



HORTONWORKS



목차

1

Hortonworks 설치

2

Hortonworks 접속

3

Hortonworks Web GUI (Ambari)

4

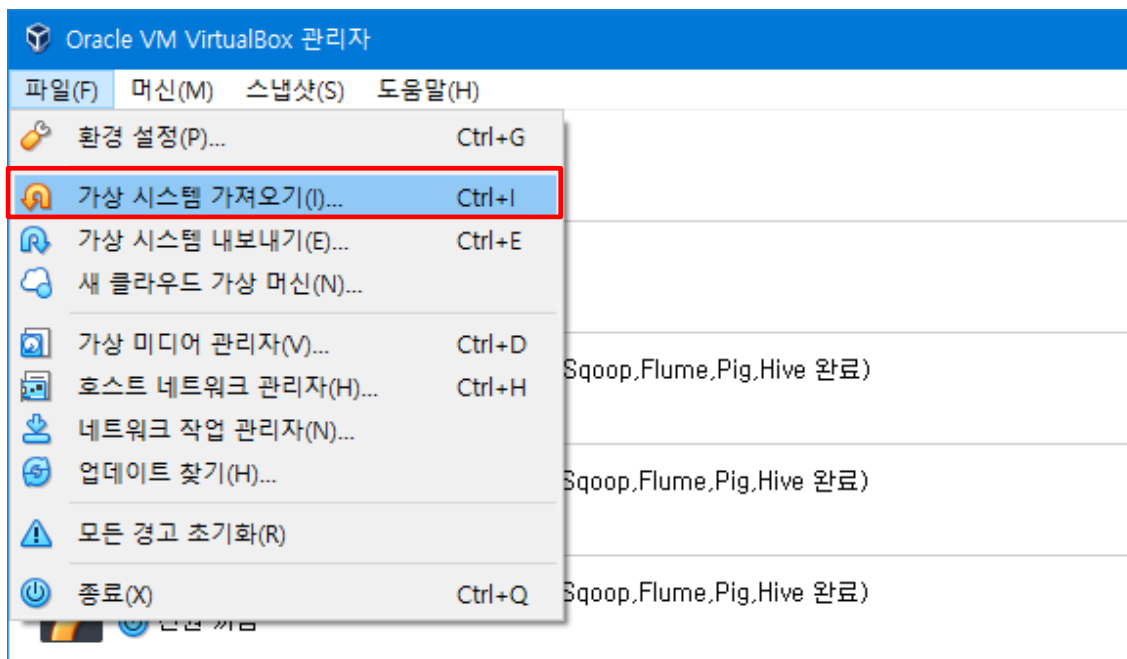
데이터 복사



PART 1. HORTONWORKS 설치

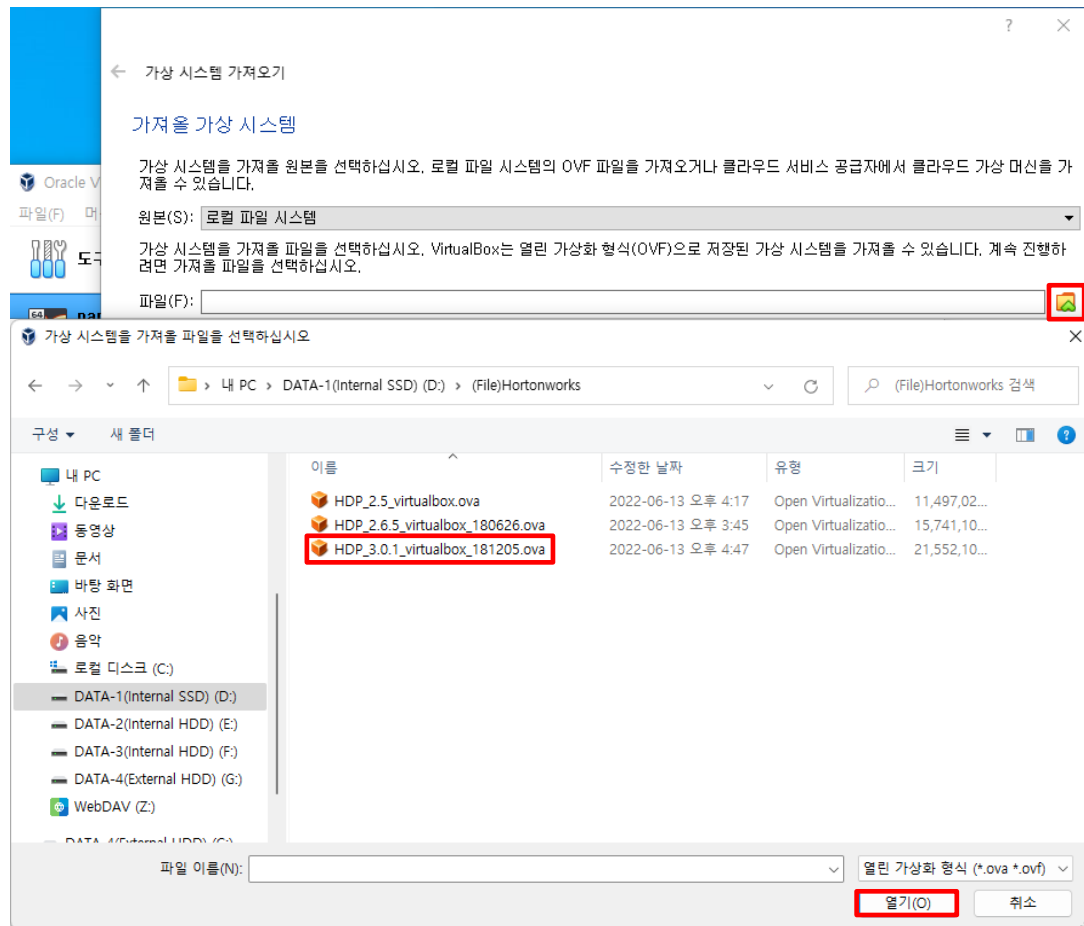


PART 1. HORTONWORKS 설치



- OVA 파일 로드
 - VirtualBox 상단메뉴의 파일 클릭
 - 가상 시스템 가져오기 클릭

PART 1. HORTONWORKS 설치



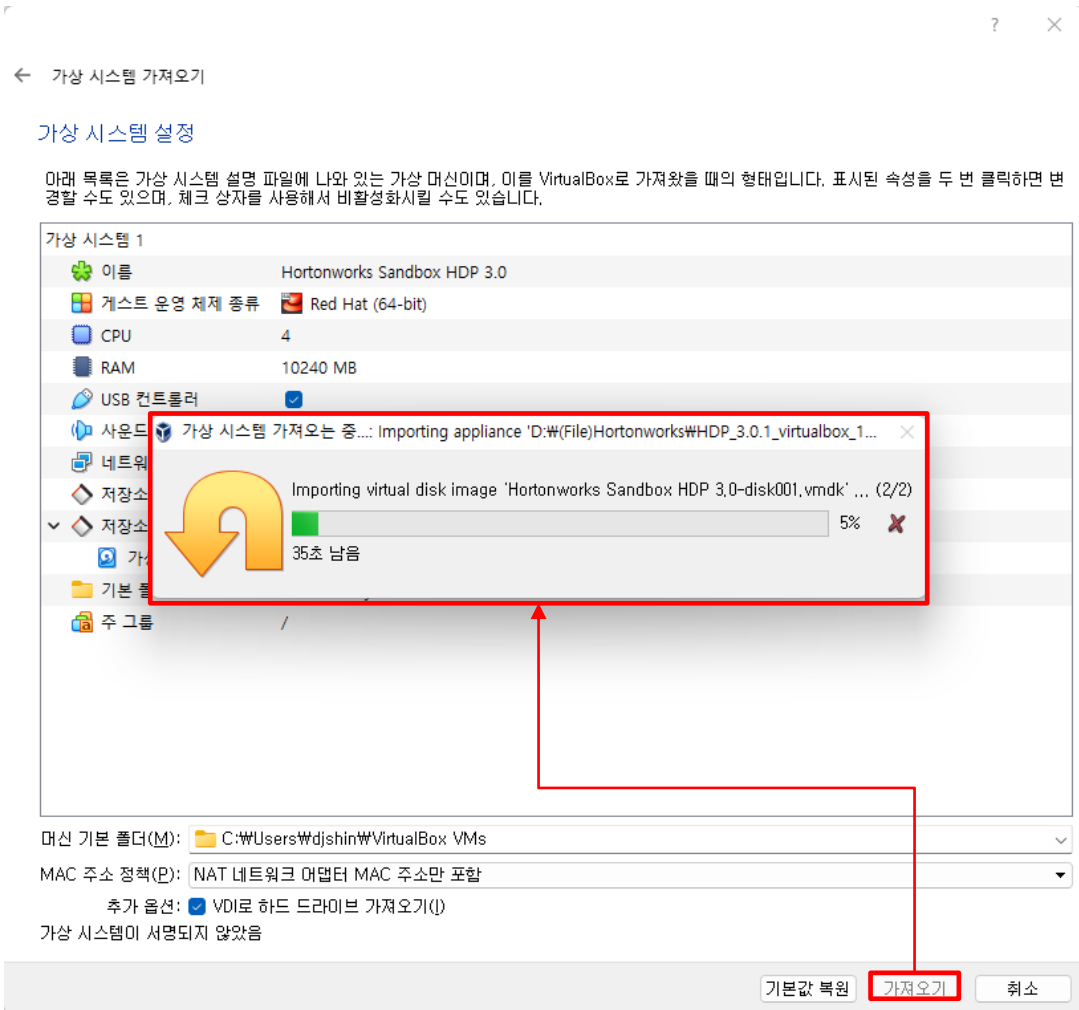
- OVA 파일 로드
 - 파일 란 우측 파일 탐색기 클릭
 - HDP_3.0.1_virtualbox_181205.ova 파일 선택 후 열기

PART 1. HORTONWORKS 설치



- OVA 파일 로드
 - 하단 다음 버튼 클릭

PART 1. HORTONWORKS 설치



■ OVA 파일 로드

- 하단 가져오기 버튼 클릭
- 상단의 가상 시스템 가져오는 중 출력됨

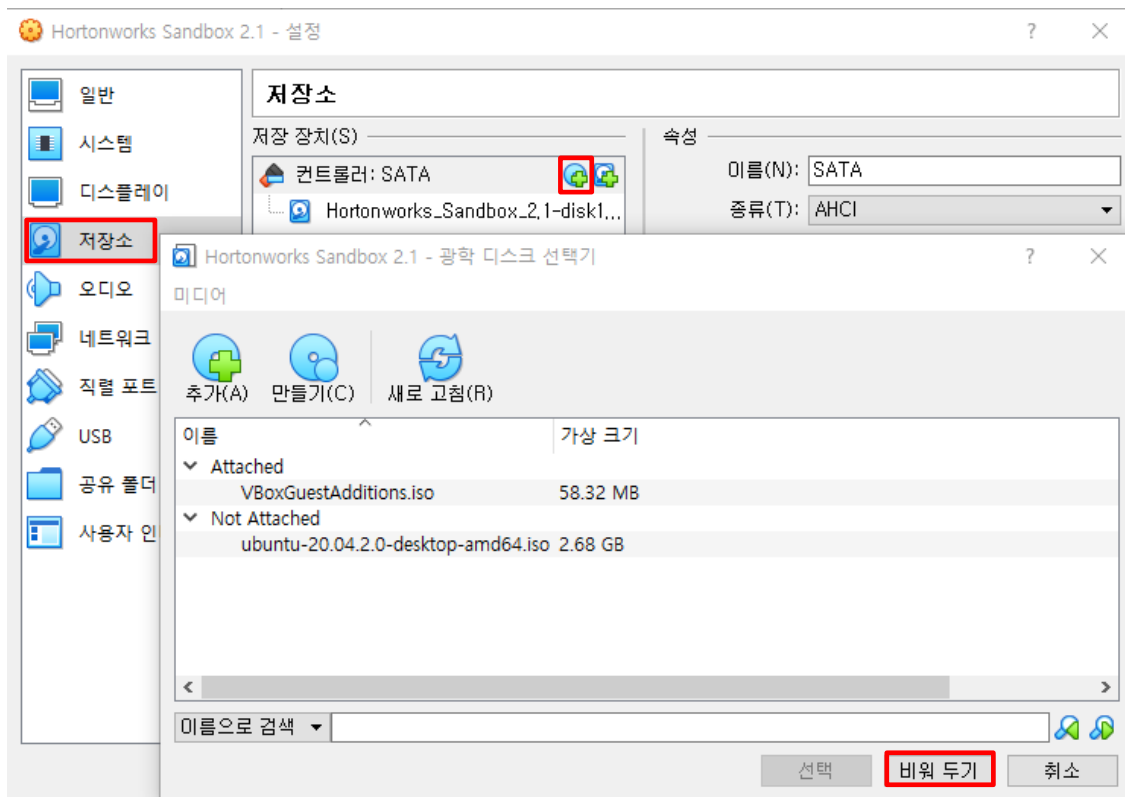
PART 1. HORTONWORKS 설치



■ 설정

- 로드가 완료된 SandBox의 오른쪽 마우스 키 클릭
- 설정 클릭

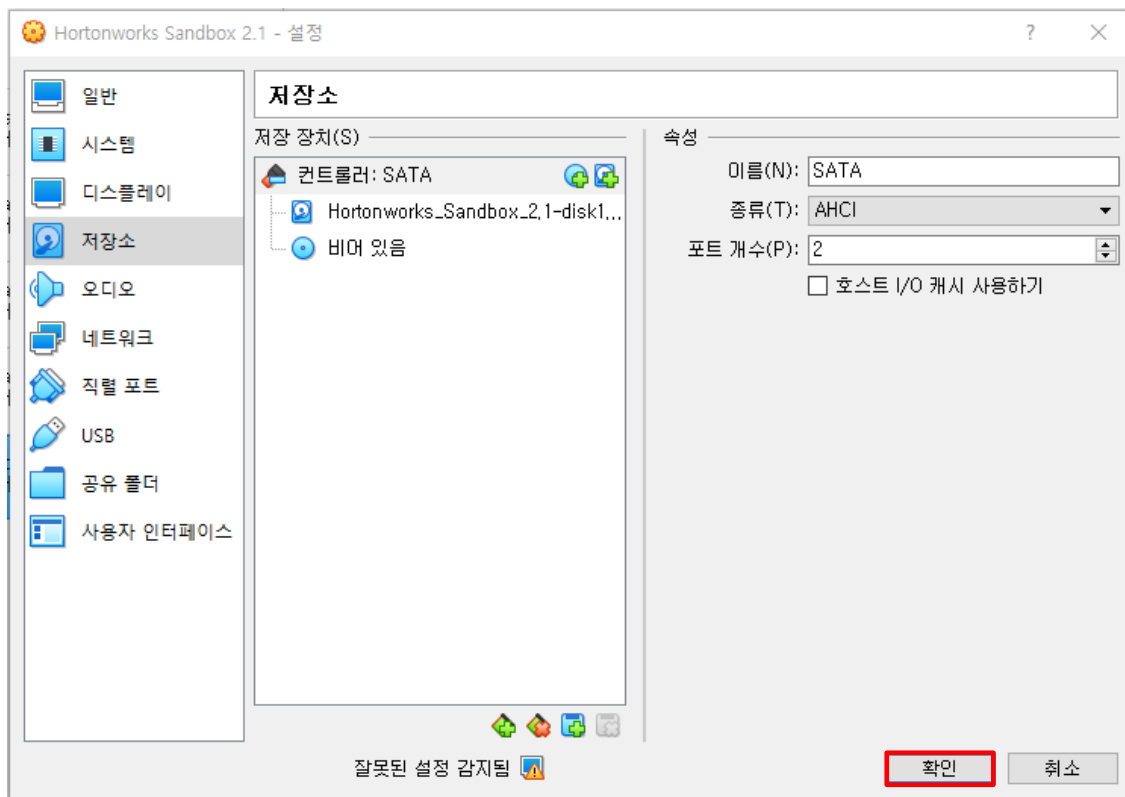
PART 1. HORTONWORKS 설치



■ 설정

- 왼쪽 메뉴 중 저장소 클릭
- 저장 장치 탭의 컨트롤러: SATA 우측의 CD 디스크 아이콘 클릭
- 하단 비워 두기 클릭

PART 1. HORTONWORKS 설치



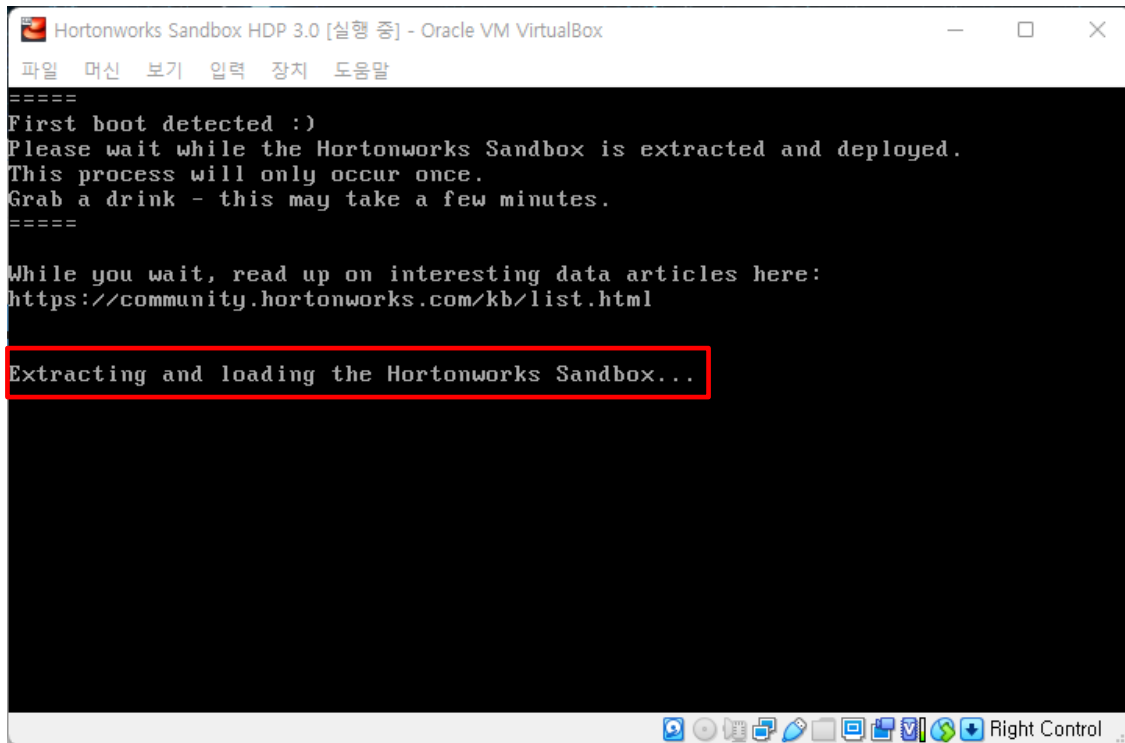
- 설정
 - 확인 클릭

PART 1. HORTONWORKS 설치



■ 시작 클릭

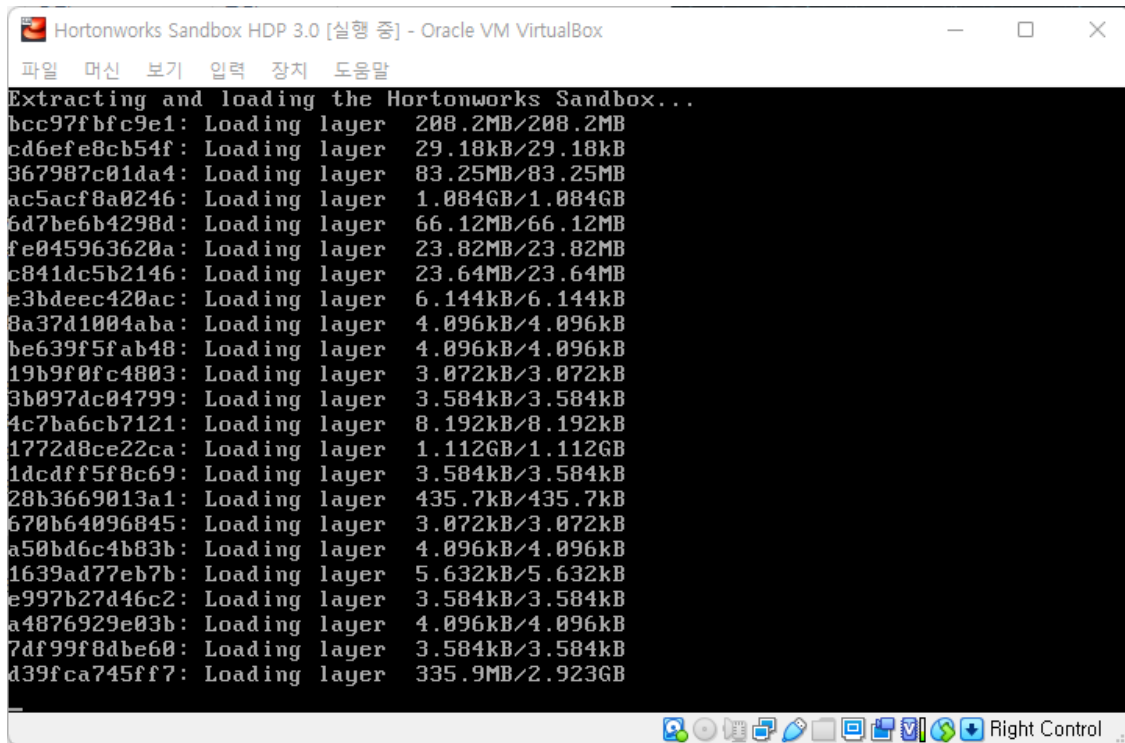
PART 1. HORTONWORKS 설치



```
====  
First boot detected :)  
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...
```

- Extracting and loading the Hortonworks Sandbox... 메시지가 나오면 대기
 - 필요한 파일을 설치하고, 내부 환경을 설정하는 과정
 - 20~30분 소요

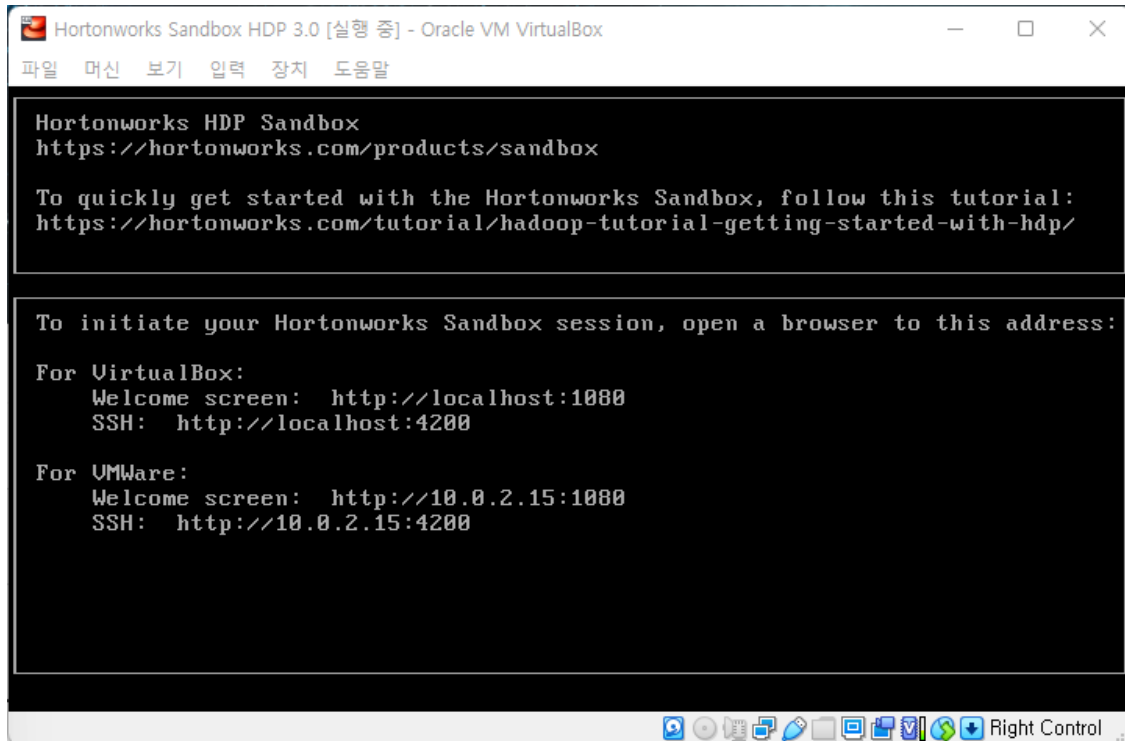
PART 1. HORTONWORKS 설치



```
Hortonworks Sandbox HDP 3.0 [실행 중] - Oracle VM VirtualBox
파일  머신  보기  입력  장치  도움말
Extracting and loading the Hortonworks Sandbox...
bcc97fbfc9e1: Loading layer 208.2MB/208.2MB
cd6efe8cb54f: Loading layer 29.18kB/29.18kB
367987c01da4: Loading layer 83.25MB/83.25MB
ac5acf8a0246: Loading layer 1.084GB/1.084GB
6d7be6b4298d: Loading layer 66.12MB/66.12MB
fe045963620a: Loading layer 23.82MB/23.82MB
c841dc5b2146: Loading layer 23.64MB/23.64MB
e3bdeec420ac: Loading layer 6.144kB/6.144kB
8a37d1004aba: Loading layer 4.096kB/4.096kB
be639f5fab48: Loading layer 4.096kB/4.096kB
19b9f0fc4803: Loading layer 3.072kB/3.072kB
3b097dc04799: Loading layer 3.584kB/3.584kB
4c7ba6cb7121: Loading layer 8.192kB/8.192kB
1772d8ce22ca: Loading layer 1.112GB/1.112GB
1dcdf5f8c69: Loading layer 3.584kB/3.584kB
28b3669013a1: Loading layer 435.7kB/435.7kB
670b64096845: Loading layer 3.072kB/3.072kB
a50bd6c4b83b: Loading layer 4.096kB/4.096kB
1639ad77eb7b: Loading layer 5.632kB/5.632kB
e997b27d46c2: Loading layer 3.584kB/3.584kB
a4876929e03b: Loading layer 4.096kB/4.096kB
7df99f8dbe60: Loading layer 3.584kB/3.584kB
d39fca745ff7: Loading layer 335.9MB/2.923GB
```

- 필요한 파일을 설치하고, 내부 환경 설정 중

PART 1. HORTONWORKS 설치



```
Hortonworks HDP Sandbox
https://hortonworks.com/products/sandbox

To quickly get started with the Hortonworks Sandbox, follow this tutorial:
https://hortonworks.com/tutorial/hadoop-tutorial-getting-started-with-hdp/

To initiate your Hortonworks Sandbox session, open a browser to this address:

For VirtualBox:
Welcome screen: http://localhost:1080
SSH: http://localhost:4200

For VMWare:
Welcome screen: http://10.0.2.15:1080
SSH: http://10.0.2.15:4200
```

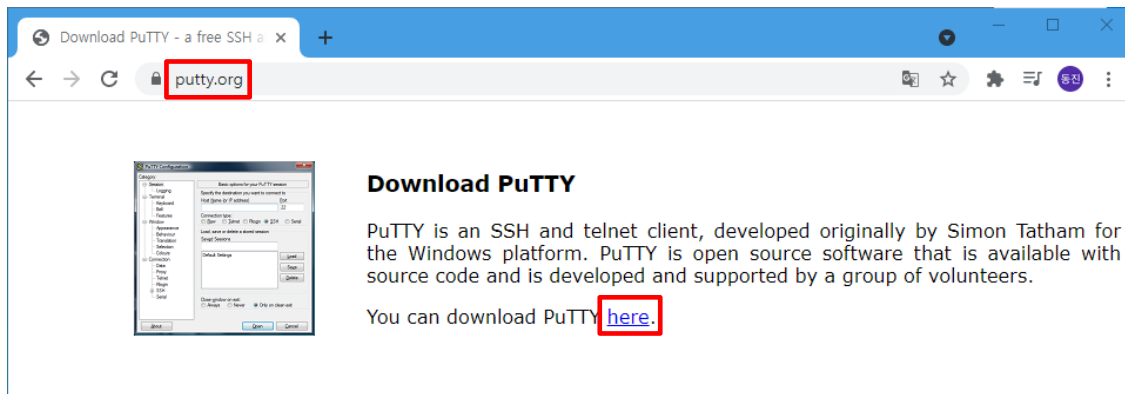
- 설치가 완료된 모습
 - Hortonworks HDP 3.0 버전 부터는 직접 로컬 로그인
이 불가능하여 Putty를 이용하여 로그인



PART 2. HORTONWORKS 접속

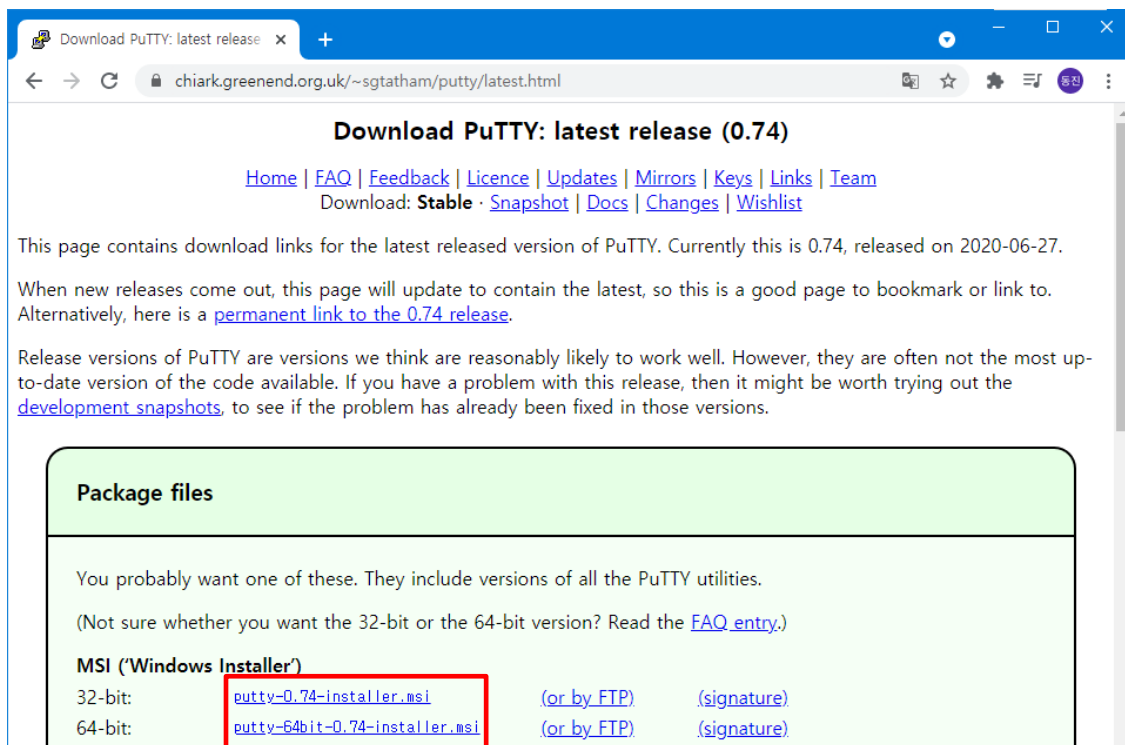


PART 2. HORTONWORKS 접속



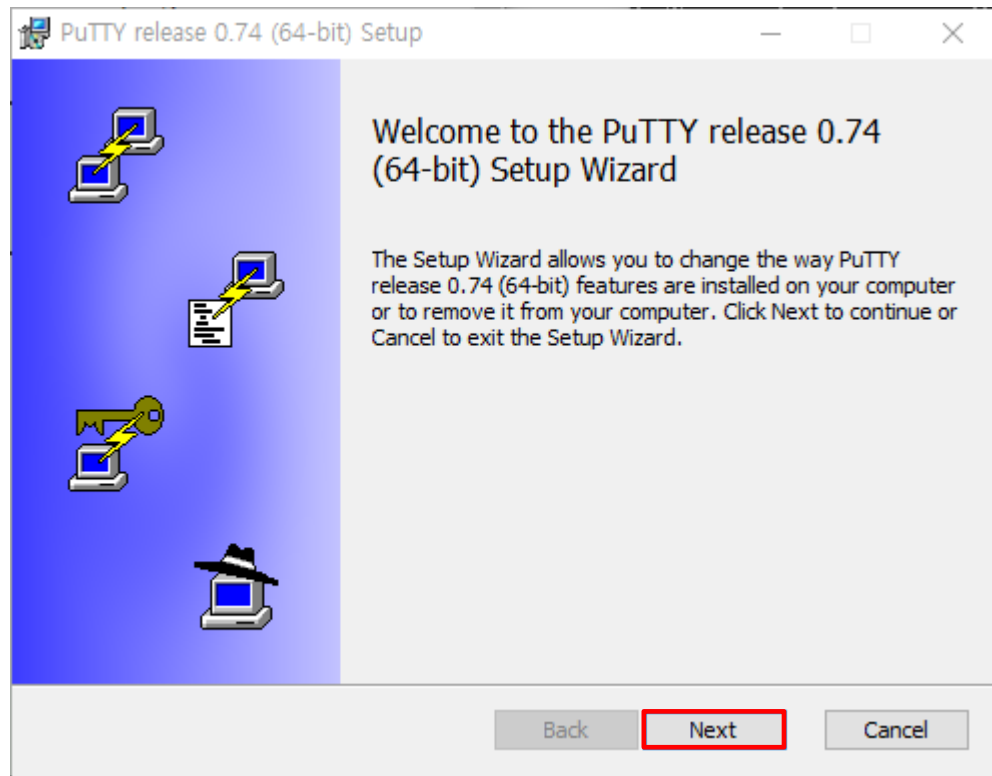
- 원격 접속 프로그램 설치
 - 인터넷 브라우저에서 putty.org 이동
 - 하단 here 링크 클릭

PART 2. HORTONWORKS 접속



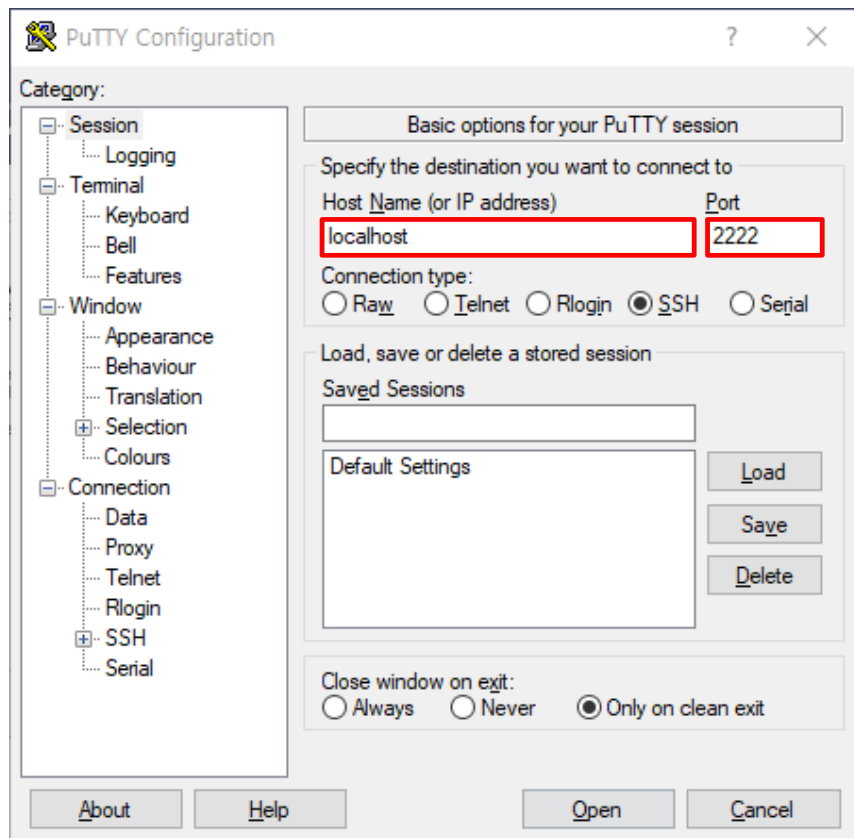
- 원격 접속 프로그램 설치
 - 아래 MSI ('Windows Installer') 해당 사용 OS의 32-bit or 64-bit에 맞춰 다운로드

PART 2. HORTONWORKS 접속



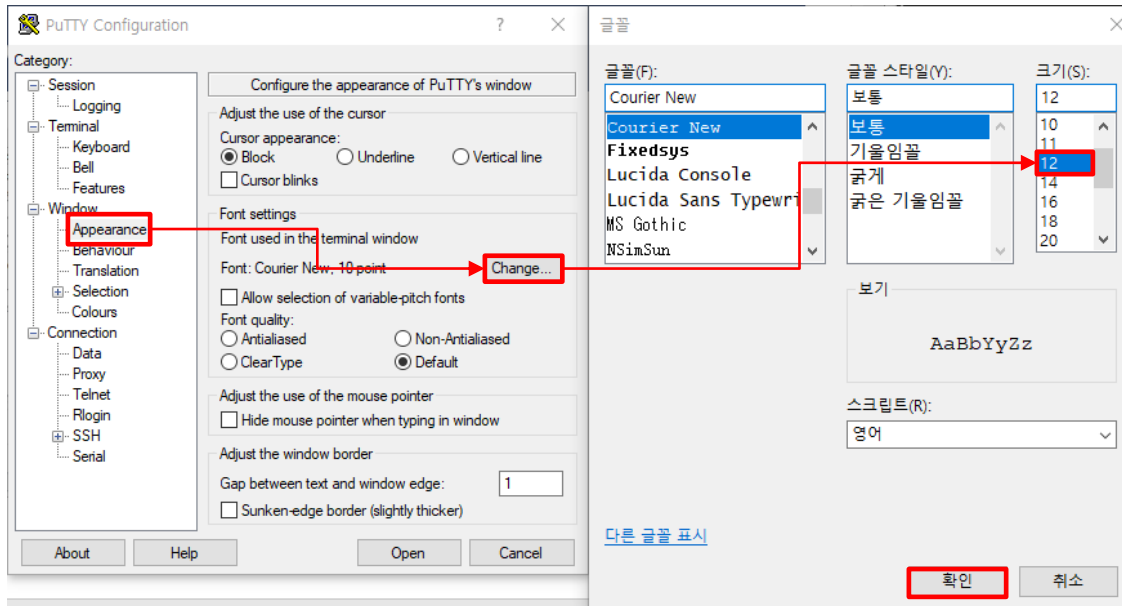
- 원격 접속 프로그램 설치
 - 연속적으로 등장하는 Next 버튼만 선택하여 설치 완료

PART 2. HORTONWORKS 접속



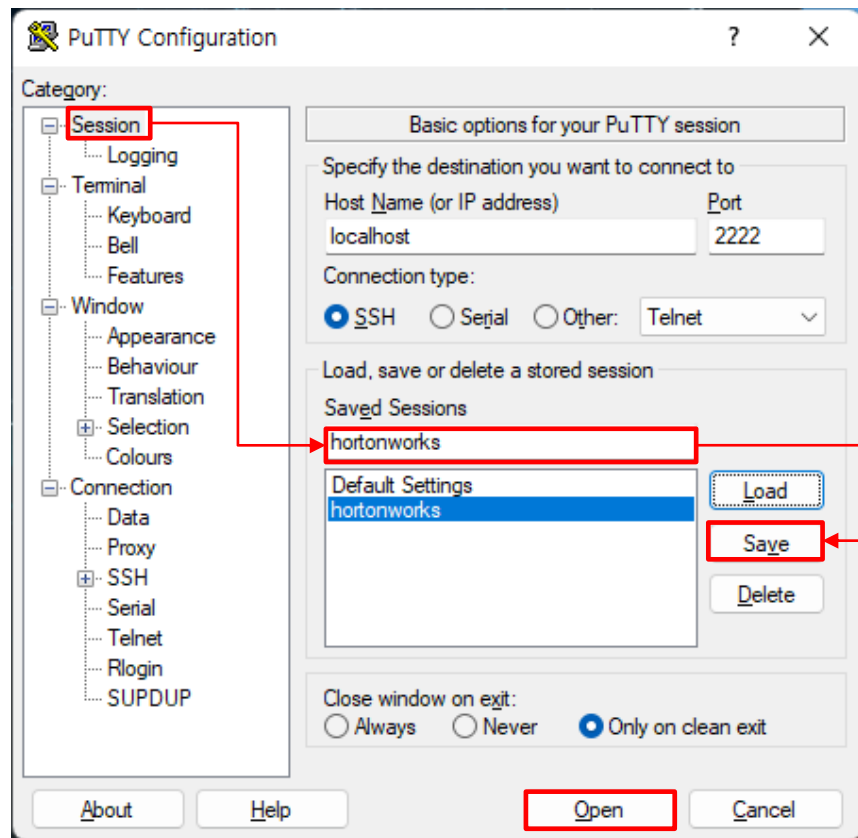
- 원격 접속 프로그램 설정
 - Host Name: localhost
 - Port: 2222

PART 2. HORTONWORKS 접속



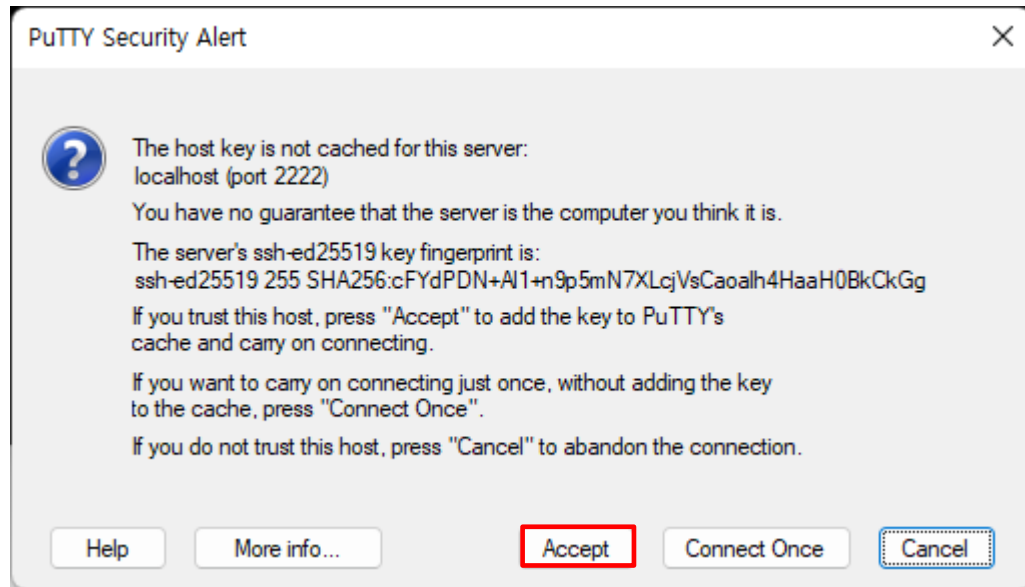
- 원격 접속 프로그램 설정
 - 좌측 Window의 Appearance 클릭
 - Change 클릭
 - 글꼴, 글꼴 스타일, 크기 설정 후 확인 버튼 클릭
 - 보통 Putty의 글자 폰트 크기만 설정

PART 2. HORTONWORKS 접속



- 원격 접속 프로그램 설정
 - 좌측 Session 클릭
 - Saved Sessions: hortonworks
 - Save 클릭
 - Open 클릭
- 위 과정을 통해 접속할 Putty 설정을 미리 해두고 나중에 설정값을 그대로 불러와 편하게 사용
 - 추후 재사용시 저장했던 Sessions 이름을 클릭하고, Load를 클릭하여 설정 불러오기

PART 2. HORTONWORKS 접속



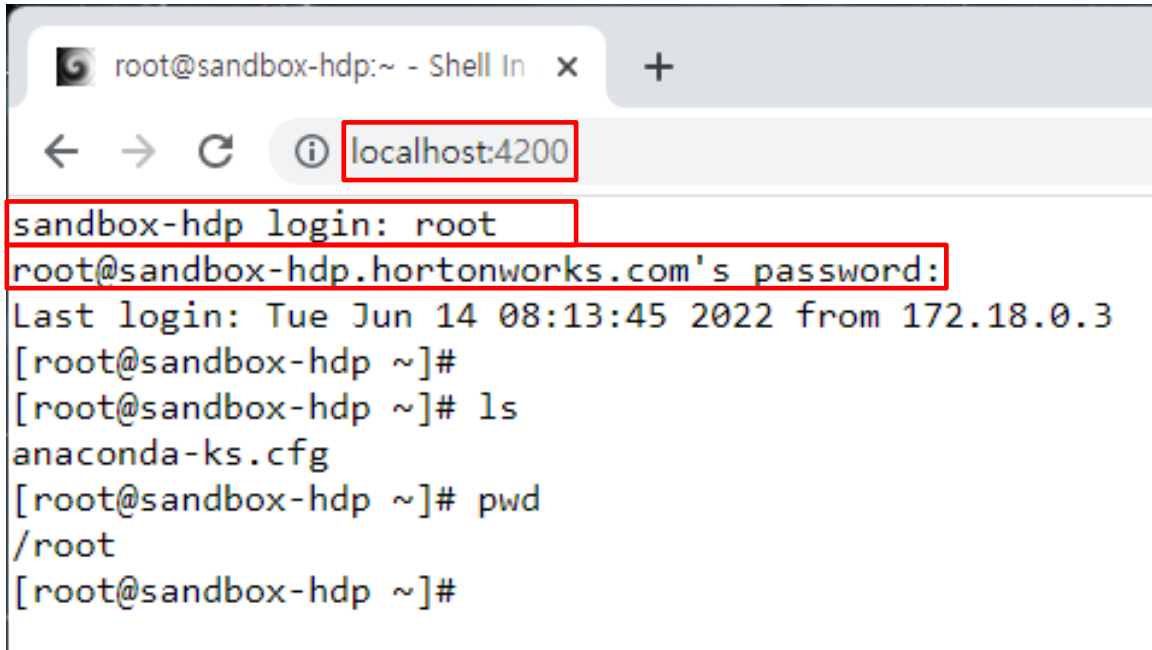
- 첫 접속이라면 하단과 같이 경고 창 출력
 - PuTTY Security Alert 창이 출력되며, 첫 SSH 프로토콜을 통해 접속하기 때문에 키를 확인하는 과정
 - 출력 시 Yes를 클릭

PART 2. HORTONWORKS 접속

```
root@sandbox-hdp:~  
login as: root  
root@localhost's password:  
You are required to change your password immediately (root enforced)  
Last login: Tue Jun 14 08:13:28 2022 from 172.18.0.3  
Changing password for root.  
(current) UNIX password:  
New password:  
Retype new password:  
[root@sandbox-hdp ~]# ls  
anaconda-ks.cfg  
[root@sandbox-hdp ~]#
```

- 로그인 접속
 - login as: root
 - root@localhost's password: hadoop
 - 보안상 첫 로그인 시 비밀번호를 바꿔야함
 - (current) UNIX password: hadoop
 - New password: bigdata12
 - Retype new password: bigdata12
- 명령어 확인
 - ls

PART 2. HORTONWORKS 접속



The screenshot shows a web browser window with a single tab titled "root@sandbox-hdp:~ - Shell In". The address bar contains "localhost:4200". The terminal content is as follows:

```
sandbox-hdp login: root
root@sandbox-hdp.hortonworks.com's password:
Last login: Tue Jun 14 08:13:45 2022 from 172.18.0.3
[root@sandbox-hdp ~]#
[root@sandbox-hdp ~]# ls
anaconda-ks.cfg
[root@sandbox-hdp ~]# pwd
/root
[root@sandbox-hdp ~]#
```

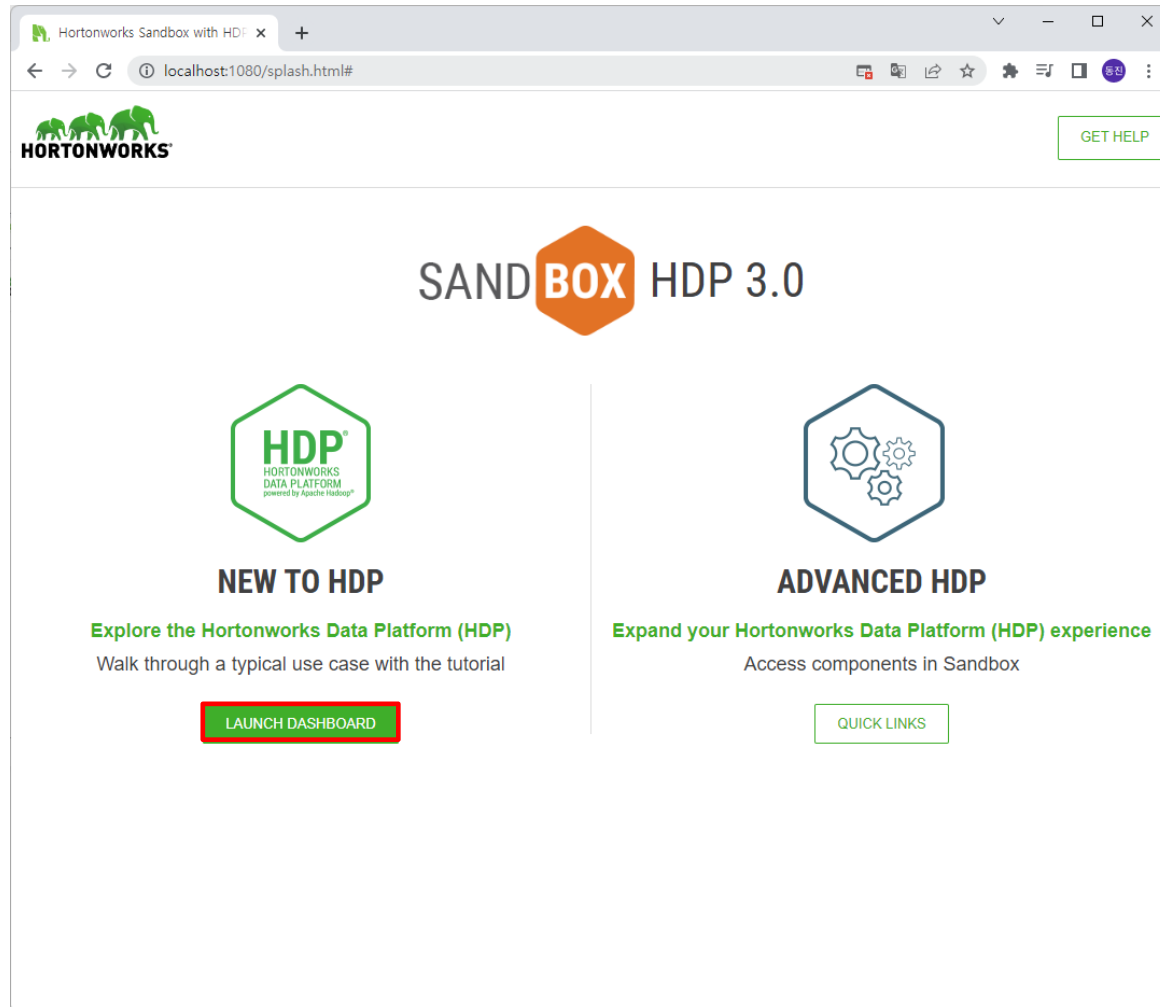
- 웹 브라우저(크롬, Edge 등)에서 터미널 접속
 - localhost:4200
 - ID, PW 입력 후 로그인
- 본인이 편한 인터페이스를 선택
 - **Putty를 추천**



PART 3. HORTONWORKS WEB GUI (AMBARI)

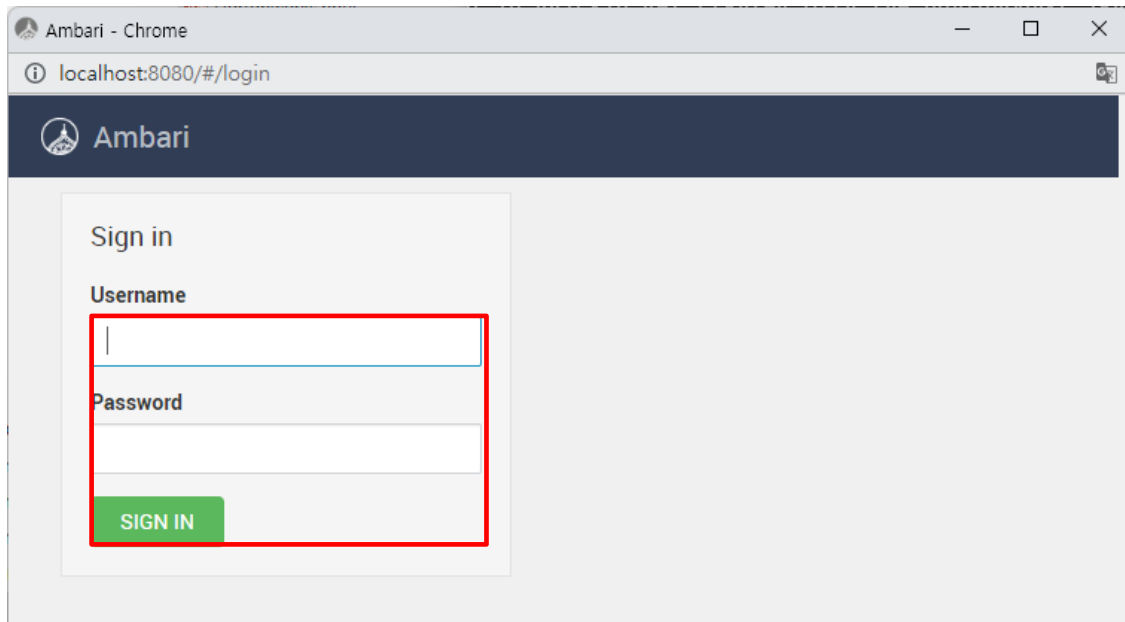


PART 3. HORTONWORKS WEB GUI (AMBARI)



- 인터넷 브라우저 실행
 - localhost:1080 접속
 - LAUNCH DASHBOARD 클릭

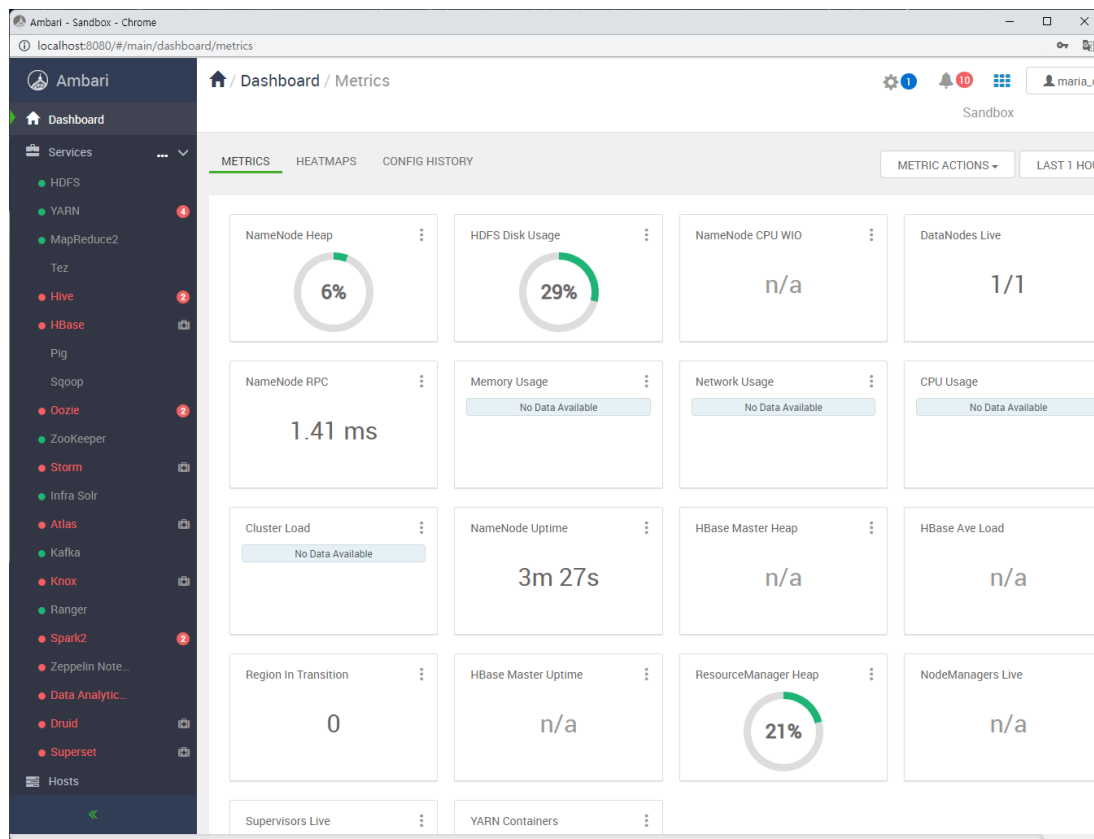
PART 3. HORTONWORKS WEB GUI (AMBARI)



The screenshot shows a web browser window titled "Ambari - Chrome" with the address bar displaying "localhost:8080/#/login". The page header features the Ambari logo and name. The main content area is titled "Sign in" and contains a login form. The form has two input fields: "Username" and "Password". A red rectangular box highlights the "Username" field, the "Password" field, and the "SIGN IN" button below them. The "SIGN IN" button is green with white text.

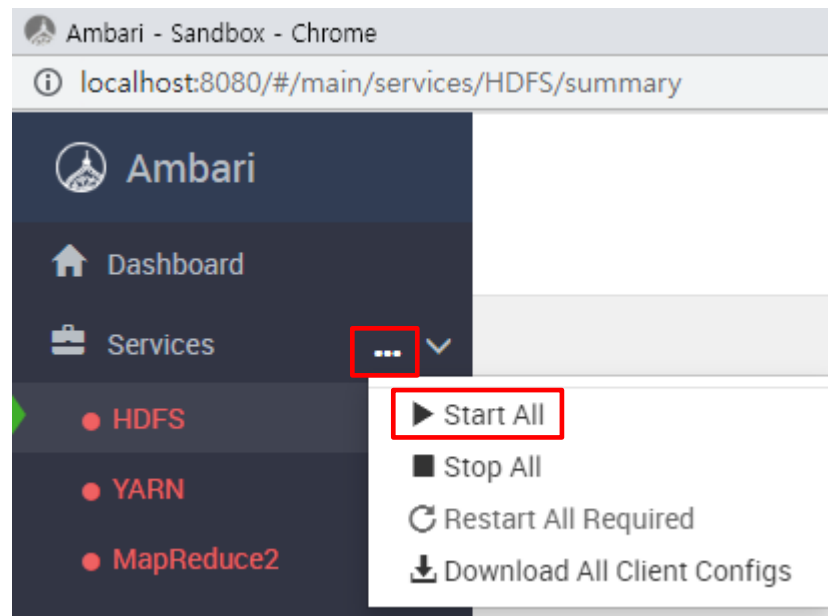
- 로그인
 - ID : maria_dev
 - PW : maria_dev
 - SIGN IN 클릭

PART 3. HORTONWORKS WEB GUI (AMBARI)



- Ambari GUI 웹 페이지 확인
 - HDFS, MapReduce, Hive, Pig 등 다양한 서비스와 현재 상태를 모니터링할 수 있음

PART 3. HORTONWORKS WEB GUI (AMBARI)



- 관련 서비스 시작
 - 부팅 시 모든 서비스가 재시작되지만, 오류나 다시 시작하고 싶을때는 왼쪽 그림과 같이 실행
 - Ambari → Services ... 클릭 → Start All 클릭
- 아래 그림과 같이 대략 10분 ~ 20분이 소요됨.
 - 컴퓨터 사양에 따라 다를 수 있음

1 Background Operation Running

Operations	Status	User	Start Time	Duration
✓ Start All Services	<div><div></div></div> 100%	maria_dev	Today 14:34	13m 38s

PART 3. HORTONWORKS WEB GUI (A)

Ambari - Sandbox - Chrome
localhost:8080/#/main/dashboard/metrics

Ambari / Dashboard / Metrics

Background Operations

1 Background Operation Running

Operations	Status	User	Start Time	Duration
Start All Services	67%	raj_ops	Today 21:50	3m

Background Operations / Start All Services

Hosts

sandbox-hdp.hortonworks.com

Background Operations / Start All Services / sandbox-hdp.hortonworks.com

Tasks

- ✓ Infra Svr Instance Start
- ✓ Range Admin Start
- ✓ Range Usersync Start
- ✓ DataNode Start

NameNode Start

■ 서비스 재시작 오류 해결

- 오랜 시간을 기다려도 서비스가 시작되지 않고, NameNode Start 서비스가 그림과 같이 반복 로그를 출력하는 경우
- 로그 내용 : Retrying after 10 seconds. Reason: Execution of '/usr/hdp/current/hadoop-hdfs-namenode/bin/hdfs dfsadmin -fs hdfs://sandbox-hdp.hortonworks.com:8020 -safemode get | grep 'Safe mode is OFF' returned 1.
- 해당 로그가 계속 출력되는건 현재 저장되어 있는 메타데이터의 정렬을 수행하여 NameNode가 안전모드인 "Safe Mode"로 진입하였기 때문에 발생
- 해결하기 위해서는 메타데이터 이미지(FSimage) 정렬을 기다려야 하지만, 빠른 실습을 위해 안전모드를 강제로 빠져 나온다.

- 아래 Putty 그림과 같이 명령어로 해결

- hdfs dfsadmin -safemode leave

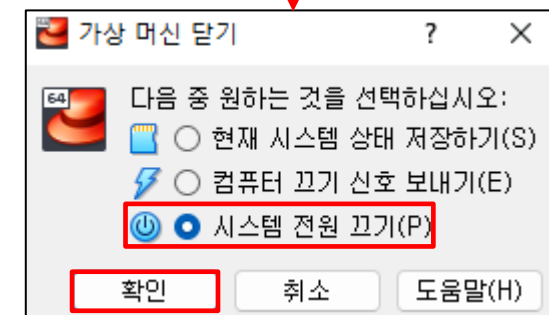
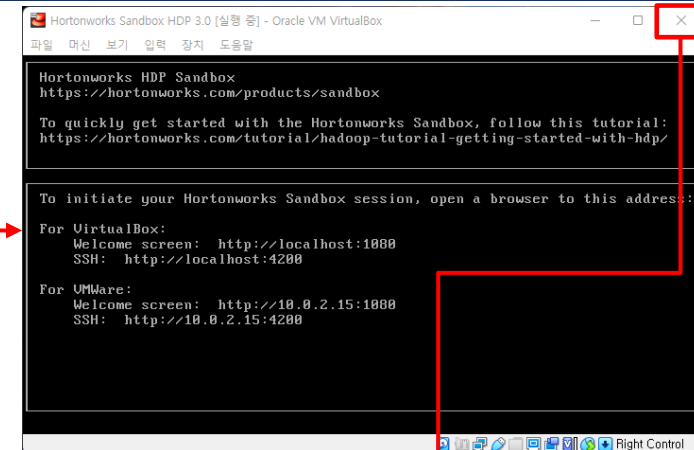
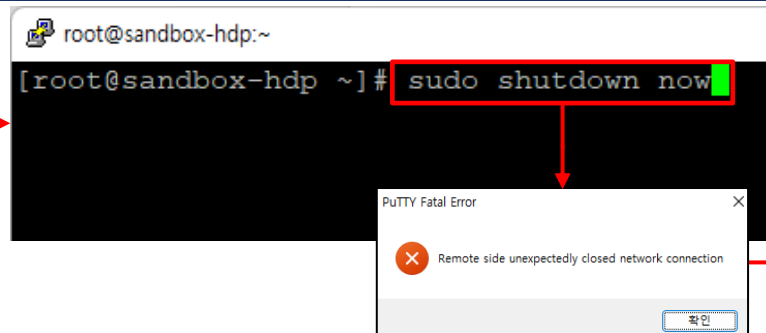
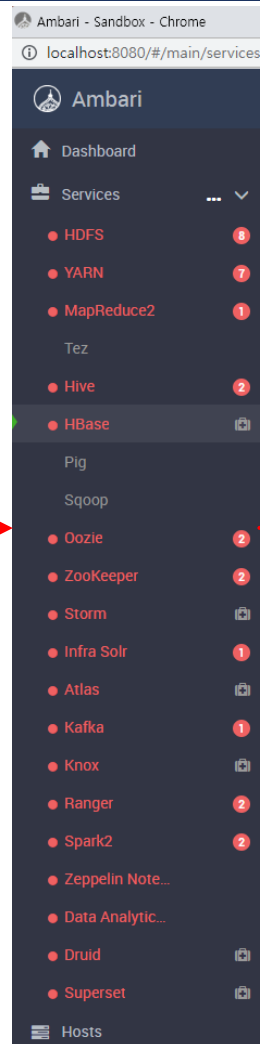
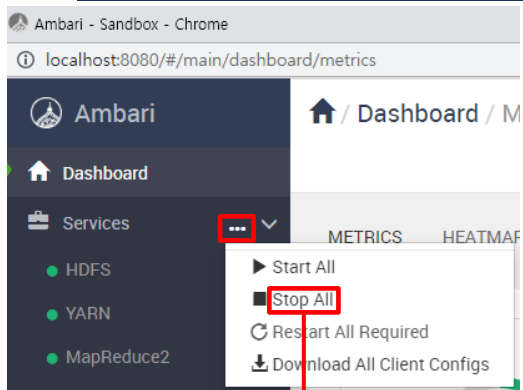
```
[root@sandbox-hdp ~]# hdfs dfsadmin -safemode leave
Safe mode is OFF
```

Background Operations / Start All Services / sandbox-hdp.hortonworks.com / ✓ NameNode Start

Task Log

```
2022-06-16 05:36:00,784 - Execute ["/usr/hdp/current/hadoop-hdfs-namenode/bin/hdfs dfsadmin -fs hdfs://sandbox-hdp.hortonworks.com:8020 -safemode get | grep 'Safe mode is OFF'"] {logoutput: True, 'tries': 115, 'user': 'hdfs', 'try_sleep': 10}
2022-06-16 05:36:03,024 - Retrying after 10 seconds. Reason: Execution of '/usr/hdp/current/hadoop-hdfs-namenode/bin/hdfs dfsadmin -fs hdfs://sandbox-hdp.hortonworks.com:8020 -safemode get | grep 'Safe mode is OFF' returned 1.
2022-06-16 05:36:14,404 - Retrying after 10 seconds. Reason: Execution of '/usr/hdp/current/hadoop-hdfs-namenode/bin/hdfs dfsadmin -fs hdfs://sandbox-hdp.hortonworks.com:8020 -safemode get | grep 'Safe mode is OFF' returned 1.
2022-06-16 05:36:25,895 - Retrying after 10 seconds. Reason: Execution of '/usr/hdp/current/hadoop-hdfs-namenode/bin/hdfs dfsadmin -fs hdfs://sandbox-hdp.hortonworks.com:8020 -safemode get | grep 'Safe mode is OFF' returned 1.
2022-06-16 05:36:37,357 - Retrying after 10 seconds. Reason: Execution of '/usr/hdp/current/hadoop-hdfs-namenode/bin/hdfs dfsadmin -fs hdfs://sandbox-hdp.hortonworks.com:8020 -safemode get | grep 'Safe mode is OFF' returned 1.
2022-06-16 05:36:48,617 - Retrying after 10 seconds. Reason: Execution of '/usr/hdp/current/hadoop-hdfs-namenode/bin/hdfs dfsadmin -fs hdfs://sandbox-hdp.hortonworks.com:8020 -safemode get | grep 'Safe mode is OFF' returned 1.
2022-06-16 05:37:00,234 - Retrying after 10 seconds. Reason: Execution of '/usr/hdp/current/hadoop-hdfs-namenode/bin/hdfs dfsadmin -fs hdfs://sandbox-hdp.hortonworks.com:8020 -safemode get | grep 'Safe mode is OFF' returned 1.
2022-06-16 05:37:11,863 - Retrying after 10 seconds. Reason: Execution of '/usr/hdp/current/hadoop-hdfs-namenode/bin/hdfs dfsadmin -fs hdfs://sandbox-hdp.hortonworks.com:8020 -safemode get | grep 'Safe mode is OFF' returned 1.
2022-06-16 05:37:23,363 - Retrying after 10 seconds. Reason: Execution of '/usr/hdp/current/hadoop-hdfs-namenode/bin/hdfs dfsadmin -fs hdfs://sandbox-hdp.hortonworks.com:8020 -safemode get | grep 'Safe mode is OFF' returned 1.
2022-06-16 05:37:34,994 - Retrying after 10 seconds. Reason: Execution of '/usr/hdp/current/hadoop-hdfs-namenode/bin/hdfs dfsadmin -fs hdfs://sandbox-hdp.hortonworks.com:8020 -safemode get | grep 'Safe mode is OFF' returned 1.
2022-06-16 05:37:46,483 - Retrying after 10 seconds. Reason: Execution of '/usr/hdp/current/hadoop-hdfs-namenode/bin/hdfs dfsadmin -fs hdfs://sandbox-hdp.hortonworks.com:8020 -safemode get | grep 'Safe mode is OFF' returned 1.
2022-06-16 05:37:57,935 - Retrying after 10 seconds. Reason: Execution of '/usr/hdp/current/hadoop-hdfs-namenode/bin/hdfs dfsadmin -fs hdfs://sandbox-hdp.hortonworks.com:8020 -safemode get | grep 'Safe mode is OFF' returned 1.
2022-06-16 05:38:09,497 - Retrying after 10 seconds. Reason: Execution of '/usr/hdp/current/hadoop-hdfs-namenode/bin/hdfs dfsadmin -fs hdfs://sandbox-hdp.hortonworks.com:8020 -safemode get | grep 'Safe mode is OFF' returned 1.
2022-06-16 05:38:20,959 - Retrying after 10 seconds. Reason: Execution of '/usr/hdp/current/hadoop-hdfs-namenode/bin/hdfs dfsadmin -fs hdfs://sandbox-hdp.hortonworks.com:8020 -safemode get | grep 'Safe mode is OFF' returned 1.
2022-06-16 05:38:32,376 - Retrying after 10 seconds. Reason: Execution of '/usr/hdp/current/hadoop-hdfs-namenode/bin/hdfs dfsadmin -fs hdfs://sandbox-hdp.hortonworks.com:8020 -safemode get | grep 'Safe mode is OFF' returned 1.
2022-06-16 05:38:43,823 - Retrying after 10 seconds. Reason: Execution of '/usr/hdp/current/hadoop-hdfs-namenode/bin/hdfs dfsadmin -fs hdfs://sandbox-hdp.hortonworks.com:8020 -safemode get | grep 'Safe mode is OFF' returned 1.
2022-06-16 05:38:55,341 - Retrying after 10 seconds. Reason: Execution of '/usr/hdp/current/hadoop-hdfs-namenode/bin/hdfs dfsadmin -fs hdfs://sandbox-hdp.hortonworks.com:8020 -safemode get | grep 'Safe mode is OFF' returned 1.
2022-06-16 05:39:06,785 - Retrying after 10 seconds. Reason: Execution of '/usr/hdp/current/hadoop-hdfs-namenode/bin/hdfs dfsadmin -fs hdfs://sandbox-hdp.hortonworks.com:8020 -safemode get | grep 'Safe mode is OFF' returned 1.
2022-06-16 05:39:18,296 - Retrying after 10 seconds. Reason: Execution of '/usr/hdp/current/hadoop-hdfs-namenode/bin/hdfs dfsadmin -fs hdfs://sandbox-hdp.hortonworks.com:8020 -safemode get | grep 'Safe mode is OFF' returned 1.
2022-06-16 05:39:29,792 - Retrying after 10 seconds. Reason: Execution of '/usr/hdp/current/hadoop-hdfs-namenode/bin/hdfs dfsadmin -fs hdfs://sandbox-hdp.hortonworks.com:8020 -safemode get | grep 'Safe mode is OFF' returned 1.
2022-06-16 05:39:41,192 - Retrying after 10 seconds. Reason: Execution of '/usr/hdp/current/hadoop-hdfs-namenode/bin/hdfs dfsadmin -fs hdfs://sandbox-hdp.hortonworks.com:8020 -safemode get | grep 'Safe mode is OFF' returned 1.
```

PART 3. HORTONWORKS WEB GUI (AMBARI)



■ 안전 종료 방법

- Ambari → Services ... 클릭 → Stop All 클릭
- 모든 서비스가 Stop된 것을 확인
- Putty에서 "sudo shutdown now" 명령어로 종료
 - Hortonworks의 이미지가 내부적으로 종료 되었기에 연결이 끊김
- VirtualBox 프로그램 오른쪽 상단의 X 표시 클릭
- "시스템 전원 끄기" 선택 후 "확인"



PART 4. 데이터 복사



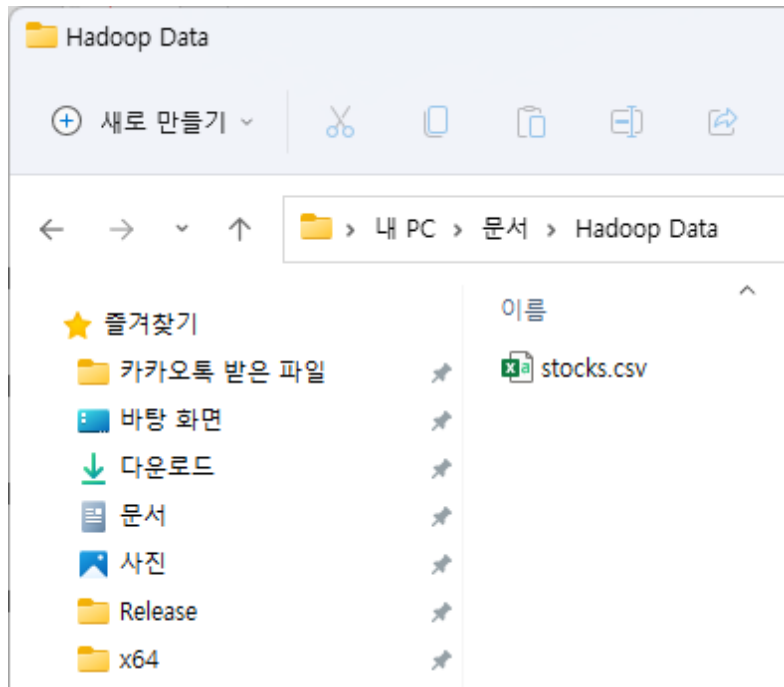
PART 4. 데이터 복사

```
root@sandbox-hdp:/
Last login: Tue Jun 14 08:58:56 2022 from 172.18.0.2
[root@sandbox-hdp ~]#
[root@sandbox-hdp ~]# ls
anaconda-ks.cfg
[root@sandbox-hdp ~]#
[root@sandbox-hdp ~]# cd /
[root@sandbox-hdp /]#
[root@sandbox-hdp /]# ls
apps      dev      lib64    proc     srv
bin       etc      media    root     sys
boot     hadoop  mnt      run      tmp
cgroups_test  home    mysql-connector-java-5.1.45  sandbox  usr
core.24120  kafka-logs  opt    sandbox-flavour  var
core.24572  lib        packer-files  sbin
[root@sandbox-hdp /]#
[root@sandbox-hdp /]# mkdir Data
[root@sandbox-hdp /]# ls
apps      Data    lib      packer-files  sbin
bin       dev     lib64    proc          srv
boot     etc     media    root          sys
cgroups_test  hadoop  mnt      run          tmp
core.24120  home    mysql-connector-java-5.1.45  sandbox  usr
core.24572  kafka-logs  opt    sandbox-flavour  var
[root@sandbox-hdp /]#
```

■ Putty 접속 후

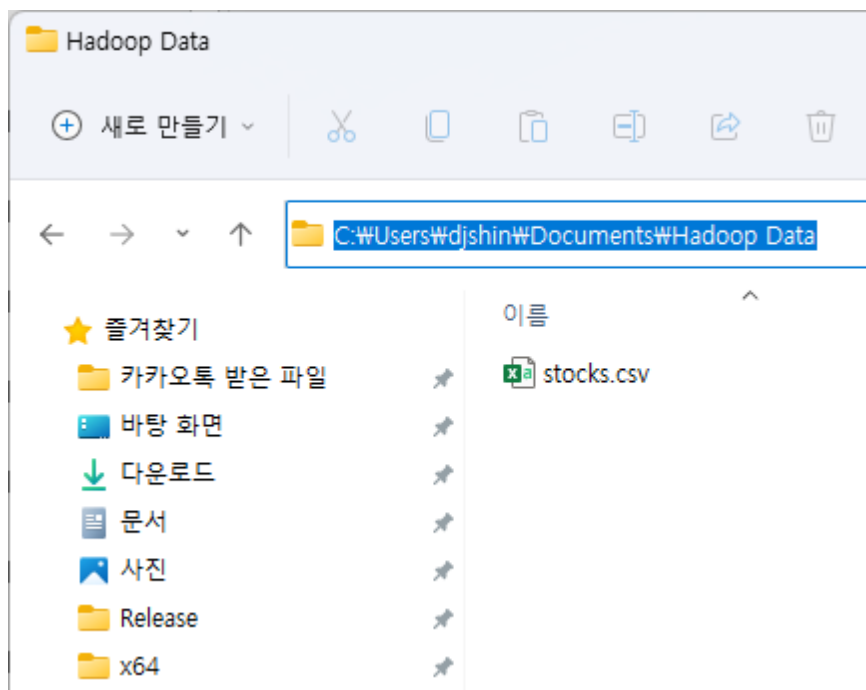
- 루트 디렉토리 이동
 - cd /
- 디렉토리 확인
 - ls
- Data 디렉토리 생성
 - mkdir Data
- 디렉토리 생성 확인
 - ls

PART 4. 데이터 복사



- 윈도우에서 실행
 - 내 PC → 문서 디렉토리에서 "Hadoop Data" 디렉토리 생성
 - 생성 후 필요한 데이터 복사

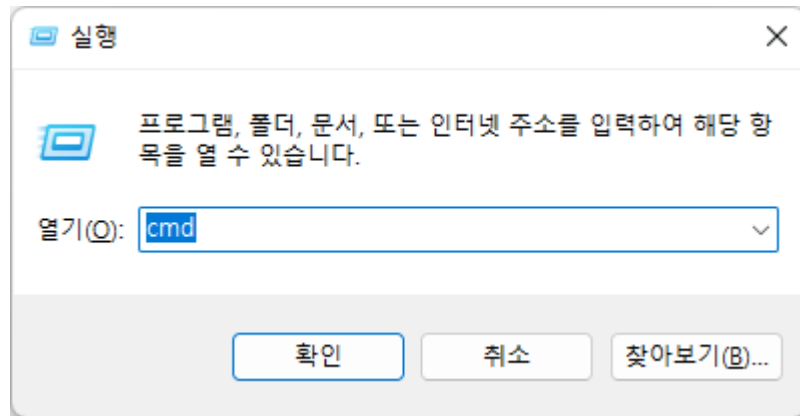
PART 4. 데이터 복사



- 윈도우에서 실행

- 해당 디렉토리에서 파일 탐색기의 현재 경로를 클릭하여 블록처리된 문자를 복사

PART 4. 데이터 복사



- 윈도우에서 실행
 - Windows 키 + R 실행창 오픈
 - cmd 입력

PART 4. 데이터 복사

```
C:\Windows\system32\cmd.exe
Microsoft Windows [Version 10.0.22000.675]
(c) Microsoft Corporation. All rights reserved.

C:\Users\djshin>cd C:\Users\djshin\Documents\Hadoop Data

C:\Users\djshin\Documents\Hadoop Data>dir
C 드라이브의 볼륨에는 이름이 없습니다.
볼륨 일련 번호: 5437-1090

C:\Users\djshin\Documents\Hadoop Data 디렉터리

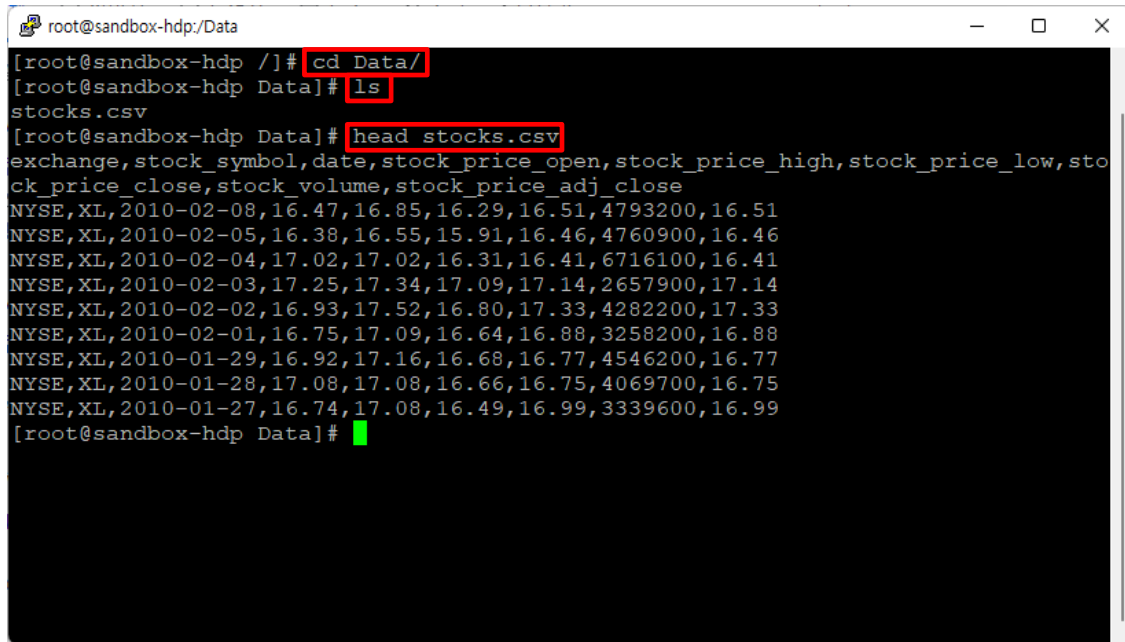
2022-06-14 오후 05:53    <DIR>
2022-06-14 오후 05:53    <DIR>
2022-06-14 오후 05:53         3,613,198  stocks.csv
                   1개 파일              3,613,198 바이트
                   2개 디렉터리 311,841,579,008 바이트 남음

C:\Users\djshin\Documents\Hadoop Data>scp -P 2222 stocks.csv root@localhost:/Data
The authenticity of host '[localhost]:2222 ([127.0.0.1]:2222)' can't be established.
ECDSA key fingerprint is SHA256:TV0krZaLbRNaLY8D/r0Fz/9o6Y212w8yZJqKmo0acV.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added '[localhost]:2222' (ECDSA) to the list of known hosts.
root@localhost's password: bigdata12
stocks.csv                                100% 3529KB  98.2MB/s   00:00

C:\Users\djshin\Documents\Hadoop Data>
```

- 윈도우에서 실행
 - cd "복사했던 명령어" 실행
 - cd C:\Users\djshin\Documents\Hadoop Data
 - 보라색으로 표시된 부분은 PC마다 다름
 - stocks.csv 파일이 존재하는지 확인
 - Hortonworks로 데이터 전송
 - scp -P 2222 stocks.csv root@localhost:/Data
 - ssh 접속 여부 확인시 : yes 입력
 - root 비밀번호 확인시 : bigdata12 입력

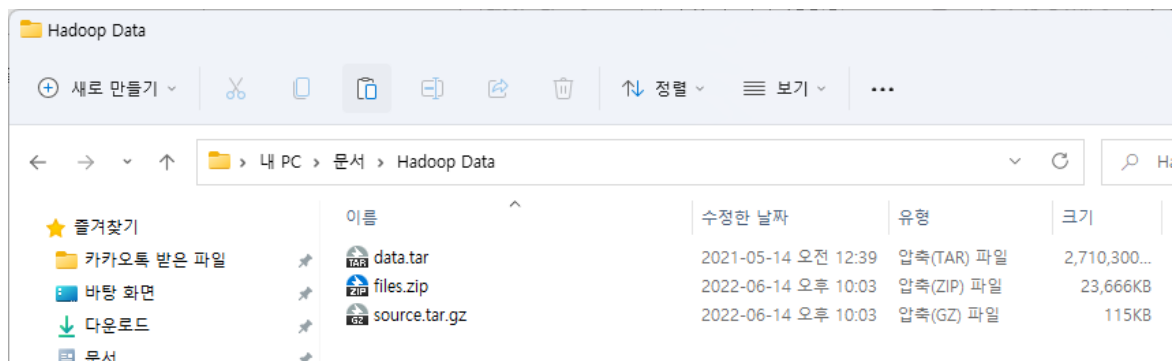
PART 4. 데이터 복사



```
root@sandbox-hdp:/Data
[root@sandbox-hdp /]# cd Data/
[root@sandbox-hdp Data]# ls
stocks.csv
[root@sandbox-hdp Data]# head stocks.csv
exchange,stock_symbol,date,stock_price_open,stock_price_high,stock_price_low,stock_price_close,stock_volume,stock_price_adj_close
NYSE,XL,2010-02-08,16.47,16.85,16.29,16.51,4793200,16.51
NYSE,XL,2010-02-05,16.38,16.55,15.91,16.46,4760900,16.46
NYSE,XL,2010-02-04,17.02,17.02,16.31,16.41,6716100,16.41
NYSE,XL,2010-02-03,17.25,17.34,17.09,17.14,2657900,17.14
NYSE,XL,2010-02-02,16.93,17.52,16.80,17.33,4282200,17.33
NYSE,XL,2010-02-01,16.75,17.09,16.64,16.88,3258200,16.88
NYSE,XL,2010-01-29,16.92,17.16,16.68,16.77,4546200,16.77
NYSE,XL,2010-01-28,17.08,17.08,16.66,16.75,4069700,16.75
NYSE,XL,2010-01-27,16.74,17.08,16.49,16.99,3339600,16.99
[root@sandbox-hdp Data]#
```

- Putty에서 실행
 - 루트 디렉토리에서 Data 디렉토리로 이동
 - `cd /Data`
 - 파일이 복사 되었는지 확인
 - `ls`
 - head 명령어로 10줄 확인
 - `head stocks.csv`
- 정상적으로 데이터가 전송된것을 확인할 수 있음
- ※ 필요한 데이터는 이와 같은 방법으로 복사하여 예제 실습할 것

PART 4. 데이터 복사



- 윈도우에서 실행
 - 실습 예제에 사용하는 파일 복사
 - data.tar
 - files.zip
 - source.tar.gz

PART 4. 데이터 복사

```
C:\Windows\system32\cmd.exe
C:\Users\djshin>cd C:\Users\djshin\Documents\Hadoop Data
C:\Users\djshin\Documents\Hadoop Data>dir
C 드라이브의 볼륨에는 이름이 없습니다.
볼륨 일련 번호: 5437-1090

C:\Users\djshin\Documents\Hadoop Data 디렉터리
2022-06-15 오후 12:52 <DIR> .
2022-06-14 오후 09:39 <DIR> ..
2021-05-14 오전 12:39 2,775,347,200 data.tar
2022-06-14 오후 10:08 24,233,819 files.zip
2022-06-14 오후 10:08 117,306 source.tar.gz
3개 파일 2,799,698,325 바이트
2개 디렉터리 307,126,857,728 바이트 남음

C:\Users\djshin\Documents\Hadoop Data>scp -P 2222 ./\* root@localhost:/Data
root@localhost's password:
data.tar 100% 2647MB 70.5MB/s 00:37
files.zip 100% 23MB 133.8MB/s 00:00
source.tar.gz 100% 115KB 38.2MB/s 00:00

C:\Users\djshin\Documents\Hadoop Data>
```

- 윈도우에서 실행
 - 데이터가 위치한 디렉토리로 이동
 - `cd C:\Users\djshin\Documents\Hadoop Data`
 - 파일 확인
 - `dir`
 - 전체 파일 전송
 - `scp -P 2222 ./* root@localhost:/Data`
 - `./*` → 현재 디렉토리 경로(`./`)의 모든 파일(*)을 의미함
 - `root@localhost's password:` 출력시 `bigdata12` 입력

PART 4. 데이터 복사

```
root@sandbox-hdp:/Data
[root@sandbox-hdp Data]# ls
data.tar  files.zip  source.tar.gz  stocks.csv

[root@sandbox-hdp Data]# tar xvf data.tar

[root@sandbox-hdp Data]# unzip files.zip -d files

[root@sandbox-hdp Data]# tar xvf source.tar.gz

[root@sandbox-hdp Data]# ls
data  data.tar  files  files.zip  source  source.tar.gz  stocks.csv
[root@sandbox-hdp Data]#
```

- Putty에서 실행
 - 데이터 디렉토리 이동
 - `cd /Data`
 - 데이터가 전송되었는지 확인
 - `ls`
 - 압축 해제
 - tar 파일 → `tar xvf data.tar`
 - zip 파일 → `unzip files.zip -d files`
 - unzip은 디렉토리를 생성하지 않고, 압축을 해제하기 때문에 `-d` 옵션을 부여하여 files 디렉토리를 생성하고 압축 해제
 - tar.gz 파일 → `tar xvf source.tar.gz`