

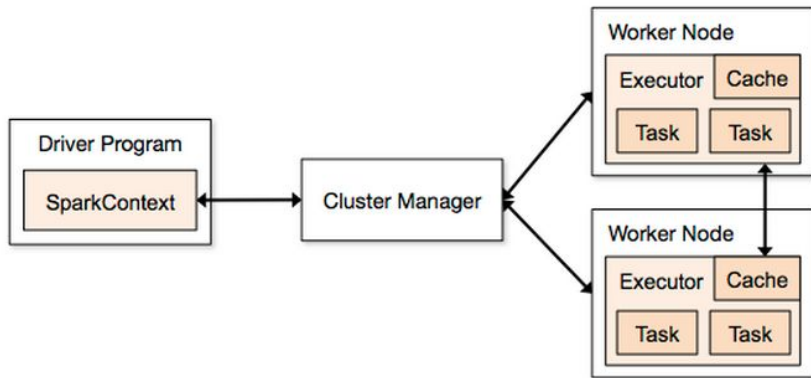


Raúl Pérez



¿Qué es?

Es un sistema de computación en cluster rápido y de propósito general.





Creación

Fue creado en la Universidad de Berkeley, California en 2009 y es considerado el primer software de código abierto que hace la programación distribuida realmente accesible a los científicos de datos.



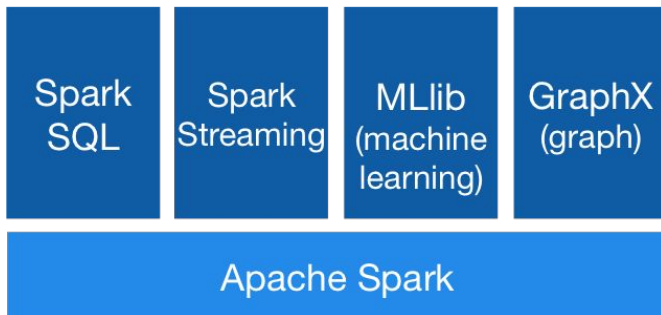
Características principales

- Integrado con Hadoop
- Trabaja en memoria (aunque también en disco)
- Procesamiento en tiempo real
- RDD (Resilient Distributed Dataset)
- API Java, Scala, Python, R



Componentes principales

- **Spark Core:** Base donde se apoya el resto de componentes
- **Spark SQL:** Procesamiento de datos estructurados y semi-estructurados
- **Spark Streaming:** Procesamiento de datos en tiempo real
- **Spark MLlib:** Librería de machine learning
- **Spark Graph:** Procesamiento de grafos. Añade DAG





RDD (Resilient Distributed Dataset)

Es un conjunto de datos distribuidos resiliente (RDD), que es una colección de elementos divididos en los nodos del clúster que pueden operarse en paralelo.




Características principal

- Es la principal abstracción de datos en Spark
- Los RDDs están particionados en los nodos del cluster
- Usan la evaluación perezosa
 - Los RDDs usan evaluación perezosa en sus transformaciones
 - Mantiene todas las transformaciones en un DAG
 - Cuando se lanza una acción se resuelve el grafo



Instalación

Ir a la página de Apache Spark, click en **Descargar**



Lightning-fast unified analytics engine

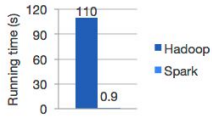
[Download](#) [Libraries](#) [Documentation](#) [Examples](#) [Community](#) [Developers](#) [Apache Software Foundation](#)

Apache Spark™ is a unified analytics engine for large-scale data processing.

Speed

Run workloads 100x faster.

Apache Spark achieves high performance for both batch and streaming data, using a state-of-the-art DAG scheduler, a query optimizer, and a physical execution engine.



Engine	Running time (s)
Hadoop	110
Spark	0.9

Logistic regression in Hadoop and Spark

Ease of Use

Write applications quickly in Java, Scala, Python, R, and SQL.

Spark offers over 80 high-level operators that make it easy to build parallel apps. And you can use it *interactively* from the Scala, Python, R, and SQL shells.


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

Spark's Python DataFrame API
Read JSON files with automatic schema inference

Latest News

- Spark+AI Summit (October 2-4th, 2018, London) agenda posted [\(Jul 24, 2018\)](#)
- Spark 2.2.2 released [\(Jul 02, 2018\)](#)
- Spark 2.1.3 released [\(Jun 29, 2018\)](#)
- Spark 2.3.1 released [\(Jun 08, 2018\)](#)

[Archive](#)



APACHECON
North America
September 24-27, 2018
Montréal, Canada

[Download Spark](#)

Built-in Libraries:

- [SQL and DataFrames](#)
- [Spark Streaming](#)
- [MLlib \(machine learning\)](#)
- [GraphX \(graph\)](#)

[Third-Party Projects](#)



Elegir la versión y click en **Descargar Spark**



Download Apache Spark™

1. Choose a Spark release: **2.3.1 (Jun 08 2018)**
2. Choose a package type: **Pre-built for Apache Hadoop 2.7 and later**
3. Download Spark: [spark-2.3.1-bin-hadoop2.7.tgz](#)
4. Verify this release using the [2.3.1 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.

Link with Spark

Spark artifacts are [hosted in Maven Central](#). You can add a Maven dependency with the following coordinates:

```
groupId: org.apache.spark  
artifactId: spark-core_2.11  
version: 2.3.1
```

Installing with PyPi

PySpark is now available in pypi. To install just run `pip install pyspark`.

Release Notes for Stable Releases

- [Spark 2.3.1](#) (Jun 08 2018)
- [Spark 2.3.0](#) (Feb 28 2018)

Latest News

Spark+AI Summit (October 2-4th, 2018, London) agenda posted (Jul 24, 2018)

Spark 2.2.2 released (Jul 02, 2018)

Spark 2.1.3 released (Jun 29, 2018)

Spark 2.3.1 released (Jun 08, 2018)

[Archive](#)



Download Spark

Built-in Libraries:

[SQL and DataFrames](#)
[Spark Streaming](#)
[MLlib \(machine learning\)](#)
[GraphX \(graph\)](#)

[Third-Party Projects](#)



Click en el primer link



Google Custom	Q
The Apache Way	
Contribute	
ASF Sponsors	

We suggest the following mirror site for your download:

<http://www-eu.apache.org/dist/spark/spark-2.3.1/spark-2.3.1-bin-hadoop2.7.tgz>

Other mirror sites are suggested below.

It is essential that you [verify the integrity](#) of the downloaded file using the PGP signature (`.asc` file) or a hash (`.md5` or `.sha*` file).

Please only use the backup mirrors to download KEYS, PGP and MD5 sigs/hashes or if no other mirrors are working.

HTTP

<http://www-eu.apache.org/dist/spark/spark-2.3.1/spark-2.3.1-bin-hadoop2.7.tgz>

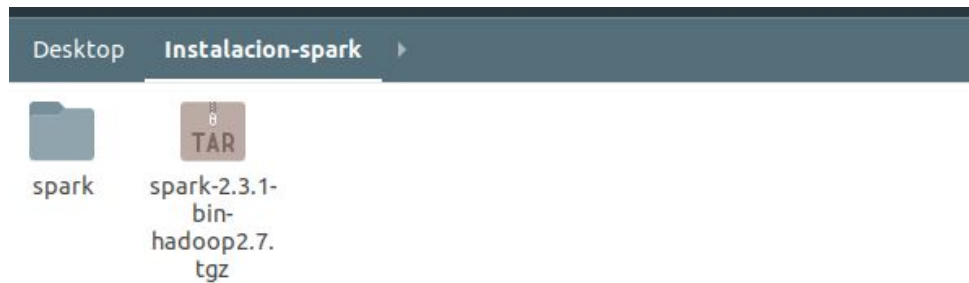
<http://www-us.apache.org/dist/spark/spark-2.3.1/spark-2.3.1-bin-hadoop2.7.tgz>

BACKUP SITES

Please only use the backup mirrors to download KEYS, PGP and MD5 sigs/hashes or if no other mirrors are working.



Descomprimir y cambiar el nombre a la carpeta (si quieren)





Abrir el archivo .bashrc y añadir Spark al path

```
# Spark environment setup
export SPARK_HOME=/home/raul/spark
export PATH=$PATH:$SPARK_HOME/bin
```