Assignment 2: Load large datasets into HDFS and analyze block distribution.

Objective: Understand how Apache Hadoop HDFS stores large files by splitting them into fixed-size blocks and placing them on DataNodes.

2.1 Step 1: Collect large Dataset (CSV/TXT)

```
su - hadoop
start-dfs.sh
start-yarn.sh
jps
```

Figure 23: Start Hadoop Services and Verify with jps

2.2 Step 2: Create HDFS directory.

```
% Created 352 MB csv file using:excel and uploaded to hdfs:
% Created exceldata directory in hdfs:
hdfs dfs -mkdir /exceldata
ls -lh mydataset.csv
hdfs dfs -put mydataset.csv /exceldata
hdfs dfs -ls /exceldata
```

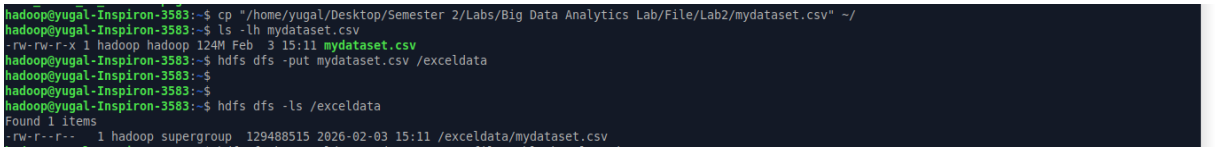

Figure 24: Created exceldata directory in hdfs and copied dataset

2.3 Step 3: Upload large dataset into HDFS

```
hdfs dfs -put mydataset.csv /exceldata
```

Figure 25: Upload large dataset into HDFS

2.4 Step 4: Verify data storage

`hdfs dfs -ls /exceldata`

```
hadoop@yugal-Inspiron-3583:~$ hdfs dfs -ls /exceldata
Found 1 items
-rw-r--r--   1 hadoop supergroup  129488515 2026-02-03 15:11 /exceldata/mydataset.csv
```

Figure 26: Verify data storage in HDFS

2.5 Step 5: Analyze block distribution

`hdfs fsck /exceldata/mydataset.csv -files -blocks -locations`

```
hadoop@yugal-Inspiron-3583:~$ hdfs fsck /exceldata/mydataset.csv -files -blocks -locations
Connecting to namenode via http://localhost:9870/fsck?ugi=hadoop&files=1&blocks=1&locations=1&path=%2Fexceldata%2Fmydataset.csv
FSCK started by hadoop (auth:SIMPLE) from /127.0.0.1 for path /exceldata/mydataset.csv at Tue Feb 03 15:18:34 IST 2026

/exceldata/mydataset.csv 369487717 bytes, replicated: replication=1, 3 block(s):  OK
0. BP-1685072063-127.0.1.1-1769537156943:blk_1073741830_1006 len=134217728 Live_repl=1  [DatanodeInfoWithStorage[127.0.0.1:9866,DS-47acaae1-e863-462a-9aec-290782a597df,
DISK]]
1. BP-1685072063-127.0.1.1-1769537156943:blk_1073741831_1007 len=134217728 Live_repl=1  [DatanodeInfoWithStorage[127.0.0.1:9866,DS-47acaae1-e863-462a-9aec-290782a597df,
DISK]]
2. BP-1685072063-127.0.1.1-1769537156943:blk_1073741832_1008 len=101052261 Live_repl=1  [DatanodeInfoWithStorage[127.0.0.1:9866,DS-47acaae1-e863-462a-9aec-290782a597df,
DISK]]


Status: HEALTHY
 Number of data-nodes:  1
 Number of racks:               1
 Total dirs:                    0
 Total symlinks:                0

Replicated Blocks:
 Total size:    369487717 B
 Total files:   1
 Total blocks (validated):      3 (avg. block size 123162572 B)
 Minimally replicated blocks:   3 (100.0 %)
 Over-replicated blocks:        0 (0.0 %)
 Under-replicated blocks:       0 (0.0 %)
 Mis-replicated blocks:         0 (0.0 %)
 Default replication factor:    1
 Average block replication:     1.0
 Missing blocks:                0
 Corrupt blocks:                0
 Missing replicas:              0 (0.0 %)
 Blocks queued for replication: 0

Erasure Coded Block Groups:
 Total size:    0 B
 Total files:   0
 Total block groups (validated):        0
 Minimally erasure-coded block groups:  0
 Over-erasure-coded block groups:       0
```

Figure 27: Analyze block distribution in HDFS

2.6 Step 6: Download processed dataset from HDFS to local system

`hdfs dfs -get /exceldata/mydataset.csv mydataset_from_hdfs.csv`

```
hadoop@yugal-Inspiron-3583:~$ hdfs dfs -get /exceldata/mydataset.csv mydataset_from_hdfs.csv
```

Figure 28: Download processed dataset from HDFS to local system

2.7 Step 7: Delete data from HDFS

`hdfs dfs -rm -r /exceldata`

```
hadoop@yugal-Inspiron-3583:~$ hdfs dfs -get /exceldata/mydataset.csv mydataset_from_hdfs.csv
hdfs dfs -rm -r /exceldata
get: `mydataset_from_hdfs.csv': File exists
Deleted /exceldata
hadoop@yugal-Inspiron-3583:~$
```

Figure 29: Delete data from HDFS

2.8 Step 8: Check file block information

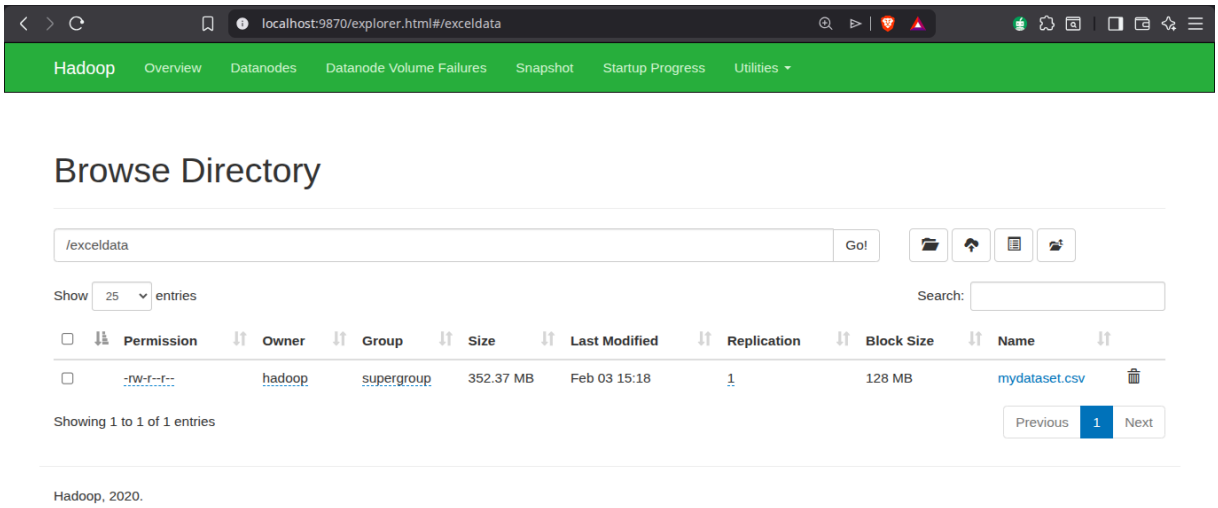

Figure 30: Check file block information in HDFS

2.9 Result

The large dataset (mydataset.csv) was successfully uploaded to HDFS under the /exceldata directory. The file was split into blocks and distributed across DataNodes, as verified by the block distribution analysis. The dataset was also downloaded back to the local system, confirming that data retrieval from HDFS works correctly. Finally, the dataset was deleted from HDFS, demonstrating proper data management and cleanup.

2.10 Conclusion

This assignment provided hands-on experience with Apache Hadoop HDFS, demonstrating how to upload large datasets, analyze block distribution, and manage data within the HDFS environment.