

Projet 8

Déployez un modèle dans le cloud



Zeynep Erdem
11-12-2023



Sommaire

- Rappel de la Problématique
- Présentation des données
- Le processus de création de l'environnement Big Data
- La réalisation de la chaîne de traitement des images dans le Cloud
- Démonstration d'exécution du script PySpark sur le Cloud
- Conclusion











































Fruits!

Rappel de la Problématique

- ❖ La start-up "Fruits!" se concentre sur l'AgriTech et la préservation de la biodiversité des fruits.
- ❖ Développer des solutions pour la récolte des chaque espèce de fruits à l'aide de robots cueilleurs intelligents.

- ❖ Créer une application mobile
 - prendre une photo d'un fruit pour obtenir des informations
 - mettre en place les premières briques de traitement sur cloud
 - une réduction de dimension de type PCA en PySpark
- ❖ Anticiper une augmentation rapide du volume de données.
- ❖ Construire une première version de l'architecture Big Data.
- ❖ Respecter des contraintes du RGPD et minimiser les coûts.

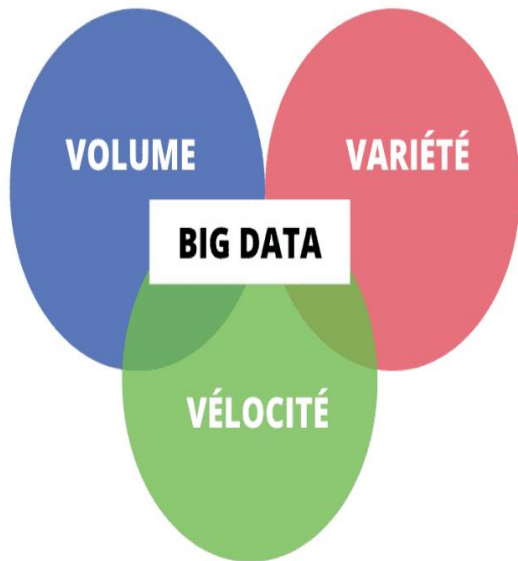
Présentation du Jeu de Données

Red Apple Category 1					
Red Apple Category 2					
Red Apple Category 3					
Red Apple Category 4					
Red Apple Category 5					
Banana					
Orange					
Pomegranate					

- Le nombre total : 90483
- Taille de train set : 67 692
- Taille de test set : 22 688
- Le nombre de classes : 131
- Taille de l'image : 100x100 pixels (jpeg)
- 5 variétés et 10 images par variétés

Décision de l'architecture Big Data

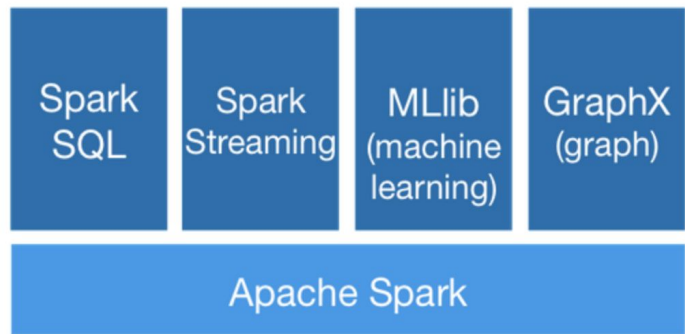
BIG DATA: LES 3Vs



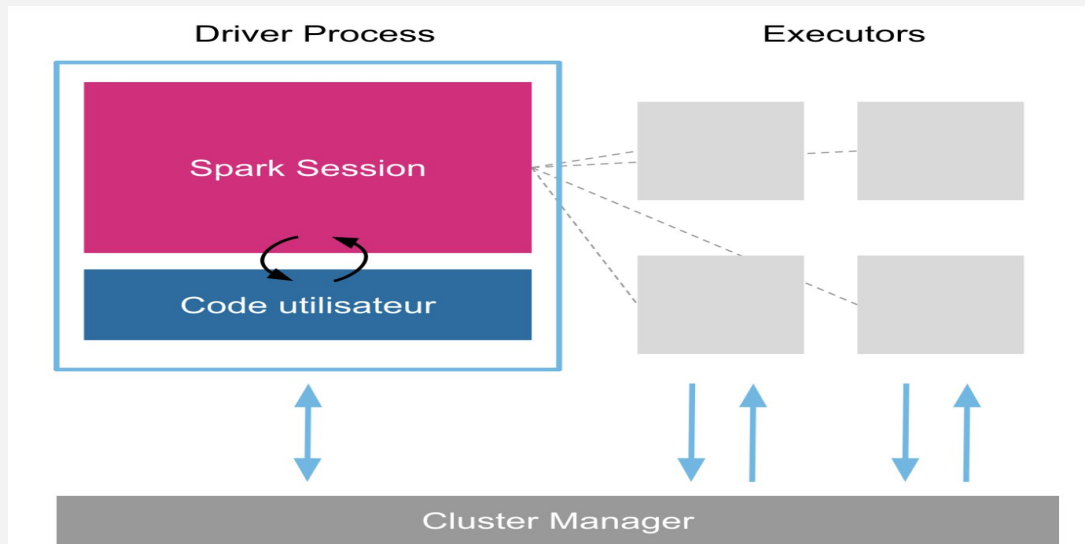
FACTORS	SPARK	HADOOP MAPREDUCE
Speed	100x times than MapReduce	Faster than traditional system
Written In	Scala	Java
Data Processing	Batch/real-time/iterative/ interactive/graph	Batch processing
Ease of Use	Compact & easier than Hadoop	Compact & lengthy
Caching	Chaches the data in-memory & enhances the system performance	Doesn't support caching of data

Calcul distribué - Résistance aux pannes - Augmenter la puissance en augmentant les noeuds - Moins de coût

Fonctionnement de Spark

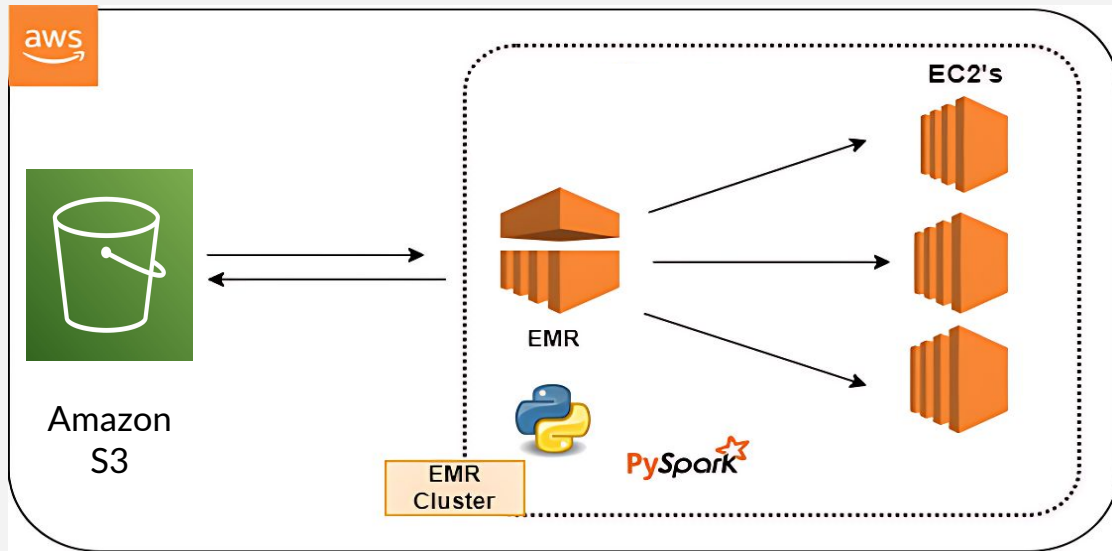
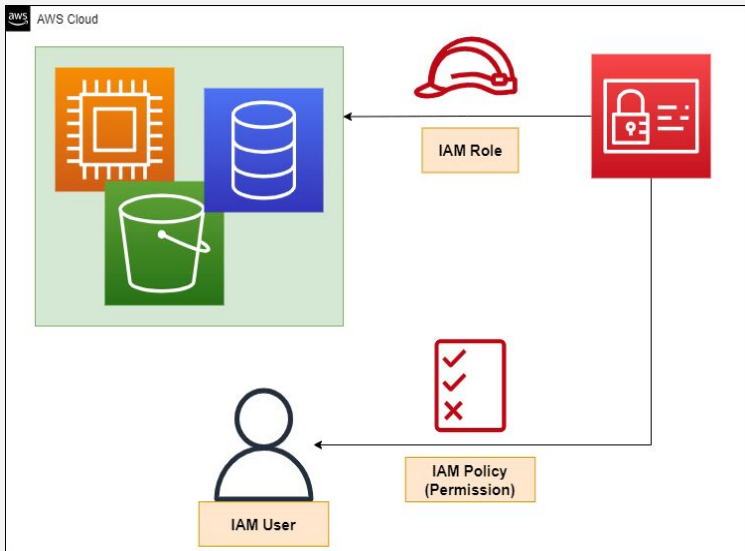


- Spark est implémenté via des API dans plusieurs langages de programmation comme PySpark
- La "lazy evaluation" retarde le calcul des transformations jusqu'à ce qu'une action soit déclenchée.



- Spark emploie un cluster manager qui assure le suivi des ressources disponibles.
- Le driver process est responsable de l'exécution le programme à travers les exécuteurs pour accomplir une tâche donnée.

Amazon Web Services



- Accès Sécurisé via IAM (Identity and Access Management)
- La gestion des utilisateurs et de leurs droits s'effectue via le service IAM

- Images
- Notebook
- Résultats
- RGPD(eu-west-3)

- Version emr-6.14.0
- 1 Primary 2 Core
- EC2 instance type
- m5Xlarge 4v core 16 gb memory

Configuration du serveur EMR et Création du tunnel ssh

Cluster info

Cluster ID

j-ZALT9AQ0DOSI

Cluster configuration

Instance groups

Capacity

1 Primary | 2 Core | 0 Task

- Nom et applications
- Groupes d'instances
- Actions d'amorçage (bootstrap)
- Sécurité et paire de clés EC2
- Instanciation du serveur



Configuration de FoxyProxy



Exécution du code



```
zeynepdem — hadoop@ip-172-31-12-136:~ — ssh -i ~/user_p8.pem -D...
--l  --l-  )
-l (  /
---\N---l---l
Amazon Linux 2 AMI

https://aws.amazon.com/amazon-linux-2/
38 package(s) needed for security, out of 63 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEEE MMMMMMM MMMMMMM RRRRRRRRRRRRRRR
E:::EEEEEEEEEEEEEEEE M:::MM M:::MM R:::RRRRRRRRRRRR
EE:::EEEEEEEEEEEEEE M:::MM M:::MM R:::RRRRRRRRRRRR
E:::E EEEEE M:::MM M:::MM R:::RRRRRRRRRRRR
E:::E M:::MM M:::MM M:::MM R:::RRRRRRRRRRRR
E:::EEEEEEEEEEEE M:::MM M:::MM R:::RRRRRRRRRRRR
E:::E EEEEE M:::MM M:::MM R:::RRRRRRRRRRRR
EE:::EEEEEEEEEEEE M:::MM M:::MM R:::RRRRRRRRRRRR
E:::EEEEEEEEEEEEEE M:::MM M:::MM R:::RRRRRRRRRRRR
EEEEEEEEEEEEEEEEEE MMMMMMM MMMMMMM RRRRRRRRRRRRR
[hadoop@ip-172-31-12-136 ~]$
```


Les étapes de traitement des images

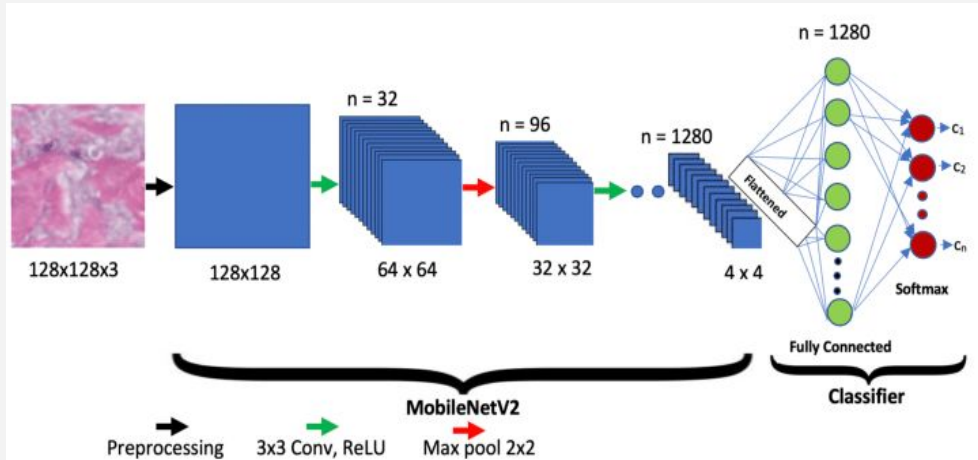


- Redimension 224x224x3
- MobileNetV2
- Utilisation de pandas UDF
- Extractions de features
- Convertir l'array en Vector

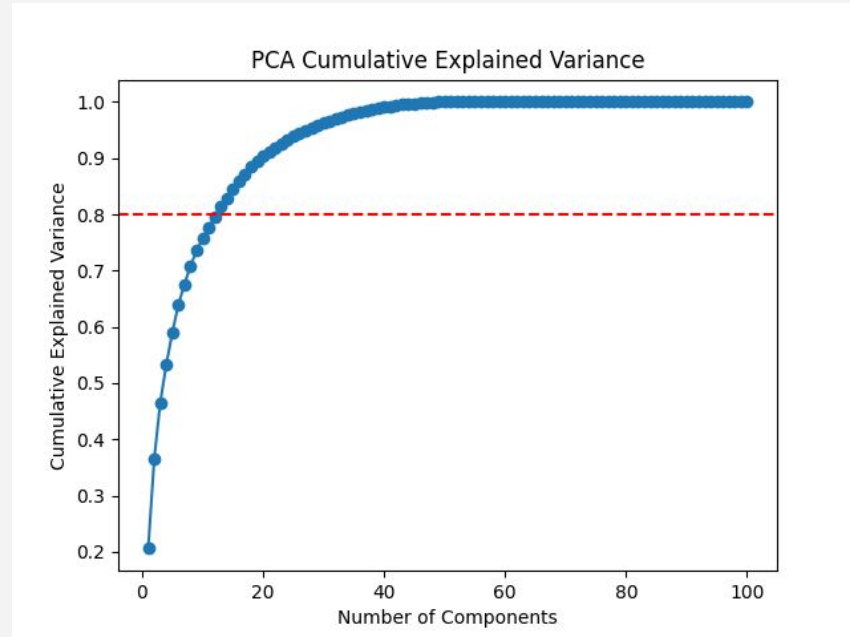
- Standard Scaler
- 1280 features → 13
- 80% de variance expliquée

- Exportation des features pca en csv et en parquet

MobileNetV2 et Réduction de dimension

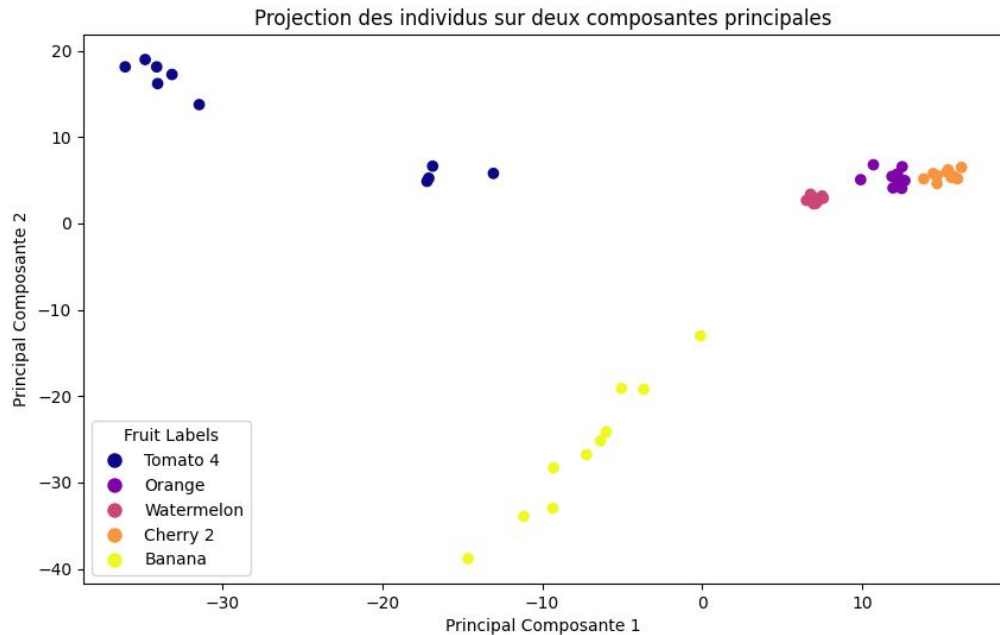


- Architecture de réseau neuronal convolutif (CNN)
- Vision par ordinateur sur des appareils mobiles.
- Elle a été développée par Google et publiée en 2018
- Une architecture légère et efficace
- Bottlenecks utilisent des opérations de séparation des canaux (depthwise separable convolutions)
- Réduire le nombre de calculs sans sacrifier significativement les performances.
- 1280 features en sortie



On peut expliquer 80 % de la variance avec 13 composantes.

Projection des individus sur deux composantes principales



On constate que les classes des fruits ont été bien distinguées.

Facturation et Cost Explorer

New cost and usage report

Recent reports ▼

Save to report library

Cost and usage graph [Info](#)

Total cost

\$6.66

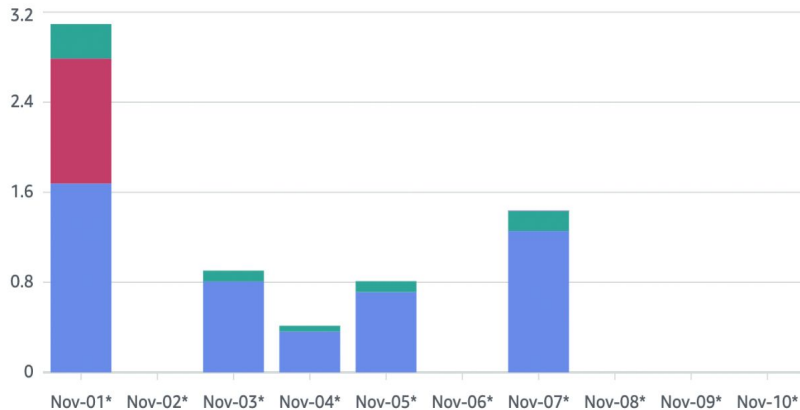
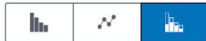
Average daily cost

\$0.67

Service count

8

Costs (\$)



■ EC2-Instances ■ Tax ■ Elastic MapReduce ■ S3 ■ EC2-Other ■ VPC ■ CloudWatch
■ Key Management Service

Report parameters



▼ Time

Date Range

2023-11-01 — 2023-11-10

Granularity

Daily ▼

▼ Group by

Dimension

[Clear](#)

Service ▼

||

▼ Filters [Info](#)

Applied filters (1)

[Clear all](#)

Service

[Clear](#)

Choose services ▼

Linked account

[Clear](#)

Choose linked accounts ▼

Region

[Clear](#)

Choose regions ▼

- Résiliation de l'instance EMR
- Cloner le serveur EMR (si besoin)

Démonstration d'exécution du script et liens pour S3



Liens IAM user

Le lien user iam aws : <https://396972527913.signin.aws.amazon.com/console>

User name : visitor

Password : visitor.P8

Conclusion et Recommandations



- Nous avons réalisé ce projet en deux temps en locale et en cloud (Spark et AWS)
 - Nous avons travaillé sur un plus petit jeu de donnée.
 - Pour valider le bon fonctionnement de la solution et diminuer les coûts.
 - Nous avons fait le choix de réaliser du transfert learning à partir du modèle MobileNetV2.
 - Nous avons effectué standardisation et réduction de dimension des features (PCA)
 - Les résultats ont été enregistrés sur s3 bucket en plusieurs partitions au format "parquet" et en "csv".
-
- On peut utiliser les features pour entraîner un modèle de classification.
 - Il faut surveiller les coûts et peut être trouver les solutions cloud moins cher et Made in France comme OVHcloud.

A close-up photograph of a dark, woven basket tipped over, spilling a variety of fresh berries onto a rustic wooden surface. The berries include bright red strawberries with green leaves, small blueberries, raspberries, and blackberries. The scene is lit with soft, natural light, creating a warm and inviting atmosphere. The word "MERCI" is overlaid in white, sans-serif capital letters on the right side of the image.

MERCI