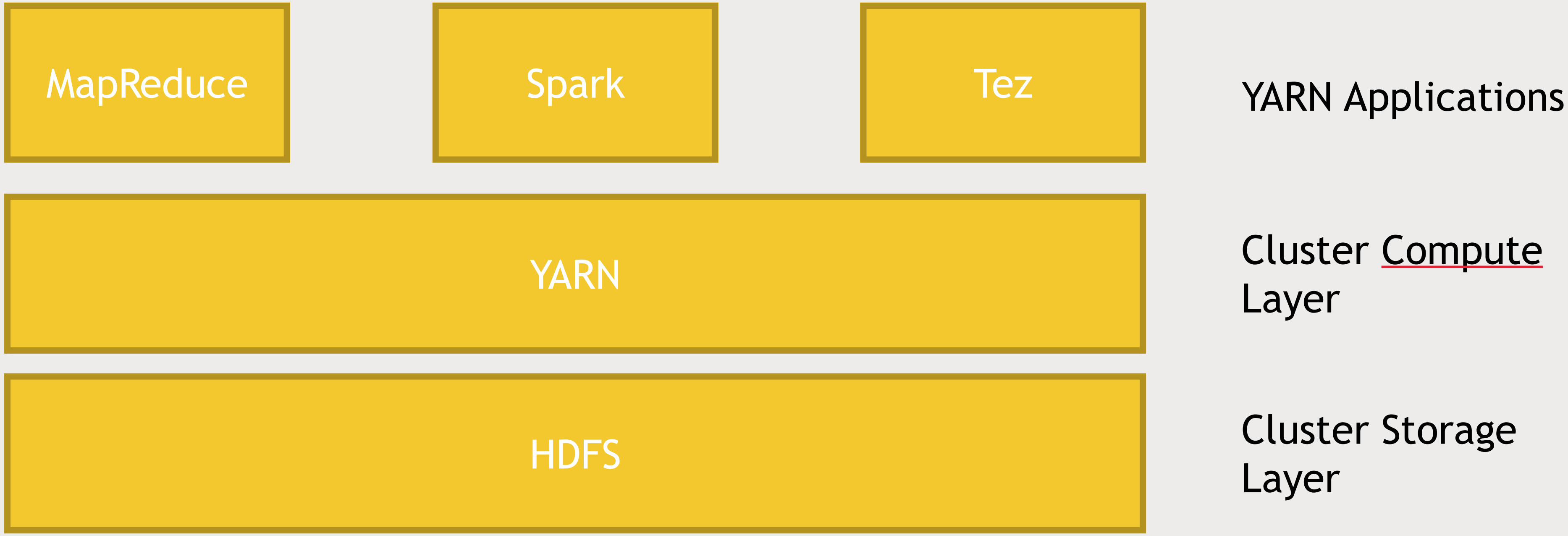
HADOOP YARN

Yet Another Resource Negotiator

What is YARN?

- Yet Another Resource Negotiator
  - *Introduced in Hadoop 2*
  - *Separates the problem of managing resources on your cluster from MapReduce*
  - *Enabled development of MapReduce alternatives (Spark, Tez) built on top of YARN*
- It’s just there, under the hood, managing the usage of your cluster
  - *I can’t think of a reason why you’d need to actually write code against it yourself in this day and age. But you can.*

Where YARN fits in



MapReduce

Spark

Tez

YARN Applications

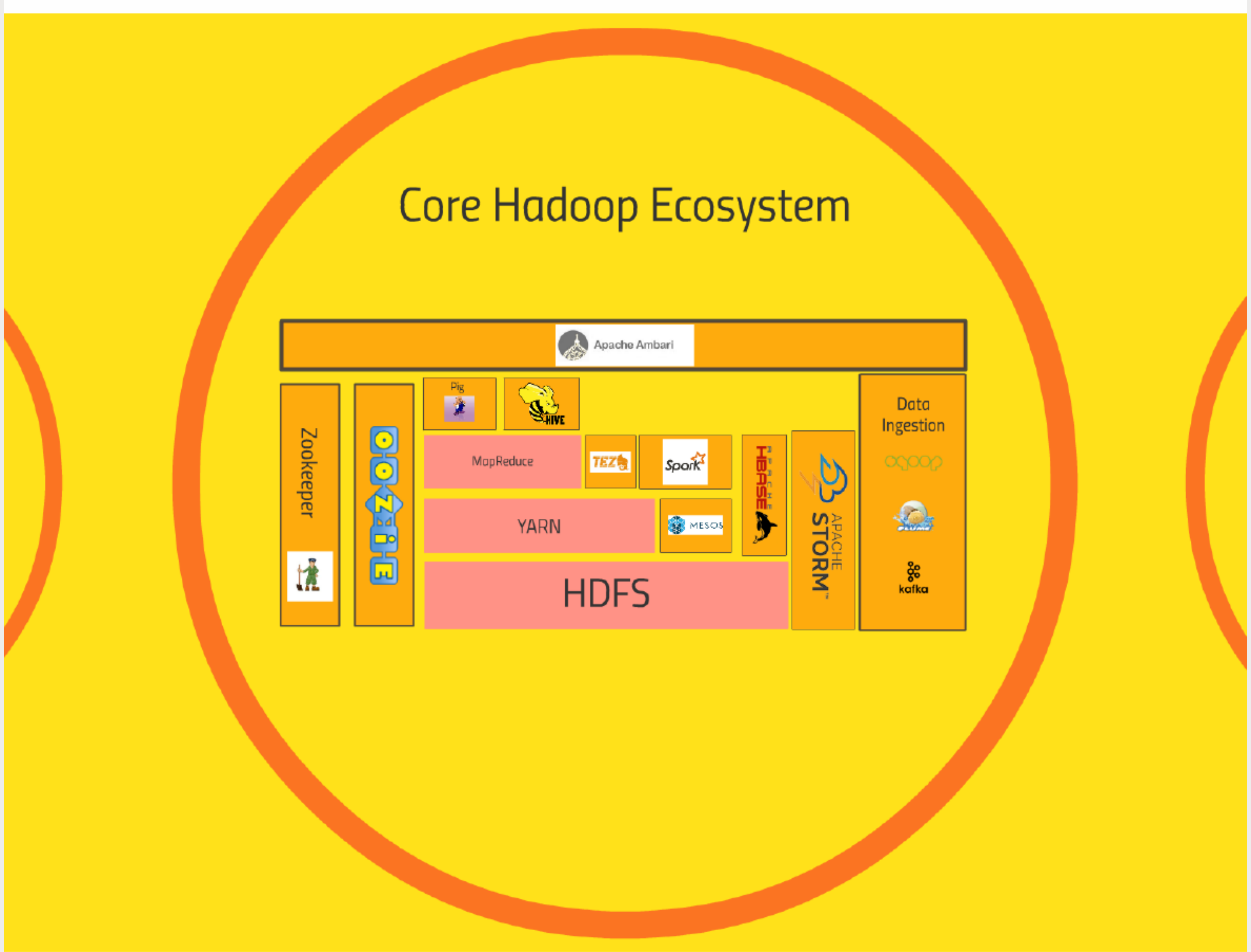
YARN

Cluster Compute Layer

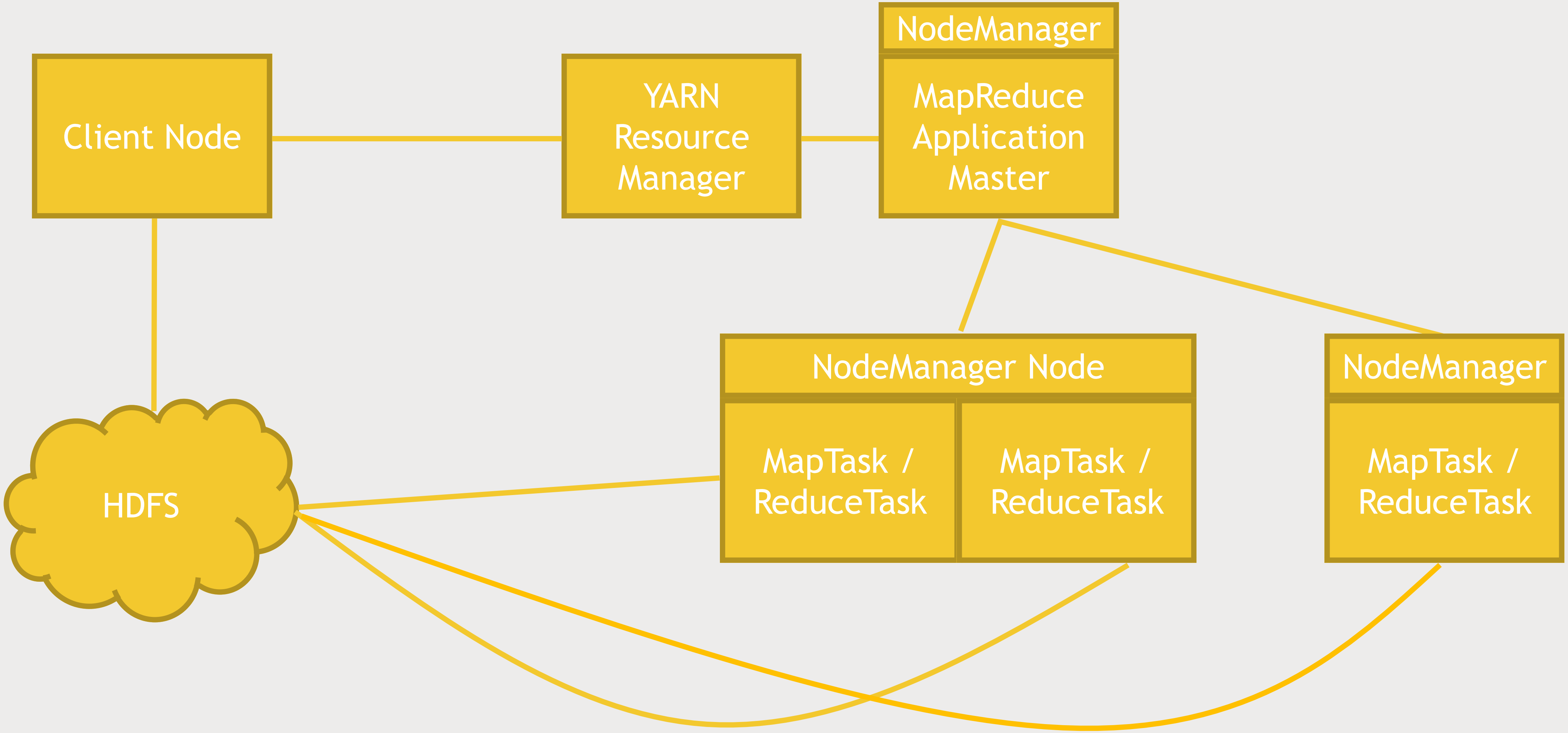
HDFS

Cluster Storage Layer

Where YARN fits in



Remember how MapReduce works



Client Node

YARN Resource Manager

NodeManager

MapReduce Application Master

HDFS

NodeManager Node

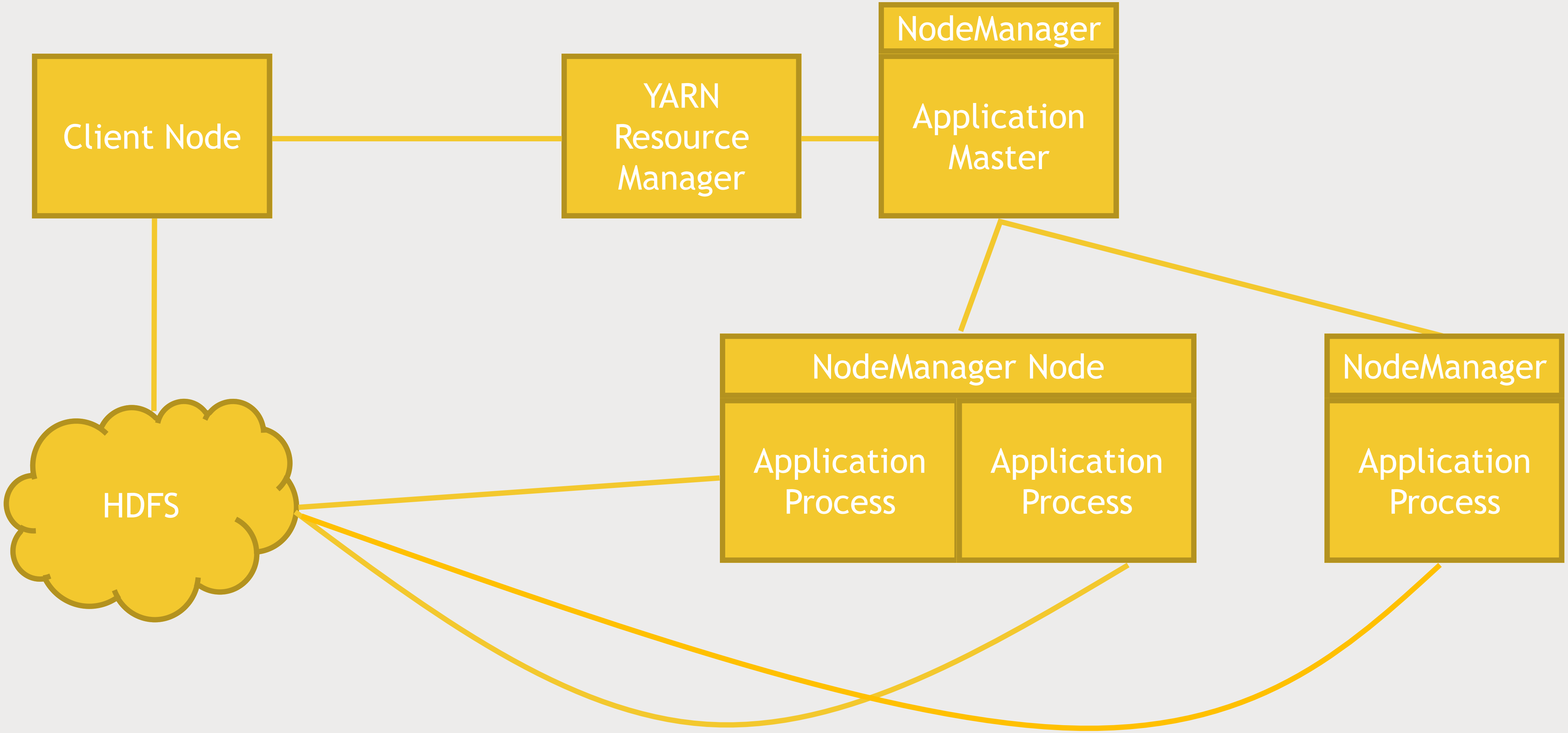
MapTask / ReduceTask

MapTask / ReduceTask

NodeManager

MapTask / ReduceTask

YARN just generalizes this



Client Node

YARN Resource Manager

NodeManager

Application Master

HDFS

NodeManager Node

Application Process

Application Process

NodeManager

Application Process

How YARN works

- Your application talks to the Resource Manager to distribute work to your cluster
- You can specify data locality – which HDFS block(s) do you want to process?
  - *YARN will try to get your process on the same node that has your HDFS blocks*
- You can specify different scheduling options for applications
  - *So you can run more than one application at once on your cluster*
  - *FIFO, Capacity, and Fair schedulers*

    first in first out

    - FIFO runs jobs in sequence, first in first out
    - Capacity may run jobs in parallel if there’s enough spare capacity
    - Fair may cut into a larger running job if you just want to squeeze in a small one

Building new YARN applications

- Why? There are so many existing projects you can just use
  - *Need a DAG*-based application? Build it on Spark or Tez*
    - (*Directed Acyclic Graph)
- But if you really really need to
  - *There are frameworks: Apache Slider, Apache Twill*
  - *And there are some books on the topic.*

And that’s really all there is to say.

- Want to practice “using YARN?” Well, we already did that with MapReduce and Spark!
- You just need to know it’s there, under the hood, managing your cluster’s resources for you
- Thanks YARN!