1. Finish mapper.py and test the code

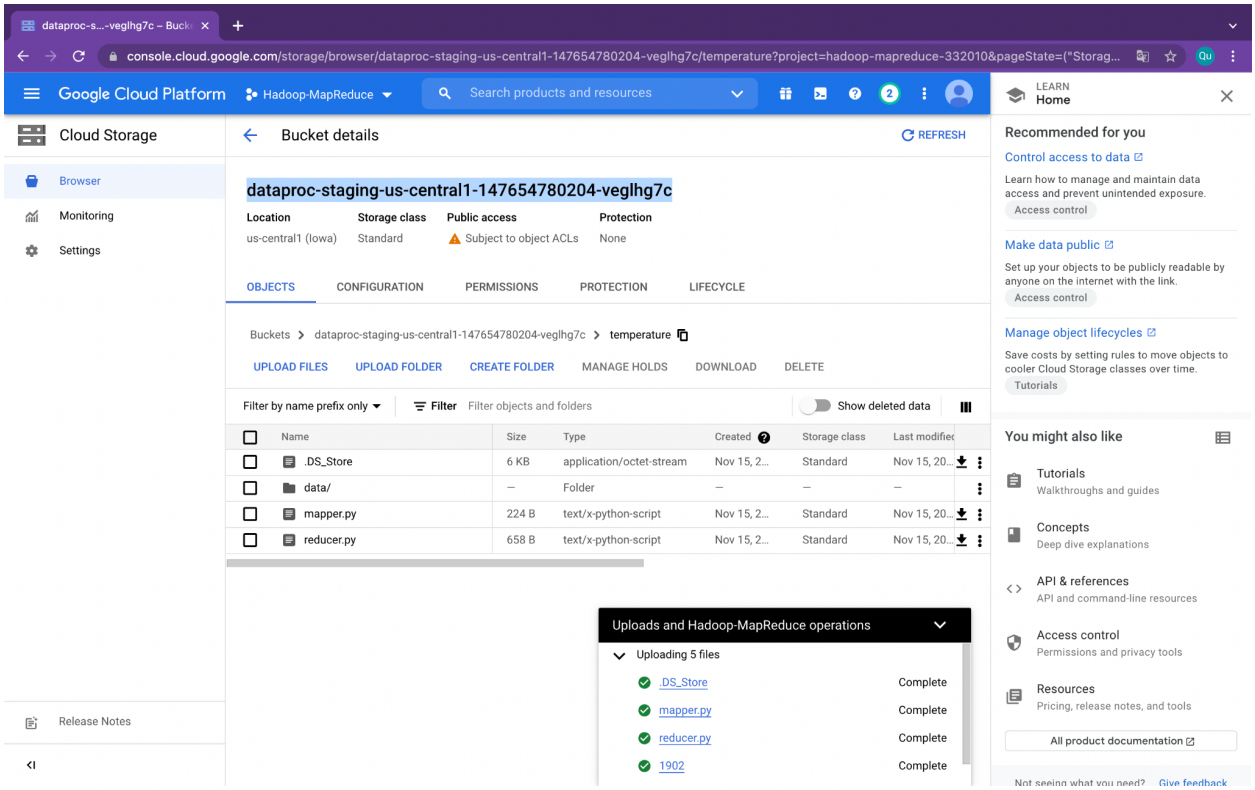
```
Last login: Mon Nov 15 11:05:17 on ttys000
(base) qmr@qumingruodeMBP ~ % /Users/qmr/Desktop/cmu_semester1/cloud
zsh: permission denied: /Users/qmr/Desktop/cmu_semester1/cloud
(base) qmr@qumingruodeMBP ~ % cd /Users/qmr/Desktop/cmu_semester1/cloud
(base) qmr@qumingruodeMBP cloud % cd temperature
(base) qmr@qumingruodeMBP temperature % cat data/1901 | python mapper.py
1901    -78
1901    -72
1901    -94
1901    -61
1901    -56
1901    -28
1901    -67
1901    -33
1901    -28
1901    -33
1901    -44
1901    -39
1901    0
1901    6
1901    0
1901    6
1901    6
1901    -11
1901    -33
1901    -50
1901    -44
1901    -28
1901    -33
1901    -33
1901    -50
1901    -33
1901    -28
1901    -44
1901    -44
1901    -44
1901    -39
1901    -50
1901    -44
1901    -39
1901    -33
1901    -22
1901    0
1901    -6
1901    -17
1901    -44
1901    -39
1901    -33
1901    -6
1901    17
1901    22
1901    28
1901    28
1901    11
1901    -17
1901    -28
1901    -56
1901    -44
1901    -44
1901    -67
1901    -44
1901    -39
1901    -22
```

2. Finish reducer.py and test the code, reduce on local machine

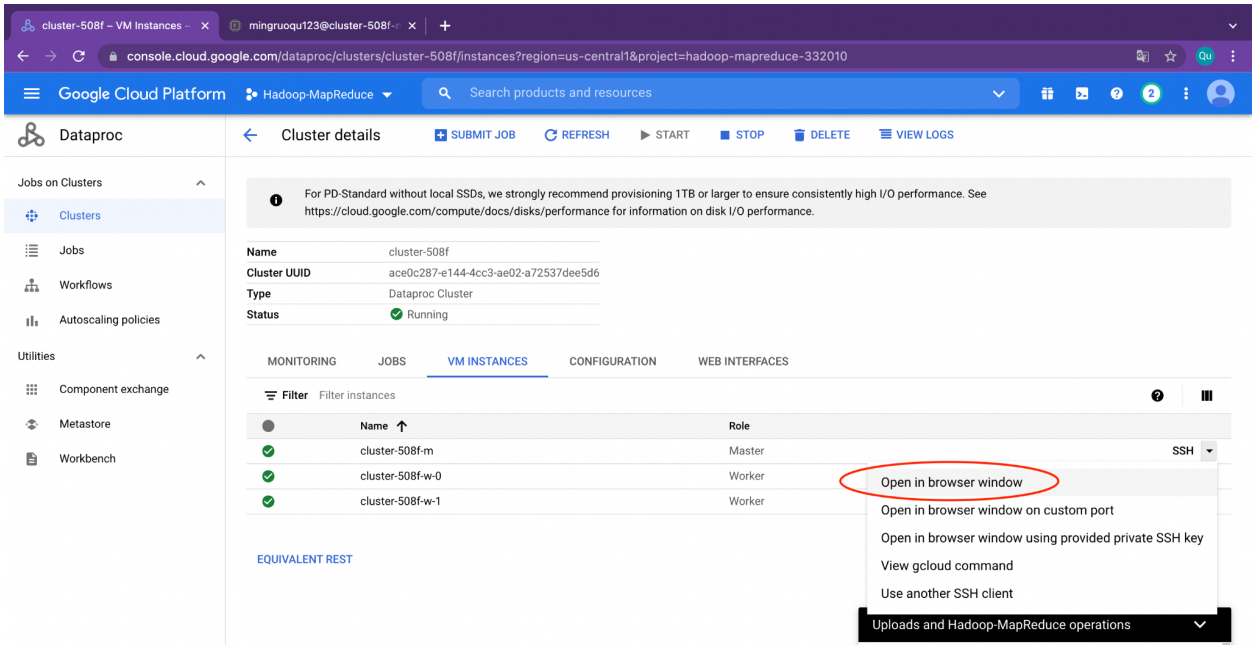
```
BrokenPipeError: [Errno 32] Broken pipe
(base) qmr@qumingruodeMacBook-Pro temperature % cat data/1902 | python mapper.py | sort -k1,1 | python reducer.py
19020101        28
19020102        11
19020103        -61
19020104        28
19020105        22
19020106        11
19020107        28
19020108        28
19020109        28
19020110        33
19020111        6
19020112        -11
19020113        -6
19020114        -22
19020115        -11
19020116        28
19020117        0
19020118        -17
19020119        -6
19020120        22
19020121        11
19020122        -6
19020123        6
19020124        28
19020125        22
19020126        17
19020127        11
19020128        -6
19020129        -6
19020130        -33
19020131        -17
19020201        22
19020202        17
19020203        -6
19020204        -17
19020205        -50
19020206        -11
19020207        -11
19020208        0
19020209        -44
19020210        -28
19020211        -61
19020212        0
19020213        -11
19020214        -33
19020215        -6
19020216        11
19020217        22
19020218        0
19020219        117
19020220        -22
19020221        -6
19020222        0
19020223        0
19020224        0
19020225        0
19020226        6
19020227        -6
19020228        6
19020301        17
19020302        22
```

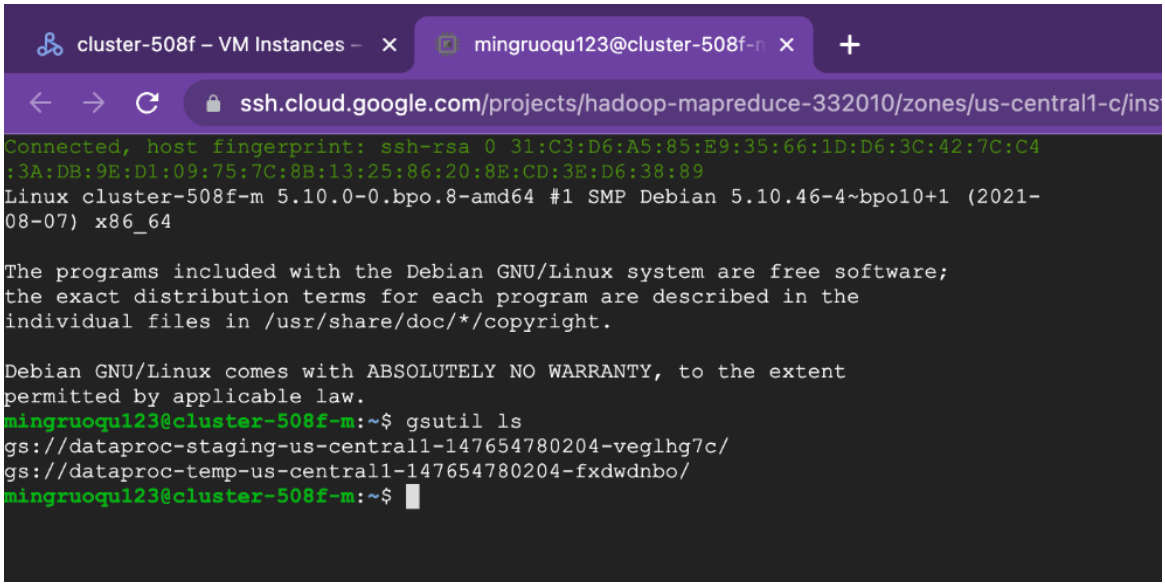
3. Create a cluster on Data Proc and upload the files we need

   bucket id: datapro-staging-us-central1-147654780204-veglhg7c

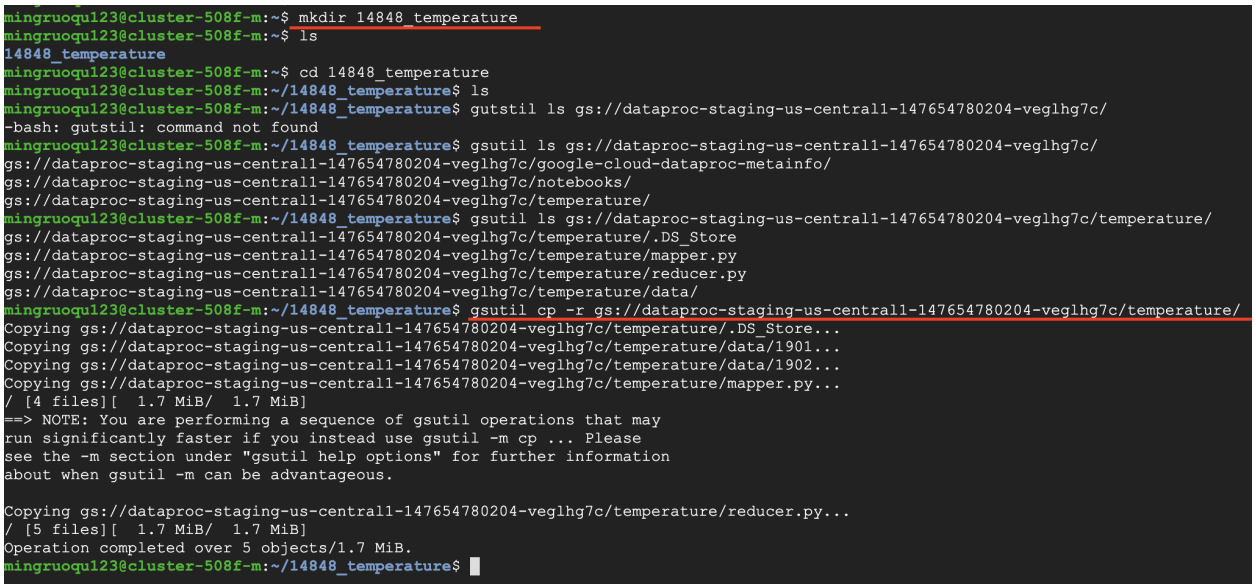

Google Cloud Platform
Hadoop-MapReduce
Search products and resources
Cloud Storage
Bucket details
REFRESH
Browser
Monitoring
Settings
dataproc-staging-us-central1-147654780204-veglhg7c
Location
us-central1 (Iowa)
Storage class
Standard
Public access
Subject to object ACLs
Protection
None
OBJECTS
CONFIGURATION
PERMISSIONS
PROTECTION
LIFECYCLE
Buckets > dataproc-staging-us-central1-147654780204-veglhg7c > temperature
UPLOAD FILES
UPLOAD FOLDER
CREATE FOLDER
MANAGE HOLDS
DOWNLOAD
DELETE
Filter by name prefix only
Filter
Filter objects and folders
Show deleted data
Name
Size
Type
Created
Storage class
.DS_Store
6 KB
application/octet-stream
Standard
data/
Folder
mapper.py
224 B
text/x-python-script
reducer.py
658 B
Uploads and Hadoop-MapReduce operations
Uploading 5 files
.DS_Store
mapper.py
reducer.py
1902
Complete
Release Notes
LEARN
Home
Recommended for you
Control access to data
Learn how to manage and maintain data access and prevent unintended exposure.
Access control
Make data public
Set up your objects to be publicly readable by anyone on the internet with the link.
Access control
Manage object lifecycles
Save costs by setting rules to move objects to cooler Cloud Storage classes over time.
Tutorials
You might also like
Tutorials
Walkthroughs and guides
Concepts
Deep dive explanations
API & references
API and command-line resources
Access control
Permissions and privacy tools
Resources
Pricing, release notes, and tools
All product documentation

4. Open SSH


Google Cloud Platform
Hadoop-MapReduce
Search products and resources
Dataproc
Cluster details
SUBMIT JOB
REFRESH
START
STOP
DELETE
VIEW LOGS
Jobs on Clusters
Clusters
Jobs
Workflows
Autoscaling policies
Utilities
Component exchange
Metastore
Workbench
For PD-Standard without local SSDs, we strongly recommend provisioning 1TB or larger to ensure consistently high I/O performance. See https://cloud.google.com/compute/docs/disks/performance for information on disk I/O performance.
Name
cluster-508f
Cluster UUID
ace0c287-e144-4cc3-ae02-a72537dee5d6
Type
Dataproc Cluster
Status
Running
MONITORING
JOBS
VM INSTANCES
CONFIGURATION
WEB INTERFACES
Filter
Filter instances
Name
Role
cluster-508f-m
Master
SSH
cluster-508f-w-0
Worker
cluster-508f-w-1
Worker
Open in browser window
Open in browser window on custom port
Open in browser window using provided private SSH key
View gcloud command
Use another SSH client
EQUIVALENT REST
Uploads and Hadoop-MapReduce operations




5. Moving the python code to Local Cluster (Namenode): create a new folder named '14848_temperature' and copy the uploaded file into the new folder through 'gsutil cp -r'




Verify it is the python file we uploaded

```
mingruoqu123@cluster-508f-m:~/14848_temperature$ gsutil cp -r gs://dataproc-staging-us-central1-147654780204-veglhg7c/temperature/ .
Copying gs://dataproc-staging-us-central1-147654780204-veglhg7c/temperature/.DS_Store...
Copying gs://dataproc-staging-us-central1-147654780204-veglhg7c/temperature/data/1901...
Copying gs://dataproc-staging-us-central1-147654780204-veglhg7c/temperature/data/1902...
Copying gs://dataproc-staging-us-central1-147654780204-veglhg7c/temperature/mapper.py...
- [4 files][  1.7 MiB/  1.7 MiB]
==> NOTE: You are performing a sequence of gsutil operations that may
run significantly faster if you instead use gsutil -m cp ... Please
see the -m section under "gsutil help options" for further information
about when gsutil -m can be advantageous.

Copying gs://dataproc-staging-us-central1-147654780204-veglhg7c/temperature/reducer.py...
- [5 files][  1.7 MiB/  1.7 MiB]
Operation completed over 5 objects/1.7 MiB.
mingruoqu123@cluster-508f-m:~/14848_temperature$ ls
temperature
mingruoqu123@cluster-508f-m:~/14848_temperature$ ls temperature/
data  mapper.py  reducer.py
mingruoqu123@cluster-508f-m:~/14848_temperature$ cd temperature/
mingruoqu123@cluster-508f-m:~/14848_temperature/temperature$ ls
data  mapper.py  reducer.py
mingruoqu123@cluster-508f-m:~/14848_temperature/temperature$ cat mapper.py
import sys
valid_quality = [0,1,4,5,9]
valid_temperature = [9999]
for line in sys.stdin:
    line = line.strip()
    date = line[15:23]
    temperature = int(line[87:92])
    quality = int(line[92])
    if (temperature not in valid_temperature and quality in valid_quality):
        print('%s\t%d' % (date, temperature))mingruoqu123@cluster-508f-m:~/14848_temperature/temperature$ cat reducer.py
```

6. create data folder

```
    print('%s\t%d' % (current_date, current_temperature))
mingruoqu123@cluster-508f-m:~/14848_temperature/temperature$ hadoop fs -ls /
Found 3 items
drwxrwxrwt   - hdfs hadoop          0 2021-11-15 22:33 /tmp
drwxrwxrwt   - hdfs hadoop          0 2021-11-15 22:33 /user
drwx-wx-wx   - hive hadoop          0 2021-11-15 22:33 /var
mingruoqu123@cluster-508f-m:~/14848_temperature/temperature$ hadoop fs -put data/ /
mingruoqu123@cluster-508f-m:~/14848_temperature/temperature$ hadoop fs -ls /
Found 4 items
drwxr-xr-x   - mingruoqu123 hadoop          0 2021-11-16 00:47 /data
drwxrwxrwt   - hdfs         hadoop          0 2021-11-15 22:33 /tmp
drwxrwxrwt   - hdfs         hadoop          0 2021-11-15 22:33 /user
drwx-wx-wx   - hive         hadoop          0 2021-11-15 22:33 /var
mingruoqu123@cluster-508f-m:~/14848_temperature/temperature$ hadoop fs -ls /data
Found 2 items
-rw-r--r--   2 mingruoqu123 hadoop     888190 2021-11-16 00:47 /data/1901
-rw-r--r--   2 mingruoqu123 hadoop     888978 2021-11-16 00:47 /data/1902
mingruoqu123@cluster-508f-m:~/14848_temperature/temperature$
```

7. the execution of Hadoop MapReduce Job in the terminal

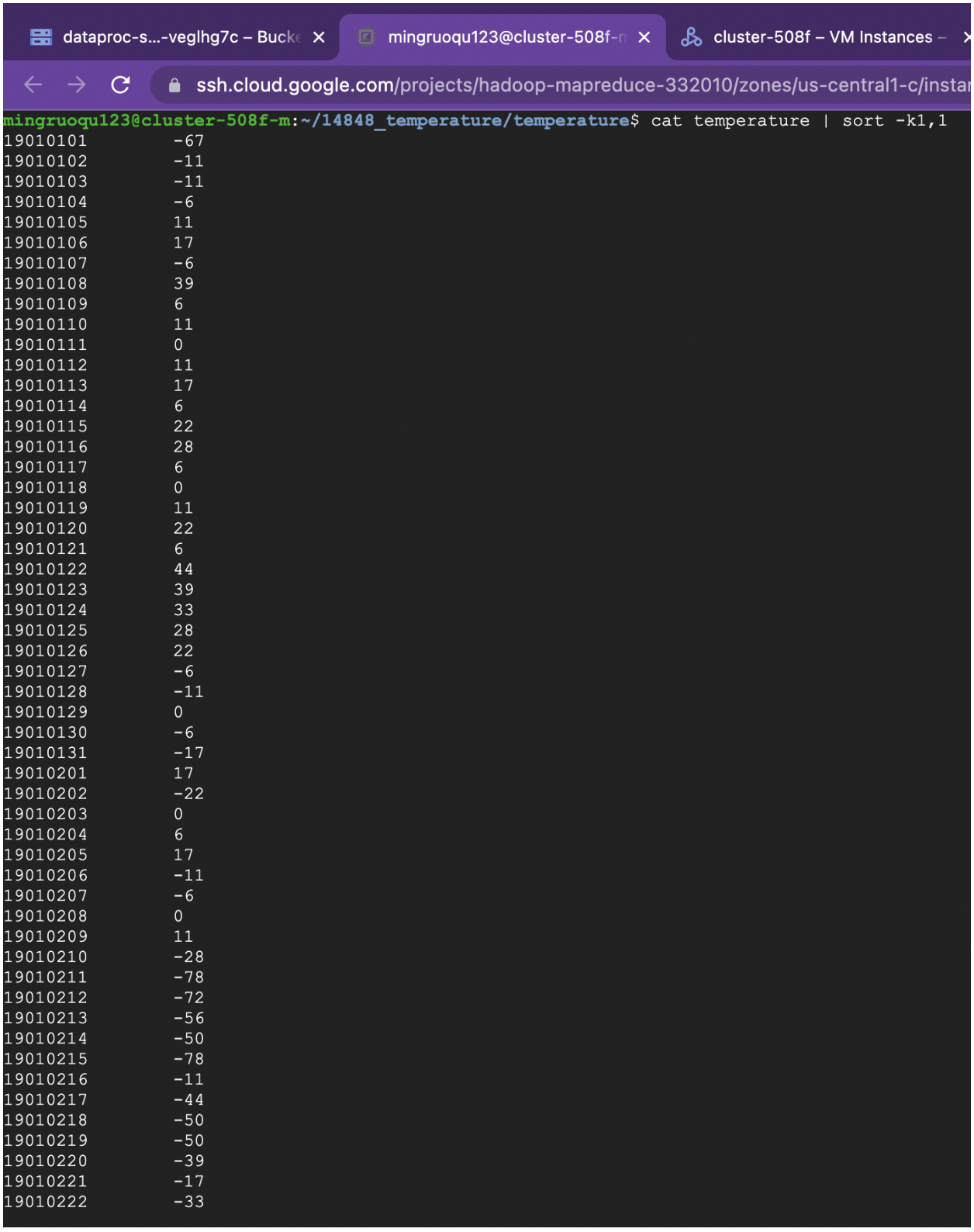
```
mingruoqu123@cluster-508f-m:~/14848_temperature/temperature$ hadoop jar /usr/lib/hadoop/hadoop-streaming.jar -file mapper.py -mapper 'python mapper.py' -file reducer.py -reducer 'pytho
n reducer.py' -input /data -output /outputFolder5
2021-11-16 02:08:35,950 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [mapper.py, reducer.py] [/usr/lib/hadoop/hadoop-streaming-3.2.2.jar] /tmp/streamjob8392252088802662895.jar tmpDir=null
2021-11-16 02:08:36,761 INFO client.RMProxy: Connecting to ResourceManager at cluster-508f-m/10.128.0.4:8032
2021-11-16 02:08:37,027 INFO client.AHSProxy: Connecting to Application History server at cluster-508f-m/10.128.0.4:10200
2021-11-16 02:08:37,519 INFO client.RMProxy: Connecting to ResourceManager at cluster-508f-m/10.128.0.4:8032
2021-11-16 02:08:37,520 INFO client.AHSProxy: Connecting to Application History server at cluster-508f-m/10.128.0.4:10200
2021-11-16 02:08:37,701 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/mingruoqu123/.staging/job_1637015590244_0003
2021-11-16 02:08:38,077 INFO mapred.FileInputFormat: Total input files to process : 2
2021-11-16 02:08:38,146 INFO mapreduce.JobSubmitter: number of splits:22
2021-11-16 02:08:38,288 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1637015590244_0003
2021-11-16 02:08:38,290 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-11-16 02:08:38,535 INFO conf.Configuration: resource-types.xml not found
2021-11-16 02:08:38,536 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-11-16 02:08:38,600 INFO impl.YarnClientImpl: Submitted application application_1637015590244_0003
2021-11-16 02:08:38,634 INFO mapreduce.Job: The url to track the job: http://cluster-508f-m:8088/proxy/application_1637015590244_0003/
2021-11-16 02:08:38,636 INFO mapreduce.Job: Running job: job_1637015590244_0003
2021-11-16 02:08:46,742 INFO mapreduce.Job: Job job_1637015590244_0003 running in uber mode : false
2021-11-16 02:08:46,743 INFO mapreduce.Job:  map 0% reduce 0%
2021-11-16 02:08:54,851 INFO mapreduce.Job:  map 14% reduce 0%
2021-11-16 02:08:55,857 INFO mapreduce.Job:  map 18% reduce 0%
2021-11-16 02:08:56,893 INFO mapreduce.Job:  map 32% reduce 0%
2021-11-16 02:08:58,903 INFO mapreduce.Job:  map 36% reduce 0%
2021-11-16 02:09:00,914 INFO mapreduce.Job:  map 45% reduce 0%
2021-11-16 02:09:03,931 INFO mapreduce.Job:  map 64% reduce 0%
2021-11-16 02:09:04,937 INFO mapreduce.Job:  map 68% reduce 0%
2021-11-16 02:09:06,948 INFO mapreduce.Job:  map 77% reduce 0%
2021-11-16 02:09:08,959 INFO mapreduce.Job:  map 82% reduce 0%
2021-11-16 02:09:09,964 INFO mapreduce.Job:  map 86% reduce 0%
2021-11-16 02:09:10,969 INFO mapreduce.Job:  map 95% reduce 0%
2021-11-16 02:09:11,974 INFO mapreduce.Job:  map 100% reduce 0%
2021-11-16 02:09:18,013 INFO mapreduce.Job:  map 100% reduce 29%
2021-11-16 02:09:20,026 INFO mapreduce.Job:  map 100% reduce 71%
2021-11-16 02:09:22,036 INFO mapreduce.Job:  map 100% reduce 100%
2021-11-16 02:09:24,053 INFO mapreduce.Job: Job job_1637015590244_0003 completed successfully
2021-11-16 02:09:24,141 INFO mapreduce.Job: Counters: 55
        File System Counters
                FILE: Number of bytes read=192588
                FILE: Number of bytes written=7590655
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1860914
                HDFS: Number of bytes written=9111
                HDFS: Number of read operations=101
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=21
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=22
                Launched reduce tasks=7
                Data-local map tasks=22
```

```
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=21
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=22
                Launched reduce tasks=7
                Data-local map tasks=22
                Total time spent by all maps in occupied slots (ms)=409891812
                Total time spent by all reduces in occupied slots (ms)=104779200
                Total time spent by all map tasks (ms)=129877
                Total time spent by all reduce tasks (ms)=33200
                Total vcore-milliseconds taken by all map tasks=129877
                Total vcore-milliseconds taken by all reduce tasks=33200
                Total megabyte-milliseconds taken by all map tasks=409891812
                Total megabyte-milliseconds taken by all reduce tasks=104779200
        Map-Reduce Framework
                Map input records=13130
                Map output records=13129
                Map output bytes=166288
                Map output materialized bytes=193470
                Input split bytes=1826
                Combine input records=0
                Combine output records=0
                Reduce input groups=730
                Reduce shuffle bytes=193470
                Reduce input records=13129
                Reduce output records=730
                Spilled Records=26258
                Shuffled Maps =154
                Failed Shuffles=0
                Merged Map outputs=154
                GC time elapsed (ms)=4327
                CPU time spent (ms)=23080
                Physical memory (bytes) snapshot=12989292544
                Virtual memory (bytes) snapshot=128415109120
                Total committed heap usage (bytes)=12656836608
                Peak Map Physical memory (bytes)=553037824
                Peak Map Virtual memory (bytes)=4429299712
                Peak Reduce Physical memory (bytes)=285155328
                Peak Reduce Virtual memory (bytes)=4431630336
        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=1859088
        File Output Format Counters
                Bytes Written=9111
2021-11-16 02:09:24,141 INFO streaming.StreamJob: Output directory: /outputFolder5
mingruoqu123@cluster-508f-m:~/14848_temperature/temperature$
```

8. open the output directory, merge all the results to one file and store it on the local cluster. You can see the results is stored as ‘temperature’

```
                Bytes Written=9111
2021-11-16 02:09:24,141 INFO streaming.StreamJob: Output directory: /outputFolder5
mingruoqu123@cluster-508f-m:~/14848_temperature/temperature$ hadoop fs -ls /
Found 7 items
drwxr-xr-x   - mingruoqu123 hadoop          0 2021-11-16 01:46 /NewTemperatureOutputFolder
drwxr-xr-x   - mingruoqu123 hadoop          0 2021-11-16 01:40 /TemperatureOutputFolder
drwxr-xr-x   - mingruoqu123 hadoop          0 2021-11-16 00:47 /data
drwxr-xr-x   - mingruoqu123 hadoop          0 2021-11-16 02:09 /outputFolder5
drwxrwxrwt   - hdfs         hadoop          0 2021-11-15 22:33 /tmp
drwxrwxrwt   - hdfs         hadoop          0 2021-11-15 22:33 /user
drwx-wx-wx   - hive         hadoop          0 2021-11-15 22:33 /var
mingruoqu123@cluster-508f-m:~/14848_temperature/temperature$ hadoop fs -ls /NewTemperatureOutputFolder
mingruoqu123@cluster-508f-m:~/14848_temperature/temperature$ hadoop fs -ls /outputFolder5
Found 8 items
-rw-r--r--   2 mingruoqu123 hadoop          0 2021-11-16 02:09 /outputFolder5/_SUCCESS
-rw-r--r--   2 mingruoqu123 hadoop       1297 2021-11-16 02:09 /outputFolder5/part-00000
-rw-r--r--   2 mingruoqu123 hadoop       1312 2021-11-16 02:09 /outputFolder5/part-00001
-rw-r--r--   2 mingruoqu123 hadoop       1299 2021-11-16 02:09 /outputFolder5/part-00002
-rw-r--r--   2 mingruoqu123 hadoop       1281 2021-11-16 02:09 /outputFolder5/part-00003
-rw-r--r--   2 mingruoqu123 hadoop       1324 2021-11-16 02:09 /outputFolder5/part-00004
-rw-r--r--   2 mingruoqu123 hadoop       1312 2021-11-16 02:09 /outputFolder5/part-00005
-rw-r--r--   2 mingruoqu123 hadoop       1286 2021-11-16 02:09 /outputFolder5/part-00006
mingruoqu123@cluster-508f-m:~/14848_temperature/temperature$ hadoop fs -getmerge /outputFolder5 temperature
mingruoqu123@cluster-508f-m:~/14848_temperature/temperature$ ls
data  mapper.py  reducer.py  temperature
mingruoqu123@cluster-508f-m:~/14848_temperature/temperature$
```

9. Results printed




10. copy the results into your bucket. Then download the results through GUI.

```
mingruoqu123@cluster-508f-m:~/14848_temperature/temperature$ gsutil ls
gs://dataproc-staging-us-central1-147654780204-veglhg7c/
gs://dataproc-temp-us-central1-147654780204-fxdwdnbo/
mingruoqu123@cluster-508f-m:~/14848_temperature/temperature$ gsutil cp temperature gs://dataproc-staging-us-central1-147654780204-veglhg7c/
Copying file://temperature [Content-Type=application/octet-stream]...
/ [1 files][  8.9 KiB/  8.9 KiB]
Operation completed over 1 objects/8.9 KiB.
mingruoqu123@cluster-508f-m:~/14848_temperature/temperature$ gsutil cp temperature gs://dataproc-staging-us-central1-147654780204-veglhg7c/temperature/
Copying file://temperature [Content-Type=application/octet-stream]...
/ [1 files][  8.9 KiB/  8.9 KiB]
Operation completed over 1 objects/8.9 KiB.
mingruoqu123@cluster-508f-m:~/14848_temperature/temperature$
```