The Hadoop Ecosystem:

So much free stuff!

OUTCOMES

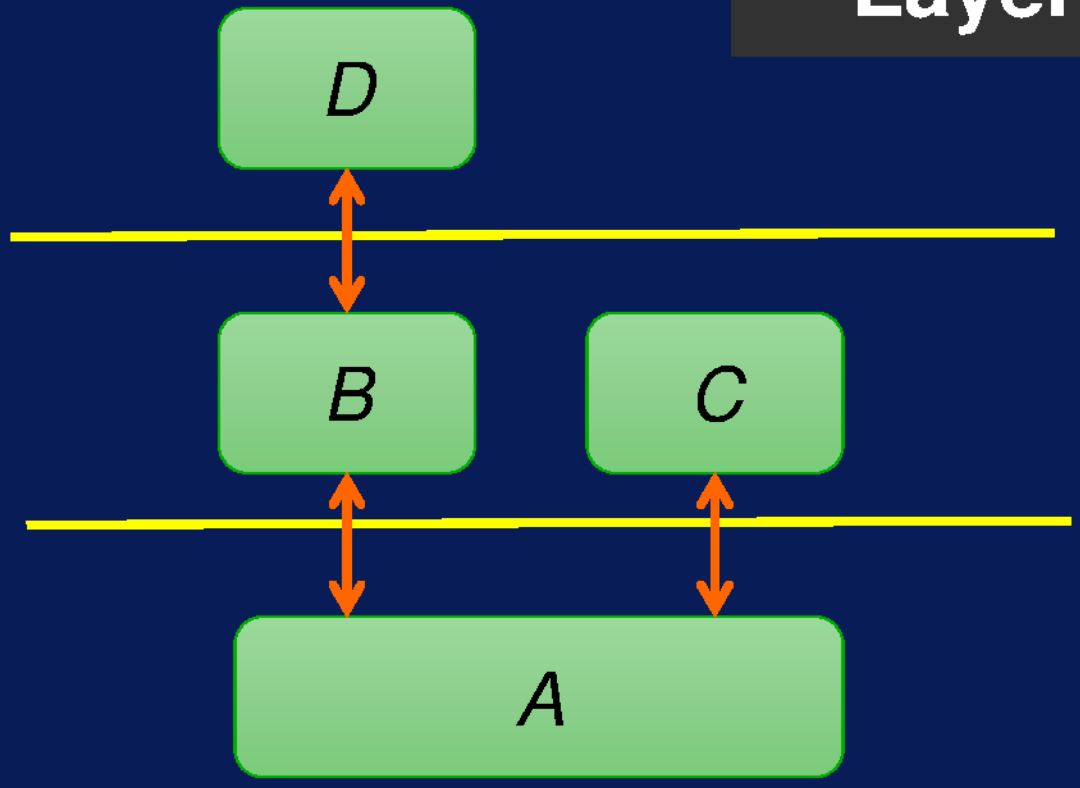
- Differentiate the major layers in the Hadoop ecosystem
- Recognize key tools of the Hadoop ecosystem including HDFS, YARN, and MapReduce
- Outline how YARN provides flexible resource management for Hadoop cluster
- Explain how YARN extends Hadoop to enable multiple frameworks such as MapReduce, Giraph, Spark and Flink

Yahoo created
Hadoop in 2005

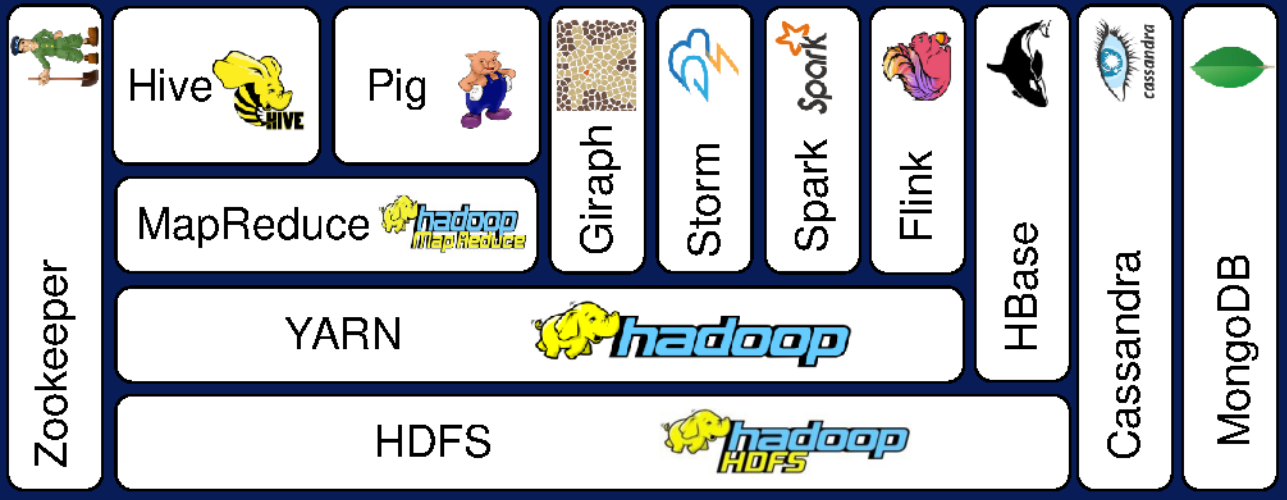
More Big Data frameworks released


Parquet
kafka
Apache
Solr
OOZIE
AVRO
mahout
Now there's over a 100!
Spark
HIVE
APACHE
HBASE
APACHE
GIRAPH
hadoop
cassandra

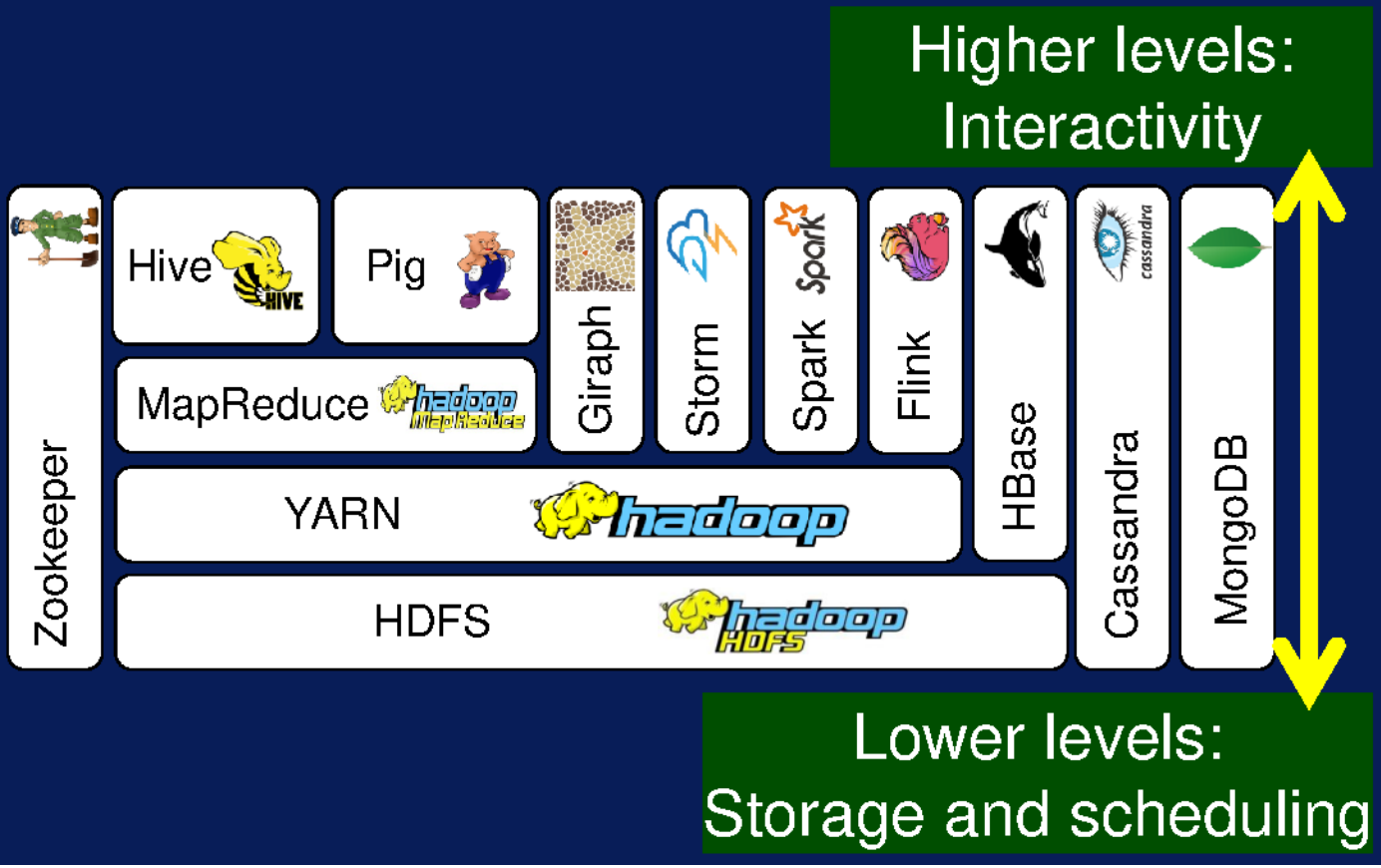
Layer Diagram



One possible layer diagram for Hadoop



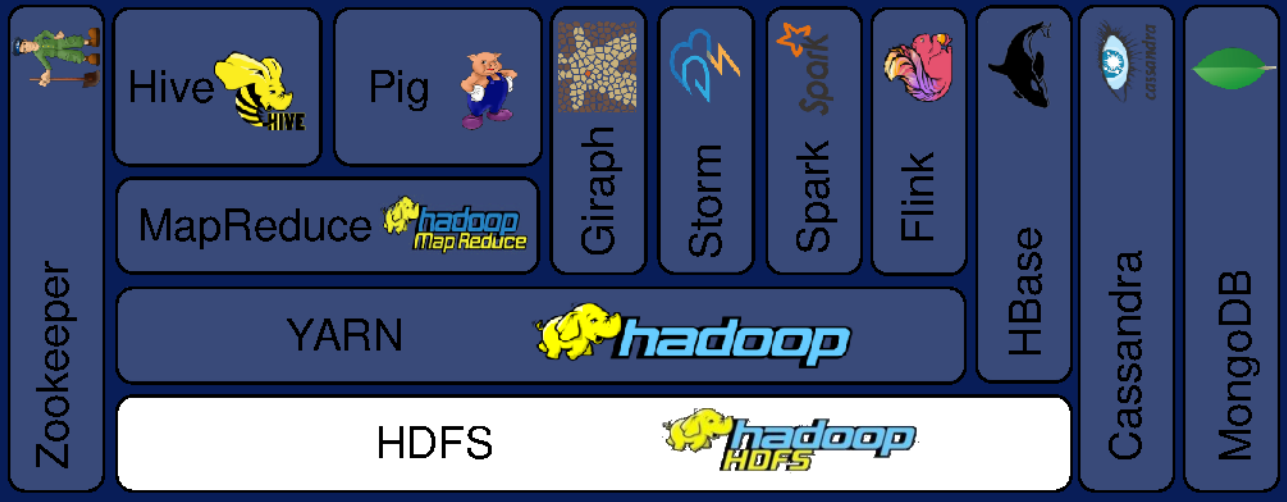
One possible layer diagram for Hadoop



Distributed file system as foundation

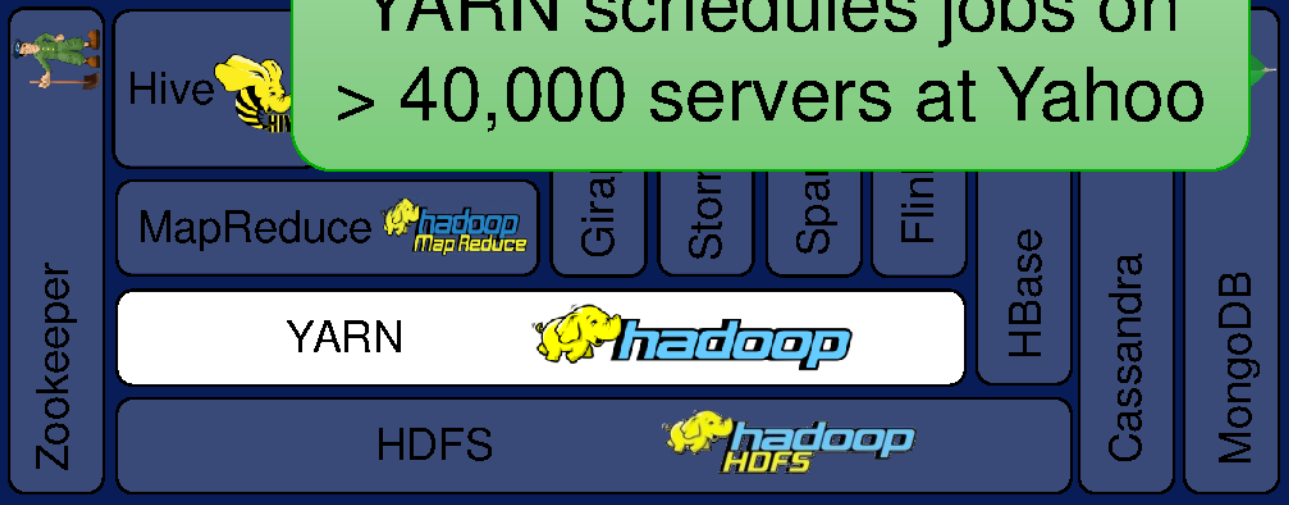
Scalable storage

Fault tolerance



Flexible scheduling and resource management

YARN schedules jobs on
> 40,000 servers at Yahoo



Simplified programming model

Map → apply()

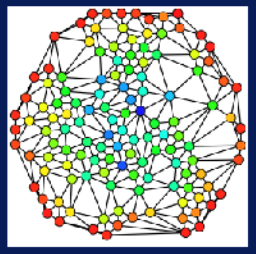
Reduce → summarize()

Higher-level programming models

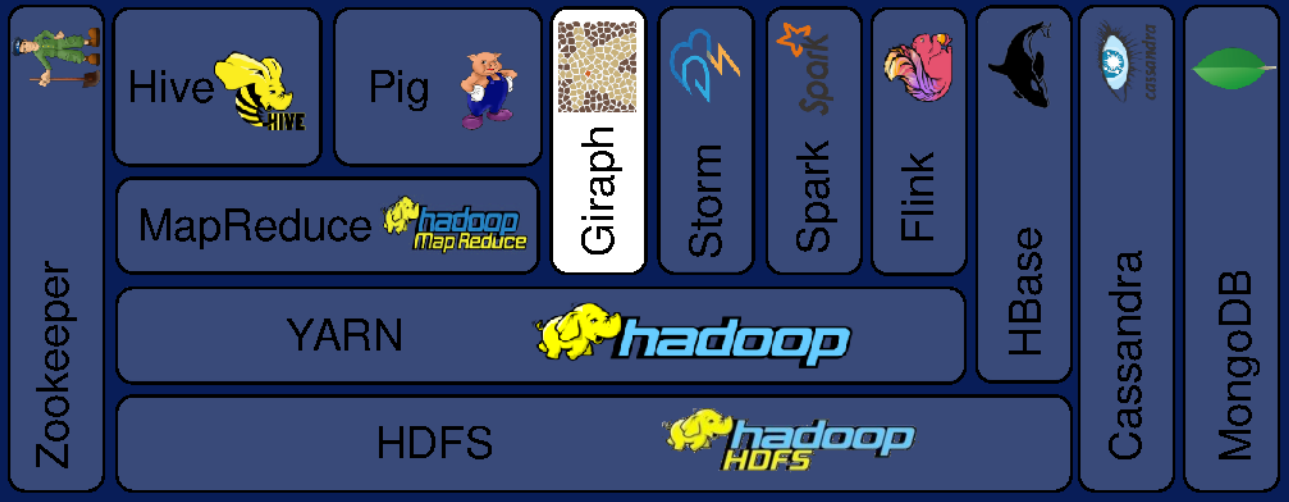
Pig = dataflow scripting

Hive = SQL-like queries

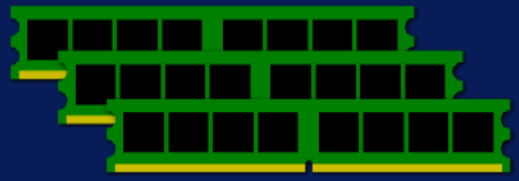
Specialized models for graph processing



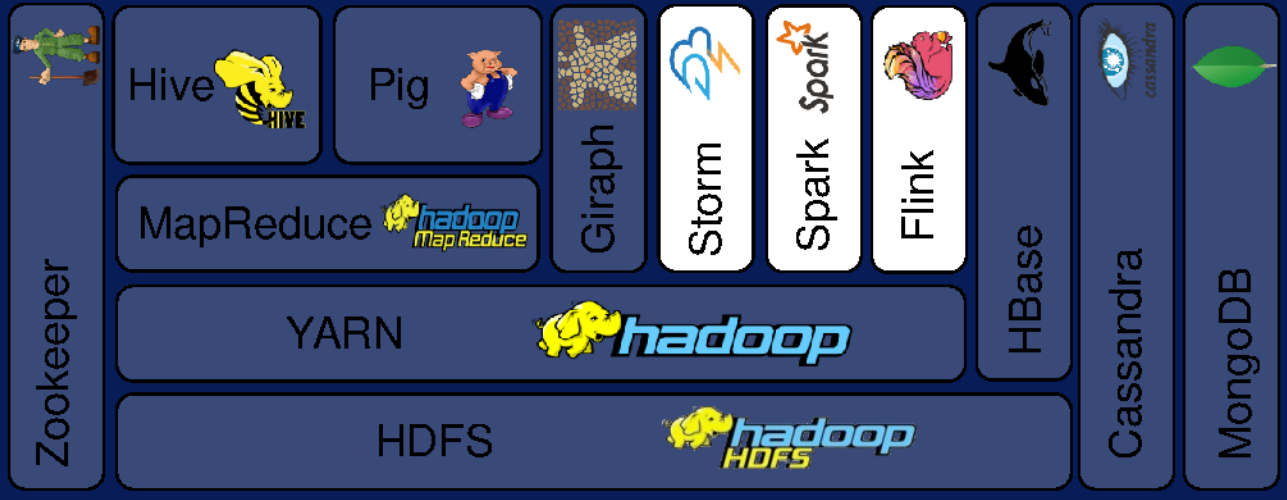
Giraph used by Facebook to analyze social graphs



Real-time and in-memory processing



In-memory → 100x faster for some tasks



NoSQL for non-files

Key-values

Sparse tables

HBase used for Facebook's Messaging Platform

Zookeeper for management

Synchronization

Configuration

High-availability

All these tools are open-source

All these tools are open-source

Large community for support

All these tools are open-source

Large community
for support

Download separately
or part of pre-built image

All these tools are open-source

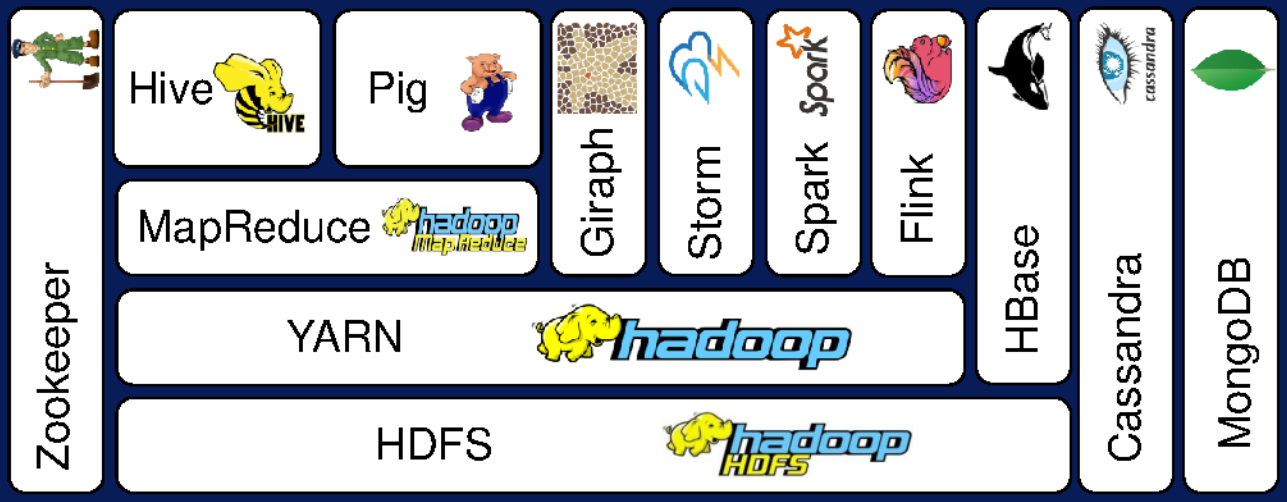
Large community
for support

Download separately
or part of pre-built image

cloudera®

MAPR®

Hortonworks



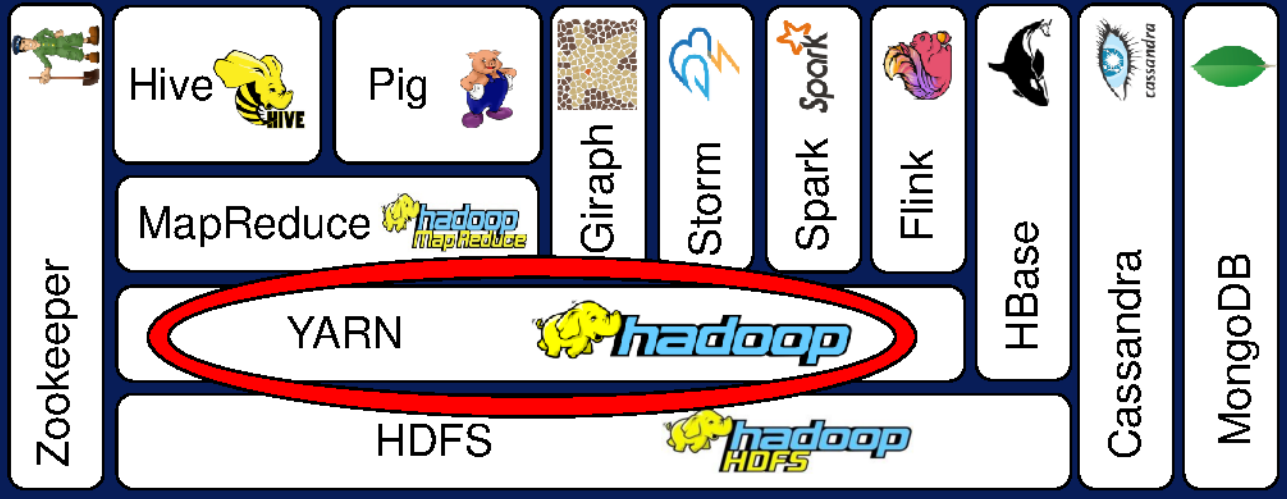
Growing number of open-source tools

YARN:

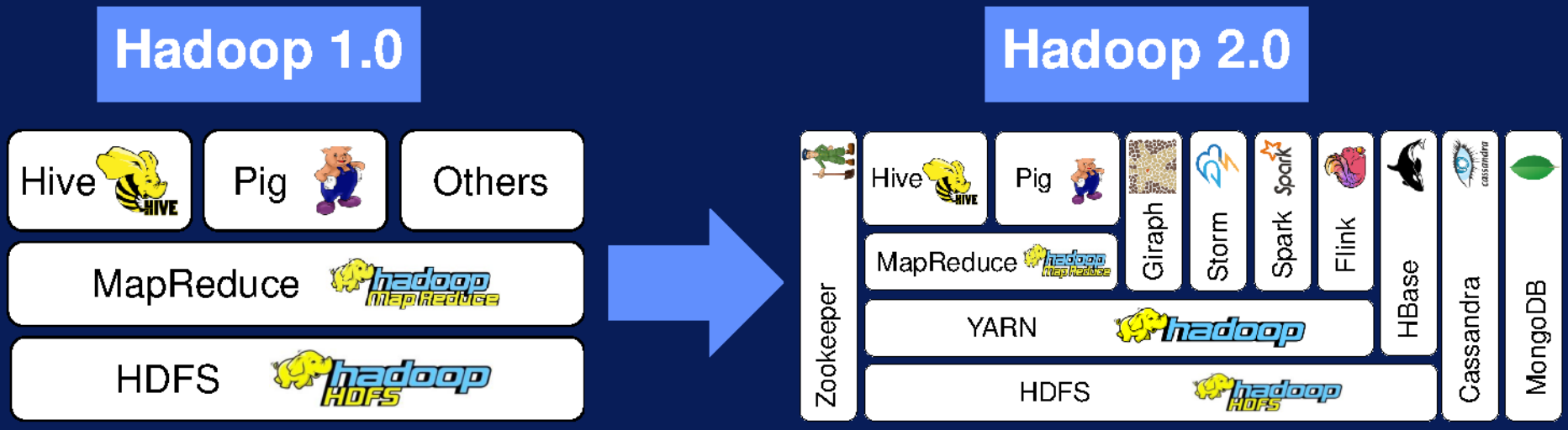
The Resource Manager for Hadoop

HDFS Cluster Utilization

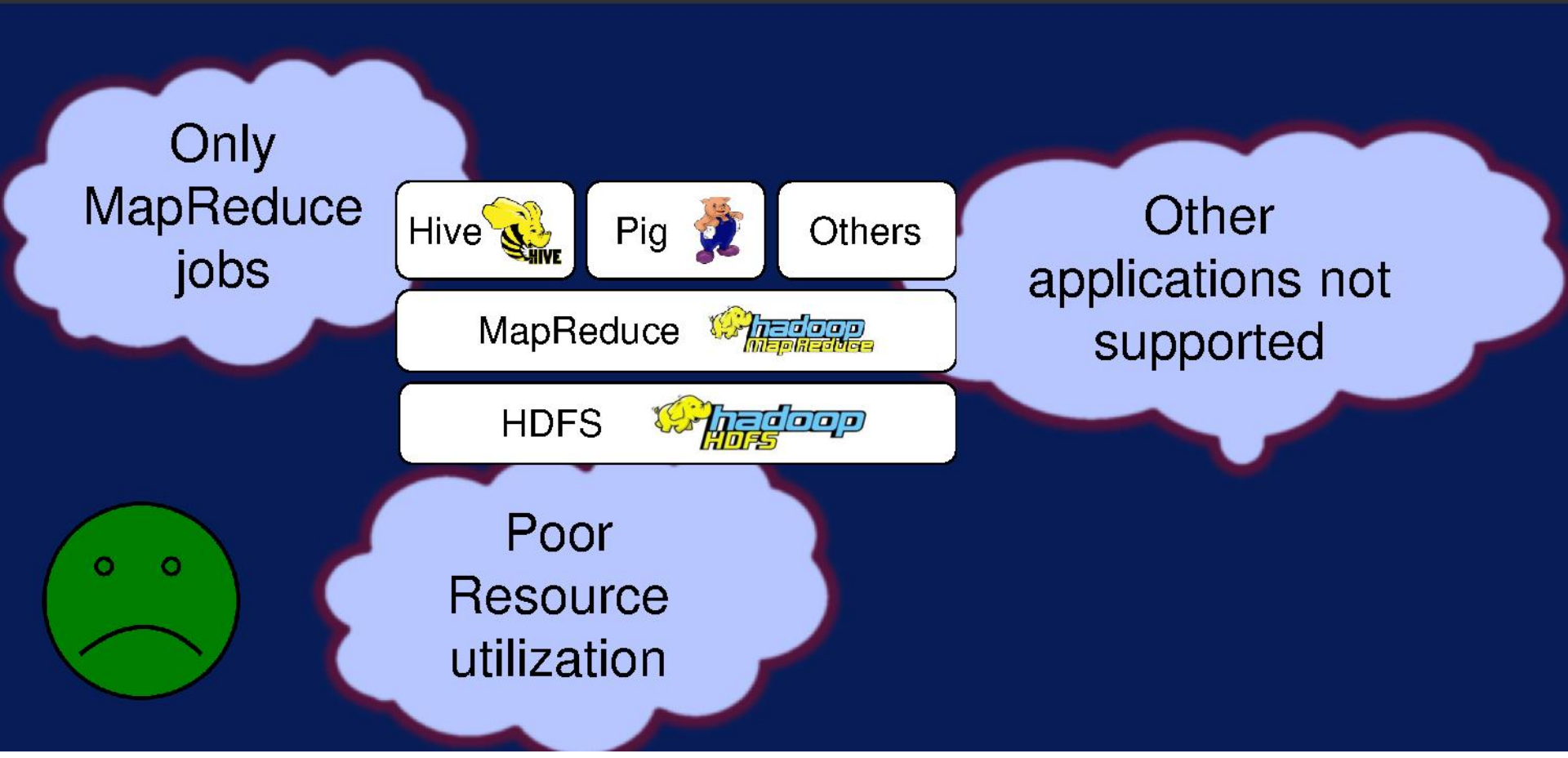
Share Hadoop across applications



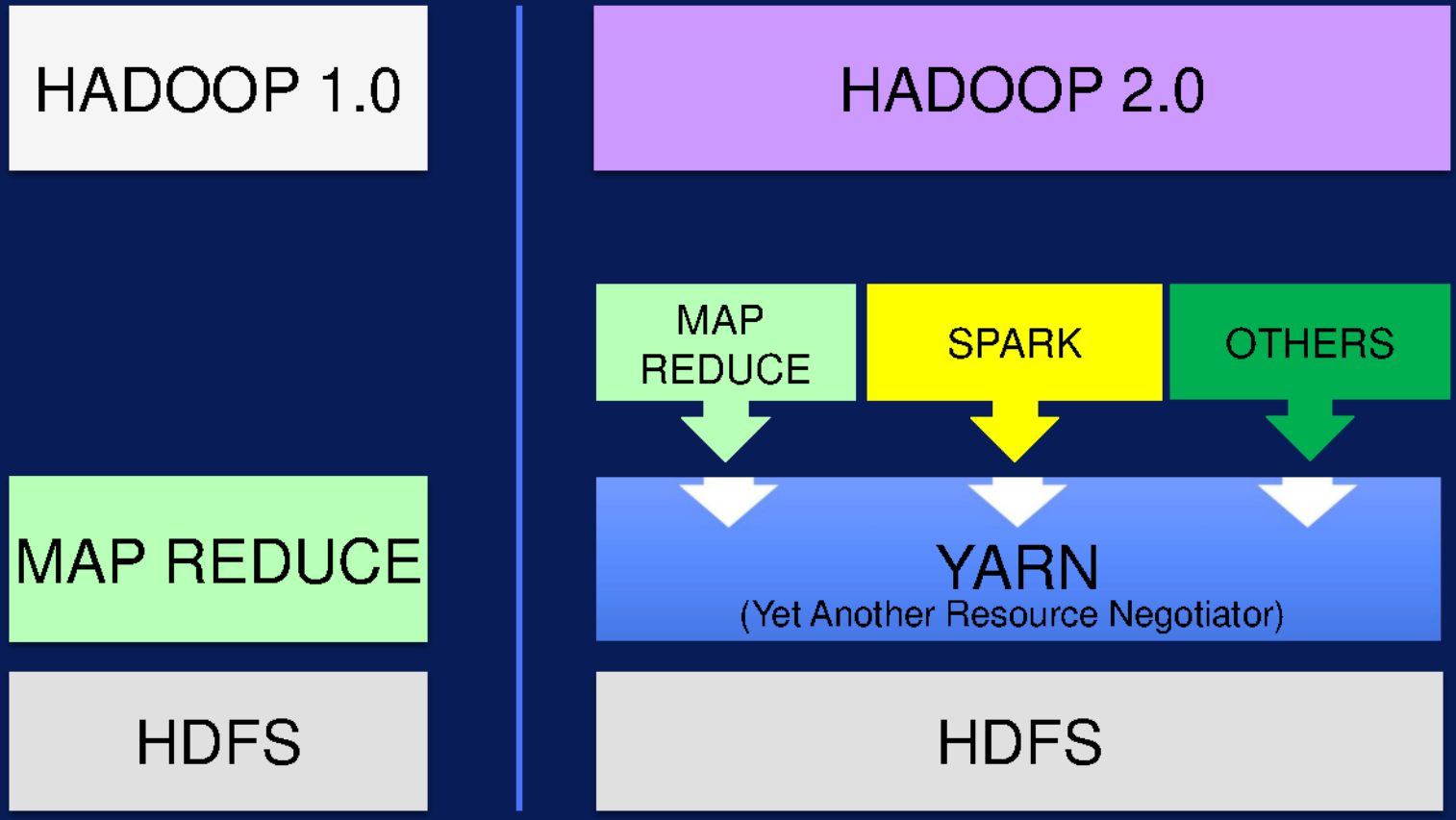
Hadoop evolved over time!



Hadoop 1.0

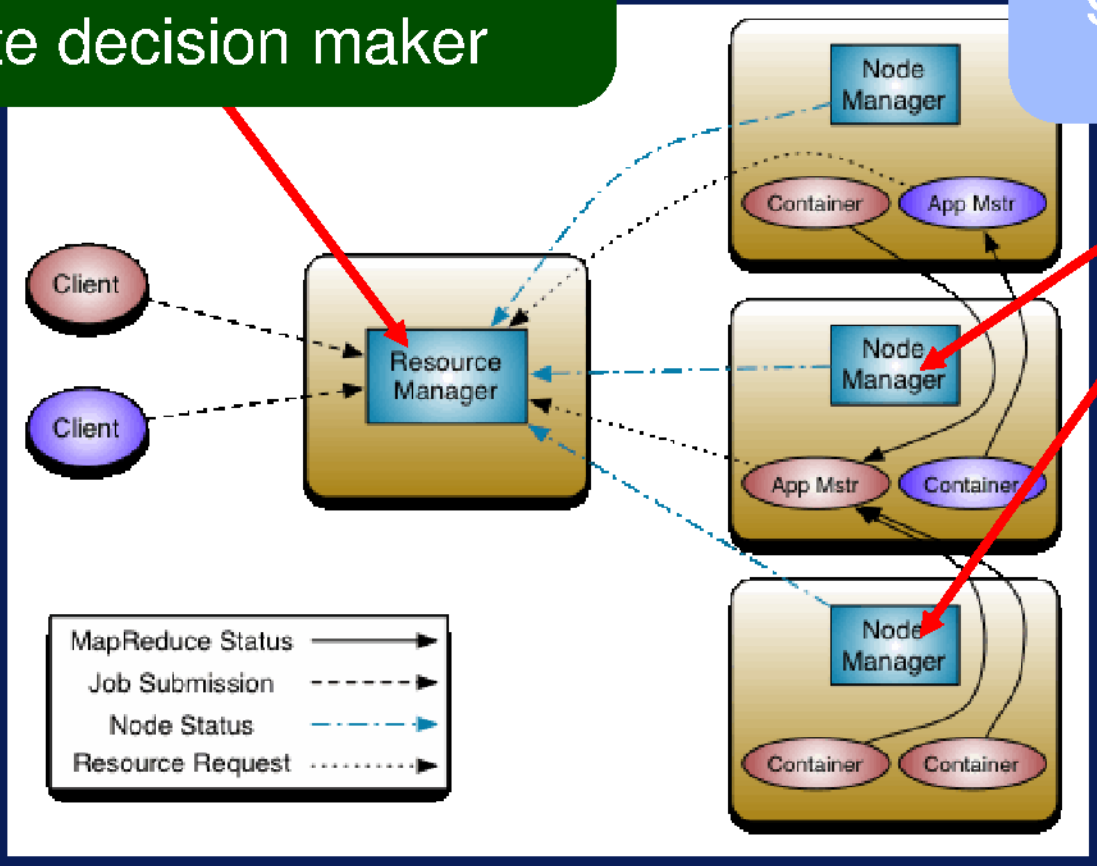

Only MapReduce jobs
Hive
Pig
Others
MapReduce
HDFS
Other applications not supported
Poor Resource utilization

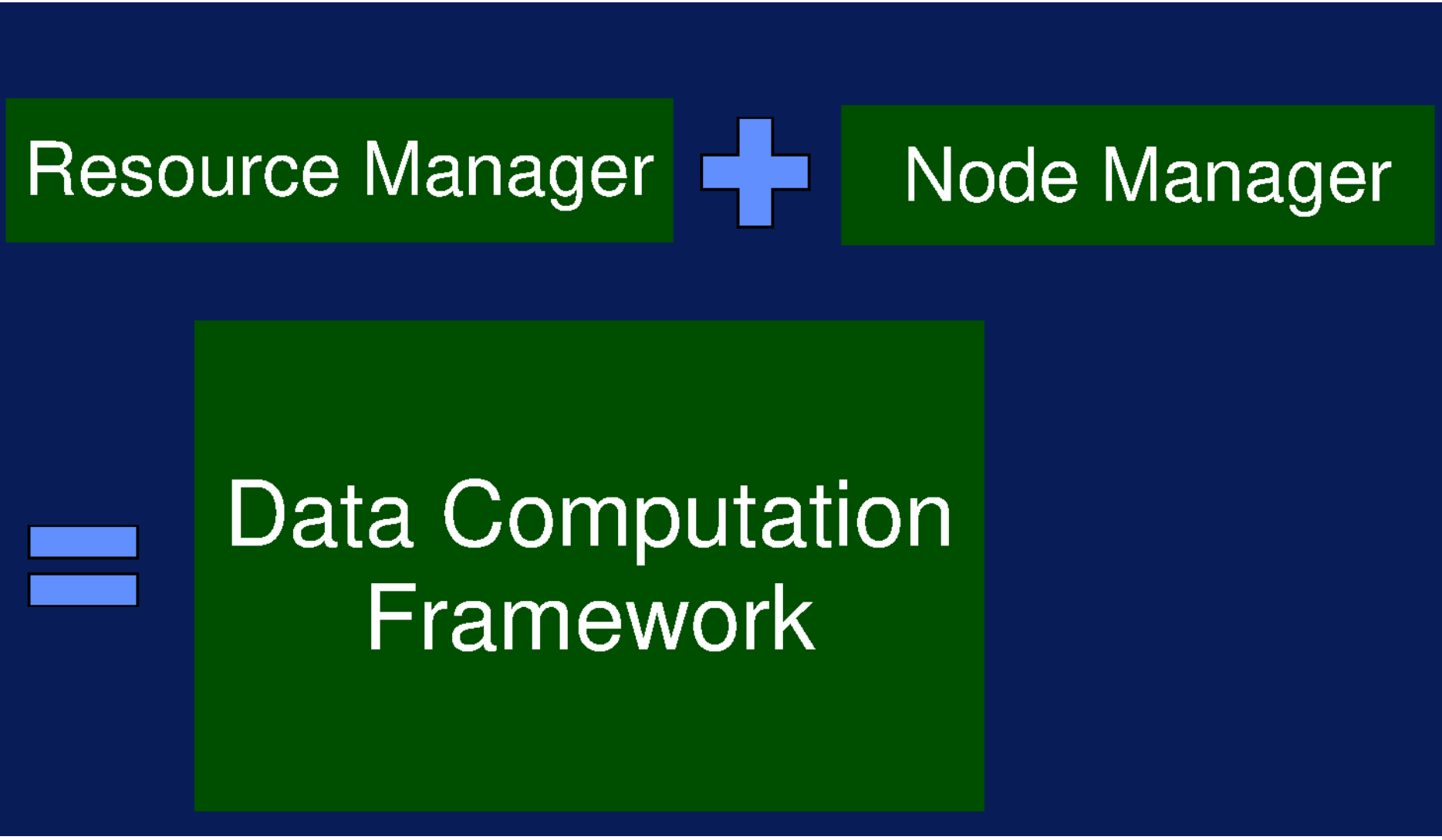
One dataset → many applications



Central Resource Manager
==
ultimate decision maker

Each machine gets a Node Manager




Resource Manager
Node Manager
Data Computation Framework

Application Master = personal negotiator

Negotiates

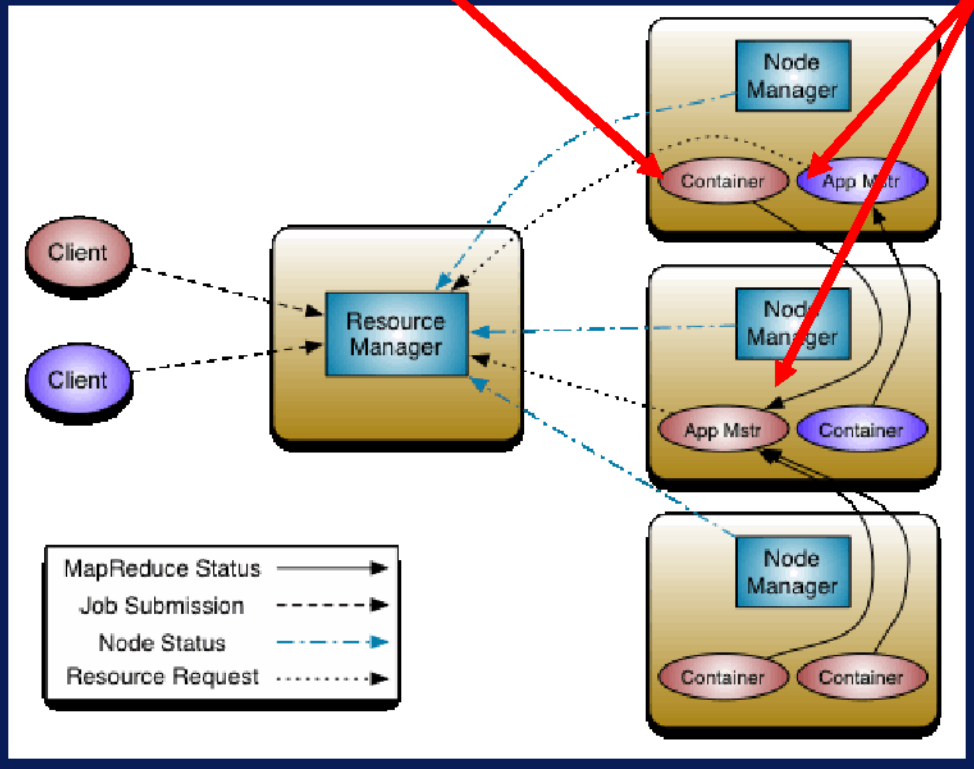
Resource Manager

Gets the job done

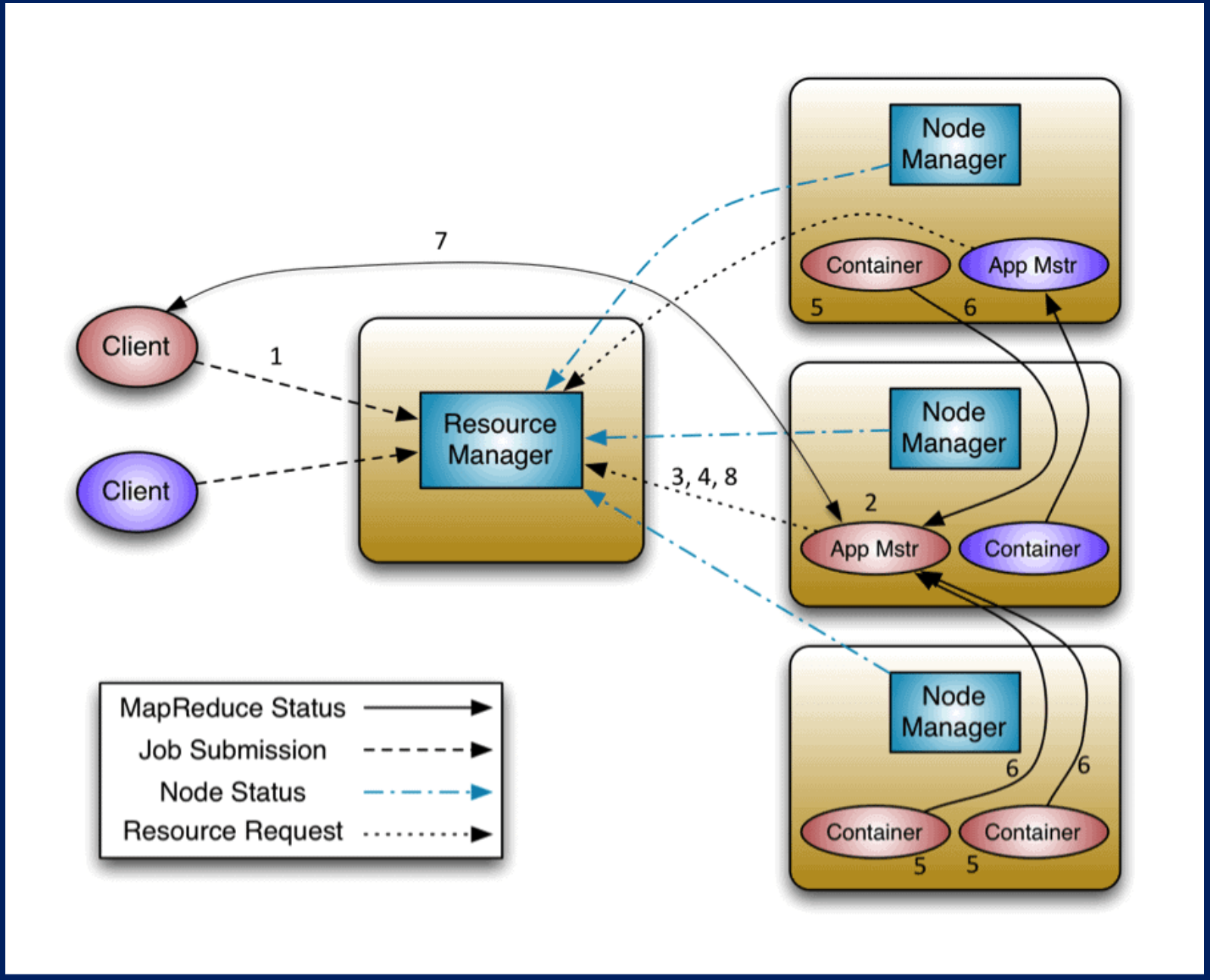
Node Manager

Container = a machine

Application Master = Personal Negotiator



An Application Execution Sequence



Essential gears in YARN engine

Resource Manager

Applications Master

Node Manager

Container

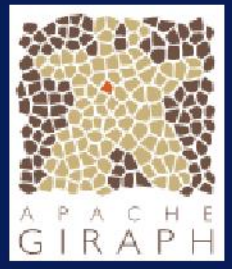
2X ↑ Jobs per day

2X ↑ CPU utilization

2.5X ↑ Number of tasks from all jobs

* Source: Apache Hadoop YARN: Yet Another Resource Negotiator." In Proceedings of the 4th Annual Symposium on Cloud Computing, 5:1–5:16. SOCC '13.

YARN → More Applications



and growing …

Data → Value

Many choices in Hadoop 2.0

One dataset → Many applications

Higher Resource Utilization → Lower Cost