

Exp. No : 3

Map Reduce program to process Weather dataset

1. Download Weather dataset.

The screenshot shows a text editor window titled 'dataset.txt' with 27 lines of data. Each line represents a day in October 2015, starting from 20151001 to 20151027. The data includes a date, time, location, and various weather metrics such as temperature, humidity, and wind speed. The data is formatted as a CSV-like structure with commas separating the fields.

2. Create mapper.py program

```

Activities Terminal Sep 14 12:40
hadoop@kiran: ~
mapper.py

GNU nano 6.2
#!/usr/bin/env python
import sys

# input comes from STDIN (standard input)
# the mapper will get daily max temperature and group it by month.
# So output will be (month, daily_max_temperature)

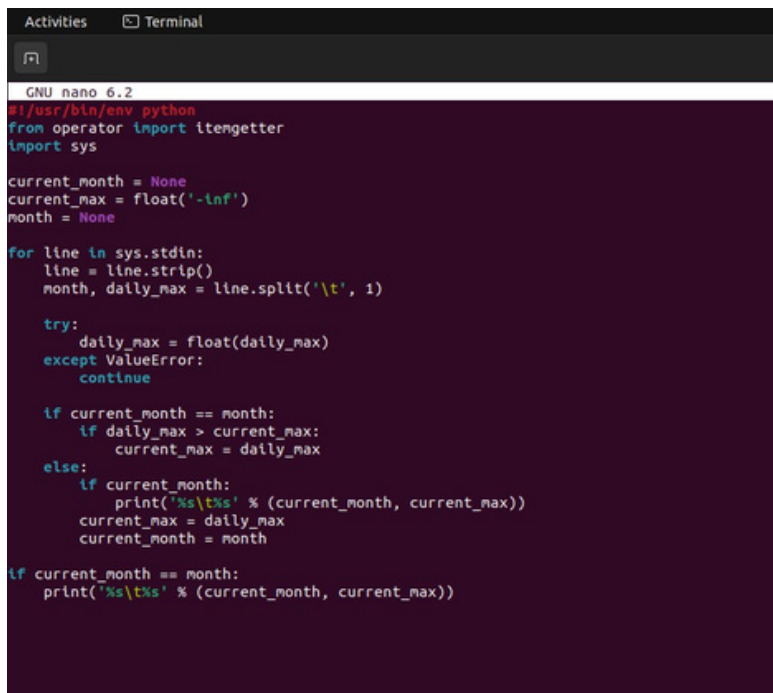
# Download the dataset (weather data)
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()

    # split the line into words
    words = line.split()

    # See the README hosted on the weather website which helps us understand how each
    # position represents a column
    month = line[10:12]
    daily_max = line[38:45]
    daily_max = daily_max.strip()

    # increase counters
    for word in words:
        # write the results to STDOUT (standard output);
        # what we output here will go through the shuffle process and then
        # be the input for the Reduce step, i.e. the input for reducer.py
        #
        # tab-delimited; month and daily max temperature as output
        print('%s\t%s' % (month, daily_max))
  
```

3. Create reducer.py



```

GNU nano 6.2
#!/usr/bin/env python
from operator import itemgetter
import sys

current_month = None
current_max = float('-inf')
month = None

for line in sys.stdin:
    line = line.strip()
    month, daily_max = line.split('\t', 1)

    try:
        daily_max = float(daily_max)
    except ValueError:
        continue

    if current_month == month:
        if daily_max > current_max:
            current_max = daily_max
    else:
        if current_month:
            print('%s\t%s' % (current_month, current_max))
            current_max = daily_max
            current_month = month

if current_month == month:
    print('%s\t%s' % (current_month, current_max))

```

5. Upload Weather dataset into HDFS Storage.

```

hadoop@koushik:~$ nano mapper.py
hadoop@koushik:~$ nano reducer.py
hadoop@koushik:~$ hdfs dfs -text /weatherdata/output/* > /home/Downloads/output/
/part-00000
bash: /home/Downloads/output/: No such file or directory
bash: /part-00000: No such file or directory
hadoop@koushik:~$ hdfs dfs -cat /weatherdata/output/part-00000 > /home/hadoop/Downloads/output.txt
hadoop@koushik:~$

```




6. Run the Map reduce program using Hadoop Streaming.

```

Activities  Terminal  Sep 14 12:42
hadoop@kiran: ~
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=90
2024-09-14 12:40:09,302 INFO mapred.LocalJobRunner: Finishing task: attempt_local458174712_0001_r_000000_0
2024-09-14 12:40:09,302 INFO mapred.LocalJobRunner: reduce task executor complete.
2024-09-14 12:40:09,913 INFO mapreduce.Job: map 100% reduce 100%
2024-09-14 12:40:09,914 INFO mapreduce.Job: Job local458174712_0001 completed successfully
2024-09-14 12:40:09,927 INFO mapreduce.Job: Counters: 36
File System Counters
  FILE: Number of bytes read=269834
  FILE: Number of bytes written=1603444
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=159136
  HDFS: Number of bytes written=90
  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=365
  Map output records=10220
  Map output bytes=81648
  Map output materialized bytes=102094
  Input split bytes=97
  Combine input records=0
  Combine output records=0
  Reduce input groups=12
  Reduce shuffle bytes=102094
  Reduce input records=10220
  Reduce output records=12
  Spilled Records=20460
  Shuffled Maps=1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=13
  Total committed heap usage (bytes)=54735672
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=79508
File Output Format Counters
  Bytes Written=90
2024-09-14 12:40:09,927 INFO streaming.StreamJob: Output directory: /weatherdata/output

```

Output :

Activities  Text Editor		
Open  		
1	01	26.5
2	02	26.6
3	03	29.1
4	04	30.8
5	05	31.1
6	06	33.6
7	07	38.5
8	08	40.2
9	09	36.5
10	10	36.9
11	11	27.6
12	12	25.9

