Projet 8:
Déployez un modèle dans le cloud

Victor BARBIER

Table of contents

Contexte

Contexte

Mission

Société “**Fruits!”**

Startup AgriTech

Création d’une application qui reconnaît les fruits

Objectif

Moteur de classification

Réalisation d’une chaîne de traitement des données

Déploiement sur une architecture Big Data

Données

Jeu de données Kaggle

Images de fruits numérisés

Jeu de données

Deux datasets Kaggle :
- Training
- Test

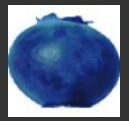
→ Utilisation de **Test**

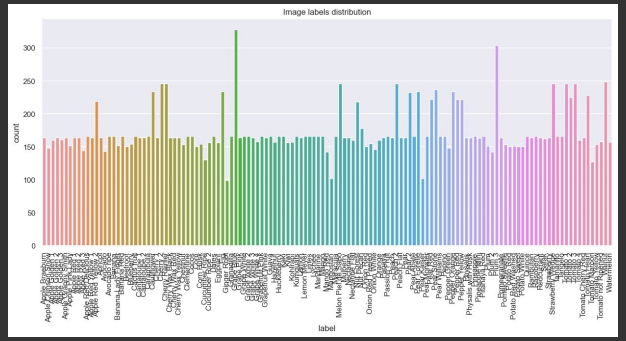
Images de fruits :
- Couleur
- 100 x 100 px
- Format *.jpg*

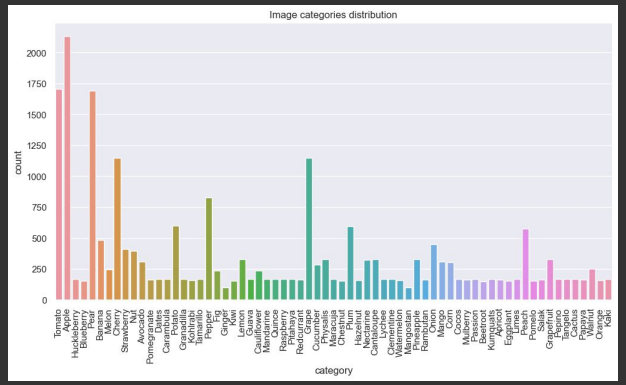
22688 **images**

131 **variétés**

67 **catégories**



Big Data

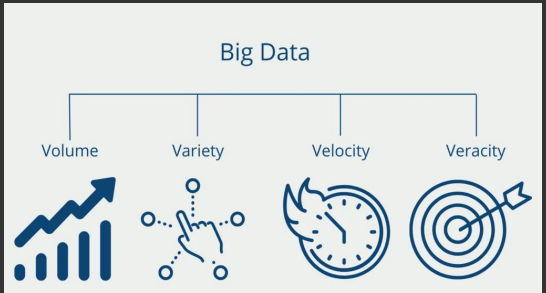
présentation

Le Big Data ?

Explosion récente de la production de données

Émergence de nouvelles technologies et méthodes de capture, stockage, recherche, partage, analyse et présentation des données

Espace virtuel : **Cloud computing**



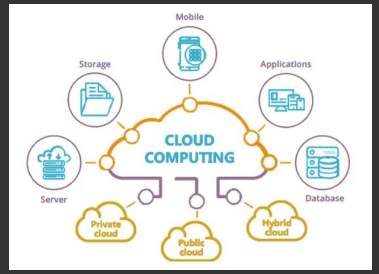
Problèmatiques

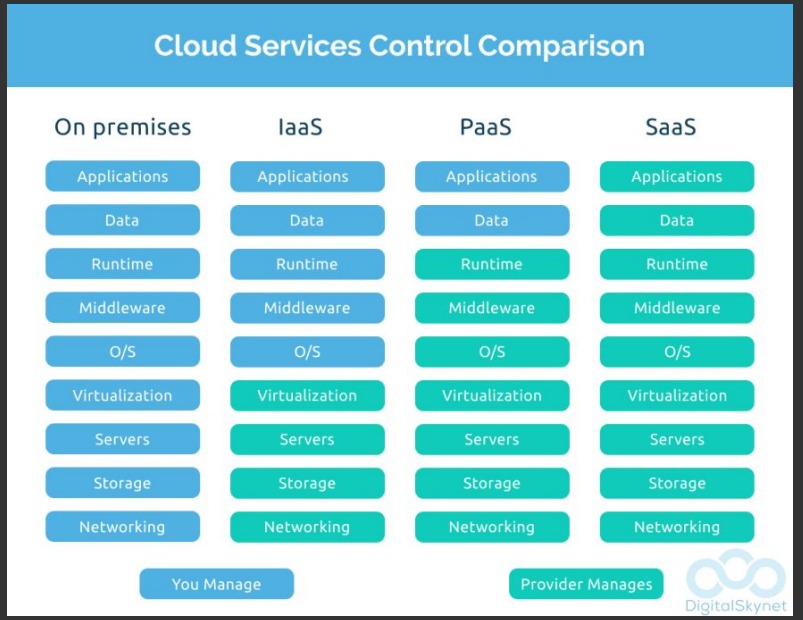
Volume	Données récoltées de plus en plus massives
Variété	Données de forme et type très variées
Vitesse	Lecture de données en continue
Véracité	Besoin de fiabilité sur les données

Le Cloud Computing

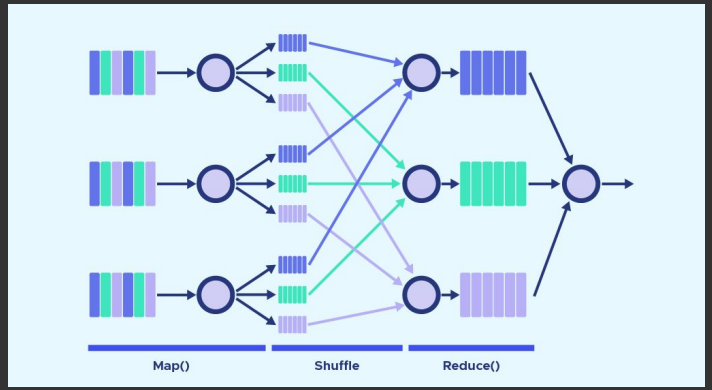
Fourniture de **serveurs informatiques** :

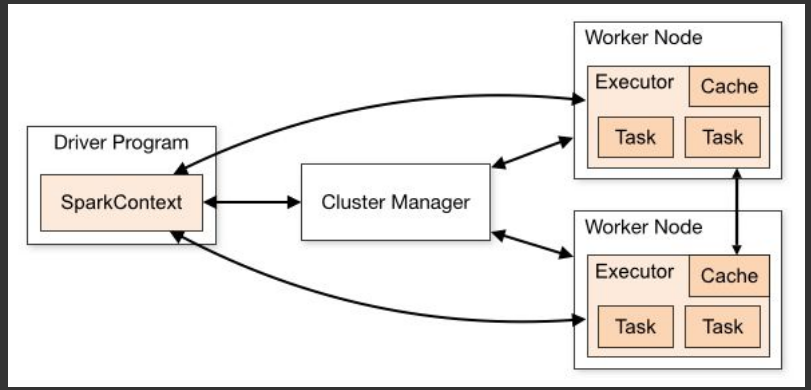
- **IAAS :** *Infrastructure as a Service*
- **PAAS :** *Platform as a Service*
- **SAAS :** *Software as a Service*





Le calcul distribué





hadoop MapReduce

APACHE Spark

Comparatif des plateformes

Service	Description	Google Cloud Platform	Amazon Web Services	Microsoft Azure
Stockage	Hébergements de fichiers dans le cloud	**Cloud Storage**	**Simple Storage Service (S3)**	**Blob Storage**
Calcul	Création de machines virtuelles	**Compute Engine**	**Elastic Compute Cloud (EC2)**	**Virtual Machines**
Traitement des données	Création de clusters pour l'exécution d'architecture Hadoop/Spark	**Dataproc**	**Elastic MapReduce (EMR)**	**Data Lake Analytics**
IAM	Identité et accès aux ressources	**Cloud Identity**	**IAM Identity Center**	**Active Directory**

Big Data

Solution retenue

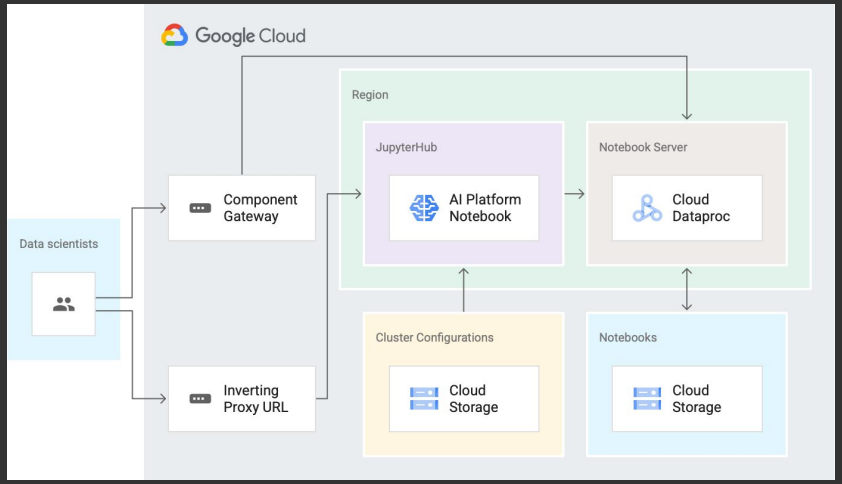
Google Cloud Platform

Stockage dans des **Buckets**

Calcul dans un **Cluster** (Dataproc)

Traitement des données via **notebook Jupyter** ou **script Pyspark**

300 euros de crédit offerts



Google Cloud storage

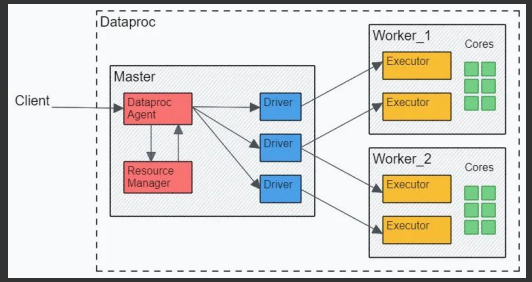
Stockage dans un **Bucket**

- Classe *Standard* ($0.023 GB/mois)
- Région *Europe-West9* (Paris)
- Accès *Fine-grained* / Non public
- Cryptage des données *Google-managed encryption key*

Google Dataproc

Calcul dans un **Cluster (VM)**

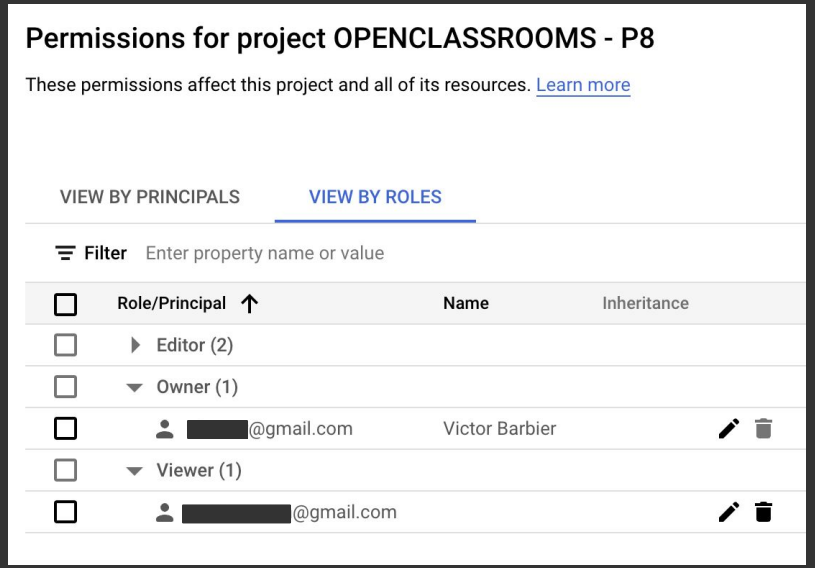
- Configuration *Standard* (*1 master - N worker*)
- Région *Europe-West9* (Paris)
- Master node : *N2 - 2 vCPU - 8 GB - 100 GB*
- Worker node : *N2 - 2 vCPU - 8 GB - 100 GB*
- VM image : *2.0.58-debian10*
- Composants additionnels : *Jupyter - Tensorflow*
- Stockage sur bucket précédent
- Autoscaling policy personnalisée



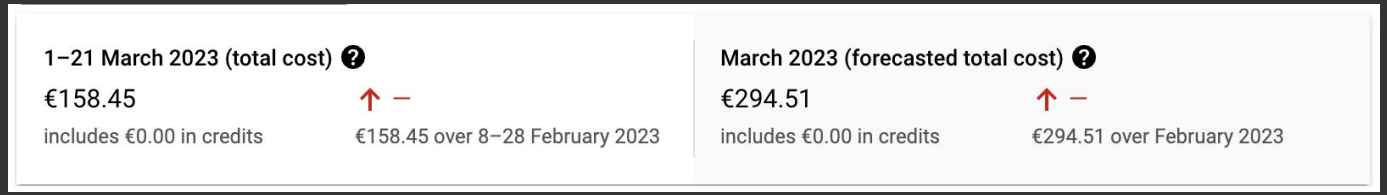
Google IAM

Contrôle des accès

- Milliers de rôles différents
- Accès au projet limité (ex **Bucket**)
- Google Cloud respecte le RGPD (General Data Protection Regulation)

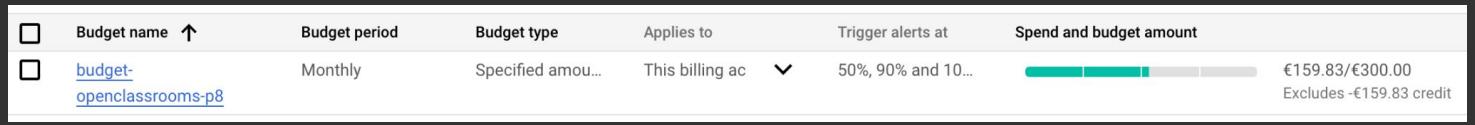


Google Billing



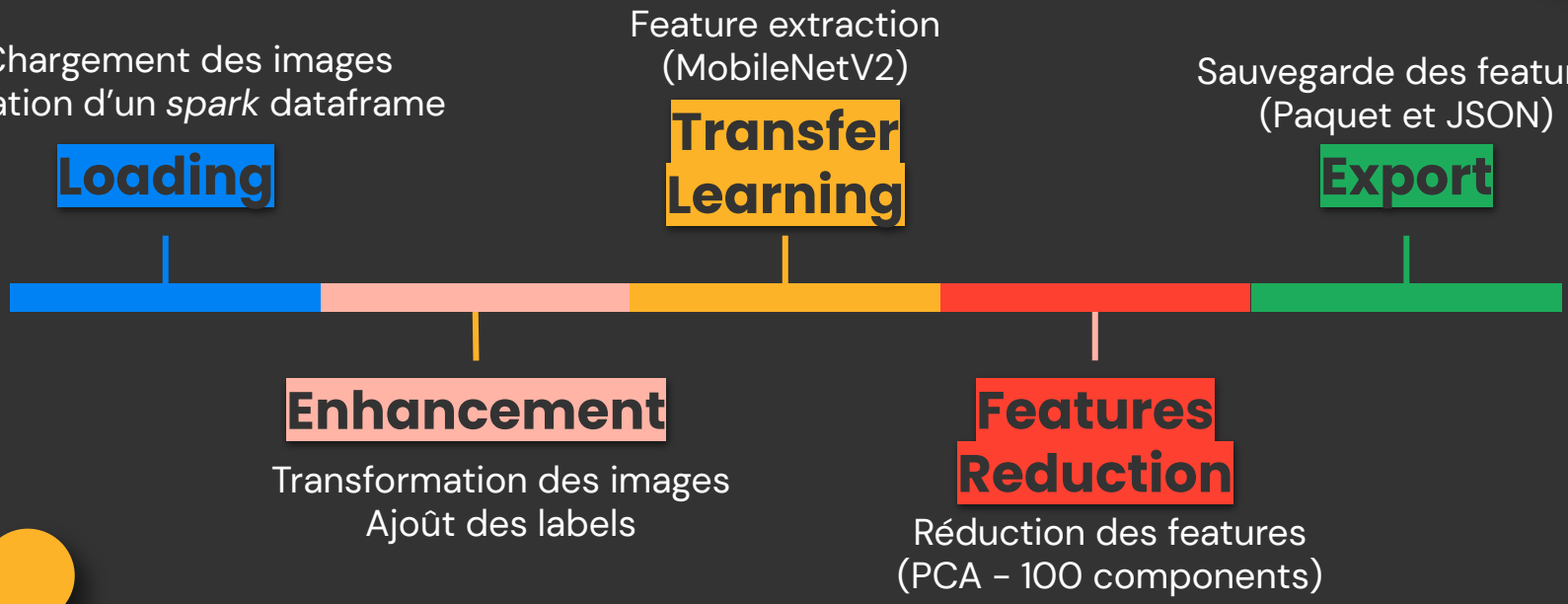
Contrôle de la tarification

- Report des coûts par service
- Création de budgets et alertes



Chaîne de traitement

Chaîne de traitement



Feature extraction

MobileNET V2

Modèle **Tensorflow**

Transfer Learning

53 couches

Dernière couche retirée (classification)

Images en **224x224**

Input	Operator	t	c	n	s
$224^2 \times 3$	conv2d	-	32	1	2
$112^2 \times 32$	bottleneck	1	16	1	1
$112^2 \times 16$	bottleneck	6	24	2	2
$56^2 \times 24$	bottleneck	6	32	3	2
$28^2 \times 32$	bottleneck	6	64	4	2
$14^2 \times 64$	bottleneck	6	96	3	1
$14^2 \times 96$	bottleneck	6	160	3	2
$7^2 \times 160$	bottleneck	6	320	1	1
$7^2 \times 320$	conv2d 1x1	-	1280	1	1
$7^2 \times 1280$	avgpool 7x7	-	-	1	-
$1 \times 1 \times 1280$	conv2d 1x1	-	k	-	

1280 **features**

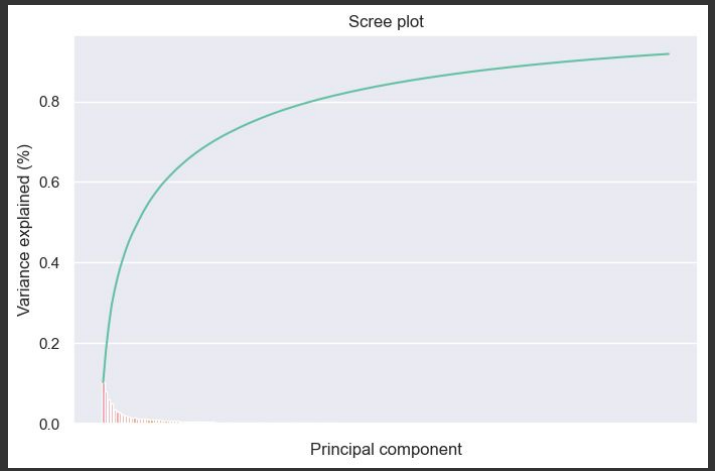
Feature reduction

PCA

Principal Component Analysis

90% **variance expliquée**

166 **features**



Exécution

Notebook

Upload dans le **Bucket**

Exécution manuelle sur le **Cluster**

Kernel **Pyspark**

Script Pyspark

Upload dans le **Bucket**

Exécution automatique sur le **Cluster**

Job **Pyspark**

Démonstration du script PySpark

Conclusion

Conclusion

Réalisation

Environnement Big Data (Google Cloud)

Preprocessing et réduction de dimension

Exécution distribuée (Pyspark)

Améliorations

Intégration CI/CD (stockage et jobs)

Clusters autozone

Bucket *lifecycle* conditions