
# Projet 9

## Réalisez un traitement dans un environnement Big Data  sur le Cloud

**Rédigé et présenté par**

**Etudiant :** Gassuc Cédric

**Mentor :** Mohammed El Abridi

# Plan

1- Choix de l'architecture cloud

2- Rôle dans l'architecture Big Data

3- Démarche de mise en oeuvre de l'environnement Big Data(EMR)

4- Etapes de la chaîne de traitement PySpark

5- Conclusion

# Comparaison des Top 3 Environnements Cloud

## AWS (Amazon Web Services)

Leader du marché avec une large gamme de services, haute fiabilité et une forte communauté d'utilisateurs.

## Microsoft Azure

Intégration parfaite avec les produits Microsoft, adapté aux entreprises et bon support hybride.

## Google Cloud Platform

Fort dans les services d'analyse et machine learning, avec une infrastructure rapide et scalable.

# Pourquoi AWS ?

## Location à la Demande

Ressources disponibles instantanément selon les besoins.
Élasticité automatique en fonction de la charge.

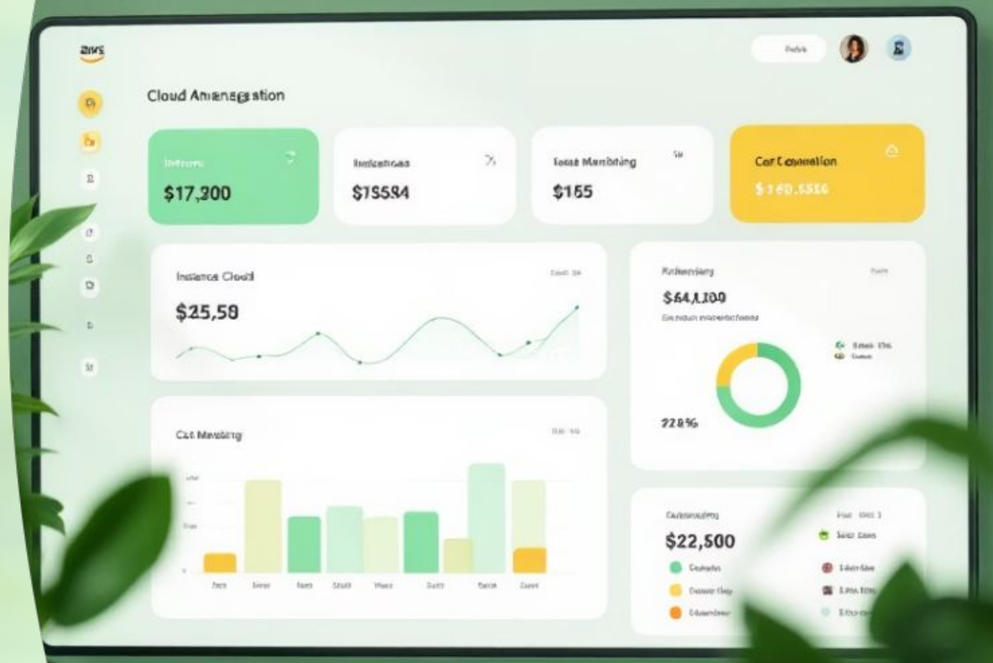
## Coût Maîtrisé

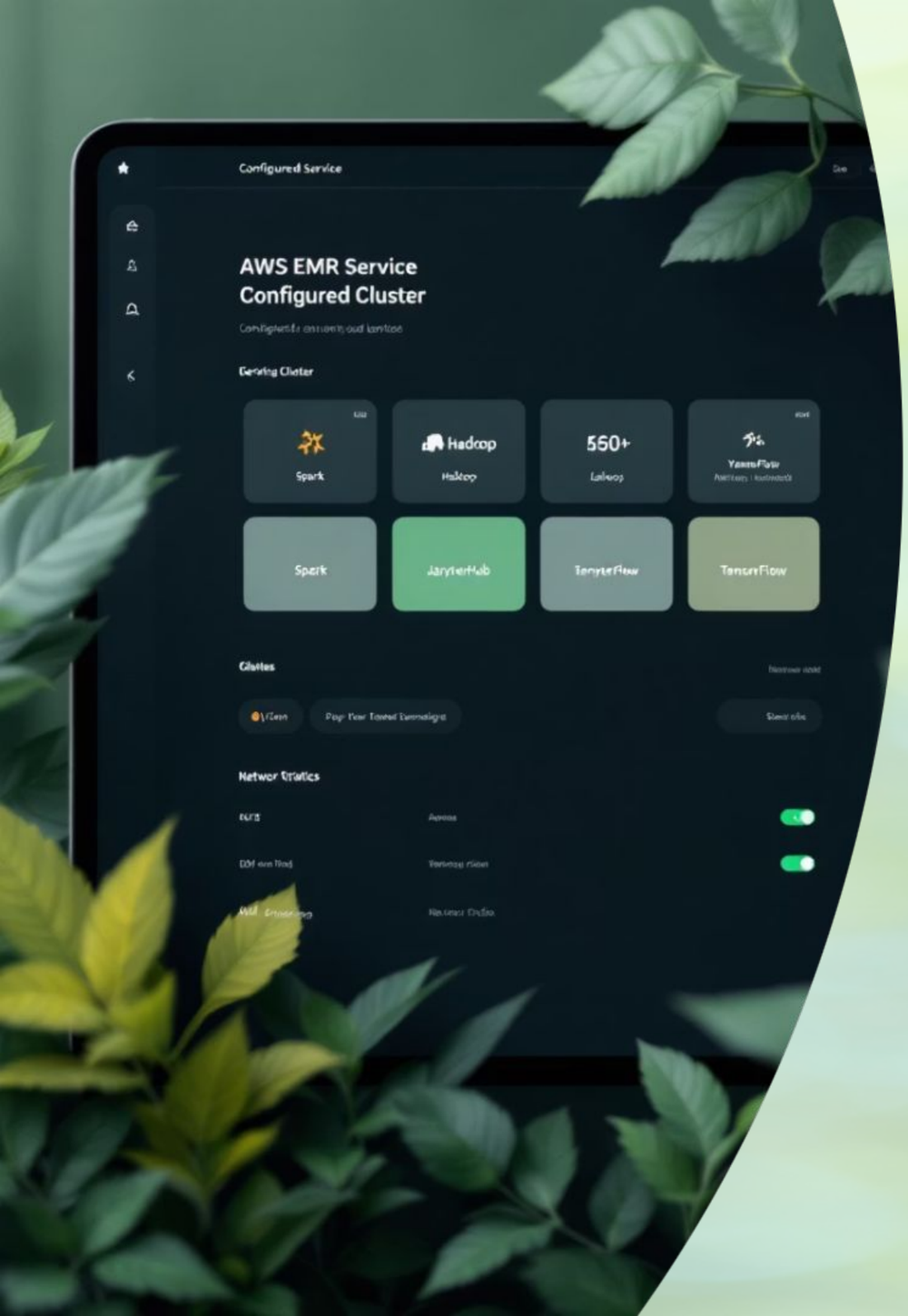
Paiement uniquement pour les ressources utilisées.
Options d'optimisation des dépenses disponibles.

## Service EC2 (IAAS)

Contrôle granulaire sur les instances.

Large choix de configurations matérielles virtualisées.






# Utilisation d'AWS EMR (PAAS)

### Apache Spark

Moteur de traitement distribué ultrarapide.

### Hadoop

Écosystème pour le stockage et traitement massif.

### JupyterHub

Environnement notebook interactif pour données.

### TensorFlow

Framework d'apprentissage automatique distribué.

# Simplicité & Flexibilité

## Compatibilité du Code

Notebook identique entre local et cloud.

Aucune modification nécessaire.

## Passage à l'Échelle

Traitement de volumes massifs de données.

Ajout dynamique de nœuds.

## Gestion des Packages

Installation automatique sur tous les nœuds.

Environnement cohérent.

# Modèle Choisi : MobileNetV2

### Architecture Légère

Optimisé pour les environnements à ressources limitées.

### Inférence Rapide

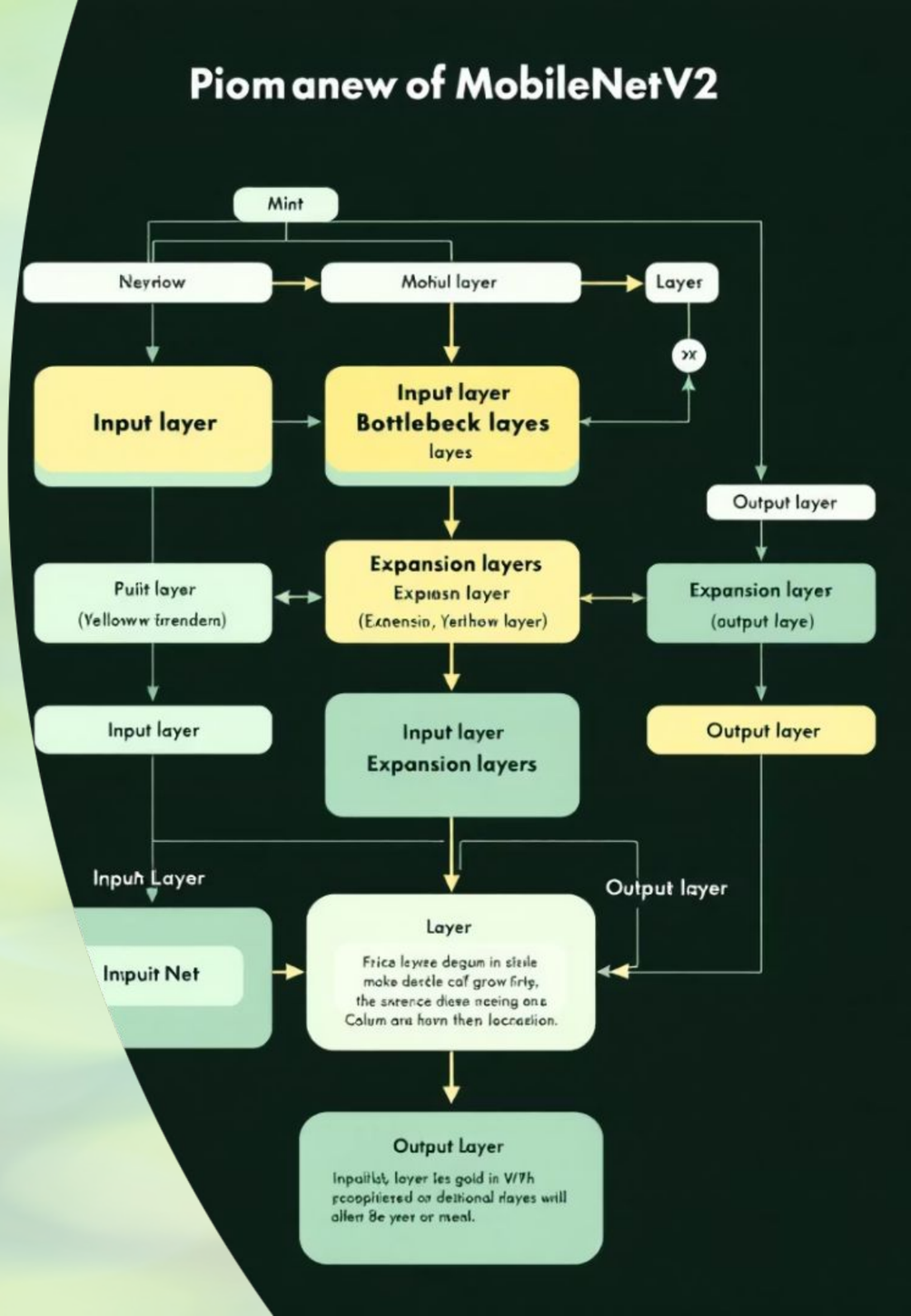
Temps de traitement réduit pour chaque image.

### Vecteur Compact

Représentation efficace pour le traitement distribué.

### Intégration Spark

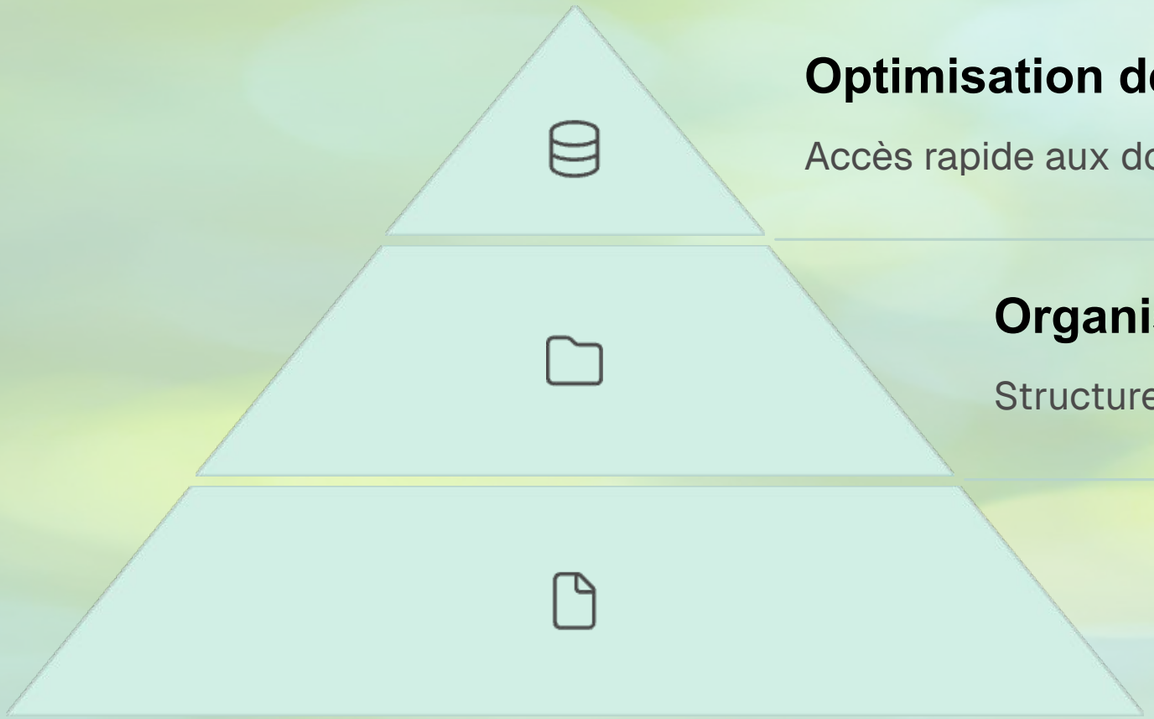
Compatibilité native avec l'écosystème Spark.



# Stockage Optimisé

- Données projet dans Amazon S3
- Espace illimité, coût en fonction de l’usage
- Exploitation efficace des données

# Stockage des Résultats



### Optimisation des requêtes

Accès rapide aux données spécifiques

### Organisation en partitions

Structure hiérarchique efficace

### Format Parquet

Compression et schéma intégrés

# Gestion efficace des résultats

Dans AWS EMR, la gestion des résultats se transforme en un moteur d’optimisation, garantissant des requêtes ultra-rapides grâce à un accès ciblé et instantané aux données.

La structuration en partitions hiérarchiques révolutionne le traitement des volumes massifs, rendant la récupération des données fluide et performante.

L’adoption du format Parquet, avec sa compression intelligente et son schéma intégré, maximise l’efficacité du stockage tout en minimisant l’espace utilisé.

# Conclusion