

WORD COUNT PROGRAM

Hadoop vs Python

Group 3

- Cecilia Inez Reva Manurung (2106636994)
- Gemilang Bagas Ramadhani (2006535205)
- Laode Alif Ma'sum Sidrajat Raja Ika (2106731213)
- Zalfy Putra Rezky (2106731453)



Link Github

<https://github.com/zalfyputra/hadoop-vs-python>

Installing Hadoop on Ubuntu 20.04

<https://medium.com/@festusmorumbasi/installing-hadoop-on-ubuntu-20-04-4610b6e0391e>

Install Java 1.8

```
● hadoop@vmbox:~$ sudo apt install openjdk-8-jdk -y
Reading package lists... Done
Building dependency tree
```



```
● hadoop@vmbox:~$ java -version
openjdk version "1.8.0_362"
OpenJDK Runtime Environment (build 1.8.0_362-8u372-ga-us1-0ubuntu1~20.04-b09)
OpenJDK 64-Bit Server VM (build 25.362-b09, mixed mode)
```

```
○ hadoop@vmbox:~$
```

Configure hadoop Env and XMLs

```
hadoop@vmbox:~$ sudo nano /home/hadoop/hadoop/etc/hadoop/hadoop-env.sh
hadoop@vmbox:~$ sudo nano /home/hadoop/hadoop/etc/hadoop/core-site.xml
hadoop@vmbox:~$ sudo nano /home/hadoop/hadoop/etc/hadoop/hdfs-site.xml
hadoop@vmbox:~$ sudo nano /home/hadoop/hadoop/etc/hadoop/mapred-site.xml
● hadoop@vmbox:~$ sudo nano /home/hadoop/hadoop/etc/hadoop/yarn-site.xml
○ hadoop@vmbox:~$
```

Install Hadoop 3.3.2

```
hadoop@vmbox:~$ wget https://downloads.apache.org/hadoop/common/hadoop-3.3.2/hadoop-3.3.2.tar.gz
--2023-06-21 13:40:45-- https://downloads.apache.org/hadoop/common/hadoop-3.3.2/hadoop-3.3.2.tar.gz
Resolving downloads.apache.org (downloads.apache.org) ... 88.99.95.219, 135.181.214.104, 2a01:4f9:*****
Connecting to downloads.apache.org (downloads.apache.org)|88.99.95.219|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 638660563 (609M) [application/x-gzip]
Saving to: 'hadoop-3.3.2.tar.gz'
```

Format HDFS NameNode

```
● hadoop@vmbox:~/hadoop/etc/hadoop$ hdfs namenode -format
WARNING: /home/hadoop/hadoop/logs does not exist. Creating.
2023-06-21 15:07:31,421 INFO namenode.NameNode: STARTUP_MSG:
*****STARTUP_MSG: Starting NameNode
STARTUP_MSG: host = vmbox/127.0.1.1
STARTUP_MSG: args = [-format]
STARTUP_MSG: version = 3.3.2
```

Install and configure OpenSSH

```
● hadoop@vmbox:~/hadoop/etc/hadoop$ sudo apt install openssh-server openssh-client -y
Reading package lists... Done
Building dependency tree
Reading state information... Done
```

```
● hadoop@vmbox:~/hadoop/etc/hadoop$ ssh-keygen -t rsa
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hadoop/.ssh/id_rsa):
Created directory '/home/hadoop/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/hadoop/.ssh/id_rsa
Your public key has been saved in /home/hadoop/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:JW4j5U3zdHTCr8hp4vAySElV300S5Y9aI4fsdjwOE0U hadoop@vmbox
The key's randomart image is:
+---[RSA 3072]---+
| oo. .o .|
| .o. +..o |
| .+ E=....|
| .+ *o+o. .|
| ...S..+= .|
| 00..o= X o|
| . . =.0 .|
| . o *+. |
| + o.. |
+---[SHA256]---+
○ hadoop@vmbox:~/hadoop/etc/hadoop$
```

```
○ hadoop@vmbox:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ECDSA key fingerprint is SHA256:QBcnDp2QJGAQydtFwoHHFS62ohvogzPncMXegvhpdgo.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
Welcome to Ubuntu 20.04.6 LTS (GNU/Linux 5.15.0-75-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:    https://landscape.canonical.com
 * Support:       https://ubuntu.com/advantage

Expanded Security Maintenance for Applications is not enabled.

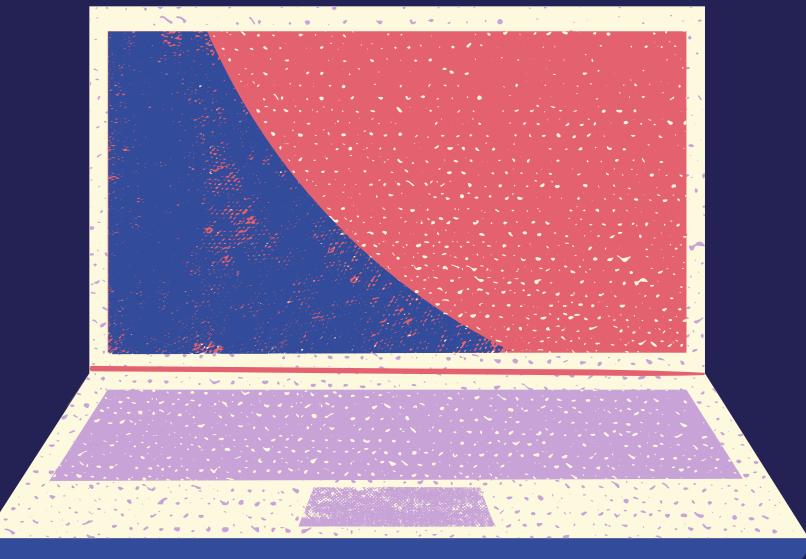
2 updates can be applied immediately.
2 of these updates are standard security updates.
To see these additional updates run: apt list --upgradable

Enable ESM Apps to receive additional future security updates.
See https://ubuntu.com/esm or run: sudo pro status

New release '22.04.2 LTS' available.
```

```
● hadoop@vmbox:~$ sudo cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
● hadoop@vmbox:~$ sudo chmod 640 ~/.ssh/authorized_keys
```

SPESIFIKASI KOMPUTER



Host

CPU: AMD Ryzen 7 5800 3.2 GHz
GPU: NVIDIA GeForce RTX 3060
RAM: 32 GB
OS: Windows 11 Home 64-bit

VirtualBox

CPU: 8-core
RAM: 28 GB
OS: Ubuntu 20.04

File Test

Name	Size
personae.txt	125,2 kB
melville.txt	2,0 MB
shakespeare.txt	12,0 MB
enwik8.txt	100,0 MB
enwik9.txt	1,0 GB

References

- <https://corpus.canterbury.ac.nz/descriptions/>
- <https://www.gutenberg.org/browse/scores/top>
- <https://mattmahoney.net/dc/textdata.html>

Start Hadoop Server

```
hadoop@vmbox:~$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [vmbox]
vmbox: Warning: Permanently added 'vmbox' (ECDSA) to the list of known hosts.
hadoop@vmbox:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hadoop@vmbox:~$ jps
7345 Jps
6392 DataNode
6828 ResourceManager
6990 NodeManager
6606 SecondaryNameNode
hadoop@vmbox:~$ █
```

Commands

- 1.Create a new folder in hdfs
- 2.Upload text file to the created folder
- 3.Run Hadoop MapReduce JAR file to count words of the text file
- 4.Store the output to a text file in local

example of 100 KB text file

```
hadoop@vmbox:~$ hdfs dfs -mkdir /100kb
hadoop@vmbox:~$ hdfs dfs -put /home/hadoop/Downloads/wordcount/personae.txt /100kb
hadoop@vmbox:~$ time hadoop jar /home/hadoop/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.2.jar wordcount /100kb /output-100kb
hadoop@vmbox:~$ hdfs dfs -get /output-100kb/part-r-00000 /home/hadoop/Downloads/wordcount/output-100kb.txt
```

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities -

Browse Directory

/output-100kb Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	0 B	Jun 09 10:01	1	128 MB	_SUCCESS
-rw-r--r--	hadoop	supergroup	48.81 KB	Jun 09 10:01	1	128 MB	part-r-00000

Showing 1 to 2 of 2 entries Previous 1 Next

Hadoop, 2022.

Hadoop

File size: 100 KB

Time: 15.821 s

+ Code + Markdown | ▶ Run All ✖ Clear All Outputs ✖ Go To ...

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL ... bash-hadoop +

```
Map output records=22960
Map output bytes=213603
Map output materialized bytes=70924
Input split bytes=105
Combine input records=22960
Combine output records=5317
Reduce input groups=5317
Reduce shuffle bytes=70924
Reduce input records=5317
Reduce output records=5317
Spilled Records=10634
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=55
CPU time spent (ms)=1200
Physical memory (bytes) snapshot=518696960
Virtual memory (bytes) snapshot=5160079360
Total committed heap usage (bytes)=424673280
Peak Map Physical memory (bytes)=314343424
Peak Map Virtual memory (bytes)=2577403904
Peak Reduce Physical memory (bytes)=204353536
Peak Reduce Virtual memory (bytes)=2582675456
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=125179
File Output Format Counters
Bytes Written=49986
real user sys
0m15,821s
0m4,307s
0m0,359s
```

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities -

Browse Directory

/output-1mb

Show 25 entries

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	0 B	Jun 09 10:09	1	128 MB	_SUCCESS
-rw-r--r--	hadoop	supergroup	483.05 KB	Jun 09 10:09	1	128 MB	part-r-00000

Showing 1 to 2 of 2 entries

Previous 1 Next

Hadoop, 2022.

Hadoop

File size: 1 MB

Time: 16.535 s

+ Code + Markdown | ▶ Run All ⌂ Clear All Outputs ⌒ Go To ...

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL ...

bash-hadoop + × ⌂ ⌒ ⌈ ⌉

```
Map output bytes=19852149
Map output materialized bytes=1826437
Input split bytes=107
Combine input records=2057725
Combine output records=124282
Reduce input groups=124282
Reduce shuffle bytes=1826437
Reduce input records=124282
Reduce output records=124282
Spilled Records=248564
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=69
CPU time spent (ms)=4250
Physical memory (bytes) snapshot=637353984
Virtual memory (bytes) snapshot=5160837120
Total committed heap usage (bytes)=557318144
Peak Map Physical memory (bytes)=423309312
Peak Map Virtual memory (bytes)=2578030592
Peak Reduce Physical memory (bytes)=214044672
Peak Reduce Virtual memory (bytes)=2582806528
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=11952731
File Output Format Counters
Bytes Written=1346479
real    0m16,535s
user    0m3,624s
sys     0m0,341s
```

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities -

Browse Directory

/output-100kb

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	0 B	Jun 09 10:01	1	128 MB	_SUCCESS
-rw-r--r--	hadoop	supergroup	48.81 KB	Jun 09 10:01	1	128 MB	part-r-00000

Showing 1 to 2 of 2 entries

Previous 1 Next

Hadoop, 2022.

Hadoop

File size: 10 MB

Time: 16.745 s

+ Code + Markdown | ▶ Run All ⌂ Clear All Outputs ⌈ Go To ...

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL ...

bash-hadoop + × ⌂ ⌂ ⌈ ⌈ ⌊ ⌉ ⌉

```
Map output bytes=3332993
Map output materialized bytes=665788
Input split bytes=103
Combine input records=340609
Combine output records=43706
Reduce input groups=43706
Reduce shuffle bytes=665788
Reduce input records=43706
Reduce output records=43706
Spilled Records=87412
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=67
CPU time spent (ms)=2350
Physical memory (bytes) snapshot=530612224
Virtual memory (bytes) snapshot=5165907968
Total committed heap usage (bytes)=476577792
Peak Map Physical memory (bytes)=325156864
Peak Map Virtual memory (bytes)=2582519808
Peak Reduce Physical memory (bytes)=205455360
Peak Reduce Virtual memory (bytes)=2584117248
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=1980046
File Output Format Counters
Bytes Written=494640
real    0m16,745s
user    0m3,424s
sys     0m0,587s
```

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities -

Browse Directory

/output-1mb

Show 25 entries

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	0 B	Jun 09 10:09	1	128 MB	_SUCCESS
-rw-r--r--	hadoop	supergroup	483.05 KB	Jun 09 10:09	1	128 MB	part-r-00000

Showing 1 to 2 of 2 entries

Previous 1 Next

Hadoop, 2022.

Hadoop

File size: 100 MB

Time: 36.35 s

+ Code + Markdown | ▶ Run All ⌂ Clear All Outputs ⌈ Go To ... ⌋ Python 3.8.10

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL ... bash-hadoop + ⌄ ⌂ ⌈ ⌋ ⌁

```
Map output bytes=151864357
Map output materialized bytes=30607972
Input split bytes=103
Combine input records=15317118
Combine output records=3453417
Reduce input groups=1439355
Reduce shuffle bytes=30607972
Reduce input records=1439355
Reduce output records=1439355
Spilled Records=4892772
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=108
CPU time spent (ms)=20440
Physical memory (bytes) snapshot=762966016
Virtual memory (bytes) snapshot=5172043776
Total committed heap usage (bytes)=722993152
Peak Map Physical memory (bytes)=504160256
Peak Map Virtual memory (bytes)=2587963392
Peak Reduce Physical memory (bytes)=258805760
Peak Reduce Virtual memory (bytes)=2584080384
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=100000000
File Output Format Counters
Bytes Written=24937465
real    0m36,350s
user    0m3,609s
sys     0m0,565s
```

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities -

Browse Directory

/output-1gb Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	0 B	Jun 09 10:16	1	128 MB	_SUCCESS
-rw-r--r--	hadoop	supergroup	170.1 MB	Jun 09 10:16	1	128 MB	part-r-00000

Showing 1 to 2 of 2 entries Previous 1 Next

Hadoop, 2022.

Hadoop

File size: 1 GB

Time: 72.668 s

+ Code + Markdown | ▶ Run All ⌂ Clear All Outputs ⌈ Go To ... ⌋ Python 3.8.10

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL ... bash-hadoop + ⌂ ⌂ ⌈ ⌋

```

Map output materialized bytes=291854430
Input split bytes=808
Combine input records=148264166
Combine output records=32298359
Reduce input groups=8859143
Reduce shuffle bytes=291854430
Reduce input records=13382052
Reduce output records=8859143
Spilled Records=45680411
Shuffled Maps =8
Failed Shuffles=0
Merged Map outputs=8
GC time elapsed (ms)=3816
CPU time spent (ms)=205470
Physical memory (bytes) snapshot=5403594752
Virtual memory (bytes) snapshot=23239303168
Total committed heap usage (bytes)=5448400896
Peak Map Physical memory (bytes)=612532224
Peak Map Virtual memory (bytes)=2595045376
Peak Reduce Physical memory (bytes)=776015872
Peak Reduce Virtual memory (bytes)=2581053440
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=1000028672
File Output Format Counters
Bytes Written=178363506
real    1m12,668s
user    0m3,974s
sys     0m0,288s

```

```
wordcount.py > ...
1 import sys
2 import csv
3 import string
4 from functools import reduce
5 import time
6 import os
7
8 def mapper(text):
9     # Remove punctuation and convert text to lowercase
10    text = text.translate(str.maketrans("", "", string.punctuation)).lower()
11
12    # Split the text into words
13    words = text.split()
14
15    # Emit each word with a count of 1
16    return [(word, 1) for word in words]
17
18 def reducer(word_counts, word_count):
19    word, counts = word_count
20
21    # Sum the counts for each word
22    if word in word_counts:
23        word_counts[word] += counts
24    else:
25        word_counts[word] = counts
26
27    return word_counts
28
29 # Check if the input file name is provided as a command-line argument
30 if len(sys.argv) < 2:
31     print("Please provide the input file name as a command-line argument.")
32     sys.exit(1)
33
34 input_file = sys.argv[1]
35
36 # Read text from file with explicit encoding
37 with open(input_file, 'r', encoding='utf-8') as file:
38     text = file.read()
39
40 # Start measuring the time
41 start_time = time.time()
42
43 # Map step: split the text into words and emit each word with a count of 1
44 mapped_data = mapper(text)
45
46 # Reduce step: combine the counts for each word
47 word_counts = reduce(reducer, mapped_data, {})
```

Features of wordcount.py

- Implements word count algorithm using map() and reduce()
- Read text from file with explicit encoding
- Write the word counts to a CSV file with proper encoding and error handling
- Construct the output file name with the iteration number

References

- <https://www.geeksforgeeks.org/python-remove-punctuation-from-string/>
- https://www.learnpython.org/en/Map%2C_Filter%2C_Reduce
- <https://stackoverflow.com/questions/436220/how-to-determine-the-encoding-of-text>
- <https://www.knowledgehut.com/blog/programming/sys-argv-python-examples>
- https://www3.ntu.edu.sg/home/ehchua/programming/webprogramming/Python_FileText.html
- <https://www.geeksforgeeks.org/how-to-iterate-over-files-in-directory-using-python/>

Python

File size: 100 KB

Time: 0.039 s

PROBLEMS 4 OUTPUT DEBUG CONSOLE TERMINAL JUPYTER

```
● hadoop@vmbox:~/Downloads/wordcount$ time python3 wordcount.py personae.txt
Word counts saved to: output-1.csv
real    0m0,039s
user    0m0,039s
sys     0m0,000s
○ hadoop@vmbox:~/Downloads/wordcount$
```

File size: 100 KB
data_1 = pd.read_csv('output-1.csv')
data_1

✓ 0.0s

	Word	Count
0	as	183
1	you	543
2	like	83
3	it	226
4	dramatis	1
...
3275	complexions	1
3276	liked	1
3277	breaths	2
3278	defied	1
3279	curtsy	1
3280 rows × 2 columns		

Python

File size: 1 MB

Time: 0.416 s

```
PROBLEMS 1 OUTPUT DEBUG CONSOLE TERMINAL JUPYTER
● hadoop@vmbox:~/Downloads/wordcount$ time python3 wordcount.py melville.txt
Word counts saved to: output-2.csv
real    0m0,416s
user    0m0,369s
sys     0m0,044s
○ hadoop@vmbox:~/Downloads/wordcount$
```

#	Word	Count
0	the	21630
1	project	102
2	gutenberg	38
3	ebook	28
4	of	10181
...
27031	obloquy	1
27032	reconcile	1
27033	lecture	1
27034	studydoor	1
27035	thisbe	1
27036	rows × 2 columns	

Python

File size: 10 MB

Time: 2.283 s

PROBLEMS 1 OUTPUT DEBUG CONSOLE TERMINAL JUPYTER

```
● hadoop@vmbox:~/Downloads/wordcount$ time python3 wordcount.py shakespeare.txt
Word counts saved to: output-3.csv

real    0m2,283s
user    0m2,187s
sys     0m0,084s
○ hadoop@vmbox:~/Downloads/wordcount$
```

```
# File size: 10 MB
data_3 = pd.read_csv('output-3.csv')
data_3
```

✓ 0.0s

	Word	Count
0	the	98677
1	project	125
2	gutenberg	41
3	ebook	13
4	of	59991
...
61601	alleluia	4
61602	chalcedony	1
61603	sardonyx	1
61604	chrysolyte	1
61605	chrysoprasus	1
61606	rows × 2 columns	

Python

File size: 100 MB
Time: 17.899 s

```
PROBLEMS 1 OUTPUT DEBUG CONSOLE TERMINAL JUPYTER

● hadoop@vmbox:~/Downloads/wordcount$ time python3 wordcount.py enwik8.txt
Word counts saved to: output-4.csv
real    0m17,899s
user    0m16,961s
sys     0m0,938s
o hadoop@vmbox:~/Downloads/wordcount$
```

File size: 100 MB

```
data_4 = pd.read_csv('output-4.csv')
data_4
```

✓ 0.5s

	Word	Count
0	mediawiki	25
1	xmlnshttpwwwmediawikiorgxmlexport03	1
2	xmlnsxihttpwww3org2001xmlschemainstance	1
3	xsischemalocationhttpwwwmediawikiorgxmlexport03	1
4	httpwwwmediawikiorgxmlexport03xsd	1
...
760425	nihongokōgo口語	1
760426	kōgo	1
760427	regiontōhoku	1
760428	kyūshū	2
760429	kansaiben	1
760430	rows × 2 columns	

Python

File size: 1 GB

Time: ± 318.04 s

Kernel killed after 2 minutes

```
PROBLEMS 1 OUTPUT DEBUG CONSOLE TERMINAL JUPYTER

④ hadoop@vmbox:~/Downloads/wordcount$ time python3 wordcount.py enwik9.txt
Killed

real    2m8,308s
user    1m42,791s
sys     0m16,089s
④ hadoop@vmbox:~/Downloads/wordcount$
```

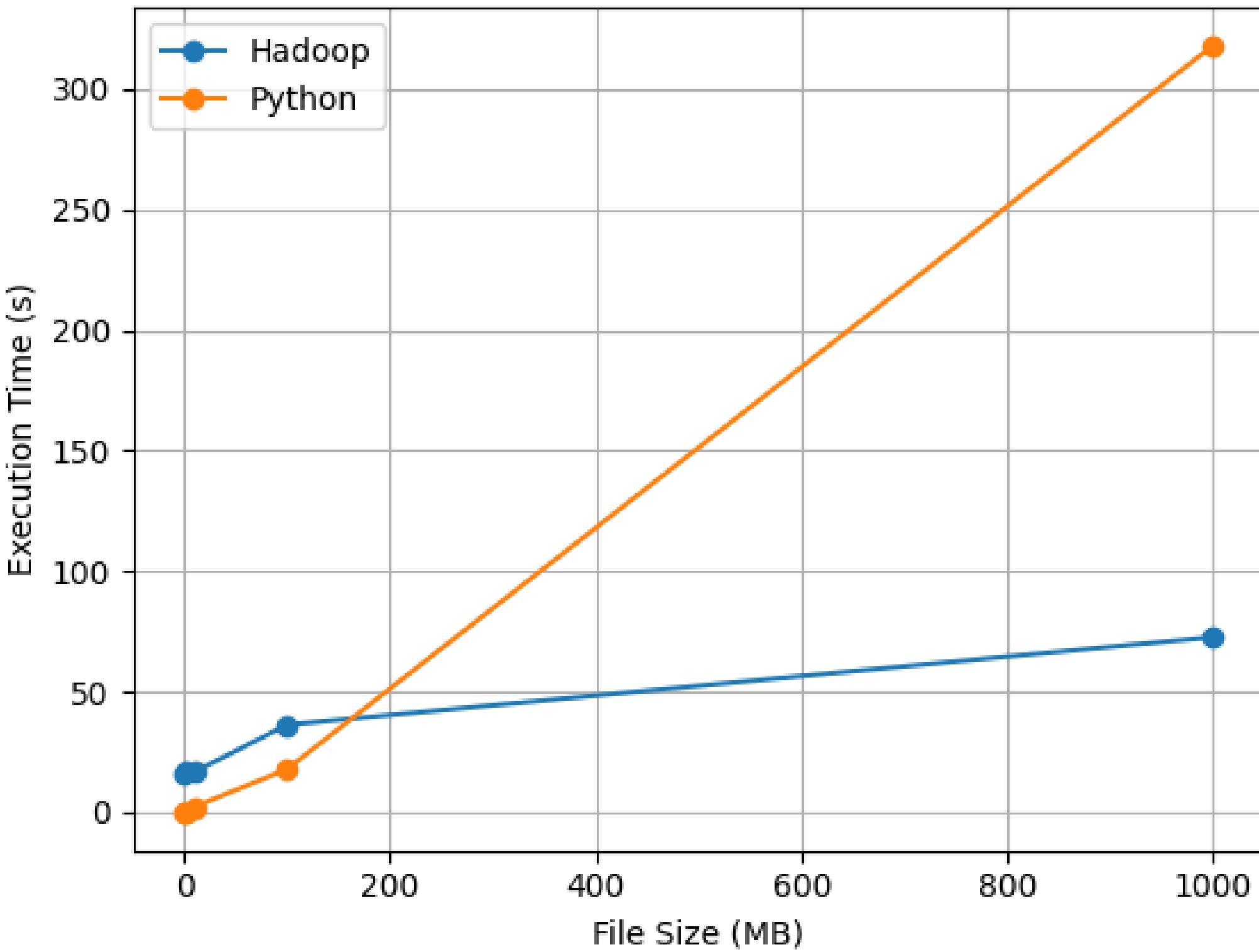
#	Word	Count
0	mediawiki	308
1	xmlnshttpwwwmediawikiorgxmlexport03	1
2	xmlsnsxihttpwww3org2001xmlschemainstance	1
3	xsischemalocationhttpwwwmediawikiorgxmlexport03	1
4	httpwwwmediawikiorgxmlexport03xsd	1
...
5133222	imagehyperbaricoxygentherapy1personchamberjgpt...	1
5133223	quotmonoplacequot	1
5133224	softsided	1
5133225	httpuhmsorg	1
5133226	uhms	1
5133227	rows × 2 columns	

Plot

	100 KB	1 MB	10 MB	100 MB	1 GB
Hadoop	15.82 s	16.53 s	16.74 s	36.35 s	72.66 s
Python	0.039 s	0.416 s	2.283 s	17.899 s	318.04 s

Waktu eksekusi program Word Count lebih cepat dilakukan oleh Python daripada Hadoop untuk file sebesar 100 KB, 1 MB, 10 MB, dan 100 MB. Namun, waktu eksekusi program Python lebih lambat daripada Hadoop untuk file sebesar 1 GB.

Comparison of Hadoop and Python Execution Times



Analisis terhadap WordCount

Jika membandingkan waktu eksekusi wordcount dengan python dan menggunakan Hadoop, maka wordcount python lebih cepat karena data yang dihitung disimpan dan dilakukan pada satu mesin atau server tunggal. Sehingga, hanya terjadi sedikit overhead terkait pengaturan dan komunikasi antar node dalam cluster seperti pada Hadoop. Dan untuk ukuran file yang relatif kecil (100 kb, 1 mb, 10 mb) maka wordcount dengan python dapat menyelesaikan tugas lebih cepat karena tidak melibatkan kompleksitas yang signifikan.

Sedangkan, hadoop dirancang untuk memproses dan menganalisis data dalam skala besar dengan membagi tugas pemrosesan ke beberapa node dalam cluster sehingga dapat terjadi overhead yang signifikan walaupun penggunaan hadoop sangat efektif, khususnya pada platform terdistribusi dan data dalam ukuran besar.

Kinerja hadoop menjadi lebih lambat dibandingkan wordcount normal pada skala kecil karena membutuhkan waktu untuk menghubungkan node-node dalam cluster dan komunikasi antar node.

Kesimpulan:

Untuk data yang besar dan membutuhkan pemrosesan dalam skala yang luas, Hadoop menjadi pilihan lebih baik karena lebih optimal. Sedangkan, pada kasus diatas, wordcount normal tidak melibatkan pengaturan dan komunikasi antar node yang kompleks sehingga kinerja lebih cepat dibandingkan menggunakan Hadoop.



THANK YOU