

FRUITS !

DÉPLOYER UNE
CHAÎNE DE
TRAITEMENT
DANS LE CLOUD

PLAN



Cadre



Le Big Data



Architecture
sélectionnée



Chaîne de
traitement



Questions -
Réponses

CADRE



Fruits!

L'IA AU SERVICE DE L'AGRITECH

🎯 Mettre en place la chaîne de prétraitement des données d'un moteur de classification pour faire face à une explosion de la demande

- Architecture et traitements Big Data

🧩 Jeu d'images de fruits et légumes

- ✓ 90483 images
- ✓ 131 classes

LE BIG DATA

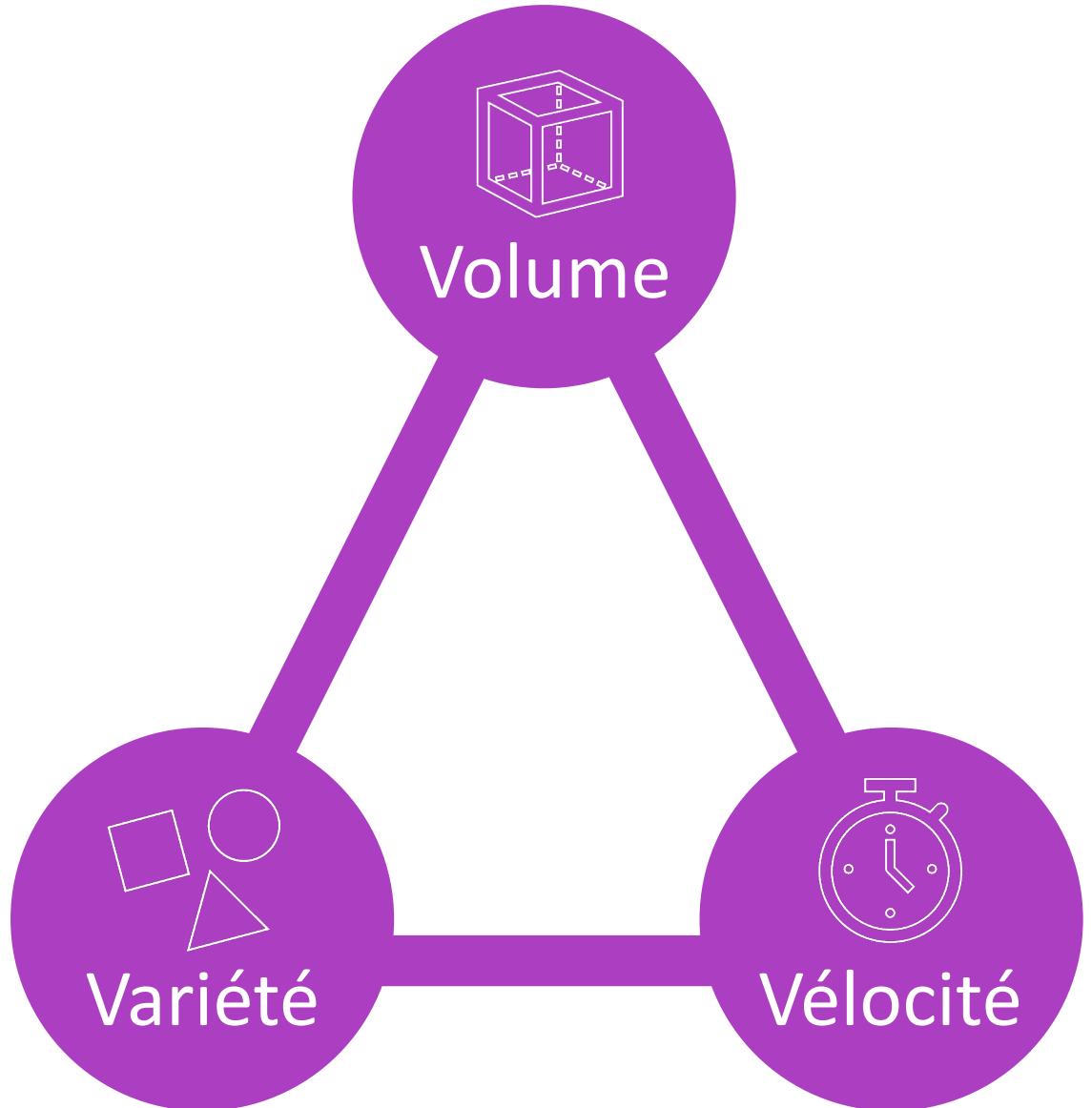
CONCEPTS – CONTRAINTES – SOLUTIONS



CONCEPTS

LE BIG DATA, C'EST QUOI ?

- Flux de données massifs riches en potentiel
- Challenges en termes de :
 - Stockage
 - Exploitation



CONTRAINTE

LES 3 V

- **Volume**
 - Des volumes variables et importants doivent pouvoir être accommodés
- **Variété**
 - Des données variées et potentiellement variables doivent pouvoir être accommodées
- **Vélocité**
 - Les données doivent être rapidement et facilement exploitables, quel que soit le volume

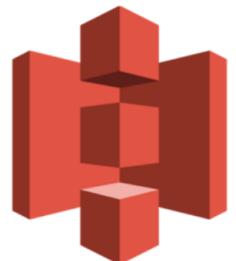
SOLUTIONS DE STOCKAGE BIG DATA



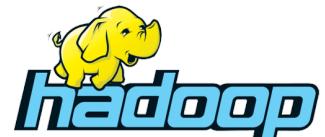
Google
Cloud Storage



Microsoft Azure
Blob Storage



Amazon Web Services
S3



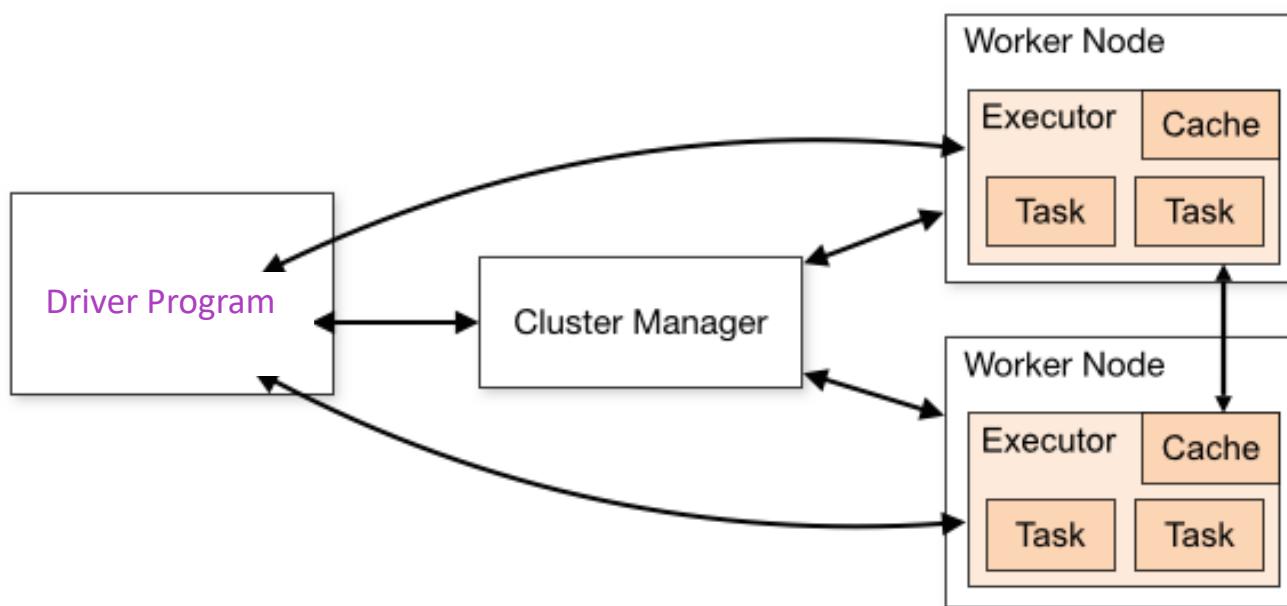
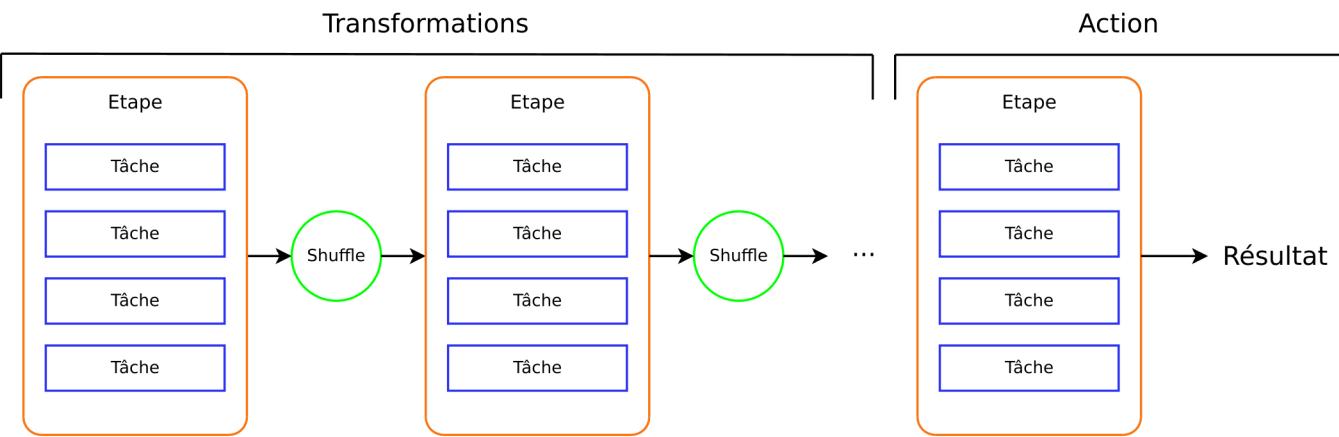
Apache
Hadoop

SOLUTION : UNE INFRASTRUCTURE DISTRIBUÉE

LE STOCKAGE DISTRIBUÉ

- Volume
 - ✓ Passage à l'échelle possible
 - ✓ Une bonne compression
- Variété
 - ✓ Capacité d'évolution
- Vélocité
 - ✓ Partitionnement
- Résilience
 - ✓ Redondance
 - ✓ Tolérance aux pannes

CLUSTER DE CALCUL - FONCTIONNEMENT



SOLUTION : UNE INFRASTRUCTURE DISTRIBUÉE LES CALCULS DISTRIBUÉS

- Vélocité
 - ✓ Partitionnement
 - ✓ Optimisation des opérations de lecture / écriture
 - ✓ Passage à l'échelle
- Résilience
 - ✓ Tolérance aux pannes

SOLUTION : UNE INFRASTRUCTURE DISTRIBUÉE

RÉCAPITULATIF

- Stockage distribué
- Calculs distribués et optimisés
- Infrastructure dédiée :
 - ✓ niveau logiciel : systèmes de fichiers et frameworks spécifiques
 - ✓ niveau matériel : agilité des serveurs

ARCHITECTURE SÉLECTIONNÉE

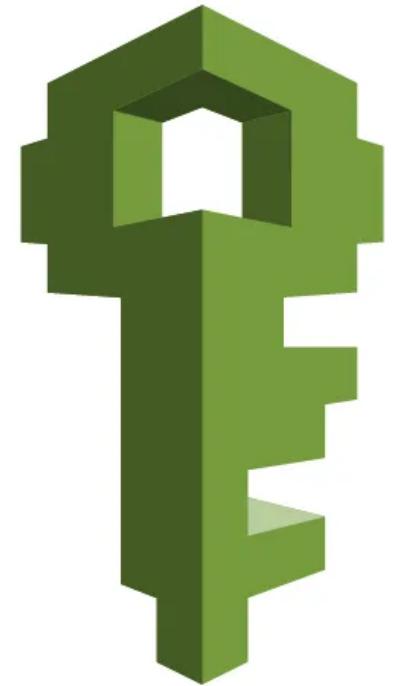
STOCKAGE – SÉCURISATION – TRAITEMENTS – TECHNOLOGIES



STOCKAGE

S3 PAR AWS

- Pas de limite de place
 - ✓ Scalabilité
- RéPLICATION possible sur plusieurs datacenters
 - ✓ Fiabilité
- Droits d'accès
 - ✓ Sécurité



SÉCURISATION

CLÉS D'ACCÈS + RÔLE IAM

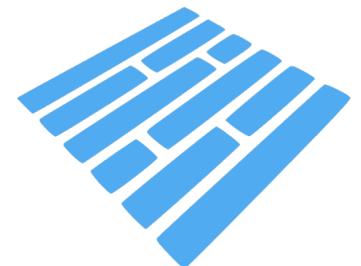
- Combinaison clé + utilisateur :
- Création de clés d'accès à S3
- Création d'un utilisateur ayant des permissions spécifiques sur S3 après utilisation des clés



TRAITEMENTS

EC2 PAR AWS

- Facile d'ajouter des instances au besoin
- ✓ Scalabilité
- Accès sécurisé



Parquet



docker

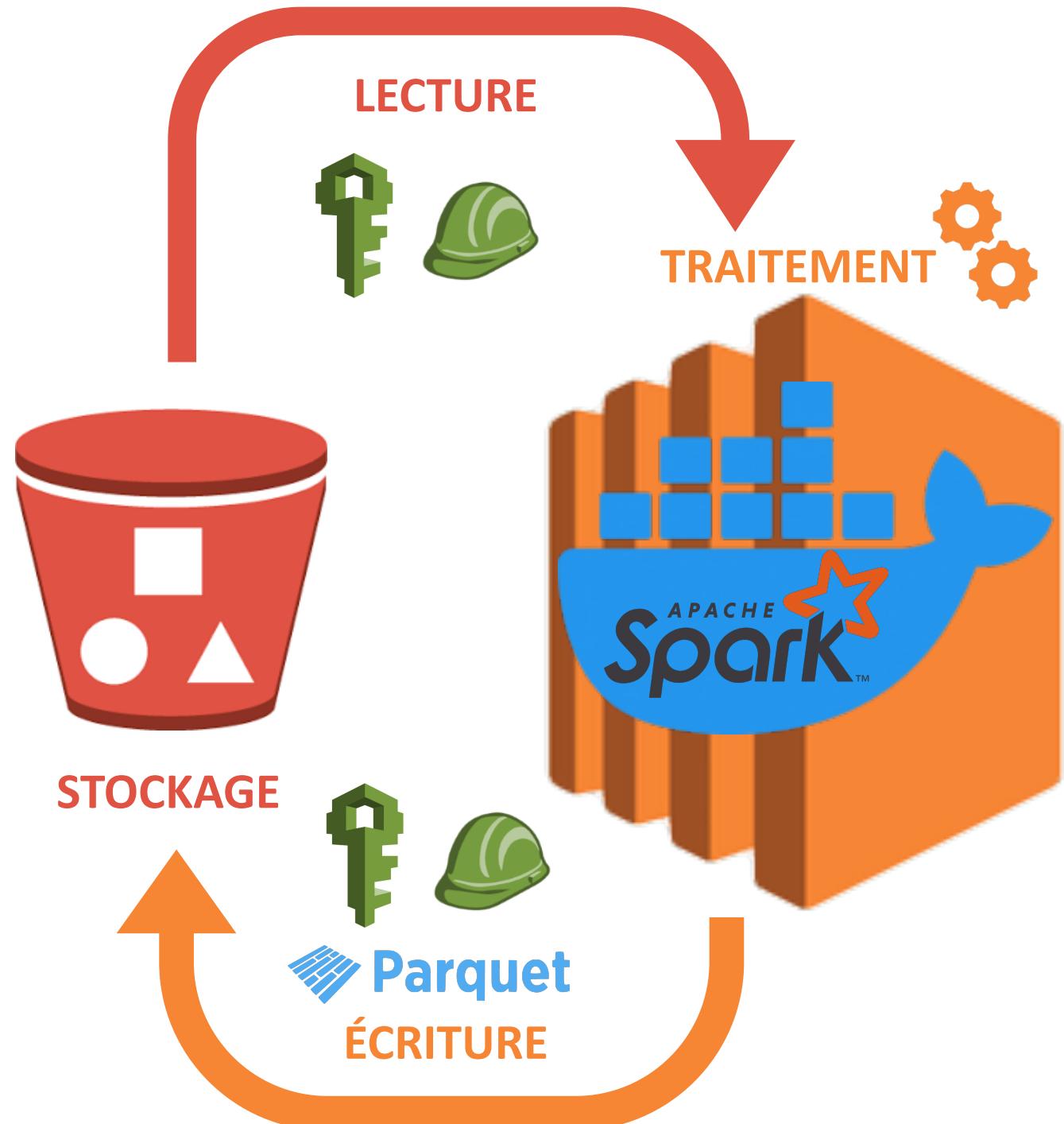
TECHNOLOGIES

SPARK ET DOCKER

- **Spark** : paralléliser les calculs
- **Parquet** : compresser et encoder les données traitées
- **Docker** : containeriser l'application

CHAÎNE DE TRAITEMENT

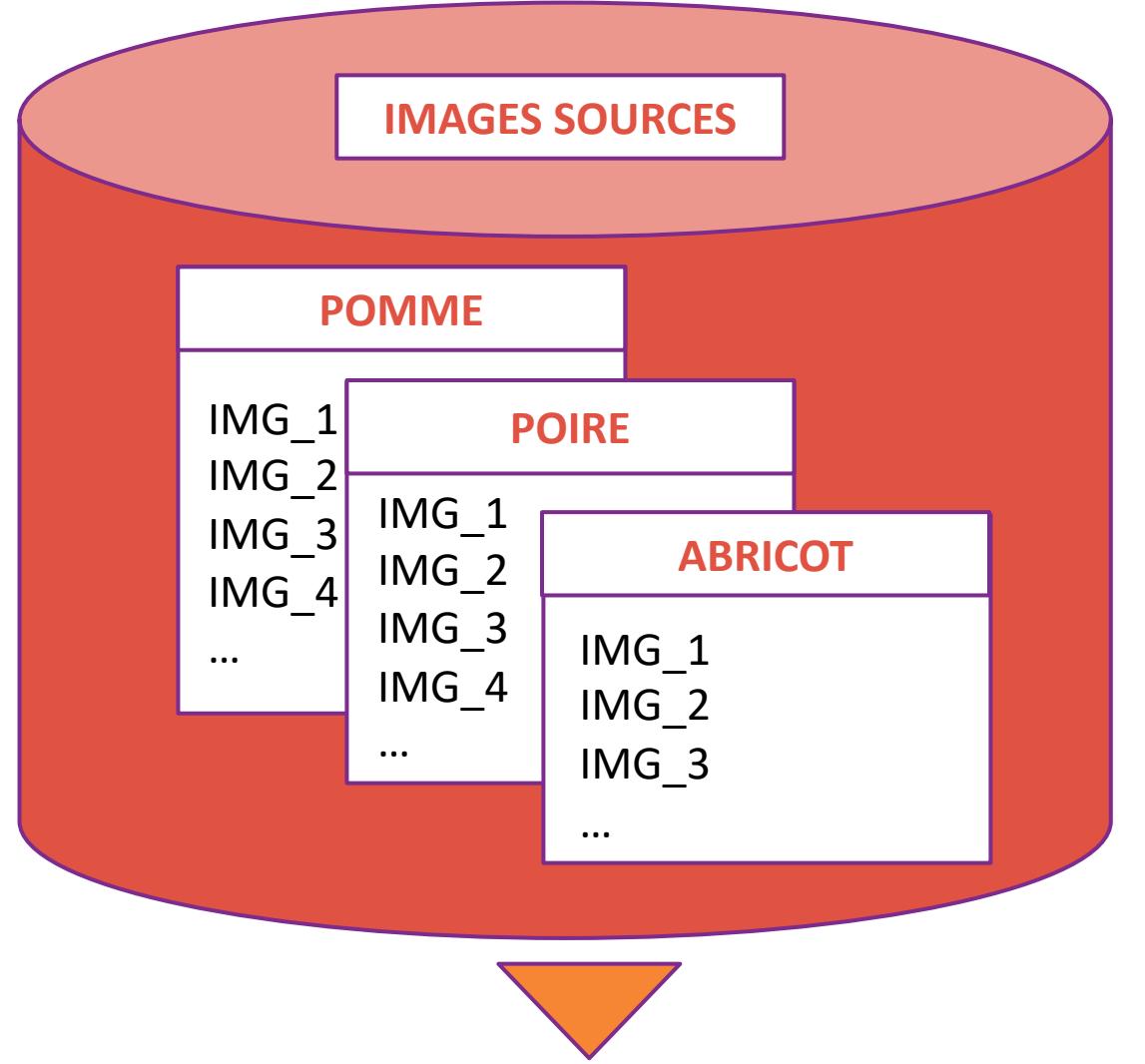
MISE EN PLACE – CHARGEMENT & LABELLISATION – RÉDUCTION DE DIMENSIONS – EXPORT



MISE EN PLACE

INFRASTRUCTURE

1. Stockage des données sur S3
2. Création clés + rôle IAM pour accès protégé
3. Création instance EC2
4. Création container Docker
5. Installation Spark



CHEMIN	CONTENU	CATÉGORIE
IMAGES_SOURCES/POMME/IMG_1	[FF D8 FF E0 00 1...]	POMME
IMAGES_SOURCES/POIRE/IMG_10	[FF D8 FF E0 00 1...]	POIRE
IMAGES_SOURCES/ABRICOT/IMG_19	[FF D8 FF E0 00 1...]	ABRICOT

CHARGEMENT & LABELLISATION

EXTRACTION DES DONNÉES

TRAITEMENTS PARALLÉLISÉS SPARK SQL :

- Images classées par catégorie
- Import des données :
 - Au format binaire pour plus de facilité dans la manipulation
 - Avec labellisation suivant le dossier de l'image

RÉDUCTION DE DIMENSIONS

EXTRACTION DES FEATURES

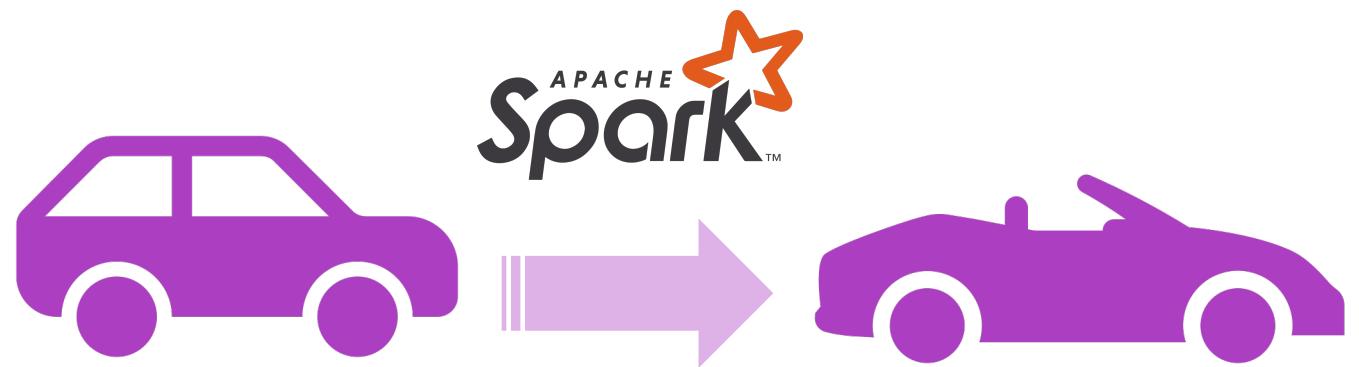
TRAITEMENTS PARALLÉLISÉS SPARK SQL :

- Réduire les images à leurs features les plus importantes pour la classification
- Approche Transfer Learning pour extraction de ces features

Application à nos images pour extraction des features les plus intéressantes

Modèle de deep learning pré-entraîné à la classification d'images

Suppression dernière couche responsable de la classification



COMPARAISON

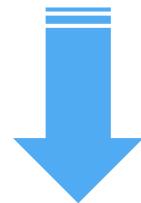
L'EFFET SPARK

TRAITEMENT PARALLÉLISÉ SPARK :

- Traitements en local sur machine quadricœur :
 - 1 cœur : 16 min
 - 4 cœurs : 4 min
- ✓ Temps de traitement divisés par le nombre de cœurs



Parquet



S3

EXPORT DES DONNÉES ENREGISTREMENT

TRAITEMENT PARALLÉLISÉ SPARK :

- Données sauvegardées sur S3 pour exploitation future pour l'apprentissage d'un nouveau modèle
- ✓ Au format Parquet pour une exploitation optimisée en mode distribué conçue pour les données massives

DÉMO

Bucket S3 (stockage)

Instance EC2 (gestion)

Instance EC2 (notebook)

CONCLUSION



■ ARCHITECTURE SÉLECTIONNÉE

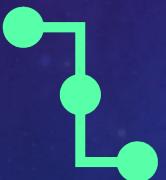
- Stockage S3 avec sécurisation via clés + rôle IAM
- Traitements dans conteneur Docker avec Spark lancé sur instance EC2

Parallélisation des calculs

Optimisation des lectures / écritures

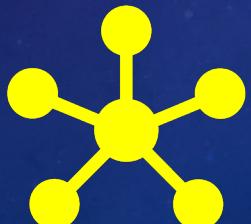
Sécurité des données

Scalabilité



■ CHAÎNE DE TRAITEMENT

- Jeu de données image labellisé et featurisé via transfer learning



■ AMÉLIORATIONS POSSIBLES

- Cluster EMR pour passage à l'échelle automatique
- Script Scala pour accélération vitesse de traitement

MERCI POUR VOTRE ATTENTION



AVEZ-VOUS DES QUESTIONS ?