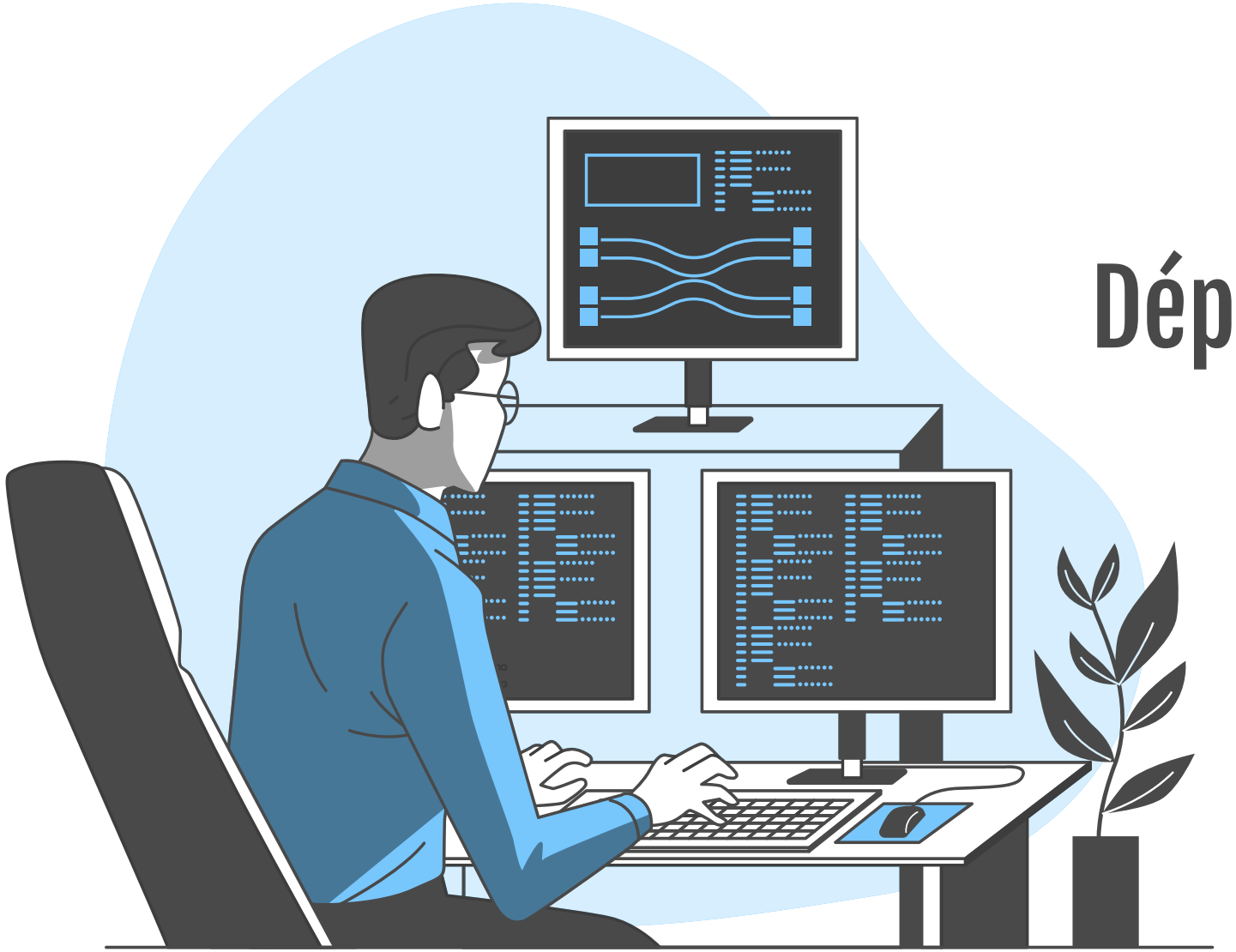


Déployez un modèle dans le cloud

OpenClassrooms – projet 8

Léo Guillaume

Roadmap

01

Introduction

Contexte, problèmes et enjeux

02

Le Big Data, quésaco ?

Éclairage sur l'éco-système du Big Data

03

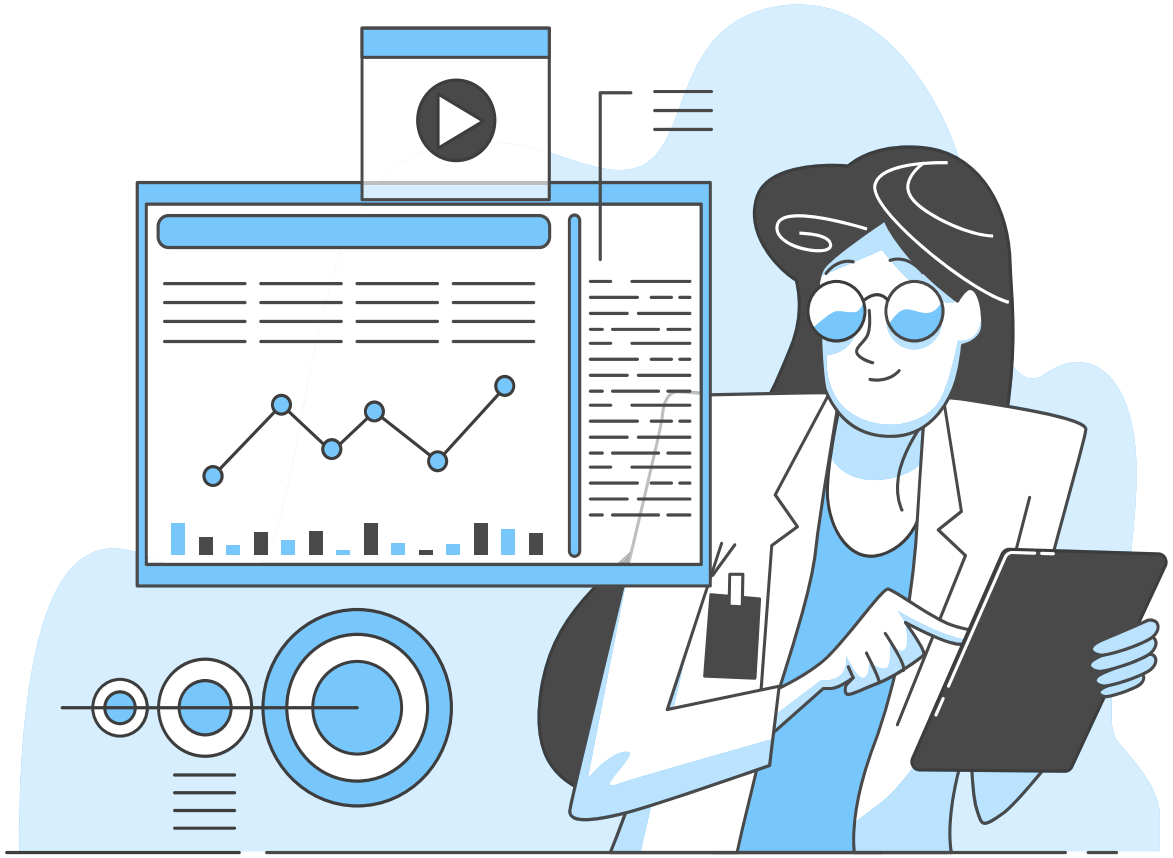
Modélisation distribuée

Données, architecture et modélisation

04

Conclusion

Résultats et perspectives



Introduction

Contexte

- La start-up **“Fruits!”** propose des solutions innovantes pour la récolte des fruits
- Souhaite un moteur de classification d’images de fruits

Enjeux

- **Volumétrie du dataset** impose d’être traité sur plusieurs machine à la fois
- L’application doit **permettre le passage à l’échelle** car sera accessible au grand publique

Problématique

- Comment développer un modèle de la classification d’image de manière distribuée dans le cloud ?

Le Big Data, quésaco ?

- **Qu'est ce que le Big Data ?**

On parle de Big Data dès lors que la quantité de données excède la faculté d'une machine à les stocker et les analyser en un temps acceptable.

- **Les 3 V du Big data :**

V

Volume

Le volume des données générées nécessite de repenser la manière dont elles sont stockées

V

Vélocité

La vélocité à laquelle nous parviennent ces données implique de mettre en place des solutions de traitement en temps réel qui ne paralysent pas le reste de l'application

V

Variété

Les données se présentent sous une grande variété de formats

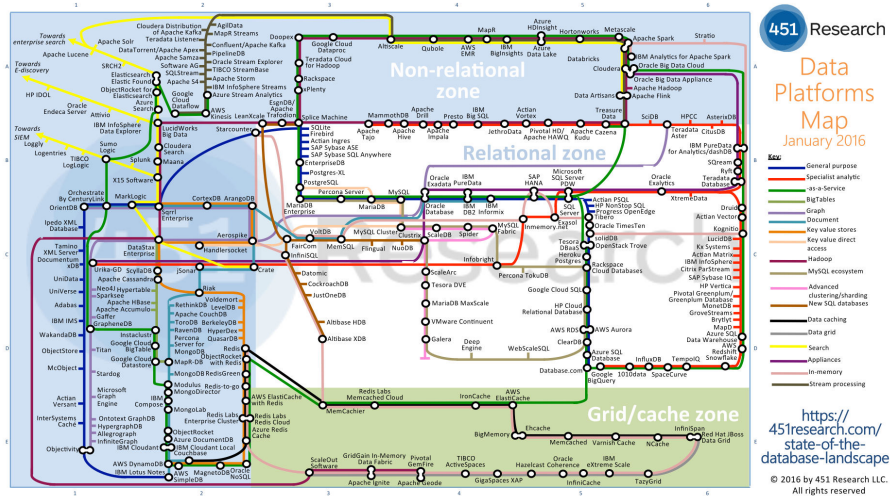
Cluster de machines

- 2 solutions : passage à l'échelle horizontale ou **verticale**

- **Qu'est ce qu'un cluster ?**
  Un cluster ensemble de machine (**noeuds**) réalisant chacun une partie d'un calcul distribué et communiquant entre eux.

- Le distribution des calculs dans un cluster implique de répondre à plusieurs enjeux :
  1. **Le répartition des ressources (scalability)**
  2. **Le stockage des données (data locality)**
  3. **La tolérance aux pannes (embracing failure)**

- Pour répondre à ces enjeux de nombreuses solutions ont été dévéloppées qui composent l'éco-système Big Data :



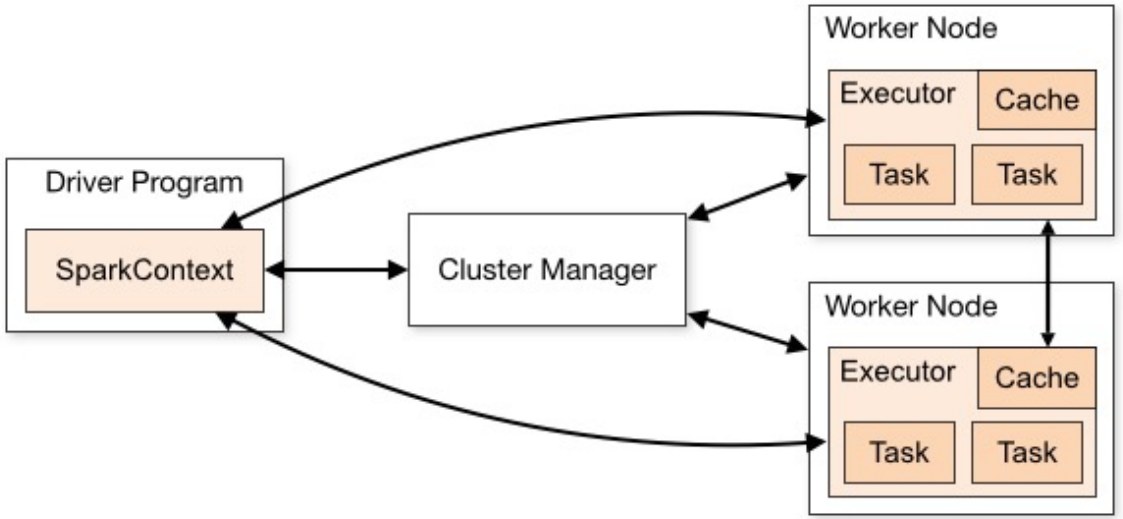
Répartitions des ressources

Concepts clef :

1. Distribuer les opérations dans un cluster : **architecteur maître esclave**
2. Diviser les opérations pour les exécuter sur plusieurs machines : **Map Reduce**

➢ Le framework **Spark** répond à ces impératifs, il présente 2 avantages :

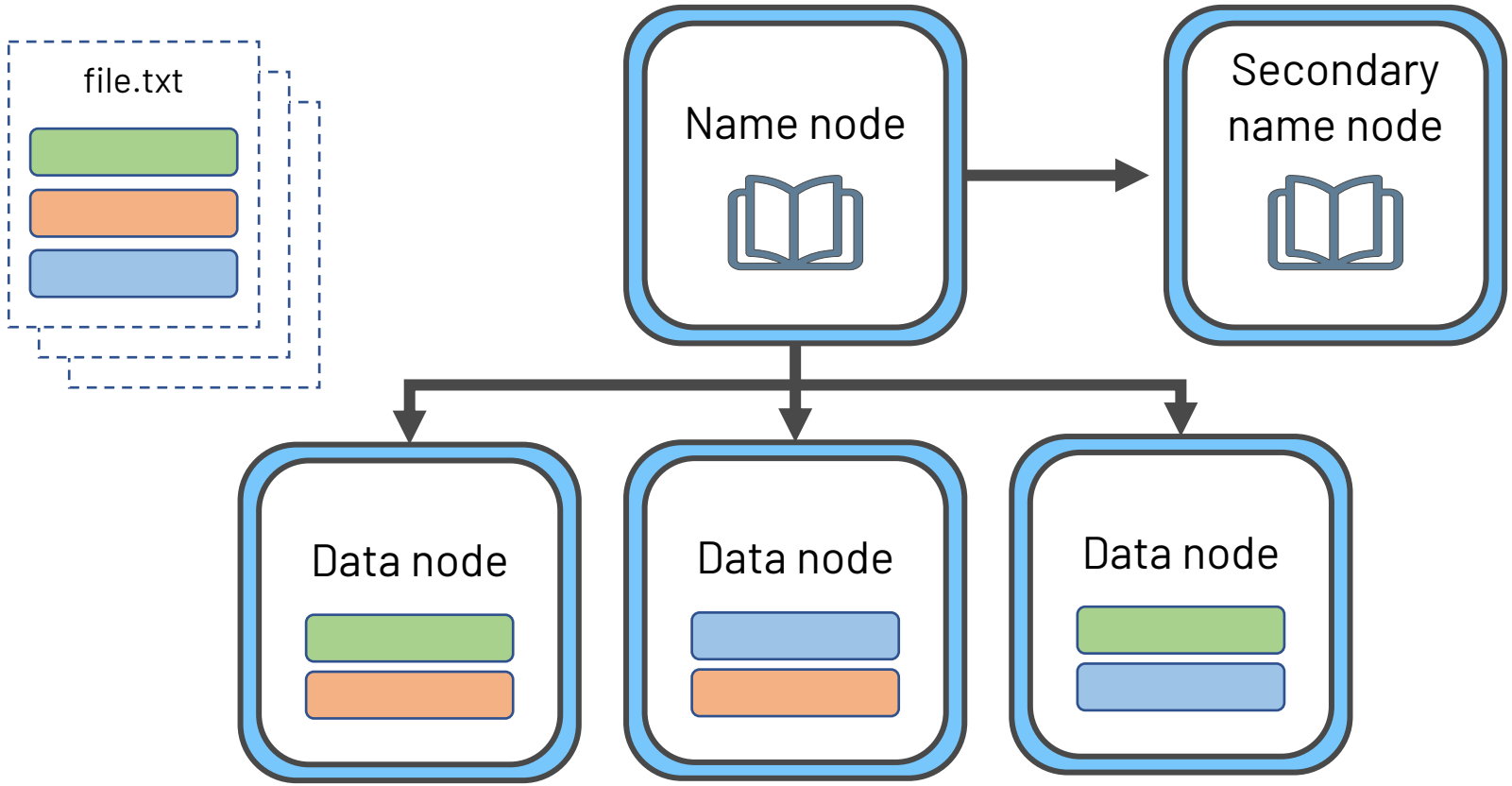
- Ecrit les données en RAM et non en disque lors de leur traitement
- Permet une abstraction des opérations de Map et Reduce



Stockage des données

Le stockage de fichiers dans un cluster nécessite **un système de fichier distribué**

- Par exemple, HDFS est un framework permettant l'écriture de manière distribuée



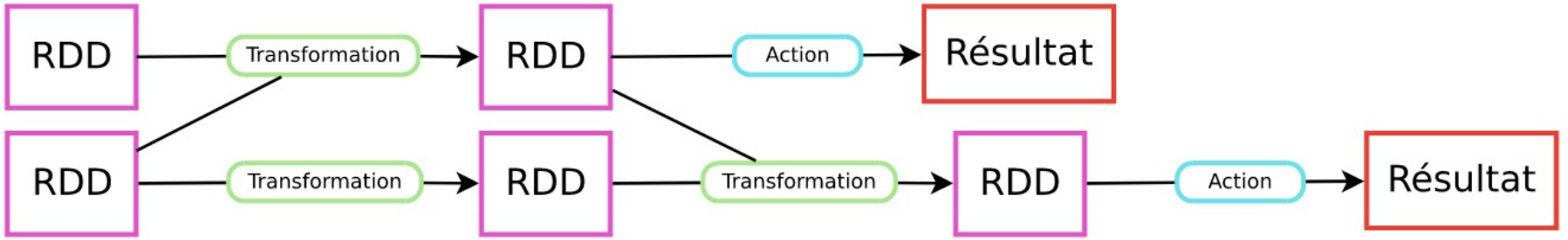
La tolérance aux pannes

Habituellement, le résultat de chaque étape de calcul est écrit en disque.

➢ Problème : avec Spark il reste en RAM

Concepts clef :

- Resilient Distributed Datasets (**RDD**)
- Directed Acyclic Graph (**DAG**)



Les données

3

Jeux de données

7

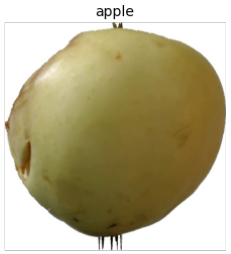
Catégories de fruits

Nombre d'images par jeu

6231 Jeu d'entrainement (50 %)

3114 Jeu de validation (25 %)

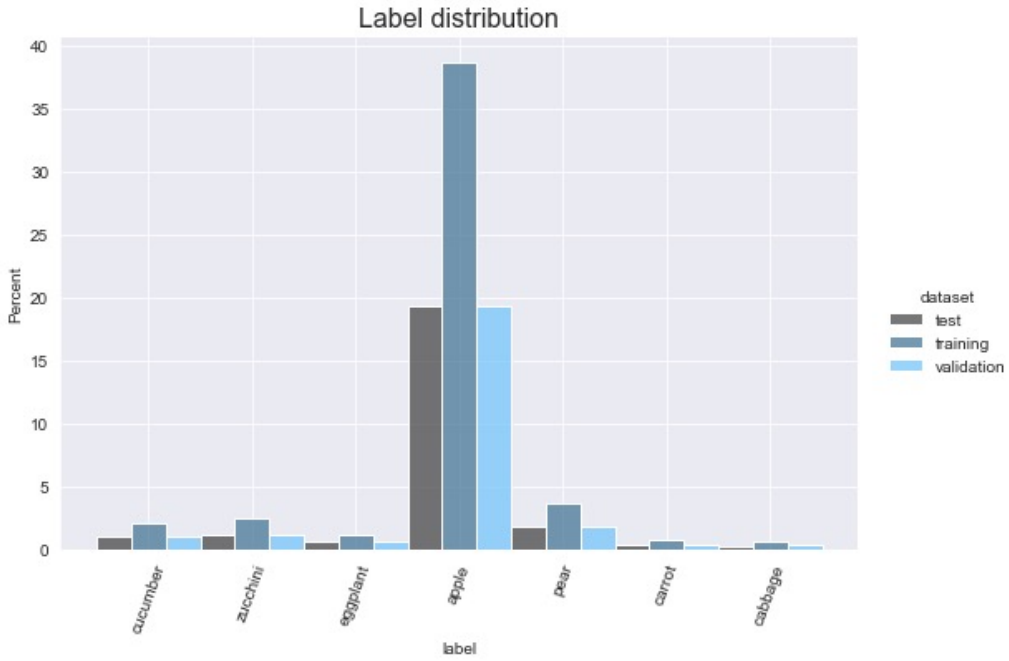
3110 Jeu de test (25 %)

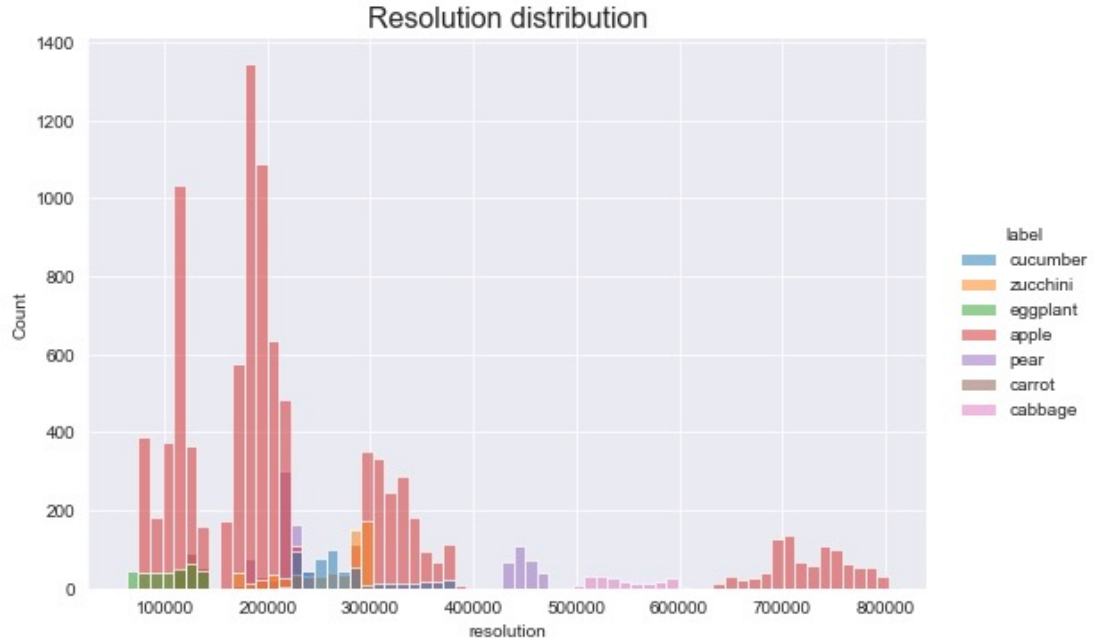


Les données

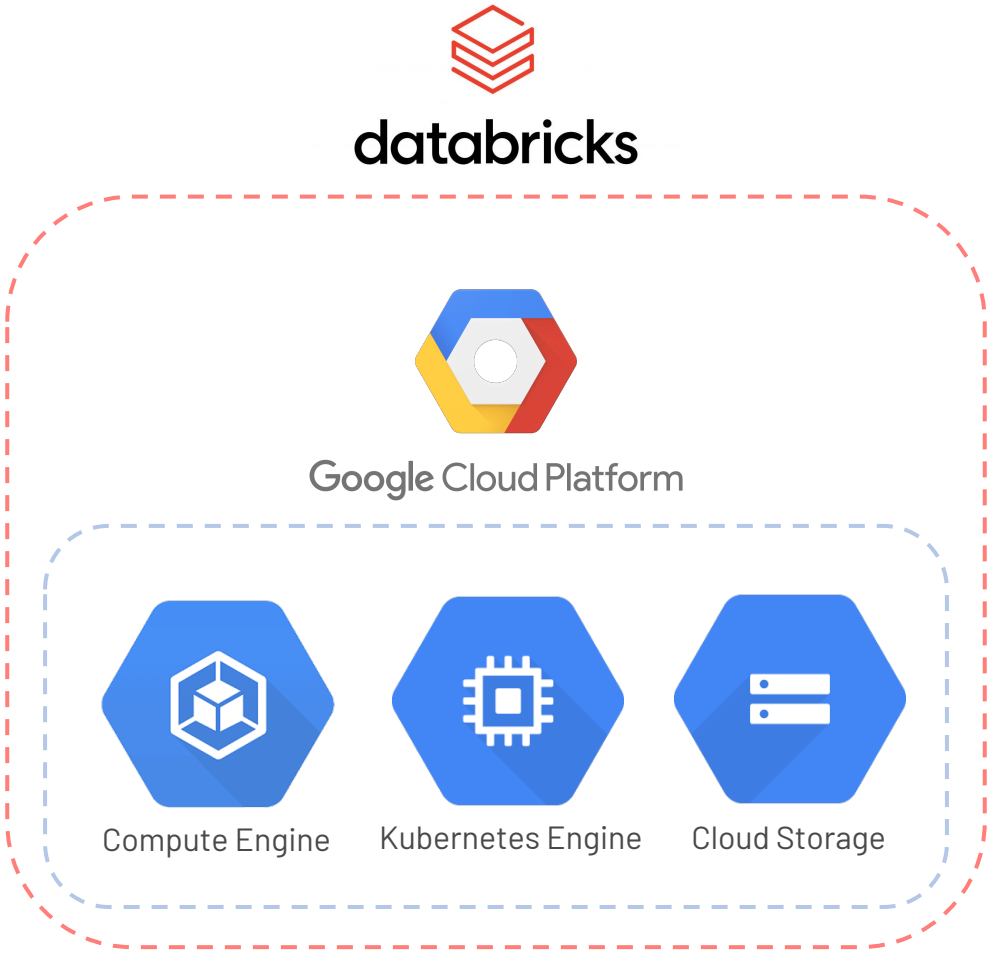
2 caractéristiques à retenir :

- **Déséquilibre des classes à prédire** : le catégorie pomme est surreprésenté dans les jeux de données
- **Résolutions hétérogènes intra et inter catégories** : les images ont des tailles différentes





Solution cloud



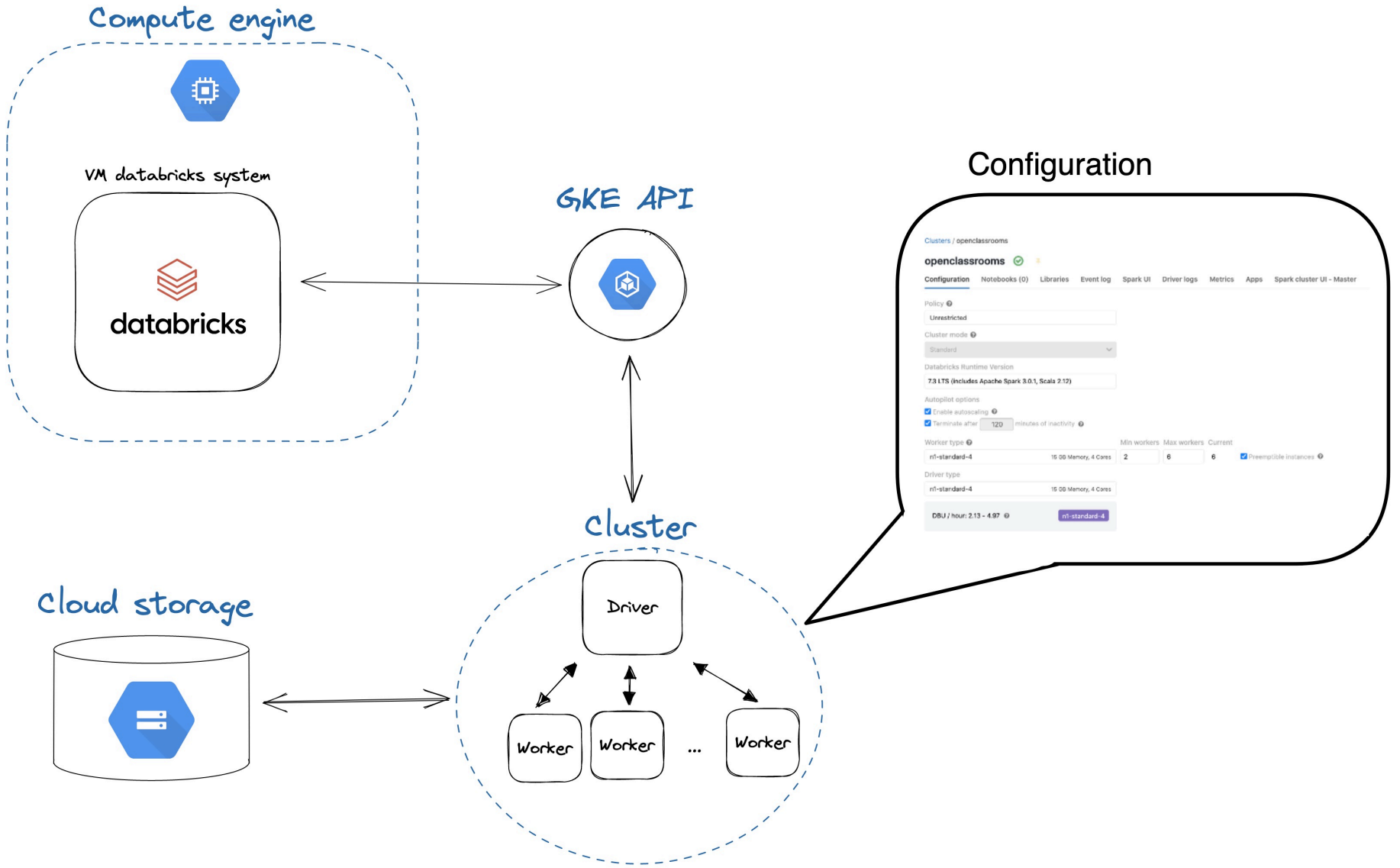
Solutions retenus :

- Utilisation de **Databricks** pour coder la pipeline de traitement des données et les différents services cloud
- Utilisation de **Google Cloud Platorm (GCP)** comme fournisseur de service cloud (machines virtuelles et stockage)

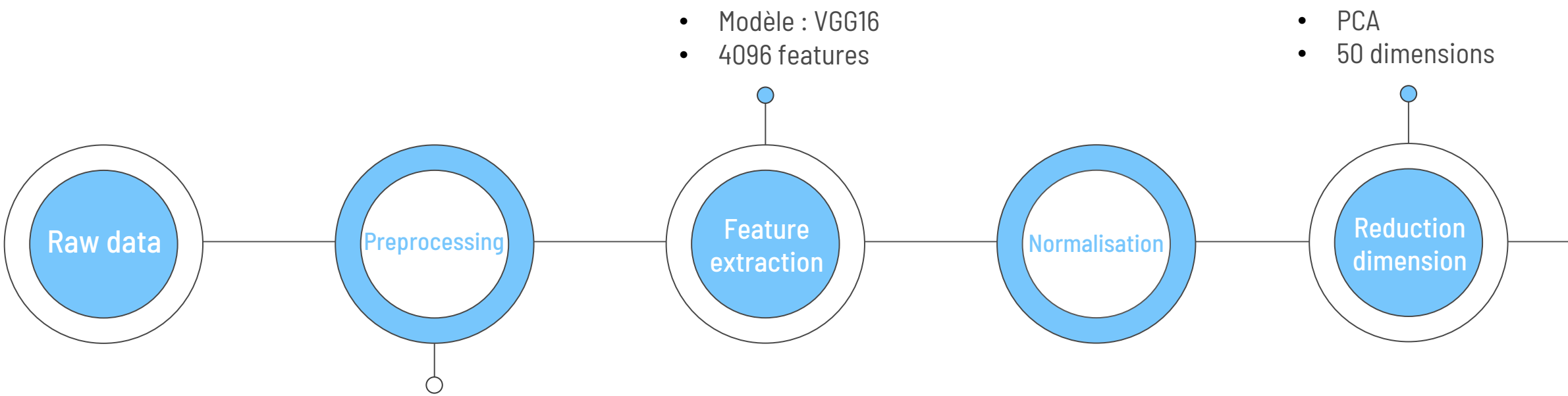
Motivations :

- Databricks permet une abstraction du déploiement
- Utilisation de GCP chez Dotaki
- 300 $ de crédit GCP offert à l'ouverture d'un compte

Architecture



Data pipeline



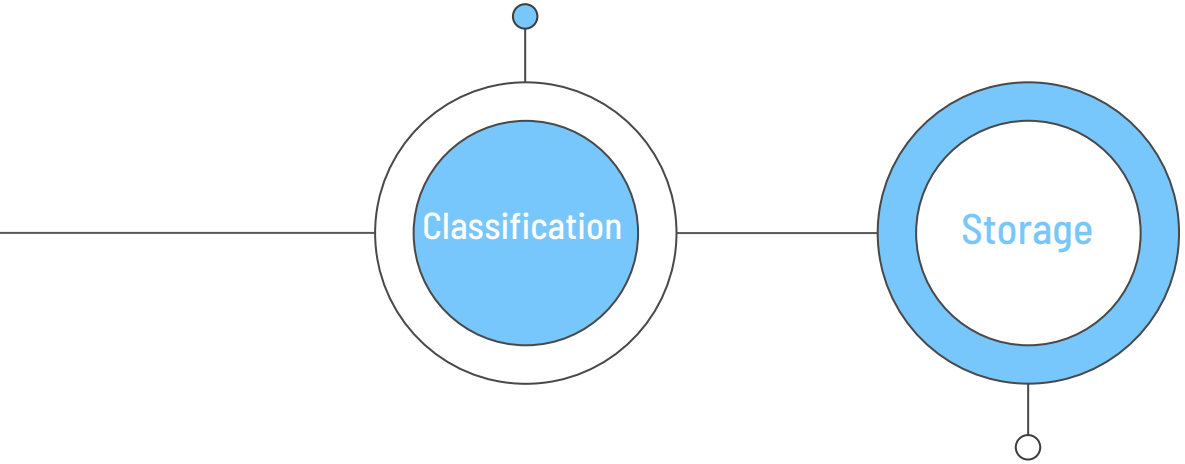
- Conversion des images en matrice
- Redimensionnement (244 x 244)

- Modèle : VGG16
- 4096 features

- PCA
- 50 dimensions

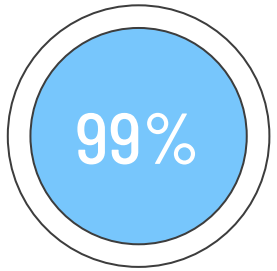
Data pipeline

- Modèle : random forest



- Ecriture des données transformées et des prédictions

Résultats



Accuracy

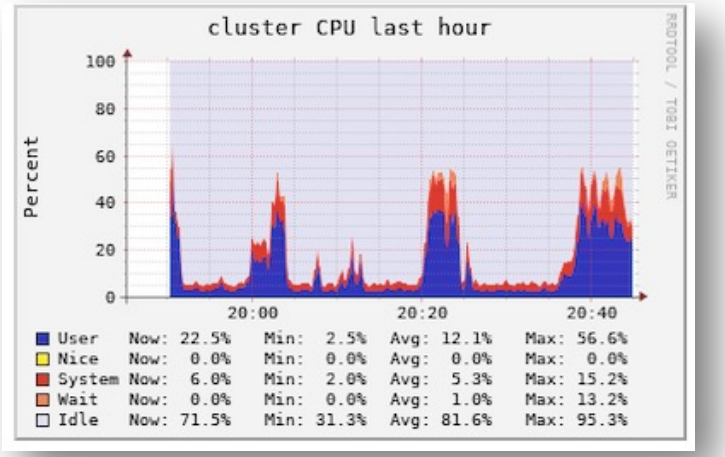
Exemple de features extraites :

	label	features
5977	apple	[0.0, 0.12709525, 0.0, 2.9404461, 0.0, 0.0, 4....
5083	cabbage	[0.0, 0.0, 0.0, 0.1813078, 0.0, 0.0, 0.4224908...
4982	eggplant	[0.0, 0.0, 1.6811758, 0.0, 0.0, 0.98431325, 0....
5675	apple	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.1742842, 0.0,...
818	apple	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...

Coûts du projet :

Projet	Identifiant du projet	Numéro du projet	Coût	Remises	Promotions et autre
sandbox	feisty-proton-348611	391676905173	246,66 €	—	-
Use of databricks endpoints da	pr-1a51326ffd3dba7f	393020353599	35,22 €	—	-

Exemple de log d'activité du cluster :



Perspectives

01

Les données

- Pre-processing des images prises par le public (pas de fond blanc)

02

La modélisation

- Mettre en place un pipeline d'entrainement de modèles

03

Le déploiement

- Tester d'autres d'autre fournisseur Cloud tel que AWS ou Microsoft Azure

Merci pour votre attention