

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.

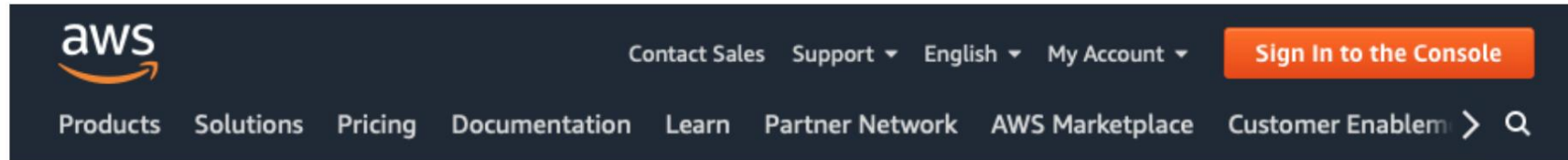
2. S3

Simple Storage Service, object storage.



Working with AWS

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!



Sign in

☒ **Root user**

Account owner that performs tasks requiring unrestricted access. [Learn more](#)

☐ **IAM user**

User within an account that performs daily tasks. [Learn more](#)

Root user email address

username@example.com

Next

— New to AWS? —

Create a new AWS 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.

Cost Management

Cost Explorer

Budgets

Budgets Reports

Savings Plans

Cost & Usage Reports

Cost Categories

Cost allocation tags

Billing

Bills

Orders and invoices

Credits

Preferences

Billing preferences

Payment methods

Consolidated billing

Tax settings

Billing Preferences

☒ **Receive PDF Invoice By Email**

Turn on this feature to receive a PDF version of your invoice by email. Invoices are generally available within the first three days of the month.

Cost Management Preferences

☒ **Receive Free Tier Usage Alerts**

Turn on this feature to receive email alerts when your AWS service usage is approaching, or has exceeded, the AWS Free Tier usage limits. If you wish to receive these alerts at an email address that is not the primary email address associated with this account, please specify the email address below.

Email Address:

☒ **Receive Billing Alerts**

Turn on this feature to monitor your AWS usage charges and recurring fees automatically, making it easier to track and manage your spending on AWS. You can set up billing alerts to receive email notifications when your charges reach a specified threshold. Once enabled, this preference cannot be disabled. [Manage Billing Alerts](#) or [try the new budgets feature!](#)

► Detailed Billing Reports [Legacy]

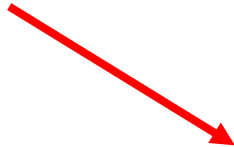
Save preferences

Working with AWS

Using EC2

In AWS Console

Select this to build
a virtual machine



AWS Management Console

AWS services









Find Services
You can enter names, keywords or acronyms.

Q Example: Relational Database Service, database, RDS

► Recently visited services

► All services

Build a solution
Get started with simple wizards and automated workflows.

Launch a virtual machine With EC2 2-3 minutes 	Build a web app With Elastic Beanstalk 6 minutes 	Build using virtual servers With Lightsail 1-2 minutes 	Register a domain With Route 53 3 minutes 
Connect an IoT device With AWS IoT 5 minutes 	Start migrating to AWS With CloudEndure Migration 1-2 minutes 	Start a development project With CodeStar 5 minutes 	Deploy a serverless microservice With Lambda, API Gateway 2 minutes 

► See more

Stay connected to your AWS resources on-the-go

Download the AWS Console Mobile App to your iOS or Android mobile device. [Learn more](#)

Explore AWS


Run Containers Not Servers
Build, Deploy, and Operate Containerized Applications with AWS Fargate. [Learn More](#)

Amazon SageMaker Autopilot
Get hands-on with this Auto-ML workshop. [Learn more](#)

Amazon Redshift RA3 Nodes
Scale your compute and storage independently and lower your costs. [Learn more](#)

Free Digital Training
Get access to 350+ self-paced online courses covering AWS products and services. [Learn more](#)

Have feedback?

 [Submit feedback](#) to tell us about your experience with the AWS Management Console.

Working with AWS

Using EC2

In AWS Console

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Add Tags 6. Configure Security Group 7. Review

Step 1: Choose an Amazon Machine Image (AMI)




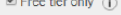






An AMI is a template that contains the software configuration (operating system, application server, and applications) required to launch your instance. You can select an AMI provided by AWS, our user community, or the AWS Marketplace; or you can select one of your own AMIs.

[Cancel and Exit](#)

Search for an AMI by entering a search term e.g. "Windows" ×

[Search by Systems Manager parameter](#)

Quick Start |< < 1 to 16 of 16 AMIs > >|

My AMIs	 Amazon Linux 2 AMI (HVM), SSD Volume Type - ami-0f919c33c90f5b58 (64-bit x86) / ami-050d581a8c1d4a570 (64-bit Arm)	Select
AWS Marketplace	 Amazon Linux 2 comes with five years support. It provides Linux kernel 4.14 tuned for optimal performance on Amazon EC2, systemd 219, GCC 7.3, Glibc 2.26, Binutils 2.29.1, and the latest software packages through extras. Root device type: ebs Virtualization type: hvm ENA Enabled: Yes	<input checked="" type="radio"/> 64-bit (x86) <input type="radio"/> 64-bit (Arm)
Community AMIs	 Amazon Linux AMI 2018.03.0 (HVM), SSD Volume Type - ami-097834fcb3081f51a	Select
	 The Amazon Linux AMI is an EBS-backed, AWS-supported image. The default image includes AWS command line tools, Python, Ruby, Perl, and Java. The repositories include Docker, PHP, MySQL, PostgreSQL, and other packages. Root device type: ebs Virtualization type: hvm ENA Enabled: Yes	64-bit (x86)
 Red Hat Enterprise Linux 8 (HVM), SSD Volume Type - ami-0a54aef4ef3b5f881 (64-bit x86) / ami-0ff50b53e6797671 (64-bit Arm)	Select	
 Red Hat Enterprise Linux version 8 (HVM), EBS General Purpose (SSD) Volume Type Root device type: ebs Virtualization type: hvm ENA Enabled: Yes	<input checked="" type="radio"/> 64-bit (x86) <input type="radio"/> 64-bit (Arm)	
 SUSE Linux Enterprise Server 15 SP1 (HVM), SSD Volume Type - ami-013d888bfc1a3962 (64-bit x86) / ami-026bf16f292030d5d (64-bit Arm)	Select	
 SUSE Linux Enterprise Server 15 Service Pack 1 (HVM), EBS General Purpose (SSD) Volume Type. Public Cloud, Advanced Systems Management, Web and Scripting, and Legacy modules enabled. Root device type: ebs Virtualization type: hvm ENA Enabled: Yes	<input checked="" type="radio"/> 64-bit (x86) <input type="radio"/> 64-bit (Arm)	
 Ubuntu Server 18.04 LTS (HVM), SSD Volume Type - ami-07c1207a9d40bc3bd (64-bit x86) / ami-0a5ee0336de62011b (64-bit Arm)	Select	
 Ubuntu Server 18.04 LTS (HVM), EBS General Purpose (SSD) Volume Type. Support available from Canonical (http://www.ubuntu.com/cloud/services). Root device type: ebs Virtualization type: hvm ENA Enabled: Yes	<input checked="" type="radio"/> 64-bit (x86) <input type="radio"/> 64-bit (Arm)	

Select Ubuntu

Made by Zack Luo

Using EC2

In AWS Console

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Add Tags 6. Configure Security Group 7. Review

Step 1: Choose an Amazon Machine Image (AMI)

An AMI is a template that contains the software configuration (operating system, application server, and applications) required to launch your instance. You can

Search for an AMI by entering a search term e.g. "Windows"

Quick Start

My AMIs

AWS Marketplace

Community AMIs



Amazon Linux 2 AMI (HVM), SSD Volume Type - ami-0f7919c33c90f5b58 (64-bit x86) / ami-0f7919c33c90f5b58
Amazon Linux 2 comes with five years support. It provides Linux kernel 4.14 tuned for optimal performance or extras.

Root device type: ebs Virtualization type: hvm ENA Enabled: Yes



Amazon Linux AMI 2018.03.0 (HVM), SSD Volume Type - ami-097834fcb3081f51a



The Amazon Linux AMI is an EBS-backed, AWS-supported image. The default image includes AWS command-line packages.

Root device type: ebs Virtualization type: hvm ENA Enabled: Yes



Red Hat Enterprise Linux 8 (HVM), SSD Volume Type - ami-0a54aef4ef3b5f881 (64-bit x86)
Red Hat Enterprise Linux version 8 (HVM), EBS General Purpose (SSD) Volume Type

Root device type: ebs Virtualization type: hvm ENA Enabled: Yes



SUSE Linux Enterprise Server 15 SP1 (HVM), SSD Volume Type - ami-013d888bfc1a3962
SUSE Linux Enterprise Server 15 Service Pack 1 (HVM), EBS General Purpose (SSD) Volume Type, Public Cloud

Root device type: ebs Virtualization type: hvm ENA Enabled: Yes



Ubuntu Server 18.04 LTS (HVM), SSD Volume Type - ami-07c1207a9d40bc3bd (64-bit x86) /
Ubuntu Server 18.04 LTS (HVM), EBS General Purpose (SSD) Volume Type. Support available from Canonical

Root device type: ebs Virtualization type: hvm ENA Enabled: Yes

Step 2: Choose an Instance Type

Amazon EC2 provides a wide selection of instance types optimized to fit different use cases. Instances are virtual servers that can run applications. A mix of resources for your applications. [Learn more](#) about instance types and how they can meet your computing needs.

Filter by:

All instance types

Current generation

Show/Hide Columns

Currently selected: t2.micro (Variable ECUs, 1 vCPUs, 2.5 GHz, Intel Xeon Family, 1 GiB memory, EBS only)

	Family	Type	vCPUs	Memory (GiB)
<input type="checkbox"/>	General purpose	t2.nano	1	0.5
<input checked="" type="checkbox"/>	General purpose	t2.micro Free tier eligible	1	1
<input type="checkbox"/>	General purpose	t2.small	1	2
<input type="checkbox"/>	General purpose	t2.medium	2	4
<input type="checkbox"/>	General purpose	t2.large	2	8

Select the one with free tier

Working with AWS

Using EC2

In AWS Console

You can assign several instances

Step 3: Configure Instance Details

Configure the instance to suit your requirements. You can launch multiple instances from the same AMI, request Spot instances to take advantage of the lower pricing, assign an access management role to the instance, and more.

Number of instances ⓘ 1 [Launch into Auto Scaling Group ⓘ](#)

Purchasing option ⓘ ☐ Request Spot instances

Network ⓘ vpc-9408d8ff (default) ↕ [Create new VPC](#)

Subnet ⓘ subnet-090ef662 | Default in us-east-2a ↕ [Create new subnet](#)
4087 IP Addresses available

Auto-assign Public IP ⓘ Use subnet setting (Enable) ↕

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:

Description:

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

Add SSH to allow connection
from outside

Working with AWS

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:

Description:

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

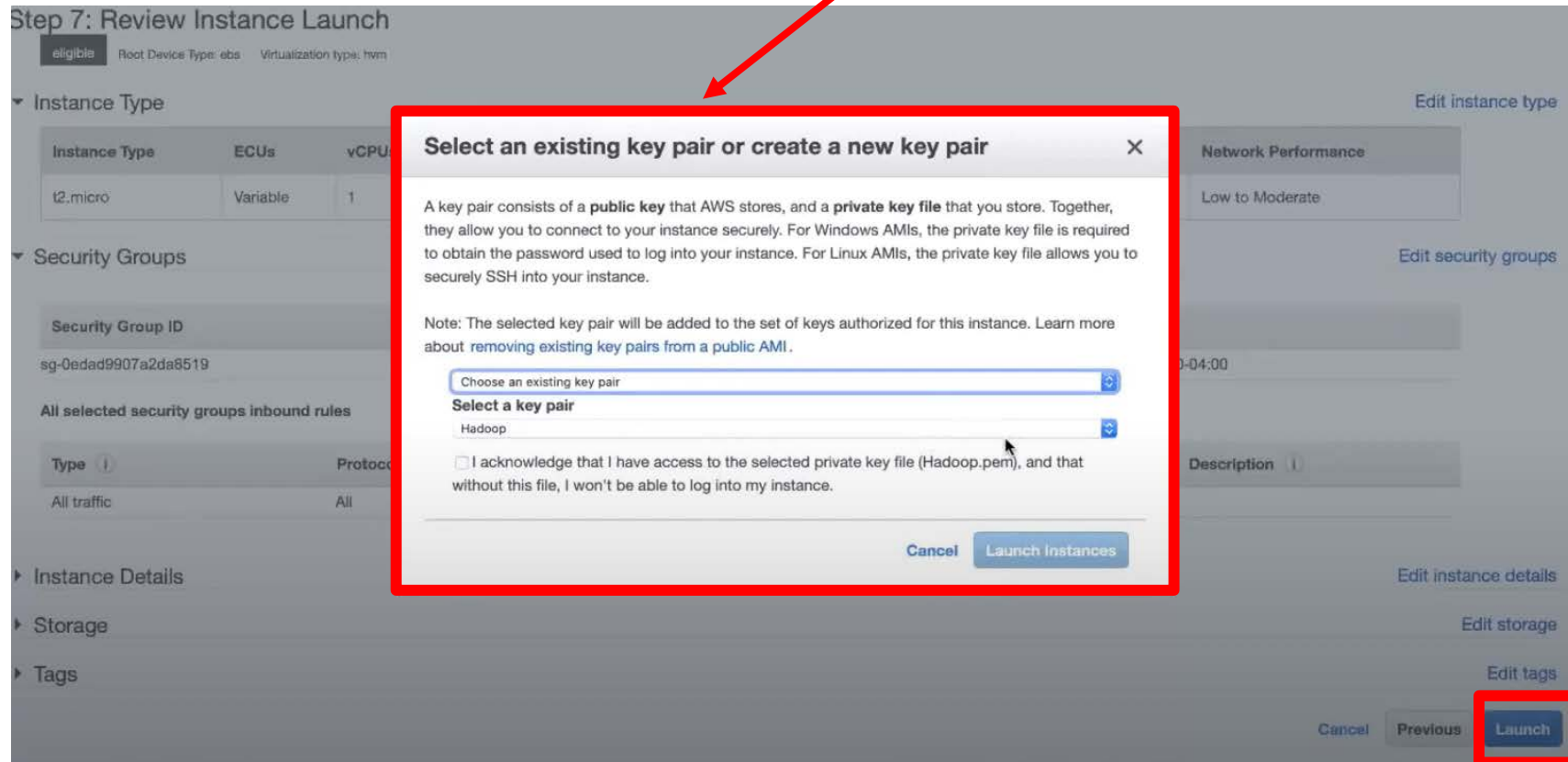
Then, follow the rest of the steps in the settings.
(eg. Adding tags, review...)

Working with AWS

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 Instance ▼

Connect

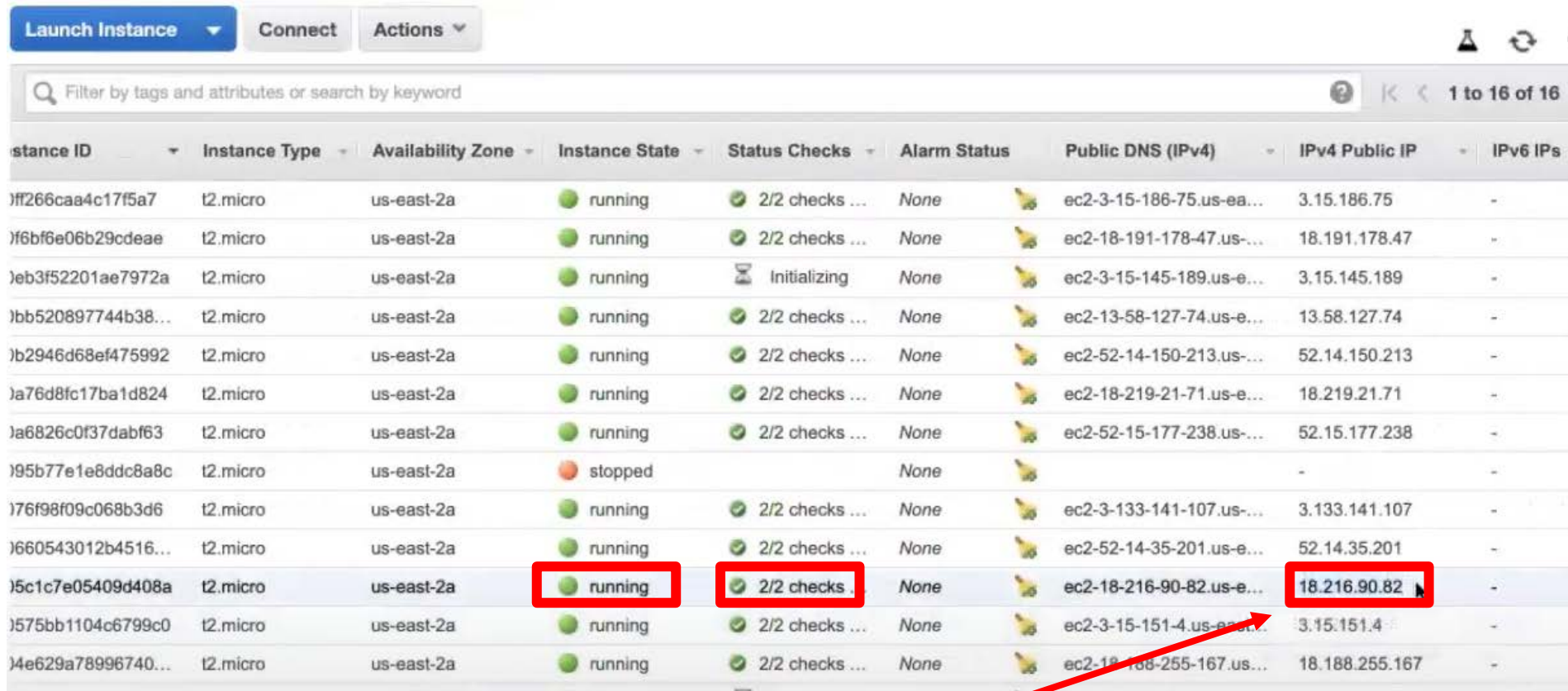
Actions ▼

<input type="checkbox"/>	Name ▼	Instance ID ▲	Instance Type ▼	Availability Zone ▼	Instance State ▼	Status Checks ▼	Alarm Status
<input type="checkbox"/>	Hadoop-NameNode	i-076f98f09c068b3d6	t2.micro	us-east-2a	● stopped		None
<input type="checkbox"/>		i-07ee842c47aa0220c	t2.micro	us-east-2a	● terminated		None
<input type="checkbox"/>	Hadoop-RM	i-0b2946d68ef475992	t2.micro	us-east-2a	● stopped		None
<input type="checkbox"/>	Hadoop-DataNode1	i-0bb520897744b38...	t2.micro	us-east-2a	● stopped		None
<input type="checkbox"/>	Hadoop-DataNode2	i-0f6bf6e06b29cdeae	t2.micro	us-east-2a	● stopped		None

Launch or stop instances

Working with AWS

Using EC2



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

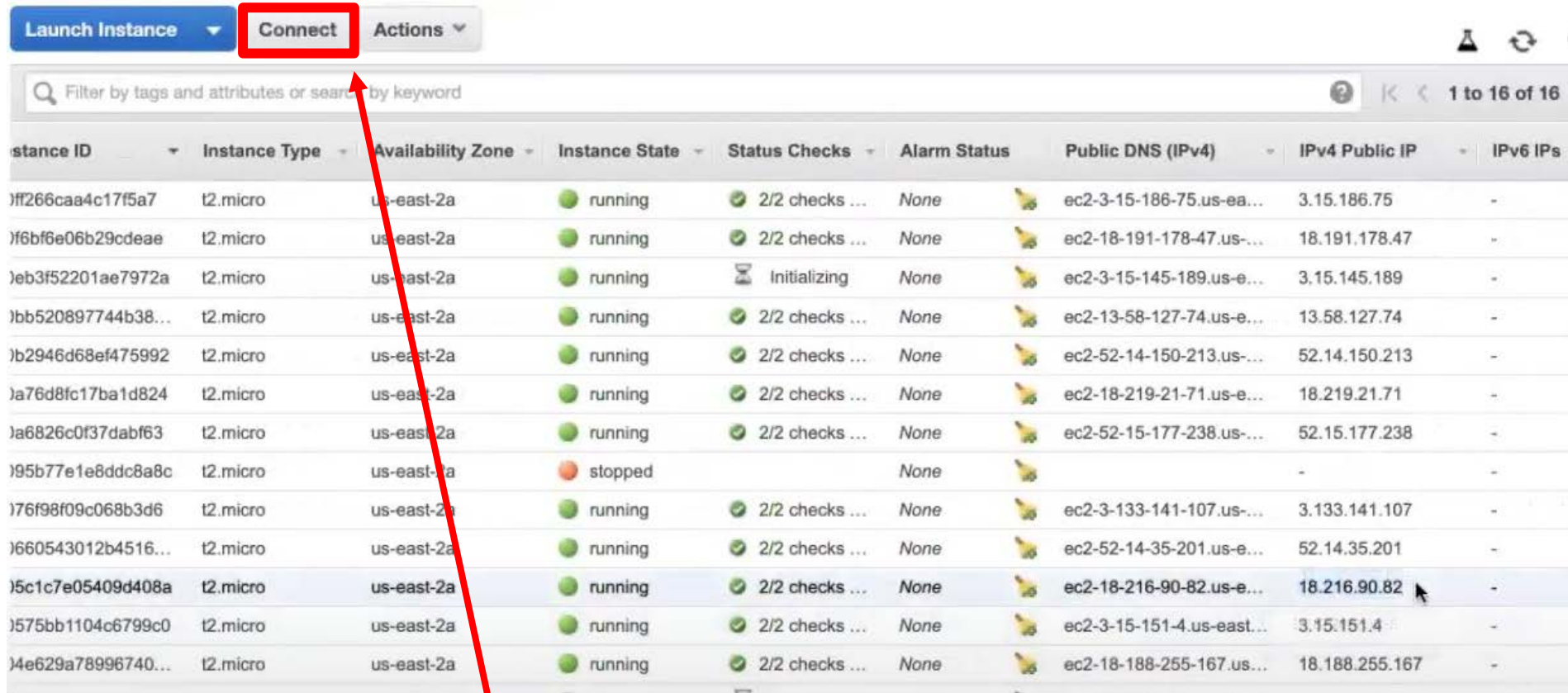
After the instance is running, use the IP address to connect to the instance with SSH.

The default user name is “ubuntu”.

Made by Zack Luo

Working with AWS

Using EC2



Launch Instance **Connect** Actions

Filter by tags and attributes or search by keyword

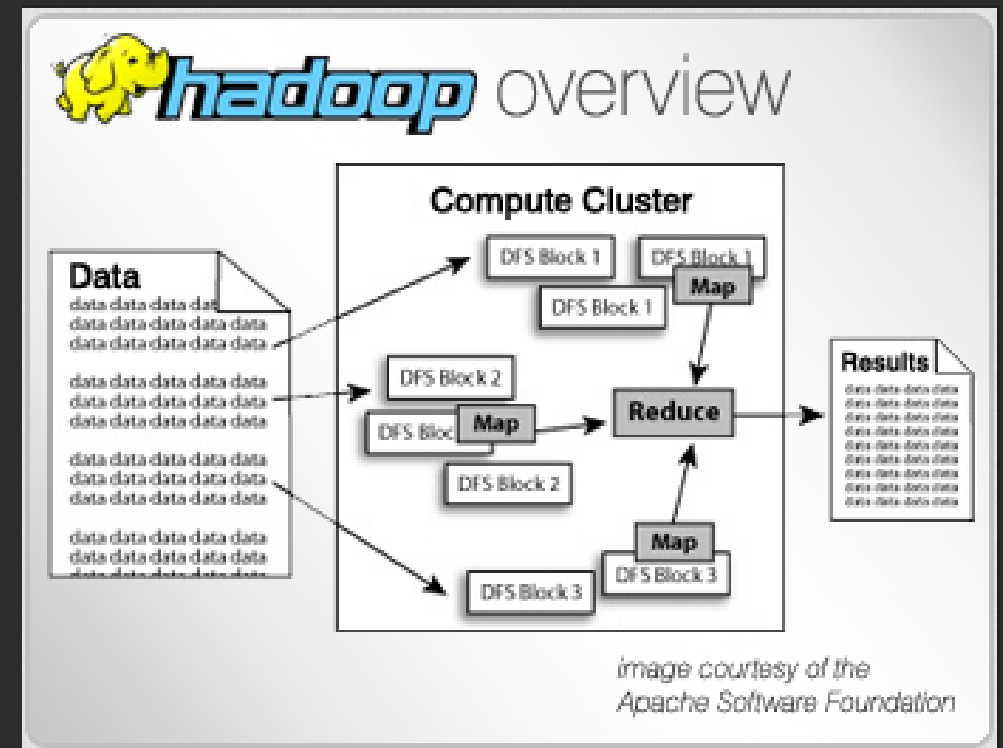
Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS (IPv4)	IPv4 Public IP	IPv6 IPs
i-ff266caa4c17f5a7	t2.micro	us-east-2a	running	2/2 checks ...	None	ec2-3-15-186-75.us-east-2.amazonaws.com	3.15.186.75	-
i-f6bf6e06b29cdeae	t2.micro	us-east-2a	running	2/2 checks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

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

Hadoop & MapReduce

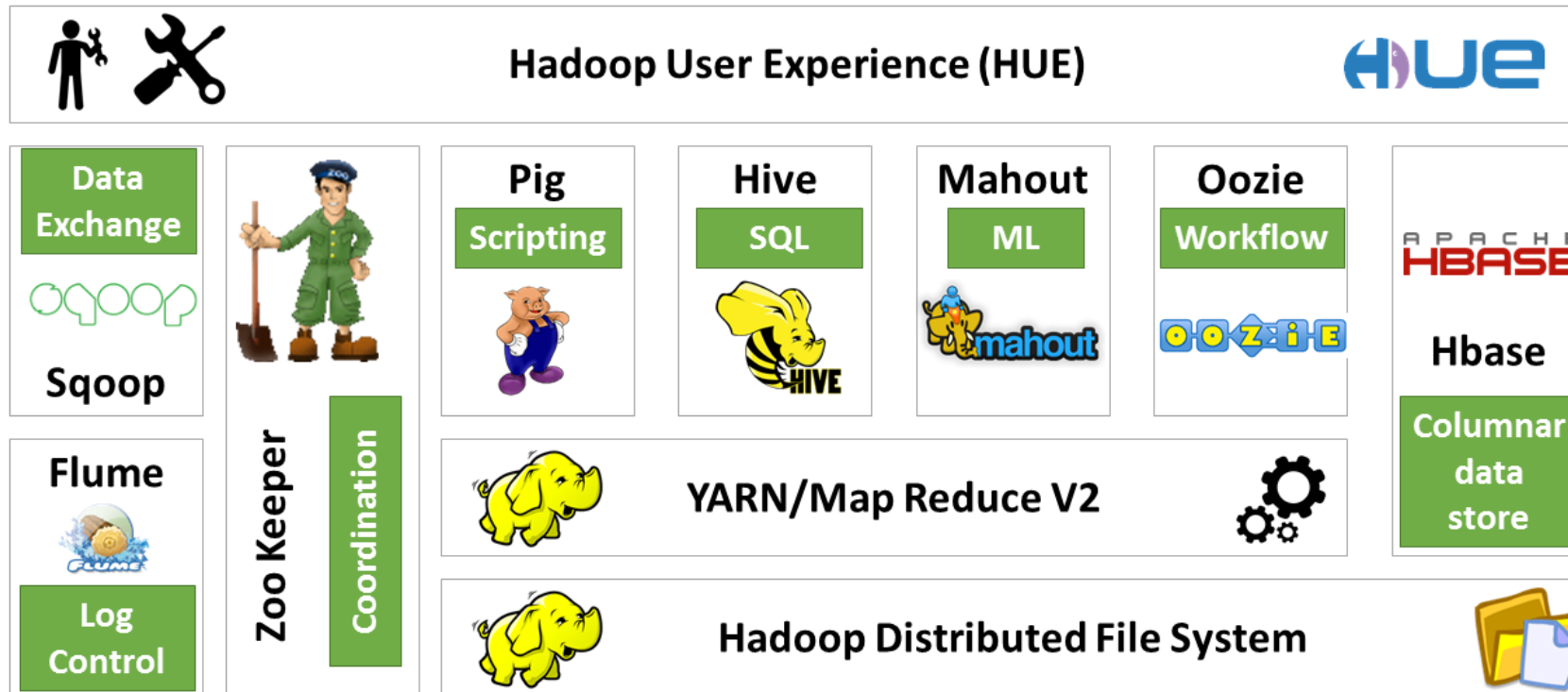
Hadoop

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

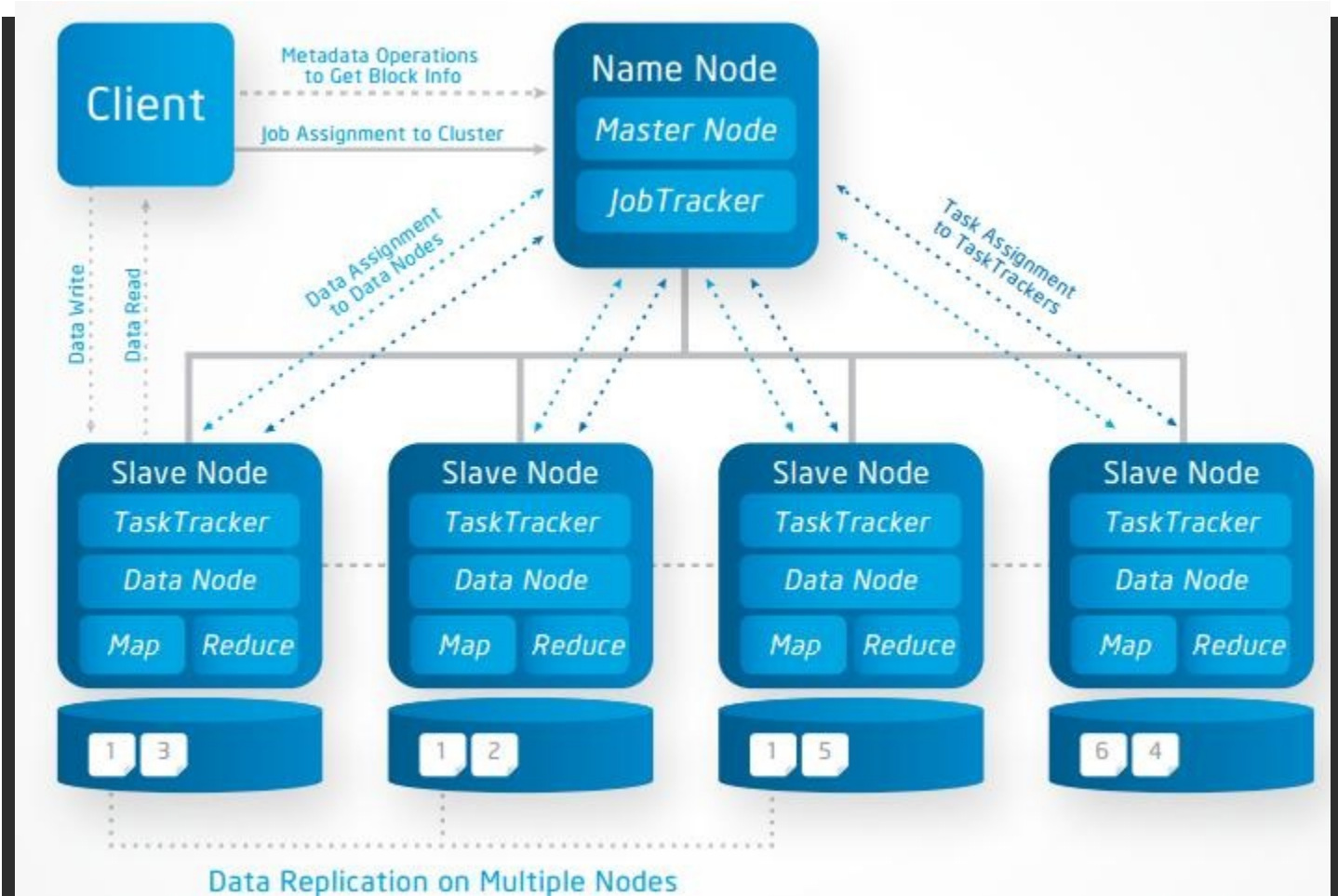
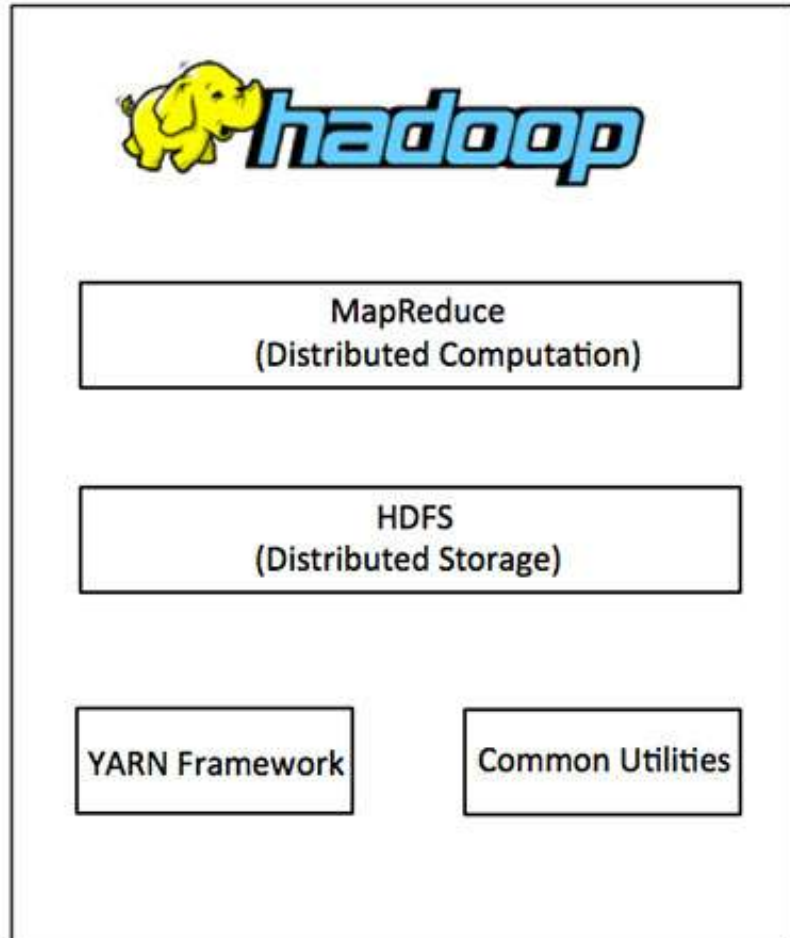


Hadoop Ecosystem

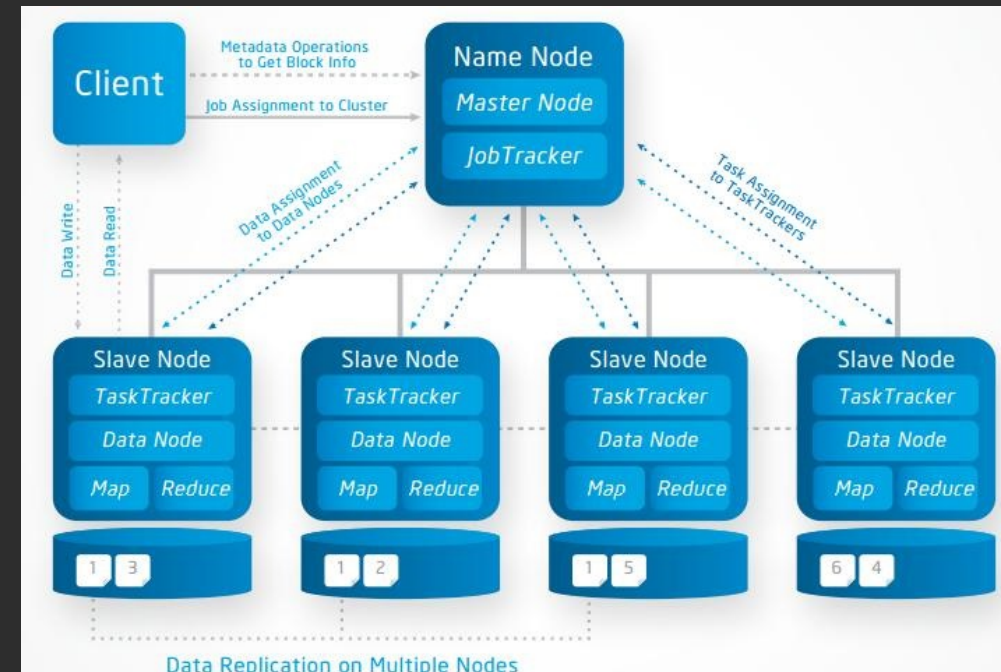
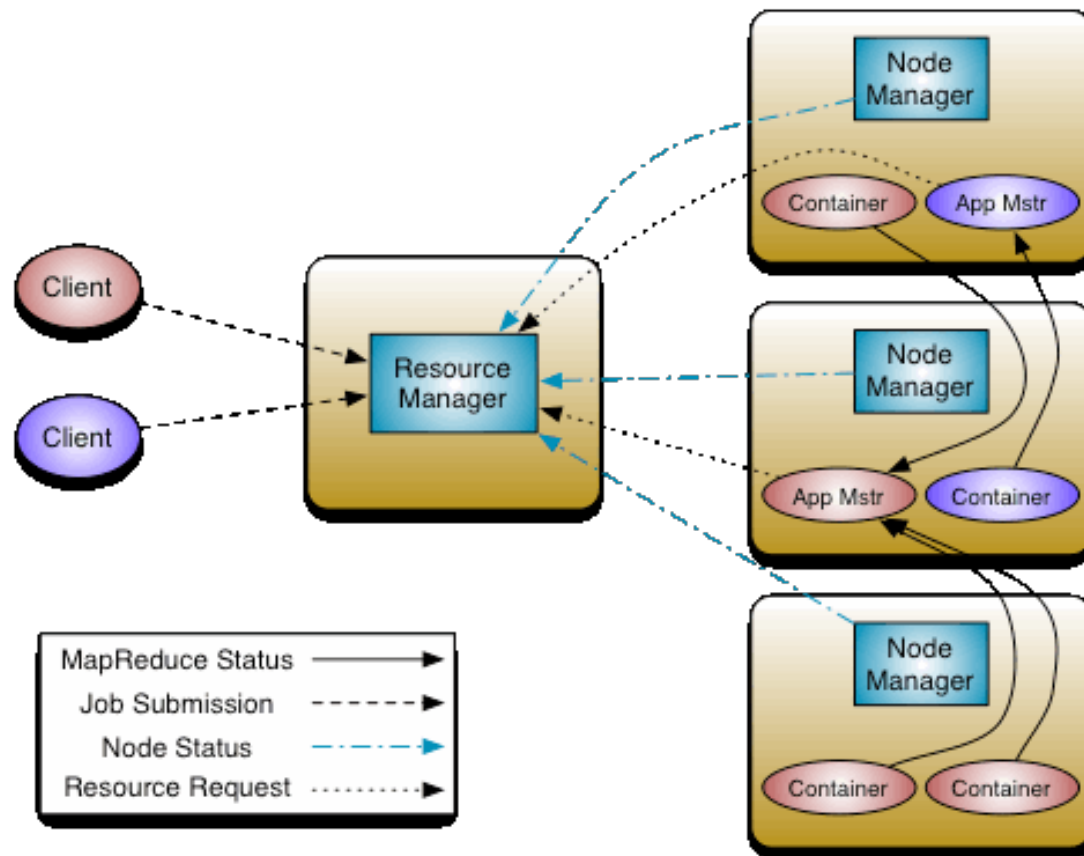
The Apache Hadoop Stack



Hadoop Structure

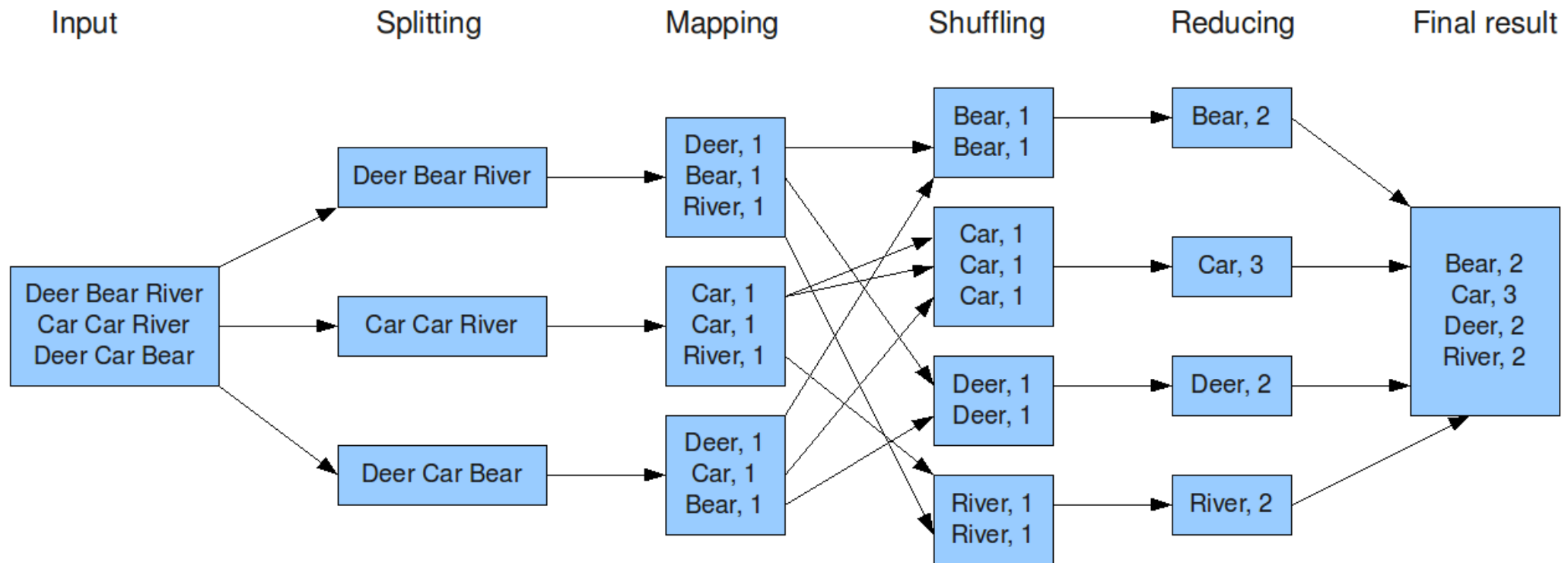


Hadoop Job Control



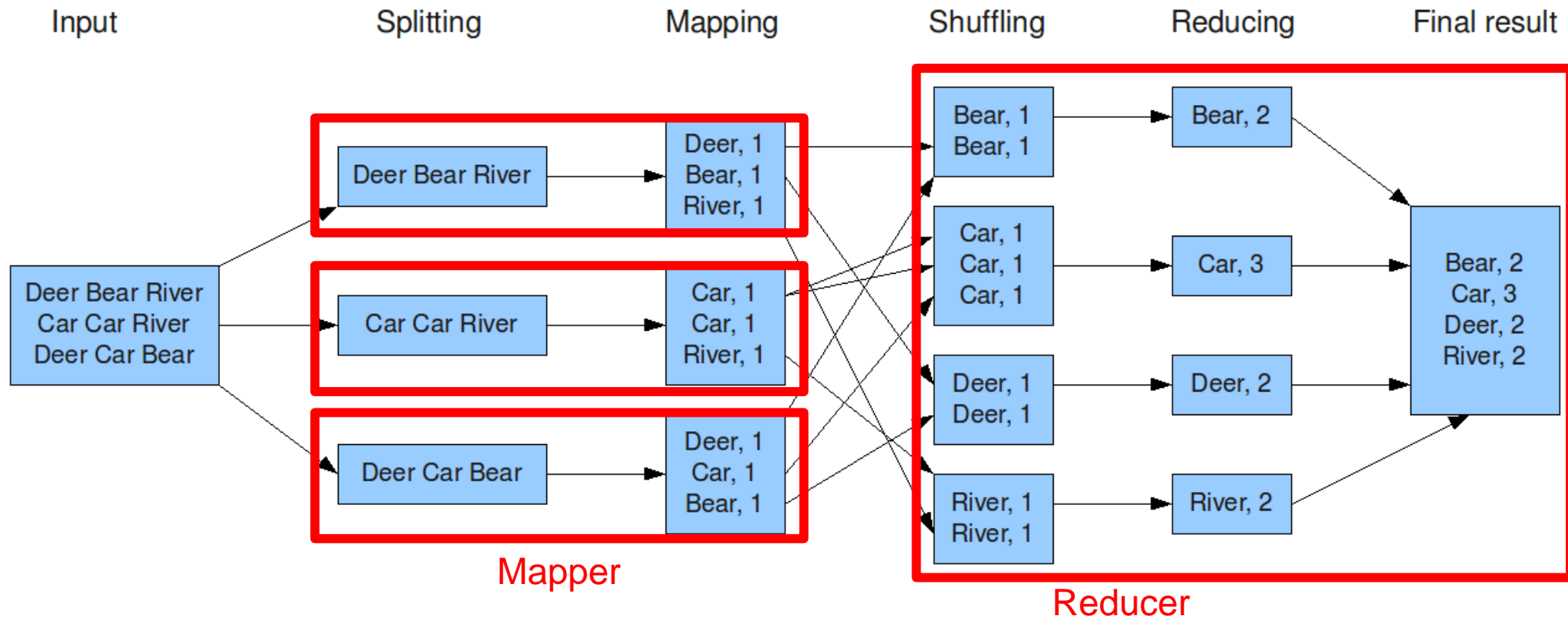
MAPREDUCE – WORD COUNT

The overall MapReduce word count process



MAPREDUCE – WORD COUNT

The overall MapReduce word count process



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