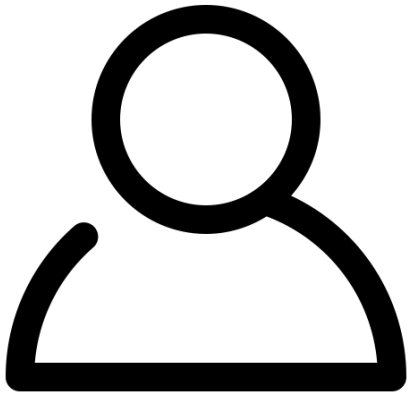
Industrialiser le Machine Learning grâce à la plateforme Azure

Paul PETON

2019

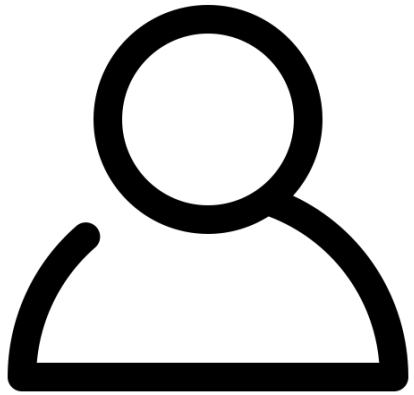
Global Azure BOOTCAMP

Merci à nos sponsors locaux



William BORDES

l'eclozr
INNOVATION MAKERS



Yoann GUILLO

Merci à nos sponsors internationaux

Skill Me UP
Powered by opsgility

JET BRAINS

SERVERLESS 360

ENZO

Cloudilonix

Microsoft

RevDeBug

kemp

Progate

Meet the Team

Lien vers l'enregistrement

Paul PETON

Lead Data Scientist

MVP Artificial Intelligence
Organisateur du meetup Club Power BI @Nantes

@paulpeton

https://www.linkedin.com/in/paul-peton-datascience

https://github.com/methodidacte/

AZEO talents & technology

L'expertise Cloud AZEO

Nous vous donnons les clés du Cloud Microsoft

Stratégie Cloud First

Expertise Azure 360°

+32 % des effectifs AZEO certifiés Azure

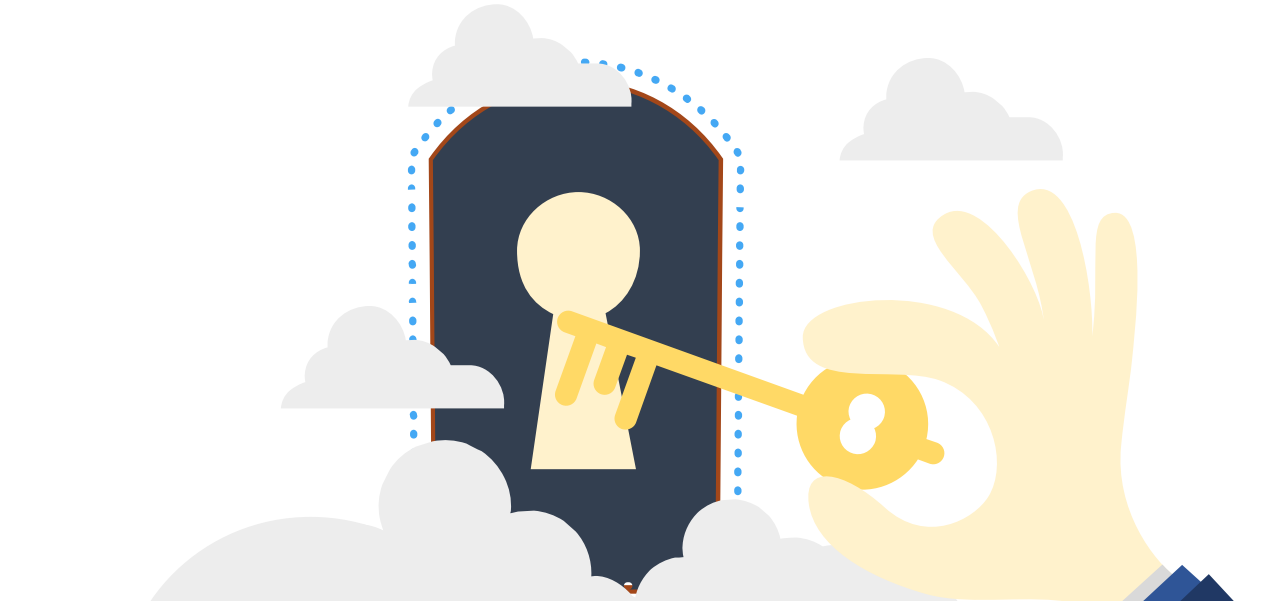
AZEO est reconnu comme l'un des influenceurs majeurs du Cloud Microsoft en France. Certifié Cloud Solution Provider par Microsoft, nous sommes suivis par l'entité services de l'éditeur comme VIP Partner Microsoft Services.

Cette reconnaissance est légitimée par nos nombreuses expériences, interventions communautaires et de par nos certifications Gold sur l'ensemble des compétences Cloud Microsoft.

AZEO, votre partenaire Cloud d'excellence :

+230 Collaborateurs
dont 190 collaborateurs salariés
50 % en Ile-De-France
50 % en réseau national

4 Implantations
Boulogne-Billancourt, Bordeaux, Nantes, Toulouse



Microsoft Partner

Microsoft

- Gold Productivité cloud
- Gold Plateforme cloud
- Gold Plate-forme de données
- Gold DevOps
- Gold Gestion de la mobilité d'entreprise

AZEO
talents & technology

Menez la **transformation numérique** grâce à nos domaines d'excellence

Nos équipes s'engagent à vos côtés autour de différents axes de transformation numérique. Appuyez vous sur les convictions technologiques de nos experts afin de mieux répondre à vos enjeux métiers.

Grâce à notre méthodologie et nos offres, vous dirigerez votre organisation vers de nouveaux succès numériques.

Infrastructure Solutions	Cloud Transformation	Digital Trust	Infrastructure Excellence
Modern Applications	DevOps & ALM	Intelligent Apps (IA & IoT)	Application Modernization
Data & AI	Modern Data Management	Self-Service BI	IA & Data-Science
Modern Workplace	Modern Desktop	Digital Worker	Digital Workplace
Business Solutions	Modern Solutions	Customer Experience	Digital Insights

AZEO
talents & technology

Agenda

Industrialiser le Machine Learning grâce à la plateforme Azure

Rapides rappels : l'Intelligence Artificielle n'existe pas
Les acteurs de notre histoire : Jayson & Cassandra
De l'impérative nécessité d'industrialiser
Cas d'usage : les vélos en location de New York
Architecture Azure : the big pict(az)ure
Conclusion : points forts et améliorations attendues

2019
Global Azure
BOOTCAMP

Paul PETON

Samedi 27 avril 2019

Rapides rappels

Levez la main si, au sein de votre organisation…

Un projet de business intelligence est en production

Un projet de data science est en production

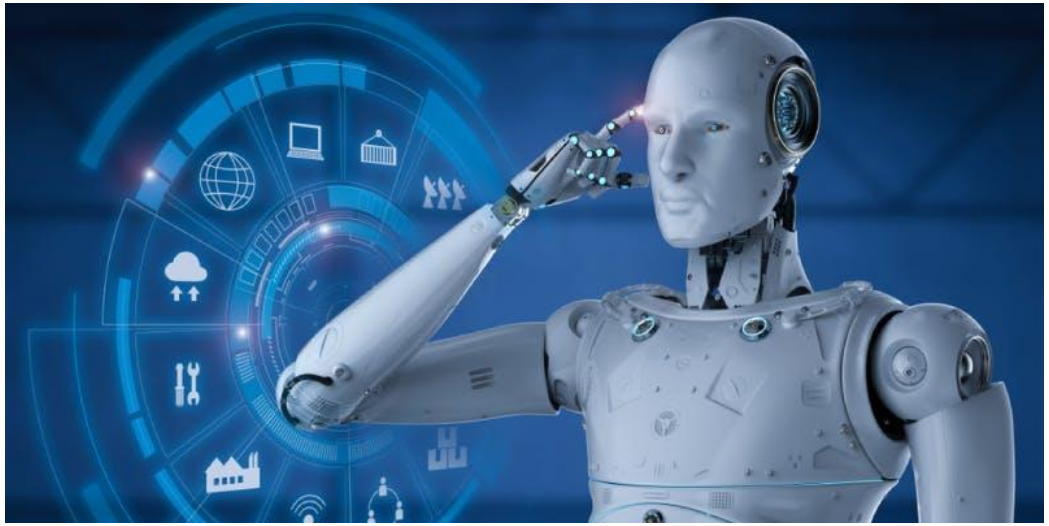
Artificial Intelligence versus Machine Learning

« If it's written in Python, it's…

Machine Learning»

« If it's written in Power Point, it's…

Artificial Intelligence»



Cette présentation parle de Machine Learning et donc de regressions linéaire, logistique, d'arbres de decision, de reseaux de neurones, etc.

Cette presentation parle surtout d'industrialisation et moins d'optimisation des modèles de Machine Learning.

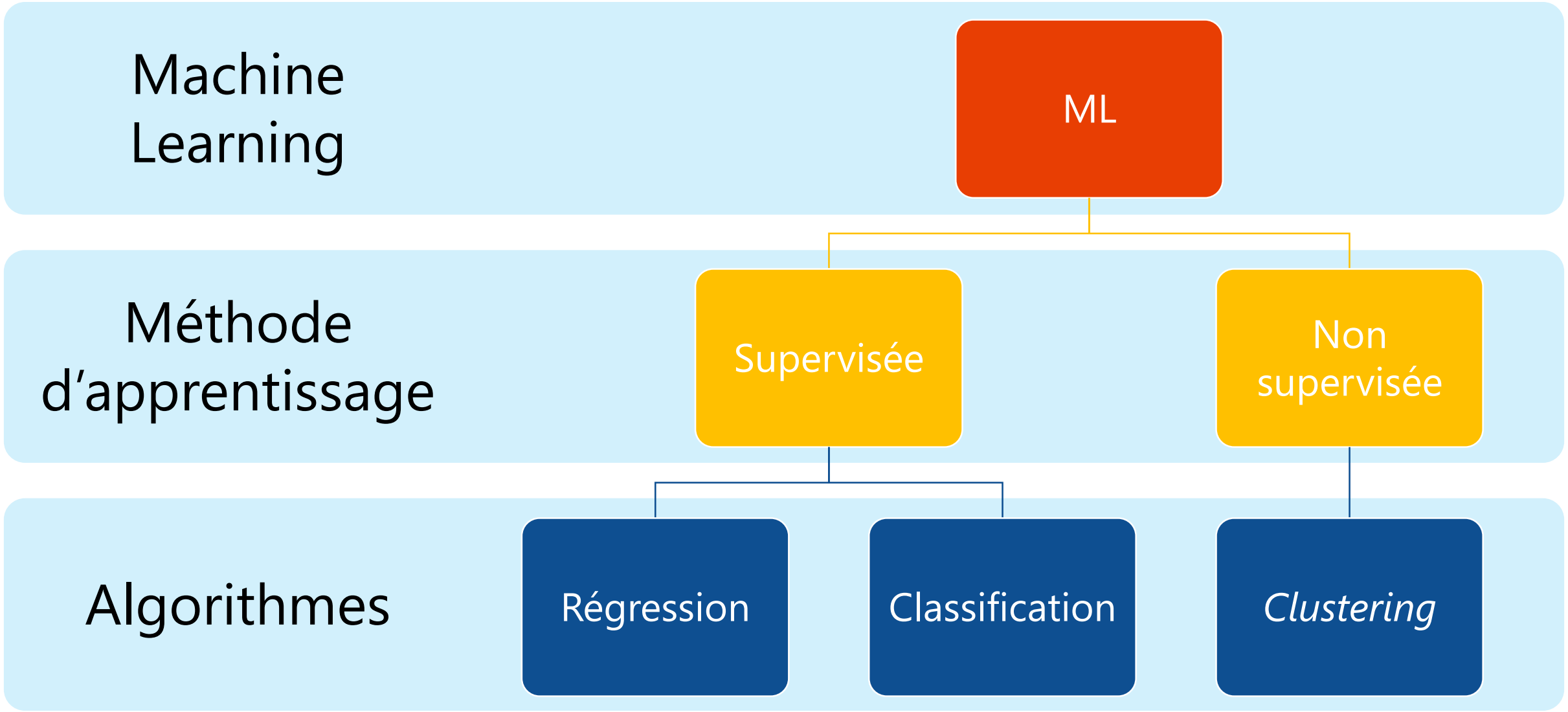
Apprentissage automatique (Machine Learning)

Définition d'Arthur Samuel (1959) :

Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.

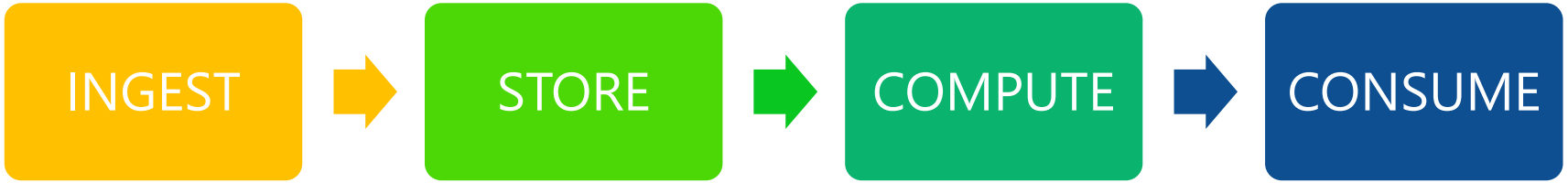
AZEO
talents & technology

Les méthodes de Machine Learning



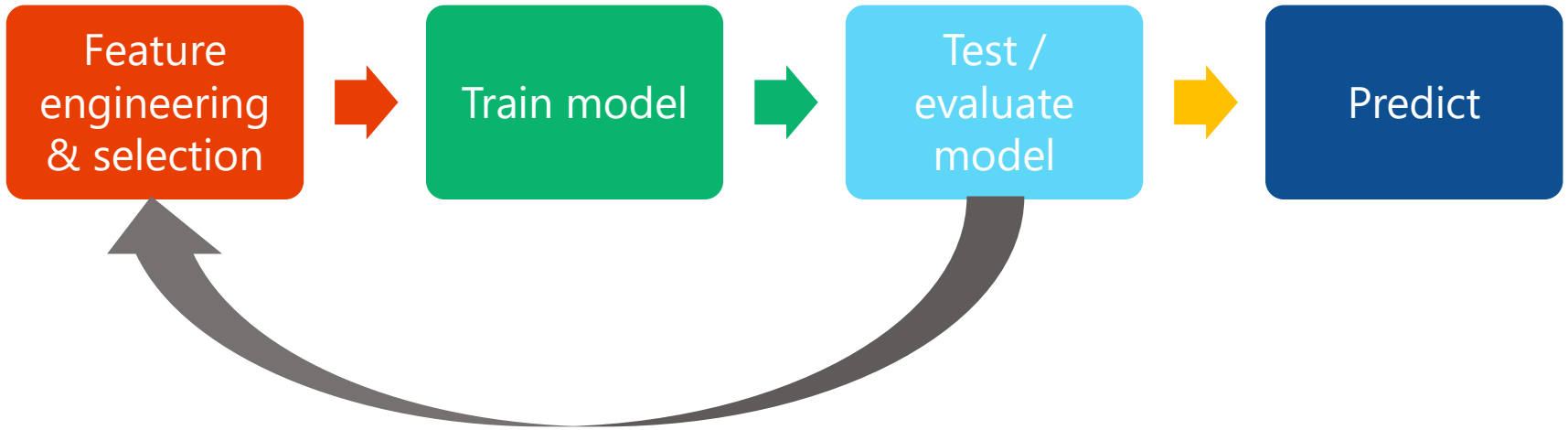
L'approche ISCC © d'un projet Data

Quatre étapes fondamentales pour un projet Data :



Focus la spécificité de la partie COMPUTE dans le cadre du Machine Learning :

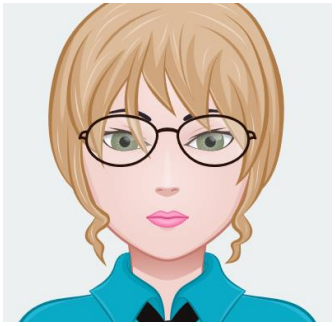
itération sur la préparation des données et l'entraînement des modèles



Nos deux protagonistes

Cette histoire est inspirée de faits et de données réelles.

Jayson est Data Scientist.



Cassandra est Data Engineer.

Tous deux travaillent sur un même projet de modélisation de la durée de trajet des vélos new-yorkais.

Jayson et Cassandra sont dans le même bateau

Les métiers de la data sont anciens par leur socle de compétences (mathématiques, algorithmie…) mais nouveaux par les technologies (NOSQL, containers…) et surtout par la dénomination des différents postes.

Chaque organisation peut avoir sa propre définition même si quelques tendances se dégagent.

Jayson : Data Scientist

A fait des études de mathématiques et de statistiques

Est proche des problématiques du métier

Adore affronter des problèmes complexes

Code en R et en Python

Exécute le code sur son laptop à partir de fichiers sources stockés en local

Cassandra : Data Engineer

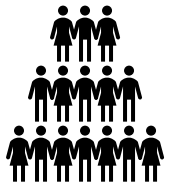
A fait des études d'informatique

A une vue d'ensemble de l'urbanisation du SI

Adore les nouvelles technos

Code en Java et en Scala

Exécute le code sur des VMs ou grâce à des services managés



Et autour d'eux :

Product Owner

Data Architect

Data Strategist

Data Analyst

Data Steewart

Chief Data Officer

Data Et Caetera

Leurs défis à relever

Jayson : Data Scientist

Sélectionner les bons facteurs explicatifs du modèle

Sélectionner le meilleur modèle

Sélectionner les meilleurs hyperparamètres du modèle

« J'explore les données et je teste différentes approches pour améliorer la précision. »

Cassandra : Data Engineer

Assurer le passage à l'échelle pour l'entrainement du modèle

Déployer le modèle et assurer son accès sécurisé

Automatiser le pipeline de Machine Learning de bout en bout

« Je m'assure de la bonne intégration du Machine Learning dans le Système d'Information. »

Pourquoi industrialiser le ML ?

Appliquer une gouvernance des données

Contrôler le *versioning*

Disposer de sauvegardes pour retour arrière

Planifier l'exécution des traitements

Monitorer les services

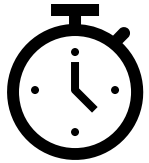
Intégrer la sécurité de l'organisation

Déployer sur des terminaux légers (*edge*)

Pistes pour une industrialisation

Un ordonnanceur pour planifier

Le nettoyage des données
L'entrainement / réentrainement du modèle
Le calcul de la prévision (en mode *batch*)

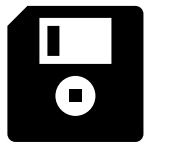


Un système de stockage pour archiver les modèles

Selon les algorithmes, selon les versions
Dans un format binaire sérialisé (et non propriétaire)

Un outil d'exposition du modèle

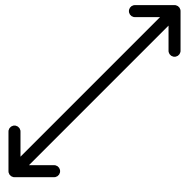
Par API REST (l*anguage-agnostic*)
D'accès sécurisé



Serving

Des ressources déployables à l'échelle

A l'aide des *containers*
Dans le *cloud*



Azure Data Science Virtual Machine

« En quelques minutes, mon terrain de jeu est prêt ! »

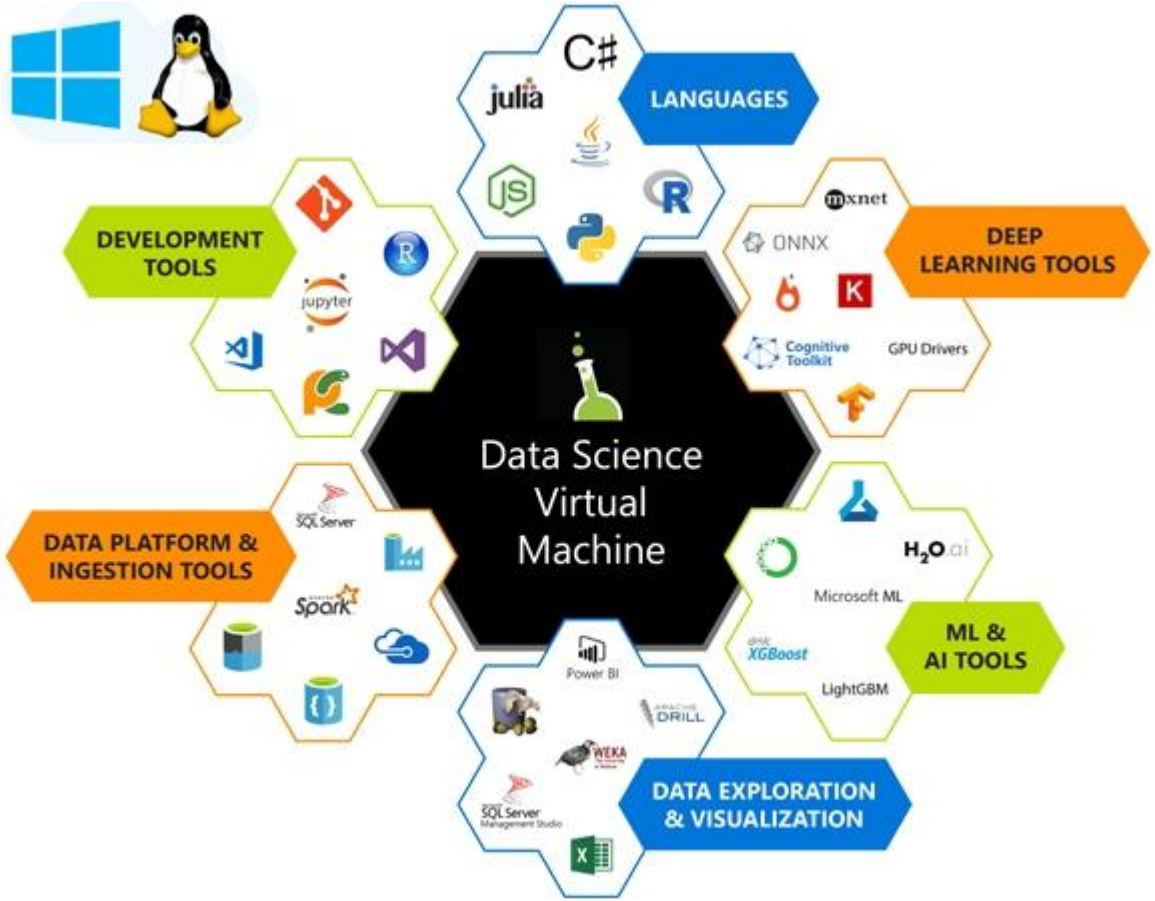
Machine virtuelle avec des outils déjà installés

- SQL Server
- R Server
- R Studio
- Anaconda
- Packages de Data Science
- Modèles pré-entrainés
- Etc.

OS Windows Server ou Linux

Version pour la production (Azure Batch)

Voir aussi la Deep Learning VM (avec GPU)



Principe des notebooks

Affichage simultané de code « live », d'images, d'équations… de manière interactive

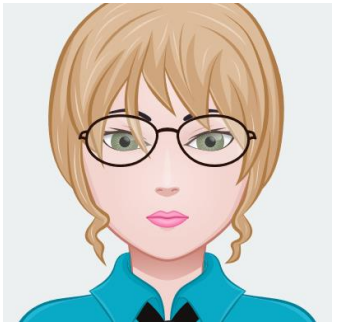
Permet le partage et le travail collaboratif

Permet d'utiliser une multitude de langages

Jupyter : Julia, Python & R

Azure Notebooks : backend paramétrable

« Super outil pour développer et tester différents scénarios »



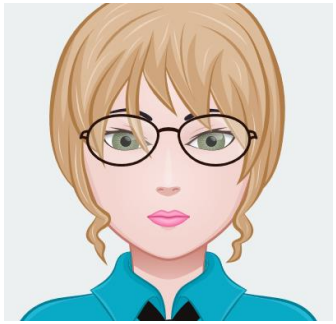
« Euh… je peux vraiment le lancer en prod directement ? »

Packages Python

Point fort du langage : sa communauté et la production d'un grand nombre de librairies Open Source

A installer par les commandes :

```
conda install NomPackage
pip install NomPackage
```



« Attention au versioning des packages ! »

Fichier d'environnement yml :

- Contient les « *dependencies* »
- = packages nécessaires
- À générer par du code

```yaml
# Conda environment specification. The dependencies defined in this file will

# be automatically provisioned for runs with userManagedDependencies=False.


# Details about the Conda environment file format:

# https://conda.io/docs/user-guide/tasks/manage-environments.html#create-env-file-manually


name: project_environment
dependencies:
  # The python interpreter version.

  # Currently Azure ML only supports 3.5.2 and later.

- python=3.6.2

- pip:
    # Required packages for AzureML execution, history, and data preparation.

  - azureml-defaults
- scikit-learn
```

Azure Python SDK

Ensemble de librairies qui facilitent l'accès à :

- Des composants de Management (VM…)
- Des composants de Runtime (ServiceBus using HTTP, Batch, Monitor)

Repo : https://github.com/Azure/azure-sdk-for-python

The full list of available packages and their latest version : https://docs.microsoft.com/fr-fr/python/api/overview/azure/?view=azure-python

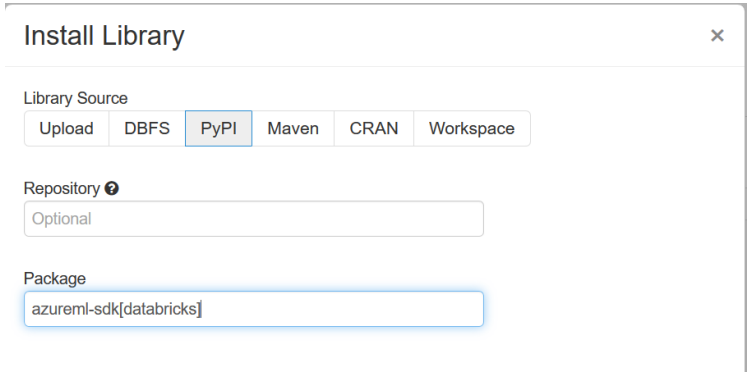
Installation par la commande :

```
$ pip install azure
```

Ou clone du reporsitory Git :

```
git clone git://github.com/Azure/azure-sdk-for-python.git
cd azure-sdk-for-python
python setup.py install
```

Ou dans l'interface utilisateur de Databricks :



New York Citibike dataset

Citi Bike is a bike sharing service available in New York City, that permits easy and affordable bike trips. They regularly release data about such trips, including starting and ending stations, starting and ending time, duration of the trip and few others variables.

This dataset is the property of NYC Bike Share, LLC and Jersey City Bike Share, LLC ("Bikeshare") operates New York City's Citi Bike bicycle sharing service for T&C

AZEO talents & technology

Cas d'usage pour le Machine Learning supervisé

Prédire la durée d'un trajet

Variable à expliquer : Trip duration (secondes)

Variable quantitative continue

Modèle candidat : régression linéaire multiple

Variables explicatives candidates :

- Trip Duration (seconds)
- Start Time and Date
- Stop Time and Date
- Start Station Name
- End Station Name
- Station ID
- Station Lat/Long
- Bike ID
- User Type (Customer = 24-hour pass or 3-day pass user; Subscriber = Annual Member)
- Gender (Zero=unknown; 1=male; 2=female)
- Year of Birth

Architecture(s) Azure

NAME	TYPE
adlsgen2sapp	Storage account
citibikedurationdeploy	Container instances
eacbadfppn	Data factory (V2)
eacbdatabircks	Azure Databricks Service
eacbmlservicew2479569759	Storage account
eacbmlservicew6648238519	Application Insights
eacbmlservicew6669147962	Key vault
eacbmlservicew8493899780	Container registry
eacbmlservicews	Machine Learning service workspace

AZEO talents & technology

Framework Apache Spark

Spark est un framework de calcul distribué

Matei Zaharia crée le projet en 2009 au cours de son doctorat au sein de l'université de Californie à Berkeley.

Première sortie sous licence Apache en 2014

Dernière version : 2.4.1 (mars 2019)

Ecrit en Scala

« Je veux dépasser les performances de MapReduce en travaillant *in memory* »

Spark dispose de deux API pour l'apprentissage statistique:

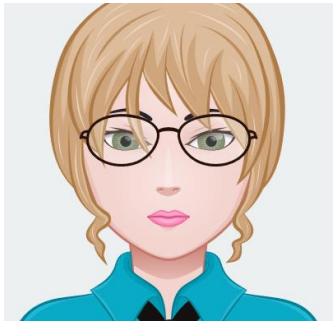
- **Mllib** qui prend en input des Resilient Distributed Datasets (RDD)
- **Spark ML** qui prend en input des Data Frames (recommandé pour l'analyse de données structurées)

Langage Scala

Langage de programmation multi-paradigme
- Orienté Objet
- Programmation fonctionnelle

SCAlable Language : langage qui peut être mis à l'échelle

Compilé en bytecode Java et utilisable par une JVM



« Enfin un langage performant et qui distribue vraiment les calculs ! »

« Euh… je ne retrouve pas l'équivalent de toutes mes librairies préférées. »

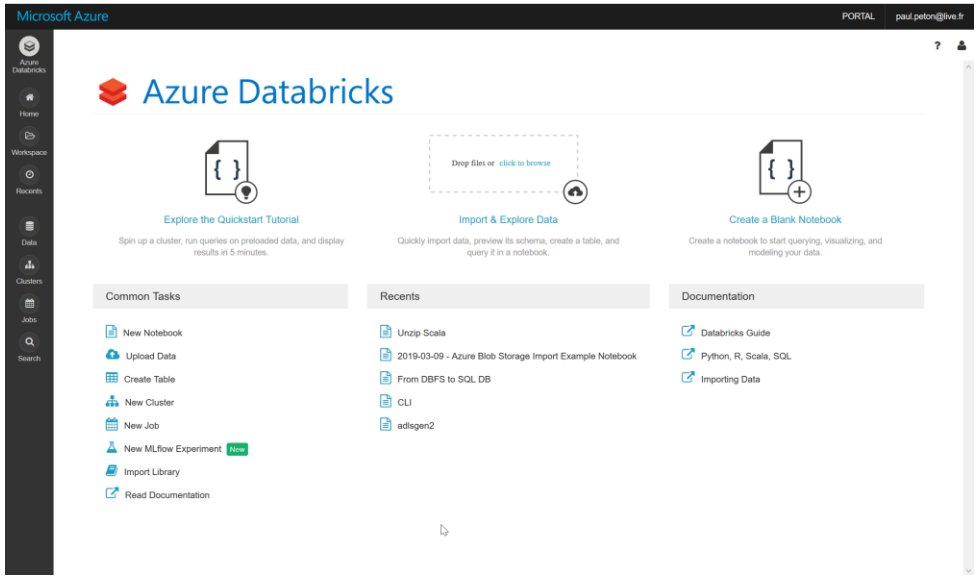
Azure Databricks

databricks

Databricks est une société fondée en 2013 par les créateurs de Apache Spark

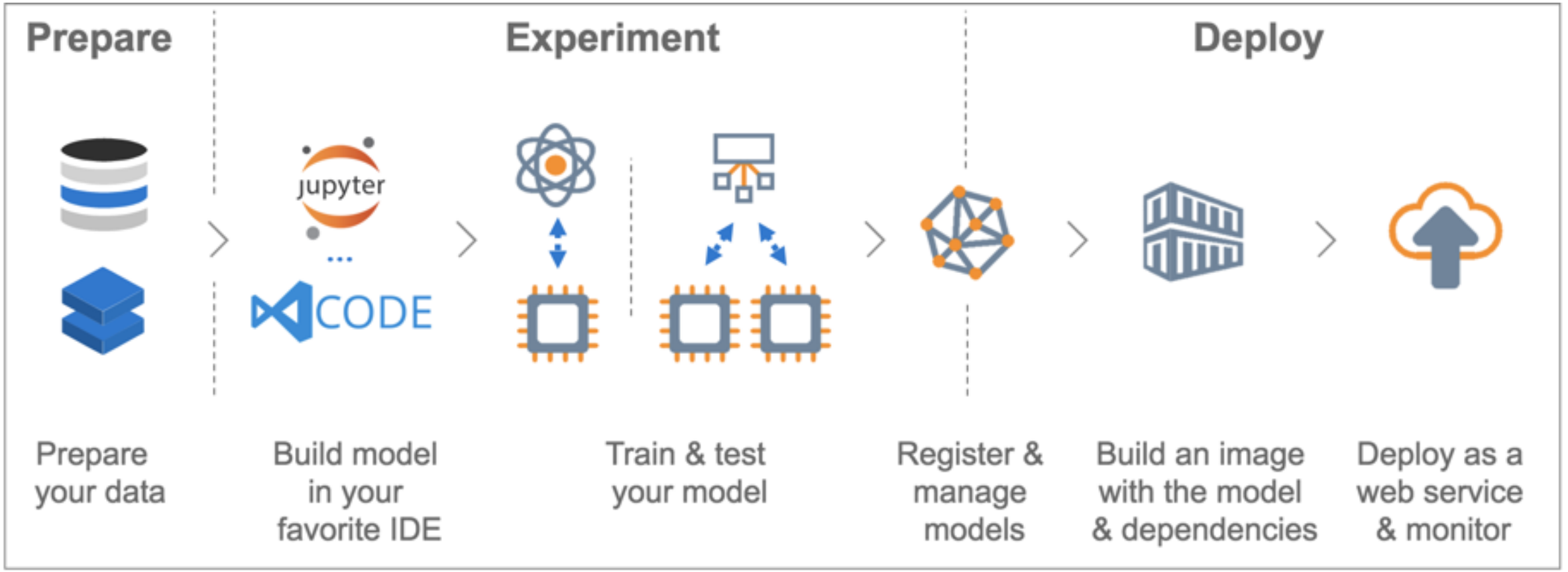
Plateforme web permettant d'exécuter des jobs Spark:

- Auto gestion du cluster dans le cloud Azure
  - Auto scalling entre un nombre de noeuds minimum et maximum
  - Arrêt du cluster après un temps d'inactivité
- Permet de faire de la datavisualisation basique sur toute la donnée
- Met à disposition des notebooks :
  - Scala
  - Python
  - R
  - Et pouvant intégrer des commandes Shell ou des requêtes SQL
- Interagit avec d'autres services Azure
  - Azure Data Lake Store (gen1 & gen2)
  - Cosmos DB
  - Azure SQL Datawarehouse
- Permet de travailler en mode collaboratif (tarification premium sous Azure)
- Existe dans une version Community gratuite sans nécessité de compte Azure
  - https://databricks.com/signup#signup/community



Azure Machine Learning Service

Service Azure permettant de gérer le workflow complet d'un projet de Machine Learning



INGEST → STORE → COMPUTE → CONSUME

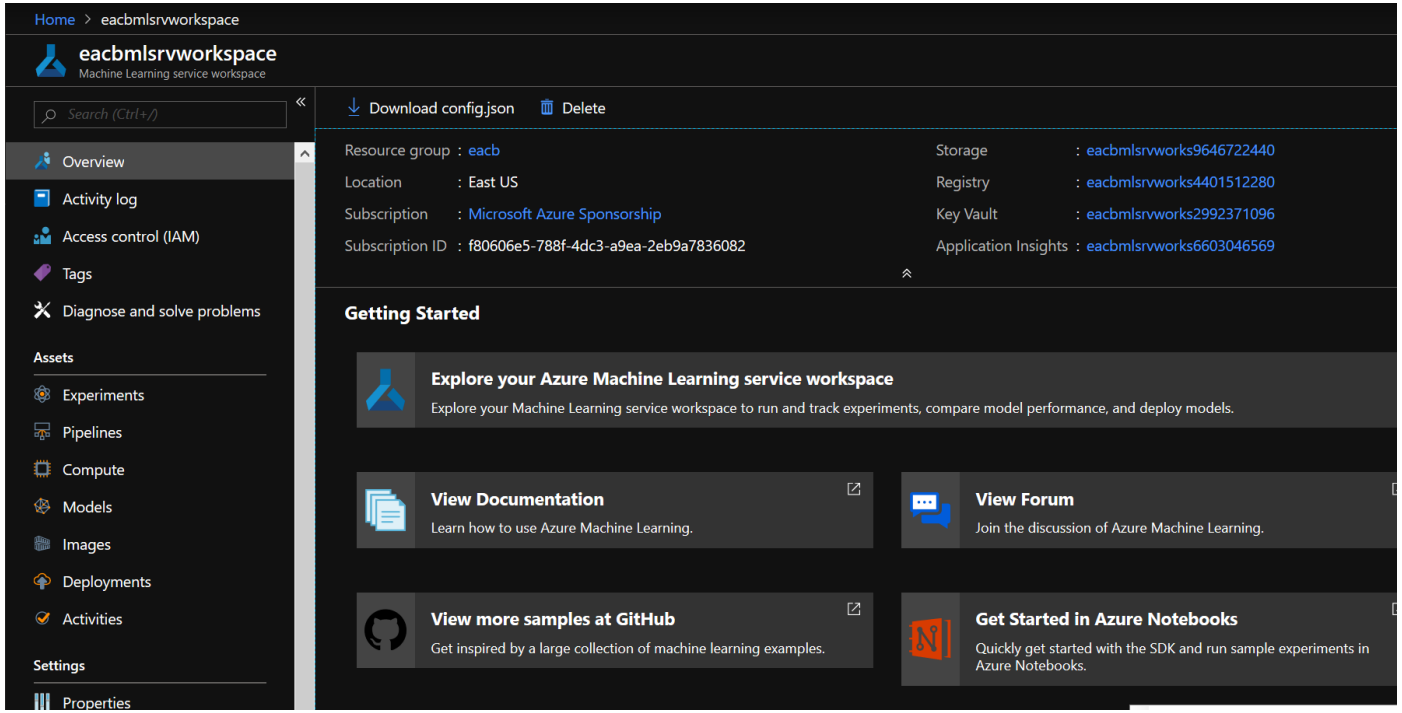
Création du service sous Azure

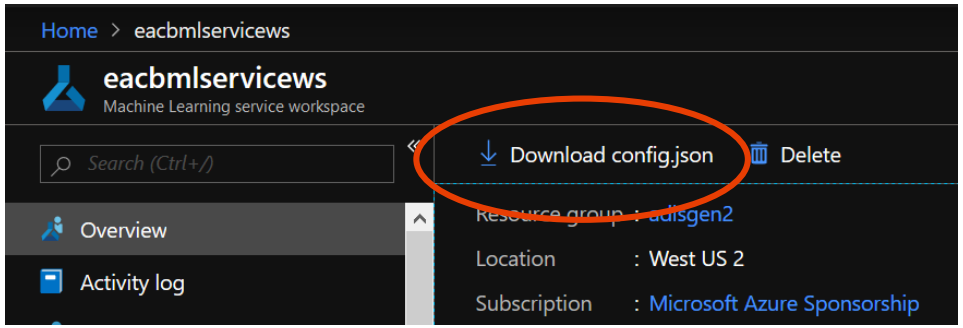
Définir :

- Un nom (unique) pour l'espace de travail
- Un groupe de ressources

Création associée :

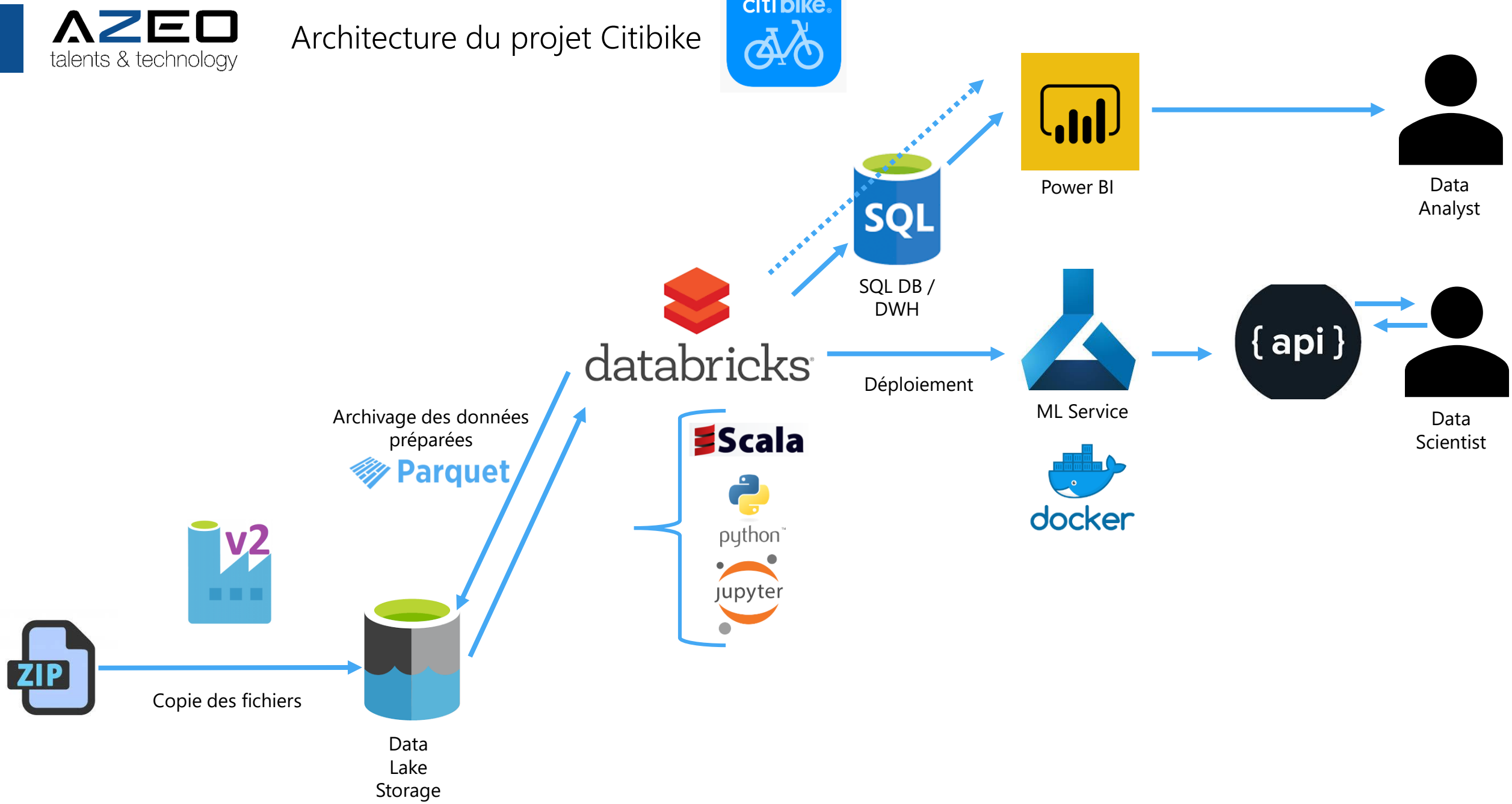
- d'un espace de stockage
- D'un container registry
- D'une key Vault
- D'Azure Application Insights






```
Config.json :
{
    "subscription_id": "XXXXXXXX-XXXX-XXXX-XXXX-XXXXXXXXXXXX",
    "resource_group": "eacb",
    "workspace_name": "eacbmlsrvworkspace"
}
```

Architecture du projet Citibike



AZEO
talents & technology

Plus que des mots…

DEMO

Enregistrement d'un modèle de Machine Learning (pickle)

Le modèle entrainé est exporté sous forme d'un binaire sérialisé.

Sources acceptées : Python et PySpark

```
from sklearn.model_selection import train_test_split

train_set, test_set = train_test_split(pandas_df, test_size=0.3, random_state=42)

X_train = train_set[['birthyear','gender','bikeid']]
X_test = test_set[['birthyear','gender','bikeid']]
duration_train_labels = train_set['tripduration']
duration_test_labels = test_set['tripduration']
```

```
from sklearn.linear_model import LinearRegression

lin_reg = LinearRegression()
lin_reg.fit(X_train, duration_train_labels)
```

```
import numpy as np
from sklearn.metrics import mean_squared_error

duration_predictions = lin_reg.predict(X_test)
lin_mse = mean_squared_error(duration_test_labels, duration_predictions)
lin_rmse = np.sqrt(lin_mse)
print(lin_rmse)
```

```
505.177307705
```

```
import pickle

# save the model to disk
filename = 'bike_duration_model_db.sav'
pickle.dump(lin_reg, open(filename, 'wb'))
```

Par défaut, le fichier s'enregistre dans « file:/databricks/driver » sur le cluster Databricks.

Identification de l'espace de travail

```python
from azureml.core import Workspace

try:
    ws = Workspace(subscription_id = subscription_id, resource_group = resource_group, workspace_name = workspace_name)
    ws.write_config()
    print('Workspace configuration succeeded. You are all set!')
except:
    print('Workspace not found. Run the cells below.')

ws.get_details()
```

```
Performing interactive authentication. Please follow the instructions on the terminal.
To sign in, use a web browser to open the page https://microsoft.com/devicelogin and enter the code BDXR36URJ to authenticate.
Interactive authentication successfully completed.
Workspace configuration succeeded. You are all set!
Out[34]:
{'location': 'westus2',
 'workspaceid': '15a21b13-4e7d-4c98-8059-dd0190942759',
 'identityType': 'SystemAssigned',
 'description': '',
 'name': 'eacbmlservicews',
 'id': '/subscriptions/f80606e5-788f-4dc3-a9ea-2eb9a7836082/resourceGroups/adlsgen2/providers/Microsoft.MachineLearningServices/workspaces/eacbmlservicews',
 'friendlyName': '',
 'creationTime': '2019-04-23T15:43:07.1380314+00:00',
 'keyVault': '/subscriptions/f80606e5-788f-4dc3-a9ea-2eb9a7836082/resourcegroups/adlsgen2/providers/microsoft.keyvault/vaults/eacbmlservicew6669147962',
 'identityPrincipalId': '41c0bea4-99cb-4305-b64d-36623d7dd52e',
 'identityTenantId': '8e2e7c2d-4702-496d-af6c-96e4bfc9f667',
 'storageAccount': '/subscriptions/f80606e5-788f-4dc3-a9ea-2eb9a7836082/resourcegroups/adlsgen2/providers/microsoft.storage/storageaccounts/eacbmlservicew2479569759',
 'applicationInsights': '/subscriptions/f80606e5-788f-4dc3-a9ea-2eb9a7836082/resourcegroups/adlsgen2/providers/microsoft.insights/components/eacbmlservicew6648238519',
 'type': 'Microsoft.MachineLearningServices/workspaces',
 'containerRegistry': '/subscriptions/f80606e5-788f-4dc3-a9ea-2eb9a7836082/resourcegroups/adlsgen2/providers/microsoft.containerregistry/registries/eacbmlservicew8493899780
'}

Command took 9.96 minutes -- by paul.peton@live.fr at 24/04/2019 à 22:17:14 on myCluster
```

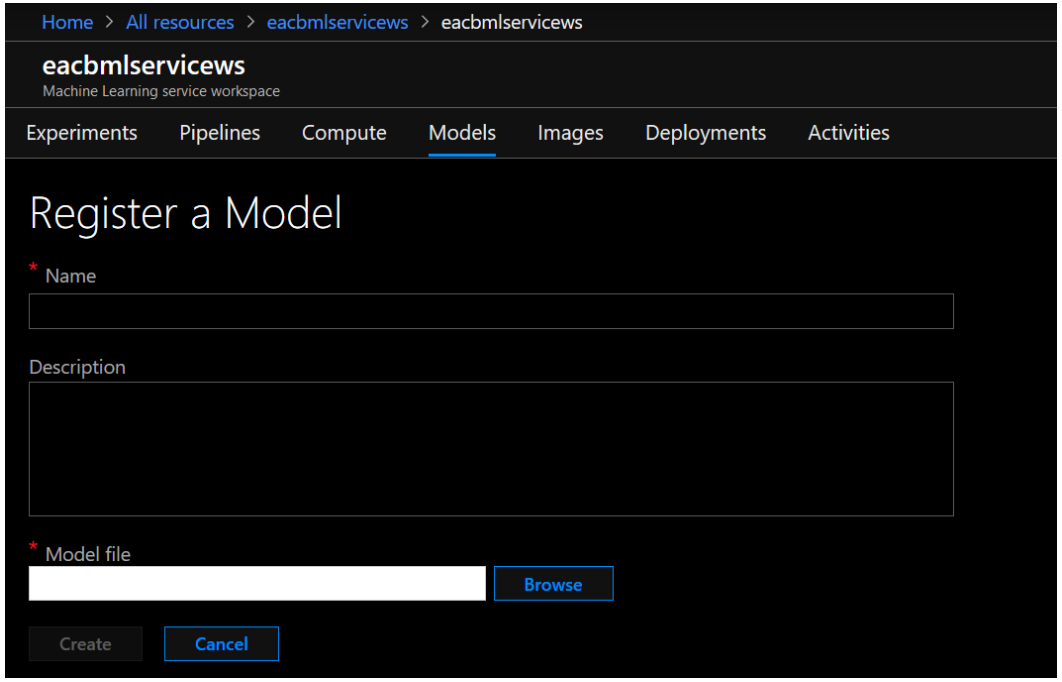
Enregistrer un modèle

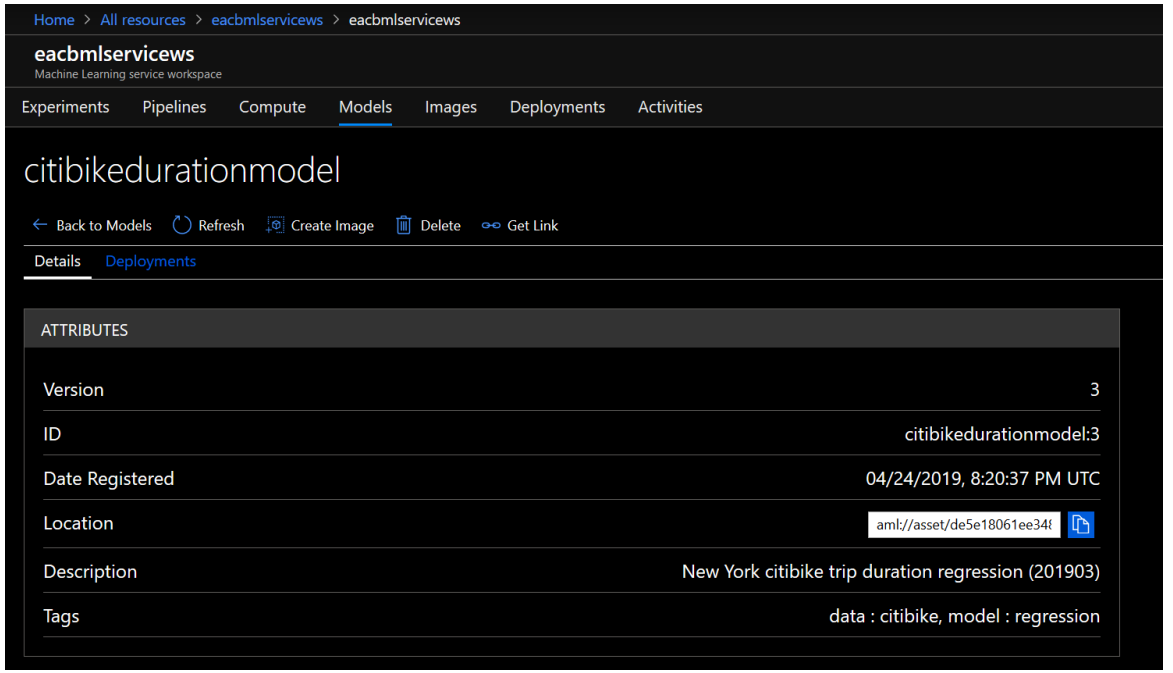
Par code ou par l'interface

```
Entrée [106]:
from azureml.core.model import Model

model_name = "citibikedurationmodel"
model = Model.register(model_path="bike_duration_model_201903.sav",
                       model_name=model_name,
                       tags={"data": "citibike", "model": "regression"},
                       description="New York citibike trip duration regression (201903)",
                       workspace=ws)
```

```
Registering model citibikedurationmodel
```

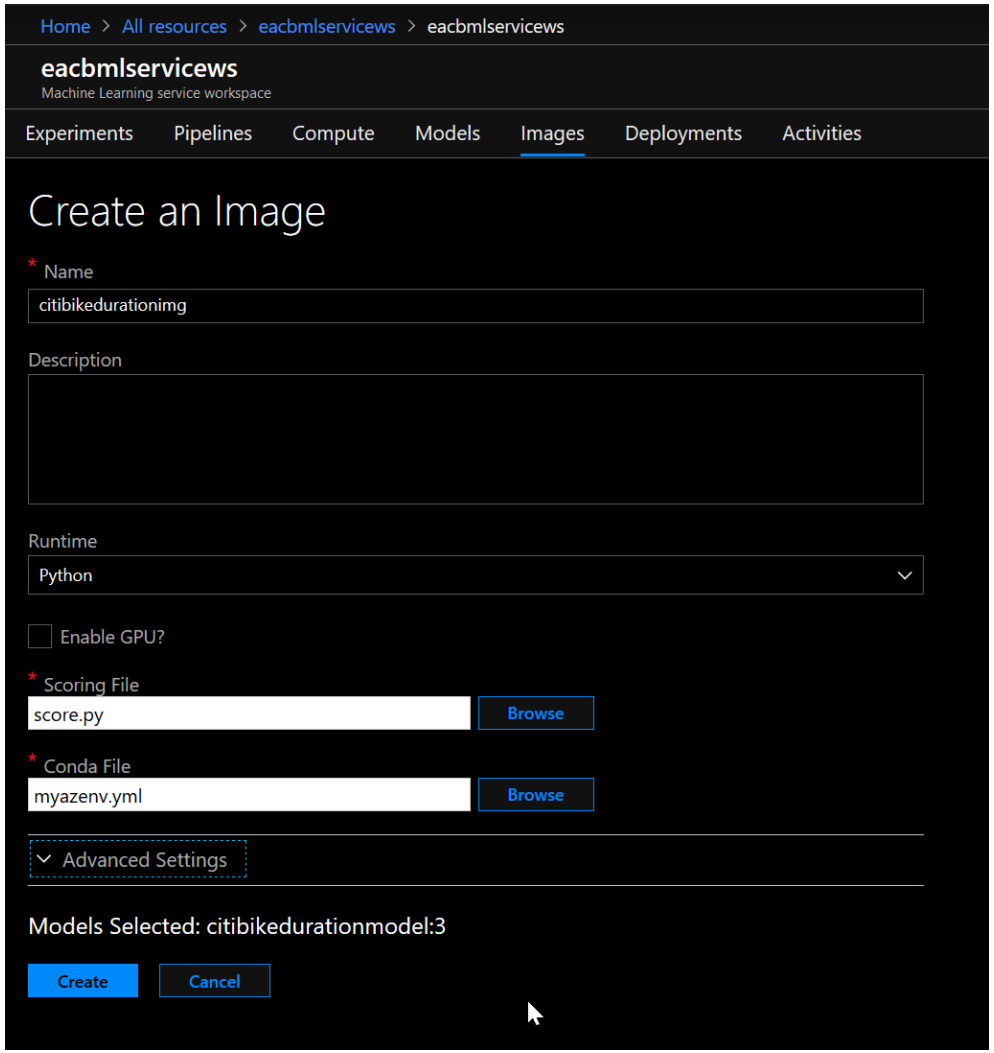




AZEO talents & technology

Créer une image

Par code ou par l'interface



Entrée [97]:

```
%%writefile score.py
import json
import numpy as np
import os
import pickle
from sklearn.externals import joblib

from azureml.core.model import Model

def init():
    global model
    # retreive the path to the model file using the model name
    model_path = Model.get_model_path('citibikedurationmodel')
    model = joblib.load(model_path)

def run(raw_data):
    data = np.array(json.loads(raw_data)['data'])
    # make prediction
    y_hat = model.predict(data)
    return json.dumps(y_hat.tolist())
```

```
Overwriting score.py
```

Entrée [98]:

```
from azureml.core.conda_dependencies import CondaDependencies

myenv = CondaDependencies()
myenv.add_conda_package("scikit-learn")

with open("myazenv.yml","w") as f:
    f.write(myenv.serialize_to_string())
```

Entrée [99]:

```
from azureml.core.webservice import AciWebservice

aciconfig = AciWebservice.deploy_configuration(cpu_cores=1,
                                               memory_gb=1,
                                               tags={"data": "citibike",  "method" : "sklearn regression"},
                                               description='Predict citibike trip duration with sklearn')
```

Déployer une image

Par code ou par l'interface

Entrée [42]:

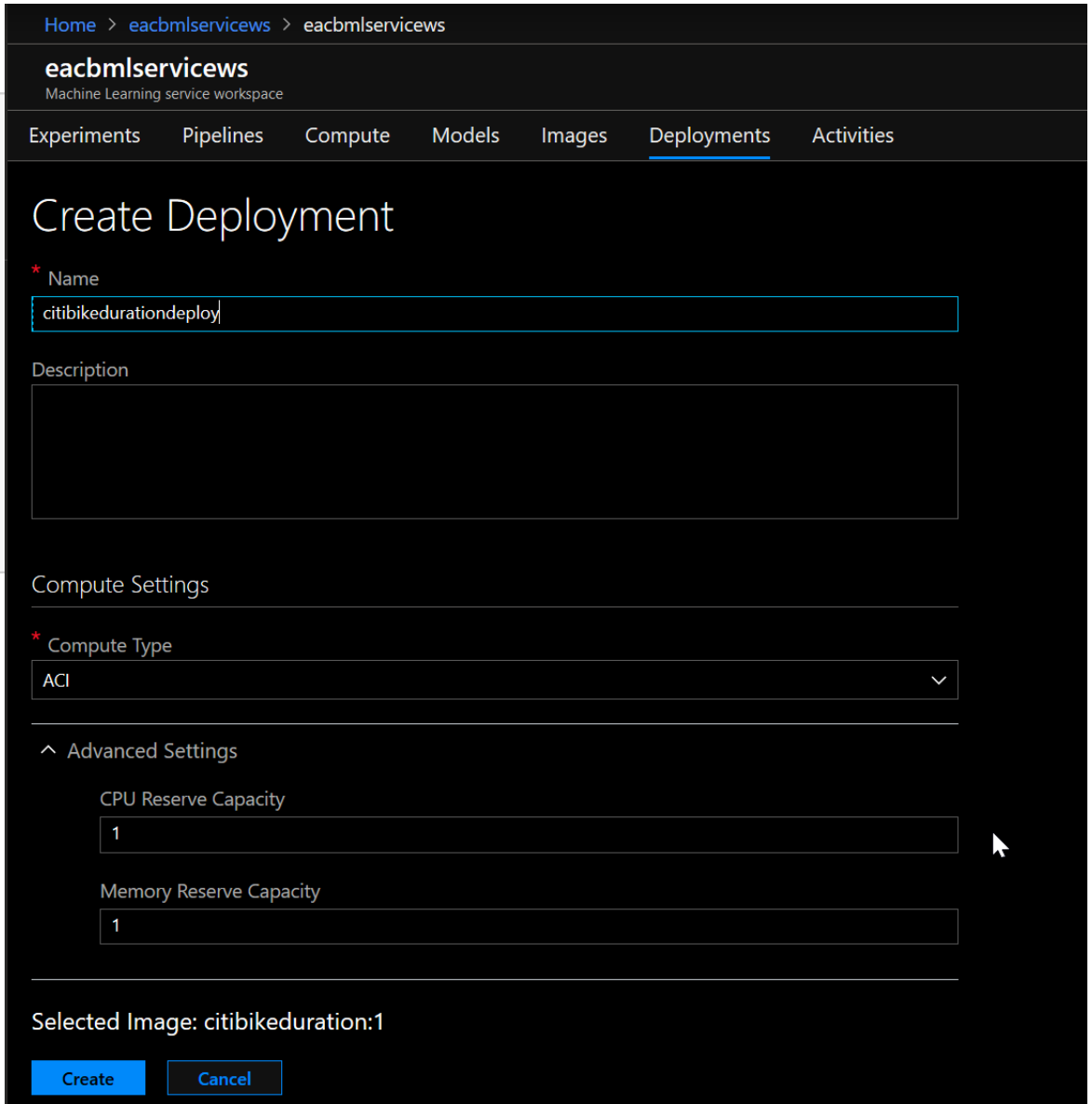
```
%%time

from azureml.core.webservice import Webservice
from azureml.core.image import ContainerImage

# configure the image
image_config = ContainerImage.image_configuration(execution_script="score.py",
                                                   runtime="python",
                                                   conda_file="myazenv.yml")

service = Webservice.deploy_from_model(workspace=ws,
                                       name='citibikedurationmodel',
                                       deployment_config=aciconfig,
                                       models=[model],
                                       image_config=image_config)

service.wait_for_deployment(show_output=True)
```



Interroger le web service

Par code

Entrée [108]:

```
import requests
import numpy as np

# send a random row from the test set to score
random_index = np.random.randint(0, len(X_test)-1)
print(random_index)
input_data = "{\"data\": [" + str(list(X_test.iloc[random_index])) + "]}"
print(input_data)
```

```
4302
{"data": [[1968, 2, 29617]]}
```

Entrée [109]:

```
headers = {'Content-Type':'application/json'}

# for AKS deployment you'd need to the service key in the header as well
# api_key = service.get_key()
# headers = {'Content-Type':'application/json',  'Authorization':('Bearer '+ api_key)}

resp = requests.post("http://13.83.243.48:80/score", input_data, headers=headers)

print("POST to url", "http://13.83.243.48:80/score")
#print("input data:", input_data)
print("label:", duration_test_labels.iloc[random_index])
print("prediction:", resp.text)
```

```
POST to url http://13.83.243.48:80/score
label: 244
prediction: "[405.00490254229214]"
```

Performances et coût global de l'architecture

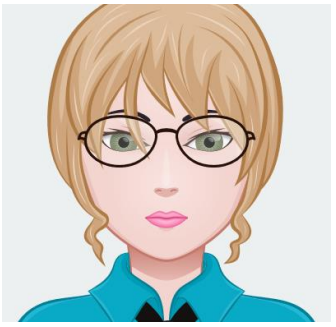
Historique : 2013/06 – 2019/03
Nombre de fichiers : 70
Nombre de lignes : 74 631 715
Nombre de champs : 11
Taille en octets :
- avant préparation (zip) : 2 589 493 115
- avant préparation (csv) : 13 987 355 730
- après préparation (parquet) : 1 637 188 856

Temps de copie : adf 49s
Temps de préparation et stockage : 20 minutes
- shell unzip : 12 min
- PySpark transformations : 4 min
- création table metastore Hive : 2 min
- sauvegarde au format parquet : 2 min

Temps de modélisation (régression) : ?
Temps de création de l'image :~15 min

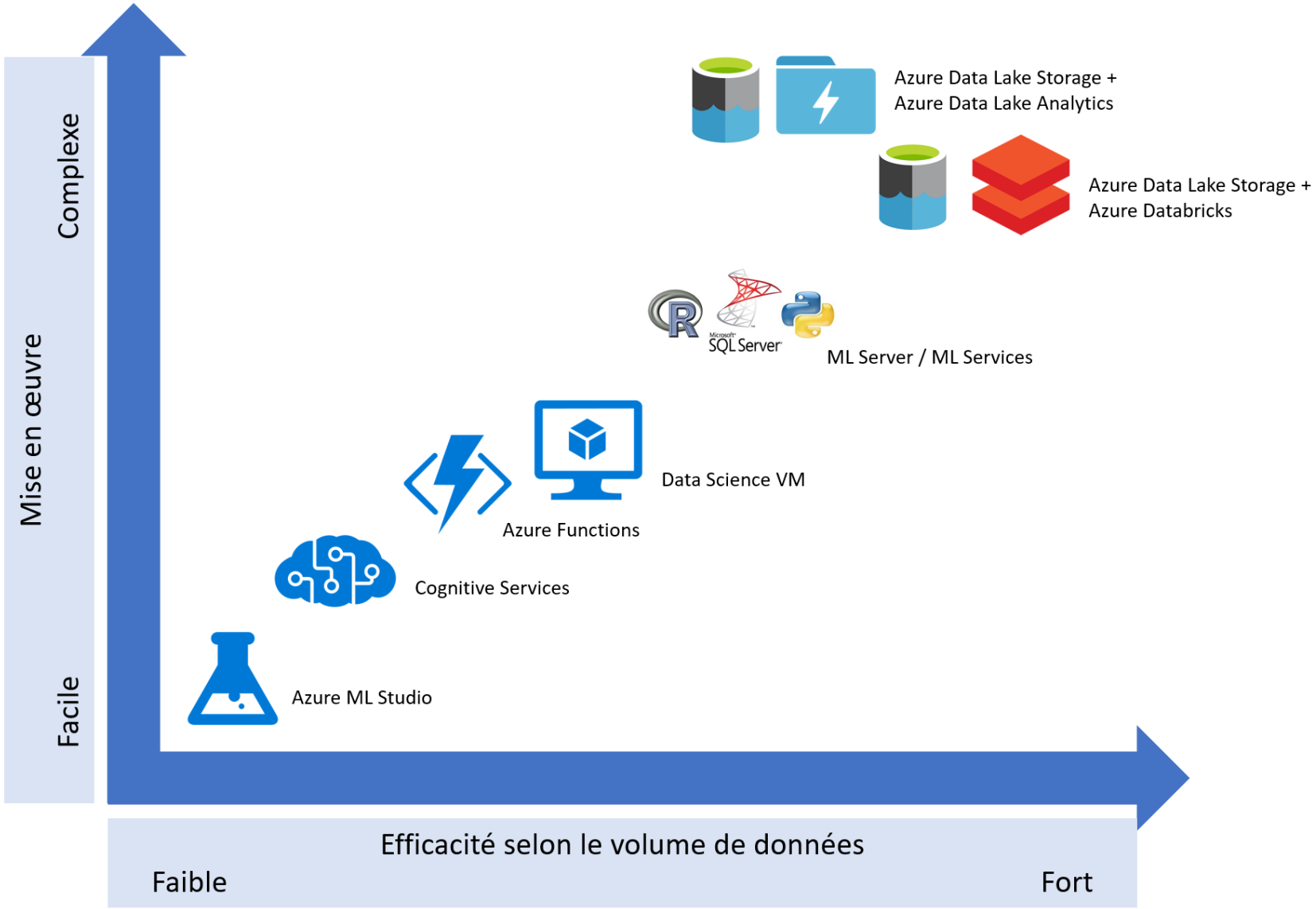
Azure Data Factory v2 : 2 €
Azure Blob Storage ou
Azure Data Lake Storage Gen2 : 2 €
Data Science Virtual Machine : selon VM
Azure Databricks : coût VM + DBU
Azure Machine Learning Service : coût VM + surcharge (~0,03€ par cœur)
1 licence Pro Power BI : 8,4€HT

« Rapide ! (sans chercher la perf du modèle »



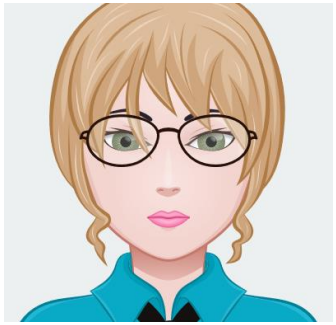
« Pas cher ! Selon les VM… »

Classement (subjectif) des solutions Microsoft pour la Data Science



Conclusion

Points forts et améliorations attendues



AZEO
talents & technology

Point de vue sur l'état de l'art de l'industrialisation du ML sur Azure

Les points forts sur lesquels s'appuyer :

Les notebooks sont partout (sauf dans le cœur des puristes du code), en local comme sur le cloud

Les services communiquent (de plus en plus) entre eux et se planifient facilement

La configuration et le déploiement des containers sont possibles par l'interface mais aussi automatisable par le code

L'histoire se finit en visualisations dans Power BI (connecteur JDBC vers les tables de Databricks)

Ce qui nous faciliterait la tâche :

Un outil de Data Preparation intégrant des connecteurs prédéfinis (Workbench ?)

Une interface pour le pipeline de ML (Studio ?)

Une séparation simple DEV / QUALIF / PROD

Une meilleure maîtrise du Software Engineering (et de Scala) par les Data Scientists…

Ou une meilleure scalabilité de PySpark

Et surtout…traiter des cas d'usage par échantillonnage

Morale de l'histoire ?

Hadley Wickham : « The Big Data Mirage »

"(many Big Data problems)

Can be reduced to a small/ medium data problem with subsetting / sampling / summarising (90%)

Can be reduced to a very large number of small data problems (9%)"

AZEO
talents & technology

VOS QUESTIONS ?

Siège Social
52, avenue André Morizet - 92100 Boulogne-Billancourt
contact@azeo.com | 01.83.62.65.54 | www.azeo.com

AZEO
talents & technology

Beaucoup de sujets connexes à aborder

Le projet Open Source mlflow

Son intégration dans Azure ML Service

Automated Machine Learning

(par code et interface à venir)

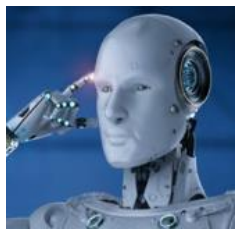
L'interprétabilité du modèle

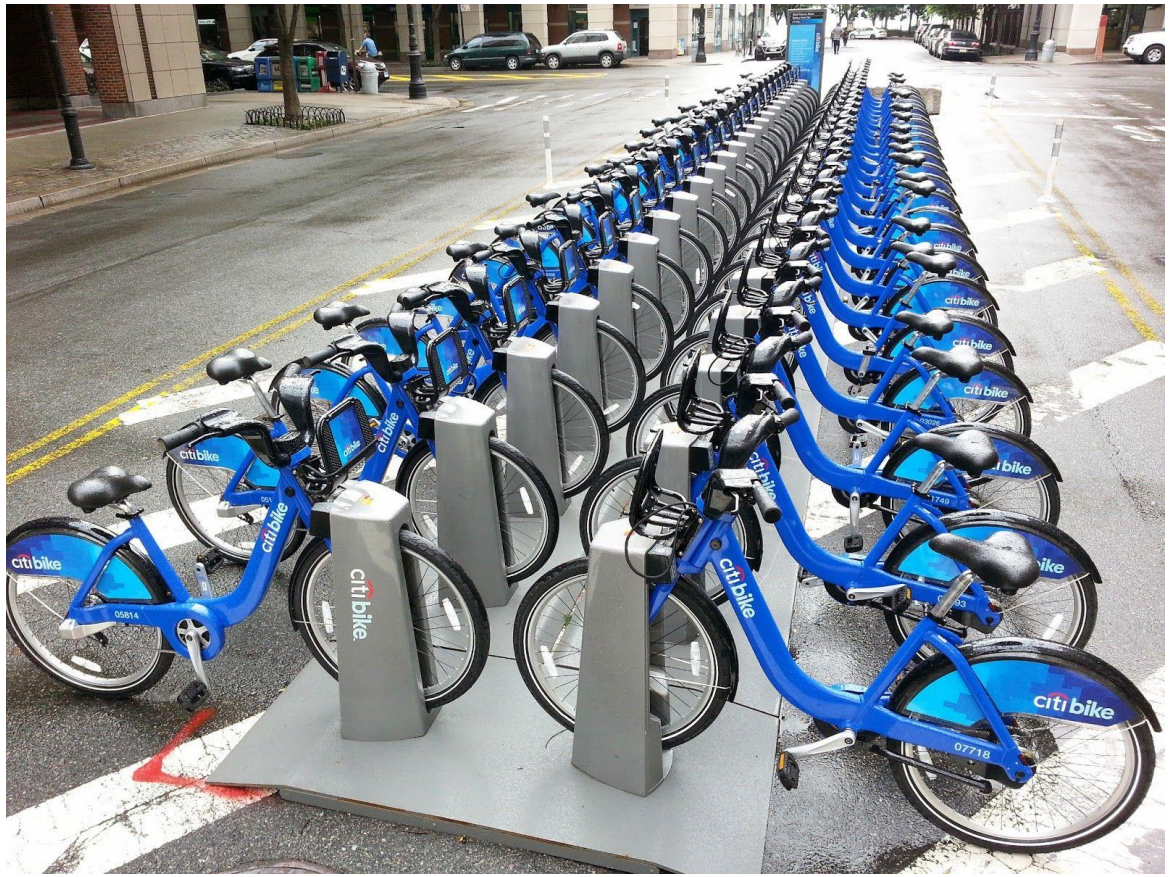
L'intégration des données temps réel

Etc.





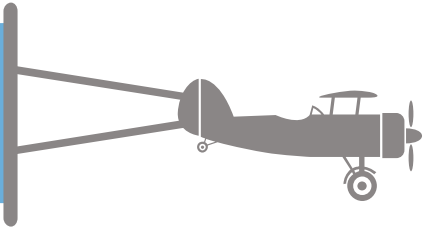


To be continued…

2019 Global Azure BOOTCAMP

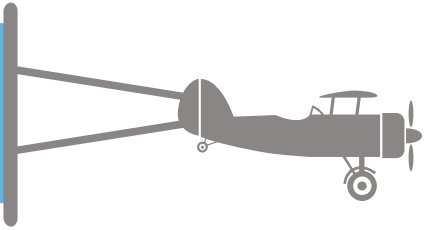
All attendees get the following :



Sponsor	Offering
Cloudmonix **https://cloudmonix.com**	Cloudmonix offers 2 months free of Ultimate or Pro plans for Azure monitoring! http://bit.ly/globalazure2019-cloudmonix Code: GAB2019
Serverless 360 **https://serverless360.com**	Serverless360 is offering a limited time Gold plan for free! http://bit.ly/globalazure2019-serverless360
KEMP **https://kemptechnologies.com**	Kemp is providing a trial of their LoadMaster (load balancer) tool, as well as a free Kemp 360 License to all attendees! http://bit.ly/globalazure2019-kemploadmaster http://bit.ly/globalazure2019-kemp360

2019 Global Azure BOOTCAMP

All attendees get the following :



Sponsor	Offering
Progate **https://progate.com**	Progate is offering a free month access to their entire platform! http://bit.ly/progateaccessgab2019
RevDeBug **https://revdebug.com/**	RevDeBug is offering an enterprise license for 3 months for ALL attendees! http://bit.ly/globalazure2019-revdebug

Merci
d'être venus

A bientôt !