



Déployer un modèle dans le cloud

Data & Analytics

Eric Blanvillain – 21-03-2022



Problématique

Le contexte et les données

Mission :

- Aider la startup Fruits a développer une application grand publique d'identification de fruits
- Permettre le passage à l'échelle à l'aide d'une infrastructure Big Data

Objectifs :

- Mettre en place une architecture Big Data
- Pré-traitement des données et classification

1
Contexte

2
Données

3
Big Data


4
Infrastructure


5
Traitement

Le contexte

Solutions innovantes pour la
récolte des fruits

Moyens :

Scripts : 

Déploiement cloud : 

Fruits! : Startup de l'AgriTech



Fruits!

Application mobile
grand public

Etape 1 :

- Mettre en place une architecture Big Data
- Anticipation : passage à l'échelle (volume de données)
- Pré-traitement des données:
 - Pré-processing
 - Réduction de dimension

AgriTech : l'IA au service de l'agriculture ++



Robots cueilleurs
intelligents

Etape 2 :

Application : photo de fruit -> informations sur le fruit
-> Faire connaître la startup

Etape 3 :

Mise en place ultérieure

Les données initiales

Jeu entrainement

- ❑ 67 692 images
- ❑ 131 classes

75%

Jeux avec étiquettes

- ❑ 90 380 images
- ❑ 131 classes

Jeu test

- ❑ 22 688 images
- ❑ 131 classes

25%

131 dossiers



Photos 360° de fruits et légumes



Fond blanc, 100x100 pixels



Un seul fruit/légume par image

Jeux sans étiquette

- ❑ multi fruits
- ❑ 103 images

+ meta-donnée

Big data

Quelles solutions pour répondre aux enjeux?

Le Big Data

Quelles solutions pour répondre aux enjeux de Fruits! ?



Volume



Variété

3 **V**
Big Data



Vitesse de création
fréquence de création
collecte et partage des données

Le Big Data c'est quoi ?

- Explosion de la quantité de données
- Le partage des données
- La recherche des données
- Le stockage des données
- Le traitement des flux de données

Les solutions ...

Stocker la donnée

Solutions existantes



Amazon S3



Google Cloud Storage



IBM Spectrum Storage



Notre solution



Amazon S3



Évolutivité :

- Pas de limite de place
- Ressources à l'échelle
- Scalabilité
- Disponibilité

Résilience :

- Redondance : copie des objets sur plusieurs systèmes
- Tolérance aux pannes

Performance :

- Durabilité
- Bonne compression

Traiter la donnée

Exécution des calculs en parallèle

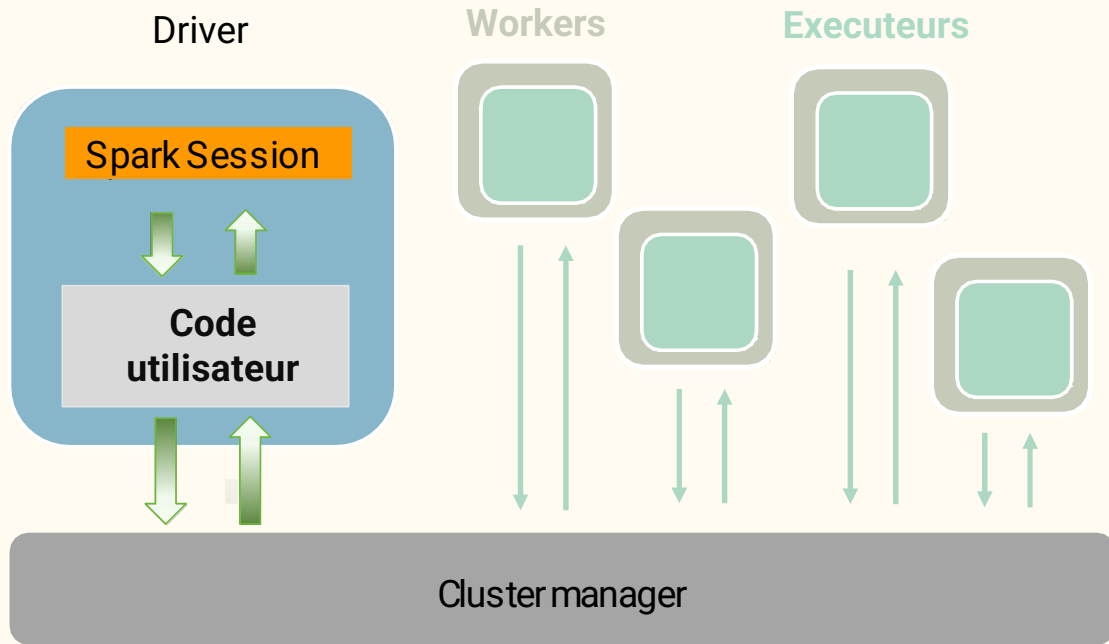
Notre solution



Calculs
distribués

Infrastructure
Données
distribuées

Configuration
Initialisation
Agrégation des calculs



Distribution des calculs
entre les workers

MapReduce : Map (transformer) Reduce (agrérer)

Architecture

The word "Architecture" is written in a black serif font. The first two letters, "Ar", are in a bold orange color. Below the "Ar" is a curved orange arrow pointing to the right, resembling the Amazon logo.

Une infrastructure distribuée

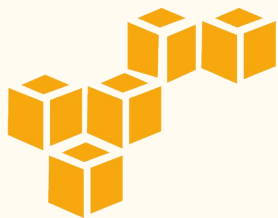
Architecture AWS



Clef IAM



Nos briques



EC2



EMR



S3



IAM - Sécurité renforcée

Identity and Access Management

Contrôlez de façon sécurisée l'accès aux services et ressources AWS.

Rôle : Utilisateur, admin, super admin

EC2 - Serveurs virtuels dans le cloud

Elastic Compute Cloud

Capacité de calcul sécurisée et redimensionnable pouvant prendre en charge quasiment tout type de charge de travail

Une infrastructure à la demande fiable et évolutive

EMR - Analyse

Elastic Map Reduce

Amazon EMR est un service qui utilise Apache Spark pour traiter et analyser de grandes quantités de données.

-> Exécutez et mettez à l'échelle facilement les cadres Apache Spark, Hive, Presto et d'autres cadres de Big Data.

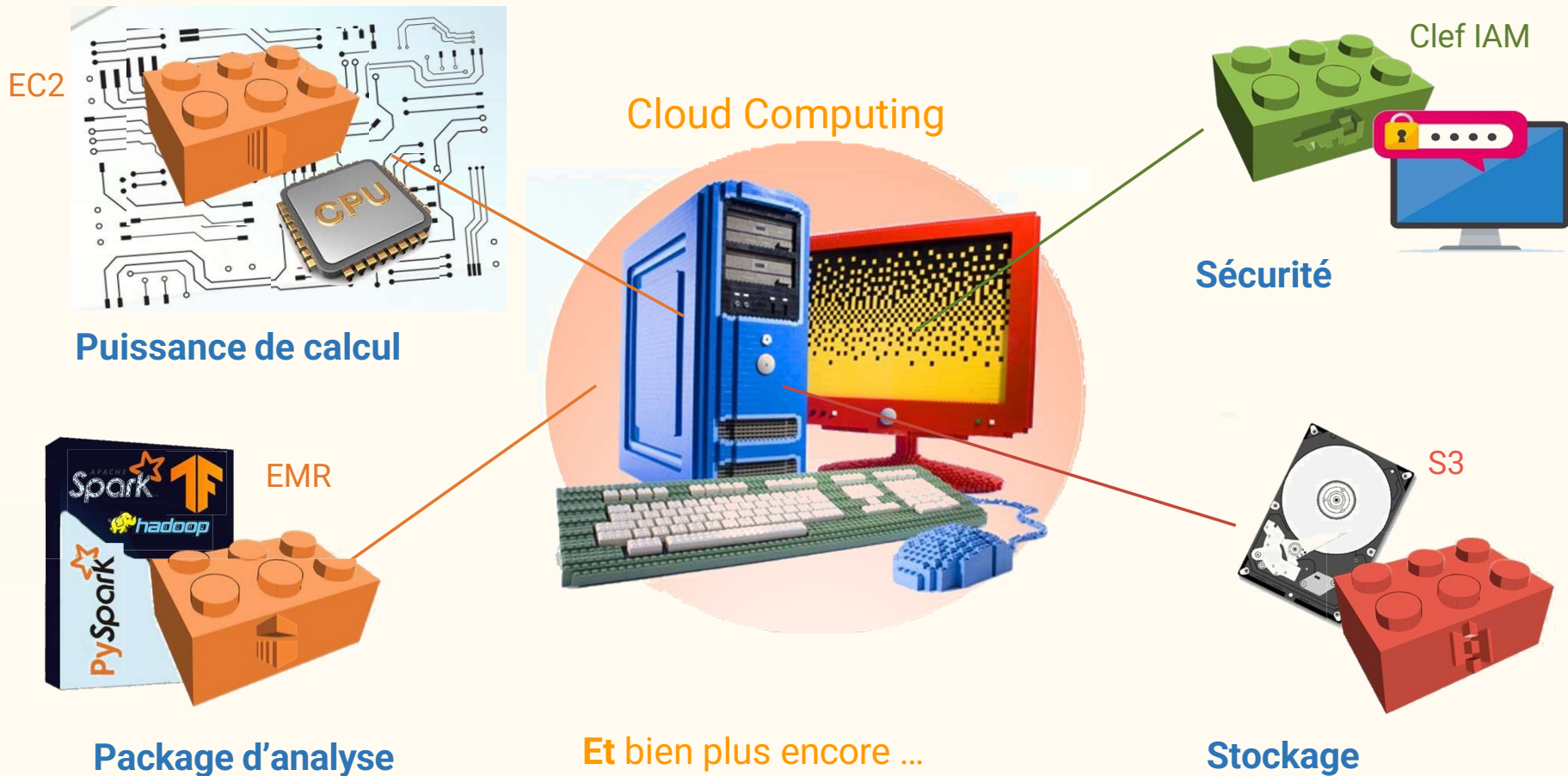
S3 - Stockage scalable dans le cloud

Simple Storage Service

Stockage d'objets conçu pour stocker et récupérer n'importe quelle quantité de données, n'importe où

-> Performances, scalabilité, disponibilité et durabilité de pointe

Un « super ordinateur virtuel »



Local vs Cloud

	Local	Cloud
Stockage	Disque dur : limité Panne possible : perte des données Données disponibles localement	Illimité Redondance : tolérance aux pannes Données disponibles partout
Puissance de calcul	Dépendante du matériel informatique à disposition	Evolutive en fonction de la charge de travail
Ethique / RGPD	Les informations restent au sein de l'entreprise Protection en interne des données à caractère confidentiel	Contraintes juridiques liées à l'hébergement des données Confidentialité des données
Sécurité	Choix de l'utilisateur	
Coût	Fixe	Variable

Chaîne de traitement

La pipeline du projet

Chaine de traitement

Un « ordinateur virtuel »

Traitement









Stockage S3



S3

Compartiment S3

Nom
 app_P8.py
 configuration.json
 emr_bootstrap.sh
 images/
 logs/
 resultat_parquet/

Données
chargées

Données
enregistrées

Données initiales

Code python

Configuration

Libraires à installer

Fichiers en format parquet

```
# Création d'une SparkSession
spark = SparkSession.builder\
    .appName('P8_preprocess_images')\
    .getOrCreate()

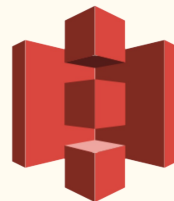
# Chargement des données
# En format "binaryFile"
df_binary = spark.read.format("binaryFile") \
    .option("pathGlobFilter", "*.jpg") \
    .option("recursiveFileLookup", "true") \
    .load(data_source)
```

```
{
  "Classification": "spark-env",
  "Configurations": {
    {
      "Classification": "export",
      "Properties": {
        "PYSPARK_PYTHON": "/usr/bin/python3"
      }
    }
  }
}
```

```
#!/bin/bash
sudo pip install numpy
sudo pip install pandas
sudo pip install Pillow
sudo pip install findspark
sudo pip install pyarrow==0.15.1
sudo python3 -m pip install numpy
sudo python3 -m pip install Pillow
sudo python3 -m pip install findspark
sudo python3 -m pip install pandas
sudo python3 -m pip install pyarrow==0.15.1
```

1_SUCCESS.txt	23/04/2021 16:16	Fichier CRC	1 Ko
part-0000-8137674-4899-4760-Safe-cs...	23/04/2021 16:16	Fichier CRC	6 Ko
part-0001-8137674-4899-4760-Safe-cs...	23/04/2021 16:16	Fichier CRC	6 Ko
part-0002-8137674-4899-4760-Safe-cs...	23/04/2021 16:16	Fichier CRC	6 Ko
part-0003-8137674-4899-4760-Safe-cs...	23/04/2021 16:16	Fichier CRC	6 Ko
part-0004-8137674-4899-4760-Safe-cs...	23/04/2021 16:16	Fichier CRC	6 Ko
part-0005-8137674-4899-4760-Safe-cs...	23/04/2021 16:16	Fichier CRC	6 Ko
part-0006-8137674-4899-4760-Safe-cs...	23/04/2021 16:16	Fichier CRC	6 Ko
part-0007-8137674-4899-4760-Safe-cs...	23/04/2021 16:16	Fichier CRC	6 Ko
part-0008-8137674-4899-4760-Safe-cs...	23/04/2021 16:16	Fichier CRC	6 Ko
part-0009-8137674-4899-4760-Safe-cs...	23/04/2021 16:16	Fichier CRC	6 Ko
part-0010-8137674-4899-4760-Safe-cs...	23/04/2021 16:16	Fichier CRC	6 Ko
part-0011-8137674-4899-4760-Safe-cs...	23/04/2021 16:16	Fichier CRC	6 Ko
part-0012-8137674-4899-4760-Safe-cs...	23/04/2021 16:16	Fichier CRC	6 Ko
part-0013-8137674-4899-4760-Safe-cs...	23/04/2021 16:16	Fichier CRC	6 Ko
part-0014-8137674-4899-4760-Safe-cs...	23/04/2021 16:16	Fichier CRC	6 Ko
part-0015-8137674-4899-4760-Safe-cs...	23/04/2021 16:16	Fichier CRC	6 Ko

S3

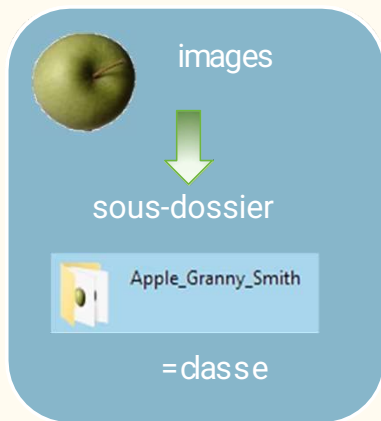


s3://p8bucket

Pipeline du projet

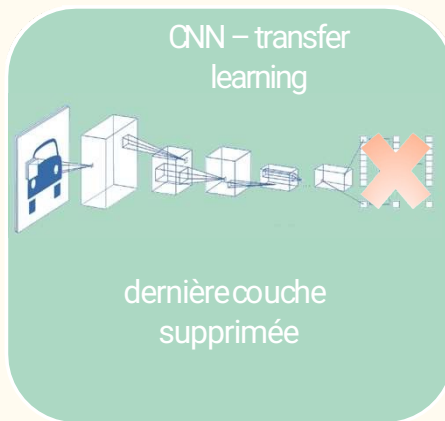


1



Lecture
Chargement

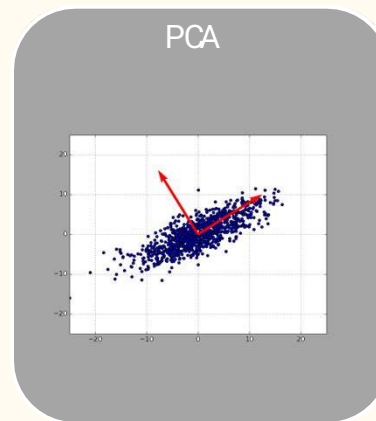
2



Extraction des
features

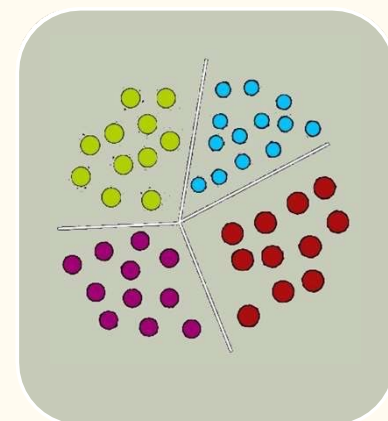
Xception
(imagenet)

3



Réduction de
dimension

4

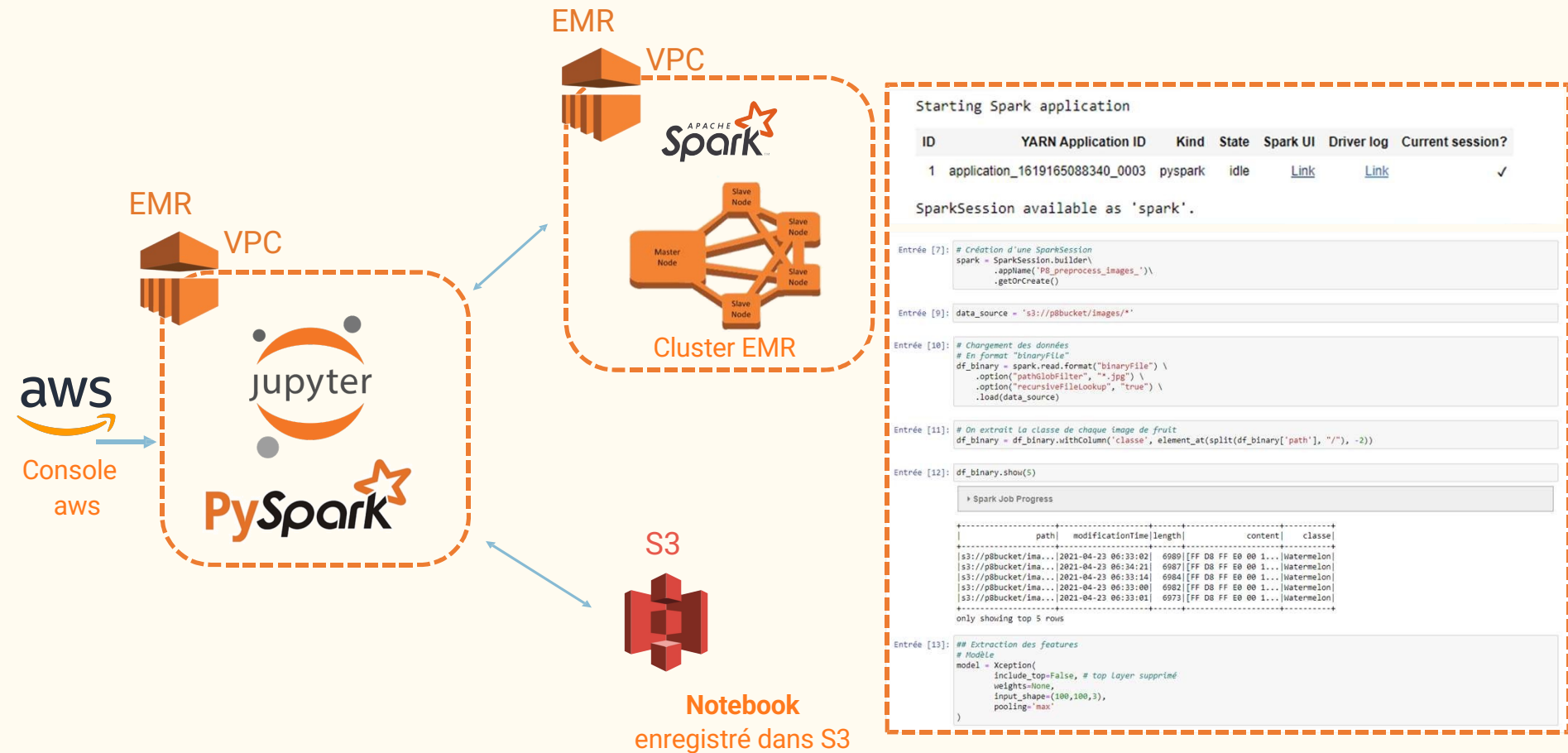


Classification

Random Forest :

- Accuracy : 0.911
- Precision : 0.925

Amazon EMR notebooks



Enregistrement des données

s3://p8bucket/resultats_parquet

Images

Classes

Features
array

Features
vectors

Vecteurs
PCA

Enregistré au format
distribué

S3



path	classe	X_features	X_vectors	X_vectors_pca
s3://p8bucket/ima...	Strawberry	[7.528077E-5, 5.7...	[7.52807682147249...	[-0.0024979452002...
s3://p8bucket/ima...	Peach	[7.164011E-5, 4.6...	[7.16401118552312...	[-0.0022347604461...
s3://p8bucket/ima...	Peach	[7.0007634E-5, 4...	[7.00076343491673...	[-0.0022332804300...
s3://p8bucket/ima...	Orange	[5.2437896E-5, 5...	[5.24378956470172...	[-0.0022457391958...
s3://p8bucket/ima...	Apple_Red_1	[6.39236E-5, 5.58...	[6.39236022834666...	[-0.0022016862054...
s3://p8bucket/ima...	Peach	[6.562176E-5, 4.3...	[6.56217598589137...	[-0.0018781298588...
s3://p8bucket/ima...	Apple_Red_1	[6.592149E-5, 6.1...	[6.592149293283E...	[-0.0022221958511...
s3://p8bucket/ima...	Strawberry	[8.3148494E-5, 5...	[8.31484940135851...	[-0.0025633449203...
s3://p8bucket/ima...	Peach	[5.6137997E-5, 4...	[5.61379965802188...	[-0.0023045842887...
s3://p8bucket/ima...	Peach	[4.3400338E-5, 5...	[4.34003377449698...	[-0.0020065563145...
s3://p8bucket/ima...	Strawberry	[5.3844553E-5, 4...	[5.38445528945885...	[-0.0024053836598...
s3://p8bucket/ima...	Peach	[7.0101734E-5, 5...	[7.01017343089915...	[-0.0020346894213...
s3://p8bucket/ima...	Peach	[7.544263E-5, 5.2...	[7.54426291678100...	[-0.0019785420552...
s3://p8bucket/ima...	Strawberry	[6.331178E-5, 5.0...	[6.33117815596051...	[-0.0023900177974...
s3://p8bucket/ima...	Peach	[4.662218E-5, 6.2...	[4.66221790702547...	[-0.0023086338846...
s3://p8bucket/ima...	Strawberry	[8.312277E-5, 6.4...	[8.31227735034190...	[-0.0025797946202...
s3://p8bucket/ima...	Pear	[6.2984975E-5, 4...	[6.29849746474064...	[-0.0018198891807...
s3://p8bucket/ima...	Strawberry	[8.780114E-5, 6.1...	[8.78011414897628...	[-0.0025548053255...
s3://p8bucket/ima...	Peach	[6.663117E-5, 5.0...	[6.66311680106446...	[-0.0023358706523...
s3://p8bucket/ima...	Peach	[5.9448852E-5, 5...	[5.94488519709557...	[-0.0023381211192...



Parquet

._SUCCESS.crc	23/04/2021 16:16	Fichier CRC	1 Ko
.part-00000-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00001-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00002-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00003-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00004-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00005-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00006-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00007-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00008-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00009-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00010-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00011-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00012-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00013-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00014-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko
.part-00015-81379a74-4d99-47e0-8dfe-c9...	23/04/2021 16:16	Fichier CRC	6 Ko

Bonne compression , conçu pour
les données massives

Nom

app_P8.py

configuration.json

emr_bootstrap.sh

images/

logs/

resultat_parquet/

Conclusion

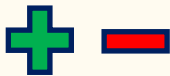
Architecture retenue - Passage à l'échelle

Clef IAM



Modifications éventuelles:
selon souhait client

EC2



Evolutions à prévoir: Passage
à l'échelle automatique (EMR)

EMR



Infrastructure:
Scripts Pyspark

S3



Aucun changement:
stockage illimité

Montée en compétence – Difficultés rencontrées

- Découverte de l'écosystème Hadoop, du moteur de traitement de données massives Spark
- Prise en main de Pyspark, et du système d'exploitation Linux (Ubuntu 20.04 LTS)
- Découverte de l'écosystème AWS
- Peu explicites pour les non initiés, nombreuses erreurs possibles
- Possibilités techniques nombreuses : difficile avec peu d'expérience d'être assuré d'avoir fait le bon choix !

Perspectives – Améliorations possibles

- Scripts en scala (meilleur performance)
- Entrainement en situation réel
- Gpu versus Cpu ! Réflexions à mener : Spark 3 Demo: Comparing Performance of GPUs vs. CPUs

<https://www.youtube.com/watch?v=tGqEZYUqexY> (video nvidia...)

Bibliographie

- Xception: Deep learning with ... - openaccess.thecvf.com. (n.d.). Retrieved March 16, 2022, from https://openaccess.thecvf.com/content_cvpr_2017/papers/Chollet_Xception_Deep_Learning_CVPR_2017_paper.pdf
- Jupyter notebook server with AWS EC2 and AWS VPC // blog // coding for entrepreneurs. Coding for Entrepreneurs. (n.d.). Retrieved March 16, 2022, from <https://www.codingforentrepreneurs.com/blog/jupyter-notebook-server-aws-ec2-aws-vpc>

Annexes



Création d'un cluster EMR

Configuration

S3



Ajout d'étape

Amorçage

Clés

Configuration des logiciels

Libérer | emr-6.1.0

<input checked="" type="checkbox"/> Hadoop 3.2.1	<input type="checkbox"/> Zeppelin 0.9.0	<input checked="" type="checkbox"/> Livy 0.7.0
<input type="checkbox"/> JupyterHub 1.1.0	<input type="checkbox"/> Tez 0.9.2	<input type="checkbox"/> Flink 1.11.0
<input type="checkbox"/> Ganglia 3.7.2	<input type="checkbox"/> HBase 2.2.5	<input type="checkbox"/> Pig 0.17.0
<input type="checkbox"/> Hive 3.1.2	<input type="checkbox"/> Presto 0.232	<input type="checkbox"/> PrestoSQL 338
<input type="checkbox"/> ZooKeeper 3.4.14	<input type="checkbox"/> MXNet 1.6.0	<input type="checkbox"/> Sqoop 1.4.7
<input type="checkbox"/> Hue 4.7.1	<input type="checkbox"/> Phoenix 5.0.0	<input type="checkbox"/> Oozie 5.2.0
<input checked="" type="checkbox"/> Spark 3.0.0	<input type="checkbox"/> HCatalog 3.1.2	<input checked="" type="checkbox"/> TensorFlow 2.1.0

Modifier les paramètres du logiciel

☐ Entrer la configuration ☒ Charger JSON à partir de S3

s3://p8fruitsbucket/configuration.json

Ajout d'étape

Type d'étape: Application Spark

Nom	Action sur échec	Emplacement JAR	Arguments
Application Spark	Continuer	command-runner.jar	spark-submit --deploy-mode client s3://p8fruitsbucket/app_P8.py --data_source s3://p8fruitsbucket/data/* --output_uri s3://p8fruitsbucket/resultats_parquets

Amorçage

Actions d'amorçage

Type d'action d'amorçage	Nom	Emplacement JAR	Arguments facultatifs
Action personnalisée	Action personnalisée	s3://p8fruitsbucket/emr_bootstrap.sh	

Ajouter une action d'amorçage: Action personnalisée

Options de sécurité

Paire de clés EC2

☒ Cluster visible pour tous les utilisateurs IAM du compte



configuration.json

Arguments

```
spark-submit --deploy-mode client s3://p8fruitsbucket/app_P8.py --data_source s3://p8fruitsbucket/data/* --output_uri s3://p8fruitsbucket/resultats_parquets
```

emr_bootstrap.sh

m5.xlarge

EC2



EMR



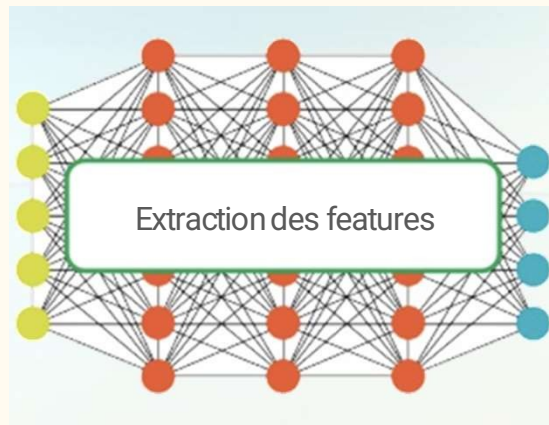
Vcpu : 4
Mémoire : 16 Gio

Pipelines - classification

Chargement
des images



Preprocessing



Régression
logistique



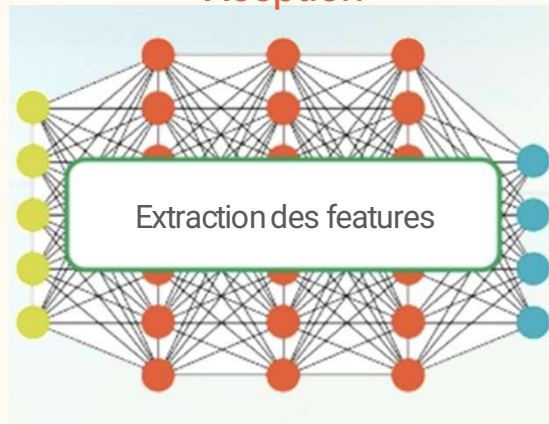
Accuracy:0.87
Précision:0.87

Xception

Chargement
des images



Preprocessing



Random
Forest



Accuracy:0.91
Précision:0.93

Logout de l'application enregistré sur S3

```
+-----+-----+-----+-----+
| path | classe | X_vectors | X_vectors_pca |
+-----+-----+-----+-----+
| s3://p8bucket/ima... | Watermelon | [4.30195686931256... | [-0.0014724724050... |
| s3://p8bucket/ima... | Watermelon | [3.85823259421158... | [-0.0014240677179... |
| s3://p8bucket/ima... | Watermelon | [2.91877313429722... | [-0.0013982374124... |
| s3://p8bucket/ima... | Watermelon | [4.27223603765014... | [-0.0013877593994... |
| s3://p8bucket/ima... | Watermelon | [3.48064459103625... | [-0.0013936819953... |
| s3://p8bucket/ima... | Clementine | [4.13163252233061... | [-0.0023987243574... |
| s3://p8bucket/ima... | Watermelon | [2.83777862932765... | [-0.0015022272896... |
| s3://p8bucket/ima... | Watermelon | [3.01670806948095... | [-0.0014527697217... |
| s3://p8bucket/ima... | Clementine | [3.88991320505738... | [-0.0024666961772... |
| s3://p8bucket/ima... | Clementine | [4.27966697316151... | [-0.0025012090477... |
| s3://p8bucket/ima... | Clementine | [4.84576812596060... | [-0.0024858020994... |
| s3://p8bucket/ima... | Clementine | [5.10514291818253... | [-0.0025488732644... |
| s3://p8bucket/ima... | Clementine | [4.07514417020138... | [-0.0023361143739... |
| s3://p8bucket/ima... | Clementine | [5.23504895681981... | [-0.0024441455624... |
| s3://p8bucket/ima... | Strawberry | [3.92002475564368... | [-0.0016503117639... |
| s3://p8bucket/ima... | Clementine | [4.92607614432927... | [-0.0023981603203... |
| s3://p8bucket/ima... | Orange | [3.83260485250502... | [-0.0020354908982... |
| s3://p8bucket/ima... | Orange | [3.81613535864744... | [-0.0020261696806... |
| s3://p8bucket/ima... | Apple_Red_1 | [4.32711349276360... | [-0.0014783001823... |
| s3://p8bucket/ima... | Strawberry | [3.16732912324368... | [-0.0016672352046... |
+-----+-----+-----+-----+
only showing top 20 rows
```

Mini-classification

```
+-----+-----+
| classe | X_vectors_pca |
+-----+-----+
| Apricot | [-0.0017985031932... |
| Lemon | [-0.0023381406432... |
| Peach | [-0.0017537706410... |
+-----+-----+
only showing top 3 rows

+-----+-----+-----+
| classe | X_vectors_pca | labelIndex |
+-----+-----+-----+
| Watermelon | [-0.0013619884910... | 9.0 |
| Watermelon | [-0.0014813502034... | 9.0 |
| Watermelon | [-0.0014462156936... | 9.0 |
+-----+-----+-----+
only showing top 3 rows

Training Dataset Count: 3396
Test Dataset Count: 1490
```

Regression logistique

```
+-----+-----+-----+
| prediction | labelIndex |
+-----+-----+-----+
| 0.0 | 0.0 |
| 0.0 | 0.0 |
| 0.0 | 0.0 |
| 0.0 | 0.0 |
| 0.0 | 0.0 |
+-----+-----+-----+
only showing top 5 rows

Test Error = 0.126846
Accuracy = 0.873154
Test Error = 0.12471
Precision = 0.87529
```

Random forest

```
+-----+-----+-----+
| prediction | labelIndex |
+-----+-----+-----+
| 0.0 | 0.0 |
| 0.0 | 0.0 |
| 0.0 | 0.0 |
| 0.0 | 0.0 |
| 0.0 | 0.0 |
+-----+-----+-----+
only showing top 5 rows

Test Error = 0.0885906
Accuracy = 0.911409
Test Error = 0.0741965
Precision = 0.925803
```