NTUA, ECE, ATDS, Installation Guide

In both VMs:

```
mkdir ./opt
mkdir ./opt/bin
wget https://dlcdn.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz
tar -xvzf hadoop-3.3.6.tar.gz
mv hadoop-3.3.6 ./opt/bin
wget https://dlcdn.apache.org/spark/spark-3.5.0/spark-3.5.0-bin-hadoop3.tgz
tar -xvzf spark-3.5.0-bin-hadoop3.tgz
mv ./spark-3.5.0-bin-hadoop3 ./opt/bin/
cd ./opt
ln -s ./bin/hadoop-3.3.6/ ./hadoop
ln -s ./bin/spark-3.5.0-bin-hadoop3/ ./spark
cd
rm hadoop-3.3.6.tar.gz
rm spark-3.5.0-bin-hadoop3.tgz
```

```
sudo nano ~/.bashrc
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64  #Value should match: dirname $(dirname $(readlink -f $(which java)))
export HADOOP_HOME=/home/user/opt/hadoop
export SPARK_HOME=/home/user/opt/spark
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME

export HADOOP_COMMON_HOME=$HADOOP_HOME

export HADOOP_HDFS_HOME=$HADOOP_HOME

export HADOOP_YARN_HOME=$HADOOP_HOME

export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native

export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin:$SPARK_HOME/bin;

export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop

export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"

export LD_LIBRARY_PATH=/home/ubuntu/opt/hadoop/lib/native:$LD_LIBRARY_PATH

export PYSPARK_PYTHON=python3

source ~/.bashrc
```

```
sudo nano $HADOOP_HOME/etc/hadoop/hadoop-env.sh

export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
```

```
sudo nano $HADOOP_HOME/etc/hadoop/core-site.xml

<?xml version="1.0" encoding="UTF-8"?>

<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>

    <property>

        <name>hadoop.tmp.dir</name>

        <value>/home/user/opt/data/hadoop</value>

        <description>Parent directory for other temporary directories.</description>
    </property>
    <property>
        <name>fs.defaultFS </name>
        <value>hdfs://master:54310</value>
        <description>The name of the default file system. </description>
    </property>
</configuration>
```

```
sudo nano $HADOOP_HOME/etc/hadoop/hdfs-site.xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
    <property>
        <name>dfs.replication</name>
        <value>1</value>
        <description>Default block replication.</description>
    </property>
    <property>
        <name>dfs.datanode.data.dir</name>
        <value>/home/user/opt/data/hdfs</value>
    </property>
</configuration>
```

```
sudo vim $HADOOP_HOME/etc/hadoop/workers
master
worker
```

```
$HADOOP_HOME/bin/hdfs namenode -format
start-dfs.sh
```

```
sudo nano $HADOOP_HOME/etc/hadoop/yarn-site.xml
<?xml version="1.0"?>
<configuration>
<!-- Site specific YARN configuration properties -->
    <property>
        <name>yarn.resourcemanager.hostname</name>
        <value>master</value>
    </property>
    <property>
        <name>yarn.resourcemanager.webapp.address</name>
        <!--Insert the public IP of your master machine here-->
        <value>83.212.80.178:8088</value>
    </property>
    <property>
        <name>yarn.nodemanager.resource.memory-mb</name>
        <value>6144</value>
    </property>
```

```
 <property>
     <name>yarn.scheduler.maximum-allocation-mb</name>
     <value>6144</value>
 </property>
 <property>
     <name>yarn.scheduler.minimum-allocation-mb</name>
     <value>128</value>
 </property>
 <property>
     <name>yarn.nodemanager.vmem-check-enabled</name>
     <value>false</value>
 </property>
<property>
     <name>yarn.nodemanager.aux-services</name>
     <value>mapreduce_shuffle,spark_shuffle</value>
 </property>
 <property>
     <name>yarn.nodemanager.aux-services.mapreduce_shuffle.class</name>
     <value>org.apache.hadoop.mapred.ShuffleHandler</value>
 </property>
 <property>
     <name>yarn.nodemanager.aux-services.spark_shuffle.class</name>
     <value>org.apache.spark.network.yarn.YarnShuffleService</value>
 </property>
```

```
  <property>
    <name>yarn.nodemanager.aux-services.spark_shuffle.classpath</name>
    <value>/home/user/opt/spark/yarn/*</value>
  </property>
</configuration>
```

```
sudo vim $SPARK_HOME/conf/spark-defaults.conf
spark.eventLog.enabled                  true
spark.eventLog.dir                      hdfs://master:54310/spark.eventLog
spark.history.fs.logDirectory           hdfs://master:54310/spark.eventLog
spark.master                            yarn
spark.submit.deployMode                 client
spark.driver.memory                     1g
spark.executor.memory                   1g
spark.executor.cores                    1
```

Start cluster:

```
start-dfs.sh
start-yarn.sh
hadoop fs -mkdir /spark.eventLog
$HADOOP_HOME/bin/hdfs namenode -format
$SPARK_HOME/sbin/start-history-server.sh
```

Stop cluster:

```
stop-dfs.sh

stop-yarn.sh

$SPARK_HOME/sbin/stop-history-server.sh
```

Data:

```
scp -r ~/datasets user@worker:.

hadoop fs -mkdir hdfs://master:54310/datasets

hadoop fs -mkdir hdfs://master:54310/datasets/income

hadoop fs -put datasets/Crime_Data_from_2010_to_2019.csv hdfs://master:54310/datasets/.

hadoop fs -put datasets/Crime_Data_from_2020_to_Present.csv hdfs://master:54310/datasets/.

hadoop fs -put datasets/revgecoding.csv hdfs://master:54310/datasets/.

hadoop fs -put datasets/LAPD_Police_Stations.csv hdfs://master:54310/datasets/.

hadoop fs -put datasets/income/LA_income_2015.csv hdfs://master:54310/datasets/income/.

hadoop fs -put datasets/income/LA_income_2017.csv hdfs://master:54310/datasets/income/.

hadoop fs -put datasets/income/LA_income_2019.csv hdfs://master:54310/datasets/income/.

hadoop fs -put datasets/income/LA_income_2021.csv hdfs://master:54310/datasets/income/.
```

WSL:

```
sudo nano /etc/resolv.conf
```

```
nameserver 8.8.8.8
```

```
sudo systemctl restart systemd-resolved.service
```

Okeanos:

```
ssh user@snf-40260.ok-kno.grnetcloud.net
```

Passwordless SSH:

```
ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
scp -r ~/.ssh/ user@worker:~/
```

Job Submission:

```
cd queries
spark-submit dataframe.py
```

Check Hadoop:

```
hadoop version
```

Check Spark:

```
spark-submit --version
spark-shell --version
spark-sql --version
```

Check python:

```
python3.8 --version
```

Check nodes:

```
jps && ssh worker jps
```

Big Datasets Github Upload:

```
git bash  ../advanced_topics_in_database_systems

git lfs install

git lfs track "Crime_Data_from_2010_to_2019.csv"

git lfs push --all origin main

git add .

git push -u origin main

git commit -m "Crime_Data_from_2010_to_2019.csv"
```