Running MapReduce WordCount Program in Python

Prepared By : Siddhartha Shakya

Before running the below commands make sure that you have :

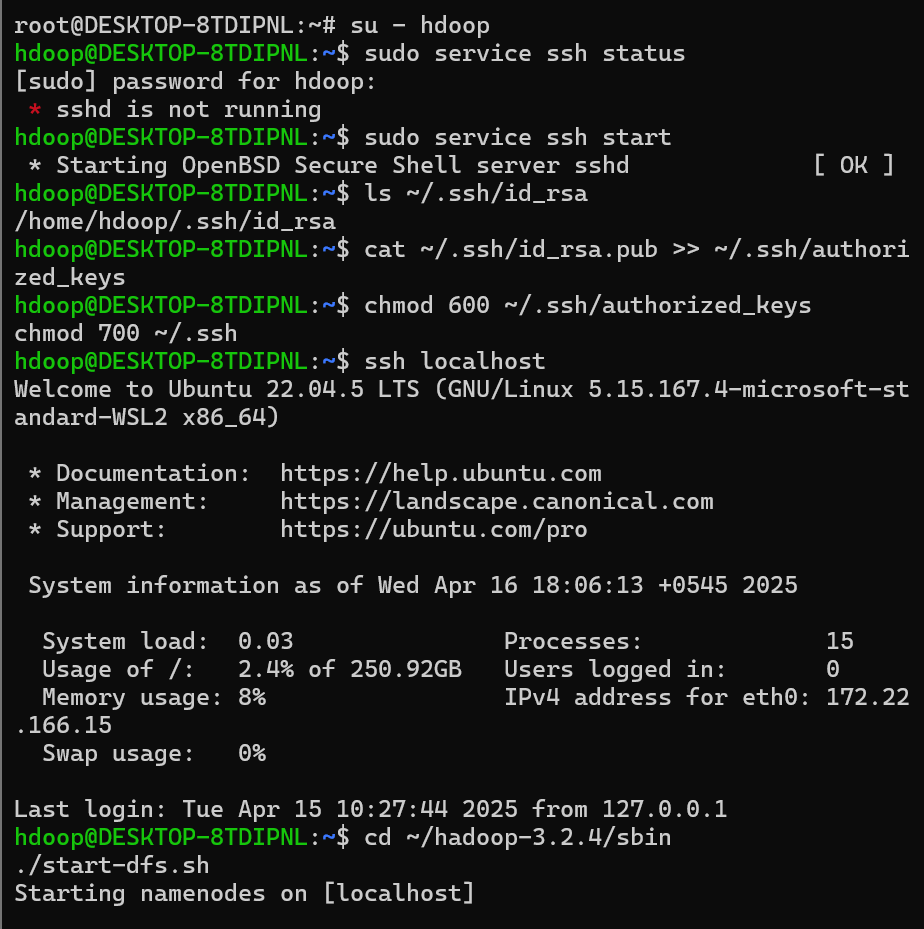
* Switched to the **hadoop user**.
* Navigated to the hadoop **Sbin** directory.
* Started the **Hadoop Distributed File System (HDFS).**
* Started the **YARN Resource Manager**.

jps

**If not** refer to the below image to complete the above tasks :

|  |
| --- |
|  |

If **jps** **(Java Processing Status)** lists all the above **running java processes**, then you are **good to continue** with the below commands**.**



A screen shot of a computer program

AI-generated content may be incorrect.

1. Python comes pre-installed in newer versions of Ubuntu. However, if you want to check which version is installed on your machine:

For python version check :

|  |
| --- |
| ~$ python3 --version |

A black background with blue text

AI-generated content may be incorrect.

1. If the python is not installed then you can install the python with the following code:

|  |
| --- |
| ~$ sudo apt install python3 |

1. Create **input\_word** folderand **insert a text file** in the **HDFS**.

|  |
| --- |
| ~$ hdfs dfs -mkdir /input\_word  ~$ hdfs dfs -put hadoop-3.2.4/LICENSE.txt /input\_word/ |

A screen shot of a computer screen

AI-generated content may be incorrect.

1. Verify the creation of the **directory** and the **file**.

|  |
| --- |
| ~$ hdfs dfs -ls /  ~$ hdfs dfs -ls /input\_word/  ~$ hdfs dfs -cat /input\_word/LICENSE.txt |

A screenshot of a computer program

AI-generated content may be incorrect.

1. Now we have to create a mapper and reducer. To create a mapper class first we have to create an empty file and add the code related.

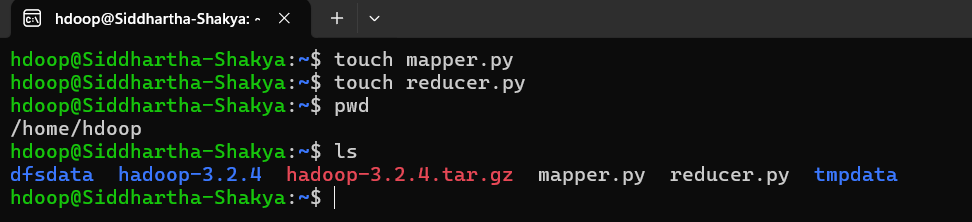
A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

1. Type “**touch mapper.py**” and press enter. Similarly for reducer “**touch reducer.py**” and press enter. You will have two files located in your **/home/hdoop** directory.



A screen shot of a computer screen

AI-generated content may be incorrect.

1. Open the mapper.py typing “**nano mapper.py**” and add the following code.

|  |
| --- |
| ~$ nano mapper.py |

Copy paste the below code :

|  |
| --- |
| *#!/usr/bin/env python3*  """mapper.py"""  import sys  *# input comes from STDIN (standard input)*  for line in sys.stdin:  *# remove leading and trailing whitespace*  line = line.strip()  *# split the line into words*  words = line.split()  *# increase counters*  for word in words:  *# write the results to STDOUT (standard output);*  *# what we output here will be the input for the*  *# Reduce step, i.e. the input for reducer.py*  *#*  *# tab-delimited; the trivial word count is 1*  **print**("%s\t%s" % (word, 1)) |

1. Now open the reducer file by typing “**reducer.py**” and press enter. add the following code in reducer.py.

|  |
| --- |
| ~$ nano reducer.py |

Copy paste the below code :

|  |
| --- |
| *#!/usr/bin/env python3*  """reducer.py"""  from operator import itemgetter  import sys  current\_word = None  current\_count = 0  word = None  *# input comes from STDIN*  for line in sys.stdin:  *# remove leading and trailing whitespace*  line = line.strip()  *# parse the input we got from mapper.py*  word, count = line.split("\t", 1)  *# convert count (currently a string) to int*  try:  count = int(count)  except ValueError:  *# count was not a number, so silently*  *# ignore/discard this line*  continue  *# this IF-switch only works because Hadoop sorts map output*  *# by key (here: word) before it is passed to the reducer*  if current\_word == word:  current\_count += count  else:  if current\_word is not None:  *# write result to STDOUT*  **print**("%s\t%s" % (current\_word, current\_count))  current\_count = count  current\_word = word  *# do not forget to output the last word if needed!*  if current\_word == word:  **print**("%s\t%s" % (current\_word, current\_count)) |

1. Now we have to give permission for the mapper and reducer. Type the following commands.

|  |
| --- |
| ~$ chmod +x /home/hdoop/mapper.py  ~$ chmod +x /home/hdoop/reducer.py |

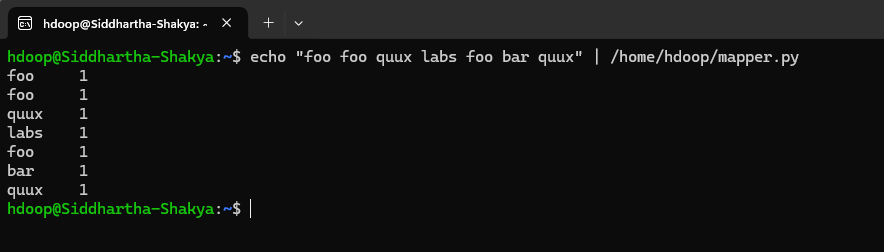
A screen shot of a computer program

AI-generated content may be incorrect.

Testing the mapper.py and reducer.py

1. To test the mapper type the following code and enter.

|  |
| --- |
| ~$ echo "foo foo quux labs foo bar quux" | /home/hdoop/mapper.py |

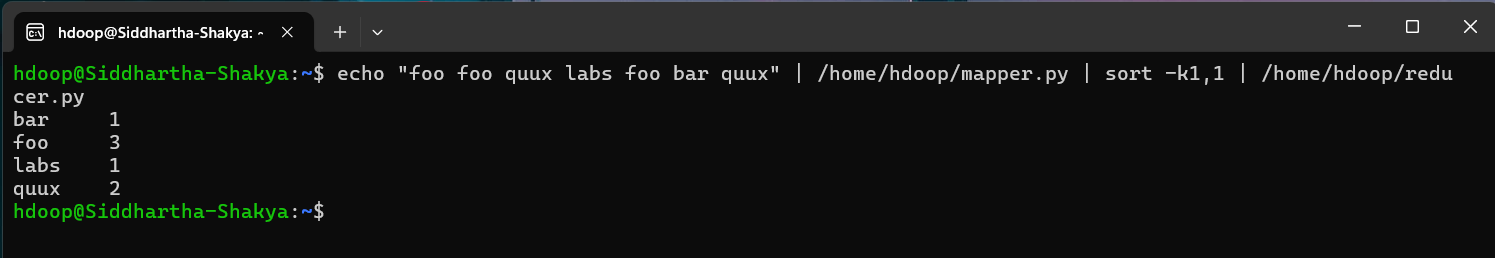


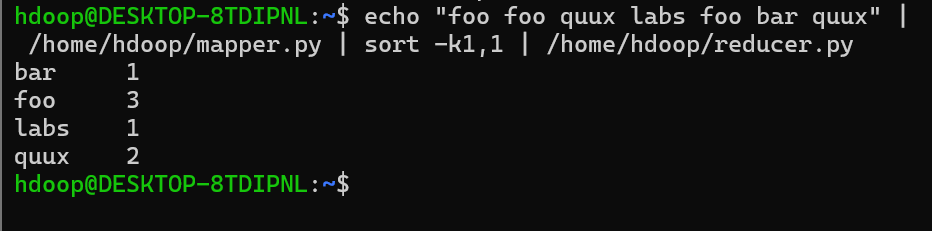
A screen shot of a computer

AI-generated content may be incorrect.

1. Now test the mapper and reducer code together

|  |
| --- |
| ~$ echo "foo foo quux labs foo bar quux" | /home/hdoop/mapper.py | sort -k1,1 | /home/hdoop/reducer.py |





Running the Mapper.py and reducer.py in HDFS

1. Now to execute the code run the following code.

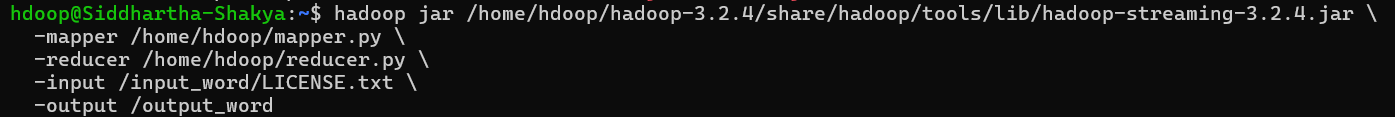
|  |
| --- |
| ~$ hadoop jar /home/hdoop/hadoop-3.2.4/share/hadoop/tools/lib/hadoop-streaming-3.2.4.jar \  -mapper /home/hdoop/mapper.py \  -reducer /home/hdoop/reducer.py \  -input /input\_word/LICENSE.txt \  -output /output\_word |

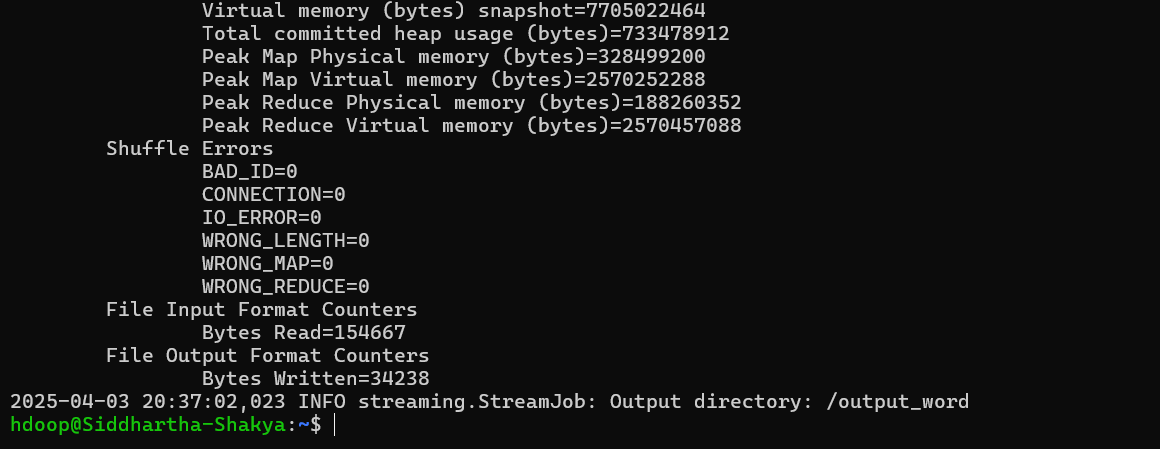
A screenshot of a computer

AI-generated content may be incorrect.

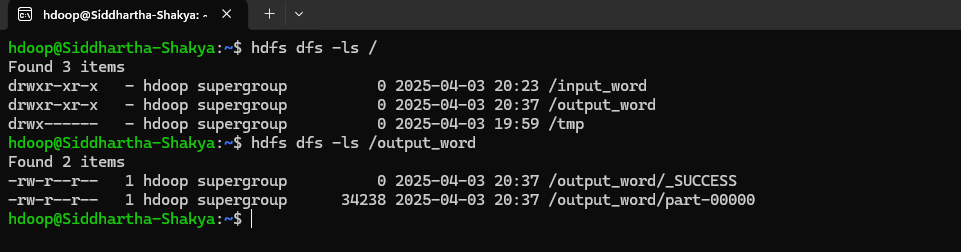
1. if you have output folder then you have to delete that

***Note: It is just the syntax. Please enter proper credentials as your system has been configured.***

******

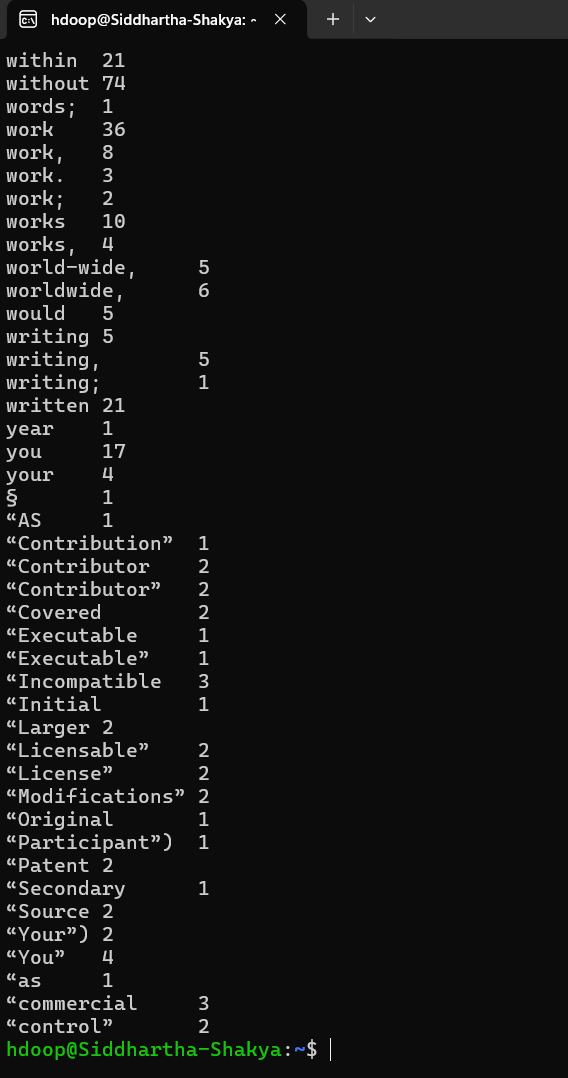


1. Now you will see the output folder and the output file. **output\_word** in my case as I have given that specific name while executing map reduce code.



1. Now you can test if the map reduce program in the python worked correctly or not by executing the following command.

|  |
| --- |
| ~$ hdfs dfs -cat /output\_word/part-00000 |



!! WELL DONE !!

1. References

* [Writing An Hadoop MapReduce Program In Python](https://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python/)