Exercise 03e: Weather data analytics using Python MapReduce

This exercise's MapReduce process is doing Weather analysis using Python

Prerequisites

Ensure that Hadoop is installed, configured and is running. More details:

Single Node Setup for first-time users.

Cluster Setup for large, distributed clusters.

Inputs and Outputs

i. Input file should be in : /weatherp/in00/

WAData.txt

Copy the content text from sample_weather.txt, Which is attached in Google classroom.

ii. Output file should be in /weatherp/wc_output/

Step 1

Create Mapping Program and Reduce Program using Python WeatherAnalysis.jar:

(i) Create MaxTempMap.py file in /home/cloudera/ folder using following code.

```
import re

import sys

for line in sys.stdin:

  val = line.strip()

  (year, temp, q) = (val[15:19], val[87:92], val[92:93])

  if (temp != "+9999" and re.match("[01459]", q)):

    print "%s\t%s" % (year, temp)
```

Output:-

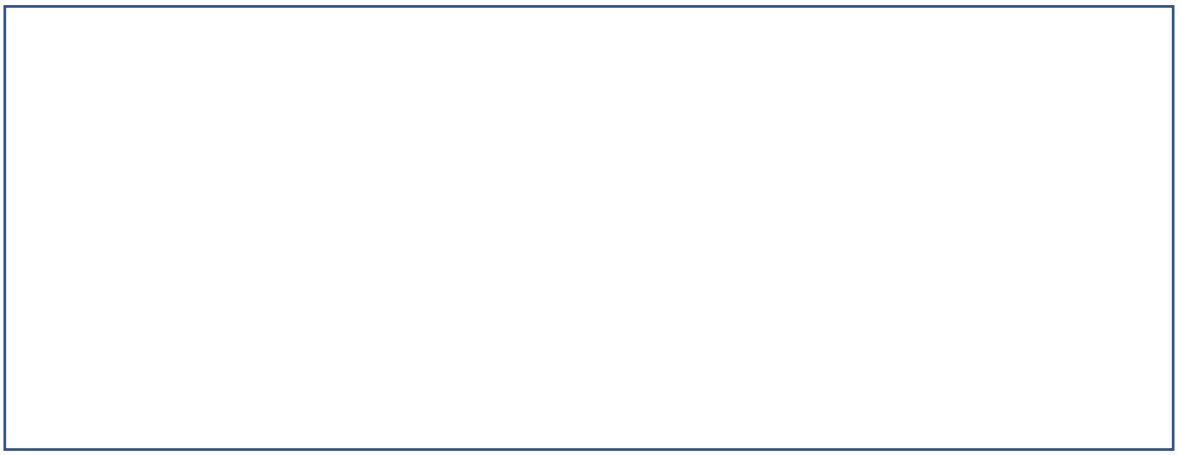
(ii) Create MaxTempReduce.py file in /home/cloudera/ floder using following code.

```
#!/usr/bin/env python


import sys


(last_key, max_val) = (None, -sys.maxint)
for line in sys.stdin:
  (key, val) = line.strip().split("\t")
  if last_key and last_key != key:
    print "%s\t%s" % (last_key, max_val)
    (last_key, max_val) = (key, int(val))
  else:
    (last_key, max_val) = (key, max(max_val, int(val)))


if last_key:
  print "%s\t%s" % (last_key, max_val)
```

Step 2

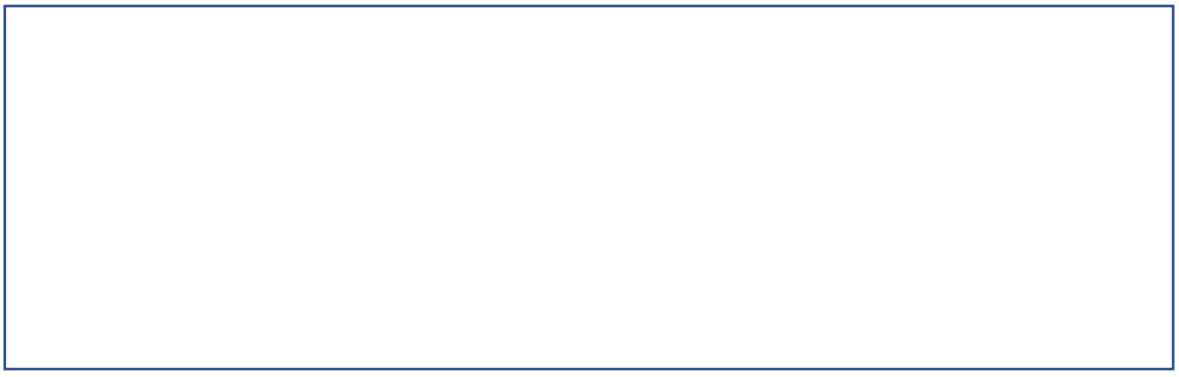
Check whether `hadoop-streaming-2.0.0-mr1-cdh4.jar` or `hadoop-streaming-2.6.0-mr1-cdh5.13.0.jar` file is available in the following path:-

"`/user/lib/hadoop-0.20-mapreduce/contrib/streaming/`**"**



Step 3

Create and copy WAData.txt-files into input folder:



[cloudera@quickstart ~]$ hdfs dfs -ls /weatherp/in00/

Found 1 items

-rw-r--r-- 1 cloudera supergroup 12054 2021-08-26 15:48 /weatherp/in00/WAData.txt

Step 4

Run the MapReduce application for python:

[cloudera@quickstart ~]$ `hadoop jar /usr/lib/hadoop-0.20-mapreduce/contrib/streaming/hadoop-streaming-2.6.0-mr1-cdh5.13.0.jar -file /home/cloudera/MaxTempMap.py /home/cloudera/MaxTempReduce.py -mapper "python map.py" -reducer "python reduce.py" -input /weatherp/in00/WAData.txt -output /weatherp/wc_output`

Show MapReduce Framework

Step 5

Output:

[cloudera@quickstart ~]$ hdfs dfs -ls /weatherp/ wc_output /

Found 2 items

-rw-r--r-- 1 cloudera supergroup 0 2021-08-26 15:50 /weatherp/wc_output/_SUCCESS

-rw-r--r-- 1 cloudera supergroup 228 2021-08-26 15:50 /weatherp/wc_output/part-r-00000

[cloudera@quickstart ~]$ hdfs dfs -cat /weatherp/wc_output/part-r-00000