



SSID: MSFTGUEST
code: msevent597rq

<https://github.com/guigirard/labDatabricksAzure>

Databricks +
Spark

Microsoft hands-on lab

Big Data : Databricks + Spark



Vos animateurs aujourd'hui



Guillaume Girard Conseiller Big Data
Expertise Spark, Hadoop, SQL, Microsoft BI



Christophe Botek Conseiller Big Data
Expertise Scala, Spark, Databricks, SQL

70

Passionnés
de données

144

Clients
satisfaits

641

Projets
livrés

482k

Heures sur
les données

Mission : En tant que centre d'excellence en Business et Data Intelligence, agileDSS aide les entreprises à devenir « **data-driven** »

2003



agileWorkflow

2010



Intelligence
d'affaires

2015



Business
consulting

2016



Big Data

2019



Data Science



Agenda

Introduction à Spark

- + Hadoop : petit rappel
- + Qu'est-ce que Spark?
- + Écosystème Spark
- + Les APIs
- + Le RDD
- + Le Dataframe
- + Le Dataset
- + Les Commandes
- + Le DAG
- + Spark SQL
- + Comment Spark s'exécute?
- + Spark & Azure



Spark

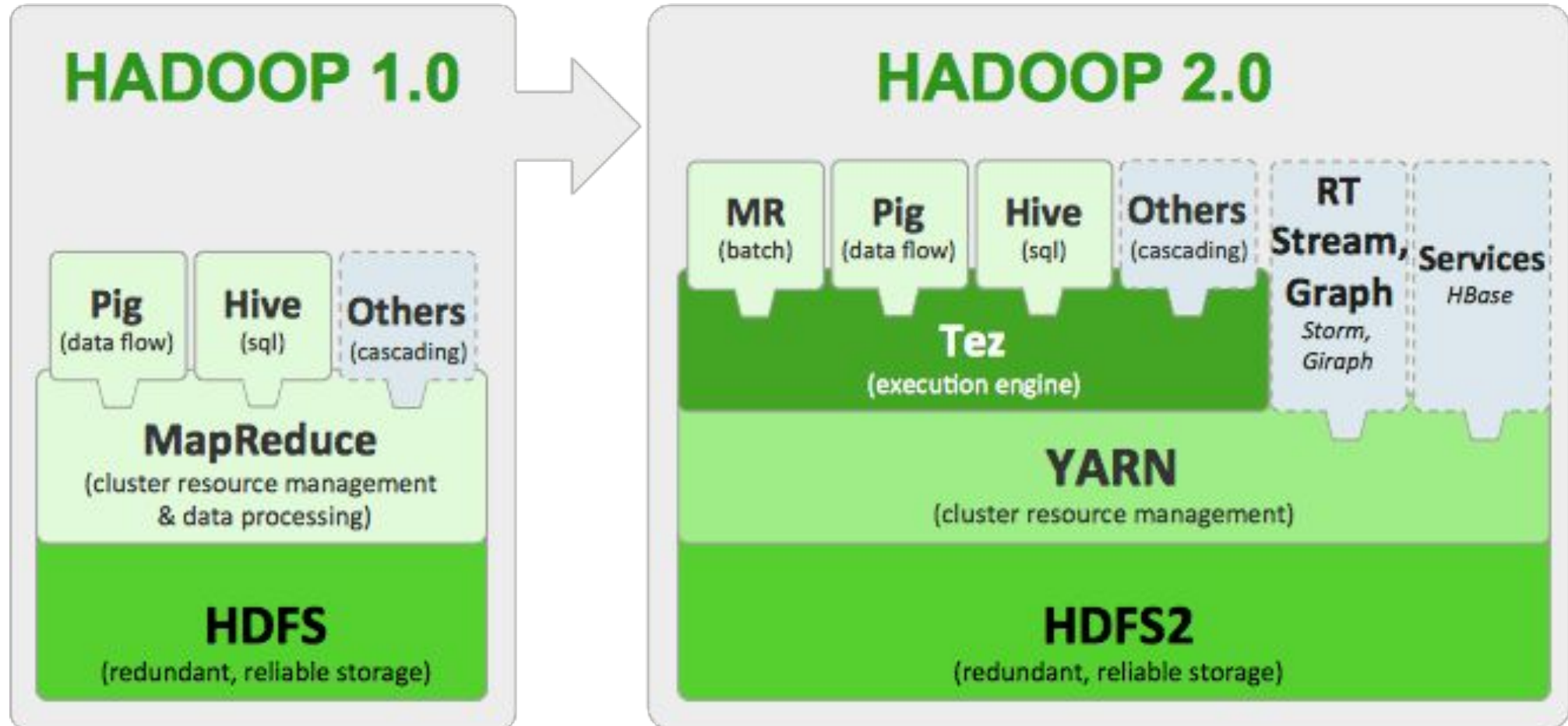
Introduction

Spark



Hadoop

Petit rappel





On a dit Spark?

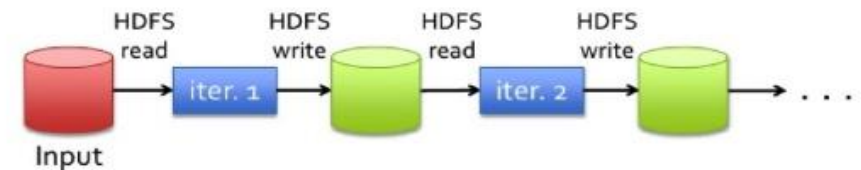
Calcul parallèle en mémoire sous stéroïdes

- + Remplacement de MapReduce
- + Remplacement de Hive
- + Forte intégration dans l'écosystème Hadoop
- + APIs haut niveau inspirées de la programmation fonctionnelle

Développement

- + Création en 2009
- + V1.0 en 2014

HADOOP MAPREDUCE VS SPARK



Écosystème

Langages de programmation

- + La librairie *core* est écrit en Scala
- + APIs disponible pour Scala, Python, Java, R, .Net

Différences

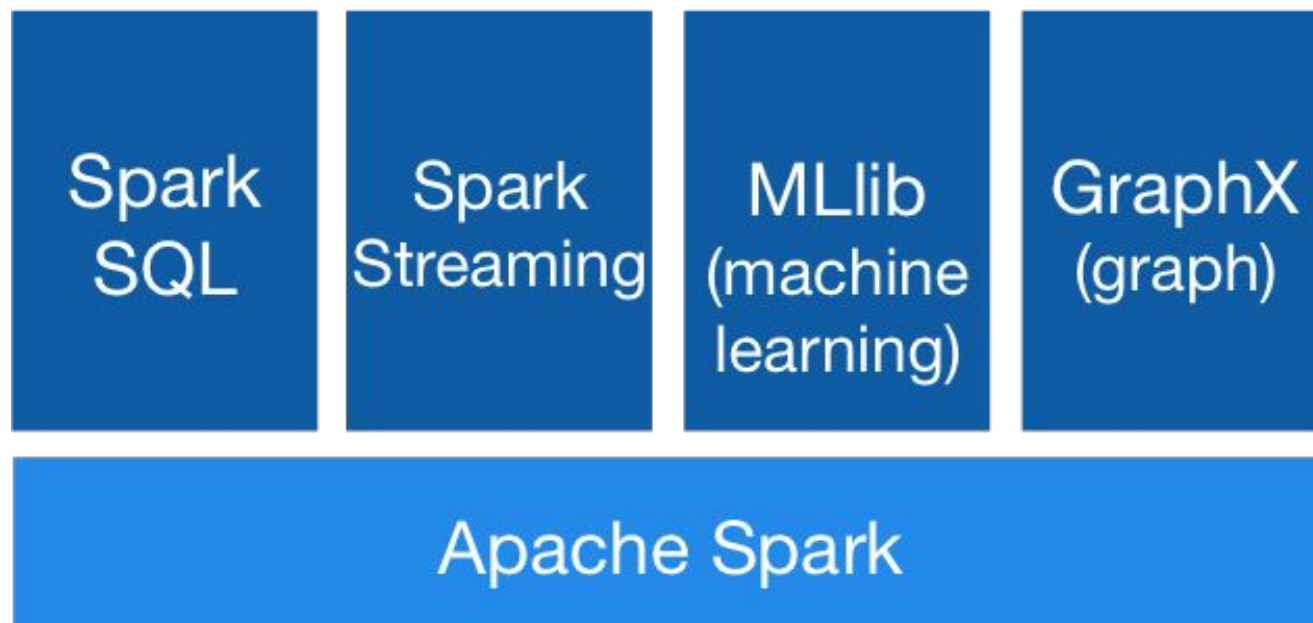
- + Scala et Java offre les meilleurs performances
- + Python offre la meilleur intégration avec les outils de data science
- + R et .Net sont moins populaires

Écosystème

- + Spark Core
- + Spark SQL
- + Spark Streaming
- + Spark ML
- + GraphX

Notebooks

- + Spark Shell
- + Apache Zeppelin
- + Jupyter
- + Databricks





Spark : les Structures de données

Trois niveaux

- + RDD : données non structurées
 - Structure historique
- + DataFrame : données structurées, sans types
 - Proche d'une table SQL
- + DataSet : données structurées, avec types
 - Permet de typer la structure de données

Spark : le RDD

| RDD : Resilient Distributed Dataset

- + Abstraction d'une collection distribuée et résiliente
- + Ne contient pas de données!
- + On peut appliquer des transformations sur cette structure
- + Structure de donnée immuable



Spark

Le Dataframe

Dataframe

| Dataframe :

- + Jeu de données organisées en colonnes nommées
- + Se rapproche le plus d'une table relationnelle traditionnelle
- + Très similaire au dataset, mais sans typage de données



Spark

Le Dataset

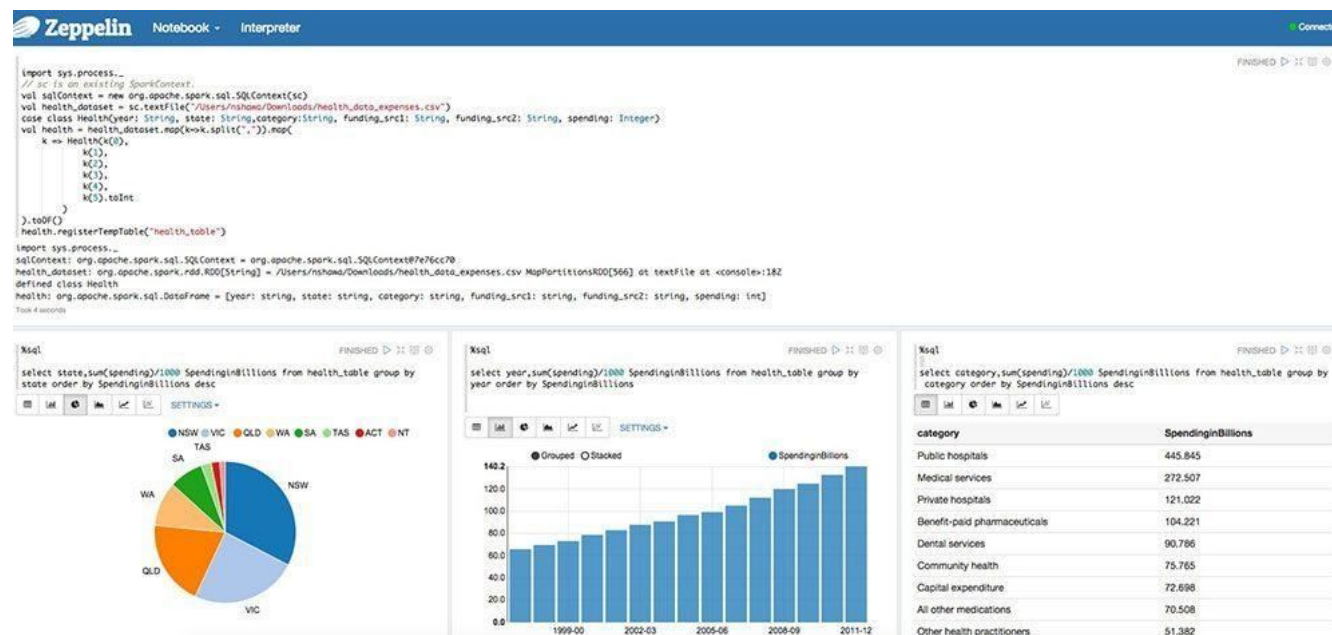
Dataset

Dataset:

- + Jeu de données organisées en collection d'objets typés
- + Permet toujours l'utilisation de fonction SQL
- + Permet à certaines erreurs de remonter de l'exécution à la compilation

Spark SQL

- + Pour les données structurées
- + Inférence du schéma
- + Moteur d'optimisation Catalyst
- + API Spark ou HiveQL





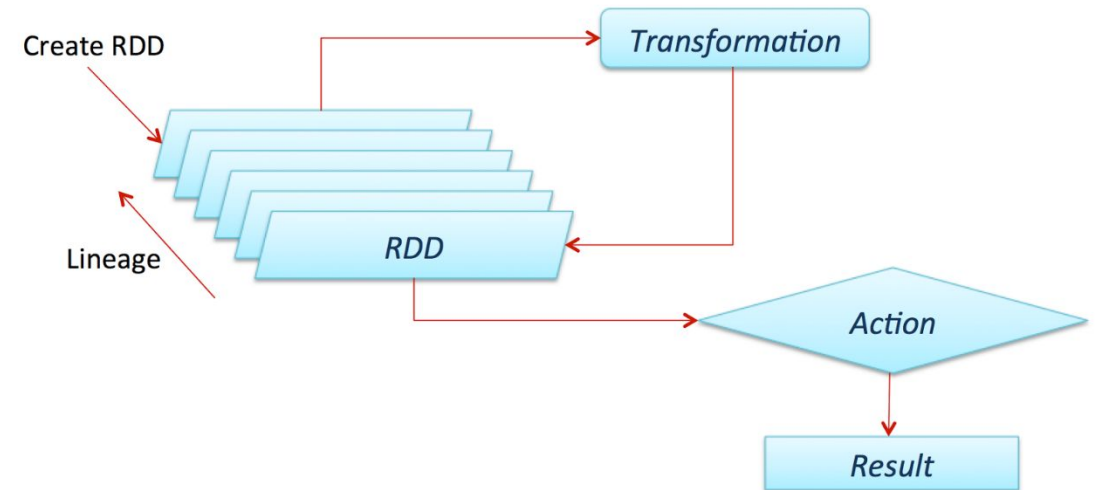
Spark : les commandes

Transformations

- + map() filter() groupByKey() join()...
- + Ne font rien!

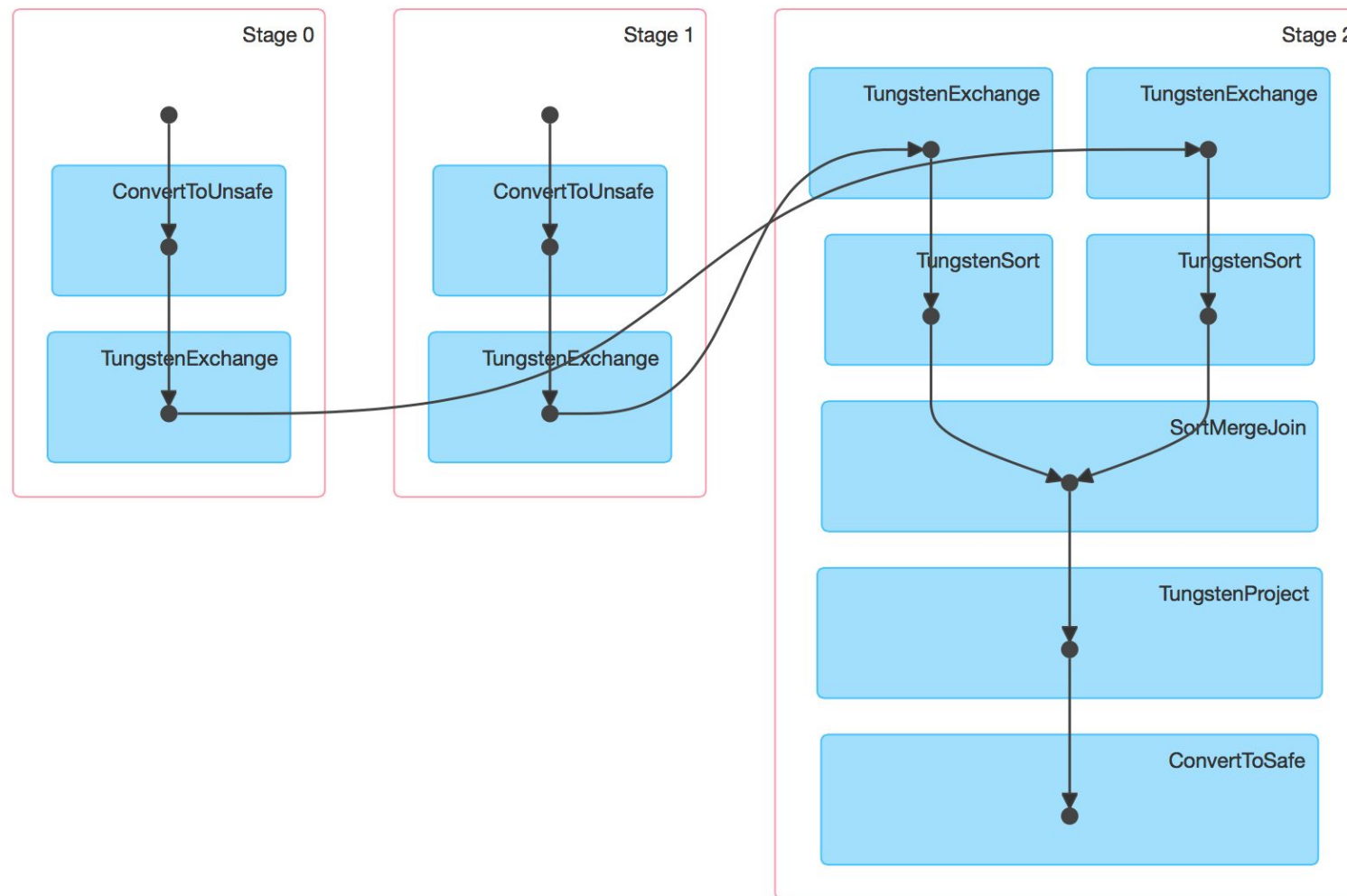
Actions

- + count() save() collect() foreach()
- + Déclenche les transformations



Spark : le DAG

(Directed Acyclic Graph)



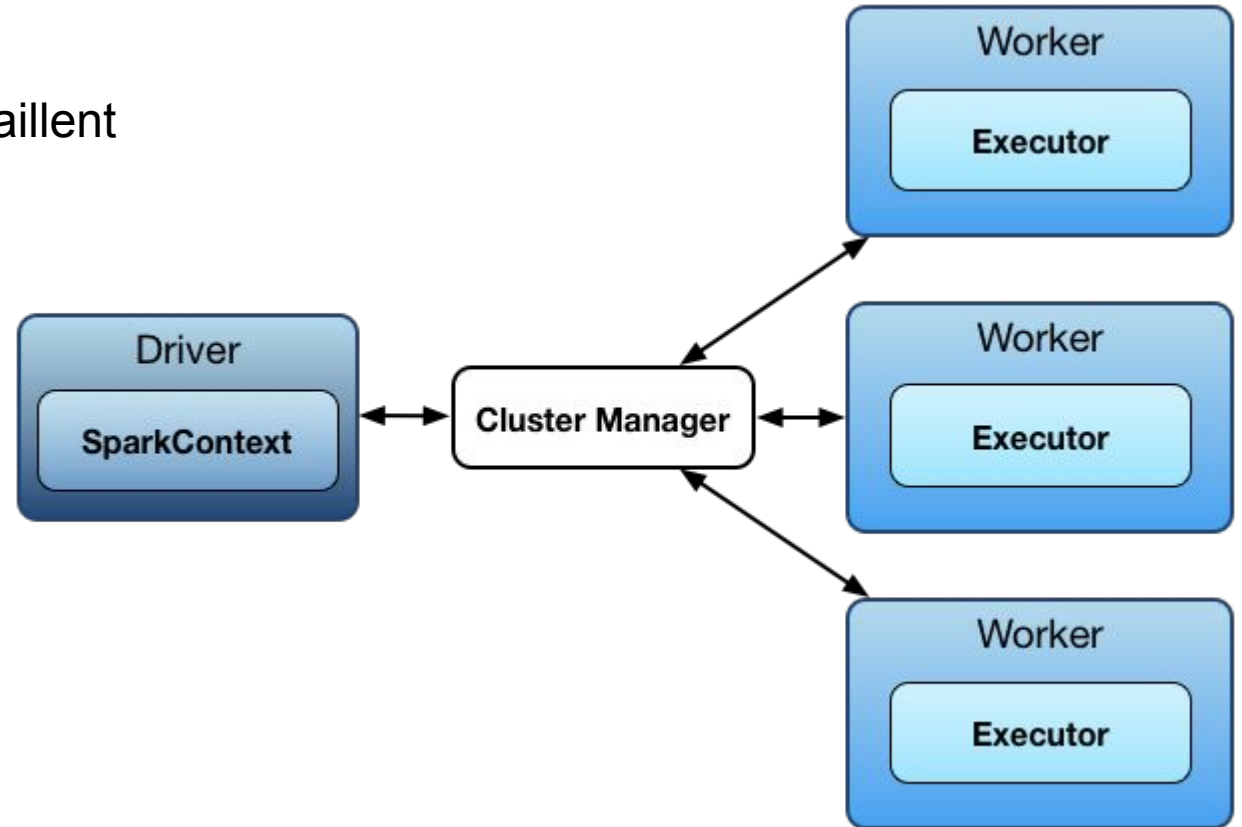


Spark

Comment Spark s'exécute?

Comment Spark s'exécute?

- + Un driver : le chef d'orchestre
- + Plusieurs exécuteurs : ceux qui travaillent
- + Un gestionnaire de ressources

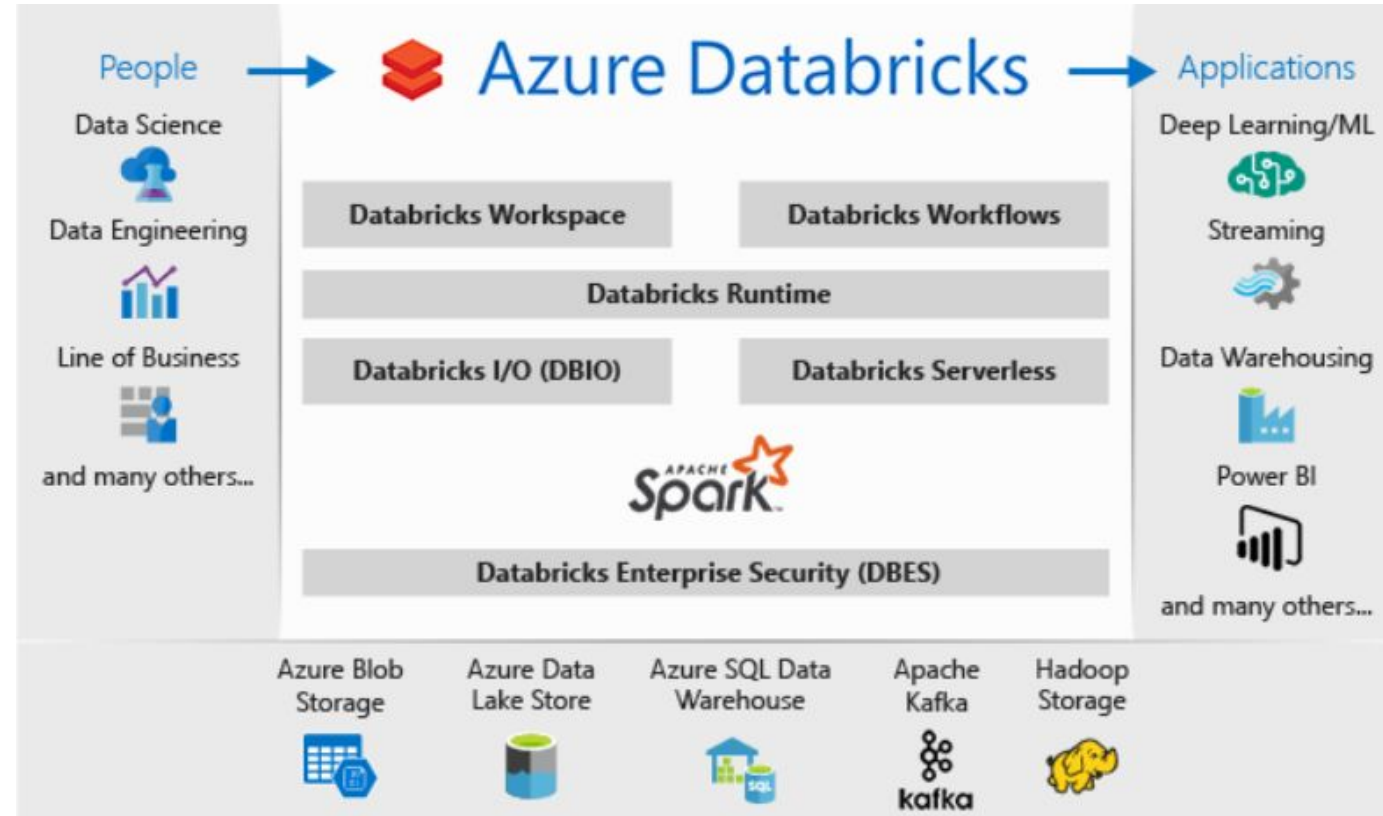




Spark

Spark & Azure

Azure Databricks Spark



Principaux avantages de Spark dans Azure

- + Facilité de déploiement
- + Élasticité
- + Intégration des notebooks





Lab

Big Data : Databricks + Spark

Lab

Présentation du sujet d'étude

Présentation du Lab

- + Création du workspace Databricks dans Azure
- + Création du cluster Spark dans Databricks

Visite guidée de Spark sur Databricks

- + Importation du lab dans l'environnement
- + Hello, world!

Analyse de données - Malware dataset

- + Analyse de jeux de données
- + Projection en SQL et Python

Visualisation

- + Intégration de Spark avec Power BI
- + Réalisation d'un rapport interactif



Téléchargement du Lab

Microsoft Azure

Télécharger le document suivant:

+ LabAgiledss-SparkDatabricks.dbc

Présent sur GitHub:

+ <https://github.com/guigirard/labDatabricksAzure>



Création du compte Azure

Microsoft Azure

Si vous avez une souscription active dans Azure

- + veuillez vous connecter à celle-ci et attendre

Si vous n'avez pas de souscription dans Azure

- + Créer une nouvelle adresse email sur <https://outlook.live.com/owa/>
- + Aller sur <https://signup.azure.com/> puis suivre les instruction
- + Azure vous demanderas une carte de crédit mais vous ne serez pas chargé



Installation de Databricks

Une fois connecté au portal azure

- + Entrer "Databricks" dans la barre de recherche
- + Puis cliquer sur "Azure Databricks"





vérifier que la souscription est activé



Installation de Databricks

Cliquer sur “Add”

The screenshot shows the Microsoft Azure portal interface for managing Azure Databricks subscriptions. The top navigation bar is blue with the Microsoft Azure logo. Below it, the breadcrumb trail shows 'Home > Azure Databricks'. The main heading is 'Azure Databricks' with 'AgileDSS' as a subtitle. A toolbar contains several icons: a plus sign for 'Add', a list icon for 'Edit columns', a circular arrow for 'Refresh', and a shield icon for 'As'. Below the toolbar, the text 'Subscriptions: All 2 selected' is displayed, followed by a link to 'Don't see a subscription'. A search bar labeled 'Filter by name...' is present. At the bottom, it says '2 items' and shows a table header with a checkbox, 'Name' with an up/down arrow, and 'Type' with an up/down arrow.

<input type="checkbox"/>	Name ↑↓	Type ↑↓
--------------------------	---------	---------



Installation de Databricks

Remplir les champs avec vos informations

Puis valider en cliquant sur “Create”

Attendre quelques minutes

Microsoft Azure

Home > Azure Databricks > Azure Databricks Service

Azure Databricks Service

Workspace name *

MyWorkSpacename

Subscription * ⓘ

Visual Studio Enterprise avec MSDN

Resource group * ⓘ

☒ Create new ☐ Use existing

MyResourceGroup

Location *

Canada Central

Pricing Tier * ⓘ

Trial (Premium - 14-Days Free DBUs)

Deploy Azure Databricks workspace in your own Virtual Network (VNet)


☐ Yes ☒ No



Installation de Databricks

Cliquer sur votre workspace

Puis sur “Launch Workspace”

 **MyWorkSpacename**
Azure Databricks Service

Overview

Activity log

Access control (IAM)

Tags

Settings

Virtual Network Peerings

Locks

Export template

Support + troubleshooting

New support request

Delete

Resource group (change) : [MyResourceGroup](#)

Subscription (change) : [Visual Studio Enterprise avec MSDN](#)


Subscription ID : [ae6e4f30-ecd0-4f79-86d5-781ac8351e5f](#)

Tags (change) : [Click here to add tags](#)

Managed Resource Group : [databricks-r](#)

URL : [https://cana](#)

Pricing Tier : [Trial \(Premiu](#)



Launch Workspace

Upgrade to Premium



Installation de Databricks

Microsoft Azure

Azure Databricks

Home

Workspace

Recents


Data

Clusters

Jobs

Search


Azure Databricks



Explore the Quickstart Tutorial


Spin up a cluster, run queries on preloaded data, and display results in 5 minutes.

Drop files or [click to browse](#)



Import & Explore Data








Quickly import data, preview its schema, create a table, and query it in a notebook.







Create a Blank Notebook

Create a notebook to start querying, visualizing, and modeling your data.





Common Tasks

-  New Notebook
-  Create Table
-  New Cluster
-  New Job
-  New MLflow Experiment New
-  Import Library
-  Read Documentation

Recents

-  2-01 Data Science Tools
-  0-00 Introduction
-  2-02 Spark MLlib
-  1-03 Managed Delta Lake

Documentation

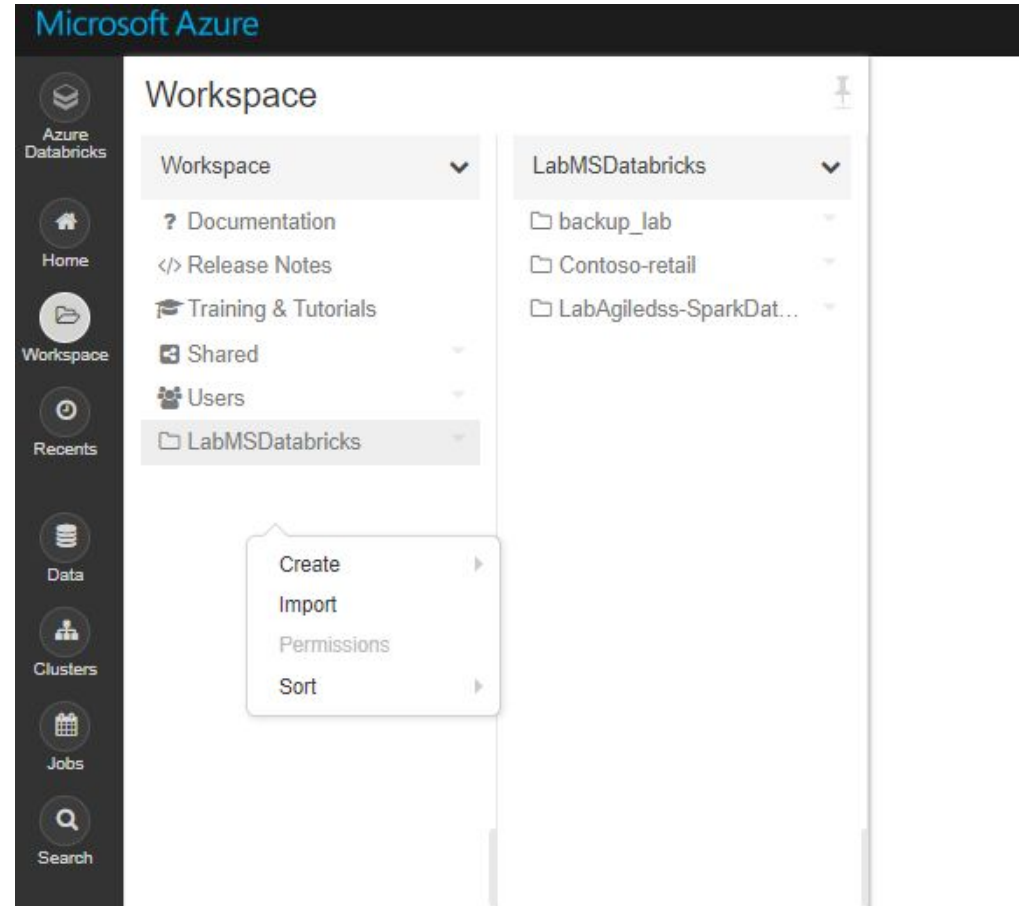
-  Databricks Guide
-  Getting Started
-  SparkR Overview
-  Importing Data



Installation de Databricks

Pour importer le lab

- + Cliquer sur "Workspace"
- + Puis Cliquez droit en dessous de "Users"
- + Puis "Import"





Installation de Databricks

Pour créer un cluster Spark

- + Cliquer sur “Clusters”
- + Puis sur “Create Cluster”
- + Entrer le nom du cluster
- + Sélectionner Runtime 6.2
- + Désélectionner “Enable autoscaling”
- + Réduire le nombre de “Workers” à 1

Microsoft Azure

Create Cluster

New Cluster

[Cancel](#) [Create Cluster](#) 1 Workers: 14.0 GB Memory, 4 Cores, 0.75 DBU
1 Driver: 14.0 GB Memory, 4 Cores, 0.75 DBU

Cluster Name
Mycluster

Cluster Mode ?
Standard

Pool ?
None

Databricks Runtime Version ? [Learn more](#)
Runtime: 6.2 (Scala 2.11, Spark 2.4.4)

New This Runtime version supports only Python 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

Workers
1

Driver Type
Same as worker 14.0 GB Memory, 4 Cores, 0.75 DBU

► Advanced Options



Références

Zoiner Tejada

| Mastering Azure Analytics: Architecting in the Cloud with Azure Data Lake, HDInsight, and Spark (ISBN: 9781491956656)

Bill Chambers, Matei Zaharia

| Spark: The Definitive Guide (ISBN: 9781491912218)

Denny Lee, Tomasz Drabas

| Learning PySpark (ISBN: 9781786463708)

Ritchie King, Nate Silver

| <https://projects.fivethirtyeight.com/flights/>

Lab :
À vous de jouer!



agileDSS : votre partenaire Big Data

Pour aller plus loin

Offre aux
participants



Offres Big Data

Exploitez tout le potentiel de vos données

Nos offres Big Data

Formation
écosystème Big
Data & Azure

Définition de **use
case & preuve de
concept**

Architecture Big
Data dans le
cloud / on prem

Développement
& mise en place
environnement
« **AI ready** »

**Visualisation &
Data Science** sur
des systèmes Big
Data

Merci

