Setting Up Hadoop

You have to go into your hadoop user on your master machine

.profile setting

Edit /home/hadoop/.profile

```
PATH=/home/hadoop/hadoop/bin:/home/hadoop/hadoop/sbin:$PATH
PATH=/home/hadoop/spark/bin:$PATH
export HADOOP_CONF_DIR=/home/hadoop/hadoop/etc/hadoop
export SPARK_HOME=/home/hadoop/spark
export LD_LIBRARY_PATH=/home/hadoop/hadoop/lib/native:$LD_LIBRARY_PATH
export PATH=$PATH:$SPARK_HOME/bin
export HADOOP_CONF_DIR=/opt/hadoop/etc/hadoop
export LD_LIBRARY_PATH=/opt/hadoop/lib/native:$LD_LIBRARY_PATH
```

.bashrc settings

Edit /home/hadoop/.bashrc

```
# java
export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk
export PATH=$JAVA_HOME/bin:$PATH
# hadoop
export HADOOP_HOME=/home/hadoop/hadoop
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_CONF_DIR=${HADOOP_HOME}/etc/hadoop
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
export HADOOP_OPTS=-Djava.net.preferIOv4stack=true
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"
```

Install Java and Hadoop

```
sudo apt-get install openjdk-8-jdk
wget https://dlcdn.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz
tar -xzf hadoop-3.3.6.tar.gz
mv hadoop-3.3.6 hadoop
```

Editing the Hadoop Settings

~/hadoop/etc/hadoop/hadoop-env.sh

```
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64/jre
```

hadoop/etc/hadoop/core-site.xml

```
<configuration>
        <property>
                <name>fs.defaultFS</name>
                <value>hdfs://node-master:9000</value>
        </property>
</configuration>
```

~/hadoop/etc/hadoop/hdfs-site.xml

```
<configuration>
        <property>
                <name>dfs.namenode.name.dir</name>
                <value>/home/hadoop/data/nameNode</value>
        </property>

        <property>
                <name>dfs.datanode.data.dir</name>
                <value>/home/hadoop/data/dataNode</value>
        </property>

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

~/hadoop/etc/hadoop/mapred-site.xml

```
<configuration>
        <property>
                <name>mapreduce.framework.name</name>
                <value>yarn</value>
        </property>
        <property>
                <name>yarn.app.mapreduce.am.env</name>
                <value>HADOOP_MAPRED_HOME=$HADOOP_HOME</value>
```

```
	</property>
	<property>
		<name>mapreduce.map.env</name>
		<value>HADOOP_MAPRED_HOME=$HADOOP_HOME</value>
	</property>
	<property>
		<name>mapreduce.reduce.env</name>
		<value>HADOOP_MAPRED_HOME=$HADOOP_HOME</value>
	</property>
	<property>
		<name>yarn.app.mapreduce.am.resource.mb</name>
		<value>512</value>
	</property>
	<property>
		<name>mapreduce.map.memory.mb</name>
		<value>256</value>
	</property>
	<property>
		<name>mapreduce.reduce.memory.mb</name>
		<value>256</value>
	</property>
</configuration>
```

~/hadoop/etc/hadoop/yarn-site.xml
The ip is the ip address of the node-master

```
<configuration>
	<property>
	<name>yarn.acl.enable</name>
	<value>0</value>
	</property>

	<property>
	<name>yarn.resourcemanager.hostname</name>
	<value>192.168.0.230</value>
	</property>

	<property>
	<name>yarn.nodemanager.aux-services</name>
	<value>mapreduce_shuffle</value>
	</property>
	<property>
	<name>yarn.nodemanager.resource.memory-mb</name>
	<value>1536</value>
	</property>

	<property>
	<name>yarn.scheduler.maximum-allocation-mb</name>
	<value>1536</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>
</configuration>
```

~/hadoop/etc/hadoop/workers

```
isel-slave1
Isel-slave2
```

```
scp hadoop-*.tar.gz isel-slave1:/home/hadoop
scp hadoop-*.tar.gz isel-slave2:/home/hadoop
```

```
tar -xzf hadoop-3.3.6.tar.gz
mv hadoop-3.3.6 hadoop
```

```
for node in isel-slave1 isel-slave2; do
   scp ~/hadoop/etc/hadoop/* $node:/home/hadoop/hadoop/etc/hadoop/;
done
```

Setting up Spark

Download Spark

```
wget https://dlcdn.apache.org/spark/spark-3.4.3/spark-3.4.3-bin-hadoop3.tgz
tar -xvf spark-3.4.3-bin-hadoop3.tgz
mv spark-3.4.3-bin-hadoop3 spark
```

Start the server

```
start-dfs.sh
start-yarn.sh
```

Yarn
http://192.168.0.230:8088
Node master web ui
192.168.0.230:9870

```
mv $SPARK_HOME/conf/spark-defaults.conf.template $SPARK_HOME/conf/spark-defaults.conf
```

Edit `$SPARK_HOME/conf/spark-defaults.conf` and set `spark.master` to `yarn`:

```
spark.master     yarn
spark.driver.memory      1g
spark.yarn.am.memory  512m
spark.executor.memory             512m
spark.eventLog.enabled  true
spark.eventLog.dir hdfs://isel-master:9000/spark-logs
spark.history.provider                org.apache.spark.deploy.history.FsHistoryProvider
spark.history.fs.logDirectory   hdfs://isel-master:9000/spark-logs
spark.history.fs.update.interval  10s
spark.history.ui.port       18080
```

```
hdfs dfs -mkdir -p /user/hadoop
hdfs dfs -mkdir /spark-logs
cd /home/hadoop
wget -O alice.txt https://www.gutenberg.org/files/11/11-0.txt
hdfs dfs -mkdir inputs
hdfs dfs -put alice.txt inputs
```

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

http://isel-master:18080

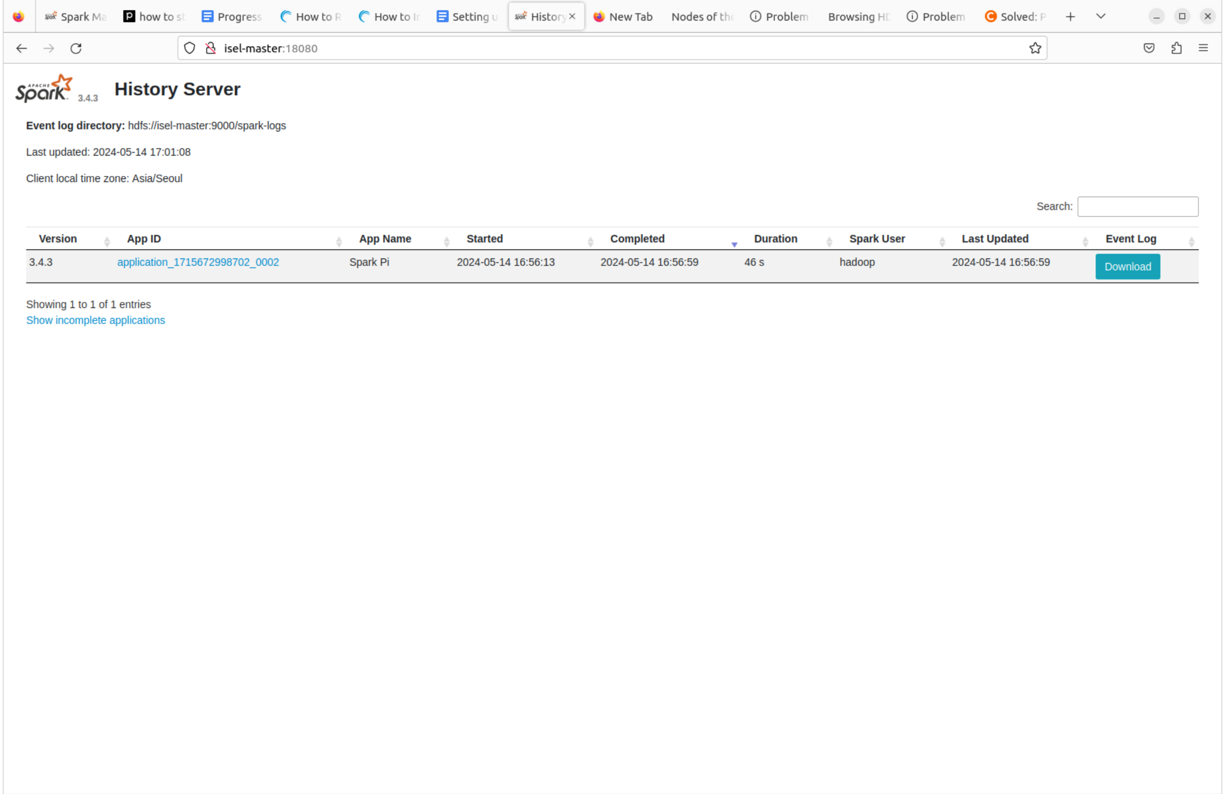
```
hdfs namenode -format
```

```
curl -sS 'http://isel-master:14000/webhdfs/v1?op=gethomedirectory&user.name=hdfs'
```

```
spark-submit --deploy-mode client \
               --class org.apache.spark.examples.SparkPi \
               $SPARK_HOME/examples/jars/spark-examples_2.11-2.2.0.jar 10
spark-submit --deploy-mode client \
             --class org.apache.spark.examples.SparkPi \ $SPARK_HOME/examples/jars/spark-examples_2.12-3.4.3.jar 2
```

```
spark-submit --deploy-mode cluster $SPARK_HOME/examples/src/main/python/pi.py 5
```

```
spark-submit --deploy-mode cluster                    --class org.apache.spark.examples.SparkPi
           $SPARK_HOME/examples/jars/spark-examples_2.12-3.4.3.jar 3
```



History Server

Event log directory: hdfs://isel-master:9000/spark-logs

Last updated: 2024-05-14 17:01:08

Client local time zone: Asia/Seoul

Search:

Version	App ID	App Name	Started	Completed	Duration	Spark User	Last Updated	Event Log
3.4.3	application_1715672998702_0002	Spark Pi	2024-05-14 16:56:13	2024-05-14 16:56:59	46 s	hadoop	2024-05-14 16:56:59	Download

Showing 1 to 1 of 1 entries
Show incomplete applications

```
spark-submit --deploy-mode cluster $SPARK_HOME/examples/src/main/python/wordcount.py
/user/hadoop/inputs/alice.txt
```