YARN

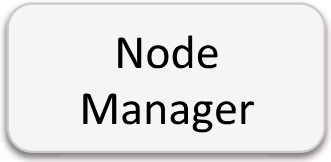
What is YARN?

- **YARN = Yet Another Resource Negotiator**
- **YARN is the Hadoop processing layer that contains**
  - A resource manager
  - A job scheduler
- **YARN allows multiple data processing engines to run on a single Hadoop cluster**
  - Batch programs (e.g. Spark, MapReduce)
  - Interactive SQL (e.g. Impala)
  - Advanced analytics (e.g. Spark, Impala)
  - Streaming (e.g. Spark Streaming)

YARN Daemons

- **Resource Manager (RM)**
  - Runs on master node
  - Global resource scheduler
  - Arbitrates system resources between competing applications
  - Has a pluggable scheduler to support different algorithms (capacity, fair scheduler, etc.)

- **Node Manager (NM)**
  - Runs on slave nodes
  - Communicates with RM



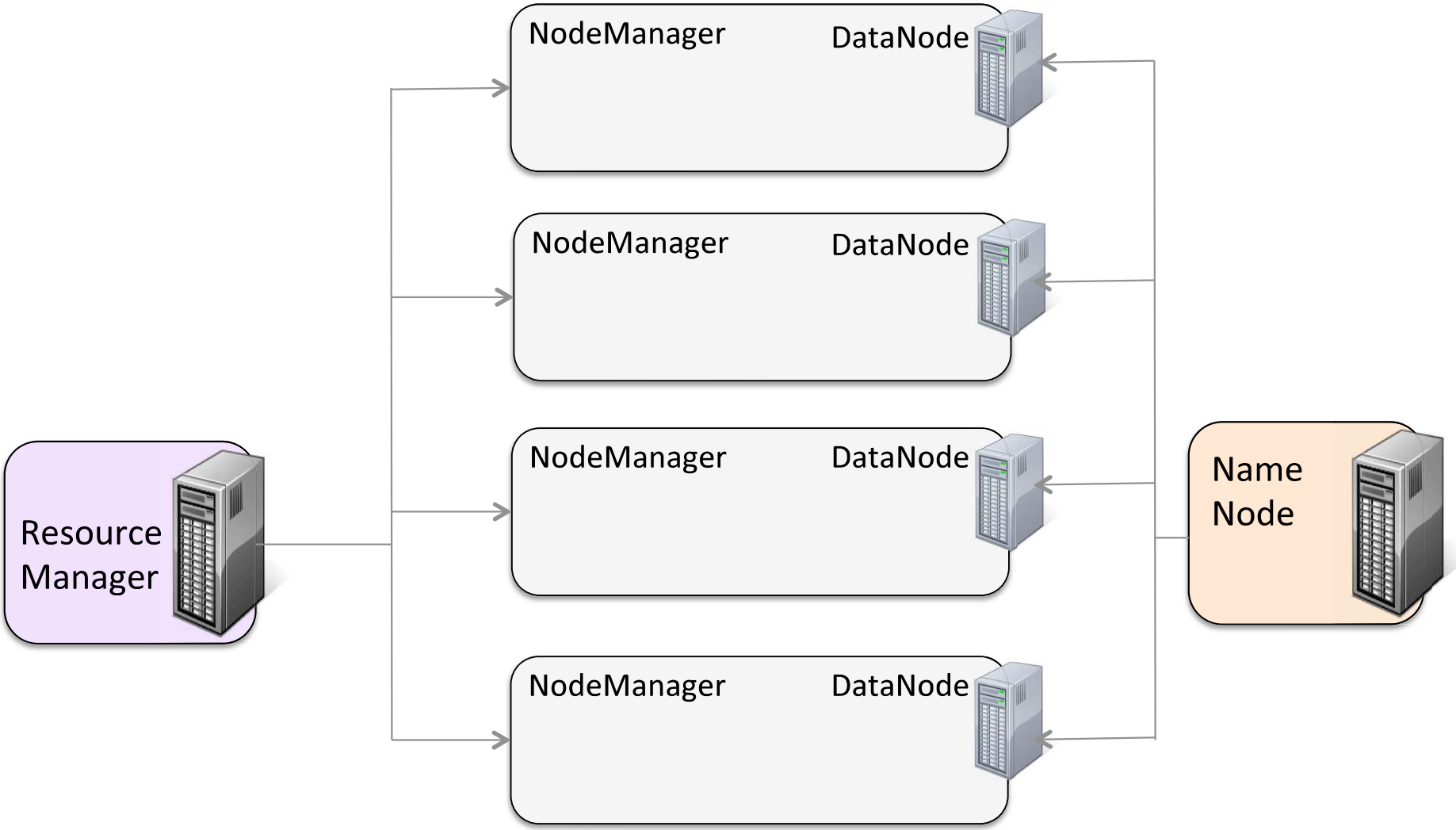
What are HDFS coun

Running an Application in YARN

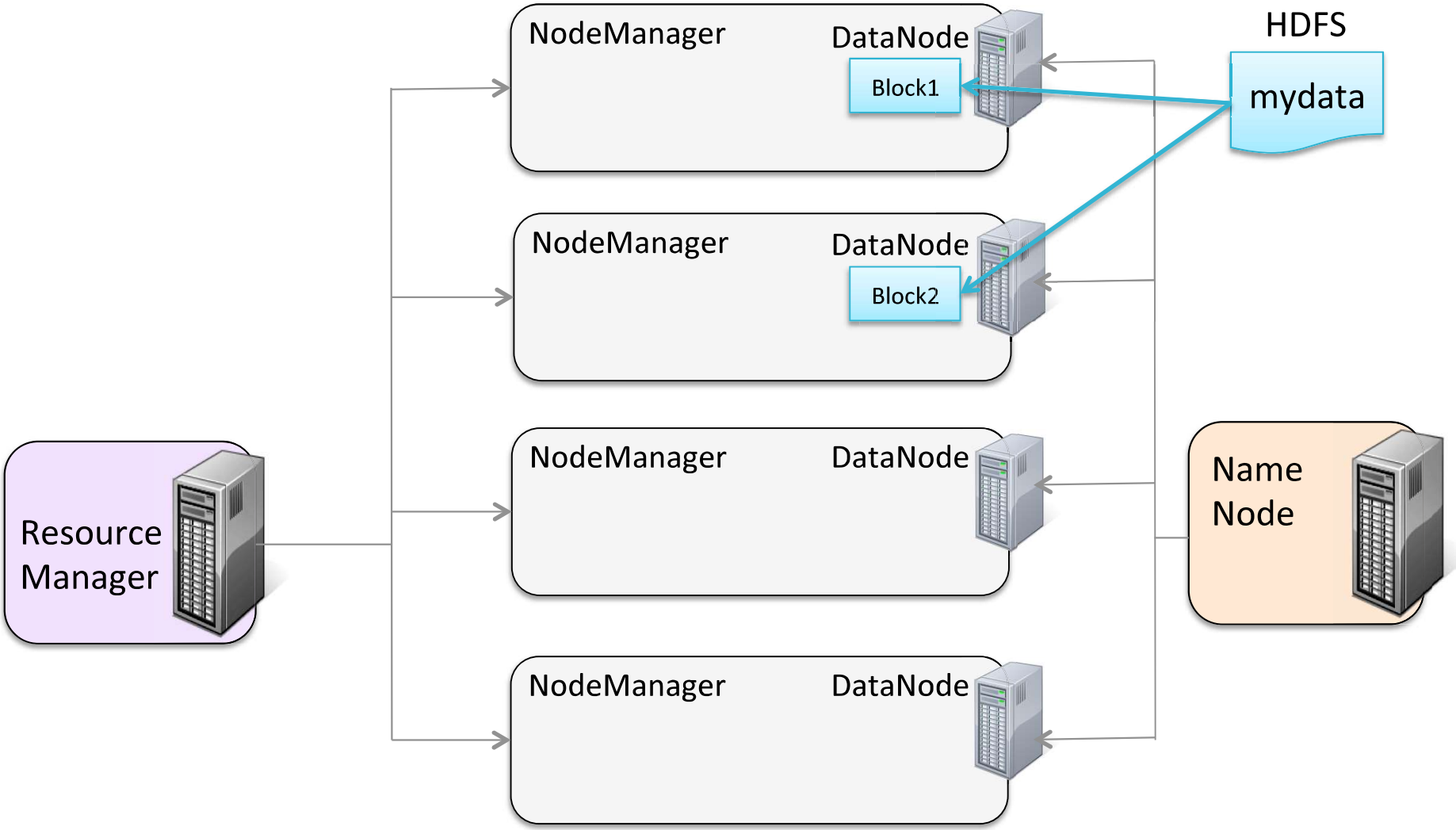
- **Containers**
  - Created by the RM upon request
  - Allocate a certain amount of resources (memory, CPU) on a slave node
  - Applications run in one or more containers

- **Application Master (AM)**
  - One per application
  - Framework/application specific
  - Runs in a container
  - Requests more containers to run application tasks

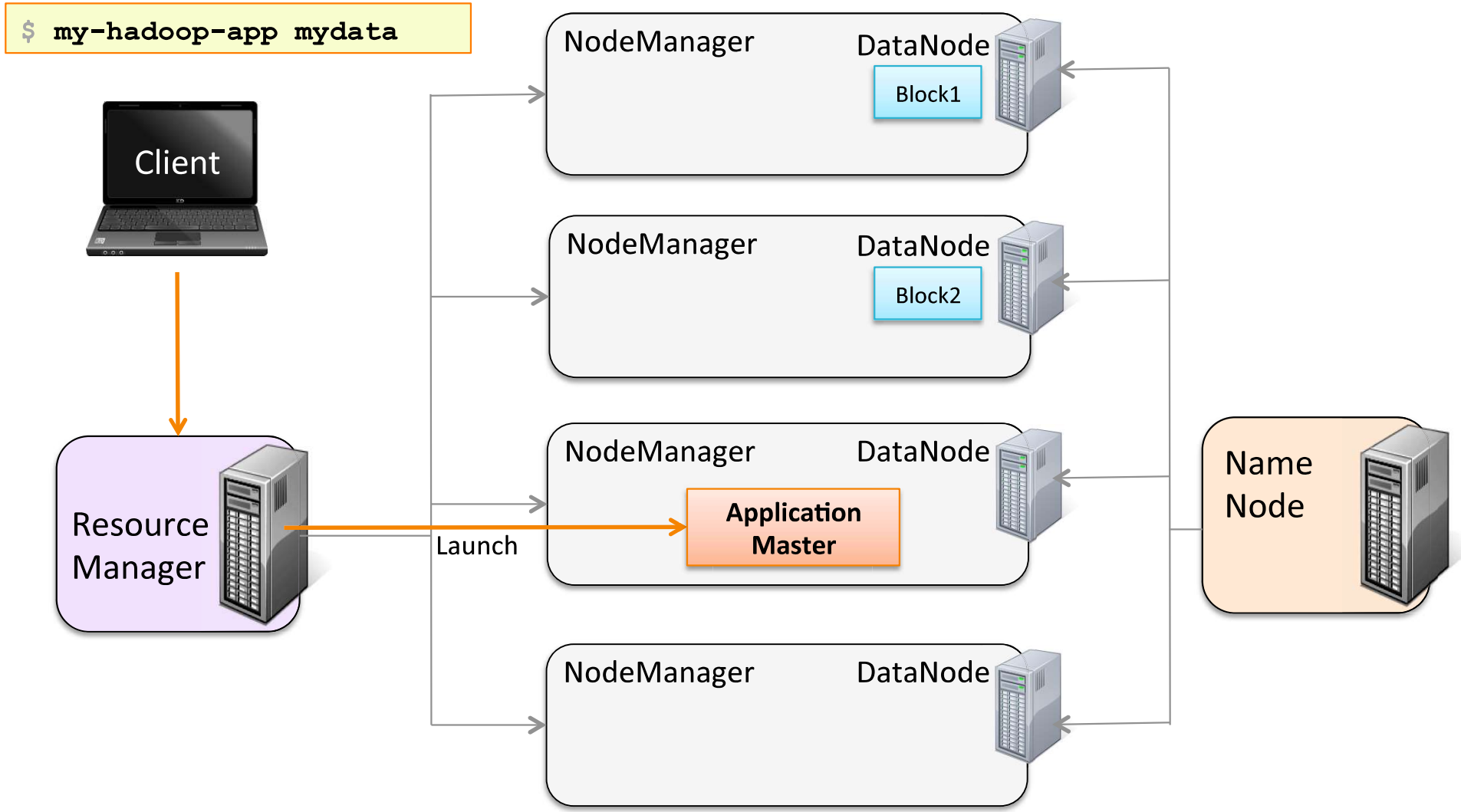
Running an Application on YARN (1)



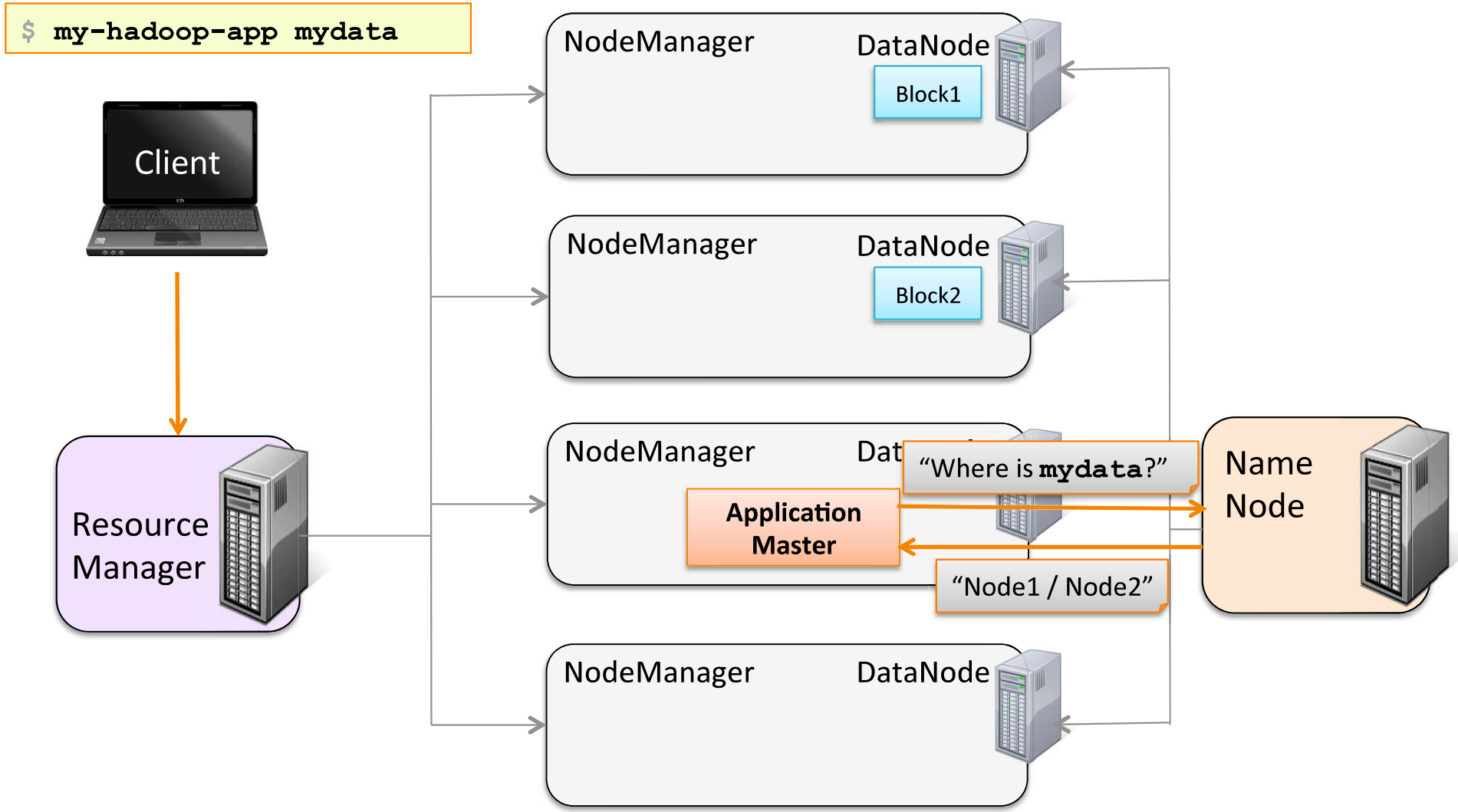
Running an Application on YARN (2)



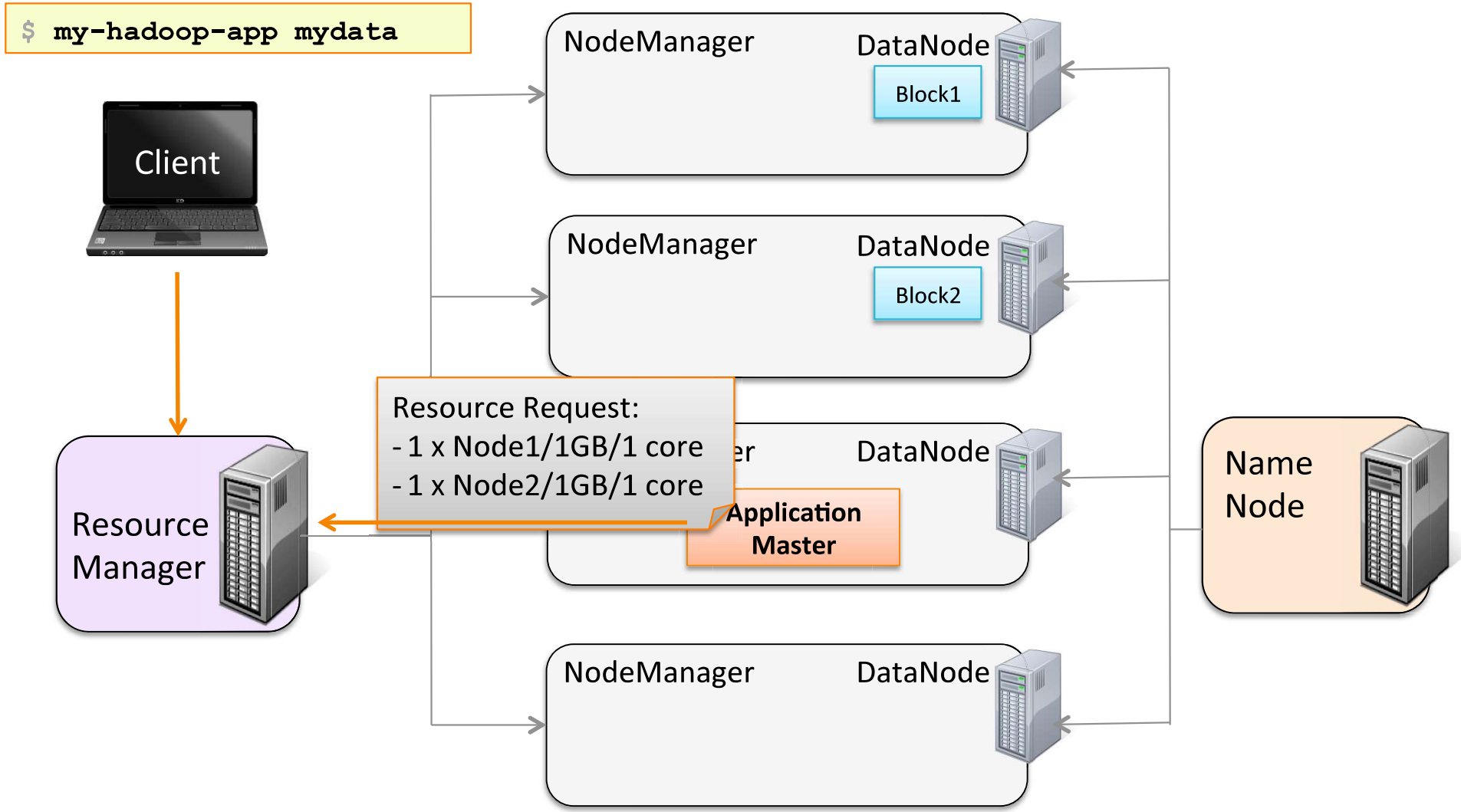
Running an Application on YARN (3)



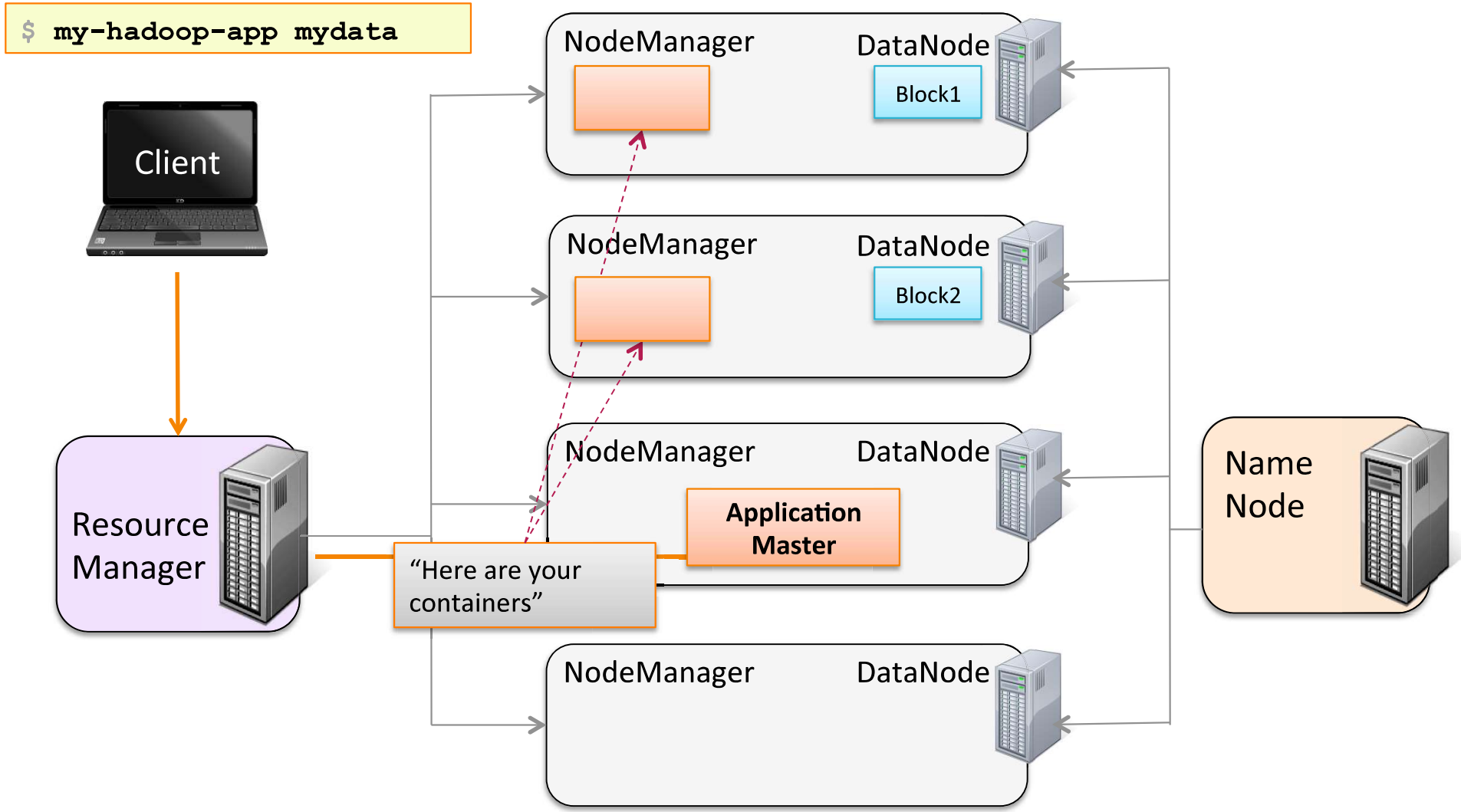
Running an Application on YARN (4)



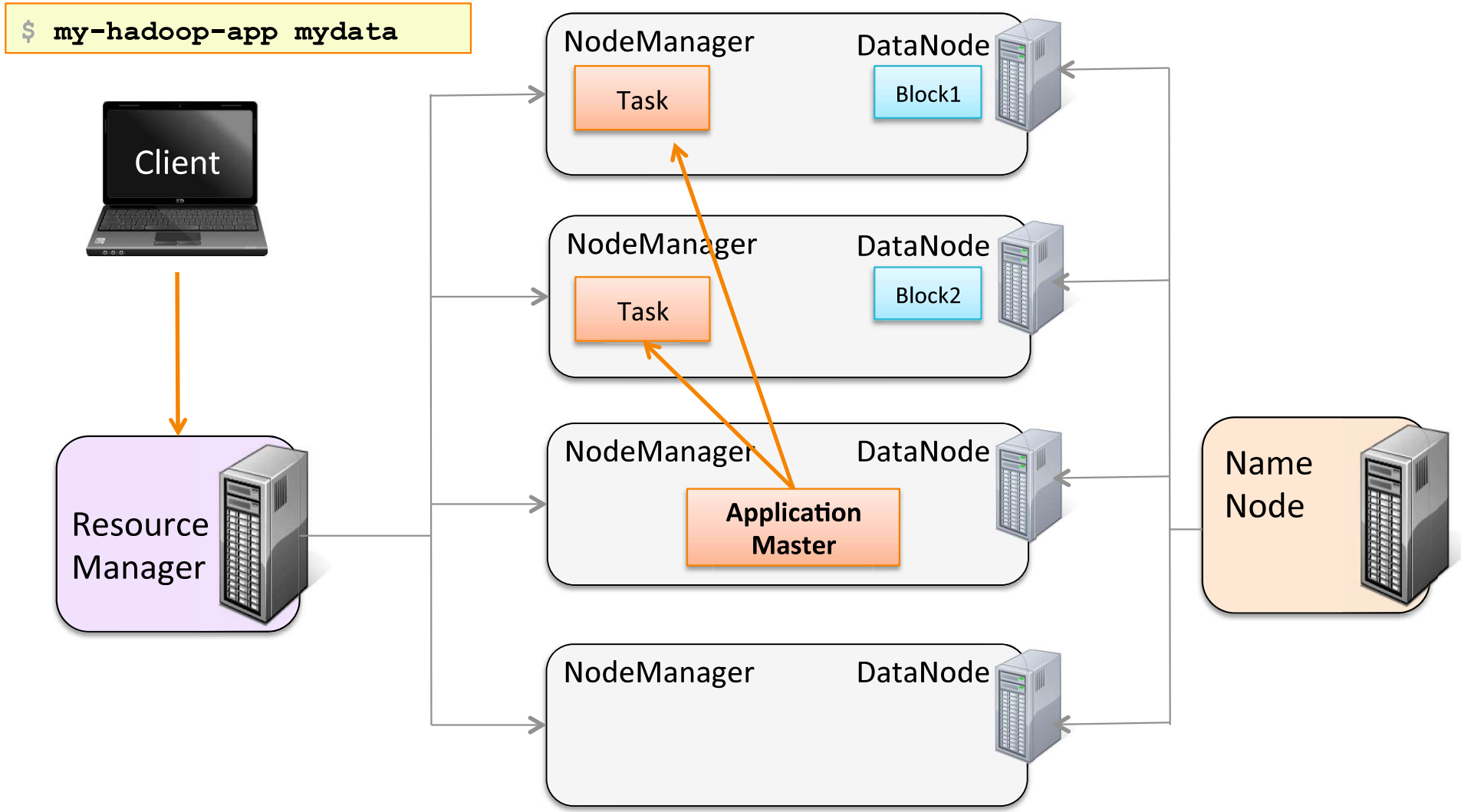
Running an Application on YARN (5)



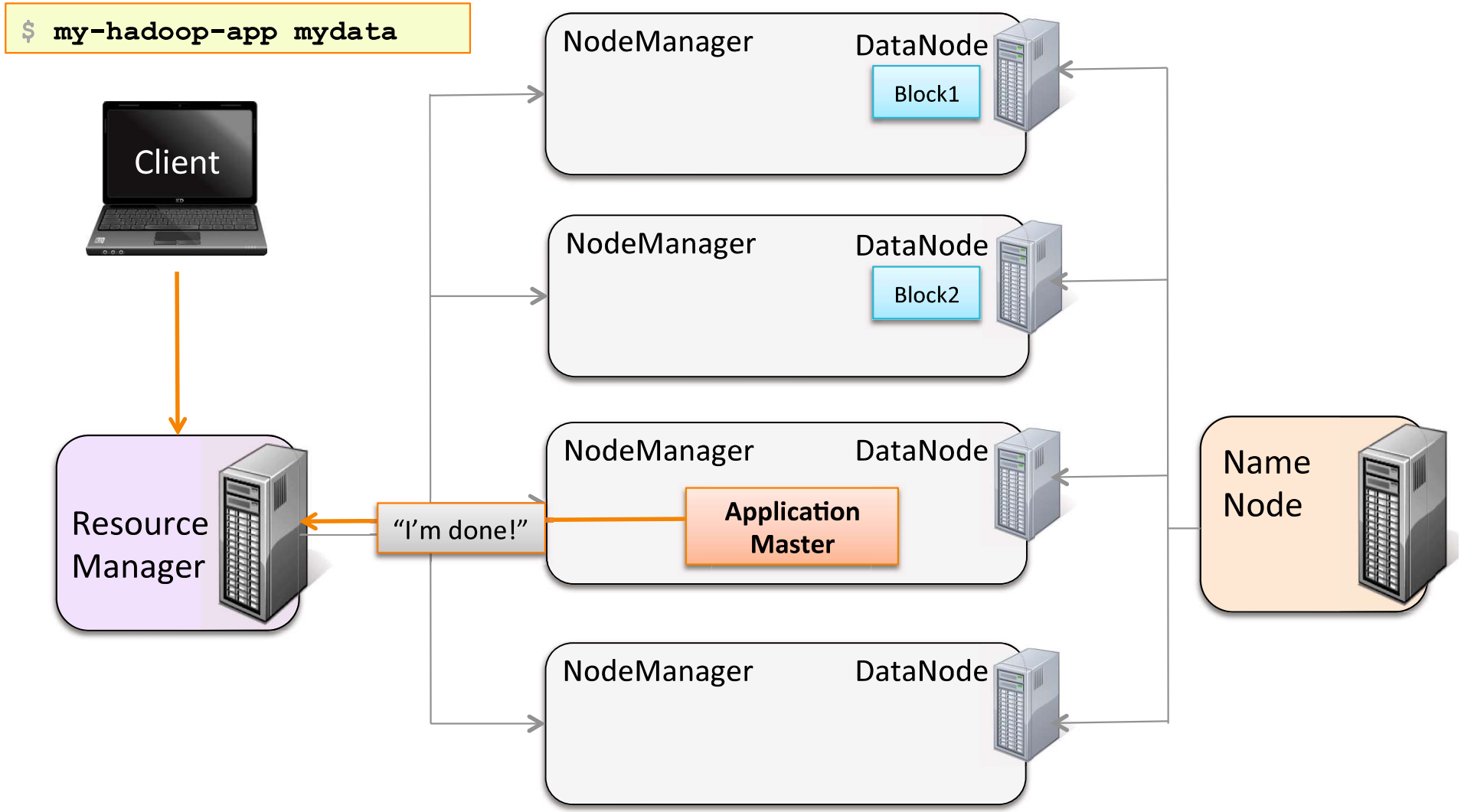
Running an Application on YARN (6)



Running an Application on YARN (7)



Running an Application on YARN (8)

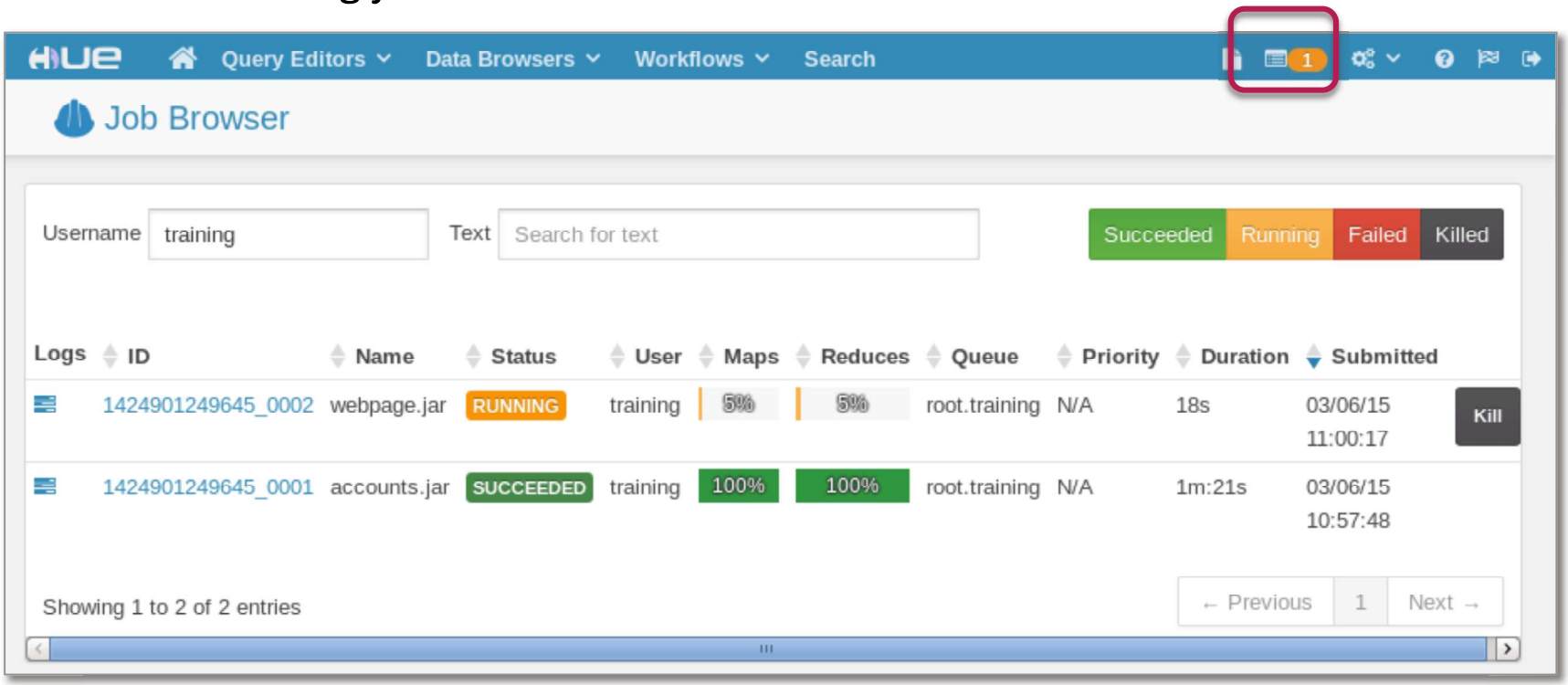


Working With YARN

- **Developers need to be able to**
  - Submit jobs (applications) to run on the YARN cluster
  - Monitor and manage jobs
- **Hadoop includes three major YARN tools for developers**
  - The Hue Job Browser
  - The YARN Web UI
  - The YARN command line
- **YARN administrators can use Cloudera Manager**
  - May also be helpful for developers
  - Included in Cloudera Express and Cloudera Enterprise
  - Not covered in this course

The Hue Job Browser

- **The Hue Job Browser allows you to**
  - Monitor the status of a job
  - View the logs
  - Kill a running job

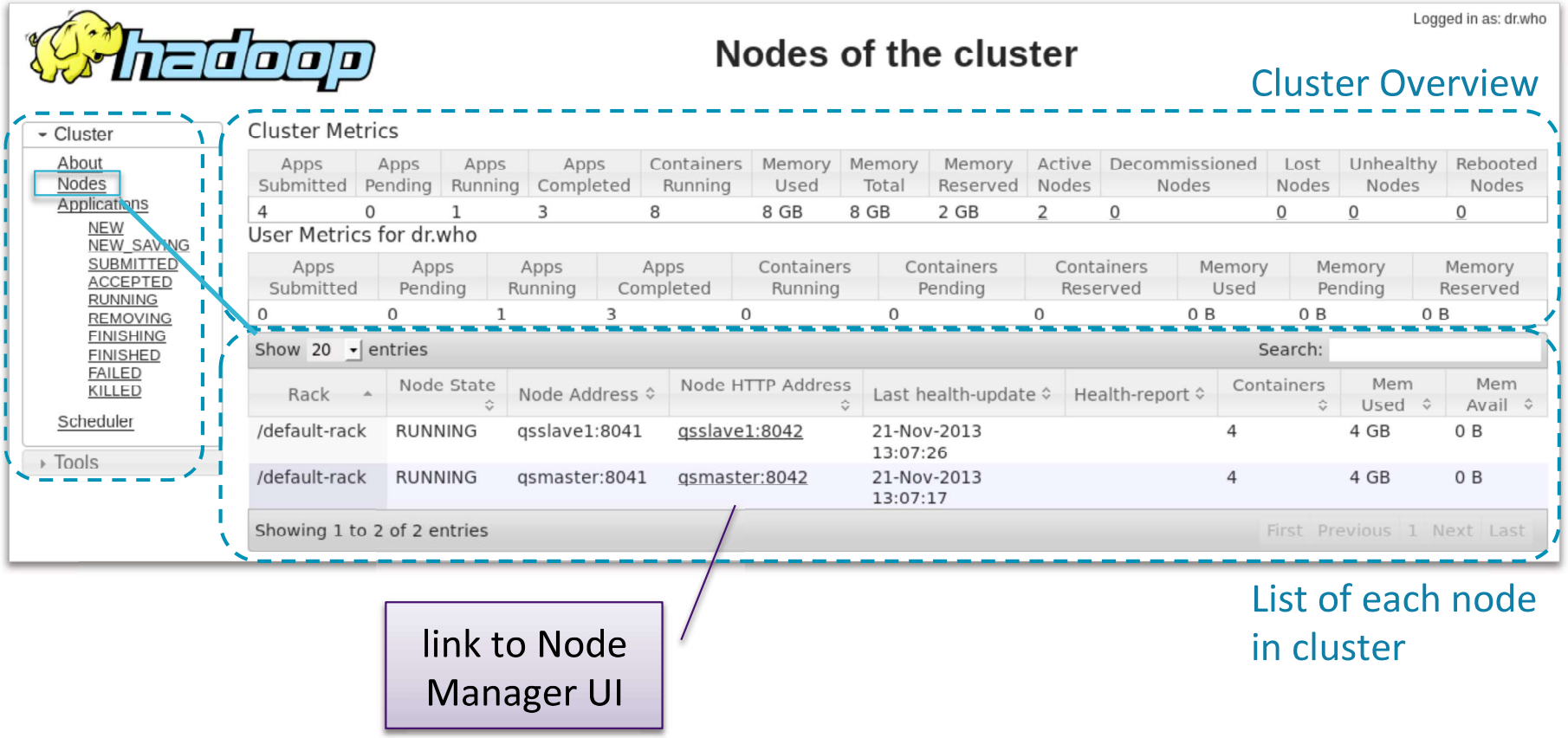


The YARN Web UI

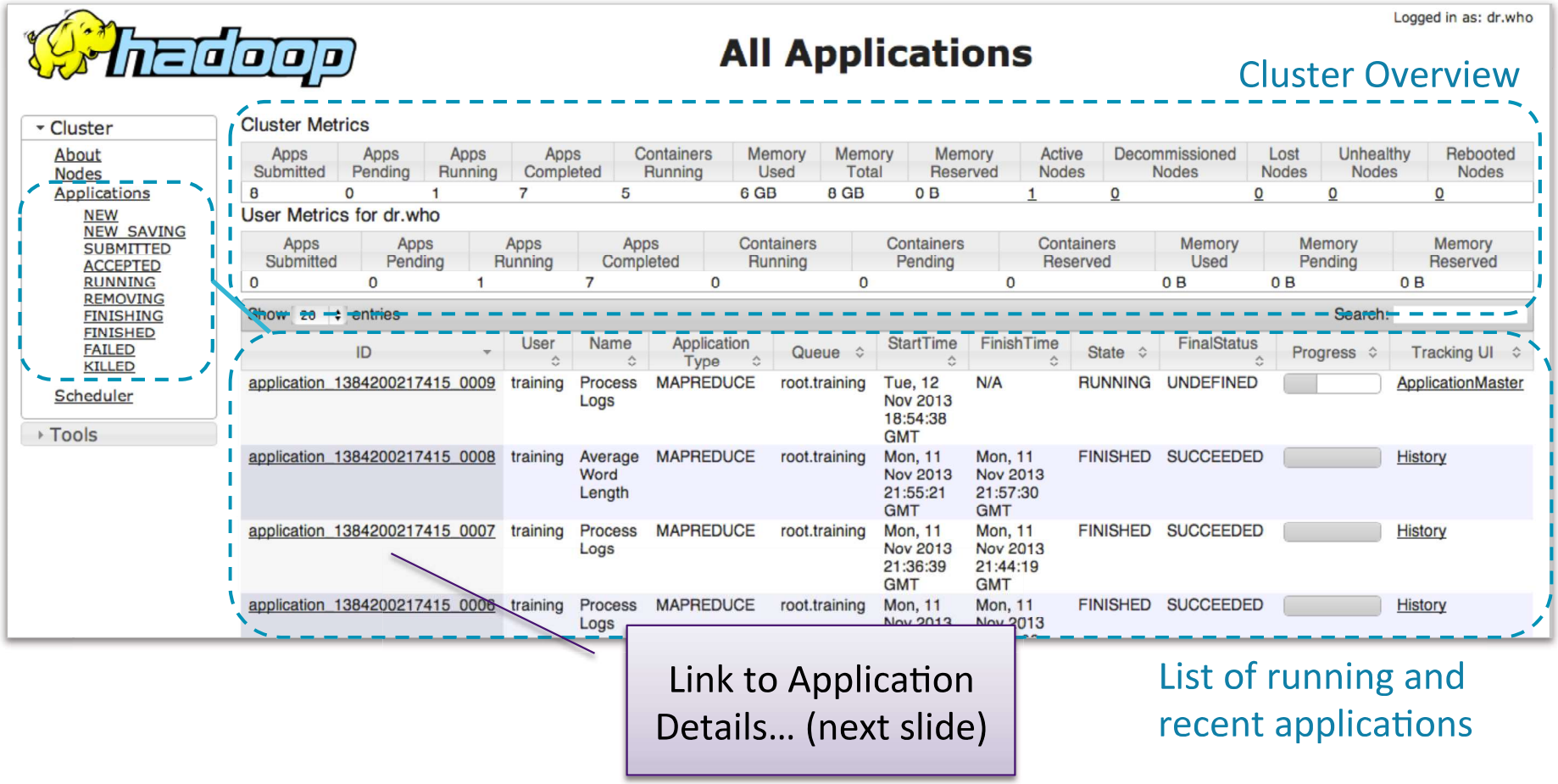
- **Resource Manager UI is the main entry point**
  - Runs on the RM host on port 8080 by default
- **Provides more detailed view than Hue**
- **Does not provide any control or configuration**

also try 8088 (in rece

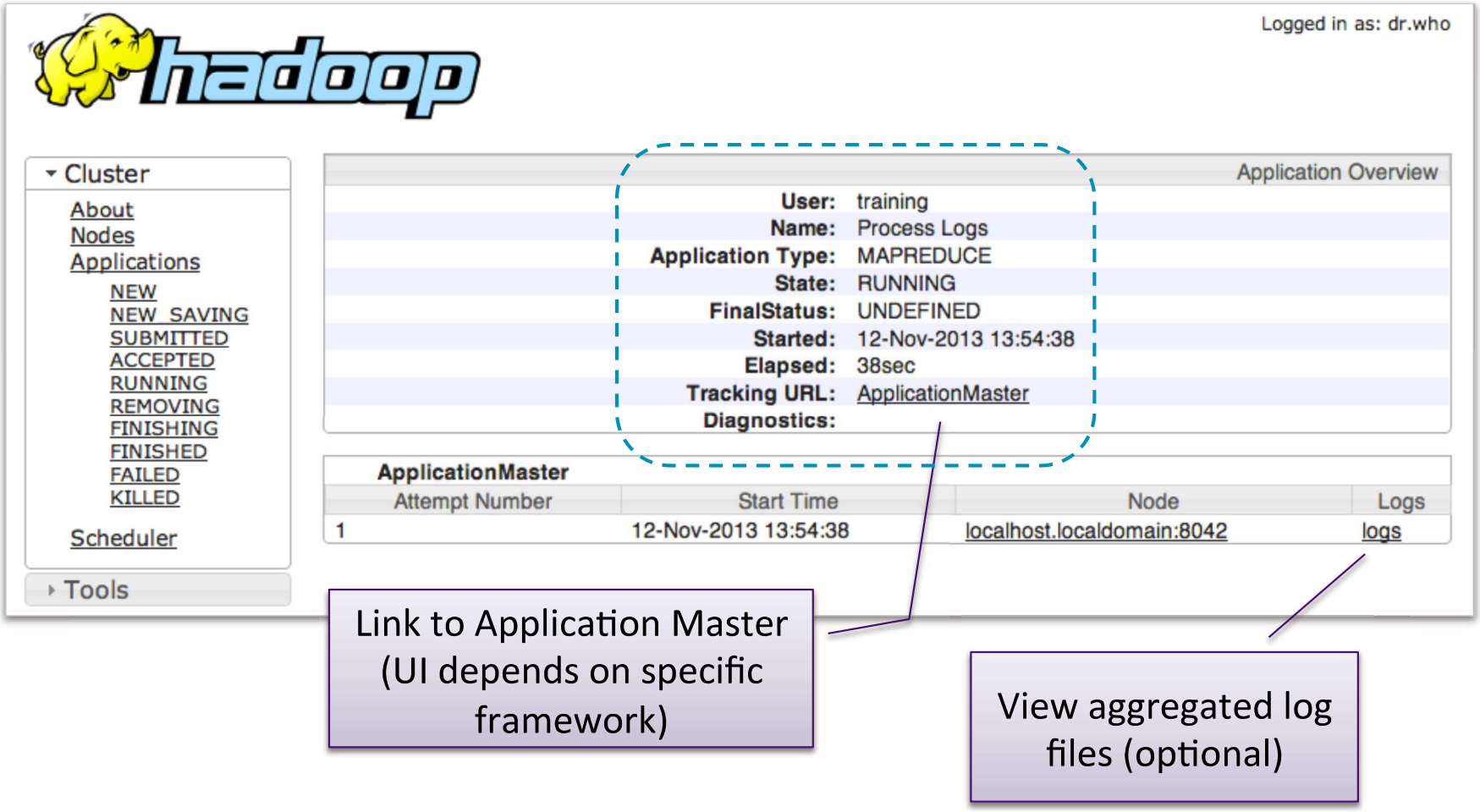
Resource Manager UI: Nodes



Resource Manager UI: Applications

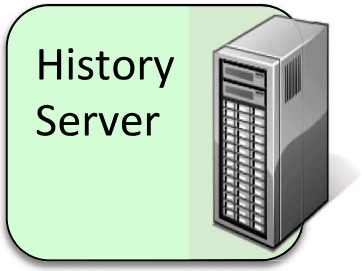


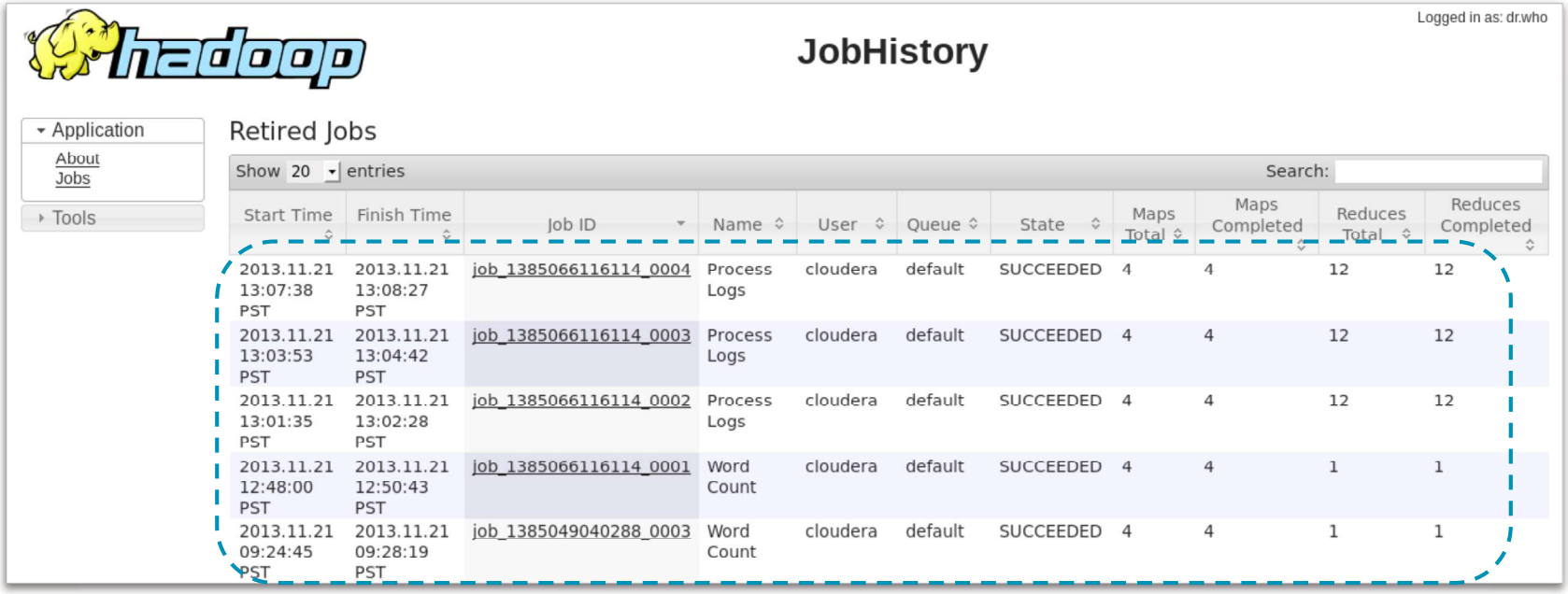
Resource Manager UI: Application Detail



Job History Server

- **YARN does not keep track of job history**
- **Spark and MapReduce each provide a Job History Server**
  - Archives job's metrics and metadata
  - Can be accessed through Job History UI or Hue





YARN Command Line

- **Command to configure and view information about the YARN cluster**
  - **`yarn <command>`**
- **Most YARN command line tools are for administrators rather than developers**
- **Some helpful commands for developers**
  - **`yarn application`**
    - Use **`-list`** to see running applications
    - Use **`-kill`** to kill a running application
  - **`yarn logs -applicationId <app-id>`**
    - View the logs of the specified application

$ yarn app -list

Bibliography

The following offer more information on topics discussed in this chapter

- ***Hadoop Application Architectures: Designing Real-World Big Data Applications*** **(published by O'Reilly)**
  - `http://tiny.cloudera.com/archbook`
- **HDFS User Guide**
  - `http://tiny.cloudera.com/hdfsuser`
- **YARN documentation**
  - `http://tiny.cloudera.com/yarndocs`
- **Cloudera Engineering Blog YARN articles**
  - `http://tiny.cloudera.com/yarnblog`