IT 609 - Big Data Processing Lab 3

ID - 202218061
NAME - JATAN SAHU

Aim

- To implement the Wordcount algorithm in MapReduce
- To compare the execution time in multimode and single-node clusters.

Step 1:

1. First, we have to create a code for the mapper and reducer
2. Save the file in .txt format

In our case

a) mapper.py
b) reducer.py
c) wordcountdata.txt

mapper	26-02-2023 21:53	Python File	1 KB
reducer	26-02-2023 21:54	Python File	2 KB
wordcountdata	26-02-2023 17:26	Text Document	5 KB

A) Code for Mapper.py

```
#!/usr/bin/env python
"""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))
```

B) Code for reducer.py

```
#!/usr/bin/env python
"""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:
```

```
            # 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))
```

STEP 02:-

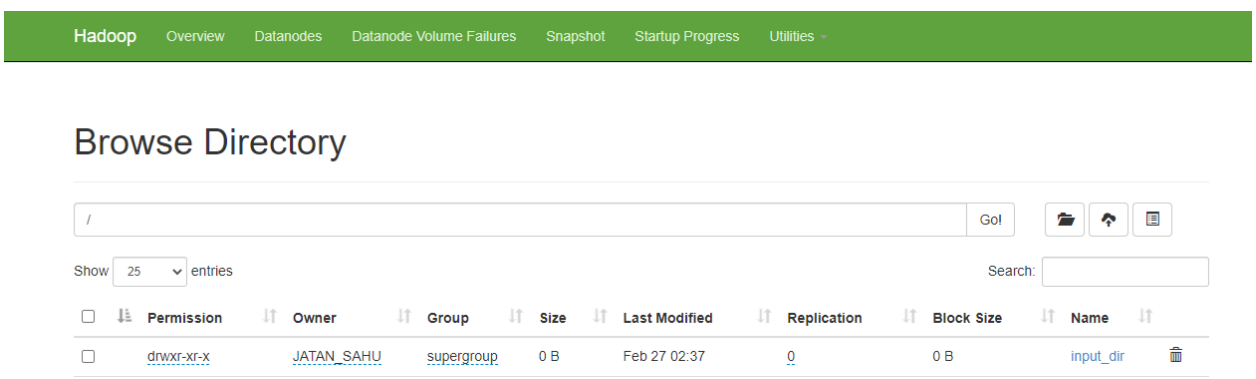
1.Start Hadoop using command prompt
>start-all.cmd

Instead

```
C:\Hadoop\hadoop-2.9.2>start-dfs.cmd

C:\Hadoop\hadoop-2.9.2>start-yarn.cmd
starting yarn daemons
```

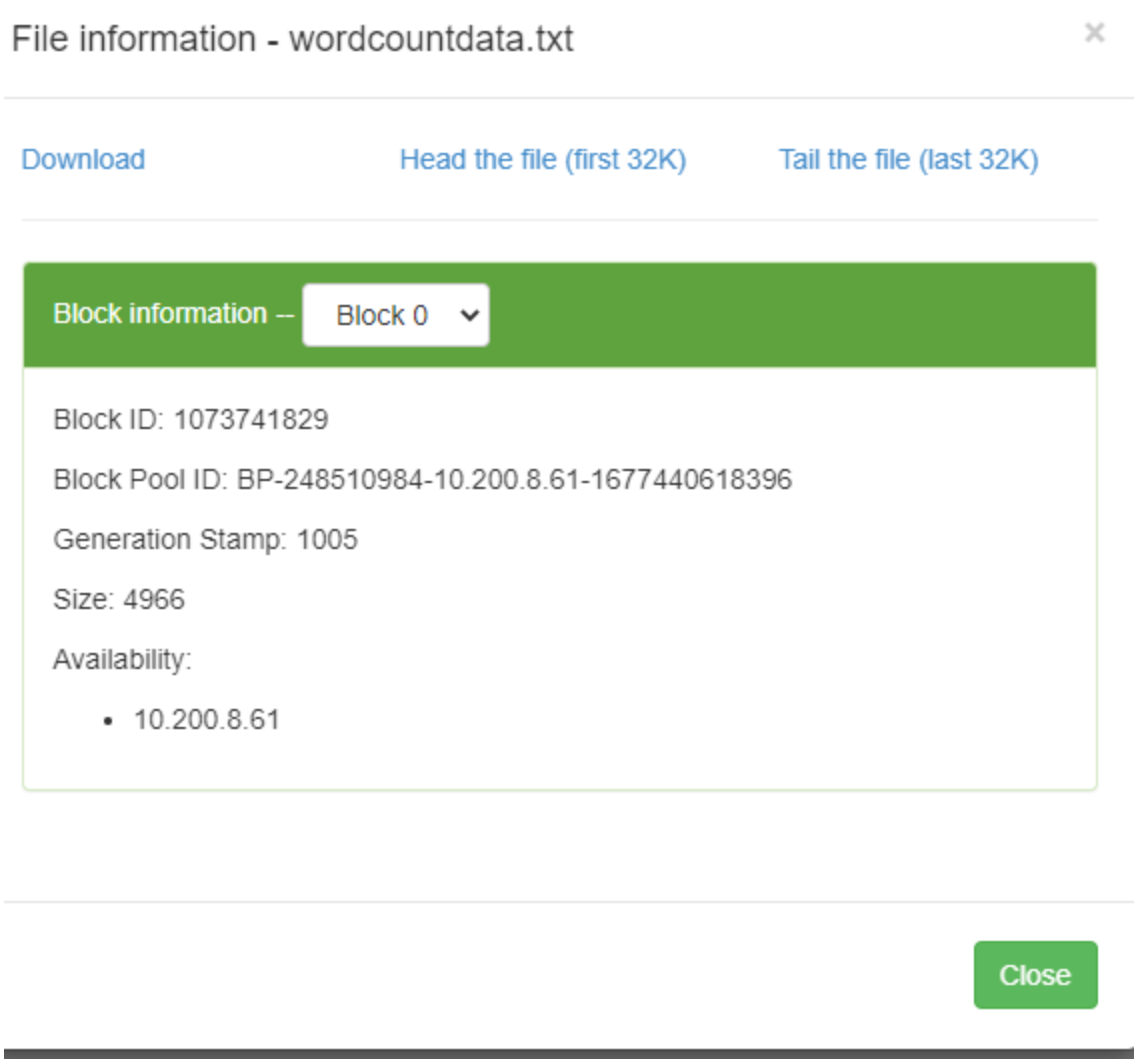
2.After installing hadoop successfully create input and output directory using command
>hadoop fs -mkdir /**<input directory name> (input_dir** in our case)

```
C:\Hadoop\hadoop-2.9.2>hadoop fs -mkdir /input_dir

C:\Hadoop\hadoop-2.9.2>hadoop fs -put C:/Hadoop/lab3/wordcountdata.txt /input_dir
```



3. Put data file(wordcount) in the input directory
>hadoop fs -put **<PATH OF DATA FILE> /input_dir**

```
C:\Hadoop\hadoop-2.9.2>hadoop fs -put C:/Hadoop/lab3/wordcountdata.txt /input_dir
```



4.Download streaming jar for 2.9.2 from below link

https://jar-download.com/artifacts/org.apache.hadoop/hadoop-streaming?p=2

5.Run mapper and reducer file in hadoop

>hadoop jar **<streaming jar_path>** -file **<mapper_path>** -mapper "python **mapper.py**" -file **<reducer_path>** -reducer "python **reducer.py**" -input <**input_path>** -output <**output_path>**

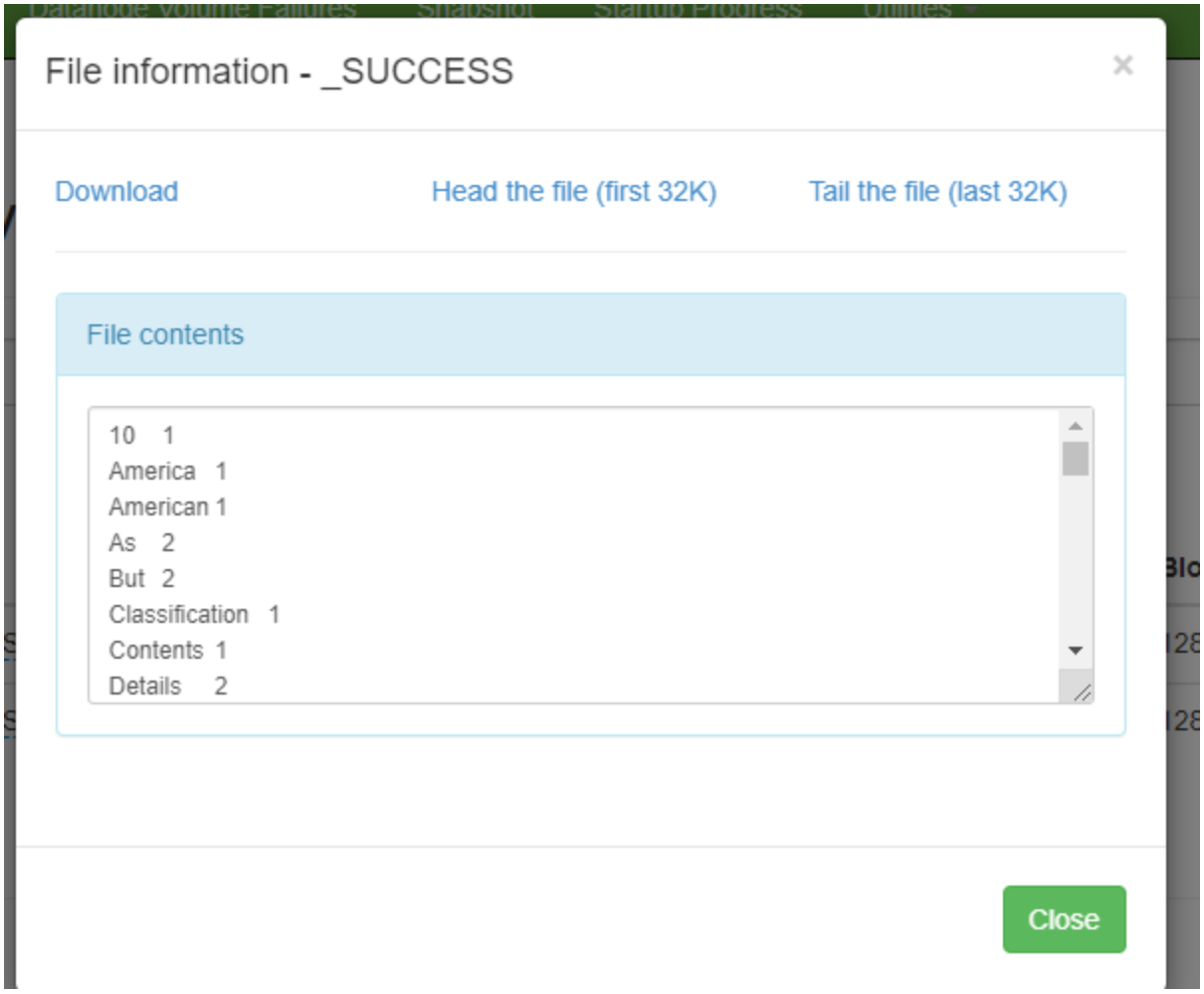
```
C:\Windows\system32>start-all.cmd
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons

C:\Windows\system32>hdfs dfs -ls /
Found 3 items
drwxr-xr-x   - JATAN_SAHU supergroup          0 2023-02-27 02:37 /input_dir
drwx------   - JATAN_SAHU supergroup          0 2023-02-27 02:12 /tmp
drwxr-xr-x   - JATAN_SAHU supergroup          0 2023-02-27 01:37 /user

C:\Windows\system32>hadoop jar C:\Hadoop\lab3\hadoop-streaming-2.9.2.jar -file "C:/Hadoop/lab3/mapper.py" -mapper "python mapper.py" -file "C:/Hadoop/lab3/reducer.py" -
reducer "python reducer.py" -input /input_dir -output /output
```

6. TOTAL TIME

```
Job Counters
        Launched map tasks=2
        Launched reduce tasks=1
        Data-local map tasks=2
        Total time spent by all maps in occupied slots (ms)=23435
        Total time spent by all reduces in occupied slots (ms)=6622
        Total time spent by all map tasks (ms)=23435
        Total time spent by all reduce tasks (ms)=6622
```

OUTPUT : -

```
23/02/27 12:02:55 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1677479480537_0001
23/02/27 12:02:56 INFO impl.YarnClientImpl: Submitted application application_1677479480537_0001
23/02/27 12:02:56 INFO mapreduce.Job: The url to track the job: http://DESKTOP-Q6LO070:8088/proxy/application_1677479480537_0001/
23/02/27 12:02:56 INFO mapreduce.Job: Running job: job_1677479480537_0001
23/02/27 12:03:11 INFO mapreduce.Job: Job job_1677479480537_0001 running in uber mode : false
23/02/27 12:03:11 INFO mapreduce.Job:  map 0% reduce 0%
23/02/27 12:03:26 INFO mapreduce.Job:  map 100% reduce 0%
23/02/27 12:03:35 INFO mapreduce.Job:  map 100% reduce 100%
23/02/27 12:03:36 INFO mapreduce.Job: Job job_1677479480537_0001 completed successfully
23/02/27 12:03:36 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=8168
                FILE: Number of bytes written=626390
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=7651
                HDFS: Number of bytes written=3643
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=23435
                Total time spent by all reduces in occupied slots (ms)=6622
                Total time spent by all map tasks (ms)=23435
                Total time spent by all reduce tasks (ms)=6622
```


File information - _SUCCESS
Download
Head the file (first 32K)
Tail the file (last 32K)
File contents
10 1
America 1
American 1
As 2
But 2
Classification 1
Contents 1
Details 2
Close

File information - part-00000

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

Block information -- Block 0

Block ID: 1073741844

Block Pool ID: BP-248510984-10.200.8.61-1677440618396

Generation Stamp: 1020

Size: 3643

Availability:

- DESKTOP-Q6LO070

File contents

```
10  1
America  1
American 1
As  2
But  2
Classification  1
Contents 1
Details  2
```

```
23/02/27 12:03:35 INFO mapreduce.Job:  map 100% reduce 100%
23/02/27 12:03:36 INFO mapreduce.Job: Job job_1677479480537_0001 completed success
23/02/27 12:03:36 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=8168
                FILE: Number of bytes written=626390
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=7651
                HDFS: Number of bytes written=3643
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=23435
                Total time spent by all reduces in occupied slots (ms)=6622
                Total time spent by all map tasks (ms)=23435
                Total time spent by all reduce tasks (ms)=6622
                Total vcore-milliseconds taken by all map tasks=23435
                Total vcore-milliseconds taken by all reduce tasks=6622
                Total megabyte-milliseconds taken by all map tasks=23997440
                Total megabyte-milliseconds taken by all reduce tasks=6780928
        Map-Reduce Framework
                Map input records=44
                Map output records=810
                Map output bytes=6542
                Map output materialized bytes=8174
                Input split bytes=202
                Combine input records=0
                Combine output records=0
                Reduce input groups=381
                Reduce shuffle bytes=8174
                Reduce input records=810
                Reduce output records=381
                Spilled Records=1620
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=381
                CPU time spent (ms)=3886
                Physical memory (bytes) snapshot=737505280
                Virtual memory (bytes) snapshot=868007936
```

```
23/02/27 12:03:36 INFO streaming.StreamJob: Output directory: /output

C:\Windows\system32>hdfs dfs -ls /output
Found 2 items
-rw-r--r--   1 JATAN_SAHU supergroup          0 2023-02-27 12:03 /output/_SUCCESS
-rw-r--r--   1 JATAN_SAHU supergroup       3643 2023-02-27 12:03 /output/part-00000

C:\Windows\system32>hdfs dfs -cat /output/part-00000
10      1
America 1
American        1
```


```
10      1
America 1
American        1
```


```
As      2
But     2
Classification  1
Contents        1
Details 2
Different       1
External        1
Fantasy 1
Fiction 1
However 3
Hyphenated      1
In      4
Jane    1
JavaScript      1
Modern  1
Month   1
Most    2
National        1
Nebula  1
Novel   2
Novelette       1
Novelist        1
Novella 1
Numerous        1
Ph.D.   1
Please  1
References      1
Science 1
See     1
Short   1
Smiley  2
Software        2
Sources 1
The     5
There   1
These   1
This    2
To      1
Unix    1
Unsourced       1
Usually 1
Variations      1
When    1
```

```
Unsourced       1
Usually 1
Variations      1
When    1
Wikipedia       1
Word    4
Writers 1
Writing 1
a       28
about   1
abstracts       1
academia        1
academic        1
accept  1
acceptable      2
accordingly     1
across  1
adding  1
adjective       1
advent  1
advertising     1
already 1
also    5
an      4
and     23
any     2
application     1
applications    2
arbiter 1
arbitrary       1
are     4
articles        1
as      9
assignments     1
at      3
automatically   1
average 1
award   1
barring 1
be      8
because 2
behavior        1
being   1
between 2
bibliographies  1
bookmarklet     1
books   1
bottom  1
boundaries      1
boundary        1
broad   1
broadly 1
```

```
broad   1
broadly 1
browsers        1
but     2
by      5
calculate       1
can     5
captions        1
case    1
categories      1
categorise      1
category        1
certain 2
challenged      1
chapter 1
character       2
characters      2
charge  1
children        1
choice  1
citations       1
cite    1
client  1
commonly        1
compounds       1
conjunctions    1
consensus       3
consistent      1
converting      1
costless        1
could   1
count   11
counted 2
counting        6
counts  3
define  1
defined 1
definition      3
definitions     3
dependent       1
depending       2
depends 1
details 2
```

```
toward  2
trait   1
translation     1
translators     1
tremendously    1
typewriters     1
typical 1
typically       1
typing  1
under   1
universities    1
up      2
update  1
used    4
users   1
usually 3
variation       1
variations      2
varies  1
various 1
vary    2
varying 2
via     2
was     1
watch   1
wayside 1
wc      1
web     1
website 1
were    3
what    1
when    2
whether 1
which   4
while   3
whitespace      1
widespread      1
with    2
within  1
word    24
words   21
work    1
workers 1
writer  1

C:\Windows\system32>s
```