

# Deployer un model dans le cloud

# Contents

1. Introduction
2. Problématique et jeu de données
3. Création de l'environnement Big Data
4. Traitement des images dans un environnement Big Data
5. Conclusion



# Introduction

Nous sommes appelés en tant que Data Scientist pour un projet initié par la Start Up “Fruits” dans le domaine de l’Agritech.

L’objectif de l’entreprise est de promouvoir la biodiversité des Fruits en développant des robots cueilleurs intelligents.



# II. Problématique et jeux de données



## A. PROBLEMATIQUE

- Mettre en place une application mobile capable de donner des informations pertinentes sur un fruit en prenant une photo.
- Développer un moteur de classification des images de fruits en intégrant une architecture Big Data.

## B. MISSION

Utiliser comme point de départ le notebook que l'alternant a mis en place, se l'approprier et compléter le travail par une réduction de dimension.

Mettre en place un script Pyspark.

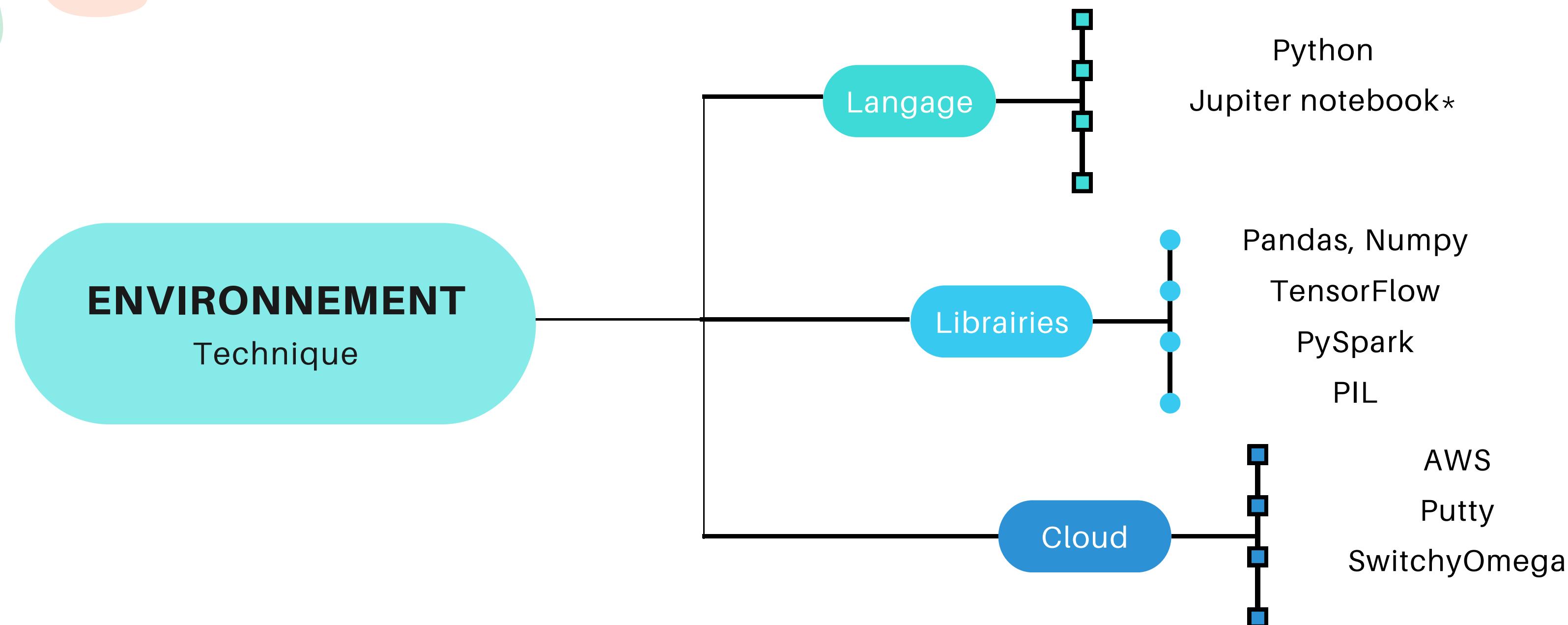
# D. Problématique et jeux de données

## JEU DE DONNEES

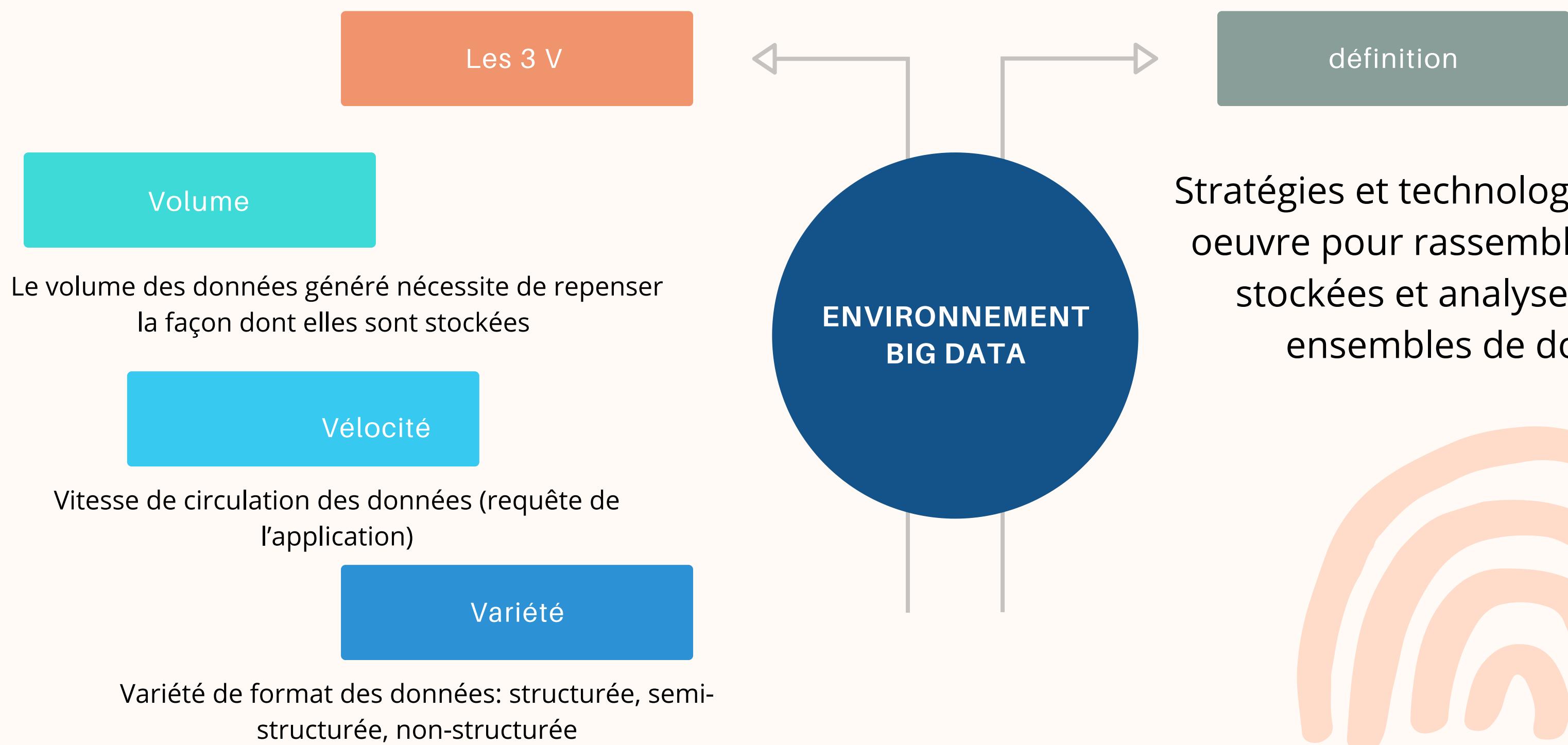
- Jeux de données issu de Fruits 360 Kaggle.
- 90423 images de fruits.
- 131 classes: Apple golden, banana, Kiwi...
- un répertoire par classe avec plusieurs photos du même fruit pris sous différents angles.
- un jeu de données entraînement, validation et test.
- taille des images 100\*100 pixels.
- images en couleur sur fond blanc uniformisé



# M. Création de l'environnement Big Data

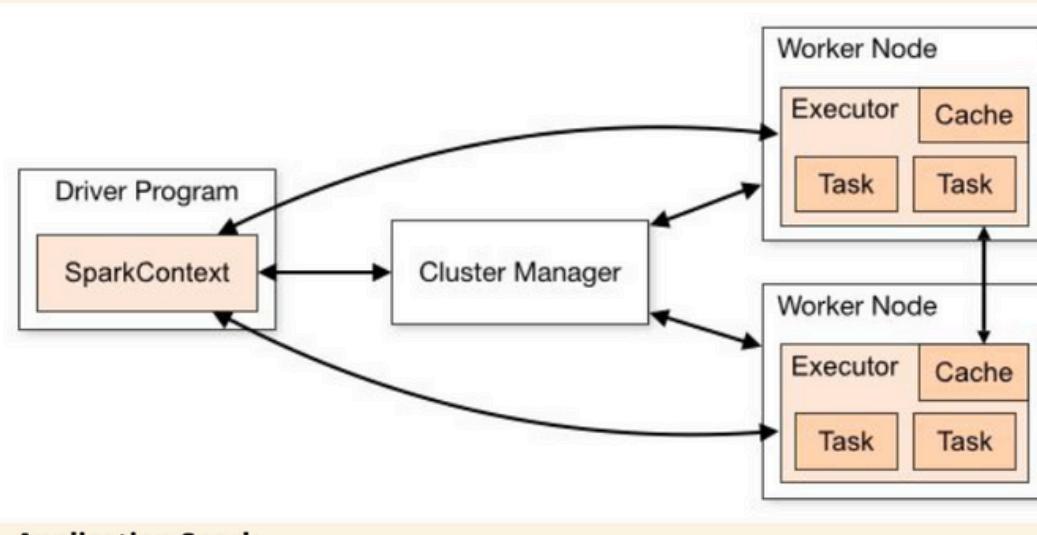


# M. Création de l'environnement Big Data

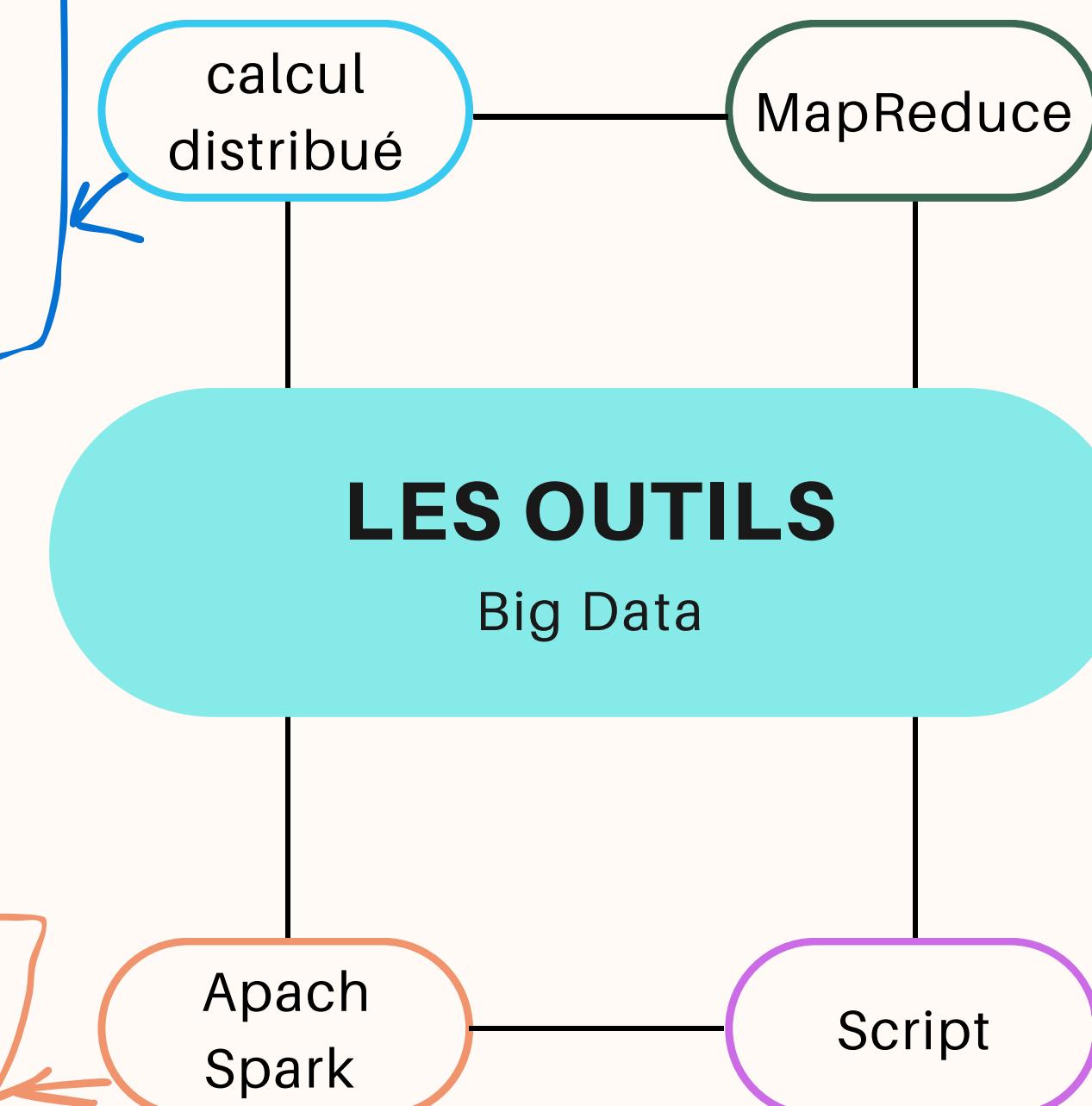


# M. Création de l'environnement Big Data

Création de plusieurs unités de calcul réparti en Cluster au profit d'un seul projet, le but étant de réduire le temps de calcul lors d'une requête



Framework opensource capable de traiter des données massives en utilisant un calcul distribué



Algorithme Mapreduce  
-utilisé pour le traitement distribué sur un volume important de donné.  
-capable de diviser les données en de plus ensembles repartis en cluster (MAP), superviser les calculs sur les petits ensembles et faire une agrégation et synthétiser les résultats (REDUCE)

Script PySpark afin de communiquer avec Spark

# M. Création de l'environnement Big Data

## LE PRESTATAIRE CLOUD : AWS

### S3: SIMPLE STORAGE SERVICE

Service de stockage et de distribution de données, on l'utilisera pour stocker nos images



### EC2: ELASTIC COMPUTE CLOUD

Gérer des serveurs sous forme de machines virtuelles, rôle d'ordinateur dans lequel on va pourvoir installer tous nos logiciels: anaconda, jupyterhub, Spark, Tensorflow...

### UTILISATEUR IAM

Contrôle des accès  
Connexion via le protocole SSH à l'instance EC2

### EMR

sous service de EC2

# M. Création de l'environnement Big Data

IAM

S3

Cluster avec EMR

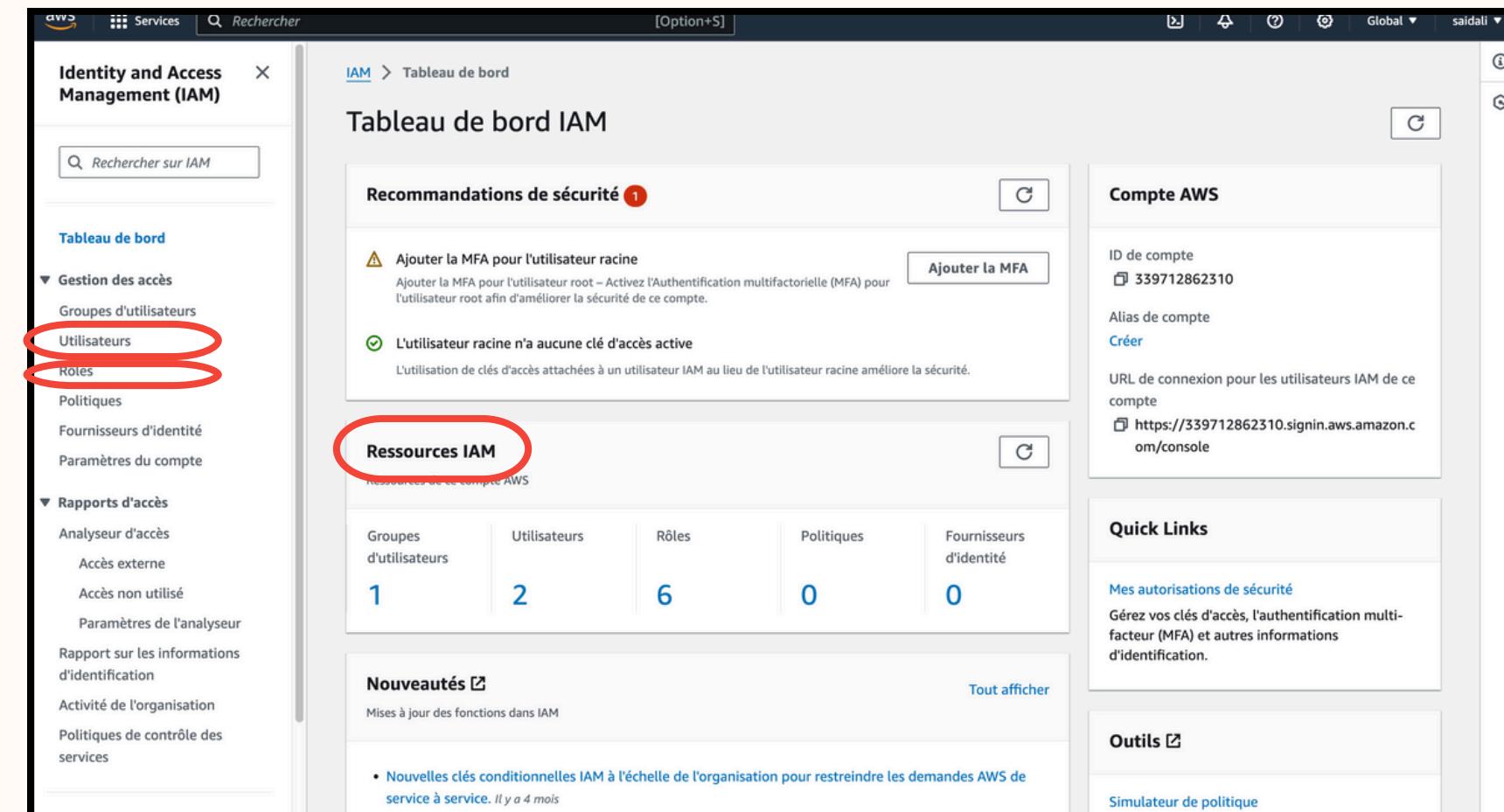
logiciel

materiel

bootstrapping

sécurité

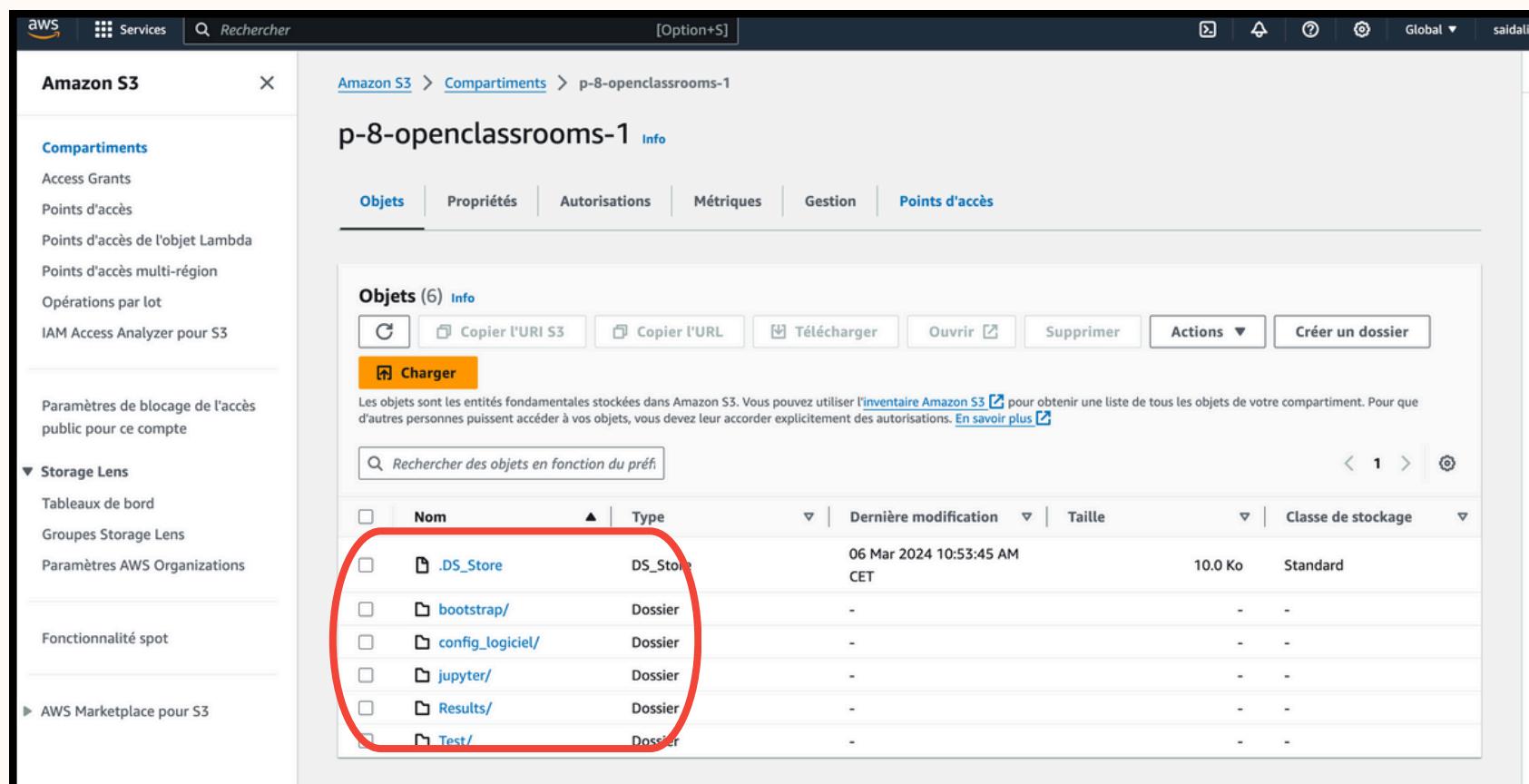
SSH



- IAM: Identity and Access Management
  - création d'utilisateurs
  - contrôle des accès des différents services
  - création d'une paire de clé SSH permettant un accès sans authentification: login et mdp
- AWS CLI
  - utilisation du terminal de commande pour un accès au services AWS

# M. Création de l'environnement Big Data

IAM      S3      Cluster avec EMR      logiciel      matériel      bootstrapping      sécurité      SSH



## S3: Simple Storage Service

- Stockage d'une grande variété de fichies
- Indépendant de EC2
- Accès aux données très rapide
- Possibilité de chiffrer les données

## S3: Simple Storage Service

- fichier bootstrap
- config logiciel
- jupyterhub
- dossier Test
- résultats

# M. Création de l'environnement Big Data

IAM            S3            **Cluster avec EMR**            logiciel            matériel            bootstrapping            sécurité            SSH



## EMR: Elastic MapReduce

Un service qui permet d'exécuter un traitement de données distribuées à grande échelle en utilisant des librairies comme Hadoop et Spark

Utilisation d'instance EC2



Avantages: Flexibilité et Gestion simplifiée

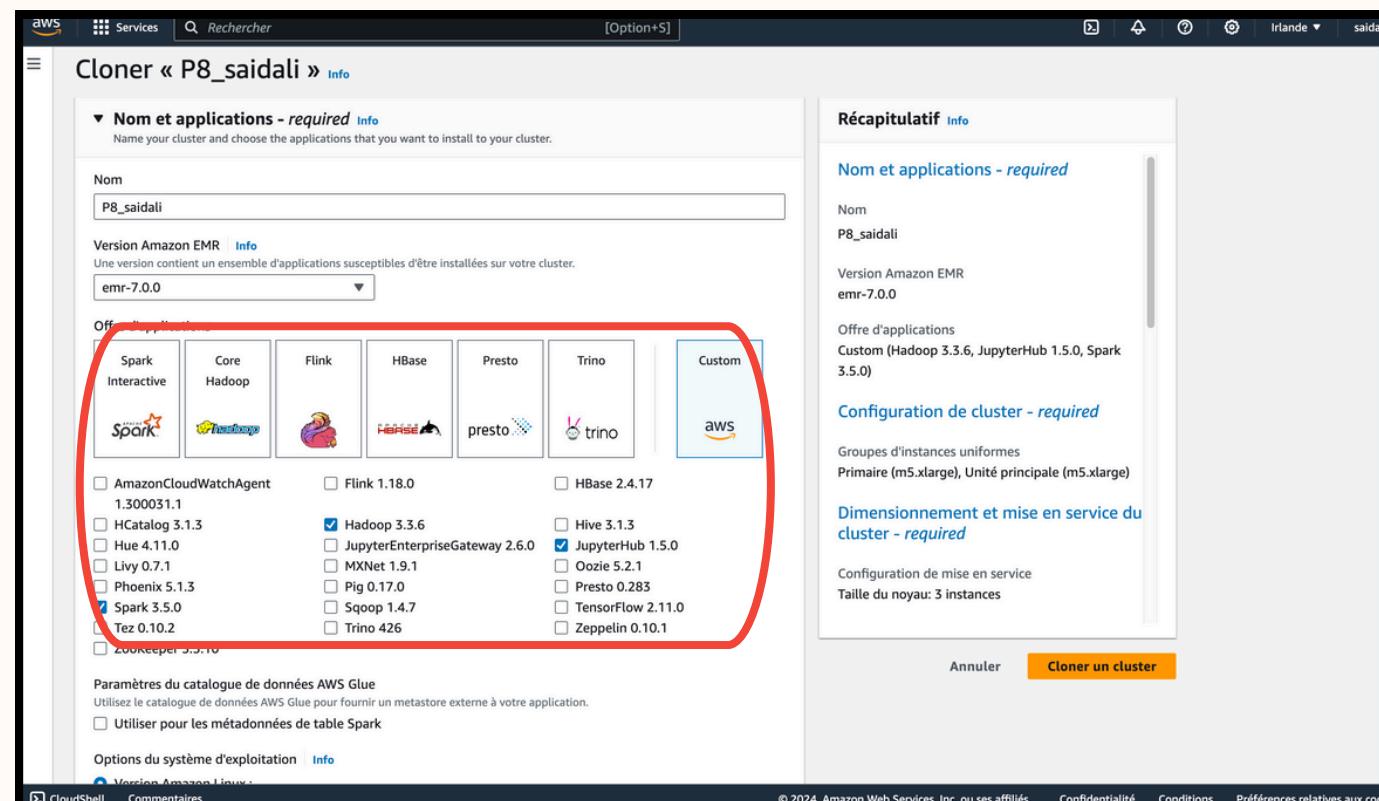
Service entièrement géré par AWS

### Création du Cluster EMR en 4 étapes :

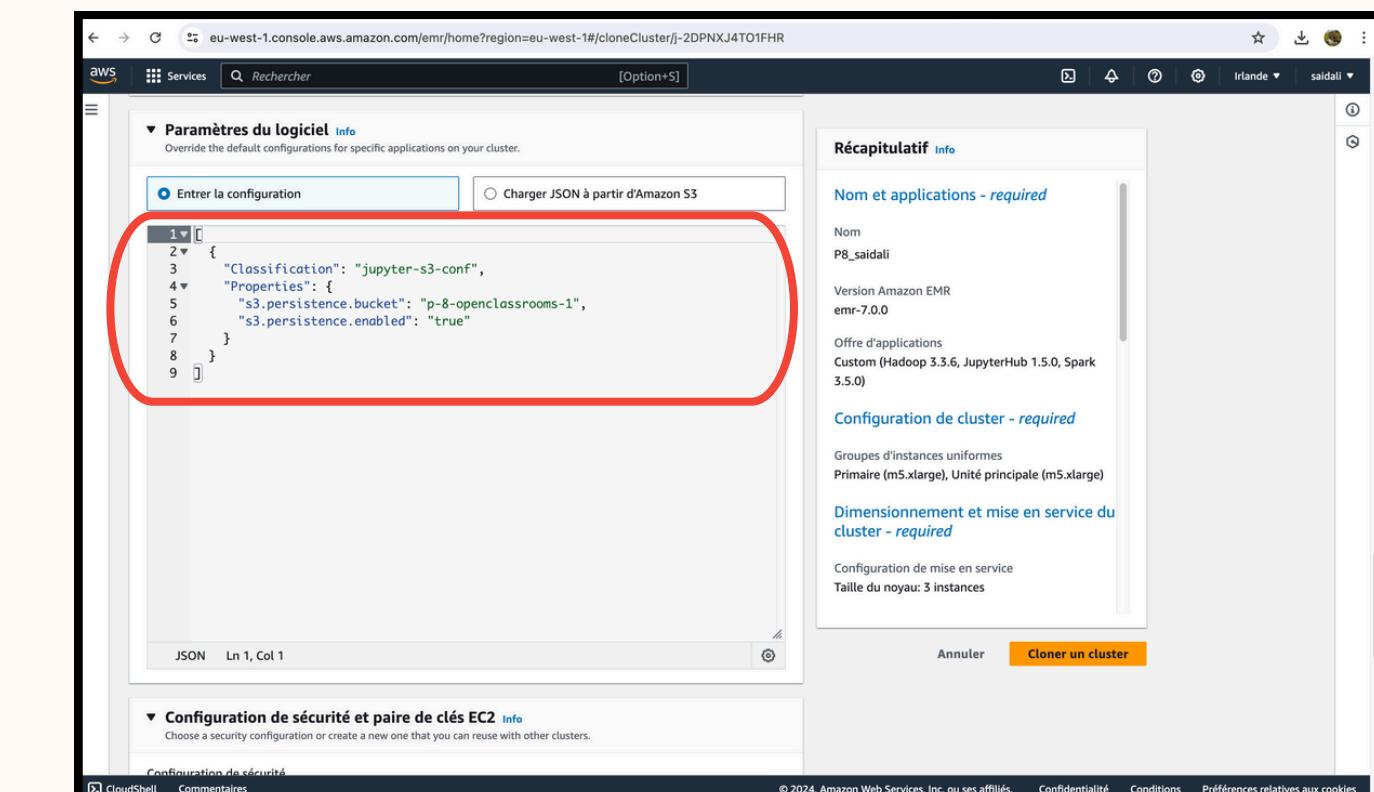
1. config logiciel
2. config matériel
3. action d'amorçage
4. sécurité

# M. Création de l'environnement Big Data

IAM      S3      Cluster avec EMR      logiciel      matériel      bootstrapping      sécurité      SSH



**Paramétrage de la persistance:**  
au niveau des notebooks créés et ouvert via JupyterHub  
Config logiciel et fichier JSON



## Les logiciels

- Spark et Hadoop: Pour le calcul distribué
- JupiterHub: Exécution des scripts PySpark du notebook

# M. Création de l'environnement Big Data

IAM      S3      Cluster avec EMR      logiciel      matériel      bootstrapping      sécurité      SSH



The screenshot shows the AWS EMR configuration interface. On the left, under 'Configuration de cluster - required', there are two options: 'Groupes d'instances uniformes' (selected) and 'Flottes d'instances flexibles'. The 'Groupes d'instances uniformes' section is highlighted with a red box. It shows a 'Primaire' group with an 'm5.xlarge' instance selected. Below it is a checkbox for 'Utiliser la haute disponibilité'. Under 'Unité principale', another 'm5.xlarge' instance is selected. On the right, the 'Récapitulatif' section shows the cluster configuration: 'Nom et applications - required' (Nom: P8\_saidali, Version Amazon EMR: emr-7.0.0, Offre d'applications: Custom (Hadoop 3.3.6, JupyterHub 1.5.0, Spark 3.5.0)), 'Configuration de cluster - required' (Groupes d'instances uniformes: Primaire (m5.xlarge), Unité principale (m5.xlarge)), and 'Dimensionnement et mise en service du cluster - required' (Taille du noyau: 3 instances). A yellow 'Cloner un cluster' button is at the bottom.

## Config Matériel: Les Instances

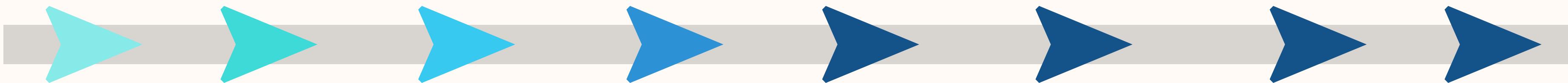
1 instance Maitre: **Driver**

3 instances principales: **Worker**

Instance de type **M5**, ce sont des instances qui équilibrées et **xlarge** (maîtrise des couts)

# M. Création de l'environnement Big Data

IAM      S3      Cluster avec EMR      logiciel      matériel      bootstrapping      sécurité      SSH



**Configuration de cluster - required**

- Groupes d'instances uniformes**: Choisissez le même type d'instance EC2 et la même option d'achat (à la demande ou Spot) pour tous les nœuds de votre groupe de nœuds.
- Flettes d'instances flexibles**: Choisissez parmi la plus grande variété d'options de provisionnement pour les instances EC2 de votre cluster. Diversifiez les types d'instances et les options d'achat, et utilisez une stratégie d'allocation.

**Groupes d'instances uniformes**

**Primaire**: Choisir un type d'instance EC2: m5.xlarge (4 vCore, 16 GiB mémoire, EBS uniquement stockage).  
 Utiliser la haute disponibilité: Lancez des clusters hautement disponibles et plus résilients avec trois nœuds primaires sur des instances à la demande. Cette configuration s'applique pendant toute la durée de vie de votre cluster.

**Unité principale**: Choisir un type d'instance EC2: m5.xlarge (4 vCore, 16 GiB mémoire, EBS uniquement stockage).

**Récapitulatif**

Nom: P8\_saidali  
 Version Amazon EMR: emr-7.0.0  
 Offre d'applications: Custom (Hadoop 3.3.6, JupyterHub 1.5.0, Spark 3.5.0)

**Configuration de cluster - required**

Groupes d'instances uniformes: Primaire (m5.xlarge), Unité principale (m5.xlarge)

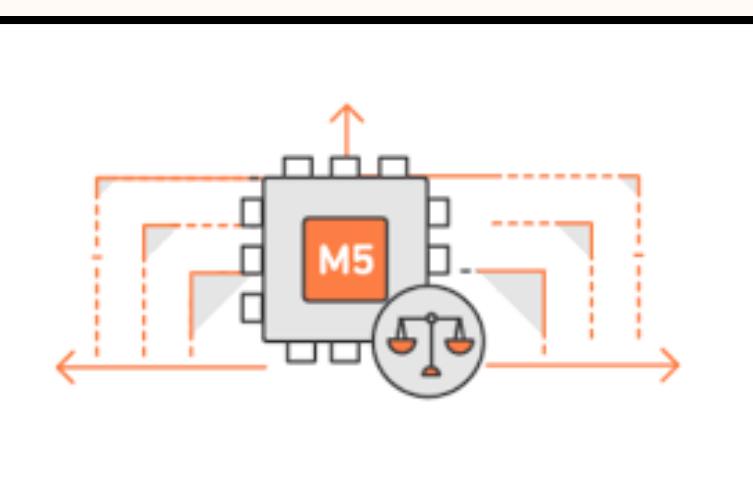
**Dimensionnement et mise en service du cluster - required**

Configuration de mise en service: Taille du noyau: 3 instances

## Config Matériel: Les Instances

1 instance Maitre: **Driver**  
 3 instances principales: **Worker**

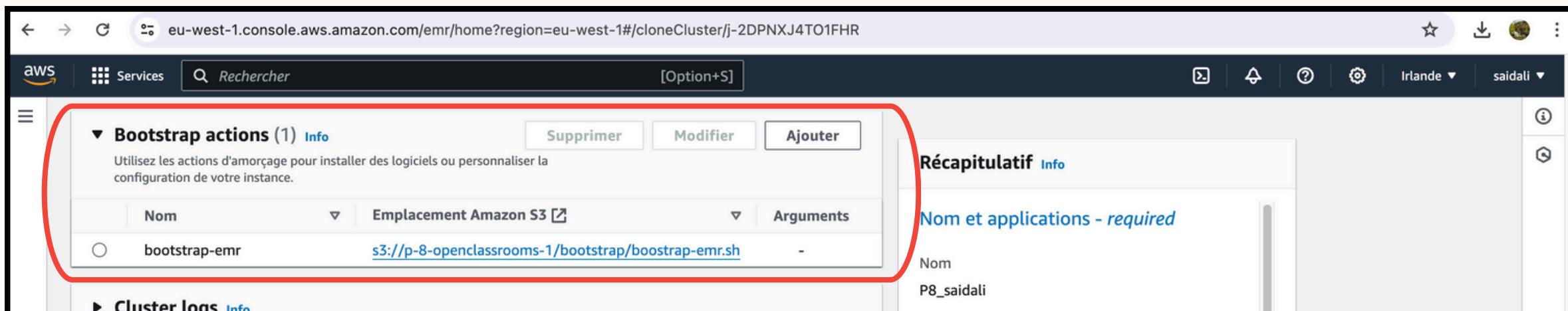
Instance de type **M5**, ce sont des instances qui équilibrées et **xlarge** (maîtrise des couts)



vCPU: 4/mémoire 16 GiO

# M. Création de l'environnement Big Data

IAM      S3      Cluster avec EMR      logiciel      matériel      **bootstrapping**      sécurité      SSH



## Bootstrapping ou Actions d'amorçage

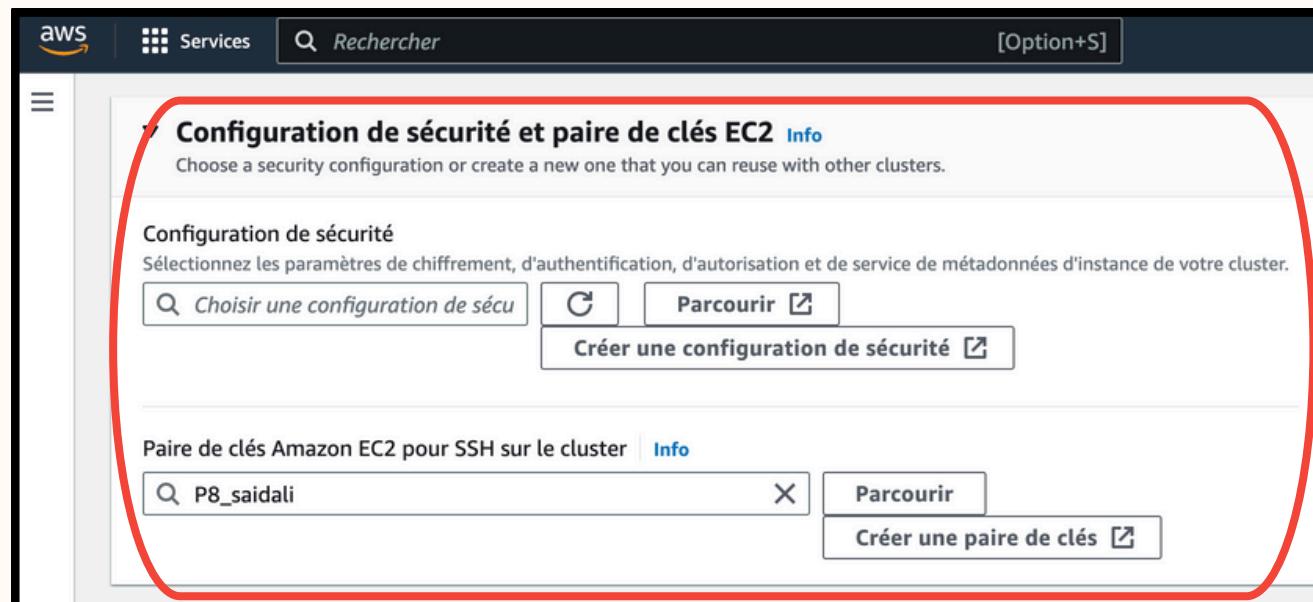
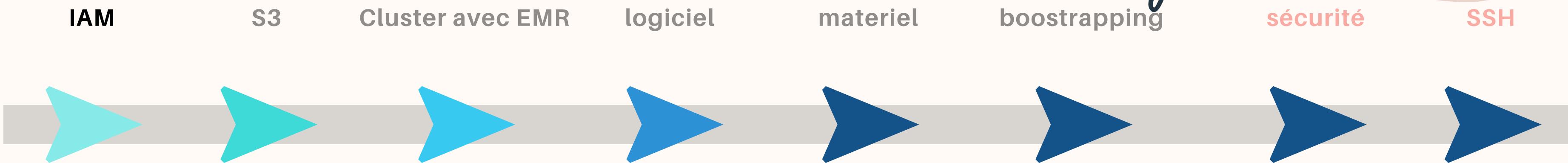
Installer les librairies et les framework manquantes pour le Script PySpark

Installation des librairies à l'initialisation des des serveurs --> présentes sur l'ensemble des machines.

```
projet_8 > $ bootstrap-emr.sh
1 #!/bin/bash
2 sudo python3 -m pip install -U setuptools
3 sudo python3 -m pip install -U pip
4 sudo python3 -m pip install wheel
5 sudo python3 -m pip install pillow
6 sudo python3 -m pip install pyarrow
7 sudo python3 -m pip install boto3
8 sudo python3 -m pip install pandas==1.2.5
9 sudo python3 -m pip install s3fs
10 sudo python3 -m pip install fsspec
11 sudo python3 -m pip install tensorflow
```

fichier bootstrap-emr stocké dans S3

# M. Création de l'environnement Big Data



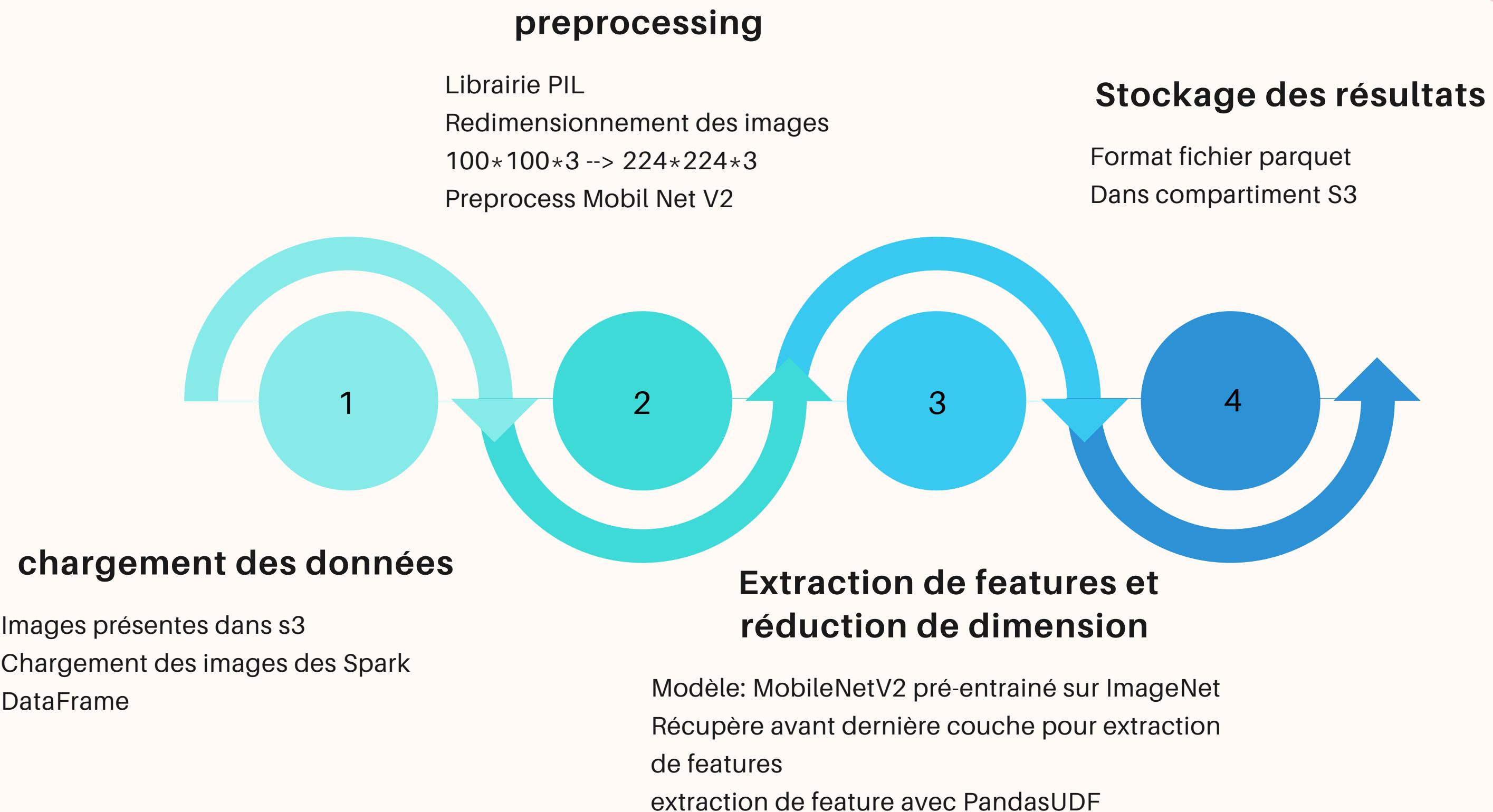
# Clé SSH

# Création d'un tunnel SSH pour connexion au driver

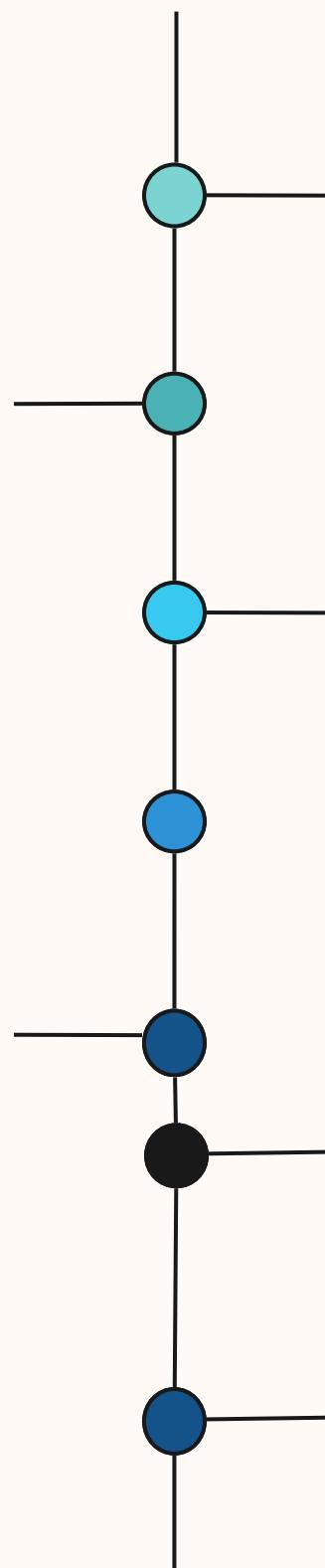
# Clé SSH

Permet de se connecter aux instances EC2 sans avoir besoin d'authentification avec login et mdp

# N. Traitement des images dans un environnement Big Data et dans le Cloud



# N. Traitement des images dans un environnement Big Data et dans le Cloud



# MobileNetV2

## Modele CNN:Réseau de Neurones Convolutifs

- pré-entraîné sur la base ImageNet
  - Détection de features et classification d'images
  - Rapidité d'exécution et faible dimensionnalité du vecteur en sortie

# Diffusion des poids

Avec `SparkContext.broadcast()` de PySpark

- Chargement du modèle sur le driver et diffusion des poids sur les workers
  - Distribuer des variables à travers le cluster pour qu'il soit disponible pour tous les noeuds de calcul

# Transfer Learning

Utiliser un modèle qui a appris et l'adapter à notre cas présent

# Préparation du modèle

Création d'un nouveau modèle en récupérant l'avant dernière couche --> extraction de features images  
Dimension de vecteur de sortie (1,1,1280)

## extraction de features

Utilisation de PandasUDF ---> répartition des données et applications itérative puis stockage des résultats dans S3  
Traitement des données en parallèle sur les différents noeuds

## réduction de dimension PCA

Analyse en Composante Principale en conservant un maximum d'informations --> stockage des résultats dans S3

# Conclusion

Les principales actions effectuées:

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

- EMR
- S3
- IAM

2. Appropriation de la chaîne de traitement des images

3. Utilisation d'un environnement Big Data au profit de la société "Fruits"

