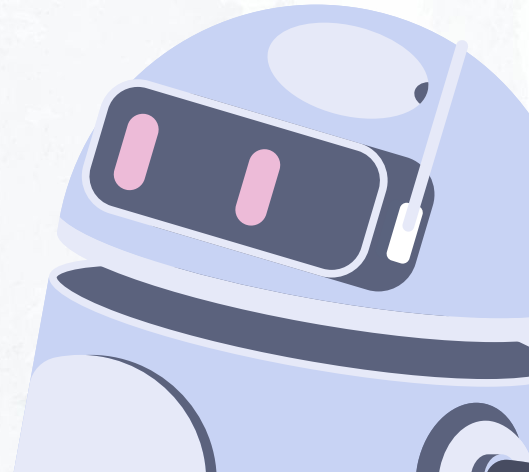


Compare the wordcount processing time using Hadoop and without using Hadoop.






Kelompok 2 :






- Muhammad Cavan Naufal Azizi (2106702730)
- RBS Kresna Ramdani Galih (2106702610)
- Jeremy Ganda Pandapotan (2106731573)
- Fahrezy H (2106731466)



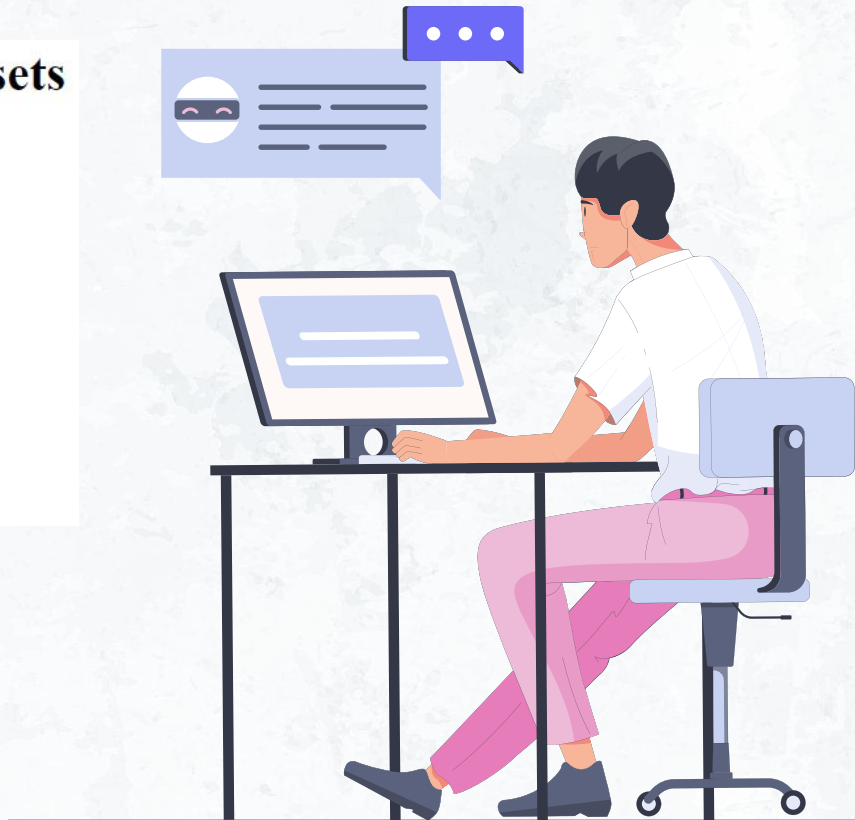
Source file .txt

Index of /~jplozi/hadooplab_lsds_2015/datasets

Name	Last modified	Size	Description
 Parent Directory	-		
 gutenberg-1G.txt.gz	2016-02-06 13:39	384M	
 gutenberg-100M.txt.gz	2016-02-06 13:38	37M	
 gutenberg-200M.txt.gz	2016-02-06 13:38	75M	
 gutenberg-500M.txt.gz	2016-02-06 13:39	187M	
 ncdc-2013-sorted.csv.gz	2016-02-06 13:40	118M	

 1G.txt	11/14/2014 5:05 AM	Text Document	1,048,576 ...
 1M.txt	6/9/2023 9:31 AM	Text Document	1,099 KB
 100M.txt	2/6/2016 7:38 PM	Text Document	102,400 KB
 200M.txt	11/14/2014 5:05 AM	Text Document	204,800 KB
 500M.txt	11/14/2014 5:05 AM	Text Document	512,000 KB

https://www.i3s.unice.fr/~jplozi/hadooplab_lsds_2015/datasets/



Source code (Python)

```
wordcount.py
1  import time
2
3  # Creating a dictionary to store the count of each word
4  word_counts = {}
5
6  # Opening our text file in read-only mode using the open() function
7  with open(r'16.txt', 'r', encoding='latin-1') as file:
8      # Reading the content of the file using the read() function and storing them in a new variable
9      data = file.read()
10
11     # Splitting the data into separate words using the split() function
12     words = data.split()
13
14     # Counting the occurrences of each word
15     for word in words:
16         if word in word_counts:
17             word_counts[word] += 1
18         else:
19             word_counts[word] = 1
20
21     # Printing the count of each word
22     for word, count in word_counts.items():
23         print(f"{word}: {count}")
```

<https://www.geeksforgeeks.org/python-program-to-count-words-in-text-file/>

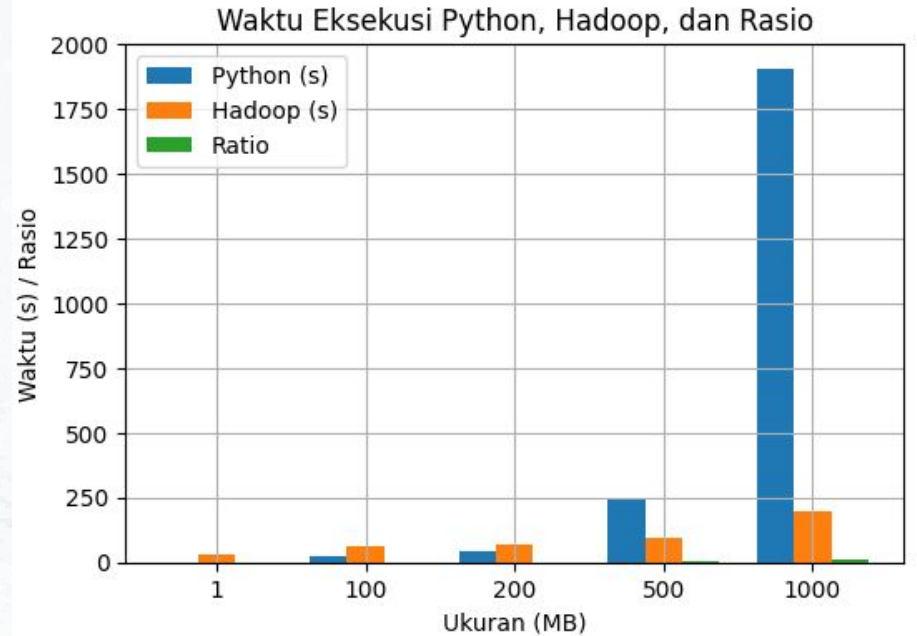
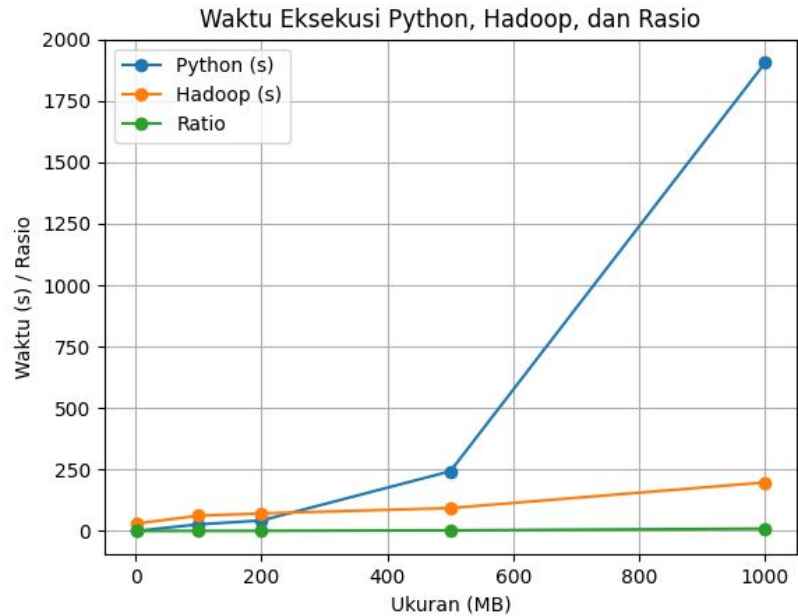


Time Compare



Size / Type	Python (s)	Hadoop (s)	Ratio
1 MB	0.137	30.242	0.00453
100 MB	27.021	62.245	0.43411
200 MB	42.617	71.460	0.59637
500 MB	243.039	93.246	2.60643
1000 MB	1905.303	197.669	9.63885

Graph



Python

1 MB

```
real    0m0.137s
user    0m0.000s
sys     0m0.000s
```

100 MB

```
real    0m27.021s
user    0m0.000s
sys     0m0.015s
```

200 MB

```
real    0m42.617s
user    0m0.000s
sys     0m0.000s
```

500 MB

```
real    4m3.039s
user    0m0.000s
sys     0m0.015s
```

1000 MB

```
real    31m45.303s
user    0m0.016s
sys     0m0.125s
```

Hadoop

1 MB

```
2023-06-09 18:31:52,049 INFO client.RMPProxy: Connecting to ResourceManager at /0.0.0.0:8032
2023-06-09 18:31:52,726 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement
the Tool interface and execute your application with ToolRunner to remedy this.
2023-06-09 18:31:52,737 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/
Asus/.staging/job_1686307953675_0005
2023-06-09 18:31:52,928 INFO input.FileInputFormat: Total input files to process : 1
2023-06-09 18:31:53,005 INFO mapreduce.JobSubmitter: number of splits:1
2023-06-09 18:31:53,562 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1686307953675_0005
2023-06-09 18:31:53,564 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-06-09 18:31:53,728 INFO conf.Configuration: resource-types.xml not found
2023-06-09 18:31:53,729 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-06-09 18:31:53,797 INFO impl.YarnClientImpl: Submitted application application_1686307953675_0005
2023-06-09 18:31:53,836 INFO mapreduce.Job: The url to track the job: http://LAPTOP-7QH5FHQ:8088/proxy/application_1686
307953675_0005/
2023-06-09 18:31:53,837 INFO mapreduce.Job: Running job: job_1686307953675_0005
2023-06-09 18:32:06,002 INFO mapreduce.Job: Job job_1686307953675_0005 running in uber mode : false
2023-06-09 18:32:06,003 INFO mapreduce.Job:  map 0% reduce 0%
2023-06-09 18:32:12,090 INFO mapreduce.Job:  map 100% reduce 0%
2023-06-09 18:32:17,147 INFO mapreduce.Job:  map 100% reduce 100%
2023-06-09 18:32:22,207 INFO mapreduce.Job: Job job_1686307953675_0005 completed successfully
2023-06-09 18:32:22,291 INFO mapreduce.Job: Counters: 54
```


Hadoop

100 MB

```
2023-06-09 17:53:54,285 INFO client.RMPProxy: Connecting to ResourceManager at /0.0.0.0:8032
2023-06-09 17:53:54,887 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2023-06-09 17:53:54,900 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-p-yarn/staging/Asus/.staging/job_1686307953675_0001
2023-06-09 17:53:55,089 INFO input.FileInputFormat: Total input files to process : 1
2023-06-09 17:53:55,147 INFO mapreduce.JobSubmitter: number of splits:1
2023-06-09 17:53:55,260 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1686307953675_0001
2023-06-09 17:53:55,262 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-06-09 17:53:55,425 INFO conf.Configuration: resource-types.xml not found
2023-06-09 17:53:55,426 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-06-09 17:53:55,665 INFO impl.YarnClientImpl: Submitted application application_1686307953675_0001
2023-06-09 17:53:55,712 INFO mapreduce.Job: The url to track the job: http://LAPTOP-7QH5HQ:8088/proxy/application_1686307953675_0001/
2023-06-09 17:53:55,712 INFO mapreduce.Job: Running job: job_1686307953675_0001
2023-06-09 17:54:08,946 INFO mapreduce.Job: Job job_1686307953675_0001 running in uber mode : false
2023-06-09 17:54:08,947 INFO mapreduce.Job:  map 0% reduce 0%
2023-06-09 17:54:32,236 INFO mapreduce.Job:  map 39% reduce 0%
2023-06-09 17:54:38,290 INFO mapreduce.Job:  map 51% reduce 0%
2023-06-09 17:54:44,343 INFO mapreduce.Job:  map 100% reduce 0%
2023-06-09 17:54:51,387 INFO mapreduce.Job:  map 100% reduce 100%
2023-06-09 17:54:56,462 INFO mapreduce.Job: Job job_1686307953675_0001 completed successfully
2023-06-09 17:54:56,530 INFO mapreduce.Job: Counters: 54
```


Hadoop

200 MB

```
2023-06-09 18:06:16,878 INFO client.RMPProxy: Connecting to ResourceManager at /0.0.0.0:8032
2023-06-09 18:06:19,550 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2023-06-09 18:06:19,565 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/Asus/.staging/job_1686307953675_0002
2023-06-09 18:06:19,754 INFO input.FileInputFormat: Total input files to process : 1
2023-06-09 18:06:19,821 INFO mapreduce.JobSubmitter: number of splits:2
2023-06-09 18:06:19,992 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1686307953675_0002
2023-06-09 18:06:19,994 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-06-09 18:06:20,202 INFO conf.Configuration: resource-types.xml not found
2023-06-09 18:06:20,202 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-06-09 18:06:20,280 INFO impl.YarnClientImpl: Submitted application application_1686307953675_0002
2023-06-09 18:06:20,343 INFO mapreduce.Job: The url to track the job: http://LAPTOP-7QHLF5HQ:8088/proxy/application_1686307953675_0002/
2023-06-09 18:06:20,344 INFO mapreduce.Job: Running job: job_1686307953675_0002
2023-06-09 18:06:32,489 INFO mapreduce.Job: Job job_1686307953675_0002 running in uber mode : false
2023-06-09 18:06:32,490 INFO mapreduce.Job:  map 0% reduce 0%
2023-06-09 18:06:50,767 INFO mapreduce.Job:  map 39% reduce 0%
2023-06-09 18:06:56,834 INFO mapreduce.Job:  map 53% reduce 0%
2023-06-09 18:06:59,894 INFO mapreduce.Job:  map 70% reduce 0%
2023-06-09 18:07:02,944 INFO mapreduce.Job:  map 75% reduce 0%
2023-06-09 18:07:09,047 INFO mapreduce.Job:  map 80% reduce 0%
2023-06-09 18:07:15,119 INFO mapreduce.Job:  map 83% reduce 0%
2023-06-09 18:07:19,176 INFO mapreduce.Job:  map 100% reduce 0%
2023-06-09 18:07:20,194 INFO mapreduce.Job:  map 100% reduce 34%
2023-06-09 18:07:21,212 INFO mapreduce.Job:  map 100% reduce 100%
2023-06-09 18:07:28,264 INFO mapreduce.Job: Job job_1686307953675_0002 completed successfully
2023-06-09 18:07:28,338 INFO mapreduce.Job: Counters: 55
```

Hadoop

500 MB

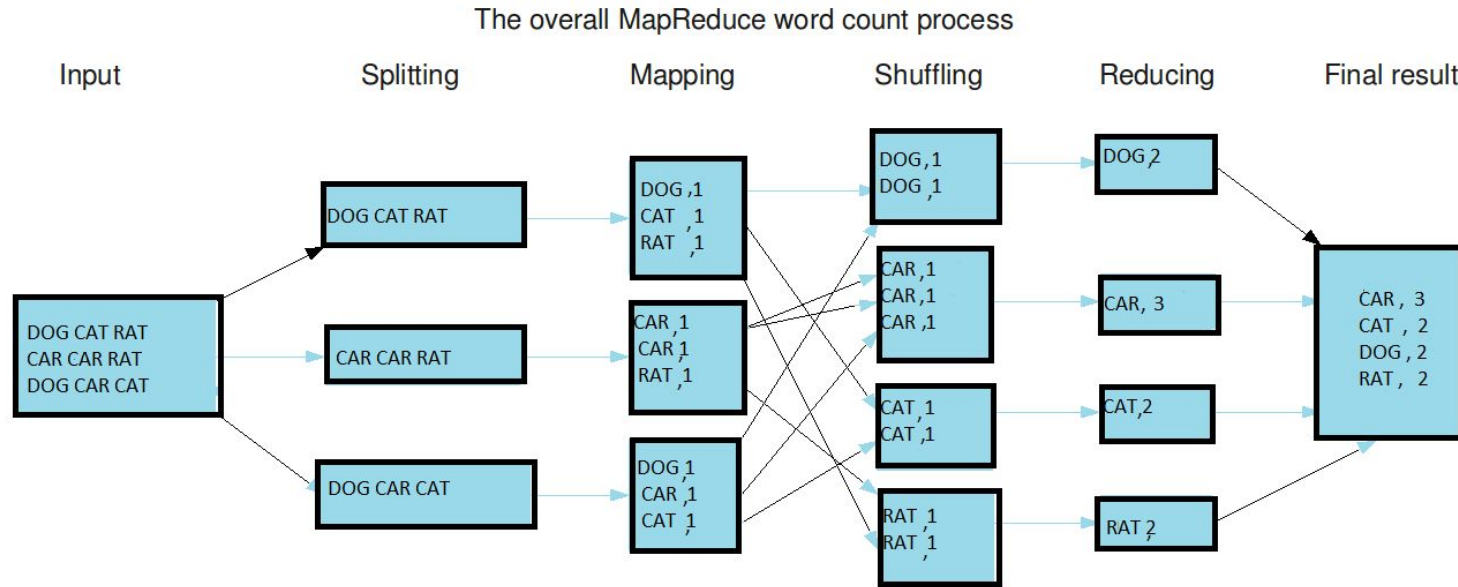
```
2023-06-09 18:13:40,435 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2023-06-09 18:13:41,159 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2023-06-09 18:13:41,171 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/Asus/.staging/job_1686307953675_0003
2023-06-09 18:13:41,350 INFO input.FileInputFormat: Total input files to process : 1
2023-06-09 18:13:41,829 INFO mapreduce.JobSubmitter: number of splits:4
2023-06-09 18:13:41,943 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1686307953675_0003
2023-06-09 18:13:41,946 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-06-09 18:13:42,117 INFO conf.Configuration: resource-types.xml not found
2023-06-09 18:13:42,118 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-06-09 18:13:42,185 INFO impl.YarnClientImpl: Submitted application application_1686307953675_0003
2023-06-09 18:13:42,224 INFO mapreduce.Job: The url to track the job: http://LAPTOP-7QHLF5HQ:8088/proxy/application_1686307953675_0003/
2023-06-09 18:13:42,224 INFO mapreduce.Job: Running job: job_1686307953675_0003
2023-06-09 18:13:55,450 INFO mapreduce.Job: Job job_1686307953675_0003 running in uber mode : false
2023-06-09 18:13:55,451 INFO mapreduce.Job: map 0% reduce 0%
2023-06-09 18:14:15,818 INFO mapreduce.Job: map 21% reduce 0%
2023-06-09 18:14:21,925 INFO mapreduce.Job: map 25% reduce 0%
2023-06-09 18:14:22,940 INFO mapreduce.Job: map 29% reduce 0%
2023-06-09 18:14:28,021 INFO mapreduce.Job: map 32% reduce 0%
2023-06-09 18:14:29,040 INFO mapreduce.Job: map 36% reduce 0%
2023-06-09 18:14:34,107 INFO mapreduce.Job: map 41% reduce 0%
2023-06-09 18:14:35,123 INFO mapreduce.Job: map 42% reduce 0%
2023-06-09 18:14:40,199 INFO mapreduce.Job: map 49% reduce 0%
2023-06-09 18:14:41,209 INFO mapreduce.Job: map 51% reduce 0%
2023-06-09 18:14:46,273 INFO mapreduce.Job: map 58% reduce 0%
2023-06-09 18:14:47,285 INFO mapreduce.Job: map 61% reduce 0%
2023-06-09 18:14:52,388 INFO mapreduce.Job: map 64% reduce 0%
2023-06-09 18:14:53,401 INFO mapreduce.Job: map 66% reduce 0%
2023-06-09 18:14:54,428 INFO mapreduce.Job: map 74% reduce 0%
2023-06-09 18:14:57,463 INFO mapreduce.Job: map 82% reduce 0%
2023-06-09 18:14:58,478 INFO mapreduce.Job: map 92% reduce 0%
2023-06-09 18:15:00,507 INFO mapreduce.Job: map 100% reduce 0%
2023-06-09 18:15:07,597 INFO mapreduce.Job: map 100% reduce 100%
2023-06-09 18:15:13,632 INFO mapreduce.Job: Job job_1686307953675_0003 completed successfully
2023-06-09 18:15:13,699 INFO mapreduce.Job: Counters: 54
```


Hadoop

1000 MB

```
2023-06-09 18:18:03,793 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2023-06-09 18:18:21,777 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2023-06-09 18:18:21,789 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/Asus/.staging/job_1686307953675_0004
2023-06-09 18:18:22,008 INFO input.FileInputFormat: Total input files to process : 1
2023-06-09 18:18:22,081 INFO mapreduce.JobSubmitter: number of splits:8
2023-06-09 18:18:22,199 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1686307953675_0004
2023-06-09 18:18:22,202 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-06-09 18:18:22,372 INFO conf.Configuration: resource-types.xml not found
2023-06-09 18:18:22,372 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-06-09 18:18:22,438 INFO impl.YarnClientImpl: Submitted application application_1686307953675_0004
2023-06-09 18:18:22,482 INFO mapreduce.Job: The url to track the job: http://LAPTOP-7QHLSHQ:8088/proxy/application_1686307953675_0004/
2023-06-09 18:18:22,483 INFO mapreduce.Job: Running job: job_1686307953675_0004
2023-06-09 18:18:34,694 INFO mapreduce.Job: Job job_1686307953675_0004 running in uber mode : false
2023-06-09 18:18:34,695 INFO mapreduce.Job: map 0% reduce 0%
2023-06-09 18:18:57,682 INFO mapreduce.Job: map 3% reduce 0%
2023-06-09 18:18:58,691 INFO mapreduce.Job: map 9% reduce 0%
2023-06-09 18:19:03,794 INFO mapreduce.Job: map 11% reduce 0%
2023-06-09 18:19:04,811 INFO mapreduce.Job: map 16% reduce 0%
2023-06-09 18:19:10,921 INFO mapreduce.Job: map 17% reduce 0%
2023-06-09 18:19:12,016 INFO mapreduce.Job: map 19% reduce 0%
2023-06-09 18:19:16,137 INFO mapreduce.Job: map 20% reduce 0%
2023-06-09 18:19:17,158 INFO mapreduce.Job: map 22% reduce 0%
2023-06-09 18:19:18,185 INFO mapreduce.Job: map 23% reduce 0%
2023-06-09 18:19:23,340 INFO mapreduce.Job: map 24% reduce 0%
2023-06-09 18:19:25,455 INFO mapreduce.Job: map 28% reduce 0%
2023-06-09 18:19:29,542 INFO mapreduce.Job: map 30% reduce 0%
2023-06-09 18:19:35,615 INFO mapreduce.Job: map 32% reduce 0%
2023-06-09 18:19:37,676 INFO mapreduce.Job: map 35% reduce 0%
2023-06-09 18:19:41,734 INFO mapreduce.Job: map 37% reduce 0%
2023-06-09 18:19:49,920 INFO mapreduce.Job: map 43% reduce 0%
2023-06-09 18:19:50,936 INFO mapreduce.Job: map 45% reduce 0%
2023-06-09 18:19:56,056 INFO mapreduce.Job: map 46% reduce 0%
2023-06-09 18:20:02,176 INFO mapreduce.Job: map 50% reduce 0%
2023-06-09 18:20:07,396 INFO mapreduce.Job: map 58% reduce 0%
2023-06-09 18:20:08,408 INFO mapreduce.Job: map 73% reduce 0%
2023-06-09 18:20:09,424 INFO mapreduce.Job: map 75% reduce 0%
2023-06-09 18:20:26,761 INFO mapreduce.Job: map 80% reduce 0%
2023-06-09 18:20:27,777 INFO mapreduce.Job: map 80% reduce 25%
2023-06-09 18:20:32,840 INFO mapreduce.Job: map 82% reduce 25%
2023-06-09 18:20:38,918 INFO mapreduce.Job: map 83% reduce 25%
2023-06-09 18:20:45,017 INFO mapreduce.Job: map 85% reduce 25%
2023-06-09 18:20:51,076 INFO mapreduce.Job: map 87% reduce 25%
2023-06-09 18:20:57,135 INFO mapreduce.Job: map 90% reduce 25%
2023-06-09 18:21:03,195 INFO mapreduce.Job: map 92% reduce 25%
2023-06-09 18:21:08,241 INFO mapreduce.Job: map 100% reduce 25%
2023-06-09 18:21:10,256 INFO mapreduce.Job: map 100% reduce 45%
2023-06-09 18:21:15,304 INFO mapreduce.Job: map 100% reduce 100%
2023-06-09 18:21:21,360 INFO mapreduce.Job: Job job_1686307953675_0004 completed successfully
2023-06-09 18:21:21,426 INFO mapreduce.Job: Counters: 55
```

MapReduce process (wordcount)



Source : <https://www.pluralsight.com/guides/getting-started-with-hadoop-mapreduce>

Analisis

Size / Type	Python (s)	Hadoop (s)	Ratio
1 MB	0.137	30.242	0.00453
100 MB	27.021	62.245	0.43411
200 MB	42.617	71.460	0.59637
500 MB	243.039	93.246	2.60643
1000 MB	1905.303	197.669	9.63885



Efektifitas meningkat seiring ukuran file bertambah

Dapat dilihat dari grafik dan tabel diatas hadoop memakan waktu yang lebih lama dibandingkan dengan python untuk ukuran file yang kecil. Hadoop baru menunjukan efektivitasnya pada file yang berukuran 500mb dengan rasio lebih dari 2.6 kali lipat lebih cepat dari python.

Mengapa Hal ini dapat terjadi?

Masalah dalam menangani file kecil dalam Hadoop terutama terjadi dalam dua aspek utama: Hadoop Distributed File System (HDFS) dan tugas MapReduce.

Analisis

Masalah dengan HDFS:

HDFS tidak dirancang untuk menangani banyak file kecil. Setiap file, direktori, dan blok dalam HDFS direpresentasikan sebagai objek dalam memori namenode. Setiap objek tersebut memakan sekitar 150 byte. Jika terdapat banyak file kecil, jumlah objek yang harus direpresentasikan di memori namenode menjadi sangat besar, menyebabkan peningkatan penggunaan memori yang signifikan. Hal ini menjadi sulit diatasi dengan perangkat keras saat ini. Jumlah file yang sangat besar, seperti miliaran file, tidak dapat ditangani dengan efisien.

Masalah pada tugas MapReduce:


Tugas MapReduce biasanya memproses blok masukan dalam satu waktu (menggunakan `FileInputFormatDefault`). Jika file sangat kecil dan jumlahnya banyak, setiap tugas pemetaan (map task) akan memproses sedikit masukan, dan jumlah tugas pemetaan menjadi sangat banyak. Hal ini mengakibatkan tambahan overhead manajemen yang signifikan. Sebagai perbandingan, jika terdapat satu file berukuran 1GB yang terbagi menjadi 16 blok berukuran 64MB, dan terdapat 10.000 file berukuran 100KB, maka 10.000 file tersebut akan menggunakan satu tugas pemetaan masing-masing, dan waktu pemrosesan akan menjadi puluhan atau ratusan kali lebih lambat dibandingkan dengan file masukan tunggal yang lebih besar.

Tutorial Install Hadoop

- Masuk ke website oracle berikut :

<https://www.oracle.com/java/technologies/javase/javase-jdk8-downloads.html>

- Download Java Development Kit 8 (JDK 8) windows x64 kemudian install.

Linux	macOS	Solaris	Windows
Product/file description		File size	Download
x86 Installer		136.77 MB	 jdk-8u371-windows-i586.exe
x64 Installer		145.50 MB	 jdk-8u371-windows-x64.exe

- Set Environment Variables (Java)

Tambahkan pada system variable dengan nama JAVA_HOME, kemudian dengan value yaitu path dari folder java yang telah terinstall.

System variables

Variable	Value
ComSpec	C:\WINDOWS\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
HADOOP_HOME	C:\hadoop-3.2.2
JAVA_HOME	C:\Program Files\Java\jdk-1.8
NUMBER_OF_PROCESSORS	8
OS	Windows_NT
Path	C:\Program Files (x86)\Common Files\Oracle\Java\javapath;C:\Prog...

- Kemudian tambahkan /bin pada variable Path.

```
C:\Program Files\nodejs\  
C:\Program Files\Java\jdk-1.8\bin  
C:\hadoop-3.2.2\bin
```

- Verifikasi Java

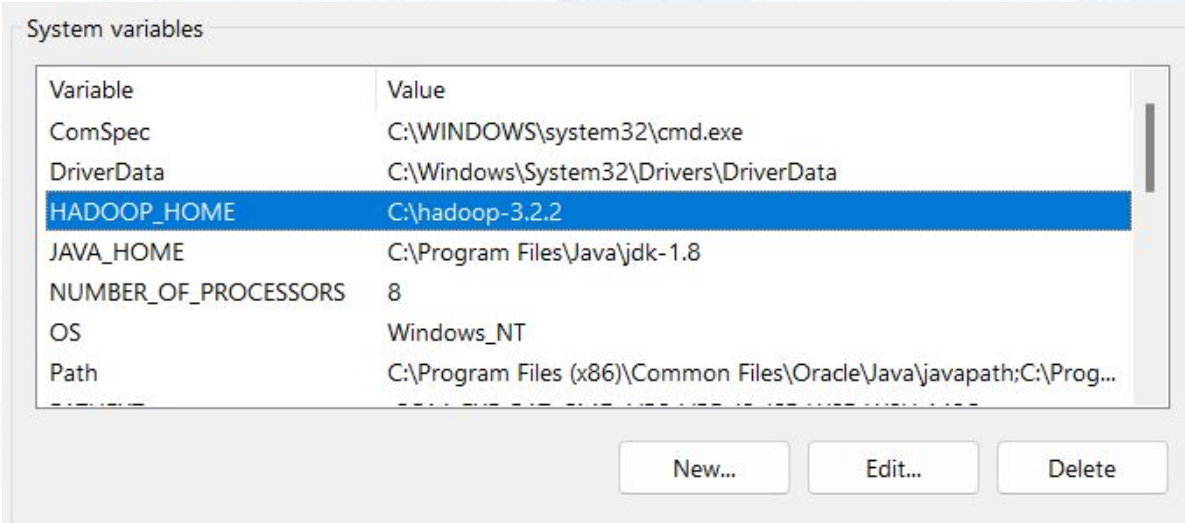
```
C:\Users\Asus>java -version  
java version "1.8.0_371"  
Java(TM) SE Runtime Environment (build 1.8.0_371-b11)  
Java HotSpot(TM) 64-Bit Server VM (build 25.371-b11, mixed mode)
```


- Download Hadoop (Recommended ver. 3.2.2)

<https://archive.apache.org/dist/hadoop/common/>

	hadoop-3.2.2-src.tar.gz.sha512	2021-01-13 18:48	154
	hadoop-3.2.2.tar.gz	2021-01-13 18:48	377M
	hadoop-3.2.2.tar.gz.asc	2021-01-13 18:48	833

- Pilih hadoop-3.2.2.tar.gz. kemudian extract file hadoop yang telah di download
- Set Environment Variables (Hadoop)



- Kemudian tambahkan /bin dan /sbin pada variable Path.

C:\Program Files\Java\jdk-1.8\bin


C:\hadoop-3.2.2\bin

C:\hadoop-3.2.2\sbin

- Tambahkan folder data pada file hadoop yang telah di extract.

 bin	5/25/2023 12:32 AM	File folder
 data	5/25/2023 12:23 AM	File folder
 etc	5/25/2023 12:13 AM	File folder

- kemudian di dalam folder data tersebut buat dua folder dengan nama namenode dan datanode.

 datanode	5/25/2023 12:35 AM	File folder
 namenode	5/25/2023 12:35 AM	File folder

- Edit core-site.xml, hdfs-site.xml, mapred-site.xml, dan yarn-site.xml.

etc/hadoop/core-site.xml :

```
<configuration>
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://localhost:9000</value>
  </property>
</configuration>
```

- etc/hadoop/hdfs-site.xml :

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
</configuration>
```

- etc/hadoop/mapred-site.xml :

```
<configuration>
  <property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
  </property>
  <property>
    <name>mapreduce.application.classpath</name>

    <value>$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/*:$HADOOP_M
APRED_HOME/share/hadoop/mapreduce/lib/*</value>
  </property>
</configuration>
```

- etc/hadoop/yarn-site.xml :

```
<configuration>
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>
  <property>
    <name>yarn.nodemanager.env-whitelist</name>



    <value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,
HADOOP_CONF_DIR,CLASSPATH_PREPEND_DISTCACHE,HADOOP_YAR
N_HOME,HADOOP_HOME,PATH,LANG,TZ,HADOOP_MAPRED_HOME</va
lue>
  </property>
</configuration>
```


- Edit Hadoop Env pada hadoop env.cmd dengan path folder jdk yang telah didownload




```
@rem The java implementation to use. Required.  
set JAVA_HOME=C:"\Program Files\Java\jdk-1.8"
```

- Download Patch file hadoop untuk Windows

<https://github.com/cdarlint/winutils>

	hadoop-3.2.1/bin	add 321 winutils	4 years ago
	hadoop-3.2.2/bin	compile hadoop-3.2.2	2 years ago

- Extract file bin tersebut kemudian pindahkan pada folder bin yang ada pada folder hadoop.

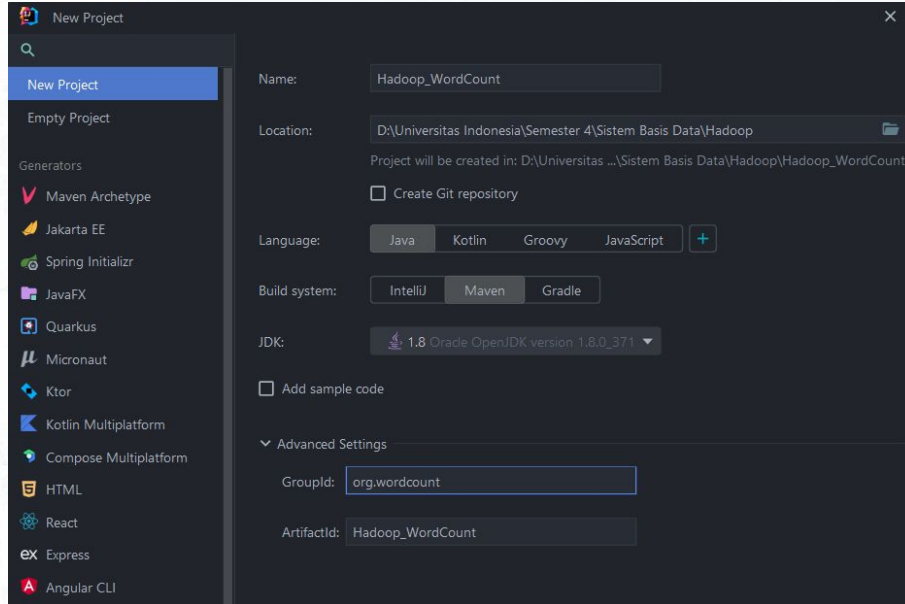
	bin	5/25/2023 12:32 AM	File folder
	data	5/25/2023 12:23 AM	File folder
	etc	5/25/2023 12:13 AM	File folder

- Verifikasi Hadoop

```
C:\Users\Asus>hadoop version
Hadoop 3.2.2
Source code repository Unknown -r 7a3bc90b05f257c8ace2f76d74264906f0f7a932
Compiled by hexiaoqiao on 2021-01-03T09:26Z
Compiled with protoc 2.5.0
From source with checksum 5a8f564f46624254b27f6a33126ff4
This command was run using /C:/hadoop-3.2.2/share/hadoop/common/hadoop-common-3.2.2.jar
```

Hadoop WordCount (Java)

- Create Maven Project (Using IntelliJ) dengan Language Java dan Build System Maven.



- Tambahkan dependencies pada pom.xml

```
<dependencies>
  <dependency>
    <groupId>org.apache.hadoop</groupId>
    <artifactId>hadoop-common</artifactId>
    <version>3.2.2</version>
  </dependency>
  <dependency>
    <groupId>org.apache.hadoop</groupId>
    <artifactId>hadoop-mapreduce-client-core</artifactId>
    <version>3.2.2</version>
  </dependency>
</dependencies>
```

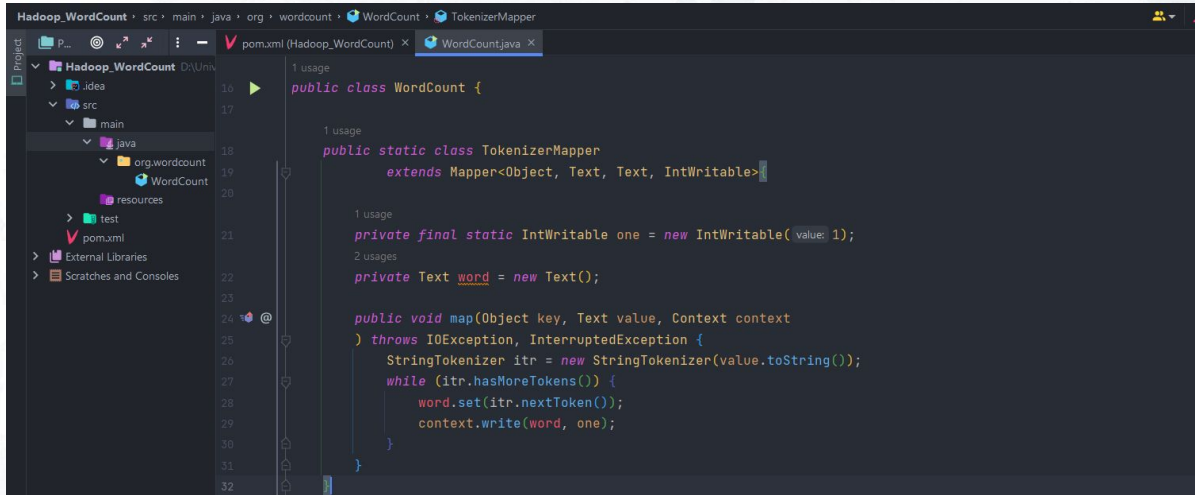

- Create Package dan didalam package tersebut buat file java berisi kode wordcount.



- kodenya dapat dilihat pada link :

[https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html#](https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html#Source_Code)

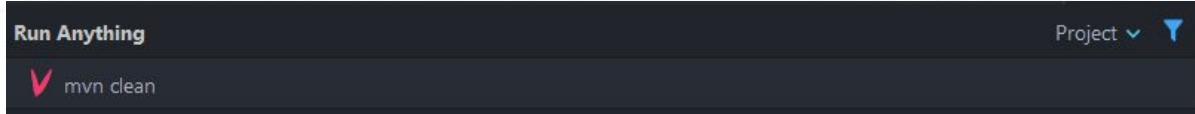
Source Code

A screenshot of an IDE showing the source code of 'WordCount.java'. The left sidebar shows the project structure: 'Hadoop_WordCount' > 'src' > 'main' > 'java' > 'org.wordcount' > 'WordCount'. The main editor displays the following code:

```
16 public class WordCount {
17
18     1 usage
19     public static class TokenizerMapper
20         extends Mapper<Object, Text, Text, IntWritable> {
21
22         1 usage
23         private final static IntWritable one = new IntWritable( value: 1);
24         2 usages
25         private Text word = new Text();
26
27         public void map(Object key, Text value, Context context
28             ) throws IOException, InterruptedException {
29             StringTokenizer itr = new StringTokenizer(value.toString());
30             while (itr.hasMoreTokens()) {
31                 word.set(itr.nextToken());
32                 context.write(word, one);
33             }
34         }
35     }
36 }
```

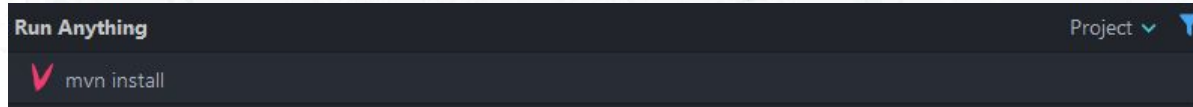
- Kemudian membuat jar file dengan maven.

mvn clean



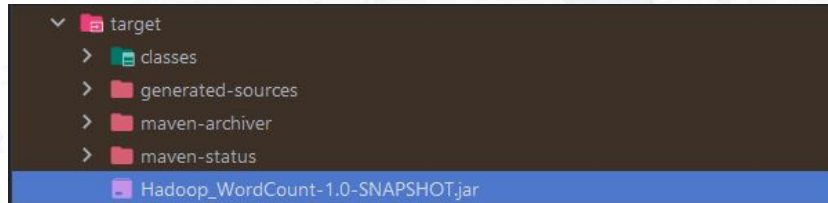
```
[INFO] -----< org.wordcount:Hadoop_WordCount >-----  
[INFO] Building Hadoop_WordCount 1.0-SNAPSHOT  
[INFO] -----[ jar ]-----  
[INFO]  
[INFO] --- maven-clean-plugin:2.5:clean (default-clean) @ Hadoop_WordCount ---  
[INFO] -----  
[INFO] BUILD SUCCESS  
[INFO] -----  
[INFO] Total time: 0.290 s  
[INFO] Finished at: 2023-06-21T13:21:54+07:00  
[INFO] -----
```

- **mvn install**



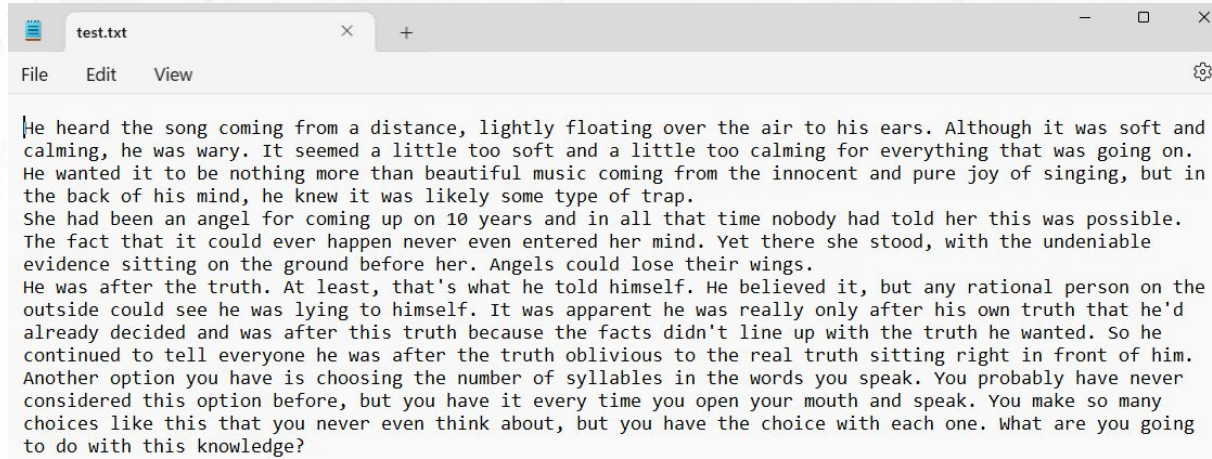
- ```
[INFO] --- maven-install-plugin:2.4:install (default-install) @ Hadoop_WordCount ---
[INFO] Installing D:\Universitas Indonesia\Semester 4\Sistem Basis Data\Hadoop\Hadoop_WordCount\target\Hadoop_WordCount-1.0-SNAPSHOT.jar to C:\Users\Asus\.m2\repository\com\example\Hadoop_WordCount\1.0-SNAPSHOT\Hadoop_WordCount-1.0-SNAPSHOT.jar
[INFO] Installing D:\Universitas Indonesia\Semester 4\Sistem Basis Data\Hadoop\Hadoop_WordCount\pom.xml to C:\Users\Asus\.m2\repository\com\example\Hadoop_WordCount\1.0-SNAPSHOT\Hadoop_WordCount-1.0-SNAPSHOT.pom
[INFO] -----
[INFO] BUILD SUCCESS
[INFO] -----
[INFO] Total time: 4.160 s
[INFO] Finished at: 2023-06-21T13:22:56+07:00
[INFO] -----
```

- Maka pada target folder terdapat file jar.



- Buat input file (.txt)

example :



- run cmd as administrator kemudian cd ke folder sbin

```
C:\Windows\System32>cd C:\hadoop-3.2.2\sbin\

C:\hadoop-3.2.2\sbin>_
```



- Menjalankan Hadoop serta jps untuk memastikan hadoop berjalan semestinya.

Start-all.cmd

```
C:\hadoop-3.2.2\sbin>start-all.cmd
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons

C:\hadoop-3.2.2\sbin>jps
13008
288 RemoteMavenServer36
11236 NameNode
7332 NodeManager
9224 Jps
13612 DataNode
3180 ResourceManager

C:\hadoop-3.2.2\sbin>
```

- `hadoop fs -mkdir /testinp`

```
C:\hadoop-3.2.2\sbin>hadoop fs -mkdir /testinp
```

- `hadoop fs -put "D:\Universitas Indonesia\Semester 4\Sistem Basis Data\Hadoop\1M.txt" /testinp`

```
C:\hadoop-3.2.2\sbin>hadoop fs -put "D:\Universitas Indonesia\Semester 4\Sistem Basis Data\Hadoop\1M.txt" /testinp
```

- `hadoop jar "D:\Universitas Indonesia\Semester 4\Sistem Basis Data\Hadoop\Hadoop_WordCount\target\Hadoop_WordCount-1.0-SNAPSHOT.jar" org.wordcount.WordCount /testinp /output_test`

```
C:\hadoop-3.2.2\sbin>hadoop jar D:\Universitas Indonesia\Semester 4\Sistem Basis Data\Hadoop\Hadoop_WordCount\target\Hadoop_WordCount-1.0-SNAPSHOT.jar org.wordcount.WordCount /testinp /output_test
JAR does not exist or is not a normal file: D:\Universitas

C:\hadoop-3.2.2\sbin>hadoop jar "D:\Universitas Indonesia\Semester 4\Sistem Basis Data\Hadoop\Hadoop_WordCount\target\Hadoop_WordCount-1.0-SNAPSHOT.jar" org.wordcount.WordCount /testinp /output_test
2023-06-21 13:37:54,318 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2023-06-21 13:37:55,072 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2023-06-21 13:37:55,086 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/Asus/.staging/job_1687329215005_0001
2023-06-21 13:37:55,332 INFO input.FileInputFormat: Total input files to process : 1
2023-06-21 13:37:55,397 INFO mapreduce.JobSubmitter: number of splits:1
2023-06-21 13:37:55,560 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1687329215005_0001
2023-06-21 13:37:55,562 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-06-21 13:37:55,752 INFO conf.Configuration: resource-types.xml not found
2023-06-21 13:37:55,753 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-06-21 13:37:56,036 INFO impl.YarnClientImpl: Submitted application application_1687329215005_0001
2023-06-21 13:37:56,092 INFO mapreduce.Job: The url to track the job: http://LAPTOP-7QHLSHQ:8088/proxy/application_1687329215005_0001/
2023-06-21 13:37:56,092 INFO mapreduce.Job: Running job: job_1687329215005_0001
2023-06-21 13:38:10,335 INFO mapreduce.Job: Job job_1687329215005_0001 running in uber mode : false
2023-06-21 13:38:10,336 INFO mapreduce.Job: map 0% reduce 0%
2023-06-21 13:38:17,460 INFO mapreduce.Job: map 100% reduce 0%
2023-06-21 13:38:24,540 INFO mapreduce.Job: map 100% reduce 100%
```

- Output

hadoop fs -cat /output\_test/part-r-00000

```
C:\hadoop-3.2.2\sbin>hadoop fs -cat /output_test/part-r-00000
"Begin 60
"I'll 60
"It 60
"You 61
>Your 61
"bridge", 60
"easy" 60
(I 60
-- 120
1 61
10 61
100 121
100-foot 60
24 60
24-hour 60
25 61
A 182
Actually, 60
After 60
All 120
Although 61
```

- Output  
stop-all.cmd

```
C:\hadoop-3.2.2\sbin>stop-all.cmd
This script is Deprecated. Instead use stop-dfs.cmd and stop-yarn.cmd
SUCCESS: Sent termination signal to the process with PID 9240.
SUCCESS: Sent termination signal to the process with PID 6824.
stopping yarn daemons
SUCCESS: Sent termination signal to the process with PID 11268.
SUCCESS: Sent termination signal to the process with PID 15348.

INFO: No tasks running with the specified criteria.
```



# Thanks!

