

Guia de instalacion Apache Spark

En la siguiente guia aprenderemos como instalar Apache Spark en una maquina con sistema operativo Windows.

Descarga de Apache Spark

Como primer paso necesitamos bajar la version mas reciente de apache spark

[Descargar Spark \(http://spark.apache.org/downloads.html\)](http://spark.apache.org/downloads.html)

Download Apache Spark™

1. Choose a Spark release:
2. Choose a package type:
3. Download Spark: [spark-3.0.0-preview-bin-hadoop2.7.tgz](#)
4. Verify this release using the 3.0.0-preview [signatures](#), [checksums](#) and [project release KEYS](#).

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

Cuando finalice la descarga creamos un folder donde descomprimir nuestro archivo de Spark por ejemplo

C:\Spark

Descarga de Java 8

Para la ejecucion de Spark como requisito necesitamos una maquina virtual de java corriendo en nuestra computadora por lo que necesitamos instalar Java 8.

[Descargar Java 8 \(https://www.oracle.com/technetwork/java/javase/jre8-downloads-2133155.html\)](https://www.oracle.com/technetwork/java/javase/jre8-downloads-2133155.html)

Java SE Runtime Environment 8u231

You must accept the [Oracle Technology Network License Agreement for Oracle Java SE](#) to download this software.

☐ Accept License Agreement
 ☒ Decline License Agreement














Product / File Description	File Size	Download
Linux x86	67.52 MB	jre-8u231-linux-i586.rpm
Linux x86	83.26 MB	jre-8u231-linux-i586.tar.gz
Linux x64	66.62 MB	jre-8u231-linux-x64.rpm
Linux x64	82.44 MB	jre-8u231-linux-x64.tar.gz
Mac OS X x64	79.91 MB	jre-8u231-macosx-x64.dmg
Mac OS X x64	71.46 MB	jre-8u231-macosx-x64.tar.gz
Solaris SPARC 64-bit	52.15 MB	jre-8u231-solaris-sparcv9.tar.gz
Solaris x64	49.97 MB	jre-8u231-solaris-x64.tar.gz
Windows x86 Online	1.97 MB	jre-8u231-windows-i586-iftw.exe
Windows x86 Offline	64.93 MB	jre-8u231-windows-i586.exe
Windows x86	67.39 MB	jre-8u231-windows-i586.tar.gz
Windows x64	72.8 MB	jre-8u231-windows-x64.exe
Windows x64	72.45 MB	jre-8u231-windows-x64.tar.gz

Descargar Winutils

Para la correcta ejecucion de Spark necesitamos obtener el paquete de Winutils de Hadoop 2.7, esto es necesario para asegurarnos de que spark funcione correctamente en windows, en el siguiente link se explica el por que la necesidad de este paquete [Windows problems](https://cwiki.apache.org/confluence/display/HADOOP2/WindowsProblems) (<https://cwiki.apache.org/confluence/display/HADOOP2/WindowsProblems>)

Ocupamos descargar el repositorio de [Aqui](https://github.com/steveloughran/winutils) (<https://github.com/steveloughran/winutils>)

y luego copiamos la carpeta de nuestra version a una ruta conocida por ejemplo "C:\Hadoop-2.7.1"

Name	Date modified	Type	Size
 hadoop-2.6.0	8/1/2019 5:54 AM	File folder	
 hadoop-2.6.3	8/1/2019 5:54 AM	File folder	
 hadoop-2.6.4	8/1/2019 5:54 AM	File folder	
 hadoop-2.7.1	8/1/2019 5:54 AM	File folder	
 hadoop-2.8.0-RC3	8/1/2019 5:54 AM	File folder	
 hadoop-2.8.1	8/1/2019 5:54 AM	File folder	
 hadoop-2.8.3	8/1/2019 5:54 AM	File folder	
 hadoop-3.0.0	8/1/2019 5:54 AM	File folder	
 .gitattributes	8/1/2019 5:54 AM	Text Document	1 KB
 .gitignore	8/1/2019 5:54 AM	Text Document	0 KB
 KEYS	8/1/2019 5:54 AM	File	26 KB
 LICENSE	8/1/2019 5:54 AM	File	12 KB
 README.md	8/1/2019 5:54 AM	MD File	7 KB

Variables de Entorno

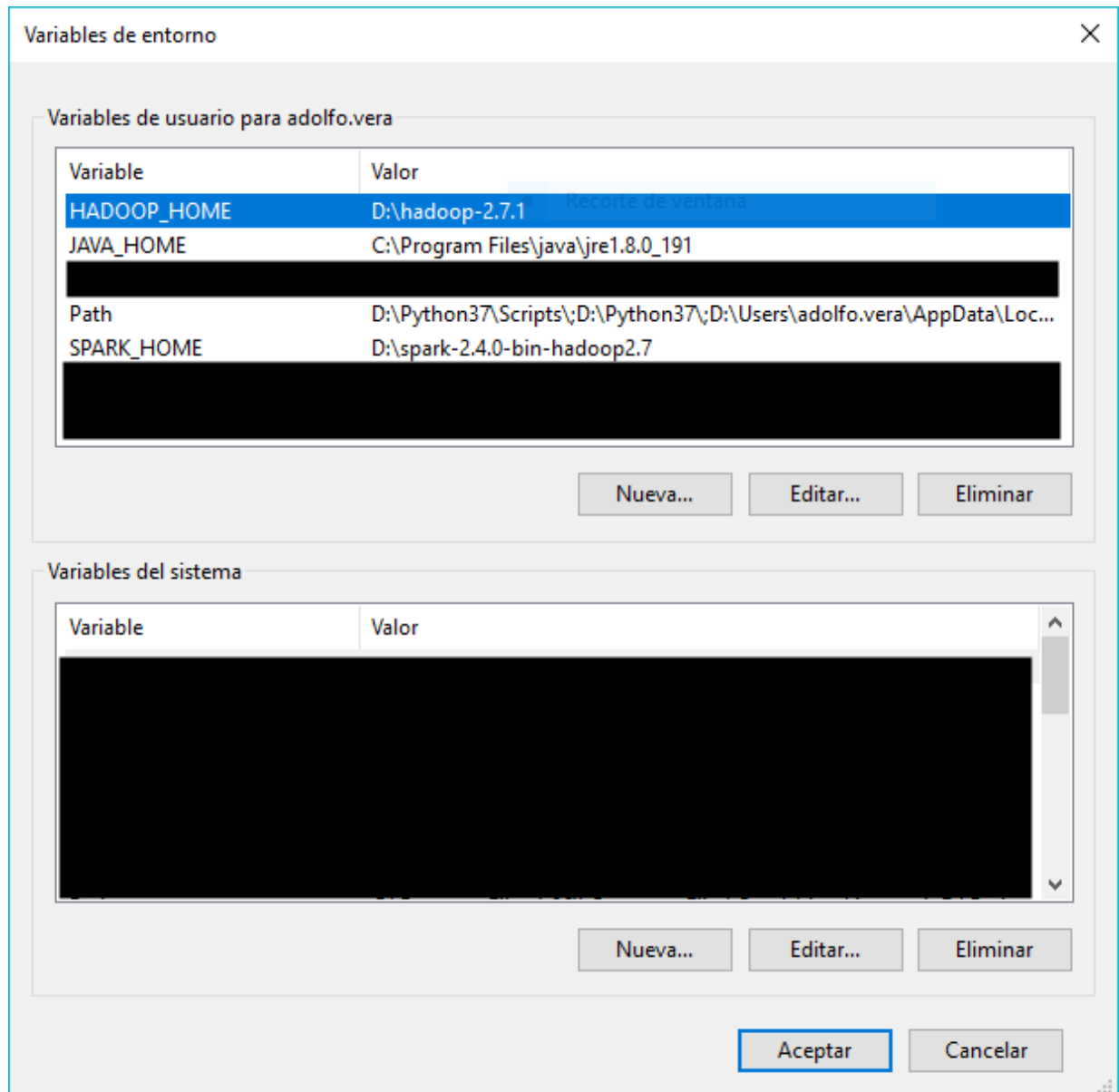
Variables de entorno

En concreto vamos a crear tres variables de entorno...

SPARK_HOME. Ruta al directorio donde hemos descomprimido el paquete de Apache Spark.

HADOOP_HOME. Apunta al directorio donde hemos copiado la carpeta con la Winutils.

JAVA_HOME. Es el directorio donde se ha instalado el JRE de Java 8



Iniciando Apache Spark

Todo lo que necesitamos para arrancar nuestra instalación de Apache Spark se encuentra dentro de la carpeta bin de Apache Spark.

Ahora en una terminal nos dirigimos al folder donde tenemos nuestro Spark instalado y ejecutamos el siguiente comando

spark-class org.apache.spark.deploy.master.Master

```

PS C:\Spark\spark-3.0.0-preview-bin-hadoop2.7\bin> spark-class org.apache.spark.deploy.master.Master
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
19/11/24 14:54:18 INFO Master: Started daemon with process name: 14184@DESKTOP-8TUDRNH
19/11/24 14:54:20 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
19/11/24 14:54:20 INFO SecurityManager: Changing view acls to: Percy
19/11/24 14:54:20 INFO SecurityManager: Changing modify acls to: Percy
19/11/24 14:54:20 INFO SecurityManager: Changing view acls groups to:
19/11/24 14:54:20 INFO SecurityManager: Changing modify acls groups to:
19/11/24 14:54:20 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(Percy); groups:
19/11/24 14:54:23 INFO Utils: Successfully started service 'sparkMaster' on port 7077.
19/11/24 14:54:23 INFO Master: Starting Spark master at spark://192.168.2.24:7077
19/11/24 14:54:23 INFO Master: Running Spark version 3.0.0-preview
19/11/24 14:54:23 INFO Utils: Successfully started service 'MasterUI' on port 8080.
19/11/24 14:54:23 INFO MasterWebUI: Bound MasterWebUI to 0.0.0.0, and started at http://DESKTOP-8TUDRNH:8080
19/11/24 14:54:24 INFO Master: I have been elected leader! New state: ALIVE
19/11/24 15:02:25 INFO Master: Registering worker 192.168.2.24:50398 with 8 cores, 14.9 GiB RAM

```

ahora lo que nos interesa es esta URL donde nuestro master de spark esta corriendo, lo necesitamos para iniciar un **worker** nuestro clueter de Spark con el siguiente comando

spark-class org.apache.spark.deploy.worker.Worker spark://192.168.2.24:7077

```

PS C:\Spark\spark-3.0.0-preview-bin-hadoop2.7\bin> spark-class org.apache.spark.deploy.worker.Worker spark://192.168.2.24:7077
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
19/11/24 15:02:22 INFO Worker: Started daemon with process name: 8420@DESKTOP-8TUDRNH
19/11/24 15:02:22 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
19/11/24 15:02:23 INFO SecurityManager: Changing view acls to: Percy
19/11/24 15:02:23 INFO SecurityManager: Changing modify acls to: Percy
19/11/24 15:02:23 INFO SecurityManager: Changing view acls groups to:
19/11/24 15:02:23 INFO SecurityManager: Changing modify acls groups to:
19/11/24 15:02:23 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(Percy); groups:
19/11/24 15:02:24 INFO Utils: Successfully started service 'sparkWorker' on port 50398.
19/11/24 15:02:24 INFO Worker: Starting Spark worker 192.168.2.24:50398 with 8 cores, 14.9 GiB RAM
19/11/24 15:02:24 INFO Worker: Running Spark version 3.0.0-preview
19/11/24 15:02:24 INFO Worker: Spark home: C:\Spark\spark-3.0.0-preview-bin-hadoop2.7
19/11/24 15:02:24 INFO ResourceUtils: =====
19/11/24 15:02:24 INFO ResourceUtils: Resources for spark.worker:
=====
19/11/24 15:02:24 INFO ResourceUtils: =====
19/11/24 15:02:24 INFO Utils: Successfully started service 'WorkerUI' on port 8081.
19/11/24 15:02:24 INFO WorkerWebUI: Bound WorkerWebUI to 0.0.0.0, and started at http://DESKTOP-8TUDRNH:8081
19/11/24 15:02:24 INFO Worker: Connecting to master 192.168.2.24:7077...
19/11/24 15:02:24 INFO TransportClientFactory: Successfully created connection to /192.168.2.24:7077 after 47 ms (0 ms spent in bootstraps)
19/11/24 15:02:25 INFO Worker: Successfully registered with master spark://192.168.2.24:7077

```

Accesando la direccion <http://desktop-8tudrnh:8080/> (<http://desktop-8tudrnh:8080/>) podremos ver el UI de nuestro cluster de Spark

Spark Master at spark://192.168.2.24:7077

URL: spark://192.168.2.24:7077

Active Workers: 1

Cores in use: 8 Total, 0 Used

Memory in use: 14.9 GiB Total, 0 B Used

Resources in use:

Applications: 0 Running, 0 Completed

Drivers: 0 Running, 0 Completed

Status: ALIVE

Workers (1)

Worker Id	Address	State	Cores	Memory	Resources
worker-20191124150224-192.168.2.24-50398	192.168.2.24:50398	ALIVE	8 (0 Used)	14.9 GiB (0 B Used)	

Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------

Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------

Instalar FindSpark

Como parte del ambiente que necesitamos preparar para el estudio de Spark necesitamos instalar el siguiente paquete que utilizaremos en nuestros notebooks.

Abriremos una terminal de nuestro ambiente virtual en Anaconda y ejecutaremos lo siguiente.

conda install -c conda-forge findspark

```
(py3spark) C:\Users\Percy\conda install -c conda-forge findspark
Collecting package metadata (repodata.json): done
Solving environment: done

## Package Plan ##

  environment location: C:\Users\Percy\Anaconda3\envs\py3spark
added / updated specs:
- findspark

The following packages will be downloaded:

package | build | size | channel
-----|-----|-----|-----
ca-certificates-2019.09.11 | hecc5488_0 | 181 KB | conda-forge
certifi-2019.9.11 | py37_0 | 147 KB | conda-forge
findspark-1.3.0 | py_1 | 6 KB | conda-forge
openssl-1.1.1d | hf4de2cd_0 | 4.7 MB | conda-forge
-----|-----|-----|-----
Total: | 5.0 MB

The following NEW packages will be INSTALLED:
findspark conda-forge/noarch::findspark-1.3.0-py_1

The following packages will be UPDATED:
ca-certificates pkgs/main::ca-certificates-2019.5.15-1 --> conda-forge::ca-certificates-2019.09.11-hecc5488_0
certifi pkgs/main::certifi-2019.6.16-py37_1 --> conda-forge::certifi-2019.9.11-py37_0
openssl pkgs/main::openssl-1.1.1c-he774522_1 --> conda-forge::openssl-1.1.1d-hf4de2cd_0

proceed ([y]/n)? y

Downloading and Extracting Packages
ca-certificates-2019.09.11 | 181 KB | ##### 100%
findspark-1.3.0 | 6 KB | ##### 100%
openssl-1.1.1d | 4.7 MB | ##### 100%
certifi-2019.9.11 | 147 KB | ##### 100%
Preparing transaction: done
Verifying transaction: done
Executing transaction: done
```

Instalar PySpark

Para nuestro ambiente de desarrollo y analisis tambien necesitamos del paquete PySpark, el cual instalamos con el siguiente comando.

conda install -c conda-forge pyspark

```
(py3spark) C:\Users\Percy\conda install -c conda-forge pyspark
Collecting package metadata (repodata.json): done
Solving environment: done

## Package Plan ##

  environment location: C:\Users\Percy\Anaconda3\envs\py3spark
added / updated specs:
- pyspark

The following packages will be downloaded:

package | build | size | channel
-----|-----|-----|-----
py4j-0.10.7 | py_2 | 177 KB | conda-forge
pyspark-2.4.4 | py_0 | 284.9 MB | conda-forge
-----|-----|-----|-----
Total: | 285.1 MB

The following NEW packages will be INSTALLED:
py4j conda-forge/noarch::py4j-0.10.7-py_1
pyspark conda-forge/noarch::pyspark-2.4.4-py_0

proceed ([y]/n)? y

Downloading and Extracting Packages
pyspark-2.4.4 | 284.9 MB | ##### 100%
py4j-0.10.7 | 177 KB | ##### 100%
Preparing transaction: done
Verifying transaction: done
Executing transaction: done
```

Prueba

Para comprobar que la instalacion fue correcta por favor ejecutar las dos celdas de codigo siguientes que utilizan y levantan un ambiente de trabajo con los paquetes antes instalados.

Si no tenemos errores el proceso de instalacion fue satisfactorio.

```
In [2]: from pyspark.sql import SparkSession

spark=SparkSession.builder.appName('data_processing').getOrCreate()

import pyspark.sql.functions as F
from pyspark.sql.types import *
```

```
In [8]: import findspark
findspark.init("C:\\Spark\\spark-3.0.0-preview-bin-hadoop2.7")

from datetime import datetime
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, date_format, udf
from pyspark.sql.types import DateType
```

In [9]:

In [10]:

In []:

In []: