

Exp. No : 3

Map Reduce program to process Weather dataset

1. Download Weather dataset.

Activities Text Editor Oct 12 09:54

dataset.txt -EXPERIMENTS/Ex 3

Save

1	23907	20150101	2.423	-98.08	30.62	2.2	-0.6	0.8	0.9	7.0	1.47	C	3.7	1.1	2.5	99.9	85.4	97.2	0.369	0.308	-99.000	-99.000	-99.000	7.0	8.1	-9999.0	-9999.0
2	23907	20150102	2.423	-98.08	30.62	3.5	1.3	2.4	2.2	10.2	1.43	C	4.9	2.3	3.1	100.0	98.8	99.8	0.391	0.327	-99.000	-99.000	-99.000	7.1	7.9	-9999.0	-9999.0
3	23907	20150103	2.423	-98.08	30.62	15.9	2.3	9.1	7.5	3.1	11.00	C	16.4	2.9	7.3	100.0	34.8	73.7	0.450	0.397	-99.000	-99.000	-99.000	7.6	7.9	-9999.0	-9999.0
4	23907	20150104	2.423	-98.08	30.62	9.2	-1.3	3.9	4.2	0.0	13.24	C	12.4	-0.5	4.9	82.0	40.6	61.7	0.413	0.352	-99.000	-99.000	-99.000	7.3	7.9	-9999.0	-9999.0
5	23907	20150105	2.423	-98.08	30.62	10.9	-3.7	3.6	2.6	0.0	13.37	C	14.7	-3.0	3.8	77.9	33.3	57.4	0.399	0.340	-99.000	-99.000	-99.000	6.3	7.0	-9999.0	-9999.0
6	23907	20150106	2.423	-98.08	30.62	20.2	2.9	11.6	10.9	0.0	12.90	C	22.0	1.6	9.9	67.7	30.2	49.3	0.395	0.335	-99.000	-99.000	-99.000	8.0	8.0	-9999.0	-9999.0
7	23907	20150107	2.423	-98.08	30.62	10.9	-3.4	3.8	4.5	0.0	12.68	C	12.4	-2.1	5.5	82.7	36.5	55.7	0.387	0.328	-99.000	-99.000	-99.000	7.6	8.3	-9999.0	-9999.0
8	23907	20150108	2.423	-98.08	30.62	0.6	-7.9	-3.6	-3.3	0.0	4.98	C	3.9	-4.8	-0.5	57.7	37.6	48.1	0.372	0.316	-99.000	-99.000	-99.000	4.7	6.1	-9999.0	-9999.0
9	23907	20150109	2.423	-98.08	30.62	2.0	0.1	1.0	0.8	0.0	2.52	C	4.1	1.2	2.5	87.8	48.9	64.4	0.368	0.312	-99.000	-99.000	-99.000	5.4	6.2	-9999.0	-9999.0
10	23907	20150110	2.423	-98.08	30.62	0.5	-2.0	-0.8	-0.6	3.9	2.11	C	2.5	-0.1	1.4	99.9	47.7	85.8	0.373	0.314	-99.000	-99.000	-99.000	5.1	6.0	-9999.0	-9999.0
11	23907	20150111	2.423	-98.08	30.62	10.9	0.0	5.4	4.4	2.6	6.38	C	12.7	1.3	5.8	100.0	77.8	97.1	0.420	0.362	-99.000	-99.000	-99.000	6.5	6.7	-9999.0	-9999.0
12	23907	20150112	2.423	-98.08	30.62	6.5	1.4	4.0	4.3	0.0	1.55	C	6.9	2.7	5.1	100.0	89.4	97.8	0.412	0.350	-99.000	-99.000	-99.000	7.3	7.5	-9999.0	-9999.0
13	23907	20150113	2.423	-98.08	30.62	3.0	-0.7	1.1	1.2	0.0	3.26	C	5.6	0.7	2.9	99.7	80.7	90.7	0.401	0.337	-99.000	-99.000	-99.000	6.1	6.8	-9999.0	-9999.0
14	23907	20150114	2.423	-98.08	30.62	2.9	0.9	1.9	1.8	0.7	1.88	C	4.7	2.0	3.1	99.6	90.8	97.9	0.395	0.331	-99.000	-99.000	-99.000	6.1	6.7	-9999.0	-9999.0
15	23907	20150115	2.423	-98.08	30.62	13.2	1.2	7.2	6.4	0.0	13.37	C	16.4	1.4	6.7	98.9	46.7	73.4	0.395	0.333	-99.000	-99.000	-99.000	6.7	7.0	-9999.0	-9999.0
16	23907	20150116	2.423	-98.08	30.62	16.7	3.5	10.1	9.9	0.0	13.68	C	19.2	1.3	8.7	80.2	38.1	50.2	0.391	0.330	-99.000	-99.000	-99.000	7.3	7.4	-9999.0	-9999.0
17	23907	20150117	2.423	-98.08	30.62	19.5	5.0	12.2	12.3	0.0	10.96	C	20.9	3.3	10.6	87.7	30.4	55.7	0.388	0.327	-99.000	-99.000	-99.000	8.7	8.4	-9999.0	-9999.0
18	23907	20150118	2.423	-98.08	30.62	20.9	7.6	14.3	13.7	0.0	15.03	C	23.4	3.5	11.9	45.9	14.6	31.4	0.383	0.325	-99.000	-99.000	-99.000	9.5	9.2	-9999.0	-9999.0
19	23907	20150119	2.423	-98.08	30.62	23.9	6.7	15.3	14.3	0.0	14.10	C	25.6	3.8	12.6	65.3	26.8	45.6	0.376	0.321	-99.000	-99.000	-99.000	9.9	9.5	-9999.0	-9999.0
20	23907	20150120	2.423	-98.08	30.62	26.0	9.5	17.8	15.9	0.0	14.57	C	27.9	6.5	14.5	88.4	16.1	50.2	0.373	0.320	-99.000	-99.000	-99.000	10.9	10.4	-9999.0	-9999.0
21	23907	20150121	2.423	-98.08	30.62	11.0	6.9	8.9	8.9	1.7	2.71	C	13.1	6.8	9.7	99.2	68.0	88.1	0.369	0.317	-99.000	-99.000	-99.000	10.7	10.6	-9999.0	-9999.0
22	23907	20150122	2.423	-98.08	30.62	8.6	3.5	6.1	5.6	40.0	1.28	C	9.1	4.1	6.3	99.6	95.2	98.0	0.546	0.418	-99.000	-99.000	-99.000	9.0	9.3	-9999.0	-9999.0
23	23907	20150123	2.423	-98.08	30.62	9.4	2.2	5.8	4.2	7.5	6.58	C	11.1	2.0	4.8	98.4	58.8	86.5	0.554	0.409	-99.000	-99.000	-99.000	7.6	8.1	-9999.0	-9999.0
24	23907	20150124	2.423	-98.08	30.62	16.0	1.4	8.7	8.0	0.0	14.26	C	18.8	0.4	7.7	92.0	33.0	63.0	0.494	0.381	-99.000	-99.000	-99.000	7.7	7.9	-9999.0	-9999.0
25	23907	20150125	2.423	-98.08	30.62	20.2	6.4	13.3	12.7	0.0	14.99	C	22.0	4.4	11.0	69.2	18.9	43.8	0.456	0.357	-99.000	-99.000	-99.000	9.1	8.9	-9999.0	-9999.0
26	23907	20150126	2.423	-98.08	30.62	21.5	7.2	14.4	14.1	0.0	12.01	C	22.9	5.5	12.2	56.8	23.7	40.6	0.433	0.349	-99.000	-99.000	-99.000	10.0	9.7	-9999.0	-9999.0
27	23907	20150127	2.423	-98.08	30.62	26.5	10.7	18.6	17.5	0.0	15.18	C	28.9	8.1	15.5	52.2	21.4	36.8	0.420	0.344	-99.000	-99.000	-99.000	11.4	10.8	-9999.0	-9999.0

Loading file "/home/hadoop/EXPERIMENTS/Ex 3/dataset.txt"...

Plain Text Tab Width: 8 Ln 1, Col 1 INS

2. Create mapper.py program

```
GNU nano 6.2 mapper.py
#!/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))

```

4. Run the Map reduce program using Hadoop Streaming.



```

-input /weatherdata/dataset.txt \
-output /weatherdata/output \
-file ~/mapper.py \
-file ~/reducer.py \
-hadoop "python3 mapper.py" \
-reducer "python3 reducer.py"
2024-09-14 12:39:56,313 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/hadoop/mapper.py, /home/hadoop/reducer.py] [] /tmp/streamjob826185063505703659.jar tmpDir=null
2024-09-14 12:39:57,567 INFO InPl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2024-09-14 12:39:57,764 INFO InPl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2024-09-14 12:39:57,780 WARN InPl.MetricsSystemImpl: JobTracker metrics system already initialized!
2024-09-14 12:39:58,108 INFO Mapred.FileInputFormat: Total input files to process : 1
2024-09-14 12:39:58,192 INFO Mapreduce.JobSubmitter: number of splits:1
2024-09-14 12:39:58,395 INFO Mapreduce.JobSubmitter: Submitting tokens for job: job_local458174712_0001
2024-09-14 12:39:58,395 INFO Mapreduce.JobSubmitter: Executing with tokens: []
2024-09-14 12:39:58,670 INFO Mapred.LocalDistributedCacheManager: Localized file:/home/hadoop/mapper.py as file:/tmp/hadoop-hadoop/mapred/local/job_local458174712_0001_2bba3c2d-2b34-4464-be51-ffbba723f3826/mapper.py
2024-09-14 12:39:58,722 INFO Mapred.LocalDistributedCacheManager: Localized file:/home/hadoop/reducer.py as file:/tmp/hadoop-hadoop/mapred/local/job_local458174712_0001_722ffb92-ba2a-447b-b6fa-6dfiab4dc169/reducer.py
2024-09-14 12:39:58,843 INFO Mapreduce.Job: The url to track the job: http://localhost:8080/
2024-09-14 12:39:58,845 INFO Mapred.LocalJobRunner: OutputCommitter set in config null
2024-09-14 12:39:58,845 INFO Mapreduce.Job: Running Job: job_local458174712_0001
2024-09-14 12:39:58,848 INFO Mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
2024-09-14 12:39:58,852 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-09-14 12:39:58,853 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup_temporary folders under output directory:false, ignore cleanup failures: false
2024-09-14 12:39:58,912 INFO Mapred.LocalJobRunner: Waiting for map tasks
2024-09-14 12:39:58,916 INFO Mapred.LocalJobRunner: Starting task: attempt_local458174712_0001_m_000000_0
2024-09-14 12:39:58,955 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-09-14 12:39:58,955 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup_temporary folders under output directory:false, ignore cleanup failures: false
2024-09-14 12:39:58,973 INFO Mapred.Task: Using ResourceCalculatorProcessTree : [ ]
2024-09-14 12:39:58,984 INFO Mapred.MapTask: Processing split: hdfs://localhost:9000/weatherdata/dataset.txt:0+79568
2024-09-14 12:39:59,011 INFO Mapred.MapTask: numReduceTasks: 1
2024-09-14 12:39:59,045 INFO Mapred.MapTask: (EquiJoin) 0 kvi 26214396(104857584)
2024-09-14 12:39:59,045 INFO Mapred.MapTask: mapreduce.task.io.sort.mb: 100
2024-09-14 12:39:59,045 INFO Mapred.MapTask: soft limit at 83886080
2024-09-14 12:39:59,045 INFO Mapred.MapTask: bufstart = 0; bufvoid = 104857600
2024-09-14 12:39:59,045 INFO Mapred.MapTask: kvstart = 26214396; length = 6553600
2024-09-14 12:39:59,049 INFO Mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2024-09-14 12:39:59,063 INFO streaming.PipelineMapred: PipelineMapred exec (/usr/bin/python3, mapper.py)
2024-09-14 12:39:59,067 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output.dir
2024-09-14 12:39:59,067 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
2024-09-14 12:39:59,068 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
2024-09-14 12:39:59,068 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.length
2024-09-14 12:39:59,068 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
2024-09-14 12:39:59,068 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
2024-09-14 12:39:59,069 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
2024-09-14 12:39:59,070 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
2024-09-14 12:39:59,070 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
2024-09-14 12:39:59,070 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
2024-09-14 12:39:59,071 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
2024-09-14 12:39:59,071 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
2024-09-14 12:39:59,207 INFO streaming.PipelineMapred: R/W/S-I/O/B ln:NA [rec/s] out:NA [rec/s]
2024-09-14 12:39:59,207 INFO streaming.PipelineMapred: R/W/S-I/O/B ln:NA [rec/s] out:NA [rec/s]
2024-09-14 12:39:59,208 INFO streaming.PipelineMapred: R/W/S-I/O/B ln:NA [rec/s] out:NA [rec/s]

```

```
Shuffle errors
BAD_ID=0
CONNECTION=0
ID_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Output Format Counters
  Bytes Written=96
2024-09-14 12:40:00,302 INFO mapred.LocalJobRunner: Finishing task: attempt_local458174712_0001_r_000000_0
2024-09-14 12:40:00,302 INFO mapred.LocalJobRunner: reduce task executor complete.
2024-09-14 12:40:00,913 INFO mapreduce.Job: map 100% reduce 100%
2024-09-14 12:40:00,914 INFO mapreduce.Job: Job JobLocal458174712_0001 completed successfully
2024-09-14 12:40:00,927 INFO mapreduce.Job: Counters: 36
  File System Counters
    FILE: Number of bytes read=209034
    FILE: Number of bytes written=1093444
    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=96
    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=20440
    Shuffled Maps=1
    Failed Shuffles=0
    Merged Map outputs=1
    GC time elapsed (ms)=13
    Total committed heap usage (bytes)=547356672
  Shuffle Errors
    BAD_ID=0
    CONNECTION=0
    ID_ERROR=0
    WRONG_LENGTH=0
    WRONG_MAP=0
    WRONG_REDUCE=0
  File Input Format Counters
    Bytes Read=72568
  File Output Format Counters
    Bytes Written=96
2024-09-14 12:40:00,927 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