Roll No. - 18

Srn No. - 201700110

Name – Kunal Agarwal

Experiment No - 18

Aim - Implement MR program that processes a weather dataset.

Theory -

Weather dataset examples read text files and give minimum and maximum temperature. The input is text files and the output is text files, consisting of minimum and maximum temperatures and city names.

Each mapper takes a line as input and breaks it into words.It then emits a city name, date and minimum and maximum temperature along with time of that particular date and city.

Each reducer takes a line as input and breaks it into words.It then emits a city name, date, global minimum and global maximum temperature.

As an optimization, the reducer is also used as a combiner on the map outputs. This reduces the amount of data sent across the network by combining each word into a single record.

Program:

Input File:

(File is Huge. Only added 2 rows)

CA\_25-Jan-2014 00:12:345 15.7 01:19:345 23.1 02:34:542 12.3 03:12:187 16 04:00:093 -14 05:12:345 35.7 06:19:345 23.1 07:34:542 12.3 08:12:187 16 09:00:093 -7 10:12:345 15.7 11:19:345 23.1 12:34:542 -22.3 13:12:187 16 14:00:093 -7 15:12:345 15.7 16:19:345 23.1 19:34:542 12.3 20:12:187 16 22:00:093 -7

CA\_26-Jan-2014 00:54:245 15.7 01:19:345 23.1 02:34:542 12.3 03:12:187 16 04:00:093 -14 05:12:345 55.7 06:19:345 23.1 07:34:542 12.3 08:12:187 16 09:00:093 -7 10:12:345 15.7 11:19:345 23.1 12:34:542 12.3 13:12:187 16 14:00:093 -7 15:12:345 15.7 16:19:345 23.1 19:34:542 12.3 20:12:187 16 22:00:093 -7

Mapper Program:

#!/usr/bin/env python3

import sys

temperatures = {}

for lines in sys.stdin:

   line = lines.strip().split()

   city, date = line[0].split('\_')

   for i in range(1,len(line),2):

       temperatures[line[i]] = float( line[i+1] )

   minTemperature= min(temperatures , key=temperatures.get)

   maxTemperature= max(temperatures , key=temperatures.get)

   min\_max\_temperature\_time = f"{city} {date} Time: {minTemperature} MinTemp: {temperatures[minTemperature]} Time: {maxTemperature} MaxTemp: {temperatures[maxTemperature]}"

   print(min\_max\_temperature\_time)

Reducer Program :

#!/usr/bin/env python3

import sys

from pprint import pprint

weather\_info = []

for lines in sys.stdin:

   line = lines.strip().split()

   weather\_info.append(line)

minIndex, maxIndex,minimumTemp, maximumTemp = -1, -1, 10000, -10000

for i in range(len(weather\_info)):

   if minimumTemp > float(weather\_info[i][5]):

       minIndex = i

       minimumTemp = float(weather\_info[i][5] )

   if maximumTemp < float(weather\_info[i][9]):

       maxIndex = i

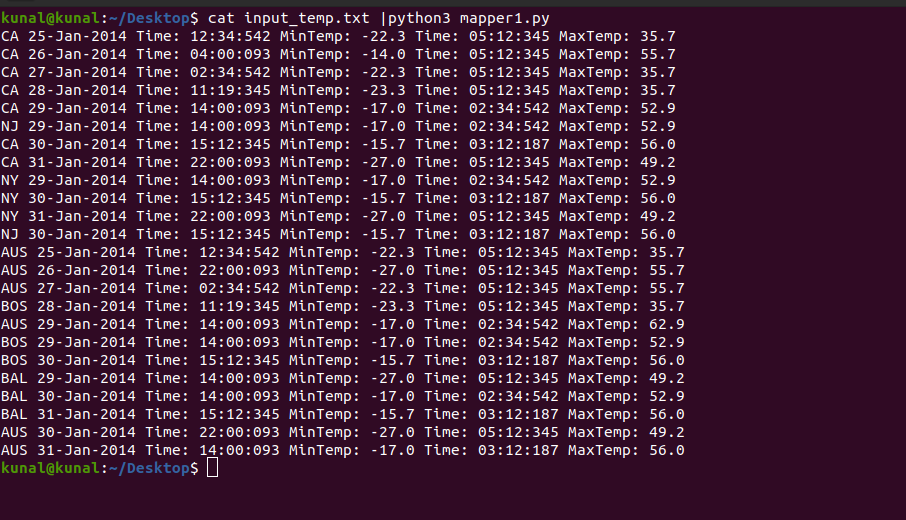
       maximumTemp = float( weather\_info[i][9] )

print(f"{weather\_info[minIndex][0]} {weather\_info[minIndex][1]} Time: {weather\_info[minIndex][3]} MinTemp : {weather\_info[minIndex][5]}")

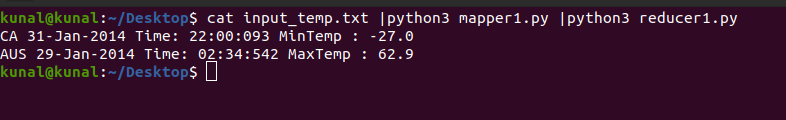
print(f"{weather\_info[maxIndex][0]} {weather\_info[maxIndex][1]} Time: {weather\_info[maxIndex][7]} MaxTemp : {weather\_info[maxIndex][9]}")

Output:

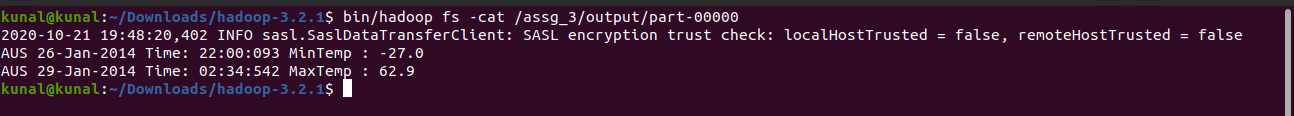
Mapper Output:



Reducer Output:



Hdfs Output:



Conclusion : We have got the minimum and maximum temperature from the weather dataset using the MapReduce algorithm in python with help of HDFS.