CIA-1: NCDC dataset

Q: Implement map reduce program for the NCDC dataset using hadoop to find minimum and max temperature. upload the code and output screenshots as a zip file.

To start Hadoop Commands:

1. su - hadoop
2. sudo service ssh start
3. ssh localhost
4. hdfs dfs -mkdir cia1 #created directory in hadoop cluster

Step 1: Download the NCDC dataset

mkdir cia1 #creating a directory locally
cd cia1
hdfs dfs -put /home/shruthimohan/ncdc_dataset.csv /cia1/input
#storing it in the hadoop cluster

Step 2: Write the mapper.py and reducer.py scripts

Make the mapper.py and reducer.py executable
chmod 777 mapper.py reducer.py

Step 3: LOCALLY RAN IT:

```
hadoop@EDITH:~/cia1$ cat /home/shruthimohan/ncdc_dataset.csv | python3 mapper.py | sort -k1,1 | python3 reducer.py
Minimum Temperature: -10.11
Maximum Temperature: 40.0
```

Output!

Step 4: Run the program on hadoop cluster:

hadoop jar /home/shruthimohan/hadoop-streaming-3.3.6.jar -input /cia1/input/ncdc_dataset.csv -output /user/hadoop/cia1/output -mapper /home/hadoop/cia1/mapper.py -reducer /home/hadoop/cia1/reducer.py

```
hadoop@EDITH:~/cia1$ hadoop jar /home/shruthimohan/hadoop-streaming-3.3.6.jar -input /cia1/input/ncdc_dataset.csv -output /user/hadoo
p/cia1/output -mapper /home/hadoop/cia1/mapper.py -reducer /home/hadoop/cia1/reducer.py
packageJobJar: [/tmp/hadoop-unjar1137096878937458949 3/] [] /tmp/streamjob16136205416904201330.jar tmpDir=null
2024-09-09 21:16:58,258 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-09-09 21:16:58,343 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-09-09 21:16:58,523 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.stagi
ng/job_1725890821876_0010
2024-09-09 21:16:58,730 INFO mapred.FileInputFormat: Total input files to process : 1
2024-09-09 21:16:58,800 INFO mapreduce.JobSubmitter: number of splits:2
2024-09-09 21:16:58,936 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1725890821876_0010
2024-09-09 21:16:58,936 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-09-09 21:16:59,054 INFO conf.Configuration: resource-types.xml not found
2024-09-09 21:16:59,054 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-09-09 21:16:59,107 INFO impl.YarnClientImpl: Submitted application application_1725890821876_0010
2024-09-09 21:16:59,143 INFO mapreduce.Job: The url to track the job: http://EDITH.localdomain:8088/proxy/application_1725890821876_0
010/
2024-09-09 21:16:59,144 INFO mapreduce.Job: Running job: job_1725890821876_0010
2024-09-09 21:17:04,220 INFO mapreduce.Job: Job job_1725890821876_0010 running in uber mode : false
2024-09-09 21:17:04,222 INFO mapreduce.Job:  map 0% reduce 0%
2024-09-09 21:17:09,287 INFO mapreduce.Job:  map 100% reduce 0%
2024-09-09 21:17:13,318 INFO mapreduce.Job:  map 100% reduce 100%
2024-09-09 21:17:14,334 INFO mapreduce.Job: Job job_1725890821876_0010 completed successfully
2024-09-09 21:17:14,391 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=93897
                FILE: Number of bytes written=1023584
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=94024
                HDFS: Number of bytes written=56
                HDFS: Number of read operations=11
```

OUTPUT:

```
hadoop@EDITH:~/cia1$ hdfs dfs -cat cia1/output/part-00000
Minimum Temperature: -10.11
Maximum Temperature: 40.0
```

Part-00000: first reducer’s output