



PROJET 8 – « DÉPLOYEZ UN MODÈLE DANS LE CLOUD »

Tewodros Cherenet DEBELA

Formation: Data Scientist , Openclassrooms

Avril 2022

Plan de présentation

Partie 1

- Rappel de la problématique
- Présentation du jeu de données

Partie 2

- Présentation du Big Data
- Choix d'architecture de cloud

Partie 3

- Solution d'architecture Big data
- Présentation de les étape de la chaîne de traitement

Partie 4

- Conclusion et point d'amélioration

Partie 1

Rappel de la
Problématique

Présentation du
jeu de données

CONTEXTE & MISSION

- L'entreprise :



Fruits!

- ⇒ start-up de l'AgriTech
 - ⇒ Application Mobile de reconnaissance d'images
 - ⇒ Sensibiliser le grand public à la biodiversité des fruits
- Déploiement d'un environnement « Big Data »
 - ⇒ prétraitements & réduction dimensionnelle
 - ⇒ accessibilité des données et résultats dans le *cloud*

DONNÉES INITIALES



82'223 images

- Jeu d'entraînement : 61'488
- Jeu de test : 20'662

120 catégories

Encodées dans le chemin d'accès :
\Training\Apple_Golden_3\101_100.jpg

Taille 100 x 100 pixels

- 10'000 dimensions
- Résolution : 96 x 96 dpi
 - Format .jpg
- Couleurs : 24 bits de profondeur



Partie 2

Présentation de Big data

Présentation du cloud

Présentation de la
réalisation de la chaîne
de traitement des images

Big Data: Le trois « V »

- **Le Volume** des données générées nécessite de repenser la manière dont elles sont stockées.
 - Dépassement de la capacité de RAM
 - Dépassement des capacités de stockage
- **La Vitesse** à laquelle nous parvenons ces données sans paralyser le reste de l'application.
- Les données se présentent sous une grande **Variété** de formats. Ex...Structurées (documents **JSON**), semi-structurées (**fichiers de log**) ou non structurées (**textes, images**).

Fournisseur de service cloud



Sélection du cloud

- AWS (Amazon Web Services)



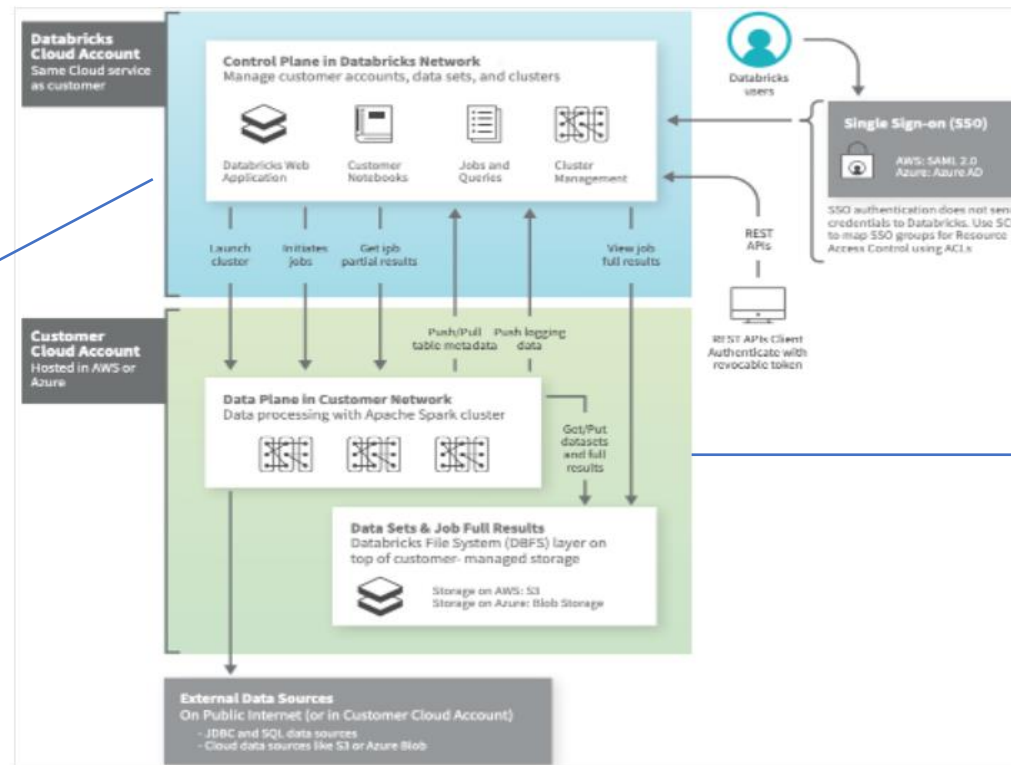
- Databricks  databricks

- Fondée par les créateurs d'Apache Spark
- Databricks développe une plate-forme Web pour travailler avec Spark qui fournit automatiquement **cluster management et IPython-style notebooks.**
- Databricks- SaaS (software as a service)-logiciel qui permettra d'utiliser le cloud à distance.

Plateforme Databricks

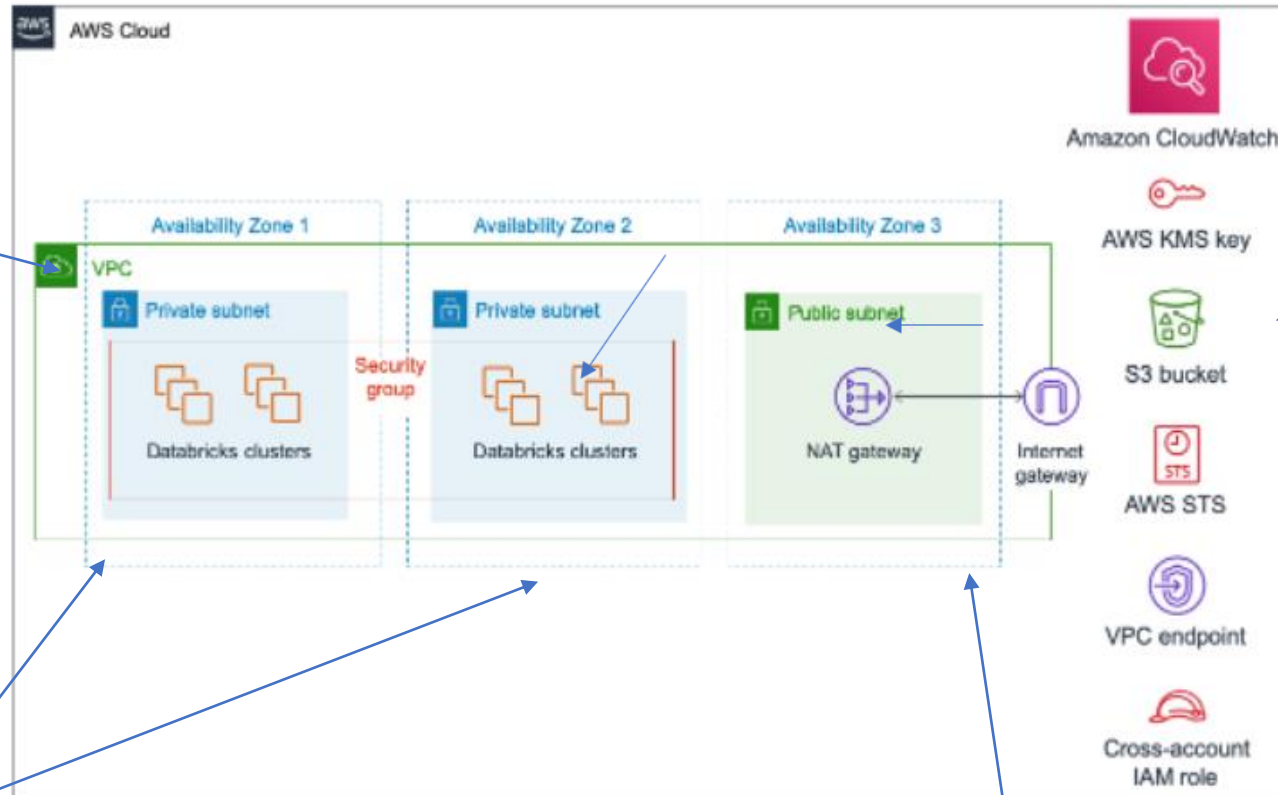
Plane de contrôle :

- Services backend géré par Databricks dans son propre compte AWS.
- Les requêtes SQL de Databricks, les commandes du **notebook** et de **nombreuses autres configurations de l'espace de travail** sont **stockées** dans le plan de contrôle et cryptées au repos.



Plane de données
C'est là que les données sont traitées à l'aide du cluster Apache Spark

Platform de Databricks



Virtual Private Cloud :

Le VPC est configuré avec des sous-réseaux privés et un sous-réseau public, conformément aux bonnes pratiques AWS, afin de disposer de son propre réseau virtuel sur AWS.

une connexion Sous-réseaux privés :

- Clusters Databricks d'instances Amazon Elastic Compute Cloud (Amazon EC2).
- Un ou plusieurs groupes de sécurité pour permettre écurisée au cluster

Sous-réseau public : Une passerelle NAT (Network address translation) pour autoriser un accès Internet sortant.

Journaux d'instance de l'instance Workspace Databricks

clé pour chiffrer le Notebook










Stockage d'objets (journaux de cluster ...)

AWS Security Service
Permettra de demander information pour l'authentification

Point d'accès aux journaux et artefact S3

Un rôle AWS Identity and Access Management permettant de déployer des clusters dans le VPC pour le nouvel espace de travail

Databricks: création de cluster



Create Cluster

New Cluster | Cancel Create Cluster

0 Workers: 0 GB Memory, 0 Cores, 0 DBU
1 Driver: 15.3 GB Memory, 2 Cores, 1 DBU ?

Cluster name

Pr8_openclassrooms

Databricks runtime version ?

Runtime: 10.4 LTS (Scala 2.12, Spark 3.2.1) | v

Instance

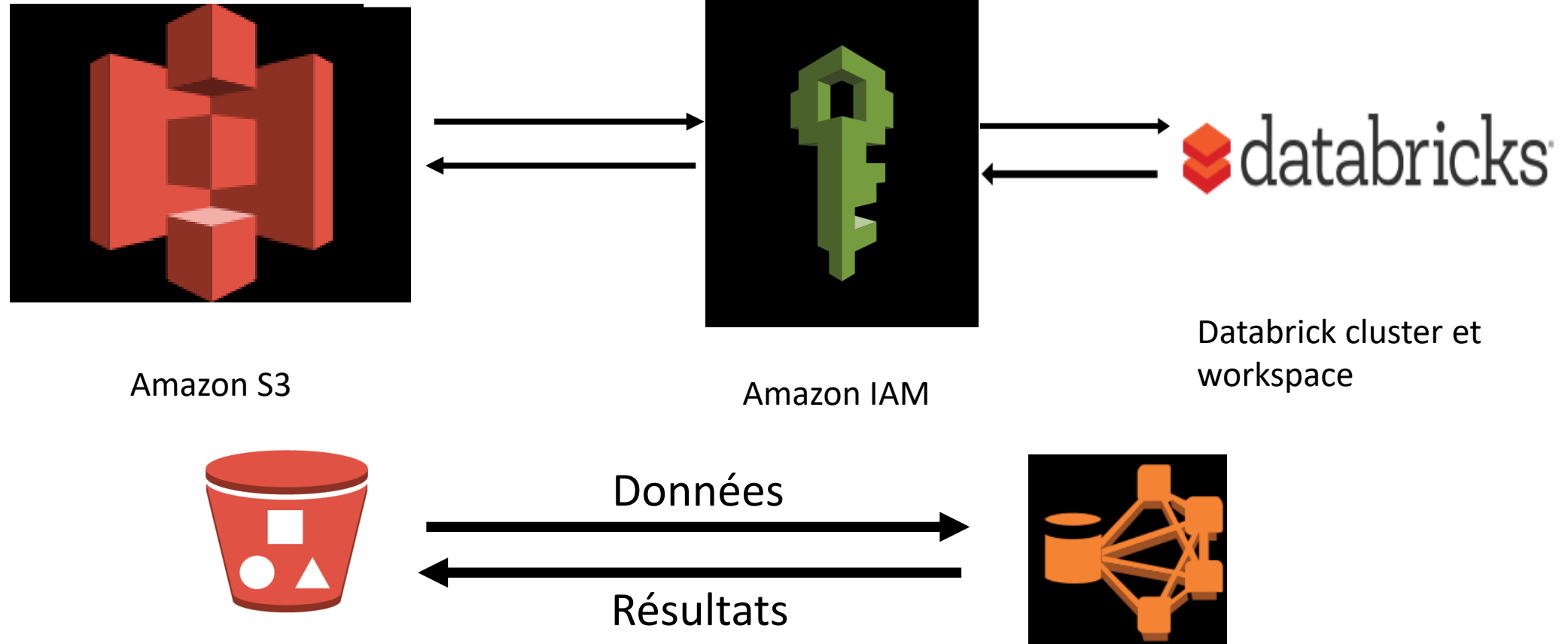
Free 15 GB Memory: As a Community Edition user, your cluster will automatically terminate after an idle period of two hours. For [more configuration options](#), please [upgrade your Databricks subscription](#).

Instances **Spark**

Availability zone ?

auto | v

DÉPLOIEMENT DANS LE CLOUD



Aws

- * EC2 – Création d'Instance-Connection avec SSH
- * S3 – Stockage de données (trois catégories de fruits avec 10 images de chacune)

Amazon S3 > Compartiments > pr8aws > fruits/

fruits/ Copier l'URI S3

Objets Propriétés

Objets (3)

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 compartiment. Pour que d'autres personnes puissent accéder à vos objets, vous devez leur accorder explicitement des autorisations. [En savoir plus](#)

Copier l'URI S3 Copier l'URL Télécharger Ouvrir Supprimer Actions Créer un dossier

Charger

Rechercher des objets en fonction du préfixe

	Nom	Type	Dernière modification	Taille	Classe de stockage
<input type="checkbox"/>	Avocado/	Dossier	-	-	-
<input type="checkbox"/>	Banana/	Dossier	-	-	-
<input type="checkbox"/>	Walnut/	Dossier	-	-	-

Amazon S3 > Compartiments > pr8aws > fruits/ > Avocado/

Avocado/ Copier l'URI S3

Objets Propriétés

Objets (10)

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 compartiment. Pour que d'autres personnes puissent accéder à vos objets, vous devez leur accorder explicitement des autorisations. [En savoir plus](#)

Copier l'URI S3 Copier l'URL Télécharger Ouvrir Supprimer Actions Créer un dossier Charger

Rechercher des objets en fonction du préfixe

	Nom	Type	Dernière modification	Taille	Classe de stockage
<input type="checkbox"/>	151_100.jpg	Objet	05 Apr 2022 11:21:17 AM CEST	4.1 KiB	Standard
<input type="checkbox"/>	207_100.jpg	Objet	05 Apr 2022 11:21:17 AM CEST	4.3 KiB	Standard
<input type="checkbox"/>	234_100.jpg	Objet	05 Apr 2022 11:21:16 AM CEST	4.3 KiB	Standard
<input type="checkbox"/>	27_100.jpg	Objet	05 Apr 2022 11:21:16 AM CEST	4.1 KiB	Standard
<input type="checkbox"/>	188_100.jpg	Objet	05 Apr 2022 11:21:16 AM CEST	4.2 KiB	Standard
<input type="checkbox"/>	31_100.jpg	Objet	05 Apr 2022 11:21:16 AM CEST	4.6 KiB	Standard
<input type="checkbox"/>	337_100.jpg	Objet	05 Apr 2022 11:21:16 AM CEST	4.4 KiB	Standard
<input type="checkbox"/>	277_100.jpg	Objet	05 Apr 2022 11:21:15 AM CEST	4.3 KiB	Standard
<input type="checkbox"/>	292_100.jpg	Objet	05 Apr 2022 11:21:17 AM CEST	4.4 KiB	Standard
<input type="checkbox"/>	316_100.jpg	Objet	05 Apr 2022 11:21:15 AM CEST	4.7 KiB	Standard

Partie 3

**Solution d'architecture
Big data**

**Présentation de les étape
chaîne de traitement des
images**

Apache Spark

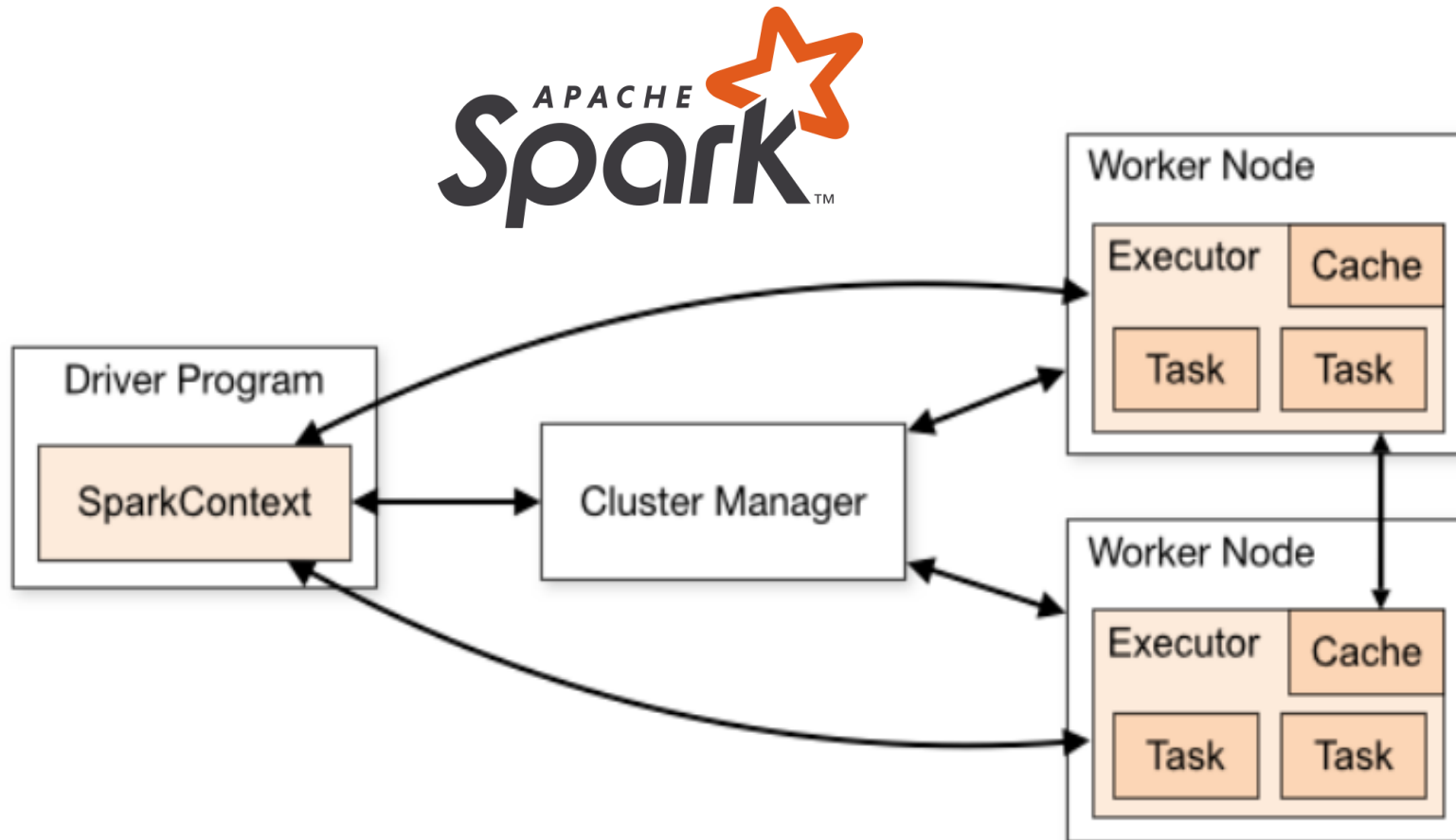


- **Apache Spark** : Plateforme (*framework*) multi-langage et ensemble de bibliothèques pour le traitement parallélisé de données sur des grappes (*clusters*) d'ordinateurs.
- **PySpark**: une interface pour Apache **Spark** en Python.
 - **Open-source**
 - **Multi-langage** : Scala, Java, Python, R...
 - **Stocke les résultats intermédiaires en RAM**

ARCHITECTURE SPARK

Capacités de calcul : Traitement par calculs distribués (MapReduce)

- * Diviser les opérations en micro opérations distribuables entre différentes machines, réalisables en parallèle
- * Agrégat les résultats sur une même machine

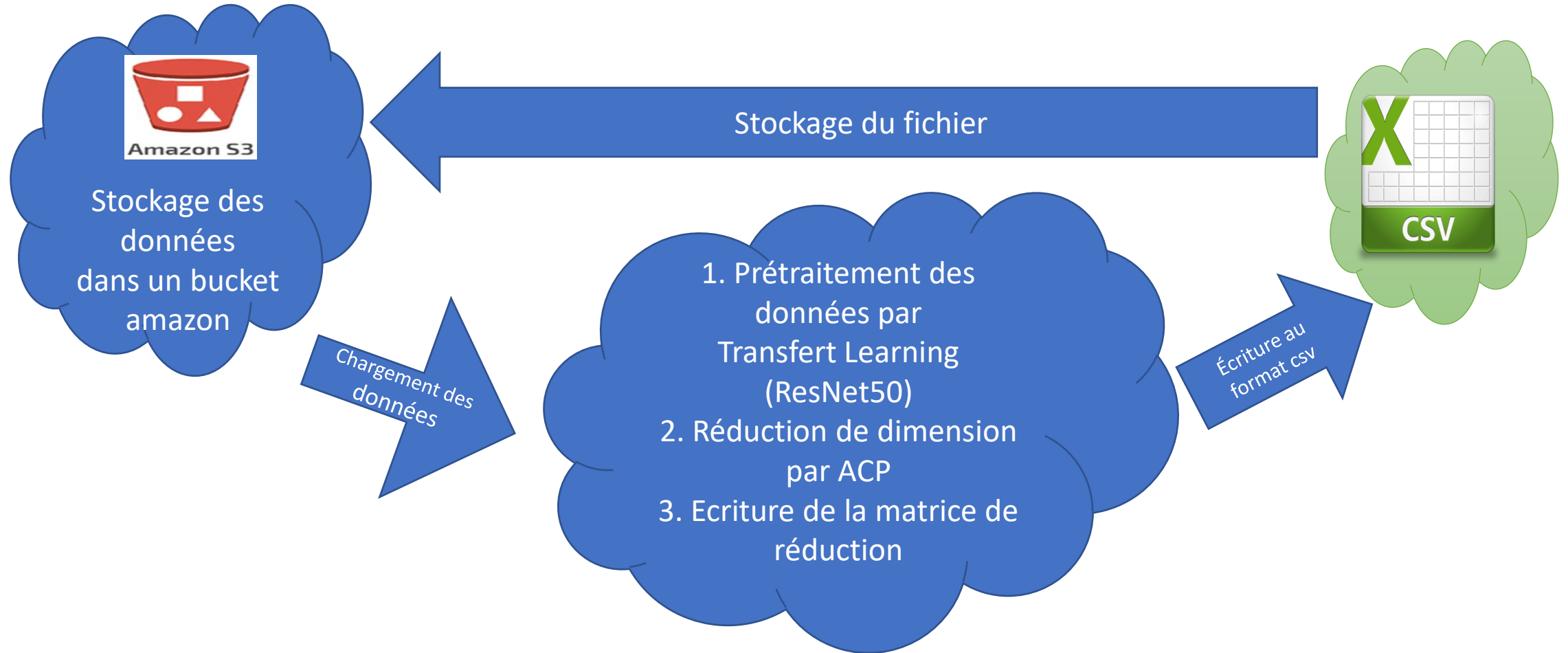


APPLICATION SPARK



- 1 application Spark = 1 ensemble de **jobs** Spark
- 1 job Spark = ensemble d'étapes (**stages**). Il se termine par 1 **action** (résultat)
- 1 étape (*stage*) Spark = 1 ensemble de **tâches** se terminant par 1 redistribution (***shuffle***)

Partie 3.2: Les étapes de la chaîne de traitement



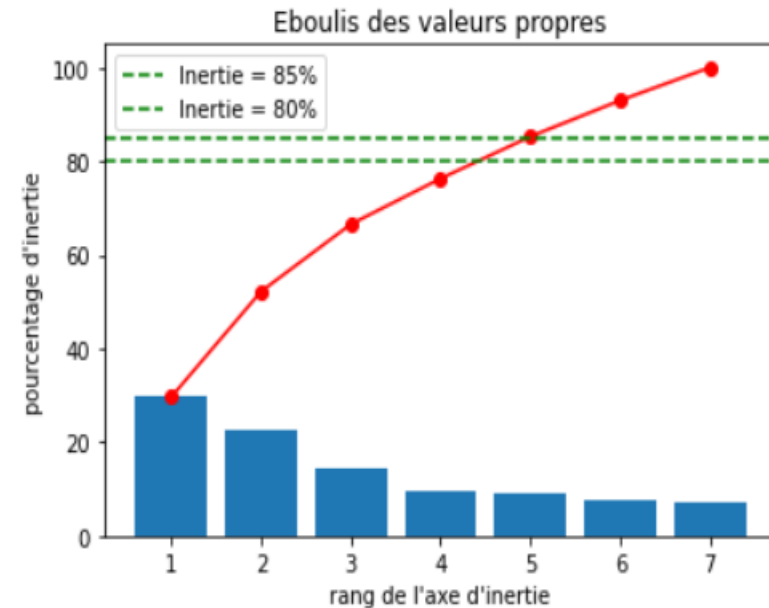
Features Extraction avec ResNet50 et composante principale après PCA

```
: features_df.show()
```

path	categorie	features
dbfs:/mnt/section...	Walnet	[0.95179075002670...
dbfs:/mnt/section...	Walnet	[0.90917086601257...
dbfs:/mnt/section...	Walnet	[0.94820779561996...
dbfs:/mnt/section...	Walnet	[0.91118246316909...
dbfs:/mnt/section...	Walnet	[0.0,0.6958804130...
dbfs:/mnt/section...	Walnet	[0.0,0.5577178001...
dbfs:/mnt/section...	Walnet	[0.0,0.5439043045...
dbfs:/mnt/section...	Walnet	[0.0,0.5718667507...
dbfs:/mnt/section...	Walnet	[0.0,0.4307540953...
dbfs:/mnt/section...	Walnet	[0.0,0.4147444069...
dbfs:/mnt/section...	Avocado	[0.0,0.5756284594...
dbfs:/mnt/section...	Avocado	[0.0,0.5745477080...
dbfs:/mnt/section...	Avocado	[0.10691066086292...
dbfs:/mnt/section...	Avocado	[0.18213319778442...
dbfs:/mnt/section...	Avocado	[0.05700669437646...
dbfs:/mnt/section...	Avocado	[0.0,3.1730124950...
dbfs:/mnt/section...	Avocado	[4.04228830337524...
dbfs:/mnt/section...	Avocado	[2.61954736709594...
dbfs:/mnt/section...	Avocado	[4.01343488693237...
dbfs:/mnt/section...	Avocado	[3.76430559158325...

only showing top 20 rows

```
pca_plot = display_scee_plot(modelpca)
pca_plot
```



Ecriture de la matrice de réduction

Resultats

← → ↻ s3.console.aws.amazon.com/s3/buckets/pr8aws?region=eu-west-3&prefix=df_result.csv/&showversions=false

Amazon S3

Compartiments

Points d'accès

Points d'accès de l'objet Lambda

Points d'accès multi-région

Opérations par lot

Analysateur d'accès pour S3

Paramètres de blocage de l'accès public pour ce compte

▼ Storage Lens

Tableaux de bord

Paramètres AWS Organizations

Fonctionnalité spot

► AWS Marketplace pour S3

Amazon S3 > Compartiments > pr8aws > df_result.csv/

df_result.csv/

Objets Propriétés

Objets (11)

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 compartiment. Pour que d'autres personnes puissent accéder à vos objets, vous devez leur accorder explicitement des autorisations. [En savoir plus](#)

🔍 Rechercher des objets en fonction du préfixe

<input type="checkbox"/>	Nom	Type	Dernière modification	Taille	Classe de stockage
<input type="checkbox"/>	_committed_7812980284244228796	-	03 Apr 2022 06:49:33 PM CEST	744.0 o	Standard
<input type="checkbox"/>	_started_7812980284244228796	-	03 Apr 2022 06:48:35 PM CEST	0 o	Standard
<input type="checkbox"/>	_SUCCESS	-	03 Apr 2022 06:49:35 PM CEST	0 o	Standard
<input type="checkbox"/>	part-00000-tid-7812980284244228796-d2186417-025a-4e60-acab-aeae88ff6b9e-1540-1-c000.csv	csv	03 Apr 2022 06:49:31 PM CEST	671.0 o	Standard
<input type="checkbox"/>	part-00001-tid-7812980284244228796-d2186417-025a-4e60-acab-aeae88ff6b9e-1541-1-c000.csv	csv	03 Apr 2022 06:49:27 PM CEST	672.0 o	Standard
<input type="checkbox"/>	part-00002-tid-7812980284244228796-d2186417-025a-4e60-acab-aeae88ff6b9e-1542-1-c000.csv	csv	03 Apr 2022 06:49:27 PM CEST	683.0 o	Standard
<input type="checkbox"/>	part-00003-tid-7812980284244228796-d2186417-025a-4e60-acab-aeae88ff6b9e-1543-1-c000.csv	csv	03 Apr 2022 06:49:32 PM CEST	680.0 o	Standard
<input type="checkbox"/>	part-00004-tid-7812980284244228796-d2186417-025a-4e60-acab-aeae88ff6b9e-1544-1-c000.csv	csv	03 Apr 2022 06:49:27 PM CEST	678.0 o	Standard
<input type="checkbox"/>	part-00005-tid-7812980284244228796-d2186417-025a-4e60-acab-aeae88ff6b9e-1545-1-c000.csv	csv	03 Apr 2022 06:49:31 PM CEST	669.0 o	Standard
<input type="checkbox"/>	part-00006-tid-7812980284244228796-d2186417-025a-4e60-acab-aeae88ff6b9e-1546-1-c000.csv	csv	03 Apr 2022 06:49:27 PM CEST	683.0 o	Standard
<input type="checkbox"/>	part-00007-tid-7812980284244228796-d2186417-025a-4e60-acab-aeae88ff6b9e-1547-1-c000.csv	csv	03 Apr 2022 06:49:20 PM CEST	375.0 o	Standard

Activités LibreOffice Calc

part-00000-tid-7812980284244228796-d2186417-025a-4e60-acab-aeae88ff6b9e-1540-1-c000.csv - LibreOffice Calc

File Edit View Insert Format Styles Sheet Data Tools Window Help

Libération Sa 10 B I U A

A:13:AMU43 fx Σ =

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	path																
2	df_result/sectionfruits/00000000_100_100.jpg	Volunt	-4.794763094319145	-10.0403899663608	-4.146476376525703	12.74897462181854	-37.3987663701477										
3	df_result/sectionfruits/00000000_163_100.jpg	Volunt	-6.79948249066895	-14.83081206051549	-4.91032443217986	11.1634032207492	-38.5666580399323										
4	df_result/sectionfruits/00000000_49_100.jpg	Volunt	-1.62025393228872	-4.83889223147938	-3.49848436973178	15.5566651598248	-38.624906261876										
5	df_result/sectionfruits/00000000_132_100.jpg	Volunt	-7.29260136020455	-15.2528657696144	-6.00800191781721	11.9768385929155	-39.176778728599										
6																	
7																	
8																	
9																	
10																	
11																	
12																	
13																	
14																	
15																	
16																	
17																	
18																	
19																	
20																	
21																	
22																	
23																	
24																	
25																	
26																	
27																	
28																	
29																	
30																	
31																	
32																	
33																	
34																	
35																	
36																	
37																	
38																	
39																	
40																	
41																	
42																	
43																	

Sheet 1 of 1

Selected: 1 row, 1,024 columns

Default

English (USA)

Average: Sum: 0

100%

Partie 4: Conclusion et points d'amélioration

Conclusion

- Mission de start-up: Introduire une mobile application qui 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.
- Mission de projet: Développer une première chaîne de traitement des données dans un environnement Big Data qui comprendra le preprocessing et une étape de réduction de dimension.
- AWS avec Plateforme Databricks
- Apach Spark
- Prétraitements de données et Réduction de dimension

Points d'amélioration

- Evoquer pour regarder s'il existe une **autre offre de cloud** afin de comparer avec Aws et databricks bien que AWS est l'un des meilleurs.
- Appliquer EC2 directement sur AWS.
- **Limitation du stockage** d'images sur s3 car il a une implication dans le temps de calcul lors l'extraction de features et la réduction de dimension.

Merci pour
votre attention

