

Hadoop 설치

on Ubuntu

Ubuntu 기본 설치

: 실습

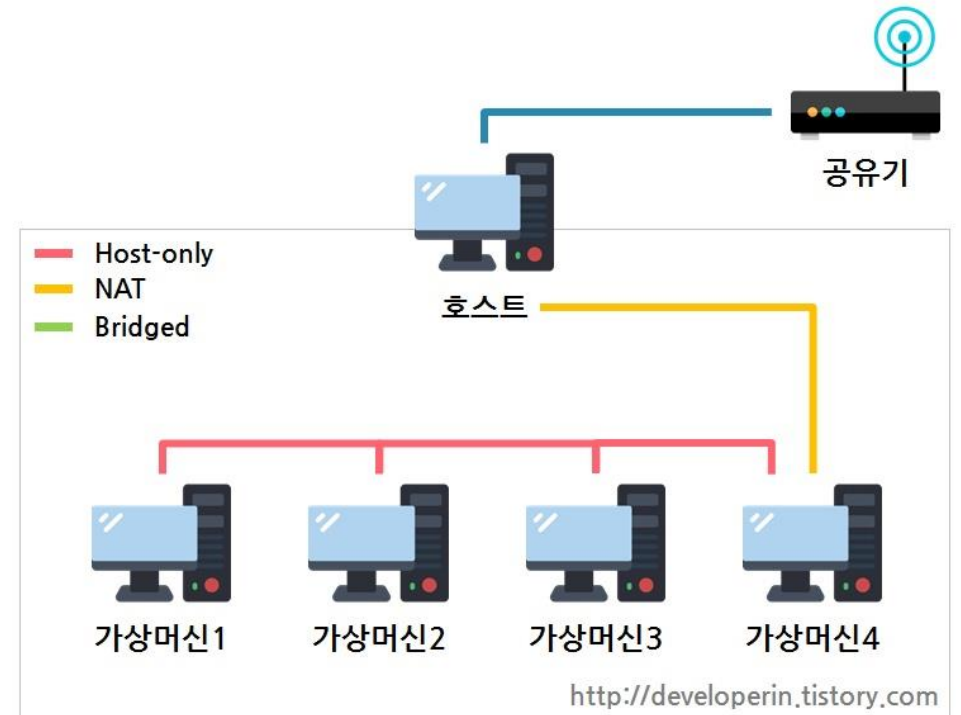
▶ 다음과 같이 VirtualBox에 Ubuntu 가상 머신을 만들어 봅니다

- ▶ 이름 : Ubuntu
- ▶ 종류 : Linux
- ▶ 버전 : Ubuntu (64-bit)
- ▶ 메모리 : 4096MB(4GB)
- ▶ 하드디스크 : 64GB(VDI, 동적 할당)

▶ 네트워크 어댑터 2개를 설정합니다.

- ▶ 어댑터 1: 호스트 전용 어댑터
- ▶ 어댑터 2: NAT 네트워크
 - ▶ 고급: 무작위 모드 - 모두 허용

▶ 설치 디스크를 넣고 가상머신을 시동하여 설치 작업을 진행합니다.



Ubuntu 기본 설치

- ▶ 설치가 끝난 후 기본적인 업데이트 작업과 기본적인 패키지 설치 작업을 진행합니다.

```
$ sudo apt update
$ sudo apt upgrade
$ sudo apt install openssh-server rdate net-tools
$ sudo apt install build-essential
```

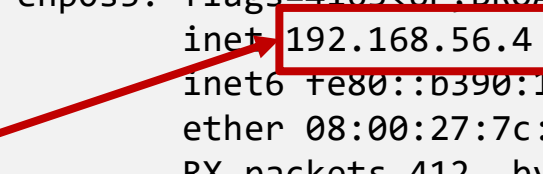
- ▶ 가상 머신의 네트워크를 확인해 봅니다.

```
$ ifconfig
```

- ▶ SSH 클라이언트로 접속해 봅니다.

```
enp0s3: flags=4163<UP,BROADCAST,RUNNING,MULTICAST> mtu 1500
    inet 192.168.56.4 netmask 255.255.255.0 broadcast 192.168.56.255
    inet6 fe80::b390:11a7:7860:1f84 prefixlen 64 scopeid 0x20<link>
    ether 08:00:27:7c:91:83 txqueuelen 1000 (Ethernet)
    RX packets 412 bytes 46078 (46.0 KB)
    RX errors 0 dropped 0 overruns 0 frame 0
    TX packets 129 bytes 18329 (18.3 KB)
    TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0

...
```



Ubuntu 기본 설치

: OpenJDK 설치

- ▶ JDK 설치(여기서는 OpenJDK 8을 설치합니다)

```
$ sudo apt search jdk # 설치할 수 있는 jdk 확인  
$ sudo apt install openjdk-8-jdk
```

- ▶ 손쉬운 접근을 위해 심볼릭 링크 설정

```
$ cd /usr/lib/jvm  
$ ln -s java-8-openjdk-amd64 jdk
```

- ▶ 환경 변수에 등록

```
export JAVA_HOME=/usr/lib/jvm/jdk
```

- ▶ 쉘에서 환경변수 적용 및 확인(~/.bashrc에 했을 경우)

```
$ source ~/.bashrc  
$ echo $JAVA_HOME
```

Hadoop 설치

- ▶ 독자 모드(Standalone mode)
 - ▶ 하둡의 기본 모드로 **HDFS**를 사용하지 않음. 다른 노드와 통신할 필요 없음
 - ▶ 테스트 및 디버깅 용도로 사용하는 모드. 일명 로컬 모드
- ▶ 가상 분산 모드(Pseudo-Distributed mode, Single Node Cluster mode)
 - ▶ 단일 노드에서 클러스터를 구성
 - ▶ 한 대의 컴퓨터에 모든 노드를 설치하여 노드(혹은 데몬)간 통신을 통해 **HDFS**를 사용
 - ▶ 실제 운영에 앞서 실제 **HDFS** 내에서의 구동을 확인하기 위한 모드
- ▶ 멀티 분산 모드(Multi-Distributed mode, Full-Distributed mode)
 - ▶ 두 대 이상의 노드를 클러스터로 묶어 구성
 - ▶ 여러 컴퓨터에 각각의 노드들을 설치하여 노드(혹은 데몬)간 통신을 통해 **HDFS**를 사용
 - ▶ 실제 운영에서 사용하는 모드
- ▶ 본 실습에서는 가상 분산 모드로 **HDFS**를 구성해 볼 예정

Hadoop 설치

: Download and Install

- ▶ <http://hadoop.apache.org>에서 링크 복사 후 다운로드

```
$ wget http://mirror.apache-kr.org/hadoop/common/hadoop-3.3.1/hadoop-3.3.1.tar.gz
```

- ▶ 압축 해제

```
$ tar zxvf hadoop-3.3.1.tar.gz
```

- ▶ 실제 실행할 위치로 이동

```
$ sudo mv hadoop-2.9.2 /usr/local/hadoop
```

```
# Hadoop-2.9.2 디렉토리를 /usr/local/Hadoop 디렉토리로 이동
```

Hadoop 설치

: 기본 설정

- ▶ /usr/local/hadoop/etc/hadoop으로 이동하여 hadoop-env.sh 파일을 수정

```
export JAVA_HOME=/usr/lib/jvm/jdk
```

- ▶ .bashrc 수정(HADOOP_HOME 정보 설정)

```
HADOOP_HOME=/usr/local/hadoop ← 하둡이 설치된 실제 경로  
PATH=$PATH:$HADOOP_HOME/bin  
PATH=$PATH:$HADOOP_HOME/sbin  
export PATH
```

- ▶ 설정 적용 및 하둡 실행 확인

```
$ source ~/.bashrc  
$ echo $HADOOP_HOME  
$ Hadoop version
```

Hadoop 설치

: 네임 노드 설정

- ▶ /usr/local/hadoop/etc/hadoop으로 이동하여 core-site.xml 파일을 수정

```
<configuration>
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://localhost:9000</value>
  </property>
  <property>
    <name>hadoop.tmp.dir</name>
    <value>/usr/local/hadoop/temp</value>
  </property>
</configuration>
```

← 가급적 설정해 주도록 합니다.
value에 설정된 디렉터리는 실제 존재해야 합니다.

Hadoop 설치

: 하둡 파일 시스템 설정

- ▶ /usr/local/hadoop/etc/hadoop으로 이동하여 hdfs-site.xml 파일을 수정

```
<configuration>
  <property>
    <name>dfs.replication</name> ← 복제본의 개수 설정
    <value>1</value>               : 현재 실행 모드는 Pseudo Distributed Mode이므로
  </property>                     1이어야 합니다.
  <property>
    <name>dfs.data.dir</name>
    <value>/usr/local/hadoop/infra/hdfs/datanode</value> ← 데이터 노드의 저장 위치
    <final>true</final>
  </property>
  <property>
    <name>dfs.name.dir</name>
    <value>/usr/local/hadoop/infra/hdfs/namenode</value> ← 네임 노드의 저장 위치
    <final>true</final>
  </property>
</configuration>
```

Hadoop 설치

: MapReduce 설정

- ▶ mapred-site.xml 파일을 수정

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

Hadoop 설치

: YARN 설정

- ▶ /usr/local/hadoop/etc/hadoop으로 이동하여 yarn-site.xml 파일을 수정

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

- ▶ Resource Manager와 Node Manager에 대한 설정을 잡는 과정

Hadoop 설치

: SSH 설정

- ▶ 현재 설치하는 방법은 가상 분산모드이므로 실제 컴퓨터가 1대일지라도 네임 노드와 데이터 노드는 네트워크를 이용하여 서로 통신을 주고 받는다
 - ▶ 서버에 로그인할 때 계정과 암호를 입력하는 절차를 생략하기 위해 상호 신뢰를 위한 키를 교환하여야 한다

```
$ cd ~/.ssh  
$ ssh-keygen -t rsa  
$ cat id_rsa.pub >> authorized_keys
```

- ▶ 로그인 없이 서버 접속이 가능한지 테스트

```
$ ssh localhost
```

Hadoop 설치

: HDFS 포맷 및 실행

- ▶ HDFS를 사용하기 위해 포맷 작업을 수행

```
$ hdfs namenode -format
```

- ▶ 하둡 데몬 실행

```
$ start-dfs.sh  
# 노드 연결을 위해 yes 입력(처음 실행시에만 요구함)  
$ start-yarn.sh
```

- ▶ 데몬 실행 확인

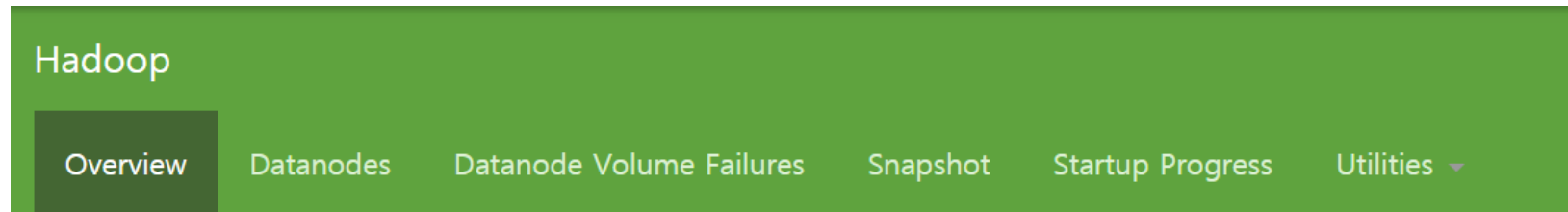
```
$ jps  
11991 DataNode  
12809 Jps  
12618 ResourceManager  
11836 NameNode  
12780 NodeManager  
12174 SecondaryNameNode
```

Hadoop 설치

: Web Interface의 활용

▶ Hadoop을 활용하기 위한 Web Interface

```
http://localhost:9870 # HDFS 네임노드 확인  
http://localhost:8042 # YARN의 노드 매니저 웹 인터페이스  
http://localhost:8088 # YARN 리소스 매니저의 웹 인터페이스(구 Job Tracker)
```



Overview 'localhost:9000' (active)

Started:	Tue Nov 05 17:15:59 +0900 2019
Version:	2.8.5, r0b8464d75227fcee2c6e7f2410377b3d53d3d5f8
Compiled:	Mon Sep 10 12:32:00 +0900 2018 by jdu from branch-2.8.5