

Sistemas Paralelos y Distribuidos
Máster en Ciencia y Tecnología Informática
Diseño de Sistemas Distribuidos
Máster en Ingeniería Informática

Curso 2022-2023

**Sistemas escalables en entornos distribuidos.
Introducción a Spark**

Alejandro Calderón Mateos, Jaime Pons Bailly-Bailliere,
acaldero@inf.uc3m.es jaime@lab.inf.uc3m.es

Félix García Carballeira
fgcarball@inf.uc3m.es



Contenidos



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

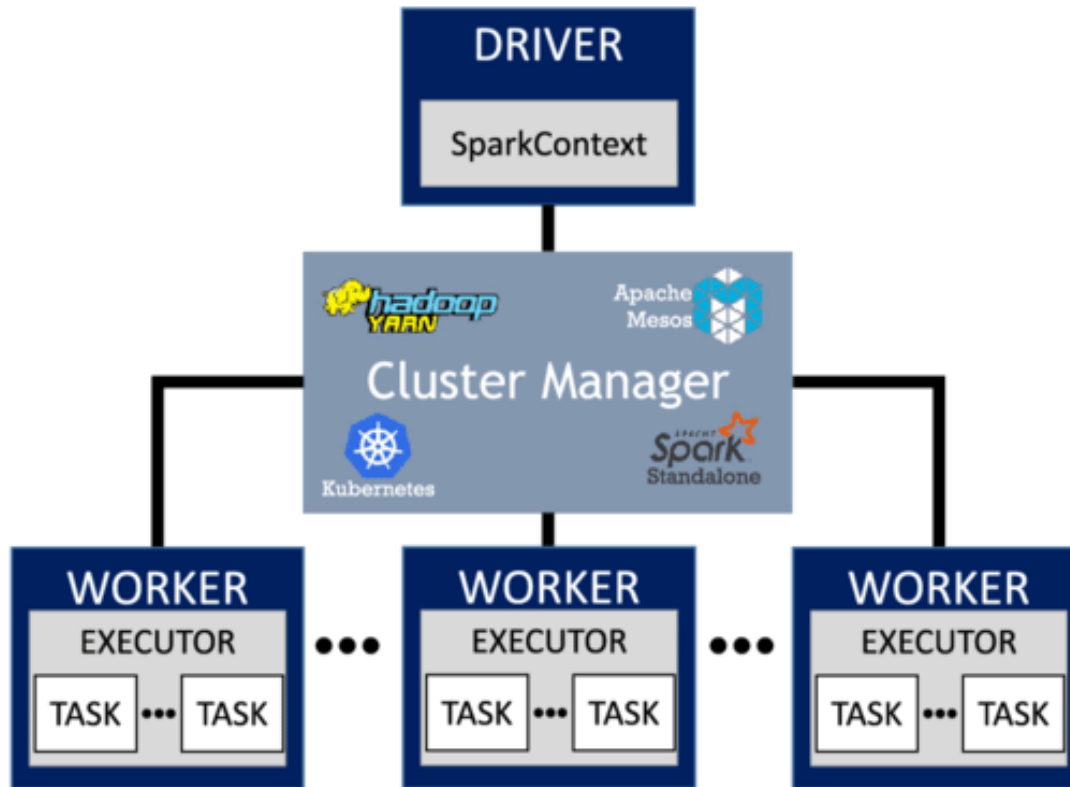


Arquitectura: capas

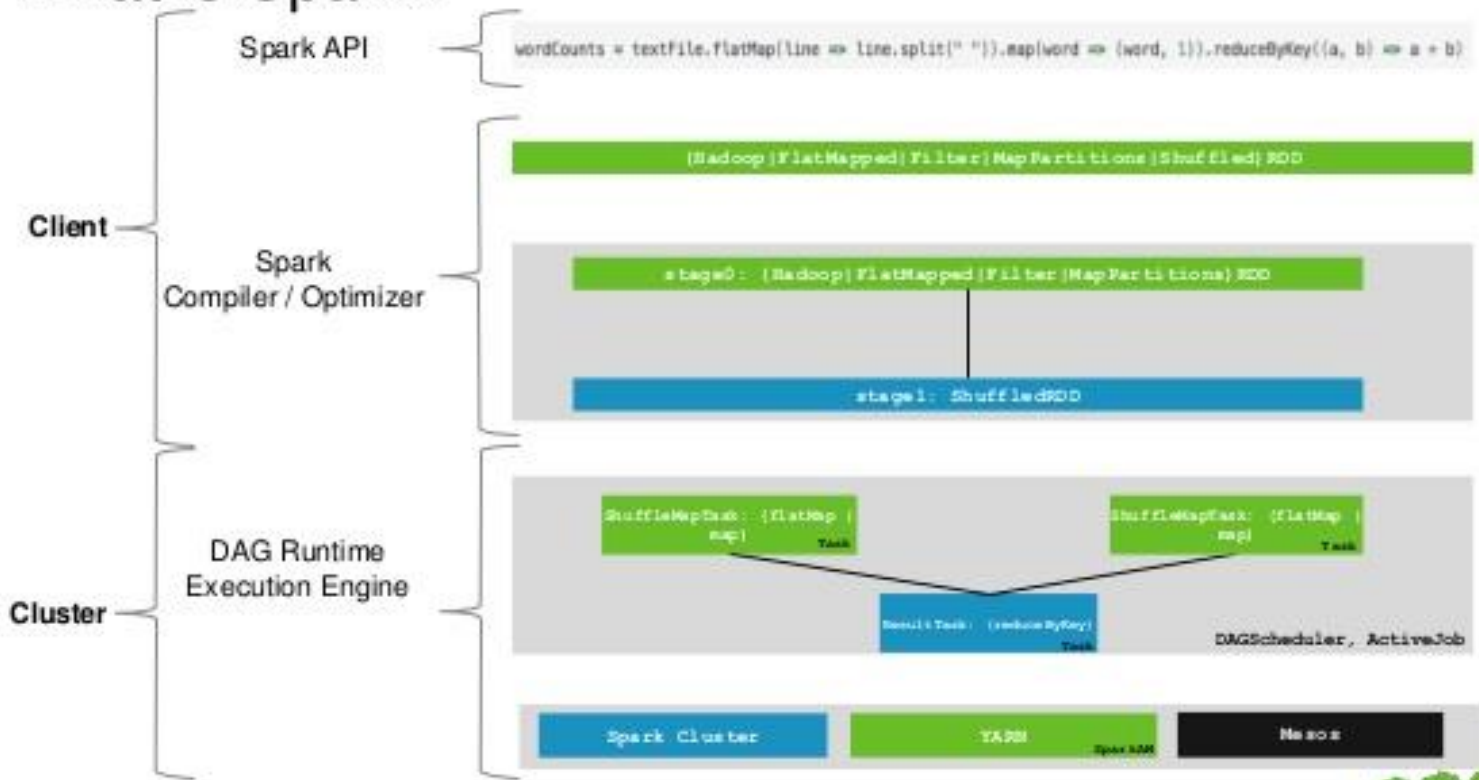


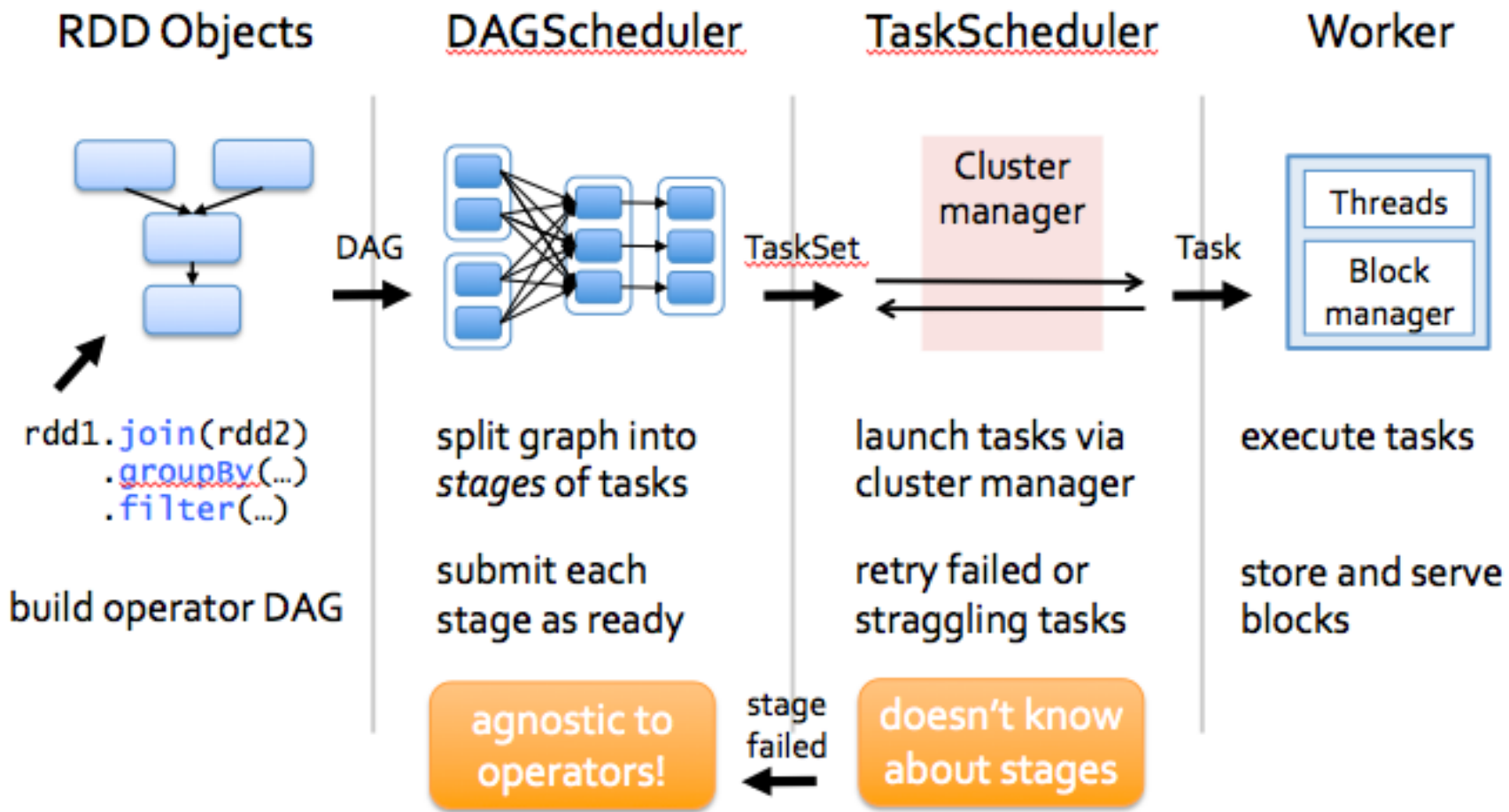


Arquitectura: despliegue



What is Spark?





Contenidos



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

Spark, Anaconda y Jupyter

Prerequisitos

Instalación

Prueba básica



```
acaldero@h1:~$ du -mh -s .  
3,9G .
```


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



Download

Libraries ▾

Documentation ▾

Examples

Community ▾

Developers ▾

Download Apache Spark™

1. Choose a Spark release: ▾
2. Choose a package type: ▾
3. Download Spark: [spark-3.2.0-bin-hadoop3.2.tgz](#)
4. Verify this release using the 3.2.0 [signatures](#), [checksums](#) and [project release KEYS](#).

Note that, Spark 2.x is pre-built with Scala 2.11 except version 2.4.2, which is pre-built with Scala 2.12. Spark 3.0+ is pre-built with Scala 2.12.

Spark

Prerequisitos

Instalación

Prueba básica



```
acaldero@h1:~$ wget https://d1cdn.apache.org/spark/spark-3.3.0/spark-3.3.0-bin-hadoop3.tgz
```

...

```
2021-11-12 12:40:44 (6,02 MB/s) - "spark-3.3.0-bin-hadoop3.tgz" guardado [...]
```



```
acaldero@h1:~$ tar xzf spark-3.3.0-bin-hadoop3.tgz
```

```
acaldero@h1:~$ ls -las spark-3.3.0-bin-hadoop3
```

```
total 164
```

```
 4 drwxr-xr-x 14 dsd dsd  4096 nov 10 15:06 .
 4 drwxr-xr-x 22 dsd dsd  4096 nov 14 03:34 ..
 4 drwxr-xr-x  2 dsd dsd  4096 oct  6 15:18 bin
 4 drwxr-xr-x  2 dsd dsd  4096 oct  6 15:18 conf
 4 drwxr-xr-x  5 dsd dsd  4096 oct  6 15:18 data
 4 drwxr-xr-x  4 dsd dsd  4096 oct  6 15:18 examples
16 drwxr-xr-x  2 dsd dsd 16384 oct  6 15:18 jars
 4 drwxr-xr-x  4 dsd dsd  4096 oct  6 15:18 kubernetes
24 -rw-r--r--  1 dsd dsd 22878 oct  6 15:18 LICENSE
 4 drwxr-xr-x  2 dsd dsd  4096 oct  6 15:18 licenses
 4 drwxrwxr-x  2 dsd dsd  4096 nov 10 15:07 logs
60 -rw-r--r--  1 dsd dsd 57677 oct  6 15:18 NOTICE
 4 drwxr-xr-x  9 dsd dsd  4096 oct  6 15:18 python
 4 drwxr-xr-x  3 dsd dsd  4096 oct  6 15:18 R
 8 -rw-r--r--  1 dsd dsd  4512 oct  6 15:18 README.md
 4 -rw-r--r--  1 dsd dsd   167 oct  6 15:18 RELEASE
 4 drwxr-xr-x  2 dsd dsd  4096 oct  6 15:18 sbin
 4 drwxr-xr-x  2 dsd dsd  4096 oct  6 15:18 yarn
```

Spark

Prerequisitos

Instalación

Prueba básica



```
acaldero@h1:~/spark-3.3.0-bin-hadoop3$ ./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 3.2.0
```

```
17/10/17 01:02:42 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using  
buitin-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
```

```
...
```

```
17/10/17 01:02:45 INFO DAGScheduler: Job 0 finished: reduce at SparkPi.scala:38, took 0,687226 s
```

```
Pi is roughly 3.1418622837245676
```

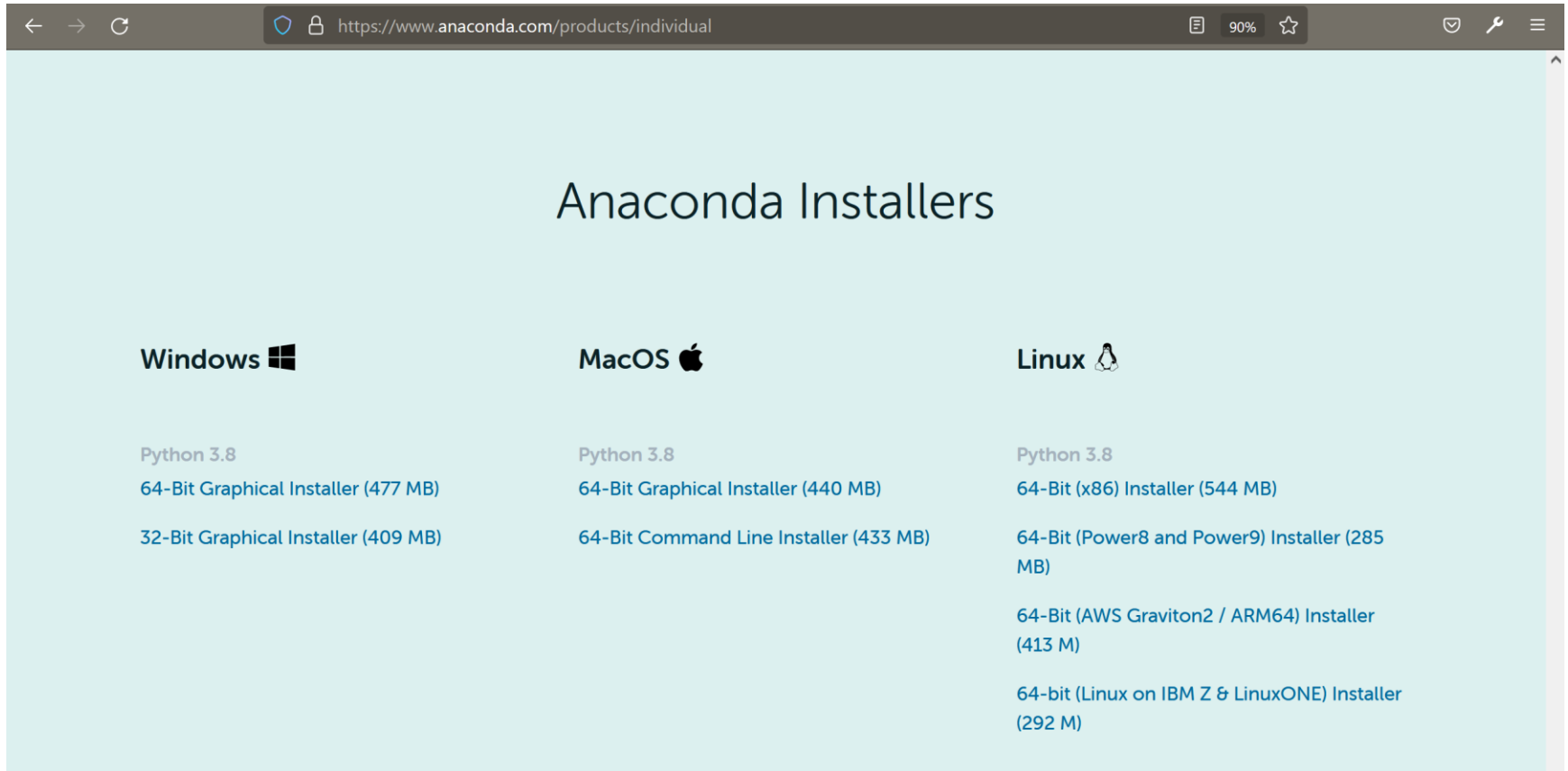
```
17/10/17 01:02:45 INFO SparkUI: Stopped Spark web UI at http://10.0.2.15:4040
```

```
...
```

Anaconda

Instalación

Prueba básica



The screenshot shows a web browser window with the URL <https://www.anaconda.com/products/individual>. The page content is titled "Anaconda Installers" and is organized into three columns for different operating systems: Windows, MacOS, and Linux. Each column lists available installers for Python 3.8, including graphical and command-line versions with their respective file sizes.

Operating System	Python Version	Installer Type	File Size
Windows	Python 3.8	64-Bit Graphical Installer	477 MB
		32-Bit Graphical Installer	409 MB
MacOS	Python 3.8	64-Bit Graphical Installer	440 MB
		64-Bit Command Line Installer	433 MB
Linux	Python 3.8	64-Bit (x86) Installer	544 MB
		64-Bit (Power8 and Power9) Installer	285 MB
		64-Bit (AWS Graviton2 / ARM64) Installer	413 M
		64-bit (Linux on IBM Z & LinuxONE) Installer	292 M

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

Anaconda

Instalación

Prueba básica



```
acaldero@h1:~$ wget https://repo.anaconda.com/archive/Anaconda3-2021.05-Linux-x86_64.sh
```

...

```
2018-11-18 15:12:23 (5,57 MB/s) - "Anaconda3-2021.05-Linux-x86_64.sh" guardado [...]
```



```
acaldero@h1:~$ chmod a+x Anaconda3-2021.05-Linux-x86_64.sh
```

```
acaldero@h1:~$ ./Anaconda3-2021.05-Linux-x86_64.sh
```

```
Welcome to Anaconda3 2021.05 (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: .....
```

...

Jupyter

Instalación

Prueba básica



```
acaldero@h1:~$ conda install jupyter
```

```
Fetching package metadata .....
```

```
Solving package specifications: .....
```

```
# All requested packages already installed.
```

```
# packages in environment at /home/acaldero/anaconda2:
```

```
#
```

```
jupyter                1.0.0                py27h9d2e098_2 ...
```



```
acaldero@h1:~$ jupyter notebook
```

```
[I 18:32:31.686 NotebookApp] [nb_conda_kernels] enabled, 2 kernels found
```

```
[I 18:32:31.792 NotebookApp] ✓ nbpresent HTML export ENABLED
```

```
[W 18:32:31.792 NotebookApp] X nbpresent PDF export DISABLED: No module named nbbrowserpdf.exporters.pdf
```

```
[I 18:32:31.796 NotebookApp] [nb_conda] enabled
```

```
[I 18:32:32.336 NotebookApp] [nb_anacondacloud] enabled
```

```
[I 18:32:32.338 NotebookApp] Serving notebooks from local directory: /home/acaldero
```

```
[I 18:32:32.338 NotebookApp] 0 active kernels
```

```
[I 18:32:32.338 NotebookApp] The Jupyter Notebook is running at: http://localhost:8888/
```

```
[I 18:32:32.338 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
```

```
...
```

Spark, Anaconda y Jupyter

Configuración



```
acaldero@h1:~$ ln -s spark-3.3.0-bin-hadoop3 spark
acaldero@h1:~$ echo "export PATH=$PATH:/home/acaldero/spark/bin" >> .profile
acaldero@h1:~$ echo "export PYSARK_DRIVER_PYTHON=ipython" >> .profile
acaldero@h1:~$ echo "export PYSARK_DRIVER_PYTHON_OPTS='notebook' pyspark">> .profile
acaldero@h1:~$ source .profile
```

...

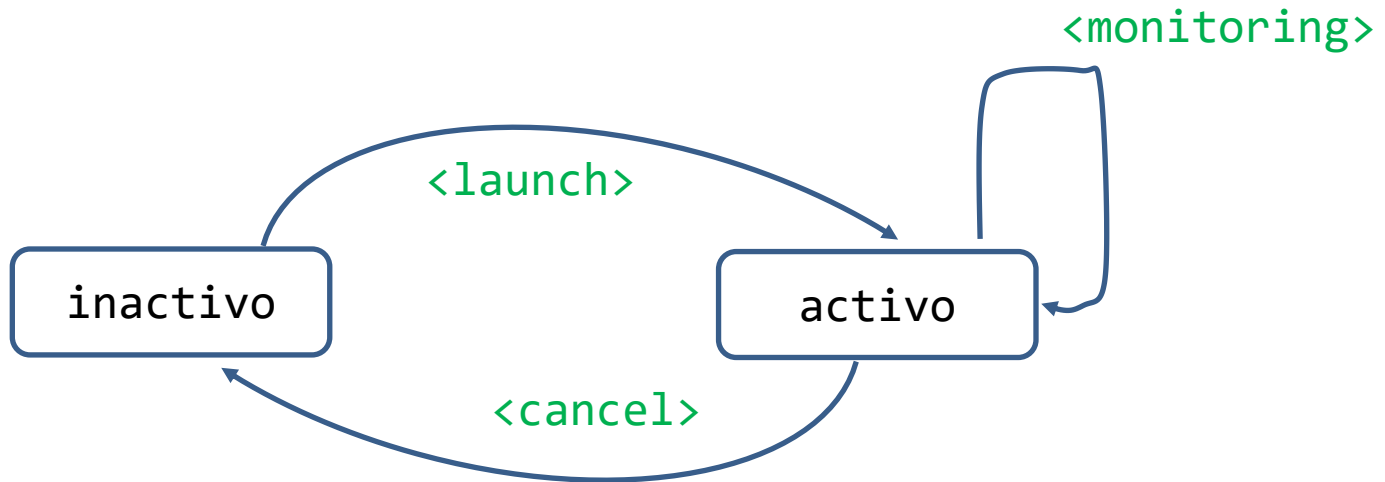
Contenidos



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

Spark

Funcionamiento General

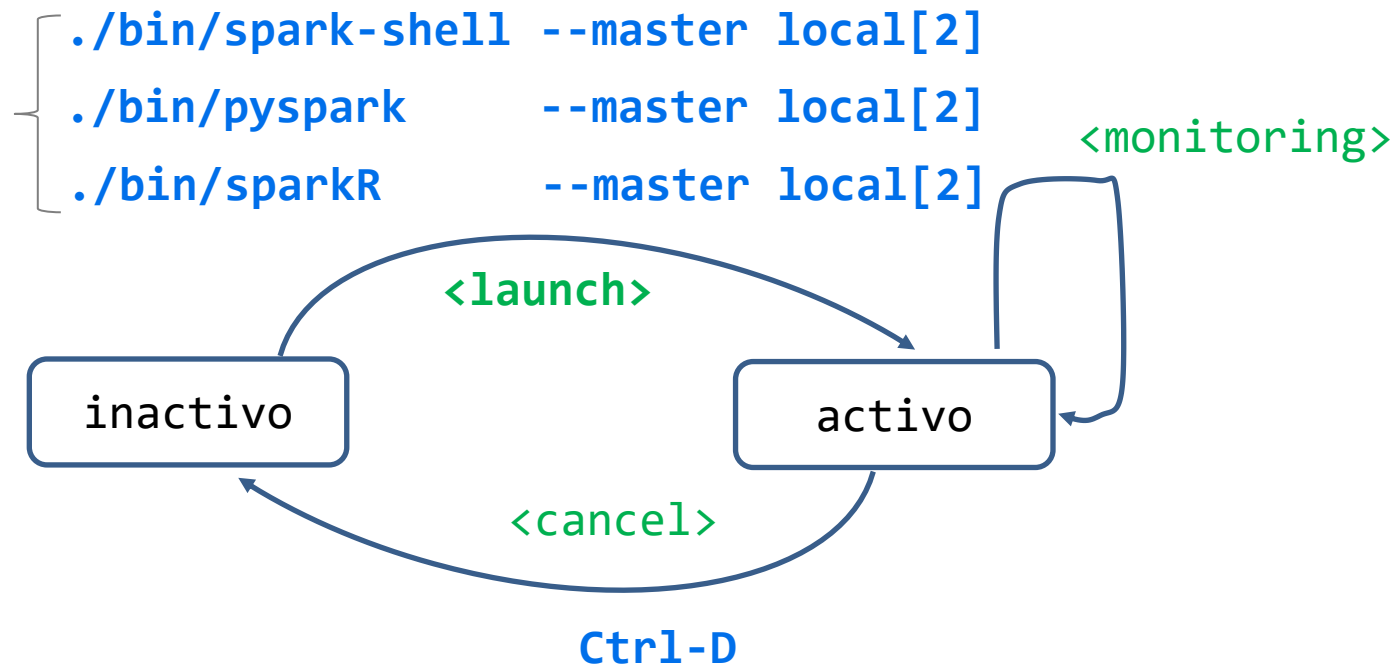


Spark: nodo autónomo

shell-interactivo

submit

libro-interactivo



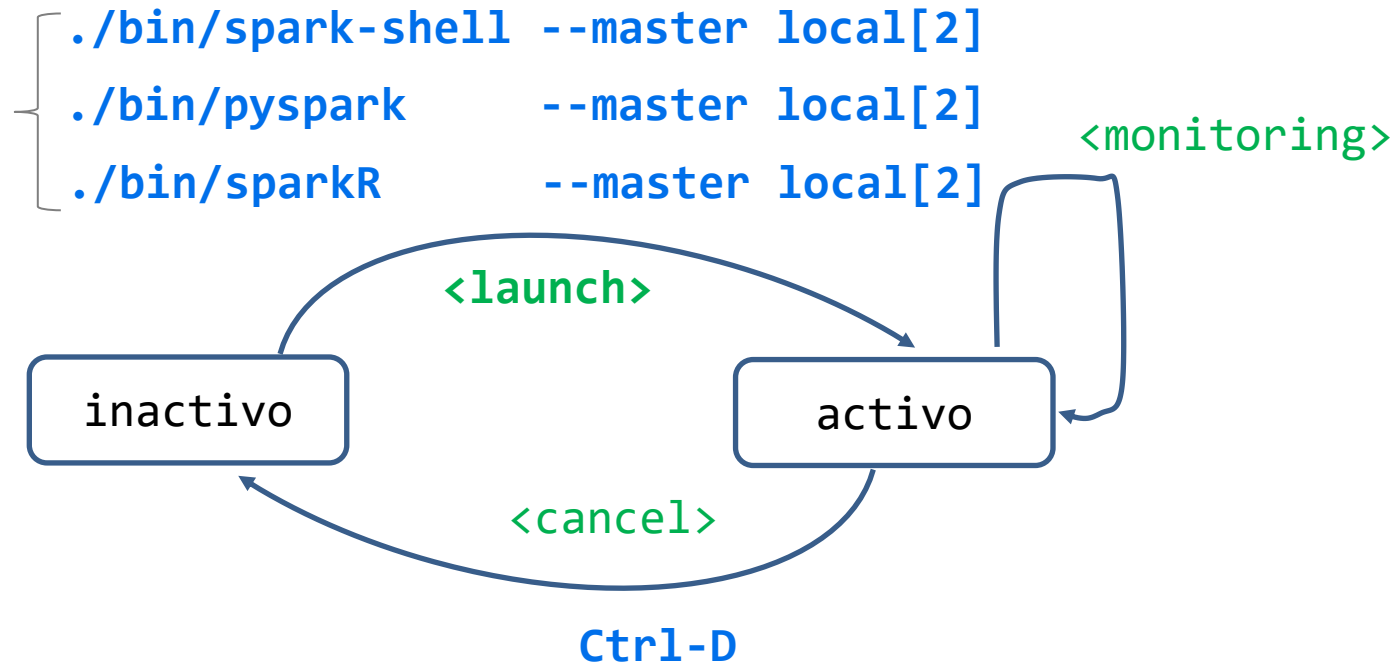
Spark: nodo autónomo

shell-interactivo

submit

libro-interactivo

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



Spark: nodo autónomo

shell- interactivo

submit

libro- interactivo



```
acaldero@h1:~/spark$ ./bin/pyspark
```

```
Python 3.8.12 (default, Oct 12 2021, 13:49:34)
[GCC 7.5.0] :: Anaconda, Inc. on linux
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).
21/11/14 04:04:48 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using
    builtin-java classes where applicable
Welcome to

    ____
   /  _ \__  __  ___  /  _ \
  /  \ \  / /_ /_/  \ /  \ \
 /_ / / ._/ \ ,_ /_ /_ /_ / \ \
    /_/

       version 3.2.0

Using Python version 3.8.12 (default, Oct 12 2021 13:49:34)
Spark context Web UI available at http://10.0.2.15:4040
Spark context available as 'sc' (master = local[*], app id = local-1636859089934).
SparkSession available as 'spark'.
>>>
```

Spark: nodo autónomo

shell-interactivo

submit

libro-interactivo

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()
>>>
```

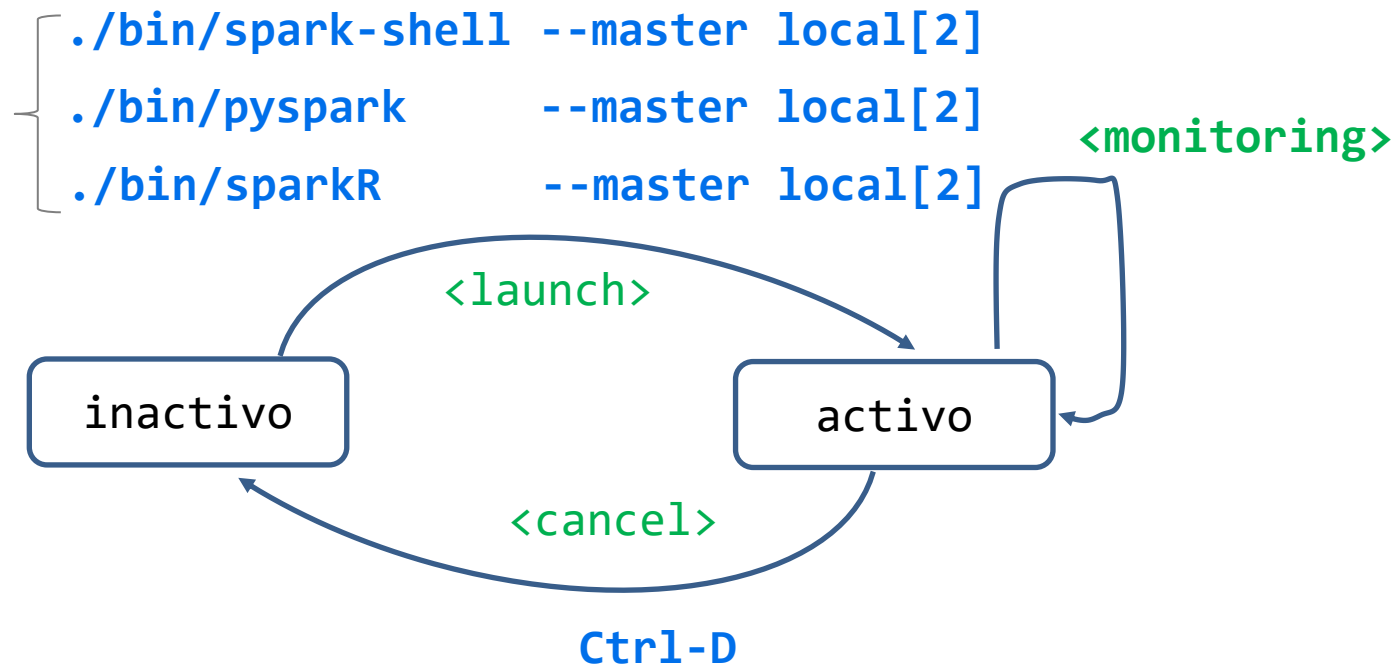


Spark: nodo autónomo

shell-interactivo

submit

libro-interactivo



Spark: nodo autónomo

shell- interactivo

submit

libro- interactivo

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
..

Spark: nodo autónomo

shell-iteractivo

submit

libro-iteractivo

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-iteractivo

submit

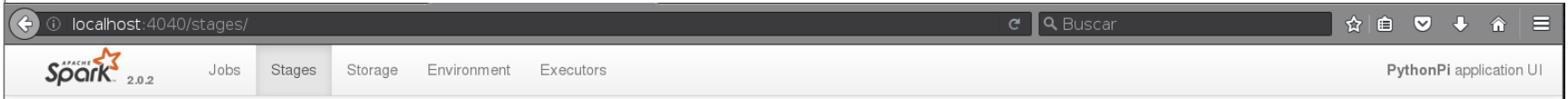
libro-iteractivo



localhost:4040/storage/ PythonPi application UI

APACHE spark 2.0.2 Jobs Stages Storage Environment Executors

Storage



localhost:4040/stages/ PythonPi application UI

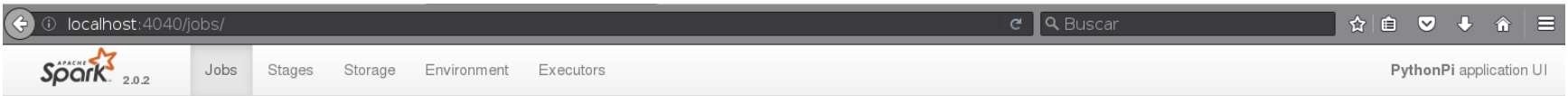
APACHE spark 2.0.2 Jobs Stages Storage Environment Executors

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/ PythonPi application UI

APACHE spark 2.0.2 Jobs Stages Storage Environment Executors

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

Executors	Stages	Time
Added	Completed	27 November 14:49
Removed	Failed	
	Active	
		34
		35
		36
		37
		38
		39
		40
	reduce	41

DAG Visualization

```

  graph TD
    subgraph Stage_0 [Stage 0]
      A[parallelize] --> B[ ]
    end
  
```

Completed Stages (1)

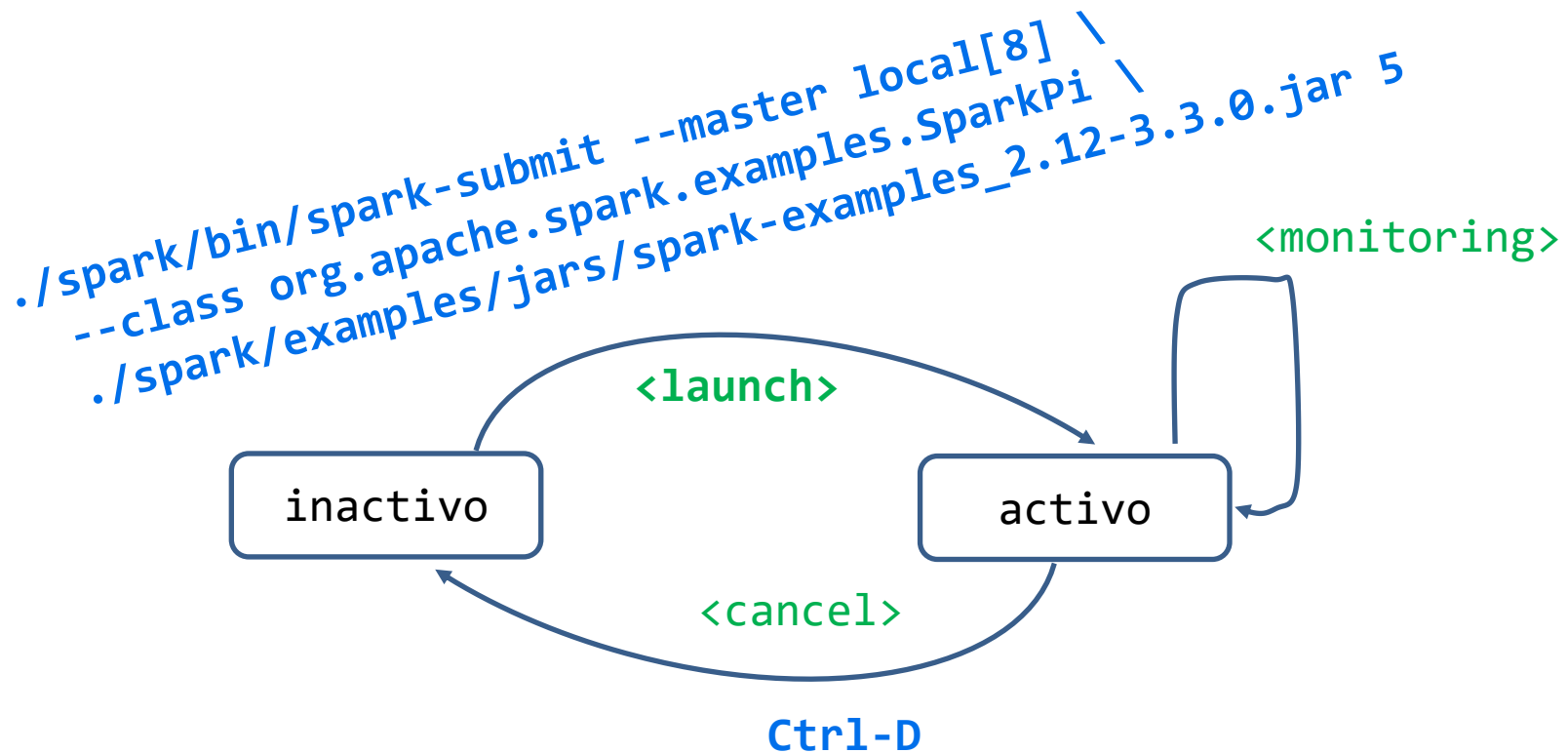
Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
0	reduce at <stdin>:1	2016/11/27 14:49:41	0,3 s	2/2				

Spark: nodo autónomo

shell-interactivo

submit

libro-interactivo



Spark: nodo autónomo

shell-interactivo

submit

libro-interactivo



```
acaldero@h1:~$ mkdir work
acaldero@h1:~$ cd work
acaldero@h1:~$ wget https://www.gutenberg.org/files/2000/2000-0.txt
```



```
acaldero@h1:~$ pyspark
[TerminalPythonApp] WARNING | Subcommand `ipython notebook` is deprecated and will be removed in future versions.
[TerminalPythonApp] 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] X nbpresent PDF export DISABLED: No module named nbbrowserpdf.exporters.pdf
[I 18:48:15.018 NotebookApp] [nb_conda] enabled
...
```

Spark: nodo autónomo

shell-interactivo

submit

libro-iteractivo



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

The screenshot shows a Jupyter Notebook interface in a browser window. The address bar shows 'localhost:8888/notebooks/test-1.ipynb'. The notebook title is 'test-1' with a note 'Last Checkpoint: a few seconds ago (unsaved changes)'. The interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help) and a toolbar with icons for file operations and execution. The main content area shows a code cell with the input 'sc' and its output: 'SparkContext'. The output is displayed as a dictionary-like structure with keys: 'Spark UI', 'Version' (v3.0.1), 'Master' (local[*]), 'AppName', and 'PySparkShell'. Below the code cell is an empty input field for the next command.

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")
  
```

The screenshot shows a Jupyter Notebook running in a browser at localhost:8888. The notebook is titled 'test-1' and shows two code cells. The first cell contains the code 'sc', which has been executed, resulting in an output of 'SparkContext' with details like 'Version v3.0.1' and 'Master local[*]'. The second cell contains the code from the previous block, which has also been executed. The interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help) and a toolbar with icons for running and saving code.

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])
```

The screenshot shows a Jupyter Notebook interface with the following content:

```
localhost:8888/notebooks/test-1.ipynb 130% ... ☆
jupyter test-1 Last Checkpoint: 3 minutes ago (unsaved changes) Logout
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3
+ ↻ ↺ ↻ ⬆ ⬇ ▶ Run ■ C ▶ Code
```

```
In [1]: sc
Out[1]: SparkContext
Spark UI
Version
v3.0.1
Master
local[*]
AppName
PySparkShell

In [3]: myRDD = sc.textFile("file:///home/dsd/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])
Out[3]: [('que', 19429),
('de', 17988),
('y', 15894),
('la', 10200),
('a', 9575),
(' ', 9504),
('el', 7957),
('en', 7898),
('no', 5611),
('se', 4690)]

In [ ]:
```


Contenidos



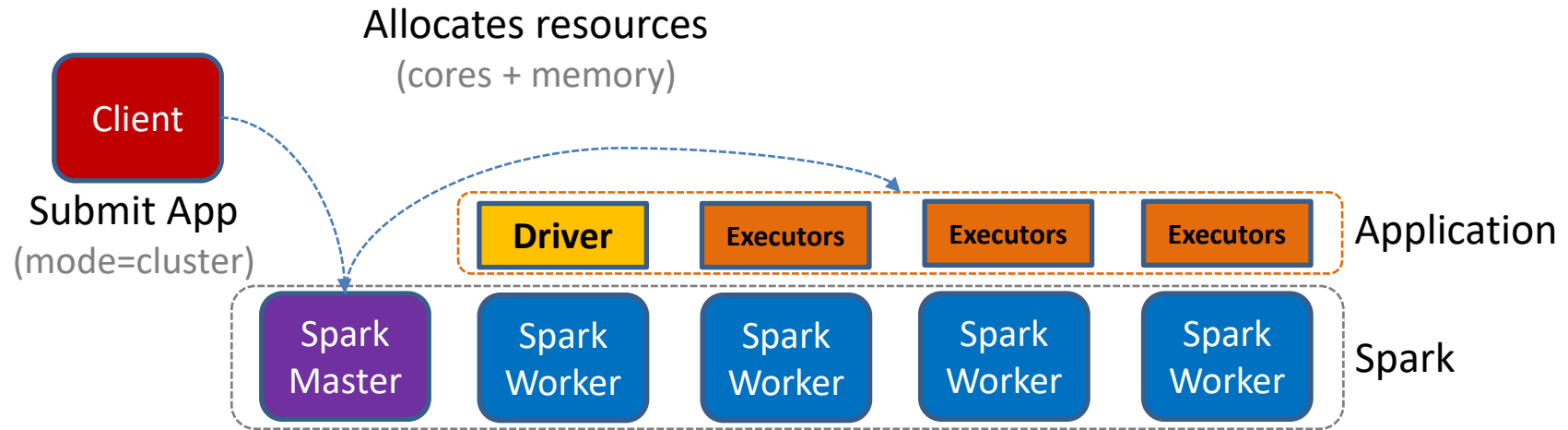
- Introducción
- ***Hand-on***
 - Pre-requisitos e instalación
 - Nodo autónomo
 - **Cluster**
- *Benchmarking*

Spark: cluster privado

Prerequisitos

Instalación

Uso básico

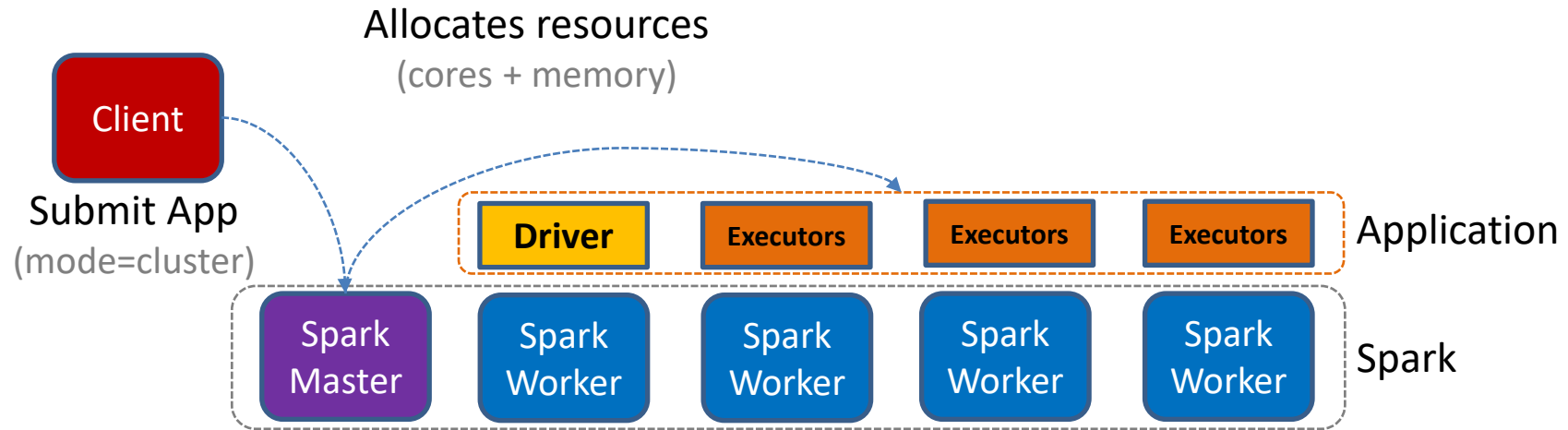


Spark: cluster privado

Prerequisitos

Instalación

Uso básico



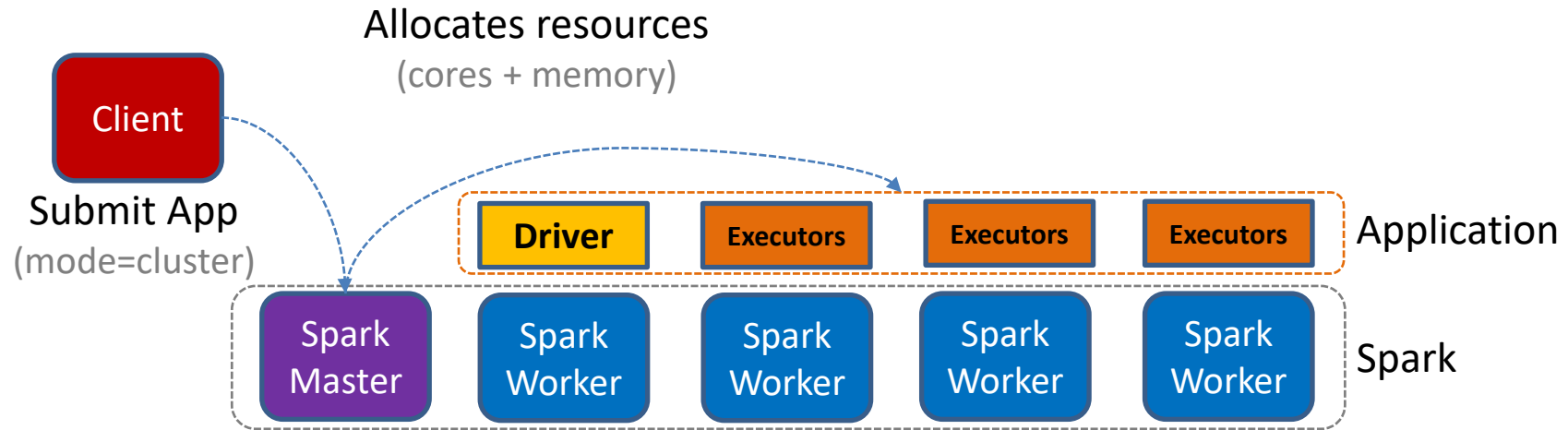
```
acaldero@h1:~$ echo "127.0.0.1 master1" >> /etc/hosts
acaldero@h1:~$ echo "127.0.0.1 worker1" >> /etc/hosts
acaldero@h1:~$ echo "127.0.0.1 worker2" >> /etc/hosts
```

Spark: cluster privado

Prerequisitos

Instalación

Uso básico



```
acaldero@h1:~$ echo "worker1" >> spark/conf/workers  
acaldero@h1:~$ echo "worker2" >> spark/conf/workers
```



```
acaldero@h1:~$ : Spark en todos los nodos (si fuera necesario)  
acaldero@h1:~$ scp -r spark acaldero@worker[1-2]:~/  
...
```

Spark: cluster privado

Prerequisites

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

Prerequisites

Instalación

Uso básico



```
acaldero@h1:/home/acaldero$ scp .ssh/id_rsa.pub acaldero@worker2:~/.ssh/authorized_keys  
Password:
```

...



```
acaldero@h1:/home/acaldero$ ssh worker2  
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.
```

...



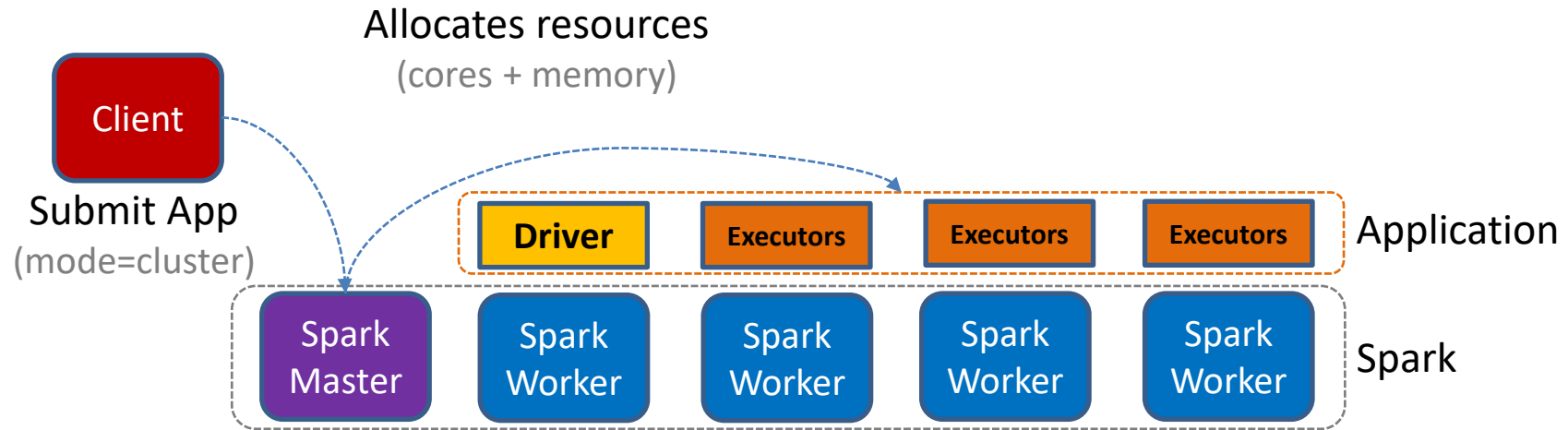
```
acaldero@worker2:~$ exit  
logout
```

Spark: cluster privado

Prerequisitos

Instalación

Uso básico



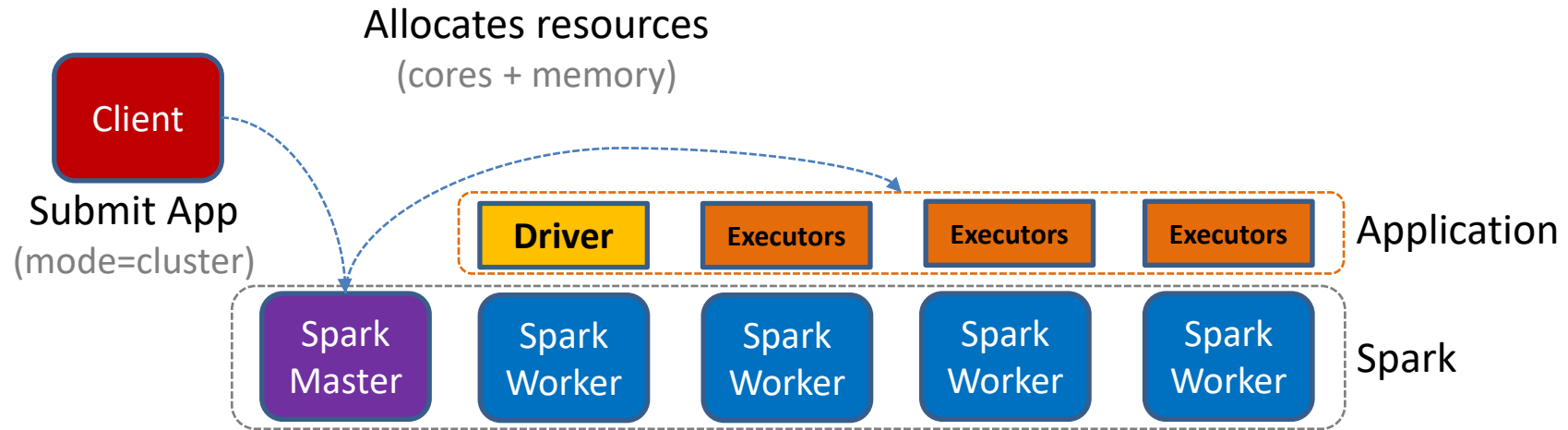
```
acaldero@h1:~$ : Ir al nodo master
acaldero@h1:~$ ssh acaldero@master1
acaldero@master1:~$ ./spark/sbin/start-all.sh
localhost: starting org.apache.spark.deploy.worker.Worker, logging to /home/acaldero/spark/logs/spark-acaldero-org.apache.spark.deploy.worker.Worker-1-ws1.out
localhost: starting org.apache.spark.deploy.worker.Worker,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

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
```

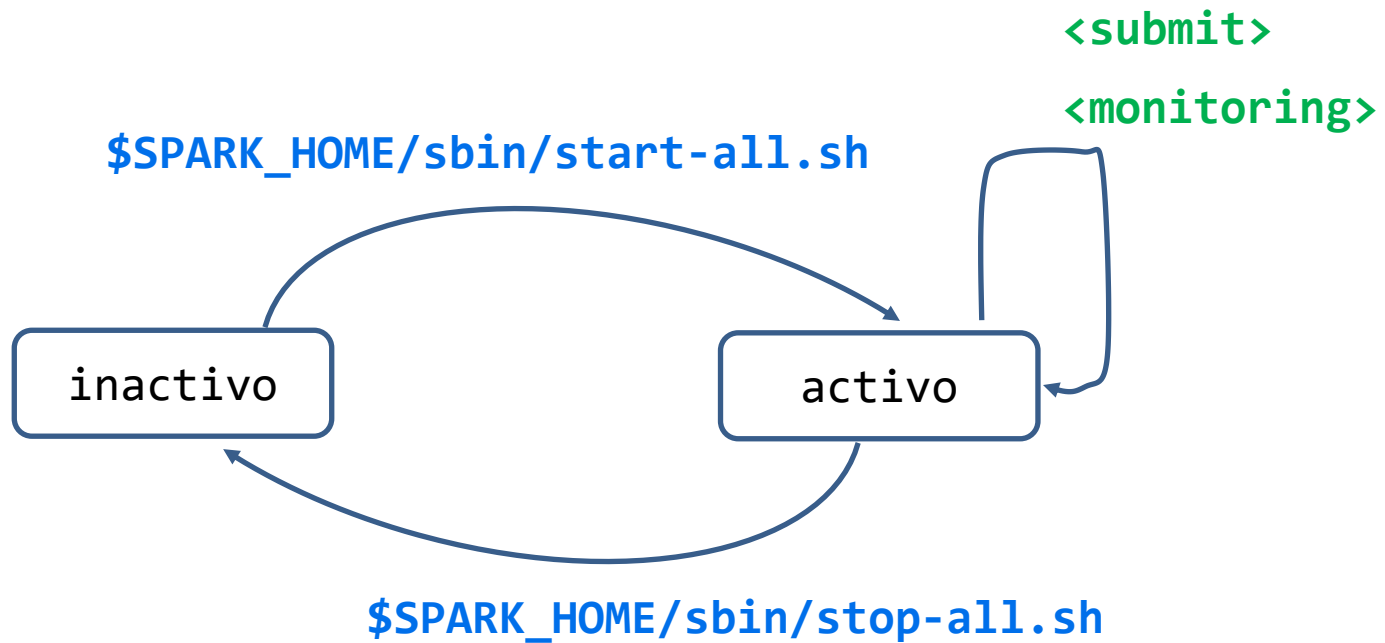
```
stopping org.apache.spark.deploy.master.Master
```


Spark: cluster privado

Prerequisitos

Instalación

Uso básico

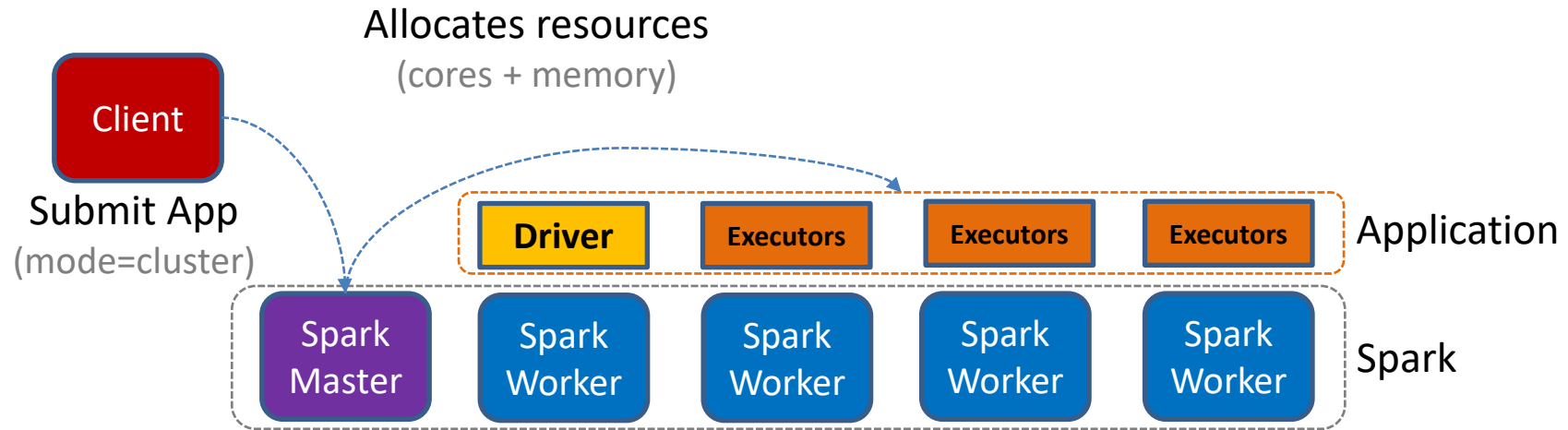


Spark: cluster privado

Prerequisites

Installation

Use basic



```
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...
```

```
...
```

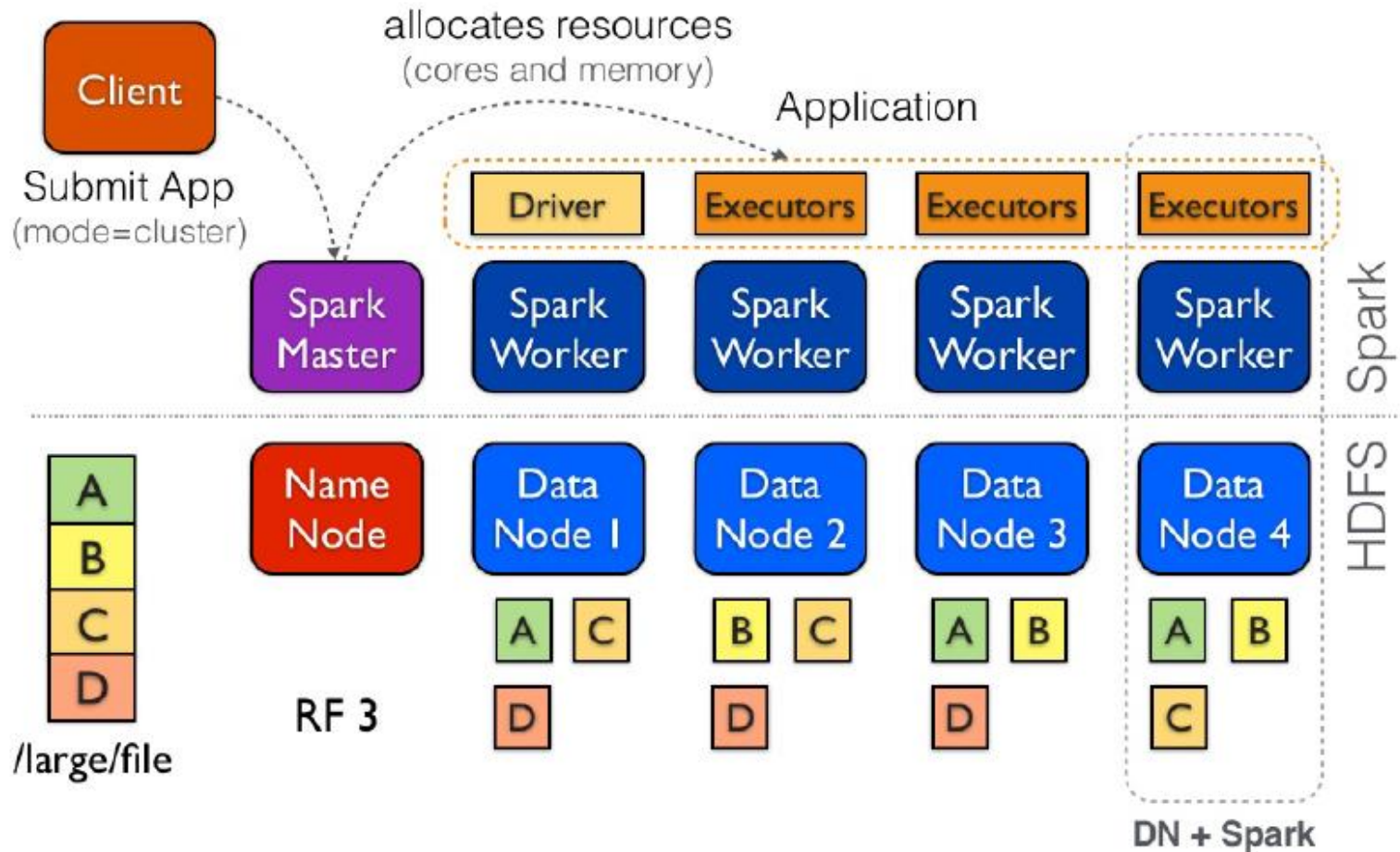
```
scala> exit
```

Spark: cluster privado

Prerequisites

Instalación

Uso básico



Contenidos



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

Benchmarking

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

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
 - ...

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)

Bibliografía: tutoriales

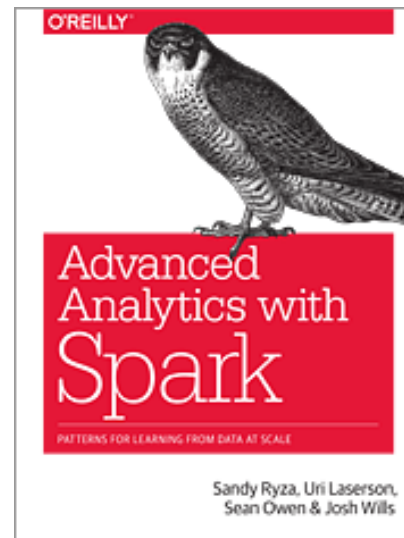
- Página Web oficial:
 - <http://spark.apache.org/>
- Introducción a cómo funciona Spark:
 - <http://spark.apache.org/docs/latest/quick-start.html>
- Tutorial de cómo instalar y usar Spark:
 - <http://spark.apache.org/docs/latest/index.html>
 - <http://spark.apache.org/docs/latest/configuration.html>

Bibliografía: libro

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



Holden Karau, Andy Konwinski,
Patrick Wendell & Matei Zaharia



Sandy Ryza, Uri Laserson,
Sean Owen & Josh Wills

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.



Sistemas Paralelos y Distribuidos
Máster en Ciencia y Tecnología Informática
Diseño de Sistemas Distribuidos
Máster en Ingeniería Informática

Curso 2022-2023

**Sistemas escalables en entornos distribuidos.
Introducción a Spark**

Alejandro Calderón Mateos, Jaime Pons Bailly-Bailliere,
acaldero@inf.uc3m.es jaime@lab.inf.uc3m.es

Félix García Carballeira
fgcarball@inf.uc3m.es

