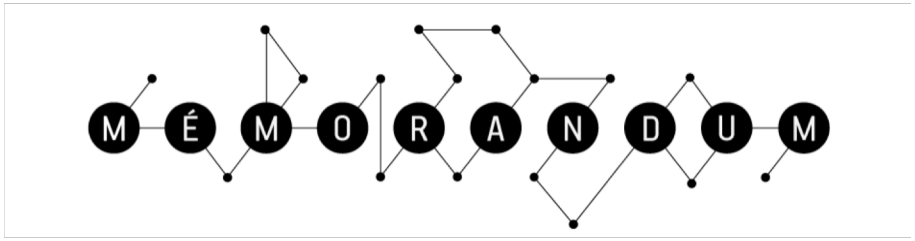
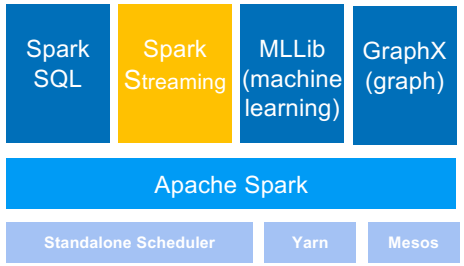


Spark Streaming en Python

Romain Jouin
Romain.jouin@memorandum.pro **06 52 86 87 30**

Streaming et Spark



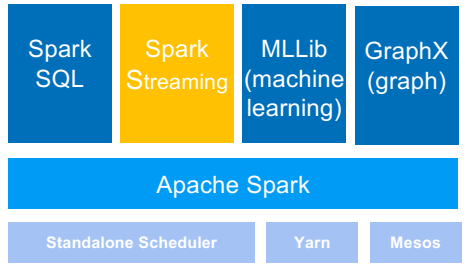
Challenges – Réponse de Spark

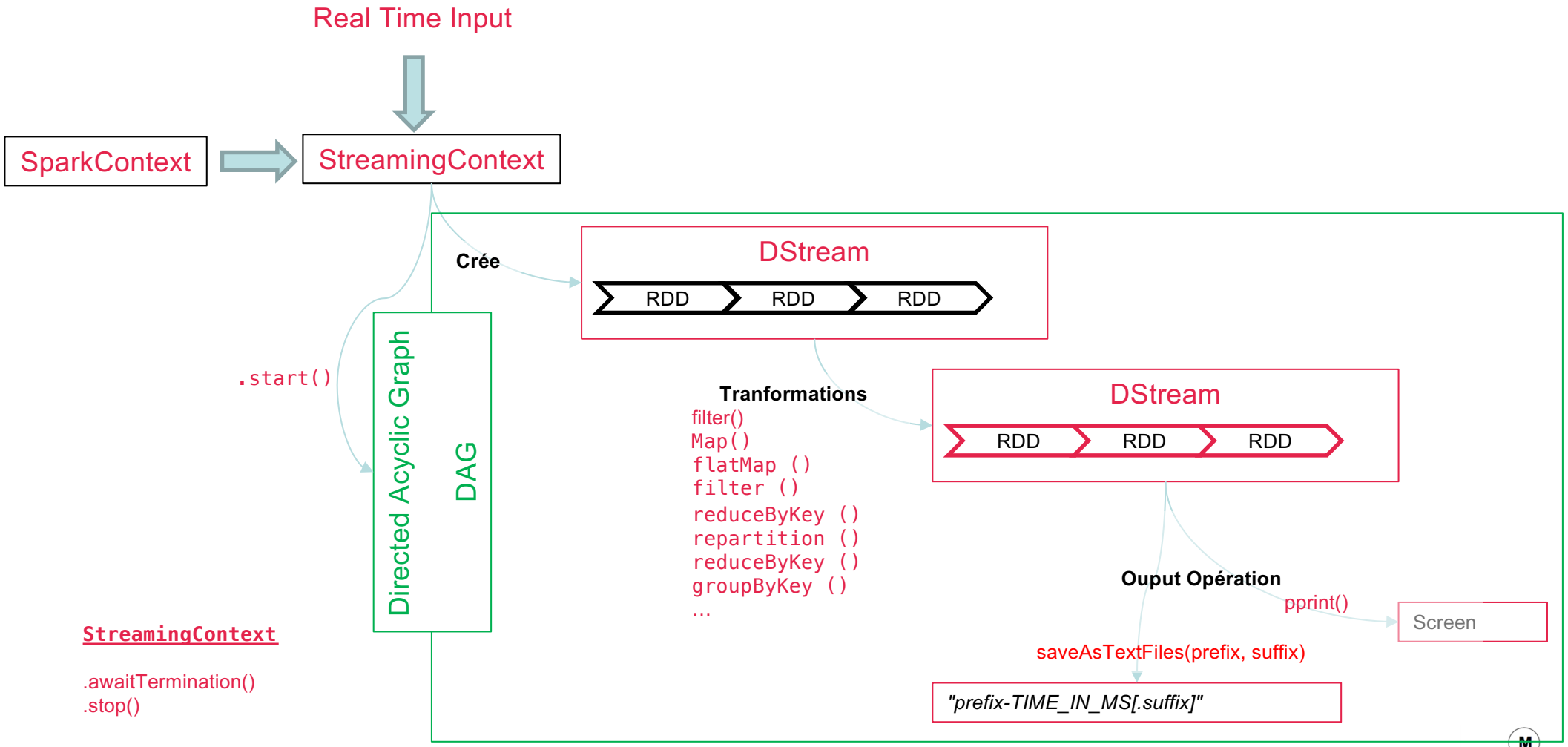
Données changeantes	Discretized Stream	DStream
Latence < ms	Spark peut descendre à des précisions de l'ordre de la milliseconde	
Presque temps réel	Micro Batching	Processé comme un enregistrement classique
Résilience	Mécanisme de « write ahead »	
Extensible	By Design !	
En mémoire	By Design !	

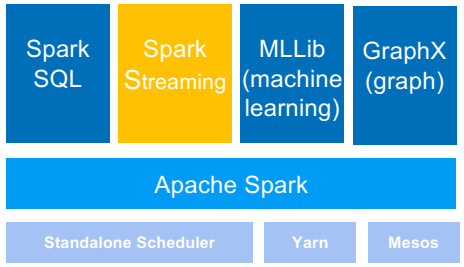
Concepts

Input Data Streams	**Basic Data Sources**	connecteurs intégrés à Spark \-> csv json parquet file hdfs
	Advanced Data Sources	besoin de librairie externe \-> mongoDB, Cassandra, Kinesis…
Batch	Durée d'un Dstream (lot de RDD)	
Output Data Stream	Point de sortie	fichier, base de données, web socket

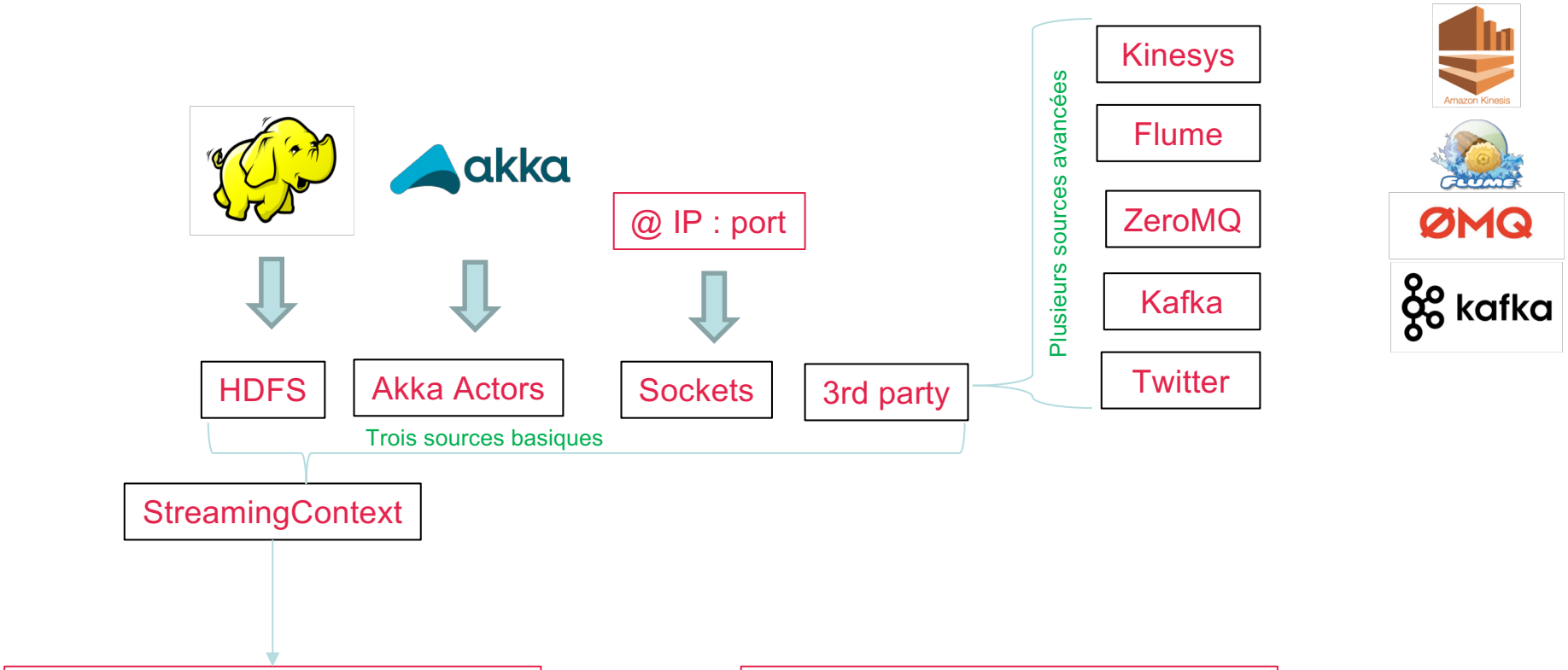
Spark Streaming 1.6 – Concepts généraux

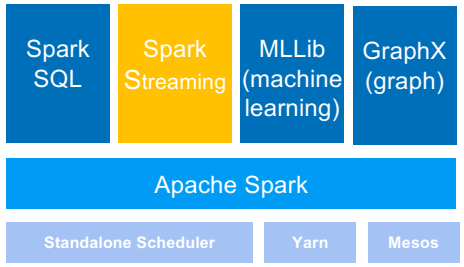






Spark Streaming 1.6 – Type d'entrées





Spark Streaming – Démo 1

from pyspark.streaming import StreamingContext

Conf spark

```
try    : sc.stop()
except : pass
try    : ssc.stop()
except : pass
conf = SparkConf()
conf.setAppName("Streaming hdfs logs")
conf.set("spark.executor.memory","1g")
conf.set("spark.driver.memory","1g")
conf.set("spark.cores.max","5")

sc   = SparkContext(conf=conf)
ssc = StreamingContext(sc, 10)
```

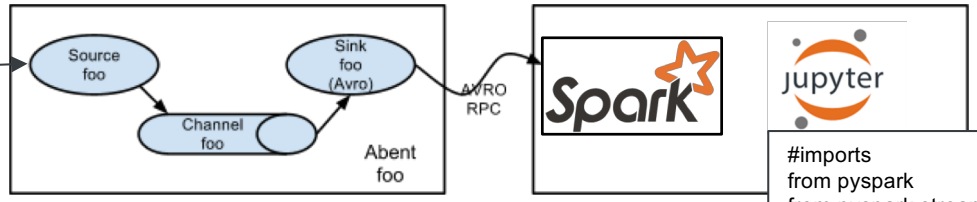
```
dstream = ssc.socketTextStream("127.0.0.1", 9999)
```

```
mots   = dstream.flatMap( lambda line     : line.split(" ") )
k_v    = mots.map       ( lambda word     : (word, 1)       )
counts = k_v.reduceByKey( lambda somme,un : somme + un      )
counts.pprint()
```

```
ssc.start()
```

```
ssc.awaitTermination()
```

Spark Streaming – Démo 2



notebook

```
#imports
from pyspark                    import SparkContext
from pyspark.streaming          import StreamingContext
from pyspark.streaming.flume    import FlumeUtils

#parametres
appname         = "PythonStreamingFlumeWordCount"
hostname, port = ("MacBook-Pro-de-oursin.local", "49490")
logdir          = "/Users/romain/Informatique/zoo/spark-1.6.0-bin-hadoop2.4/logs/dstream/"
outputpath      = logdir + appname

# initialisation
#sc             = SparkContext(appName=appname)
ssc             = StreamingContext(sc, 1)
kvs             = FlumeUtils.createStream(ssc, hostname, int(port))

# algo
lines           = kvs.map(lambda x: x[1])
counts          = lines.flatMap(lambda line: line.split(" ")) \
                    .map(lambda word: (word, 1)) \
                    .reduceByKey(lambda a, b: a+b)
counts.saveAsTextFiles(outputpath)

# starting
ssc.start()

# listening during 15 seconds
time.sleep(15)
ssc.stop()
```

Terminal 1

```
./bin/pyspark
--master
spark://MacBook-Pro-de-oursin.local:7077
--total-executor-cores
1
--packages
com.databricks:spark-csv_2.10:1.4.0
--jars
/Users/romain/Informatique/zoo/spark-1.6.0-bin-hadoop2.4/lib/spark-streaming-flume-sink_2.10-1.6.0.jar
--jars
/Users/romain/Informatique/zoo/spark-1.6.0-bin-hadoop2.4/lib/spark-streaming-flume_2.10-1.6.0.jar
--jars
/Users/romain/Informatique/zoo/spark-1.6.0-bin-hadoop2.4/lib/spark-streaming-flume-assembly_2.10-1.6.0.jar

> ./logs/pyspark.log 2>&1
```

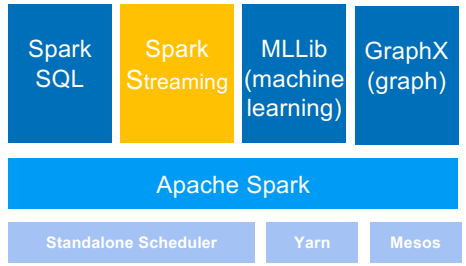
Terminal 2

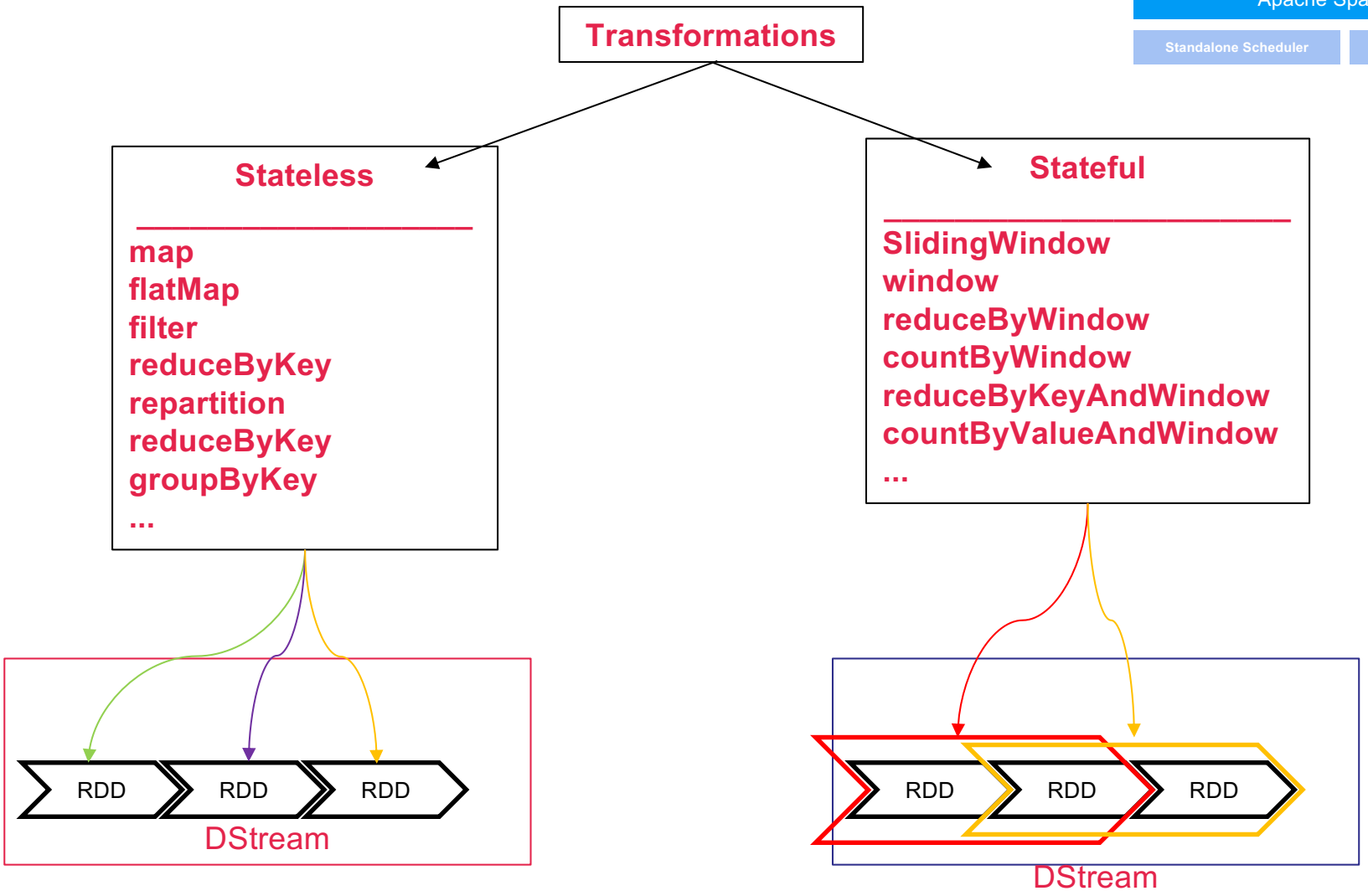
Dans apache-flume-1.6.0-bin :

```
bin/flume-ng agent
--conf conf
--conf-file ./conf/avro.conf
--name agent1
-Dflume.root.logger=INFO,console
```

```
agent1.sources = source1
agent1.channels= channel1
agent1.sinks   = spark


agent1.channels.channel1.type                = memory
agent1.channels.channel1.capacity            = 2000000
agent1.channels.channel1.transactionCapacity = 1000000

agent1.sources.source1.type     = exec
agent1.sources.source1.channels = channel1
agent1.sources.source1.command  = tail -f /Users/romain/Informatique/zoo/spark-1.6.0-bin-hadoop2.4/logs/pyspark.log

agent1.sinks                    = avroSink
agent1.sinks.avroSink.type      = avro
agent1.sinks.avroSink.channel   = channel1
agent1.sinks.avroSink.hostname  = MacBook-Pro-de-oursin.local
agent1.sinks.avroSink.port      = 49490
```

Spark Streaming – Transformations de D-Stream





Spark Streaming – Concepts généraux