AWS and Hadoop/MapReduce

ECE 6363 Lab 4

LAB4 OBJECTIVES

- A peek of the **Amazon AWS** and managing VMs on cloud
- Understand the **MapReduce** concept.
- Get familiar with the Hadoop framework.
- Experience working a small Hadoop cluster with VMs.

Basics of AWS

What is AWS?

- Amazon Web Service hosts servers at their facilities.
- On-demand use & pay.
- **Availability Zones**: Contains multiple datacenters close together to run simultaneously for availability and failure-proof
- **Regions**: Divided by country/regulatory boundaries. Have different pricing and services offerings because of the regulatory discrepancies.

Basics of AWS

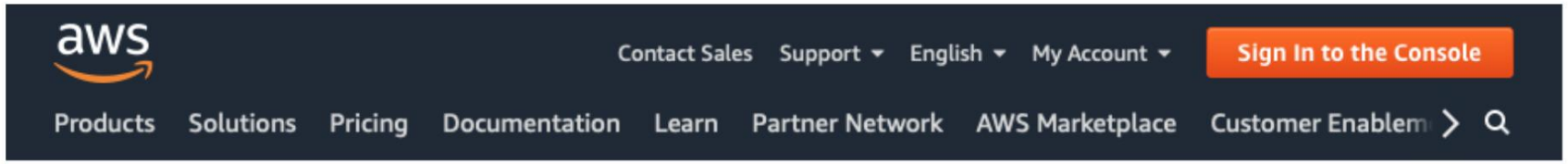
What is AWS?

Most important building blocks:

1. EC2
   Elastic Cloud Computing, Auto Scaling to automatically scale upon demand.

2. S3
   Simple Storage Service, object storage.

AWS Set-up



Create account:
Use your information to create account on AWS. Select personal account. Add your credit card information.

Sign-in to Console:
To start using your AWS services, log in on with the orange button.

AWS Set-up

Security:

Root user vs. IAM user.

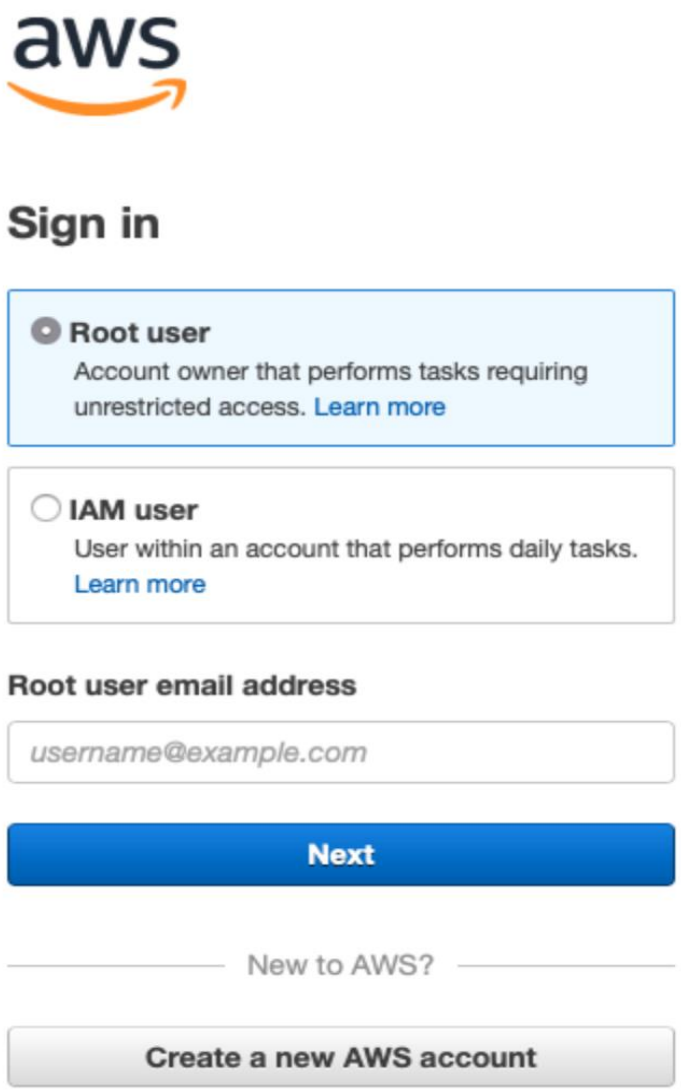
Access Key and **Secret Key** are needed for CLI

Secret key can only be shown ONCE!

- Store it properly!!

Once access key is leaked out, you need to delete and generate a new one.

Do **NOT** generate access key for root account!



AWS Set-up

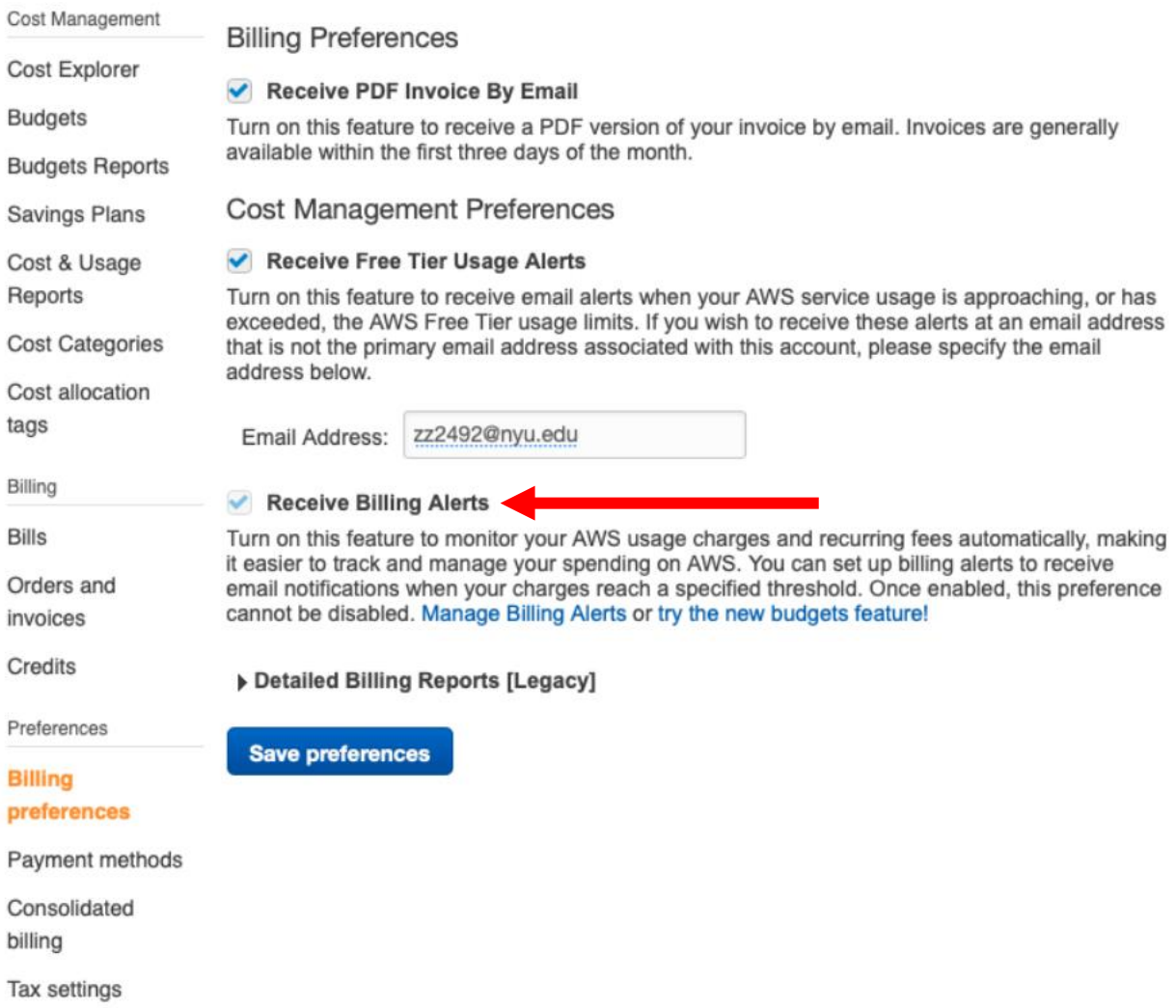
Billing:

- Free-tier use!

  https://aws.amazon.com/free/?all-free-tier.sort-by=item.additionalFields.SortRank&all-free-tier.sort-order=asc

Under Billing Preference:

- Turn on notifications and email alerts for your usage and billing communications to avoid being charged and not knowing.

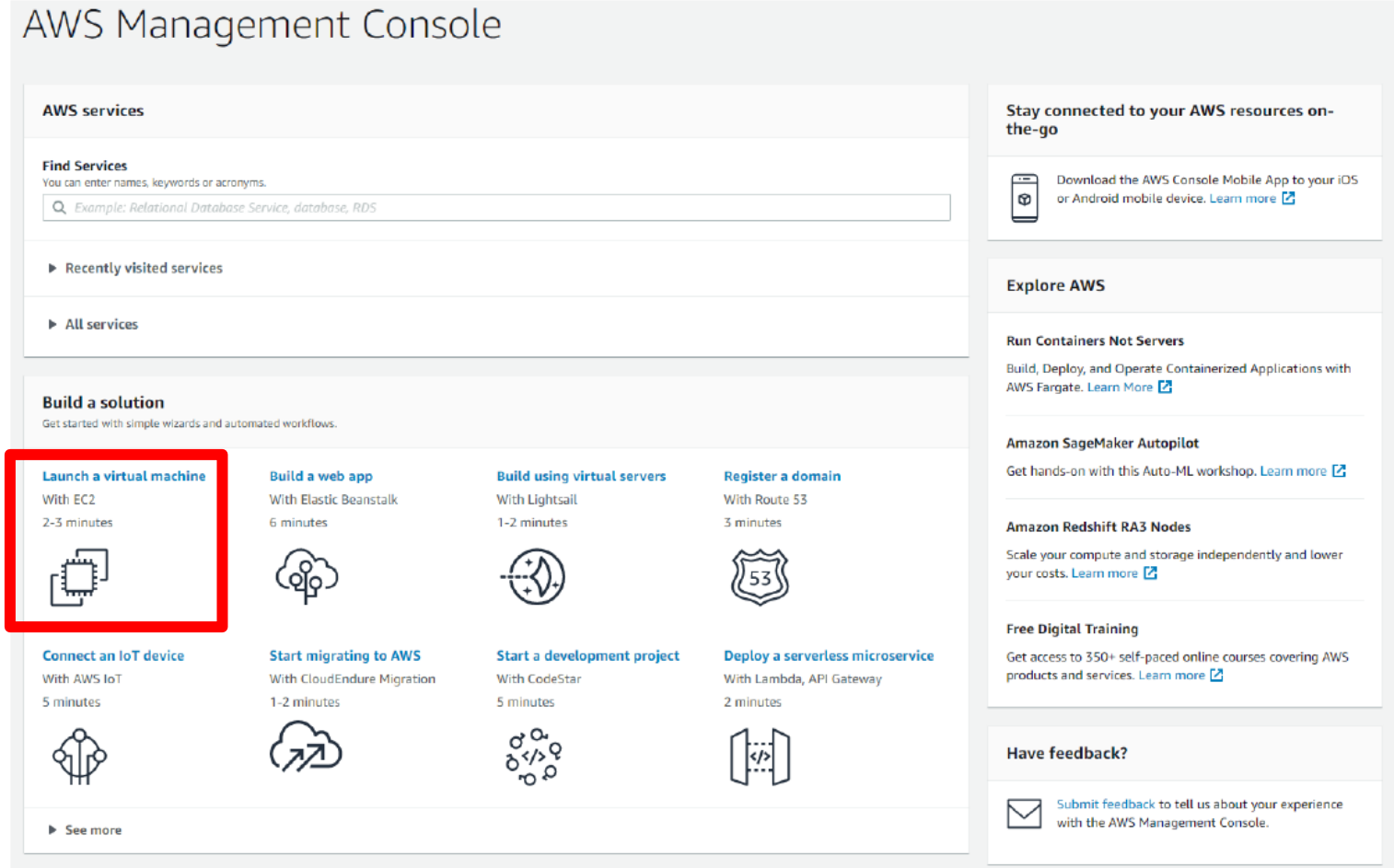


Working with AWS

Using EC2

In AWS Console

Select this to build a virtual machine

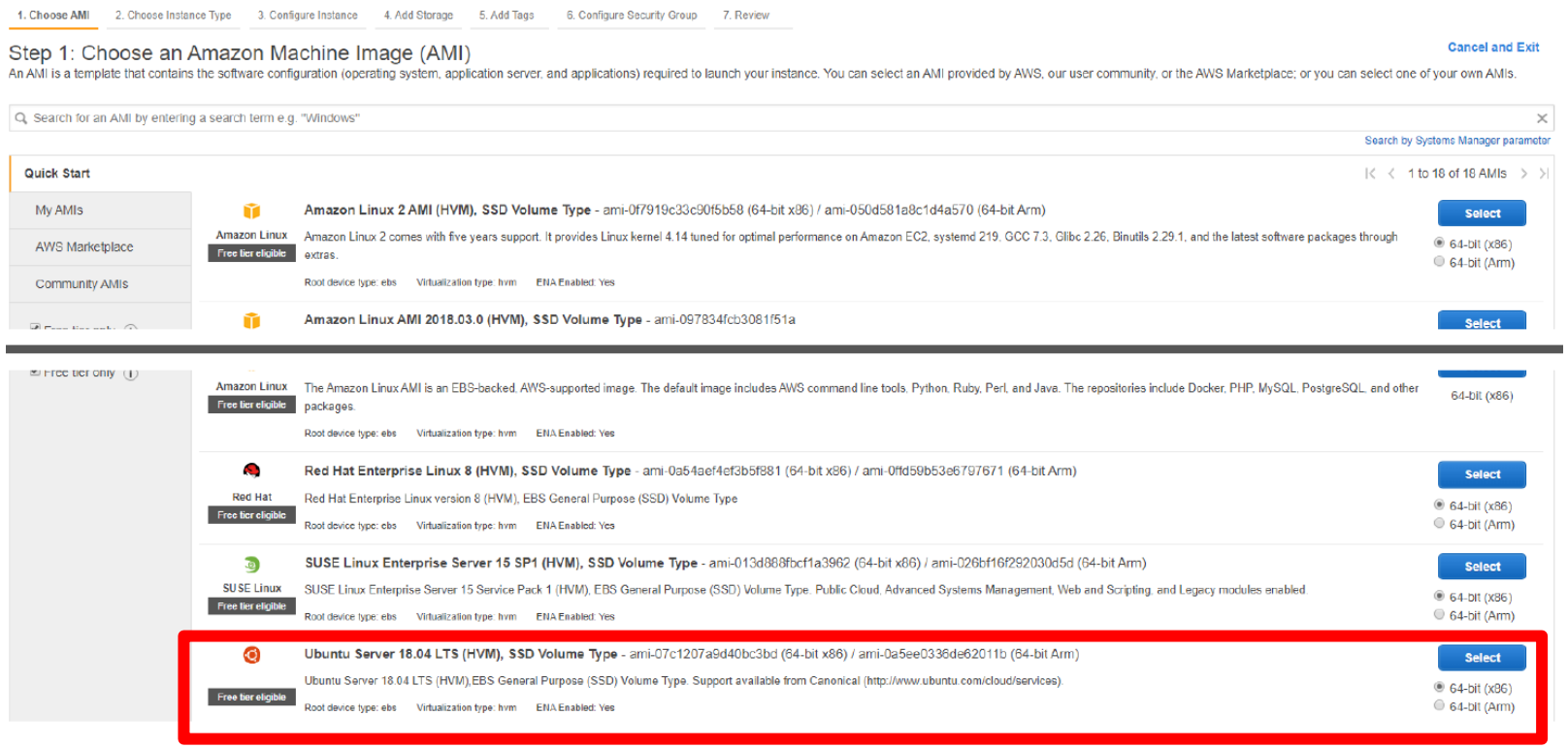


Made by Zack Luo

Working with AWS

Using EC2

In AWS Console



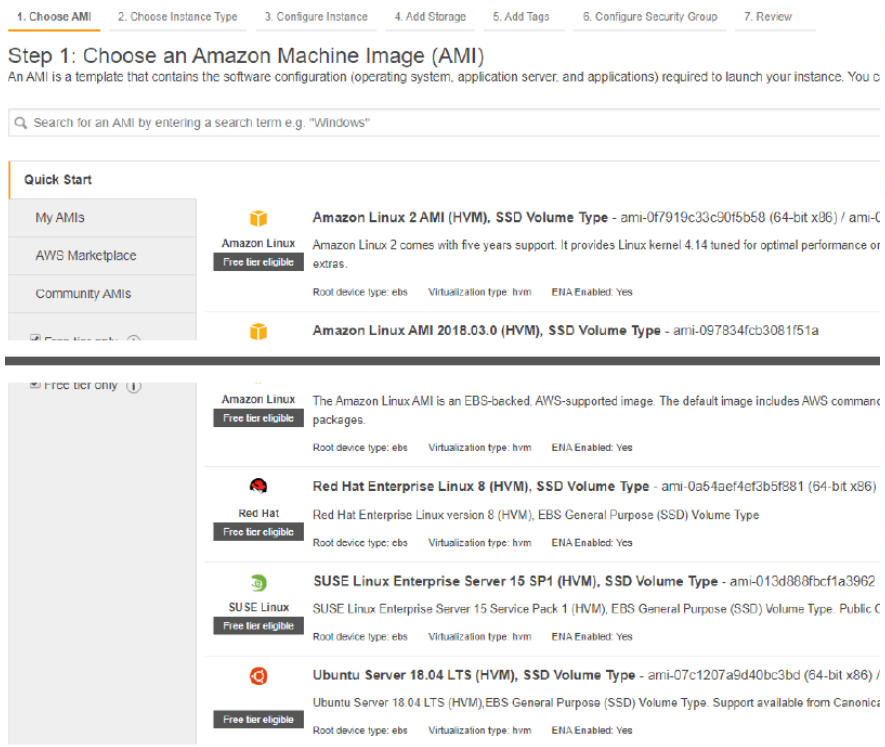
Select Ubuntu

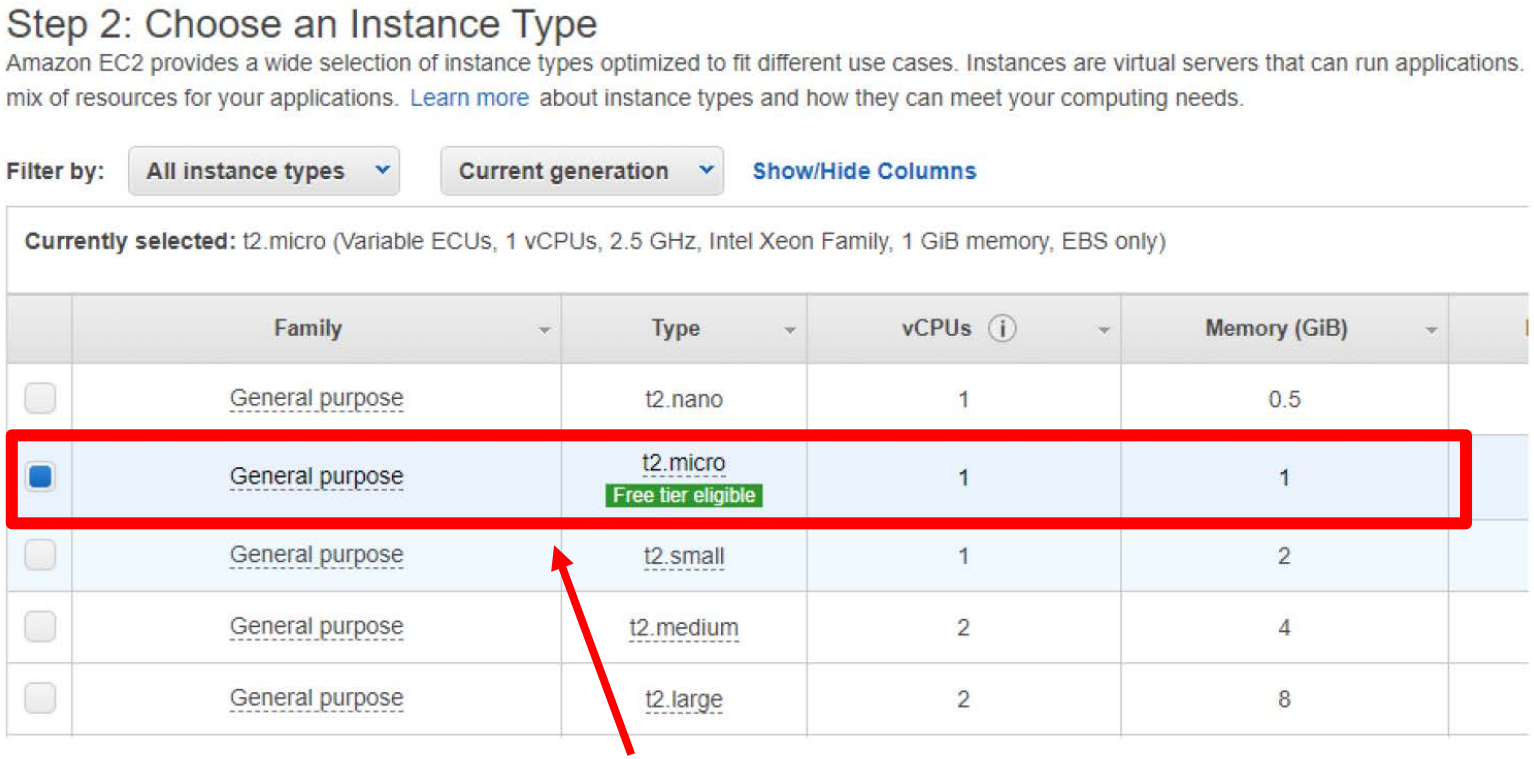
Made by Zack Luo

Working with AWS

Using EC2

In AWS Console





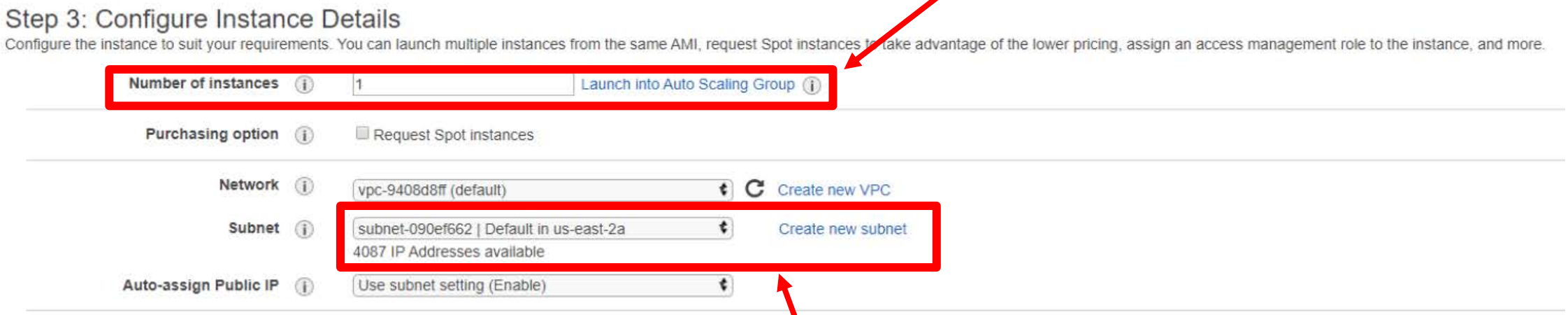
Select the one with free tier

Made by Zack Luo

Using EC2

In AWS Console

You can assign several instances



You can select a subnet for inter connection between VMS

Using EC2

In AWS Console

Step 6: Configure Security Group

A security group is a set of firewall rules that control the traffic for your instance. On this page, you can add rules to allow specific traffic to reach your instance. For example, if you want to set up a web server and allow Internet traffic to reach your instance, add rules that allow unrestricted access to the HTTP and HTTPS ports. You can create a new security group or select from an existing one below. Learn more about Amazon EC2 security groups.

Assign a security group: Create a **new** security group

Select an **existing** security group

Security group name: launch-wizard-1

Description: launch-wizard-1 created 2020-05-16T16:58:34.561-04:00

Type	Protocol	Port Range	Source	Description
SSH	TCP	22	Custom 0.0.0.0/0	e.g. SSH for Admin Desktop

Add Rule

Add SSH to allow connection from outside

Using EC2

In AWS Console

Step 6: Configure Security Group

A security group is a set of firewall rules that control the traffic for your instance. On this page, you can add rules to allow specific traffic to reach your instance. For example, if you want to set up a web server and allow Internet traffic to reach your instance, add rules that allow unrestricted access to the HTTP and HTTPS ports. You can create a new security group or select from an existing one below. Learn more about Amazon EC2 security groups.

Assign a security group: Create a **new** security group

Select an **existing** security group

Security group name: launch-wizard-1

Description: launch-wizard-1 created 2020-05-16T16:58:34.561-04:00

Type	Protocol	Port Range	Source	Description
SSH	TCP	22	Custom 0.0.0.0/0	e.g. SSH for Admin Desktop

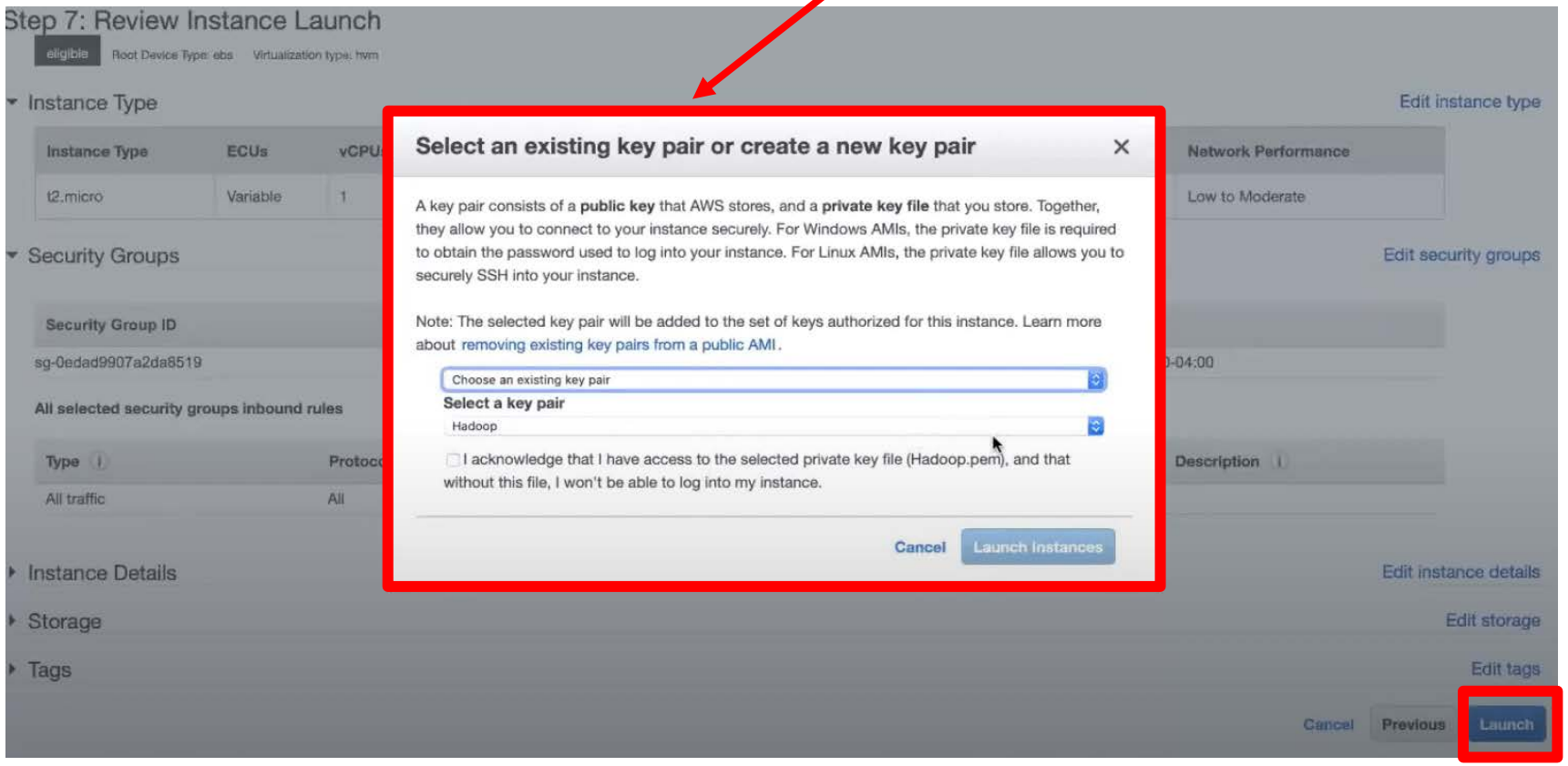
Add Rule

Then, follow the rest of the steps in the settings.
(eg. Adding tags, review…)

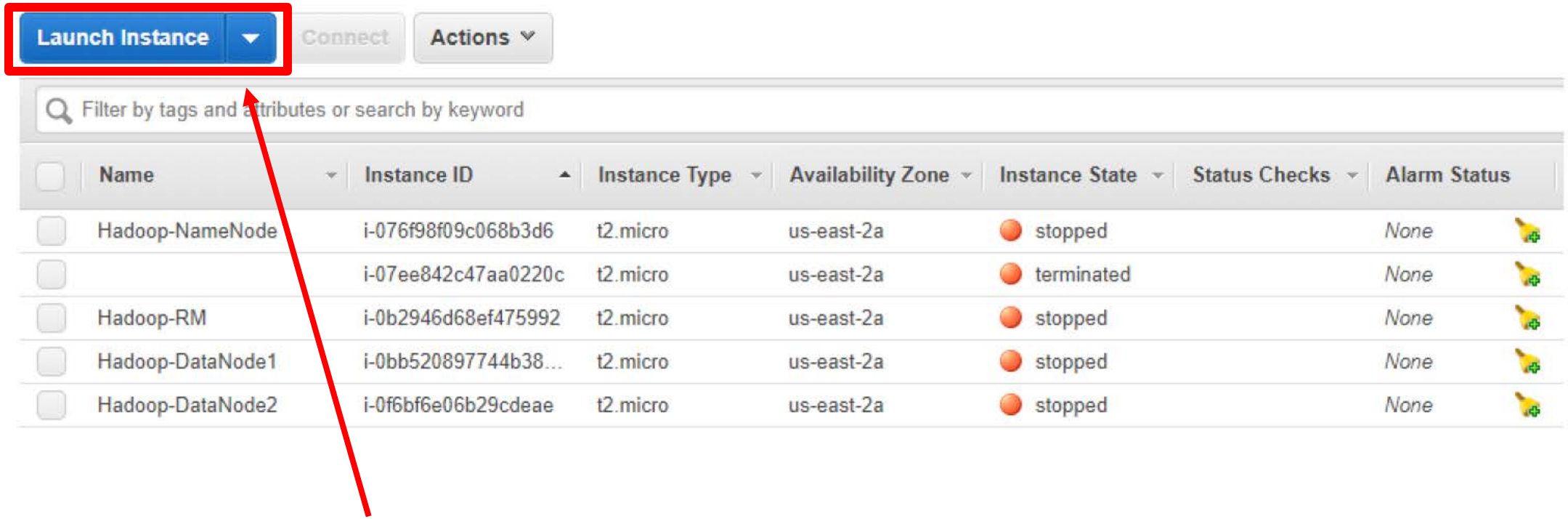
Using EC2

In AWS Console

After you click “launch”, this will pop up. Create a new key pair for SSH connection, (**as you did in LAB 1**)



Using EC2



Launch or stop instances

Using EC2

Launch Instance | Connect | Actions

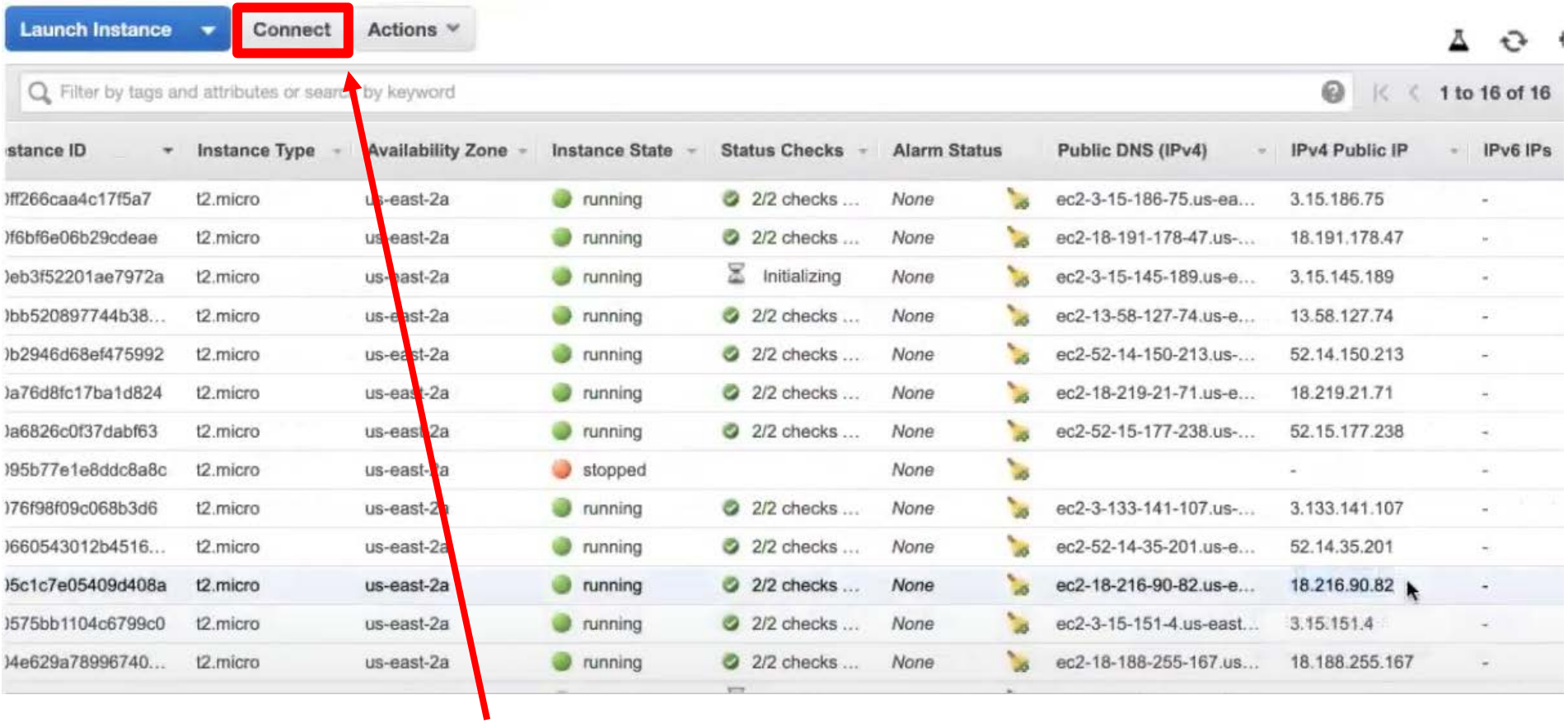
Filter by tags and attributes or search by keyword — 1 to 16 of 16

stance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS (IPv4)	IPv4 Public IP	IPv6 IPs
)ff266caa4c17f5a7	t2.micro	us-east-2a	running	2/2 checks ...	None	ec2-3-15-186-75.us-ea...	3.15.186.75	-
)f6bf6e06b29cdeae	t2.micro	us-east-2a	running	2/2 checks ...	None	ec2-18-191-178-47.us-...	18.191.178.47	-
)eb3f52201ae7972a	t2.micro	us-east-2a	running	Initializing	None	ec2-3-15-145-189.us-e...	3.15.145.189	-
)bb520897744b38...	t2.micro	us-east-2a	running	2/2 checks ...	None	ec2-13-58-127-74.us-e...	13.58.127.74	-
)b2946d68ef475992	t2.micro	us-east-2a	running	2/2 checks ...	None	ec2-52-14-150-213.us-...	52.14.150.213	-
)a76d8fc17ba1d824	t2.micro	us-east-2a	running	2/2 checks ...	None	ec2-18-219-21-71.us-e...	18.219.21.71	-
)a6826c0f37dabf63	t2.micro	us-east-2a	running	2/2 checks ...	None	ec2-52-15-177-238.us-...	52.15.177.238	-
)95b77e1e8ddc8a8c	t2.micro	us-east-2a	stopped		None		-	-
)76f98f09c068b3d6	t2.micro	us-east-2a	running	2/2 checks ...	None	ec2-3-133-141-107.us-...	3.133.141.107	-
)660543012b4516...	t2.micro	us-east-2a	running	2/2 checks ...	None	ec2-52-14-35-201.us-e...	52.14.35.201	-
)5c1c7e05409d408a	t2.micro	us-east-2a	running	2/2 checks ...	None	ec2-18-216-90-82.us-e...	18.216.90.82	-
)575bb1104c6799c0	t2.micro	us-east-2a	running	2/2 checks ...	None	ec2-3-15-151-4.us-east...	3.15.151.4	-
)4e629a78996740...	t2.micro	us-east-2a	running	2/2 checks ...	None	ec2-18-188-255-167.us...	18.188.255.167	-

After the instance is running, use the IP address to connect to the instance with SSH.

The default user name is “ubuntu”.

Using EC2

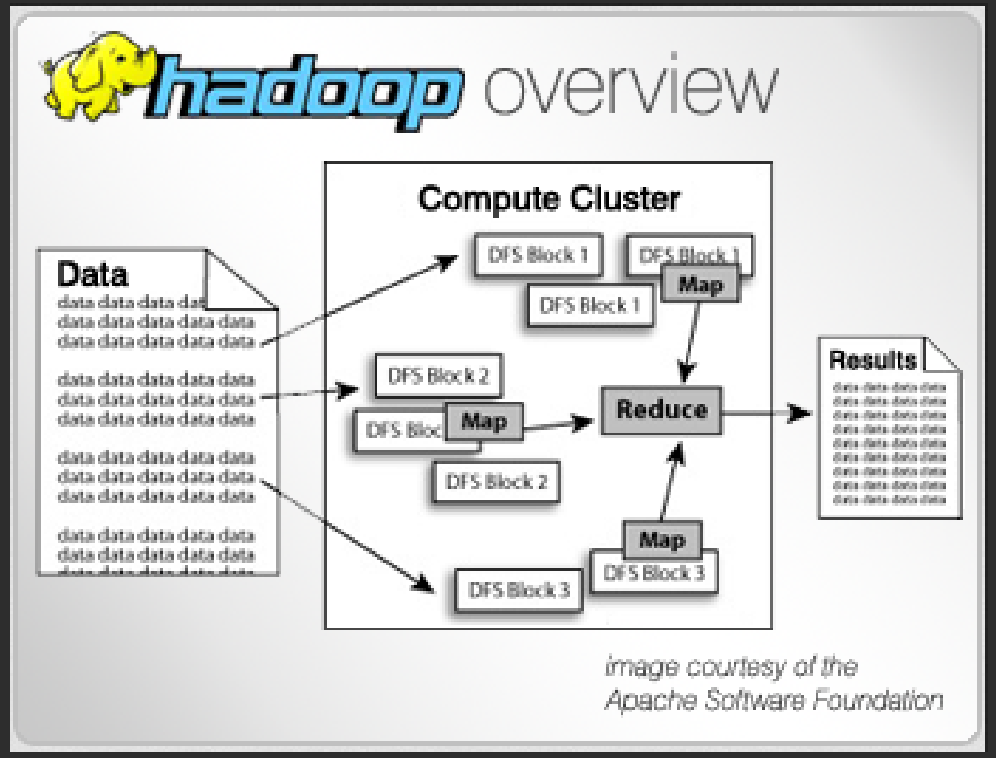


If you forget how to do so, there is a hint here.

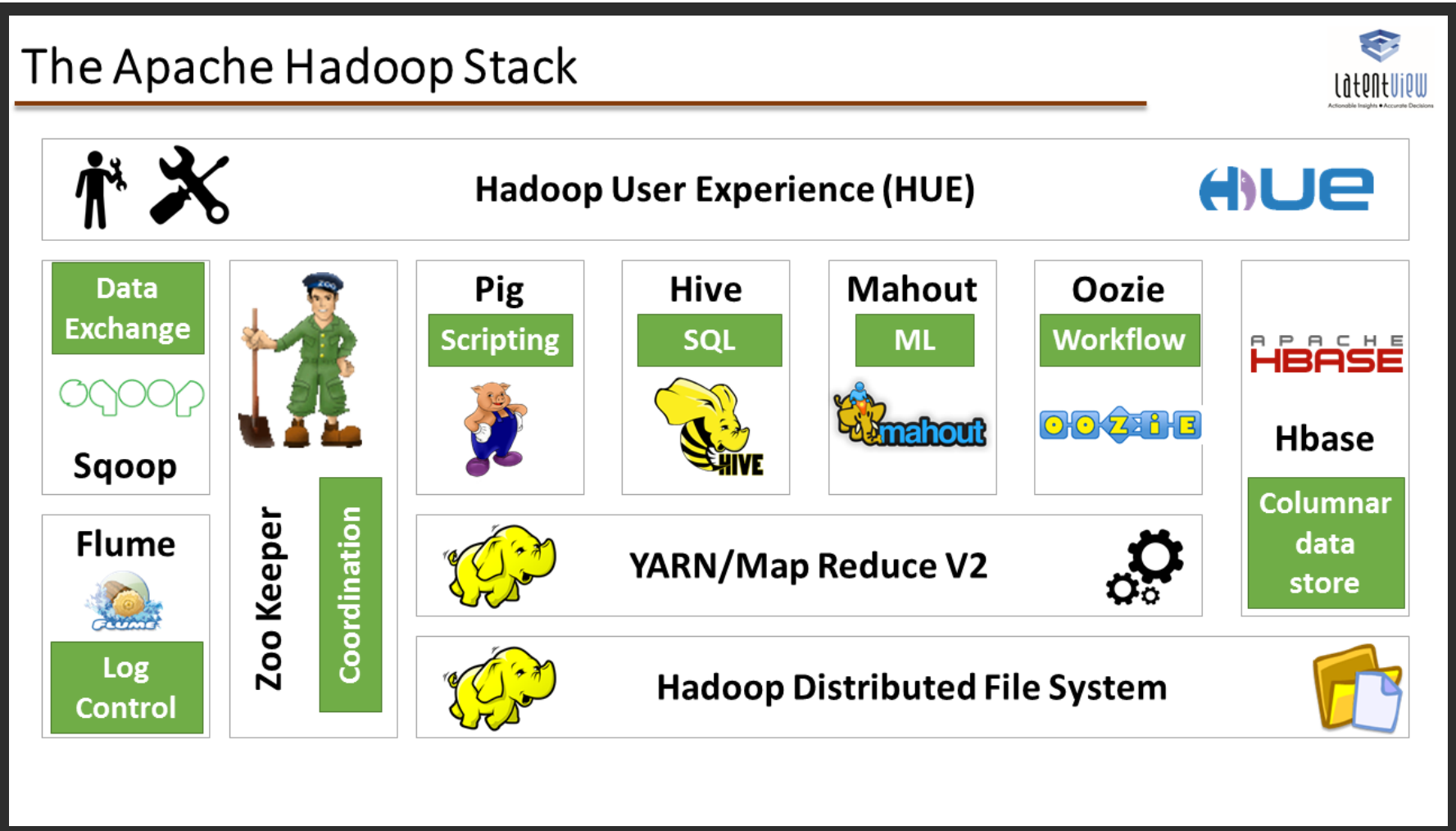
Hadoop & MapReduce

Hadoop

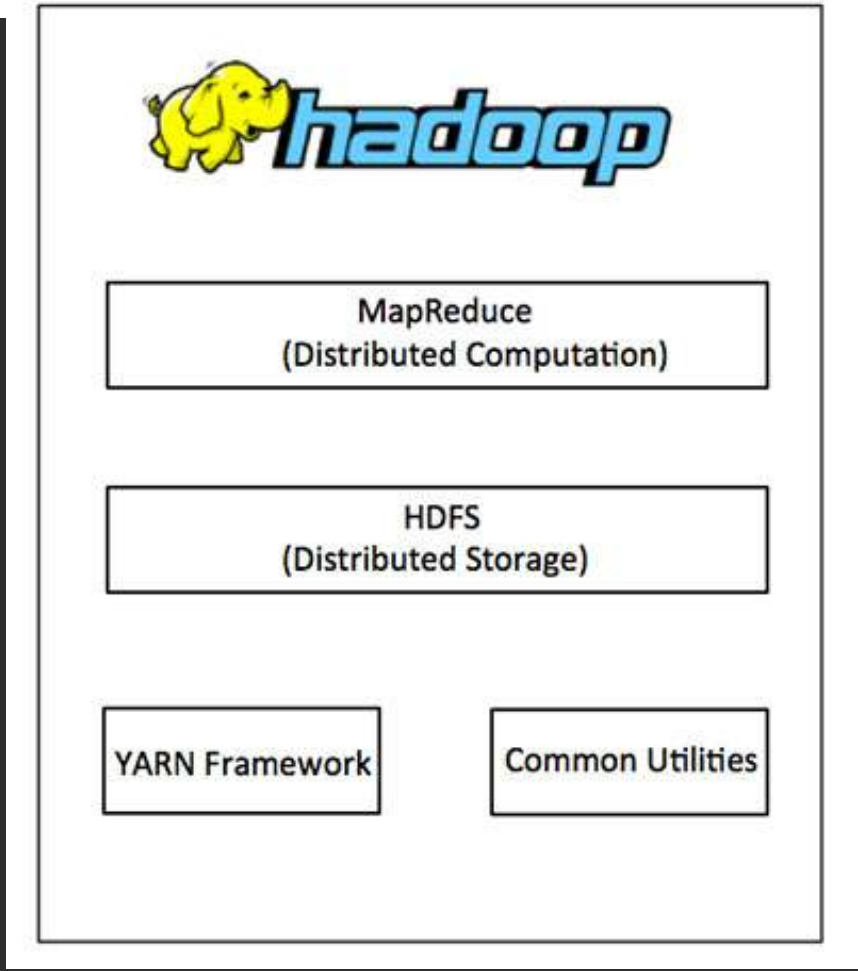
- Hadoop is an open source implementation of Google Map-Reduce.
- It uses a distributed file system: HDFS
- Runs Map-Reduce jobs.

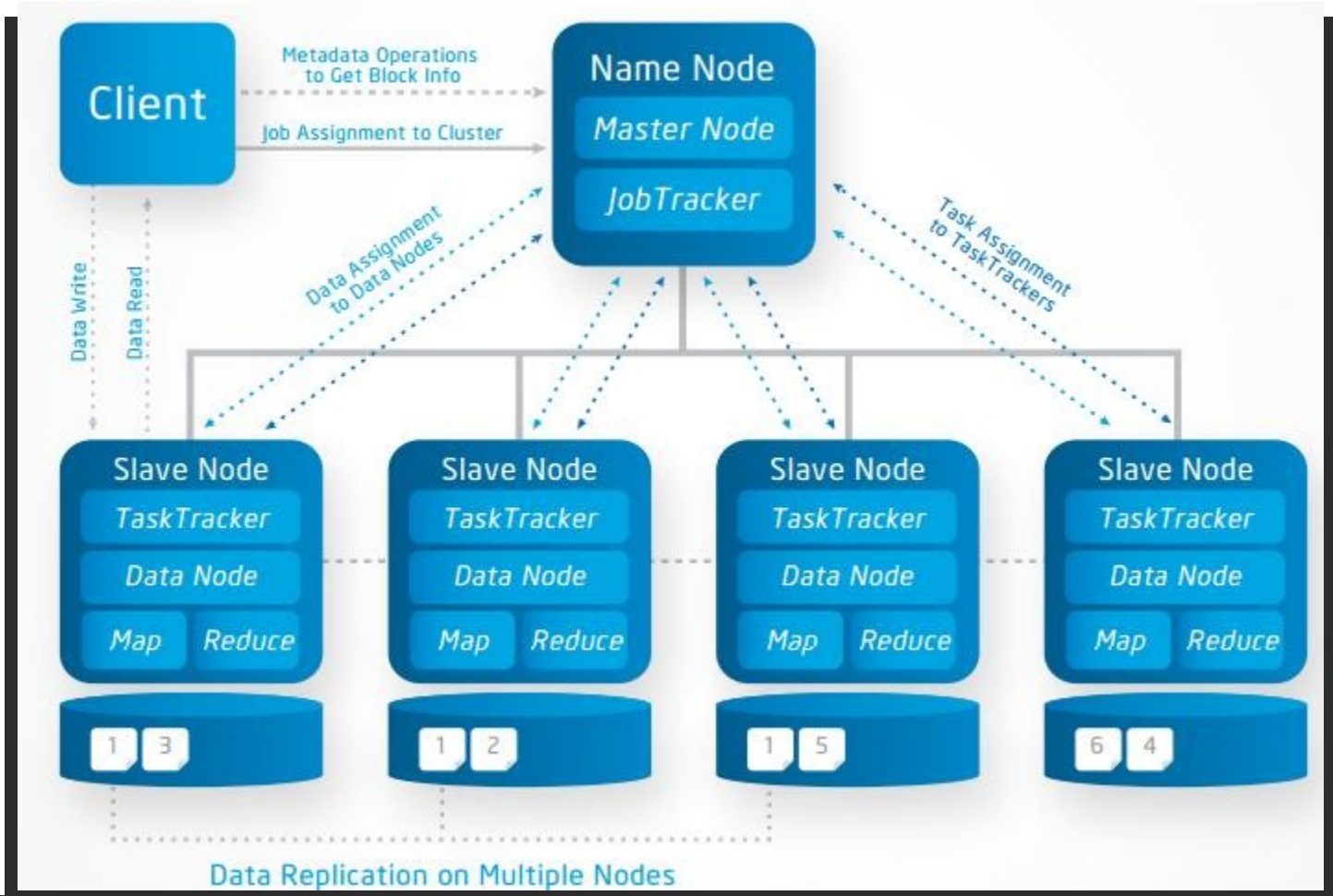


Hadoop Ecosystem

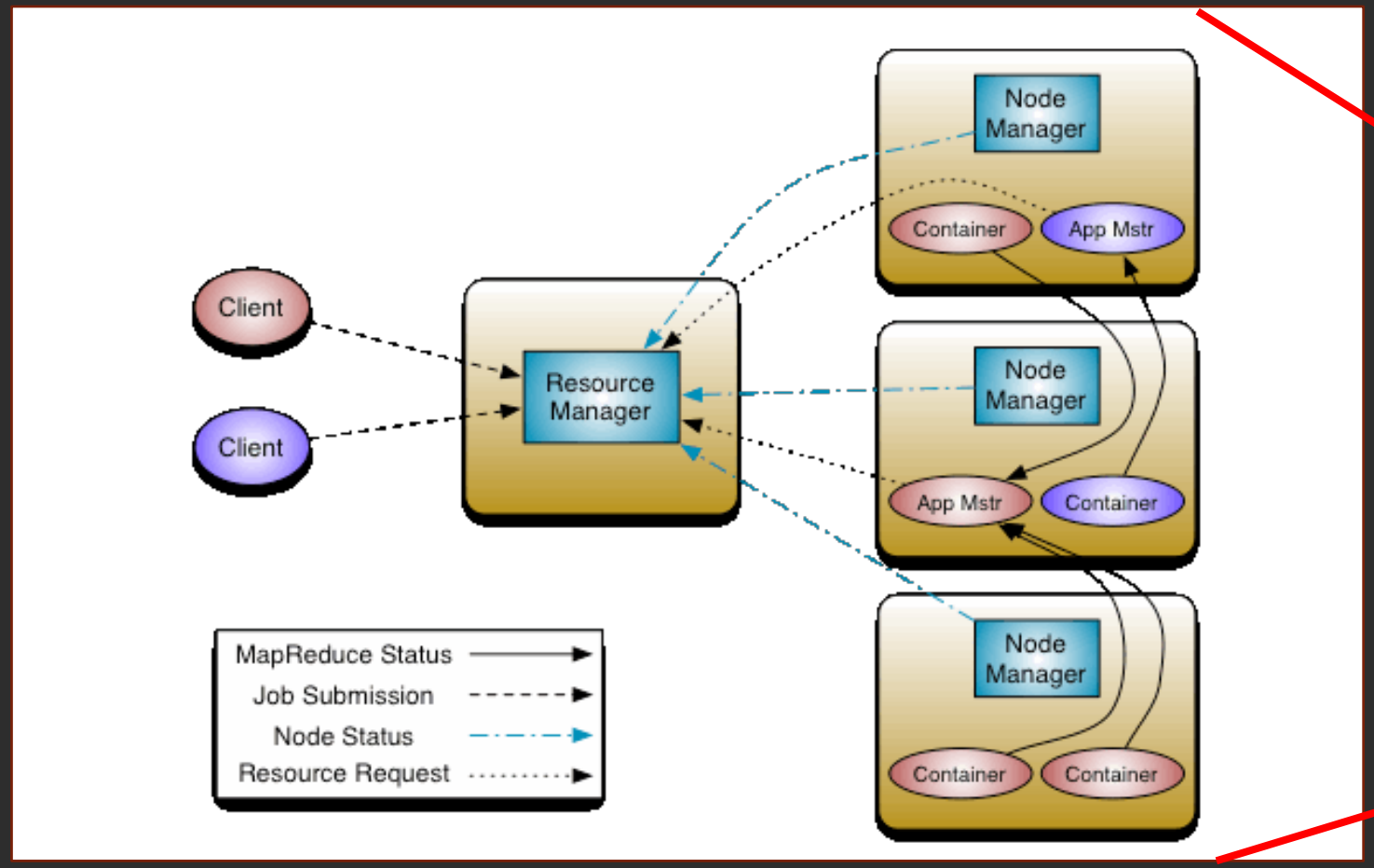


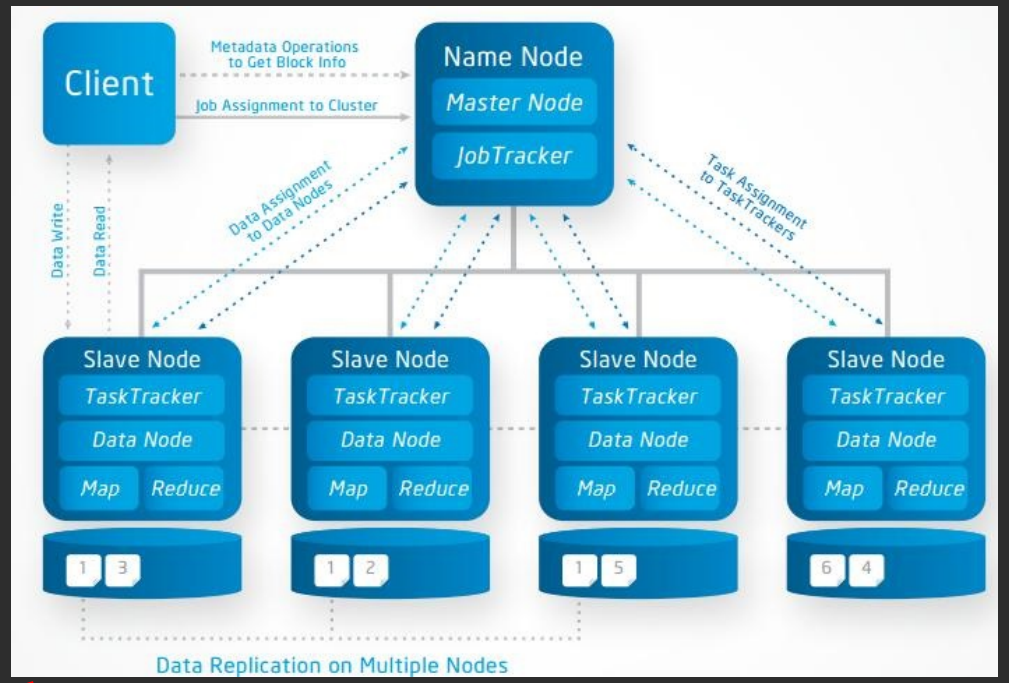
Hadoop Structure



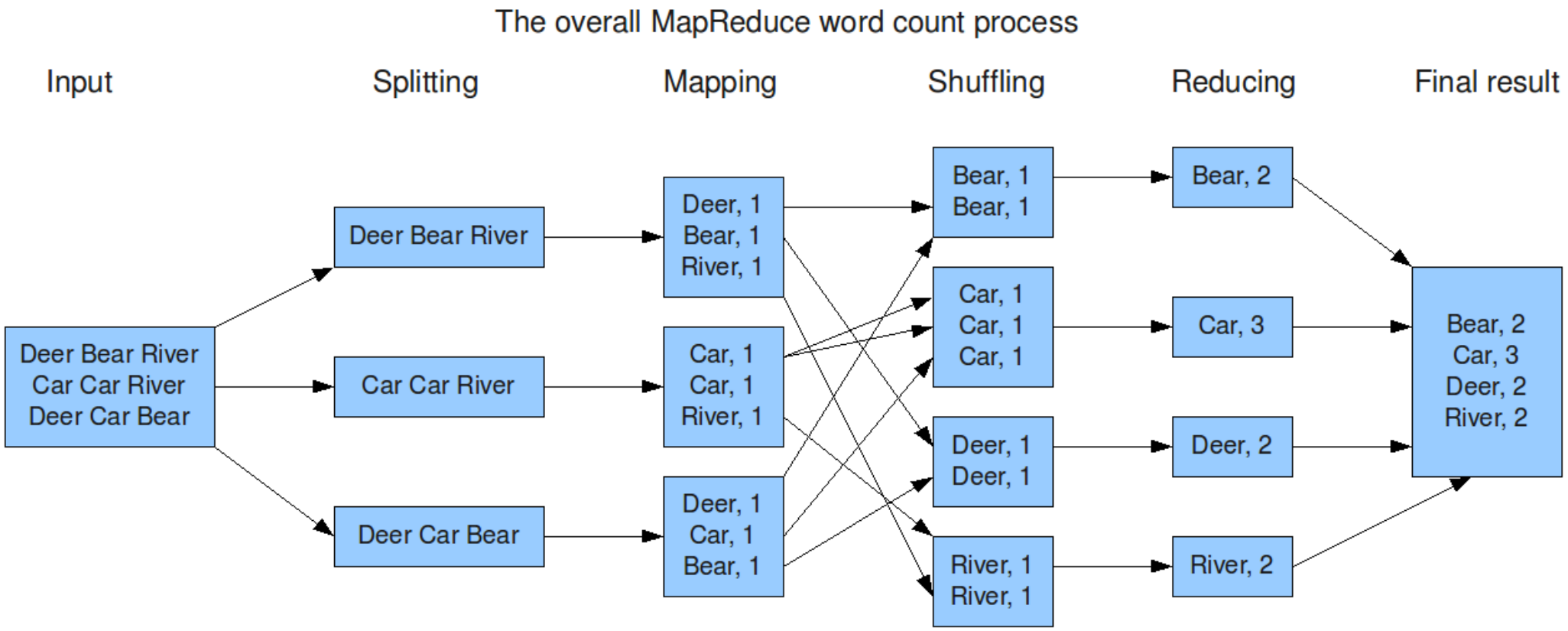


Hadoop Job Control

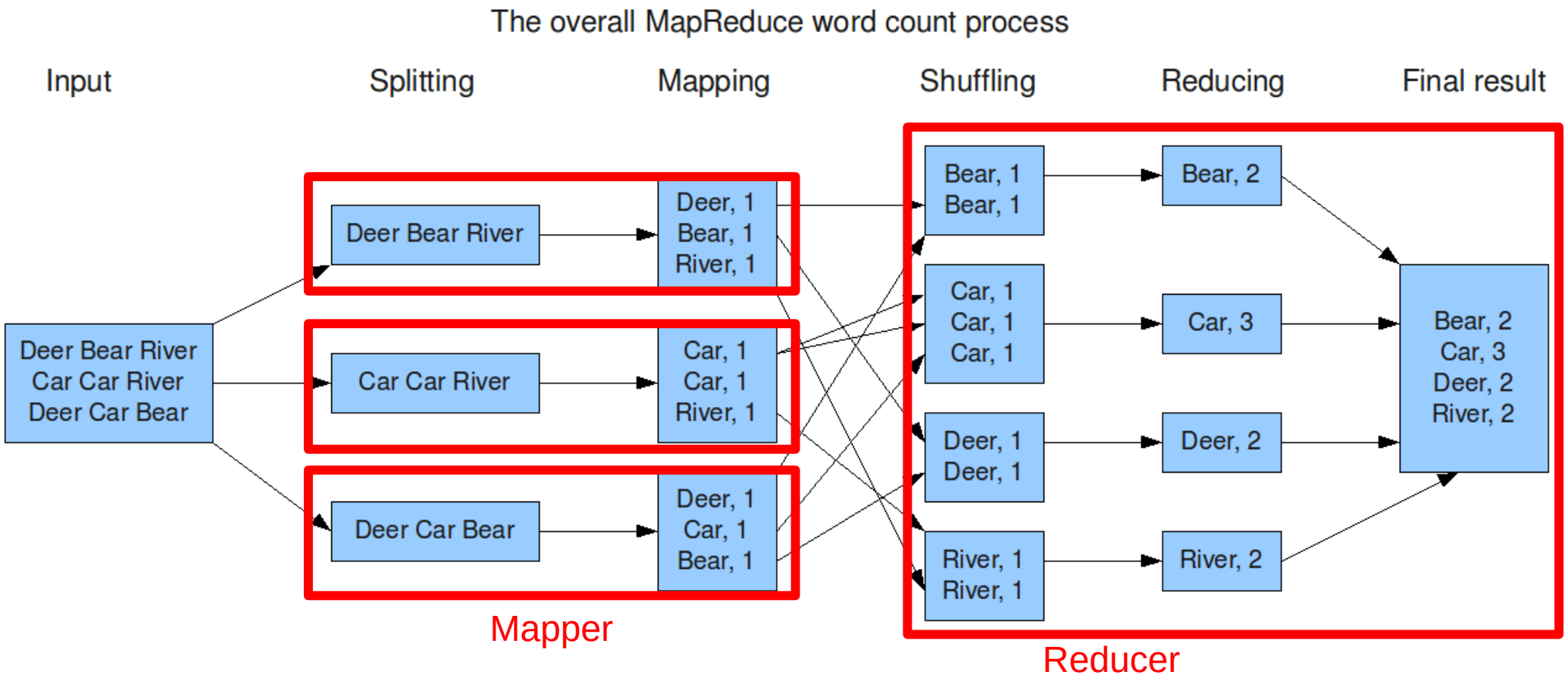




MAPREDUCE – WORD COUNT



MAPREDUCE – WORD COUNT



Installing Hadoop

- Installation
  - Follow the instructions in http://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html
  - Make sure that you go through these steps
    - **Prerequisites**
    - **Download**
    - **Prepare to Start the Hadoop Cluster**
    - **Standalone Operation**

Hadoop Pre-Requirement

- Make sure that you have installed the packets with apt-get:
  - default-jre
  - default-jdk
  - ssh
  - rsync

If you forget how to install the packets in Ubuntu, please review Lab 1

Installing Hadoop

- Installation
  - Follow the instructions in http://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html
  - Make sure that you go through these steps
    - **Prerequisites**
    - **Download**
    - **Prepare to Start the Hadoop Cluster**
    - **Standalone Operation**

At this point, you install Hadoop in a single node

Hadoop- Run MapReduce

- **Two files are provided to run the MapReduce word count example**
  - WordCount.java: word count MapReduce program
  - wordCountText.txt: input text file to the WordCount program

- **Put the text file into HDFS**
  - (Depends on the path you install Hadoop, and the version of Hadoop)
  - bin/hadoop fs -mkdir input
  - bin/hadoop fs -put wordCountText.txt input

Hadoop- Run MapReduce

- **Compile WordCount program and produce the jar file**
  - mkdir wordcount_classes
  - javac -classpath ${HADOOP_CLASSPATH} -d wordcount_classes WordCount.java
    - (you can find out the **classpath** by issuing bin/hadoop classpath)
  - jar -cvf wc.jar -C wordcount_classes/ .
- **Run MapReduce with the produced jar file**
  - bin/hadoop jar wc.jar WordCount input output
  - Get the result from the output, then you can get the word counts
  - You need to report top-5 frequent words in the given text file

Hadoop in Cluster Mode (Bonus)

- You'll need two or more VMs
- Enable network between the master and slave VM,
  - there are different ways to do this, one way to do this in VirtualBox,
    - In VirtualBox general setting->Network: create a host-only network: vboxnet
    - for each VM under Setting->Network: enable adapter 2 and attach to the host-only network created
    - login the VMs and configure the IP addresses,
      - ex: sudo ifconfig enp0s8 10.0.0.5.1 netmask 255.255.255.0
    - test if the VMs can ping each other

Hadoop in cluster mode - Configuration

- You can use the following references for configuration
  - http://pingax.com/install-apache-hadoop-ubuntu-cluster-setup/
  - http://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/ClusterSetup.html
- Files that need to be configured (for both master and slave VMs)
  - /etc/hosts
  - hadoop-2.7.2/etc/hadoop/hadoop-env.sh
  - hadoop-2.7.2/etc/hadoop/core-site.xml
  - hadoop-2.7.2/etc/hadoop/hdfs-site.xml
  - hadoop-2.7.2/etc/hadoop/yarn-site.xml
  - hadoop-2.7.2/etc/hadoop/mapred-site.xml.template
  - hadoop-2.7.2/etc/hadoop/masters
  - hadoop-2.7.2/etc/hadoop/slaves

Hadoop in cluster mode – start and run

- Format the namenode (on master)
  - bin/hdfs namenode -format
- Start HDFS daemons and MapReduce daemons (on master)
  - sbin/start-all.sh
- Use jps command to check running daemons
  - on master: NameNode, SecondaryNameNode, Jps, ResourceManager
  - on slave: Jps, DataNode, NodeManager
- Run **WordCount** and see if it is faster