

Déployez un modèle dans le Cloud



Sommaire

- Rappel problématique et présentation jeu de données
- Présentation des services Big Data
 - Présentation services AWS (S3, IAM, EC2, EMR)
 - Présentation de Hadoop et Spark
- Présentation de la chaîne de traitement
 - Description des étapes de connexion et création EMR
 - Présentation des étapes du script PySpark
- Démonstration de description du script PySpark
- Synthèse et conclusion



Problématique et Présentation du jeu de données

- Problématique

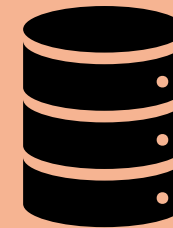
Une application mobile qui fournit des informations sur le fruit pris en photo.

Mettre en place les premières briques de traitement pour un passage à l'échelle en termes de volumes de données.

Contraintes:

- Diffusion des poids du modèle Tensorflow
- Réduction de dimensions de type PCA en PySpark
- Respect des contraintes RGPD

- Jeu de données



- 131 fruits différents
- Entre 297 et 984 fichiers par fruits
(~68 000 fichiers au total)

Présentation des services Big Data



Présentation services AWS

Amazon S3



Service de stockage en ligne

- Stocker et récupérer des données depuis n'importe où sur le Web.
- Chiffrement

Amazon EC2



Service de calcul sur Cloud

- Calcul distribué via instance de MV
- Tarification à la demande
- Sécurité renforcée et flexibilité

- AWS IAM



Service de gestion des utilisateurs et utilisateurs

- Gérer les autorisations
- Gérer les groupes

Amazon EMR



Service traitement de données massives

- Plusieurs frameworks (Hadoop, Spark,...)
- Infrastructure évolutive



Présentation Hadoop et Spark



Une plateforme open-source qui fournit des outils pour stocker et traiter des grands volumes de données en utilisant un cluster de serveurs.

Traitement en lots

- HDFS (hadoop Distributed File System) pour le stockage de données
- MapReduce pour traiter les données de manière parallèle



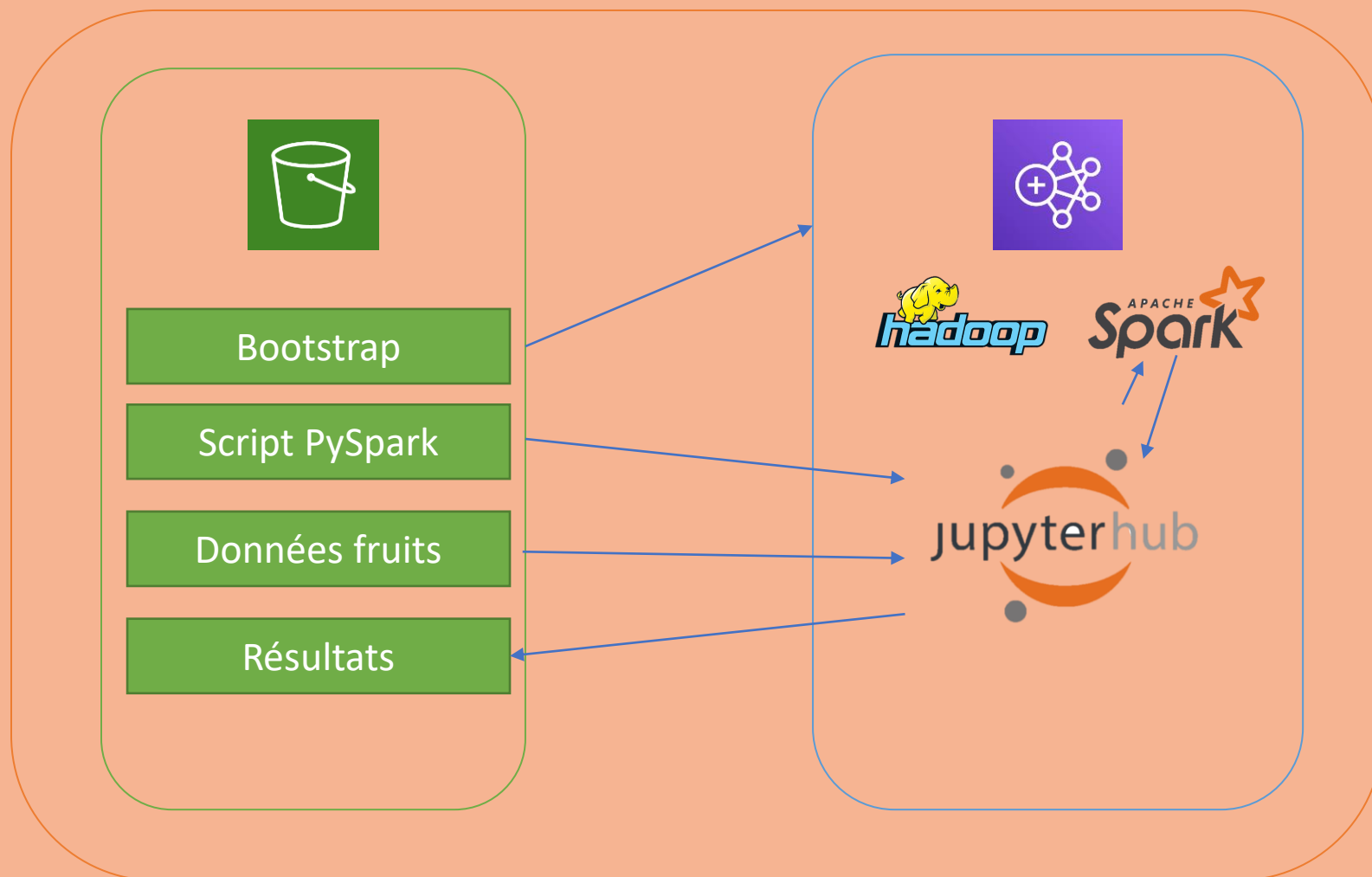
Une plateforme open-source qui permet de traiter des données massives en temps réel.

- Utilise la mémoire vive (RAM)
- Peut gérer le stream processing et le machine learning



Processus de création et connexion EMR





Etapes création cluster EMR

Pré-requis:

- Création d'une paire de clé pour une connexion SSH
- Création d'un fichier d'amorçage

Paire de clés

Une paire de clés, composée d'une clé privée et d'une clé publique, est un ensemble d'informations d'identification de sécurité que vous utilisez pour prouver votre identité lors de la connexion à une instance.

Nom

Windows_key

Le nom peut avoir un maximum de 255 caractères ASCII. Il ne peut pas inclure d'espaces avant ou après.

Type de paire de clés [Informations](#)

☒ RSA

☐ ED25519

Format de fichier de clé privée

☐ .pem
À utiliser avec OpenSSH

☒ .ppk
À utiliser avec PuTTY

Balises - *facultatif*

Aucune balise n'est associée à cette ressource.

[Ajouter une balise](#)

Vous pouvez ajouter jusqu'à 50 identifications supplémentaires.

[Annuler](#) [Créer une paire de clés](#)

```
#!/bin/bash
sudo python3 -m pip install -U pip
sudo python3 -m pip install wheel
sudo python3 -m pip install pillow
sudo python3 -m pip install pandas
sudo python3 -m pip install -U scikit-learn
sudo python3 -m pip install matplotlib
sudo python3 -m pip install pyspark
sudo python3 -m pip install pyarrow
sudo python3 -m pip install boto3
sudo python3 -m pip install s3fs
sudo python3 -m pip install fsspec
```



- Définition du tunnel SSH du driver

Nom du groupe de sécurité
ElasticMapReduce-master

ID du groupe de sécurité
sg-028dfbdc97c96bce7

Description
Master group for Elastic MapReduce created on 2023-04-14T17:56:25.714Z

ID de VPC
vpc-01b79ac521c030633

Propriétaire
409889106312

Nombre de règles entrantes
9 Entrées d'autorisation

Nombre de règles sortantes
1 Entrée d'autorisation

Règles entrantes

Règles sortantes

Balises

Règles entrantes (9)

Gérer les balises

Modifier les règles entrantes

Filtrer les règles des groupes de sécurité

< 1 >

<input type="checkbox"/>	Name	ID de règle de grou...	Version IP	Type	Protocole	Plage de ports
<input type="checkbox"/>	-	sgr-0cde54c49738659c8	-	Tous les UDP	UDP	0 - 65535
<input type="checkbox"/>	-	sgr-0c664ee16fed8f44	-	Tous les TCP	TCP	0 - 65535
<input type="checkbox"/>	-	sgr-0e12e9f30c50c4e41	IPv6	SSH	TCP	22
<input type="checkbox"/>	-	sgr-0c52dc664720581...	-	Tous les ICMP - IPv4	ICMP	Tous
<input type="checkbox"/>	-	sgr-0a6ed51e7ccb859b7	IPv4	SSH	TCP	22

- Création du cluster













▼ Personnaliser votre offre d'applications

Applications incluses dans l'offre

☐ Flink 1.16.0

☐ HBase 2.4.15

☒ Hadoop 3.3.3

☐ Hue 4.10.0

☒ JupyterHub 1.5.0

☐ MXNet 1.9.1

☐ Phoenix 5.1.2

☐ Presto 0.278

☐ Sqoop 1.4.7

☐ Tez 0.10.2

☐ Zeppelin 0.10.1

☐ Ganglia 3.7.2

☐ HCatalog 3.1.3

☐ Hive 3.1.3

☐ JupyterEnterpriseGateway 2.6.0

☐ Livy 0.7.1

☐ Oozie 5.2.1

☐ Pig 0.17.0

☒ Spark 3.3.1

☒ TensorFlow 2.11.0

☐ Trino 403

☐ ZooKeeper 3.5.10

▼ 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	Arguments
<input type="radio"/>	Bootstrap	s3://p8datalubinstephane/bootstrap-emr.sh	-

Configuration de sécurité et autorisations [Info](#)

Configuration de sécurité - facultatif

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

Choisir une configuration de ...

Parcourir

Créer une configuration de sécurité

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

Regis-P8

Parcourir

Créer une paire de clés

r5.2xlarge

8 vCore 64 GiB mémoire

EBS uniquement stockage

Prix à la demande : 0.592 USD par instance/heure

Prix Spot le plus bas : \$0.142 (eu-west-3c)

Statut

🔄 Démarrage en cours

Statut

🔄 Action d'amorçage

Statut

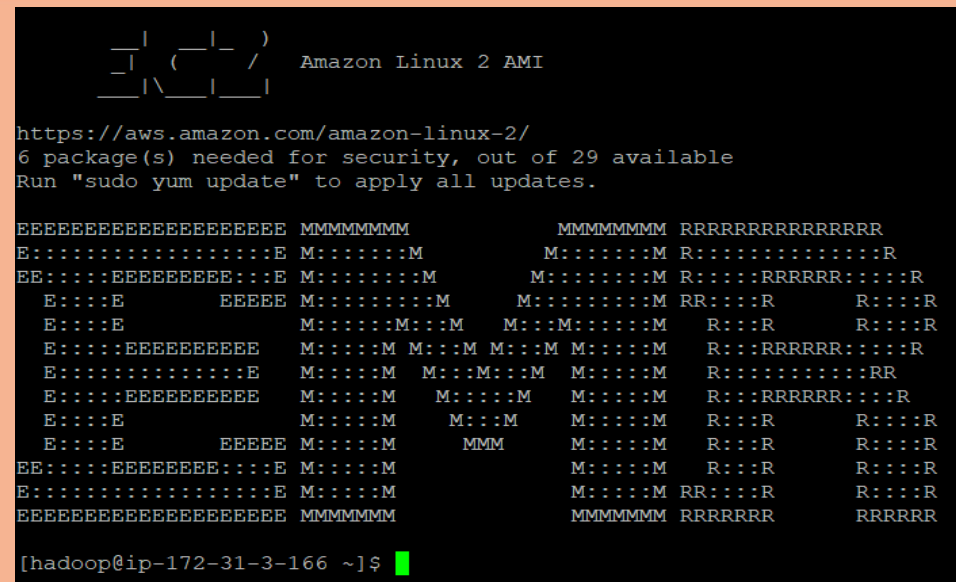
✅ En attente

P8 - Déployer un modèle dans le cloud - Stéphane LUBIN

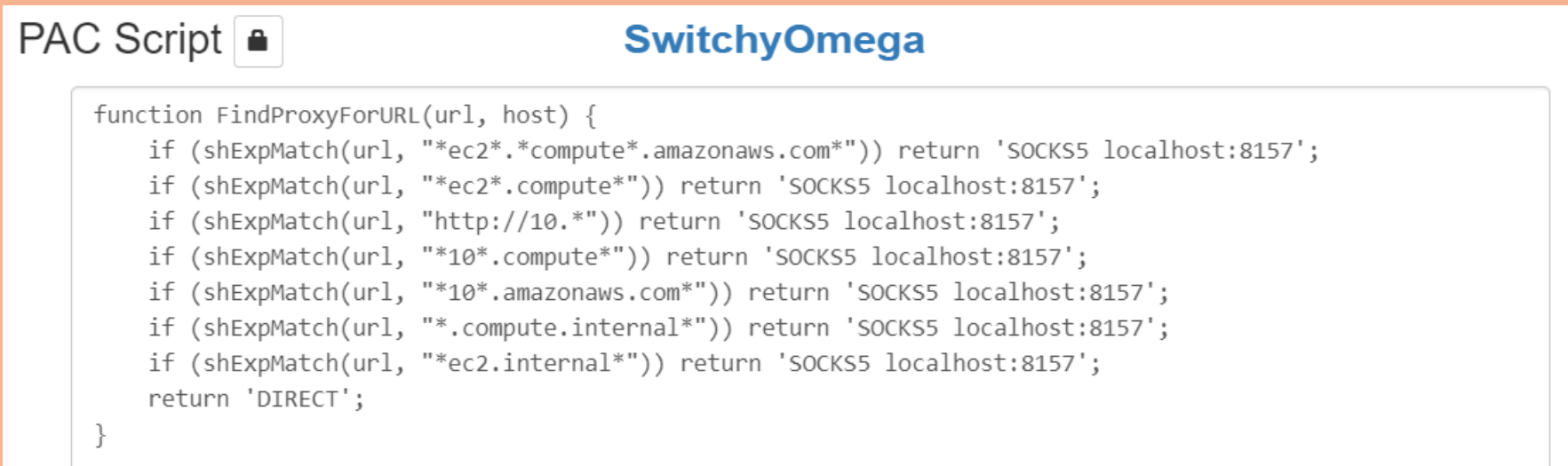
Fruits!

- Connexion SSH





- Connexion via PuTTY
- Les clé de connexion
- Les ports dynamiques utilisées

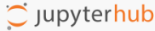


- **Proxy Navigateur (Google Chrome)**



• Connexion Jupyter Hub et Serveur Historique Spark

Application	URL de l'interface utilisateur 
Gestionnaire de ressources	 http://ec2-13-38-83-117.eu-west-3.compute.amazonaws.com:8088/
JupyterHub	 https://ec2-13-38-83-117.eu-west-3.compute.amazonaws.com:9443/
Nom du nœud HDFS	 http://ec2-13-38-83-117.eu-west-3.compute.amazonaws.com:9870/
Spark History Server	 http://ec2-13-38-83-117.eu-west-3.compute.amazonaws.com:18080/




Sign in

Username:

Password:

Sign in



3.3.1-amzn-0

History Server

Event log directory: hdfs:///var/log/spark/apps

Last updated: 2023-05-09 09:58:02

Client local time zone: Europe/Paris

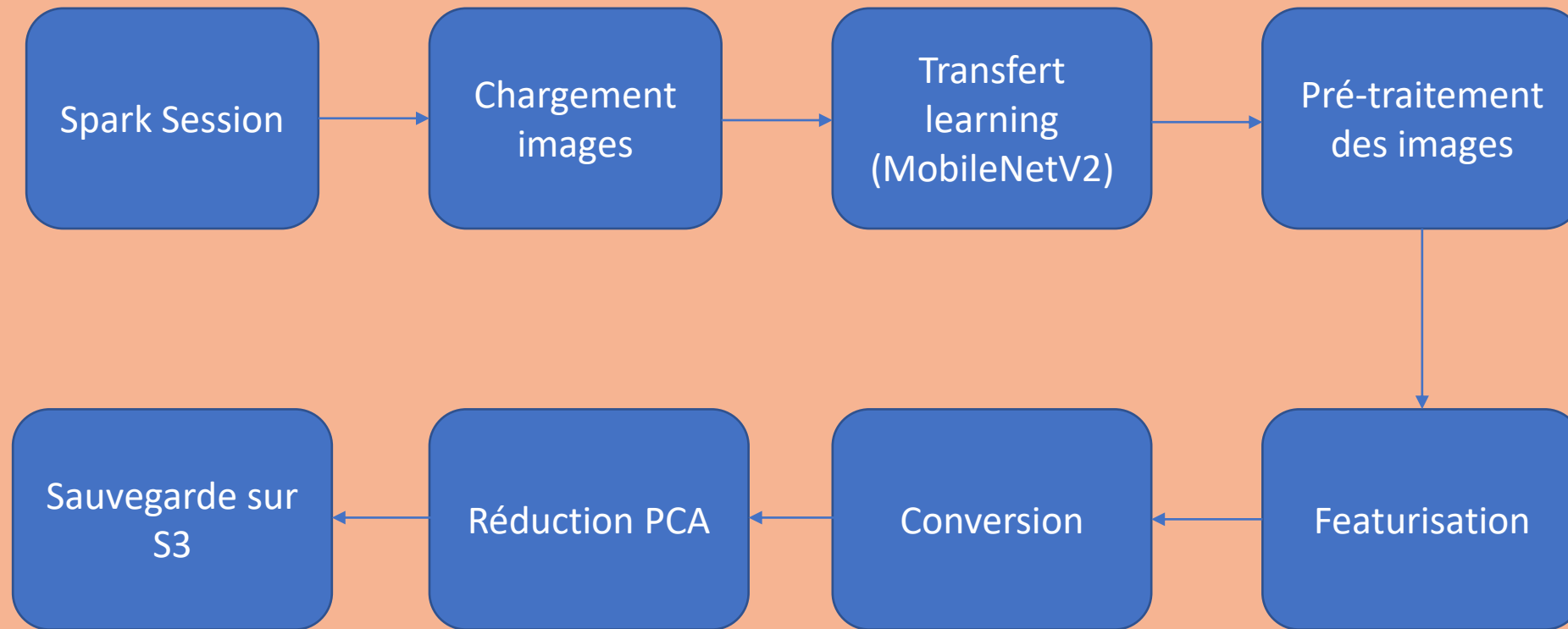
No completed applications found!

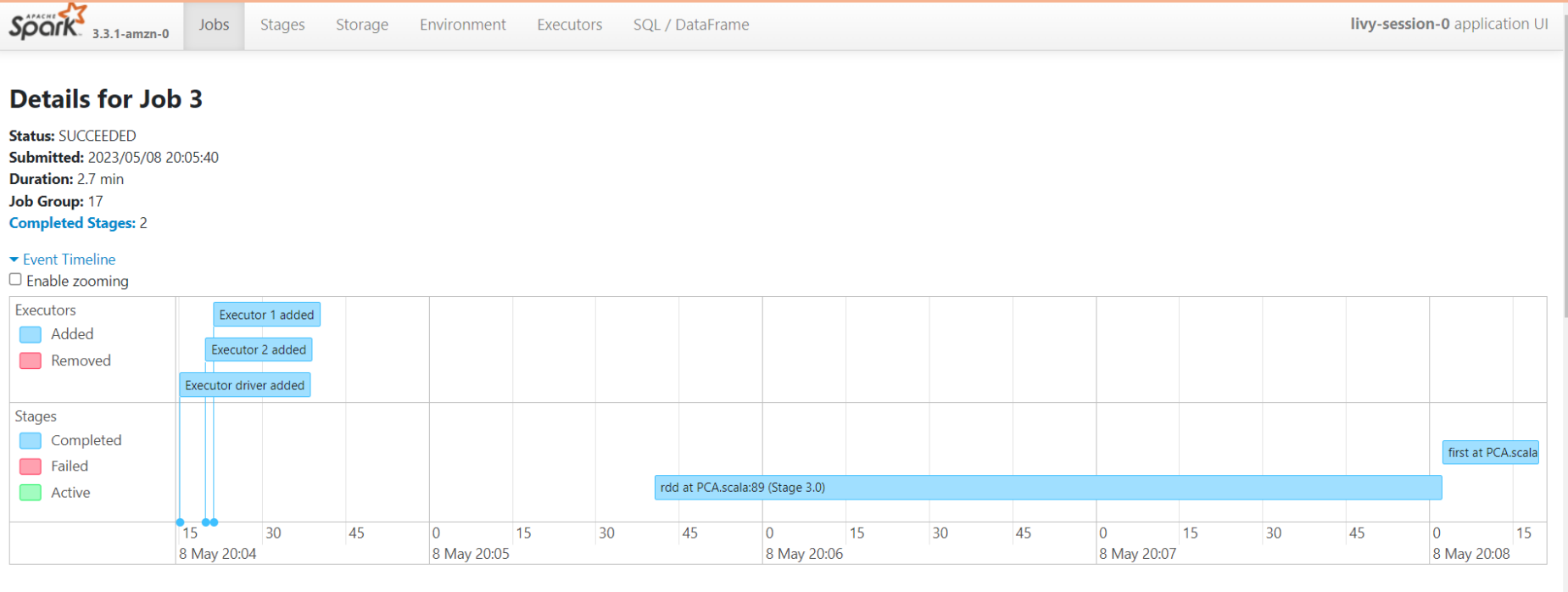
Did you specify the correct logging directory? Please verify your setting of `spark.history.fs.logDirectory` listed above and whether you have the permissions to access it. It is also possible that your application did not run to completion or did not stop the SparkContext.









[Show incomplete applications](#)



Etapes du Script PySpark





<input type="checkbox"/>	Nom	▲	Type ▼	Dernière modification ▼	Taille ▼	Classe de stockage
<input type="checkbox"/>	 _SUCCESS		-	08 May 2023 10:37:55 PM CEST	0 o	Standard
<input type="checkbox"/>	 part-00000-7b20f26f-9ec5-40c9-9f0c-2c40e6707148-c000.snappy.parquet		parquet	08 May 2023 10:37:52 PM CEST	3.1 Mo	Standard
<input type="checkbox"/>	 part-00001-7b20f26f-9ec5-40c9-9f0c-2c40e6707148-c000.snappy.parquet		parquet	08 May 2023 10:37:53 PM CEST	3.1 Mo	Standard
<input type="checkbox"/>	 part-00002-7b20f26f-9ec5-40c9-9f0c-2c40e6707148-c000.snappy.parquet		parquet	08 May 2023 10:37:37 PM CEST	3.1 Mo	Standard
<input type="checkbox"/>	 part-00003-7b20f26f-9ec5-40c9-9f0c-2c40e6707148-c000.snappy.parquet		parquet	08 May 2023 10:37:36 PM CEST	3.2 Mo	Standard
<input type="checkbox"/>	 part-00004-7b20f26f-9ec5-40c9-9f0c-2c40e6707148-c000.snappy.parquet		parquet	08 May 2023 10:37:53 PM CEST	3.2 Mo	Standard
<input type="checkbox"/>	 part-00005-7b20f26f-9ec5-40c9-9f0c-2c40e6707148-c000.snappy.parquet		parquet	08 May 2023 10:37:42 PM CEST	3.1 Mo	Standard
<input type="checkbox"/>	 part-00006-7b20f26f-9ec5-40c9-9f0c-2c40e6707148-c000.snappy.parquet		parquet	08 May 2023 10:37:39 PM CEST	3.1 Mo	Standard

Démonstration du script PySpark



Synthèse et conclusion

La combinaison des différents services AWS permettent de gérer les éventuels problèmes:

- La gestion de données massives qui évoluent
- L'environnement de travail
- La capacité matérielle
- Le coût

MERCI DE VOTRE ATTENTION

