EX.NO.3 210701307

IMPLEMENT A MAPREDUCE PROGRAM TO PROCESS A WEATHER DATASET

AIM:

To implement a MapReduce python program to process a weather dataset in Hadoop.

PROCEDURE:

1. Open command prompt as administrator and start the Hadoop by using the

command: start-all.cmd

2. Create a new directory in the Hadoop file systems using the

command: hadoop fs -mkdir /weather

3. Upload the input text file into the weather directory using the command:

hadoop fs -put

C:/Users/mercy/OneDrive/Documents/DataAnalytics/WeatherPrediction/sample_weather.txt/weather

- 4. Create the mapper and reducer files.
- 5. To execute the files with Hadoop streaming run the following command:

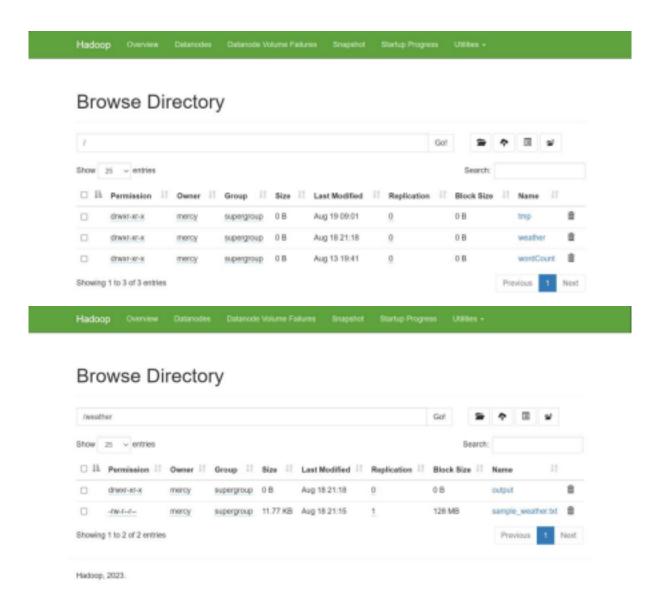
hadoop jar C:/hadoop-3.3.6/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar ^ -file C:/Users/mercy/Documents/DataAnalytics/WeatherPrediction/mapper.py ^ -file C:/Users/mercy/Documents/DataAnalytics/WeatherPrediction/reducer.py ^ -input /weather/sample_weather.txt ^ -output /weather/output ^ -mapper "python mapper.py" ^ -reducer "python reducer.py"

MAPPER.PY:

```
#!C:/ProgramData/chocolatey/bin/python3.exe
import sys
def map1():
    for line in sys.stdin:
        tokens = line.strip().split()
        if len(tokens) < 13:
            continue</pre>
```

```
station = tokens[0]
    if "STN" in station:
       continue
    date_hour = tokens[2]
    temp = tokens[3]
    dew = tokens[4]
    wind = tokens[12]
    if temp == "9999.9" or dew == "9999.9" or wind == "999.9":
       continue
    hour = int(date hour.split(" ")[-1])
    date = date hour[:date hour.rfind(" ")-2]
    if 4 < hour <= 10:
       section = "section1"
    elif 10 < hour <= 16:
       section = "section2"
    elif 16 < hour <= 22:
       section = "section3"
    else:
       section = "section4"
    key out = f"{station} {date} {section}"
    value out = f''\{temp\} \{dew\} \{wind\}''
    print(f"{key out}\t{value out}")
if name == " main ":
  map1()
REDUCER.PY:
#!C:/ProgramData/chocolatey/bin/python3.exe
import sys
def reduce1():
  current key = None
```

```
sum temp, sum dew, sum wind = 0, 0, 0
  count = 0
  for line in sys.stdin:
    key, value = line.strip().split("\t")
    temp, dew, wind = map(float, value.split())
    if current key is None:
       current key = key
    if key == current key:
       sum temp += temp
       sum dew += dew
       sum wind += wind
       count += 1
    else:
       avg temp = sum temp / count
       avg dew = sum dew / count
       avg wind = sum wind / count
       print(f"{current key}\t{avg temp} {avg dew} {avg wind}")
       current key = key
       sum_temp, sum_dew, sum_wind = temp, dew, wind
       count = 1
  if current key is not None:
    avg temp = sum temp / count
    avg dew = sum dew / count
    avg wind = sum wind / count
    print(f"{current key}\t{avg temp} {avg dew} {avg wind}")
if __name__ == "__main__":
  reduce1()
  OUTPUT:
```



RESULT:

Thus the implementation of the MapReduce python program to process a weather dataset in Hadoop is executed successfully.