Week 10
Distributed Computing With Hadoop Introduction

Mastering Cloud Computing

Coleman Kane
(based on material by Paul Talaga)

UNIVERSITY OF Cincinnati

Distributed Computing

Recall W03.2 “Distributed Computing” lecture

Needs management of distribution of work:

- Compute resources (CPU/RAM)
- Long-Term Storage (Filesystem)
- User-supplied “algorithm” or “work”
- Scheduling/dependency

UNIVERSITY OF Cincinnati

Hadoop Platform

Provides framework/API for distributed computing

- http://hadoop.apache.org/
- https://en.wikipedia.org/wiki/Apache_Hadoop

Hides away distributed back-end

Provide consolidated abstractions for:

- workload distribution
- storage
- application development

UNIVERSITY OF Cincinnati

Hadoop Core

Hadoop is built atop some core components:

- Hadoop Common: Common shared libraries, runtime
- Hadoop Distributed File System (HDFS) - Storage abstraction across multiple nodes providing a single filesystem environment
- Hadoop YARN (Yet Another Resource Negotiator): Compute management layer, or “distributed OS” for your Hadoop cluster

Hadoop Architecture

At its core - Java application, platform independent

Data-centric

Core tools & frameworks

Cluster built out of “nodes” that provide resources - execution time, disk space, etc.

Application development system, to build distributed solutions to user data problems

UNIVERSITY OF Cincinnati

Hadoop Common

Basic primitives for programming environment:

- Core data types
- Exceptions
- Data structures
- IPC primitives
- Logic, math, computation runtime
- Data handling/encoding/transfer

HDFS - File System

Filesystem abstraction for Hadoop clusters

Manage storage efficiently for:

- Very large files
- Streaming data access
- Built on commodity hardware
- High availability
- Concurrent access

UNIVERSITY OF Cincinnati

HDFS - Limitations

Doesn’t work as well for

- Many small files
- Low-latency data access

In these cases, you may always use another storage abstraction or engine

HDFS - Architecture

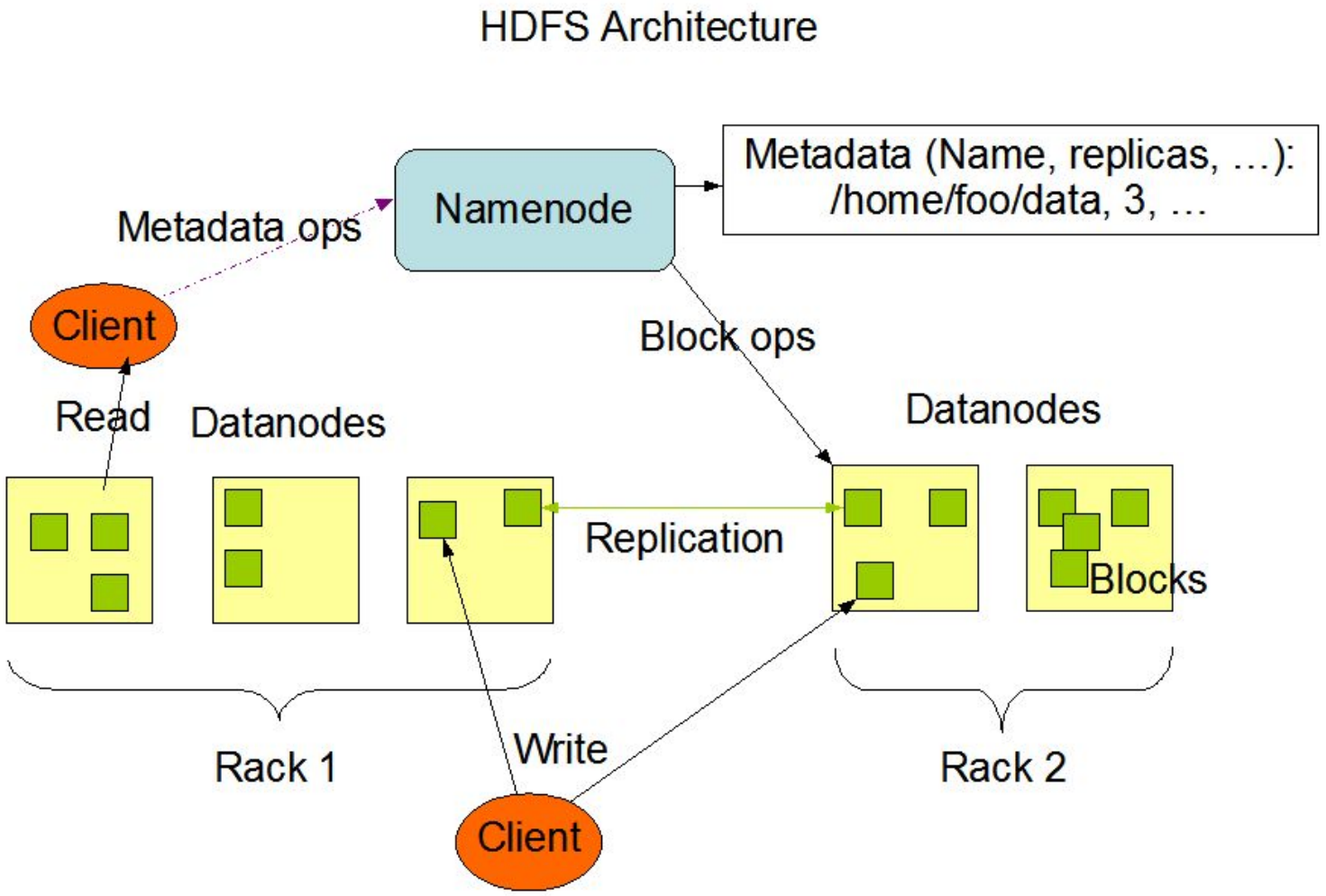
Built from nodes, which are server processes deployed to nodes in your cluster

namenode - HDFS metadata stored here, connects namespace to blocks in data store

datanode - The content of your files are broken into “blocks” and stored across these

UNIVERSITY OF Cincinnati

HDFS - Architecture Visual



Taken from: http://hortonworks.com/apache/hdfs/#section_2

UNIVERSITY OF Cincinnati

YARN - Concepts

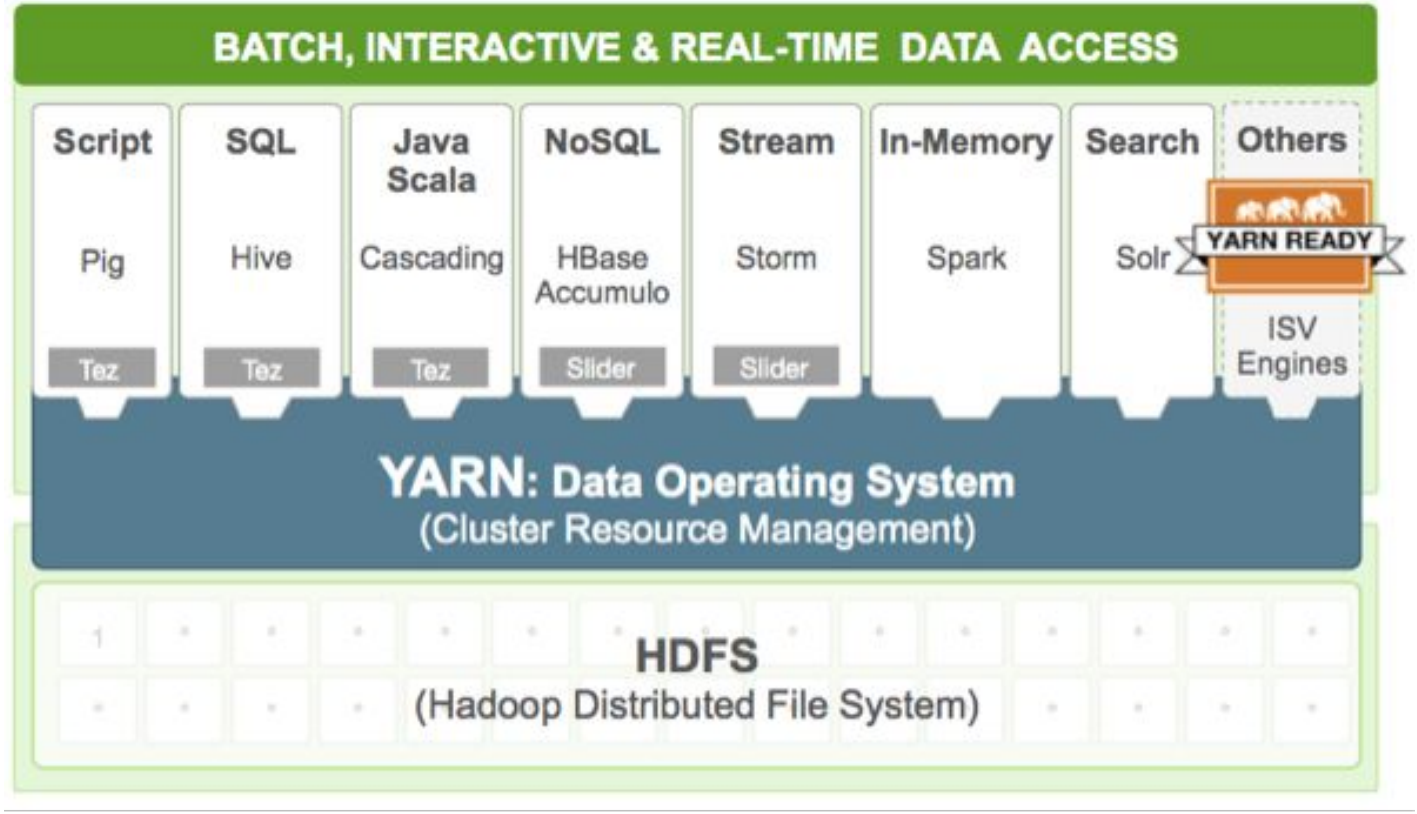
The platform development layer for Hadoop - or “Cluster OS”

In the above regard, it almost acts as a means of IaaS(M)

PaaS-like layers built atop this to provide “application interfaces” to Hadoop users.

Examples include: MapReduce, Spark

Hadoop, HDFS, YARN - Visual



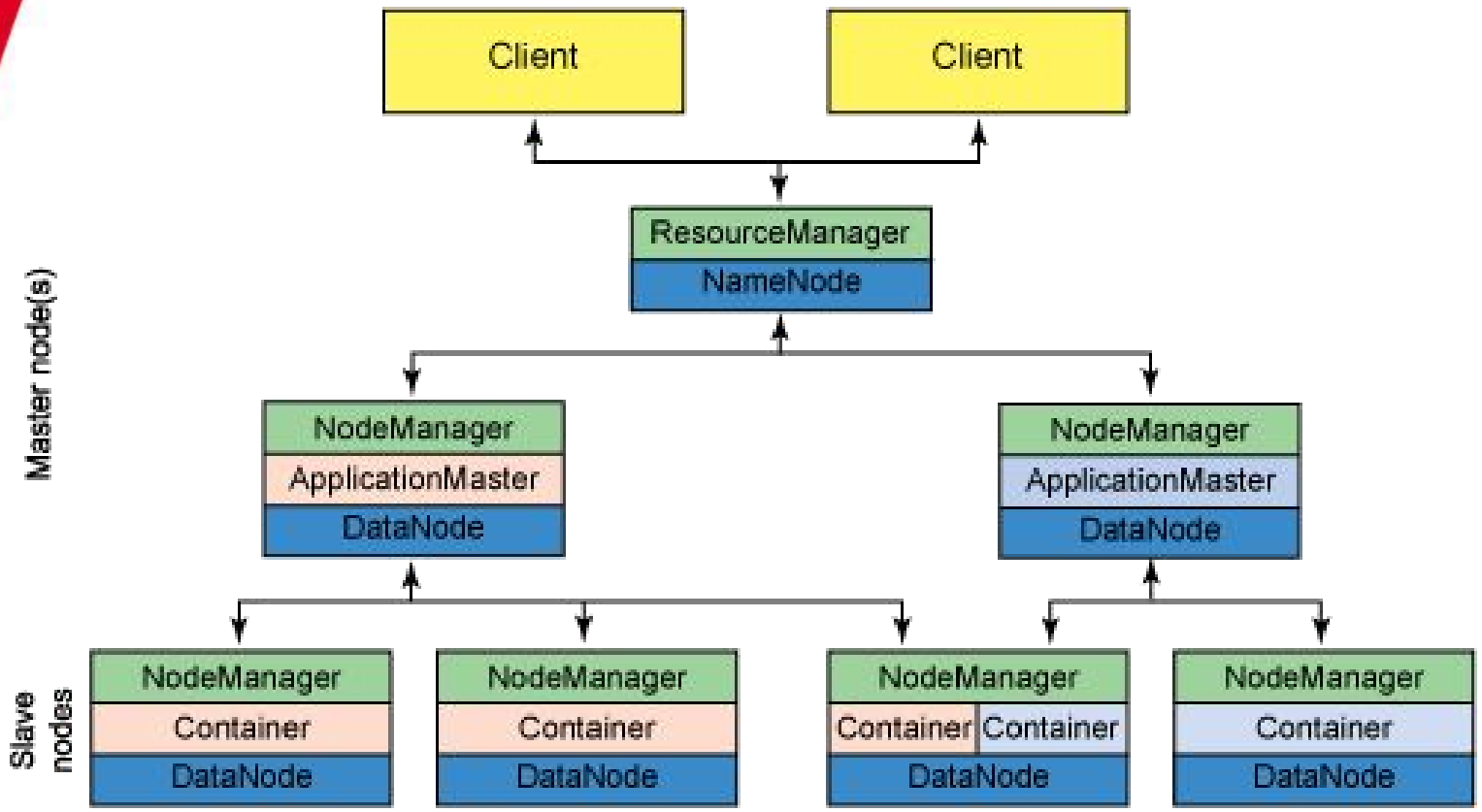
UNIVERSITY OF Cincinnati

YARN - Architecture

ResourceManager - Manages the compute resources in the cluster, receives work from clients

NodeManager - Manages workload within an execution node in the cluster. Instantiates containers to execute workloads, distributes to other NodeManagers on-demand

YARN - Architecture



UNIVERSITY OF Cincinnati