Diseño de Sistemas Distribuidos

Máster en Ciencia y Tecnología Informática

Curso 2018-2019

Sistemas escalables en entornos distribuidos. Introducción a Spark

Alejandro Calderón Mateos & Jaime Pons Bailly-Bailliere

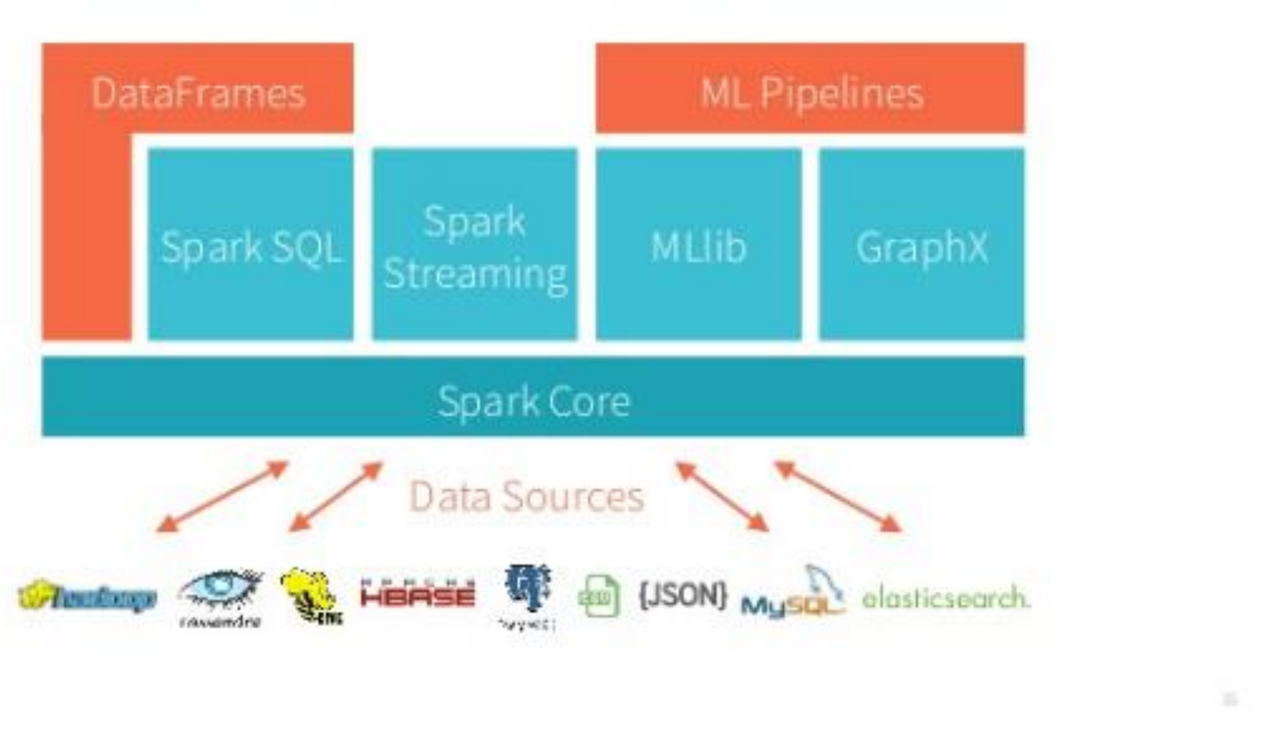
acaldero@inf.uc3m.es jaime@lab.inf.uc3m.es

Contenidos

- **Introducción**
- *Hand-on*
  - Pre-requisitos e instalación
  - Nodo autónomo
  - Cluster
- *Benchmarking*

http://Spark.apache.org

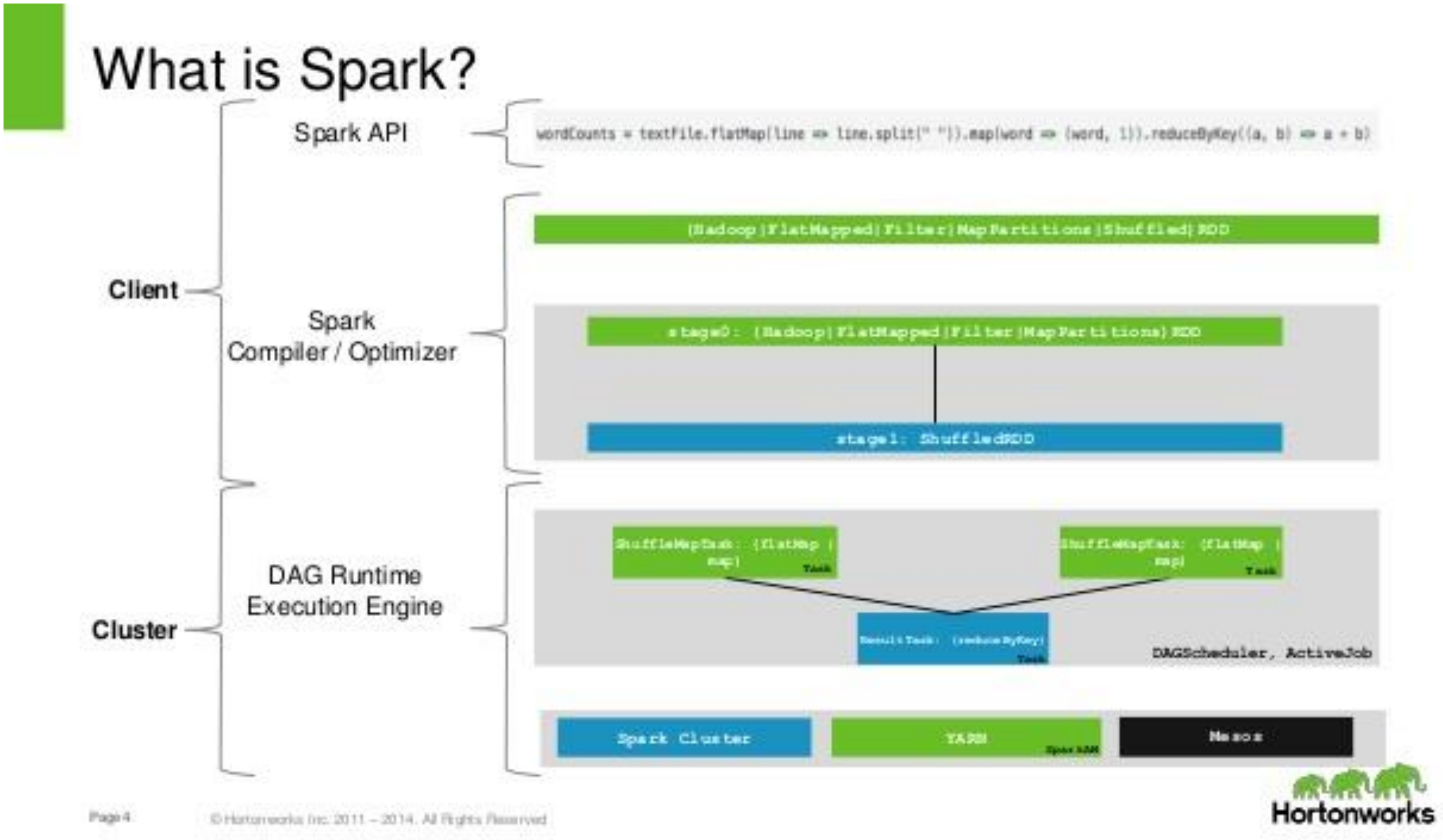
Arquitectura



https://www.mapr.com/sites/default/files/blogimages/Spark-core-stack-DB.jpg

Spark

Arquitectura



What is Spark?

Client

Cluster

Spark API

```
wordCounts = textFile.flatMap(line => line.split(" ")).map(word => (word, 1)).reduceByKey((a, b) => a + b)
```

Spark Compiler / Optimizer

(Hadoop|FlatMapped|Filter|MapPartitions|Shuffled)RDD

stage0: (Hadoop|FlatMapped|Filter|MapPartitions)RDD

stage1: ShuffledRDD

DAG Runtime Execution Engine

ShuffleMapTask: (flatMap | map)

Task

ShuffleMapTask: (flatMap | map)

Task

ResultTask: (reduceByKey)

Task

DAGScheduler, ActiveJob

Spark Cluster

YARN

Spark AM

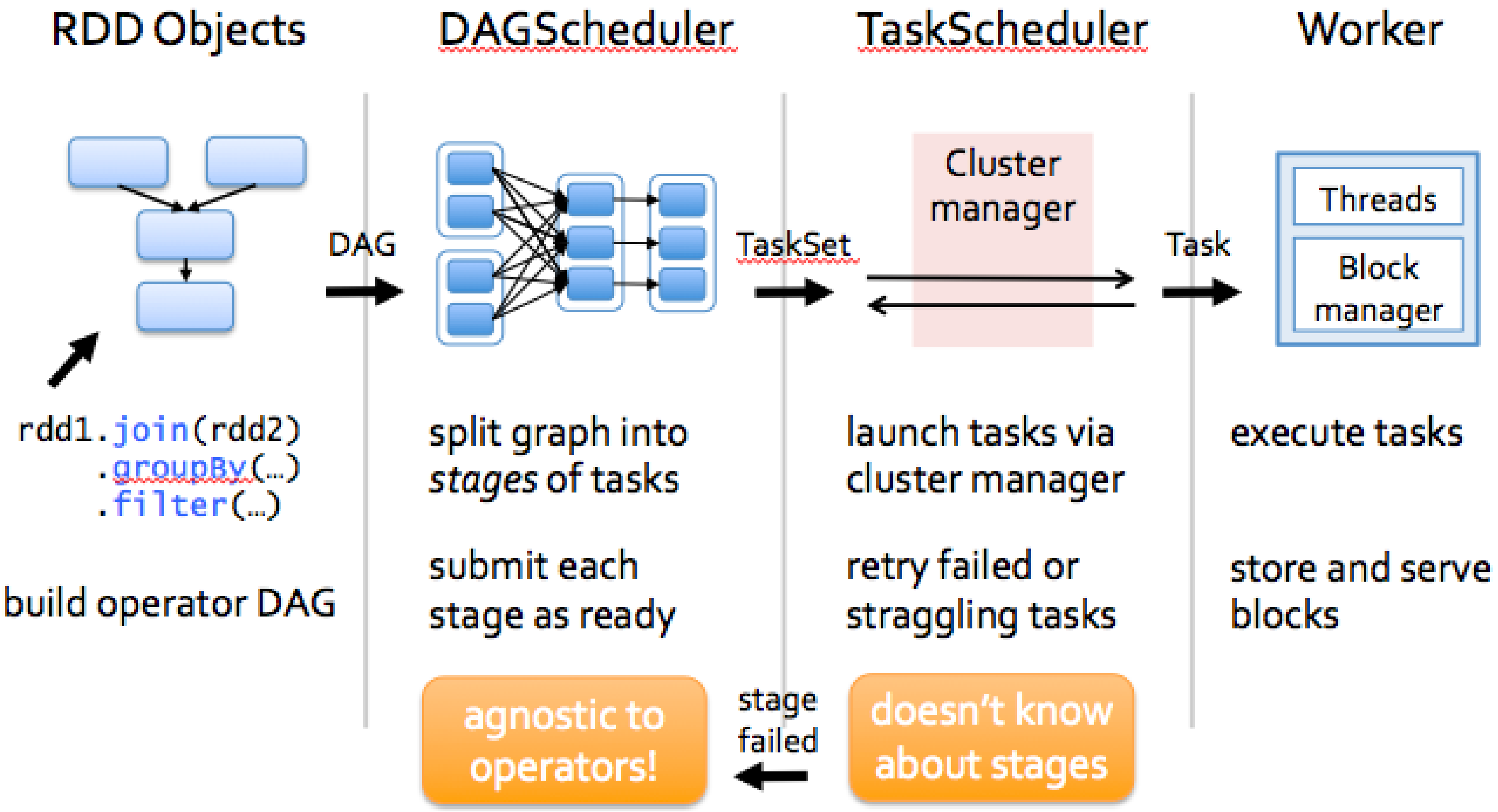
Mesos

Page 4

Hortonworks

http://image.slidesharecdn.com/sparkwebinaroctober-141002152140-phpapp02/95/yarn-ready-apache-spark-4-638.jpg?cb=1412263553

Arquitectura



```
rdd1.join(rdd2)
    .groupBy(…)
    .filter(…)
```

build operator DAG

split graph into *stages* of tasks

submit each stage as ready

agnostic to operators!

launch tasks via cluster manager

retry failed or straggling tasks

doesn't know about stages

execute tasks

store and serve blocks

https://sigmoid.com/wp-content/uploads/2015/03/Apache_Spark1.png

Contenidos

- Introducción
- ***Hand-on***
  - **Pre-requisitos e instalación**
  - Nodo autónomo
  - Cluster
- *Benchmarking*

http://Spark.apache.org

Spark, Anaconda y Jupyter

Prerequisitos Instalación Prueba básica

```
acaldero@h1:~$ du -mh -s .
2,8G .
```

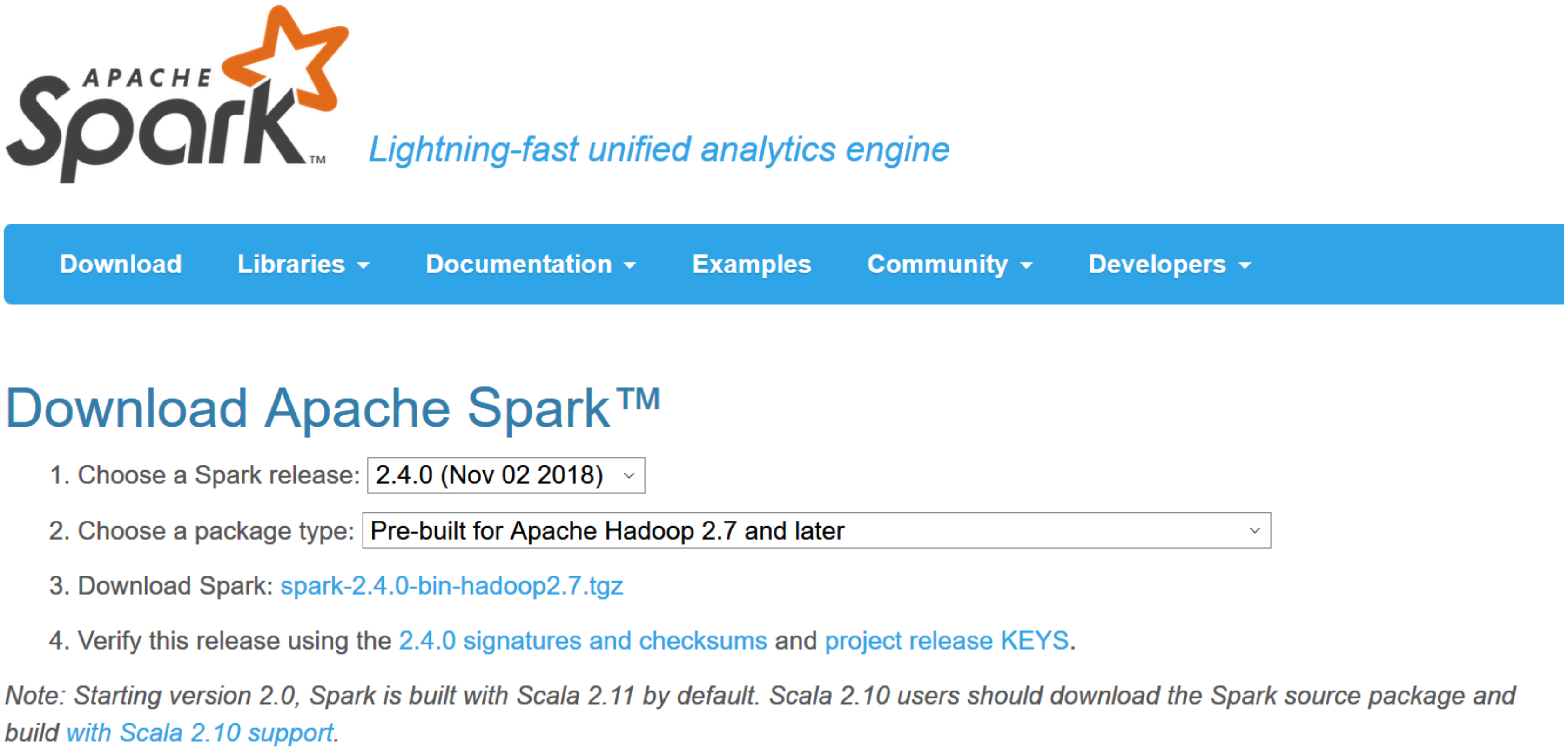
Spark

Prerequisitos Instalación Prueba básica

```
acaldero@h1:~$ sudo apt-get install ssh rsync
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following NEW packages will be installed:
  rsync ssh
…

acaldero@h1:~$ sudo apt-get install default-jdk
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following extra packages will be installed:
  libice-dev libpthread-stubs0-dev libsm-dev libx11-dev libx11-doc
  libxau-dev libxcb1-dev libxdmcp-dev libxt-dev openjdk-7-jdk
…
```

Spark

Prerequisitos **Instalación** Prueba básica



Lightning-fast unified analytics engine

Download Apache Spark™

1. Choose a Spark release: 2.4.0 (Nov 02 2018)
2. Choose a package type: Pre-built for Apache Hadoop 2.7 and later
3. Download Spark: spark-2.4.0-bin-hadoop2.7.tgz
4. Verify this release using the 2.4.0 signatures and checksums and project release KEYS.

Note: Starting version 2.0, Spark is built with Scala 2.11 by default. Scala 2.10 users should download the Spark source package and build with Scala 2.10 support.

http://spark.apache.org/downloads.html

Spark

Prerequisitos **Instalación** Prueba básica

```
acaldero@h1:~$ wget https://www.apache.org/dyn/closer.lua/spark/spark-2.4.0/spark-2.4.0-bin-hadoop2.7.tgz
…
2018-11-18 12:40:44 (6,02 MB/s) - “spark-2.2.0-bin-hadoop2.7.tgz” guardado [...]


acaldero@h1:~$ tar zxf spark-2.4.0-bin-hadoop2.7.tgz
acaldero@h1:~$ ls -las spark-2.4.0-bin-hadoop2.7
total 96
 4 drwxr-xr-x 12 acaldero acaldero  4096 jul  1 01:09 .
 4 drwx------ 39 acaldero acaldero  4096 oct 17 00:50 ..
 4 drwxr-xr-x  2 acaldero acaldero  4096 jul  1 01:09 bin
 4 drwxr-xr-x  2 acaldero acaldero  4096 jul  1 01:09 conf
 0 drwxr-xr-x  5 acaldero acaldero    47 jul  1 01:09 data
 0 drwxr-xr-x  4 acaldero acaldero    27 jul  1 01:09 examples
12 drwxr-xr-x  2 acaldero acaldero  8192 jul  1 01:09 jars
20 -rw-r--r--  1 acaldero acaldero 17881 jul  1 01:09 LICENSE
 4 drwxr-xr-x  2 acaldero acaldero  4096 jul  1 01:09 licenses
28 -rw-r--r--  1 acaldero acaldero 24645 jul  1 01:09 NOTICE
 4 drwxr-xr-x  8 acaldero acaldero  4096 jul  1 01:09 python
 0 drwxr-xr-x  3 acaldero acaldero    16 jul  1 01:09 R
 4 -rw-r--r--  1 acaldero acaldero  3809 jul  1 01:09 README.md
 4 -rw-r--r--  1 acaldero acaldero   128 jul  1 01:09 RELEASE
 4 drwxr-xr-x  2 acaldero acaldero  4096 jul  1 01:09 sbin
 0 drwxr-xr-x  2 acaldero acaldero    41 jul  1 01:09 yarn
```

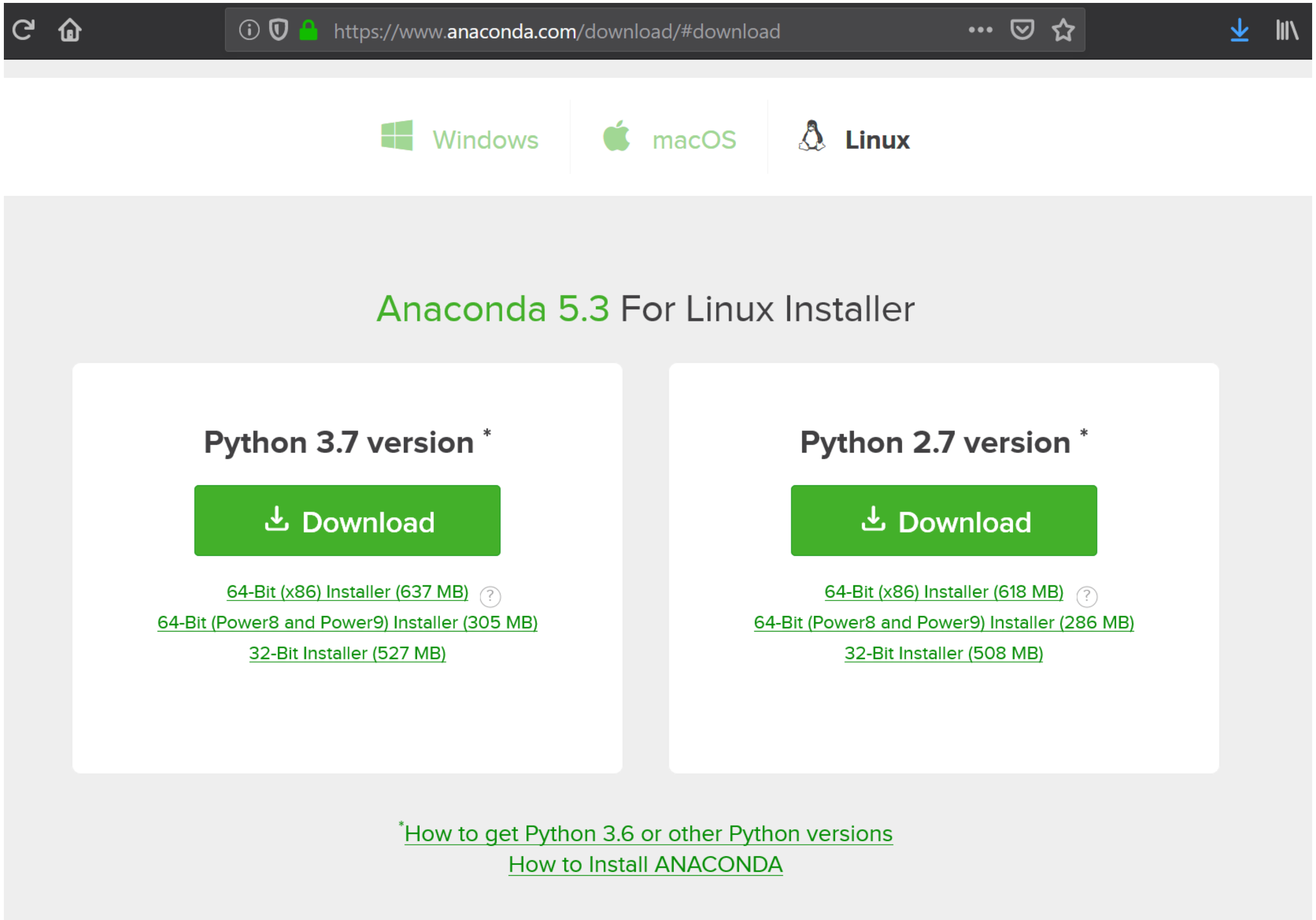
Spark

Prerequisitos Instalación **Prueba básica**

```
acaldero@h1:~$ ./bin/run-example SparkPi 5
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
17/10/17 01:02:41 INFO SparkContext: Running Spark version 2.2.0
17/10/17 01:02:42 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using
     builtin-java classes where applicable
17/10/17 01:02:42 INFO SparkContext: Submitted application: Spark Pi
17/10/17 01:02:42 INFO SecurityManager: Changing view acls to: acaldero
17/10/17 01:02:42 INFO SecurityManager: Changing modify acls to: acaldero
17/10/17 01:02:42 INFO SecurityManager: Changing view acls groups to:
17/10/17 01:02:42 INFO SecurityManager: Changing modify acls groups to:
17/10/17 01:02:42 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users
     with view permissions: Set(acaldero); groups with view permissions: Set(); users  with modify
     permissions: Set(acaldero); groups with modify permissions: Set()
17/10/17 01:02:42 INFO Utils: Successfully started service 'sparkDriver' on port 39281.
17/10/17 01:02:42 INFO SparkEnv: Registering MapOutputTracker
17/10/17 01:02:42 INFO SparkEnv: Registering BlockManagerMaster
…
```

Anaconda

Instalación Prueba básica



Anaconda 5.3 For Linux Installer

Python 3.7 version *

64-Bit (x86) Installer (637 MB)
64-Bit (Power8 and Power9) Installer (305 MB)
32-Bit Installer (527 MB)

Python 2.7 version *

64-Bit (x86) Installer (618 MB)
64-Bit (Power8 and Power9) Installer (286 MB)
32-Bit Installer (508 MB)

*How to get Python 3.6 or other Python versions
How to Install ANACONDA

https://www.anaconda.com/download/

Anaconda

Instalación Prueba básica

```
acaldero@h1:~$ wget https://repo.continuum.io/archive/Anaconda3-5.3.0-Linux-x86_64.sh

…
2018-11-18 15:12:23 (5,57 MB/s) - “Anaconda3-5.3.0-Linux-x86_64.sh” guardado [...]

acaldero@h1:~$ chmod a+x Anaconda3-5.3.0-Linux-x86_64.sh
acaldero@h1:~$ ./ Anaconda3-5.3.0-Linux-x86_64.sh
Welcome to Anaconda3 5.3.0 (by Continuum Analytics, Inc.)

In order to continue the installation process, please review the license
agreement.
Please, press ENTER to continue
>>>
…

acaldero@h1:~$ bash
acaldero@h1:~$ conda update --all
Fetching package metadata .......
Solving package specifications: .........
…
```

http://jupyter.readthedocs.io/en/latest/install.html#existing-python-new-jupyter

Spark, Anaconda y Jupyter

Configuración

```
acaldero@h1:~$ ln -s spark-2.2.0-bin-hadoop2.7 spark
acaldero@h1:~$ echo "export PATH=$PATH:/home/acaldero/spark/bin"          >> .profile
acaldero@h1:~$ echo "export PYSPARK_DRIVER_PYTHON=ipython"                >> .profile
acaldero@h1:~$ echo "export PYSPARK_DRIVER_PYTHON_OPTS='notebook' pyspark">> .profile
acaldero@h1:~$ source .profile

…
```

https://gist.github.com/tommycarpi/f5a67c66a8f2170e263c

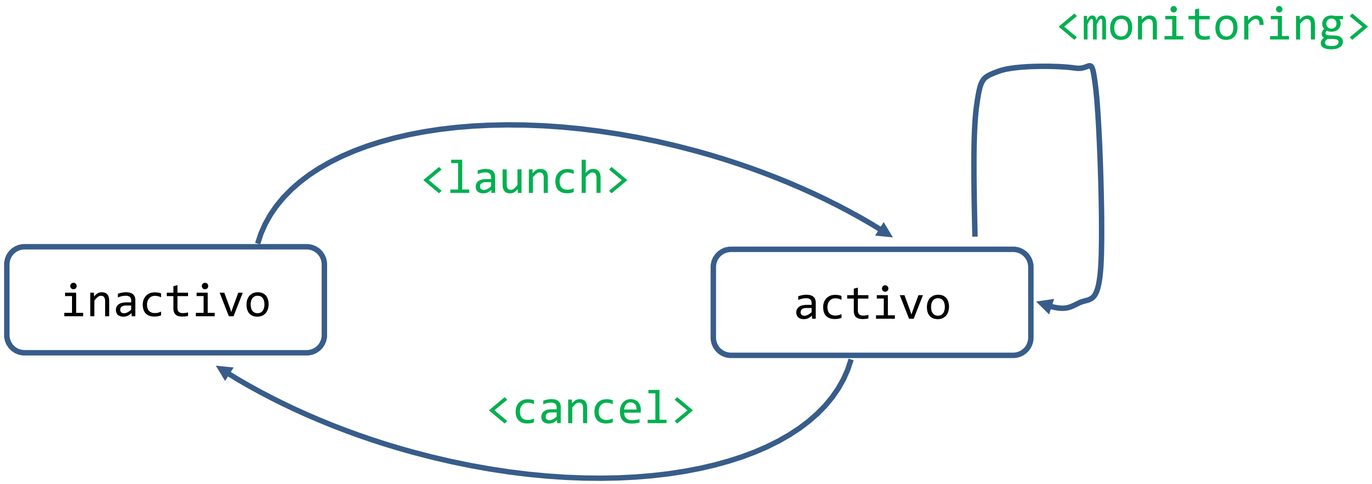
Contenidos

- Introducción
- ***Hand-on***
  - Pre-requisitos e instalación
  - **Nodo autónomo**
  - Cluster
- *Benchmarking*

http://Spark.apache.org

Spark

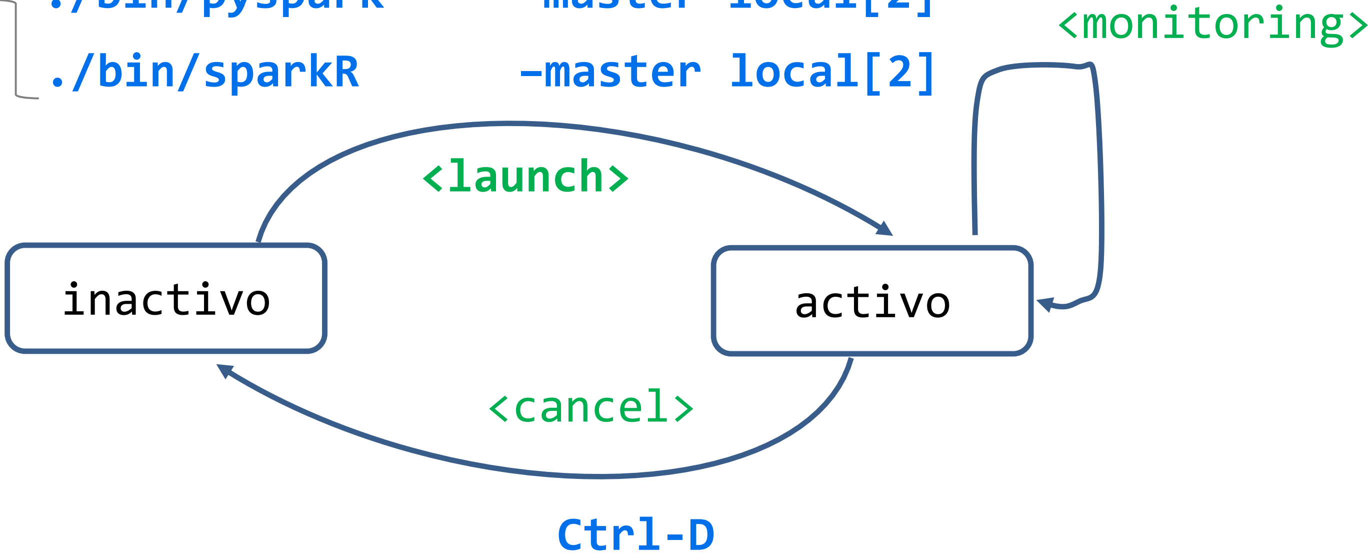
Funcionamiento General



Spark: nodo autónomo

shell-interactivo submit libro-interactivo

```
./bin/spark-shell –master local[2]
./bin/pyspark     –master local[2]
./bin/sparkR      –master local[2]
```

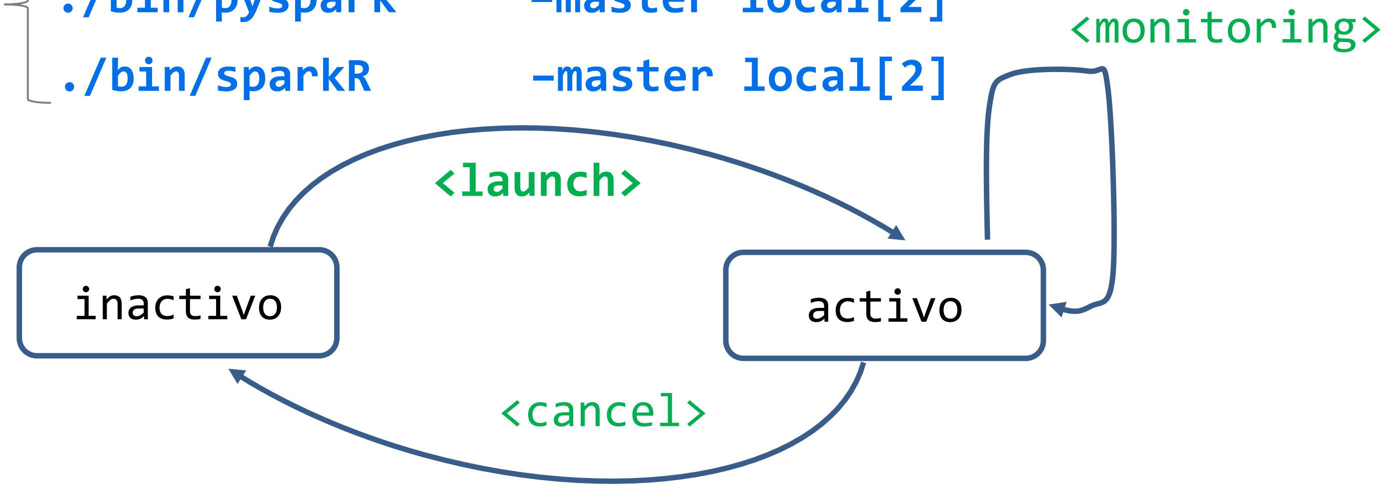


Spark: nodo autónomo

shell-interactivo submit libro-interactivo

local	-> 1 thread
local[N]	-> N threads
local[*]	-> as many threads as cores are

```
./bin/spark-shell –master local[2]
./bin/pyspark     –master local[2]
./bin/sparkR      –master local[2]
```



Ctrl-D

Spark: nodo autónomo

shell-interactivo submit libro-interactivo

```
acaldero@h1:~$ ./bin/pyspark
Python 2.7.13 (default, Jan 19 2017, 14:48:08)
[GCC 6.3.0 20170118] on linux2
Type "help", "copyright", "credits" or "license" for more information.
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
17/10/17 01:08:04 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using
     builtin-java classes where applicable
17/10/17 01:08:12 WARN ObjectStore: Version information not found in metastore.
     hive.metastore.schema.verification is not enabled so recording the schema version 1.2.0
17/10/17 01:08:12 WARN ObjectStore: Failed to get database default, returning NoSuchObjectException
17/10/17 01:08:13 WARN ObjectStore: Failed to get database global_temp, returning NoSuchObjectException
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.2.0
      /_/

Using Python version 2.7.13 (default, Jan 19 2017 14:48:08)
SparkSession available as 'spark'.
>>>
```

Spark: nodo autónomo

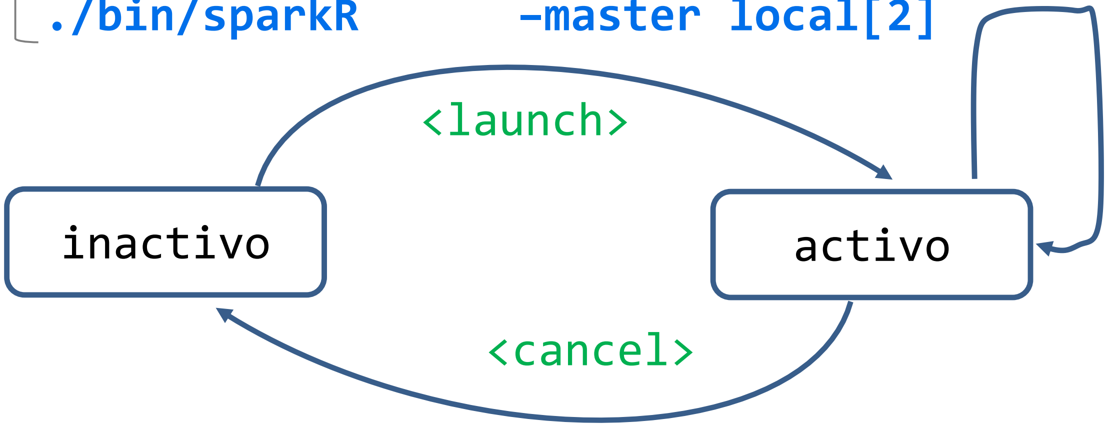
shell-interactivo submit libro-interactivo

```
Using Python version 2.7.13 (default, Jan 19 2017 14:48:08)
SparkSession available as 'spark'.
>>> import sys
>>> from random import random
>>> from operator import add
>>> from pyspark.sql import SparkSession
>>>
>>> partitions = 2
>>> n = 100000 * partitions
>>> def f(_):
...     x = random() * 2 - 1
...     y = random() * 2 - 1
...     return 1 if x ** 2 + y ** 2 < 1 else 0
...
>>> spark = SparkSession.builder.appName("PythonPi").getOrCreate()
>>> count = spark.sparkContext.parallelize(range(1, n + 1), partitions).map(f).reduce(add)
16/11/27 14:08:13 WARN TaskSetManager: Stage 0 contains a task of very large size (368 KB). The maximum
     recommended task size is 100 KB.
>>> print("Pi is roughly %f" % (4.0 * count / n))
Pi is roughly 3.139500
>>> spark.stop()
>>>
```

Spark: nodo autónomo

shell-interactivo submit libro-interactivo

```
./bin/spark-shell –master local[2]
./bin/pyspark     –master local[2]
./bin/sparkR      –master local[2]
```



Ctrl-D

Spark: nodo autónomo

shell-interactivo submit libro-interactivo

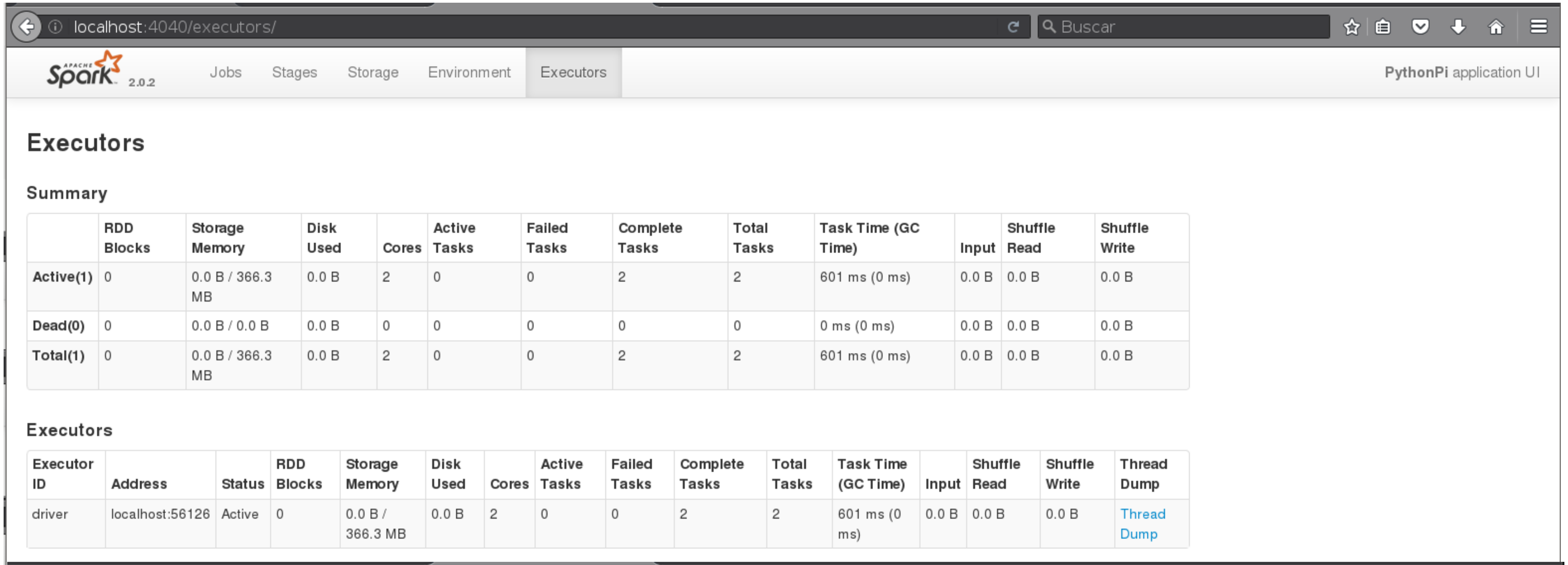
```
Using Python version 2.7.9 (default, Jun 29 2016 13:08:31)
SparkSession available as 'spark'.
>>> import sys
>>> from random import random
>>> from operator import add
>>> from pyspark.sql import SparkSession
>>>
>>> partitions = 2
>>> n = 100000 * partitions
>>> def f(_):
...     x = random() * 2 - 1
...     y = random() * 2 - 1
...     return 1 if x ** 2 + y ** 2 < 1 else 0
...
>>> spark = SparkSession.builder.appName("PythonPi").getOrCreate()
>>> count = spark.sparkContext.parallelize(range(1, n + 1), partitions).map(f).reduce(add)
16/11/27 14:08:13 WARN TaskSetManager: Stage 0 contains a task of very large size (368 KB).
    The maximum recommended task size is 100 KB.
>>> print("Pi is roughly %f" % (4.0 * count / n))
Pi is roughly 3.139500
>>> spark.stop()
>>>
```

http://<ip>:4040
http://<ip>:4041
...

http://spark.apache.org/docs/latest/monitoring.html

Spark: nodo autónomo

shell-interactivo submit libro-interactivo



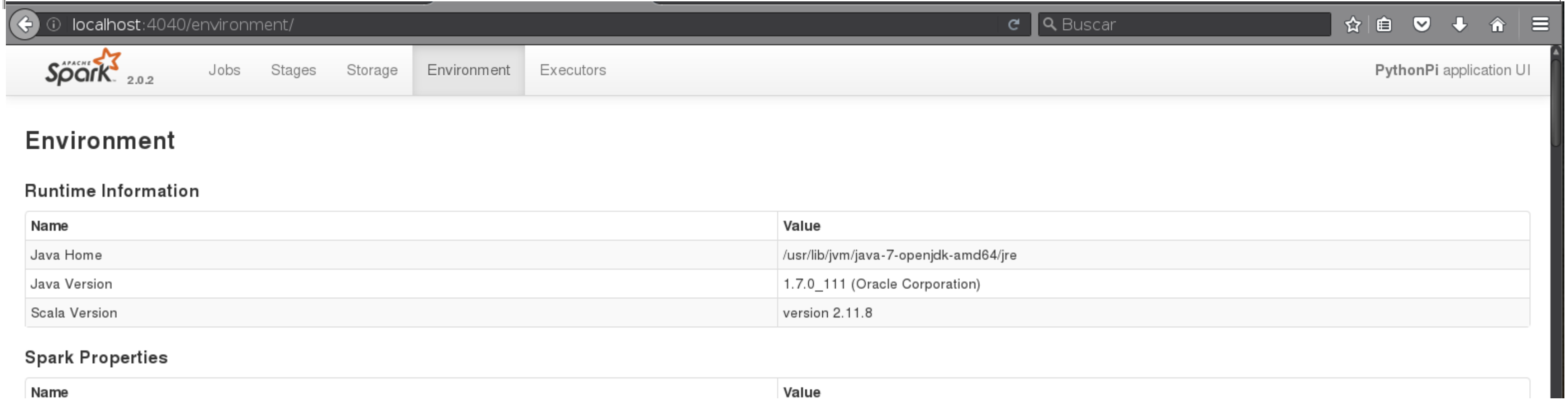
Executors

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write
Active(1)	0	0.0 B / 366.3 MB	0.0 B	2	0	0	2	2	601 ms (0 ms)	0.0 B	0.0 B	0.0 B
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0 ms (0 ms)	0.0 B	0.0 B	0.0 B
Total(1)	0	0.0 B / 366.3 MB	0.0 B	2	0	0	2	2	601 ms (0 ms)	0.0 B	0.0 B	0.0 B

Executors

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Thread Dump
driver	localhost:56126	Active	0	0.0 B / 366.3 MB	0.0 B	2	0	0	2	2	601 ms (0 ms)	0.0 B	0.0 B	0.0 B	Thread Dump



Environment

Runtime Information

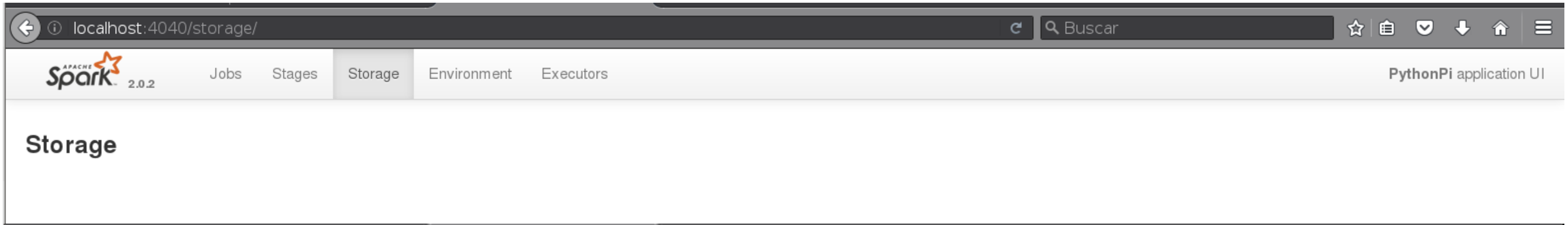
Name	Value
Java Home	/usr/lib/jvm/java-7-openjdk-amd64/jre
Java Version	1.7.0_111 (Oracle Corporation)
Scala Version	version 2.11.8

Spark Properties

Name	Value

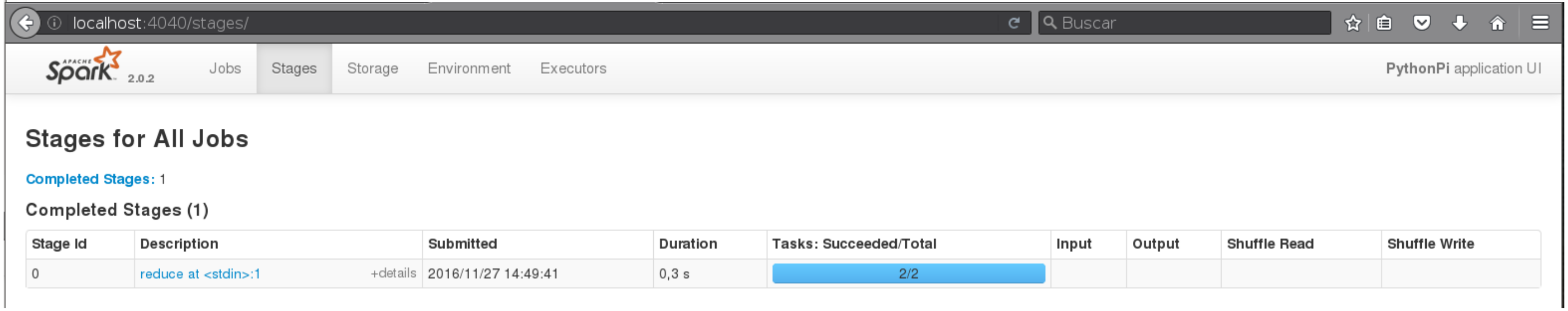
Spark: nodo autónomo

shell-interactivo submit libro-interactivo



localhost:4040/storage/

Storage



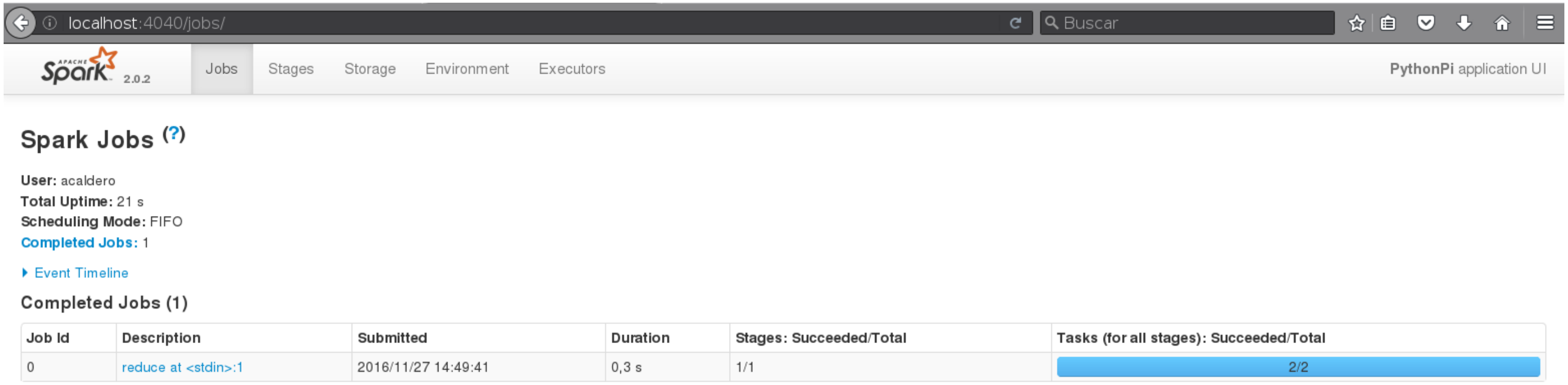
localhost:4040/stages/

Stages for All Jobs

Completed Stages: 1

Completed Stages (1)

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
0	reduce at <stdin>:1 +details	2016/11/27 14:49:41	0,3 s	2/2				



localhost:4040/jobs/

Spark Jobs (?)

User: acaldero
Total Uptime: 21 s
Scheduling Mode: FIFO
Completed Jobs: 1

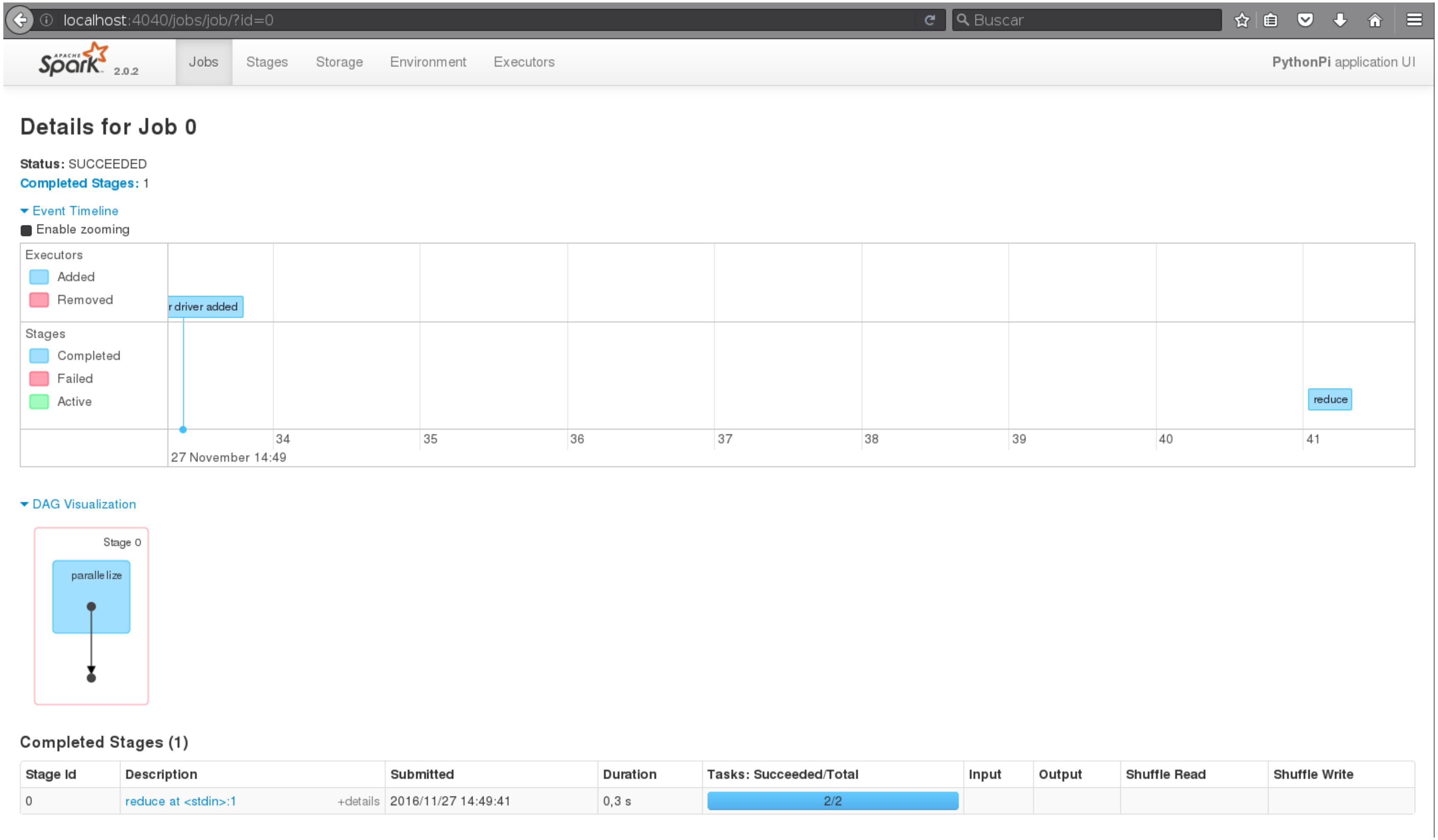
▸ Event Timeline

Completed Jobs (1)

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
0	reduce at <stdin>:1	2016/11/27 14:49:41	0,3 s	1/1	2/2

Spark: nodo autónomo

shell-interactivo submit libro-interactivo



Details for Job 0

Status: SUCCEEDED
Completed Stages: 1

▾ Event Timeline
■ Enable zooming

▾ DAG Visualization

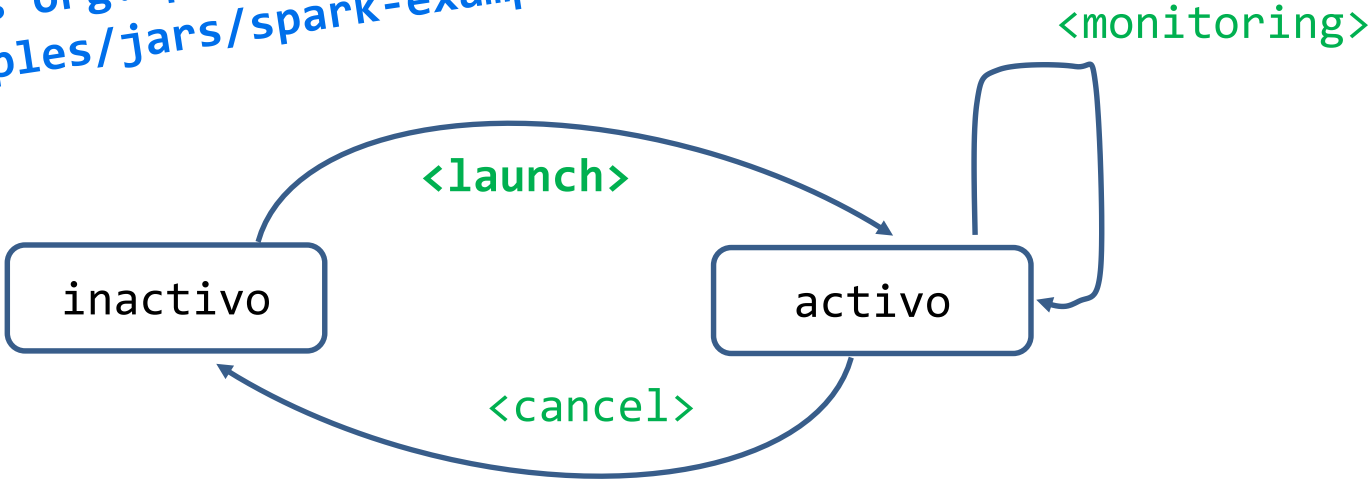
Completed Stages (1)

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
0	reduce at <stdin>:1 +details	2016/11/27 14:49:41	0,3 s	2/2				

Spark: nodo autónomo

shell-interactivo **submit** libro-interactivo

```
./bin/spark-submit --master local[8] \
  --class org.apache.spark.examples.SparkPi \
  ./examples/jars/spark-examples_2.11-2.2.0.jar \
  5
```



Ctrl-D

Spark: nodo autónomo

shell-interactivo submit **libro-interactivo**

```
acaldero@h1:~$ mkdir work
acaldero@h1:~$ cd work
acaldero@h1:~$ wget http://www.gutenberg.org/cache/epub/2000/pg2000.txt
```

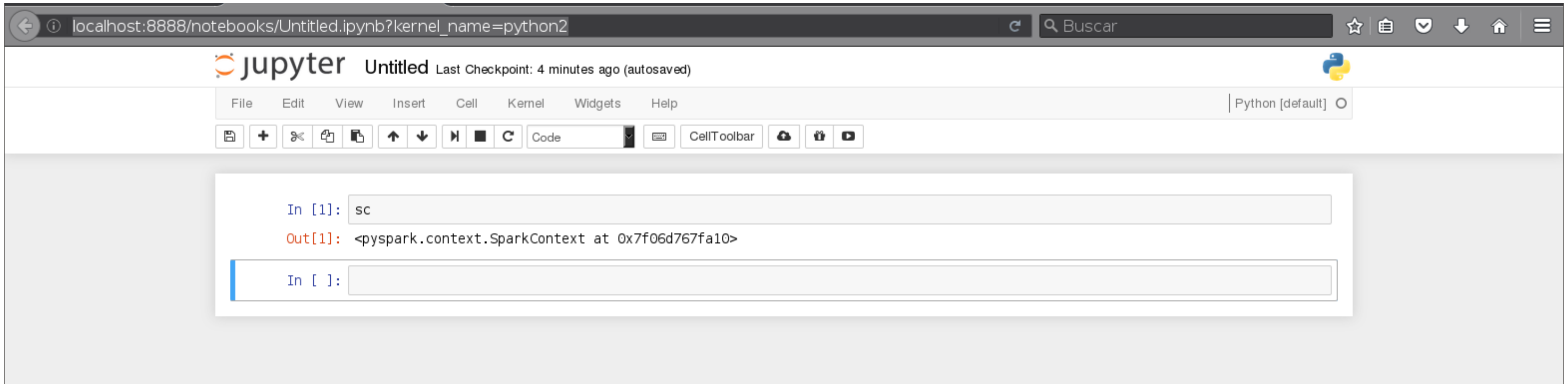
```
acaldero@h1:~$ pyspark
[TerminalIPythonApp] WARNING | Subcommand `ipython notebook` is deprecated and will be removed in future versions.
[TerminalIPythonApp] WARNING | You likely want to use `jupyter notebook` in the future
[I 18:48:14.980 NotebookApp] [nb_conda_kernels] enabled, 2 kernels found
[I 18:48:15.016 NotebookApp] ✓ nbpresent HTML export ENABLED
[W 18:48:15.016 NotebookApp] ✗ nbpresent PDF export DISABLED: No module named nbbrowserpdf.exporters.pdf
[I 18:48:15.018 NotebookApp] [nb_conda] enabled
…
```

https://gist.github.com/tommycarpi/f5a67c66a8f2170e263c

Spark: nodo autónomo

shell-interactivo submit **libro-interactivo**

```
acaldero@h1:~$ firefox http://localhost:8888/
ps# sc + <shift + enter>
```



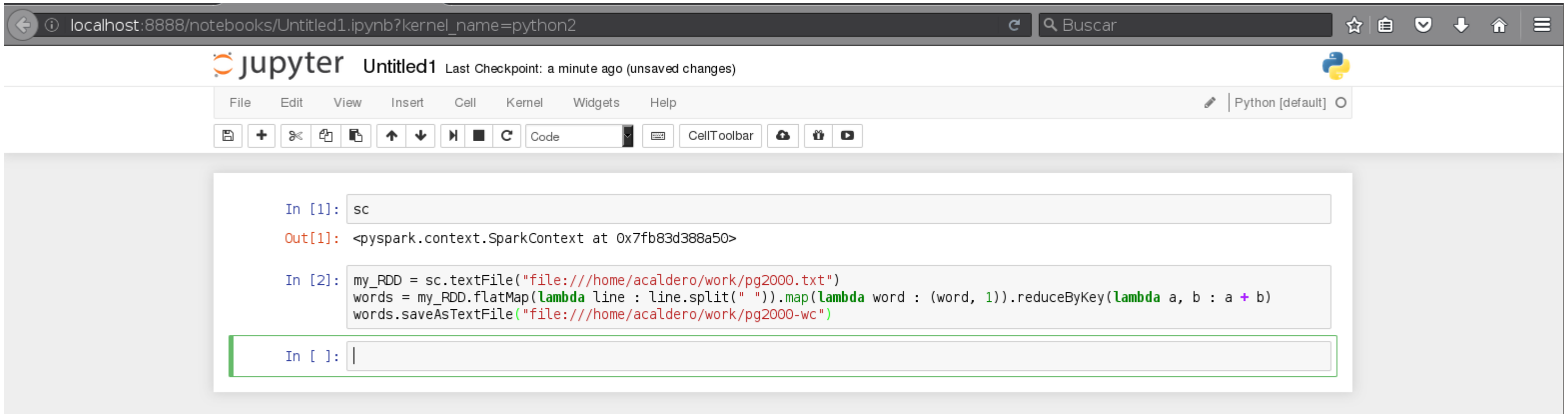
localhost:8888/notebooks/Untitled.ipynb?kernel_name=python2
Buscar
jupyter
Untitled Last Checkpoint: 4 minutes ago (autosaved)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Python [default]
Code
CellToolbar

```
In [1]: sc
Out[1]: <pyspark.context.SparkContext at 0x7f06d767fa10>

In [ ]:
```

Spark: nodo autónomo

shell-interactivo submit **libro-interactivo**

```
myRDD = sc.textFile("file:///home/acaldero/work/pg2000.txt")
words = myRDD.flatMap(lambda line : line.split(" ")).map(lambda word : (word,
    1)).reduceByKey(lambda a, b : a + b)
words.saveAsTextFile("file:///home/acaldero/work/pg2000-wc")
```



localhost:8888/notebooks/Untitled1.ipynb?kernel_name=python2
Buscar
jupyter
Untitled1
Last Checkpoint: a minute ago (unsaved changes)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Python [default]
Code
CellToolbar

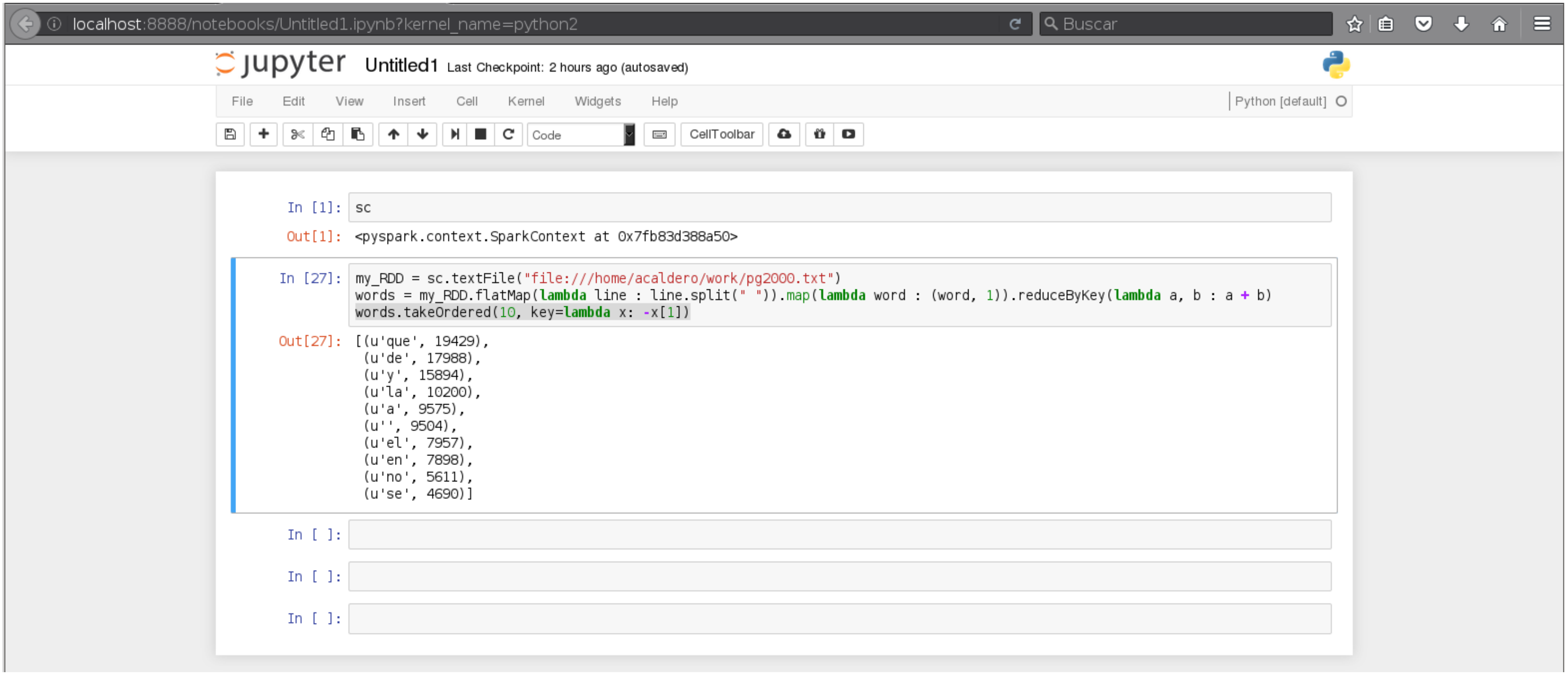
```
In [1]: sc
Out[1]: <pyspark.context.SparkContext at 0x7fb83d388a50>

In [2]: my_RDD = sc.textFile("file:///home/acaldero/work/pg2000.txt")
        words = my_RDD.flatMap(lambda line : line.split(" ")).map(lambda word : (word, 1)).reduceByKey(lambda a, b : a + b)
        words.saveAsTextFile("file:///home/acaldero/work/pg2000-wc")

In [ ]:
```

https://geekytheory.com/apache-spark-que-es-y-como-funciona/

Spark: nodo autónomo

shell-interactivo submit **libro-interactivo**

```
myRDD = sc.textFile("file:///home/acaldero/work/pg2000.txt")
words = myRDD.flatMap(lambda line : line.split(" ")).map(lambda word : (word,
   1)).reduceByKey(lambda a, b : a + b)
words.takeOrdered(10, key=lambda x: -x[1])
```



localhost:8888/notebooks/Untitled1.ipynb?kernel_name=python2
Buscar
jupyter Untitled1 Last Checkpoint: 2 hours ago (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Python [default]
Code
CellToolbar

```
In [1]: sc
Out[1]: <pyspark.context.SparkContext at 0x7fb83d388a50>

In [27]: my_RDD = sc.textFile("file:///home/acaldero/work/pg2000.txt")
         words = my_RDD.flatMap(lambda line : line.split(" ")).map(lambda word : (word, 1)).reduceByKey(lambda a, b : a + b)
         words.takeOrdered(10, key=lambda x: -x[1])
Out[27]: [(u'que', 19429),
          (u'de', 17988),
          (u'y', 15894),
          (u'la', 10200),
          (u'a', 9575),
          (u'', 9504),
          (u'el', 7957),
          (u'en', 7898),
          (u'no', 5611),
          (u'se', 4690)]

In [ ]:

In [ ]:

In [ ]:
```

http://stackoverflow.com/questions/24656696/spark-get-collection-sorted-by-value

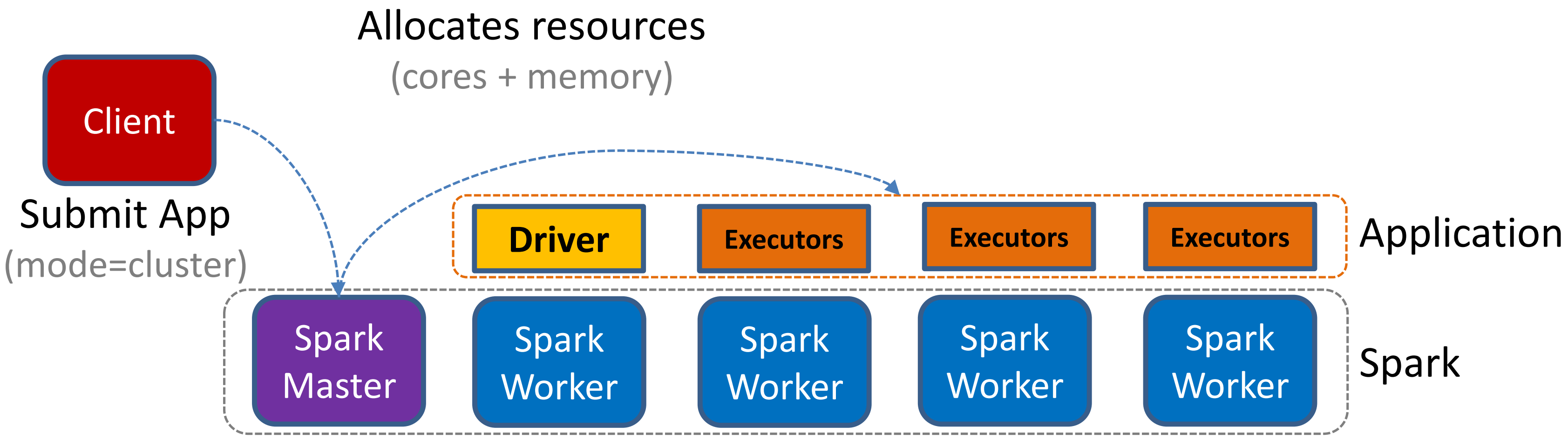
Contenidos

- Introducción
- ***Hand-on***
  - Pre-requisitos e instalación
  - Nodo autónomo
  - **Cluster**
- *Benchmarking*

http://Spark.apache.org

Spark: cluster privado

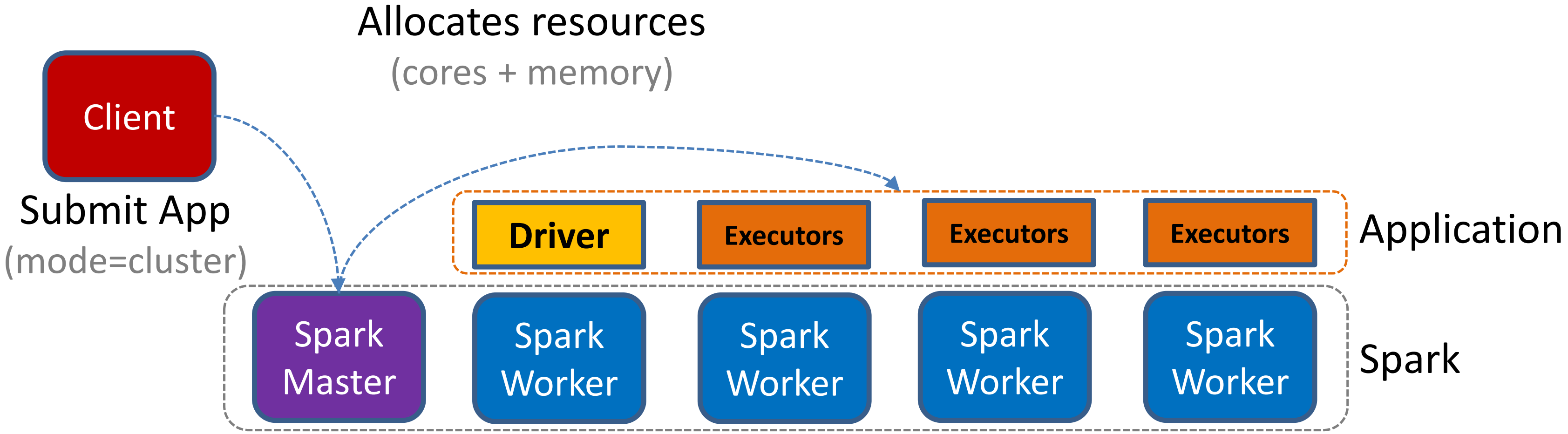
Prerequisitos Instalación Uso básico



https://trongkhoanguyenblog.wordpress.com/2014/11/23/a-gentle-introduction-to-apache-spark/

Spark: cluster privado

Prerequisitos **Instalación** Uso básico

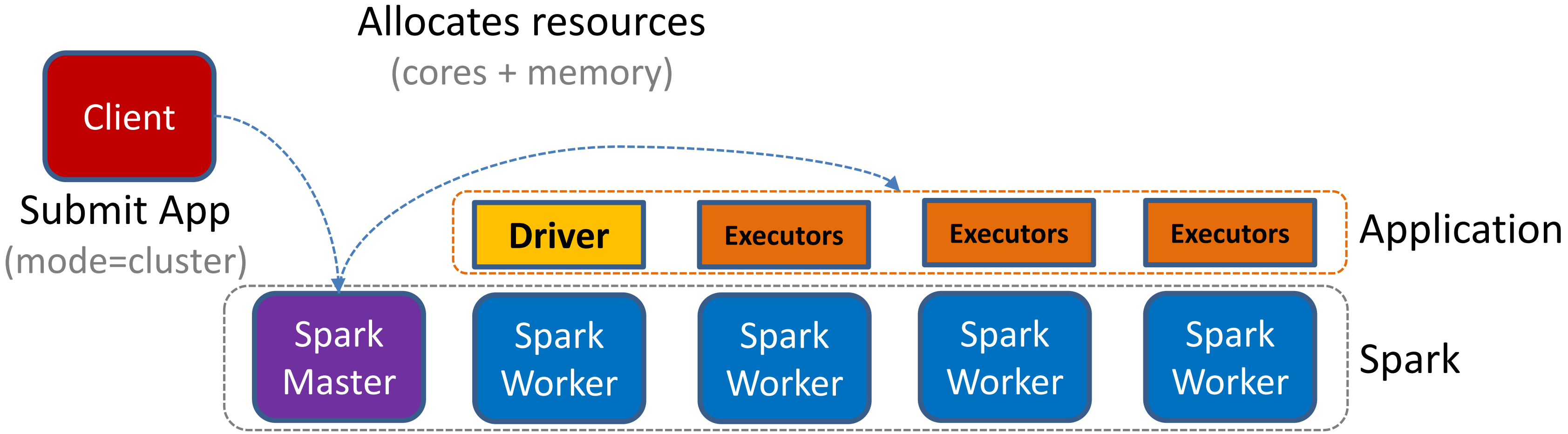


```
acaldero@h1:~$ echo "127.0.0.1  master1" >> /etc/hosts
acaldero@h1:~$ echo "127.0.0.1  slave1"  >> /etc/hosts
acaldero@h1:~$ echo "127.0.0.1  slave2"  >> /etc/hosts
```

http://spark.apache.org/docs/latest/spark-standalone.html

Spark: cluster privado

Prerequisitos **Instalación** Uso básico



```
acaldero@h1:~$ echo "node1" >> spark/conf/slaves
acaldero@h1:~$ echo "node2" >> spark/conf/slaves
```

```
acaldero@h1:~$ : Spark en todos los nodos (si fuera necesario)
acaldero@h1:~$ scp -r spark acaldero@node1:~/
…
```

http://spark.apache.org/docs/latest/spark-standalone.html

Spark: cluster privado

Prerequisitos **Instalación** Uso básico

```
acaldero@h1:/home/acaldero$ ssh-keygen -t rsa -P ""
Generating public/private rsa key pair.
Enter file in which to save the key (/home/acaldero/.ssh/id_rsa):
Created directory '/home/acaldero/.ssh'.
Your identification has been saved in /home/acaldero/.ssh/id_rsa.
Your public key has been saved in /home/acaldero/.ssh/id_rsa.pub.
The key fingerprint is:
f0:14:95:a1:0b:78:57:0b:c7:65:47:43:39:b2:2f:8a acaldero@ws1
The key's randomart image is:
+---[RSA 2048]----+
|        oo=+oo=. |
|     .   *oo..o. |
…
```

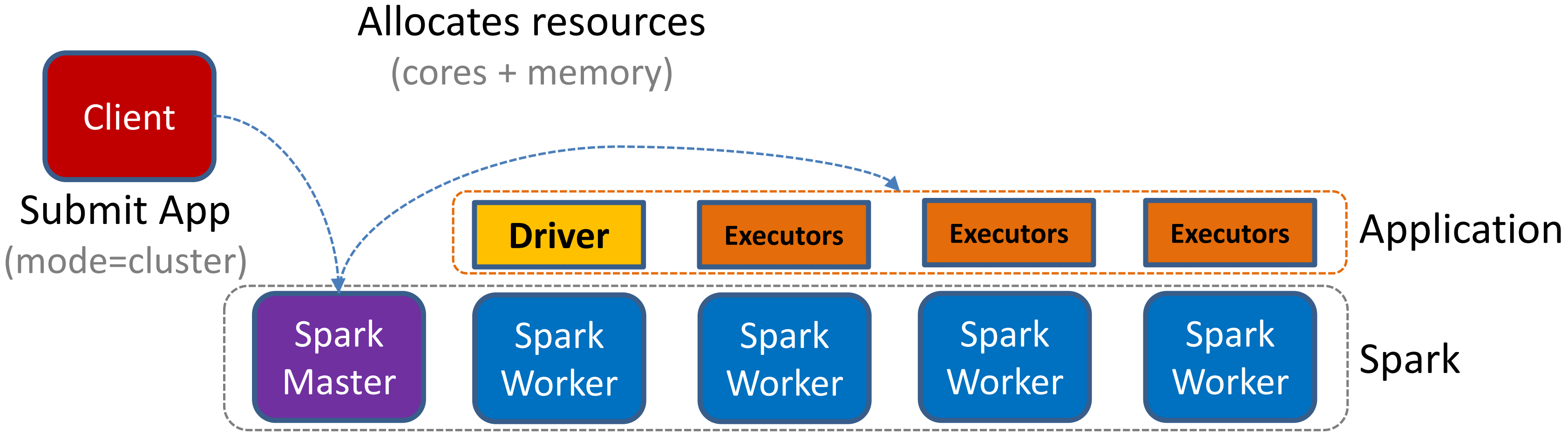
Spark: cluster privado

Prerequisitos **Instalación** Uso básico

```
acaldero@h1:/home/acaldero$ scp .ssh/id_rsa.pub acaldero@node1:~/.ssh/authorized_keys
Password:
…

acaldero@h1:/home/acaldero$ ssh node1
The authenticity of host 'localhost (::1)' can't be established.
ECDSA key fingerprint is bb:85:4c:6a:ff:e4:34:f8:ac:82:bf:56:a6:79:d8:80.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
…

acaldero@node1:~$ exit
logout
```

Spark: cluster privado

Prerequisitos Instalación **Uso básico**

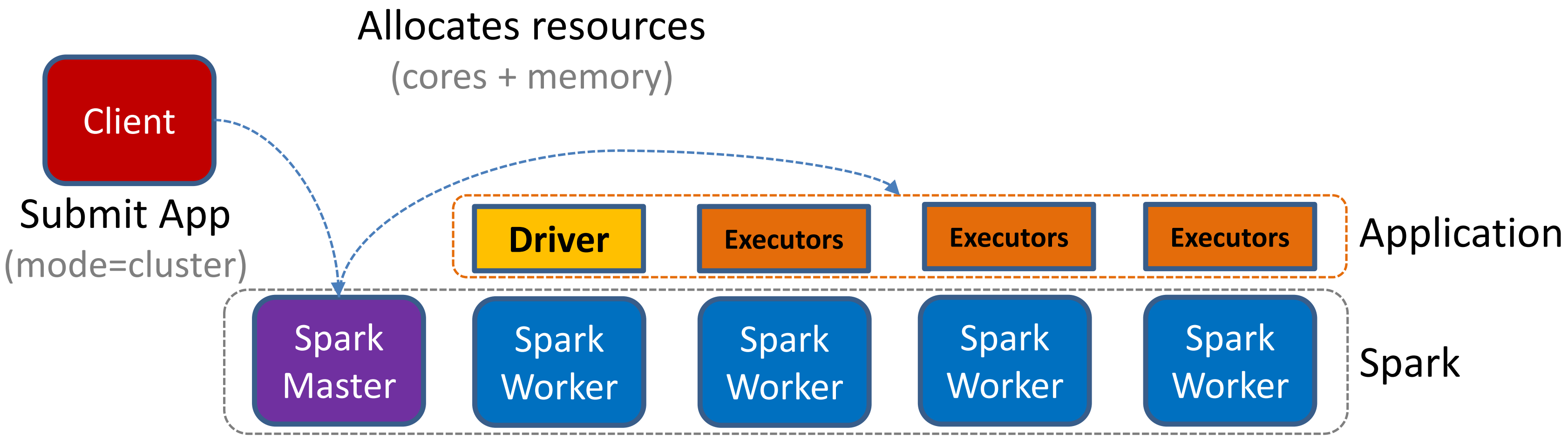


```
acaldero@h1:~$ : Ir al nodo master
acaldero@h1:~$ ssh acaldero@master1
acaldero@master1:~$ ./spark/sbin/start-all.sh
localhost: starting org.apache.spark.deploy.worker.Worker, logging to /home/acaldero/spark/logs/spark-acaldero-org.apache.spark.deploy.worker.Worker-1-ws1.out
localhost: starting org.apache.spark.deploy.worker.Worker, logging to /home/acaldero/spark/logs/spark-acaldero-org.apache.spark.deploy.worker.Worker-1-ws1.out
localhost: starting org.apache.spark.deploy.worker.Worker, logging to /home/acaldero/spark/logs/spark-acaldero-org.apache.spark.deploy.worker.Worker-1-ws1.out
…
```

http://spark.apache.org/docs/latest/spark-standalone.html

Spark: cluster privado

Prerequisitos Instalación **Uso básico**

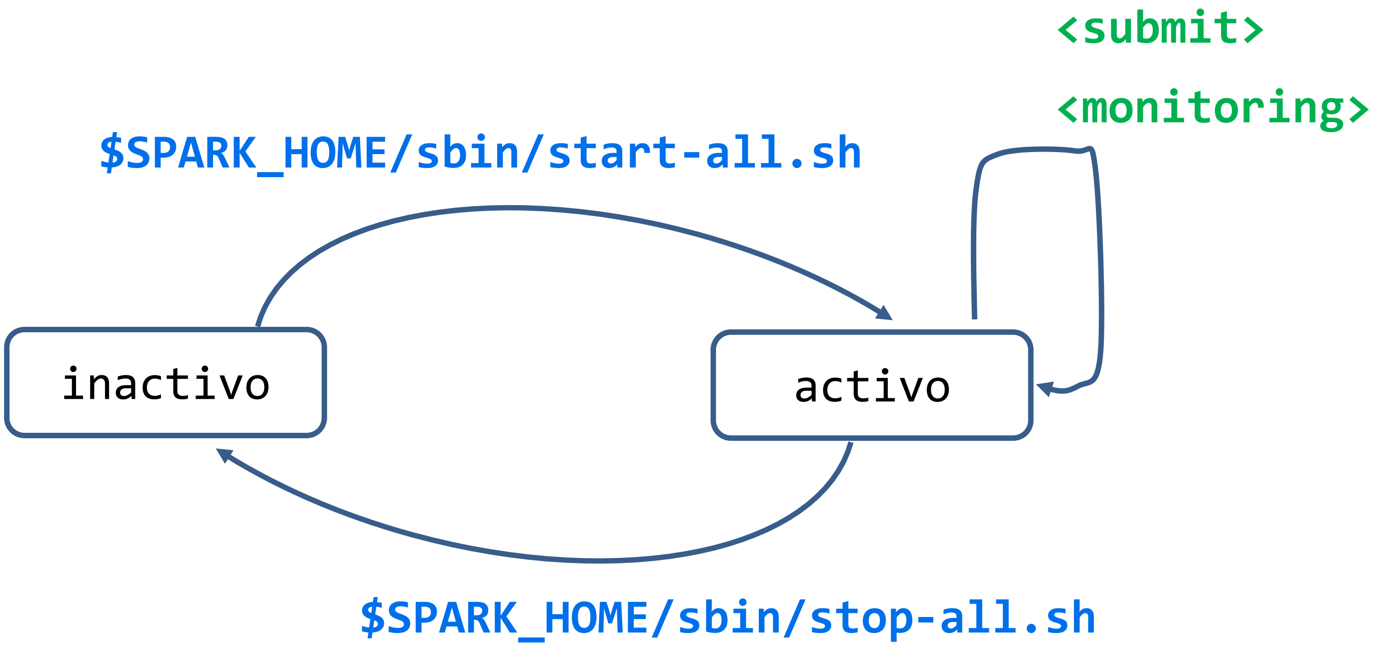


```
acaldero@master1:~$ ./spark/sbin/stop-all.sh
acaldero@master1:~$ exit
acaldero@h1:~$ : Regresar al cliente
localhost: stopping org.apache.spark.deploy.worker.Worker
localhost: stopping org.apache.spark.deploy.worker.Worker
localhost: stopping org.apache.spark.deploy.worker.Worker
stopping org.apache.spark.deploy.master.Master
```

http://spark.apache.org/docs/latest/spark-standalone.html

Spark: cluster privado

Prerequisitos Instalación **Uso básico**



<submit>

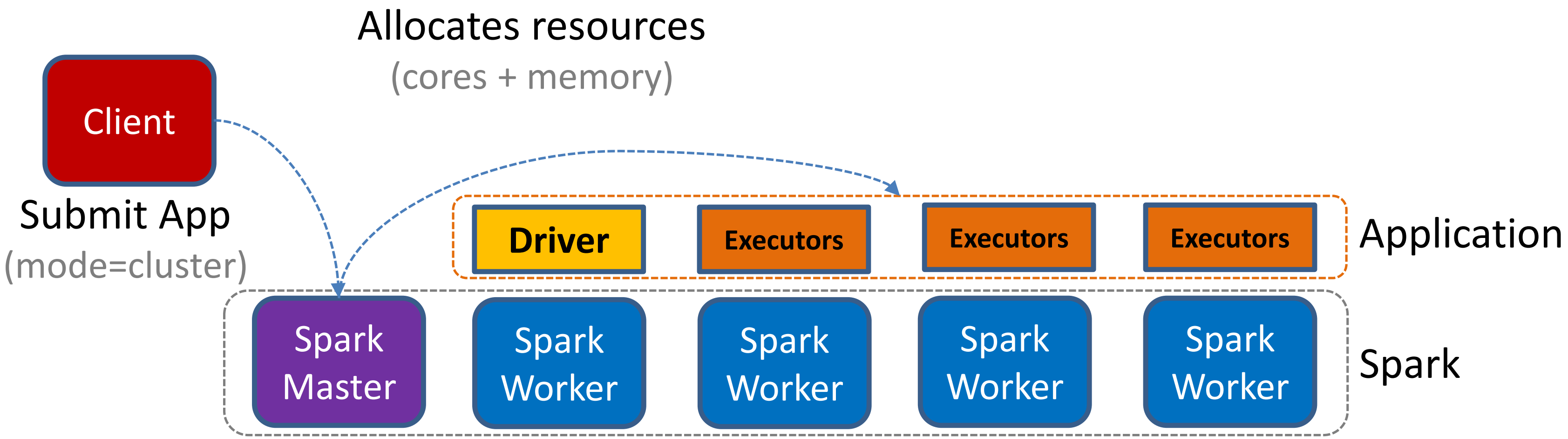
<monitoring>

$SPARK_HOME/sbin/start-all.sh

$SPARK_HOME/sbin/stop-all.sh

Spark: cluster privado

Prerequisitos Instalación **Uso básico**

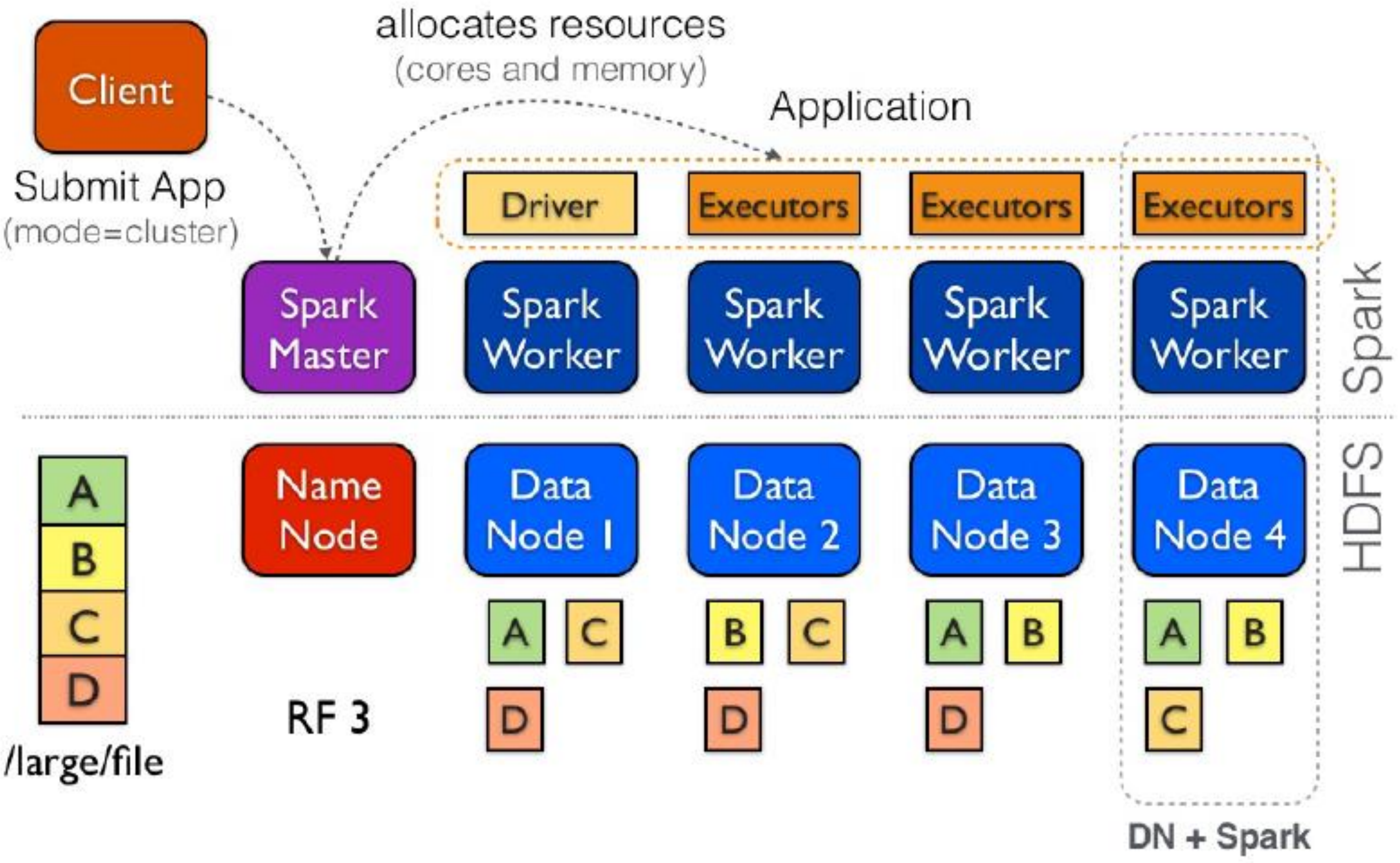


```
acaldero@h1:~$ ./spark/bin/spark-shell --master spark://master1:7077
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel).
16/11/27 23:13:55 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform...
…
scala> exit
```

http://spark.apache.org/docs/latest/spark-standalone.html

Spark: cluster privado

Prerequisitos **Instalación** Uso básico



https://trongkhoanguyenblog.wordpress.com/2014/11/23/a-gentle-introduction-to-apache-spark/

Contenidos

- Introducción
- *Hand-on*
  - Pre-requisitos e instalación
  - Nodo autónomo
  - Cluster
- ***Benchmarking***

http://Spark.apache.org

Benchmarking

- HiBench
  - https://github.com/intel-hadoop/HiBench
- Spark-perf
  - https://github.com/databricks/spark-perf

http://01org.github.io/sparkscore/about.html

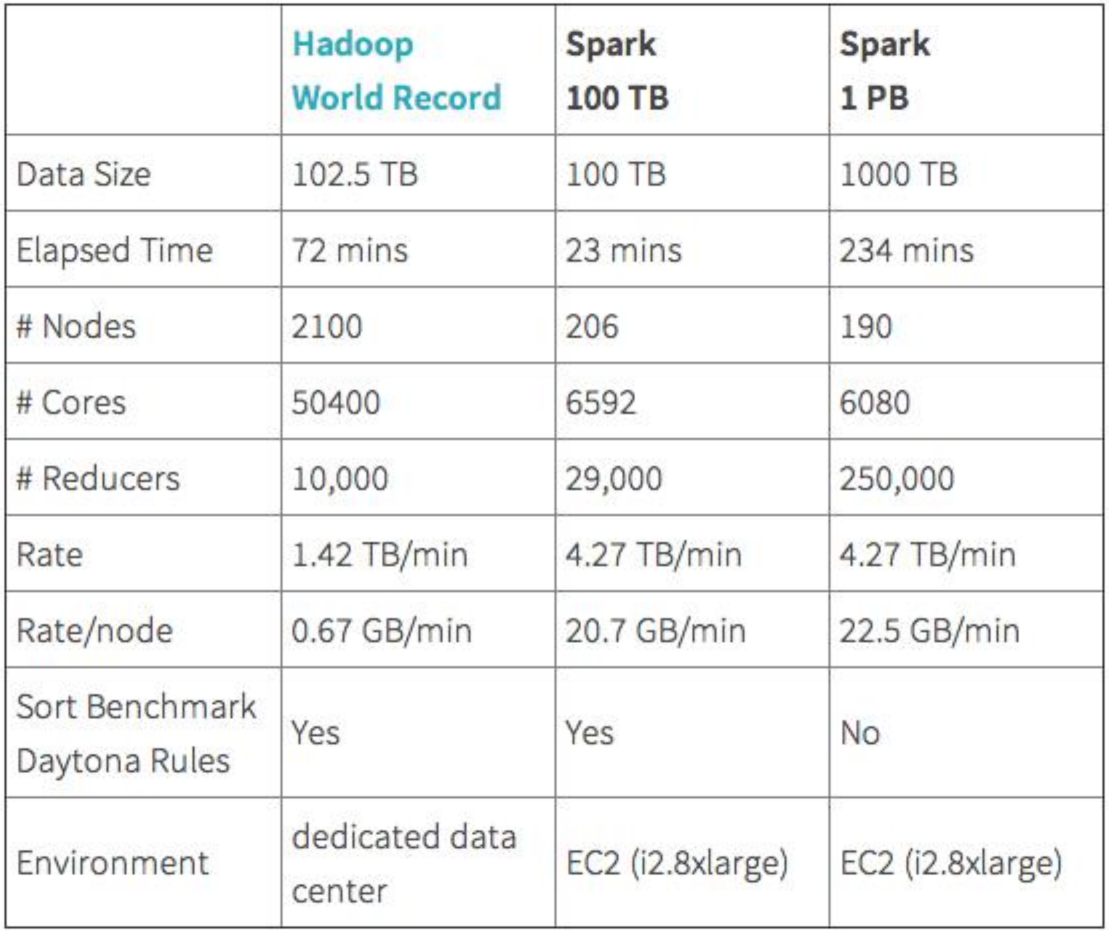
Benchmarking

- TeraSort
  - Elevada entrada y salida, y comunicación intermedia
- WordCount, PageRank
  - Contar referencias de palabras, enlaces, etc.

- SQL
  - Scan, Join, Aggregate
  - …
- Machine Learning
  - Bayesian Classification
  - K-means clustering
  - …

https://www.oreilly.com/ideas/investigating-sparks-performance

TeraSort (2014)



	Hadoop World Record	Spark 100 TB	Spark 1 PB
Data Size	102.5 TB	100 TB	1000 TB
Elapsed Time	72 mins	23 mins	234 mins
# Nodes	2100	206	190
# Cores	50400	6592	6080
# Reducers	10,000	29,000	250,000
Rate	1.42 TB/min	4.27 TB/min	4.27 TB/min
Rate/node	0.67 GB/min	20.7 GB/min	22.5 GB/min
Sort Benchmark Daytona Rules	Yes	Yes	No
Environment	dedicated data center	EC2 (i2.8xlarge)	EC2 (i2.8xlarge)

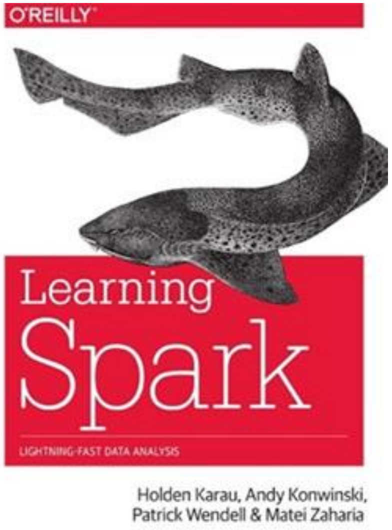
https://gigaom.com/2014/10/10/databricks-demolishes-big-data-benchmark-to-prove-spark-is-fast-on-disk-too/

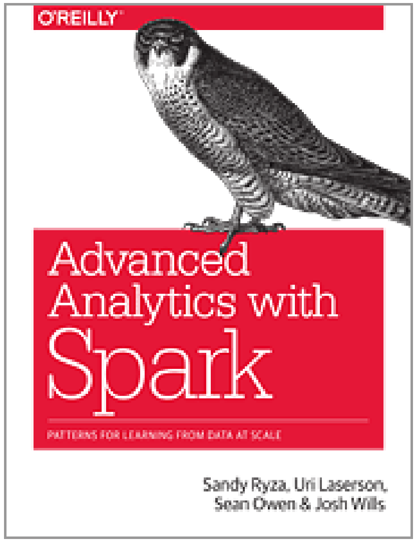
Bibliografía: tutoriales

- Página Web oficial:
  - http://spark.apache.org/
- Introducción a cómo funciona Spark:
  - http://spark.apache.org/docs/latest/quick-start.html
- Tutorial de cómo instalar y usar Spark:
  - http://spark.apache.org/docs/latest/index.html
  - http://spark.apache.org/docs/latest/configuration.html

Bibliografía: libro

- Learning Spark, Advanced Analytics with Spark:
  - http://shop.oreilly.com/product/0636920028512.do
  - http://shop.oreilly.com/product/0636920035091.do


O'REILLY
Learning Spark
LIGHTNING-FAST DATA ANALYSIS
Holden Karau, Andy Konwinski, Patrick Wendell & Matei Zaharia


O'REILLY
Advanced Analytics with Spark
PATTERNS FOR LEARNING FROM DATA AT SCALE
Sandy Ryza, Uri Laserson, Sean Owen & Josh Wills

https://github.com/databricks/learning-spark

Agradecimientos

- Por último pero no por ello menos importante, agradecer al personal del Laboratorio del Departamento de Informática todos los comentarios y sugerencias para esta presentación.

Diseño de Sistemas Distribuidos

Máster en Ciencia y Tecnología Informática

Curso 2018-2019

Sistemas escalables en entornos distribuidos. Introducción a Spark

Alejandro Calderón Mateos & Jaime Pons Bailly-Bailliere

acaldero@inf.uc3m.es jaime@lab.inf.uc3m.es