



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

| 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

PLATFORME

AWS : propose EMR avec mise à l'échelle à la demande
(coûts maîtrisés)

STOCKAGE

Bucket S3

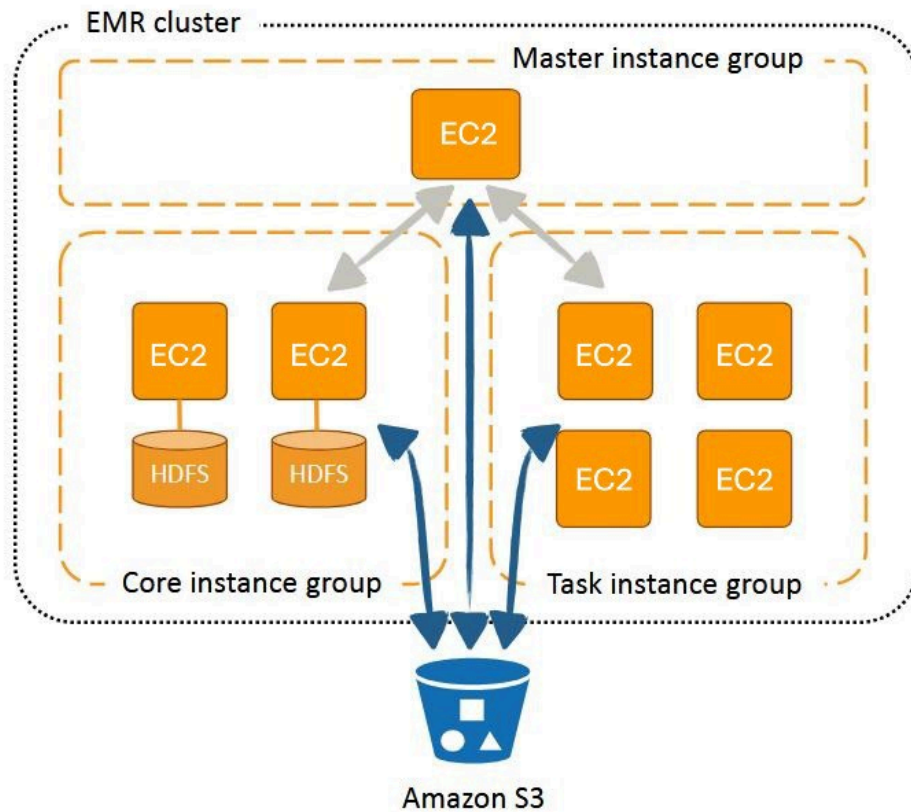
TRANSFER LEARNING

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

▼ **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"/> bootstrap-emr	s3://legal-bucket/bootstrap-emr.sh	-

- Mise à l'échelle automatique

▼ **Dimensionnement et mise en service du cluster - requis** [Info](#)

Choisissez la manière dont Amazon EMR doit dimensionner votre cluster.

Choisir une option

☐ Définir manuellement la taille du cluster
Utilisez cette option si vous connaissez vos modèles de charge de travail à l'avance.

☒ Utiliser la mise à l'échelle gérée par EMR
Surveillez les principales métriques de charges de travail afin qu'EMR puisse optimiser la taille du cluster et l'utilisation des ressources.

☐ Utiliser un autoscaling personnalisé
Pour dimensionner de manière programmatique les unités principales et les nœuds de tâches, créez des politiques d'autoscaling personnalisées.

Configuration du dimensionnement

Taille minimale du cluster instance(s) Taille maximale du cluster instance(s)

Nombre maximal de nœuds principaux dans le cluster
Limitez le nombre de nœuds principaux dans votre cluster.
 instance(s)

Nombre maximal d'instances à la demande dans le cluster
Pour mettre en service le nœud primaire afin d'utiliser la tarification à la demande et les autres nœuds du cluster afin d'utiliser la tarification Spot, définissez cette valeur à 1. Pour mettre en service l'ensemble du cluster afin d'utiliser la tarification à la demande, utilisez la même valeur que la taille maximale de votre cluster.
 instance(s)

Configuration de mise en service
Définissez la taille de votre noyau et tâchesgroupes d'instance. Amazon EMR tente de fournir cette capacité lorsque vous lancez votre cluster.

Nom	Type d'instance	Taille de l'instance(s)	Utiliser
Task_m5.xlarge_ON_DEMAND_By_Managed_Scaling	m5.xlarge	<input type="text" value="1"/>	<input type="checkbox"/>
Secondaire	m5.xlarge	<input type="text" value="30"/>	<input checked="" type="checkbox"/>
Unité principale	m5.xlarge	<input type="text" value="3"/>	<input type="checkbox"/>

S3

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

jlegal-bucket [Info](#)

[Objets](#) | [Propriétés](#) | [Autorisations](#) | [Métriques](#) | [Gestion](#) | [Points d'accès](#)

Objets (6)

[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](#)

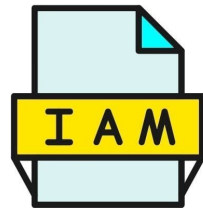
< 1 > ⚙

<input type="checkbox"/>	Nom	Type	Dernière modification	Taille	Classe de stockage
<input type="checkbox"/>	bootstrap-emr.sh	sh	07 Feb 2025 10:12:23 AM CET	708.0 o	Standard
<input type="checkbox"/>	jupyter/	Dossier	-	-	-
<input type="checkbox"/>	logs/	Dossier	-	-	-
<input type="checkbox"/>	model/	Dossier	-	-	-
<input type="checkbox"/>	Results/	Dossier	-	-	-
<input type="checkbox"/>	Test/	Dossier	-	-	-

| IAM ET SECURITÉ

IAM – Identity and Access Management

- Utilisateurs
- Groupes
- Rôles
- Politiques



Sécurité

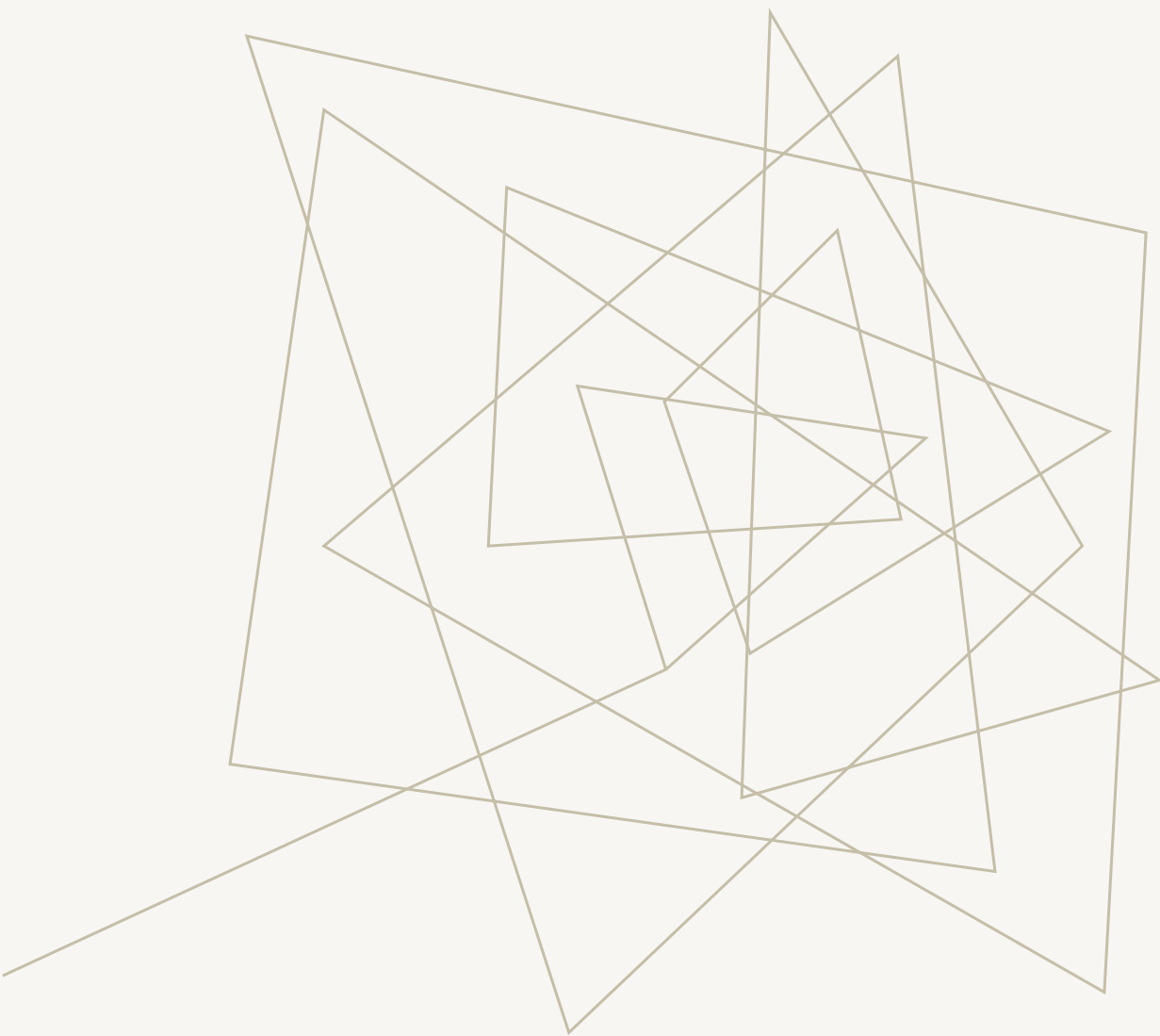
- Paires de clés
- Tunnel SSH
- EBS Snapshots

| MODÉLISATION



Calcul distribué

- Données au format « binaire »
- Preprocessing
 - ❑ resize 224x224
- Pandas_UDF
- Modèle pré-entraîné (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'entraînement 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

A series of thin, light brown lines forming an abstract, overlapping geometric pattern on the left side of the slide. The lines intersect to create various polygonal shapes, some of which are filled with a very light brown color.

MERCI

ANNEXES

Fichiers de sortie au
format « parquet »

Results/ [Copier l'URI S3](#)

Objets | Propriétés

Objets (132) [Copier l'URI S3](#) [Copier l'URL](#) [Télécharger](#) [Ouvrir](#) [Supprimer](#) [Actions](#) [Créer un dossier](#) [Charger](#)

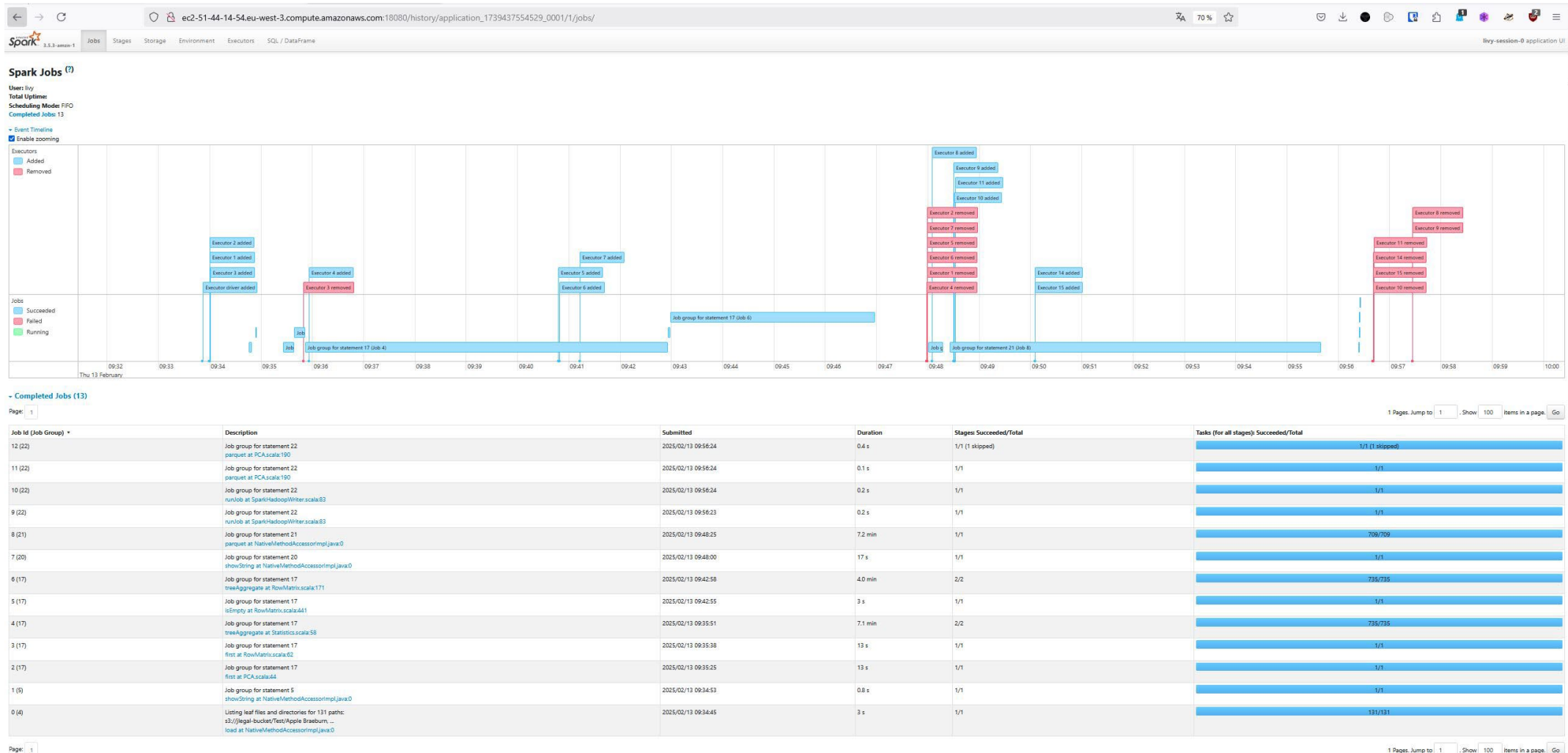
Les objets sont les entités fondamentales stockées dans Amazon S3. Vous pouvez utiliser l'[inventaire Amazon S3](#) pour obtenir une liste de tous les objets de votre compartiment. Pour que d'autres personnes puissent accéder à vos objets, vous devez leur accorder explicitement des autorisations. [En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

Rechercher des objets en fonction du préfixe

<input type="checkbox"/>	Nom	Type	Dernière modification	Taille	Classe de stockage
<input type="checkbox"/>	_SUCCESS	-	13 Feb 2025 10:55:41 AM CET	0 o	Standard
<input type="checkbox"/>	label=Apple Braeburn/	Dossier	-	-	-
<input type="checkbox"/>	label=Apple Crimson Snow/	Dossier	-	-	-
<input type="checkbox"/>	label=Apple Golden 1/	Dossier	-	-	-
<input type="checkbox"/>	label=Apple Golden 2/	Dossier	-	-	-
<input type="checkbox"/>	label=Apple Golden 3/	Dossier	-	-	-
<input type="checkbox"/>	label=Apple Granny Smith/	Dossier	-	-	-
<input type="checkbox"/>	label=Apple Pink Lady/	Dossier	-	-	-
<input type="checkbox"/>	label=Apple Red 1/	Dossier	-	-	-
<input type="checkbox"/>	label=Apple Red 2/	Dossier	-	-	-
<input type="checkbox"/>	label=Apple Red 3/	Dossier	-	-	-
<input type="checkbox"/>	label=Apple Red Delicious/	Dossier	-	-	-
<input type="checkbox"/>	label=Apple Red Yellow 1/	Dossier	-	-	-
<input type="checkbox"/>	label=Apple Red Yellow 2/	Dossier	-	-	-
<input type="checkbox"/>	label=Apricot/	Dossier	-	-	-
<input type="checkbox"/>	label=Avocado ripe/	Dossier	-	-	-
<input type="checkbox"/>	label=Avocado/	Dossier	-	-	-
<input type="checkbox"/>	label=Banana Lady Finger/	Dossier	-	-	-
<input type="checkbox"/>	label=Banana Red/	Dossier	-	-	-
<input type="checkbox"/>	label=Banana/	Dossier	-	-	-
<input type="checkbox"/>	label=Beetroot/	Dossier	-	-	-
<input type="checkbox"/>	label=Blueberry/	Dossier	-	-	-
<input type="checkbox"/>	label=Cactus fruit/	Dossier	-	-	-
<input type="checkbox"/>	label=Cantaloupe 1/	Dossier	-	-	-
<input type="checkbox"/>	label=Cantaloupe 2/	Dossier	-	-	-
<input type="checkbox"/>	label=Carambola/	Dossier	-	-	-
<input type="checkbox"/>	label=Cauliflower/	Dossier	-	-	-
<input type="checkbox"/>	label=Cherry 1/	Dossier	-	-	-
<input type="checkbox"/>	label=Cherry 2/	Dossier	-	-	-

ANNEXES

Séquencement des « travaux » de Spark



Gestionnaire des ressources

ec2-51-44-14-54.eu-west-3.compute.amazonaws.com:8088/cluster/apps/RUNNING
90%

Logged in as: dr.who

RUNNING Applications

- [Cluster](#)
- [About Nodes](#)
- [Node Labels](#)
- [Applications](#)
- [NEW](#)
- [NEW SAVING](#)
- [SUBMITTED](#)
- [ACCEPTED](#)
- [RUNNING](#)
- [FINISHED](#)
- [FAILED](#)
- [KILLED](#)
- [Scheduler](#)

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	Total Resources	Reserved Resources	Physical Mem Used %	Physical VCores Used %
1	0	1	0	8	<memory: 67.35 GB, vCores: 7>	<memory: 84 GB, vCores: 28>	<memory: 11 GB, vCores: 1>	38	28

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
7	0	0	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority	Scheduler Busy %	RM Dispatcher EventQueue Size	Scheduler Dispatcher EventQueue Size
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory: 1, vCores: 1>	<memory: 12288, vCores: 4>	0	0	0	0

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Allocated GPUs	Reserved CPU VCores	Reserved Memory MB	Reserved GPUs	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes
application_1739437554529_0001	livy	livy-session-0	SPARK	livy-session-0-nr4v2yof	root default	0	Thu Feb 13 10:33:39 +0100 2025	Thu Feb 13 10:33:40 +0100 2025	N/A	RUNNING	UNDEFINED	7	7	68968	-1	1	11264	-1	80.2	80.2	<div style="width: 100%;"></div>	ApplicationMaster	0

Showing 1 to 1 of 1 entries

Résumé HDFS

ANNEXES

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 Committed 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