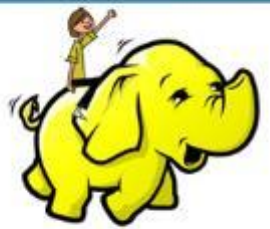
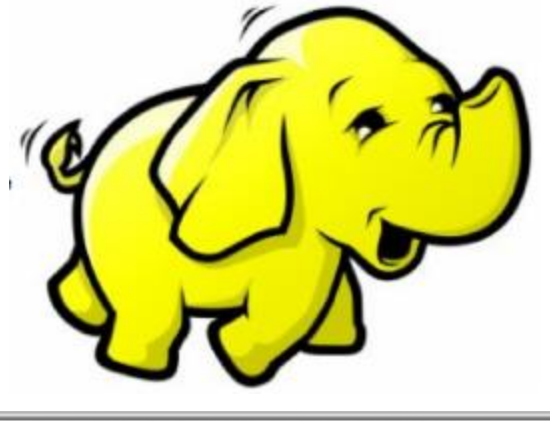
edureka!

Hadoop Administration



Hadoop Administration



Module 6: Advanced Topics: QJM, HDFS Federation and Security

Course Topics

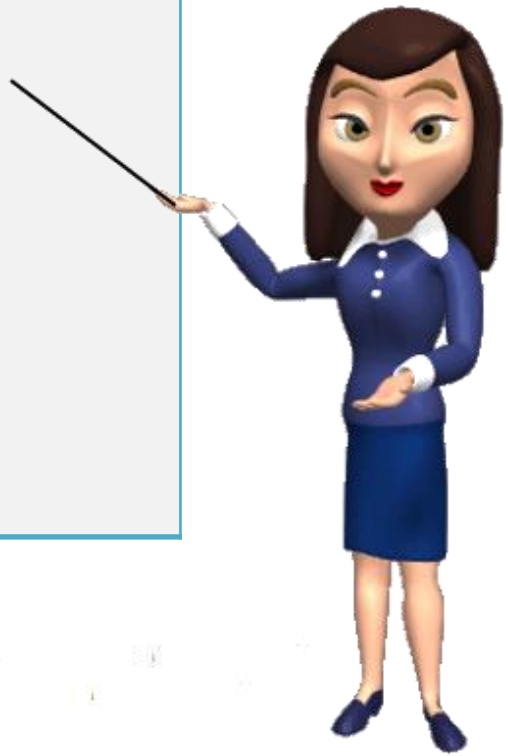
edureka!

- **Module 1**
  - Understanding Big Data
  - Hadoop Components
- **Module 2**
  - Different Hadoop Server Roles
  - Hadoop Cluster Configuration
- **Module 3**
  - Hadoop Cluster Planning
  - Job Scheduling
- **Module 4**
  - Securing your Hadoop Cluster
  - Backup and Recovery
- **Module 5**
  - Hadoop 2.0 New Features
  - HDFS High Availability
- **Module 6**
  - **Quorum Journal Manager (QJM)**
  - **Hadoop 2.0 - YARN**
- **Module 7**
  - Oozie Workflow Scheduler
  - Hive and Hbase Administration
- **Module 8**
  - Hadoop Cluster Case Study
  - Hadoop Implementation

Topics of the day

edureka!

- Hadoop 2.0 New Features
  - HDFS High Availability - QJM
  - YARN and MRv2
- YARN and Hadoop ecosystem
- YARN Components
- Job Tracker and Job Submission
- MR Application Execution in YARN



http://www.pvsd.k12.ca.us/site/default.aspx?PageID=129

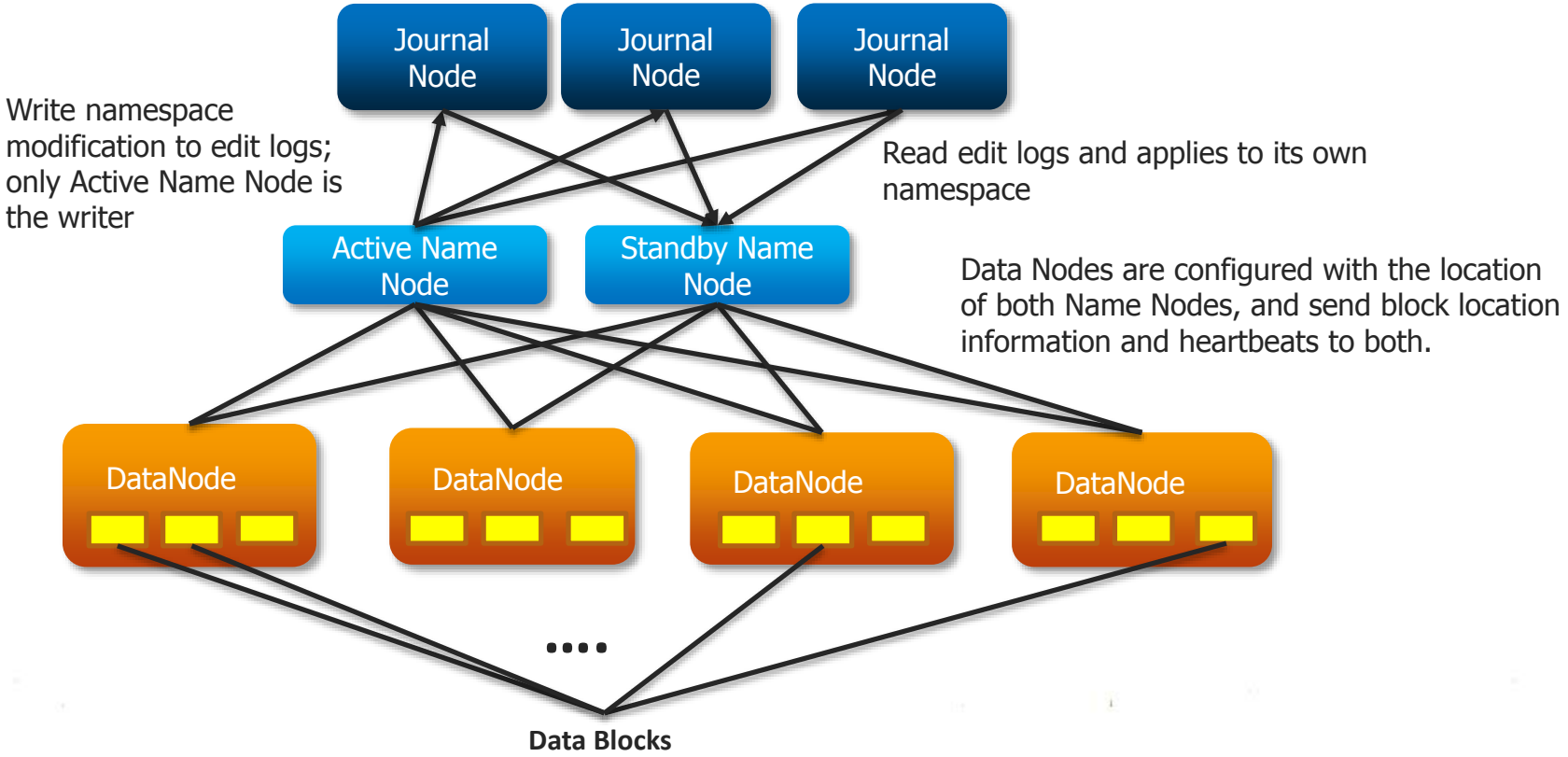
Let’s Revise

edureka!

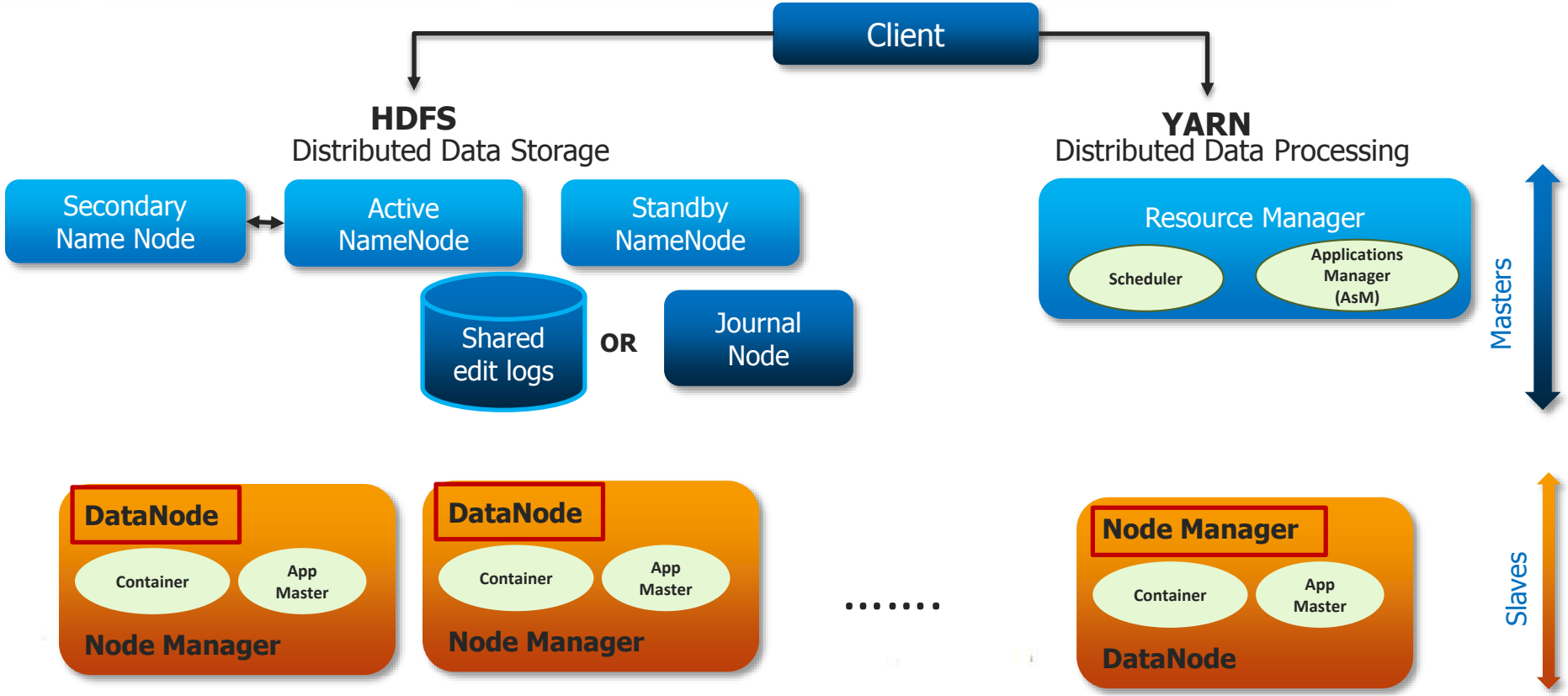
- ✓ Hadoop 1.0 Vs. Hadoop 2.0
- ✓ Hadoop 2.0 Cluster daemons

NN High Availability – Quorum Journal Manager

edureka!



Hadoop 2.0 Server Roles



Job Tracker – Overburdened

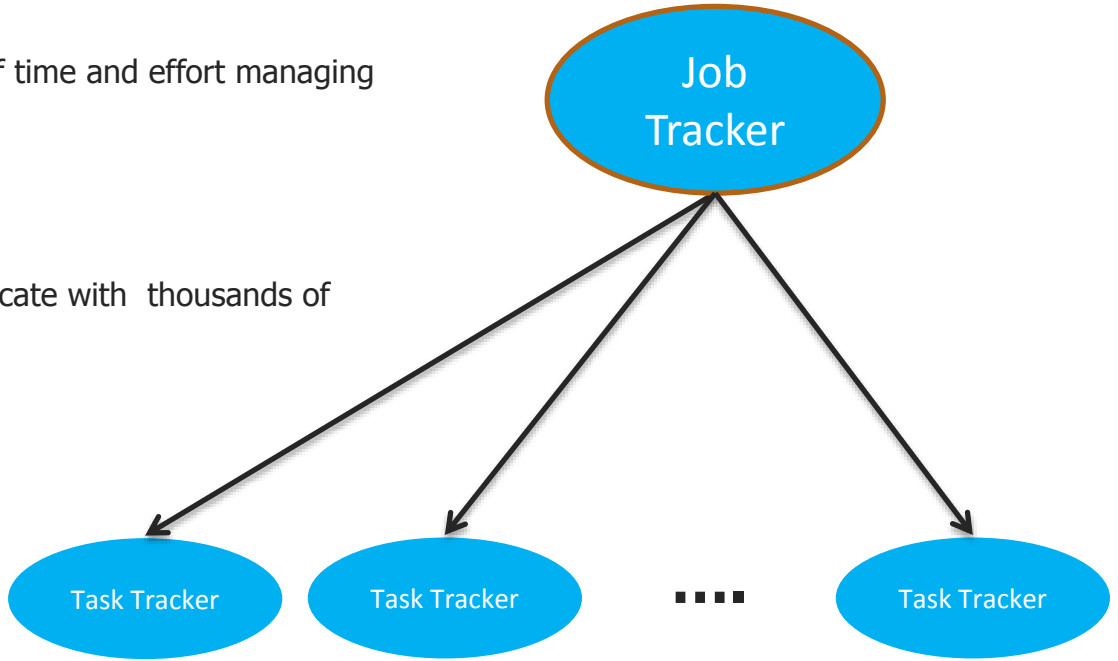
edureka!

CPU

- ✓ Spends a very significant portion of time and effort managing the life cycle of applications

Network

- ✓ Single Listener Thread to communicate with thousands of Map and Reduce Jobs



MRv1 – Unpredictability in Large Clusters

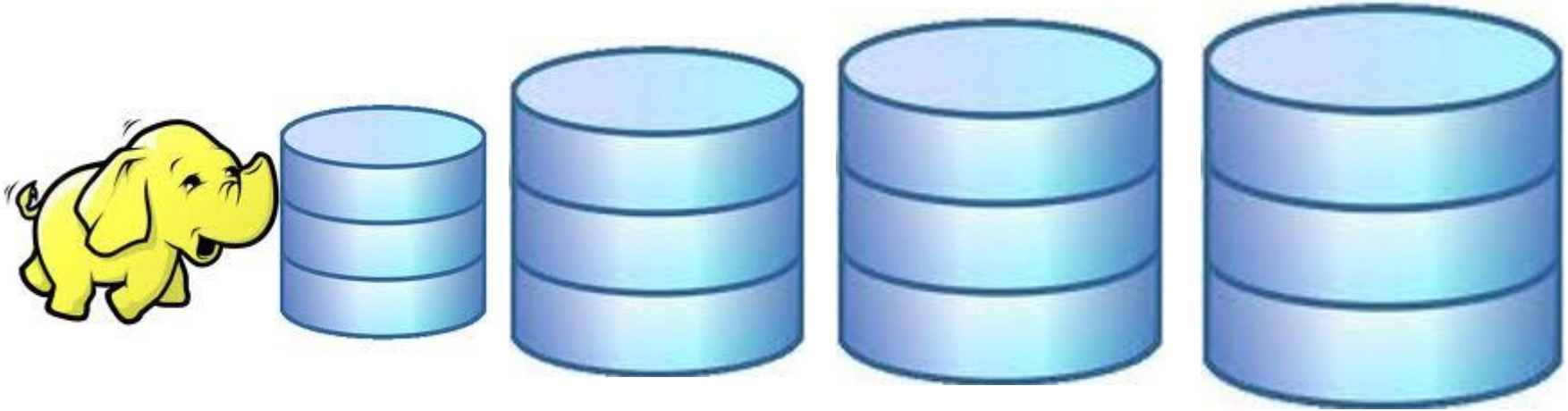
edureka!

As the cluster size grow and reaches to 4000 Nodes

- ✓ **Cascading Failures**
  - ✓ The DataNode failures results in a serious deterioration of the overall cluster performance because of attempts to replicate data and overload live nodes, through network flooding.
- ✓ **Multi-tenancy**
  - ✓ As clusters increase in size, you may want to employ these clusters for a variety of models. MRv1 dedicates its nodes to Hadoop and cannot be re-purposed for other applications and workloads in an Organization. With the growing popularity and adoption of cloud computing among enterprises, this becomes more important.

Unutilized Data in HDFS

edureka!

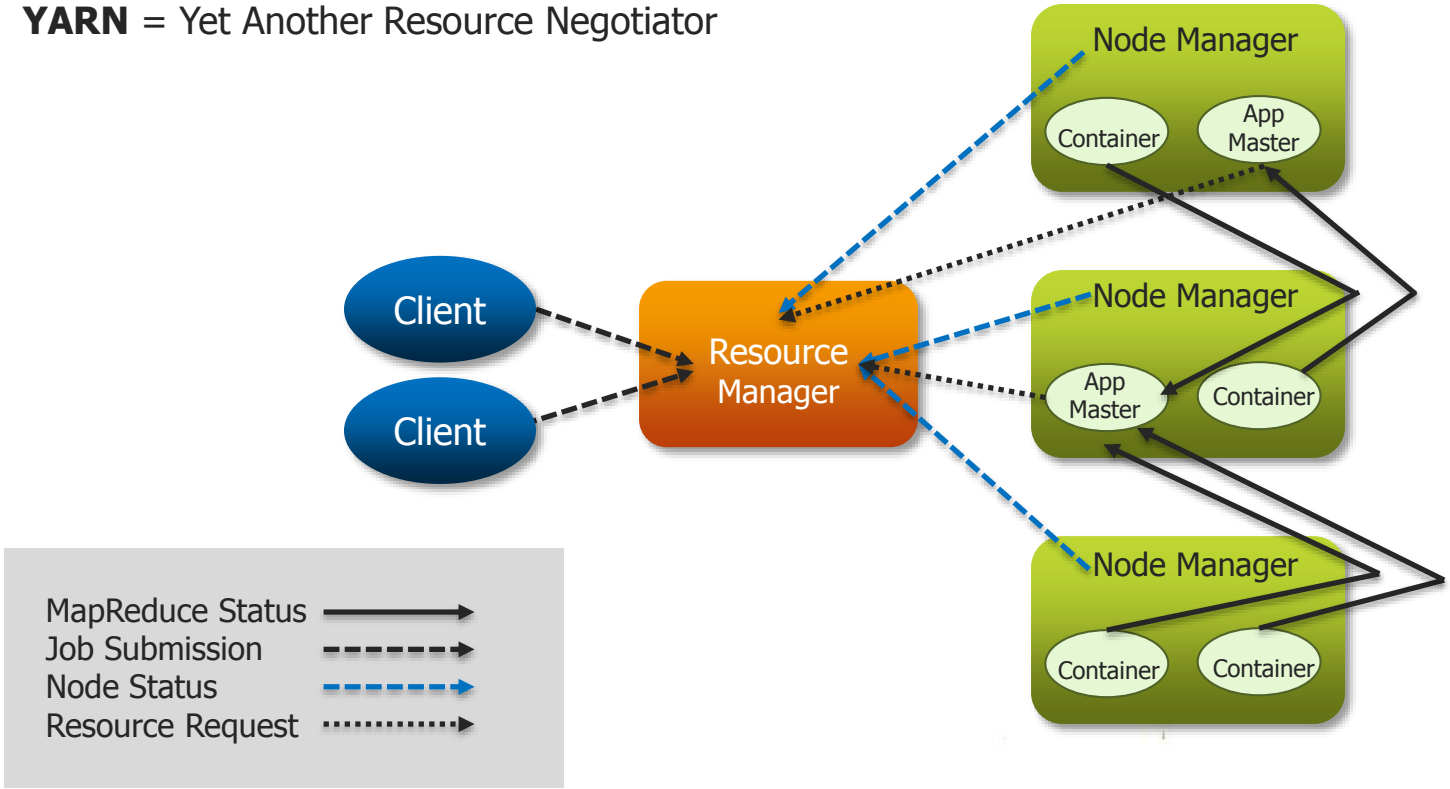


- ✓ Terabytes and Petabytes of data in HDFS can be used only for MapReduce processing

Hadoop 2.0 : YARN

edureka!

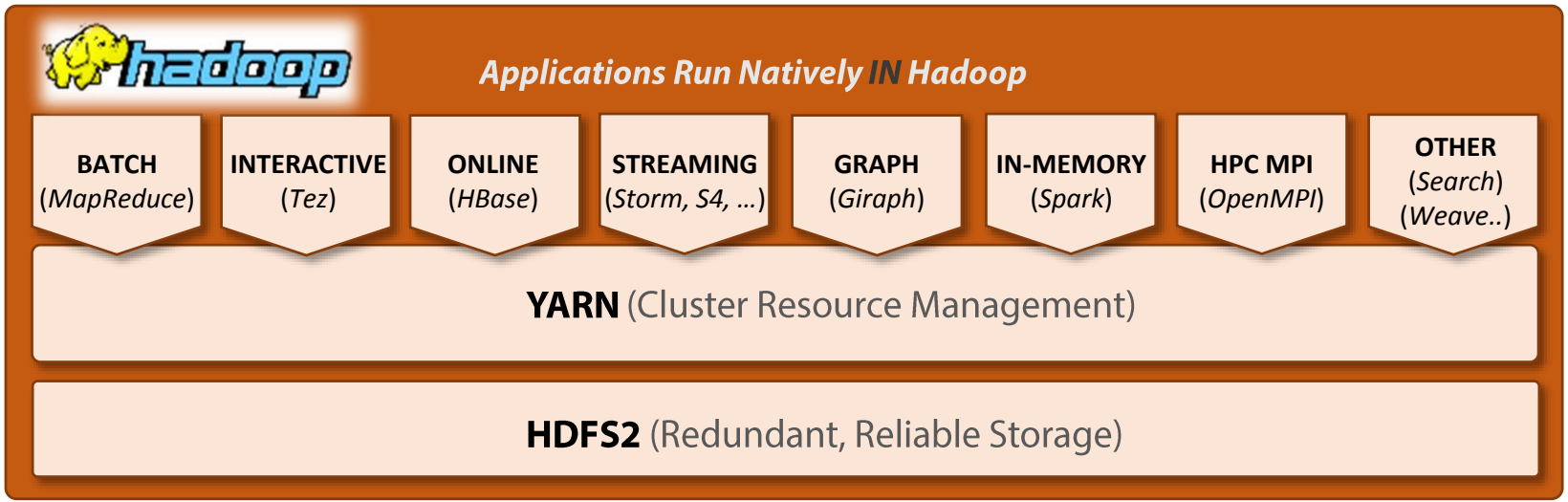
YARN = Yet Another Resource Negotiator



http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html

YARN – Moving beyond MapReduce

edureka!



http://hadoop.apache.org/docs/stable2/hadoop-yarn/hadoop-yarn-site/YARN.html

Multi-tenancy - Capacity Scheduler

edureka!

- ✓ Organizes jobs into queues
- ✓ Queue shares as %'s of cluster
- ✓ FIFO scheduling within each queue
- ✓ Data locality-aware Scheduling

Multi-tenancy - Capacity Scheduler

edureka!

- ✓ **Hierarchical Queues**
  To manage the resource within an organization.
- ✓ **Capacity Guarantees**
  A fraction to the total available capacity allocated to each Queue.
- ✓ **Security**
  To safeguard applications from other users.
- ✓ **Elasticity**
  Resources are available in a predictable and elastic manner to queues.
- ✓ **Multi-tenancy**
  Set of limit to prevent over-utilization of resources by a single application.
- ✓ **Operability**
  Runtime configuration of Queues.
- ✓ **Resource-based scheduling**
  If needed, Applications can request more resources than the default.

Hadoop 2.0: Configure the Capacity Scheduler

edureka!

✓ Edit **yarn-site.xml** to enable the Capacity Scheduler

Property	Value
yarn.resourcemanager.scheduler.class	org.apache.hadoop.yarn.server.resourcemanager.scheduler.capacity.CapacityScheduler

✓ Configure **'queues'** in capacity-scheduler.xml

Annie's Question

edureka!

The Resource Manager has two main components
a) NameNode and SNN
b) Scheduler and Applications Manager
c) Manager and Application manager4

Annie's Answer

edureka!

Answer: Scheduler and Applications Manager

Annie's Question

edureka!

YARN enables a user to interact with all data in multiple ways simultaneously, making Hadoop a true multi-use data platform and allowing it to take its place in a modern data architecture.

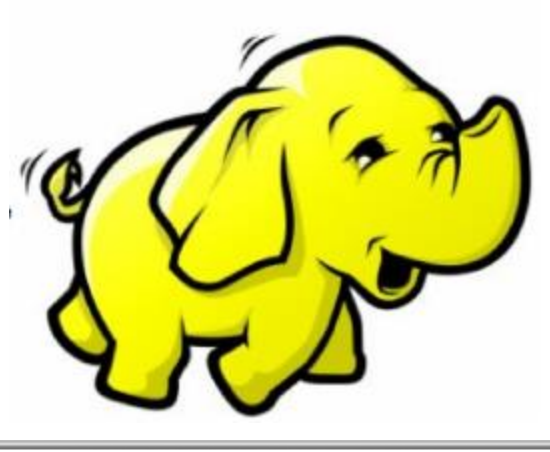
a) True
b) False

Annie's Answer

edureka!

Answer: True

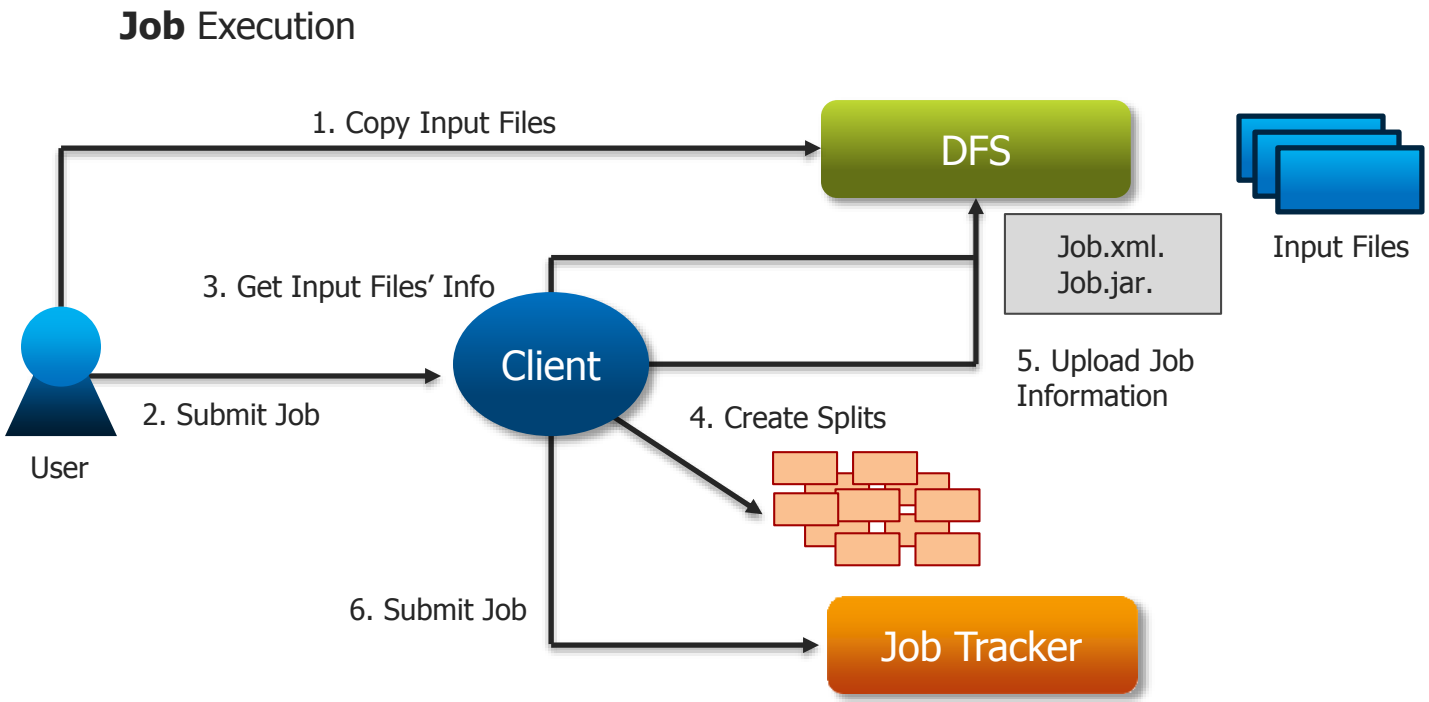
Hadoop Administration



Executing MapReduce Application on YARN

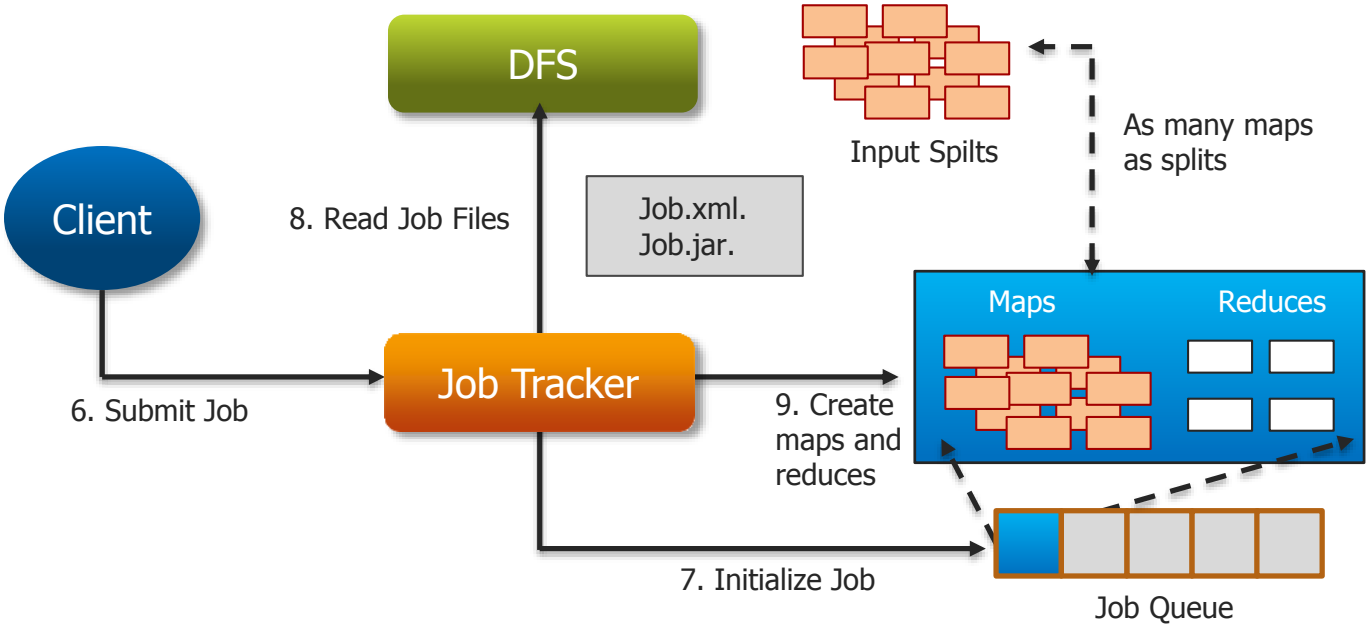
Job Tracker

edureka!



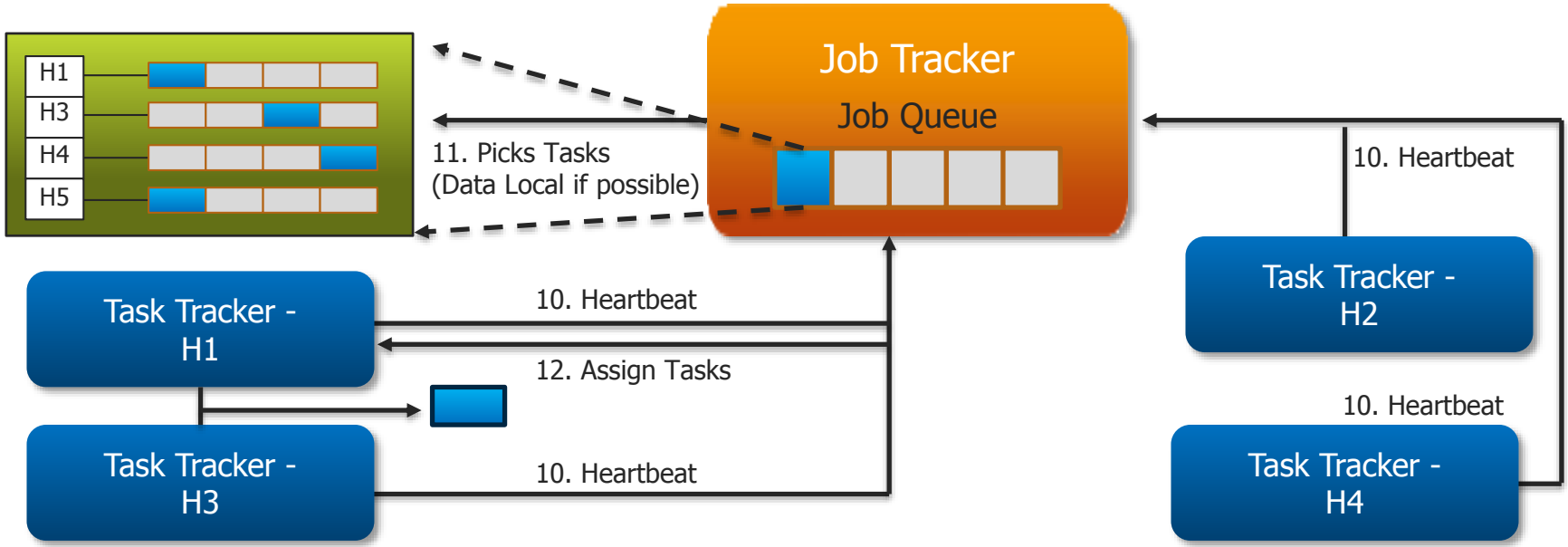
Job Tracker (Contd.)

edureka!



Job Tracker (Contd.)

edureka!

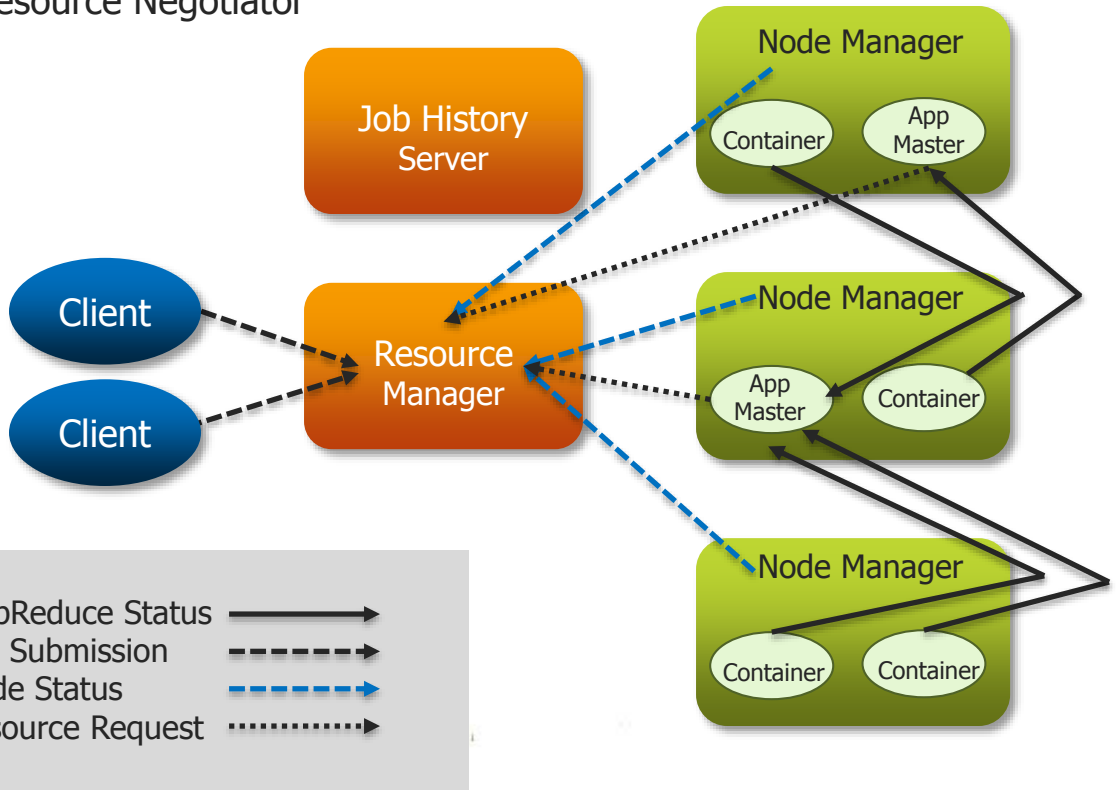


YARN – Beyond MapReduce

edureka!

YARN = Yet Another Resource Negotiator

- ✓ **Resource Manager**
  - ✓ Cluster Level resource manager
  - ✓ Long Life, High Quality Hardware
- ✓ **Node Manager**
  - ✓ One per Data Node
  - ✓ Monitors resources on Data Node
- ✓ **Application Master**
  - ✓ One per application
  - ✓ Short life
  - ✓ Manages task /scheduling



YARN MR Application Execution Flow

edureka!

Application Vs. Job Execution

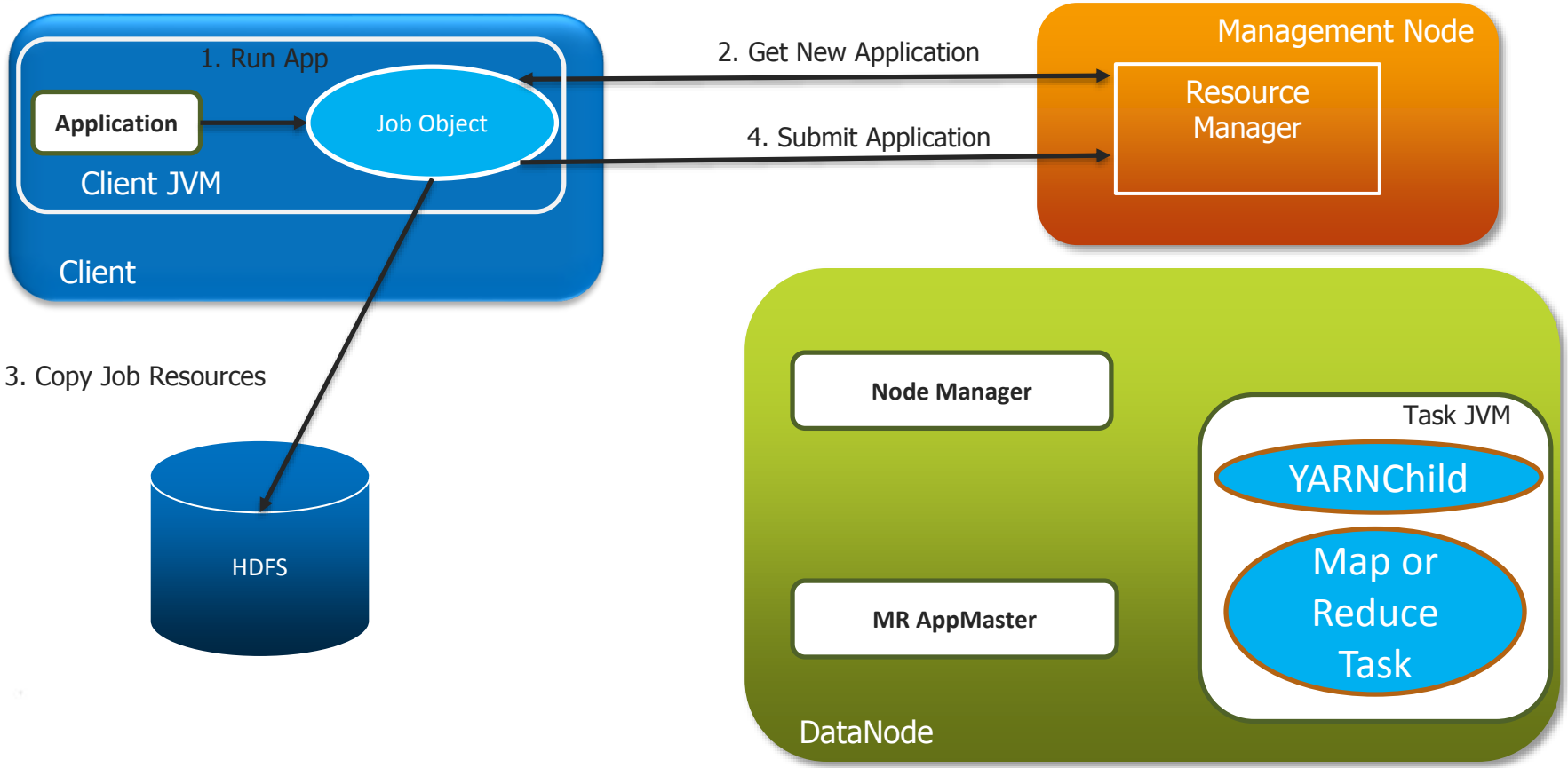
- ✓ MapReduce **Application** Execution
  - ✓ Submission
  - ✓ Initialization
  - ✓ Tasks Assignment
  - ✓ Tasks' Memory
  - ✓ Status Updates
  - ✓ Failure Recovery

YARN MR Application Execution Flow

edureka!

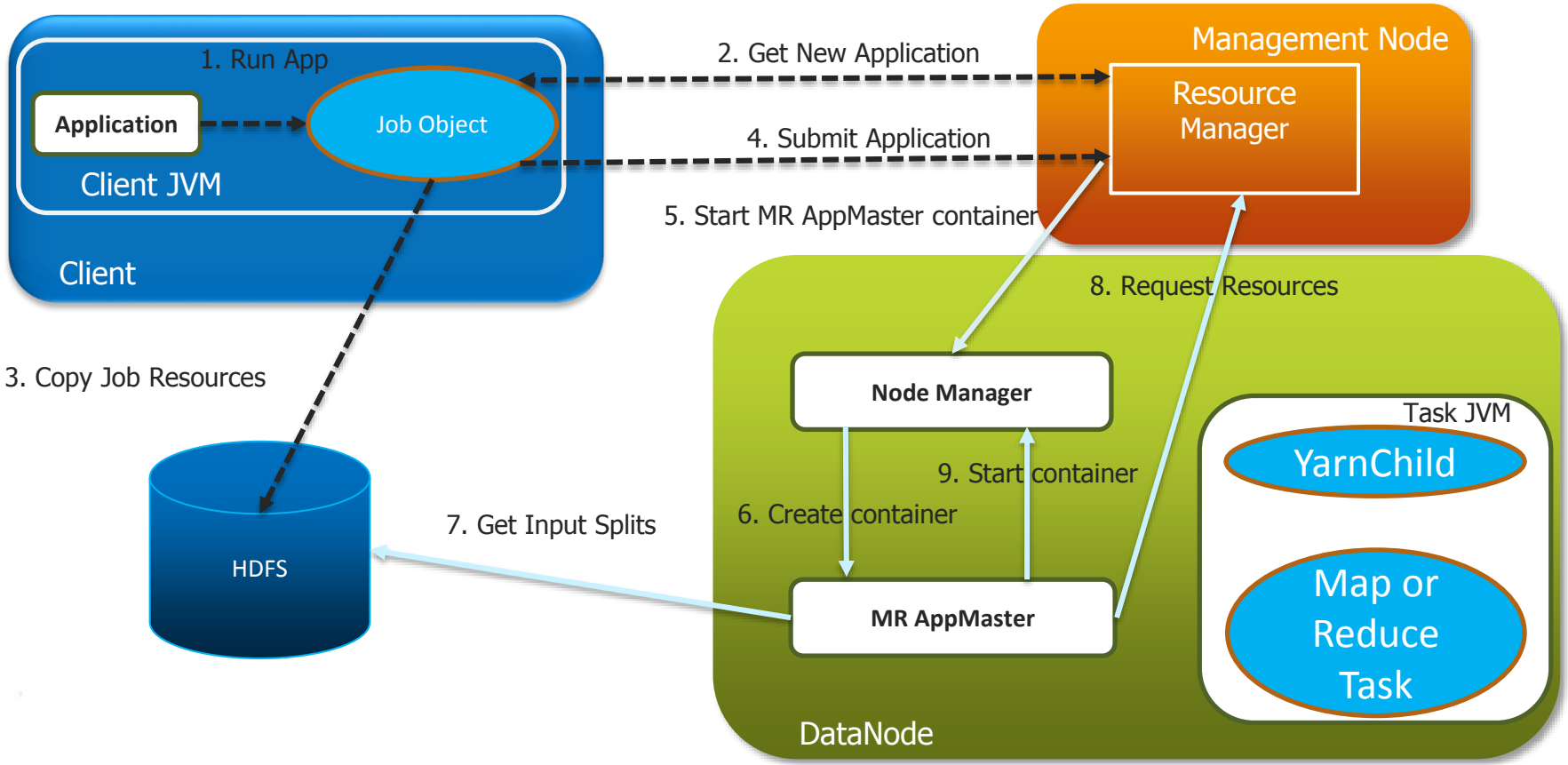
- ✓ **Client**
  - ✓ Submit a MapReduce Application
- ✓ **Resource Manager**
  - ✓ Manage the resource utilization across
  - ✓ Hadoop Cluster
- ✓ **Node Manager**
  - ✓ Runs on each Data Node
  - ✓ Creates execution container
  - ✓ Monitors Container's usage
- ✓ **MapReduce Application Master**
  - ✓ Coordinates and Manages MapReduce tasks
  - ✓ Negotiates with Resource Manager to schedule asks
  - ✓ The tasks are started by Node Manager(s)
- ✓ **HDFS**
  - ✓ shares resources and task's artefacts among YARN components

YARN MR Application Execution Flow

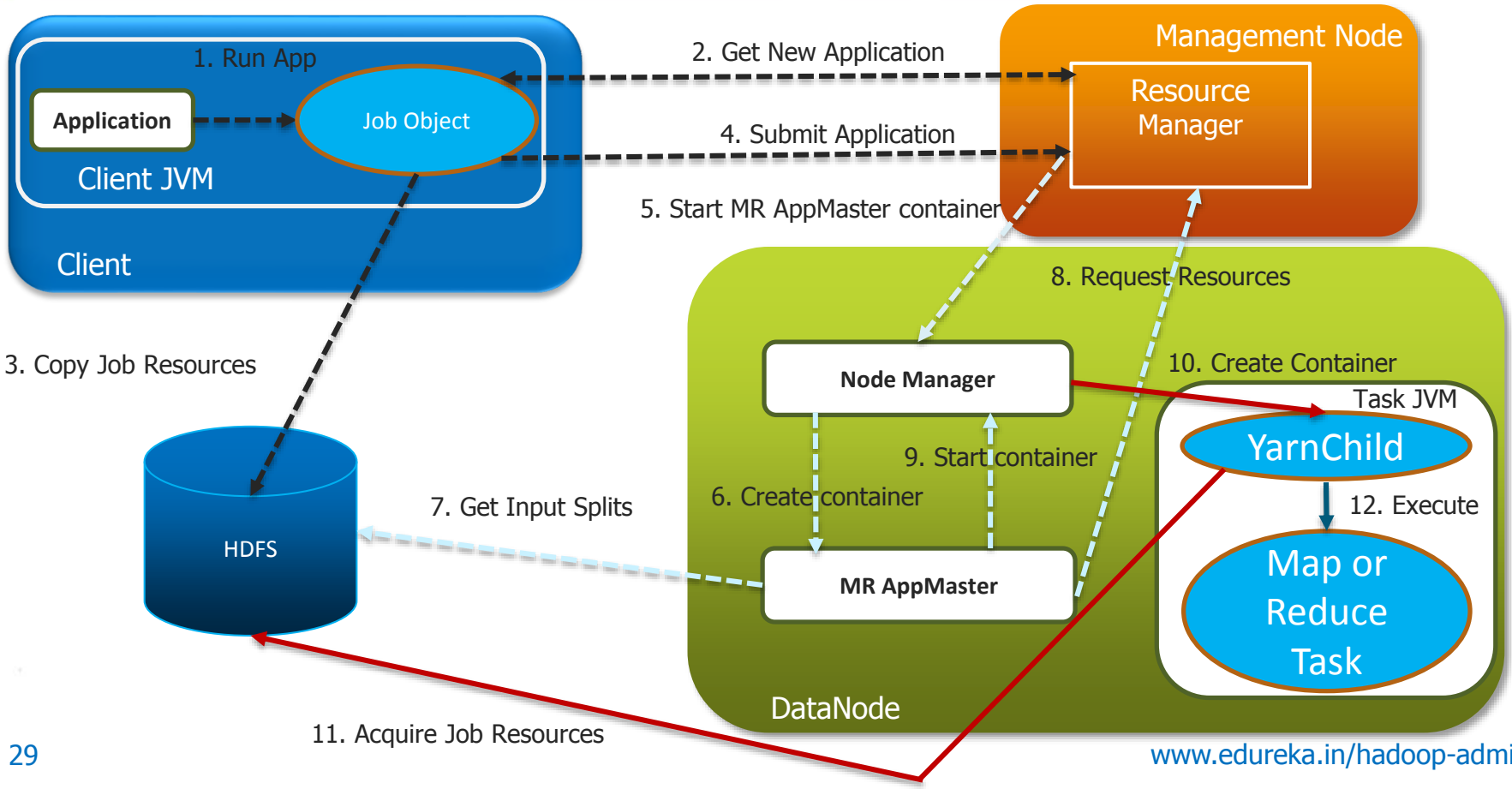


YARN MR Application Execution Flow

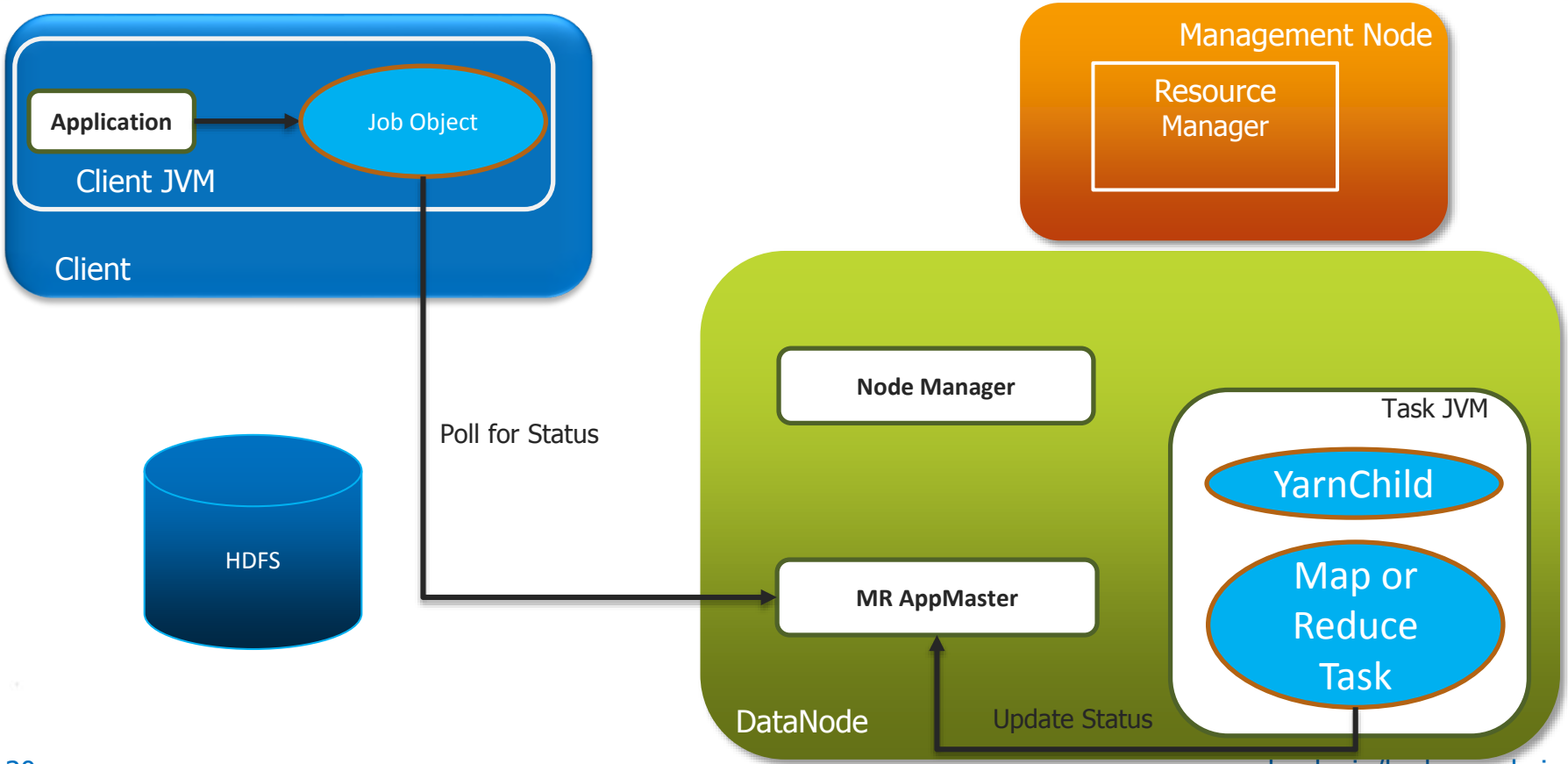
edureka!



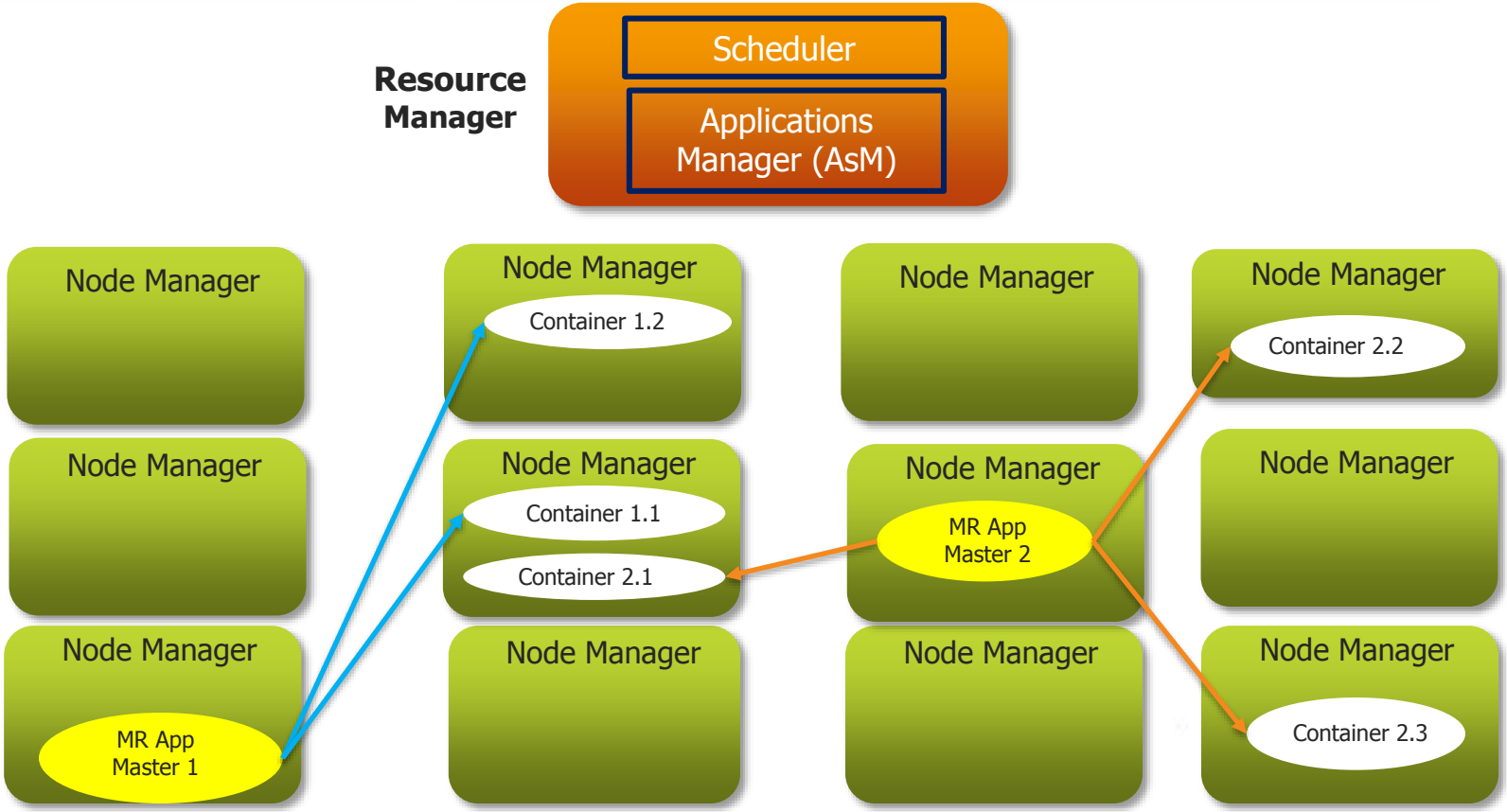
YARN MR Application Execution Flow



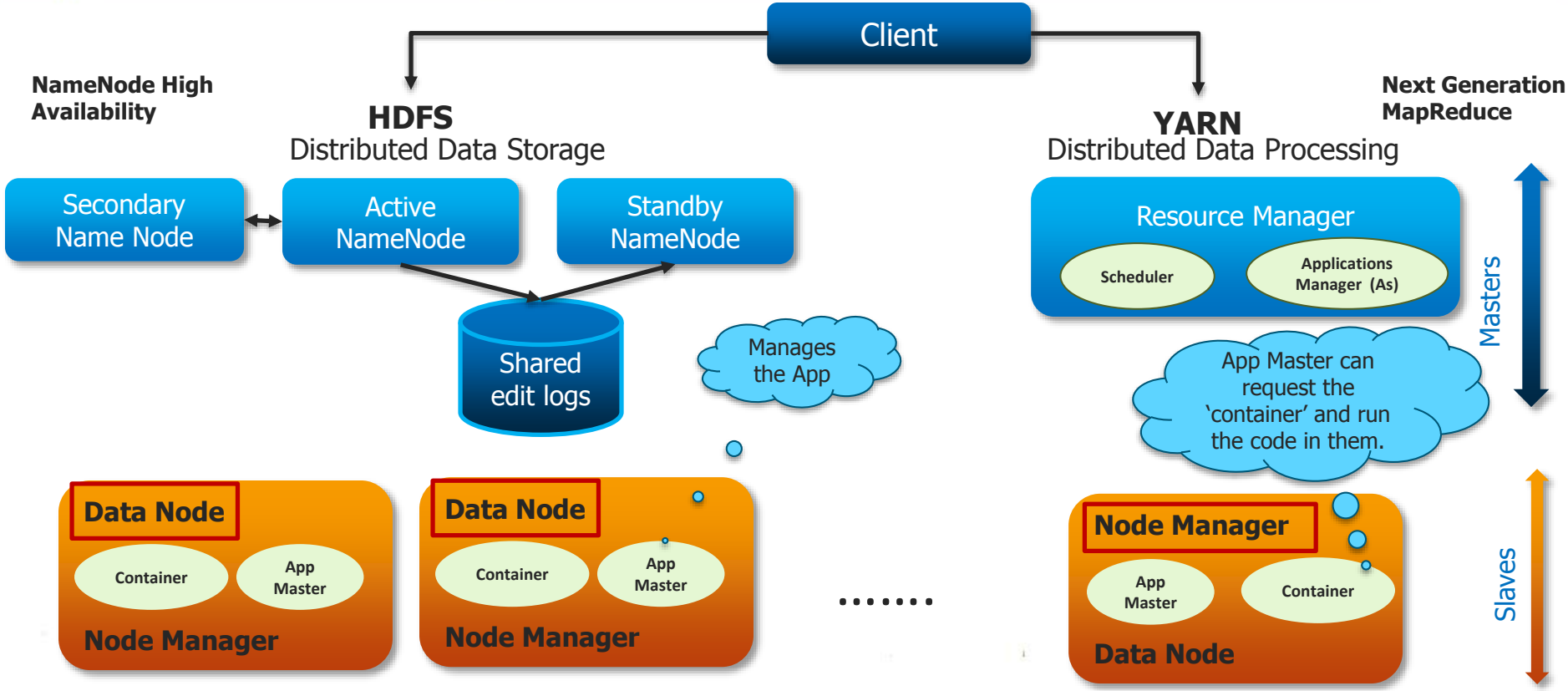
YARN MR Application Execution Flow



Hadoop 2.0 : YARN Workflow



Hadoop 2.0 – In Summary



Annie's Question

edureka!

YARN was developed to overcome the following disadvantage in MRv1?

a) Single Point of Failure of NameNode
b) Only one version can be run in classic MapReduce
c) Too much burden on Job Tracker

Annie's Answer

edureka!

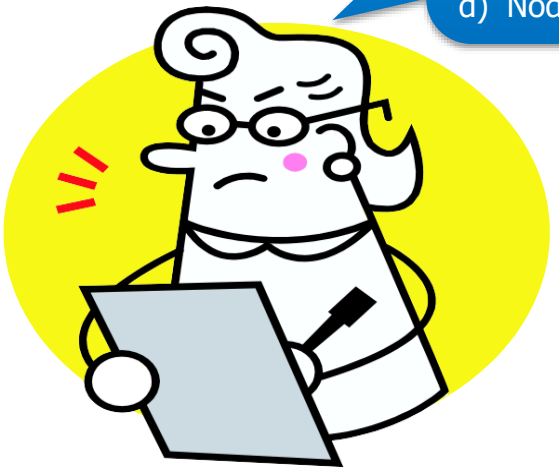
Answer: Too much burden on Job Tracker

Annie's Question

edureka!

In YARN, the functionality of Job Tracker has been replaced by which of the following YARN features:

a) Job Scheduling
b) Task Monitoring
c) Resource Management
d) Node management



Annie's Answer

edureka!

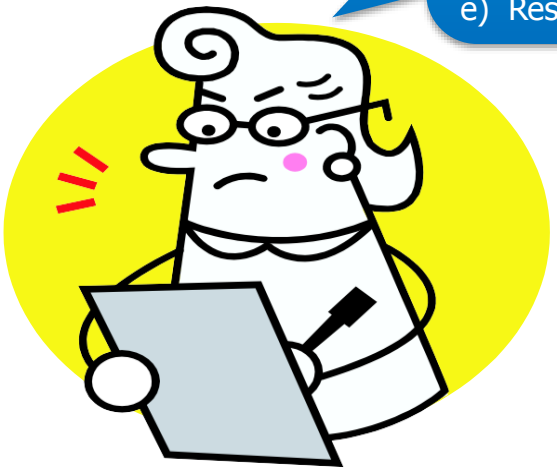
Task Monitoring and Resource Management. The fundamental idea of MRv2 is to split up the two major functionalities of the Job Tracker, i.e. resource management and job scheduling/monitoring, into separate daemons. A global **Resource Manager (RM)** for resources and per-application **Application Master (AM)** for task monitoring.

Annie's Question

edureka!

In YARN, which of the following daemons takes care of the container and the resource utilization by the applications?

a) Node Manager
b) Job Tracker
c) Task tracker
d) Application Master
e) Resource manager



Annie's Answer

edureka!

Answer: Application Master

Annie's Question

edureka!

Can we run MRv1 Jobs in a YARN enabled Hadoop Cluster?
a) Yes
b) No

Annie's Answer

edureka!

Answer: Yes. You need recompile the Jobs in MRv2 after enabling YARN to run the Job successfully on a YARN enabled Hadoop Cluster.

Annie's Question

edureka!

Which of the following YARN daemon is responsible for launching the tasks?

a) Task Tracker
b) Resource Manager
c) Application Master
d) Application Master Server

Annie's Answer

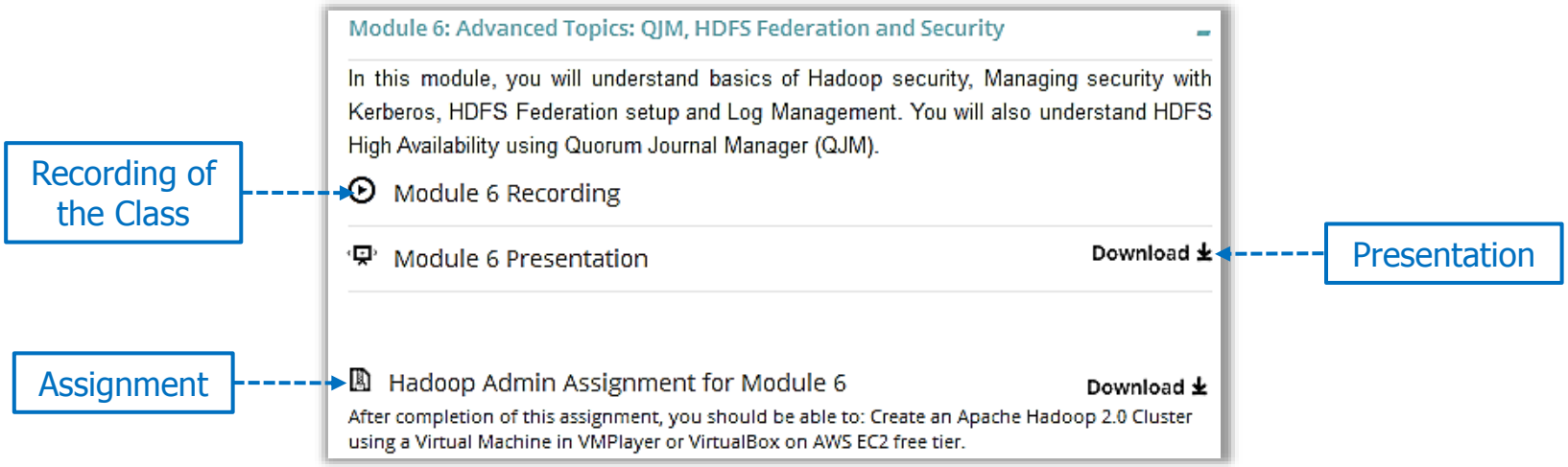
edureka!

Answer: Application Master

Tasks for you

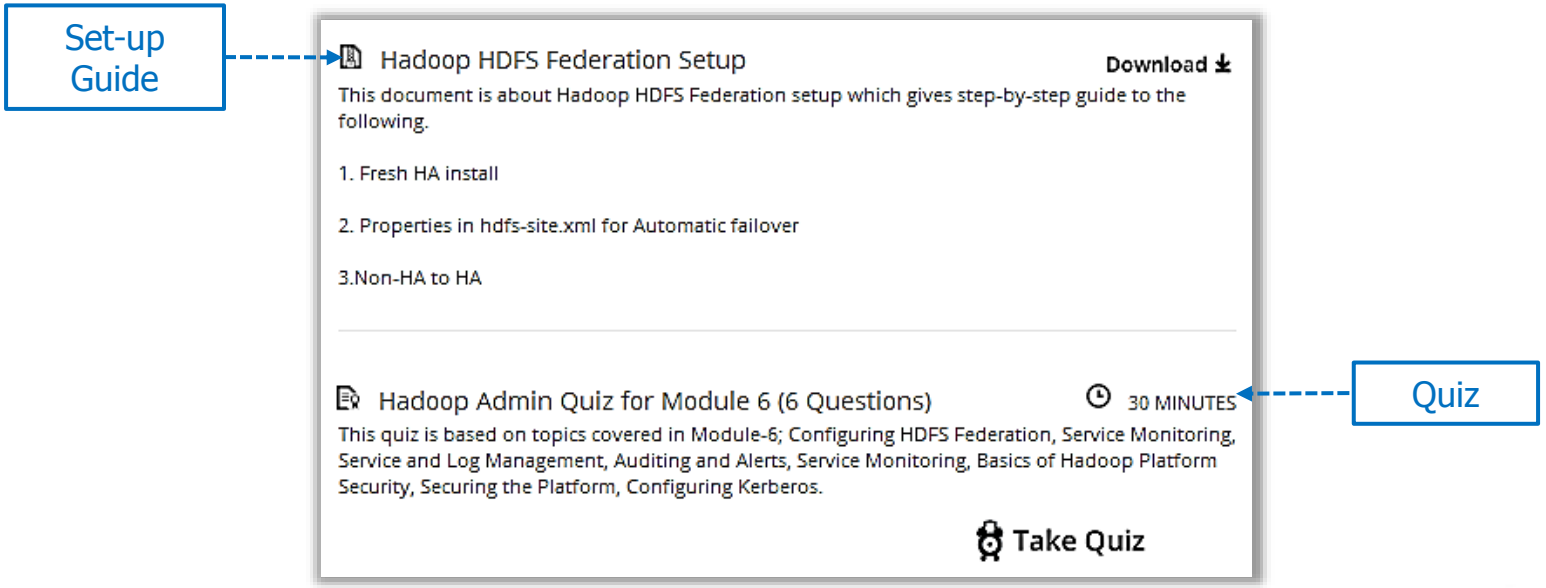
- **Attempt the following Assignments using the documents present in the LMS:**
  - Create an Apache Hadoop 2.0 Cluster using a Virtual Machine in VMPlayer or VirtualBox on AWS EC2 free tier.

What's Within the LMS?



What's Within the LMS?

edureka!



edureka!

Thank You

See You in Class Next Week