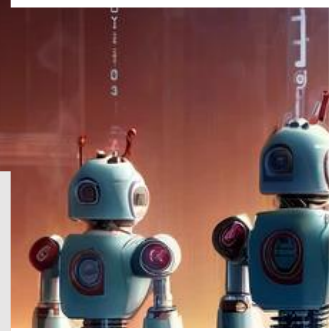
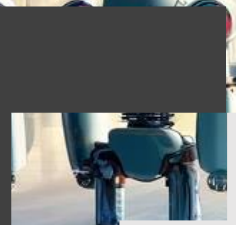
THE BRIDGE

Big Data (IV)
Frameworks: Spark





… y llega Spark

La gracia de Spark es que hace que programar sobre un cluster se parezca más a programar sobre un único ordenador (casi te puedes olvidar de que estás en un cluster, casi…)

… y llega Spark

La gracia de Spark es que hace que programar sobre un cluster se parezca más a programar sobre un único ordenador (casi te puedes olvidar de que estás en un cluster, casi…)

¿Qué es Apache Spark?

Apache Spark es un motor unificado de analíticas para procesar datos a gran escala que integra módulos para SQL, streaming, aprendizaje automático y procesamiento de grafos.

Apache Spark is an open-source, distributed processing system used for big data workloads. It utilizes in-memory caching, and optimized query execution for fast analytic queries against data of any size. It provides development APIs in Java, Scala, Python and R, and supports code reuse across multiple workloads—batch processing, interactive queries, real-time analytics, machine learning, and graph processing.

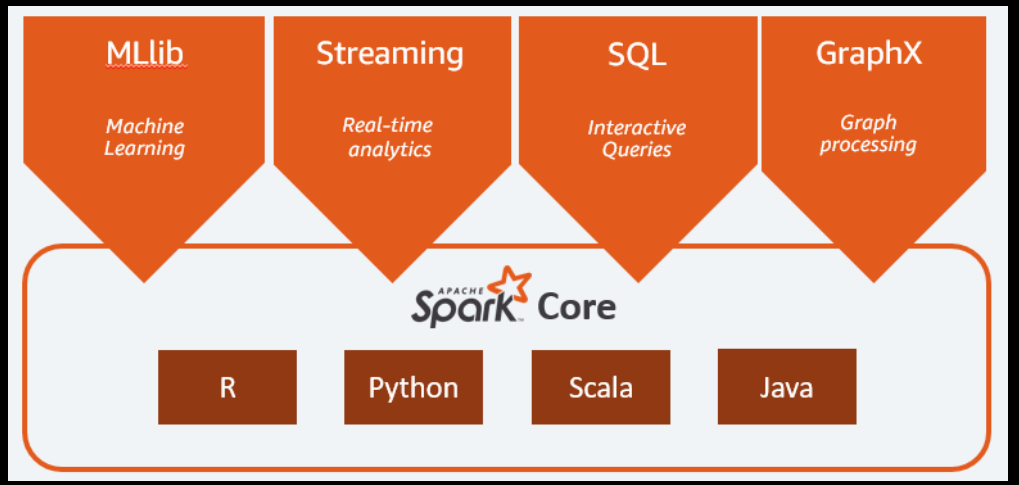
… y llega Spark

La gracia de Spark es que hace que programar sobre un cluster se parezca más a programar sobre un único ordenador (casi te puedes olvidar de que estás en un cluster, casi…)

¿Qué es Apache Spark? Google Cloud

Apache Spark es un motor unificado de analíticas para procesar datos a gran escala que integra módulos para SQL, streaming, aprendizaje automático y procesamiento de grafos.

Apache Spark is an open-source, distributed processing system used for big data workloads. It utilizes in-memory caching, and optimized query execution for fast analytic queries against data of any size. It provides development APIs in Java, Scala, Python and R, and supports code reuse across multiple workloads—batch processing, interactive queries, real-time analytics, machine learning, and graph processing. aws



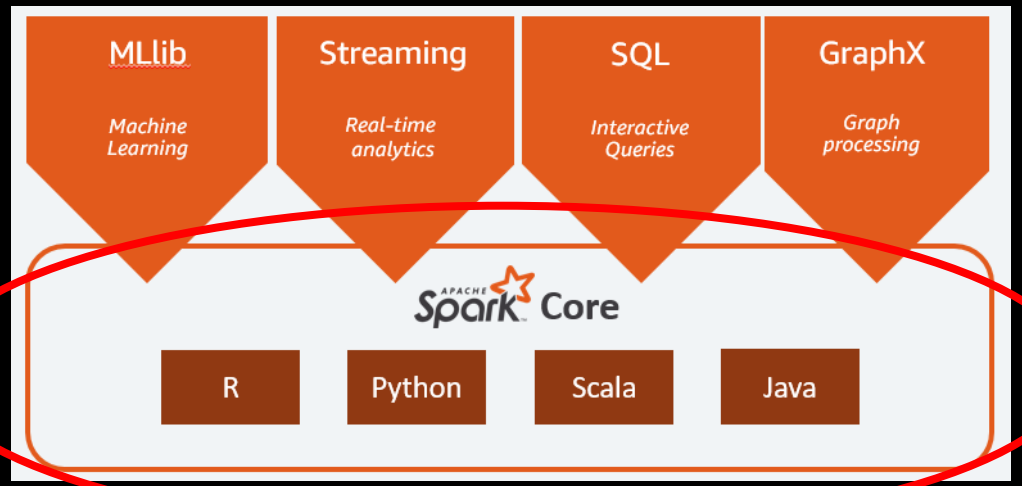
… y llega Spark

La gracia de Spark es que hace que programar sobre un cluster se parezca más a programar sobre un único ordenador (casi te puedes olvidar de que estás en un cluster, casi…)

¿Qué es Apache Spark?

Apache Spark es un motor unificado de analíticas para procesar datos a gran escala que integra módulos para SQL, streaming, aprendizaje automático y procesamiento de grafos.

Apache Spark is an open-source, distributed processing system used for big data workloads. It utilizes in-memory caching, and optimized query execution for fast analytic queries against data of any size. It provides development APIs in Java, Scala, Python and R, and supports code reuse across multiple workloads—batch processing, interactive queries, real-time analytics, machine learning, and graph processing.



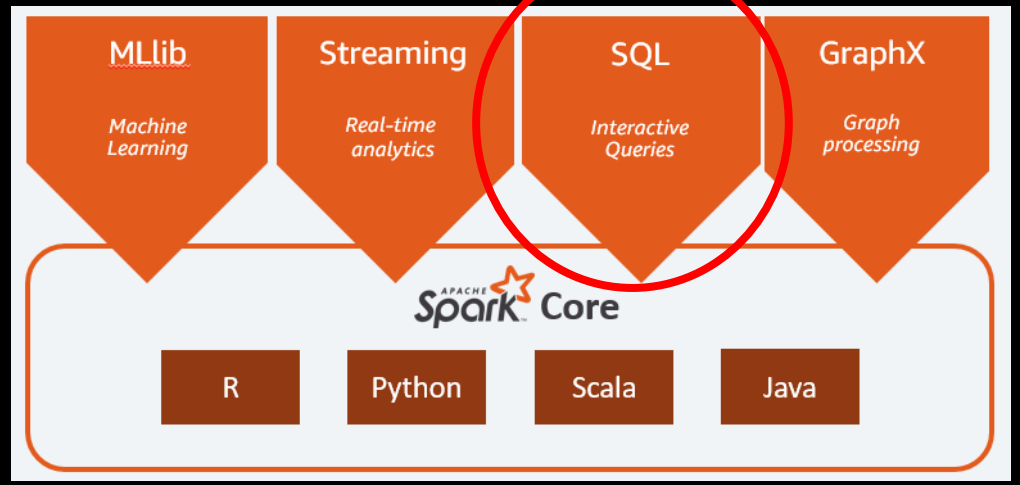
… y llega Spark

La gracia de Spark es que hace que programar sobre un cluster se parezca más a programar sobre un único ordenador (casi te puedes olvidar de que estás en un cluster, casi…)

¿Qué es Apache Spark? Google Cloud

Apache Spark es un motor unificado de analíticas para procesar datos a gran escala que integra módulos para SQL, streaming, aprendizaje automático y procesamiento de grafos.

Apache Spark is an open-source, distributed processing system used for big data workloads. It utilizes in-memory caching, and optimized query execution for fast analytic queries against data of any size. It provides development APIs in Java, Scala, Python and R, and supports code reuse across multiple workloads—batch processing, interactive queries, real-time analytics, machine learning, and graph processing. aws

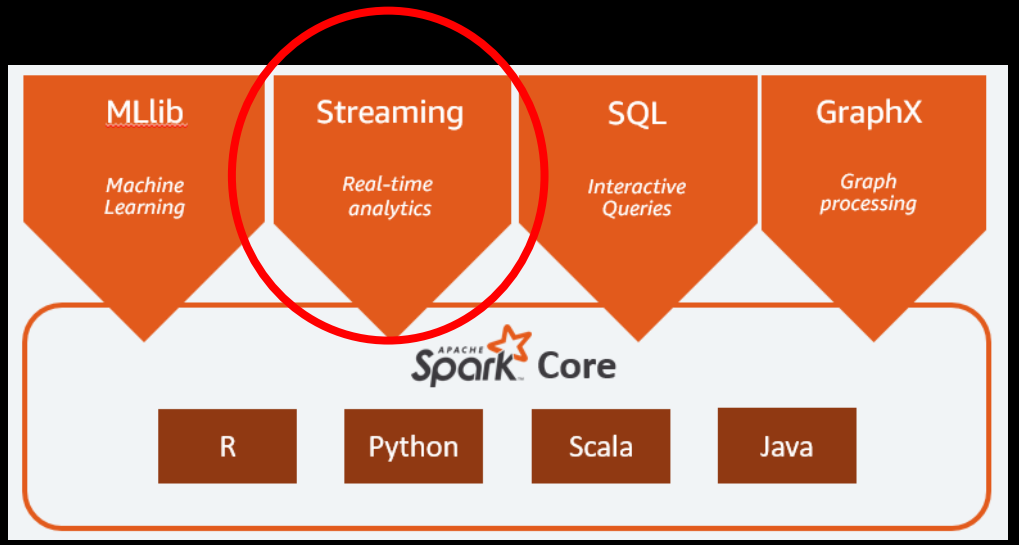


… y llega Spark

La gracia de Spark es que hace que programar sobre un cluster se parezca más a programar sobre un único ordenador (casi te puedes olvidar de que estás en un cluster, casi…)

¿Qué es Apache Spark?

Apache Spark es un motor unificado de analíticas para procesar datos a gran escala que integra módulos para SQL, streaming, aprendizaje automático y procesamiento de grafos.



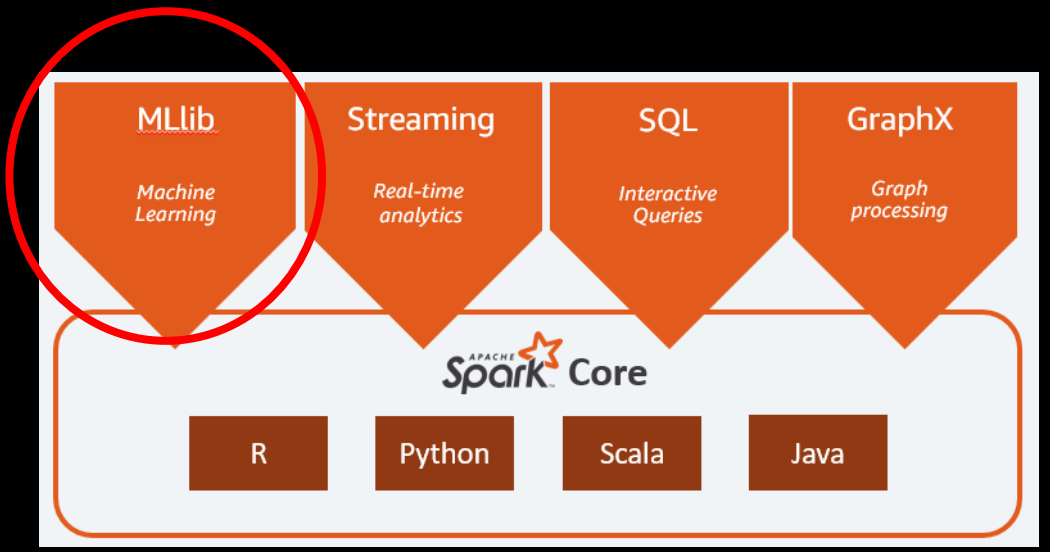
Apache Spark is an open-source, distributed processing system used for big data workloads. It utilizes in-memory caching, and optimized query execution for fast analytic queries against data of any size. It provides development APIs in Java, Scala, Python and R, and supports code reuse across multiple workloads—batch processing, interactive queries, real-time analytics, machine learning, and graph processing.

… y llega Spark

La gracia de Spark es que hace que programar sobre un cluster se parezca más a programar sobre un único ordenador (casi te puedes olvidar de que estás en un cluster, casi…)

¿Qué es Apache Spark?

Apache Spark es un motor unificado de analíticas para procesar datos a gran escala que integra módulos para SQL, streaming, aprendizaje automático y procesamiento de grafos.



Apache Spark is an open-source, distributed processing system used for big data workloads. It utilizes in-memory caching, and optimized query execution for fast analytic queries against data of any size. It provides development APIs in Java, Scala, Python and R, and supports code reuse across multiple workloads—batch processing, interactive queries, real-time analytics, machine learning, and graph processing.

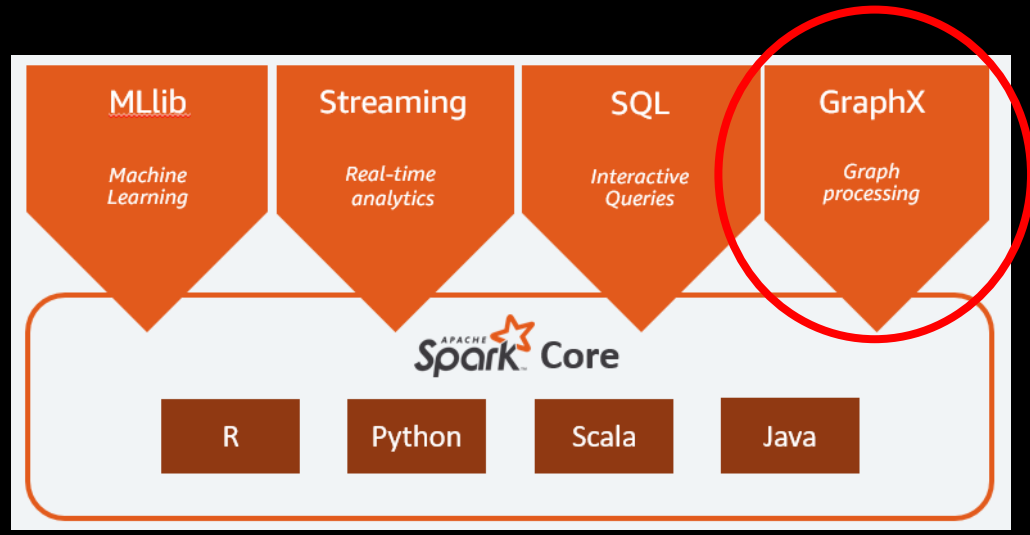
… y llega Spark

La gracia de Spark es que hace que programar sobre un cluster se parezca más a programar sobre un único ordenador (casi te puedes olvidar de que estás en un cluster, casi…)

¿Qué es Apache Spark? Google Cloud

Apache Spark es un motor unificado de analíticas para procesar datos a gran escala que integra módulos para SQL, streaming, aprendizaje automático y procesamiento de grafos.

Apache Spark is an open-source, distributed processing system used for big data workloads. It utilizes in-memory caching, and optimized query execution for fast analytic queries against data of any size. It provides development APIs in Java, Scala, Python and R, and supports code reuse across multiple workloads—batch processing, interactive queries, real-time analytics, machine learning, and graph processing. aws



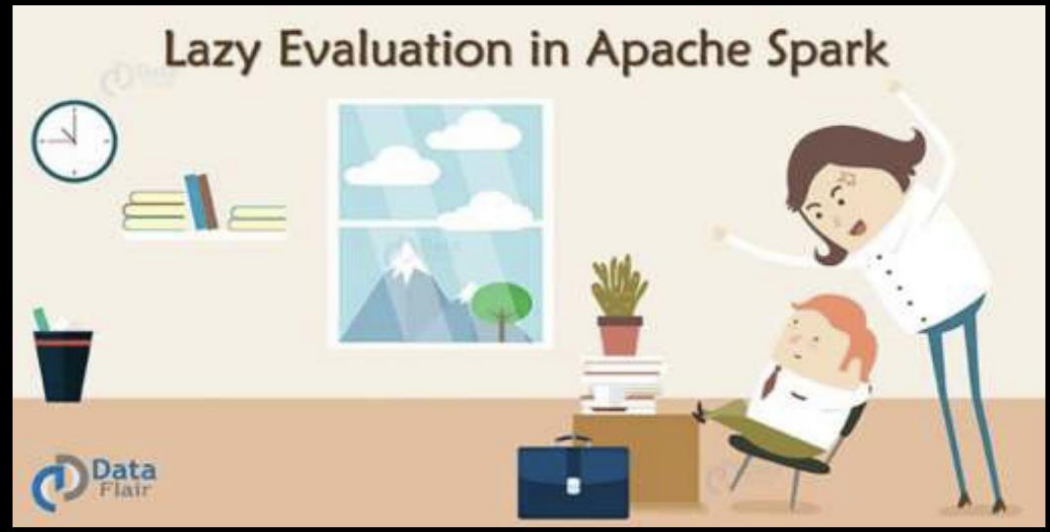
Spark

Nociones básicas

Spark

Nociones básicas

Transformación: operaciones que una vez ejecutadas crearían un nuevo conjunto de datos a partir de un conjunto de datos existente (Grafo DAG)

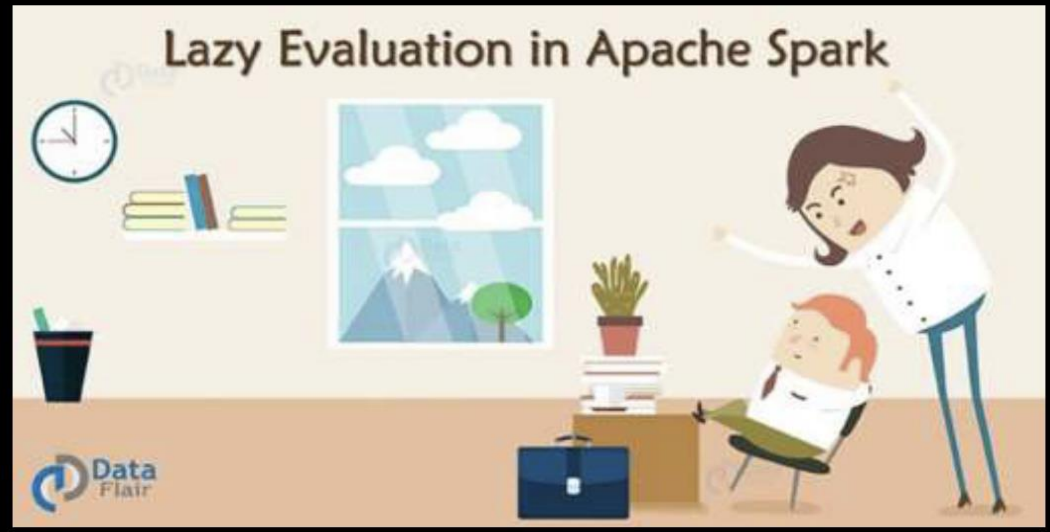


Spark

Nociones básicas

Transformación: operaciones que una vez ejecutadas crearían un nuevo conjunto de datos a partir de un conjunto de datos existente

Acción: operaciones que devuelven un valor al programa controlador (al programa principal y este quiere mostrarlos) o escriben datos a un sistema de almacenamiento externo



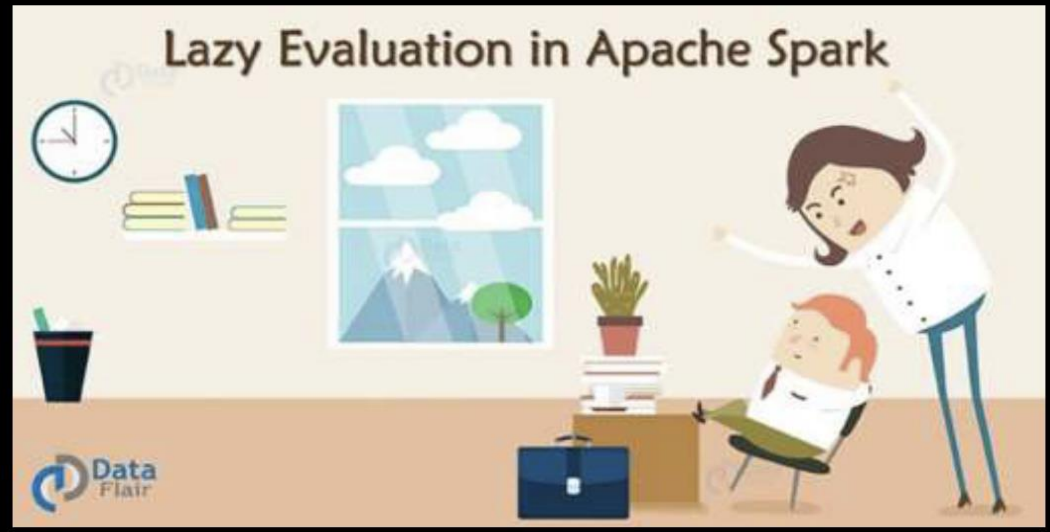
Spark

Nociones básicas

Transformación: operaciones que una vez ejecutadas crearían un nuevo conjunto de datos a partir de un conjunto de datos existente

Acción: operaciones que devuelven un valor al programa controlador (al programa principal y este quiere mostrarlos) o escriben datos a un sistema de almacenamiento externo

Lazy Evaluation: No se ejecutan las transformaciones pendientes hasta llegar a una acción. Permite analizar y optimizar la ejecución.



Para más: http://ashkrit.blogspot.com/2018/09/anatomy-of-apache-spark-job.html

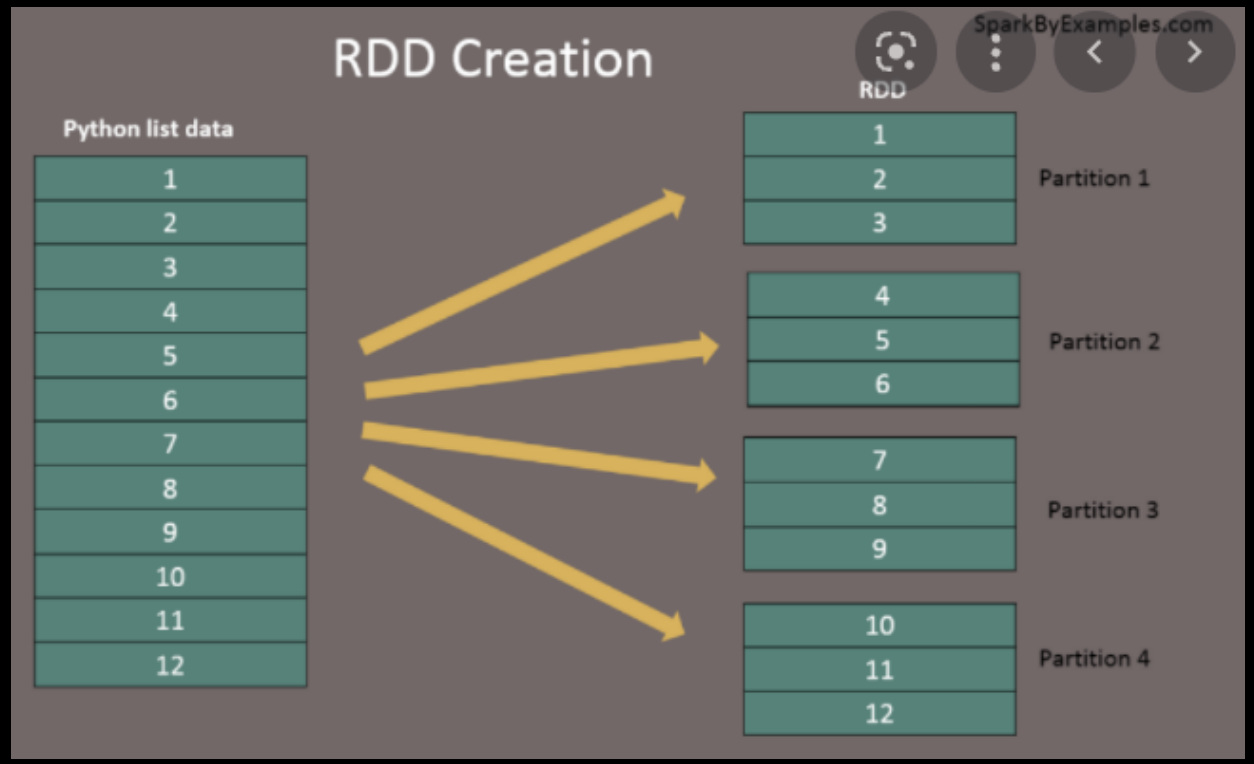
Spark

Nociones básicas

Spark

Nociones básicas

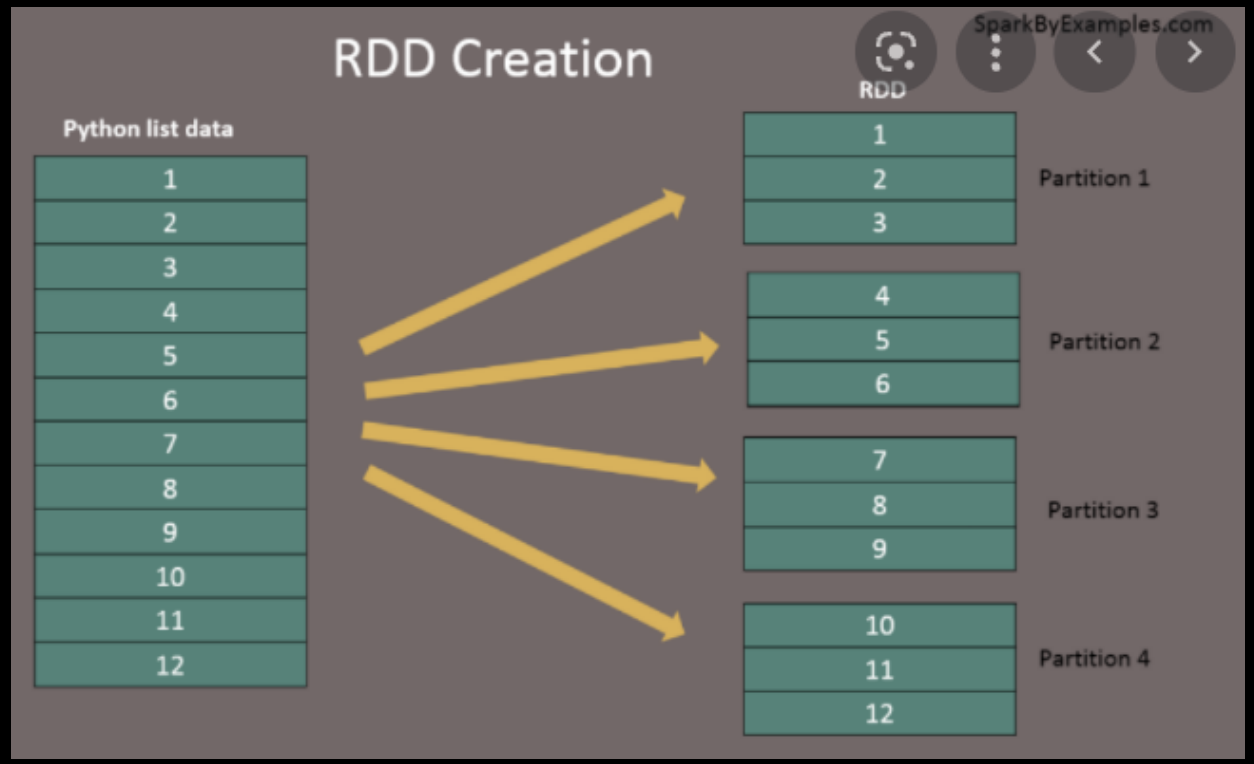
RDD = Resilient Distributed Dataset, estructura básica de datos (particionada y que se puede mantener en memoria)



Spark

Nociones básicas

RDD = Resilient Distributed Dataset, estructura básica de datos (particionada y que se puede mantener en memoria)

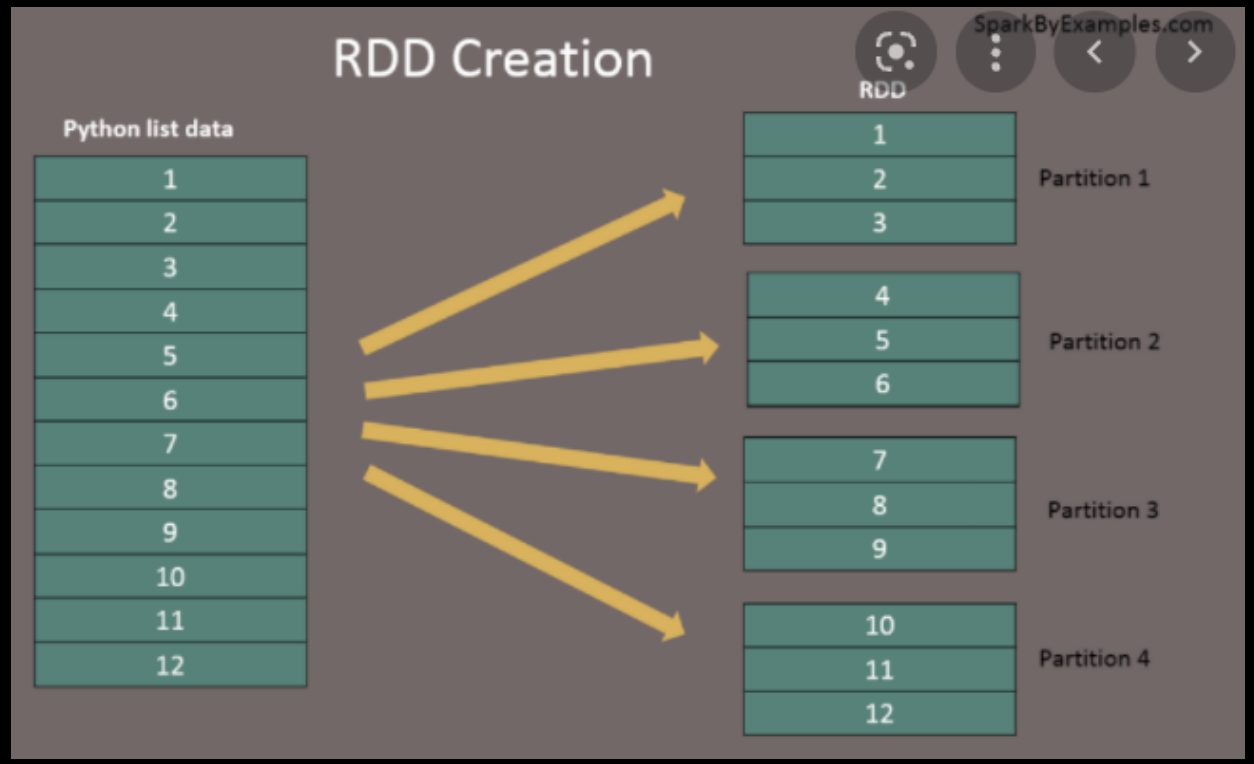


- Es una estructura que internamente está dividida en bloques denominados particiones de forma que se puedan procesar en paralelo.

Spark

Nociones básicas

RDD = Resilient Distributed Dataset, estructura básica de datos (particionada y que se puede mantener en memoria)

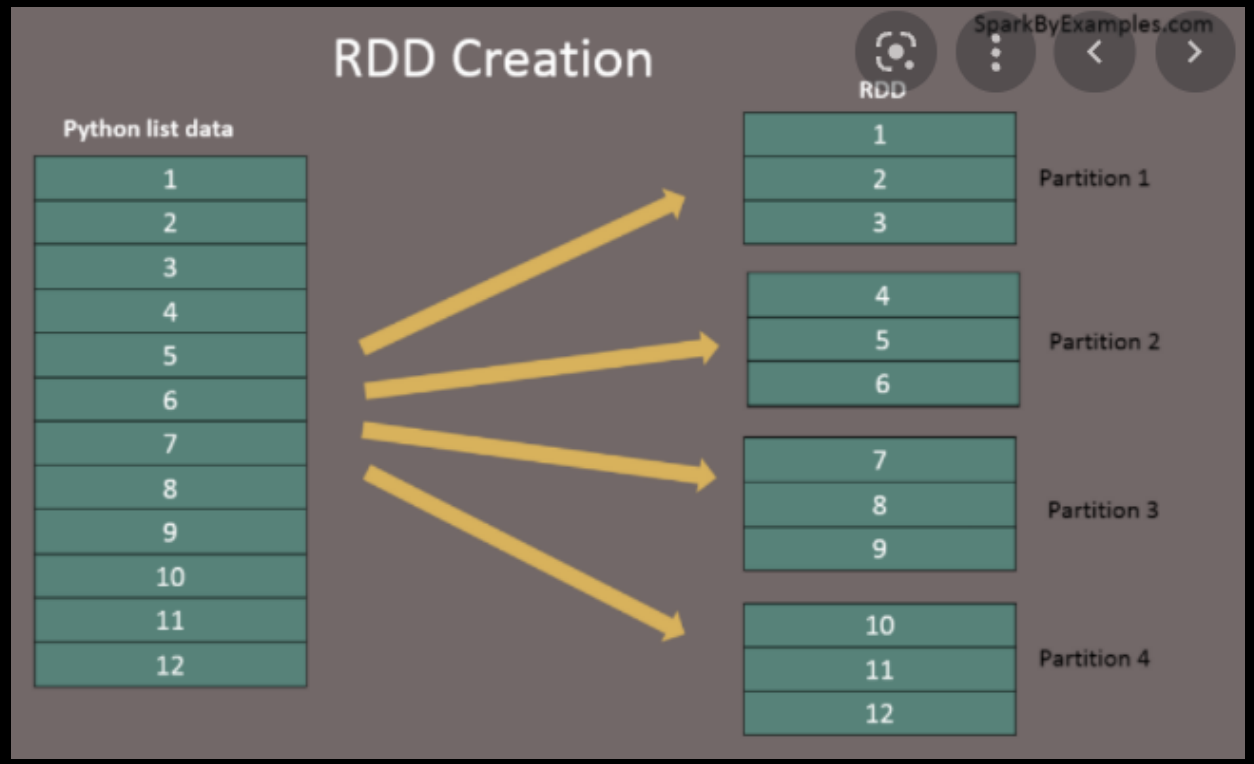


- Es una estructura que internamente está dividida en bloques denominados particiones de forma que se puedan procesar en paralelo.
- Al estar particionado se pueden replicar las particiones en diferentes nodos, permitiendo la redundancia y de ahí la resilencia

Spark

Nociones básicas

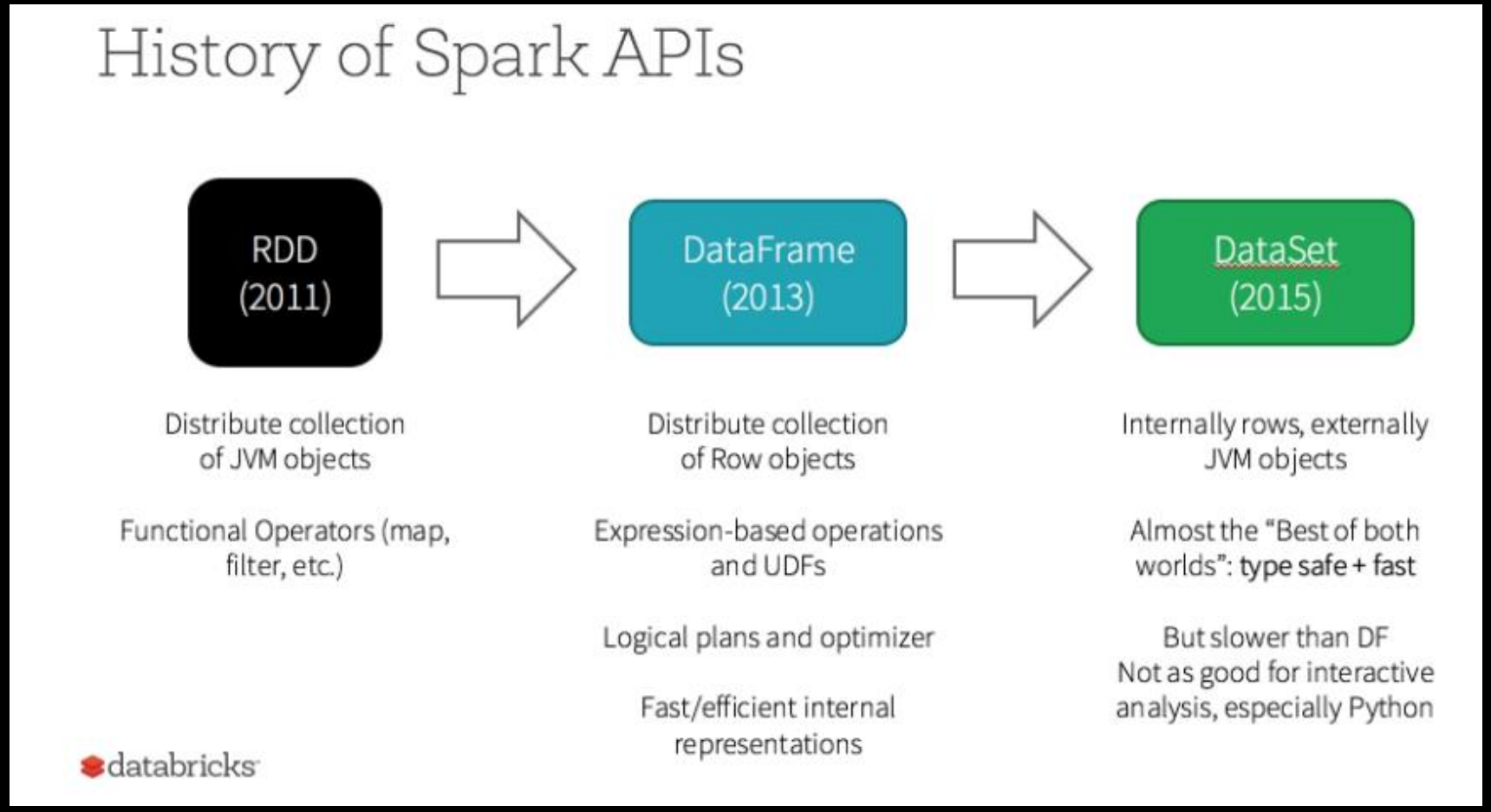
RDD = Resilient Distributed Dataset, estructura básica de datos (particionada y que se puede mantener en memoria)



- Es una estructura que internamente está dividida en bloques denominados particiones de forma que se puedan procesar en paralelo.
- Al estar particionado se pueden replicar las particiones en diferentes nodos, permitiendo la redundancia y de ahí la resilencia
- El particionamiento permite trabajar con las diferentes particiones en memoria y así trabajar con estructuras grandes en ordenadores “pequeños”

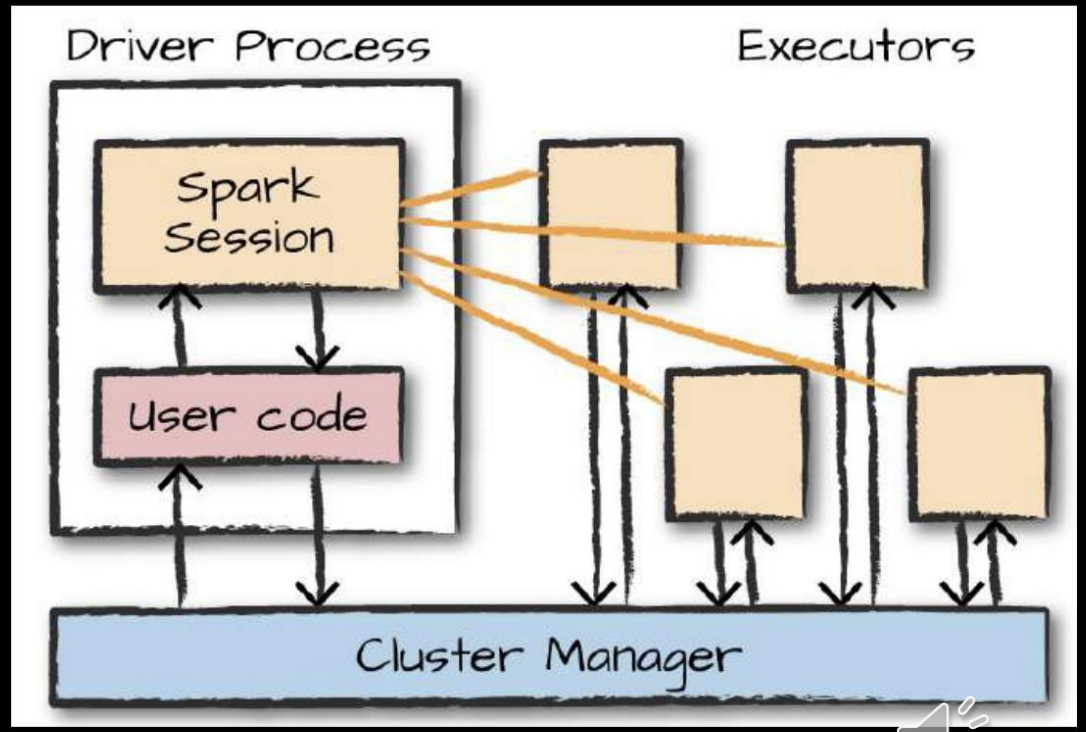
Spark

Estructuras



Spark

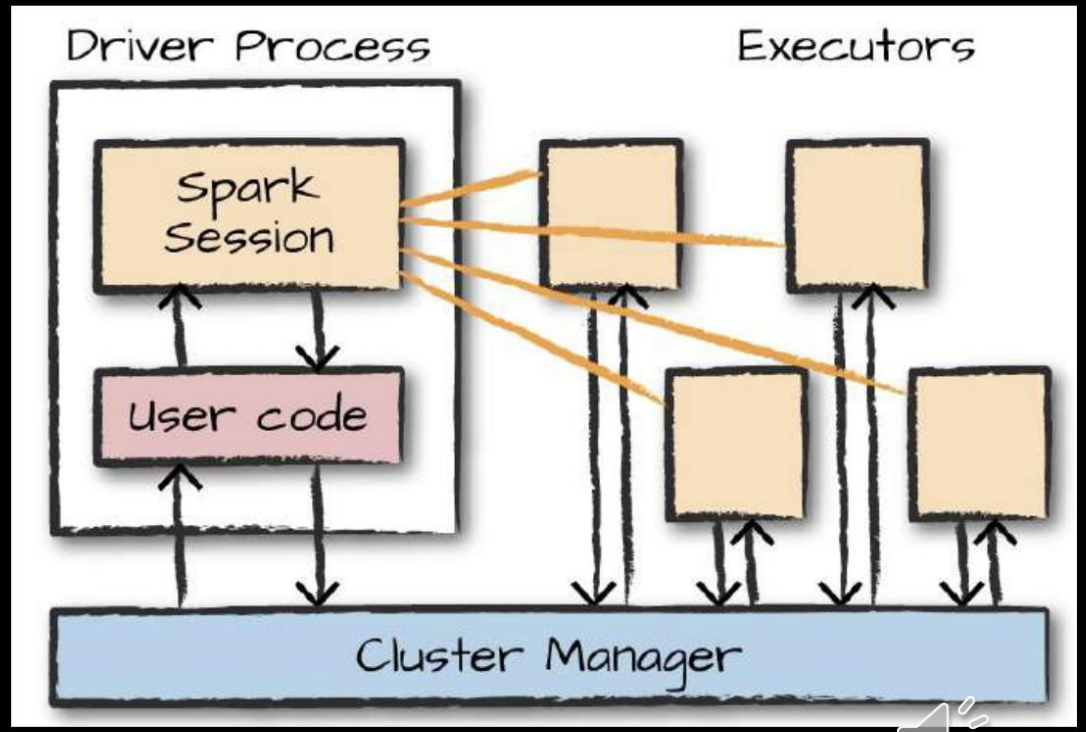
Orquestación



Spark

Orquestación

Spark Context: Objeto especial que es necesario crear para interactuar con un cluster

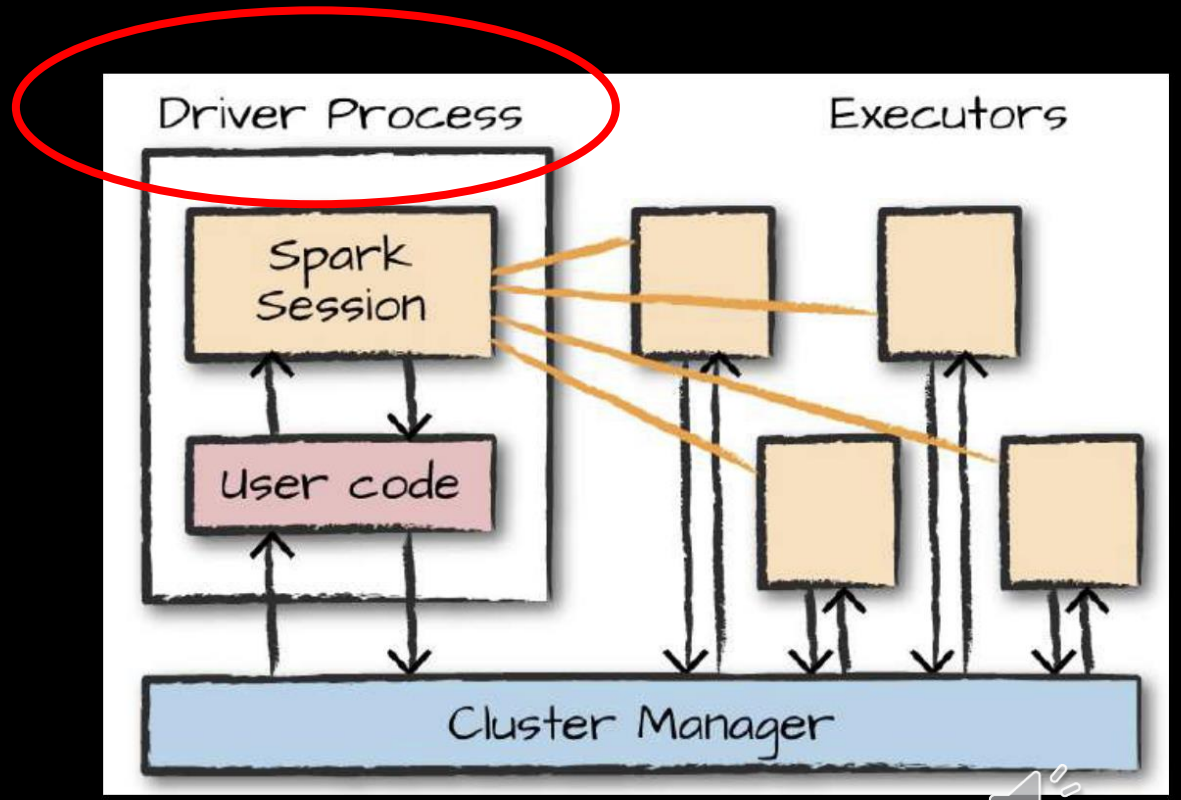


Spark

Orquestación

Spark Context: Objeto especial que es necesario crear para interactuar con un cluster

Driver: Programa principal creado por el Spark Context que divide y coordina la ejecución



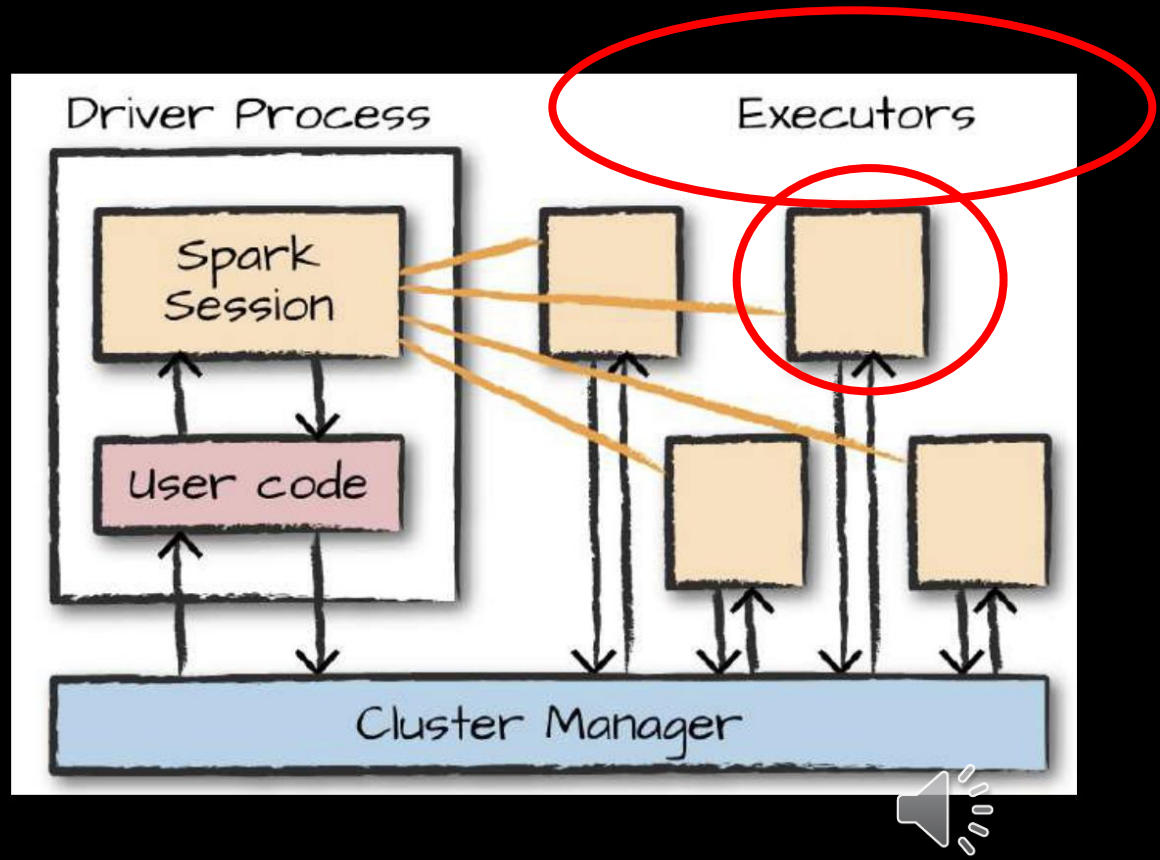
Spark

Orquestación

Spark Context: Objeto especial que es necesario crear para interactuar con un cluster

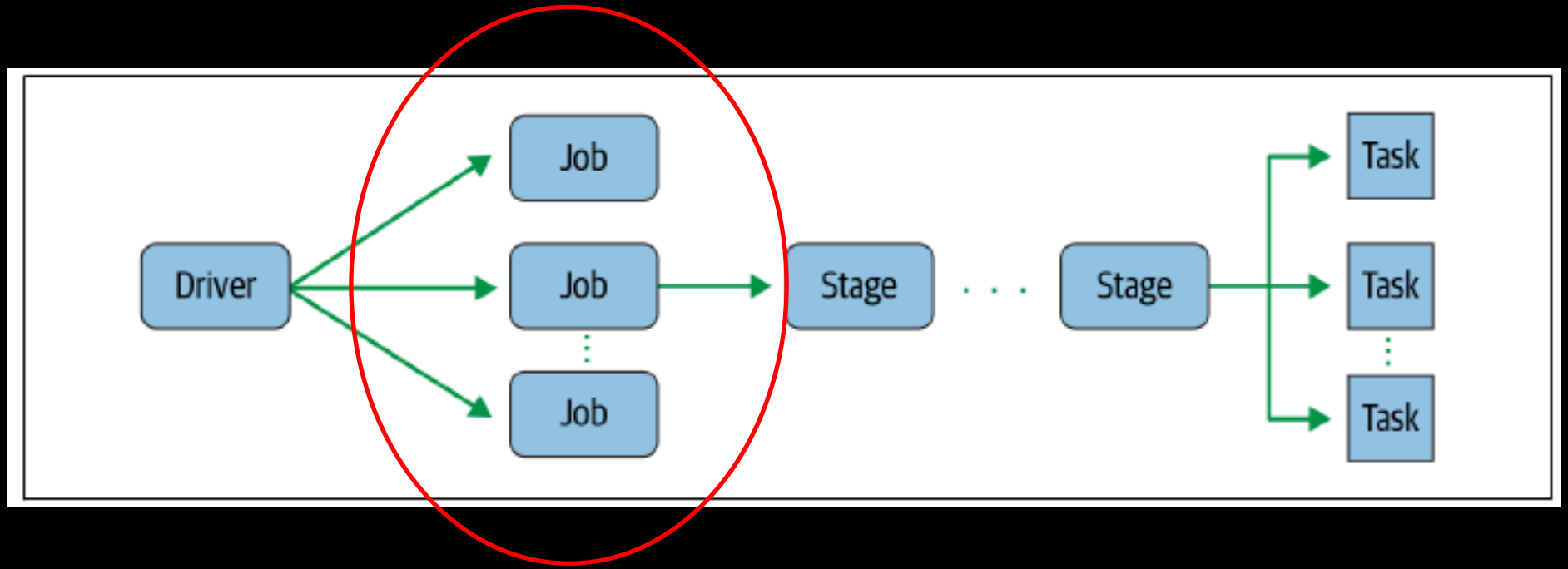
Driver: Programa principal creado por el Spark Context que divide y coordina la ejecución

Ejecutor/Executor: Cada una de las pieza de código que ejecutan trabajos en los nodos worker



Spark

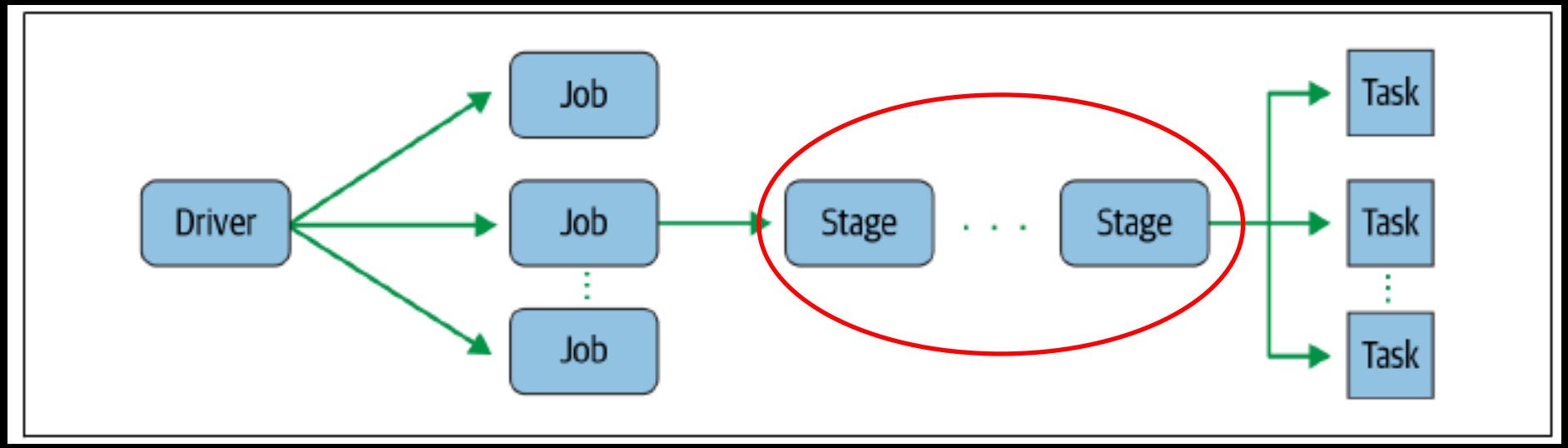
Orquestación



Job: Transformación o acción completa sobre un RDD (en general)

Spark

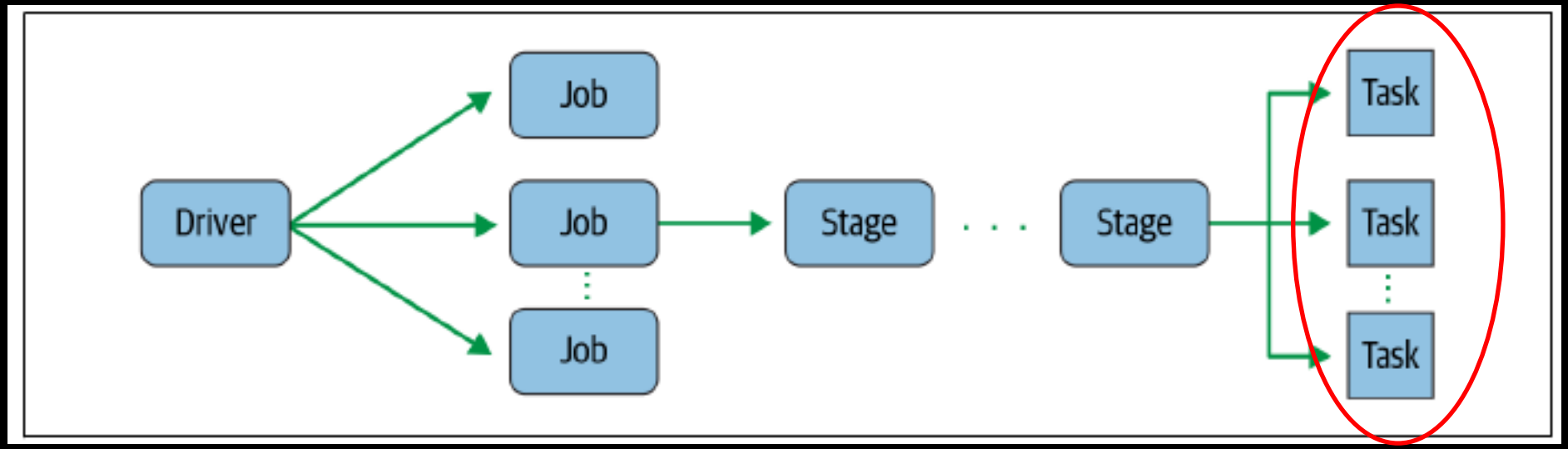
Orquestación



Stage: Etapas secuenciales en las que se descompone un job

Spark

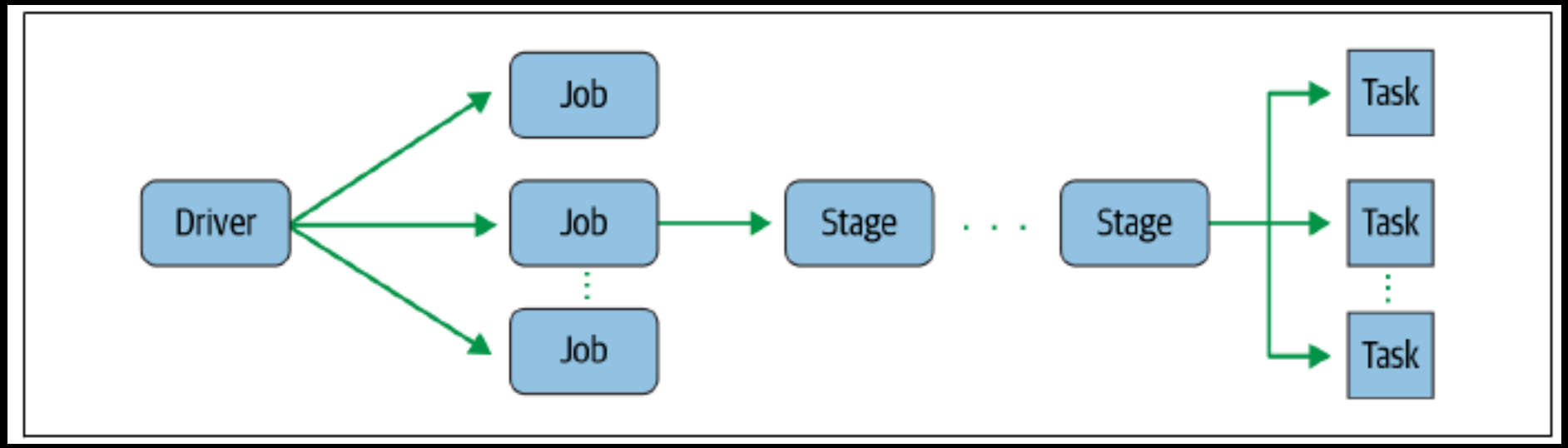
Orquestación



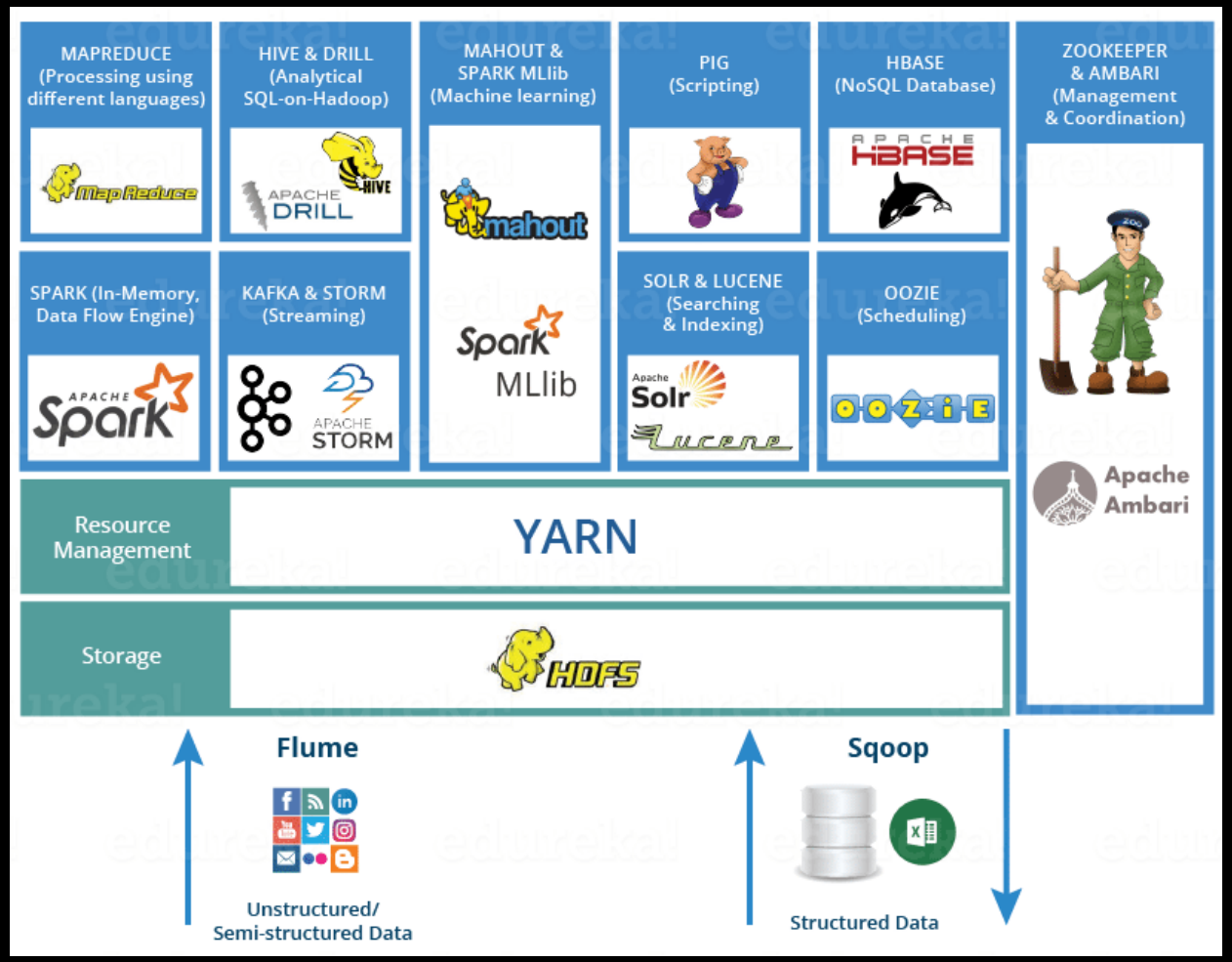
Task: Tareas en las que se divide una etapa o stage y que pueden ejecutarse de forma paralela

Spark

Orquestación



Hadoop Ecosystem



Los destacados (para mí :-)):

- HDFS
- Yarn
- Spark (Spark Mlib)
- Hive
- Kafka

Para más:
https://www.edureka.co/blog/hadoop-ecosystem


THE
BRIDGE