Fruits!

EA Stephen
Data scientist

Déployer un modèle dans le cloud

Fruits!

SOMMAIRE

Fruits!

1. Introduction

1.1 Mission
1.2 Le jeu de données

Fruits!

1.1 Mission

Fruits est une start-up de l'AgriTech souhaitant mettre à disposition une application mobile permettant au grand public d'avoir un moteur de classification d'images de fruits/légumes.

Développer dans un environnement Big data une première chaîne de traitement des données (preprocessing des images et réduction dimension)

Il faut prendre en compte que le volume des données va augmenter rapidement après la livraison de ce projet.

Il faut donc adapter une architecture cloud en fonction du volume et du script pyspark pour la première chaîne de traitement

1.2 Le jeu de données

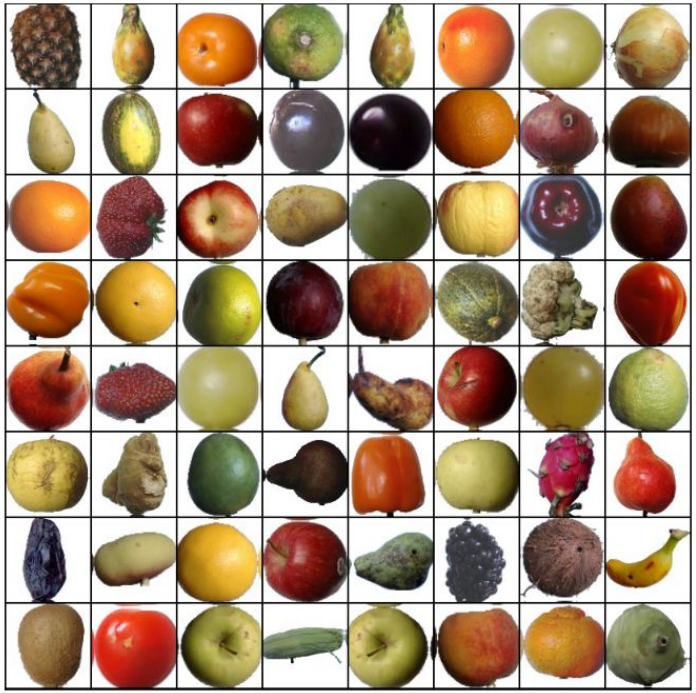
Le nombre total d'images : 90483.

Taille de l'ensemble d'entraînement : 67 692 images (un fruit ou un légume par image).

Taille de l'ensemble de test : 22 688 images (un fruit ou un légume par image).

Le nombre de classes : 131 (fruits et légumes).

Taille de l'image : 100 x 100 pixels.



Fruits!

2. Scalabilité

2.1 Evaluation du besoin de FRUITS par rapport au 3V’s
2.2 Evolution des architectures proposées pour FRUITS

Fruits!

2.1 Evaluation du besoin de FRUITS par au 3V's

Volume

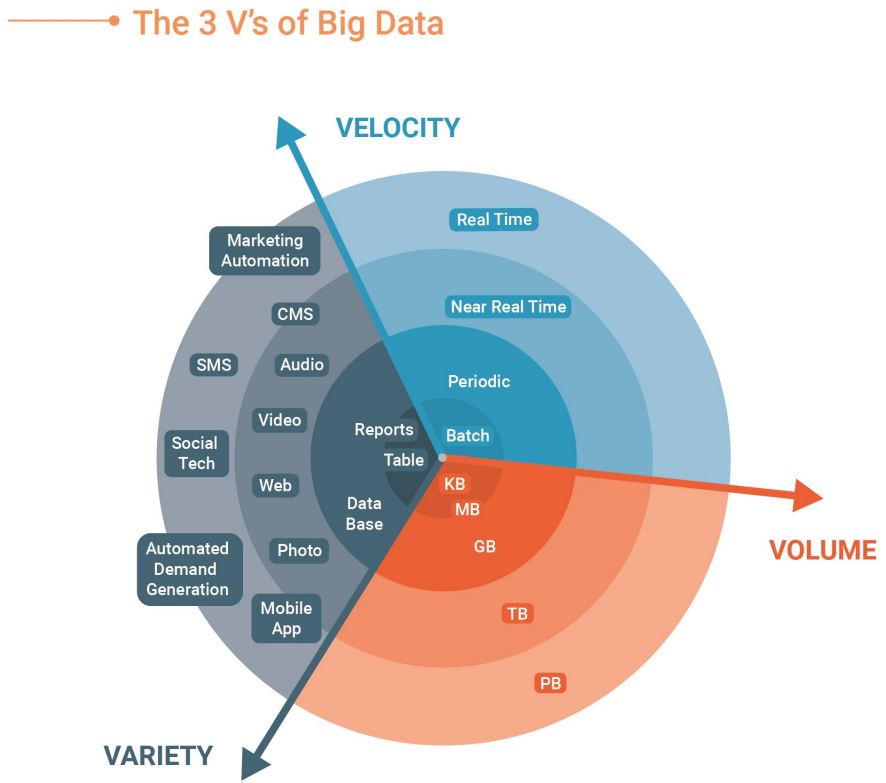
La quantité de données astronomiques générées par les entreprises est en constante augmentation

Vélocité

La vélocité fait référence à la vitesse à laquelle les données sont générées. Les données à grande vitesse sont générées à un rythme tel qu'elles nécessitent des techniques de traitement distribuées (hadoop,Spark).

Variété

Les données ont des types différents structurés (databases),semi-structurés (Csv, Json, Xml), non structurés (images, textes, vidéo, audios)



Fruits!

2.2 Evolution des architectures proposées pour FRUITS

single instance

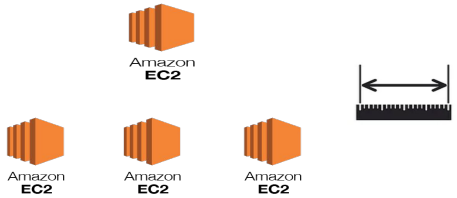
Scaling Up: Augmentation du type d'instance(besoins de calcul, de mémoire, de réseaux ou de stockage.)

Avantage : Le coût

Désavantage: Possible latence

Désavantage : pas tolérance au pannes, distribution local spark, Il ne nécessite aucun gestionnaire de ressources.(Spark s'exécute localement avec autant de threads de travail que de cœurs logiques sur votre machine.)

Cluster EC2



Scaling out : Augmentation cluster besoin de calcul, de mémoire, de réseaux ou de stockage.

Avantages : Le coût comparé à serverless

tolérance au pannes , moins de latence, distribution des calculs sur plusieurs machines (hadoop, spark), réplications des données partitionnées

Désavantages : Mis en oeuvre plus complexe (docker swarm, Kubernetes)

EMR (serverless)

Scaling out : Augmentation cluster besoins de calcul, de mémoire, réseaux ou de stockage.

Désavantages : le coût

Avantages: moins de latence, distribution des calculs sur plusieurs machines (hadoop, spark)

Tolérance au pannes , réplications des données partitionnées

Préconfiguré (pas de gestion serveurs, pas d'installation , pas configuration, pas de maintenance)

Fruits!

3. **Architecture et environnement**

3.1 EC2 single
3.2 EC2 cluster
3.3 EMR

Fruits!

3.1 EC2 single

3.1.1 L’instance
3.1.2 Environnement containeriser
3.1.3 La configuration de l’environnement big data

Fruits!

3.1.1 L'instance

EC2 description

- AMI (Ubuntu server 22.04)

- Type de l'instance (t2.xlarge)
  - 4 vCPU
  - 16 Gio Mémoire
  - 30 Gio stockage

- Paire de clés (connexion SSH)

- Groupe de sécurité port
  - Port 22 (SSH)
  - Port 8888 (Jupyterlab)
  - Protocole TCP
  - Source IP 0.0.0.0/0 (entrante, sortante)

- Rôle IAM
  - Full Access Aws service et ressource

- Non
  - IP dynamique

- Région
  - eu-west-3

3.1.2 Environnement containeriser

1

```
# Copyright (c) Jupyter Development Team.
# Distributed under the terms of the Modified BSD License.
ARG OWNER=jupyter
ARG BASE_CONTAINER=$OWNER/scipy-notebook
FROM $BASE_CONTAINER

LABEL maintainer="Jupyter Project <jupyter@googlegroups.com>"

# Fix DL4006
SHELL ["/bin/bash", "-o", "pipefail", "-c"]

USER root

# Spark dependencies
# Default values can be overridden at build time
# (ARGS are in lower case to distinguish them from ENV)
ARG spark_version="3.2.1"
ARG hadoop_version="3.2"
ARG spark_checksum="145ADACF189FECF05FBA3A69841D2804DD66546B11D14FC181AC49D89F3CB5E4FECD9B25F56F0AF767155419CD430838FB651992AEB37D3A6F91E7E009D1F9AE"
ARG openjdk_version="8"

ENV APACHE_SPARK_VERSION="${spark_version}" \
    HADOOP_VERSION="${hadoop_version}"
ENV JUPYTER_ENABLE_LAB=yes
RUN apt-get update --yes && \
    apt-get install --yes --no-install-recommends \
    "openjdk-${openjdk_version}-jre-headless" \
    ca-certificates-java && \
    apt-get clean && rm -rf /var/lib/apt/lists/*

# Spark installation
WORKDIR /tmp
RUN wget -q "https://archive.apache.org/dist/spark/spark-${APACHE_SPARK_VERSION}/spark-${APACHE_SPARK_VERSION}-bin-hadoop${HADOOP_VERSION}.tgz" && \
    echo "${spark_checksum} *spark-${APACHE_SPARK_VERSION}-bin-hadoop${HADOOP_VERSION}.tgz" | sha512sum -c - && \
    tar xzf "spark-${APACHE_SPARK_VERSION}-bin-hadoop${HADOOP_VERSION}.tgz" -C /usr/local --owner root --group root --no-same-owner && \
    rm "spark-${APACHE_SPARK_VERSION}-bin-hadoop${HADOOP_VERSION}.tgz"

WORKDIR /usr/local
```

Image jupyter-notebook

Spécifier version spark, hadoop, Installer java (compatible dépendances)

Installer spark

2

```
# Configure Spark
ENV SPARK_HOME=/usr/local/spark
ENV SPARK_OPTS="--driver-java-options=-Xms1024M --driver-java-options=-Xmx4096M --driver-java-options=-Dlog4j.logLevel=info" \
    PATH="${PATH}:${SPARK_HOME}/bin"

RUN ln -s "spark-${APACHE_SPARK_VERSION}-bin-hadoop${HADOOP_VERSION}" spark && \
    # Add a link in the before_notebook hook in order to source automatically PYTHONPATH
    mkdir -p /usr/local/bin/before-notebook.d && \
    ln -s "${SPARK_HOME}/sbin/spark-config.sh" /usr/local/bin/before-notebook.d/spark-config.sh

# Fix Spark installation for Java 11 and Apache Arrow library
# see: https://github.com/apache/spark/pull/27356, https://spark.apache.org/docs/latest/#downloading
RUN cp -p "${SPARK_HOME}/conf/spark-defaults.conf.template" "${SPARK_HOME}/conf/spark-defaults.conf" && \
    echo 'spark.driver.extraJavaOptions -Dio.netty.tryReflectionSetAccessible=true' >> "${SPARK_HOME}/conf/spark-defaults.conf" && \
    echo 'spark.executor.extraJavaOptions -Dio.netty.tryReflectionSetAccessible=true' >> "${SPARK_HOME}/conf/spark-defaults.conf"

WORKDIR /jars
ENV SPARK_jars=/usr/local/spark/jars

RUN wget https://repo1.maven.org/maven2/com/amazonaws/aws-java-sdk-bundle/1.12.226/aws-java-sdk-bundle-1.12.226.jar

# Get Hadoop-AWS Jar

RUN wget https://repo1.maven.org/maven2/org/apache/hadoop/hadoop-aws/3.3.1/hadoop-aws-3.3.1.jar

# Get jets3t JAR
RUN wget https://repo1.maven.org/maven2/net/java/dev/jets3t/jets3t/0.9.4/jets3t-0.9.4.jar

USER ${NB_UID}
```

Configuration spark

Télécharger fichier .jar pour accéder à S3 avec les fonctionnalités de spark (aws-sdk-bundle, hadoop-aws, jets3t)

3

```
# Install pyarrow
RUN mamba install --quiet --yes \
    'pyarrow' && \
    mamba clean --all -f -y && \
    fix-permissions "${CONDA_DIR}" && \
    fix-permissions "/home/${NB_USER}"

RUN pip3 install awscli
RUN pip3 install s3fs
RUN pip3 install fsspec
RUN pip3 install boto3
RUN pip3 install botocore

WORKDIR "${HOME}"
```

Install autres packages

3.1.3 La configuration de l'environnement big data

Amazon EC2

docker Image

RUN Container avec credentials Aws

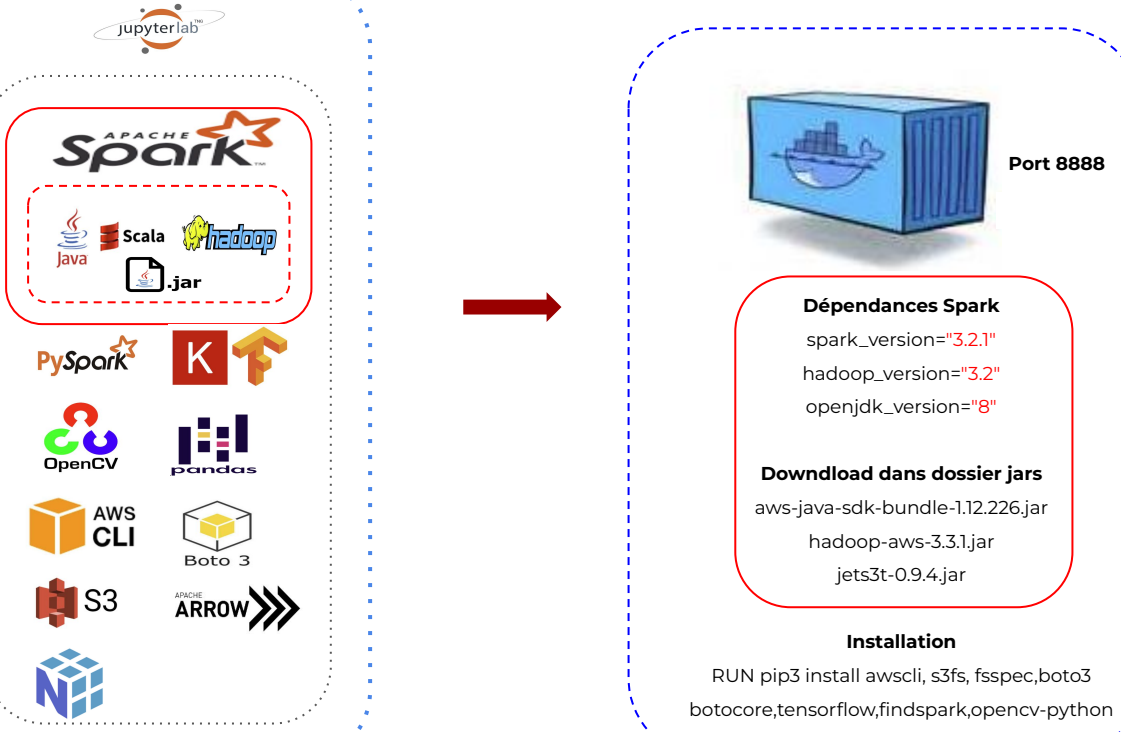
Visual Studio Code

SSH

Paire de clés connexion SSH

DNS IPV4 public (dynamique)

Port 22 (SSH)



Port 8888

Dépendances Spark

spark_version="3.2.1"

hadoop_version="3.2"

openjdk_version="8"

Downdload dans dossier jars

aws-java-sdk-bundle-1.12.226.jar

hadoop-aws-3.3.1.jar

jets3t-0.9.4.jar

Installation

RUN pip3 install awscli, s3fs, fsspec,boto3 botocore,tensorflow,findspark,opencv-python

Fruits!

3.2 EC2 cluster

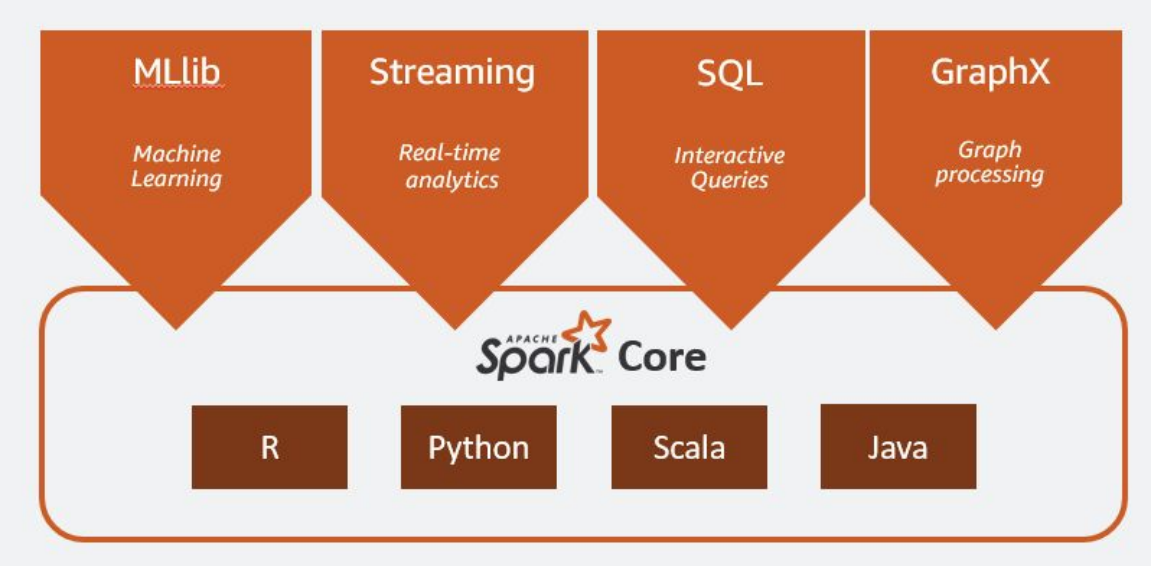
3.2.1 Apache Spark eco system
3.2.2 Les structures de données dans Spark
3.2.3 Comment fonctionne Spark
3.2.4 L’architecture big data

Fruits!

3.2.1 Apache Spark eco system

Le framework Spark comprend :

- Spark Core comme base de la plateforme
- Spark SQL pour les requêtes interactives
- Spark Streaming pour des analyses en temps réel
- Spark MLlib pour l'apprentissage automatique
- Spark GraphX pour le traitement de graphes

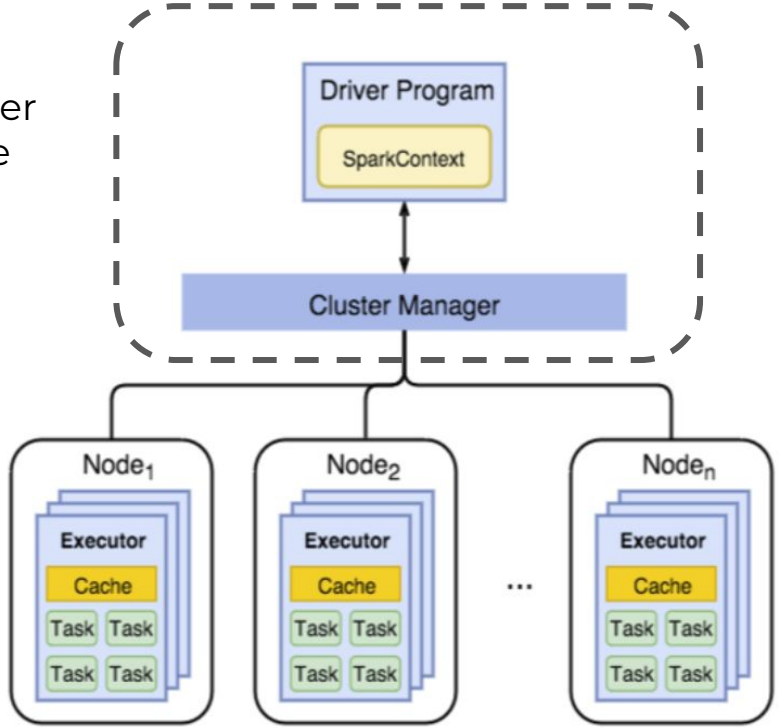


Fruits!

3.2.2 Les structures de données dans Spark

	RDD	Dataframe	Dataset
Immutable	**Oui**	**Oui**	**Oui**
Tolérance panne	**Oui**	**Oui**	**Oui**
Type safe	**Oui**	**Non**	**Oui**
Optimizer	**Non**	**Catalyst Engine**	**Catalyst Engine**
Memory usage	**High**	Faible en raison de la mémoire hors tas binaire Tungsten	Faible en raison de la mémoire hors tas binaire Tungsten
Exécution	**Lente**	Augmenté grâce à des plans d'exécution optimisés à l'aide d'un catalyseur Engine	Augmenté grâce à des plans d'exécution optimisés à l'aide d'un catalyseur Engine
Lazy evaluation	**Oul**	**Oui**	**Oui**
Schéma	Définir manuellement schéma	**Automatique**	**Automatique**

3.2.3 Fonctionnement de spark



Driver program (pilote)

- Le pilote est composé du programme, et une session Spark (entrée vers les fonctionnalités de spark). Il est le responsable de l'exécution du code sur les différentes machines.

Cluster manager

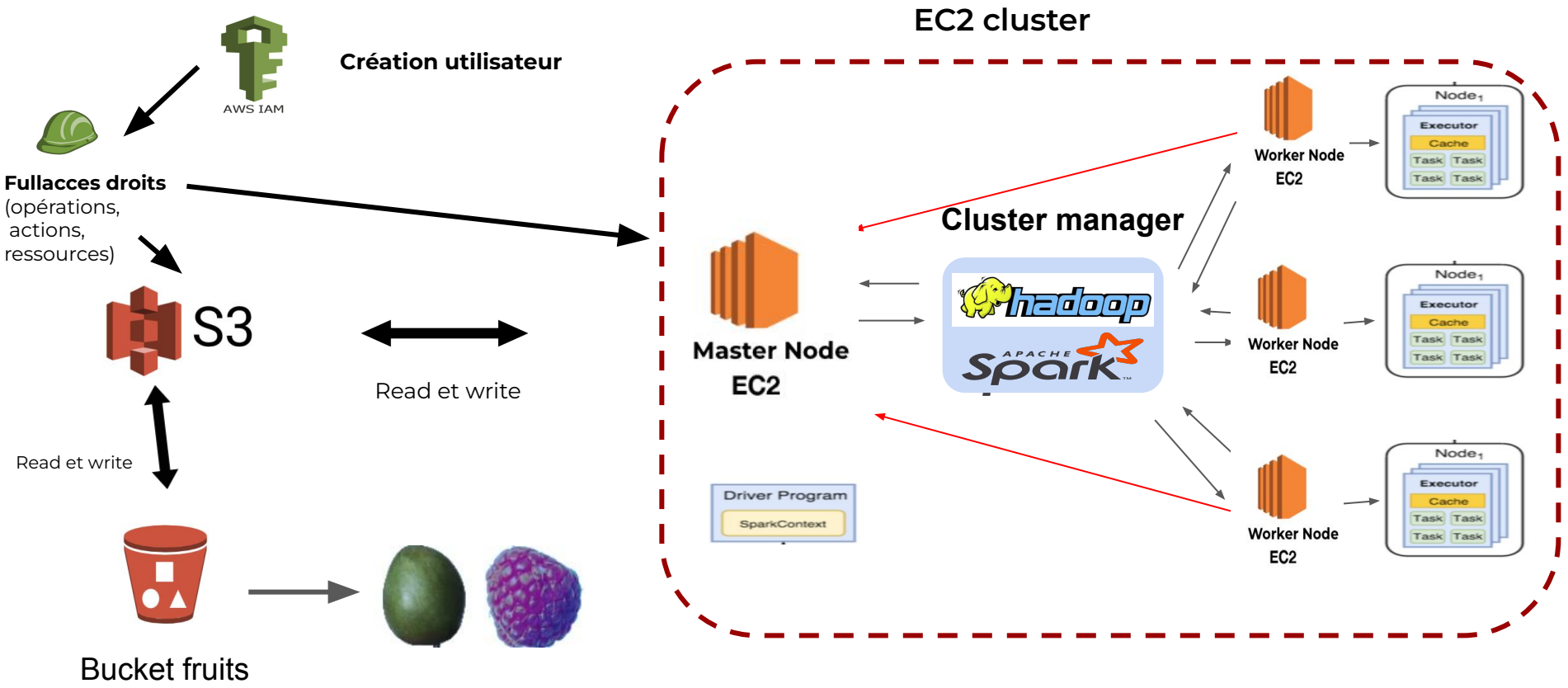
- Gère l'allocation de ressources, la division des programmes, l'exécution du programme (Standalone, Yarn, Mesos)

Workers nodes

- Chaque exécuteur, ou nodes Worker, reçoit une tâche du pilote et exécute cette tâche.
- Les données sont répliquées et conservées en mémoire vive (cache) afin que l'application soit tolérante au panne(Spark s'exécute 100 fois plus vite en mémoire et 10 fois plus vite sur disque que hadoop)

Fruits!

3.2.4 L'architecture big data



Fruits!

3.3 EMR

3.3.1 L’architecture big data EMR
3.3.2 Auto scaling EMR

3.3.1 L'architecture big data EMR

Détails de configuration

Étiquette de version : emr-6.7.0

Distribution Hadoop : Amazon

Applications : TensorFlow 2.4.1, Spark 3.2.1, JupyterEnterpriseGateway 2.1.0, JupyterHub 1.4.1

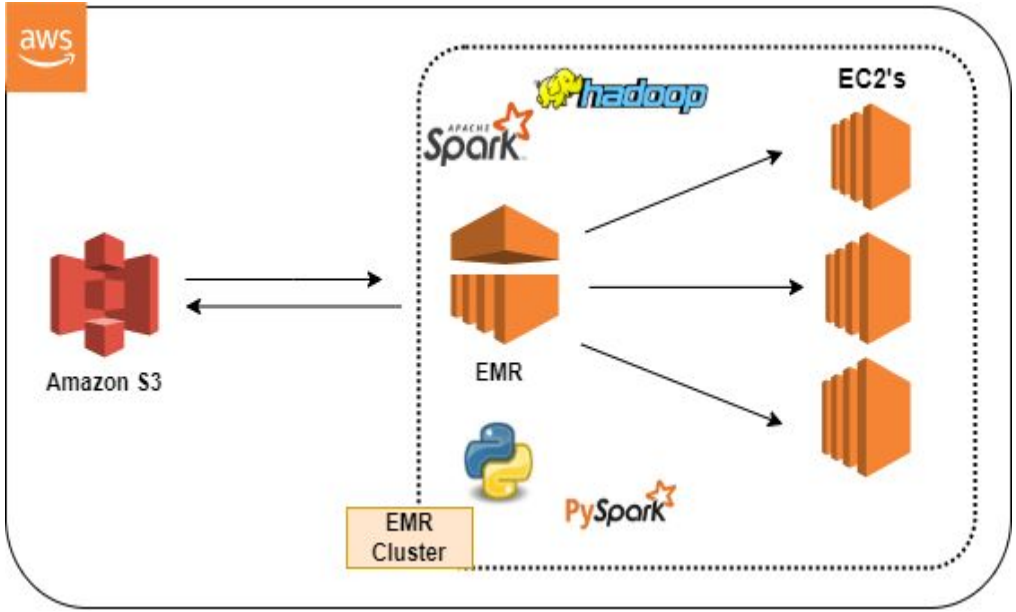
URI de connexion : s3://aws-logs-933635246190-eu-west-3/elasticmapreduce/

Vue cohérente EMRFS : Désactivé

ID d'AMI personnalisée : --

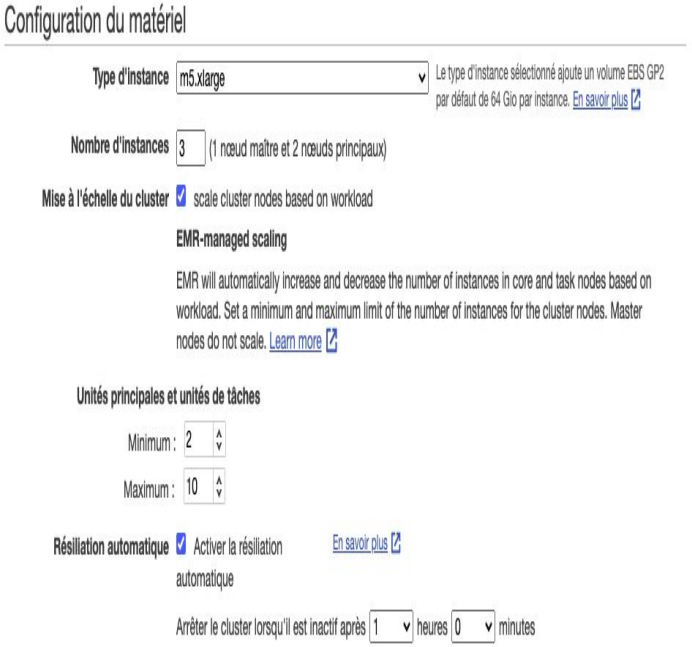
Version d'Amazon Linux : 2.0.20220719.0 En savoir plus

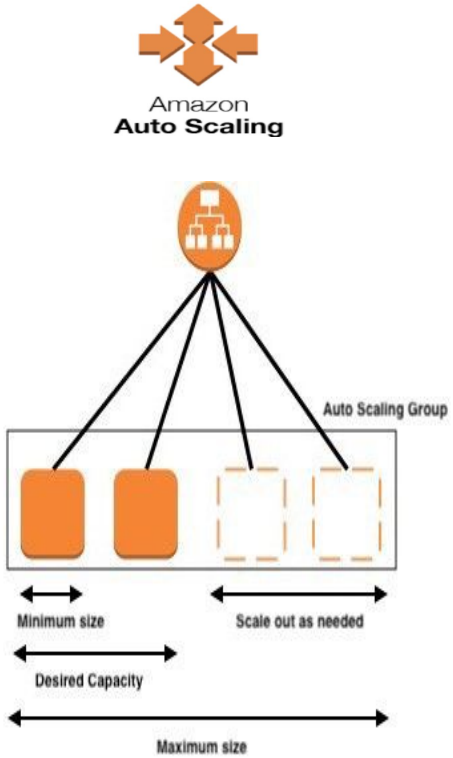
Nom et type de nœud	Type d'instance
MASTER Groupe d'instances maître - 1	**m5.xlarge** 4 Cœurs virtuels, 16 Gio de mémoire, stockage EBS uniquement Stockage sur EBS : 64 Gio
CORE Groupe d'instances principal - 2	**m5.xlarge** 4 Cœurs virtuels, 16 Gio de mémoire, stockage EBS uniquement Stockage sur EBS : 64 Gio



Fruits!

3.2.2 Auto scaling EMR





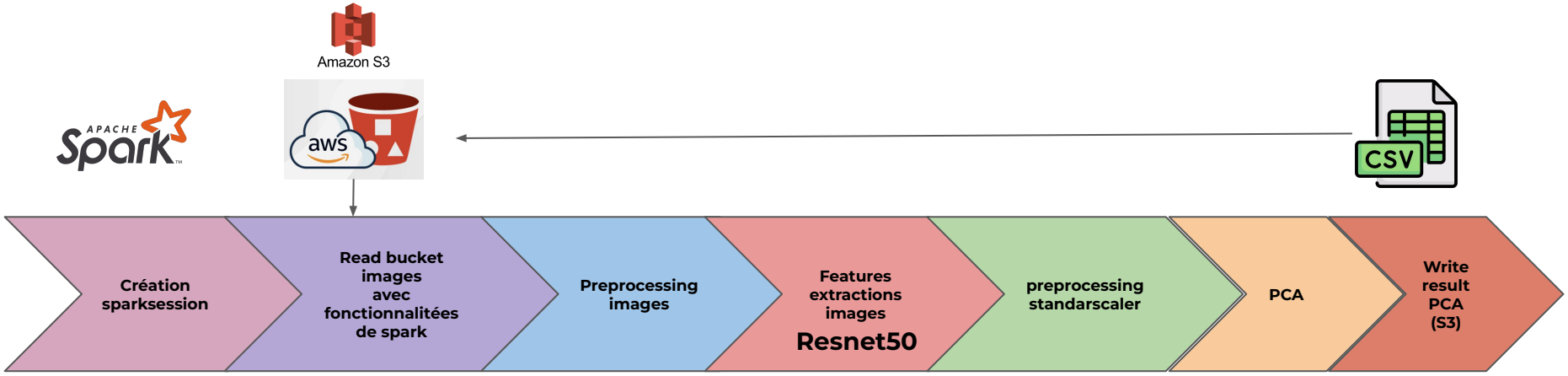
EMR augmentera et diminuera automatiquement le nombre d'instances dans les nœuds principaux et de tâches en fonction de la charge de travail. Définissez une limite minimale et maximale du nombre d'instances pour les nœuds de cluster. Les nœuds maîtres ne sont pas mis à l'échelle.

Fruits!

4. **Chaîne de traitement**

4.1 Les étapes de la chaîne de traitement
4.2 Features extraction Resnet50
4.3 Visualisation PCA
4.4 Write résultat PCA dans data lake (S3)

4.1 Les étapes de la chaîne de traitement



IDENTIFIANT	ID d'application YARN	Gentil	État	Interface utilisateur Spark	Journal du conducteur	Utilisateur	Session actuelle?
0	application_1663597751956_0001	pyspark	inactif	Lien	Lien	Aucun	✔

```
def read_S3_buckets_img(s3path):

    s3path = s3path

    # Read images avec le format binaryFile
    df_images = spark.read.format("binaryFile") \
      .option("pathGlobFilter", "*.jpg") \
      .option("recursiveFileLookup", "true") \
      .load(s3path)
    df_images.printSchema()

    return df_images
```

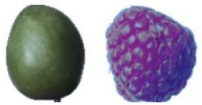
Resize

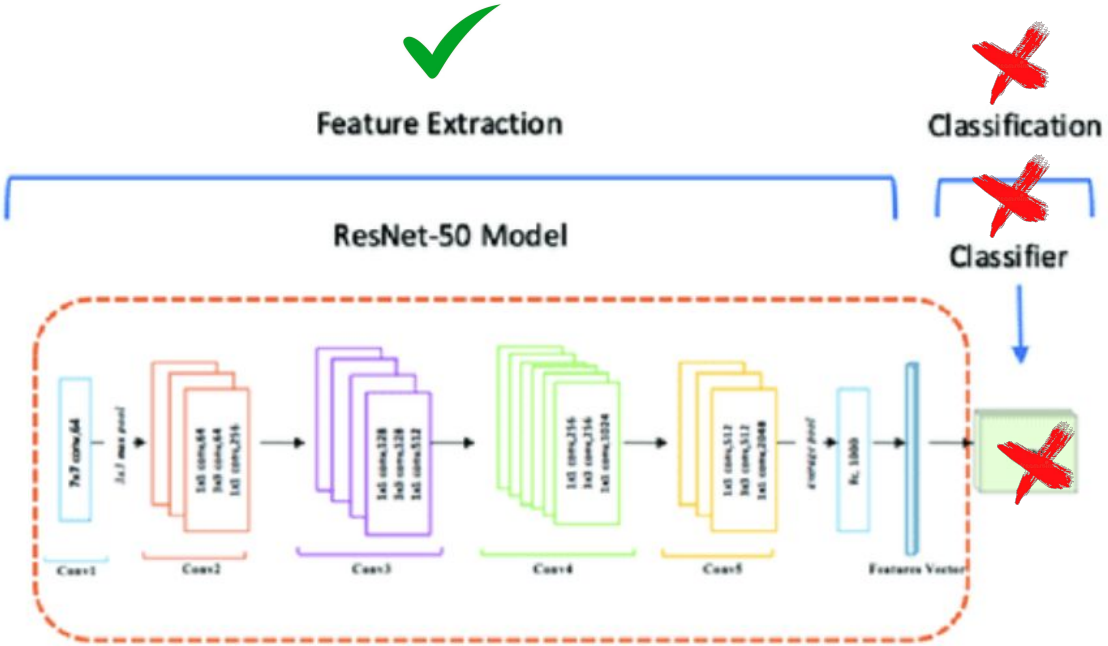
100x100

224x224

path	label	features_extraction_ResNet50	features_scaled_StandardScaler	features_pca
s3a://ocfruitsp8/...	raspberry	[0.0,0.0,0.0,0.0,...	[0.0,0.0,0.0,0.0,...	[98.3656270320505...
s3a://ocfruitsp8/...	raspberry	[0.0,0.0,0.0,0.0,...	[0.0,0.0,0.0,0.0,...	[88.7803466332628...
s3a://ocfruitsp8/...	raspberry	[0.0,0.0,0.0,0.0,...	[0.0,0.0,0.0,0.0,...	[95.7223947966706...
s3a://ocfruitsp8/...	raspberry	[0.0,0.0,0.0,0.0,...	[0.0,0.0,0.0,0.0,...	[104.156610510264...
s3a://ocfruitsp8/...	raspberry	[0.0,0.0,0.0,0.0,...	[0.0,0.0,0.0,0.0,...	[87.7651299642067...
s3a://ocfruitsp8/...	raspberry	[0.0,0.0,0.0,0.0,...	[0.0,0.0,0.0,0.0,...	[90.9680860429685...
s3a://ocfruitsp8/...	raspberry	[0.0,0.0,0.0,0.0,...	[0.0,0.0,0.0,0.0,...	[89.8776923367681...
s3a://ocfruitsp8/...	raspberry	[0.0,0.0,0.0,0.0,...	[0.0,0.0,0.0,0.0,...	[92.0226157636022...
s3a://ocfruitsp8/...	raspberry	[0.0,0.0,0.0,0.0,...	[0.0,0.0,0.0,0.0,...	[89.4623227934633...
s3a://ocfruitsp8/...	raspberry	[0.0,0.0,0.0,0.0,...	[0.0,0.0,0.0,0.0,...	[95.6694779208875...
s3a://ocfruitsp8/...	raspberry	[0.0,0.0,0.0,0.0,...	[0.0,0.0,0.0,0.0,...	[89.8530285023048...
s3a://ocfruitsp8/...	raspberry	[0.0,0.0,0.0,0.0,...	[0.0,0.0,0.0,0.0,...	[91.8247326805036...
s3a://ocfruitsp8/...	raspberry	[0.0,0.0,0.0,0.0,...	[0.0,0.0,0.0,0.0,...	[94.3949962773765...
s3a://ocfruitsp8/...	raspberry	[0.0,0.0,0.0,0.0,...	[0.0,0.0,0.0,0.0,...	[93.8360389146309...
s3a://ocfruitsp8/...	raspberry	[0.0,0.0,0.0,0.0,...	[0.0,0.0,0.0,0.0,...	[91.4965047227552...
s3a://ocfruitsp8/...	mango	[0.0,0.0,0.0,0.0,...	[0.0,0.0,0.0,0.0,...	[-84.965653169756...
s3a://ocfruitsp8/...	mango	[0.0,0.0,0.0,0.0,...	[0.0,0.0,0.0,0.0,...	[-86.810272337467...
s3a://ocfruitsp8/...	mango	[0.0,0.0,0.0,0.0,...	[0.0,0.0,0.0,0.0,...	[-84.4281326938 28...
s3a://ocfruitsp8/...	mango	[0.0,0.0,0.0,0.0,...	[0.0,0.0,0.0,0.0,...	[-88.516856670406...
s3a://ocfruitsp8/...	mango	[0.0,0.0,0.0,0.0,...	[0.0,0.0,0.0,0.0,...	[-106.13608345547...

only showing top 20 rows





4.2 Features extractions Resnet50

Input Images 224x224

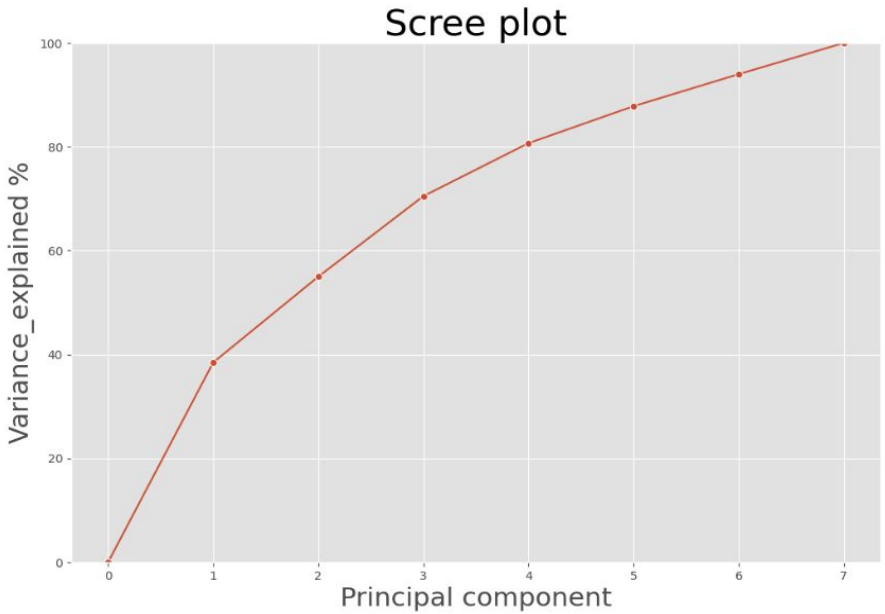
Output feature extraction

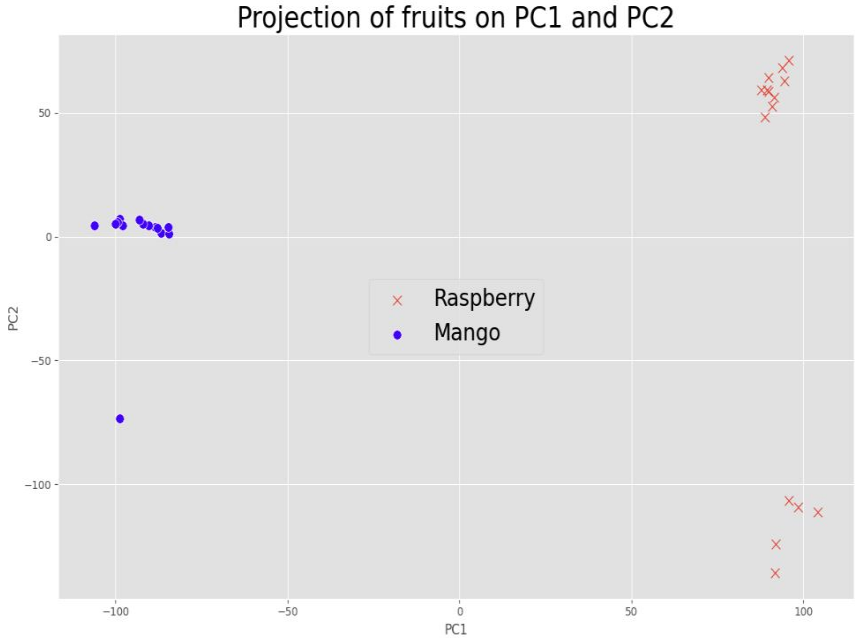
```
+-------------------+---------+--------------------------+
|               path|    label|features_extraction_ResNet50|
+-------------------+---------+--------------------------+
|s3a://ocfruitsp8/...|raspberry|      [0.0,0.0,0.0,0.0,...|
|s3a://ocfruitsp8/...|raspberry|      [0.0,0.0,0.0,0.0,...|
|s3a://ocfruitsp8/...|raspberry|      [0.0,0.0,0.0,0.0,...|
|s3a://ocfruitsp8/...|raspberry|      [0.0,0.0,0.0,0.0,...|
|s3a://ocfruitsp8/...|raspberry|      [0.0,0.0,0.0,0.0,...|
|s3a://ocfruitsp8/...|raspberry|      [0.0,0.0,0.0,0.0,...|
|s3a://ocfruitsp8/...|raspberry|      [0.0,0.0,0.0,0.0,...|
|s3a://ocfruitsp8/...|raspberry|      [0.0,0.0,0.0,0.0,...|
|s3a://ocfruitsp8/...|raspberry|      [0.0,0.0,0.0,0.0,...|
|s3a://ocfruitsp8/...|raspberry|      [0.0,0.0,0.0,0.0,...|
|s3a://ocfruitsp8/...|raspberry|      [0.0,0.0,0.0,0.0,...|
|s3a://ocfruitsp8/...|raspberry|      [0.0,0.0,0.0,0.0,...|
|s3a://ocfruitsp8/...|raspberry|      [0.0,0.0,0.0,0.0,...|
|s3a://ocfruitsp8/...|raspberry|      [0.0,0.0,0.0,0.0,...|
|s3a://ocfruitsp8/...|raspberry|      [0.0,0.0,0.0,0.0,...|
|s3a://ocfruitsp8/...|    mango|      [0.0,0.0,0.0,0.0,...|
|s3a://ocfruitsp8/...|    mango|      [0.0,0.0,0.0,0.0,...|
|s3a://ocfruitsp8/...|    mango|      [0.0,0.0,0.0,0.0,...|
|s3a://ocfruitsp8/...|    mango|      [0.0,0.0,0.0,0.0,...|
|s3a://ocfruitsp8/...|    mango|      [0.0,0.0,0.0,0.0,...|
+-------------------+---------+--------------------------+
only showing top 20 rows
```

Supprimer couches fully connected

Fruits!

4.3 Visualisation PCA



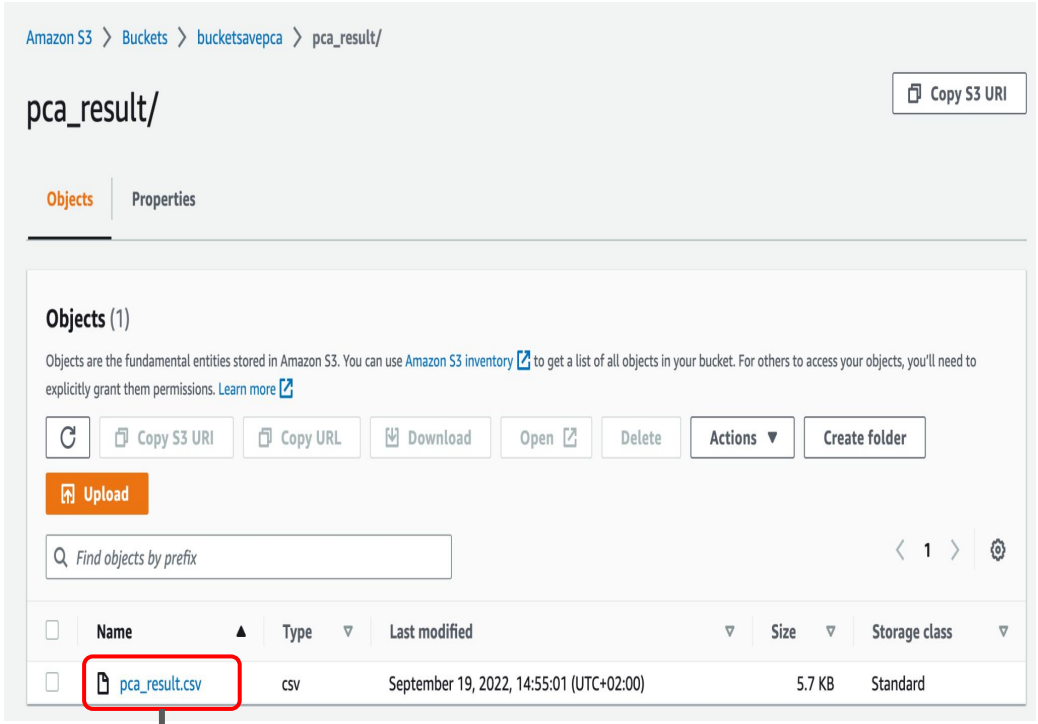


La variance expliquée sur un plan 2d (2 components) est de 55%

La variance pour chaques composantes : [0.3847107 0.54991919 0.70502387 0.80705559 0.87827463 0.93989402 1.]

Fruits!

4.4 Résultat sauvegarder dans S3



pca_result

path	label	features_pca
s3a://ocfruitsp8/fruits_emr/raspberry/10_100.jpg	raspberry	[98.36562703205053,-109.32981338579955,-25.630564756127914,-68.16995752607688,-7.013184573234865,62.64742396792836,103.52329083032754]
s3a://ocfruitsp8/fruits_emr/raspberry/122_100.jpg	raspberry	[88.78034663326284,48.41495427042843,12.025074053467311,-6.881883339088025,104.49574646740214,-4.299783485124397,6.4878242652442415]
s3a://ocfruitsp8/fruits_emr/raspberry/11_100.jpg	raspberry	[95.72239479667067,-106.50639570287791,-25.273432003063455,-73.17201881686206,-12.438760932559578,38.67947328428462,65.72690622580195]
s3a://ocfruitsp8/fruits_emr/raspberry/12_100.jpg	raspberry	[104.15661051026427,-111.18585027992576,-22.207237823926402,-150.67660374366002,-9.631762177403524,-76.28099702425423,-114.59589263445892]
s3a://ocfruitsp8/fruits_emr/raspberry/120_100.jpg	raspberry	[87.76512996420676,59.188563921846644,13.871435447545327,-1.4832735149657166,59.945045053181914,-3.7036716853420 76,1.3240461308406446]
s3a://ocfruitsp8/fruits_emr/raspberry/121_100.jpg	raspberry	[90.96080604296855,52.525098751537343,13.29833658985439 4,-5.3692029615629 79,108.40813049040389,-4.329432008525379,5.788148075481941]
s3a://ocfruitsp8/fruits_emr/raspberry/119_100.jpg	raspberry	[89.87769233676881,58.557588381805196,13.195911850530626,1.3439625199923875,28.716901102355287,-3.6201719244778023,1.6434188973039756]
s3a://ocfruitsp8/fruits_emr/raspberry/1_100.jpg	raspberry	[92.02261576360 22,-124.20349047817432,-26.195782732489 06,125.79681026327515,2.5893727816005176,-16.51603926273867,-27.822336296261174]
s3a://ocfruitsp8/fruits_emr/raspberry/118_100.jpg	raspberry	[89.46232279346339,59.124612826911 7,12.846322242075805,1.8540332685348666,25.408641988334693,-1.4483721211813576,-0.5581890463168727]
s3a://ocfruitsp8/fruits_emr/raspberry/116_100.jpg	raspberry	[85.66947792088759,71.29659793219349,16.331522331316666,9.869535411027455,-47.74885588218835,-4.415364908360798,-4.878447870613723]
s3a://ocfruitsp8/fruits_emr/raspberry/115_100.jpg	raspberry	[89.85302850230487,64.12540514611187,14.743419527216 27,6.004771193650225,-50.088865864474984,-1.8368678114175747,-3.5426356233238967]
s3a://ocfruitsp8/fruits_emr/raspberry/0_100.jpg	raspberry	[91.82473268050362,-135.87607391789806,-28.887140401170694,140.79775293509056,8.263178624470745,-14.248359311108098,-13.972032613427 91]
s3a://ocfruitsp8/fruits_emr/raspberry/113_100.jpg	raspberry	[94.39496627737656,62.706447626369 32,16.248763344741 18,9.689047607062863,-83.50905883363063,3.4153482142218 24,0.2796223796069283]
s3a://ocfruitsp8/fruits_emr/raspberry/117_100.jpg	raspberry	[93.83603891463093,68.20368414185 33,15.851379445304014,12.988050620233505,-41.76109825081 49,1.0433442049202695,-4.086516133725723]
s3a://ocfruitsp8/fruits_emr/raspberry/114_100.jpg	raspberry	[91.49650472725526,56.14919511284 32,12.786698949736488,11.043653034859405,-79.90311151688904,4.315050884749294,1.4830588766527961]
s3a://ocfruitsp8/fruits_emr/mango/102_100.jpg	mango	[-84.96566531697565 1,2.272814159916297 3,-17.089824674561445,-1.1585650831654646,2.740643605267987 2,40.115126116135094,-23.985927862849785]
s3a://ocfruitsp8/fruits_emr/mango/101_100.jpg	mango	[-86.81027233746703,1.439923793670355,-14.885221404476592,-1.6384378655488705,4.933057184301963,54.05099393039577,-34.406088490102 68]
s3a://ocfruitsp8/fruits_emr/mango/100_100.jpg	mango	[-84.42813269382806,1.157248425461345 5,-19.014214897945 32,-1.5880389449448869,3.913034500650115 3,55.224290609443834,-33.936997773709381]
s3a://ocfruitsp8/fruits_emr/mango/104_100.jpg	mango	[-88.51685667040643,0.800107805905372,-22.795598304575613,-0.9435849287913335,2.7948811726134 67,45.874274255068929,-28.753039375653 99]
s3a://ocfruitsp8/fruits_emr/mango/113_100.jpg	mango	[-106.13608345547654,4.571158242155822,-19.959946907429373,1.5571904611777862,-4.922569388405416 5,-92.54223414374337,56.740595196654 7]
s3a://ocfruitsp8/fruits_emr/mango/106_100.jpg	mango	[-90.32784156932 44,4.528717972676062,-23.310081972214384,-0.995466870647984 3,0.230784131126921 95,20.247302539103362,-12.621710662469875]
s3a://ocfruitsp8/fruits_emr/mango/103_100.jpg	mango	[-87.88448694338 49,3.643921221320385,-21.015395679753883,-1.942031413978144,2.335021518637214 3,46.025502971383 29,-28.613633638717534]
s3a://ocfruitsp8/fruits_emr/mango/105_100.jpg	mango	[-84.71707630843981,3.928576081461960 4,-25.780870553706554,-1.2827643791786 8,0.568203422401708,26.422292918962 31,-16.853432610530895]
s3a://ocfruitsp8/fruits_emr/mango/110_100.jpg	mango	[-98.79319580623418,7.127576697238392,-27.202469978859313,-0.4359849188714934,-4.183031176375559,-45.210175394734 61,27.744748165025565]
s3a://ocfruitsp8/fruits_emr/mango/108_100.jpg	mango	[-98.04789699302512,4.434198398756520 6,-22.469441414664 03,-0.029165096088471264,-3.144259152320103 6,-18.587634404403612,9.878035183035763]
s3a://ocfruitsp8/fruits_emr/mango/111_100.jpg	mango	[-99.43484973255185,5.947361462154818,-25.960975578679 91,-1.6376531368119718,-4.571452805642 42,-50.78055633778399,31.17499803049984]
s3a://ocfruitsp8/fruits_emr/mango/107_100.jpg	mango	[-92.16145902164794,5.203360889919983,-26.893581441622 4,-0.3047497459274513 5,-0.5414448799813 3,8.242227346480348,-6.602841518371755]
s3a://ocfruitsp8/fruits_emr/mango/109_100.jpg	mango	[-93.12754820813 08,6.793301689730298,-28.137858410645983,-0.9598440283894252,-1.338765994428201 8,-11.411796471910387,6.103476235654138]
s3a://ocfruitsp8/fruits_emr/mango/112_100.jpg	mango	[-100.04650559370708,5.296415435250045,-22.081694375610 63,-0.7297396493358013,-3.612303070831263,-56.388700025583 58,34.85439236569904]
s3a://ocfruitsp8/fruits_emr/mango/113_100.jpg	mango	[-98.79774638833773,-73.335206662626 9,303.5924732303218 4,-1.5460411505792655,-0.934111563190294 8,-0.6824949224529582,2.477161252798119]

5 Conclusion

La mission

Les architectures utilisées

- EC2 single (mode Local spark)
- EC2 cluster (mode standalone spark)
- EMR (préconfiguré Yarn)

Preprocessing et PCA

- Les images sont séparées lors de la projection en 2D

Amélioration possible

- Mise à l'échelle autoscaling EMR
- Preprocessing images
- Tester d'autres modèles features extractions et hyperparamètres
- Entraîner un modèle de classification avec les features extractions
- Approfondir connaissance Spark et service AWS