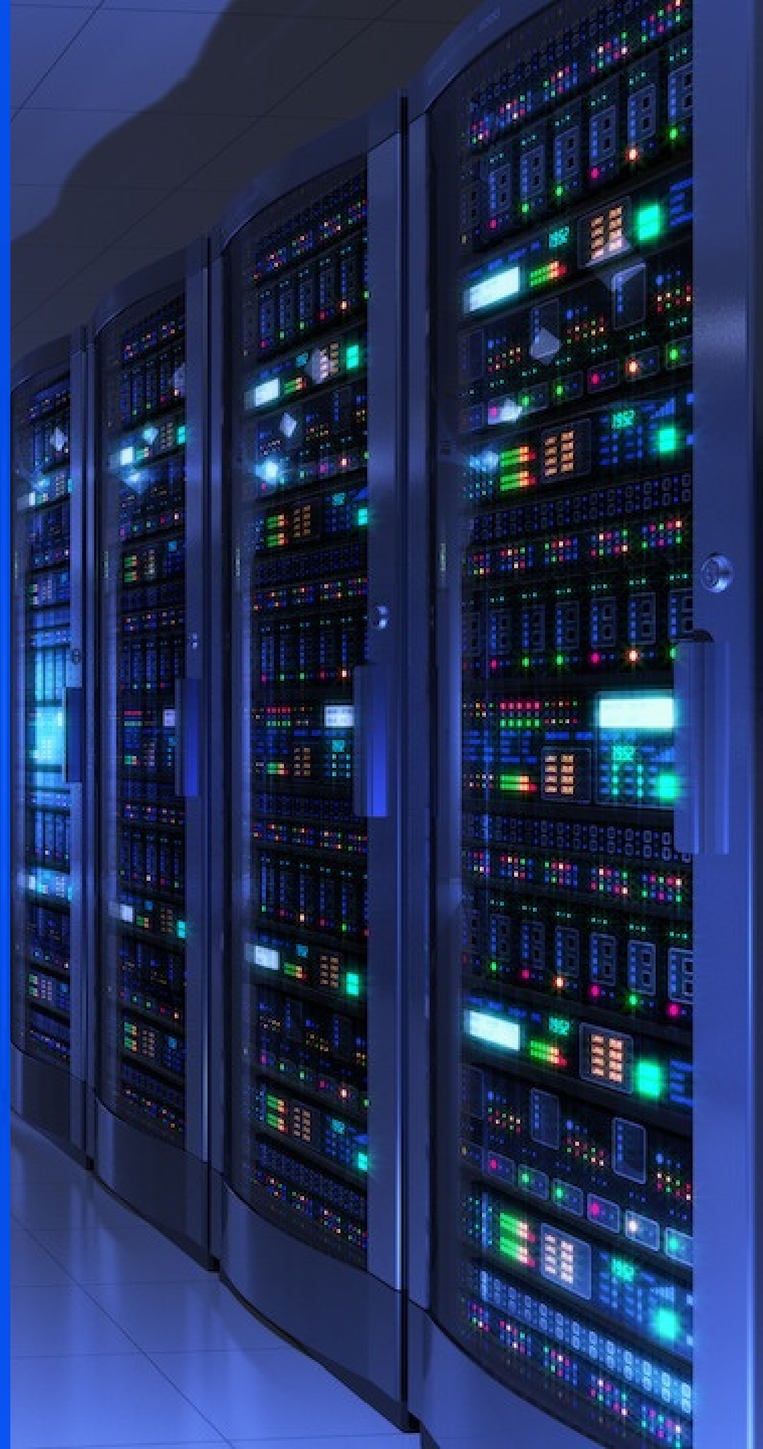
OPENCLASSROOMS - DATA SCIENTIST

P8 : DÉPLOYEZ UN MODÈLE DANS LE CLOUD

ROMAIN VAILLANT - MAI 2022



P8 : Déployez un modèle dans le cloud

SOMMAIRE

1.Problématique

APPLICATION MOBILE

Start-up. Développement d'une **application** de **reconnaissance instantanée** de fruits & légumes.

DONNEES

Le **jeu de données** est constitué des **images** de fruits et leurs **labels** associés.

DEVELOPPEMENT INITIAL

Une première **chaîne de traitement** des données incluant le **preprocessing** et une étape de **réduction de dimension**.

PASSAGE A L'ECHELLE

Le volume de données va augmenter. Besoin d'une **utilisation du cloud** pour profiter d’une **architecture Big Data**.

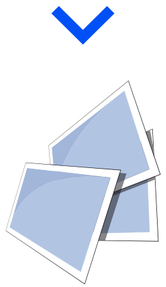
2.Jeu de données

131 fruits et légumes
90380 images (360°)

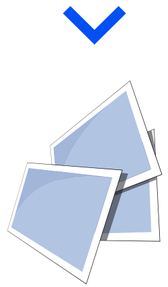
100x100
RGB
.jpg

(échantillons de 10 images, 3 fruits)

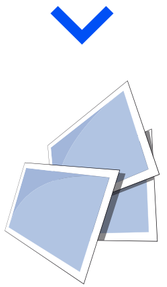
fruits-360



apple
(apple-pink-lady)



avocado



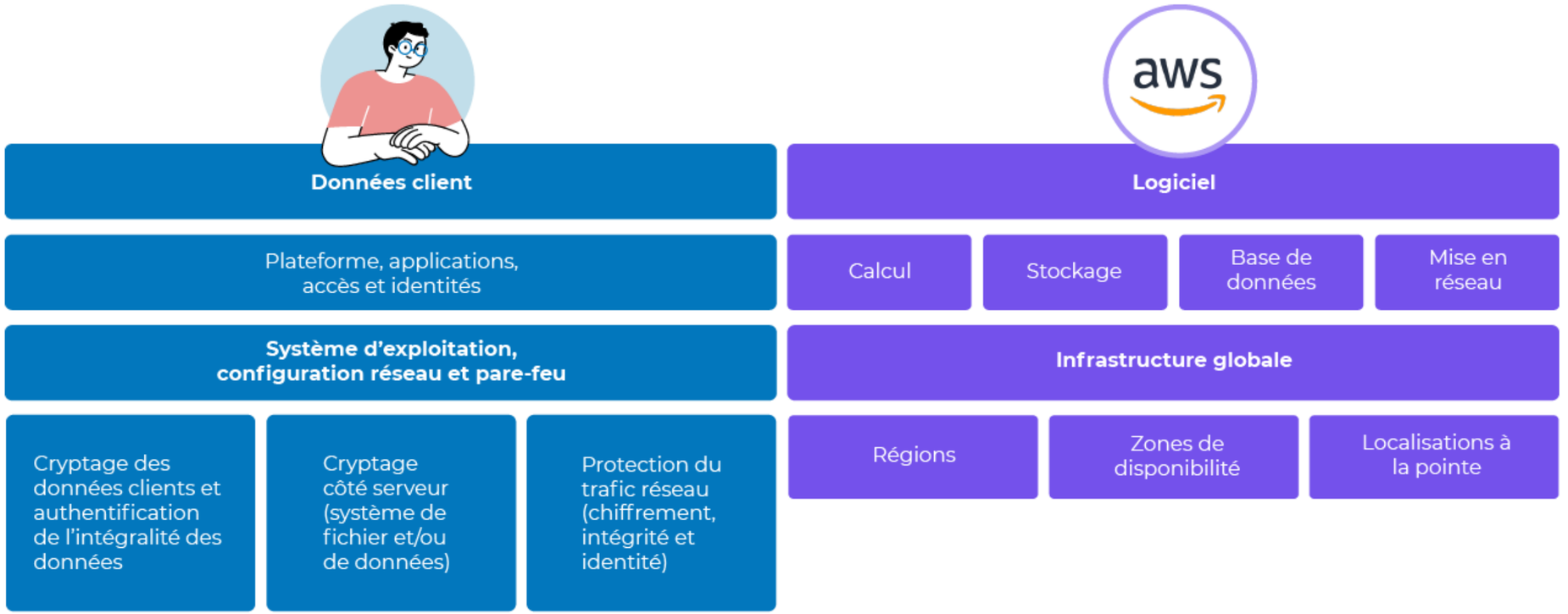
pineapple

3.Cloud et environnement Big Data

CLOUD
AMAZON WEB SERVICES (AWS)

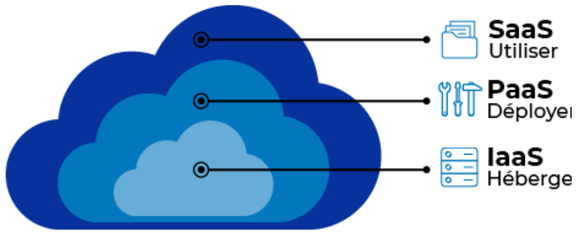
ARCHITECTURE BIG DATA
APACHE-SPARK
PYSPARK

CLOUD
AMAZON WEB SERVICES



Le principe de responsabilité partagée chez AWS

CLOUD
AMAZON WEB SERVICES



Type de services

UTILISATEUR IAM

Création de la clé privée de l'utilisateur pour la connexion SSH au serveur EC2

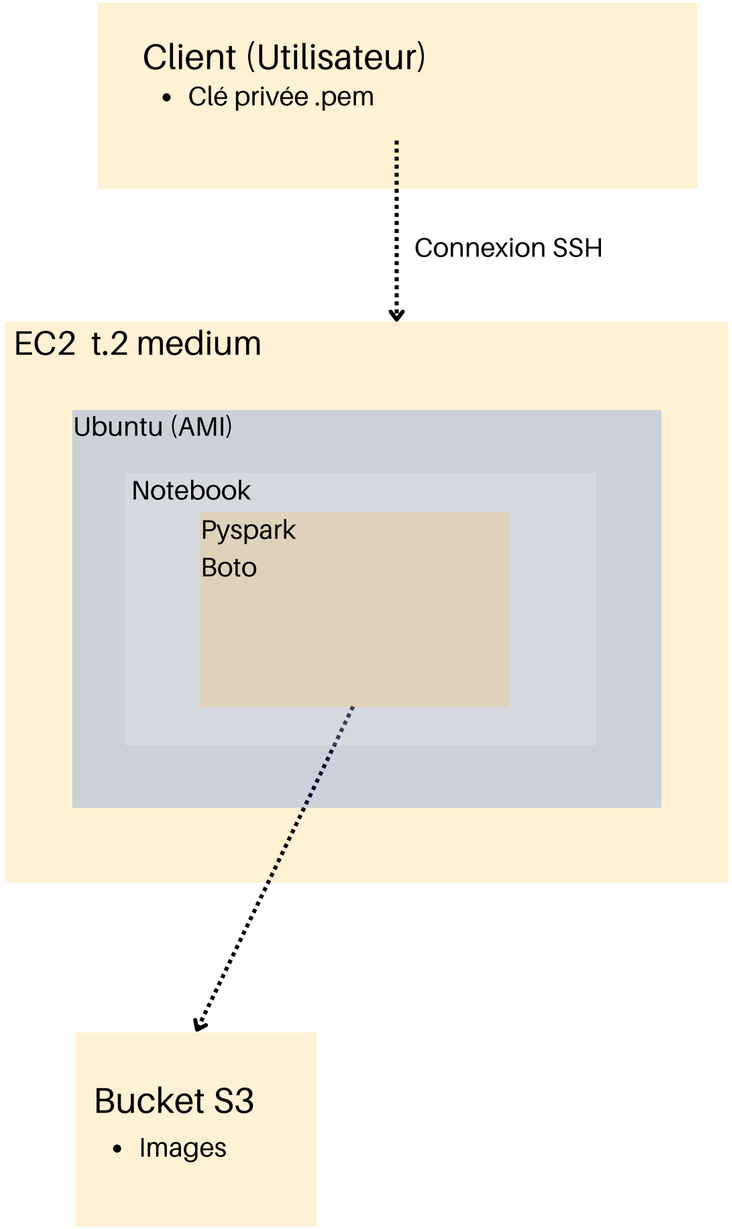
SERVEUR EC2 (IAAS)

Serveur de développement:
Ubuntu (AMI)
Anaconda
Pyspark

Via connexion SSH

BUCKET S3

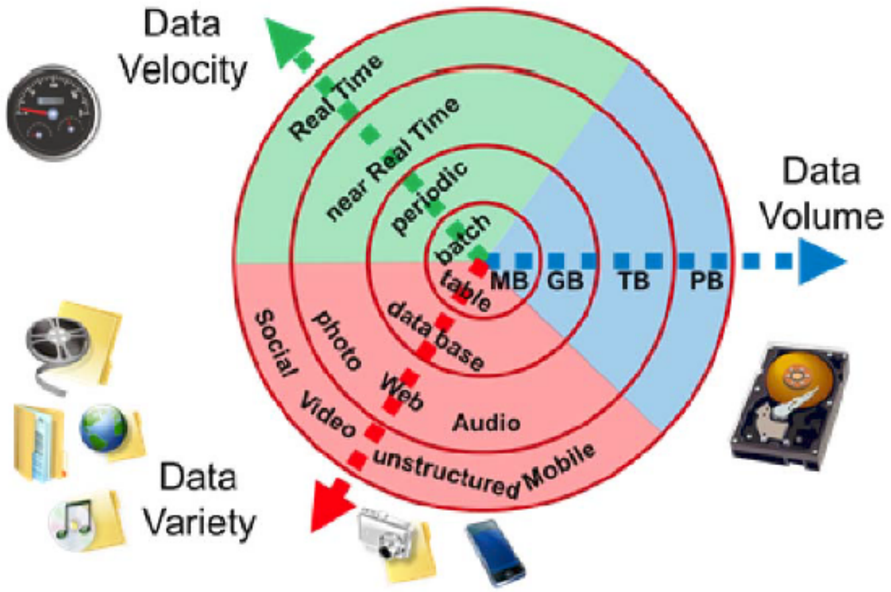
Stockage des images



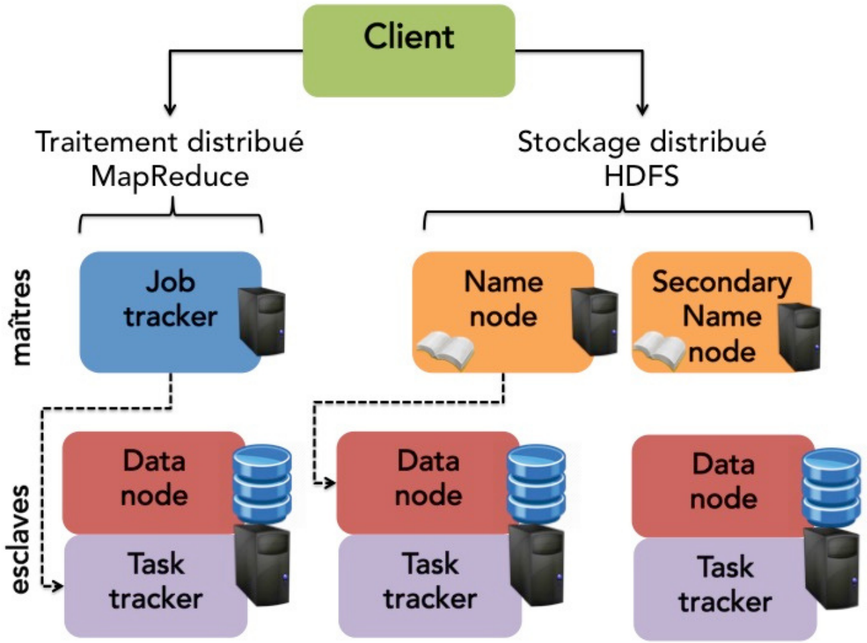
BIG DATA

- Terme apparu en 1997 dans un article sur la visualisation des données dans le cadre de la 8 ème conférence IEEE (Institute of Electrical and Electronics Engineers)
-Taille des données > RAM disponible
- Progrès des capacités des systèmes de stockage

Les 3 V du Big Data



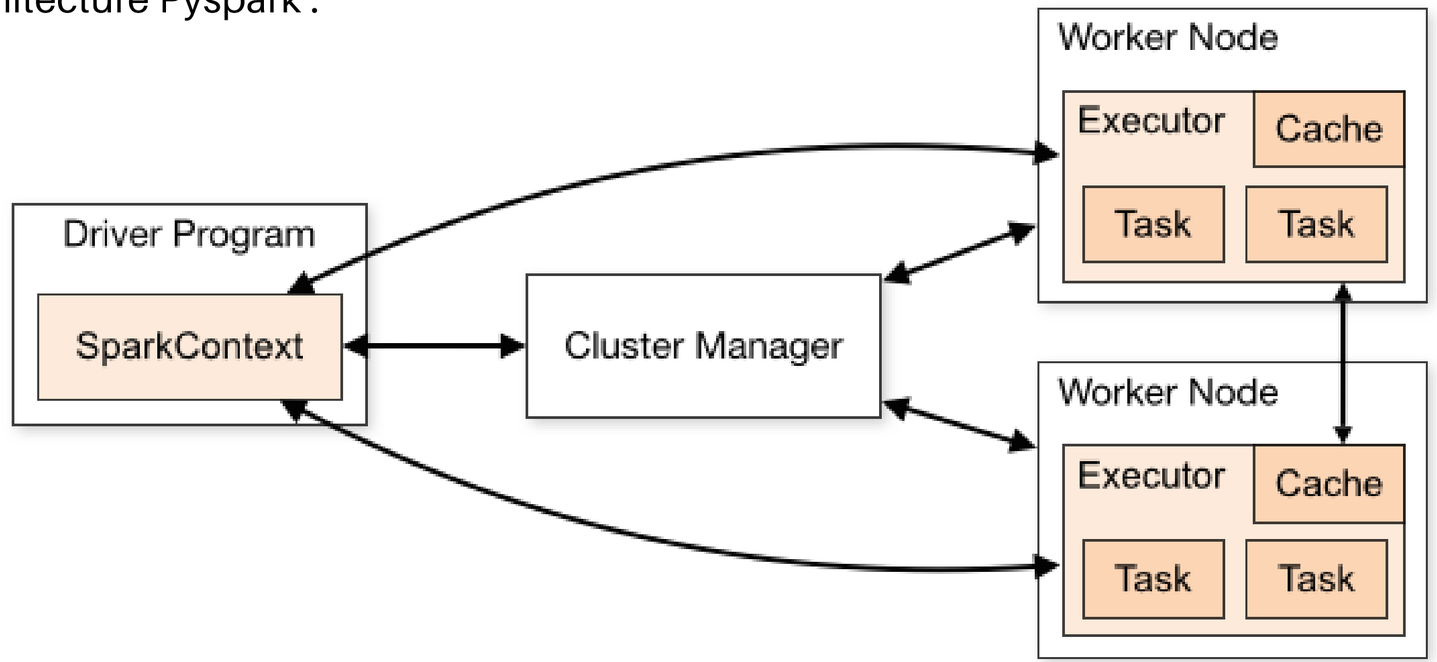
Architecture Hadoop



PYSPARK

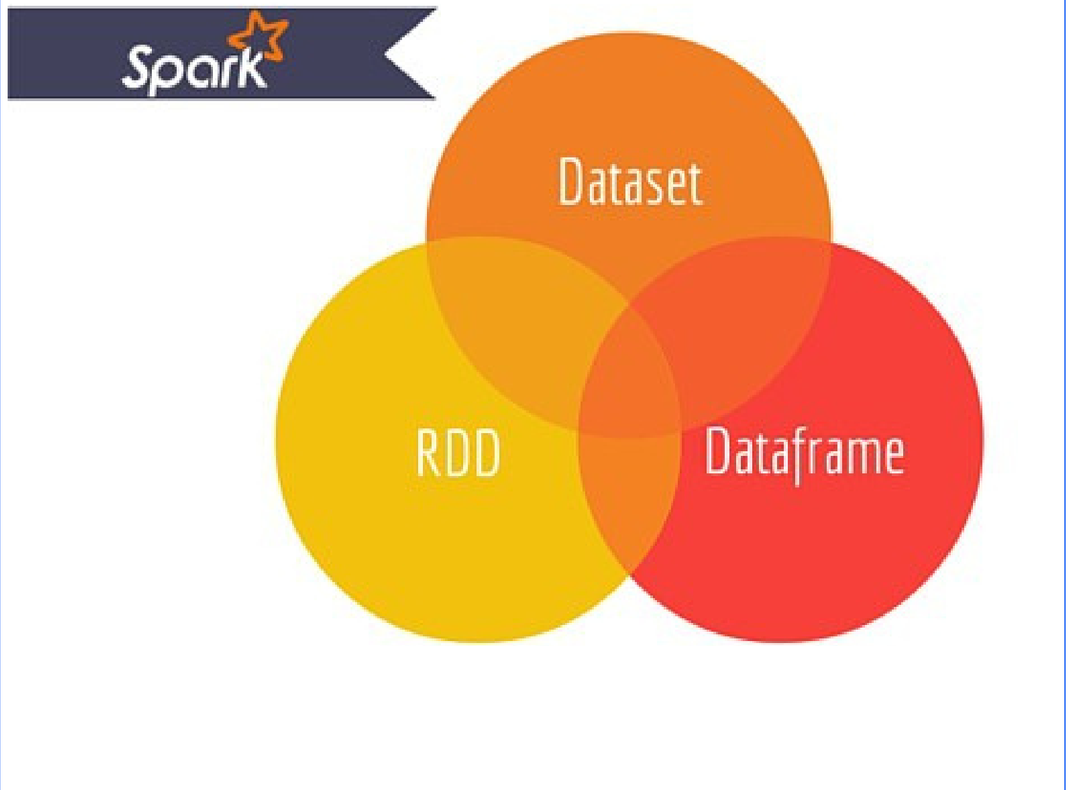
PySpark permet à python de s'interfacer dynamiquement avec des **objets JVM** à travers la **librairie Py4j**

Architecture Pyspark :



PYSPARK APIs

APIs



Features

RDD (2011)

Distribute collection of JVM objects

Functional Operators (map, filter, etc.)

DataFrame (2013)

Distribute collection of Row objects

Expression-based operations and UDFs

Logical plans and optimizer

Fast/efficient internal representations

databricks

4. Chaîne de traitement

CHARGEMENT DES DONNEES

Boto
resource()

PRE-PROCESSING

Keras
preprocess_input()

VECTORISATION PAR TRANSFER LEARNING

Keras
ResNet50()

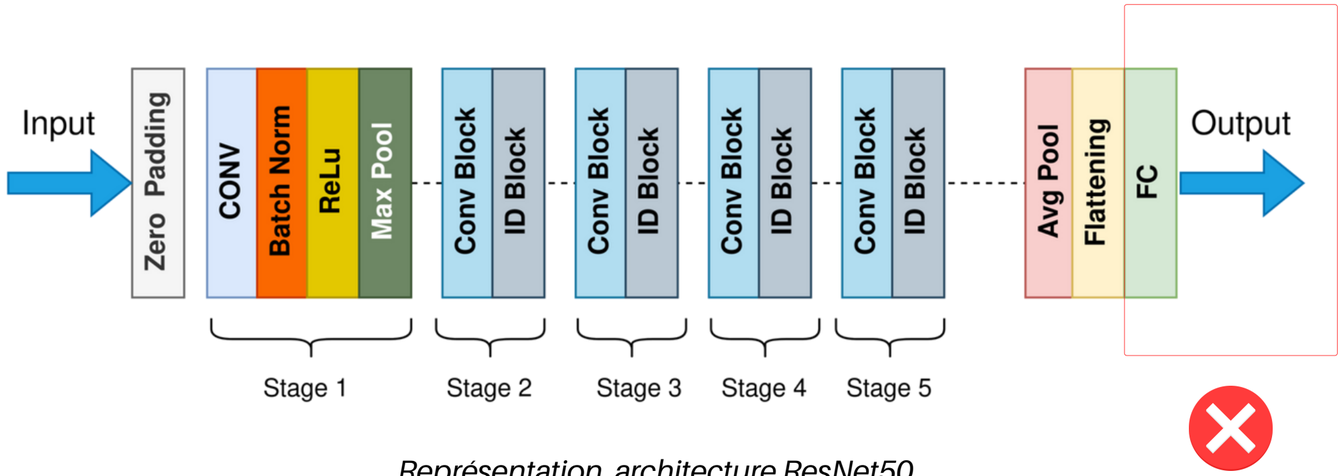
RÉDUCTION DE DIMENSION

Pyspark
Pipeline:
StandardScaler()
PCA()

TRANSFER LEARNING

RESNET50

Suppression des couches **fully connected** et **softmax**



Représentation architecture ResNet50

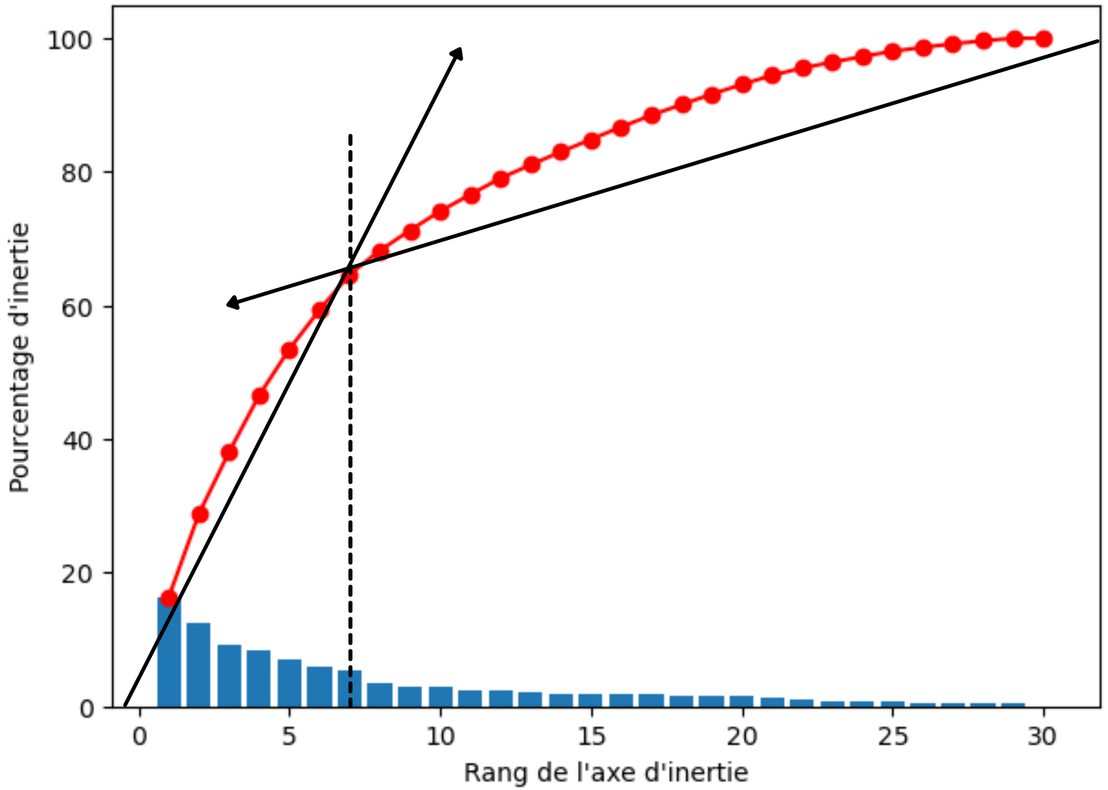
REDUCTION DE DIMENSION (PCA)

Vecteurs creux :

[0.0, 0.0, 0.0, **0.66**, 0.0, 0.0, **0.27**, ...]

Méthode du coude :

Composantes **k=7**



Ebouli des valeurs propres

5.Conclusion et recommandations

PUISSANCE DU BIG DATA

Parallélisation (Pyspark)

UTILITE DU CLOUD

AWS

IAM, EC2, EC2

PRETRAITEMENT

Transfert Learning - Extraction des features

Réduction de dimension (PCA)

PASSAGE A LECHELLE

Première étude faisabilité

Besoin machine + puissante (4 Go / 2 coeurs) ?

Modification des paramètres de configuration

Merci de votre écoute

Séance de questions