Setting up HDFS
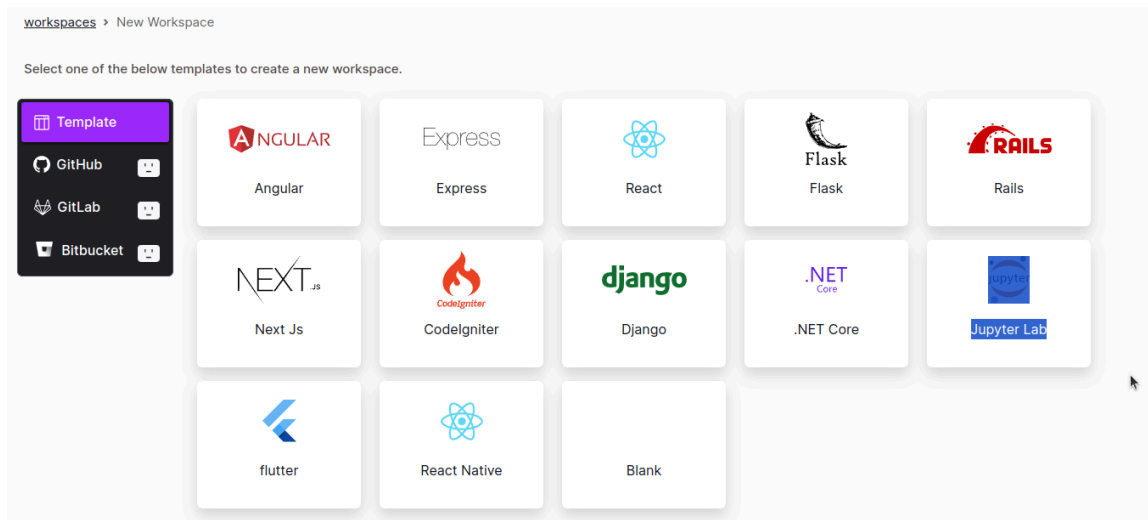
1 - Sign up/ Log in to runcode.io

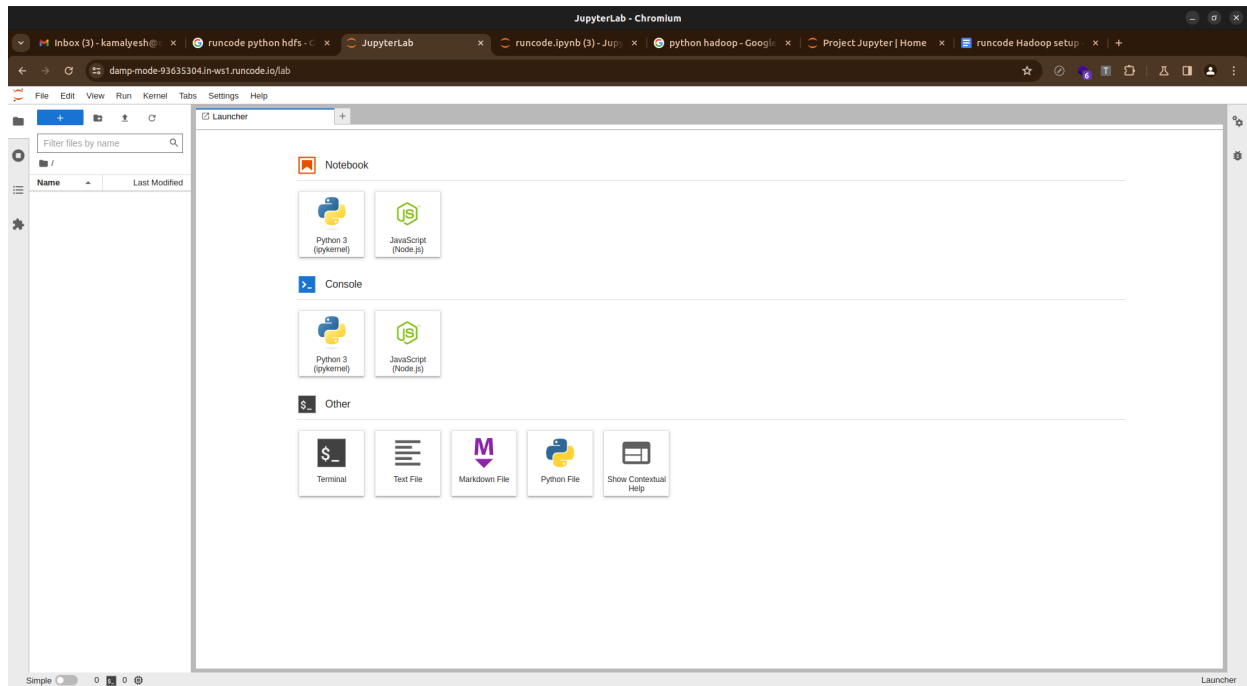2 - Create new workspace https://runcode.io/ide/new/

1. Select configuration. I am using Tiny for this trial
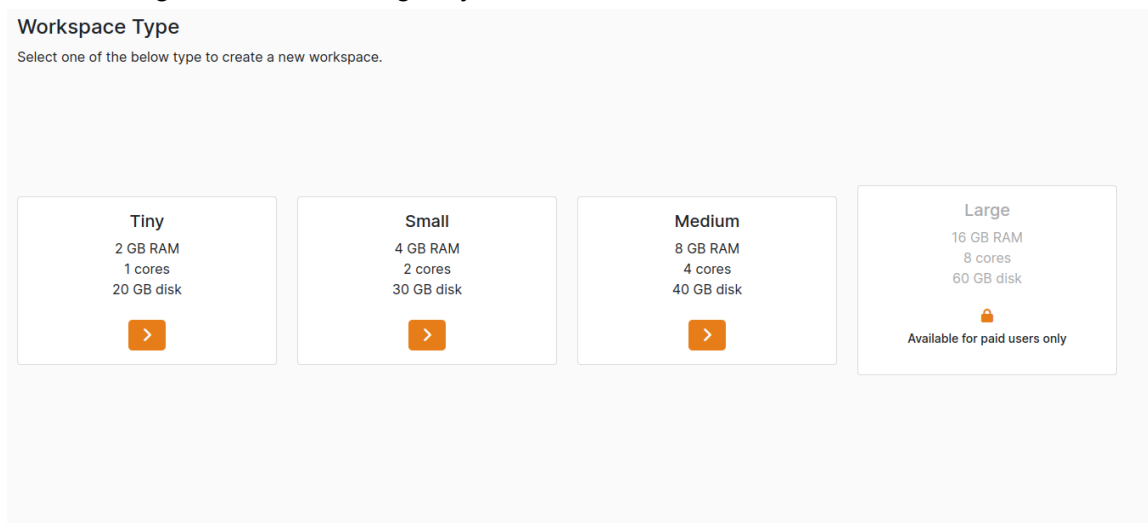
It will take some time in loading and show an error page.
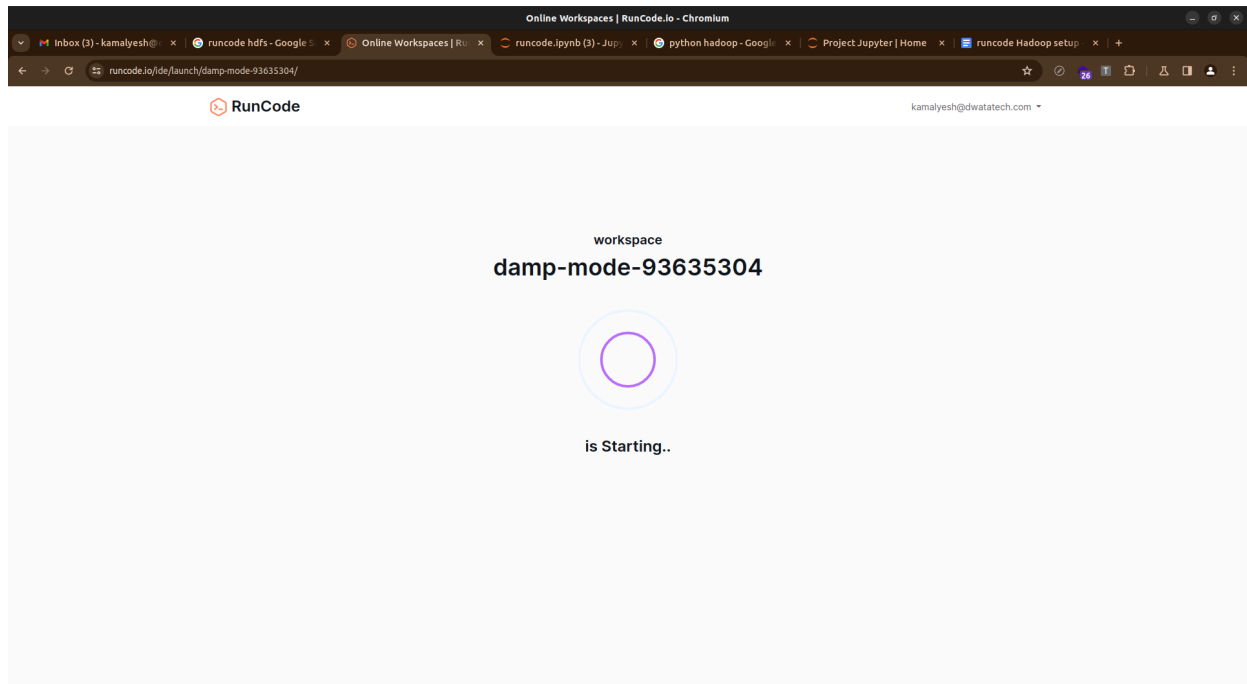
Refresh the page and you will see the IDE. Open terminal

Click on the file → New → Terminal

Then start executing the below codes

For setting up HDFS follow this link - https://sites.google.com/view/dksbin/tutorials/hadoop

1 - **sudo apt-get update**
2 - **sudo apt install openjdk-8-jdk**
3 - **java -version**
4 - **update-alternatives --config java**

Adding a dedicated Hadoop user

5 - **sudo addgroup hadoop**
6 - **sudo adduser --ingroup hadoop hduser**
7 - **groups hduser**
8 - **sudo adduser hduser sudo**

Installing SSH

9 - **sudo apt-get install ssh**
10 - **which ssh**
11 - **which sshd**
12 - **su hduser**
13 - **ssh-keygen**
14 - **cat $HOME/.ssh/id_rsa.pub >> $HOME/.ssh/authorized_keys**
15 - **ssh localhost**

Download Hadoop

16 - **wget https://archive.apache.org/dist/hadoop/core/hadoop-2.9.0/hadoop-2.9.0.tar.gz**
17 - **tar xvzf hadoop-2.9.0.tar.gz**
18 - **cd hadoop-2.9.0**
19 - **sudo mkdir -p /usr/local/hadoop**
20 - **sudo chown -R hduser:hadoop /usr/local/hadoop**

Setup Configuration Files

The following files need to be configured one by one

21 - **sudo nano ~/.bashrc**

Confirm the java home path according to your system

```
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export HADOOP_INSTALL=/usr/local/hadoop
export PATH=$PATH:$HADOOP_INSTALL/bin
export PATH=$PATH:$HADOOP_INSTALL/sbin
export HADOOP_MAPRED_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_HOME=$HADOOP_INSTALL
export HADOOP_HDFS_HOME=$HADOOP_INSTALL
export YARN_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_INSTALL/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_INSTALL/lib"
#export HADOOP_HOME=/usr/local/hadoop/sbin
```

Add this lines to the end of the file

```
    . /usr/share/bash-completion/bash_completion
  elif [ -f /etc/bash_completion ]; then
    . /etc/bash_completion
  fi
fi
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export HADOOP_INSTALL=/usr/local/hadoop
export PATH=$PATH:$HADOOP_INSTALL/bin
export PATH=$PATH:$HADOOP_INSTALL/sbin
export HADOOP_MAPRED_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_HOME=$HADOOP_INSTALL
export HADOOP_HDFS_HOME=$HADOOP_INSTALL
export YARN_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_INSTALL/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_INSTALL/lib"
#export HADOOP_HOME=/usr/local/hadoop/sbin
```

Like this the press Ctrl+O to overwrite and press enter, don't add the file name then press Ctrl+X to close it

22 - **source ~/.bashrc**
To save you file

23 - **javac -version**
24 - **which javac**
25 - **readlink -f /usr/bin/javac**
26 - **nano /usr/local/hadoop/etc/hadoop/hadoop-env.sh**
This is very important step to setup the java path in your hadoop environment
You have to paste this line in the above file

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

```
# Unless required by applicable law or agreed to in writing, software
# distributed under the License is distributed on an "AS IS" BASIS,
# WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
# See the License for the specific language governing permissions and
# limitations under the License.

# Set Hadoop-specific environment variables here.

# The only required environment variable is JAVA_HOME.  All others are
# optional.  When running a distributed configuration it is best to
# set JAVA_HOME in this file, so that it is correctly defined on
# remote nodes.

# The java implementation to use.
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64

# The jsvc implementation to use. Jsvc is required to run secure datanodes
# that bind to privileged ports to provide authentication of data transfer
# protocol.  Jsvc is not required if SASL is configured for authentication of
# data transfer protocol using non-privileged ports.
#export JSVC_HOME=${JSVC_HOME}

export HADOOP_CONF_DIR=${HADOOP_CONF_DIR:-"/etc/hadoop"}

# Extra Java CLASSPATH elements.  Automatically insert capacity-scheduler.
for f in $HADOOP_HOME/contrib/capacity-scheduler/*.jar; do
  if [ "$HADOOP_CLASSPATH" ]; then
    export HADOOP_CLASSPATH=$HADOOP_CLASSPATH:$f
```

In the same place it will be commented just remove the hash and place your JAVA_HOME path there and

Like this the press Ctrl+O to overwrite and press enter, don't add the file name then press Ctrl+X to close it

27 - **sudo mkdir -p /app/hadoop/tmp**
28 - **sudo chown hduser:hadoop /app/hadoop/tmp**
29 - **nano /usr/local/hadoop/etc/hadoop/core-site.xml**

Add the following lines to core-site.xml

```
<configuration>
 <property>
  <name>hadoop.tmp.dir</name>
  <value>/app/hadoop/tmp</value>
  <description>A base for other temporary
directories.</description>
 </property>
<property>
 <name>fs.default.name</name>
 <value>HDFS://localhost:54310</value>
 <description>The name of the default file system.  A URI whose
 scheme and authority determine the FileSystem implementation.
The
 uri's scheme determines the config property (fs.SCHEME.impl)
naming
 the FileSystem implementation class.  The uri's authority is
used to
 determine the host, port, etc. for a filesystem.</description>
</property>
</configuration>
```

At the end of the file paste it `</configuration>` it will be written already don't paste again
Like this the press Ctrl+O to overwrite and press enter, don't add the file name then press Ctrl+X to close it

```xml
<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
  <name>hadoop.tmp.dir</name>
  <value>/app/hadoop/tmp</value>
  <description>A base for other temporary directories.</description>
 </property>
<property>
  <name>fs.default.name</name>
  <value>hdfs://localhost:54310</value>
  <description>The name of the default file system.  A URI whose
  scheme and authority determine the FileSystem implementation.  The
  uri's scheme determines the config property (fs.SCHEME.impl) naming
  the FileSystem implementation class.  The uri's authority is used to
  determine the host, port, etc. for a filesystem.</description>
 </property>
</configuration>
```

Mapred-site.xml

30 - cp /usr/local/hadoop/etc/hadoop/mapred-site.xml.template /usr/local/hadoop/etc/hadoop/mapred-site.xml

31 - nano /usr/local/hadoop/etc/hadoop/mapred-site.xml

Add the following lines to mapred-site.xml

```xml
<configuration>
 <property>
  <name>mapred.job.tracker</name>
  <value>localhost:54311</value>
  <description>The host and port that the MapReduce job tracker
runs
  at.  If "local", then jobs are run in-process as a single map
  and reduce task.
  </description>
 </property>
</configuration>
```

```
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
  <name>mapred.job.tracker</name>
  <value>localhost:54311</value>
  <description>The host and port that the MapReduce job tracker runs
  at.  If "local", then jobs are run in-process as a single map
  and reduce task.
  </description>
 </property>
</configuration>
```

At the end of the file paste it `</configuration>` it will be written already don't paste again

Like this the press Ctrl+O to overwrite and press enter, don't add the file name then press Ctrl+X to close it

32 - **sudo mkdir -p /usr/local/hadoop_store/HDFS/namenode**

33 - **sudo mkdir -p /usr/local/hadoop_store/HDFS/datanode**

34 - **sudo chown -R hduser:hadoop /usr/local/hadoop_store**

HDFS-site.xml

35 - **nano /usr/local/hadoop/etc/hadoop/HDFS-site.xml**

Add the following lines to HDFS-site.xml

```
<configuration>

<property>

    <name>dfs.replication</name>

    <value>1</value>

    <description>Default block replication.

    The actual number of replications can be specified when the

file is     eeecreated.
```

```
    The default is used if replication is not specified in create
time.
    </description>
</property>
<property>
  <name>dfs.namenode.name.dir</name>
  <value>file:/usr/local/hadoop_store/HDFS/namenode</value>
</property>
<property>
  <name>dfs.datanode.data.dir</name>
  <value>file:/usr/local/hadoop_store/HDFS/datanode</value>
</property>
</configuration>
```

```xml
<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
  <name>dfs.replication</name>
  <value>1</value>
  <description>Default block replication.
  The actual number of replications can be specified when the file is created.
  The default is used if replication is not specified in create time.
  </description>
 </property>
<property>
   <name>dfs.namenode.name.dir</name>
   <value>file:/usr/local/hadoop_store/hdfs/namenode</value>
 </property>
<property>
```

Dont add this eee it should be in the same line and it was not coming thats why it is like that

At the end of the file paste it `</configuration>` it will be written already don't paste again

Like this the press Ctrl+O to overwrite and press enter, don't add the file name then press Ctrl+X to close it

Add the following lines to yarn-site.xml

36 - nano /usr/local/hadoop/etc/hadoop/yarn-site.xml

```
<configuration>
<property>
<name>yarn.nodemanager.aux-services</name>
 <value>mapreduce_shuffle</value>
 </property>
<property>
      <name>yarn.nodemanager.aux-services</name>
      <value>mapreduce_shuffle</value>
  </property>
</configuration>
```

```
  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->
<configuration>

<property>
<name>yarn.nodemanager.aux-services</name>
 <value>mapreduce_shuffle</value>
 </property>
<property>
      <name>yarn.nodemanager.aux-services</name>
      <value>mapreduce_shuffle</value>
  </property>
</configuration>
```

At the end of the file paste it `</configuration>` it will be written already don't paste again
Like this the press Ctrl+O to overwrite and press enter, don't add the file name then press Ctrl+X to close it

Format the New Hadoop Filesystem

37 - hadoop namenode -format

38 - **start-dfs.sh**

39 - **start-yarn.sh**

40 - **jps**

In jps this all should run

hduser@runcode:~/hadoop-2.9.0$ jps

19777 ResourceManager

19953 NodeManager

20198 Jps

19257 NameNode

19434 DataNode

Otherwise you HDFS is not running properly

Now your HDFS setup is done Wooohoooo

We will see how to **get** and **put** file in HDFS

First will check any file is there in HDFS or not
hadoop fs -ls /
This command will show you the files present in HDFS with their permissions and location

Put command
hadoop fs -copyFromLocal /home/ubuntu/workspace/finalF.csv /

This is put command in which you have to give the file location from your hadoop cluster and at the end you have to specify the filename which you want to store in HDFS

Get command
hadoop fs -get /finalF.csv /home/ubuntu/workspace/new

This is the get command (fetching the file) after **-get** you have to give the file name which you want to extract from HDFS and **new** is the folder I have created to store the extracted file into the folder.

But

For the first time you will get the permission denied error like this
hduser@runcode:~/hadoop-2.9.0$ hadoop fs -get /finalF.csv
/home/ubuntu/workspace/new
24/05/08 06:15:46 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
get: /home/ubuntu/workspace/new/finalF.csv._COPYING_ (Permission denied)

Then just run this command
sudo chown hduser:hadoop -R /home/ubuntu/workspace/new

new is the filename which I have given in my local

And you are set with the HDFS

If any issues are there am attaching some websites links which we referred while setting up HDFS and also our logs and codes how we setup HDFS for the first time

How to download HDFS reference links
1 - https://cloudzy.com/blog/install-hadoop-on-ubuntu/
2 -
https://www.digitalocean.com/community/tutorials/how-to-install-hadoop-on-ubuntu-13-10#start-hadoop
    It is the most important link
3 - https://phoenixnap.com/kb/install-hadoop-ubuntu
    It is also one of the most imp links
4 - https://www.guru99.com/how-to-install-hadoop.html
    For certain errors it might help
5 - https://sites.google.com/view/dksbin/tutorials/hadoop
    From this we have got it sucessful and some commands which we have written some different commands it is in day 2 txt file

The bold links are the most important ones and on the top you will get the link from which this setup is done

Day 1 - 📄 Day 1 HDFS Setup

Day 2 - 📄 Day 2 HDFS Setup

Day 3 - 📄 Day 3 HDFS Setup