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 /**

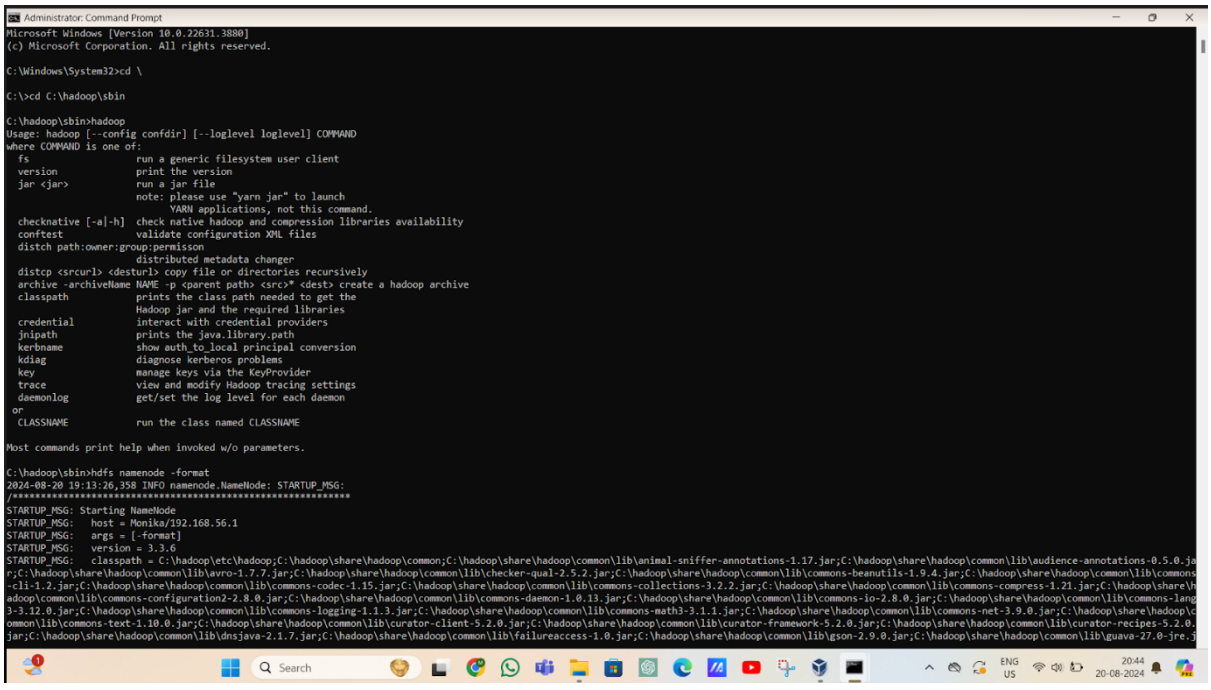
C:\> **cd hadoop**

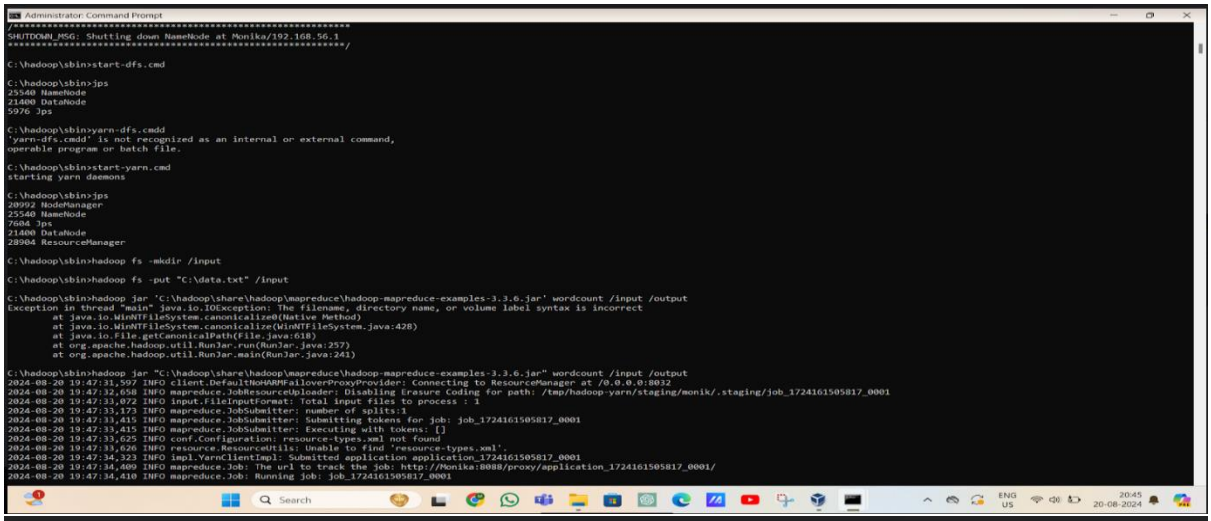
C:\hadoop> **hadoop fs -mkdir input**

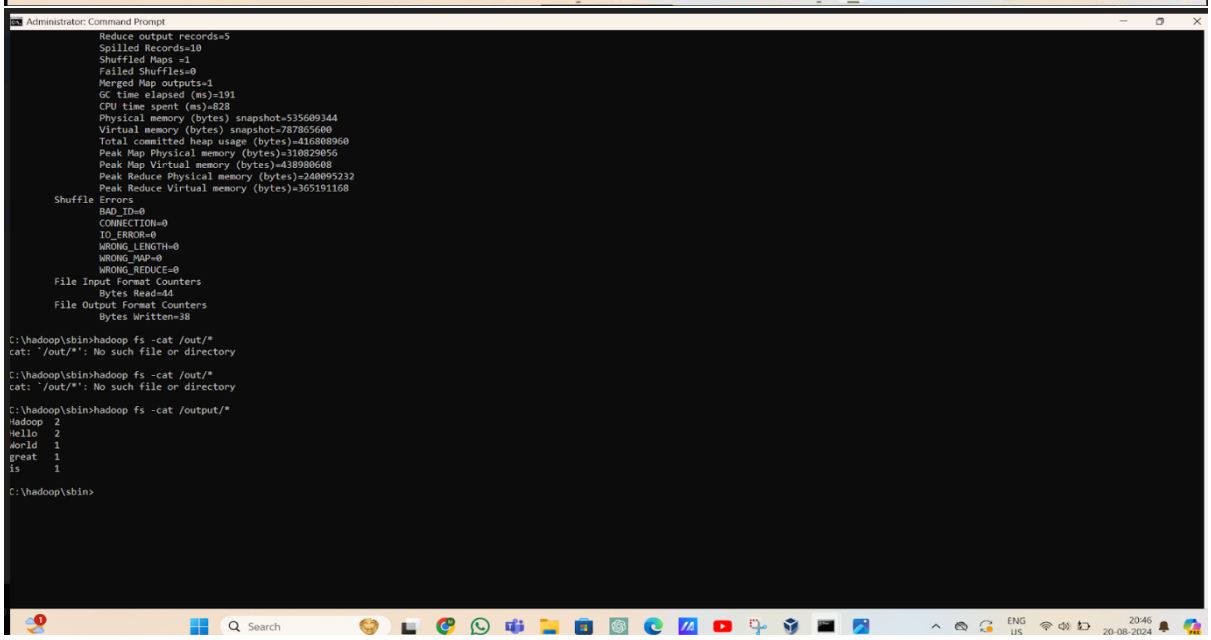
C:\hadoop> **hadoop fs -put C:/Users/monik/Documents/wordcount/data.txt /input1**

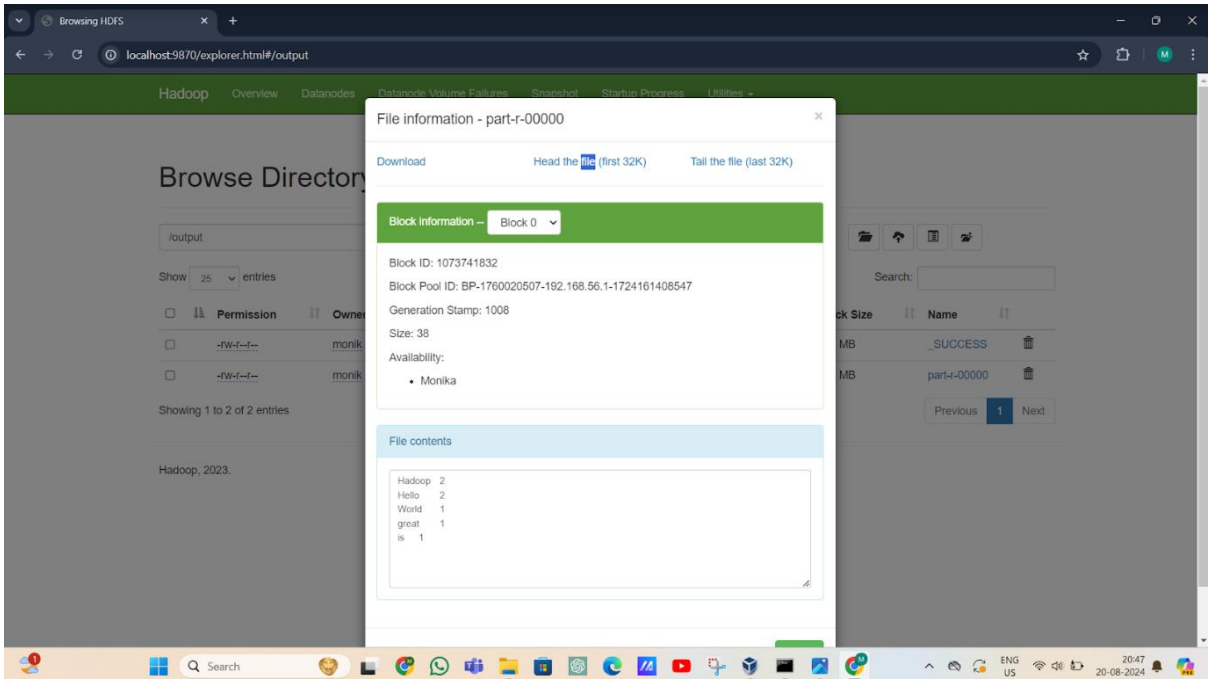
C:\hadoop> **hadoop jar C:\hadoop\share\hadoop\tools\lib\hadoop-streaming-3.3.6.jar -input /user/input/inpfile.txt -output /user/output -mapper " C:\Users\monik\Documents\wordcount\mapper.py" -reducer " C:\Users\monik\Documents\wordcount\reducer.py"**

OUTPUT:




Administrator: Command Prompt
/**
SHUTDOWN_MSG: Shutting down NameNode at Monika/192.168.56.1
**/
C:\hadoop\sbin>start-dfs.cmd
C:\hadoop\sbin>jps
25540 NameNode
21400 DataNode
5976 Jps
C:\hadoop\sbin>yarn-dfs.cmdd
'yarn-dfs.cmdd' is not recognized as an internal or external command,
operable program or batch file.
C:\hadoop\sbin>start-yarn.cmd
starting yarn daemons
C:\hadoop\sbin>jps
20992 NodeManager
25540 NameNode
7604 Jps
21400 DataNode
28984 ResourceManager
C:\hadoop\sbin>hadoop fs -mkdir /input
C:\hadoop\sbin>hadoop fs -put "C:\data.txt" /input
C:\hadoop\sbin>hadoop jar 'C:\hadoop\share\hadoop\mapreduce\hadoop-mapreduce-examples-3.3.6.jar' wordcount /input /output
Exception in thread "main" java.io.IOException: The filename, directory name, or volume label syntax is incorrect
at java.io.WinNTFileSystem.canonicalize0(Native Method)
at java.io.WinNTFileSystem.canonicalize(WinNTFileSystem.java:428)
at java.io.File.getCanonicalPath(File.java:618)
at org.apache.hadoop.util.RunJar.run(RunJar.java:257)
at org.apache.hadoop.util.RunJar.main(RunJar.java:241)
C:\hadoop\sbin>hadoop jar "C:\hadoop\share\hadoop\mapreduce\hadoop-mapreduce-examples-3.3.6.jar" wordcount /input /output
2024-08-20 19:47:31,597 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-08-20 19:47:32,658 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/monik/.staging/job_1724161505817_0001
2024-08-20 19:47:33,072 INFO input.FileInputFormat: Total input files to process : 1
2024-08-20 19:47:33,173 INFO mapreduce.JobSubmitter: number of splits:1
2024-08-20 19:47:33,415 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1724161505817_0001
2024-08-20 19:47:33,415 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-08-20 19:47:33,625 INFO conf.Configuration: resource-types.xml not found
2024-08-20 19:47:33,626 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-08-20 19:47:34,323 INFO impl.YarnClientImpl: Submitted application application_1724161505817_0001
2024-08-20 19:47:34,409 INFO mapreduce.Job: The url to track the job: http://Monika:8088/proxy/application_1724161505817_0001/
2024-08-20 19:47:34,410 INFO mapreduce.Job: Running job: job_1724161505817_0001
Search
ENG US
20:45
20-08-2024


Administrator: Command Prompt
Reduce output records=5
Spilled Records=10
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=191
CPU time spent (ms)=828
Physical memory (bytes) snapshot=535609344
Virtual memory (bytes) snapshot=787865600
Total committed heap usage (bytes)=416808960
Peak Map Physical memory (bytes)=310829056
Peak Map Virtual memory (bytes)=438980608
Peak Reduce Physical memory (bytes)=240095232
Peak Reduce Virtual memory (bytes)=365191168
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=44
File Output Format Counters
Bytes Written=38
C:\hadoop\sbin>hadoop fs -cat /out/*
cat: `/out/*': No such file or directory
C:\hadoop\sbin>hadoop fs -cat /out/*
cat: `/out/*': No such file or directory
C:\hadoop\sbin>hadoop fs -cat /output/*
Hadoop 2
Hello 2
World 1
great 1
is 1
C:\hadoop\sbin>
Search
ENG US
20:46
20-08-2024


Browsing HDFS
localhost:9870/explorer.html#/output
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/output
Show 25 entries
Search:
Permission
Owner
Size
Name
-rw-r--r--
monik
MB
_SUCCESS
-rw-r--r--
monik
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-1760020507-192.168.56.1-1724161408547
Generation Stamp: 1008
Size: 38
Availability:
Monika
File contents
Hadoop 2
Hello 2
World 1
great 1
is 1
Search
ENG US
20:47
20-08-2024