

의사(모조) 분산모드 실습

라. 실습2 - Single Node Cluster(모조 분산 모드)

1) ssh 설치

ssh 설치

```
yum install openssh*
```

ssh 서비스 실행

```
/usr/sbin/sshd
```

비밀번호를 생략한 ssh 로그인 설정
공개키와 비밀키 생성

```
ssh-keygen -t rsa -P ""
```

공개키를 ssh의 인증키로 등록

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

인증키의 permission 설정(생략 가능)

```
chmod 600 ~/.ssh/authorized_keys
```

ssh 접속 확인

```
ssh localhost
```

하둡 자체가 분산처리 시스템이기 때문에 여러대의 컴퓨터가 데이터를 주고받게 된다. 따라서 컴퓨터간 신뢰관계가 필요하게 된다.

보안통신인 SSH(secure shell)을 사용한다.

비밀번호를 사용해서 계속해서 로그인을 하기보다는 공개키와 비밀키를 자동으로 생성하는 기능이 ssh에서 제공되기 때문에 이를 실행할 것이다.

```
ssh-keygen -t rsa -P ""
```

```
[root@localhost ~]# ssh-keygen -t rsa -P ""
Generating public/private rsa key pair.
Enter file in which to save the key (/root/.ssh/id_rsa):
```

공개키와 비밀키를 어디에 저장할지 물어본다.

기본값으로 실습하도록 하자.

```
[root@localhost ~]# ssh-keygen -t rsa -P ""
Generating public/private rsa key pair.
Enter file in which to save the key (/root/.ssh/id_rsa):
Created directory '/root/.ssh'.
Your identification has been saved in /root/.ssh/id_rsa.
Your public key has been saved in /root/.ssh/id_rsa.pub.
The key fingerprint is:
SHA256: Q6ba+nM1T2hjn13o7svf5mSBKM6NDK2Zg2yKx0D1Ygg root@localhost.localdomain
The key's randomart image is:
+---[RSA 2048]-----+
|
|      o
|      +
| E . . . So o . o |
| = o oo. O.@ . . o|
| = o.= = O B + . o|
| . o + . . o +.ooo|
| . . . . . o o==+|
+---[SHA256]-----+
[root@localhost ~]#
```

실행할때마다 다른 키가 설정된다.

이것을 인증키로 등록을 해야한다.

```
[root@localhost ~]# cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
```

방금 생성된 rsa.pub을 authorized_keys 즉, 인증키에 추가.

리눅스의 사용 권한

Read 4

Write 2

execute 1

현재사용자 / 그룹사용자 / 기타사용자
6 0 0

이는,

read는 읽기권한

write는 쓰기권한

execute는 실행권한이다.

```
chmod 600 ~/.ssh/authorized_keys
```

명령같은 경우는 현재사용자는 쓰기 읽기가 가능하고, 그룹사용자, 기타사용자는 아무 권한을 갖지못하게 설정하는 것이다.

```
[root@localhost ~]# ssh localhost
The authenticity of host 'localhost (:::1)' can't be established.
ECDSA key fingerprint is SHA256: zwUDu9GV9KWYKKTW9X9vXrHLKxD83STLj7YctD8vNLI.
ECDSA key fingerprint is MD5: 44: 0b: e0: 1e: 88: f2: ff: e8: 39: 8e: a5: 32: 6b: 96: 52: 5f.
Are you sure you want to continue connecting (yes/no)? █
```

처음접속이기 때문에 key를 물어본다. 이러한 사용자가 접속을 했는데 허용할까요?라고한다.
이때 yes를 해야만 다음부터 비밀번호를 물어보지 않는다.

```
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
Last login: Fri May 31 00:53:25 2019
[root@localhost ~]#
```

승인을 해주면 permanently 즉, 영구적으로 접속가능자로 등록되었다고 나온다.

```
[root@localhost ~]# ssh localhost
Last login: Fri May 31 01:41:10 2019 from localhost
[root@localhost ~]# █
```

따라서 다시 접속을하면 바로 접속이 된다.

2) hadoop-env.sh 수정

```
gedit $HADOOP_HOME/etc/hadoop/hadoop-env.sh
```

25번 라인 JDK 경로 수정 : **export JAVA_HOME=/usr/local/jdk1.8**

텍스트 편집기의 줄 번호가 표시되도록 설정(텍스트 편집기 - 기본 설정 - 줄 번호 표시 체크)

하둡은 java위에서 돌아가는 프로그램이기 때문에, 자바 홈이 어디인지를 설정해주어야한다.

```
# The java implementation to use.
export JAVA_HOME=/usr/local/jdk1.8|
```

sh ▼ 탭 너비: 8 ▼

25행, 35열

다음을 설정하지 않으면 실행이 되지 않는다.

3) core-site.xml 수정(네임노드를 설정하는 파일)

```
gedit $HADOOP_HOME/etc/hadoop/core-site.xml
```

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

네임노드가 어떤것인지를 설정해야한다.

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

입력

4) hdfs-site.xml 수정(파일 복제 옵션)

```
gedit $HADOOP_HOME/etc/hadoop/hdfs-site.xml
```

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
</configuration>
```

파일을 몇개의 파일로 복제를 할 것인가를 설정하는 것이다.

즉, 파일을 몇덩어리로 나눌것인가. 이다.

일단 하나의 컴퓨터에서 테스트를 진행할 것이기 때문에, 1로 하도록 한다.

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

<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
</configuration>
```

5) 네임노드 포맷

```
hdfs namenode -format
```



6) 하둡 클러스터 시작

```
start-dfs.sh
```

```
[root@localhost ~]# start-dfs.sh
Starting namenodes on [localhost]
localhost: starting namenode, logging to /home/centos/hadoop-2.9.2/logs/hadoop-root@localhost.localdomain.out
localhost: starting datanode, logging to /home/centos/hadoop-2.9.2/logs/hadoop-root@localhost.localdomain.out
Starting secondary namenodes [0.0.0.0]
The authenticity of host '0.0.0.0 (0.0.0.0)' can't be established.
ECDSA key fingerprint is SHA256: zwUDu9GV9KWYKKTW9X9vXrHLKxD83STLj7YctD8vNLI.
ECDSA key fingerprint is MD5: 44: 0b: e0: 1e: 88: f2: ff: e8: 39: 8e: a5: 32: 6b: 96: 52: 5f.
Are you sure you want to continue connecting (yes/no)? █
```

ssh에서 키를 만들었기 때문에, datanode와 namenode간에 신뢰관계를 형성해야한다.
따라서 이를 허락받는 과정이다.
yes를 입력하자.

7) 프로세스 확인

```
jps
```

```
[root@localhost ~]# jps
3380 NameNode
3508 DataNode
3672 SecondaryNameNode
3789 Jps
```

```
[root@localhost ~]# jps
65185 Jps
64724 NameNode
64820 DataNode
65060 SecondaryNameNode
[root@localhost ~]# █
```

현재 컴퓨터는 한대지만, 3개의 서버장치가 동시에 구동되듯이 실행되고 있다는 뜻이다.

8) 웹브라우저에서 확인

<http://localhost:50070/>



Overview 'localhost:9000' (active)	
Name:	Tue May 13 02:08:43 +0900 2019
Version:	2.9.2, r826afbeae31ca687bc2f8471dc841b66ed2c6704
Compiled:	Tue Nov 13 21:42:00 +0900 2018 by ajsaka from branch-2.9.2
Cluster ID:	CID-e4834164-f153-490a-8d73-3e6cb47662a1
Block Pool ID:	BP-2005977525-127.0.0.1-1559236112239

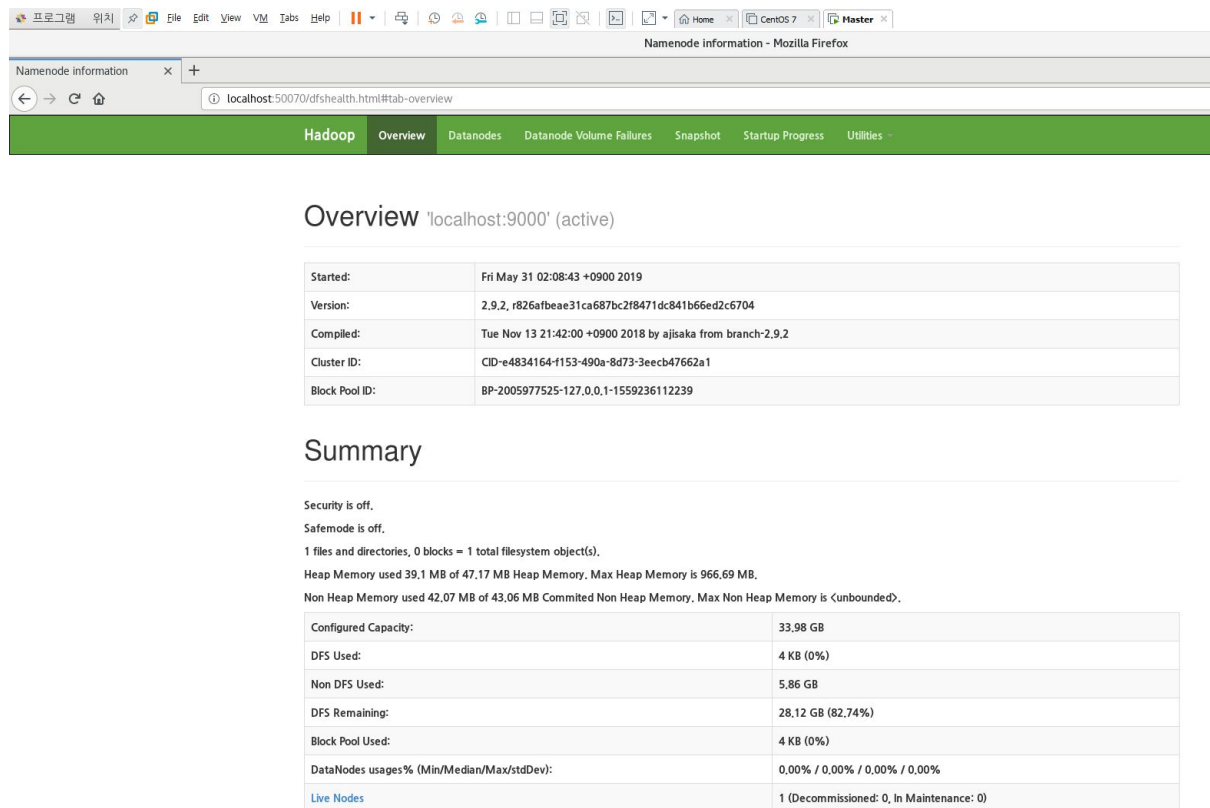
Summary	
Security is off.	
1 files and directories, 0 blocks = 1 total filesystem object(s).	
Heap Memory used 39.1 MB of 47.17 MB Heap Memory. Max Heap Memory is 966.69 MB.	
Non Heap Memory used 42.07 MB of 43.06 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.	

Resource Usage	
Configured Capacity:	33.98 GB
DFS Used:	4 KB (0%)
Non DFS Used:	5.86 GB
DFS Remaining:	28.12 GB (82.74%)
Block Pool Used:	4 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%

9) ResourceManager와 NodeManager 시작

```
start-yarn.sh
```

firefox에서 50070포트로 들어가면, hadoop의 설정을 볼 수 있다.



Overview 'localhost:9000' (active)	
Started:	Fri May 31 02:08:43 +0900 2019
Version:	2.9.2, r826afbeae31ca687bc2f8471dc841b66ed2c6704
Compiled:	Tue Nov 13 21:42:00 +0900 2018 by ajsaka from branch-2.9.2
Cluster ID:	CID-e4834164-f153-490a-8d73-3e6cb47662a1
Block Pool ID:	BP-2005977525-127.0.0.1-1559236112239

Summary	
Security is off.	
Safemode is off.	
1 files and directories, 0 blocks = 1 total filesystem object(s).	
Heap Memory used 39.1 MB of 47.17 MB Heap Memory. Max Heap Memory is 966.69 MB.	
Non Heap Memory used 42.07 MB of 43.06 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.	

Resource Usage	
Configured Capacity:	33.98 GB
DFS Used:	4 KB (0%)
Non DFS Used:	5.86 GB
DFS Remaining:	28.12 GB (82.74%)
Block Pool Used:	4 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%

```
[root@localhost ~]# start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /home/centos/hadoop-2.9.2/logs/yarn-root-resourc
anager-localhost.localdomain.out
localhost: starting nodemanager, logging to /home/centos/hadoop-2.9.2/logs/yarn-root-
demanager-localhost.localdomain.out
[root@localhost ~]#
```

```
[root@localhost ~]# jps
65745 NodeManager
65650 ResourceManager
64724 NameNode
64820 DataNode
65060 SecondaryNameNode
66458 Jps
[root@localhost ~]#
```

ResourceManager가 실행중인 것을 알 수 있다.

10) 웹브라우저에서 ResourceManager 실행 확인

<http://localhost:8088>

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed
0	0	0	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Dec
1	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type
Capacity Scheduler	[MEMORY]

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime
----	------	------	------------------	-------	----------------------	-----------	------------

11) 분석 프로그램 실행(wordcount)

hadoop-env.sh 파일의 단어 갯수 분석

맵리듀스 job을 실행하기 위해서는 HDFS 디렉토리를 만들어야 함

```
hdfs dfs -mkdir /user
hdfs dfs -mkdir /user/root
hdfs dfs -mkdir /user/root/conf
```

```
hdfs dfs -mkdir /input

hdfs dfs -copyFromLocal /home/centos/hadoop-2.9.2/README.txt /input

hdfs dfs -ls /input

rm -rf wordcount_output

hadoop jar
$HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.
jar wordcount /input/README.txt ~/wordcount-output

hdfs dfs -ls ~/wordcount-output
```

이제 한대에서 설정을 끝낸, 모조 분산처리를 실습할 예정인데,
그전에 기본 작업 디렉토리를 만들어야한다.

```
hdfs dfs -mkdir /user
```

```
hdfs dfs -mkdir /user/root
```

```
hdfs dfs -mkdir /user/root/conf
```

mkdir =>make directory이다.

즉, 이러한 폴더들을 만들고

```
hdfs dfs -copyFromLocal /home/centos/hadoop-2.9.2/README.txt /input
```

local에 있는 README.txt파일을 하둡에 /input에 올리겠다.

즉, local에 있는 파일을 hadoop 분산파일 시스템에 올리겠습니다.

```
hdfs dfs -ls /input
```

올라갔는지 확인해 보고


```
rm -rf wordcount_output
```

아까전에 실행했던 wordcount파일을 삭제한 후

```
hadoop jar  
$HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.  
jar wordcount /input/README.txt ~/wordcount-output
```

```
hdfs dfs -ls ~/wordcount-output
```

```
hadoop jar  
$HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar  
wordcount /input/README.txt ~/wordcount-output
```

local이 아니라, /input즉, 하둡에서 가져와 실행을 하고, 실행한다.
그리고 목록확인

실행결과 확인

```
hdfs dfs -cat ~/wordcount-output/part-r-00000
```

```
(BIS), 1  
(ECCN) 1  
(TSU) 1  
(see 1  
5D002.C.1, 1  
740.13) 1  
<http://www.wassenaar.org/> 1  
Administration 1  
Apache 1
```