# 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 \

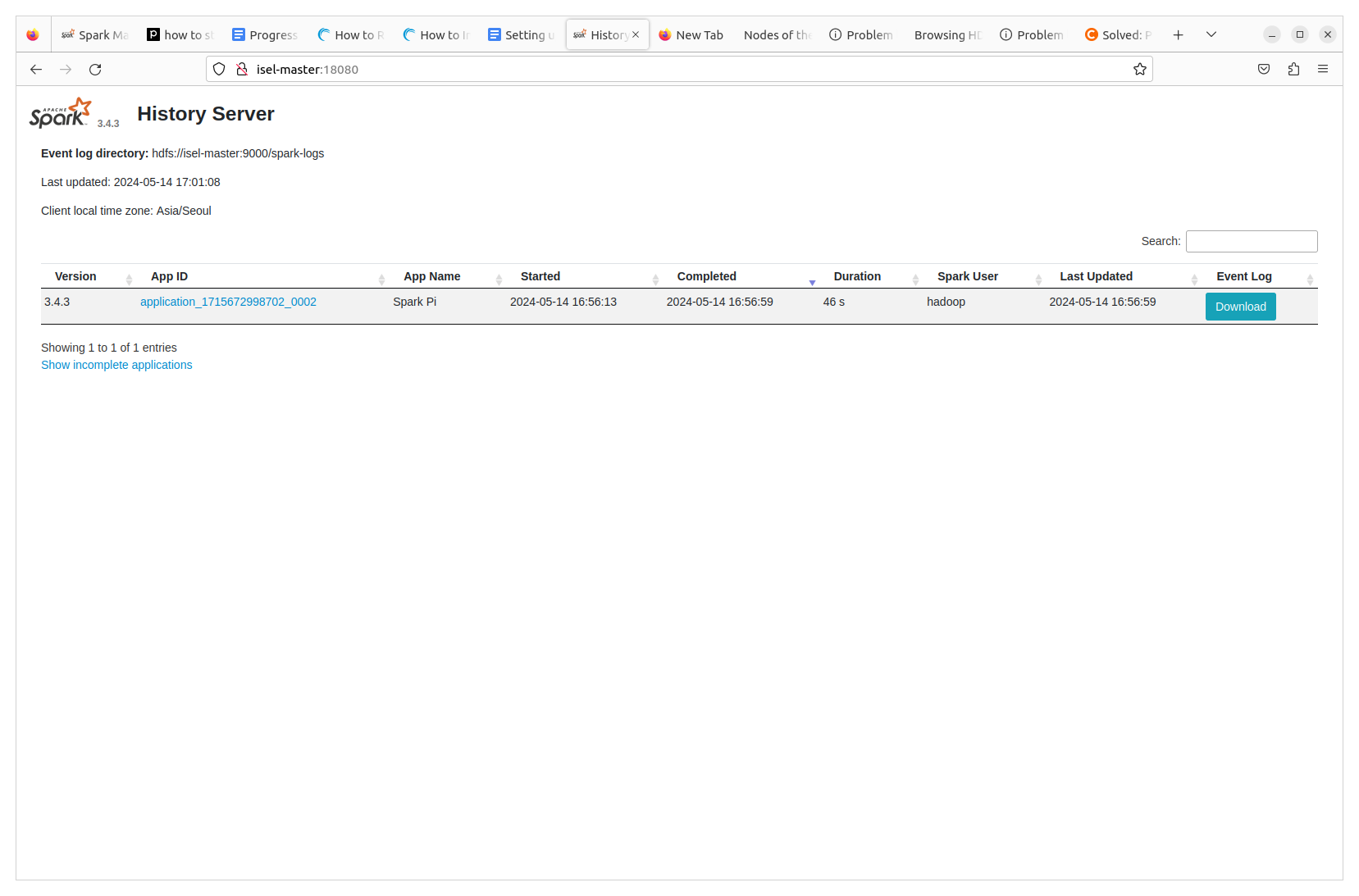
$SPARK\_HOME/examples/jars/spark-examples\_2.11-2.2.0.jar 10

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



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