

Déployez un modèle dans le cloud

Formation data scientist
Projet 8 | Alexis Marceau | Octobre 2022

Sommaire

- I Contexte et problématique
- II Présentation du jeu de données
- III Chaîne de traitement des images
- IV Conclusion



I. Contexte et problématique

Contexte



Fruits!

- "Fruits!": start-up de l'Agritech
- Produit :
 - Développement de robots cueilleurs intelligents
 - Application mobile grand public de reconnaissance de fruit et affichage d'informations

Problématique



Fruits!

- Mission : Développer une **chaine de traitement d'images** incluant **preprocessing** et **réduction de dimension** dans un environnement **Big Data**
- Objectif : Prévoir le passage à l'échelle pour des volumes de données massifs

II. Présentation du jeu de données

Jeu de données

- Origine : Kaggle

- Images de 131 variétés de fruits et légumes labélisés (Fruits 360, Mihai Oltean)
- Plusieurs variétés d'un même fruit



Fruits 360

A dataset with 90380 images of 131 fruits and vegetables



- Caractéristiques :

- Images 100x100 JPEG RGB
- Fruits centrés sur fond blanc et sous tous les angles
- Plus de 90 000 images

III. Chaîne de traitement des images

Rappel : Big Data

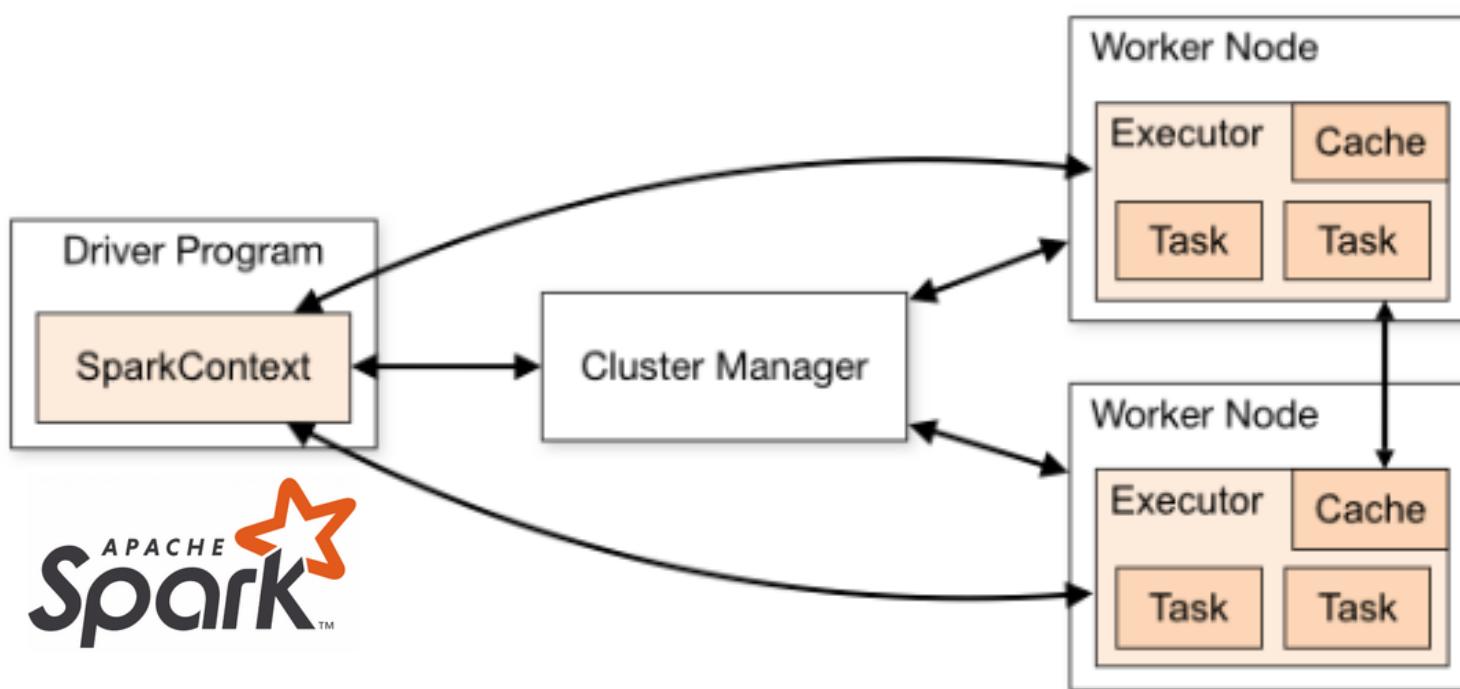
- Défini par les trois « V » du Big Data :
 - **Volume** : gros volumes de données, trop important pour être manipulées sur une seule machine
 - **Vitesse** : vitesse de circulation des données
 - **Variété** : différents types de données



Quelles solutions pour répondre à ces problématiques ?

Calculs distribués

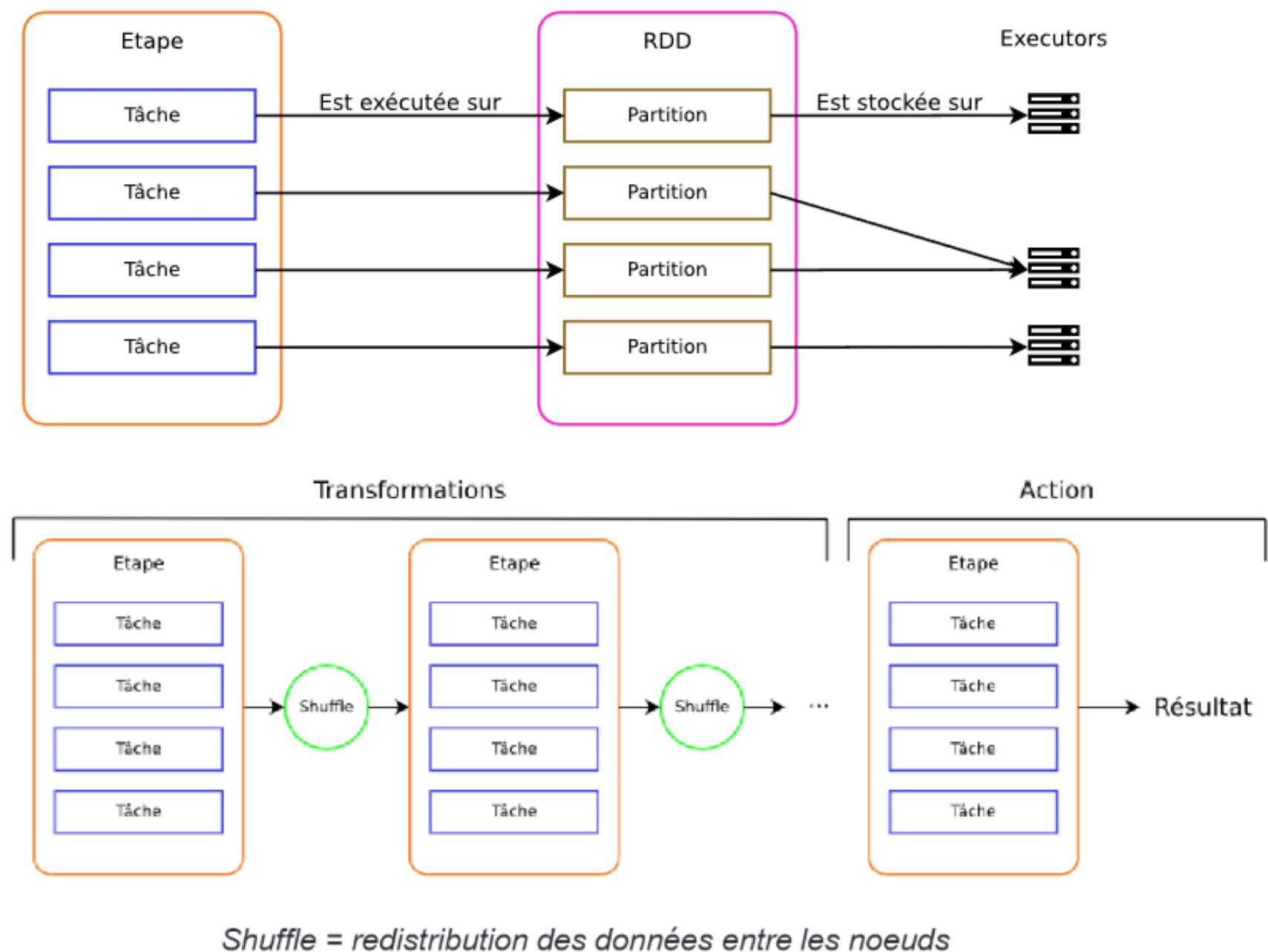
- Diviser les opérations en micro opérations distribuables entre différentes machines, réalisables en parallèle
- Agréger les résultats sur une même machine



- **Driver** : chargé de répartir les tâches sur les différents executors.
- **Cluster manager** : chargé d'instancier les différents workers.
- **Un ou plusieurs workers** : chaque worker instancie un executor chargé d'exécuter les différentes tâches de calculs.

Fonctionnement d'un cluster de calcul

- RDD (Resilient Distributed Datasets) principale innovation de Spark
- RDD = collection d'éléments partitionnés et répartis entre les nœuds du cluster
- Permet d'effectuer des calculs parallèles en mémoire sur un cluster de façon complètement tolérante aux pannes.
- Chaque tâche s'exécute sur une partition différente des données et ces partitions sont créées par les RDD



Rappel problématique

Développer une **chaine de traitement d'images** incluant **preprocessing** et
réduction de dimension dans un environnement **Big Data**

Outils de résolution



PySpark est une interface pour Apache Spark en Python.



Emulation BASH pour communiquer avec la machine virtuelle en ligne de commande

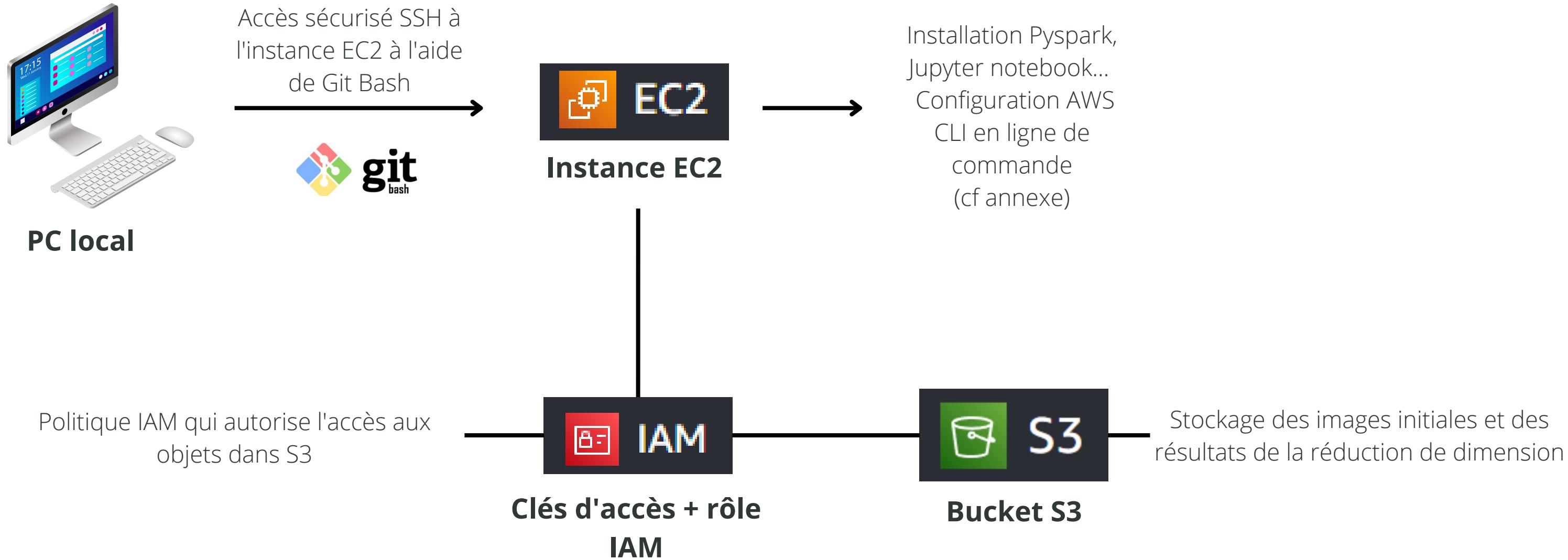


Amazon Web Services (AWS) plateforme cloud proposant des services de stockage (S3), serveur virtuel (EC2)...

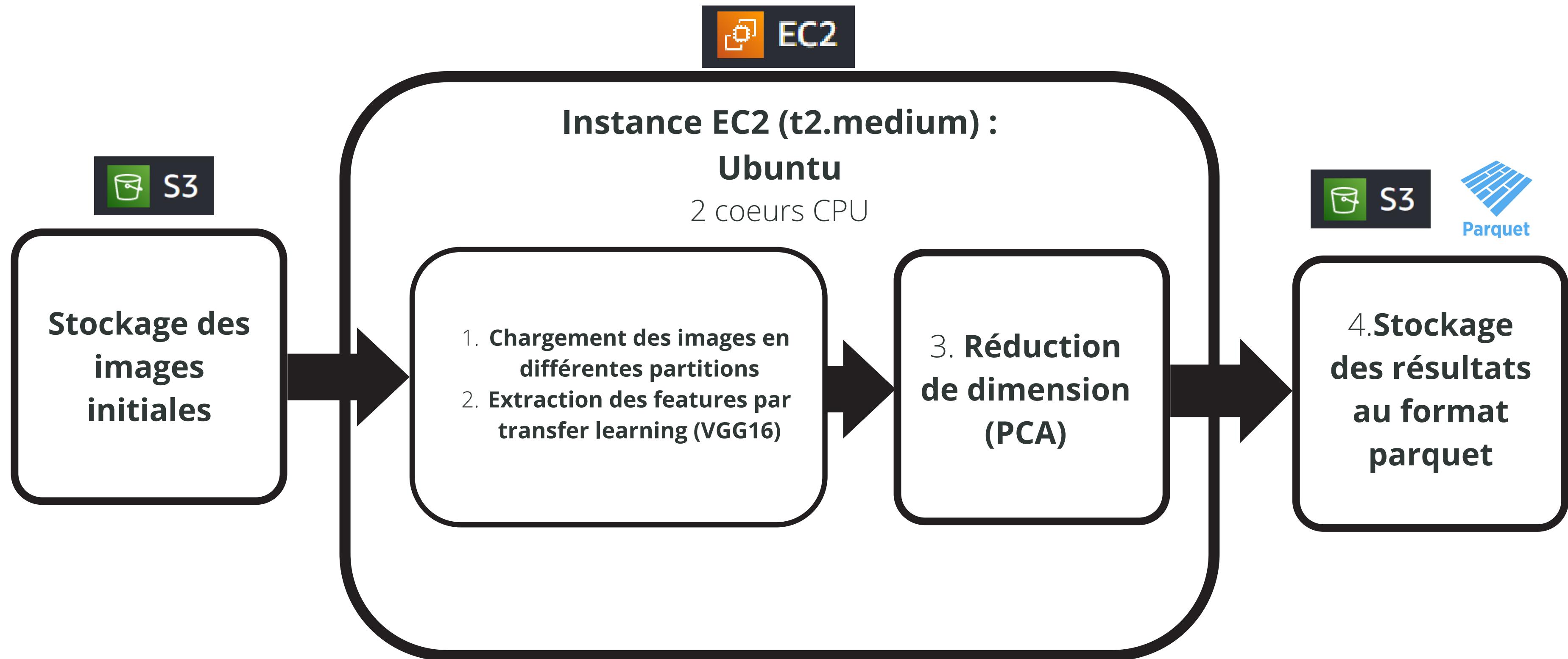


Format de fichier pour une exploitation optimisée en mode distribué conçue pour les données massives

Architecture Big Data aws



Chaine de traitement



Rappels VGG16, PCA

VGG16

VGG16 est un modèle de réseau de neurones à convolution pré-entraîné sur 14 millions d'images appartenant à 1000 classes. Il comprend 16 couches profondes.

PCA (ou ACP)

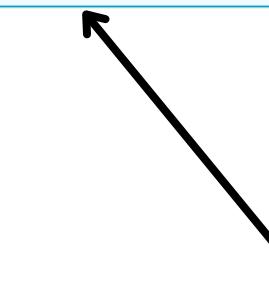
L'analyse en composantes principales (ACP) est une technique populaire pour l'analyse de grands ensembles de données contenant un nombre élevé de dimensions/caractéristiques par observation.

Représente les données dans un espace de plus petite dimension en conservant au maximum la variance.

Captures



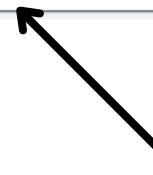
| Instances (1/1) Informations | | | | | | | | | | | | | | | | | |
|---|-----------|---|---------------------|-----------------------------------|---|-----------------|---|--------------------|------------------|---------------------------|---|--------------------------|---|-----------------|---|--------------|---|
| <input type="text"/> Find instance by attribute or tag (case-sensitive) | | | | | | | | | | | | | | | | | |
| <input checked="" type="checkbox"/> | Name | ▼ | ID d'instance | État de l'insta... | ▼ | Type d'insta... | ▼ | Contrôle des st... | Statut d'alar... | Zone de dispon... | ▼ | DNS IPv4 public | ▼ | Adresse IPv4... | ▼ | IP élastique | ▼ |
| <input checked="" type="checkbox"/> | p8_medium | | i-0efe40eada8040459 | En cours d'exécution | | t2.medium | | - | Aucune al... | + eu-west-3c | | ec2-13-37-250-245.eu-... | | 13.37.250.245 | | - | |



Etat actuel de l'instance. Eteinte après chaque utilisation pour limiter les couts d'utilisation.



| Compartiments (2) Info | | | | |
|---|----------------------|---------------------|------------------------------|--|
| Les compartiments sont des conteneurs pour les données stockées dans S3. En savoir plus ? | | | | |
| <input type="text"/> Rechercher des compartiments par nom | | | | |
| Nom | Région AWS | Accéder | Date de création | |
| imagesp8alexis | EU (Paris) eu-west-3 | Public | 26 Sep 2022 10:44:52 AM CEST | |
| p8alexis | EU (Paris) eu-west-3 | Public | 28 Sep 2022 05:17:45 PM CEST | |



Accès aux images et résultats rendu public

Captures

▼ Règles entrantes

| ID de règle du groupe de ... | Plage de ports | Protocole | Source | Groupes de sécurité |
|------------------------------|----------------|-----------|----------------|---------------------|
| sgr-0444949ebf4b79a14 | 22 | TCP | 90.████████/32 | launch-wizard-4 |
| sgr-0a81625271ba2a975 | 4040 | TCP | 90.████████/32 | launch-wizard-4 |
| sgr-08acb5f486e9bb227 | 8888 | TCP | 90.████████/32 | launch-wizard-4 |

Diagramme des règles entrantes :

- Port 22 : Accès SSH (Port SSH pour se connecter à la machine virtuelle)
- Port 4040 : Accès spark UI
- Port 8888 : Accès jupyter notebook

```
graph TD; A[Accès jupyter notebook] --> B[8888]; C[Accès spark UI] --> D[4040]; E[Port SSH pour se connecter à la machine virtuelle] --> F[22];
```

Captures

Jupyter notebook

Déploiement d'un modèle dans le cloud

Ce notebook est une première chaîne de traitement d'images qui comprend le preprocessing et une étape de réduction de dimension pour un outil de classification d'image

This screenshot shows a Jupyter notebook interface. The title bar indicates it's a 'solution_p8' notebook, last saved 10 minutes ago. The toolbar includes standard options like File, Edit, View, Insert, Cell, Kernel, Help, and execution controls. The main content area displays a section titled 'Déploiement d'un modèle dans le cloud' with a descriptive text about image processing and dimensionality reduction.

Environnement virtuel où est installé jupyter notebook et les différentes librairies (pyspark, tensorflow...)

Notebook pour les essais tensorflow avec image locale

Jupyter notebook contenant la solution

Image locale

jupyter

Files Running Clusters

Select items to perform actions on them.

| Name | Last Modified | File size |
|------------------------|------------------------|-----------|
| 0 | il y a 11 jours | |
| certs | il y a 5 jours | |
| pyenv | il y a 6 jours | 24.7 kB |
| essai_tensorflow.ipynb | Actif il y a 4 minutes | 94.8 kB |
| solution_p8.ipynb | il y a 6 jours | 4.95 kB |
| 0_100.jpg | | |

This screenshot shows the Jupyter file manager interface. It lists several files and directories: '0', 'certs', 'pyenv', 'essai_tensorflow.ipynb', 'solution_p8.ipynb', and '0_100.jpg'. Arrows from the left side of the image point to specific items in the file list: one arrow points to the '0' directory, another to 'essai_tensorflow.ipynb', and a third to '0_100.jpg'.

Captures

Dataframe Spark après traitements

| image | category | cnn_features | scaled_features | PCA_scaled_features |
|----------------------|------------------|----------------------|----------------------|----------------------|
| {s3a://imagesp8al... | Apricot | [0.34555277228355... | [-1.3051236662838... | [19.7924763401514... |
| {s3a://imagesp8al... | Fig | [0.43078595399856... | [0.29667724873002... | [-15.021246715174... |
| {s3a://imagesp8al... | Apricot | [0.33527731895446... | [-1.4982319160702... | [20.1023038825190... |
| {s3a://imagesp8al... | AppleGrannySmith | [0.41182667016983... | [-0.0596276265991... | [13.8924662689332... |
| {s3a://imagesp8al... | Banana | [0.45888328552246... | [0.82471492653149... | [-14.998285024456... |
| {s3a://imagesp8al... | Banana | [0.46727916598320... | [0.98250006060434... | [-19.732085826666... |
| {s3a://imagesp8al... | Apricot | [0.33720248937606... | [-1.4620518781926... | [19.5415504641457... |
| {s3a://imagesp8al... | Apricot | [0.31869816780090... | [-1.8098065745962... | [19.3937144705535... |
| {s3a://imagesp8al... | Fig | [0.42452642321586... | [0.17904087832195... | [-14.793826908495... |
| {s3a://imagesp8al... | Fig | [0.44485887885093... | [0.56115199262522... | [-15.655091885204... |
| {s3a://imagesp8al... | Banana | [0.47459599375724... | [1.12000638370110... | [-16.864074542526... |
| {s3a://imagesp8al... | AppleGrannySmith | [0.41704300045967... | [0.03840370888729... | [13.4793330963494... |
| {s3a://imagesp8al... | Apricot | [0.33236104249954... | [-1.5530379692990... | [19.7857451435443... |
| {s3a://imagesp8al... | Apricot | [0.34890440106391... | [-1.2421359666679... | [19.4981801842992... |
| {s3a://imagesp8al... | AppleGrannySmith | [0.40586775541305... | [-0.1716144716028... | [14.3048741280020... |
| {s3a://imagesp8al... | Banana | [0.49525517225265... | [1.50825765335364... | [-16.034574260345... |
| {s3a://imagesp8al... | Apricot | [0.34523218870162... | [-1.3111484451849... | [19.1629881841080... |
| {s3a://imagesp8al... | AppleGrannySmith | [0.41303160786628... | [-0.0369830383544... | [13.1352527216536... |
| {s3a://imagesp8al... | Fig | [0.45605725049972... | [0.77160479507570... | [-15.961338081706... |
| {s3a://imagesp8al... | Fig | [0.45045301318168... | [0.66628346031660... | [-15.172182620221... |

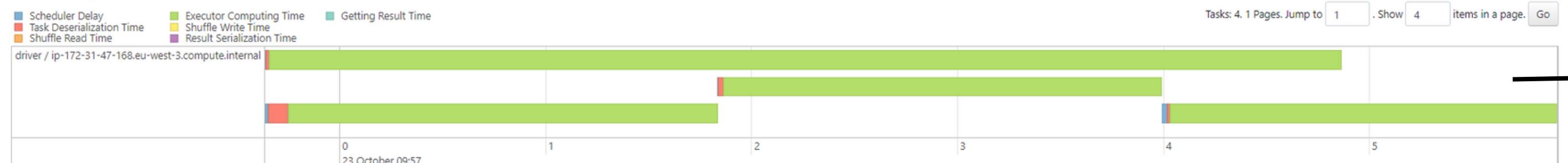
Captures

Spark UI

Details for Stage 8 (Attempt 0)

Resource Profile Id: 0
Total Time Across All Tasks: 11 s
Locality Level Summary: Node local: 4
Shuffle Read Size / Records: 371.6 KiB / 18
Associated Job Ids: 5

- ▶ DAG Visualization
- ▶ Show Additional Metrics
- ▼ Event Timeline
- Enable zooming



Parallélisation des tâches sur les deux coeurs du CPU

Summary Metrics for 4 Completed Tasks

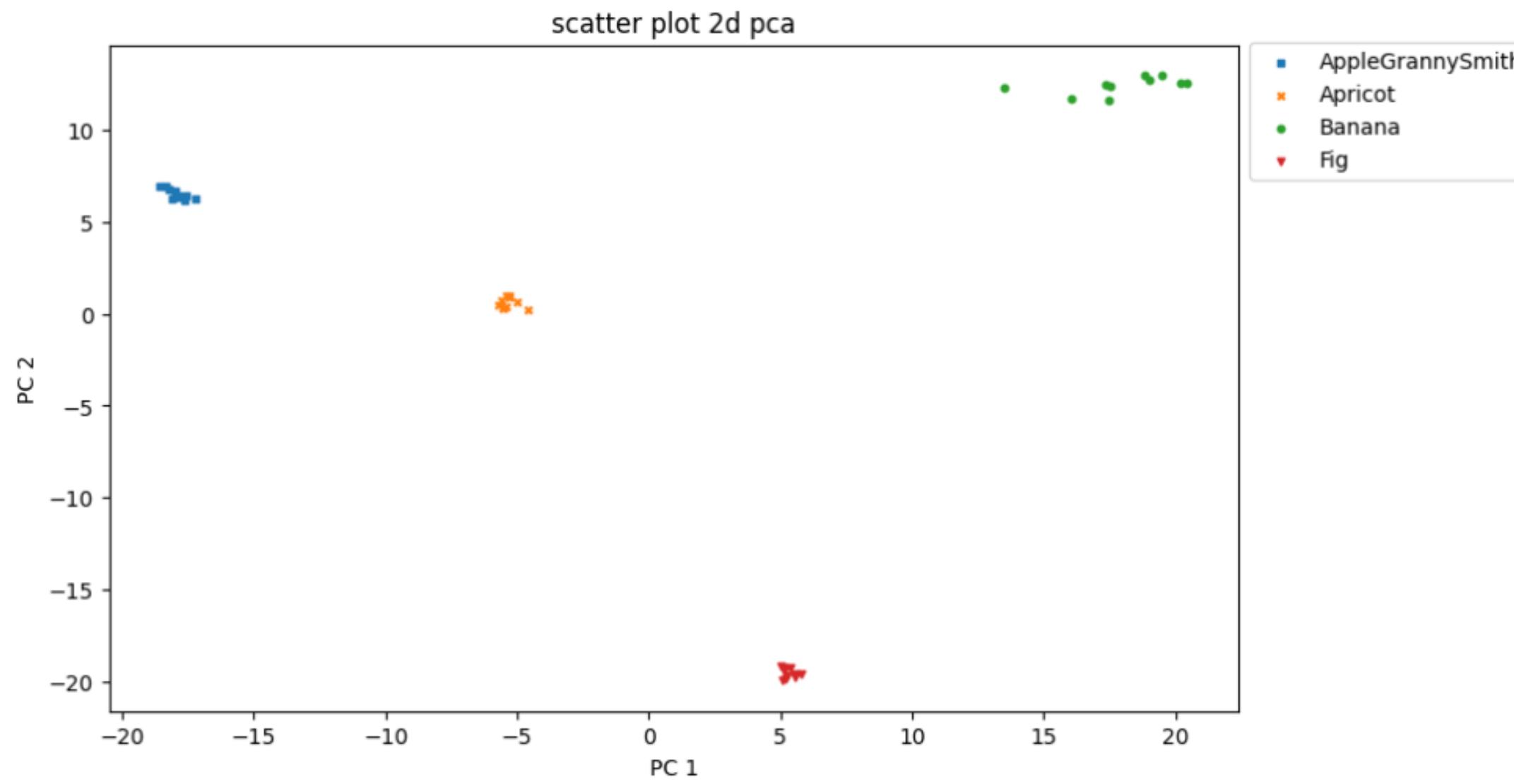
| Metric | Min | 25th percentile | Median | 75th percentile | Max |
|-----------------------------|--------------|-----------------|--------------|-----------------|---------------|
| Duration | 2 s | 2 s | 2 s | 5 s | 5 s |
| GC Time | 0.0 ms | 0.0 ms | 0.0 ms | 0.0 ms | 0.0 ms |
| Shuffle Read Size / Records | 76.3 KiB / 4 | 83.5 KiB / 4 | 87.9 KiB / 4 | 123.9 KiB / 6 | 123.9 KiB / 6 |

Aggregated Metrics by Executor

Tasks (4)

| Index | Task ID | Attempt | Status | Locality level | Executor ID | Host | Logs | Launch Time | Duration | GC Time | Shuffle Read Size / Records | Errors |
|-------|---------|---------|---------|----------------|-------------|---|------|---------------------|----------|---------|-----------------------------|--------|
| 0 | 8 | 0 | SUCCESS | NODE_LOCAL | driver | ip-172-31-47-168.eu-west-3.compute.internal | | 2022-10-23 11:56:59 | 5 s | | 123.9 KiB / 6 | |
| 1 | 9 | 0 | SUCCESS | NODE_LOCAL | driver | ip-172-31-47-168.eu-west-3.compute.internal | | 2022-10-23 11:56:59 | 2 s | | 87.9 KiB / 4 | |
| 2 | 10 | 0 | SUCCESS | NODE_LOCAL | driver | ip-172-31-47-168.eu-west-3.compute.internal | | 2022-10-23 11:57:01 | 2 s | | 83.5 KiB / 4 | |
| 3 | 11 | 0 | SUCCESS | NODE_LOCAL | driver | ip-172-31-47-168.eu-west-3.compute.internal | | 2022-10-23 11:57:03 | 2 s | | 76.3 KiB / 4 | |

Captures



La visualisation 2d des deux premières composantes de la PCA permet de voir l'efficacité du VGG16 pour distinguer les échantillons de fruits.

Captures



LIENS OUVERT BUCKET : <https://imagep8alexis.s3.eu-west-3.amazonaws.com/>

Amazon S3 > Compartiments > p8alexis > results.parquet/

results.parquet/

Objets Propriétés

Objets (9)

Les objets sont les entités fondamentales stockées dans Amazon S3. Vous pouvez utiliser l'inventaire Amazon S3 pour obtenir une liste de tous les objets de votre compartiment. Pour que d'autres personnes puissent accéder à vo

Copier l'URI S3 Copier l'URL Télécharger Ouvrir Supprimer Actions ▾ Créer un dossier Charger

Rechercher des objets en fonction du préfixe

| <input type="checkbox"/> | Nom | Type | Dernière modification |
|--------------------------|---|---------|------------------------------|
| <input type="checkbox"/> | _SUCCESS | - | 23 Oct 2022 11:57:31 AM CEST |
| <input type="checkbox"/> | part-00000-310a328c-1a95-4229-8c77-e931e69a9beb-c000.snappy.parquet | parquet | 23 Oct 2022 11:57:29 AM CEST |
| <input type="checkbox"/> | part-00001-310a328c-1a95-4229-8c77-e931e69a9beb-c000.snappy.parquet | parquet | 23 Oct 2022 11:57:30 AM CEST |
| <input type="checkbox"/> | part-00002-310a328c-1a95-4229-8c77-e931e69a9beb-c000.snappy.parquet | parquet | 23 Oct 2022 11:57:29 AM CEST |
| <input type="checkbox"/> | part-00003-310a328c-1a95-4229-8c77-e931e69a9beb-c000.snappy.parquet | parquet | 23 Oct 2022 11:57:27 AM CEST |
| <input type="checkbox"/> | part-00004-310a328c-1a95-4229-8c77-e931e69a9beb-c000.snappy.parquet | parquet | 23 Oct 2022 11:57:28 AM CEST |
| <input type="checkbox"/> | part-00005-310a328c-1a95-4229-8c77-e931e69a9beb-c000.snappy.parquet | parquet | 23 Oct 2022 11:57:28 AM CEST |
| <input type="checkbox"/> | part-00006-310a328c-1a95-4229-8c77-e931e69a9beb-c000.snappy.parquet | parquet | 23 Oct 2022 11:57:31 AM CEST |
| <input type="checkbox"/> | part-00007-310a328c-1a95-4229-8c77-e931e69a9beb-c000.snappy.parquet | parquet | 23 Oct 2022 11:57:30 AM CEST |

```
#lecture depuis le S3 pour vérification
spark.read.parquet("s3a://p8alexis/results.parquet").show(20)
```

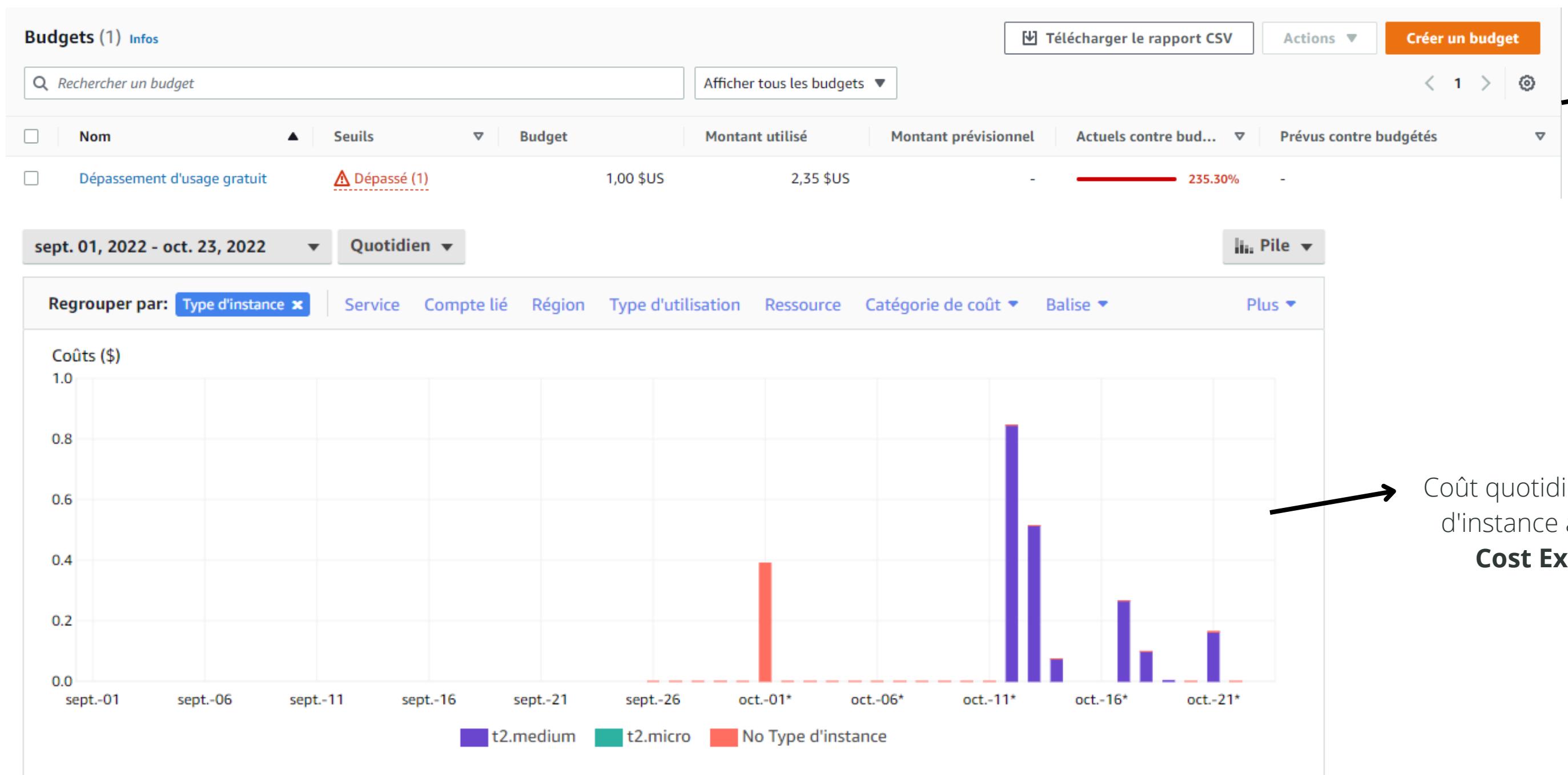
| PCA_scaled_features | category |
|----------------------|------------------|
| [19.7924763401514... | Apricot |
| [-15.021246715174... | Fig |
| [20.1023038825190... | Apricot |
| [13.8924662689332... | AppleGrannySmith |
| [-14.998285024456... | Banana |
| [-19.732085826666... | Banana |
| [19.068888999963,... | Apricot |
| [-14.844720721926... | Fig |
| [13.5888146129590... | AppleGrannySmith |
| [-16.706050269590... | Banana |
| [14.1982429300848... | AppleGrannySmith |
| [-14.810123622129... | Fig |
| [19.5415504641457... | Apricot |
| [19.3937144705535... | Apricot |
| [-14.793826908495... | Fig |
| [-15.655091885204... | Fig |
| [-16.864074542526... | Banana |
| [13.4793330963494... | AppleGrannySmith |
| [13.8436557407524... | AppleGrannySmith |
| [13.0319204379996... | AppleGrannySmith |

only showing top 20 rows

La sauvegarde au format parquet permet de conserver la structure partitionnée

Captures

Gestion des coûts



Définition d'un budget avec **AWS Budgets**

Coût quotidien par type d'instance avec **AWS Cost Explorer**

Conclusion

- Apprentissage
 - Pyspark
 - Ecosystème AWS
 - Administration d'un serveur Linux par SSH
- Difficultés
 - Superposition Spark/Java
 - Complexité architecture data



MERCI DE VOTRE
ATTENTION

Annexe

Installation jupyter notebook EC2

```
sudo apt-get update
```

```
sudo apt-get install python3-pip
```

```
sudo apt-get install python3-venv
```

```
python3 -m venv pyenv
```

```
source pyenv/bin/activate
```

```
pip3 install jupyterlab
```

```
jupyter notebook --generate-config
```

```
jupyter server password
```

```
jupyter notebook password
```

```
mkdir certs
```

```
cd certs/
```

```
openssl req -x509 -nodes -days 365 -newkey rsa:2048 -keyout mykey.key -out mycert.pem
```

Ahnexe

Installation jupyter notebook EC2

```
cd ~/jupyter/  
vim jupyter_notebook_config.py  
i
```

```
c.IPKernelApp.pylab = 'inline'  
c.NotebookApp.certfile = u'/home/ubuntu/certs/mycert.pem'  
c.NotebookApp.keyfile= u'/home/ubuntu/certs/mykey.key'  
c.NotebookApp.ip = '0.0.0.0'  
c.NotebookApp.open_browser = False  
c.NotebookApp.port = 8888  
c.NotebookApp.allow_root = True
```

Esc

:wq

Annexe

Installation jupyter notebook EC2

```
sudo apt-get install awscli
```

```
aws configure
```

key id :

secret acces key : ...

region : eu-west-3

format : json

```
aws s3 ls
```

```
sudo apt-get install default-jre
```

```
sudo apt-get install scala
```

```
pip install py4j
```

```
wget http://archive.apache.org/dist/spark/spark-3.3.0/spark-3.3.0-bin-hadoop3.tgz
```

```
sudo tar -zxvf spark-3.3.0-bin-hadoop3.tgz
```

Annexe

Installation jupyter notebook EC2

pip3 install findspark

pip3 install pyspark

pip3 install pyarrow

pip3 install pandas

pip3 install numpy

pip3 install pillow

pip install tensorflow --no-cache-dir ou tensorflow cpu

jupyter notebook

<https://....:8888>

thisisunsafe

<https://mvnrepository.com/artifact/org.apache.hadoop/hadoop-aws/>

<https://mvnrepository.com/artifact/com.amazonaws/aws-java-sdk-bundle/1.12.317/jar>

hadoop-aws-3.3.2.jar

aws-java-sdk-bundle-1.12.317.jar

upload les dossiers dans pyenv/lib/python3.10/site-packages/pyspark/jars