Spark

engineer / 1234

hadoop 설치

```
sudo apt update -y
sudo apt upgrade -y

sudo apt install vim -y
sudo apt install openssh-server ssh-askpass -y

###### 굳이 안해도 됨! #####
sudo vim /etc/hostname

engineer

sudo vim /etc/hosts

engineer

reboot
##########################

ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys

# java 설치
wget https://corretto.aws/downloads/latest/amazon-corretto-11-x64-linux-
jdk.tar.gz
tar xvzf amazon-corretto-11-x64-linux-jdk.tar.gz
mv amazon-corretto-11.0.12.7.1-linux-x64 java

# hadoop 설치
wget https://dlcdn.apache.org/hadoop/common/hadoop-3.3.1/hadoop-3.3.1.tar.gz
tar xvzf hadoop-3.3.1.tar.gz
mv hadoop-3.3.1 hadoop

# path 등록
sudo vim ~/.bashrc

# java
export JAVA_HOME=/home/engineer/java
export PATH=$PATH:$JAVA_HOME/bin

# hadoop
export HADOOP_HOME=/home/engineer/hadoop
export HADOOP_CONF_DIR=/home/engineer/hadoop/etc/hadoop
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin

# hadoop user
export HDFS_NAMENODE_USER=engineer
export HDFS_DATANODE_USER=engineer
```

```
export HDFS_SECONDARYNAMENODE_USER=engineer
export YARN_RESOURCEMANAGER_USER=engineer
export YARN_NODEMANAGER_USER=engineer

source ~/.bashrc
# 확인
java -version
javac -version

# hadoop 설정
cd $HADOOP_CONF_DIR
vim hadoop-env.sh
export JAVA_HOME=/home/engineer/java
export HADOOP_HOME=/home/engineer/hadoop
export HADOOP_CONF_DIR=/home/engineer/hadoop/etc/hadoop
```

hadoop 설정

`vim core-site.xml`

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

`vim hdfs-site.xml`

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

`mapred-site.xml`

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

```
# format
hdfs namenode -format
hdfs datanode -format

# 실행
start-dfs.sh
start-yarn.sh
```

```
jps
hdsf dfsadmin -report

# browser 열어서
localhost:9870

stop-dfs.sh
stop-yarn.sh
```

spark 설치

```
sudo apt install python3-pip -y
pip install numpy

wget https://dlcdn.apache.org/spark/spark-3.1.2/spark-3.1.2-bin-without-
hadoop.tgz
tar xvzf spark-3.1.2-bin-without-hadoop.tgz
mv spark-3.1.2-bin-without-hadoop spark

# path
sudo vim ~/.bashrc

# spark
export SPARK_HOME=/home/engineer/spark
export PATH=$PATH:$SPARK_HOME/bin
export PATH=$PATH:$SPARK_HOME/sbin
export SPARK_DIST_CLASSPATH=$(${HADOOP_HOME}/bin/hadoop classpath)

source ~/.bashrc

cd $SPARK_HOME/conf
cp spark-env.sh.template spark-env.sh
vim spark-env.sh

export JAVA_HOME=/home/engineer/java
export HADOOP_CONF_DIR=/home/engineer/hadoop/etc/hadoop
export YARN_CONF_DIR=/home/engineer/hadoop/etc/hadoop
export SPARK_DIST_CLASSPATH=$(/home/engineer/hadoop/bin/hadoop classpath)

cp workers.template workers
cp spark-defaults.conf.template spark-defaults.conf

spark.master            yarn


start-dfs.sh
start-yarn.sh

pyspark
```