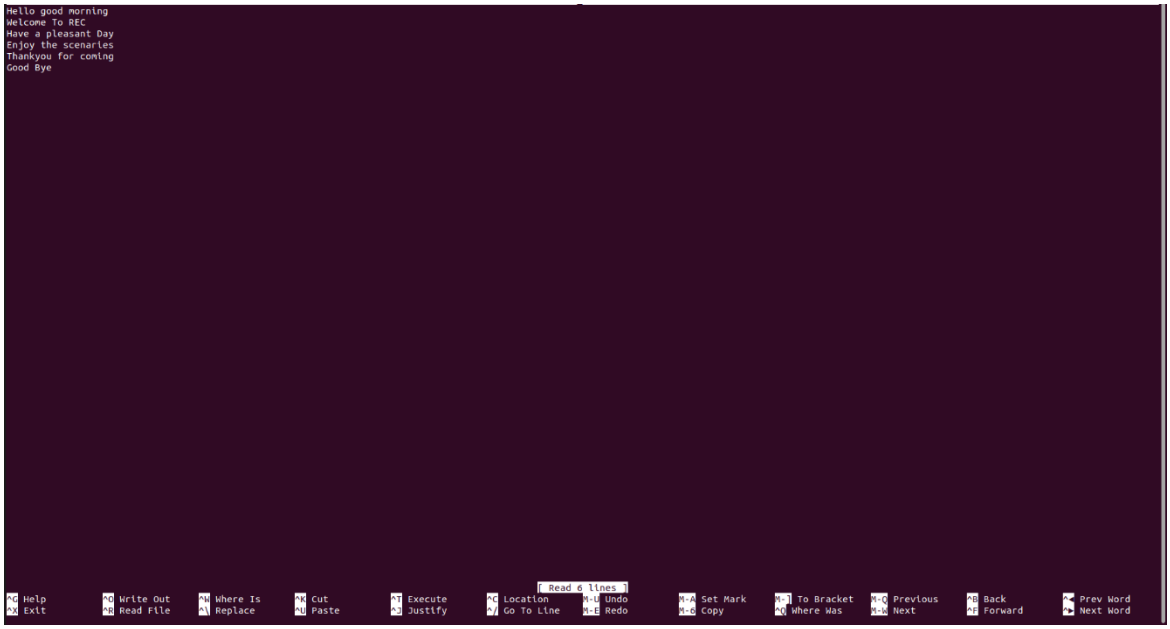
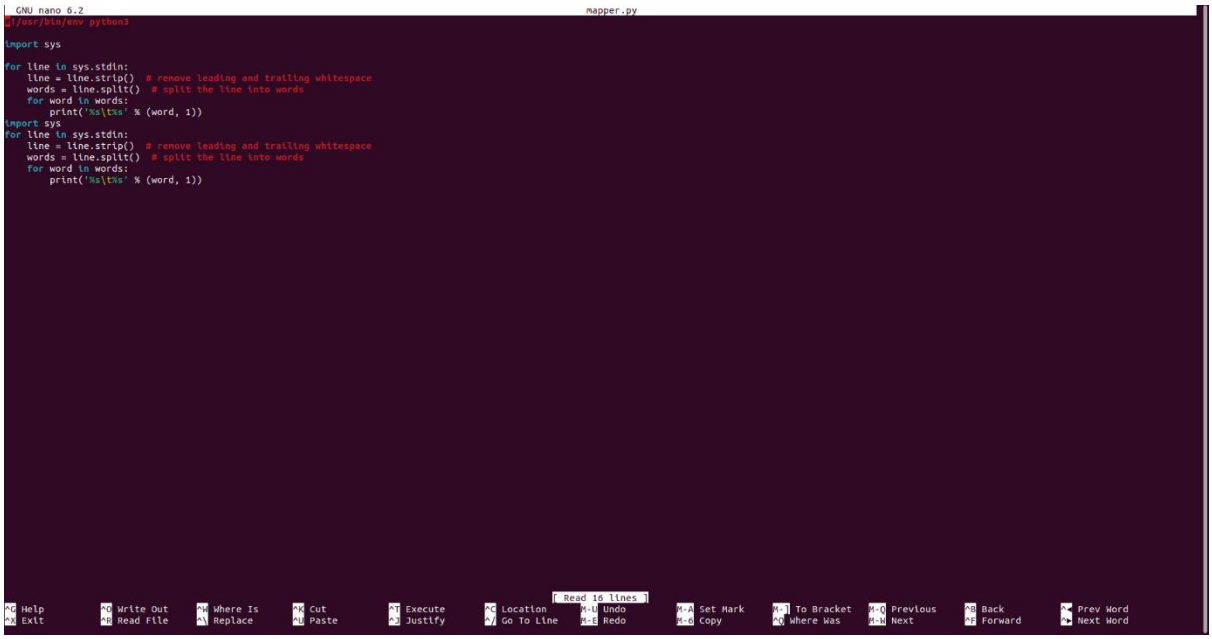
Exp. No : 2

Word Count Map Reduce program

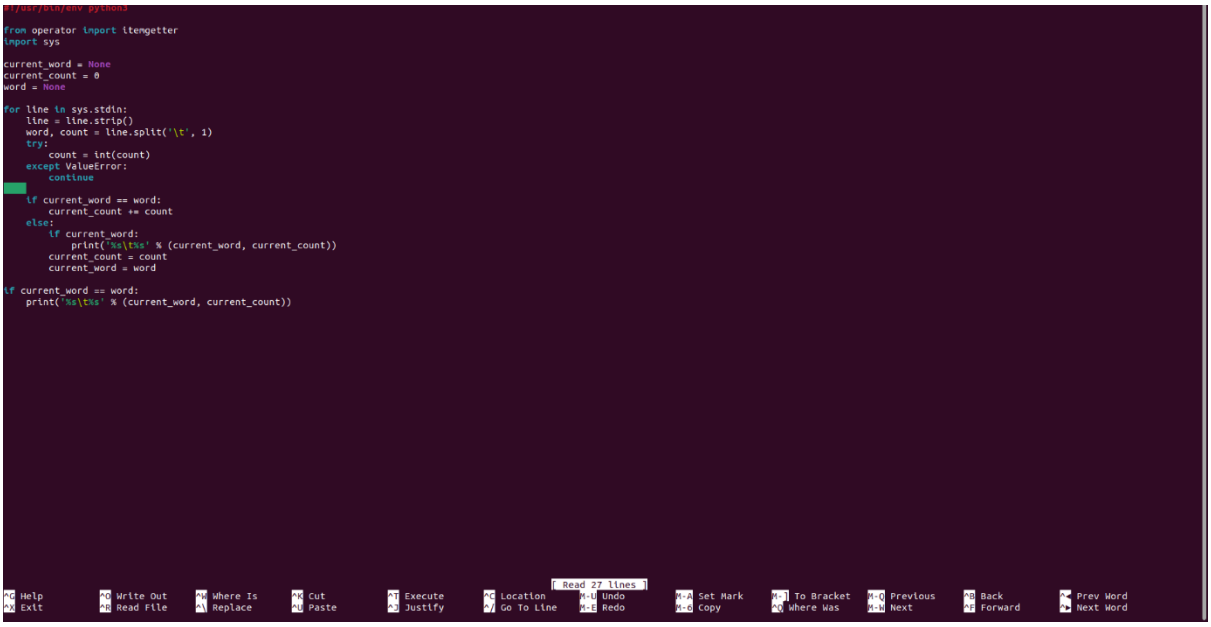
1. Create word_count.txt file



2. Create mapper.py program



3. Create reducer.py program.



4. Storing the word_count.txt in HDFS Storage.

5. Running the Word Count program using Hadoop Streaming.

```
-input /word_count_in_python/word_count.txt \
-output /word_count_in_python/new_output \
-mapper ~/mapper.py \
-reducer ~/reducer.py
2024-09-14 10:13:36,684 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2024-09-14 10:13:36,802 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2024-09-14 10:13:36,802 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2024-09-14 10:13:36,821 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2024-09-14 10:13:37,109 INFO mapred.FileInputFormat: Total input files to process : 1
2024-09-14 10:13:37,192 INFO mapreduce.JobSubmitter: number of splits:1
2024-09-14 10:13:37,333 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local682529798_0001
2024-09-14 10:13:37,333 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-09-14 10:13:37,527 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2024-09-14 10:13:37,530 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
2024-09-14 10:13:37,534 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2024-09-14 10:13:37,536 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-09-14 10:13:37,538 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2024-09-14 10:13:37,537 INFO mapreduce.Job: Running job: job_local682529798_0001
2024-09-14 10:13:37,602 INFO mapred.LocalJobRunner: Waiting for map tasks
2024-09-14 10:13:37,608 INFO mapred.LocalJobRunner: Starting task: attempt_local682529798_0001_m_000000_0
2024-09-14 10:13:37,644 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-09-14 10:13:37,645 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2024-09-14 10:13:37,679 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2024-09-14 10:13:37,692 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/word_count_in_python/word_count.txt:0+103
2024-09-14 10:13:37,728 INFO mapred.MapTask: numReduceTasks: 1
2024-09-14 10:13:37,792 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2024-09-14 10:13:37,793 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2024-09-14 10:13:37,793 INFO mapred.MapTask: soft limit at 83886080
2024-09-14 10:13:37,793 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2024-09-14 10:13:37,793 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2024-09-14 10:13:37,796 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2024-09-14 10:13:37,799 INFO streaming.PipeMapRed: PipeMapRed exec [/home/hadoop/mapper.py]
2024-09-14 10:13:37,803 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output.dir
2024-09-14 10:13:37,804 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
2024-09-14 10:13:37,804 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
2024-09-14 10:13:37,805 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.length
2024-09-14 10:13:37,805 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
2024-09-14 10:13:37,805 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
2024-09-14 10:13:37,806 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
2024-09-14 10:13:37,807 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
2024-09-14 10:13:37,807 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
2024-09-14 10:13:37,807 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
2024-09-14 10:13:37,807 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
2024-09-14 10:13:37,808 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
2024-09-14 10:13:37,945 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
2024-09-14 10:13:37,948 INFO streaming.PipeMapRed: Records R/W=6/1
2024-09-14 10:13:37,949 INFO streaming.PipeMapRed: MRErrorThread done
2024-09-14 10:13:37,950 INFO streaming.PipeMapRed: mapRedFinished
2024-09-14 10:13:37,952 INFO mapred.LocalJobRunner:
2024-09-14 10:13:37,953 INFO mapred.MapTask: Starting flush of map output
2024-09-14 10:13:37,953 INFO mapred.MapTask: Spilling map output
```

```
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Output Format Counters
                Bytes Written=139
2024-09-14 10:13:38,349 INFO mapred.LocalJobRunner: Finishing task: attempt_local682529798_0001_r_000000_0
2024-09-14 10:13:38,350 INFO mapred.LocalJobRunner: reduce task executor complete.
2024-09-14 10:13:38,547 INFO mapreduce.Job: Job job_local682529798_0001 running in uber mode : false
2024-09-14 10:13:38,549 INFO mapreduce.Job:  map 100% reduce 100%
2024-09-14 10:13:38,552 INFO mapreduce.Job: Job job_local682529798_0001 completed successfully
2024-09-14 10:13:38,563 INFO mapreduce.Job: Counters: 36
        File System Counters
                FILE: Number of bytes read=283292
                FILE: Number of bytes written=1572231
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=206
                HDFS: Number of bytes written=139
                HDFS: Number of read operations=15
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=6
                Map output records=18
                Map output bytes=139
                Map output materialized bytes=181
                Input split bytes=109
                Combine input records=0
                Combine output records=0
                Reduce input groups=18
                Reduce shuffle bytes=181
                Reduce input records=18
                Reduce output records=18
                Spilled Records=36
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=16
                Total committed heap usage (bytes)=505413632
        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=103
        File Output Format Counters
                Bytes Written=139
2024-09-14 10:13:38,563 INFO streaming.StreamJob: Output directory: /word_count_in_python/new_output
hadoop@kiran:~$ hdfs dfs -cat /word_count_in_python/new_output/part-00000
```