Mapreduce – Paradigm

Sundharakumar KB

Department of Computer Science and Engineering
School of Engineering

Shiv Nadar University Chennai

SHIV NADAR UNIVERSITY CHENNAI

Mapreduce Programming Model

Input.txt	Map Phase	Shuffle & Sort	Reducer	output
apple, banana	apple, 1	apple,1		
banana, orange	banana,1	apple,1		
orange, orange	banana,1	apple,1		
apple, orange	orange,1	banana,1	apple,3	apple,3
orange, apple	orange,1	banana,1	banana,5	banana,5
banana, banana	orange,1	banana,1	orange,6	orange,6
orange, banana	apple,1	banana,1		
	banana,1	banana,1		
	banana,1	orange,1		
	orange,1	orange,1		
	banana,1	orange,1		
		orange,1		
		orange,1		
		orange,1		
		orange,1		
		orange,1		

Anatomy of a file read

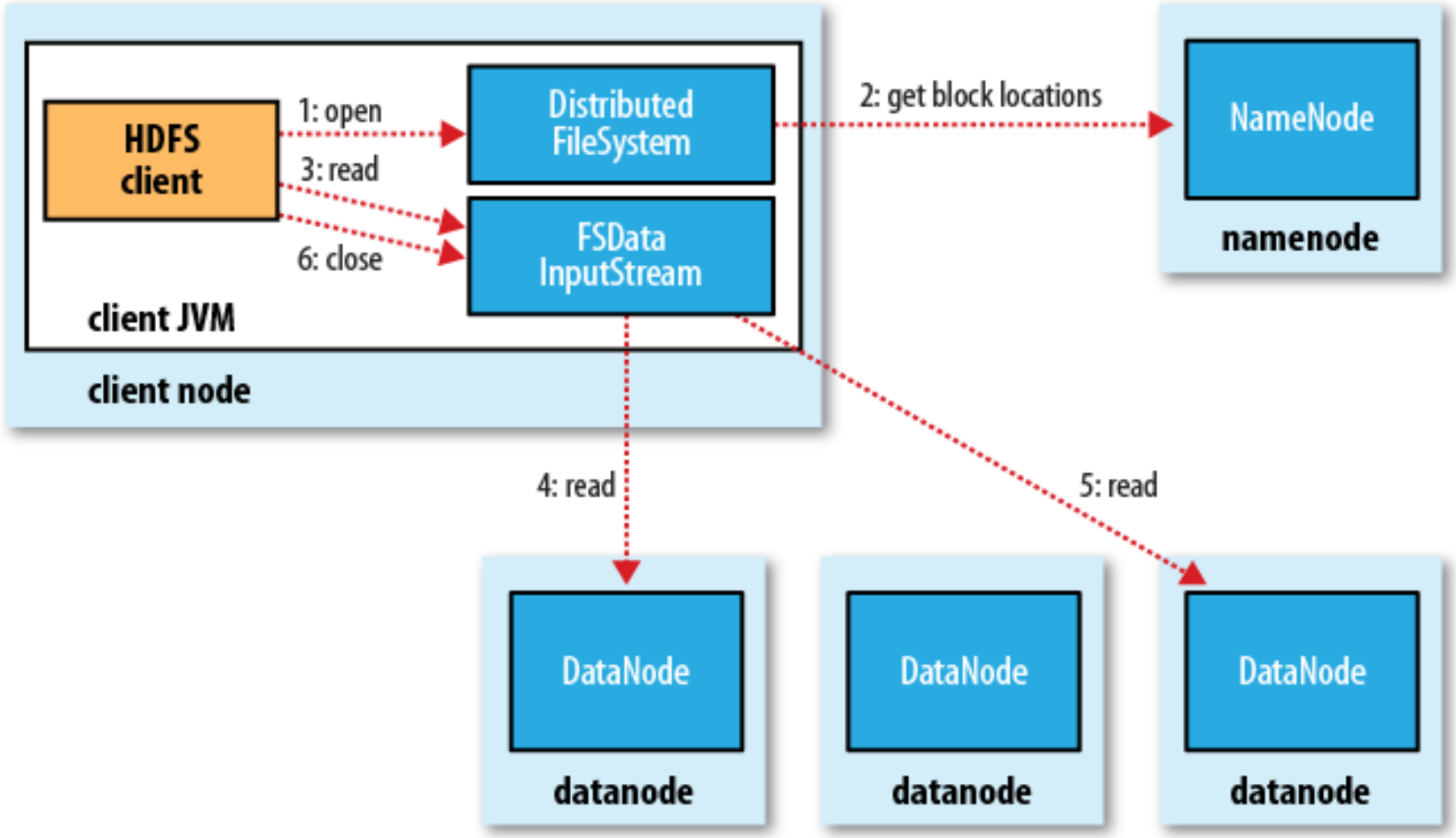


Image ref: Hadoop the definitive guide

Anatomy of a file write

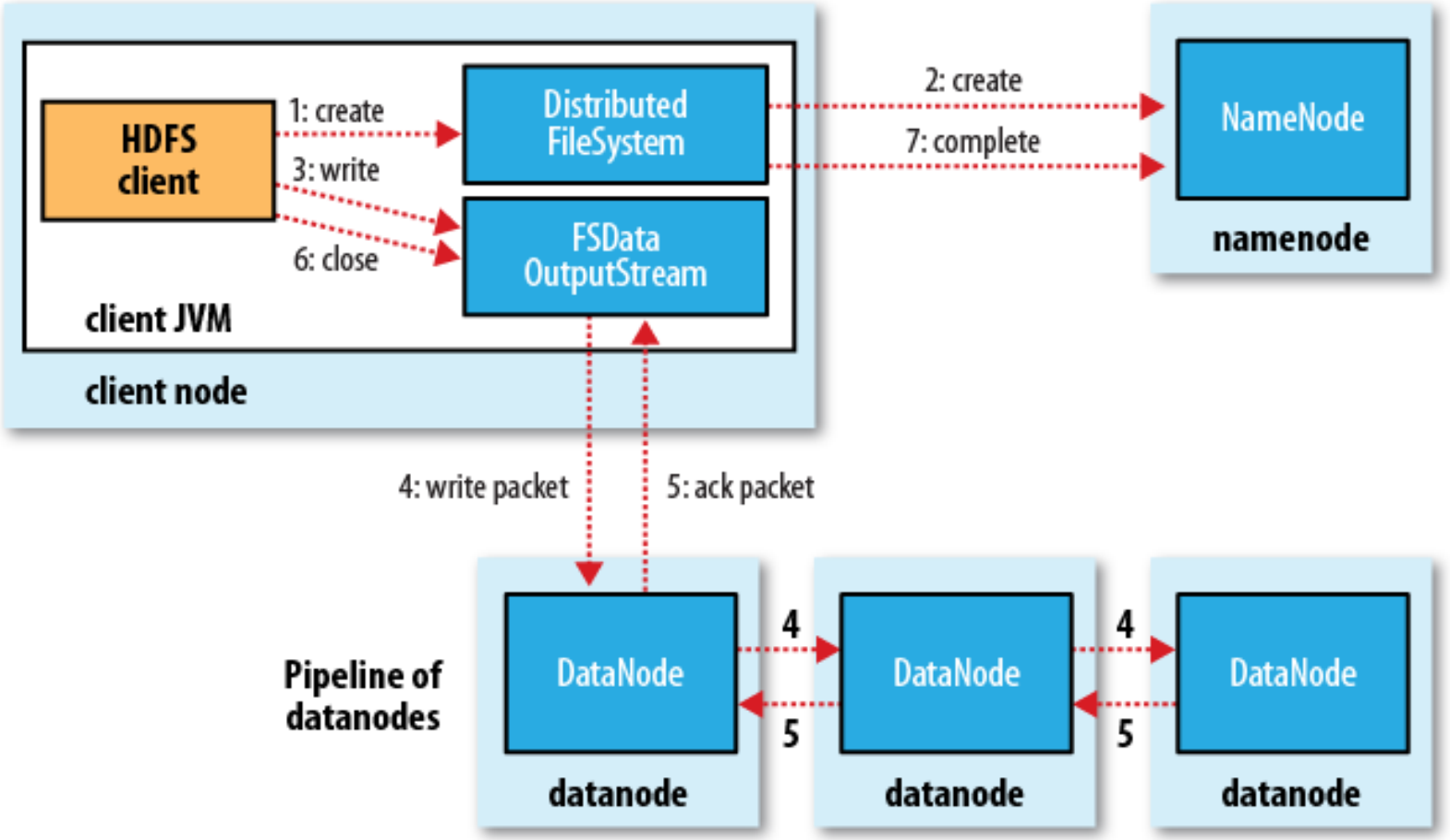
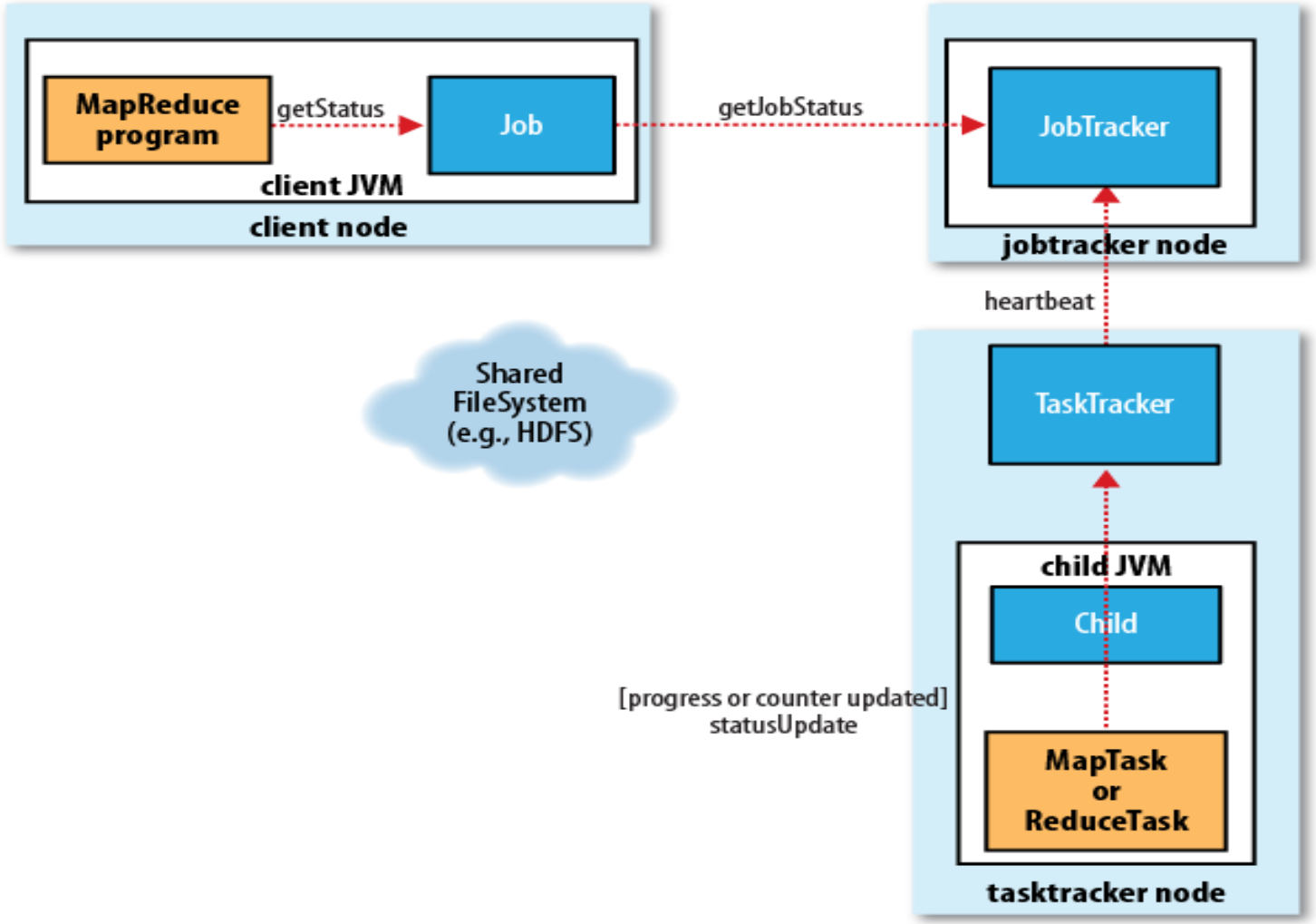


Image ref: Hadoop the definitive guide

Job Completion



Job Completion

- Task failure
- Task tracker failure
- Job tracker failure
- Namenode failure

HDFS components – version 2.x

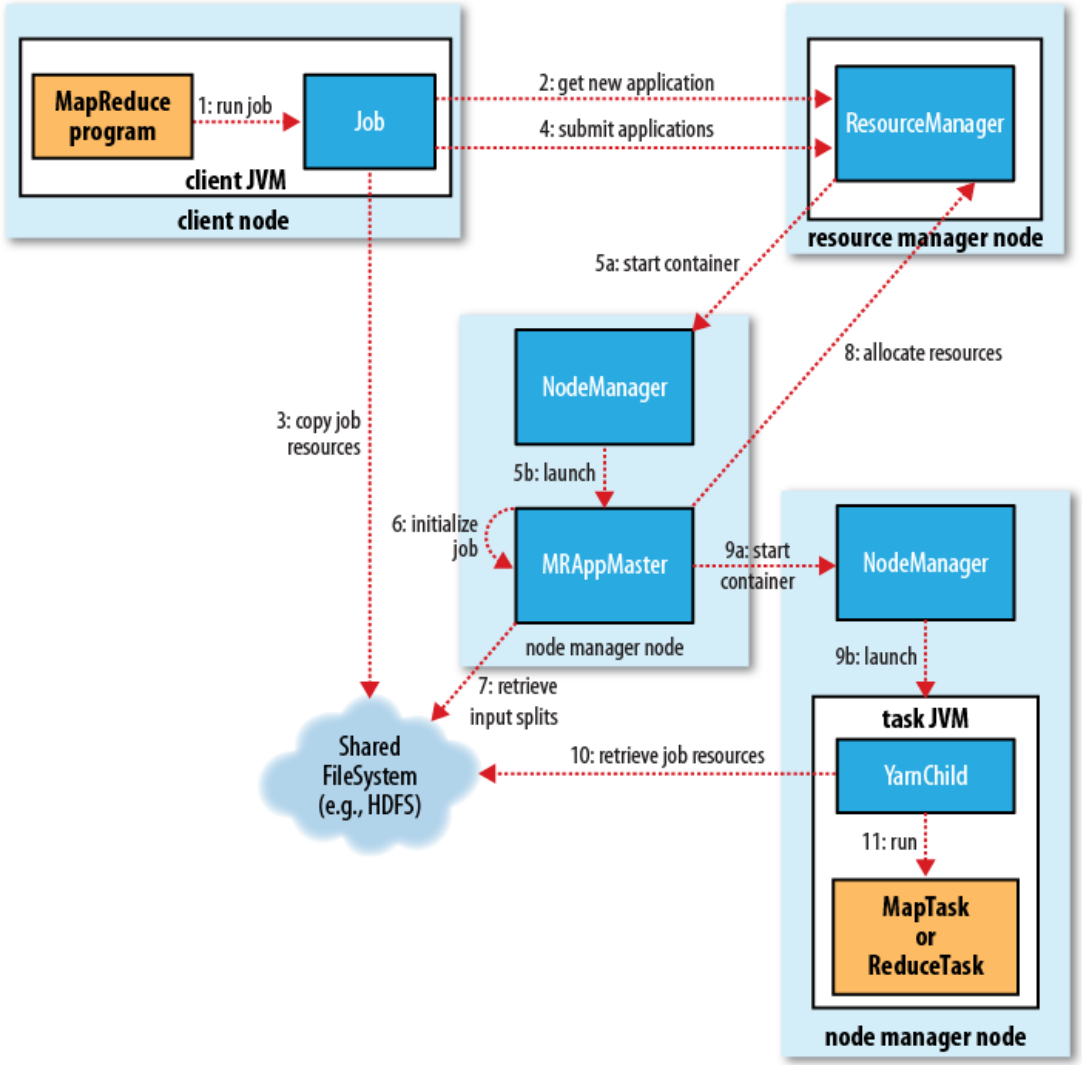
- NameNode (High availability)
- YARN (yet another resource negotiator)
- Resource Manager.

YARN

- YARN (Yet Another Resource Negotiator) is Hadoop cluster's resource management system.
- It was mainly introduced in Hadoop v2 mainly to help in MapReduce implementation but in general it can help in any distributed programming paradigms.
- With the introduction of YARN, the roles of job tracker and task trackers are removed and instead Node manager, Application master and resource manager are introduced

MR1	YARN
Job Tracker	Resource manager, application master
Task Tracker	Node Manager

Anatomy of Job run in YARN



SHIV NADAR UNIVERSITY CHENNAI

Job completion in YARN