Projet 8
Déployer un modèle dans le cloud

Réaliser un modèle de scoring

- Présentation du projet et des données
- Création de l’architecture Big Data
- Présentation du script spark
- Conclusion

Présentation du projet

- Votre start-up souhaite se faire connaître en mettant à disposition du public une application qui permettrait aux utilisateurs de prendre en photo un fruit et d'obtenir des informations sur ce fruit.

- Cette application permettra de de mettre en place une première version du moteur de classification des images de fruits et de construire une première version de l'architecture Big Data nécessaire.

- Vous êtes donc chargé de développer dans un environnement Big Data une première chaîne de traitement des données

Contraintes

- Le volume des donnée va grandements augmenté au cours du temps
- Dépassement des capacités de RAM
- Dépassement des capacité de stockages
- Augmentation du temps de traitement

Objectifs

- Dévelloper une architecture Big Data dans le cloud
  - AWS
  - EC2
  - S3
  - IAM

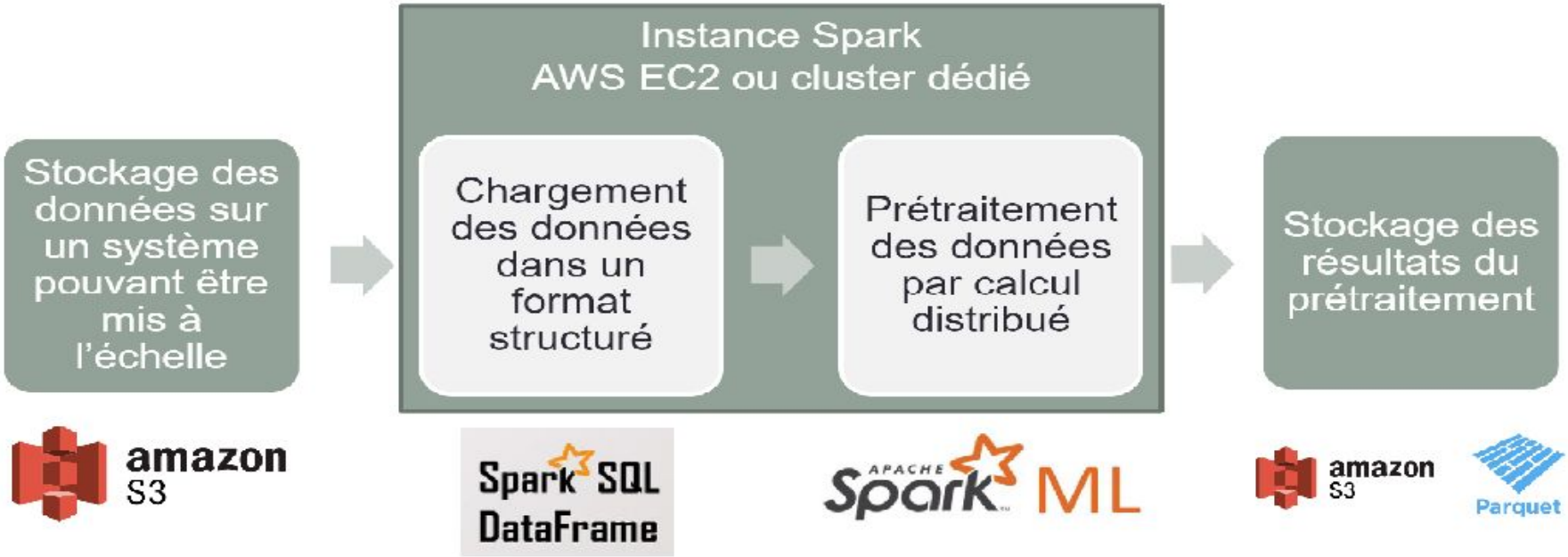
- Créer une première chaines de traitements des données avec Pyspark

Présentation des données

- Trois jeux de données
  - Jeu d’entrainement : 53 000 images
  - Jeu de Test : 18 000 images
  - Jeu multi fruits : 103 images
  -
- 120 Variétés de fruits réparti dans des dossiers différents

- Formats Jpeg RGB 100 x 100

Architecture Big



IAM

- Utilisateur racine qui possède l'accès total à la console aws
- Utilisateur IAM

- Rôles d’instances

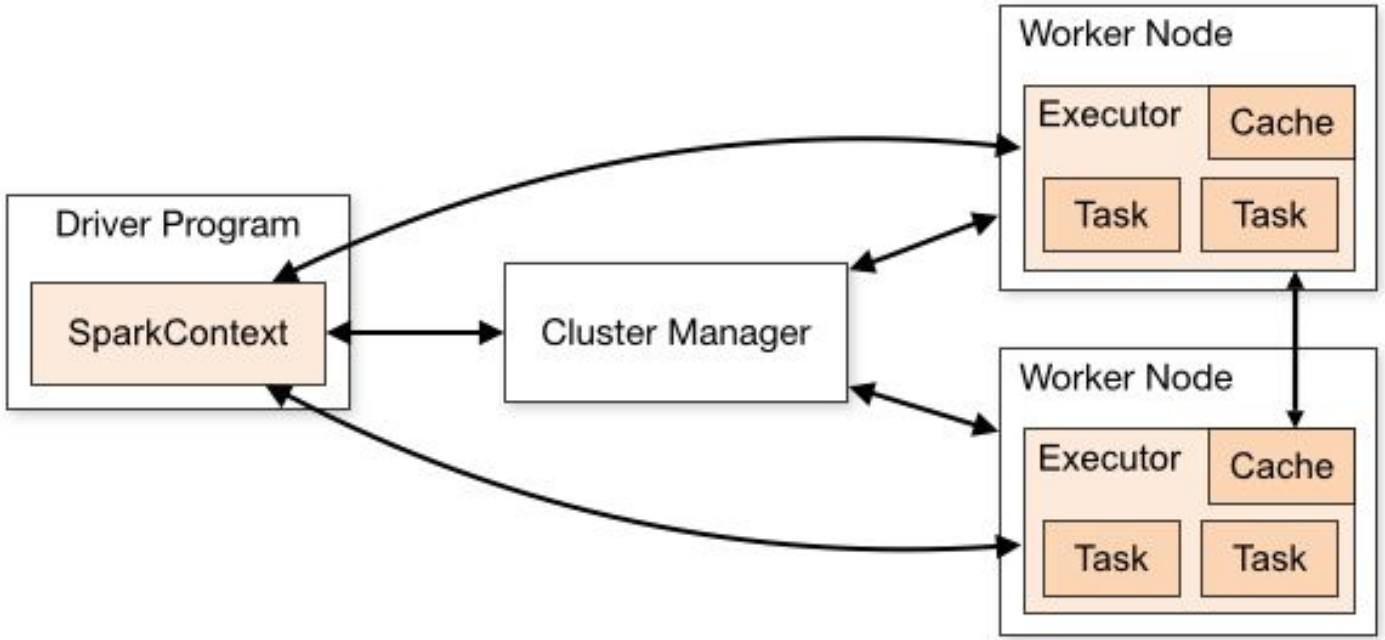
EC2

- Type :
  - Amazon Linux 2 Kernel 5.10 AMI 2.0.20221004.0 x86_64 HVM gp2
  - T3 medium
  - eu-west-3

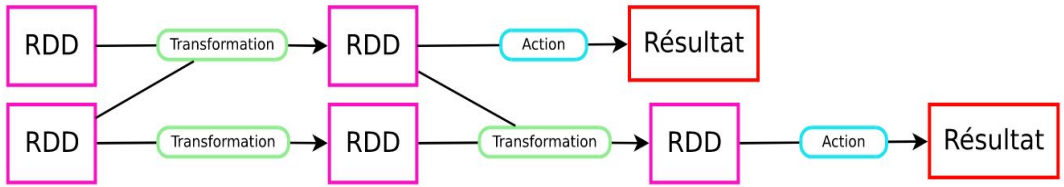
- Configuration:
  - Anaconda 2022.10
  - Python 3.9
  - Pyspark 3.1.1
  - Tensorflow 2.10.0
  - OpenJDK 11.0.16

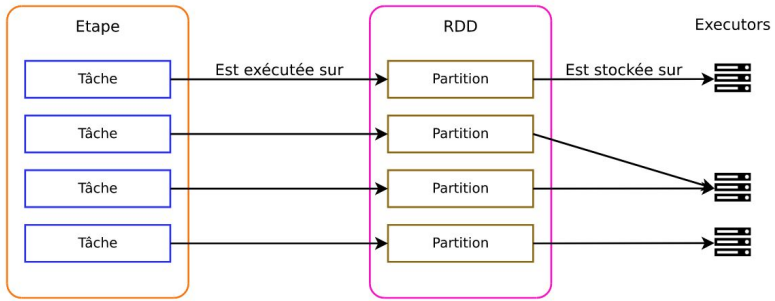
- Création d’un tunnel ssh pour accéder à jupyter notebook sur un navigateur

Pyspark

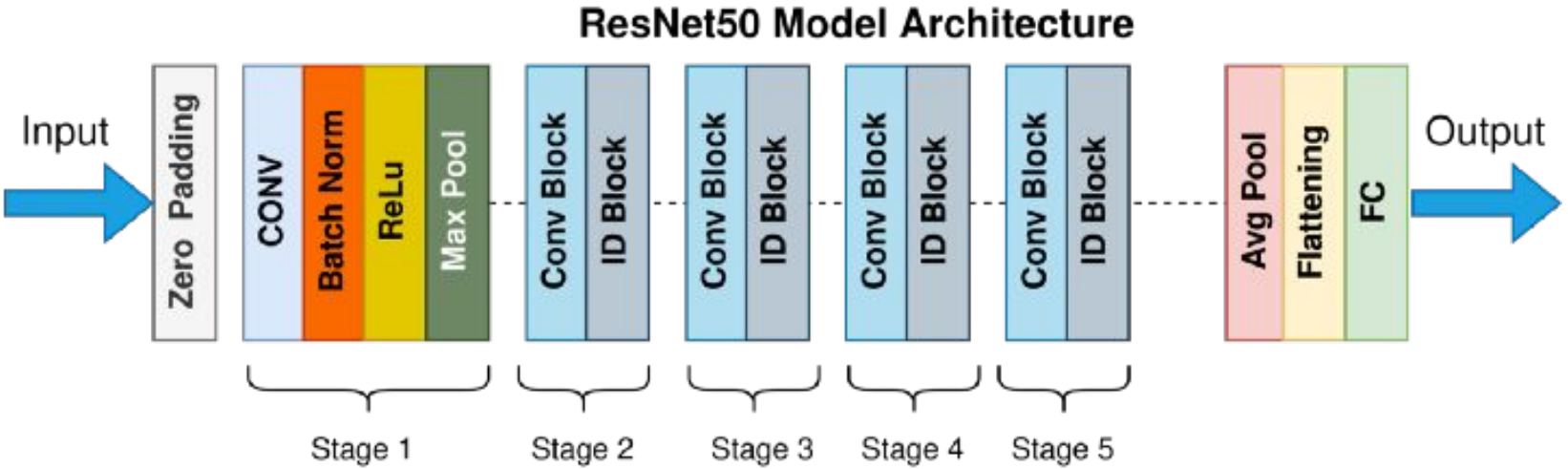

Worker Node
Executor
Cache
Task
Task
Driver Program
SparkContext
Cluster Manager
Worker Node
Executor
Cache
Task
Task

RDD


RDD
Transformation
RDD
Action
Résultat
RDD
Transformation
RDD
Transformation
RDD
Action
Résultat


Etape
Tâche
Tâche
Tâche
Tâche
Est exécutée sur
RDD
Partition
Partition
Partition
Partition
Est stockée sur
Executors

Modèle


ResNet50 Model Architecture
Input
Zero Padding
CONV
Batch Norm
ReLu
Max Pool
Conv Block
ID Block
Conv Block
ID Block
Conv Block
ID Block
Conv Block
ID Block
Avg Pool
Flattening
FC
Output
Stage 1
Stage 2
Stage 3
Stage 4
Stage 5

Script Spark

- Configuration de l’instance (Clés d’accès et secrète,région , package)

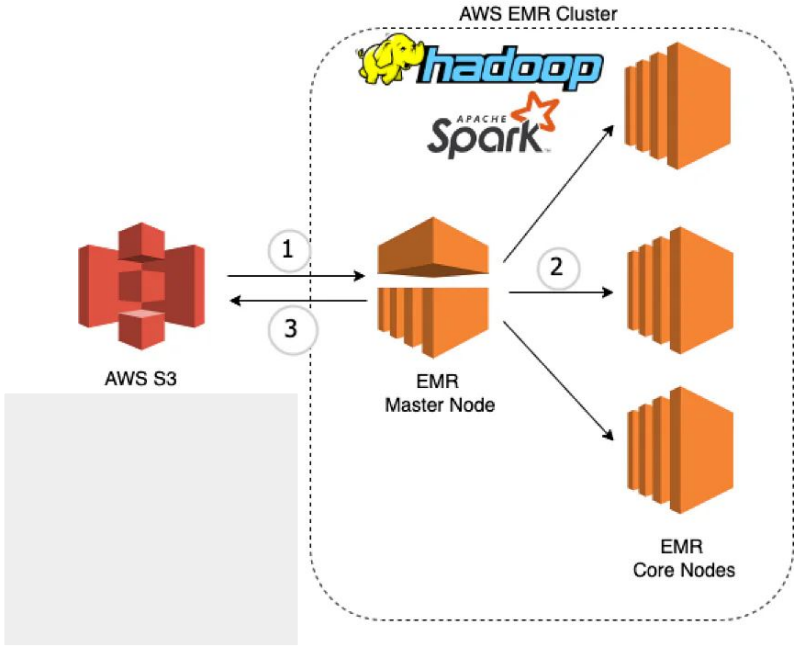
```
'--packages com.amazonaws:aws-java-sdk-bundle:1.11.874,org.apache.hadoop:hadoop-aws:3.2.0 pyspark-shell'
```

- Création de l’instance spark
- Lecture des fichier sur s3 : spark.read()
- Prédiction avec le modèle definie
- Sauvegarde au format parquet sur S3 : spark.write()

Conclusion

Passage à grande échelle

- Script Spark:
  - Remplacer du modèle par le modèle préentrainé
- Infrastructure:
  - Remplacement du l'instance EC2 par une
  - instance plus puissante
  - Utilisation d'un cluster EMR avec plusieurs instances ec2
- Stockage :
  - Définir un cycle de vie



Problème rencontrés

- Installation de jupyter notebook et création d’un tunnel ssh.
  - Modification des groupes de securités
  - Configuration jupyter
  - Configuration de l’environnement spark

- Configuration de l’environnement spark
  - aws-jdk-bundle
  - hadoop-aws
  - compatibilité tensorflow - pyArrow