# Plataformas para BD

## Sistemas de Big Data

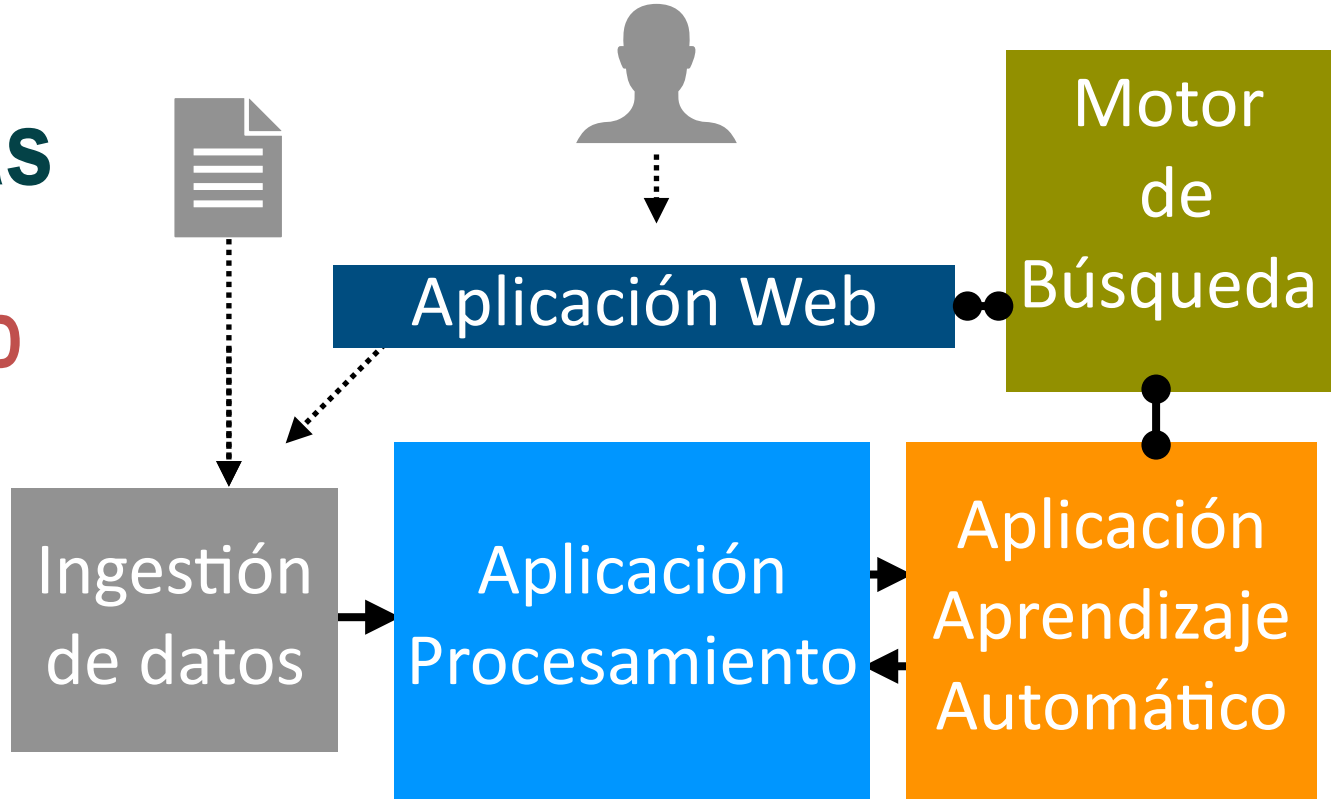
Ricardo García Ródenas
Ricardo.Garcia@uclm.es

Sistemas de
Big_Data

Curso Especialización
Inteligencia_Artificial y
Big_Data

# Plataformas BD

## Arquitectura BD



# Plataformas BD


## Características

- *Apache Software Foundation (Código abierto)*
- Procesamiento **batch**
- **MapReduce.** Capacidad de procesar gran cantidad datos

# Plataformas BD


## Características

- Escalabilidad
- Tolerancia a fallos
- Flexibilidad

# Plataformas BD


## Retos

- No **recurisvidad** ni **interactividad**
  - Coste de **inicio**
  - **Múltiples** fases Map/Reduce
  - Múltiples **archivos**

# Plataformas BD


## Retos

- Falta **programadores**
  - Conocimientos **Java**
  - **SQL+ Hadoop**
- **Administración arte y ciencia**

# Plataformas BD


## Retos

- **Seguridad** de datos
- **Gestión** y **gobierno** de los datos

# Plataformas BD


## Módulos

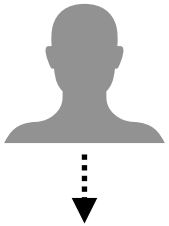
- **HDFS (***Hadoop distributed file system)*
- **YARN** *(Yet Another Resource Negotiator)*

# Plataformas BD


## Módulos

- **HDFS (*Hadoop distributed file system)***
  - Sistema archivos **distribuido**
  - **Replicación  bloques**
- **YARN** *(Yet Another Resource Negotiator)*
  - *ResourceManager*
  - *ApplicationMaster (tareas MapReduce)*

# Motor

Motor de Búsqueda

Aplicación Web


Cassandra

Aplicación Procesamiento


HDFS

YARN MapReduce

Aplicación Aprendizaje Automático

MAHOUT

# Plataformas BD


## Características

- Procesamiento **batch**, **streaming** y **en linea**
  - **Código abierto**
  - Procesamiento **iterativo** de algoritmos **aprendizaje automático**

# Plataformas BD


## Características

- **Alto rendimiento**
  - Computación en **memoria. Eliminación** del uso del disco de los resultados intermedio
  - **DAG (optimización** etapas Map/Reduce**)**

# Plataformas BD


## Características

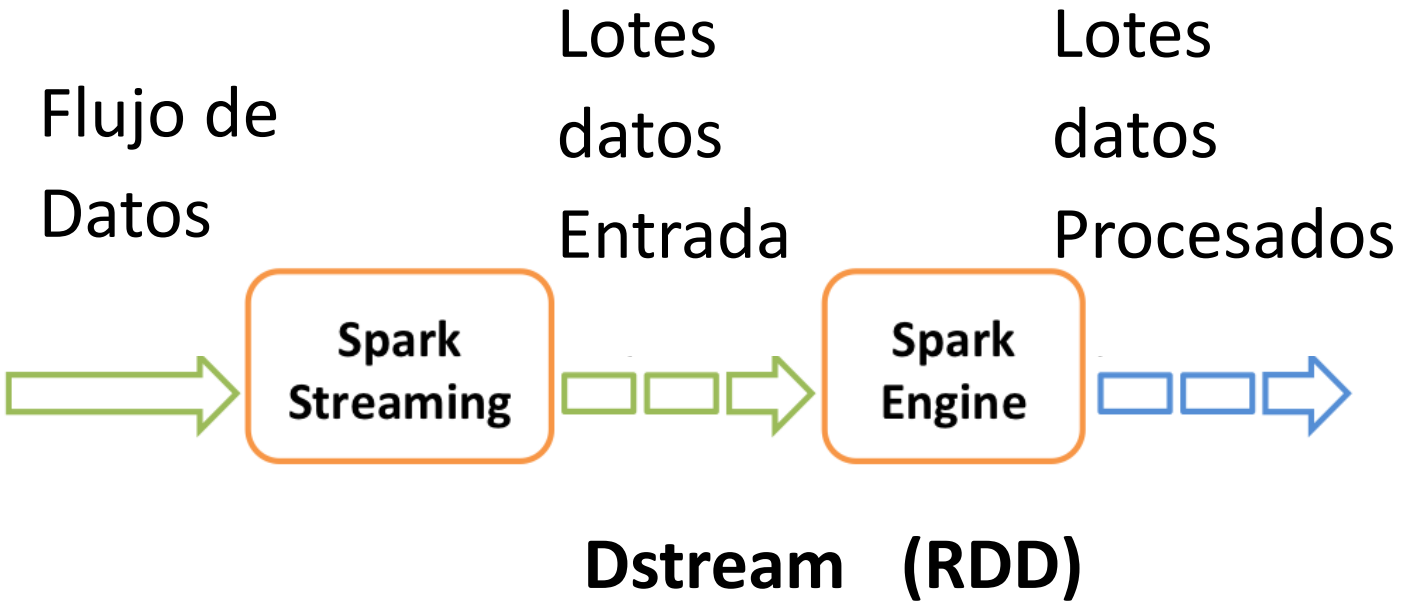
- Facilidad de **programación**
  - **Python**
  - **R**
  - **Scala**
  - **Java**
  - **SQL**

# Plataformas BD


## RDD

- **RDD** *Resilent Distributed Datasets*
  - **Estructura de datos distribuida**
  - **Partición** (unidad atómica situada en un nodo)
  - RDD es **colección** de particiones
  - Coloca automáticamente operaciones en RDDs
  - **Tolerante a fallos**

# Spark Streaming



# Ecosistema


Spark SQL+ DataFrame

Streaming

MLlib

GraphX

**Spark Engine API**

R

Python

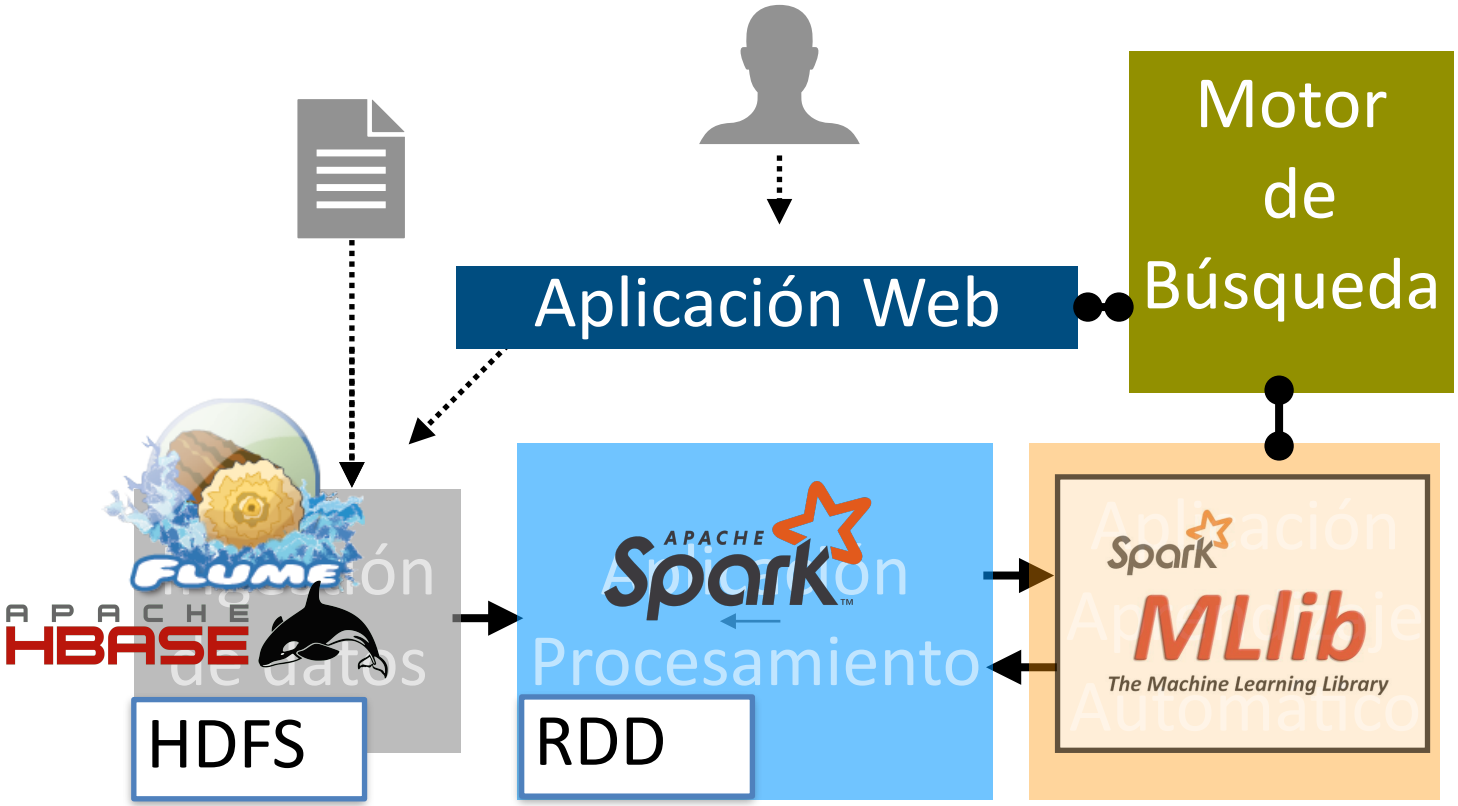
Scala

Java

SQL

# Motor


APACHE
Spark


Aplicación Web
Motor de Búsqueda
FLUME
APACHE
HBASE
HDFS
APACHE
Spark
Procesamiento
RDD
Spark
MLlib
The Machine Learning Library

# Similitudes/ diferencias


- **Características comunes**
  - **Localidad datos**
  - **Ejecución en etapas**
  - **HDFS** para la persistencia en disco
- **Diferencias**
  - **En memoria / en disco**
  - **DAG** optimizado / manual

# Plataformas para BD



## Sistemas de Big Data

Ricardo García Ródenas
Ricardo.Garcia@uclm.es

Sistemas de
Big_Data

Curso Especialización
Inteligencia_Artificial y
Big_Data