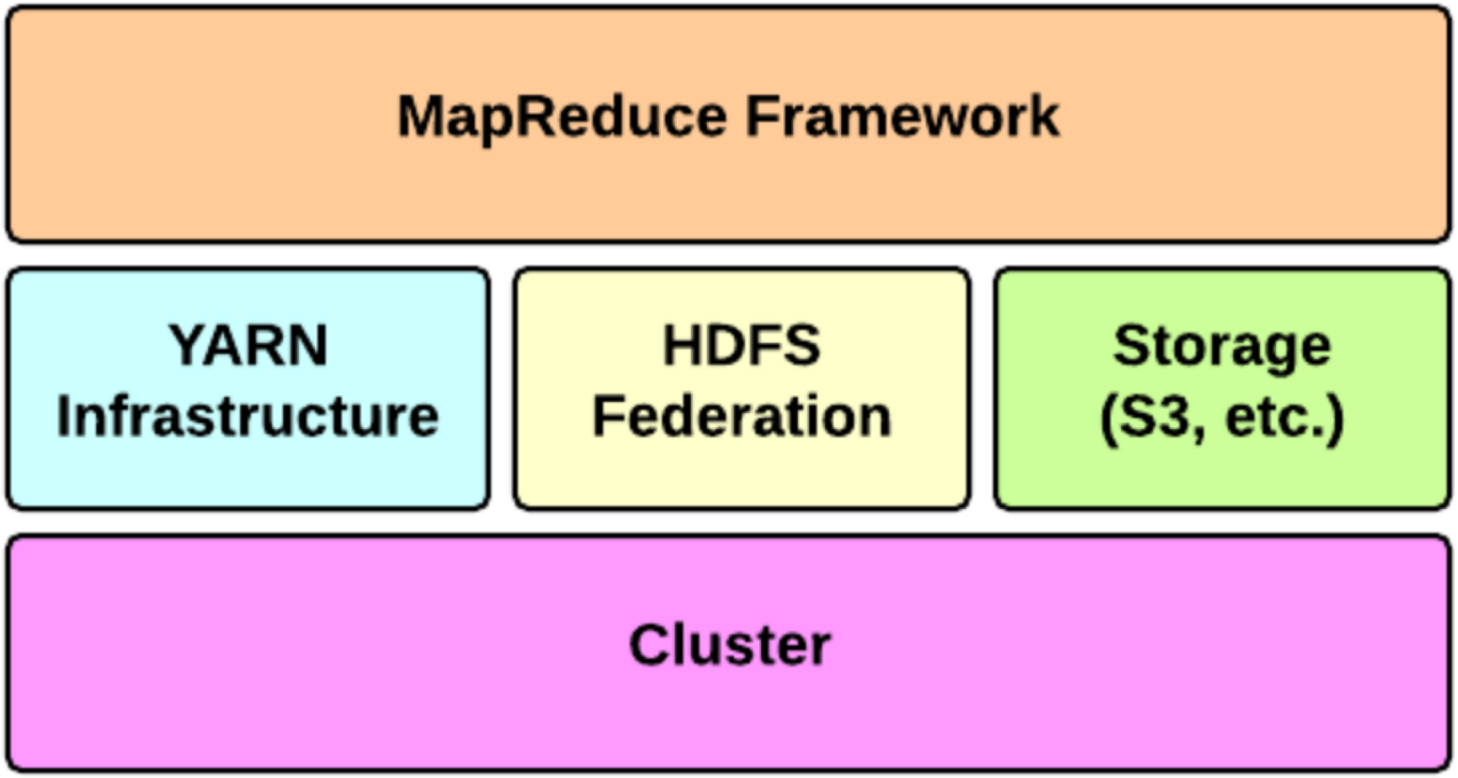
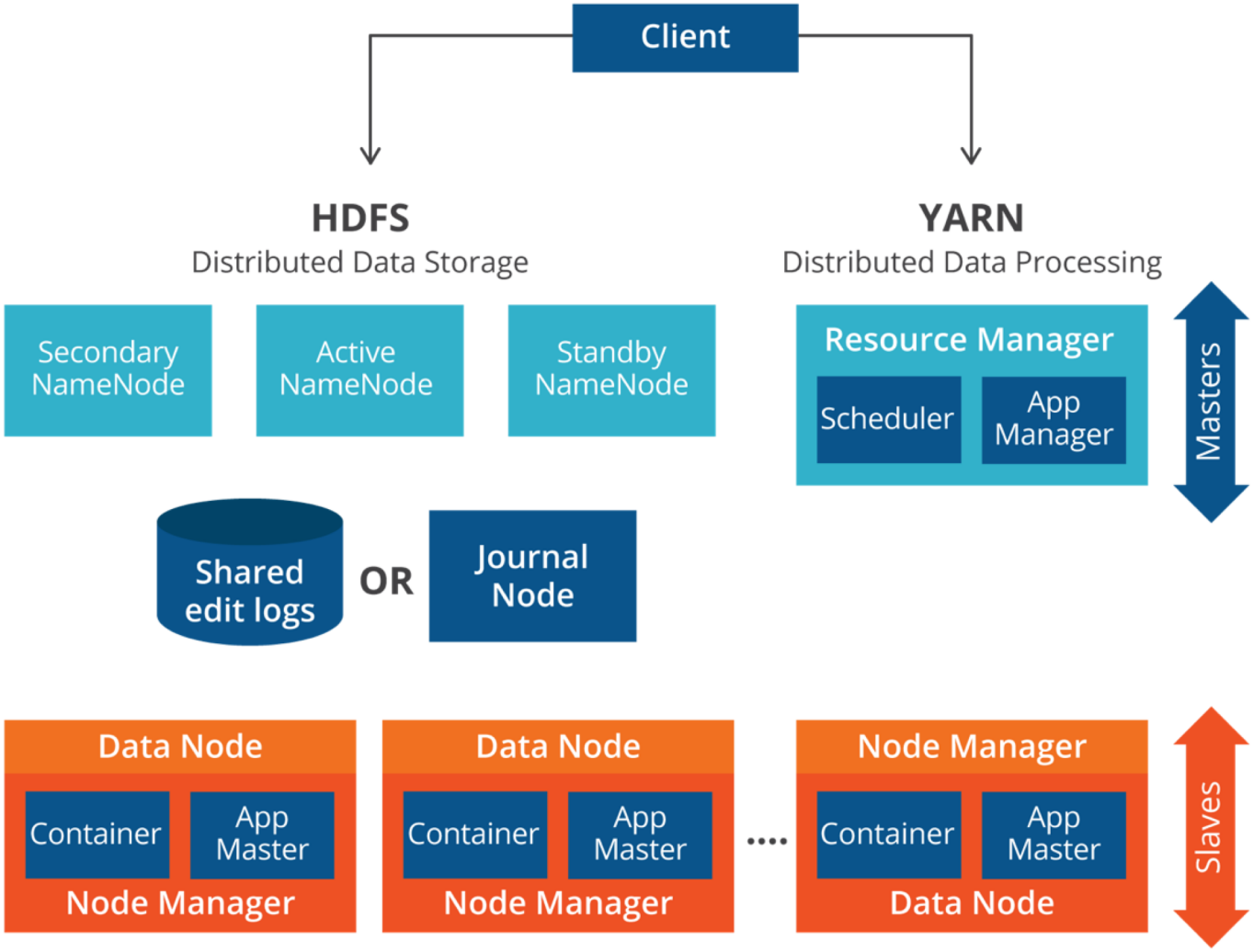
Section 4 – Hadoop MapReduce

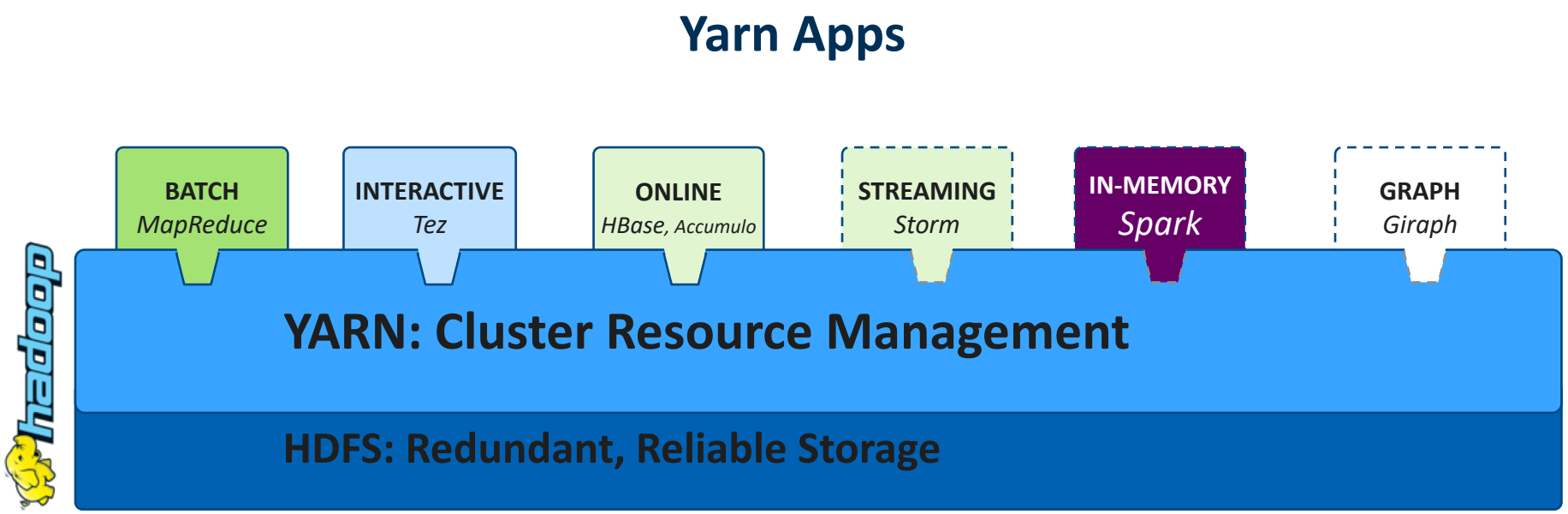
Hadoop Architecture



Apache Hadoop 2.0 and YARN



YARN – Yet Another Resource Negotiator



Cloudera Implementation

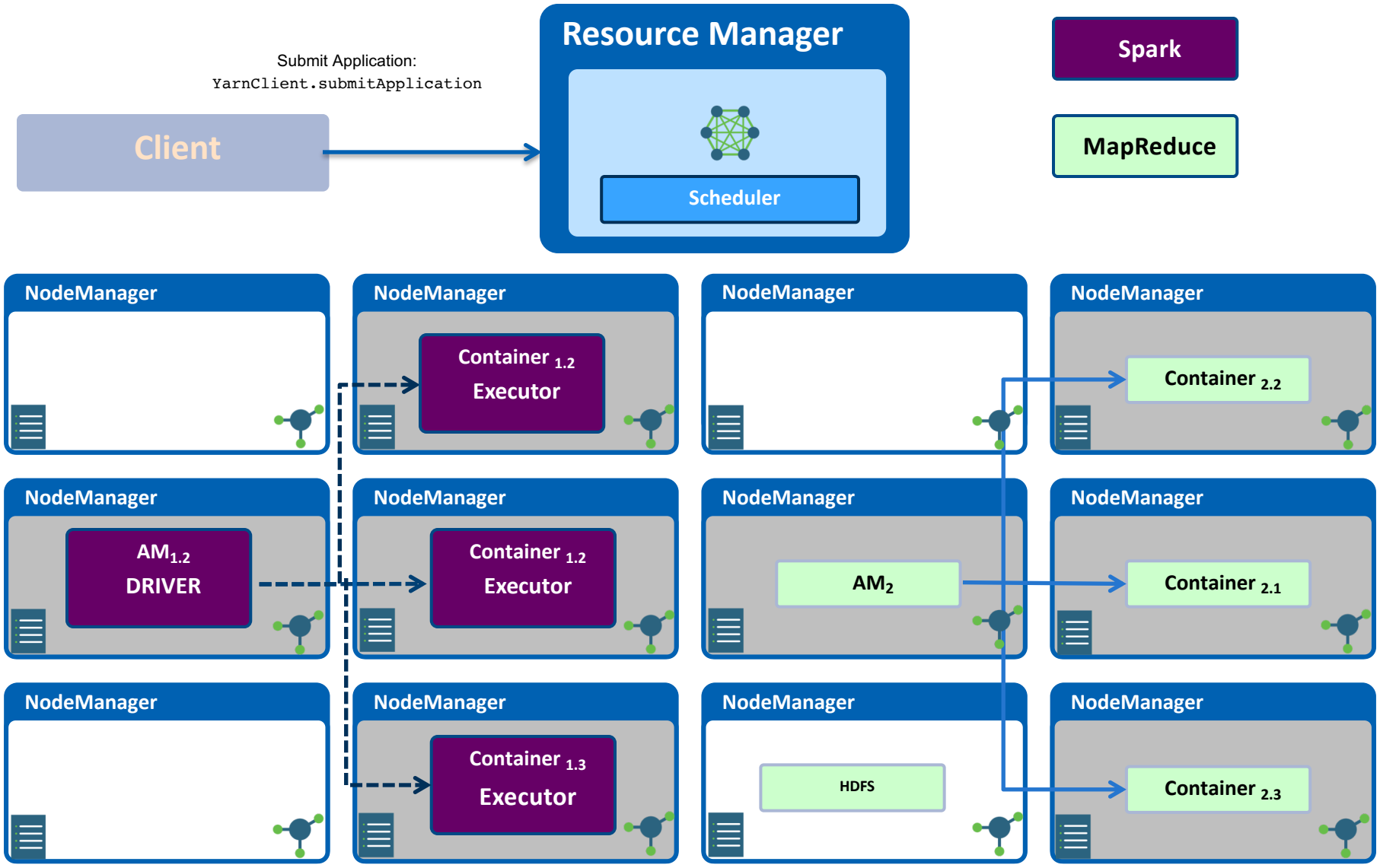
Master Nodes
ResourceManager
NameNode
Standby NameNode
JournalNodes
ZooKeeper

Worker Nodes
HDFS DataNode
YARN NodeManager
HBase RegionServer
Impala Daemons
Solr Servers

Utility Nodes
Cloudera Manager
JournalNode
ZooKeeper
Oozie
Hive Server
Impala Catalog Server
Impala State Store
Job History Server
Cloudera Management Services

Edge Nodes
Hadoop command-line client
Hive command-line client
Impala command-line client
Flume agents
Hue Server
HBase REST proxy
HBase Thrift proxy

Yarn Architecture – Cluster Mode



EMR – Component Versions

https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-release-5x.html#emr-5260-app-versions

Python - Running MR job

1. Run locally

cat demo.txt | ./hours_mapper_demo.py | sort | ./hours_reducer_demo.py

2. Run in the cluster (AWS EMR) – HDFS and S3

a) Make sure source and Data is copied to Cluster

scp -i ~/AppDevelopment/AmazonSecInfo/pem/ec2rbjamazon.pem *.txt hadoop@ec2-52-90-133-41.compute-1.amazonaws.com:

b) Make sure Code is copied

scp -i ~/AppDevelopment/AmazonSecInfo/pem/ec2rbjamazon.pem *.py hadoop@ec2-52-90-133-41.compute-1.amazonaws.com:

Python - Running MR job

Data can be in S3 or HDFS

a). Make the Input and O/P HDFS directories and Copy Data to HDFS

```
hadoop fs –mkdir Inputfiles
hadoop fs –put *.csv Inputfiles
```

b). Run the command on Cluster

```
hadoop jar /usr/lib/hadoop/hadoop-streaming.jar -file ./hours_mapper_demo.py -mapper ./hours_mapper_demo.py -file ./hours_reducer_demo.py -reducer ./hours_reducer_demo.py -input Inputfiles/demo.txt -output MyOutputfiles
```

c). Validate O/P was created

```
hadoop fs –ls MyOutputfiles
```

d). Copy to MasterNode

```
hadoop fs –copyToLocal MyOutputfiles .
```

Note for S3:
-input S3://Inputfiles/demo.txt
-output S3://Outputfiles/count-output.txt

Java Env Setup

1. Run locally
   - pom.xml or gradle build - add dependencies - make sure to include hadoop-client.jar
   - Run from Eclipse

2. Run in the cluster (AWS EMR)
   - Build JAR in Eclipse or from command line
   - hadoop jar count_jobs.jar org.cscie88.hadoopmr.HoursCounter input_logs output_hr

Thank you.

Questions?