MODUL 10
HDFS dan MapReduce

A. Tujuan :

1. Mampu mengimplementasikan HDFS
2. Mampu mengimplementasikan MapReduce

B. Dasar Teori

Hadoop merupakan sebuah framework yang terus dikembangkan untuk melakukan pemrosesan big data. Berikut produk utama yang dikembangkan dalam Hadoop.

1. Hadoop Common

Hadoop Common adalah library-library umum yang mendukung library lainnya untuk dapat digunakan. Ini terkait perintah-perintah dasar yang ada pada Hadoop.

2. Hadoop Distributed File System (HDFS™)

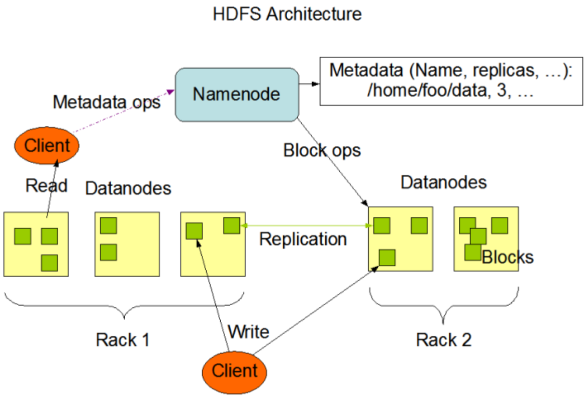
Berbeda dengan system file data pada umumnya yaitu FAT32 dan NTFS yang dapat menyimpan 1 file data berkisaran antara 4 GB hingga 16 TB. HDFS adalah format sistem file yang dapat menampung 1 file data yang sangat besar dengan mengecilkan cluster sekelompok host data storage.

3. Hadoop YARN

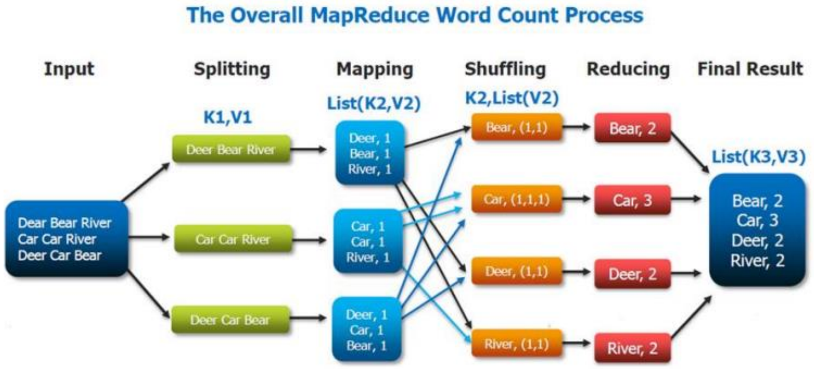
Hadoop YARN adalah framework yang digunakan untuk mengatur pekerjaan secara terjadwal (schedule) dan manajemen cluster data.

4. Hadoop MapReduce

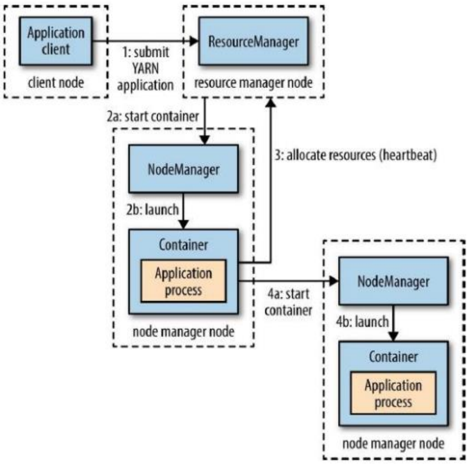
Hadoop MapReduce adalah paradigma pemrosesan data yang mengambil spesifikasi big data untuk menentukan bagaimana data tersebut dijadikan input dan output untuk diterapkan. MapReduce terintegrasi erat dengan HDFS untuk menyimpan data yang diperlukan.



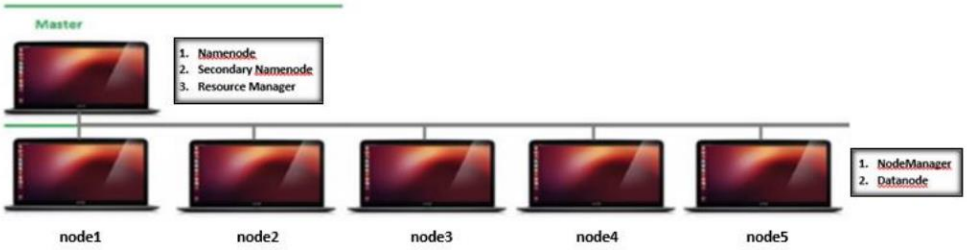
Gambar 1. Arsitektur HDFS



Gambar 2. Cara kerja pemrograman pada MapReduce



Gambar 3. Cara kerja YARN pada sebuah aplikasi



Gambar 4. Teknologi Hadoop Multinode

C. Tugas Pendahuluan

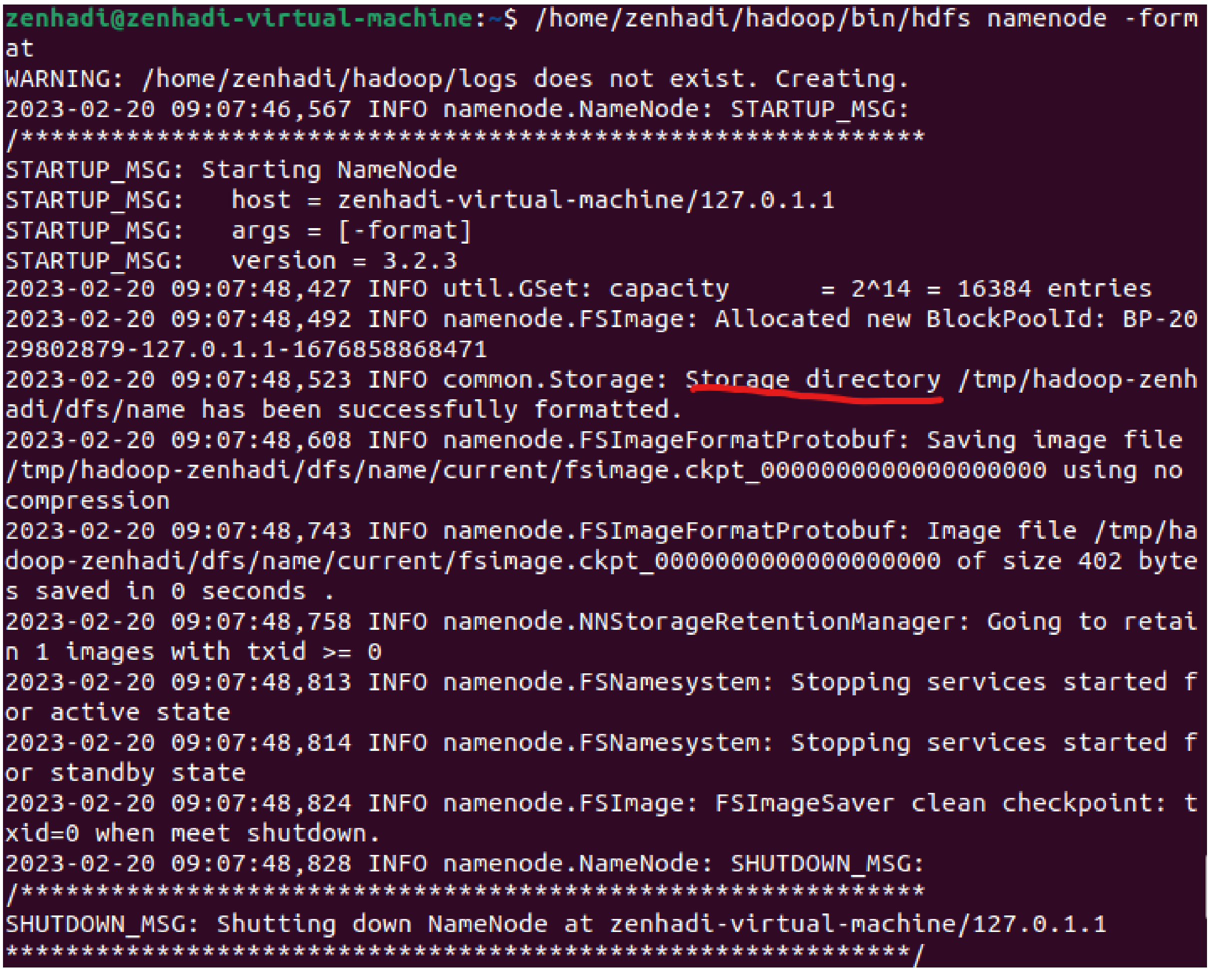
Pelajari konsep Hadoop dengan baik.

D. Percobaan

D.1. Konfigurasi HDFS

1. Jalankan hadoop

```
$ /home/zenhadi/hadoop/bin/hdfs namenode -format
```

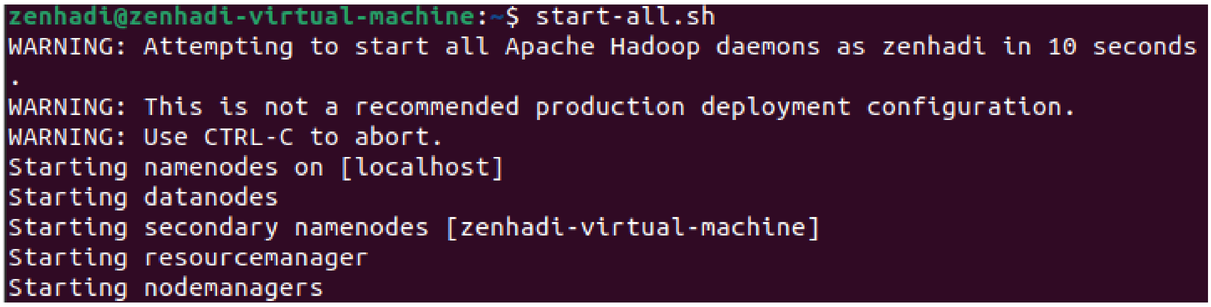


```
zenhadi@zenhadi-virtual-machine:~$ /home/zenhadi/hadoop/bin/hdfs namenode -form
at
WARNING: /home/zenhadi/hadoop/logs does not exist. Creating.
2023-02-20 09:07:46,567 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = zenhadi-virtual-machine/127.0.1.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 3.2.3
2023-02-20 09:07:48,427 INFO util.GSet: capacity      = 2^14 = 16384 entries
2023-02-20 09:07:48,492 INFO namenode.FSImage: Allocated new BlockPoolId: BP-20
29802879-127.0.1.1-1676858868471
2023-02-20 09:07:48,523 INFO common.Storage: Storage directory /tmp/hadoop-zenh
adi/dfs/name has been successfully formatted.
2023-02-20 09:07:48,608 INFO namenode.FSImageFormatProtobuf: Saving image file
/tmp/hadoop-zenhadi/dfs/name/current/fsimage.ckpt_0000000000000000000 using no
compression
2023-02-20 09:07:48,743 INFO namenode.FSImageFormatProtobuf: Image file /tmp/ha
doop-zenhadi/dfs/name/current/fsimage.ckpt_0000000000000000000 of size 402 byte
s saved in 0 seconds .
2023-02-20 09:07:48,758 INFO namenode.NNStorageRetentionManager: Going to retai
n 1 images with txid >= 0
2023-02-20 09:07:48,813 INFO namenode.FSNamesystem: Stopping services started f
or active state
2023-02-20 09:07:48,814 INFO namenode.FSNamesystem: Stopping services started f
or standby state
2023-02-20 09:07:48,824 INFO namenode.FSImage: FSImageSaver clean checkpoint: t
xid=0 when meet shutdown.
2023-02-20 09:07:48,828 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at zenhadi-virtual-machine/127.0.1.1
************************************************************/
```

Simpan file di /home/hduser/hadoop_dir/namenode-dir/

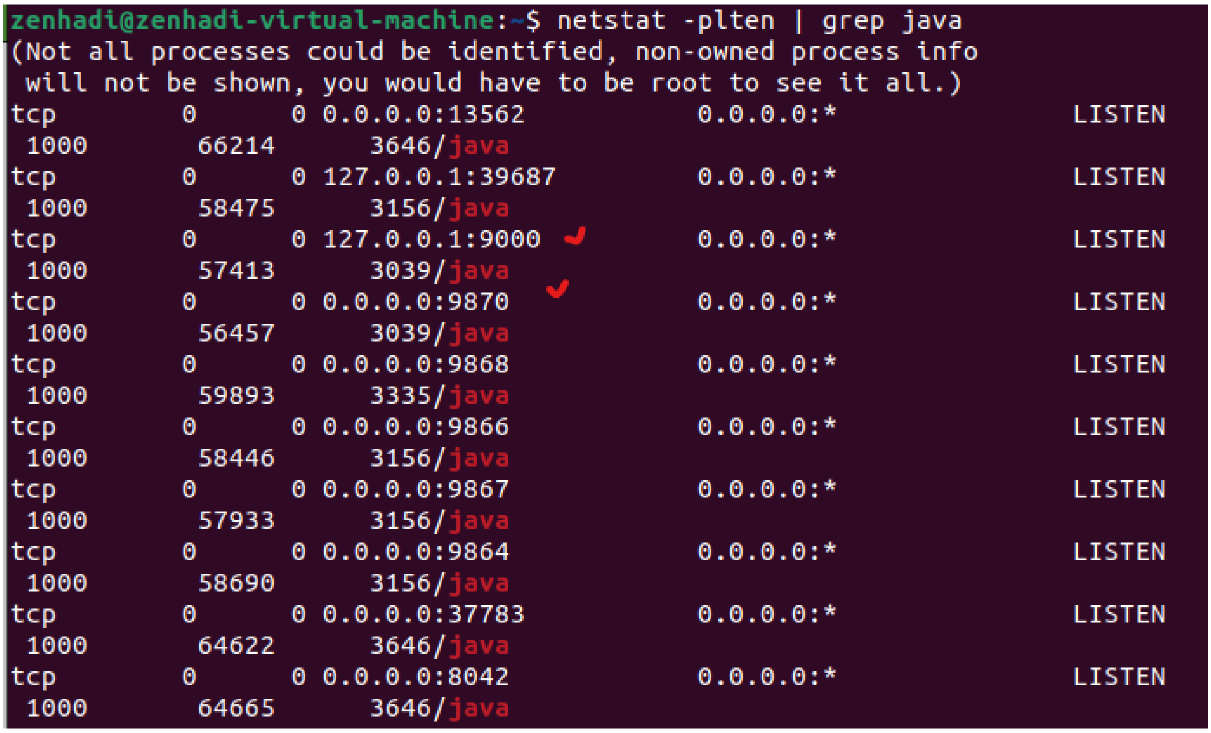
2. Mulai Hadoop

$ start-all.sh



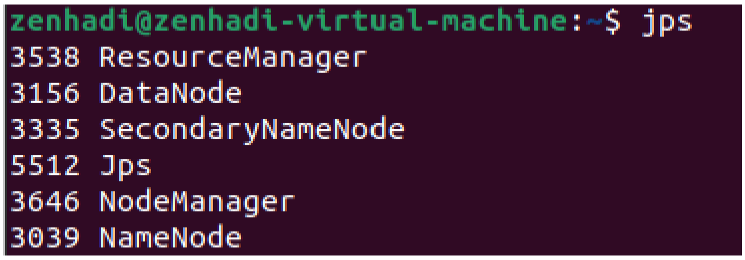
```
zenhadi@zenhadi-virtual-machine:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as zenhadi in 10 seconds
.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [zenhadi-virtual-machine]
Starting resourcemanager
Starting nodemanagers
```

Untuk memverifikasi bahwa daemon namenode dan datanode berjalan, jalankan perintah diatas di terminal. Ini menampilkan proses Java yang sedang berjalan pada sistem.



```
zenhadi@zenhadi-virtual-machine:~$ netstat -plten | grep java
(Not all processes could be identified, non-owned process info
 will not be shown, you would have to be root to see it all.)
tcp        0      0 0.0.0.0:13562           0.0.0.0:*               LISTEN
 1000       66214      3646/java
tcp        0      0 127.0.0.1:39687         0.0.0.0:*               LISTEN
 1000       58475      3156/java
tcp        0      0 127.0.0.1:9000          0.0.0.0:*               LISTEN
 1000       57413      3039/java
tcp        0      0 0.0.0.0:9870            0.0.0.0:*               LISTEN
 1000       56457      3039/java
tcp        0      0 0.0.0.0:9868            0.0.0.0:*               LISTEN
 1000       59893      3335/java
tcp        0      0 0.0.0.0:9866            0.0.0.0:*               LISTEN
 1000       58446      3156/java
tcp        0      0 0.0.0.0:9867            0.0.0.0:*               LISTEN
 1000       57933      3156/java
tcp        0      0 0.0.0.0:9864            0.0.0.0:*               LISTEN
 1000       58690      3156/java
tcp        0      0 0.0.0.0:37783           0.0.0.0:*               LISTEN
 1000       64622      3646/java
tcp        0      0 0.0.0.0:8042            0.0.0.0:*               LISTEN
 1000       64665      3646/java
```

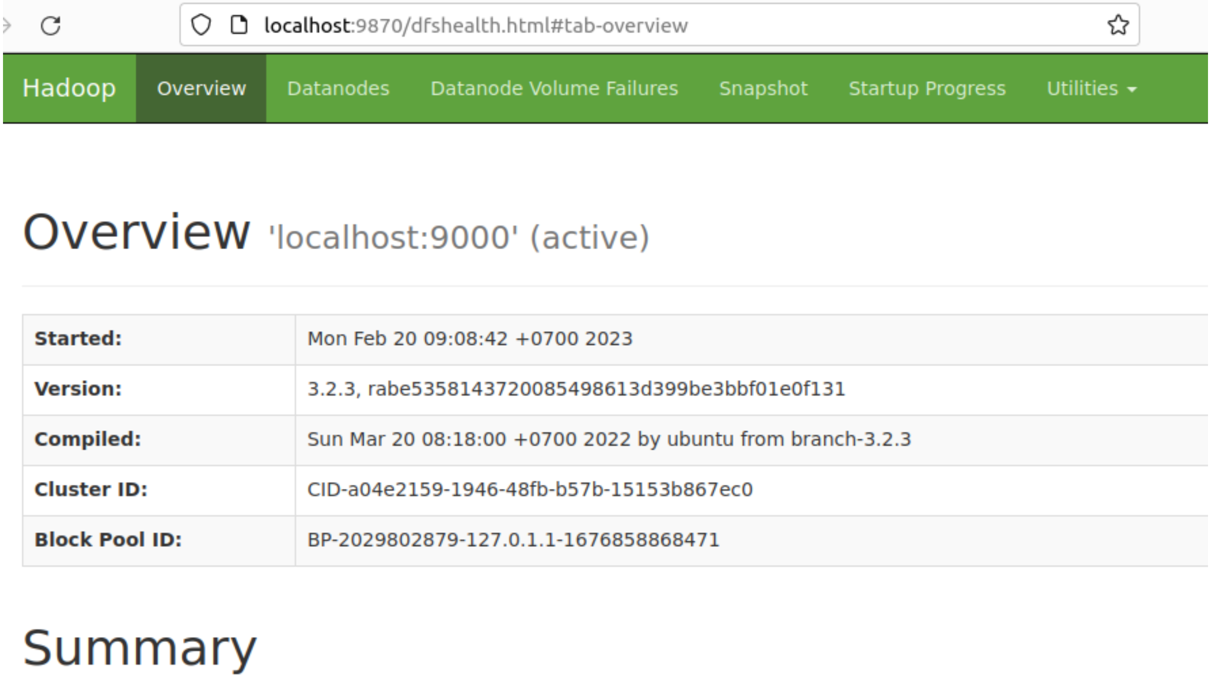
JPS: Java Virtual Machine Process Status



```
zenhadi@zenhadi-virtual-machine:~$ jps
3538 ResourceManager
3156 DataNode
3335 SecondaryNameNode
5512 Jps
3646 NodeManager
3039 NameNode
```

Terlihat bahwa datanode dan namenode terletak di server yang sama saat diaplikasikan pada single node Hadoop. Saat berjalan di cluster, namenode tidak mengandung datanode. Jika namenode atau datanode belum berjalan, lihat file log selama start-dfs.sh berjalan.

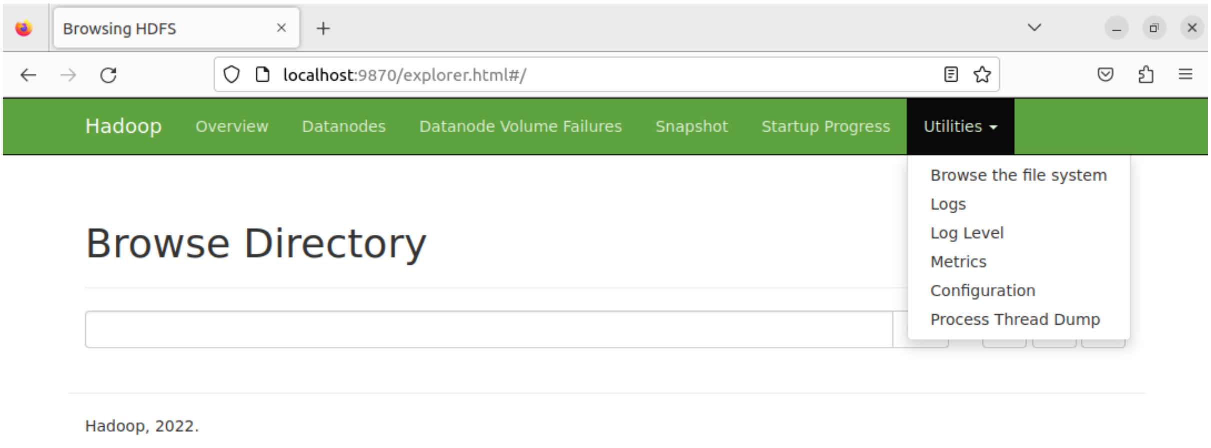
3. Jalankan Hadoop pada browser



Overview 'localhost:9000' (active)

Started:	Mon Feb 20 09:08:42 +0700 2023
Version:	3.2.3, rabe5358143720085498613d399be3bbf01e0f131
Compiled:	Sun Mar 20 08:18:00 +0700 2022 by ubuntu from branch-3.2.3
Cluster ID:	CID-a04e2159-1946-48fb-b57b-15153b867ec0
Block Pool ID:	BP-2029802879-127.0.1.1-1676858868471

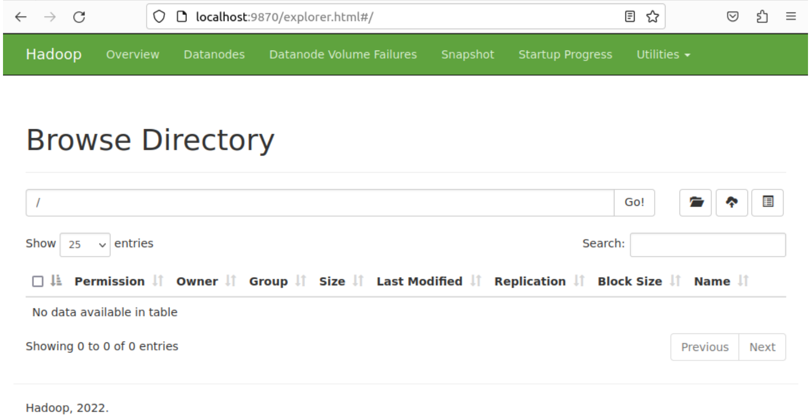
Summary



Browse Directory

Hadoop, 2022.

Klik bagian **Browse the file system**, hasilnya akan terlihat seperti berikut:



Browse Directory

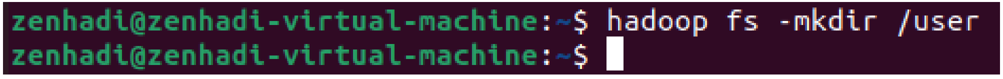
/

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
No data available in table							

Showing 0 to 0 of 0 entries

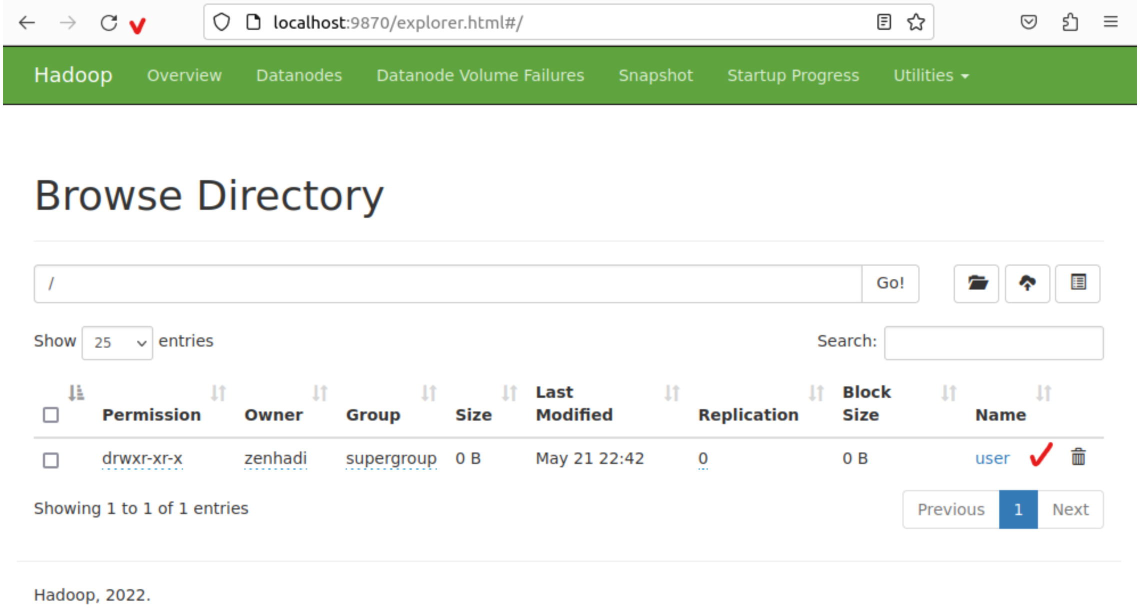
Hadoop, 2022.

4. Buat folder baru
   $hadoop fs -mkdir /user



```
zenhadi@zenhadi-virtual-machine:~$ hadoop fs -mkdir /user
zenhadi@zenhadi-virtual-machine:~$ 
```

5. Pastikan folder user telah terbentuk



Browse Directory

/ Go!

Show 25 entries Search:

☐	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
☐	drwxr-xr-x	zenhadi	supergroup	0 B	May 21 22:42	0	0 B	user ✓

Showing 1 to 1 of 1 entries

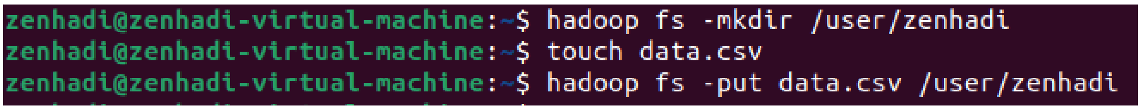
Hadoop, 2022.

6. Tambahkan folder dan file yang baru di dalam folder **user**

a. Buat folder baru: $ hadoop fs -mkdir /user/zenhadi

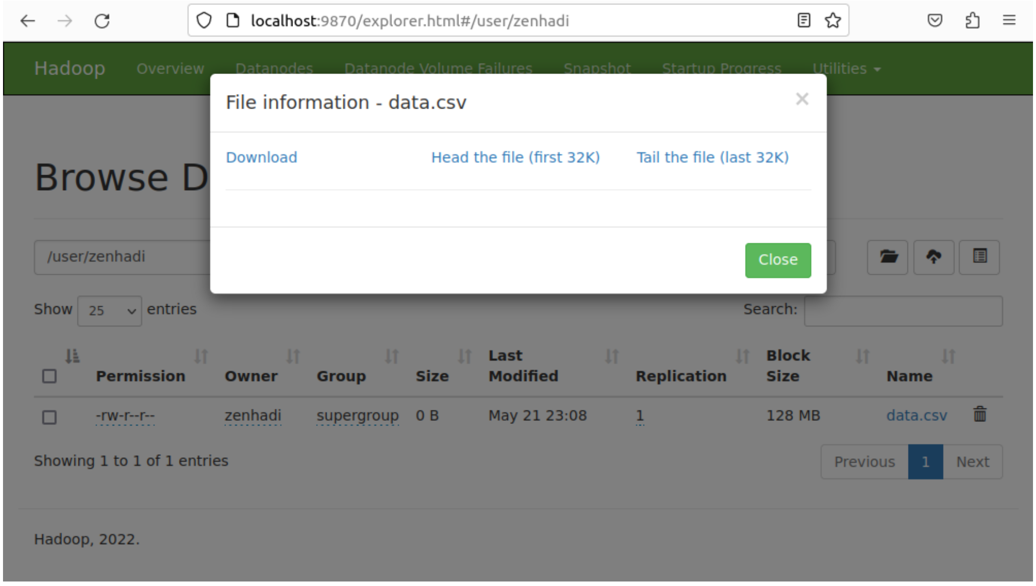
b. Buat file baru: $ touch data.csv

c. Masukkan file baru ke folder /user/zenhadi:
   $hadoop fs -put data.csv /user/zenhadi



```
zenhadi@zenhadi-virtual-machine:~$ hadoop fs -mkdir /user/zenhadi
zenhadi@zenhadi-virtual-machine:~$ touch data.csv
zenhadi@zenhadi-virtual-machine:~$ hadoop fs -put data.csv /user/zenhadi
```

d. Refresh kembali browser



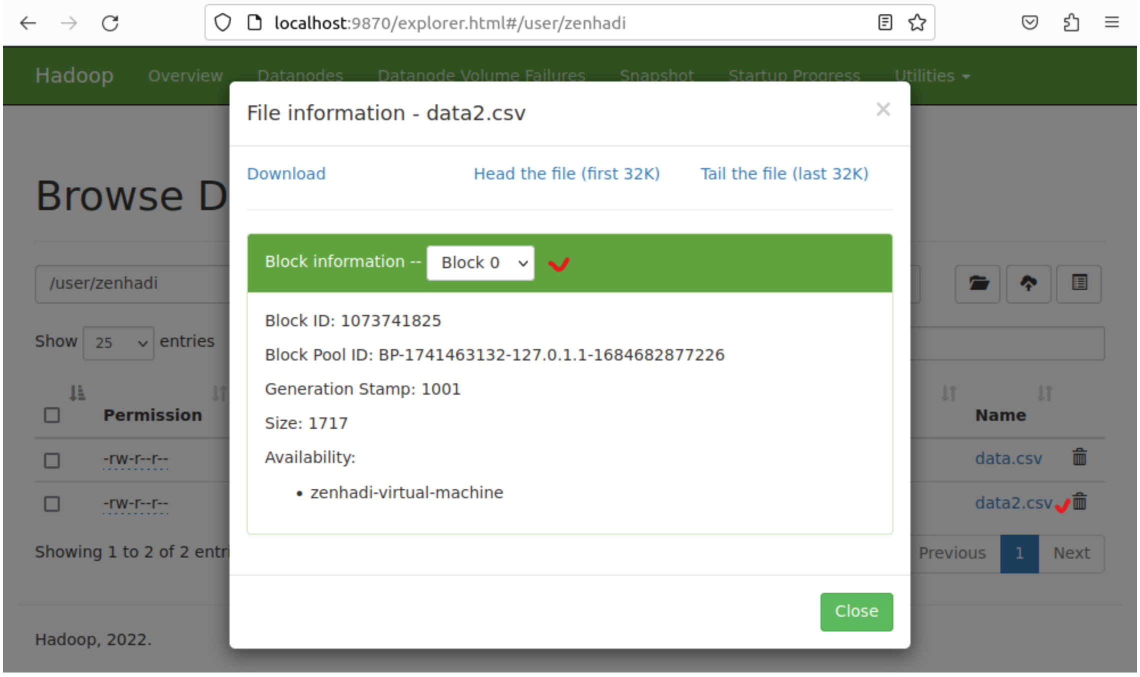
File information - data.csv

/user/zenhadi

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
□	-rw-r--r--	zenhadi	supergroup	0 B	May 21 23:08	1	128 MB	data.csv

Showing 1 to 1 of 1 entries

e. Buatlah file yang berisi sebuah data, simpan di /user/zenhadi



File information - data2.csv

/user/zenhadi

Block information -- Block 0 ✓

Block ID: 1073741825

Block Pool ID: BP-1741463132-127.0.1.1-1684682877226

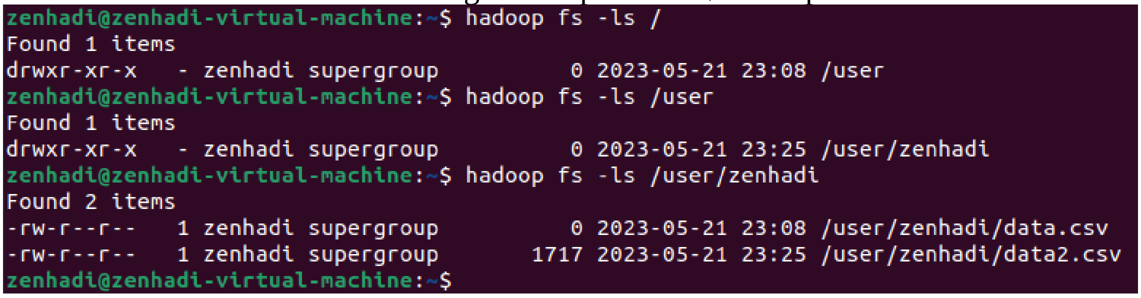
Generation Stamp: 1001

Size: 1717

Availability:

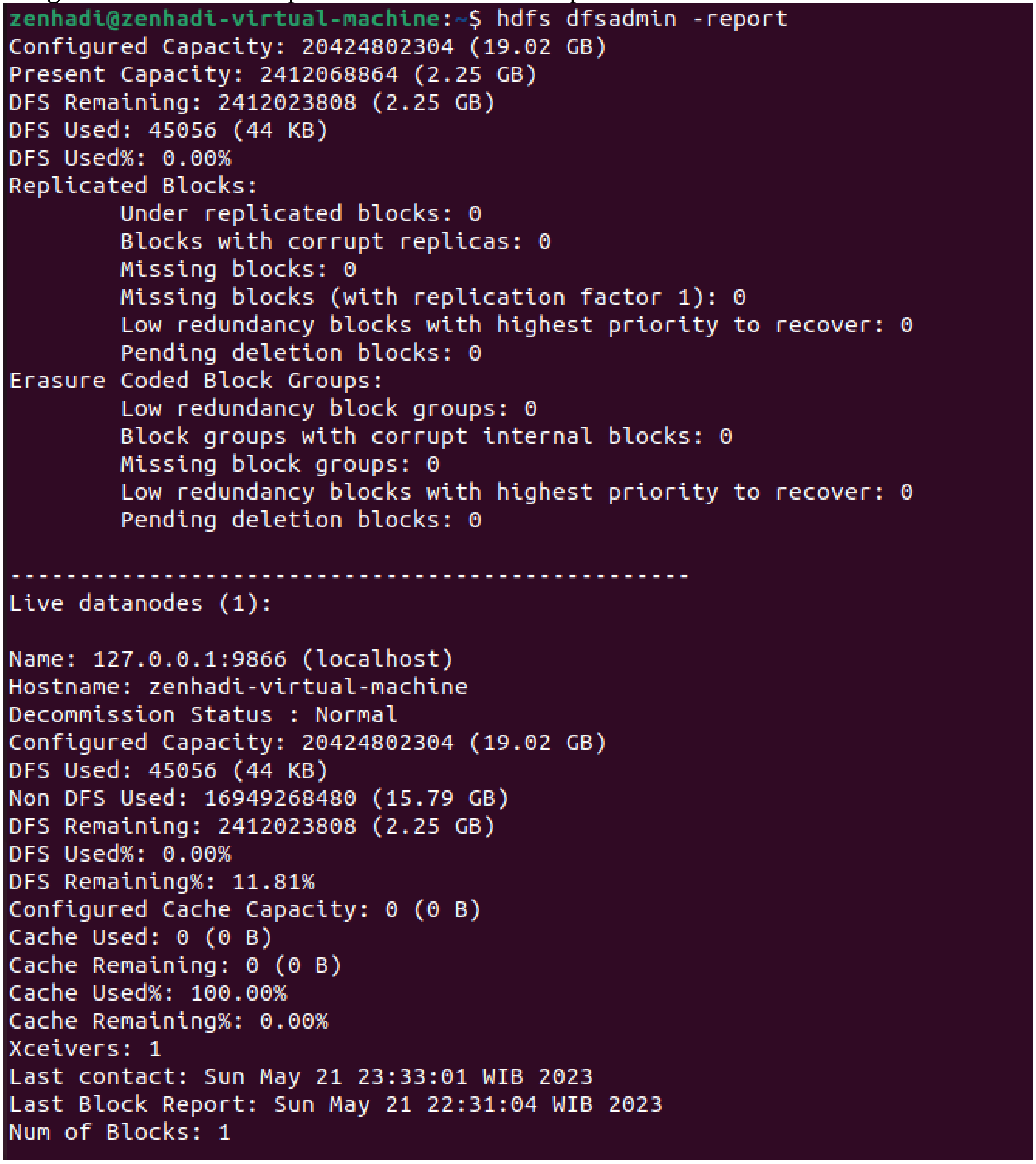
- zenhadi-virtual-machine

f. Untuk melihat dari terminal gunakan perintah: $ hadoop fs -ls



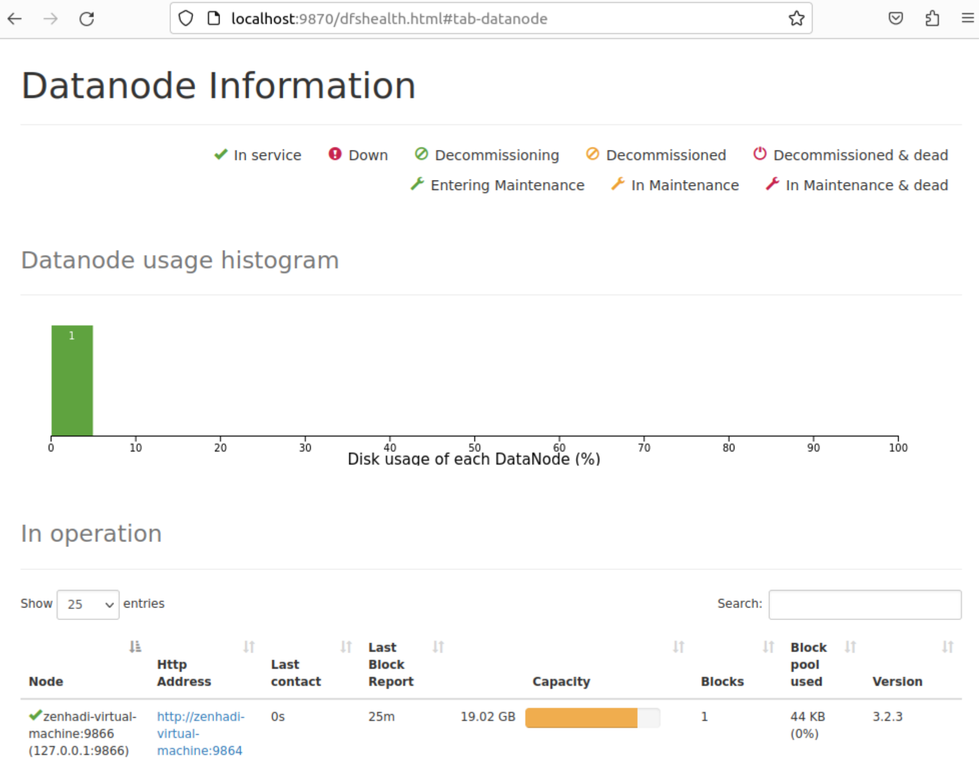
```
zenhadi@zenhadi-virtual-machine:~$ hadoop fs -ls /
Found 1 items
drwxr-xr-x   - zenhadi supergroup          0 2023-05-21 23:08 /user
zenhadi@zenhadi-virtual-machine:~$ hadoop fs -ls /user
Found 1 items
drwxr-xr-x   - zenhadi supergroup          0 2023-05-21 23:25 /user/zenhadi
zenhadi@zenhadi-virtual-machine:~$ hadoop fs -ls /user/zenhadi
Found 2 items
-rw-r--r--   1 zenhadi supergroup          0 2023-05-21 23:08 /user/zenhadi/data.csv
-rw-r--r--   1 zenhadi supergroup       1717 2023-05-21 23:25 /user/zenhadi/data2.csv
zenhadi@zenhadi-virtual-machine:~$
```

g. Untuk melihat report: $hdfs dfsadmin -report



```
zenhadi@zenhadi-virtual-machine:~$ hdfs dfsadmin -report
Configured Capacity: 20424802304 (19.02 GB)
Present Capacity: 2412068864 (2.25 GB)
DFS Remaining: 2412023808 (2.25 GB)
DFS Used: 45056 (44 KB)
DFS Used%: 0.00%
Replicated Blocks:
        Under replicated blocks: 0
        Blocks with corrupt replicas: 0
        Missing blocks: 0
        Missing blocks (with replication factor 1): 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0
Erasure Coded Block Groups:
        Low redundancy block groups: 0
        Block groups with corrupt internal blocks: 0
        Missing block groups: 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0

-------------------------------------------------
Live datanodes (1):

Name: 127.0.0.1:9866 (localhost)
Hostname: zenhadi-virtual-machine
Decommission Status : Normal
Configured Capacity: 20424802304 (19.02 GB)
DFS Used: 45056 (44 KB)
Non DFS Used: 16949268480 (15.79 GB)
DFS Remaining: 2412023808 (2.25 GB)
DFS Used%: 0.00%
DFS Remaining%: 11.81%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Sun May 21 23:33:01 WIB 2023
Last Block Report: Sun May 21 22:31:04 WIB 2023
Num of Blocks: 1
```

Informasi ini sama dengan di browser menu **Overview** dan **Datanodes**.



localhost:9870/dfshealth.html#tab-datanode

Datanode Information

✔ In service ● Down ⊘ Decommissioning ⊘ Decommissioned ⏻ Decommissioned & dead
Entering Maintenance In Maintenance In Maintenance & dead

Datanode usage histogram

1
0
10
20
30
40
50
60
70
80
90
100
Disk usage of each DataNode (%)

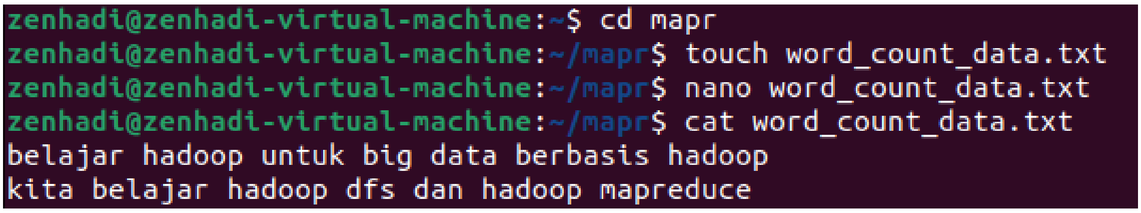
In operation

Show 25 entries Search:

Node	Http Address	Last contact	Last Block Report	Capacity	Blocks	Block pool used	Version
✔zenhadi-virtual-machine:9866 (127.0.0.1:9866)	http://zenhadi-virtual-machine:9864	0s	25m	19.02 GB	1	44 KB (0%)	3.2.3

D.2. MAPREDUCE

1. Buat direktori mapr untuk menyimpan semua file yang diperlukan:
   $ mkdir mapr
2. Buat file teks:
   $ touch word_count_data.txt
3. Gunakan nano atau gedit untuk memasukkan data di file tersebut seperti dalam contoh berikut.

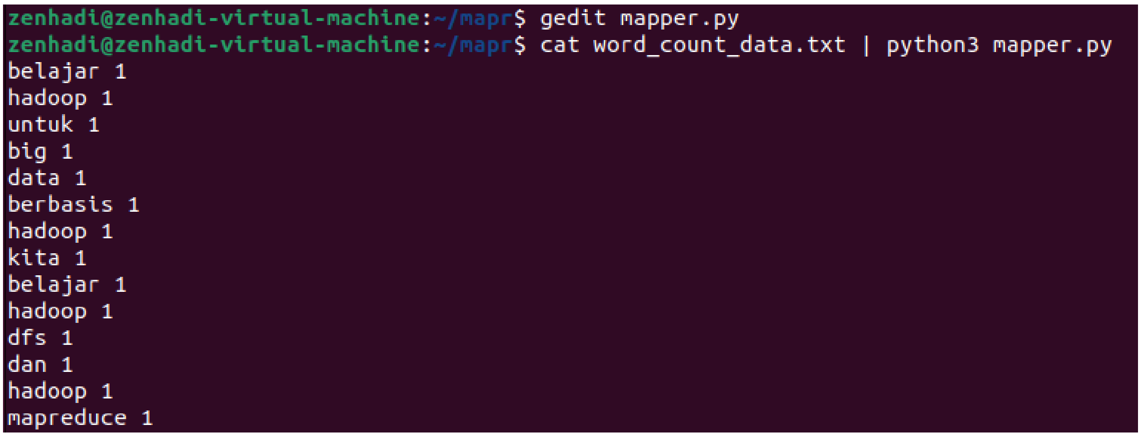


```
zenhadi@zenhadi-virtual-machine:~$ cd mapr
zenhadi@zenhadi-virtual-machine:~/mapr$ touch word_count_data.txt
zenhadi@zenhadi-virtual-machine:~/mapr$ nano word_count_data.txt
zenhadi@zenhadi-virtual-machine:~/mapr$ cat word_count_data.txt
belajar hadoop untuk big data berbasis hadoop
kita belajar hadoop dfs dan hadoop mapreduce
```

4. Buat file python mapper.py

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

# import sys because we need to read and write data to STDIN and
STDOUT
import sys

# reading entire line from STDIN (standard input)
for line in sys.stdin:
        # to remove leading and trailing whitespace
        line = line.strip()
        # split the line into words
        words = line.split()

        # we are looping over the words array and printing the word
        # with the count of 1 to the STDOUT
        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
                print (word, 1)
```

5. Jalankan file python mapper.py dengan input dari file teks:
   $ cat word_count_data.txt | python3 mapper.py

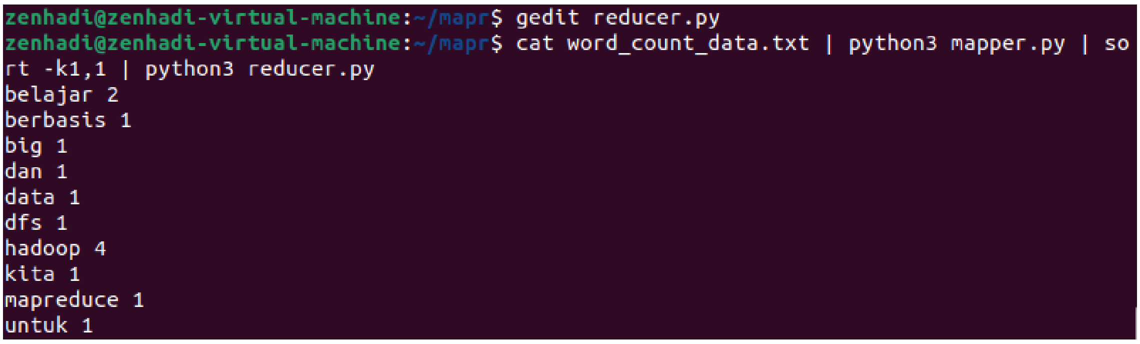


```
zenhadi@zenhadi-virtual-machine:~/mapr$ gedit mapper.py
zenhadi@zenhadi-virtual-machine:~/mapr$ cat word_count_data.txt | python3 mapper.py
belajar 1
hadoop 1
untuk 1
big 1
data 1
berbasis 1
hadoop 1
kita 1
belajar 1
hadoop 1
dfs 1
dan 1
hadoop 1
mapreduce 1
```

6. Buat file reducer.py

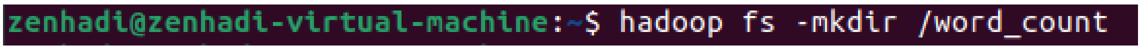
```
#!/usr/bin/env python

from operator import itemgetter
import sys

current_word = None
current_count = 0
word = None

# read the entire line from STDIN
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()
    # splitting the data on the basis of tab we have provided in mapper.py
    word, count = line.split(' ', 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 (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 (current_word, current_count)
```

7. Jalankan file python reducer.py dengan input dari file teks:
   $cat word_count_data.txt | python3 mapper.py | sort -k1,1 | ptyhon3 reducer.py



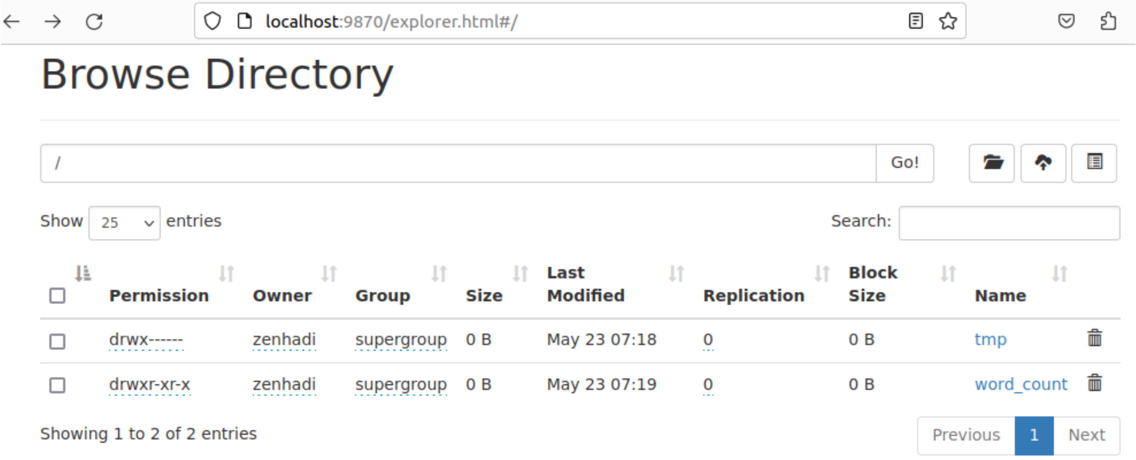
```
zenhadi@zenhadi-virtual-machine:~/mapr$ gedit reducer.py
zenhadi@zenhadi-virtual-machine:~/mapr$ cat word_count_data.txt | python3 mapper.py | so
rt -k1,1 | python3 reducer.py
belajar 2
berbasis 1
big 1
dan 1
data 1
dfs 1
hadoop 4
kita 1
mapreduce 1
untuk 1
```

8. Buat direktori di hadoop
   $hadoop fs -mkdir /word_count



```
zenhadi@zenhadi-virtual-machine:~$ hadoop fs -mkdir /word_count
```

9. Cek hasilnya di web browser:
   http://localhost:9870/



Browse Directory

/

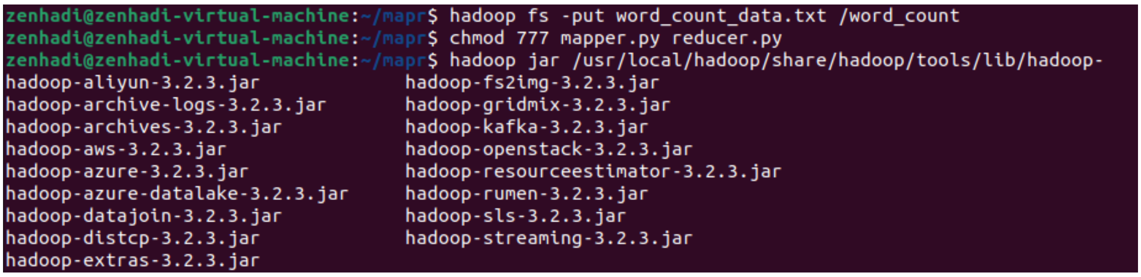
Show 25 entries

Search:

☐	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
☐	drwx------	zenhadi	supergroup	0 B	May 23 07:18	0	0 B	tmp
☐	drwxr-xr-x	zenhadi	supergroup	0 B	May 23 07:19	0	0 B	word_count

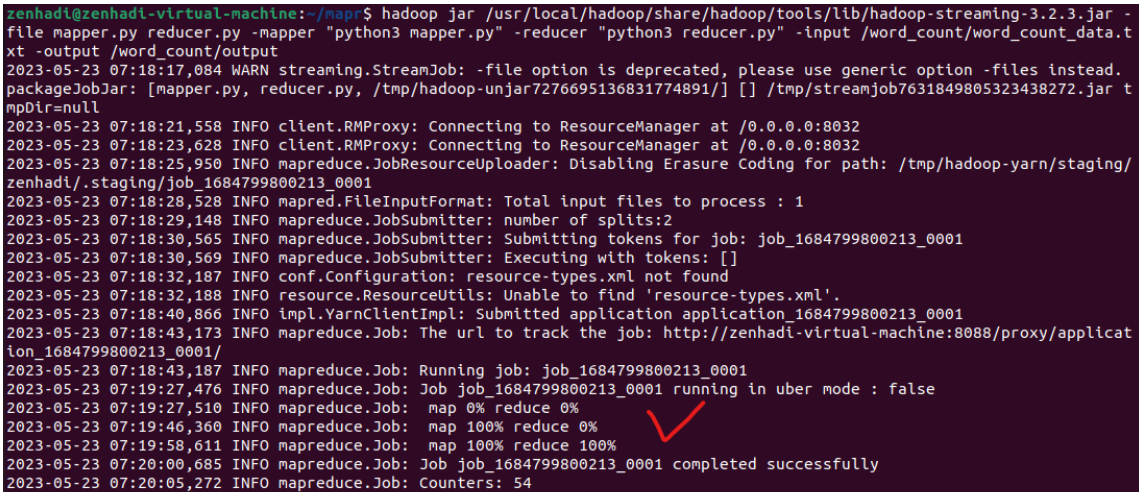
Showing 1 to 2 of 2 entries

10. Lakukan langkah berikut:
    a. Kirim file word_count_data.txt ke hadoop
       $ hadoop fs -put word_count_data.txt /word_count
    b. Rubah mode file mapper.py dan reducer.py
       $ chmod 777 mapper.py reducer.py

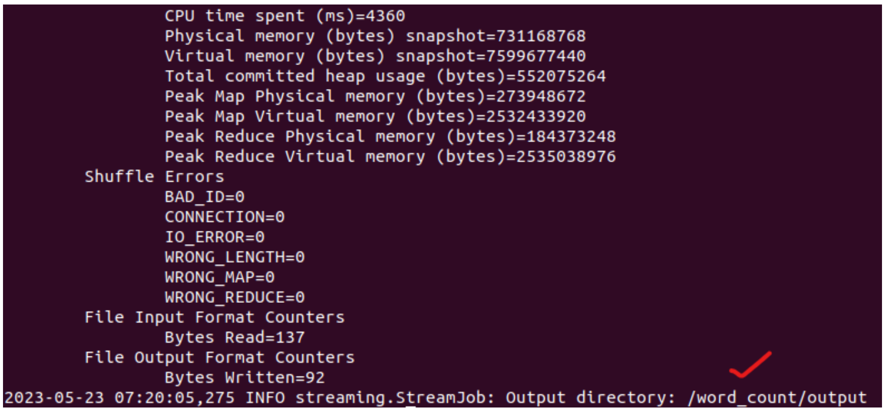


```
zenhadi@zenhadi-virtual-machine:~/mapr$ hadoop fs -put word_count_data.txt /word_count
zenhadi@zenhadi-virtual-machine:~/mapr$ chmod 777 mapper.py reducer.py
zenhadi@zenhadi-virtual-machine:~/mapr$ hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-
hadoop-aliyun-3.2.3.jar              hadoop-fs2img-3.2.3.jar
hadoop-archive-logs-3.2.3.jar        hadoop-gridmix-3.2.3.jar
hadoop-archives-3.2.3.jar            hadoop-kafka-3.2.3.jar
hadoop-aws-3.2.3.jar                 hadoop-openstack-3.2.3.jar
hadoop-azure-3.2.3.jar               hadoop-resourceestimator-3.2.3.jar
hadoop-azure-datalake-3.2.3.jar      hadoop-rumen-3.2.3.jar
hadoop-datajoin-3.2.3.jar            hadoop-sls-3.2.3.jar
hadoop-distcp-3.2.3.jar              hadoop-streaming-3.2.3.jar
hadoop-extras-3.2.3.jar
```

11. Jalan mapreduce di hadoop dengan perintah berikut:
    $ hadoop jar
    /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.2.3.jar
    -file mapper.py reducer.py
    -mapper “python3 mapper.py”
    -reducer “python3 reducer.py”
    -input /word_count/word_count_data.txt
    -output /word_count/output



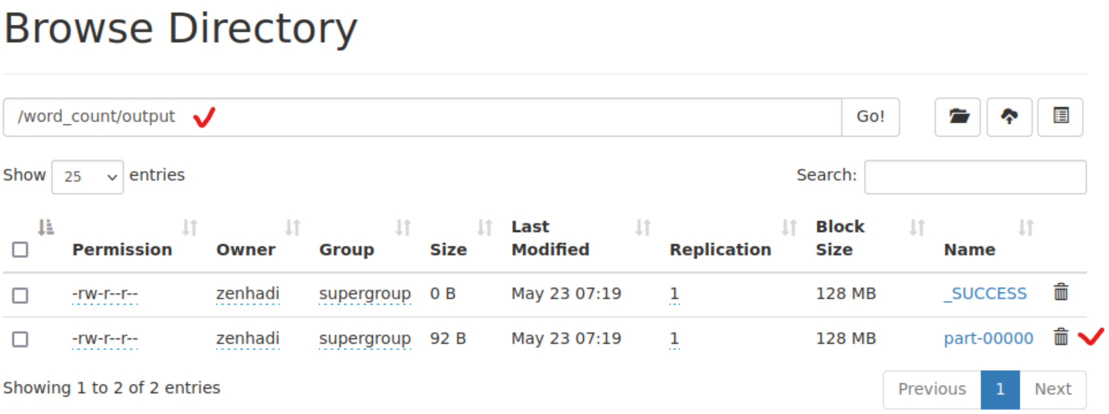
```
zenhadi@zenhadi-virtual-machine:~/mapr$ hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.2.3.jar -
file mapper.py reducer.py -mapper "python3 mapper.py" -reducer "python3 reducer.py" -input /word_count/word_count_data.t
xt -output /word_count/output
2023-05-23 07:18:17,084 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [mapper.py, reducer.py, /tmp/hadoop-unjar7276695136831774891/] [] /tmp/streamjob7631849805323438272.jar t
mpDir=null
2023-05-23 07:18:21,558 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2023-05-23 07:18:23,628 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2023-05-23 07:18:25,950 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/
zenhadi/.staging/job_1684799800213_0001
2023-05-23 07:18:28,528 INFO mapred.FileInputFormat: Total input files to process : 1
2023-05-23 07:18:29,148 INFO mapreduce.JobSubmitter: number of splits:2
2023-05-23 07:18:30,565 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1684799800213_0001
2023-05-23 07:18:30,569 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-05-23 07:18:32,187 INFO conf.Configuration: resource-types.xml not found
2023-05-23 07:18:32,188 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-05-23 07:18:40,866 INFO impl.YarnClientImpl: Submitted application application_1684799800213_0001
2023-05-23 07:18:43,173 INFO mapreduce.Job: The url to track the job: http://zenhadi-virtual-machine:8088/proxy/applicat
ion_1684799800213_0001/
2023-05-23 07:18:43,187 INFO mapreduce.Job: Running job: job_1684799800213_0001
2023-05-23 07:19:27,476 INFO mapreduce.Job: Job job_1684799800213_0001 running in uber mode : false
2023-05-23 07:19:27,510 INFO mapreduce.Job:  map 0% reduce 0%
2023-05-23 07:19:46,360 INFO mapreduce.Job:  map 100% reduce 0%
2023-05-23 07:19:58,611 INFO mapreduce.Job:  map 100% reduce 100%
2023-05-23 07:20:00,685 INFO mapreduce.Job: Job job_1684799800213_0001 completed successfully
2023-05-23 07:20:05,272 INFO mapreduce.Job: Counters: 54
```



```
                CPU time spent (ms)=4360
                Physical memory (bytes) snapshot=731168768
                Virtual memory (bytes) snapshot=7599677440
                Total committed heap usage (bytes)=552075264
                Peak Map Physical memory (bytes)=273948672
                Peak Map Virtual memory (bytes)=2532433920
                Peak Reduce Physical memory (bytes)=184373248
                Peak Reduce Virtual memory (bytes)=2535038976
        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=137
        File Output Format Counters
                Bytes Written=92
2023-05-23 07:20:05,275 INFO streaming.StreamJob: Output directory: /word_count/output
```

Amati proses yang berhasil dijalankan.

12. Amati proses yang di browser di direktori: /word_count/output
    a. Klik pada file: part-00000
    b. Klik pada bagian: Head the file (first 32K)
    c. Hasil akan muncul di bagian bawah.



Browse Directory

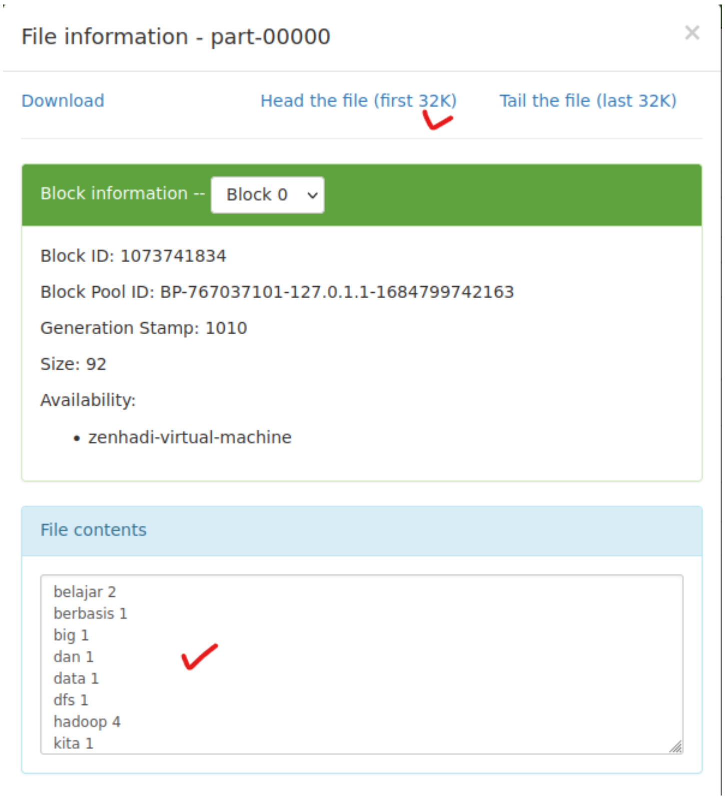
/word_count/output ✓ Go!

Show 25 entries Search:

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
□	-rw-r--r--	zenhadi	supergroup	0 B	May 23 07:19	1	128 MB	_SUCCESS
□	-rw-r--r--	zenhadi	supergroup	92 B	May 23 07:19	1	128 MB	part-00000 ✓

Showing 1 to 2 of 2 entries

Previous 1 Next



File information - part-00000

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

Block information -- Block 0

Block ID: 1073741834

Block Pool ID: BP-767037101-127.0.1.1-1684799742163

Generation Stamp: 1010

Size: 92

Availability:

- zenhadi-virtual-machine

File contents

```
belajar 2
berbasis 1
big 1
dan 1
data 1
dfs 1
hadoop 4
kita 1
```

E. Laporan Resmi :

1. Analisalah semua langkah-langkah instalasi diatas dan buat kesimpulan.