# ImplementaMapReduceprogramtoprocessa weatherdataset

**Steps:**

1. Opencommandpromptandrunasadministrator Go to hadoopsbin directory



Note:

* 1. Checkhadoop/data/datanodeandhadoop/data/namenodeandifbothfolders are empty, type “hdfsnamenode -format”.
  2. Checkpythonversionwith“python --version”.
  3. Check“C:\Python39\”isaddedinEnvironment variables>Systemvariables

>Path, if not add your python path.

* 1. CheckEnvironmentvariables>Systemvariables>HADOOP\_HOMEisset as “C:\Hadoop”.

1. StartHadoopServicesstart-dfs.cmdstart-yarn.cmd



1. Openthebrowserandgo totheURL“localhost:9870”



1. CreateaDirectoryinHDFS hadoopfs-mkdir /user/weather



1. CopytheInputFiletoHDFShdfsdfs-put

C:\Users\monid\OneDrive\Documents\DataAnalytics\sample\_weather.txt/user/weather



# Note:mapper.py:

#!/usr/bin/envpythonimport sys

def map1(): for line in sys.stdin:

tokens=line.strip().split() if len(tokens) < 13:

continue

station=tokens[0] if "STN" in station:

continue

date\_hour = tokens[2] temp = tokens[3] dew= tokens[4] wind = tokens[12]

iftemp=="9999.9"ordew=="9999.9"orwind=="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" elif16 <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}")

ifname ==" main ":

map1()

# reducer.py: #!

/usr/bin/env pythonimport sys

def reduce1(): current\_key = None sum\_temp, sum\_dew, sum\_wind = 0, 0, 0

count=0

forlinein sys.stdin:

key, value = line.strip().split("\t") temp, dew, wind = map(float, value.split()) if current\_key is None:

current\_key= key

ifkey== 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()

1. RuntheHadoopStreaming Job

hadoop jar C:\hadoop\share\hadoop\tools\lib\hadoop-streaming-3.3.1.jar ^ -files"/Users/monid/OneDrive/Documents/DataAnalytics/mapper2.py,/Users/monid/OneDrive/Documents/DataAnalytics/reducer2.py" ^ -input

/user/weather/sample\_weather.txt ^ -output /user/output1 ^ -mapper "pythonC:/Users/monid/OneDrive/Documents/DataAnalytics/mapper2.py" ^ -reducer"pythonC:/Users/monid/OneDrive/Documents/DataAnalytics/reducer2.py "



1. ViewtheOutputhdfsdfs -cat/user/output1/part-00000



1. Oncethemapreduceoperationsareperformedsuccessfully,the output will be present in the specified directory.

“/user/output1/part-00000”



1. StopHadoopServicesstop-dfs.cmdstop-yarn.cmd



**RESULT:**

ThustheimplementationoftheMapReducepythonprogramaweatherdatasetinHadoopisexecuted successfully.