EXP 2: Run a basic Word Count Map Reduce program to understand Map Reduce Paradigm.

AIM:

To run a basic Word Count MapReduce program.

PROCEDURE:

Step 1: Create Data File
1. Log in with your Hadoop user.
2. Create a file named `word_count_data.txt`.
3. Populate the file with the text data you wish to analyze.

Step 2: Mapper Logic
1. Create a file named `mapper.py`.
2. Write the logic to read input, split lines into words, and output each word with a count.

Step 3: Reducer Logic
1. Create a file named `reducer.py`.
2. Write the logic to aggregate the occurrences of each word and generate the final count.

Step 4: Prepare Hadoop Environment
1. Start Hadoop daemons by running the necessary command.
2. Create a directory in HDFS to store your data.

Step 5: Upload Data to HDFS
1. Copy your `word_count_data.txt` file from the local file system to HDFS.

Step 6: Make Python Files Executable
1. Grant executable permissions to the `mapper.py` and `reducer.py` files.

Step 7: Run Word Count with Hadoop Streaming
1. Download the Hadoop Streaming JAR file.
2. Run the Word Count program by specifying the input data, output directory, and the mapper and reducer files.

Step 8: Check Output
1. Check the output of the Word Count program in the specified HDFS output directory.

Commands:

C:\hadoop\sbin> **start-all.cmd**

C:\hadoop\sbin> **jps**

C:\hadoop\sbin> **cd /**

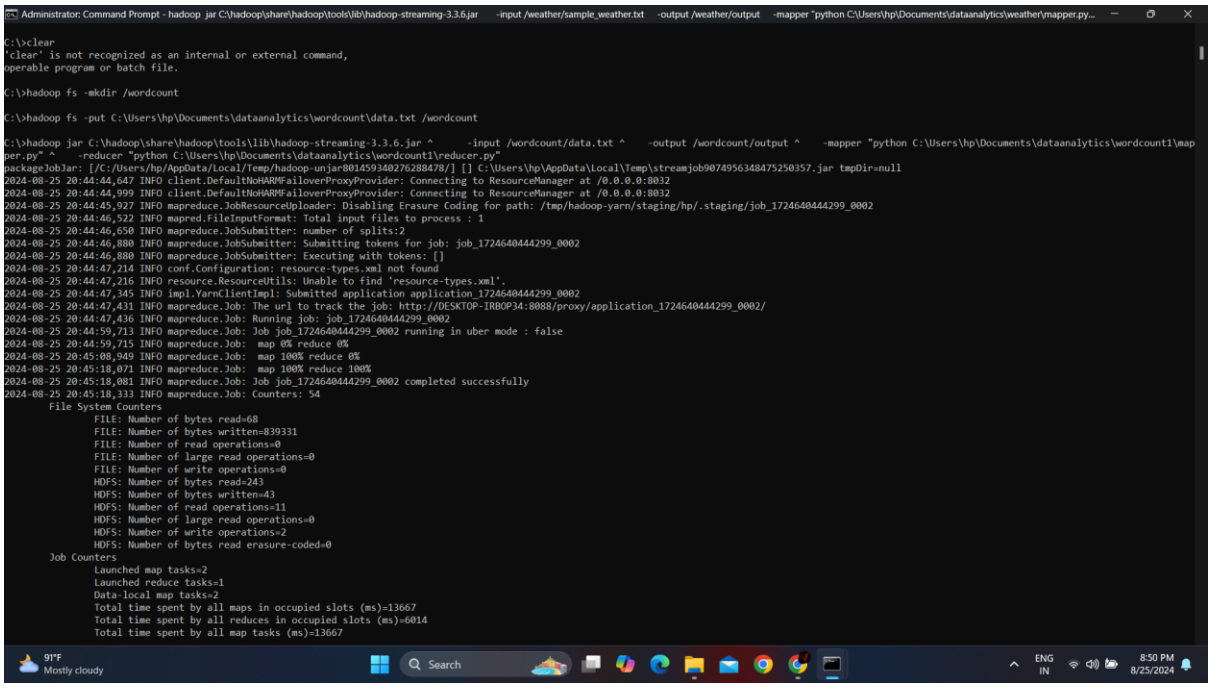
C:\> **cd hadoop**

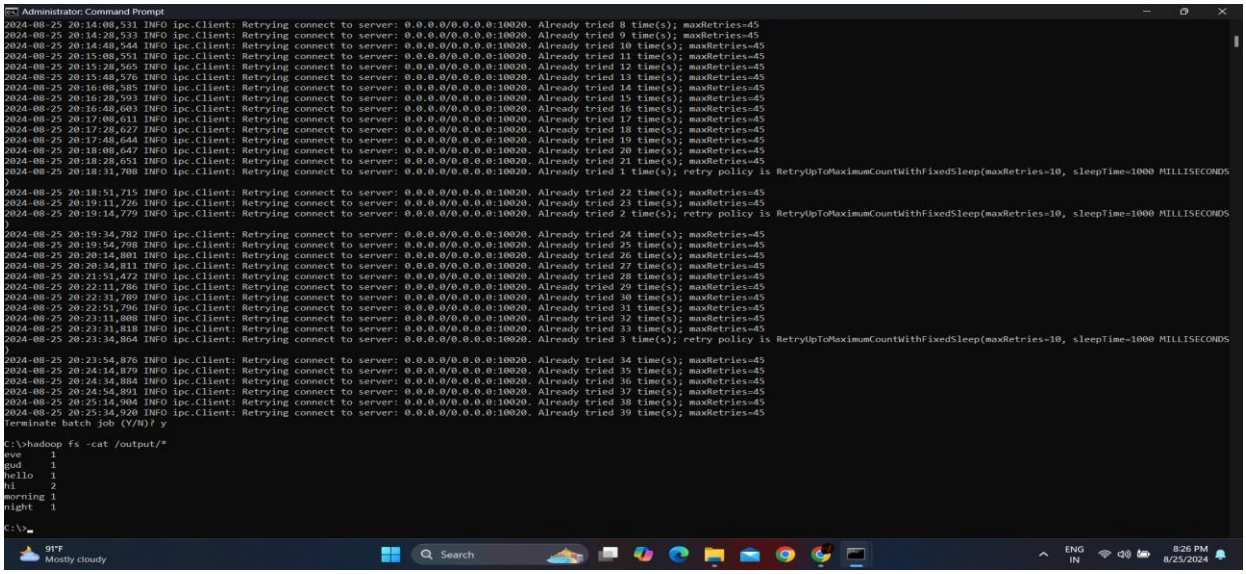
C:\hadoop> **hadoop fs -mkdir input**

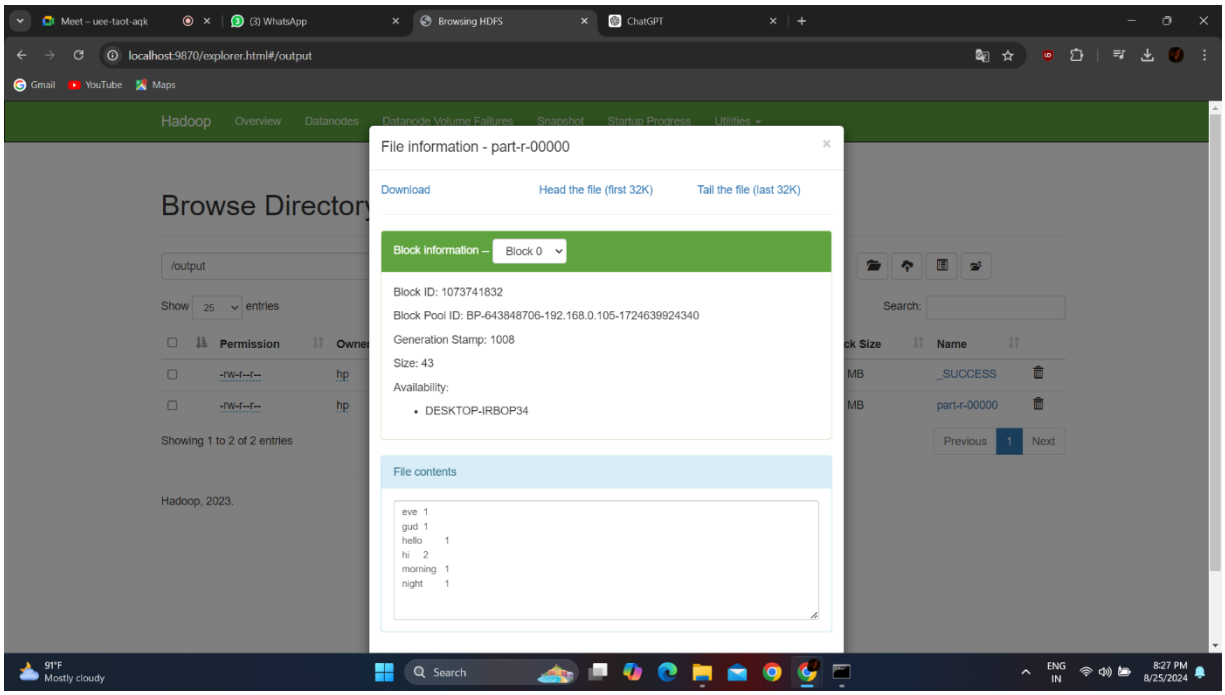
C:\hadoop> **hadoop fs -put C:/Users/hp/Documents/wordcount/data.txt /input1**

C:\hadoop> **hadoop jar C:\hadoop\share\hadoop\tools\lib\hadoop-streaming-3.3.6.jar -input /user/input/inpfile.txt -output /user/output -mapper " C:\Users\hp\Documents\wordcount\mapper.py" -reducer " C:\Users\hp\Documents\wordcount\reducer.py"**

OUTPUT:




Administrator: Command Prompt
2024-08-25 20:14:08,531 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 8 time(s); maxRetries=45
2024-08-25 20:14:28,533 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 time(s); maxRetries=45
2024-08-25 20:14:48,544 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 10 time(s); maxRetries=45
2024-08-25 20:15:08,551 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 11 time(s); maxRetries=45
2024-08-25 20:15:28,565 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 12 time(s); maxRetries=45
2024-08-25 20:15:48,576 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 13 time(s); maxRetries=45
2024-08-25 20:16:08,585 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 14 time(s); maxRetries=45
2024-08-25 20:16:28,593 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 15 time(s); maxRetries=45
2024-08-25 20:16:48,603 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 16 time(s); maxRetries=45
2024-08-25 20:17:08,611 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 17 time(s); maxRetries=45
2024-08-25 20:17:28,627 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 18 time(s); maxRetries=45
2024-08-25 20:17:48,644 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 19 time(s); maxRetries=45
2024-08-25 20:18:08,647 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 20 time(s); maxRetries=45
2024-08-25 20:18:28,651 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 21 time(s); maxRetries=45
2024-08-25 20:18:31,708 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 1 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS
)
2024-08-25 20:18:51,715 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 22 time(s); maxRetries=45
2024-08-25 20:19:11,726 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 23 time(s); maxRetries=45
2024-08-25 20:19:14,779 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 2 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS
)
2024-08-25 20:19:34,782 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 24 time(s); maxRetries=45
2024-08-25 20:19:54,798 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 25 time(s); maxRetries=45
2024-08-25 20:20:14,801 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 26 time(s); maxRetries=45
2024-08-25 20:20:34,811 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 27 time(s); maxRetries=45
2024-08-25 20:21:51,472 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 28 time(s); maxRetries=45
2024-08-25 20:22:11,786 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 29 time(s); maxRetries=45
2024-08-25 20:22:31,789 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 30 time(s); maxRetries=45
2024-08-25 20:22:51,796 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 31 time(s); maxRetries=45
2024-08-25 20:23:11,808 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 32 time(s); maxRetries=45
2024-08-25 20:23:31,818 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 33 time(s); maxRetries=45
2024-08-25 20:23:34,864 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 3 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS
)
2024-08-25 20:23:54,876 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 34 time(s); maxRetries=45
2024-08-25 20:24:14,879 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 35 time(s); maxRetries=45
2024-08-25 20:24:34,884 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 36 time(s); maxRetries=45
2024-08-25 20:24:54,891 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 37 time(s); maxRetries=45
2024-08-25 20:25:14,904 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 38 time(s); maxRetries=45
2024-08-25 20:25:34,920 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 39 time(s); maxRetries=45
Terminate batch job (Y/N)? y
C:\>hadoop fs -cat /output/*
eve 1
gud 1
hello 1
hi 2
morning 1
night 1
C:\>
91°F
Mostly cloudy
Search
ENG
IN
8:26 PM
8/25/2024


Meet – uee-taot-aqk
(3) WhatsApp
Browsing HDFS
ChatGPT
localhost:9870/explorer.html#/output
Gmail
YouTube
Maps
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/output
Show 25 entries
Search:
Permission
Owner
Block Size
Name
-rw-r--r--
hp
MB
_SUCCESS
-rw-r--r--
hp
MB
part-r-00000
Showing 1 to 2 of 2 entries
Previous
1
Next
Hadoop, 2023.
File information - part-r-00000
Download
Head the file (first 32K)
Tail the file (last 32K)
Block information –
Block 0
Block ID: 1073741832
Block Pool ID: BP-643848706-192.168.0.105-1724639924340
Generation Stamp: 1008
Size: 43
Availability:
DESKTOP-IRBOP34
File contents
eve 1
gud 1
hello 1
hi 2
morning 1
night 1
91°F
Mostly cloudy
Search
ENG
IN
8:27 PM
8/25/2024