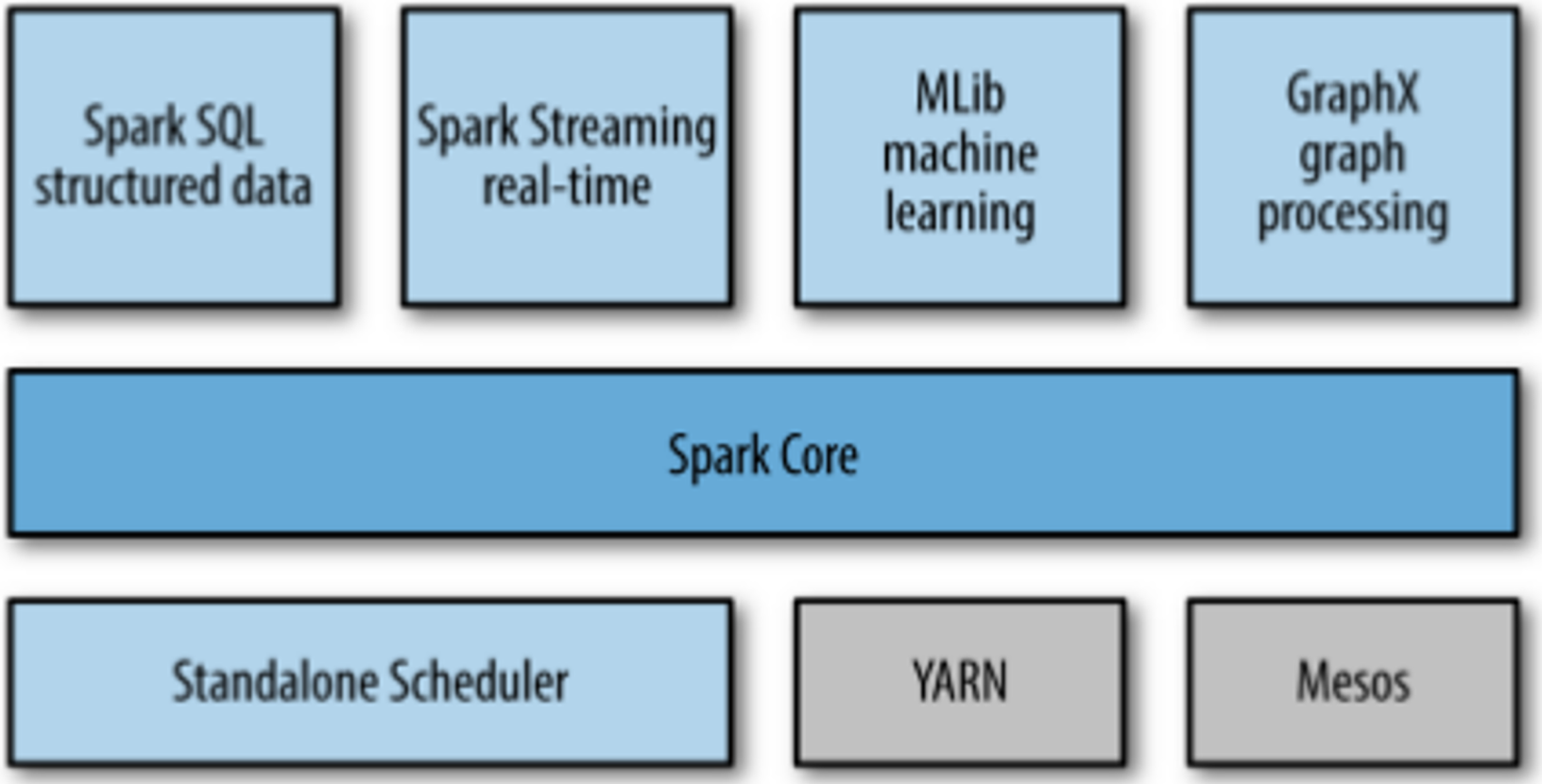
Spark Définition

- Est un framework de traitements Big Data open source construit pour effectuer des analyses sophistiquées et conçu pour la rapidité et la facilité d'utilisation
- Est une ré-ecriture du paradigme map-reduce en mémoire
- Spark peut être jusqu'à 100 fois plus rapide que Hadoop
- Permet de lancer des requêtes en mode intéractif
- Spark offre plusieurs API : Java, Python, Scala, R et SQL
- Spark est basé sur la notion de RDD « Resilient distributed datasets » et DataFrame

Spark Architecture



Spark écosystème

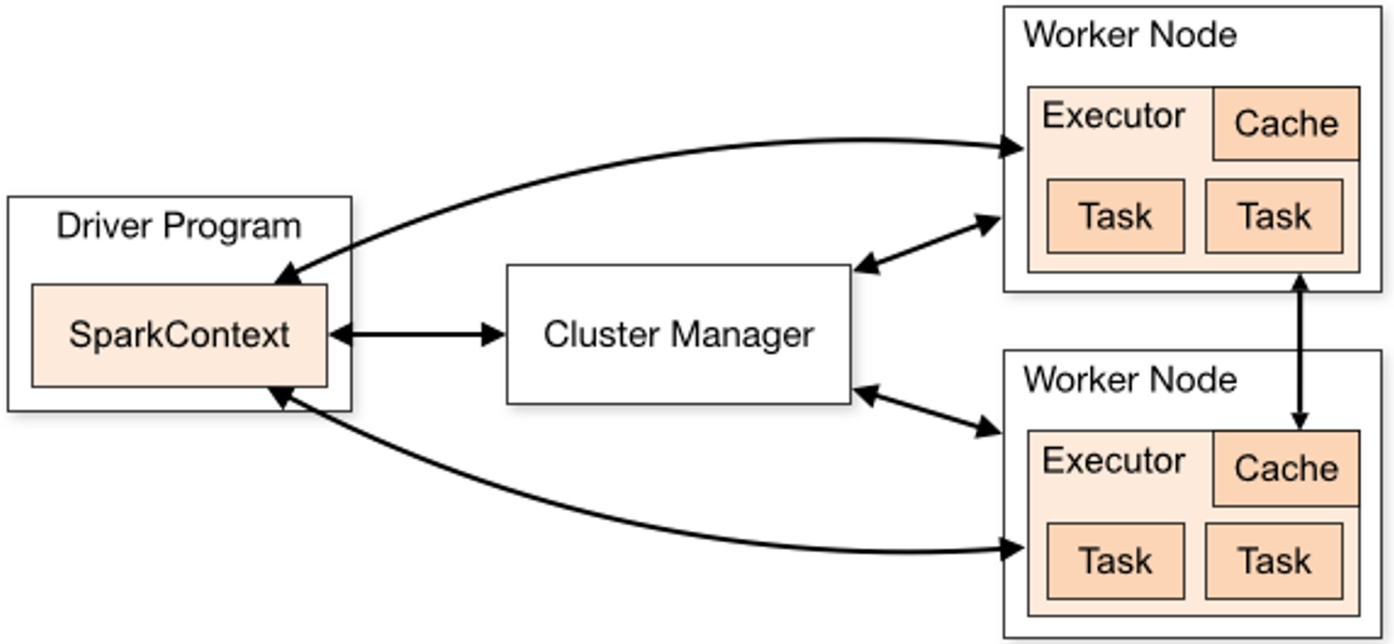
Spark Streaming : peut être utilisé pour traitement temps-réel des données en flux. Il s'appuie sur un mode de traitement en "micro batch" et utilise pour les données temps-réel DStream, c'est-à-dire une série de RDD (Resilient Distributed Dataset).

Spark SQL : permet d'exposer les jeux de données Spark via API JDBC et d'exécuter des requêtes de type SQL en utilisant les outils BI et de visualisation traditionnels. Spark SQL permet d'extraire, transformer et charger des données sous différents formats (JSON, Parquet, base de données) et les exposer pour des requêtes ad-hoc.

Spark MLlib : est une librarie de machine learning qui contient tous les algorithmes et utilitaires d'apprentissage classiques, comme la classification, la régression, le clustering..

Spark GraphX : est une API pour les traitements de graphes et de parallélisation de graphes.

Spark Workflow



Un programme spark est composé de deux applications :

- Un driver : tourne sur le master
- Des workers : ils tournent sur les nœuds ‘slaves’ du cluster

Spark RDD

- Un RDD est une collection partitionnée d'enregistrements en lecture seule qui ne peut être créée que par des opérations déterministes :
  - soit à partir de données présentes dans un stockage stable,
  - soit à partir d'autres RDDs.
- Ces opérations sont appelées **transformations** pour les différentier des autres opérations:
  - map, filter, flatMap…
- Les RDDs n'ont pas besoin d'être matérialisés tout du long puisqu'un RDD dispose de suffisamment d'informations sur la façon dont il a été produit à partir d'un autre ensemble de données pour pouvoir être recalculé. Ainsi, un programme ne peut faire référence à un RDD s'il n'est pas capable de le reconstruire suite à une panne.

Spark RDD

- Les RDD supportent deux types d’opérations

- Opérations de transformation Il crée un nouveau RDD Spark à partir de l'existant. De plus, il transmet l'ensemble de données à la fonction et retourne un nouvel ensemble de données.
- Opérations d'action Dans Apache Spark, Action renvoie le résultat final au driver

https://spark.apache.org/docs/latest/rdd-programming-guide.html

Spark RDD Workflow

Rappel fonction lambda

- Petite fonctions non associé à un nom

```
lambda a , b : a + b
```

Cette fonction renvoie la somme de deux arguments

- Peuvent être utilisable partout où il y’ a des objets de type fonction
- Ne peut calculer qu’une seule expression

Spark RDD Input:Outpput

Function name	Purpose	Example
sc.parallelize	Transform a python list to an RDD	list = [1,2,3,4] rdd = sc.parallelize(list)
sc.textFile	Create an RDD from a text file	file_name = /documents/book.txt rdd = sc.textFile(file_name)
saveAsTextFile	Save the RDD as a text file	file_name = /documents/book.txt rdd.saveAsTextFile(file_name)

Spark RDD Transformations

Function name	Purpose	Example	Result
map	Apply a function to each element in the RDD and return an RDD of the result	rdd = [1,2,3,4] rdd.map(lambda x: x +1)	[2,3,4,5]
flatMap	Apply a function to each element in the RDD and return an RDD of the contents of the iterators returned	rdd = [[1,2],[3,4]] rdd.flatMap(lambda x: x +1)	[2,3,4,5]
filter	Retrun an RDD consitent of only elements that pass the condition	rdd = [1,2,3,4] rdd.filter(lambda x: x != 1)	[2,3,4]
distinct	Remove duplicates	rdd = [1,2,2,3] rdd.distinct()	[1,2,3]
sortBy	Sort the element of an RDD	rdd = [3,2,4,1] rdd.sort()	[1,2,3,4]
reduceBykey	Apply a transformation and return a new RDD based on the reduce function	rdd = [(Paris, 13), (Paris, 24), (Lyon,12), (Lille, 40)] rdd.reduceByKey(lambda a, b: a + b)	[(Paris, 37), (Lyon, 12), (Lille, 40)]

Spark RDD Actions

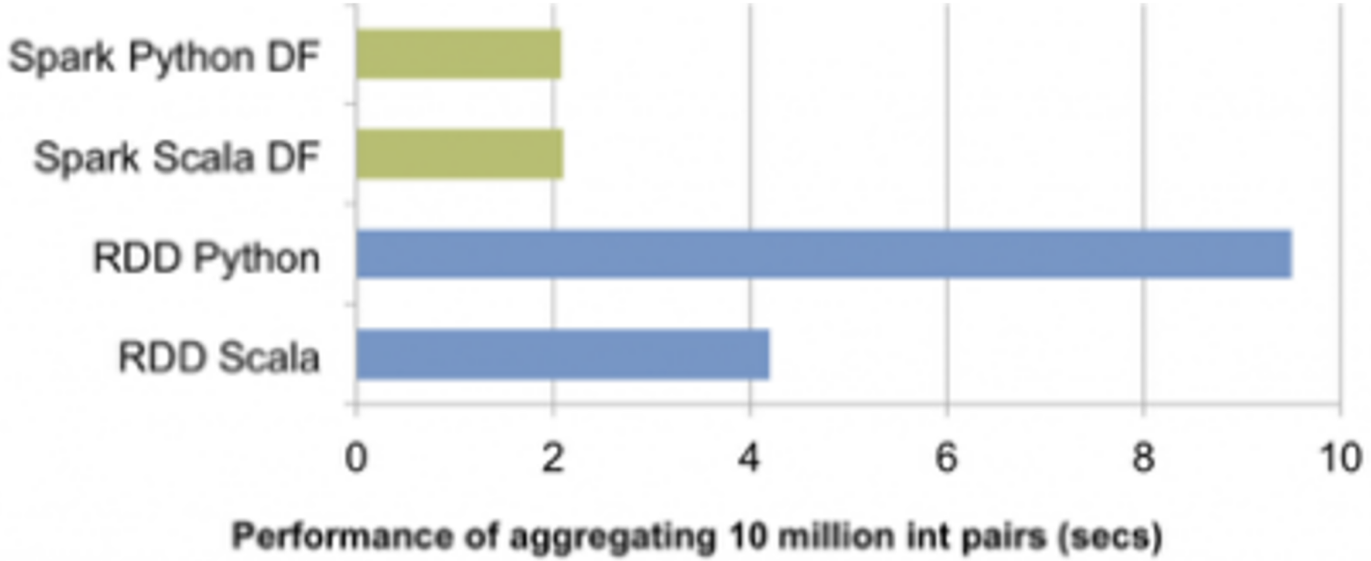
Function name	Purpose	Example	Result
collect	Return all the elements of the dataset as an array at the driver program	rdd = [1,2,3,4] rdd.collect()	[1,2,3,4]
take	Return an array with the first *n* elements of the dataset.	rdd = [1,2,3,4] rdd.take(2)	[1,2]
first	Return the first element of the dataset (similar to take(1)).	rdd = [1,2,3,4] rdd.first()	1
count	Return the number of elements in the dataset.	rdd = [1,2,3,4] rdd.count()	4
reduce	Aggregate the elements of the dataset using a function *func* (which takes two arguments and returns one)	rdd = [1,2,3,4] rdd.reduce(lambda a, b: a + b)	10

Spark Dataframe

- Un DataFrame est une abstraction de programmation dans le module Spark SQL.
- Les DataFrames ressemblent à des tables de bases de données relationnelles ou à des feuilles de calcul Excel avec des en-têtes: les données résident dans des lignes et des colonnes de différents types de données.
- Les dataframes sont basés sur la notion du RDD en rajoutant:
  - Une structure de données
  - Un typage

Spark Dataframe vs RDD

Faster than RDDs



Benefits from Catalyst optimizer