Introduction à Apache Spark (PySpark)

Principes de bases

par **Fabien Barbaud** - **@BarbaudFabien**

Apache Spark

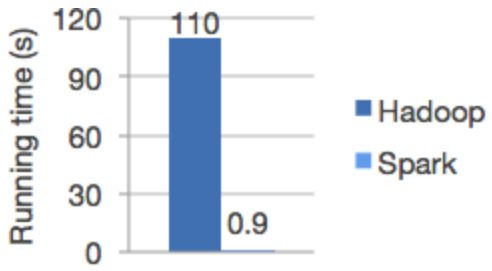
En résumé

Spark (ou **Apache Spark**) est un **framework** open source de **calcul distribué**. Il s'agit d'un ensemble d'outils et de composants logiciels structurés selon une architecture définie. Développé à **l'université de Californie à Berkeley** par AMPLab3, Spark est aujourd'hui un projet de la fondation **Apache**. Ce produit est un cadre applicatif de traitements **big data** pour effectuer des **analyses complexes à grande échelle**.
Wikipedia

Apache Spark

Rapidité

Régression logique sur Hadoop VS Spark



100x plus rapide

Apache Spark

Simple

Python

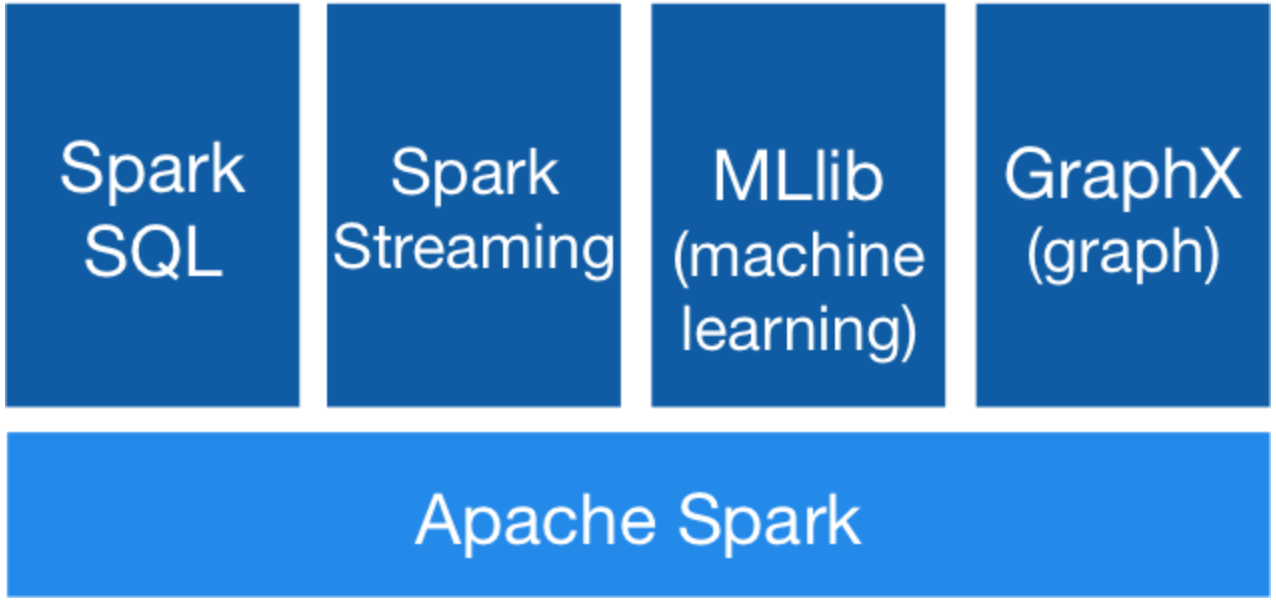
```
df = spark.read.json("logs.json")
df.where("age > 21")
  .select("name.first").show()
```

Pouvoir rapidement et simplement déployer une application parallélisée de traitement dans les langages Scala, Python, R.

Apache Spark

Généraliste

Stack Spark



Combiner du SQL, du Streaming, du machine learning, ... en une seule application avec Spark

Apache Spark

Resilient Distributed Datasets (RDDs)

le RDD est une **collection** d'éléments **partitionnés** et **répartis** entre les nœuds du cluster et **accessible uniquement** en **lecture-seule**.

RDD

Apache Spark

Resilient Distributed Datasets (RDDs)

Paralléliser une collection

```
data = [1, 2, 3, 4, 5]
distData = sc.parallelize(data)
```

RDD

Apache Spark

Resilient Distributed Datasets (RDDs)

Charger de la donnée externe

```
distFile = sc.textFile("data.txt")
```

Le fichier peut être en local mais Spark support les systèmes de fichier distribué : *hdfs://*, *s3a://*, ...

RDD

Apache Spark

Resilient Distributed Datasets (RDDs)

Sauvegarder un RDD

```
sc.saveAsTextFile("data.txt")
```

RDD

Apache Spark

Resilient Distributed Datasets (RDDs)

Exemple simple

```
lines = sc.textFile("data.txt")
lineLengths = lines.map(lambda s: len(s))
totalLength = lineLengths.reduce(lambda a, b: a + b)
```

RDD

Apache Spark

Resilient Distributed Datasets (RDDs)

Transformations

- map
- filter
- groupByKey
- reduceByKey
- join
- ...

RDD

Apache Spark

Resilient Distributed Datasets (RDDs)

Actions

- reduce
- collect
- count
- take
- ...

RDD

Apache Spark

Resilient Distributed Datasets (RDDs)

Exemple Wordcount

```
words = sc.textFile("/zeppelin/files/wordcount/conseil-tenu-par-les-rats.txt").flatMap(lambda line: line.split(" "))
wordCounts = words.map(lambda word: (word, 1)).reduceByKey(lambda a,b:a +b)
```

RDD

Apache Spark

Resilient Distributed Datasets (RDDs)

Tester avec Zeppelin

```
$ git clone https://github.com/fabienbarbaud/apache-zeppelin.git
$ cd apache-zeppelin
$ docker-compose up -d
```

http://localhost:8080/#/notebook/2FUS99C8T

Apache Spark

DataFrames

Un DataFrame est une **collection distribuée** de données organisées en **colonnes nommées**. Il est conceptuellement équivalent à une **table dans une base de données relationnelle**.

DataFrames

Apache Spark

DataFrames

Exemple simple

```
df = spark.read.json("/zeppelin/files/dataframe/MOCK_DATA.json")
df.filter(df['gender'] == "Female")
```

DataFrames

Apache Spark

DataFrames

Aussi en CSV

```
df = spark.read.option("header",True).csv("/zeppelin/files/dataframe/MOCK_DATA.csv")
df.filter(df['gender'] == "Male")
```

DataFrames

Apache Spark

DataFrames

Avec un groupement

```
df.groupBy("gender").count()
```

DataFrames

Apache Spark

DataFrames

Mais aussi en SQL

```
df = spark.read.option("header",True).csv("/zeppelin/files/dataframe/MOCK_DATA.csv")
df.createOrReplaceTempView("people")
sqlDF = spark.sql("SELECT * FROM people")
```

DataFrames

Apache Spark

DataFrames

Tester avec Zeppelin

http://localhost:8080/#/notebook/2FUAYT7SC

Apache Spark

DataFrames

Exemple Wordcount

```
from pyspark.sql.functions import split, explode

df = spark.read.text("/zeppelin/files/wordcount/conseil-tenu-par-les-rats.txt")
df.select(explode(split(df.value, '\s+')).alias('word')).groupBy("word").count()
```

http://localhost:8080/#/notebook/2FUZMQVNB

Apache Zeppelin