hadoop 구축

내용유무	☑
FE/BE/INFRA	data
작성자	승엽 종호
중요도	⭐⭐⭐⭐⭐

hadoop 설치

```
wget https://dlcdn.apache.org/hadoop/common/hadoop-3.3.5/hadoop-3.3.5.tar.gz

tar -xvzf hadoop-3.3.5.tar.gz
```

hadoop 설정 파일 작성

- 아래 실행하는 모든 파일 전부 복사 권장
- master, slave1, slave2 일괄 적용
  - core-site.xml

```
<configuration>
    <property>
        <name>fs.defaultFS</name>
        <value>hdfs://마스터url:설정포트</value>
    </property>
</configuration>
```

  - mapred-site.xml

```
<configuration>
    <property>
        <name>mapreduce.framework.name</name>
        <value>yarn</value>
    </property>
    <property>
        <name>mapreduce.jobhistory.address</name>
        <value>마스터url:설정포트</value>
    </property>
    <property>
        <name>mapreduce.jobhistory.webapp.address</name>
        <value>마스터url:설정포트</value>
    </property>
</configuration>
```

  - yarn-site.xml

```
<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_H
        </property>
        <property>
            <name>yarn.resourcemanager.hostname</name>
            <value>마스터url</value>
        </property>
        <property>
            <name>yarn.resourcemanager.address</name>
            <value>마스터url:설정포트</value>
        </property>
```

```
        <property>
            <name>yarn.resourcemanager.webapp.address</name>
            <value>마스터url:설정포트</value>
        </property>
</configuration>
```

- master
  - hdfs-site.xml

```
<configuration>
    <property>
        <name>dfs.replication</name>
        <value>3</value>
    </property>
    <property>
        <name>dfs.namenode.name.dir</name>
        <value>/home/ubuntu/data/hdfs/namenode</value>
    </property>
    <property>
        <name>dfs.datanode.data.dir</name>
        <value>/home/ubuntu/data/hdfs/datanode</value>
    </property>
    <property>
        <name>dfs.namenode.http-address</name>
        <value>마스터url:설정포트</value>
    </property>
    <property>
        <name>dfs.namenode.secondary.http-address</name>
        <value>세컨더리 네임노드 서버(슬레이브1)url:설정포트</value>
    </property>
    <property>
        <name>dfs.client.use.datanode.hostname</name>
        <value>true</value>
    </property>
</configuration>
```

  - workers

```
$ cd $HADOOP_CONF_DIR

$ vim workers
# 이하 입력 후 저장

master url
slave1 url
slave2 url
```

- slave1, slave2
  - hdfs-site.xml

```
<configuration>
    <property>
        <name>dfs.replication</name>
        <value>3</value>
    </property>
    <property>
        <name>dfs.namenode.name.dir</name>
        <value>/home/ubuntu/data/hdfs/namenode</value>
    </property>
    <property>
        <name>dfs.datanode.data.dir</name>
        <value>/home/ubuntu/data/hdfs/datanode</value>
    </property>
</configuration>
```

트러블 슈팅

1. slave1, slave2의 datanode가 활성화되지 않음
   a. master의 서버 설정을 localhost로 하여 발생한 문제로 예상
      i. etc/hosts 에 ip와 서버 도메인 기록 후 설정 변경

2. 방화벽 문제
    a. datanode는 활성화 되었지만 접근이 불가능했음
        i. 설정 포트를 모두 열었지만 timeout error
        ii. netstat -nltp를 통해 사용하는 포트를 확인하니 20000~50000번대의 포트가 랜덤으로 사용하는 것 발견
        iii. ufw allow from [아이피]를 통해 클러스터 간 포트 개방
            1. 해결되어 여러번 시도를 통해 사용하지 않는 포트 삭제