PROJET_09 :

Réalisez un traitement dans un environnement Big data sur le cloud

Jérôme LE GAL
Etudiant OpenClassRooms – parcours Data Scientist
Le 16/02/2025

OPENCLASSROOMS

CONTEXTE ET OBJECTIFS

- Mission pour la startup : ***Fruits***
- Application mobile : informations sur les fruits

- Modèle sur le Cloud
- Architecture Big Data
- Respect des contraintes RGPD

DESCRIPTION DU JEU DE DONNÉES

Jeu de test :

Images de fruits

- Quantité : 22688 (131 variétés)
- Format : jpg
- Résolution : 100x100 pixels
- Auteur : Mihai Oltean
- Licence : MIT License (permission : à titre gratuit)

ENVIRONNEMENT **BIG DATA**

Pourquoi ?

- Traiter gros volume de données
- Traiter rapidement

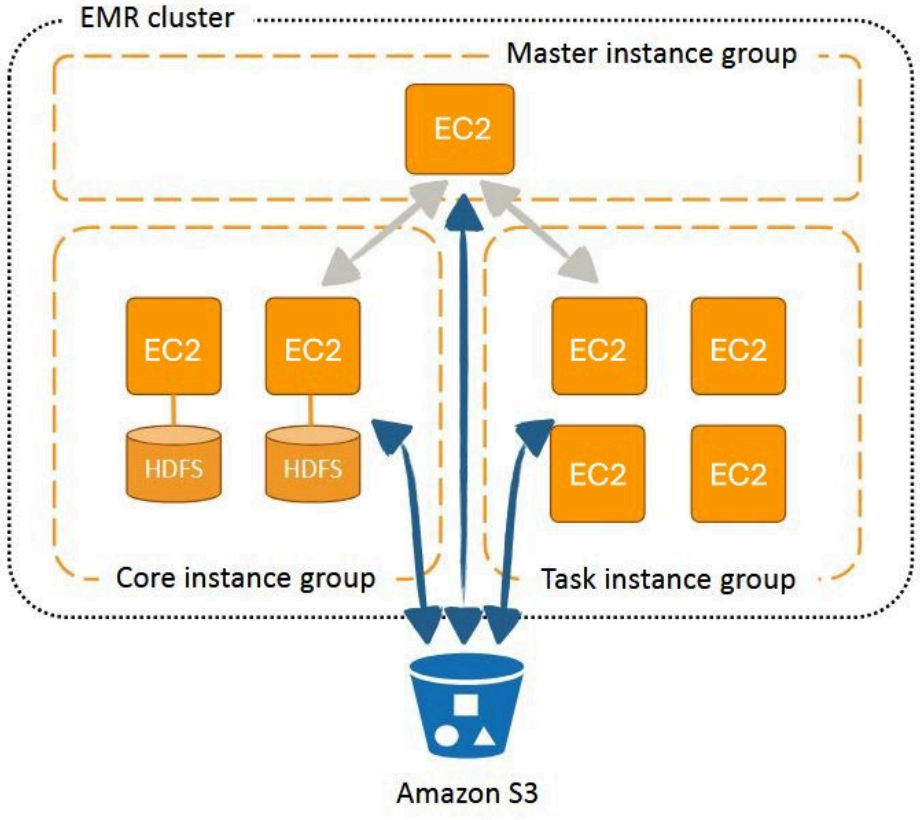
Contraintes :

- Plateforme Cloud Big Data
- Coûts de fonctionnement
- Adapter les scripts
- Règles RGPD

CHOIX TECHNIQUES

TRAITEMENT DES DONNÉES	SPARK : permet le calcul distribué
PLATEFORME	AWS : propose EMR avec mise à l'échelle à la demande (coûts maîtrisés)
STOCKAGE	Bucket S3
TRANSFER LEARNING	Modèle léger : MobileNetV2
FORMAT DE SORTIE	Fichiers Parquet

EMR



- Master instance group
  - NameNode (HDFS)
  - Yarn
  - Job History
  - Spark Driver
- Core instance group
- Task instance group

EC2 = Serveur virtuel dans le Cloud

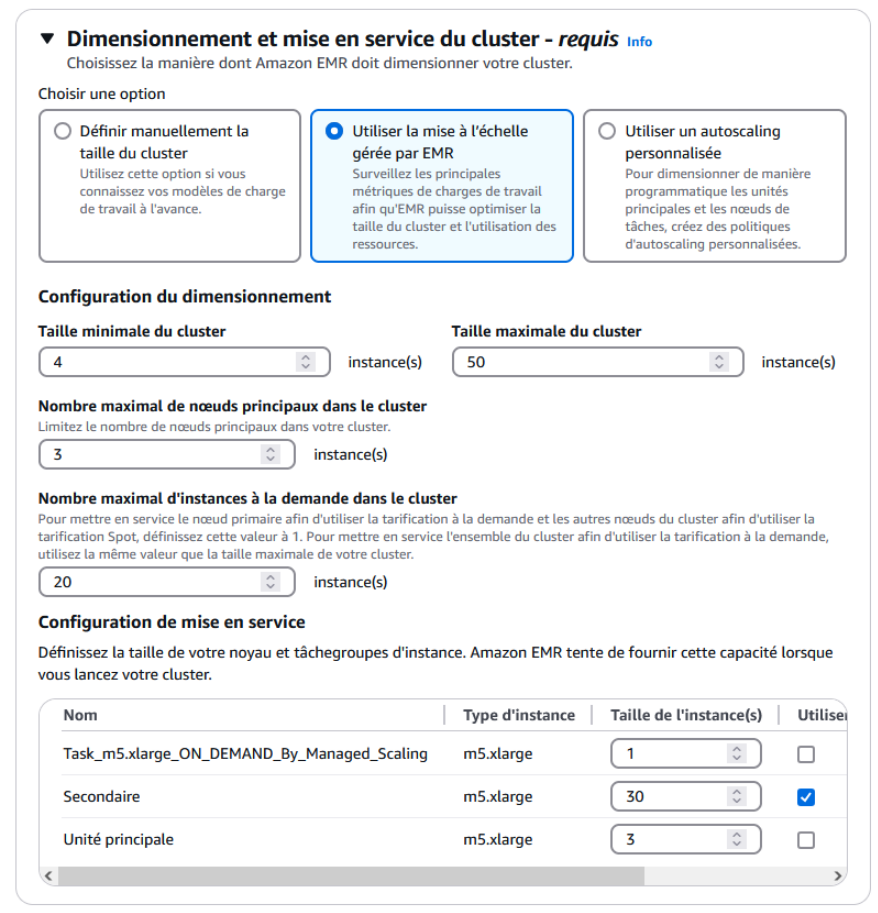
m5.xlarge

4 vCore, 16 Gio de mémoire, Stockage EBS uniquement

CONFIGURATION EMR

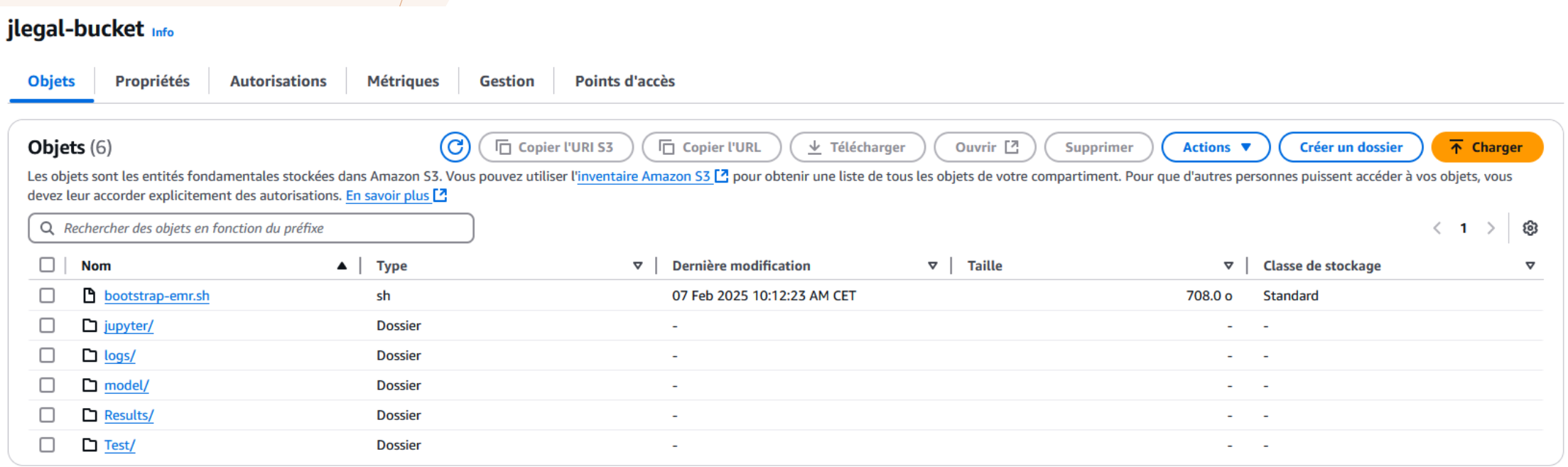
- Pré-installation :
  - TensorFlow
  - Spark
  - JupyterHub
- Bootstrap
- Réseau privé
- Connexion sécurisée par tunnel SSH
- Logs sur S3
- Persistance des Notebooks

- Mise à l'échelle automatique



S3

- Bootstrap
- Notebooks en persistant
- Logs EMR
- Sauvegarde du modèle
- Fichiers de sortie
- Fichiers d’entrée



IAM ET SECURITÉ

IAM – Identity and Access Management

- Utilisateurs
- Groupes
- Rôles
- Politiques

Sécurité

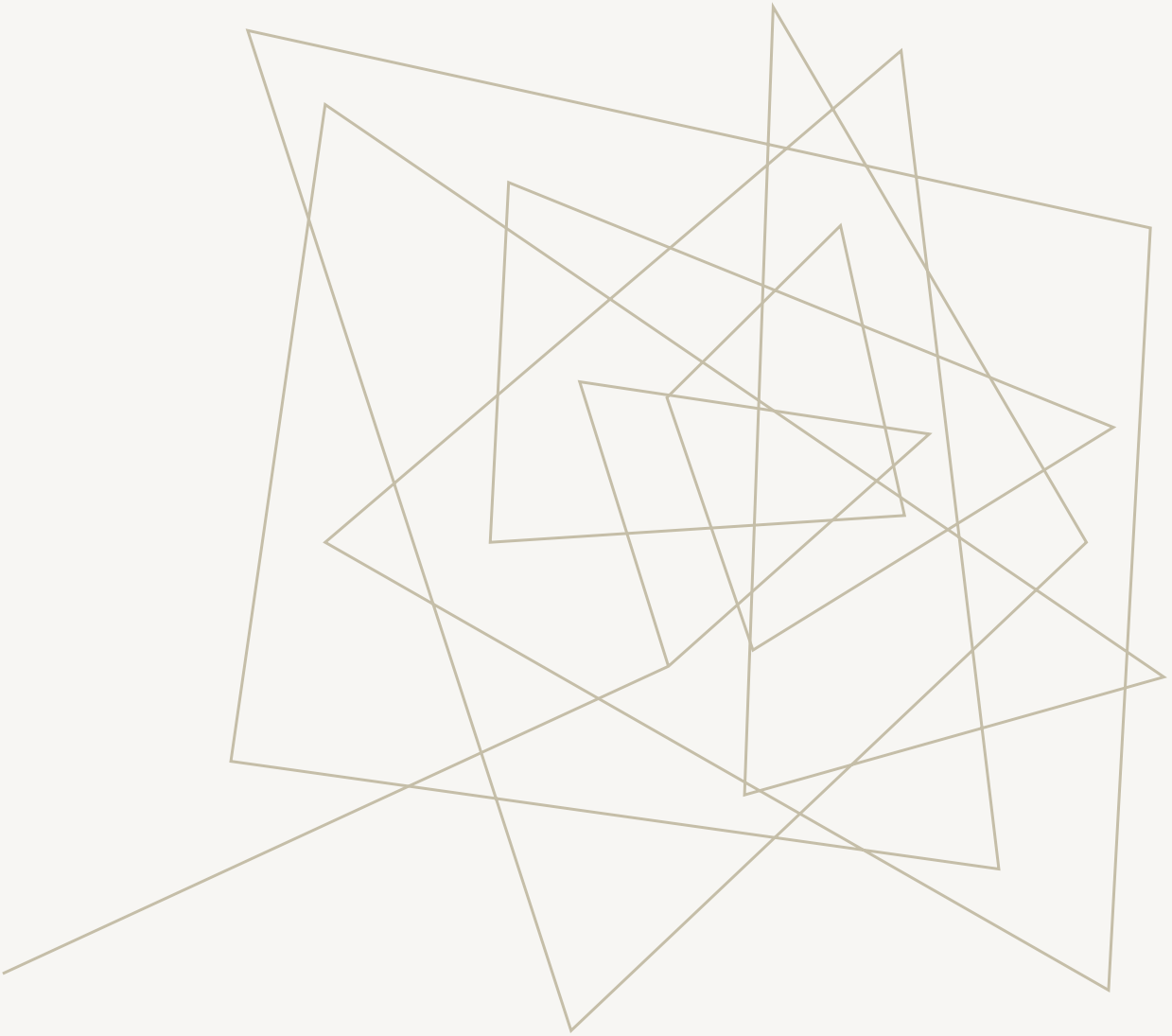
- Paires de clés
- Tunnel SSH
- EBS Snapshots

Identity and Access Management

MODÉLISATION

Calcul distribué

- Données au format « binaire »
- Preprocessing
  - resize 224x224
- Pandas_UDF
- Modèle pré-entrainé (MobileNetV2)
  - Output (1,1,1280)
- PCA
  - 90% de variance expliquée (1,1,186)
- ✓ Création d'un pipeline



DÉMONSTRATION SUR LE CLOUD

RÈGLES RGPD

Licéité,
Loyauté,
Transparence

Limitation
des finalités

Minimisation
des données

Exactitude

Limitation
de la
conservation

- Les images utilisées servent exclusivement au test du modèle
- Serveurs et Stockage en France (Paris)
- Images d’entrainement libre de droits et non conservées
- Seules les images sont utilisées, pas d’informations personnelles

CONCLUSION

- Calcul distribué avec Spark
- EMR : bonne base simple à mettre en œuvre
  - Haute disponibilité
  - Redondance
  - Coûts réduits avec optimisation
  - Mise à l’échelle automatique

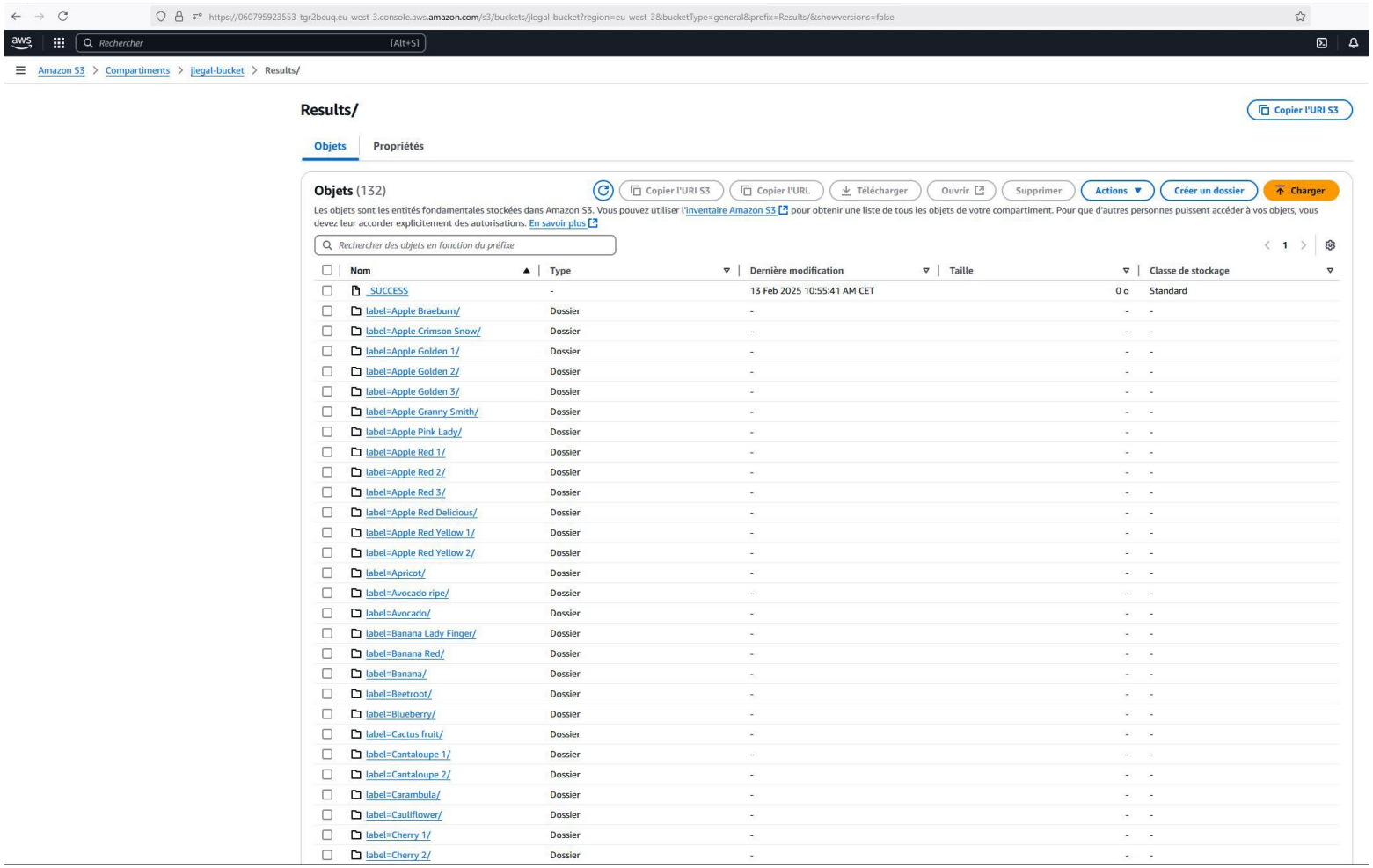
AMÉLIORATIONS

- Augmenter les capacités des EC2
- Passage aux GPUs
- Production : Master en HD
- Spark Streaming pour appli en temps réel

MERCI

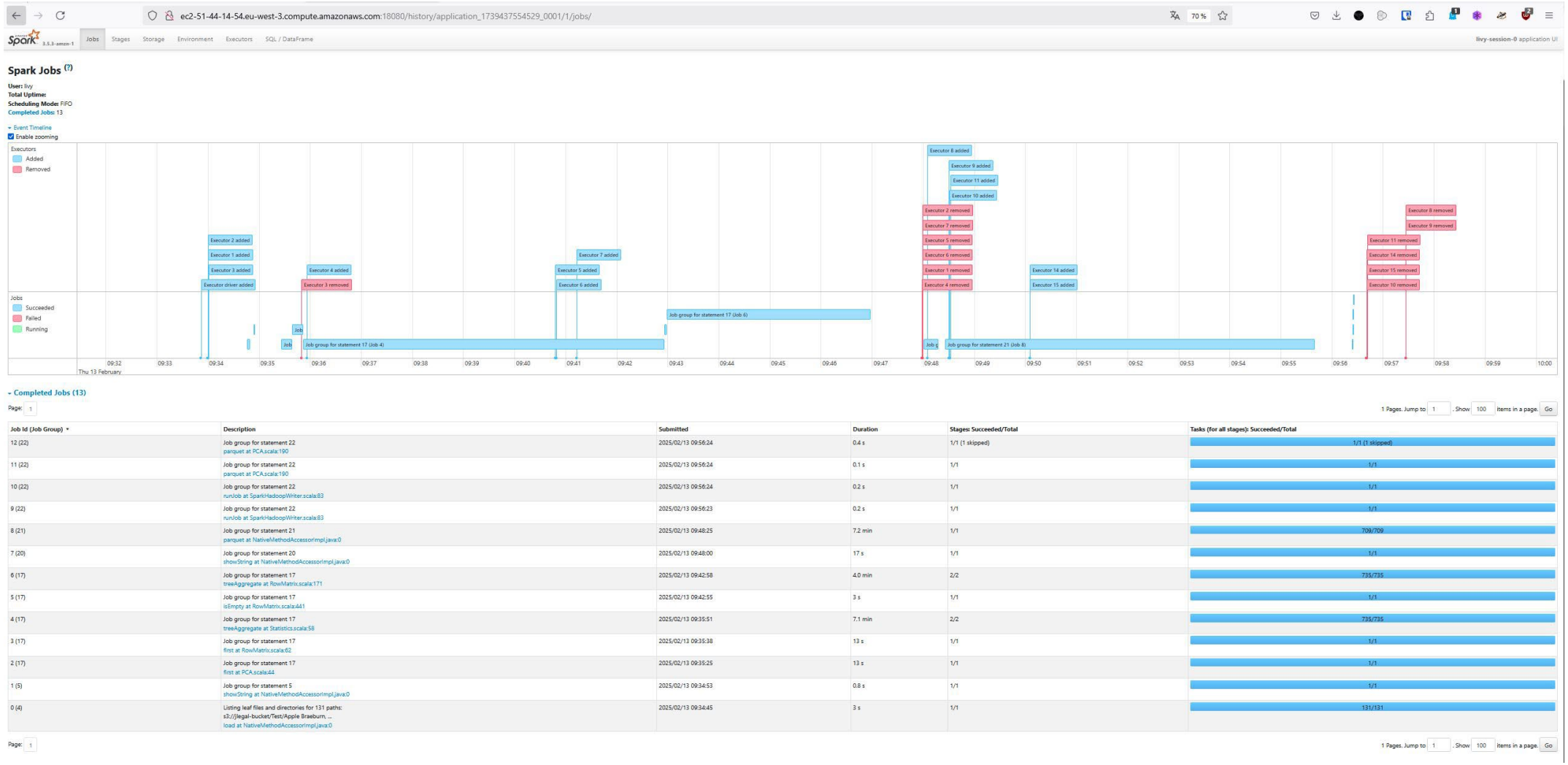
ANNEXES

Fichiers de sortie au format « parquet »



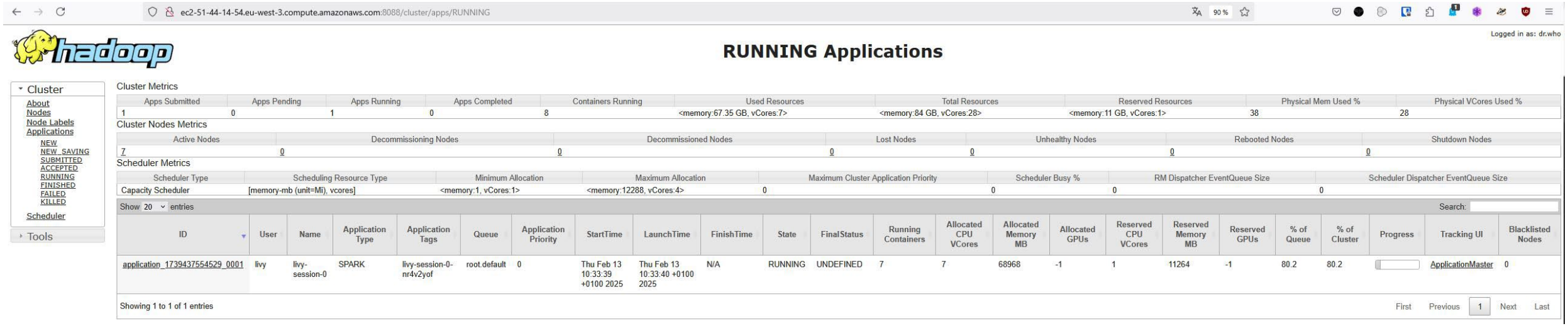
ANNEXES

Séquencement des « travaux » de Spark



ANNEXES

Gestionnaire des ressources



ANNEXES

Résumé HDFS

Summary

Security is off.

Safemode is off.

22 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 22 total filesystem object(s).

Heap Memory used 201.18 MB of 250 MB Heap Memory. Max Heap Memory is 1.8 GB.

Non Heap Memory used 63.73 MB of 65.44 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	173.62 GB
Configured Remote Capacity:	0 B
DFS Used:	24 KB (0%)
Non DFS Used:	0 B
DFS Remaining:	173.62 GB (100%)
Block Pool Used:	24 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	3 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion (including replicas)	0
Block Deletion Start Time	Thu Feb 13 10:04:58 +0100 2025
Last Checkpoint Time	Thu Feb 13 10:04:55 +0100 2025
Last HA Transition Time	Never
Enabled Erasure Coding Policies	RS-6-3-1024k