



Réalisez un traitement dans un environnement Big Data sur le Cloud

Réaliser par:Kaoutar EL Mardi

jury:Emmanuel Goudot



Table of contents

01

**Problématique et jeu
de données**

02

**Processus de création
de l'environnement
Big Data**

03

**Chaîne de traitement
des images**

04

**Exécution du script
PySpark sur le Cloud**

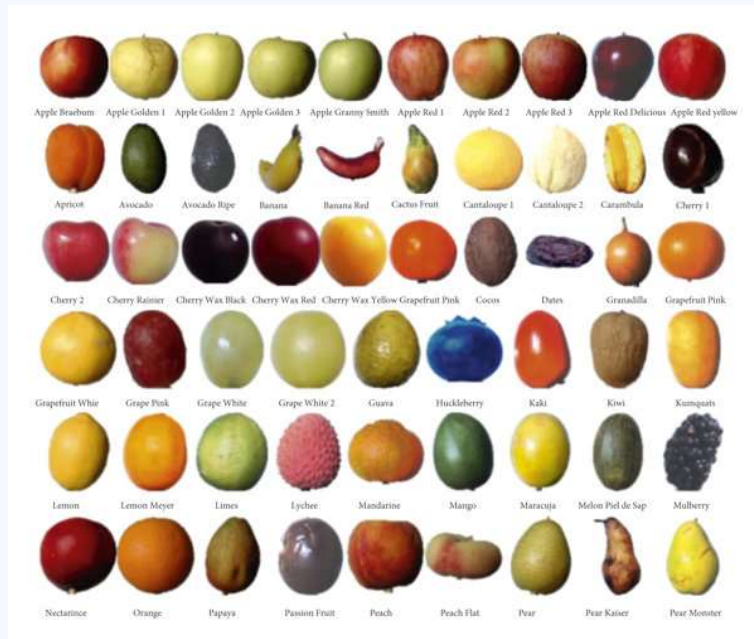
01 Problématique et données

The background of the slide is a light blue-grey color. It features decorative elements: a grid of small blue dots in the upper right and lower left corners, and stylized circuit board traces in purple, orange, and grey that meander across the slide. A horizontal line underlines the text, ending in a small circle.

Problématique

Fruits!" développe des robots cueilleurs intelligents pour préserver la biodiversité. Pour sensibiliser le public, l'entreprise lance une **application mobile d'identification des fruits**. Cette application servira à construire **une architecture Big Data** évolutive, respectant le RGPD et optimisant les coûts.

Jeu de données



Caractéristiques principales :

- **Images** : Plus de 70 000 images de fruits.
- **Catégories** : Représente plus de 100 catégories de fruits et légumes.
- **Diversité** : Images prises sous différents angles et conditions d'éclairage.
- **Étiquetage** : Chaque image est soigneusement étiquetée, facilitant la reconnaissance précise des fruits.

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

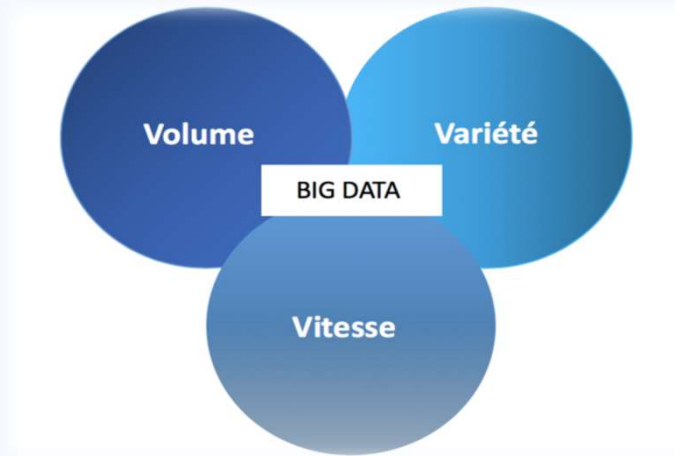


02

Processus de Création de de l'environnement Big Data

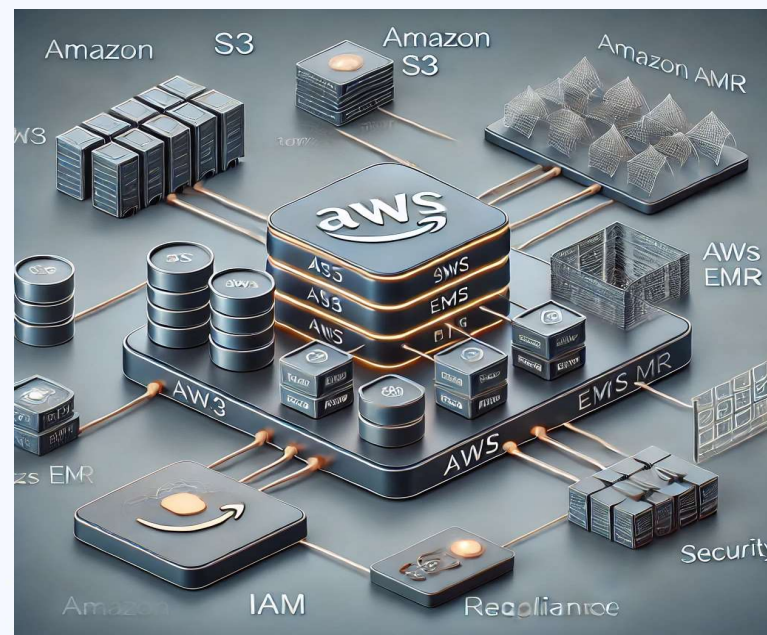
Pourquoi Environnement Big Data

L'environnement Big Data est indispensable pour gérer la croissance rapide des données de l'application "Fruits!". Les trois dimensions du Big Data — Volume, Variété, et Vitesse — nécessitent une architecture robuste

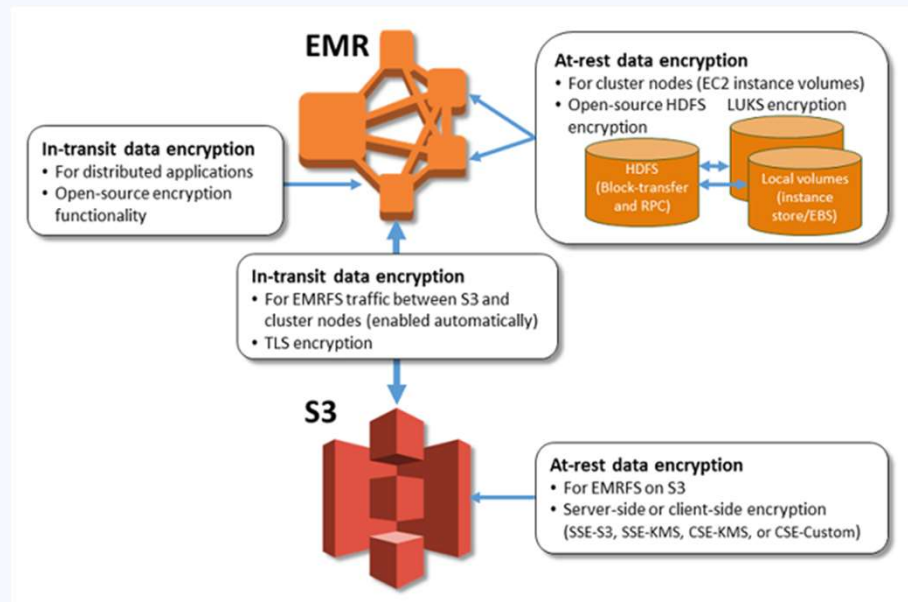


Outils et déploiement

Amazon AWS offre les outils nécessaires pour traiter ces données efficacement tout en assurant la conformité avec le RGPD.



processus de création de l'environnement Big Data



AWS EMR

- Scalabilité automatique
 - Intégration facile
 - Gestion simplifiée

Amazon S3

- Durabilité et disponibilité
 - Tarification flexible
 - Conformité RGPD

AWS IAM

- Sécurité renforcée
- Contrôle granulaire
- Conformité RGPD

"Configuration IAM : Sécurisation des ressources AWS avec l'authentification multifactorielle (MFA) et la création de rôles pour un contrôle d'accès précis, en conformité avec le RGPD."

Identity and Access Management (IAM)

Rechercher sur IAM

Tableau de bord

Gestion des accès

Groupes d'utilisateurs

Utilisateurs

Rôles

Politiques

Fournisseurs d'identité

Paramètres du compte

Rapports d'accès

Analyseur d'accès

Accès externe

Accès non utilisé

Paramètres de l'analyseur

Rapport sur les informations d'identification

Activité de l'organisation

Politiques de contrôle des services

Consoles connexes

IAM Identity Center

AWS Organizations

IAM > Tableau de bord

Tableau de bord IAM

Recommandations de sécurité 2

Ajouter la MFA pour l'utilisateur racine

Ajouter la MFA pour l'utilisateur root – Activez l'Authentification multifactorielle (MFA) pour l'utilisateur root afin d'améliorer la sécurité de ce compte.

Ajouter la MFA

Désactiver ou supprimer les clés d'accès pour l'utilisateur racine

Désactivez ou supprimez les clés d'accès de l'utilisateur racine et utilisez plutôt les clés d'accès attachées à un utilisateur IAM pour renforcer la sécurité.

Gérer les clés d'accès

Ressources IAM

Ressources de ce compte AWS

Groupes d'utilisateurs	Utilisateurs	Rôles	Politiques	Fournisseurs d'identité
1	1	5	1	0

Nouveautés

Mises à jour des fonctions dans IAM

- Rôles Anywhere IAM prend désormais en charge la modification du mappage des attributs des certificats. Il y a 4 mois
- IAM Roles Anywhere propose désormais des informations d'identification valides jusqu'à 12 heures. Il y a 5 mois
- Amazon EKS introduit des contrôles simplifiés pour la gestion de l'accès aux clusters IAM. Il y a 8 mois
- IAM Access Analyzer simplifie désormais l'inspection des accès non utilisés pour vous guider dans l'approche du moindre privilège. Il y a 9 mois

plus

Tout afficher

Configuration S3 : Création de compartiments sécurisés dans les régions européennes pour le chargement et le stockage des données, assurant la conformité RGPD, avec un contrôle d'accès géré via IAM pour protéger les données sensibles.

Amazon S3

Compartiments

Access Grants

Points d'accès

Points d'accès de l'objet Lambda

Points d'accès multi-région

Opérations par lot

IAM Access Analyzer pour S3

Paramètres de blocage de l'accès public pour ce compte

Storage Lens

Tableaux de bord

Groupe Storage Lens

Paramètres AWS Organizations

Fonctionnalité spot

AWS Marketplace pour S3

Amazon S3

► 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 répertoire

Compartiments à usage général (4)

Info

Toutes les régions AWS

↺

Copier l'ARN

Vider

Supprimer

Créer un compartiment

Rechercher des compartiments par nom

< 1 > {

	Nom	Région AWS	Analyseur d'accès IAM	Date de création
<input type="radio"/>	aws-logs-025066242574-eu-north-1	Europe (Stockholm) eu-north-1	Afficher l'analyseur pour eu-north-1	07 Aug 2024 04:25:47 PM CEST
<input type="radio"/>	aws-logs-025066242574-eu-west-1	Europe (Irlande) eu-west-1	Afficher l'analyseur pour eu-west-1	07 Aug 2024 09:40:41 PM CEST
<input type="radio"/>	my-projet9-open	Europe (Irlande) eu-west-1	Afficher l'analyseur pour eu-west-1	07 Aug 2024 12:28:30 PM CEST
<input type="radio"/>	oc-cal	Europe (Irlande) eu-west-1	Afficher l'analyseur pour eu-west-1	06 Aug 2024 05:12:32 PM CEST

Creation de EMR

▼ Nom et applications - *requis* [Info](#)

Donnez un nom à votre cluster et choisissez les applications que vous voulez y installer.

Nom

Moncluster6

Version Amazon EMR [Info](#)

Une version contient un ensemble d'applications susceptibles d'être installées sur votre cluster.

emr-7.0.0

Offre d'applications



- | | | |
|---|---|--|
| <input type="checkbox"/> AmazonCloudWatchAgent 1.300031.1 | <input type="checkbox"/> Flink 1.18.0 | <input type="checkbox"/> HBase 2.4.17 |
| <input type="checkbox"/> HCatalog 3.1.3 | <input checked="" type="checkbox"/> Hadoop 3.3.6 | <input type="checkbox"/> Hive 3.1.3 |
| <input type="checkbox"/> Hue 4.11.0 | <input type="checkbox"/> JupyterEnterpriseGateway 2.6.0 | <input checked="" type="checkbox"/> JupyterHub 1.5.0 |
| <input type="checkbox"/> Livy 0.7.1 | <input type="checkbox"/> MXNet 1.9.1 | <input type="checkbox"/> Oozie 5.2.1 |
| <input type="checkbox"/> Phoenix 5.1.3 | <input type="checkbox"/> Pig 0.17.0 | <input type="checkbox"/> Presto 0.283 |
| <input checked="" type="checkbox"/> Spark 3.5.0 | <input type="checkbox"/> Sqoop 1.4.7 | <input type="checkbox"/> TensorFlow 2.11.0 |
| <input type="checkbox"/> Tez 0.10.2 | <input type="checkbox"/> Trino 426 | <input type="checkbox"/> Zeppelin 0.10.1 |
| <input type="checkbox"/> ZooKeeper 3.5.10 | | |

Paramètres du catalogue de données AWS Glue

Utilisez le catalogue de données AWS Glue pour fournir un metastore externe à votre application.

☐ Utiliser pour les métadonnées de table Spark

Options du système d'exploitation [Info](#)

- ☒ Version Amazon Linux :
- ☐ Amazon Machine Image (AMI) personnalisée
- ☒ Appliquez automatiquement les dernières mises à jour Amazon Linux

Cluster EMR
Version : emr-7.0.0

Applications sélectionnées :

Spark : Pour le traitement des données.

Hadoop : Pour la gestion des données massives.

JupyterHub : Pour l'analyse interactive et le développement.

▼ Configuration de cluster - *requis* [Info](#)

Choisissez une méthode de configuration pour les groupes de nœuds primaires, principaux et de tâches de votre cluster.

☒ Groupes d'instances uniformes

Choisissez le même type d'instance EC2 et la même option d'achat (à la demande ou Spot) pour tous les nœuds de votre groupe de nœuds. [En savoir plus](#)

☐ Flottes d'instances flexibles

Choisissez parmi la plus grande variété d'options de provisionnement pour les instances EC2 de votre cluster. Diversifiez les types d'instances et les options d'achat, et utilisez une stratégie d'allocation. [En savoir plus](#)

Groupes d'instances uniformes

Primaire

Choisir un type d'instance EC2

m5.xlarge
4 vCore 16 GiB mémoire EBS uniquement stockage
Prix à la demande : 0.214 USD par instance/heure
Prix Spot le plus bas : 0.091 USD (eu-west-1b)

Actions ▼

☐ Utiliser la haute disponibilité

Lancez des clusters hautement disponibles et plus résilients avec trois nœuds primaires sur des instances à la demande. Cette configuration s'applique pendant toute la durée de vie de votre cluster. [En savoir plus](#)

► Configuration de nœud - *facultatif*

Unité principale

Choisir un type d'instance EC2

m5.xlarge
4 vCore 16 GiB mémoire EBS uniquement stockage
Prix à la demande : 0.214 USD par instance/heure
Prix Spot le plus bas : 0.091 USD (eu-west-1b)

Actions ▼

► Configuration de nœud - *facultatif*

Tâche 1 sur 1

Nom

Tâche - 1

Retirer le groupe d'instances

Choisir un type d'instance EC2

m5.xlarge
4 vCore 16 GiB mémoire EBS uniquement stockage
Prix à la demande : 0.214 USD par instance/heure
Prix Spot le plus bas : 0.091 USD (eu-west-1b)

Actions ▼

► Configuration de nœud - *facultatif*

Ajouter un groupe d'instances de tâches

Vous pouvez ajouter jusqu'à 47 autres groupes d'instances de tâches.

Volume racine EBS

Le volume racine EBS s'applique aux systèmes d'exploitation et aux applications que vous installez sur le cluster. [Contraintes relatives au rapport de volume racine EBS](#)

Taille (Gio)

15

15 - 100 GiB par volume
Volume SSD polyvalent (gp3)

IOPS

3000

3000- 16000 IOPS par volume.
Choisissez un rapport maximum de 500:1 entre les IOPS et la taille du volume.

Débit (Mio/s)

125

125- 1000 MiB/s par volume.
Choisissez un rapport maximum de 0.25:1 entre le débit et les IOPS.

- Type d'instances** : m5.xlarge (4 vCPU, 16 Go de RAM) pour les nœuds principaux et les tâches.
- Volume racine EBS** : 15 Go de stockage avec 3000 IOPS pour une performance optimale.
- Débit** : 125 MiB/s pour assurer une vitesse de transfert de données adéquate.

▼ **Actions d'amorçage (1)** [Info](#) Supprimer Modifier Ajouter

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

	Nom ▼	Emplacement Amazon S3 ↗ ▼	Arguments
<input type="radio"/>	bib	s3://my-projet9-open/bootstrap-emr.sh	-

Mise à jour : Le script commence par mettre à jour setuptools et pip.

Installation des bibliothèques :
Installe les bibliothèques nécessaires comme pandas, boto3, et tensorflow.

Configuration de Spark :
Configure Spark History Server pour écrire les logs dans un compartiment S3 spécifique.

```
#!/bin/bash
# Mise à jour de setuptools et pip
sudo python3 -m pip install -U setuptools
sudo python3 -m pip install -U pip

# Installation des bibliothèques nécessaires
sudo python3 -m pip install wheel
sudo python3 -m pip install pillow
sudo python3 -m pip install pandas==1.2.5
sudo python3 -m pip install pyarrow
sudo python3 -m pip install boto3
sudo python3 -m pip install s3fs
sudo python3 -m pip install fsspec
sudo python3 -m pip install tensorflow
```

```
# Configuration de Spark History Server pour écrire les logs dans S3
sudo tee -a /etc/spark/conf/spark-defaults.conf <<EOF
spark.history.fs.logDirectory=s3://my-projet9-open/spark-logs/
EOF
```


[illegible]

03

Chaîne de traitement des images



Shéma

Chargement
des données

Préparation
de modèle

Extraction
des features
et PCA

Stockage

6.10.5.1. 4.10.5.1 Chargement des données

```
j> images = spark.read.format("binaryFile") \
  .option("pathGlobFilter", "*.jpg") \
  .option("recursiveFileLookup", "true") \
  .load(PATH_Data)

FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),...)

j> images.show(5)

FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),...)
+-----+-----+-----+-----+
| path | modificationTime | length | content |
+-----+-----+-----+-----+
|s3://p8-data/Test...|2021-07-03 09:00:08| 7353|[FF D8 FF E0 00 1...|
|s3://p8-data/Test...|2021-07-03 09:00:08| 7350|[FF D8 FF E0 00 1...|
|s3://p8-data/Test...|2021-07-03 09:00:08| 7349|[FF D8 FF E0 00 1...|
|s3://p8-data/Test...|2021-07-03 09:00:08| 7348|[FF D8 FF E0 00 1...|
|s3://p8-data/Test...|2021-07-03 09:00:09| 7328|[FF D8 FF E0 00 1...|
+-----+-----+-----+-----+
only showing top 5 rows
```

une colonne contenant les labels de chaque image :

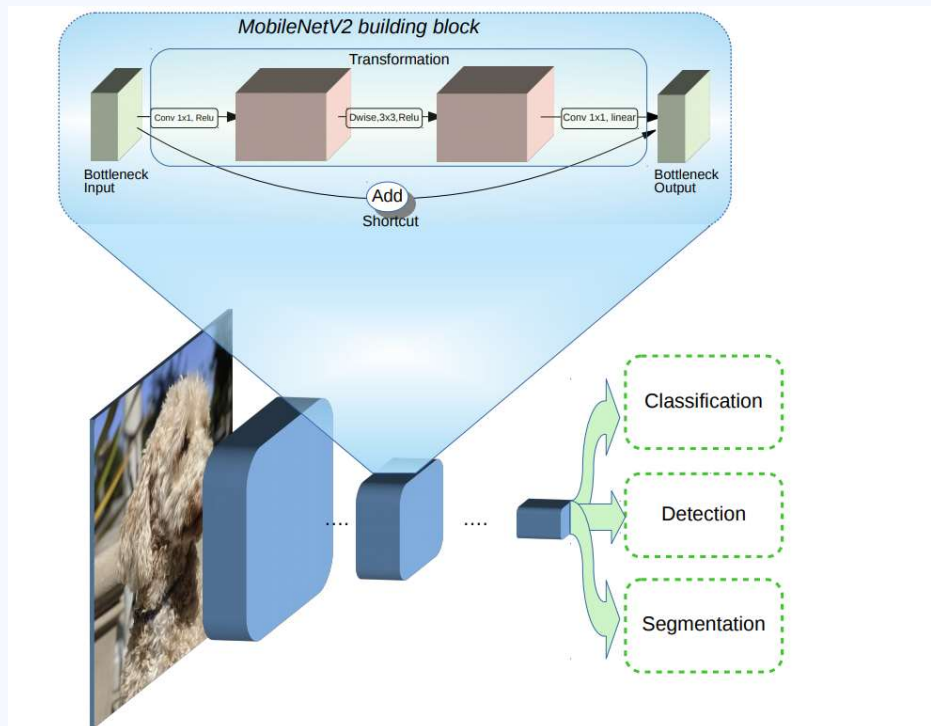
```
[7]: images = images.withColumn('label', element_at(split(images['path'], '/'), -2))
print(images.printSchema())
print(images.select('path', 'label').show(5, False))

FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),...)
root
 |-- path: string (nullable = true)
 |-- modificationTime: timestamp (nullable = true)
 |-- length: long (nullable = true)
 |-- content: binary (nullable = true)
 |-- label: string (nullable = true)

None
+-----+-----+
| path | label |
+-----+-----+
|s3://p8-data/Test/Watermelon/r_106_100.jpg|Watermelon|
|s3://p8-data/Test/Watermelon/r_109_100.jpg|Watermelon|
|s3://p8-data/Test/Watermelon/r_108_100.jpg|Watermelon|
|s3://p8-data/Test/Watermelon/r_107_100.jpg|Watermelon|
|s3://p8-data/Test/Watermelon/r_95_100.jpg|Watermelon|
+-----+-----+
only showing top 5 rows

None
```

MobileNetV2



MobileNetV2 est une architecture de réseau neuronal optimisée pour les dispositifs mobiles. L'image montre un bloc clé qui utilise des convolutions et un chemin de raccourci pour capturer efficacement les caractéristiques. Ce design permet de réaliser des tâches comme la classification, la détection, et la segmentation d'images tout en restant léger et performant.

Traiter des images avec Pandas UDFs et PySpark, en trois étapes :

- **Prétraitement** : `preprocess(content)`: Charge et redimensionne les images à 224x224 pixels.
- **Optimisation avec UDF** : `featurize_udf(content_series_iter)`: Applique l'extraction sur des lots d'images avec un modèle chargé une seule fois.
- **Extraction des caractéristiques** : `featurize_series(model, content_series)`: Utilise un modèle pour extraire et aplatir les caractéristiques des images.

l'extraction et la transformation des caractéristiques (features) d'images, suivi d'une réduction de dimension avec PCA :

- **Configuration Spark** : Ajustement des paramètres Spark pour optimiser le traitement des lots d'images.
- **Extraction des caractéristiques** : Les caractéristiques des images sont extraites avec une UDF et enregistrées dans un DataFrame.
- **Enregistrement** : Les caractéristiques extraites sont enregistrées au format Parquet.

- **Conversion en vecteurs denses** : Les caractéristiques sont converties en vecteurs denses pour une meilleure compatibilité avec les algorithmes de machine learning.
- **Application du PCA** : Le PCA est appliqué pour réduire la dimensionnalité des vecteurs de caractéristiques, ce qui facilite le stockage et l'analyse ultérieure.
- **Enregistrement des résultats** : Les résultats du PCA sont sauvegardés dans un fichier Parquet.

Les données extraites sont chargées à partir de fichiers Parquet à l'aide de `spark.read.parquet`.

- **Chargement des données** : Les données réduites par PCA sont également chargées à partir d'un chemin spécifié.
- **Validation des résultats** : Affichage des premières lignes du DataFrame pour vérifier le contenu et la structure des données.
Validation de la dimension des caractéristiques après réduction par PCA.

Pyspark

Spark Jobs (?)

User: livy

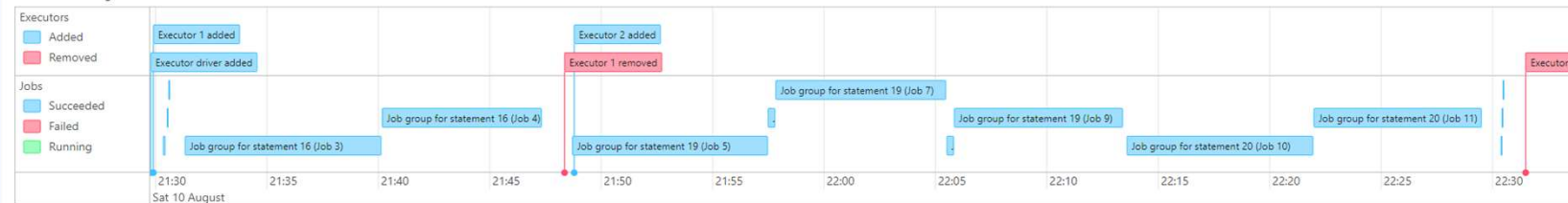
Total Uptime:

Scheduling Mode: FIFO

Completed Jobs: 15

▼ Event Timeline

☐ Enable zooming



▼ Completed Jobs (15)

Page: 1

1 Pages. Jump to 1. Show 100 items in a page. Go

Job Id (Job Group) ▼	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
14 (25)	Job group for statement 25 head at <stdin>:1	2024/08/10 22:30:27	0.4 s	1/1	1/1
13 (24)	Job group for statement 24 parquet at NativeMethodAccessorImpl.java:0	2024/08/10 22:30:24	88 ms	1/1	1/1
12 (23)	Job group for statement 23 parquet at NativeMethodAccessorImpl.java:0	2024/08/10 22:30:21	0.2 s	1/1	1/1
11 (20)	Job group for statement 20 parquet at NativeMethodAccessorImpl.java:0	2024/08/10 22:21:58	7.5 min	1/1 (1 skipped)	24/24 (709 skipped)
10 (20)	Job group for statement 20 parquet at NativeMethodAccessorImpl.java:0	2024/08/10 22:13:34	8.4 min	1/1	709/709
9 (19)	Job group for statement 19	2024/08/10 22:05:49	7.6 min	2/2 (1 skipped)	28/28 (709 skipped)

04

Démonstration d'exécution dans le cloud



jupyterhub

P9_Notebook_EMR_PySpark_V1.0

Dernière Sauvegarde : il y a une minute (modifié)

LogoutControl Panel

FileEditViewInsertCellKernelWidgetsHelp

Non tablePySpark

Exécuter

Code

4.10.1 Démarrage de la session Spark

Entrée [1]:

L'exécution de cette cellule démarre l'application Spark
from pyspark.sql import SparkSession
spark = SparkSession.builder.appname('P9').getOrCreate()

Starting Spark application

ID	YARN Application ID	Kind	State	Spark UI	Driver log	Current session?
0	application_1628050279029_0001	pyspark	idle	Lien	Lien	✓

FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'))...
SparkSession available as 'spark'.
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'))...

Affichage des informations sur la session en cours et liens vers Spark UI :

Entrée [2]:

%xInfo

Current session configs: {'driverMemory': '1000M', 'executorCores': 2, 'proxyUser': 'jovyan', 'kind': 'pyspark'}

ID	YARN Application ID	Kind	State	Spark UI	Driver log	Current session?
0	application_1628050279029_0001	pyspark	idle	Lien	Lien	✓

Informations sur le cluster

ID de cluster

j-XPKHHRHA2YAEF

Configuration de cluster

Groupes d'instances

Capacité

1 primaire(s) 1 unité(s) principale(s) 1 tâche(s)

Applications

Version d'Amazon EMR

emr-7.0.0

Applications installées

Hadoop 3.3.6, JupyterHub 1.5.0, Spark 3.5.0

Gestion des clusters

Destination des journaux dans Amazon S3

[aws-logs-025066242574-eu-west-1/elasticmapreduce](#)

Interfaces utilisateur d'application persistantes

Serveur d'historique Spark

[Lien](#)

Serveur de chronologie YARN

[Lien](#)

DNS public du nœud primaire

[ec2-52-211-178-211.eu-west-1.compute.amazonaws.com](#)

Connexion au nœud primaire à l'aide de SSH

[Lien](#)

Connexion au nœud primaire à l'aide de SSH

[Lien](#)

Statut et heure

Statut

En attente

Heure de création

11 août 2024 05:49 (UTC+02:00)

Temps écoulé

10 minutes, 21 secondes

Propriétés

Actions d'amorçage

Instances (Matériel)

Étapes

Applications

Configurations

Surveillance

Évènements

Identifications (0)

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

[Lien](#)

Interface utilisateur du serveur d'historique Spark

[Lien](#)

Interfaces utilisateur d'application sur le nœud primaire

Celles-ci nécessitent l'activation du tunneling SSH.

Application

Gestionnaire de ressources

JupyterHub

Nom du nœud HDFS

Serveur d'historique Spark

URL de l'interface utilisateur

[http://ec2-52-211-178-211.eu-west-1.compute.amazonaws.com:8088/](#)

[https://ec2-52-211-178-211.eu-west-1.compute.amazonaws.com:9443/](#)

[http://ec2-52-211-178-211.eu-west-1.compute.amazonaws.com:9870/](#)

[http://ec2-52-211-178-211.eu-west-1.compute.amazonaws.com:18080/](#)

Interface utilisateur d'apolication sur les nœuds princiaux et nœuds de tâches



Conclusion



Thanks!

Do you have any questions?

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**
