# Deploying a model in the Cloud

AUGUST 2022



NAJWA MOULINE

# Presentation Outline

1. Objectives
2. Dataset Overview
3. Image processing
4. Conclusion and recommendations

# Objectives

## Context

- A young AgriTech start-up offering innovative solutions for fruit harvesting
- Innovative solution considered: development of intelligent picking robots.

## Business Problem

- First step to be known: make available to the general public a mobile application for fruit recognition from a photo
- Implementation of a first version:
  - Fruit image classification engine
  - Scalable big data architecture

## Mission

- Set up a Big Data environment
- Develop a first image processing chain including:
  - Pre-processing
  - Dimension reduction
  - Modeling

# Dataset Overview

## Dataset overview

### Provided dataset

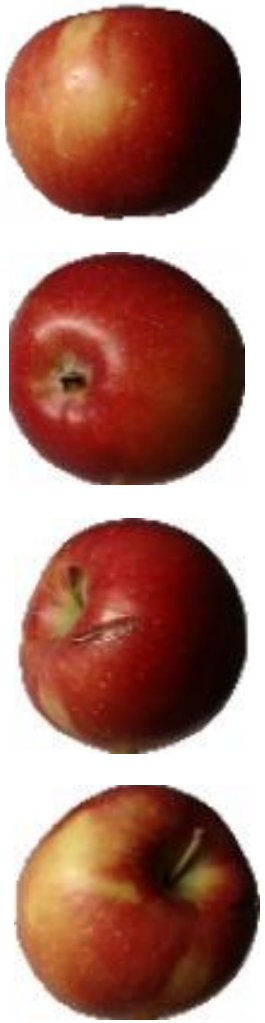
- Dataset from Kaggle: more than 80,000 images of fruits/vegetables classified into 131 categories
- Colour photos representing the fruit/vegetable 360° (100x100 pixels) on a white background, in jpeg format
- Several categories of the same fruit (apple)

### Considered dataset

- 928 photos / 4 fruit categories
- 2 categories of related fruits (apple)

```
+---------------+-----+
|          label|count|
+---------------+-----+
| Apple_Braeburn|  246|
|Apple_Pink_Lady|  228|
|            Fig|  234|
|           Kiwi|  220|
+---------------+-----+
```

### Photos' example

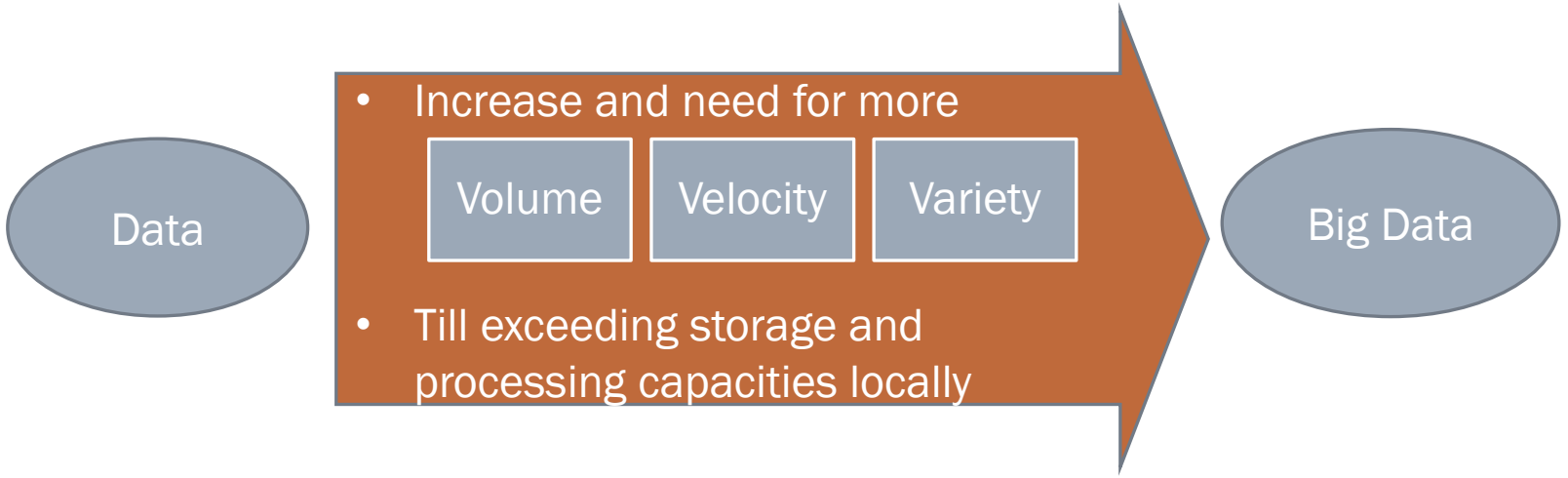


# Image processing

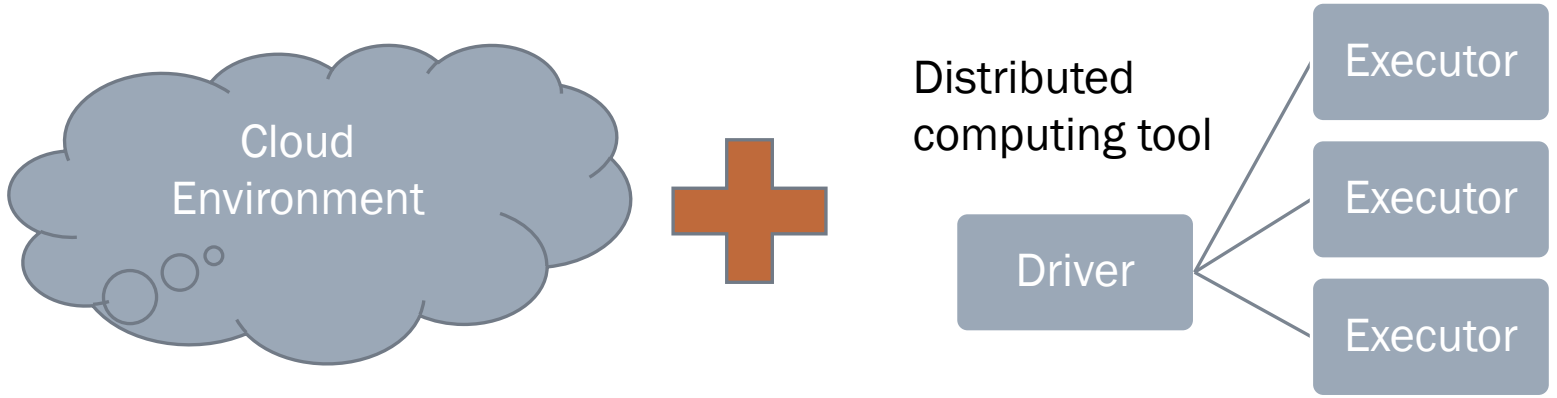
## Architecture Big Data

### The importance of Big Data

#### Big Data




#### Solution: a scalable, resilient and distributed infrastructure



# Image processing

## Architecture Big Data

### The blocks of the architecture

#### Solution: a scalable, resilient and distributed infrastructure


#### Cloud Environment - AWS

Amazon S3

- S3 (Simple Storage Service)
- Low-cost, unlimited, distributed, and resilient storage service
- Data stored in buckets

Amazon EC2

- Secure and resizable compute web service
- Service for managing servers as virtual machines in the cloud
- Configuration of the operating system, processor, storage, ...

AWS IAM

- Access Management Identification Service
- Definition of uses, groups and roles

# Image processing

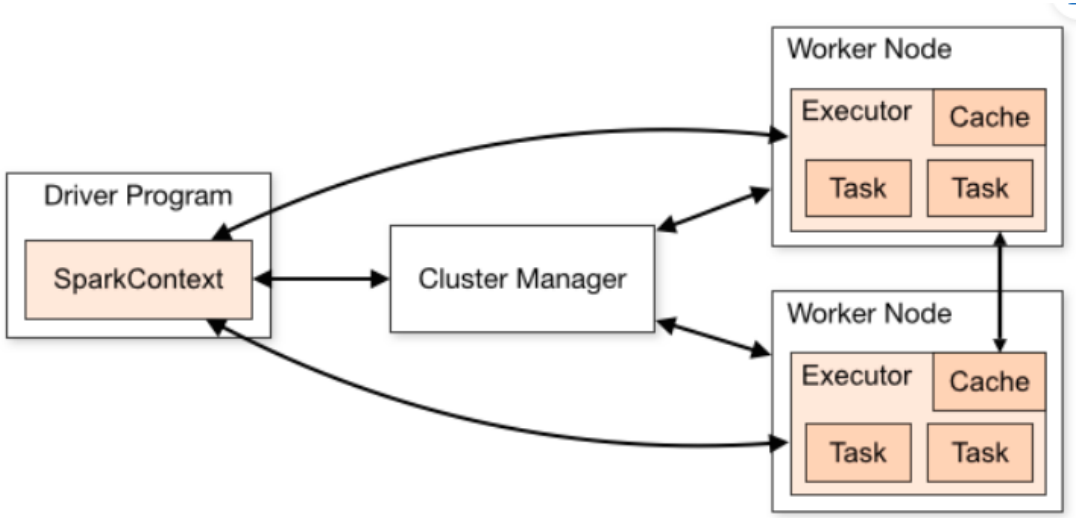
## Architecture Big Data

### The blocks of the architecture

#### Solution: a scalable, resilient and distributed infrastructure


#### Distributed Computing Tool – Spark

Spark or Apache Spark is an open source distributed computing framework




Graph from databricks

# Image processing

## Architecture Big Data

### Setting up the architecture

#### EC2 instance configuration via ssh

- Server update/upgrade
- Python/pip installation
- Virtual environment creation
- Installing the Jupyter notebook
- Spark installation (Java, Scala, Spark 3.3.0 with Hadoop, findspark)
- Installation of libraries (boto3, pandas, …)

#### Service S3 configuration through AWS console

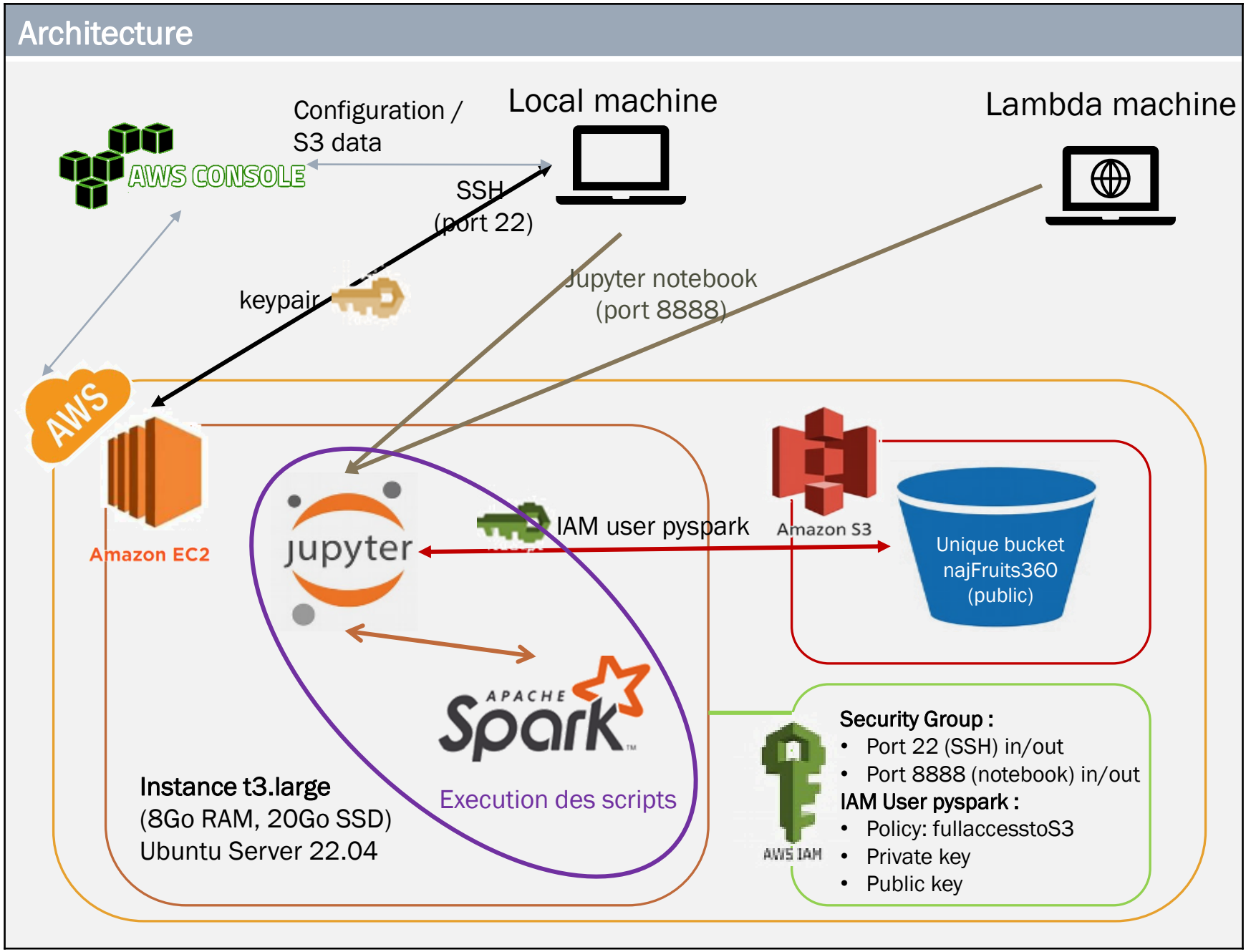
- Creating an S3 bucket
- Loading data into the bucket with interface

#### Configuration IAM service through AWS console

- Definition of Security group – access control to the instance
- Setting Users/Groups - S3 Bucket Access Control

# Image processing

## Architecture Big Data



# Image processing

## Processing chain

### Methodology

#### Process

- Creating a Spark session
- Loading the Data
- Pre-processing / Feature Extraction
- Reducing dimensions
- Modeling

#### Specific Big Data Libraries

PySpark, Boto 3, Spark SQL, Apache Spark ML

# Image processing

## Processing chain

### Loading the Data

#### Processus

Creating a Spark session

Retrieving image paths

Creating a Spark DataFrame

Extracting fruit categories from the path

```
+---------------+-----+
|          label|count|
+---------------+-----+
| Apple_Braeburn|  246|
|Apple_Pink_Lady|  228|
|            Fig|  234|
|           Kiwi|  220|
+---------------+-----+
```

```
+--------------------+--------------+
|            img_path|         label|
+--------------------+--------------+
|Apple_Braeburn/23...|Apple_Braeburn|
|Apple_Braeburn/23...|Apple_Braeburn|
|Apple_Braeburn/23...|Apple_Braeburn|
|Apple_Braeburn/23...|Apple_Braeburn|
|Apple_Braeburn/23...|Apple_Braeburn|
|Apple_Braeburn/23...|Apple_Braeburn|
```

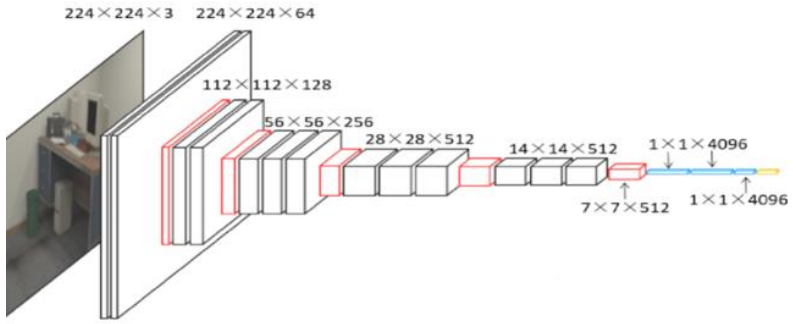
# Image processing

## Processing chain

### Pre-processing / Feature extraction by VGG16 transfer learning

#### Process

- Loading the model – VGG16 without last layer
- Preprocessing images
- Predicting features
- Creating a Spark DataFrame with vectorised features and associated labels




```
+--------------+--------------------+
|         label|        features_vec|
+--------------+--------------------+
|Apple_Braeburn|[0.0,0.0,0.0,0.90...|
|Apple_Braeburn|[0.0,0.0,0.0,0.65...|
|Apple_Braeburn|[0.0,0.0,0.0,1.26...|
|Apple_Braeburn|[0.0,0.0,0.0,0.93...|
|Apple_Braeburn|[0.0,0.0,0.0,0.64...|
|Apple_Braeburn|[0.0,0.0,0.0,0.39...|
```

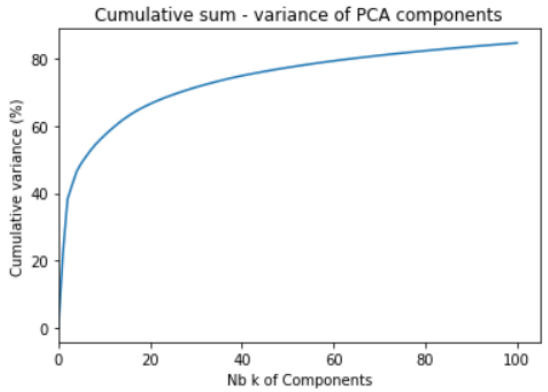
# Image processing

## Processing chain

### Reducing dimensions with PCA

#### Process

- Indexing categories
- Applying Standard Scaler
- Applying PCA
- Converting to Pandas DataFrame
- Saving the csv format file in S3




```
+--------+--------------------+
|category|        features_pca|
+--------+--------------------+
|     0.0|[3.51591850674599...|
|     0.0|[0.68652618962347...|
|     0.0|[1.04571548325630...|
|     0.0|[2.37953706247352...|
|     0.0|[2.28412267169469...|
```

# Image processing

## Processing chain

## Modeling with Kmeans

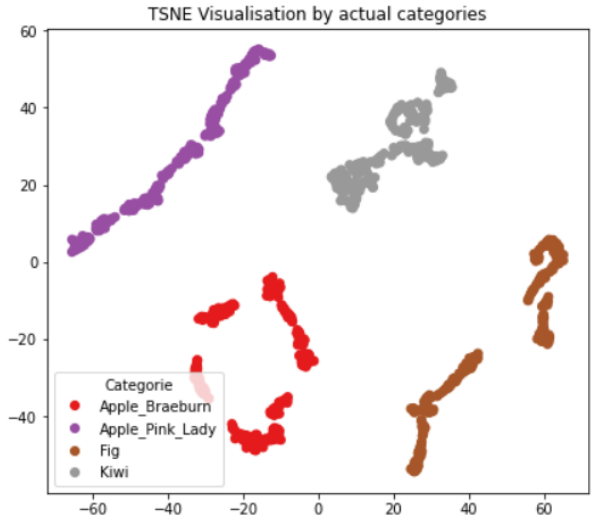
### Process

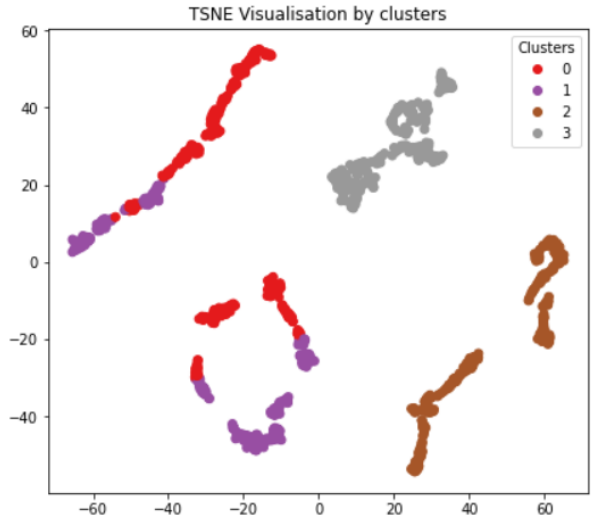
Modeling with KMeans

Visualising with TSNE – Real categories

- Transformation features vector to string
- Separation of PCA features (10 dimensions)

Visualising with TSNE – Clusters



# Conclusion

## Conclusion

- Setting up a big data environment with AWS (EC2, S3) and Spark
- Scaling up may require the review of EC2 instance suitability (processing, memory)

## Recommendations

- At Big Data level:
  - Choosing a more powerful EC2 instance to model the entire dataset
- At data processing level:
  - Input images more ‘real’ (with non white standard rear)
  - Improved pre-processing (other image pre-processing techniques and transfer learning models)

*Thank you for your attention!*