

DÉPLOYEZ UN MODÈLE DANS LE CLOUD

KONAN KOFFI

PLAN DE PRÉSENTATION

I/ Problématique et jeu de données

II/ Étapes de la chaîne de traitement

- *Preprocessing et réduction de dimensions*

III/ Différentes briques d'architecture choisies sur le cloud et leur rôle dans l'architecture Big Data

IV/ Conclusion et recommandation

I PROBLÉMATIQUE ET JEU DE DONNÉES

Start-up de l'AgriTech



Objectifs

➤ I Reconnaissance de fruit



ORANGE

Riche en vitamine C

- *action antioxydante*
- *favorise l'absorption du fer*
- ...

➤ II Introduire de l'IA dans la récolte de fruit tout en respectant la biodiversité des fruits



I PROBLÉMATIQUE ET JEU DE DONNÉES

Problématique

- *Comment pourrait-on traiter de gros volumes de données (volumétrie extrême) ?*

Mission

- *Mise en place de premières briques de traitement des images*
- *Développement de scripts adaptés en termes de volume de données et de vitesse dans un environnement Big Data*



Jeu de données*

- *Un jeu de données avec 90380 images de 131 fruits et légumes*



fruits-360_dataset



fruits-360-original-size

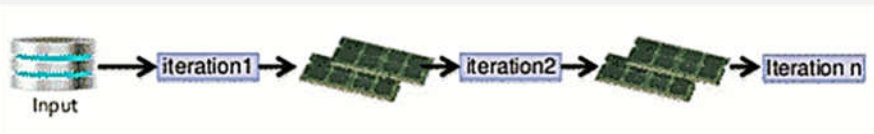
- **Plus d'I Go de Données**

* <https://www.kaggle.com/datasets/moltean/fruits>

II ÉTAPES DE LA CHAÎNE DE TRAITEMENT



- Une infrastructure de calcul de cluster open-source avec des analyses **en mémoire (in-memory)**



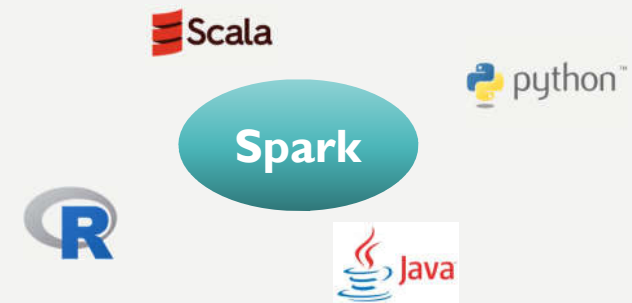
- Plus rapide que certains systèmes informatiques en cluster (Hadoop)
- 100 fois plus vite en mémoire et 10 fois plus vite même sur le disque
- Idéale pour l'analyse de données à grande échelle

<https://meritis.fr/larchitecture-framework-spark/>

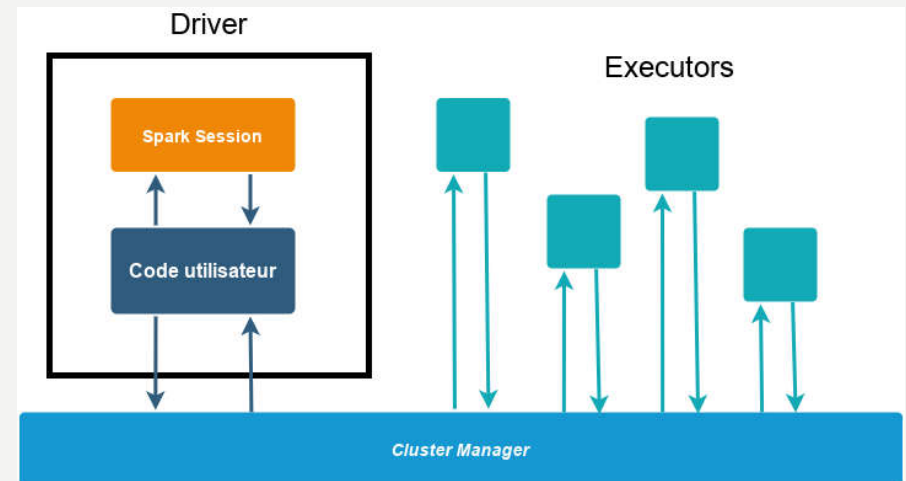
<https://www.lebigdata.fr/apache-spark-tout-savoir>

<https://datascientest.com/apprendre-a-utiliser-lapi-python-pour-spark>

<http://b3d.bdpedia.fr/spark-batch.html>



Architecture globale Apache Spark



II ÉTAPES DE LA CHAÎNE DE TRAITEMENT

Mise en place de la SparkSession dans le notebook Jupyter



API Python pour Spark

SparkSession : commencer à programmer avec
Spark SQL & DataFrame,
Spark Datasets,
Spark MLlib.



DataFrame

```
+-----+-----+-----+-----+-----+-----+
|              origin|height|width|nChannels|mode|              data|
+-----+-----+-----+-----+-----+-----+
|s3a://dmc-in/Trai...|  798|  324|      3|  16|[FF FF FF FF FF F...|
|s3a://dmc-in/Trai...|  797|  325|      3|  16|[FF FF FF FF FF F...|
|s3a://dmc-in/Trai...|  793|  335|      3|  16|[FF FF FF FF FF F...|
|s3a://dmc-in/Trai...|  796|  328|      3|  16|[FF FF FF FF FF F...|
|s3a://dmc-in/Trai...|  787|  339|      3|  16|[FF FF FF FF FF F...|
+-----+-----+-----+-----+-----+
only showing top 5 rows

root
|-- image: struct (nullable = true)
|   |-- origin: string (nullable = true)
|   |-- height: integer (nullable = true)
|   |-- width: integer (nullable = true)
|   |-- nChannels: integer (nullable = true)
|   |-- mode: integer (nullable = true)
|   |-- data: binary (nullable = true)
```

Conversion des code de couleurs et modification de la taille de nos images à 224 × 224

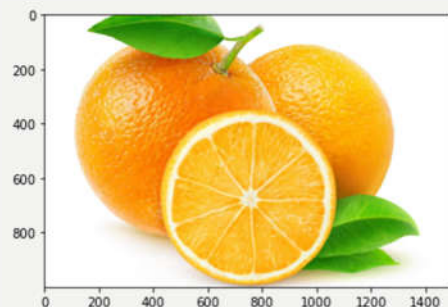
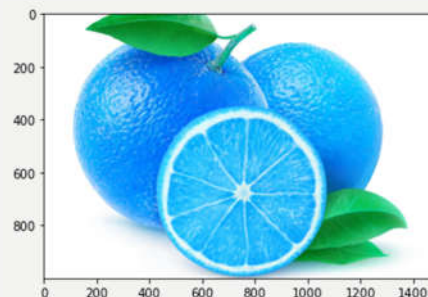


Image réelle (RGB)

Spark



Entrée
attendue :
BGRA



**Image bleue
après traitement Spark**

Transformations

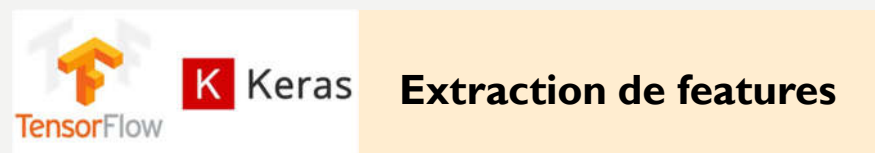


- RGB → BGR
- Échelle réduite



224 × 224

II ÉTAPES DE LA CHAÎNE DE TRAITEMENT



Vectorisation des features d'images

Transfer Learning via ResNet50 model

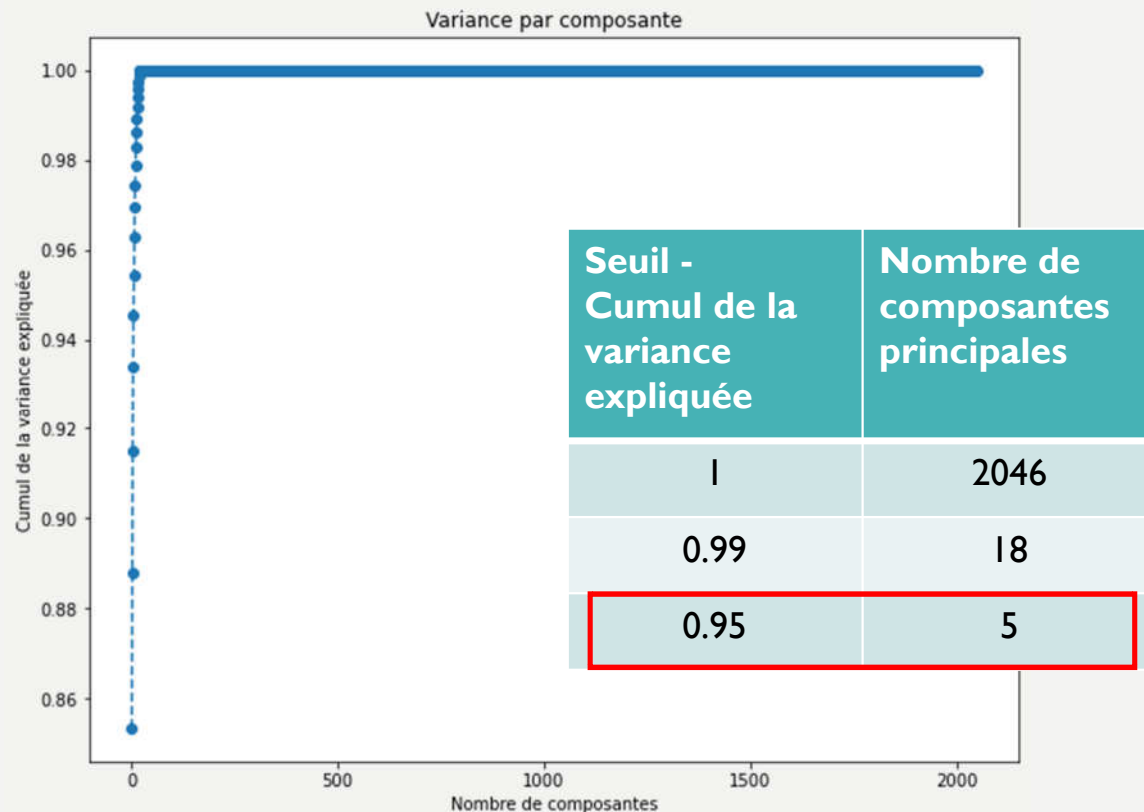


- *Suppression de la dernière couche du réseau de neurones*
- *Récupération des poids*

- *Vecteurs denses*
- *Nécessaire pour MLlib*

II ÉTAPES DE LA CHAÎNE DE TRAITEMENT

Réduction de dimensionnalité PCA (intégré à Spark)



DataFrame avec PCA(features) + Label

pca_features	label
[14.5507688986928...	cucumber_1
[14.6421758454457...	cucumber_1
[14.0862501664891...	cucumber_1
[15.6812748471574...	cucumber_1
[12.4223389816891...	cucumber_1

only showing top 5 rows

Stockage de DataFrame

III DIFFÉRENTES BRIQUES D'ARCHITECTURE CHOISIES SUR LE CLOUD ET LEUR RÔLE DANS L'ARCHITECTURE BIG DATA

Architecture Big Data AWS

- basée sur un serveur EC2 Linux



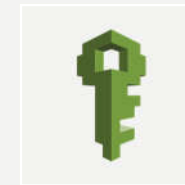
S3

- *Stockage de données*



EC2

- *Calculs*
- *Déploiement de modèle*



IAM

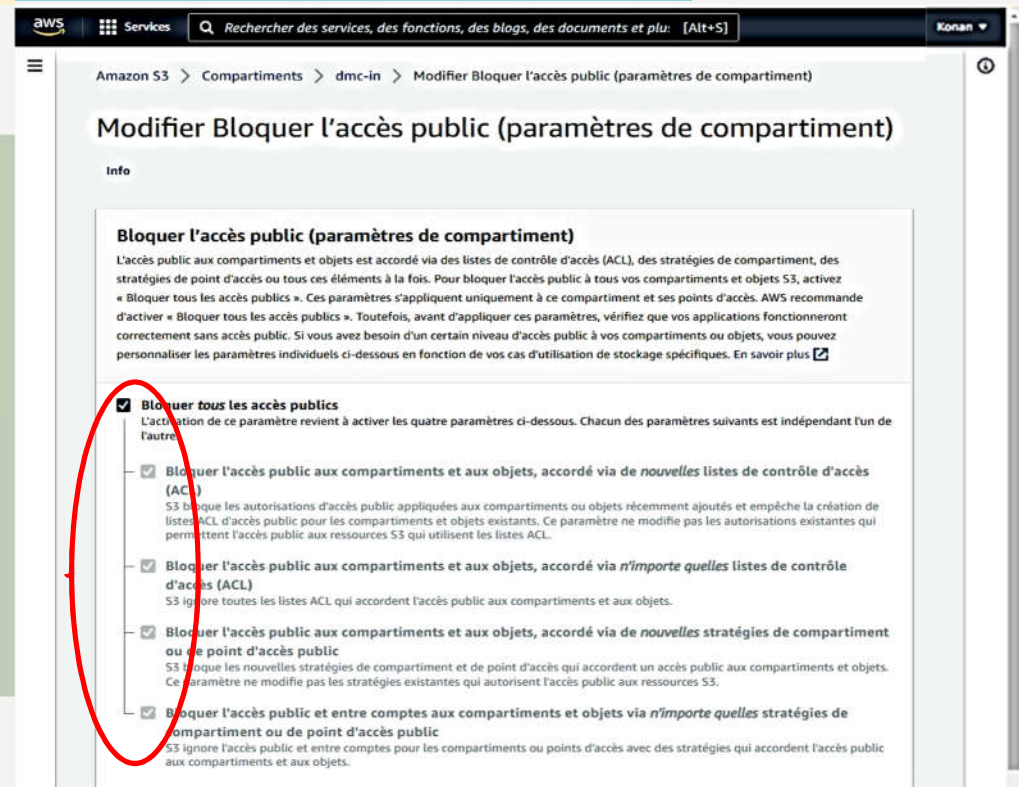
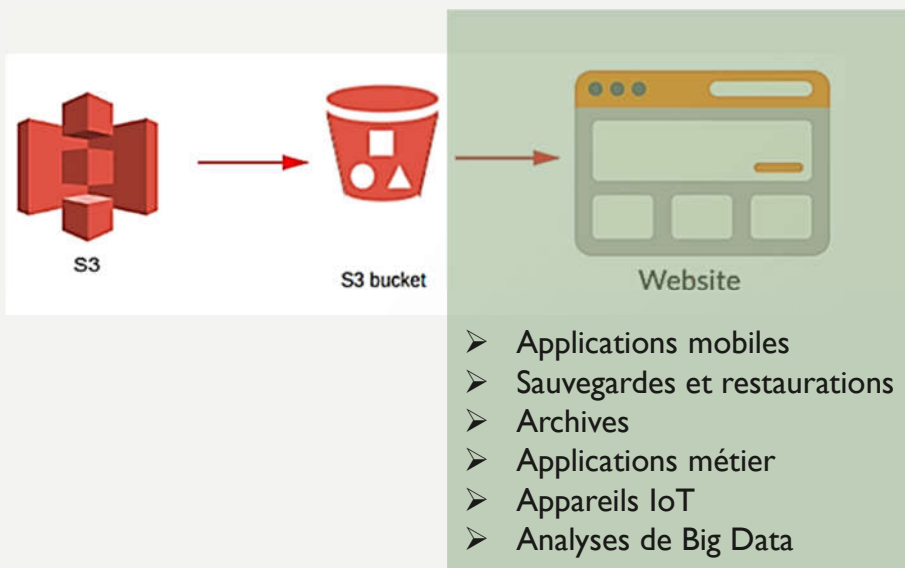
- *Control d'accès*

III DIFFÉRENTES BRIQUES D'ARCHITECTURE CHOISIES SUR LE CLOUD ET LEUR RÔLE DANS L'ARCHITECTURE BIG DATA

S3 (Simple Storage Service)

Évolutivité de stockage d'objets,
Disponibilité des données

Sécurité et des performances de pointe



III DIFFÉRENTES BRIQUES D'ARCHITECTURE CHOISIES SUR LE CLOUD ET LEUR RÔLE DANS L'ARCHITECTURE BIG DATA



S3, Buckets et Objets

- Fonctions de gestion pour **optimisation**, **organisation** et pour **configuration** de l'accès aux données (des structurées aux non-structurées)
- Répondre à des exigences spécifiques

aws Services Rechercher des services, des fonctions, des blogs, des documents et plu. [Alt+S]

Amazon S3 > Compartiments

Instantané de compte Afficher le tableau de bord de Storage Lens

Storage Lens offre une visibilité sur l'utilisation du stockage et les tendances d'activité. En savoir plus

Compartiments (3) Info Copier l'ARN Vider Supprimer Créer un compartiment

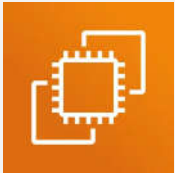
Rechercher des compartiments par nom

Nom	Région AWS	Accéder	Date de création
dmc-in	EU (Paris) eu-west-3	Compartiment et objets non publics	29 Aug 2022 11:41:07 AM CEST
dmc-out	EU (Paris) eu-west-3	Compartiment et objets non publics	29 Aug 2022 12:00:38 PM CEST
mawsbucket-test	EU (Paris) eu-west-3	Compartiment et objets non publics	03 Sep 2022 05:08:07 PM CEST

Stockage de Données d'images

Stockage Features d'images après la réduction de dimension PCA

III DIFFÉRENTES BRIQUES D'ARCHITECTURE CHOISIES SUR LE CLOUD ET LEUR RÔLE DANS L'ARCHITECTURE BIG DATA



(EC2)

(Elastic Compute Cloud)

Servers pour exécuter des applications web, ...

Offre IaaS d'Amazon Web Services à l'usage courant.

aws Services Rechercher des services, des fonctions, des blogs, des documents et plus [Alt+S]

Démarrage rapide

Amazon Linux
aws

Ubuntu
ubuntu

Windows
Microsoft

Red Hat
Red Hat

SUSE Linux
SUSE

Parcourir d'autres AMI
Y compris des AMI provenant d'AWS, de Marketplace et de la communauté

Amazon Machine Image (AMI)

Ubuntu Server 22.04 LTS (HVM), SSD Volume Type Eligible à l'offre gratuite

ami-09e513e9eacab10c1 (64 bits (x86)) / ami-0d86e312c6fb92cd5 (64 bits (Arm))
Virtualisation: hvm ENA activée: true Type d'appareil racine: ebs

Description
Canonical, Ubuntu, 22.04 LTS, amd64 jammy image build on 2022-06-09

Architecture 64 bits (x86) ID de l'AMI ami-09e513e9eacab10c1

Type d'instance Informations

Type d'instance
t2.medium
Famille: t2 2 vCPU 4 Gio Mémoire
À la demande Linux tarification: 0.0528 USD par heure
À la demande Windows tarification: 0.0708 USD par heure

Comparer les types d'instance

Paire de clés (connexion) Informations

Vous pouvez utiliser une paire de clés pour vous connecter en toute sécurité à votre instance. Assurez-vous d'avoir accès à la paire de clés sélectionnée avant de lancer l'instance.

Nom de la paire de clés - obligatoire
Sélectionner

Créer une paire de

Résumé

Nombre d'instances Informations
1

Image logicielle (AMI)
Canonical, Ubuntu, 22.04 LTS, ...en savoir plus
ami-09e513e9eacab10c1

Type de serveur virtuel (type d'instance)
t2.medium

Pare-feu (groupe de sécurité)
Nouveau groupe de sécurité

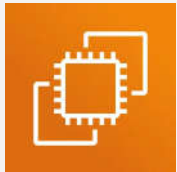
Stockage (volumes)
1 volume(s) - 30 Gio

Offre gratuite : La première année inclut 750 heures d'utilisation mensuelle des instances t2.micro (ou t3.micro dans les régions où t2.micro n'est pas disponible) sur les AMI de l'offre gratuite, 30 Gio de stockage EBS, 2 millions d'I/O, 1 Go d'instantanés et 100 Go de bande passante vers Internet

Annuler Lancer l'instance

Elastic Block Store :
Périphérique racine
Stockage du système
d'exploitation

III DIFFÉRENTES BRIQUES D'ARCHITECTURE CHOISIES SUR LE CLOUD ET LEUR RÔLE DANS L'ARCHITECTURE BIG DATA



(EC2)

Création de l'instance EC2 server Linux

Création obligatoire d'une paire de clés publique/privée [SSH](#) ?

- Protocole de **connexions** sécurisées entre deux systèmes

Clé privée



Clé publique

Architecture client/server

Clé privée



Clé publique

Échange de clés sécurisées



Local



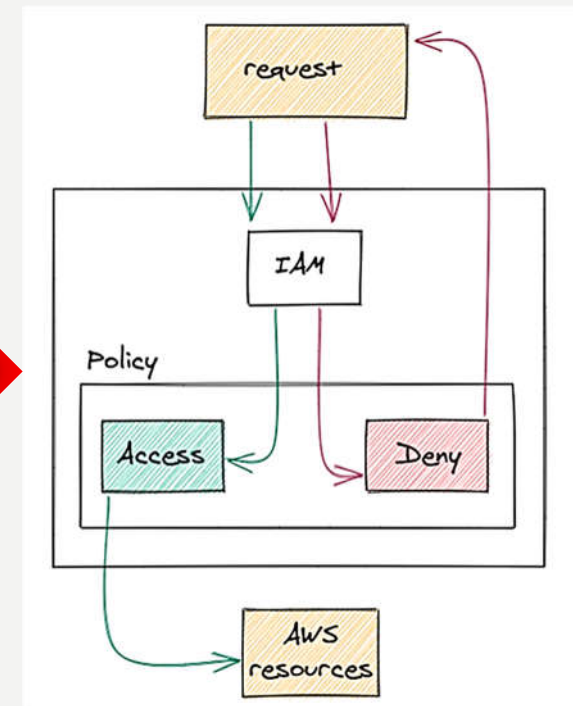
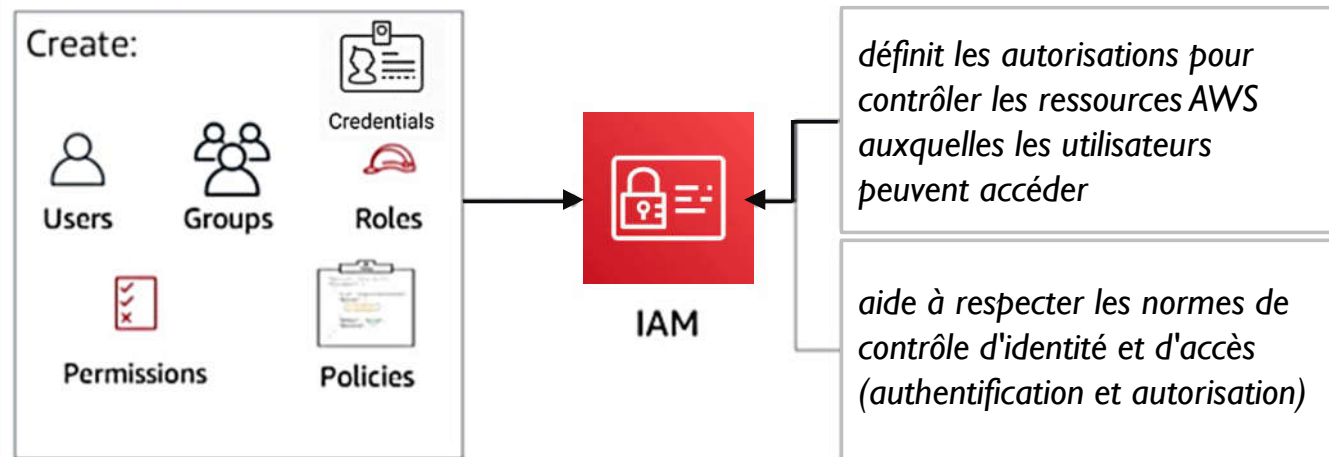
Systèmes hôte
servers **AWS**
EC2 Linux

III DIFFÉRENTES BRIQUES D'ARCHITECTURE CHOISIES SUR LE CLOUD ET LEUR RÔLE DANS L'ARCHITECTURE BIG DATA



IAM (Identity and Access Management)

Rappel : Service de gestion des identités et des accès d'AWS



III DIFFÉRENTES BRIQUES D'ARCHITECTURE CHOISIES SUR LE CLOUD ET LEUR RÔLE DANS L'ARCHITECTURE BIG DATA



IAM Role

Tableau de bord de la stratégie attachée au rôle IAM

IAM > Rôles > ec2-s3-full-access

ec2-s3-full-access

Allows EC2 instances to call AWS services on your behalf.

Récapitulatif

Date de création September 02, 2022, 20:54 (UTC+02:00)	ARN arn:aws:iam::388483445354:role/ec2-s3-full-access	ARN de profil d'instance arn:aws:iam::388483445354:instance-profile/ec2-s3-full-access
Dernière activité Il y a 2 jours	Durée maximale de la session 1 heure	

Autorisations Relations d'approbation Balises Access Advisor Révoquer les séances

Politiques des autorisations (1)
Vous pouvez attacher jusqu'à 10 politiques gérées.

Filter les stratégies par nom de propriété ou de stratégie et appuyer sur Entrée

<input type="checkbox"/>	Nom de la politique	Type	Description
<input type="checkbox"/>	AmazonS3FullAccess	Gérées par AWS	Provides full access to all buckets via the AWS Management Console.

Limite d'autorisations - (not set)
Définissez une limite d'autorisations pour contrôler le nombre maximum d'autorisations que ce rôle peut avoir. Il ne s'agit pas d'un paramètre courant, mais il peut être utilisé pour déléguer la gestion des autorisations à d'autres utilisateurs.

Définir une limite d'autorisations

Autorisation d'un accès total aux ressources S3

III DIFFÉRENTES BRIQUES D'ARCHITECTURE CHOISIES SUR LE CLOUD ET LEUR RÔLE DANS L'ARCHITECTURE BIG DATA

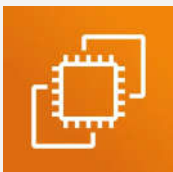


Tableau de bord de notre instance EC2 et configuration du notebook Jupyter sur AWS

aws Services Rechercher des services, des fonctions, des blogs, des documents et plu: [Alt+S]

Instances (1/1) Informations

Se connecter État de l'instance Actions Lancer des instances

Recherche

Name	ID d'instance	État de l'instance	Type d'instance	C...	Statut d'alar...	Nom du groupe de s...	Heure de lancement
notebook-dev...	i-0b3f3bfb28a07da06	Arrêt(e)	t2.medium	-	Aucune al...	jupyter-notebook	2022/09/05 17:09 GMT+2

Instance : i-0b3f3bfb28a07da06 (notebook-dev-instance)

Détails Sécurité Mise en réseau Stockage Vérifications de statut Surveillance Balises

▼ Détails de sécurité

Rôle IAM
ec2-s3-full-access

ID du propriétaire
388483445354

Heure de lancement
Mon Sep 05 2022 17:09:00 GMT+0200 (heure d'été d'Europe centrale)

Groupes de sécurité
sg-0c0227f674ed7fe0c (jupyter-notebook)

▼ Règles entrantes

ID de règle du groupe de ...	Plage de ports	Protocole	Source	Groupes de sécurité
sgr-0cbf5f4567a3aede1	8888	TCP	0.0.0.0/0	jupyter-notebook
sgr-0d0374e19d53c7470	443	TCP	0.0.0.0/0	jupyter-notebook
sgr-0790f2f9cf35f2e03	22	TCP	0.0.0.0/0	jupyter-notebook

▼ Règles sortantes

ID de règle du groupe de ...	Plage de ports	Protocole	Destination	Groupes de sécurité
sgr-0a53efc0c449a75f7	Tout	Tout	0.0.0.0/0	jupyter-notebook

+
Clé privée
(.pem)

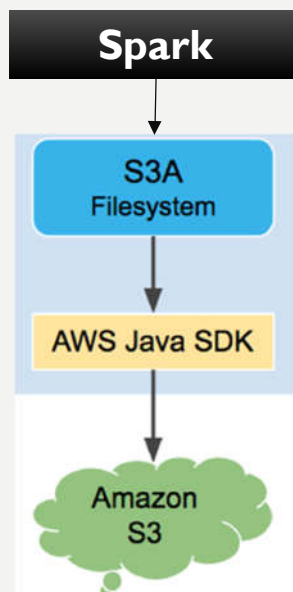
III DIFFÉRENTES BRIQUES D'ARCHITECTURE CHOISIES SUR LE CLOUD ET LEUR RÔLE DANS L'ARCHITECTURE BIG DATA

Communication entre Spark et AWS S3

Autorisation d'accès aux fichiers dans S3 à EC2

Complément de dépendances dans le répertoire Spark

Dépendances java « *module.jar* » :



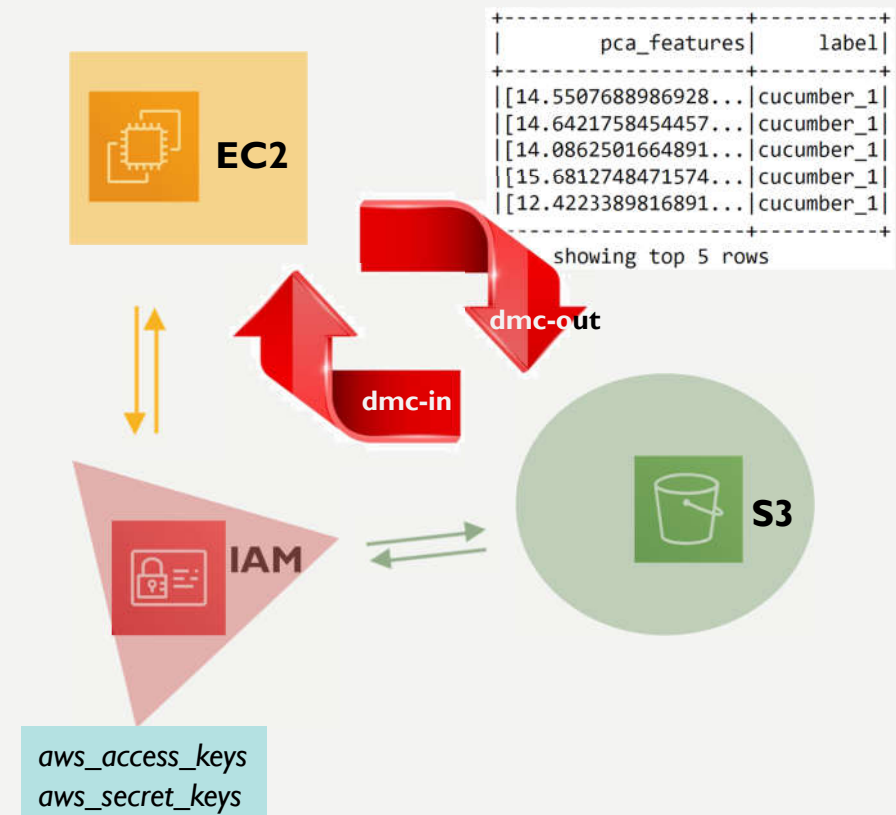
hadoop-aws-3.3.1.jar

- Intégration de Spark (et autres applications de l'écosystème Hadoop) avec **AWS**

Le connecteur S3A permet de lire et d'écrire des fichiers stockés dans Amazon S3 grâce avec une URL de préfixe : `s3a://`

aws-java-sdk-bundle-1.11.901.jar

- Prise en compte du cycle de vie des API,
- Gestion des informations d'identification, les tentatives, l'organisation des données et la sérialisation



IV/ CONCLUSION ET RECOMMANDATION

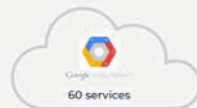
Tout projet qui consiste à étudier des millions voire des milliards de données doit judicieusement se tourner vers des outils adaptés au Big Data.

➤ Scripts



Une place de choix dans le calcul distribué accélérant principalement la vitesse d'exécution de traitement sur des bases de données d'une extrême volumétrie.

➤ Les ressources Big Data



Recommandation au traitement de données évolutives

➤ Amazon EC2 + **Auto Scaling**

Bon complément à
l'équilibrage de charge

➤ Amazon EMR (Elastic Map Reduce) mise en place automatique d'un cluster Spark qui gère naturellement l'auto scaling en fonction de son paramétrage.

Merci pour votre attention