

Big Data Aplicado

Josep Garcia

garcia_jos19@ieseduardoprimo.es



IES EDUARDO
PRIMO MARQUÉS



SPARK



- Apache Spark es un entorno de procesamiento distribuido y paralelo que trabaja en memoria.
- Permite el análisis de grandes conjuntos de datos.
- Integra diferentes entornos como Bases de Datos NoSQL, RealTime, machine learning, o análisis de grafos, etc
- Es mucho más rápido que MapReduce.
- Compatible con Hadoop.

<https://spark.apache.org/>

SPARK

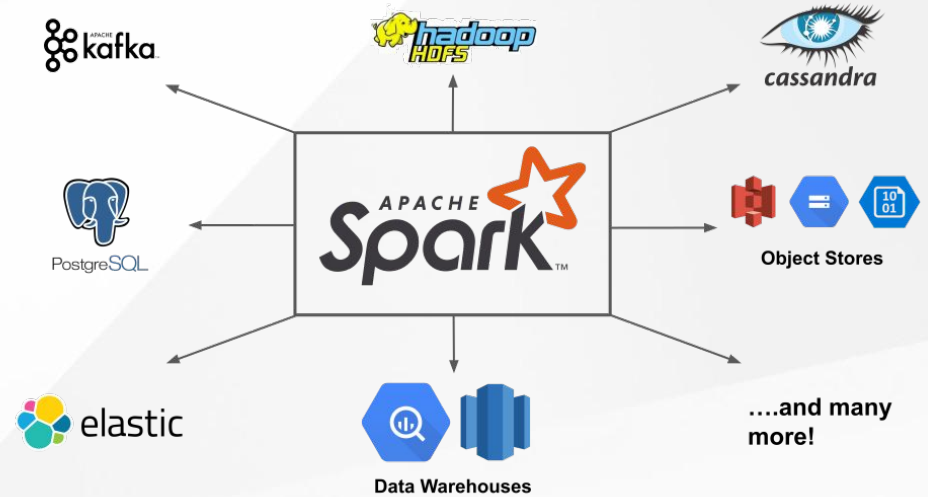


- Al contrario que Hadoop Map Reduce que trabaja sobre todo con procesos de tipo Batch, Spark está orientado al trabajo in-memory y el procesamiento en real.
- Mientras MapReduce trabaja secuencialmente, Spark lo hace en paralelo.
- Compatible con Hadoop:
 - Se puede ejecutar sobre HDFS.
 - MapReduce. Se puede usar en el mismo clúster que MapReduce.
 - YAR: una aplicación Spark se puede lanzar sobre YARN.
 - Se puede mezclar aplicaciones Spark y MapReduce para trabajar con batch y Real Time.

SPARK

- Soporta múltiples fuentes de datos:

- Hive
- Json
- Cassandra
- CSV
- RDBMS...

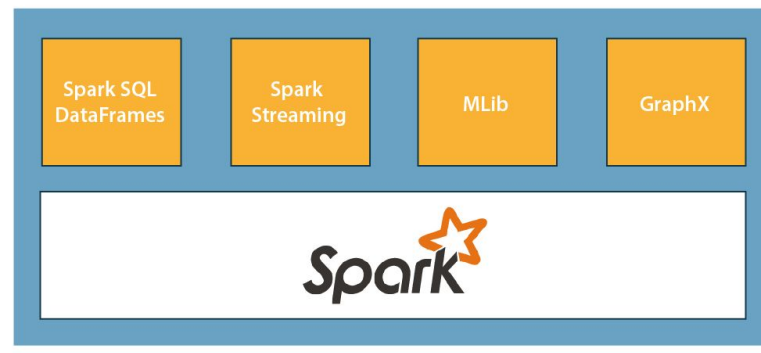


- **Construido en Scala**, pero se pueden escribir aplicaciones en Java, Python y R.
- Dispone de un **Shell interactivo**.

SPARK



- Consiste en un Core y un conjunto de librerías



- Spark core
 - El motor para el procesamiento en escala y distribuido.
 - Construido en Scala pero hay APIs para Python, Java y R.
 - Se encarga de: Gestión de memoria, recuperación ante fallos, planificación, distribución de trabajos en el clúster, Monitorización de trabajo, acceso a los sistemas de almacenamiento.

SPARK



Spark Core RDD

- Usa una estructura de datos especial denominada RDD: **R**esilient **D**istributed **D**atasets.
- Resilient Distributed Datasets permite realizar procesos - fault tolerant 'in-memory'.
- Los RDD son colecciones de registros inmutables y particionadas que además pueden ser manejadas en paralelo.
- Los RDDs pueden contener cualquier clase de objetos Python, Scala, Java o personalizados.
- Los RDD se crean habitualmente transformándolos de otros RDD o cargando los datos de una Fuente externa, como por ejemplo HDFS o HBase.

SPARK



¿Qué son los RDD?

A screenshot of a YouTube video player. The video title is "QUÉ SON LOS RDD" by "OpenWebinars". The video is at 1:40. The main content area shows a slide titled "Evaluación perezosa" (Lazy Evaluation) with three bullet points: "Los RDDs usan evaluación perezosa en sus transformaciones.", "Mantiene todas las transformaciones en un DAG.", and "Cuando se lanza una acción, se resuelve el grafo.". On the right side of the video, there is a small inset video of a man, Abraham Requena, with his Twitter handle @AbrahamReqMes. The video player controls at the bottom show a progress bar at 2:05 / 7:27 and the text "Características principales".

<https://openwebinars.net/blog/que-son-los-rdd/>



IES EDUARDO
PRIMO MARQUÉS



SPARK



Spark Streaming

- Se usa para procesar fuentes de datos en tiempo real (streaming data)
- Permite procesar con una alta tolerancia a fallos y un gran rendimiento las fuentes “vivas” de información que le suministremos.
- Su unidad fundamental de trabajo es el Dstream (serie de RDDs, que veremos posteriormente)

SPARK



Spark SQL

- Permite integrar comandos y componentes relacionales junto con la programación funcional de Spark.
- Podemos usar SQL o Hive Query Language
- Permite el acceso a múltiples fuentes de datos
- Dispone de 4 librería básicas
 - Data Source
 - DataFrame
 - Interpreter and Optimizer
 - Sql Service
- Permite el acceso por JDBC o ODBC

SPARK



Spark GraphX

- Es el API para procesamiento paralelo en **grafos**.
- Spark GraphX implementa Resilient Distributed Graph (RDG- una abstracción de los RDD's).
- RDG's asocia registros con los vértices y bordes de un grafo. Sin embargo, se pueden seguir viendo como colecciones tradicionales de RDD.
- Se dispone de una gran cantidad de librerías con algoritmos preparados, que permiten agilizar el proceso de construcción de aplicaciones y mejora el rendimiento y velocidad.

SPARK



Spark Mlib

- Se utiliza para **machine learning** en Spark.
- Se dispone de una variedad de algoritmos y otros procesos como “data cleaning”:
 - Clasificación, clustering, regression, extracción etc...
- Permite su ejecución sobre HDFS, HBase, etc...

SPARK



Descarga e instalación <https://spark.apache.org/>

The screenshot shows the Apache Spark download page. At the top is a dark blue navigation bar with the Apache Spark logo on the left and links for Download, Libraries, Documentation, Examples, Community, and Developers in the center. On the right of the bar is the Apache Software Foundation logo. Below the navigation bar, the main content area has a light blue header "Download Apache Spark™". Under this header, there are four numbered steps: 1. Choose a Spark release (3.5.0 (Sep 13 2023)), 2. Choose a package type (Pre-built for Apache Hadoop 3.3 and later), 3. Download Spark (spark-3.5.0-bin-hadoop3.tgz), and 4. Verify this release using the 3.5.0 signatures, checksums and project release KEYS by following these procedures. Below these steps is a note about Scala versions. To the right of the main content is a "Latest News" section with a list of recent releases and an "Archive" link. Below the news section is a "COMMUNITY CODE" logo and a red "DOWNLOAD SPARK" button. Further down, there is a "Link with Spark" section with Maven coordinates, an "Installing with PyPi" section, and a "Convenience Docker Container Images" section.

Download Apache Spark™

1. Choose a Spark release:
2. Choose a package type:
3. Download Spark: [spark-3.5.0-bin-hadoop3.tgz](#)
4. Verify this release using the 3.5.0 [signatures](#), [checksums](#) and [project release KEYS](#) by following these [procedures](#).

Note that Spark 3 is pre-built with Scala 2.12 in general and Spark 3.2+ provides additional pre-built distribution with Scala 2.13.

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_2.12  
version: 3.5.0
```

Installing with PyPi

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

Convenience Docker Container Images

[Spark Docker Container images](#) are available from [DockerHub](#), these images contain non-ASF software and may be subject to

Latest News

- Spark 3.3.4 released (Dec 16, 2023)
- Spark 3.4.2 released (Nov 30, 2023)
- Spark 3.5.0 released (Sep 13, 2023)
- Spark 3.3.3 released (Aug 21, 2023)

[Archive](#)

COMMUNITY CODE

DOWNLOAD SPARK

Built-in Libraries:

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

SPARK vs HADOOP



- Spark es un producto distinto a hadoop, se integran bien, son compatible y muchas veces se utilizan conjuntamente, pero son productos distintos.
- Para lo que está diseñado, en determinadas ocasiones es mucho más rápido que hadoop.
- Mapreduce es bueno en procesos batch,spark en procesos de tiempo real.

