* Définition de Spark
  + Définition
  + Utilité
  + Exemple d’utilisation
* Le système de stockage
  + RDD
  + Dataframe
  + Dataset
* Les fonctions map
* Zeppelin
* Debugging
* Optimisation
  + Fonctionnement du moteur Spark (stockage, calcul, …)
  + Partitionnement
  + Shuffling
  + Caching / Broadcasting
* Spark Stream
  + Spark Streaming
  + Spark Structured Streaming
  + Spark Structured Streaming avec Kafka

Programme cours OCTO : <https://www.octo.academy/fr/formation/144-data-science-niveau-avance>

Cours OpenClassRoom : [Réalisez des calculs distribués sur des données massives - OpenClassrooms](https://openclassrooms.com/fr/courses/4297166-realisez-des-calculs-distribues-sur-des-donnees-massives)

Avantage / Inconvénient : <https://thenewstack.io/the-good-bad-and-ugly-apache-spark-for-data-science-work/>

Définition de Spark : <https://mapr.com/blog/spark-101-what-it-what-it-does-and-why-it-matters/>

# Définition de Spark

## Loin

Définition de Spark

* Framework open source
* Calcul distribué
* Calcul en mémoire

Que fait Spark en pratique :

* Data processing, ETL (général)
* Streaming

Les API Spark

* SparkSQL (Data processing ETL)
* Streaming
* Machine Learning
* GraphX

Les distributions de Spark :

* HortonWorks
* DataBricks
* MapR

Langages:

* Scala
* Python
* R
* …

## Fonctionnement côté machine :

* RDD
* Disque, Cœur
* Lazy
* Lien RDD et fichier disque
* Driver, spark context, master/worker,
* Optimiseur de tâche + DAG (rapide)

# Pratique

## Rdd

Notebook !!!!

* Chargement des données
  + Créer une parallelize
  + Charger
    - depuis un csv
    - un Hive
* map(), flatMap, filter, distinct, collect, first, take, count, reduce saveAsText…

## Dataframe

* Chargement des données
* filter, where, groupby, join...
* withColumn
* udf

## Dataset

# Optimisation

## Fonctionnement côté machine (suite) : Optimiser

Définition de Job, Stage, Task

Shuffling, partition

Spark UI

<https://dzone.com/articles/apache-spark-on-yarn-performance-and-bottlenecks>

# Missing

* Type de donnée
* Configuration pour créer une application

# Streaming

* Spark Streaming (Spark 1)
* Structure Streaming (Spark 2)