



Fruits!

#8 Déployez un modèle dans le Cloud

Soutenance Emilie Groschêne le 7/07/2023

Evaluateur: Panayotis Papoutsis

Mentor: Léa Naccache

Sommaire

I

Problématique et jeu de données

II

Processus de création de l'environnement Big Data

III

Chaîne de traitement des images

IV

Démonstration d'exécution du script sur le Cloud

V

Synthèse et conclusion

I. PROBLEMATIQUE ET JEU DE DONNEES

I. Présentation de la problématique



Fruits!

Fruits est une start-up de l'agri-tech qui a pour volonté de **préserver la biodiversité des fruits** en développant des **robots cueilleurs intelligents** qui appliqueraient des **traitements** spécifiques à chaque espèce de fruits lors de la récolte.

❑ **Mission:**

Pour se faire **connaître auprès du grand public**, elle souhaite mettre à sa disposition une **application mobile** qui permettrait aux utilisateurs de **prendre en photo un fruit ou un légume** et d'obtenir des **informations** sur ce dernier.

❑ **Objectifs :**

- sensibiliser le grand public à la biodiversité des fruits
- mettre en place une **première version du moteur de classification** des images de fruits
- construire une **première version de l'architecture Big Data** nécessaire

❑ **Contraintes :**

- le **volume des données va augmenter très rapidement** après la livraison de ce projet
- les **serveurs** doivent être situés sur le **territoire européen**

I. Présentation du jeu de données

Le dataset est composé de:

- **90 483** images en **haute qualité** représentant chacune **un fruit ou un légume (131 variétés différentes)**
- **75%** des images seront utilisés pour **entraîner** le modèle et **25%** pour le **tester**
- toutes les images sont de taille identique: **100x100 pixels couleur sur fond blanc**

Chaque fruit ou légume a été placé sur un moteur à faible vitesse (3 tours par minute) puis une courte vidéo de 20 secondes a été prise:

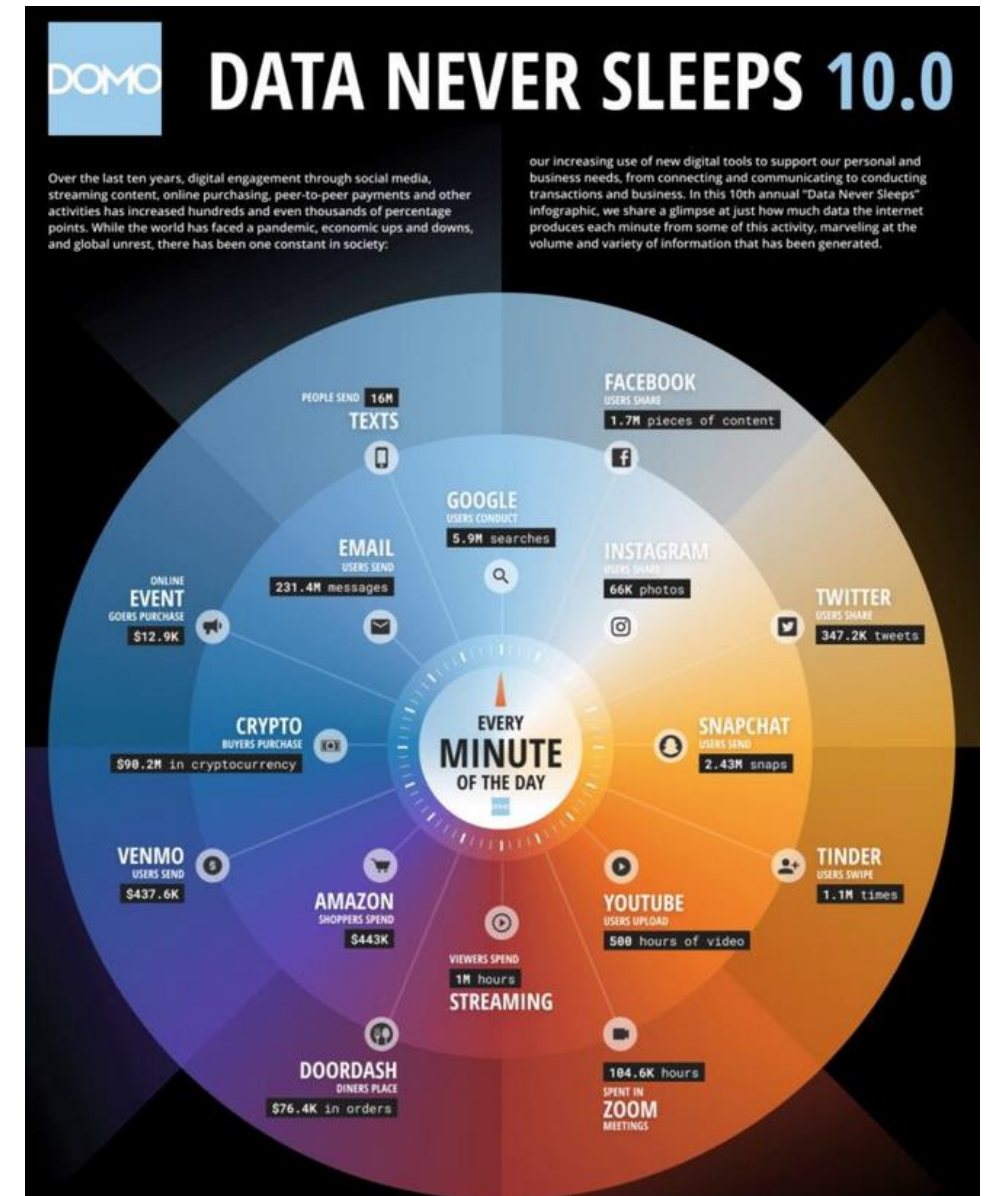
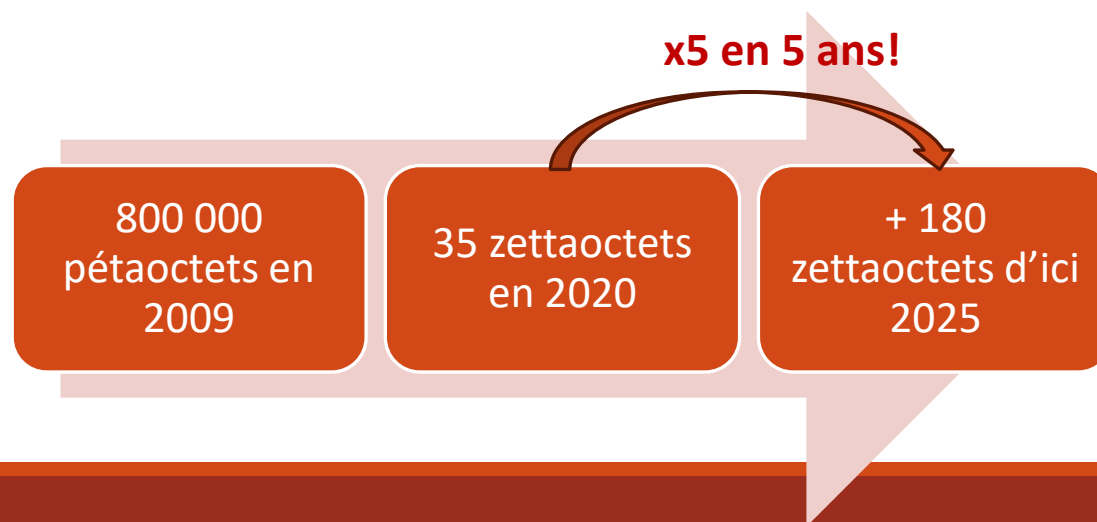


A gauche: image originale avec son arrière plan (feuille blanche) et l'axe moteur.
A droite: image retraitée de son arrière plan et du moteur puis réduite à 100x100 pixels

II. PROCESSUS DE CREATION DE L'ENVIRONNEMENT BIG DATA

II. Qu'est ce que le Big Data (1/2)

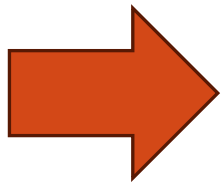
- Utilisation massive d'internet et des objets connectés
- production de quantités astronomiques de données
- problématiques de stockage et de traitement approprié pour les entreprises



II. Qu'est ce que le Big Data (2/2)

Il faut donc des **technologies innovantes** capables de traiter:

- des **volumes de données énormes** et en **constante augmentation**
- des données provenant de **sources multiples** et de **natures diverses**
- des **besoins analytiques** vitaux à fournir dans des délais impartis



Le choix de l'architecture de données à mettre en place est donc essentiel



II. Architecture des données retenue: Databricks sur AWS (1/2)

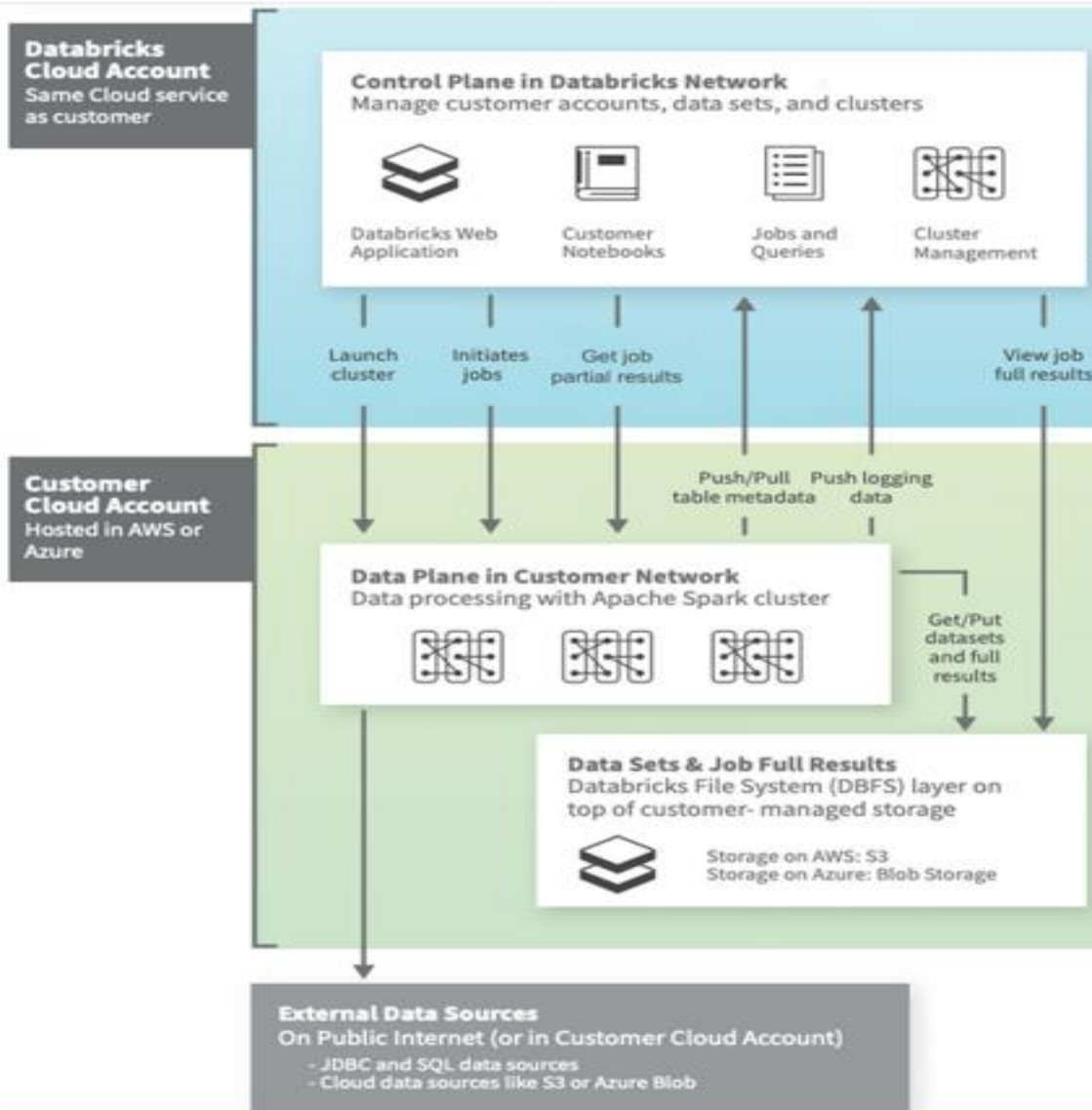
❑ Avantages:



Environnement PaaS permettant d'accéder aux données et aux ressources de calcul

- **simplicité**: une seule architecture de données unifiée sur S3 (solution de stockage des données sur Amazon) pour l'analytique SQL, la data science et le machine learning. De plus, aucune expérience avec les systèmes d'exploitation Linux ou Unix n'est demandée
- **rapport performance / prix**: performances du data warehouse au prix d'un data lake grâce à des clusters de calcul optimisés par SQL
- contrôle du **cycle de vie de la donnée entièrement géré sur le cloud**
- permet la **collaboration**
- **réputation**: des clients prestigieux

II. Architecture des données retenue: Databricks sur AWS (2/2)



❑ Control plane

Services backend que Databricks gère dans son propre compte AWS.

Gestion des comptes utilisateurs, des workspaces, des données et des clusters.

❑ Data plane:

Endroit où les données sont traitées et où se trouvent les **ressources de calcul** (cluster Apache Spark).

Fournit de la puissance de calcul et du stockage.

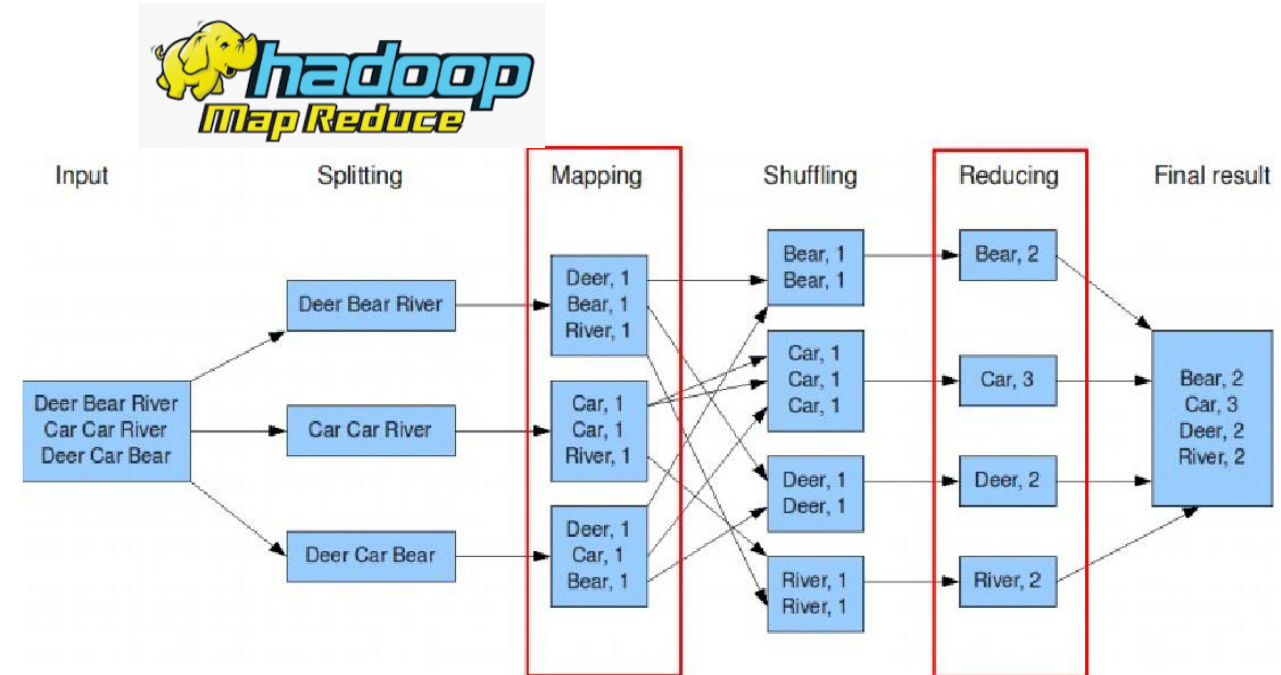
II. Objectif de traitement des données à grande échelle: de MapReduce à Apache Spark

MapReduce est un **modèle de programmation** qui **décompose le traitement d'une opération** (« job ») en plusieurs étapes:

- **MAP:** génération des paires clé / valeur
- **SHUFFLE AND SORT:** trie et consolide les données issues de la phase Map
- **REDUCE:** produit le résultat final (ici le comptage des occurrences des valeurs)

❑ 3 concepts clés:

- les **données** sont **distribuées** lorsqu'elles sont **stockées**
- Le **calcul** est **exécuté** là où les **données** sont **stockées** (data locality)
- **Ordre séquentiel** et non itératif (Map, Suffle, Reduce)



❑ Ajouts de SPARK:

- **Données traitées en mémoire** pour une utilisation plus rapide
- **Plan d'exécution** pour organiser les « jobs » (lazy evaluation)
- **API de haut niveau**

II. Processus de création de l'environnement S3 et Databricks (1/3)

1

Création de l'utilisateur avec Programmatic Access

- Création de l'utilisateur sur AWS IAM (Identity and Access Management)
- Ajout de la stratégie **AmazonS3FullAccess**
- Téléchargement des identifiants (credentials) au format .csv



IAM > Utilisateurs

Utilisateurs (1) Infos						
Un utilisateur IAM est une identité avec des informations d'identification à long terme utilisées pour interagir avec AWS dans un compte.						
<input type="text" value="Rechercher des utilisateurs par nom d'utilisateur ou clé d'accès"/>						
<input type="checkbox"/>	Nom d'utilisateur	Groupes	Dernière activité	MFA	Âge du mot de pas...	Identifiant de la clé d'a
<input type="checkbox"/>	databricks	Aucun	Il y a 5 heures	Aucun	Aucun	Active - AKIAWKMHQ

2

Création d'un workspace sur Databricks

- CloudFormation template qui va créer un rôle IAM et un bucket S3

Let's set up your workspace

We're going to send you to your AWS Console to configure your account.

Once you sign in, we'll pre-populate a CloudFormation template that creates an IAM role and S3 bucket for you, then deploys your workspace.

If you encounter any errors during the process, visit the [Databricks Community](#) for troubleshooting guidance.

CloudFormation > Piles > databricks-workspace-stack-b90ac

Workspace Name:

Human readable name for your workspace

AWS Region of the Databricks workspace:

AWS Region where the workspace will be created

Piles (3)

Actif	Afficher imbriqué
<input checked="" type="radio"/>	<input checked="" type="checkbox"/>

Piles
<div>databricks-workspace-stack-b90ac</div> <div>30-06-2023 11:04:37 UTC+0200</div> <div>CREATE_COMPLETE</div>
<div>databricks-workspace-stack-2420b</div> <div>15-06-2023 10:45:58 UTC+0200</div> <div>CREATE_COMPLETE</div>
<div>databricks-workspace-stack-</div> <div></div> <div></div>

databricks-workspace-stack-b90ac

Supprimer Mettre à jour Actions de pile Créer une pile

Informations sur la pile Événements Ressources Sorties Paramètres Modèle J

Présentation

ID de la pile	Description
arn:aws:cloudformation:eu-west-3:434612633663:stack/databricks-workspace-stack-b90ac/f94c1b00-1724-11ee-a866-0aa12bf32802	Set up resources and deploy a Databricks workspace in your AWS account. If you encounter any errors during the process, reach out to quickstart-support@databricks.com . Visit this Databricks Community post for troubleshooting guidance: https://dbricks.co/AWSQuickStartHelp .
Statut	Motif du statut
CREATE_COMPLETE	-
Pile racine	Pile parent
-	-
Heure de création	Heure de suppression
30-06-2023 11:04:37 UTC+0200	-

II. Processus de création de l'environnement S3 et Databricks (2/3)

3

Chargement des données sur S3



Amazon S3 > Compartiments > fruits-pictures

fruits-pictures Infos

Objets | Propriétés | Autorisations | Métriques

Objets (2)
Les objets sont les entités fondamentales stockées dans Amazon S3. Vous pouvez permettre à d'autres personnes d'accéder à vos objets, vous devez leur accorder explicitement des permissions.

<input type="checkbox"/>	Nom	Type
<input type="checkbox"/>	fruits360_photos_features/	Dossier
<input type="checkbox"/>	test_S3/	Dossier

4

Stockage de la clé d'accès et clé secrète sur Databricks

databricks Search data, notebooks, recents, and more... CTRL + P

DBFS

Upload File DBFS

DBFS Target Directory ⓘ
/FileStore/tables/ (optional)

Files uploaded to DBFS are accessible by everyone who has access to this workspace. [Learn more](#)

Files ⓘ

databricks_ac...

99 b
[Remove file](#)

✓ File uploaded to /FileStore/tables/databricks_accessKeys-1.csv

5

Création d'un cluster Spark

Summary

1 Driver 30.5 GB Memory, 4 Cores
Runtime 9.1.x-cpu-ml-scala2.12

i3.xlarge 1 DBU/h

II. Processus de création de l'environnement S3 et Databricks (3/3)

6

Création et lancement du script Pyspark pour relier Databricks au bucket S3

```
1 file_type = "csv"
2 first_row_header = "true"
3 delimiter = ","
4
5 # Read the CSV file to spark dataframe
6 aws_keys_df = spark.read.format(file_type)\
7 .option("header", first_row_header)\
8 .option("sep", delimiter)\
9 .load("/FileStore/tables/databricks_accessKeys.csv")
```

► (1) Spark Jobs

► aws_keys_df: pyspark.sql.dataframe.DataFrame = [Access_key_ID: string, Secret_access_key: string]

Command took 4.53 seconds -- by emgroschene@gmail.com at 30/06/2023 11:29:54 on [default]basic-starter-cluster

Cmd 3

```
1 # To send keys to AWS
2
3 from pyspark.sql.functions import *
4 import urllib
5
6 # Collect access and secret key from spark dataframe (dbfs:/FileStore/tables/)
7 ACCESS_KEY = aws_keys_df.select('Access_key_ID').collect()[0]['Access_key_ID']
8 SECRET_KEY = aws_keys_df.select('Secret_access_key').collect()[0]['Secret_access_key']
9
10 # Encode secret key (safe="" means every character in the secret key is encoded)
11 ENCODED_SECRET_KEY = urllib.parse.quote(SECRET_KEY,"")
```

```
1 # Mount the S3 bucket
2
3 # AWS S3 bucket name
4 AWS_S3_BUCKET = "fruits-pictures"
5 # Mount name for the bucket
6 MOUNT_NAME = "/mnt/fruits-pictures"
7 # Source url
8 SOURCE_URL = "s3n://{0}:{1}@{2}".format(ACCESS_KEY, ENCODED_SECRET_KEY, AWS_S3_BUCKET)
9 # Mount the drive
10 dbutils.fs.mount(SOURCE_URL, MOUNT_NAME)
```

ut[5]: True

Command took 11.97 seconds -- by emgroschene@gmail.com at 30/06/2023 11:30:55 on [default]basic-starter-cluster

5

```
1 # Read data from the mounted S3 bucket
2
3 # Check if the AWS S3 bucket was mounted successfully
4 # %fs ls "/mnt/fruits-pictures/"
5 display(dbutils.fs.ls("/mnt/fruits-pictures/test_local/Watermelon/"))
```

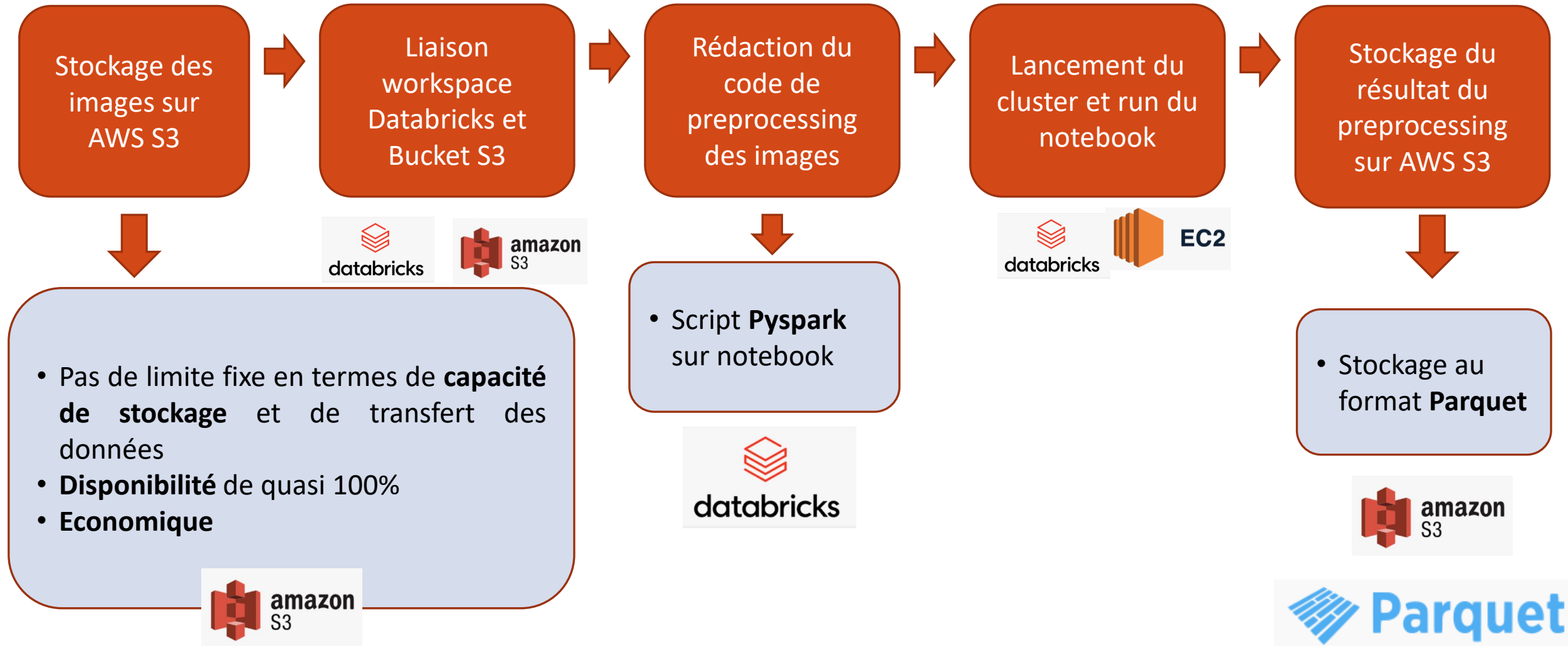
► (2) Spark Jobs

Table +

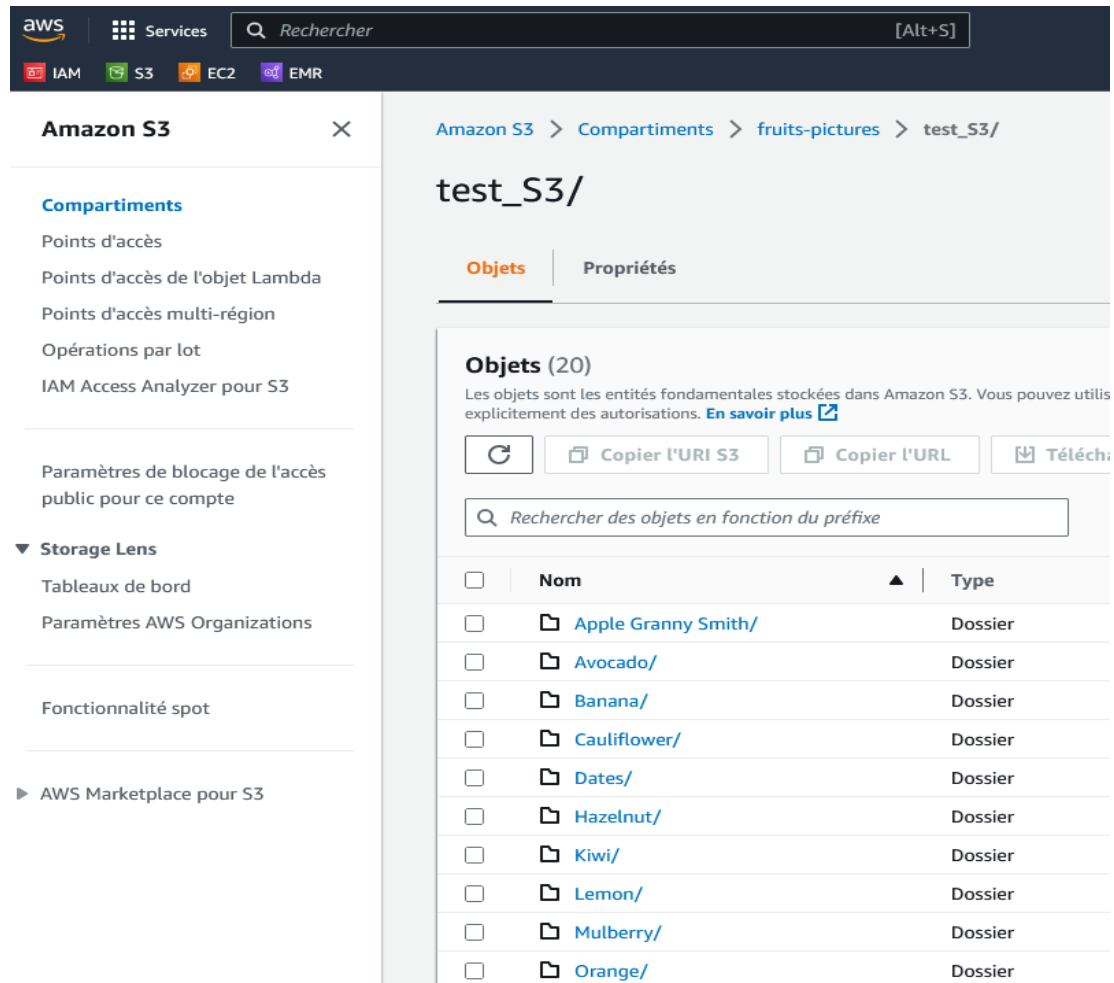
	path	name	size	modificationTime
1	dbfs:/mnt/fruits-pictures/test_local/Watermelon/125_100.jpg	125_100.jpg	6855	1688117519000
2	dbfs:/mnt/fruits-pictures/test_local/Watermelon/r_113_100.jpg	r_113_100.jpg	7018	1688117520000
3	dbfs:/mnt/fruits-pictures/test_local/Watermelon/r_183_100.jpg	r_183_100.jpg	7018	1688117521000
4	dbfs:/mnt/fruits-pictures/test_local/Watermelon/r_47_100.jpg	r_47_100.jpg	7071	1688117522000
5	dbfs:/mnt/fruits-pictures/test_local/Watermelon/r_63_100.jpg	r_63_100.jpg	7251	1688117522000

III. CHAINE DE TRAITEMENT DES IMAGES

III. Chaîne de traitement des images



III. Chaîne de traitement des images: stockage des images sur AWS S3

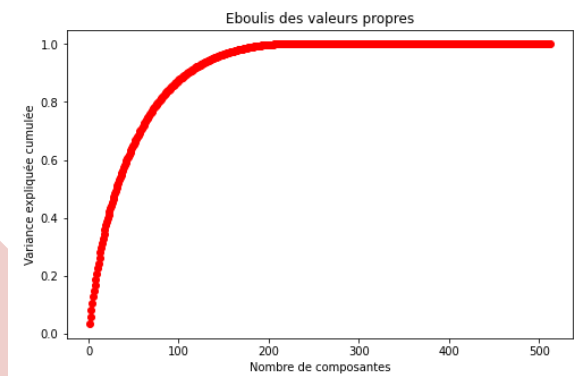
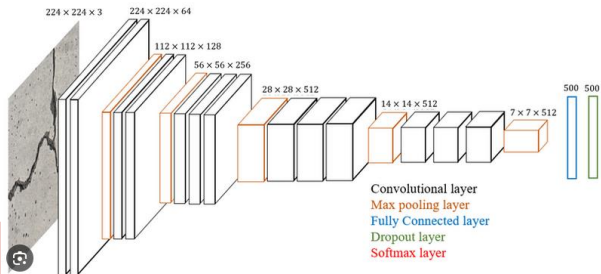


Stockage du dossier contenant les images à preprocesser dans le **bucket S3** rattaché au workspace de Databricks.

La création d'un bucket permet de **choisir la région** où les images seront stockées et de **réduire les délais de traitement et les coûts d'utilisation**.



III. Chaîne de traitement des images: Preprocessing des données



Chargement
des images
au format
binary file

Chargement
du modèle
de transfer
learning

Broadcasting
des poids

Extraction
des features

Transformation
des features en
vecteur dense

Standardisation
+ ACP

Enregistrement
sur le bucket
S3 au format
parquet

path	modificationTime	length	content
dbfs:/mnt/fruits-...	2023-07-03 08:56:10	7231	[FF D8 FF E0 00 1...
dbfs:/mnt/fruits-...	2023-07-03 08:56:09	7226	[FF D8 FF E0 00 1...
dbfs:/mnt/fruits-...	2023-07-03 08:56:08	7224	[FF D8 FF E0 00 1...
dbfs:/mnt/fruits-...	2023-07-03 08:56:08	7100	[FF D8 FF E0 00 1...
dbfs:/mnt/fruits-...	2023-07-03 08:56:07	7012	[FF D8 FF E0 00 1...

features
[53.791595, 20.18...
[35.181335, 0.0, ...
[19.833202, 0.0, ...
[24.497759, 11.62...
[30.44961, 6.8856...

std_features	pcaFeatures
[-0.5292908654997...	[8.99227389760900...
[-0.2090919187030...	[7.44657047451287...
[-0.2688848630991...	[5.99492786066021...
[0.83298584323146...	[-9.1931090447816...
[-0.6274642586441...	[4.39778728061732...

III. Chaîne de traitement des images: allocation des ressources pour le traitement des images

Compute > **UI preview** Provide feedback

P8_ML_Cluster  

Configuration	Notebooks (0)	Libraries	Event log	Spark UI	Driver logs	Metrics	Apps
---------------	---------------	-----------	-----------	----------	-------------	---------	------

☒ Multi node ☐ Single node

Access mode Single user access

Emilie GROSCHE

Performance

Databricks Runtime Version

12.2 LTS ML (includes Apache Spark 3.3.2, Scala 2.12)

☐ Use Photon Acceleration ?

Worker type

61 GB Memory, 8 Cores

Min workers	Max workers	Current
1	1	1
1	2	2
1	3	3
1	4	4
1	5	5
1	6	6
1	7	7
1	8	8
1	9	9
1	10	10
1	11	11
1	12	12
1	13	13
1	14	14
1	15	15
1	16	16
1	17	17
1	18	18
1	19	19
1	20	20
1	21	21
1	22	22
1	23	23
1	24	24
1	25	25
1	26	26
1	27	27
1	28	28
1	29	29
1	30	30
1	31	31
1	32	32
1	33	33
1	34	34
1	35	35
1	36	36
1	37	37
1	38	38
1	39	39
1	40	40
1	41	41
1	42	42
1	43	43
1	44	44
1	45	45
1	46	46
1	47	47
1	48	48
1	49	49
1	50	50
1	51	51
1	52	52
1	53	53
1	54	54
1	55	55
1	56	56
1	57	57
1	58	58
1	59	59
1	60	60
1	61	61
1	62	62
1	63	63
1	64	64
1	65	65
1	66	66
1	67	67
1	68	68
1	69	69
1	70	70
1	71	71
1	72	72
1	73	73
1	74	74
1	75	75
1	76	76
1	77	77
1	78	78
1	79	79
1	80	80
1	81	81
1	82	82
1	83	83
1	84	84
1	85	85
1	86	86
1	87	87
1	88	88
1	89	89
1	90	90
1	91	91
1	92	92
1	93	93
1	94	94
1	95	95
1	96	96
1	97	97
1	98	98
1	99	99
1	100	100

8

8

2

Driver type

61 GB Memory, 8 Cores

☒ Enable autoscaling ☐ Enable autoscaling local storage ?

☒ Terminate after minutes of inactivity ?

Types d'instances Amazon EC2

Amazon EC2 fournit un vaste éventail de types d'instances optimisés pour différents cas d'utilisation. Ces types d'instances correspondent à différentes capacités de CPU, de mémoire, de stockage et de mise en réseau. Vous pouvez ainsi choisir un ensemble de ressources parfaitement adapté à vos besoins. Par exemple, vous pouvez choisir une instance conçue pour exécuter des applications web, ou une instance conçue pour exécuter des applications de calcul intensif. Vous pouvez également choisir une instance conçue pour exécuter des applications de calcul intensif, ou une instance conçue pour exécuter des applications de calcul intensif. Vous pouvez également choisir une instance conçue pour exécuter des applications de calcul intensif, ou une instance conçue pour exécuter des applications de calcul intensif.

 Usage général

Calcul optimisé

Mémoire optimisée

Calcul accéléré

Stockage optimisé

Usage général

Les instances à usage général assurent l'équilibre entre les ressources informatiques, la mémoire aussi être utilisées pour un certain nombre de charges de travail diverses. Ces instances sont idé: ressources en proportions égales, à l'instar des serveurs Web et des référentiels de code.

Mac	T4g	T3	T3a	T2	M6g	M6i	M6a	M5	M5a
-----	-----	----	-----	----	-----	-----	-----	----	-----

Les instances T3 représentent un type d'instance à usage général, extensible, de nouvelle génération de CPU de base avec la possibilité d'étendre l'utilisation de CPU à tout moment et aussi longtemps qu'il y a de la demande. Elles offrent un équilibre entre les ressources de calcul, de mémoire et de réseau et sont conçues pour

Les principaux types d'instances

III. Chaîne de traitement des images: enregistrement des résultats

Enregistrement de la matrice
au format **Parquet** sur AWS S3

Amazon S3 > Compartiments > fruits-pictures > fruits360_photos_features/

fruits360_photos_features/

Objets

Propriétés

Objets (19)

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 bucket. [En savoir plus](#)

🔄

Copier l'URI S3

Copier l'URL

Télécharger







Ouvrir

Supprimer

Actions ▼

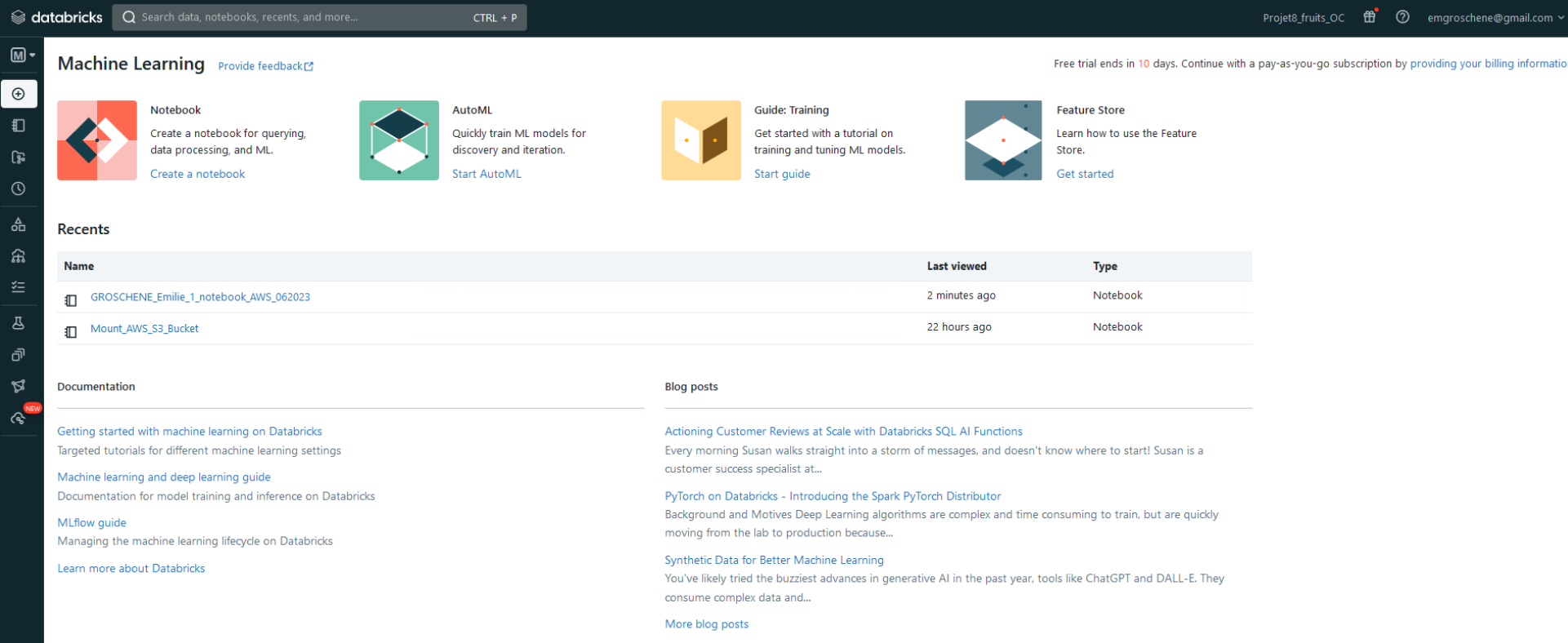
🔍

Rechercher des objets en fonction du préfixe

<input type="checkbox"/>	Nom ▲	Type ▼	Dernière modification
<input type="checkbox"/>	 _committed_29641652051109451	-	04 Jul 2023 10:42:30 AM CEST
<input type="checkbox"/>	 _started_29641652051109451	-	04 Jul 2023 10:42:13 AM CEST
<input type="checkbox"/>	 _SUCCESS	-	04 Jul 2023 10:42:30 AM CEST
<input type="checkbox"/>	 part-00000-tid-29641652051109451-1a0269a8-1828-49a1-bee3-b8781d8c7bb2-268-1-c000.snappy.parquet	parquet	04 Jul 2023 10:42:30 AM CEST
<input type="checkbox"/>	 part-00001-tid-29641652051109451-1a0269a8-1828-49a1-bee3-b8781d8c7bb2-269-1-c000.snappy.parquet	parquet	04 Jul 2023 10:42:27 AM CEST
<input type="checkbox"/>	 part-00002-tid-29641652051109451-1a0269a8-1828-49a1-bee3-b8781d8c7bb2-270-1-c000.snappy.parquet	parquet	04 Jul 2023 10:42:30 AM CEST

IV. DEMONSTRATION D'EXECUTION DU SCRIPT SUR LE CLOUD


IV. Démonstration d'exécution du script sur le Cloud





The screenshot displays the Databricks Machine Learning interface. At the top, a dark navigation bar includes the Databricks logo, a search bar, and user information. Below this, the 'Machine Learning' section features five main cards: 'Notebook' (with a 'Create a notebook' link), 'AutoML' (with a 'Start AutoML' link), 'Guide: Training' (with a 'Start guide' link), and 'Feature Store' (with a 'Get started' link). A 'Recents' table lists two notebooks: 'GROSCHENE_Emilie_1_notebook_AWS_062023' viewed 2 minutes ago and 'Mount_AWS_S3_Bucket' viewed 22 hours ago. The bottom section is divided into 'Documentation' and 'Blog posts', each with several links to guides and articles.


Machine Learning [Provide feedback](#)

Free trial ends in 10 days. Continue with a pay-as-you-go subscription by [providing your billing information](#).



**Notebook**
Create a notebook for querying, data processing, and ML.
[Create a notebook](#)

**AutoML**
Quickly train ML models for discovery and iteration.
[Start AutoML](#)

**Guide: Training**
Get started with a tutorial on training and tuning ML models.
[Start guide](#)

**Feature Store**
Learn how to use the Feature Store.
[Get started](#)

Recents

Name	Last viewed	Type
 GROSCHENE_Emilie_1_notebook_AWS_062023	2 minutes ago	Notebook
 Mount_AWS_S3_Bucket	22 hours ago	Notebook

Documentation

[Getting started with machine learning on Databricks](#)
Targeted tutorials for different machine learning settings

[Machine learning and deep learning guide](#)
Documentation for model training and inference on Databricks

[MLflow guide](#)
Managing the machine learning lifecycle on Databricks

[Learn more about Databricks](#)

Blog posts

[Actioning Customer Reviews at Scale with Databricks SQL AI Functions](#)
Every morning Susan walks straight into a storm of messages, and doesn't know where to start! Susan is a customer success specialist at...

[PyTorch on Databricks - Introducing the Spark PyTorch Distributor](#)
Background and Motives Deep Learning algorithms are complex and time consuming to train, but are quickly moving from the lab to production because...

[Synthetic Data for Better Machine Learning](#)
You've likely tried the buzziest advances in generative AI in the past year, tools like ChatGPT and DALL-E. They consume complex data and...

[More blog posts](#)

V. SYNTHESE ET CONCLUSION

V. Synthèse et conclusion



- Mise en place d'un **environnement Big Data** conforme aux normes **RGPD** en vigueur avec manipulation de la donnée
- Découverte de **Databricks** et **AWS**
- Elaboration de script en **Pyspark**, notion de **paralléliser** les calculs

Retour critique sur la solution retenue:



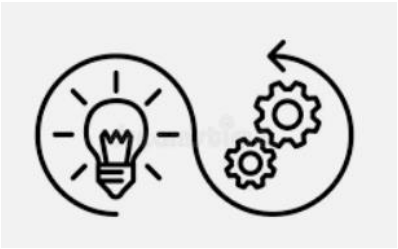
- Facile d'utilisation, pas de machine virtuelle à installer, paramétrage simple (QuickStart), accès aux librairies de ML
- Puissance de calcul facilement modifiable / configurable
- Outil collaboratif
- Documentation complète avec de nombreux notebooks d'exemples
- Possibilité de relier le compte Github pour le versionning (repos)



- Stockage peu coûteux
- Bucket S3 très simple à relier à Databricks

V. Synthèse et conclusion

Axes d'amélioration:



- Paramétrage du **cluster** à approfondir (worker type, driver type, autoscaling etc...)
- Optimisation des **jobs** Spark (format parquet, mise en cache des résultats intermédiaires)
- Attention à la **facturation**: bien vérifier que le cluster est résilié! (mise en place d'alerte + maîtrise des outils de facturation sur AWS)
- Tester d'autres modèles de transfer learning plus rapides (MobileNetV2 par exemple)

MERCI
