Hdoop command:

$ hdfs dfs –copyFromLocal input.txt /WordCountFiles

$ hadoop jar ./WordCount.jar WordCount /WordCountFiles/input.txt /WordCountFiles/output

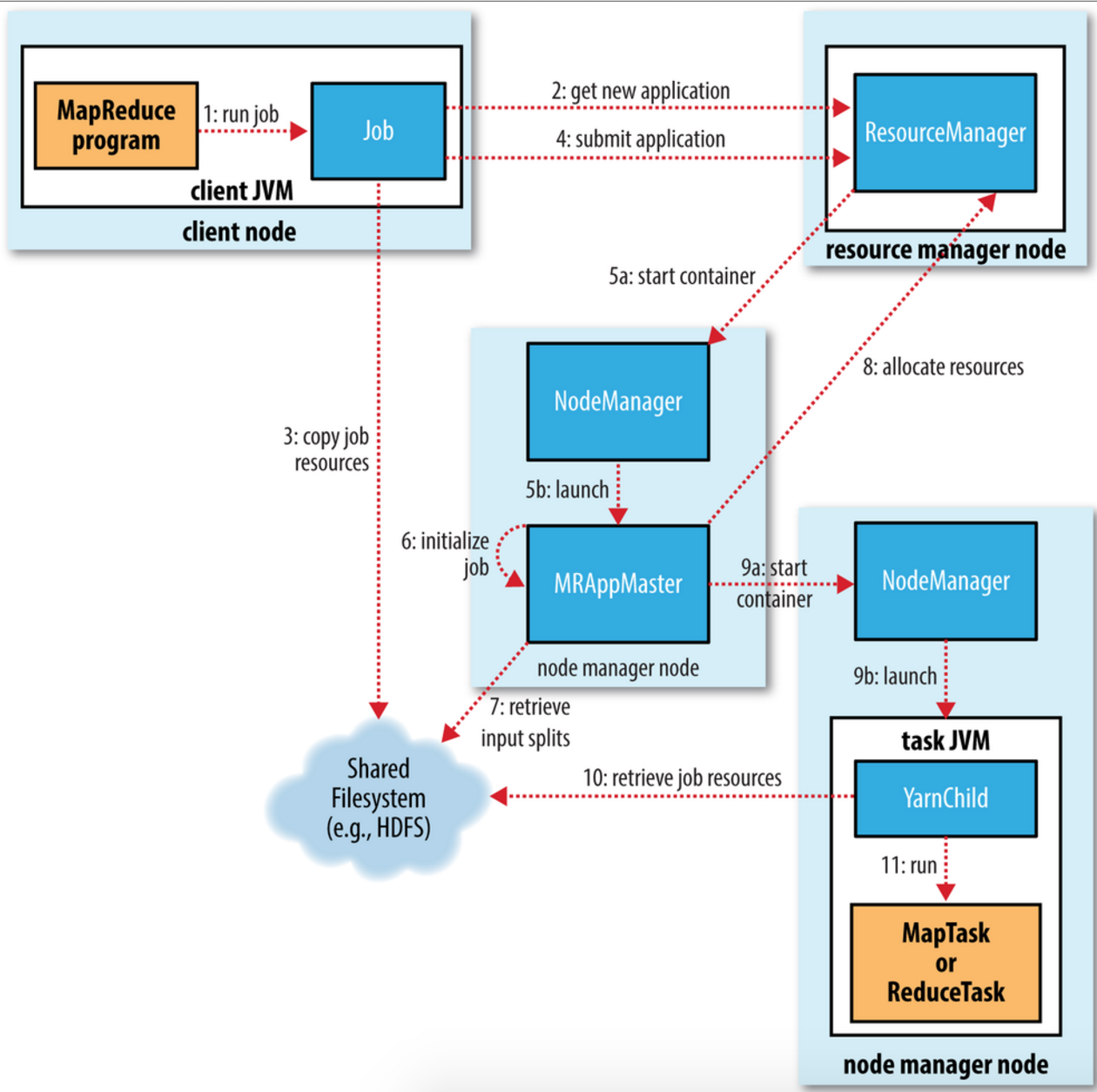
Application Scene:

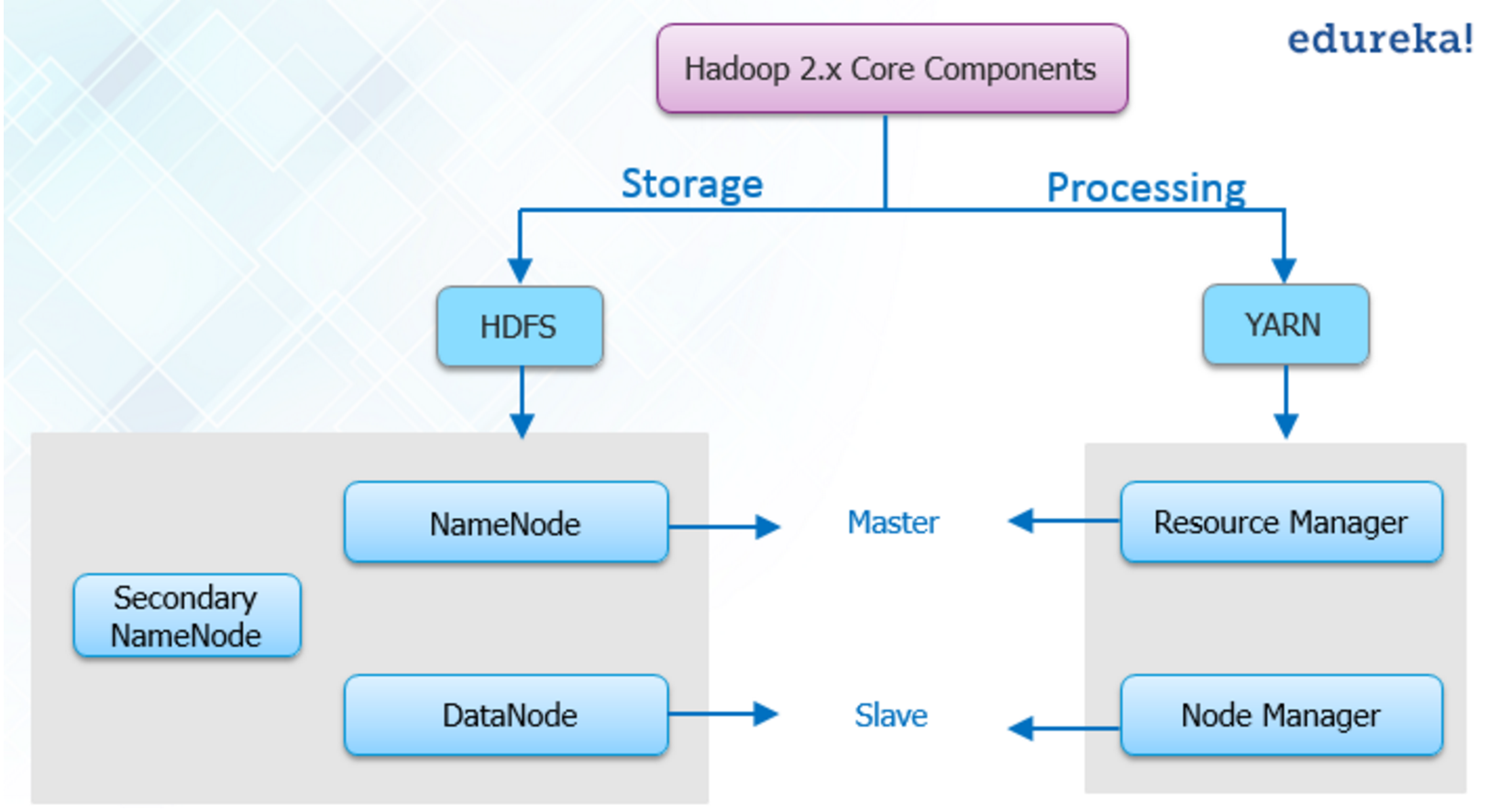
* OTA
* Mobile Data
* E-Commerce
* Energy Mining
* Save Energy
* Base Framework Regulation
* Image Processing
* Cheating inspection
* Medical

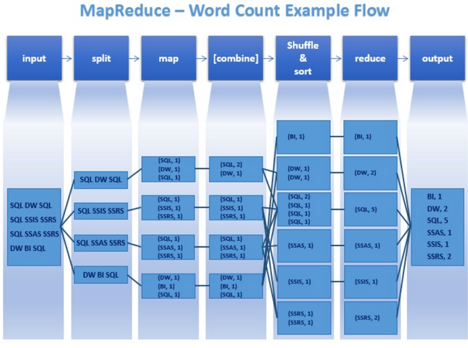
Why Hadoop

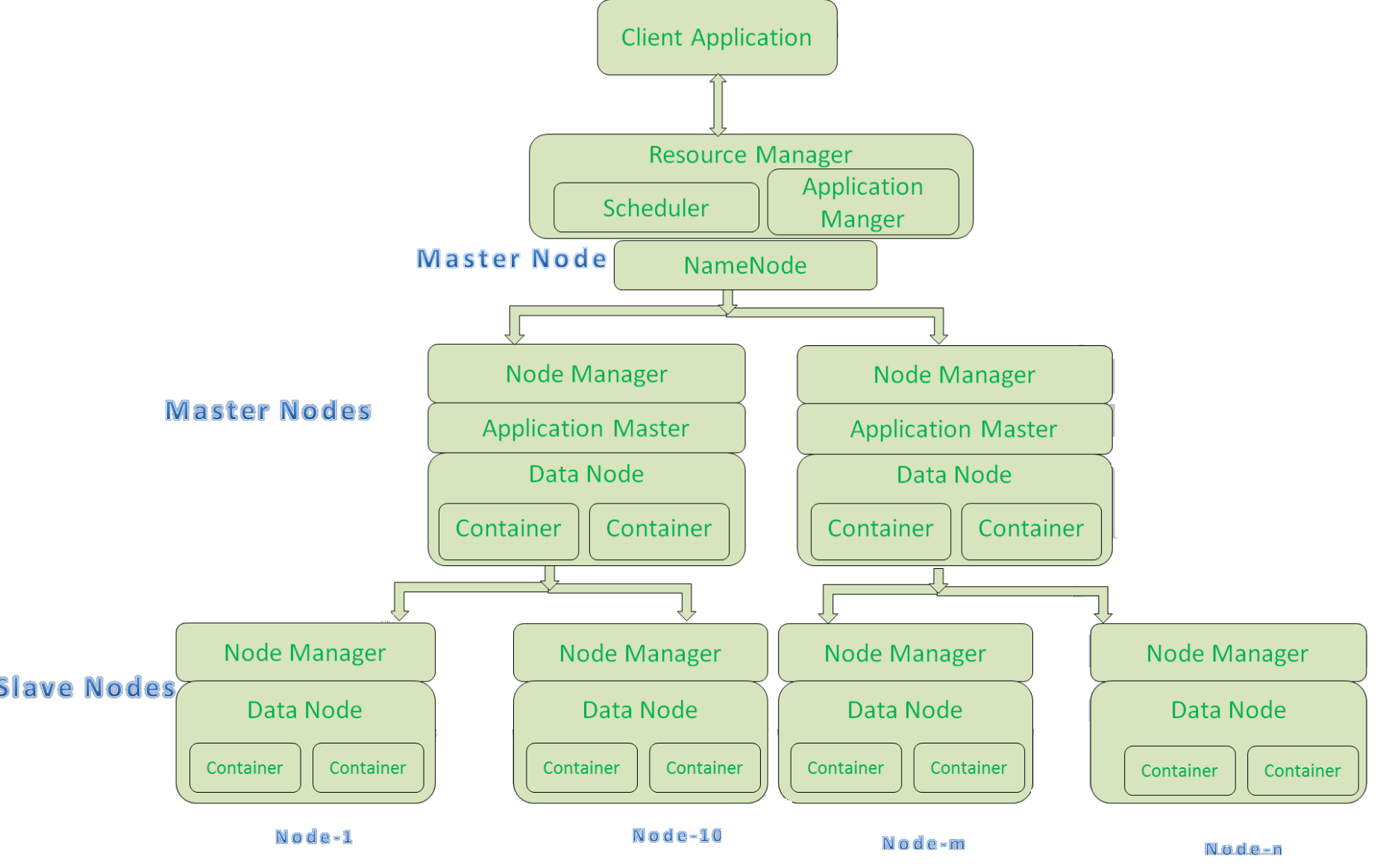
* Scalability
* Efficiency
* Flexibility

How Hadoop runs a MapReduce job









Problem Solving

* DataNode cannot run:

Plan 1:

1. Stop dfs
2. Open hdfs-site.xml
3. Remove the data.dir and name.dir properties from hdfs-site.xml and -format namenode again.
4. Then remove the hadoopdata directory and add the data.dir and name.dir in hdfs-site.xml and again format namenode.
5. Then start dfs again.

**Plan 2:**

* 1. Delete dfs/data
* DataNode cannot show in webpage:

1. Check localhost

Hadoop Cluster Installation

Master: Ubuntu14.04, 4G, 40G

Slave01: Ubuntu14.04, 4G, 40G

Slave02: Ubuntu14.04, 4G, 40G

Slave03: Ubuntu14.04, 4G, 40G

1. without password login node

base on master node:

ssh-keygen –t rsa –p

base on slave node:

mkdir ~/.ssh

copy to all slaves:

cat ~/id\_rsa.pub ~/.ssh/authorized\_keys

scp /home/hadoop/.ssh/ authorized\_keys hadoop@slave01.node:~/

1. install java environment

<http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html>

download linux x64 jdk-8u91-linux-x64.tar.gz and export

modify environment variable

sudo mkdir /usr/java/jdk1.8

sudo mv jdk1.8.0\_91/ /usr/java/jdk1.8

sudo nano /etc/profile

export JAVA\_HOME=/usr/java/jdk1.8

export CLASSPATH=.:$JAVA\_HOME/lib:$JAVA\_HOME/jre/lib:$CLASSPATH

export PATH=$JAVA\_HOME/bin:$JAVA\_HOME/jre/bin:$PATH:$HADOOP\_HOME/bin:$HADOOP\_HOME/sbin

export HADOOP\_HOME=/usr/hadoop-2.6.4

source /etc/profile

java –version

1. install Hadoop

**<http://mirror.csclub.uwaterloo.ca/apache/hadoop/common/hadoop-2.6.4/hadoop-2.6.4.tar.gz>**

change dirctoory

sudo mv Hadoop-2.6.4 /usr

modify /etc/hostname file

sudo nano /etc/hostname

enter password, and eg.change username to slave02

sudo nano /etc/hosts

enter all members ip + name

modify 7 files

hadoop-env.sh

yarn-env.sh

slaves

core-site.xml

hdfs-site.xml

mapred-site.xml

yarn-site.xml

core-site.xml

   <property>  
        <name>fs.defaultFS</name>  
        <value>hdfs://master:9000</value>  
    </property>  
    <property>  
        <name>io.file.buffer.size</name>  
        <value>131072</value>  
    </property>  
    <property>  
        <name>hadoop.tmp.dir</name>  
        <value>file:/usr/ hadoop-2.6.4/tmp</value>  
        <description>Abase for other temporary directories.</description>  
    </property>  
    <property>  
        <name>hadoop.proxyuser.hduser.hosts</name>  
        <value>\*</value>  
    </property>  
    <property>  
        <name>hadoop.proxyuser.hduser.groups</name>  
        <value>\*</value>  
    </property>  
  
</configuration>  
  
 hdfs-site.xml

<configuration>  
    <property>  
        <name>dfs.namenode.secondary.http-address</name>  
        <value>master:9001</value>  
    </property>  
    <property>  
        <name>dfs.namenode.name.dir</name>  
        <value>file:/usr/hadoop-2.6.4/dfs/name</value>  
    </property>  
    <property>  
        <name>dfs.datanode.data.dir</name>  
        <value>file:/usr/ hadoop-2.6.4/dfs/data</value>  
    </property>  
    <property>  
        <name>dfs.replication</name>  
        <value>3</value>  
    </property>  
    <property>  
        <name>dfs.webhdfs.enabled</name>  
        <value>true</value>  
    </property>  
  
</configuration>  
  
mapred-site.xml

<configuration>  
    <property>  
        <name>mapreduce.framework.name</name>  
        <value>yarn</value>  
    </property>  
    <property>  
        <name>mapreduce.jobhistory.address</name>  
        <value>master:10020</value>  
    </property>  
    <property>  
        <name>mapreduce.jobhistory.webapp.address</name>  
        <value>master:19888</value>  
    </property>  
  
</configuration>  
  
yarn-site.xml

<configuration>  
    <property>  
        <name>yarn.nodemanager.aux-services</name>  
        <value>mapreduce\_shuffle</value>  
    </property>  
    <property>  
        <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>  
        <value>org.apache.hadoop.mapred.ShuffleHandler</value>  
    </property>  
    <property>  
        <name>yarn.resourcemanager.address</name>  
        <value>master:8032</value>  
    </property>  
    <property>  
        <name>yarn.resourcemanager.scheduler.address</name>  
        <value>master:8030</value>  
    </property>  
    <property>  
        <name>yarn.resourcemanager.resource-tracker.address</name>  
        <value>master:8031</value>  
    </property>  
    <property>  
        <name>yarn.resourcemanager.admin.address</name>  
        <value>master:8033</value>  
    </property>  
    <property>  
        <name>yarn.resourcemanager.webapp.address</name>  
        <value>master:8088</value>  
    </property>  
  
</configuration>

1. start service

initial namenode(only need once):

hadoop namenode –format

Start Hadoop:

./start-all.sh

jps

stop-all.sh