Apache YARN

FUNDAMENTOS TECNOLÓGICOS PARA BIG DATA

Isaac Esau Rubio Torres 2018

Sistemas de Procesamiento Batch

Apache Yarn

YARN

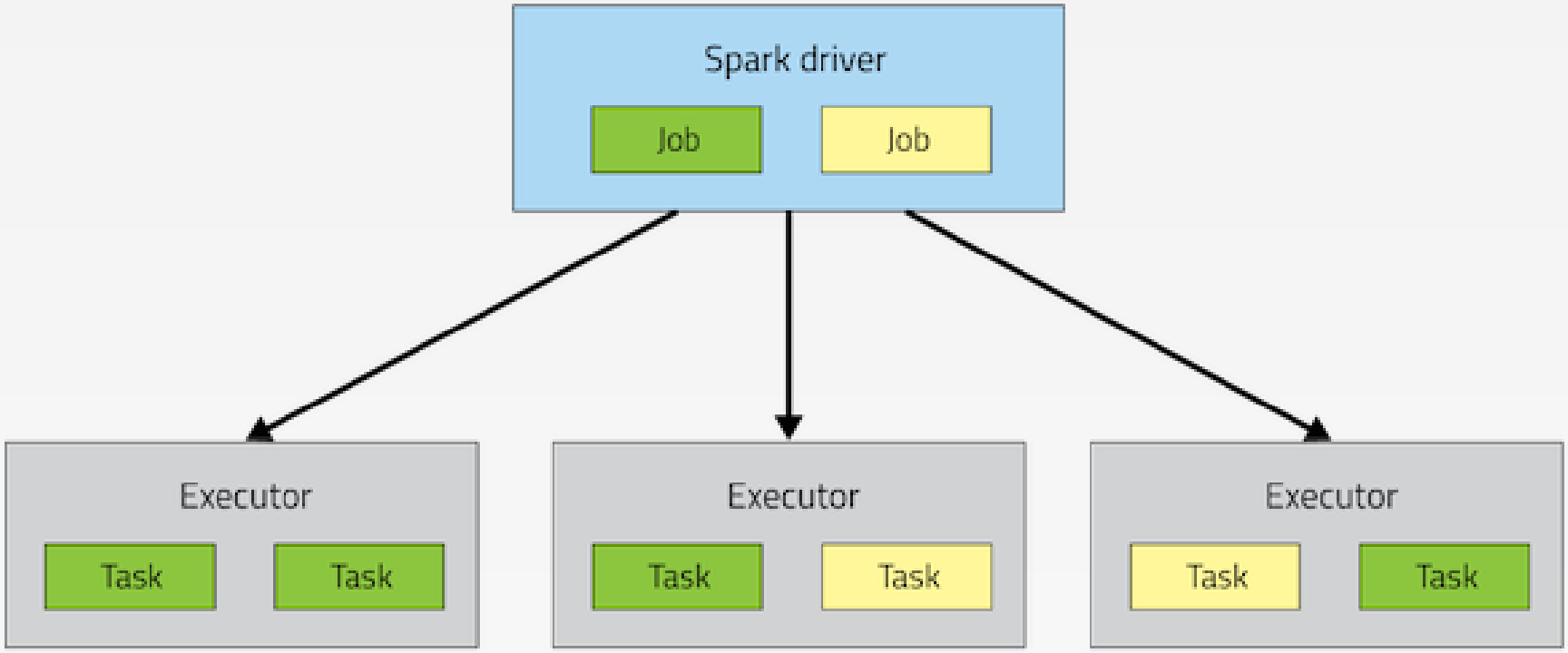
- Apache Hadoop YARN (por las siglas en inglés de “otro negociador de recursos”) es una tecnología de administración de clústeres.
- YARN se caracteriza como un sistema operativo distribuido, a gran escala, para aplicaciones de Big Data.
- YARN combina un administrador central de recursos que reconcilia la forma en que las aplicaciones utilizan los recursos del sistema de Hadoop con los agentes de administración de nodo que monitorean las operaciones de procesamiento de nodos individuales del clúster.

Resumen de conceptos

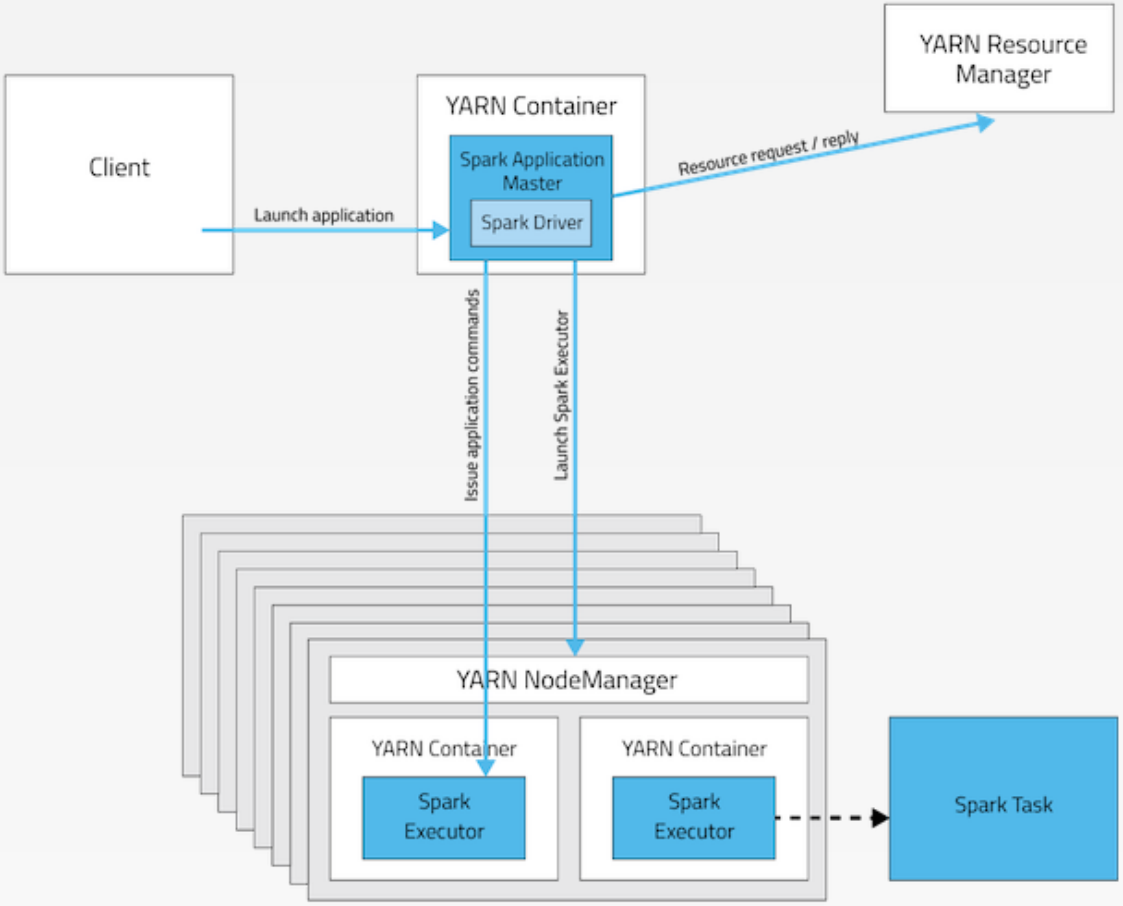
- **Aplicación**: puede ser un solo puesto de trabajo, un conjunto de nodos, un servicio de larga ejecución para emitir nuevos comandos o una sesión de exploración interactiva.
- **Spark Driver**: El controlador Spark es el proceso que se ejecuta el contexto de Spark (que representa la sesión de aplicación). Este controlador es responsable de convertir la aplicación a un grafo dirigido de pasos individuales (DAG) para ejecutar en el clúster. Un controlador por aplicación.
- **Spark Application Master**: se encarga de las solicitudes de recursos de negociación realizadas por el executor y de encontrar un conjunto adecuado de nodos/ contenedores en los que se ejecutan las aplicaciones de Spark. Hay un maestro por aplicación.
- **Spark Executor**: Una sola instancia de JVM en un nodo que sirve a una sola aplicación Spark. El ejecutor ejecuta varias tareas durante su vida útil, y múltiples tareas al mismo tiempo. Un nodo puede tener varios ejecutores de Spark y hay muchos nodos que ejecutan ejecutores para cada aplicación cliente.

- **Tareas**: una tarea Spark representa una unidad de trabajo en una partición de un conjunto de datos distribuidos.

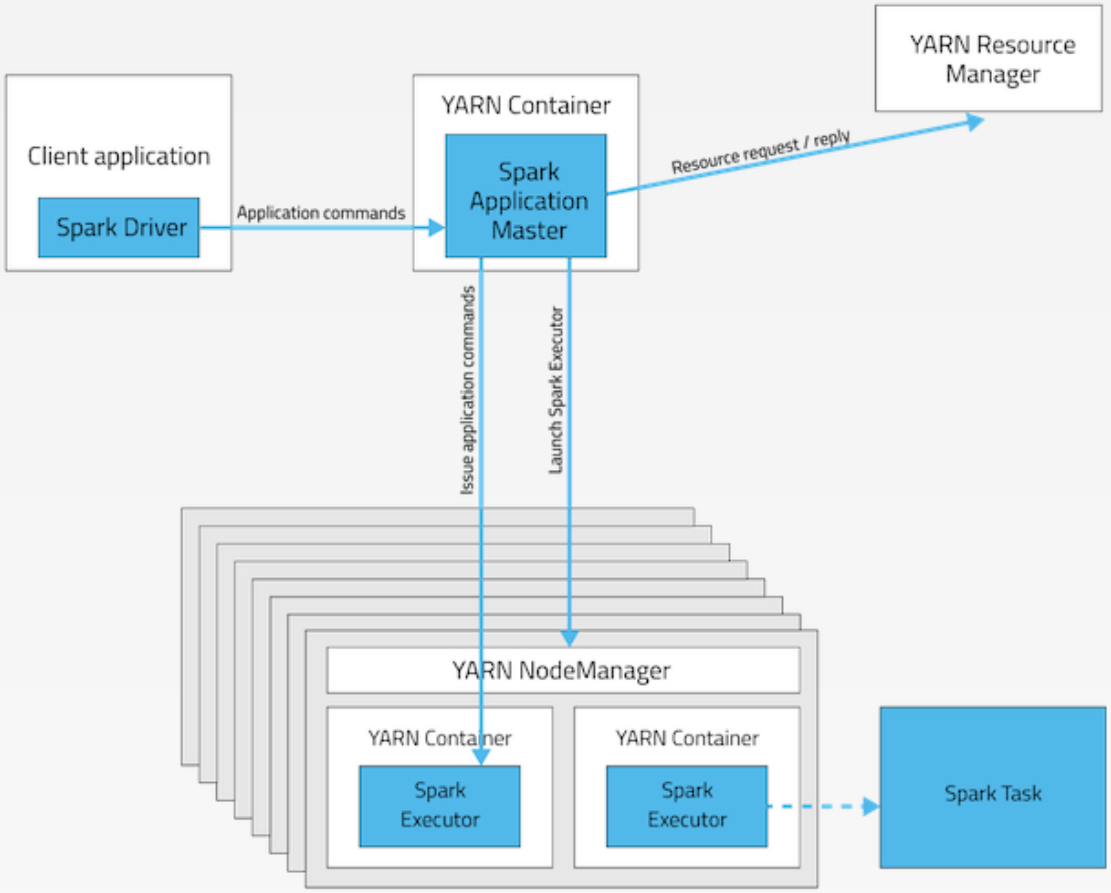
Modelo de ejecución de Spark



Modo Yarn cluster



Modo Yarn client



Diferencias entre modos

	Yarn Cluster	Yarn Client	SparkStandalone
Controlador/driver	Application Master	Cliente	Cliente
¿Quién solicita los recursos?	Application Master	Application Master	Cliente
¿Quién comienza los procesos “executors”?	YARN NodeManager	YARN NodeManager	Spark Slave
Servicios persistentes	YARN ResourceManager y NodeManagers	YARN ResourceManager y NodeManagers	Spark Master y Workers
Soporte para Spark Shell	No	Si	Si