

Déployez un modèle dans le cloud

PROJET 08/ Openclassrooms
Gulsum Kapanoglu



Dans ce Project..

- ✓ Problématique
- ✓ Objectifs dans ce projet
- ✓ Présentation des données
- ✓ Big data
- ✓ Spark/Pyspark
- ✓ Architecture big data
- ✓ Application de la solution sur en local
- ✓ Application de la solution sur le Cloud (AWS)
- ✓ Conclusion

Dans ce Project..

- **Déroulement des étapes du projet**
Le projet va être réalisé en 2 temps, dans deux environnements différents.
Dans un premier temps développer et exécuter code en local,
Déploiement de la solution dans le cloud

Problématique

Problématique

«Fruits» start-up de l'AgriTech souhaite proposer une solution innovante de récolte des fruits avec des robots cueilleurs intelligents.

Mettre en place une application mobile de reconnaissance des fruits.



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.

Fruits!

Objectifs dans ce projet

- Développer une première chaîne de traitement des données qui Il n'est pas nécessaire d'entraîner un modèle pour le moment. 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
 - à utiliser le cloud AWS pour profiter d'une architecture Big Data (EMR, S3, IAM)

Objectifs dans ce projet

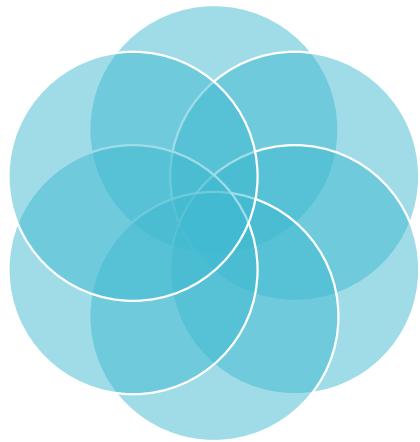
Ce qui est attendu :

- développer une première architecture Big Data :
- ✓ Préprocessing des images et réduction de dimension
- ✓ Anticipation du passage à l'échelle

Présentation des données

Fruits extrait de la
photo par un
algorithme de
Machine Learning

Fruits filmés
à **360°**,
sur **fond blanc**

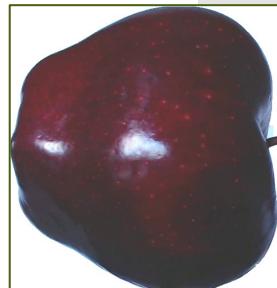


Taille des images:
100x100 pixels

Nombre total
d'images: **22688**

Nombre d'images
à traiter : **762**

Nombres de
classes : **131**



Modele

- Je décide de partir sur une solution de **transfert learning**.
- Simplement, le **transfert learning** consiste à utiliser la connaissance déjà acquise par un modèle entraîné (ici **MobileNetV2**) pour l'adapter à notre problématique

Big data



Big data

Définition : stratégies et technologies mises en œuvre pour rassembler, organiser, stocker et analyser de vastes ensembles de données.

Vitesse

Analyse en temps réel

Variété

Différence de format

Volume

Très grande quantité de données

Véracité : Cohérence, fiabilité et qualité

Solutions de stockage big data



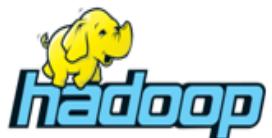
Google
Cloud Storage



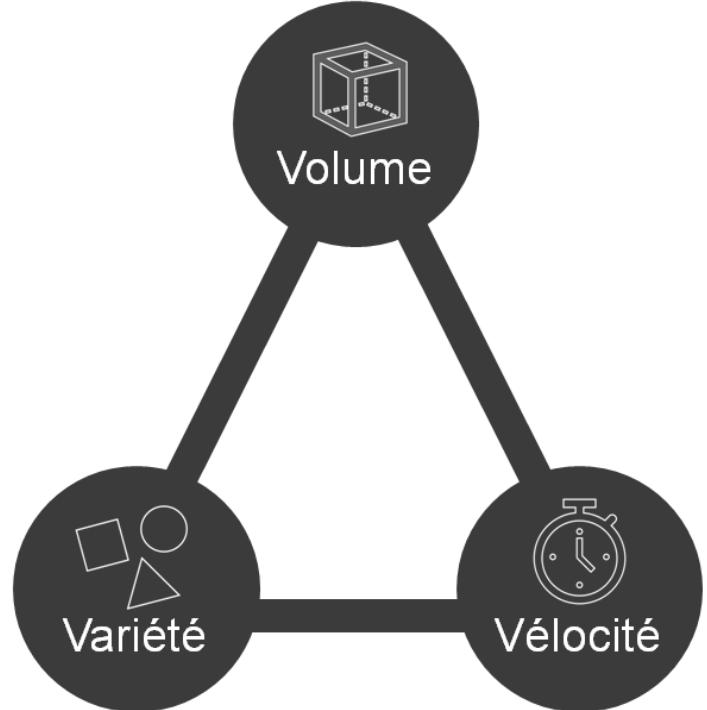
Microsoft Azure
Blob Storage

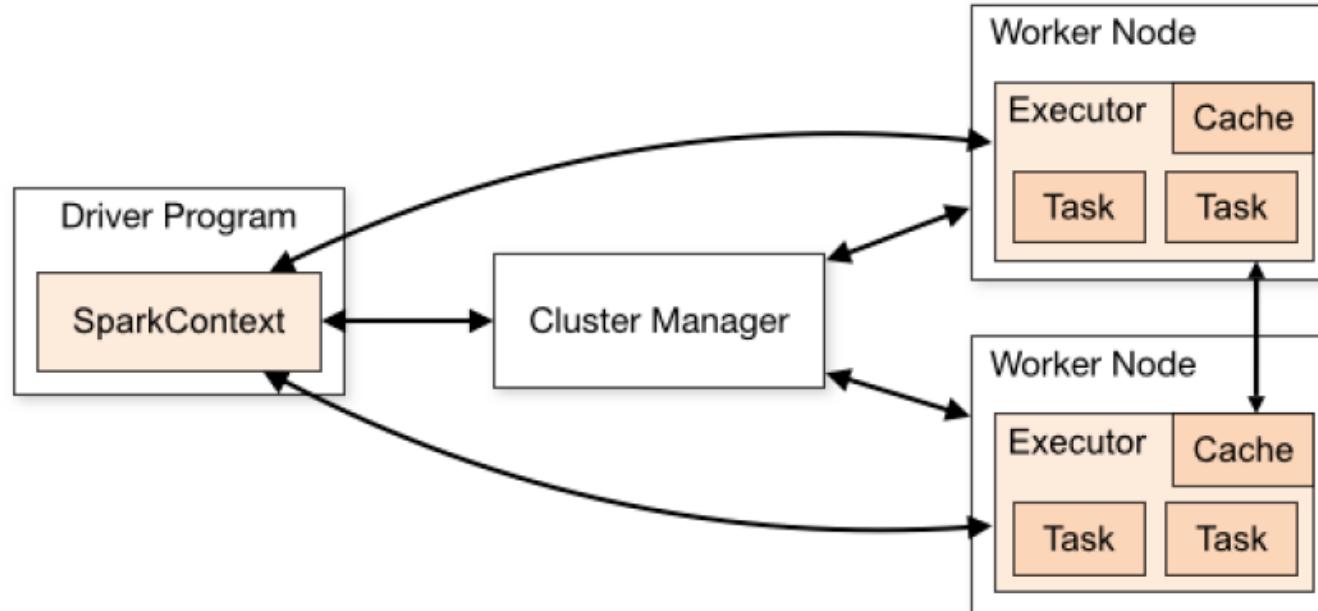


Amazon Web Services
S3



Apache
Hadoop





Application maître :
Configuration /
Initialisation
Aggrégation des calculs

Cluster Manager :
Gestion des ressources
Distribution des calculs
entre les workers

Workers :
Exécution des tâches
en parallèle

CALCULS DISTRIBUÉS

- Diviser les opérations en micro opérations distribuables entre différentes machines, réalisables en parallèle
- Agréger les résultats sur une même machine

Spark Principe de fonctionnement

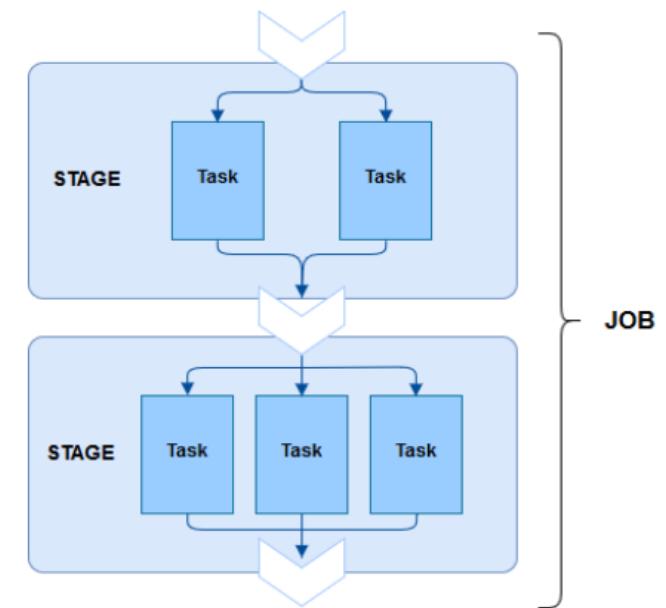
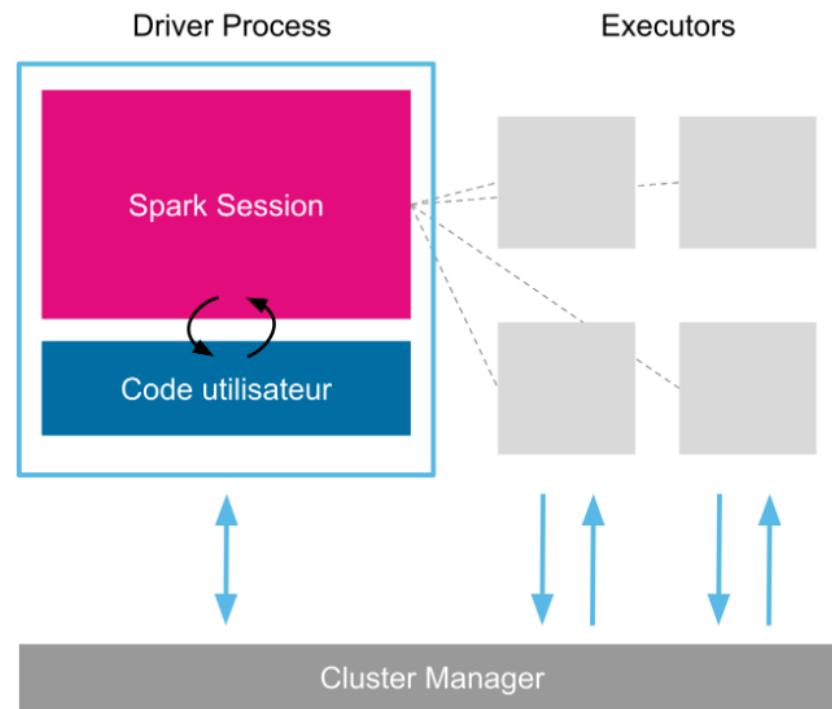
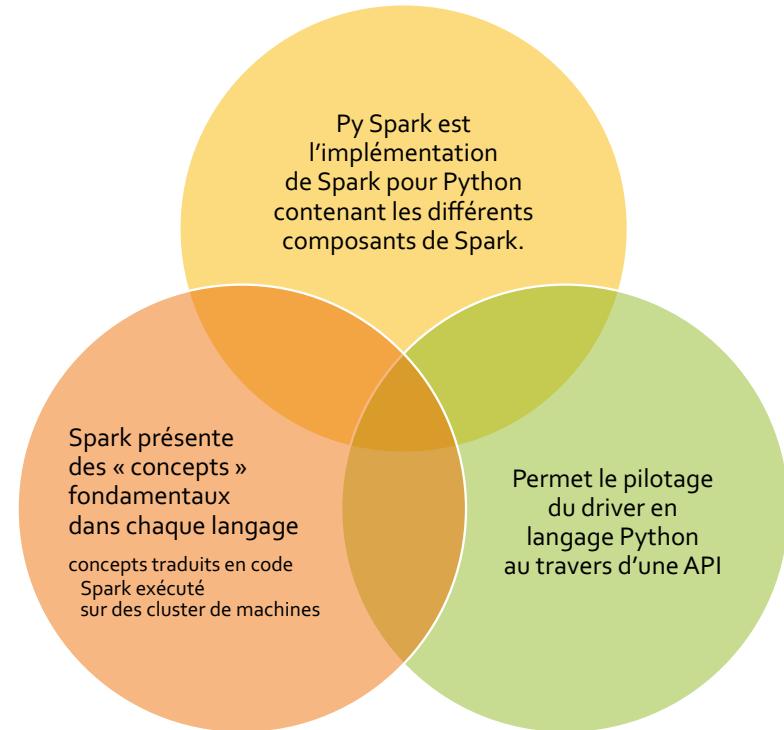


Schéma de l'architecture logique d'exécution



Spark

Principe de fonctionnement

Pandas UDF (User Defined Function)

Fonction définie par l'utilisateur

Permet des opérations vectorisées

Apache Arrow pour transférer des données

Pandas pour travailler avec les données

Performances jusqu'à 100 fois supérieures aux UDF Python

Architecture big data



AWS IAM



Amazon S3



Amazon EC2



AWS –Qu'est que c'est ?

- Service de cloud computing à la demande ,les plus populaires sont Elastic Compute Cloud (EC2) et Simple Storage Service(S3) et IAM (Identity and Access Management)
- **EC2** : est une Infrastructure as a Service (IaaS), car Amazon fournit un accès à une partie de ses serveurs mais c'est à l'utilisateur de gérer le système opératif, run time et data
- **S3**: permet de stocker des données de manière infinie. La tarification s'adapte automatiquement à l'utilisation.
- **IAM** : Permet de gérer les services AWS accessibles à un compte IAM (utilisateur).
 - La possibilité d'avoir du “pay as you go” est l'une des forces du cloud. Vous payez à la hauteur de votre consommation réelle.

La mise en œuvre d'une architecture Big Data

- 1) Crée un compte sur AWS
- 2) Lancer un instance sur EC2

The screenshot shows the AWS EC2 Home page for the Europe (Paris) region. The left sidebar includes links for Tableau de bord EC2, Instances, Images, Elastic Block Store, Réseau et sécurité, and New EC2 Experience. The main content area displays resource counts: 0 Instances (en cours d'exécution), 0 Adresses IP Elastic, 0 Équilibreurs de charge, 0 Groupes de placement, 5 Groupes de sécurité, 0 Hôtes dédiés, 1 Instances, 0 Instantanés, 1 Paires de clés, and 1 Volumes. A callout box highlights the "Lancer une instance" button. The right sidebar contains sections for Attributs du compte (Platformes prises en charge, VPC par défaut, Paramètres, Chiffrement EBS, Zones, Console de série EC2, Spécification des crédits par dé, Expériences de console), Informations supplémentaires (Guide de démarrage, Documentation, Toutes les ressources EC2, Forums, Tarification, Nous contacter), and Rubriques d'aide (Pourquoi ne puis-je pas me connecter à mon instance Amazon).

Il faut faire attention pour région

(EC2) Elastic Compute Cloud

The screenshot shows the AWS EC2 Instances page. At the top left, there's a sidebar with options like 'New EC2 Experience', 'EC2 Dashboard', 'EC2 Global View', 'Events', and 'Tags'. The main area displays a table titled 'Instances (1) Info' with one row. The row contains the instance ID 'i-0db8bc74145ab46c0', its state 'Stopped', type 't2.micro', and other details. A 'Launch instances' button is at the top right of the table.

Catalogue des AMI

Une AMI est un modèle qui contient la configuration logicielle (système d'exploitation, serveur d'applications et applications) requise pour lancer votre instance. Vous pouvez sélectionner une AMI fournie par AWS, notre communauté d'utilisateurs ou AWS Marketplace, ou bien l'une de vos propres AMI.

AMI

AMI sélectionnée: (ami-08a32bf9aa25ef9da)

[Créer un modèle avec une AMI](#)

[Lancer une instance avec une AMI](#)

tensorflow

[AMI à démarrage rapide \(3\)](#)

AMI couramment utilisées

[Mes AMI \(0\)](#)

Crées par moi-même

[AMI d'AWS Marketplace \(82\)](#)

AWS et AMI tierces de confiance

[AMI de la communauté \(477\)](#)

Publiées par n'importe qui

Affiner les résultats

[Effacer tous les filtres](#)

- Offre gratuite Informations uniquement
- Catégorie de système d'exploitation
 - Tous les Linux/Unix
 - Tous les Windows
- Architecture

tensorflow (3 filtré, 3 non filtré)

ubuntu

Ubuntu

Fournisseur vérifié

Deep Learning AMI GPU TensorFlow 2.10.0 (Ubuntu 20.04) 20220927

ami-08a32bf9aa25ef9da (64 bits (x86))

Built with AWS optimized TensorFlow, NVIDIA CUDA, cuDNN, NCCL, GPU Driver, Docker, NVIDIA-Docker and EFA support. For a fully managed experience, check: <https://aws.amazon.com/sagemaker>

Plateforme: ubuntu

Type d'appareil racine: ebs

Virtualisation: hvm

ENA activée: Oui

[Sélectionner](#)

64 bits (x86)

ubuntu

Ubuntu

Fournisseur vérifié

Deep Learning AMI GPU TensorFlow 2.9.2 (Ubuntu 20.04) 20221005

ami-0226deb962c1d9d7e (64 bits (x86))

Built with AWS optimized TensorFlow, NVIDIA CUDA, cuDNN, NCCL, GPU Driver, Docker, NVIDIA-Docker and EFA support. For a fully managed experience, check: <https://aws.amazon.com/sagemaker>

Plateforme: ubuntu

Type d'appareil racine: ebs

Virtualisation: hvm

ENA activée: Oui

[Sélectionner](#)

64 bits (x86)

aws

Amazon Linux

Deep Learning AMI GPU TensorFlow 2.10.0 (Amazon Linux 2) 20220927

ami-0251711462e893767 (64 bits (x86))

Built with AWS optimized TensorFlow, NVIDIA CUDA, cuDNN, NCCL, GPU Driver, Docker, NVIDIA-Docker and EFA support. For a fully managed experience, check: <https://aws.amazon.com/sagemaker>

Plateforme: amazonlinux

Type d'appareil racine: ebs

Virtualisation: hvm

ENA activée: Oui

[Sélectionner](#)

64 bits (x86)

[Voir tous les résultats](#)

La mise en œuvre d'une architecture Big Data

AWS Services Rechercher [Alt+S] Paris Gulsum

New EC2 Experience Tell us what you think X

Tableau de bord EC2

Vue globale EC2

Événements

Balises

Limites

Instances

- Instances
- Types d'instances
- Modèles de lancement
- Demandes Spot
- Savings Plans
- Instances réservées
- Hôtes dédiés
- Réservations de capacité

Images

- AMI
- Catalogue des AMI

Elastic Block Store

- Volumes
- Instantanés
- Gestionnaire de cycle de vie

Réseau et sécurité

- Groupes de sécurité
- Adresses IP élastiques

EC2 Instances i-0db8bc74145ab46c0

Résumé de l'instance pour i-0db8bc74145ab46c0 Informations Mis à jour il y a less than a minute

ID d'instance	Adresse IPv4 publique	Adresses IPv4 privées
i-0db8bc74145ab46c0	-	172.31.45.140
Adresse IPv6	État de l'instance	DNS IPv4 public
-	Arrêté(e)	-
Type de nom d'hôte	Nom DNS de l'IP privé (IPv4 uniquement)	Adresses IP élastiques
Nom de l'adresse IP: ip-172-31-45-140.eu-west-3.compute.internal	ip-172-31-45-140.eu-west-3.compute.internal	-
Réponse à un nom DNS de ressource privée	Type d'instance	Recherche d'AWS Compute Optimizer
IPv4 (A)	t2.micro	Inscrivez-vous à AWS Compute Optimizer pour obtenir des recommandations.
Adresse IP attribuée automatiquement	ID de VPC	En savoir plus
-	vpc-0a787ae454b364e59	
Rôle IAM	ID de sous-réseau	Nom du groupe Auto Scaling
-	subnet-0028f001e7abcfed6	-

Détails Sécurité Mise en réseau Stockage Vérifications de statut Surveillance Balises

▼ Détails de l'instance Informations

Plateforme	ID AMI	Surveillance
Ubuntu (déduit)	ami-03c476a1ca8e3ebdc	désactivé
Informations sur la plateforme	Nom de l'AMI	Protection de la résiliation
Linux/UNIX	ubuntu/images/hvm-ssd/ubuntu-jammy-22.04-amd64-server-20221206	Désactivé
Protection contre l'arrêt	Heure de lancement	Emplacement de l'AMI
Désactivé	Sun Jan 08 2023 11:26:37 GMT+0100 (heure normale d'Europe centrale) (2 days)	amazon/ubuntu/images/hvm-ssd/ubuntu-jammy-22.04-amd64-server-20221206
Récupération automatique de l'instance	Cycle de vie	Comportement Arrêt - Mise en veille prolongée
Par défaut	normal	désactivé

La mise en œuvre d'une architecture Big Data

Création de clusters dans EMR

The screenshot shows the AWS EMR console interface. At the top, there are two tabs: "EMR console" (selected) and "EMR - AWS Console". The browser address bar displays the URL: "eu-west-3.console.aws.amazon.com/elasticmapreduce/home?optOut=true®ion=eu-west-3". The AWS navigation bar includes links for various services like Lambda, CloudWatch, and S3, along with a search bar and user information.

The main content area is titled "Amazon EMR" and features a sidebar with options: "Créer un cluster" (highlighted in blue), "Afficher les détails", "Cloner", and "Résilier". Below this, there's a filter section with "Filtre : Tous les clusters" and a count of "43 clusters (tous chargés)". A table header is visible, listing columns: "Nom", "ID", "Statut", "Heure de création (UTC+1)", "Temps écoulé", and "Heures d'instances normalisées".

A prominent message banner at the top states: "Amazon EMR has launched a new console experience. Learn more or switch to the new console".

EMR

EMR console x EMR - AWS Console x +

eu-west-3.console.aws.amazon.com/elasticmapreduce/home?optOut=true®ion=eu-west-3#create-cluster:

Services Rechercher [Alt+S]

Amazon EMR has launched a new console experience. Learn more or switch to the new console.

Créer un cluster - Options avancées Accéder aux options rapides

Étape 1 : Logiciels et étapes

Étape 2 : Matériel

Étape 3 : Paramètres de cluster généraux

Étape 4 : Sécurité

Configuration des logiciels

Libérer [emr-5.7.0]

Logiciel	Version
Hadoop 3.2.1	Zeppelin 0.10.0
JupyterHub 1.4.1	Tez 0.9.2
Ganglia 3.7.2	HBase 2.4.4
Hive 3.1.3	Presto 0.272
JupyterEnterpriseGateway 2.1.0	MXNet 1.8.0
Hue 4.10.0	Phoenix 5.1.2
Oozie 5.2.1	Spark 3.2.1
TensorFlow 2.4.1	Livy 0.7.1
	Flink 1.14.2
	Pig 0.17.0
	ZooKeeper 3.5.7
	Sqoop 1.4.7
	Trino 378
	HCatalog 3.1.3

Plusieurs nœuds principaux (facultatif)

Utilisez plusieurs nœuds principaux pour améliorer la disponibilité du cluster. En savoir plus

Paramètres de AWS Glue Data Catalog (facultatif)

Utiliser pour les métadonnées de table Hive

Utiliser pour les métadonnées de table Spark

Modifier les paramètres du logiciel

Entrer la configuration Charger JSON à partir de S3

```
classification=config-file-name,properties=[myKey1=myValue1,myKey2=myValue2]
```

Ajouter des étapes (facultatif)

Une étape est une unité de travail que vous soumettez au cluster. Par exemple, une étape peut contenir une ou plusieurs tâches Hadoop ou Spark. Vous pouvez également soumettre des étapes supplémentaires à un cluster après le début de son exécution. En savoir plus

Concurrency: Run multiple steps at the same time to improve cluster utilization

After last step completes: Clusters enters waiting state

EMR

Amazon EMR has launched a new console experience. [Learn more](#) or switch to the new console.

Étape 1 : Logiciels et étapes

Étape 2 : Matériel

Étape 3 : Paramètres de cluster généraux

Étape 4 : Sécurité

Options générales

Nom du cluster p8

Journalisation

Dossier S3 `s3://aws-logs-948883924704-eu-west-3/elastictmapreduc...`

Chiffrement de journaux

Protection de la résiliation

Balises

Clé	Valeur (facultative)
Ajouter une clé pour créer une balise	

Options supplémentaires

Vue cohérente EMRFS

Options du système d'exploitation

Version d'Amazon Linux 2.0.20221210.1

ID d'AMI personnalisée Entrer une ID AMI

Mettre à jour tous les packages au redémarrage (recommandé)

▶ Actions d'amorçage

Annuler Précédent Suivant

Amazon EMR has launched a new console experience. [Learn more](#) or switch to the new console.

Créer un cluster - Options avancées [Accéder aux options rapides](#)

Étape 1 : Logiciels et étapes

Étape 2 : Matériel

Étape 3 : Paramètres de cluster généraux

Étape 4 : Sécurité

Options de sécurité

Paire de clés EC2 gism

Cluster visible pour tous les utilisateurs IAM du compte

Autorisations

Par défaut Personnalisé

Utilisez les rôles IAM par défaut. Si des rôles sont absents, ils seront créés automatiquement pour vous avec des stratégies gérées pour les mises à jour automatiques de stratégies.

Rôle EMR [EMR_DefaultRole](#) Use EMR_DefaultRole_V2

Profil d'instance EC2 [EMR_E2_DefaultRole](#)

Rôle Auto Scaling [EMR_AutoScaling_DefaultRole](#)

▶ Configuration de la sécurité

▶ Groupes de sécurité EC2

Annuler Précédent Crée un cluster

EMR

us-east-1.console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#cluster-details:j-RQMGOWLHW5RO

Amazon EMR has launched a new console experience. [Learn more](#) or switch to the new console.

Clone Terminate AWS CLI export

Cluster: p8 Waiting Cluster ready to run steps.

Summary Application user interfaces Monitoring Hardware Configurations Events Steps Bootstrap actions

Add task instance group

Instance groups

Filter: Filter instance groups ... 1 instance group (all loaded) C

ID	Status	Node type & name	Instance type	Instance count	Purchasing c
ig-2UI37N6P406N	Running	MASTER Master - 1	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB	1 Instances	On-demand (

Cluster Scaling Policy Edit

No scaling enabled

Auto-termination Edit

Select a time to have the cluster terminate after the cluster becomes idle. Choose a minimum of 1 minute or a max of 24 hours. [Learn more](#)

Auto-termination: Terminate if idle for 1 hour

SSH

The screenshot shows a browser window with two main components:

- PuTTY Configuration Dialog:** A modal window titled "PuTTY Configuration". It has a sidebar with categories like Session, Logging, Terminal, Window, Appearance, Connection, and Features. The "Session" category is expanded, showing fields for "Host Name (or IP address)" set to "100.eu-west-3.compute.amazonaws.com" and "Port" set to "22". The "Connection type" radio button is selected for "SSH". Below these are "Load", "Save", and "Delete" buttons. At the bottom are "Open" and "Cancel" buttons.
- Modal Window: configurer la connexion Web**: This window provides instructions for setting up a tunnel SSH to the master node of an Amazon EMR cluster. It includes sections for "Étape 1 : Ouvrez un tunnel SSH vers le nœud maître Amazon EMR" and "Étape 2 : Configurer un outil de gestion de proxy". The "Étape 1" section is detailed below.

Étape 1 : Ouvrez un tunnel SSH vers le nœud maître Amazon EMR - [En savoir plus](#)

Windows Mac/Linux

1. Téléchargez PuTTY.exe sur votre ordinateur à partir de : <http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html>
2. Lancez PUTTY.
3. Dans la liste Category, cliquez sur Session
4. Dans le champ Host Name, tapez **hadoop@ec2-35-180-42-100.eu-west-3.compute.amazonaws.com**
5. Dans la liste Category, développez Connection > SSH > Auth
6. Pour le fichier de clés privées utilisé pour l'authentification, cliquez sur Browse et sélectionnez le fichier de clés privées (**gism.ppk**) utilisé pour lancer le cluster.
7. Dans la liste Category, développez Connection > SSH, puis cliquez sur Tunnels.
8. Dans le champ Source port, tapez **8157** (port local inutilisé choisi de façon aléatoire).
9. Sélectionnez les options Dynamic et Auto.
10. Laissez le champ Destination vide, puis cliquez sur Add.
11. Cliquez sur Open.
12. Cliquez sur Yes pour ignorer l'alerte de sécurité.

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

Chrome Firefox

1. Go to <https://chrome.google.com/webstore/category/extensions>, search for **Proxy SwitchyOmega**, and add it to Chrome.
2. Choose **New profile** and enter **emr-socks-proxy** as the profile name.
3. Choose **PAC profile** and then **Create Proxy Auto-Configuration (PAC)** files help you define an allow list for browser requests that should be forwarded to a web proxy server.
4. In the **PAC Script** field, replace the contents with the following script that defines which URLs should be forwarded through your web proxy server. If you specified a different port number when you set up your SSH tunnel, replace 8157 with your port number.

```
function FindProxyForURL(url, host) {  
    if (shExpMatch(url, "*ec2*.amazonaws.com*")) return 'SOCKS5 localhost:8157';  
    if (shExpMatch(url, "*ec2*.compute*")) return 'SOCKS5 localhost:8157';  
}
```

Fermer

JupyterHub

The screenshot shows the AWS EMR console interface. On the left, there is a sidebar with navigation links for Amazon EMR, EMR Studio, EMR Serverless (New), EMR on EC2 (selected), Clusters, Notebooks, Git repositories, Security configurations, Block public access, VPC subnets, and Events. Below that is a section for EMR on EKS with Virtual clusters. At the bottom of the sidebar are Help and What's new links.

The main content area has a header with tabs: Summary (selected), Application user interfaces, Monitoring, Hardware, Configurations, Events, Steps, and Bootstrap actions. A message box at the top says "Amazon EMR has launched a new console experience. [Learn more](#) or [switch to the new console](#)".

Persistent application user interfaces

Applications installed on the Amazon EMR cluster publish user interfaces (UI) as web sites to monitor cluster activity. Persistent UI logs are available for 30 days after an application ends. Persistent UI don't require SSH tunneling. They are hosted off the cluster.

Application user interface

- YARN timeline server
- Spark history server

On-cluster application user interfaces

On-cluster UI are available only while clusters are running. Because they are hosted on the master node, on-cluster UI require a connection via SSH tunneling. Set up SSH tunneling before accessing these application UI. [Learn more](#)

Application	User interface URL	Status
HDFS Name Node	http://ec2-18-209-227-151.compute-1.amazonaws.com:9870/	SSH tunnel not enabled
JupyterHub	https://ec2-18-209-227-151.compute-1.amazonaws.com:9443/	SSH tunnel not enabled
Spark History Server	http://ec2-18-209-227-151.compute-1.amazonaws.com:18080/	SSH tunnel not enabled
Resource Manager	http://ec2-18-209-227-151.compute-1.amazonaws.com:8088/	SSH tunnel not enabled

The following table lists web interfaces you can view on the task nodes:

Application	User interface URL
HDFS Data Node	http://ec2-000-000-000-000.compute-1.amazonaws.com:9864/

JUPYTERHUB

The screenshot shows a JupyterHub interface with two main windows. The top window is a file browser titled "jupyterhub" showing files in a directory. The bottom window is a notebook titled "Kapanoglu_Gulsum_Notebook_13122022.ipynb" containing code and output.

File Browser (Top Window):

- Files: Untitled.ipynb, jupyterhub-proxy.pid, jupyterhub.sqlite, jupyterhub_cookie_secret
- Running: 28 dakika önce (28 minutes ago)
- bir saat önce (1 hour ago)
- birkaç saniye sonra (in a few seconds)
- bir saat önce (1 hour ago)

Notebook Session (Bottom Window):

In [7]:

```
PATH = 's3://oc-gulsump8'
PATH_Data = PATH + '/Test'
PATH_Result = PATH + '/Results'
print('PATH: ' + PATH)
print('PATH_Data: ' + PATH_Data)
print('PATH_Result: ' + PATH_Result)
```

PATH: s3://oc-gulsump8
PATH_Data: s3://oc-gulsump8/Test
PATH_Result: s3://oc-gulsump8/Results

In [8]:

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

In [9]:

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

path	modificationTime	length	content
s3://oc-gulsump8/...	2022-12-31 12:48:10	7353	[FF D8 FF E0 00 1...]
s3://oc-gulsump8/...	2022-12-31 12:48:13	7350	[FF D8 FF E0 00 1...]
s3://oc-gulsump8/...	2022-12-31 12:48:12	7349	[FF D8 FF E0 00 1...]
s3://oc-gulsump8/...	2022-12-31 12:48:11	7348	[FF D8 FF E0 00 1...]
s3://oc-gulsump8/...	2022-12-31 12:15:25	7328	[FF D8 FF E0 00 1...]

only showing top 5 rows

In [10]:

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

root

8°C Nuageux 01:00 09/01/2023

EMR-A LA FIN

The screenshot shows the AWS EMR console interface. On the left, there is a sidebar with the following navigation options:

- Amazon EMR
- EMR Studio
- EMR Serverless New
- MR on EC2
- Clusters
- Notebooks
- Git repositories
- Security configurations
- Block public access
- /PC subnets
- Events
- MR on EKS
- /virtual clusters

Below the sidebar, there are two sections: "Help" and "What's new".

The main content area displays information about a cluster named "p8" which is currently "Waiting". It includes tabs for "Summary", "Application user interfaces", "Monitoring", "Hardware", "Configurations", "Events", "Steps", and "Bootstrap actions".

Summary tab details:

- ID: j-RQMGOWLHW5RO
- Creation date: 2023-01-08 16:03 (UTC+1)
- Elapsed time: 8 hours, 37 minutes
- After last step completes: Cluster waits
- Termination protection: On Change
- Tags: -- [View All / Edit](#)
- Master public DNS: ec2-18-209-227-151.compute-1.amazonaws.com [Connect to the Master Node Using SSH](#)

Application user interfaces tab details:

- Persistent user interfaces: Spark history server, YARN timeline server
- On-cluster user interfaces: Not Enabled [Enable an SSH Connection](#)

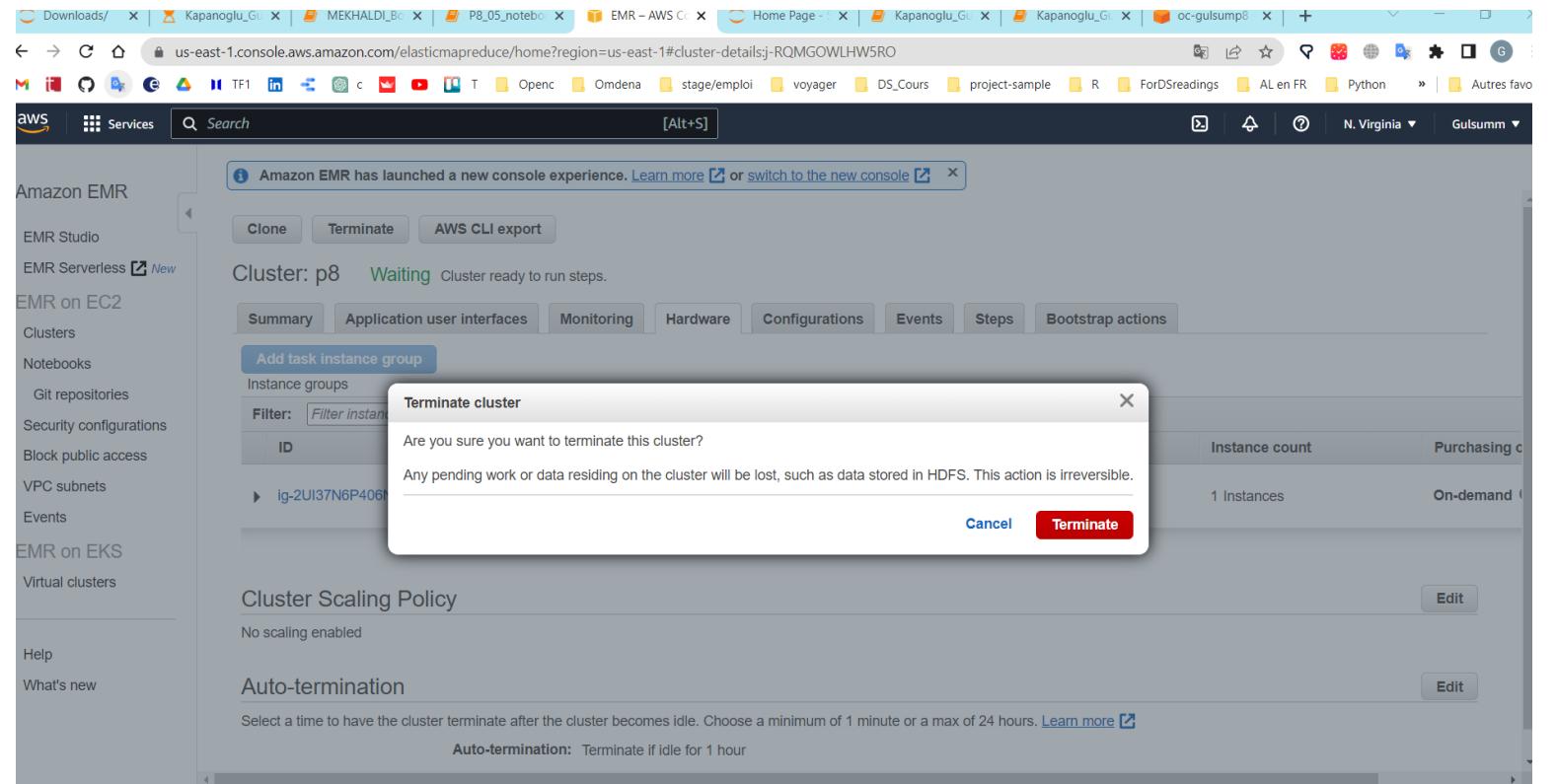
Configuration details tab details:

- Release label: emr-6.7.0
- Hadoop distribution: Amazon 3.2.1
- Applications: JupyterHub 1.4.1, TensorFlow 2.4.1, Spark 3.2.1
- Log URI: s3://aws-logs-948883924704-us-east-1/elasticmapreduce/
- EMRFS consistent view: Disabled
- Custom AMI ID: --
- Amazon Linux Release: 2.0.20221210.1 [Learn more](#)

Network and hardware tab details:

- Availability zone: us-east-1b
- Subnet ID: subnet-002eb4c6b72035a9d
- Master: Running 1 m5.xlarge
- Core: --
- Task: --
- Cluster scaling: Not enabled
- Auto-termination: Terminate if idle for 1 hour

EMR- RESILIATION



S3 (Simple Storage Service)

console.aws.amazon.com/s3/buckets?region=eu-west-3®ion=eu-west-3

Services Search [Alt+S] Global Gulsumm

Amazon S3

Buckets

- Access Points
- Object Lambda Access Points
- Multi-Region Access Points
- Batch Operations
- Access analyzer for S3

Block Public Access settings for this account

Storage Lens

- Dashboards
- AWS Organizations settings

Amazon S3 > Buckets

Account snapshot

Storage lens provides visibility into storage usage and activity trends. Learn more

View Storage Lens dashboard

Buckets (2) Info

Buckets are containers for data stored in S3. Learn more

Name	AWS Region	Access	Creation date
aws-logs-948883924704-us-east-1	US East (N. Virginia) us-east-1	Objects can be public	January 8, 2023, 16:03:24 (UTC+01:00)
oc-gulsump8	EU (Paris) eu-west-3	Objects can be public	December 30, 2022, 14:43:12 (UTC+01:00)

console.aws.amazon.com/s3/buckets/oc-gulsump8?region=eu-west-3&tab=objects

Services Search [Alt+S] Global Gulsumm

oc-gulsump8

Objects Properties Permissions Metrics Management Access Points

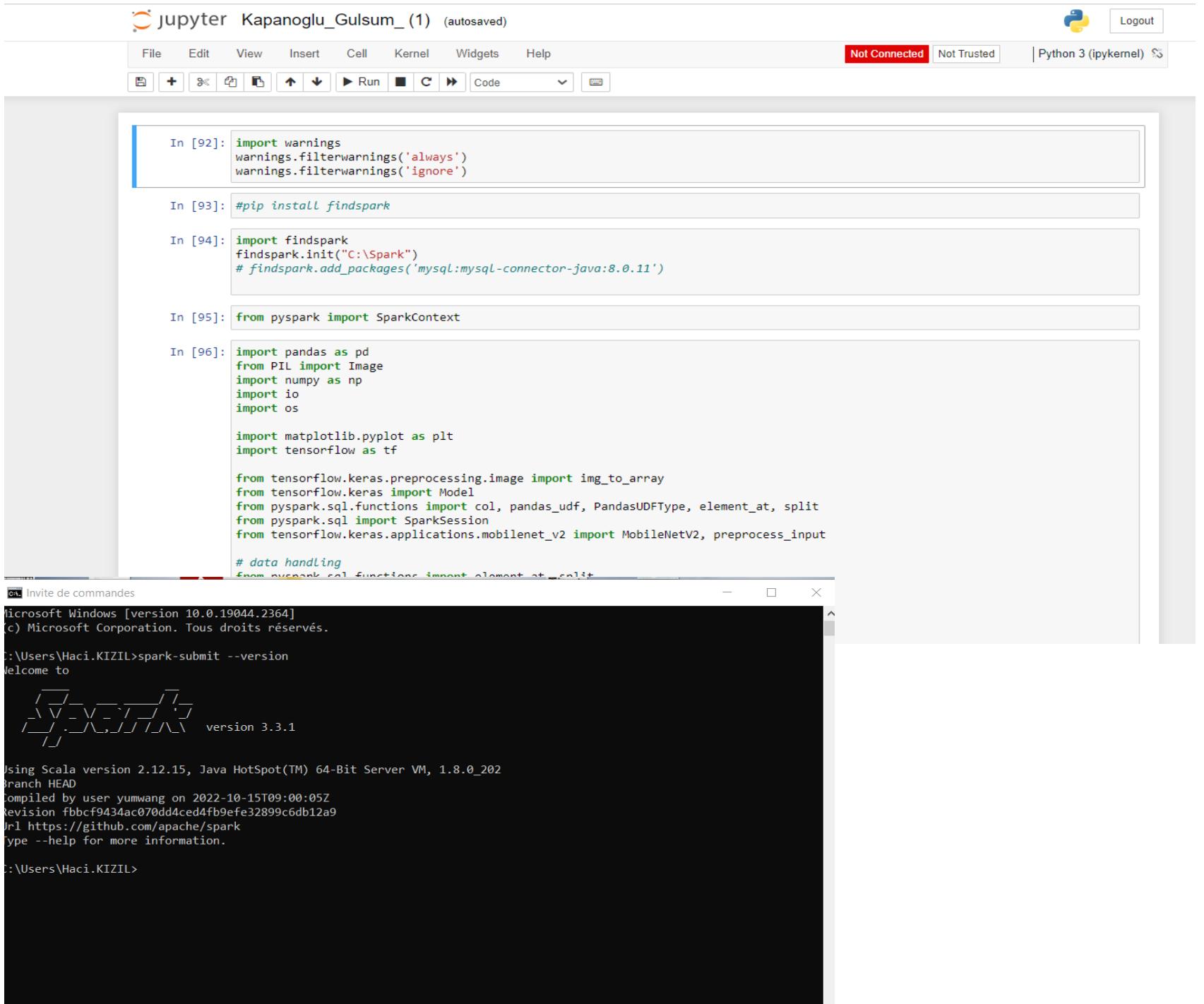
Objects (14)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

Name	Type	Last modified	Size	Storage class
Csv_resultsp8.csv	csv	January 9, 2023, 00:51:01 (UTC+01:00)	71.5 MB	Standard
j-12U52JW829A6/	Folder	-	-	-
j-BBBIQSAHNWDM6/	Folder	-	-	-
j-3BZMNWCIEIO28/	Folder	-	-	-
j-5FXC2PGSY1G06/	Folder	-	-	-
j-3PGKMT9CGY8PZ/	Folder	-	-	-
j-B53BO8FGMZKL/	Folder	-	-	-
j-Fi1DP09INWJJ/	Folder	-	-	-
j-XLKBFKEN4H5A/	Folder	-	-	-
jupyter/	Folder	-	-	-
Kapanoglu_Gulsump8.ipynb	ipynb	January 2, 2023, 15:38:38 (UTC+01:00)	60.4 KB	Standard
P8_06_bootstrap-emr.sh	sh	January 7, 2023, 18:32:07 (UTC+01:00)	270.0 B	Standard
Results/	Folder	-	-	-
Test/	Folder	-	-	-

Application de la solution sur en local

Spark en local



Model

Préparation du modèle

EXTRACTION DES FEATURES

```
In [109]: model = MobileNetV2(weights='imagenet',
                             include_top=True,
                             input_shape=(224, 224, 3))

In [110]: new_model = Model(inputs=model.input,
                         outputs=model.layers[-2].output)

In [111]: new_model.summary()
```

Model: "model_2"

Layer (type)	Output Shape	Param #	Connected to
input_3 (InputLayer)	[None, 224, 224, 3 0)]	0	[]
Conv1 (Conv2D)	(None, 112, 112, 32 864)	864	['input_3[0][0]']
bn_Conv1 (BatchNormalization)	(None, 112, 112, 32 128)	128	['Conv1[0][0]']
Conv1_relu (ReLU)	(None, 112, 112, 32 0)	0	['bn_Conv1[0][0]']
expanded_conv_depthwise (Depth	(None, 112, 112, 32 288)	288	['Conv1_relu[0][0]']

- J'ai utilisé un modèle entraîné **MobileNetV2** pour l'adapter à notre problématique.
- J'ai fourni au modèle les images, et récupérer l'avant dernière couche du modèle. la dernière couche de modèle est une couche softmax qui permet la classification des images ce que nous ne souhaitons pas dans ce projet. L'avant dernière couche correspond à un **vecteur réduit** de dimension (1,1,1280).

Cela permettra de réaliser une première version du moteur pour la classification des images des fruits.

Etapes de la chaîne de traitement

```
In [147]: pca_matrix.show(20)
```

path	label	cnn_vectors	features_pca
file:/C:/Users/Ha...	Apple Golden 1 [0.0,0.0479974746...	[4.98089584321073...	
file:/C:/Users/Ha...	Apple Golden 1 [0.00111107924021...	[5.50980016257361...	
file:/C:/Users/Ha...	Apple Golden 1 [0.03722351416945...	[6.39840192490234...	
file:/C:/Users/Ha...	Apple Golden 1 [0.0,0.0,0.684601...	[5.43004893616666...	
file:/C:/Users/Ha...	Apple Golden 1 [0.0,0.0,0.366576...	[3.98251790524833...	
file:/C:/Users/Ha...	Apple Golden 1 [0.0,0.0274292603...	[7.58072670909864...	
file:/C:/Users/Ha...	Apple Golden 1 [0.0,0.0785900354...	[4.70645256629018...	
file:/C:/Users/Ha...	Apple Golden 1 [0.0,0.0660749003...	[4.77716108318939...	
file:/C:/Users/Ha...	Apple Golden 1 [0.0,0.0,0.527887...	[5.98091586841501...	
file:/C:/Users/Ha...	Apple Golden 1 [0.30108457803726...	[6.65606951884041...	
file:/C:/Users/Ha...	Apple Golden 1 [0.04646725580096...	[6.20499713395209...	
file:/C:/Users/Ha...	Apple Golden 1 [0.0,0.0,0.183206...	[6.58447506776186...	
file:/C:/Users/Ha...	Apple Golden 1 [0.05055416002869...	[5.90109476390204...	
file:/C:/Users/Ha...	Apple Golden 1 [0.00387673475779...	[5.88031374705795...	
file:/C:/Users/Ha...	Apple Crimson Snow [0.0,0.0,0.0,0.0,...	[-2.8030696421467...	
file:/C:/Users/Ha...	Apple Golden 1 [0.0,0.0241741072...	[6.98971067519546...	
file:/C:/Users/Ha...	Apple Golden 1 [0.0,0.6196982860...	[5.24937609188189...	
file:/C:/Users/Ha...	Apple Golden 1 [0.0,0.0567283257...	[6.49499667546596...	
file:/C:/Users/Ha...	Apple Braeburn [0.01309019234031...	[-12.442065620066...	
file:/C:/Users/Ha...	Apple Golden 1 [0.09267588704824...	[7.10548674052478...	

```
only showing top 20 rows
```

```
In [152]: final_df.show(10, True)
```

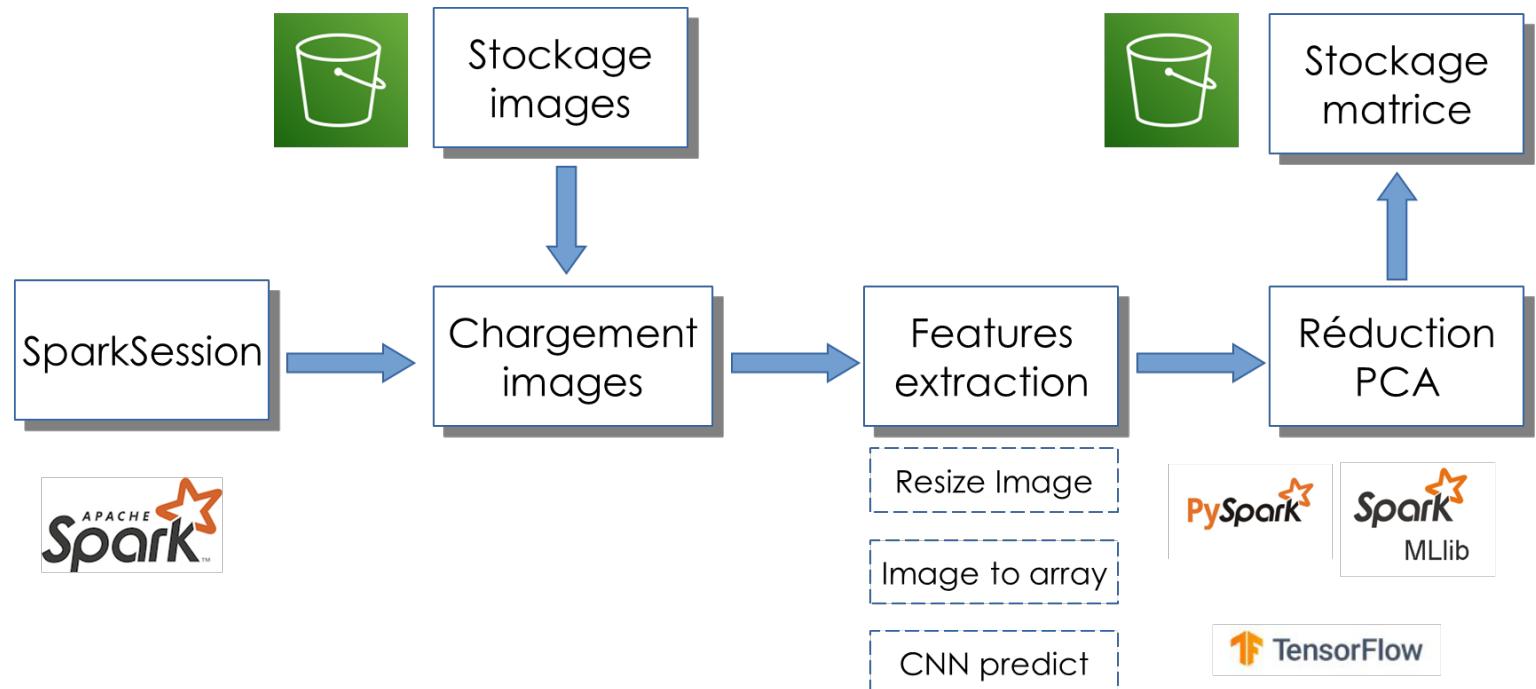
path	label	cnn_vectors	features_pca	features
file:/C:/Users/Ha...	Apple Golden 1 [0.0,0.0479974746...	[4.98089584321073...	[4.980896, 5.4048...	
file:/C:/Users/Ha...	Apple Golden 1 [0.00111107924021...	[5.50980016257361...	[5.5098, 4.626695...	
file:/C:/Users/Ha...	Apple Golden 1 [0.03722351416945...	[6.39840192490234...	[6.3984017, 4.591...	
file:/C:/Users/Ha...	Apple Golden 1 [0.0,0.0,0.684601...	[5.43004893616666...	[5.430049, 4.8585...	
file:/C:/Users/Ha...	Apple Golden 1 [0.0,0.0,0.366576...	[3.98251790524833...	[3.982518, 3.7568...	
file:/C:/Users/Ha...	Apple Golden 1 [0.0,0.0274292603...	[7.58072670909864...	[7.5807266, 2.013...	
file:/C:/Users/Ha...	Apple Golden 1 [0.0,0.0785900354...	[4.70645256629018...	[4.7064524, 6.249...	
file:/C:/Users/Ha...	Apple Golden 1 [0.0,0.0660749003...	[4.77716108318939...	[4.777161, 6.3284...	
file:/C:/Users/Ha...	Apple Golden 1 [0.0,0.0,0.527887...	[5.98091586841501...	[5.980916, 4.6318...	
file:/C:/Users/Ha...	Apple Golden 1 [0.30108457803726...	[6.65606951884041...	[6.6560698, 4.061...	

```
only showing top 10 rows
```

Application de la solution sur le Cloud (AWS)

Etapes de la chaîne de traitement

Schéma du traitement des images dans notebook Jupyter



Spark Job

Non sécurisé | ip-172-31-14-107.ec2.internal:20888/proxy/application_1673190521502_0001/

User: livy
Total Uptime: 36 min
Scheduling Mode: FIFO
Active Jobs: 1
Completed Jobs: 3

Event Timeline

Active Jobs (1)

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
3 (19)	Job group for statement 19 showString at NativeMethodAccessorImpl.java:0	2023/01/08 15:28:18 (kill)	29 min	0/1	225/709 (3 running)

Completed Jobs (3)

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
2 (10)	Job group for statement 10 showString at NativeMethodAccessorImpl.java:0	2023/01/08 15:26:07	0.2 s	1/1	1/1
1 (9)	Job group for statement 9 showString at NativeMethodAccessorImpl.java:0	2023/01/08 15:25:55	4 s	1/1	1/1
0 (8)	Listing leaf files and directories for 131 paths: s3://oc-gulsump8/Test/Apple Braeburn, ... load at NativeMethodAccessorImpl.java:0	2023/01/08 15:25:21	23 s	1/1	131/131

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

Etapes de la chaîne de traitement

```
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),...  
[26]: final_df = pca_matrix.withColumn('features', vector_to_array_udf('features_pca'))  
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),...  
[27]: final_df.show(10, True)
```

```
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),...  
+-----+-----+-----+-----+-----+  
|       path|    label|  cnn_vectors| features_pca|      features|  
+-----+-----+-----+-----+-----+  
| s3://oc-gulsump8/...| Watermelon|[0.04279927909374...|[-2.0967874076741...|[-2.0967875, 5.78...|  
| s3://oc-gulsump8/...| Watermelon|[0.01487588509917...|[-1.8770516056425...|[-1.8770516, 3.28...|  
| s3://oc-gulsump8/...|Pineapple Mini|[0.0,4.7234764099...|[-5.9249884518216...|[-5.9249883, 3.36...|  
| s3://oc-gulsump8/...|Pineapple Mini|[0.0,4.7648377418...|[-4.8288816936532...|[-4.8288817, 2.31...|  
| s3://oc-gulsump8/...| Watermelon|[0.11024109274148...|[-2.6555744153131...|[-2.6555743, 2.17...|  
| s3://oc-gulsump8/...| Watermelon|[0.11004009842872...|[-2.8012731267059...|[-2.801273, 2.576...|  
| s3://oc-gulsump8/...|Pineapple Mini|[0.0,4.7182641029...|[-6.2835548731391...|[-6.283555, 4.230...|  
| s3://oc-gulsump8/...| Watermelon|[0.05197056382894...|[-4.3869267867072...|[-4.3869267, 5.74...|  
| s3://oc-gulsump8/...| Watermelon|[0.01286508701741...|[-1.9200305572885...|[-1.9200306, 3.89...|  
| s3://oc-gulsump8/...| Cauliflower|[0.0,0.4670273065...|[-4.8778844295835...|[-4.8778844, 2.46...|  
+-----+-----+-----+-----+  
only showing top 10 rows
```

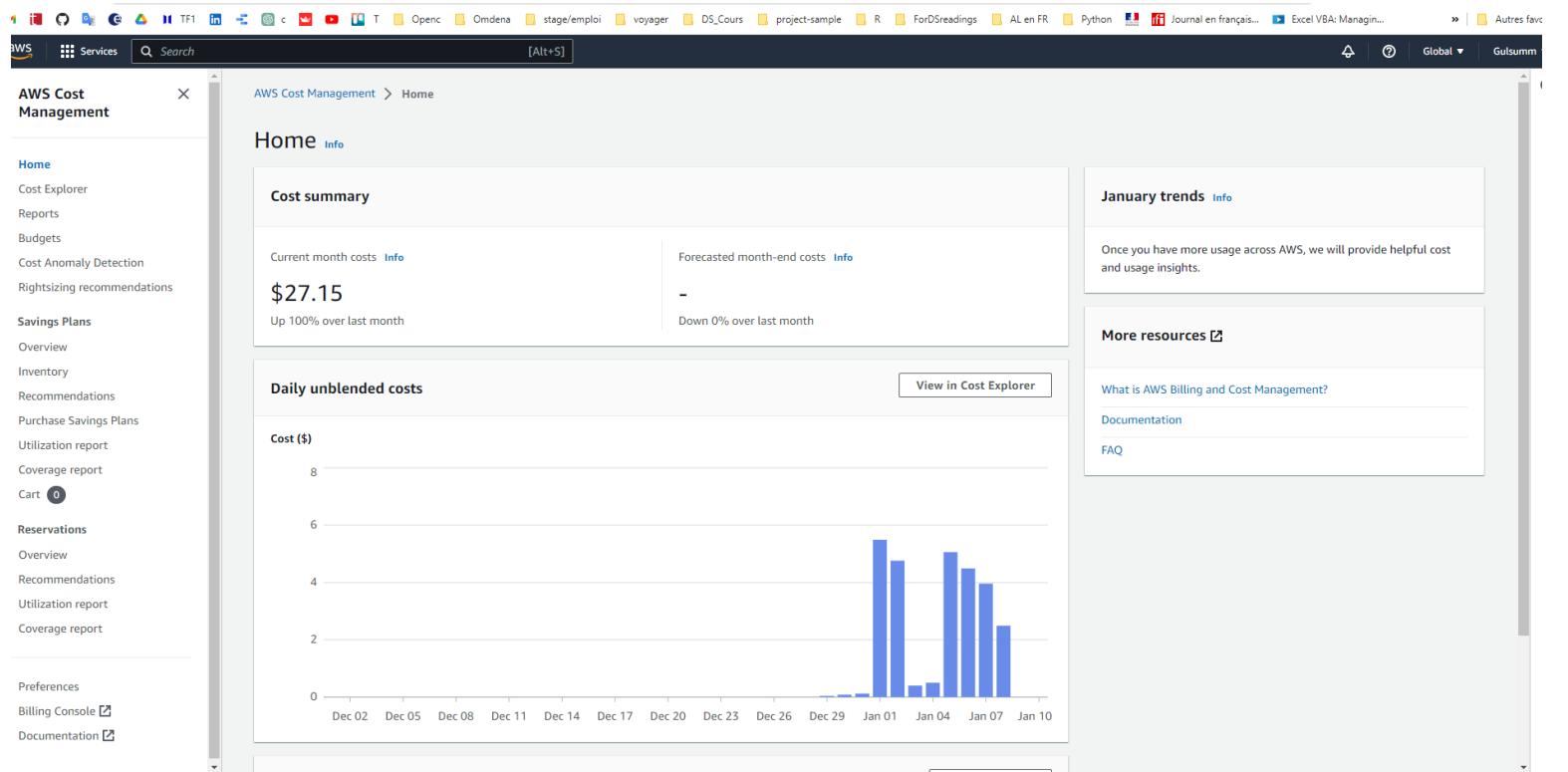
```
In [ ]:
```

Results

The screenshot shows the AWS S3 console interface. The URL in the browser is `console.aws.amazon.com/s3/buckets/oc-gulsump8?region=eu-west-3&prefix=Results/&showversions=false`. The page title is "Results/" under the "Amazon S3 > Buckets > oc-gulsump8 > Results/". The "Objects" tab is selected. There are 25 objects listed:

Name	Type	Last modified	Size	Storage class
_SUCCESS	-	January 9, 2023, 00:26:12 (UTC+01:00)	0 B	Standard
part-00000-9139a3a8-14cc-4515-93ef-b6e77ba9884b-c000.snappy.parquet	parquet	January 9, 2023, 00:19:25 (UTC+01:00)	6.1 MB	Standard
part-00001-9139a3a8-14cc-4515-93ef-b6e77ba9884b-c000.snappy.parquet	parquet	January 9, 2023, 00:19:25 (UTC+01:00)	6.2 MB	Standard
part-00002-9139a3a8-14cc-4515-93ef-b6e77ba9884b-c000.snappy.parquet	parquet	January 9, 2023, 00:19:58 (UTC+01:00)	6.1 MB	Standard
part-00003-9139a3a8-14cc-4515-93ef-b6e77ba9884b-c000.snappy.parquet	parquet	January 9, 2023, 00:19:57 (UTC+01:00)	6.1 MB	Standard
part-00004-9139a3a8-14cc-4515-93ef-b6e77ba9884b-c000.snappy.parquet	parquet	January 9, 2023, 00:20:26 (UTC+01:00)	6.1 MB	Standard
part-00005-9139a3a8-14cc-4515-93ef-b6e77ba9884b-c000.snappy.parquet	parquet	January 9, 2023, 00:20:39 (UTC+01:00)	6.1 MB	Standard
part-00006-9139a3a8-14cc-4515-93ef-b6e77ba9884b-c000.snappy.parquet	parquet	January 9, 2023, 00:21:11 (UTC+01:00)	6.1 MB	Standard
part-00007-9139a3a8-14cc-4515-93ef-b6e77ba9884b-c000.snappy.parquet	parquet	January 9, 2023, 00:21:16 (UTC+01:00)	6.1 MB	Standard
part-00008-9139a3a8-14cc-4515-93ef-b6e77ba9884b-c000.snappy.parquet	parquet	January 9, 2023, 00:21:42 (UTC+01:00)	6.1 MB	Standard
part-00009-9139a3a8-14cc-4515-93ef-b6e77ba9884b-c000.snappy.parquet	parquet	January 9, 2023, 00:21:48 (UTC+01:00)	6.1 MB	Standard
part-00010-9139a3a8-14cc-4515-93ef-b6e77ba9884b-c000.snappy.parquet	parquet	January 9, 2023, 00:22:16 (UTC+01:00)	6.1 MB	Standard
part-00011-9139a3a8-14cc-4515-93ef-b6e77ba9884b-c000.snappy.parquet	parquet	January 9, 2023, 00:22:21 (UTC+01:00)	6.1 MB	Standard
part-00012-9139a3a8-14cc-4515-93ef-b6e77ba9884b-c000.snappy.parquet	parquet	January 9, 2023, 00:22:49 (UTC+01:00)	6.1 MB	Standard

AWS COST



S3 Pour jury acces

- Bonjour,
-
- Vous avez maintenant accès à AWS Management Console pour le compte se terminant par 4704.
-
- URL de connexion :
<https://948883924704.signin.aws.amazon.com/console>
- Nom d'utilisateur : jury
-
- Votre mot de passe sera fourni séparément par votre administrateur de compte AWS.

Conclusion

Conclusion

Big Data solution dans le cloud :

- Mise en place d'une instance EC2 et d'un BucketS3
- Gestion des droits sur S3
- Configuration de session et contexte Spark

Difficultés :

- Mise en place d'un environnement Spark fonctionnel
- Choix complexes, nombreuses combinaisons techniques possibles
- Debugging compliqué dû à des erreurs explicites (superposition Spark/Java)
- Incompatibilité de library

Merci!