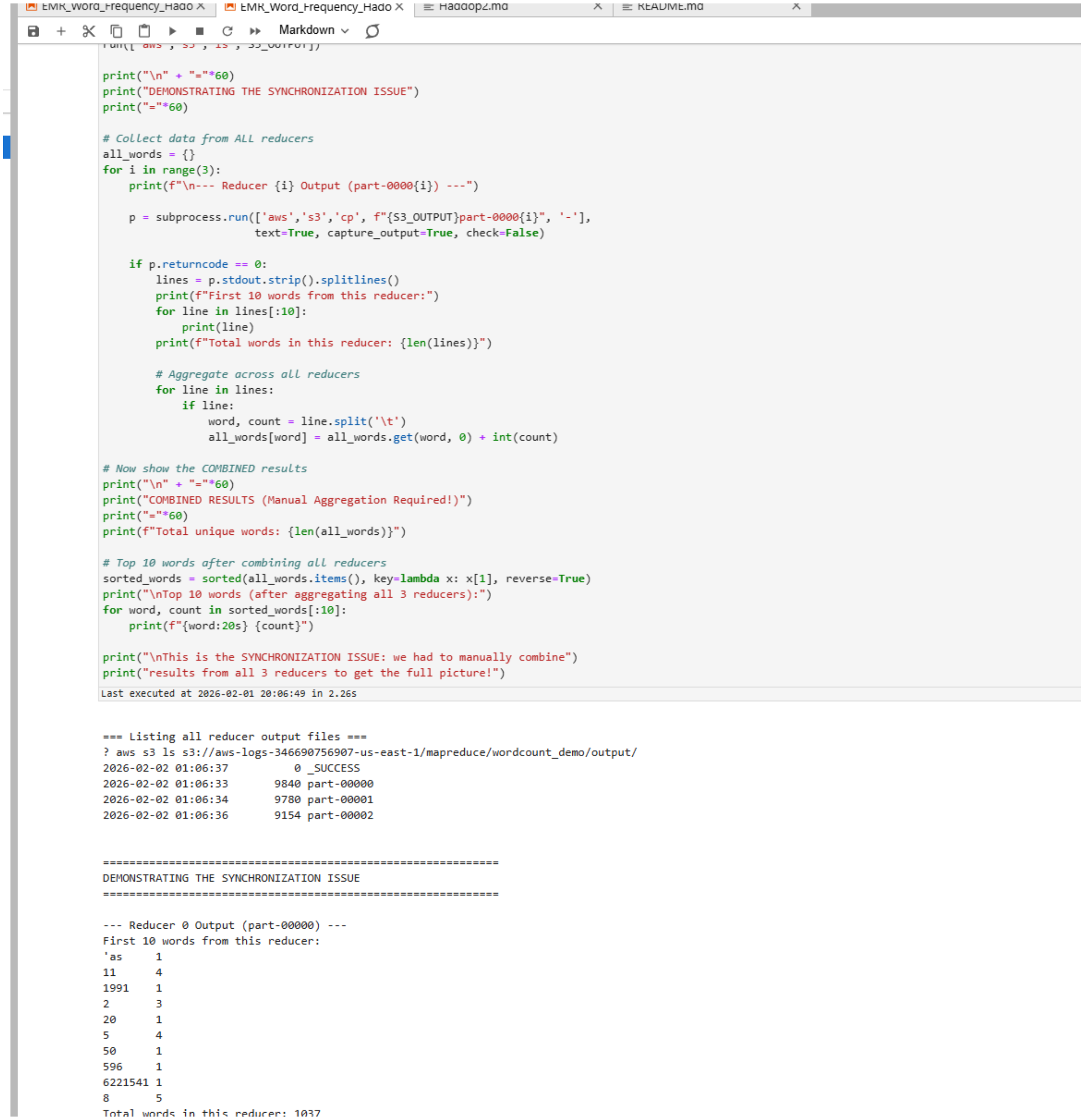

```python
run(['aws','s3','ls',S3_OUTPUT])

print("\n" + "="*60)
print("DEMONSTRATING THE SYNCHRONIZATION ISSUE")
print("="*60)

# Collect data from ALL reducers
all_words = {}
for i in range(3):
    print(f"\n--- Reducer {i} Output (part-0000{i}) ---")

    p = subprocess.run(['aws','s3','cp', f"{S3_OUTPUT}part-0000{i}", '-'],
                       text=True, capture_output=True, check=False)

    if p.returncode == 0:
        lines = p.stdout.strip().splitlines()
        print(f"First 10 words from this reducer:")
        for line in lines[:10]:
            print(line)
        print(f"Total words in this reducer: {len(lines)}")

        # Aggregate across all reducers
        for line in lines:
            if line:
                word, count = line.split('\t')
                all_words[word] = all_words.get(word, 0) + int(count)

# Now show the COMBINED results
print("\n" + "="*60)
print("COMBINED RESULTS (Manual Aggregation Required!)")
print("="*60)
print(f"Total unique words: {len(all_words)}")

# Top 10 words after combining all reducers
sorted_words = sorted(all_words.items(), key=lambda x: x[1], reverse=True)
print("\nTop 10 words (after aggregating all 3 reducers):")
for word, count in sorted_words[:10]:
    print(f"{word:20s} {count}")

print("\nThis is the SYNCHRONIZATION ISSUE: we had to manually combine")
print("results from all 3 reducers to get the full picture!")
```

Last executed at 2026-02-01 20:06:49 in 2.26s

```
=== Listing all reducer output files ===
? aws s3 ls s3://aws-logs-346690756907-us-east-1/mapreduce/wordcount_demo/output/
2026-02-02 01:06:37          0 _SUCCESS
2026-02-02 01:06:33       9840 part-00000
2026-02-02 01:06:34       9780 part-00001
2026-02-02 01:06:36       9154 part-00002


============================================================
DEMONSTRATING THE SYNCHRONIZATION ISSUE
============================================================

--- Reducer 0 Output (part-00000) ---
First 10 words from this reducer:
'as     1
11      4
1991    1
2       3
20      1
5       4
50      1
596     1
6221541 1
8       5
Total words in this reducer: 1037
```

```
--- Reducer 2 Output (part-00002) ---
First 10 words from this reducer:
1       47
2020    1
4       5
64      1
7       3
a       695
absence 1
accept  1
accessed        1
accessible      1
Total words in this reducer: 974

============================================================
COMBINED RESULTS (Manual Aggregation Required!)
============================================================
Total unique words: 3048

Top 10 words (after aggregating all 3 reducers):
the                  1839
and                  942
to                   811
a                    695
of                   638
it                   610
she                  553
i                    546
you                  486
said                 462

This is the SYNCHRONIZATION ISSUE: we had to manually combine
results from all 3 reducers to get the full picture!
```

[ ]:

```
## 8) Experiment with Combiner

A combiner pre-aggregates data on mapper nodes before sending to reducers.
This reduces network traffic and speeds up the job.
```

[14]:

```python
# Add new S3 output path at the top of your notebook (in Section 0)
S3_OUTPUT_COMBINER = f"{S3_BASE}/output_combiner/"

# Run job with combiner
run(['aws','s3','rm', S3_OUTPUT_COMBINER, '--recursive'], check=False)

cmd = [
    'hadoop','jar', str(STREAMING_JAR),
    '-D','mapreduce.job.name=wordcount-with-combiner',
    '-D','mapreduce.job.reduces=3',
    '-files','mapper.py,reducer.py',
    '-mapper','mapper.py',
    '-reducer','reducer.py',
    '-combiner','reducer.py',  # THIS IS THE NEW LINE!
    '-input', S3_INPUT,
    '-output', S3_OUTPUT_COMBINER,
]
run(cmd)
```

Last executed at 2026-02-01 20:12:02 in 29.35s

```
? aws s3 rm s3://aws-logs-346690756907-us-east-1/mapreduce/wordcount_demo/output_combiner/ --recursive
delete: s3://aws-logs-346690756907-us-east-1/mapreduce/wordcount_demo/output_combiner/_SUCCESS
delete: s3://aws-logs-346690756907-us-east-1/mapreduce/wordcount_demo/output_combiner/part-00001
delete: s3://aws-logs-346690756907-us-east-1/mapreduce/wordcount_demo/output_combiner/part-00002
delete: s3://aws-logs-346690756907-us-east-1/mapreduce/wordcount_demo/output_combiner/part-00000
```

```
alized bytes=57700\n\t\tInput split bytes=1160\n\t\tCombine input records=30684\n\t\tCombine output records=6786\n\t\tReduce input groups=3048\n\t\tRed
\t\tFailed Shuffles=0\n\t\tMerged Map outputs=27\n\t\tGC time elapsed (ms)=389\n\t\tCPU time spent (ms)=13460\n\t\tPhysical memory (bytes) snapshot=663
(bytes)=617000960\n\t\tPeak Map Virtual memory (bytes)=7036334080\n\t\tPeak Reduce Physical memory (bytes)=406142976\n\t\tPeak Reduce Virtual memory (b
=0\n\tFile Input Format Counters \n\t\tBytes Read=342592\n\tFile Output Format Counters \n\t\tBytes Written=28774\n2026-02-02 01:12:01,588 INFO streami
FO impl.MetricsSystemImpl: Stopping s3a-file-system metrics system...\n2026-02-02 01:12:01,592 INFO impl.MetricsSystemImpl: s3a-file-system metrics sys
```

[15]:

```python
print("=== Combiner Job Results ===")
run(['aws','s3','ls', S3_OUTPUT_COMBINER])

# Aggregate combiner results
combiner_words = {}
for i in range(3):
    p = subprocess.run(['aws','s3','cp', f"{S3_OUTPUT_COMBINER}part-0000{i}", '-'],
                       text=True, capture_output=True, check=False)
    if p.returncode == 0:
        for line in p.stdout.strip().splitlines():
            if line:
                word, count = line.split('\t')
                combiner_words[word] = combiner_words.get(word, 0) + int(count)

print(f"Total unique words with combiner: {len(combiner_words)}")

sorted_combiner = sorted(combiner_words.items(), key=lambda x: x[1], reverse=True)
print("\nTop 10 words with combiner:")
for word, count in sorted_combiner[:10]:
    print(f"{word:20s} {count}")

print(f"\nResults match without combiner: {all_words == combiner_words}")
print("The combiner gives SAME results but with better performance!")
```

Last executed at 2026-02-01 20:12:09 in 2.25s

```
=== Combiner Job Results ===
? aws s3 ls s3://aws-logs-346690756907-us-east-1/mapreduce/wordcount_demo/output_combiner/
2026-02-02 01:12:00          0 _SUCCESS
2026-02-02 01:11:56       9840 part-00000
2026-02-02 01:11:56       9780 part-00001
2026-02-02 01:12:00       9154 part-00002

Total unique words with combiner: 3048

Top 10 words with combiner:
the                  1839
and                  942
to                   811
a                    695
of                   638
it                   610
she                  553
i                    546
you                  486
said                 462

Results match without combiner: True
The combiner gives SAME results but with better performance!
```

Troubleshooting