DÉPLOYER UN MODÈLE DANS LE CLOUD

INTRODUCTION

_PLAN

1. Introduction:

- Contexte/Objectifs
- Jeu de données

2. Architecture Big Data:

- Big Data ?
- Use Case
- AWS
- PySpark

3. Chaine de traitement:

- Instance Spark
- VGG16
- Passage à l'echelle

4. Conclusion:

- Apprentissage
- Améliorations

INTRODUCTION

_CONTEXTE/OBJECTIFS

- **Fruits!: Start-up de l'Agritech**
- **Applications:**
  - Outils de reconnaissance de fruits
  - Robots cueilleurs intelligents
- **Objectifs: Mise en place d'une architecture Big Data**
  - Pre-processing et réduction de dimension
  - Passage à l'echelle

INTRODUCTION

_JEU DE DONNÉES

IMAGES

- **Origine: Kaggle**
  - Train: 65 000 images / Test: 20 000 images
  - 131 dossiers/variétés
  - 1 dossier multi-fruits
  - Labels

- **Caractéristiques: Données sur les précédents prêts**
  - Images en couleurs
  - Photos 360 sur fond blanc
  - Taille 100x100

INTRODUCTION

_JEU DE DONNÉES

IMAGES

- **Apple Golden**

- **Madarine:**

ARCHITECTURE BIG DATA

_BIG DATA ?

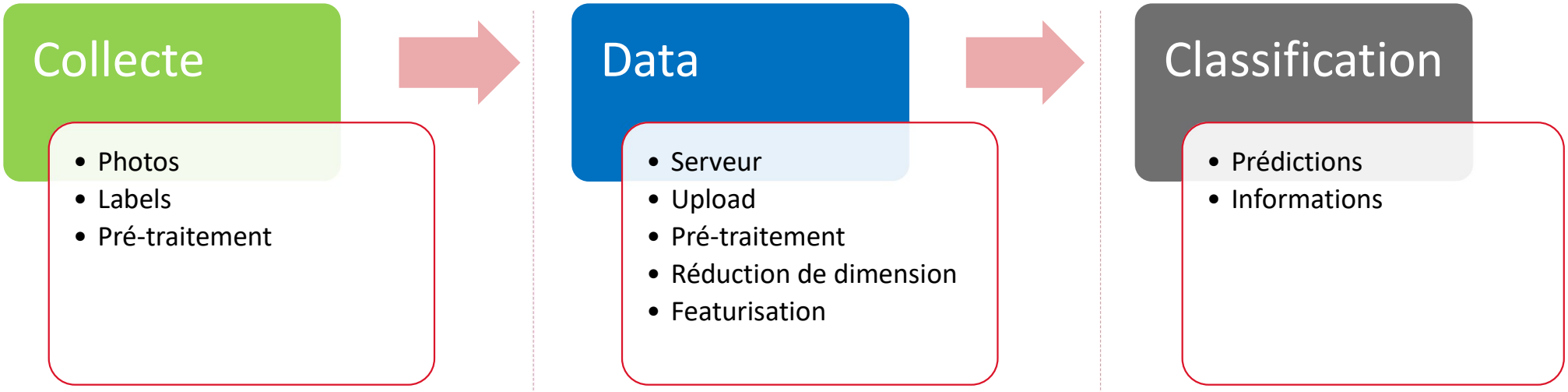
BIG DATA

Définition: 3V

- **Volume: Trop important**
  - Stockage
  - Capacité de traitement (RAM, CPU)
- **Vitesse: Fréquence d'emission ou reception**
- **Variété: Textes, images, photos, vidéos, traitées, structures…**

ARCHITECTURE BIG DATA

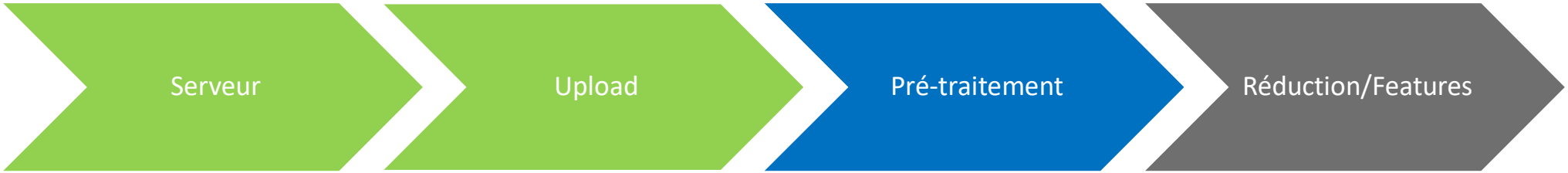
_USE CASE



PROJET 8

ARCHITECTURE BIG DATA

_AWS



Amazon EC2

EC2:

T2 large

Ubuntu Server

IAM / clé SSH

IP Elastic

S3:

Buckets

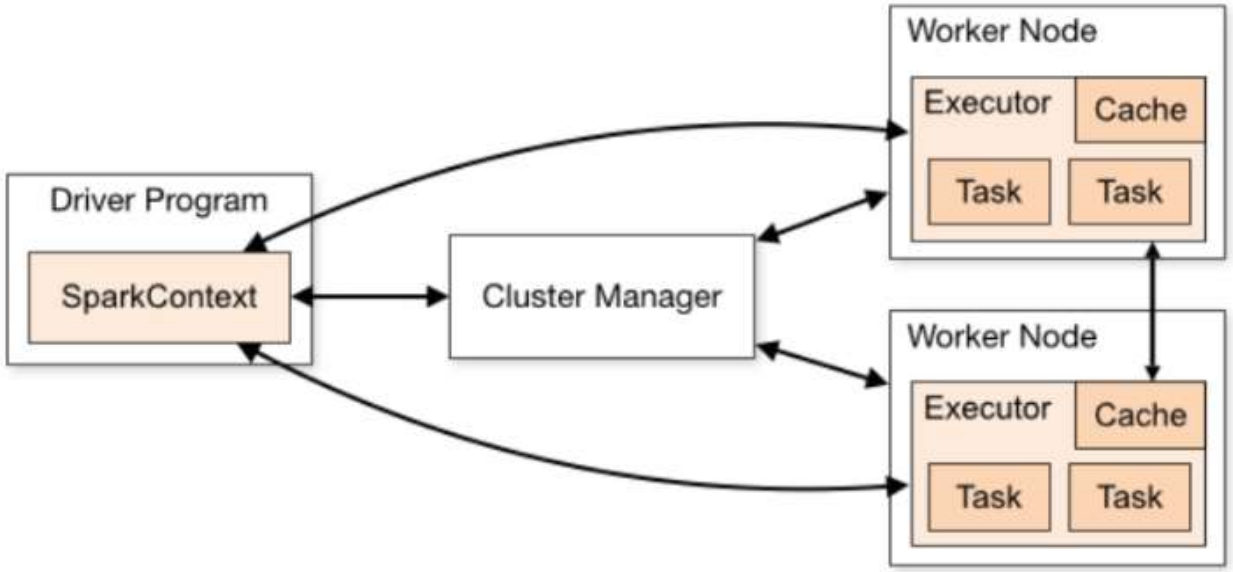
Lecture via Spark

ARCHITECTURE BIG DATA

_PYSPARK

PYSPARK

Calcul Distribué
API python Spark



Cluster Spark:

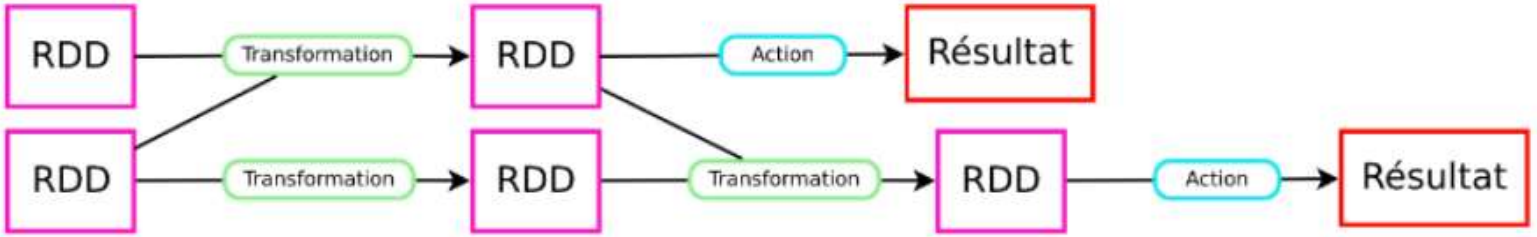
- **Workers** -> Instancie un executor
- **Driver** -> Répartie les taches aux executors
- **Cluster manager** -> Instancie les workers

ARCHITECTURE BIG DATA

_PYSPARK

RDD

Resilient Distributed Dataset



RDD:

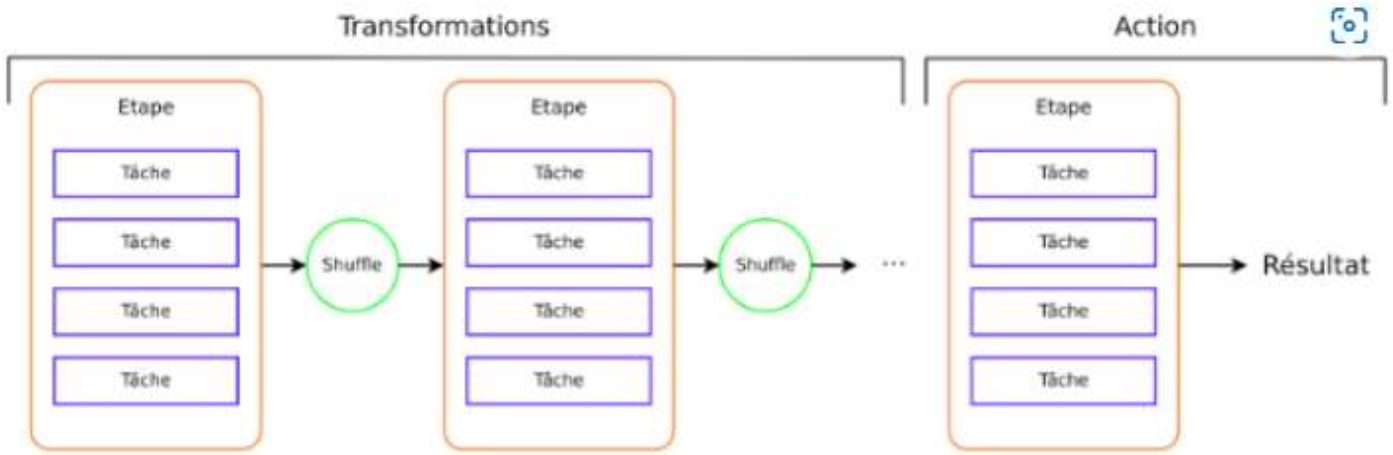
- **Transformations qui donnent en sortie des RDD**
- **Actions qui donnent en sortie un résultat**

ARCHITECTURE BIG DATA

_PYSPARK

CALCULS

Executors



Job Spark:

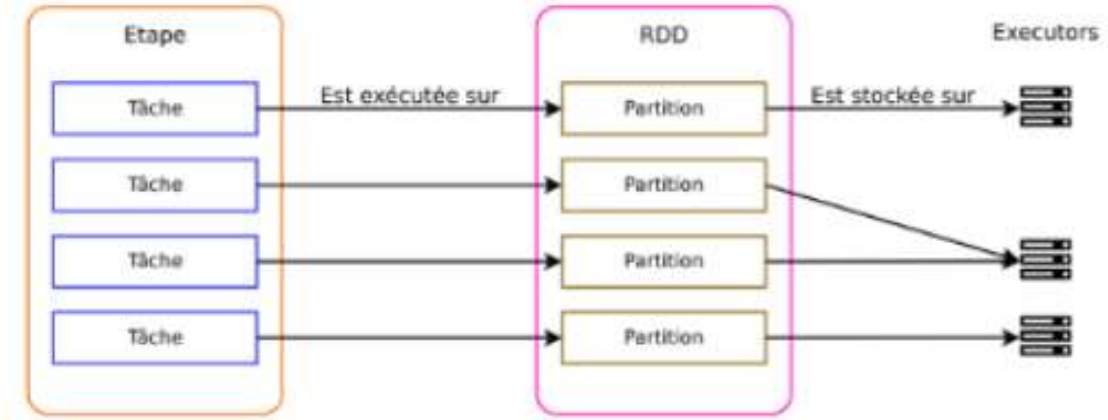
- **Ensemble d'étapes, elles-mêmes constitués d'un ensemble de tâches.**
- **Action sur RDD, plusieurs étapes séparées par des shuffles**

ARCHITECTURE BIG DATA

_PYSPARK

CALCULS

Executors



Fruits!

CHAINE DE TRAITEMENT

_INSTANCE SPARK

INSTANCE SPARK

- **VGG16: Feature extractor**
- **Création: SparkContext.getOrCreate()**
  - .master()
  - .config()
  - .appName()
- **Spark Dataframe: spark.createDataFrame()**
- **Array to Vector: udf(lambda l: Vectors.dense(l), VectorUDT())**

CHAINE DE TRAITEMENT

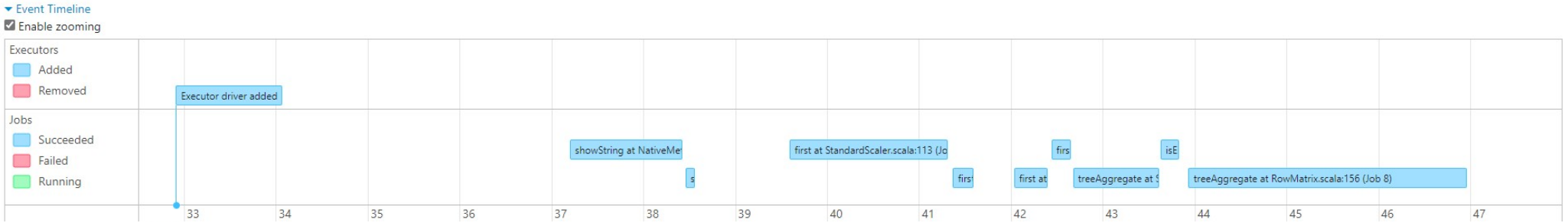
_INSTANCE SPARK

INSTANCE SPARK

- Scaler: StandardScaler()
- PCA: PCA()
- Vector to array: udf()
- Spark Dataframe to Pandas DF: .toPandas()

CHAINE DE TRAITEMENT

_INSTANCE SPARK



Completed Jobs (9)

Page: 1

1 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
8	treeAggregate at RowMatrix.scala:156 treeAggregate at RowMatrix.scala:156	2022/06/08 09:37:43	3 s	1/1	2/2
7	isEmpty at RowMatrix.scala:426 isEmpty at RowMatrix.scala:426	2022/06/08 09:37:43	0.2 s	1/1	1/1
6	treeAggregate at Statistics.scala:58 treeAggregate at Statistics.scala:58	2022/06/08 09:37:42	0.9 s	1/1	2/2
5	first at RowMatrix.scala:62 first at RowMatrix.scala:62	2022/06/08 09:37:42	0.2 s	1/1	1/1
4	first at PCA.scala:44 first at PCA.scala:44	2022/06/08 09:37:42	0.4 s	1/1	1/1
3	first at StandardScaler.scala:113 first at StandardScaler.scala:113	2022/06/08 09:37:41	0.2 s	1/1 (1 skipped)	1/1 (2 skipped)
2	first at StandardScaler.scala:113 first at StandardScaler.scala:113	2022/06/08 09:37:39	2 s	1/1	2/2
1	showString at NativeMethodAccessorImpl.java:0 showString at NativeMethodAccessorImpl.java:0	2022/06/08 09:37:38	96 ms	1/1	1/1
0	showString at NativeMethodAccessorImpl.java:0 showString at NativeMethodAccessorImpl.java:0	2022/06/08 09:37:37	1 s	1/1	1/1

Fruits!

CHAINE DE TRAITEMENT

_VGG16

FEATURES EXTRACTOR

- **Model: VGG16**
- **Couches: Sans classification**

```
VGG_extractor = VGG16(
                weights=None,
                include_top=False,
                input_shape=(100, 100, 3))
```

- **Sans entrainement**

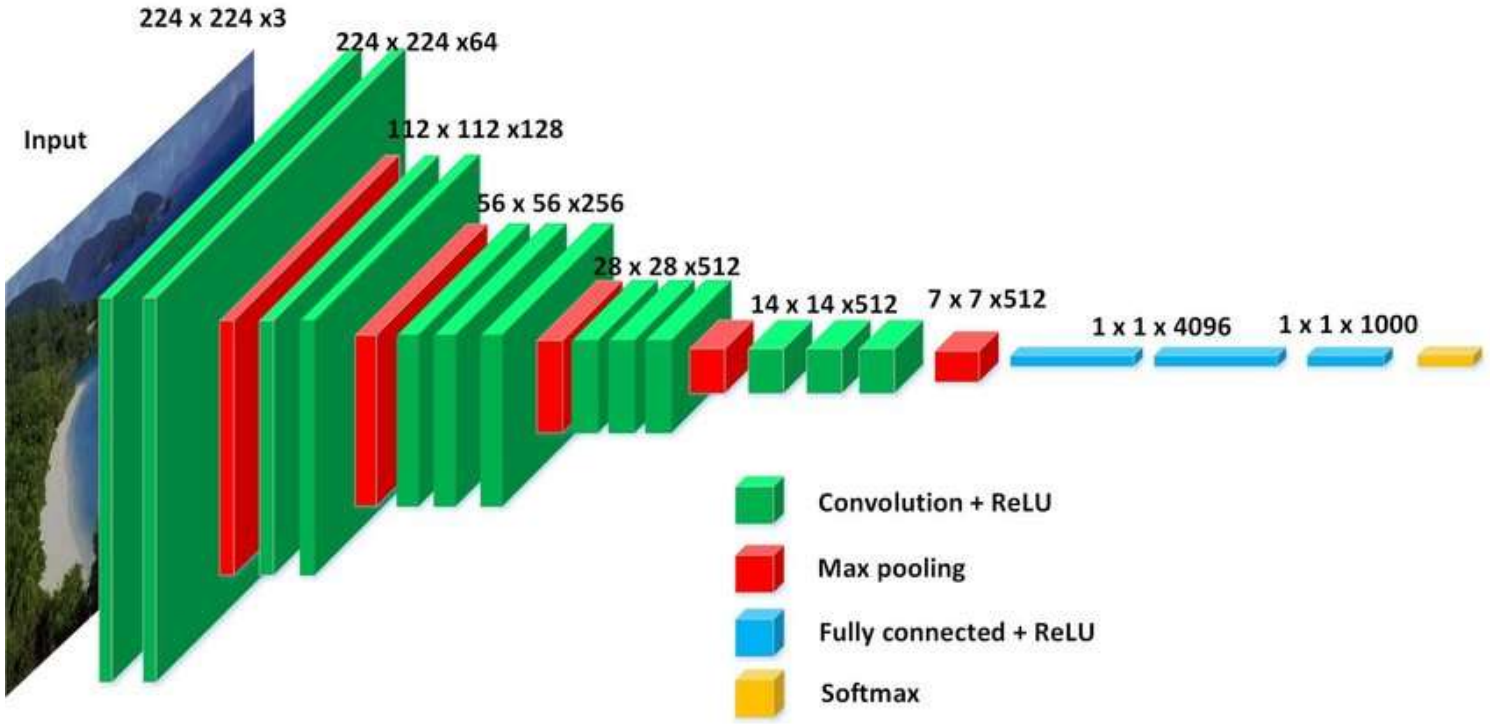
```
1 for layer in VGG_model.layers:
2     layer.trainable = False
```

```
=================================
Total params: 14,714,688
Trainable params: 0
Non-trainable params: 14,714,688
_________________________________
```

CHAINE DE TRAITEMENT

_VGG16

FEATURES EXTRACTOR



CHAINE DE TRAITEMENT

_VGG16

PASSAGE ECHELLE

- **Notebook: Aucun changement**
- **Buckets S3: OK**
- **Instance EC2: A augmenter pour une plus grande capacité de calcul (RAM, Processeur)**
- **Spark: Optimisation des instances**

CONCLUSION

_APPRENTISSAGE

APPRENTISSAGE

CONCLUSION

_AMÉLIORATIONS

AMÉLIORATIONS

- **Images:**
  - Pré-traitement pour cas réel
  - Multi-fruits
  - Modèle entrainé
- **Architecture: Passage à l'echelle**
- **Use case: Identification améliorée**
  - Maturité des fruits
  - Maladies