Projet 8 : Déployez un modèle dans le cloud

Présenté par :
Bourama FANE
Etudiant Data Scientist

Dirigé par :
Babou M'BAYE
Mentor chez OpenClassrooms

08 Janvier 2024

Sommaire

Plan de la présentation

Contexte

Vous êtes **Data Scientist** dans une très jeune **start-up de l'AgriTech**, nommée "**Fruits !**", qui cherche à proposer des solutions innovantes pour la récolte des fruits.

La volonté de lentreprise est de préserver la biodiversité des fruits en permettant des traitements spécifiques pour chaque espèce de fruits en développant des robots cueilleurs intelligents.

Problématique

Votre start-up souhaite dans un premier temps se faire connaître en mettant à disposition du grand public une application mobile qui permettrait aux utilisateurs de prendre en photo un fruit et d'obtenir des informations sur ce fruit.

Pour la start-up, cette application permettrait de sensibiliser le grand public à la biodiversité des fruits et de mettre en place une première version du moteur de classification des images de fruits.

Objectifs

- ☞ Développer une première chaîne de traitement des données qui comprendra le **preprocessing** et une étape de **réduction de dimension**.

- ☞ Tenir compte du fait que le volume de données va augmenter très rapidement après la livraison de ce projet, ce qui implique de :

  - ✓ Déployer le traitement des données dans un **environnement Big Data**
  - ✓ Développer les scripts en **pyspark** pour effectuer du **calcul distribué**

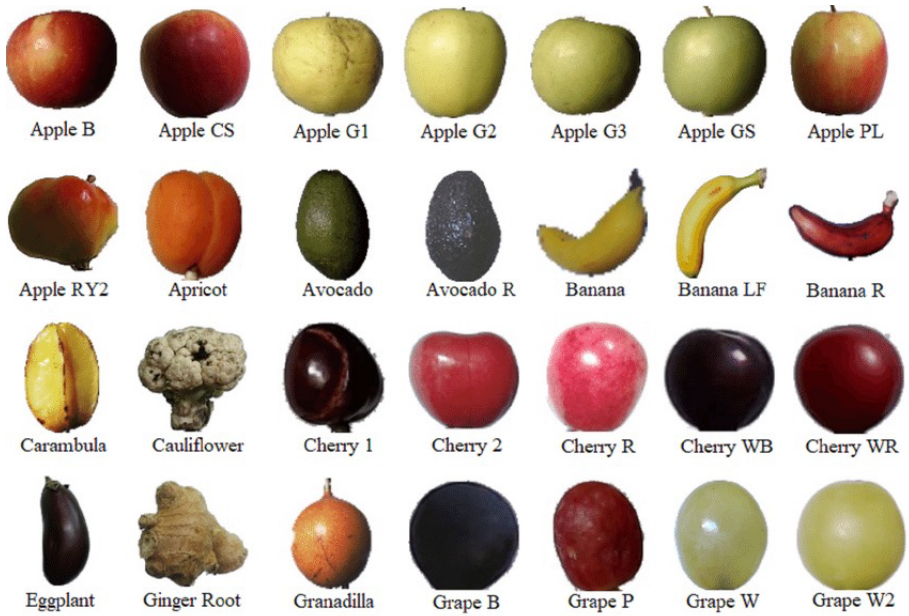
Différentes étapes du projet

- ✓ Liste des choix techniques généraux retenus
- ✓ Déploiement de la solution en local
- ✓ Déploiement de la solution dans le cloud

Données

☞ **"Fruit 360"** contient **90 483 images**

- ✓ Training : 67 692 images
- ✓ Test : **22 688 images**



☞ Répartition des images en **131 dossiers**

- ✓ 1 dossier = 1 fruit
- ✓ Photos prises sous différents angles
- ✓ Images de 100X100 pixels au format : JPG, RGB

Plan de la présentation

Choix techniques généraux retenus

Calcul distribué

- ☞ Dans les deux environnements (Local et Cloud) nous utiliserons donc Spark et nous lexploiterons à travers des scripts python grâce à **PySpark**.
- ☞ Dans la version cloud nous réaliserons les opérations sur un cluster de machine.

Transfert Learning

- ☞ Nous allons utilisons un **MobileNetV2** ;
- ☞ L'avant dernière couche correspond à un vecteur réduit de dimension (1,1,1280) ;
- ☞ Rapidité d'exécution.

Plan de la présentation

Deploiement local

- ✓ Préparer nos données
  - ☞ Importer les images dans un dataframe pandas UDF
  - ☞ Associer aux images leur label
  - ☞ Préprocessing
- ✓ Préparer notre modèle
  - ☞ Importer le modèle MobileNetV2
  - ☞ Créer un nouveau modèle dépourvu de la dernière couche de **MobileNetV2**
- ✓ Exécuter les actions d'extraction de features
- ✓ Enregistrer le résultat de nos actions
- ✓ Tester le bon fonctionnement en chargeant les données enregistrées

Plan de la présentation

Quest ce que le BigData

BigData

« 3 caractéristiques »sont associées au Big Data nommés les « **3V** » :

- **Volume** : Le Big Data implique des quantités massives de données. Datas ne tiennent pas en Ram (>million de gigaoctets)
- **Vélocité** : Les Data sont générées et modifiées très rapidement.
- **Variété** : Data hétérogènes en types et formats. (données structurées, données semi-structurées (comme le format JSON) et données non structurées (texte, vidéos, images, etc.).

Pourquoi le cloud

- Coût (Pay as You Go)
- Evite les problème lié à linfrastructure et à sa maintenance
- Facilité pour passer à léchelle

Prestataire cloud : AWS

L'objectif premier est de pouvoir, grâce à AWS, louer de la puissance de calcul à la demande. L'idée étant de pouvoir, quel que soit la charge de travail, obtenir suffisamment de puissance de calcul pour pouvoir traiter nos images, même si le volume de données venait à fortement augmenter.

De plus, la capacité d'utiliser cette puissance de calcul à la demande permet de diminuer drastiquement les coûts si l'on compare les coûts d'une location de serveur complet sur une durée fixe (1 mois, 1 année par exemple).

Solution & Stockage

Service EMR

✓ la solution **PAAS** en choisissant d'utiliser le service **EMR** d'Amazon Web Services.

EC2

✗ Espace limité ;

✗ Risque de saturation (ralentissements, disfonctionnements) ;

✗ Pertes de données en cas de résiliation (Coûts) ;

Amazon S3

✓ Utiliser **Amazon S3** permet de s'affranchir de toutes ces problématiques. L'espace disque disponible est illimité, et il est indépendant de nos serveurs EC2. L'accès aux données est très rapide car nous restons dans **l'environnement d'AWS**

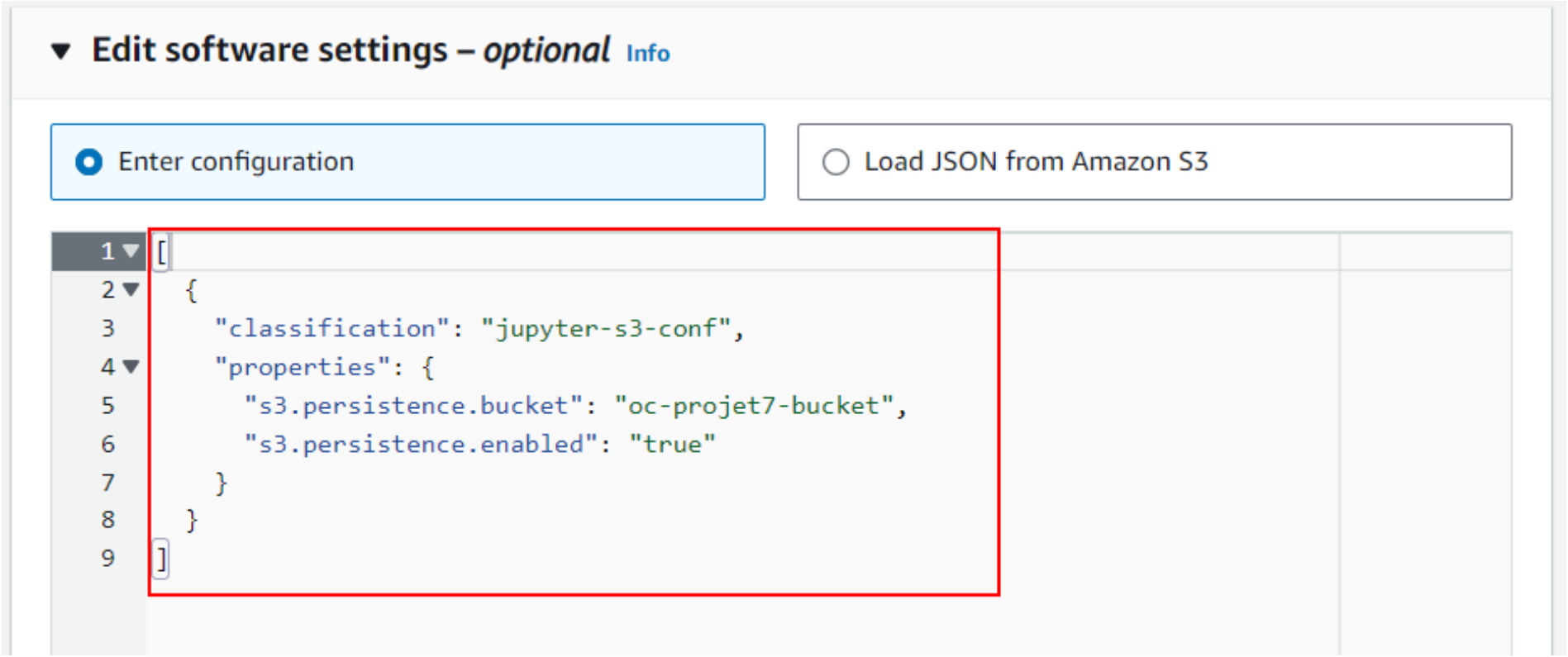
Configurations EMR : choix des logiciels à installer

Application bundle

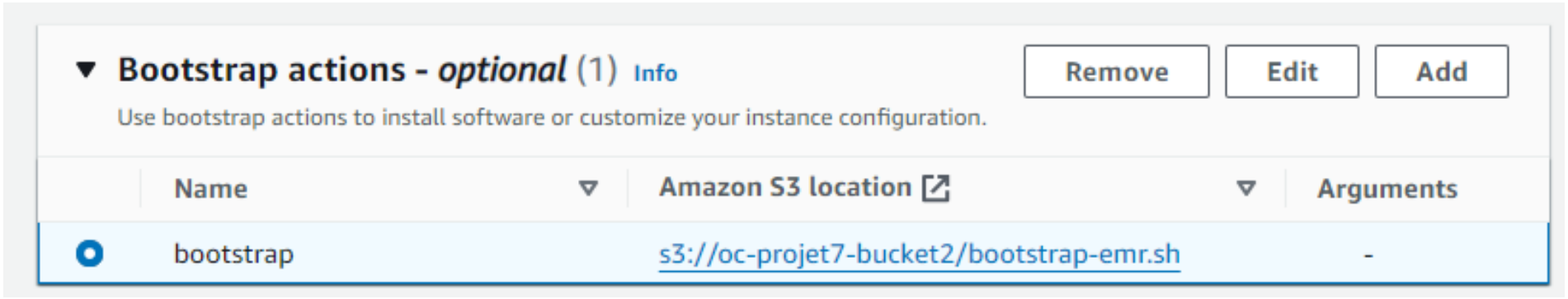
Spark Interactive	Core Hadoop	Flink	HBase	Presto	Trino	Custom

- [ ] AmazonCloudWatchAgent 1.300031.1
- [ ] HCatalog 3.1.3
- [ ] Hue 4.11.0
- [ ] Livy 0.7.1
- [ ] Phoenix 5.1.3
- [x] Spark 3.5.0
- [ ] Tez 0.10.2
- [ ] ZooKeeper 3.5.10
- [ ] Flink 1.18.0
- [x] Hadoop 3.3.6
- [ ] JupyterEnterpriseGateway 2.6.0
- [ ] MXNet 1.9.1
- [ ] Pig 0.17.0
- [ ] Sqoop 1.4.7
- [ ] Trino 426
- [ ] HBase 2.4.17
- [ ] Hive 3.1.3
- [x] JupyterHub 1.5.0
- [ ] Oozie 5.2.1
- [ ] Presto 0.283
- [ ] TensorFlow 2.11.0
- [ ] Zeppelin 0.10.1

Configurations EMR : Persistance des notebooks

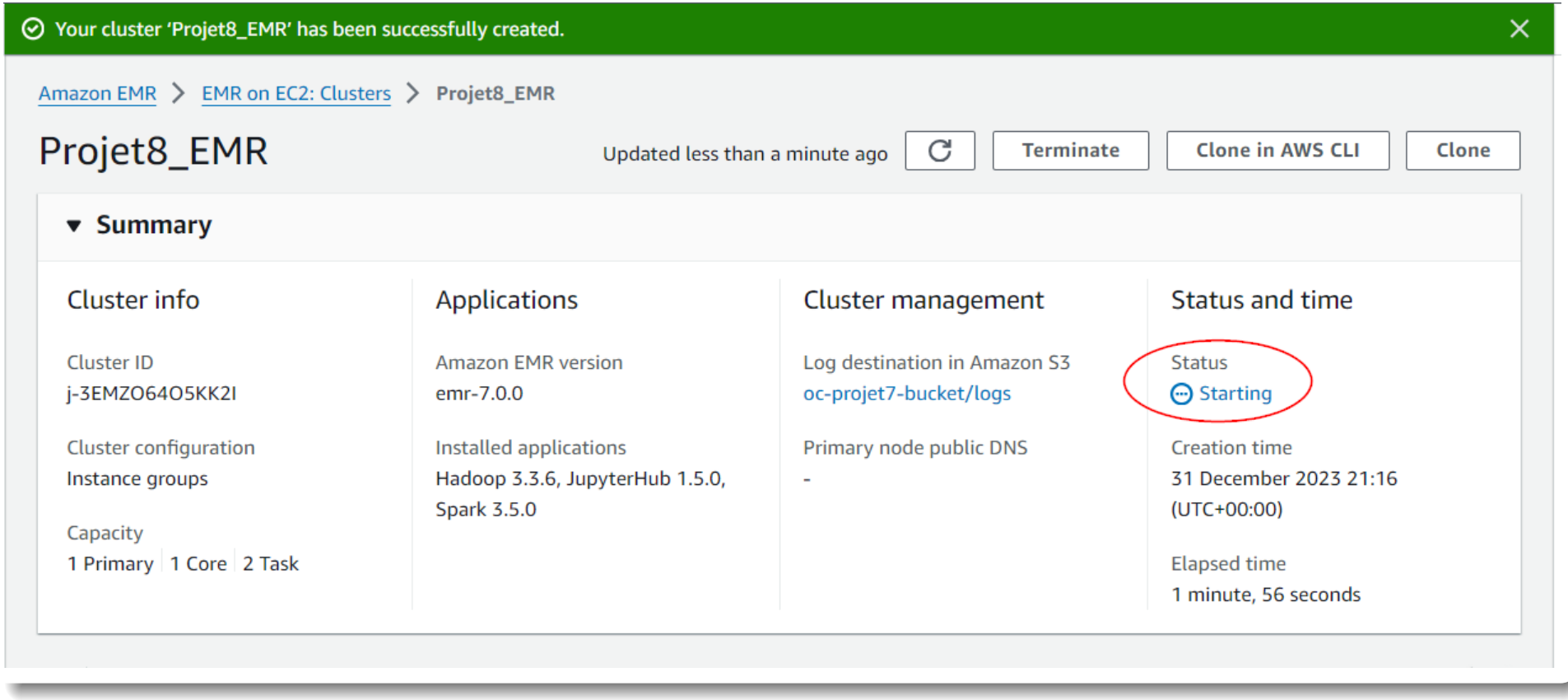


Configurations EMR : Actions d'amorçage




```
#!/bin/bash
sudo python3 -m pip install -U setuptools pip
sudo python3 -m pip install  wheel pillow pandas numpy pyarrow boto3 s3fs fsspec
sudo python3 -m pip install  tensorflow
```

Configurations EMR : Instanciation du serveur



Configurations EMR : Tunnel SSH

hadoop@ip-172-31-22-72:~

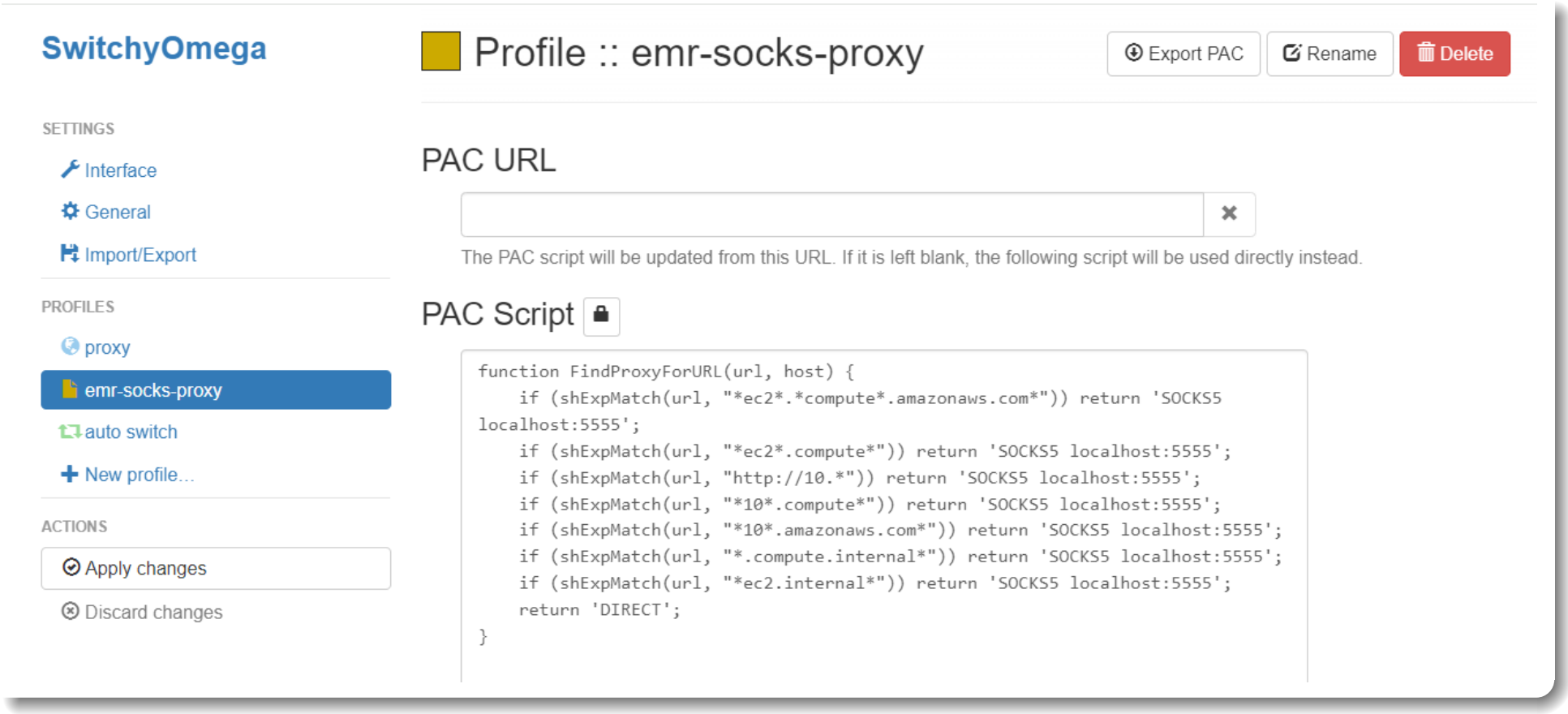
```
PS C:\Users\Fane0763\OpenClassroom\OC Projet 8> ssh -i ./Projet8KeyEC2Naby.pem  -D 5555 hadoop@ec2-54-167-31-13.compute-1.amazonaws.com
Warning: Identity file ./Projet8KeyEC2Naby.pem not accessible: No such file or directory.
hadoop@ec2-54-167-31-13.compute-1.amazonaws.com: Permission denied (publickey,gssapi-keyex,gssapi-with-mic).
PS C:\Users\Fane0763\OpenClassroom\OC Projet 8> ssh -i ./Projet8keyNaby.pem  -D 5555 hadoop@ec2-54-167-31-13.compute-1.amazonaws.com

A newer release of "Amazon Linux" is available.
  Version 2023.3.20231218:
Run "/usr/bin/dnf check-release-update" for full release and version update info
   ,     #_
   ~\_  ####_        Amazon Linux 2023
  ~~  \_#####\
  ~~     \###|
  ~~       \#/ ___   https://aws.amazon.com/linux/amazon-linux-2023
   ~~       V~' '->
    ~~~         /
      ~~._.   _/
         _/ _/
       _/m/'
Last login: Tue Jan  2 08:46:01 2024

EEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE::::::EEEEEEEEE:::E M::::::::M       M::::::::M R::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR:::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R::R      R::::R
  E::::::EEEEEEEEEE  M:::::M M:::M M:::M M:::::M   R::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E::::::EEEEEEEEEE  M:::::M   M:::::M   M:::::M   R::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R::R      R::::R
EE::::::EEEEEEEE:::E M:::::M             M:::::M   R::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR:::R      R::::R
EEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-22-72 ~]$
```

Configurations EMR : SwitchyOmega



Configurations EMR : Connexion à Jupyter

Application UIs

Spark History Server UI

Application UIs on the primary node

Enable an SSH connection

These require SSH tunneling to be enabled.

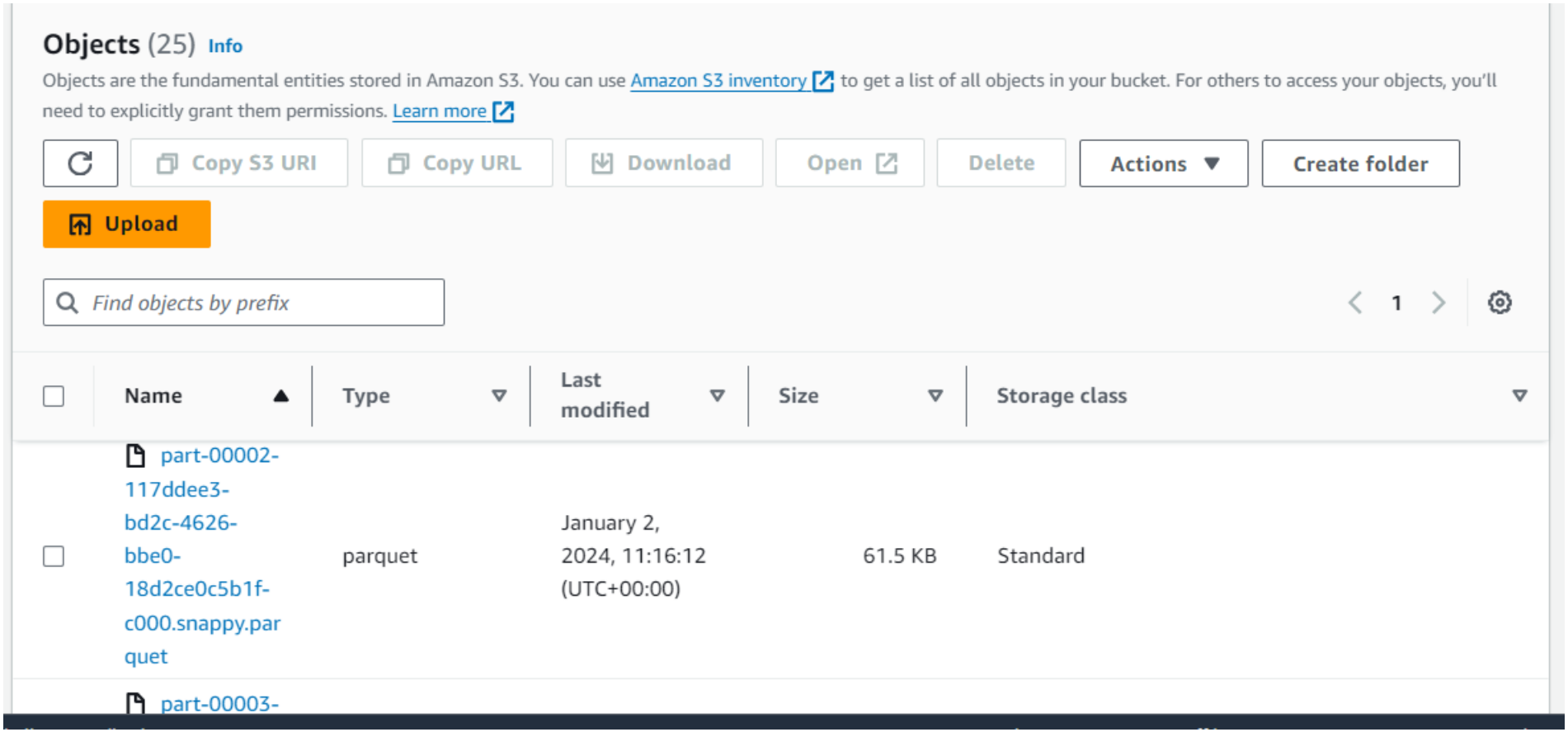
Application	UI URL
HDFS Name Node	http://ec2-54-167-31-13.compute-1.amazonaws.com:9870/
JupyterHub	https://ec2-54-167-31-13.compute-1.amazonaws.com:9443/
Resource Manager	http://ec2-54-167-31-13.compute-1.amazonaws.com:8088/
Spark History Server	http://ec2-54-167-31-13.compute-1.amazonaws.com:18080/

Application UIs on the core and task nodes

Application	UI URL
HDFS Data Node	http://ec2-000-000-000-000.compute-1.amazonaws.com:9864/
Node Manager	http://ec2-000-000-000-000.compute-1.amazonaws.com:8042/

Résultats

On peut également constater la présence des fichiers



Plan de la présentation

Conclusion

- ☞ La solution a parfaitement fonctionné en mode local.
- ☞ La deuxième phase a consisté à créer un réel cluster de calculs (AWS & EMR).
- ☞ Amazon S3 pour stocker les données de notre projet
- ☞ Il nous sera f**acile de faire face à une monté de la charge** de travail en **redimensionnant** simplement notre cluster de machines.

MERCI POUR VOTRE AIMABLE ATTENTION