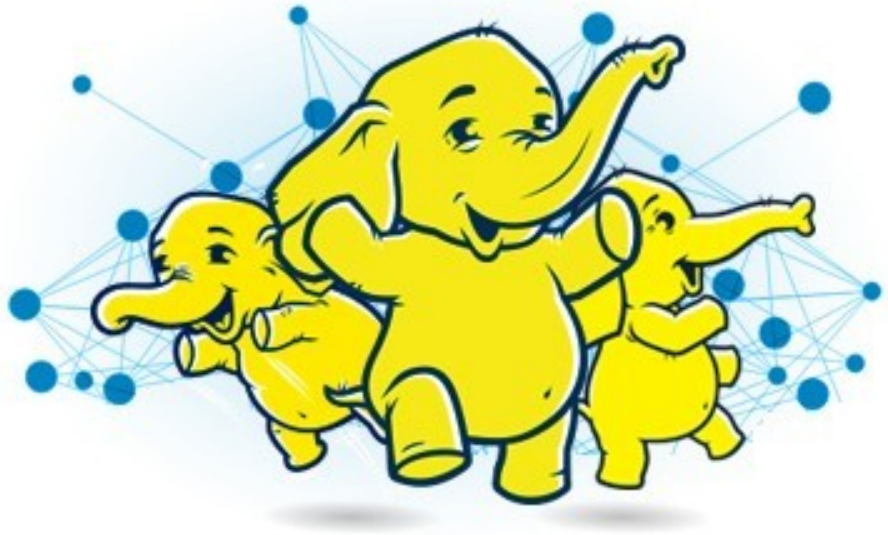


Big Data, Hadoop &Spark

CCT SIL

Traitement et Manipulation de la donnée à l'aide des technologies Big Data

30 Juin 2017

SOMMAIRE

HISTORIQUE DE SPARK

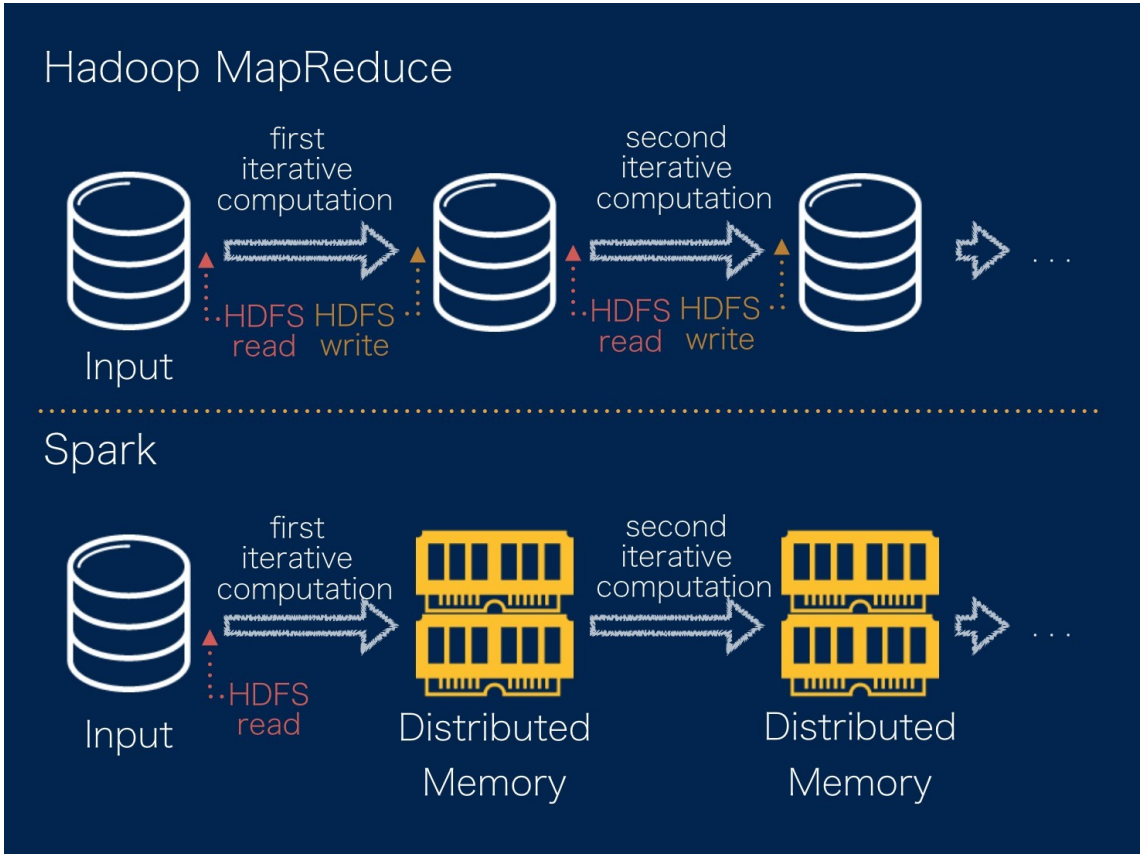
- Développé par AMPLab, de l'Université UC Berkeley, en 2009

- Passé OpenSource en 2010 sous forme de projet Apache
  - Release 1.0 –Mai 2014
  - Release 2.0 –mi 2016

- Juin 2013 : Top Apache Project (Apache Spark)

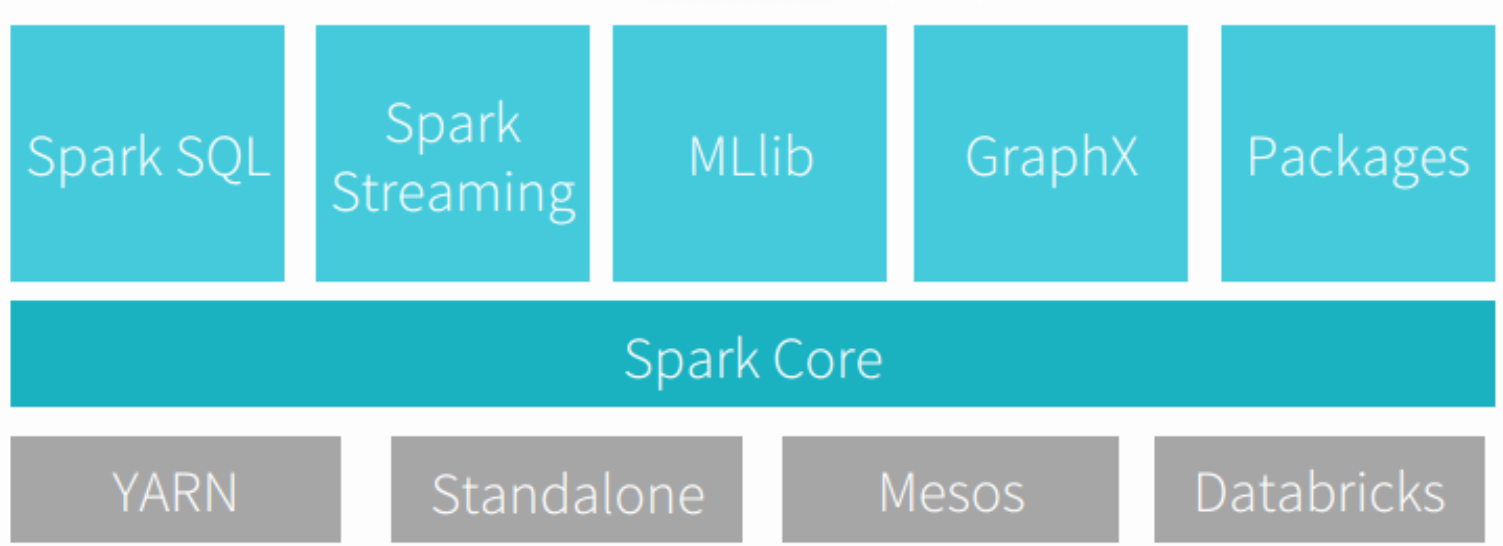
- Extension du modèle MapReduce (plus performant, in-memory)

SPARK VS MAP-REDUCE



Alternative in-memory plus rapide que MapReduce de Hadoop
(100 x plus rapide en mémoire & 10 x plus vite sur disque)

INTRODUCTION À SPARK



kubernetes

Scala Java Programming python R

Framework généraliste / API en Scala, Java, Python et R
Ecosystème riche (SparkSQL, Spark Streaming, MLlib, GraphX)

RDD
Resilient Distributed DataSet

- Collection d'objets distribués
- Données non structurées en entrée
- Structure de donnée Immutable
- In memory par défaut
- Manipulés par des opérateurs : transformations / actions

Transformations

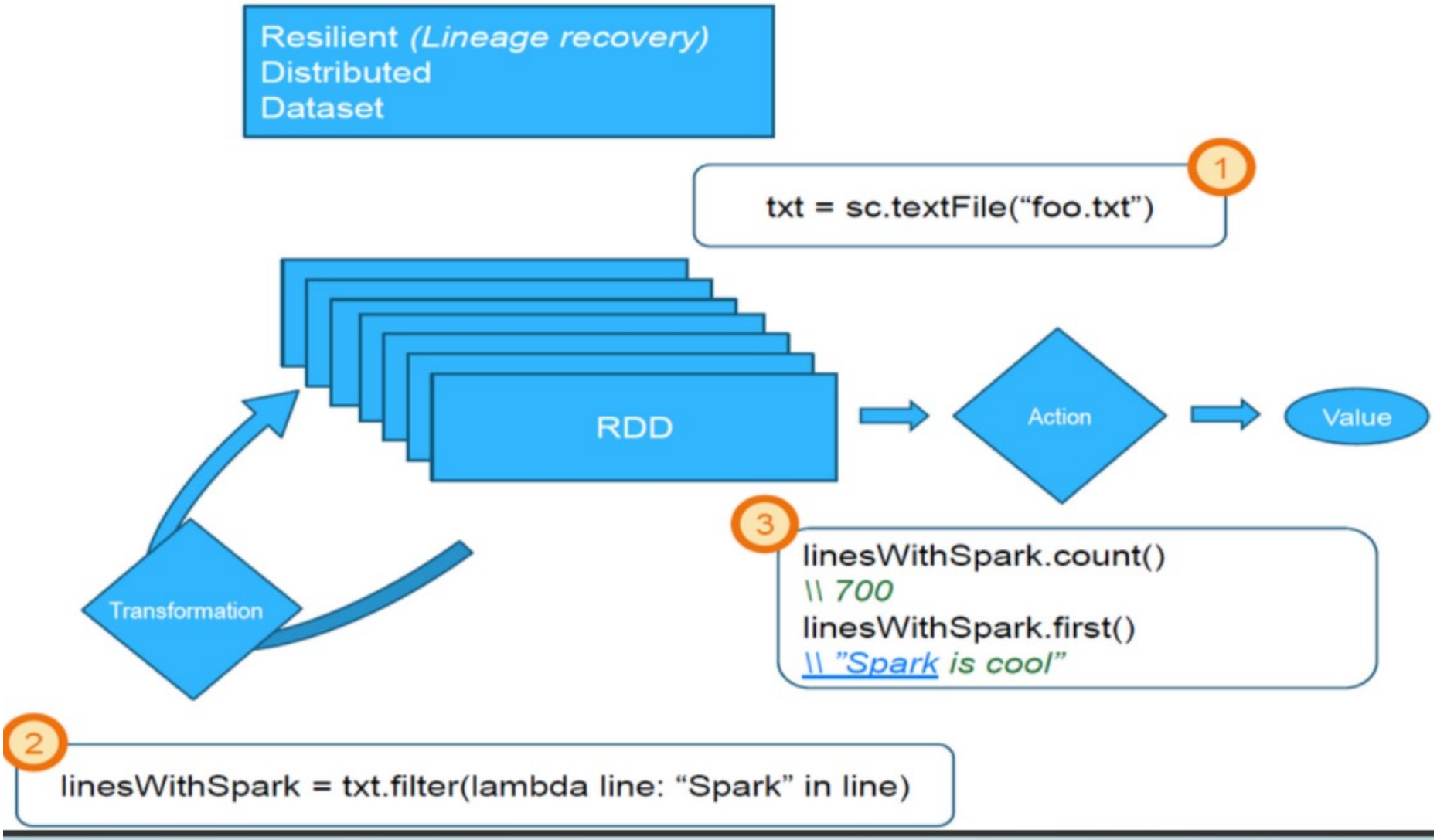
- Creation d'un jeu de donnée
- Lazy par nature. N'est exécuté que lorsque d'une action est effectuée
- Exemple :
  - Map(func)
  - Filter(func)
  - Distinct()

Actions

- Retourne au driver programme une valeur ou exporte les données vers un systeme de stockage
- Exemple:
  - Count()
  - Reduce(funct)
  - Collect
  - Take()

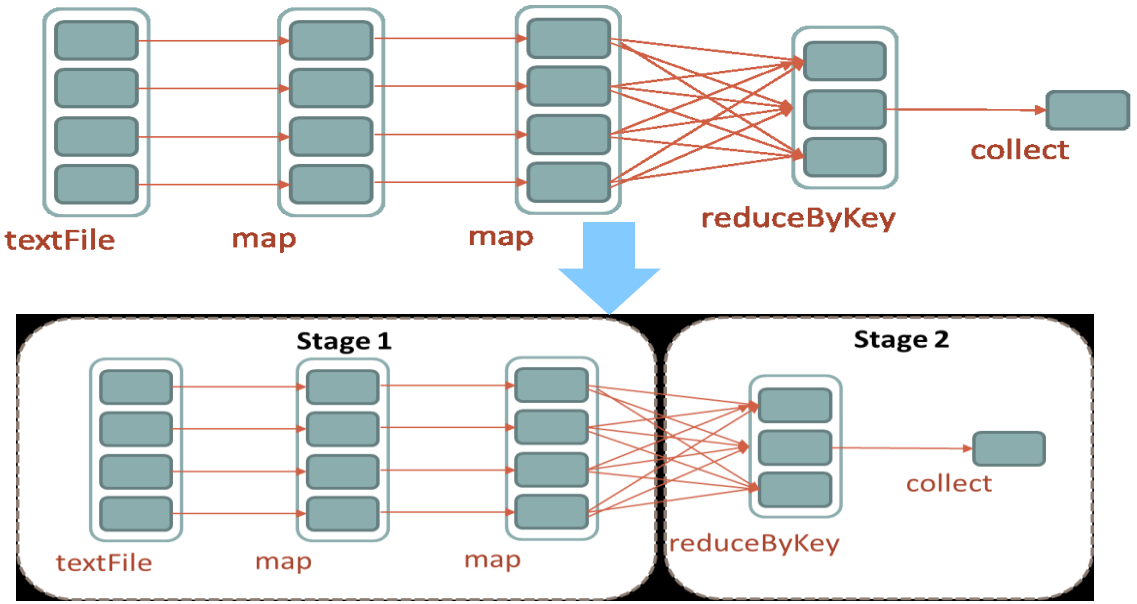
- Tolérants aux pannes : un RDD sait comment recréer et recalculer son ensemble de données

Transformations et Actions



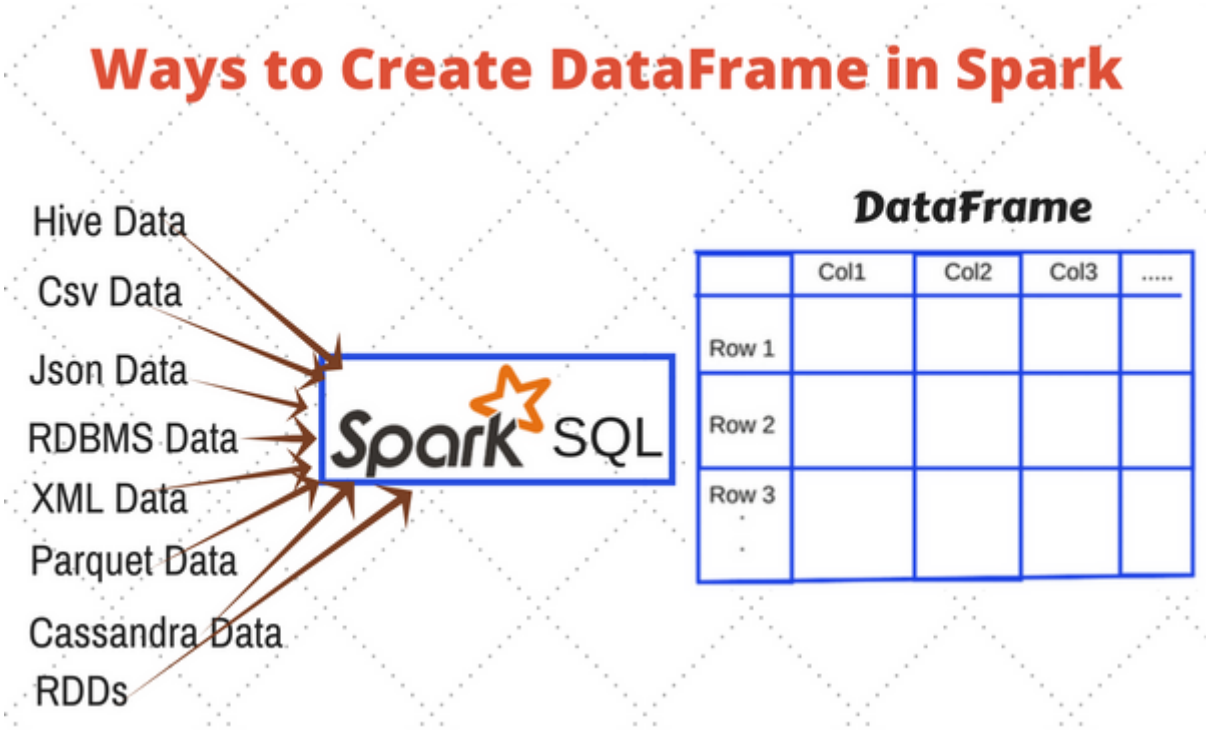
PLAN D'EXÉCUTION DE SPARK

- Les taches sont les unités fondamentales d'exécution
- Les stages
  - ensemble de taches qui peuvent être exécutés en parallèles
  - ensemble de séquences de RDD sans Shuffle (tri par clé)
- Le shuffle est appliqué entre les stages



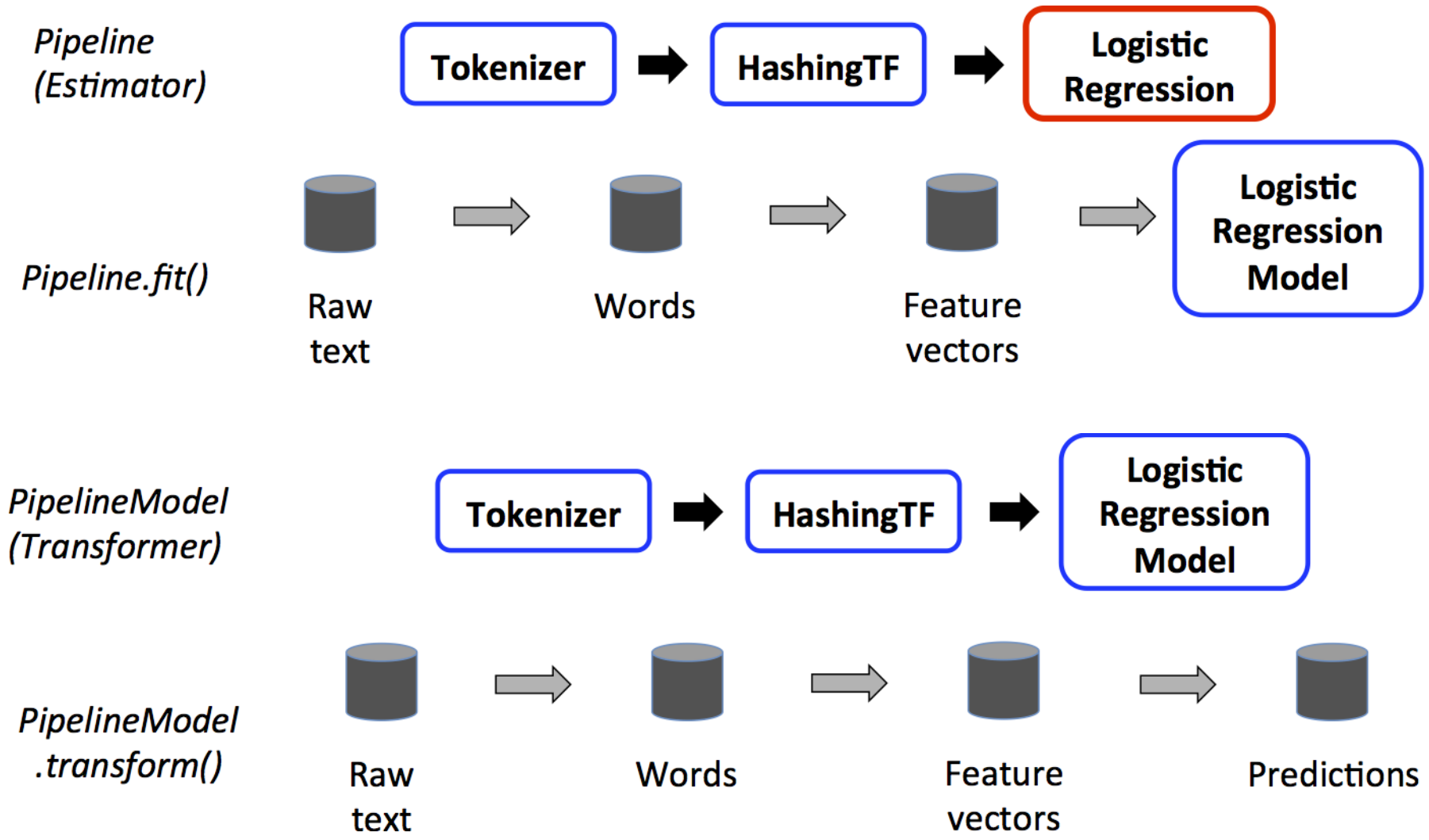
Spark SQL

- Manipulation de Dataframes (proches des pandas Dataframes)
- Données tructurées en entrée (base de données, fichiers CSV ...)
- Opération type SQL (filter, join, group)
- Hérite des propriétés RDD
- Catalyst Optimizer : optimisation de requêtes.

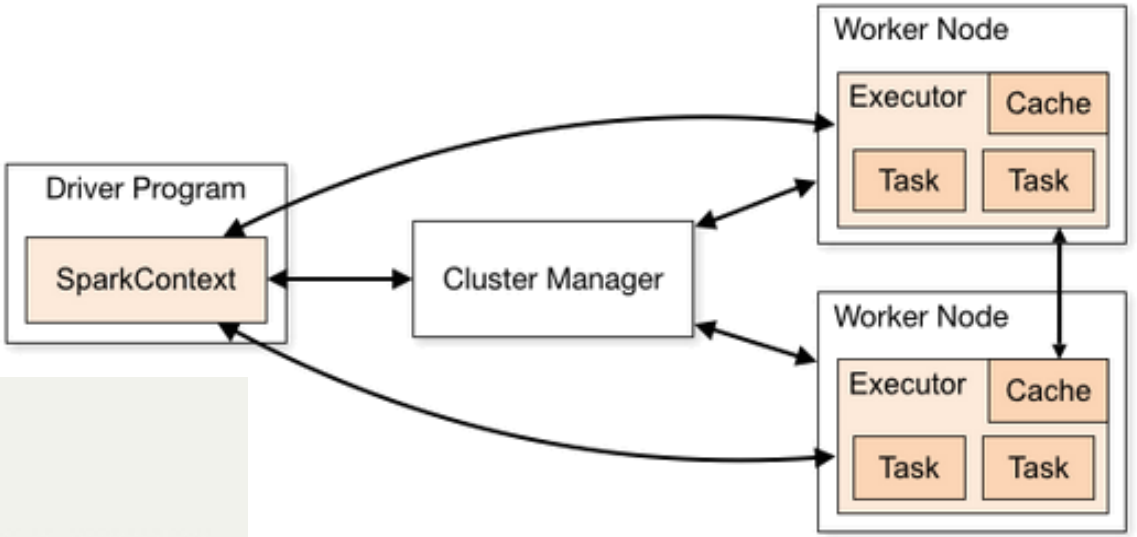


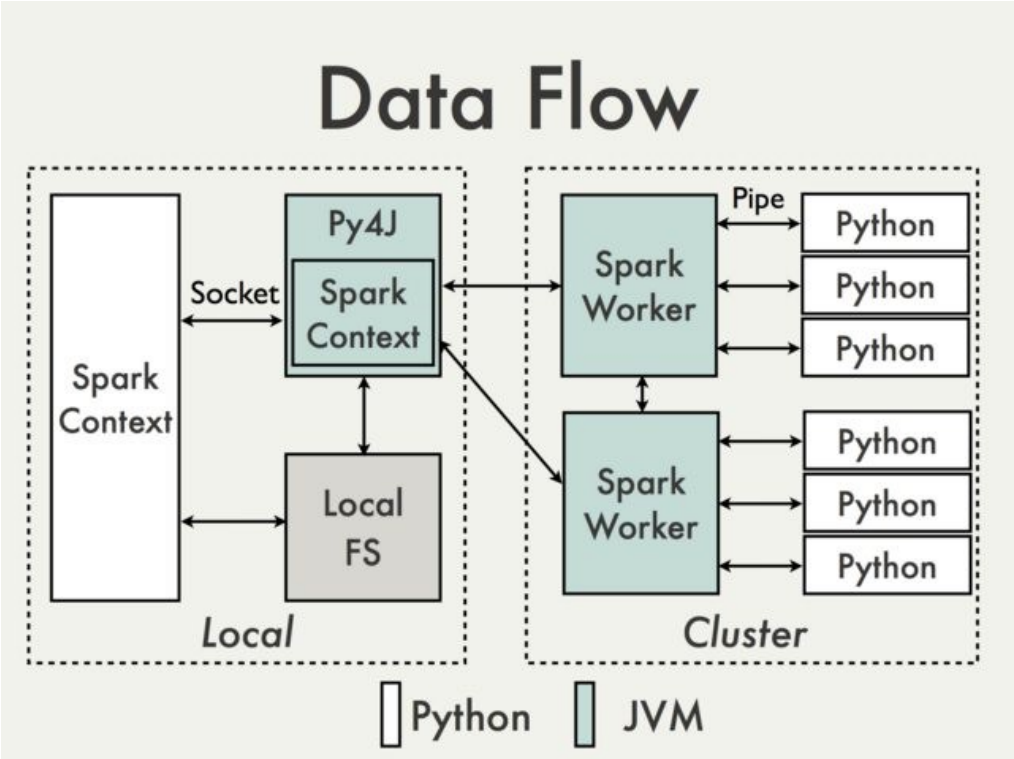
Spark STREAMING

Spark MLLIb


Pipeline
(Estimator)
Tokenizer
HashingTF
Logistic
Regression
Pipeline.fit()
Raw
text
Words
Feature
vectors
Logistic
Regression
Model
PipelineModel
(Transformer)
Tokenizer
HashingTF
Logistic
Regression
Model
PipelineModel
.transform()
Raw
text
Words
Feature
vectors
Predictions

ÉXECUTION DE SPARK





Dashboard Spark

Spark | Stages | Storage | Environment | Executors | **Spark UI Tester** application UI

Spark Stages

Total Duration: 20.3 s
Scheduling Mode: FIFO
Active Stages: 1
Completed Stages: 4
Failed Stages: 1

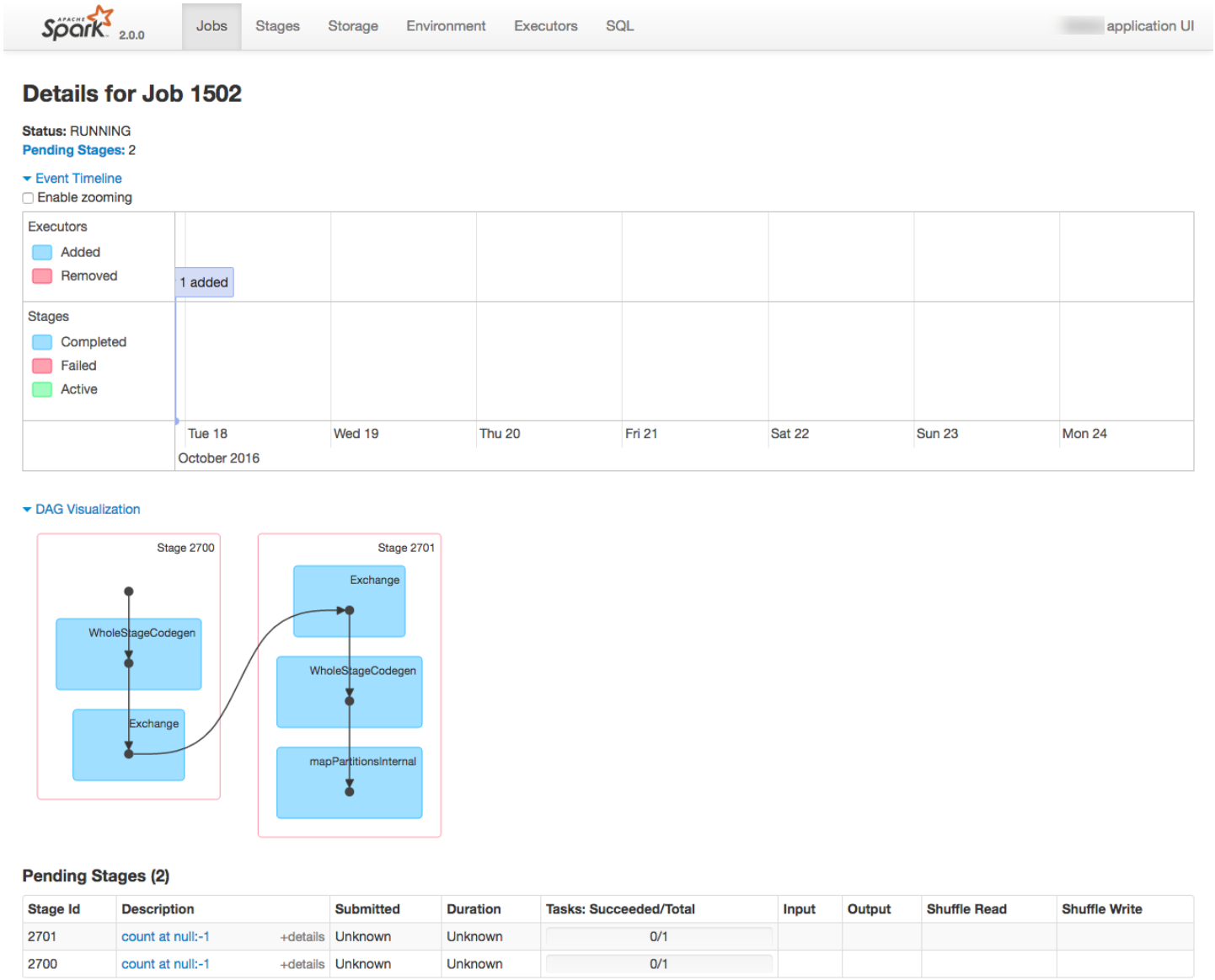
Active Stages (1)

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Shuffle Read	Shuffle Write
5	*Partially failed phase* count at UIWorkloadGenerator.scala:72	2013/09/25 13:02:09	64 ms	15/100 (3 failed)		

Completed Stages (4)

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Shuffle Read	Shuffle Write
2	*Single Shuffle* count at UIWorkloadGenerator.scala:63	2013/09/25 13:02:00	1.8 s	100/100		
3	*Single Shuffle* reduceByKey at UIWorkloadGenerator.scala:63	2013/09/25 13:01:59	1.4 s	100/100		151.2 KB
1	*Cache and Count*	2013/09/25 13:01:54	1.0 s	100/100		

Dashboard Spark



Pending Stages (2)

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
2701	count at null:-1 +details	Unknown	Unknown	0/1				
2700	count at null:-1 +details	Unknown	Unknown	0/1				

Merci,

Des Questions ?