UNIT 2 – Data Storage Technologies

Activity 2: MapReduce Optimization Experiment

Task 1: Program Development

1.Develop a MapReduce application for NOAA weather datasets analysis

2.Implement mapper to parse weather records into key-value pairs

3.Implement reducer to aggregate weather metrics

4.Inclusion of combiners when beneficial

Task 2: Optimization Experiments Trials on Hadoop

1.Execution of multiple trial runs, Vary the number of mappers and reducers.

2. Enable or disable combiners to observe differences.

3. Adjust block sizes or input split configurations.

4.Compare execution times and resource usage before and after tuning.

Task 3: Performance Monitoring and Analysis

1.Monitor job execution using Hadoop JobHistoryServer

2.Track time per phase (map, shuffle, reduce)

3.Record number of mappers and reducers allocated

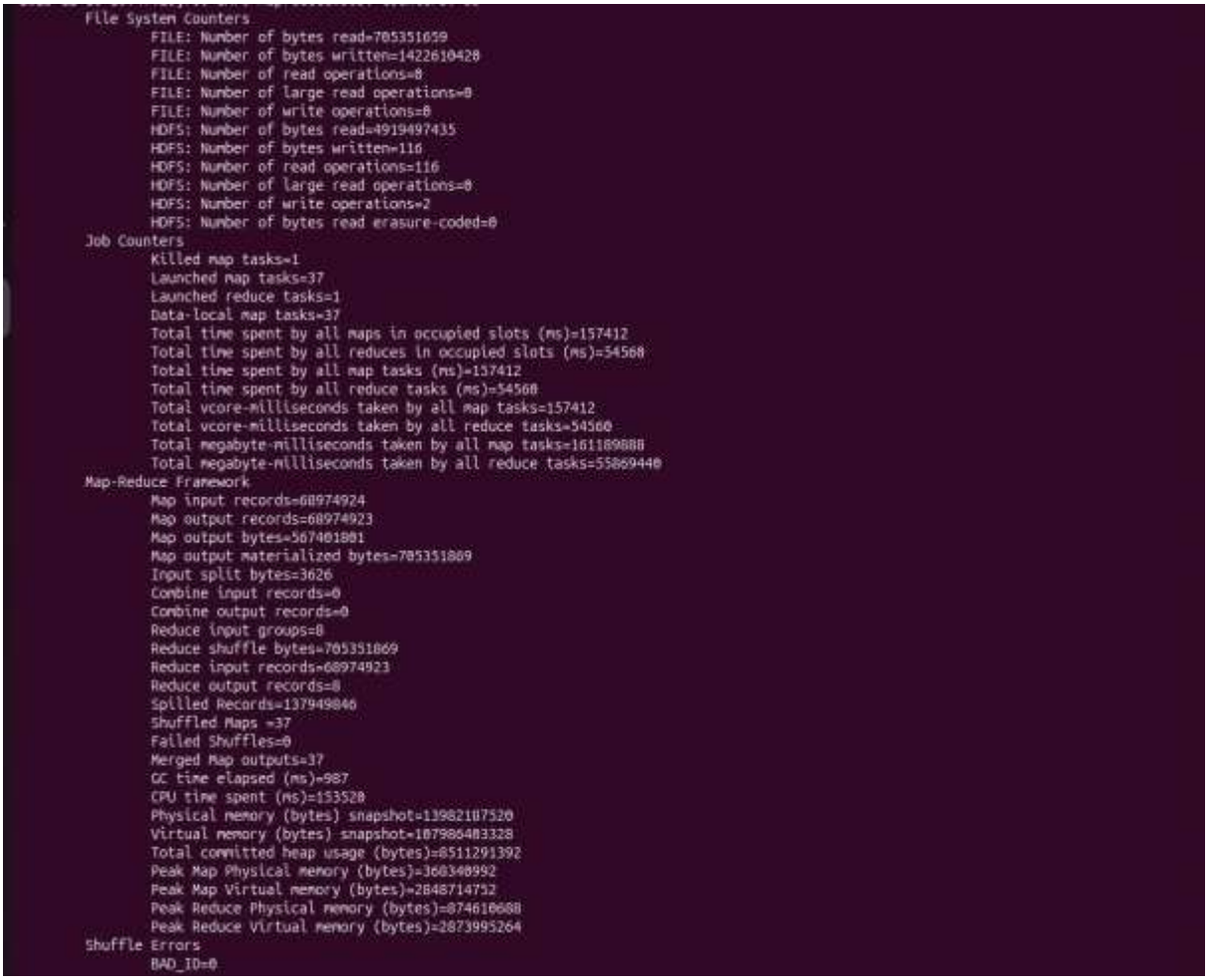
4.Identify performance bottlenecks (CPU, I/O, network)

5.Analyze logs for optimization opportunities

When Mapper block size is 128 mb with and without Combiner

```
2025-12-16 10:03:18,996 INFO streaming.StreamJob: Output directory: /cloud/output/Pinky
hduser@master:~$ hdfs dfs -cat /cloud/output/Pinky
cat: `/cloud/output/Pinky': Is a directory
hduser@master:~$ hdfs dfs -ls /cloud/output/Pinky
Found 2 items
-rw-r--r--   2 hduser supergroup          0 2025-12-16 10:03 /cloud/output/Pinky/_SUCCESS
-rw-r--r--   2 hduser supergroup        101 2025-12-16 10:03 /cloud/output/Pinky/part-00000
hduser@master:~$ hdfs dfs -cat /cloud/output/Pinky/part-00000
clear   3432
cloudy  3797
drizzle 166
fog     154
haze    16
rain    664
snow    527
snow pellets    1
thunderstorms   27
hduser@master:~$
```

Without Combiner:-



With Combiner:-

```
File System Counters
        FILE: Number of bytes read=4486
        FILE: Number of bytes written=11941418
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=4919497435
        HDFS: Number of bytes written=116
        HDFS: Number of read operations=116
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
        HDFS: Number of bytes read erasure-coded=0
Job Counters
        Killed map tasks=1
        Launched map tasks=37
        Launched reduce tasks=1
        Data-local map tasks=37
        Total time spent by all maps in occupied slots (ms)=159013
        Total time spent by all reduces in occupied slots (ms)=21267
        Total time spent by all map tasks (ms)=159013
        Total time spent by all reduce tasks (ms)=21267
        Total vcore-milliseconds taken by all map tasks=159013
        Total vcore-milliseconds taken by all reduce tasks=21267
        Total megabyte-milliseconds taken by all map tasks=162829312
        Total megabyte-milliseconds taken by all reduce tasks=21777408
Map-Reduce Framework
        Map input records=68974924
        Map output records=68974923
        Map output bytes=567401081
        Map output materialized bytes=4622
        Input split bytes=3626
        Combine input records=68974923
        Combine output records=296
        Reduce input groups=8
        Reduce shuffle bytes=4622
        Reduce input records=296
        Reduce output records=8
        Spilled Records=592
        Shuffled Maps =37
        Failed Shuffles=0
        Merged Map outputs=37
        GC time elapsed (ms)=768
        CPU time spent (ms)=106870
        Physical memory (bytes) snapshot=13448446464
        Virtual memory (bytes) snapshot=108022033152
        Total committed heap usage (bytes)=8010072064
        Peak Map Physical memory (bytes)=372330496
        Peak Map Virtual memory (bytes)=2848497664
        Peak Reduce Physical memory (bytes)=219353088
        Peak Reduce Virtual memory (bytes)=2846709152
Shuffle Errors
        BAD_ID=0
```

When Mapper block size is 64 mb with and without Combiner

```
hduser@master:~$ hadoop fs -cat /cloud/output/Pinky4/part-00000
clear    26949008
cloudy   29815349
drizzle 400466
fog      3345100
haze     125642
rain     4397305
snow     3926349
thunderstorms    15704
```

```
        File System Counters
                FILE: Number of bytes read=4406
                FILE: Number of bytes written=11948392
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=4919497435
                HDFS: Number of bytes written=116
                HDFS: Number of read operations=116
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=37
                Launched reduce tasks=1
                Data-local map tasks=37
                Total time spent by all maps in occupied slots (ms)=177766
                Total time spent by all reduces in occupied slots (ms)=9884
                Total time spent by all map tasks (ms)=177766
                Total time spent by all reduce tasks (ms)=9884
                Total vcore-milliseconds taken by all map tasks=177766
                Total vcore-milliseconds taken by all reduce tasks=9884
                Total megabyte-milliseconds taken by all map tasks=182032384
                Total megabyte-milliseconds taken by all reduce tasks=10121216
        Map-Reduce Framework
                Map input records=68974924
                Map output records=68974923
                Map output bytes=567401801
                Map output materialized bytes=4622
                Input split bytes=3626
                Combine input records=68974923
                Combine output records=296
                Reduce input groups=8
                Reduce shuffle bytes=4622
                Reduce input records=296
                Reduce output records=8
                Spilled Records=592
                Shuffled Maps =37
                Failed Shuffles=0
                Merged Map outputs=37
                GC time elapsed (ms)=1097
                CPU time spent (ms)=111720
                Physical memory (bytes) snapshot=13345013760
                Virtual memory (bytes) snapshot=107956510720
                Total committed heap usage (bytes)=8804829184
                Peak Map Physical memory (bytes)=378601984
                Peak Map Virtual memory (bytes)=2844217344
                Peak Reduce Physical memory (bytes)=221843456
                Peak Reduce Virtual memory (bytes)=2850045952
```

When Reducer is 1 with and without Combiner

```
2025-12-16 10:03:18,996 INFO streaming.StreamJob: Output directory: /cloud/output/Pinky
hduser@master:~$ hdfs dfs -cat /cloud/output/Pinky
cat: `/cloud/output/Pinky': Is a directory
hduser@master:~$ hdfs dfs -ls /cloud/output/Pinky
Found 2 items
-rw-r--r--   2 hduser supergroup          0 2025-12-16 10:03 /cloud/output/Pinky/_SUCCESS
-rw-r--r--   2 hduser supergroup        101 2025-12-16 10:03 /cloud/output/Pinky/part-00000
hduser@master:~$ hdfs dfs -cat /cloud/output/Pinky/part-00000
clear   3432
cloudy  3797
drizzle 166
fog     154
haze    16
rain    664
snow    527
snow pellets    1
thunderstorms   27
hduser@master:~$
```

Without Combiner:-

```
        File System Counters
                FILE: Number of bytes read=705351059
                FILE: Number of bytes written=1422610428
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=4919497435
                HDFS: Number of bytes written=116
                HDFS: Number of read operations=116
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=37
                Launched reduce tasks=1
                Data-local map tasks=37
                Total time spent by all maps in occupied slots (ms)=157412
                Total time spent by all reduces in occupied slots (ms)=54560
                Total time spent by all map tasks (ms)=157412
                Total time spent by all reduce tasks (ms)=54560
                Total vcore-milliseconds taken by all map tasks=157412
                Total vcore-milliseconds taken by all reduce tasks=54560
                Total megabyte-milliseconds taken by all map tasks=161189888
                Total megabyte-milliseconds taken by all reduce tasks=55869440
        Map-Reduce Framework
                Map input records=68974924
                Map output records=68974923
                Map output bytes=567401801
                Map output materialized bytes=705351869
                Input split bytes=3626
                Combine input records=0
                Combine output records=0
                Reduce input groups=8
                Reduce shuffle bytes=705351869
                Reduce input records=68974923
                Reduce output records=8
                Spilled Records=137949846
                Shuffled Maps =37
                Failed Shuffles=0
                Merged Map outputs=37
                GC time elapsed (ms)=987
                CPU time spent (ms)=153520
                Physical memory (bytes) snapshot=13982187520
                Virtual memory (bytes) snapshot=107986483328
                Total committed heap usage (bytes)=8511291392
                Peak Map Physical memory (bytes)=360340992
                Peak Map Virtual memory (bytes)=2848714752
                Peak Reduce Physical memory (bytes)=874610688
                Peak Reduce Virtual memory (bytes)=2873995264
        Shuffle Errors
                BAD_ID=0
```

With Combiner:-

```
File System Counters
        FILE: Number of bytes read=4486
        FILE: Number of bytes written=1194141B
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=4919497435
        HDFS: Number of bytes written=116
        HDFS: Number of read operations=116
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
        HDFS: Number of bytes read erasure-coded=0
Job Counters
        Killed map tasks=1
        Launched map tasks=37
        Launched reduce tasks=1
        Data-local map tasks=37
        Total time spent by all maps in occupied slots (ms)=159013
        Total time spent by all reduces in occupied slots (ms)=21267
        Total time spent by all map tasks (ms)=159013
        Total time spent by all reduce tasks (ms)=21267
        Total vcore-milliseconds taken by all map tasks=159013
        Total vcore-milliseconds taken by all reduce tasks=21267
        Total megabyte-milliseconds taken by all map tasks=162829312
        Total megabyte-milliseconds taken by all reduce tasks=21777408
Map-Reduce Framework
        Map input records=68974924
        Map output records=68974923
        Map output bytes=567401081
        Map output materialized bytes=4622
        Input split bytes=3626
        Combine input records=68974923
        Combine output records=296
        Reduce input groups=8
        Reduce shuffle bytes=4622
        Reduce input records=296
        Reduce output records=8
        Spilled Records=592
        Shuffled Maps =37
        Failed Shuffles=0
        Merged Map outputs=37
        GC time elapsed (ms)=768
        CPU time spent (ms)=106870
        Physical memory (bytes) snapshot=13448446464
        Virtual memory (bytes) snapshot=108022033152
        Total committed heap usage (bytes)=8010072064
        Peak Map Physical memory (bytes)=372338496
        Peak Map Virtual memory (bytes)=2848497664
        Peak Reduce Physical memory (bytes)=219353088
        Peak Reduce Virtual memory (bytes)=2846709152
Shuffle Errors
        BAD_ID=0
```

When Reducer is 2 Without Combiner and With Combiner

```
hduser@master:~$ hadoop fs -cat /cloud/output/Pinky3/part-00000
clear   26949008
cloudy  29815349
drizzle 400466
fog     3345100
haze    125642
rain    4397305
snow    3926349
thunderstorms   15704
```

Without Combiner:-

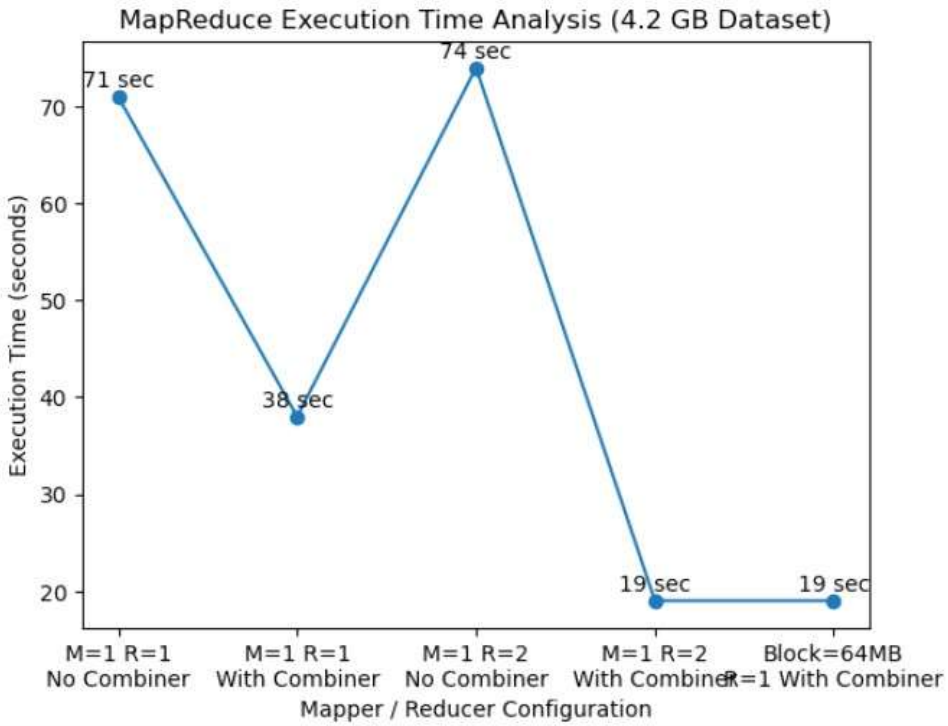

```
        File System Counters
                FILE: Number of bytes read=705351659
                FILE: Number of bytes written=1422923957
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=4919497435
                HDFS: Number of bytes written=116
                HDFS: Number of read operations=121
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Killed reduce tasks=1
                Launched map tasks=37
                Launched reduce tasks=3
                Data-local map tasks=37
                Total time spent by all maps in occupied slots (ms)=170737
                Total time spent by all reduces in occupied slots (ms)=128793
                Total time spent by all map tasks (ms)=170737
                Total time spent by all reduce tasks (ms)=128793
                Total vcore-milliseconds taken by all map tasks=170737
                Total vcore-milliseconds taken by all reduce tasks=128793
                Total megabyte-milliseconds taken by all map tasks=174834688
                Total megabyte-milliseconds taken by all reduce tasks=123692032
        Map-Reduce Framework
                Map input records=68974924
                Map output records=68974923
                Map output bytes=567401081
                Map output materialized bytes=705352091
                Input split bytes=3626
                Combine input records=0
                Combine output records=0
                Reduce input groups=8
                Reduce shuffle bytes=705352091
                Reduce input records=68974923
                Reduce output records=8
                Spilled Records=137949846
                Shuffled Maps =74
                Failed Shuffles=0
                Merged Map outputs=74
                GC time elapsed (ms)=1209
                CPU time spent (ms)=167980
                Physical memory (bytes) snapshot=14081712128
                Virtual memory (bytes) snapshot=110012504064
                Total committed heap usage (bytes)=8624537600
                Peak Map Physical memory (bytes)=370929664
                Peak Map Virtual memory (bytes)=2046994432
                Peak Reduce Physical memory (bytes)=619884544
                Peak Reduce Virtual memory (bytes)=2876809216
        Shuffle Errors
```

With Combiner:-


```
025-12-16 17:09:54,295 INFO mapreduce.Job: Counters: 55
        File System Counters
                FILE: Number of bytes read=4412
                FILE: Number of bytes written=12255359
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=4919497435
                HDFS: Number of bytes written=116
                HDFS: Number of read operations=121
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=37
                Launched reduce tasks=2
                Data-local map tasks=37
                Total time spent by all maps in occupied slots (ms)=181442
                Total time spent by all reduces in occupied slots (ms)=13732
                Total time spent by all map tasks (ms)=181442
                Total time spent by all reduce tasks (ms)=13732
                Total vcore-milliseconds taken by all map tasks=181442
                Total vcore-milliseconds taken by all reduce tasks=13732
                Total megabyte-milliseconds taken by all map tasks=185796608
                Total megabyte-milliseconds taken by all reduce tasks=14061568
        Map-Reduce Framework
                Map input records=68974924
                Map output records=68974923
                Map output bytes=567481801
                Map output materialized bytes=4844
                Input split bytes=3626
                Combine input records=68974923
                Combine output records=296
                Reduce input groups=0
                Reduce shuffle bytes=4844
                Reduce input records=296
                Reduce output records=0
                Spilled Records=592
                Shuffled Maps =74
                Failed Shuffles=0
                Merged Map outputs=74
                GC time elapsed (ms)=1896
                CPU time spent (ms)=123070
                Physical memory (bytes) snapshot=13700193792
                Virtual memory (bytes) snapshot=110851907584
                Total committed heap usage (bytes)=8096055296
                Peak Map Physical memory (bytes)=373399552
                Peak Map Virtual memory (bytes)=2851651584
                Peak Reduce Physical memory (bytes)=225587200
                Peak Reduce Virtual memory (bytes)=2849181696
        Shuffle Errors
```

Task 4: Visualization and Reporting

1.Graphs comparing execution times



2.Visual summaries of mapper/reducer variations

Mapper	Reducer	Combiner	Execution Time
1	1	No	1min 11 sec
1	1	Yes	38sec
1	2	No	1min 14 sec
1	2	Yes	19 sec
Block size = 128mb	1	No	1min 11 sec
Block size = 128mb	1	Yes	38sec
Block size = 64 mb	1	Yes	19 sec

Pipeline:-

Input --> Mapper --> Combiner --> Shuffle & Sort --> Reducer --> Output

3.A structured 300-word document stored on GitHub

4.Inclusion of experiment notes, visuals, and program code

Link= https://github.com/rajsanodiya122/Fobd-Unit-2-Act2

Short Summary :-

When Mapper = 1 :- The input dayaset is processed as a single map task. Data processing is **sequential**, not parallel. CPU and cluster resources are **underutilized**. Execution time is **higher** for large datasets (like my 4.2 GB data). Network and reducer stages wait for one mapper to finish.

When Mapper = 2 :- The data is split into two parts, each part is processed in parallel. Improved **parallel processing**. Better **CPU and I/O utilization**. Reduced mapper execution time. Faster availability of intermediate data for reducers

When Reducer = 1:- All intermediate key-value pairs produced by mappers are sent to a **single reduce task**. Data aggregation happens **sequentially.** Shuffle phase sends **all intermediate data to one node.** Can become a **bottleneck** for large datasets**.** Simpler output handling (single output file)

In my experiment:
For a 4.2 GB dataset **without a combiner**, reducer = 1 completed in **71 seconds**, showing moderate performance but limited scalability.

When **Reducer = 2:-** Intermediate data is **partitioned** and processed in **parallel** by two reduce tasks. Improved **parallelism** in the reduce phase**.** Reduced load on a single reducer**.** Multiple output files are generated**.** Extra overhead due to partitioning and task coordination

In my experiment:
Without a combiner, reducer = 2 took **74 seconds**, slightly **more than reducer = 1**, because the overhead of managing an extra reducer outweighed the benefit for this dataset size.

Without Combiner:- All intermediate key–value pairs generated by the mappers are **directly sent to the reducer** through the shuffle phase.Large volume of intermediate data**.** High **network I/O during shuffle.** Reducers take longer to aggregate data**.** Execution time increases significantly**.** Cluster resources are less efficiently utilized

In my experiment:

Without a combiner, execution time was **71–74 seconds**, even with different reducer configurations, due to heavy data transfer between mapper and reducer.

With Combiner:- A **combiner** performs **local aggregation at the mapper level** before data is sent to reducers. Intermediate data size is **greatly reduced.** Shuffle phase becomes much faster**.** Network traffic is minimized**.** Reducers process fewer records**.** Overall execution time is drastically reduced

In my experiment:
With a combiner enabled, execution time dropped to **38 seconds** and further to **19 seconds**, showing a **major performance improvement**.