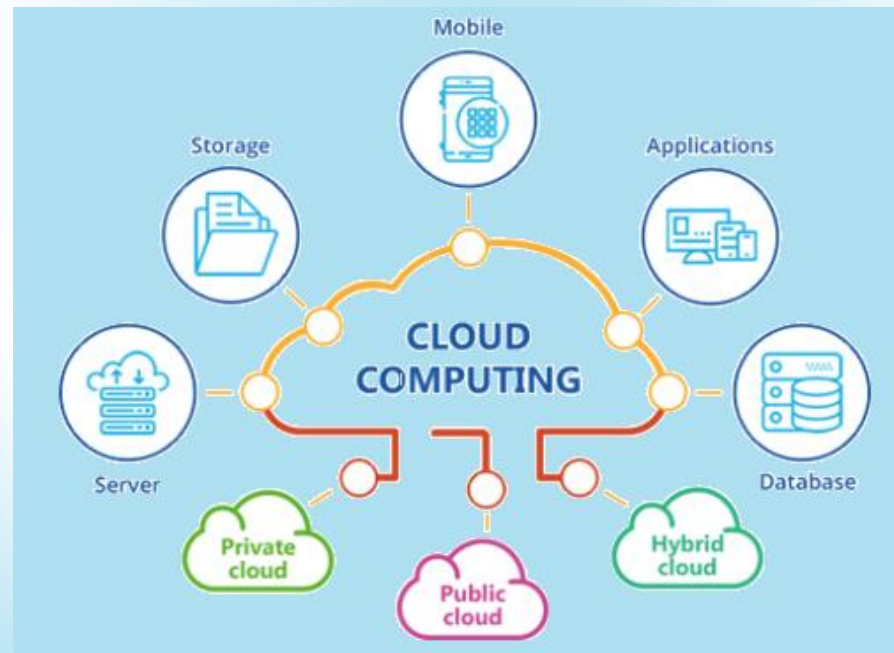


Déployez un modèle dans le cloud



-  1 Problématique
-  2 Données
-  3 Big Data
-  4 Chaîne de traitement
-  5 Conclusions

-  1 Problématique
-  2 Données
-  3 Big Data
-  4 Chaîne de traitement
-  5 Conclusions



Fruits!



Entreprise *Fruits!*

- Startup de l'AgriTech
- L'IA au service de l'agriculture



Etape 1 :

application mobile grand public de reconnaissance de fruit par photographie

- Classification d'images (volume exponentiel d'images)



Etape 2 :

robots cueilleurs intelligents

- Dans une mission ultérieure

Mission

- ☐ Mettre en place une **architecture Big Data**
- ☐ Préparer les données :
 - **Pré-processing**
 - **Réduction de dimension**

Contraintes

- ☐ Anticiper **passage à l'échelle**
(volume exponentiel, calculs distribués)
- ☐ Scripts **PySpark**
- ☐ Déploiement cloud 

Objectifs

- ☐ Faire connaître la startup
- ☐ Classification d'images pour application mobile

- 1 Problématique
- 2 Données
- 3 Big Data
- 4 Chaîne de traitement
- 5 Conclusions

90380 images
131 classes

Jeux avec étiquettes



- ☐ Images de 1 fruit ou 1 légume.
- ☐ 120 variétés différentes.
- ☐ Fond blanc, 100x100 pixels, en couleur, centré.



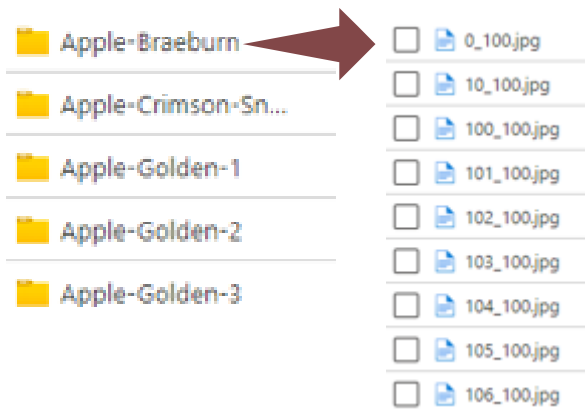
Photos fruit de cactus 360° sur 3 axes

67692 images
131 classes

Jeu entraînement

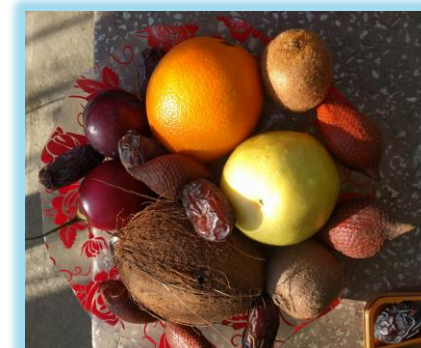
22688 images
131 classes

Jeu test



103 images
Multi fruits

Jeu sans étiquettes

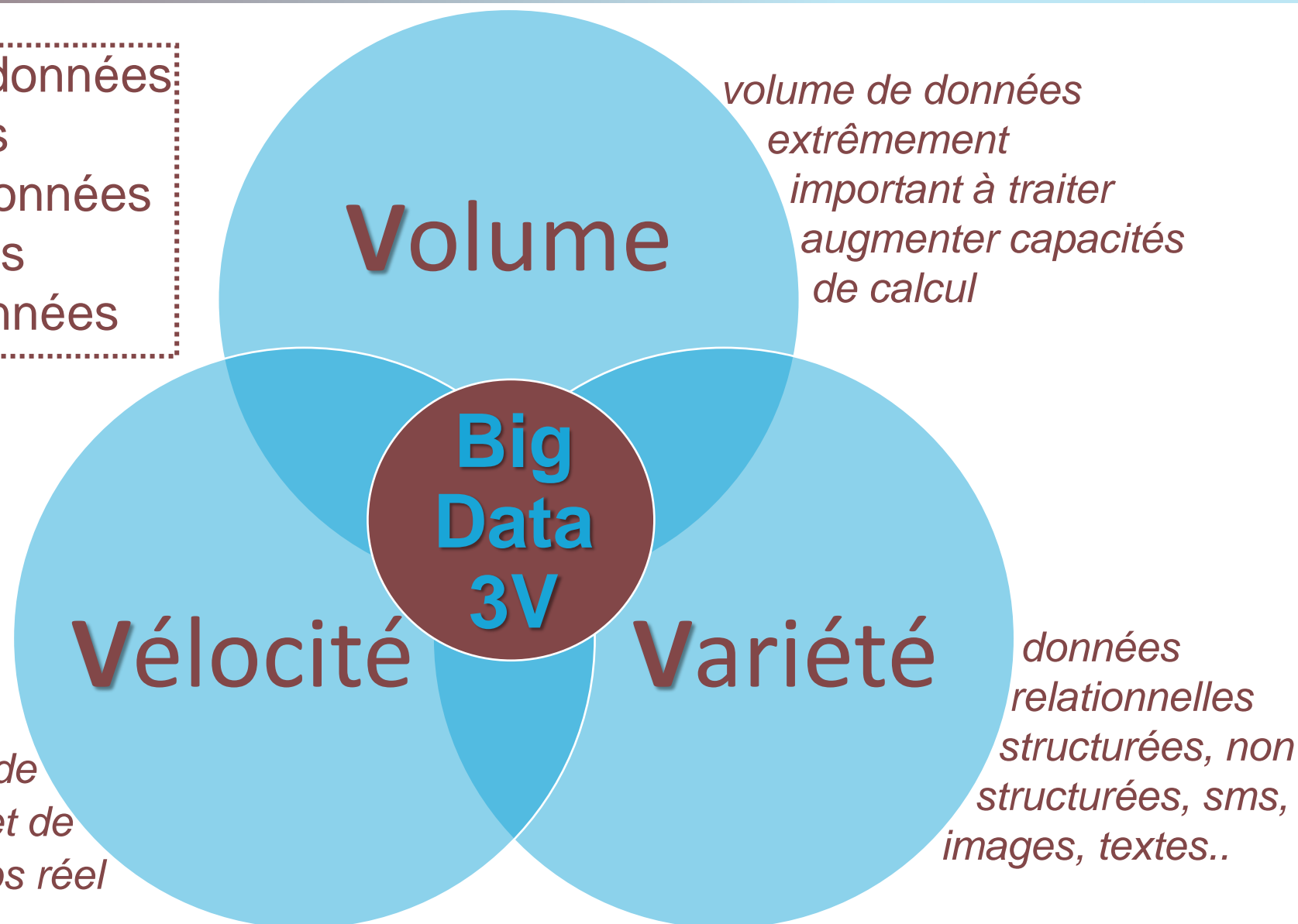


- 1 Problématique
- 2 Données
- 3 Big Data**
- 4 Chaîne de traitement
- 5 Conclusions

Explosion de la quantité de données
Partage des données
Analyse/visualisation des données
Stockage des données
Traitement des flux de données

Big Data

*besoin de capacités de
diffusion en continu et de
traitement en temps réel*

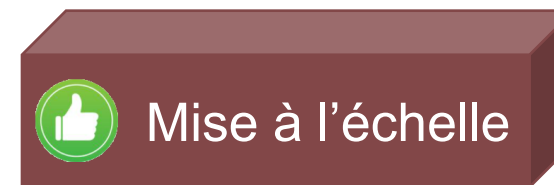
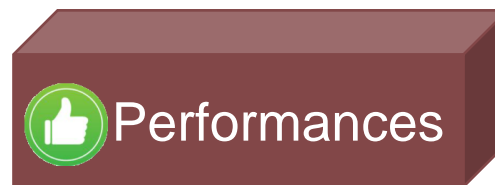
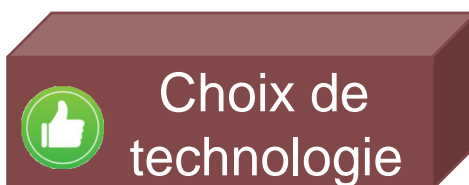
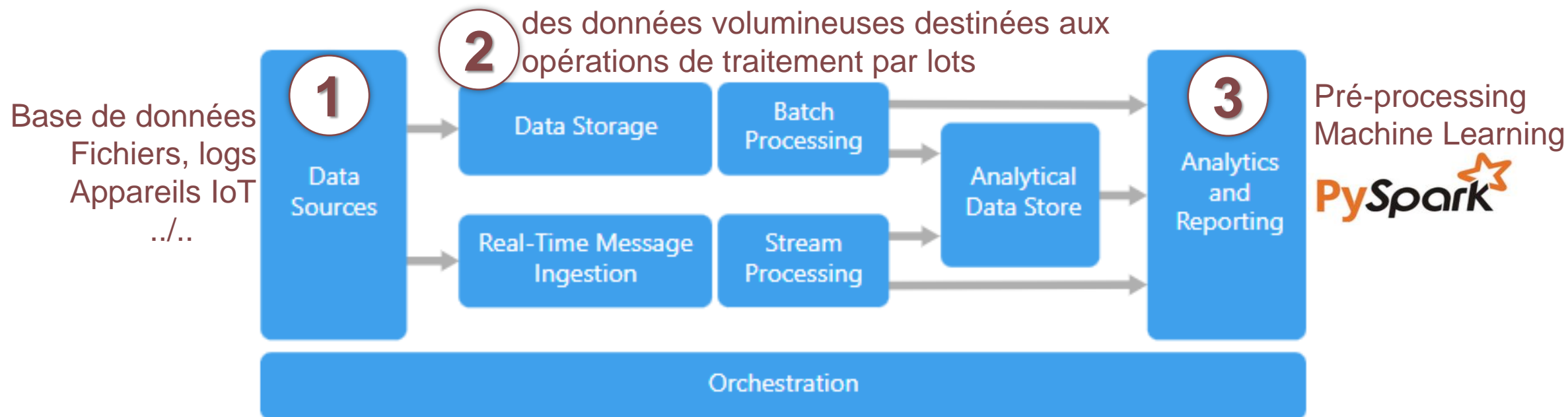


3 Big Data – Architecture



Fruits!

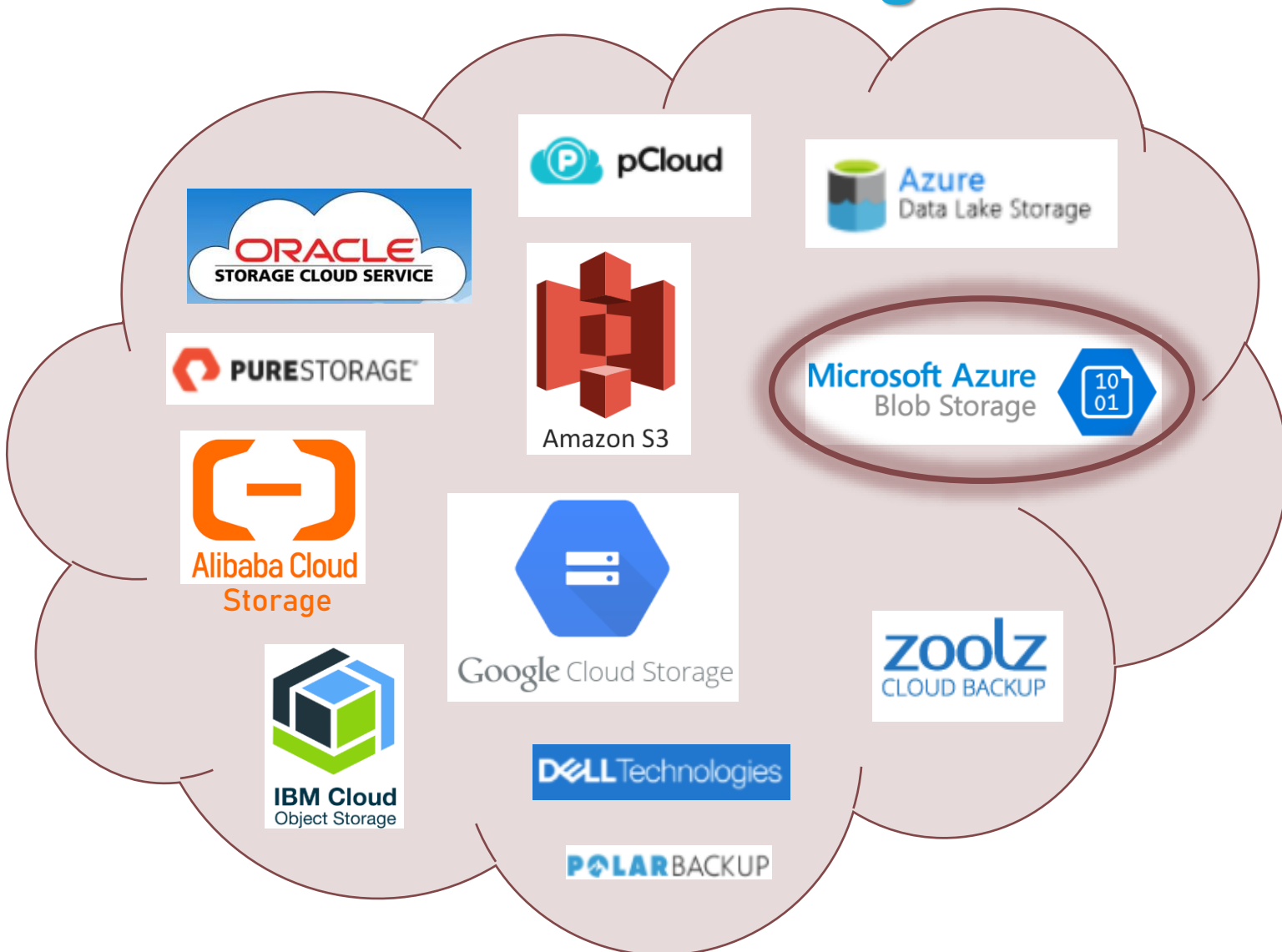
Stockage dans un magasin de
fichiers distribués : « Lac de données »
des données volumineuses destinées aux
opérations de traitement par lots



Pay As You Go



Fournisseurs Stockage Cloud



Données

Échelle illimitée
Durabilité, disponibilité
Géo-réplication



Sécurité

Contrôle de l'accès, authentification (rôles)
Chiffrement et contrôle réseau



Service

Ingestion scalabilité, traitement, visualisation des données
Frameworks analytiques courants supportés



Coûts

Mise à l'échelle indépendante du stockage et du calcul
Stratégie de cycle de vie
Pay as you go

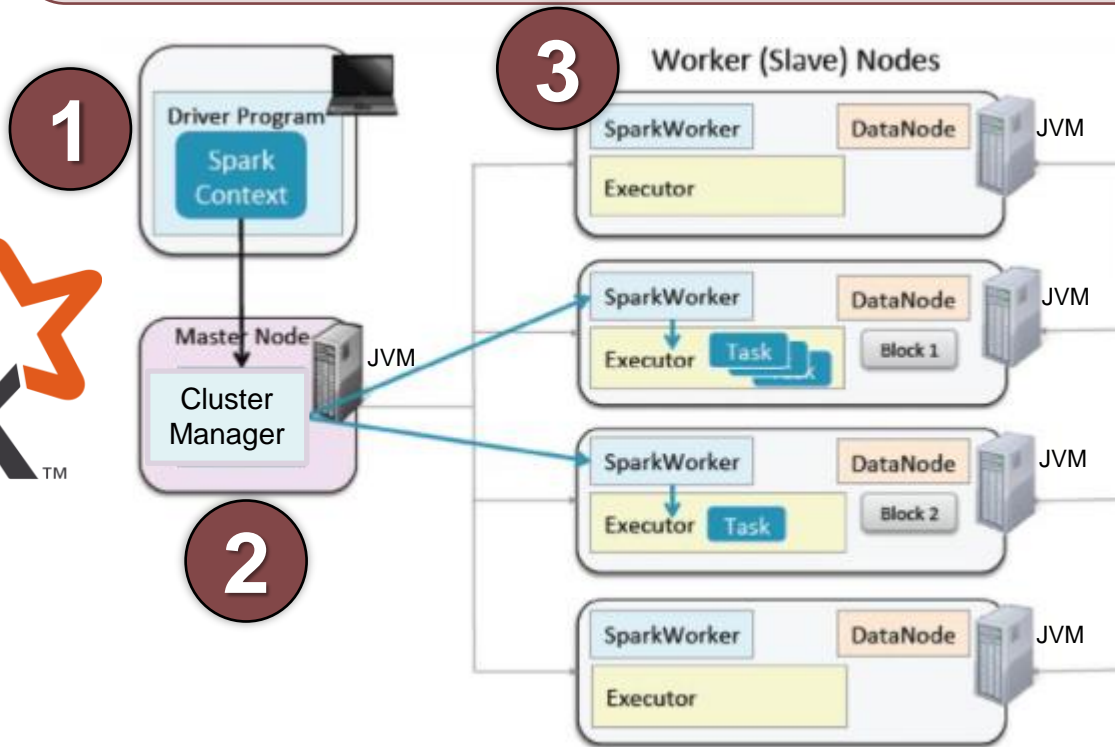
Parallélisation des calculs sur plusieurs machines.

Comment :



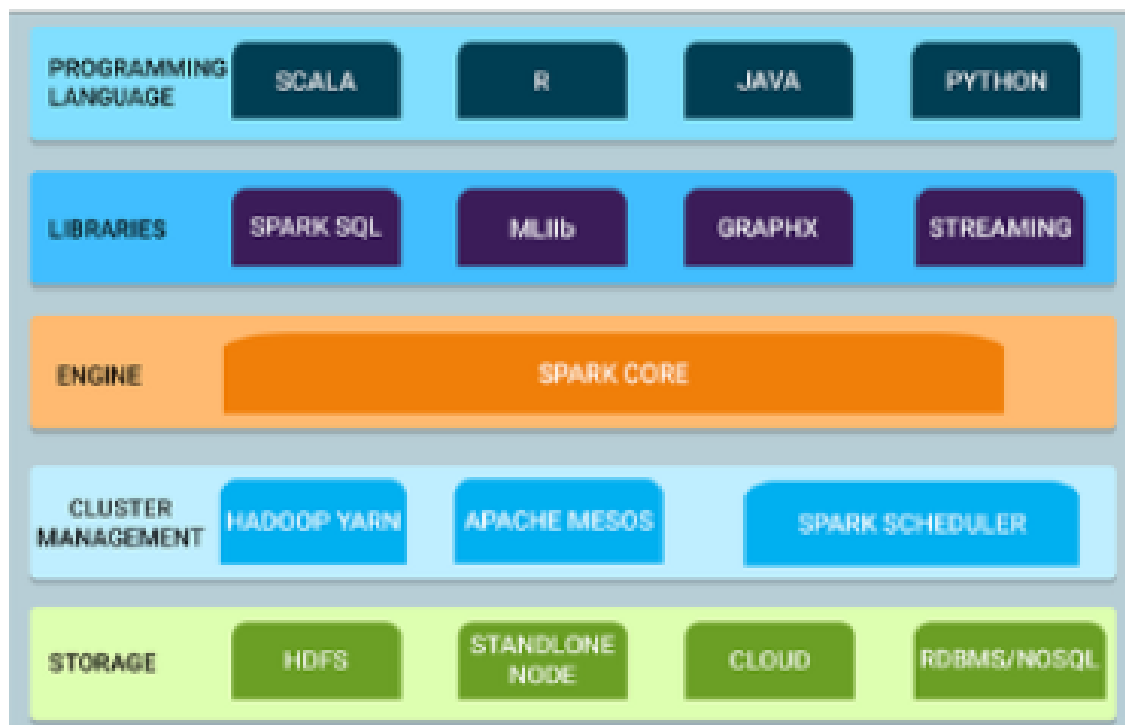
distribuer les calculs entre les machines?
agréger les résultats des différentes machines?
maîtriser les coûts, gérer les pannes...?

Utiliser :

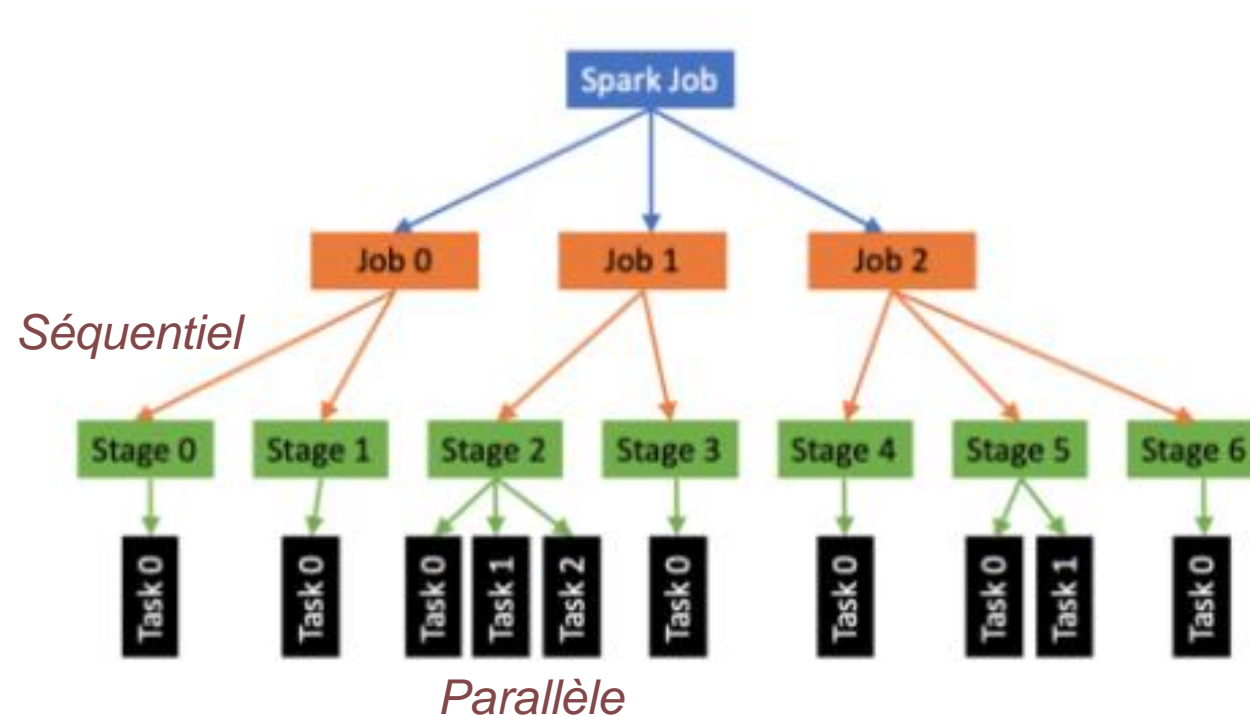


- 1 Configuration, initialisation
Agrégation des calculs
- 2 Gestion des ressources
Distribution des calculs
entre workers
- 3 Exécution des tâches en
parallèles

Framework Spark



Spark job



Basé sur Hadoop map/reduce + traitement “in memory”.
Basé sur Resilient Distributed Datasets RDD + Spark DataFrames.
→ permettant la parallélisation des opérations (transformations ou actions).
→ Tolérant aux pannes grâce aux graphes acycliques orientés.

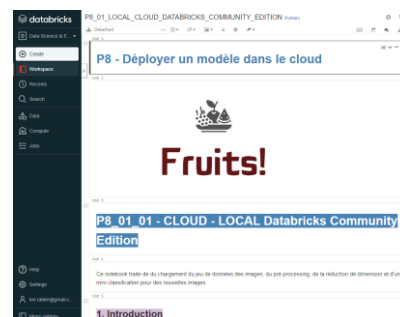


3 Architecture de développement retenue



Fruits!

 databricks
Community Edition



APACHE
SparkTM

PySpark

Version gratuite de la plateforme databricks
Spark basée sur le cloud.

Accès à :

- un micro-cluster,
- un gestionnaire notebook IPython
- un environnement partageable pour prototyper des applications simples.
- Stockage

**Plateforme
(as a Service)**

Données et accès

Applications

Runtime

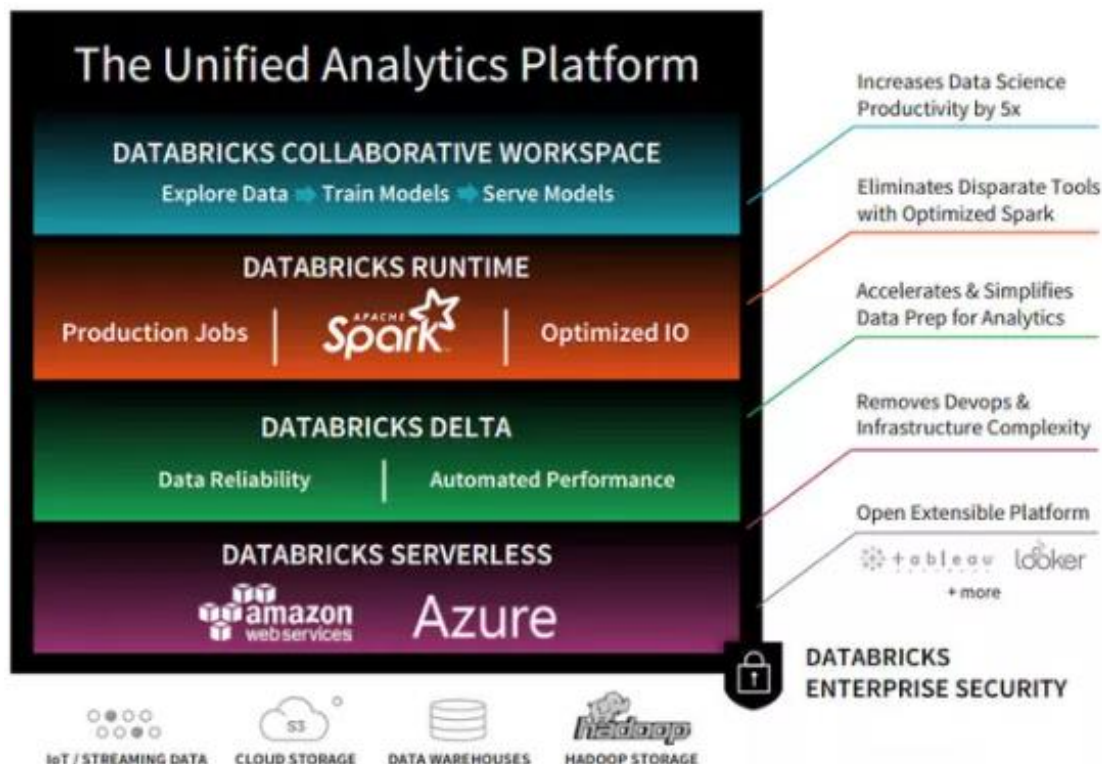
Système
d'exploitation

Machine virtuelle

Calcul

Réseaux

Comptes



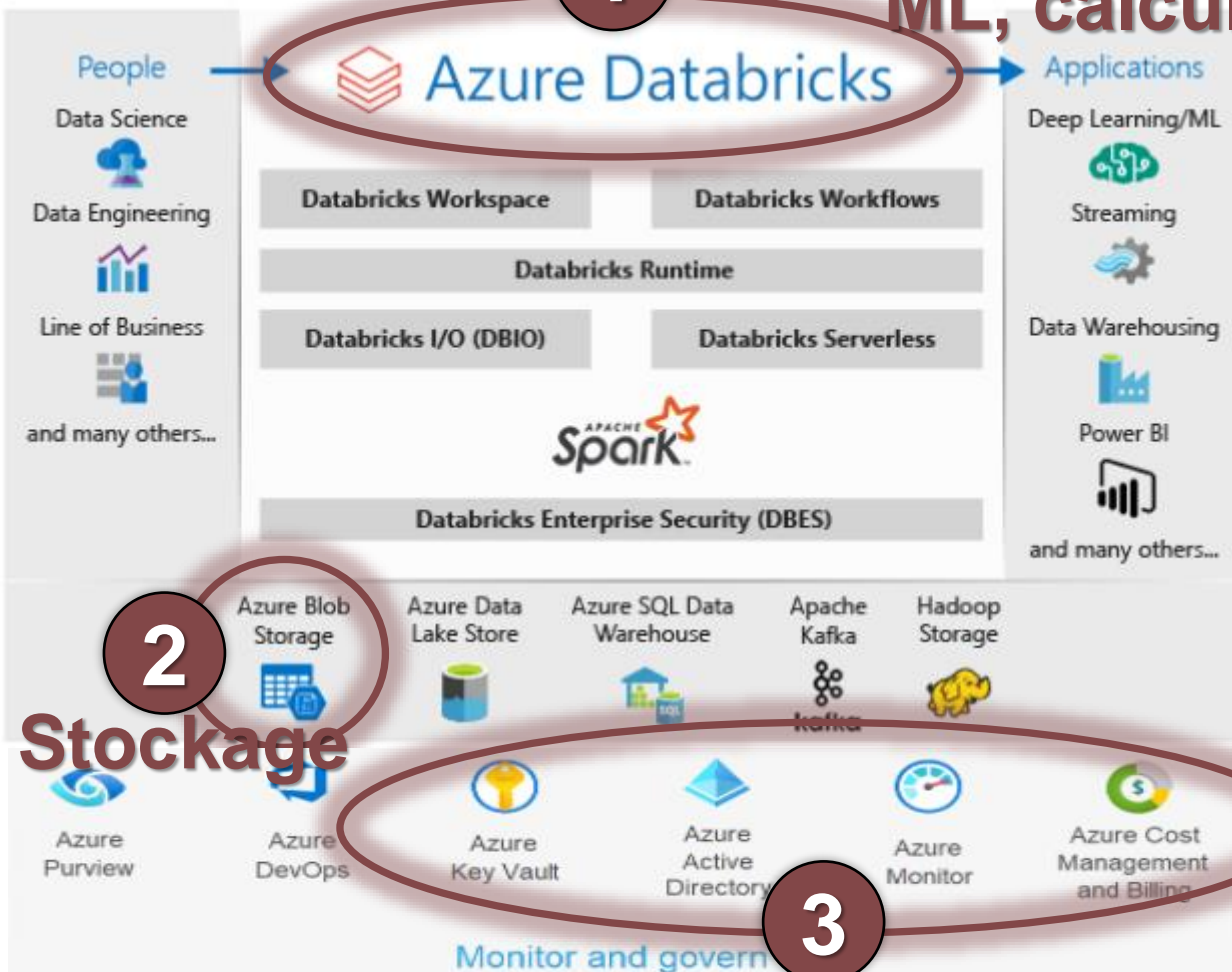
3 Architecture Cloud revenue



Fruits!

1

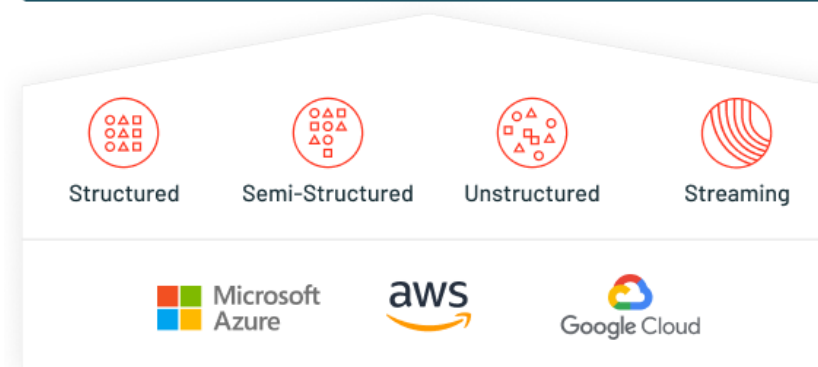
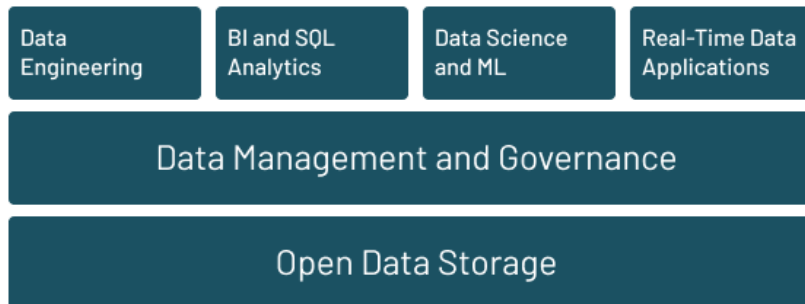
Cluster, notebook,
ML, calculs



2
Stockage

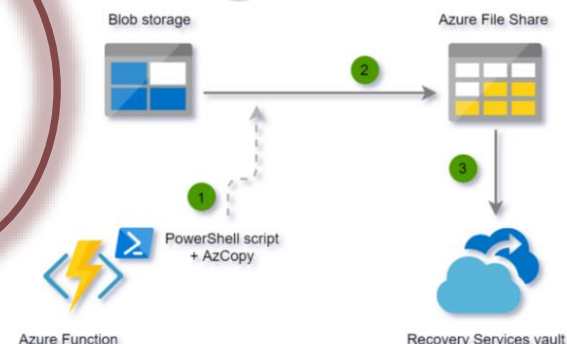
3

Authentification, contrôle



4

Chargement



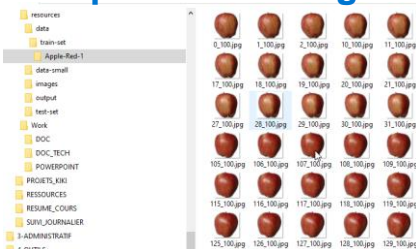
- 1 Problématique
- 2 Données
- 3 Big Data
- 4 Chaîne de traitement**
- 5 Conclusions

4 Chaîne de traitement



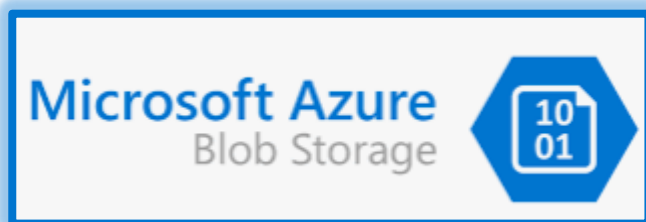
Fruits!

Disque local : images



Script
powerShell

Stockage
Cloud

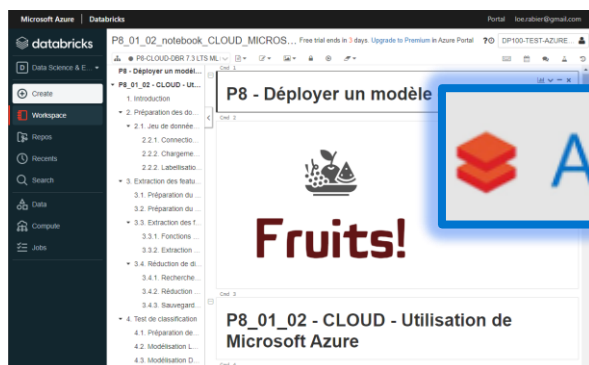


Signature d'accès partagé

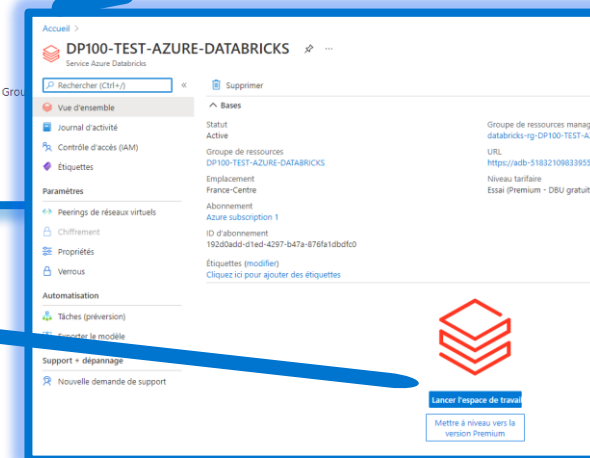
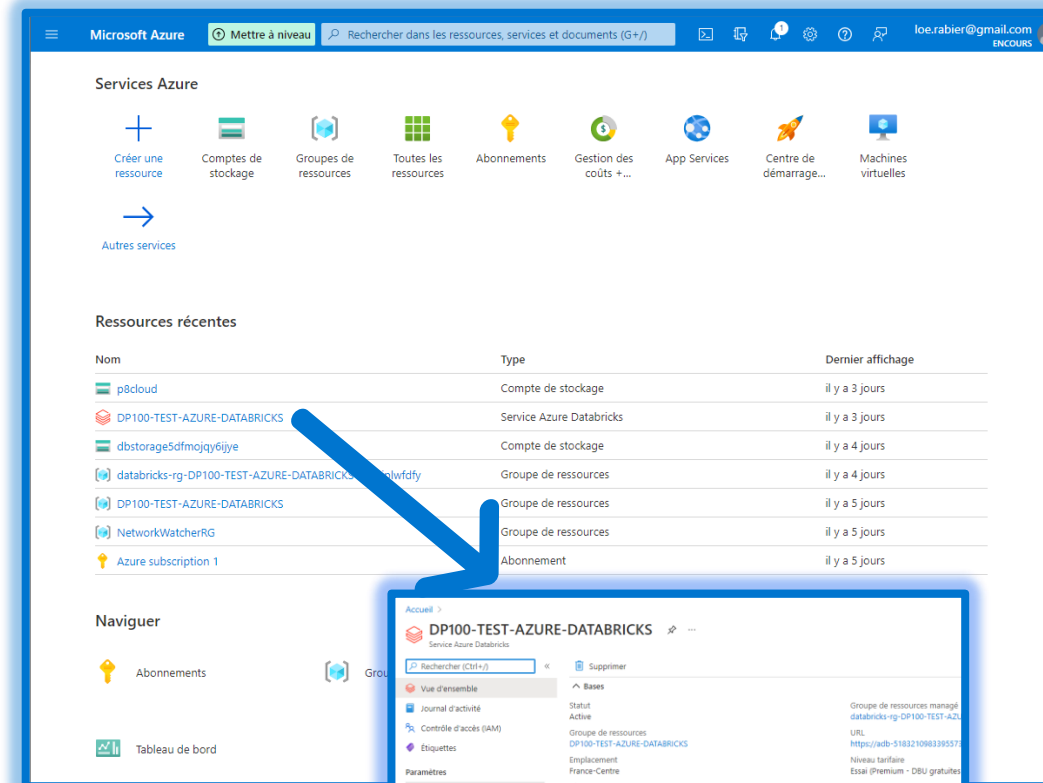
Écriture

SAS
Token

Lecture



Portail Microsoft Azure



4 Chaîne de traitement - Portail Microsoft Azure



Fruits!

Microsoft Azure

Rechercher dans les ressources, services et documents (G+/)

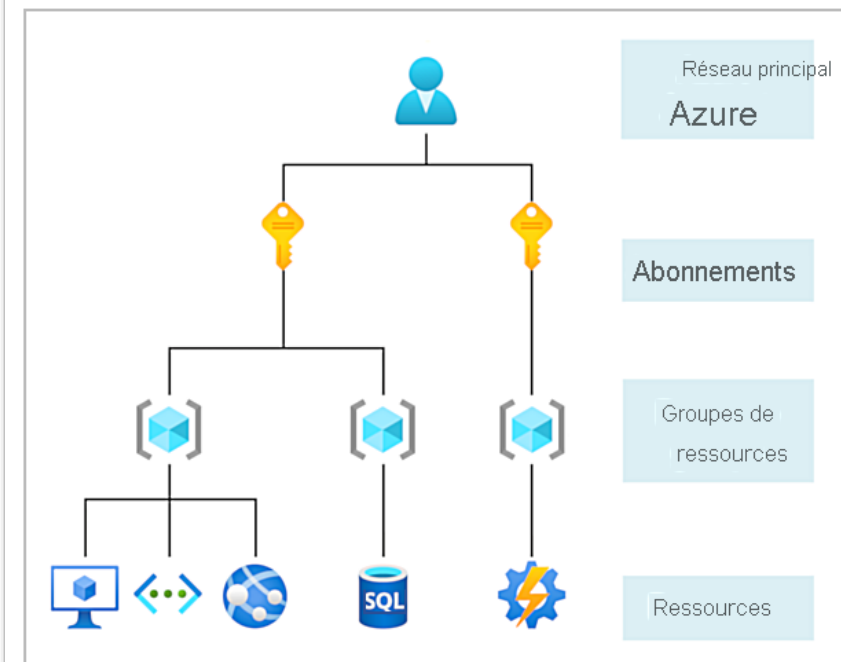
Services Azure

- Créer une ressource
- Comptes de stockage
- Groupes de ressources
- Toutes les ressources
- Abonnements
- Gestion des coûts +...
- App Services
- Centre de démarrage...
- Machines virtuelles

Autres services

Ressources récentes

Nom	Type	Dernier affichage
DP100-TEST-AZURE-DATABRICKS	Service Azure Databricks	il y a 5 minutes
p8cloud	Compte de stockage	il y a 19 minutes
dbstorage5dfmojqy6ijye	Compte de stockage	il y a 5 jours
databricks-rg-DP100-TEST-AZURE-DATABRICKS-kziip1wfdy	Groupe de ressources	il y a 5 jours
DP100-TEST-AZURE-DATABRICKS	Groupe de ressources	il y a 5 jours
NetworkWatcherRG	Groupe de ressources	il y a 5 jours
Azure subscription 1	Abonnement	il y a 5 jours



4 Chaîne de traitement – Stockage Azure Blob



Fruits!

Microsoft Azure

Rechercher dans les ressources, services et documents (G+/)

lo.e.rabier@gmail.com ENCOURS

Accueil > p8cloud > Compte de stockage

Rechercher (Ctrl+Q)

Abonnement (modifier) : Azure subscription 1

ID d'abonnement : 192d0add-d1ed-4297-b47a-87...

État du disque : Principal : Disponible, secondai...

État de provisionnement : Réussite

Créé : 09/09/2021, 08:51:35

Étiquettes (modifier) :

Propriétés Supervision Fonctionnalités (5) Recommandations Tutoriels Outils de développement

Data Lake Storage

Espace de noms hiérarchique	Activé
Niveau d'accès par défaut	Cool
Accès public aux objets blob	Activé
Suppression réversible d'objet blob	Désactivé
Suppression réversible de conteneur	Activé (7 jours)
Gestion des versions	Désactivé
Flux de modification	Désactivé
NFS v3	Désactivé

Service Fichier

Grand partage de fichiers	Désactivé
Active Directory	Non configuré
Suppression réversible	Activé (7 jours)
Capacité de partage	5 Tio

Service File d'attente

Prise en charge de CMK	Désactivé
------------------------	-----------

Service Table

Prise en charge de CMK	Désactivé
------------------------	-----------

Sécurité

Exiger un transfert sécurisé pour les opérations d'API REST	Activé
Accès de clé de compte de stockage	Activé
Version TLS minimale	Version 1.2
Chiffrement d'infrastructure	Désactivé

Réseau

Autoriser l'accès à partir de	Tous les réseaux
Nombre de connexions de point de terminaison privé	0
Routage réseau	Routage réseau Microsoft
Accès pour les services Microsoft approuvés	Oui

Microsoft Azure

Rechercher dans les ressources, services et documents (G+/)

lo.e.rabier@gmail.com ENCOURS

Accueil > p8cloud > Conteneur

Rechercher (Ctrl+Q)

Charger + Ajouter un répertoire Actualiser Renommer Supprimer Modifier le niveau

Méthode d'authentification : Clé d'accès (Basculer sur un compte d'utilisateur Azure AD)

Emplacement : p8-cloud / resources / data / train-set / Apricot

Rechercher les objets blobs par préfixe (respect de la casse)

Nom	Modifié	Niveau d'accès	Type d'objet blob	Taille
[.]				
0_100.jpg	09/09/2021, 10:30:36	Réduit (déduit)	Objet blob de blocs	4.35 Ki
10_100.jpg	09/09/2021, 10:30:36	Réduit (déduit)	Objet blob de blocs	4.37 Ki
100_100.jpg	09/09/2021, 10:30:25	Réduit (déduit)	Objet blob de blocs	4.02 Ki
101_100.jpg	09/09/2021, 10:30:29	Réduit (déduit)	Objet blob de blocs	4.04 Ki
102_100.jpg	09/09/2021, 10:30:37	Réduit (déduit)	Objet blob de blocs	3.99 Ki
103_100.jpg	09/09/2021, 10:30:31	Réduit (déduit)	Objet blob de blocs	3.99 Ki
104_100.jpg	09/09/2021, 10:30:37	Réduit (déduit)	Objet blob de blocs	4.07 Ki
105_100.jpg	09/09/2021, 10:30:28	Réduit (déduit)	Objet blob de blocs	4.03 Ki
106_100.jpg	09/09/2021, 10:30:24	Réduit (déduit)	Objet blob de blocs	4.06 Ki
107_100.jpg	09/09/2021, 10:30:26	Réduit (déduit)	Objet blob de blocs	4.02 Ki
108_100.jpg	09/09/2021, 10:30:35	Réduit (déduit)	Objet blob de blocs	4.02 Ki
109_100.jpg	09/09/2021, 10:30:26	Réduit (déduit)	Objet blob de blocs	3.98 Ki
11_100.jpg	09/09/2021, 10:30:31	Réduit (déduit)	Objet blob de blocs	4.38 Ki
110_100.jpg	09/09/2021, 10:30:38	Réduit (déduit)	Objet blob de blocs	3.97 Ki
111_100.jpg	09/09/2021, 10:30:29	Réduit (déduit)	Objet blob de blocs	3.99 Ki
112_100.jpg	09/09/2021, 10:30:26	Réduit (déduit)	Objet blob de blocs	4.1 KiB
113_100.jpg	09/09/2021, 10:30:24	Réduit (déduit)	Objet blob de blocs	4.07 Ki
114_100.jpg	09/09/2021, 10:30:33	Réduit (déduit)	Objet blob de blocs	4.07 Ki

4 Chaîne de traitement – Azure Databricks



Fruits!

Microsoft Azure | Recherchez dans les ressources, services et documents (G+)

Accueil >

DP100-TEST-AZURE-DATABRICKS Service Azure Databricks

Rechercher (Ctrl+J) Supprimer

Vue d'ensemble

- Journal d'activité
- Contrôle d'accès (IAM)
- Étiquettes

Paramètres

- Peerings de réseaux virtuels
- Chiffrement
- Propriétés
- Verrous

Automatisation

- Tâches (préversion)
- Exporter le modèle

Support + dépannage

- Nouvelle demande de support

Bases

Statut: Active

Groupe de ressources: DP100-TEST-AZURE-DATABRICKS

Emplacement: France-Centre

Abonnement: Azure subscription 1

ID d'abonnement: 192d0add-d1ed-4297-b47a-876fa1dbdfc0

Étiquettes (modifier): Cliquez ici pour ajouter des étiquettes

Groupe de ressources managé: databricks-rg-DP100-TEST-AZURE-DATABRICKS-kziip1wdfy

URL: https://adb-5183210983395573.13.azuredatabricks.net

Niveau tarifaire: Essai (Premium - DBU gratuites pendant 14 jours)

Lancer l'espace de travail

Mettre à niveau vers la version Premium

Microsoft Azure | Databricks

Portal | loe.rabier@gmail.com

databricks

Data Science & E...

Create

Workspace

Repos

Recents

Search

Data

Compute

Jobs

P8_01_02_notebook_CLOUD_MICROS... Free trial ends in 3 days. Upgrade to Premium in Azure Portal

P8 - Déployer un modèle dans le cloud

1. Introduction

2. Préparation des données

2.1. Jeu de données

2.2.1. Connexion

2.2.2. Chargement

2.2.2. Labellisation

3. Extraction des caractéristiques

3.1. Préparation du jeu de données

3.2. Préparation du modèle

3.3. Extraction des caractéristiques

3.3.1. Fonctions de réduction de dimension

3.3.2. Extraction des caractéristiques

3.4. Réduction de dimension

3.4.1. Recherche de la meilleure réduction de dimension

3.4.2. Réduction de dimension

3.4.3. Sauvegarde du modèle

4. Test de classification

4.1. Préparation des données

4.2. Modélisation L...

4.3. Modélisation D...

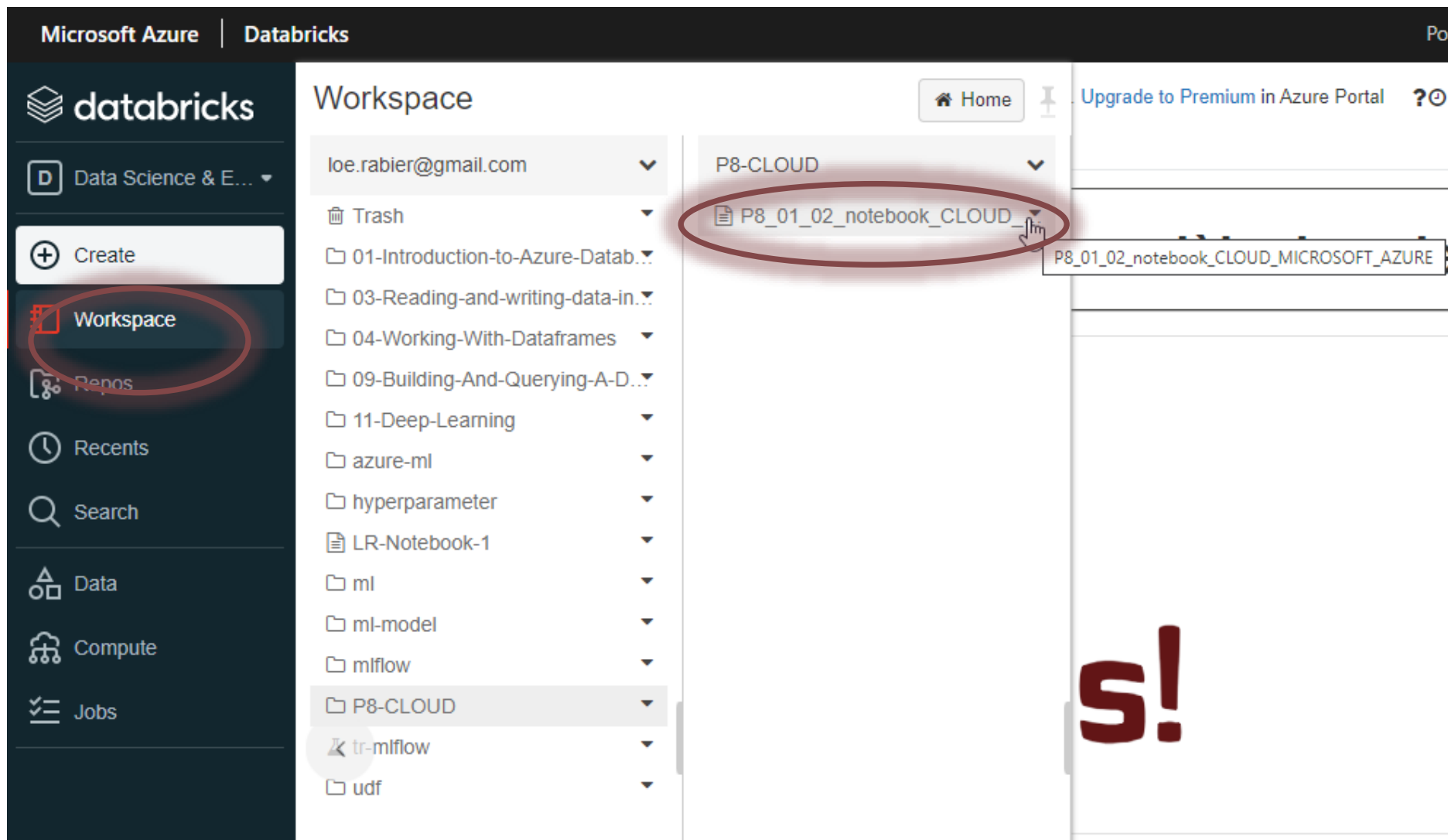
4.4. Modélisation R...

Fruits!

P8_01_02 - CLOUD - Utilisation de Microsoft Azure

Ce notebook traite de du chargement du jeu de données des images, du pré-processing, de la réduction de dimension et d'une mini classification pour des nouvelles images en utilisant l'outil **Microsoft Databricks Azure** pour la partie "compute" (calculs distribués) et un **container blob de stockage "Data Lake Storage"**.

Workspace - Notebooks



Microsoft Azure | Databricks

Workspace

loe.rabier@gmail.com

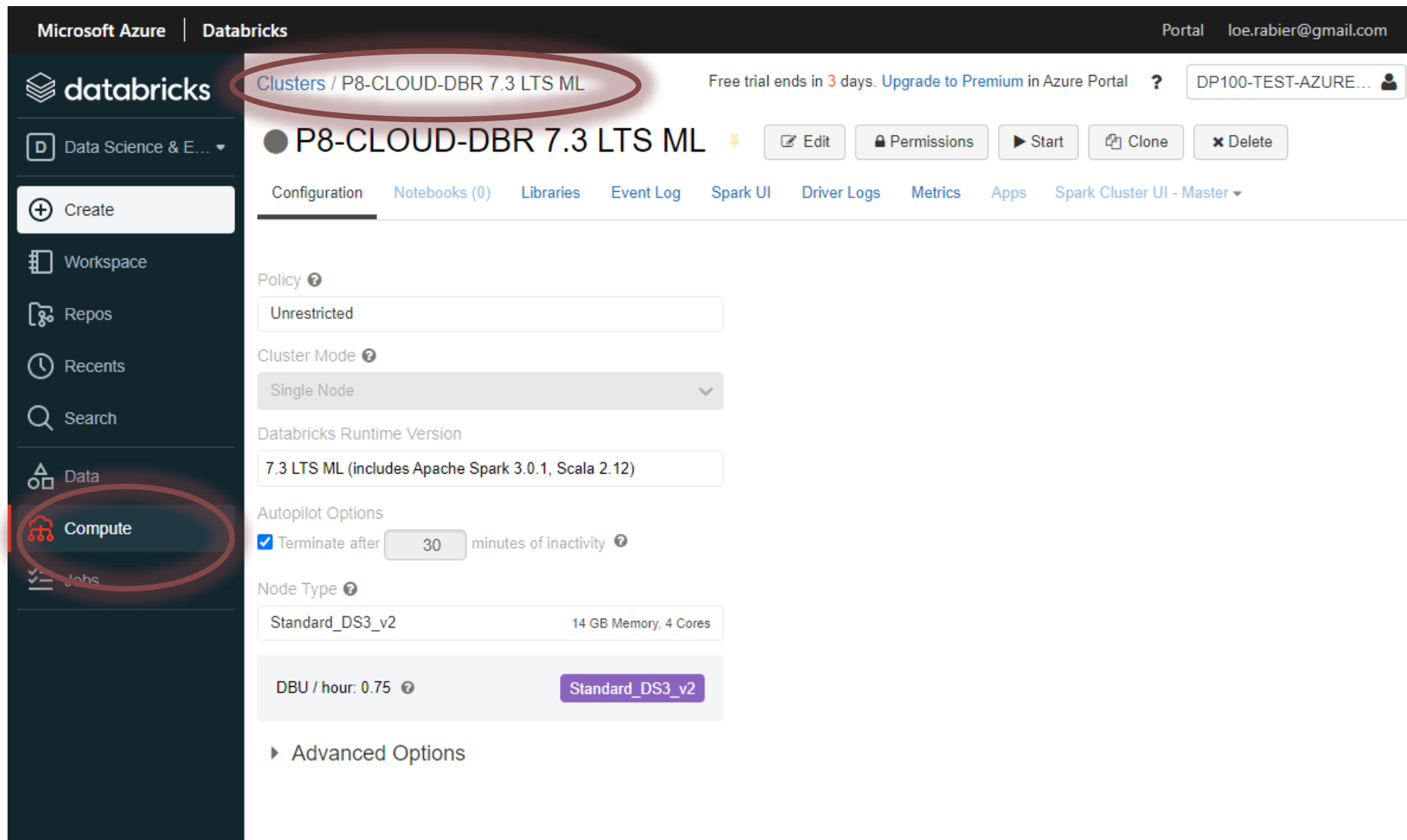
P8-CLOUD

P8_01_02_notebook_CLOUD

P8_01_02_notebook_CLOUD_MICROSOFT_AZURE

S!

Cluster



The screenshot shows the Azure Databricks interface. The top navigation bar includes 'Microsoft Azure | Databricks', a 'Portal' link, and the user email 'loe.rabier@gmail.com'. The left sidebar contains navigation items: 'Data Science & E...', 'Create', 'Workspace', 'Repos', 'Recents', 'Search', 'Data', 'Compute', and 'Jobs'. The 'Compute' item is highlighted with a red circle. The main content area displays the 'Clusters' page for a specific cluster named 'P8-CLOUD-DBR 7.3 LTS ML', which is also circled in red. The breadcrumb path is 'Clusters / P8-CLOUD-DBR 7.3 LTS ML'. The cluster details include: 'P8-CLOUD-DBR 7.3 LTS ML' with buttons for 'Edit', 'Permissions', 'Start', 'Clone', and 'Delete'; tabs for 'Configuration', 'Notebooks (0)', 'Libraries', 'Event Log', 'Spark UI', 'Driver Logs', 'Metrics', 'Apps', and 'Spark Cluster UI - Master'; 'Policy' set to 'Unrestricted'; 'Cluster Mode' set to 'Single Node'; 'Databricks Runtime Version' set to '7.3 LTS ML (includes Apache Spark 3.0.1, Scala 2.12)'; 'Autopilot Options' with 'Terminate after 30 minutes of inactivity' checked; 'Node Type' set to 'Standard_DS3_v2' with '14 GB Memory, 4 Cores'; and 'DBU / hour: 0.75' with a 'Standard_DS3_v2' button. An 'Advanced Options' link is at the bottom.

4 Chaîne de traitement – Azure DataBricks



Fruits!

Données

Microsoft Azure | Databricks

Portal loe.rabier@gmail.com

Free trial ends in 3 days. [Upgrade to Premium](#) in Azure Portal ? DP100-TEST-AZURE...

Create New Table

Data source ?

[Upload File](#) **[DBFS](#)** [Other Data Sources](#) [Partner Integrations](#)

Select a file from DBFS ?

FileStore	hive	cvPipelineModel	_SUCCESS
ml	loe.rabier@gmail.com	deep_learning	_committed_74366304
mlflow		delta	_started_74366304811
mnt		modelPredictions.parquet	part-00000-tid-7436630
tmp		pageviews_by_second.parquet	part-00001-tid-7436630
user		people.parquet	part-00002-tid-7436630
			part-00003-tid-7436630
			part-00004-tid-7436630
			part-00005-tid-7436630
			part-00006-tid-7436630
			part-00007-tid-7436630

/user/loe.rabier@gmail.com/people.parquet

[Create Table with UI](#) [Create Table in Notebook](#) ?

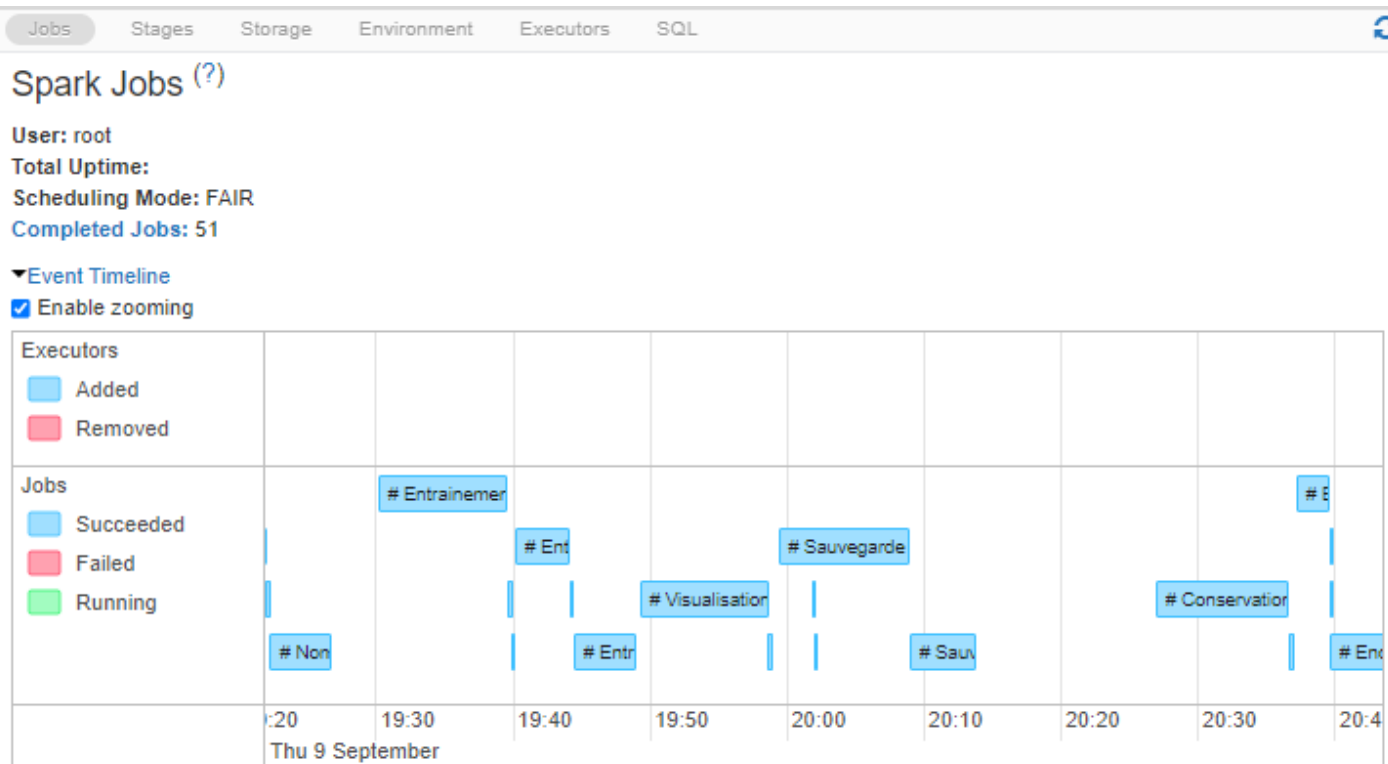
data

4 Chaîne de traitement – Spark UI

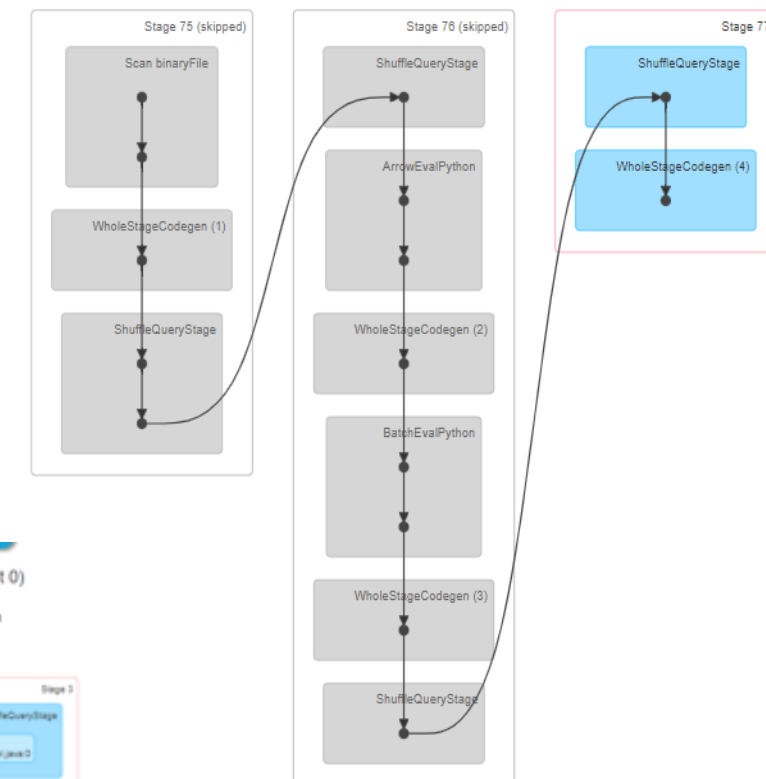


Fruits!

EventTimeline



DAG



Stage

Details for Stage 3 (Attempt 0)

Total Time Across All Tasks: 79 ms
Locality Level Summary: Process local: 1
Shuffle Read: 1045 B / 10

▼ DAG Visualization

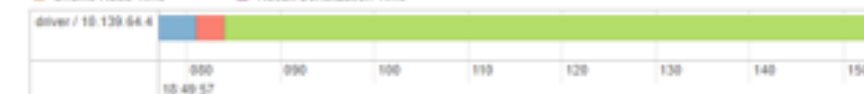


► Show Additional Metrics

▼ Event Timeline

☐ Enable zooming

Scheduler Delay
Task Deserialization Time
Shuffle Read Time
Executor Computing Time
Shuffle Write Time
Result Serialization Time
Getting Result Time

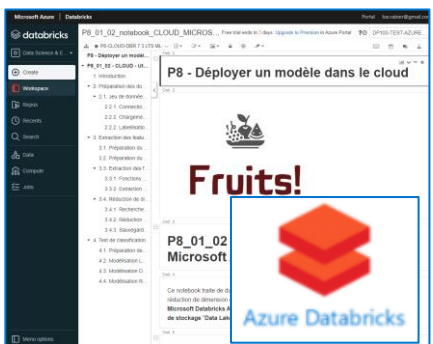
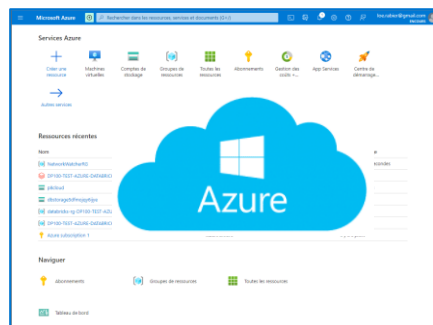


4 Chaîne de traitement – Processus



Fruits!

ENVIRONNEMENT TRAVAIL



CHARGEMENT IMAGES

Accès services partagés

Configuration AzCopy

Stockage Blob

- 1 répertoire = 1 classe
- Toutes les images chargées dans le cloud

Item	Modèle	Nombre d'images	Type d'origine	Taille
12	12	10,000	Image (Blob)	10,000
13	13	10,000	Image (Blob)	10,000
14	14	10,000	Image (Blob)	10,000
15	15	10,000	Image (Blob)	10,000
16	16	10,000	Image (Blob)	10,000
17	17	10,000	Image (Blob)	10,000
18	18	10,000	Image (Blob)	10,000
19	19	10,000	Image (Blob)	10,000
20	20	10,000	Image (Blob)	10,000
21	21	10,000	Image (Blob)	10,000
22	22	10,000	Image (Blob)	10,000
23	23	10,000	Image (Blob)	10,000
24	24	10,000	Image (Blob)	10,000
25	25	10,000	Image (Blob)	10,000
26	26	10,000	Image (Blob)	10,000
27	27	10,000	Image (Blob)	10,000
28	28	10,000	Image (Blob)	10,000
29	29	10,000	Image (Blob)	10,000
30	30	10,000	Image (Blob)	10,000
31	31	10,000	Image (Blob)	10,000
32	32	10,000	Image (Blob)	10,000
33	33	10,000	Image (Blob)	10,000
34	34	10,000	Image (Blob)	10,000
35	35	10,000	Image (Blob)	10,000
36	36	10,000	Image (Blob)	10,000
37	37	10,000	Image (Blob)	10,000
38	38	10,000	Image (Blob)	10,000
39	39	10,000	Image (Blob)	10,000
40	40	10,000	Image (Blob)	10,000
41	41	10,000	Image (Blob)	10,000
42	42	10,000	Image (Blob)	10,000
43	43	10,000	Image (Blob)	10,000
44	44	10,000	Image (Blob)	10,000
45	45	10,000	Image (Blob)	10,000
46	46	10,000	Image (Blob)	10,000
47	47	10,000	Image (Blob)	10,000
48	48	10,000	Image (Blob)	10,000
49	49	10,000	Image (Blob)	10,000
50	50	10,000	Image (Blob)	10,000
51	51	10,000	Image (Blob)	10,000
52	52	10,000	Image (Blob)	10,000
53	53	10,000	Image (Blob)	10,000
54	54	10,000	Image (Blob)	10,000
55	55	10,000	Image (Blob)	10,000
56	56	10,000	Image (Blob)	10,000
57	57	10,000	Image (Blob)	10,000
58	58	10,000	Image (Blob)	10,000
59	59	10,000	Image (Blob)	10,000
60	60	10,000	Image (Blob)	10,000
61	61	10,000	Image (Blob)	10,000
62	62	10,000	Image (Blob)	10,000
63	63	10,000	Image (Blob)	10,000
64	64	10,000	Image (Blob)	10,000
65	65	10,000	Image (Blob)	10,000
66	66	10,000	Image (Blob)	10,000
67	67	10,000	Image (Blob)	10,000
68	68	10,000	Image (Blob)	10,000
69	69	10,000	Image (Blob)	10,000
70	70	10,000	Image (Blob)	10,000
71	71	10,000	Image (Blob)	10,000
72	72	10,000	Image (Blob)	10,000
73	73	10,000	Image (Blob)	10,000
74	74	10,000	Image (Blob)	10,000
75	75	10,000	Image (Blob)	10,000
76	76	10,000	Image (Blob)	10,000
77	77	10,000	Image (Blob)	10,000
78	78	10,000	Image (Blob)	10,000
79	79	10,000	Image (Blob)	10,000
80	80	10,000	Image (Blob)	10,000
81	81	10,000	Image (Blob)	10,000
82	82	10,000	Image (Blob)	10,000
83	83	10,000	Image (Blob)	10,000
84	84	10,000	Image (Blob)	10,000
85	85	10,000	Image (Blob)	10,000
86	86	10,000	Image (Blob)	10,000
87	87	10,000	Image (Blob)	10,000
88	88	10,000	Image (Blob)	10,000
89	89	10,000	Image (Blob)	10,000
90	90	10,000	Image (Blob)	10,000
91	91	10,000	Image (Blob)	10,000
92	92	10,000	Image (Blob)	10,000
93	93	10,000	Image (Blob)	10,000
94	94	10,000	Image (Blob)	10,000
95	95	10,000	Image (Blob)	10,000
96	96	10,000	Image (Blob)	10,000
97	97	10,000	Image (Blob)	10,000
98	98	10,000	Image (Blob)	10,000
99	99	10,000	Image (Blob)	10,000
100	100	10,000	Image (Blob)	10,000

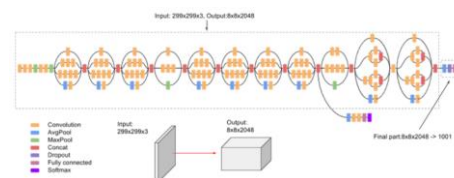
FEATURES EXTRACTION

Images

- Chargement Binary File
- Extraction de la classe

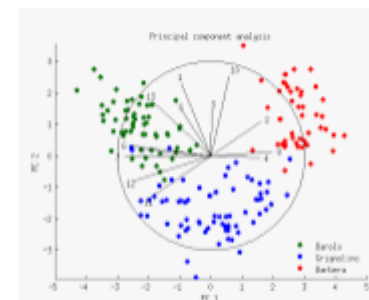
Transfert learning

- Extraction features
- CNN InceptionV3



RÉDUCTION DIMENSION

ACP



Sauvegarde Données Cloud

BONUS CLASSIFICATION

LogisticRegression

DecisionTreeClassifier

RandomForest

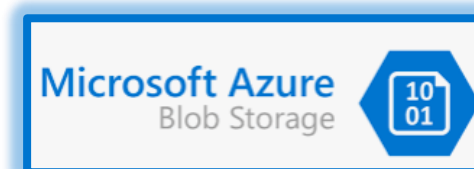
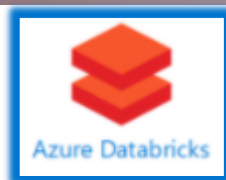
DANS LE CADRE DU PROJET P8

HORS PROJET 24

4 Chaîne de traitement – Ecriture Blob parquet



Fruits!



Microsoft Azure | Databricks

P8_01_02_notebook_CLOUD_MICRO... Free trial ends in 3 days. Upgrade to Premium in Azure Portal DP100-TEST-AZURE...

P8 - Déployer un modèle...

P8_01_02 - CLOUD - Ut...

1. Introduction

2. Préparation des données

2.1. Jeu de données

2.2.1. Connexion

2.2.2. Chargement

2.2.2. Labellisation

3. Extraction des features

3.1. Préparation des features

3.2. Préparation des features

3.3. Extraction des features

3.3.1. Fonctions

3.3.2. Extraction des features

3.4. Réduction de dimension

3.4.1. Recherche

3.4.2. Réduction de dimension

3.4.3. Sauvegarde

4. Test de classification

4.1. Préparation des données

4.2. Modélisation

4.3. Modélisation

4.4. Modélisation

```
df_reduit: pyspark.sql.dataframe.DataFrame = [path: string, Classe: string ... 4 more fields]
```

Command took 0.69 seconds -- by loe.rabier@gmail.com at 09/09/2021, 21:49:09 on P8-CLOUD-DBR 7.3 LTS ML

Cmd 63

```
# Visualisation du dataframe réduit
df_reduit.show()
```

Spark Jobs

features_scaled	path	Classe	features	features_vectors
features_scaled	vectors_pca			
[dbfs:/mnt/p8-clou...]	Cucumber-Ripe	[3.3073008, 1.298...]	[3.30730080604553...]	[1.73852559282914...]
[dbfs:/mnt/p8-clou...]	Apple-Red-Delicious	[0.6085658, 0.392...]	[0.60856580734252...]	[-1.1990166623205...]
[dbfs:/mnt/p8-clou...]	Cucumber-Ripe	[2.3520522, 1.510...]	[2.35205221176147...]	[0.69874839676122...]
[dbfs:/mnt/p8-clou...]	Cucumber-Ripe	[0.74117935, 0.83...]	[0.74117934703826...]	[-1.0546683374022...]
[dbfs:/mnt/p8-clou...]	Apple-Red-Delicious	[0.64716136, 0.57...]	[0.64716136455535...]	[-1.1570058386464...]
[dbfs:/mnt/p8-clou...]	Apple-Red-Delicious	[0.14471096, 0.65...]	[0.14471095800399...]	[-1.7039173735215...]
[dbfs:/mnt/p8-clou...]	Apple-Red-Delicious	[1.05698, 0.64239...]	[1.05698001384735...]	[-0.7109229146756...]
[dbfs:/mnt/p8-clou...]	Apple-Red-Delicious	[0.97037876, 0.87...]	[0.97037875652313...]	[-0.7109229146756...]

Command took 9.71 minutes -- by loe.rabier@gmail.com at 09/09/2021, 21:49:15 on P8-CLOUD-DBR 7.3 LTS ML

vectors_pca

=

features extraction des images réduites

Microsoft Azure

Rechercher dans les ressources, services et documents (G+ /)

Accueil > Comptes de stockage > p8cloud >

p8-cloud

Conteneur

Rechercher (Ctrl+/)

Charger Ajouter un répertoire Actualiser Renommer Supprimer Modifier le niveau

Vue d'ensemble

Diagnostiquer et résoudre les problèmes

Contrôle d'accès (IAM)

Paramètres

Jetons d'accès partagé

Gérer l'ACL

Stratégie d'accès

Propriétés

Métadonnées

Nom	Modifié	Niveau d'accès	Type d'objet blob	Taille
[.]				
[_committed_6530750328417404008]	09/09/2021, 22:13:48	Réduit (réduit)	Objet blob de blocs	1.59 Mo
[_started_6530750328417404008]	09/09/2021, 22:13:48	Réduit (réduit)	Objet blob de blocs	0 B
[_SUCCESS]	09/09/2021, 22:13:48	Réduit (réduit)	Objet blob de blocs	0 B
part-00000-tid-6530750328417404008-4537d998-cfc7...	09/09/2021, 22:13:41	Réduit (réduit)	Objet blob de blocs	13.59 Mo
part-00001-tid-6530750328417404008-4537d998-cfc7...	09/09/2021, 22:13:35	Réduit (réduit)	Objet blob de blocs	9.07 Mo
part-00002-tid-6530750328417404008-4537d998-cfc7...	09/09/2021, 22:11:32	Réduit (réduit)	Objet blob de blocs	9.08 Mo
part-00003-tid-6530750328417404008-4537d998-cfc7...	09/09/2021, 22:12:37	Réduit (réduit)	Objet blob de blocs	9.03 Mo
part-00004-tid-6530750328417404008-4537d998-cfc7...	09/09/2021, 22:12:29	Réduit (réduit)	Objet blob de blocs	9.03 Mo
part-00005-tid-6530750328417404008-4537d998-cfc7...	09/09/2021, 22:12:38	Réduit (réduit)	Objet blob de blocs	9.04 Mo
part-00006-tid-6530750328417404008-4537d998-cfc7...	09/09/2021, 22:13:43	Réduit (réduit)	Objet blob de blocs	9.03 Mo
part-00007-tid-6530750328417404008-4537d998-cfc7...	09/09/2021, 22:13:41	Réduit (réduit)	Objet blob de blocs	9.03 Mo
part-00008-tid-6530750328417404008-4537d998-cfc7...	09/09/2021, 22:13:37	Réduit (réduit)	Objet blob de blocs	9.03 Mo
part-00009-tid-6530750328417404008-4537d998-cfc7...	09/09/2021, 22:13:45	Réduit (réduit)	Objet blob de blocs	9.03 Mo
part-00010-tid-6530750328417404008-4537d998-cfc7...	09/09/2021, 22:11:40	Réduit (réduit)	Objet blob de blocs	9.04 Mo
part-00011-tid-6530750328417404008-4537d998-cfc7...	09/09/2021, 22:11:30	Réduit (réduit)	Objet blob de blocs	9.04 Mo
part-00012-tid-6530750328417404008-4537d998-cfc7...	09/09/2021, 22:10:35	Réduit (réduit)	Objet blob de blocs	13.59 Mo
part-00013-tid-6530750328417404008-4537d998-cfc7...	09/09/2021, 22:10:35	Réduit (réduit)	Objet blob de blocs	13.59 Mo
part-00014-tid-6530750328417404008-4537d998-cfc7...	09/09/2021, 22:10:35	Réduit (réduit)	Objet blob de blocs	13.58 Mo
part-00015-tid-6530750328417404008-4537d998-cfc7...	09/09/2021, 22:10:36	Réduit (réduit)	Objet blob de blocs	13.59 Mo

enregistrées au format parquet (réduit) dans le conteneur Blob Azure dans le cloud

- 1 Problématique
- 2 Données
- 3 Big Data
- 4 Chaîne de traitement
- 5 Conclusions**

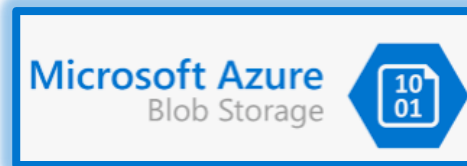
Architecture retenue



Très simple à paramétrer, calculs distribués puissants, pay as you go.

A revoir :

Choix du cluster (GPU, nombre de workers, passage à l'échelle automatique (une case à cocher)).



Peu coûteux, très simple à créer et à connecter (mount, SAS Token) avec databricks.

A revoir :

Choix du tiers et choix de la région, réplication (pour le coût de stockage)

Montée en compétence

Langage Pyspark.

Framework Apache Spark (dataframe Spark, RDD, job spark, cycle de vie, format parquet, token...).

Début certification Microsoft Azure Examen DP-100.

Améliorations

Utilisation du cache, GPU, scripts en scala, debug plus poussé (erreurs mal catchées).

Travail sur les coûts, alertes.

Hyperparamétrage des modèles de classification sur toutes les images.

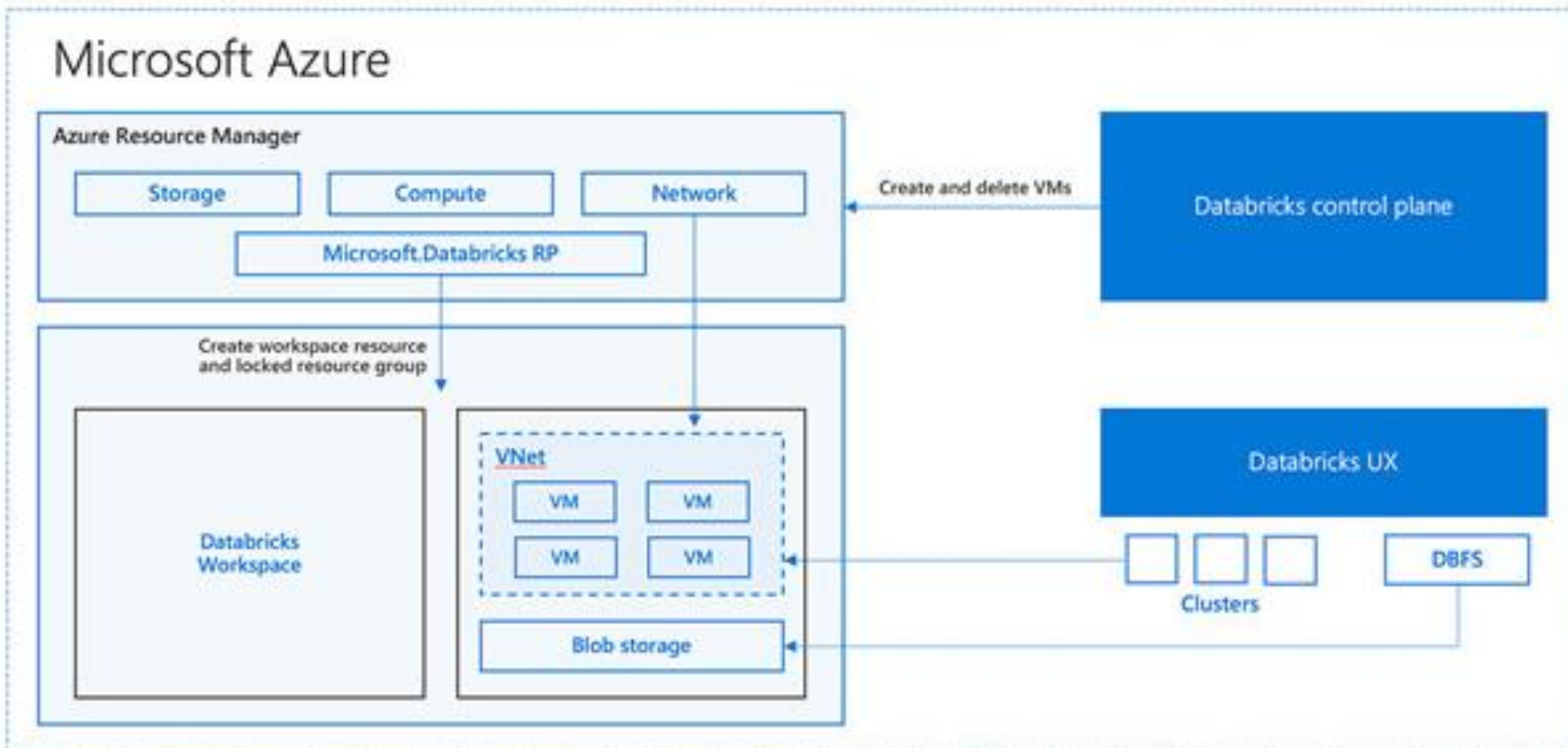
Prendre en compte la maturité des fruits sur les photos, multi-fruits pour entraîner les modèles.



Annexes



Annexe – Architecture Azure Databricks









(Architecture Big Data d'Azure Databricks - Source Azure)

Microsoft Azure

Mettre à niveau

Rechercher dans les ressources, services et documents (G+)



loe.rabier@gmail.com
ENCOURS

Accueil > Abonnements >

Abonnements

Encours

+ Ajouter

Affichez la liste des abonnements pour lesquels vous disposez d'autorisations de contrôle d'accès en fonction du rôle (RBAC) pour gérer les ressources Azure. Afin d'afficher les abonnements pour lesquels vous disposez d'un accès à la facturation, [cliquez ici](#)

Affichage des abonnements dans l'annuaire Encours. Un abonnement ne s'affiche pas ? [Changer les annuaires](#)

Mon rôle ⓘ
8 sélectionné

État ⓘ
3 sélectionné

Appliquer

Affichage 1 sur 1 abonnements ☒ filtre des
Afficher uniquement les abonnements sélectionnés dans [abonnements généraux](#) ⓘ

Rechercher

Nom de l'abonnement ↑↓
Azure subscription 1 ...

Azure subscription 1

Abonnement

Rechercher (Ctrl+J)


Annuler l'abonnement Renommer → Modifier le répertoire Commentaires

Votre crédit gratuit restant de 176,97 \$US expire dans 18 jours. Mettez à niveau pour continuer à utiliser votre compte. →

Bases

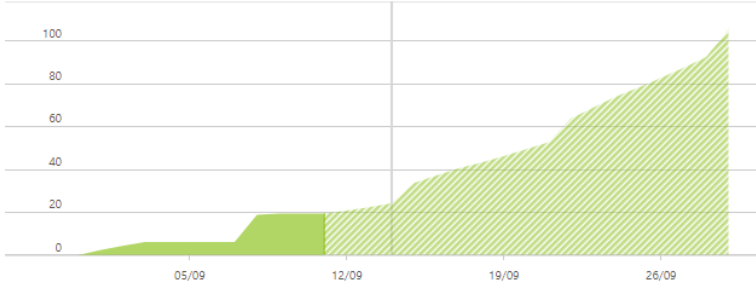
ID d'abonnement	: 192d0add-d1ed-4297-b47a-876fa1dbdfc0	Nom de l'abonnement	: Azure subscription 1
Répertoire	: Encours (Encours524.onmicrosoft.com)	Mon rôle	: Propriétaire
État	: Actif	Plan	: Plan Azure
Groupe d'administration ...	: ---	Niveau de sécurité	: Indisponible

Coûts par ressource ⓘ Afficher les détails >



p8cloud	8,53 €
6b29349a496045d1a8359b692903fd16	2,76 €
fb037ccc2b2e4b11a9f61a0ad2b873b4	1,26 €
Autres	6,92 €

Taux et prévision des dépenses ⓘ Afficher les détails >



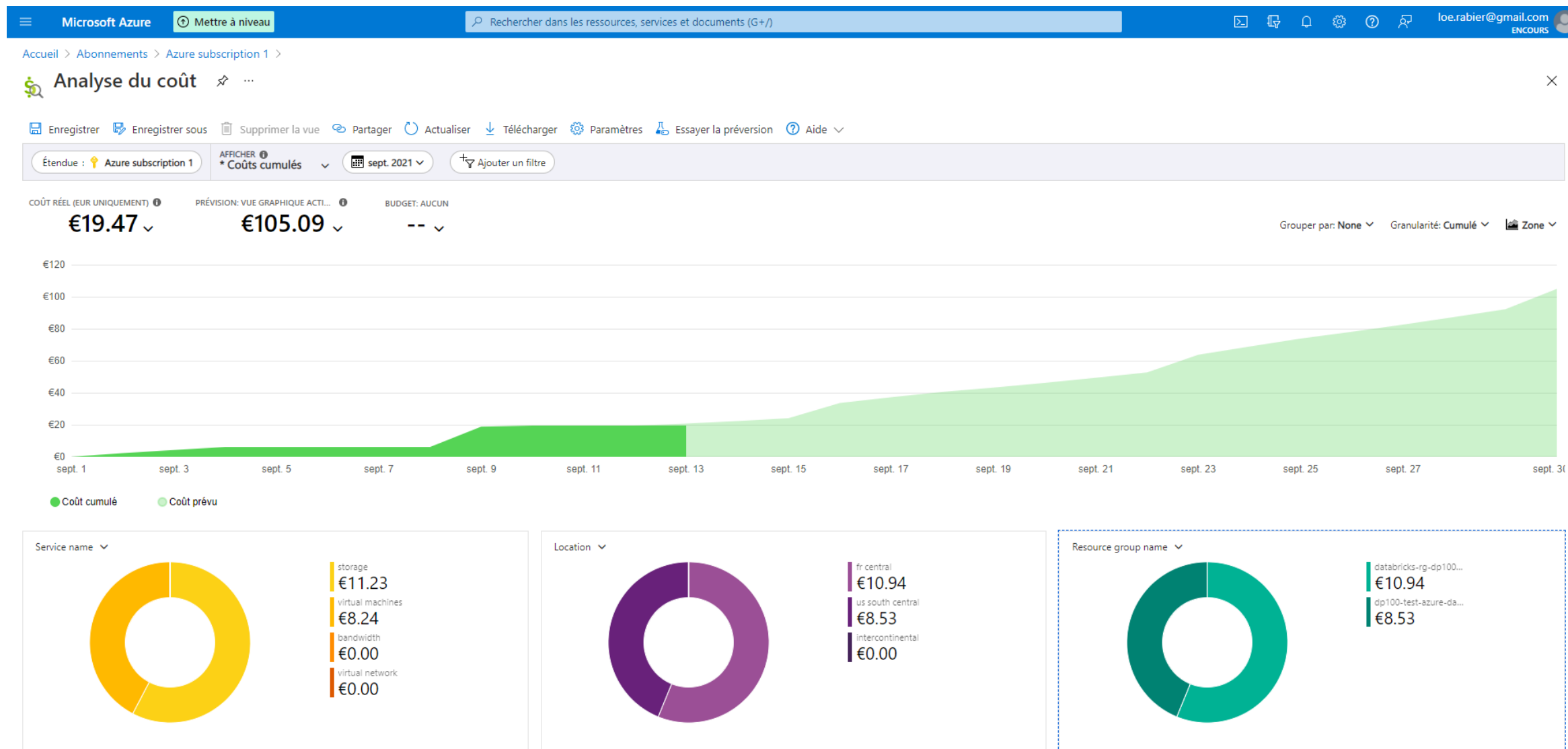
Coût actuel	19,47 €	Prévision	105,09 €
-------------	---------	-----------	----------

Services gratuits pendant 12 mois ⓘ

Période d'utilisation : 01/09/2021 - 30/09/2021

Compteur	↑↓ Utilisation/Limite	↑↓ État
Storage, Tiered Block Blob, Hot Read Operations	51% 1,02 / 2 10K	Dépassement peu probable

30





Microsoft Azure

Mettre à niveau

Rechercher dans les ressources, services et documents (G+/)

loerabier@gmail.com

ENCOURS

Accueil > Groupes de ressources >

Groupes de ressou...

Encours

+ Créer

Gérer la vue

Filtrer un champ...

Nom ↑↓

databricks-rg-DP100-TEST-AZURE-DAT...

DP100-TEST-AZURE-DATABRICKS

NetworkWatcherRG

DP100-TEST-AZURE-DATABRICKS

Groupe de ressources

Rechercher (Ctrl+/)

+ Créer

Modifier les colonnes

Supprimer le groupe de ressources

Actualiser

Exporter au format CSV

Ouvrir une requête

Attribuer des étiquettes

Déplacer

Vue d'ensemble

Journal d'activité

Contrôle d'accès (IAM)

Étiquettes

Visualiseur de ressources

Événements

Paramètres

Déploiements

Sécurité

Stratégies

Propriétés

Verrous

Gestion des coûts

Analyse du coût

Alertes de coût (préversion)

Budgets

Recommandations du conseiller

Supervision

Insights (préversion)

Alertes

Métriques

Paramètres de diagnostic

Journaux

^ Bases

Abonnement (modifier) : Azure subscription 1

ID d'abonnement : 192d0add-d1ed-4297-b47a-876fa1dbdfc0

Étiquettes (modifier) : Cliquez ici pour ajouter des étiquettes

Déploiements : 2 Réussite

Emplacement : France-Centre

Ressources

Recommandations

Filtrer un champ...

Type == tout

Emplacement == tout

Ajouter un filtre

Affichage de 1 à 2 sur 2 enregistrements.

Afficher les types masqués

Aucun regroupement


Vue liste

<input type="checkbox"/> Nom ↑↓	Type ↑↓	Emplacement ↑↓	
<input type="checkbox"/> DP100-TEST-AZURE-DATABRICKS	Service Azure Databricks	France-Centre	...
<input type="checkbox"/> p8cloud	Compte de stockage	USA Centre Sud	...







< Précédent

Page 1 sur 1

Suivant >

 **Microsoft Azure** Mettre à niveau

Rechercher dans les ressources, services et documents (G+)

      loe.rabier@gmail.com
ENCOURS


Accueil >

Toutes les ressources

✧ ...

Encours

+ Créer

 Gérer la vue

Actualiser

Exporter au format CSV

Ouvrir une requête

Attribuer des étiquettes

Supprimer

Commentaires

Filtrer un champ...

Abonnement == tout

Groupe de ressources == tout

Type == tout







Emplacement == tout

Ajouter un filtre

Affichage de 1 à 6 sur 6 enregistrements. ☐ Afficher les types masqués

Aucun regroupement

Vue liste

<input type="checkbox"/> Nom ↑↓	Type ↑↓	Groupe de ressources ↑↓	Emplacement ↑↓	Abonnement ↑↓	
<input type="checkbox"/>  dbstorage5dfmojqy6ijye	Compte de stockage	databricks-rg-DP100-TEST-AZURE-DATABRICKS-kz...	France-Centre	Azure subscription 1	...
<input type="checkbox"/>  DP100-TEST-AZURE-DATABRICKS	Service Azure Databricks	DP100-TEST-AZURE-DATABRICKS	France-Centre	Azure subscription 1	...
<input type="checkbox"/>  NetworkWatcher_francecentral	Observateur réseau	NetworkWatcherRG	France-Centre	Azure subscription 1	...
<input type="checkbox"/>  p8cloud	Compte de stockage	DP100-TEST-AZURE-DATABRICKS	USA Centre Sud	Azure subscription 1	...
<input type="checkbox"/>  workers-sg	Groupe de sécurité réseau	databricks-rg-DP100-TEST-AZURE-DATABRICKS-kz...	France-Centre	Azure subscription 1	...
<input type="checkbox"/>  workers-vnet	Réseau virtuel	databricks-rg-DP100-TEST-AZURE-DATABRICKS-kz...	France-Centre	Azure subscription 1	...



1. Création container blob Data Lake Storage

Depuis le portail Azure :

<https://p8cloud.blob.core.windows.net/p8-cloud>

2. Installation outils de chargement des images en masse

car depuis Azure sur le container le chargement est effectué image par image, mais environ 68000 images!

<https://docs.microsoft.com/en-us/azure/storage/common/storage-use-azcopy-blobs-uploadazure>

2.1. download win64

2.2. ouvrir cmd.exe

2.3. se placer dans le répertoire

`cd C:\4-OUTILS\azcopy`

2.4. lancer le programme : azcopy

2.5. créer une clé SAS token sur le container P8-cloud, menu signature d'accès partagé, sélectionner blob container objet et générer la clé.

2.6. Récupérer la clé SAS token pour blob

`https://p8cloud.blob.core.windows.net/?sv=2020-08-`

`04&ss=bfqt&srt=sco&sp=rwdlacupx&se=2021-09-09T15:46:46Z&st=2021-09-`

`09T07:46:46Z&spr=https&sig=QH%2Fi7GR9vUqoV0loWIW2gQGkaRUcLzdJumCVgoOfkEo%3D`

2.7. Lancer la commande dans le nom du container suivi des bons répertoires:

`azcopy copy "C:\2.DATA_SCIENCE\PARCOURS_DATA_SCIENTIST\PROJET_8-`

`DEPLOYEZ_MODELE_DANS_CLOUD\resources\data"`

`"https://p8cloud.blob.core.windows.net/p8-cloud/resources?sv=2020-08-`

`04&ss=bfqt&srt=sco&sp=rwdlacupx&se=2021-09-09T15:46:46Z&st=2021-09-`

`09T07:46:46Z&spr=https&sig=QH%2Fi7GR9vUqoV0loWIW2gQGkaRUcLzdJumCVgoOfkEo%3D" --`

`recursive`

3. Création ressource Microsoft Azure Databricks

Depuis le portail Azure :

3.1. + create , rechercher databricks

3.2. Importer le notebook local au format .dbc archive.

3.3. Créer un cluster pour les calculs ==> compute databrick runtime

3.4. Attacher le cluster au notebook et démarrer le cluster

4. Connection data lake storage (stockage) à azure databricks (compute - calcul distribué)

Depuis le portail Azure :

<https://docs.databricks.com/data/data-sources/azure/azure-storage.html>

```
dbutils.fs.mount(source = "wasbs://<container-name>@<storage-account-name>.blob.core.windows.net", mount_point = "/mnt/<mount-name>", extra_configs = {"<conf-key>":"<key-name>"})
```

- avec : <storage-account-name> = le nom du container account qui contient tous les blobs : depuis le portail azure, ressources = compte de stockage et c'est son nom p8cloud.
- avec : <container-name> = le nom du container account qui contient tous les blobs : depuis le portail azure, cliquer sur le compte de stockage p8cloud et depuis menu Stockage des données, accéder au 'Conteneurs' et c'est le nom de notre conteneur : p8-cloud.
- avec : <mount-name> = le nom à choisir librement où on veut stocker les données.
- avec : <conf-key> = fs.azure.account.key.<storage-account-name>.blob.core.windows.net
- avec : <key-name> = Clé d'accès du conteneur de stockage depuis le portail azure, cliquer sur le compte de stockage p8cloud puis dans le menu Sécurité + réseau, cliquer sur Clés d'accès puis Afficher les clés et copier la première clé dans key1.

cf vidéo youtube bien faite : <https://www.youtube.com/watch?v=zwMksSEjNvU>

pour accéder au ressources : `df.spark.read("/mnt/<mount-name>/.....jpg...csv...")`

5. Copie des fichiers du blob sur le c: pour soutenance

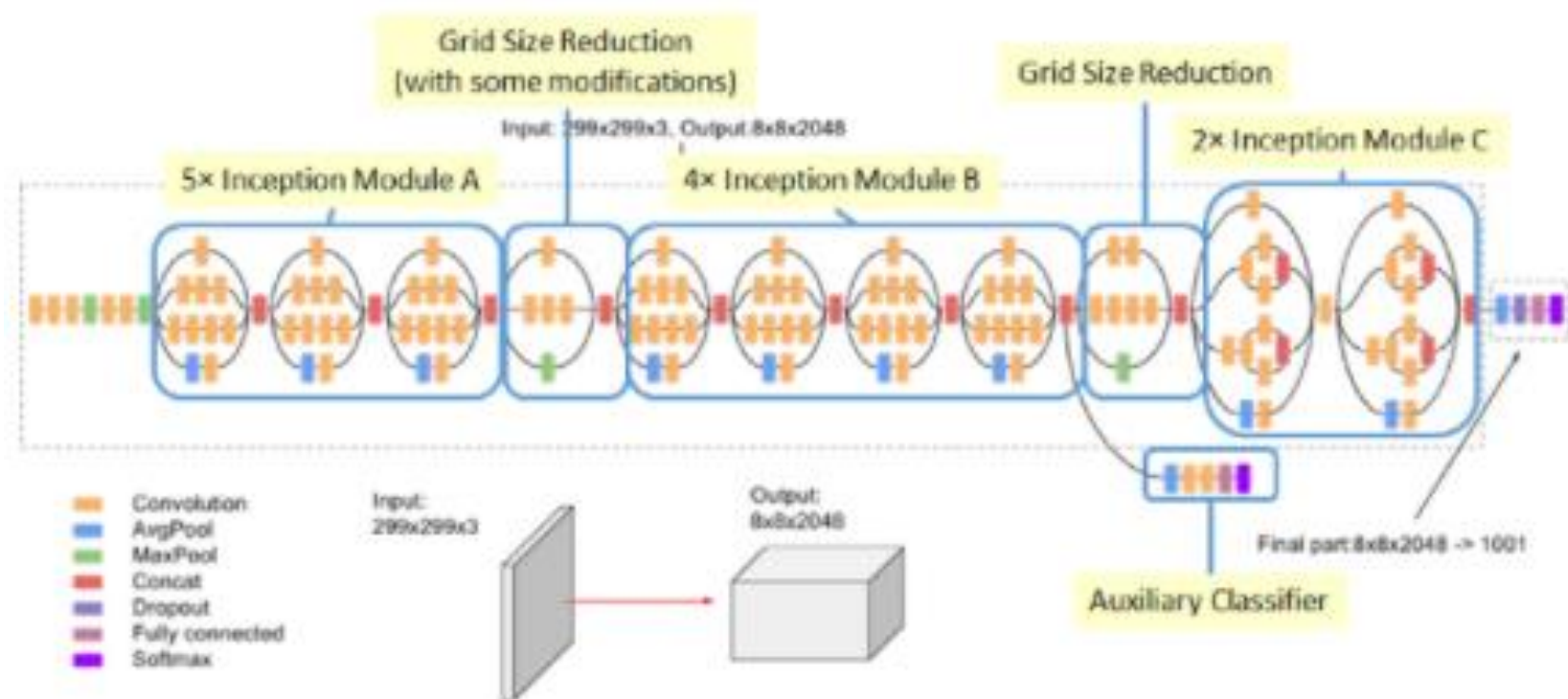
Depuis le portail Azure : depuis compte de stockage, container, cliquer sur le conteneur blob p8-cloud, se placer dans le répertoire où sont stockés les fichiers parquets et clic droit télécharger

ou utiliser l'outil azcopy :

```
azcopy copy "https://p8cloud.blob.core.windows.net/p8-cloud/resources/output/resultats_features_parquet?sv=2020-08-04&ss=bfqt&srt=sco&sp=rwdlacupx&se=2021-09-09T15:46:46Z&st=2021-09-09T07:46:46Z&spr=https&sig=QH%2Fi7GR9vUqoV0loWIW2gQGkaRUcLzdJumCVgoOfkEo%3D" "C:\2.DATA_SCIENCE\PARCOURS_DATA_SCIENTIST\PROJET_8-DEPLOYEZ_MODELE_DANS_CLOUD\resources\output" --recursive
```

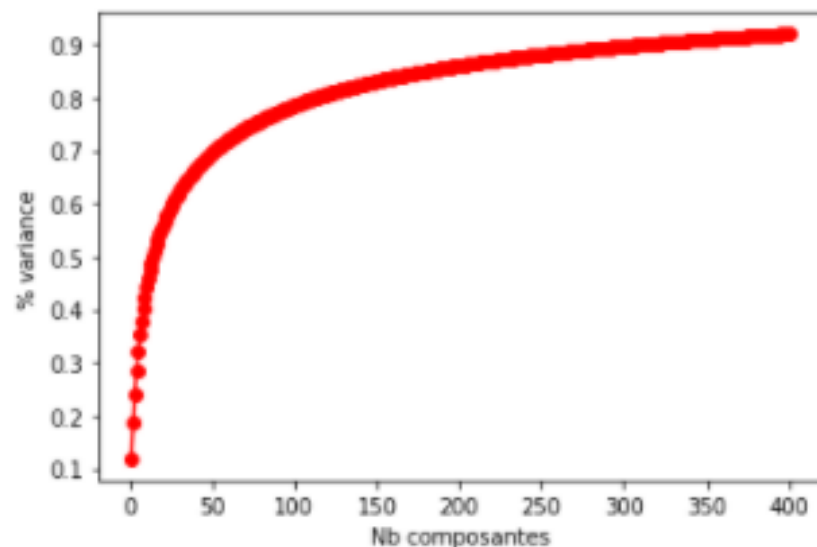
Utilisation d'un modèle CNN de transfert learning : InceptionV3

- Redimensionnement des images en 299x299 pixels.
- Instanciation du modèle entraîné avec les images de la librairie 'Imagenet' en supprimant la dernière couche (qui effectue la classification) de l'application Keras de la librairie Tensorflow.
- En sortie, vecteur de taille (8, 8, 2048), donc 2048 features.

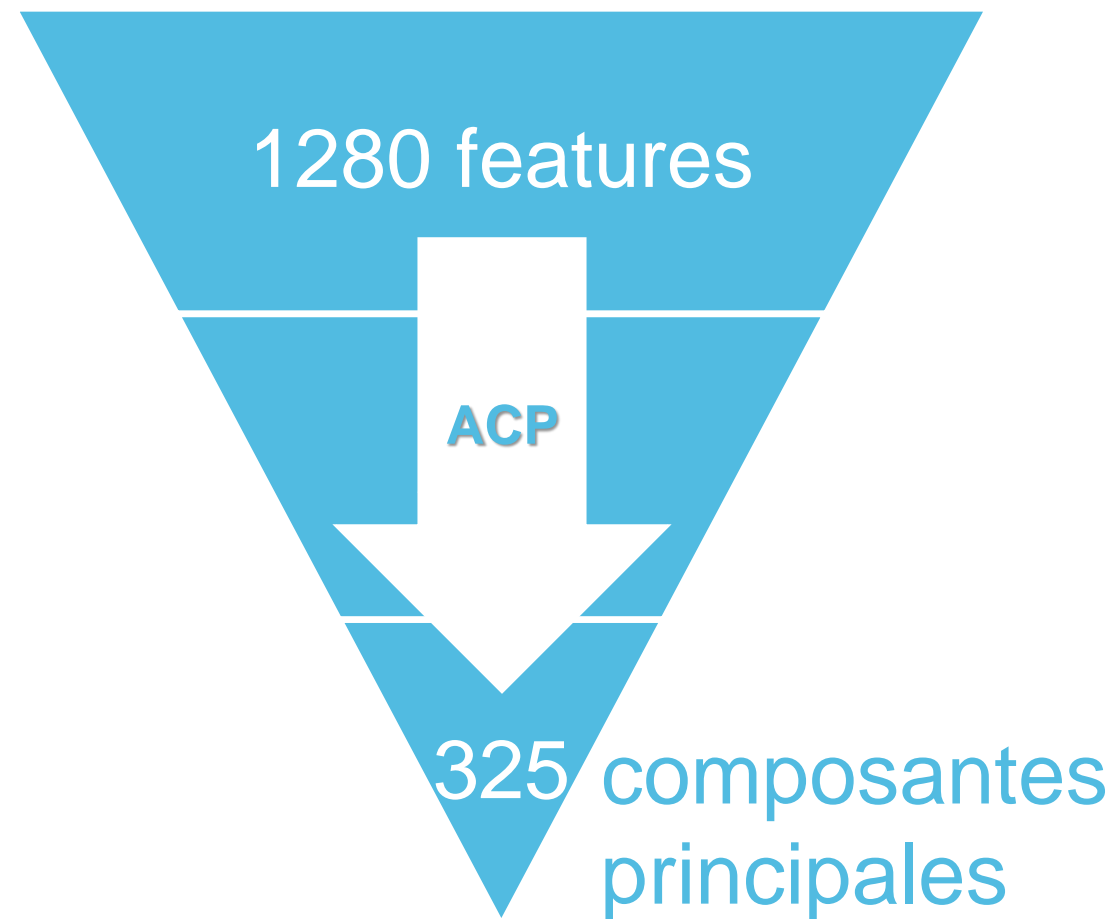




Réduction de dimension en utilisant l'analyse en composantes principales (ACP).



325 composantes expliquent 95% de la variance totale



4484 images

LogisticRegression

Accuracy = 1

DecisionTreeClassifier

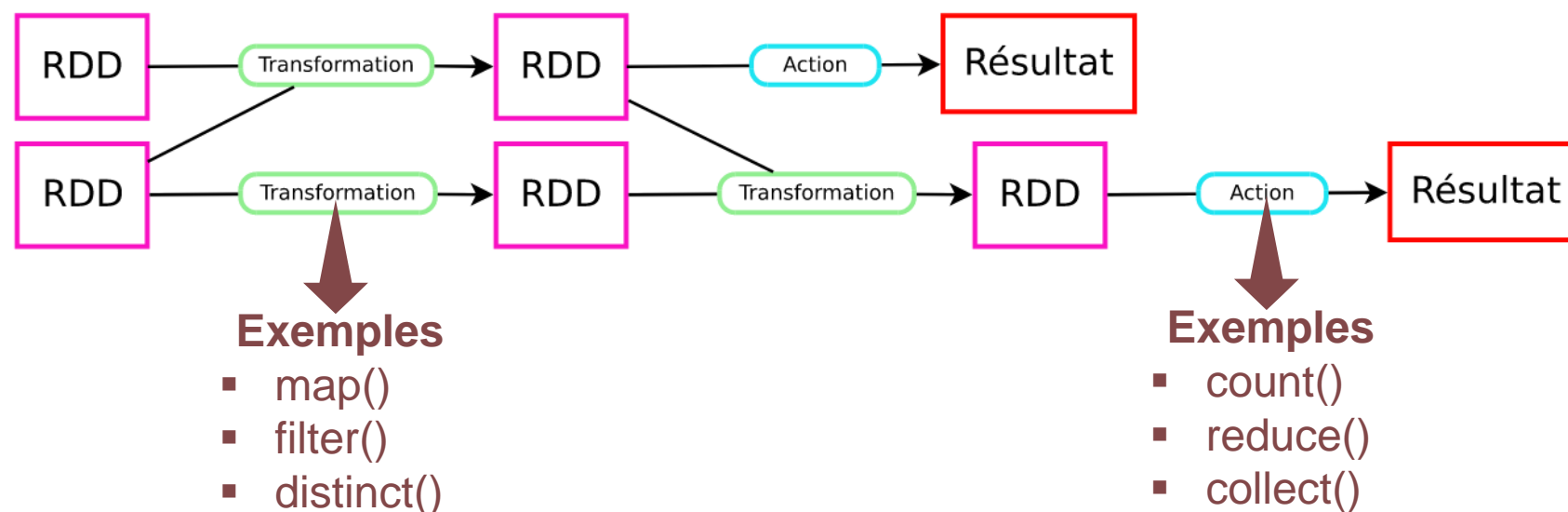
Accuracy = 0,96767

RandomForestClassifier

Accuracy = 0,989967

Les **R**esilient **D**istributed **D**ataset (RDD)

Dans une application Spark, les **transformations** et les **actions** réalisées sur les RDD permettent de construire un **graphe acyclique orienté (DAG : "directed acyclic graph")**



Lorsqu'un nœud devient indisponible, il peut être régénéré à partir de ses nœuds parents. C'est précisément ce qui permet la **tolérance aux pannes** des applications Spark.

Spark utilise une **évaluation paresseuse**, ce qui signifie qu'il ne fait aucun travail jusqu'à ce que vous demandiez un résultat.

Un **job** Spark correspond à une action sur un RDD et est composé de plusieurs **étapes** séparées par des **shuffles**.

Partitions :

découpage des données

Tâche :

traitement d'une partition

Étape :

ensemble de tâches réalisées en parallèle

Shuffle :

redistribution des données entre les nœuds

