# Lab1: Working with hadoop

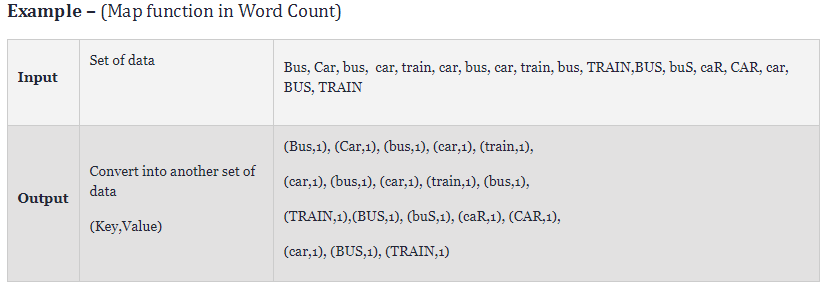
Introduction:

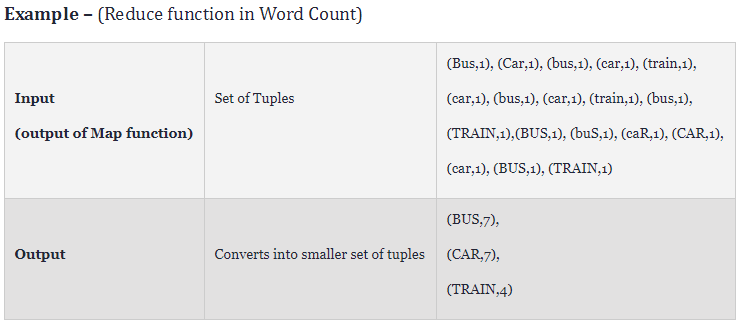
In Hadoop, MapReduce is a computation that decomposes large manipulation jobs into individual tasks that can be executed in parallel across a cluster of servers. The results of tasks can be joined together to compute final results.

MapReduce consists of 2 steps:

**Map Function** – It takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (Key-Value pair).

**Reduce Function** – Takes the output from Map as an input and combines those data tuples into a smaller set of tuples.





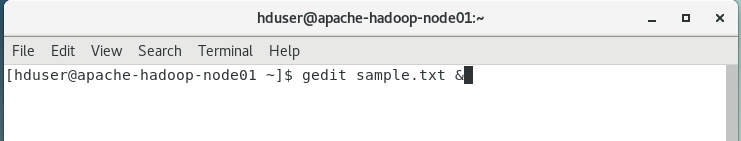
**Pre-requisites:**

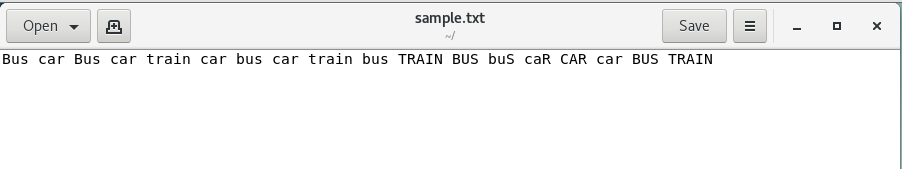
1. Hadoop should be installed
2. Hadoop cluster should be up and running

Steps:

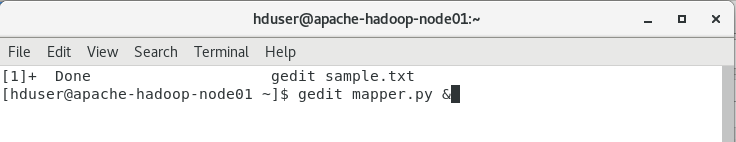
1. Create a file with the following content and name it sample.txt.

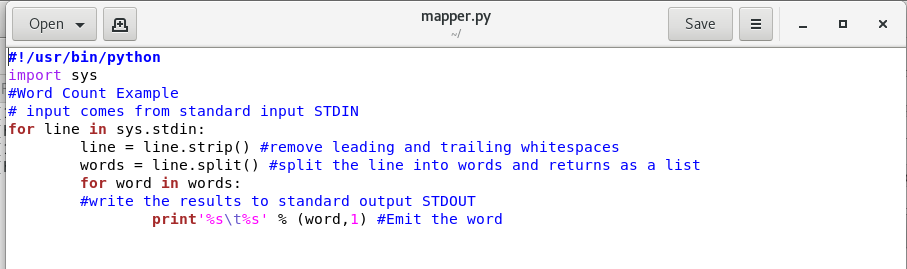
Bus car Bus car train car bus car train bus TRAIN BUS buS caR CAR car BUS TRAIN



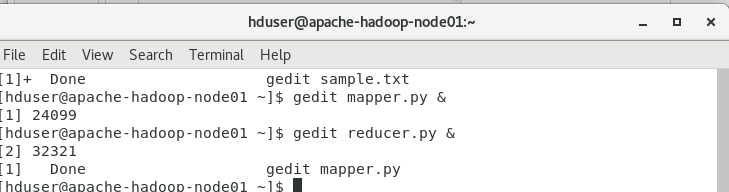


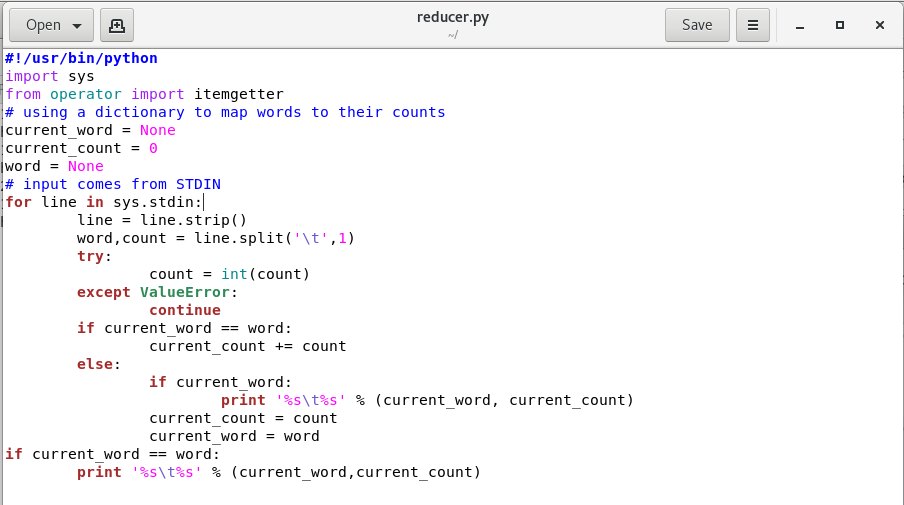
1. Write the mapper code in mapper.py





1. Write the reducer code in reducer.py

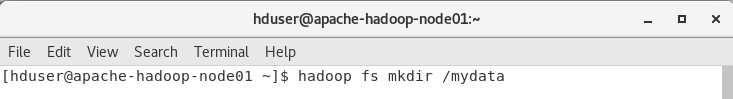
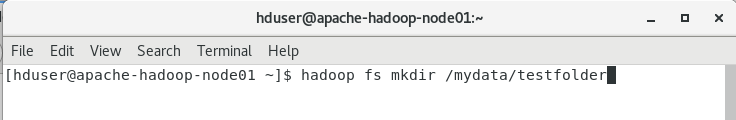


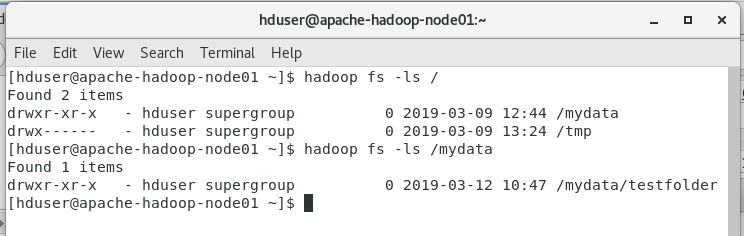


1. Create directory in Hadoop file system where you want to store your data files

hadoop fs -mkdir /mydata

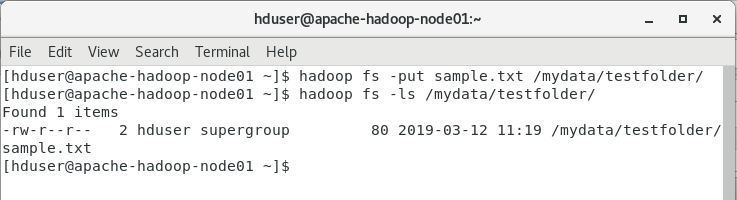
hadoop fs -mkdir /mydata/testfolder



1. Move the data file i.e. sample.txt to Hadoop file system

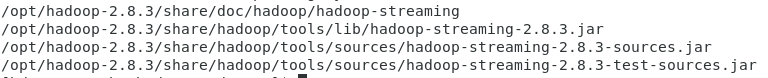
hadoop fs -put sample.txt /mydata/testfolder/



1. Find Hadoop streaming location

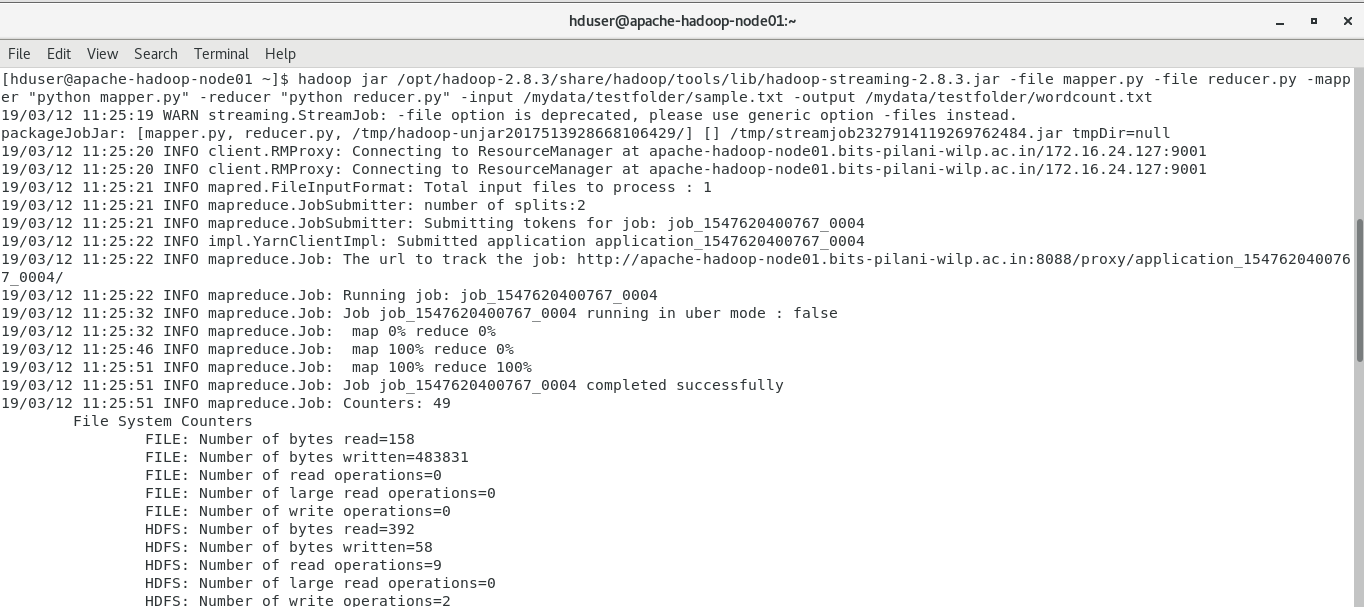
find / -name Hadoop\*stream\*

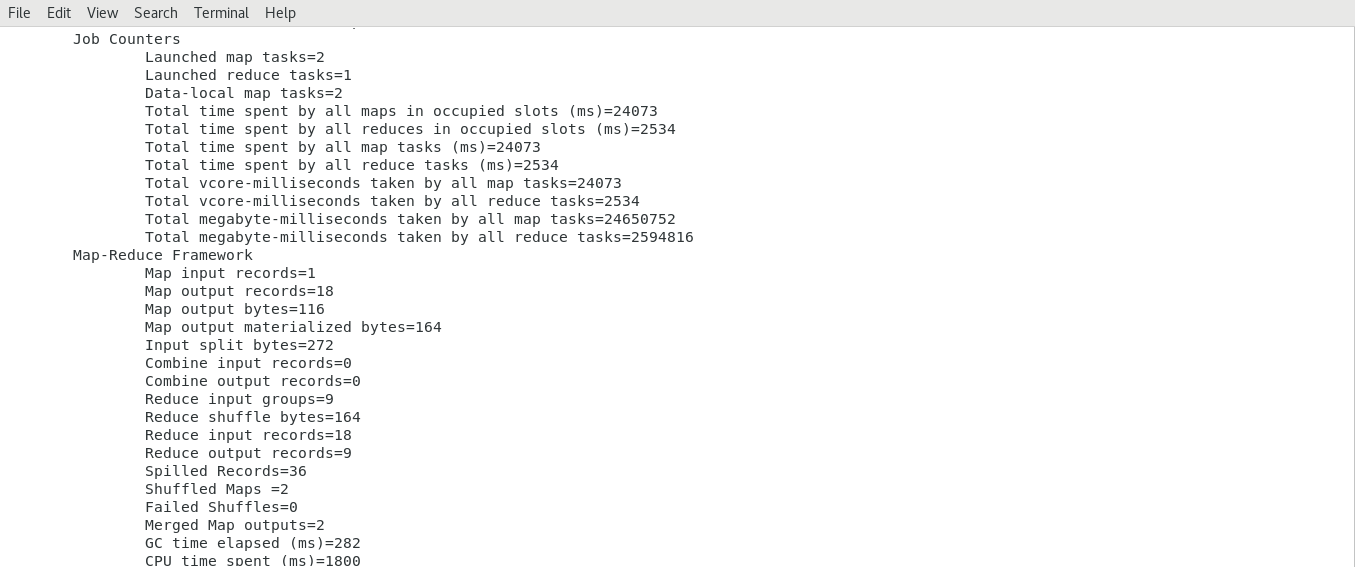


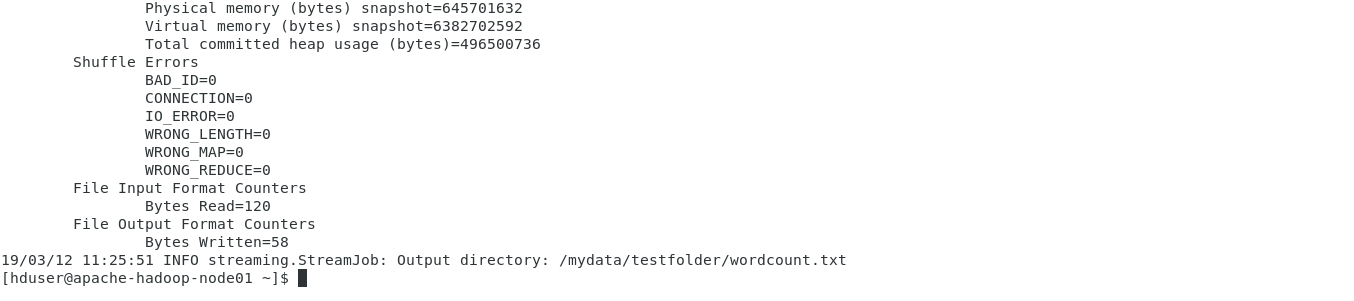


1. Run map reduce using streaming

hadoop jar /opt/hadoop-2.8.3/share/hadoop/tools/lib/hadoop-streaming-2.8.3.jar -file mapper.py -file reducer.py -mapper "python mapper.py" -reducer "python reducer.py" -input /mydata/testfolder/sample.txt -output /mydata/testfolder/wordcount.txt

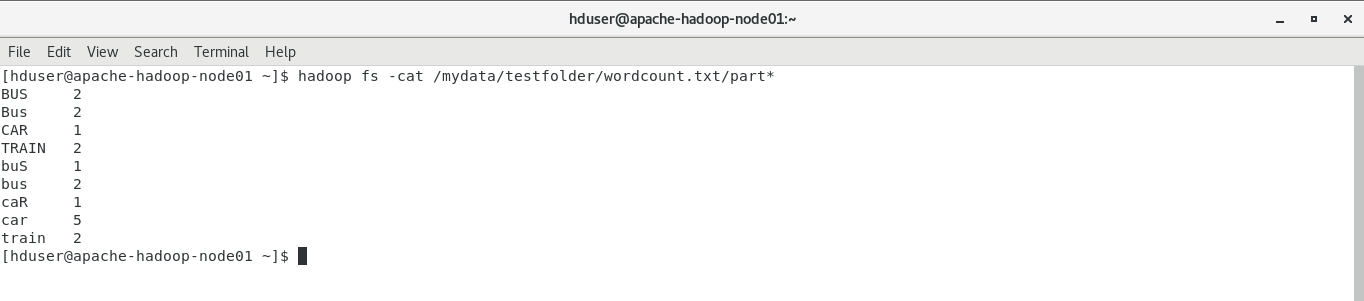






1. Check the output

hadoop fs -cat /mydata/testfolder/wordcount.txt/part\*



1. Check the Hadoop application information

