Introduction to Big Data

ECAP456

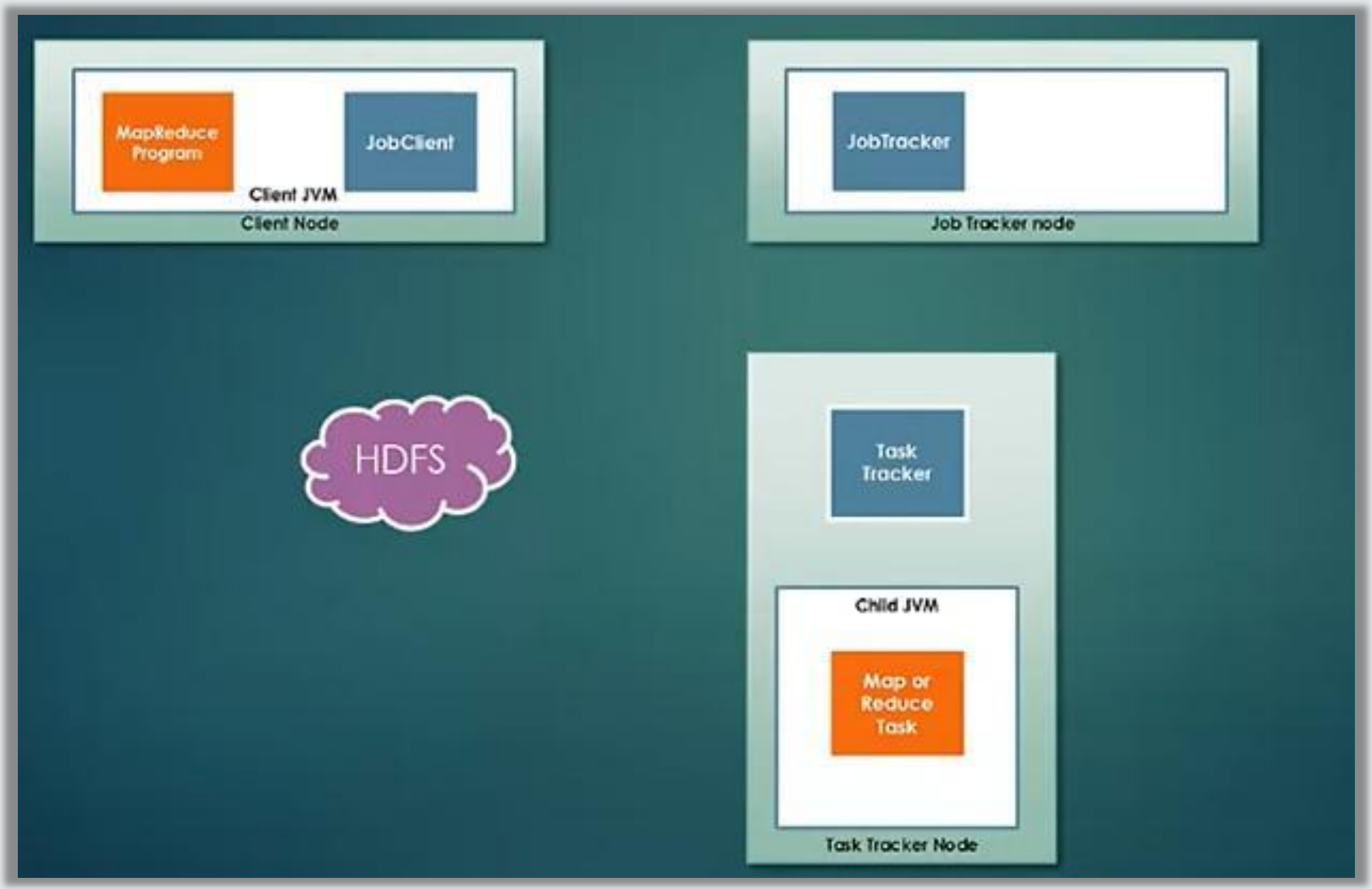
Dr. Rajni Bhalla

Associate Professor

Learning Outcomes

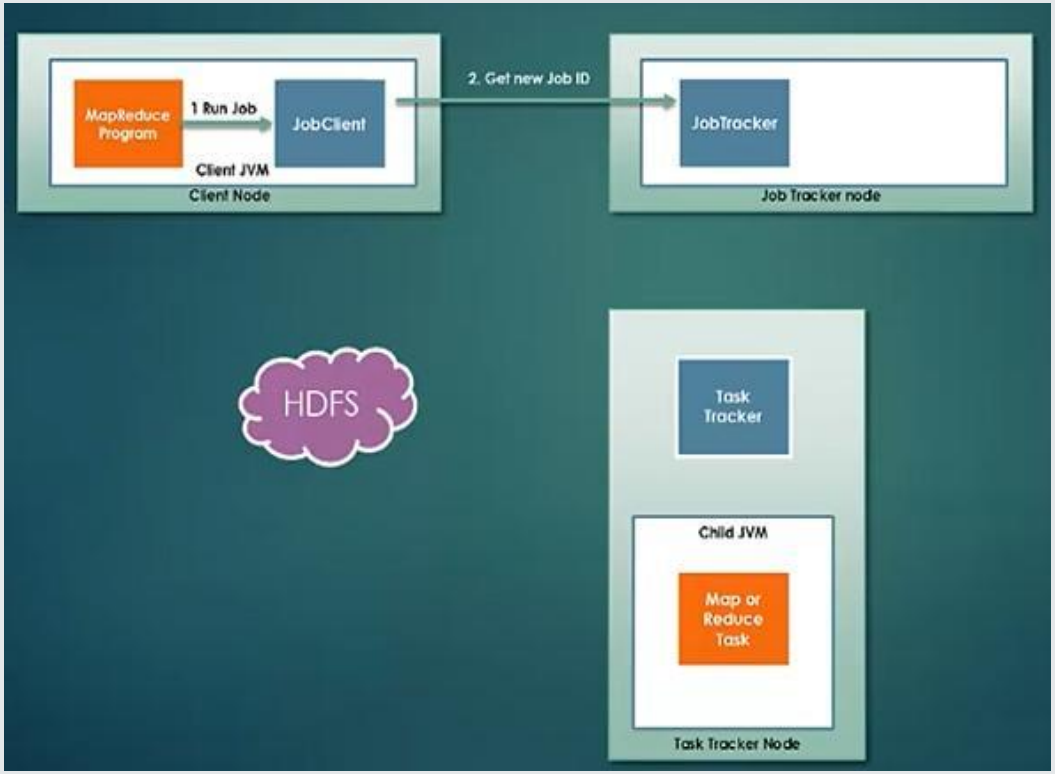
After this lecture, you will be able to

- Learn how job execution is carried out in classic MapReduce.

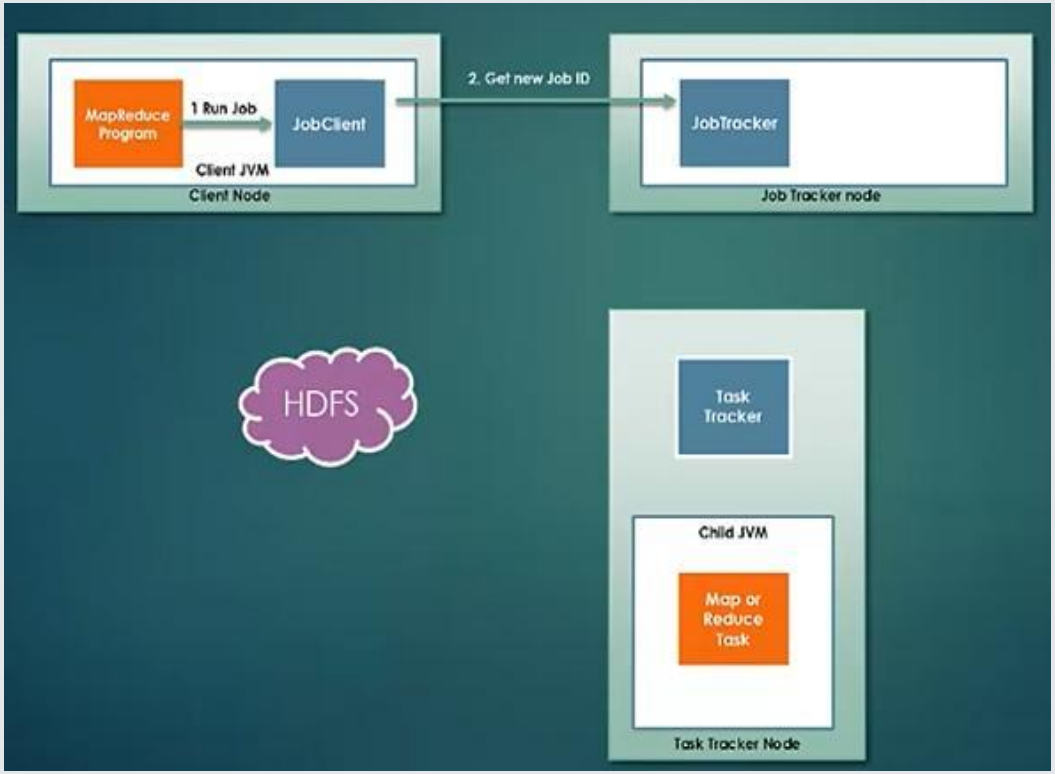

MapReduce Program
JobClient
Client JVM
Client Node
JobTracker
Job Tracker node
HDFS
Task Tracker
Child JVM
Map or Reduce Task
Task Tracker Node

Job Submission

JobClient get new job ID.



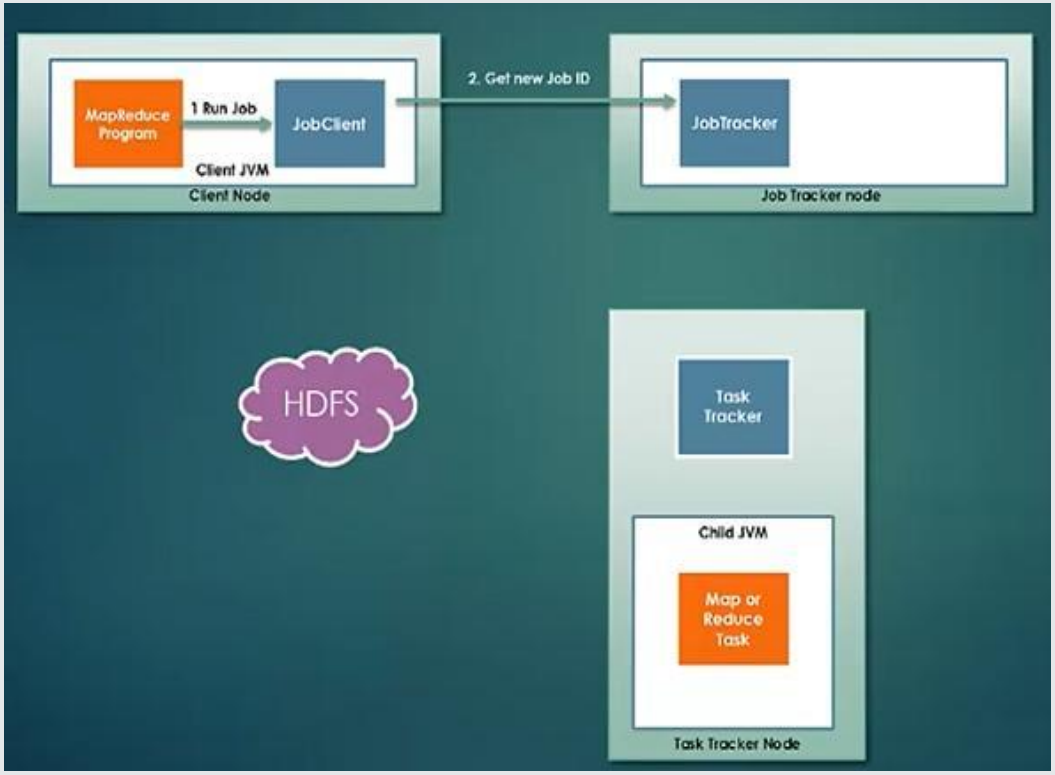
Job Submission



JobClient get new job ID.

Output has been created or not

Job Submission

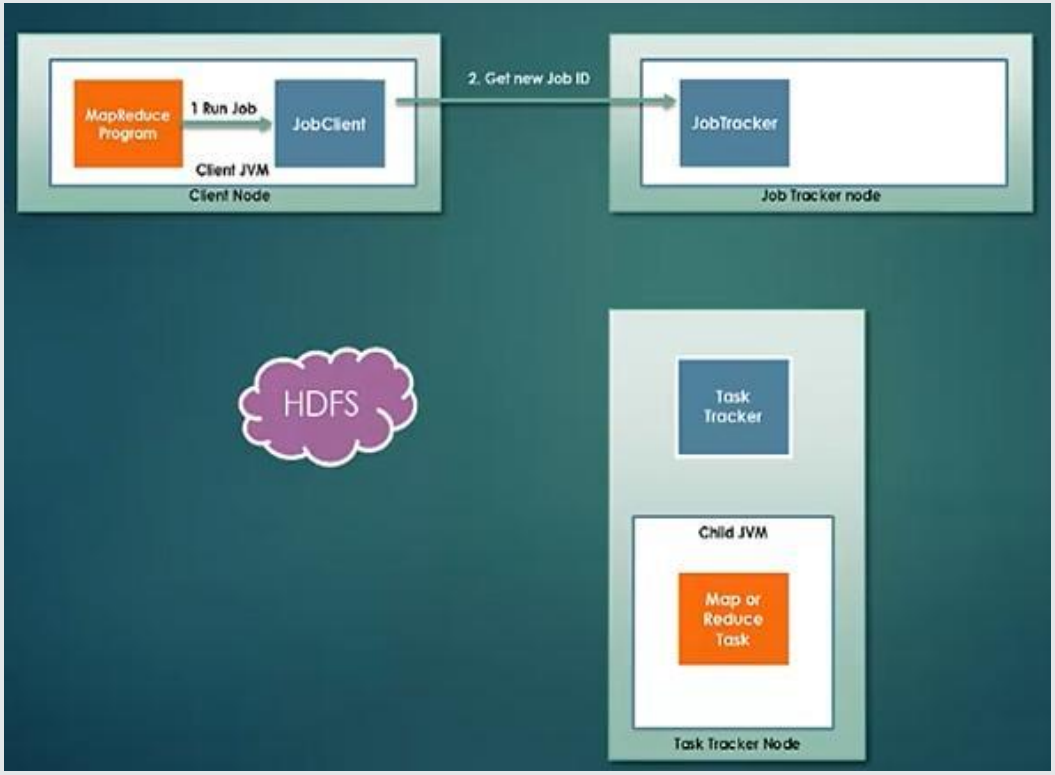


JobClient get new job ID.

Output has been created or not

Calculate the splits

Job Submission



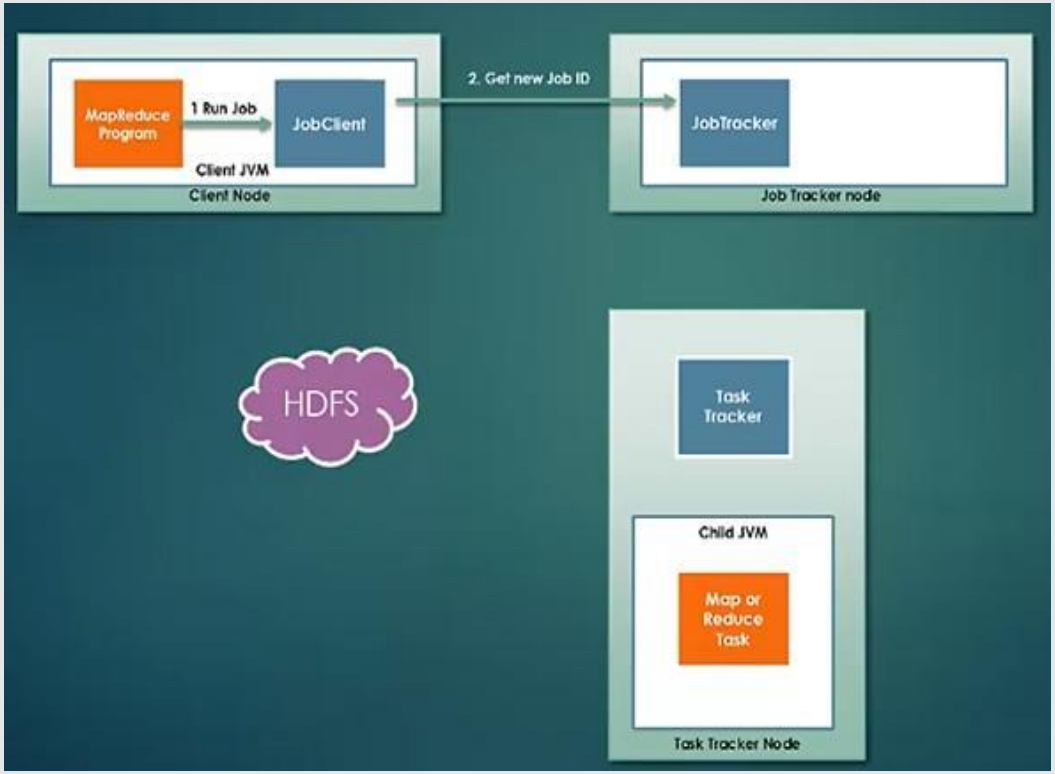
JobClient get new job ID.

Output has been created or not

Calculate the splits

Copies the resources to HDFS

Job Submission



JobClient get new job ID.

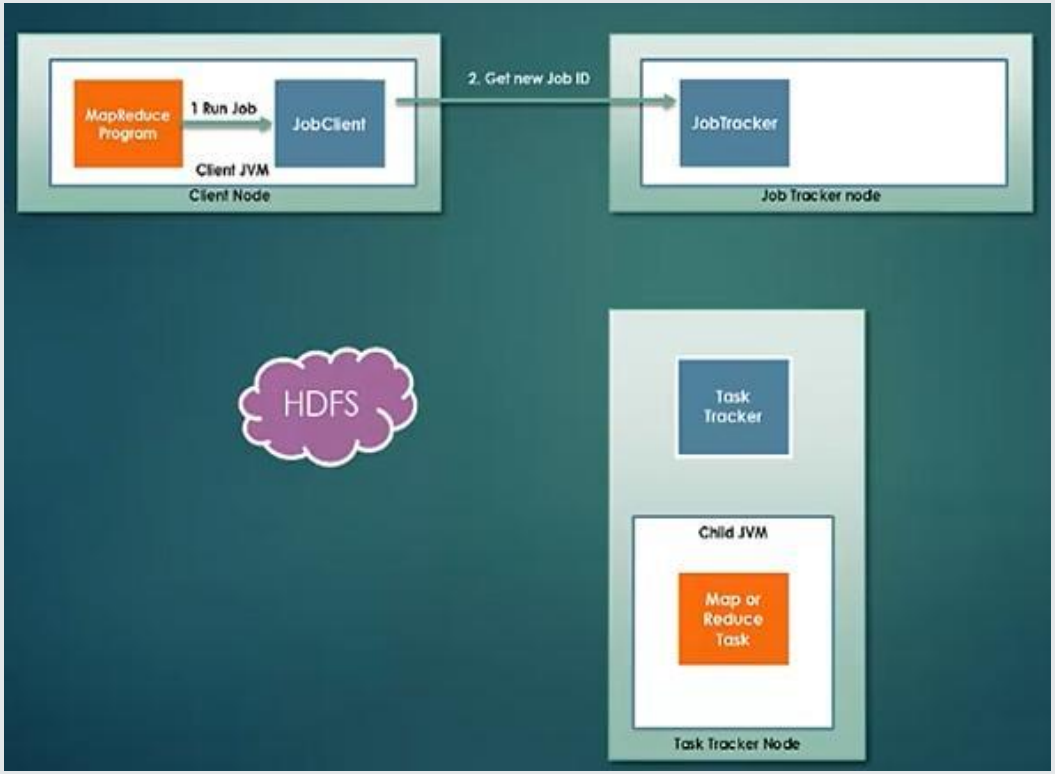
Output has been created or not

Calculate the splits

Copies the resources to HDFS

Submits the job

Job Submission



JobClient get new job ID.

Output has been created or not

Calculate the splits

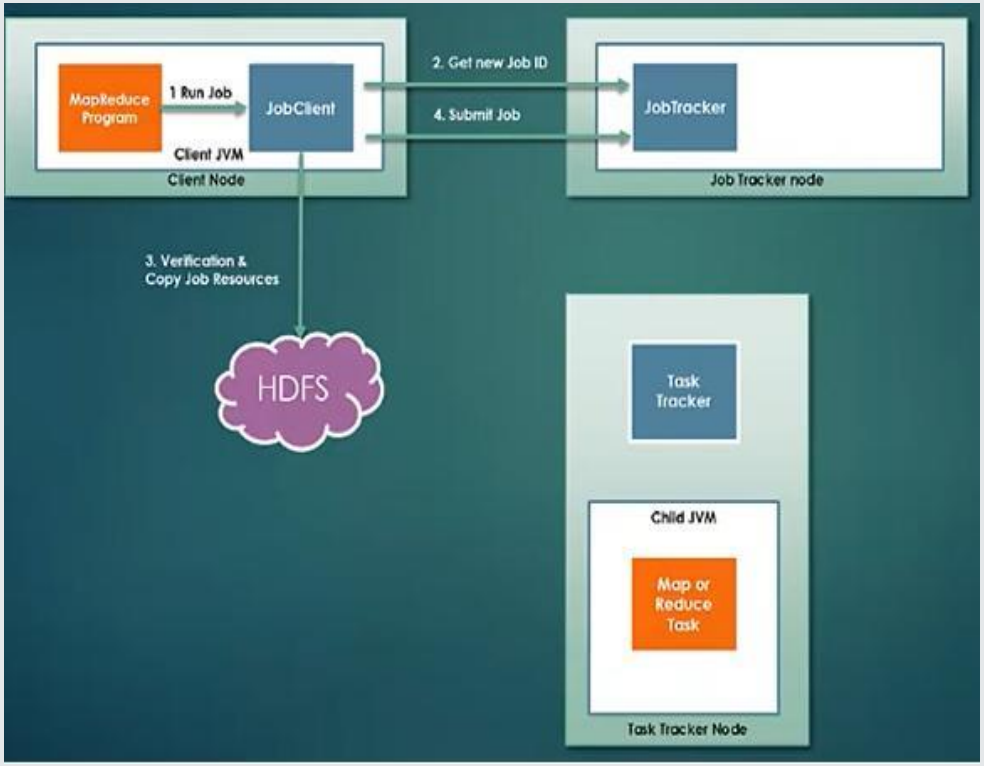
Copies the resources to HDFS

Submits the job

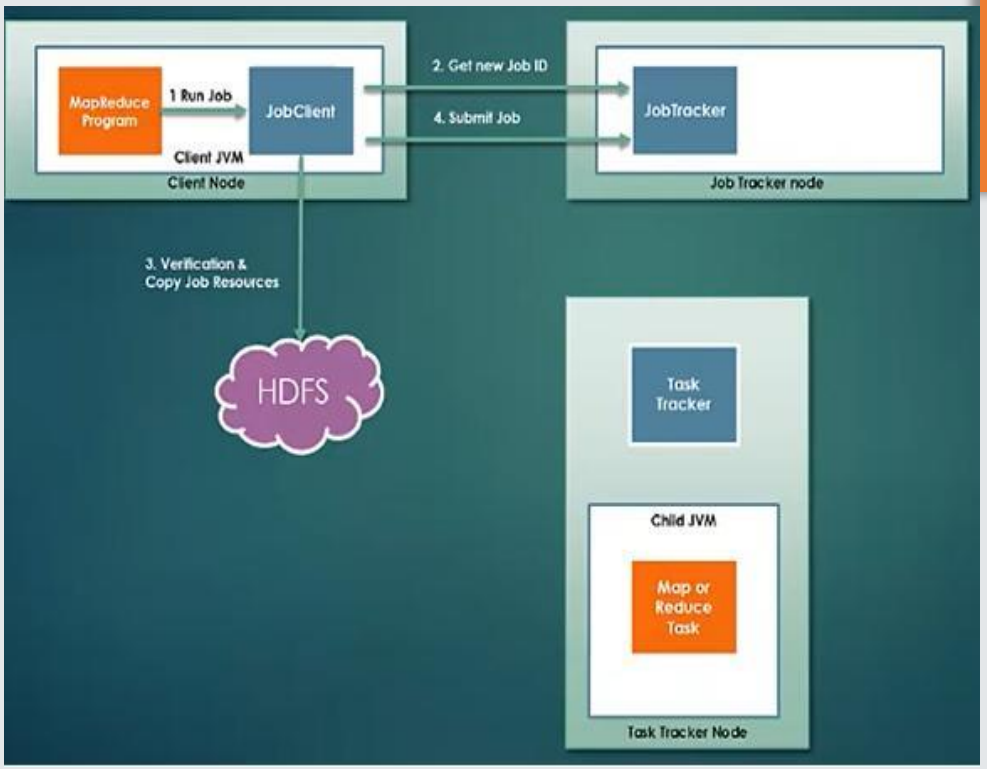
JobClient creates an instance of JobSubmitter

Job Initialization

Scheduler picks the job from the queue.



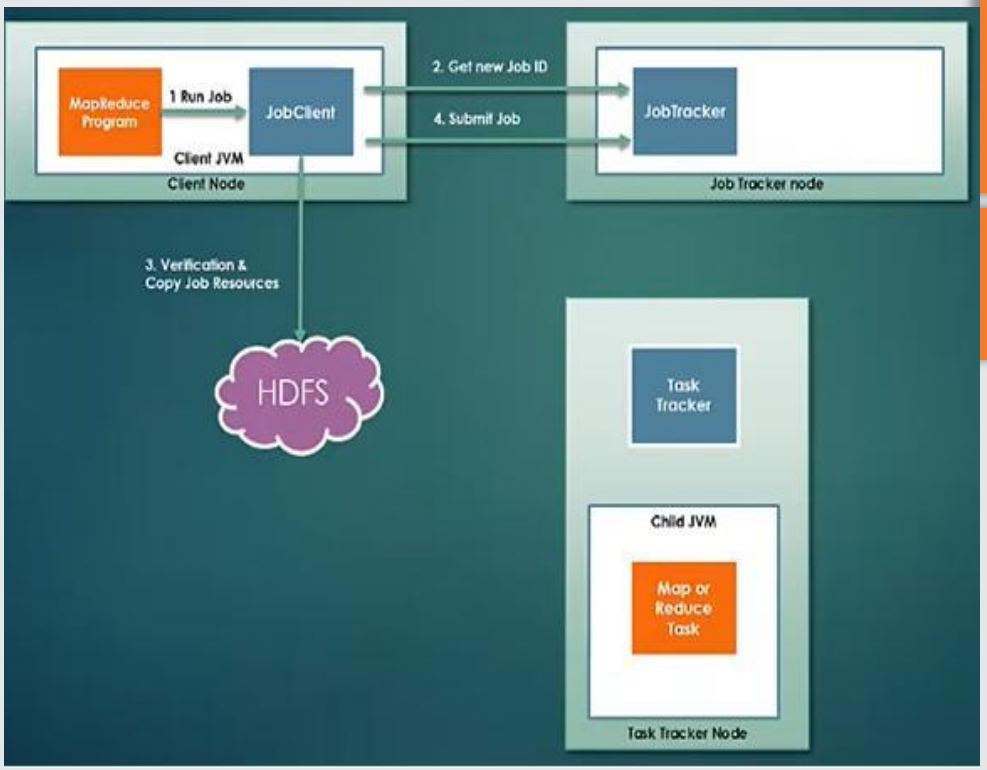
Job Initialization



Scheduler picks the job from the queue.

Creates an object which encapsulates its tasks and does book keeping

Job Initialization

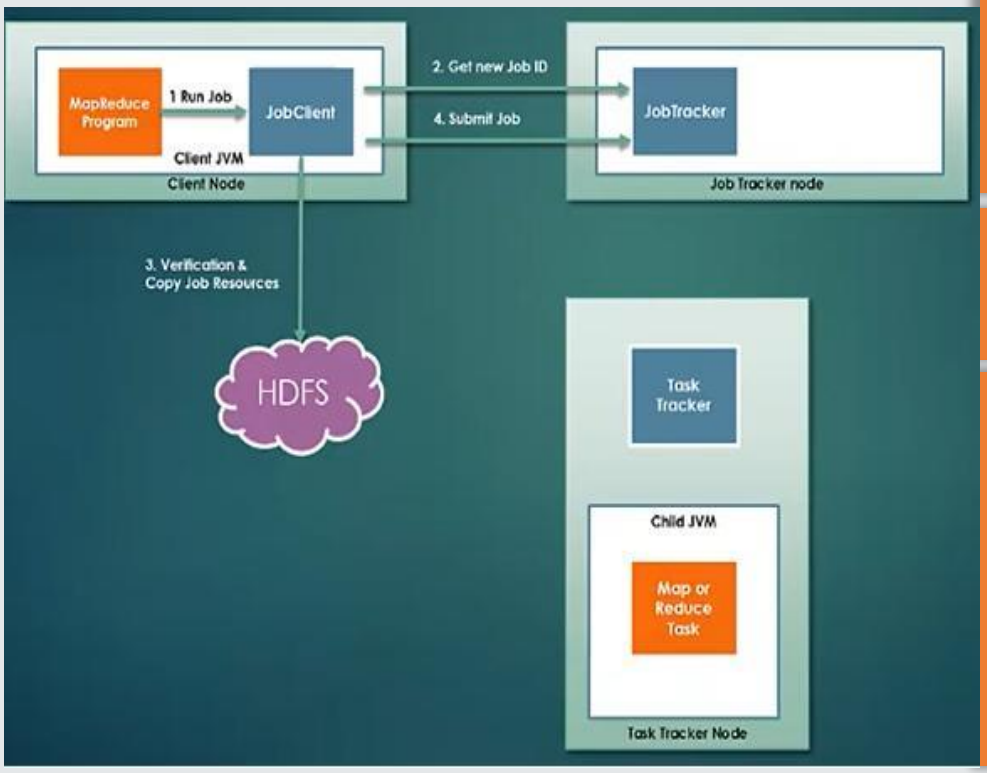


Scheduler picks the job from the queue.

Creates an object which encapsulates its tasks and does book keeping

Creates 1 Map task per split.

Job Initialization



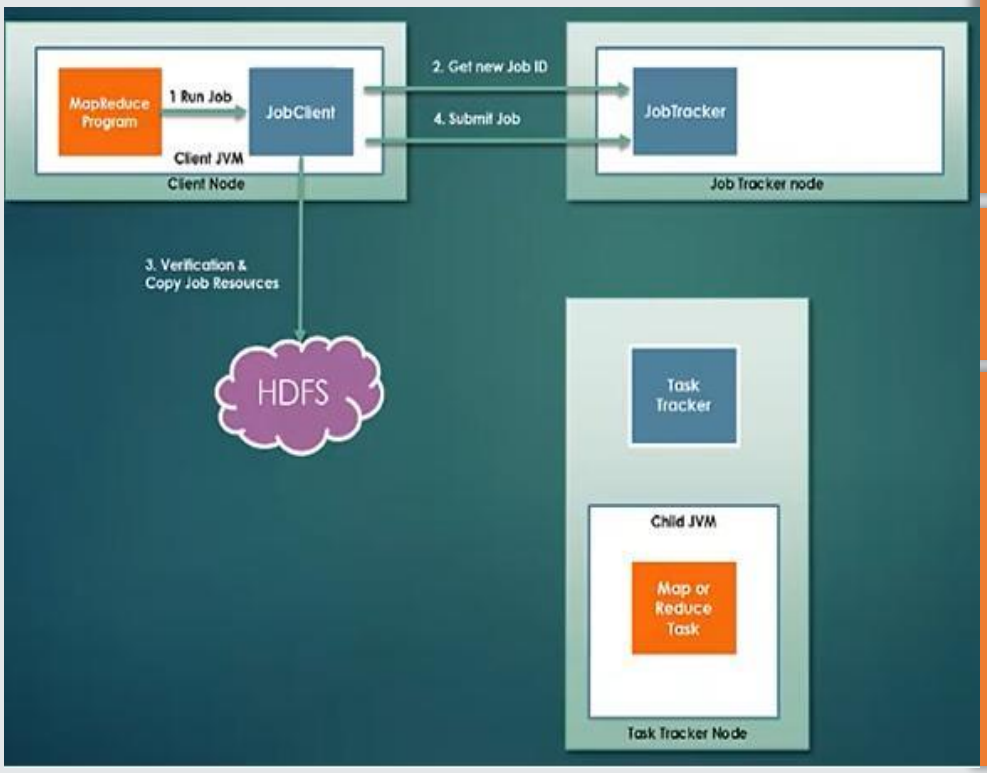
Scheduler picks the job from the queue.

Creates an object which encapsulates its tasks and does book keeping

Creates 1 Map task per split.

Number of reducers are decided by
“Mapred.reduce.tasks”
“job.setNumReduceTasks()”

Job Initialization



Scheduler picks the job from the queue.

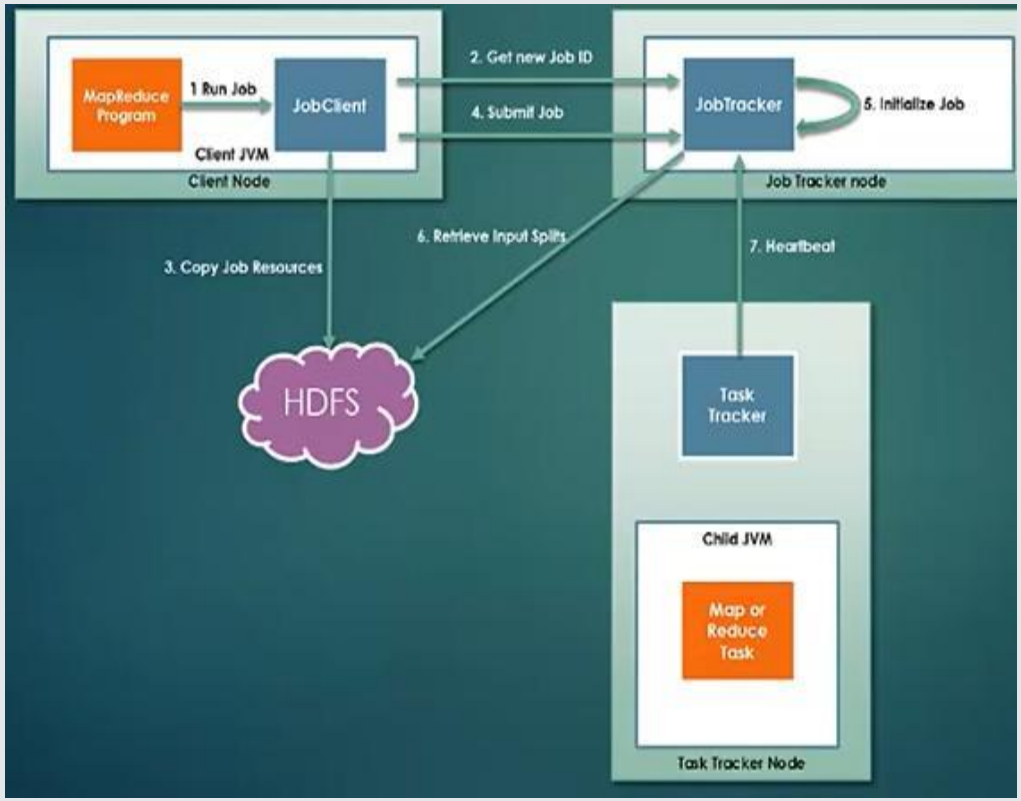
Creates an object which encapsulates its tasks and does book keeping

Creates 1 Map task per split.

Number of reducers are decided by
“Mapred.reduce.tasks”
“job.setNumReduceTasks()”

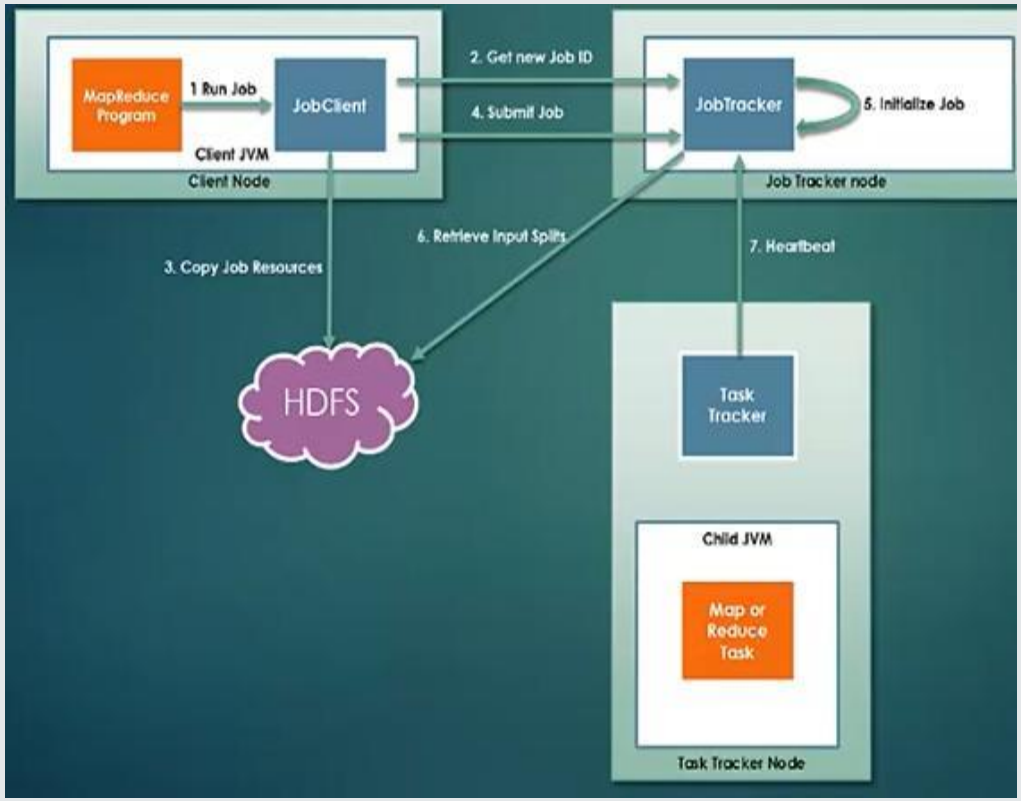
Creates setup and cleanup jobs on task trackers.

Task Assignment



TaskTracker runs a simple loop that periodically sends heartbeat.

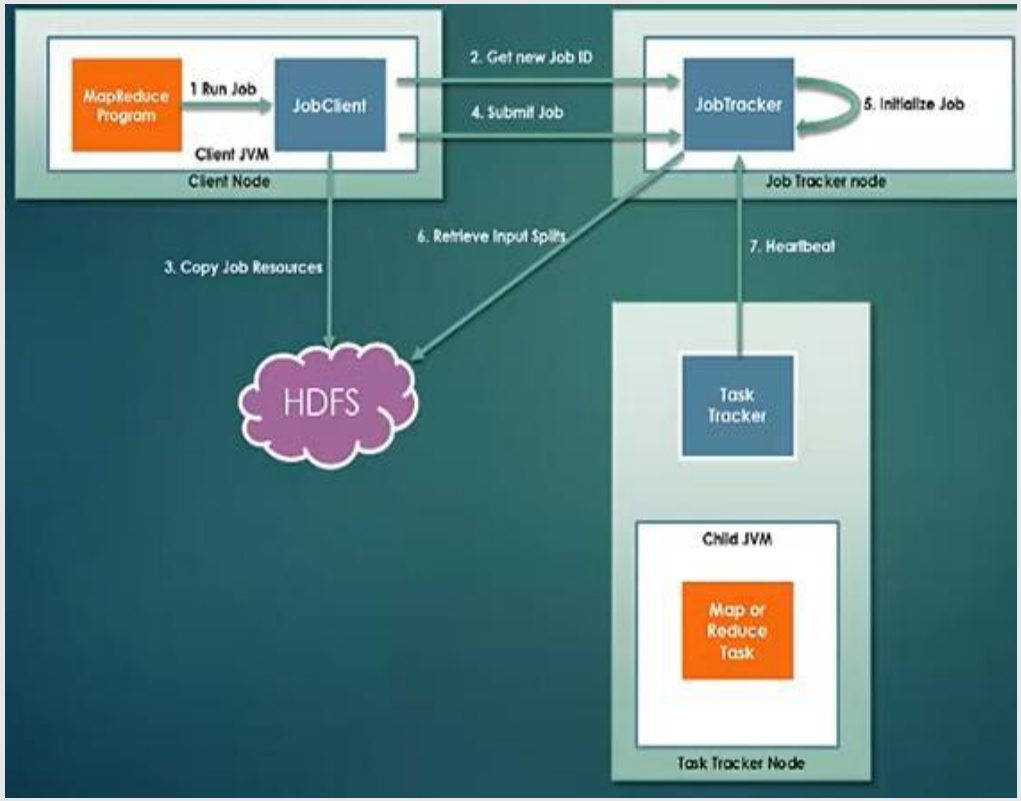
Task Assignment



TaskTracker runs a simple loop that periodically sends heartbeat.

TaskTracker have fixed number of slots to run Map and Reduce Tasks

Task Assignment

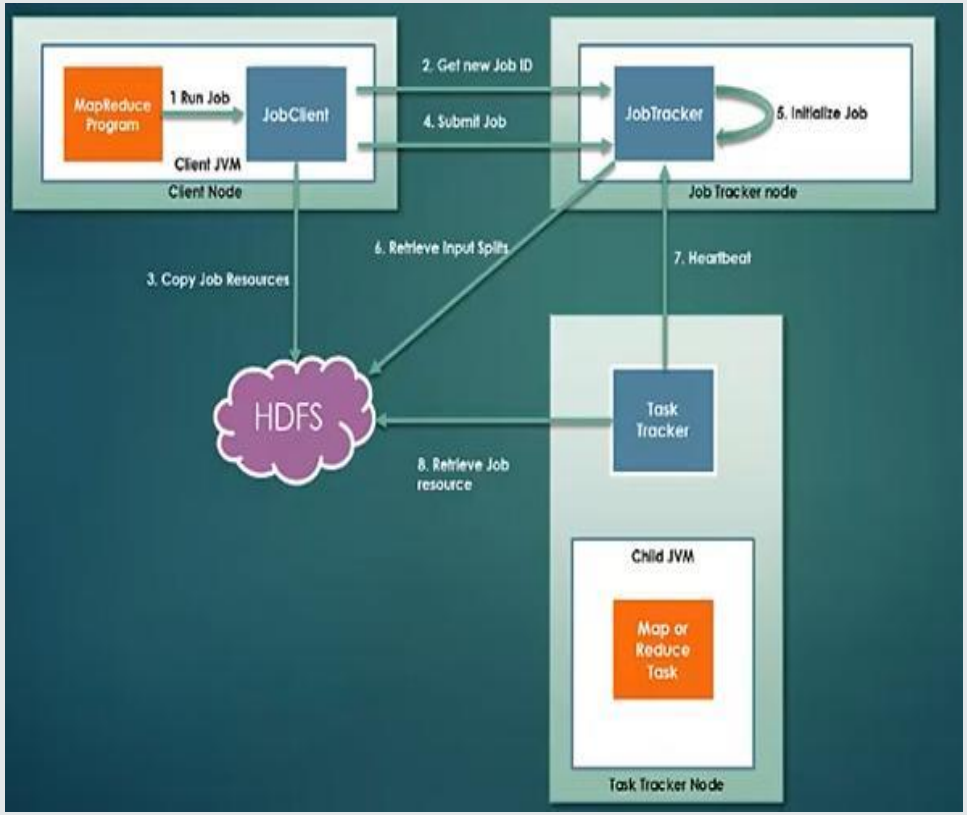


TaskTracker runs a simple loop that periodically sends heartbeat.

TaskTracker have fixed number of slots to run Map and Reduce Tasks

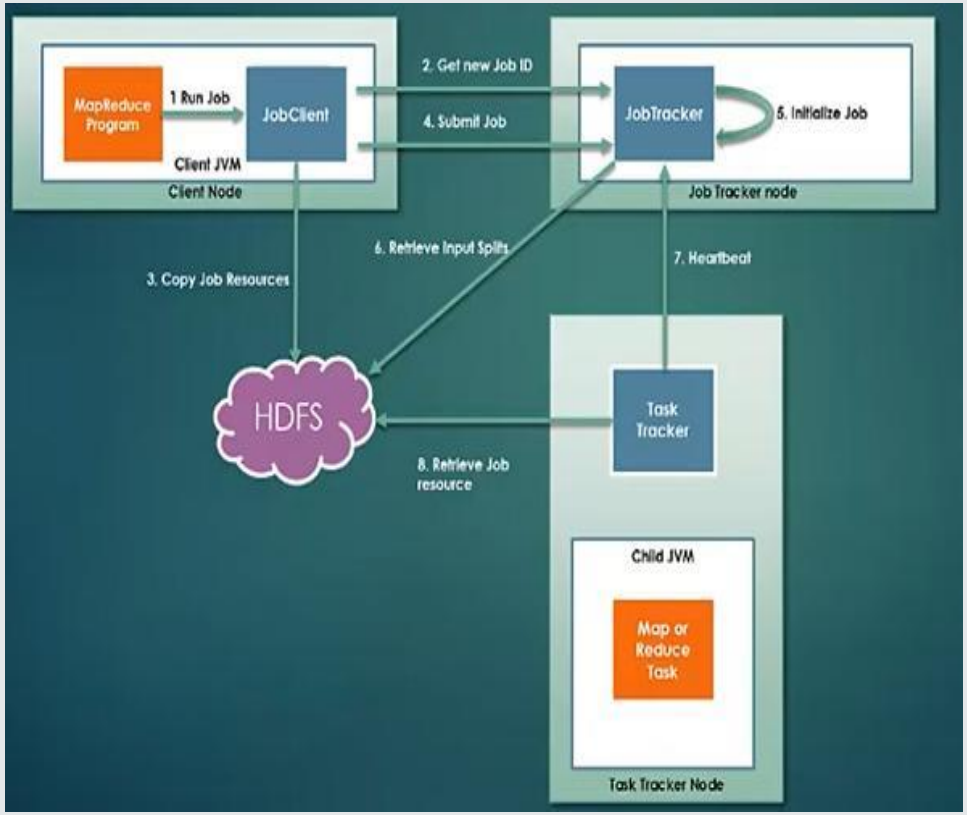
The number of tasks that can run be modified considering memory an cores of the processing node.

Task Execution



Retrieve the JAR file from the HDFS.

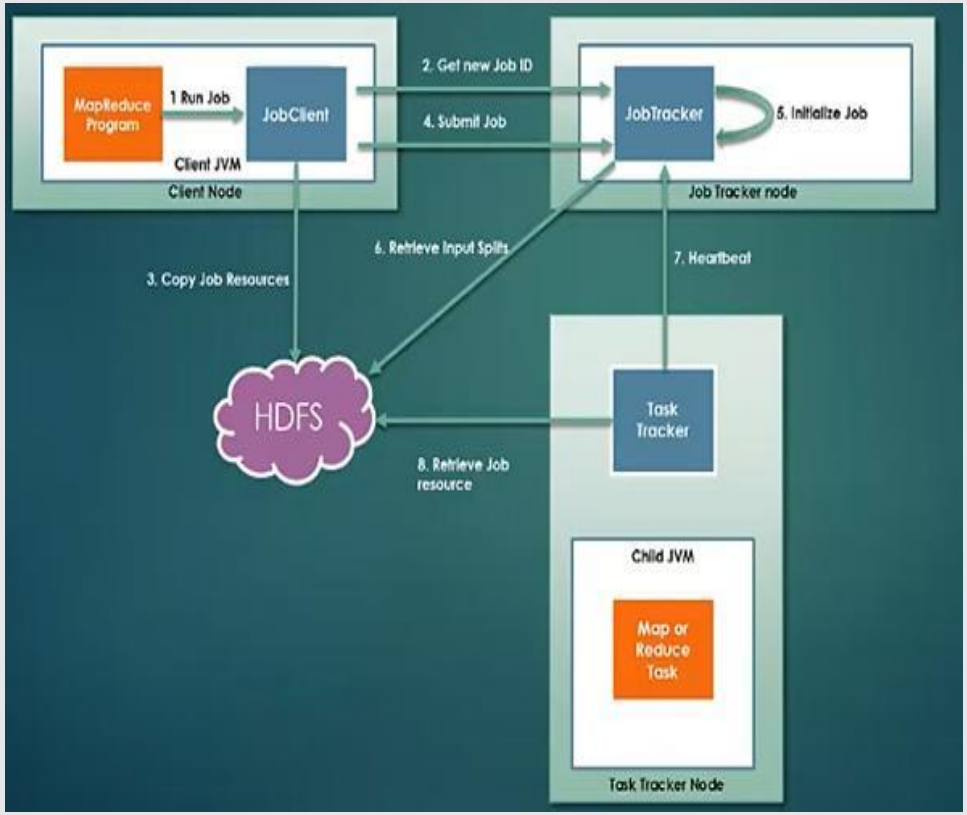
Task Execution



Retrieve the JAR file from the HDFS.

Launches JVM to run

Task Execution

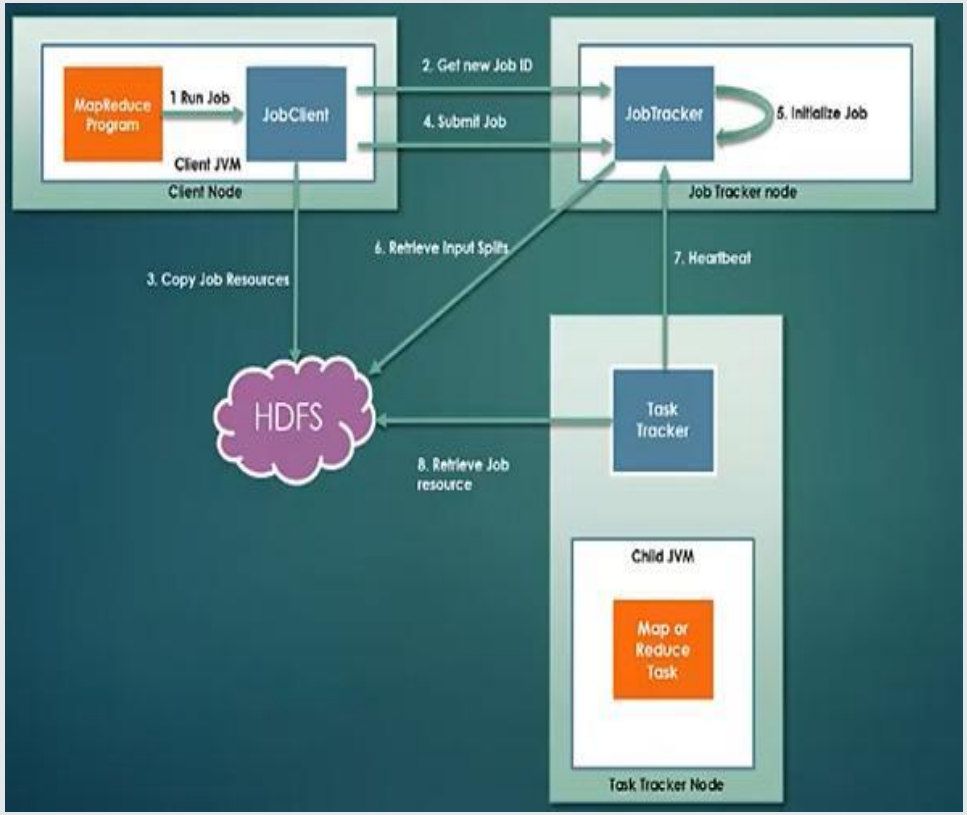


Retrieve the JAR file from the HDFS.

Launches JVM to run

TaskTracker sends regular updates to JobTracker about the statuses.

Task Execution



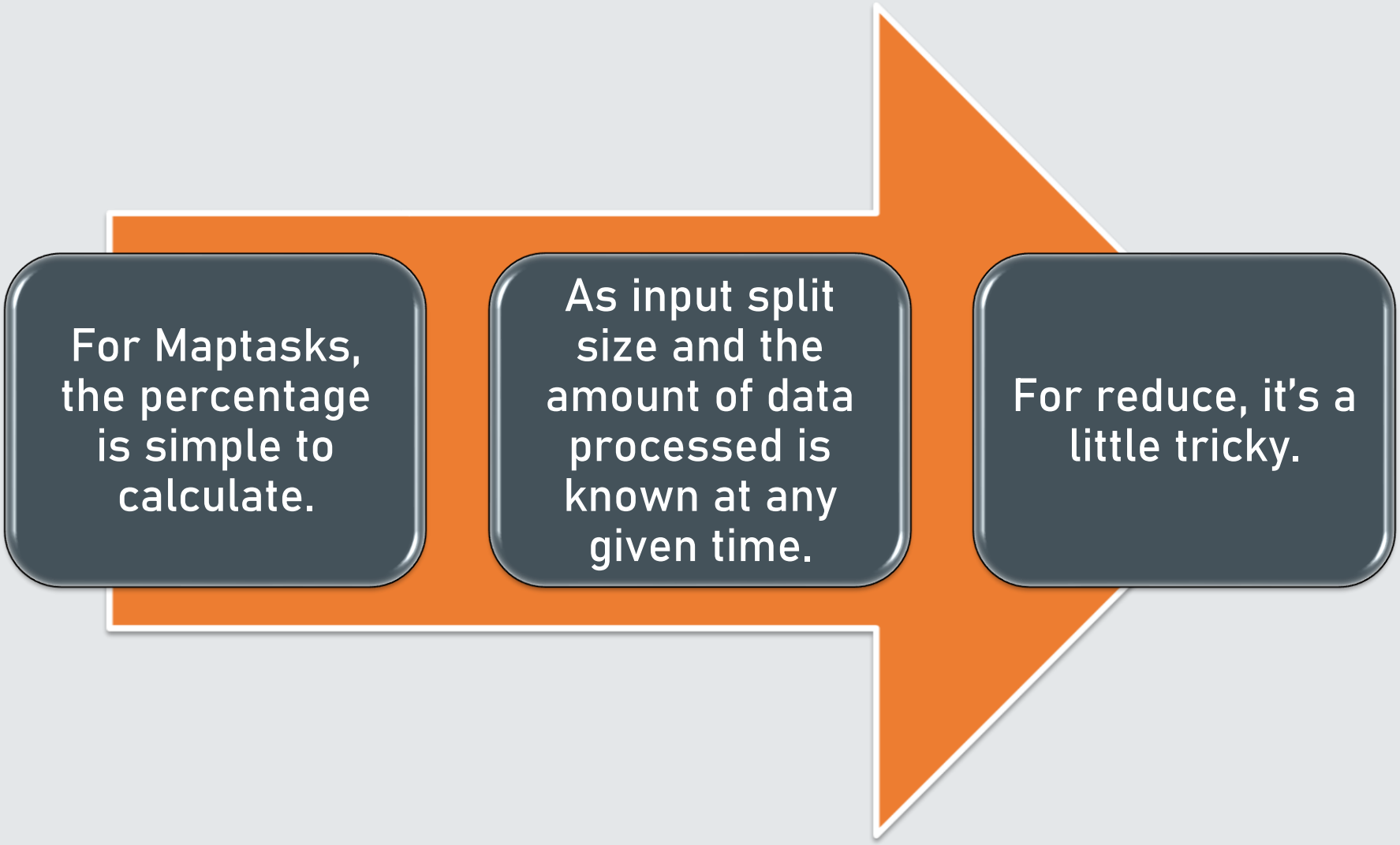
Retrieve the JAR file from the HDFS.

Launches JVM to run

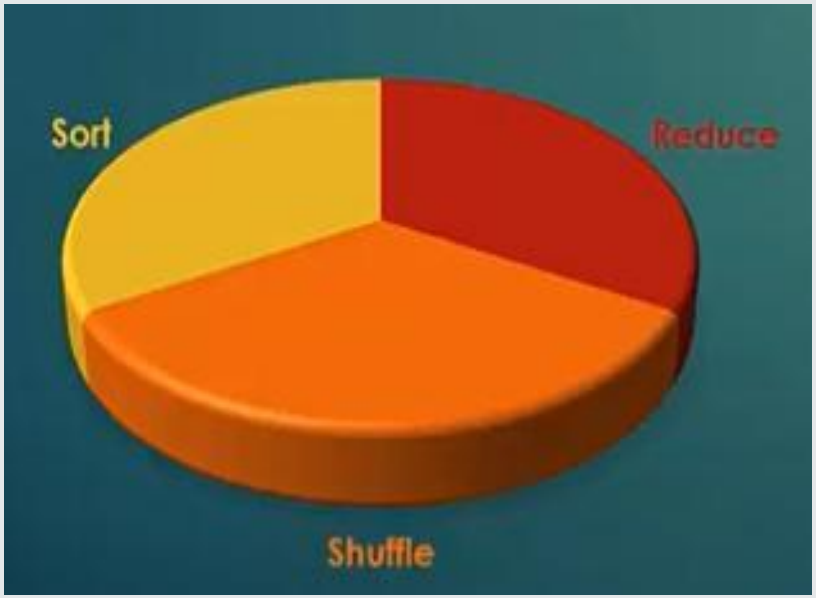
TaskTracker sends regular updates to JobTracker about the statuses.

Cleanup job run after completion of reduce tasks

Progress and Update Calculation

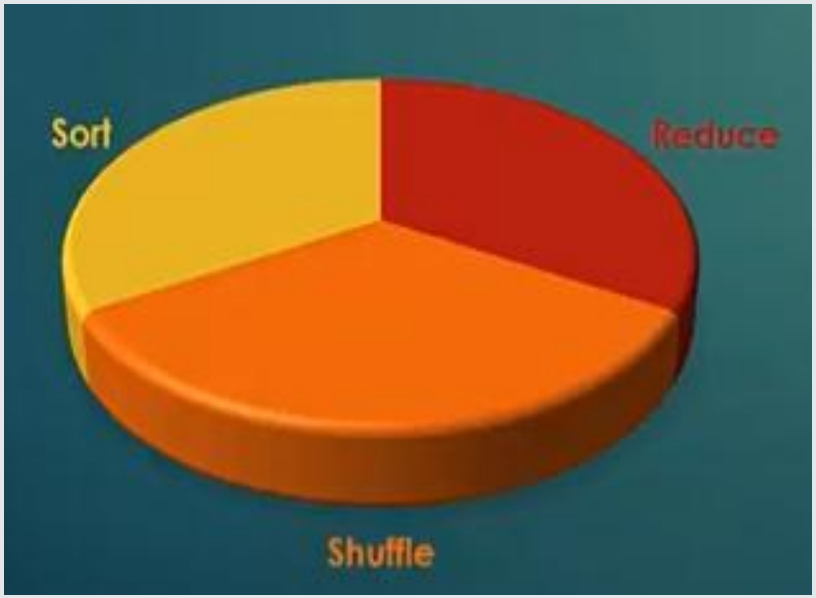


Progress and Update Calculation



If just started with reduce phase, completion would be:

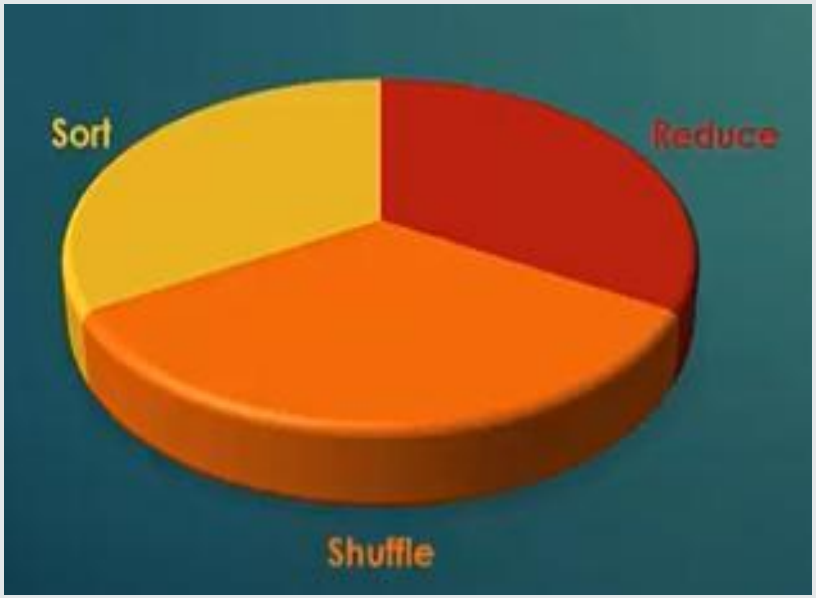
Progress and Update Calculation



If just started with reduce phase, completion would be:

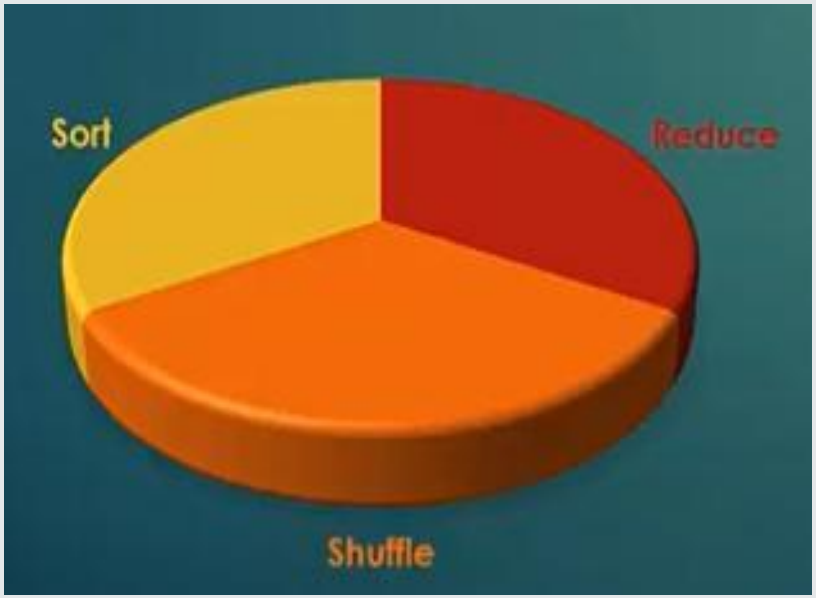
1/3 (Sort) + 1/3 (Shuffle) = 2/3 or 67%

Progress and Update Calculation



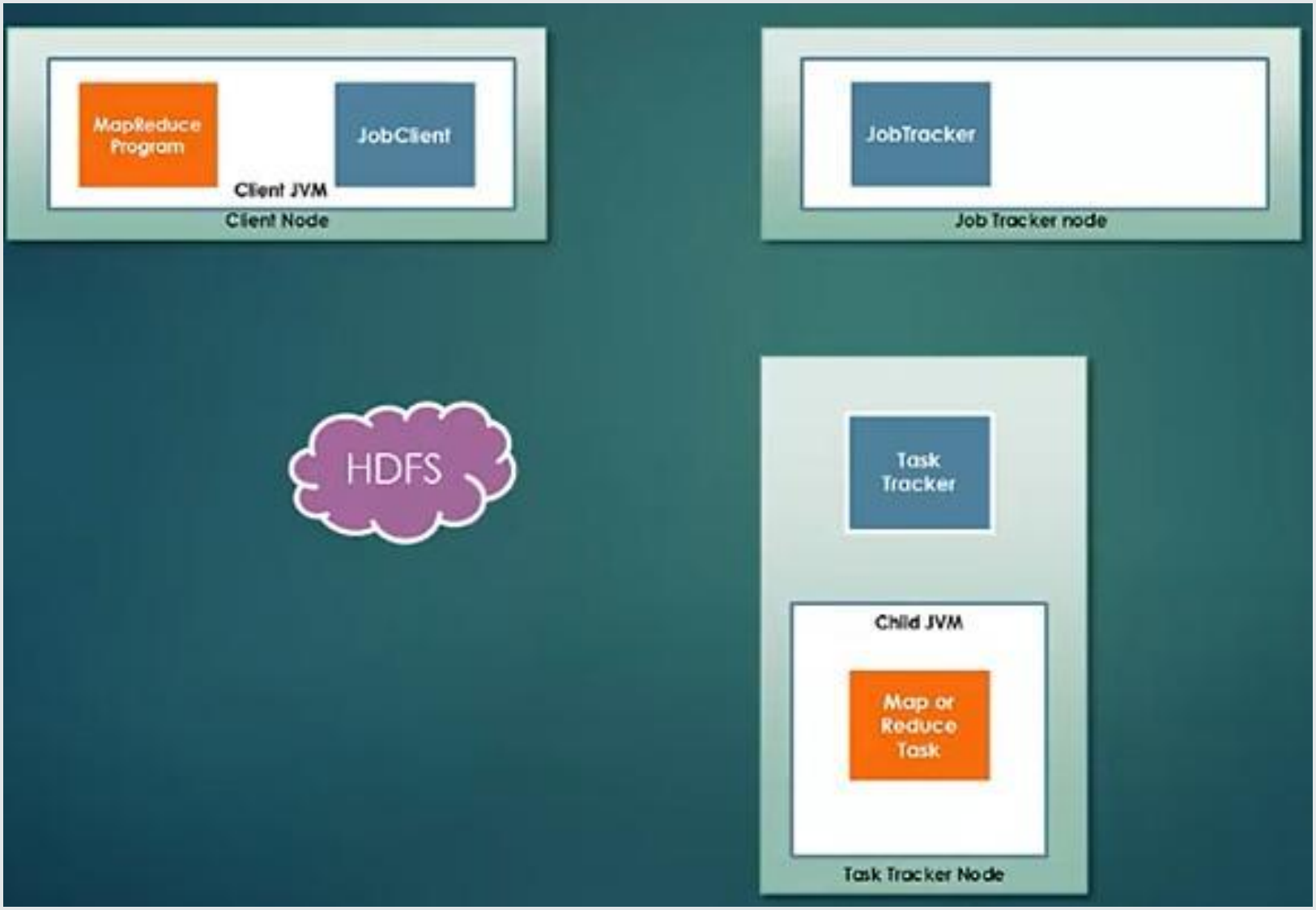
If just half of reduce phase is done, then completion would be:

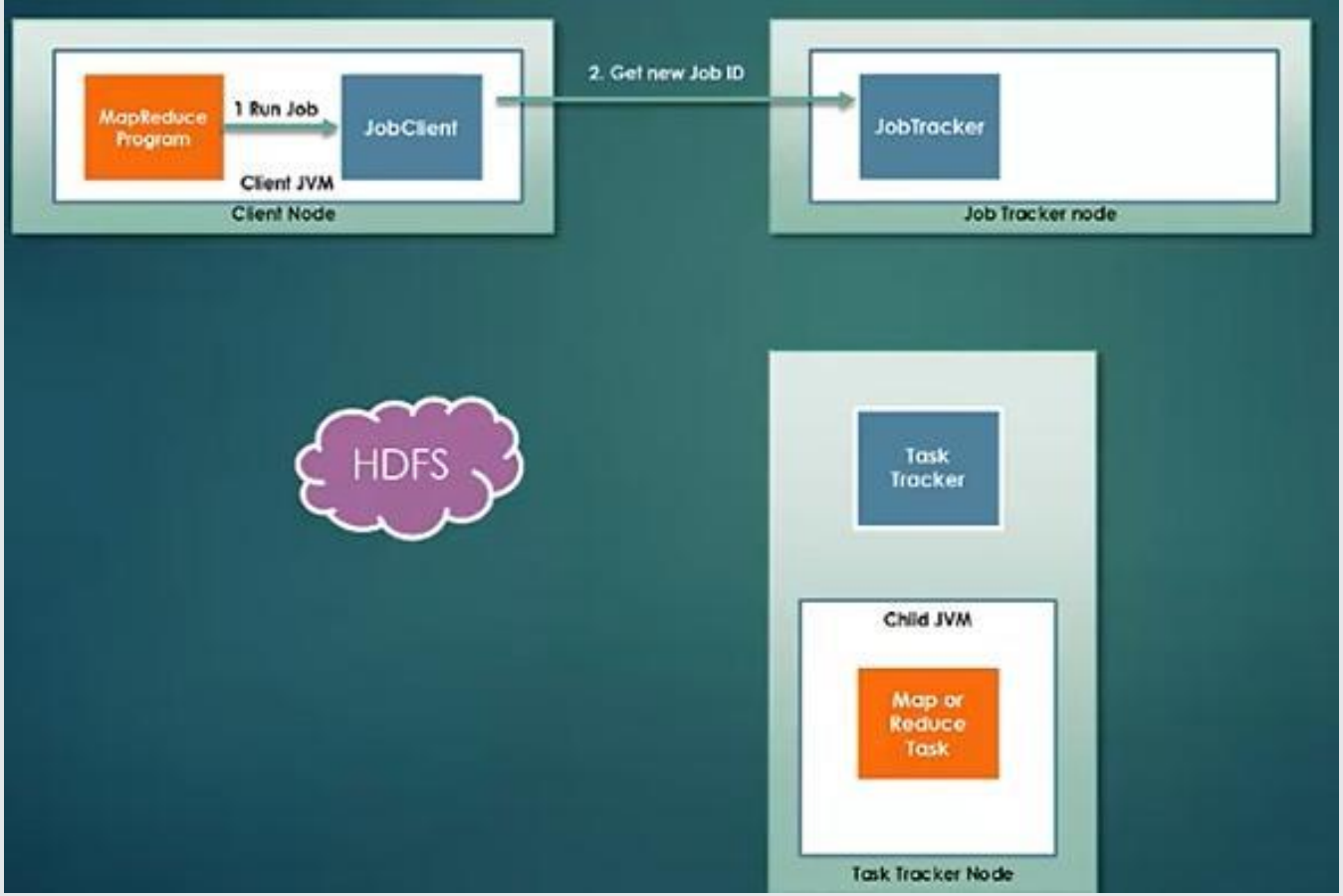
Progress and Update Calculation



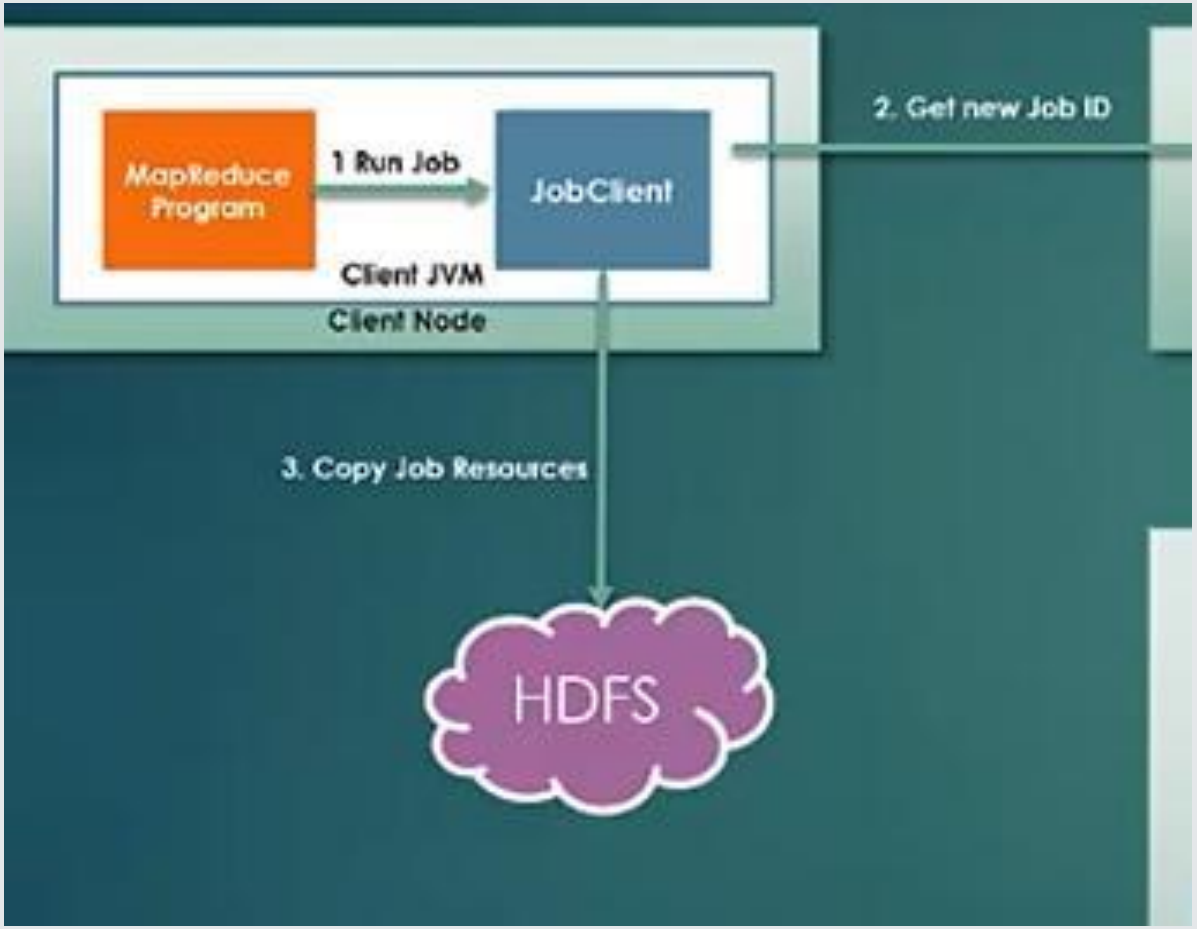
If just half of reduce phase is done, then completion would be:

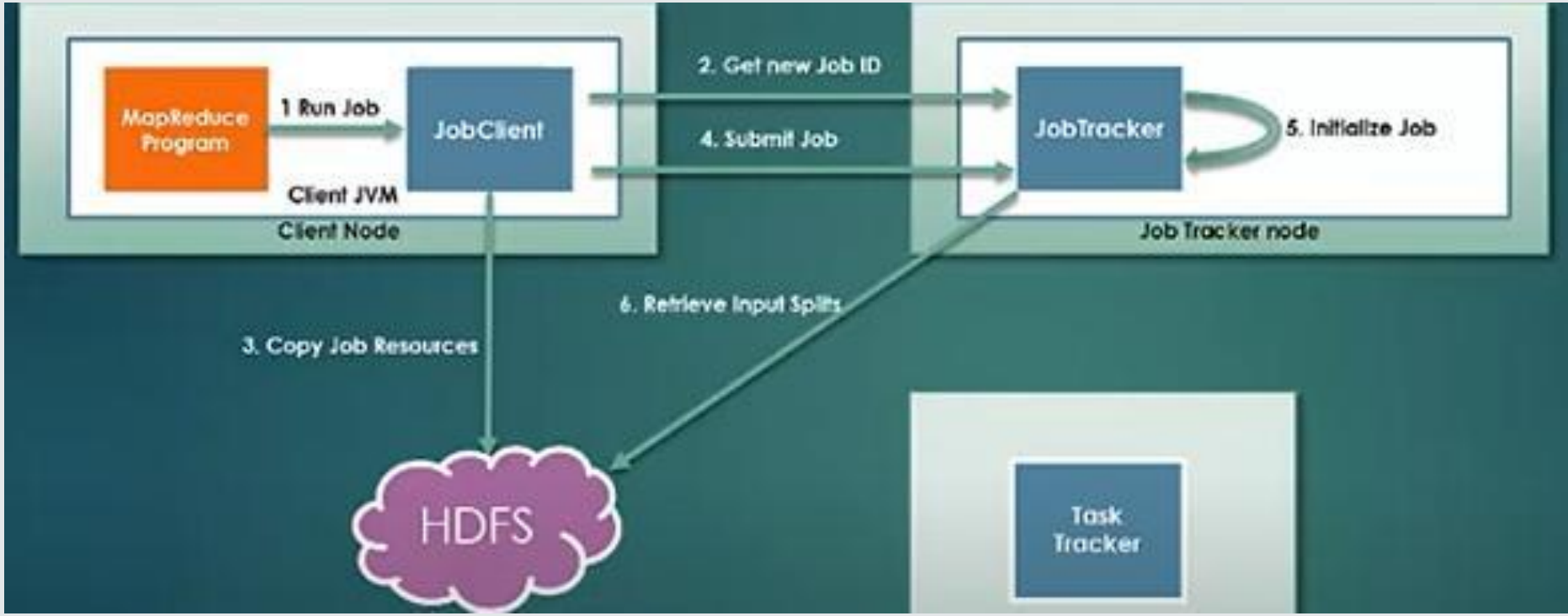
1/3 (Sort) + 1/3 (Shuffle) +1/6 (Reduce) = 5/6 or 83%


MapReduce Program
JobClient
Client JVM
Client Node
JobTracker
Job Tracker node
HDFS
Task Tracker
Child JVM
Map or Reduce Task
Task Tracker Node

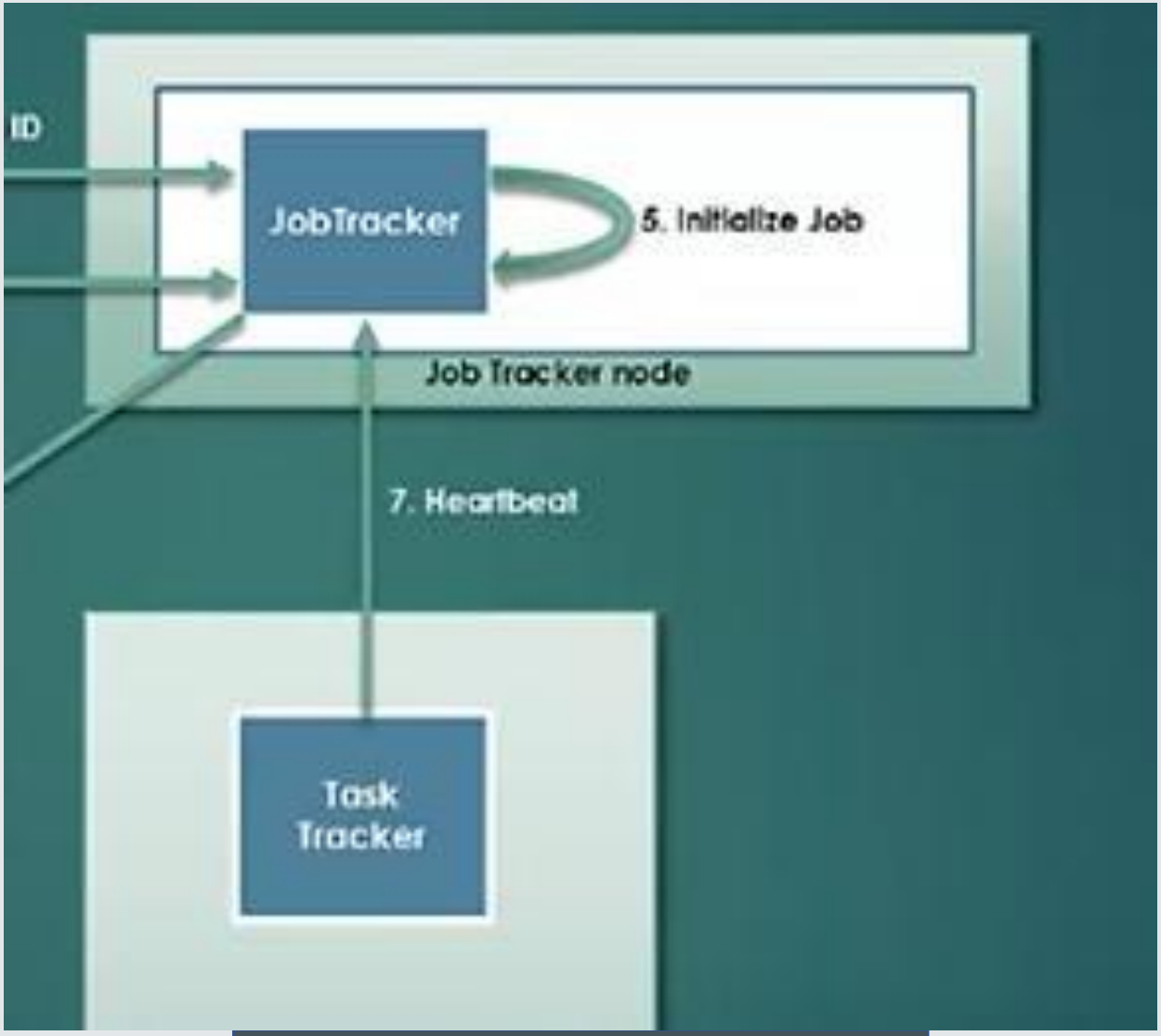


Job Submission

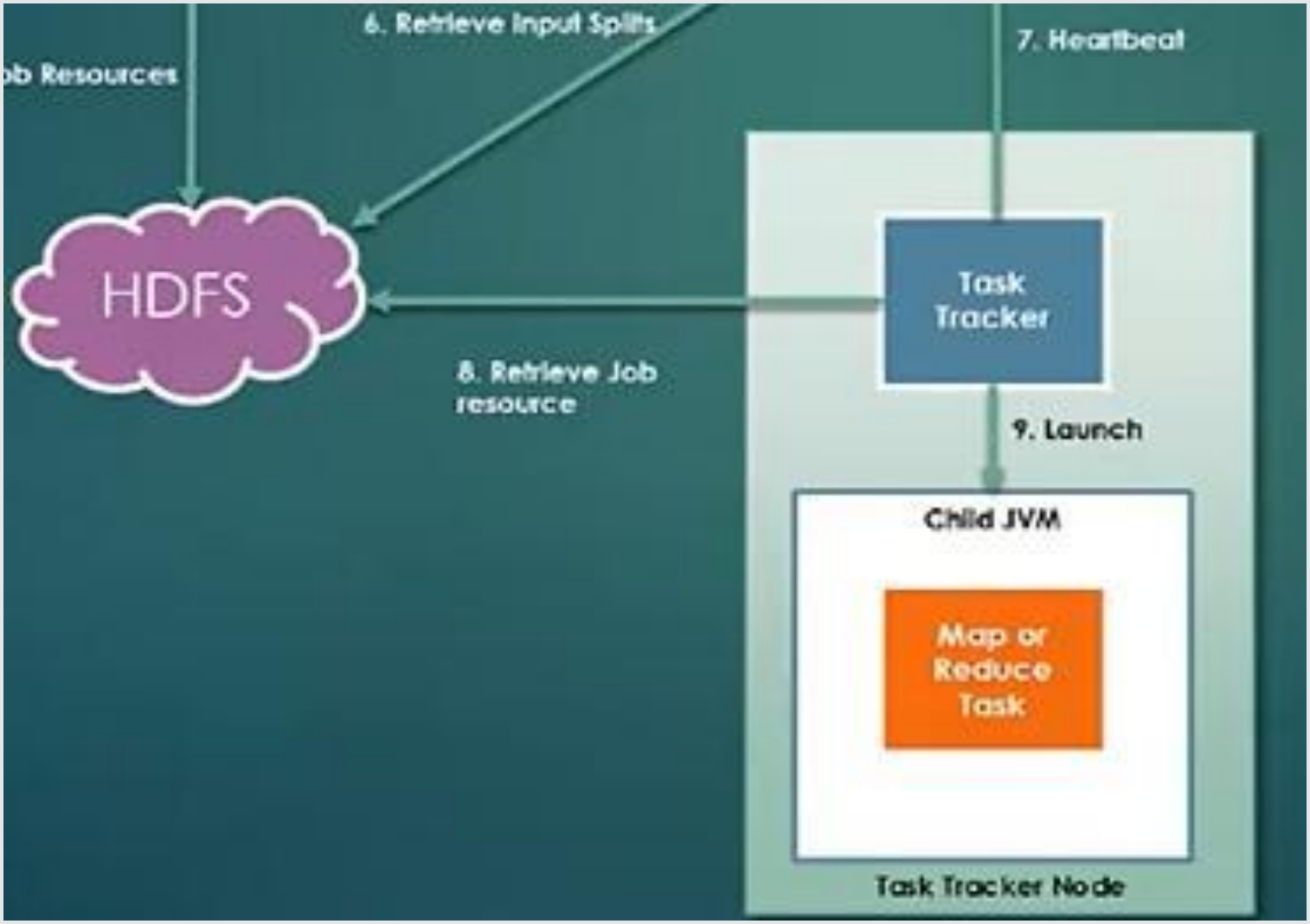

MapReduce Program
1 Run Job
JobClient
2. Get new Job ID
Client JVM
Client Node
3. Copy Job Resources
HDFS



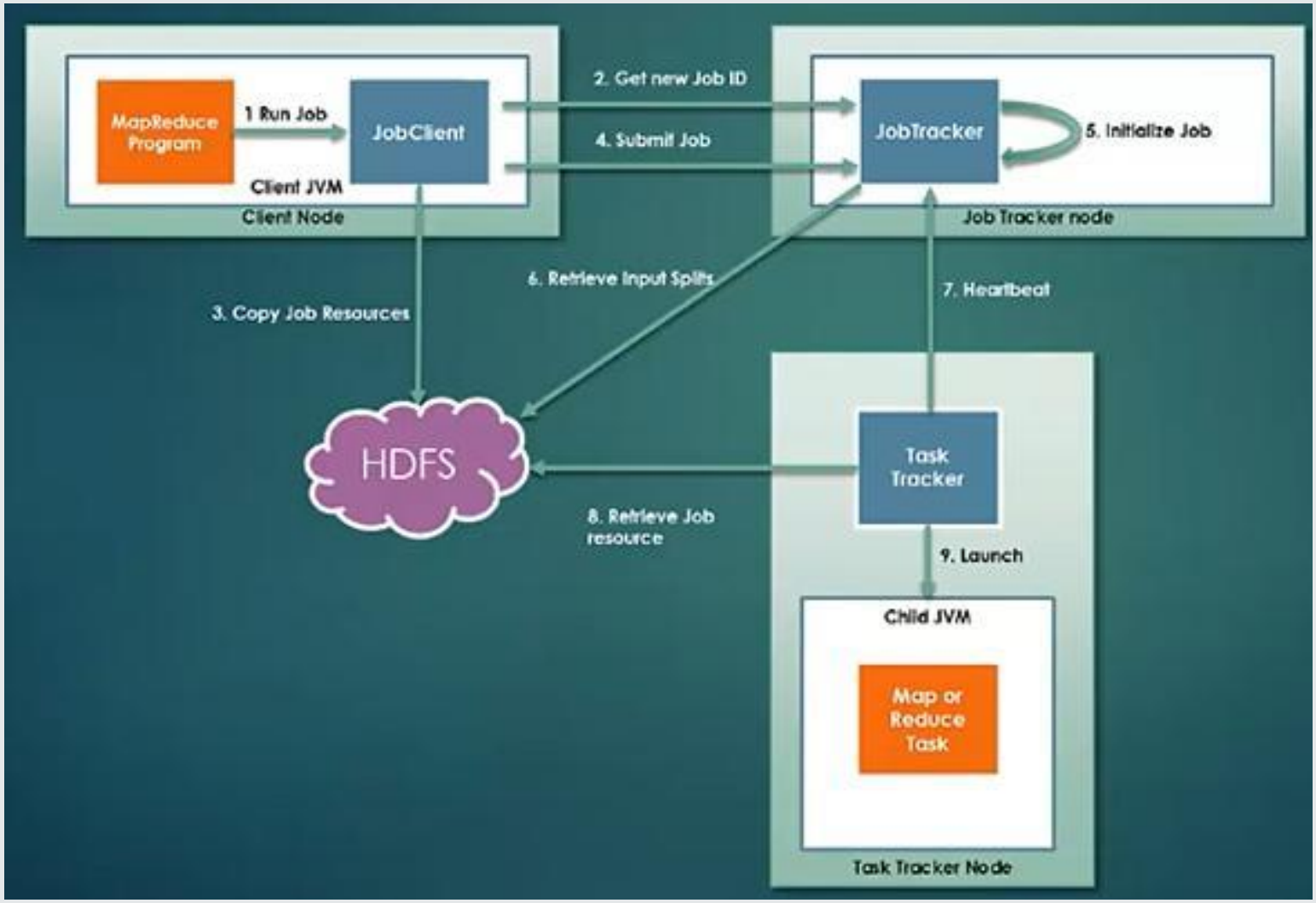
Task Intialization



Task Assignment



Task Execution



Job Completion


That’s all for now…