OPENCLASSROOMS

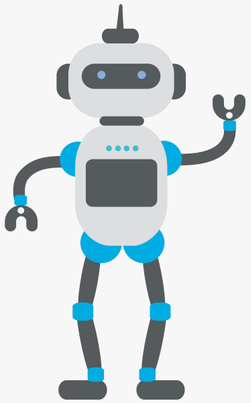
Kevin

Parcours Data Scientist

Problématique

L'entreprise

- Start-up AgriTech
- Proposition innovante pour la récolte de fruits

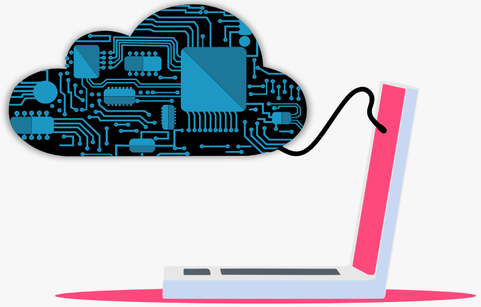


Mission

- Application mobile
- Algorithme de classification

Contraintes

- PySpark
- Architecture Big Data EC2



Jeu de données

Process

Conclusion

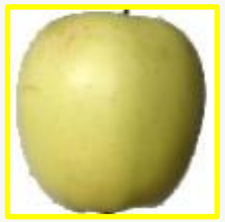
1

Jeu de données

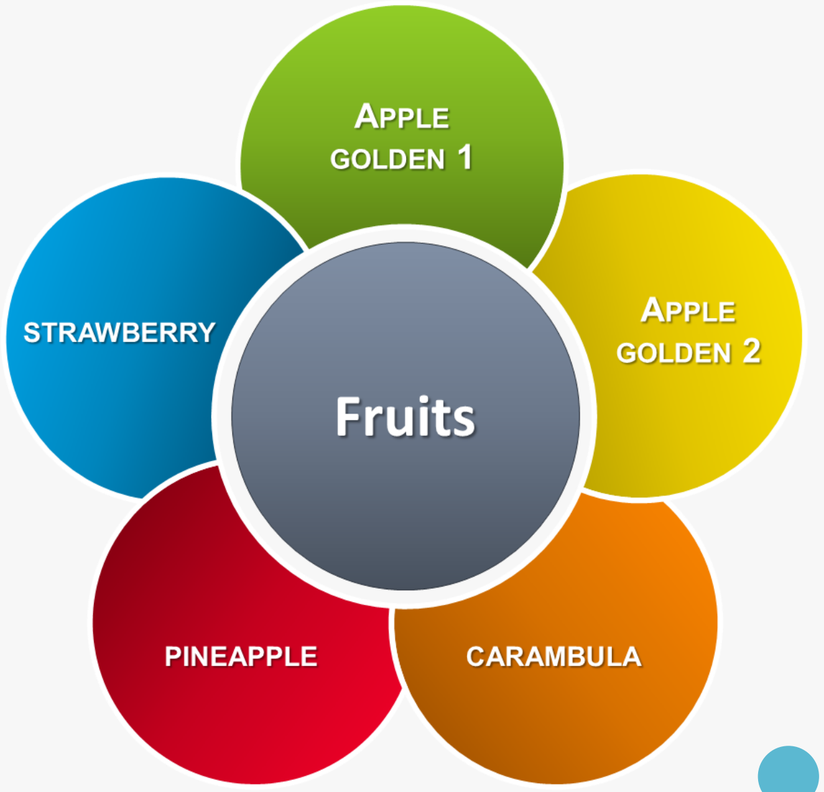
Informations générales

Dataset

- Partie d'entrainement / test
- 130+ fruits différents
- 53000 images d'entrainement
- 400 images par fruit
- Fruits sous différents angles


Détail



Dataset

Format

Dimension 100x100

Format JPG

Couleurs

Angles

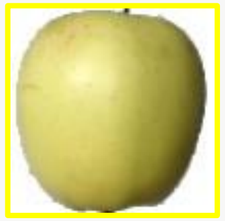
Timelapse

Rotation 3 axes

Fond et Focus

Centré sur le fruit

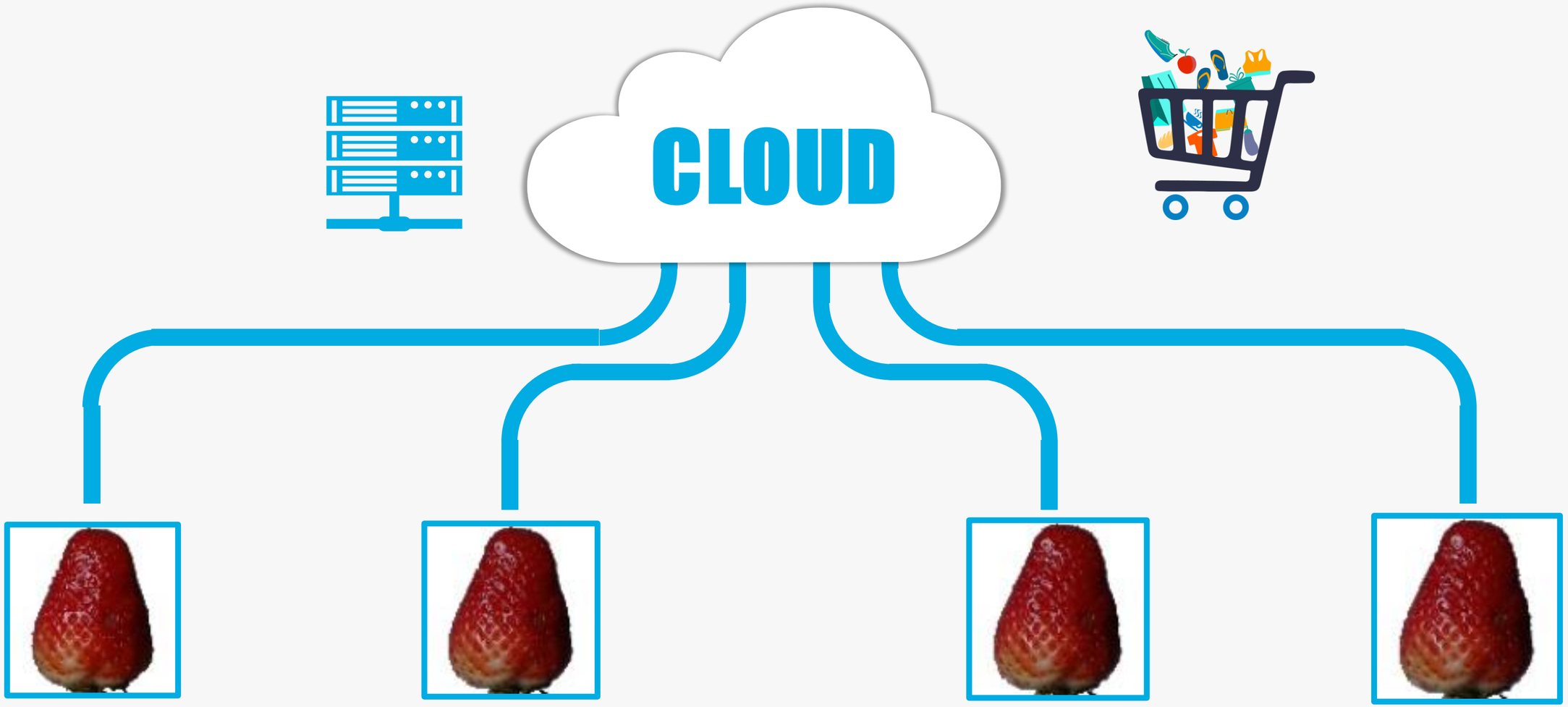
Fond blanc



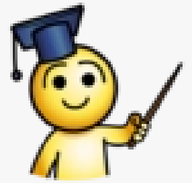
Informations

400+ images

Dataset restreint



Le Big Data

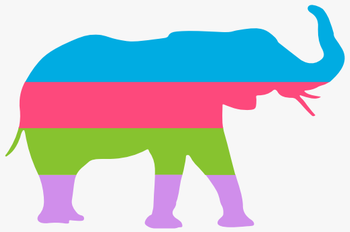


Volume

Big Data

Variété

Definition



Vitesse

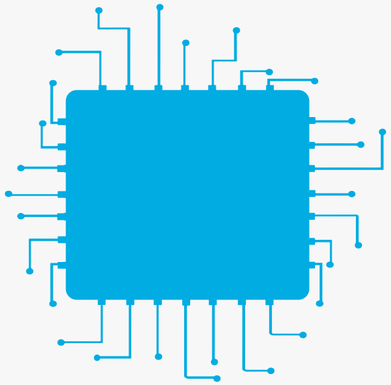
Véracité / Valeur

PySpark

Big Data

Proches de Panda

DataFrame

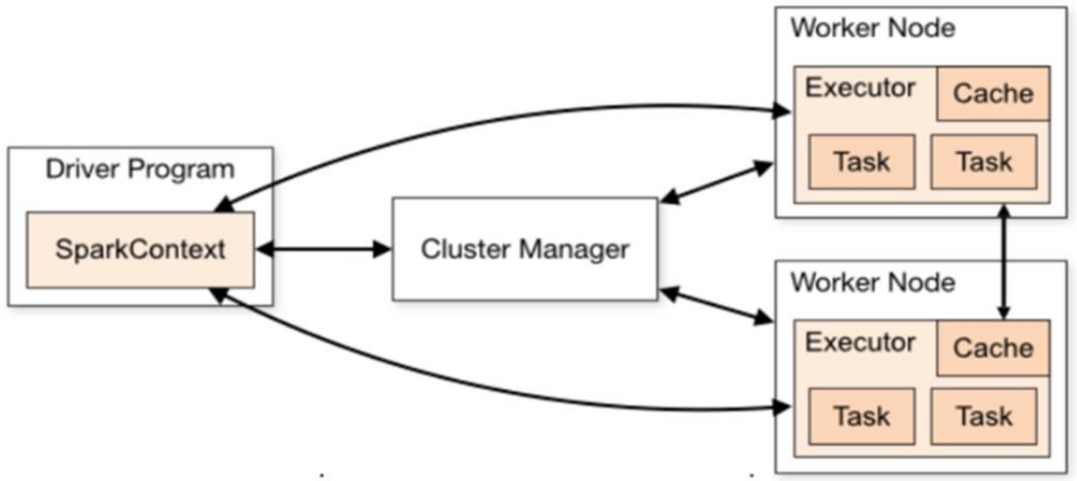


python™ + Apache Spark

=

PySpark

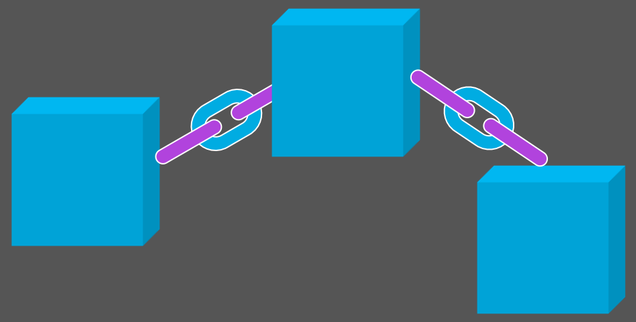
PySpark



Application maître :
Configuration / Initialisation
Aggrégation des calculs

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

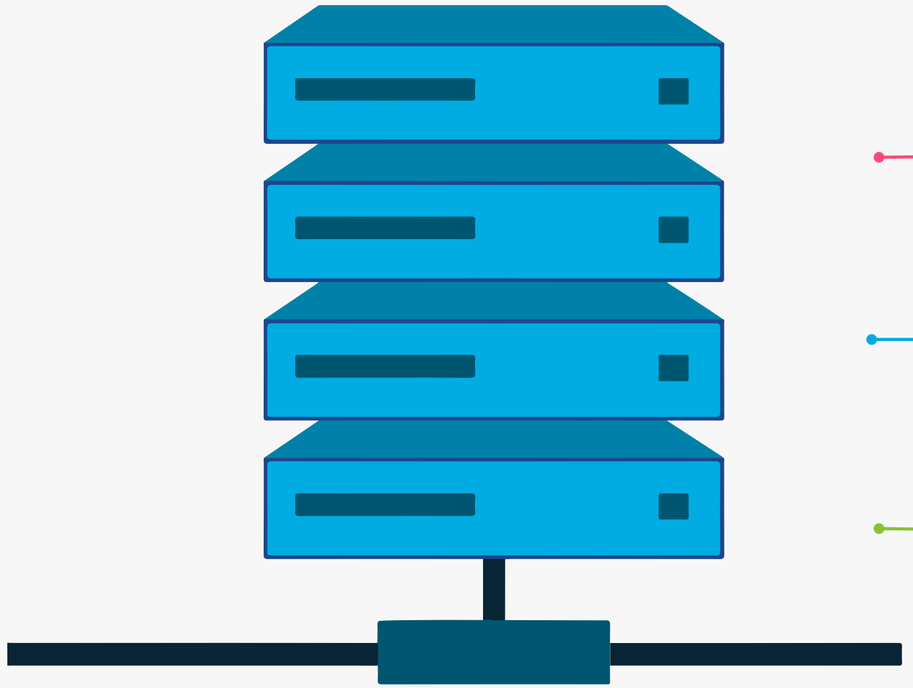
Workers :
Exécution des tâches en parallèle



- Adapté au Big Data
- Division des opérations
- Stockage
- Agrégation


2
Process

Cloud AWS



Architecture
Big DATA

EC2 Linux

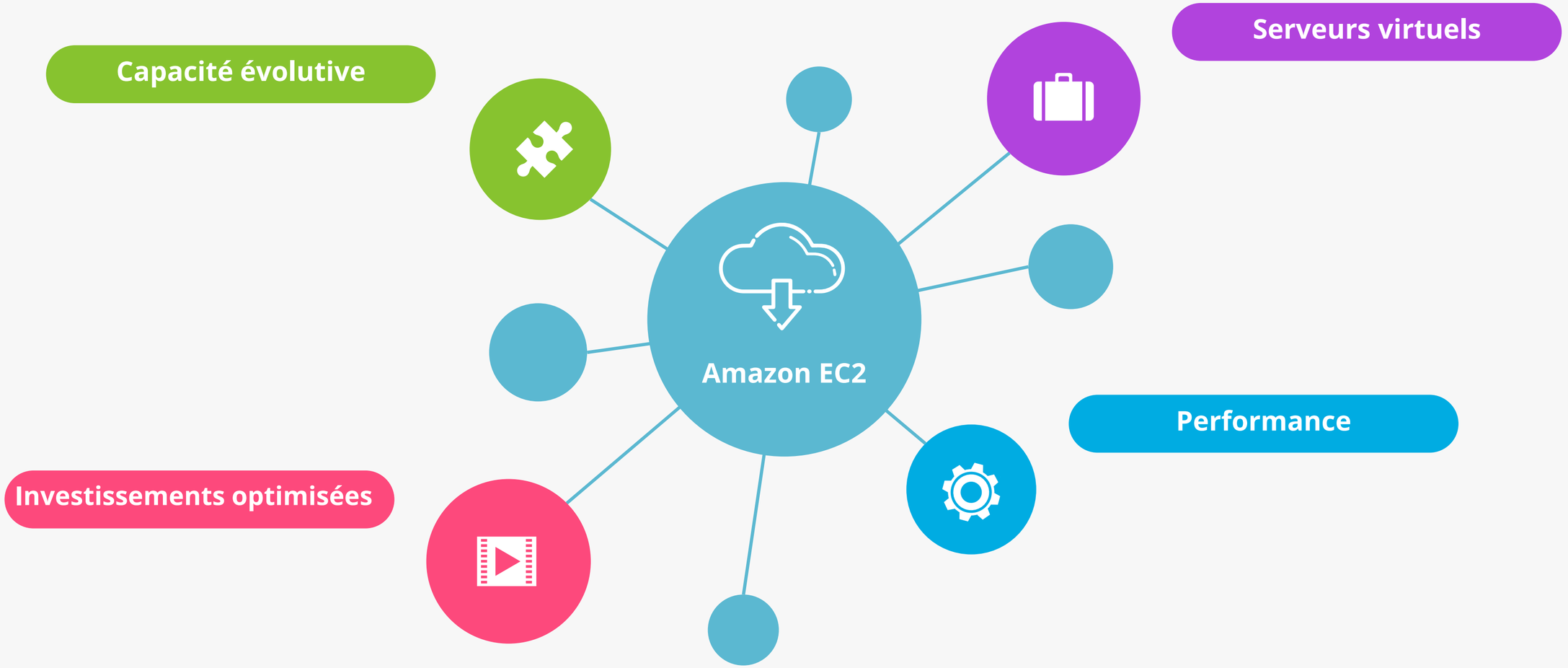
Stockage

S3

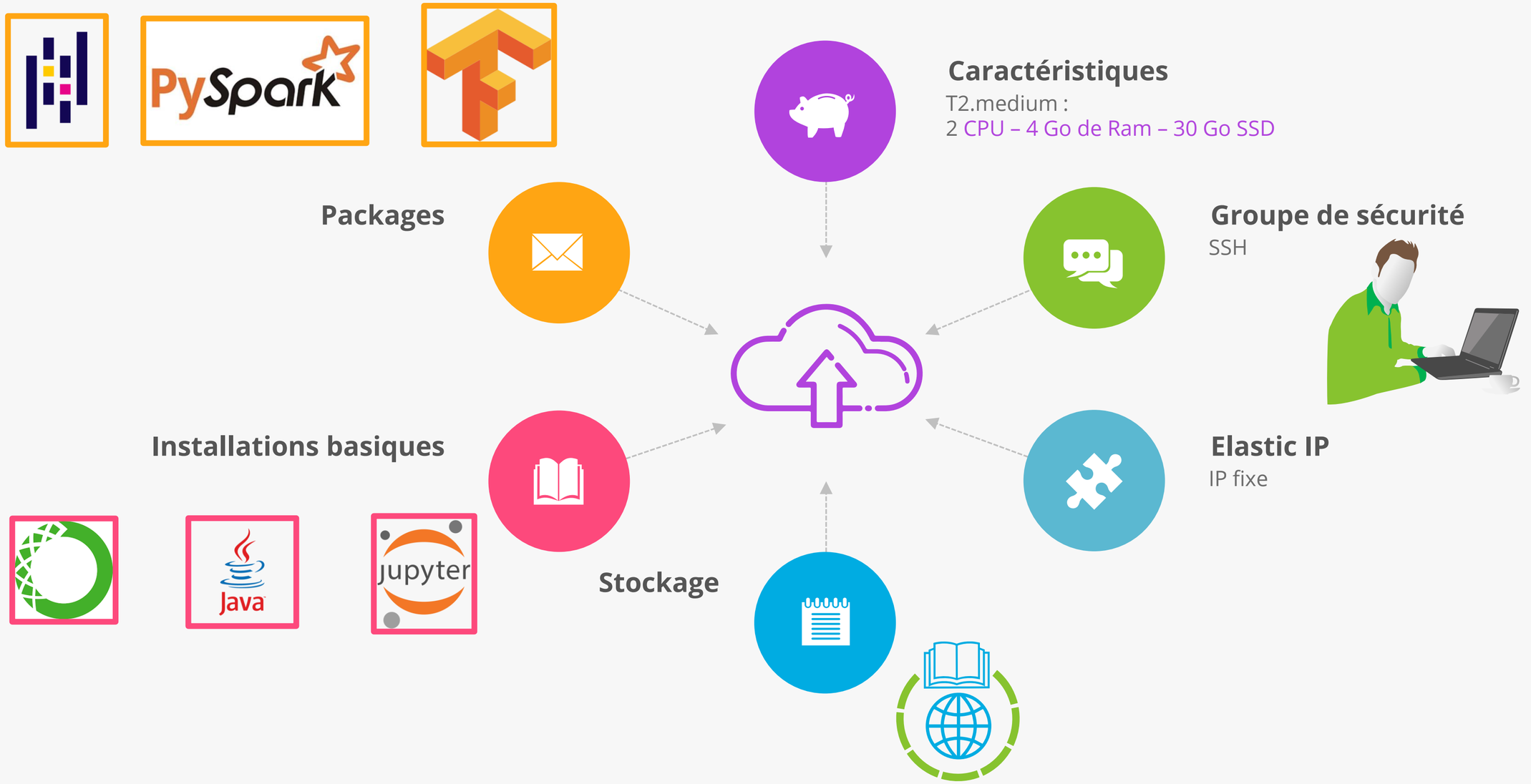
Access
Management

IAM

EC2


Capacité évolutive
Serveurs virtuels
Amazon EC2
Performance
Investissements optimisées

Instances


PySpark
Caractéristiques
T2.medium :
2 CPU – 4 Go de Ram – 30 Go SSD
Packages
Groupe de sécurité
SSH
Installations basiques
Elastic IP
IP fixe
Java
jupyter
Stockage

Utilisation de l’instance

3 étapes

Configurer l’accès web

Groupe de sécurité

01

Configurer Jupyter

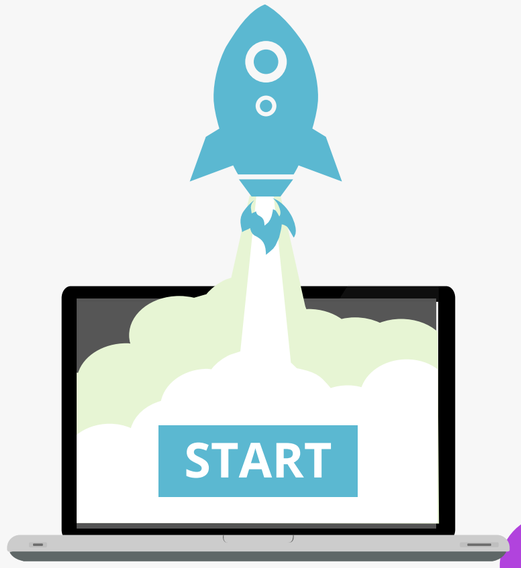
Accès distants

02

Lancer le notebook

Téléverser le notebook

03



Stockage : **Amazon S3**

Avantages

Durable

Sécurité

Simple

Rapide

IAM

Management des accès aux services et ressources

Upload du dataset

Création d’un bucket

Process Cloud

6 steps

Session PySpark

Chargement S3

Modèle Resnet

Features

DataFrame

Sauvegarde

```
+--------------+--------------------+
|         Label|               Image|
+--------------+--------------------+
|Apple Golden 1|[FF D8 FF E0 00 1...|
|Apple Golden 1|[FF D8 FF E0 00 1...|
|Apple Golden 1|[FF D8 FF E0 00 1...|
|Apple Golden 1|[FF D8 FF E0 00 1...|
|Apple Golden 2|[FF D8 FF E0 00 1...|
|Apple Golden 2|[FF D8 FF E0 00 1...|
|Apple Golden 2|[FF D8 FF E0 00 1...|
|Apple Golden 2|[FF D8 FF E0 00 1...|
```

- Accéder au S3
- Télécharger les données
- Création du DataFrame

Données téléchargées :

- URL de l'image : Label au format texte
- Image au format binaire

Simulations

DataFrame

Stockage des résultats

Application

Pré-processing : Redimensionner l'image

Array

Extraction des features obtenues

Resnet50

Réseau de neurones pré-entrainés

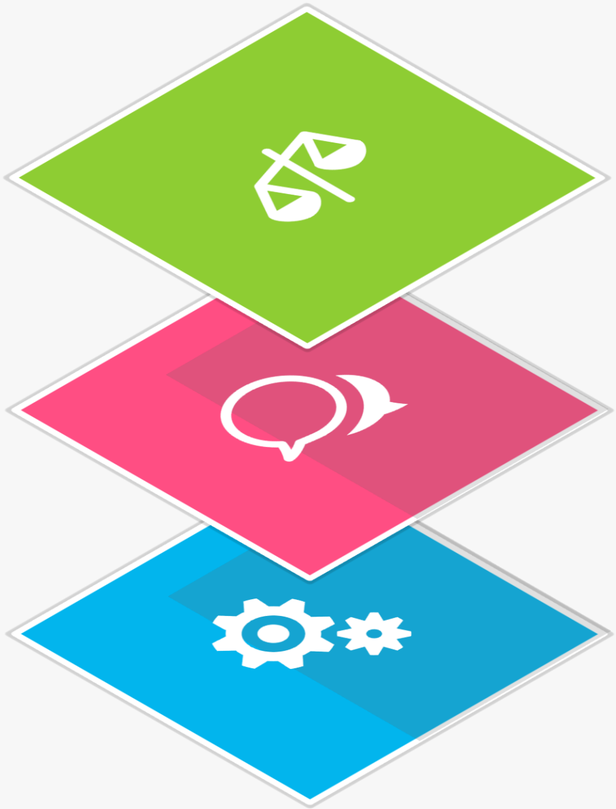
Dernières couches retirées

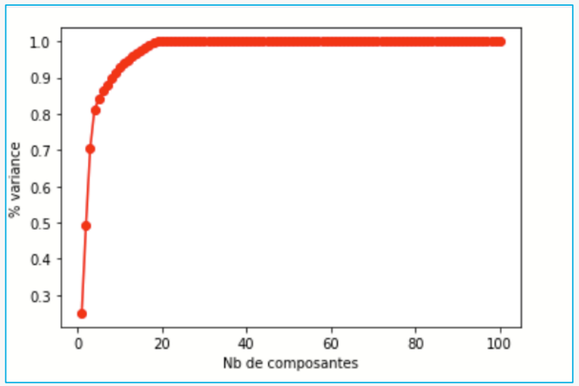
Manipulation Features

DenseVector

StandardScaler

PCA





Segmentation personnalisée

```
+--------------+--------------------+--------------------+--------------------+
|         Label|            features|      ScaledFeatures|         PCAFeatures|
+--------------+--------------------+--------------------+--------------------+
|Apple Golden 1|[0.10101263970136...|[-0.6368764608745...|[25.6296319881354...|
|Apple Golden 2|[0.42274096608161...|[-0.3263586233441...|[18.9299228690892...|
|Apple Golden 1|[0.39658847451210...|[-0.3515998422929...|[26.2584256959103...|
|     Carambula|[2.95661902427673...|[2.11922751465298...|[6.84219591157919...|
|Apple Golden 2|[0.35389965772628...|[-0.3928011851284...|[19.3471739277303...|
|    Strawberry|[0.56954985857009...|[-0.1846652252568...|[-21.757498342526...|
|    Strawberry|[0.49779418110847...|[-0.2539206061759...|[-22.837205721816...|
|     Carambula|[2.78103446960449...|[1.94976113028455...|[5.54254530701162...|
|     Pineapple|[0.0,0.0613296367...|[-0.7343693574317...|[-31.304020620157...|
|     Carambula|[2.92338633537292...|[2.08715280488205...|[5.86885746353973...|
```


```
{'ResponseMetadata': {'RequestId': '5T9005XB820JAW3P',
  'HostId': '4V1+By5P/8yDtnpKBj4eeWuIZF9WUAoS8Nvi/7Tj7uaeKcI+GIZLJaTZnmyy2Qa8qGR7DxkTbRw=',
  'HTTPStatusCode': 200,
  'HTTPHeaders': {'x-amz-id-2': '4V1+By5P/8yDtnpKBj4eeWuIZF9WUAoS8Nvi/7Tj7uaeKcI+GIZLJaTZnmyy2Qa8qGR7DxkTbRw=',
   'x-amz-request-id': '5T9005XB820JAW3P',
   'date': 'Fri, 30 Sep 2022 09:35:32 GMT',
   'etag': '"676d6072740ba261ab7dc27624d42fe2"',
   'server': 'AmazonS3',
   'content-length': '0'},
  'RetryAttempts': 0},
 'ETag': '"676d6072740ba261ab7dc27624d42fe2"'}
```


Vecteurs
Anciennes variables

String
Nouvelles variables

3

Conclusion et recommandation

Personnel

Zone de confort 01

- Découverte Linux

AWS & PySpark 02

- Découverte d'un écosystème
- Curiosité

Difficultés 03

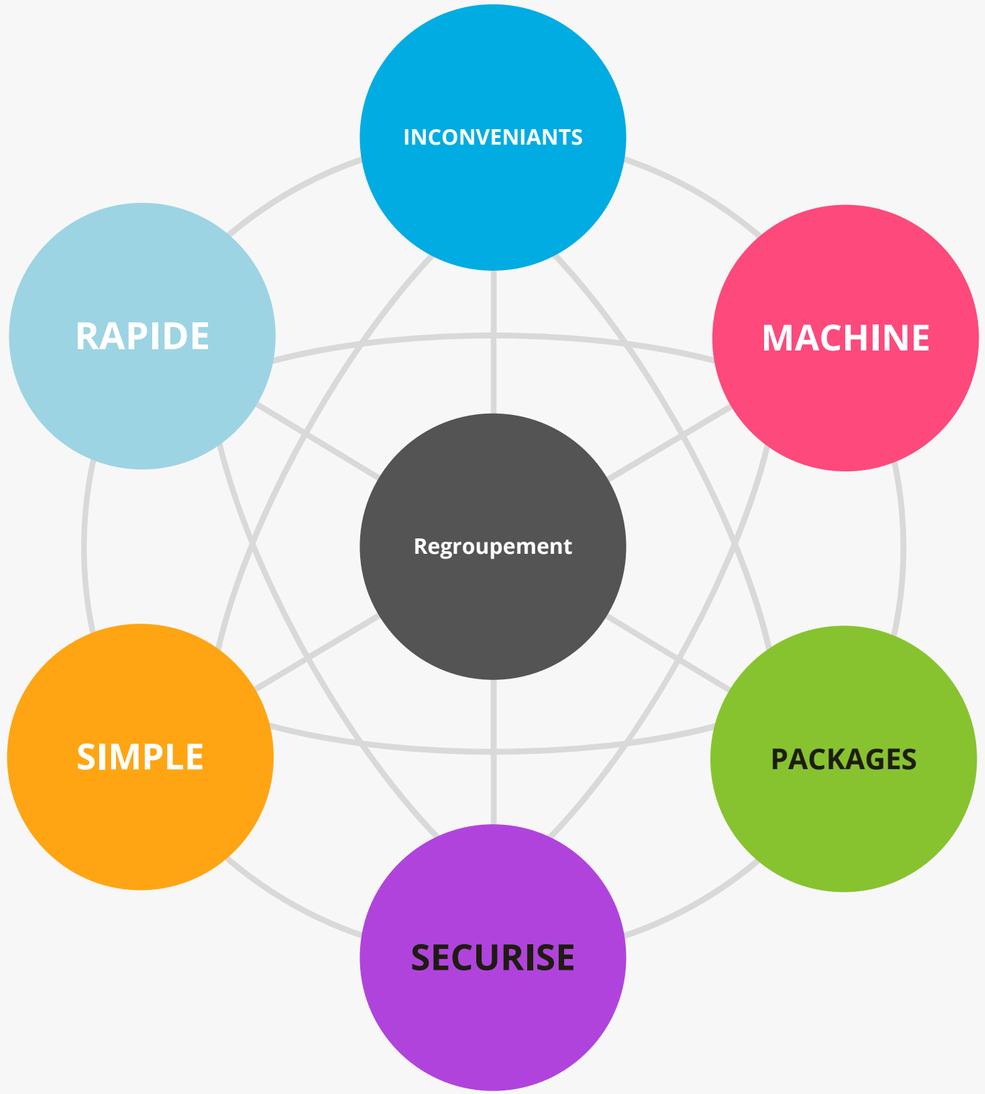
- Multiples façons de faire
- Debug délicat
- PC qui crash
- Contraintes techniques / budgétaires

SageMaker

Rapide

Simple

Sécurisé



Inconveniants

Version de spark fixe

Machine

Pas besoin de local

Packages

Peu à installer

Réflexion

Optimisation

- Etudier les couts et besoins
- Optimisation algorithme et données
- Monitoring
- Versions des librairies

Interrogation

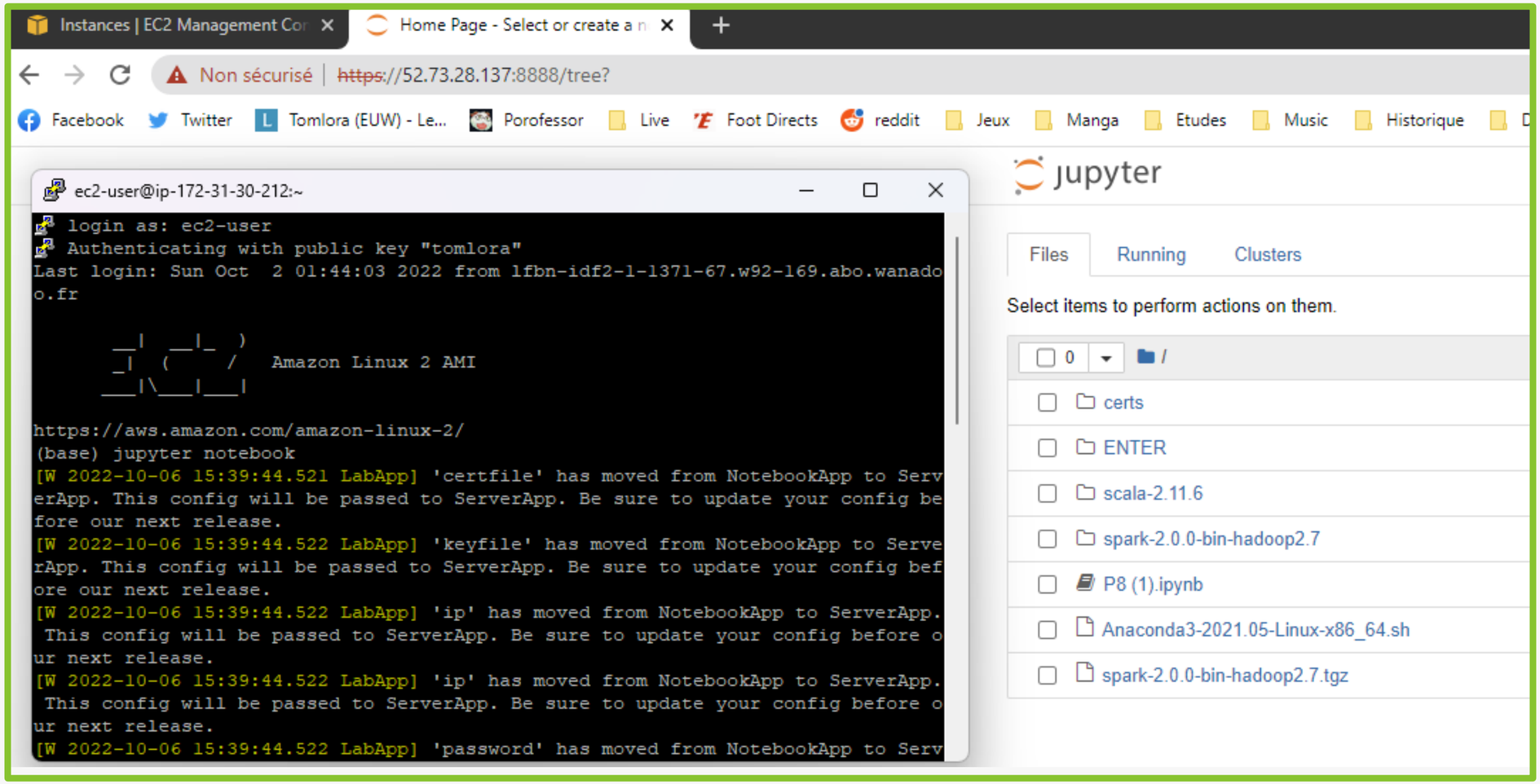
- Image différente
- Maturité du fruit
- Pathologie

4

ANNEXES

Instance

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	Adresses IP I...	Surveillance	Nom du groupe de s...	Nom de clé
–	i-02d30f73ef00cf9ad	Arrêté(e)	t2.medium	–	Aucune al...	us-east-1b	ec2-52-73-28-137.com...	52.73.28.137	52.73.28.137	–	disabled	Notebook,launch-wizar...	tomlora



Stockage (Bucket)