SOUTENANCE PROJET 8

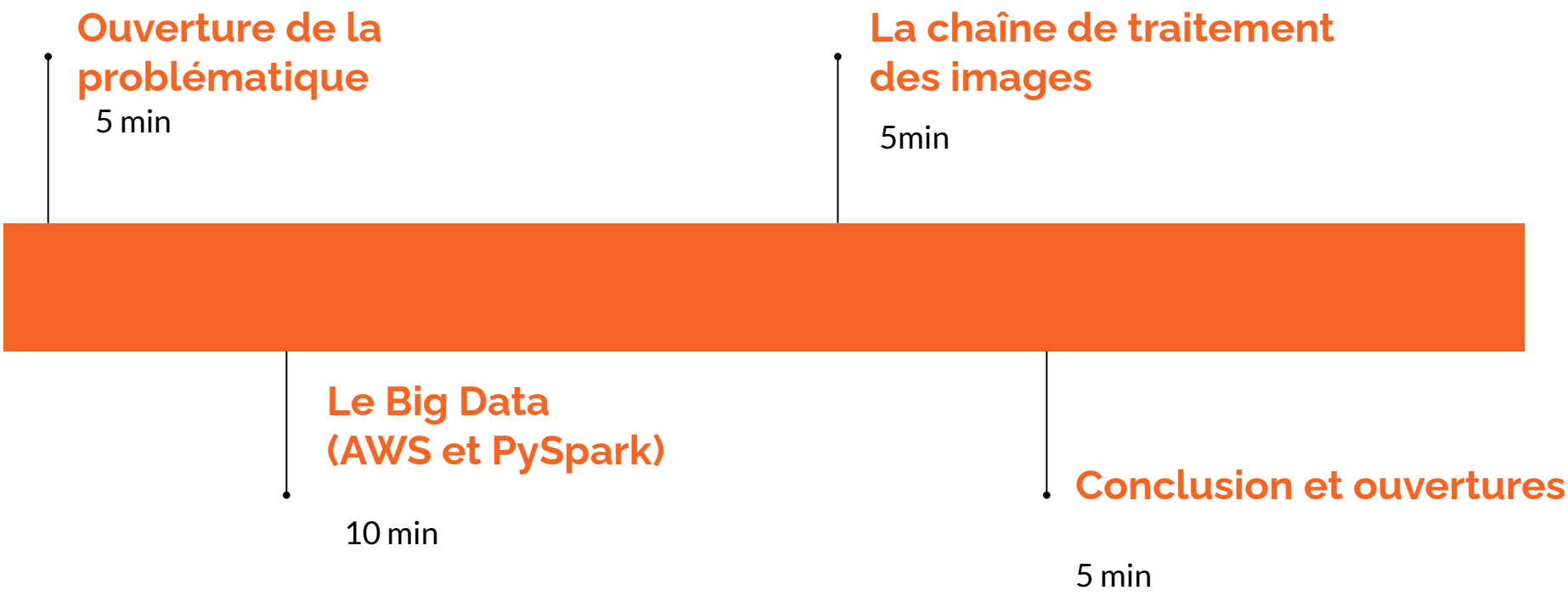
Déployez un modèle dans le cloud

GAUTHIER RAULT
PARCOURS DATA SCIENTIST
CHEZ OPENCLASSROOMS

EVALUATEUR
MONSIEUR Amosse EDOUARD

"Tout va être connecté au cloud et aux données... Tout cela sera médiatisé par des logiciels."
Par Satya NADELLA

ORDRE DU JOUR



OUVERTURE DE LA PROBLÉMATIQUE

Start-up de l'AgriTech «Fruits» souhaite proposer une solution innovante de récolte de fruits avec des robots cueilleurs intelligents.
Pour se faire connaître, elle souhaite mettre à disposition du grand public une application mobile de classification de fruits.

<u>Objectifs principaux</u>

- Anticipation du passage à l'échelle
    - Environnement Big Data
    - Script en PySpark
- Première chaîne de prétraitement des données
    - Preprocessing
    - Réduction de dimension

<u>Moyens pour y parvenir:</u>

- Base de données d'images de fruits

Définition du big data (ChatGPT)

RA définition du big data

Le big data est un terme utilisé pour décrire une grande quantité de données, généralement plus grande que celle que les outils traditionnels de traitement de données sont capables de manipuler et de traiter efficacement. Le big data peut être structuré, semi-structuré ou non structuré, et il peut provenir de différentes sources, telles que les bases de données, les journaux d'événements, les capteurs, les appareils connectés et les médias sociaux. Les entreprises et les organisations utilisent souvent des outils et des technologies spécialisés pour analyser le big data afin d'en extraire des insights et des informations précieuses.

Le Big Data - Les 3V

Volume

Gros volume de données

Vitesse/Vélocité

Les données doivent être traitées et analysées rapidement, parfois en temps quasi réel.

Variété

Différence de format

Véracité

Cohérence, fiabilité et qualité

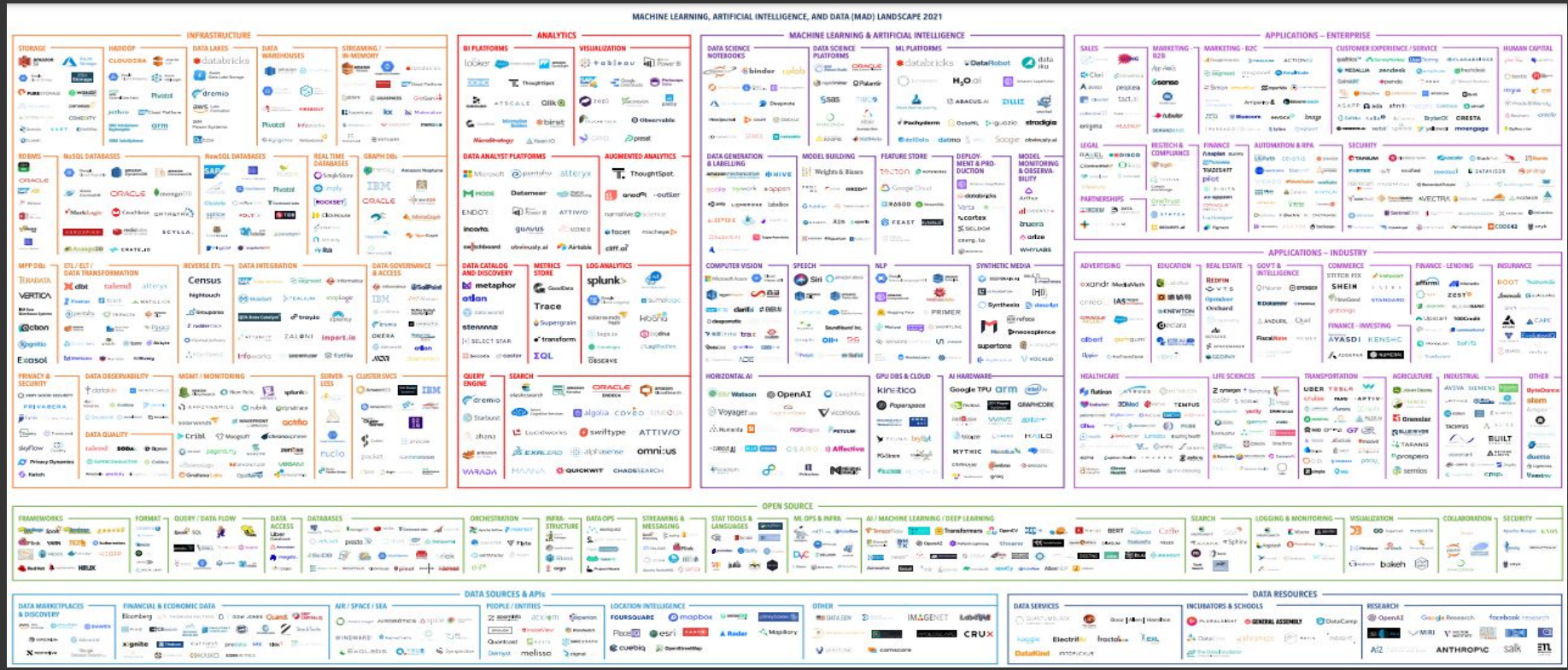
Valeur

Valeur que le Big Data apporte à l'entreprise

Variabilité

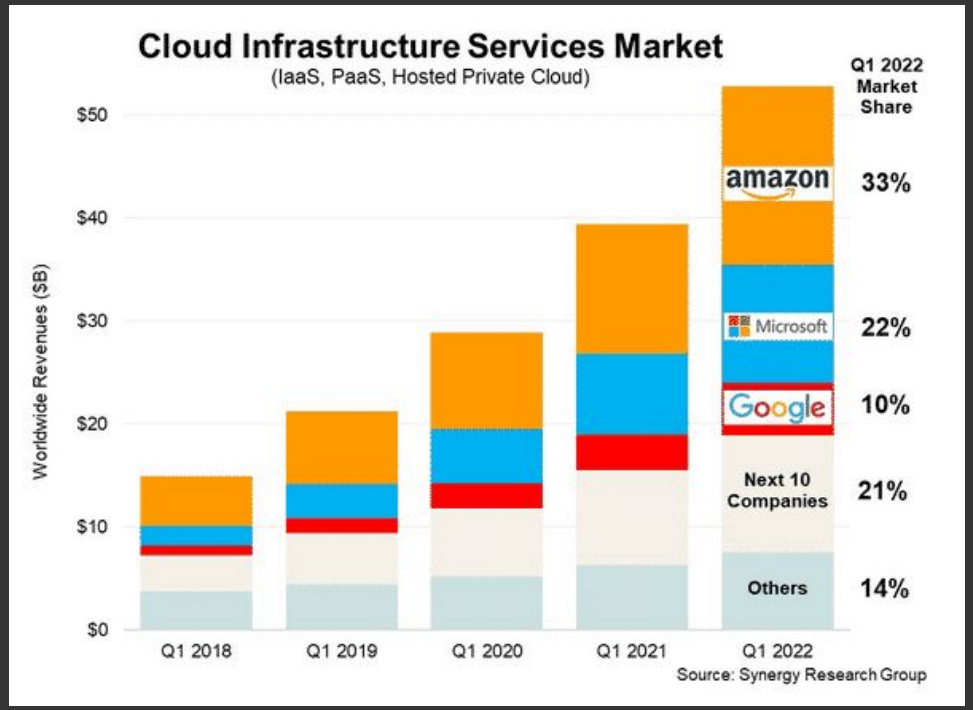
Formatés différemment d'une source de données à une autre

Large écosystème Big data en 2021



Leader in the Gartner 2021 Magic Quadrant for Data Science and Machine Learning Platforms

Leaders des solutions Cloud big data



Fournisseur AWS choisi

Solutions d’AWS sélectionnées

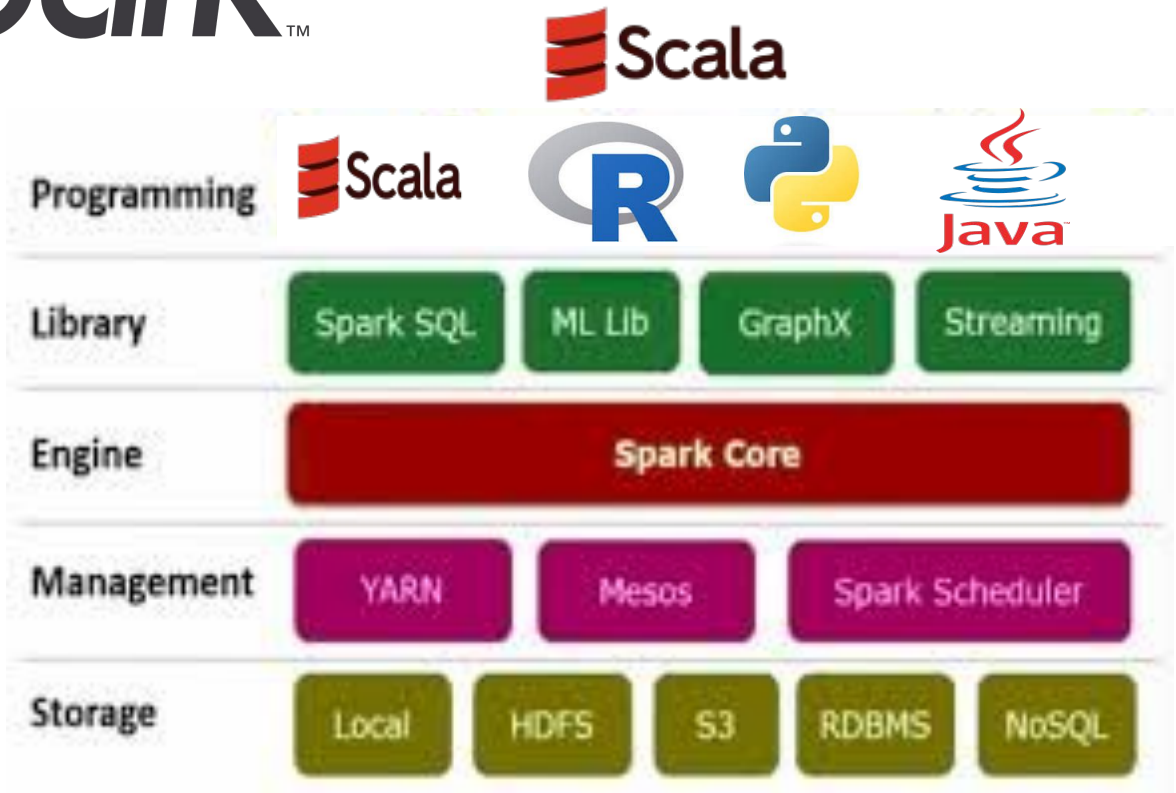
IAM (Identity and Access Management) : pour la gestion des droits d’accès/authentification

S3 (Simple Storage Service) : pour le stockage

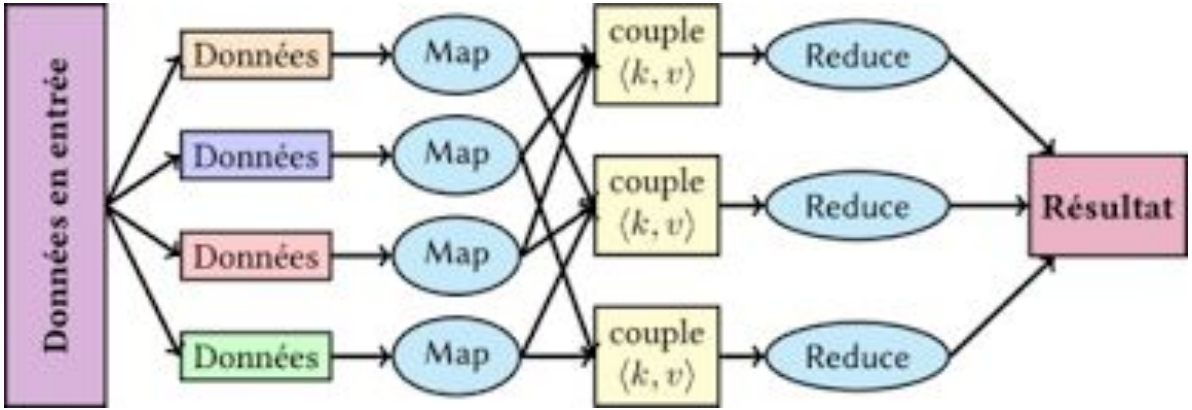
EC2 (Elastic Compute Cloud) : pour les calculs


APACHE
Spark
TM


APACHE
Spark
TM


Scala
Programming
Scala
R
Java
Library
Spark SQL
ML Lib
GraphX
Streaming
Engine
Spark Core
Management
YARN
Mesos
Spark Scheduler
Storage
Local
HDFS
S3
RDBMS
NoSQL

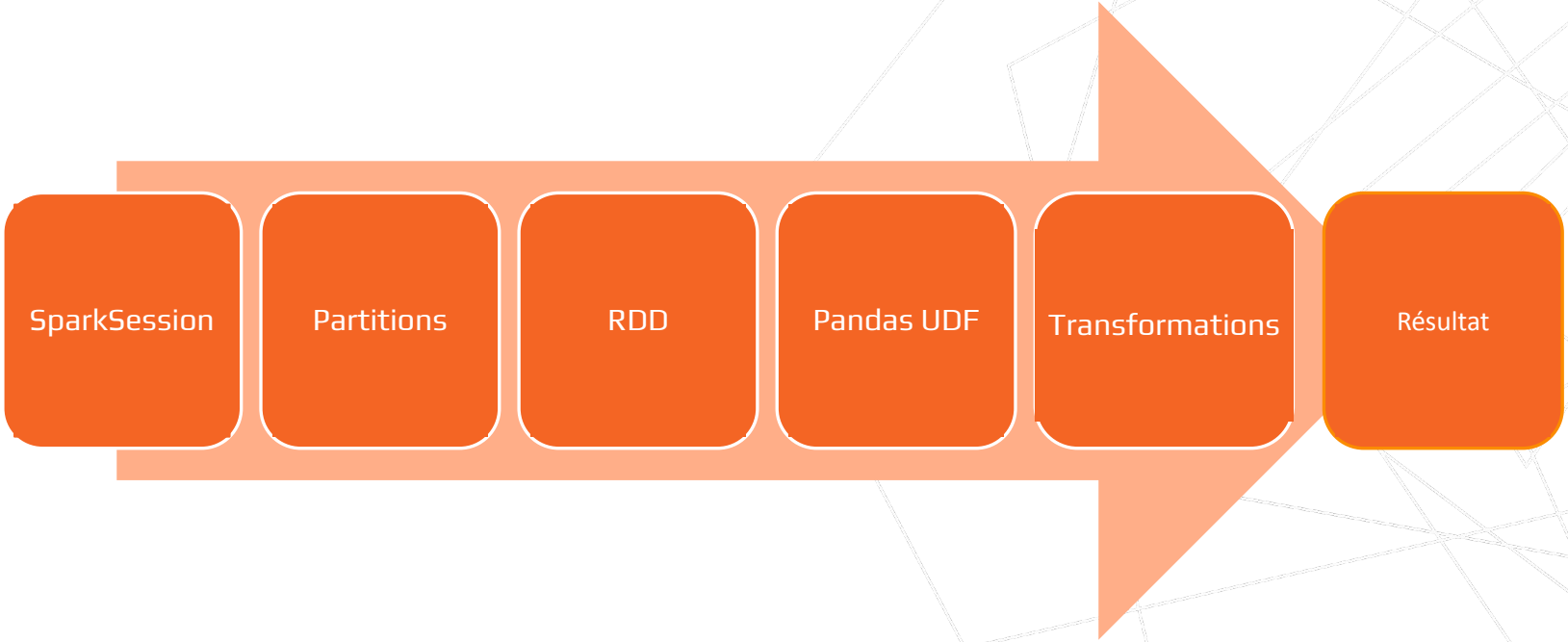
Basé sur Map/Reduce avec traitement “in memory”



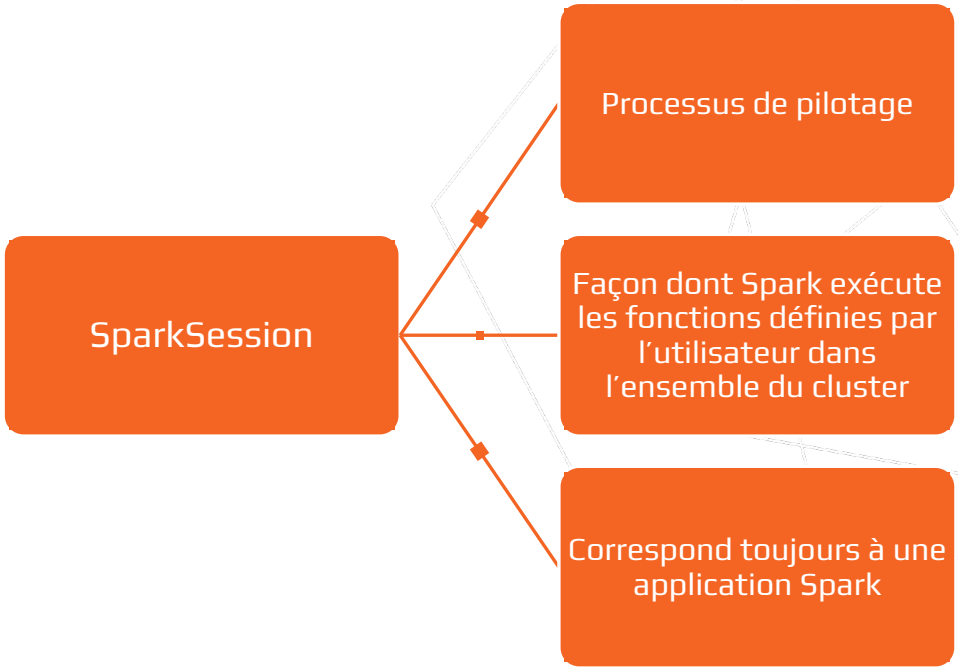
CALCULS DISTRIBUÉS

- Diviser les opérations en micro opérations distribuables entre différentes machines, réalisables en parallèle
- Agréger les résultats sur une même machine

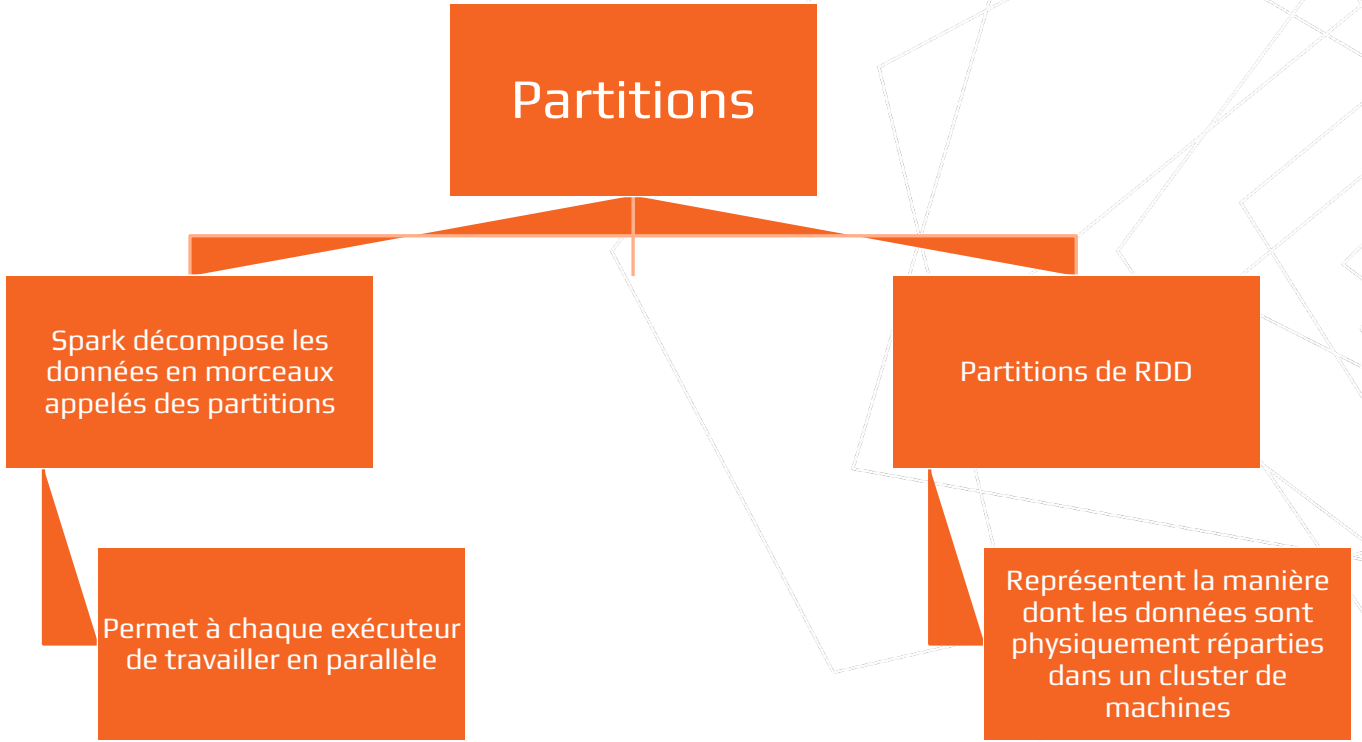
Spark: Concepts principaux



Spark: Concepts principaux



Spark: Concepts principaux

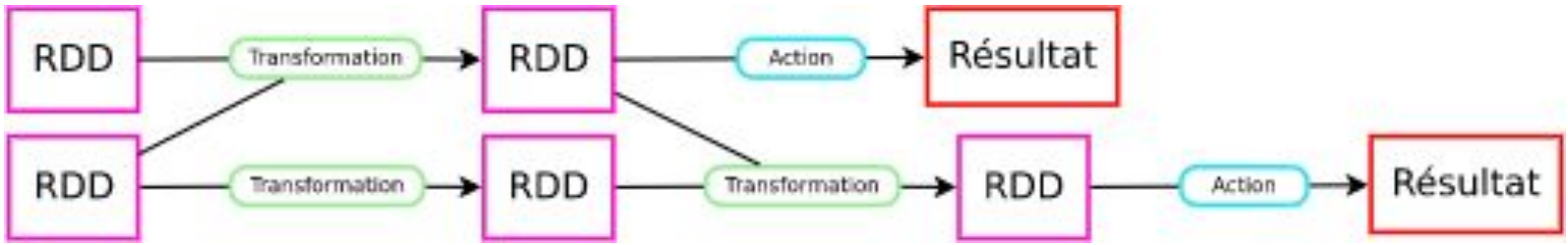


Tolérance aux pannes:

- ➢ Resilient Distributed Datasets (RDD)
  - ○ Division des données en partitions

Graphe Acyclique Orienté (DAG) :

- ➢ Panne : Régénération à partir des noeuds parents
- ➢ Noeuds (RDD ou Résultats) : liés par des actions et transformations



Spark: Concepts principaux

Transformations

★☆☆ **RDD**: Résilient Distributed Dataset
Structures fondamentales de Spark
Objets Immuables

★★☆ Passage d'un RDD en un autre RDD = Transformation

★★★ Ne renvoient aucun résultat

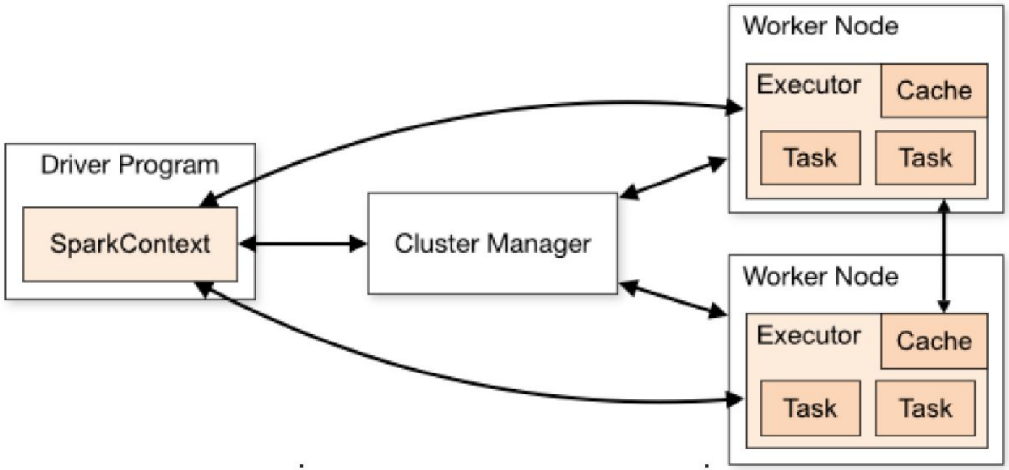
Lazy Evaluation
Spark n'agira pas sur les transformations tant que nous n'aurons pas appelé un résultat

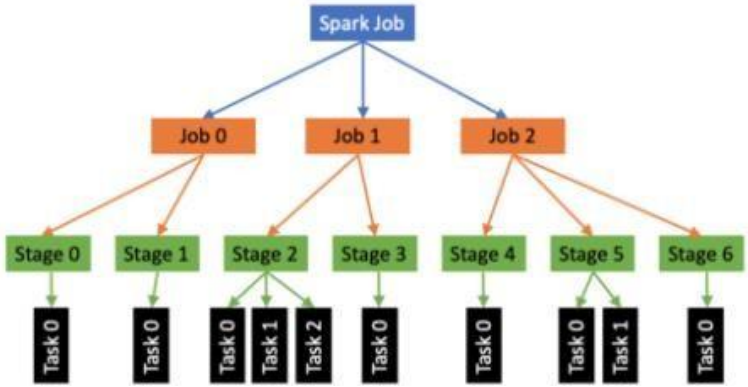
Spark: Concepts principaux

Pandas UDF (User Defined Function)

Permet des opérations vectorisées

Performances jusqu'à 100 fois supérieures aux UDF Python



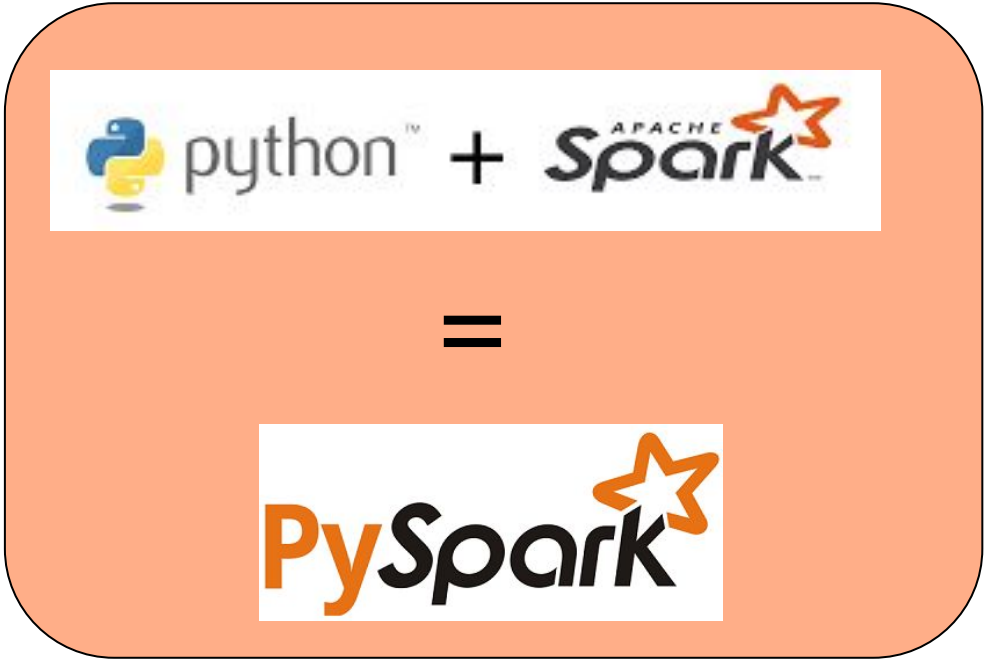


Application maître :
Configuration / Initialisation
Aggrégation des calculs

Cluster Manager :
Gestion des ressources
Distribution des calculs entre les workers

Workers :
Exécution des tâches en parallèle

PySpark



Données à disposition

kaggle

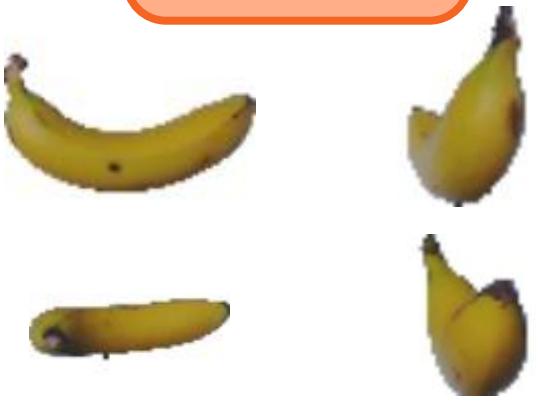
Nombre total d'images avec label:
90380 avec 131 variétés

67692 images
(Training set)

22688 images
(Test set)

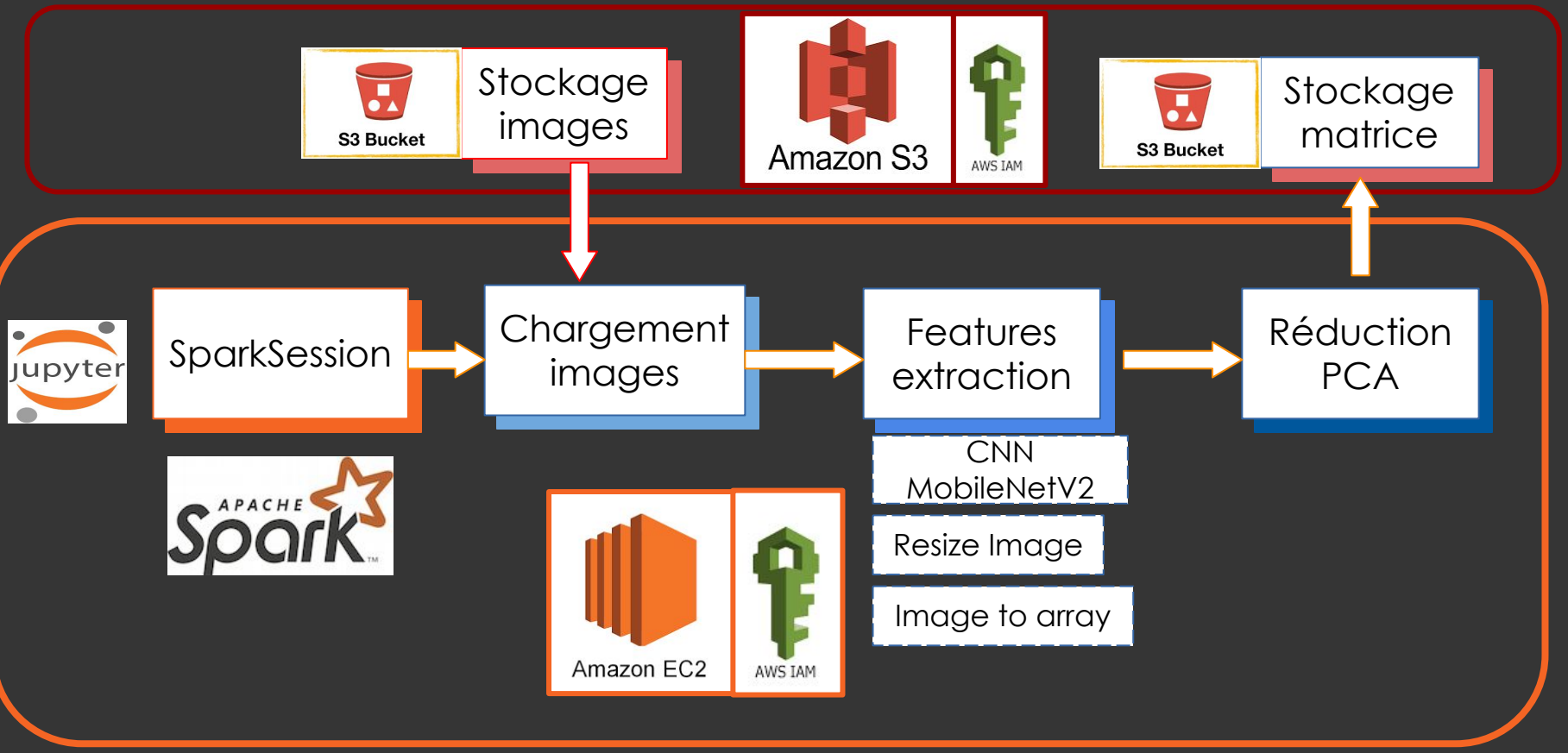
103 images multifruits sans label

Échantillon de 4 images sur 5 catégories



Fond blanc
En couleur
Centré
100*100 pixels
Avec différents angles

Chaîne de traitement



Désagréments rencontrés

1. Installation et paramétrage (Spark et AWS)
2. Spark étant écrit en Scala, sous PySpark les messages d'erreurs ne sont pas prégnant
3. AWS est très riches (plus de 200 solutions), on peut faire beaucoup de chose mais cela devient vite compliqué
4. Les coûts difficilements maîtrisables

Améliorations et Ouvertures

1. **Le code**
   a. Transposer les scripts en Scala
   b. Tester d'autres CNN (VGG, Resnet)
2. **Plateforme préconfiguré pour la Data Science**
   a. Amazon SageMaker
   b. Microsoft Databricks
3. **Business**
   a. Application grand public
      i. Proposer une analyse de l'apport énergétique du fruits
      ii. Évaluer le stade de maturation du fruits (forme de Date Limite de consommation)
   b. Robots cueilleurs (échanger avec des experts/agriculteurs)
      i. En relation avec les clients, récolter les produits par calibre, maturité (supply chain)
      ii. Identifier les maladies pouvant se répandre et appliquer un remède local

Merci beaucoup pour votre attention