제2부 : Hadoop & Spark

1. Hadoop 개요 및 Master 서버 설정

목 차

1. Hadoop 시스템 개요
2. Master 서버 생성 및 환경 설정

Hadoop+Hive+Spark 수업 목적

- 최신 버전 Linux server 기반 Cluster 구축
- Hadoop 설치와 환경설정 및 구동
- Hive SQL 활용 Hadoop 데이터 연동
- Spark 기반 분산 데이터 처리

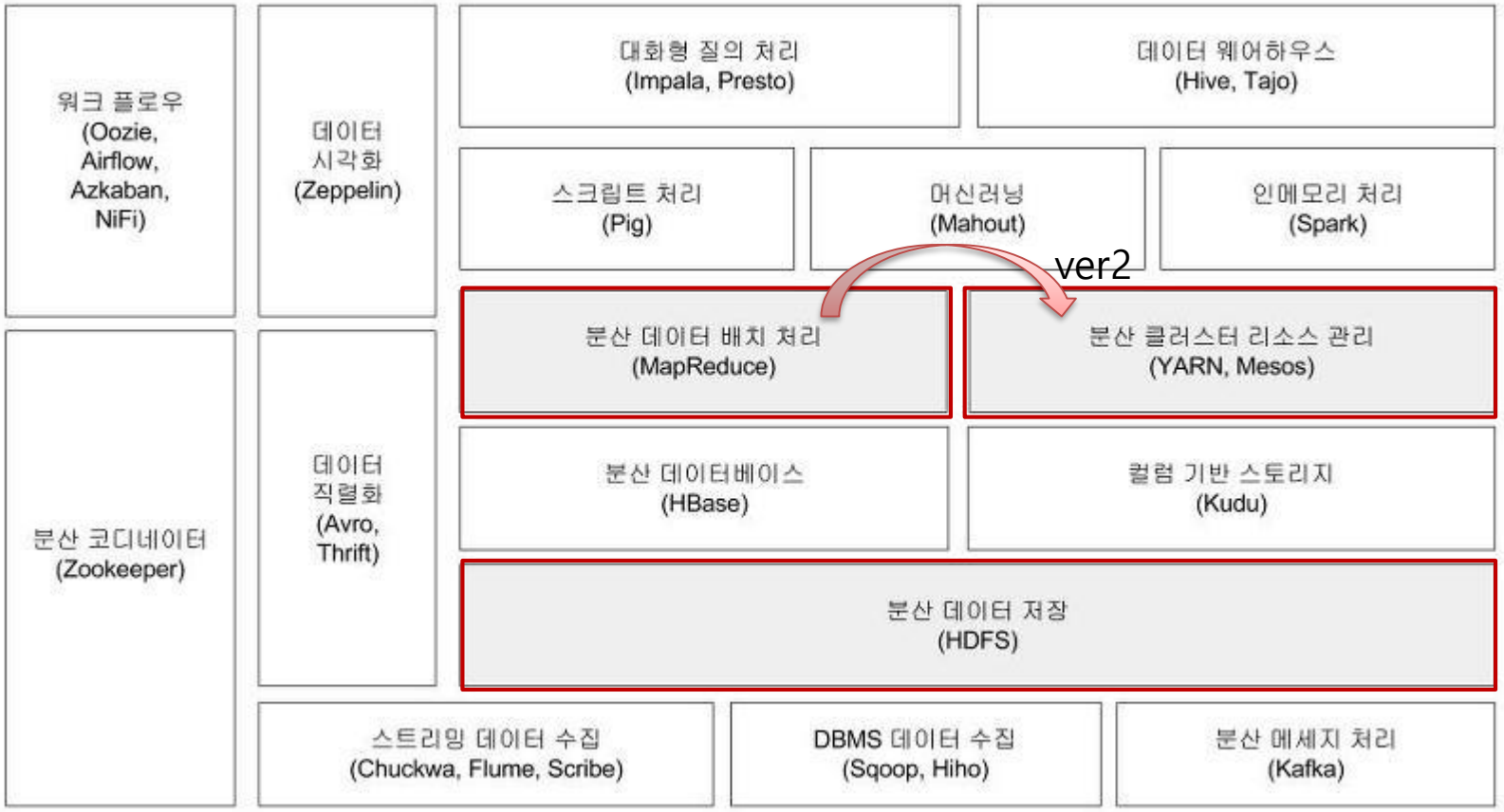
Hadoop 시스템 개요

- 구글 : GFS(Google File System)와 맵리듀스(MapReduce)
- 2005년 더그커팅 구현
- 2008년 아파치 최상위 프로젝트 승격
- 주요 구성
  - **HDFS(Hadoop Distributed File System) : 저장소**
  - **Map Reduce : 분산처리 시스템**

Hadoop 이점

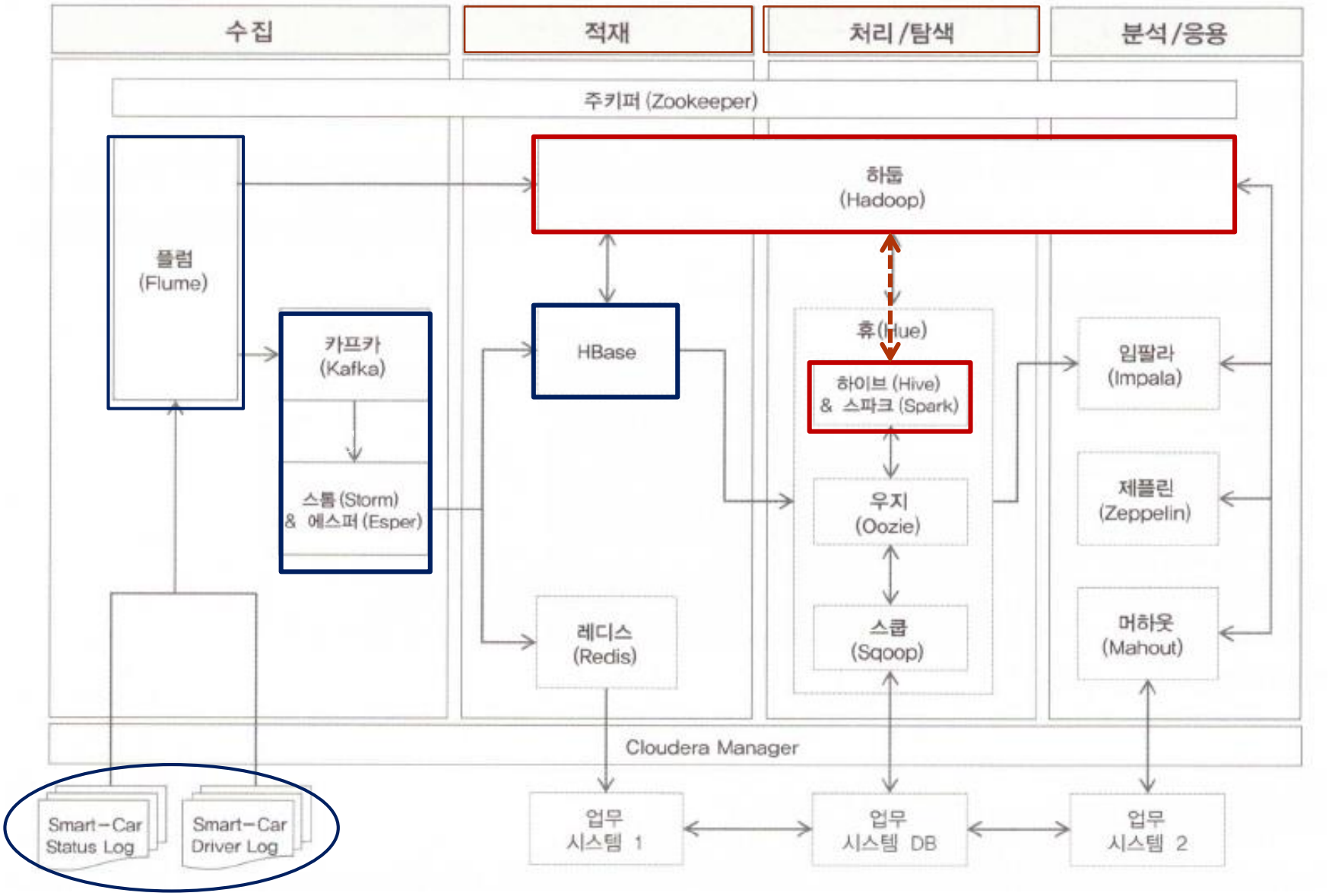
- 여러 대 서버에 데이터 저장
- 각 서버에서 동시 데이터 처리(분산처리)
- 기존의 RDBMS 대체
- 고가 장비 대신 리눅스 서버로 대체
- 오픈 소스
  - SW 라이선스, 고가 HW 비용 절감

Hadoop 에코 시스템

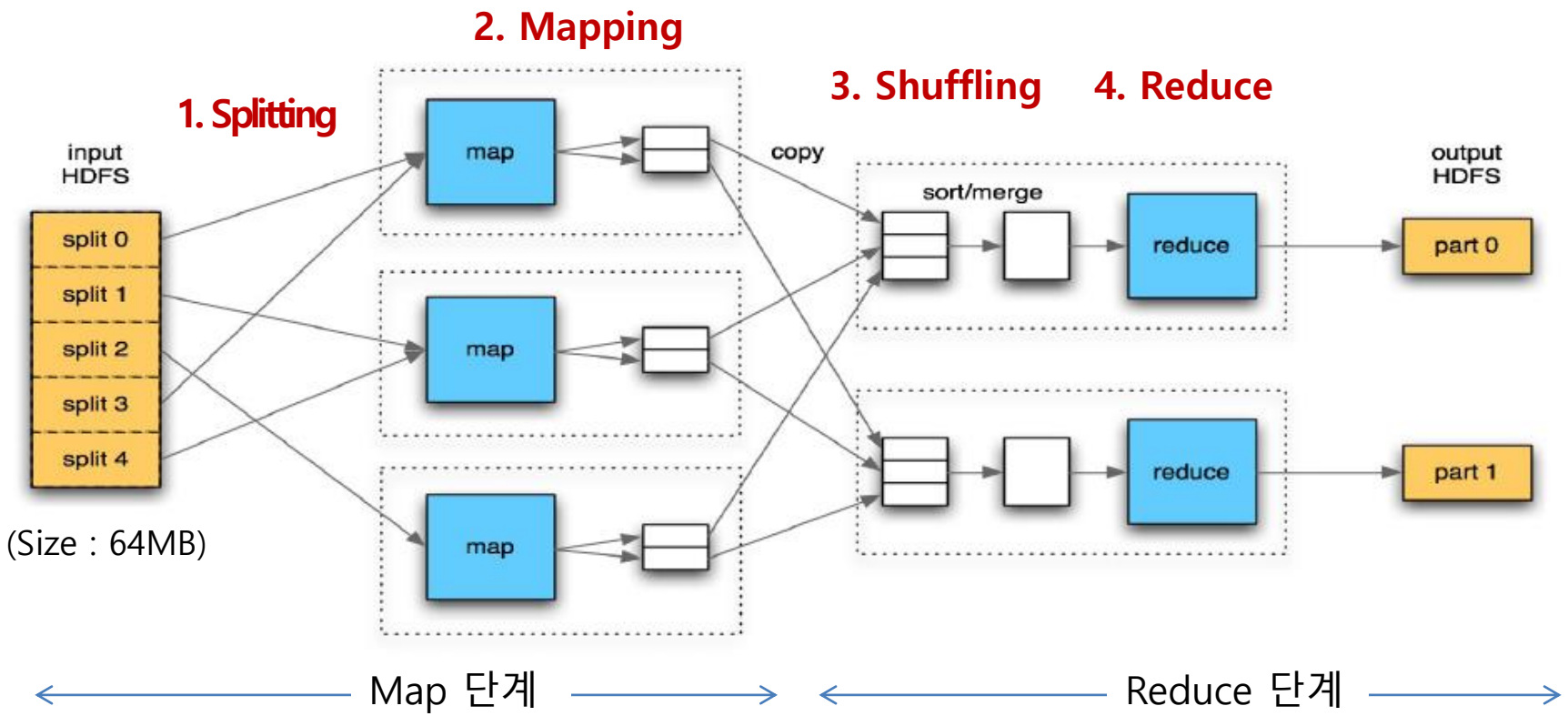


시작하세요. 하둡프로그래밍 – 위키북스 참고

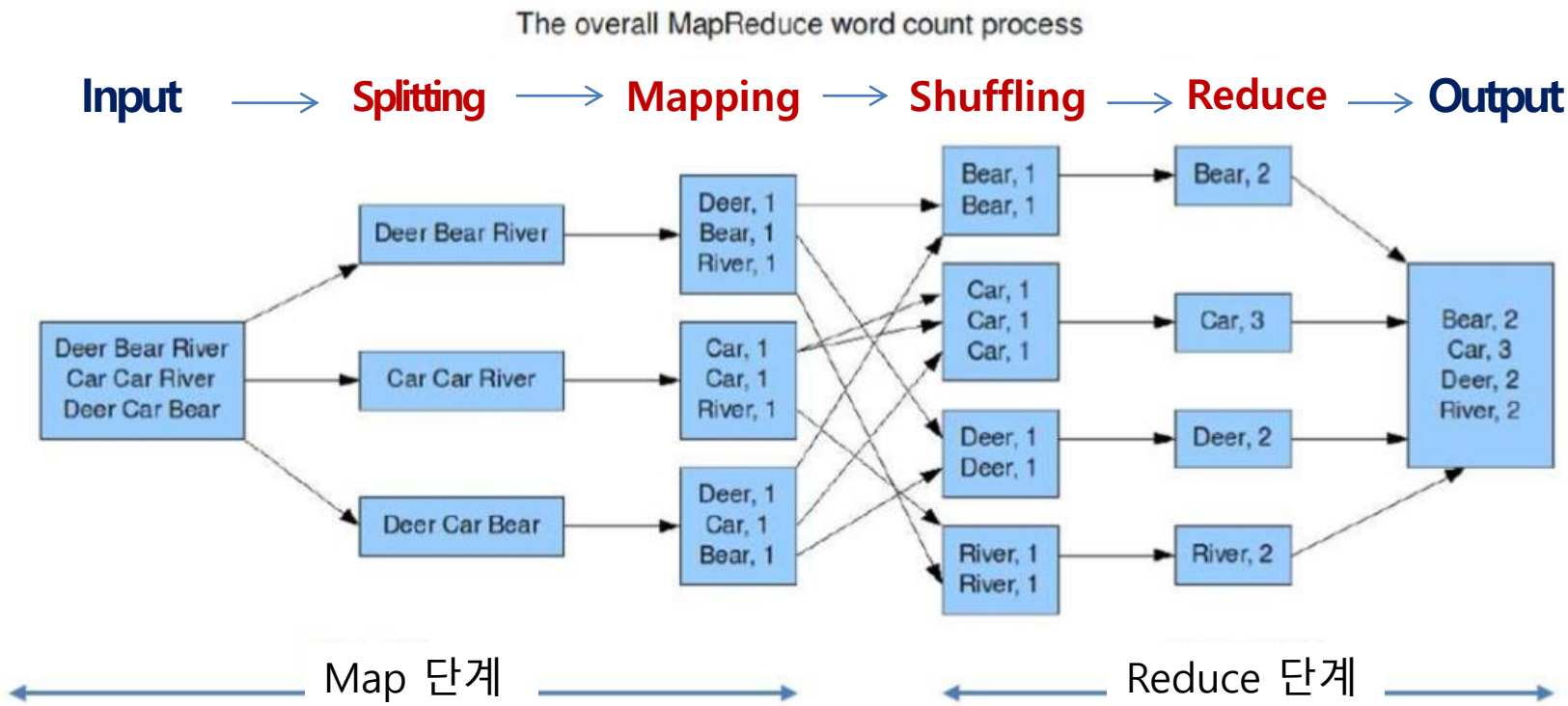
Hadoop+Hive+Spark 소프트웨어 아키텍처



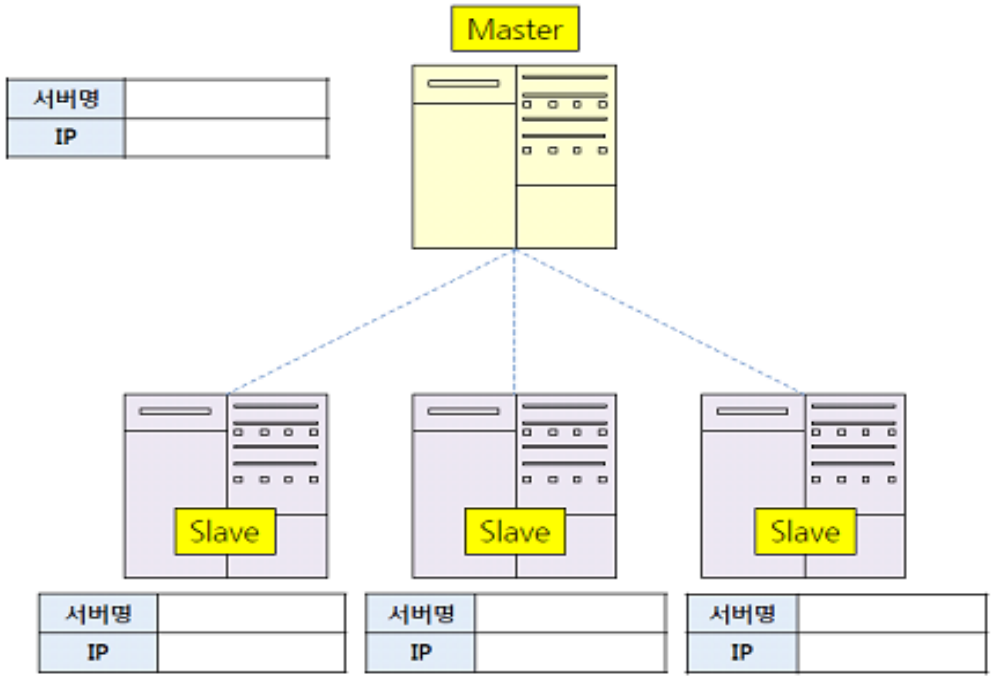
Hadoop 분산 병렬 처리



Word Count 예


The overall MapReduce word count process
Input
Splitting
Mapping
Shuffling
Reduce
Output
Deer Bear River
Car Car River
Deer Car Bear
Deer Bear River
Car Car River
Deer Car Bear
Deer, 1
Bear, 1
River, 1
Car, 1
Car, 1
River, 1
Deer, 1
Car, 1
Bear, 1
Bear, 1
Bear, 1
Car, 1
Car, 1
Car, 1
Deer, 1
Deer, 1
River, 1
River, 1
Bear, 2
Car, 3
Deer, 2
River, 2
Bear, 2
Car, 3
Deer, 2
River, 2
Map 단계
Reduce 단계

Hadoop 클러스터 구축 시스템 구성도



IPV4 : 192.168.0.35

First Name Server
- hostname : master
- ip **192.168.0.5**

Secondary Name Server, Data1
- hostname : slave1
- ip **192.168.0.10**

Data2
- hostname : slave2
- ip **192.168.0.11**

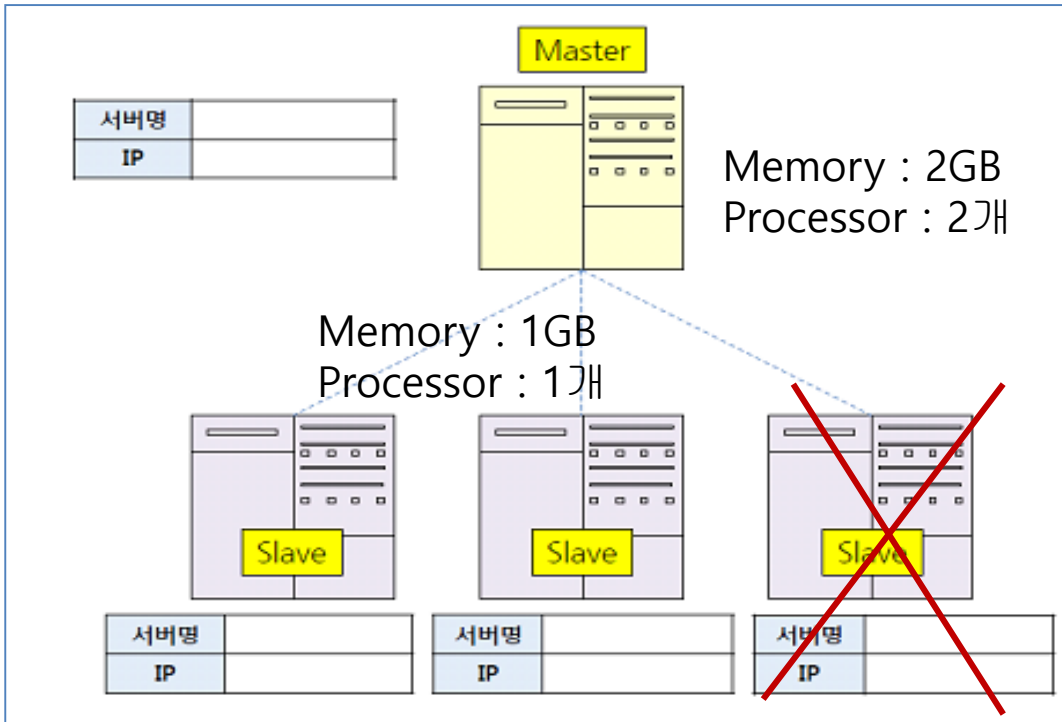
Data3
- hostname : slave3
- ip **192.168.0.12**

네트워크 카드 설정
- netmask : **255.255.255.0**
- gateway : **기본 라우팅 참고**

네임서버(DNS)

기본 네임서버 참고

Hadoop 클러스터 구축 시스템 구성도



Linux
실제 IP

IPV4 : 192.168.0.35

First Name Server
- hostname : master
- ip **192.168.0.5**

Secondary Name Server, Data1
- hostname : slave1
- ip **192.168.0.10**

Data2
- hostname : slave2
- ip **192.168.0.11**

Data3
- ~~hostname : slave3~~
- ~~ip : 10.0.2.12~~

네트워크 카드 설정

netmask : **255.255.255.0**

gateway : **기본 라우팅 참고**

네임서버(DNS)

기본 네임서버 참고

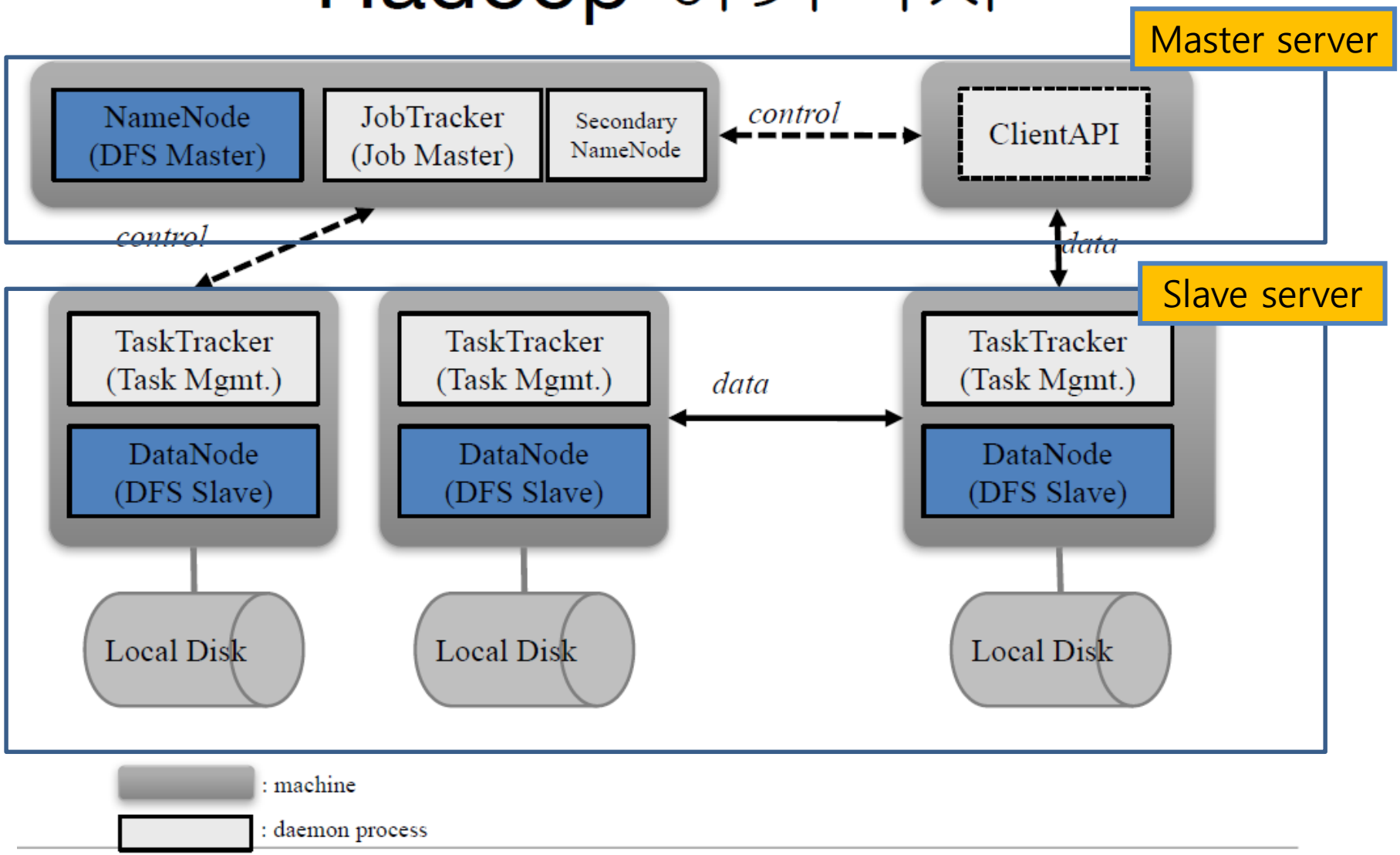
Master vs Slave server

Master Server	Slave Server
➢ Name node, job Tracker, 보조 네임노드 설치 서버 ➢ 디스크 : 2 ~ 4개 ➢ 메모리 : 32GB ~ 128GB ✓ 64GB : 1억 파일 저장 ➢ CPU : 코어 수 16~24개	➢ Data node, Task Tracker 설치 서버 ➢ 디스크 : 4 ~ 12개(1TB~3TB) ➢ 메모리 : 24GB ~ 48GB ➢ CPU : 쿼드 코어 2개 이상

✓ Job Tracker : Job 실행 요청 받고, Task Tracker에 할당

✓ Task Tracker : 실제 Task 실행, Job Tracker 진행 상태 보고

Hadoop 아키텍처


Master server
NameNode
(DFS Master)
JobTracker
(Job Master)
Secondary
NameNode
control
ClientAPI
control
data
Slave server
TaskTracker
(Task Mgmt.)
DataNode
(DFS Slave)
TaskTracker
(Task Mgmt.)
DataNode
(DFS Slave)
data
TaskTracker
(Task Mgmt.)
DataNode
(DFS Slave)
Local Disk
Local Disk
Local Disk
: machine
: daemon process

Name node(Master Server)

- HDFS에서의 NameNode(master)는 분산환경에서 Job 분배, 지시 및 감독
- 작업의 대상이 되는 파일을 블록(block)단위로 나누어서 slave node 분배
- 메타데이터 저장
  - ✓ 파일 별 블록, 각 블록 별 정보(DataNode 위치 등)
  - ✓ 실제 파일의 블록은 DataNode에 저장됨
- 전체 분산 파일시스템 이상 유무 체크, DataNode(slave) 데이터 입출력 작업 (low-level I/O tasks) 지시
- NameNode 문제 발생 사 Hadoop 클러스터 전체 시스템 차단
  - ✓ Secondary NameNode 이용(NameNode 이중화) 문제 해결

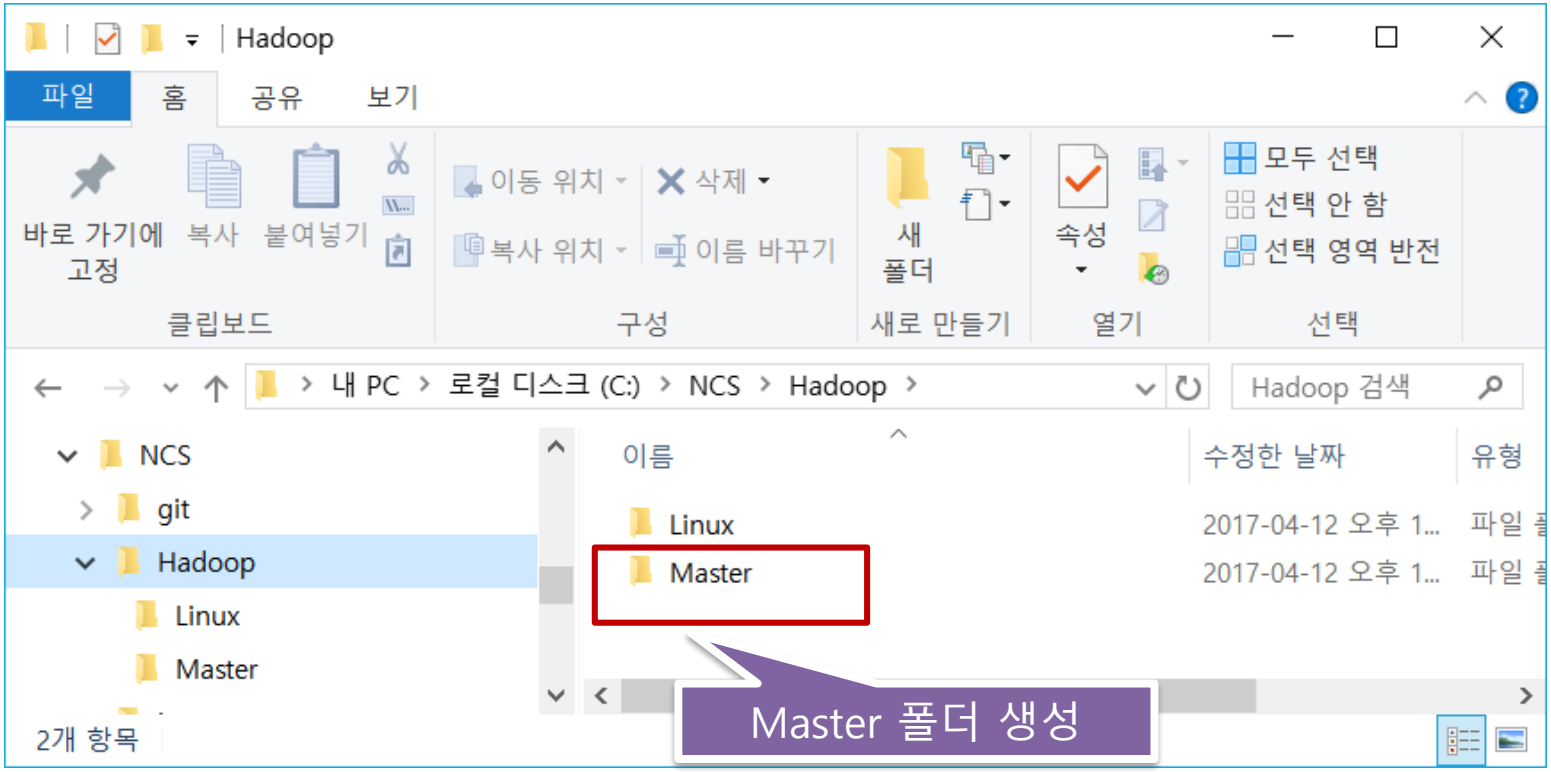
Data node(Slave Server)

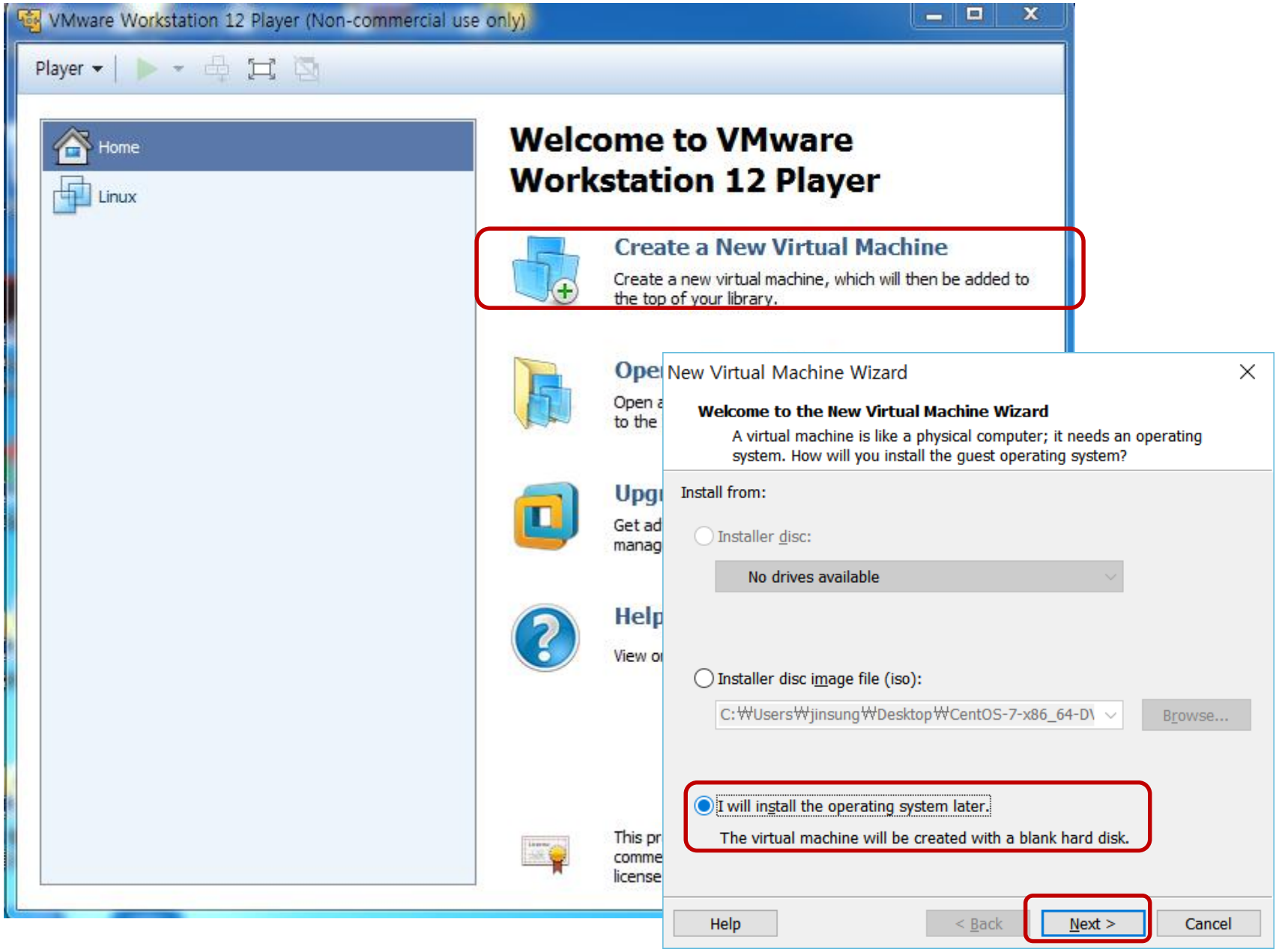
- 실제 HDFS(분산파일시스템) 대상으로 읽기 및 쓰기의 모든 작업 수행
- 모든 작업은 대상 파일을 random하게 블록(block) 단위로 나누어 진행
- DataNode 작업은 NameNode에 수시 보고, 메타데이터 형식 저장
- 데이터를 읽어 들이는 즉시 각 Data node에 분배(Name node)
  - ✓ HDFS는 큰 데이터 파일을 여러 개로 분리시켜서 각 node 처리
- 각각의 조각(chunk) 는 여러 대의 컴퓨터에 중복적으로 복제
  - ✓ 한 컴퓨터에서 장애가 발생해도 다른 컴퓨터를 통해 데이터 이용
- 모든 파일조각들은 하나의 namespace를 공유
  - ✓ 클러스터 내의 모든 node들은 공유된 파일 이용

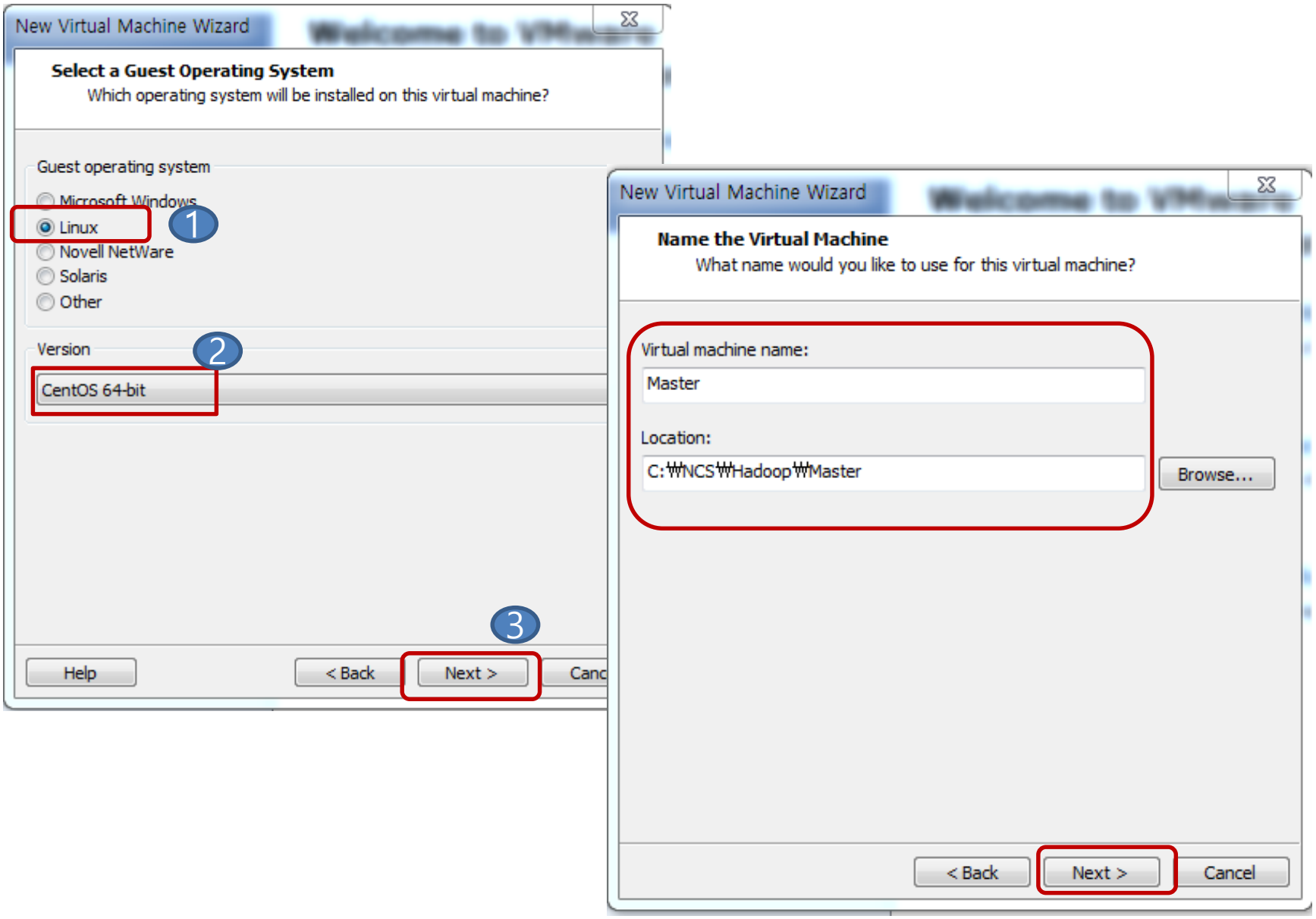
2. Master 서버 생성/환경 설정

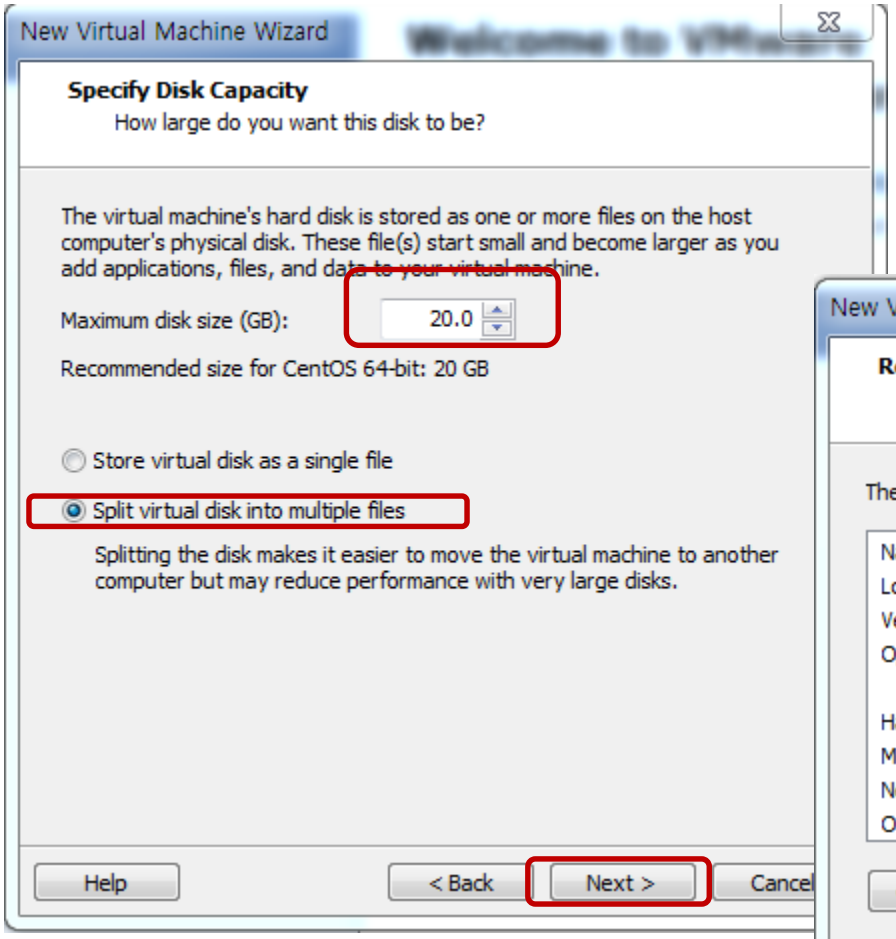
1) Master 서버 생성
2) 방화벽 제거
3) 네트워크 설정
4) 호스트 네임설정
5) Java 설치
6) Slave 폴더 생성(Master 폴더 복사)

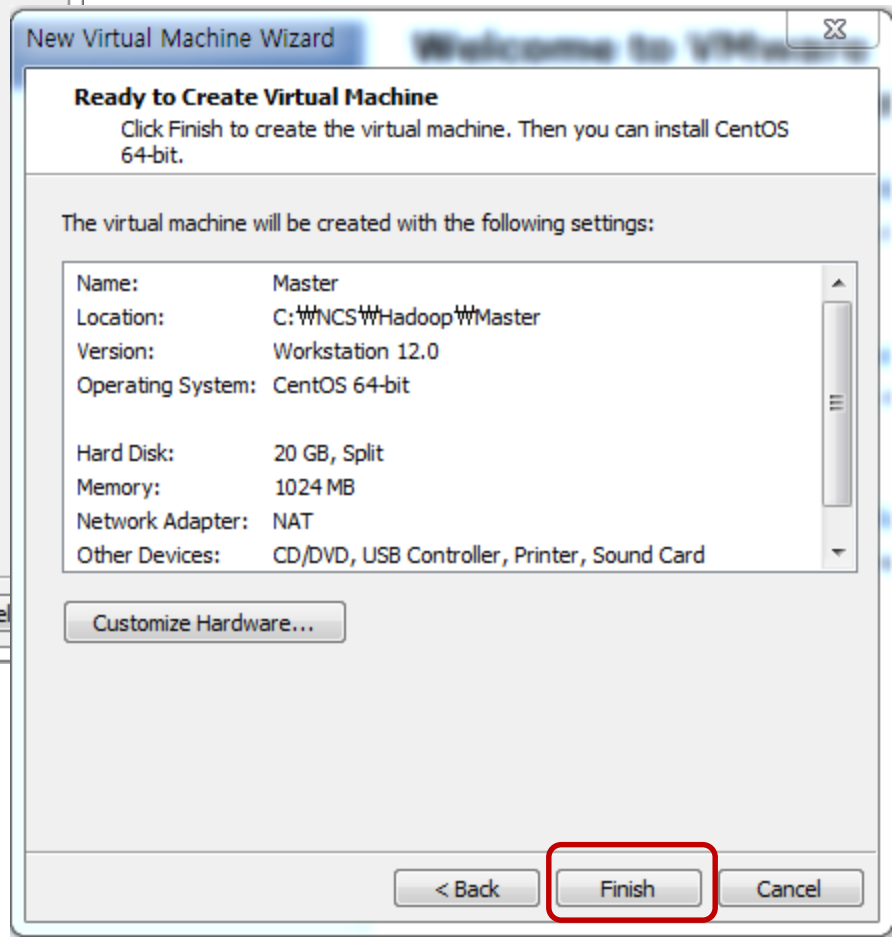
1) Master 서버 생성

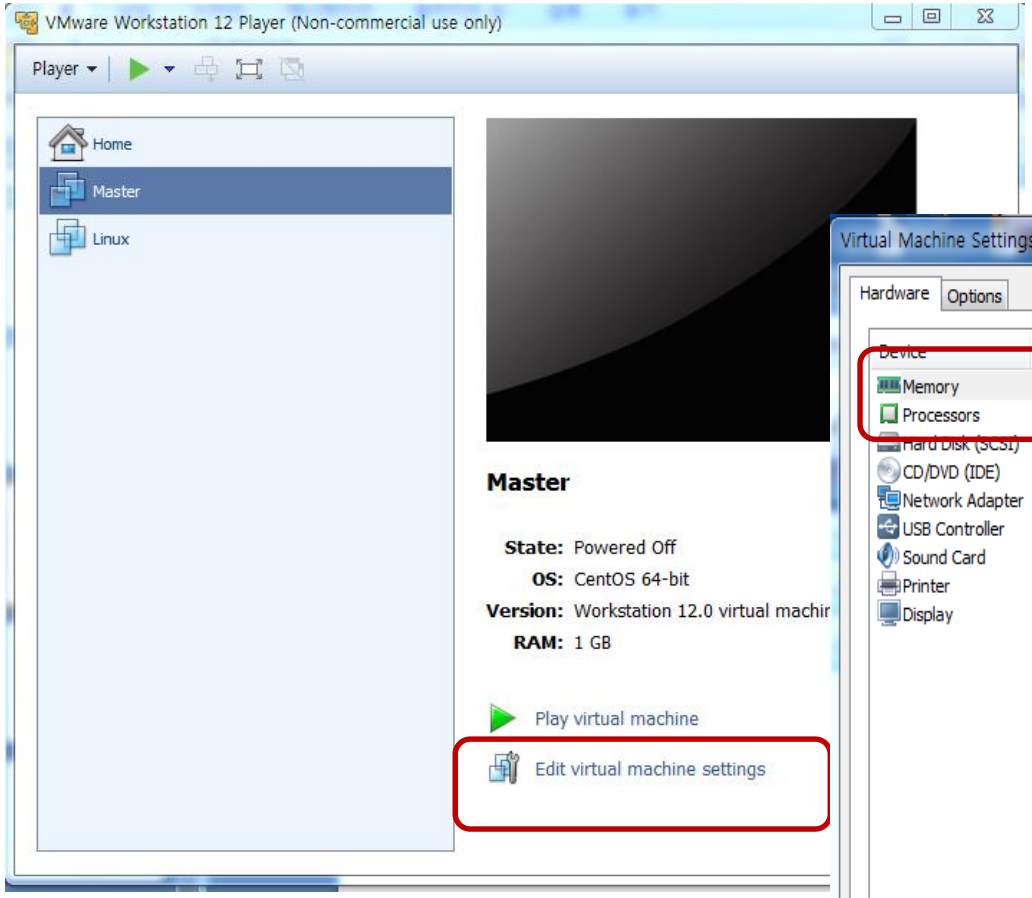



VMware Workstation 12 Player (Non-commercial use only)
Player
Home
Linux
Welcome to VMware Workstation 12 Player
Create a New Virtual Machine
Create a new virtual machine, which will then be added to the top of your library.
New Virtual Machine Wizard
Welcome to the New Virtual Machine Wizard
A virtual machine is like a physical computer; it needs an operating system. How will you install the guest operating system?
Install from:
Installer disc:
No drives available
Installer disc image file (iso):
C:₩Users₩jinsung₩Desktop₩CentOS-7-x86_64-D
Browse...
I will install the operating system later.
The virtual machine will be created with a blank hard disk.
Help
< Back
Next >
Cancel

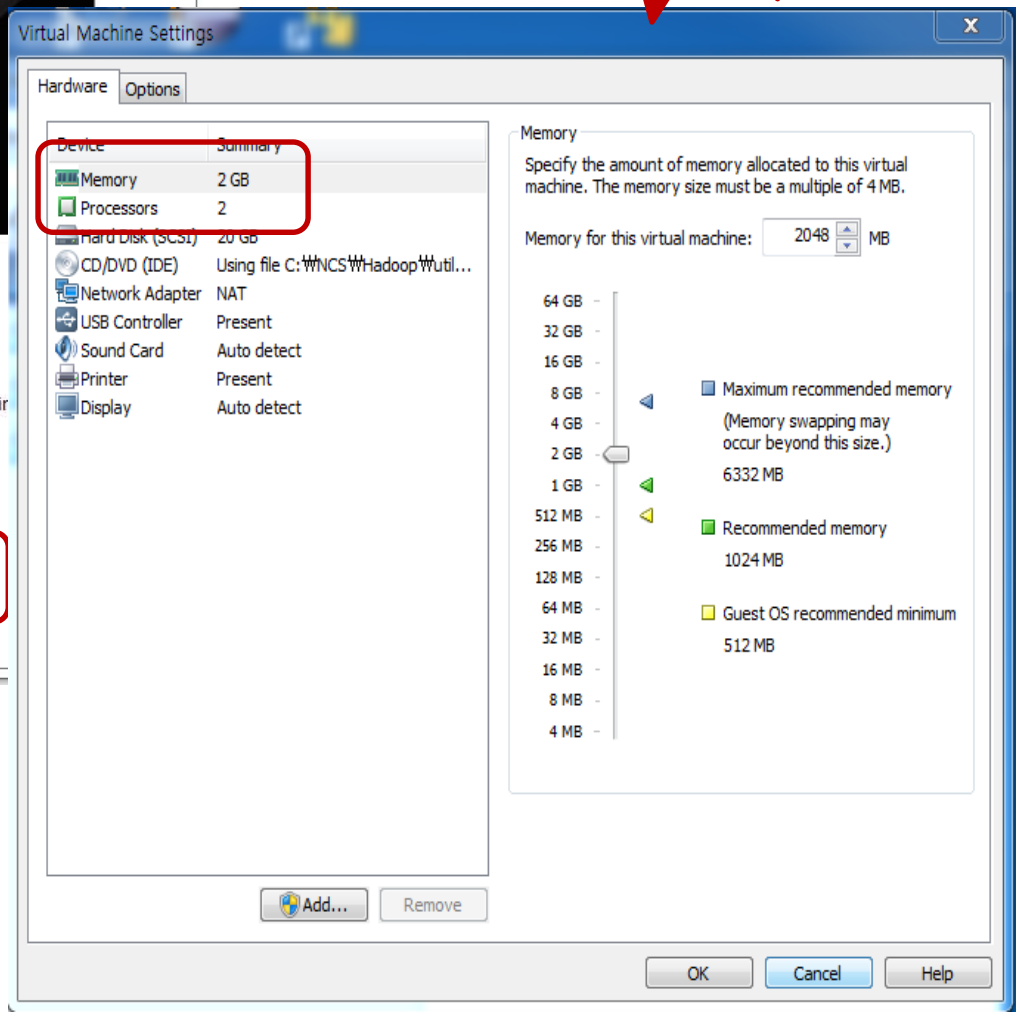

New Virtual Machine Wizard
Select a Guest Operating System
Which operating system will be installed on this virtual machine?
Guest operating system
Microsoft Windows
Linux
Novell NetWare
Solaris
Other
1
Version
2
CentOS 64-bit
3
Help
< Back
Next >
Canc
New Virtual Machine Wizard
Name the Virtual Machine
What name would you like to use for this virtual machine?
Virtual machine name:
Master
Location:
C:₩NCS₩Hadoop₩Master
Browse...
< Back
Next >
Cancel

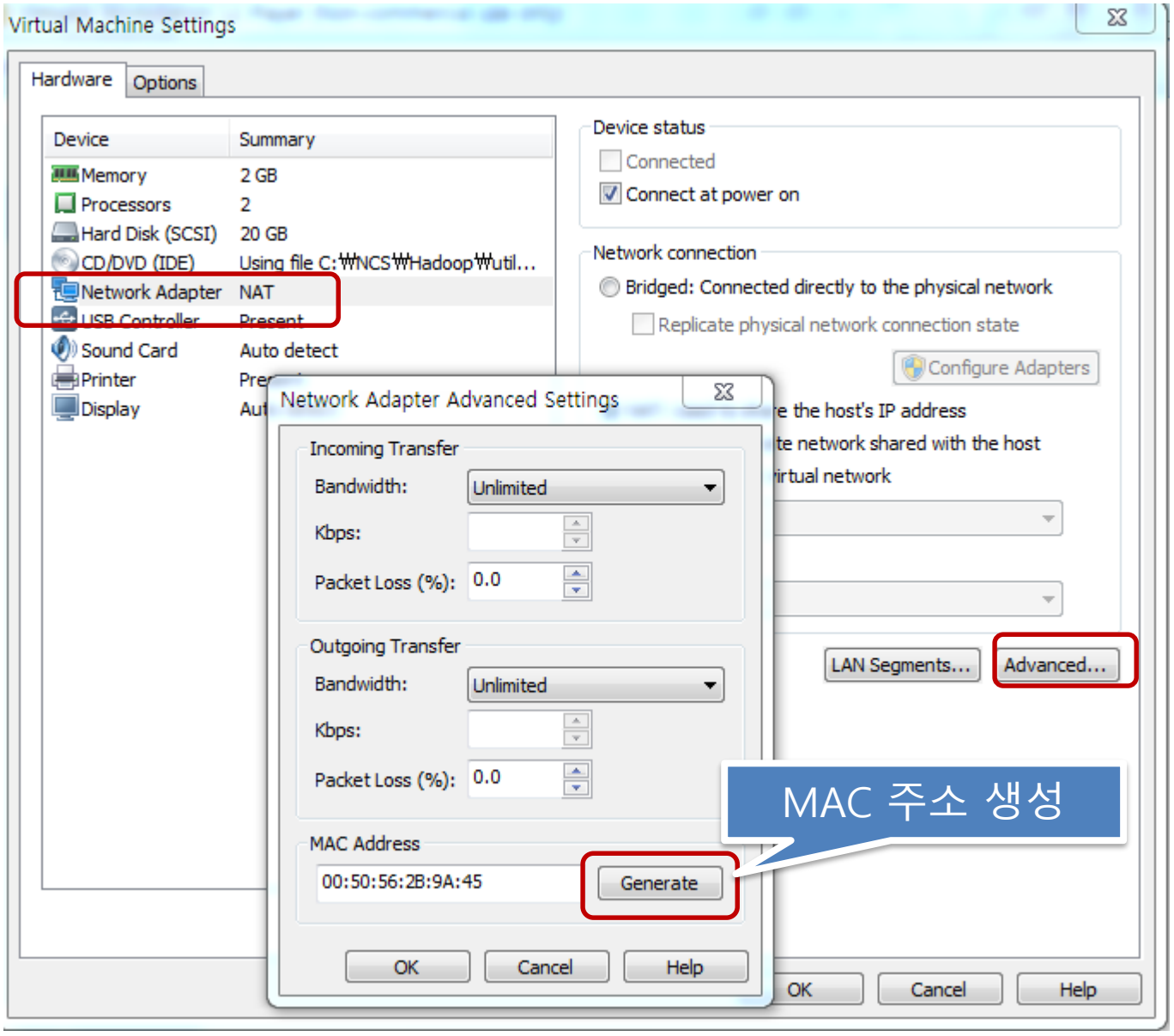

New Virtual Machine Wizard
Specify Disk Capacity
How large do you want this disk to be?
The virtual machine's hard disk is stored as one or more files on the host computer's physical disk. These file(s) start small and become larger as you add applications, files, and data to your virtual machine.
Maximum disk size (GB):
20.0
Recommended size for CentOS 64-bit: 20 GB
Store virtual disk as a single file
Split virtual disk into multiple files
Splitting the disk makes it easier to move the virtual machine to another computer but may reduce performance with very large disks.
Help
< Back
Next >
Cancel

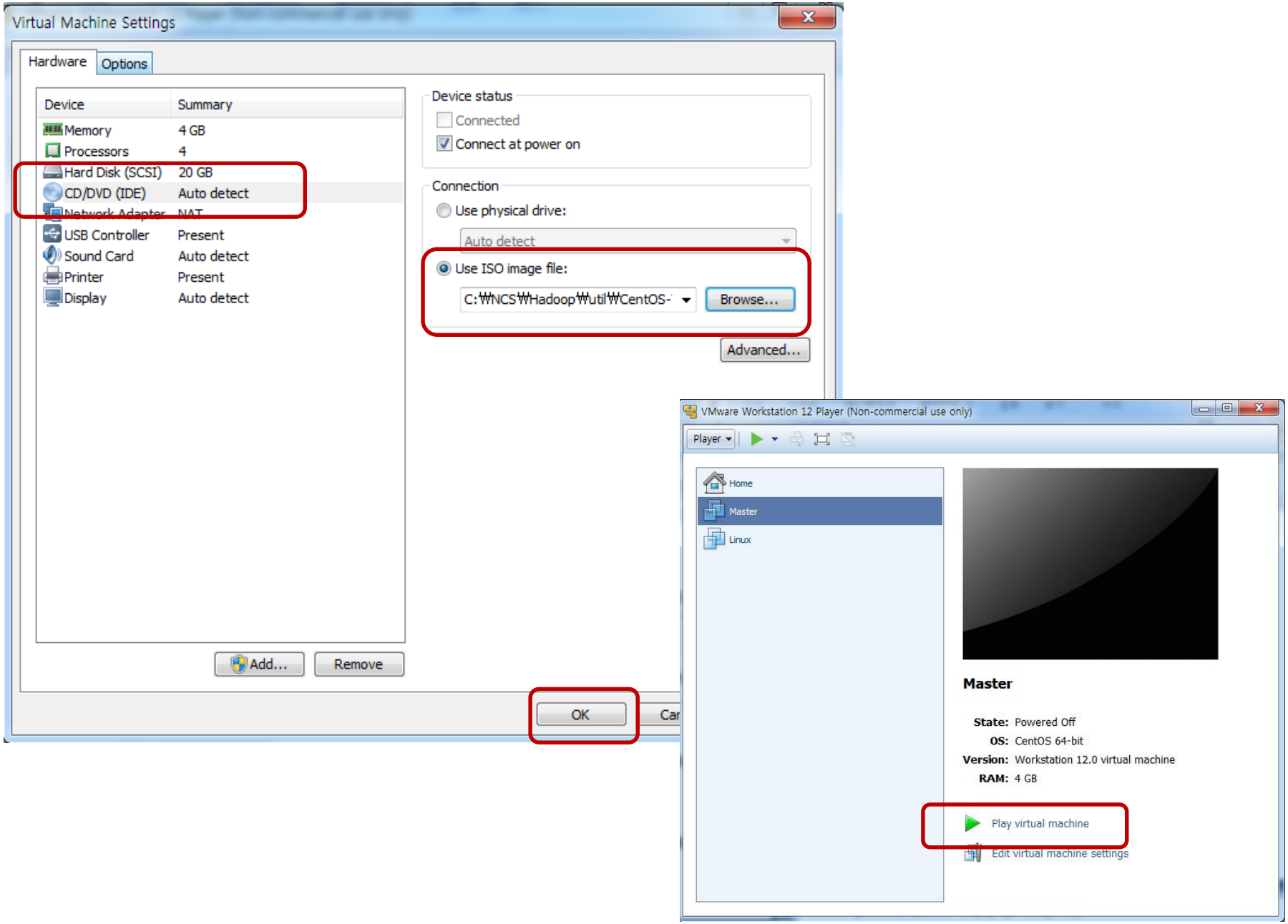

New Virtual Machine Wizard
Ready to Create Virtual Machine
Click Finish to create the virtual machine. Then you can install CentOS 64-bit.
The virtual machine will be created with the following settings:
Name: Master
Location: C:₩NCS₩Hadoop₩Master
Version: Workstation 12.0
Operating System: CentOS 64-bit
Hard Disk: 20 GB, Split
Memory: 1024 MB
Network Adapter: NAT
Other Devices: CD/DVD, USB Controller, Printer, Sound Card
Customize Hardware...
< Back
Finish
Cancel

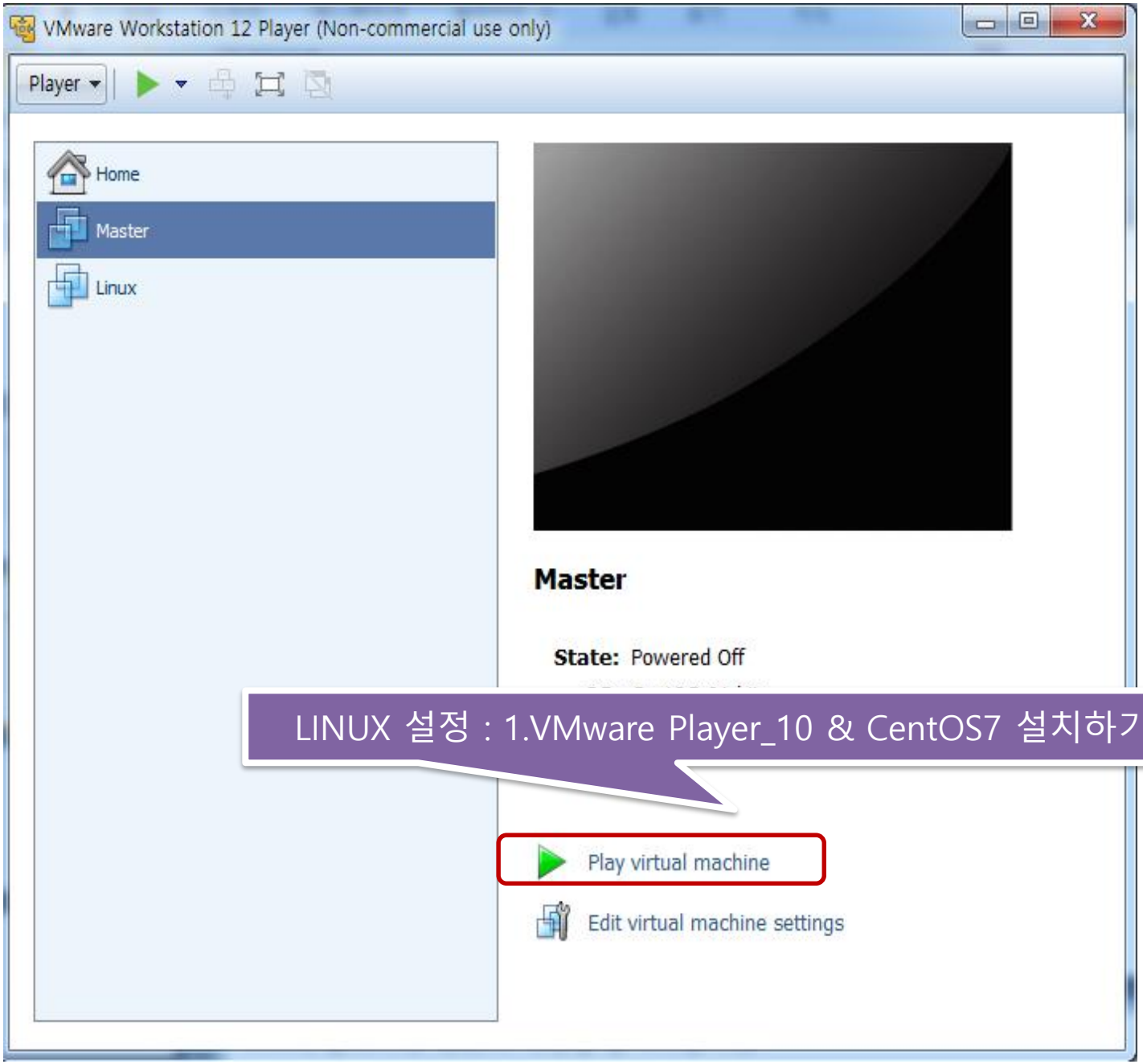


Memory 2GB
Processors 2

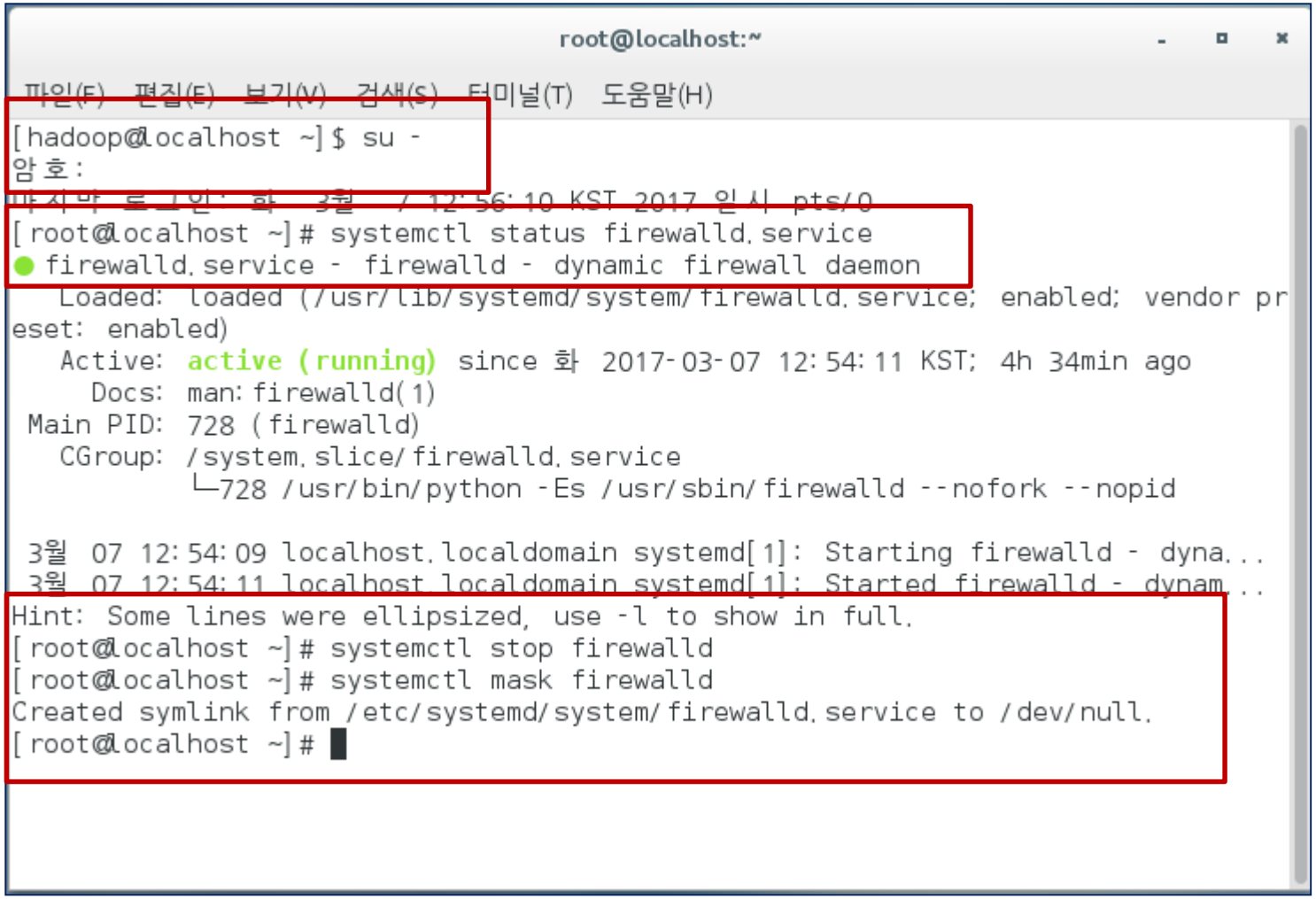



Virtual Machine Settings
Hardware
Options
Device
Summary
Memory
2 GB
Processors
2
Hard Disk (SCSI)
20 GB
CD/DVD (IDE)
Using file C:₩NCS₩Hadoop₩util...
Network Adapter
NAT
USB Controller
Present
Sound Card
Auto detect
Printer
Display
Device status
Connected
Connect at power on
Network connection
Bridged: Connected directly to the physical network
Replicate physical network connection state
Configure Adapters
LAN Segments...
Advanced...
OK
Cancel
Help
Network Adapter Advanced Settings
Incoming Transfer
Bandwidth:
Unlimited
Kbps:
Packet Loss (%):
0.0
Outgoing Transfer
Bandwidth:
Unlimited
Kbps:
Packet Loss (%):
0.0
MAC Address
00:50:56:2B:9A:45
Generate
OK
Cancel
Help
MAC 주소 생성

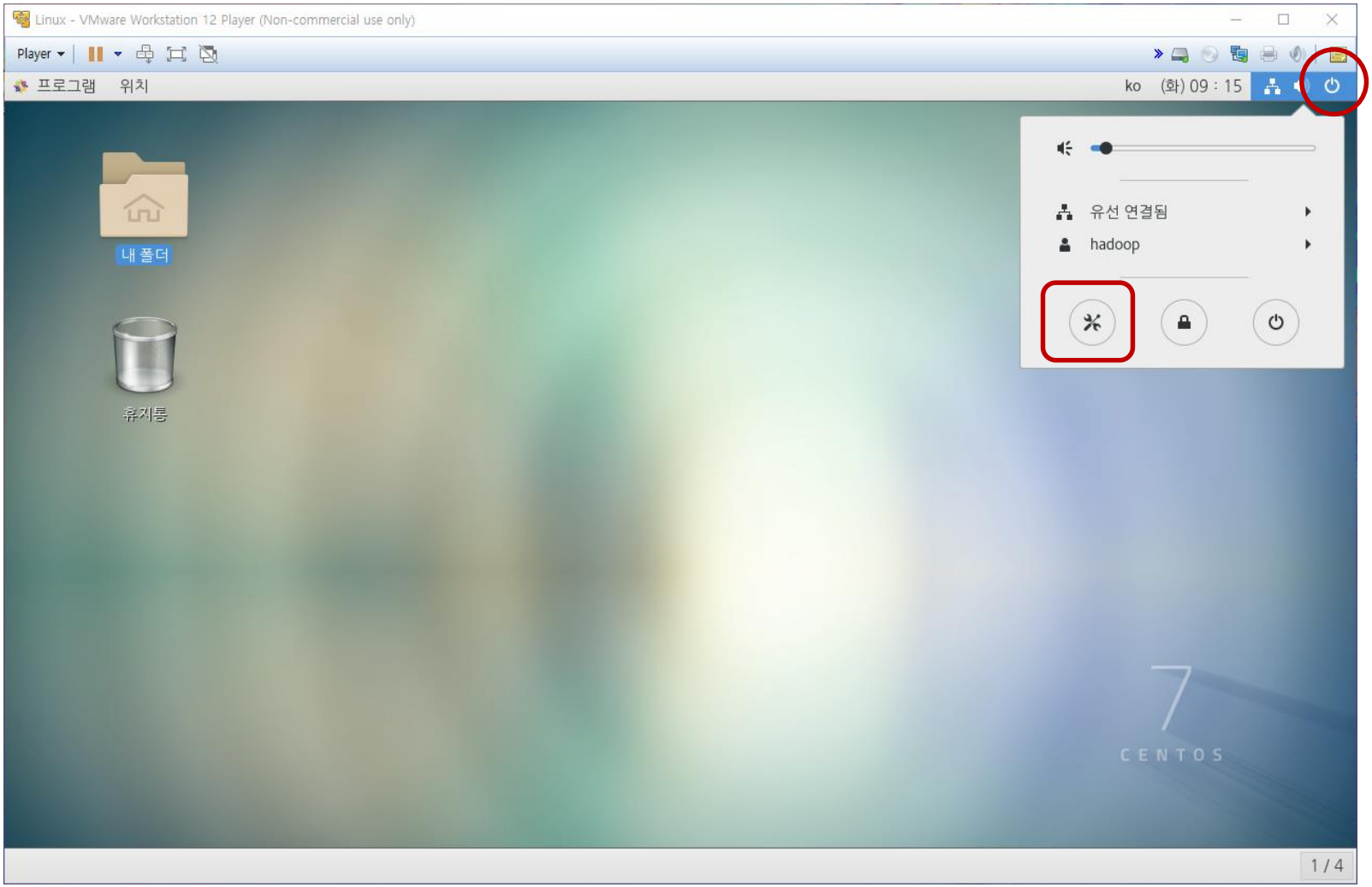

Virtual Machine Settings
Hardware
Options
Device
Summary
Memory
4 GB
Processors
4
Hard Disk (SCSI)
20 GB
CD/DVD (IDE)
Auto detect
Network Adapter
NAT
USB Controller
Present
Sound Card
Auto detect
Printer
Present
Display
Auto detect
Add...
Remove
Device status
Connected
Connect at power on
Connection
Use physical drive:
Auto detect
Use ISO image file:
C:₩NCS₩Hadoop₩util₩CentOS-
Browse...
Advanced...
OK
VMware Workstation 12 Player (Non-commercial use only)
Player
Home
Master
Linux
Master
State: Powered Off
OS: CentOS 64-bit
Version: Workstation 12.0 virtual machine
RAM: 4 GB
Play virtual machine
Edit virtual machine settings


VMware Workstation 12 Player (Non-commercial use only)
Player
Home
Master
Linux
Master
State: Powered Off
LINUX 설정 : 1.VMware Player_10 & CentOS7 설치하기 참고
Play virtual machine
Edit virtual machine settings

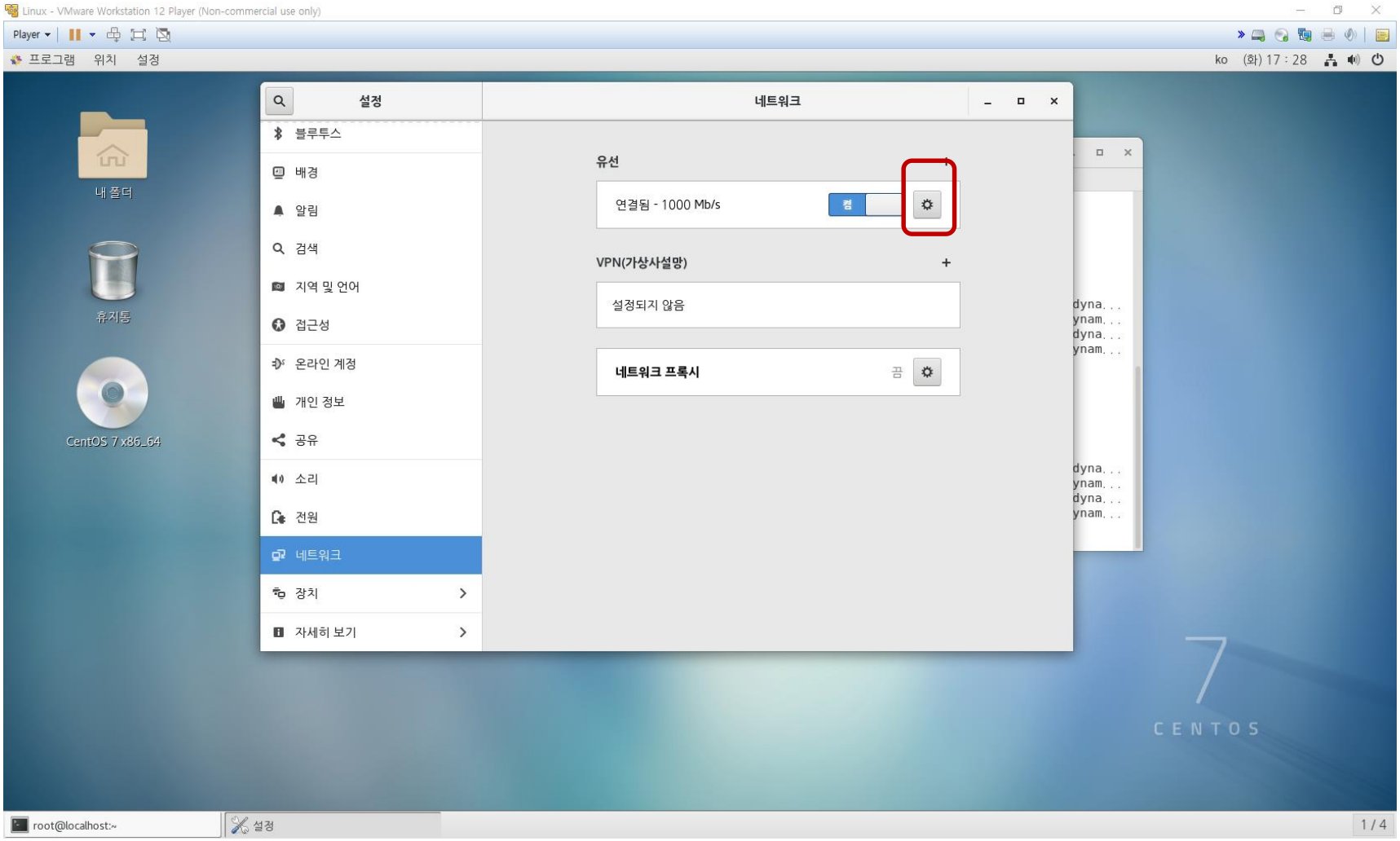
2) 방화벽 제거


root@localhost:~
파일(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)
[hadoop@localhost ~]$ su -
암호:
마지막 로그인: 화 3월 7 12:56:10 KST 2017 일시 pts/0
[root@localhost ~]# systemctl status firewalld.service
● firewalld.service - firewalld - dynamic firewall daemon
Loaded: loaded (/usr/lib/systemd/system/firewalld.service; enabled; vendor pr
eset: enabled)
Active: active (running) since 화 2017-03-07 12:54:11 KST; 4h 34min ago
Docs: man:firewalld(1)
Main PID: 728 (firewalld)
CGroup: /system.slice/firewalld.service
└─728 /usr/bin/python -Es /usr/sbin/firewalld --nofork --nopid
3월 07 12:54:09 localhost.localdomain systemd[1]: Starting firewalld - dyna...
3월 07 12:54:11 localhost.localdomain systemd[1]: Started firewalld - dynam...
Hint: Some lines were ellipsized, use -l to show in full.
[root@localhost ~]# systemctl stop firewalld
[root@localhost ~]# systemctl mask firewalld
Created symlink from /etc/systemd/system/firewalld.service to /dev/null.
[root@localhost ~]#

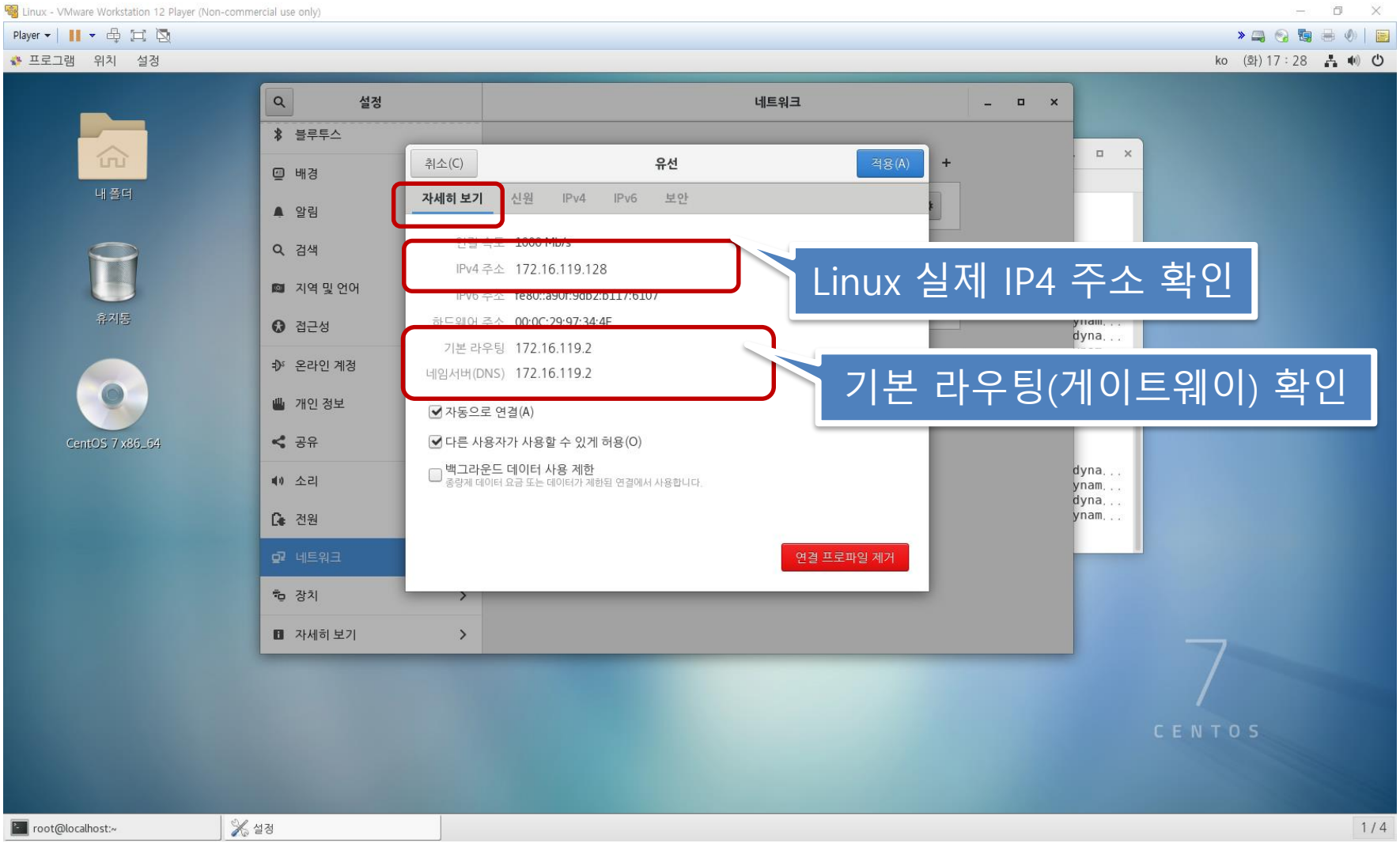
3) 네트워크 설정


Linux - VMware Workstation 12 Player (Non-commercial use only)
Player
프로그램
위치
ko
(화) 09 : 15
내 폴더
휴지통
유선 연결됨
hadoop
CENTOS
1 / 4

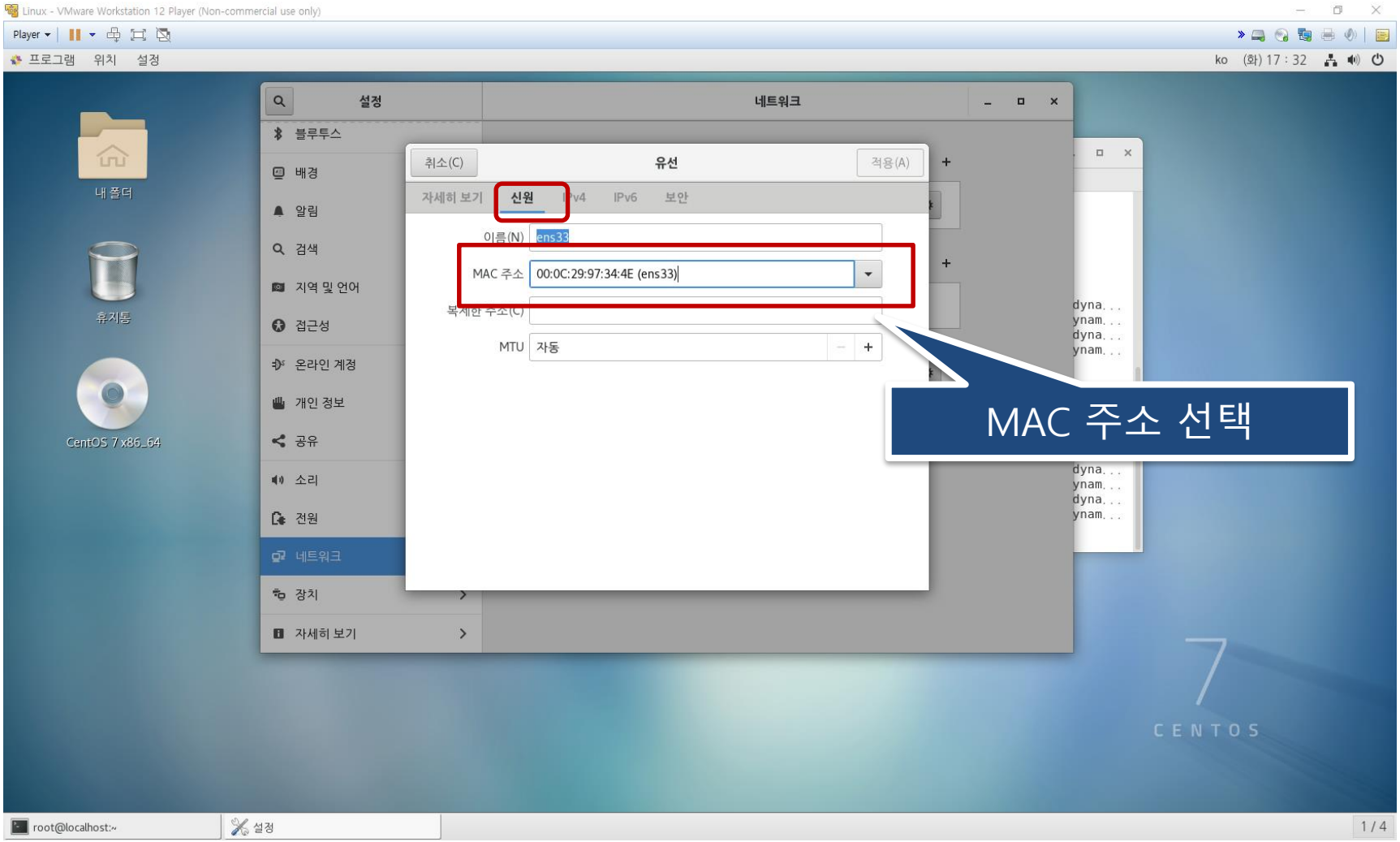
3) 네트워크 설정


Linux - VMware Workstation 12 Player (Non-commercial use only)
Player
프로그램 위치 설정
ko (화) 17 : 28
내 폴더
휴지통
CentOS 7 x86_64
설정
네트워크
블루투스
배경
알림
검색
지역 및 언어
접근성
온라인 계정
개인 정보
공유
소리
전원
네트워크
장치
자세히 보기
유선
연결됨 - 1000 Mb/s
켬
VPN(가상사설망)
설정되지 않음
네트워크 프록시
끔
CENTOS
root@localhost:~
설정
1 / 4

3) 네트워크 설정

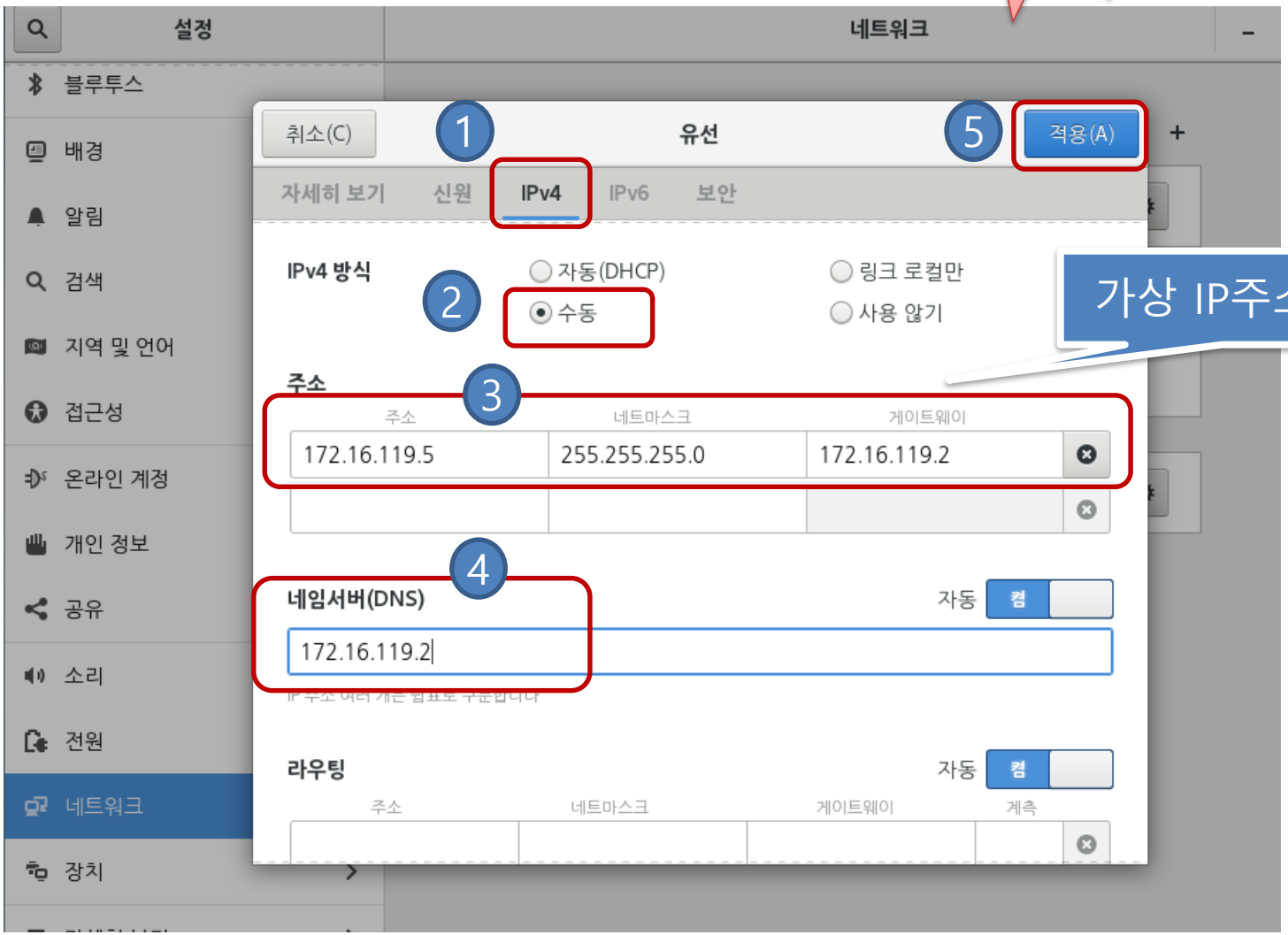

Linux - VMware Workstation 12 Player (Non-commercial use only)
Player
프로그램 위치 설정
ko (화) 17 : 28
내 폴더
휴지통
CentOS 7 x86_64
설정
네트워크
블루투스
배경
알림
검색
지역 및 언어
접근성
온라인 계정
개인 정보
공유
소리
전원
네트워크
장치
자세히 보기
취소(C)
유선
적용(A)
자세히 보기
신원
IPv4
IPv6
보안
IPv4 주소 172.16.119.128
기본 라우팅 172.16.119.2
네임서버(DNS) 172.16.119.2
자동으로 연결(A)
다른 사용자가 사용할 수 있게 허용(O)
백그라운드 데이터 사용 제한
연결 프로파일 제거
Linux 실제 IP4 주소 확인
기본 라우팅(게이트웨이) 확인
CENTOS
root@localhost:~
설정
1 / 4

3) 네트워크 설정

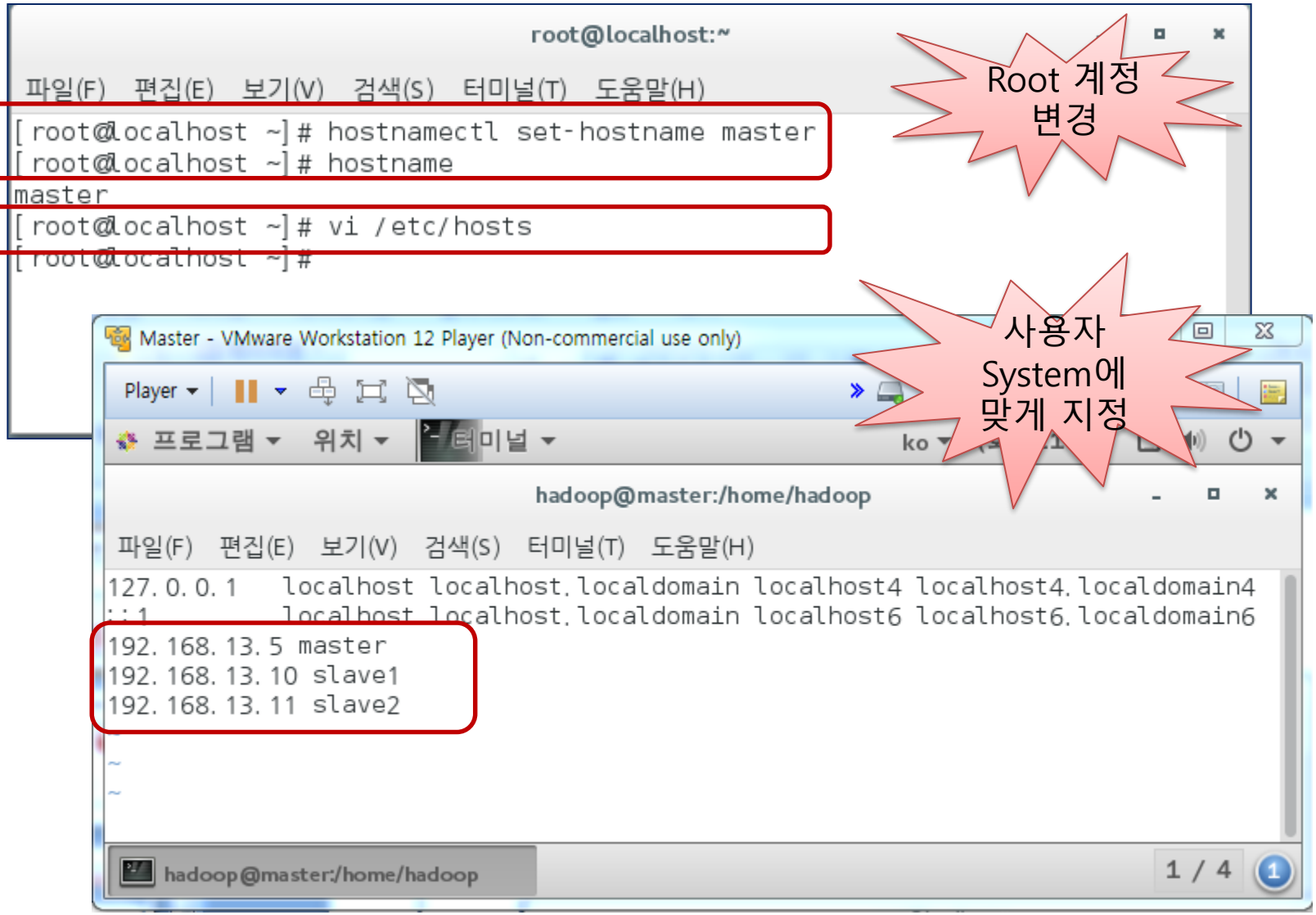

Linux - VMware Workstation 12 Player (Non-commercial use only)
Player
프로그램 위치 설정
ko (화) 17 : 32
내 폴더
휴지통
CentOS 7 x86_64
설정
네트워크
블루투스
배경
알림
검색
지역 및 언어
접근성
온라인 계정
개인 정보
공유
소리
전원
네트워크
장치
자세히 보기
취소(C)
유선
적용(A)
자세히 보기
신원
IPv4
IPv6
보안
이름(N) ens33
MAC 주소 00:0C:29:97:34:4E (ens33)
복제한 주소(C)
MTU 자동
MAC 주소 선택
CENTOS
root@localhost:~
설정
1 / 4

주소 : 실제IP에서 4번째 숫자 변경
게이트웨이 : 기본 라우팅 동일
네임서버 : 기본 네임 서버 동일

사용자 System에 맞게 지정



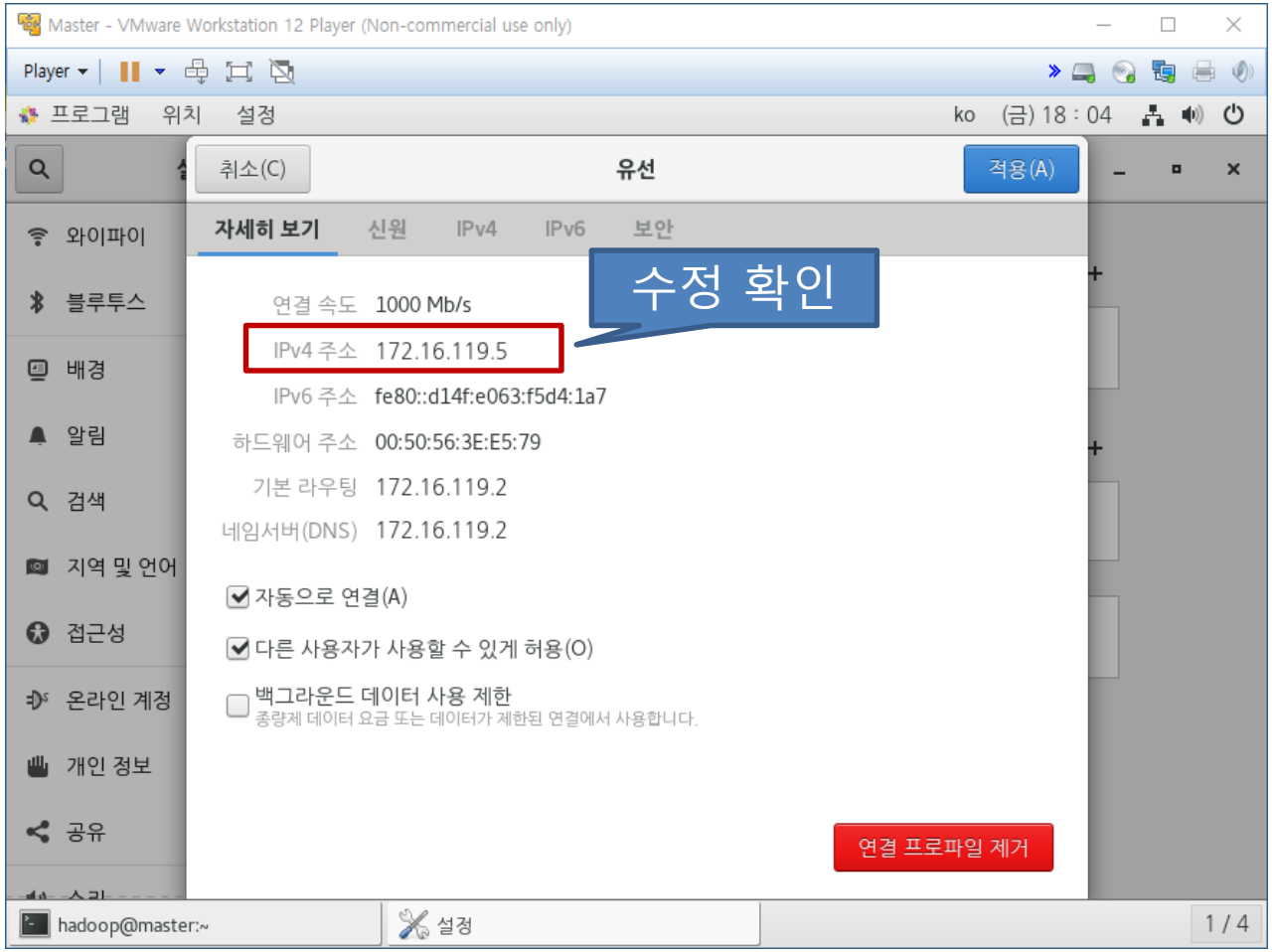
4) 호스트 네임 설정


root@localhost:~
파일(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)
[root@localhost ~]# hostnamectl set-hostname master
[root@localhost ~]# hostname
master
[root@localhost ~]# vi /etc/hosts
[root@localhost ~]#
Root 계정 변경
Master - VMware Workstation 12 Player (Non-commercial use only)
Player
프로그램 위치 터미널
ko
사용자 System에 맞게 지정
hadoop@master:/home/hadoop
파일(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)
127.0.0.1 localhost localhost.localdomain localhost4 localhost4.localdomain4
::1 localhost localhost.localdomain localhost6 localhost6.localdomain6
192.168.13.5 master
192.168.13.10 slave1
192.168.13.11 slave2
hadoop@master:/home/hadoop
1 / 4

인터넷 연결 확인


Linux - VMware Workstation 12 Player (Non-commercial use only)
Player
프로그램
위치
Firefox 웹 브라우저
ko
(금) 09:32
NAVER – Mozilla Firefox
NAVER
https://www.naver.com
검색
NAVER
메일 카페 블로그 지식iN 쇼핑 Pay TV 사전 뉴스 증권 부동산 지도 영화 뮤직 책 웹툰 더보기
야근에 지친 30대에게 떠먹여주는 혜택
BMW 30분 2천 원대
SOCAR
Connect with people
Forgot Username or Pa
연합뉴스 > 내집마련 돕는 '디딤돌대출' 최대 2조 늘어난다
네이버뉴스 연예 스포츠 경제 랭킹
08.04. (금) 증시
뉴스스탠드 > 전체 언론사 MY 뉴스
코스피
2395.00 ▲8.15
증권 오늘의 주요 시황뉴
파이낸셜뉴스
중앙일보
MBC
한겨레
데일리안
SBS
NAVER – Mozilla Firefox
1 / 4

서버 재부팅 후 네트워크 설정 확인



5) Java 설치(jdk)

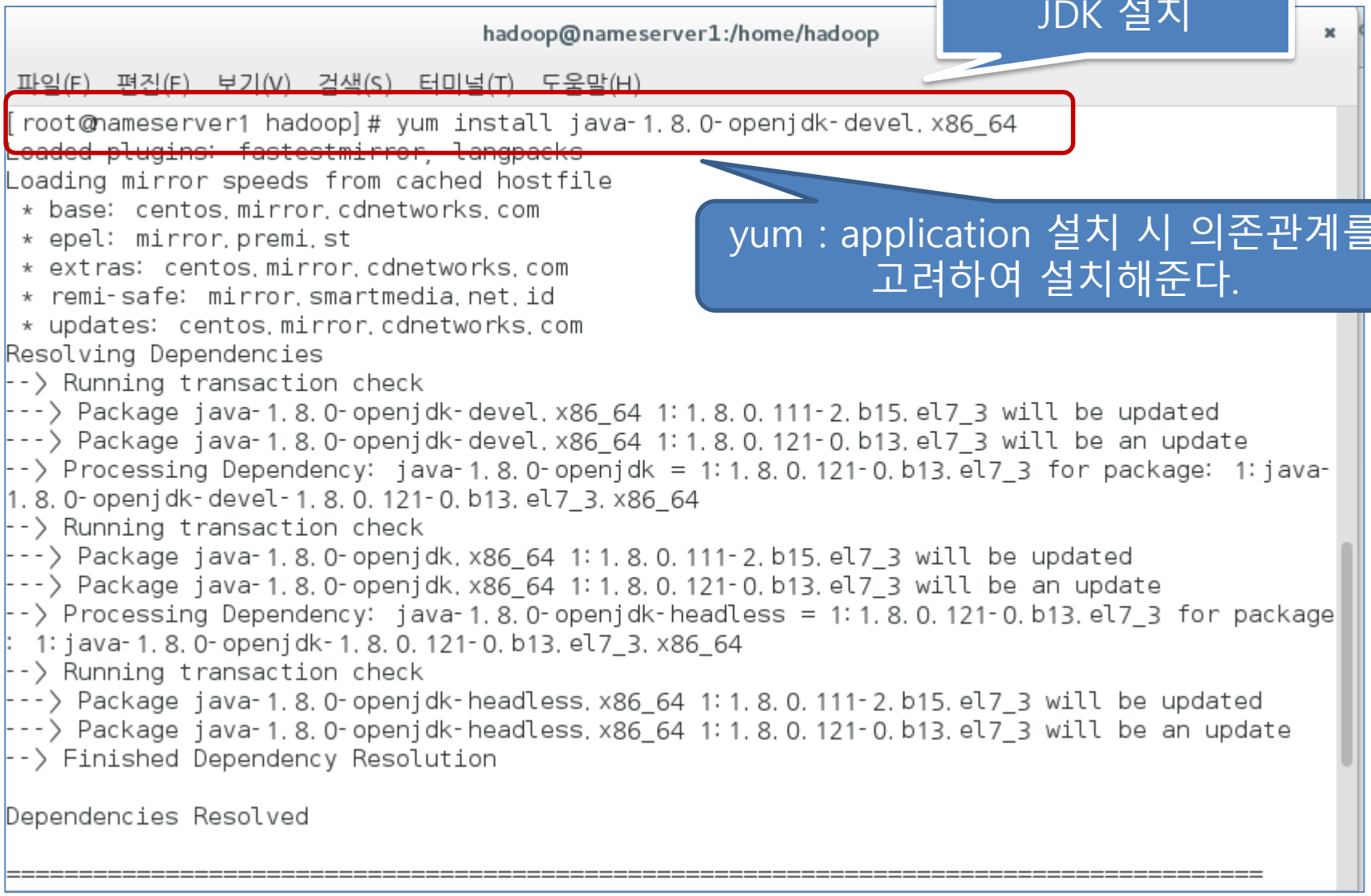
1) Java 설치 확인

✓ yum으로 JDK를 설치하기 위해서는 먼저 JRE 설치를 확인 한다.

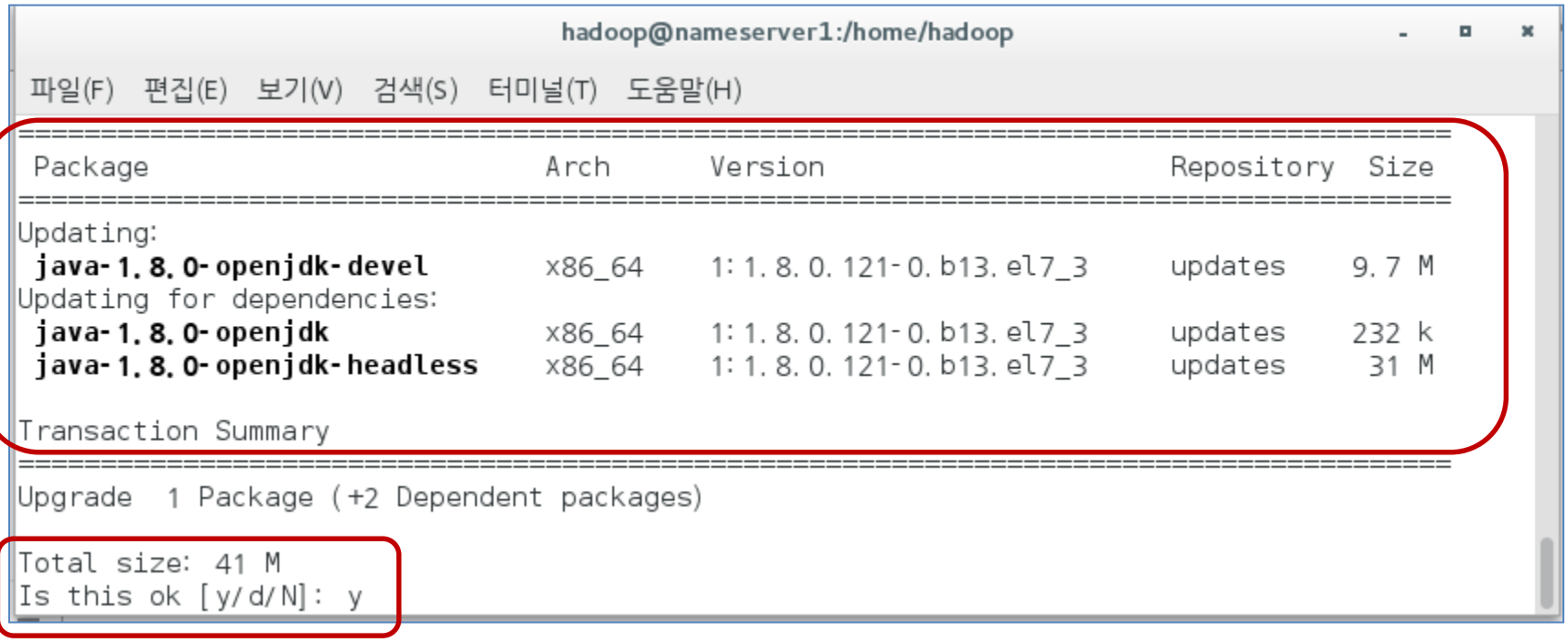
```
hadoop@nameserver1:/home/hadoop
파일(F)  편집(E)  보기(V)  검색(S)  터미널(T)  도움말(H)
[root@nameserver1 hadoop]# yum list java*jdk-devel
Loaded plugins: fastestmirror, langpacks
Loading mirror speeds from cached hostfile
 * base: centos.mirror.cdnetworks.com
 * epel: mirror.premi.st
 * extras: centos.mirror.cdnetworks.com
 * remi-safe: mirror.smartmedia.net.id
 * updates: centos.mirror.cdnetworks.com
Installed Packages
java-1.8.0-openjdk-devel.x86_64          1:1.8.0.111-2.b15.el7_3          @updates
Available Packages
java-1.6.0-openjdk-devel.x86_64          1:1.6.0.41-1.13.13.1.el7_3       updates
java-1.7.0-openjdk-devel.x86_64          1:1.7.0.121-2.6.8.0.el7_3        updates
java-1.8.0-openjdk-devel.i686            1:1.8.0.121-0.b13.el7_3          updates
java-1.8.0-openjdk-devel.x86_64          1:1.8.0.121-0.b13.el7_3          updates
[root@nameserver1 hadoop]#
```

현재 1.6, 1.7, 1.8 버전 설치가 가능하다. 여기서는 1.8 버전을 설치한다.

2) JDK 설치


JDK 설치
hadoop@nameserver1:/home/hadoop
파일(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)
[root@nameserver1 hadoop]# yum install java-1.8.0-openjdk-devel.x86_64
Loaded plugins: fastestmirror, langpacks
Loading mirror speeds from cached hostfile
* base: centos.mirror.cdnetworks.com
* epel: mirror.premi.st
* extras: centos.mirror.cdnetworks.com
* remi-safe: mirror.smartmedia.net.id
* updates: centos.mirror.cdnetworks.com
Resolving Dependencies
--> Running transaction check
---> Package java-1.8.0-openjdk-devel.x86_64 1:1.8.0.111-2.b15.el7_3 will be updated
---> Package java-1.8.0-openjdk-devel.x86_64 1:1.8.0.121-0.b13.el7_3 will be an update
--> Processing Dependency: java-1.8.0-openjdk = 1:1.8.0.121-0.b13.el7_3 for package: 1:java-1.8.0-openjdk-devel-1.8.0.121-0.b13.el7_3.x86_64
--> Running transaction check
---> Package java-1.8.0-openjdk.x86_64 1:1.8.0.111-2.b15.el7_3 will be updated
---> Package java-1.8.0-openjdk.x86_64 1:1.8.0.121-0.b13.el7_3 will be an update
--> Processing Dependency: java-1.8.0-openjdk-headless = 1:1.8.0.121-0.b13.el7_3 for package: 1:java-1.8.0-openjdk-1.8.0.121-0.b13.el7_3.x86_64
--> Running transaction check
---> Package java-1.8.0-openjdk-headless.x86_64 1:1.8.0.111-2.b15.el7_3 will be updated
---> Package java-1.8.0-openjdk-headless.x86_64 1:1.8.0.121-0.b13.el7_3 will be an update
--> Finished Dependency Resolution
Dependencies Resolved
yum : application 설치 시 의존관계를 고려하여 설치해준다.

JDK설치 정보 제공화면



3) JDK 설치 확인

현재 JDK1.8.0_252 최신 버전이 설치되었다.
JRE는 1.8.0_252 최신 버전으로 upgrade 되었다.

➢ **최신 버전은 변경될 수 있음**

6) Slave 폴더 생성(Master 폴더 복사)