BDA Lab 6 Tasks

Notes:

- Keep in mind that output directory name will be your roll number else no marks would be awarded
- You are required to run your WordCount on the text file provided with this lab named "TheseWords.txt" which contains all the words in the tutorial.
- Must change “i19-1207” with your roll number at all places.
- Must change “SaadNaeem” at all places with your name.

Task-1

Creating your first WordCount MapReduce program in python. Use the pdf provided with this lab or follow the tutorial at:

- https://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python/
- You need to make changes in the code according to your problem statement and fix any bugs that are there in the code (Because there are few points that need to be fixed in the code).

Extra Help related to Task-1:

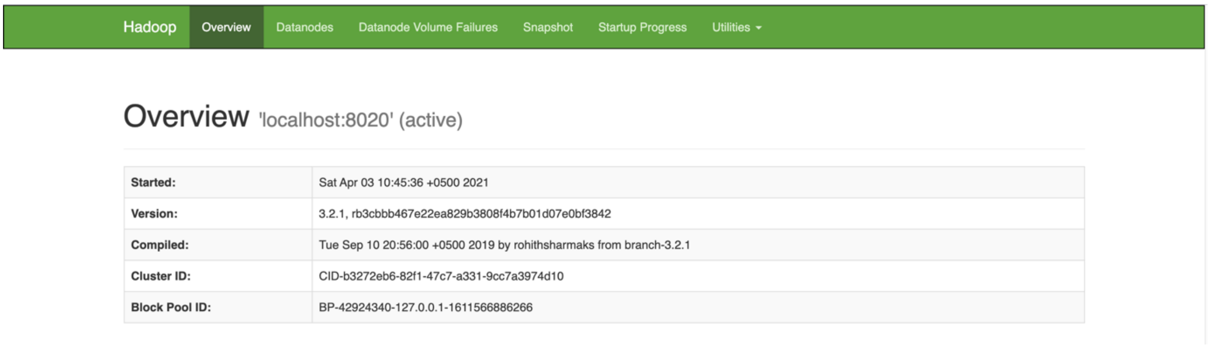
Information about hadoop streaming can be found at:
https://hadoop.apache.org/docs/r1.2.1/streaming.html

Hint-1: you need to get the streaming api jar file for your hadoop version
Hint-2: I have put the streaming api jar file in my hadoop installation at: /usr/local/Cellar/hadoop/3.2.1/libexec/share/hadoop/mapreduce/hadoop-streaming-3.2.1.jar
so when i run the MapReduce program in python i give the above path to "hadoop jar /usr/local/Cellar/hadoop/3.2.1/libexec/share/hadoop/mapreduce/hadoop-streaming-3.2.1.jar ...."

Debugging Tips:
No Spaces in file path

```
echo "foo foo quux labs foo bar quux" | python mapper.py
echo "foo foo quux labs foo bar quux" | python mapper.py | sort -k1,1 | python reducer.py
```

Task-2 (20% Marks)



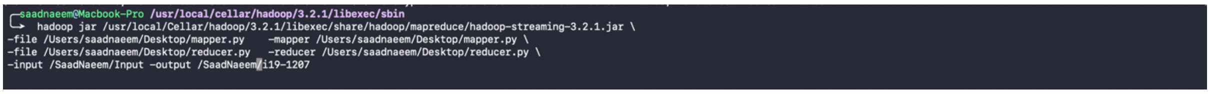
Overview 'localhost:8020' (active)

Started:	Sat Apr 03 10:45:36 +0500 2021
Version:	3.2.1, rb3cbbb467e22ea829b3808f4b7b01d07e0bf3842
Compiled:	Tue Sep 10 20:56:00 +0500 2019 by rohithsharmaks from branch-3.2.1
Cluster ID:	CID-b3272eb6-82f1-47c7-a331-9cc7a3974d10
Block Pool ID:	BP-42924340-127.0.0.1-1611566886266

Requirement 1:i19-1207-image-1.png

You are required to submit i19-1207-Image-1.png where i19-1207 is your roll number that shows that your installation is up and running and shows your unique block id and cluster id.

Task-3 (20% Marks)

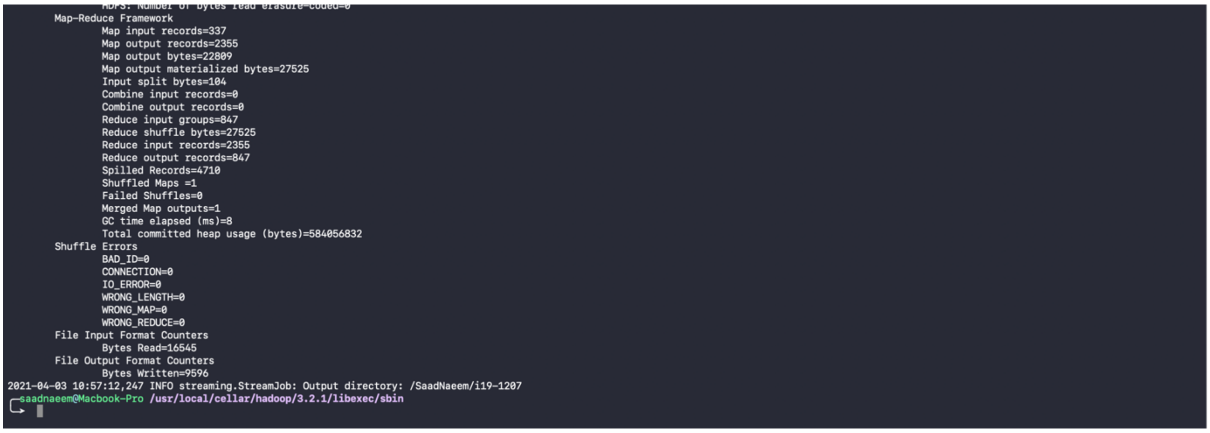


```
┌─saadnaeem@Macbook-Pro /usr/local/cellar/hadoop/3.2.1/libexec/sbin
└→  hadoop jar /usr/local/Cellar/hadoop/3.2.1/libexec/share/hadoop/mapreduce/hadoop-streaming-3.2.1.jar \
-file /Users/saadnaeem/Desktop/mapper.py    -mapper /Users/saadnaeem/Desktop/mapper.py \
-file /Users/saadnaeem/Desktop/reducer.py   -reducer /Users/saadnaeem/Desktop/reducer.py \
-input /SaadNaeem/Input -output /SaadNaeem/i19-1207
```

Requirement 2:i19-1207-image2.png

Submit “i19-1207-Image-2.png” that shows where were you when you executed the command to run map reduce in python and what was the command itself.

Task-4 (20% Marks)

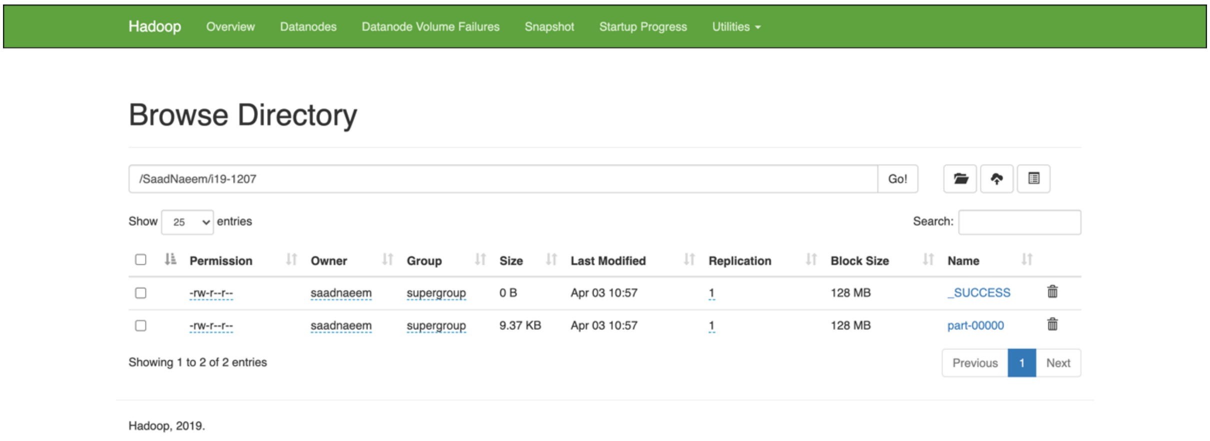


```
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=337
                Map output records=2355
                Map output bytes=22809
                Map output materialized bytes=27525
                Input split bytes=104
                Combine input records=0
                Combine output records=0
                Reduce input groups=847
                Reduce shuffle bytes=27525
                Reduce input records=2355
                Reduce output records=847
                Spilled Records=4710
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=8
                Total committed heap usage (bytes)=584056832
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=16545
        File Output Format Counters
                Bytes Written=9596
2021-04-03 10:57:12,247 INFO streaming.StreamJob: Output directory: /SaadNaeem/i19-1207
╭─saadnaeem@Macbook-Pro /usr/local/cellar/hadoop/3.2.1/libexec/sbin
╰─➤ █
```

Requirement 3:i19-1207-image-3.png

Submit “i19-1207-Image-3.png” that shows the user name of the system, the success of map reduce on terminal, and if you notice carefully the streaming API shows the output directory where the results are stored.

Task-5 (20% Marks)



Browse Directory

/SaadNaeem/i19-1207 Go!

Show 25 entries Search:

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
□	-rw-r--r--	saadnaeem	supergroup	0 B	Apr 03 10:57	1	128 MB	_SUCCESS
□	-rw-r--r--	saadnaeem	supergroup	9.37 KB	Apr 03 10:57	1	128 MB	part-00000

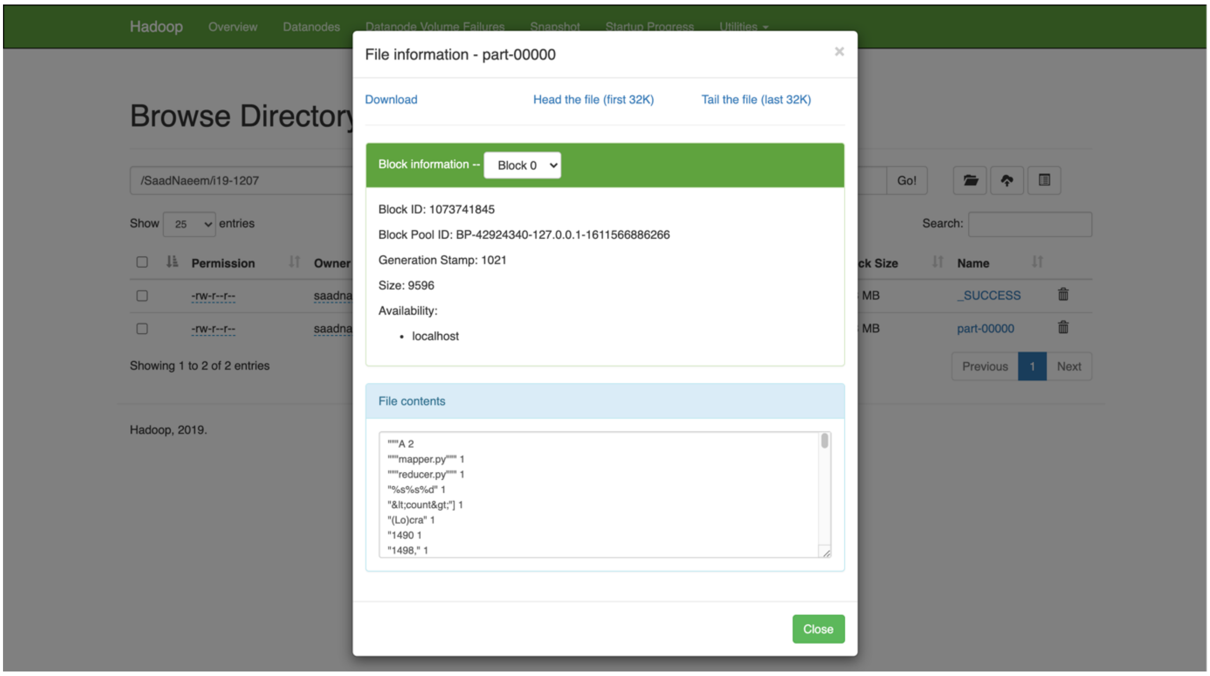
Showing 1 to 2 of 2 entries

Hadoop, 2019.

Requirement 4:i19-1207-image-4.png

i19-1207-image-4.png shows that your Datanode is working properly and the MapReduce success in terminal has indeed reflected here.

Task-6 (20% Marks)



File information - part-00000

Download | Head the file (first 32K) | Tail the file (last 32K)

Block information -- Block 0

Block ID: 1073741845

Block Pool ID: BP-42924340-127.0.0.1-1611566886266

Generation Stamp: 1021

Size: 9596

Availability:

- localhost

File contents

```
""""A 2
""""mapper.py"""" 1
""""reducer.py"""" 1
"%s%s%d" 1
"&lt;count&gt;"] 1
"(Lo)cra" 1
"1490 1
"1498," 1
```

Requirement 5: i19-1207-image-5

Moment of truth " i19-1207-image-5.png" shows the actual results and verifies your MapReduce logic.

Bonus Task

Task: There is a redundancy in Mapper/Reducer logic you need to find it and fix it. This requires and understanding of what "Mapper" does and what "Reducer" does. And what "streaming.jar" is doing.

Requirement: **You can only do this task if you have done all the mandatory tasks &** Only submit your "mapper.py" and "reducer.py" if you completed this task.