PA6 –

Used - AWS EMR for this assignment

We need provide command line argument to the program as per requirement:

MONTH

YEAR

SEASON

In my dataset Station id was not present (provided by TA) then so I have taken the filename which contain the station id it self

Data used : <ftp://ftp.ncdc.noaa.gov/pub/data/noaa/isd-lite/>

ftp://ftp.ncdc.noaa.gov/pub/data/normals/1981-2010/source-datasets/isdlite-normals.tar.gz

Below are the output :



Run my program on number of combination :

<https://s3-us-west-2.amazonaws.com/test-weather-pa6/final_MONTH_map_2_reduce_2/part-00000>

<https://s3-us-west-2.amazonaws.com/test-weather-pa6/final_MONTH_map_2_reduce_2/part-00001>

* Even I have run on multiple instances :

For example :

bin/hadoop jar contrib/streaming/hadoop-streaming.jar -D mapread.map.tasks=10 -D mapred.reduce.tasks=2 -mapper "mapper.py SEASON " -file mapper.py -reducer "reducer.py" -file reducer.py -input s3://test-weather-data-pa6 -output s3://test-weather-pa6/final\_season\_map\_10\_reduce\_2

logs:

