

# Déployez un modèle dans le cloud



**OPENCLASSROOMS**



**Fruits!**

# Sommaire

- **Problématique**
- **Présentation des données**
- **Big Data?**
- **Traitement des images**
- **Conclusions**

# Introduction

## Fruits !

Souhaite proposer des solutions innovantes pour la récolte des fruits.  
Développer des robots cueilleurs intelligent à l'aide d'une application qui permettra aux utilisateurs de prendre en photo un fruit et d'obtenir des informations sur ce fruit.

## Mission

Développer un environnement Big Data  
Réaliser une première chaîne de traitement des données avec le pré-processing et une étape de réduction de dimension.

# Présentation des données

- Ensemble de données contenant des images de haut qualité de fruit avec les labels associés
- 22700 images au format JPG 100x100 pixels
- 131 variétés différentes
- Chaque fruit est photographié sous différents angles.

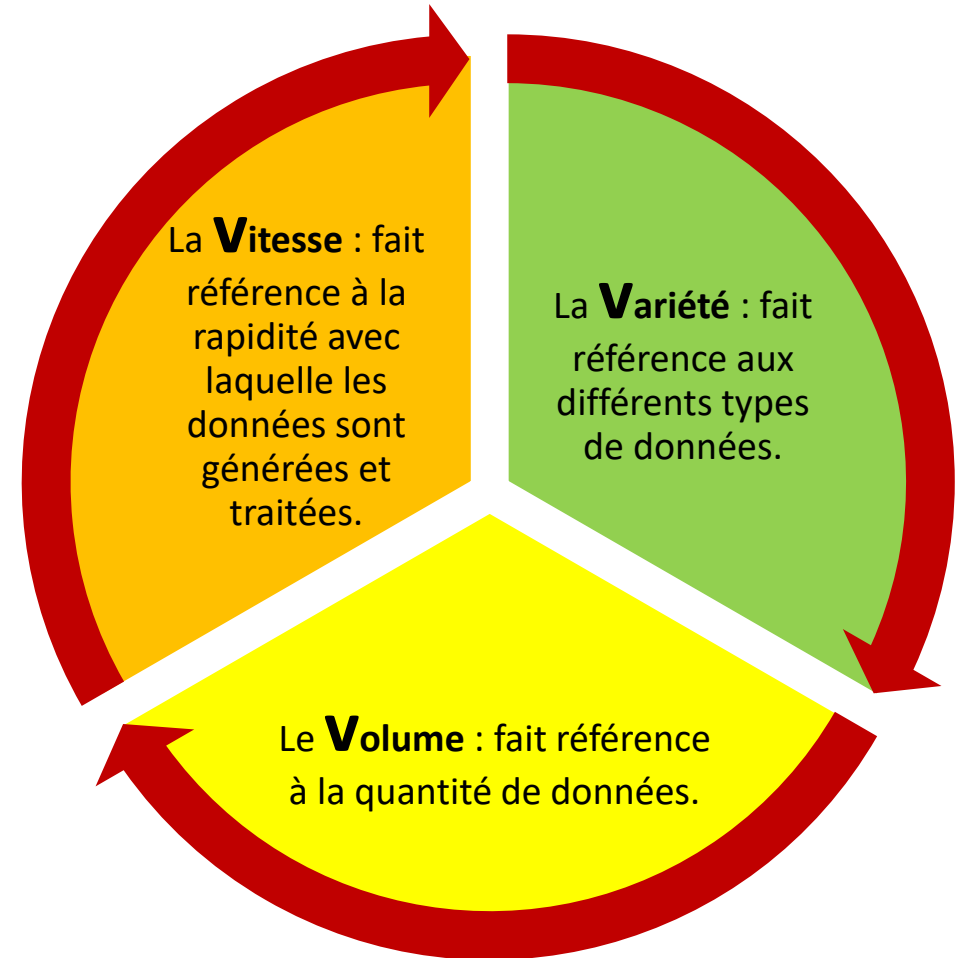


# Big Data

## 1- Qu'est-ce que le Big Data ?

Le **Big Data** est un terme utilisé pour décrire des ensembles de données volumineux, complexes et hétérogènes qui sont collectés à partir de diverses sources et qui ne peuvent pas être traités par les méthodes traditionnelles de gestion de données.

Les données du Big Data sont caractérisées par les 3V:  
le volume, la variété et la vitesse.

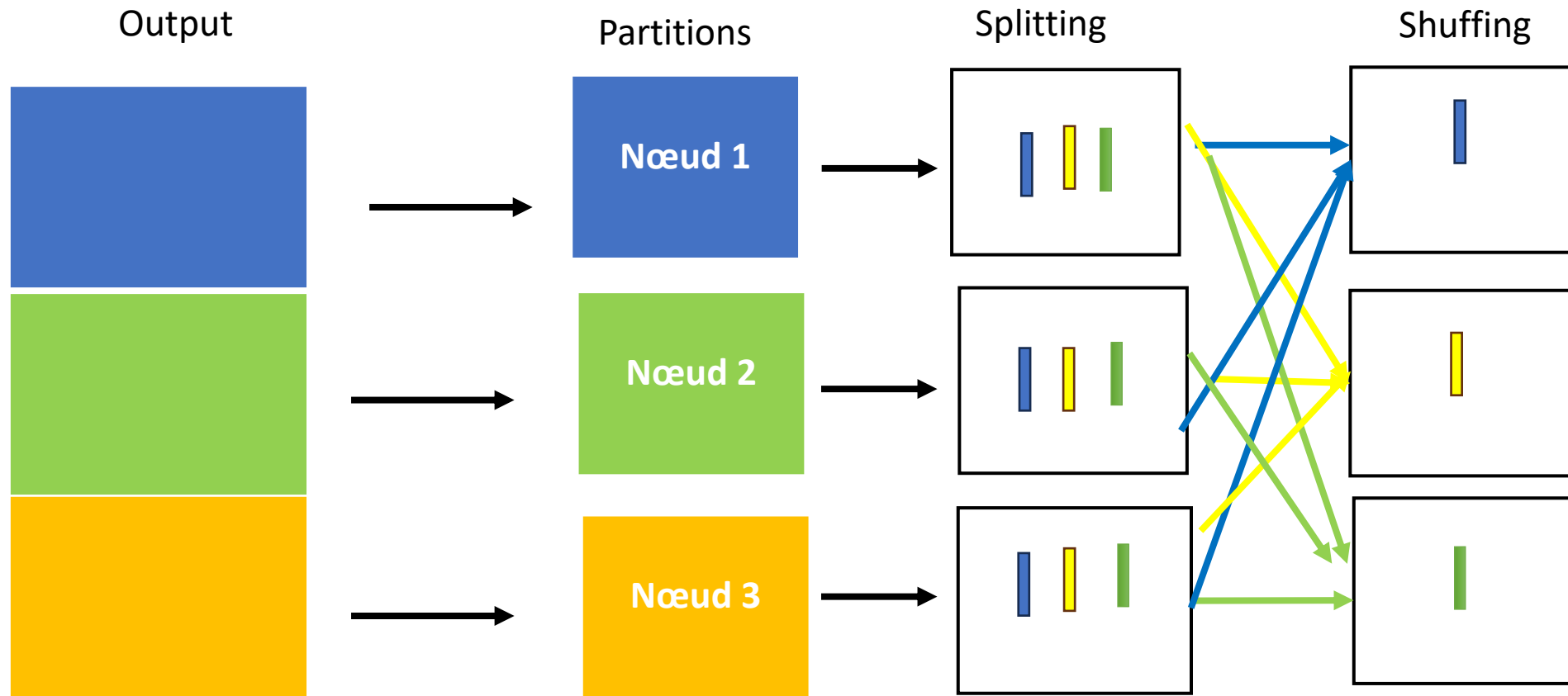


### 2- Big Data Frameworks



	Hadoop	Spark
<b>Architecture</b>	Stocke et traite les données sur un stockage externe	Stocke et traite les données dans la mémoire interne
<b>Traitement de données</b>	Par lots	En temps réel
<b>Performances</b>	Moins rapides	Plus rapides
<b>Machine learning</b>	Bibliothèques externes	Bibliothèques intégrées
<b>Sécurité</b>	Kerberos	authentification avec un mot de passe secret
<b>Base de données</b>	HDFS	HDFS,...

### 3- Comment Spark traite les données



### 3- Amazon Web Services



Le Cloud AWS est une plateforme de services cloud développée par le géant américain Amazon. AWS regroupe plus de 200 services répartis en diverses catégories telles que le stockage cloud, la puissance de calcul, l'analyse de données, l'intelligence artificielle ou même le développement de jeux vidéo.



**4- Configuration de la console AWS**

**Etape1:**

Création d'un bucket sur S3 dans lequel je télécharge le contenu du dossier Test, le fichier d'amorçage (zh\_bootstrap-emr.sh) et la création d'un dossier(OC\_Result) pour télécharger les parquets

<input type="radio"/>	zh-p8-bucket	Europe (Paris) eu-west-3	Compartiment et objets non publics	13 Nov 2023 02:34:10 PM CET		
<input type="checkbox"/>	OC_Result/	Dossier	-	-	-	
<input type="checkbox"/>	Test/	Dossier	-	-	-	
<input type="checkbox"/>	zh_bootstrap-emr.sh	sh	22 Nov 2023 09:26:57 AM CET	439.0 o	Standard	

### Etape2:

Création du cluster EMR dans Instances EC2 situées en France (eu-west-3b)

The screenshot shows the Amazon EMR console interface. At the top, the AWS logo and 'Services' menu are visible. The search bar contains 'Rechercher' and the keyboard shortcut '[Alt+S]'. The region is set to 'Paris' and the account ID is 'zh\_oc'. The breadcrumb navigation shows 'Amazon EMR > EMR sur EC2: Clusters'. The main section is titled 'Clusters (18)' with an 'Info' link. Below this, there are buttons for 'Afficher les détails', 'Résilier', 'Cloner', and 'Créer un cluster'. The 'Créer un cluster' button is highlighted with a red box. Below the buttons, there are filters: 'Filtrer les clusters par statut', 'Rechercher des clusters', and 'Filtrer les clusters par date et heure de création'. The 'Nom et applications' section is expanded, showing a 'Nom' field with the value 'P8-oc-cluster' entered. The 'Créer un cluster' button and the 'Nom' field are both highlighted with red boxes.

Big Data - suite

Offre d'applications

Spark



Core Hadoop



HBase



Presto



Trino



Custom



☐ Flink 1.16.0

☐ HCatalog 3.1.3

☐ Hue 4.10.0

☐ Livy 0.7.1

☐ Phoenix 5.1.2

☒ Spark 3.3.1

☐ Tez 0.10.2

☐ ZooKeeper 3.5.10

☐ Ganglia 3.7.2

☒ Hadoop 3.3.3

☐ JupyterEnterpriseGateway 2.6.0

☐ MXNet 1.9.1

☐ Pig 0.17.0

☐ Sqoop 1.4.7

☐ Trino 403

☐ HBase 2.4.15

☐ Hive 3.1.3

☒ JupyterHub 1.5.0

☐ Oozie 5.2.1

☐ Presto 0.278

☒ TensorFlow 2.11.0

☐ Zeppelin 0.10.1

J'ai choisi spark, tensorflow et jupyterhub comme application

Configuration de mise en service

Définissez la taille de votre noyau et tâchegroupes d'instance. Amazon EMR tente de fournir cette capacité lorsque vous lancez votre cluster.

Nom	Type d'instance	Taille de l'instance(s)
Unité principale	m5.xlarge	<input type="text" value="2"/>
Tâche - 1	m5.xlarge	<input type="text" value="1"/>

Je sélectionne 2 instances Principales et une instances maitres => 3 instances EC2

## Big Data - suite

### ▼ Actions d'amorçage – facultatif [Info](#)

Utilisez les actions d'amorçage pour installer des logiciels ou personnaliser la configuration de votre instance.

#### Actions d'amorçage (1)

[Supprimer](#)[Modifier](#)[Ajouter](#)

	Nom	Emplacement Amazon S3 <a href="#">↗</a>	Arguments
<input type="radio"/>	zh-p8-bucket	<a href="s3://zh-p8-bucket/zh_bootstrap-emr.sh">s3://zh-p8-bucket/zh_bootstrap-emr.sh</a>	-

**zh-bootstrap-emr.sh** est  
fichier pour installer les  
bibliothèques manquantes

### Configuration de sécurité et paire de clés EC2 – facultatif [Info](#)

#### Configuration de sécurité

Sélectionnez les paramètres de chiffrement, d'authentification, d'autorisation et de service de métadonnées d'instance de votre cluster.

[Parcourir \[↗\]\(#\)](#)[Créer une configuration de sécurité \[↗\]\(#\)](#)

#### Paire de clés Amazon EC2 pour SSH sur le cluster [Info](#)

[Parcourir](#)[Créer une paire de clés \[↗\]\(#\)](#)

A cette étape nous sélectionnons  
la paire de clés EC2 créé précédemment  
Elle nous permettra de nous connecter  
en ssh à nos instances EC2  
sans avoir à entrer nos login/mot de  
passe.

## Big Data - suite

Il ne nous reste plus qu'à attendre que le serveur soit prêt.  
Cette étape peut prendre entre **15 et 20 minutes**.

Amazon EMR > EMR sur EC2: Clusters > P8-oc-cluster

### P8-oc-cluster

Mis à jour il y a 9 minutes 🔄 Résilier Cloner dans AWS CLI Cloner

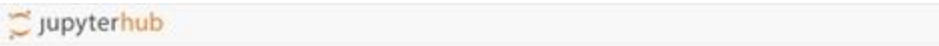
▼ Récapitulatif

Informations sur le cluster	Applications	Gestion des clusters	Statut et heure
<p>ID de cluster j-CF1DCROJWE7C</p> <p>Configuration de cluster Groupes d'instances</p> <p>Capacité 1 primaire(s) 3 unité(s) principale(s) 1 tâche(s)</p>	<p>Version d'Amazon EMR emr-6.10.0</p> <p>Applications installées Hadoop 3.3.3, JupyterHub 1.5.0, Spark 3.3.1, TensorFlow 2.11.0</p>	<p>Destination des journaux dans Amazon S3 <a href="#">aws-logs-134225938989-eu-west-3/elasticmapreduce</a></p> <p>Interfaces utilisateur d'application persistantes <a href="#">Serveur d'historique Spark</a> <a href="#">Serveur de chronologie YARN</a></p> <p>DNS public du nœud primaire <a href="#">ec2-35-181-59-142.eu-west-3.compute.amazonaws.com</a></p> <p><a href="#">Connexion au nœud primaire à l'aide de SSH</a> <a href="#">Connexion au nœud primaire à l'aide de</a></p>	<p>Statut <span>🟢 En attente</span></p> <p>Heure de création 23 novembre 2023 14:16 (UTC+01:00)</p> <p>Temps écoulé 1 heure, 2 minutes</p>

```
hadoop@ip-172-31-36-164:~  
  
 _ _ | _ _ )  
 _ | ( _ | /  Amazon Linux 2 AMI  
 _ | \ _ | _ |  
  
https://aws.amazon.com/amazon-linux-2/  
58 package(s) needed for security, out of 88 available  
Run "sudo yum update" to apply all updates.  
  
EEEEEEEEEEEEEEEEEEEE MMMMMMM MMMMMMM RRRRRRRRRRRRRR  
E::::::::::::::::::::E M::::::::M M::::::::M R::::::::R  
EE::::::::EEEEEEEE::::E M::::::::M M::::::::M R::::RRRRRR::::R  
E::::E EEEEE M::::::::M M::::::::M RR::::R R::::R  
E::::E M::::::::M M::M::::M R::R R::::R  
E::::EEEEEEEE M::::M M::M M::M M::M R::RRRRR::::R  
E::::::::::::E M::::M M::M::::M M::M R::::::::RR  
E::::EEEEEEEE M::::M M::M M::M R::RRRRR::::R  
E::::E M::::M M::M M::M R::R R::::R  
E::::E EEEEE M::::M MMM M::M R::R R::::R  
EE::::::::EEEEEEEE::::E M::::M M::M R::R R::::R  
E::::::::::::E M::::M M::M RR::::R R::::R  
EEEEEEEEEEEEEEEEEEEE MMMMMMM MMMMMMM RRRRRRR RRRRRR  
  
[hadoop@ip-172-31-36-164 ~]$
```

Nous avons correctement établi le tunnel **ssh** avec le driver sur le port "5555"

Etape3:



Sign in

Username:  
jovyan

Password:  
\*\*\*\*\*

Sign in

Objets (2)

Les objets sont les entités fondamentales stockées dans Amazon S3. Vous pouvez utiliser l'[inventaire Amazon S3](#) pour obtenir une liste de tous les objets de votre compartiment. Pour que d'autres personnes puissent accéder à vos objets, vous devez leur accorder explicitement des autorisations. [En savoir plus](#)

Copier l'URI S3

Copier l'URL

Télécharger

Ouvrir

Supprimer

Actions ▼

Créer un dossier

Charger

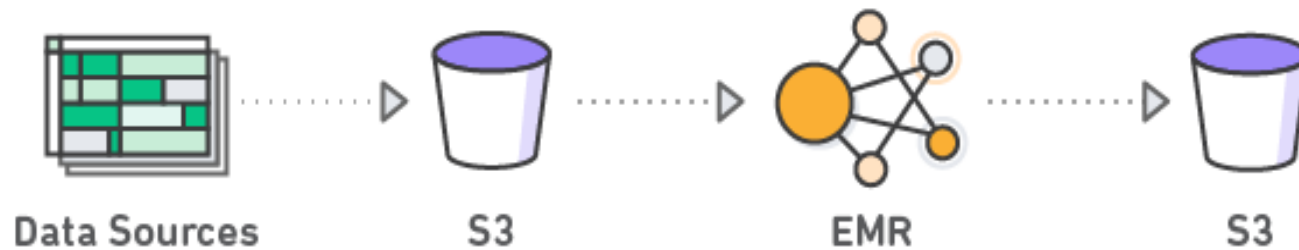
Rechercher des objets en fonction du préfixe

☐ Afficher les versions

< 1 >

<input type="checkbox"/>	Nom ▲	Type ▼	Dernière modification ▼	Taille ▼	Classe de stockage ▼
<input type="checkbox"/>	.s3keep	s3keep	20 Nov 2023 09:42:39 AM CET	0 o	Standard
<input type="checkbox"/>	notebook.ipynb	ipynb	20 Nov 2023 02:17:12 PM CET	50.9 Ko	Standard

# Traitement des images



<https://www.kaggle.com/datasets/moltean/fruits>

Télécharger  
Les données  
Dans S3

Utilisez Amazon  
EMR pour  
Traitement des  
images + Extraction  
Features +  
Réduction de  
Dimension (ACP)

Charger des  
données au  
format parquet  
dans S3

## Traitement d'image - suite

- 1280 descripteur par image
- Conversion au format vecteur dense
- Standardisation

path	label	features	features_vectorized	scaledFeatures
s3://zh-p8-bucket...	Pineapple Mini	[0.0, 5.068508, 0...	[0.0,5.0685081481...	[-0.7568072935344...
s3://zh-p8-bucket...	Pineapple Mini	[0.015393238, 4.6...	[0.01539323758333...	[-0.7282748230803...
s3://zh-p8-bucket...	Watermelon	[0.11004015, 0.05...	[0.11004015058279...	[-0.5528399781295...
s3://zh-p8-bucket...	Raspberry	[0.6104076, 0.644...	[0.61040759086608...	[0.37462697797761...
s3://zh-p8-bucket...	Raspberry	[0.079946205, 0.5...	[0.07994620501995...	[-0.6086212657221...
s3://zh-p8-bucket...	Cauliflower	[0.0, 0.63743114,...	[0.0,0.6374311447...	[-0.7568072935344...

### PCA finale

label	features_pca
Pineapple Mini	[-9.3577244748706...
Pineapple Mini	[-6.8753150180104...
Pineapple Mini	[-13.269800796270...
Raspberry	[-2.4566974559442...
Raspberry	[-6.0856911357879...
Cauliflower	[-4.4442835489278...

only showing top 6 rows

```
cumsum = 0
for i in pca.explainedVariance.cumsum():
    cumsum += 1
    if(i > 0.8):
        print(
            '{} composantes expliquent 80% de la variance'.format(cumsum))
        break
```

161 composantes expliquent 80% de la variance

```
pca = PCA(
    k=cumsum,
    inputCol='scaledFeatures',
    outputCol='features_pca')
model_pca = pca.fit(df_scaled)
df_final = model_pca.transform(df_scaled)
```



### Historique de session Spark

Job Id (Job Group)	Description	Submitted	Duration ▾	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
23 (20)	Job group for statement 20 first at PCA.scala:44	2023/12/03 22:06:05	4.2 min	2/2	711/711
28 (25)	Job group for statement 25 first at PCA.scala:44	2023/12/03 22:18:08	4.1 min	2/2	711/711
35 (30)	Job group for statement 30 first at PCA.scala:44	2023/12/03 22:39:34	4.0 min	2/2	711/711
4 (7)	Job group for statement 7 showString at NativeMethodAccessorImpl.java:0	2023/12/03 21:24:04	4.0 min	1/1	710/710
9 (9)	Job group for statement 9 showString at NativeMethodAccessorImpl.java:0	2023/12/03 21:29:35	3.8 min	1/1	710/710
33 (29)	Job group for statement 29 showString at NativeMethodAccessorImpl.java:0	2023/12/03 22:31:03	3.7 min	1/1	710/710
11 (10)	Job group for statement 10 first at StandardScaler.scala:113	2023/12/03 21:33:58	3.6 min	1/1	710/710

Sauvegarde sur S3

Objets (21) [Info](#)

Les objets sont les entités fondamentales stockées dans Amazon S3. Vous pouvez utiliser l'[inventaire Amazon S3](#) pour obtenir une liste de tous les objets de votre compartiment. Pour que d'autres personnes puissent accéder à vos objets, vous devez leur accorder explicitement des autorisations. [En savoir plus](#)

↻

📄 Copier l'URI S3

📄 Copier l'URL

⬇️ Télécharger

Ouvrir [↗](#)

Supprimer

Actions ▼

Créer un dossier

📁 Charger

🔍 Rechercher des objets en fonction de

🔍

🔌 Afficher les versions

< 1 >

⚙️

<input type="checkbox"/>	Nom ▲	Type ▼	Dernière modification ▼	Taille ▼	Classe de stockage ▼
<input type="checkbox"/>	<div>📄 <a href="#">_SUCCESS</a></div>	-	04 Dec 2023 12:00:00 AM CET	0 o	Standard
<input type="checkbox"/>	<div>📄 <a href="#">part-00000-24818291-8ee1-4e72-8106-a1dc6494652d-c000.snappy.parquet</a></div>	parquet	03 Dec 2023 11:58:18 PM CET	1.4 Mo	Standard
<input type="checkbox"/>	<div>📄 <a href="#">part-00001-24818291-8ee1-4e72-8106-</a></div>	parquet	03 Dec 2023 .....	1.4 Mo	Standard

Zouheir HMIDI

18

# Conclusions

- ❖ Le projet a permis de déployer un développement ML sur le cloud en environnement Big Data, en utilisant :
  - Apache Spark et Pyspark pour les traitements distribués
  - AWS IAM pour la gestion des utilisateurs et des autorisations
  - AWS S3 pour le stockage des données
  
- ❖ Limites :
  - outil payant