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 /**

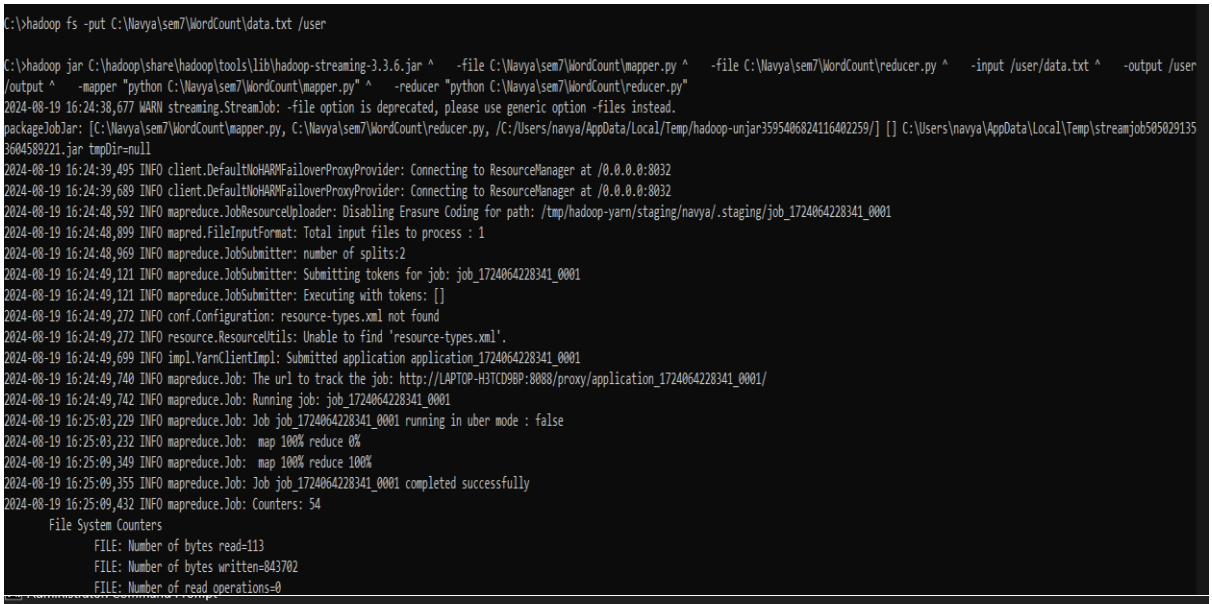
C:\> **cd hadoop**

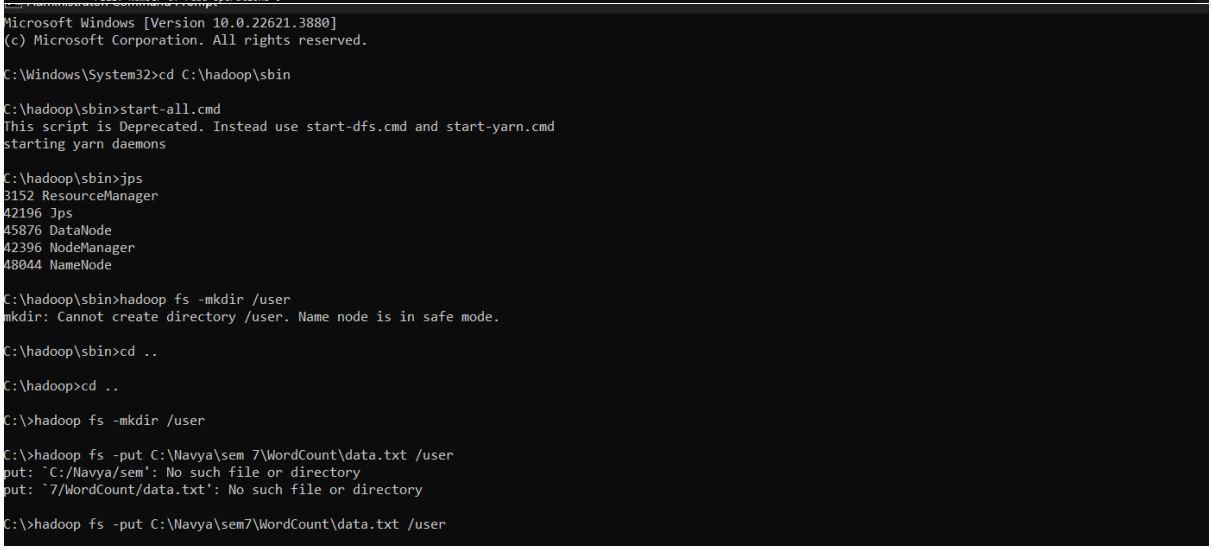
C:\hadoop> **hadoop fs -mkdir input**

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:





```
                Virtual memory (bytes) snapshot=1206562816
                Total committed heap usage (bytes)=662175744
                Peak Map Physical memory (bytes)=342204416
                Peak Map Virtual memory (bytes)=401985536
                Peak Reduce Physical memory (bytes)=241143808
                Peak Reduce Virtual memory (bytes)=405164032
        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=105
        File Output Format Counters
                Bytes Written=42
2024-08-19 16:25:09,432 INFO streaming.StreamJob: Output directory: /user/output
```