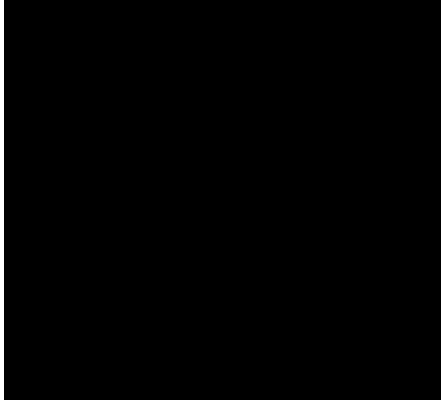
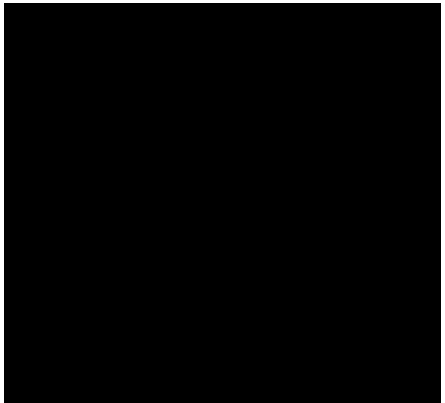
Projet 8
Déployez un modèle dans le Cloud

Sébastien MAVIEL – Formation Data Scientist
Projet n°8

Plan de la Présentation





Présentation du projet

Contexte

- L’entreprise Fruits cherche à développer une application mobile qui permettra aux utilisateurs de prendre en photo un fruit et d'obtenir des informations sur ce fruit.

Ma mission

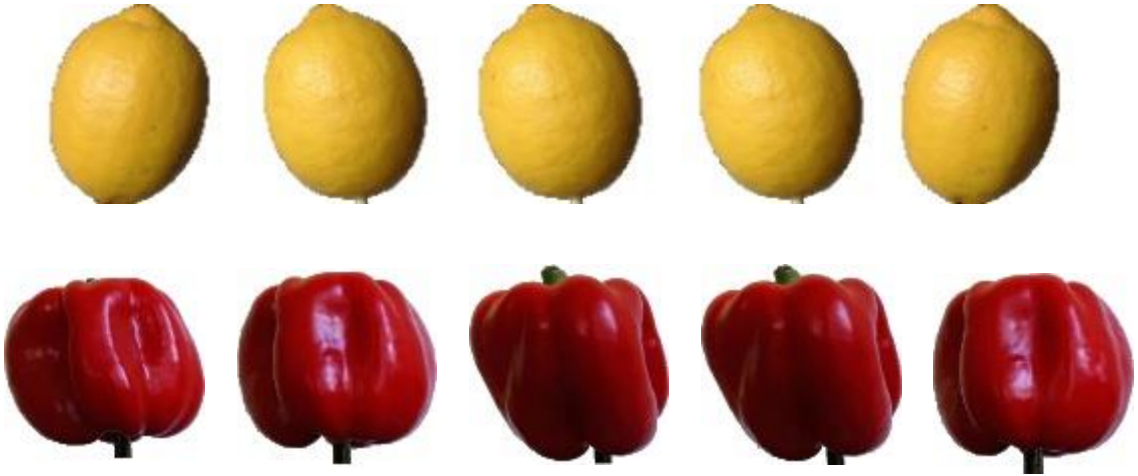
- Développer dans un environnement Big data une première chaine de traitement des données (Pre-processing et une étape de réduction de dimension)

Fruits!

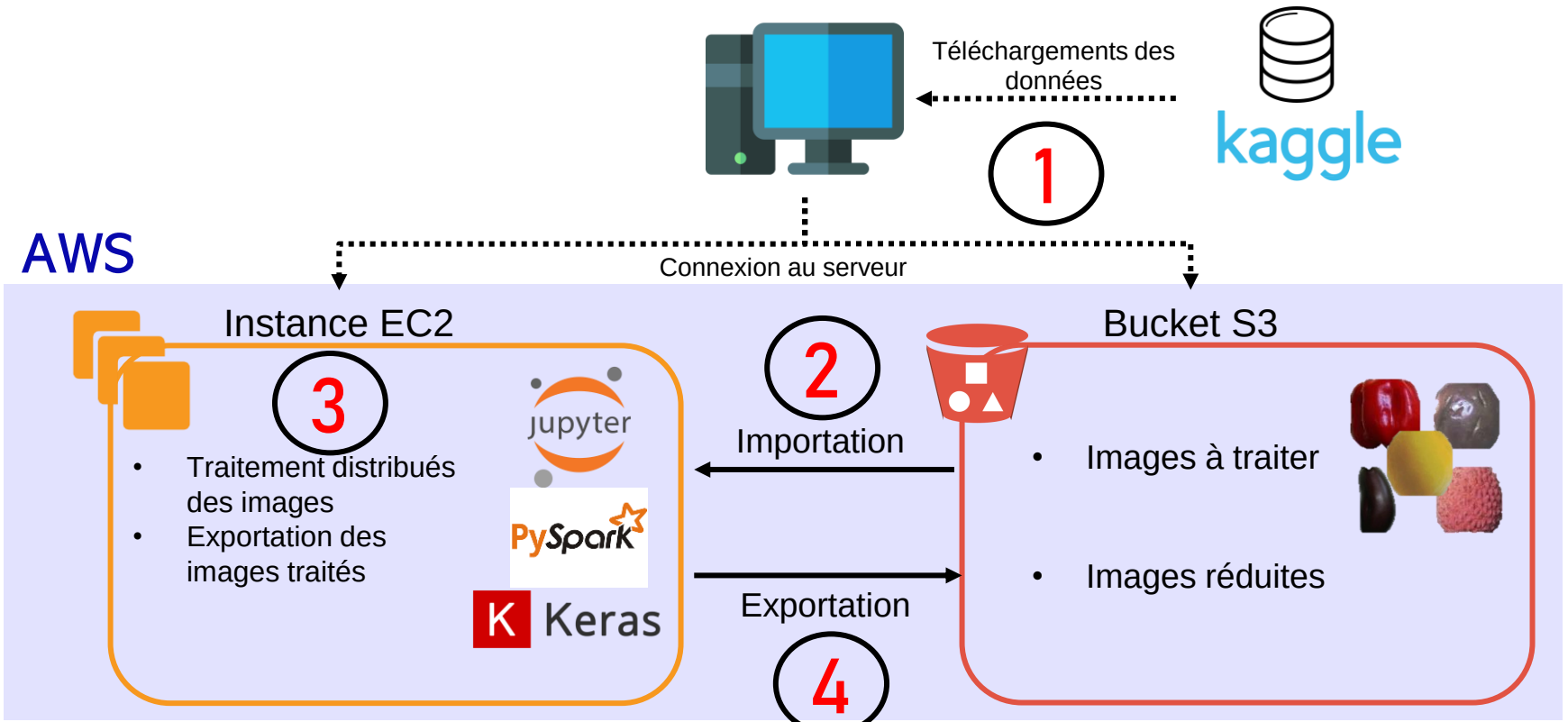
Présentation des données

Détails des données

- 67692 images 100x100 avec fond blanc
- 131 fruits
- Vue 360°

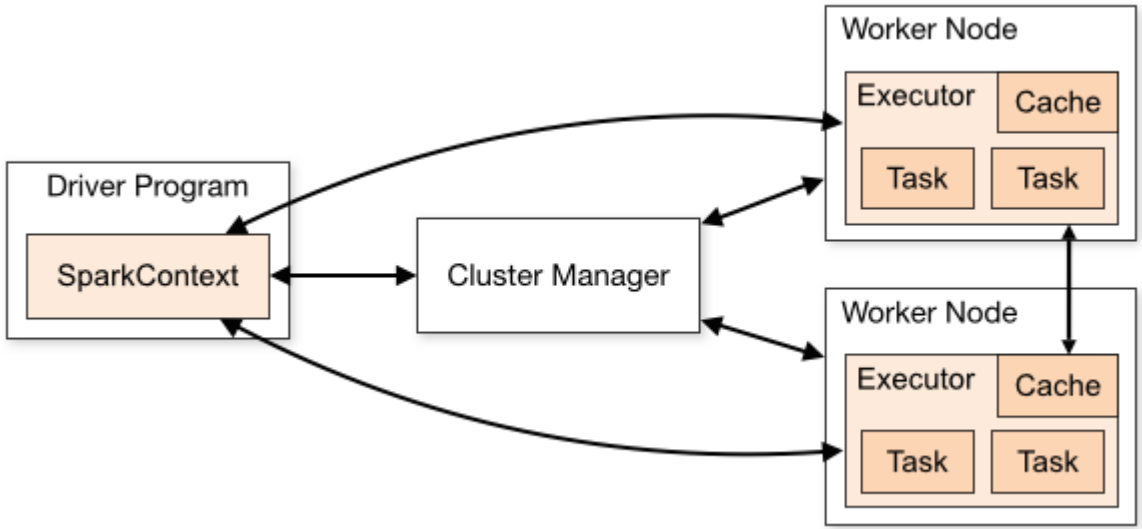


Architecture mise en place



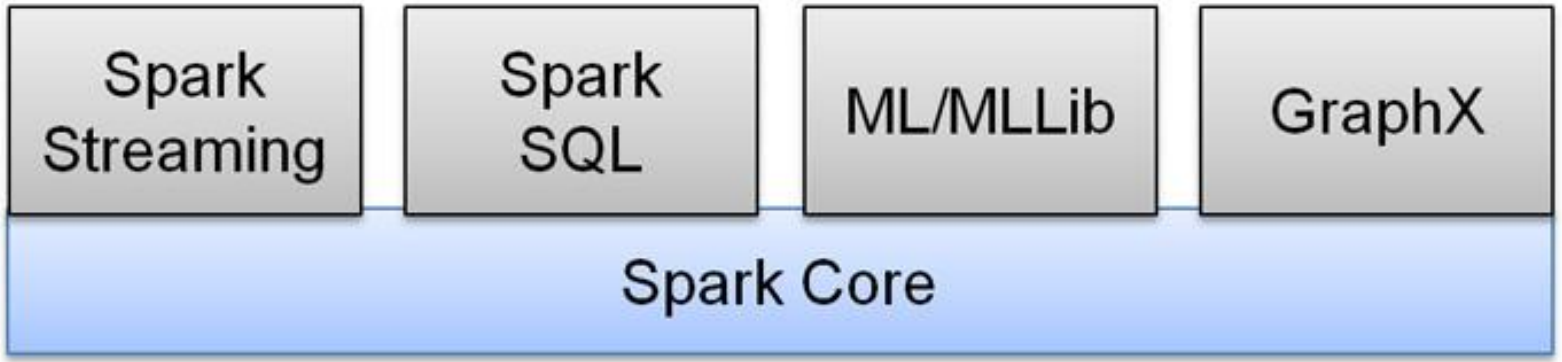
Spark et PySpark

- Spark est un framework open-source de calcul distribué.
- PySpark est une API Python pour utilisé Spark (nativement en Scala)



Spark et PySpark

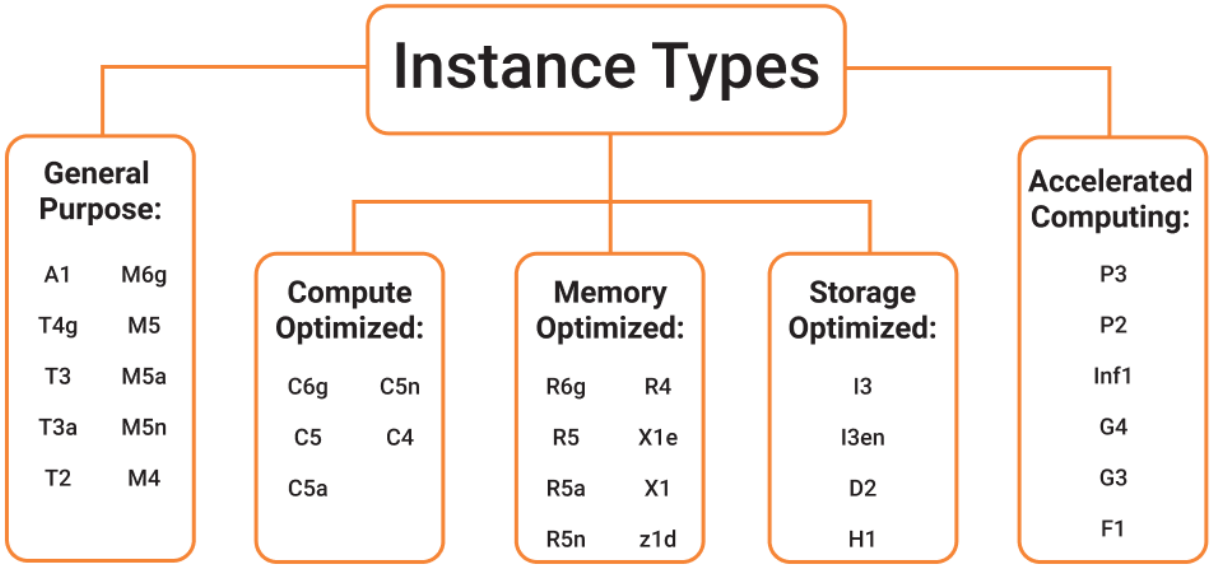
- 4 librairies construites par-dessus la Core API



Amazon Web Services

- Service de cloud computing à la demande
- AWS est le leader sur ce marché
- S3 et EC2 ont été utilisés pour ce projet

Elastic Cloud Computing (EC2)



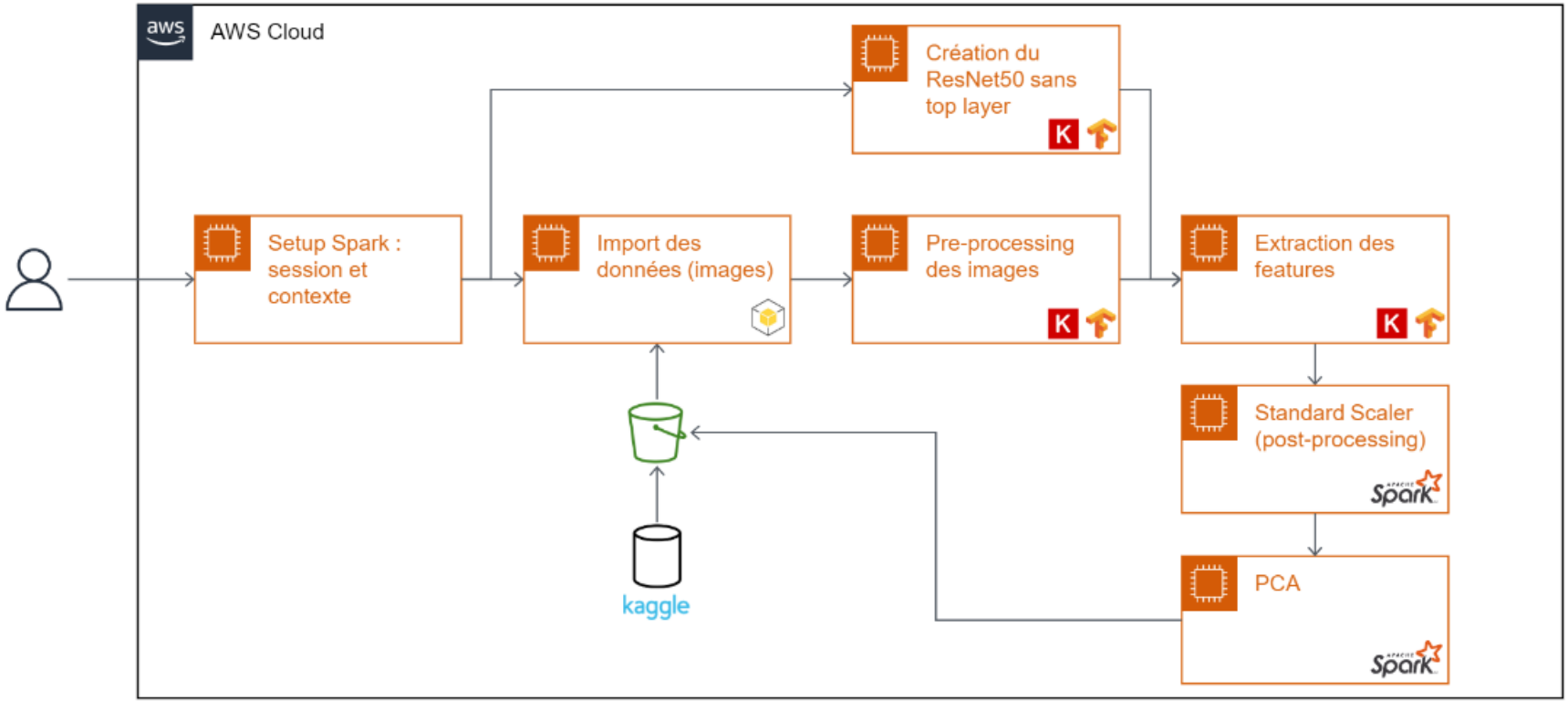
Instance T2: Pas cher et extensible

Simple Storage Service (S3)

Bucket S3

- Service de stockage et de distributions de fichiers
- Contrôle d'accès aux sous-dossiers et aux fichiers avec IAM
- Accès à la librairie avec boto3 (AWS SDK pour Python)

Chaîne de traitements des données



Conclusion

Travail réalisé

- Mise en place d'une instance EC2 et d'un bucket S3
- Administration et configuration d'un serveur Linux par SSH
- Configuration de session et contexte Spark
- Développement d'une chaine de pré-traitements

Difficultés rencontrées

- Mise en place de l'environnement Spark
- Débugging parfois complexe

Améliorations possibles

- Ajouter des workers Spark
- Ajouter le modèle de classification