Apache Spark

Apache Spark

- Sistema de procesamiento distribuido.

APIs

- APIs en:
  - Java
  - Scala
  - Python
  - R

APIs

- APIs de alto nivel:
  - RDD API
  - DataFrame API
  - Spark SQL
  - MLlib
  - GraphX
  - GraphFrames
  - Spark Streaming

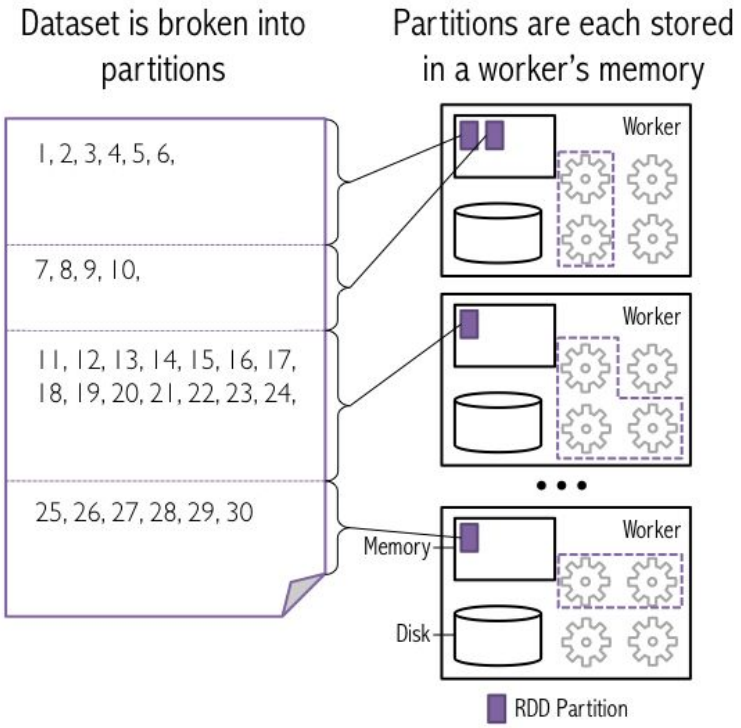
Arquitectura

- Comunicación entre un driver y una serie de ejecutores (executors).
- Tareas (jobs) del driver se convierten en tareas para los executors.
- Los resultados de esas tareas vuelven al driver.

Resilent Distributed Datasets (RDDs)

- Colecciones particionadas en un cluster.
- Guardados en memoria o disco.
- Reconstruidos automáticamente frente a fallos de máquinas o demoras en un job.
- Creados a partir de datos externos.

Resilent Distributed Datasets (RDDs)



Operaciones sobre RDD

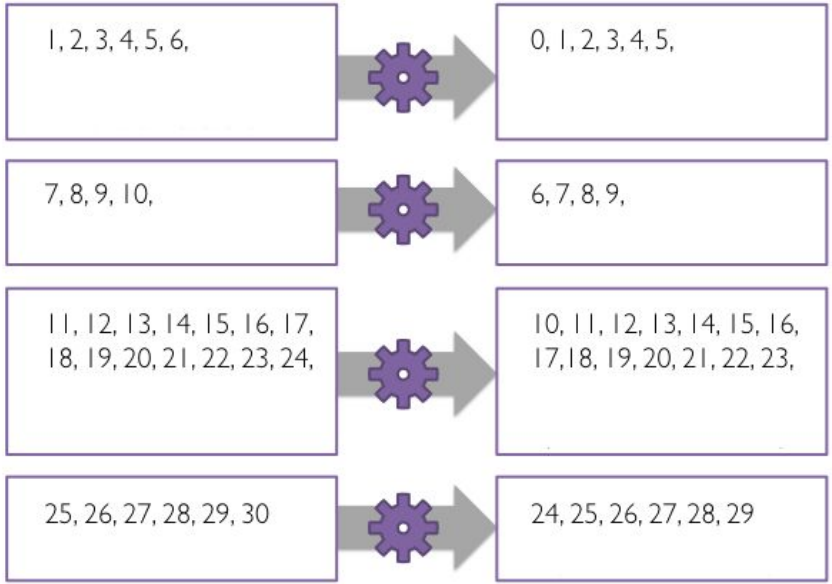
- Transformaciones
- Acciones

Transformaciones

- Crean un nuevo RDD a partir de otro existente.
- Lazy.
- Cache (ram o disco)

Transformaciones: Ejecución

`map(f)`: Each task makes a new partition by calling `f(e)` on each entry `e` in the original partition



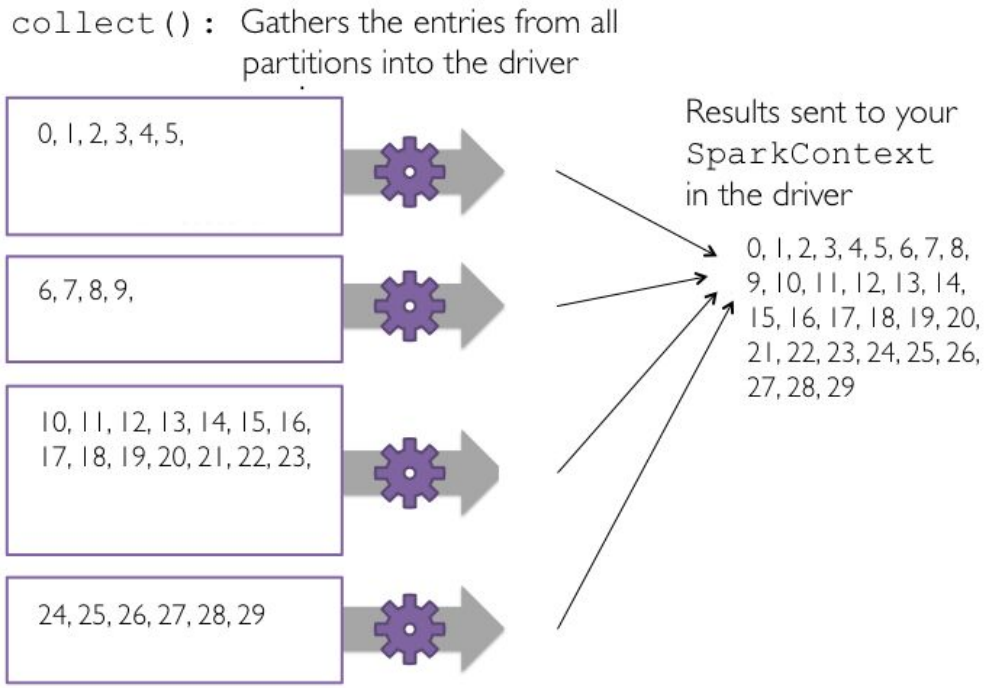
Transformaciones

- Map
- Filter
- FlatMap
- ReduceByKey
- GroupByKey
- Join

Acciones

- Devuelven un valor al driver luego de procesar los datos.

Acciones: Ejecución



Acciones

- Reduce
- Collect
- Count
- Take
- TakeOrdered
- First

RDD Programming Guide

https://spark.apache.org/docs/latest/rdd-programming-guide.html