

OPENCLASSROOMS



Kevin

Parcours Data Scientist

Problématique

L'entreprise

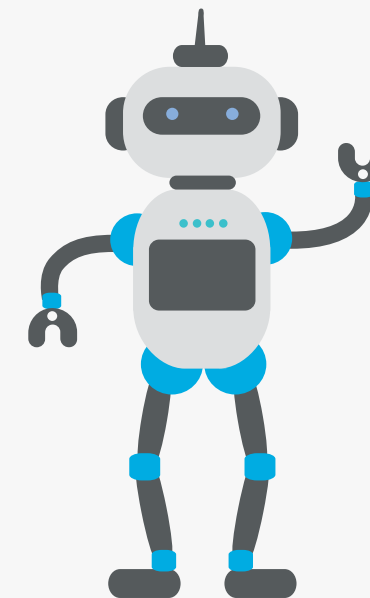
- Start-up AgriTech
- Proposition innovante pour la récolte de fruits

Mission

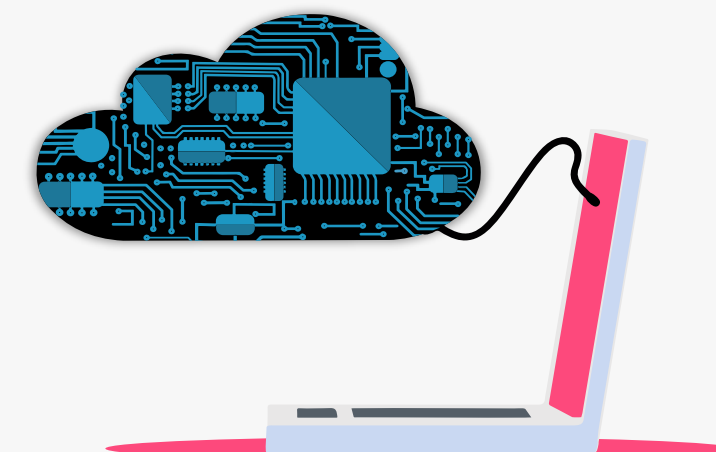
- Application mobile
- Algorithme de classification

Contraintes

- PySpark
- Architecture Big Data EC2

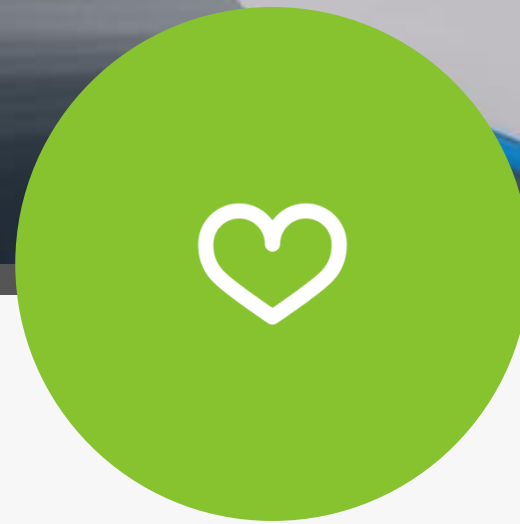


Fruits!

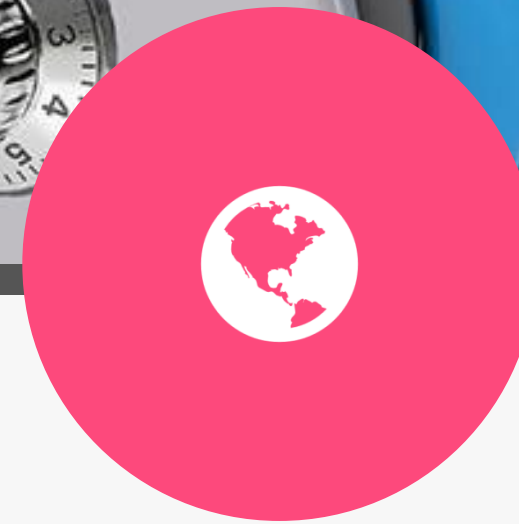




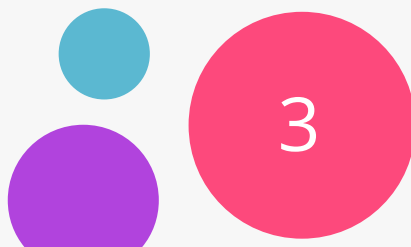
Jeu de données



Process



Conclusion



1



Jeu de données

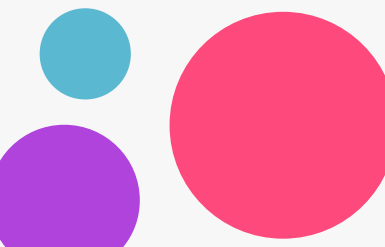
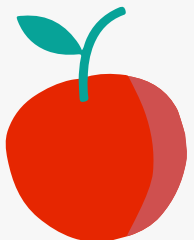
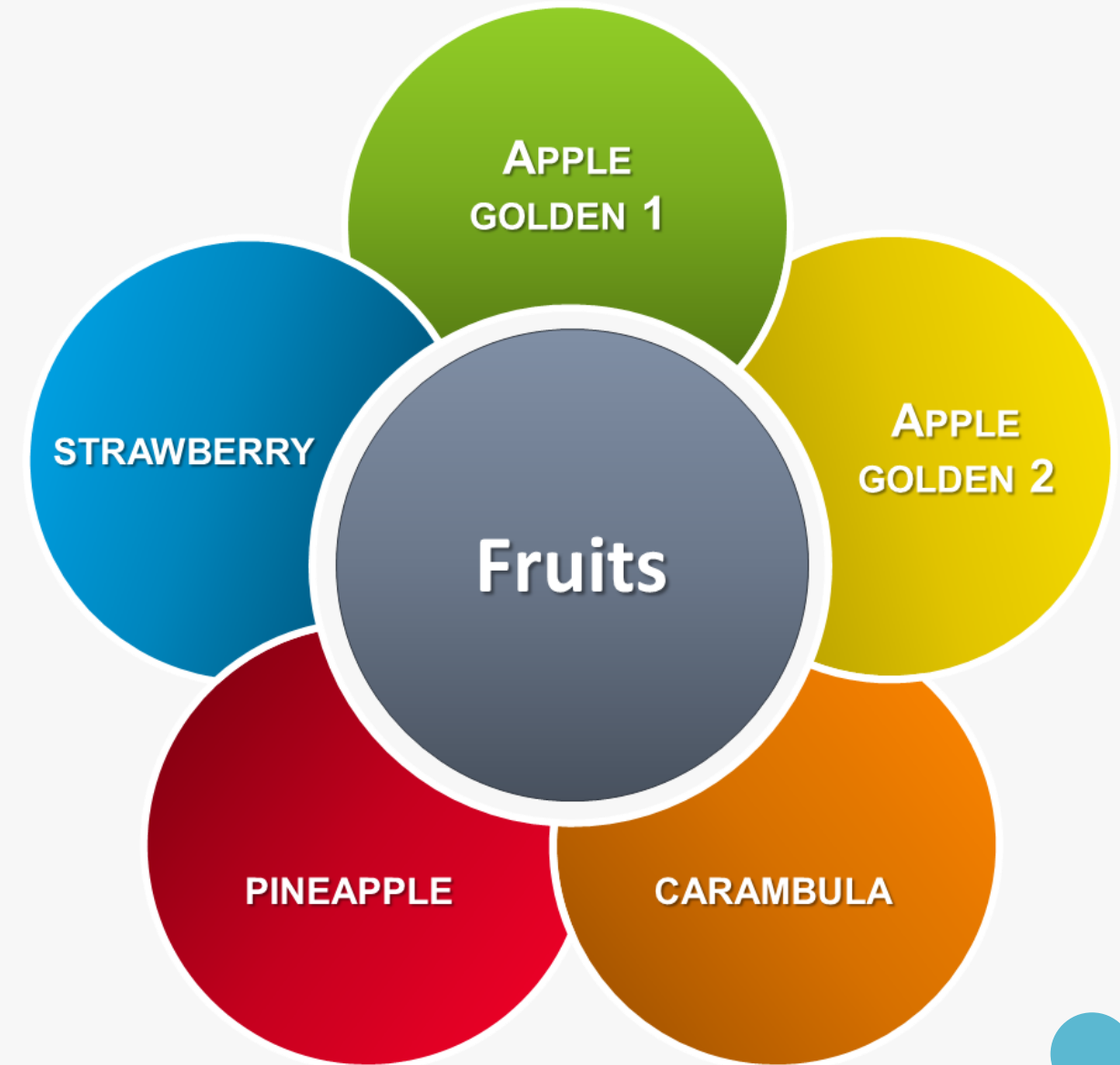
Informations générales

Dataset

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



Détail



Dataset

Format

Dimension 100x100

Format JPG

Couleurs

Fond et Focus

Centré sur le fruit

Fond blanc

Angles

Timelapse

Rotation 3 axes

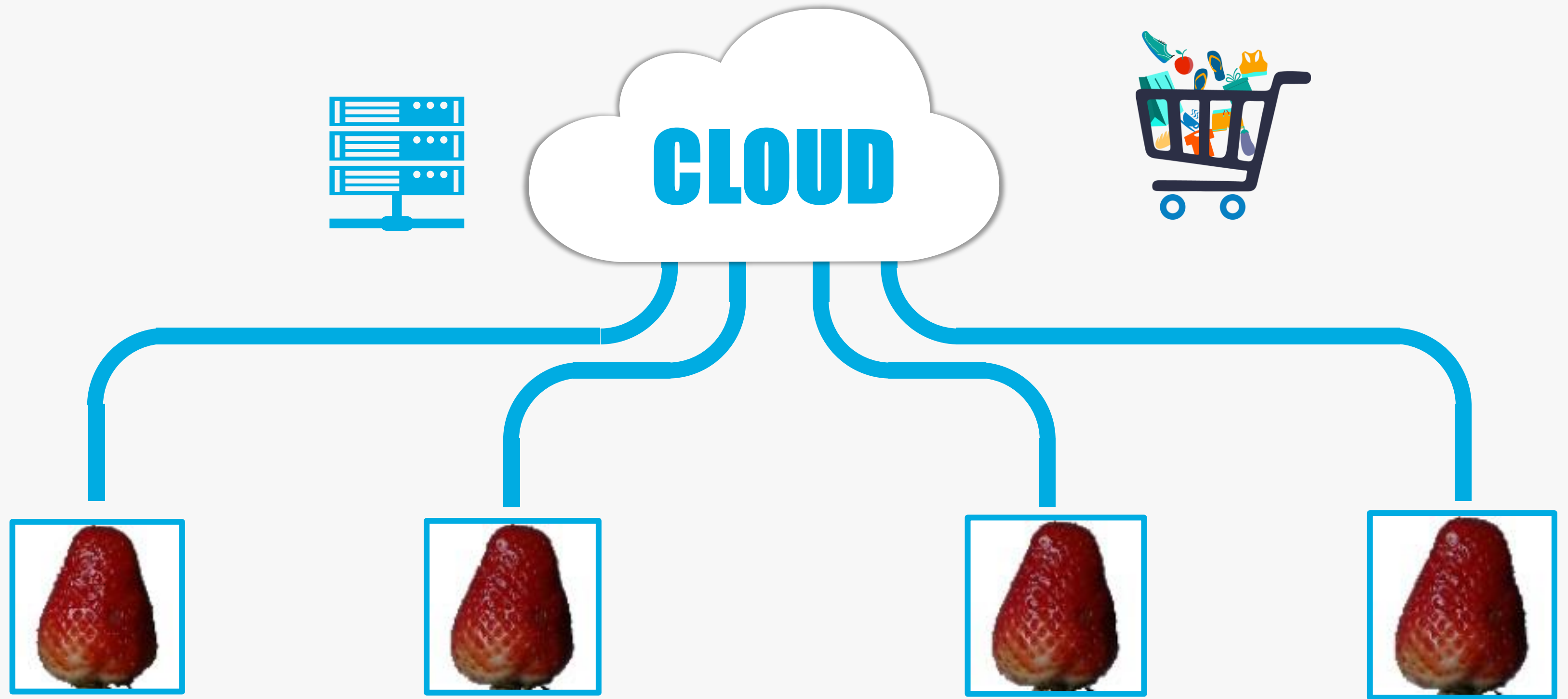
Informations

400+ images

Caractéristiques



Dataset restraint



Le Big Data



PySpark

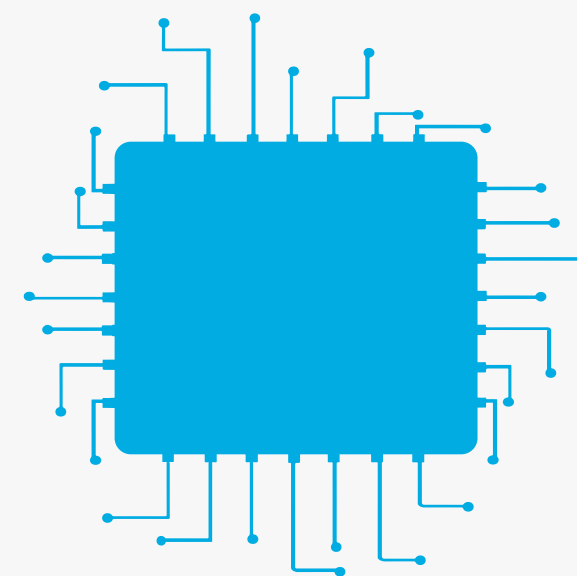


Big Data

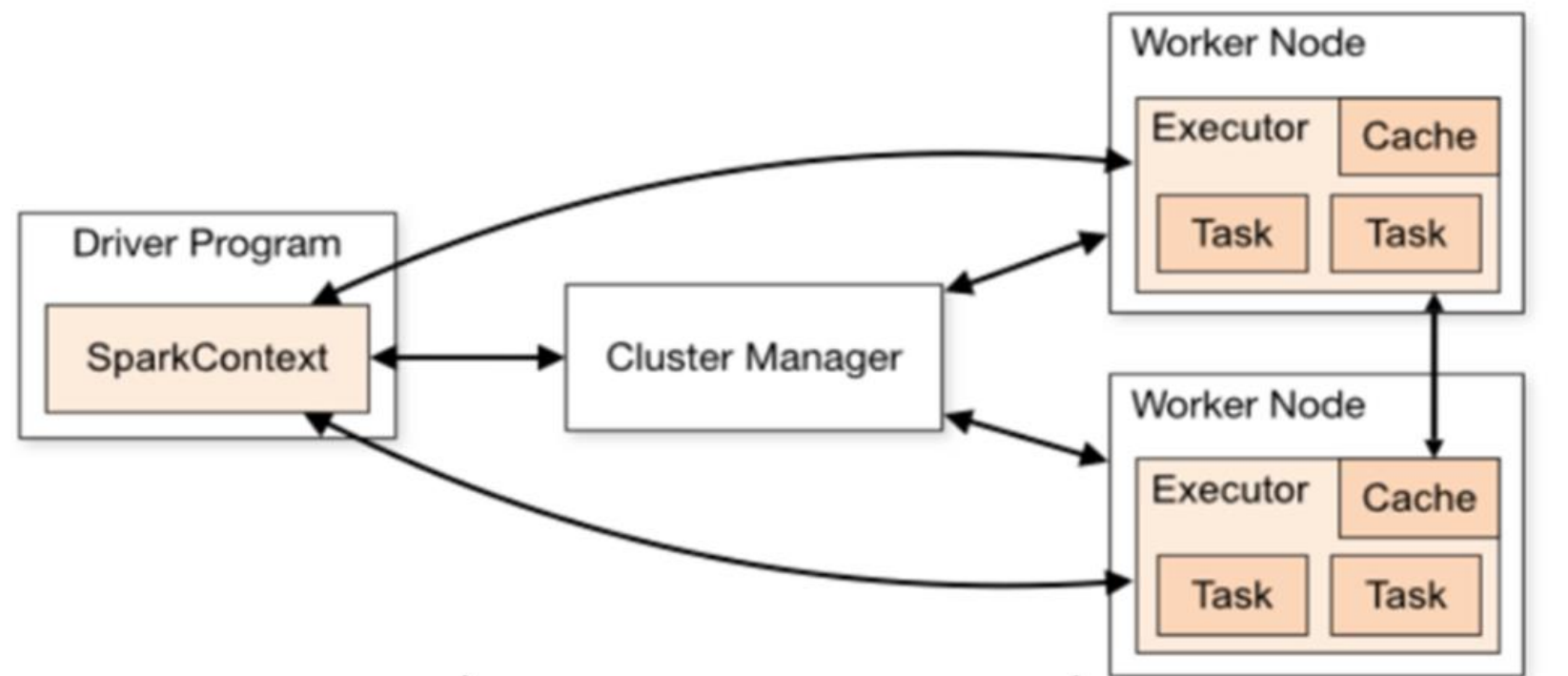
Proches de Panda

DataFrame

 python™ + 
=

PySpark



Application maître :
Configuration /
Initialisation
Agrégation des calculs

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

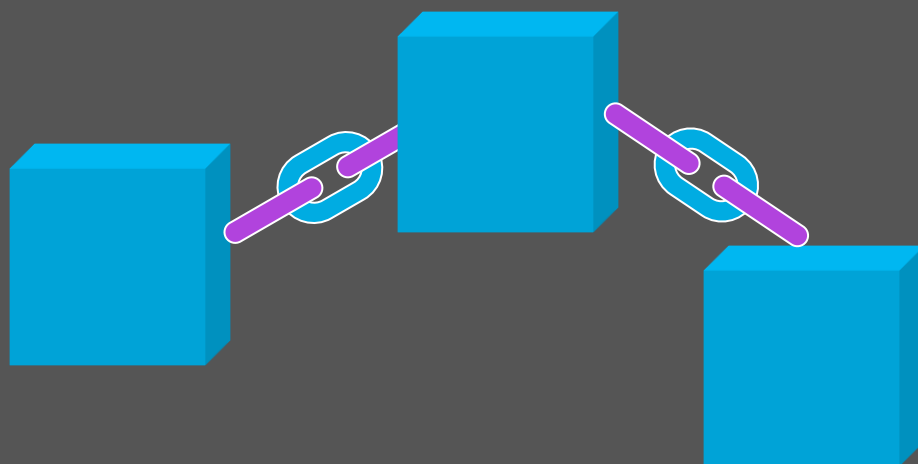
Workers :
Exécution des tâches
en parallèle

Adapté au Big Data

Division des opérations

Stockage

Agrégation

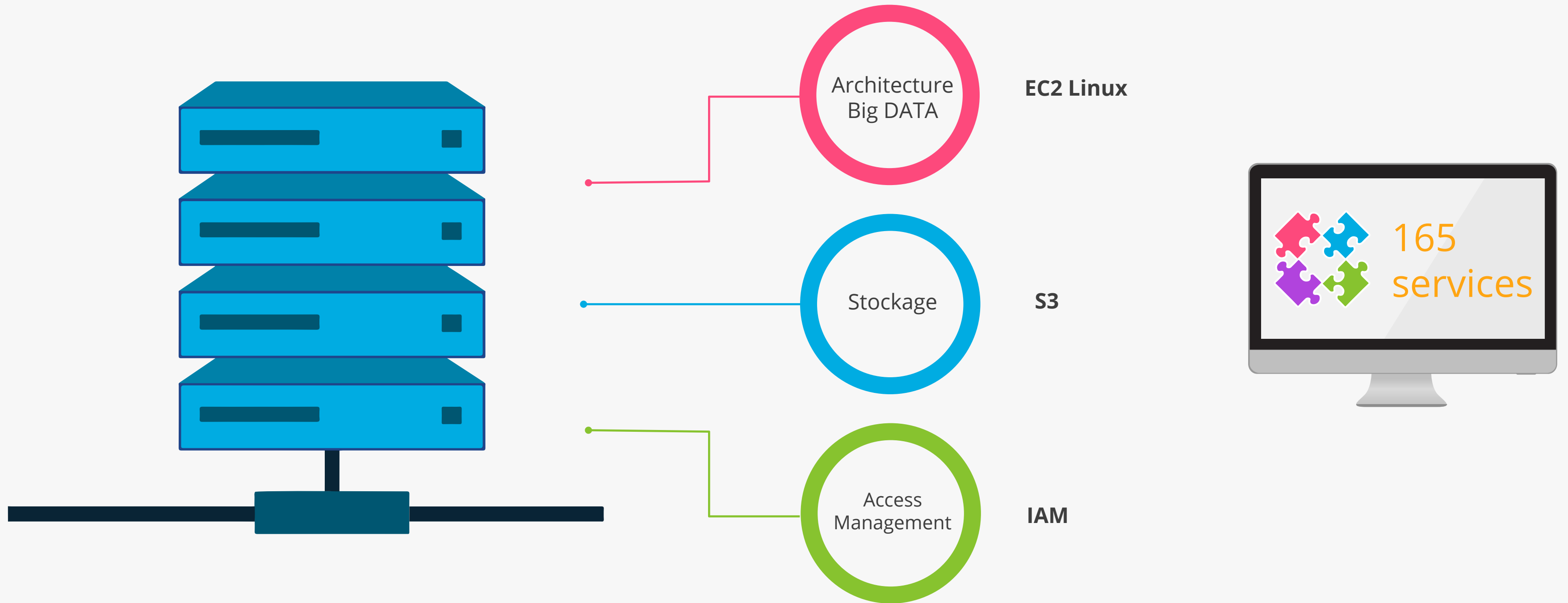


2

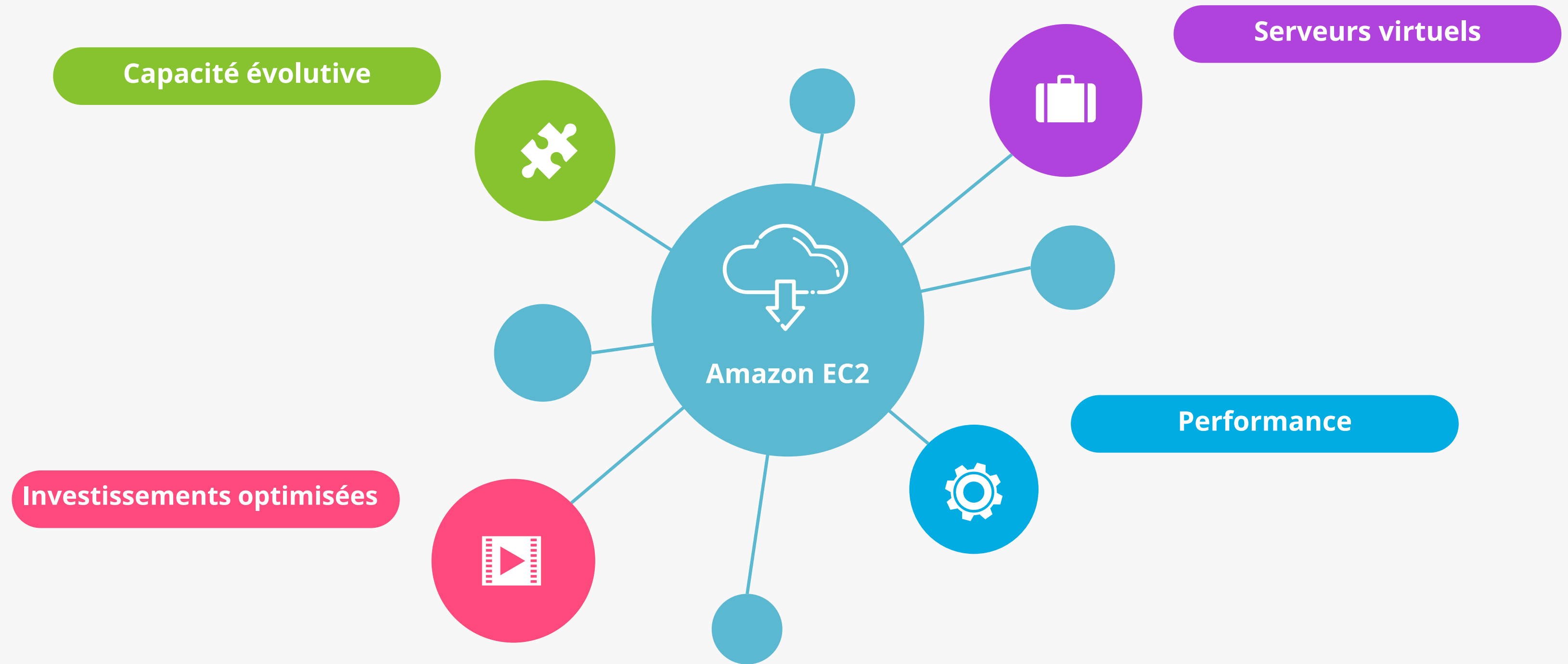


Process

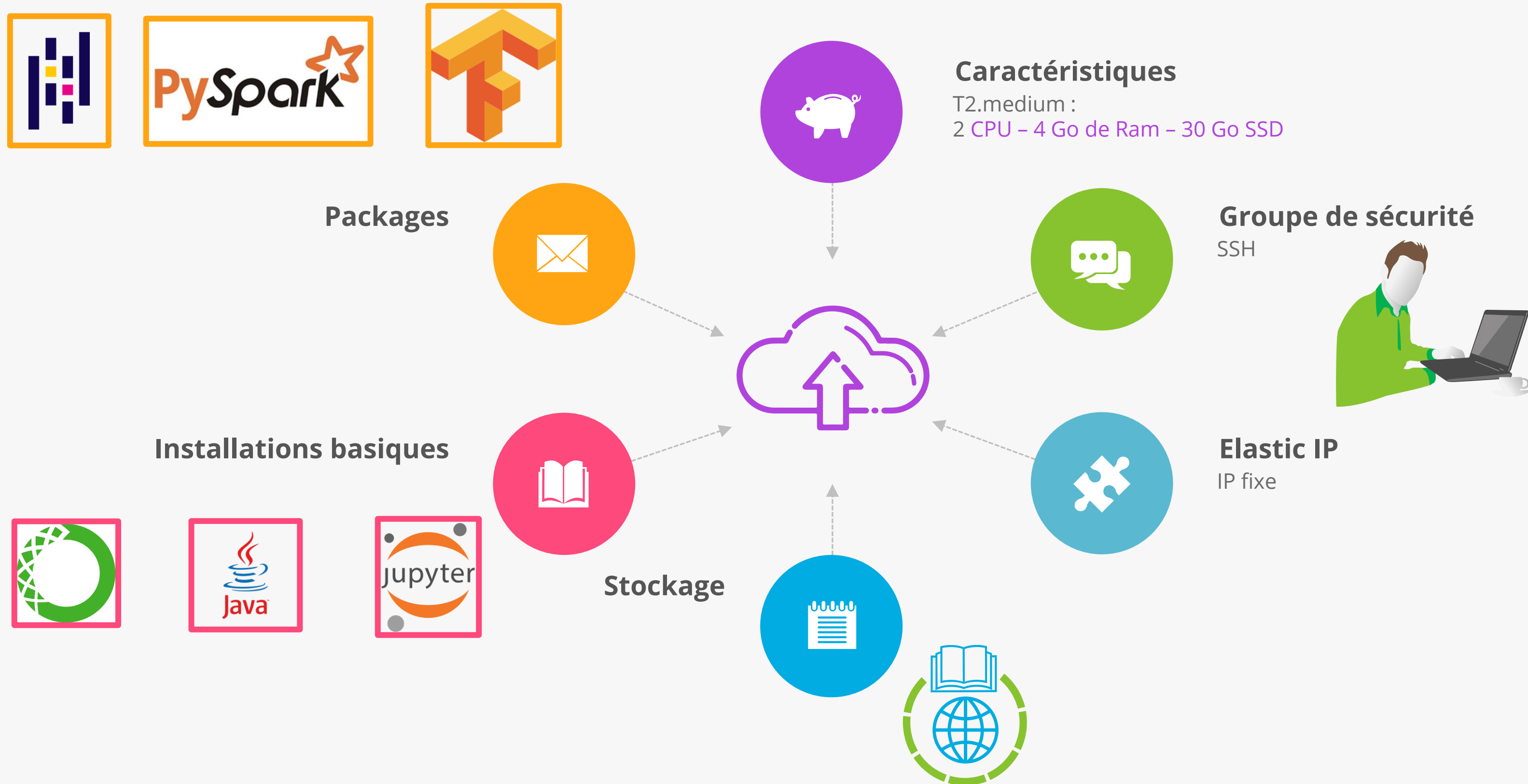
Cloud AWS



EC2



Instances



Utilisation de l'instance

3 étapes



Configurer l'accès web

Groupe de sécurité

Configurer Jupyter

Accès distants

Lancer le notebook

Téléverser le notebook

01

02

03



Stockage : **Amazon S3**



Avantages

Durable

Sécurité

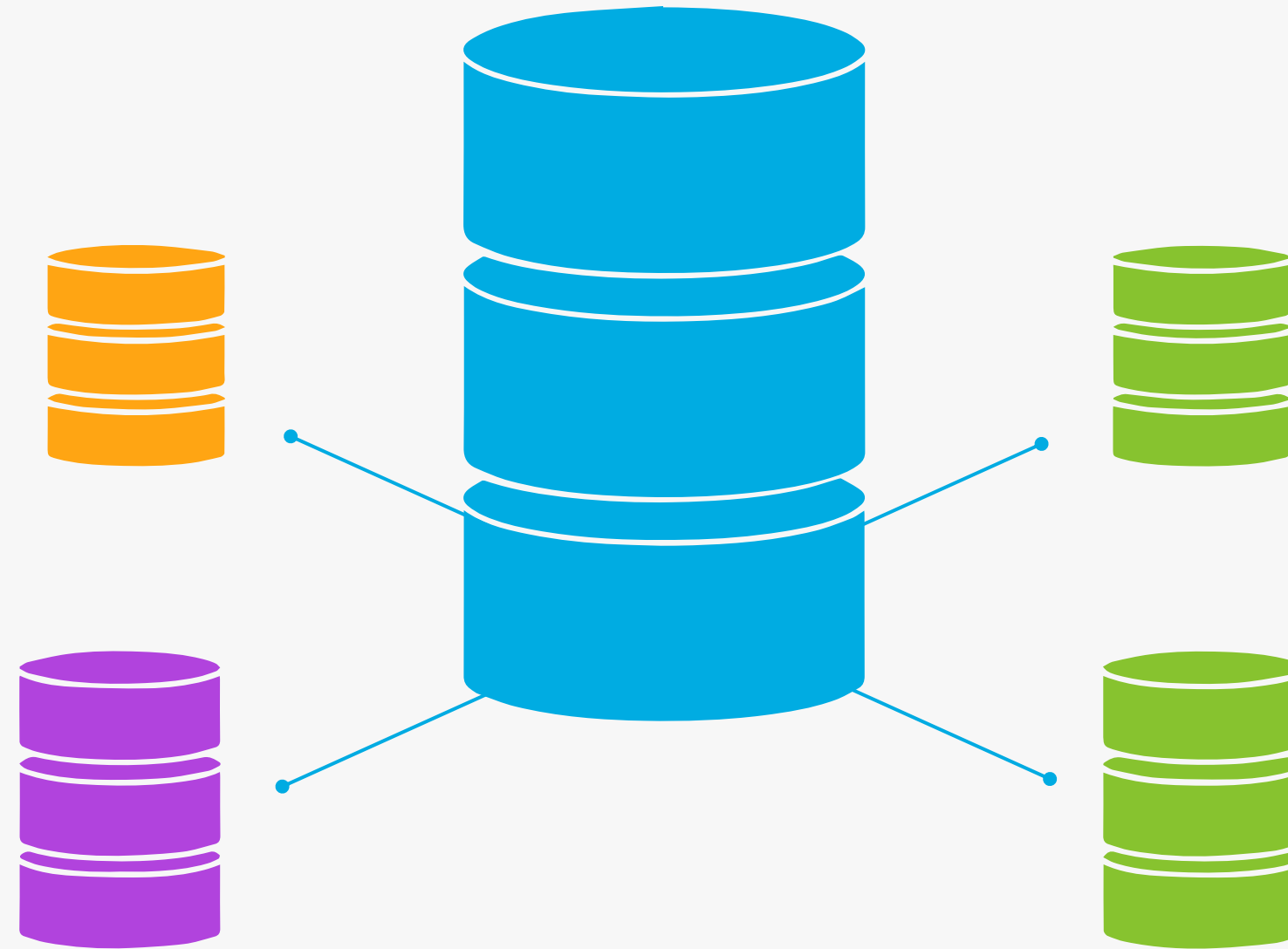
Simple

Rapide



IAM

Management des accès aux
services et ressources



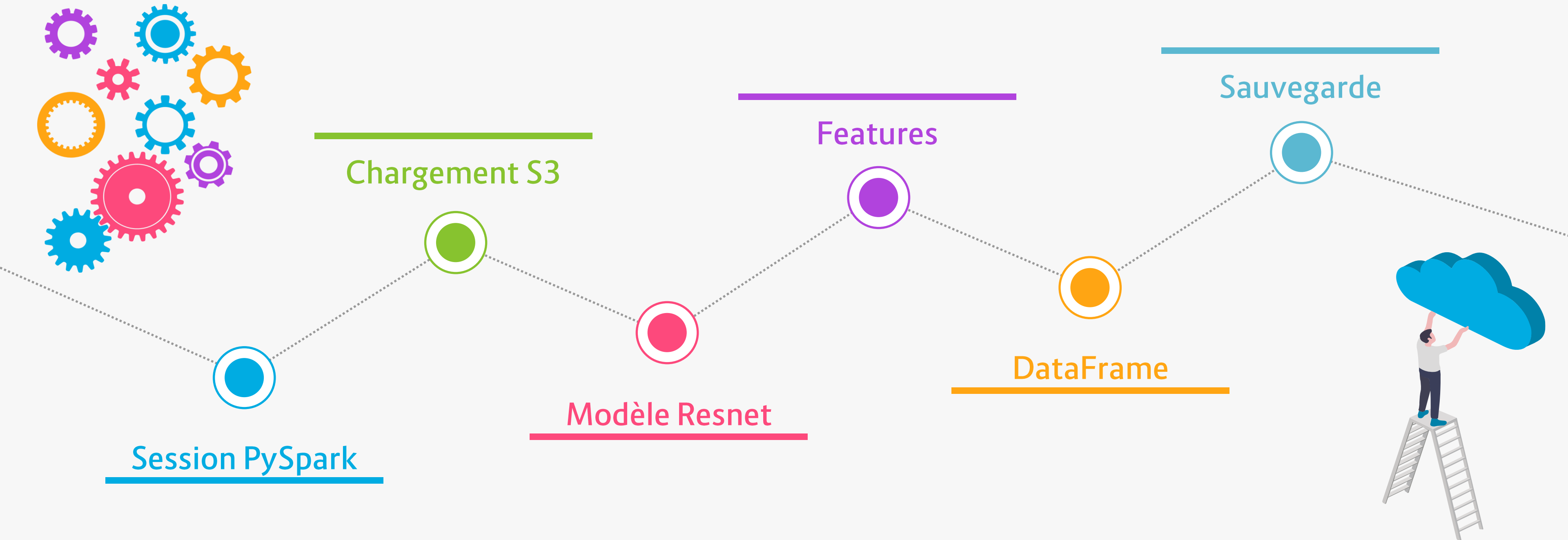
Upload du dataset

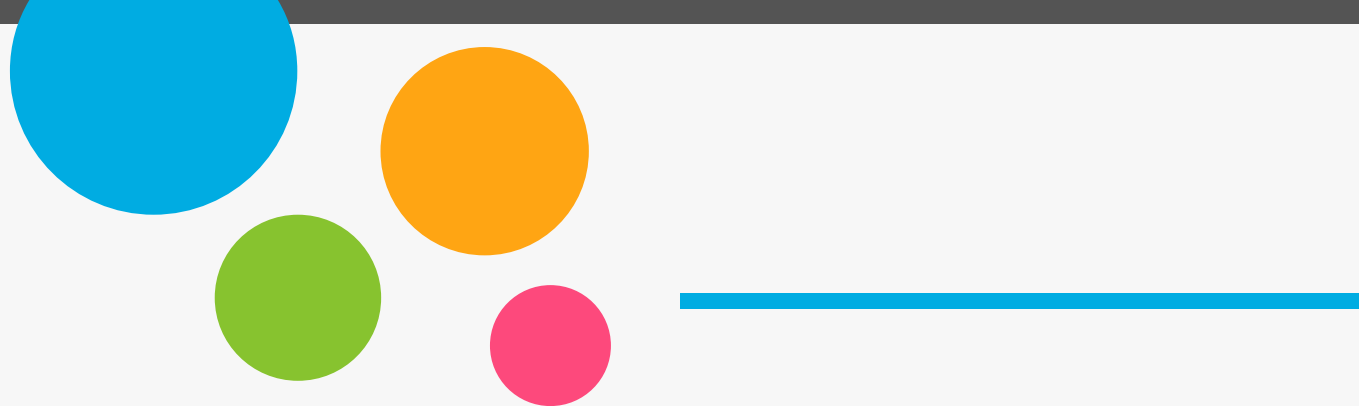


Création d'un bucket

Process Cloud

6 steps





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-entraînés

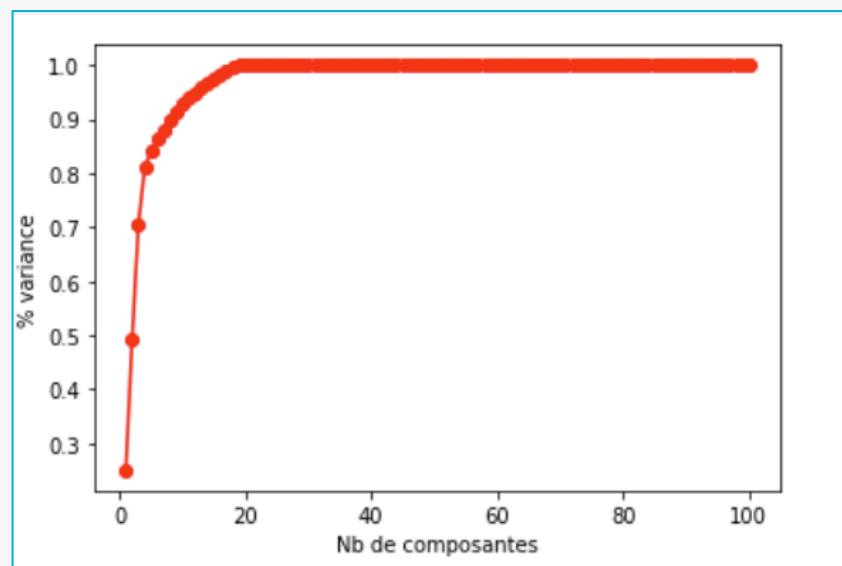
Dernières couches retirées

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

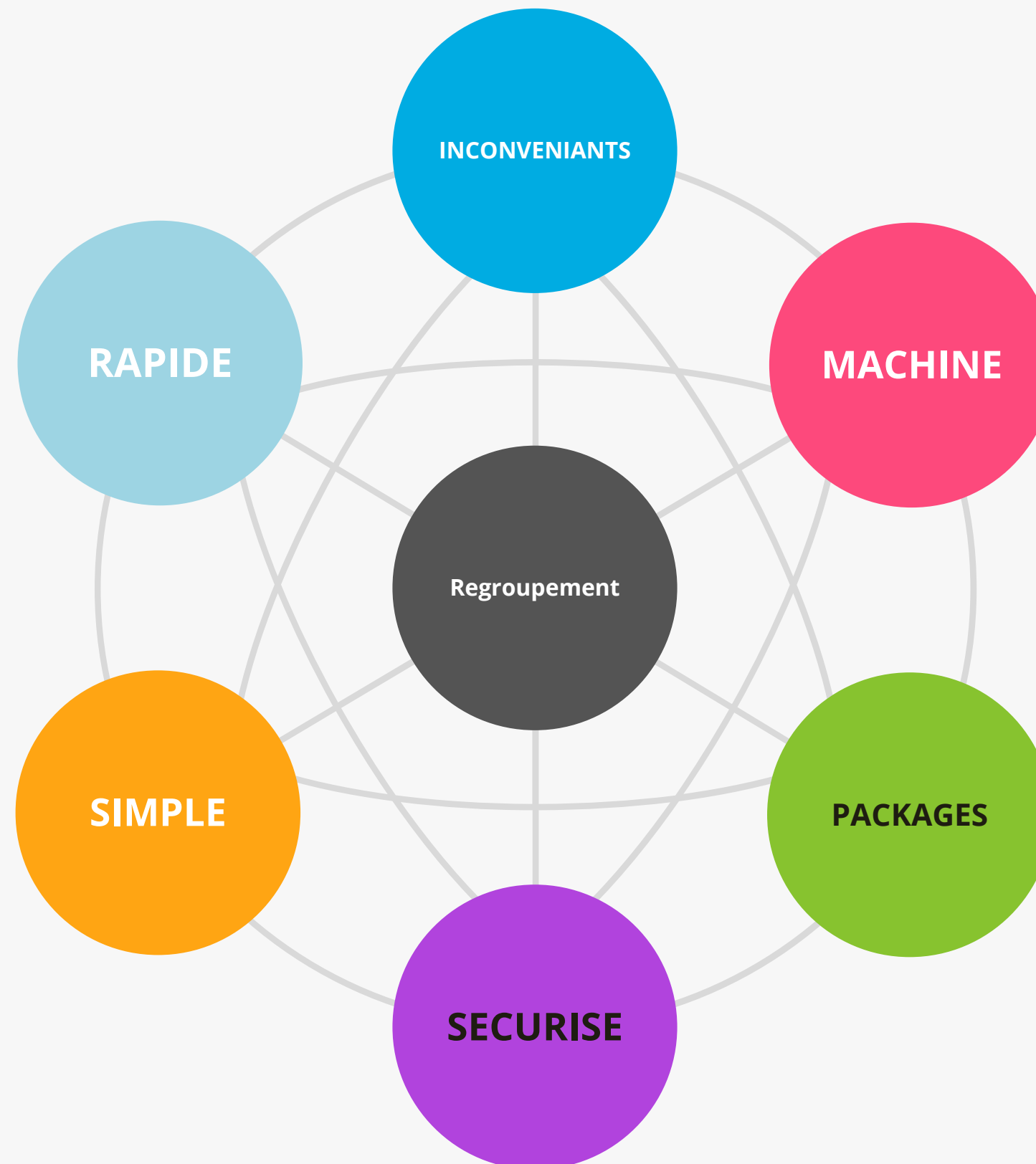


SageMaker

Rapide

Simple

Sécurisé



Inconvenients

Version de spark fixe

Machine

Pas besoin de local

Packages

Peu à installer

Réflexion

Optimisation

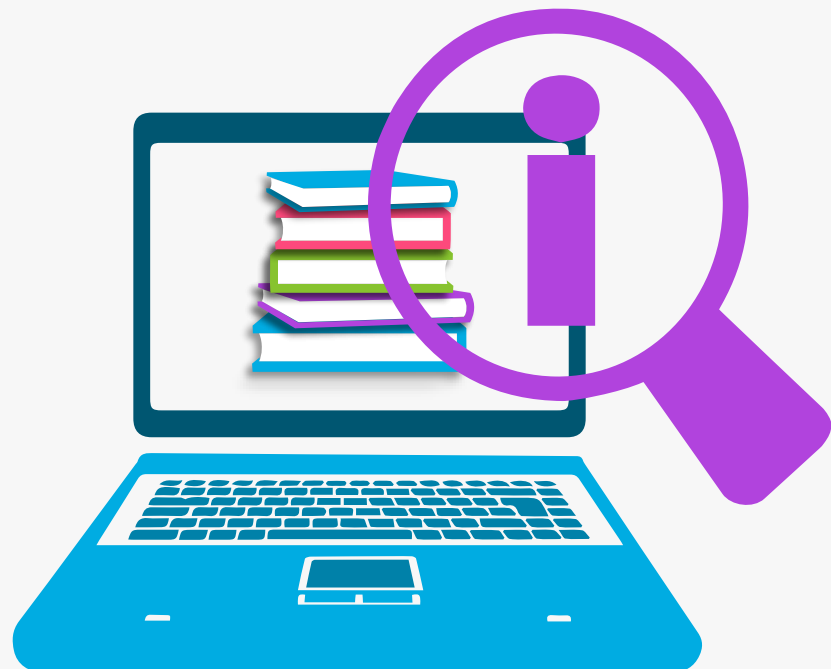


Etudier les couts et besoins

Optimisation algorithme et données

Monitoring

Versions des librairies



Interrogation

Image différente

Maturité du fruit

Pathologie





ANNEXES

4

Instance

<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	Adresses IP I...	Surveillance	Nom du groupe de s...	Nom de clé
<input checked="" type="checkbox"/>	-	i-02d30f73ef00cf9ad	Arrêté(e)	t2.medium	-	Aucune alar...	us-east-1b	ec2-52-73-28-137.com...	52.73.28.137	52.73.28.137	-	disabled	Notebook,launch-wizar...	tomlora

Instances | EC2 Management Console

Home Page - Select or create a new instance

Non sécurisé | https://52.73.28.137:8888/tree?

Facebook Twitter Tomlora (EUW) - Le... Porofessor Live Foot Directs reddit Jeux Manga Etudes Music Historique

ec2-user@ip-172-31-30-212:~

login as: ec2-user
Authenticating with public key "tomlora"
Last login: Sun Oct 2 01:44:03 2022 from lfbn-idf2-1-1371-67.w92-169.abo.wanadoo.fr

Amazon Linux 2 AMI

https://aws.amazon.com/amazon-linux-2/
(base) jupyter notebook
[W 2022-10-06 15:39:44.521 LabApp] 'certfile' has moved from NotebookApp to ServerApp. This config will be passed to ServerApp. Be sure to update your config before our next release.
[W 2022-10-06 15:39:44.522 LabApp] 'keyfile' has moved from NotebookApp to ServerApp. This config will be passed to ServerApp. Be sure to update your config before our next release.
[W 2022-10-06 15:39:44.522 LabApp] 'ip' has moved from NotebookApp to ServerApp. This config will be passed to ServerApp. Be sure to update your config before our next release.
[W 2022-10-06 15:39:44.522 LabApp] 'ip' has moved from NotebookApp to ServerApp. This config will be passed to ServerApp. Be sure to update your config before our next release.
[W 2022-10-06 15:39:44.522 LabApp] 'password' has moved from NotebookApp to ServerApp. This config will be passed to ServerApp. Be sure to update your config before our next release.

jupyter

Files Running Clusters

Select items to perform actions on them.

☐ 0 /

☐ certs

☐ ENTER

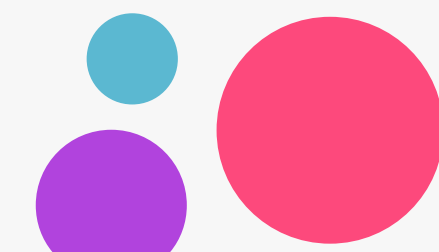
☐ scala-2.11.6

☐ spark-2.0.0-bin-hadoop2.7

☐ P8 (1).ipynb

☐ Anaconda3-2021.05-Linux-x86_64.sh

☐ spark-2.0.0-bin-hadoop2.7.tgz





Stockage (Bucket)

Compartiments (1) [Info](#)

Les compartiments sont des conteneurs pour les données stockées dans S3. [En savoir plus](#)

↻

Copier l'ARN

Vider

Supprimer

Créer un compartiment

Rechercher des compartiments par nom

< 1 > ⚙

	Nom ▲	Région AWS ▼	Accéder ▼	Date de création ▼
<input type="radio"/>	tomlora	USA Est (Virginie du Nord) us-east-1	Les objets peuvent être publics	30 Sep 2022 12:03:47 AM CEST

Objets (2)

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](#)

↻

Copier l'URI S3

Copier l'URL

Télécharger

Ouvrir

Supprimer

Actions ▼

Créer un dossier

Charger

Rechercher des objets en fonction du préfixe

< 1 > ⚙

<input type="checkbox"/>	Nom ▲	Type ▼	Dernière modification ▼	Taille ▼	Classe de stockage ▼
<input type="checkbox"/>	📁 result/	Dossier	-	-	-
<input type="checkbox"/>	📁 test/	Dossier	-	-	-

