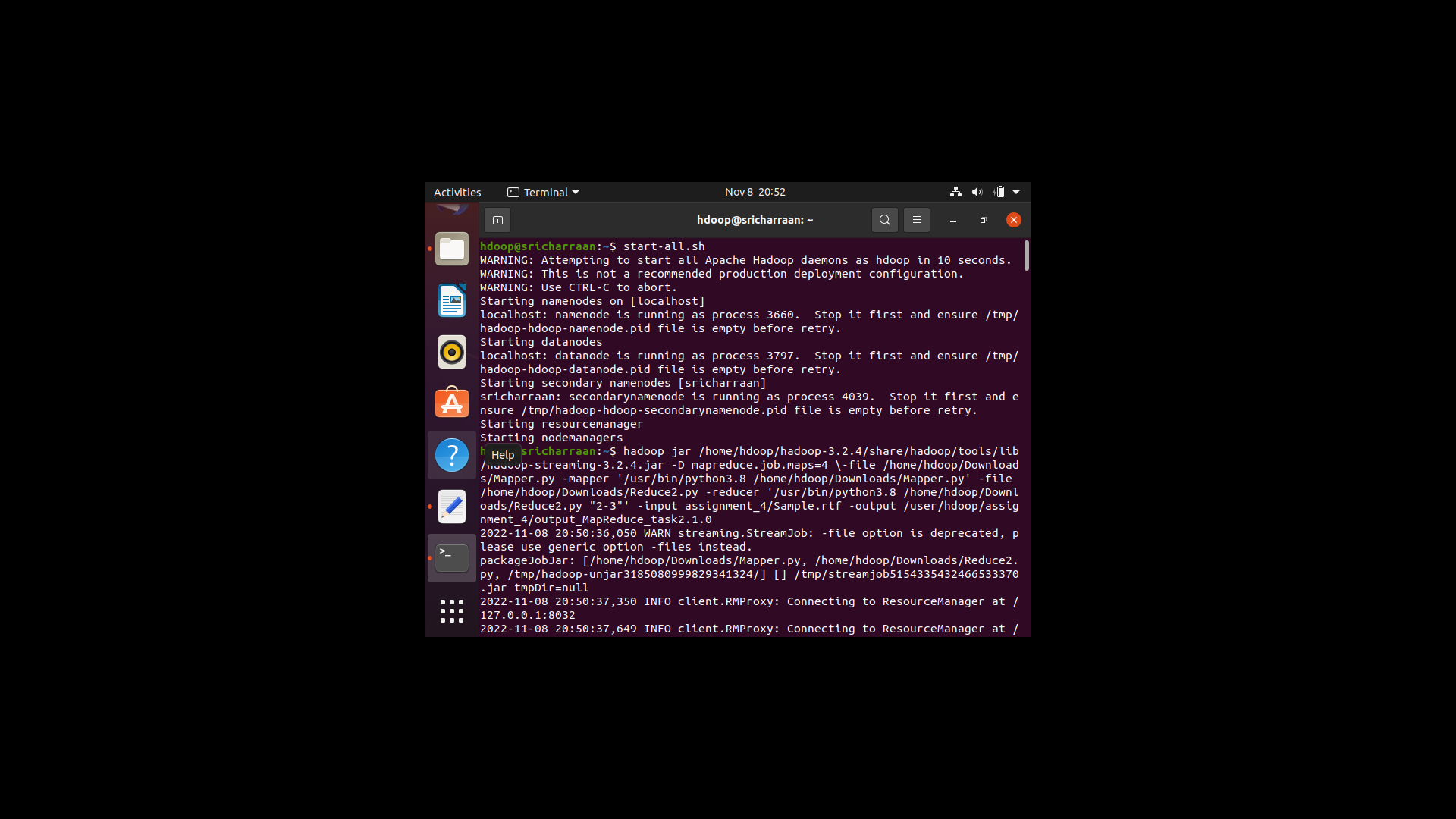
**ENGR-E: 516 - Engineering Cloud Computing**

**Assignment 1: MapReduce using Hadoop**

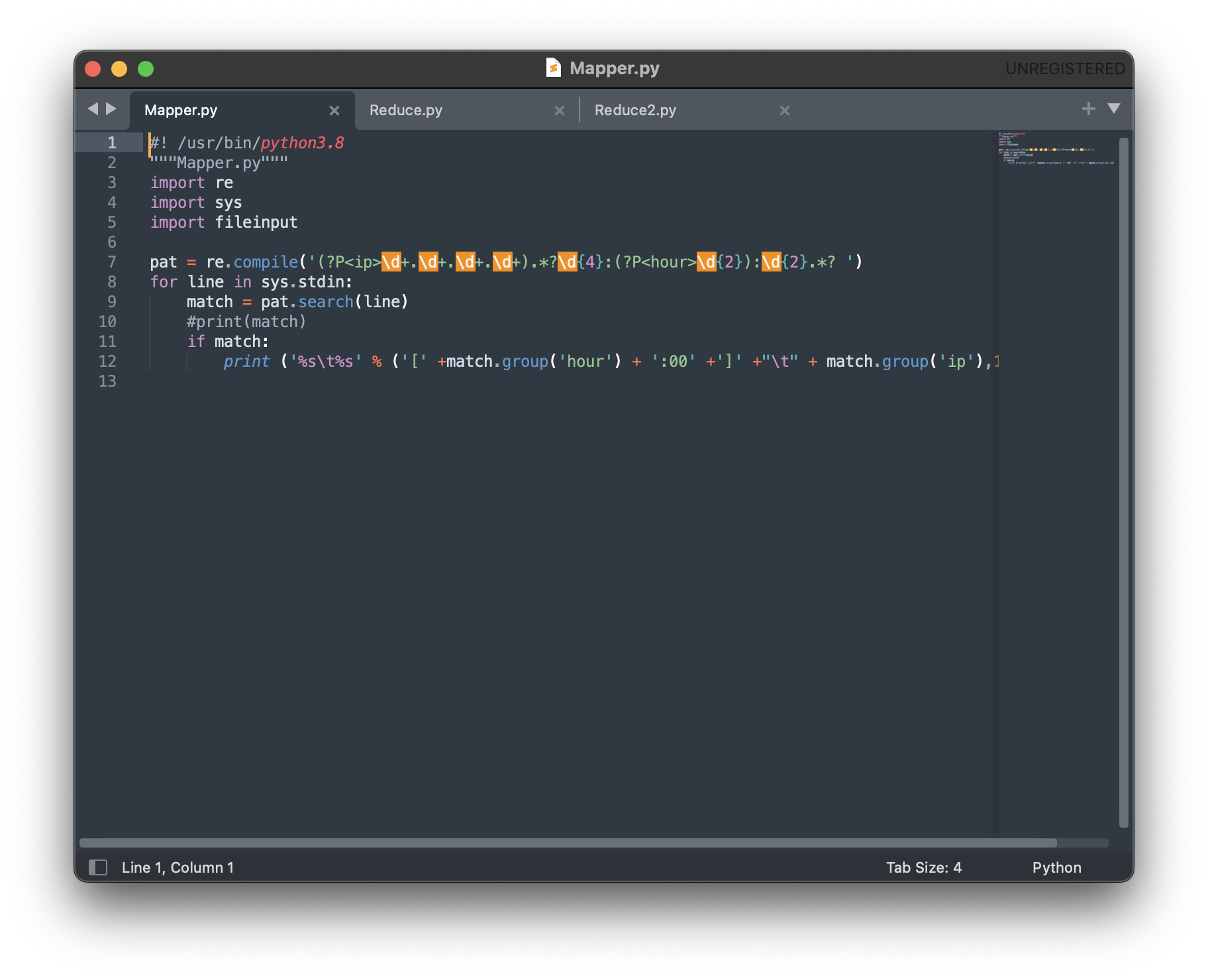
* 1. ***Setting Up: Hadoop Environment***
* After Setting up Ubuntu environment I downloaded the latest version of jdk
* Then, the setup of SSH localhost is done
* Then, download the Apache Hadoop version 3.2.4
* Then, setting path of different .xml files was done.
* Then, after allocating the path the whole file is run and Hadoop is setup
* Then to start Hadoop the command “start-all.sh” is used as one can see in the *Fig. 1.1*
* The whole process was referenced, and the links of reference are mentioned in *1.6*



*Fig. 1.1: Starting Hadoop Environment*

* 1. ***Coding Mapper and Reducer Function***

The *Fig.1.2.1* shows the Mapper function of the MapReduce function in which the time and IP address is scrapped from the log file and each match of time and IP address is given the count of 1.

******

*Fig. 1.2.1: Mapper Function*

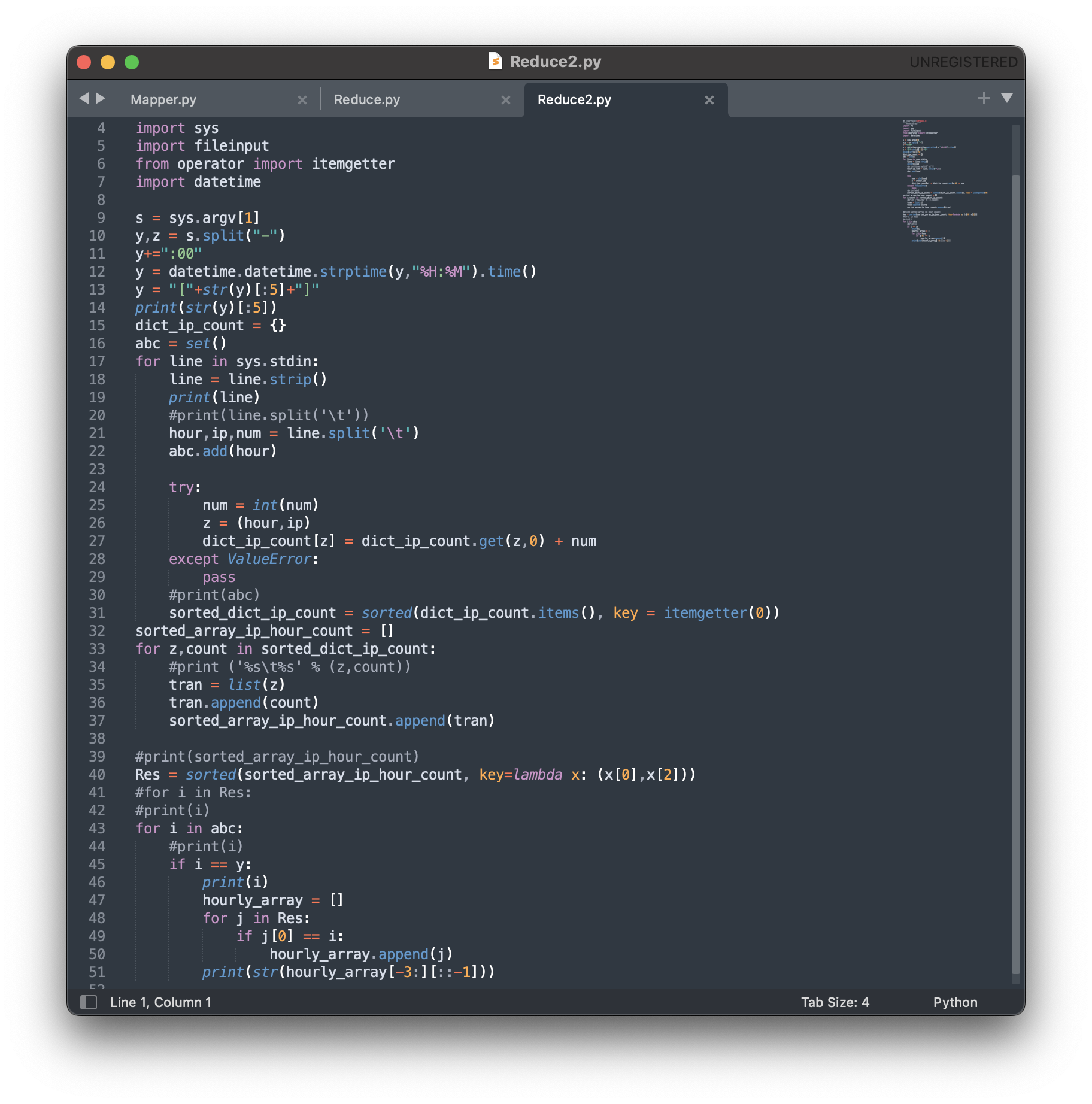
The *Fig.1.2.2* shows the Reducer function of the MapReduce function in which the counting of repetition of IP address is done according to the time and reporting the **top 3-IP addresses for every hour**.

*Text

Description automatically generated*

*Fig. 1.2.2: Reducer Function for top 3 IP Addresses*

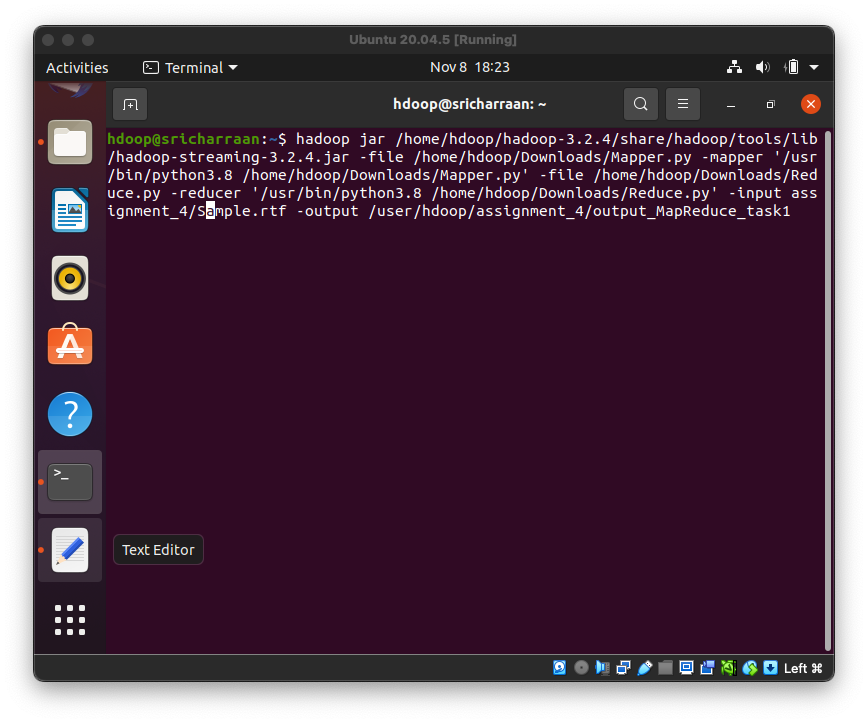
The *Fig.1.2.3* shows the Reducer function of the MapReduce function in which the counting of repetition of IP address is done according to the time and only reporting the top 3-IP addresses of the user mentioned timeframe used for *Database Search*.

**

*Fig. 1.2.3: Reducer Function for Database Search*

* 1. ***Output the top 3 IP addresses with the granularity of an hour***

The MapReduce cluster is inbuilt function in Hadoop to build it one must call the Hadoop streaming file in which the location of mapper and reducer function must be mentioned, and input file should be uploaded in the Hadoop cluster and its location for smooth reading of the input and the new location of where the output file should be stored.

****

*Fig. 1.3.1: MapReduce command for top 3 IP Address*

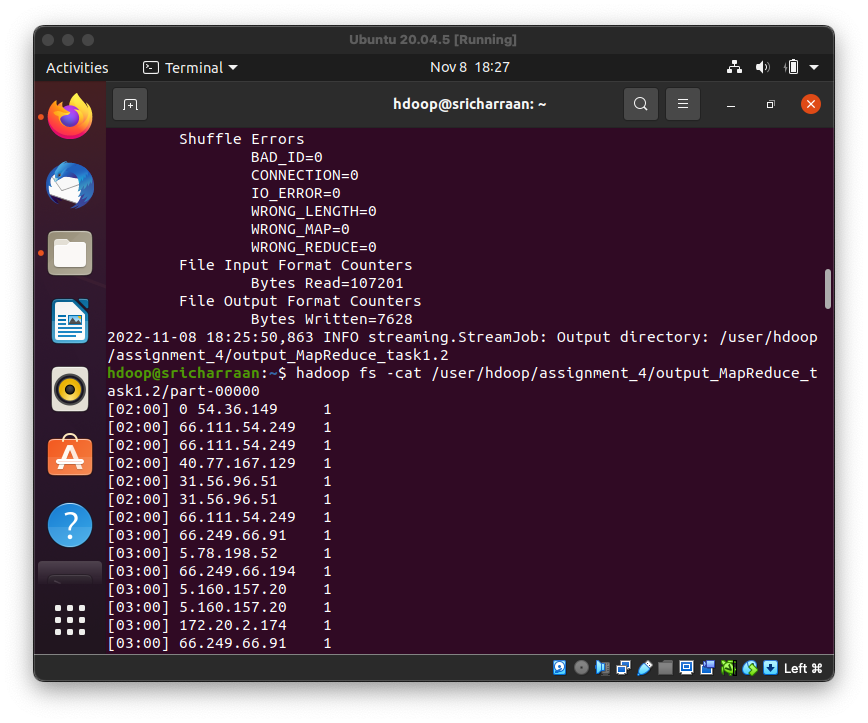
After successfully giving all the details and location of the required documents when can see in *Fig.1.3.2* the successful running of mapper and reducer function.

**Graphical user interface, text, chat or text message

Description automatically generated**

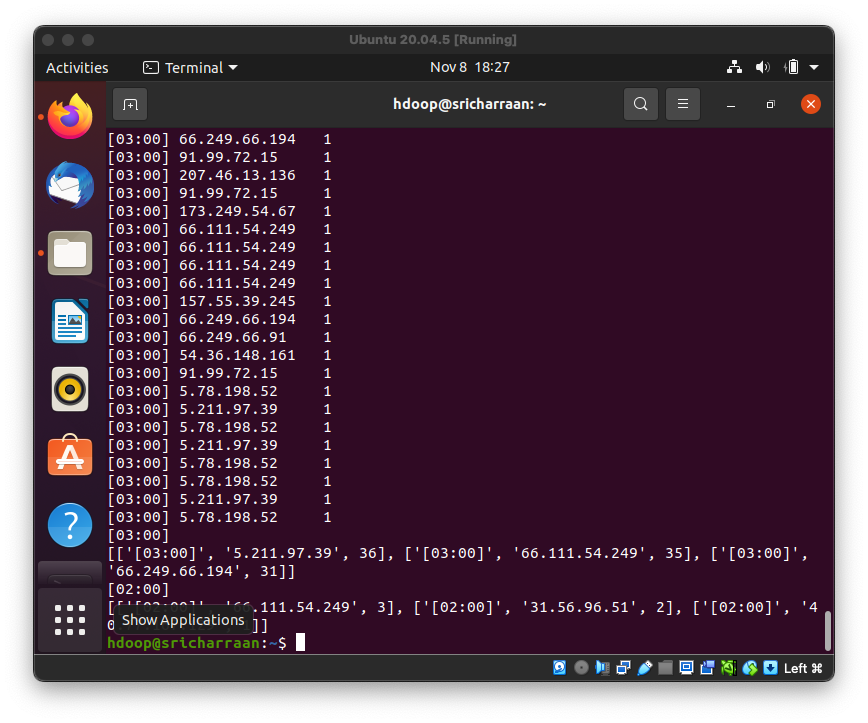
*Fig. 1.3.2: Successfully running of Mapper and Reducer*

After successful completion of the whole MapReduce function the output file stored in the designated location mentioned, and to access it from the Hadoop cluster we need to use the ***-cat*** function.

****

*Fig. 1.3.3: End MapReduce function and accessing output file*

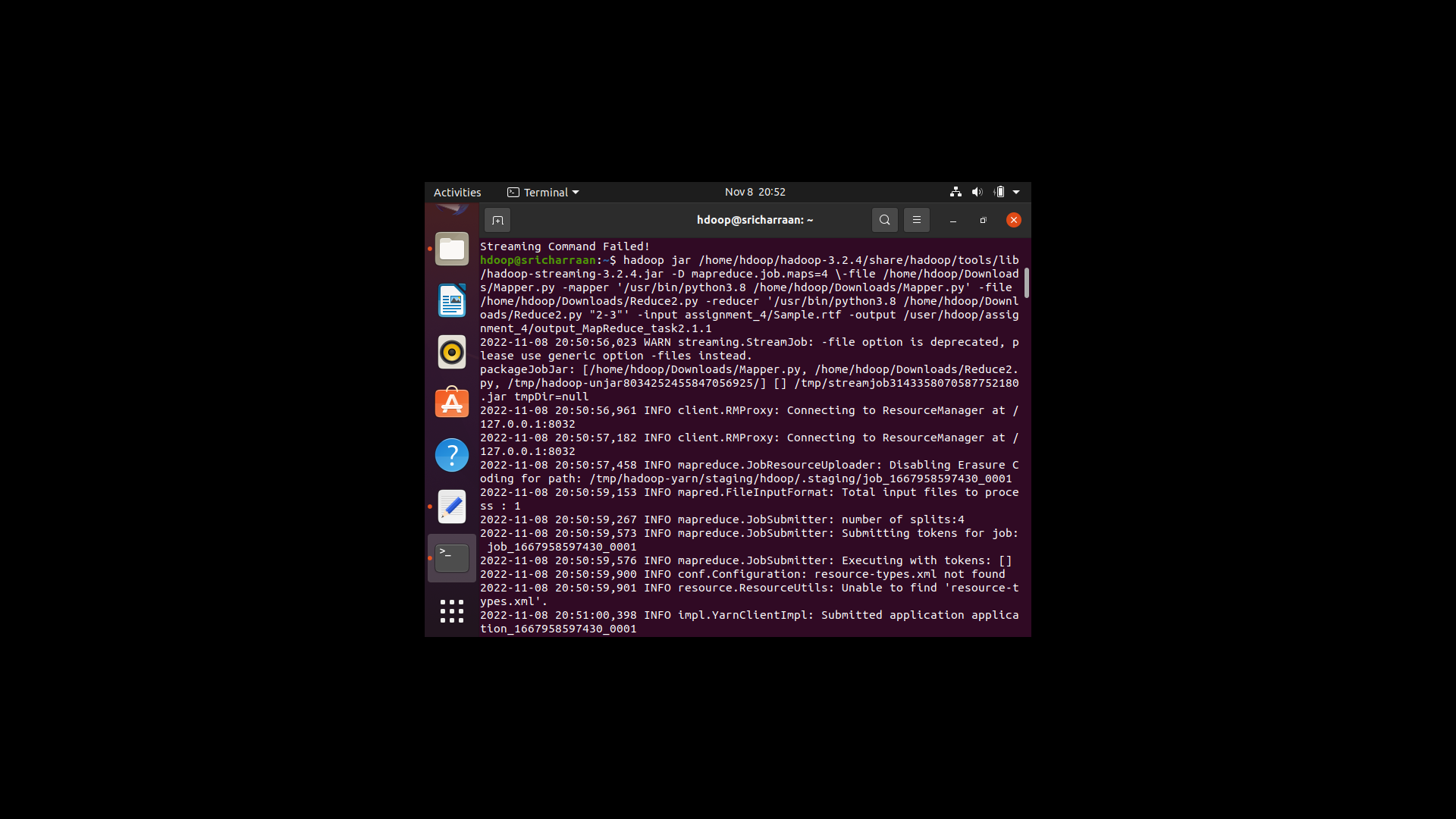
The output of the whole MapReduce function can be seen in the *Fig. 1.3.4*

****

*Fig. 1.3.4: Output of top 3-IP address with granularity of an hour*

* 1. ***Database Search***

The MapReduce cluster is inbuilt function in Hadoop to build it one must call the Hadoop streaming file in which the location of mapper and reducer function must be mentioned, and input file should be uploaded in the Hadoop cluster and its location for smooth reading of the input and the new location of where the output file should be stored. The Database is searched for 2:00 - 3:00



*Fig. 1.4.1: MapReduce command for Database Search from 2:00 – 3:00*

After successfully giving all the details and location of the required documents when can see in *Fig.1.4.2* the successful running of mapper and reducer function.

*Graphical user interface, text, application

Description automatically generated*

*Fig. 1.4.2: Successfully running of Mapper and Reducer*

After successful completion of the whole MapReduce function the output file stored in the designated location mentioned, and to access it from the Hadoop cluster we need to use the ***-cat*** function.

Graphical user interface

Description automatically generated

*Fig. 1.4.3: End MapReduce function and accessing output file*

The output of the whole MapReduce function can be seen in the *Fig. 1.4.4.* The output shows the top 3 IP address from 2:00 – 3:00.

Graphical user interface, application

Description automatically generated

*Fig. 1.4.4: Output of Database Search for 2:00 - 3:00*

* 1. ***GitHub Assignment Link***

<https://github.com/sricharraan/ENGR-E-516-Cloud-Computing/tree/main/Assignment1>

***1.5 References:***

1. <https://phoenixnap.com/kb/install-hadoop-ubuntu>
2. <https://www.youtube.com/watch?v=mhisPejz-6c&t=788s>
3. <https://www.youtube.com/watch?v=9CW8rD_MF_o&list=PLylBsOS1de9igxKhZJNpB32sxLTKY43_j&index=15&t=764s>