



OPENCLASSROOMS

## Projet 9

-

Réalisez un traitement dans un  
environnement Big Data sur le Cloud



CentraleSupélec

Pierrick BERTHE

Formation Expert en Data Science  
*Openclassrooms – CentraleSupélec*

*août 2023 → juin 2024*



# Problématique

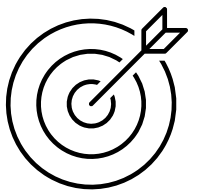
La jeune startup « Fruit » qui veut créer une application mobile qui permettant aux utilisateurs de prendre en photo un fruit et d'obtenir des informations sur ce fruit.



**=> Le développement de cette application nécessite un traitement Big Data pour la reconnaissance d'image.**

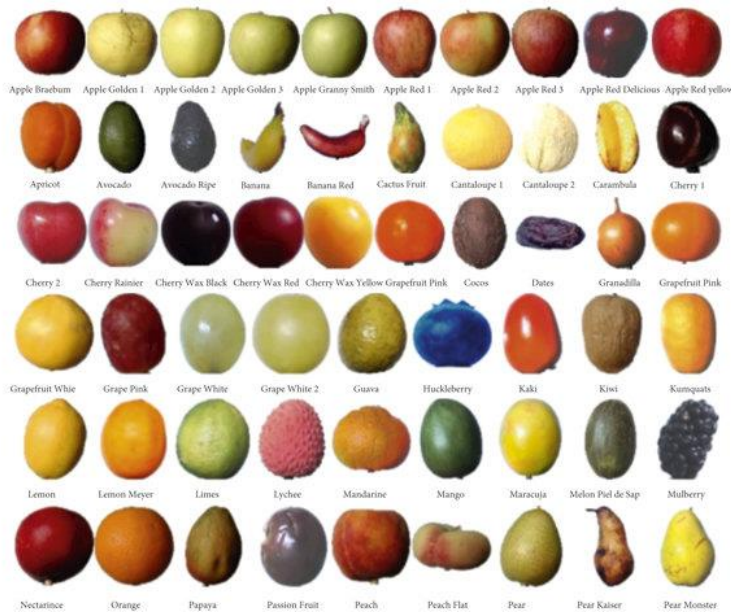
## Missions :

1. Expliquer la chaîne de traitement des données
2. Détailler l'architecture Big Data retenue
3. Veiller au respect des contraintes RGPD (serveur dans l'UE)
4. Apporter un retour critique sur la solution proposée





# Présentation du jeu de données



Ce jeu de données comporte des images de fruits :

- 22\_688 images
- 131 catégories de fruits
- 100 x 100 pixels

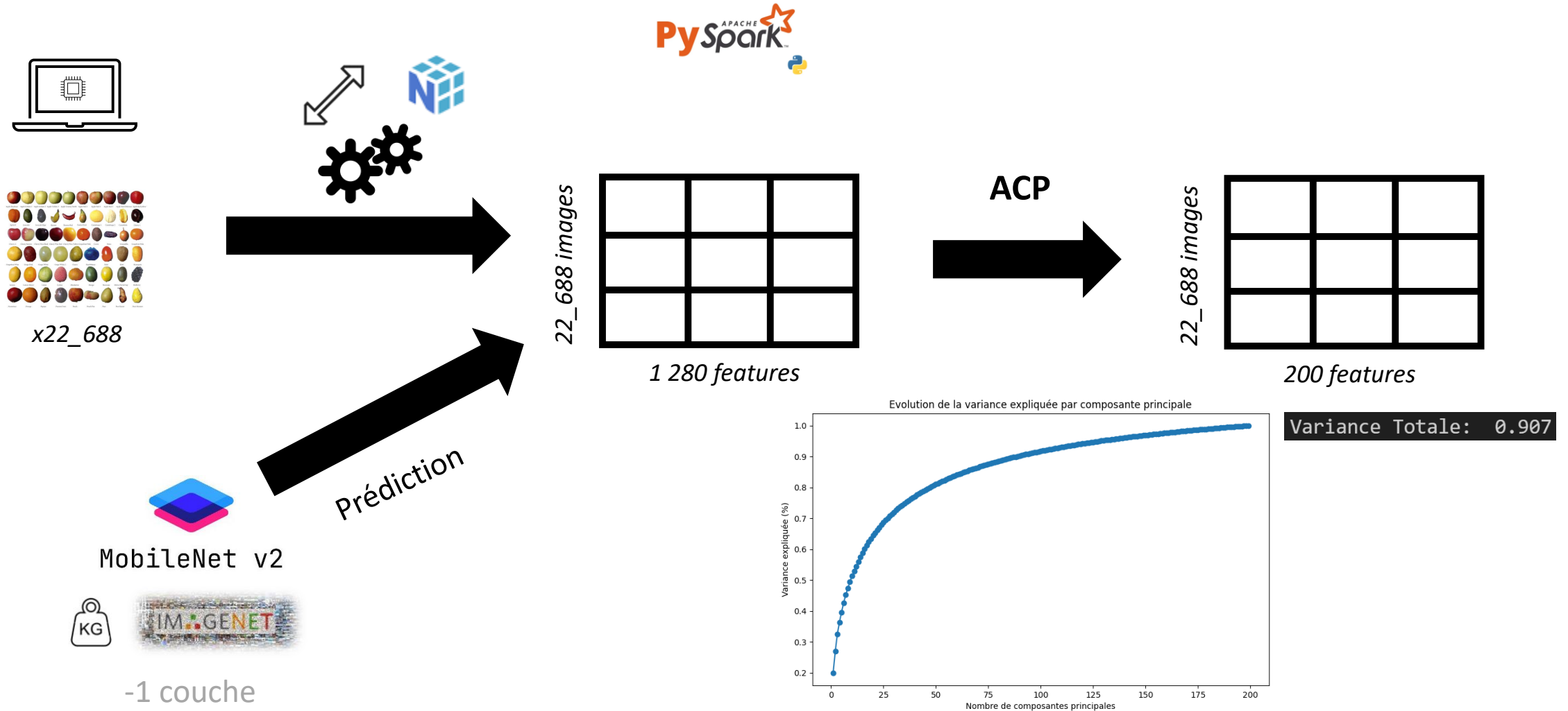
|-- Apple Golden 1/  
| |-- 63\_100.jpg  
| |-- ...

|-- Apple Golden 2/  
| |-- 3\_100.jpg  
| |-- ...

|-- Apple Golden 3/  
| |-- ...

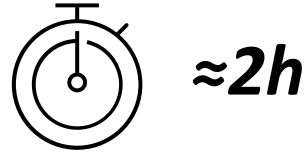
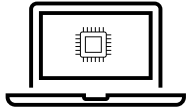


# Chaine de traitement de données





# Problématique Big Data



- Temps de calcul important
- Grand nombre d'image à traiter
- Mémoire et capacité calcul limité en local
- Risque de panne

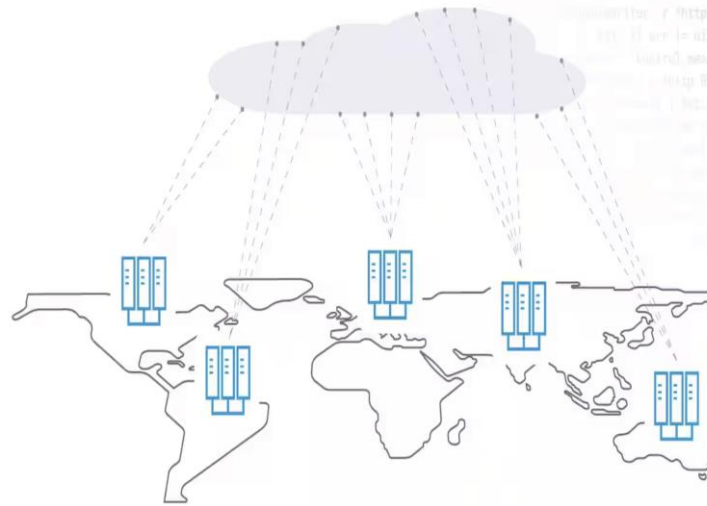




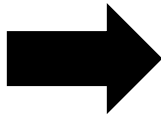
# Solution Big Data



Le **calcul distribué** consiste à répartir une énorme tâche de calcul sur un volume de données gigantesque avec sur différents ordinateurs à travers le monde



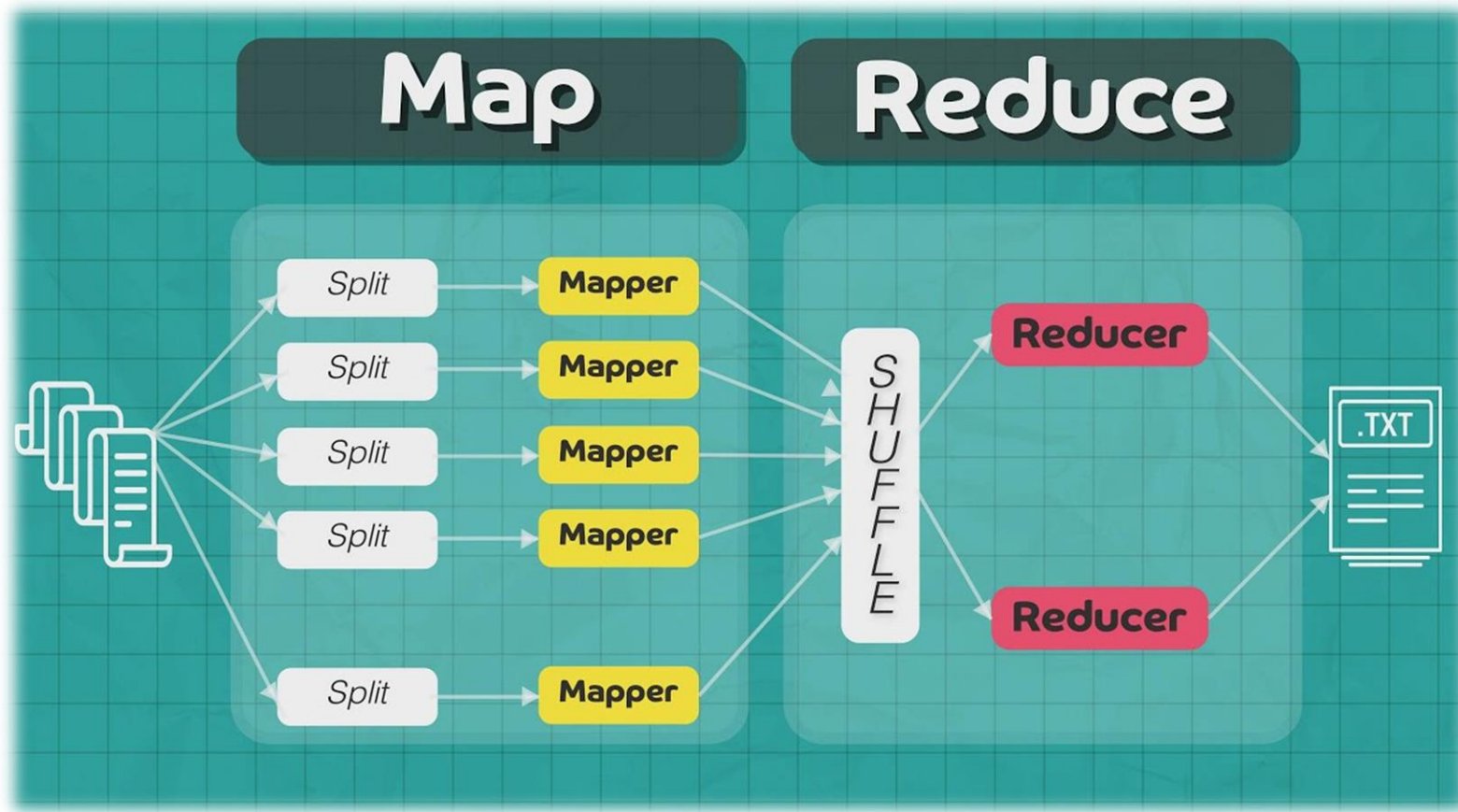
#DeFi



- Optimisation des transferts disques et réseau
- Scalabilité pour adapter la puissance au besoin
- Tolérance aux pannes



# Solution Big Data

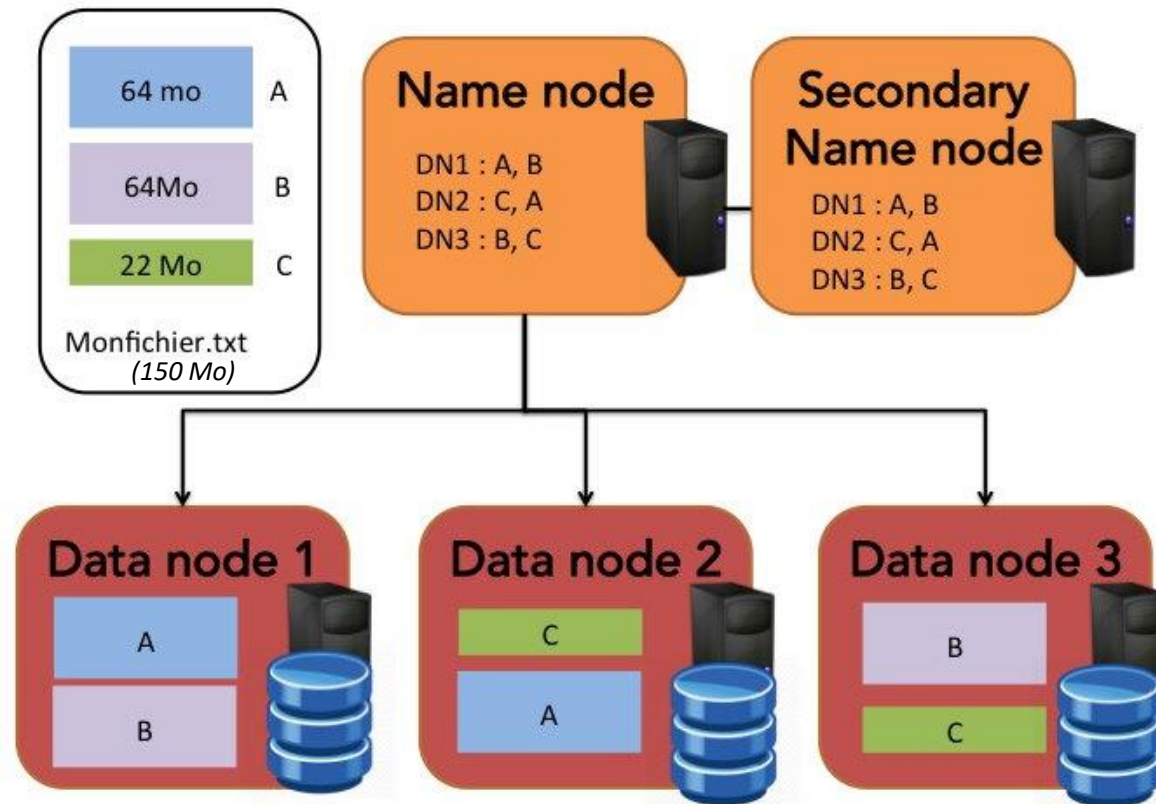




# Solution Big Data



*Hadoop Distributed File System*



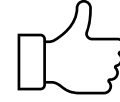







# Solution Big Data



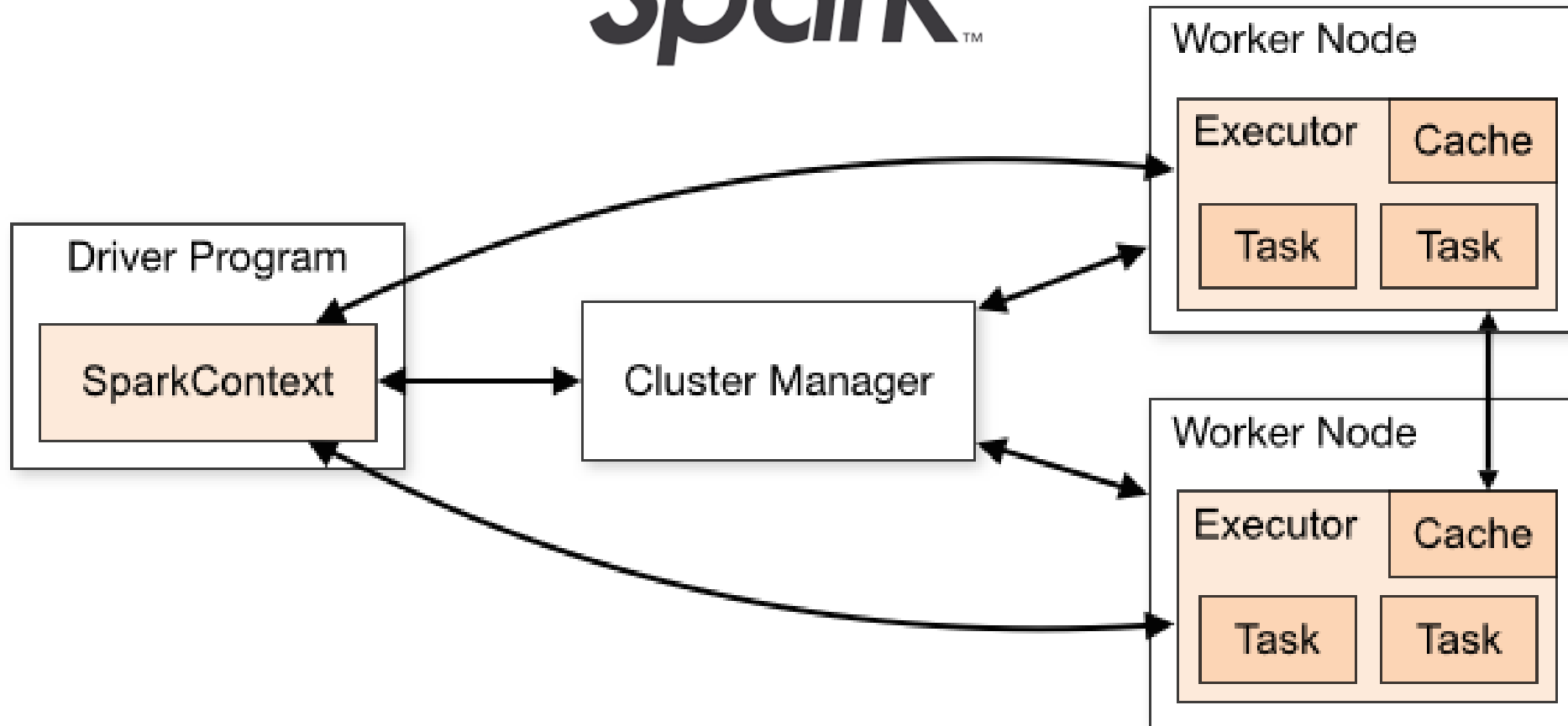
- Stockage et traitement sur disque
- Traitement des données par lot
- ML par l'intégration de bibliothèques externes



- Stockage et traitement sur mémoire vive (RAM) 
- Traitement des données par lot / temps réel / itérations 
- Bibliothèque ML intégrée (SparkML) 



# Solution Big Data

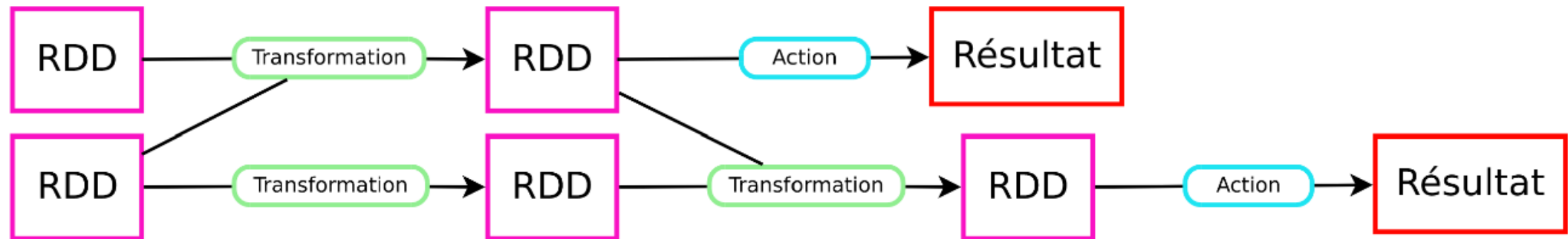




# Solution Big Data



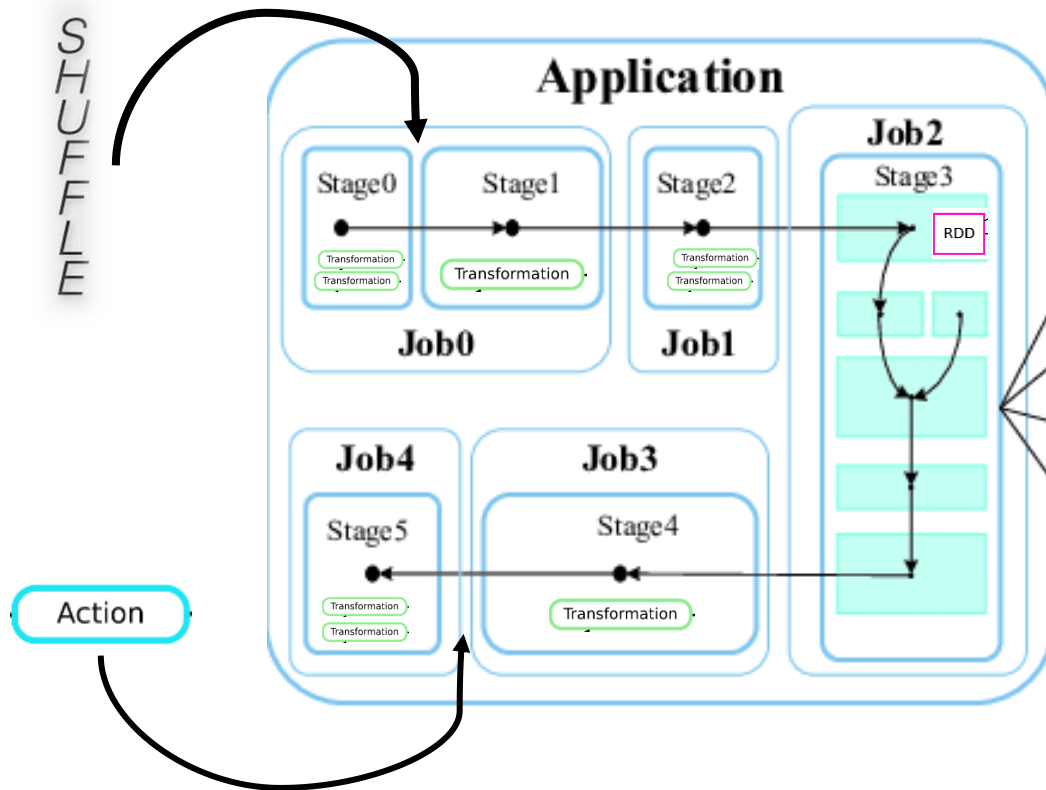
- Transformation :  $\text{RDD} \Rightarrow \text{RDD}$
- Action :  $\text{RDD} \Rightarrow \text{RDD}$
- Lazy evaluation



*Résilient  
Distributed  
Dataset*



# Solution Big Data



**DAG**

*Directed  
Acyclic  
Graph*



# Architecture Big Data

Louer de la puissance de  
calcul à la demande



Coût faible

Leader du marché

Nombre de produits  
importants



# Architecture Big Data



IAM

Utilisateur  
avec tous les accès :  
- *S3FullAccess*  
- *EMRFullAccess*  
- *EC2FullAccess*



amazon  
S3

Images stockées sur cloud

Lecture / écriture fichiers

Résultats stockés sur le cloud



amazon  
EMR

IaaS / PaaS

Cluster préconfiguré



Installation librairies supp



Amazon  
EC2

**4 instances :**

- 1 instance maître
- 1 instance core
- 2 instances workers

m5.xlarge

Connexion SSH (Foxy proxy)



# Architecture Big Data



IAM



```
C:\Users\pierr>aws configure
AWS Access Key ID [*****4LRJ]: [REDACTED]4LRJ
AWS Secret Access Key [*****Na7P]: 9rsIwy[REDACTED]IAxG3GeONq7P
Default region name [eu-west-3]: eu-west-3
Default output format [json]: json
```

=> Paire de clef sécurisant l'accès de l'utilisateur



IAM > Utilisateurs > Pierrick

**Pierrick** Infos Supprimer

**Récapitulatif**

ARN arn:aws:iam::[REDACTED]:user/Pierrick	Accès par console ⚠ Activé sans l'authentification MFA	Clé d'accès 1 AKIA[REDACTED]4LRJ - Active ✔ Utilisé Hier. Hier ancien.
Création May 29, 2024, 09:54 (UTC+02:00)	Dernière connexion à la console ✔ Hier	Clé d'accès 2 <a href="#">Créer une clé d'accès</a>

**Autorisations** | Groupes (1) | Balises (1) | Informations d'identification de sécurité | Access Advisor

**Politiques des autorisations (1)** 🔄 Supprimer Ajouter des autorisations ▼

Les autorisations sont définies par des politiques attachées à l'utilisateur directement ou via des groupes.

Rechercher  Filtrer par Type Tous les types ▼ < 1 > ⚙

<input type="checkbox"/>	Nom de la politique <a href="#">🔗</a>	Type ▼	Attaché via <a href="#">🔗</a>
<input type="checkbox"/>	AdministratorAccess	Gérées par AWS – fonction professionnelle	Groupe <a href="#">admin</a>



# Architecture Big Data



```
C:\Users\pierr\VSC_Projects\Projet9_OCR_DataScientist\dossier_sync_AWS_S3>aws s3 mb s3://pb-ocr-p9
```



**Compartiments à usage général (1)** [Info](#) Toutes les régions AWS ↻ Copier l'ARN Vider Supprimer Créer un compartiment

Les compartiments sont des conteneurs pour les données stockées dans S3.

Nom	Région AWS	Analyseur d'accès IAM	Date de création
<input type="radio"/> <a href="#">pb-ocr-p9</a>	Europe (Paris) eu-west-3	<a href="#">Afficher l'analyseur pour eu-west-3</a>	29 May 2024 10:47:04 AM CEST



```
C:\Users\pierr\VSC_Projects\Projet9_OCR_DataScientist\dossier_sync_AWS_S3>aws s3 sync . s3://pb-ocr-p9
```



**Objets (131)** [Info](#) ↻ Copier l'URI S3 Copier l'URL Télécharger Ouvrir Supprimer Actions Créer un dossier Charger

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](#)

<input type="checkbox"/>	Nom	Type	Dernière modification	Taille	Classe de stockage
<input type="checkbox"/>	<a href="#">Apple Braeburn/</a>	Dossier	-	-	-
<input type="checkbox"/>	<a href="#">Apple Crimson Snow/</a>	Dossier	-	-	-
<input type="checkbox"/>	<a href="#">Apple Golden 1/</a>	Dossier	-	-	-
<input type="checkbox"/>	<a href="#">Apple Golden 2/</a>	Dossier	-	-	-
<input type="checkbox"/>	<a href="#">Apple Golden 3/</a>	Dossier	-	-	-





# Architecture Big Data



```
dossier_sync_AWS_S3 > {} jupyter-s3-conf.json > ...
1  [
2    {
3      "classification": "jupyter-s3-conf",
4      "properties": {
5        "s3.persistence.bucket": "pb-ocr-p9",
6        "s3.persistence.enabled": "true"
7      }
8    }
9  ]
```

```
$ bootstrap-emr.sh
dossier_sync_AWS_S3 > $ bootstrap-emr.sh
1  sudo python3 -m pip install -U setuptools
2  sudo python3 -m pip install -U pip
3  sudo python3 -m pip install wheel pillow tensorflow pyarrow boto3 s3fs fsspec
   ipython
4  sudo python3 -m pip install pandas==1.2.5
5  sudo python3 -m pip install matplotlib==3.4.3
```



```
C:\Users\pierr\VSC_Projects\Projet9_OCR_DataScientist\dossier_sync_AWS_S3>aws s3 sync . s3://pb-ocr-p9
upload: .\bootstrap-emr.sh to s3://pb-ocr-p9/bootstrap-emr.sh
upload: .\jupyter-s3-conf.json to s3://pb-ocr-p9/jupyter-s3-conf.json
upload: .\Berthe_Pierrick_1_notebook_052024.ipynb to s3://pb-ocr-p9/Berthe_Pierrick_1_notebook_052024.ipynb
```



Amazon S3 > Compartiments > pb-ocr-p9

pb-ocr-p9 Info

Objets Propriétés Autorisations Métriques Gestion Points d'accès

Objets (3) Info Copier l'URI S3 Copier l'URL Télécharger Ouvrir Supprimer Actions Créer un dossier Charger

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](#)

Rechercher des objets en fonction du préfixe

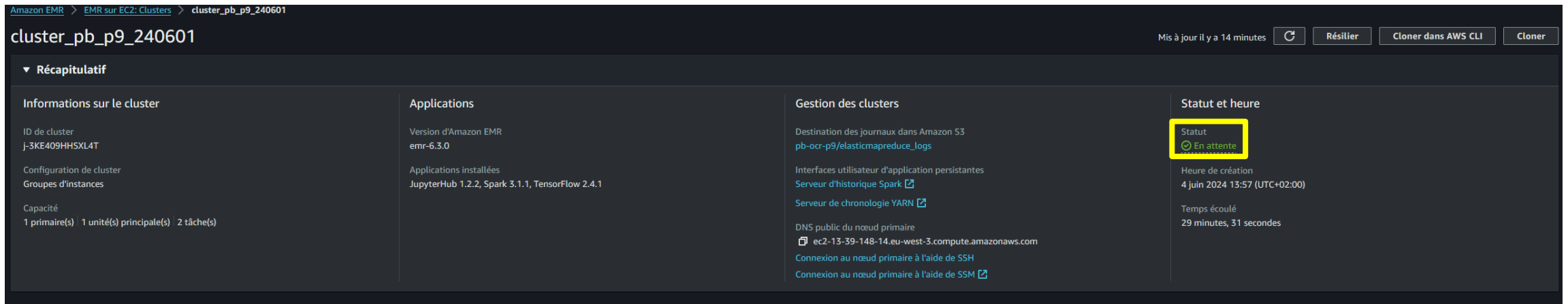
	Nom	Type	Dernière modification	Taille	Classe de stockage
<input type="checkbox"/>	bootstrap-emr.sh	sh	31 May 2024 12:21:55 PM CEST	161.0 o	Standard
<input type="checkbox"/>	jupyter-s3-conf.json	json	31 May 2024 12:21:56 PM CEST	155.0 o	Standard
<input type="checkbox"/>	Test/	Dossier	-	-	-



# Architecture Big Data



- Version AWS EMR : **emr-6.3.0**
- Applications : **JupyterHub 1.2.2, Spark 3.1.1, TensorFlow 2.4.1**
- Instance EC2: **1 Primaire** (m5.xlarge), **1 Unité principale** (m5.xlarge), **2 Tâches** (m5.xlarge)
- Résilier le cluster après le temps d'inactivité: **1 heure**
- Action d'amorçage : « **s3://pb-ocr-p9/bootstrap-emr.sh** »
- Paramètre du logiciel : « **jupyter-s3-conf.json** »
- Paire de clé EC2 : « **clef\_ssh\_aws\_ec2.ppk** »



Amazon EMR > EMR sur EC2: Clusters > cluster\_pb\_p9\_240601

cluster\_pb\_p9\_240601

Mis à jour il y a 14 minutes [Régénérer](#) [Résilier](#) [Cloner dans AWS CLI](#) [Cloner](#)

▼ Récapitulatif

Informations sur le cluster	Applications	Gestion des clusters	Statut et heure
<p>ID de cluster j-3KE409HHSXL4T</p> <p>Configuration de cluster Groupes d'instances</p> <p>Capacité 1 primaire(s) 1 unité(s) principale(s) 2 tâche(s)</p>	<p>Version d'Amazon EMR emr-6.3.0</p> <p>Applications installées JupyterHub 1.2.2, Spark 3.1.1, TensorFlow 2.4.1</p>	<p>Destination des journaux dans Amazon S3 <a href="#">pb-ocr-p9/elasticmapreduce_logs</a></p> <p>Interfaces utilisateur d'application persistantes <a href="#">Serveur d'historique Spark</a></p> <p><a href="#">Serveur de chronologie YARN</a></p> <p>DNS public du nœud primaire <a href="#">ec2-13-39-148-14.eu-west-3.compute.amazonaws.com</a></p> <p><a href="#">Connexion au nœud primaire à l'aide de SSH</a></p> <p><a href="#">Connexion au nœud primaire à l'aide de SSM</a></p>	<p>Statut <b>En attente</b></p> <p>Heure de création 4 juin 2024 13:57 (UTC+02:00)</p> <p>Temps écoulé 29 minutes, 31 secondes</p>



# Architecture Big Data



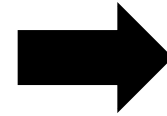
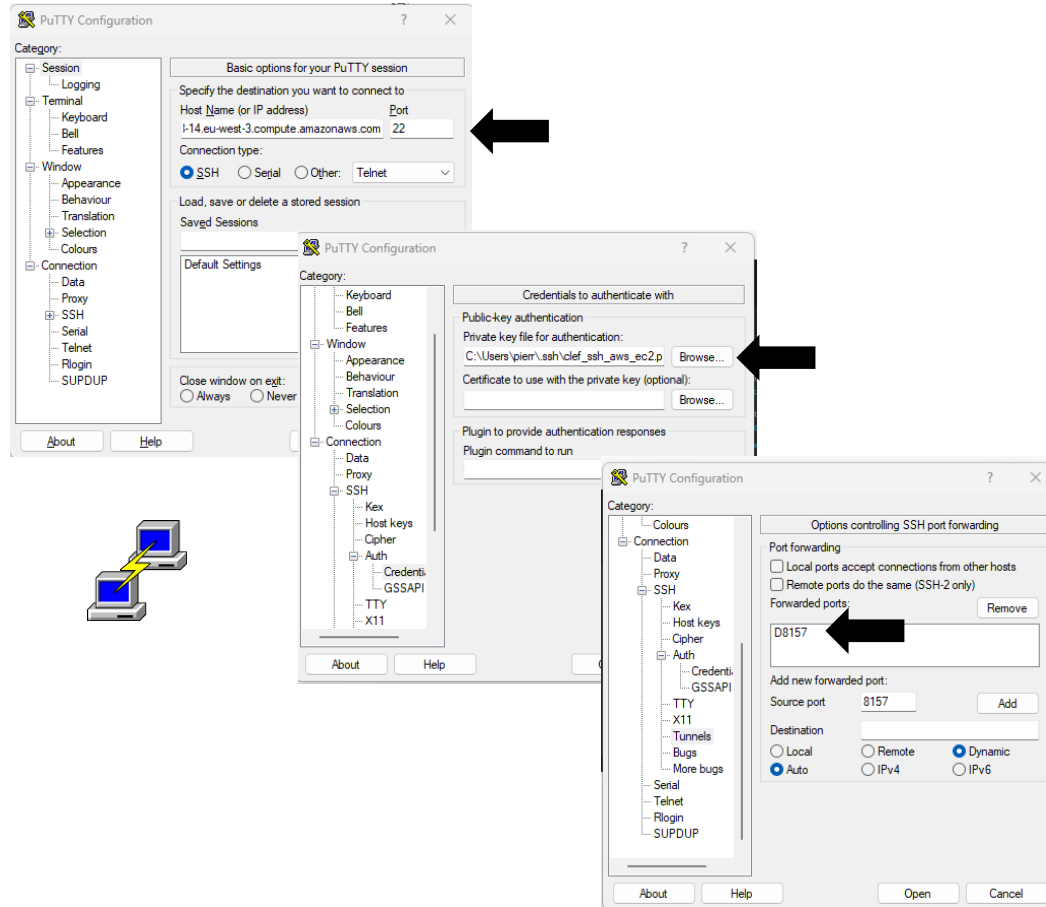
amazon  
EMR

## Modifier groupe de sécurité EC2 du driver

Règles entrantes									
Règles entrantes (8)									
<input type="text" value="Recherche"/>									
<div><span>↻</span> <span>Gérer les balises</span> <span>Modifier les règles entrantes</span></div>									
<div>&lt; 1 &gt; ⚙</div>									
<input type="checkbox"/>	Name	ID de règle de grou...	Version IP	Type	Protocole	Plage de ports	Sou		
<input type="checkbox"/>	-	sgr-0bc8cbcd066f300e2	-	Tous les UDP	UDP	0 - 65535	sg-C		
<input type="checkbox"/>	-	sgr-07cab499d5be1d7f	-	Tous les ICMP - IPv4	ICMP	Tous	sg-C		
<input type="checkbox"/>	-	sgr-0ca942ee0b47637...	IPv4	SSH	TCP	22	109.		
<input type="checkbox"/>	-	sgr-01f8628f44f1352c0	-	TCP personnalisé	TCP	8443	<a href="#">pl-4</a>		
<input type="checkbox"/>	-	sgr-0f00802f0be7a6ae4	-	Tous les ICMP - IPv4	ICMP	Tous	sg-C		
<input type="checkbox"/>	-	sgr-054ab6969f685e4d8	-	Tous les UDP	UDP	0 - 65535	sg-C		
<input type="checkbox"/>	-	sgr-09242bcee37f86ad4	-	Tous les TCP	TCP	0 - 65535	sg-C		
<input type="checkbox"/>	-	sgr-06cbcb6165ef36106	-	Tous les TCP	TCP	0 - 65535	sg-C		



# Architecture Big Data



```
hadoop@ip-172-31-46-55:~$  
Amazon Linux 2 AMI  
https://aws.amazon.com/amazon-linux-2/  
83 package(s) needed for security, out of 132 available  
Run "sudo yum update" to apply all updates.  
EEEEEEEEEEEEEEEEEEEE MMMMMMM MMMMMMM RRRRRRRRRRRRRR  
E:::::EEEEEEEEEEEE M:::::M M:::::M R:::::R  
EE:::::EEEEEEEEEEEE M:::::M M:::::M R:::::R  
E:::::E EEEEE M:::::M M:::::M RR:::R R:::R  
E:::::E M:::::M M:::::M R:::R R:::R  
E:::::EEEEEEEEEE M:::::M M:::::M R:::R R:::R  
E:::::EEEEEEEEEE M:::::M M:::::M R:::R R:::R  
E:::::E M:::::M M:::::M R:::R R:::R  
E:::::E EEEEE M:::::M M M M:::::M R:::R R:::R  
EE:::::EEEEEEEEEE M:::::M M:::::M R:::R R:::R  
E:::::EEEEEEEEEE M:::::M M:::::M RR:::R R:::R  
EEEEEEEEEEEEEEEEEEEE MMMMMMM MMMMMMM RRRRRRR RRRRRR  
[hadoop@ip-172-31-46-55 ~]$
```



# Architecture Big Data



Propriétés | Actions d'amorçage | Instances (Matériel) | Étapes | Applications | Configurations | Surveillance | Événements | identifications (1)

### Interfaces utilisateur d'application Info

Les applications installées sur votre cluster Amazon EMR publient des interfaces utilisateur en tant que sites web. Vous pouvez les utiliser pour surveiller l'activité du cluster.

☒ Interfaces utilisateur d'application sur le cluster  
Les interfaces utilisateur sur le cluster sont disponibles uniquement pendant l'exécution de votre cluster. Utilisez les liens suivants pour démarrer. Pour accéder à toutes les interfaces utilisateur d'application, configurez le tunneling SSH.

☐ Interfaces utilisateur d'application persistantes  
Les interfaces utilisateur persistantes ne nécessitent pas de tunneling SSH. Elles sont hébergées hors du cluster et sont disponibles pendant 30 jours après la fin d'une application.

### Interfaces utilisateur d'application en direct

Ces interfaces utilisateur d'application sur cluster sont disponibles sans tunneling SSH.

Interfaces utilisateur d'application [🔗](#)

[Interface utilisateur du serveur d'historique Spark](#)

### Interfaces utilisateur d'application sur le nœud primaire

Celles-ci nécessitent l'activation du tunneling SSH. [Activer une connexion SSH](#)

Application	URL de l'interface utilisateur <a href="#">🔗</a>
Gestionnaire de ressources	<a href="http://ec2-13-39-148-14.eu-west-3.compute.amazonaws.com:8088/">http://ec2-13-39-148-14.eu-west-3.compute.amazonaws.com:8088/</a>
JupyterHub	<a href="https://ec2-13-39-148-14.eu-west-3.compute.amazonaws.com:9443/">https://ec2-13-39-148-14.eu-west-3.compute.amazonaws.com:9443/</a>
Nom du nœud HDFS	<a href="http://ec2-13-39-148-14.eu-west-3.compute.amazonaws.com:9870/">http://ec2-13-39-148-14.eu-west-3.compute.amazonaws.com:9870/</a>
Serveur d'historique Spark	<a href="http://ec2-13-39-148-14.eu-west-3.compute.amazonaws.com:18080/">http://ec2-13-39-148-14.eu-west-3.compute.amazonaws.com:18080/</a>



Options | Proxies | Importer les paramètres | Pattern Tester | Journal | Aide | A propos

Ajouter

Get Location

EMR

Nom ou Description (optionnel)

EMR

Hostname

localhost

Type

socks5

Port

8157

Country

Nom d'utilisateur (optionnel)

username

City

city

Mot de passe (optionnel)

\*\*\*\*

Couleur

PAG URL

Proxy DNS

☐

Save Location

Save

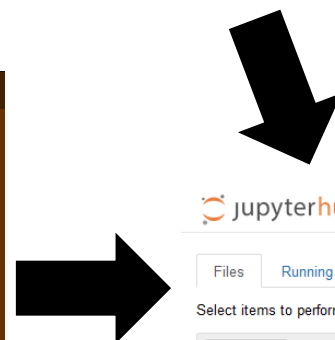
Quick Add

Include

Type

Nom ou Description (optionnel)

Modèles



Logout Control Panel

Files Running Clusters

Select items to perform actions on them.

☐

0

/

☐

Berthe\_Pierrick\_1\_notebook\_052024.ipynb

Activé il y a 7 minutes



# Architecture Big Data



jupyterhub Berthe\_Pierrick\_1\_notebook\_052024 (auto-sauvegardé) Logout Control Panel

File Edit View Insert Cell Kernel Widgets Help Fiable PySpark

Exécuter

s3://pb-ocr-p9/Te...	Watermelon	[0.03737948, 0.0...	[0.03737948089838...
s3://pb-ocr-p9/Te...	Watermelon	[0.119276375, 0.0...	[0.11927637457847...
s3://pb-ocr-p9/Te...	Raspberry	[0.61040884, 0.64...	[0.61040884256362...
s3://pb-ocr-p9/Te...	Cauliflower	[0.028072517, 1.5...	[0.02807251736521...
s3://pb-ocr-p9/Te...	Cauliflower	[0.0, 1.253267, 1...	[0.0, 1.2532670497...
s3://pb-ocr-p9/Te...	Pineapple	[0.15377553, 3.00...	[0.15377552807331...
s3://pb-ocr-p9/Te...	Cauliflower	[0.0, 2.1522408, ...]	[0.0, 2.1522407531...
s3://pb-ocr-p9/Te...	Pineapple Mini	[1.921151E-4, 4.7...	[1.92115097888745...
s3://pb-ocr-p9/Te...	Onion White	[0.08827547, 0.02...	[0.08827546983957...
s3://pb-ocr-p9/Te...	Apple Golden 1	[0.0, 0.014880168...	[0.0, 0.0148801682...
s3://pb-ocr-p9/Te...	Onion White	[0.0, 0.036529582...	[0.0, 0.0365295819...
s3://pb-ocr-p9/Te...	Rambutan	[0.28802907, 3.56...	[0.28802907466888...
s3://pb-ocr-p9/Te...	Onion White	[0.0, 0.76288563, ...]	[0.0, 0.7628856301...
s3://pb-ocr-p9/Te...	Pear Red	[0.0052123987, 0...	[0.00521239871159...
s3://pb-ocr-p9/Te...	Pear Forelle	[0.24709865, 0.0...	[0.24709865450859...
s3://pb-ocr-p9/Te...	Pineapple	[0.0, 3.5519376, ...]	[0.0, 3.5519375801...
s3://pb-ocr-p9/Te...	Pear Red	[0.018603958, 0.0...	[0.01860395818948...

only showing top 20 rows

Réduction de dimension ACP

Entrée [\*]:

```
# Nombre de composantes de L'ACP
nbr_composante = 200

# Créer un objet PCA, entraînement et transformation
pca = PCA(
    k=nbr_composante,
    inputCol="features_array",
    outputCol="pca_features"
)
model_fitted = pca.fit(features_df)
pca_result_df = model_fitted.transform(features_df)

# Calculer la variance totale et la variance cumulative expliquée
explained_variance_table = calculer_variance_expliquee(
    model_fitted,
    nbr_composante
)

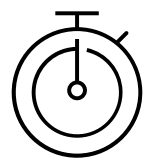
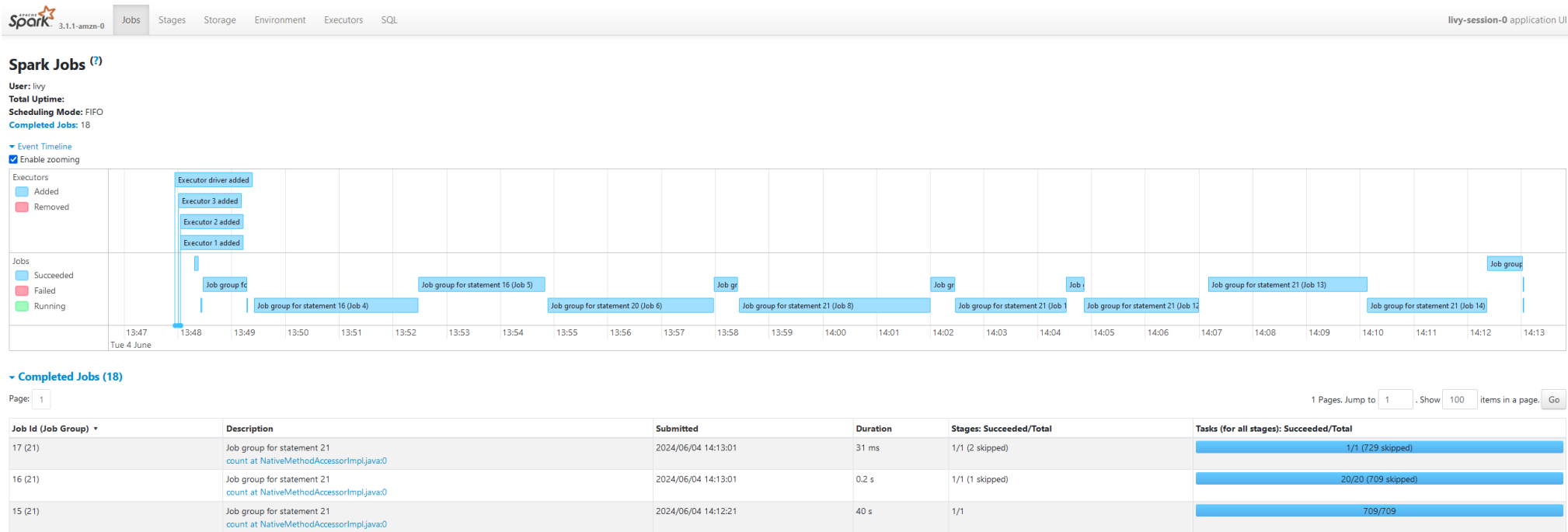
# Enregistrer Les résultats de L'ACP en format parquet
pca_result_df.write.mode('overwrite').parquet(PATH_Result_pca_parquet)
print("\nLes résultats de L'ACP ont été sauvegardés")

# Afficher Les dimensions du dataframe des résultats de L'ACP
print("Number of rows: ", pca_result_df.count())
print("Number of columns: ", len(pca_result_df.columns))
print(pca_result_df.columns)
```

Progress:



# Architecture Big Data



≈26 min



# Architecture Big Data



Amazon S3 > Compartiments > pb-ocr-p9 > Results/

## Results/

Copier l'URI S3

Objets Propriétés

Objets (3) 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](#)

Rechercher des objets en fonction du préfixe

<input type="checkbox"/>	Nom	Type	Dernière modification
<input type="checkbox"/>	features_parquet/	Dossier	-
<input type="checkbox"/>	pca_csv/	Dossier	-
<input type="checkbox"/>	pca_parquet/	Dossier	-

Amazon S3 > Compartiments > pb-ocr-p9 > Results/ > features\_parquet/

## features\_parquet/

Copier l'URI S3

Objets Propriétés

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](#)

Rechercher des objets en fonction du préfixe

<input type="checkbox"/>	Nom	Type	Dernière modification
<input type="checkbox"/>	_SUCCESS	-	04 Jun 2024 03:54:50 PM CEST
<input type="checkbox"/>	part-00000-d82eabe0-d207-46a1-b7b7-7644a612ad5a-c000.snappy.parquet	parquet	04 Jun 2024 03:53:10 PM CEST
<input type="checkbox"/>	part-00001-d82eabe0-d207-46a1-b7b7-7644a612ad5a-c000.snappy.parquet	parquet	04 Jun 2024 03:53:11 PM CEST

Amazon S3 > Compartiments > pb-ocr-p9 > Results/ > pca\_parquet/

## pca\_parquet/

Copier l'URI S3

Objets Propriétés

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](#)

Rechercher des objets en fonction du préfixe

<input type="checkbox"/>	Nom	Type	Dernière modification
<input type="checkbox"/>	_SUCCESS	-	04 Jun 2024 04:12:22 PM CEST
<input type="checkbox"/>	part-00000-8e64afce-f619-47f9-bbf3-f91ea408ea71-c000.snappy.parquet	parquet	04 Jun 2024 04:10:50 PM CEST
<input type="checkbox"/>	part-00001-8e64afce-f619-47f9-bbf3-f91ea408ea71-c000.snappy.parquet	parquet	04 Jun 2024 04:10:51 PM CEST

Amazon S3 > Compartiments > pb-ocr-p9 > Results/ > pca\_csv/

## pca\_csv/

Copier l'URI S3

Objets Propriétés

Objets (1) 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](#)

Rechercher des objets en fonction du préfixe

<input type="checkbox"/>	Nom	Type	Dernière modification
<input type="checkbox"/>	values_pca.csv	csv	04 Jun 2024 04:13:10 PM CEST





# Respect RGPD



```
C:\Users\pierr>aws configure
AWS Access Key ID [*****4LRJ]: *****4LRJ
AWS Secret Access Key [*****Nw7D]: *****Nw7D
Default region name [eu-west-3]: eu-west-3
Default output format [json]: json
```



Amazon EMR > EMR sur EC2: Clusters > cluster\_pb\_p9\_240601

## cluster\_pb\_p9\_240601

Mis à jour il y a 14 minutes [🔄](#) [Réinitialiser](#) [Cloner dans AWS CLI](#) [Cloner](#)

▼ Récapitulatif

### Informations sur le cluster

ID de cluster  
j-3KE409HHSXL4T

Configuration de cluster  
Groupes d'instances

Capacité  
1 primaire(s) | 1 unité(s) principale(s) | 2 tâche(s)

### Applications

Version d'Amazon EMR  
emr-6.3.0

Applications installées  
JupyterHub 1.2.2, Spark 3.1.1, TensorFlow 2.4.1

### Gestion des clusters

Destination des journaux dans Amazon S3  
pb-ocr-p9/elasticmapreduce\_logs

Interfaces utilisateur d'application persistantes  
[Serveur d'historique Spark](#)

[Serveur de chronologie YARN](#)

DNS public du nœud primaire  
[ec2-13-39-148-14.eu-west-3.compute.amazonaws.com](#)

[Connexion au nœud primaire à l'aide de SSM](#)

### Statut et heure

Statut  
En attente

Heure de création  
4 juin 2024 13:57 (UTC+02:00)

Temps écoulé  
29 minutes, 31 secondes



AWS Services | Rechercher [Alt+S]

## EC2 Global View

Explorateur de régions | Recherche globale | Mis à jour il y a 1 minute

Explorateur de régions  
[Recherche globale](#)  
Paramètres Nouveau

### Recherche globale (4)

Effectuer une recherche globale pour rechercher des ressources spécifiques dans toutes les régions pour lesquelles votre compte est activé

Find resources by attribute or tag

Type de ressource = Instance [×](#) [Clear filters](#)

	Name	ID de ressource	Type de resso...	Région
<input type="radio"/>	-	i-013d046838665ddf8	Instance	eu-west-3
<input type="radio"/>	-	i-07812b34c93eca51a	Instance	eu-west-3
<input type="radio"/>	-	i-0ed5dc1f3d2c587da	Instance	eu-west-3
<input type="radio"/>	-	i-022f283a80ae3de86	Instance	eu-west-3



AWS Services | Rechercher [Alt+S]

## Amazon S3 > Compartiments

Aperçu du compte : mis à jour toutes les 24 heures [Toutes les régions AWS](#) [Afficher le tableau de bord de Storage Lens](#)

Compartiments à usage général | Compartiments de répertoires

Compartiments à usage général (1) [Toutes les régions AWS](#) [Copier l'ARN](#) [Vider](#) [Supprimer](#) [Créer un compartiment](#)

Les compartiments sont des conteneurs pour les données stockées dans S3.

Rechercher des compartiments par nom

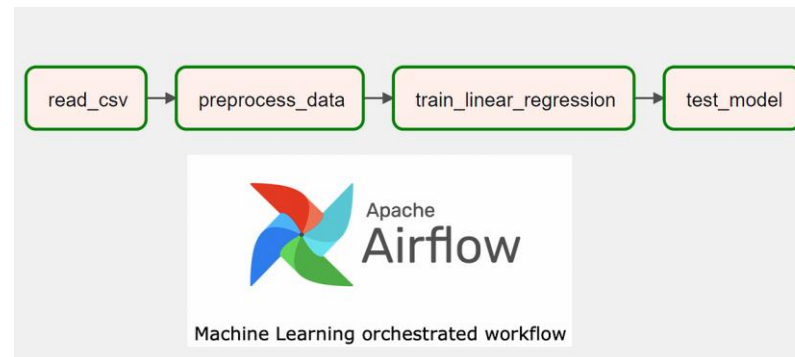
	Nom	Région AWS	Anal.
<input type="radio"/>	pb-ocr-p9	Europe (Paris) eu-west-3	<a href="#">Afficher</a>



# Retour critique

=> **Optimisation de l'utilisation des clusters pour optimiser les coûts :**








- Alertes de facturation (surtout si cluster adaptatif)
- Configuration de Spark
- Librairie Spark EMRFS optimisée pour AWS
- Mise en place de workflow automatique avec Airflow

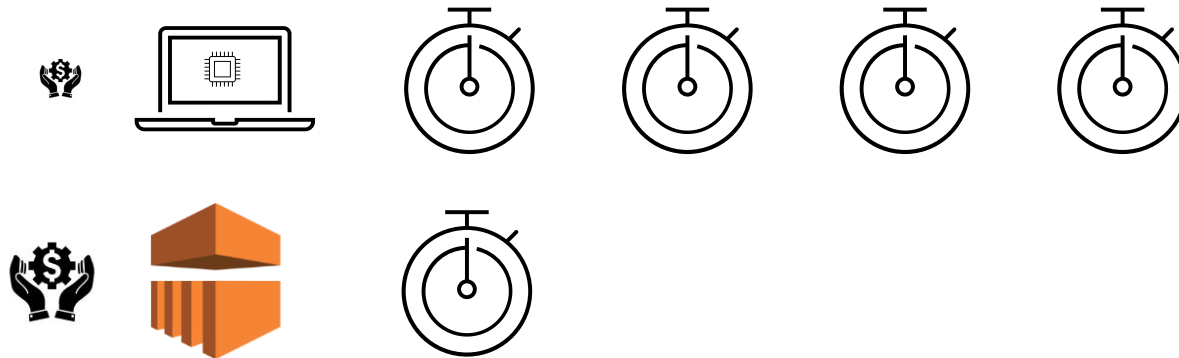
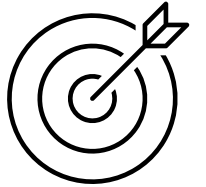




# Conclusion

## Missions :

1. Expliquer la chaîne de traitement des données ✓ 
2. Détailler l'architecture Big Data retenue ✓     
3. Veiller au respect des contraintes RGPD (serveur dans l'UE) ✓ 
4. Apporter un retour critique sur la solution proposée ✓





OPENCLASSROOMS

Merci pour votre attention



CentraleSupélec

Pierrick BERTHE

Formation Expert en Data Science  
*Openclassrooms – CentraleSupélec*

août 2023 → juin 2024