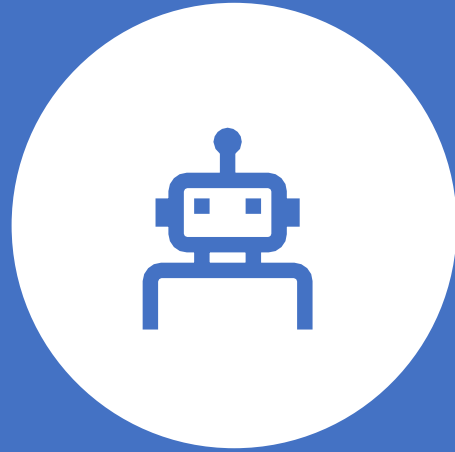


Workshop Cloud Experts

« Vos analyses de Machine Learning avec Azure Databricks »

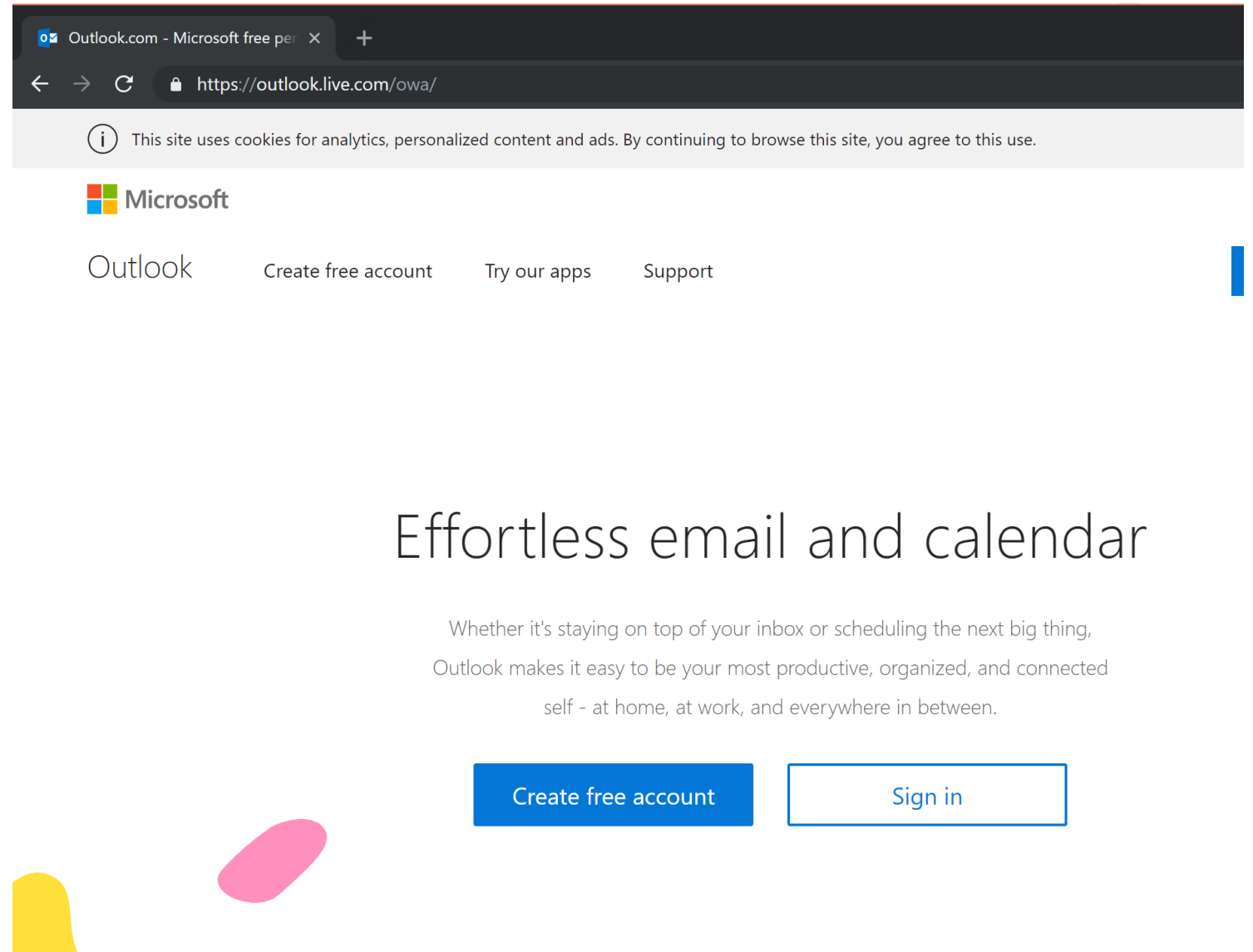
Mercredi 5 juin 2019

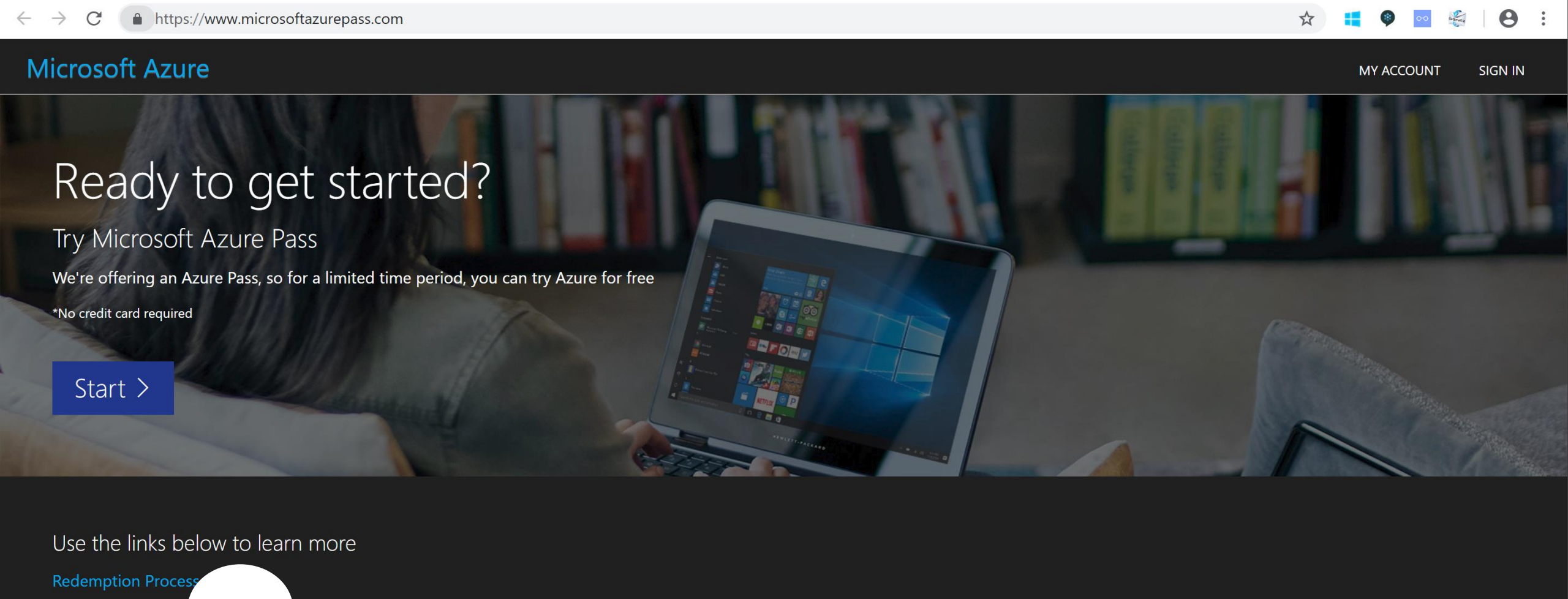




Création d'un compte Azure via un redeem code

1) Création adresse email outlook





2) Activation Azure Pass <https://www.microsoftazurepass.com>

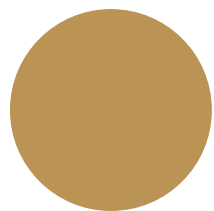
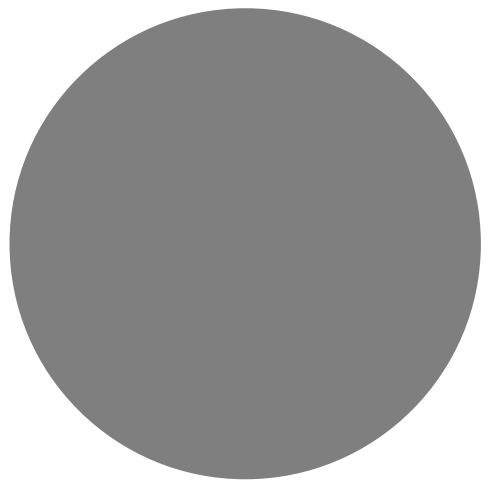
2) Activation Azure Pass



Se connecter
avec le compte
email Microsoft
créé



Redeem code



Provisionnement du service Azure Databricks

1. Provisionnement du service Azure Databricks

The screenshot shows the Microsoft Azure portal interface. On the left is a dark sidebar with navigation options: 'Create a resource', 'Home', 'Dashboard', 'All services', 'FAVORITES', 'All resources', 'Resource groups', 'App Services', 'Function Apps', 'SQL databases', 'Azure Cosmos DB', 'Virtual machines', 'Load balancers', and 'Storage accounts'. The main area is titled 'Microsoft Azure' and contains a search bar with the text 'Search resources, services, and docs'. Below the search bar, the breadcrumb path is 'Home > New > Marketplace > Everything'. The 'Marketplace' section is active, showing a list of categories: 'My Saved List' (0 items), 'Everything' (selected), 'Compute', 'Networking', 'Storage', 'Web', 'Mobile', 'Containers', 'Databases', and 'Analytics'. The 'Everything' view displays search results for 'databricks'. There are filters for 'Pricing' (All), 'Operating System' (All), and 'Publisher' (All). The results table shows one entry: 'Azure Databricks' by 'Microsoft' in the 'Analytics' category. This entry is highlighted with a red rounded rectangle.

Preview Microsoft Azure Report a bug Search resources, services, and docs

Home > New > Marketplace > Everything

Marketplace

My Saved List 0

Everything

Compute

Networking

Storage

Web

Mobile

Containers

Databases


Analytics

Everything

databricks

Pricing All Operating System All Publisher All

Results

NAME	PUBLISHER	CATEGORY
 Azure Databricks	Microsoft	Analytics

1. Provisionnement du service Azure Databricks

The screenshot displays the Azure Databricks service overview page. The left sidebar contains navigation options: 'Add', 'Edit columns', and 'More'. Below these is a search bar labeled 'Filter by name...' and a list of resources, with 'Databricks' selected. The main content area is divided into two sections. The top section, titled 'Databricks Azure Databricks Service', shows the 'Overview' tab selected in the left-hand menu. This section displays key information: 'Resource group (change) AzureDB', 'Subscription (change) Microsoft Azure Internal Consumption', and 'Subscription ID' (redacted). To the right, a 'Managed Resource Group' section shows the 'URL' as 'https://westeurope.azuredatabricks.net' and the 'Pricing Tier' as 'premium'. Below this, a large red Databricks logo is centered above a blue 'Launch Workspace' button. The bottom section features a grid of links: 'Documentation', 'Getting Started', 'Import Data from File', 'Import Data from Azure Storage', 'Notebook', and 'Admin Guide', each accompanied by a small icon.

Dashboard > Azure Databricks > Databricks

Azure Databricks Microsoft

+ Add Edit columns More

Filter by name...

NAME ↑↓

Databricks

Databricks Azure Databricks Service

Search (Ctrl+/)

Overview

Activity log

Access control (IAM)

Tags

Settings

Virtual Network Peerings

Locks

Automation script

Support + troubleshooting

New support request

Delete

Resource group (change) AzureDB

Subscription (change) Microsoft Azure Internal Consumption

Subscription ID

Managed Resource Group

URL https://westeurope.azuredatabricks.net

Pricing Tier premium

Launch Workspace

Documentation

Getting Started

Import Data from File

Import Data from Azure Storage

Notebook

Admin Guide

2. Création du Cluster Spark Databricks

The screenshot displays the Microsoft Azure Databricks interface. On the left is a dark sidebar with navigation icons for Azure Databricks, Home, Workspace, Recents, Data, Clusters (highlighted), Jobs, and Search. The main content area is titled 'Clusters / Cluster DB'. It features a green status indicator and a title 'Cluster DB' with a pin icon. Action buttons include 'Edit', 'Clone', 'Restart', 'Terminate', and 'Delete'. Below these are tabs for 'Configuration', 'Notebooks (0)', 'Libraries', 'Event Log', 'Spark UI', 'Driver Logs', and 'Spark Cluster UI - Master'. The 'Configuration' tab is active, showing settings for 'Cluster Mode' (Standard), 'Databricks Runtime Version' (5.3), and 'Python Version' (3). The 'Autopilot Options' section is highlighted with a red rectangle and contains two checked options: 'Enable autoscaling' and 'Terminate after 120 minutes of inactivity'. Below this, the 'Worker Type' is set to 'Standard_DS3_v2' with '14.0 GB Memory, 4 Cores, 0.75 DBU', and 'Min Workers' is 2, 'Max Workers' is 8. The 'Driver Type' is also 'Standard_DS3_v2' with the same specifications. An 'Advanced Options' link is at the bottom.

Microsoft Azure

Clusters / Cluster DB

Cluster DB

Edit Clone Restart Terminate Delete

Configuration Notebooks (0) Libraries Event Log Spark UI Driver Logs Spark Cluster UI - Master

Cluster Mode

Standard

Databricks Runtime Version

5.3 (includes Apache Spark 2.4.0, Scala 2.11)

Python Version

3

Autopilot Options

☒ Enable autoscaling

☒ Terminate after 120 minutes of inactivity

Worker Type

Standard_DS3_v2 14.0 GB Memory, 4 Cores, 0.75 DBU

Min Workers Max Workers

2 8

Driver Type

Standard_DS3_v2 14.0 GB Memory, 4 Cores, 0.75 DBU

Advanced Options

Bien vérifier que les options *Enable autoscaling* et *Terminate after X minutes of activity* soient activées.

Cela permet de mettre en place *l'autoscaling* entre les *workers* du cluster Spark et de désactiver automatiquement le cluster en cas d'inactivité spécifiée par un timeout.

2. Création du Cluster Spark Databricks

Microsoft Azure

Azure Databricks

Home

Workspace

Recents

Clusters

[+ Create Cluster](#)

All Created by me Accessible by

▼ Interactive Clusters

Name	State	Nodes	Driver	Worker	Runtime	Creator	
Cluster DB	Running	3	Standard_DS3_v2	Standard_DS3_v2	5.3 (includes Apache...	seretkow@microso...	0
Cluster DB 54	Terminated	-	Standard_DS3_v2	Standard_DS3_v2	5.4 Beta (includes A...	seretkow@microso...	0
Cluster GPU	Terminated	-	Standard_NC12 (b...	Standard_NC12 (b...	5.4 Beta (includes A...	seretkow@microso...	0

3. Importation des librairies Azure ML service dans Databricks

Microsoft Azure

Azure Databricks

Home

Workspace

Recents

Data

Clusters

Jobs

Search

Create Library

Library Source

Upload DBFS **PyPI** Maven CRAN

Repository ?

Optional

Package

azureml-sdk[databricks]

Create Cancel

azureml-sdk[databricks]

azureml-sdk[automl_databricks]

4. Association des librairies Azure ML service au cluster Databricks

Microsoft Azure



Azure Databricks



Home



Workspace



Recents

Clusters / Cluster DB

Cluster DB

Edit

Clone

Restart

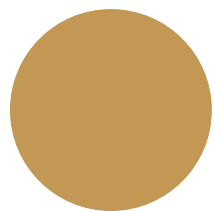
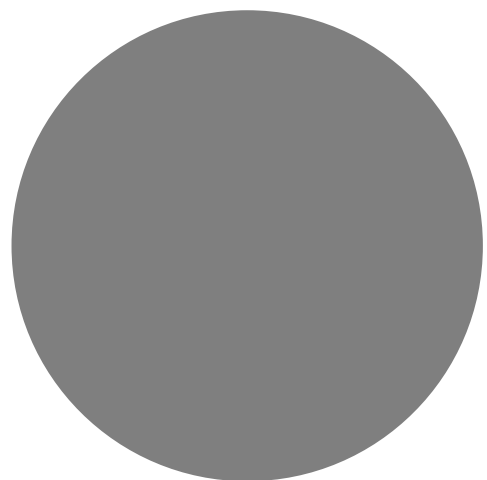
Terminate

Delete

Configuration Notebooks (0) Libraries Event Log Spark UI Driver Logs Spark Cluster UI - Master ▼

Uninstall Install New

<input type="checkbox"/>	Name	Type	Status	Source
<input type="checkbox"/>	azureml-sdk[automl_databricks]	PyPI	● Installed	
<input type="checkbox"/>	azureml-sdk[databricks]	PyPI	● Installed	



Provisionnement du service Azure ML service

5. Provisionnement du service Azure ML service

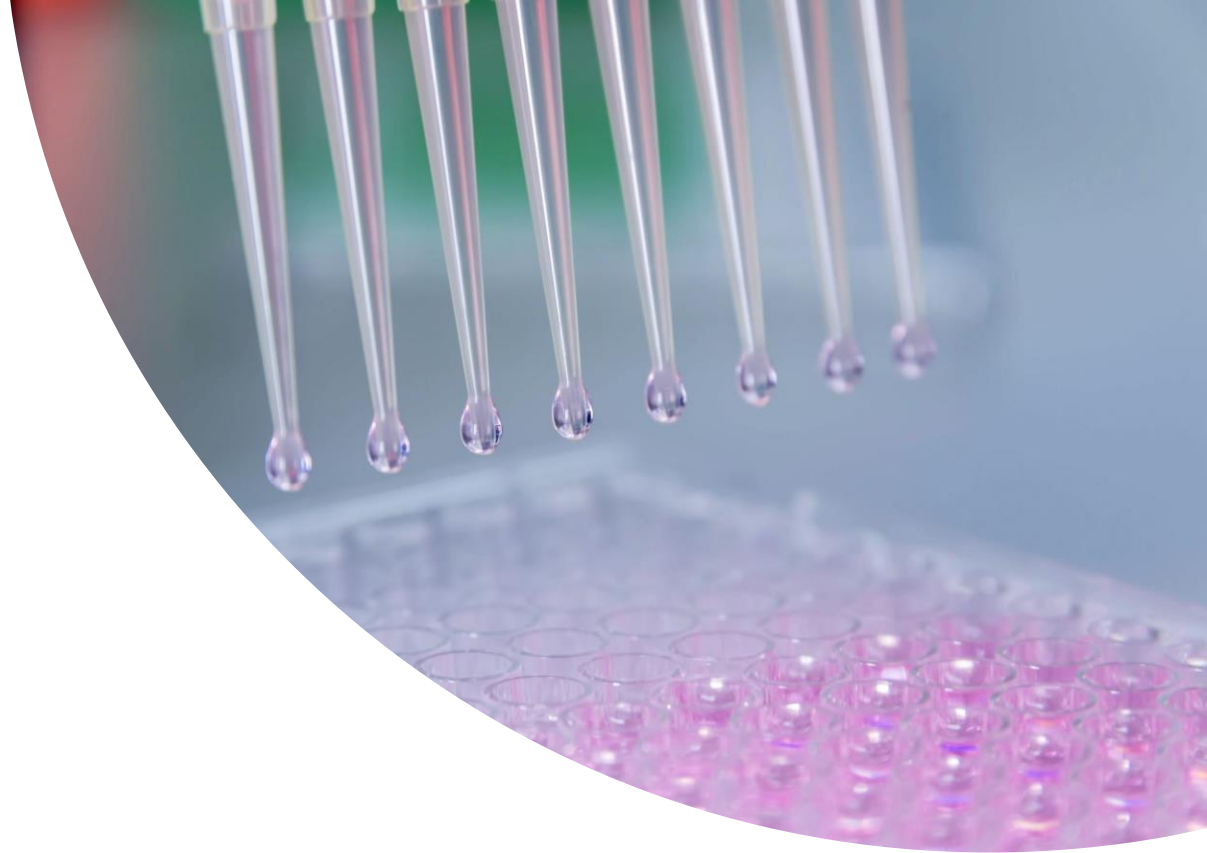
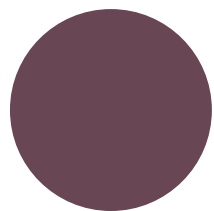
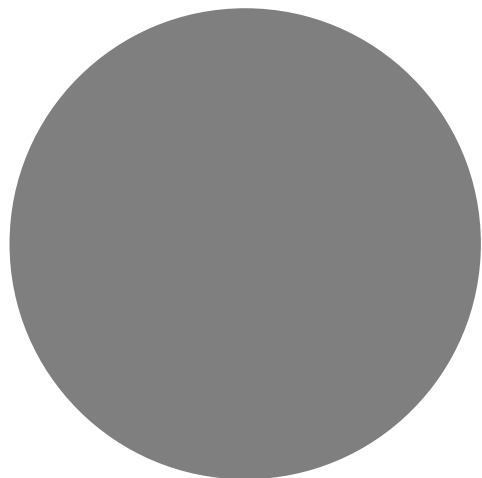
The screenshot shows the Azure portal interface for creating a new ML service workspace. The left sidebar contains navigation links such as 'Create a resource', 'Home', 'Dashboard', 'All services', 'FAVORITES', 'All resources', 'Resource groups', 'App Services', 'Function Apps', 'SQL databases', 'Azure Cosmos DB', 'Virtual machines', 'Load balancers', 'Storage accounts', 'Virtual networks', 'Azure Active Directory', 'Monitor', 'Advisor', 'Security Center', 'Cost Management + Billing', and 'Help + support'. The main content area is titled 'ML service workspace' and includes the following fields:

- Workspace name:** A text input field with the placeholder 'Enter the workspace name'.
- Subscription:** A dropdown menu currently showing 'Microsoft Azure Internal Consumption'.
- Resource group:** A dropdown menu with 'Select existing...' and a link to 'Create new'.
- Location:** A dropdown menu currently showing 'West Europe'.

Below these fields, an information box states: 'For your convenience, these resources are added automatically to the workspace, if regionally available: [Azure Container Registry](#), [Azure storage](#), [Azure Application Insights](#) and [Azure Key Vault](#).' At the bottom, there is a blue 'Create' button and a link to 'Automation options'.

Une fois le service provisionné il sera nécessaire de retenir le nom du workspace, la localisation, le ressource groupe ainsi que l'ID du service.

Ces valeurs seront nécessaires pour assurer une connexion au workspace Azure ML service depuis un notebook Azure Databricks.



Exercices du lab

Exercices

Notebooks & données

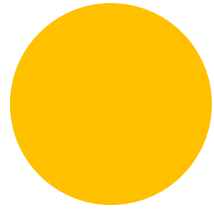
Les données sont dans Data.zip

Le Fichier .DBC est une archive Databricks qui contient tous les notebooks.

The screenshot shows the GitHub repository page for 'retkowsky / Cloud_Workshop_AzureDatabricks'. The repository has 2 commits, 1 branch, 0 releases, and 1 contributor. The 'Code' tab is selected, showing a list of files. The files include 01. Installation et Configuration.html, 02. Introduction Apache Spark avec Databricks.html, 03. Data Exploration.html, 04. Clustering.html, 05. Regression.html, 06. Classification.html, 07. Classification Evaluation.html, 08. Cross Validation.html, 09. Parameter Tuning.html, 10. AutoML avec Azure ML service.html, 11. Déploiement avec AKS.html, 12. Azure Databricks et MLFlow - Introduction.html, 13. Azure Databricks et MLFlow - Scikit Learn.html, 14. Azure Databricks et MLFlow - Batch.html, Cloud Workshop.dbc, README.md, and index.html. The 'Cloud Workshop.dbc' file is highlighted.

File Name	Action	Time
01. Installation et Configuration.html	Add files via upload	6 minutes ago
02. Introduction Apache Spark avec Databricks.html	Add files via upload	6 minutes ago
03. Data Exploration.html	Add files via upload	6 minutes ago
04. Clustering.html	Add files via upload	6 minutes ago
05. Regression.html	Add files via upload	6 minutes ago
06. Classification.html	Add files via upload	6 minutes ago
07. Classification Evaluation.html	Add files via upload	6 minutes ago
08. Cross Validation.html	Add files via upload	6 minutes ago
09. Parameter Tuning.html	Add files via upload	6 minutes ago
10. AutoML avec Azure ML service.html	Add files via upload	6 minutes ago
11. Déploiement avec AKS.html	Add files via upload	6 minutes ago
12. Azure Databricks et MLFlow - Introduction.html	Add files via upload	6 minutes ago
13. Azure Databricks et MLFlow - Scikit Learn.html	Add files via upload	6 minutes ago
14. Azure Databricks et MLFlow - Batch.html	Add files via upload	6 minutes ago
Cloud Workshop.dbc	Add files via upload	6 minutes ago
README.md	Initial commit	7 minutes ago
index.html	Add files via upload	6 minutes ago

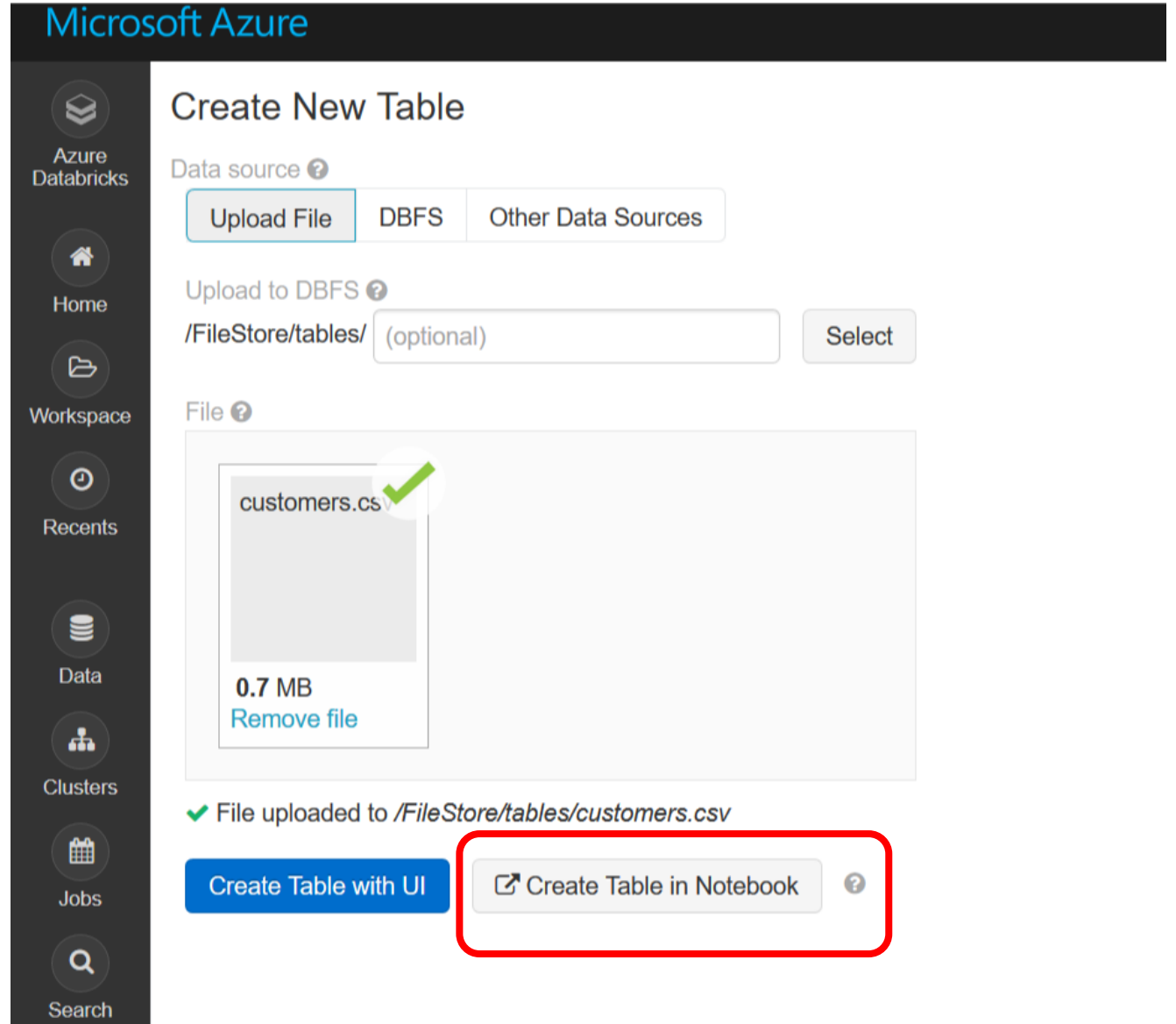
<https://aka.ms/WorkshopDatabricks>



Chargement des
données

Chargement des données

1. Télécharger le fichier Data.zip
2. Dézipper ce fichier sur votre poste de travail pour extraire les différents fichiers .CSV
3. Charger chaque fichier .CSV dans DBFS (Databricks File System) à l'aide de l'assistant suivant et cliquer ensuite sur *Create Table in Notebook* pour visualiser le code Python correspondant.



Microsoft Azure

Create New Table

Data source ?

Upload File DBFS Other Data Sources

Upload to DBFS ?

/FileStore/tables/ (optional) Select

File ?

customers.csv ✓

0.7 MB
Remove file

✓ File uploaded to /FileStore/tables/customers.csv

Create Table with UI Create Table in Notebook ?

6. Chargement des données

Remplacer les valeurs initiales **false** par **true** du code Python pour inférer automatiquement les propriétés des variables et récupérer le nom de chaque variable en première ligne.

Un *dataframe* est ainsi créé.

Microsoft Azure

2019-03-11 - DBFS Example (5) (Python)

Detached File View: Code Permissions Run All Clear

Cmd 1

Overview

This notebook will show you how to create and query a table or DataFrame that you uploaded to DBFS. [DBFS](#) is a Data Lake File System that assumes that you have a file already inside of DBFS that you would like to read from.

This notebook is written in **Python** so the default cell type is Python. However, you can use different languages by using the language dropdown in the top right corner.

Cmd 2

```
# File location and type
file_location = "/FileStore/tables/flights.csv"
file_type = "csv"

# CSV options
infer_schema = "false"
first_row_is_header = "false"
delimiter = ","

# The applied options are for CSV files. For other file types, these will be ignored.
df = spark.read.format(file_type) \
    .option("inferSchema", infer_schema) \
    .option("header", first_row_is_header) \
    .option("sep", delimiter) \
    .load(file_location)

display(df)
```

