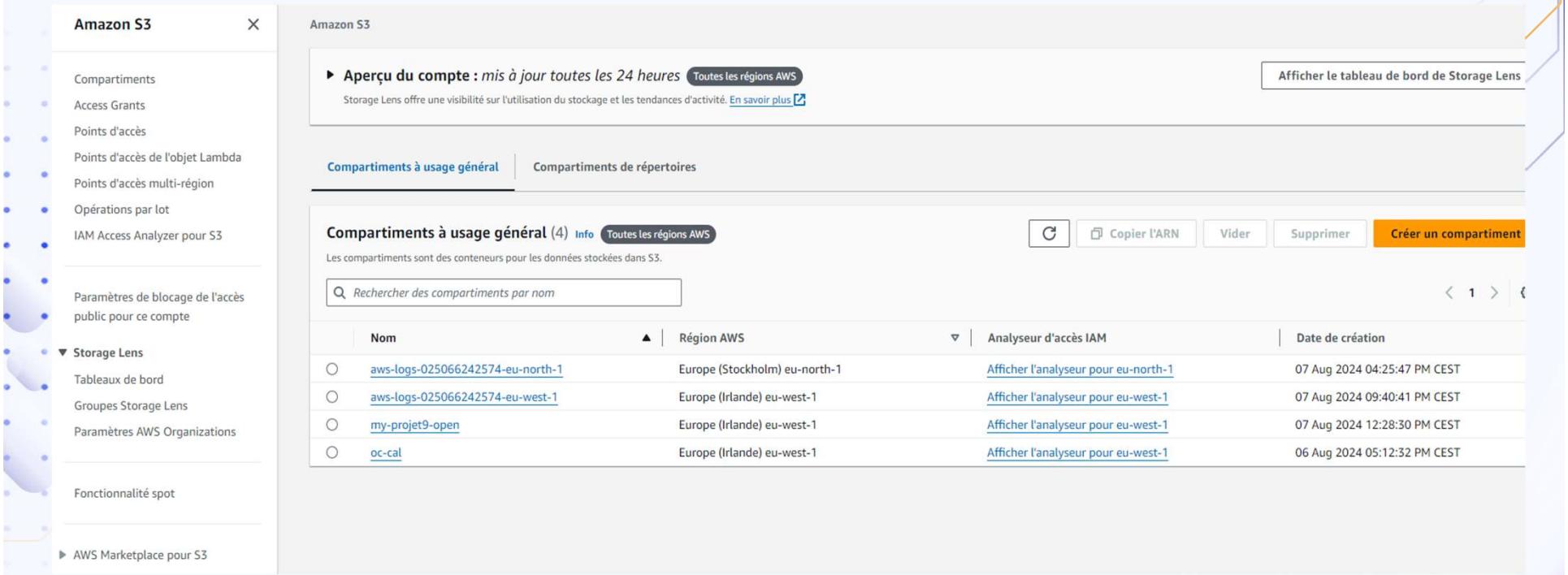



AWS EMR
• Scalabilité automatique
• Intégration facile
• Gestion simplifiée
Amazon S3
• Durabilité et disponibilité
• Tarification flexible
• Conformité RGPD
AWS IAM
• Sécurité renforcée
• Contrôle granulaire
• Conformité RGPD

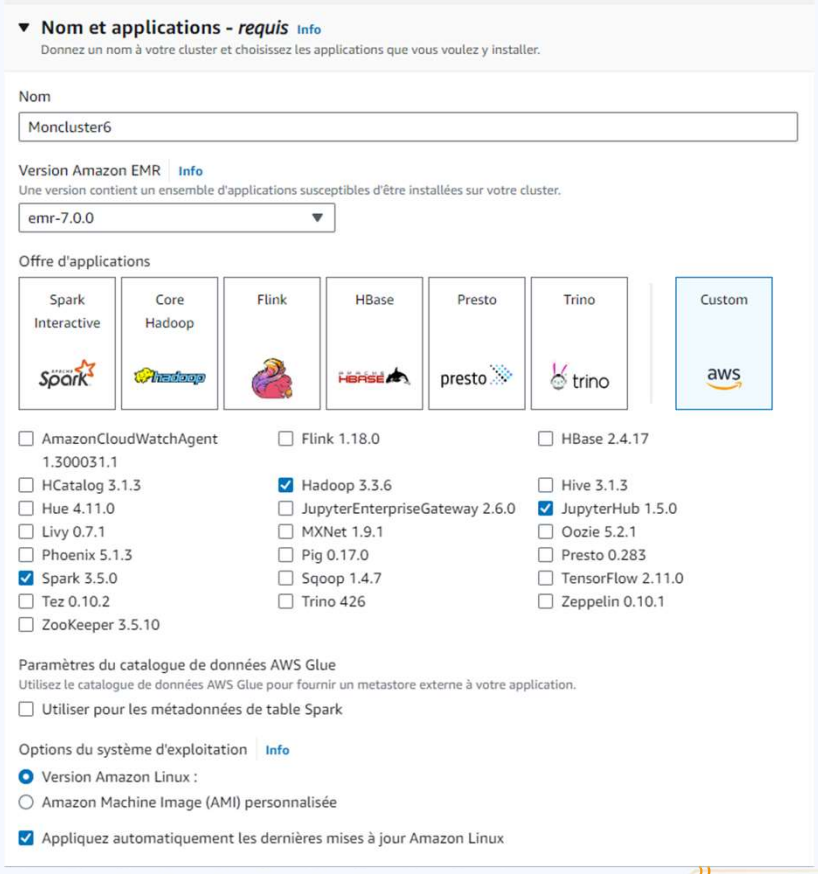
"Configuration IAM : Sécurisation des ressources AWS avec l'authentification multifactorielle (MFA) et la création de rôles pour un contrôle d'accès précis, en conformité avec le RGPD."



Configuration S3 : Création de compartiments sécurisés dans les régions européennes pour le chargement et le stockage des données, assurant la conformité RGPD, avec un contrôle d'accès géré via IAM pour protéger les données sensibles.

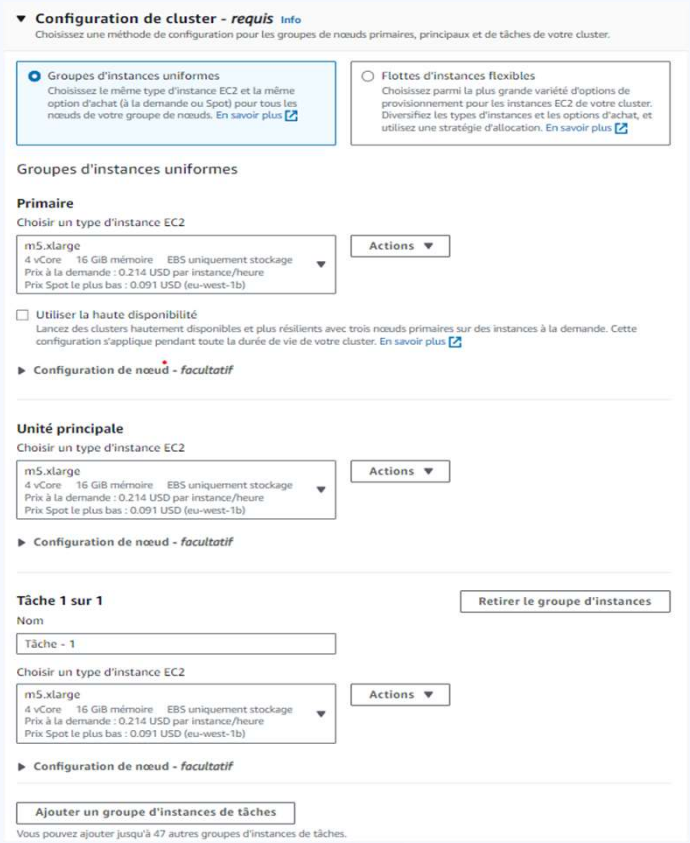


Creation de EMR

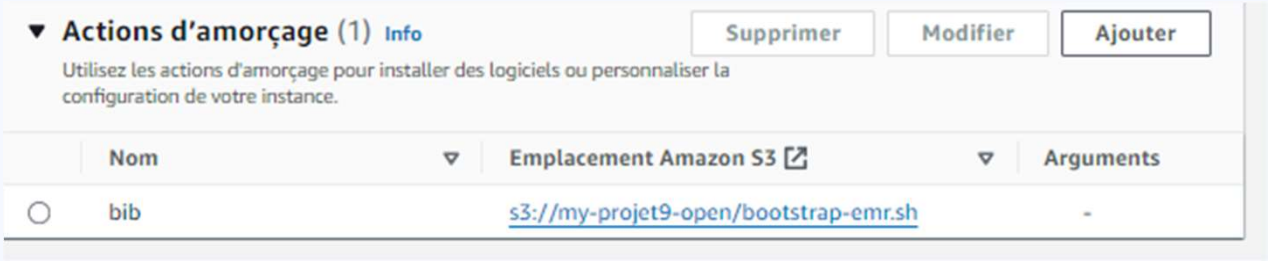


Cluster EMR
Version : emr-7.0.0

Applications sélectionnées :
Spark : Pour le traitement des données.
Hadoop : Pour la gestion des données massives.
JupyterHub : Pour l'analyse interactive et le développement.



- **Type d'instances :** m5.xlarge (4 vCPU, 16 Go de RAM) pour les nœuds principaux et les tâches.
- **Volume racine EBS :** 15 Go de stockage avec 3000 IOPS pour une performance optimale.
- **Débit :** 125 MiB/s pour assurer une vitesse de transfert de données adéquate.



Mise à jour : Le script commence par mettre à jour setuptools et pip.

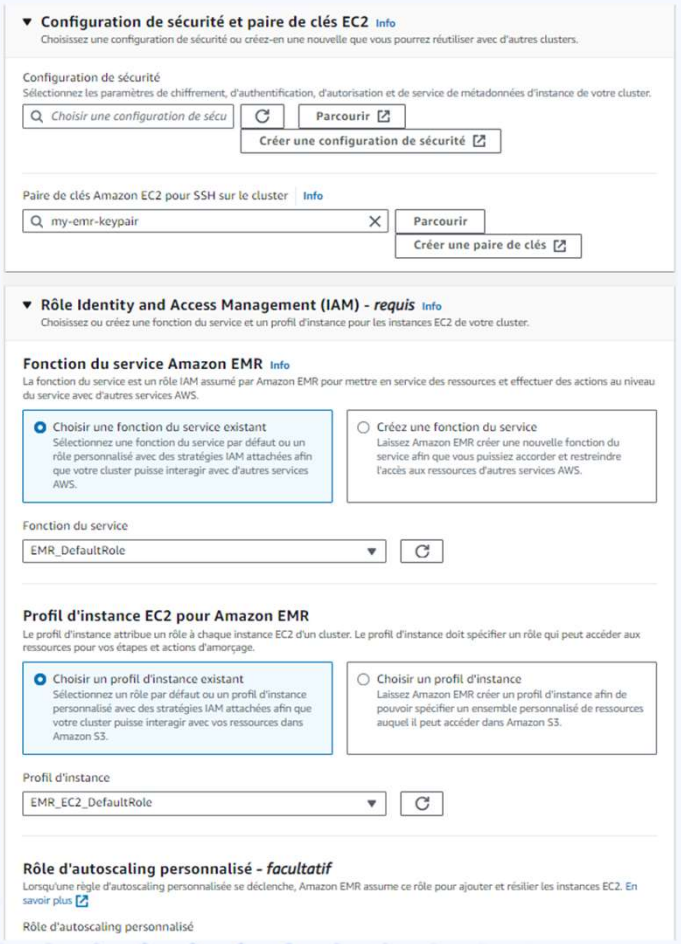
Installation des bibliothèques : Installe les bibliothèques nécessaires comme pandas, boto3, et tensorflow.

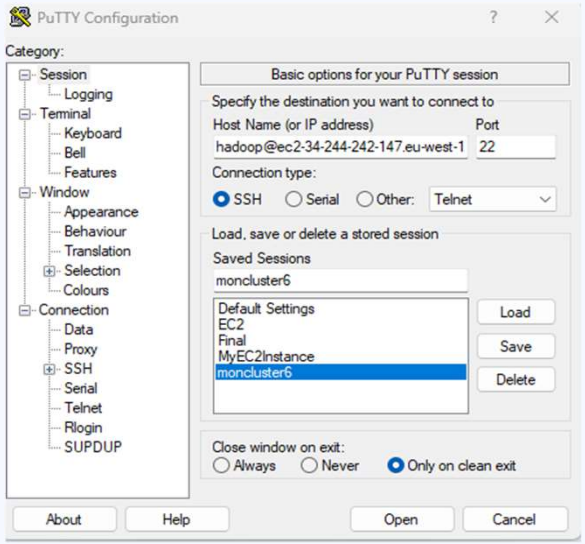
Configuration de Spark : Configure Spark History Server pour écrire les logs dans un compartiment S3 spécifique.

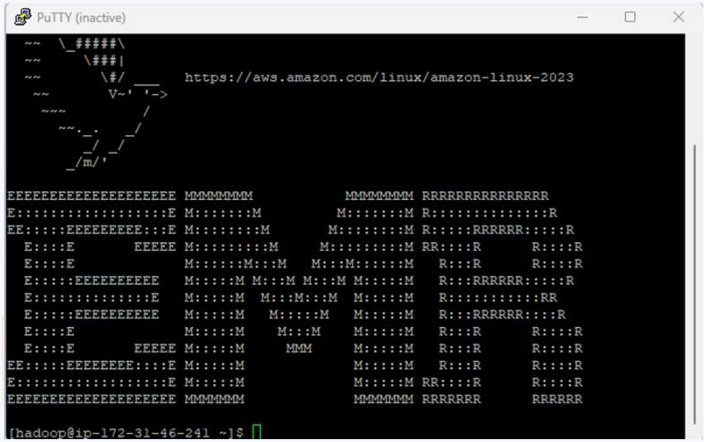
```
#!/bin/bash
# Mise à jour de setuptools et pip
sudo python3 -m pip install -U setuptools
sudo python3 -m pip install -U pip

# Installation des bibliothèques nécessaires
sudo python3 -m pip install wheel
sudo python3 -m pip install pillow
sudo python3 -m pip install pandas==1.2.5
sudo python3 -m pip install pyarrow
sudo python3 -m pip install boto3
sudo python3 -m pip install s3fs
sudo python3 -m pip install fsspec
sudo python3 -m pip install tensorflow


# Configuration de Spark History Server pour écrire les logs dans S3
sudo tee -a /etc/spark/conf/spark-defaults.conf <<EOF
spark.history.fs.logDirectory=s3://my-projet9-open/spark-logs/
EOF
```


Configuration de sécurité et paire de clés EC2 Info
Choisissez une configuration de sécurité ou créez-en une nouvelle que vous pourrez réutiliser avec d'autres clusters.
Configuration de sécurité
Sélectionnez les paramètres de chiffrement, d'authentification, d'autorisation et de service de métadonnées d'instance de votre cluster.
Choisir une configuration de sécu
Parcourir
Créer une configuration de sécurité
Paire de clés Amazon EC2 pour SSH sur le cluster Info
my-emr-keypair
Parcourir
Créer une paire de clés
Rôle Identity and Access Management (IAM) - requis Info
Choisissez ou créez une fonction du service et un profil d'instance pour les instances EC2 de votre cluster.
Fonction du service Amazon EMR Info
La fonction du service est un rôle IAM assumé par Amazon EMR pour mettre en service des ressources et effectuer des actions au niveau du service avec d'autres services AWS.
Choisir une fonction du service existant
Sélectionnez une fonction du service par défaut ou un rôle personnalisé avec des stratégies IAM attachées afin que votre cluster puisse interagir avec d'autres services AWS.
Créez une fonction du service
Laissez Amazon EMR créer une nouvelle fonction du service afin que vous puissiez accorder et restreindre l'accès aux ressources d'autres services AWS.
Fonction du service
EMR_DefaultRole
Profil d'instance EC2 pour Amazon EMR
Le profil d'instance attribue un rôle à chaque instance EC2 d'un cluster. Le profil d'instance doit spécifier un rôle qui peut accéder aux ressources pour vos étapes et actions d'amorçage.
Choisir un profil d'instance existant
Sélectionnez un rôle par défaut ou un profil d'instance personnalisé avec des stratégies IAM attachées afin que votre cluster puisse interagir avec vos ressources dans Amazon S3.
Choisir un profil d'instance
Laissez Amazon EMR créer un profil d'instance afin de pouvoir spécifier un ensemble personnalisé de ressources auquel il peut accéder dans Amazon S3.
Profil d'instance
EMR_EC2_DefaultRole
Rôle d'autoscaling personnalisé - facultatif
Lorsqu'une règle d'autoscaling personnalisée se déclenche, Amazon EMR assume ce rôle pour ajouter et résilier les instances EC2. En savoir plus
Rôle d'autoscaling personnalisé

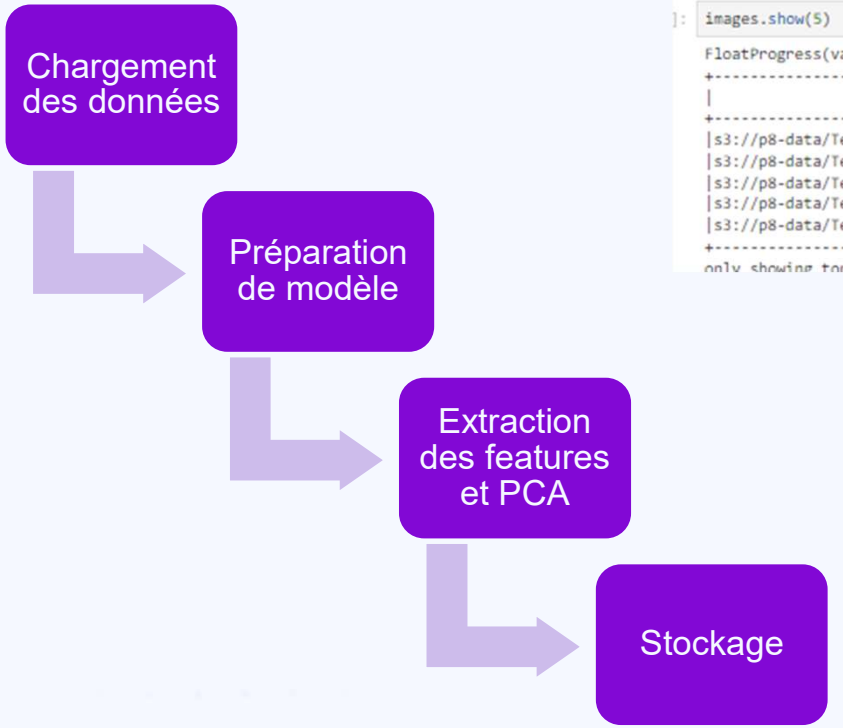

PuTTY Configuration
Category:
Session
Logging
Terminal
Keyboard
Bell
Features
Window
Appearance
Behaviour
Translation
Selection
Colours
Connection
Data
Proxy
SSH
Serial
Telnet
Rlogin
SUPDUP
Basic options for your PuTTY session
Specify the destination you want to connect to
Host Name (or IP address)
Port
hadoop@ec2-34-244-242-147.eu-west-1
22
Connection type:
SSH
Serial
Other:
Telnet
Load, save or delete a stored session
Saved Sessions
moncluster6
Default Settings
EC2
Final
MyEC2Instance
moncluster6
Load
Save
Delete
Close window on exit:
Always
Never
Only on clean exit
About
Help
Open
Cancel


PuTTY (inactive)
https://aws.amazon.com/linux/amazon-linux-2023
[hadoop@ip-172-31-46-241 ~]$

03

Chaîne de traitement des images

Shéma



6.10.5.1. 4.10.5.1 Chargement des données

```
images = spark.read.format("binaryFile") \
  .option("pathGlobFilter", "*.jpg") \
  .option("recursiveFileLookup", "true") \
  .load(PATH_Data)
```

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

```
images.show(5)
```

```
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…
+--------------------+-------------------+------+--------------------+
|                path|   modificationTime|length|             content|
+--------------------+-------------------+------+--------------------+
|s3://p8-data/Test...|2021-07-03 09:00:08|  7353|[FF D8 FF E0 00 1...|
|s3://p8-data/Test...|2021-07-03 09:00:08|  7350|[FF D8 FF E0 00 1...|
|s3://p8-data/Test...|2021-07-03 09:00:08|  7349|[FF D8 FF E0 00 1...|
|s3://p8-data/Test...|2021-07-03 09:00:08|  7348|[FF D8 FF E0 00 1...|
|s3://p8-data/Test...|2021-07-03 09:00:09|  7328|[FF D8 FF E0 00 1...|
+--------------------+-------------------+------+--------------------+
only showing top 5 rows
```

une colonne contenant les **labels** de chaque image :

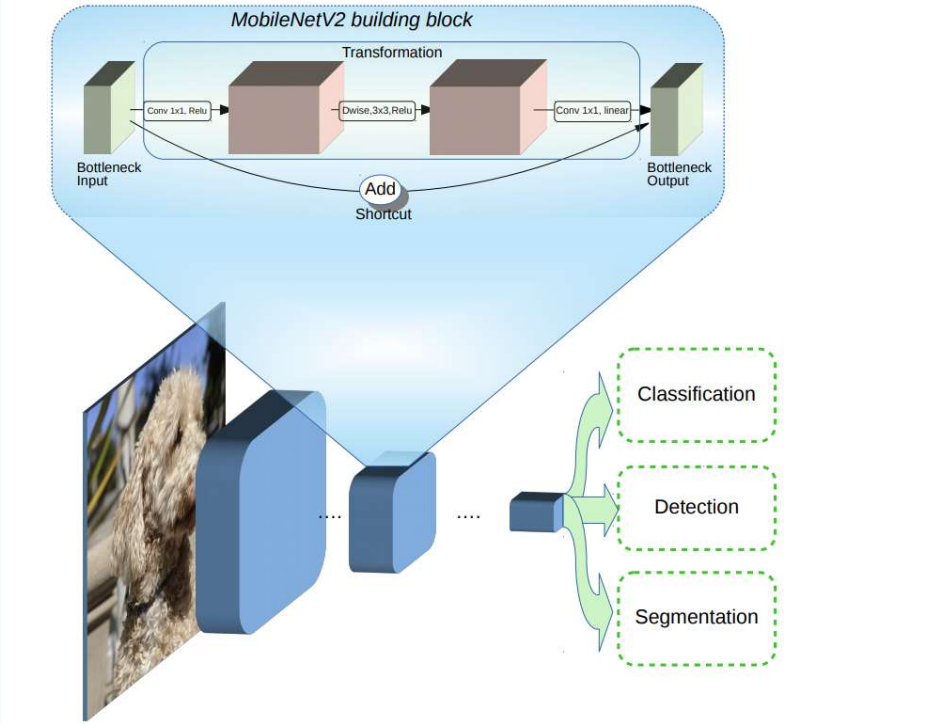
```
images = images.withColumn('label', element_at(split(images['path'], '/'),-2))
print(images.printSchema())
print(images.select('path','label').show(5,False))
```

```
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', wid
root
 |-- path: string (nullable = true)
 |-- modificationTime: timestamp (nullable = true)
 |-- length: long (nullable = true)
 |-- content: binary (nullable = true)
 |-- label: string (nullable = true)

None
+----------------------------------------+----------+
|path                                    |label     |
+----------------------------------------+----------+
|s3://p8-data/Test/Watermelon/r_106_100.jpg|Watermelon|
|s3://p8-data/Test/Watermelon/r_109_100.jpg|Watermelon|
|s3://p8-data/Test/Watermelon/r_108_100.jpg|Watermelon|
|s3://p8-data/Test/Watermelon/r_107_100.jpg|Watermelon|
|s3://p8-data/Test/Watermelon/r_95_100.jpg |Watermelon|
+----------------------------------------+----------+
only showing top 5 rows

None
```

MobileNetV2



MobileNetV2 est une architecture de réseau neuronal optimisée pour les dispositifs mobiles. L'image montre un bloc clé qui utilise des convolutions et un chemin de raccourci pour capturer efficacement les caractéristiques. Ce design permet de réaliser des tâches comme la classification, la détection, et la segmentation d'images tout en restant léger et performant.

Traiter des images avec Pandas UDFs et PySpark, en trois étapes :

- **Prétraitement :**preprocess(content): Charge et redimensionne les images à 224x224 pixels.
- **Optimisation avec UDF :**featurize_udf(content_series_iter): Applique l'extraction sur des lots d'images avec un modèle chargé une seule fois.
- **Extraction des caractéristiques :**featurize_series(model, content_series): Utilise un modèle pour extraire et aplatir les caractéristiques des images.

l'extraction et la transformation des caractéristiques (features) d'images, suivi d'une réduction de dimension avec PCA :

- **Configuration Spark :** Ajustement des paramètres Spark pour optimiser le traitement des lots d'images.
- **Extraction des caractéristiques :** Les caractéristiques des images sont extraites avec une UDF et enregistrées dans un DataFrame.
- **Enregistrement :** Les caractéristiques extraites sont enregistrées au format Parquet.

- **Conversion en vecteurs denses :** Les caractéristiques sont converties en vecteurs denses pour une meilleure compatibilité avec les algorithmes de machine learning.
- **Application du PCA :** Le PCA est appliqué pour réduire la dimensionnalité des vecteurs de caractéristiques, ce qui facilite le stockage et l'analyse ultérieure.
- **Enregistrement des résultats :** Les résultats du PCA sont sauvegardés dans un fichier Parquet.

Les données extraites sont chargées à partir de fichiers Parquet à l'aide de spark.read.parquet.

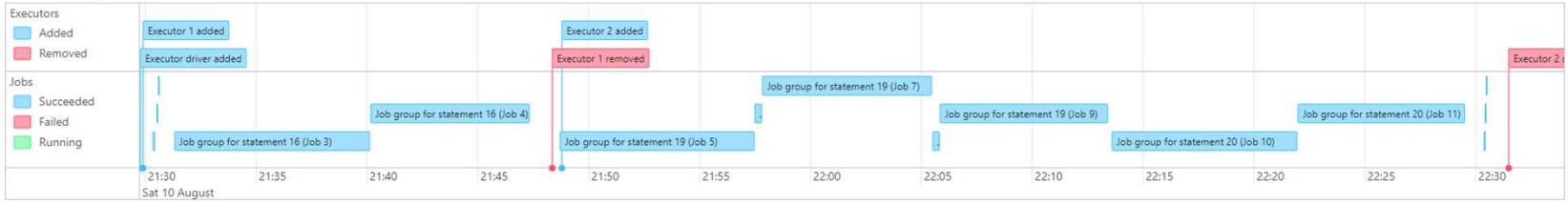
- **Chargement des données :**Les données réduites par PCA sont également chargées à partir d'un chemin spécifié.
- **Validation des résultats :**Affichage des premières lignes du DataFrame pour vérifier le contenu et la structure des données.
  Validation de la dimension des caractéristiques après réduction par PCA.

Pyspark

Spark Jobs (?)

User: livy
Total Uptime:
Scheduling Mode: FIFO
Completed Jobs: 15

▾ Event Timeline
☐ Enable zooming



▾ Completed Jobs (15)

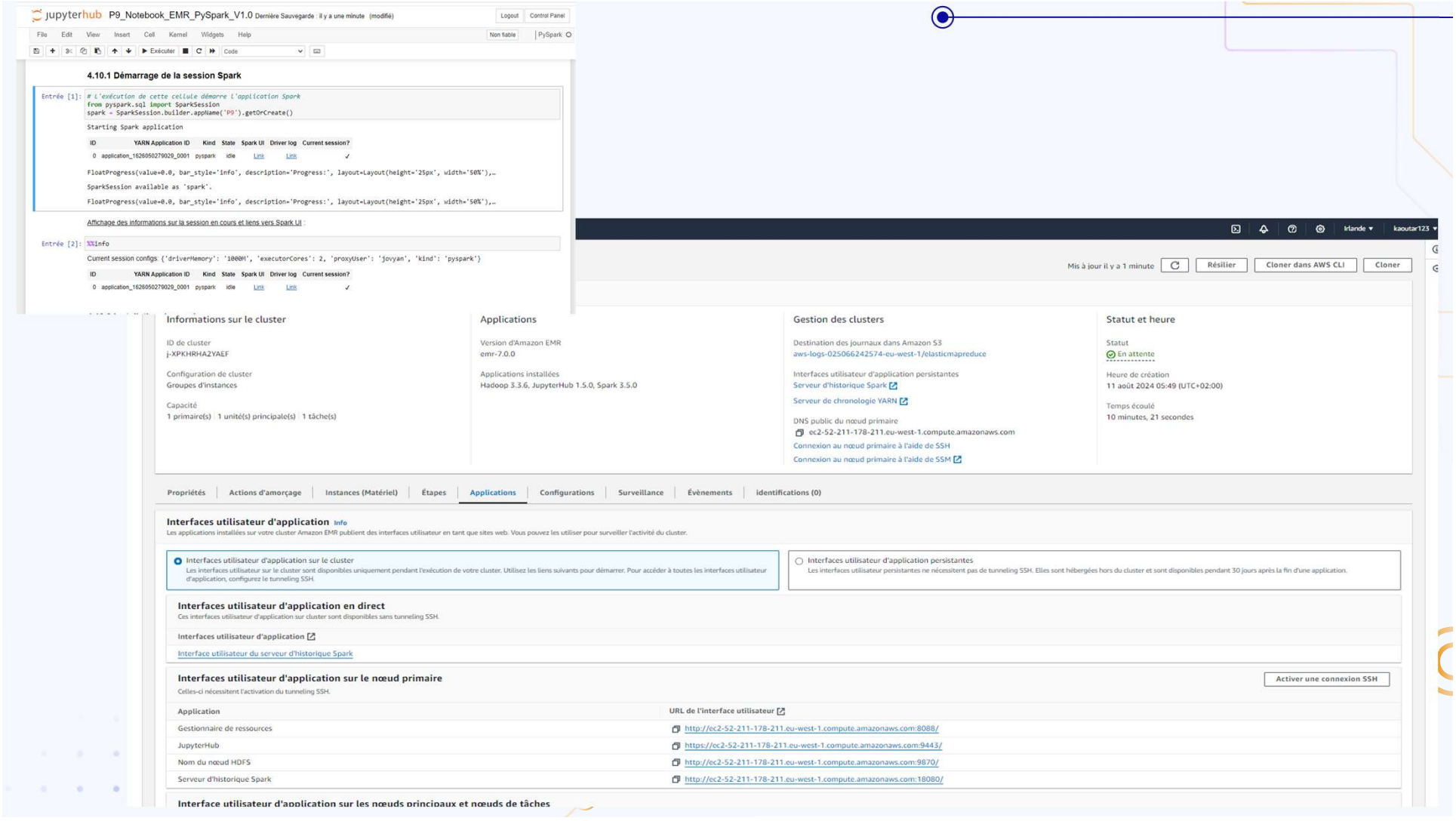
Page: 1

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
14 (25)	Job group for statement 25 head at <stdin>:1	2024/08/10 22:30:27	0.4 s	1/1	1/1
13 (24)	Job group for statement 24 parquet at NativeMethodAccessorImpl.java:0	2024/08/10 22:30:24	88 ms	1/1	1/1
12 (23)	Job group for statement 23 parquet at NativeMethodAccessorImpl.java:0	2024/08/10 22:30:21	0.2 s	1/1	1/1
11 (20)	Job group for statement 20 parquet at NativeMethodAccessorImpl.java:0	2024/08/10 22:21:58	7.5 min	1/1 (1 skipped)	24/24 (709 skipped)
10 (20)	Job group for statement 20 parquet at NativeMethodAccessorImpl.java:0	2024/08/10 22:13:34	8.4 min	1/1	709/709
9 (19)	Job group for statement 19	2024/08/10 22:05:49	7.6 min	2/2 (1 skipped)	28/28 (709 skipped)

04

Démonstration d’execution dans le cloud


jupyterhub P9_Notebook_EMR_PySpark_V1.0 Dernière Sauvegarde : il y a une minute (modifié)
Logout
Control Panel
File Edit View Insert Cell Kernel Widgets Help
Non fiable
PySpark
Exécuter
Code
4.10.1 Démarrage de la session Spark
Entrée [1]:
# L'exécution de cette cellule démarre l'application Spark
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName('P9').getOrCreate()
Starting Spark application
ID YARN Application ID Kind State Spark UI Driver log Current session?
0 application_1626050279029_0001 pyspark idle Link Link ✓
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…
SparkSession available as 'spark'.
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…
Affichage des informations sur la session en cours et liens vers Spark UI :
Entrée [2]:
%%info
Current session configs: {'driverMemory': '1000M', 'executorCores': 2, 'proxyUser': 'jovyan', 'kind': 'pyspark'}
ID YARN Application ID Kind State Spark UI Driver log Current session?
0 application_1626050279029_0001 pyspark idle Link Link ✓
Irlande
kaoutar123
Mis à jour il y a 1 minute
Résilier
Cloner dans AWS CLI
Cloner
Informations sur le cluster
ID de cluster
j-XPKHRHA2YAEF
Configuration de cluster
Groupes d'instances
Capacité
1 primaire(s) | 1 unité(s) principale(s) | 1 tâche(s)
Applications
Version d'Amazon EMR
emr-7.0.0
Applications installées
Hadoop 3.3.6, JupyterHub 1.5.0, Spark 3.5.0
Gestion des clusters
Destination des journaux dans Amazon S3
aws-logs-025066242574-eu-west-1/elasticmapreduce
Interfaces utilisateur d'application persistantes
Serveur d'historique Spark
Serveur de chronologie YARN
DNS public du nœud primaire
ec2-52-211-178-211.eu-west-1.compute.amazonaws.com
Connexion au nœud primaire à l'aide de SSH
Connexion au nœud primaire à l'aide de SSM
Statut et heure
Statut
En attente
Heure de création
11 août 2024 05:49 (UTC+02:00)
Temps écoulé
10 minutes, 21 secondes
Propriétés
Actions d'amorçage
Instances (Matériel)
Étapes
Applications
Configurations
Surveillance
Évènements
identifications (0)
Interfaces utilisateur d'application Info
Les applications installées sur votre cluster Amazon EMR publient des interfaces utilisateur en tant que sites web. Vous pouvez les utiliser pour surveiller l'activité du cluster.
Interfaces utilisateur d'application sur le cluster
Les interfaces utilisateur sur le cluster sont disponibles uniquement pendant l'exécution de votre cluster. Utilisez les liens suivants pour démarrer. Pour accéder à toutes les interfaces utilisateur d'application, configurez le tunneling SSH.
Interfaces utilisateur d'application persistantes
Les interfaces utilisateur persistantes ne nécessitent pas de tunneling SSH. Elles sont hébergées hors du cluster et sont disponibles pendant 30 jours après la fin d'une application.
Interfaces utilisateur d'application en direct
Ces interfaces utilisateur d'application sur cluster sont disponibles sans tunneling SSH.
Interfaces utilisateur d'application
Interface utilisateur du serveur d'historique Spark
Interfaces utilisateur d'application sur le nœud primaire
Celles-ci nécessitent l'activation du tunneling SSH.
Activer une connexion SSH
Application
URL de l'interface utilisateur
Gestionnaire de ressources
http://ec2-52-211-178-211.eu-west-1.compute.amazonaws.com:8088/
JupyterHub
https://ec2-52-211-178-211.eu-west-1.compute.amazonaws.com:9443/
Nom du nœud HDFS
http://ec2-52-211-178-211.eu-west-1.compute.amazonaws.com:9870/
Serveur d'historique Spark
http://ec2-52-211-178-211.eu-west-1.compute.amazonaws.com:18080/
Interface utilisateur d'application sur les nœuds principaux et nœuds de tâches

Conclusion

Thanks!

Do you have any questions?