#mapper

```
hadoop@ubuntu22:~$ cat mapp.py
import sys
for line in sys.stdin:
    line = line.strip()
    parts = line.split(",")
    if len(parts) == 2:
        year, temp = parts
        print(f"{year}\t{temp}")
```

#reducer

```
hadoop@ubuntu22:~$ cat redu.py
import sys
dmax = {}
dmin = {}
for line in sys.stdin:
    line = line.strip()
    parts = line.split("\t")
    if len(parts) == 2:
        year = int(parts[0])
        temp = int(parts[1])
        if year not in dmax:
            dmax[year] = temp
            dmin[year] = temp
        else:
            dmax[year] = max(dmax[year], temp)
            dmin[year] = min(dmin[year], temp)
print("Year,MaxTemp,MinTemp")
for year in sorted(dmax.keys()):
    print(f"{year},{dmax[year]},{dmin[year]}")
```

#data

```
2005   39
2006   39
2007   46
2008   45
2009   34
2010   38
2011   30
2012   22
2013   34
2014   38
2015   31
2016   14
2017   27
2018   39
2019   40
```

2020	36
2021	34
2022	37
2023	39
2024	40
2025	27
2008	30
2012	29
2013	0
2010	37
2015	60
2017	60
2009	60
2006	43
2007	38
2024	32
2019	32
2018	21
2011	6
2020	36
2023	41
2005	40
2006	58
2016	56
2025	29
2014	39
2008	38
2010	42
2019	47
2021	27
2022	35
2024	44
2017	44
2006	46
2013	47
2018	54
2011	55
2020	44
2024	45
2009	44
2014	49
2017	55
2015	46
2012	39
2007	51
2023	51
2019	42
2025	14
2011	28
2022	-10
2008	27
2016	14

```
 2014   26
 2021   38
 2006   34
 2013   37
 2020   37
 2018   63
 2024   76
 2007   56
 2005   56
 2006   79
 2007   76
 2008   65
 2009   48
 2010   53
 2011   66
 2012   69
 2013   74
 2014   51
 2015   45
 2016   42
 2017   48
 2018   47
 2019   56
 2020   73
 2021   73
 2022   71
 2023   72
 2024   70
 2025   15
 2008   21
 2012    8
 2013   25
```

#code

```
hadoop@ubuntu22:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [ubuntu22]
Starting resourcemanager
Starting nodemanagers
hadoop@ubuntu22:~$ jps
10689 Jps
9970 SecondaryNameNode
10323 NodeManager
9715 DataNode
9561 NameNode
10189 ResourceManager
```

```
hadoop@ubuntu22:~$ hdfs dfs -put weathertext.csv  /160122771091
hadoop@ubuntu22:~$ hdfs dfs -put mapp.py  /160122771091
hadoop@ubuntu22:~$ hdfs dfs -put redu.py  /160122771091
hadoop@ubuntu22:~$ hdfs dfs -ls  /160122771091
Found 7 items
-rw-r--r--   1 hadoop supergroup        170 2025-02-04 15:23 /160122771091/mapp.py
-rw-r--r--   1 hadoop supergroup        104 2025-01-28 14:45 /160122771091/mapper.py
drwxr-xr-x   - hadoop supergroup          0 2025-01-28 15:10 /160122771091/output3.txt
-rw-r--r--   1 hadoop supergroup        498 2025-02-04 15:24 /160122771091/redu.py
-rw-r--r--   1 hadoop supergroup        392 2025-01-28 14:46 /160122771091/reducer.py
-rw-r--r--   1 hadoop supergroup        796 2025-02-04 15:23 /160122771091/weathertext.csv
-rw-r--r--   1 hadoop supergroup         47 2025-01-28 14:38 /160122771091/wordcount.txt
hadoop@ubuntu22:~$ hadoop jar /home/hadoop/hadoop/share/hadoop/tools/lib/hadoop-
streaming-3.3.6.jar -file mapp.py -mapp "python3 mapp.py" -file redu.py -reducer "python3
redu.py" -input /160122771091/weathertext.csv -output /160122771091/output.txt
2025-02-04 15:25:52,728 WARN streaming.StreamJob: -file option is deprecated, please use
generic option -files instead.
File: file:/home/hadoop/-mapp does not exist.
Try -help for more information
Streaming Command Failed!
hadoop@ubuntu22:~$ hadoop jar /home/hadoop/hadoop/share/hadoop/tools/lib/hadoop-streaming-
3.3.6.jar -file mapp.py -mapper "python3 mapp.py" -file redu.py -reducer "python3 redu.py" -input
/160122771091/weathertext.csv -output /160122771091/output.txt
2025-02-04 15:26:50,741 WARN streaming.StreamJob: -file option is deprecated, please use
generic option -files instead.
packageJobJar: [mapp.py, redu.py] [] /tmp/streamjob1418201042988821765.jar tmpDir=null
2025-02-04 15:26:51,733 INFO impl.MetricsConfig: Loaded properties from hadoop-
metrics2.properties
2025-02-04 15:26:51,869 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10
second(s).
2025-02-04 15:26:51,869 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2025-02-04 15:26:51,881 WARN impl.MetricsSystemImpl: JobTracker metrics system already
initialized!
2025-02-04 15:26:52,158 INFO mapred.FileInputFormat: Total input files to process : 1
2025-02-04 15:26:52,248 INFO mapreduce.JobSubmitter: number of splits:1
2025-02-04 15:26:52,502 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_local1040376759_0001
2025-02-04 15:26:52,503 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-02-04 15:26:52,775 INFO mapred.LocalDistributedCacheManager: Localized
file:/home/hadoop/mapp.py as file:/tmp/hadoop-
hadoop/mapred/local/job_local1040376759_0001_042abce3-9b33-48ae-8343-
ab8b9edab250/mapp.py
2025-02-04 15:26:52,795 INFO mapred.LocalDistributedCacheManager: Localized
file:/home/hadoop/redu.py as file:/tmp/hadoop-
hadoop/mapred/local/job_local1040376759_0001_dcacf153-2ee6-4074-9210-
2c1b4a308ed3/redu.py
2025-02-04 15:26:52,928 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2025-02-04 15:26:52,929 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2025-02-04 15:26:52,930 INFO mapreduce.Job: Running job: job_local1040376759_0001
2025-02-04 15:26:52,931 INFO mapred.LocalJobRunner: OutputCommitter is
org.apache.hadoop.mapred.FileOutputCommitter
```

```
2025-02-04 15:26:52,938 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-02-04 15:26:52,939 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-02-04 15:26:53,005 INFO mapred.LocalJobRunner: Waiting for map tasks
2025-02-04 15:26:53,007 INFO mapred.LocalJobRunner: Starting task: attempt_local1040376759_0001_m_000000_0
2025-02-04 15:26:53,035 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-02-04 15:26:53,035 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-02-04 15:26:53,058 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2025-02-04 15:26:53,078 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/160122771091/weathertext.csv:0+796
2025-02-04 15:26:53,106 INFO mapred.MapTask: numReduceTasks: 1
2025-02-04 15:26:53,153 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2025-02-04 15:26:53,153 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2025-02-04 15:26:53,153 INFO mapred.MapTask: soft limit at 83886080
2025-02-04 15:26:53,153 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2025-02-04 15:26:53,153 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2025-02-04 15:26:53,157 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2025-02-04 15:26:53,162 INFO streaming.PipeMapRed: PipeMapRed exec [/usr/bin/python3, mapp.py]
2025-02-04 15:26:53,166 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output.dir
2025-02-04 15:26:53,166 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
2025-02-04 15:26:53,167 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
2025-02-04 15:26:53,167 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.length
2025-02-04 15:26:53,168 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
2025-02-04 15:26:53,168 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
2025-02-04 15:26:53,170 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
2025-02-04 15:26:53,170 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
2025-02-04 15:26:53,171 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
2025-02-04 15:26:53,171 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
2025-02-04 15:26:53,172 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
2025-02-04 15:26:53,173 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
2025-02-04 15:26:53,348 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
2025-02-04 15:26:53,348 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
```

```
2025-02-04 15:26:53,349 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
2025-02-04 15:26:53,352 INFO streaming.PipeMapRed: Records R/W=100/1
2025-02-04 15:26:53,355 INFO streaming.PipeMapRed: MRErrorThread done
2025-02-04 15:26:53,356 INFO streaming.PipeMapRed: mapRedFinished
2025-02-04 15:26:53,358 INFO mapred.LocalJobRunner:
2025-02-04 15:26:53,358 INFO mapred.MapTask: Starting flush of map output
2025-02-04 15:26:53,358 INFO mapred.MapTask: Spilling map output
2025-02-04 15:26:53,358 INFO mapred.MapTask: bufstart = 0; bufend = 796; bufvoid = 104857600
2025-02-04 15:26:53,358 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214000(104856000); length = 397/6553600
2025-02-04 15:26:53,369 INFO mapred.MapTask: Finished spill 0
2025-02-04 15:26:53,383 INFO mapred.Task: Task:attempt_local1040376759_0001_m_000000_0 is done. And is in the process of committing
2025-02-04 15:26:53,399 INFO mapred.LocalJobRunner: Records R/W=100/1
2025-02-04 15:26:53,400 INFO mapred.Task: Task 'attempt_local1040376759_0001_m_000000_0' done.
2025-02-04 15:26:53,413 INFO mapred.Task: Final Counters for attempt_local1040376759_0001_m_000000_0: Counters: 23
        File System Counters
                FILE: Number of bytes read=1427
                FILE: Number of bytes written=651308
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=796
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=5
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=1
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=100
                Map output records=100
                Map output bytes=796
                Map output materialized bytes=1002
                Input split bytes=102
                Combine input records=0
                Spilled Records=100
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=271581184
        File Input Format Counters
                Bytes Read=796
2025-02-04 15:26:53,414 INFO mapred.LocalJobRunner: Finishing task: attempt_local1040376759_0001_m_000000_0
2025-02-04 15:26:53,414 INFO mapred.LocalJobRunner: map task executor complete.
2025-02-04 15:26:53,420 INFO mapred.LocalJobRunner: Waiting for reduce tasks
2025-02-04 15:26:53,420 INFO mapred.LocalJobRunner: Starting task: attempt_local1040376759_0001_r_000000_0
```

```
2025-02-04 15:26:53,443 INFO output.FileOutputCommitter: File Output Committer Algorithm
version is 2
2025-02-04 15:26:53,446 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup
_temporary folders under output directory:false, ignore cleanup failures: false
2025-02-04 15:26:53,447 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2025-02-04 15:26:53,460 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin:
org.apache.hadoop.mapreduce.task.reduce.Shuffle@636731bd
2025-02-04 15:26:53,463 WARN impl.MetricsSystemImpl: JobTracker metrics system already
initialized!
2025-02-04 15:26:53,483 INFO reduce.MergeManagerImpl: MergerManager:
memoryLimit=1272394496, maxSingleShuffleLimit=318098624, mergeThreshold=839780416,
ioSortFactor=10, memToMemMergeOutputsThreshold=10
2025-02-04 15:26:53,489 INFO reduce.EventFetcher: attempt_local1040376759_0001_r_000000_0
Thread started: EventFetcher for fetching Map Completion Events
2025-02-04 15:26:53,536 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map
attempt_local1040376759_0001_m_000000_0 decomp: 998 len: 1002 to MEMORY
2025-02-04 15:26:53,543 INFO reduce.InMemoryMapOutput: Read 998 bytes from map-output for
attempt_local1040376759_0001_m_000000_0
2025-02-04 15:26:53,545 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of
size: 998, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->998
2025-02-04 15:26:53,549 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
2025-02-04 15:26:53,550 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-02-04 15:26:53,551 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory
map-outputs and 0 on-disk map-outputs
2025-02-04 15:26:53,565 INFO mapred.Merger: Merging 1 sorted segments
2025-02-04 15:26:53,565 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left
of total size: 991 bytes
2025-02-04 15:26:53,567 INFO reduce.MergeManagerImpl: Merged 1 segments, 998 bytes to disk
to satisfy reduce memory limit
2025-02-04 15:26:53,568 INFO reduce.MergeManagerImpl: Merging 1 files, 1002 bytes from disk
2025-02-04 15:26:53,569 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from
memory into reduce
2025-02-04 15:26:53,570 INFO mapred.Merger: Merging 1 sorted segments
2025-02-04 15:26:53,573 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left
of total size: 991 bytes
2025-02-04 15:26:53,574 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-02-04 15:26:53,580 INFO streaming.PipeMapRed: PipeMapRed exec [/usr/bin/python3,
redu.py]
2025-02-04 15:26:53,583 INFO Configuration.deprecation: mapred.job.tracker is deprecated.
Instead, use mapreduce.jobtracker.address
2025-02-04 15:26:53,586 INFO Configuration.deprecation: mapred.map.tasks is deprecated.
Instead, use mapreduce.job.maps
2025-02-04 15:26:53,644 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
2025-02-04 15:26:53,645 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA
[rec/s]
2025-02-04 15:26:53,646 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA
[rec/s]
2025-02-04 15:26:53,647 INFO streaming.PipeMapRed: Records R/W=100/1
2025-02-04 15:26:53,657 INFO streaming.PipeMapRed: MRErrorThread done
2025-02-04 15:26:53,657 INFO streaming.PipeMapRed: mapRedFinished
2025-02-04 15:26:53,765 INFO mapred.Task: Task:attempt_local1040376759_0001_r_000000_0 is
done. And is in the process of committing
```

```
2025-02-04 15:26:53,768 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-02-04 15:26:53,769 INFO mapred.Task: Task attempt_local1040376759_0001_r_000000_0 is
allowed to commit now
2025-02-04 15:26:53,809 INFO output.FileOutputCommitter: Saved output of task
'attempt_local1040376759_0001_r_000000_0' to hdfs://localhost:9000/160122771091/output.txt
2025-02-04 15:26:53,809 INFO mapred.LocalJobRunner: Records R/W=100/1 > reduce
2025-02-04 15:26:53,810 INFO mapred.Task: Task 'attempt_local1040376759_0001_r_000000_0'
done.
2025-02-04 15:26:53,810 INFO mapred.Task: Final Counters for
attempt_local1040376759_0001_r_000000_0: Counters: 30
        File System Counters
                FILE: Number of bytes read=3463
                FILE: Number of bytes written=652310
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=796
                HDFS: Number of bytes written=272
                HDFS: Number of read operations=10
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=3
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Combine input records=0
                Combine output records=0
                Reduce input groups=21
                Reduce shuffle bytes=1002
                Reduce input records=100
                Reduce output records=22
                Spilled Records=100
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=17
                Total committed heap usage (bytes)=310378496
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Output Format Counters
                Bytes Written=272
2025-02-04 15:26:53,811 INFO mapred.LocalJobRunner: Finishing task:
attempt_local1040376759_0001_r_000000_0
2025-02-04 15:26:53,811 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-02-04 15:26:53,936 INFO mapreduce.Job: Job job_local1040376759_0001 running in uber
mode : false
2025-02-04 15:26:53,937 INFO mapreduce.Job:  map 100% reduce 100%
2025-02-04 15:26:53,940 INFO mapreduce.Job: Job job_local1040376759_0001 completed
successfully
```

```
2025-02-04 15:26:53,944 INFO mapreduce.Job: Counters: 36
        File System Counters
                FILE: Number of bytes read=4890
                FILE: Number of bytes written=1303618
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1592
                HDFS: Number of bytes written=272
                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=100
                Map output records=100
                Map output bytes=796
                Map output materialized bytes=1002
                Input split bytes=102
                Combine input records=0
                Combine output records=0
                Reduce input groups=21
                Reduce shuffle bytes=1002
                Reduce input records=100
                Reduce output records=22
                Spilled Records=200
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=17
                Total committed heap usage (bytes)=581959680
        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=796
        File Output Format Counters
                Bytes Written=272
2025-02-04 15:26:53,948 INFO streaming.StreamJob: Output directory: /160122771091/output.txt
hadoop@ubuntu22:~$ hdfs dfs -ls  /160122771091
Found 8 items
-rw-r--r--   1 hadoop supergroup        170 2025-02-04 15:23 /160122771091/mapp.py
-rw-r--r--   1 hadoop supergroup        104 2025-01-28 14:45 /160122771091/mapper.py
drwxr-xr-x   - hadoop supergroup          0 2025-02-04 15:26 /160122771091/output.txt
drwxr-xr-x   - hadoop supergroup          0 2025-01-28 15:10 /160122771091/output3.txt
-rw-r--r--   1 hadoop supergroup        498 2025-02-04 15:24 /160122771091/redu.py
-rw-r--r--   1 hadoop supergroup        392 2025-01-28 14:46 /160122771091/reducer.py
-rw-r--r--   1 hadoop supergroup        796 2025-02-04 15:23 /160122771091/weathertext.csv
```

```
-rw-r--r--   1 hadoop supergroup         47 2025-01-28 14:38 /160122771091/wordcount.txt
hadoop@ubuntu22:~$ hdfs dfs -cat  /160122771091/output.txt/part-00000
Year,MaxTemp,MinTemp
2005,56,39
2006,79,34
2007,76,38
2008,65,21
2009,60,34
2010,53,37
2011,66,6
2012,69,8
2013,74,0
2014,51,26
2015,60,31
2016,56,14
2017,60,27
2018,63,21
2019,56,32
2020,73,36
2021,73,27
2022,71,-10
2023,72,39
2024,76,32
2025,29,14
hadoop@ubuntu22:~$
```