

Big Data Aplicado

Josep Garcia

garcia_jos19@ieseduardoprimo.es

SPARK



- Apache Spark es una 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/>

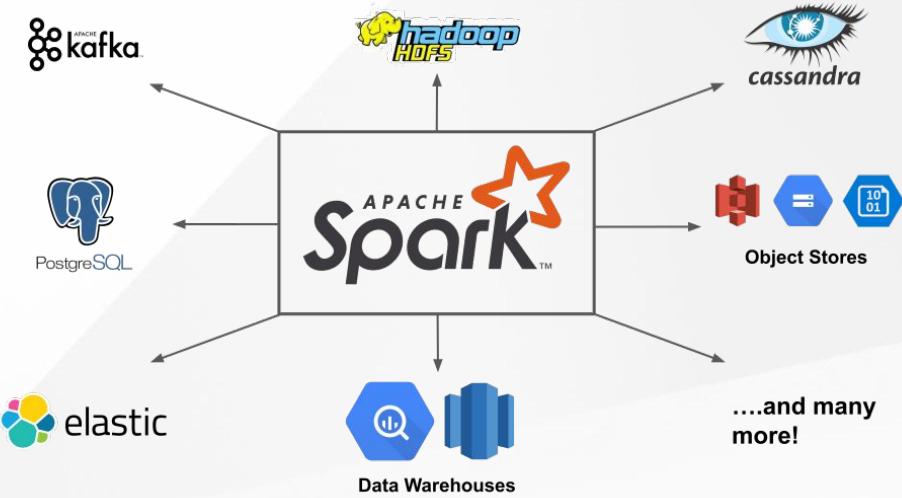
SPARK



- 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.

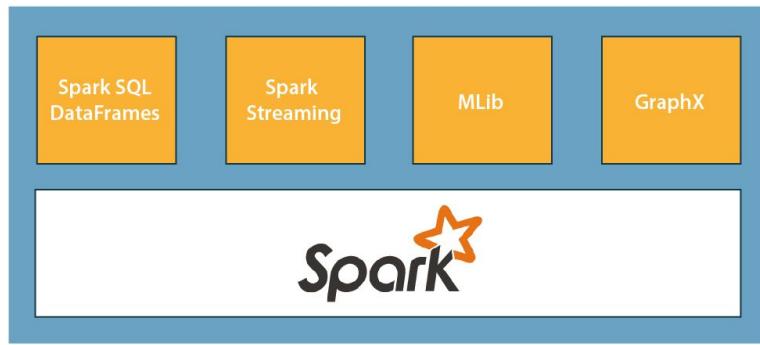
SPARK

- 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

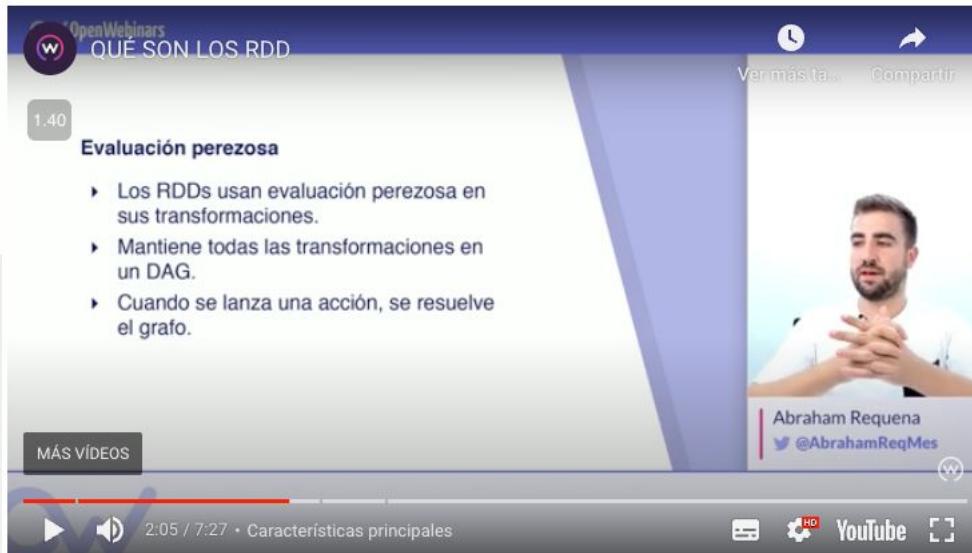


- Spark core
 - El motor para el procesamiento en escala y distribuido.
 - Construido en Scala pero hay APIs para Python, Java y R.
 - Se encarga de: Gestión de memoria, recuperación ante fallos, planificación, distribución de trabajos en el clúster, Monitorización de trabajo, acceso a los sistemas de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

Spark Core RDD

- Usa una estructura de datos especial denominada RDD: **Resilient Distributed Datasets**.
- 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.

¿Qué son los RDD?



OpenWebinars QUÉ SON LOS RDD

1.40

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.

MÁS VÍDEOS

2:05 / 7:27 • Características principales

Abraham Requena
@AbrahamReqMes

W

YouTube

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



IES EDUARDO
PRIMO MARQUÉS



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 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 GraphX

- Es el API para procesamientos 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 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/>

 Apache Spark

Download Libraries Documentation Examples Community Developers Apache Software Foundation

Download Apache Spark™

- Choose a Spark release: [3.5.0 \(Sep 13 2023\)](#)
- Choose a package type: [Pre-built for Apache Hadoop 3.3 and later](#)
- Download Spark: [spark-3.5.0-bin-hadoop3.tgz](#)
- 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.

