Assignment 6: Hadoop; hands-on activity I

1. Create Instance and Setup Hadoop

Login to the Jetstream VM using ssh (I have created a security group by creating **SSH Public Key** on Jetstream)

```
Last login: Sat Nov  9 20:13:05 on ttys009
(base) mukundkomati@Mukunds-MacBook-Pro ~ % ssh -i .ssh/msk_hadoop_id_rsa exouser@149.165.154.230
The authenticity of host '149.165.154.230 (149.165.154.230)' can't be established.
ED25519 key fingerprint is SHA256:1z7KFdzD7IAG0yeq/zNjfg68wkgFq8vhmq1tHr26xY0.
This key is not known by any other names.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added '149.165.154.230' (ED25519) to the list of known hosts.

 System information as of Sun Nov 10 04:53:06 UTC 2024

  System load:  1.0               Processes:             349
  Usage of /:   31.2% of 57.97GB  Users logged in:       0
  Memory usage: 6%                IPv4 address for ens3: 10.3.5.247
  Swap usage:   0%

=======================https://jetstream.status.io/=======================

Overall Jetstream2 Status:   Operational



=========================================================================

Last login: Sun Nov 10 04:40:11 2024
exouser@bda-hadoop:~$
```

Format the namenode

```
exouser@bda-hadoop:~$ ~/hadoop-3.4.0/bin/hadoop namenode -format
WARNING: Use of this script to execute namenode is deprecated.
WARNING: Attempting to execute replacement "hdfs namenode" instead.

2024-11-10 04:58:05,711 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
```

Output shows successfully formatted

```
2024-11-10 04:58:06,947 INFO common.Storage: Storage directory /tmp/hadoop-exouser/dfs/name has been successfully formatted.
```

Start all Hadoop services

```
exouser@bda-hadoop:~$ ~/hadoop-3.4.0/sbin/start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as exouser in 10 second
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
```

Ensure that all the services are up and running

```
exouser@bda-hadoop:~$ jps
129217 Bootstrap
173559 Jps
172532 ResourceManager
172772 NodeManager
171770 DataNode
171503 NameNode
172124 SecondaryNameNode
```

Check Cluster Health

Overview 'localhost:9000' (✔active)

Started:	Sun Nov 10 04:59:53 +0000 2024
Version:	3.4.0, rbd8b77f398f626bb7791783192ee7a5dfaeec760
Compiled:	Mon Mar 04 06:35:00 +0000 2024 by root from (HEAD detached at release-3.4.0-RC3)
Cluster ID:	CID-6ad7c867-f14f-4d29-b0fb-5400efa67217
Block Pool ID:	BP-671428153-10.3.5.247-1731214686918

Summary

Security is off.

Safemode is off.

1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).

Heap Memory used 339.88 MB of 825 MB Heap Memory. Max Heap Memory is 6.53 GB.

Non Heap Memory used 55.91 MB of 57.28 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	57.97 GB
Configured Remote Capacity:	0 B
DFS Used:	24 KB (0%)
Non DFS Used:	18.82 GB
DFS Remaining:	39.13 GB (67.5%)
Block Pool Used:	24 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	1 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion (including replicas)	0
Block Deletion Start Time	Sun Nov 10 04:59:53 +0000 2024

NameNode Journal Status

Current transaction ID: 1

Journal Manager	State
FileJournalManager(root=/tmp/hadoop-exouser/dfs/name)	EditLogFileOutputStream(/tmp/hadoop-exouser/dfs/name/current/edits_inprogress_0000000000000000001)

NameNode Storage

Storage Directory	Type	State
/tmp/hadoop-exouser/dfs/name	IMAGE_AND_EDITS	Active

DFS Storage Types

Storage Type	Configured Capacity	Capacity Used	Capacity Remaining	Block Pool Used	Nodes In Service
DISK	57.97 GB	24 KB (0%)	39.1 GB (67.46%)	24 KB	1

2. Working with HDFS Commands

Task1 : Create a directory on HDFS

```
exouser@bda-hadoop:~$ hdfs dfs -mkdir /assignment_data
exouser@bda-hadoop:~$ hdfs dfs -ls /
Found 1 items
drwxr-xr-x   - exouser supergroup          0 2024-11-10 05:12 /assignment_data
```

Task2 : Upload files to HDFS

```
(base) mukundkomati@Mukunds-MacBook-Pro hands_on_hadoop1 % scp file1.txt file2.csv
 exouser@149.165.154.230:~/
exouser@149.165.154.230's password:
file1.txt                                     100%  320    32.9KB/s   00:00
file2.csv                                     100%  479    ((((((((((((((((((((
```

Task 3: List files in HDFS

```
exouser@bda-hadoop:~$ hdfs dfs -put file1.txt /assignment_data/
exouser@bda-hadoop:~$ hdfs dfs -put file2.csv /assignment_data/
exouser@bda-hadoop:~$ hadoop fs -ls /assignment_data
Found 2 items
-rw-r--r--   1 exouser supergroup        320 2024-11-10 05:18 /assignment_data/file1.txt
-rw-r--r--   1 exouser supergroup        479 2024-11-10 05:18 /assignment_data/file2.csv
```

Task 4: View file content in HDFS

```
exouser@bda-hadoop:~$ hadoop fs -cat /assignment_data/file1.txt
Hadoop is a framework for distributed storage and processing.
MapReduce is a programming model for processing large data sets.
Hadoop Streaming allows us to use any programming language for mapper and reducer.
Python is commonly used with Hadoop for data processing.
Hadoop can handle large amounts of data efficiently.
exouser@bda-hadoop:~$ hadoop fs -cat /assignment_data/file2.csv
cement,blast_furnace_slag,fly_ash,water,superplasticizer,coarse_aggregate,fine_aggregate ,age,concrete_compressive_strength
540,0,0,162,2.5,1040,676,28,79.99
540,0,0,162,2.5,1055,676,28,61.89
332.5,142.5,0,228,0,932,594,270,40.27
332.5,142.5,0,228,0,932,594,365,41.05
198.6,132.4,0,192,0,978.4,825.5,360,44.3
266,114,0,228,0,932,670,90,47.03
380,95,0,228,0,932,594,365,43.7
380,95,0,228,0,932,594,28,36.45
266,114,0,228,0,932,670,28,45.85
475,0,0,228,0,932,594,28,39.29exouser@bda-hadoop:~$
```

Task 5: Create a new directory in HDFS

```
exouser@bda-hadoop:~$ hdfs dfs -mkdir /assignment_data/docs
exouser@bda-hadoop:~$ hadoop fs -ls /assignment_data
Found 3 items
drwxr-xr-x   - exouser supergroup          0 2024-11-10 05:21 /assignment_data/docs
-rw-r--r--   1 exouser supergroup        320 2024-11-10 05:18 /assignment_data/file1.txt
-rw-r--r--   1 exouser supergroup        479 2024-11-10 05:18 /assignment_data/file2.csv
```

Task 6: Move files to a different directory in HDFS

```
exouser@bda-hadoop:~$ hdfs dfs -mv /assignment_data/file1.txt /assignment_data/docs/
exouser@bda-hadoop:~$ hdfs dfs -mv /assignment_data/file2.csv /assignment_data/docs/
exouser@bda-hadoop:~$ hadoop fs -ls /assignment_data/docs
Found 2 items
-rw-r--r--   1 exouser supergroup        320 2024-11-10 05:18 /assignment_data/docs/file1.txt
-rw-r--r--   1 exouser supergroup        479 2024-11-10 05:18 /assignment_data/docs/file2.csv
```

Task 7: Delete files from HDFS

```
exouser@bda-hadoop:~$ hdfs dfs -rm /assignment_data/docs/file1.txt
Deleted /assignment_data/docs/file1.txt
exouser@bda-hadoop:~$ hadoop fs -ls /assignment_data/docs
Found 1 items
-rw-r--r--   1 exouser supergroup        479 2024-11-10 05:18 /assignment_data/docs/file2.csv
```

Task 8: Check HDFS file status

```
exouser@bda-hadoop:~$ hdfs fsck /assignment_data/docs/file2.csv -files -blocks -locations
Connecting to namenode via http://localhost:9870/fsck?ugi=exouser&files=1&blocks=1&locations
=1&path=%2Fassignment_data%2Fdocs%2Ffile2.csv
```

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

Replicated Blocks:
 Total size:    479 B
 Total files:   1
 Total blocks (validated):      1 (avg. block size 479 B)
 Minimally replicated blocks:   1 (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
 Under-erasure-coded block groups:      0
 Unsatisfactory placement block groups: 0
 Average block group size:      0.0
 Missing block groups:          0
 Corrupt block groups:          0
 Missing internal blocks:       0
 Blocks queued for replication: 0
FSCK ended at Sun Nov 10 05:34:35 UTC 2024 in 7 milliseconds


The filesystem under path '/assignment_data/docs/file2.csv' is HEALTHY
```

HDFS fsck Command Summary for file2.csv

- ***File Size:***
  - *479 bytes: Total size of file2.csv in HDFS.*
- ***Replication Factor:***
  - *replication=1: The file has one copy stored across DataNodes, set by default for single-replica needs.*
- ***Blocks:***
  - *1 block(s): Entire file fits in one block, as it is smaller than the default 128MB HDFS block size.*
- ***Block Information:***
  - *BP-671428153-10.3.5.247-1731214686918:blk_1073741826_1002: Block ID for the file.*
  - *len=479: Block size is 479 bytes, identical to file size.*
  - *Live_repl=1: One live replica, matching the replication factor.*
- ***DataNode Location:***
  - *[127.0.0.1:9866, DISK]: Block stored on DataNode at IP 127.0.0.1 with port 9866, stored on disk.*
- ***Health Check Summary:***
  - ***Status:*** *HEALTHY*
    - *One active DataNode, matching the replication factor.*
    - *No missing or corrupt blocks, ensuring data integrity and accessibility.*

This health check confirms that file2.csv is securely stored and fully accessible in HDFS.

Task 9: Delete a directory from HDFS

```
exouser@bda-hadoop:~$ hadoop fs -rm -r /assignment_data/docs/
Deleted /assignment_data/docs
exouser@bda-hadoop:~$ hadoop fs -ls /assignment_data
exouser@bda-hadoop:~$
```

3. Assess your understanding and proficiency in Mapred

Task 10: Dataset Overview

Social Recommendation Data

This reviews dataset includes ratings as well as social (or trust) relationships between users. Data are from LibraryThing (a book review website).

Description of the Dataset

•**File Format** : .txt

- The dataset is structured as a dictionary in Python, with nested dictionaries for each review.
- Reviews are organized by keys that are tuples, where:
  - The first element is a product/work ID.
  - The second element is the user ID.

•**Size**:

- The dataset is large, around 1.66 GB.

•**Type of Data**:

- Contains reviews for products or works (e.g., books or media).
- Each review has the following fields:
  - **'comment'**: Review text describing user experience.
  - **'nhelpful'**: Number of helpful votes the review received.
  - **'unixtime'**: Unix timestamp of the review submission.
  - **'work'**: Unique identifier for the product/work reviewed.
  - **'flags'**: List of moderation flags (empty in this case).
  - **'user'**: ID of the user who wrote the review.
  - **'stars'**: Rating given by the user (numeric value).
  - **'time'**: Date when the review was submitted.

Summary: The dataset contains user reviews with detailed attributes, including review text, rating, and helpfulness. It is structured for easy access and can be used for sentiment analysis, recommendation systems, or understanding user preferences.

Download the data on local computer and copy it to local Jetstream instance using ssh

```
(base) mukundkomati@Mukunds-MacBook-Pro lthing_data % scp reviews.txt exouser@149.165.154.230:~/
exouser@149.165.154.230's password:
reviews.txt                                   100% 1589MB   2.9MB/s   09:16
```

Upload the reviews data to HDFS

```
exouser@bda-hadoop:~$ hdfs dfs -put reviews.txt /assignment_data/
exouser@bda-hadoop:~$
```

Check the status of the ”reviews.txt" in terms of file size, replication factor, and block locations


```
exouser@bda-hadoop:~$ hdfs fsck /assignment_data/reviews.txt -files -blocks -locations
Connecting to namenode via http://localhost:9870/fsck?ugi=exouser&files=1&blocks=1&locations=1&path=%2Fassignment_data%2Freviews.txt
FSCK started by exouser (auth:SIMPLE) from /127.0.0.1 for path /assignment_data/reviews.txt at Sun Nov 10 16:16:59 UTC 2024

/assignment_data/reviews.txt 1665980007 bytes, replicated: replication=1, 13 block(s):  OK
0. BP-671428153-10.3.5.247-1731214686918:blk_1073741942_1118 len=134217728 Live_repl=1  [DatanodeInfoWithStorage[127.0.0.1:9866,DS-9
1. BP-671428153-10.3.5.247-1731214686918:blk_1073741943_1119 len=134217728 Live_repl=1  [DatanodeInfoWithStorage[127.0.0.1:9866,DS-9
2. BP-671428153-10.3.5.247-1731214686918:blk_1073741944_1120 len=134217728 Live_repl=1  [DatanodeInfoWithStorage[127.0.0.1:9866,DS-9
3. BP-671428153-10.3.5.247-1731214686918:blk_1073741945_1121 len=134217728 Live_repl=1  [DatanodeInfoWithStorage[127.0.0.1:9866,DS-9
4. BP-671428153-10.3.5.247-1731214686918:blk_1073741946_1122 len=134217728 Live_repl=1  [DatanodeInfoWithStorage[127.0.0.1:9866,DS-9
5. BP-671428153-10.3.5.247-1731214686918:blk_1073741947_1123 len=134217728 Live_repl=1  [DatanodeInfoWithStorage[127.0.0.1:9866,DS-9
6. BP-671428153-10.3.5.247-1731214686918:blk_1073741948_1124 len=134217728 Live_repl=1  [DatanodeInfoWithStorage[127.0.0.1:9866,DS-9
7. BP-671428153-10.3.5.247-1731214686918:blk_1073741949_1125 len=134217728 Live_repl=1  [DatanodeInfoWithStorage[127.0.0.1:9866,DS-9
8. BP-671428153-10.3.5.247-1731214686918:blk_1073741950_1126 len=134217728 Live_repl=1  [DatanodeInfoWithStorage[127.0.0.1:9866,DS-9
9. BP-671428153-10.3.5.247-1731214686918:blk_1073741951_1127 len=134217728 Live_repl=1  [DatanodeInfoWithStorage[127.0.0.1:9866,DS-9
10. BP-671428153-10.3.5.247-1731214686918:blk_1073741952_1128 len=134217728 Live_repl=1  [DatanodeInfoWithStorage[127.0.0.1:9866,DS-
11. BP-671428153-10.3.5.247-1731214686918:blk_1073741953_1129 len=134217728 Live_repl=1  [DatanodeInfoWithStorage[127.0.0.1:9866,DS-
12. BP-671428153-10.3.5.247-1731214686918:blk_1073741954_1130 len=55367271 Live_repl=1  [DatanodeInfoWithStorage[127.0.0.1:9866,DS-9


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

Replicated Blocks:
 Total size:    1665980007 B
 Total files:   1
 Total blocks (validated):      13 (avg. block size 128152308 B)
 Minimally replicated blocks:   13 (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
 Under-erasure-coded block groups:      0
 Unsatisfactory placement block groups: 0
 Average block group size:      0.0
 Missing block groups:          0
 Corrupt block groups:          0
 Missing internal blocks:       0
 Blocks queued for replication: 0
FSCK ended at Sun Nov 10 16:16:59 UTC 2024 in 1 milliseconds


The filesystem under path '/assignment_data/reviews.txt' is HEALTHY
```

File Path: /assignment_data/reviews.txt
Location of the file within HDFS.
Status: HEALTHY
The file is intact with no issues detected.
Total Size: 1.66 GB (1,665,980,007 bytes)
Total size of the file in bytes.
Replication Factor: 1
Only one replica of each block exists.
Total Blocks: 13
The file is split across 13 blocks for storage.
Average Block Size: ~128 MB (128,152,308 bytes)
Each block is approximately 128 MB in size.
Data Nodes: 1
There is a single data node holding the file blocks.
Racks: 1
All blocks are stored on a single rack.
Block Details:
All 13 blocks are stored on the data node at 127.0.0.1:9866.

Replication and Integrity Check:
•**Minimally Replicated Blocks**: 100% (13/13)
All blocks meet the minimum replication level.

Overall Health Check Result:
The HDFS file /assignment_data/reviews.txt is confirmed healthy with no missing, corrupt, or under-replicated blocks, ensuring reliable storage and data integrity.

Task 11: Word Count

Write mapper and reducer functions for the Word Count Program

```
exouser@bda-hadoop:~$ vi word_count_mapper.py
exouser@bda-hadoop:~$ vi word_count_reducer.py
```

mapper function

```
#!/usr/bin/env python
import sys
import re

# Input comes from standard input (stdin)
for line in sys.stdin:
    # Remove punctuation and numbers, and convert to lowercase
    line = re.sub(r'[^a-zA-Z\s]', '', line).lower()

    # Split the line into words
    words = line.split()

    # Output each word with count 1
    for word in words:
        print(f"{word}\t1")
```

reducer function

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

current_word = None
current_count = 0

# Input comes from standard input (stdin)
for line in sys.stdin:
    # Read each line, which contains a word and its count
    word, count = line.strip().split('\t')
    count = int(count)

    if current_word == word:
        current_count += count
    else:
        if current_word:
            # Output the count for the previous word
            print(f"{current_word}\t{current_count}")
        current_word = word
        current_count = count

# Output the count for the last word
if current_word == word:
    print(f"{current_word}\t{current_count}")
```

Add execute permission to the .py files

```
exouser@bda-hadoop:~$ chmod +x word_count_mapper.py
exouser@bda-hadoop:~$ chmod +x word_count_reducer.py
```

Submit the mapreduce job using Hadoop streaming

```
exouser@bda-hadoop:~$ hadoop jar /home/exouser/hadoop-3.4.0/share/hadoop/tools/lib/hadoop-streaming-3.4.0.jar \
    -input /assignment_data/reviews.txt \
    -output /assignment_output/word_count \
    -mapper "/usr/bin/python3 word_count_mapper.py" \
    -reducer "/usr/bin/python3 word_count_reducer.py" \
    -file word_count_mapper.py \
    -file word_count_reducer.py
2024-11-10 16:24:26,356 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [word_count_mapper.py, word_count_reducer.py, /tmp/hadoop-unjar6383017485394013305/] [] /tmp/streamjob8697413192479995578.jar tmpDir=null
2024-11-10 16:24:27,048 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-11-10 16:24:27,170 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-11-10 16:24:27,367 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/exouser/.staging/job_1731214806101_0011
2024-11-10 16:24:27,628 INFO mapred.FileInputFormat: Total input files to process : 1
2024-11-10 16:24:27,670 INFO mapreduce.JobSubmitter: number of splits:13
2024-11-10 16:24:27,764 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1731214806101_0011
2024-11-10 16:24:27,764 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-11-10 16:24:27,923 INFO conf.Configuration: resource-types.xml not found
2024-11-10 16:24:27,923 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-11-10 16:24:27,989 INFO impl.YarnClientImpl: Submitted application application_1731214806101_0011
2024-11-10 16:24:28,021 INFO mapreduce.Job: The url to track the job: http://bda-hadoop:8088/proxy/application_1731214806101_0011/
2024-11-10 16:24:28,022 INFO mapreduce.Job: Running job: job_1731214806101_0011
```

Status % of mapreduce jobs

```
2024-11-10 16:27:57,551 INFO mapreduce.Job:  map 100% reduce 87%
2024-11-10 16:28:03,571 INFO mapreduce.Job:  map 100% reduce 89%
2024-11-10 16:28:09,593 INFO mapreduce.Job:  map 100% reduce 91%
2024-11-10 16:28:15,614 INFO mapreduce.Job:  map 100% reduce 92%
2024-11-10 16:28:21,641 INFO mapreduce.Job:  map 100% reduce 93%
2024-11-10 16:28:27,662 INFO mapreduce.Job:  map 100% reduce 96%
2024-11-10 16:28:33,681 INFO mapreduce.Job:  map 100% reduce 97%
2024-11-10 16:28:39,700 INFO mapreduce.Job:  map 100% reduce 99%
2024-11-10 16:28:44,720 INFO mapreduce.Job:  map 100% reduce 100%
2024-11-10 16:28:44,724 INFO mapreduce.Job: Job job_1731214806101_0011 completed successfully
```

Final Status of the job on YARN resource manager Web UI

Application application_1731214806101_0011

User:	exouser
Name:	streamjob8697413192479995578.jar
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	root.default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Sun Nov 10 16:24:27 +0000 2024
Launched:	Sun Nov 10 16:24:28 +0000 2024
Finished:	Sun Nov 10 16:28:43 +0000 2024
Elapsed:	4mins, 15sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

Displaying the first 5 words from the word count output

```
exouser@bda-hadoop:~$ hadoop fs -ls /assignment_output/word_count
Found 2 items
-rw-r--r--   1 exouser supergroup          0 2024-11-10 16:28 /assignment_output/word_count/_SUCCESS
-rw-r--r--   1 exouser supergroup   41137392 2024-11-10 16:28 /assignment_output/word_count/part-00000
exouser@bda-hadoop:~$ hadoop fs -cat /assignment_output/word_count/part-00000 | head -n 5
a       6358648
aa      714
aaa     85
aaaa    7
aaaaa   3
cat: Unable to write to output stream.
```

MapReduce Workflow

1.Input Data

- Input: Large text file (reviews.txt) in HDFS.

2.Map Step

- Reads each line, removes punctuation, converts to lowercase, and splits into words.
- Outputs key-value pairs: <word> <1>.

3.Shuffle and Sort Step

- Groups identical words together across mappers.

4.Reduce Step

- Sums the occurrences for each word and outputs the count. Note that the words are already groupec together from the short and shuffle step.

5.Final Output

- Result: <word> <count> written to HDFS.

6.Execution

- The job is executed in parallel across the cluster, with mappers processing data and reducers aggregating results.

Task 12: Top N words

The mapper function remains same

```
exouser@bda-hadoop:~$ cp word_count_mapper.py topn_mapper.py
exouser@bda-hadoop:~$ vi topn_reducer.py
```

Write the reducer function for Top N words

```
#!/usr/bin/env python3
import sys
from collections import defaultdict, Counter

def reduce_function(K=5):
    """Reduce function to read word count pairs and find the top K words."""
    word_count = defaultdict(int)

    # Count occurrences of each word
    for line in sys.stdin:
        word, count = line.strip().split('\t')
        word_count[word] += int(count)

    # Sort words by frequency (descending)
    sorted_word_count = sorted(word_count.items(), key=lambda x: x[1], reverse=True)

    # Output the top K words and their counts
    for word, count in sorted_word_count[:K]:
        print(f"{word}\t{count}")

if __name__ == "__main__":
    # Check if K is provided via command-line arguments
    if len(sys.argv) > 1:
        K = int(sys.argv[1])  # Retrieve K from command-line arguments
    else:
        K = 5  # Default value if not provided

    reduce_function(K)
```

Add execute permissions to the .py files

```
exouser@bda-hadoop:~$ chmod +x topn_mapper.py
exouser@bda-hadoop:~$ chmod +x topn_reducer.py
```

Submit the mapreduce job using Hadoop streaming for top 3 most frequent words from the data

```
exouser@bda-hadoop:~$ hadoop jar /home/exouser/hadoop-3.4.0/share/hadoop/tools/lib/hadoop-streaming-3.4.0.jar \
-Dmapreduce.job.name="TopN" \
    -input /assignment_data/reviews.txt \
    -output /assignment_output/topn \
    -mapper "/usr/bin/python3 topn_mapper.py" \
    -reducer "/usr/bin/python3 topn_reducer.py 3" \
    -file topn_mapper.py \
    -file topn_reducer.py
2024-11-10 16:34:30,353 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [topn_mapper.py, topn_reducer.py, /tmp/hadoop-unjar6787351379380709531/] [] /tmp/streamjob3680203936111490762.jar tmpDir=null
2024-11-10 16:34:31,116 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-11-10 16:34:31,251 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-11-10 16:34:31,468 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/exouser/.staging/job_1731214806101_0012
2024-11-10 16:34:31,757 INFO mapred.FileInputFormat: Total input files to process : 1
2024-11-10 16:34:31,807 INFO mapreduce.JobSubmitter: number of splits:13
2024-11-10 16:34:31,909 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1731214806101_0012
2024-11-10 16:34:31,909 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-11-10 16:34:32,081 INFO conf.Configuration: resource-types.xml not found
2024-11-10 16:34:32,082 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-11-10 16:34:32,150 INFO impl.YarnClientImpl: Submitted application application_1731214806101_0012
2024-11-10 16:34:32,189 INFO mapreduce.Job: The url to track the job: http://bda-hadoop:8088/proxy/application_1731214806101_0012/
2024-11-10 16:34:32,190 INFO mapreduce.Job: Running job: job_1731214806101_0012
2024-11-10 16:34:37,295 INFO mapreduce.Job: Job job_1731214806101_0012 running in uber mode : false
```

Status % of mapreduce jobs

```
2024-11-10 16:38:24,689 INFO mapreduce.Job:  map 100% reduce 96%
2024-11-10 16:38:30,709 INFO mapreduce.Job:  map 100% reduce 97%
2024-11-10 16:38:36,728 INFO mapreduce.Job:  map 100% reduce 99%
2024-11-10 16:38:42,742 INFO mapreduce.Job:  map 100% reduce 100%
2024-11-10 16:38:42,747 INFO mapreduce.Job: Job job_1731214806101_0012 completed successful
```

Final Status of the job on YARN resource manager Web UI

Application application_1731214806101_0012

User:	exouser
Name:	TopN
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	root.default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Sun Nov 10 16:34:32 +0000 2024
Launched:	Sun Nov 10 16:34:32 +0000 2024
Finished:	Sun Nov 10 16:38:41 +0000 2024
Elapsed:	4mins, 9sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

Displaying the top 3 words from the data

```
exouser@bda-hadoop:~$ hadoop fs -ls /assignment_output/topn
Found 2 items
-rw-r--r--   1 exouser supergroup          0 2024-11-10 16:38 /assignment_output/topn/_SUCCESS
-rw-r--r--   1 exouser supergroup         36 2024-11-10 16:38 /assignment_output/topn/part-00000
exouser@bda-hadoop:~$ hadoop fs -cat /assignment_output/topn/part-00000
the     12398778
and     7117250
of      6389641
```

Submit a job for the top 10 most frequent words from the data to test input functionality

```
exouser@bda-hadoop:~$ hadoop jar /home/exouser/hadoop-3.4.0/share/hadoop/tools/lib/hadoop-streaming-3.4.0.jar \
-Dmapreduce.job.name="TopN" \
    -input /assignment_data/reviews.txt \
    -output /assignment_output/topn \
    -mapper "/usr/bin/python3 topn_mapper.py" \
    -reducer "/usr/bin/python3 topn_reducer.py 10" \
    -file topn_mapper.py \
    -file topn_reducer.py
2024-11-10 16:52:48,573 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [topn_mapper.py, topn_reducer.py, /tmp/hadoop-unjar859943638536217995/] [] /tmp/streamjob6171959584055143455.jar tmpDir=null
2024-11-10 16:52:49,282 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-11-10 16:52:49,408 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-11-10 16:52:49,608 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/exouser/.staging/job_1731214806101_0014
2024-11-10 16:52:49,874 INFO mapred.FileInputFormat: Total input files to process : 1
2024-11-10 16:52:49,915 INFO mapreduce.JobSubmitter: number of splits:13
2024-11-10 16:52:50,409 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1731214806101_0014
2024-11-10 16:52:50,409 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-11-10 16:52:50,566 INFO conf.Configuration: resource-types.xml not found
2024-11-10 16:52:50,567 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-11-10 16:52:50,627 INFO impl.YarnClientImpl: Submitted application application_1731214806101_0014
2024-11-10 16:52:50,657 INFO mapreduce.Job: The url to track the job: http://bda-hadoop:8088/proxy/application_1731214806101_0014/
2024-11-10 16:52:50,658 INFO mapreduce.Job: Running job: job_1731214806101_0014
2024-11-10 16:52:55,736 INFO mapreduce.Job: Job job_1731214806101_0014 running in uber mode : false
```

Application application_1731214806101_0014

User:	exouser
Name:	TopN
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	root.default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Sun Nov 10 16:52:50 +0000 2024
Launched:	Sun Nov 10 16:52:50 +0000 2024
Finished:	Sun Nov 10 16:56:56 +0000 2024
Elapsed:	4mins, 5sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

*Note that the total runtime of the topk words without combiner is **4mins 5sec***

```
exouser@bda-hadoop:~$ hadoop fs -cat /assignment_output/topn/part-00000
the     12398778
and     7117250
of      6389641
a       6358648
to      5935163
is      3802221
in      3801718
i       3518781
that    2676368
it      2527595
```

MapReduce Workflow

Map Phase:
•The mapper reads input lines (from HDFS or standard input), processes each line, and outputs a key-value pair for every word (word\t1).
Shuffle and Sort Phase (Managed by the Hadoop framework):
•The framework groups the output of all mappers by key (the word) and sorts it. All occurrences of the same word are placed together, and the key-value pairs are sorted by key.
Reduce Phase:
•The reducer receives the grouped key-value pairs (words with their respective counts) from the mappers. It aggregates the counts, sorts the words by frequency, and outputs the top K most frequent words.

Task 12: Top N words with combiner

The mapper and reducer functions remain the same

```
exouser@bda-hadoop:~$ vi topn_combiner.py
exouser@bda-hadoop:~$ chmod +x topn_combiner.py
```

Write the combiner function and add permission to the combiner python file

```
#!/usr/bin/env python
import sys
from collections import defaultdict

# Dictionary to store local counts of words
local_word_count = defaultdict(int)

# Input comes from standard input (stdin)
for line in sys.stdin:
    word, count = line.strip().split('\t')
    local_word_count[word] += int(count)

# Output each word with its local count
for word, count in local_word_count.items():
    print(f"{word}\t{count}")
~
```

Submit the mapreduce job using Hadoop streaming for top 3 most frequent words from the data
Note that combiner is passed as a parameter while submitting the job

```
exouser@bda-hadoop:~$ hadoop jar /home/exouser/hadoop-3.4.0/share/hadoop/tools/lib/hadoop-streaming-3.4.0.jar \
    -Dmapreduce.job.name="TopNCombiner" \
    -input /assignment_data/reviews.txt \
    -output /assignment_output/topn_combiner \
    -mapper "/usr/bin/python3 topn_mapper.py" \
    -combiner "/usr/bin/python3 topn_combiner.py" \
    -reducer "/usr/bin/python3 topn_reducer.py 3" \
    -file topn_mapper.py \
    -file topn_combiner.py \
    -file topn_reducer.py
2024-11-10 16:42:24,942 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [topn_mapper.py, topn_combiner.py, topn_reducer.py, /tmp/hadoop-unjar5877037369855075129/] [] /tmp/streamjob6408743125476204987.jar tmpDir=
2024-11-10 16:42:25,665 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-11-10 16:42:25,792 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-11-10 16:42:25,989 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/exouser/.staging/job_1731214806101
2024-11-10 16:42:26,297 INFO mapred.FileInputFormat: Total input files to process : 1
2024-11-10 16:42:26,341 INFO mapreduce.JobSubmitter: number of splits:13
2024-11-10 16:42:26,839 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1731214806101_0013
2024-11-10 16:42:26,839 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-11-10 16:42:27,000 INFO conf.Configuration: resource-types.xml not found
2024-11-10 16:42:27,001 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-11-10 16:42:27,063 INFO impl.YarnClientImpl: Submitted application application_1731214806101_0013
2024-11-10 16:42:27,095 INFO mapreduce.Job: The url to track the job: http://bda-hadoop:8088/proxy/application_1731214806101_0013/
2024-11-10 16:42:27,097 INFO mapreduce.Job: Running job: job_1731214806101_0013
2024-11-10 16:42:32,206 INFO mapreduce.Job: Job job_1731214806101_0013 running in uber mode : false
```

Status % of mapreduce jobs

```
2024-11-10 16:44:00,150 INFO mapreduce.Job:  map 78% reduce 0%
2024-11-10 16:44:01,158 INFO mapreduce.Job:  map 79% reduce 0%
2024-11-10 16:44:02,163 INFO mapreduce.Job:  map 85% reduce 0%
2024-11-10 16:44:03,171 INFO mapreduce.Job:  map 92% reduce 0%
2024-11-10 16:44:16,241 INFO mapreduce.Job:  map 92% reduce 31%
2024-11-10 16:44:17,245 INFO mapreduce.Job:  map 100% reduce 31%
2024-11-10 16:44:22,262 INFO mapreduce.Job:  map 100% reduce 84%
2024-11-10 16:44:24,270 INFO mapreduce.Job:  map 100% reduce 100%
2024-11-10 16:44:24,275 INFO mapreduce.Job: Job job_1731214806101_0013 completed successfully
```

Final Status of the job on YARN resource manager Web UI

Application application_1731214806101_0013

User:	exouser
Name:	TopNCombiner
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	root.default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Sun Nov 10 16:42:27 +0000 2024
Launched:	Sun Nov 10 16:42:27 +0000 2024
Finished:	Sun Nov 10 16:44:23 +0000 2024
Elapsed:	1mins, 56sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

Displaying the top 3 words from the data

```
exouser@bda-hadoop:~$ hadoop fs -ls /assignment_output/topn_combiner
Found 2 items
-rw-r--r--   1 exouser supergroup          0 2024-11-10 16:44 /assignment_output/topn_combiner/_SUCCESS
-rw-r--r--   1 exouser supergroup         36 2024-11-10 16:44 /assignment_output/topn_combiner/part-00000
exouser@bda-hadoop:~$ hadoop fs -cat /assignment_output/topn_combiner/part-00000
the     12398778
and     7117250
of      6389641
```

Submit a job for the top 10 most frequent words from the data to test input functionality

```
exouser@bda-hadoop:~$ hadoop jar /home/exouser/hadoop-3.4.0/share/hadoop/tools/lib/hadoop-streaming-3.4.0.jar \
    -Dmapreduce.job.name="TopNCombiner" \
    -input /assignment_data/reviews.txt \
    -output /assignment_output/topn_combiner \
    -mapper "/usr/bin/python3 topn_mapper.py" \
    -combiner "/usr/bin/python3 topn_combiner.py" \
    -reducer "/usr/bin/python3 topn_reducer.py 10" \
    -file topn_mapper.py \
    -file topn_combiner.py \
    -file topn_reducer.py
2024-11-10 17:11:40,390 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [topn_mapper.py, topn_combiner.py, topn_reducer.py, /tmp/hadoop-unjar7394744815625031677/] [] /tmp/streamjob3039332507576471108.jar tmpDir=null
2024-11-10 17:11:41,107 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-11-10 17:11:41,230 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-11-10 17:11:41,427 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/exouser/.staging/job_1731214806101_0015
2024-11-10 17:11:41,707 INFO mapred.FileInputFormat: Total input files to process : 1
2024-11-10 17:11:41,749 INFO mapreduce.JobSubmitter: number of splits:13
2024-11-10 17:11:42,245 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1731214806101_0015
2024-11-10 17:11:42,245 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-11-10 17:11:42,410 INFO conf.Configuration: resource-types.xml not found
2024-11-10 17:11:42,410 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-11-10 17:11:42,477 INFO impl.YarnClientImpl: Submitted application application_1731214806101_0015
2024-11-10 17:11:42,509 INFO mapreduce.Job: The url to track the job: http://bda-hadoop:8088/proxy/application_1731214806101_0015/
2024-11-10 17:11:42,511 INFO mapreduce.Job: Running job: job_1731214806101_0015
2024-11-10 17:11:47,597 INFO mapreduce.Job: Job job_1731214806101_0015 running in uber mode : false
```

User:	exouser
Name:	TopNCombiner
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	root.default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Sun Nov 10 17:11:42 +0000 2024
Launched:	Sun Nov 10 17:11:42 +0000 2024
Finished:	Sun Nov 10 17:13:50 +0000 2024
Elapsed:	2mins, 8sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

Note that the total runtime of the topk words with combiner is ***2mins 8sec***

The output with combiner is same as without combiner. The combiner does NOT change the result

```
exouser@bda-hadoop:~$ hadoop fs -cat /assignment_output/topn_combiner/part-00000
the     12398778
and     7117250
of      6389641
a       6358648
to      5935163
is      3802221
in      3801718
i       3518781
that    2676368
it      2527595
```

Combiner vs. No Combiner:

•Without a Combiner:

- All mapper outputs are sent to the reducers, even if they contain duplicate keys (like the word "hello").
- The reducer has to do all the aggregation work (counting each word's total occurrences).
- This leads to more network traffic and slower performance.

•With a Combiner:

- The mapper's output is pre-aggregated locally by the combiner, so duplicate keys (e.g., multiple ”the 1"s) are merged before they are sent to the reducer.
- This reduces the amount of data transferred across the network, speeding up the shuffle phase.
- The reducer gets fewer key-value pairs to aggregate, reducing its workload.

The job's runtime with the combiner is approximately 2 minutes faster due to the improved performance in the shuffle phase. Additionally, the reducer phase executes more quickly as it processes fewer key-value pairs.