

Déployez un modèle dans le cloud

Projet 8 : Openclassroom

Plan de la présentation

A. Introduction

- a. Rappel de la problématique
- b. Présentation du jeu de données

B. Présentation de l'architecture AWS

- a. Préparation des données
- b. Amazon EMR
- c. Architecture adoptée
- d. Chaîne de traitement

C. Conclusion

A. Introduction

a. Rappel de la problématique

- Startup Agritech “Fruits” => application pour la sensibilisation à la biodiversité
- Objectif long terme => Robots cueilleurs intelligents
- Quelle évolution pour les données ?
- Quel environnement choisir ?

b. Présentation du jeu de données

Data Explorer

758.39 MB

- ▼ fruits-360
 - Test
 - Training
 - papers
 - test-multiple_fruits
 - LICENSE
 - readme.md

Data Explorer

758.39 MB

- ▼ fruits-360
 - Test
 - ▼ Training
 - Apple Braeburn
 - Apple Crimson Sn...
 - Apple Golden 1
 - Apple Golden 2
 - Apple Golden 3
 - Apple Granny Smith
 - Apple Pink Lady
 - Apple Red 1
 - Apple Red 2
 - Apple Red 3
 - Apple Red Delicious
 - Apple Red Yellow 1
 - Apple Red Yellow 2
 - Apricot
 - Avocado
 - Avocado ripe
 - Banana
 - Banana Lady Finger
 - Banana Red

b. Présentation du jeu de données

- Données d'entraînement : 67692
- Données de test : 22688
- Données avec plusieurs fruits / légumes : 103
- 131 classes de fruits / légumes
- Taille des images : 100 * 100 pixels

b. Présentation du jeu de données



B. Présentation de l'architecture AWS

a. Préparation des données

- Comment utiliser les données sur internet depuis le cloud?
 - Base de données
 - Serveur sur le cloud + HDFS
 - S3
 - Téléchargement à la demande

a. Préparation des données



Amazon S3 Access Points

Create Access Points for each application and/or user that requires access to objects in your new or existing bucket

Configure S3 Access Points

Configure permissions per Access Point to limit public access, and restrict access by object prefixes, and object tags

Limit Access to VPC

You can create Access Points that limit all S3 storage access to a Virtual Private Cloud (VPC)

Easily scale your access

Access Points are easy to scale as you build more applications for your large shared data sets

a. Préparation des données

Autorisations

Groupes

Balises (1)

Informations d'identification de sécurité

Access Advisor

▼ Permissions policies (1 stratégie appliquée)

Ajouter des autorisations

+ Ajouter une stratégie en ligne

Nom de la stratégie ▼

Type de stratégie ▼

Attachée directement

▼  AmazonS3FullAccess Stratégie gérée par AWS ×

Récapitulatif de la stratégie

{ } JSON

Simuler la stratégie

Q Filtre

Service ▼	Niveau d'accès	Ressource	Condition de demande
Autoriser (1 de 234 services) Afficher les 233 restants			
S3	Accès complet	Toutes les ressources	Aucun

► Permissions boundary (not set)

a. Préparation des données

oc-p8-salah

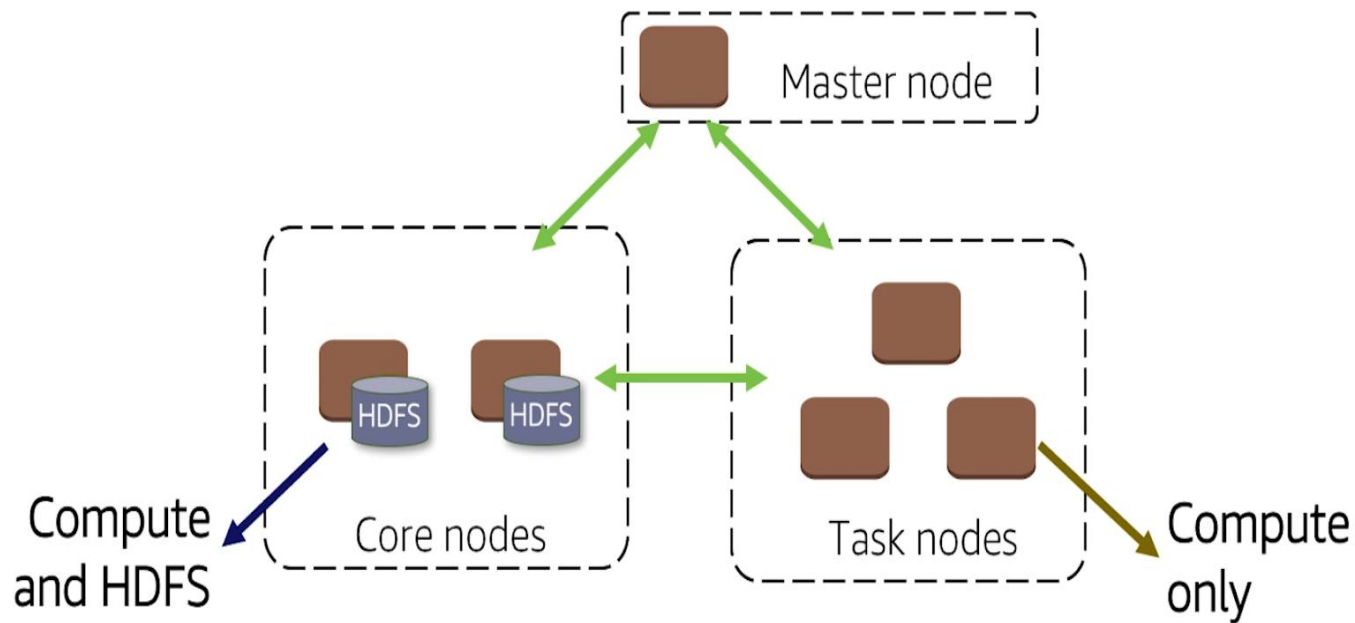
Présentation Propriétés Autorisations Gestion Points d'accès

Q Saisir un préfixe et appuyer sur Entrée pour lancer la recherche. Pour annuler, appuyer sur ESC.

Charger + Créer un dossier Télécharger Actions ▾

<input type="checkbox"/>	Nom ▾
<input type="checkbox"/>	📁 Cherry_Wax_Yellow
<input type="checkbox"/>	📁 Kohlrabi
<input type="checkbox"/>	📁 Mulberry
<input type="checkbox"/>	📁 Peach
<input type="checkbox"/>	📁 Peach_Flat
<input type="checkbox"/>	📁 Pear
<input type="checkbox"/>	📁 Pear_Kaiser

b. Amazon EMR



b. Amazon EMR

- Intérêts:
 - Simple
 - Prix
 - Découplage calcul / stockage
 - Ségrégation des serveurs
 - Remplacement automatique d'instances
 - S3 code propriétaire vs S3a pour EC2

c. Architecture adoptée

Étape 1 : Logiciels et étapes

Étape 2 : Matériel

Étape 3 : Paramètres de cluster généraux

Étape 4 : Sécurité

Configuration des logiciels

Libérer ⓘ

- | | | |
|--|---|---|
| <input checked="" type="checkbox"/> Hadoop 2.8.5 | <input type="checkbox"/> Zeppelin 0.8.2 | <input checked="" type="checkbox"/> Livy 0.7.0 |
| <input checked="" type="checkbox"/> JupyterHub 1.1.0 | <input type="checkbox"/> Tez 0.9.2 | <input type="checkbox"/> Flink 1.10.0 |
| <input checked="" type="checkbox"/> Ganglia 3.7.2 | <input type="checkbox"/> HBase 1.4.13 | <input type="checkbox"/> Pig 0.17.0 |
| <input type="checkbox"/> Hive 2.3.6 | <input type="checkbox"/> Presto 0.232 | <input type="checkbox"/> ZooKeeper 3.4.14 |
| <input type="checkbox"/> MXNet 1.5.1 | <input type="checkbox"/> Sqoop 1.4.7 | <input type="checkbox"/> Mahout 0.13.0 |
| <input type="checkbox"/> Hue 4.6.0 | <input type="checkbox"/> Phoenix 4.14.3 | <input type="checkbox"/> Oozie 5.2.0 |
| <input checked="" type="checkbox"/> Spark 2.4.5 | <input type="checkbox"/> HCatalog 2.3.6 | <input checked="" type="checkbox"/> TensorFlow 1.14.0 |

Prise en charge multimédiaire

c. Architecture adoptée

Networking

Use a Virtual Private Cloud (VPC) to process sensitive data or connect to a private network. Launch the cluster into a VPC with a public, private or shared subnet. Subnets may be associated with and AWS Outpost or AWS Local Zone.

Launch the cluster into a VPC with a public, private, or shared subnet. Subnets may be associated with an AWS Outpost or AWS Local Zone.

Réseau [Créer un VPC](#)

Sous-réseau EC2

Cluster Nodes and Instances

Choisissez le type d'instance, le nombre d'instances et une option d'achat. Vous pouvez choisir d'utiliser les instances à la demande et/ou les instances Spot. Le type d'instance et l'option d'achat s'appliquent à toutes les instances EC2 de chaque groupe d'instances, et vous pouvez uniquement spécifier ces options pour un groupe d'instances lors de sa création. [En savoir plus sur les options d'achat d'instance](#)

Console options for automatic scaling have changed. [Learn more](#)

Node type	Type d'instance	Nombre d'instances	Option d'achat
Maître Groupe d'instances maître - 1	m4.large 4 Cœurs virtuels, 8 GIO de mémoire, stockage EBS uniquement Stockage sur EBS : 32 Gio Ajouter des paramètres de configuration	1 Instances	<input checked="" type="radio"/> A la demande <input type="radio"/> Spot <input type="text" value="Utiliser le prix à la demande comm."/>
Principal Groupe d'instances principal - 2	m4.large 4 Cœurs virtuels, 8 GIO de mémoire, stockage EBS uniquement Stockage sur EBS : 32 Gio Ajouter des paramètres de configuration	<input type="text" value="2"/> Instances	<input checked="" type="radio"/> A la demande <input type="radio"/> Spot <input type="text" value="Utiliser le prix à la demande comm."/>
Tâche Groupe d'instances de tâches - 3	m4.large 4 Cœurs virtuels, 8 GIO de mémoire, stockage EBS uniquement Stockage sur EBS : 32 Gio Ajouter des paramètres de configuration	<input type="text" value="0"/> Instances	<input checked="" type="radio"/> A la demande <input type="radio"/> Spot <input type="text" value="Utiliser le prix à la demande comm."/>

c. Architecture adoptée

Étape 1 : Logiciels et étapes

Étape 2 : Matériel


Étape 3 : Paramètres de cluster généraux

Étape 4 : Sécurité

Options générales

Nom du cluster

☒ Journalisation ⓘ

Dossier S3 

☐ Log encryption ⓘ

☒ Débogage ⓘ

☒ Protection de la résiliation ⓘ

Balises ⓘ

c. Architecture adoptée

Étape 1 : Logiciels et étapes

Étape 2 : Matériel

Étape 3 : Paramètres de cluster généraux

Étape 4 : Sécurité

Options de sécurité

Paire de clés EC2 ⓘ

☒ Cluster visible pour tous les utilisateurs IAM du compte ⓘ

Autorisations ⓘ

☒ Par défaut ☐ Personnalisé

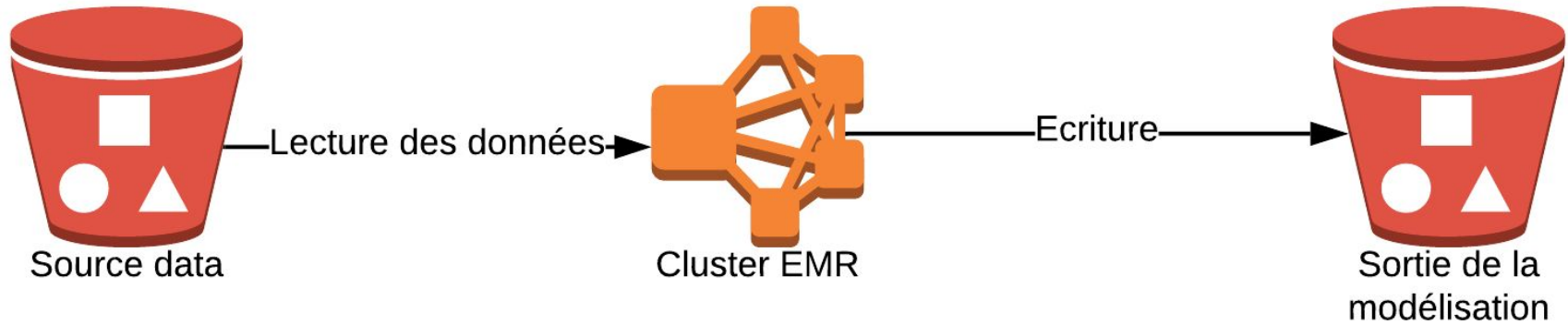
Utilisez les rôles IAM par défaut. Si des rôles sont absents, ils seront créés automatiquement pour vous avec des stratégies gérées pour les mises à jour automatiques de stratégies.

c. Architecture adoptée

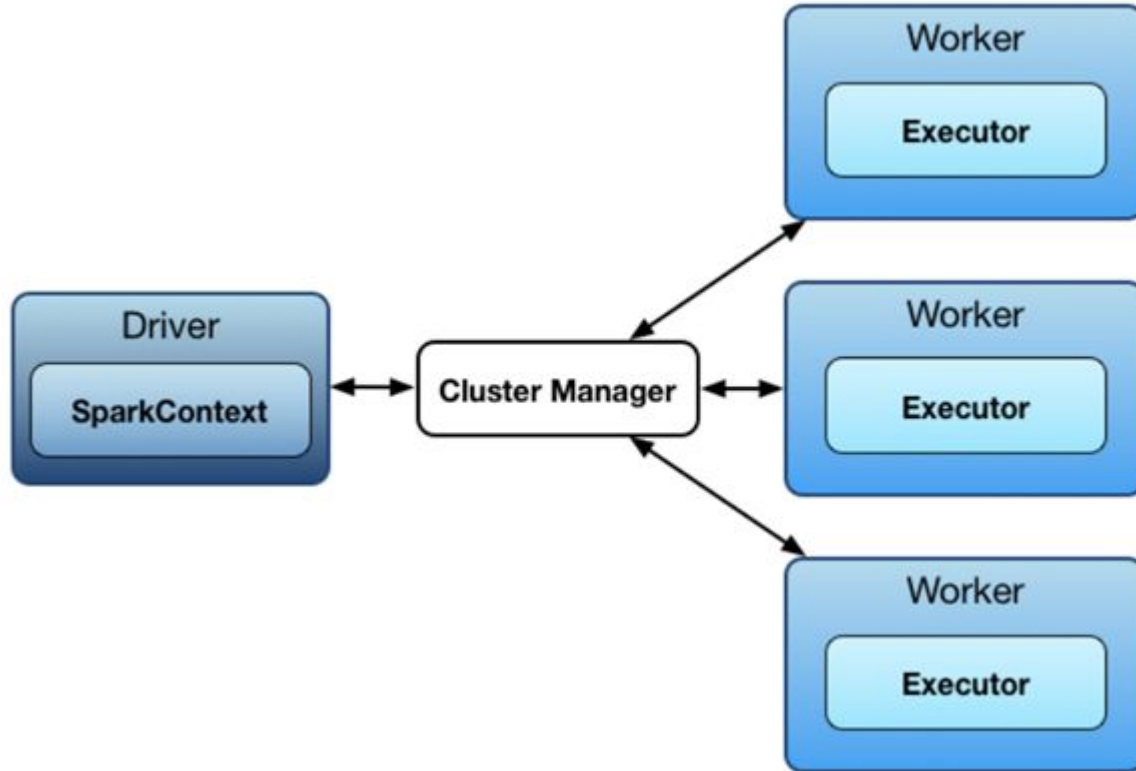
- Configuration de sécurité

TCP personnalisé	TCP	8888
SSH	TCP	22

c. Architecture adoptée

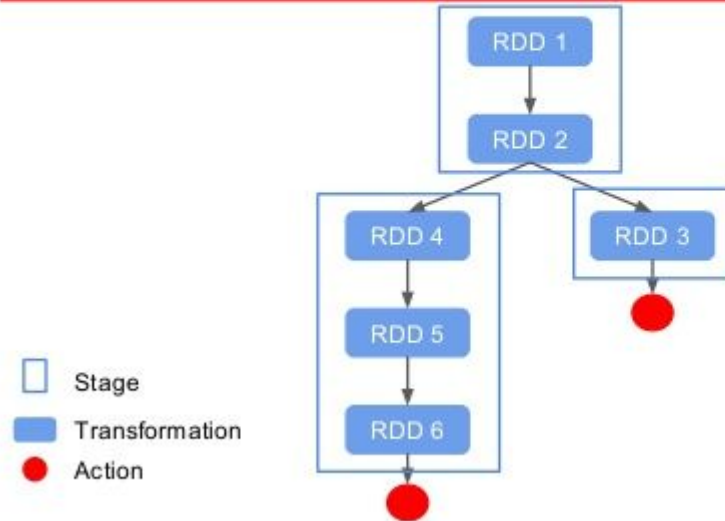


c. Architecture adoptée



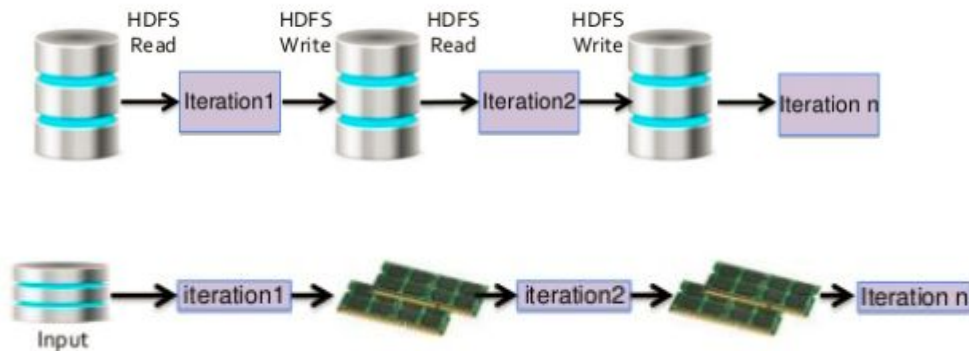
c. Architecture adoptée

Spark DAG



c. Architecture adoptée

- Avantages Spark :
 - In memory
 - Lazy evaluation
- Inconvénients:
 - Plus cher



d. Chaîne de traitement



d. Chaîne de traitement

- SparkContext

```
Entrée [1]: ► sc
Out[1]: SparkContext

Spark UI
Version
v2.4.5-amzn-0
Master
yarn
AppName
PySparkShell
```

d. Chaîne de traitement

- Dataframe d'entrée

```
+-----+-----+
|          image|          label|
+-----+-----+
|[s3a://oc-p8-sala...|Cherry_Wax_Yellow|
|[s3a://oc-p8-sala...|Cherry_Wax_Yellow|
|[s3a://oc-p8-sala...|Cherry_Wax_Yellow|
+-----+-----+
only showing top 3 rows
```

d. Chaîne de traitement

- Extraction des features

```
+-----+-----+-----+
|          image|          label|          feature|
+-----+-----+-----+
|[s3a://oc-p8-sala...|Cherry_Wax_Yellow|[0.15766188502311...|
|[s3a://oc-p8-sala...|Cherry_Wax_Yellow|[0.35756713151931...|
|[s3a://oc-p8-sala...|Cherry_Wax_Yellow|[0.52175581455230...|
+-----+-----+-----+
only showing top 3 rows
```

d. Chaîne de traitement

- Sauvegarde en format parquet

oc-p8-salah





Présentation

🔍 Saisir un préfixe et appuyer sur Entrée pour lancer la recherche. Pour annuler, appuyer sur ESC.

⬇️ Charger + Créer un dossier Télécharger Actions ▾

USA Est (Ohio) ↺

Affichage 1 à 91

<input type="checkbox"/> Nom ▾	Dernière modification ▾	Taille ▾	Classe de stockage ▾
<input type="checkbox"/>  _SUCCESS	juil. 20, 2020 4:29:32 AM GMT+0200	0 o	Standard
<input type="checkbox"/>  part-00000-a1452c2d-77c3-4613-87b7-08d5d30c5959-c000.snappy.parquet	juil. 20, 2020 4:28:58 AM GMT+0200	39.1 Ko	Standard
<input type="checkbox"/>  part-00001-a1452c2d-77c3-4613-87b7-08d5d30c5959-c000.snappy.parquet	juil. 20, 2020 4:28:58 AM GMT+0200	40.0 Ko	Standard
<input type="checkbox"/>  part-00002-a1452c2d-77c3-4613-87b7-08d5d30c5959-c000.snappy.parquet	juil. 20, 2020 4:28:59 AM GMT+0200	40.0 Ko	Standard

Conclusion

Conclusion

- Pour aller plus loin :
 - continuez plus loin sur la modélisation
 - Activer l'auto scaling sur le cluster
 - Utiliser Hadoop au lieu de S3
 - Monitoring
 - Détecter les fruits mûrs, pourris ...
 - Politique de réduction de prix ?