**Hadoop (Linux)**

Apache Hadoop is an open source framework that is used to efficiently store and process large datasets ranging in size from gigabytes to petabytes of data. Instead of using one large computer to store and process the data, Hadoop allows clustering multiple computers to analyze massive datasets in parallel more quickly.

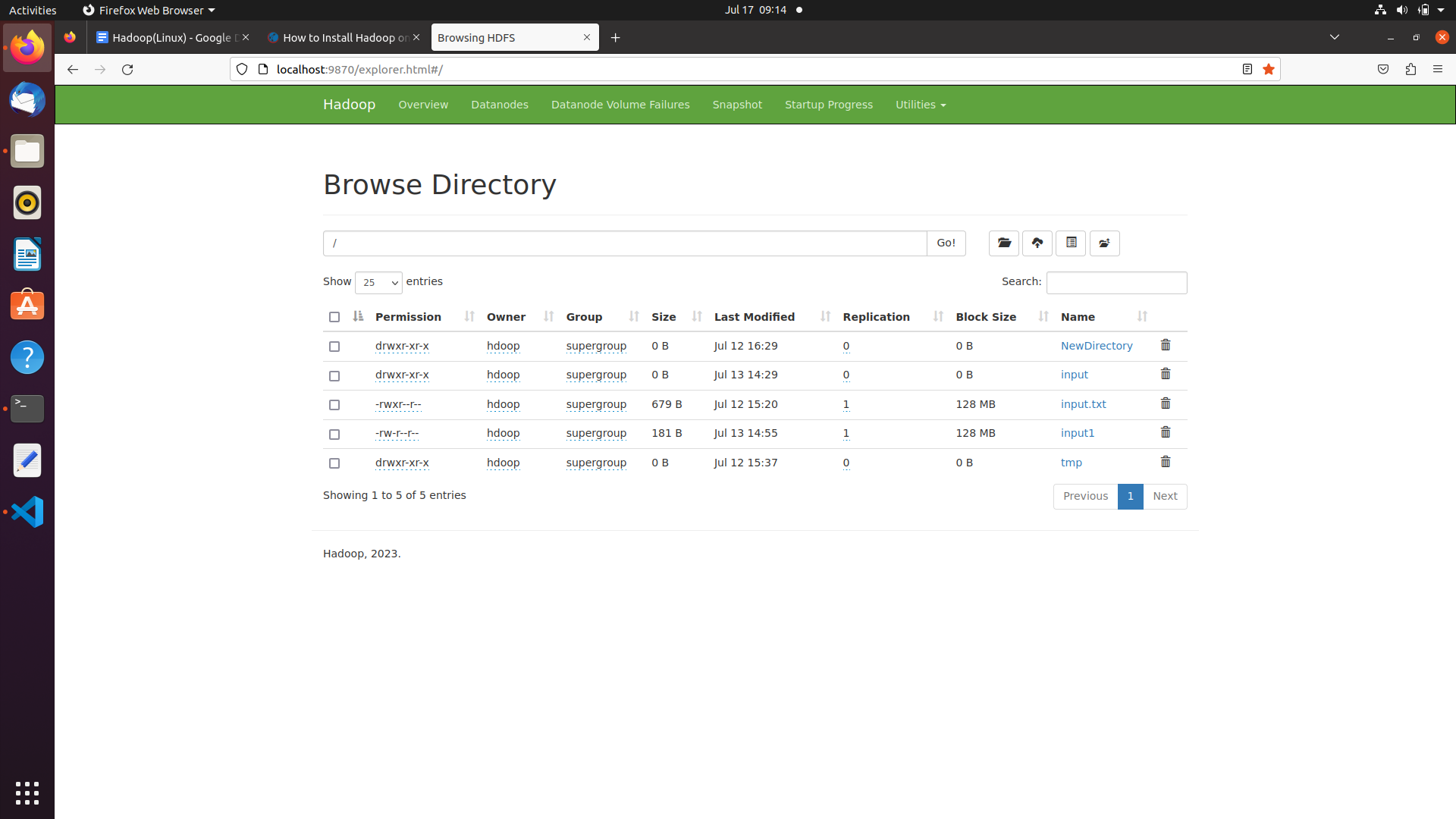
Hadoop consists of four main modules:

* Hadoop Distributed File System (HDFS) – A distributed file system that runs on standard or low-end hardware. HDFS provides better data throughput than traditional file systems, in addition to high fault tolerance and native support of large datasets.
* Yet Another Resource Negotiator (YARN) – Manages and monitors cluster nodes and resource usage. It schedules jobs and tasks.
* MapReduce – A framework that helps programs do the parallel computation on data. The map task takes input data and converts it into a dataset that can be computed in key value pairs. The output of the map task is consumed by reduce tasks to aggregate output and provide the desired result.
* Hadoop Common – Provides common Java libraries that can be used across all modules.

**Hadoop Setup:**

Follow the steps from below website link to install and setup hadoop for single node cluster.

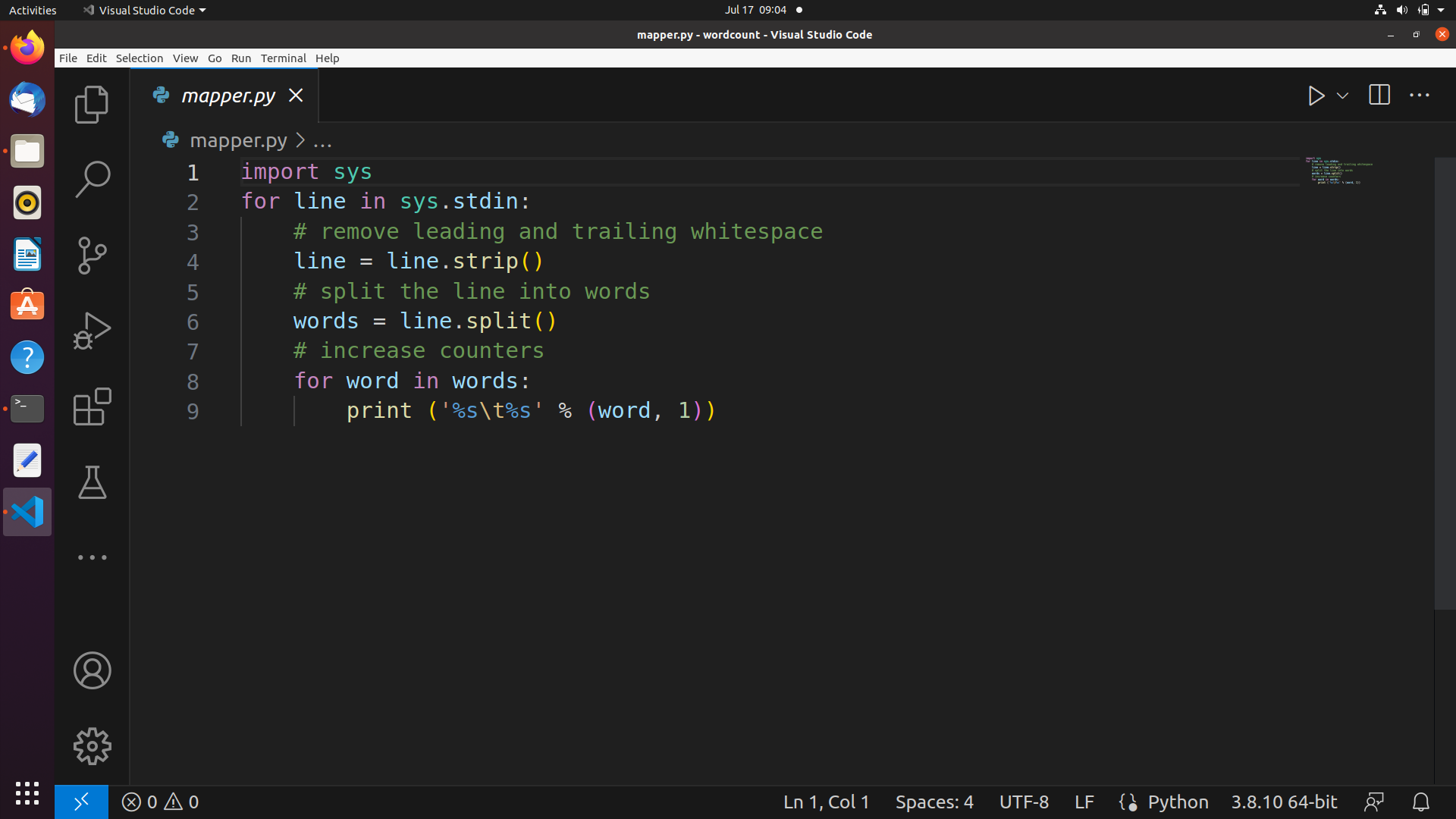
**https://phoenixnap.com/kb/install-hadoop-ubuntu**

****

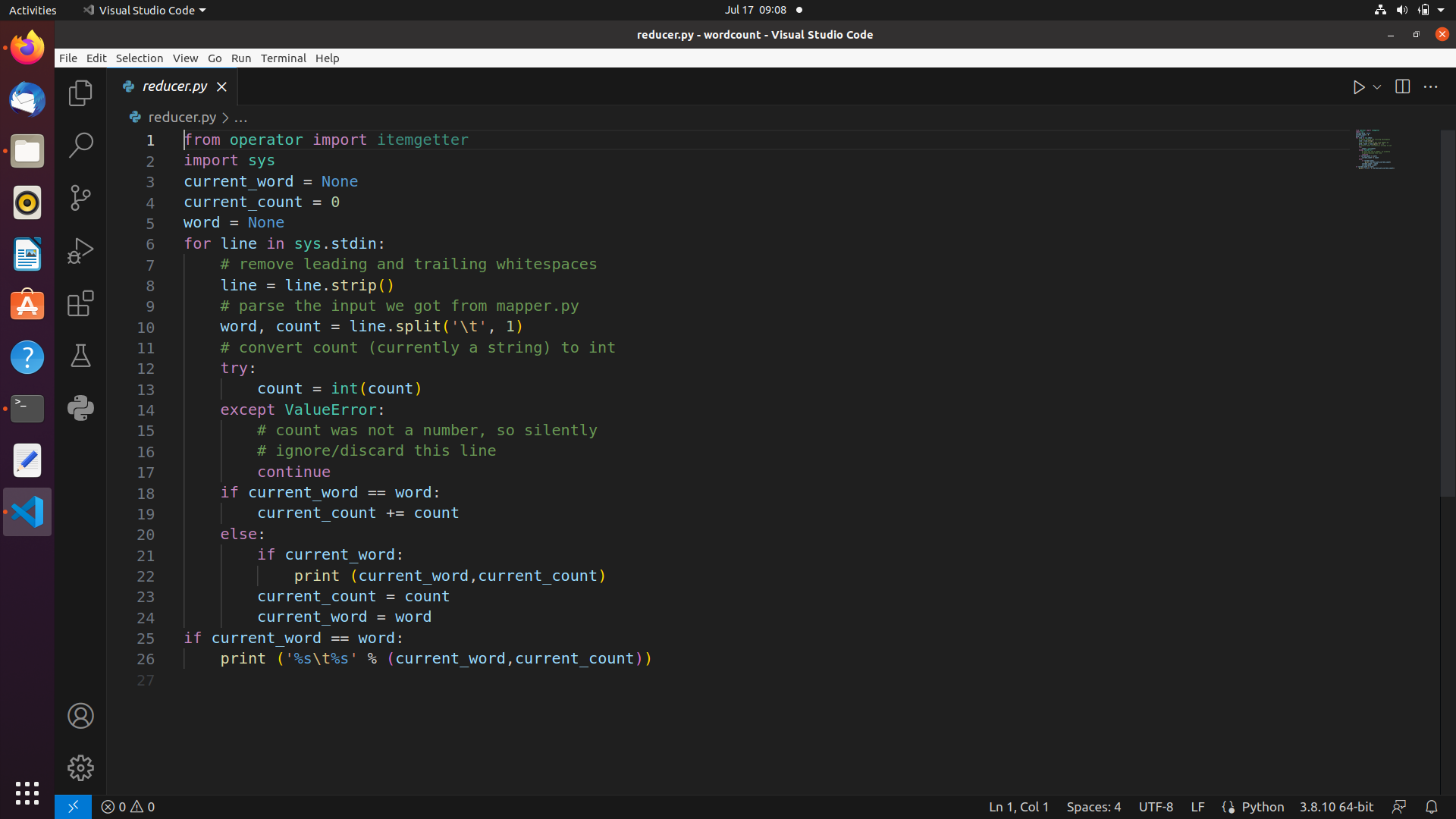
**Word Count Problem in Hadoop using MapReduce:**

**Note: Have an input file in hadoop containing words.**

1. Create a mapper.py file in python (keep the files in local system).



1. Create a reducer.py file in python



1. Run the following stream command in the Terminal to execute and store the output in hadoop (single node cluster).

**Hadoop stream Command:**

**hadoop jar /home/hdoop/hadoop-3.3.6/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar -file /home/hdoop/Documents/python/wordcount/mapper.py -mapper "python3 mapper.py" -file /home/hdoop/Documents/python/wordcount/reducer.py -reducer "python3 reducer.py" -input /input.txt -output /NewDirectory/Output**

**Output:**

