How to Access Big Data Cluster?
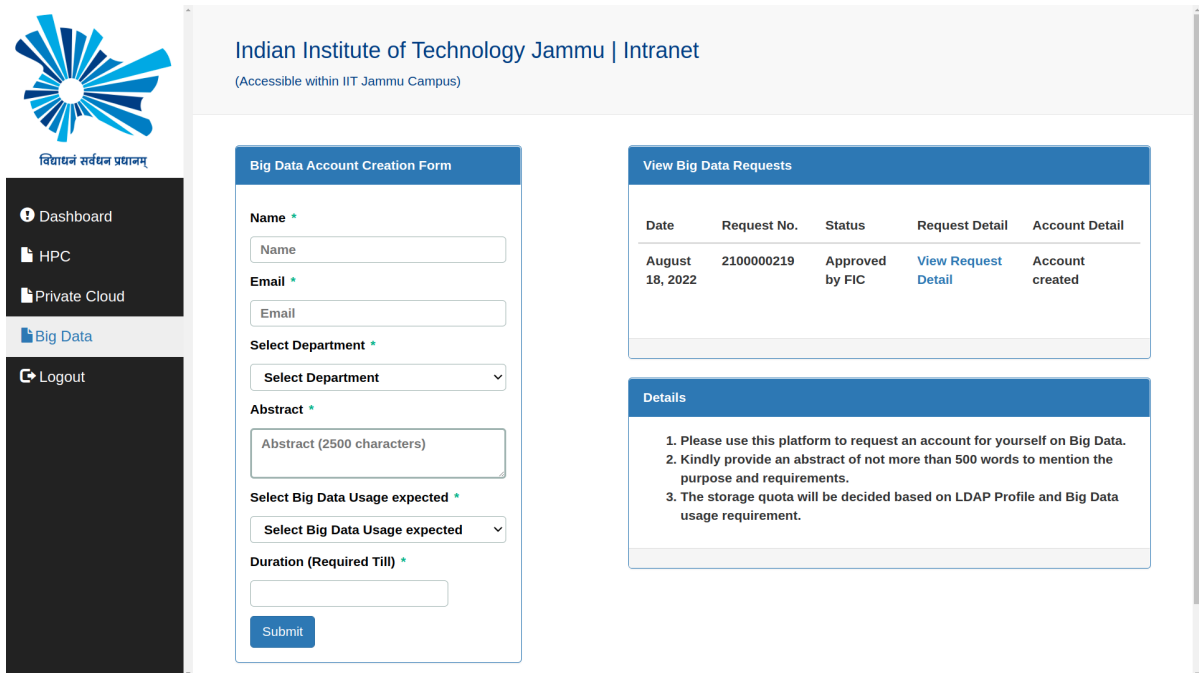
1. Request for the Big data user access credentials :
   Link: intranet.iitjammu.ac.in

2. You will be provided user credentials after approval from authorities to access your Big data user account:
   Credentials: USERNAME, PASSWORD, IP_Address.
3. SSH the machine using Moba xterm.
   https://mobaxterm.mobatek.net/download.html

   Click on -> Session -> SSH -> IP and Username -> Enter password on terminal
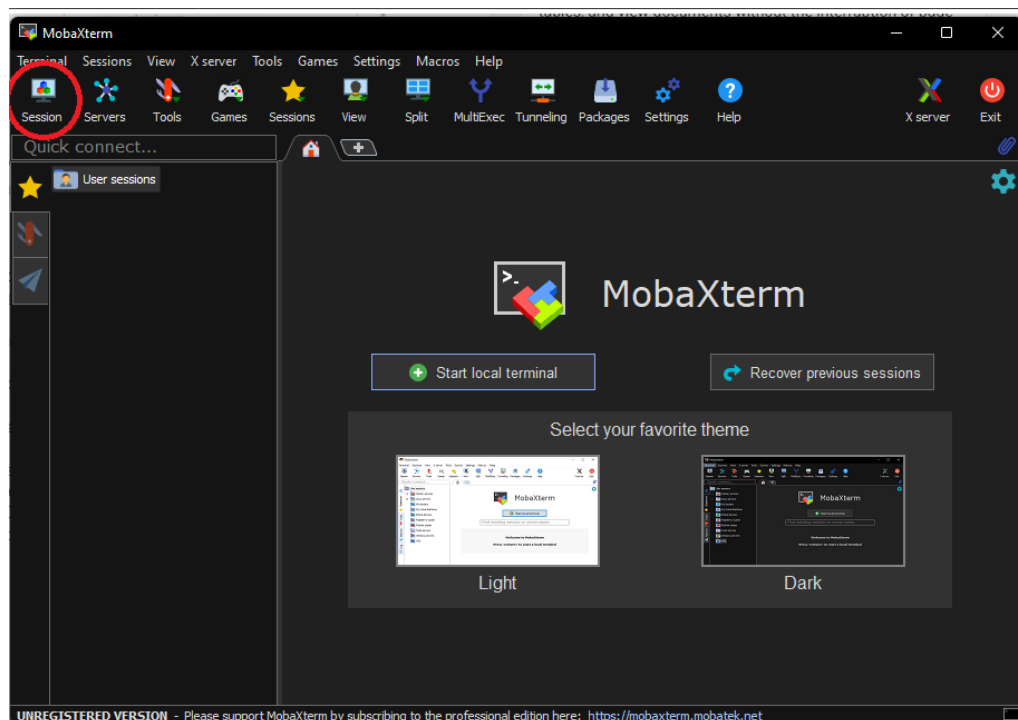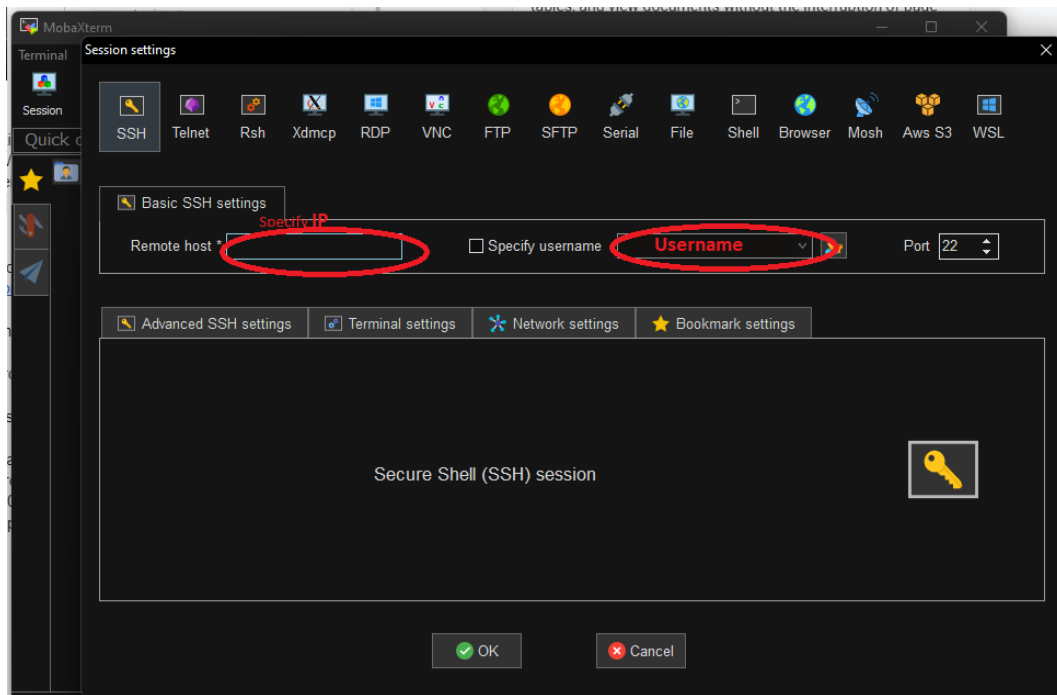
4. Write the mapper and reducer code that you want to run on the Hadoop cluster.
   You can follow the link for the word count program.
   :https://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python/
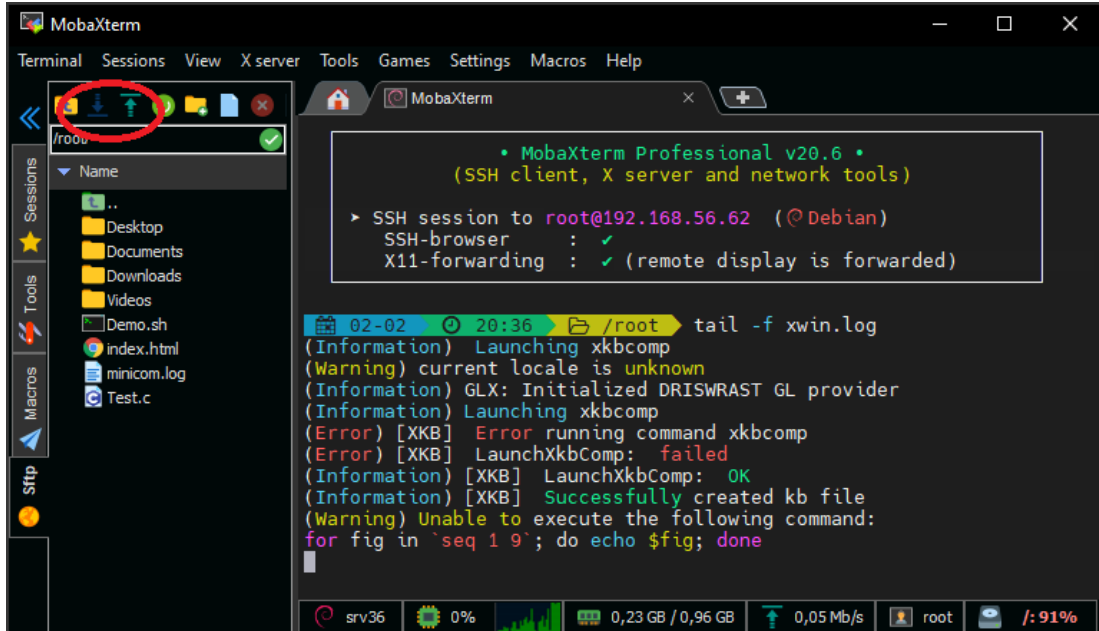
5. Upload the mapper and reducer to your BigData user space using mobaXterm.

Run map reduce program for wordcount
1. Create a file data.txt with sample data,
2. Upload data.txt to HDFS(hadoop distributed file system),

   A. Create a directory in HDFS

```
      hdfs dfs -mkdir /data
```
B. Put local data.txt into that folder in HDFS.
```
      hdfs dfs -put /home/username/data.txt /data
```

.

3. Run mapper.py and reducer.py on data.txt to perform word count.
```
hadoop jar
/usr/hdp/3.1.4.0-315/hadoop-mapreduce/hadoop-streaming.ja
r -input data.txt  -output output001/ou5 -mapper
mapper.py -reducer reducer.py
```

Brief Explanation:

`hadoop jar` <location of hadoop jar file>

`-input` < location of input data in HDFS>

`-output` <location of output folder(unique name) in HDFS>

`-mapper` <Location of mapper in local>

`-reducer` <location of reducer in local>

.