

Hadoop 개요

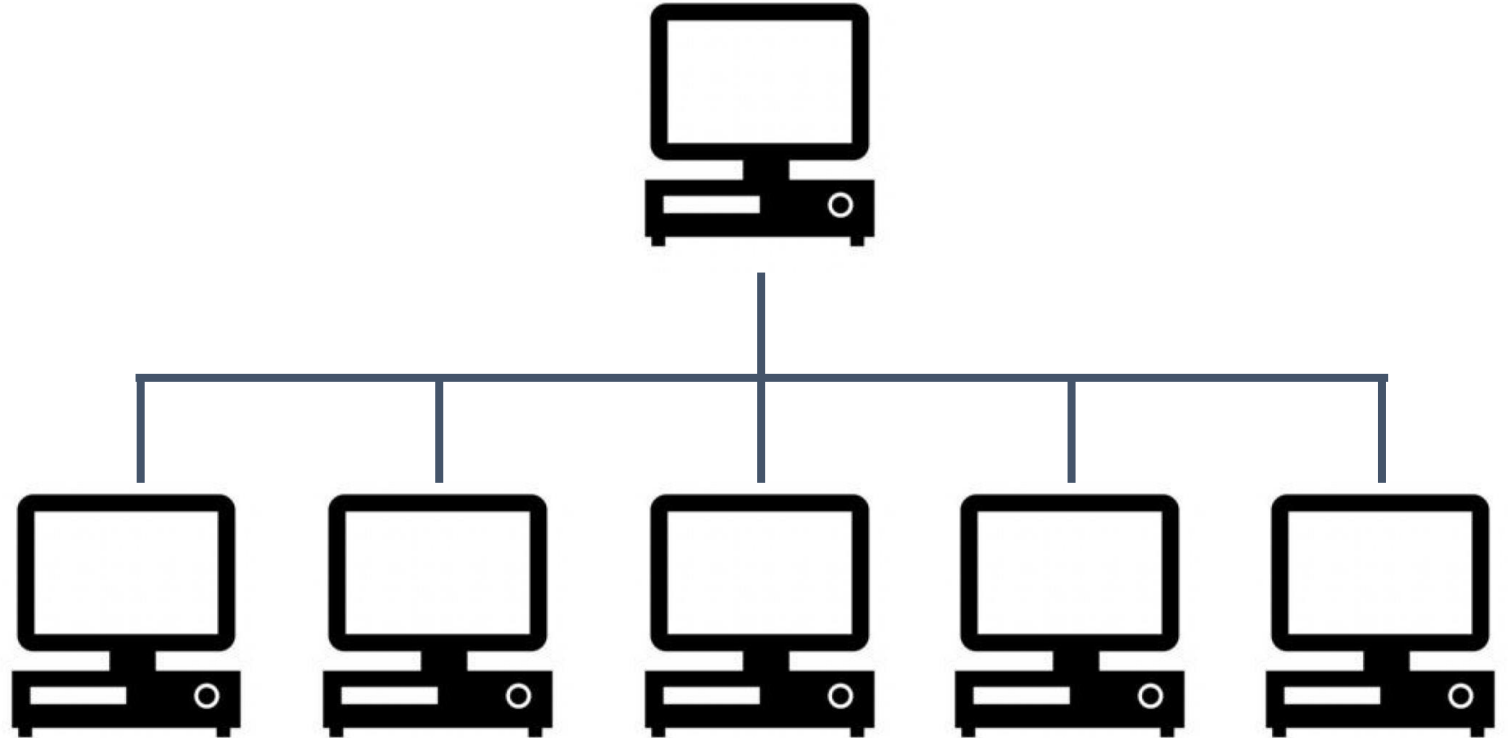
로컬시스템 vs 분산시스템

- 로컬 컴퓨터에서는 RAM 용량(예: 32GB) 범위의 데이터를 다룰 수 있습니다.
- 더 큰 데이터셋은 SQL 데이터베이스를 사용하여 스토리지를 사용하거나, 여러대의 컴퓨터로 구성된 분산시스템을 사용해야 합니다.

Local



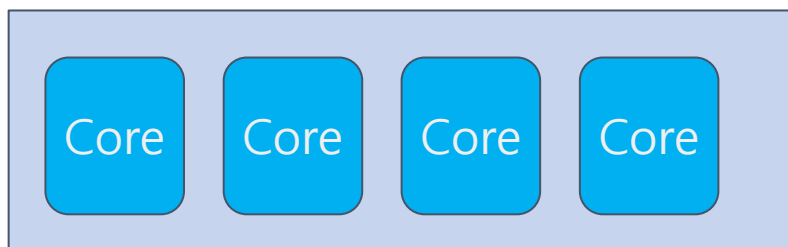
Distributed



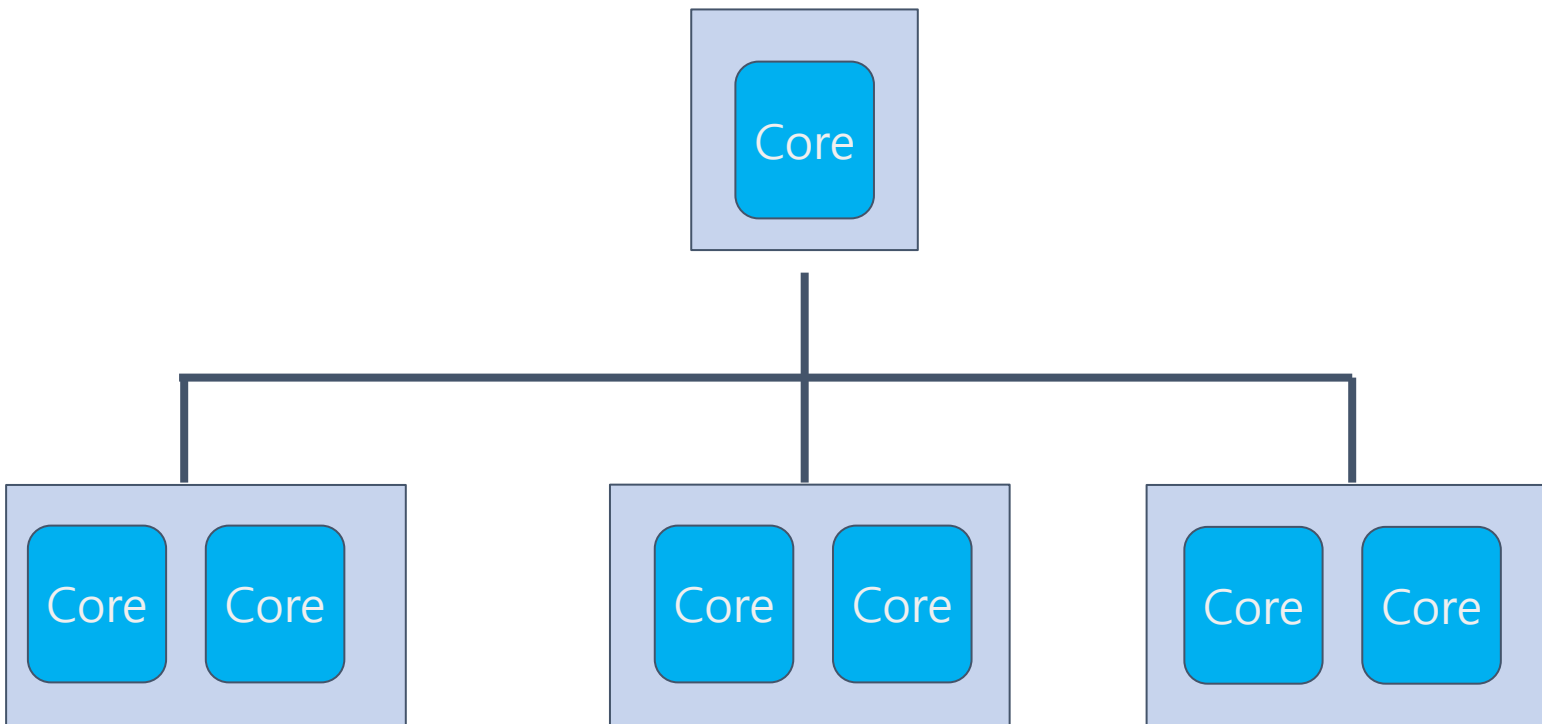
로컬시스템 vs 분산시스템

- 로컬 프로세스는 단일 시스템의 컴퓨팅 리소스를 사용합니다.
- 분산 프로세스는 네트워크를 통해 연결된 여러 머신의 컴퓨팅 리소스를 액세스 할 수 있습니다.
- 단일 머신을 Scale Up 하는 것보다 여러대의 머신으로 확장(Scale Out)하는 것이 더 쉽습니다.
- 분산시스템에서는 한 대의 시스템에 장애가 발생해도 전체 네트워크가 계속 작동 할 수 있습니다.

Local



Distributed



Hadoop



Hadoop Architecture

Application Layer



Other
Applications

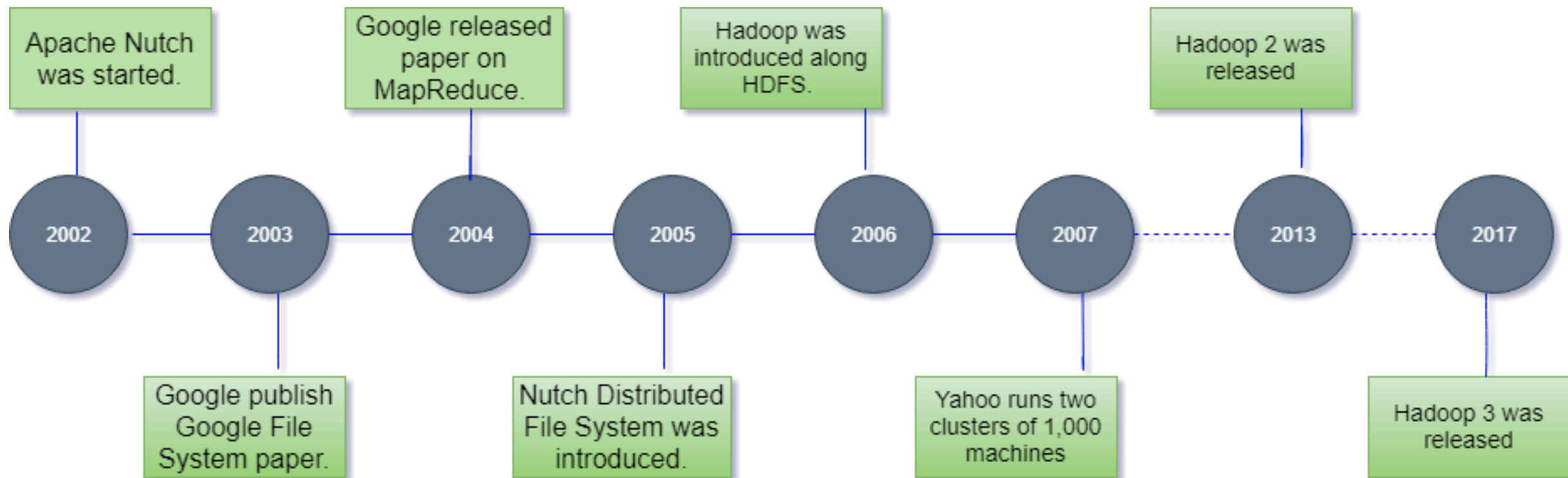
Resource Management
Layer



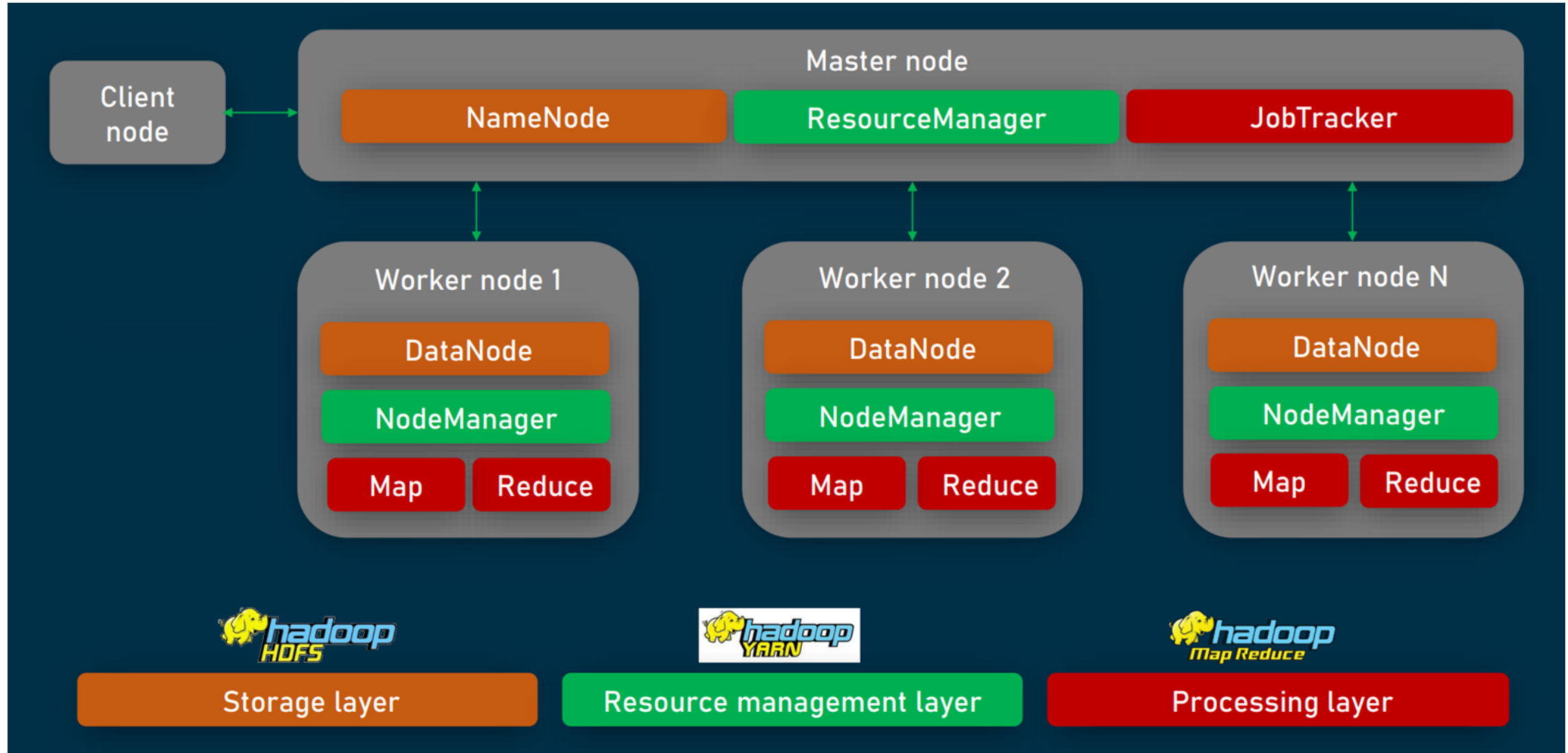
Storage Layer



Hadoop의 역사

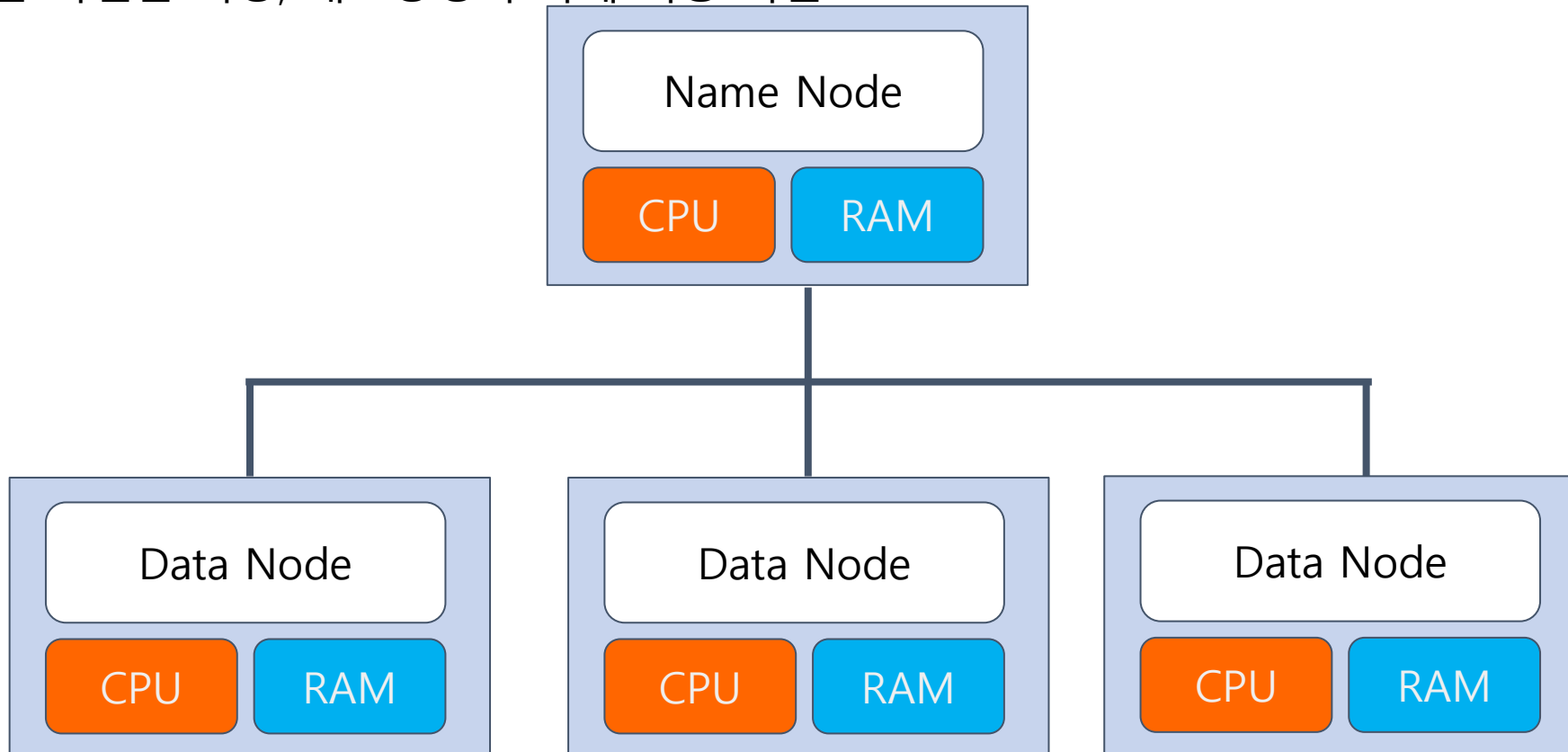


Hadoop 클러스터 아키텍처



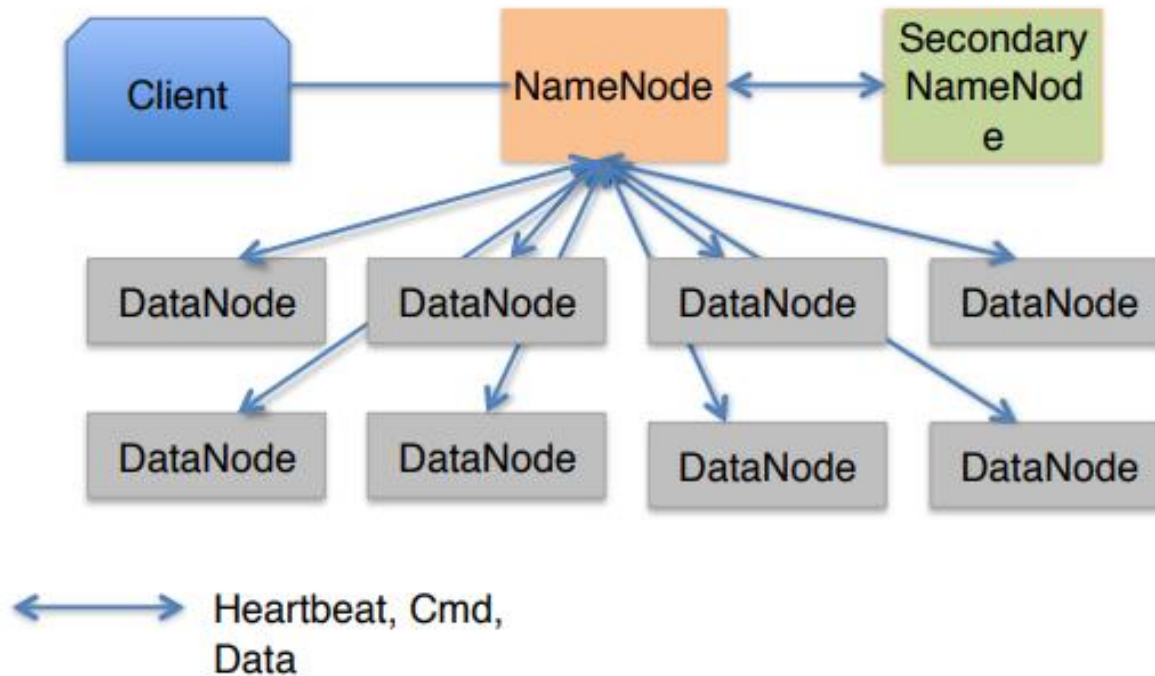
HDFS(Hadoop Distributed File System)

- HDFS는 기본적으로 128MB 크기의 데이터 블록(Block)을 사용하며, 각 Block은 최소 3개로 복제됩니다.
- Block은 내고장성(Fault Tolerance)을 지원하는 방식으로 분산(distributed) 됩니다.
- Block 의 여러 복사본이 노드 장애로 인한 데이터 손실을 방지합니다.
- Block 단위 처리 이점 : 탐색비용 최소화, 메타 데이터 크기 감소, 스토리지 관리 단순화, 물리 디스크 크기를 초과하는 파일을 저장, 내고장성과 복제 기능 지원



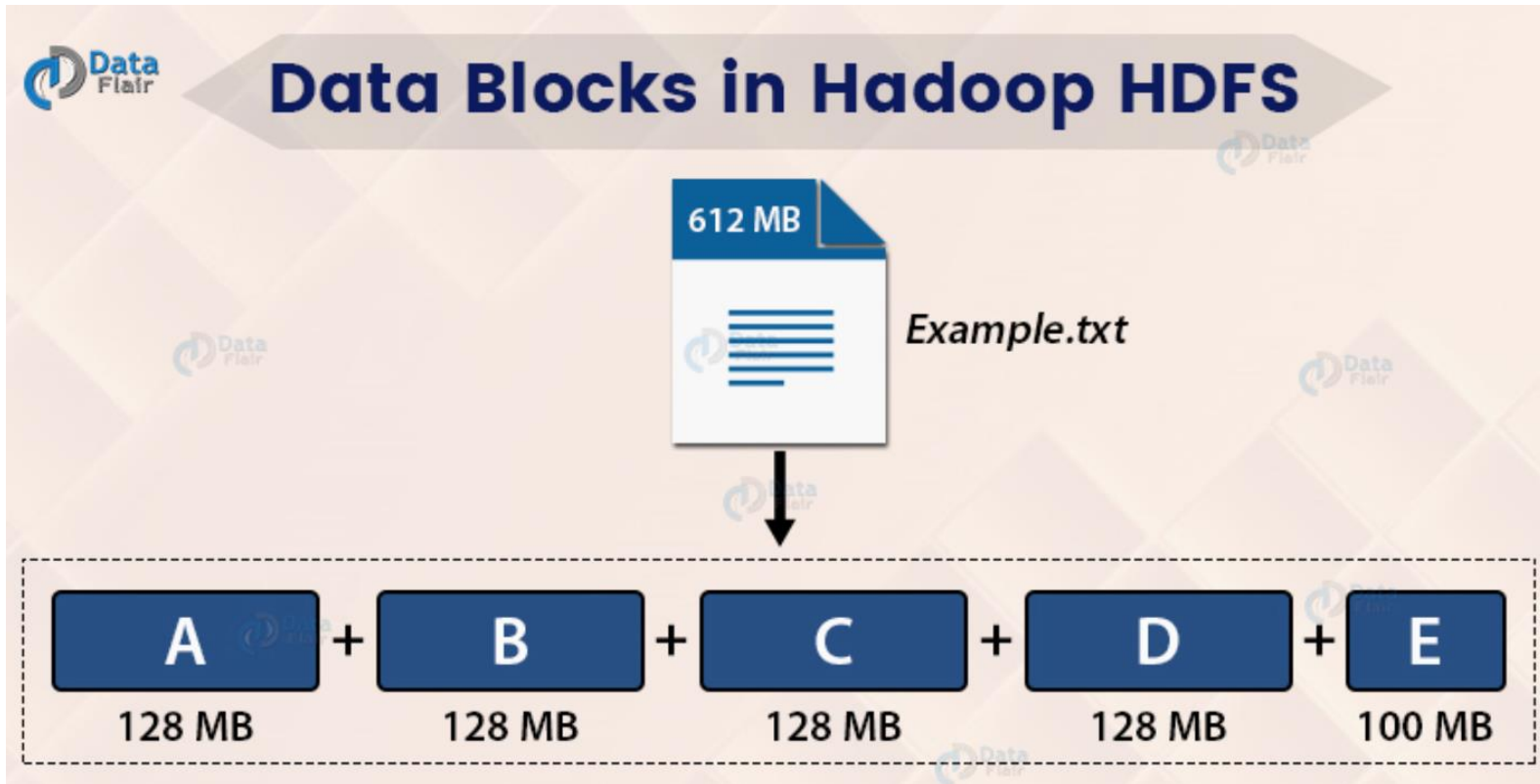
HDFS - 아키텍처

- HDFS는 마스터 슬레이브 구조로 NameNode와 여러 개의 DataNode로 구성됩니다.
- NameNode는 메타데이터를 가지고 있고, 데이터는 블록 단위로 나누어 데이터노드에 저장됩니다.
- 메타데이터는 파일이름, 파일크기, 파일생성시간, 파일접근권한, 파일 소유자 및 그룹 소유자, 파일이 위치한 블록의 정보 등으로 구성됩니다.
- 사용자는 네임노드를 이용해 데이터를 쓰고, 읽을 수 있습니다.



HDFS - 파일 저장 방식

- HDFS는 파일을 분산 저장하기 위해서 먼저 파일의 메타 데이터와 콘텐츠 데이터를 분리합니다.
- 메타 데이터 : 파일의 접근 권한, 생성일, 수정일, 네임 스페이스 등 파일에 대해 설명하는 정보
- 콘텐츠 데이터 : 실제 파일에 저장된 데이터
- 파일의 메타데이터는 네임 노드에 저장되며 콘텐츠 데이터는 블록 단위로 쪼개져서 데이터 노드에 저장 됩니다.



HDFS - NameNode

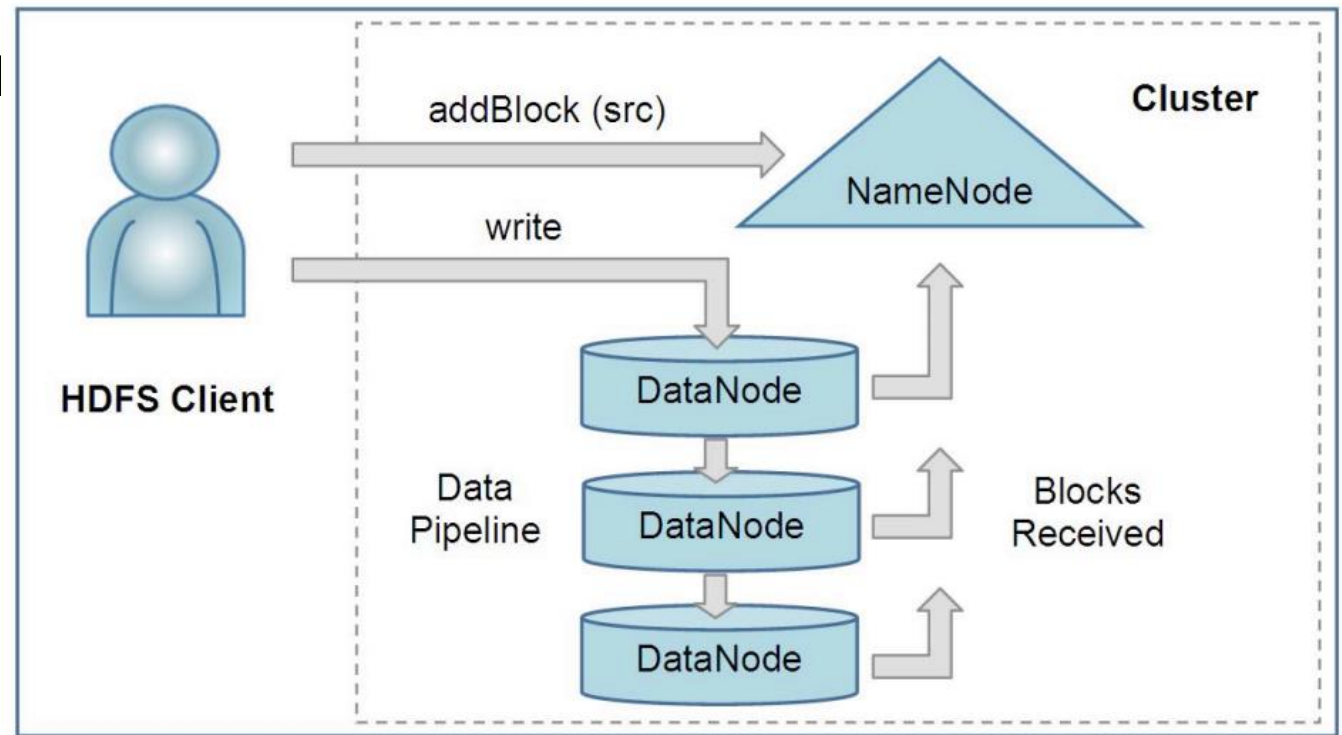
- NameNode는 파일의 메타 데이터를 inode(파일을 기술하는 디스크 상의 데이터 구조)에 저장합니다.
- NameNode에는 파일 구성하는 블록들의 목록과 위치 정보가 저장되어 있습니다.
- NameNode는 HDFS에 파일을 읽거나 쓰는 작업의 시작점 역할을 수행합니다.

■ 파일 읽기 작업

1. 클라이언트는 NameNode에 파일의 네임 스페이스에 해당하는 블록들의 목록과 주소를 요청
2. 클라이언트는 가장 가까이 위치한 DataNode에서 블록을 읽어옴

■ 파일 쓰기 작업

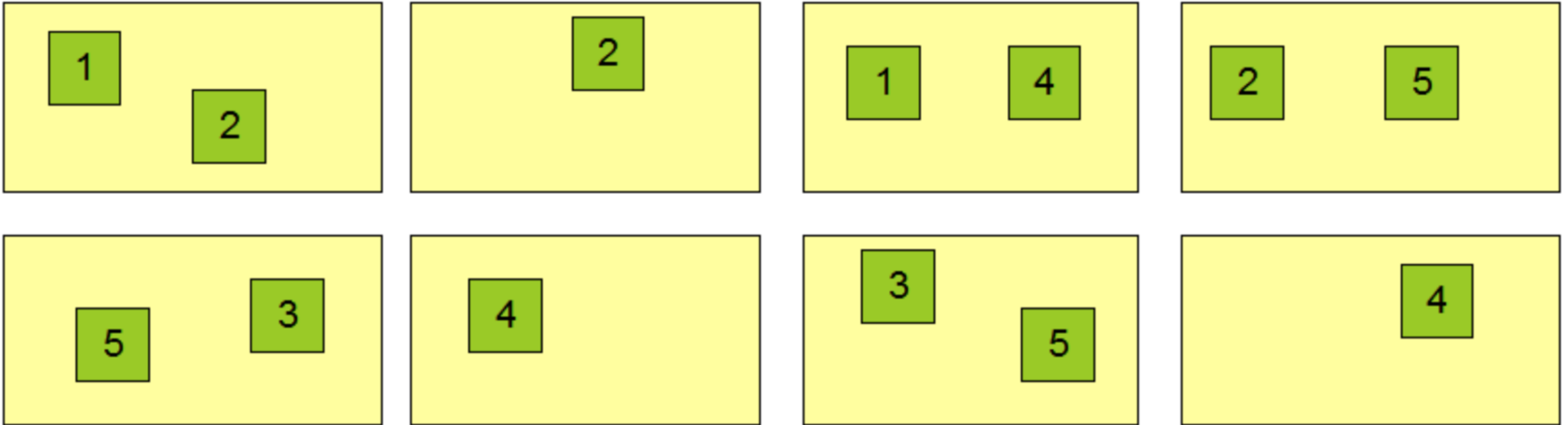
1. 클라이언트는 NameNode에게 어느 Data Node에 블록을 쓰면 좋을지 요청
2. NameNode가 DataNode를 할당
3. 클라이언트는 블록을 쓰기 위한 데이터 파이프라인 생성
4. 데이터 파이프라인을 이용해서 블록 쓰기 작업 수행



HDFS - DataNode

- 파일의 콘텐츠 데이터는 블록 단위로 나뉘며, 블록의 기본 크기는 128MB이며 최소 3개의 복사본을 생성하여 분산 저장합니다.
- DataNode는 그 중 하나의 복사본을 저장하는 것입니다.

Datanodes



MapReduce

- 맵리듀스는 구글에서 대용량 데이터 처리를 분산 병렬 컴퓨팅에서 처리하기 위한 목적으로 제작하여 2004년 발표한 소프트웨어 프레임워크입니다.
- 이 프레임워크는 페타바이트 이상의 대용량 데이터를 신뢰도가 낮은 컴퓨터로 구성된 클러스터 환경에서 병렬 처리를 지원하기 위해서 개발되었습니다.
- 이 프레임워크는 함수형 프로그래밍에서 일반적으로 사용되는 Map과 Reduce라는 함수 기반으로 주로 구성됩니다.

MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat

jeff@google.com, sanjay@google.com

Google, Inc.

Abstract

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a *map* function that processes a key/value pair to generate a set of intermediate key/value pairs, and a *reduce* function that merges all intermediate values associated with the same intermediate key. Many real world tasks are expressible in this model, as shown in the paper.

given day, etc. Most such computations are conceptually straightforward. However, the input data is usually large and the computations have to be distributed across hundreds or thousands of machines in order to finish in a reasonable amount of time. The issues of how to parallelize the computation, distribute the data, and handle failures conspire to obscure the original simple computation with large amounts of complex code to deal with these issues.

As a reaction to this complexity, we designed a new

MapReduce

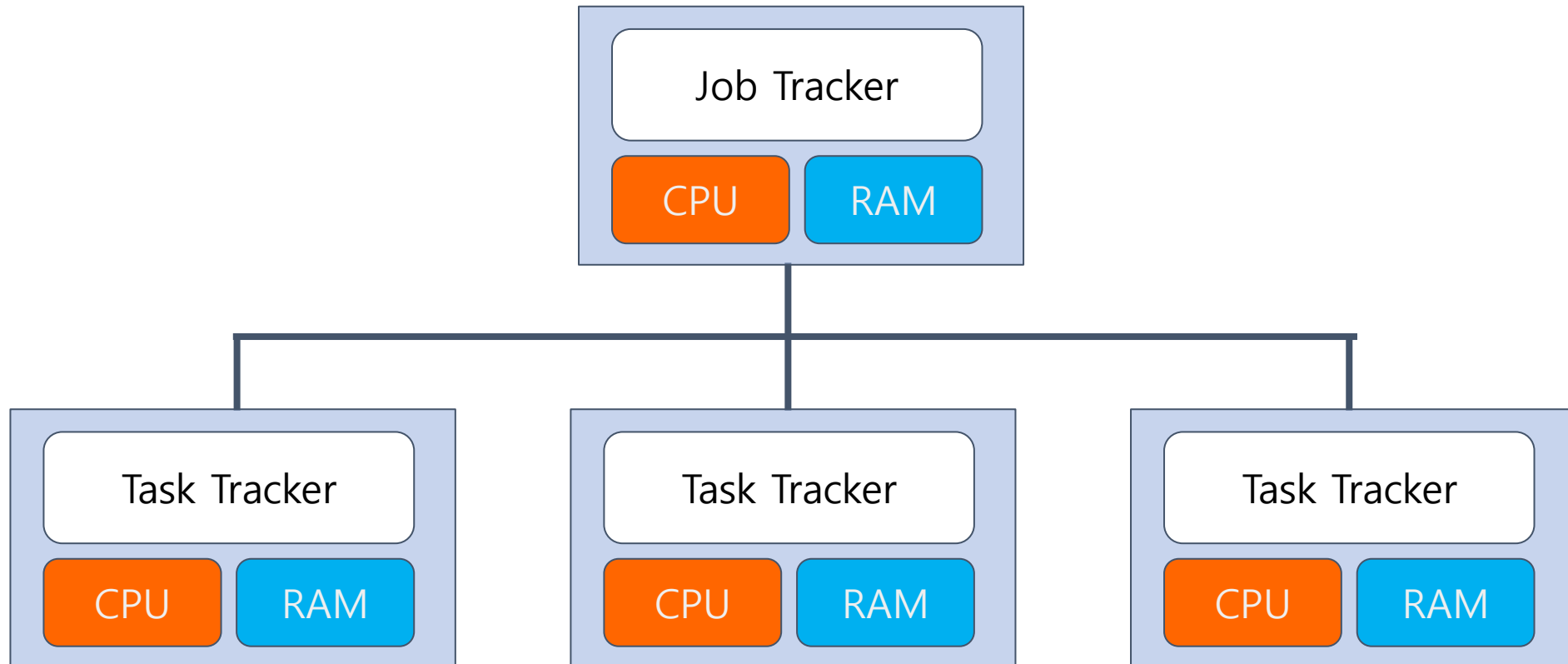
- 맵 리듀스는 구글 내부에서 크롤링 된 문서, 로그 등 방대한 양의 raw data를 분석하는 과정에서 느낀 불편함에서 출발했습니다.
- 프로그램 로직 자체는 단순한데 입력 데이터의 크기가 워낙 커서 연산을 하나의 물리 머신에서 수행할 수가 없었습니다.
- 거대한 인풋 데이터를 쪼개어 수많은 머신들에게 분산시켜서 로직을 수행한 다음 결과를 하나로 합치자는 것이 핵심 아이디어 입니다.
- MapReduce 프레임워크에서 개발자가 코드를 작성하는 부분은 map과 reduce 두 가지 함수입니다.
- map은 전체 데이터를 쪼갠 청크에 대해서 실제로 수행할 로직입니다.
- reduce는 분산되어 처리된 결과 값들을 다시 하나로 합쳐주는 과정이며, 이 역시 분산된 머신들에서 병렬적으로 수행됩니다.

■ MapReduce 수행 절차

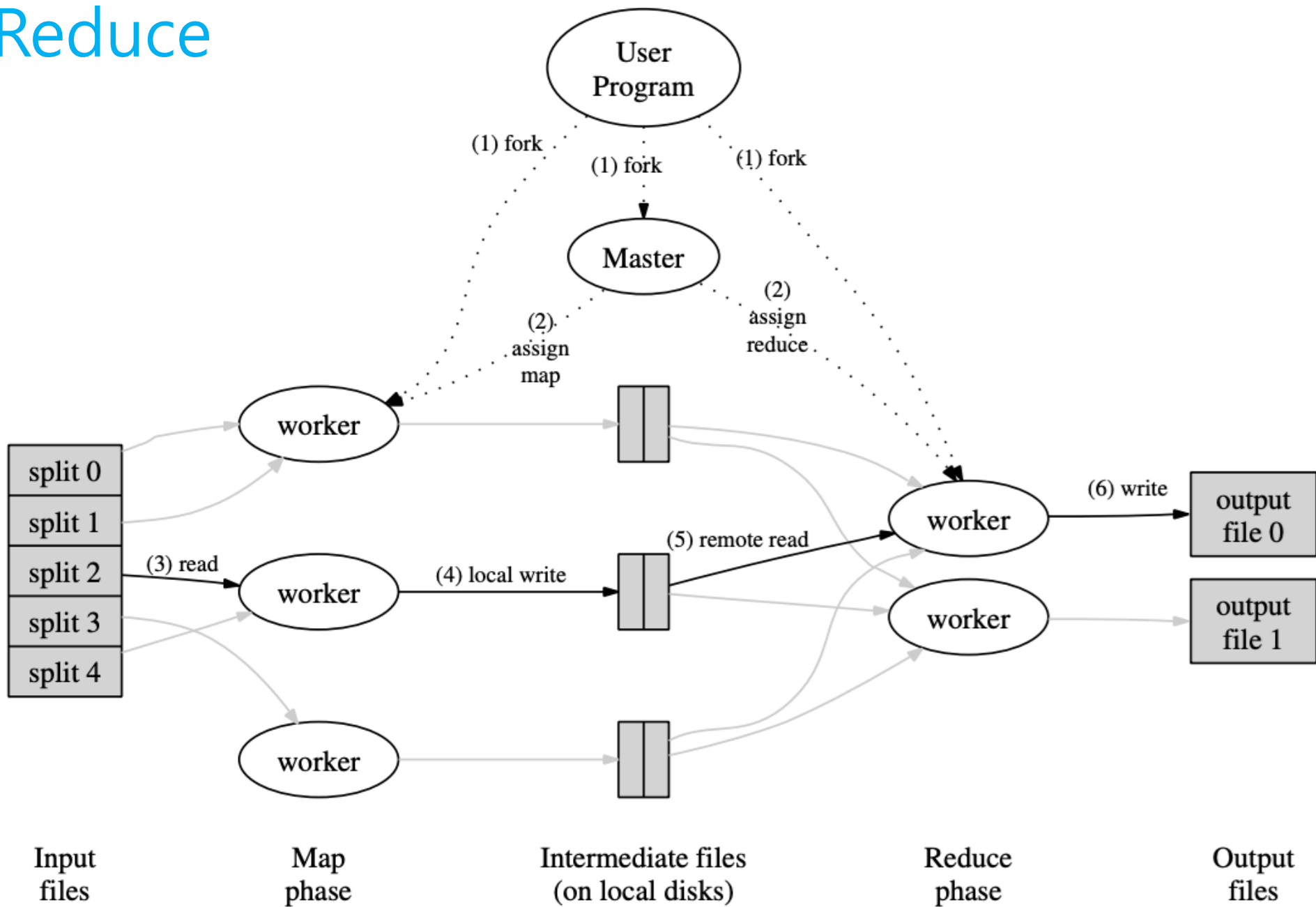
1. 쪼개기(Split): 크기가 큰 인풋 파일을 작은 단위의 청크들로 나누어 분산 파일 시스템(ex. HDFS)에 저장합니다.
2. 데이터 처리하기(Map): 잘게 쪼개어진 파일을 인풋으로 받아서 데이터를 분석하는 로직을 수행합니다.
3. 처리된 데이터 합치기(Reduce): 처리된 데이터를 다시 합칩니다.

MapReduce

- MapReduce는 Computation Task를 분산 된 파일셋(예 : HDFS)으로 분할하는 방법입니다..
- Job Tracker와 여러 Task Tracker로 구성됩니다.
- Job Tracker는 맵리듀스 Job이 수행되는전체 과정을 조정하며, Job 에 대한 마스터(Master) 역할 수행합니다.
- Task Tracker는 Job 에 대한 분할된Task 를 수행하며, 실질적인 Data Processing 의 주체입니다.
- Task Tracker는 Task에 CPU와 메모리를 할당하고 Worker 노드의 Task 들을 모니터링 합니다.

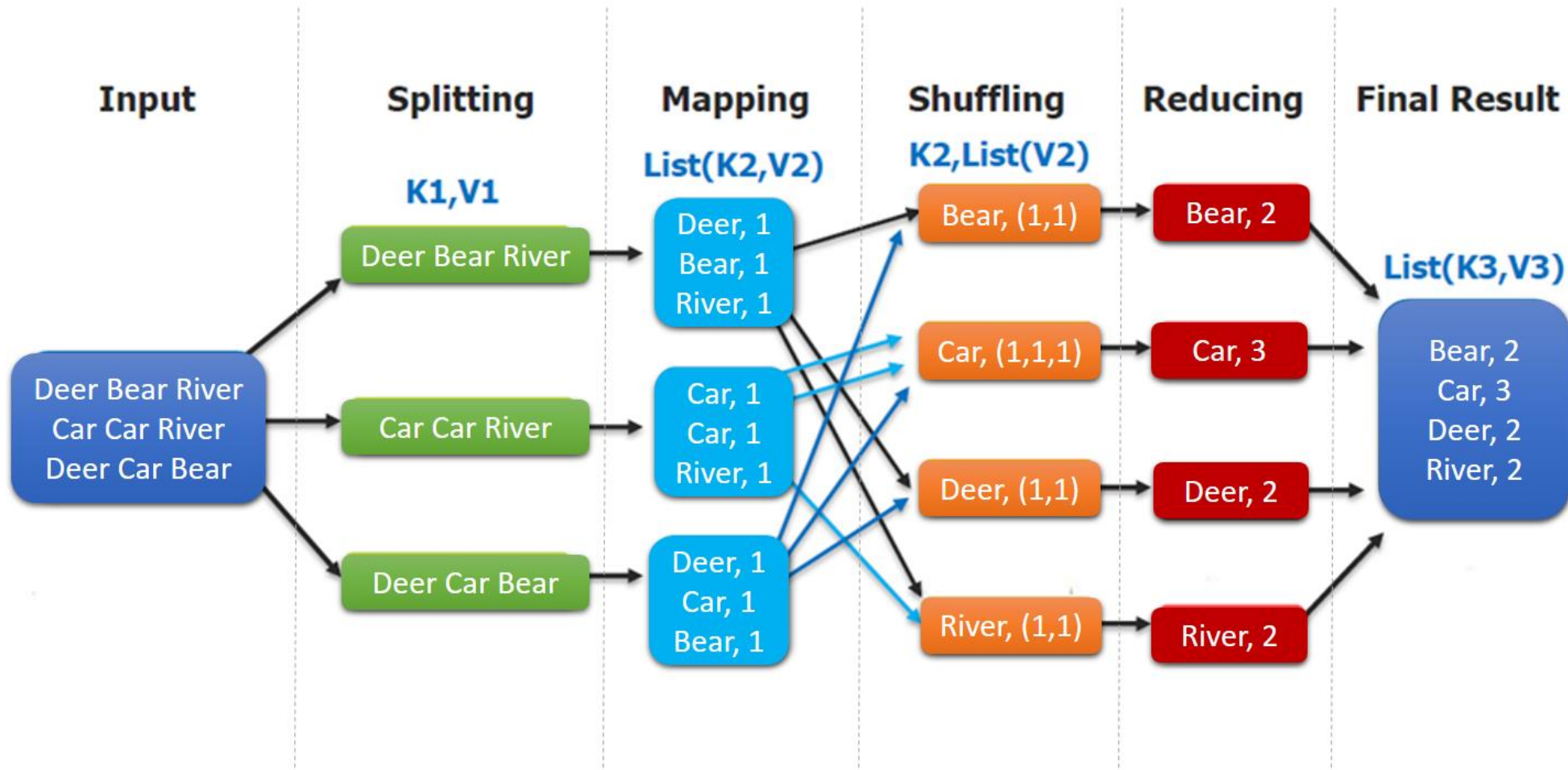


MapReduce



MapReduce

3대의 Mapper와 4대의 Reducer 노드로 이루어진 클러스터에서 워드 카운팅을 수행하는 예시



MapReduce Word Count Process

리눅스(우분투) 가상머신 설치

VirtualBox 설치

<https://www.virtualbox.org/>



■ 설치방법 참고 : <https://zoosso.tistory.com/1214>

VirtualBox 7.0.10 platform packages

- [Windows hosts](#)
- [macOS / Intel hosts](#)
- [Linux distributions](#)
- [Solaris hosts](#)
- [Solaris 11 IPS hosts](#)

VirtualBox 7.0.10 Oracle VM VirtualBox Extension Pack

- [All supported platforms](#)

Support VirtualBox RDP, disk encryption, NVMe and PXE boot for Intel cards. See [this chapter from the User Manual](#) for an introduction to this Extension Pack. The Extension Pack binaries are released under the [VirtualBox Personal Use and Evaluation License \(PUEL\)](#). Please install the same version extension pack as your installed version of VirtualBox.

Ubuntu 설치 – Ubuntu 22.04 LTS 다운로드

<https://ubuntu.com/#download>

The screenshot shows the Ubuntu website's download section. The browser's address bar displays 'ubuntu.com/#download'. The navigation bar includes the Canonical logo, the Ubuntu logo, and links for Enterprise, Developer, Community, and Download. The Download dropdown menu is open, showing options for Desktop, Server, IoT, and Cloud. Under the Desktop section, the '22.04 LTS' version is highlighted with a red box, next to the '22.10' version. The Server section features a 'Get Ubuntu Server' button. The IoT and Cloud sections provide links to various hardware and cloud providers. The footer contains links to tutorials, documentation, other download methods, and Ubuntu flavors.

Enterprise Open Source and Lin x +

ubuntu.com/#download

CANONICAL We are hiring Products ▾

ubuntu® Enterprise ▾ Developer ▾ Community ▾ Download ▴ Search 🔍 Sign in

Ubuntu Desktop ▸

Download Ubuntu desktop and replace your current operating system whether it's Windows or Mac OS, or, run Ubuntu alongside it.

22.04 LTS 22.10

Ubuntu Server ▸

The most popular server Linux in the cloud and data centre, you can rely on Ubuntu Server and its five years of guaranteed free upgrades.

Get Ubuntu Server

Mac and Windows
ARM
IBM Power
s390x

Ubuntu for IoT ▸

Are you a developer who wants to try snappy Ubuntu Core or classic Ubuntu on an IoT board?

Raspberry Pi
Intel IoT platforms
Intel NUC
KVM
Qualcomm Dragonboard 410c
Intel IEI TANK 870
AMD-Xilinx Evaluation kits & SOMs
RISC-V platforms

Ubuntu Cloud ▸

Use Ubuntu optimised and certified server images on most major clouds.

Get started on Amazon AWS, Microsoft Azure, Google Cloud Platform and more...

Download cloud images for local development and testing

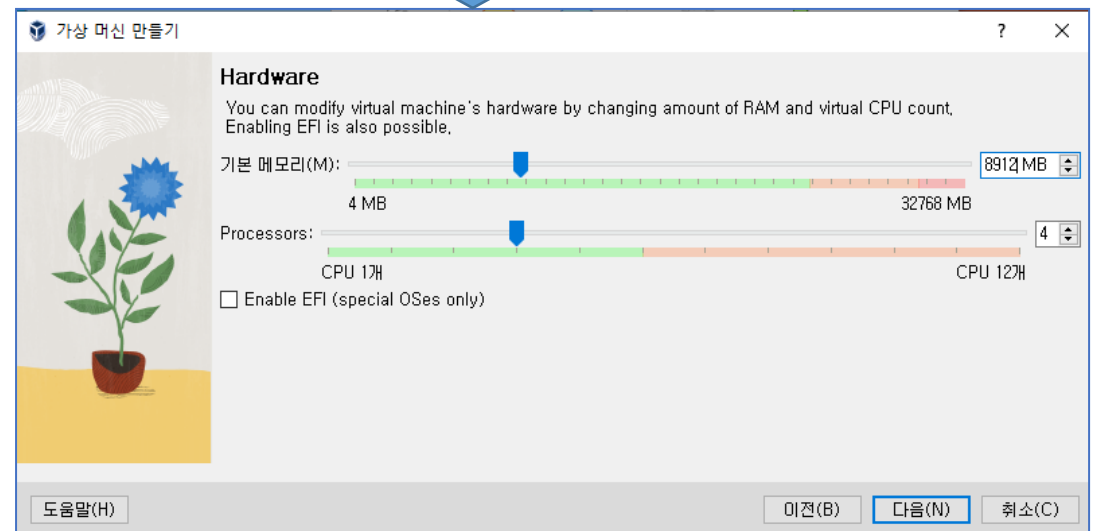
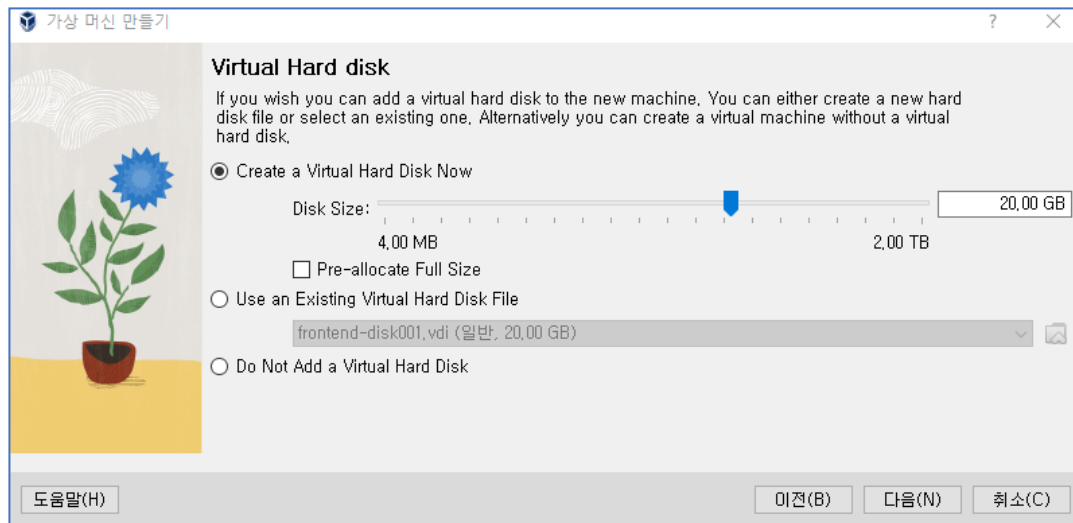
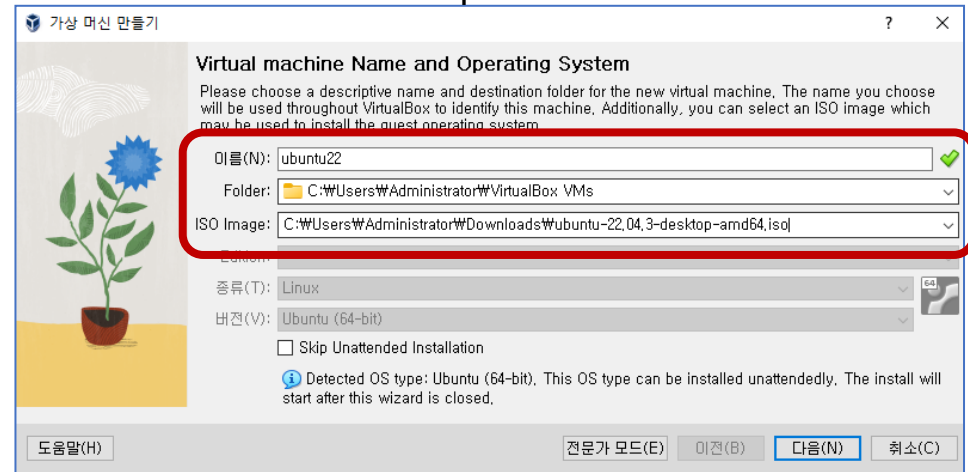
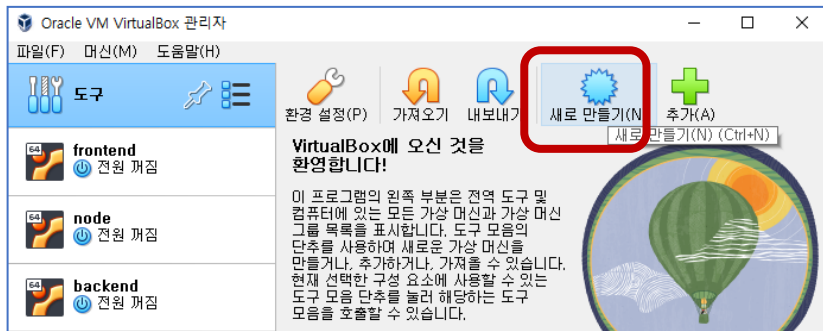
TUTORIALS READ THE DOCS OTHER WAYS TO DOWNLOAD UBUNTU FLAVOURS

If you are already running Ubuntu - you can [upgrade](#) with the Software Read the official docs for [Ubuntu Desktop](#), [Ubuntu Server](#), and [Ubuntu](#) Ubuntu is available via [BitTorrents](#) and via a minimal [network installer](#) Find new ways to experience Ubuntu, each with their own choice

Ubuntu 설치 - 새로 만들기

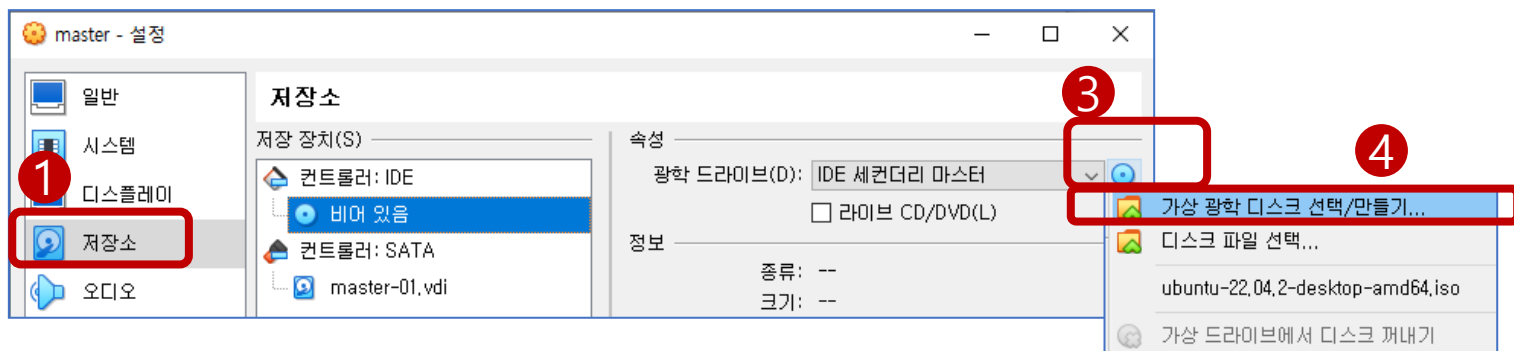
서버 이름 입력
설치 위치 선택
ubuntu-22.04.3-desktop-amd64.iso 선택

VirtualBox 관리자에서 [새로만들기] 클릭

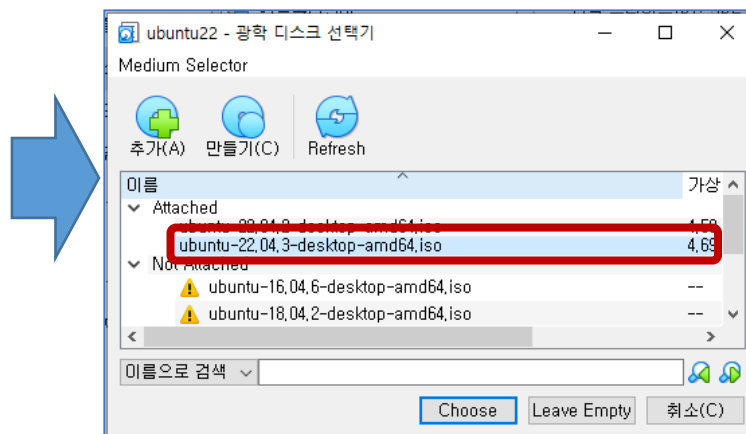


Ubuntu 설치 - 광학 드라이브 설정

생성한 가상 머신에서 [설정] 클릭



다운로드한
[우분투 ISO 파일 선택] 선택



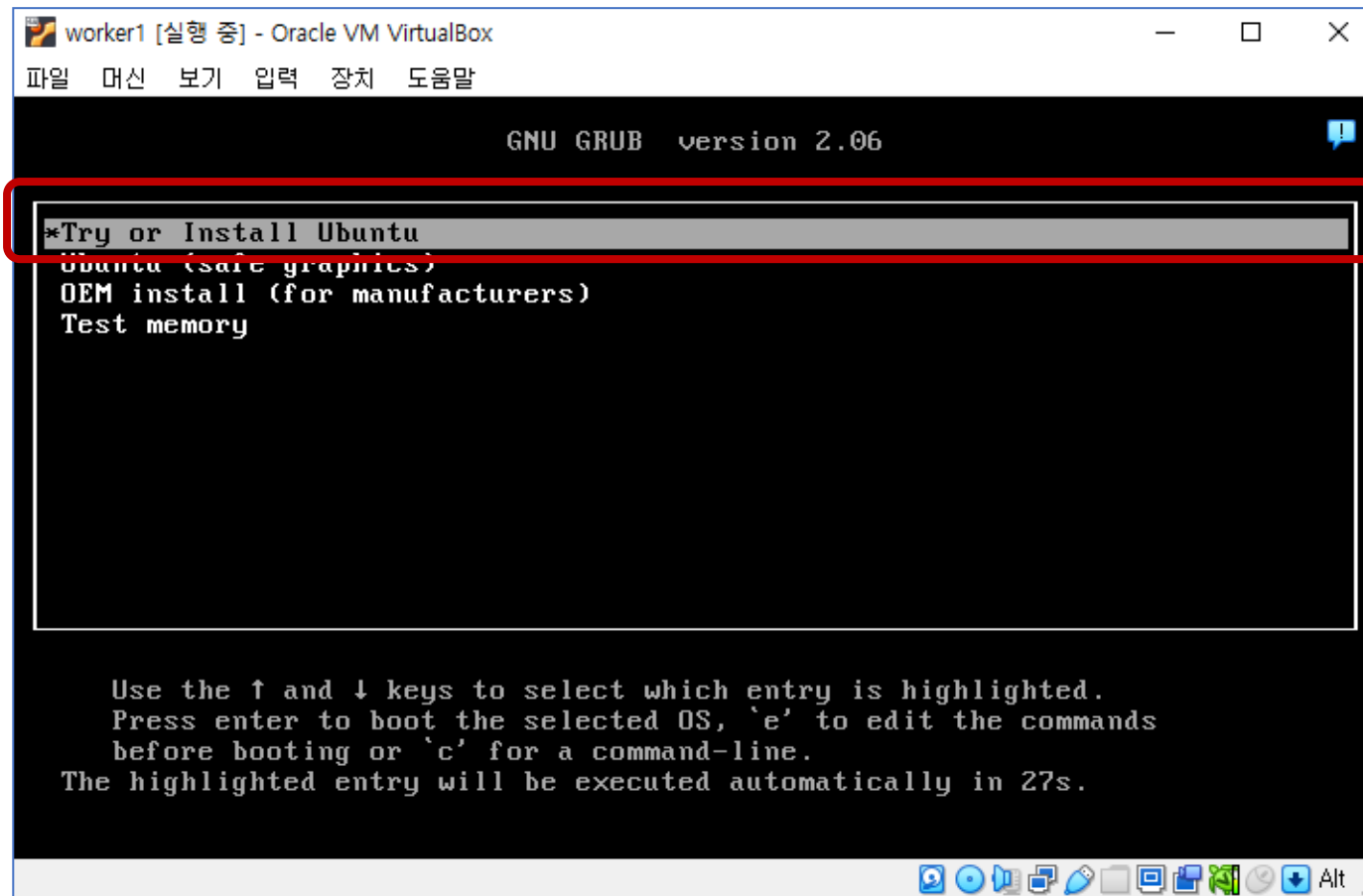
- ① [저장소] 클릭
- ② [비어있음] 클릭
- ③ 아이콘 클릭
- ④ [가상 광학 디스크 선택/만들기...] 클릭

Ubuntu 설치 - Install Ubuntu

[시작(T)] 버튼 클릭

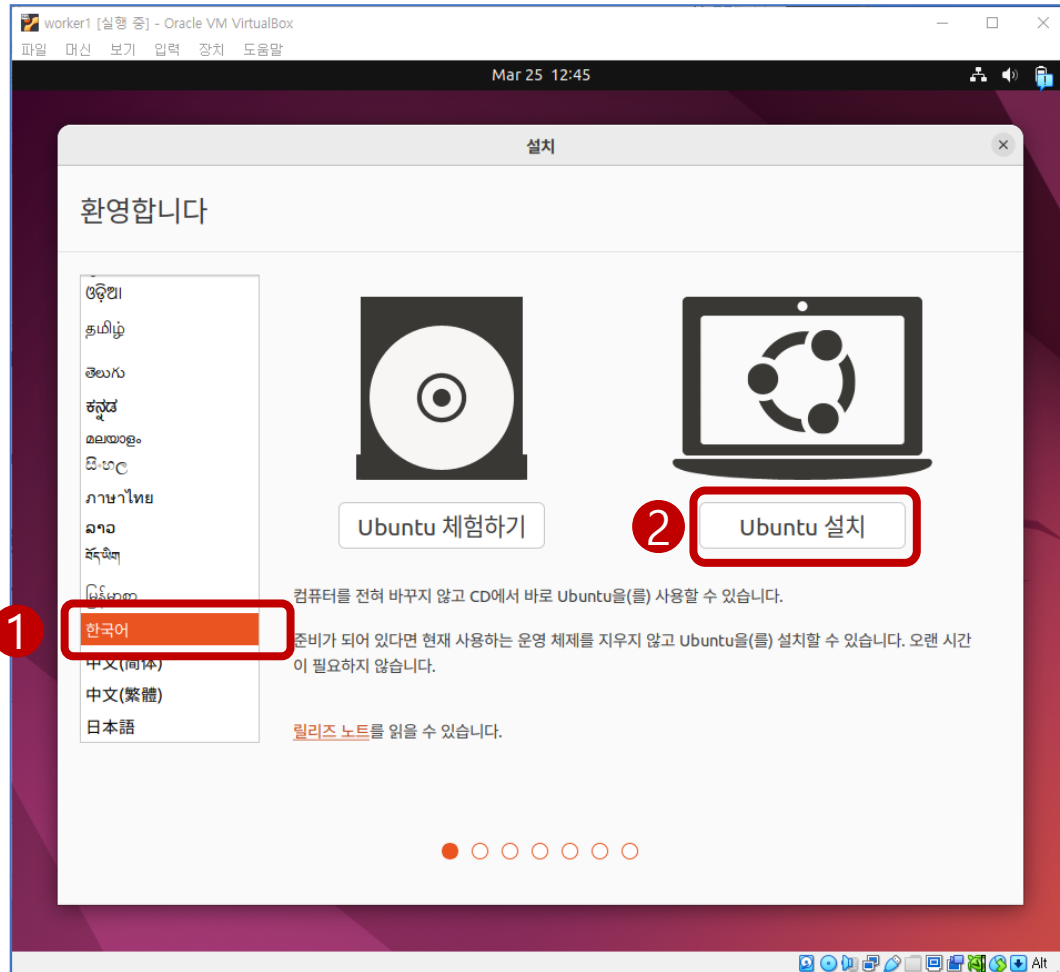


[Try or Install Ubuntu] 선택

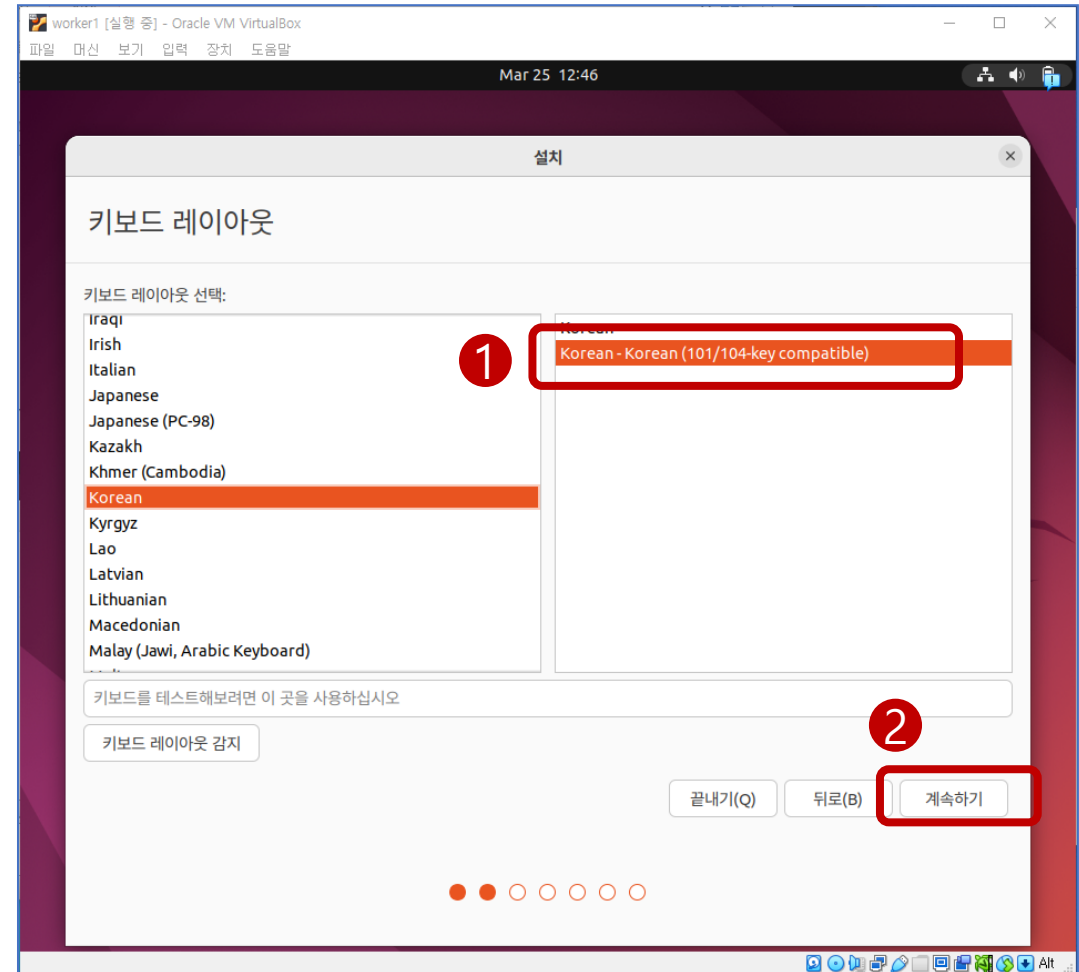


Ubuntu 설치 - 언어 및 키보드 선택

- ① [한국어] 선택
- ② [ubuntu 설치] 선택하고 [Enter Key] 누름

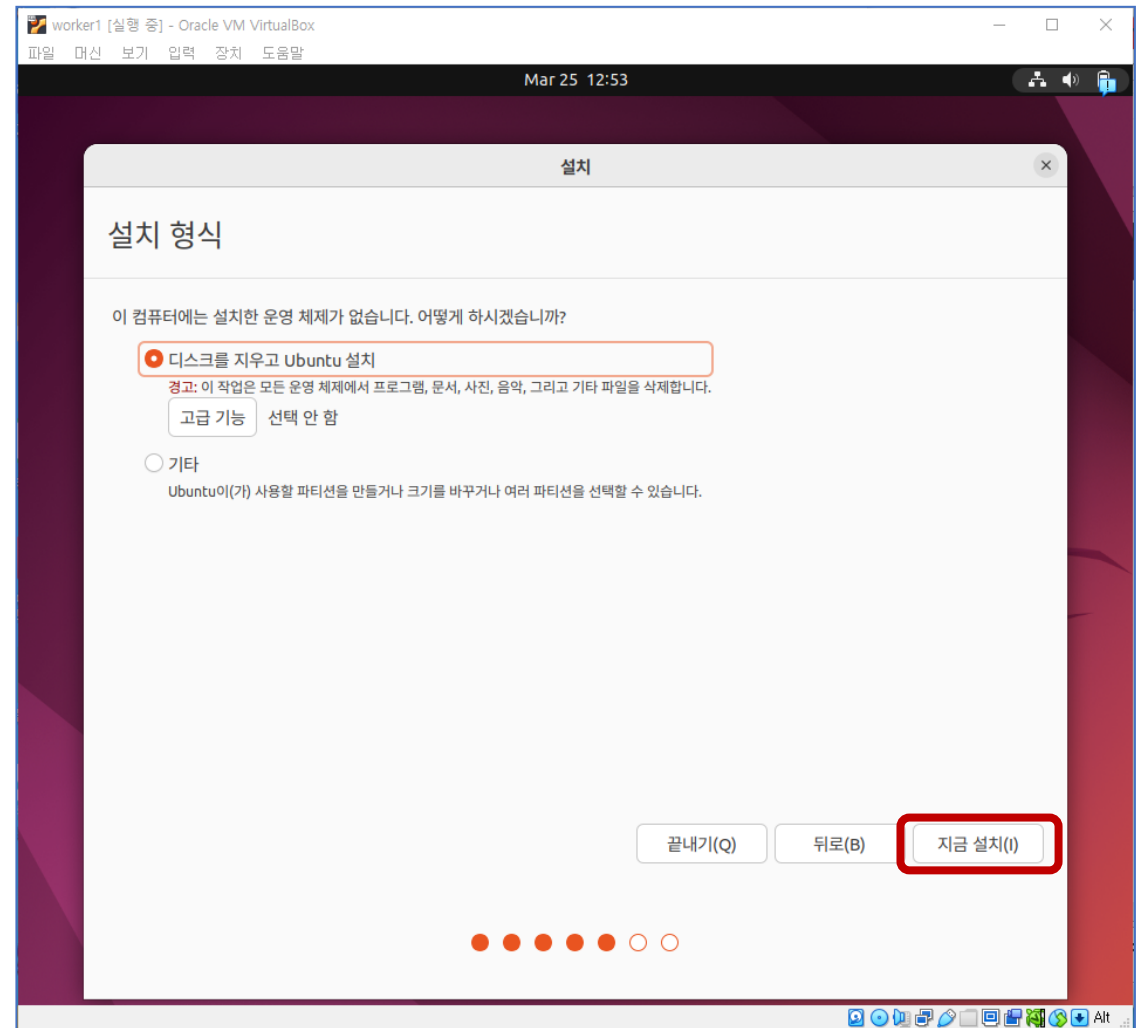
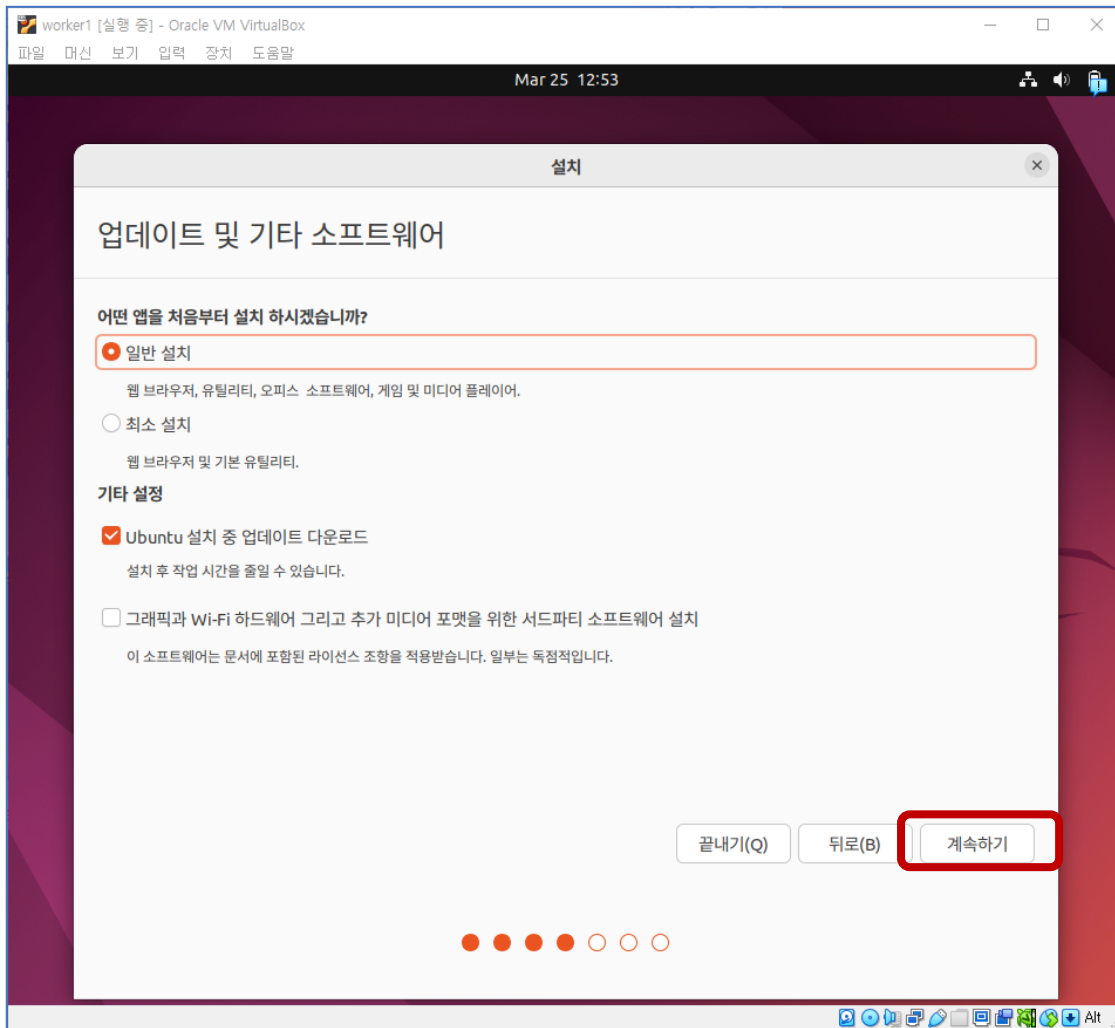


- ① 키보드 레이아웃 선택
- ② [계속하기] 클릭

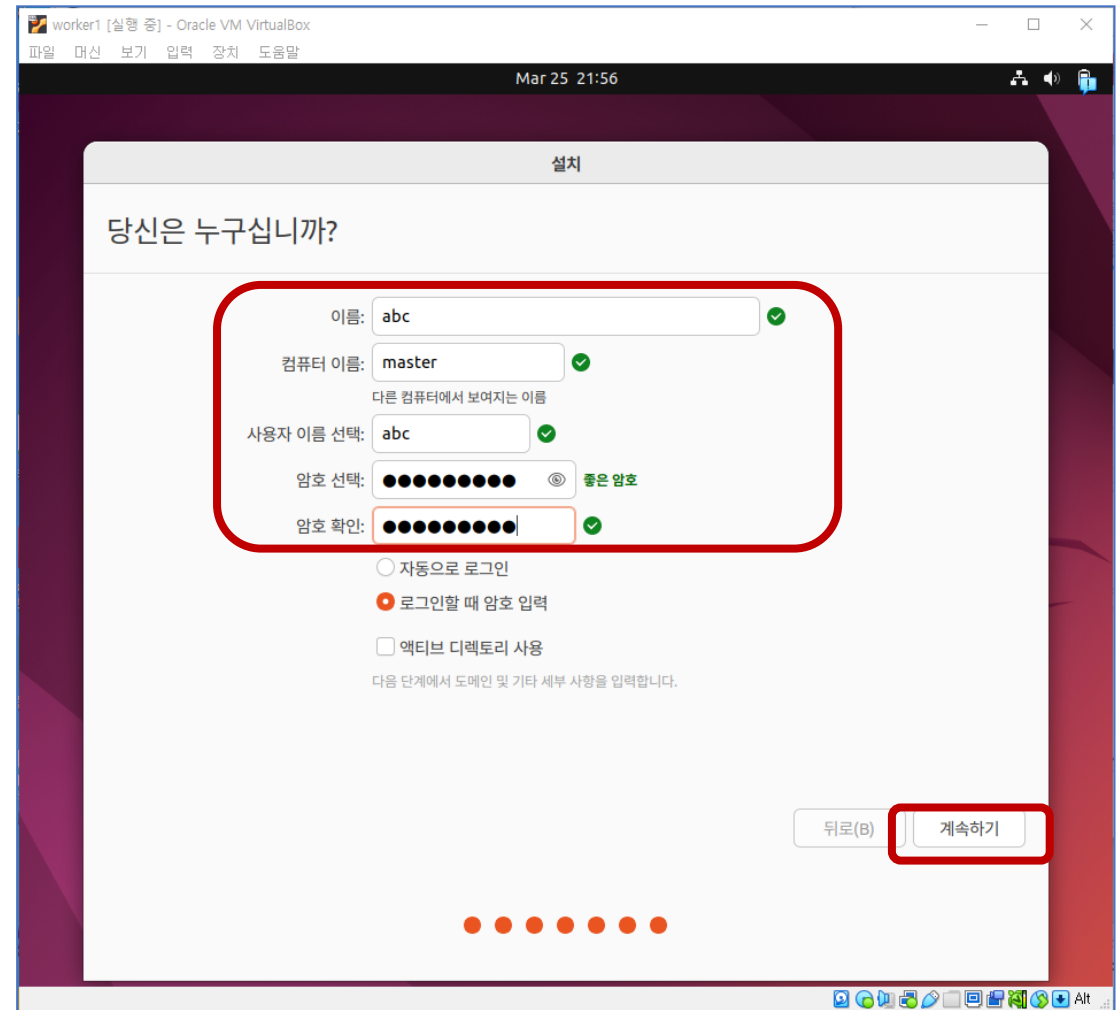
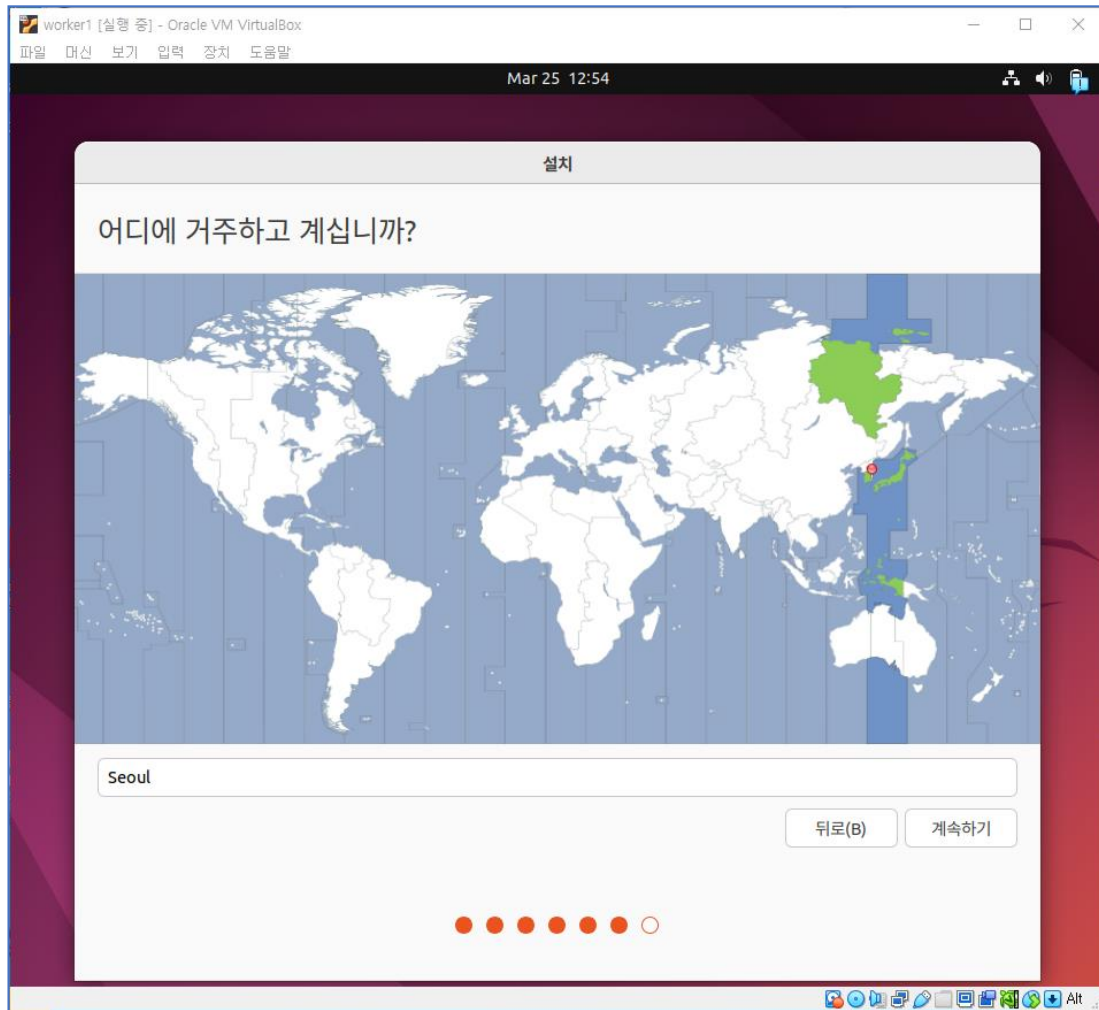


※ Ubuntu 설치시 화면 짚림 현상 발생시 해결 방법
Alt + F7 → 화살표키 or 마우스 이동

Ubuntu 설치 - 업데이트 및 설치 형식 선택

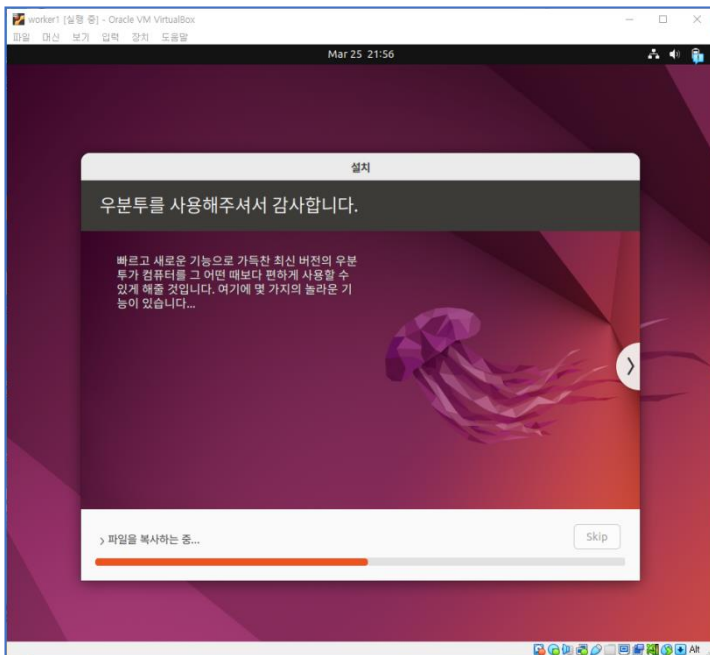


Ubuntu 설치 - 지역 선택 및 계정 설정

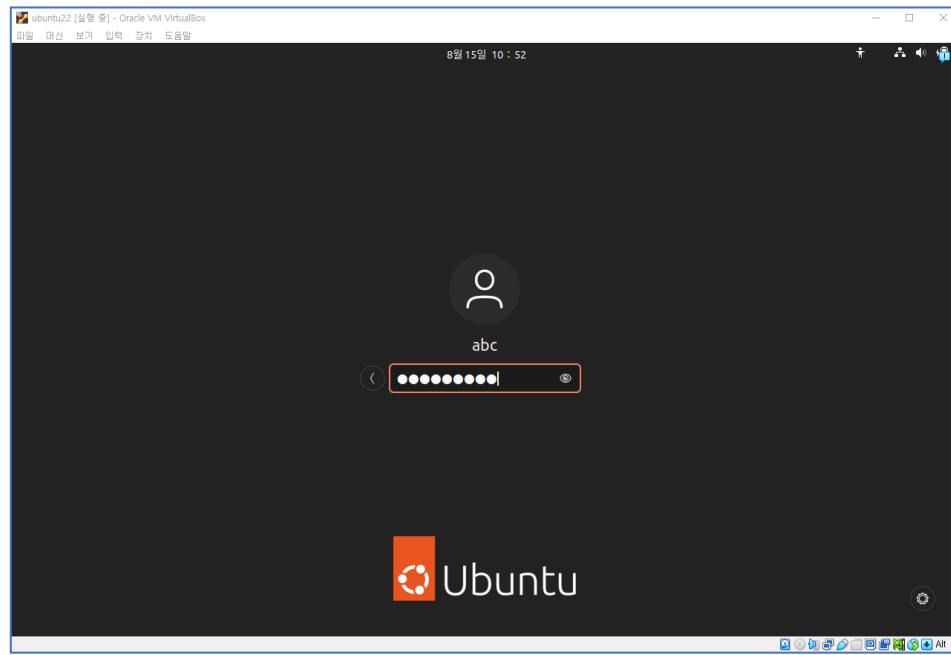


Ubuntu 설치 - 설치 완료 및 재부팅

설치 진행 화면



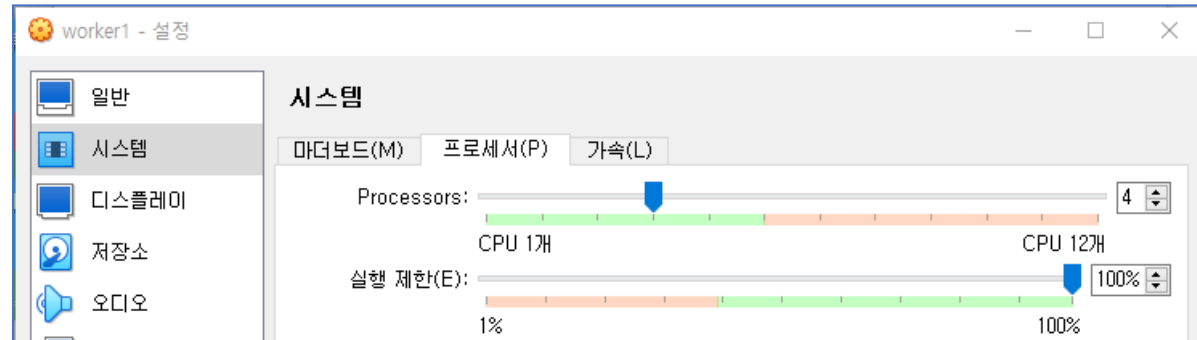
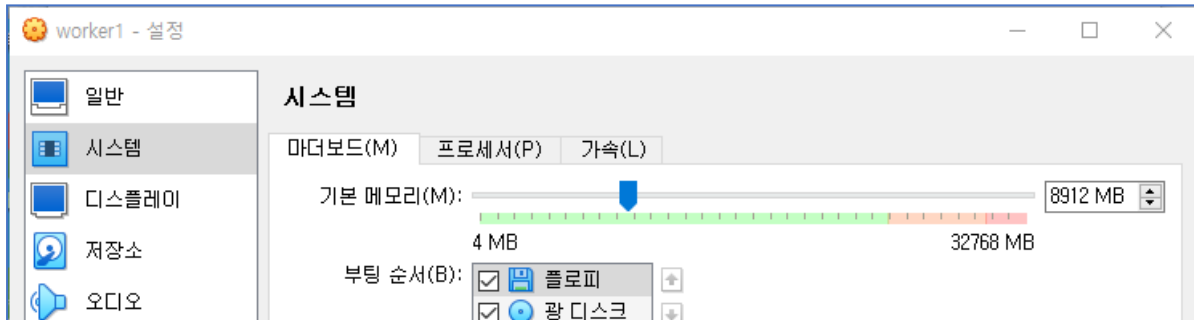
로그인 화면



가상머신 운영 최소 권장 사항

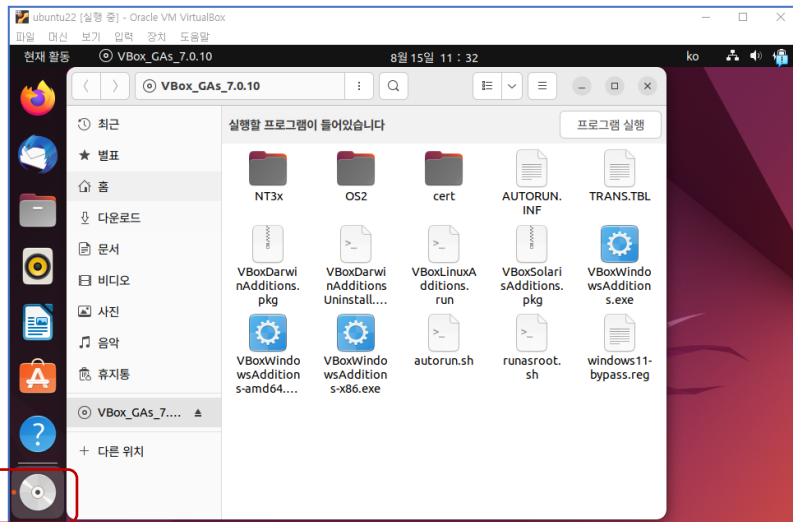
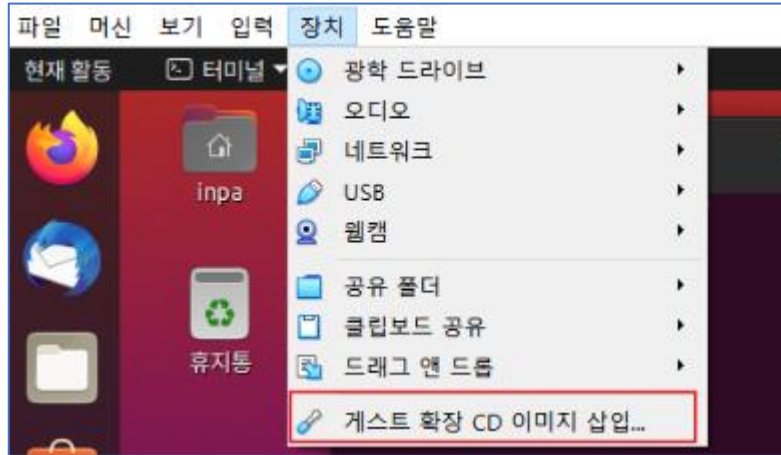
메모리 : 8G 이상

CPU : 4개 이상



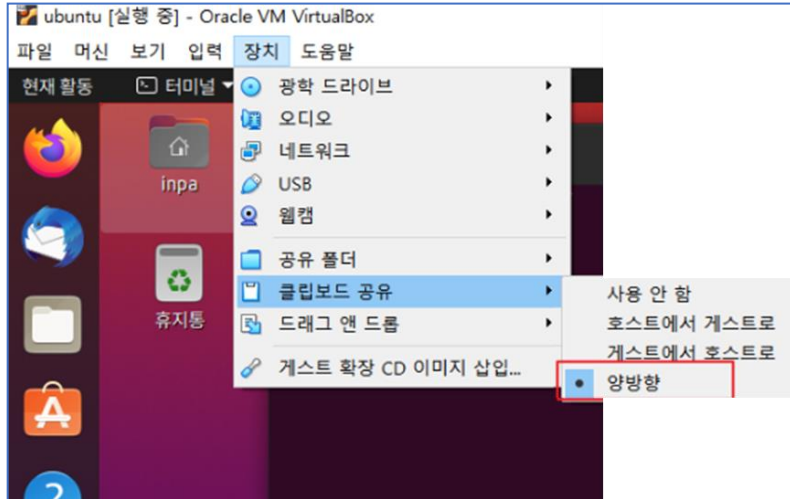
Ubuntu 설치 - 복사/붙여넣기 설정

■ 게스트 확장 설치



■ 클립보드 공유 설정

장치 → 클립보드 공유 → 양방향



■ 복사-붙여넣기 단축키

■ 가상머신에서 복사 및 붙여넣기

가상머신에서 복사 Ctrl + Insert

가상머신에 붙여넣기 Shift + Insert

■ 윈도우에서 가상머신으로 복사하기

윈도우에서 복사 : Ctrl + C

가상머신에서 붙여넣기 : Shift + Insert

■ 가상머신에서 윈도우로 복사하기

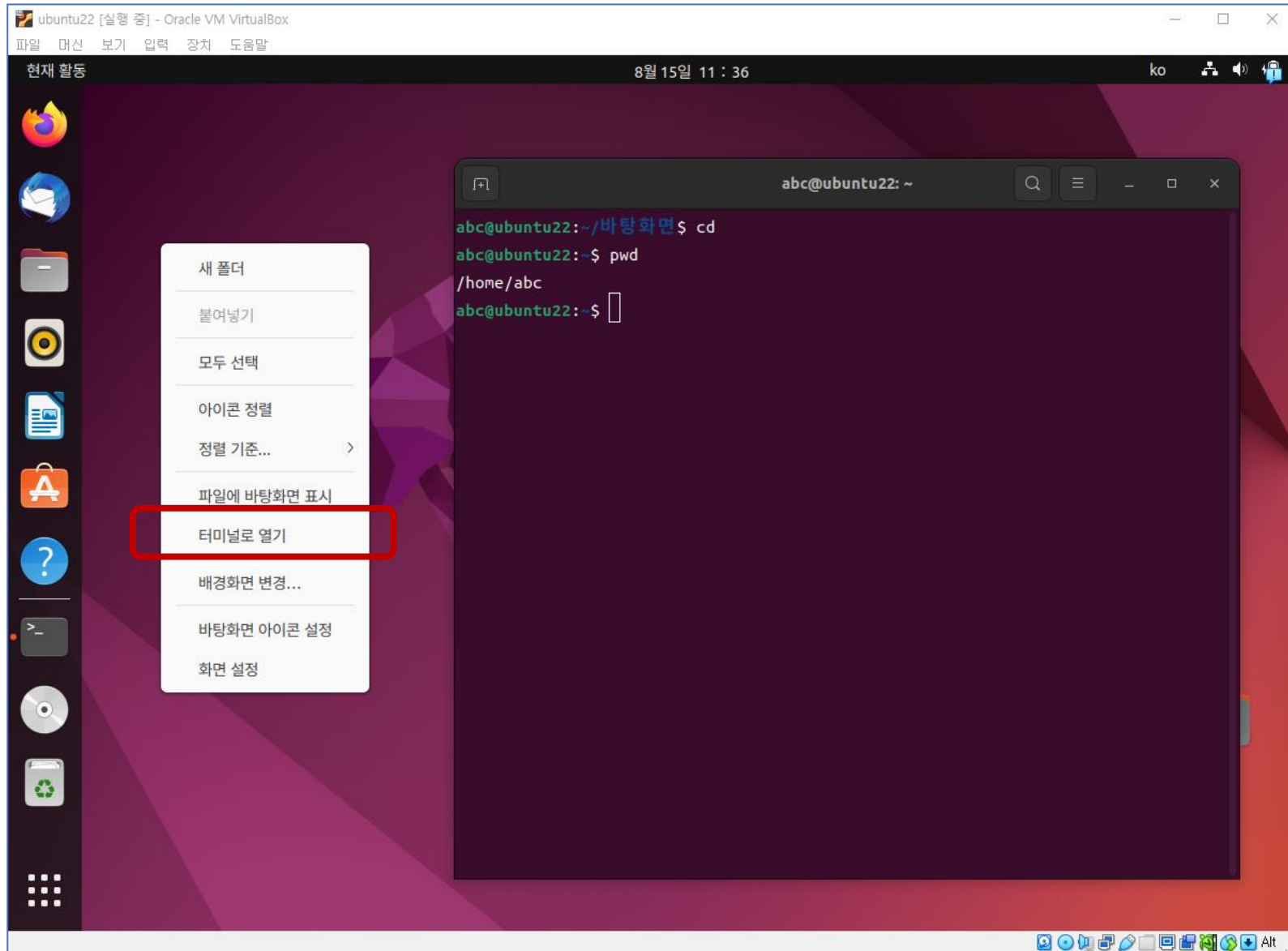
가상머신에서 복사 : Ctrl + Insert

윈도우에 붙여넣기 : Ctrl + V

참고 : <https://shorturl.at/avM89>

Hadoop 설치

Ubuntu 터미널 실행



Hadoop 설치

■ 시스템 패키지 업데이트 및 필요 프로그램 설치

```
sudo apt update  
sudo apt upgrade -y
```

■ Java 설치

```
sudo apt install openjdk-8-jdk  
sudo apt install maven  
java -version
```

```
abc@ubuntu22:~$ java -version  
openjdk version "1.8.0_382"  
OpenJDK Runtime Environment (build 1.8.0_382-8u382-ga-1~22.04.1-b05)  
OpenJDK 64-Bit Server VM (build 25.382-b05, mixed mode)
```

■ Java HOME 환경 설정

```
cd
```

```
vi .bashrc
```

```
-----  
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64  
-----
```

```
source .bashrc
```

■ SSH 서버 설치 및 설정(password 없이 ssh 연결하기)

```
sudo apt install openssh-server  
service ssh start  
service ssh status
```

```
sudo ufw allow 22  
ssh-keygen -t rsa
```

```
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys  
cat ~/.ssh/authorized_keys  
chmod 0660 ~/.ssh/authorized_keys
```

```
ssh localhost  
exit
```

Hadoop 설치

■ Apache Hadoop 다운로드 및 설치

<https://hadoop.apache.org/releases.html>

```
cd
wget https://downloads.apache.org/hadoop/common/hadoop-3.3.4/hadoop-3.3.4.tar.gz
tar zxvf hadoop-3.3.4.tar.gz
mv hadoop-3.3.4 hadoop3
```

■ Hadoop HOME 환경 설정

```
vi .bashrc
```

```
-----
export HADOOP_HOME=/home/abc/hadoop3
export PATH=$PATH:$HADOOP_HOME/bin
-----
```

```
source .bashrc
```

```
hadoop version
```


Hadoop 설정 : core-site.xml (클러스터 내 네임노드에서 실행되는 하둡 데몬에 관한 설정)

```
cd ~/hadoop3/etc/hadoop/
```

```
vi core-site.xml
```

```
-----  
<configuration>  
  <property>  
    <name>fs.defaultFS</name>  
    <value>hdfs://localhost:9000/</value>  
  </property>  
  <property>  
    <name>hadoop.http.staticuser.user</name>  
    <value>abc</value>  
  </property>  
</configuration>  
-----
```

Hadoop 설정 : hdfs-site.xml : (하둡 파일시스템에 관한 설정)

```
cd ~/hadoop3
```

```
mkdir dfs
```

```
mkdir dfs/name
```

```
mkdir dfs/namesecondary
```

```
mkdir dfs/data
```

```
vi ~/hadoop3/etc/hadoop/hdfs-site.xml
```

```
-----  
<configuration>  
  <property>  
    <name>dfs.replication</name>  
    <value>1</value>  
  </property>  
  <property>  
    <name>dfs.namenode.name.dir</name>  
    <value>file:///home/abc/hadoop3/dfs/name</value>  
  </property>  
  <property>  
    <name>dfs.datanode.data.dir</name>  
    <value>file:///home/abc/hadoop3/dfs/data</value>  
  </property>  
</configuration>  
-----
```

Hadoop 설정 : mapred-site.xml (맵리듀스에 관한 설정)

```
vi ~/hadoop3/etc/hadoop/mapred-site.xml
```

```
-----  
<configuration>  
  <property>  
    <name>mapreduce.framework.name</name>  
    <value>yarn</value>  
  </property>  
  <property>  
    <name>yarn.app.mapreduce.am.env</name>  
    <value>HADOOP_MAPRED_HOME=/home/abc/hadoop3</value>  
  </property>  
  <property>  
    <name>mapreduce.map.env</name>  
    <value>HADOOP_MAPRED_HOME=/home/abc/hadoop3</value>  
  </property>  
  <property>  
    <name>mapreduce.reduce.env</name>  
    <value>HADOOP_MAPRED_HOME=/home/abc/hadoop3</value>  
  </property>  
  <property>  
    <name>mapreduce.application.classpath</name>  
    <value>>[hadoop classpath 를 넣음]</value>  
  </property>  
</configuration>
```

[hadoop classpath 를 넣음] 이라고 된 부분에는
echo \$(hadoop classpath) 명령으로 출력된 결과를 넣음

Hadoop 설정 - yarn-site.xml (Resource Manager에 관한 설정)

```
vi ~/hadoop3/etc/hadoop/yarn-site.xml
```

```
-----  
<configuration>
```

```
  <property>
```

```
    <name>yarn.nodemanager.aux-services</name>
```

```
    <value>mapreduce_shuffle</value>
```

```
  </property>
```

```
</configuration>  
-----
```

Hadoop 설정 : hadoop-env.sh

```
vi ~/hadoop3/etc/hadoop/hadoop-env.sh
```

```
-----  
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64  
export HADOOP_CONF_DIR=${HADOOP_HOME}/etc/hadoop  
export HADOOP_MAPRED_HOME=${HADOOP_HOME}  
export HADOOP_COMMON_HOME=${HADOOP_HOME}  
export HADOOP_LOG_DIR=${HADOOP_HOME}/logs  
export HADOOP_PID_DIR=${HADOOP_HOME}/pids  
export HDFS_NAMENODE_USER="abc"  
export HDFS_DATANODE_USER="abc"  
export YARN_RESOURCEMANAGER_USER="abc"  
export YARN_NODEMANAGER_USER="abc"
```

■ NameNode 포맷

```
$HADOOP_HOME/bin/hdfs namenode -format
```

Hadoop 실행

■ Hadoop 실행

\$HADOOP_HOME/sbin/start-dfs.sh

\$HADOOP_HOME/sbin/start-yarn.sh

\$HADOOP_HOME/sbin/mr-jobhistory-daemon.sh start historyserver

■ Hadoop 실행 확인

jps

<http://localhost:9870/>

<http://localhost:8088/>

<http://localhost:8042/>

■ Hadoop 종료

\$HADOOP_HOME/sbin/stop-dfs.sh

\$HADOOP_HOME/sbin/stop-yarn.sh

\$HADOOP_HOME/sbin/mr-jobhistory-daemon.sh stop historyserver

```
abc@abc-VirtualBox:~/hadoop3$ jps
6739 DataNode
7144 ResourceManager
6616 NameNode
7259 NodeManager
5597 SecondaryNameNode
7741 Jps
7598 JobHistoryServer
abc@abc-VirtualBox:~/hadoop3$
abc@abc-VirtualBox:~/hadoop3$
```

Overview 'localhost:9000' (✓active)	
Started:	Thu Mar 23 11:22:00 +0900 202
Version:	3.3.4, ra585a73c3e02ac62350c
Compiled:	Fri Jul 29 21:32:00 +0900 2022
Cluster ID:	CID-59e14b76-24fe-4bde-bbc3-5
Block Pool ID:	BP-449612038-127.0.1.1-16795

Cluster Metrics

Apps Submitted	A
0	0

Cluster Nodes Metrics

Active Nodes
1

Scheduler Metrics

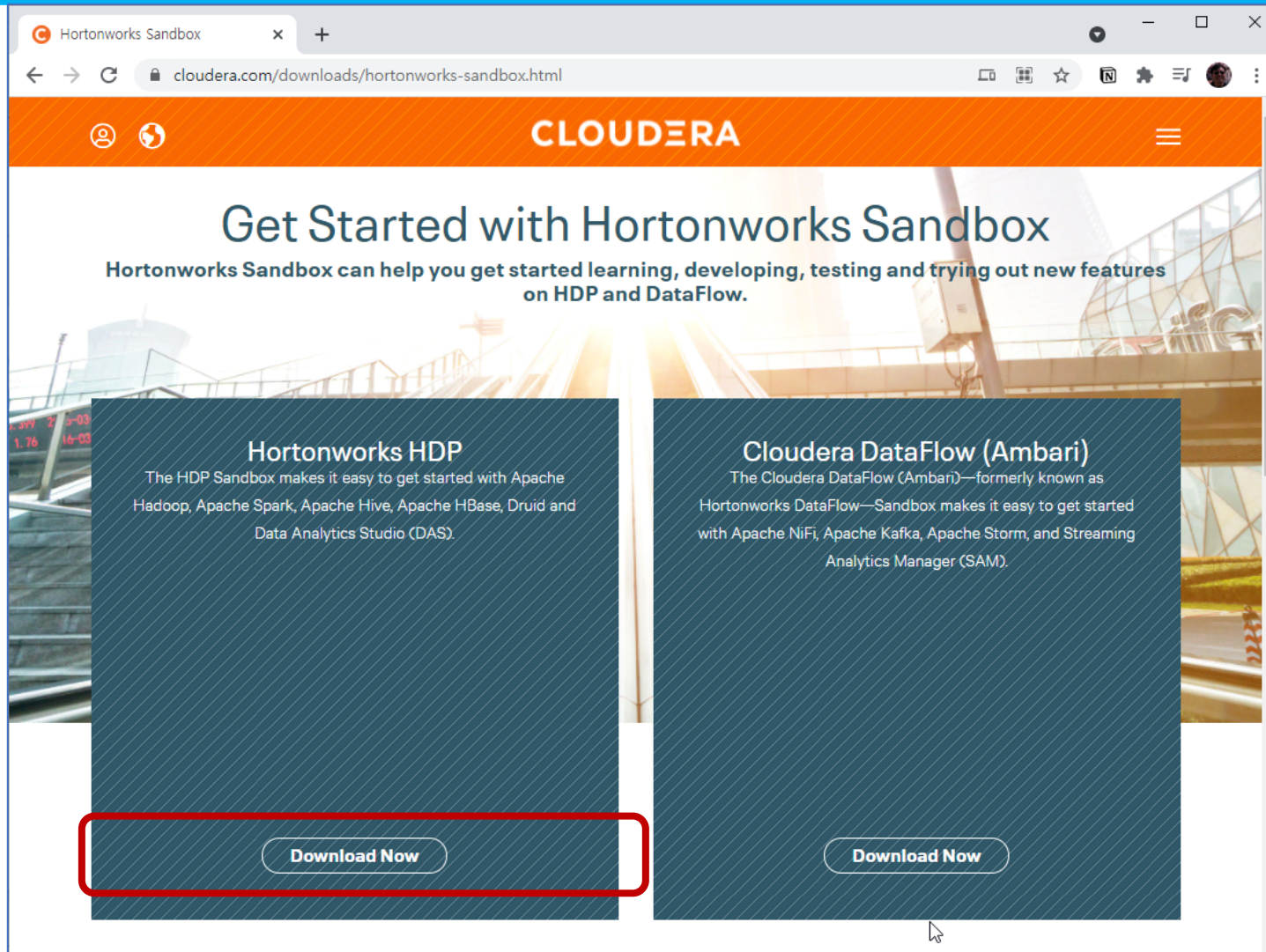
Scheduler Type
Capacity Scheduler

■ Hadoop 완전분산모드(Fully-Distributed) Cluster 구성 : <https://bit.ly/40nN35E> 참고

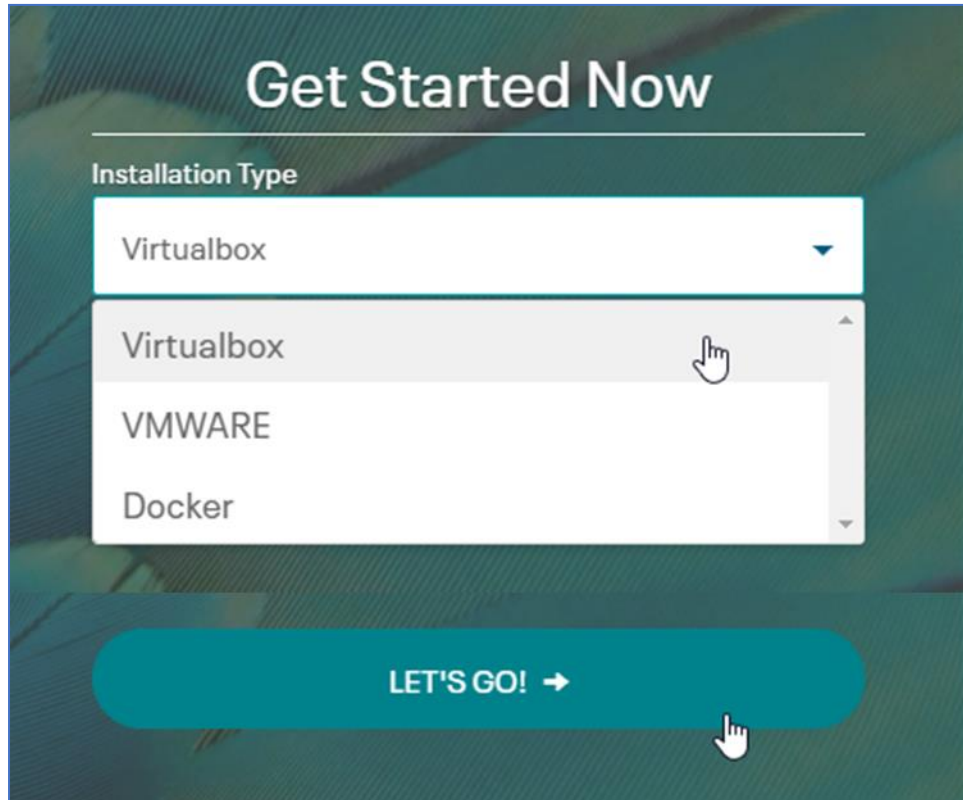
HDP(Hortonworks Data Platform) 설치

HDP 다운로드

<https://www.cloudera.com/downloads/hortonworks-sandbox.html>



HDP 다운로드



Sign in or complete our product interest form to continue.

Sign In

For self-learning

First Name: Danny

Last Name: Park

Business Email:

Company:

Job Title:

Phone:

HDP 다운로드

Thank you for choosing Hortonworks Data Platform (HDP) on Sandbox

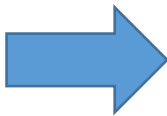
Sandbox HDP Virtualbox Downloads

HDP Sandbox 3.0.1 (Latest)

[Install Guide on VirtualBox](#)

Older Versions

- **2.6.5**
- 2.5.0

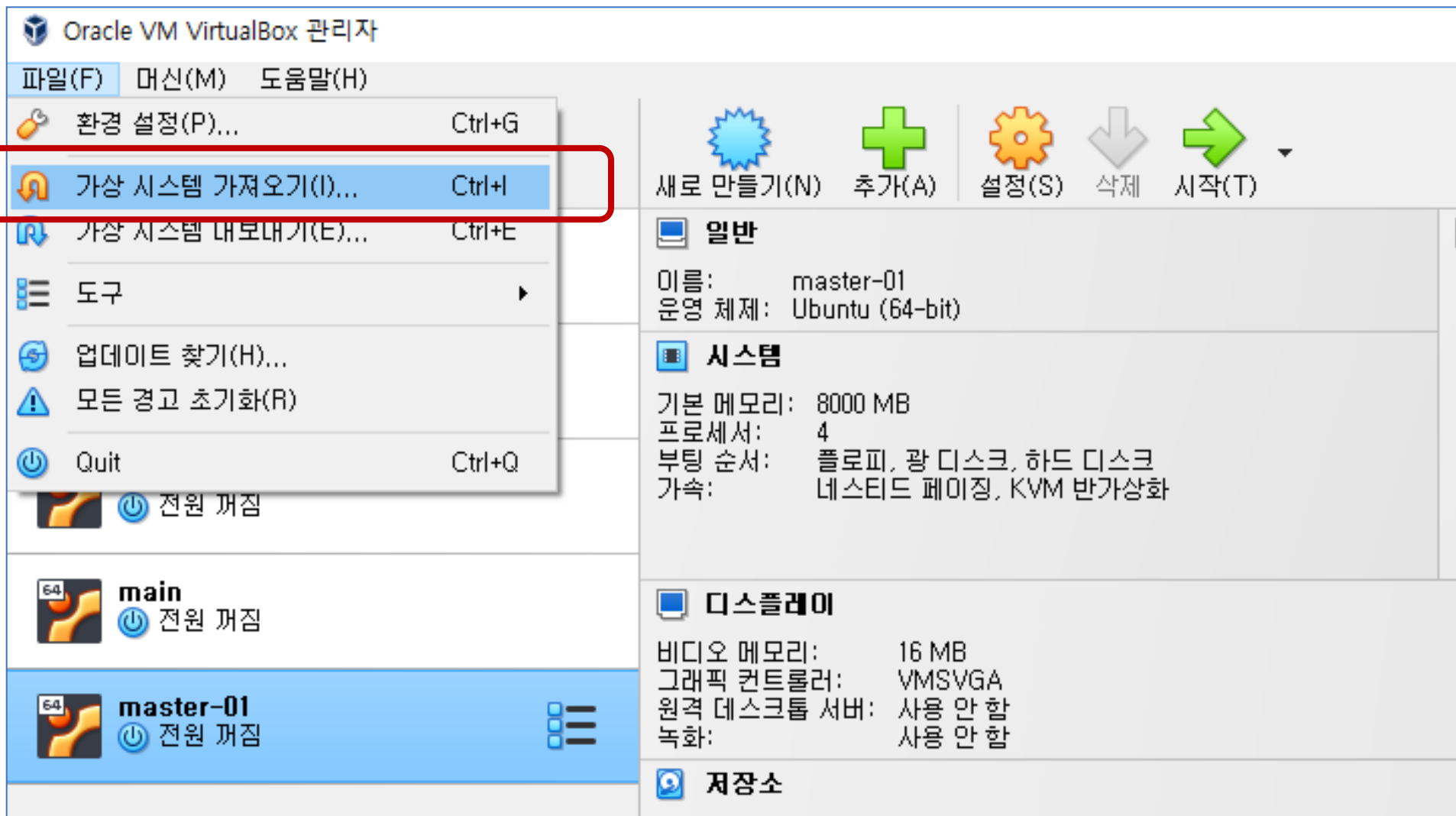


HDP_2.6.5_virtualbox_180626.ova 15.0GB

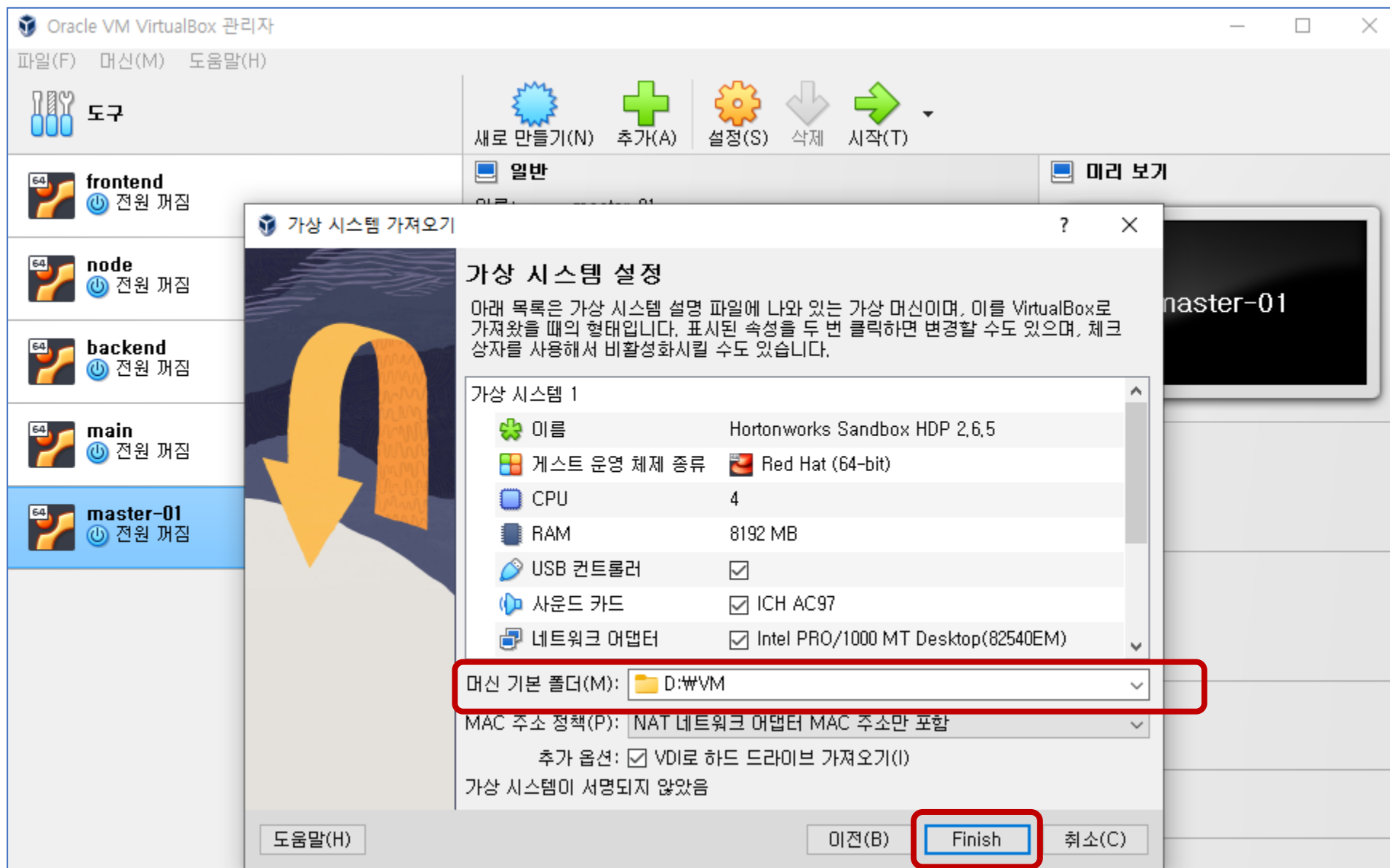
Direct Download Link : https://archive.cloudera.com/hwx-sandbox/hdp/hdp-2.6.5/HDP_2.6.5_virtualbox_180626.ova

HDP 설치

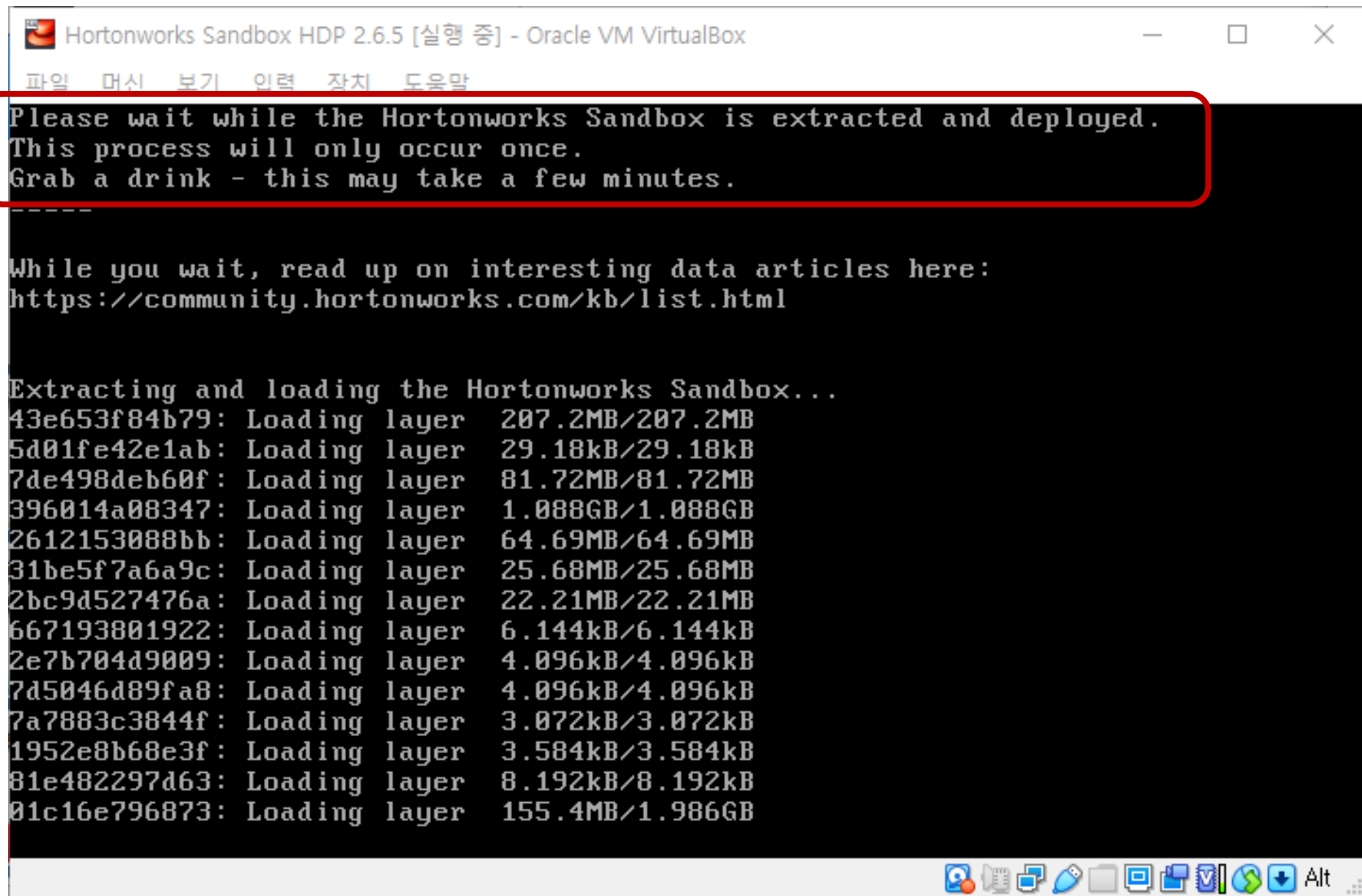
VirtualBox 실행



HDP 설치



HDP 설치



```
Hortonworks Sandbox HDP 2.6.5 [실행 중] - Oracle VM VirtualBox
파일  머신  보기  입력  장치  도움말

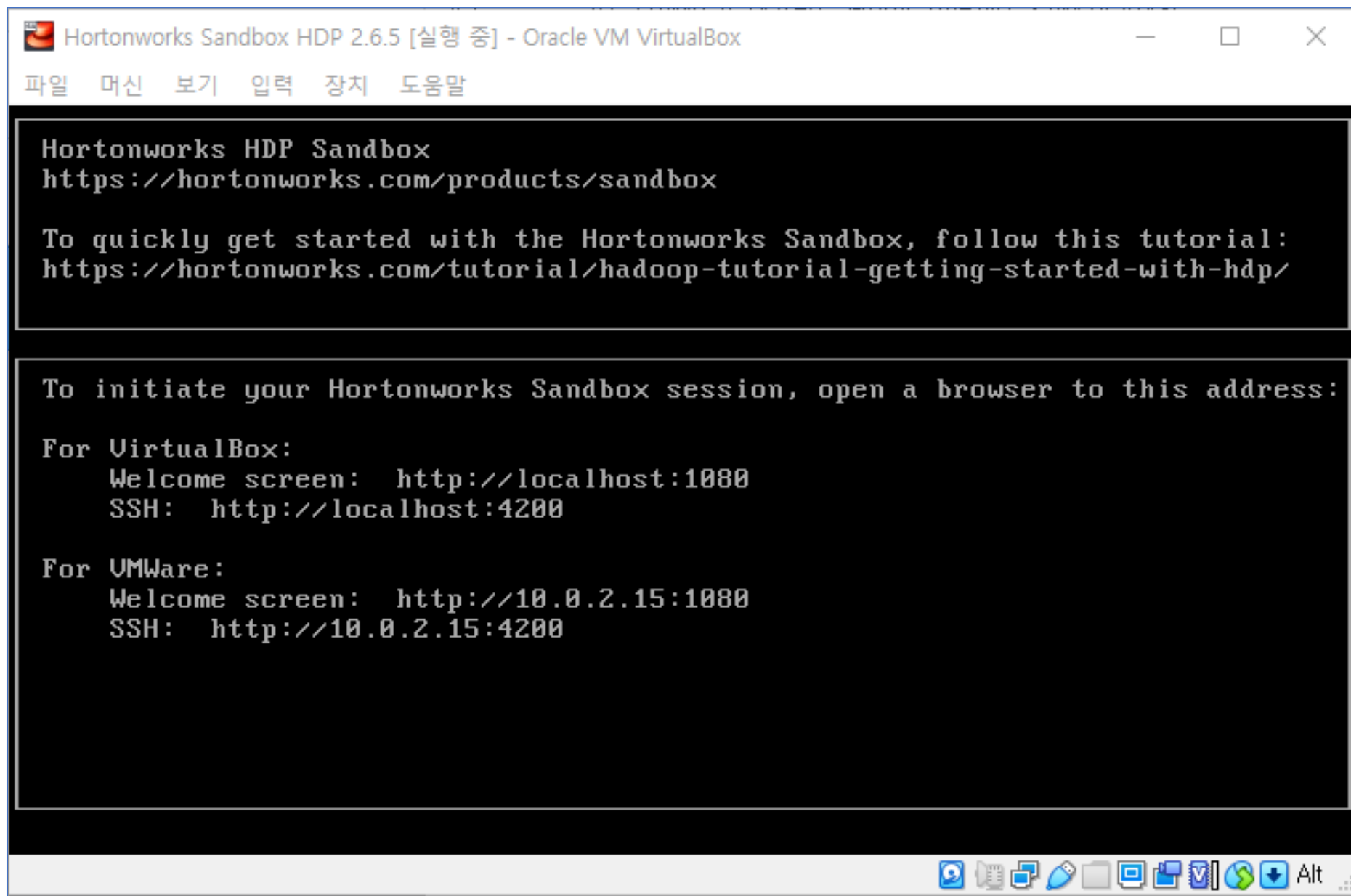
Please wait while the Hortonworks Sandbox is extracted and deployed.
This process will only occur once.
Grab a drink - this may take a few minutes.

-----

While you wait, read up on interesting data articles here:
https://community.hortonworks.com/kb/list.html

Extracting and loading the Hortonworks Sandbox...
43e653f84b79: Loading layer 207.2MB/207.2MB
5d01fe42e1ab: Loading layer 29.18kB/29.18kB
7de498deb60f: Loading layer 81.72MB/81.72MB
396014a08347: Loading layer 1.088GB/1.088GB
2612153088bb: Loading layer 64.69MB/64.69MB
31be5f7a6a9c: Loading layer 25.68MB/25.68MB
2bc9d527476a: Loading layer 22.21MB/22.21MB
667193801922: Loading layer 6.144kB/6.144kB
2e7b704d9009: Loading layer 4.096kB/4.096kB
7d5046d89fa8: Loading layer 4.096kB/4.096kB
7a7883c3844f: Loading layer 3.072kB/3.072kB
1952e8b68e3f: Loading layer 3.584kB/3.584kB
81e482297d63: Loading layer 8.192kB/8.192kB
01c16e796873: Loading layer 155.4MB/1.986GB
```

HDP 설치 완료



Thank you