

Projet 9 : Réalisez un traitement dans un environnement Big Data sur le Cloud

30/11/2024

Soukaina GUAOUA ELJADDI

**Parcours Data Scientist
OpenClassrooms**

Plan:

- ❑ Problématique et jeu de données
- ❑ Processus de création de l'environnement Big Data, S3 et EMR
- ❑ Réalisation de la chaîne de traitement des images dans un environnement Big Data dans le cloud
- ❑ Synthèse et Conclusion

Problématique

Contexte : “Fruits !” : jeune start-up de l’**AgriTech** qui cherche à proposer des **solutions innovantes** pour la **récolte de fruits**.

Objectifs

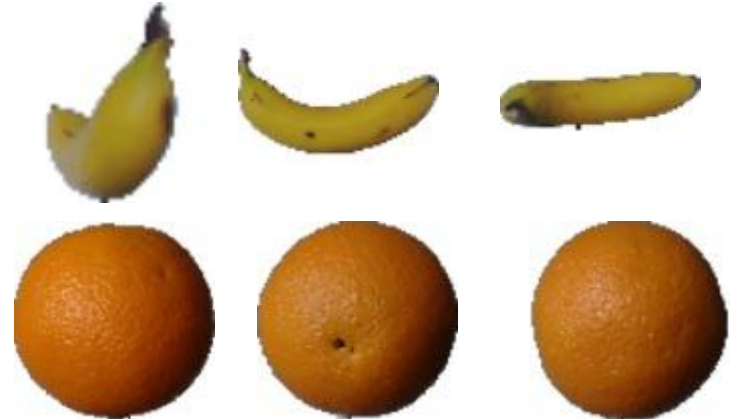
- développer des **robots cueilleurs** intelligents.
- se faire connaître grâce à une **application** grand public.

Missions :

- Mettre en place l’architecture Big Data.
- s’approprier la chaîne de traitement d’images et la compléter par une étape de réduction de dimensions.

Jeu de données 'Fruits'

- Dataset d'images Fruits-360 sur Kaggle.
- Jeu de test comprenant 23619 images de fruits (un fruit par image)
- 141 classes : Apple Red 3, Banana, Orange.....
- Un répertoire par classe, avec plusieurs photos du même fruit sous différents angles.
- Taille des images : 100x100 pixels.
- Sur fond blanc uniformisé.



Processus de création de l'environnement Big Data, S3 et EMR

Environnement Big Data :

AWS : plateforme cloud offrant des outils et services pour gérer et analyser de grandes quantités de données (Big Data)

- Amazon **EMR** (Elastic MapReduce)
- Amazon **S3** (Simple Storage Service)
- AWS **Glue**
- Évolutivité et conformité



Processus de création de l'environnement Big Data, S3 et EMR

Environnement Big Data :

PySpark : API Python d'Apache Spark, un framework de calcul distribué conçu pour traiter efficacement de grandes quantités de données.

- Traitement distribué
- Traitement en mémoire
- Bibliothèques Big Data
- Facilité d'utilisation



Processus de création de l'environnement Big Data, S3 et EMR

The screenshot shows the Amazon S3 console interface. The top navigation bar includes the AWS logo, a search bar, and user information. The left sidebar contains navigation links for Amazon S3, Compartiments, Access Grants, Points d'accès, Points d'accès de l'objet Lambda, Points d'accès multi-région, Opérations par lot, IAM Access Analyzer pour S3, Paramètres de blocage de l'accès public pour ce compte, Storage Lens, Tableaux de bord, Groupes Storage Lens, Paramètres AWS Organizations, Fonctionnalité spot, and AWS Marketplace pour S3.

The main content area displays the 'Compartiments à usage général' section. It includes a header with 'Aperçu du compte : mis à jour toutes les 24 heures' and a button 'Afficher le tableau de bord de Storage Lens'. Below this, there are tabs for 'Compartiments à usage général' and 'Compartiments de répertoires'. The 'Compartiments à usage général' tab is active, showing a table of buckets.

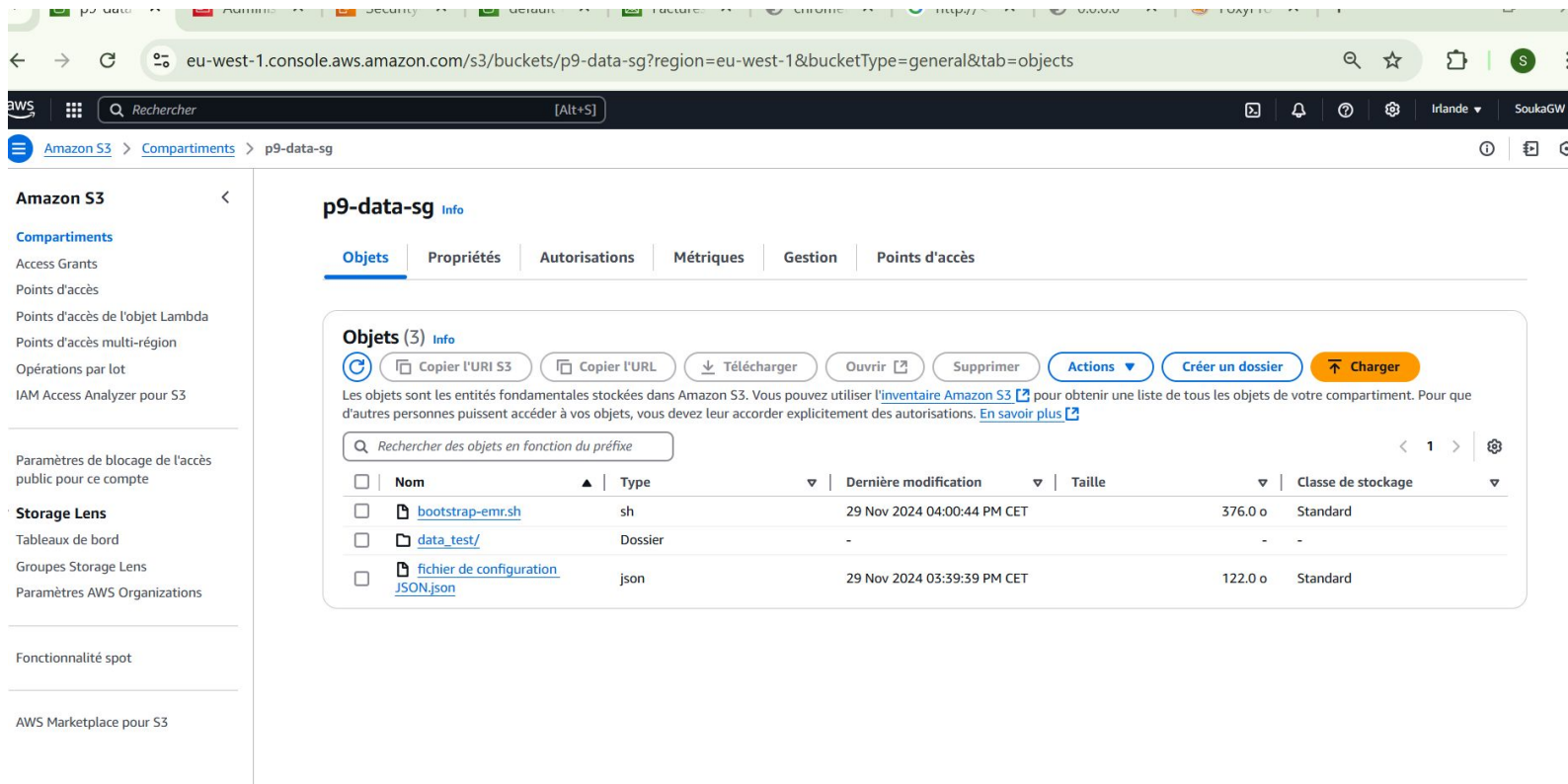
Compartiments à usage général (2) Info Toutes les régions AWS

Les compartiments sont des conteneurs pour les données stockées dans S3.

Rechercher des compartiments par nom

	Nom	Région AWS	Analyseur d'accès IAM	Date de création
<input type="radio"/>	aws-logs-324037287221-eu-west-1	Europe (Irlande) eu-west-1	Afficher l'analyseur pour eu-west-1	29 Nov 2024 02:46:04 PM CET
<input type="radio"/>	p9-data-sg	Europe (Irlande) eu-west-1	Afficher l'analyseur pour eu-west-1	29 Nov 2024 02:11:59 PM CET

Processus de création de l'environnement Big Data, S3 et EMR



The screenshot displays the AWS S3 console interface for the bucket 'p9-data-sg' in the 'eu-west-1' region. The left sidebar shows the navigation menu with 'Amazon S3' selected. The main content area shows the 'Objets' (Objects) tab, which lists three objects in a table. The table columns are 'Nom' (Name), 'Type', 'Dernière modification' (Last modified), 'Taille' (Size), and 'Classe de stockage' (Storage class). The objects are 'bootstrap-emr.sh' (sh, 376.0 o, Standard), 'data_test/' (Dossier, -, -), and 'fichier de configuration JSON.json' (json, 122.0 o, Standard). The interface includes a search bar, a list of actions (Copier l'URI S3, Copier l'URL, Télécharger, Ouvrir, Supprimer, Actions, Créer un dossier, Charger), and a description of objects in Amazon S3.

eu-west-1.console.aws.amazon.com/s3/buckets/p9-data-sg?region=eu-west-1&bucketType=general&tab=objects

Amazon S3 > Compartiments > p9-data-sg

Amazon S3

- Compartiments
- Access Grants
- Points d'accès
- Points d'accès de l'objet Lambda
- Points d'accès multi-région
- Opérations par lot
- IAM Access Analyzer pour S3

Paramètres de blocage de l'accès public pour ce compte

Storage Lens

- Tableaux de bord
- Groupes Storage Lens
- Paramètres AWS Organizations

Fonctionnalité spot

AWS Marketplace pour S3

p9-data-sg Info

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

Objets (3) Info

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](#)

Rechercher des objets en fonction du préfixe

<input type="checkbox"/>	Nom	Type	Dernière modification	Taille	Classe de stockage
<input type="checkbox"/>	bootstrap-emr.sh	sh	29 Nov 2024 04:00:44 PM CET	376.0 o	Standard
<input type="checkbox"/>	data_test/	Dossier	-	-	-
<input type="checkbox"/>	fichier de configuration JSON.json	json	29 Nov 2024 03:39:39 PM CET	122.0 o	Standard

Processus de création de l'environnement Big Data, S3 et EMR

The screenshot shows the AWS Management Console interface for an EMR cluster. The browser tabs include 'bootstrap-emr.sh - Objet', 'IAM Management Console', 'AmazonEMR-InstanceProfile', 'Propriétés > Mon cluster', and 'Factures | Billing and Cost'. The address bar shows the URL 'eu-west-1.console.aws.amazon.com/emr/home?region=eu-west-1#/clusterDetails/j-3IIHJ5UKIYSIJ'. The console header includes the AWS logo, a search bar, and navigation links for 'Amazon EMR' and 'EMR sur EC2: Clusters'. A green notification bar at the top states 'Votre cluster « Mon cluster2 » a été créé.' Below this, a 'Notifications' bar shows 0 errors, 0 warnings, 2 successes, and 0 info messages. The main heading is 'Mon cluster2' with a refresh icon and a timestamp 'Mise à jour il y a moins d'une minute'. Action buttons include 'Résilier', 'Cloner dans AWS CLI', and 'Cloner'. The 'Récapitulatif' section is divided into four columns: 'Informations sur le cluster', 'Applications', 'Gestion des clusters', and 'Statut et heure'. The 'Informations sur le cluster' column lists the cluster ID 'j-3IIHJ5UKIYSIJ', configuration 'Groupes d'instances', and capacity '1 primaire(s) | 1 unité(s) principale(s) | 2 tâche(s)'. The 'Applications' column lists 'Version d'Amazon EMR' as 'emr-6.10.0' and installed applications 'Hadoop 3.3.3, JupyterHub 1.5.0, Spark 3.3.1'. The 'Gestion des clusters' column shows 'Destination des journaux dans Amazon S3' as 'Journalisation non configurée' and 'DNS public du nœud primaire' as 'ec2-52-50-201-72.eu-west-1.compute.amazonaws.com'. The 'Statut et heure' column shows 'Statut' as 'Action d'amorçage' and 'Heure de création' as '29 novembre 2024 16:57 (UTC+01:00)'. The 'Temps écoulé' is '2 minutes, 30 secondes'.

bootstrap-emr.sh - Objet x IAM Management Console x AmazonEMR-InstanceProfile x Propriétés > Mon cluster x Factures | Billing and Cost x

eu-west-1.console.aws.amazon.com/emr/home?region=eu-west-1#/clusterDetails/j-3IIHJ5UKIYSIJ

aws Rechercher [Alt+S] Irlande SoukaGW

Amazon EMR > EMR sur EC2: Clusters > Mon cluster2

✓ Votre cluster « Mon cluster2 » a été créé.

Notifications 0 0 2 0

Mon cluster2 Mise à jour il y a moins d'une minute

Résilier Cloner dans AWS CLI Cloner

▼ Récapitulatif

Informations sur le cluster	Applications	Gestion des clusters	Statut et heure
ID de cluster j-3IIHJ5UKIYSIJ	Version d'Amazon EMR emr-6.10.0	Destination des journaux dans Amazon S3 Journalisation non configurée	Statut Action d'amorçage
Configuration de cluster Groupes d'instances	Applications installées Hadoop 3.3.3, JupyterHub 1.5.0, Spark 3.3.1	DNS public du nœud primaire ec2-52-50-201-72.eu-west-1.compute.amazonaws.com Connexion au nœud primaire à l'aide de SSH Connexion au nœud primaire à l'aide de SSM	Heure de création 29 novembre 2024 16:57 (UTC+01:00)
Capacité 1 primaire(s) 1 unité(s) principale(s) 2 tâche(s)			Temps écoulé 2 minutes, 30 secondes

Processus de création de l'environnement Big Data, S3 et EMR

eu-west-1.console.aws.amazon.com/emr/home?region=eu-west-1#/clusters

Amazon EMR > EMR sur EC2: Clusters

Clusters (9) Info

Filtrer les clusters par statut Rechercher des clusters Filtrer les clusters par date et heure de création

	ID de cluster	Nom du cluster	Statut	Heure de création (UTC+01:00)	Temps écoulé	Heures d'instances normalisées
<input type="checkbox"/>	j-3IIHJ5UKIYSIJ	Mon cluster2	Action d'amorçage Exécution des actions d'amorçage	29 novembre 2024 16:57	5 minutes, 46 secondes	0
<input type="checkbox"/>	j-3M39S88Q4J3R7	Mon cluster2	Résilié avec des erreurs Erreur de validation	29 novembre 2024 16:43	42 secondes	0
<input type="checkbox"/>	j-274SHUII3E650	Mon cluster2	Résilié avec des erreurs Échec d'amorçage	29 novembre 2024 16:39	3 minutes, 42 secondes	0
<input type="checkbox"/>	j-368PUQDHARW9X	Mon cluster1	Résilié avec des erreurs Erreur de validation	29 novembre 2024 16:33	43 secondes	0
<input type="checkbox"/>	j-28GT87IXS91AX	Mon cluster	Résilié avec des erreurs Erreur de validation	29 novembre 2024 16:01	14 minutes, 51 secondes	0
<input type="checkbox"/>	j-19J8Q79R8C9LH	p9-fruitsg	Résilié avec des erreurs Échec d'amorçage	29 novembre 2024 15:41	3 minutes, 58 secondes	0
<input type="checkbox"/>	j-2ALBNGTOD0B46	p9-fruits	Résilié avec des erreurs Échec d'amorçage	29 novembre 2024 15:24	3 minutes, 39 secondes	0
<input type="checkbox"/>	j-3PGUX1BJZ9DX	Mon cluster	Résilié avec des erreurs Échec d'amorçage	29 novembre 2024 15:08	3 minutes, 48 secondes	0
<input type="checkbox"/>	j-392RFNPFUOO3V	p9-SG	Résilié avec des erreurs Échec d'amorçage	29 novembre 2024 14:46	3 minutes, 44 secondes	0

Mode compact

Processus de création de l'environnement Big Data, S3 et EMR

The screenshot displays the Amazon EMR console interface for a cluster named 'Mon cluster2'. The browser address bar shows the URL: `eu-west-1.console.aws.amazon.com/emr/home?region=eu-west-1#/clusterDetails/j-3IIHJ5UKIYSIJ`. The console header includes the AWS logo, a search bar, and navigation links for 'Amazon EMR', 'EMR sur EC2: Clusters', and 'Mon cluster2'. A green notification bar at the top states: 'Votre cluster « Mon cluster2 » a été créé.' Below this, a 'Notifications' section shows a timeline of events. The main content area is titled 'Mon cluster2' and includes a 'Résumé' (Summary) section with the following details:

- Informations sur le cluster:** ID de cluster: j-3IIHJ5UKIYSIJ, Configuration de cluster: Groupes d'instances, Capacité: 1 primaire(s) | 1 unité(s) principale(s) | 2 tâche(s).
- Applications:** Version d'Amazon EMR: emr-6.10.0, Applications installées: Hadoop 3.3.3, JupyterHub 1.5.0, Spark 3.3.1.
- Gestion des clusters:** Destination des journaux dans Amazon S3: Journalisation non configurée, Interfaces utilisateur d'application persistantes: Serveur d'historique Spark, Serveur de chronologie YARN, DNS public du nœud primaire: ec2-52-50-201-72.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: 29 novembre 2024 16:57 (UTC+01:00), Temps écoulé: 8 minutes, 32 secondes.

Below the summary, a horizontal tab bar allows switching between different views: 'Propriétés', 'Actions d'amorçage', 'Instances (Matériel)', 'Étapes', 'Applications', 'Configurations', 'Surveillance', 'Événements', and 'Identifications (1)'. The 'Propriétés' tab is currently selected, showing several configuration sections:

- Système d'exploitation:** Version Amazon Linux: 2.0.20241031.0.
- Journaux de cluster:** Archiver les fichiers journaux dans Amazon S3: Désactivé, Chiffrement pour les journaux: Désactivé.
- Résiliation du cluster et remplacement des nœuds:** Option de résiliation: Résilier manuellement le cluster, Temps d'inactivité: -, Protection contre la résiliation: Désactivé, Remplacement des nœuds défectueux: Activé.
- Réseau et sécurité:** Réseau: Cloud privé virtuel (VPC) vpc-099c188e3971e039a, Sous-réseau(s) et zone(s) de disponibilité: subnet-0a0d694f8c6da2625 (eu-west-1c), Groupes de sécurité EC2 (pare-feu).
- Configuration de sécurité:** Configuration de sécurité: Aucun, Paire de clés EC2: cles-ppk.
- Autorisations:** Fonction du service pour Amazon EMR: AmazonEMR-ServiceRole-20241129T165726, Profil d'instance EC2: Administrateur, Rôle d'autoscaling personnalisé: Non configuré.

Processus de création de l'environnement Big Data, S3 et EMR

The screenshot displays the AWS Management Console for an Amazon EMR cluster. The browser address bar shows the URL: `eu-west-1.console.aws.amazon.com/emr/home?region=eu-west-1#/clusterDetails/j-3IIHJ5UKIYSIJ`. The console header includes the AWS logo, a search bar, and navigation links for Amazon EMR, EMR sur EC2: Clusters, and Mon cluster2. A green notification bar at the top states: "Votre cluster « Mon cluster2 » a été créé." Below this, a notifications bar shows 0 errors, 0 warnings, 2 successes, and 0 info messages. The main section is titled "Mon cluster2" and includes buttons for "Réinitialiser", "Cloner dans AWS CLI", and "Cloner". The cluster details are organized into four columns:

▼ Récapitulatif	Applications	Gestion des clusters	Statut et heure
Informations sur le cluster ID de cluster j-3IIHJ5UKIYSIJ Configuration de cluster Groupes d'instances Capacité 1 primaire(s) 1 unité(s) principale(s) 2 tâche(s)	Version d'Amazon EMR emr-6.10.0 Applications installées Hadoop 3.3.3, JupyterHub 1.5.0, Spark 3.3.1	Destination des journaux dans Amazon S3 Journalisation non configurée Interfaces utilisateur d'application persistantes Serveur d'historique Spark Serveur de chronologie YARN DNS public du nœud primaire ec2-52-50-201-72.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 ✓ En attente Heure de création 29 novembre 2024 16:57 (UTC+01:00) Temps écoulé 8 minutes, 32 secondes

Processus de création de l'environnement Big Data, S3 et EMR

The screenshot shows the AWS EMR console in the eu-west-1 region. A modal window titled 'Activer une connexion SSH' is open, providing instructions on how to access the EMR cluster's web interfaces via an SSH tunnel. The modal includes a list of steps for both Windows and Mac/Linux users, starting from downloading PuTTY to configuring the SSH connection and port forwarding. The background shows the EMR cluster details page with a sidebar containing links to various interfaces like JupyterHub and Hadoop.

eu-west-1.console.aws.amazon.com/emr/home?region=eu-west-1#/clusterDetails/j-3IIHJ5UKIYSU

Activer une connexion SSH

Les applications EMR publient des interfaces utilisateur sous forme de sites Web hébergés sur le nœud primaire. Pour des raisons de sécurité, ces sites Web ne sont disponibles que sur le serveur Web local du nœud primaire.

Pour accéder aux interfaces Web, vous devez établir un tunnel SSH avec le nœud primaire à l'aide d'une redirection de port dynamique ou locale. Si vous utilisez la redirection de port dynamique, vous devez également configurer un serveur proxy pour afficher les interfaces Web. [En savoir plus](#)

Étape 1: Ouvrez un tunnel SSH vers le nœud primaire Amazon EMR.

Windows

Mac/Linux

1. Téléchargez PuTTY.exe sur votre ordinateur à partir de : <https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html>
2. Démarrez PuTTY.
3. Dans la liste Category (Catégorie), cliquez sur Session.
4. Dans le champ Host Name (Nom d'hôte), entrez `hadoop@ec2-52-50-201-72.eu-west-1.compute.amazonaws.com`.
5. Dans la liste Category, développez Connection (Connexion) > SSH > Auth.
6. Pour Private key file for authentication (Fichier de clé privée pour l'authentification), cliquez sur Browse (Parcourir) et sélectionnez le fichier de clé privée (`c1es-ppk.ppk`) utilisé pour lancer le cluster.
7. Dans la liste Category, développez Connection > SSH, puis cliquez sur Tunnels.
8. Dans le champ Source port (Port source), tapez 8157 (port local non utilisé choisi au hasard).
9. Sélectionnez les options Dynamic (Dynamique) et Auto.
10. Laissez le champ Destination vide et cliquez sur Add (Ajouter).
11. Cliquez sur Open (Ouvrir).
12. Cliquez sur Yes (Oui) pour ignorer l'alerte de sécurité.

Étape 2: Configurez un outil de gestion de proxy. [En savoir plus](#)

Fermer

Processus de création de l'environnement Big Data, S3 et EMR

```
hadoop@ip-172-31-10-88:~$ ssh -i cles-ppk hadoop@ip-172-31-10-88
Using username "hadoop".
Authenticating with public key "cles-ppk"
Last login: Fri Nov 29 23:08:51 2024 from 179.172.116.78.rev.sfr.net

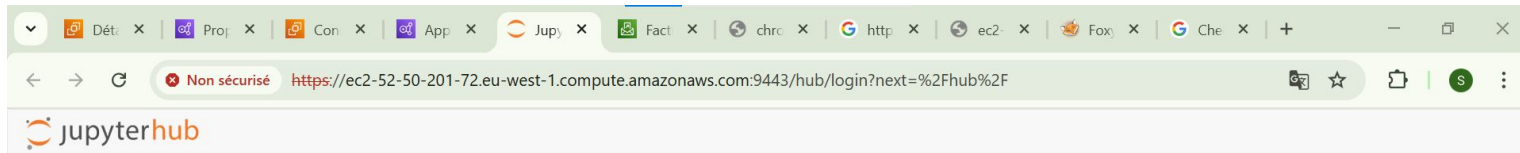
#_
###          Amazon Linux 2
#####
#####\
#####|      AL2 End of Life is 2025-06-30.
#####/
V~>
A newer version of Amazon Linux is available!
Amazon Linux 2023, GA and supported until 2028-03-15.
https://aws.amazon.com/linux/amazon-linux-2023/

4 package(s) needed for security, out of 7 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEEE MMMMMMM MRRRRRRRRRRRRRRRR
E:EEEEEEEEEEEEEEEEEE M:EEEEEE M:EEEEEE R:EEEEEE
EE:EEEEEEEEEEEEEEEEEE M:EEEEEE M:EEEEEE R:EEEEEE
E:EE EEEEE M:EEEEEE M:EEEEEE RR:EE R:EE
E:EE M:EEEE M:EEEE M:EEEE M:EEEE R:EE R:EE
E:EEEEEEEEEEEEEE M:EEEE M:EEEE M:EEEE R:EEEEEE
E:EEEEEEEEEEEEEE M:EEEE M:EEEE M:EEEE R:EEEEEE
E:EE M:EEEE M:EEEE M:EEEE M:EEEE R:EE R:EE
E:EE EEEEE M:EEEE M:EEEE M:EEEE M:EEEE R:EE R:EE
EE:EEEEEEEEEEEEEE M:EEEE M:EEEE M:EEEE R:EE R:EE
E:EEEEEEEEEEEEEE M:EEEE M:EEEE M:EEEE R:EE R:EE
EEEEEEEEEEEEEEEEEEEE MMMMMMM MRRRRRR RRRRRR

[hadoop@ip-172-31-10-88 ~]$
```

Processus de création de l'environnement Big Data, S3 et EMR



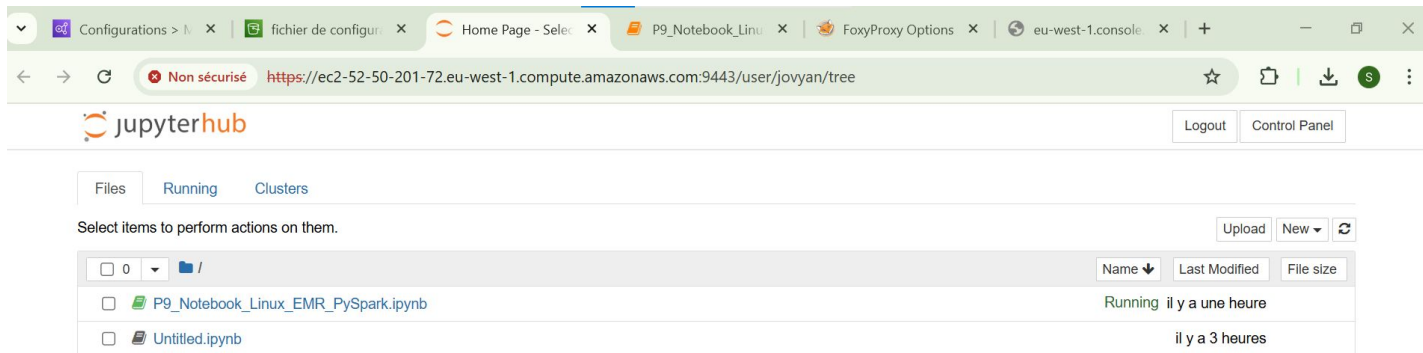
Sign in

Username:

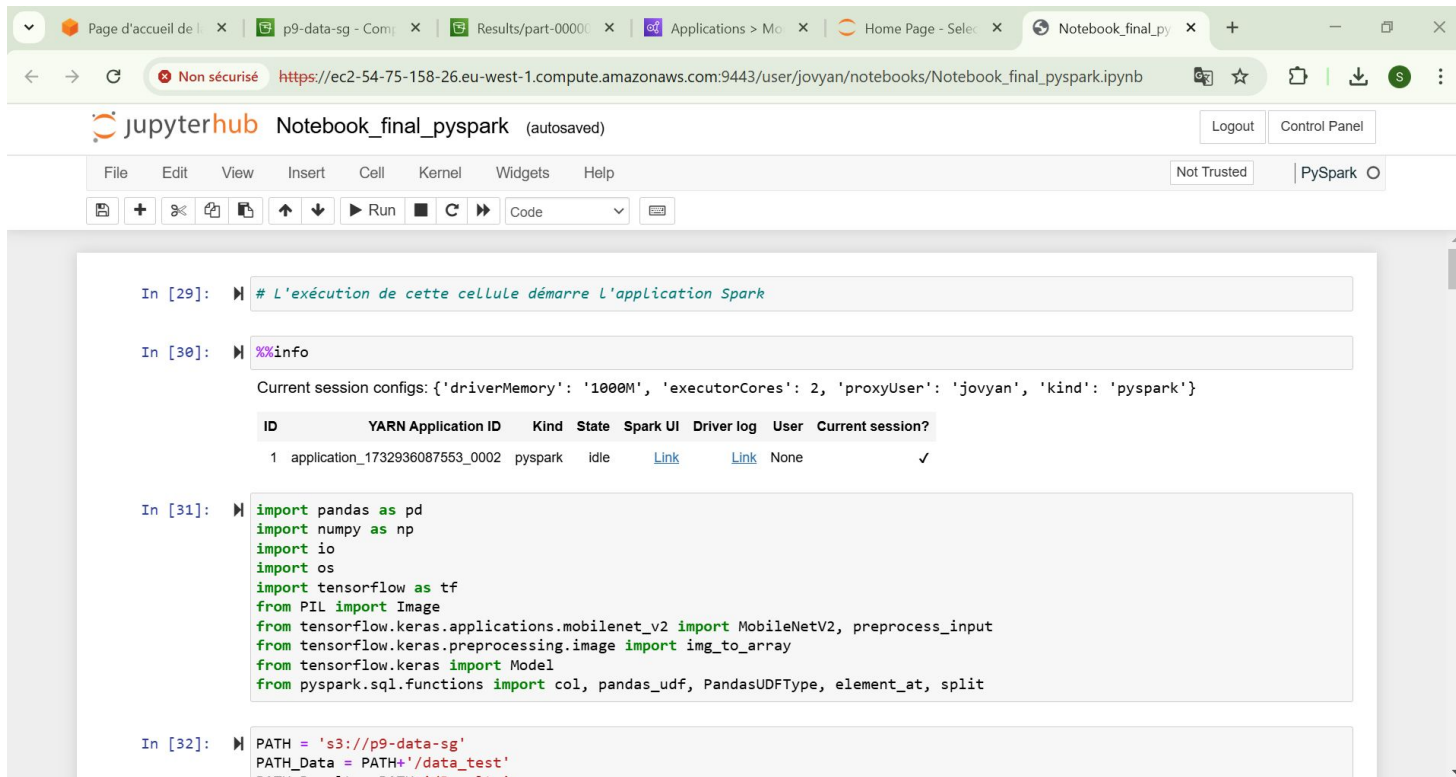
Password:

Sign in

Réalisation de la chaîne de traitement des images dans un environnement Big Data dans le cloud



Réalisation de la chaîne de traitement des images dans un environnement Big Data dans le cloud



The screenshot displays a Jupyter Notebook titled "Notebook_final_pyspark" running on a JupyterHub instance. The browser address bar shows the URL: https://ec2-54-75-158-26.eu-west-1.compute.amazonaws.com:9443/user/jovyan/notebooks/Notebook_final_pyspark.ipynb. The notebook interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help) and a toolbar with icons for file operations and execution. The code cells are as follows:

```
In [29]: # L'exécution de cette cellule démarre l'application Spark
```

```
In [30]: %%info
```

Current session configs: {'driverMemory': '1000M', 'executorCores': 2, 'proxyUser': 'jovyan', 'kind': 'pyspark'}

ID	YARN Application ID	Kind	State	Spark UI	Driver log	User	Current session?
1	application_1732936087553_0002	pyspark	idle	Link	Link	None	✓

```
In [31]: import pandas as pd
import numpy as np
import io
import os
import tensorflow as tf
from PIL import Image
from tensorflow.keras.applications.mobilenet_v2 import MobileNetV2, preprocess_input
from tensorflow.keras.preprocessing.image import img_to_array
from tensorflow.keras import Model
from pyspark.sql.functions import col, pandas_udf, PandasUDFType, element_at, split
```

```
In [32]: PATH = 's3://p9-data-sg'
PATH_Data = PATH+'/data_test'
```

Réalisation de la chaîne de traitement des images dans un environnement Big Data dans le cloud

The screenshot shows a Jupyter Notebook titled "Notebook_final_pyspark" running on a JupyterHub instance. The browser address bar shows the URL: `https://ec2-54-75-158-26.eu-west-1.compute.amazonaws.com:9443/user/jovyan/notebooks/Notebook_final_pyspark.ipynb`. The notebook interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help) and a toolbar with icons for file operations and execution. The main area displays a PySpark DataFrame with columns: path, modificationTime, length, and content. The first five rows of data are shown, representing image files in an S3 bucket. Below the data preview, the code for creating the DataFrame is visible, followed by the output of the schema command, which shows the data types for each column.

```
| path | modificationTime | length | content |
+-----+-----+-----+-----+
|s3://p9-data-sg/d...|2024-11-29 13:23:00|5492|[FF D8 FF E0 00 1...|
|s3://p9-data-sg/d...|2024-11-29 13:23:03|5482|[FF D8 FF E0 00 1...|
|s3://p9-data-sg/d...|2024-11-29 13:22:59|5479|[FF D8 FF E0 00 1...|
|s3://p9-data-sg/d...|2024-11-29 13:23:02|5467|[FF D8 FF E0 00 1...|
|s3://p9-data-sg/d...|2024-11-29 13:23:00|5454|[FF D8 FF E0 00 1...|
+-----+-----+-----+-----+
only showing top 5 rows

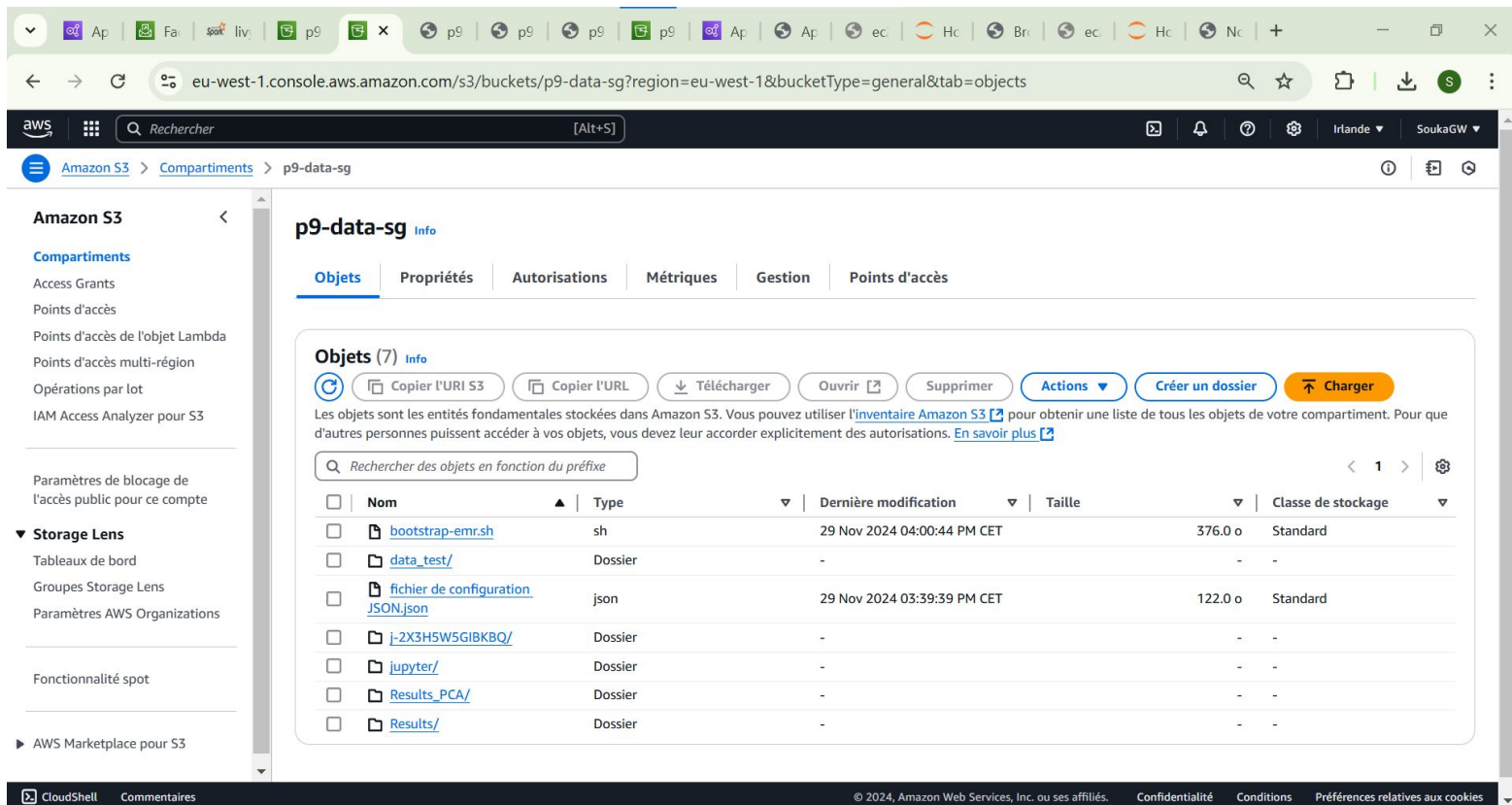
In [35]: images = images.withColumn('label', element_at(split(images['path'], '/'),-2))
print(images.printSchema())
print(images.select('path','label').show(5,False))

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://p9-data-sg/data_test/orange_sample/92_100.jpg|orange_sample|
|s3://p9-data-sg/data_test/orange_sample/r_46_100.jpg|orange_sample|
```

Réalisation de la chaîne de traitement des images dans un environnement Big Data dans le cloud



The screenshot shows the AWS S3 console interface for the bucket 'p9-data-sg' in the 'eu-west-1' region. The left sidebar contains navigation links for Amazon S3, Compartiments, Access Grants, Points d'accès, Points d'accès de l'objet Lambda, Points d'accès multi-région, Opérations par lot, IAM Access Analyzer pour S3, Paramètres de blocage de l'accès public pour ce compte, Storage Lens, Tableaux de bord, Groupes Storage Lens, Paramètres AWS Organizations, Fonctionnalité spot, and AWS Marketplace pour S3. The main content area shows the 'Objets (7)' tab with a list of objects. The objects are:

	Nom	Type	Dernière modification	Taille	Classe de stockage
<input type="checkbox"/>	bootstrap-emr.sh	sh	29 Nov 2024 04:00:44 PM CET	376.0 o	Standard
<input type="checkbox"/>	data_test/	Dossier	-	-	-
<input type="checkbox"/>	fichier de configuration JSON.json	json	29 Nov 2024 03:39:39 PM CET	122.0 o	Standard
<input type="checkbox"/>	i-2X3H5W5GIBKBQ/	Dossier	-	-	-
<input type="checkbox"/>	jupyter/	Dossier	-	-	-
<input type="checkbox"/>	Results_PCA/	Dossier	-	-	-
<input type="checkbox"/>	Results/	Dossier	-	-	-

At the bottom of the page, there is a footer with the text '© 2024, Amazon Web Services, Inc. ou ses affiliés.' and links for Confidentialité, Conditions, and Préférences relatives aux cookies.

Réalisation de la chaîne de traitement des images dans un environnement Big Data dans le cloud

The screenshot displays the AWS S3 console interface for the 'data_test' bucket. The breadcrumb navigation shows the path: Amazon S3 > Compartiments > p9-data-sg > data_test/. The left sidebar contains navigation links for Amazon S3, Compartiments, Access Grants, Points d'accès, Points d'accès de l'objet Lambda, Points d'accès multi-région, Opérations par lot, IAM Access Analyzer pour S3, Paramètres de blocage de l'accès public pour ce compte, Storage Lens, Tableaux de bord, Groupes Storage Lens, Paramètres AWS Organizations, Fonctionnalité spot, and AWS Marketplace pour S3.

The main content area shows the 'data_test/' bucket. It includes a 'Copier l'URI S3' button and tabs for 'Objets' and 'Propriétés'. The 'Objets' tab is active, showing a list of objects. Above the list are buttons for 'Copier l'URI S3', 'Copier l'URL', 'Télécharger', 'Ouvrir', 'Supprimer', 'Actions', 'Créer un dossier', and 'Charger'. A search bar is present with the text 'Rechercher des objets en fonction du préfixe'. The object list has columns for 'Nom', 'Type', 'Dernière modification', 'Taille', and 'Classe de stockage'.

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

Réalisation de la chaîne de traitement des images dans un environnement Big Data dans le cloud

eu-west-1.console.aws.amazon.com/s3/buckets/p9-data-sg?region=eu-west-1&bucketType=general&prefix=Results/&showversions=false

Amazon S3 > Compartiments > p9-data-sg > Results/

Amazon S3

- Compagniments
- Access Grants
- Points d'accès
- Points d'accès de l'objet Lambda
- Points d'accès multi-région
- Opérations par lot
- IAM Access Analyzer pour S3
- Paramètres de blocage de l'accès public pour ce compte
- Storage Lens
 - Tableaux de bord
 - Groupes Storage Lens
 - Paramètres AWS Organizations
- Fonctionnalité spot
- AWS Marketplace pour S3

Results/

Objets (16) Info

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](#)

Rechercher des objets en fonction du préfixe

	Nom	Type	Dernière modification	Taille	Classe de stockage
<input type="checkbox"/>	_SUCCESS	-	30 Nov 2024 01:53:32 PM CET		0 o Standard
<input type="checkbox"/>	part-00000-cdd375cd-e837-4b20-83a2-8eee6d2b70d1-c000.snappy.parquet	parquet	30 Nov 2024 01:53:17 PM CET	11.0 Ko	Standard
<input type="checkbox"/>	part-00001-cdd375cd-e837-4b20-83a2-8eee6d2b70d1-c000.snappy.parquet	parquet	30 Nov 2024 01:53:17 PM CET	11.2 Ko	Standard
<input type="checkbox"/>	part-00002-cdd375cd-e837-4b20-83a2-8eee6d2b70d1-c000.snappy.parquet	parquet	30 Nov 2024 01:53:19 PM CET	5.4 Ko	Standard
<input type="checkbox"/>	part-00003-cdd375cd-e837-4b20-83a2-8eee6d2b70d1-c000.snappy.parquet	parquet	30 Nov 2024 01:53:19 PM CET	5.6 Ko	Standard
<input type="checkbox"/>	part-00004-cdd375cd-e837-4b20-83a2-8eee6d2b70d1-c000.snappy.parquet	parquet	30 Nov 2024 01:53:21 PM CET	5.8 Ko	Standard
<input type="checkbox"/>	part-00005-cdd375cd-e837-4b20-83a2-8eee6d2b70d1-c000.snappy.parquet	parquet	30 Nov 2024 01:53:21 PM CET	6.1 Ko	Standard

Réalisation de la chaîne de traitement des images dans un environnement Big Data dans le cloud

The screenshot shows the AWS S3 console interface. The left sidebar contains the navigation menu with 'Storage Lens' expanded. The main content area displays the 'Results_PCA/' bucket. The 'Objets' tab is selected, showing a list of objects. The table has columns for Name, Type, Dernière modification, Taille, and Classe de stockage. Three objects are listed, all of type 'parquet' and size 39.1 Ko or 40.2 Ko.


	Nom	Type	Dernière modification	Taille	Classe de stockage
<input type="checkbox"/>	_SUCCESS	-	30 Nov 2024 06:14:43 AM CET	0 o	Standard
<input type="checkbox"/>	part-00000-e1afd1aa-3d76-4e65-830d-5fb1a40e820a-c000.snappy.parquet	parquet	30 Nov 2024 06:14:43 AM CET	39.1 Ko	Standard
<input type="checkbox"/>	part-00001-e1afd1aa-3d76-4e65-830d-5fb1a40e820a-c000.snappy.parquet	parquet	30 Nov 2024 06:14:43 AM CET	40.2 Ko	Standard

Réalisation de la chaîne de traitement des images dans un environnement Big Data dans le cloud

EMR | x p9-da x p9-da x p9-da x p9-da x p9-da x Appli x Appli x Home x Notel x +

Non sécurisé ec2-54-75-158-26.eu-west-1.compute.amazonaws.com:8088/cluster/app/application_1732936087553_0002

Logged in as: dr.who



Application application_1732936087553_0002

Cluster

[About](#)
[Nodes](#)
[Node Labels](#)
[Applications](#)
NEW
NEW_SAVING
SUBMITTED
ACCEPTED
RUNNING
FINISHED
FAILED
KILLED
[Scheduler](#)

Kill Application


Application Overview

User:	livy
Name:	livy-session-1
Application Type:	SPARK
Application Tags:	livy-session-1-zdvumipe
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	RUNNING: AM has registered with RM and started running.
Queue:	default
FinalStatus Reported by AM:	Application has not completed yet.
Started:	Sat Nov 30 04:43:05 +0000 2024
Launched:	Sat Nov 30 04:43:05 +0000 2024
Finished:	N/A
Elapsed:	8hrs, 28mins, 1sec
Tracking URL:	ApplicationMaster
Log Aggregation Status:	NOT_START
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	false
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

Réalisation de la chaîne de traitement des images dans un environnement Big Data dans le cloud

EMR | p9-da | p9-da | p9-da | p9-da | p9-da | Applic | All App | Home | Note | +

Non sécurisé ec2-54-75-158-26.eu-west-1.compute.amazonaws.com:8088/cluster

 **All Applications**

Cluster

- About
- Nodes
- Node Labels
- Applications
- NEW
- NEW SAVING
- SUBMITTED
- ACCEPTED
- RUNNING
- FINISHED
- FAILED
- KILLED
- Scheduler

Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	Total Resources	Reserved Resources	Physical Mem Used %
2	0	1	1	1	<memory:1.38 GB, vCores:1>	<memory:36 GB, vCores:12>	<memory:0 B, vCores:0>	23

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
0	0	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority
Capacity Scheduler	[memory-mb (unit-MB), vcores]	<memory:32, vCores:1>	<memory:12288, vCores:4>	0

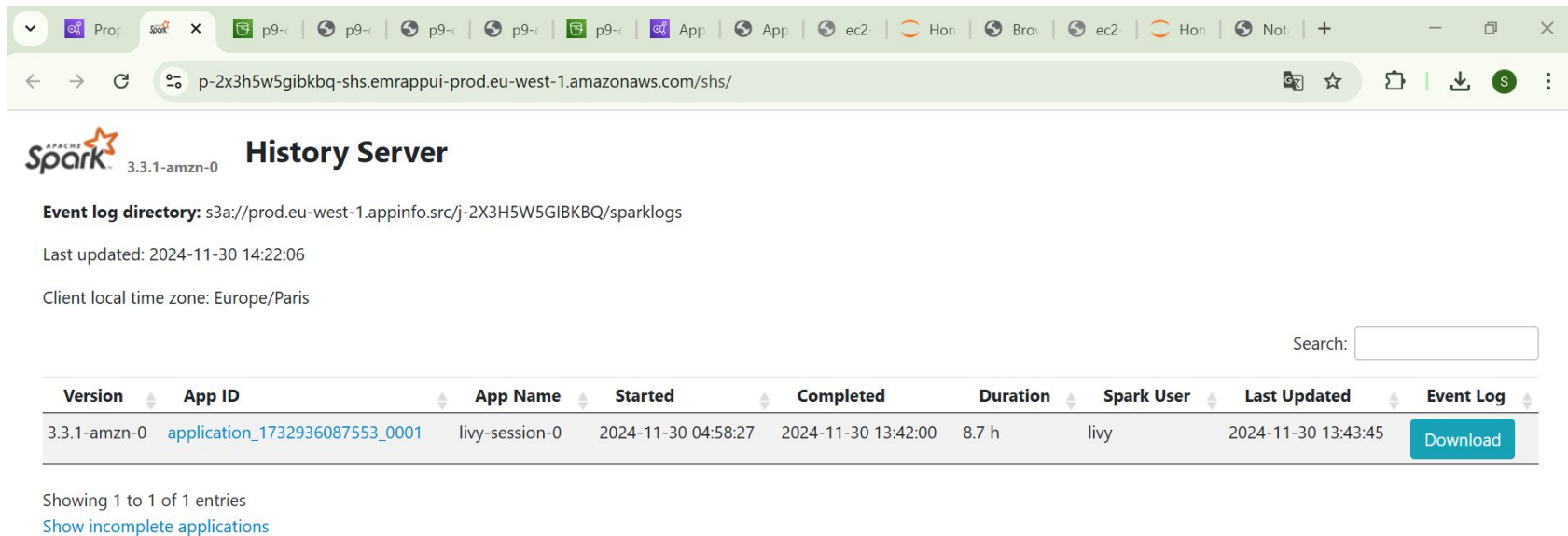
Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	Final Status	Running Containers	Allocated CPU V-Cores	Allocated Memory MB	Allocated GPUs	Reserved CPU V-Cores	Reserved Memory MB	Reserved GPUs	% of Queue	% of Cluster	Progress
application_1732936087553_0002	livy	livy-session-1	SPARK	livy-session-1-zdrvmpe	default	0	Sat Nov 30 05:43:05 +0100 2024	Sat Nov 30 05:43:05 +0100 2024	N/A	RUNNING	UNDEFINED	1	1	1408	-1	0	0	-1	3.8	3.8	<div></div>
application_1732936087553_0001	livy	livy-session-0	SPARK	livy-session-0-hzirduhf	default	0	Sat Nov 30 04:58:15 +0100 2024	Sat Nov 30 04:58:17 +0100 2024	Sat Nov 30 13:42:01 +0100 2024	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.0	0.0	<div></div>

Showing 1 to 2 of 2 entries

First

Réalisation de la chaîne de traitement des images dans un environnement Big Data dans le cloud



The screenshot shows the Apache Spark History Server interface in a web browser. The browser's address bar displays the URL: `p-2x3h5w5gibkbq-shs.emrappui-prod.eu-west-1.amazonaws.com/shs/`. The page header includes the Apache Spark logo and the text "History Server" and "3.3.1-amzn-0". Below the header, the "Event log directory" is listed as `s3a://prod.eu-west-1.appinfo.src/j-2X3H5W5GIBKBQ/sparklogs`. The "Last updated" timestamp is "2024-11-30 14:22:06", and the "Client local time zone" is "Europe/Paris". A search bar is located on the right side of the page. The main content area features a table with the following columns: Version, App ID, App Name, Started, Completed, Duration, Spark User, Last Updated, and Event Log. A single entry is shown in the table, with a "Download" button in the Event Log column. Below the table, it states "Showing 1 to 1 of 1 entries" and provides a link to "Show incomplete applications".

Event log directory: `s3a://prod.eu-west-1.appinfo.src/j-2X3H5W5GIBKBQ/sparklogs`

Last updated: 2024-11-30 14:22:06

Client local time zone: Europe/Paris

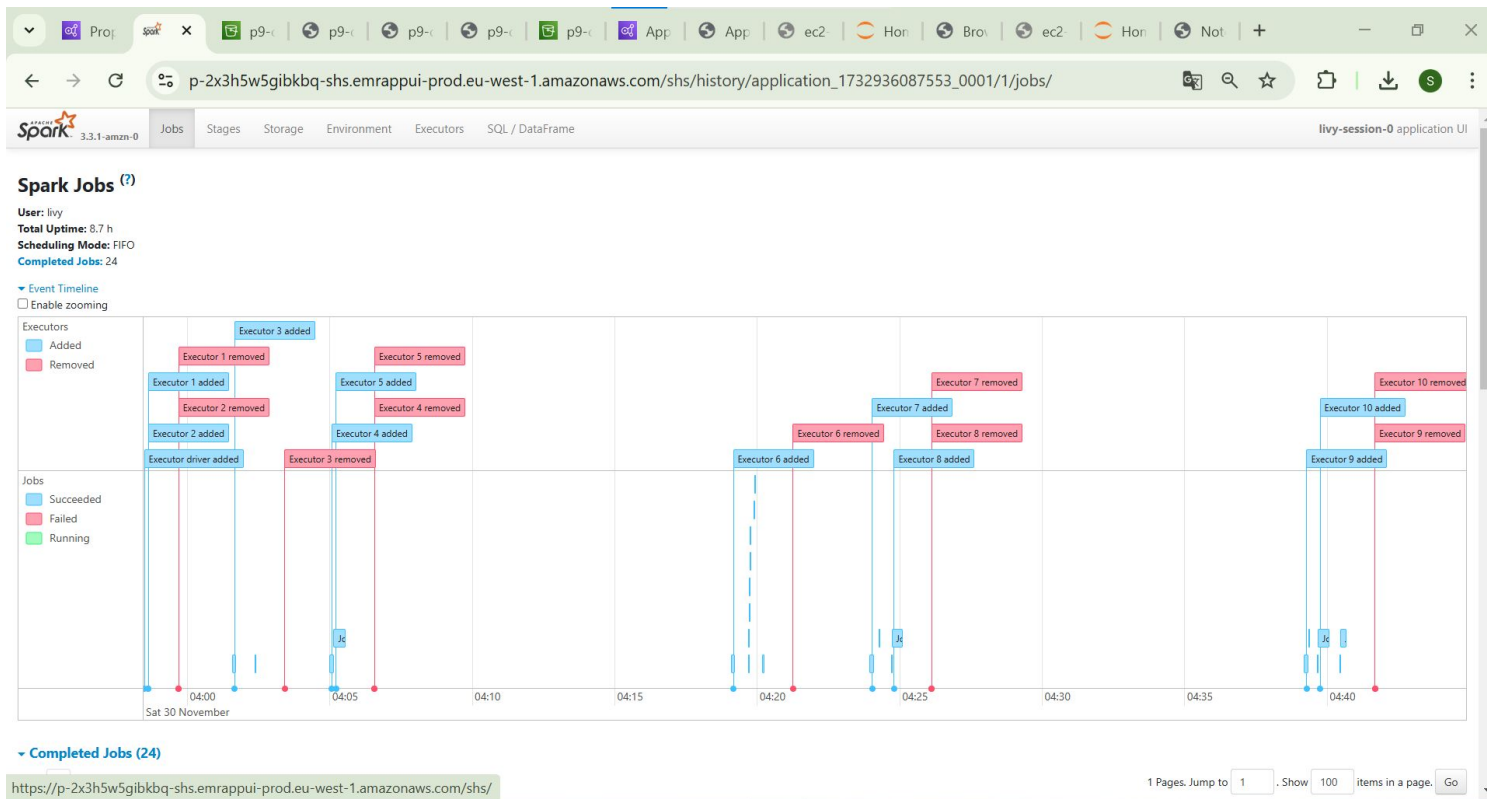
Search:

Version	App ID	App Name	Started	Completed	Duration	Spark User	Last Updated	Event Log
3.3.1-amzn-0	application_1732936087553_0001	livy-session-0	2024-11-30 04:58:27	2024-11-30 13:42:00	8.7 h	livy	2024-11-30 13:43:45	Download

Showing 1 to 1 of 1 entries

[Show incomplete applications](#)

Réalisation de la chaîne de traitement des images dans un environnement Big Data dans le cloud



Synthèse et Conclusion

1. Mise en place d'une architecture Big Data :

- Déploiement d'un cluster **EMR (Elastic MapReduce)** intégrant **Apache Spark**, permettant un traitement distribué et efficace des données volumineuses. Cette architecture inclut des outils essentiels comme Spark, Hadoop, JupyterHub et TensorFlow.
- Utilisation de **S3 (Simple Storage Service)** pour stocker les données, qu'il s'agisse des images originales ou des résultats obtenus.
- Gestion des accès et des permissions assurée via **IAM (Identity & Access Management)**.

Synthèse et Conclusion

2. Maîtrise de la chaîne de traitement d'images :

- Chargement et prétraitement des données.
- Mise en place du modèle **MobileNetV2** avec transfert d'apprentissage et ajustement des poids.
- Extraction des caractéristiques, suivie de la réduction dimensionnelle pour une analyse optimisée.

Synthèse et Conclusion

3. Avantages offerts par l'environnement Big Data pour "Fruits!" :

- **Évolutivité et performance** : L'architecture permet d'adapter facilement la charge de travail en redimensionnant le cluster en fonction des besoins.
- **Optimisation des coûts** : Bien que les coûts augmentent avec l'utilisation, ils restent inférieurs à ceux liés à l'acquisition de matériel ou à la location de serveurs dédiés.
- **Préparation à l'avenir** : Cette infrastructure pose les bases pour des fonctionnalités avancées, comme l'entraînement de modèles de classification des fruits, ouvrant la voie à des analyses plus sophistiquées.