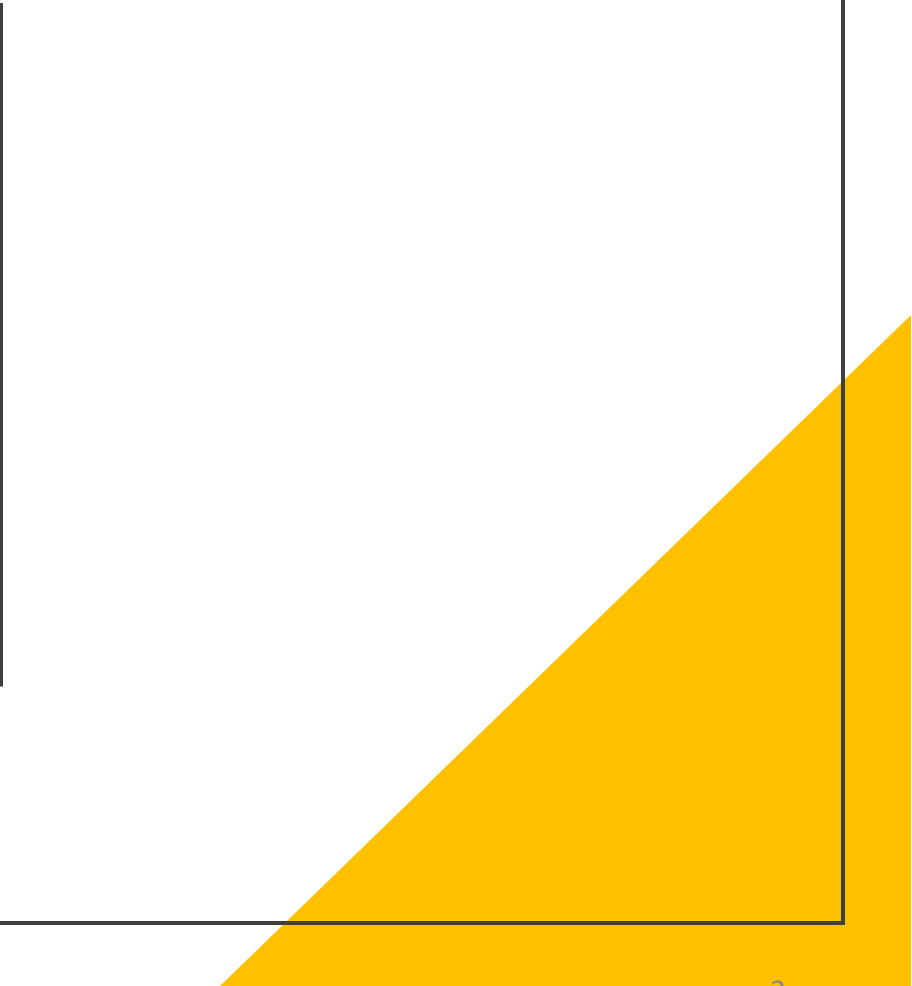


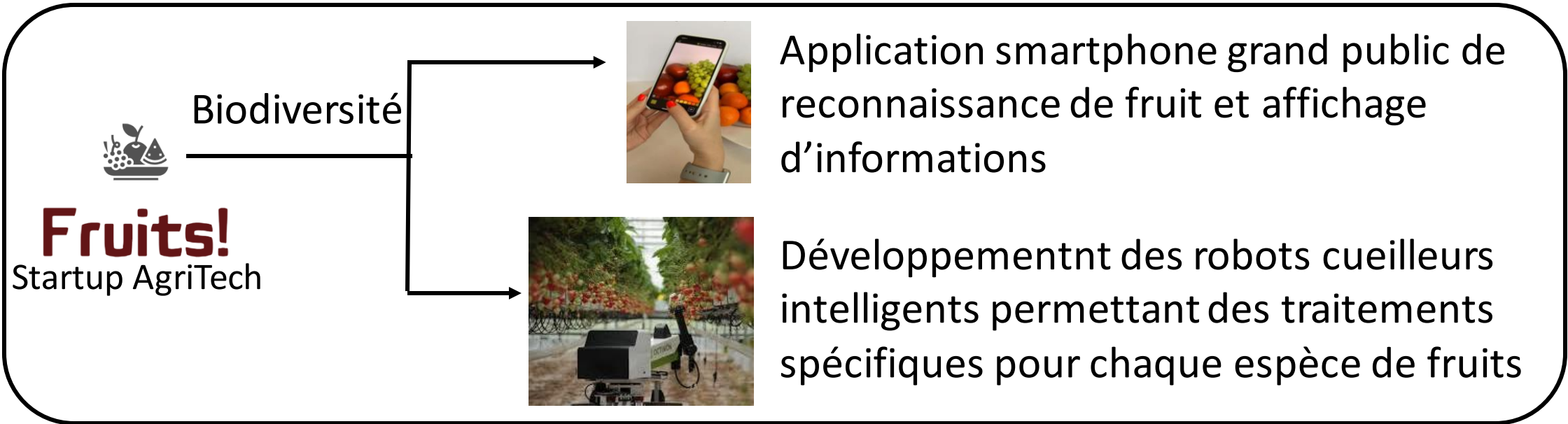
Projet 8 : Déployez un modèle dans le cloud

Fanjamalala Rajaonalison
02/2022

Présentation de l'étude

A vertical line is positioned to the right of the title. In the bottom right corner, there is a yellow triangle pointing upwards and to the left, with its hypotenuse forming a diagonal line.

Contexte



- Accès grand public
- Données massives : variées, volumineuses, vitesse d'obtention



- Développement d'une première version de l'architecture Big Data
- Mise en place d'une première version du moteur de classification des images de fruits

Dataset (source : <https://www.kaggle.com/moltean/fruits>)

Description :

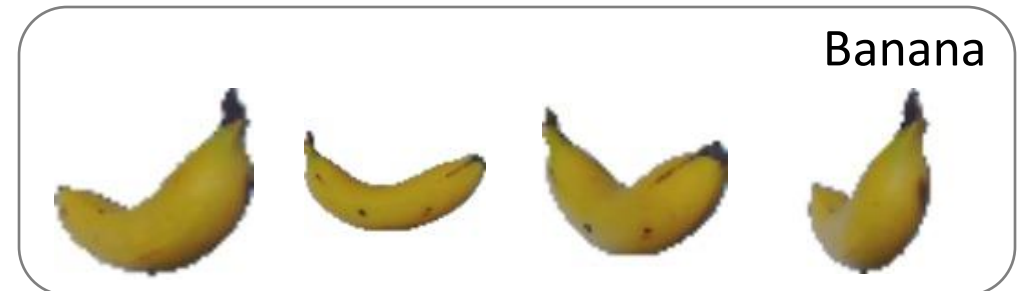
- Ensemble de données contenant des images de haute qualité de fruits et légumes, avec les labels associés
- 120 variétés de fruits (un fruit peut avoir plusieurs variétés)

Caractéristique du jeu :

- Taille : 90483 images
- Jeu de Train : 67692 images
- Jeu de Test : 22688 images

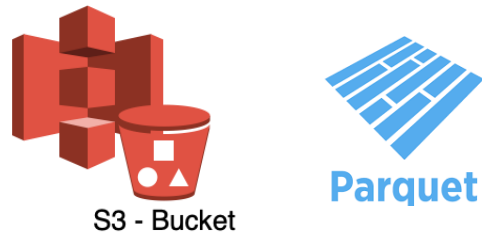
Caractéristique des images :

- Taille : 100x100 pixels
- Format : JPG RGB
- Photos sur fond blanc de fruit sous différents angles



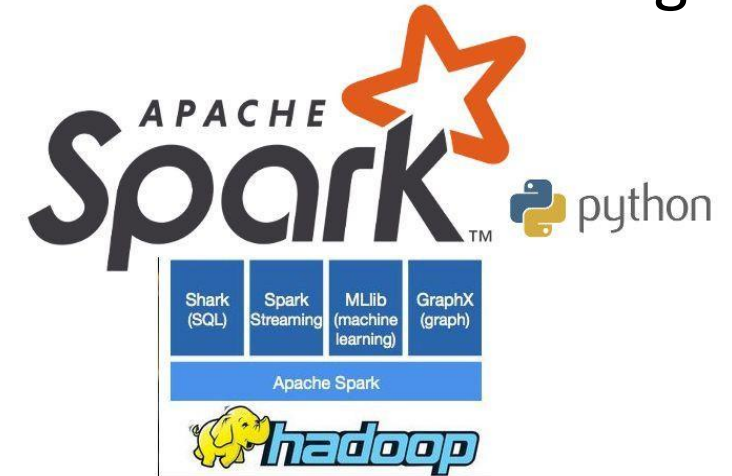
100%

1. Création de l'environnement Big Data sur le cloud



3. Sauvegarde des résultats

2. Traitement des images



Quelques Concepts utiles

A vertical line is positioned to the right of the text. A yellow triangle is located in the bottom right corner of the slide, pointing towards the top right.

AWS (Amazon Web Services)

Plateforme spécialisée dans les services de cloud computing à la demande

➔ fournisseur de services informatiques via Internet



EC2 (Elastic Compute Cloud)

- Interface web permettant de créer des machines virtuelles ou instances du serveur (partie des serveurs d'Amazon)
- Services de calculs variées, complètes et performantes



S3 - Bucket

S3 (Simple Storage Service)

- Stockage d'objets conçu pour extraire n'importe quelle quantité de données, depuis n'importe où
- Capacité de mise à l'échelle, disponibilité des données, performances et sécurité

Spark - PySpark



Framework open source de calcul distribué
➔ essentiellement dédié au Big Data et Machine Learning

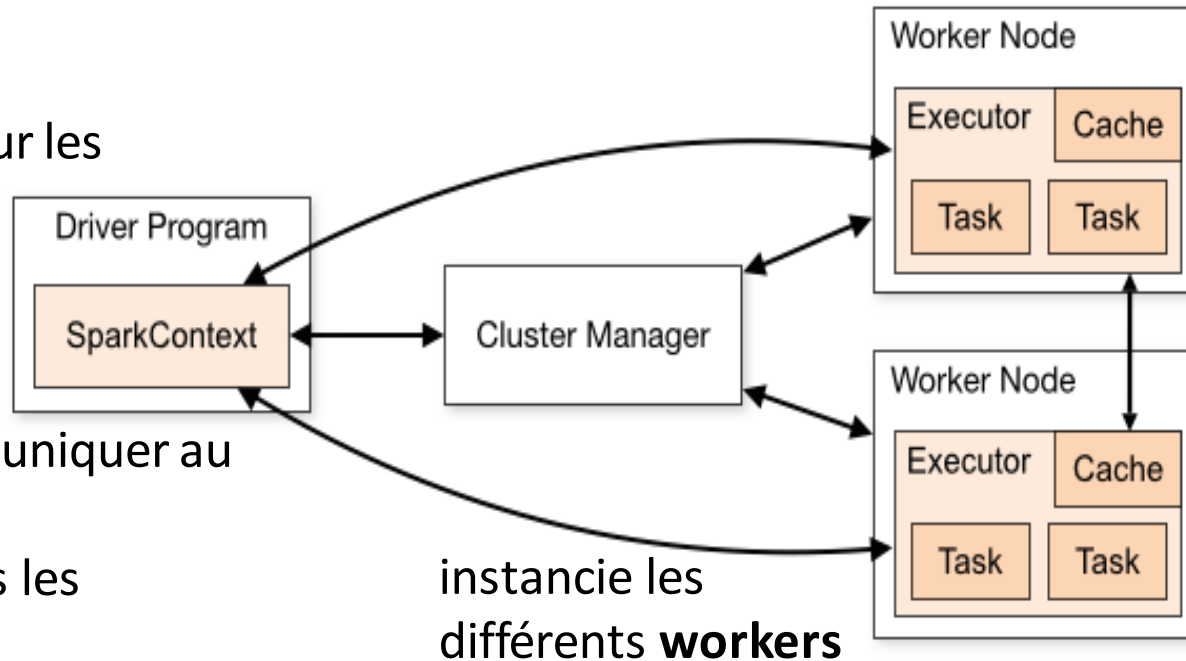


Librairie de Python pour Spark

Cluster Spark :

répartie les tâches sur les différents **executors**

- objet pour communiquer au driver
- coordonne toutes les opérations

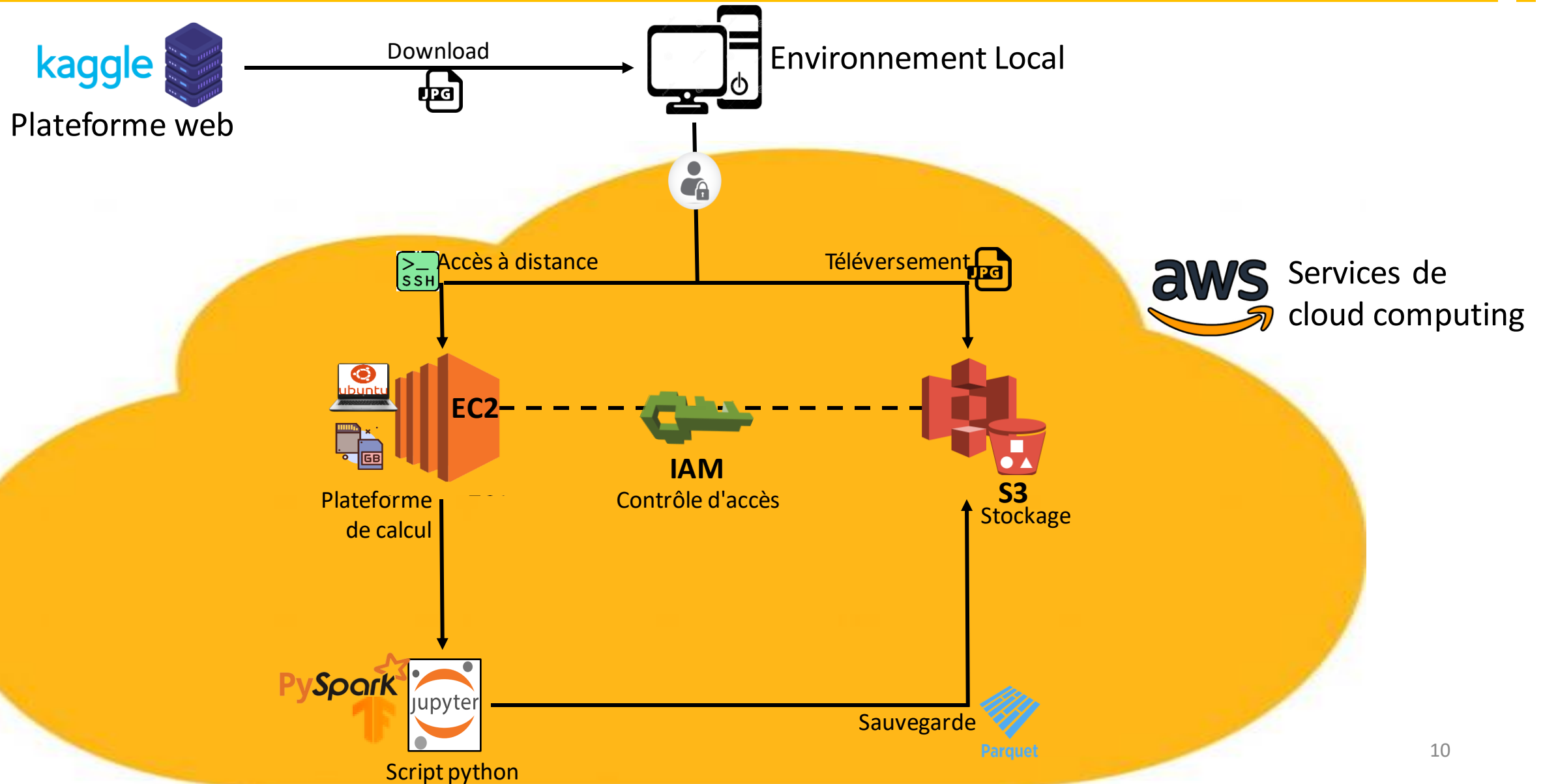


Chaque **worker** instancie un **executor** chargé d'exécuter les différentes tâches de calculs.

Environnement Big Data sur le cloud

A vertical line is positioned to the right of the text. A yellow triangle is located in the bottom right corner of the slide, pointing towards the top right.

Architecture



Instance EC2

AMI (Amazon Machine Image) :

- Image du disque, configuration logicielle



Ubuntu 20.04 LTS

Type d'instance : t2.medium

- Instance à usage général à faible coût avec la possibilité de booster en cas de besoin
- Idéal pour les bases de données, environnements de développement

Groupe de sécurité :

- **SSH**, HTTP, HTTPS, règle TCP personnalisée (port **8888**)

Stockage : 30Go

- Maximum de l'offre gratuit

Type d'instance	vCPU	Mémoire (Go)	Stockage (Go)	Performances de mise en réseau	Processeur physique
t2.medium	2	4	EBS uniquement	Faibles à modérées	Série Intel Xeon

Swap file :

fichier système qui crée un espace de stockage temporaire sur un disque SSD ou un disque dur lorsque le système manque de mémoire.



Applications : Python 3, Java 8, Spark 3, Hadoop-AWS 2.7, Anaconda, Jupyter notebook

S3 Bucket

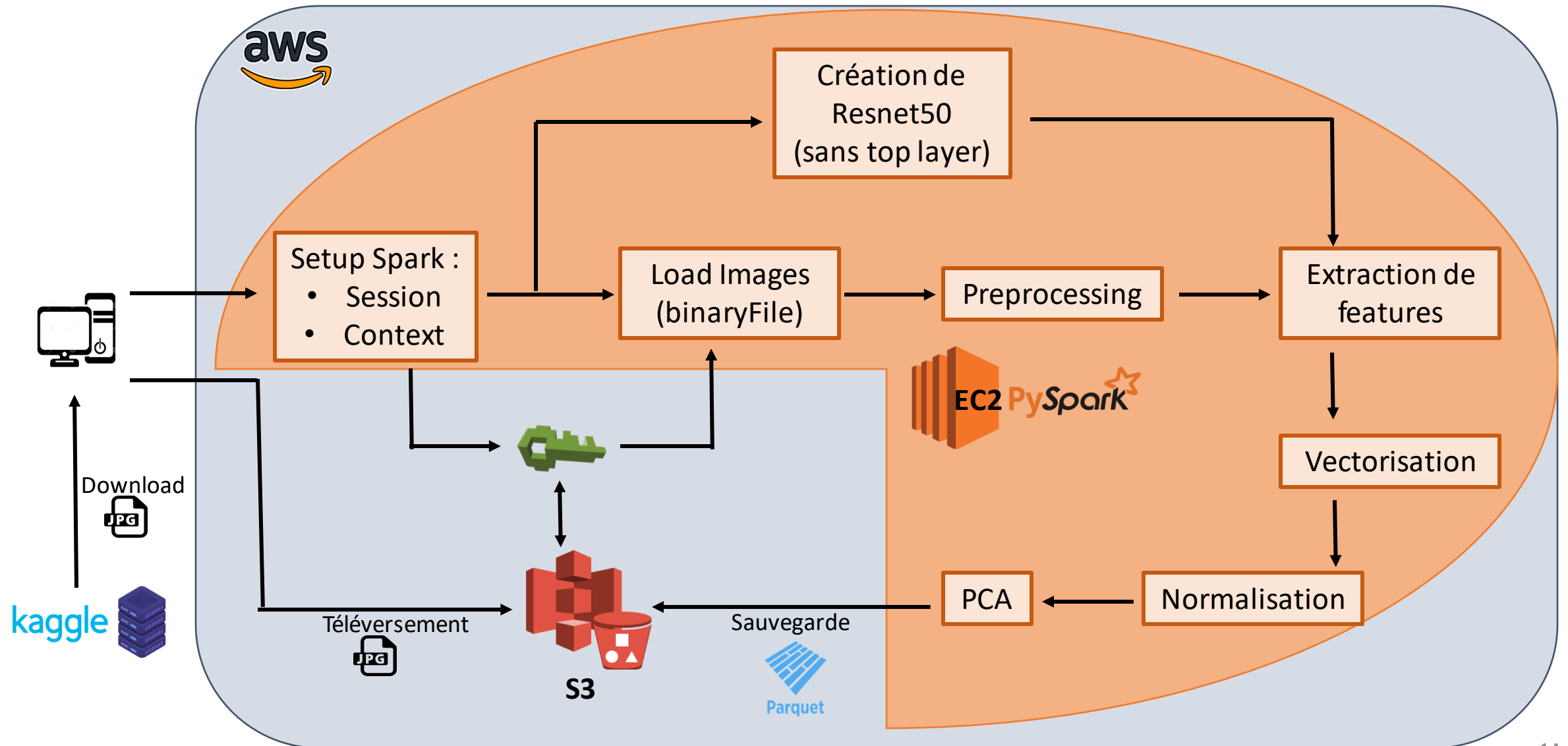
- Upload des données via Interface web ou AWS CLI (Interface de ligne de commande)
- Lecture des fichiers depuis Spark
- Enregistrement de fichier depuis Spark
- Contrôle d'accès par utilisateurs ou groupe d'utilisateurs
 - ➔ (Droit d'accès défini via resource-based Policy)
 - A partir d'identifiants de sécurité : Id, secret_key



Première Chaîne de traitements

A vertical line is positioned to the right of the text. A yellow triangle is located in the bottom right corner of the slide, pointing towards the top right.

Différentes étapes



Setup Spark & Chargement images

Setup Spark

- Localisation de Spark dans l'EC2
- Configuration de la session Spark
- Connexion à S3 : **IAM credentials**
- Création du contexte Spark capable de communiquer avec S3

Upload images

- Lecture de l'image en format **binaryFile**
- Spark Dataframe contenant le path et le contenu de l'image
- Ajout d'une colonne label obtenu par split du path

path	content	label
s3a://ocp8s3/Data...	[FF D8 FF E0 00 1...	Peach
s3a://ocp8s3/Data...	[FF D8 FF E0 00 1...	Apple
s3a://ocp8s3/Data...	[FF D8 FF E0 00 1...	Peach
s3a://ocp8s3/Data...	[FF D8 FF E0 00 1...	Apple
s3a://ocp8s3/Data...	[FF D8 FF E0 00 1...	Pear

only showing top 5 rows



Librairie utilisé

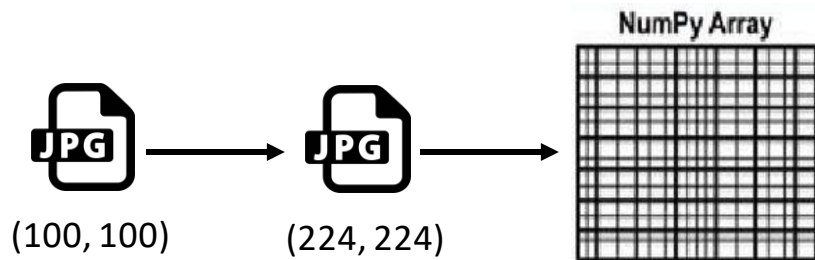
- findspark, Python Imaging Library, Pyspark



Preprocessing & Extraction Features

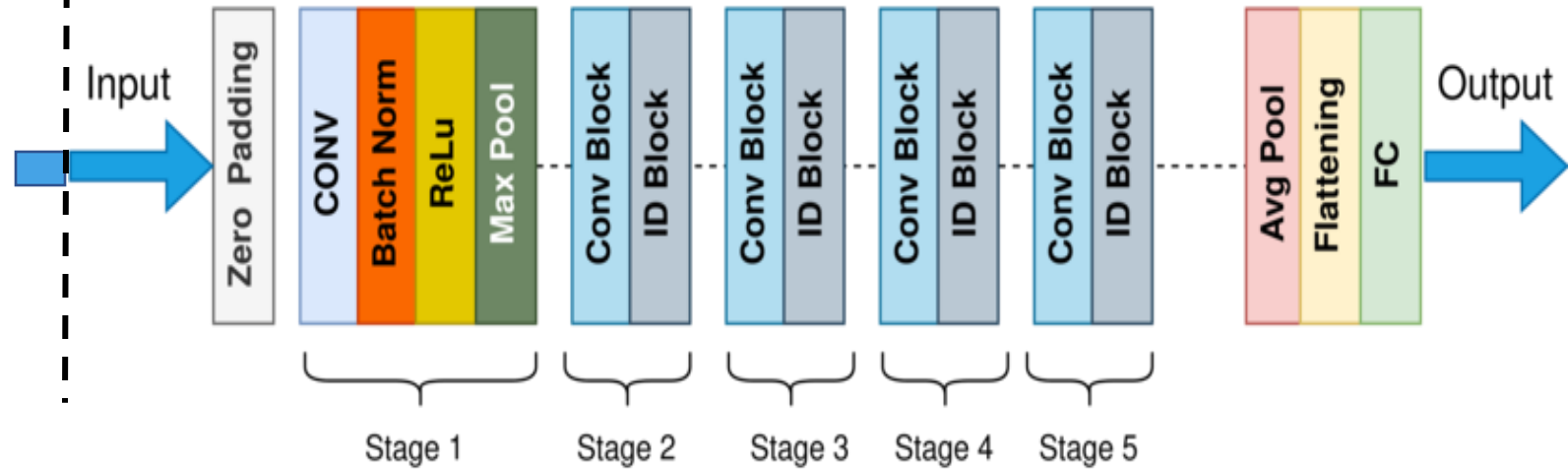
Preprocessing

- Redimensionnement des images (224, 224)
- Conversion en numpy array



Extraction des Features

- Transfert learning
- Modèle Resnet50 sans top layer
- 50 couches de neurones



Librairie utilisé

- Tensorflow, Numpy, Pyspark



Réduction dimensions & Sauvegarde

Conversion au format vecteur dense

Normalisation

- Suppression de la moyenne
- Scaling variance unitaire

features	featuresVct	featuresStd	featuresPca
[0.0, 13.786607, ...]	[0.0, 13.786606788...]	[-0.2545462489499...]	[-178.60495683676...]
[0.0, 13.263342, ...]	[0.0, 13.263341903...]	[-0.2545462489499...]	[57.3152703321728...]
[8.7571144E-4, 14...]	[8.75711441040039...]	[-0.2235875500555...]	[94.8506550363394...]
[0.0, 11.68219, 6...]	[0.0, 11.682189941...]	[-0.2545462489499...]	[186.945236921911...]
[0.0, 13.927928, ...]	[0.0, 13.927927970...]	[-0.2545462489499...]	[-118.07431465861...]

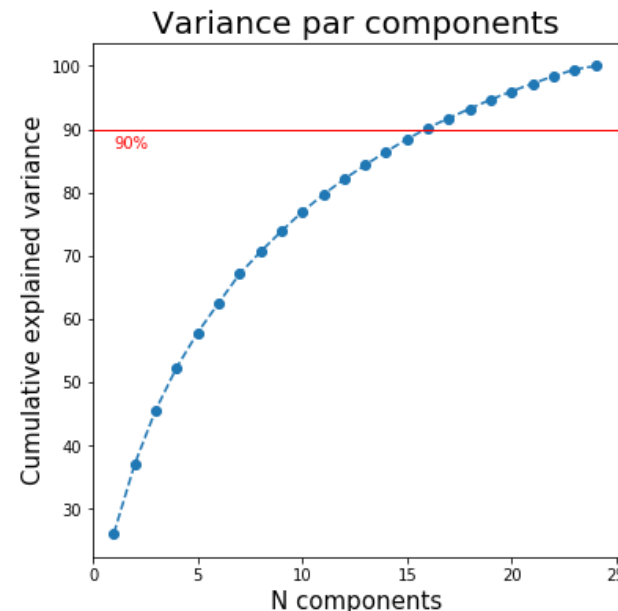
only showing top 5 rows

Transformation PCA sur les k premières composantes

- k= 25

Librairie utilisé

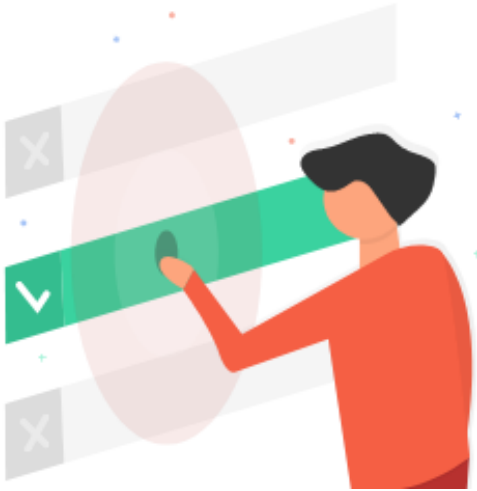
- Pyspark (ML)



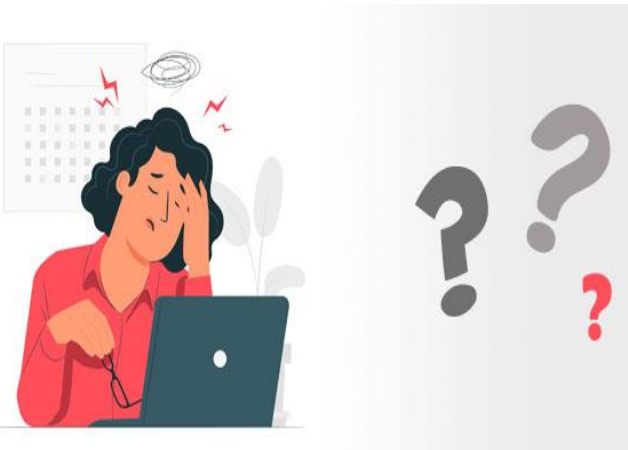
Conclusions

A vertical line is positioned to the right of the word 'Conclusions'. In the bottom right corner of the slide, there is a yellow right-angled triangle pointing towards the top-left.

Résumé de l'étude



- Accès et configuration de services AWS:
 - Mise en place d'une instance EC2 et d'un Bucket S3
 - Gestion des droits d'accès du S3
- Communication / Accès au serveur par SSH
- Configuration de session et contexte Spark
- Première chaîne de traitements des images



- Lenteur des calculs : configuration serveur pas assez puissante (t2.medium)
- Nombreuses possibilités techniques : choix complexes
- Debug compliqué : erreurs peu explicites, superposition Spark/Java/S3

Recommandations

Amélioration de l'étude

- Prétraitement d'images réels (recadrage, plusieurs fruits, arrière plan, etc.)
- Essaie d'autres modèles pour le transfer learning
- Choix d'un k adapté pour le pca
- Choix d'une instance plus performante

Passage à l'échelle

- Déploiement du modèle
- Évolution de l'infrastructure de calcul : instance EC2 de plus grande capacité RAM/Processeur
- Augmentation du nombre d'instances esclaves (nœuds) sans coupure
- Utilisation de EMR : cluster de calculs (plus adéquats pour exécuter des tâches de traitement de données distribuées à grande échelle)

Pousser le cas d'usage

- Identification de la maturité des fruits pour les cueillir au bon moment
- Identification des pathologies ou des fruits abîmés

MERCI

Questions



Incompréhensions

