

# Big Data Aplicado

Josep Garcia

[j.garciagarcia@edu.gva.es](mailto:j.garciagarcia@edu.gva.es)



IES EDUARDO  
PRIMO MARQUÉS



# HADOOP



Versiones en la parte de procesos

- MapReduce V1
- MapReduce V2 (YARN)

# Mapreduce. Yarn.

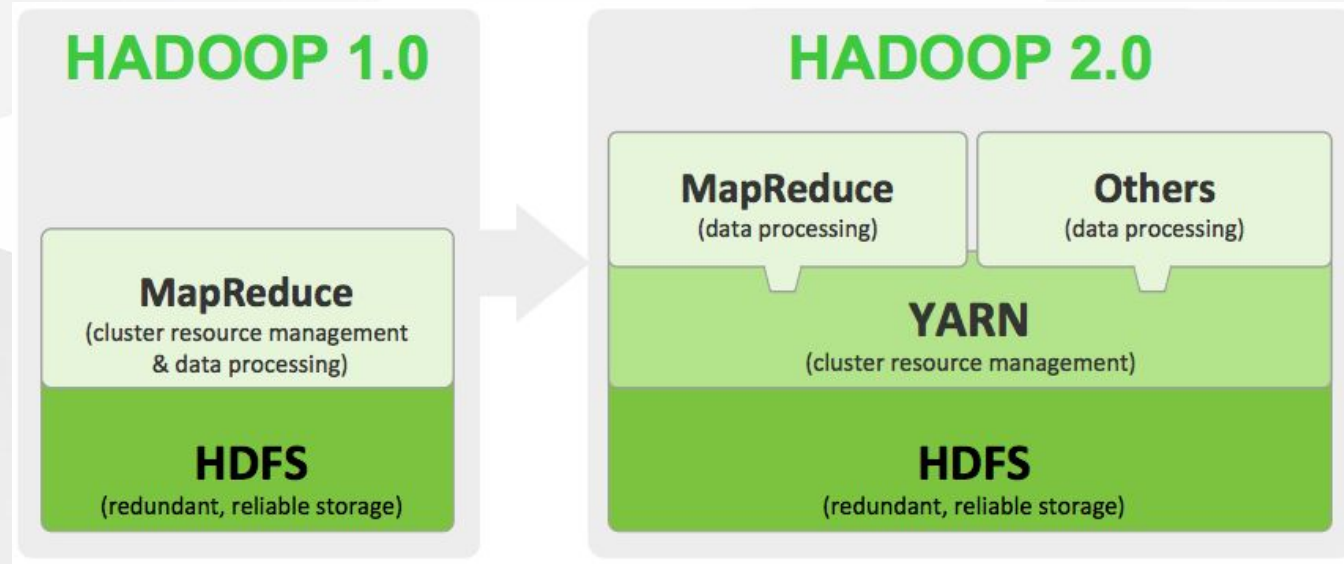
## MapReduce V1

- Pensado para procesos batch.
- Inconveniente: se encarga tanto del proceso de los datos como de la gestión del cluster.

## Yarn

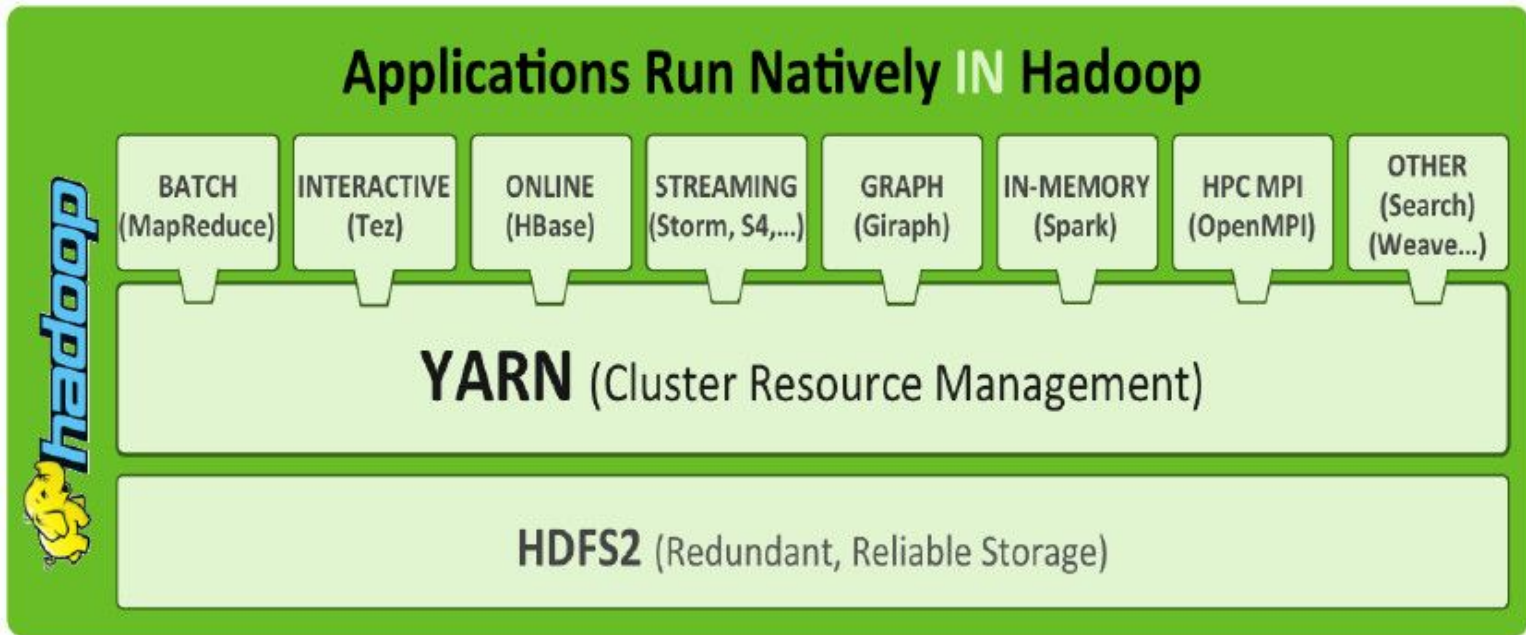
- Admite otro tipo de productos y procesos que no sean batch.
- Procesos distintos para el proceso de datos y para la gestión del cluster.

# HADOOP



En V2: YARN gestiona el clúster y MapReduce pasa a ser una de las opciones para trabajar con procesos.

# HADOOP



# HADOOP 3

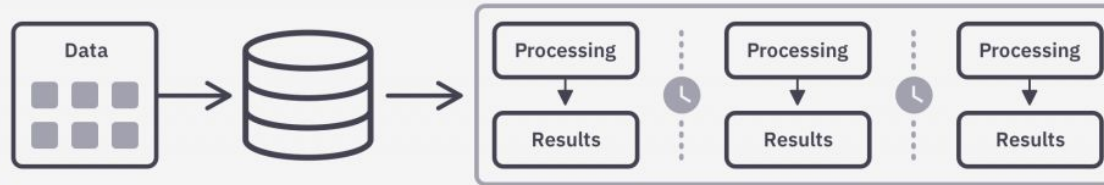


- Mejora la capacidad y eficiencia de YARN.
- Mejora el rendimiento y escalabilidad del clúster.
- Continuación de MapReduce y soporte para otros frameworks como Spark y Flink.
- Ajuste dinámico de recursos para aplicaciones.
- Mejora en resiliencia y recuperación de fallos.
- Soporte para GPU y procesadores no nativos

# Batch vs Streaming



## Batch Processing



## Data Stream Processing



# YARN

## NODO MAESTRO

Resource Manager



## NODOS HIJOS

Node Manager

Node Manager

Node Manager



IES EDUARDO  
PRIMO MARQUÉS



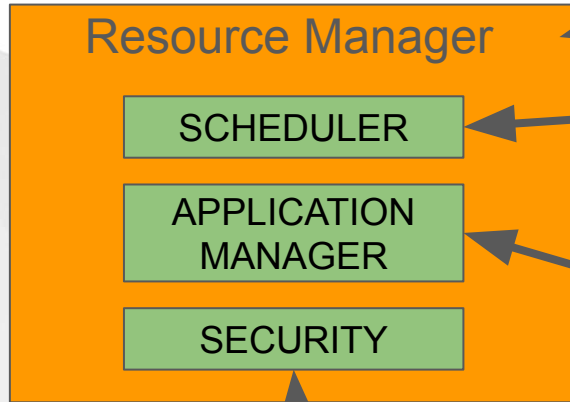


# YARN



## NODO MAESTRO

Gestor de todo YARN



Determina cómo se planifican los trabajos. Decide qué recursos tiene un nodo para ser utilizados por el app manager.

Cuando alguien solicita una aplicación, arranca un application manager en cada nodo para que se encargue.

También implementa medidas de seguridad.

# YARN



¿Cómo se lanza una aplicación?

Se solicita al **application manager** que se abra un APP MASTER.

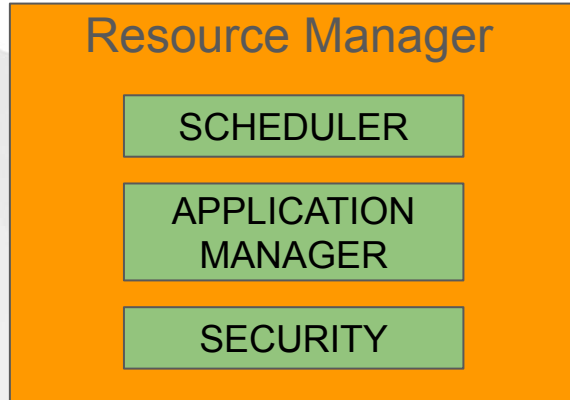
El application manager tiene en cuenta la información del **scheduler** para decidir en qué nodo abre este APP MASTER.

El APP MASTER es un coordinador de la aplicación, si tenemos 100 aplicaciones lanzadas, habrá 100 APP MASTER.

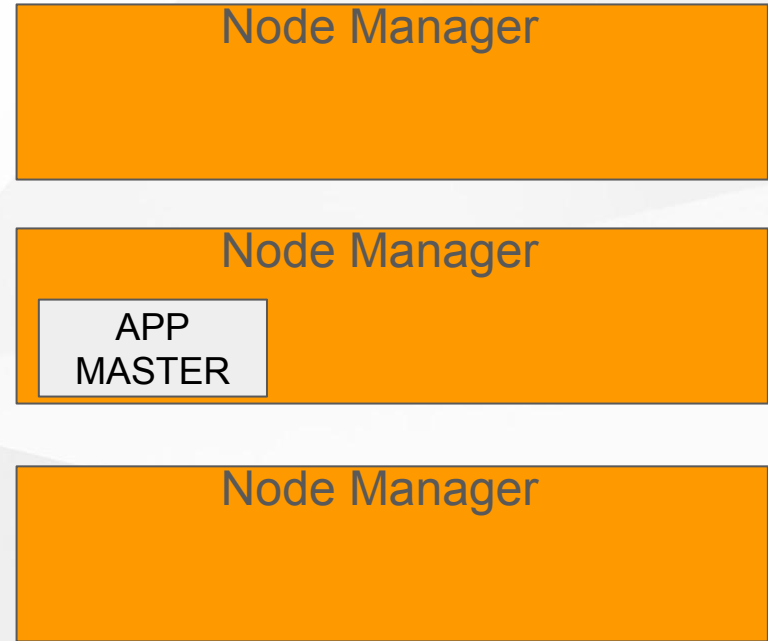
# YARN



## NODO MAESTRO



## NODOS HIJOS



# YARN



¿Qué hace el APP MASTER?

Crea los contenedores necesarios para ejecutar la aplicación.

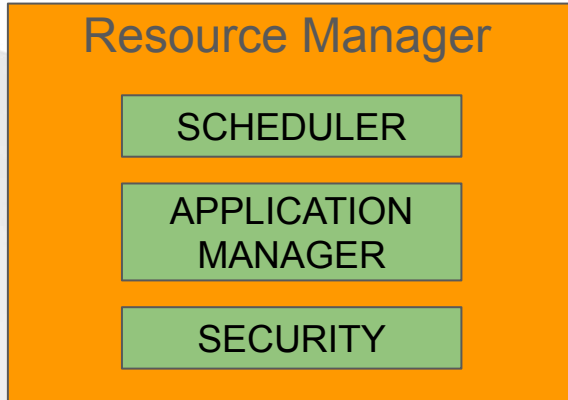
Un contenedor es un sitio donde se ejecuta algo.

En una aplicación MAPREDUCE podría crearse un escenario similar al siguiente:

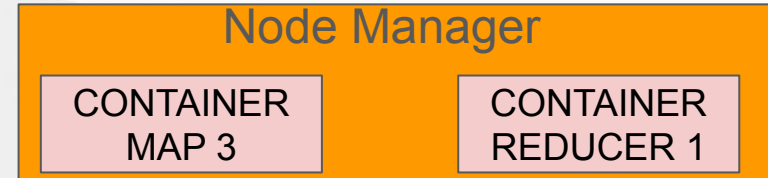
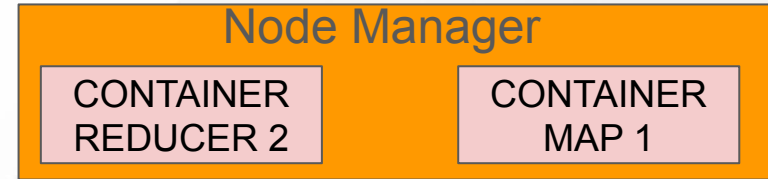
# YARN



## NODO MAESTRO



## NODOS HIJOS



# YARN



APP MASTER: según la información de scheduler determina en qué nodo se ejecuta cada contenedor.

\*\* Cada nodo tiene que acceder a su bloque de datos en **local**, por lo que el “CONTAINER MAP 1”, se ejecuta en el mismo lugar en el que se encuentran los datos.