Parcours Data Scientist | Projet 8

Fruits!

Déployer un modèle dans le cloud

Présenté par Fiacre KAKPO, novembre 2021

Plan de travail

Fruits!

I. Contexte et objectifs

II. Présentation des données disponibles

III. AWS : Architecture cloud

IV. Traitement des images et réduction dimensionnelle

V. Pistes d'amélioration et conclusions

Contexte & Objectifs

Contexte & Objectifs

Fruits!

L'entreprise : Fruits !

- ➔ Start-up de l'Agritech
- ➔ application de reconnaissance d'images
- ➔ Croissance rapide des volumes de données

Déploiement d'un environnement "Big Data"

- ➔ Prétraitements & Réduction dimensionnelle
- ➔ Elasticité du cluster de traitement
- ➔ Accessibilité des données et résultats dans le cloud

Présentation du jeu de données

Fruits!

Caractéristiques du jeu de données

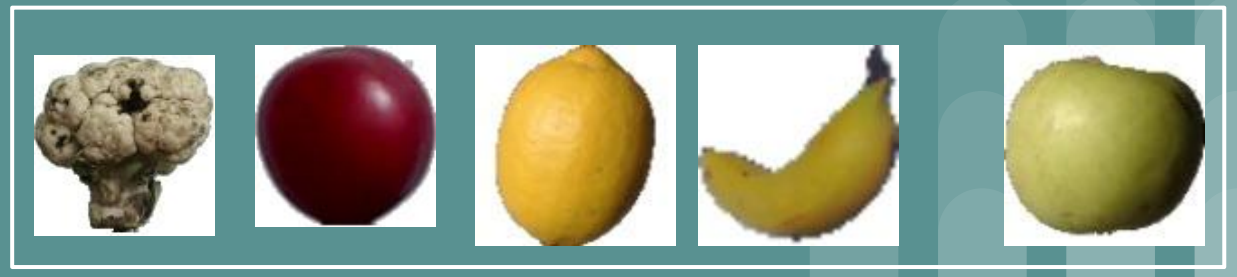
Le jeu de données contient :

- 90483 images de fruits sur fond blanc téléchargeables à ce lien : https://www.kaggle.com/moltean/fruits
- classées en 131 variétés de fruits et légumes
- Résolution : 100X100 pixels
- Des images prise sous plusieurs angles
- Format .jpg

Travail sur l'échantillon

Choix de cinq classes d'images pour le test de l'architecture big data pour des raisons de coûts (énergetiques et financiers)

(cohérence avec les valeurs de l'entreprise)



L’Architecture Big Data avec Amazon Web Services

Amazon Elastic Compute Cloud (Amazon EC2)

Fruits!

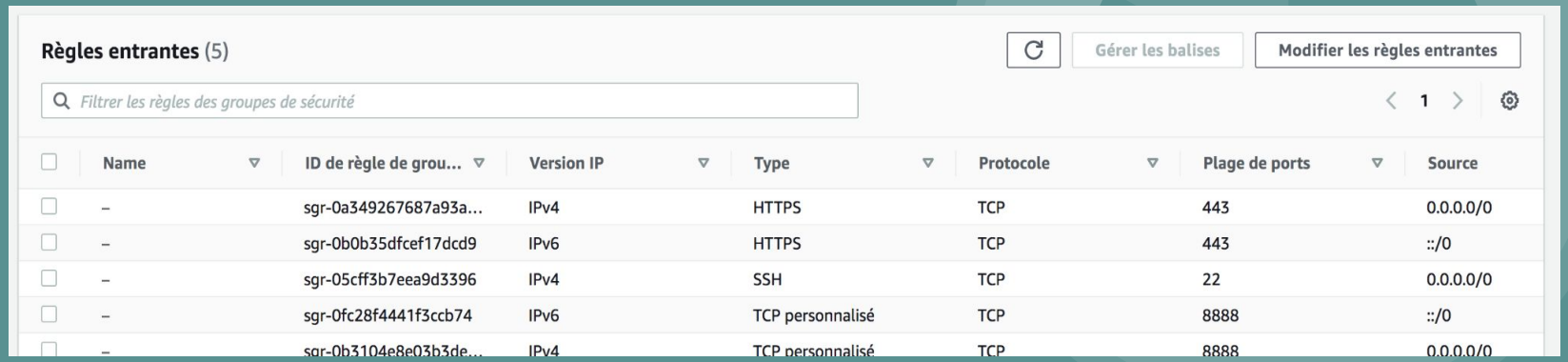
- ➔ Création et lancement d'une instance adaptée à aux contraintes techniques :
  - t2.xlarge
  - RAM: 16 Go
  - 4 CPUs
- ➔ Connexion via tunnel SSH (clé privée .pem)
- ➔ Installation des différents outils nécessaires au projet (anaconda, Spark, Pyspark, java etc.);
- ➔ Connexion à jupyter notebook via un mot de passe et début du script en pyspark.

Groupe de sécurité et accès

Fruits!

amazon web services EC2

Mise en place du groupe de sécurité



Règles entrantes (5)

Gérer les balises | Modifier les règles entrantes

Filtrer les règles des groupes de sécurité

1

Name	ID de règle de grou...	Version IP	Type	Protocole	Plage de ports	Source
–	sgr-0a349267687a93a...	IPv4	HTTPS	TCP	443	0.0.0.0/0
–	sgr-0b0b35dfcef17dcd9	IPv6	HTTPS	TCP	443	::/0
–	sgr-05cff3b7eea9d3396	IPv4	SSH	TCP	22	0.0.0.0/0
–	sgr-0fc28f4441f3ccb74	IPv6	TCP personnalisé	TCP	8888	::/0
–	sgr-0b3104e8e03b3de...	IPv4	TCP personnalisé	TCP	8888	0.0.0.0/0

Mise en place du rôle IAM (identity and access management)

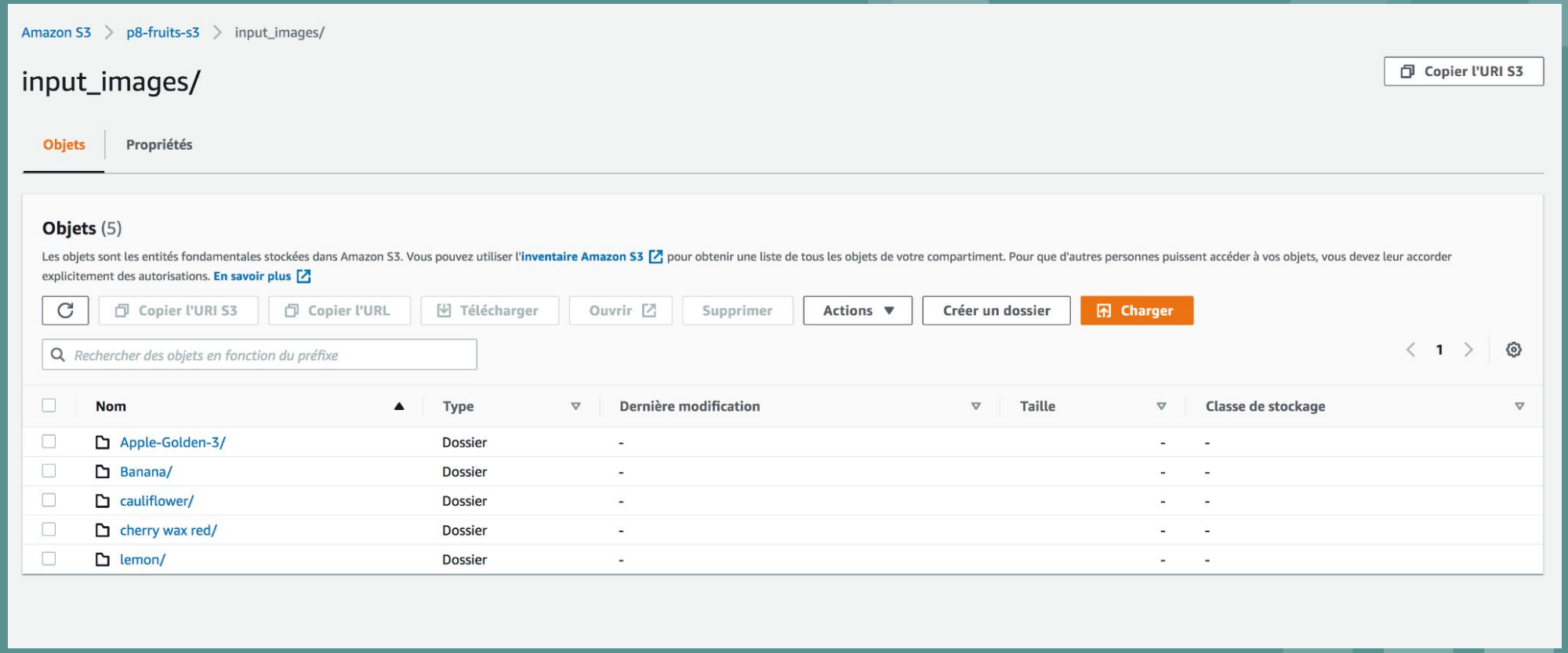
Nom de la stratégie	Type de stratégie
AmazonEC2FullAccess	Stratégie gérée par AWS
AmazonS3FullAccess	Stratégie gérée par AWS

Amazon Simple Storage Service (S3)

Fruits!

Stockage d'un échantillon d'images sur le S3 (Simple Storage Service)

Chargement de trois dossiers de tests (5 types de fruits) comportant chacun 10 images.

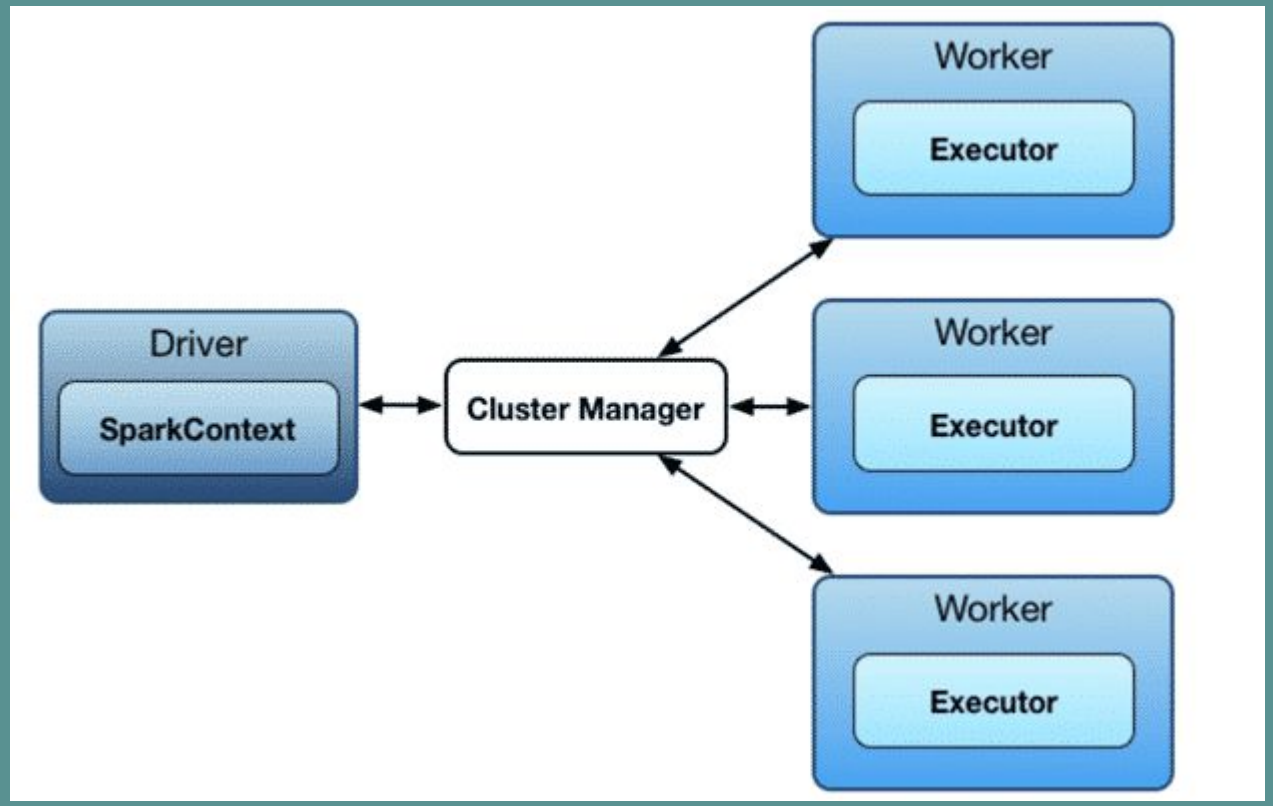


Fonctionnement général de Spark

Fruits!

Plateforme (framework) Opensource multi-langage et ensemble de bibliothèques pour le traitement parallélisé de données sur des grappes (clusters) d'ordinateurs.

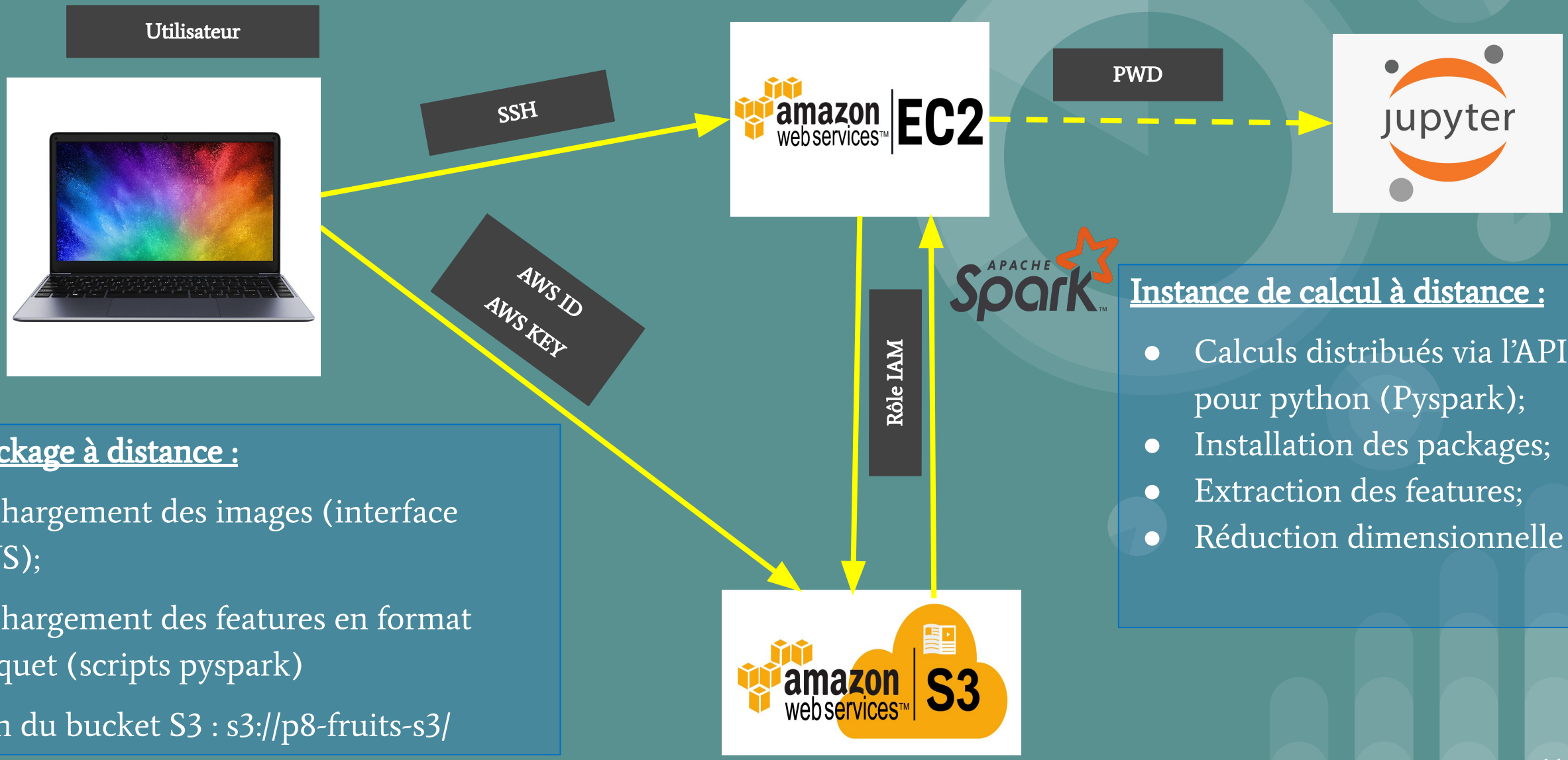
Architecture de Spark



- ➔ Le driver utilise SparkContext pour se connecter au cluster manager et lui soumettre des tâches;
- ➔ Le cluster manager alloue les ressources aux workers et contrôle l'exécution des tâches;
- ➔ Un Worker est constitué d'un ou plusieurs exécuteurs;
- ➔ L' exécuteur exécute le code qui lui est assigné par le driver (via le CM) et lui rapporte l'état d'avancement de la tâche.

Architecture globale de l'environnement Big Data

Fruits!



Stockage à distance :

- Chargement des images (interface AWS);
- Chargement des features en format parquet (scripts pyspark)

Lien du bucket S3 : s3://p8-fruits-s3/

Instance de calcul à distance :

- Calculs distribués via l'API pour python (Pyspark);
- Installation des packages;
- Extraction des features;
- Réduction dimensionnelle

Traitement des images et réduction dimensionnelle

Headlines processing & Réduction dimensionnelle

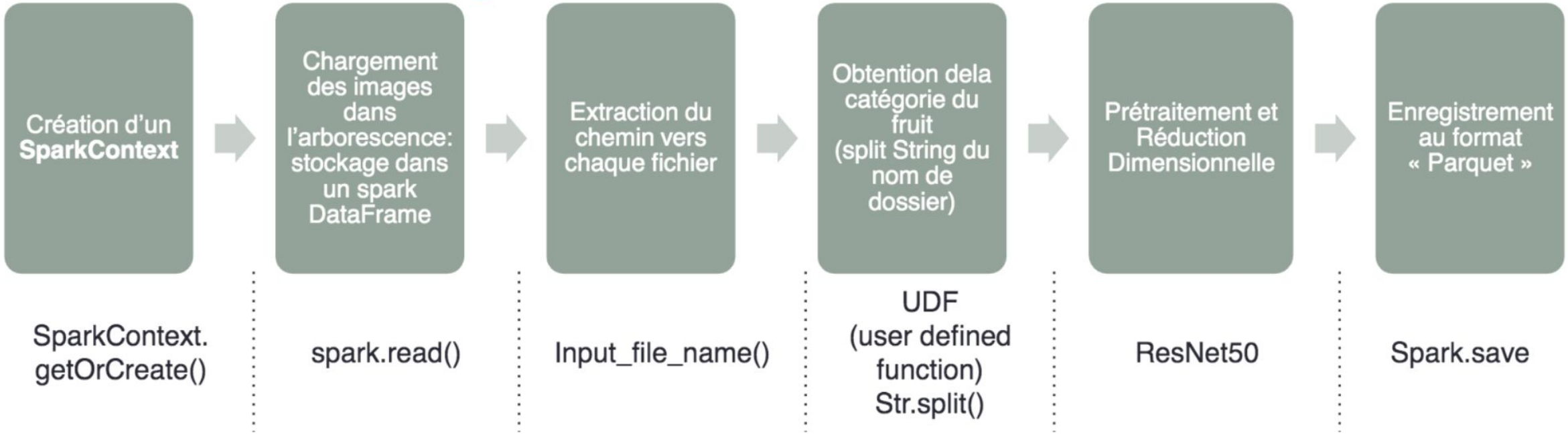
Fruits!

- Choix d'un échantillon de cinq classes (Pomme, banane et citron, cerise, chou fleur);
- Choix de dix images par classe pour le traitement
- Scripts Pyspark construits avec l'échantillon d'images (5 types de fruits, 10 images choisis à la volée pour chacun)
- Choix du réseau de neurones résiduels (ResNet) Resnet50 pour l'extraction de features.
- Réduction dimensionnelle avec PCA (Principal Component Analysis)

Traitement des images

Fruits!

Instance Spark



Extraction des features avec Resnet 50

Fruits!

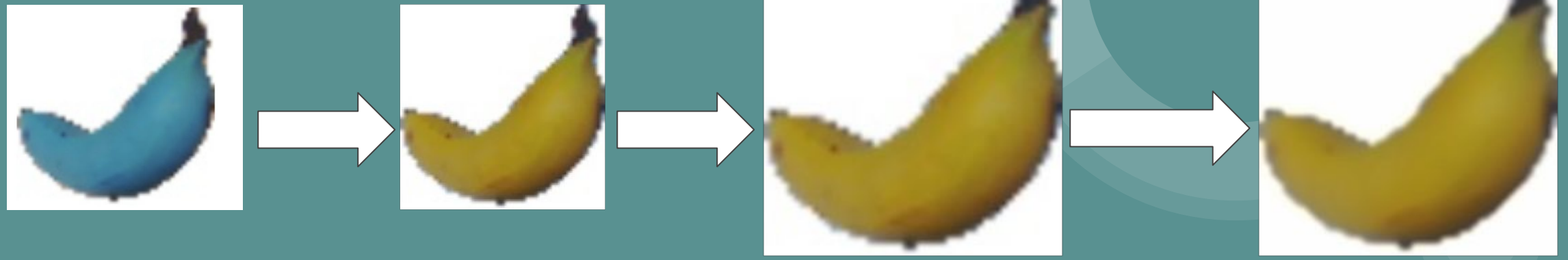


Image au format BGR

Conversion en RGB

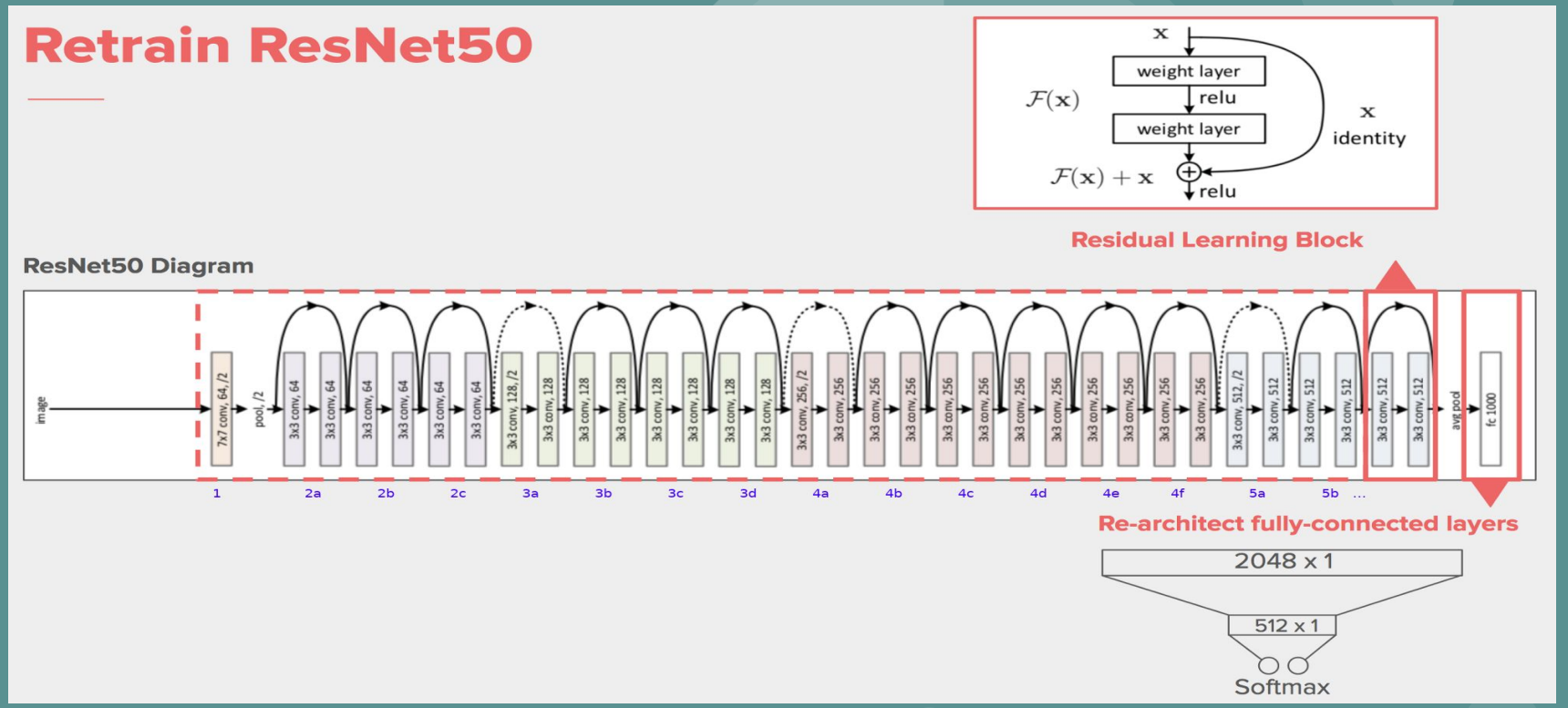
Redimensionnement 224 X 224

Réduction des bruits

Extraction des features avec Resnet 50

Fruits!

- Resnet50 est un réseau de neurones convolutifs résiduel à 50 couches.
- Pré-entraîné sur la base de données ImageNet (14 millions d'images).
- Input : image en format 224X224
- Output : vecteur de 2048 dimensions



Dans le cadre d'une extraction de features, on supprime la dernière couche : **fully connected** (couche de classification)

```
model = ResNet50(include_top=False)
```

Réduction dimensionnelle avec PCA

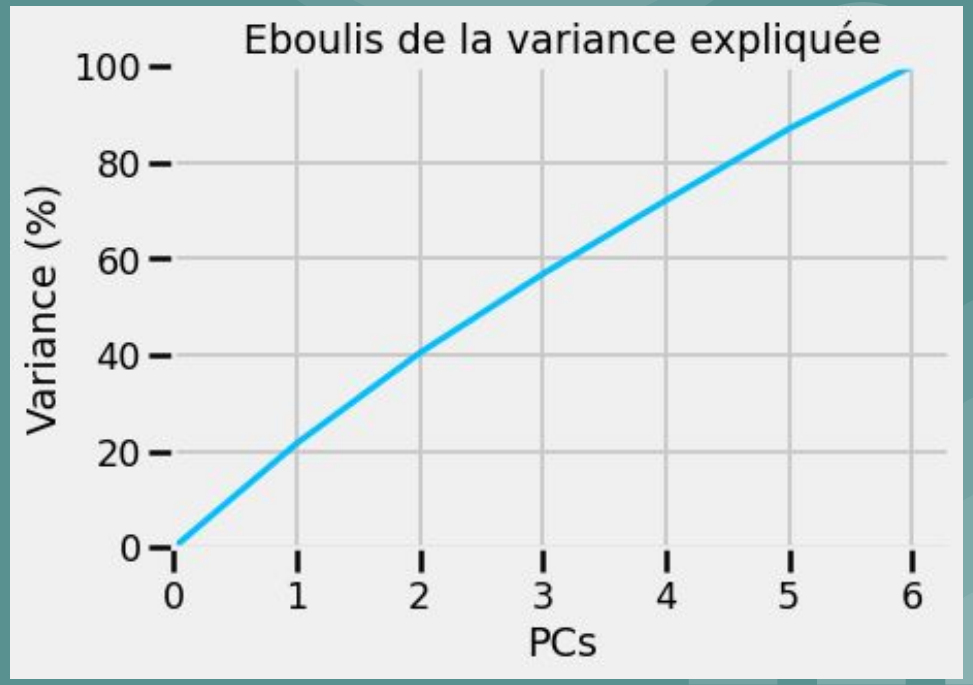
Features issus de l'extraction via ResNet50 passées au PCA après standardisation

path	label	features
s3a://p8-fruits-s...	lemon	[0.0, 0.0, 0.0, 0...
s3a://p8-fruits-s...	lemon	[0.0, 0.0, 0.0, 0...
s3a://p8-fruits-s...	lemon	[0.0, 0.0, 0.0, 0...
s3a://p8-fruits-s...	lemon	[0.0, 0.0, 0.0, 0...
s3a://p8-fruits-s...	Apple-Golden-3	[0.0, 0.0, 0.0, 0...
s3a://p8-fruits-s...	lemon	[0.0, 0.0, 0.0, 0...
s3a://p8-fruits-s...	lemon	[0.0, 0.0, 0.0, 0...
s3a://p8-fruits-s...	lemon	[0.0, 0.0, 0.0, 0...
s3a://p8-fruits-s...	lemon	[0.0, 0.0, 0.0, 0...
s3a://p8-fruits-s...	lemon	[0.0, 0.0, 0.0, 0...
s3a://p8-fruits-s...	Apple-Golden-3	[0.0, 0.0, 0.0, 0...
s3a://p8-fruits-s...	lemon	[0.0, 0.0, 0.0, 0...
s3a://p8-fruits-s...	lemon	[0.0, 0.0, 0.0, 0...
s3a://p8-fruits-s...	lemon	[0.0, 0.0, 0.0, 0...
s3a://p8-fruits-s...	lemon	[0.0, 0.0, 0.0, 0...
s3a://p8-fruits-s...	lemon	[0.0, 0.0, 0.0, 0...
s3a://p8-fruits-s...	lemon	[0.0, 0.0, 0.0, 0...
s3a://p8-fruits-s...	Apple-Golden-3	[0.0, 0.0, 0.0, 0...
s3a://p8-fruits-s...	Apple-Golden-3	[0.0, 0.0, 0.0, 0...

only showing top 20 rows

```
pca = PCA(k=6, inputCol="feats_scaled", outputCol="pca")
modelpca = pca.fit(features_df_scaled)
transformed = modelpca.transform(features_df_scaled)
```

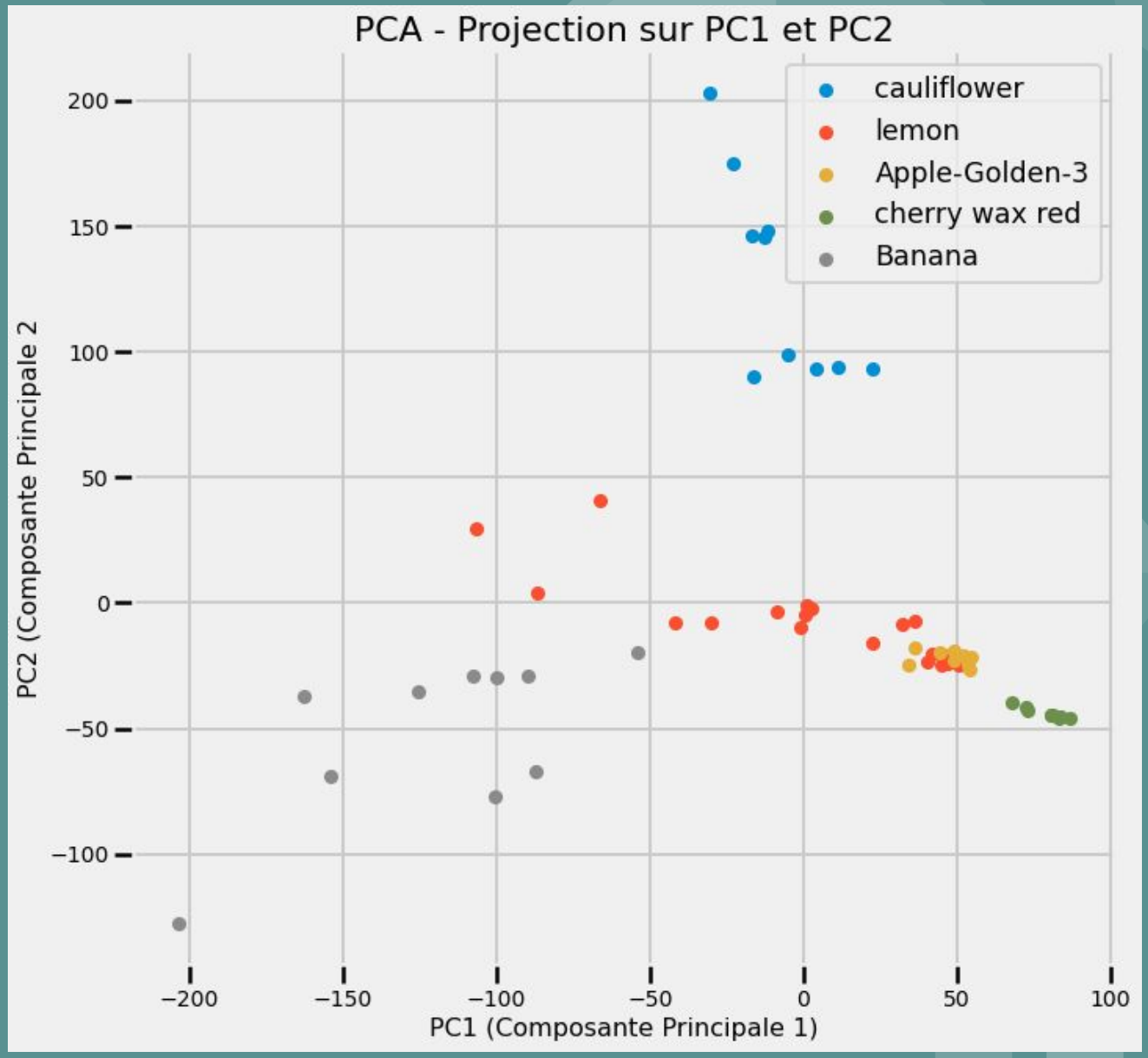
Eboulis de la variance expliquée : environ **6 composants** permettent d'expliquer plus de 95-98% de la variance.



Visualisation des résultats du PCA

Fruits!

- ➔ On observe un bon regroupement des images d'une même espèces de fruit.
- ➔ Certaines espèces particulièrement bien séparées des autres (Cauliflower - Chou-fleur, Banane), et framboise) d'autres plus mélangées (citron, Cerise et pomme)



Conclusion et perspectives

Conclusions et Perspectives

Fruits!

Notions apprises ou couvertes

- Prise en main de Spark et Pyspark
- Découverte de l'écosystème AWS
- Administration d'un serveur Linux par SSH

Perspectives

- Test avec plus d'images ! Relancer le script avec un échantillon nettement plus grand en changeant d'instance EC2 si nécessaire.
- Utiliser t-SNE à la place de PCA pour la réduction dimensionnelle ?
- Utiliser un service EMR (Elastic MapReduce) à la place d'EC2 ?

Fruits!

Merci !!!!

Annexes

Connexion à EC2 et Jupyter Notebook

Fruits!

```
Last login: Sat Nov 27 08:07:56 on ttys001
(base) MacBook-Pro-de-wick:~ macbookproal$ cd Desktop/KeyAWSEC2/
(base) MacBook-Pro-de-wick:KeyAWSEC2 macbookproal$ sudo ssh -i "p8_key_ec2.pem" ec2-user@ec2-34-244-171-139.eu-west-1.compute.amazonaws.com
Password:
The authenticity of host 'ec2-34-244-171-139.eu-west-1.compute.amazonaws.com (34.244.171.139)' can't be established.
ECDSA key fingerprint is SHA256:/tgWj2eOWg6mQx0iAr2Md0RmlS9NXT9qKCYkIDHUVrM.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'ec2-34-244-171-139.eu-west-1.compute.amazonaws.com,34.244.171.139' (ECDSA) to the list of known hosts.
Last login: Fri Nov 26 20:57:39 2021 from 102.64.130.202

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/
(base) [ec2-user@ip-172-31-26-61 ~]$ jupyter notebook
[W 2021-11-27 08:20:49.536 LabApp] 'certfile' has moved from NotebookApp to ServerApp. This config will be passed to ServerApp. Be sure to update your config before our next release.
[W 2021-11-27 08:20:49.536 LabApp] 'keyfile' has moved from NotebookApp to ServerApp. This config will be passed to ServerApp. Be sure to update your config before our next release.
[W 2021-11-27 08:20:49.536 LabApp] 'ip' has moved from NotebookApp to ServerApp. This config will be passed to ServerApp. Be sure to update your config before our next release.
[W 2021-11-27 08:20:49.537 LabApp] 'ip' has moved from NotebookApp to ServerApp. This config will be passed to ServerApp. Be sure to update your config before our next release.
[W 2021-11-27 08:20:49.537 LabApp] 'password' has moved from NotebookApp to ServerApp. This config will be passed to ServerApp. Be sure to update your config before our next release.
[W 2021-11-27 08:20:49.537 LabApp] 'port' has moved from NotebookApp to ServerApp. This config will be passed to ServerApp. Be sure to update your config before our next release.
[W 2021-11-27 08:20:49.537 LabApp] 'port' has moved from NotebookApp to ServerApp. This config will be passed to ServerApp. Be sure to update your config before our next release.
[I 2021-11-27 08:20:49.545 LabApp] JupyterLab extension loaded from /home/ec2-user/anaconda3/lib/python3.8/site-packages/jupyterlab
[I 2021-11-27 08:20:49.545 LabApp] JupyterLab application directory is /home/ec2-user/anaconda3/share/jupyter/lab
[I 08:20:49.550 NotebookApp] Serving notebooks from local directory: /home/ec2-user
[I 08:20:49.550 NotebookApp] Jupyter Notebook 6.3.0 is running at:
[I 08:20:49.550 NotebookApp] https://ip-172-31-26-61.eu-west-1.compute.internal:8888/
[I 08:20:49.550 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[W 08:21:37.196 NotebookApp] SSL Error on 10 ('102.64.130.202', 55973): [SSL: SSLV3_ALERT_CERTIFICATE_UNKNOWN] sslv3 alert certificate unknown (_ssl.c:1125)
[W 08:21:37.198 NotebookApp] SSL Error on 11 ('102.64.130.202', 55972): [SSL: SSLV3_ALERT_CERTIFICATE_UNKNOWN] sslv3 alert certificate unknown (_ssl.c:1125)
[W 08:21:37.446 NotebookApp] SSL Error on 12 ('102.64.130.202', 55974): [SSL: SSLV3_ALERT_CERTIFICATE_UNKNOWN] sslv3 alert certificate unknown (_ssl.c:1125)
[W 08:21:46.997 NotebookApp] SSL Error on 11 ('102.64.130.202', 55996): [SSL: SSLV3_ALERT_CERTIFICATE_UNKNOWN] sslv3 alert certificate unknown (_ssl.c:1125)
[W 08:21:47.005 NotebookApp] SSL Error on 10 ('102.64.130.202', 55997): [SSL: SSLV3_ALERT_CERTIFICATE_UNKNOWN] sslv3 alert certificate unknown (_ssl.c:1125)
[W 08:21:47.312 NotebookApp] SSL Error on 12 ('102.64.130.202', 55999): [SSL: SSLV3_ALERT_CERTIFICATE_UNKNOWN] sslv3 alert certificate unknown (_ssl.c:1125)
[I 08:21:47.575 NotebookApp] 302 GET / (102.64.130.202) 0.510000ms
[I 08:21:47.978 NotebookApp] 302 GET /tree? (102.64.130.202) 0.670000ms
[W 08:21:49.768 NotebookApp] SSL Error on 13 ('102.64.130.202', 56006): [SSL: SSLV3_ALERT_CERTIFICATE_UNKNOWN] sslv3 alert certificate unknown (_ssl.c:1125)
[W 08:21:49.768 NotebookApp] SSL Error on 12 ('102.64.130.202', 56008): [SSL: SSLV3_ALERT_CERTIFICATE_UNKNOWN] sslv3 alert certificate unknown (_ssl.c:1125)
[W 08:21:49.769 NotebookApp] SSL Error on 14 ('102.64.130.202', 56007): [SSL: SSLV3_ALERT_CERTIFICATE_UNKNOWN] sslv3 alert certificate unknown (_ssl.c:1125)
[W 08:21:49.773 NotebookApp] SSL Error on 15 ('102.64.130.202', 56009): [SSL: SSLV3_ALERT_CERTIFICATE_UNKNOWN] sslv3 alert certificate unknown (_ssl.c:1125)
[W 08:21:50.252 NotebookApp] SSL Error on 13 ('102.64.130.202', 56013): [SSL: SSLV3_ALERT_CERTIFICATE_UNKNOWN] sslv3 alert certificate unknown (_ssl.c:1125)
[W 08:21:50.324 NotebookApp] SSL Error on 14 ('102.64.130.202', 56014): [SSL: SSLV3_ALERT_CERTIFICATE_UNKNOWN] sslv3 alert certificate unknown (_ssl.c:1125)
[W 08:21:50.851 NotebookApp] SSL Error on 14 ('102.64.130.202', 56019): [SSL: SSLV3_ALERT_CERTIFICATE_UNKNOWN] sslv3 alert certificate unknown (_ssl.c:1125)
[I 08:21:54.124 NotebookApp] 302 POST /login?next=%2Ftree%3F (102.64.130.202) 1.100000ms
[W 08:22:00.945 NotebookApp] SSL Error on 14 ('102.64.130.202', 56049): [SSL: SSLV3_ALERT_CERTIFICATE_UNKNOWN] sslv3 alert certificate unknown (_ssl.c:1125)
[W 08:22:15.371 NotebookApp] Notebook P8_01_notebook.ipynb is not trusted
[W 08:22:15.626 NotebookApp] SSL Error on 16 ('102.64.130.202', 56086): [SSL: SSLV3_ALERT_CERTIFICATE_UNKNOWN] sslv3 alert certificate unknown (_ssl.c:1125)
[I 08:22:17.609 NotebookApp] Kernel started: 0fc44201-87d9-46f1-ab87-8b43329fee28, name: python3
[W 08:22:18.332 NotebookApp] SSL Error on 26 ('102.64.130.202', 56096): [SSL: SSLV3_ALERT_CERTIFICATE_UNKNOWN] sslv3 alert certificate unknown (_ssl.c:1125)
[W 08:22:57.789 NotebookApp] SSL Error on 36 ('54.151.6.217', 43240): [SSL: HTTP_REQUEST] http request (_ssl.c:1125)
[I 08:24:20.851 NotebookApp] Saving file at /P8_01_notebook.ipynb
[W 08:24:20.851 NotebookApp] Notebook P8_01_notebook.ipynb is not trusted
[I 08:26:19.280 NotebookApp] Saving file at /P8_01_notebook.ipynb
```

Jupyter Notebook sur EC2

Fruits!