



DATA SCIENTIST PROJET 8:

Deployer un modèle dans le cloud

Valide Dolores

Mentor : Germond Adrien



>>> Compétences évaluées



Utiliser les outils du cloud pour manipuler des données dans un environnement Big Data

Paralléliser des opérations de calcul avec Pyspark

Identifier les outils du cloud permettant de mettre en place un environnement Big Data



Sommaire

- Problématique
- Présentation des données
- Présentation de l'architecture Big Data
- Traitement des images
- Conclusions



Fruits!





Problématique





Fruits!



Fruits !

Souhaite proposer des solutions innovantes pour la récolte des fruits.

Développer des robots cueilleurs intelligent à l'aide d'une application qui permettra aux utilisateurs de prendre en photo un fruit et d'obtenir des informations sur ce fruit.



Mission

Développer un environnement Big Data
Réaliser une première chaîne de traitement des données avec le préprocessing et une étape de réduction de dimension.



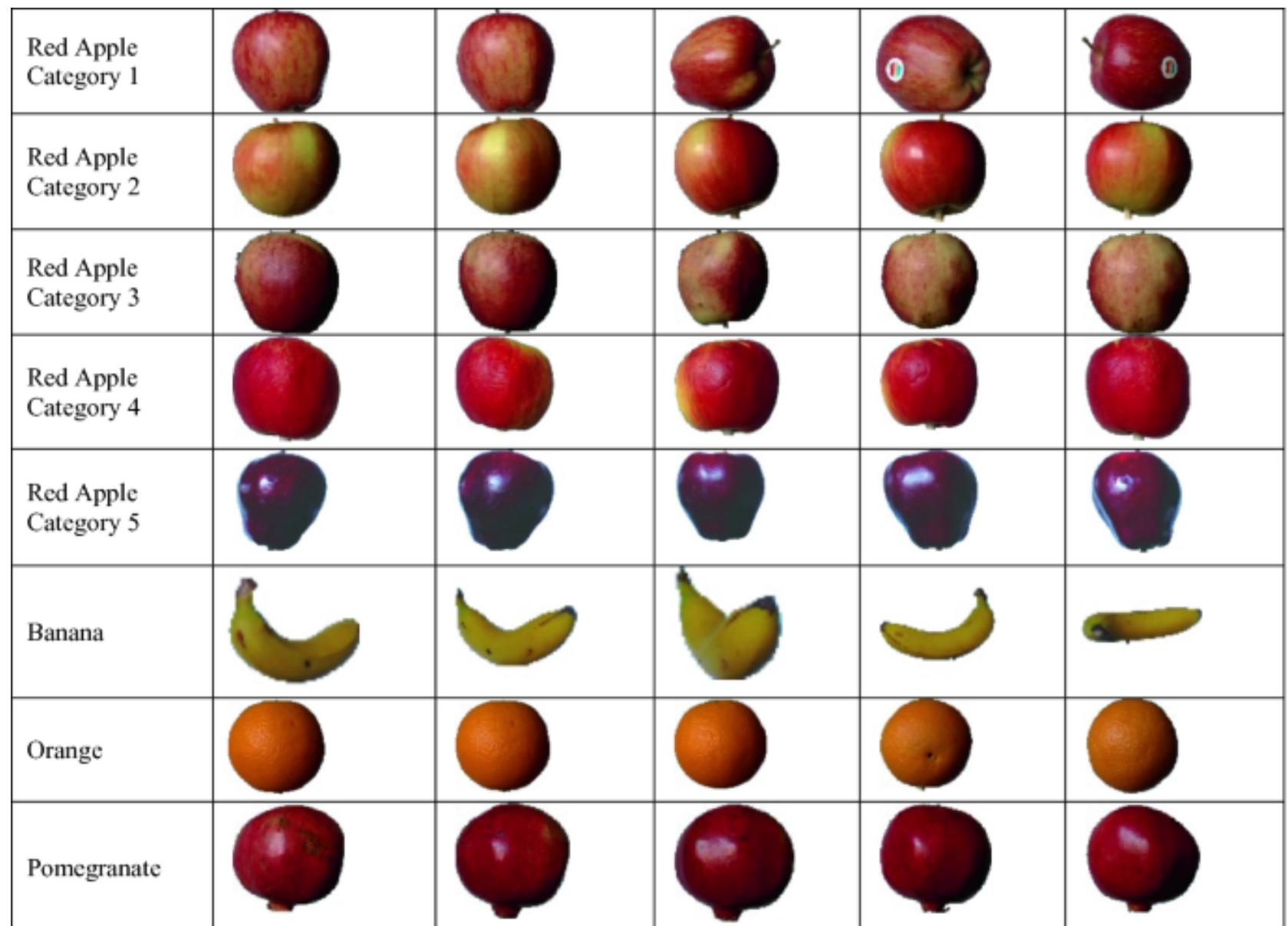
Présentation des données

Données issues d'un kernel Kaggle

- 90423 images et 131 classes
- 2 jeux de données training(67692) et test set (22688)

131 dossiers :

- représente un fruit ou un légume
- image avec fond blanc et sous 3 axes
- taille de 100 x 100 pixels en JPG RGB
- plusieurs variétés pour certains fruits

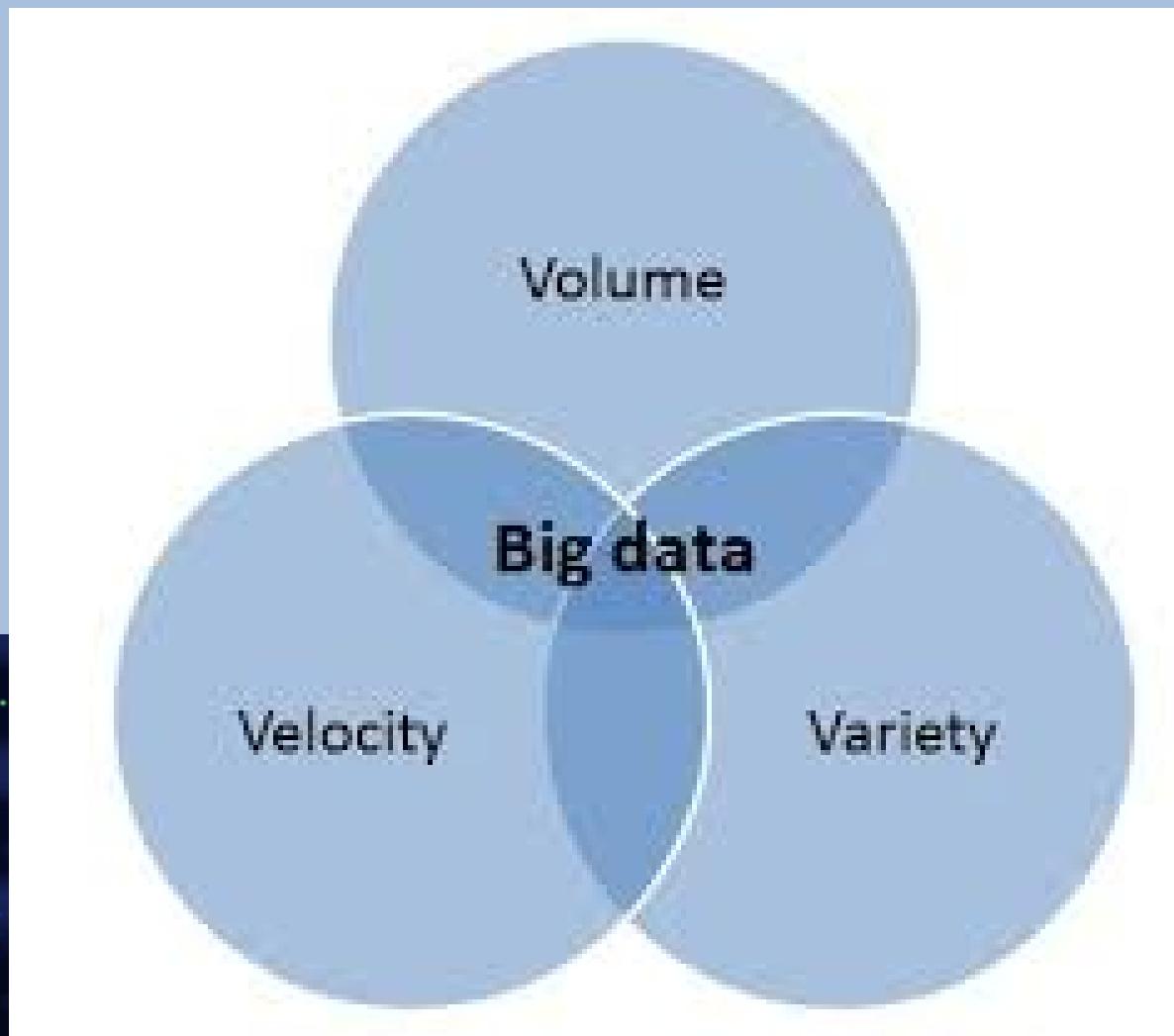


Pourquoi un environnement BIG DATA ?

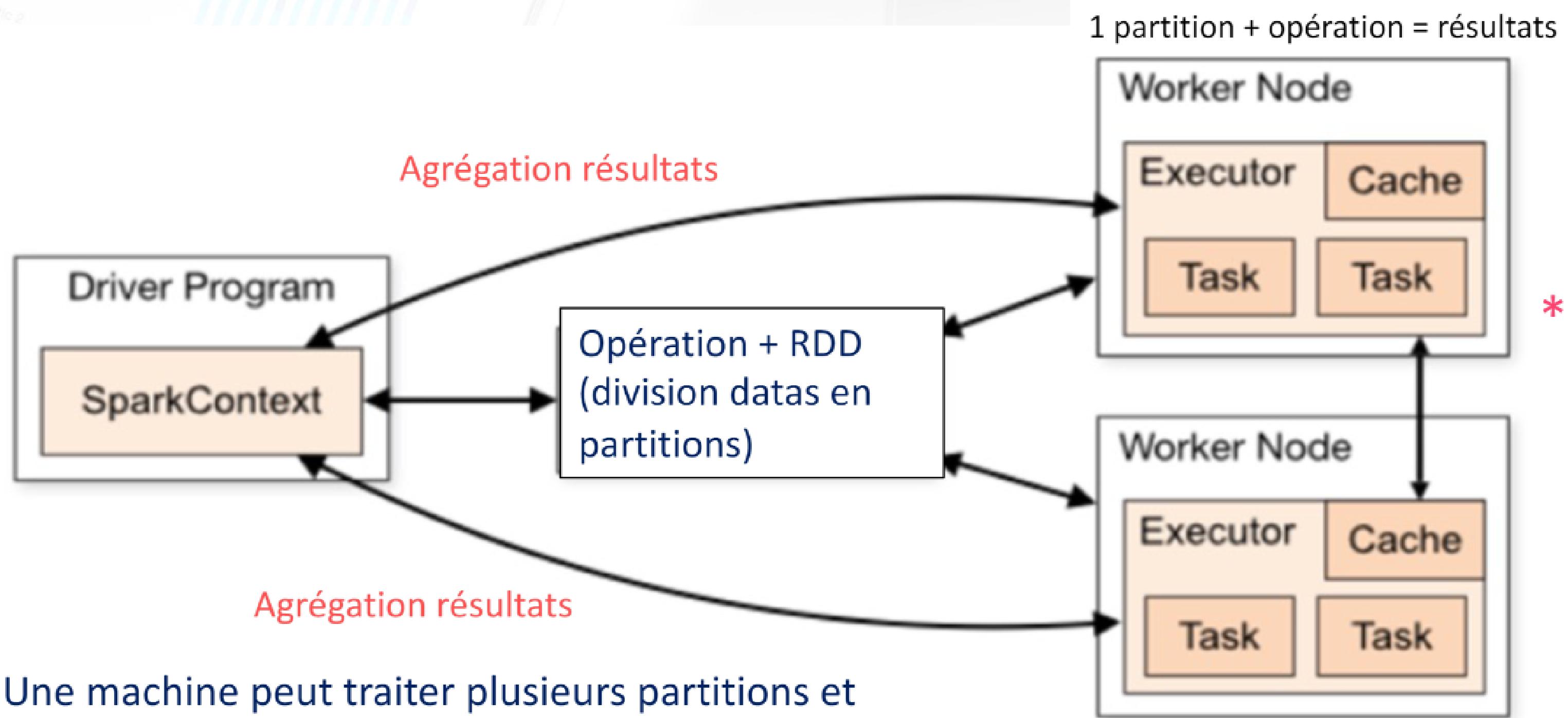
Fort volume de données

Susceptible d'augmenter rapidement

Dans divers formats (issus de l'application mobile).



Qu'est-ce que le calcul distribué?



Le service employé



Capacité de stockage:
Cloud AWS

Capacité de ram :
Serveur Cloud AWS

Calcul distribué :
MapReduce
Spark



Les étapes

Données sur le
cloud (s3)

Environnement
de travail EMR
Cluster

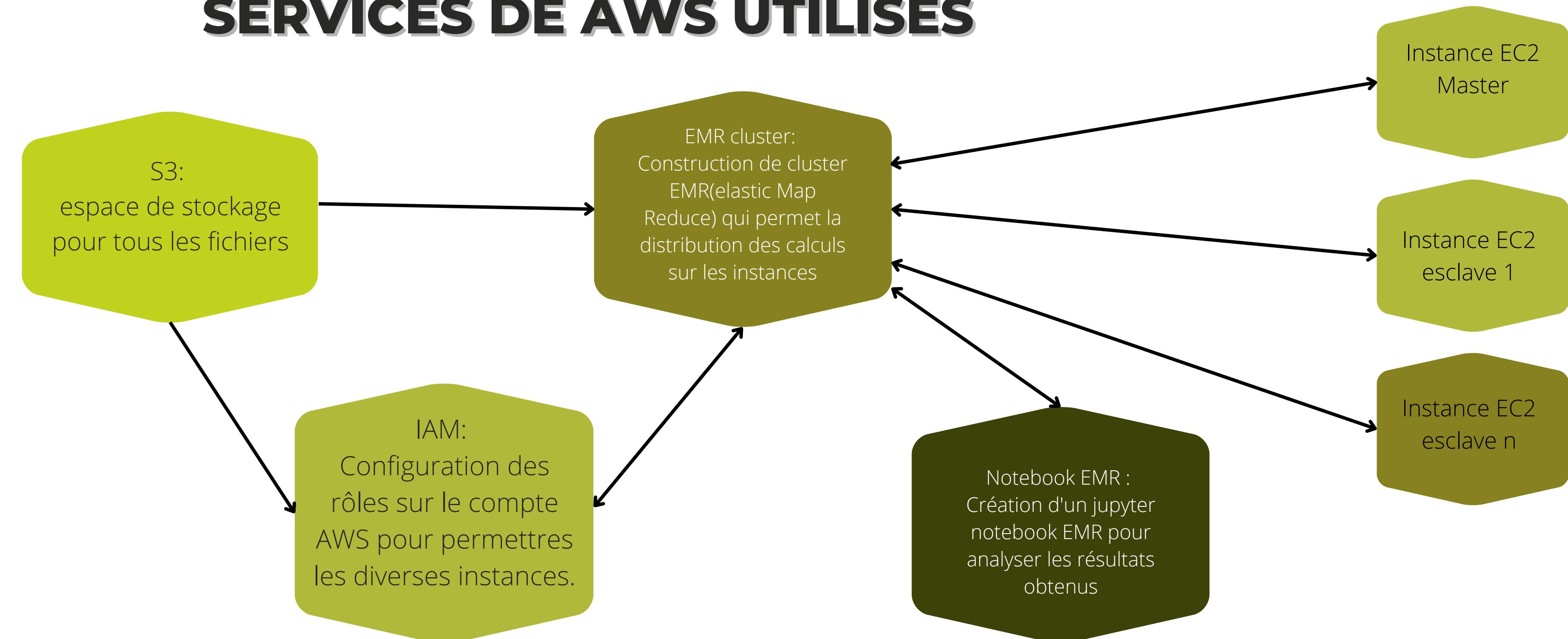
Traitements des
données ,
preprocessing
Jupiter Notebook
(EMR Workplaces)

ACP

Exportation en
csv des
features
obtenus dans
S3

- Extraction features avec vgg16
- Standardisation standard scaler
- PCA

SERVICES DE AWS UTILISÉS



Conclusions

Mise en place environnement Big Data :

AWS : EC2, IAM, EMR
Spark

Passage à l'échelle pour le développement :

Stockage : S3 (illimité)
Pas de modification du
code, juste ajout
d'instances pour le calcul



Conclusions Recommandations

Choix d'un modèle adapté avec le client.

**Avec le déploiement, bien penser l'évolution
de l'infrastructure en fonction des besoins et
du budget alloué.**

Merci de votre attention