APACHE hadoop

Hadoop For Map Reduce Kelompok 2

Abdul Fikih Kurnia	2106731200
Bintang Marsyuma R	2106731415
Ibrahim Rijal	2106633323
Stefan Agusto Hutapea	2106700744

Introduction

Hadoop is a classic example used to demonstrate the basic functionality of Hadoop MapReduce framework. It involves counting the occurrences of each word in a collection of documents.

Introduction

- The MapReduce framework takes care of task scheduling, fault tolerance, and data distribution across the cluster. It automatically manages the parallel execution of map and reduce tasks on different nodes, allowing for scalable and efficient data processing.

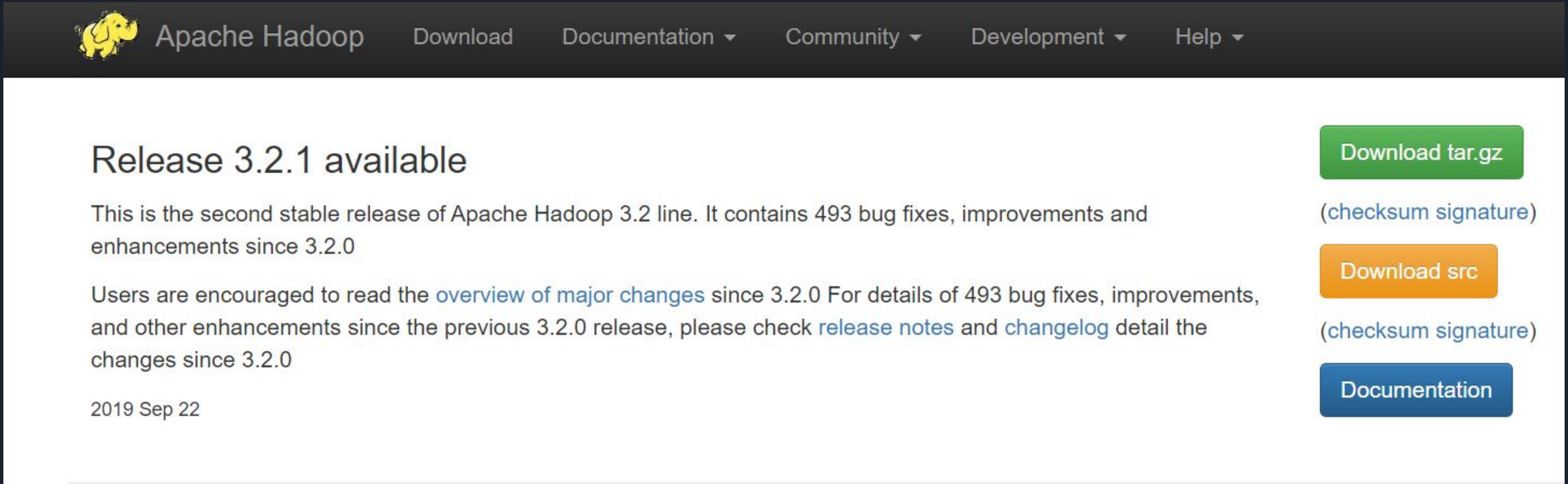
- Hadoop MapReduce is designed for batch processing of large-scale data. It is particularly useful for tasks like log analysis, data transformation, data aggregation, and large-scale computations, where the data can be processed in parallel across a cluster of machines.

INSTALLATIONS

Go to Repo for more information : https://github.com/irbirojodoh/SBD-Hadoop-Analysis

Download Hadoop 3.2.1:

Binaries can be found here:
https://hadoop.apache.org/release/3.2.1.html



Release 3.2.1 available

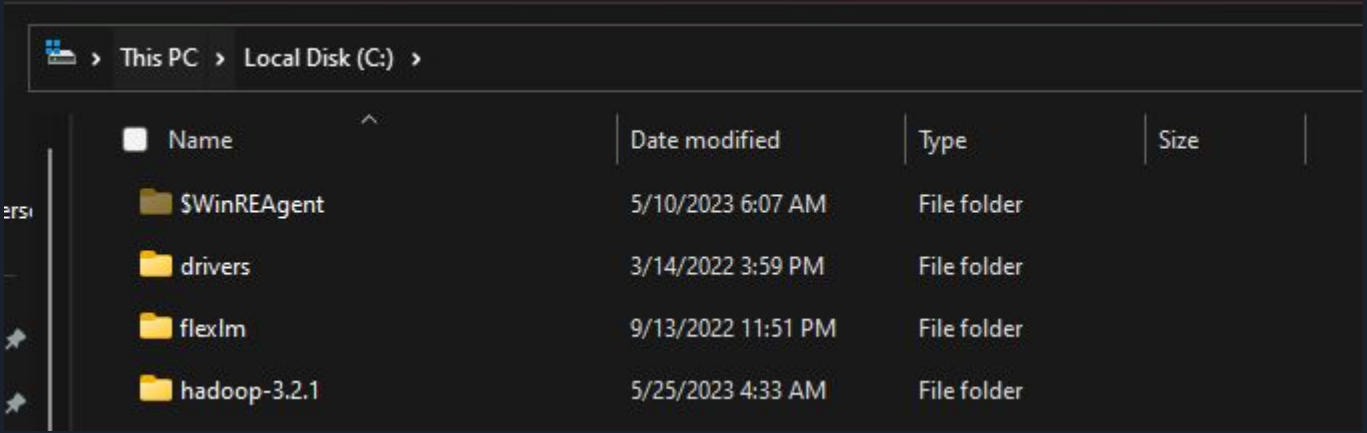
This is the second stable release of Apache Hadoop 3.2 line. It contains 493 bug fixes, improvements and enhancements since 3.2.0

Users are encouraged to read the overview of major changes since 3.2.0 For details of 493 bug fixes, improvements, and other enhancements since the previous 3.2.0 release, please check release notes and changelog detail the changes since 3.2.0

2019 Sep 22

Download the binaries, and put it in C:

Binaries can be found here: https://www.apache.org/dyn/closer.cgi/hadoop/common/hadoop-3.2.1/hadoop-3.2.1.tar.gz

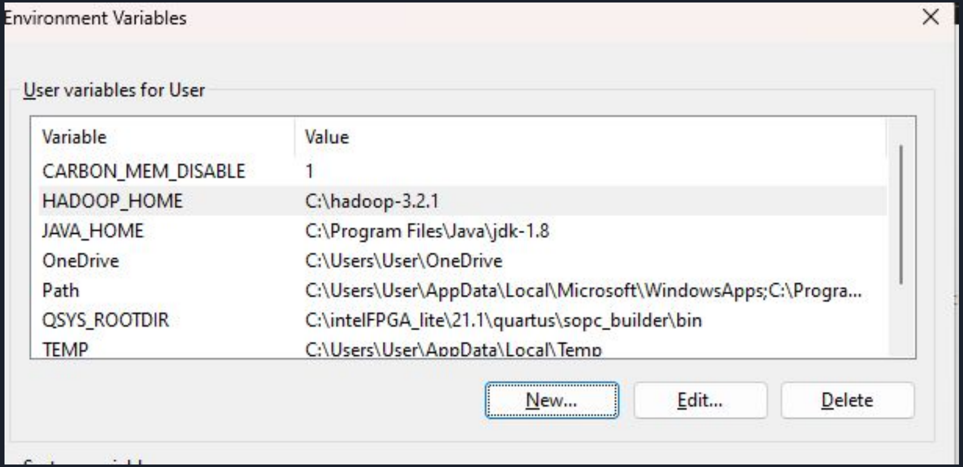


Name	Date modified	Type	Size
$WinREAgent	5/10/2023 6:07 AM	File folder	
drivers	3/14/2022 3:59 PM	File folder	
flexlm	9/13/2022 11:51 PM	File folder	
hadoop-3.2.1	5/25/2023 4:33 AM	File folder	

Setup environment

This step is needed so that Command Line can recognise hadoop commands.

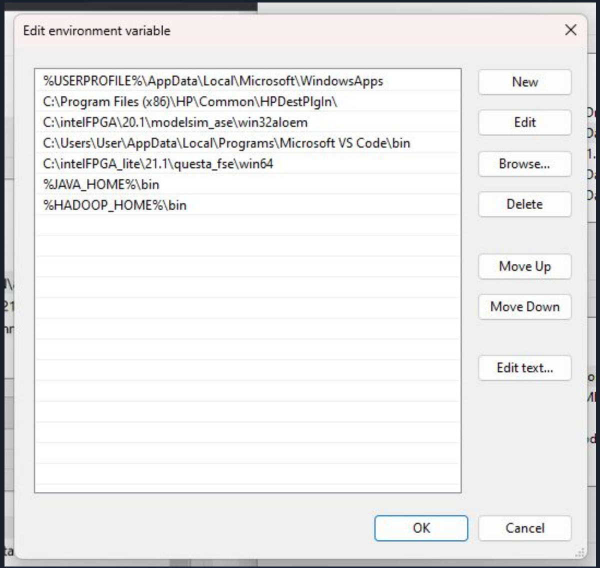
*Be careful, missteps can lead to the CMD being unusable



Environment Variables

User variables for User

Variable	Value
CARBON_MEM_DISABLE	1
HADOOP_HOME	C:\hadoop-3.2.1
JAVA_HOME	C:\Program Files\Java\jdk-1.8
OneDrive	C:\Users\User\OneDrive
Path	C:\Users\User\AppData\Local\Microsoft\WindowsApps;C:\Progra...
QSYS_ROOTDIR	C:\intelFPGA_lite\21.1\quartus\sopc_builder\bin
TEMP	C:\Users\User\AppData\Local\Temp

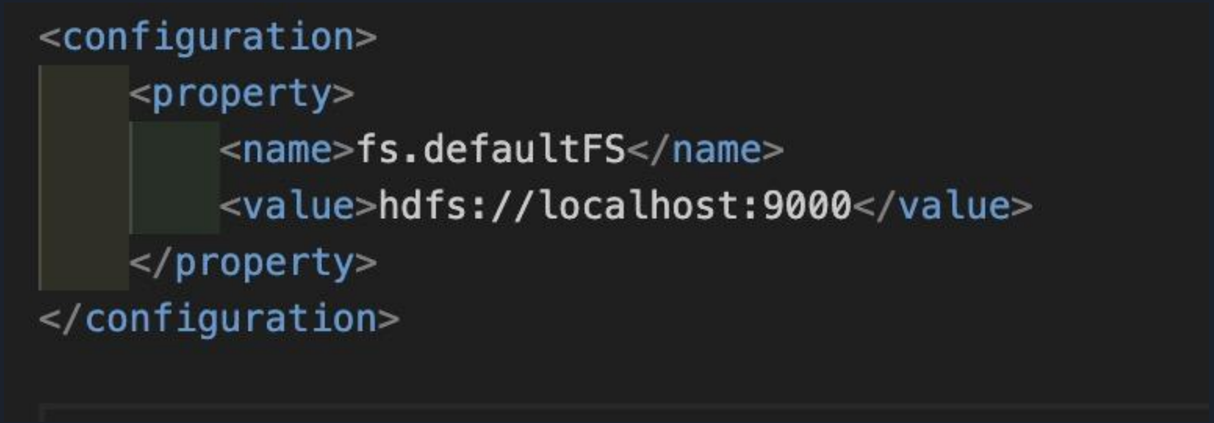


Edit environment variable

```
%USERPROFILE%\AppData\Local\Microsoft\WindowsApps
C:\Program Files (x86)\HP\Common\HPDestPlgIn\
C:\intelFPGA\20.1\modelsim_ase\win32aloem
C:\Users\User\AppData\Local\Programs\Microsoft VS Code\bin
C:\intelFPGA_lite\21.1\questa_fse\win64
%JAVA_HOME%\bin
%HADOOP_HOME%\bin
```

Setup core-site.xml

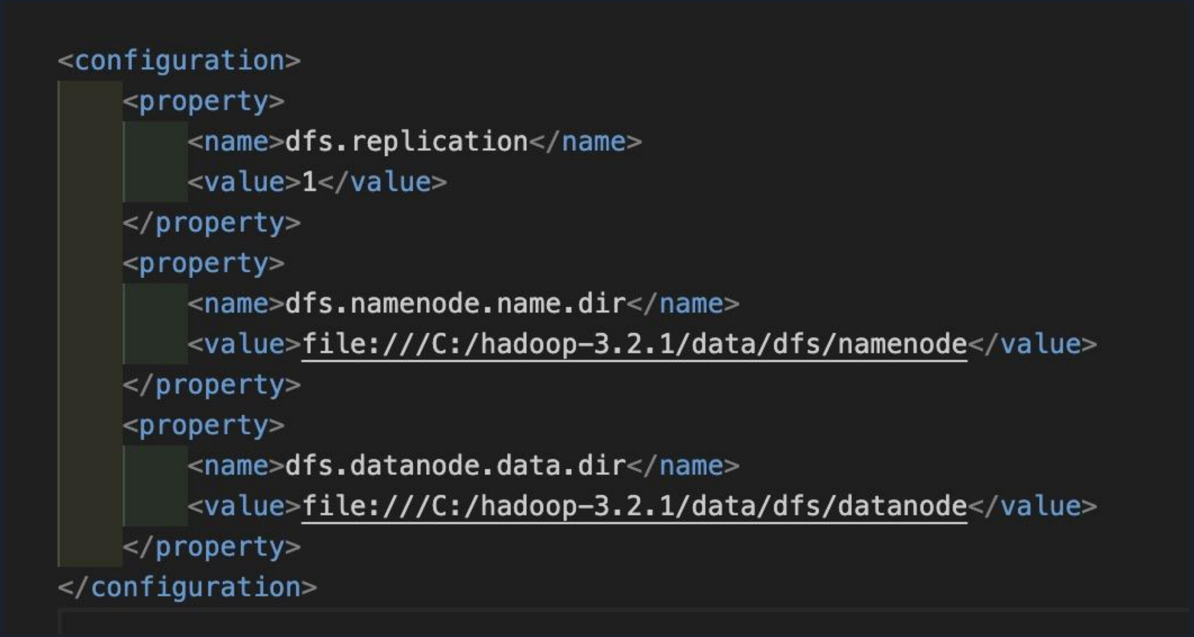
This XML configuration sets the default file system (fs.defaultFS) to HDFS running on localhost with port 9000.



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

Setup hdfs-site.xml

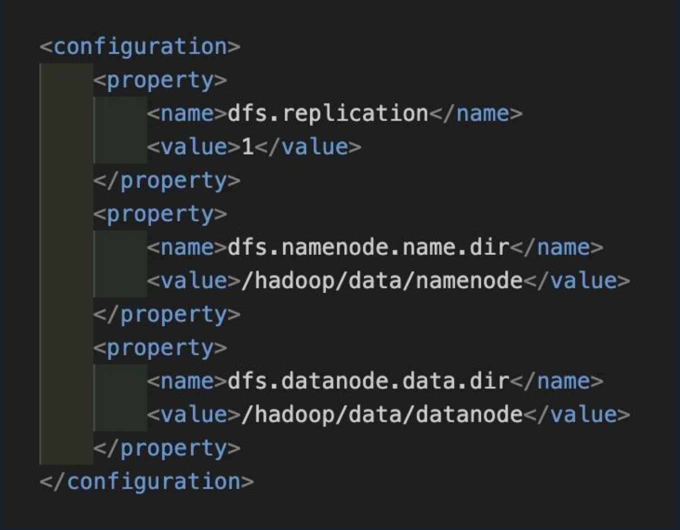
This configuration sets the replication factor (dfs.replication) to 1, specifies the directory for the namenode (dfs.namenode.name.dir) as "file:///C:/hadoop-3.2.1/data/dfs/namenode", and sets the directory for the datanode (dfs.datanode.data.dir) as "file:///C:/hadoop-3.2.1/data/dfs/datanode".



```
<configuration>
    <property>
        <name>dfs.replication</name>
        <value>1</value>
    </property>
    <property>
        <name>dfs.namenode.name.dir</name>
        <value>file:///C:/hadoop-3.2.1/data/dfs/namenode</value>
    </property>
    <property>
        <name>dfs.datanode.data.dir</name>
        <value>file:///C:/hadoop-3.2.1/data/dfs/datanode</value>
    </property>
</configuration>
```

Setup mapred-site.xml

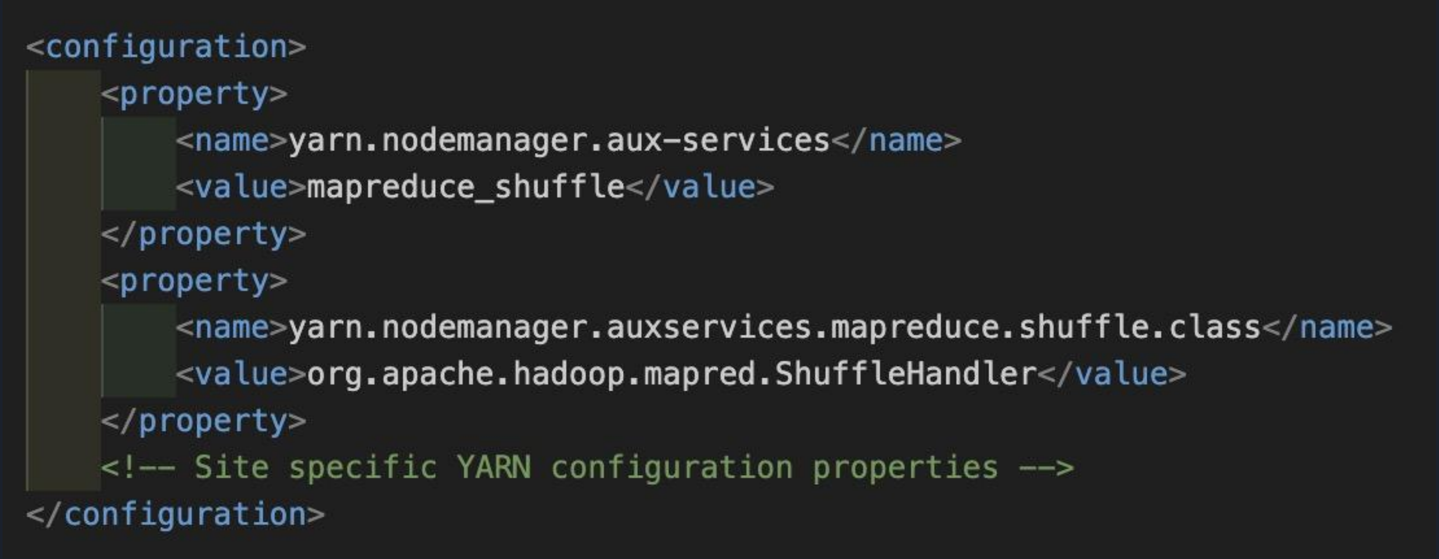
This configuration sets the replication factor (dfs.replication) to 1, specifies the directory for the namenode (dfs.namenode.name.dir) as "/hadoop/data/namenode", and sets the directory for the datanode (dfs.datanode.data.dir) as "/hadoop/data/datanode".



```
<configuration>
    <property>
        <name>dfs.replication</name>
        <value>1</value>
    </property>
    <property>
        <name>dfs.namenode.name.dir</name>
        <value>/hadoop/data/namenode</value>
    </property>
    <property>
        <name>dfs.datanode.data.dir</name>
        <value>/hadoop/data/datanode</value>
    </property>
</configuration>
```

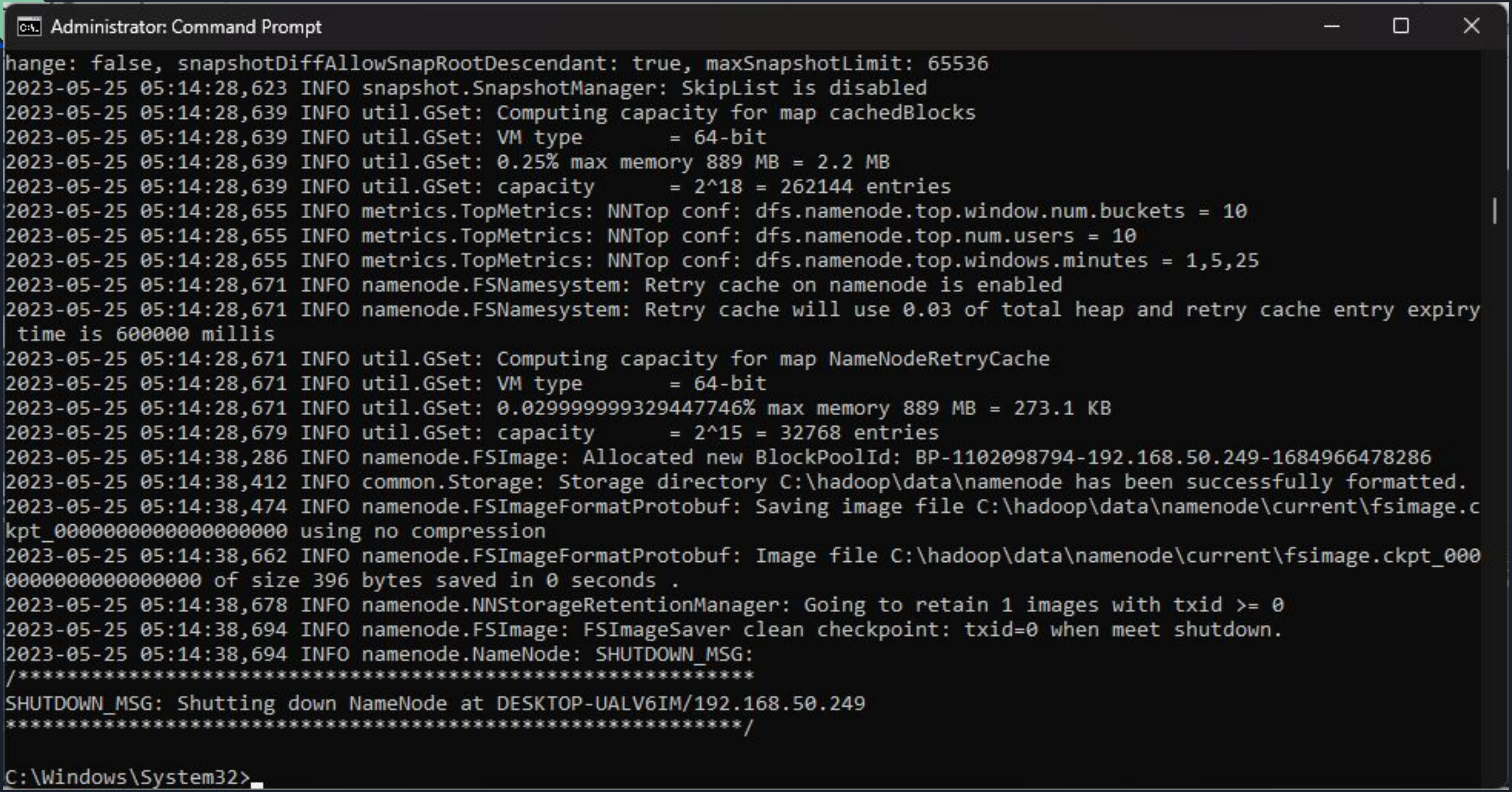
Setup yarn-site.xml

This configuration sets the auxiliary services for the nodemanager (yarn.nodemanager.aux-services) to "mapreduce_shuffle" and specifies the class for the mapreduce shuffle service (yarn.nodemanager.auxservices.mapreduce.shuffle.class) as "org.apache.hadoop.mapred.ShuffleHandler". It also includes a comment for site-specific YARN configuration properties.



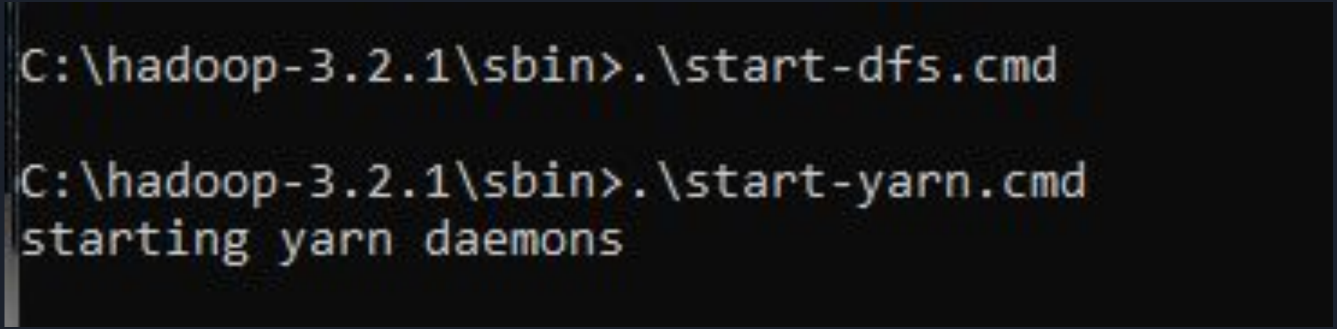
```xml
<configuration>
    <property>
        <name>yarn.nodemanager.aux-services</name>
        <value>mapreduce_shuffle</value>
    </property>
    <property>
        <name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
        <value>org.apache.hadoop.mapred.ShuffleHandler</value>
    </property>
    <!-- Site specific YARN configuration properties -->
</configuration>
```

Format NameNode

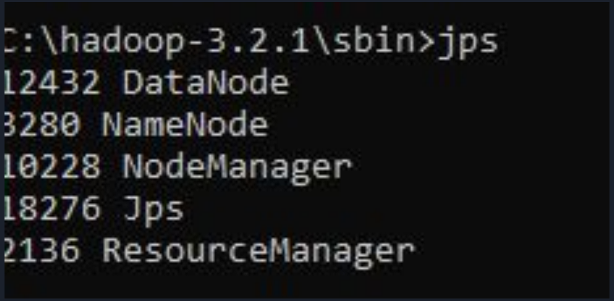


```
hange: false, snapshotDiffAllowSnapRootDescendant: true, maxSnapshotLimit: 65536
2023-05-25 05:14:28,623 INFO snapshot.SnapshotManager: SkipList is disabled
2023-05-25 05:14:28,639 INFO util.GSet: Computing capacity for map cachedBlocks
2023-05-25 05:14:28,639 INFO util.GSet: VM type       = 64-bit
2023-05-25 05:14:28,639 INFO util.GSet: 0.25% max memory 889 MB = 2.2 MB
2023-05-25 05:14:28,639 INFO util.GSet: capacity      = 2^18 = 262144 entries
2023-05-25 05:14:28,655 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
2023-05-25 05:14:28,655 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
2023-05-25 05:14:28,655 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
2023-05-25 05:14:28,671 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
2023-05-25 05:14:28,671 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry
 time is 600000 millis
2023-05-25 05:14:28,671 INFO util.GSet: Computing capacity for map NameNodeRetryCache
2023-05-25 05:14:28,671 INFO util.GSet: VM type       = 64-bit
2023-05-25 05:14:28,671 INFO util.GSet: 0.029999999329447746% max memory 889 MB = 273.1 KB
2023-05-25 05:14:28,679 INFO util.GSet: capacity      = 2^15 = 32768 entries
2023-05-25 05:14:38,286 INFO namenode.FSImage: Allocated new BlockPoolId: BP-1102098794-192.168.50.249-1684966478286
2023-05-25 05:14:38,412 INFO common.Storage: Storage directory C:\hadoop\data\namenode has been successfully formatted.
2023-05-25 05:14:38,474 INFO namenode.FSImageFormatProtobuf: Saving image file C:\hadoop\data\namenode\current\fsimage.c
kpt_0000000000000000000 using no compression
2023-05-25 05:14:38,662 INFO namenode.FSImageFormatProtobuf: Image file C:\hadoop\data\namenode\current\fsimage.ckpt_000
0000000000000000 of size 396 bytes saved in 0 seconds .
2023-05-25 05:14:38,678 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2023-05-25 05:14:38,694 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2023-05-25 05:14:38,694 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at DESKTOP-UALV6IM/192.168.50.249
************************************************************/

C:\Windows\System32>
```

Start all daemon and verify

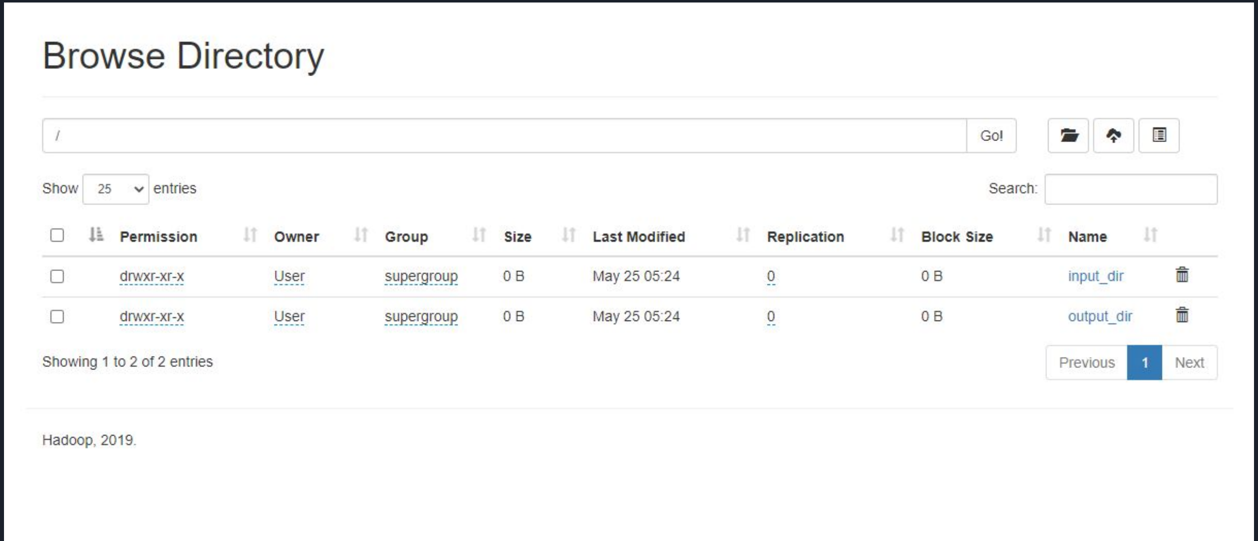


```
C:\hadoop-3.2.1\sbin>.\start-dfs.cmd

C:\hadoop-3.2.1\sbin>.\start-yarn.cmd
starting yarn daemons
```



```
C:\hadoop-3.2.1\sbin>jps
12432 DataNode
3280 NameNode
10228 NodeManager
18276 Jps
2136 ResourceManager
```

Create input and output dir



Browse Directory

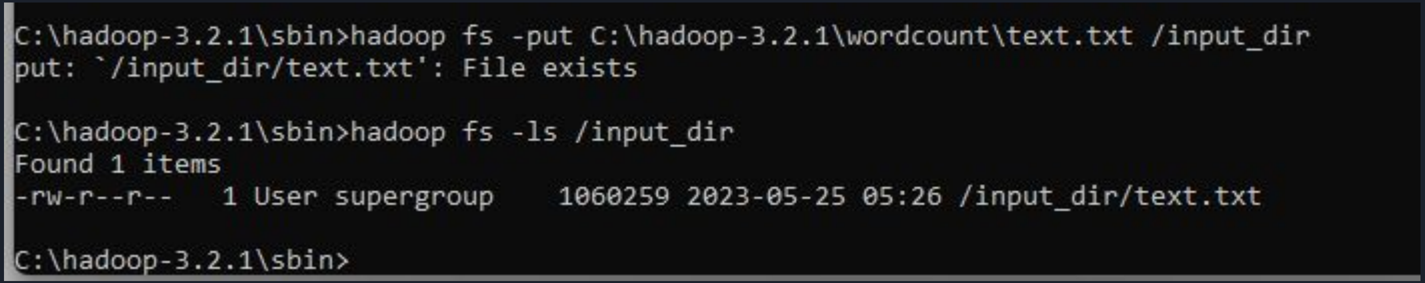
Show 25 entries Search:

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
□	drwxr-xr-x	User	supergroup	0 B	May 25 05:24	0	0 B	input_dir
□	drwxr-xr-x	User	supergroup	0 B	May 25 05:24	0	0 B	output_dir

Showing 1 to 2 of 2 entries

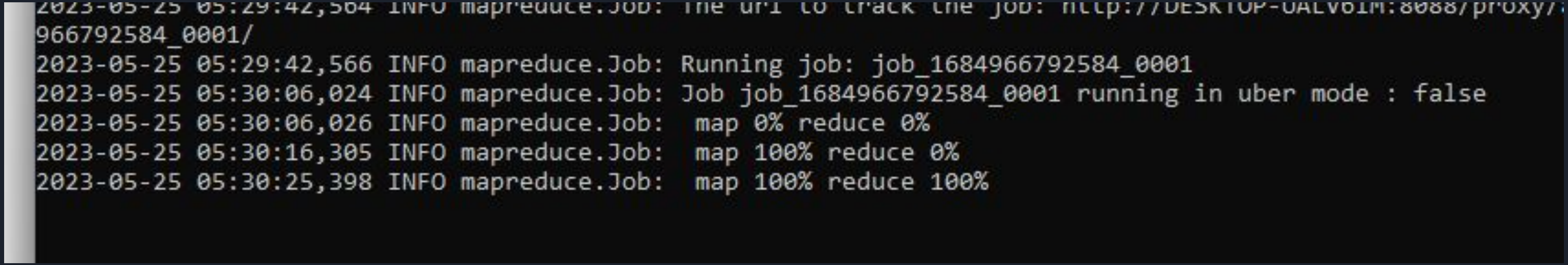
Hadoop, 2019.

Put file into input_dir



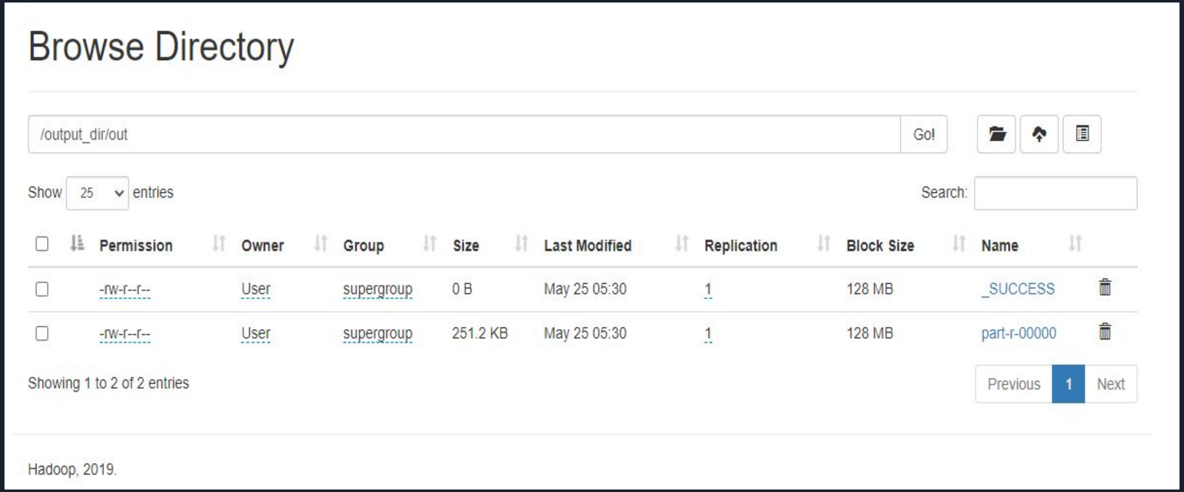
```
C:\hadoop-3.2.1\sbin>hadoop fs -put C:\hadoop-3.2.1\wordcount\text.txt /input_dir
put: `/input_dir/text.txt': File exists

C:\hadoop-3.2.1\sbin>hadoop fs -ls /input_dir
Found 1 items
-rw-r--r--   1 User supergroup    1060259 2023-05-25 05:26 /input_dir/text.txt

C:\hadoop-3.2.1\sbin>
```

Run mapreduce



```
2023-05-25 05:29:42,564 INFO mapreduce.Job: The url to track the job: http://DESKTOP-UALV6IM:8088/proxy/
966792584_0001/
2023-05-25 05:29:42,566 INFO mapreduce.Job: Running job: job_1684966792584_0001
2023-05-25 05:30:06,024 INFO mapreduce.Job: Job job_1684966792584_0001 running in uber mode : false
2023-05-25 05:30:06,026 INFO mapreduce.Job:  map 0% reduce 0%
2023-05-25 05:30:16,305 INFO mapreduce.Job:  map 100% reduce 0%
2023-05-25 05:30:25,398 INFO mapreduce.Job:  map 100% reduce 100%
```

See results



Browse Directory

/output_dir/out Go!

Show 25 entries

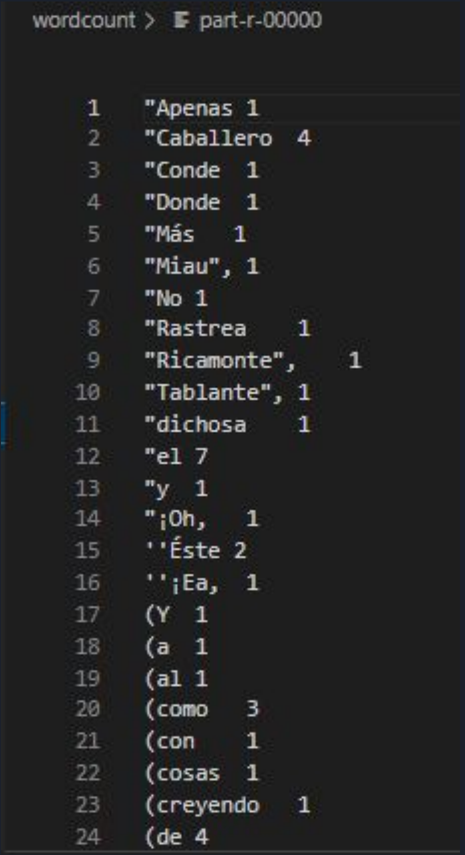
Search:

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
□	-rw-r--r--	User	supergroup	0 B	May 25 05:30	1	128 MB	_SUCCESS
□	-rw-r--r--	User	supergroup	251.2 KB	May 25 05:30	1	128 MB	part-r-00000

Showing 1 to 2 of 2 entries

Previous 1 Next

Hadoop, 2019.



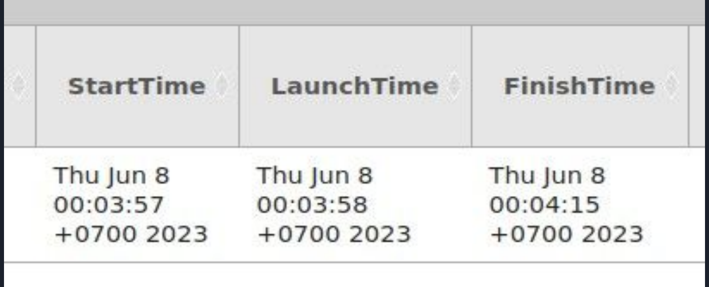
wordcount > part-r-00000

```
"Apenas 1
"Caballero	4
"Conde	1
"Donde	1
"Más	1
"Miau",	1
"No	1
"Rastrea	1
"Ricamonte",	1
"Tablante",	1
"dichosa	1
"el	7
"y	1
"¡Oh,	1
''Éste	2
''¡Ea,	1
(Y	1
(a	1
(al	1
(como	3
(con	1
(cosas	1
(creyendo	1
(de	4
```

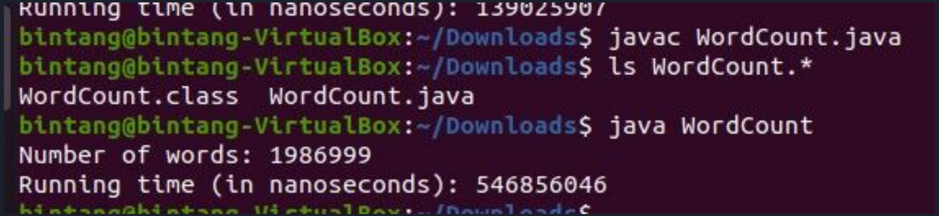
EXPERIMENTS

Comparison
Hadoop vs Native Java

1 MB Text File



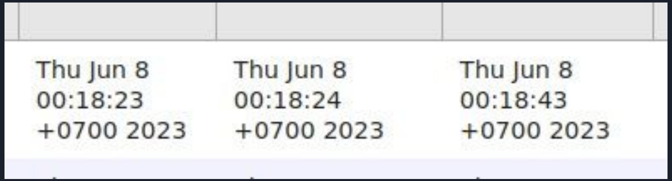
StartTime	LaunchTime	FinishTime
Thu Jun 8 00:03:57 +0700 2023	Thu Jun 8 00:03:58 +0700 2023	Thu Jun 8 00:04:15 +0700 2023



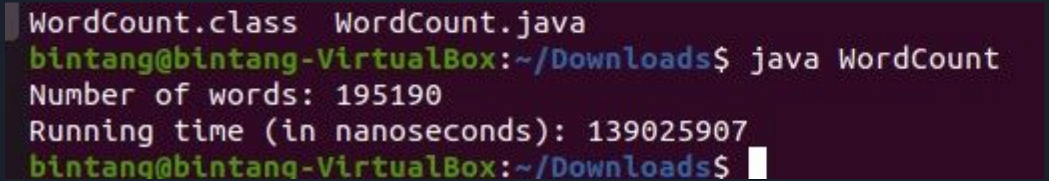
```
Running time (in nanoseconds): 139025907
bintang@bintang-VirtualBox:~/Downloads$ javac WordCount.java
bintang@bintang-VirtualBox:~/Downloads$ ls WordCount.*
WordCount.class  WordCount.java
bintang@bintang-VirtualBox:~/Downloads$ java WordCount
Number of words: 1986999
Running time (in nanoseconds): 546856046
bintang@bintang-VirtualBox:~/Downloads$
```

Comparison
Hadoop vs Native Java

10 MB Text File

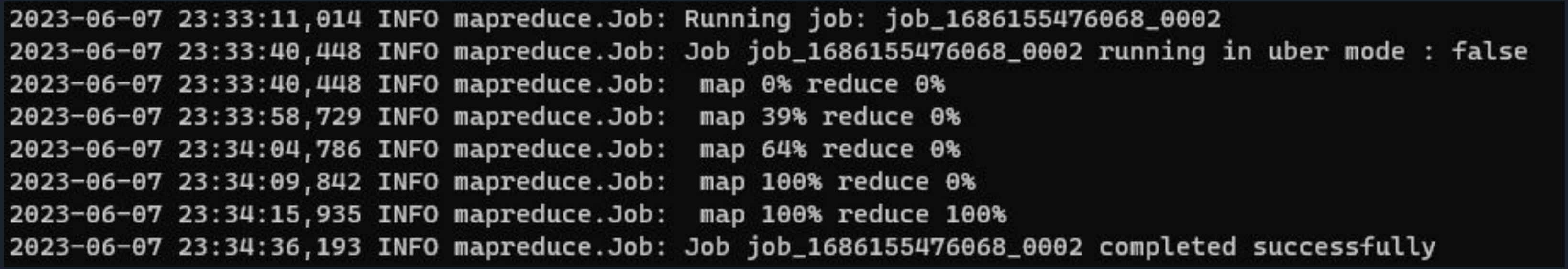


Thu Jun 8 00:18:23 +0700 2023	Thu Jun 8 00:18:24 +0700 2023	Thu Jun 8 00:18:43 +0700 2023

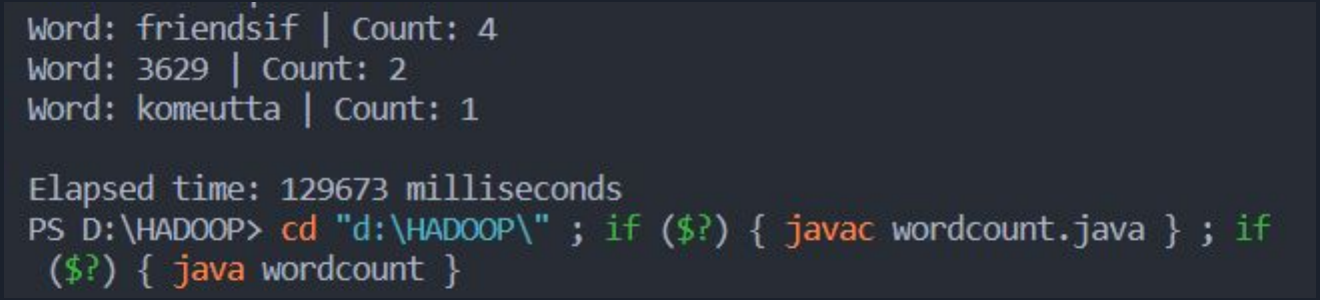


```
WordCount.class  WordCount.java
bintang@bintang-VirtualBox:~/Downloads$ java WordCount
Number of words: 195190
Running time (in nanoseconds): 139025907
bintang@bintang-VirtualBox:~/Downloads$
```

100 Mb Text File

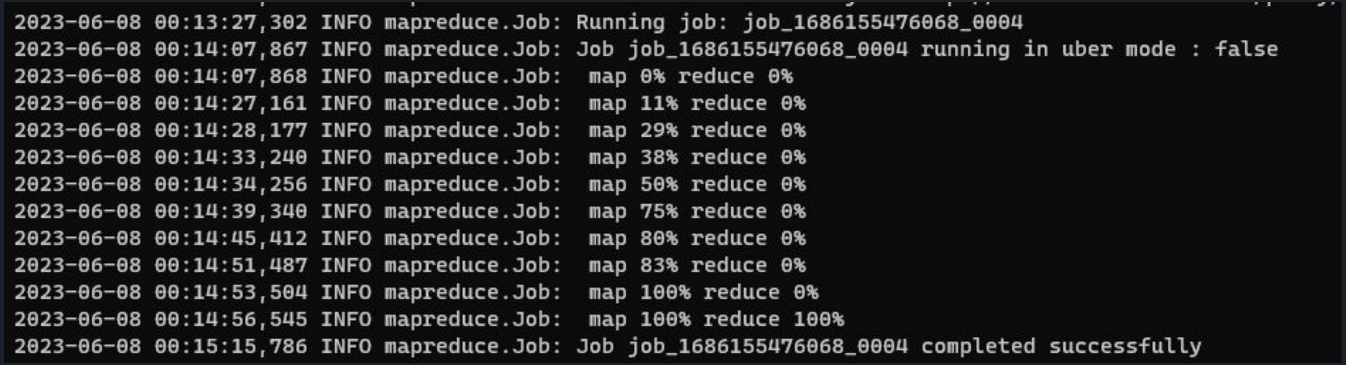


```
2023-06-07 23:33:11,014 INFO mapreduce.Job: Running job: job_1686155476068_0002
2023-06-07 23:33:40,448 INFO mapreduce.Job: Job job_1686155476068_0002 running in uber mode : false
2023-06-07 23:33:40,448 INFO mapreduce.Job:  map 0% reduce 0%
2023-06-07 23:33:58,729 INFO mapreduce.Job:  map 39% reduce 0%
2023-06-07 23:34:04,786 INFO mapreduce.Job:  map 64% reduce 0%
2023-06-07 23:34:09,842 INFO mapreduce.Job:  map 100% reduce 0%
2023-06-07 23:34:15,935 INFO mapreduce.Job:  map 100% reduce 100%
2023-06-07 23:34:36,193 INFO mapreduce.Job: Job job_1686155476068_0002 completed successfully
```

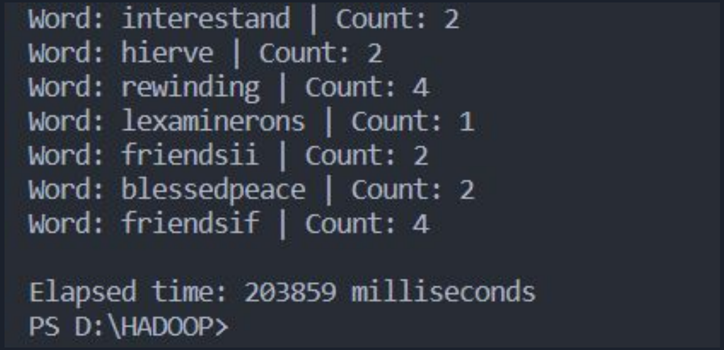


```
Word: friendsif | Count: 4
Word: 3629 | Count: 2
Word: komeutta | Count: 1

Elapsed time: 129673 milliseconds
PS D:\HADOOP> cd "d:\HADOOP\" ; if ($?) { javac wordcount.java } ; if
 ($?) { java wordcount }
```

200 MB Text File

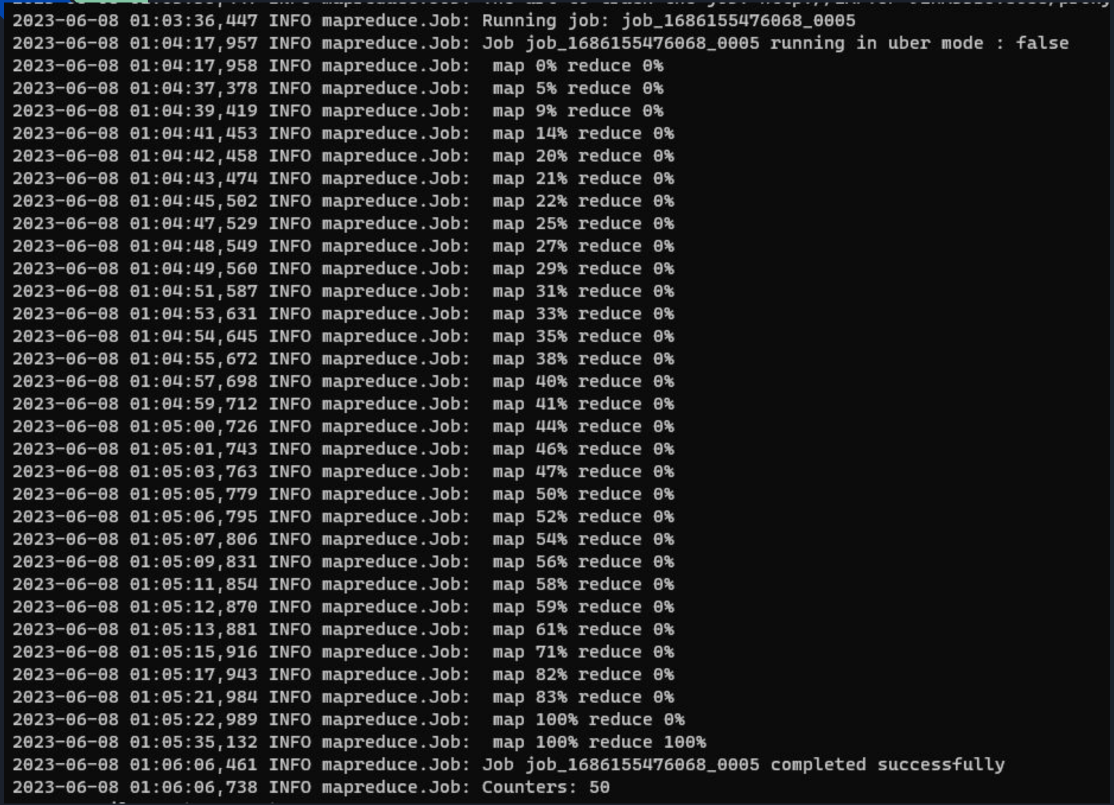


```
2023-06-08 00:13:27,302 INFO mapreduce.Job: Running job: job_1686155476068_0004
2023-06-08 00:14:07,867 INFO mapreduce.Job: Job job_1686155476068_0004 running in uber mode : false
2023-06-08 00:14:07,868 INFO mapreduce.Job:  map 0% reduce 0%
2023-06-08 00:14:27,161 INFO mapreduce.Job:  map 11% reduce 0%
2023-06-08 00:14:28,177 INFO mapreduce.Job:  map 29% reduce 0%
2023-06-08 00:14:33,240 INFO mapreduce.Job:  map 38% reduce 0%
2023-06-08 00:14:34,256 INFO mapreduce.Job:  map 50% reduce 0%
2023-06-08 00:14:39,340 INFO mapreduce.Job:  map 75% reduce 0%
2023-06-08 00:14:45,412 INFO mapreduce.Job:  map 80% reduce 0%
2023-06-08 00:14:51,487 INFO mapreduce.Job:  map 83% reduce 0%
2023-06-08 00:14:53,504 INFO mapreduce.Job:  map 100% reduce 0%
2023-06-08 00:14:56,545 INFO mapreduce.Job:  map 100% reduce 100%
2023-06-08 00:15:15,786 INFO mapreduce.Job: Job job_1686155476068_0004 completed successfully
```

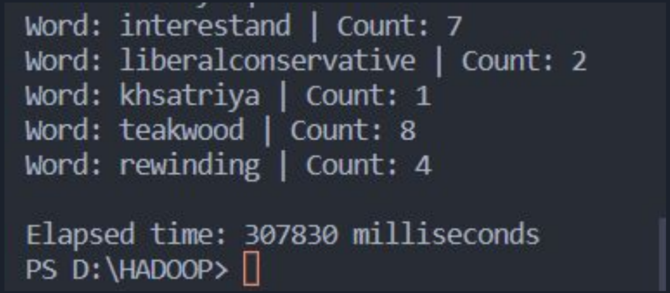


```
Word: interestand | Count: 2
Word: hierve | Count: 2
Word: rewinding | Count: 4
Word: lexaminerons | Count: 1
Word: friendsii | Count: 2
Word: blessedpeace | Count: 2
Word: friendsif | Count: 4

Elapsed time: 203859 milliseconds
PS D:\HADOOP>
```

500 MB Text File

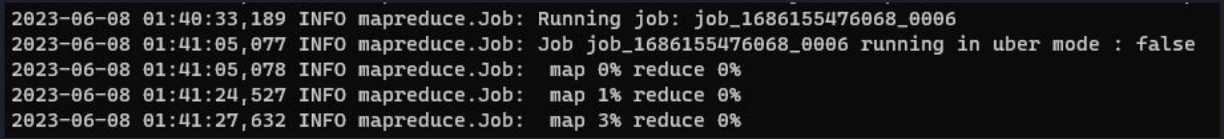


```
2023-06-08 01:03:36,447 INFO mapreduce.Job: Running job: job_1686155476068_0005
2023-06-08 01:04:17,957 INFO mapreduce.Job: Job job_1686155476068_0005 running in uber mode : false
2023-06-08 01:04:17,958 INFO mapreduce.Job:  map 0% reduce 0%
2023-06-08 01:04:37,378 INFO mapreduce.Job:  map 5% reduce 0%
2023-06-08 01:04:39,419 INFO mapreduce.Job:  map 9% reduce 0%
2023-06-08 01:04:41,453 INFO mapreduce.Job:  map 14% reduce 0%
2023-06-08 01:04:42,458 INFO mapreduce.Job:  map 20% reduce 0%
2023-06-08 01:04:43,474 INFO mapreduce.Job:  map 21% reduce 0%
2023-06-08 01:04:45,502 INFO mapreduce.Job:  map 22% reduce 0%
2023-06-08 01:04:47,529 INFO mapreduce.Job:  map 25% reduce 0%
2023-06-08 01:04:48,549 INFO mapreduce.Job:  map 27% reduce 0%
2023-06-08 01:04:49,560 INFO mapreduce.Job:  map 29% reduce 0%
2023-06-08 01:04:51,587 INFO mapreduce.Job:  map 31% reduce 0%
2023-06-08 01:04:53,631 INFO mapreduce.Job:  map 33% reduce 0%
2023-06-08 01:04:54,645 INFO mapreduce.Job:  map 35% reduce 0%
2023-06-08 01:04:55,672 INFO mapreduce.Job:  map 38% reduce 0%
2023-06-08 01:04:57,698 INFO mapreduce.Job:  map 40% reduce 0%
2023-06-08 01:04:59,712 INFO mapreduce.Job:  map 41% reduce 0%
2023-06-08 01:05:00,726 INFO mapreduce.Job:  map 44% reduce 0%
2023-06-08 01:05:01,743 INFO mapreduce.Job:  map 46% reduce 0%
2023-06-08 01:05:03,763 INFO mapreduce.Job:  map 47% reduce 0%
2023-06-08 01:05:05,779 INFO mapreduce.Job:  map 50% reduce 0%
2023-06-08 01:05:06,795 INFO mapreduce.Job:  map 52% reduce 0%
2023-06-08 01:05:07,806 INFO mapreduce.Job:  map 54% reduce 0%
2023-06-08 01:05:09,831 INFO mapreduce.Job:  map 56% reduce 0%
2023-06-08 01:05:11,854 INFO mapreduce.Job:  map 58% reduce 0%
2023-06-08 01:05:12,870 INFO mapreduce.Job:  map 59% reduce 0%
2023-06-08 01:05:13,881 INFO mapreduce.Job:  map 61% reduce 0%
2023-06-08 01:05:15,916 INFO mapreduce.Job:  map 71% reduce 0%
2023-06-08 01:05:17,943 INFO mapreduce.Job:  map 82% reduce 0%
2023-06-08 01:05:21,984 INFO mapreduce.Job:  map 83% reduce 0%
2023-06-08 01:05:22,989 INFO mapreduce.Job:  map 100% reduce 0%
2023-06-08 01:05:35,132 INFO mapreduce.Job:  map 100% reduce 100%
2023-06-08 01:06:06,461 INFO mapreduce.Job: Job job_1686155476068_0005 completed successfully
2023-06-08 01:06:06,738 INFO mapreduce.Job: Counters: 50
```

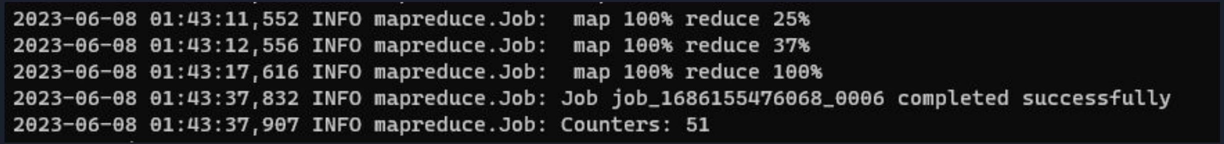


```
Word: interestand | Count: 7
Word: liberalconservative | Count: 2
Word: khsatriya | Count: 1
Word: teakwood | Count: 8
Word: rewinding | Count: 4

Elapsed time: 307830 milliseconds
PS D:\HADOOP>
```

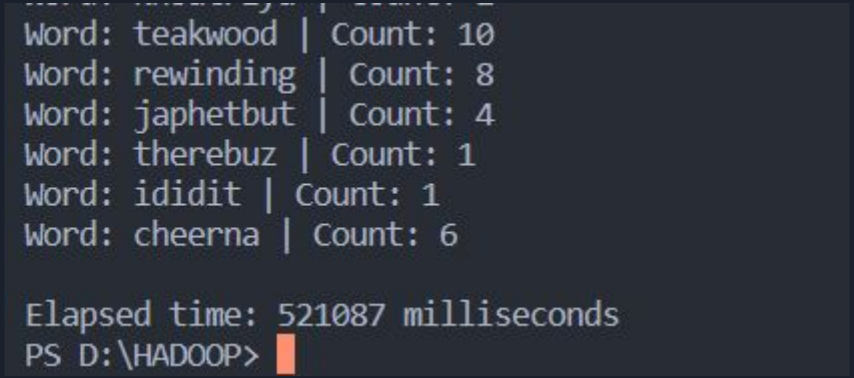
1000 MB Text File



```
2023-06-08 01:40:33,189 INFO mapreduce.Job: Running job: job_1686155476068_0006
2023-06-08 01:41:05,077 INFO mapreduce.Job: Job job_1686155476068_0006 running in uber mode : false
2023-06-08 01:41:05,078 INFO mapreduce.Job:  map 0% reduce 0%
2023-06-08 01:41:24,527 INFO mapreduce.Job:  map 1% reduce 0%
2023-06-08 01:41:27,632 INFO mapreduce.Job:  map 3% reduce 0%
```



```
2023-06-08 01:43:11,552 INFO mapreduce.Job:  map 100% reduce 25%
2023-06-08 01:43:12,556 INFO mapreduce.Job:  map 100% reduce 37%
2023-06-08 01:43:17,616 INFO mapreduce.Job:  map 100% reduce 100%
2023-06-08 01:43:37,832 INFO mapreduce.Job: Job job_1686155476068_0006 completed successfully
2023-06-08 01:43:37,907 INFO mapreduce.Job: Counters: 51
```

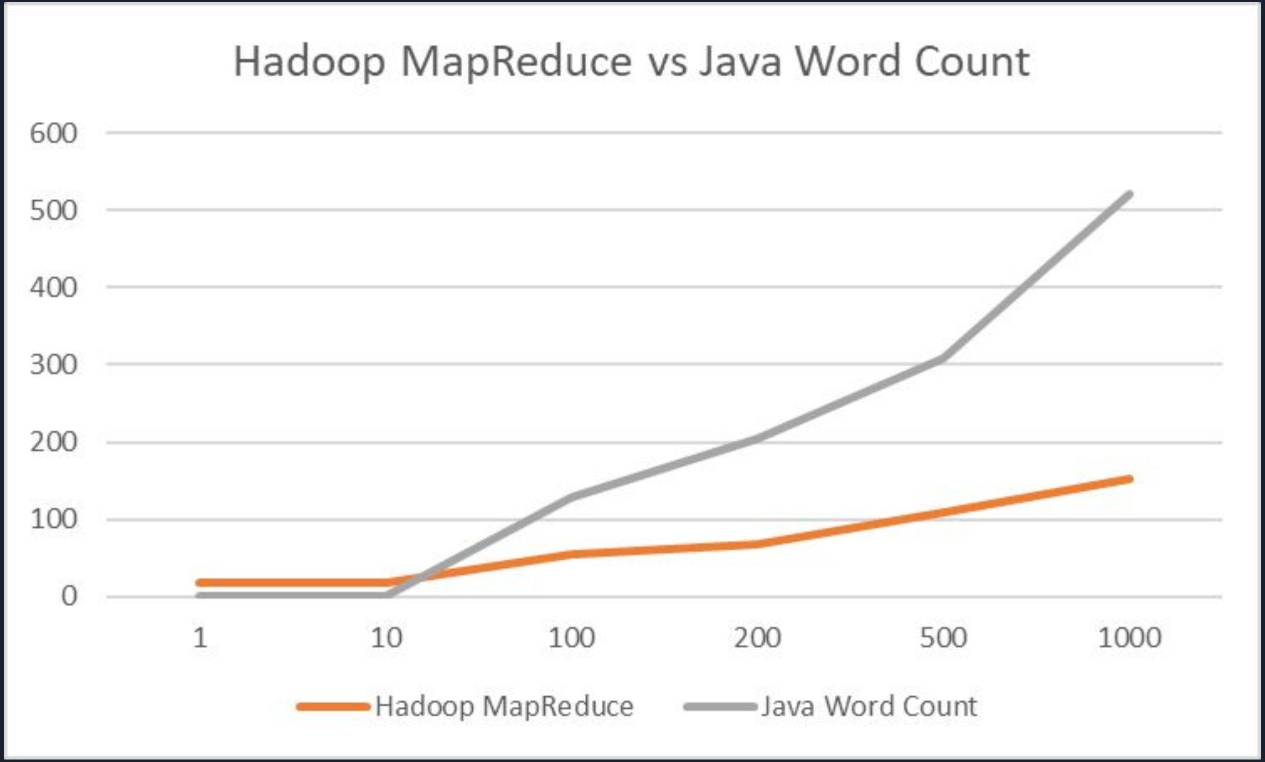


```
Word: teakwood | Count: 10
Word: rewinding | Count: 8
Word: japhetbut | Count: 4
Word: therebuz | Count: 1
Word: ididit | Count: 1
Word: cheerna | Count: 6

Elapsed time: 521087 milliseconds
PS D:\HADOOP>
```

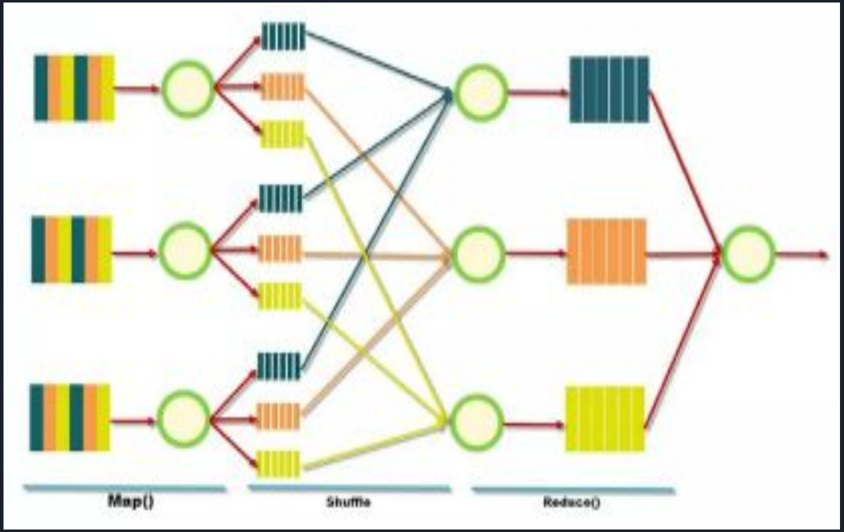
Table Compare

Size File	With Hadoop	Without Hadoop (Java)	Scalable (hadoop : java)
1 Mb	17 sec	0,546 sec	31,13
10 Mb	19 sec	1,39 sec	13,66
100 Mb	55,33 sec	129,67 sec	0,42
200 Mb	68,43 sec	203,85 sec	0,33
500 Mb	109,72 sec	307,83 sec	0,32
1000 Mb	152,86 sec	521,08 sec	0,29

Graphics Hadoop vs Java



Analyst Result



Based on the experiment results and the graph, A thorough analysis of the speed differences between Hadoop and the Java Word Count program can be explained as follows:

Hadoop is a framework for processing data designed to handle large volumes of data in a distributed manner. It uses the MapReduce approach, where data is divided into blocks, which are then processed in parallel by multiple nodes in a cluster. On the other hand, the Java Word Count program is a Java program that counts the number of words in a text without using any framework.

Analyst Result

1. In general, Hadoop is slower in performing Word Count compared to the standalone Java Word Count program. This can be observed from the recorded execution times in the table. Several factors contribute to this difference:

- a. Framework Overhead: Hadoop has additional overhead to manage during processing. This includes the process of dividing data into blocks, task scheduling, data transmission between nodes, and result synchronization. These overheads cause Hadoop to require additional time to initiate and coordinate processing tasks.

- b. Data Partitioning and Parallel Processing: While data partitioning and parallel processing in Hadoop enhance its scalability, they also introduce additional overhead in terms of data transmission between nodes and result synchronization. In some cases, this overhead can impact the performance of Hadoop, especially when dealing with relatively small datasets.

Analyst Result

2. Scalability: Despite Hadoop's slower performance in certain cases, it demonstrates an advantage in terms of scalability. Scalability refers to a system's ability to increase processing capacity as the volume of data increases. In the table, the values recorded in the "Scalable" column depict the comparison of processing speed between Hadoop and the standalone Java Word Count program as the dataset size increases. Hadoop maintains a relatively constant execution time, while the standalone Java Word Count program experiences a significant increase.

3. Dataset Size: The speed difference between Hadoop and the standalone Java Word Count program can also be influenced by the dataset size. For relatively small datasets, the standalone Java Word Count program performs better due to the overhead faced by Hadoop, which can become more significant in this context. However, as the dataset size increases, the advantages of parallel processing and Hadoop's scalability become more apparent.

Conclusion

In conclusion, Hadoop offers advantages in terms of scalability and distributed data processing but introduces additional overhead that can result in slower performance in certain cases. The standalone Java Word Count program tends to have better speed for relatively small datasets. The choice between Hadoop and the standalone Java Word Count program depends on the specific requirements of the data processing task, the dataset size, and the desired scalability.

TERIMA KASIH
SETIAP HARI PENUH CERITA