How to use Hadoop & Spark?

Contents

1. **Using Windows**
   - Anaconda installation
   - Spark installation
   - Hadoop installation

2. **Using Linux**
   - Virtual box installation
   - Spark installation
   - Hadoop installation

3. **Spark implementation**

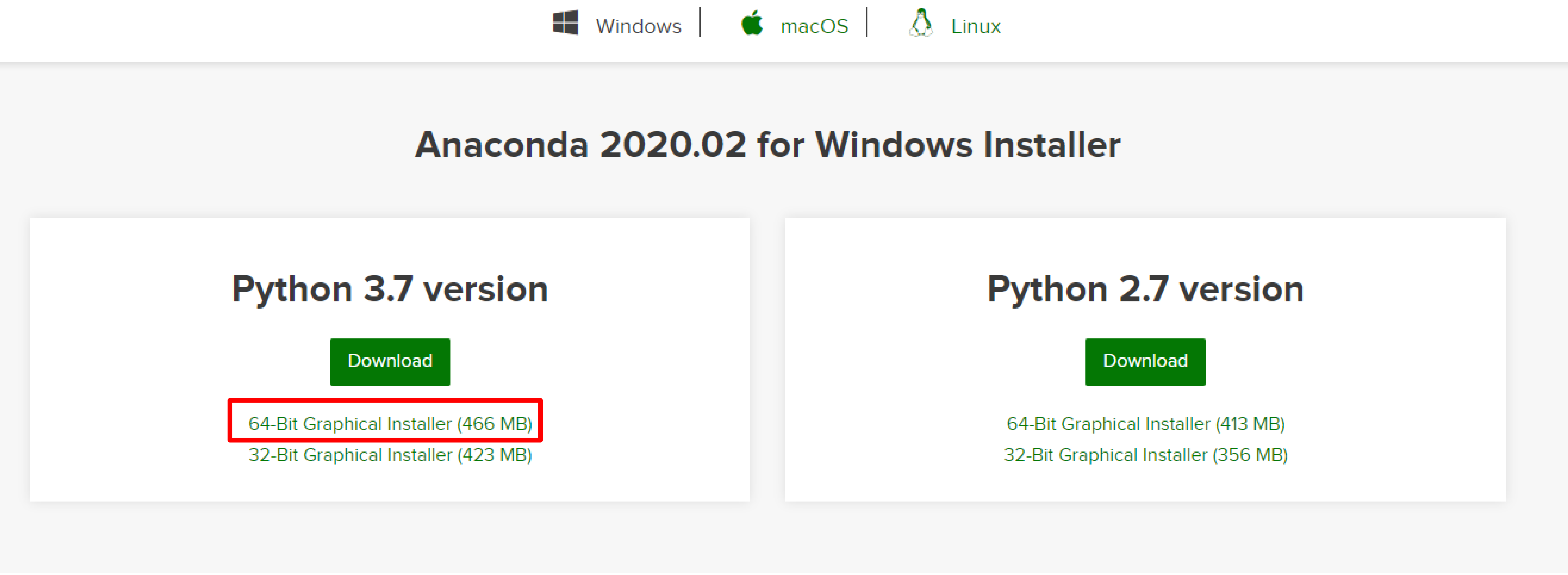
4. **Issues**

Using Windows

Anaconda Installation

Install Anaconda

- Anaconda is a tool for Python programming and command prompt



Windows | macOS | Linux

Anaconda 2020.02 for Windows Installer

Python 3.7 version	Python 2.7 version
Download	Download
64-Bit Graphical Installer (466 MB)	64-Bit Graphical Installer (413 MB)
32-Bit Graphical Installer (423 MB)	32-Bit Graphical Installer (356 MB)

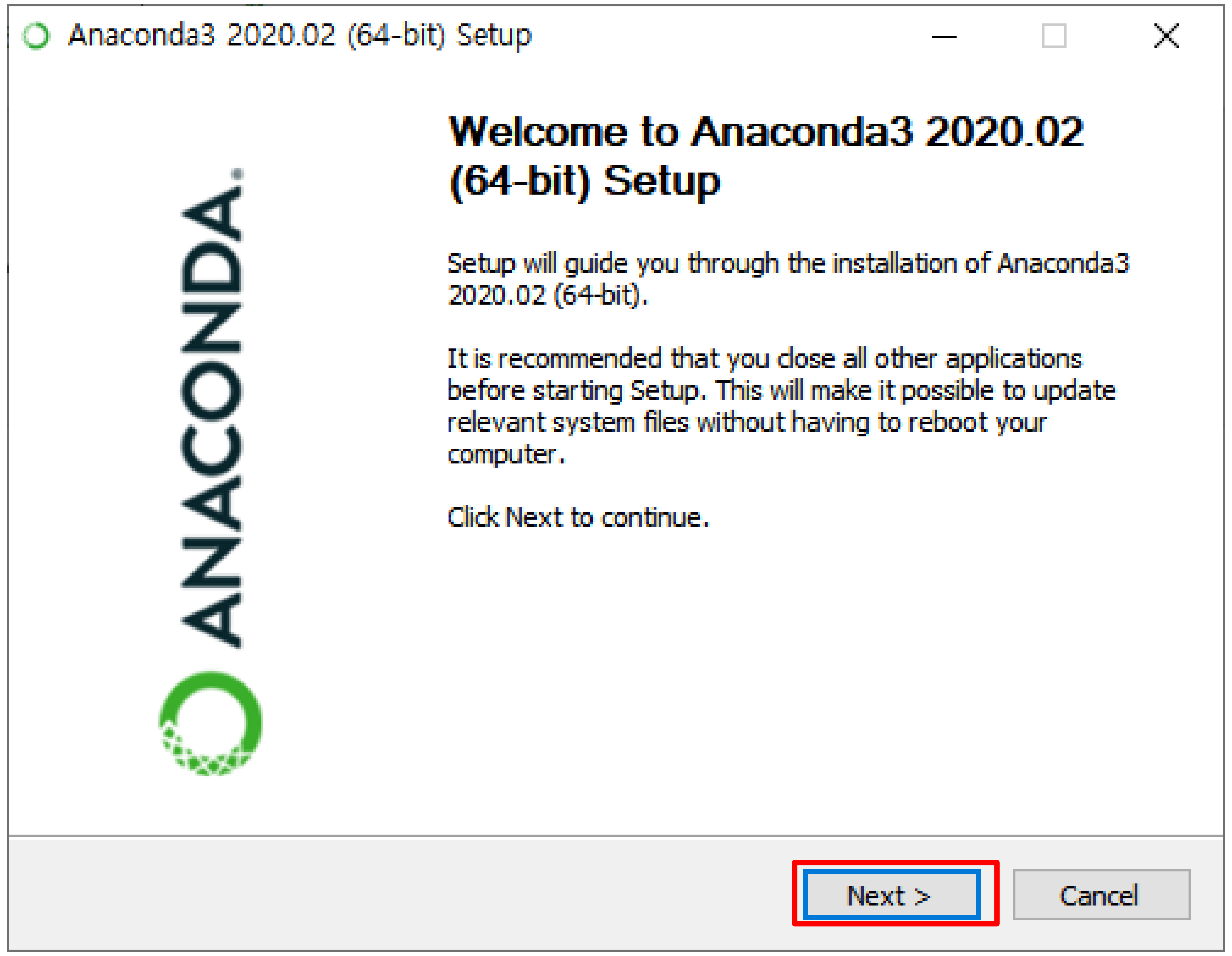
Get Started with Anaconda Individual Edition

- If your computer's OS is Windows x86-64, We recommend download Python 3.7 64-bit Graphical Installer.
- You can download another version if your computer has another OS.

https://www.anaconda.com/distribution/

Install Anaconda

➢ **Begin installment and accept the license**



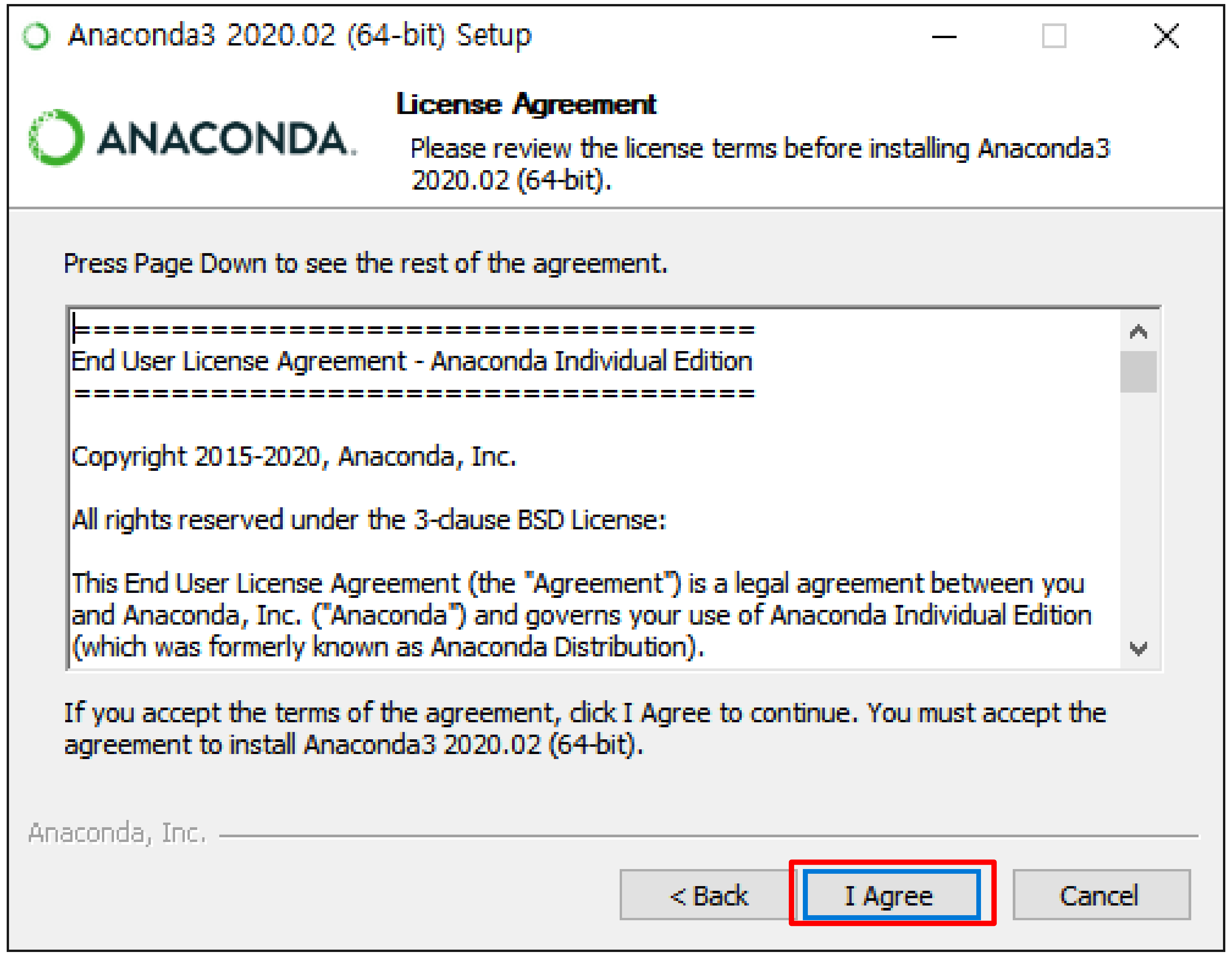
Anaconda3 2020.02 (64-bit) Setup

Welcome to Anaconda3 2020.02 (64-bit) Setup

Setup will guide you through the installation of Anaconda3 2020.02 (64-bit).

It is recommended that you close all other applications before starting Setup. This will make it possible to update relevant system files without having to reboot your computer.

Click Next to continue.



Anaconda3 2020.02 (64-bit) Setup

License Agreement

Please review the license terms before installing Anaconda3 2020.02 (64-bit).

Press Page Down to see the rest of the agreement.

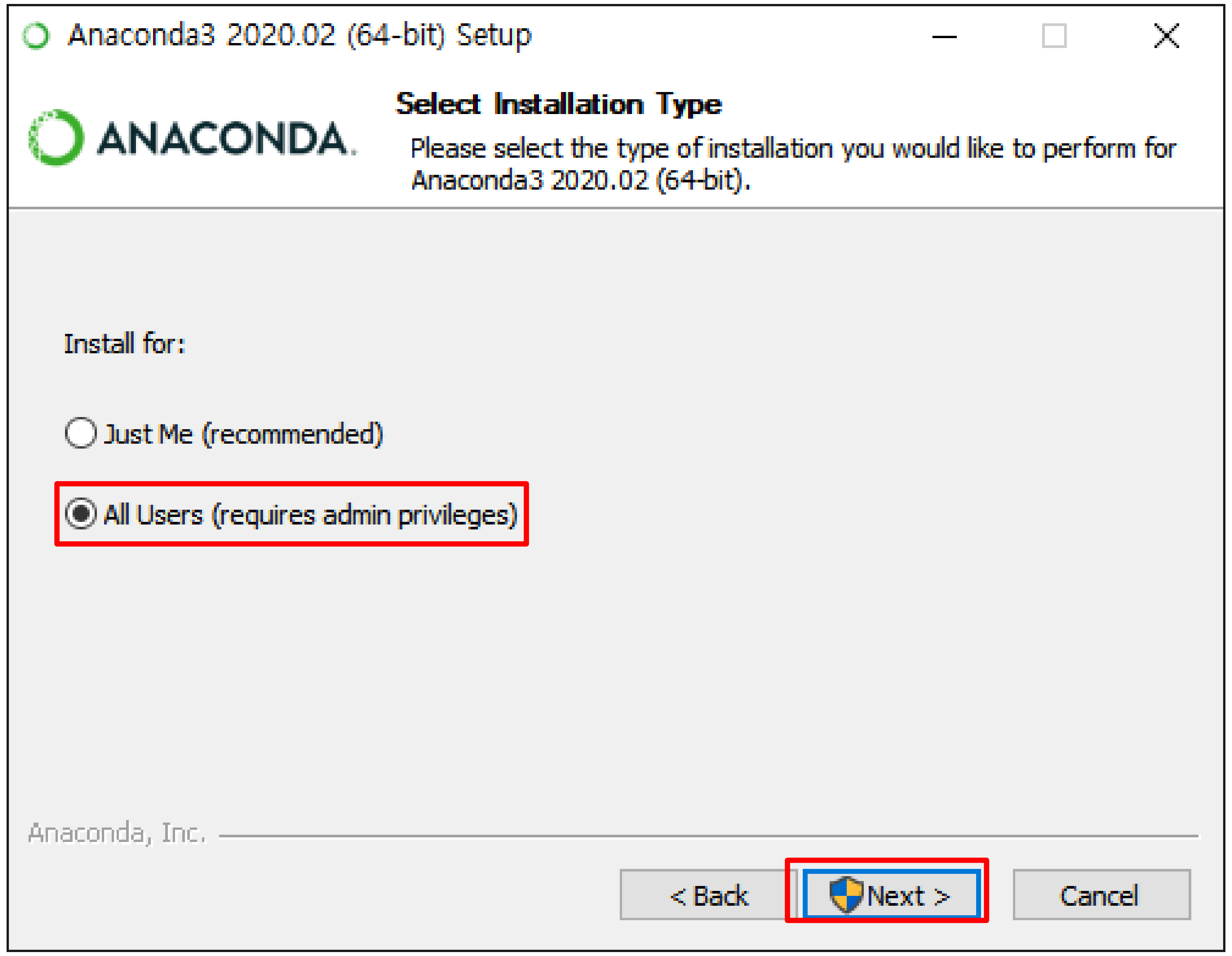
```
===================================
End User License Agreement - Anaconda Individual Edition
===================================

Copyright 2015-2020, Anaconda, Inc.

All rights reserved under the 3-clause BSD License:

This End User License Agreement (the "Agreement") is a legal agreement between you
and Anaconda, Inc. ("Anaconda") and governs your use of Anaconda Individual Edition
(which was formerly known as Anaconda Distribution).
```

If you accept the terms of the agreement, click I Agree to continue. You must accept the agreement to install Anaconda3 2020.02 (64-bit).

Anaconda, Inc.

Install Anaconda

- ➢ **Check "All Users" to avoid system authority issue**



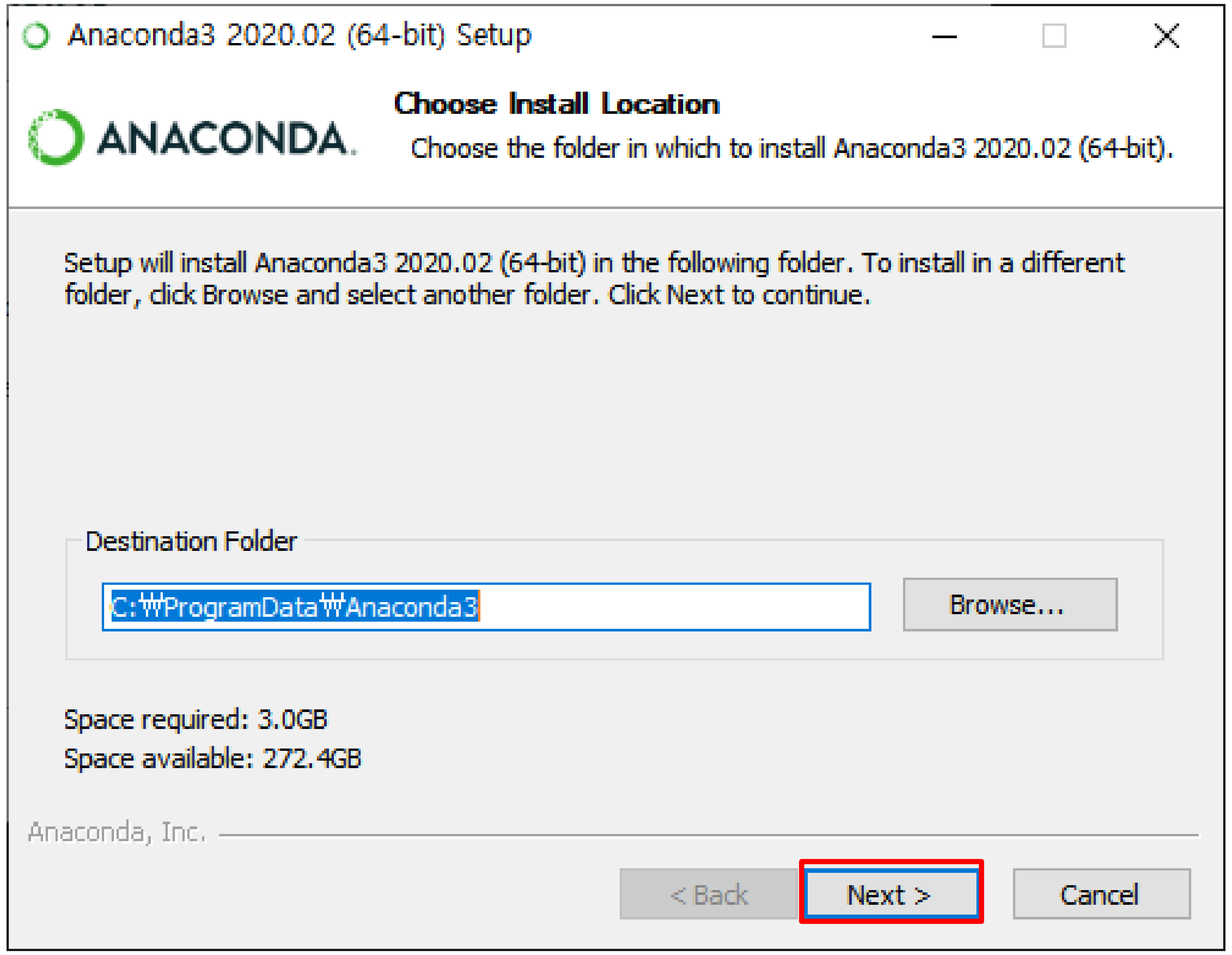
Select Installation Type

Please select the type of installation you would like to perform for Anaconda3 2020.02 (64-bit).

Install for:

○ Just Me (recommended)

● All Users (requires admin privileges)



Choose Install Location

Choose the folder in which to install Anaconda3 2020.02 (64-bit).

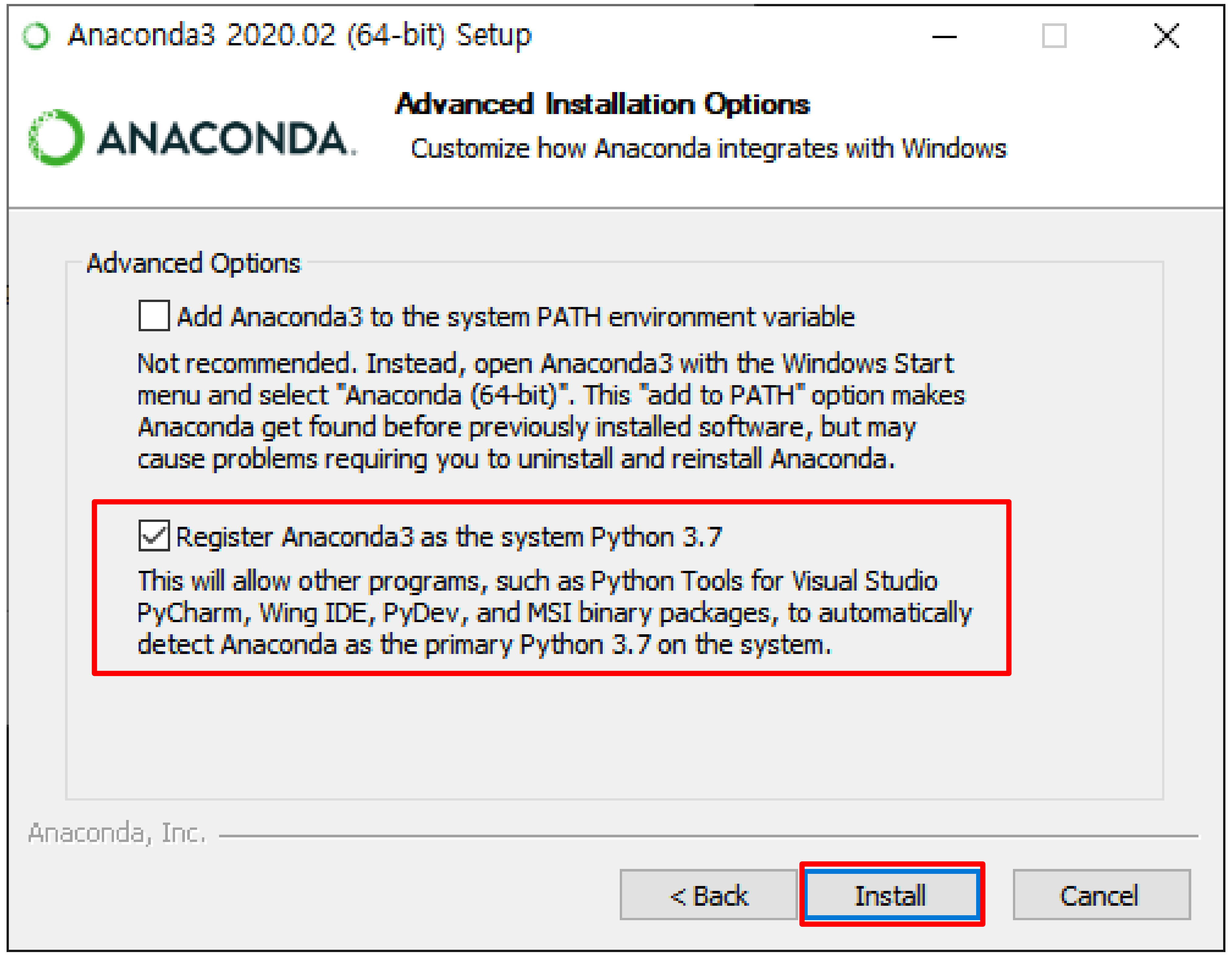
Setup will install Anaconda3 2020.02 (64-bit) in the following folder. To install in a different folder, click Browse and select another folder. Click Next to continue.

Destination Folder: C:₩ProgramData₩Anaconda3

Space required: 3.0GB
Space available: 272.4GB

Install Anaconda

- Check “Register Anaconda3 as the system Python 3.7”
- Be careful not to check “Add Anaconda3 to the system PATH environment variable.”
- Finish installment



Anaconda3 2020.02 (64-bit) Setup

Advanced Installation Options

Customize how Anaconda integrates with Windows

Advanced Options

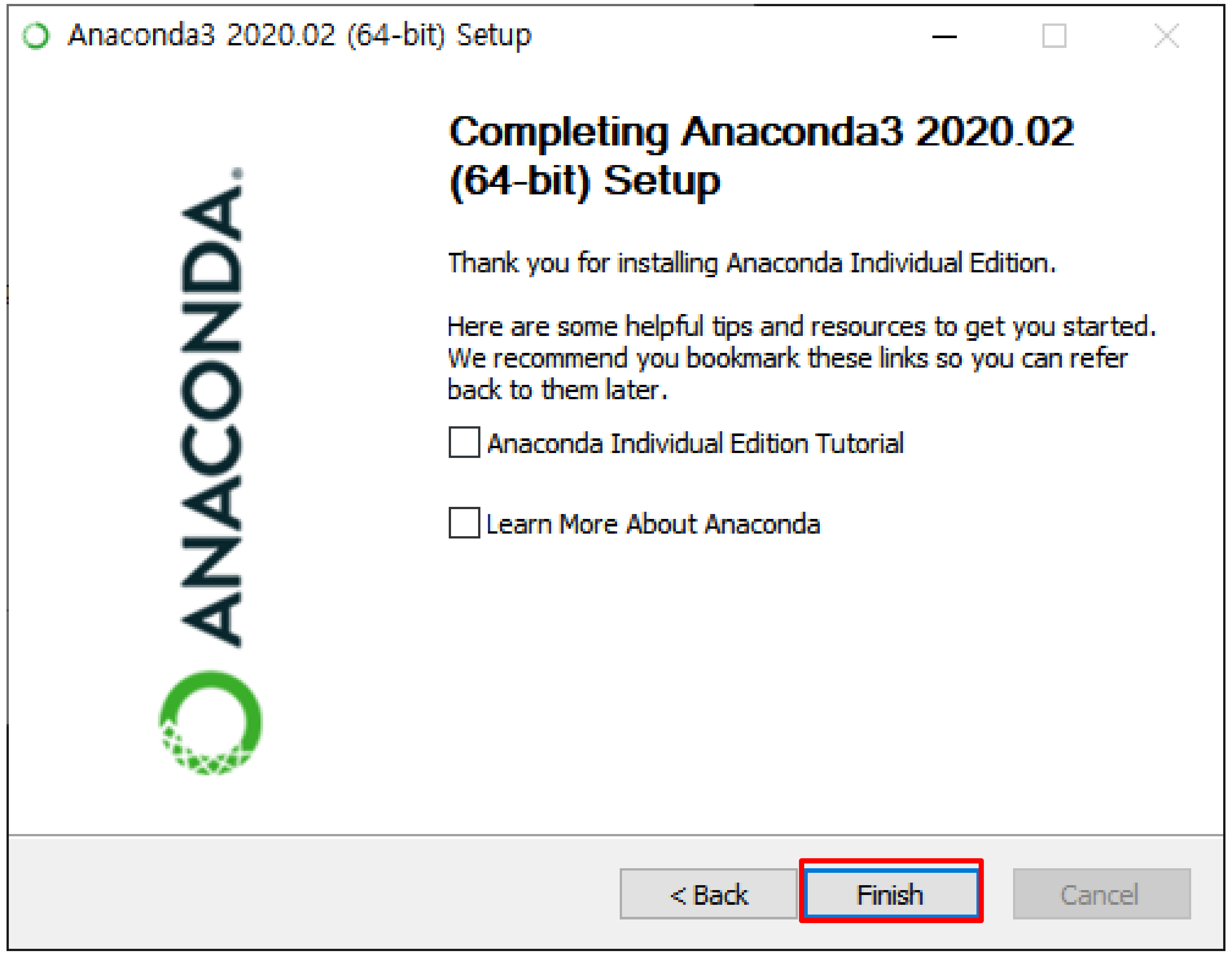
☐ Add Anaconda3 to the system PATH environment variable

Not recommended. Instead, open Anaconda3 with the Windows Start menu and select "Anaconda (64-bit)". This "add to PATH" option makes Anaconda get found before previously installed software, but may cause problems requiring you to uninstall and reinstall Anaconda.

☑ Register Anaconda3 as the system Python 3.7

This will allow other programs, such as Python Tools for Visual Studio PyCharm, Wing IDE, PyDev, and MSI binary packages, to automatically detect Anaconda as the primary Python 3.7 on the system.

Anaconda, Inc.



Anaconda3 2020.02 (64-bit) Setup

Completing Anaconda3 2020.02 (64-bit) Setup

Thank you for installing Anaconda Individual Edition.

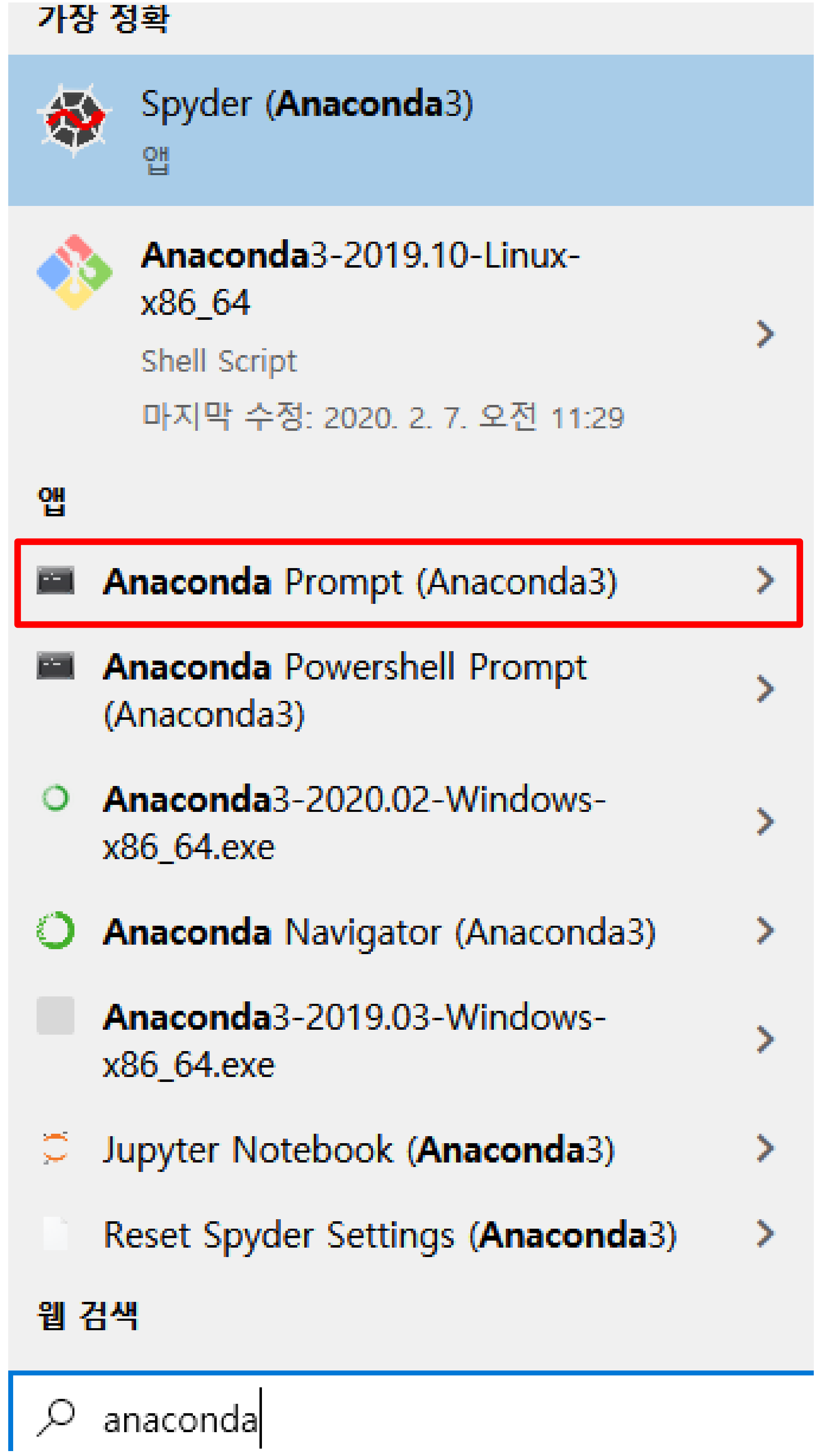
Here are some helpful tips and resources to get you started. We recommend you bookmark these links so you can refer back to them later.

☐ Anaconda Individual Edition Tutorial

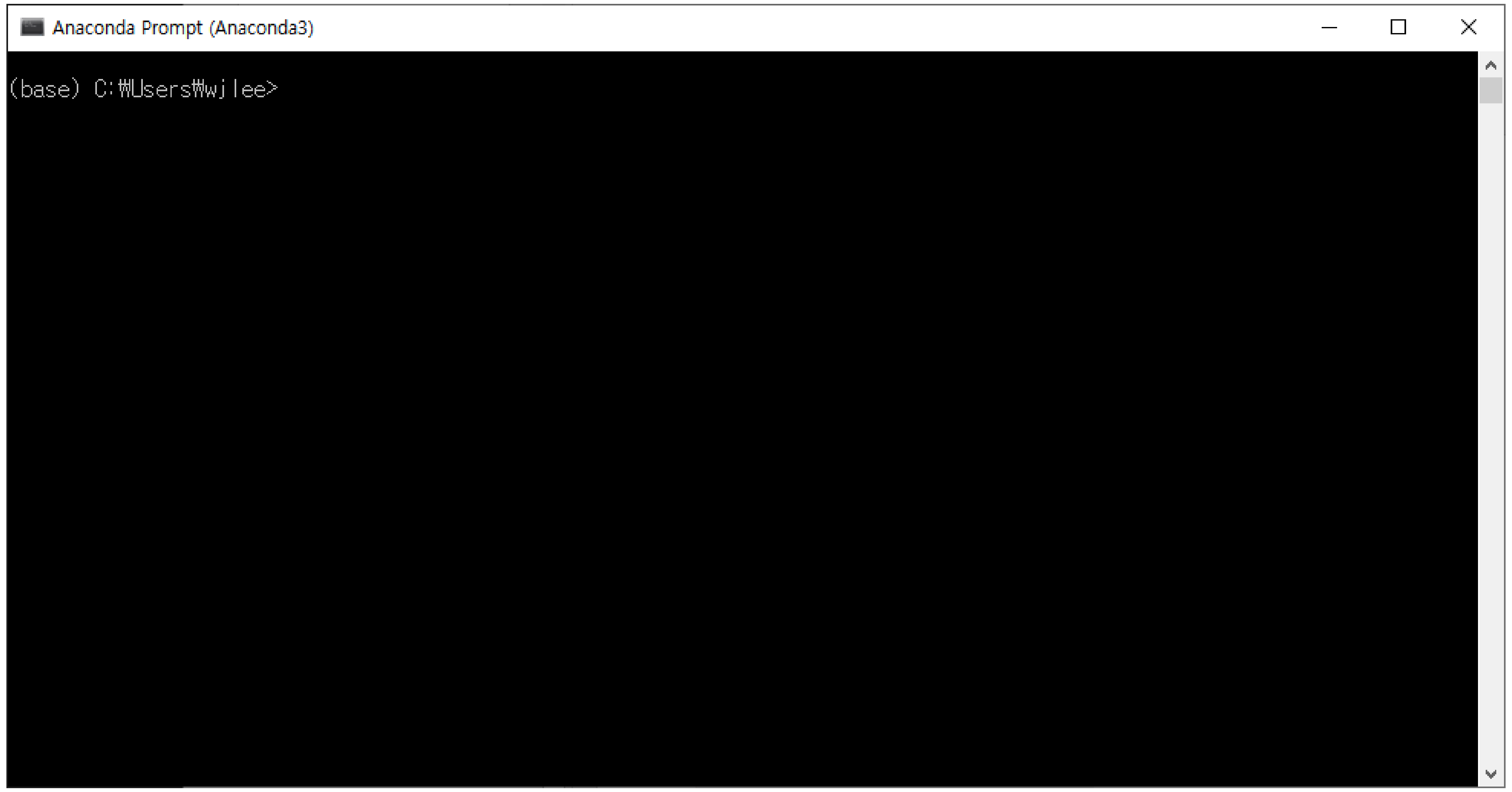
☐ Learn More About Anaconda

Anaconda

- ➢ **Search Anaconda Prompt and open it**
- ➢ **We will use Python and Spark using this tool**



가장 정확
Spyder (Anaconda3)
앱
Anaconda3-2019.10-Linux-x86_64
Shell Script
마지막 수정: 2020. 2. 7. 오전 11:29
앱
Anaconda Prompt (Anaconda3)
Anaconda Powershell Prompt (Anaconda3)
Anaconda3-2020.02-Windows-x86_64.exe
Anaconda Navigator (Anaconda3)
Anaconda3-2019.03-Windows-x86_64.exe
Jupyter Notebook (Anaconda3)
Reset Spyder Settings (Anaconda3)
웹 검색
anaconda

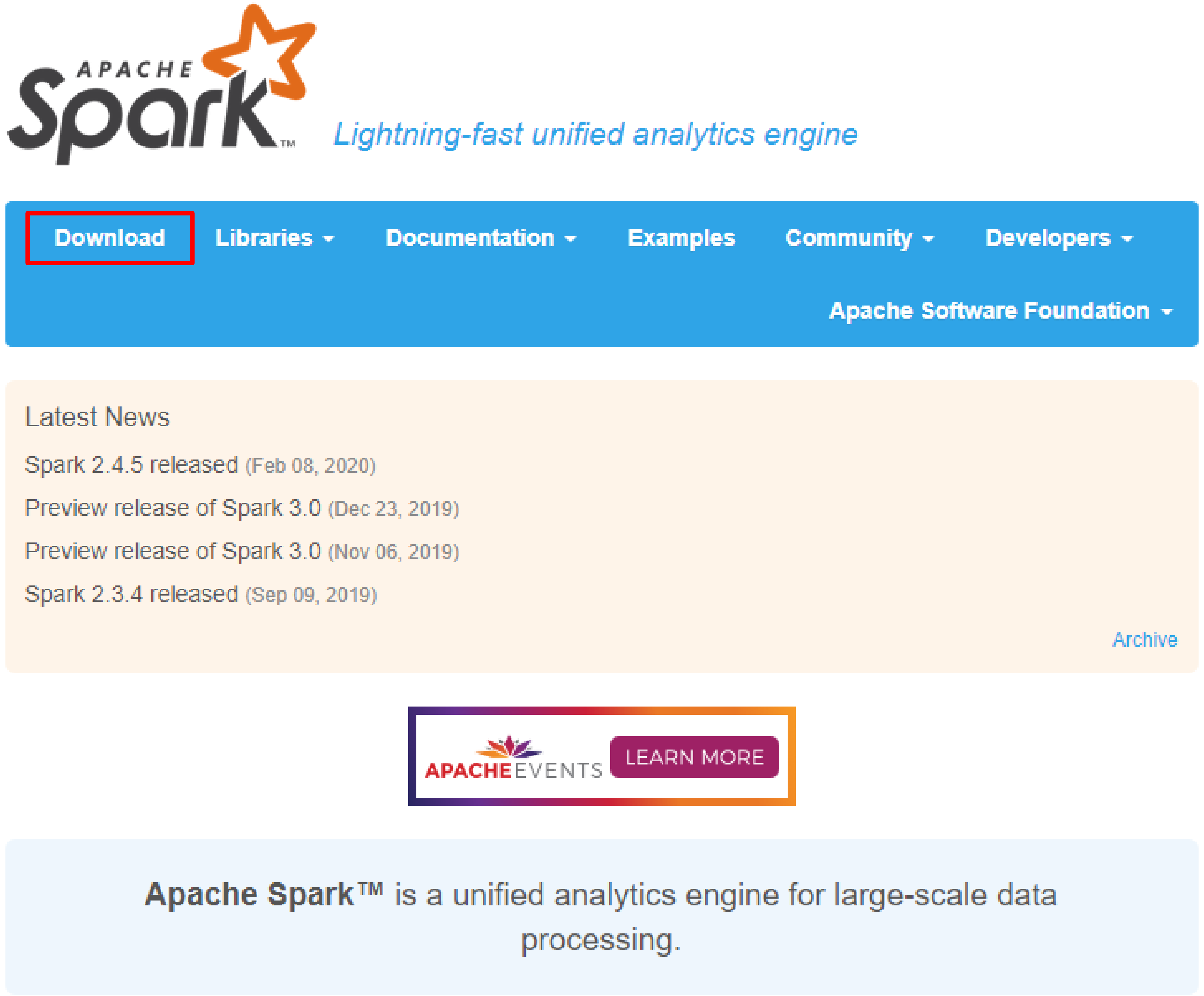


```
(base) C:₩Users₩wjlee>
```

Spark Installation

Download Spark

- ➢ **Download Spark package from web site**



Latest News

Spark 2.4.5 released (Feb 08, 2020)

Preview release of Spark 3.0 (Dec 23, 2019)

Preview release of Spark 3.0 (Nov 06, 2019)

Spark 2.3.4 released (Sep 09, 2019)

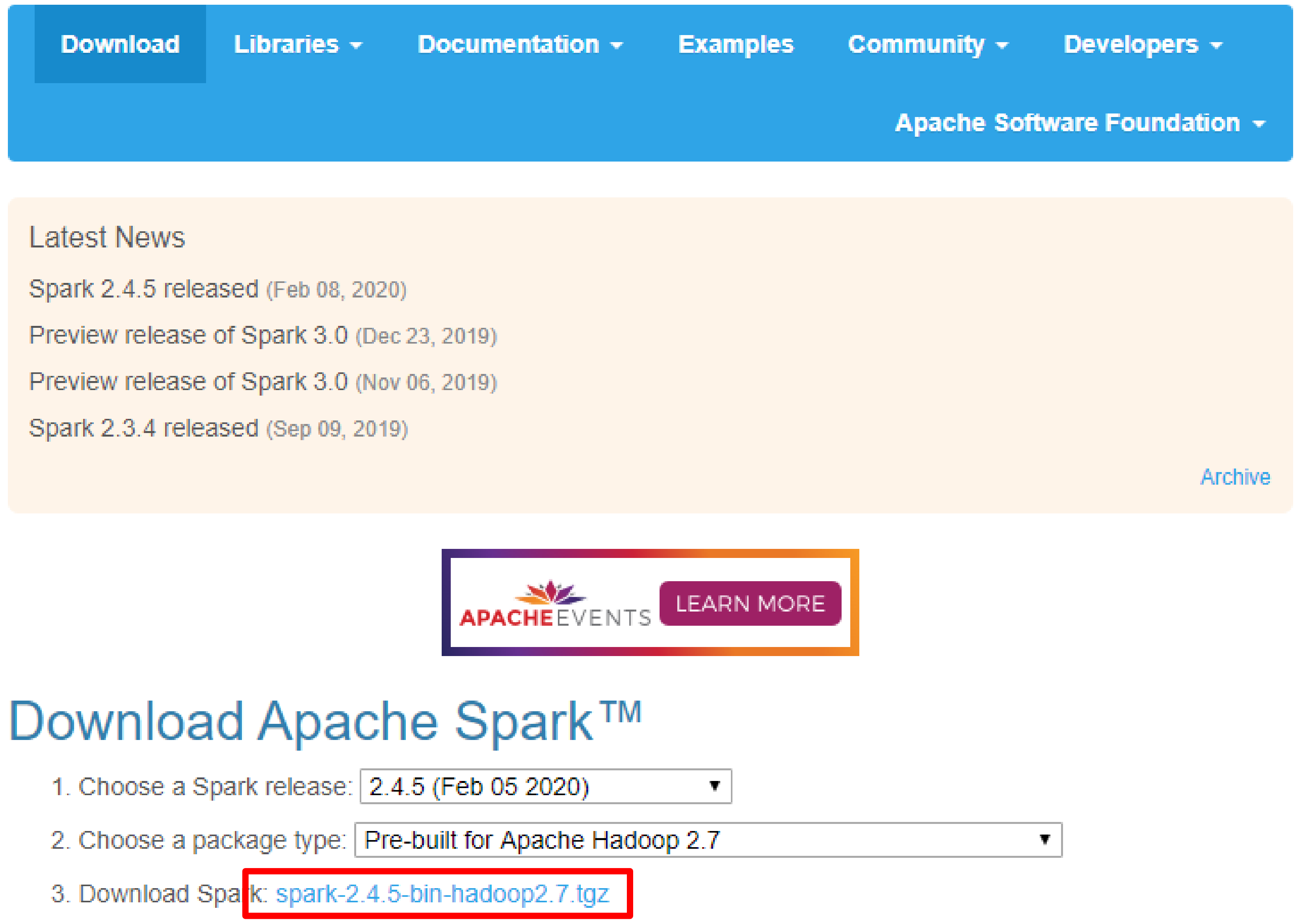
Archive

Apache Spark™ is a unified analytics engine for large-scale data processing.

http://spark.apache.org/

Download Spark

- ➢ **Choose spark 2.4.5 & Apache Hadoop 2.7**



Latest News

Spark 2.4.5 released (Feb 08, 2020)

Preview release of Spark 3.0 (Dec 23, 2019)

Preview release of Spark 3.0 (Nov 06, 2019)

Spark 2.3.4 released (Sep 09, 2019)

Archive

Download Apache Spark™

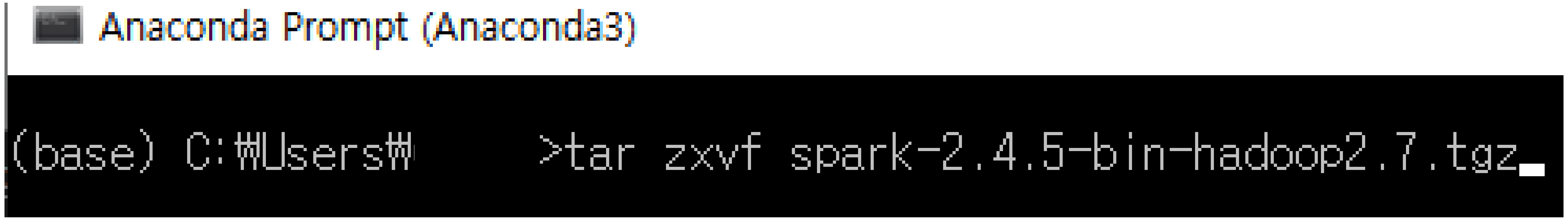
1. Choose a Spark release: 2.4.5 (Feb 05 2020)
2. Choose a package type: Pre-built for Apache Hadoop 2.7
3. Download Spark: spark-2.4.5-bin-hadoop2.7.tgz

- ➢ **Then, download “spark-2.4.5-bin-hadoop2.7.tgz”**
- ➢ **Or, you can download the file here:**

  https://www.apache.org/dyn/closer.lua/spark/spark-2.4.5/spark-2.4.5-bin-hadoop2.7.tgz

Download Spark

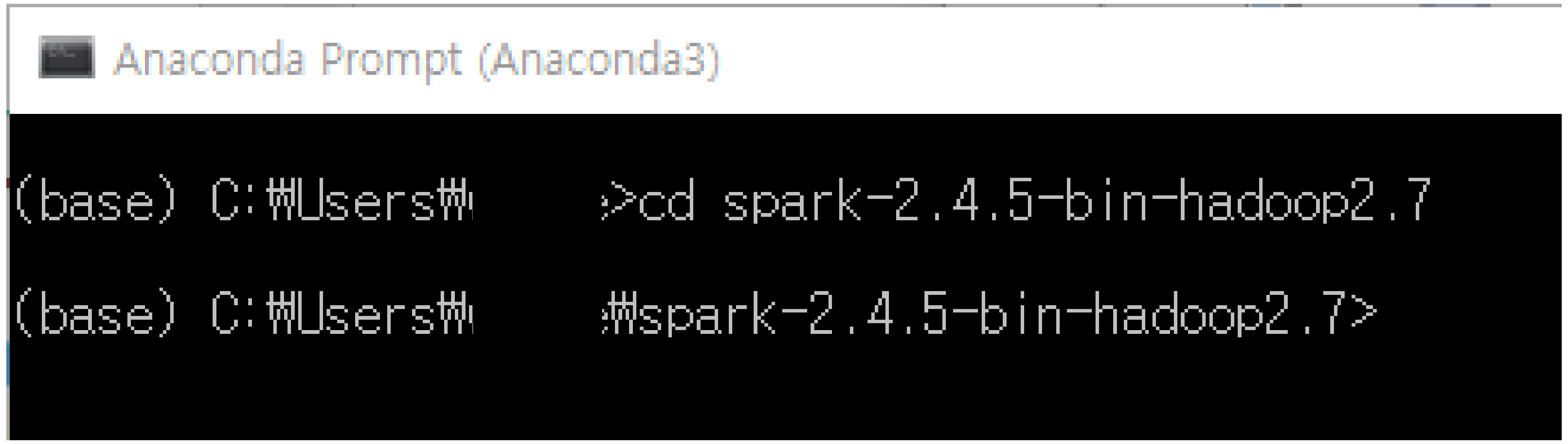
- Go to the folder where the package is installed and unzip the package for installment



Anaconda Prompt (Anaconda3)

```
(base) C:₩Users₩     >tar zxvf spark-2.4.5-bin-hadoop2.7.tgz
```

- After then, you can enter the “spark-2.4.5-bin-Hadoop.2.7” folder

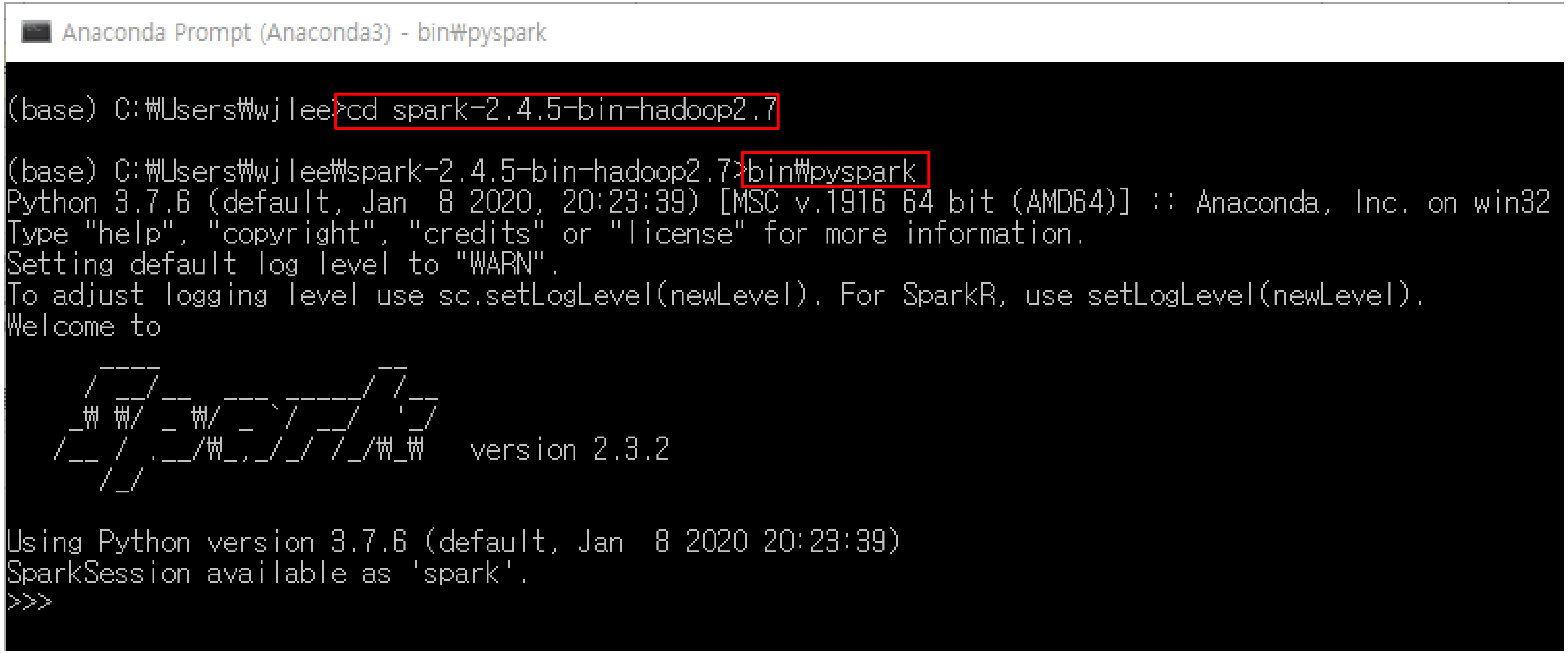


Anaconda Prompt (Anaconda3)

```
(base) C:₩Users₩     >cd spark-2.4.5-bin-hadoop2.7

(base) C:₩Users₩     ₩spark-2.4.5-bin-hadoop2.7>
```

Download Spark

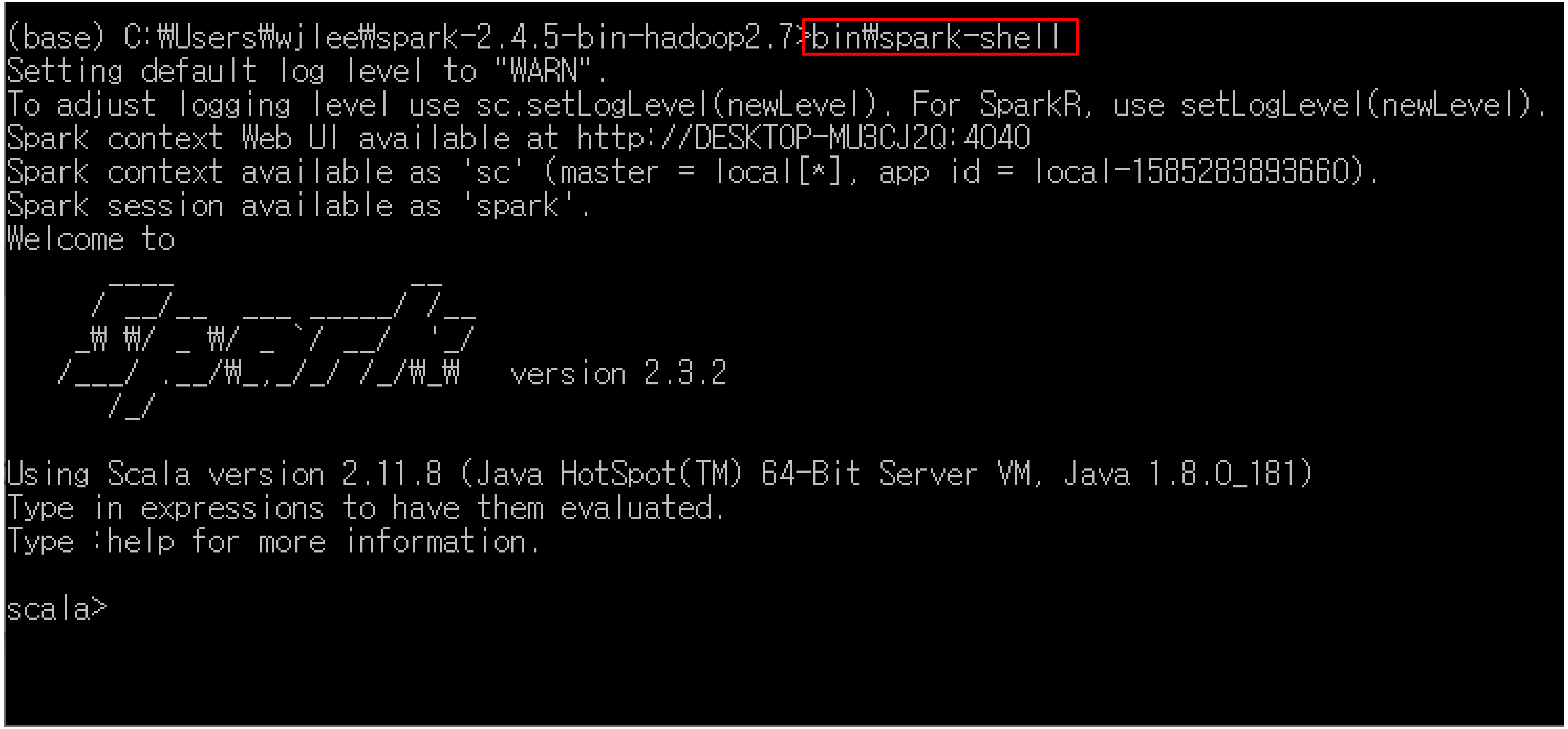
➢ **Type “bin\pyspark” and open Pyspark**



```
(base) C:₩Users₩wjlee>cd spark-2.4.5-bin-hadoop2.7

(base) C:₩Users₩wjlee₩spark-2.4.5-bin-hadoop2.7>bin₩pyspark
Python 3.7.6 (default, Jan  8 2020, 20:23:39) [MSC v.1916 64 bit (AMD64)] :: Anaconda, Inc. on win32
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _₩ ₩/ _ ₩/ _ `/ __/  '_/
   /__ / .__/₩_,_/_/ /_/₩_₩   version 2.3.2
      /_/

Using Python version 3.7.6 (default, Jan  8 2020 20:23:39)
SparkSession available as 'spark'.
>>>
```

Download Spark

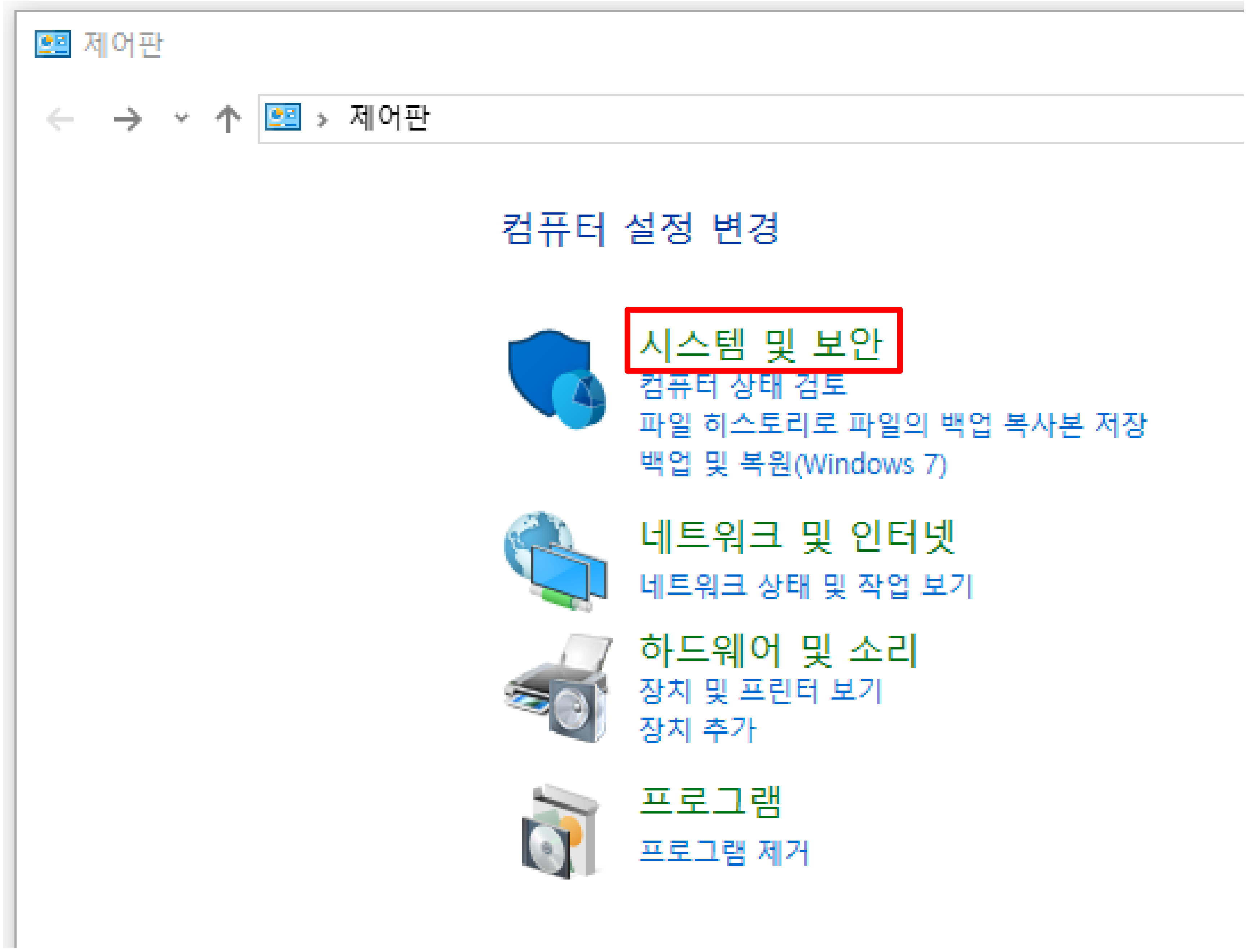
- Type “bin\spark-shell” and open Spark-Shell
- Spark-Shell is based on Scala



```
(base) C:₩Users₩wjlee₩spark-2.4.5-bin-hadoop2.7>bin₩spark-shell
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Spark context Web UI available at http://DESKTOP-MU3CJ2Q:4040
Spark context available as 'sc' (master = local[*], app id = local-1585283893660).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _₩ ₩/ _ ₩/ _ `/ __/  '_/
   /___/ .__/₩_,_/_/ /_/₩_₩   version 2.3.2
      /_/

Using Scala version 2.11.8 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_181)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

Configure System Path

- ➢ **We can configure system path for calling Spark in any folder.**
- ➢ **Enter Control Panel > System and Security > System**



제어판

← → ˅ ↑ › 제어판

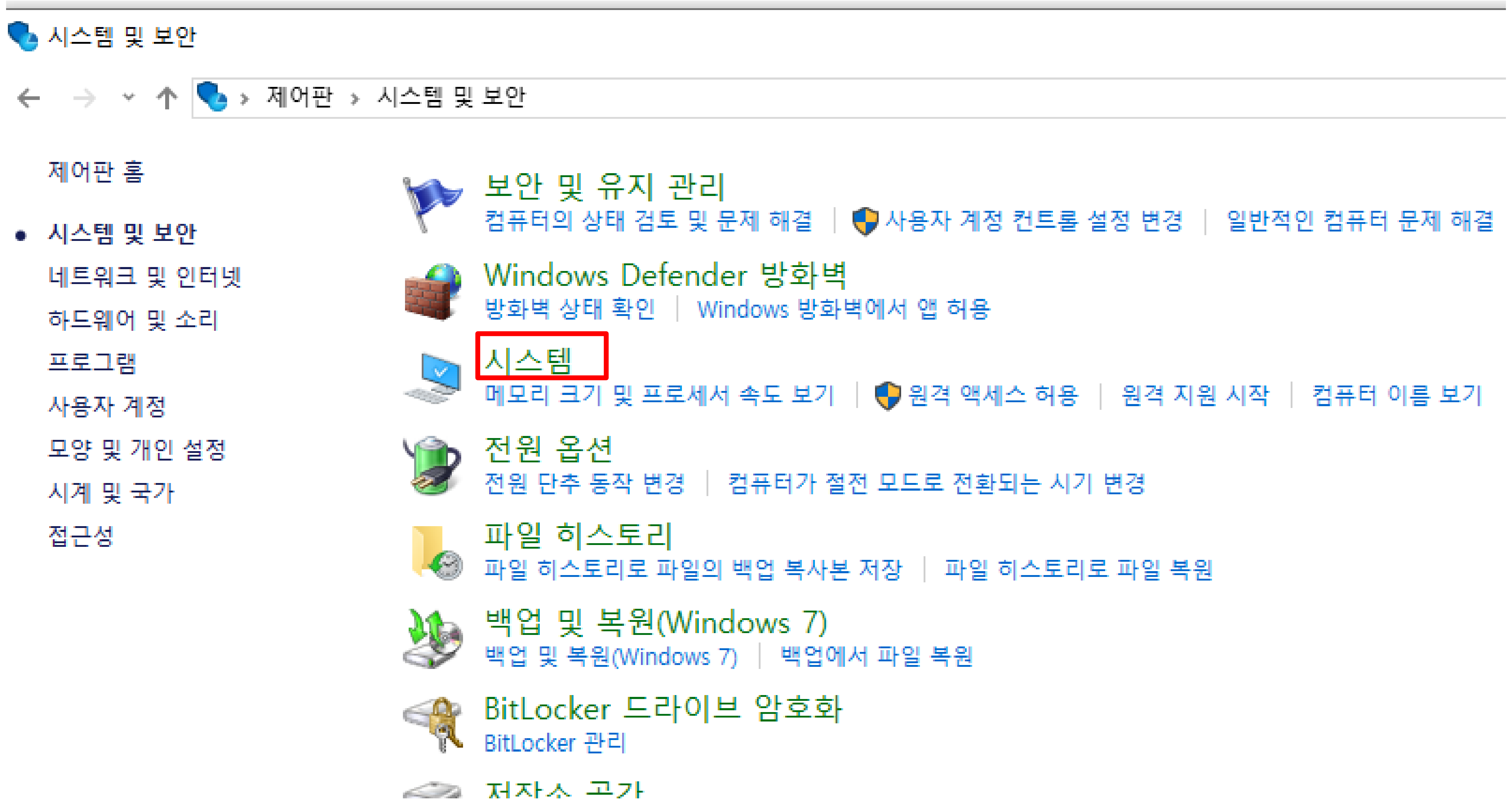
컴퓨터 설정 변경

시스템 및 보안
컴퓨터 상태 검토
파일 히스토리로 파일의 백업 복사본 저장
백업 및 복원(Windows 7)

네트워크 및 인터넷
네트워크 상태 및 작업 보기

하드웨어 및 소리
장치 및 프린터 보기
장치 추가

프로그램
프로그램 제거



시스템 및 보안

← → ˅ ↑ › 제어판 › 시스템 및 보안

제어판 홈

- 시스템 및 보안

네트워크 및 인터넷
하드웨어 및 소리
프로그램
사용자 계정
모양 및 개인 설정
시계 및 국가
접근성

보안 및 유지 관리
컴퓨터의 상태 검토 및 문제 해결 | 사용자 계정 컨트롤 설정 변경 | 일반적인 컴퓨터 문제 해결

Windows Defender 방화벽
방화벽 상태 확인 | Windows 방화벽에서 앱 허용

시스템
메모리 크기 및 프로세서 속도 보기 | 원격 액세스 허용 | 원격 지원 시작 | 컴퓨터 이름 보기

전원 옵션
전원 단추 동작 변경 | 컴퓨터가 절전 모드로 전환되는 시기 변경

파일 히스토리
파일 히스토리로 파일의 백업 복사본 저장 | 파일 히스토리로 파일 복원

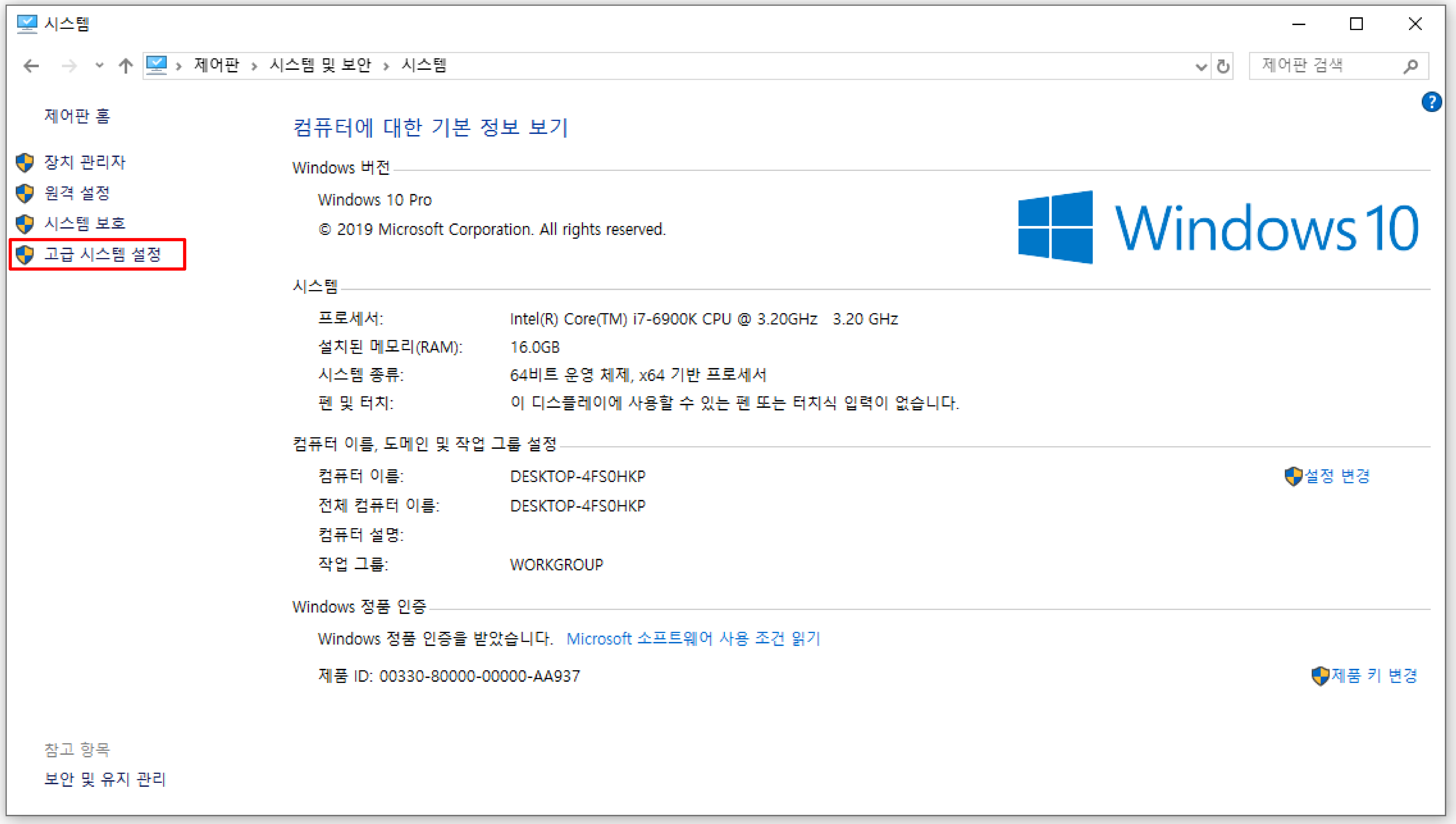
백업 및 복원(Windows 7)
백업 및 복원(Windows 7) | 백업에서 파일 복원

BitLocker 드라이브 암호화
BitLocker 관리

저장소 공간

Configure System Path

- ➢ **System > Advanced System Settings**



컴퓨터에 대한 기본 정보 보기

Windows 버전

Windows 10 Pro

© 2019 Microsoft Corporation. All rights reserved.

시스템

프로세서:	Intel(R) Core(TM) i7-6900K CPU @ 3.20GHz 3.20 GHz
설치된 메모리(RAM):	16.0GB
시스템 종류:	64비트 운영 체제, x64 기반 프로세서
펜 및 터치:	이 디스플레이에 사용할 수 있는 펜 또는 터치식 입력이 없습니다.

컴퓨터 이름, 도메인 및 작업 그룹 설정

컴퓨터 이름:	DESKTOP-4FS0HKP
전체 컴퓨터 이름:	DESKTOP-4FS0HKP
컴퓨터 설명:	
작업 그룹:	WORKGROUP

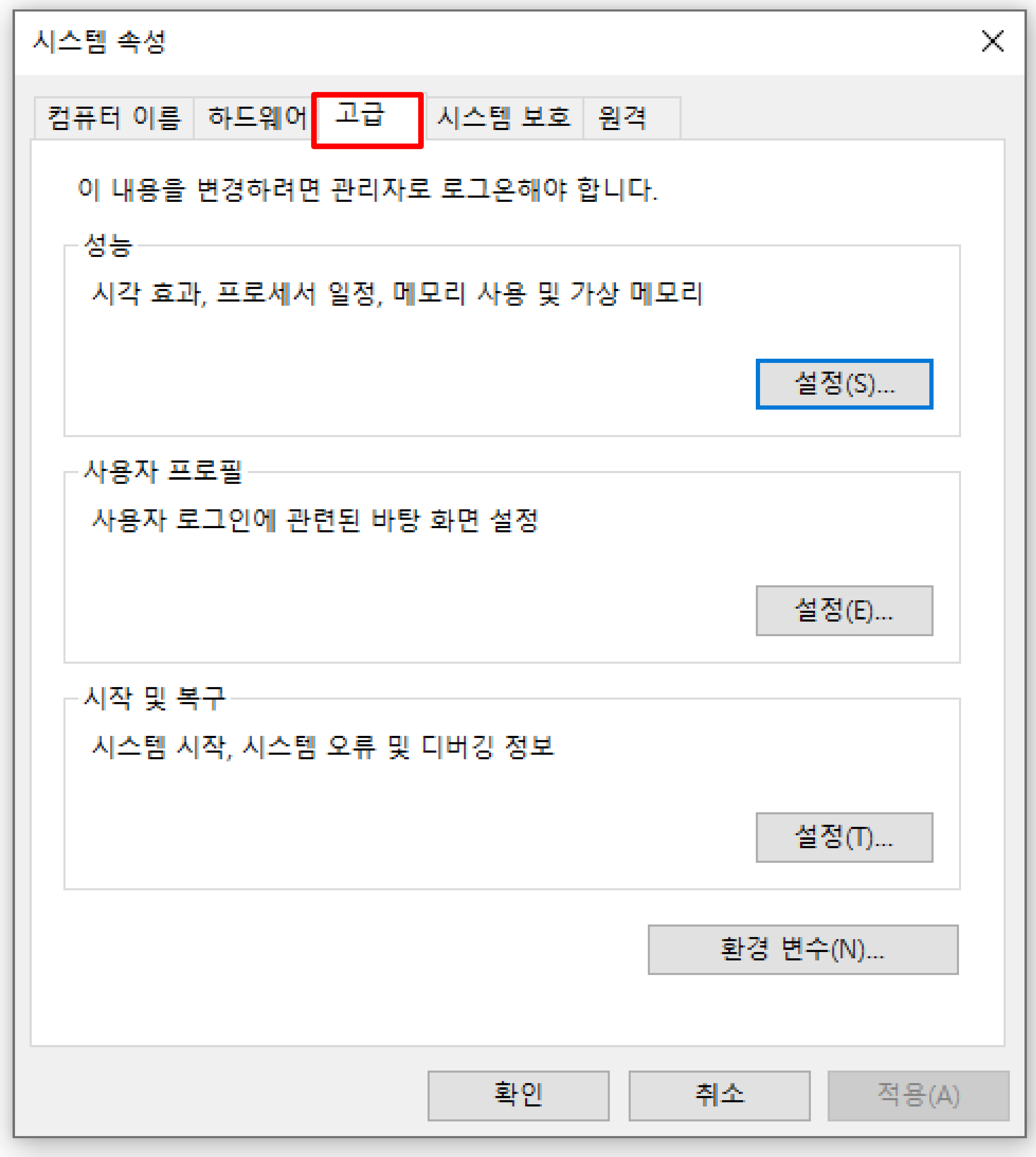
Windows 정품 인증

Windows 정품 인증을 받았습니다. Microsoft 소프트웨어 사용 조건 읽기

제품 ID: 00330-80000-00000-AA937

Configure System Path

- ➢ **“Advanced” tab > Environment Variables**
- ➢ **Add Spark folder to Path variable**



시스템 속성

이 내용을 변경하려면 관리자로 로그온해야 합니다.

성능

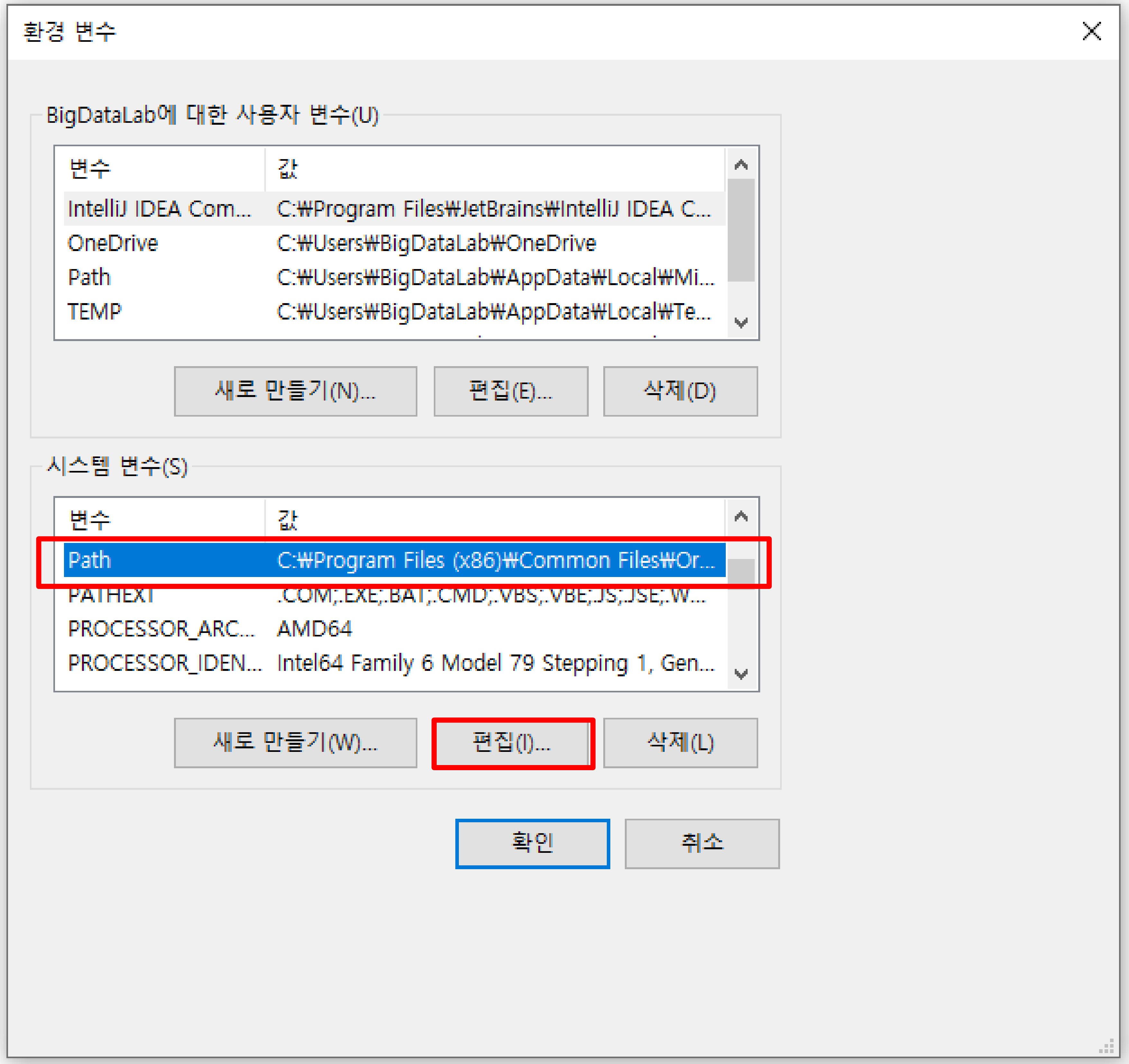
시각 효과, 프로세서 일정, 메모리 사용 및 가상 메모리

사용자 프로필

사용자 로그인에 관련된 바탕 화면 설정

시작 및 복구

시스템 시작, 시스템 오류 및 디버깅 정보



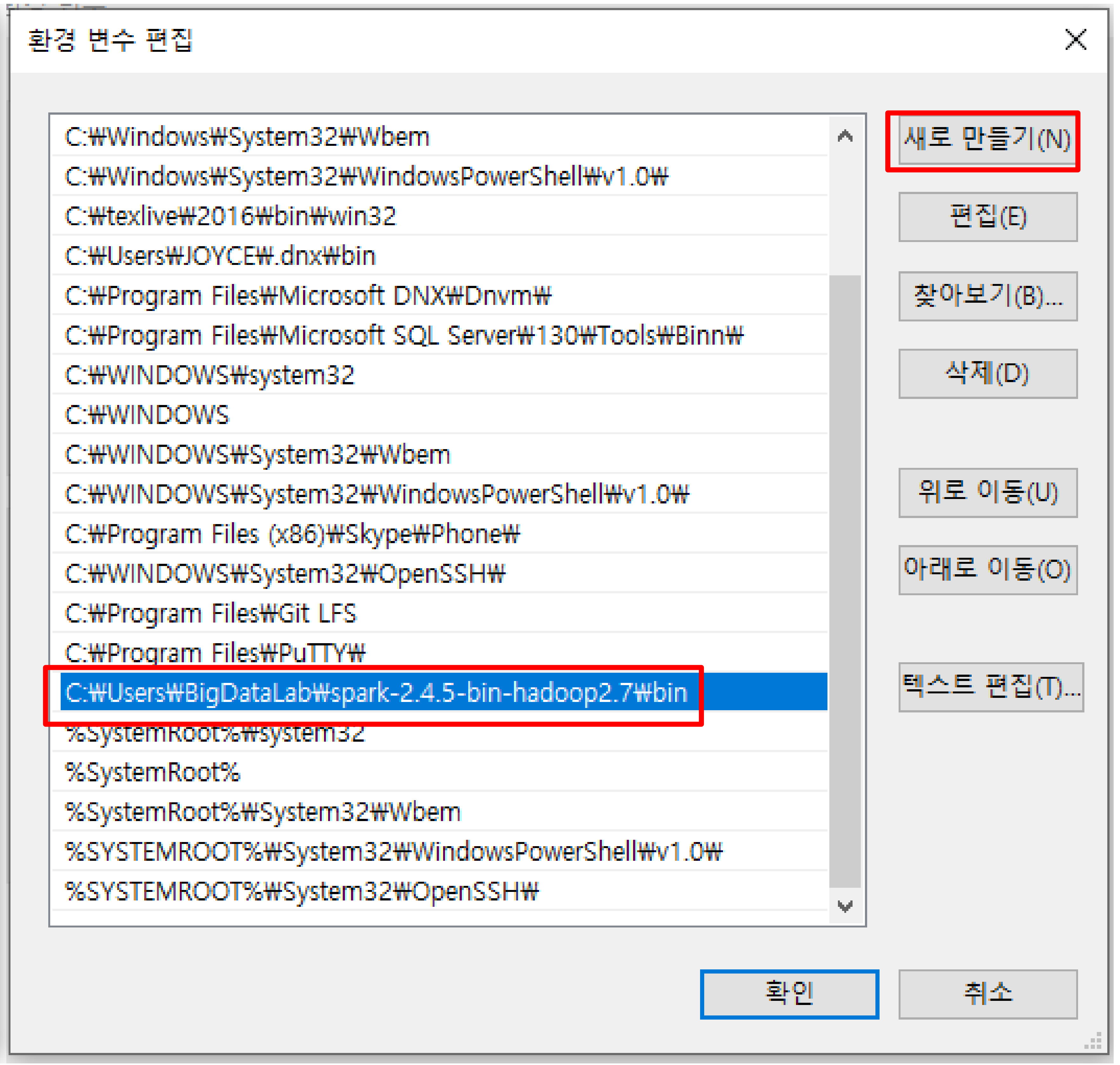
환경 변수

BigDataLab에 대한 사용자 변수(U)

변수	값
IntelliJ IDEA Com...	C:₩Program Files₩JetBrains₩IntelliJ IDEA C...
OneDrive	C:₩Users₩BigDataLab₩OneDrive
Path	C:₩Users₩BigDataLab₩AppData₩Local₩Mi...
TEMP	C:₩Users₩BigDataLab₩AppData₩Local₩Te...

시스템 변수(S)

변수	값
Path	C:₩Program Files (x86)₩Common Files₩Or...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.W...
PROCESSOR_ARC...	AMD64
PROCESSOR_IDEN...	Intel64 Family 6 Model 79 Stepping 1, Gen...



환경 변수 편집

```
C:₩Windows₩System32₩Wbem
C:₩Windows₩System32₩WindowsPowerShell₩v1.0₩
C:₩texlive₩2016₩bin₩win32
C:₩Users₩JOYCE₩.dnx₩bin
C:₩Program Files₩Microsoft DNX₩Dnvm₩
C:₩Program Files₩Microsoft SQL Server₩130₩Tools₩Binn₩
C:₩WINDOWS₩system32
C:₩WINDOWS
C:₩WINDOWS₩System32₩Wbem
C:₩WINDOWS₩System32₩WindowsPowerShell₩v1.0₩
C:₩Program Files (x86)₩Skype₩Phone₩
C:₩WINDOWS₩System32₩OpenSSH₩
C:₩Program Files₩Git LFS
C:₩Program Files₩PuTTY₩
C:₩Users₩BigDataLab₩spark-2.4.5-bin-hadoop2.7₩bin
%SystemRoot%₩system32
%SystemRoot%
%SystemRoot%₩System32₩Wbem
%SYSTEMROOT%₩System32₩WindowsPowerShell₩v1.0₩
%SYSTEMROOT%₩System32₩OpenSSH₩
```

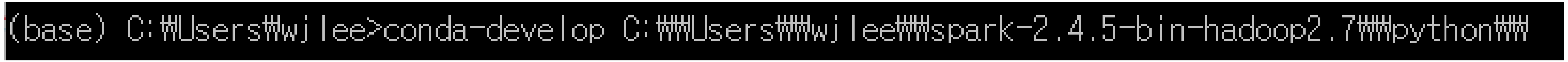
- ➢ **In our case, we add the path “C:\Users\BigdataLab\spark-2.4.5-bin-hadoop2.7\bin**

Configure System Path

- ➢ **Return to Anaconda Prompt, and configure conda path with the following command**

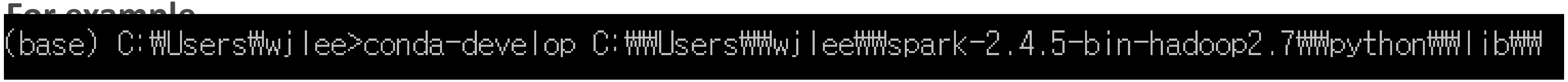
conda-develop SPARK_PATH\\python\\

conda-develop SPARK_PATH\\python\\lib\\



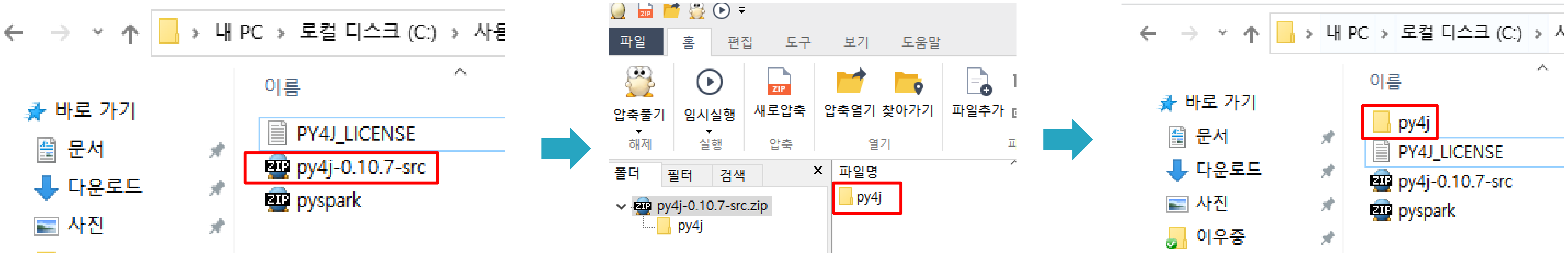
```
(base) C:₩Users₩wjlee>conda-develop C:₩₩Users₩₩wjlee₩₩spark-2.4.5-bin-hadoop2.7₩₩python₩₩
```

For example



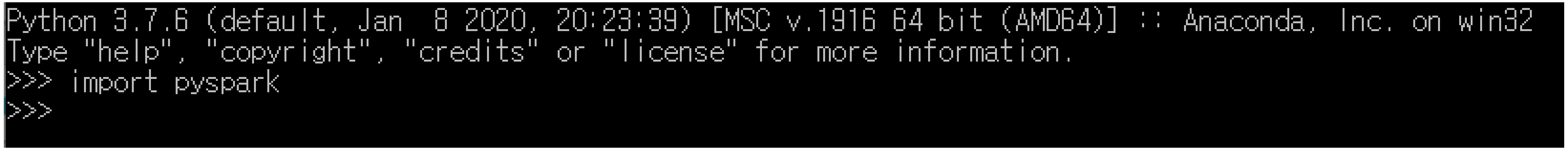
```
(base) C:₩Users₩wjlee>conda-develop C:₩₩Users₩₩wjlee₩₩spark-2.4.5-bin-hadoop2.7₩₩python₩₩lib₩₩
```

- ➢ **Go to SPARK_PATH\python\lib, then drag py4j folder out from py4j-0.10.7-scr.zip**



Configure System Path

- Then, you can import pyspark library in Python



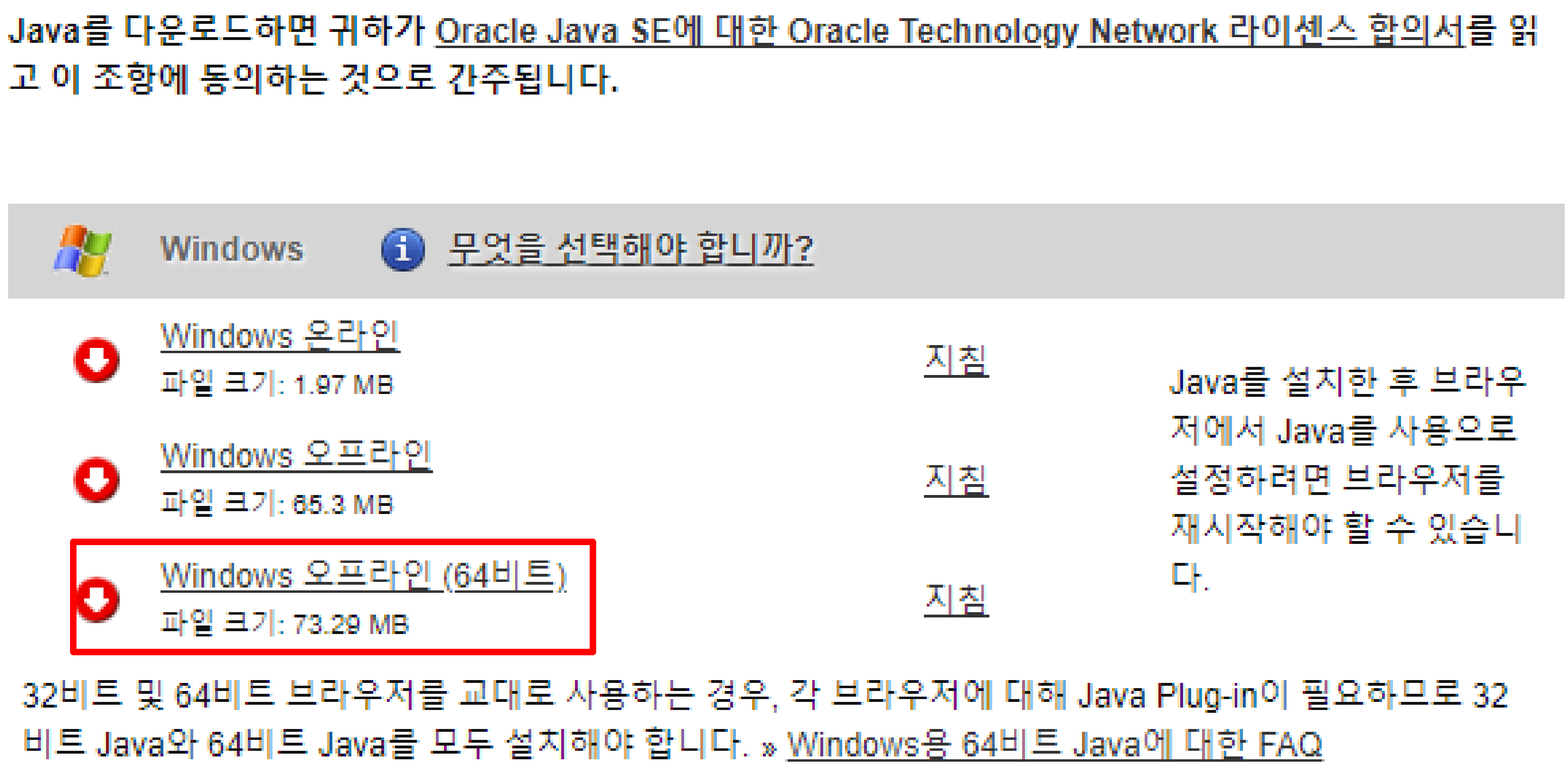
```
Python 3.7.6 (default, Jan  8 2020, 20:23:39) [MSC v.1916 64 bit (AMD64)] :: Anaconda, Inc. on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import pyspark
>>>
```

Hadoop installation

Download JAVA

- ➢ **Download JAVA installation file (64bit)**

  Refer the link on the bottom.



Java를 다운로드하면 귀하가 Oracle Java SE에 대한 Oracle Technology Network 라이센스 합의서를 읽고 이 조항에 동의하는 것으로 간주됩니다.

Windows 무엇을 선택해야 합니까?

Windows 온라인
파일 크기: 1.97 MB
지침

Windows 오프라인
파일 크기: 65.3 MB
지침

Windows 오프라인 (64비트)
파일 크기: 73.29 MB
지침

Java를 설치한 후 브라우저에서 Java를 사용으로 설정하려면 브라우저를 재시작해야 할 수 있습니다.

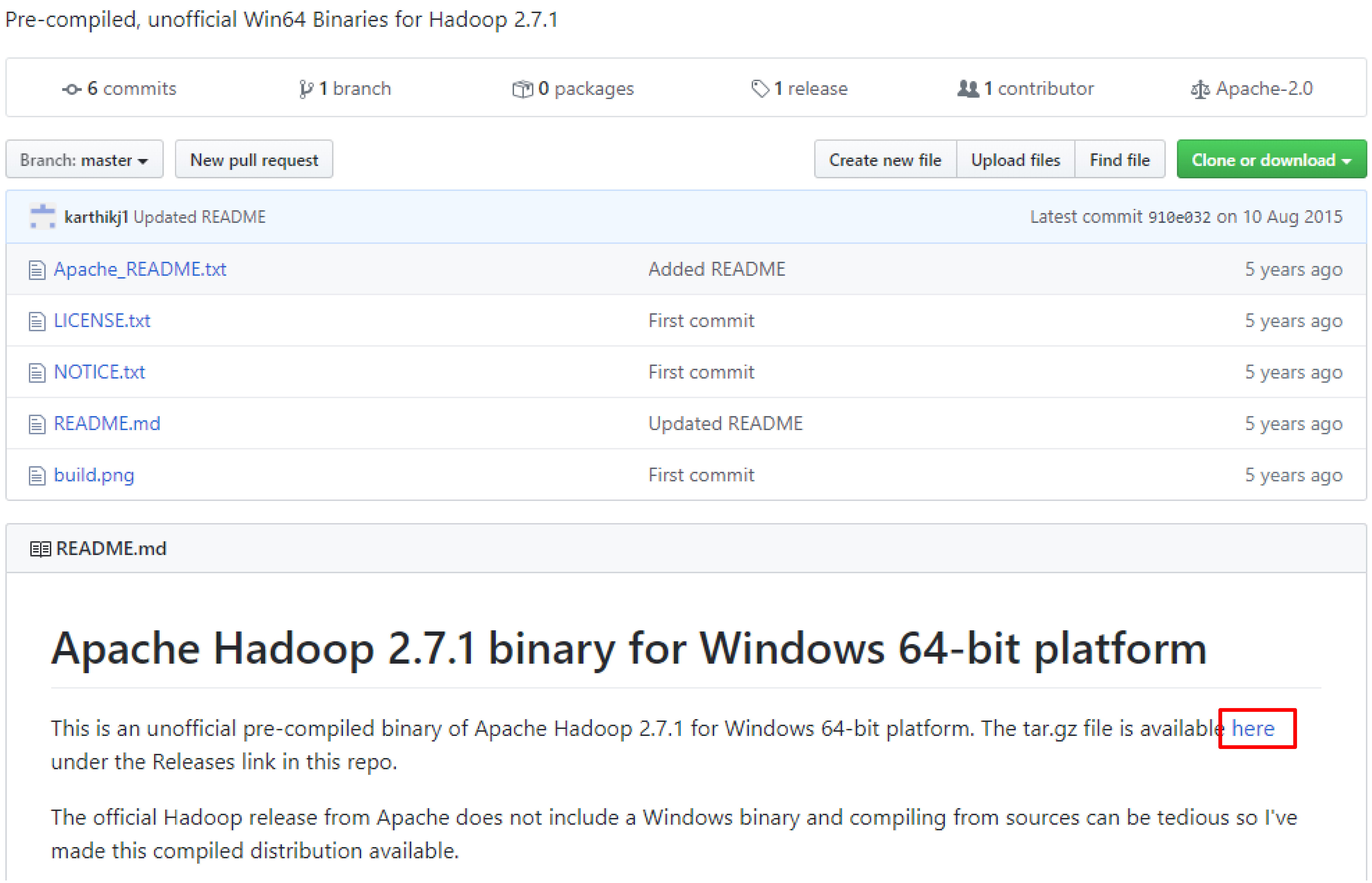
32비트 및 64비트 브라우저를 교대로 사용하는 경우, 각 브라우저에 대해 Java Plug-in이 필요하므로 32비트 Java와 64비트 Java를 모두 설치해야 합니다. » Windows용 64비트 Java에 대한 FAQ

- ➢ **In our case, we install JAVA compatible with Windows OS.**
- ➢ **If your desktop has another OS, you need to install another version of JAVA.**

https://www.java.com/ko/download/manual.jsp

Install Hadoop

- **Download pre-compiled package for Windows**



Pre-compiled, unofficial Win64 Binaries for Hadoop 2.7.1

karthikj1 Updated README — Latest commit 910e032 on 10 Aug 2015

Apache_README.txt	Added README	5 years ago
LICENSE.txt	First commit	5 years ago
NOTICE.txt	First commit	5 years ago
README.md	Updated README	5 years ago
build.png	First commit	5 years ago

README.md

Apache Hadoop 2.7.1 binary for Windows 64-bit platform

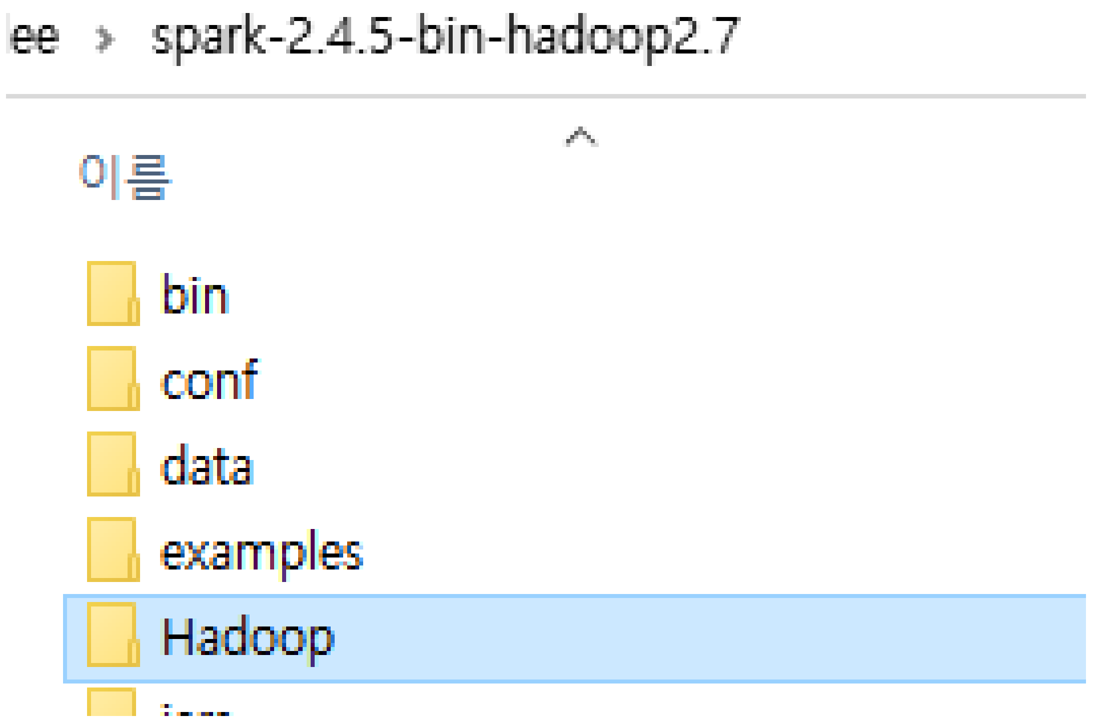
This is an unofficial pre-compiled binary of Apache Hadoop 2.7.1 for Windows 64-bit platform. The tar.gz file is available here under the Releases link in this repo.

The official Hadoop release from Apache does not include a Windows binary and compiling from sources can be tedious so I've made this compiled distribution available.

https://github.com/karthikj1/Hadoop-2.7.1-Windows-64-binaries

Install Hadoop

- In your Spark folder, make a new folder named Hadoop



ee › spark-2.4.5-bin-hadoop2.7

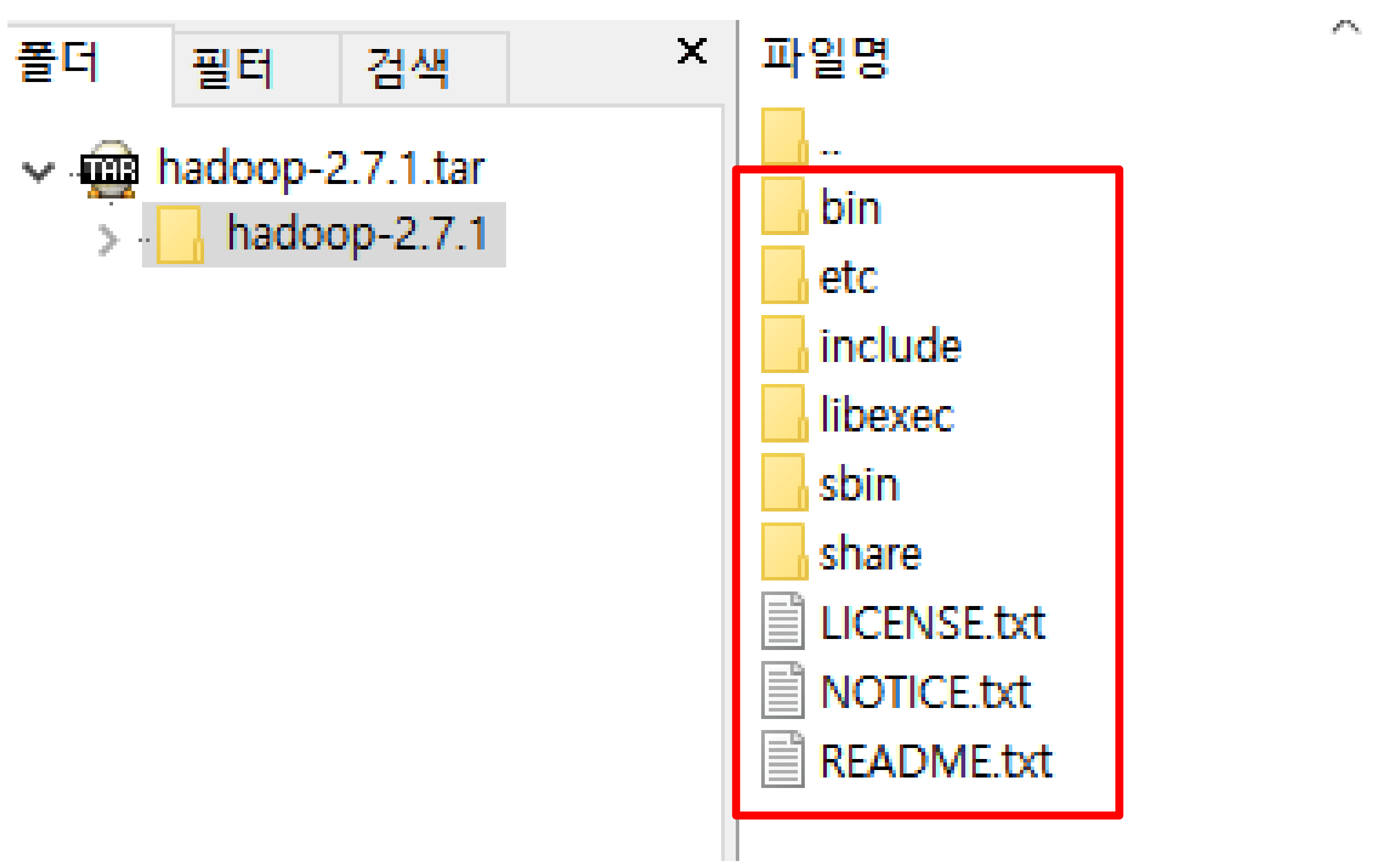
이름

bin
conf
data
examples
Hadoop

- Open the Hadoop tar.zip file and move all folder or files into Hadoop folder

NOTE: We recommend there must be no space in the path of new folder

For example "C:\Users\big data class\spark\Hadoop folder" (X)

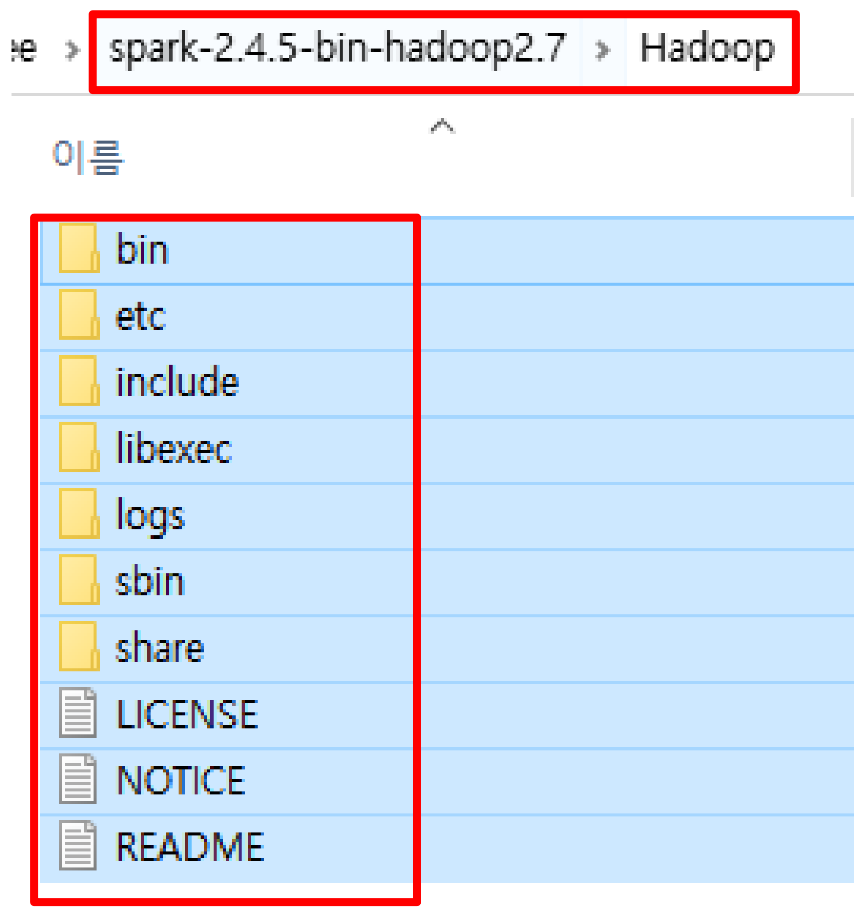


폴더 필터 검색 ✕

hadoop-2.7.1.tar
hadoop-2.7.1

파일명

..
bin
etc
include
libexec
sbin
share
LICENSE.txt
NOTICE.txt
README.txt



ee › spark-2.4.5-bin-hadoop2.7 › Hadoop

이름

bin
etc
include
libexec
logs
sbin
share
LICENSE
NOTICE
README

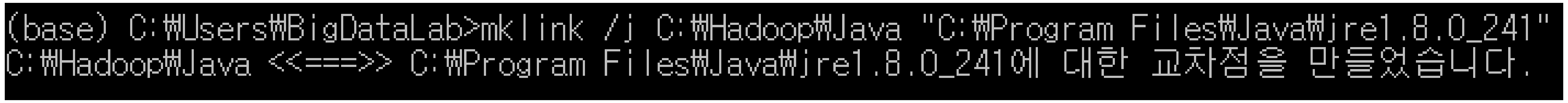
Configure PATH

- ➢ Make new folder “C:\Hadoop” and Return to Anaconda prompt
- ➢ Link java jre file to “C:\Hadoop” with the following command

NOTE: There must be no SPACE in your path!

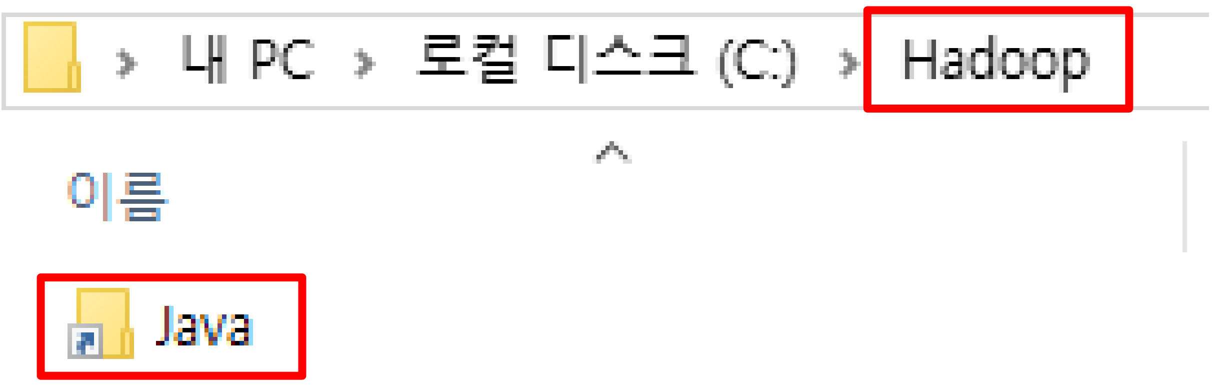
mklink /j C:\Hadoop\Java "C:\Program Files\Java\jre1.8.0_241“

- ➢ Then you can see system successfully link two paths like following



```
(base) C:₩Users₩BigDataLab>mklink /j C:₩Hadoop₩Java "C:₩Program Files₩Java₩jre1.8.0_241"
C:₩Hadoop₩Java <<===>> C:₩Program Files₩Java₩jre1.8.0_241에 대한 교차점을 만들었습니다.
```

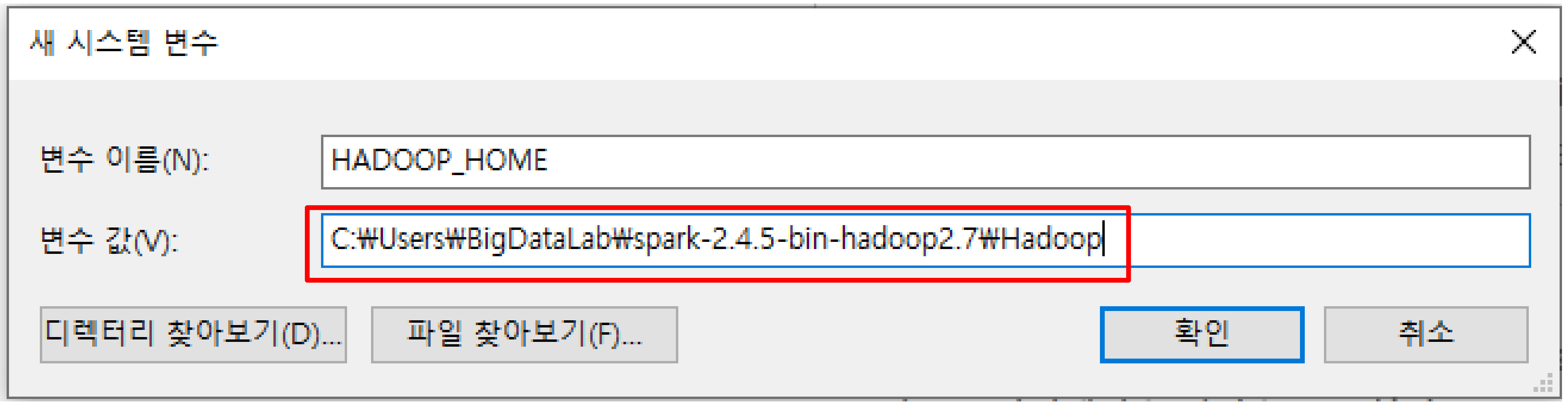
- ➢ Finally, you can see JAVA folder emerges in your “C:\Hadoop”



Configure PATH

- ➢ **Enter Control Panel > System and Security > System > Advanced System Settings > Environment variables**

- ➢ **Then, Add variable “HADOOP_HOME” and “JAVA_HOME” like following**



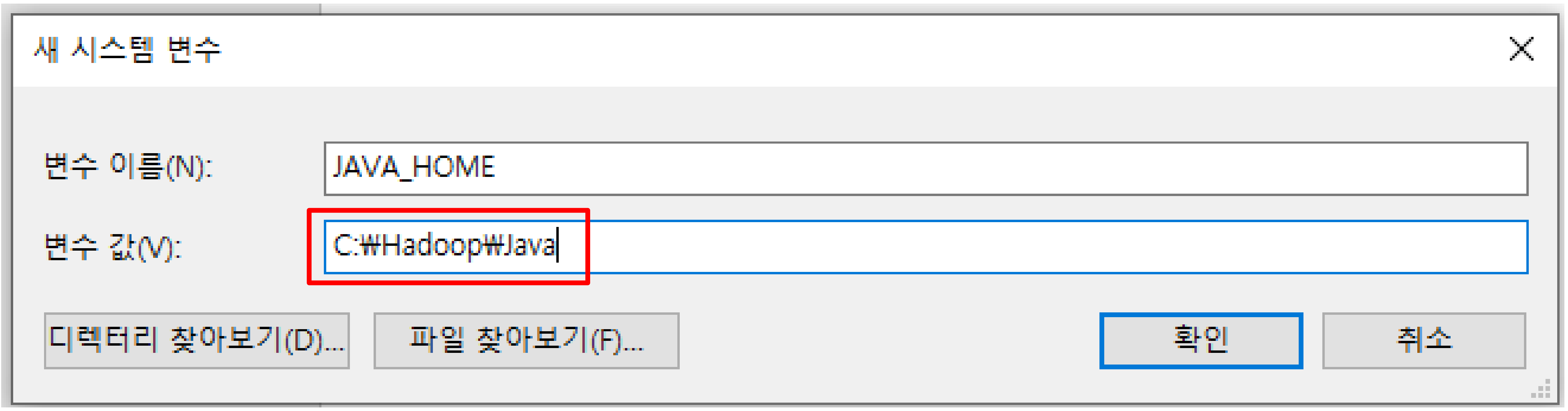
새 시스템 변수

변수 이름(N): HADOOP_HOME

변수 값(V): C:₩Users₩BigDataLab₩spark-2.4.5-bin-hadoop2.7₩Hadoop

HADOOP_HOME :

The folder where Hadoop files installed



새 시스템 변수

변수 이름(N): JAVA_HOME

변수 값(V): C:₩Hadoop₩Java

JAVA_HOME :

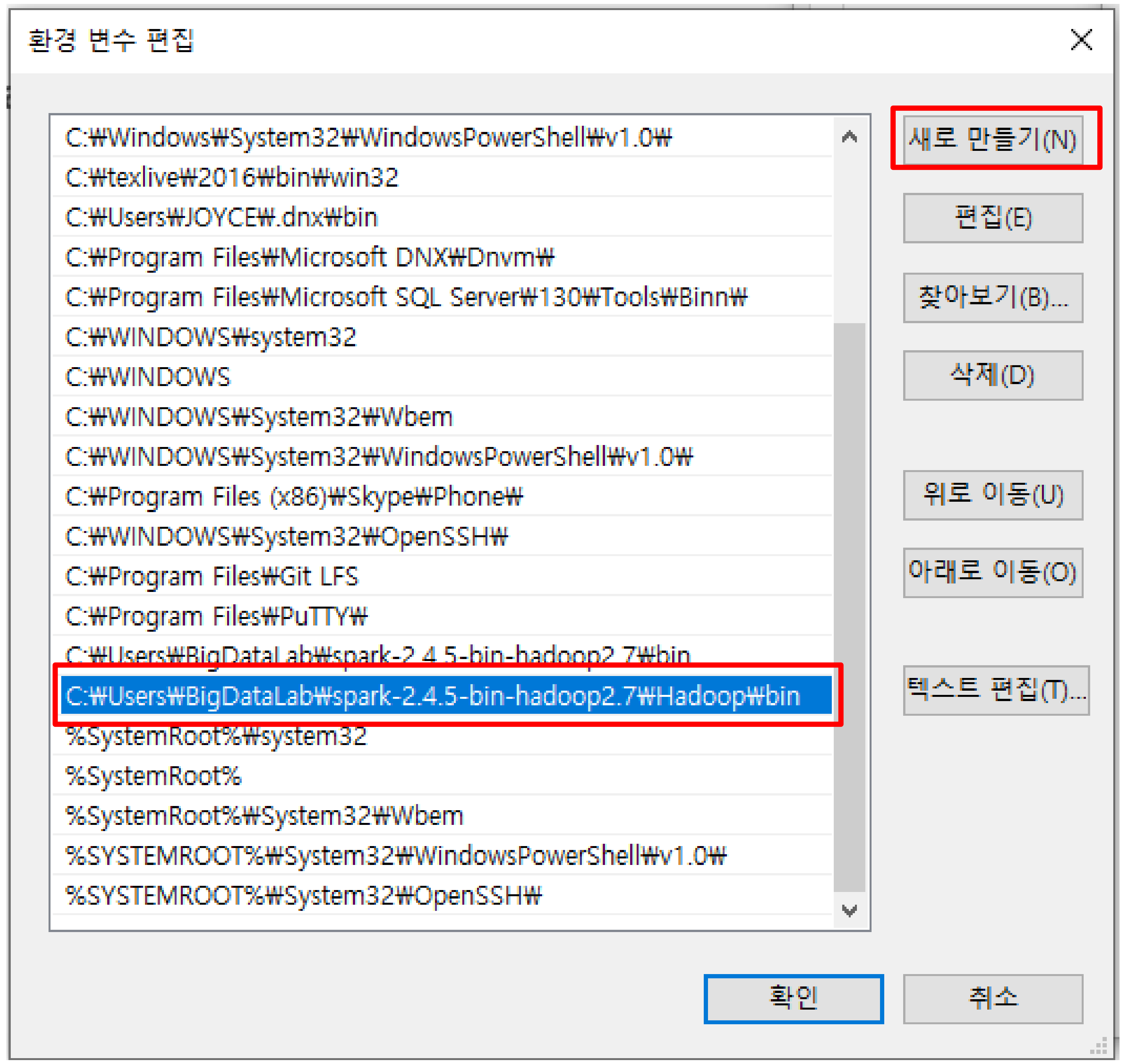
The link folder to JAVA

Configure PATH

- ➢ Add “bin” folder of Hadoop folder to “Path” variable

  Bin folder directory: *YOUR_SPARK_PATH\Hadoop\bin*

- ➢ For example, look at the following



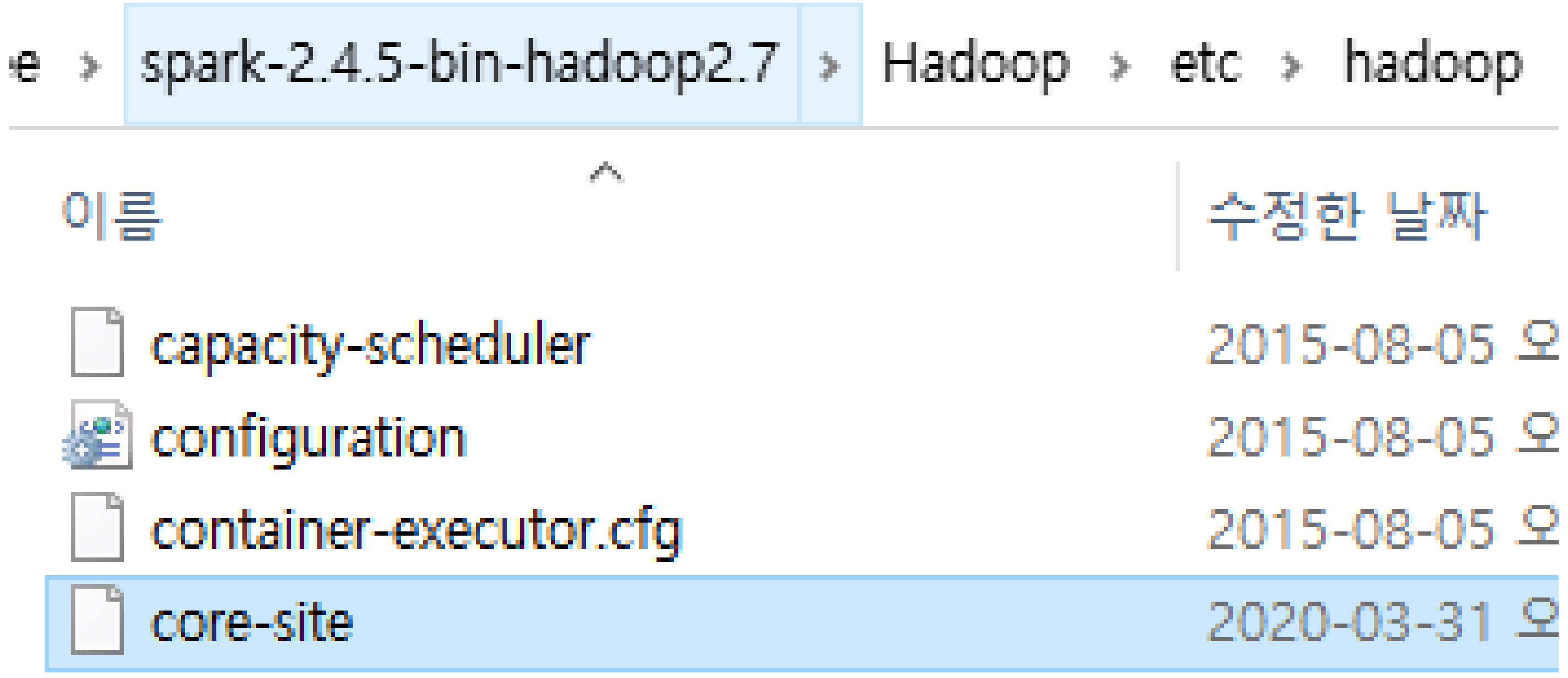
환경 변수 편집

```
C:₩Windows₩System32₩WindowsPowerShell₩v1.0₩
C:₩texlive₩2016₩bin₩win32
C:₩Users₩JOYCE₩.dnx₩bin
C:₩Program Files₩Microsoft DNX₩Dnvm₩
C:₩Program Files₩Microsoft SQL Server₩130₩Tools₩Binn₩
C:₩WINDOWS₩system32
C:₩WINDOWS
C:₩WINDOWS₩System32₩Wbem
C:₩WINDOWS₩System32₩WindowsPowerShell₩v1.0₩
C:₩Program Files (x86)₩Skype₩Phone₩
C:₩WINDOWS₩System32₩OpenSSH₩
C:₩Program Files₩Git LFS
C:₩Program Files₩PuTTY₩
C:₩Users₩BigDataLab₩spark-2.4.5-bin-hadoop2.7₩bin
C:₩Users₩BigDataLab₩spark-2.4.5-bin-hadoop2.7₩Hadoop₩bin
%SystemRoot%₩system32
%SystemRoot%
%SystemRoot%₩System32₩Wbem
%SYSTEMROOT%₩System32₩WindowsPowerShell₩v1.0₩
%SYSTEMROOT%₩System32₩OpenSSH₩
```

Hadoop Configuration

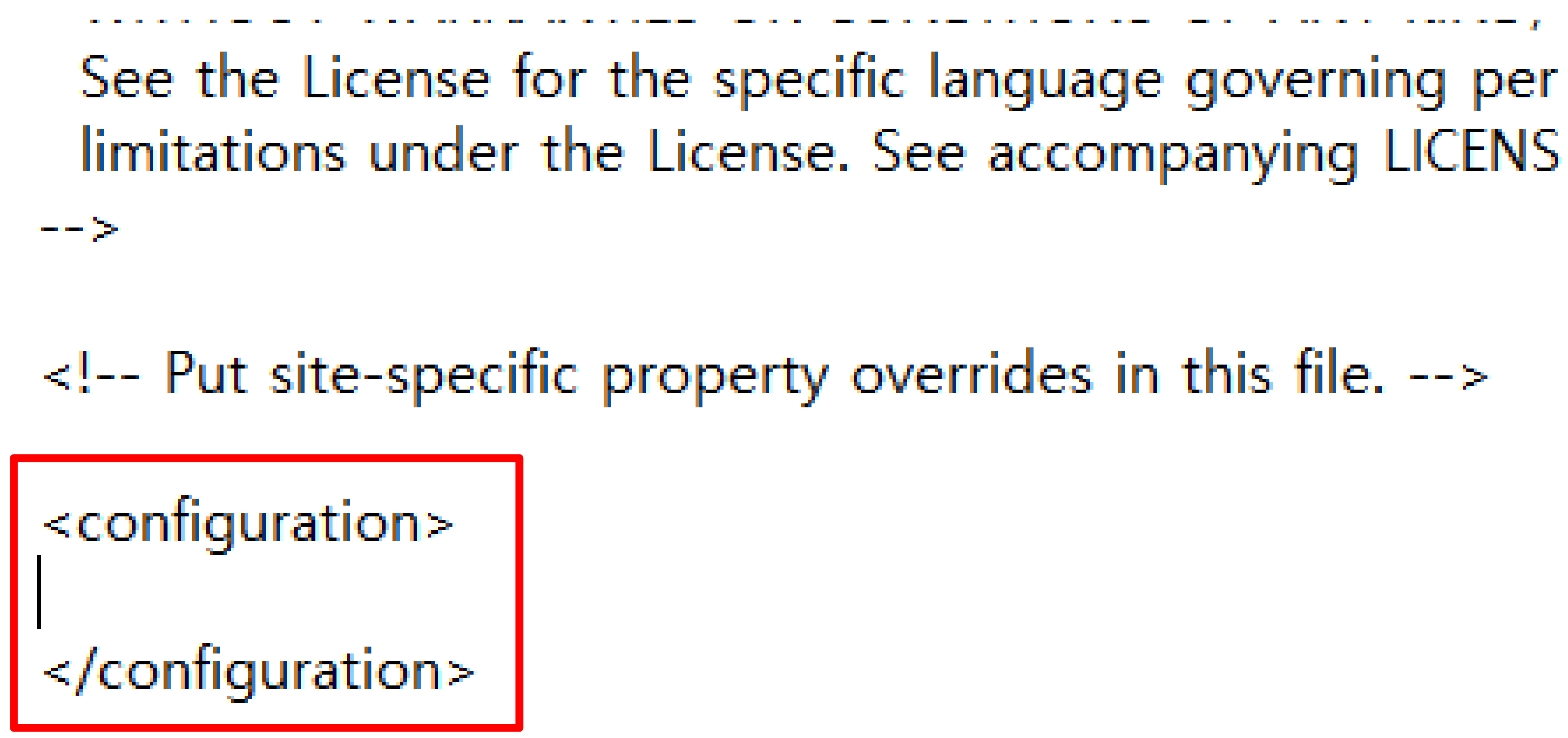
➢ **Before start Hadoop file system, we must edit some configuration file**

- **Open %HADOOP_HOME%\etc\Hadoop\core-site.xml**



e › spark-2.4.5-bin-hadoop2.7 › Hadoop › etc › hadoop

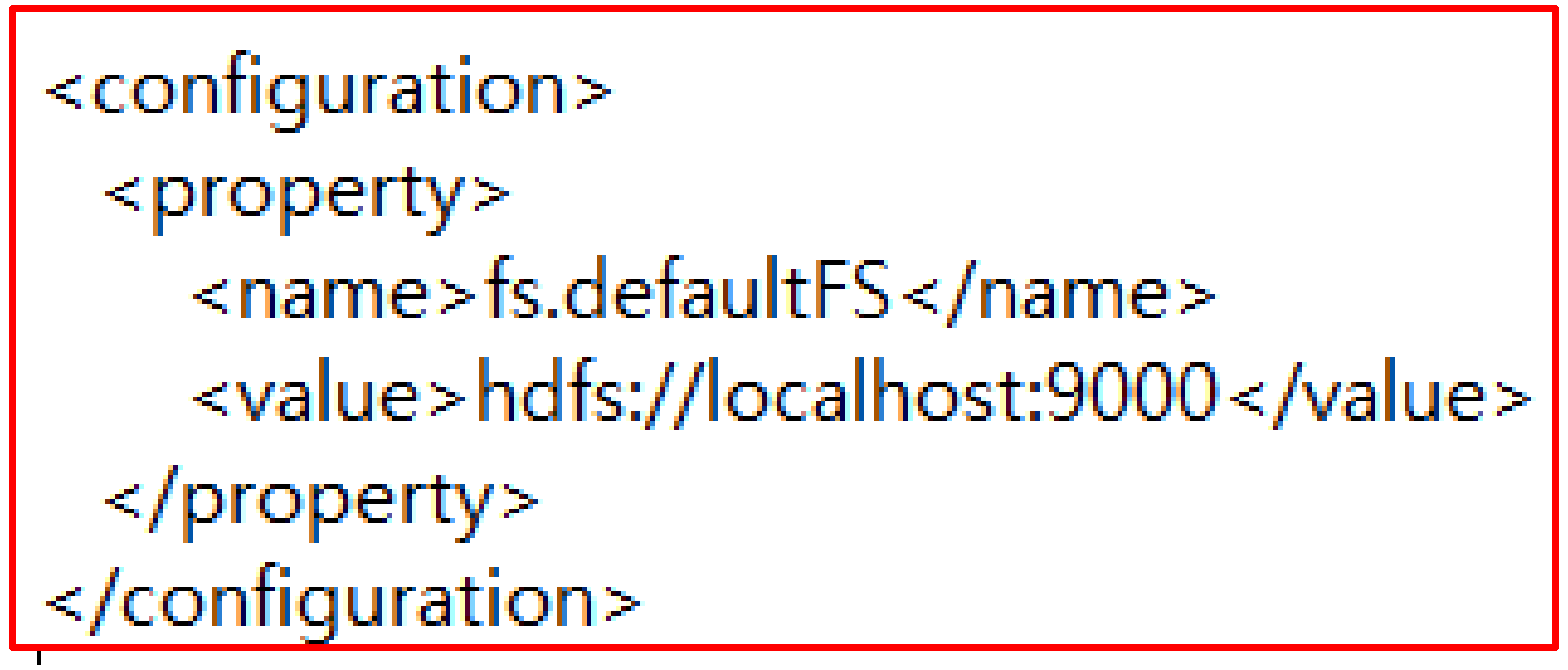
이름	수정한 날짜
capacity-scheduler	2015-08-05 오
configuration	2015-08-05 오
container-executor.cfg	2015-08-05 오
core-site	2020-03-31 오



```
  See the License for the specific language governing per
  limitations under the License. See accompanying LICENS
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>

</configuration>
```

- **Edit CONFIGURATION like following**

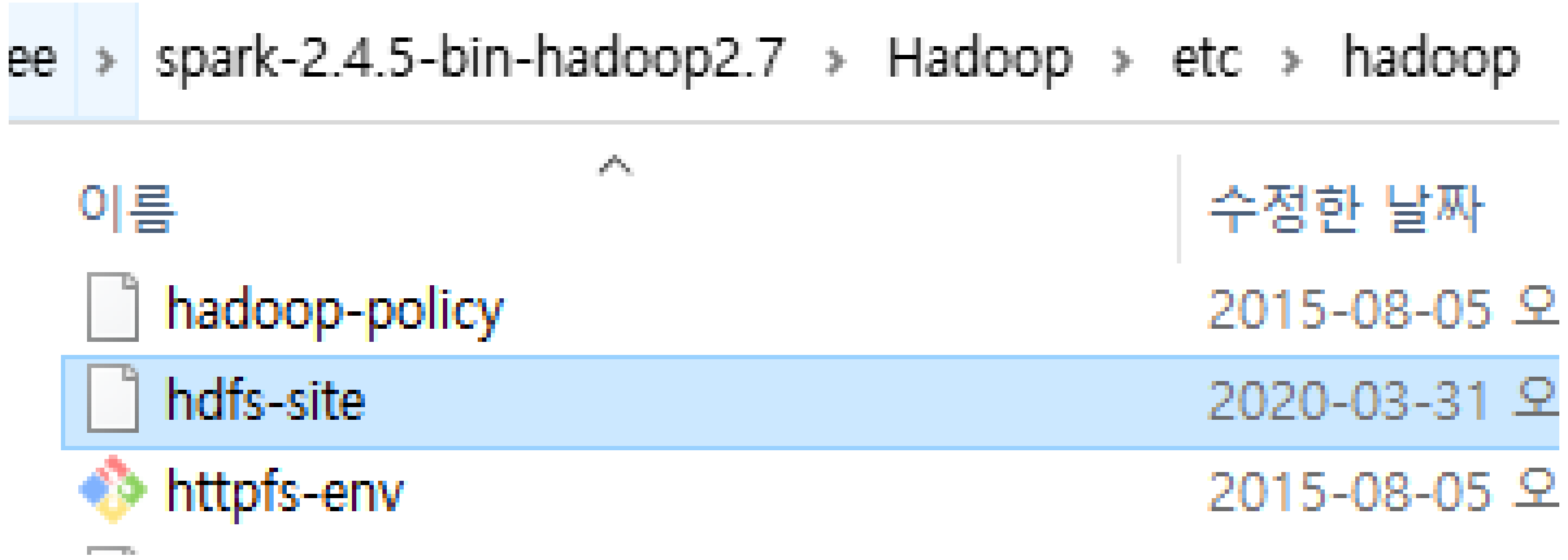
```
<configuration>
 <property>
  <name>fs.defaultFS</name>
      <value>hdfs://localhost:9000</value>
 </property>
</configuration>
```



```
<configuration>
  <property>
     <name>fs.defaultFS</name>
     <value>hdfs://localhost:9000</value>
  </property>
</configuration>
```

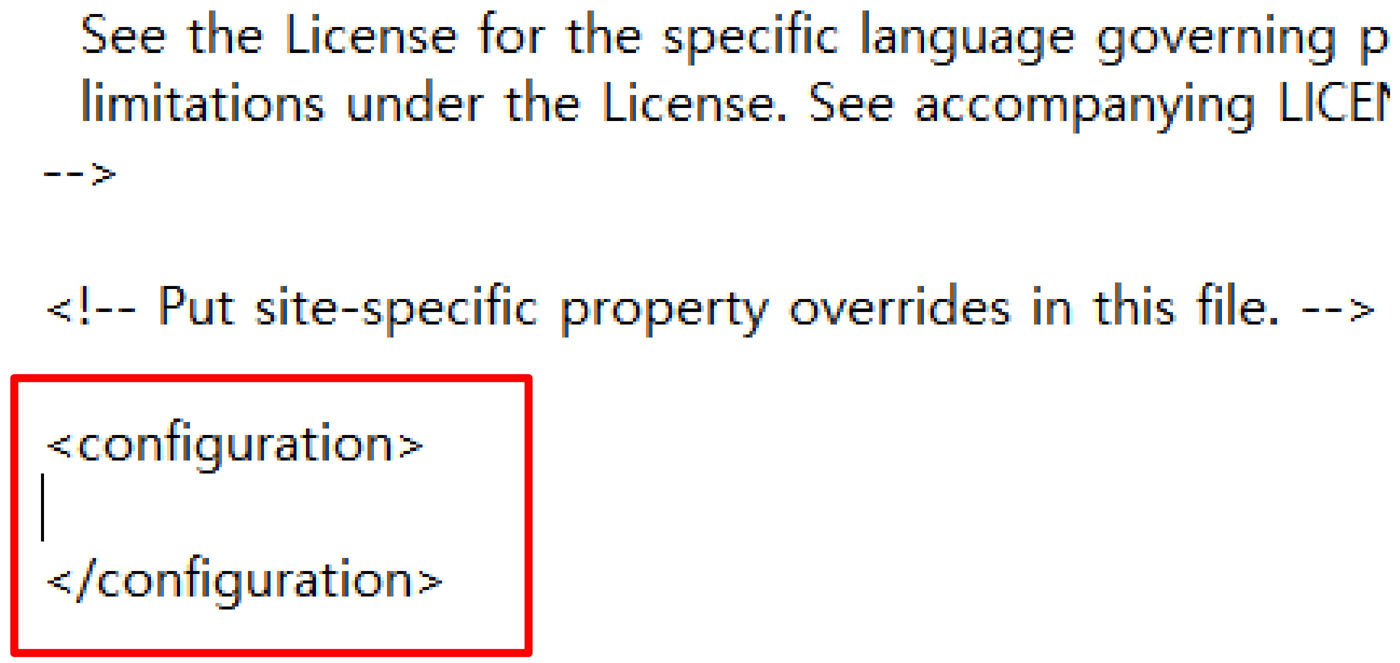
Hadoop Configuration

- Before start Hadoop file system, we must edit some configuration file
  - Open %HADOOP_HOME%\etc\Hadoop\hdfs-site.xml



ee › spark-2.4.5-bin-hadoop2.7 › Hadoop › etc › hadoop

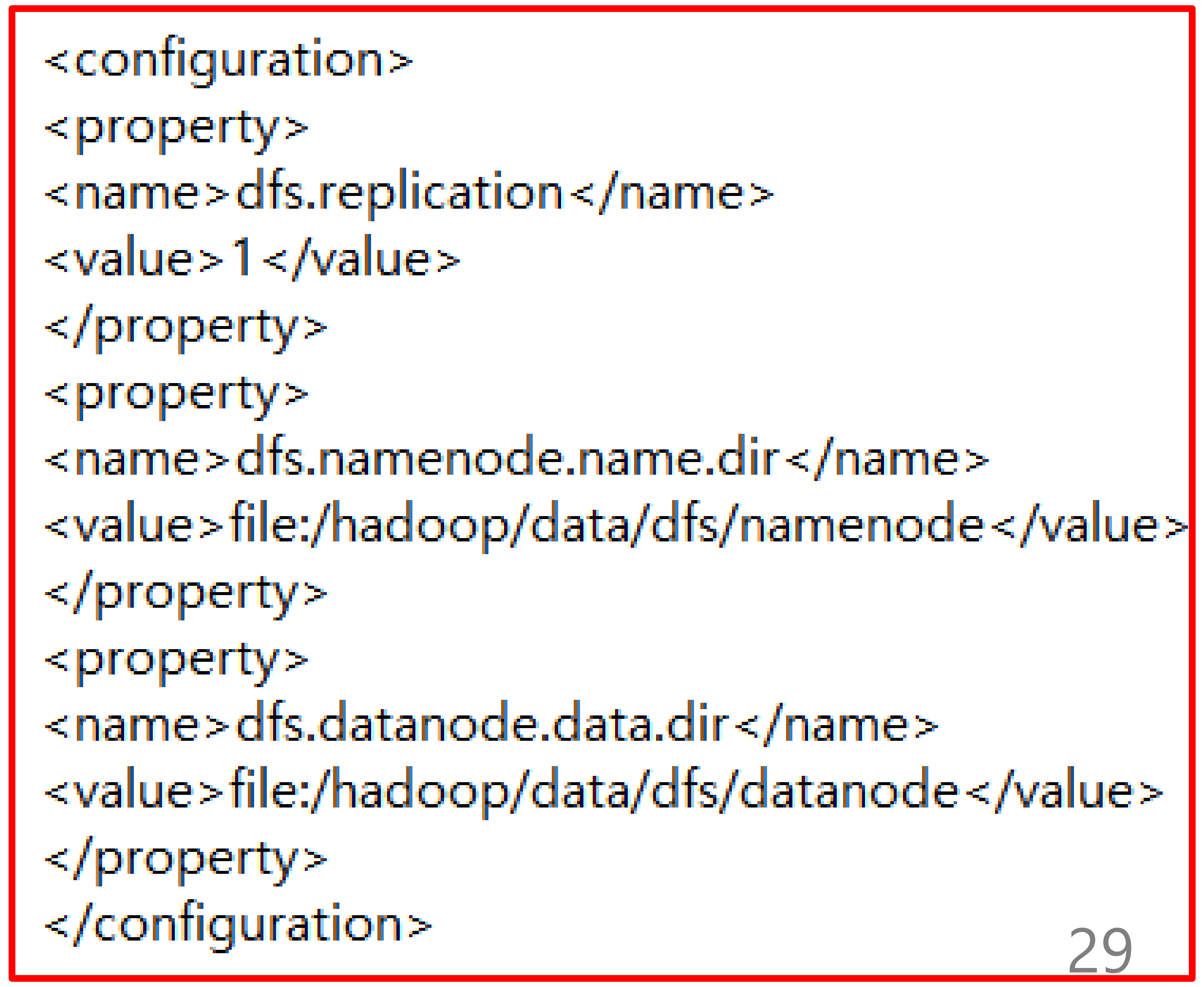
이름	수정한 날짜
hadoop-policy	2015-08-05 오
hdfs-site	2020-03-31 오
httpfs-env	2015-08-05 오



```
  See the License for the specific language governing p
  limitations under the License. See accompanying LICEI
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>

</configuration>
```

  - Edit CONFIGURATION like following

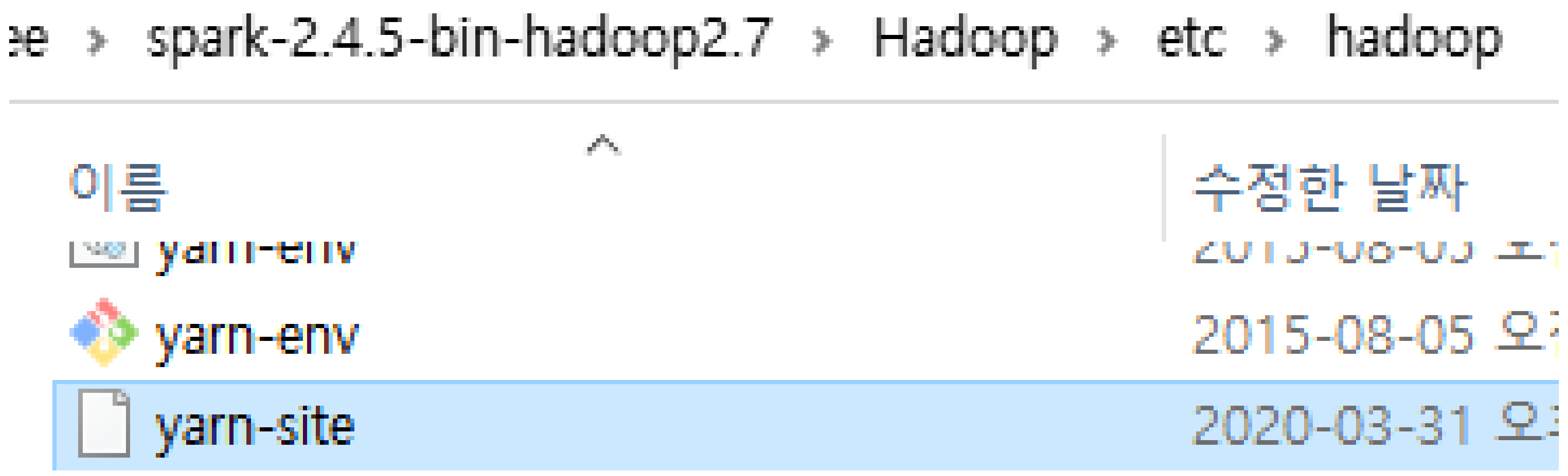
```
<configuration>
 <property> <name>dfs.replication</name> <value>1</value> </property>
 <property>
   <name>dfs.namenode.name.dir</name>  <value>file:/hadoop/data/dfs/namenode</value>
 </property>
 <property>
   <name>dfs.datanode.data.dir</name> <value>file:/hadoop/data/dfs/datanode</value>
 </property>
</configuration>
```

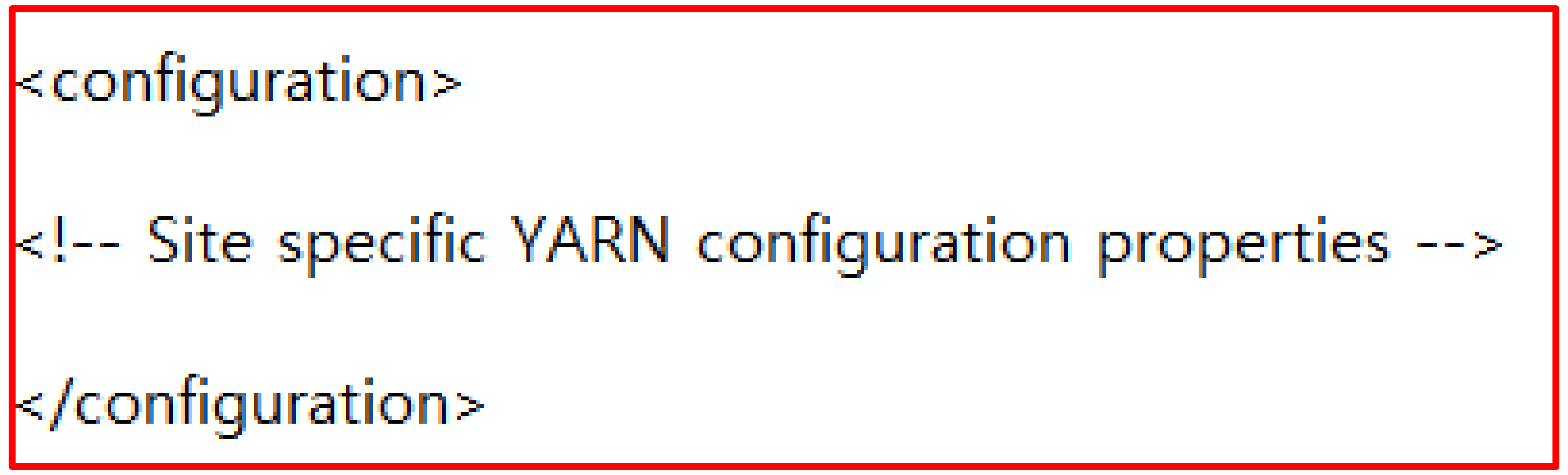


```
<configuration>
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
<property>
<name>dfs.namenode.name.dir</name>
<value>file:/hadoop/data/dfs/namenode</value>
</property>
<property>
<name>dfs.datanode.data.dir</name>
<value>file:/hadoop/data/dfs/datanode</value>
</property>
</configuration>
```

Hadoop Configuration

- Before start Hadoop file system, we must edit some configuration file
  - Open %HADOOP_HOME%\etc\Hadoop\yarn-site.xml

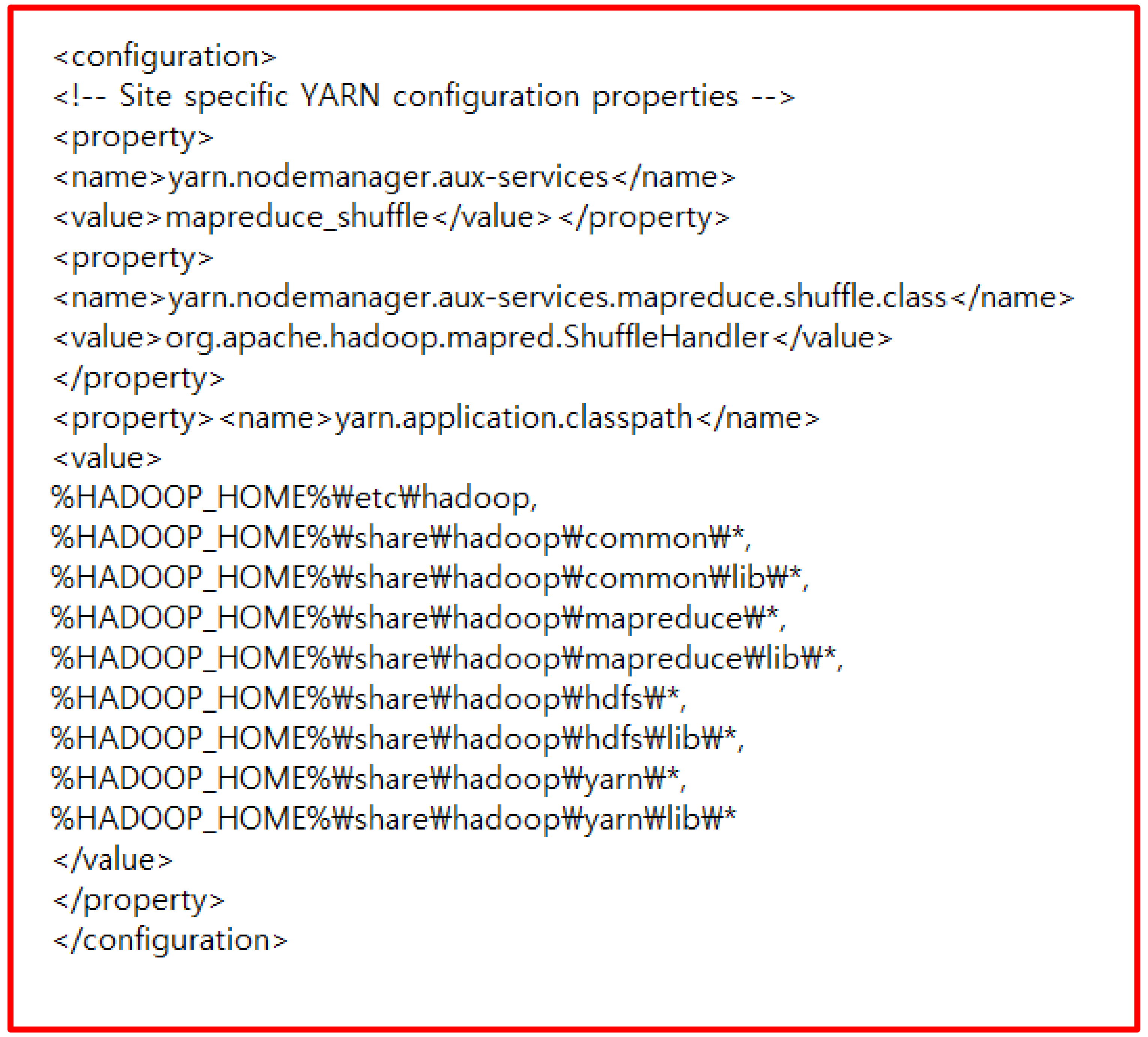




```
<configuration>

<!-- Site specific YARN configuration properties -->

</configuration>
```

  - Edit CONFIGURATION like following

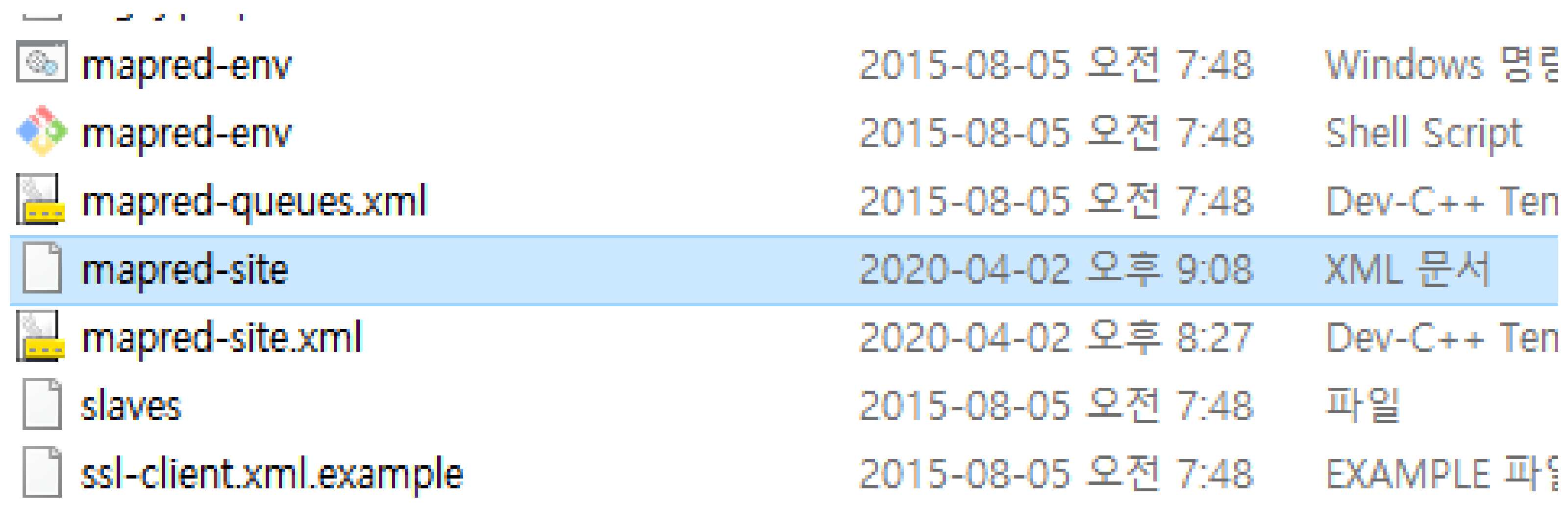
```
<configuration>
<!-- Site specific YARN configuration properties -->
<property><name>yarn.nodemanager.aux-services</name><value>mapreduce_shuffle</value></property>
<property>
<name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
<value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
<property><name>yarn.application.classpath</name>
<value>
%HADOOP_HOME%\etc\hadoop,
%HADOOP_HOME%\share\hadoop\common\*,
%HADOOP_HOME%\share\hadoop\common\lib\*,
%HADOOP_HOME%\share\hadoop\mapreduce\*,
%HADOOP_HOME%\share\hadoop\mapreduce\lib\*,
%HADOOP_HOME%\share\hadoop\hdfs\*,
%HADOOP_HOME%\share\hadoop\hdfs\lib\*,
%HADOOP_HOME%\share\hadoop\yarn\*,
%HADOOP_HOME%\share\hadoop\yarn\lib\*
</value></property></configuration>
```



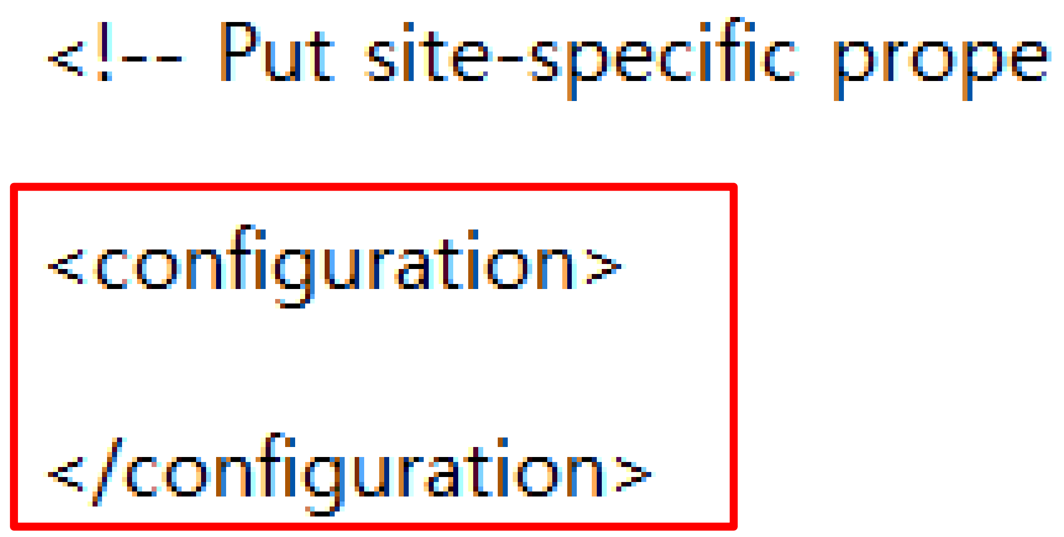
```
<configuration>
<!-- Site specific YARN configuration properties -->
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value> </property>
<property>
<name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
<value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
<property> <name>yarn.application.classpath</name>
<value>
%HADOOP_HOME%₩etc₩hadoop,
%HADOOP_HOME%₩share₩hadoop₩common₩*,
%HADOOP_HOME%₩share₩hadoop₩common₩lib₩*,
%HADOOP_HOME%₩share₩hadoop₩mapreduce₩*,
%HADOOP_HOME%₩share₩hadoop₩mapreduce₩lib₩*,
%HADOOP_HOME%₩share₩hadoop₩hdfs₩*,
%HADOOP_HOME%₩share₩hadoop₩hdfs₩lib₩*,
%HADOOP_HOME%₩share₩hadoop₩yarn₩*,
%HADOOP_HOME%₩share₩hadoop₩yarn₩lib₩*
</value>
</property>
</configuration>
```

Hadoop Configuration

- ➢ **Before start Hadoop file system, we must edit some configuration file**
  - **Open %HADOOP_HOME%\etc\Hadoop\mapred-site.xml.template**
  - **Make new file “mapred-site.xml” and copy the contents of mapred-site.xml.template**



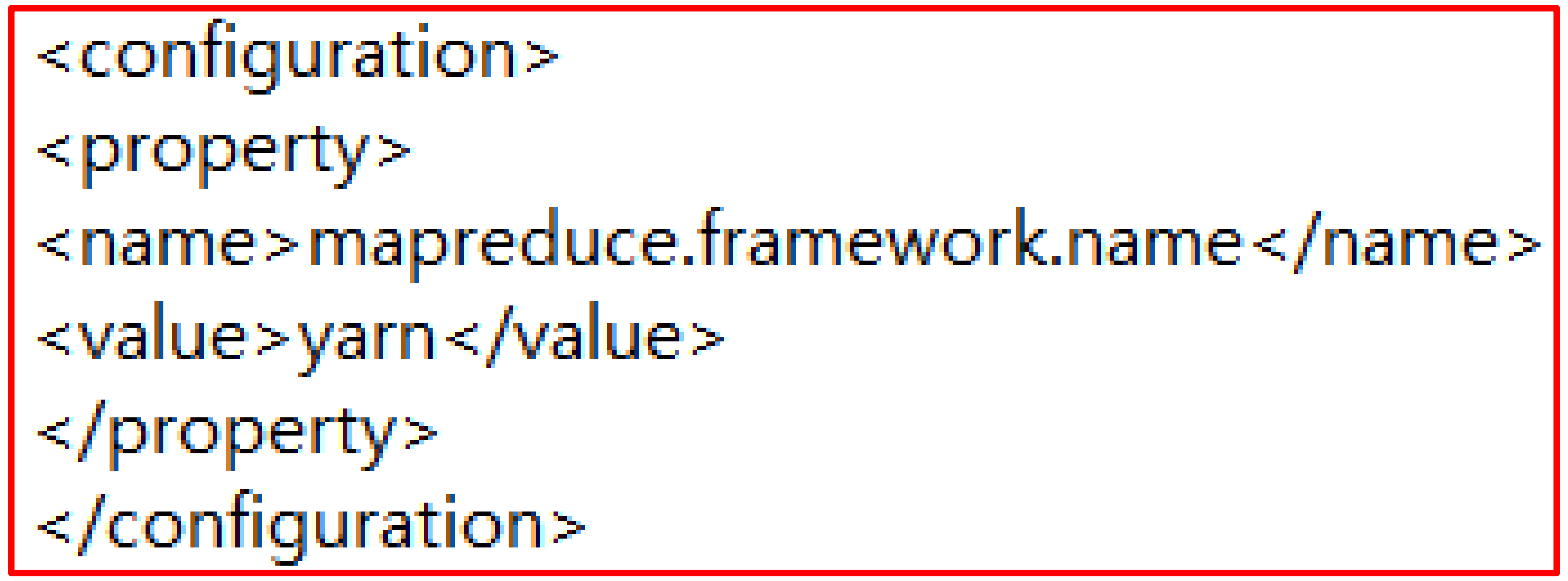
mapred-env	2015-08-05 오전 7:48	Windows 명령
mapred-env	2015-08-05 오전 7:48	Shell Script
mapred-queues.xml	2015-08-05 오전 7:48	Dev-C++ Ten
mapred-site	2020-04-02 오후 9:08	XML 문서
mapred-site.xml	2020-04-02 오후 8:27	Dev-C++ Ten
slaves	2015-08-05 오전 7:48	파일
ssl-client.xml.example	2015-08-05 오전 7:48	EXAMPLE 파일



```
<!-- Put site-specific prope

<configuration>

</configuration>
```

- **Edit CONFIGURATION in mapred-site.xml file like following**

<configuration>
<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
</property>
</configuration>



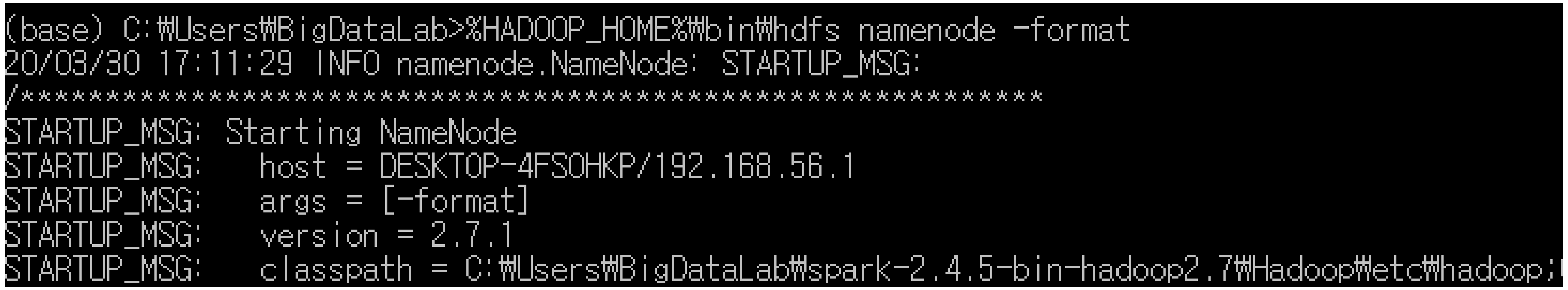
```
<configuration>
<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
</property>
</configuration>
```

Hadoop Configuration

- At last, we can use Hadoop file system
- If you want to Format Namenode, then input like following

%HADOOP_HOME%\bin\hdfs namenode -format

- For example,



```
(base) C:₩Users₩BigDataLab>%HADOOP_HOME%₩bin₩hdfs namenode -format
20/03/30 17:11:29 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = DESKTOP-4FSOHKP/192.168.56.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 2.7.1
STARTUP_MSG:   classpath = C:₩Users₩BigDataLab₩spark-2.4.5-bin-hadoop2.7₩Hadoop₩etc₩hadoop;
```

Start Hadoop File System

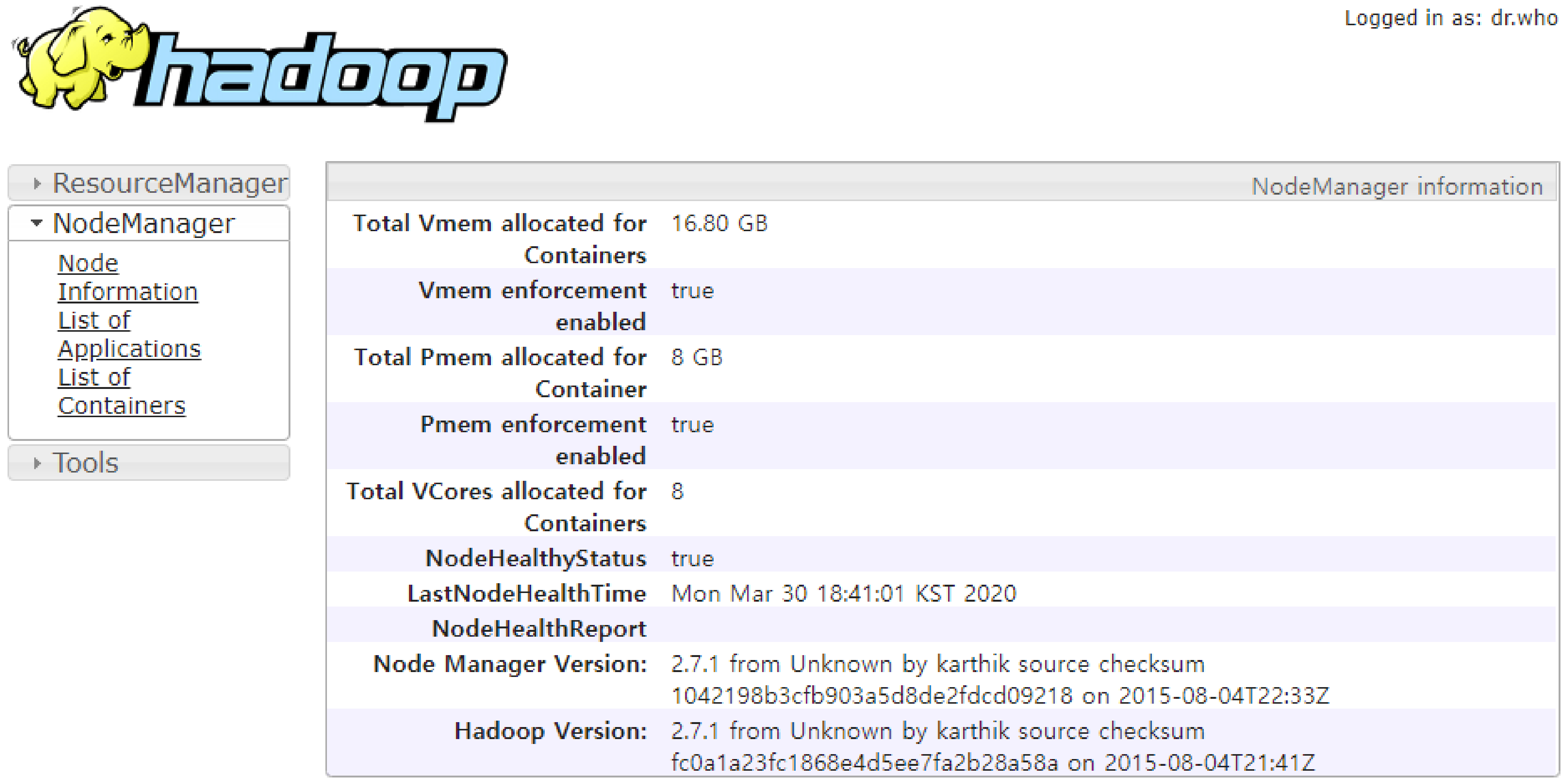
- Start DFS and YARN with the following command

```
%HADOOP_HOME%\sbin\start-dfs.cmd
%HADOOP_HOME%\sbin\start-yarn.cmd
```

- A few seconds later, enter the following links

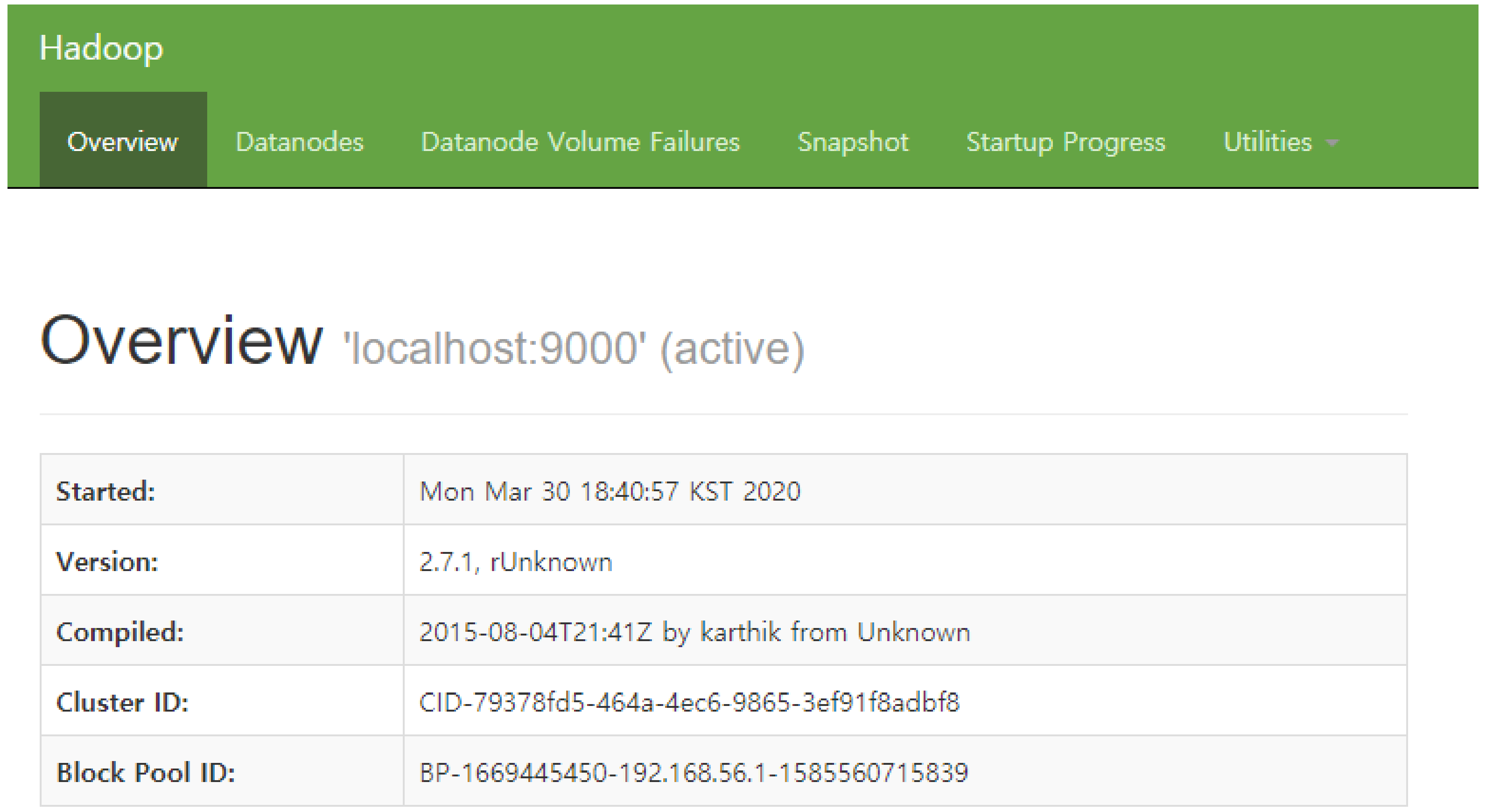
  http://localhost:8088 & http://localhost:50070

- Then you can see Hadoop websites.



NodeManager information

Total Vmem allocated for Containers	16.80 GB
Vmem enforcement enabled	true
Total Pmem allocated for Container	8 GB
Pmem enforcement enabled	true
Total VCores allocated for Containers	8
NodeHealthyStatus	true
LastNodeHealthTime	Mon Mar 30 18:41:01 KST 2020
NodeHealthReport	
Node Manager Version:	2.7.1 from Unknown by karthik source checksum 1042198b3cfb903a5d8de2fdcd09218 on 2015-08-04T22:33Z
Hadoop Version:	2.7.1 from Unknown by karthik source checksum fc0a1a23fc1868e4d5ee7fa2b28a58a on 2015-08-04T21:41Z



Overview 'localhost:9000' (active)

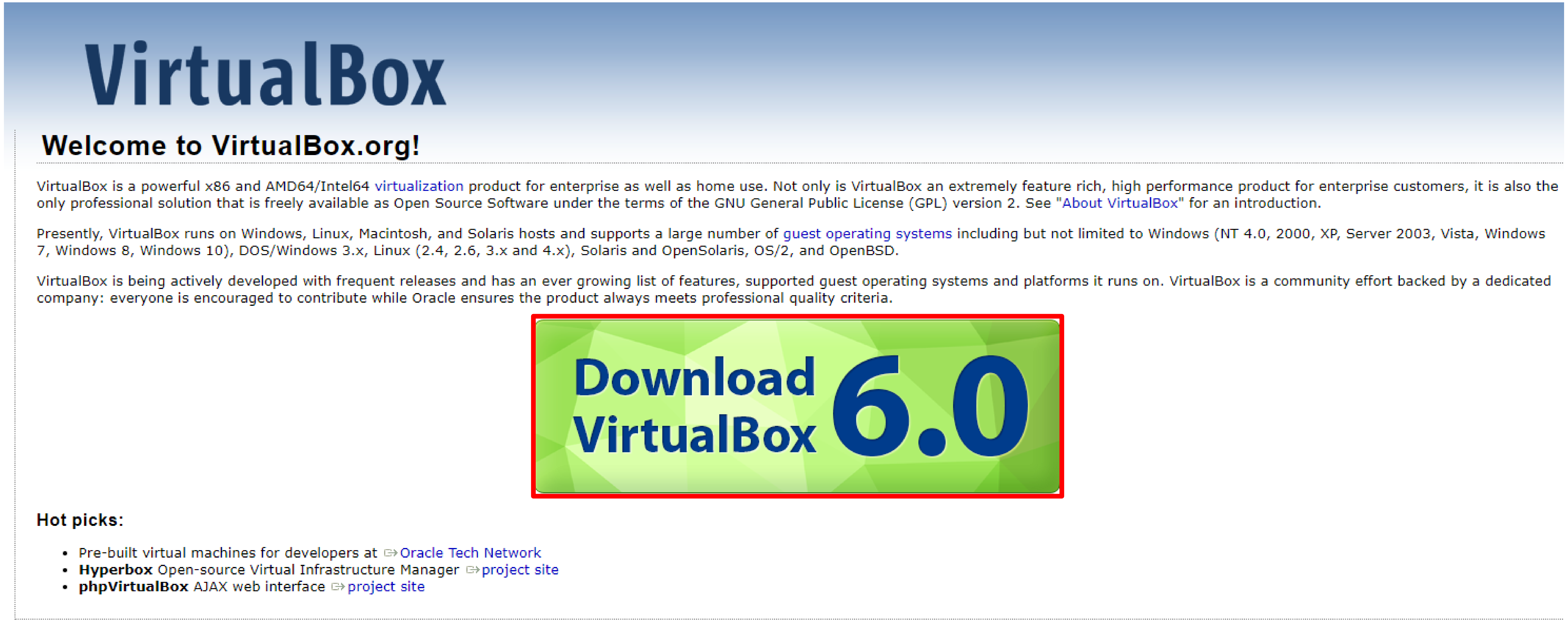
Started:	Mon Mar 30 18:40:57 KST 2020
Version:	2.7.1, rUnknown
Compiled:	2015-08-04T21:41Z by karthik from Unknown
Cluster ID:	CID-79378fd5-464a-4ec6-9865-3ef91f8adbf8
Block Pool ID:	BP-1669445450-192.168.56.1-1585560715839

Using Linux

Virtual box installation

Virtual Box

- VirtualBox is a powerful virtualization product.



VirtualBox

Welcome to VirtualBox.org!

VirtualBox is a powerful x86 and AMD64/Intel64 virtualization product for enterprise as well as home use. Not only is VirtualBox an extremely feature rich, high performance product for enterprise customers, it is also the only professional solution that is freely available as Open Source Software under the terms of the GNU General Public License (GPL) version 2. See "About VirtualBox" for an introduction.

Presently, VirtualBox runs on Windows, Linux, Macintosh, and Solaris hosts and supports a large number of guest operating systems including but not limited to Windows (NT 4.0, 2000, XP, Server 2003, Vista, Windows 7, Windows 8, Windows 10), DOS/Windows 3.x, Linux (2.4, 2.6, 3.x and 4.x), Solaris and OpenSolaris, OS/2, and OpenBSD.

VirtualBox is being actively developed with frequent releases and has an ever growing list of features, supported guest operating systems and platforms it runs on. VirtualBox is a community effort backed by a dedicated company: everyone is encouraged to contribute while Oracle ensures the product always meets professional quality criteria.

Hot picks:

- Pre-built virtual machines for developers at ⇨Oracle Tech Network
- **Hyperbox** Open-source Virtual Infrastructure Manager ⇨project site
- **phpVirtualBox** AJAX web interface ⇨project site

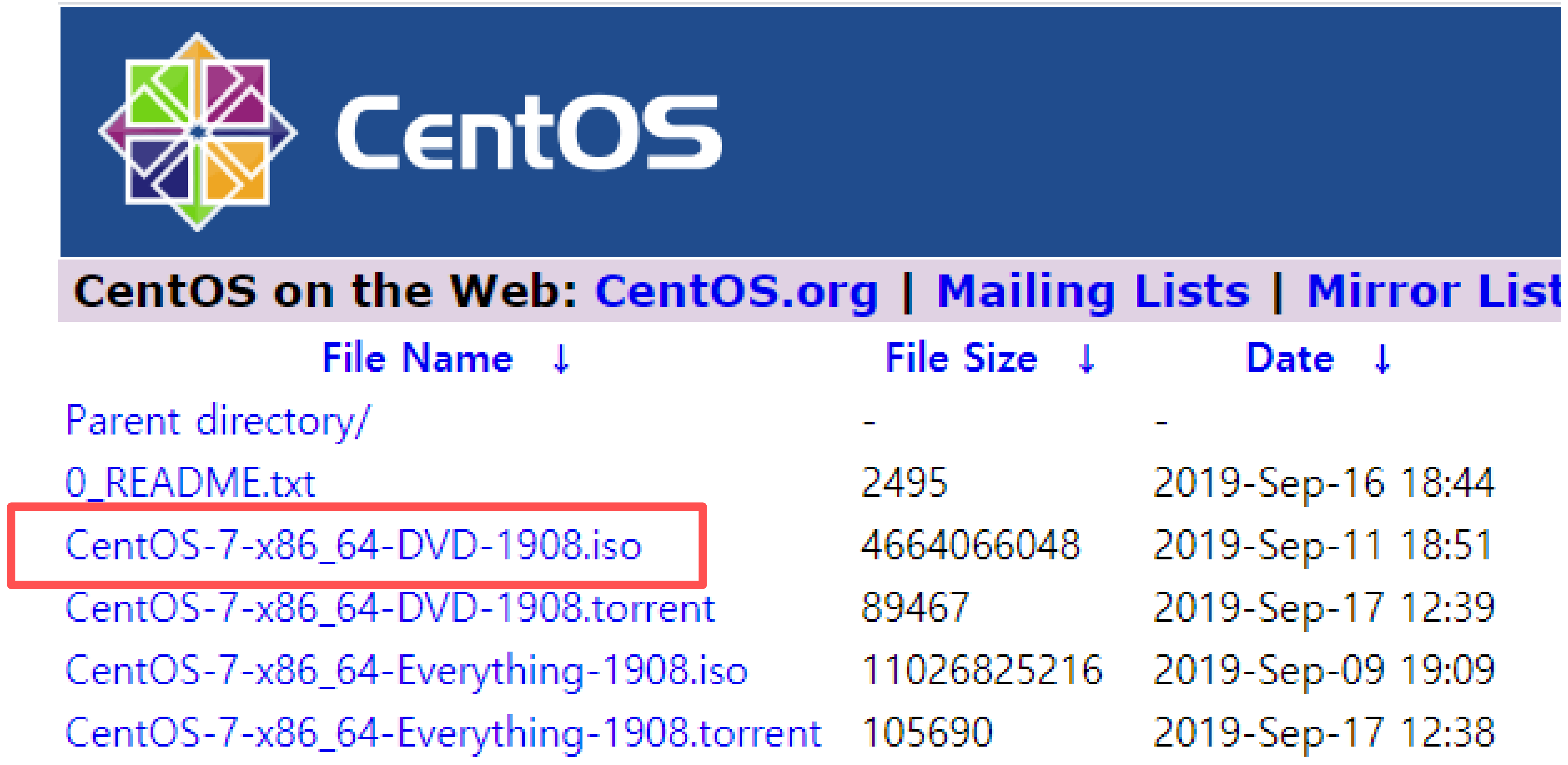
https://www.virtualbox.org/

CentOS 7

- Download CentOS 7 with the following link:

http://mirror.kakao.com/centos/7.7.1908/isos/x86_64/

- Download CentOS-x86_64-DVD-1908.iso



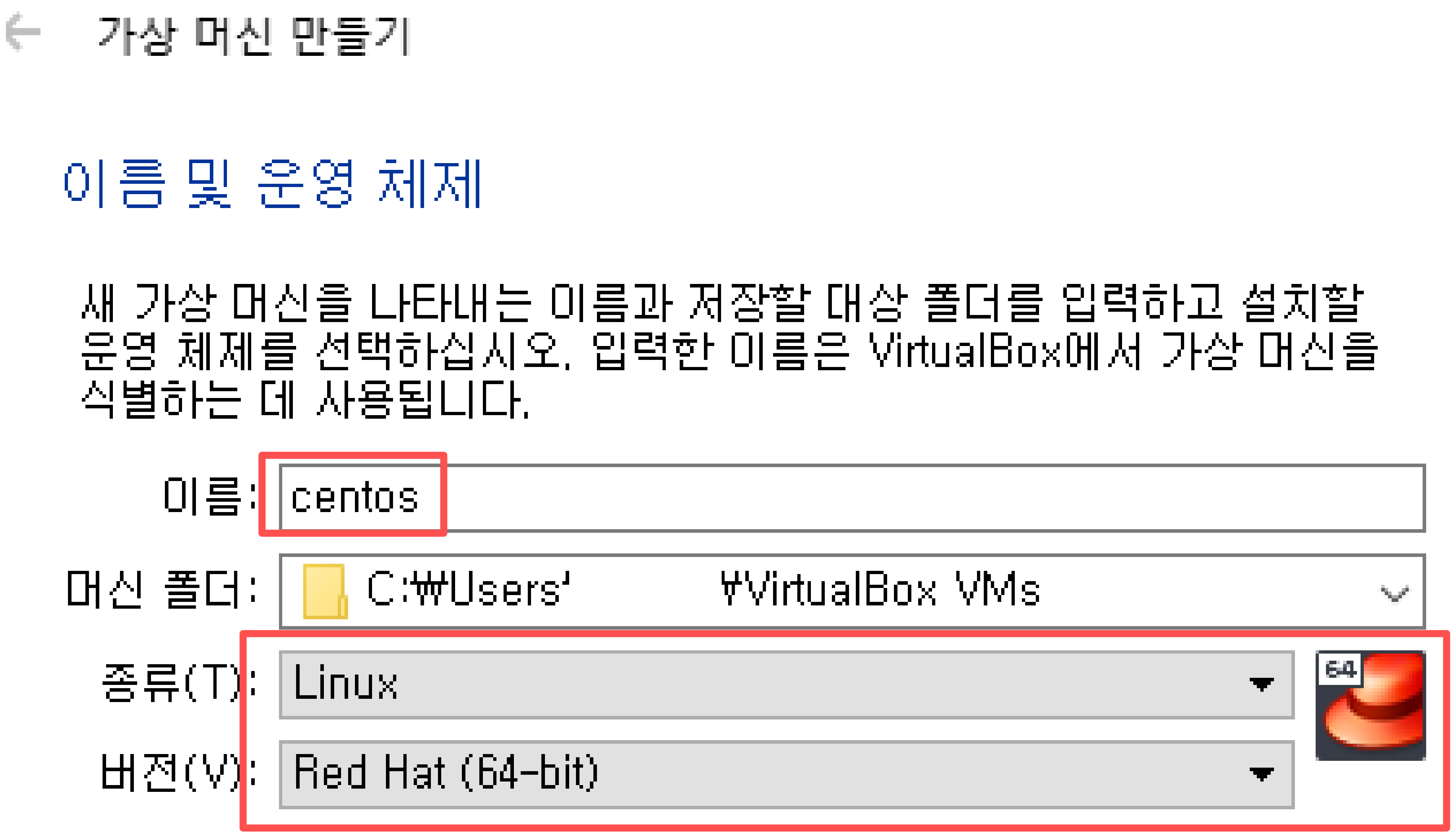
CentOS on the Web: CentOS.org | Mailing Lists | Mirror List

File Name ↓	File Size ↓	Date ↓
Parent directory/	-	-
0_README.txt	2495	2019-Sep-16 18:44
CentOS-7-x86_64-DVD-1908.iso	4664066048	2019-Sep-11 18:51
CentOS-7-x86_64-DVD-1908.torrent	89467	2019-Sep-17 12:39
CentOS-7-x86_64-Everything-1908.iso	11026825216	2019-Sep-09 19:09
CentOS-7-x86_64-Everything-1908.torrent	105690	2019-Sep-17 12:38

http://mirror.kakao.com/centos/7.7.1908/isos/x86_64/

CentOS 7

➢ **Choose operation system in detail**



가상 머신 만들기

이름 및 운영 체제

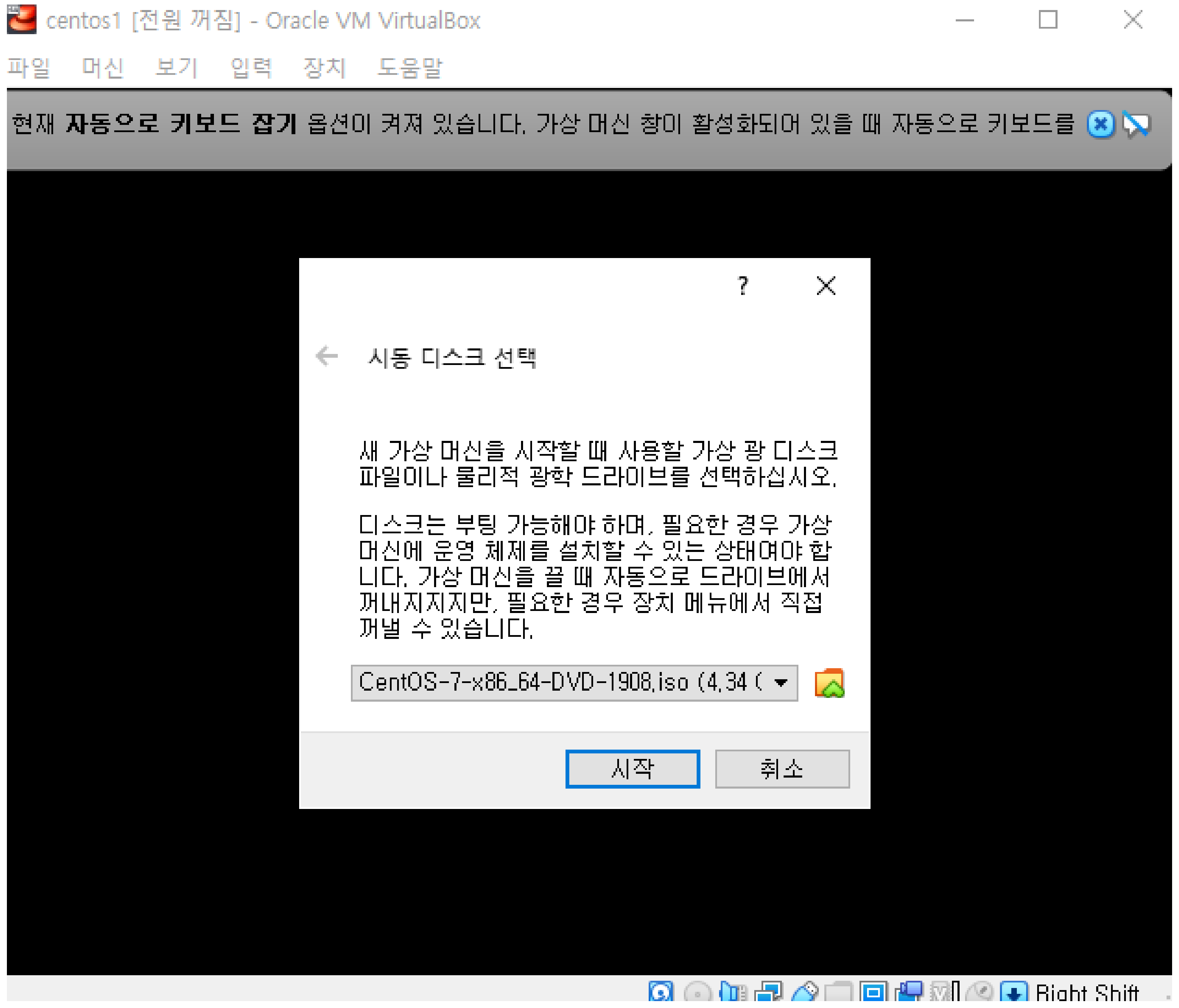
새 가상 머신을 나타내는 이름과 저장할 대상 폴더를 입력하고 설치할 운영 체제를 선택하십시오. 입력한 이름은 VirtualBox에서 가상 머신을 식별하는 데 사용됩니다.

이름: centos

머신 폴더: C:₩Users' ¥VirtualBox VMs

종류(T): Linux

버전(V): Red Hat (64-bit)



현재 **자동으로 키보드 잡기** 옵션이 켜져 있습니다. 가상 머신 창이 활성화되어 있을 때 자동으로 키보드를

시동 디스크 선택

새 가상 머신을 시작할 때 사용할 가상 광 디스크 파일이나 물리적 광학 드라이브를 선택하십시오.

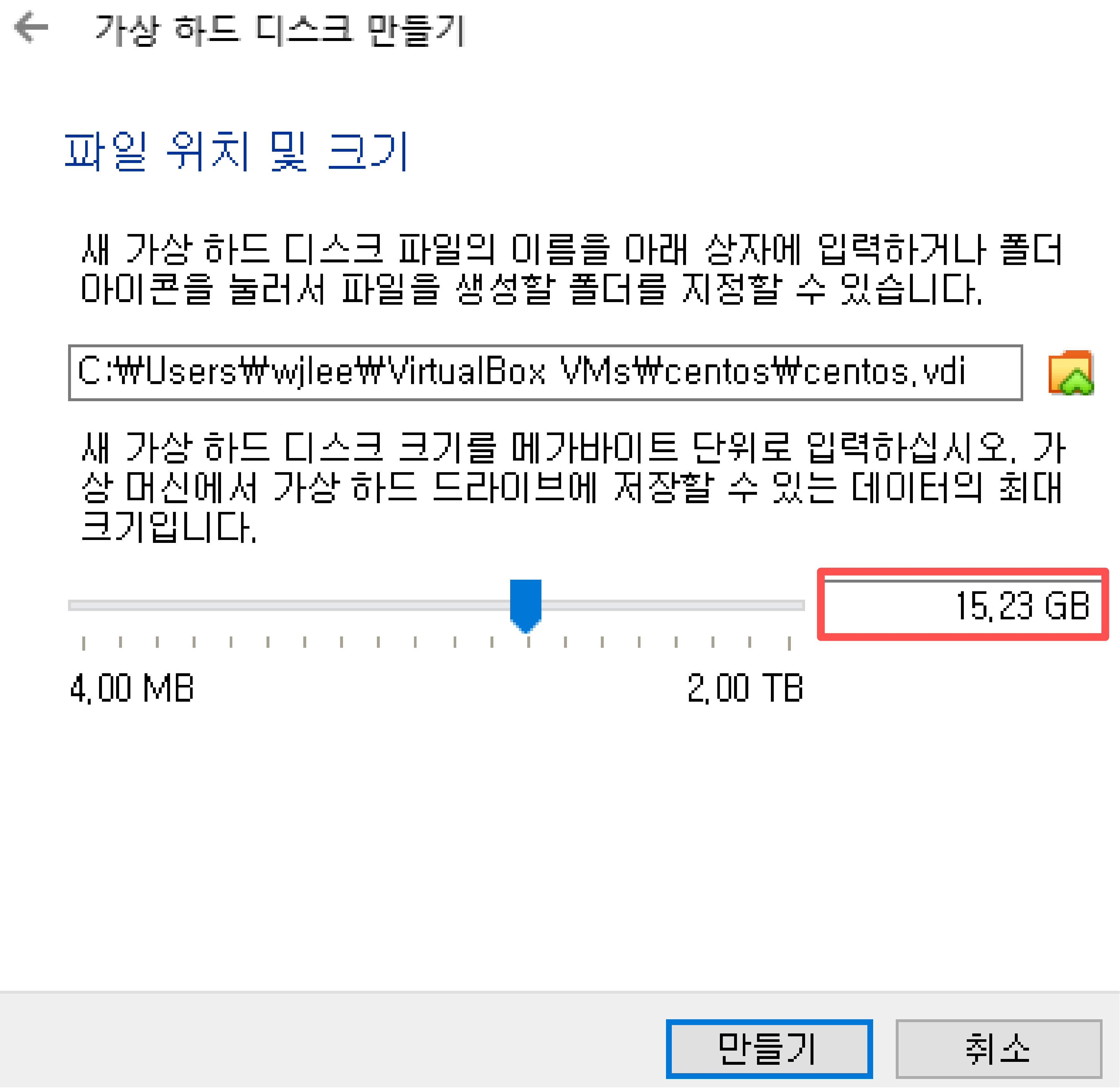
디스크는 부팅 가능해야 하며, 필요한 경우 가상 머신에 운영 체제를 설치할 수 있는 상태여야 합니다. 가상 머신을 끌 때 자동으로 드라이브에서 꺼내지지만, 필요한 경우 장치 메뉴에서 직접 꺼낼 수 있습니다.

CentOS-7-x86_64-DVD-1908.iso (4.34 (

CentOS 7

➢ Minimum volume of the hard disk

- You have to make volume of the hard disk more than 15 GB.
- Unless, you are able to get out of disk.



← 가상 하드 디스크 만들기

파일 위치 및 크기

새 가상 하드 디스크 파일의 이름을 아래 상자에 입력하거나 폴더 아이콘을 눌러서 파일을 생성할 폴더를 지정할 수 있습니다.

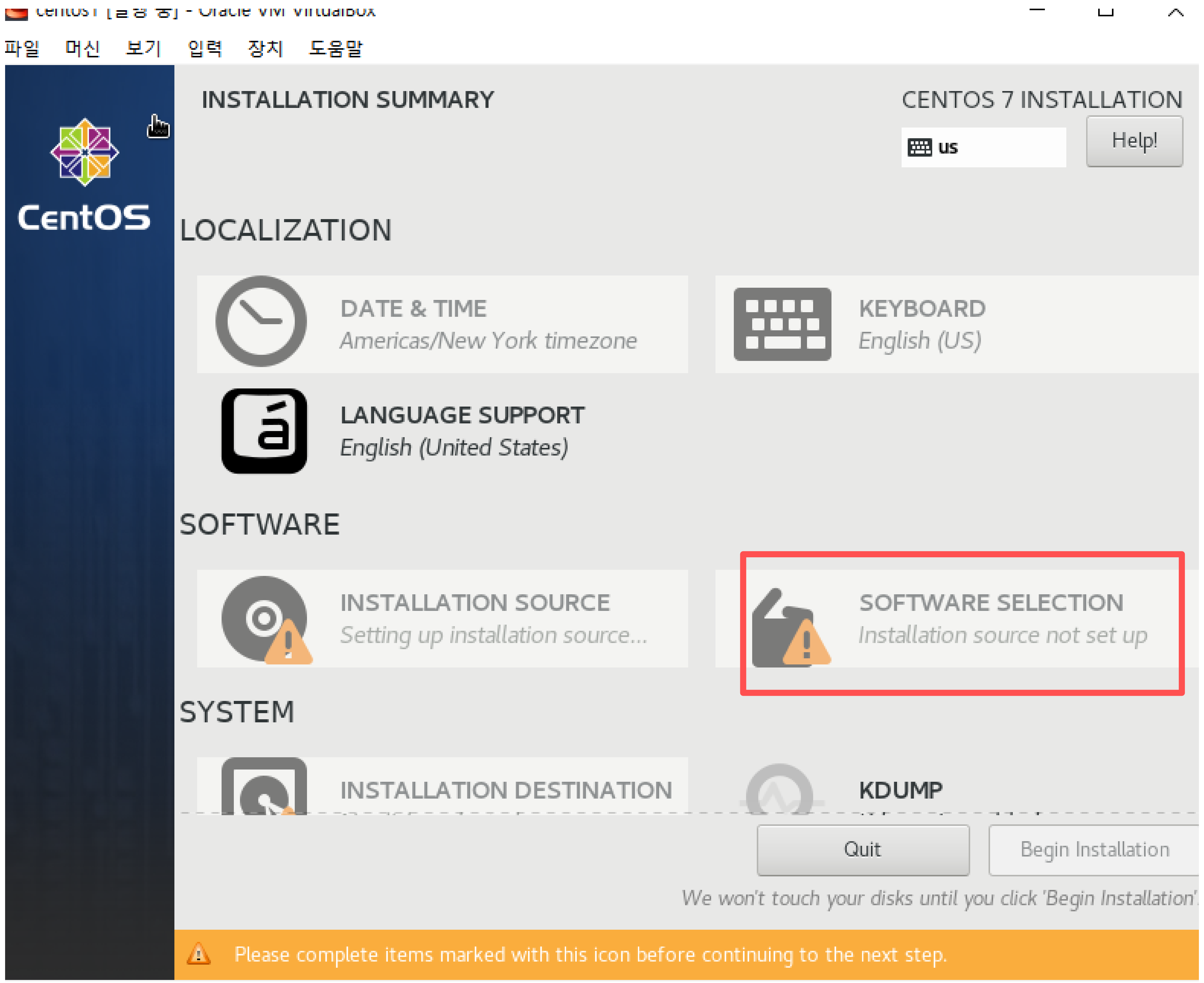
C:₩Users₩wjlee₩VirtualBox VMs₩centos₩centos.vdi

새 가상 하드 디스크 크기를 메가바이트 단위로 입력하십시오. 가상 머신에서 가상 하드 드라이브에 저장할 수 있는 데이터의 최대 크기입니다.

15.23 GB

CentOS 7

- ➢ **CentOS installation process**



INSTALLATION SUMMARY

CENTOS 7 INSTALLATION

LOCALIZATION

DATE & TIME
Americas/New York timezone

KEYBOARD
English (US)

LANGUAGE SUPPORT
English (United States)

SOFTWARE

INSTALLATION SOURCE
Setting up installation source...

SOFTWARE SELECTION
Installation source not set up

SYSTEM

INSTALLATION DESTINATION

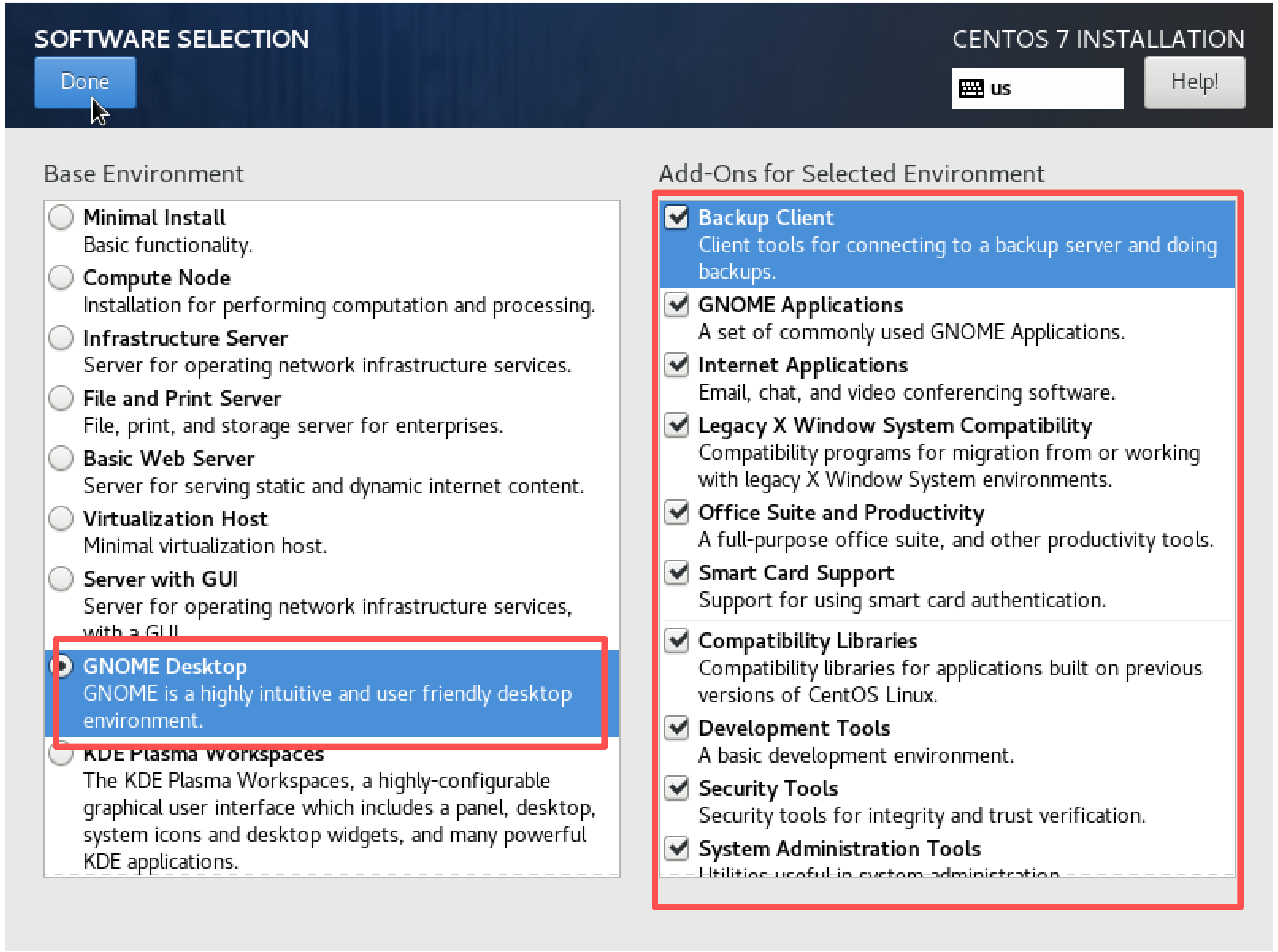
KDUMP

We won't touch your disks until you click 'Begin Installation'.

⚠ Please complete items marked with this icon before continuing to the next step.

CentOS 7

➢ CentOS installation process



Base Environment

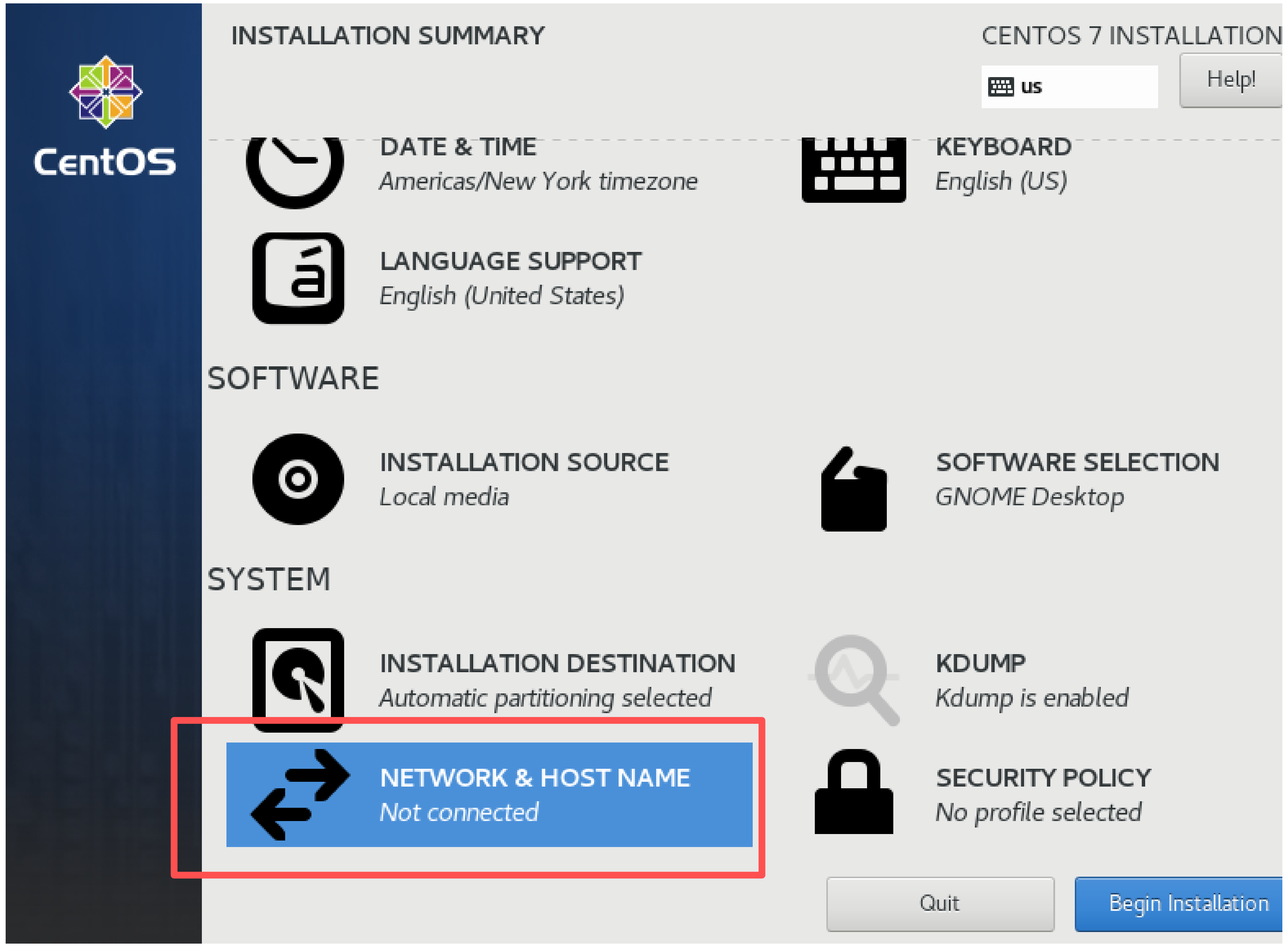
- ○ **Minimal Install**
  Basic functionality.
- ○ **Compute Node**
  Installation for performing computation and processing.
- ○ **Infrastructure Server**
  Server for operating network infrastructure services.
- ○ **File and Print Server**
  File, print, and storage server for enterprises.
- ○ **Basic Web Server**
  Server for serving static and dynamic internet content.
- ○ **Virtualization Host**
  Minimal virtualization host.
- ○ **Server with GUI**
  Server for operating network infrastructure services, with a GUI.
- ● **GNOME Desktop**
  GNOME is a highly intuitive and user friendly desktop environment.
- ○ **KDE Plasma Workspaces**
  The KDE Plasma Workspaces, a highly-configurable graphical user interface which includes a panel, desktop, system icons and desktop widgets, and many powerful KDE applications.

Add-Ons for Selected Environment

- ☑ **Backup Client**
  Client tools for connecting to a backup server and doing backups.
- ☑ **GNOME Applications**
  A set of commonly used GNOME Applications.
- ☑ **Internet Applications**
  Email, chat, and video conferencing software.
- ☑ **Legacy X Window System Compatibility**
  Compatibility programs for migration from or working with legacy X Window System environments.
- ☑ **Office Suite and Productivity**
  A full-purpose office suite, and other productivity tools.
- ☑ **Smart Card Support**
  Support for using smart card authentication.
- ☑ **Compatibility Libraries**
  Compatibility libraries for applications built on previous versions of CentOS Linux.
- ☑ **Development Tools**
  A basic development environment.
- ☑ **Security Tools**
  Security tools for integrity and trust verification.
- ☑ **System Administration Tools**
  Utilities useful in system administration.

CentOS 7

➢ CentOS installation process



INSTALLATION SUMMARY

CENTOS 7 INSTALLATION

DATE & TIME
Americas/New York timezone

KEYBOARD
English (US)

LANGUAGE SUPPORT
English (United States)

SOFTWARE

INSTALLATION SOURCE
Local media

SOFTWARE SELECTION
GNOME Desktop

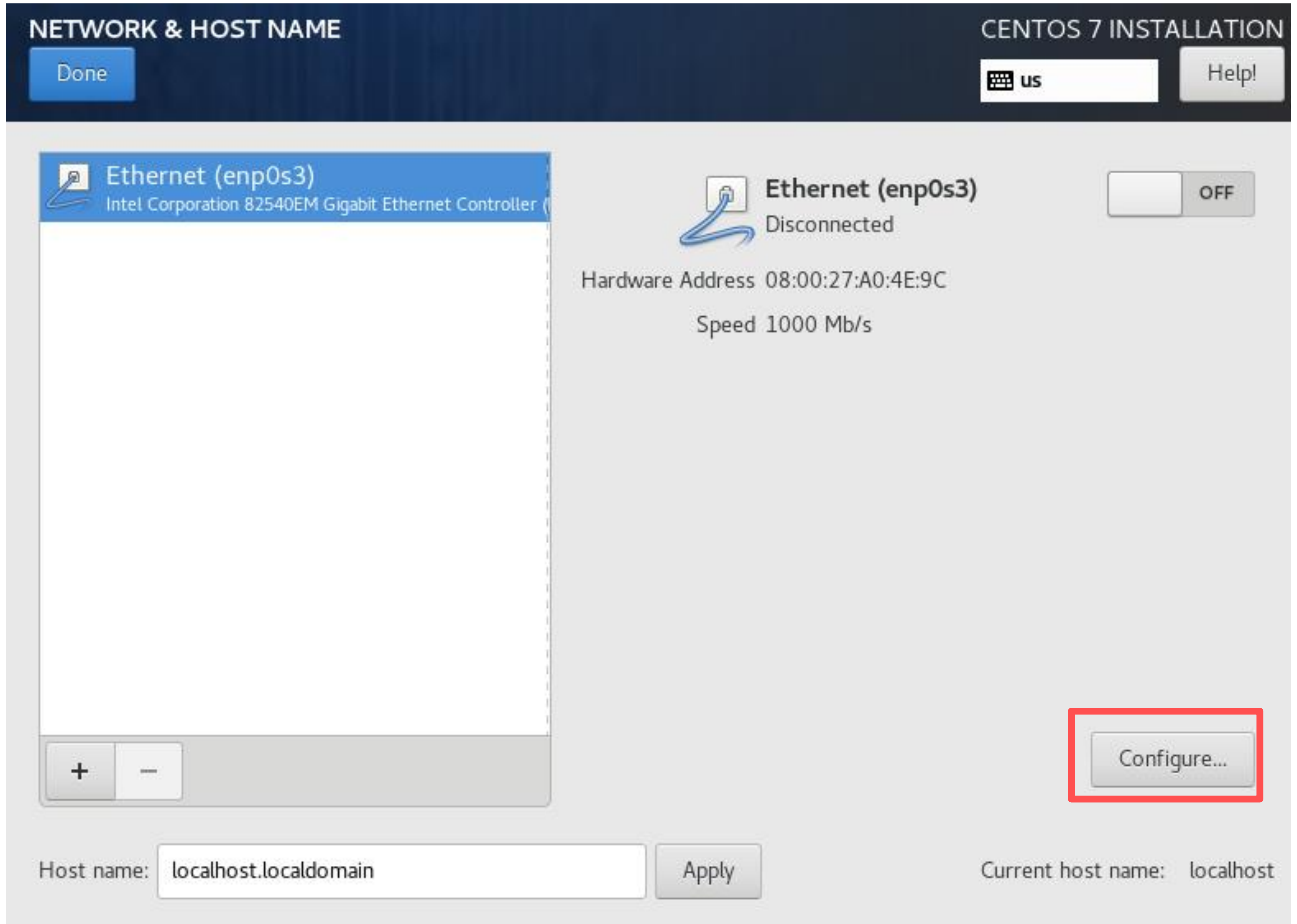
SYSTEM

INSTALLATION DESTINATION
Automatic partitioning selected

KDUMP
Kdump is enabled

NETWORK & HOST NAME
Not connected

SECURITY POLICY
No profile selected



NETWORK & HOST NAME

CENTOS 7 INSTALLATION

Ethernet (enp0s3)
Intel Corporation 82540EM Gigabit Ethernet Controller (

Ethernet (enp0s3)
Disconnected

Hardware Address 08:00:27:A0:4E:9C

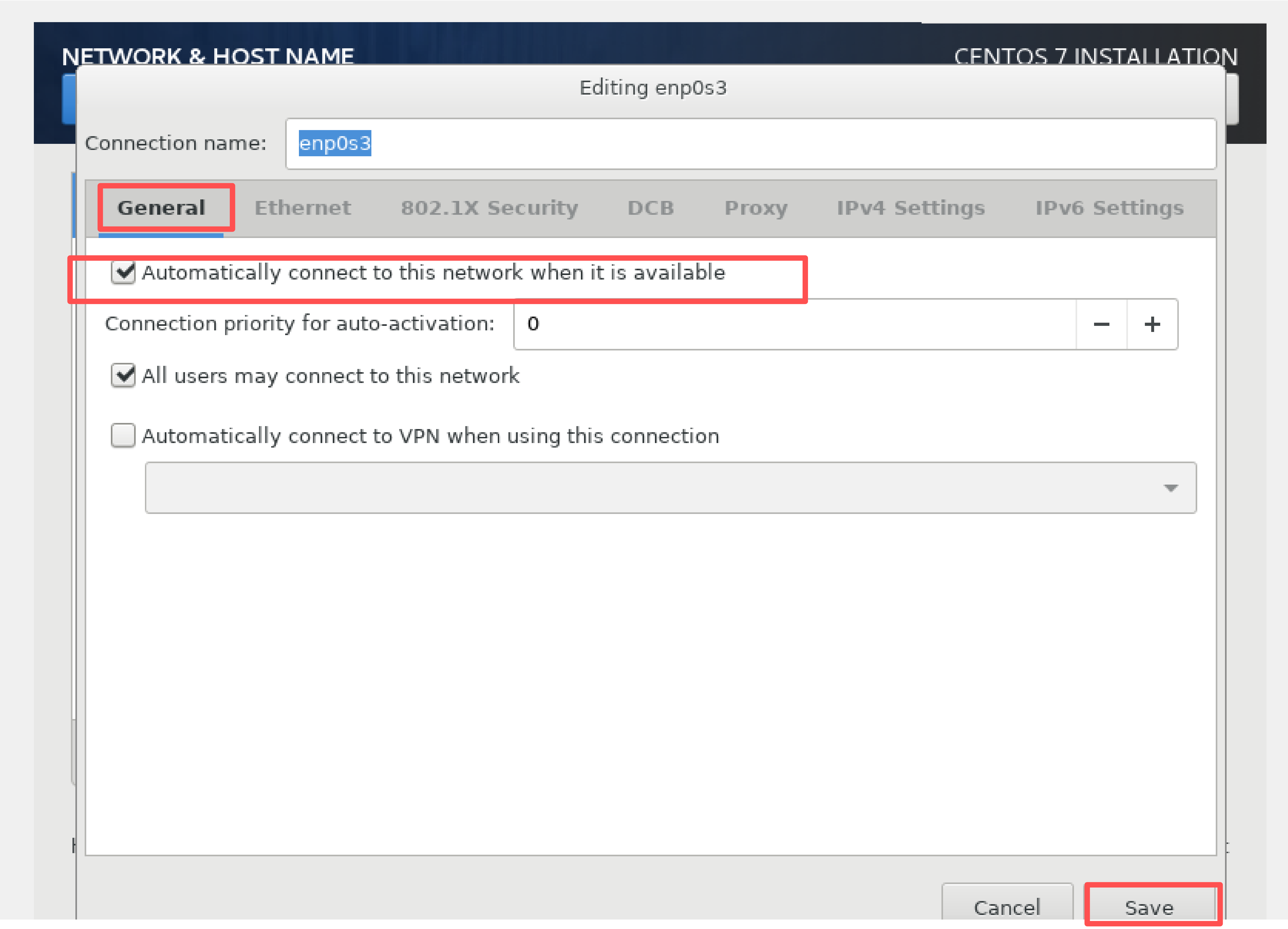
Speed 1000 Mb/s

Host name: localhost.localdomain

Current host name: localhost

CentOS 7

➢ CentOS installation process



Connection name: enp0s3

☑ Automatically connect to this network when it is available

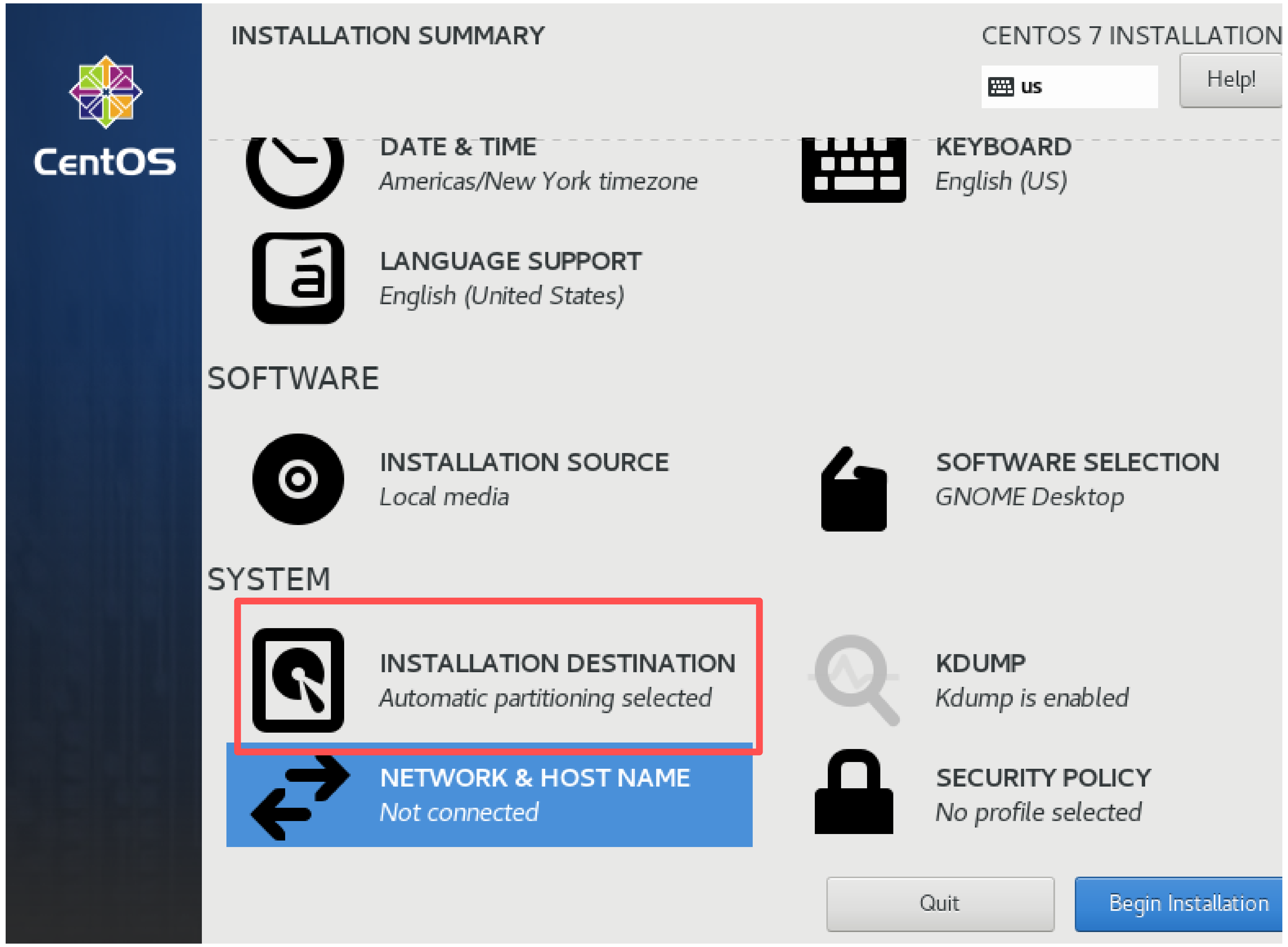
Connection priority for auto-activation: 0

☑ All users may connect to this network

□ Automatically connect to VPN when using this connection

CentOS 7

➢ CentOS installation process



INSTALLATION SUMMARY

CENTOS 7 INSTALLATION

DATE & TIME
Americas/New York timezone

KEYBOARD
English (US)

LANGUAGE SUPPORT
English (United States)

SOFTWARE

INSTALLATION SOURCE
Local media

SOFTWARE SELECTION
GNOME Desktop

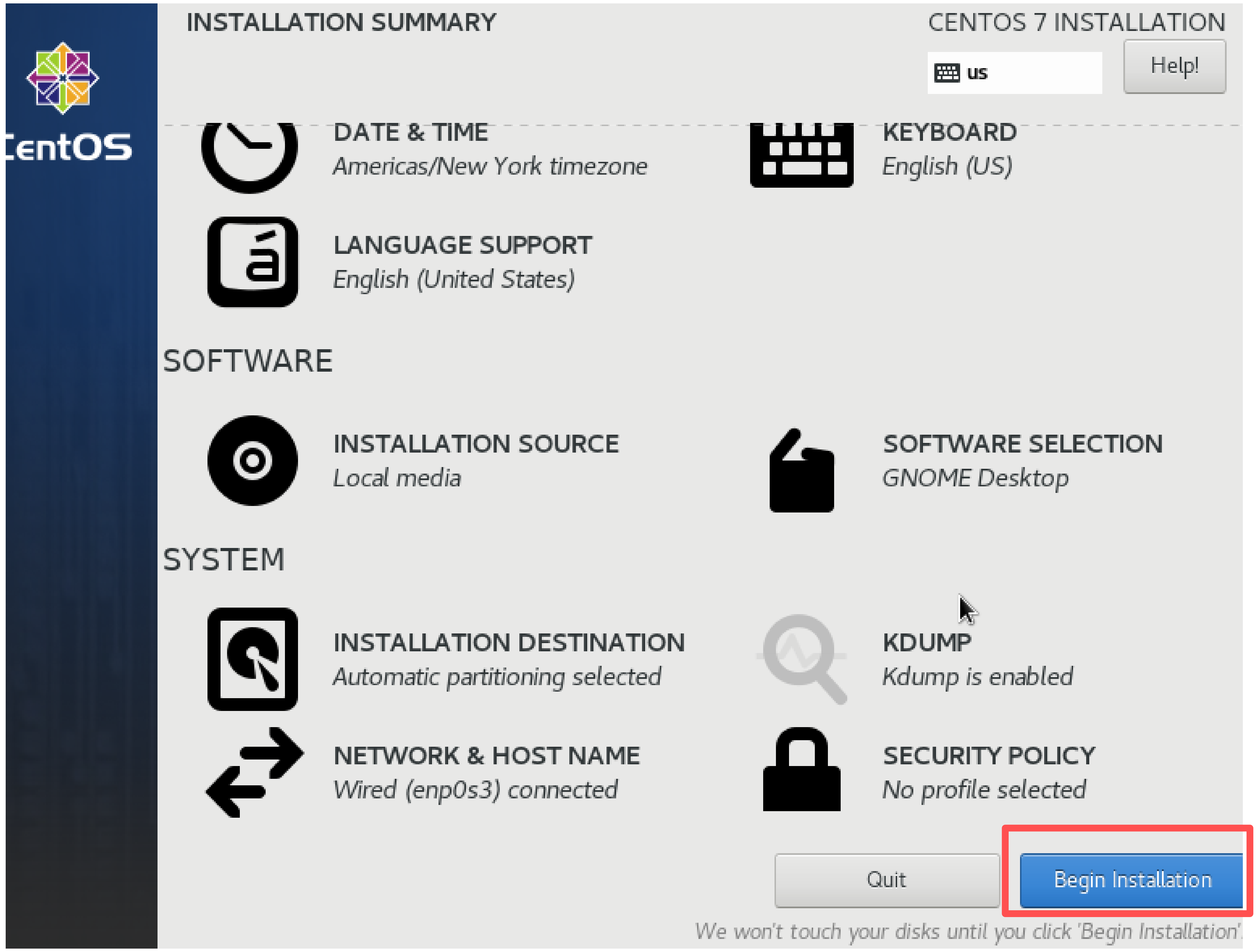
SYSTEM

INSTALLATION DESTINATION
Automatic partitioning selected

KDUMP
Kdump is enabled

NETWORK & HOST NAME
Not connected

SECURITY POLICY
No profile selected



INSTALLATION SUMMARY

CENTOS 7 INSTALLATION

DATE & TIME
Americas/New York timezone

KEYBOARD
English (US)

LANGUAGE SUPPORT
English (United States)

SOFTWARE

INSTALLATION SOURCE
Local media

SOFTWARE SELECTION
GNOME Desktop

SYSTEM

INSTALLATION DESTINATION
Automatic partitioning selected

KDUMP
Kdump is enabled

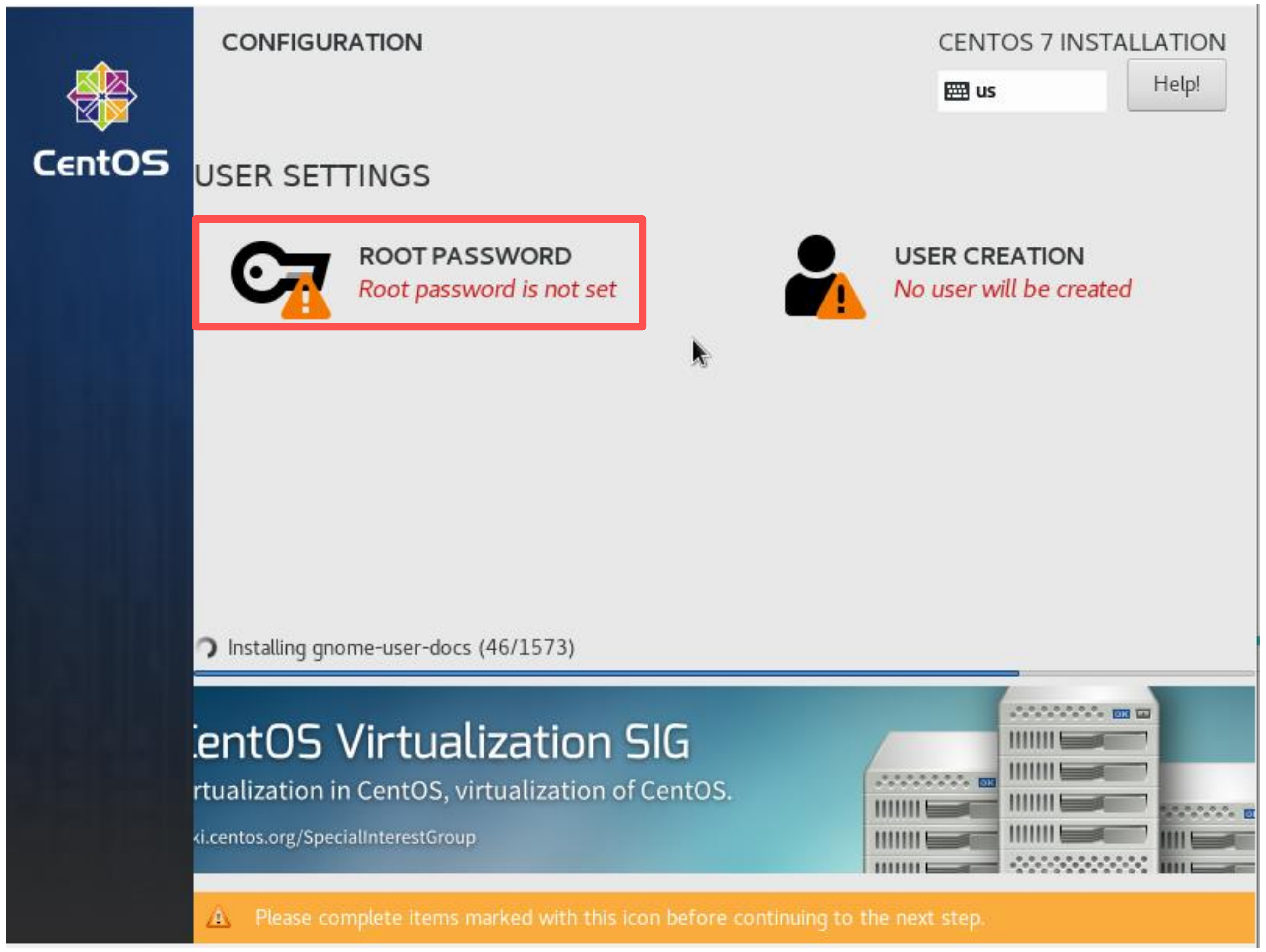
NETWORK & HOST NAME
Wired (enp0s3) connected

SECURITY POLICY
No profile selected

We won't touch your disks until you click 'Begin Installation'

CentOS 7

➢ CentOS installation process



CentOS
CONFIGURATION
CENTOS 7 INSTALLATION
us
Help!

USER SETTINGS

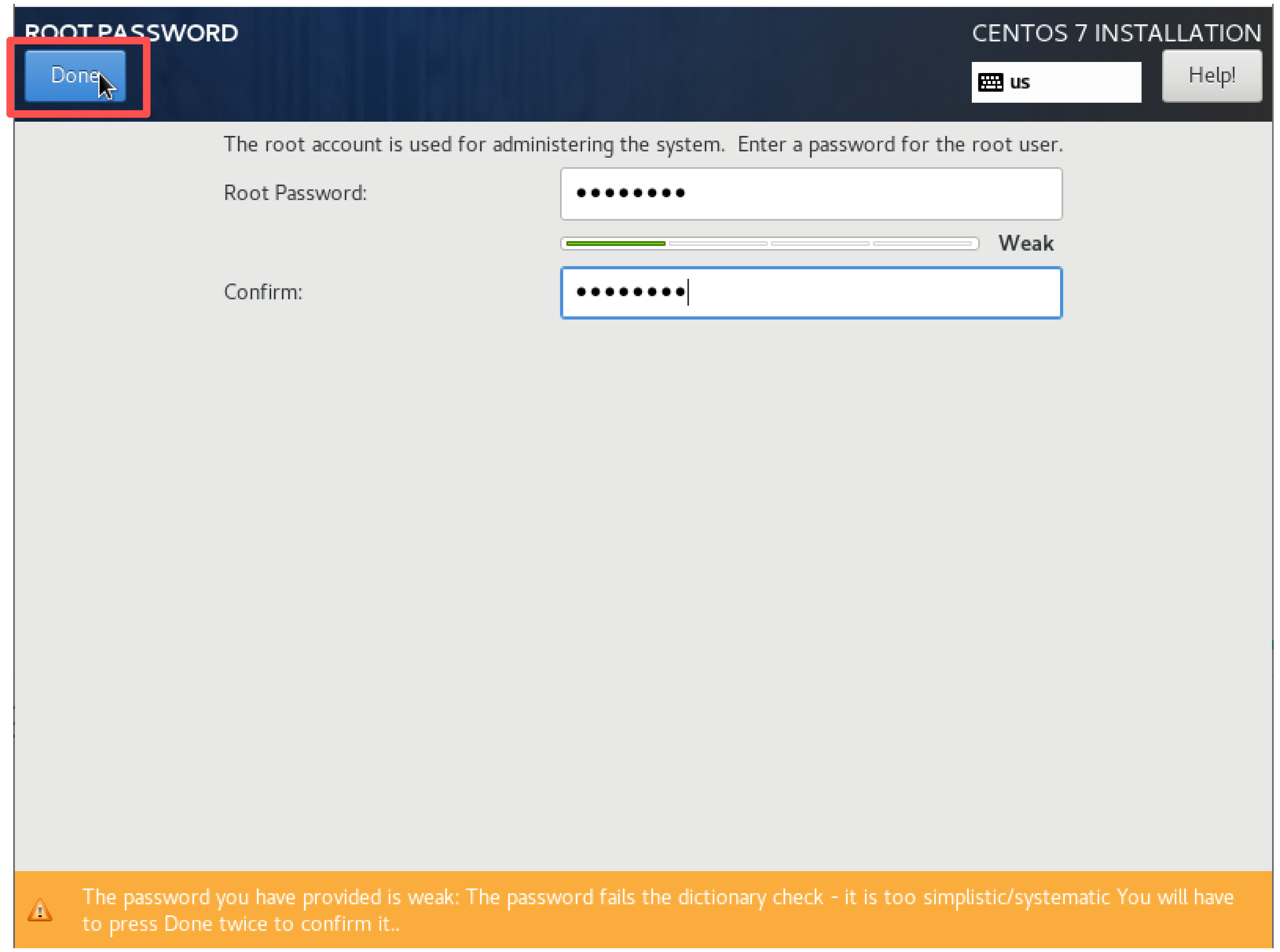
ROOT PASSWORD
Root password is not set

USER CREATION
No user will be created

Installing gnome-user-docs (46/1573)

entOS Virtualization SIG
rtualization in CentOS, virtualization of CentOS.
i.centos.org/SpecialInterestGroup

Please complete items marked with this icon before continuing to the next step.



The root account is used for administering the system. Enter a password for the root user.

Root Password: ••••••••

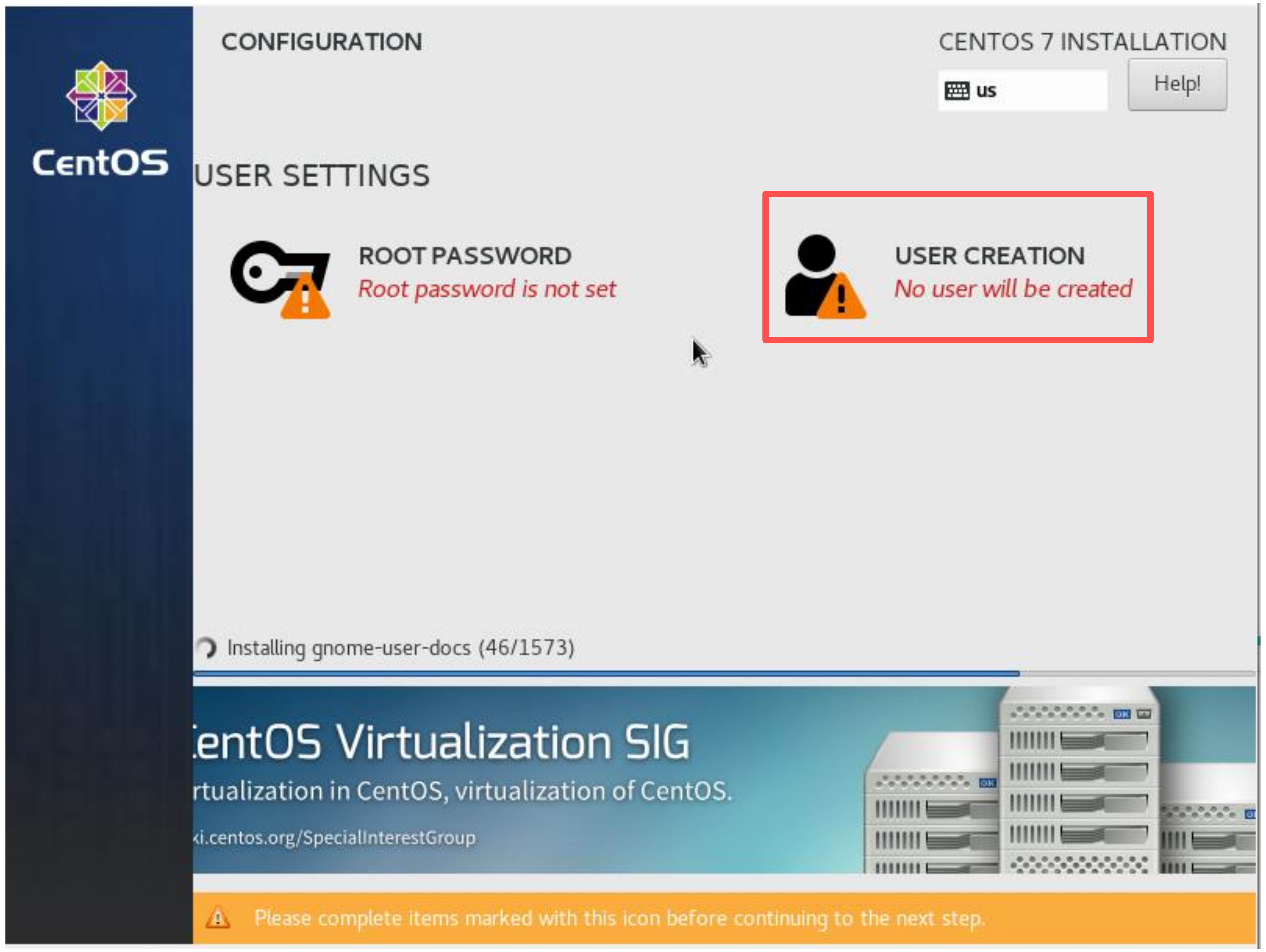
Weak

Confirm: ••••••••

The password you have provided is weak: The password fails the dictionary check - it is too simplistic/systematic You will have to press Done twice to confirm it..

CentOS 7

➢ CentOS installation process



USER SETTINGS

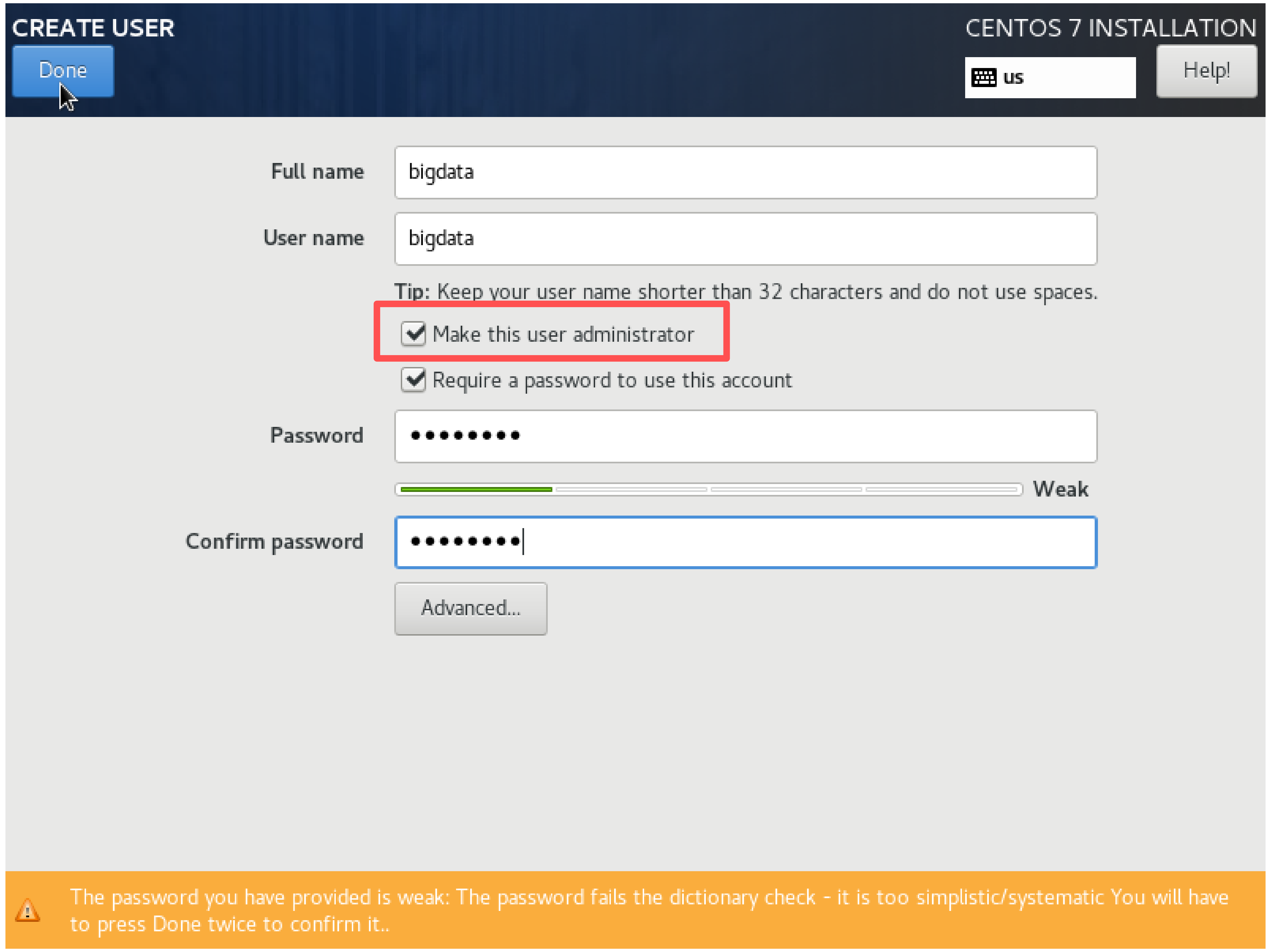
ROOT PASSWORD
Root password is not set

USER CREATION
No user will be created

Installing gnome-user-docs (46/1573)

entOS Virtualization SIG
rtualization in CentOS, virtualization of CentOS.
i.centos.org/SpecialInterestGroup

⚠ Please complete items marked with this icon before continuing to the next step.



Full name: bigdata

User name: bigdata

Tip: Keep your user name shorter than 32 characters and do not use spaces.

☑ Make this user administrator

☑ Require a password to use this account

Password: ••••••••

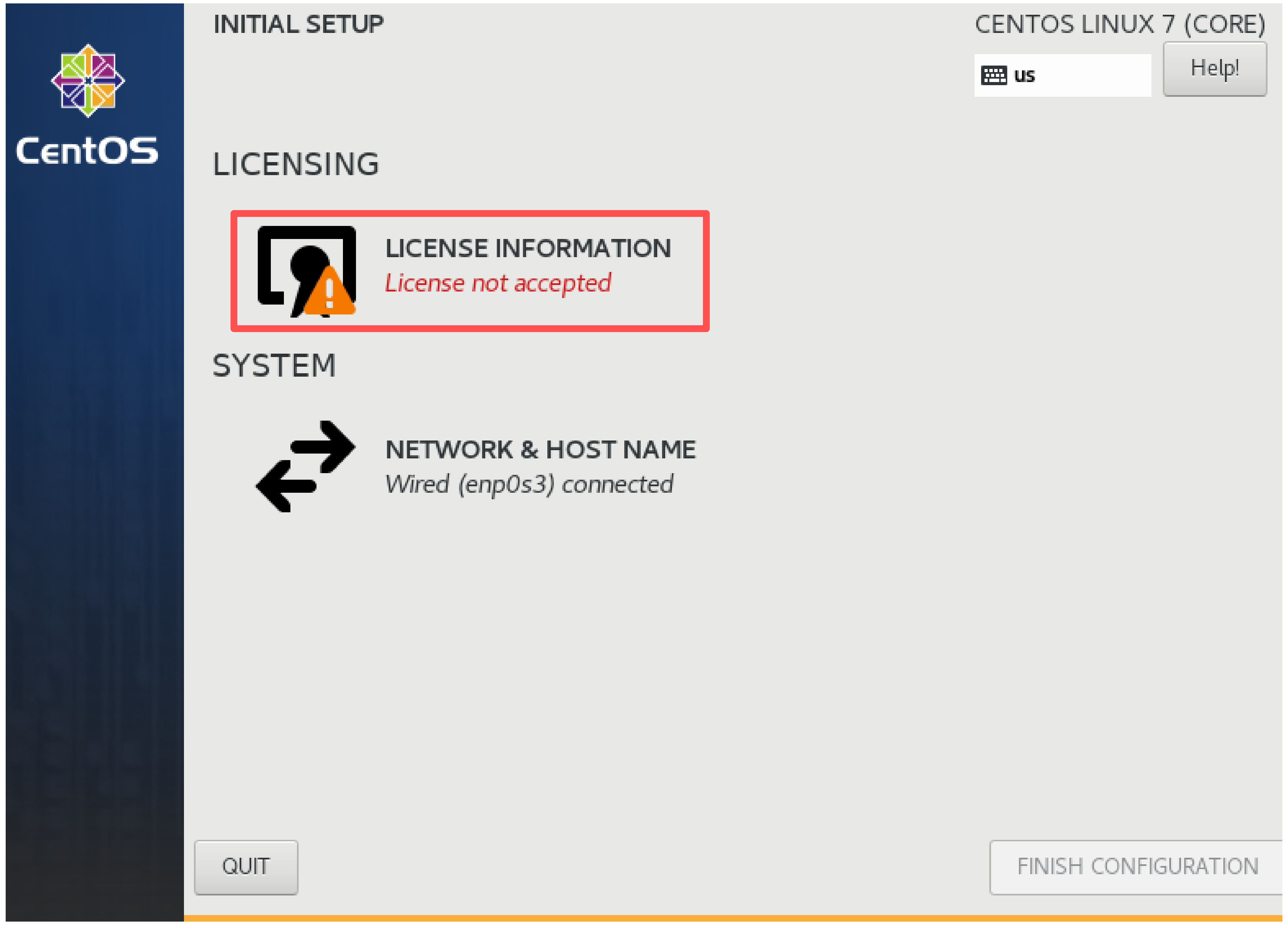
Weak

Confirm password: ••••••••

⚠ The password you have provided is weak: The password fails the dictionary check - it is too simplistic/systematic You will have to press Done twice to confirm it..

CentOS 7

➢ CentOS installation process



INITIAL SETUP

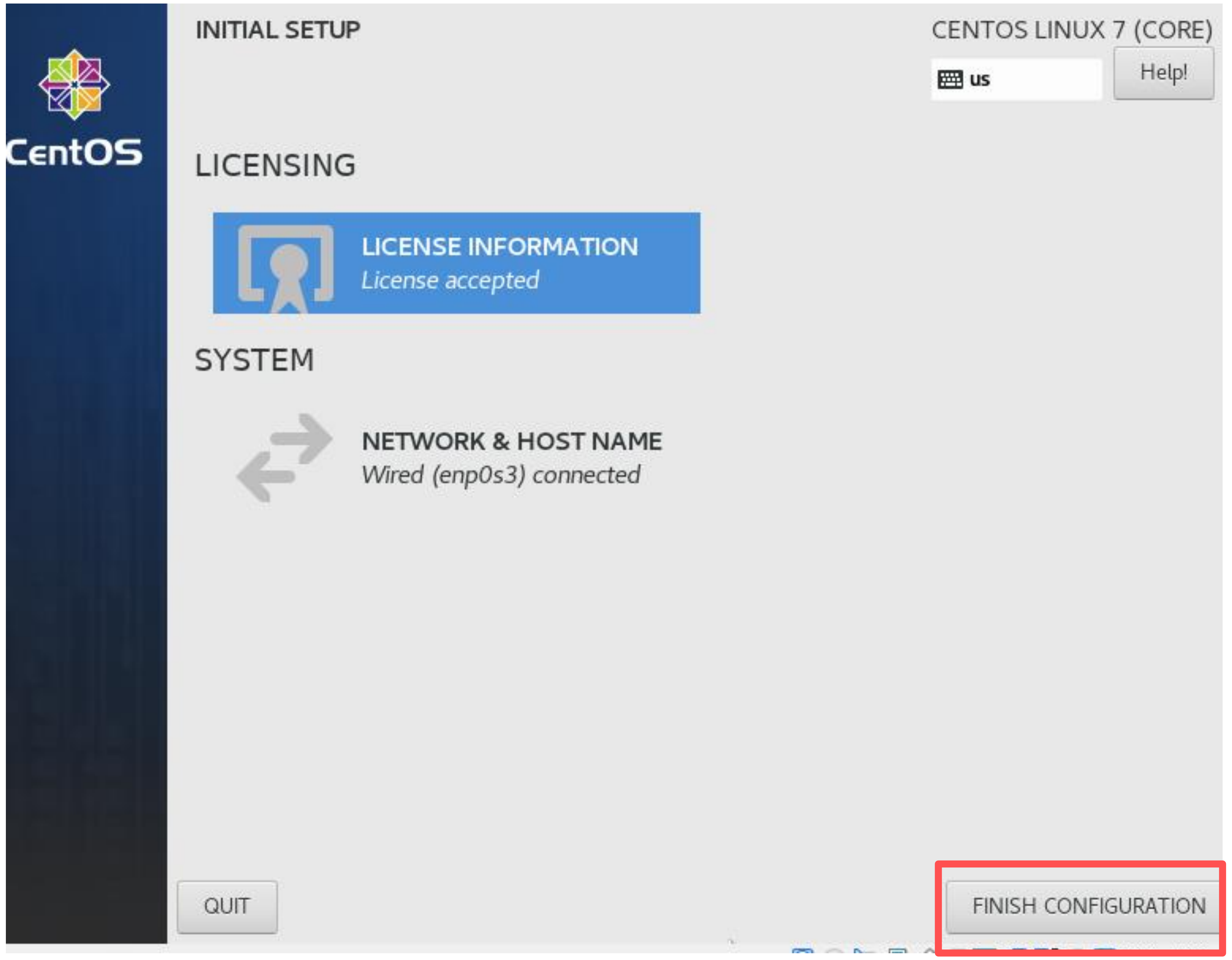
CENTOS LINUX 7 (CORE)

LICENSING

LICENSE INFORMATION
License not accepted

SYSTEM

NETWORK & HOST NAME
Wired (enp0s3) connected



INITIAL SETUP

CENTOS LINUX 7 (CORE)

LICENSING

LICENSE INFORMATION
License accepted

SYSTEM

NETWORK & HOST NAME
Wired (enp0s3) connected

Download Python3

- **Start CentOS, then move into the terminal**
- **Install python3 with the following command**

  - Add repository to yum

    ```
    sudo yum install –y https://centos7.iuscommunity.org/ius-release.rpm
    ```

  - Install the library

    ```
    sudo yum install –y python36u python36-libs python36-devel python36u-pip
    ```

  - Modify Alias

    ```
    sudo unlink /bin/python
    sudo ln -s /bin/python3.6  /bin/python
    ```

https://snowdeer.github.io/python/2018/02/20/install-python3-on-centos/

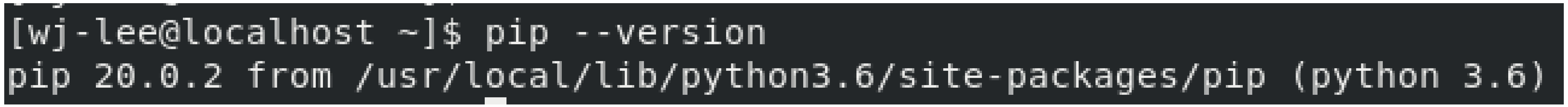
Download Pip

- Install Pip which is a tool for downloading python library

- Install Pip with the following command

  - sudo curl https://bootstrap.pypa.io/get-pip.py -o get-pip.py
  - python get-pip.py

- You can check the pip version using “pip --version” command



```
[wj-lee@localhost ~]$ pip --version
pip 20.0.2 from /usr/local/lib/python3.6/site-packages/pip (python 3.6)
```

https://linuxize.com/post/how-to-install-pip-on-centos-7/

Download Packages

- **Install libraries using pip3 with the following command:**

```
pip3 install numpy
pip3 install scikit-learn
pip3 install matplotlib
```

- **Install pyspark library with the following command:**

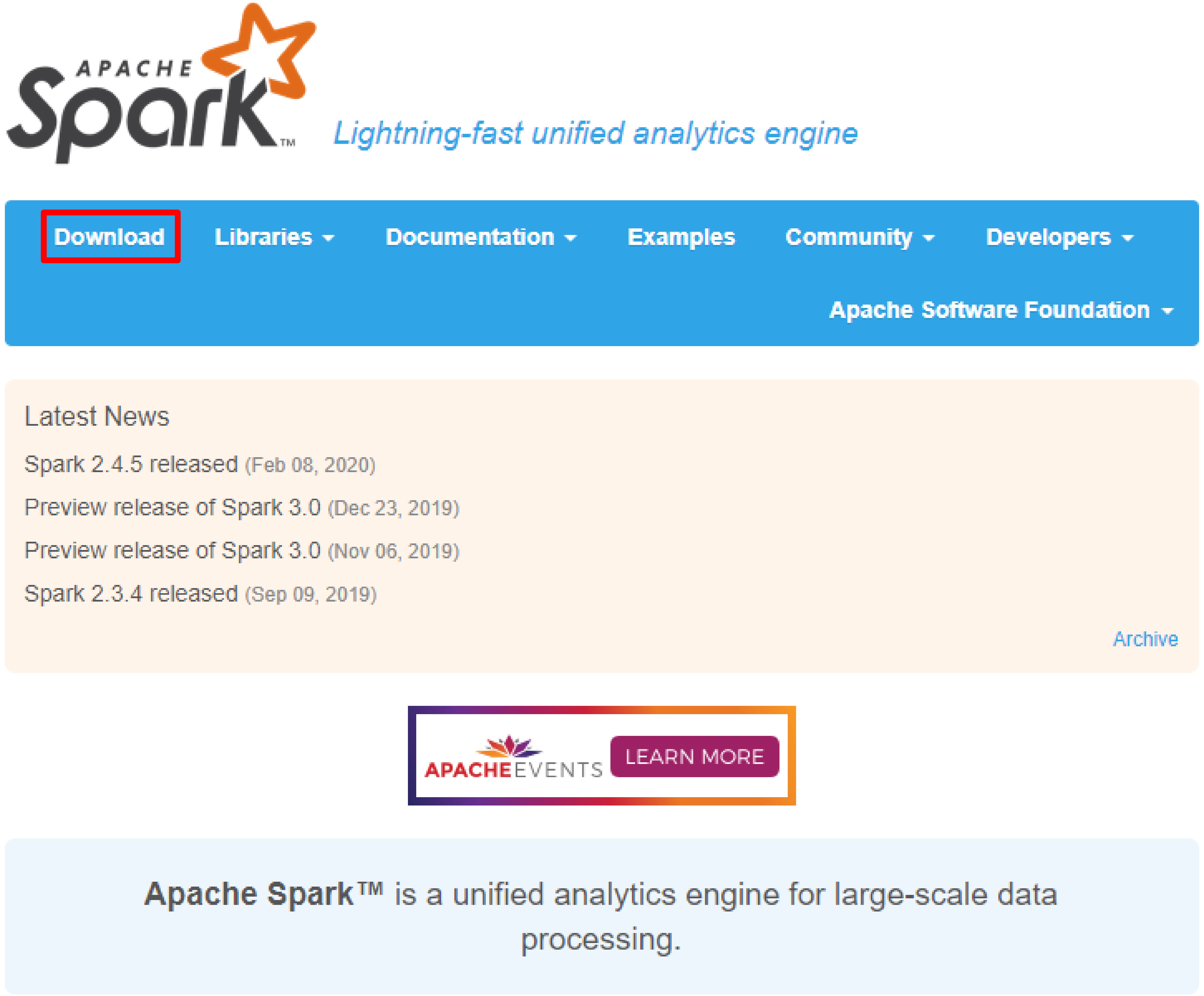
```
pip3 install pyspark –U --no-cache
```

https://snowdeer.github.io/python/2018/02/20/install-python3-on-centos/

Spark installation

Download Spark

- Go http://spark.apache.org/ and click “Download”



Latest News

Spark 2.4.5 released (Feb 08, 2020)

Preview release of Spark 3.0 (Dec 23, 2019)

Preview release of Spark 3.0 (Nov 06, 2019)

Spark 2.3.4 released (Sep 09, 2019)

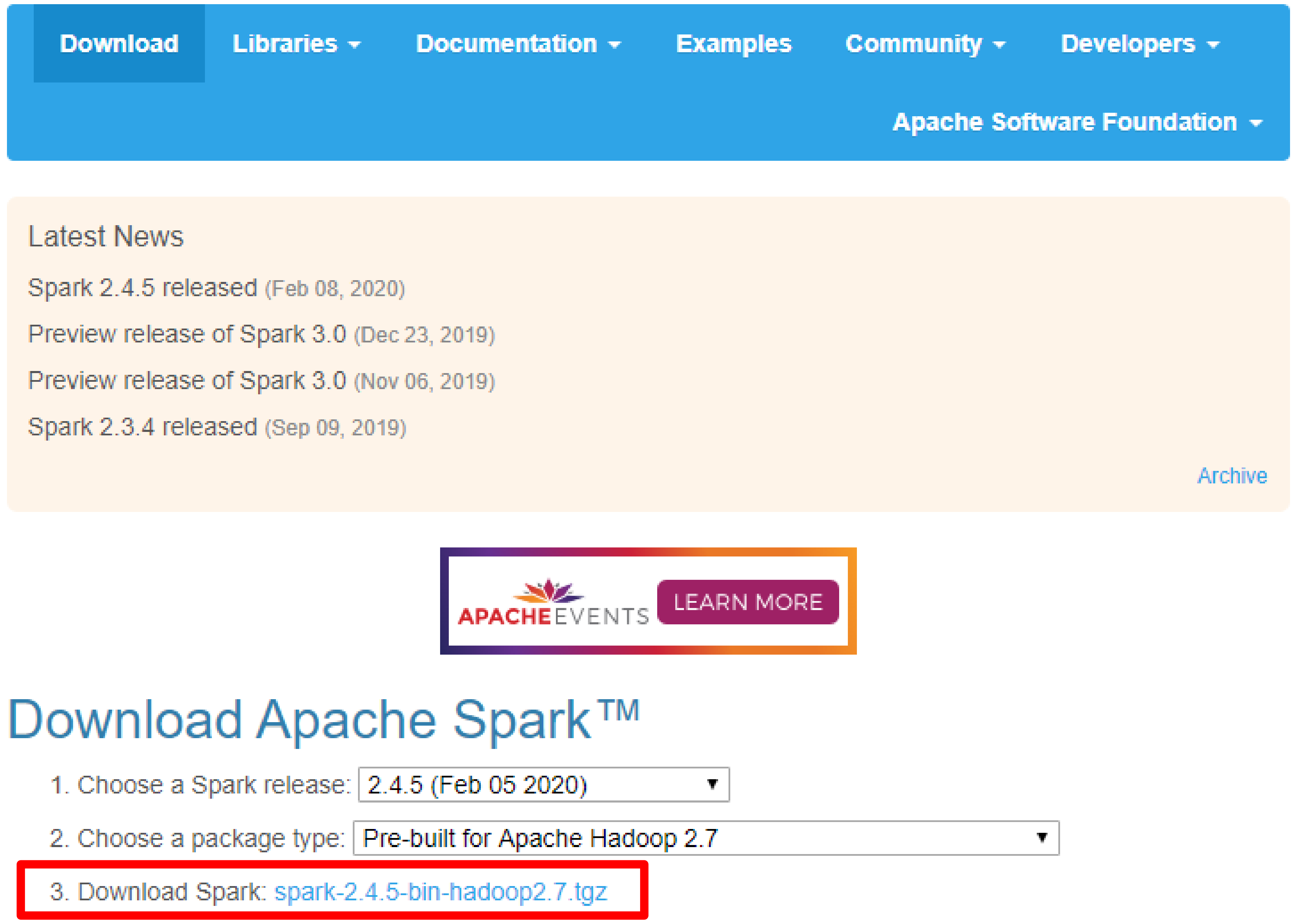
Archive

Apache Spark™ is a unified analytics engine for large-scale data processing.

http://spark.apache.org/

Download Spark

- Choose spark 2.4.5 & Apache Hadoop 2.7



Latest News

Spark 2.4.5 released (Feb 08, 2020)

Preview release of Spark 3.0 (Dec 23, 2019)

Preview release of Spark 3.0 (Nov 06, 2019)

Spark 2.3.4 released (Sep 09, 2019)

Archive

Download Apache Spark™

1. Choose a Spark release: 2.4.5 (Feb 05 2020)
2. Choose a package type: Pre-built for Apache Hadoop 2.7
3. Download Spark: spark-2.4.5-bin-hadoop2.7.tgz

- Then, download “spark-2.4.5-bin-hadoop2.7.tgz”
- Or, you can download the file here:

https://www.apache.org/dyn/closer.lua/spark/spark-2.4.5/spark-2.4.5-bin-hadoop2.7.tgz

http://spark.apache.org/

Download Spark

- **Unpack your .tgz file**

  - Move *spark-2.4.5-bin-hadoop2.7.tgz* file to your $HOME directory

    ```
    sudo mv spark-2.4.5-bin-hadoop2.7 $HOME (at your current directory where this file installed)
    ```

  - Unpack your .tgz file like following command:

    ```
    tar zxvf spark-2.4.5-bin-hadoop2.7.tgz
    ```

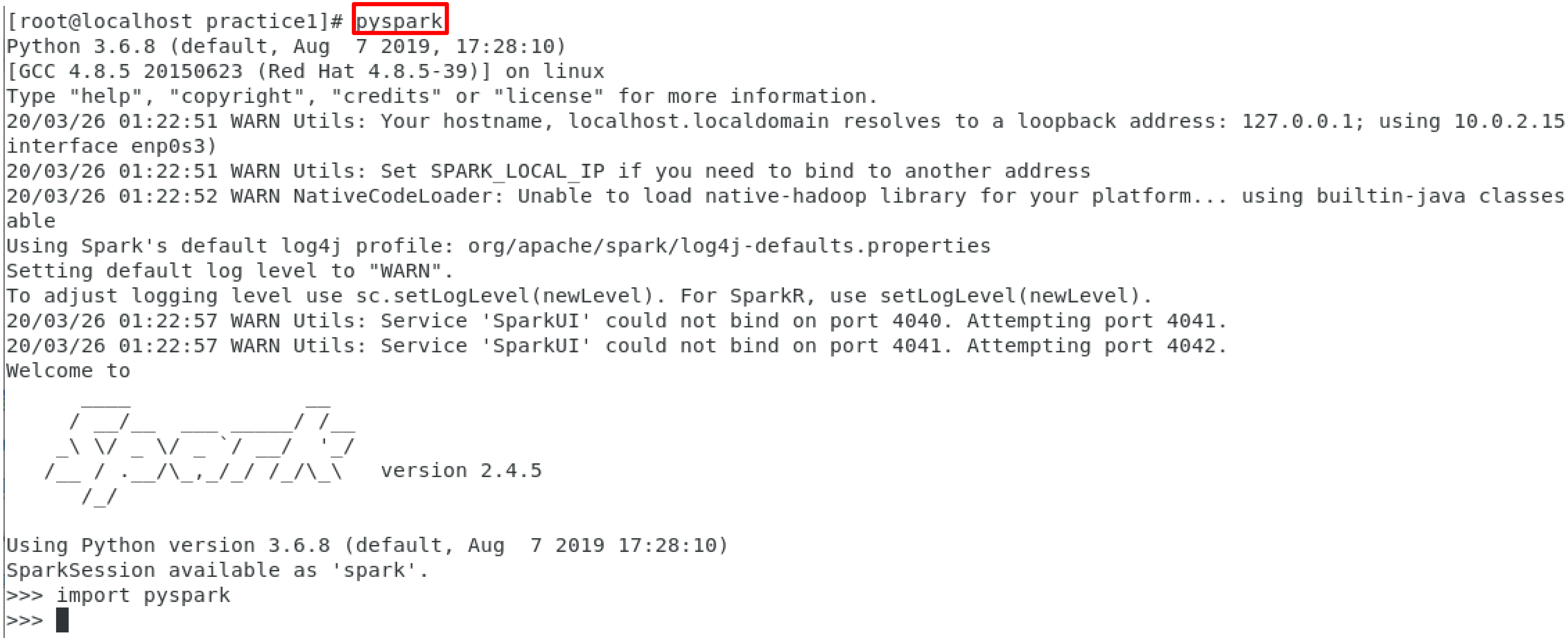
- **Set the path**

  ```
  export SPARK_HOME=$HOME/spark-2.4.5-bin-hadoop2.7
  export PATH=$PATH:$SPARK_HOME/bin
  echo ‘export SPARK_HOME=$HOME/spark-2.4.5-bin-hadoop2.7’ >> .bash_profile
  echo ‘export PATH=$PATH:$SPARK_HOME/bin’ >> .bash_profile
  echo ‘export SPARK_HOME= $HOME/spark-2.4.5-bin-hadoop2.7’ >> ~/.bashrc
  echo ‘export PATH=$PATH:$SPARK_HOME/bin’ >> ~/.bashrc
  ```

https://gist.github.com/darcyliu/d47edccb923b0f03280a4cf8b66227c1

Start Pyspark in CentOS

- ➢ **You can execute Pyspark entering just “pyspark” at command line.**
- ➢ **Then, you can import pyspark here.**



```
[root@localhost practice1]# pyspark
Python 3.6.8 (default, Aug  7 2019, 17:28:10)
[GCC 4.8.5 20150623 (Red Hat 4.8.5-39)] on linux
Type "help", "copyright", "credits" or "license" for more information.
20/03/26 01:22:51 WARN Utils: Your hostname, localhost.localdomain resolves to a loopback address: 127.0.0.1; using 10.0.2.15
interface enp0s3)
20/03/26 01:22:51 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
20/03/26 01:22:52 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
able
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
20/03/26 01:22:57 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
20/03/26 01:22:57 WARN Utils: Service 'SparkUI' could not bind on port 4041. Attempting port 4042.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.4.5
      /_/

Using Python version 3.6.8 (default, Aug  7 2019 17:28:10)
SparkSession available as 'spark'.
>>> import pyspark
>>> 
```

Hadoop installation

Install Hadoop

- Check whether JAVA is already installed : java -version
- Install ssh and and pdsh

```
sudo yum install openssh-server openssh-clients openssh-askpass
sudo yum install pdsh
```

- In case of yum exception, edit the first line of configuration files like below

```
sudo vi /usr/bin/yum
sudo vi /usr/libexec/urlgrabber-ext-down
```

```
#! /usr/bin/python
```

➡

```
#! /usr/bin/python2.7
```

❖ *In Vim editior, click "i" for editing(edit mode)*

❖ *click ESC(command mode) and ":wq" for save and exit.*

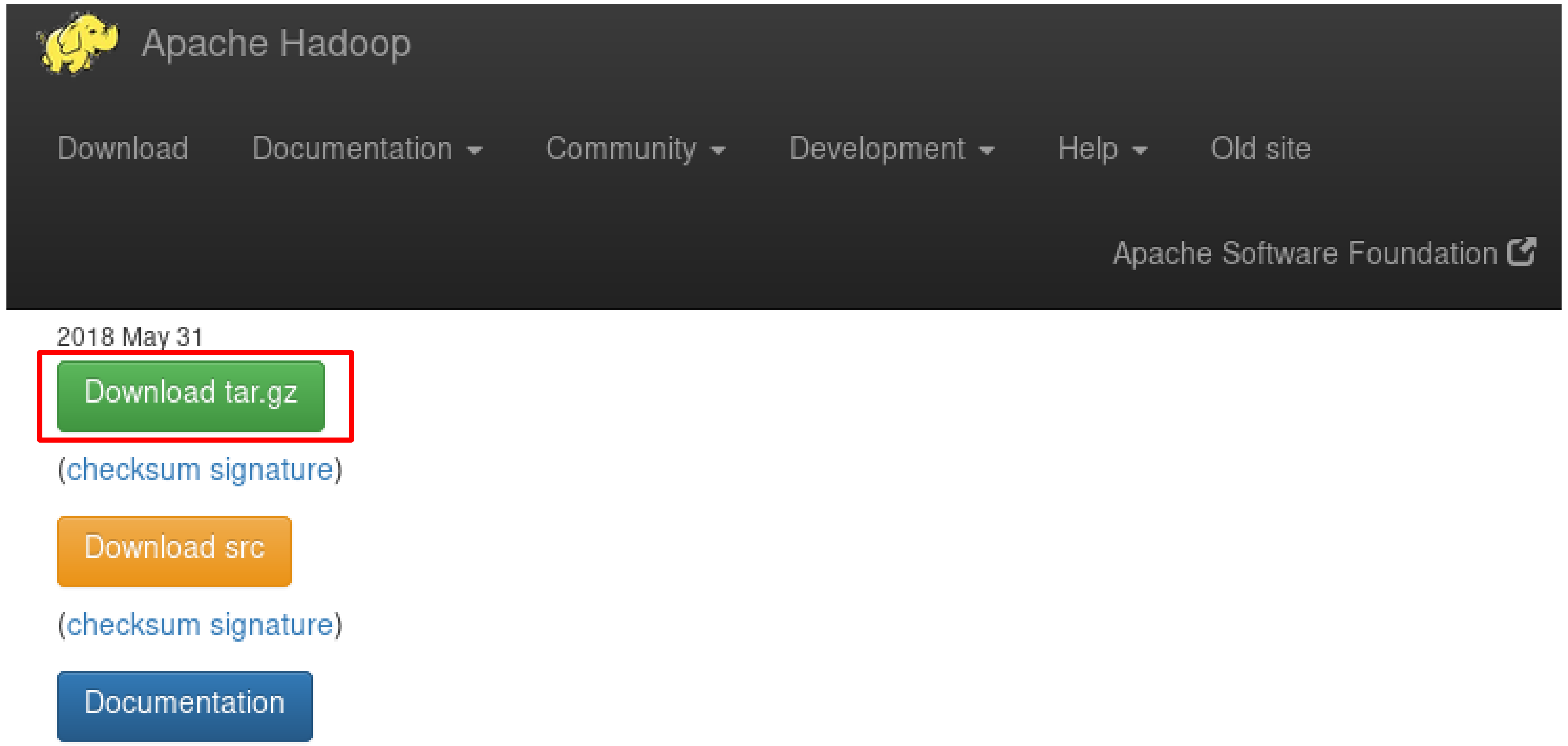
❖ *If you want to exit without save, click ":q!" and type "q!".*

Install Hadoop

- Go to Hadoop website and download package

  https://hadoop.apache.org/release/2.7.7.html

- Click “Download tar.gz”

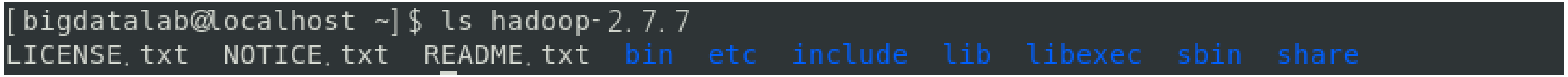


2018 May 31

Install Hadoop

- Move tar.gz package from Downloads directory to home directory
- Unpack tar.gz package and check new directory for Hadoop

```
tar zxvf hadoop-2.7.7.tar.gz
ls hadoop-2.7.7
```



```
[bigdatalab@localhost ~]$ ls hadoop-2.7.7
LICENSE.txt  NOTICE.txt  README.txt  bin  etc  include  lib  libexec  sbin  share
```

- Path configuration

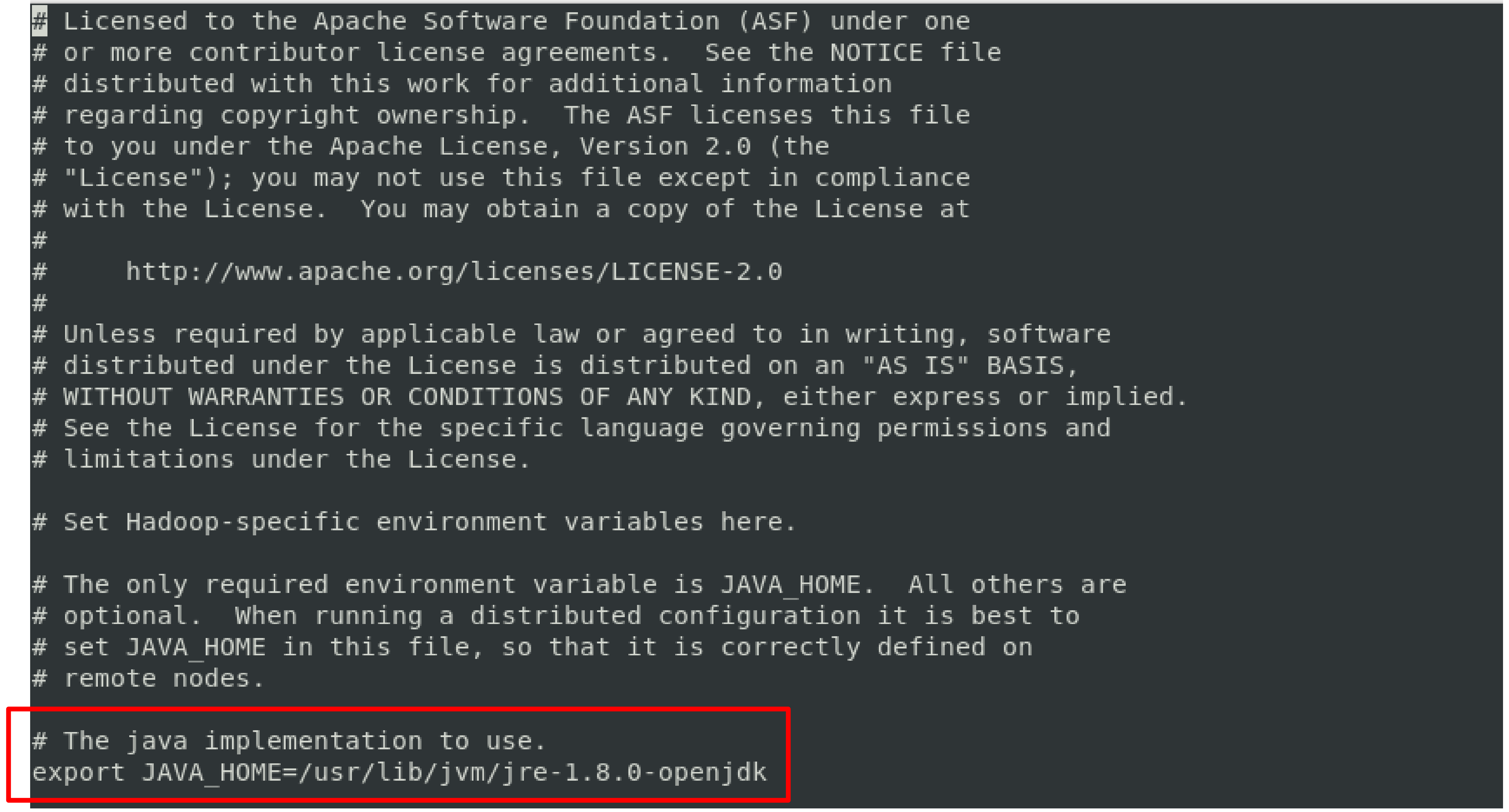
```
export HADOOP_HOME=$HOME/hadoop-2.7.7
export PATH=$PATH:$HADOOP_HOME/bin
echo ‘export HADOOP_HOME=$HOME/hadoop-2.7.7’ >> .bash_profile
echo ‘export PATH=$PATH:$HADOOP_HOME/bin’ >> .bash_profile
echo ‘export HADOOP_HOME=$HOME/hadoop-2.7.7’ >> ~/.bashrc
echo ‘export PATH=$PATH:$HADOOP_HOME/bin’ >> ~/.bashrc
```

Hadoop Configuration

- **Edit JAVA path configuration file**

  sudo vi $HADOOP_HOME/etc/hadoop/hadoop-env.sh

  edit “export JAVA_HOME=${JAVA_HOME}” to “export JAVA_HOME=/usr/lib/jvm/jre-1.8.0-openjdk“



```
# Licensed to the Apache Software Foundation (ASF) under one
# or more contributor license agreements.  See the NOTICE file
# distributed with this work for additional information
# regarding copyright ownership.  The ASF licenses this file
# to you under the Apache License, Version 2.0 (the
# "License"); you may not use this file except in compliance
# with the License.  You may obtain a copy of the License at
#
#     http://www.apache.org/licenses/LICENSE-2.0
#
# Unless required by applicable law or agreed to in writing, software
# distributed under the License is distributed on an "AS IS" BASIS,
# WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
# See the License for the specific language governing permissions and
# limitations under the License.

# Set Hadoop-specific environment variables here.

# The only required environment variable is JAVA_HOME.  All others are
# optional.  When running a distributed configuration it is best to
# set JAVA_HOME in this file, so that it is correctly defined on
# remote nodes.

# The java implementation to use.
export JAVA_HOME=/usr/lib/jvm/jre-1.8.0-openjdk
```

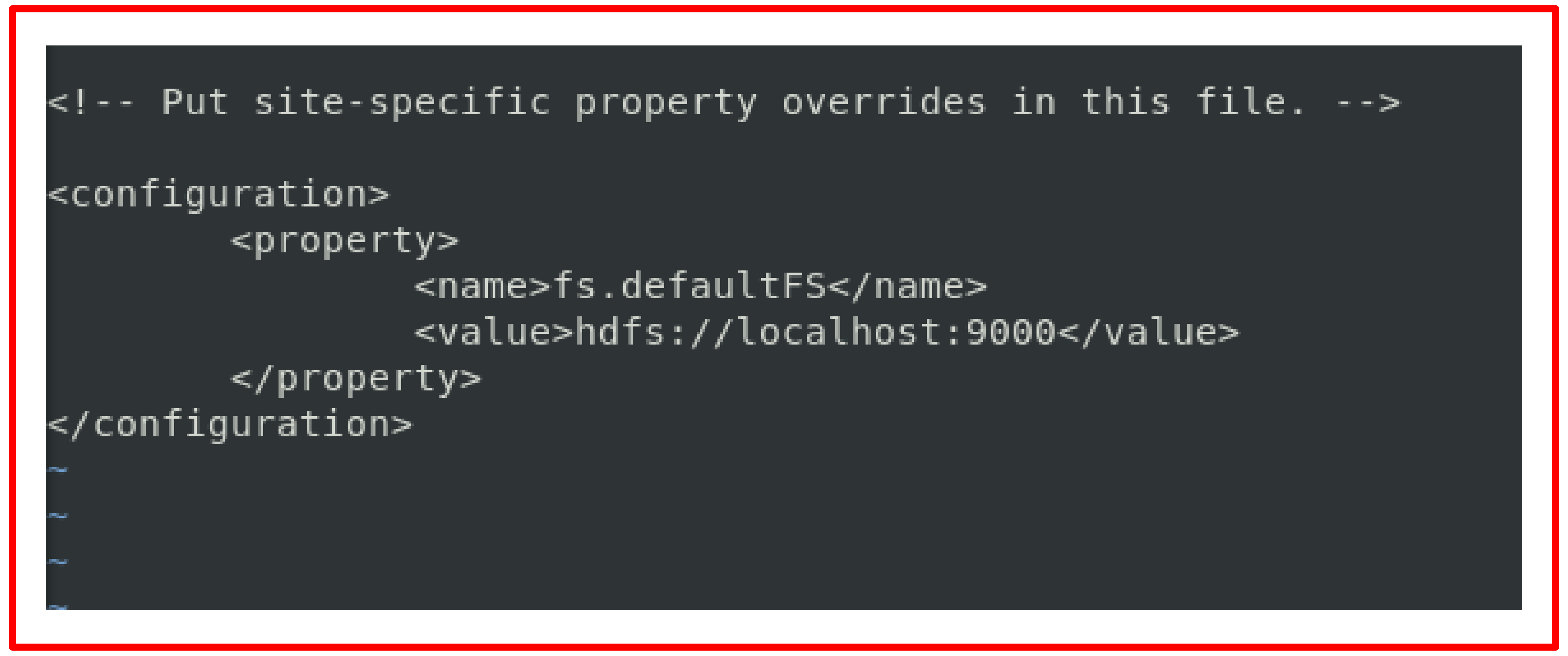
- ***In Vim editior, click “i” for editing(edit mode)***
- ***click ESC(command mode) and “:wq” for save and exit.***
- ***If you want to exit without save, click “:q!” and type “q!”.***

Hadoop Configuration

- Before start Hadoop file system, we must edit some configuration files

  - sudo vi $HADOOP_HOME/etc/hadoop/core-stie.xml
  - Edit CONFIGURATION like following

```
<configuration>
   <property>
      <name>fs.defaultFS</name>
      <value>hdfs://localhost:9000</value>
   </property>
</configuration>
```



```
<!-- Put site-specific property overrides in this file. -->

<configuration>
        <property>
                <name>fs.defaultFS</name>
                <value>hdfs://localhost:9000</value>
        </property>
</configuration>
~
~
~
```

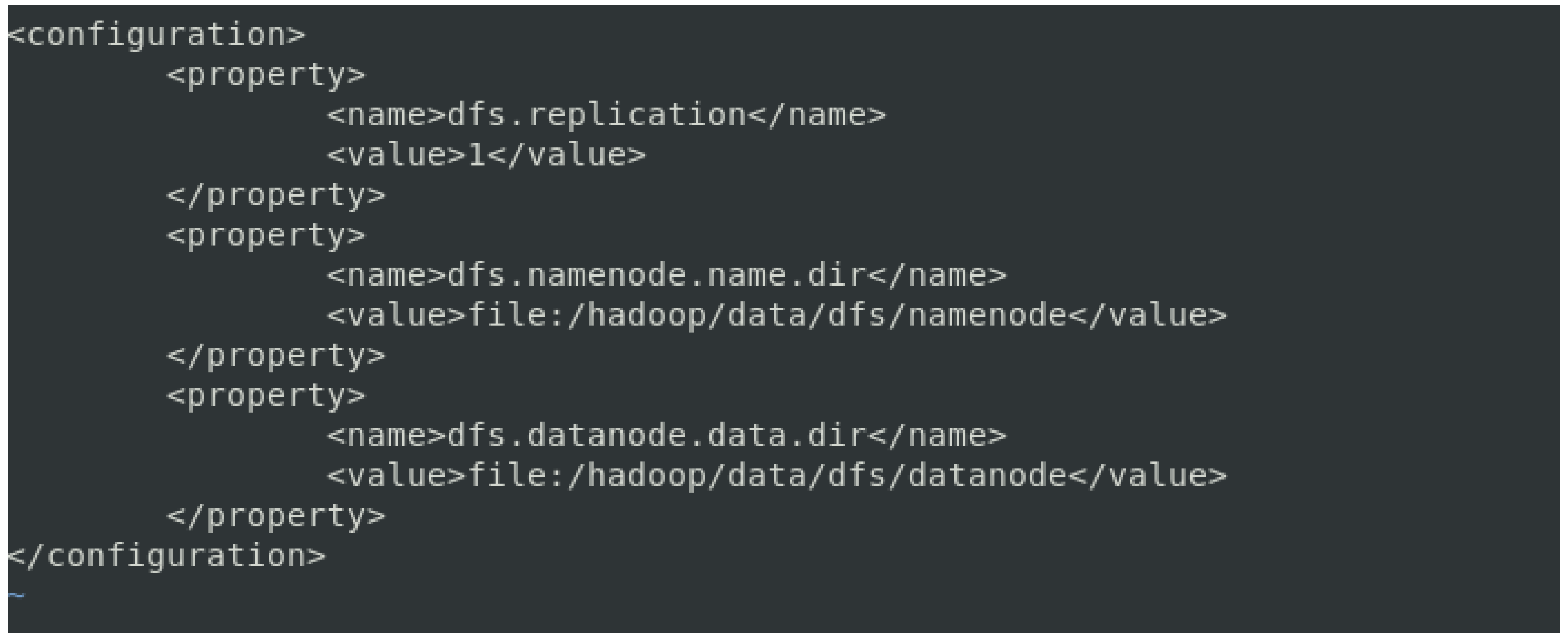
- ***In Vim editior, click "i" for editing(edit mode)***
- ***click ESC(command mode) and ":wq" for save and exit.***
- ***If you want to exit without save, click ":q!" and type "q!".***

Hadoop Configuration

- Before start Hadoop file system, we must edit some configuration files

  - sudo vi $HADOOP_HOME/etc/hadoop/hdfs-site.xml
  - Edit CONFIGURATION like following

```
<configuration>
        <property>
                <name>dfs.replication</name>
        <value>1</value>
        </property>
        <property>
                <name>dfs.namenode.name.dir</name>
                <value>file:/hadoop/data/dfs/namenode</value>
        </property>
        <property>
                <name>dfs.datanode.data.dir</name>
                <value>file:/hadoop/data/dfs/datanode</value>
        </property>
</configuration>
```



```
<configuration>
        <property>
                <name>dfs.replication</name>
                <value>1</value>
        </property>
        <property>
                <name>dfs.namenode.name.dir</name>
                <value>file:/hadoop/data/dfs/namenode</value>
        </property>
        <property>
                <name>dfs.datanode.data.dir</name>
                <value>file:/hadoop/data/dfs/datanode</value>
        </property>
</configuration>
~
```

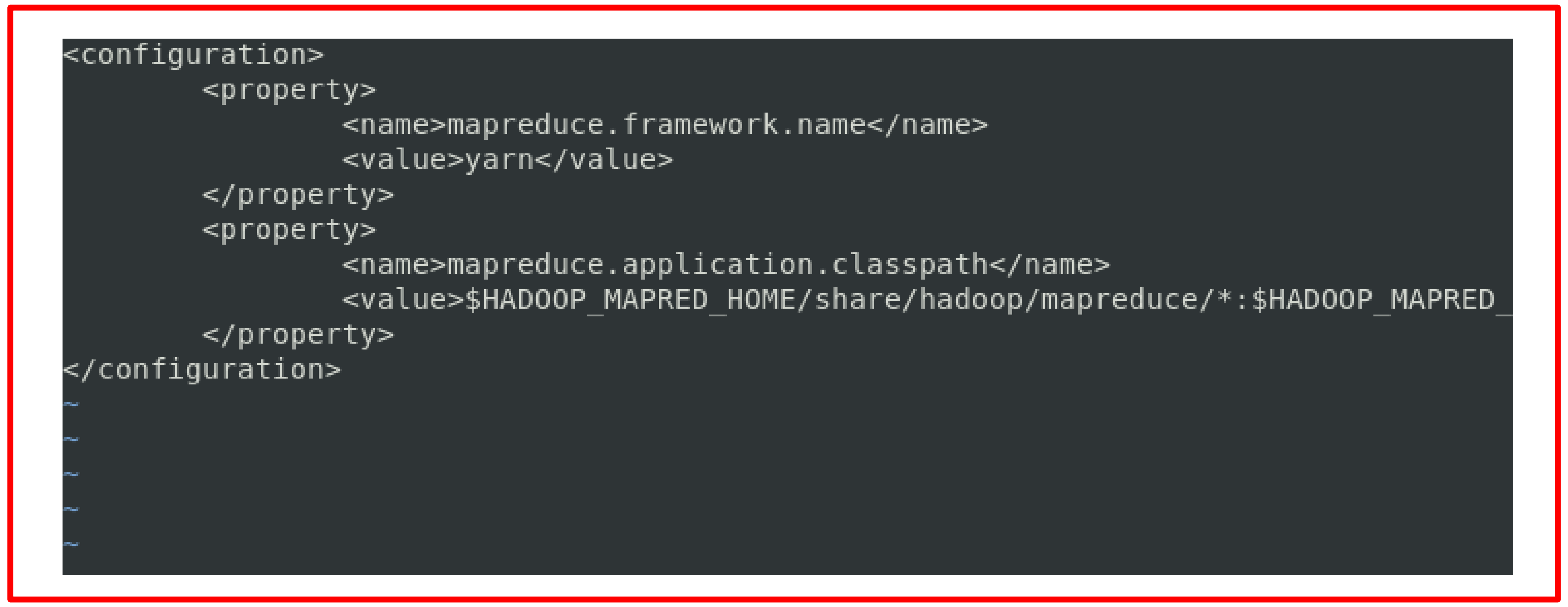
- ❖ *In Vim editior, click “i” for editing(edit mode)*
- ❖ *click ESC(command mode) and “:wq” for save and exit.*
- ❖ *If you want to exit without save, click “:q!” and type “q!”.*

Hadoop Configuration

- Before start Hadoop file system, we must edit some configuration files

  - sudo vi $HADOOP_HOME/etc/hadoop/mapred-site.xml
  - Edit CONFIGURATION like following

```
<configuration>
   <property>
      <name>mapreduce.framework.name</name>
      <value>yarn</value>
   </property>
   <property>
      <name>mapreduce.application.classpath</name>

<value>$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/*:$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/lib/*</value>
   </property>
</configuration>
```



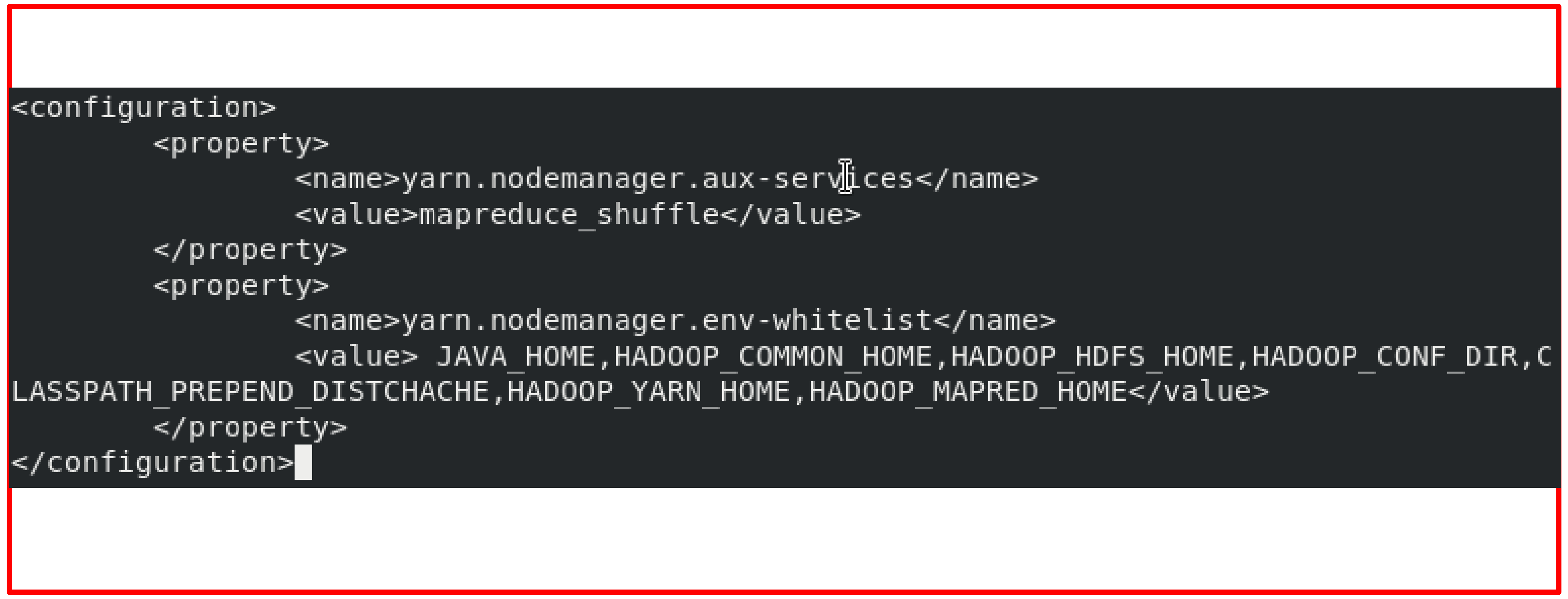
```
<configuration>
        <property>
                <name>mapreduce.framework.name</name>
                <value>yarn</value>
        </property>
        <property>
                <name>mapreduce.application.classpath</name>
                <value>$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/*:$HADOOP_MAPRED_
        </property>
</configuration>
~
~
~
~
~
```

- *In Vim editior, click “i” for editing(edit mode)*
- *click ESC(command mode) and “:wq” for save and exit.*
- *If you want to exit without save, click “:q!” and type “q!”.*

Hadoop Configuration

- Before start Hadoop file system, we must edit some configuration files
  - sudo vi $HADOOP_HOME/etc/hadoop/yarn-site.xml
  - Edit CONFIGURATION like following

```
<configuration>
   <property>
      <name>yarn.nodemanager.aux-services</name>
      <value>mapreduce_shuffle</value>
   </property>
   <property>
      <name>yarn.nodemanager.env-whitelist</name>

<value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CLASSPATH_PREPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP_MAPRED_HOME</value>
   </property>
</configuration>
```

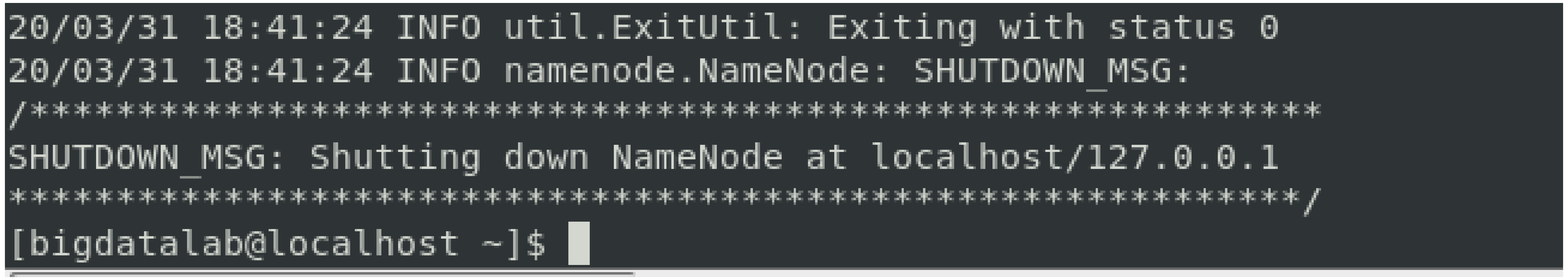


```
<configuration>
        <property>
                <name>yarn.nodemanager.aux-services</name>
                <value>mapreduce_shuffle</value>
        </property>
        <property>
                <name>yarn.nodemanager.env-whitelist</name>
                <value> JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,C
LASSPATH_PREPEND_DISTCHACHE,HADOOP_YARN_HOME,HADOOP_MAPRED_HOME</value>
        </property>
</configuration>
```

- *In Vim editior, click “i” for editing(edit mode)*
- *click ESC(command mode) and “:wq” for save and exit.*
- *If you want to exit without save, click “:q!” and type “q!”.*

Hadoop Configuration

- To format Namenode enter the command: hdfs namenode -format



```
20/03/31 18:41:24 INFO util.ExitUtil: Exiting with status 0
20/03/31 18:41:24 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at localhost/127.0.0.1
************************************************************/
[bigdatalab@localhost ~]$ 
```

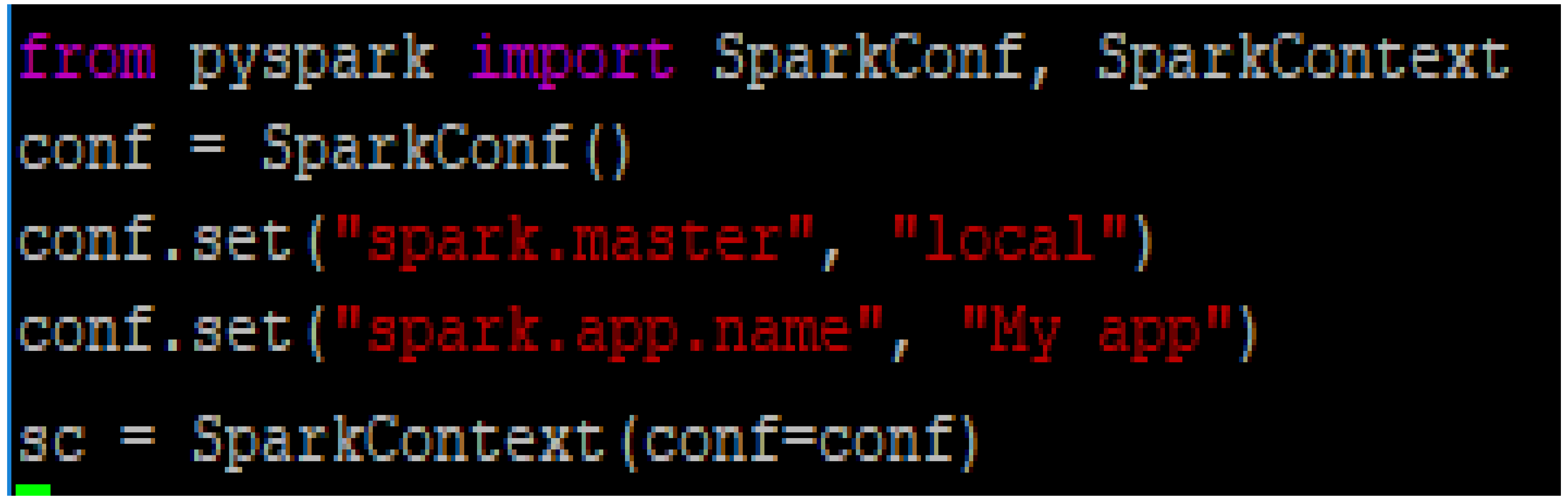
- Start DFS and YARN (You need to type password multiple times)

```
sudo $HADOOP_HOME/sbin/start-dfs.sh
sudo $HADOOP_HOME/sbin/start-yarn.sh
```

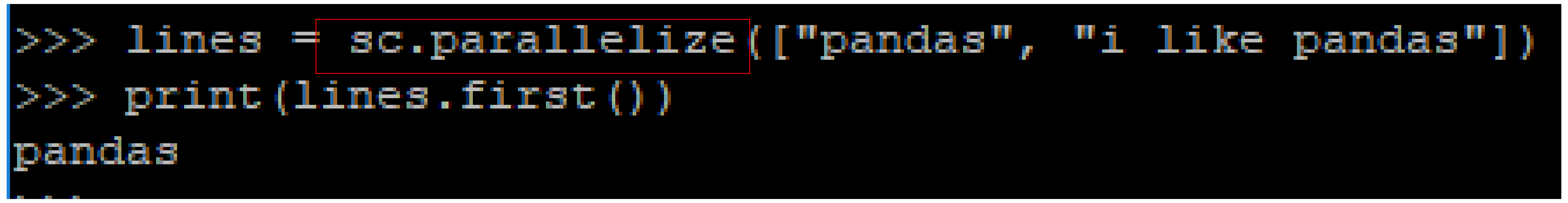
- Check “http://localhost:8088” and “http://localhost:50070”

Spark Implementation

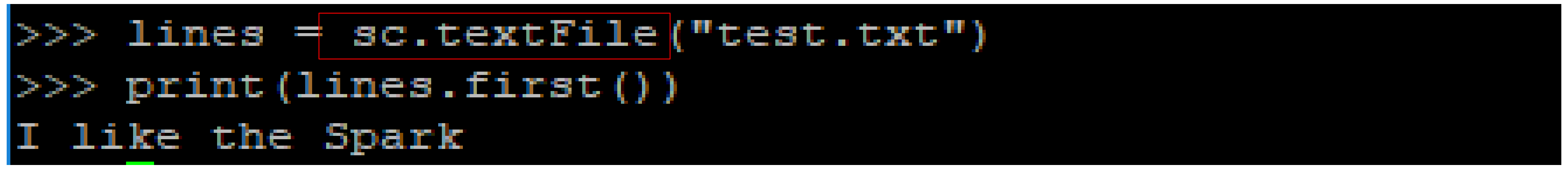
Steps for Spark Implementation



```
from pyspark import SparkConf, SparkContext
conf = SparkConf()
conf.set("spark.master", "local")
conf.set("spark.app.name", "My app")
sc = SparkContext(conf=conf)
```



```
>>> lines = sc.parallelize(["pandas", "i like pandas"])
>>> print(lines.first())
pandas
```



```
>>> lines = sc.textFile("test.txt")
>>> print(lines.first())
I like the Spark
```

1. **Initialize a SparkContext**
    - ➢ Import the Spark Package in your program
    - ➢ Configure Spark with SparkConf
    - ➢ Call *set* to add configuration values

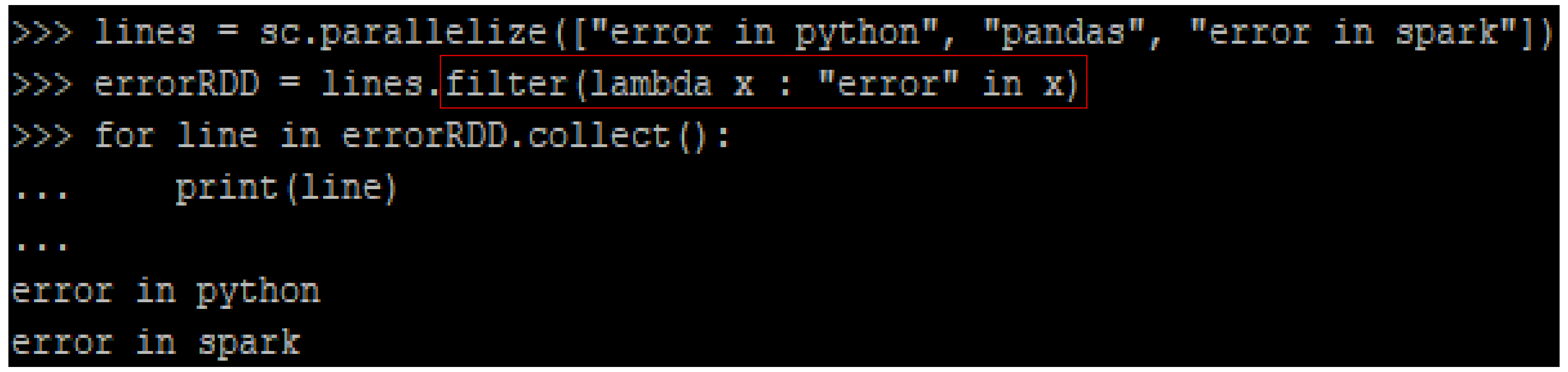
2. **Create RDDs (Import data)**
    - ➢ parallelize()
    - ➢ textFile()
        - ▪ load data from an external storage

Steps for Spark Implementation

3. RDD Operations

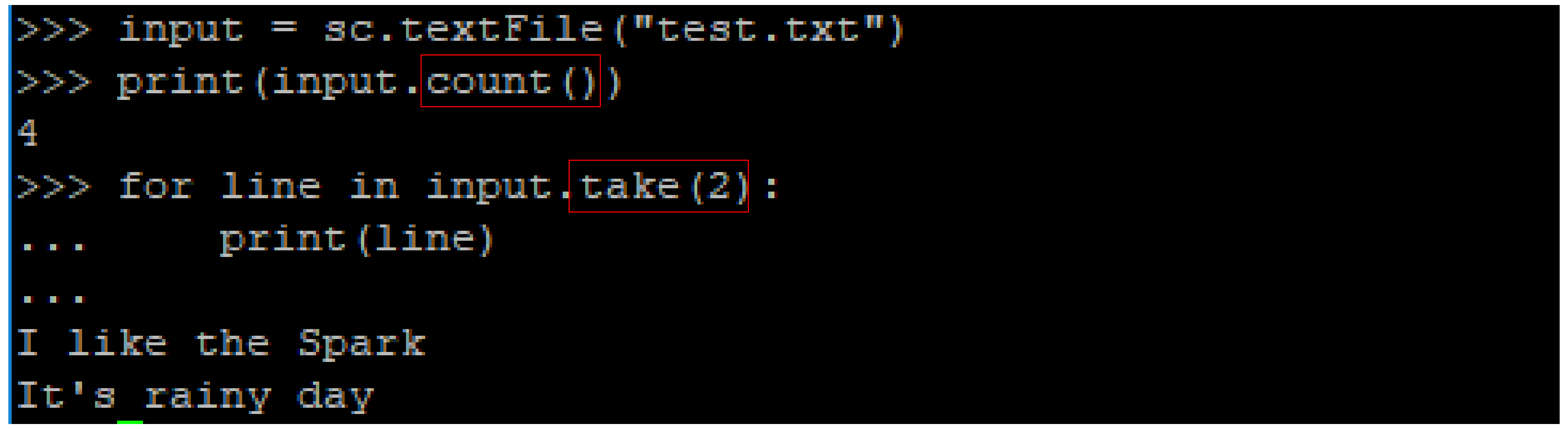
RDDs support two types of operations, *transformation* and *action*

1) Transformation
   - Operations on RDDs that return a new RDD
   - Transformed RDDs are computed only when you use them with an action function



```
>>> lines = sc.parallelize(["error in python", "pandas", "error in spark"])
>>> errorRDD = lines.filter(lambda x : "error" in x)
>>> for line in errorRDD.collect():
...     print(line)
...
error in python
error in spark
```

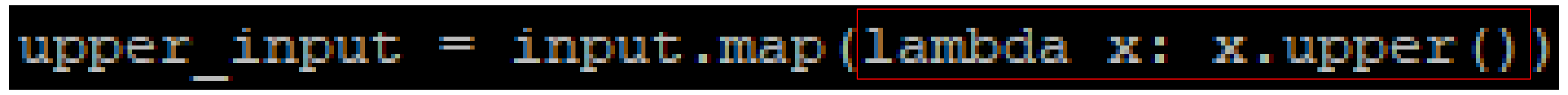
2) Action
   - Operations that return a final value to the driver program or write data to an external storage system
   - Actions force the evaluation of the transformations required for the RDD they were called on



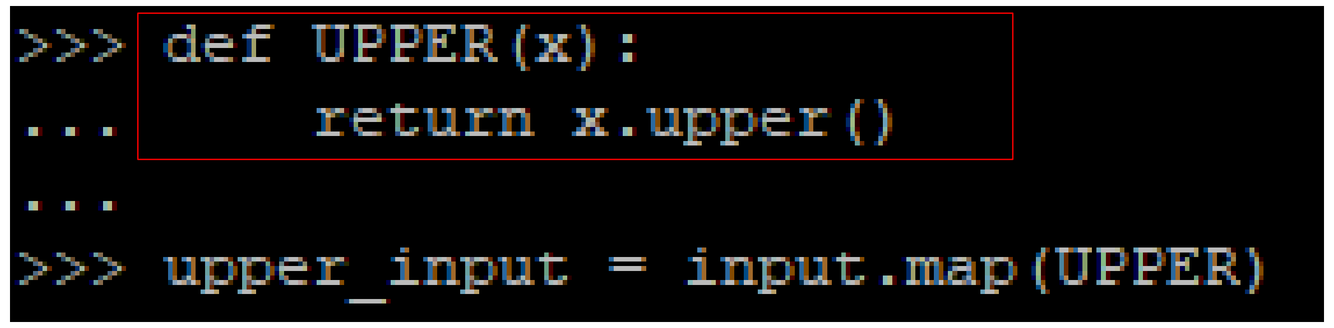
```
>>> input = sc.textFile("test.txt")
>>> print(input.count())
4
>>> for line in input.take(2):
...     print(line)
...
I like the Spark
It's rainy day
```

* ref.) https://spark.apache.org/docs/latest/rdd-programming-guide.htm

❖ Passing Function
   - Most of Spark's transformations, and some of its actions, depend on passing in function that are used to compute data
   - In python, we can pass in lambda expressions, top-level functions, or locally defined



```
upper_input = input.map(lambda x: x.upper())
```



```
>>> def UPPER(x):
...     return x.upper()
...
>>> upper_input = input.map(UPPER)
```

Steps for Spark Implementation

4. Use the Spark in an application

- In python, you simply write applications as Python scripts

- You must run them using the

  ```
  $ spark-submit ScriptName.py
  ```

- If you run your program on single machine with specific cores, use

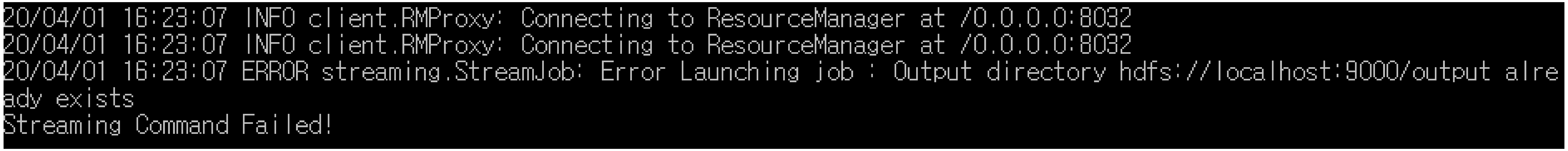
  ```
  $ spark-submit --master local[N] SctriptName.py
  ```

  * N : the number of cores you want to execute

Issues

Hadoop Issues

- ➢ **ERROR : Output directory already exists**



```
20/04/01 16:23:07 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
20/04/01 16:23:07 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
20/04/01 16:23:07 ERROR streaming.StreamJob: Error Launching job : Output directory hdfs://localhost:9000/output alre
ady exists
Streaming Command Failed!
```

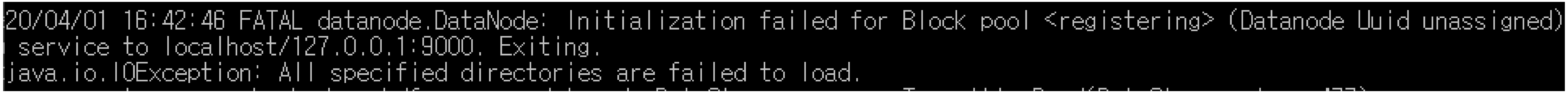
1. **Delete HDFS output directory using below command**

```
(Windows) hdfs dfs -rm /YOUR_DIRECTORY/*
          hdfs dfs -rmdir /YOUR_DIRECTORY/
(Linux) sudo %HADOOP_HOME%/bin/hdfs dfs –rm /YOUR_DIRECTORY/*
        sudo %HADOOP_HOME%/bin/hdfs dfs -rmdir /YOUR_DIRECTORY/
```

2. **Try again**

Hadoop Issues

- **ERROR : Initialization failed for Block pool**



```
20/04/01 16:42:46 FATAL datanode.DataNode: Initialization failed for Block pool <registering> (Datanode Uuid unassigned)
 service to localhost/127.0.0.1:9000. Exiting.
java.io.IOException: All specified directories are failed to load.
```

1. **Delete “data” folder.**

   (Windows) rm -r C:\Hadoop\data OR delete directly

   (Linux) sudo rm -r /hadoop/data

2. **Then, format namenode**

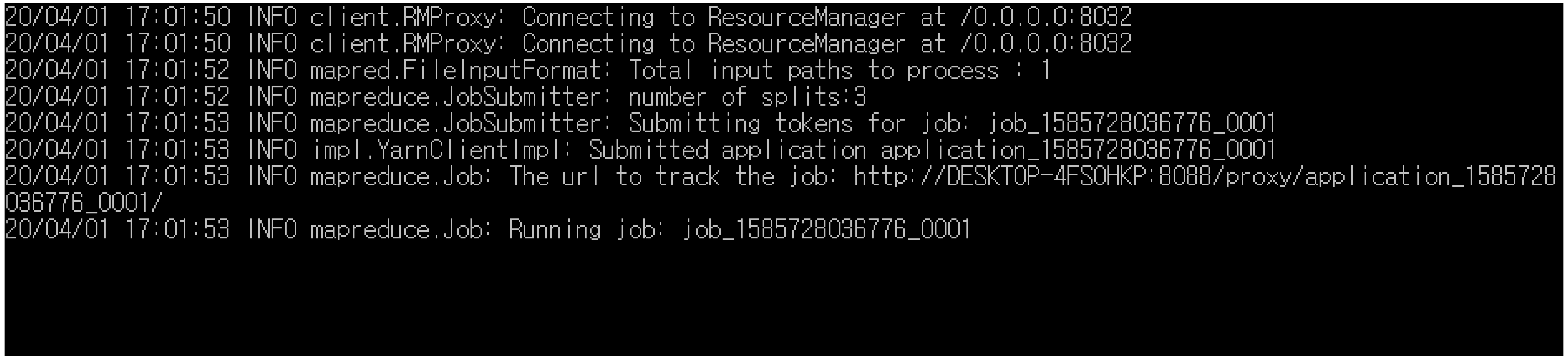
   (Windows)hdfs namenode -format

   (Linux) sudo $HADOOP_HOME/bin/hdfs namenode -format

3. **Try again.**

Hadoop Issues

➢ **ERROR : Process stopped more than a few minutes before starting Mapreduce**



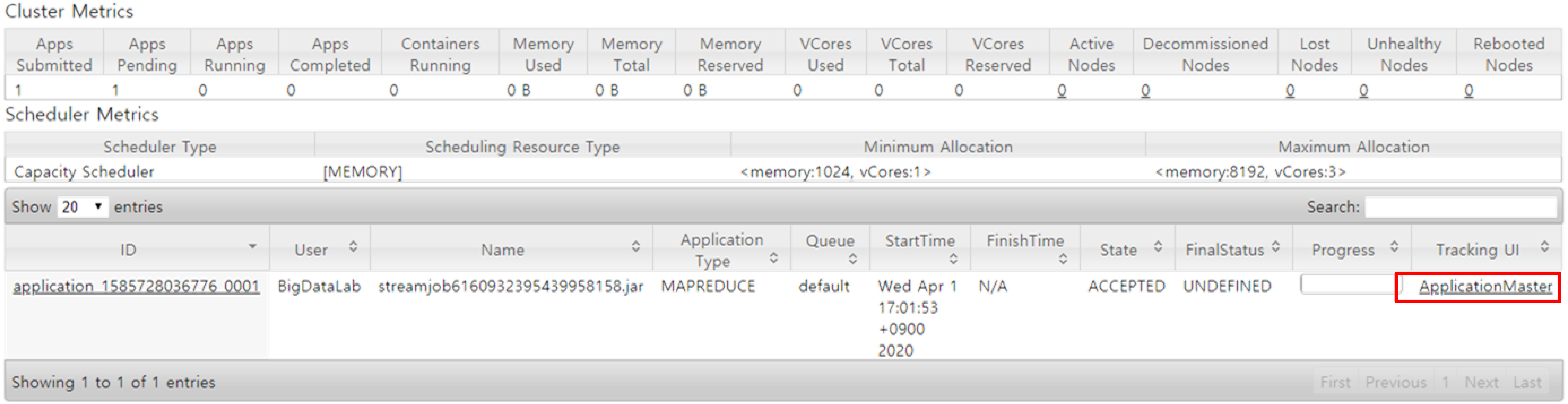
```
20/04/01 17:01:50 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
20/04/01 17:01:50 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
20/04/01 17:01:52 INFO mapred.FileInputFormat: Total input paths to process : 1
20/04/01 17:01:52 INFO mapreduce.JobSubmitter: number of splits:3
20/04/01 17:01:53 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1585728036776_0001
20/04/01 17:01:53 INFO impl.YarnClientImpl: Submitted application application_1585728036776_0001
20/04/01 17:01:53 INFO mapreduce.Job: The url to track the job: http://DESKTOP-4FSOHKP:8088/proxy/application_1585728
036776_0001/
20/04/01 17:01:53 INFO mapreduce.Job: Running job: job_1585728036776_0001
```

1. **Go to Hadoop cluster admin page**

   (Windows, Linux) http://localhost:8088/cluster

Hadoop Issues

2. **Find current application and click ApplicationMaster**



Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
1	1	0	0	0	0 B	0 B	0 B	0	0	0	0	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[MEMORY]	<memory:1024, vCores:1>	<memory:8192, vCores:3>

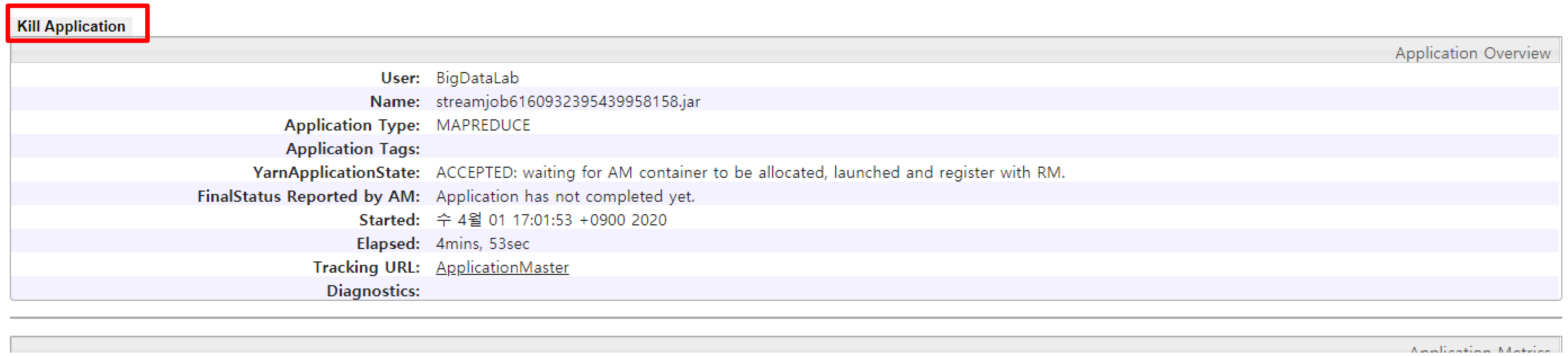
Show 20 entries

Search:

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress	Tracking UI
application_1585728036776_0001	BigDataLab	streamjob6160932395439958158.jar	MAPREDUCE	default	Wed Apr 1 17:01:53 +0900 2020	N/A	ACCEPTED	UNDEFINED		ApplicationMaster

Showing 1 to 1 of 1 entries

3. **Kill application and try again**



Application Overview

User: BigDataLab
Name: streamjob6160932395439958158.jar
Application Type: MAPREDUCE
Application Tags:
YarnApplicationState: ACCEPTED: waiting for AM container to be allocated, launched and register with RM.
FinalStatus Reported by AM: Application has not completed yet.
Started: 수 4월 01 17:01:53 +0900 2020
Elapsed: 4mins, 53sec
Tracking URL: ApplicationMaster
Diagnostics:

Application Metrics

4. **If same issue repeated, format namenode and try again**

Hadoop Issues

➢ ERROR : SAFE MODE

When dfs and yarn closed without command (sbin/stop-all),
namenode and datanode can be safe mode which inserting or deleting file is not allowed.

1. Break safe mode using command

```
(Windows) hadoop dfsadmin -safemode leave
(Linux) sudo $HADOOP_HOME/bin/hadoop dfsadmin -safemode leave
```

※ Use stop-all command when closing dfs and yarn

```
(Windows) %HADOOP_HOME%\sbin\stop-all.cmd
(Linux) sudo $HADOOP_HOME/sbin/stop-all.sh
```