DÉPLOYEZ UN MODÈLE DANS LE CLOUD

ETUDIANT

SAHEL TAHERIAN

PROJET N°8
PARCOURS « DATA SCIENTIST »

SOUTENANCE DE PROJET

03 JUIN 2022

PLAN DE LA PRÉSENTATION

I. Rappel de la problématique et Présentation du jeu de données

II. Présentation de l'environnement Big Data dans le cloud

III. Architecture retenue et chaîne de traitement

IV. Conclusion

RAPPEL DE LA PROBLÉMATIQUE ET PRÉSENTATION DU JEU DE DONNÉES

PROBLÉMATIQUE

Fruits!

L'ENTREPRISE : *Fruits !*

STARTUP AGRITECH

- **Mettre à disposition du grand public une application mobile** de reconnaissance de fruit et affichage d'informations
- **Mettre en place une première version du moteur de classification des images de fruits.**
- **Construire une première version de l'architecture Big Data nécessaire.**
- Développement des robots cueilleurs intelligents

OBJECTIF : Mettre en place l'architecture Big Data

- Prétraitement et réduction de dimension
- Accessibilité des données et des résultats dans le cloud

JEU DE DONNÉES

ORIGINE:

Fruits 360, *Mihai Oltean*

- **Un dataframe de 90483 images de 131 fruits et légumes avec les labels associés**
- **Jeu d'entraînement : 67692 images**
- **Jeu de Test : 22688 images**

CARACTÉRISTIQUES :

- **Images 100x100 JPEG RGB**
- **Photos studio sur fond blanc de fruits centrée sur le fruit**
- **Plusieurs variétés du même fruit (exemple : pomme « red » et « golden »)**
- **Photos sous tous les angles (rotation 3 axes)**

LE BIG DATA

QU'EST-CE QUE LE BIG DATA?

COMMENT RÉPONDRE À SES ENJEUX?

QU'EST-CE QUE LE BIG DATA ?
LES DONNÉES MASSIVES, LES MÉGADONNÉES

- À partir du moment où la quantité de données excède la faculté d'une machine à les stocker et les analyser en un temps acceptable, Nous considérerons que l'on fait du big data

les 3V du big data : Volume, Vélocité, Variété

  - **Volume :** une quantité importante de données pour être stockées et/ou traitées sur une seule machine avec des performances acceptables
  - **Vitesse** à laquelle les données sont produites
  - Large **Variété** de types de données

- Les outils développés pour résoudre les problèmes des big data:
  - leur facilité d'utilisation,
  - leur faculté à passer aisément à l'échelle
  - leur disponibilité auprès du grand public

Passer à l'échelle

La possibilité pour les systèmes de traitement de données d'augmenter leurs capacités de traitement au fur et à mesure que les données augmentent

COMMENT RÉPONDRE À CES ENJEUX?

CAPACITÉS DE CALCUL : Traitement par calculs distribués (MapReduce: Divisez pour distribuer)

- Diviser les opérations en micro opérations distribuables entre différentes machines, réalisables en parallèle
- Agréger l'ensemble des résultats intermédiaires obtenus pour chaque lot pour construire le résultat final.

MapReduce:

- cela demande **beaucoup d'efforts de transformer** un algorithme en MapReduce.
- les deux étapes MAP et REDUCE ne suffisent pas,

Un framework Hadoop :

MapReduce + HDFS (Hadoop Distributed File System)

- **HDFS:** un système de fichiers distribué.

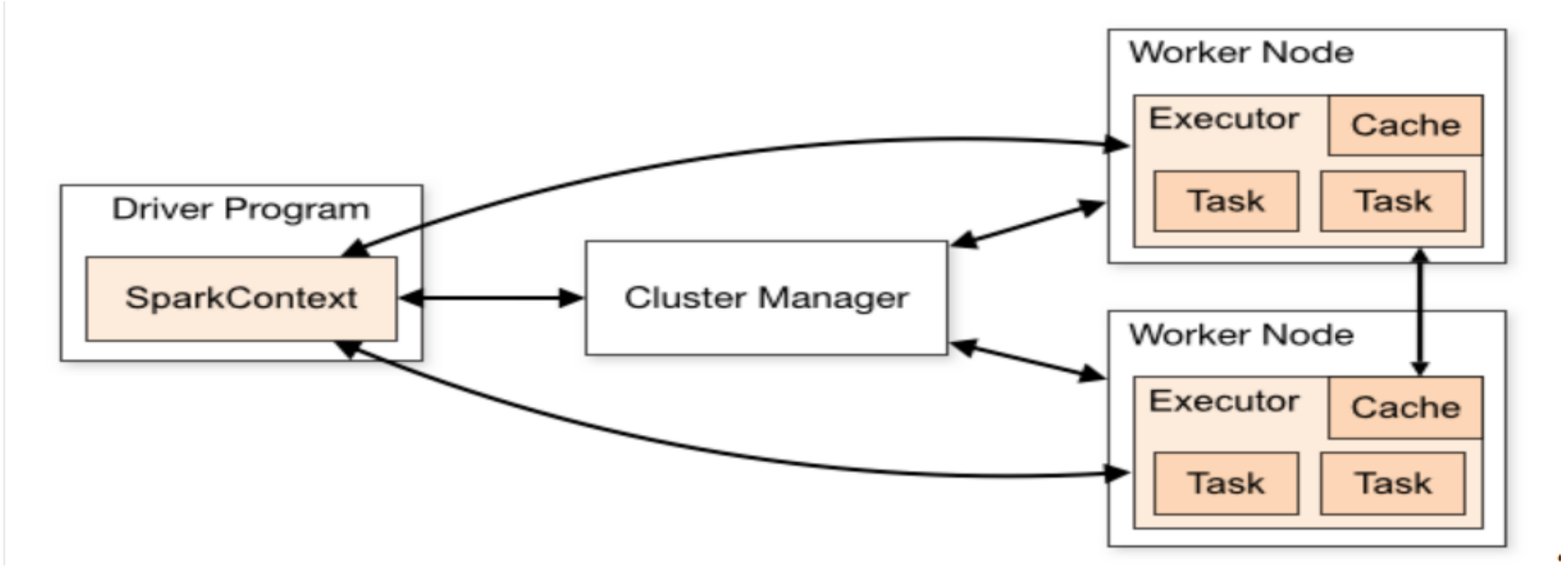
Il a été conçu pour stocker des fichiers de très grande taille, dans un cadre distribué avec une Tolérance aux pannes

- **YARN** séparer la gestion des ressources du cluster et la gestion des jobs MapReduce, permettant ainsi de généraliser cette gestion des ressources à d'autres applications

CLUSTER DE SPARK

- Spark écrit les données en RAM

Un Cluster de Spark



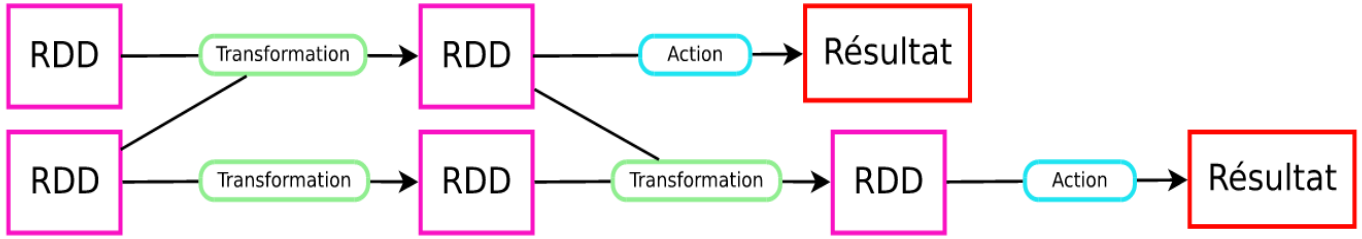
DRIVER :
Configuration
Initialisation
Agrégation des calculs

CLUSTER MANAGER :
Gestion des ressources
Distribution des calculs entre les workers

WORKERS : Charger de l'Exécution des tâches de calcules

CLUSTER DE SPARK

Graphe acyclique orienté (DAG)



- **Utilisation de Resilient Distributed Datasets (RDD)**

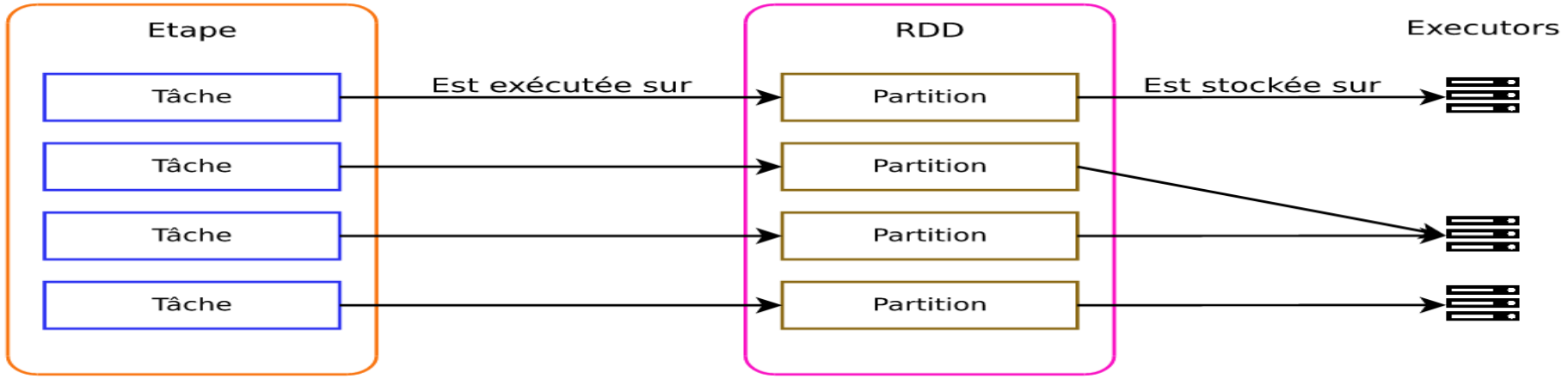
Ils dictent la manière dont les calculs vont être distribués sur les différentes machines.

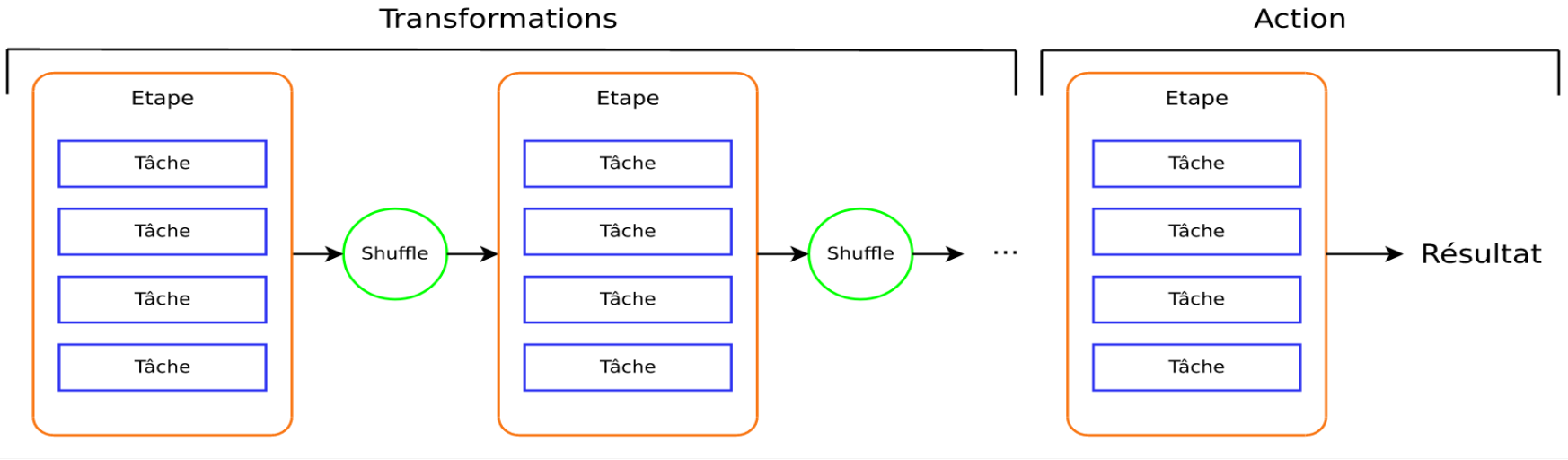
- **Division des données en partitions**
- **Duplication des données (3 machines par défaut)**
- **Tolérance aux pannes**

Graphe Acyclique Orienté (DAG) :

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

COMMENT SPARK DISTRIBUE LES CALCULS SUR LES DIFFÉRENTS EXECUTORS ?





ARCHITECTURE RETENUE ET CHAÎNE DE TRAITEMENT

PRÉTRAITEMENT

Objectif :

Préparer les images pour le Learning

Extraction des images

Réduction de dimensions

path	category	content	features	scaled_features	reduced_features
s3a://p8-sahelbuc...	cucumber_1	[FF D8 FF E0 00 1...	[6.685832, 1.2260...	[0.33976092858023...	[34.2339263611350...
s3a://p8-sahelbuc...	cucumber_1	[FF D8 FF E0 00 1...	[9.626037, 3.1773...	[1.25002302599642...	[32.2562875210084...
s3a://p8-sahelbuc...	carrot_1	[FF D8 FF E0 00 1...	[1.179364, 5.2705...	[-1.3649942877012...	[18.5147374826944...
s3a://p8-sahelbuc...	carrot_1	[FF D8 FF E0 00 1...	[1.3376092, 7.980...	[-1.3160029311091...	[15.4059220226347...
s3a://p8-sahelbuc...	carrot_1	[FF D8 FF E0 00 1...	[6.2302, 2.800575...	[0.19870110980076...	[19.9182351191773...
s3a://p8-sahelbuc...	eggplant_violet_1	[FF D8 FF E0 00 1...	[7.592759, 11.005...	[0.62053777197090...	[-19.546216527816...
s3a://p8-sahelbuc...	eggplant_violet_1	[FF D8 FF E0 00 1...	[9.540223, 6.2902...	[1.22345589652186...	[-18.427371614516...
s3a://p8-sahelbuc...	eggplant_violet_1	[FF D8 FF E0 00 1...	[6.048176, 12.184...	[0.14234804213954...	[-23.171094974965...
s3a://p8-sahelbuc...	eggplant_violet_1	[FF D8 FF E0 00 1...	[2.0192919, 9.292...	[-1.1049598202559...	[-21.687611166781...
s3a://p8-sahelbuc...	eggplant_violet_1	[FF D8 FF E0 00 1...	[8.46298, 6.25776...	[0.88995076207595...	[-20.893976123551...
s3a://p8-sahelbuc...	eggplant_violet_1	[FF D8 FF E0 00 1...	[2.7497358, 15.39...	[-0.8788206456440...	[-16.602838131263...

INSTANCE SPARK

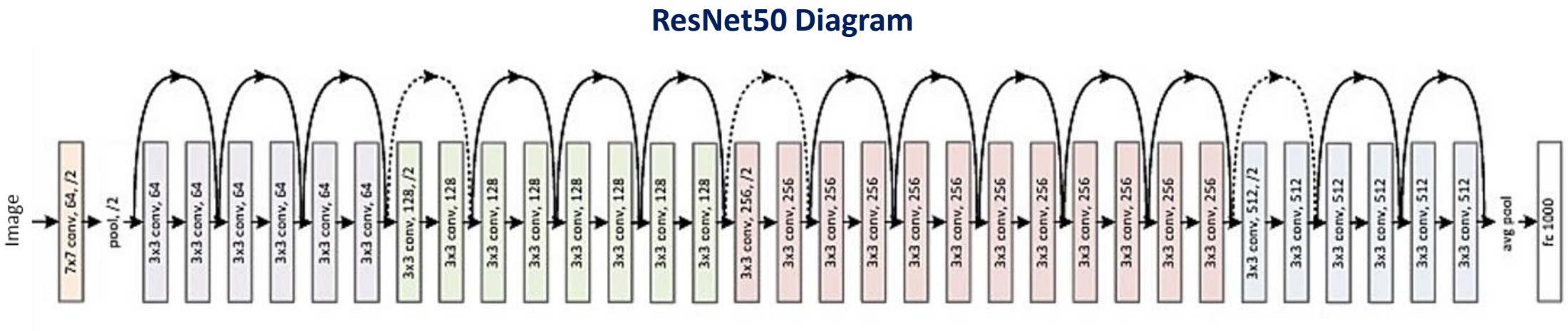
- Création d'un SparkContext
- Chargement des images dans un spark DataFrame
- Extraction des Features
- Réduction Dimensionnelle
- Enregistrement sur S3

PRÉTRAITEMENT EN DÉTAIL

MODÈLE DE CLASSIFICATION: **Réseau RESNET50**

APPROCHE : Transfer Learning

- couche supérieure enlevée (fully-connected)
- Pondération en broadcast déjà pré-entraînés



L'INFRASTRUCTURE ET PASSAGE À L'ÉCHELLE

Stockage fichiers sur S3 :

- upload via AWS CLI ou Interface Web
- Lecture des fichiers depuis Spark
- Enregistrement de fichier depuis spark vers S3
- Enregistrement de la sortie depuis spark vers S3

Instance EC2 (T2.xlarge) / OS Ubuntu Server 18.04

Configuration : Python 3 / Java 8 / Spark / Hadoop-AWS/ Spark MLLib / Pillow

Configuration sur machine distante : accès via SSH

- Chargement clés IAM / AWS
- Installation des logiciels et packages
- Mise en place Jupyter Notebook accessible à distance pour exécution du code / analyse des résultats sur DataBricks

PASSER À L'ÉCHELLE

- Des modifications de code Spark/Python à apporter
- Stockage des fichiers :
  - S3
  - DataBricks
- Instance EC2 de plus grande capacité RAM/Processeur
- Creation d'un cluster Elastic Map Reduce avec plusieurs instances EC2 (1 Maître + n esclaves)
- Configuration
- Augmentation du nombre d'instances esclaves / noeuds

CONCLUSION ET PERSPECTIVES

CONCLUSION ET PERSPECTIVES

Enseignements

- Prise en main Pyspark
- Configuartion de spark en local
- Découverte du format distribué parquet
- Découverte l'écosystème AWS
- Administration d'un serveur Linux par SSH
- Notebook des script Pyspark deployé sur DataBricks

Aller plus loin

- Prétraitement pour les cas réels (recadrage, plusieurs fruits, arrière plan, etc.)
- Entraîner le modèle (approche transfer learning)
- Déployer le modèle en production sur un cluster comme EMR
- Identifier la maturité des fruits pour les cueillir au bon moment
- Identifier les pathologies ou les fruits abîmés

MERCI DE VOTRE ATTENTION