**Aim : Weather Analysis using MapReduce.**

Here, we will write a Map-Reduce program for analyzing weather datasets to understand its data processing programming model. Weather sensors are collecting weather information across the globe in a large volume of log data. This weather data is semi-structured and record-oriented.

This data is stored in a line-oriented ASCII format, where each row represents a single record. Each row has lots of fields like longitude, latitude, daily max-min temperature, daily average temperature, etc. for easiness, we will focus on the main element, i.e. temperature. We will use the data from the National Centres for Environmental Information(NCEI). It has a massive amount of historical weather data that we can use for our data analysis.

Problem Statement :-

Analyzing weather data of Fairbanks, Alaska to find cold and hot days using MapReduce Hadoop.

Step 1 : Open terminal > Add the weather.txt file and create mapper.py and reducer.py.

[cloudera@quickstart Desktop]$ mkdir weather

[cloudera@quickstart Desktop]$ cd weather

[cloudera@quickstart weather]$ touch weather.txt

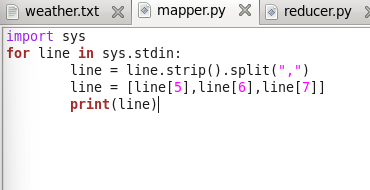
[cloudera@quickstart weather]$ ls

weather.txt

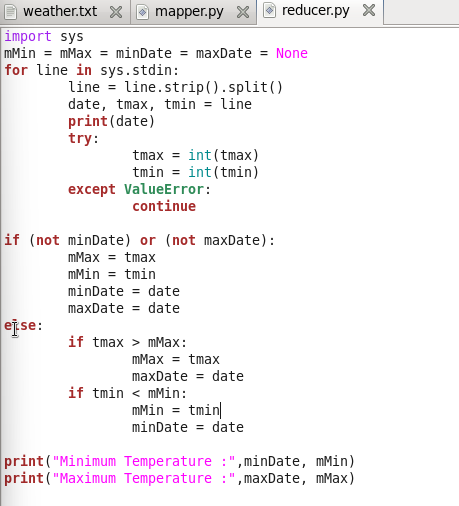
[cloudera@quickstart weather]$ touch mapper.py

[cloudera@quickstart weather]$ touch reducer.py

Step 2 : Open mapper.py and write the code.



Step 3 : Open reducer.py and write the code.



Step 4 : Running the mapper function with weather.txt file to check the output.

[cloudera@quickstart weather]$ cat weather.txt | python mapper.py

**Output:**

20100101 -178 -311

20100102 -244 -322

20100103 -194 -289

20100104 -167 -200

20100105 -133 -167

20100106 -133 -172

20100107 -150 -278

20100108 -233 -328

20100109 -233 -322

20100110 -117 -244

20100111 -67 -128

20100112 -78 -122

20100113 -17 -89

20100114 39 -72

20100115 -67 -72

20100116 22 -50

20100117 33 -44

20100118 6 -172

20100119 -56 -183

20100120 -67 -139

20100121 -67 -94

20100122 -44 -67

20100123 -6 -44

20100124 0 -11

20100125 -11 -161

20100126 -161 -233

20100127 -167 -222

20100128 -167 -283

20100129 -189 -283

20100130 -156 -267

20100131 -150 -272

Step 5: The mapper function will put these values in the stream which will be used by the reducer function and give the output

[cloudera@quickstart weather]$ cat weather.txt | python mapper.py |python reducer.py

**Output:**

['20100101', '-178', '-311']

20100101

('Minimum Temperature:', '20100101', -311)

('Maximum Temperature:', '20100101', -178)

['20100102', '-244', '-322']

20100102

('Minimum Temperature:', '20100102', -322)

('Maximum Temperature:', '20100101', -178)

['20100103', '-194', '-289']

20100103

('Minimum Temperature:', '20100102', -322)

('Maximum Temperature:', '20100101', -178)

['20100104', '-167', '-200']

20100104

('Minimum Temperature:', '20100102', -322)

('Maximum Temperature:', '20100104', -167)

['20100105', '-133', '-167']

20100105

('Minimum Temperature:', '20100102', -322)

('Maximum Temperature:', '20100105', -133)

['20100106', '-133', '-172']

20100106

('Minimum Temperature:', '20100102', -322)

('Maximum Temperature:', '20100105', -133)

['20100107', '-150', '-278']

20100107

('Minimum Temperature:', '20100102', -322)

('Maximum Temperature:', '20100105', -133)

['20100108', '-233', '-328']

20100108

('Minimum Temperature:', '20100108', -328)

('Maximum Temperature:', '20100105', -133)

['20100109', '-233', '-322']

20100109

('Minimum Temperature:', '20100108', -328)

('Maximum Temperature:', '20100105', -133)

['20100110', '-117', '-244']

20100110

('Minimum Temperature:', '20100108', -328)

('Maximum Temperature:', '20100110', -117)

['20100111', '-67', '-128']

20100111

('Minimum Temperature:', '20100108', -328)

('Maximum Temperature:', '20100111', -67)

['20100112', '-78', '-122']

20100112

('Minimum Temperature:', '20100108', -328)

('Maximum Temperature:', '20100111', -67)

['20100113', '-17', '-89']

20100113

('Minimum Temperature:', '20100108', -328)

('Maximum Temperature:', '20100113', -17)

['20100114', '39', '-72']

20100114

('Minimum Temperature:', '20100108', -328)

('Maximum Temperature:', '20100114', 39)

['20100115', '-67', '-72']

20100115

('Minimum Temperature:', '20100108', -328)

('Maximum Temperature:', '20100114', 39)

['20100116', '22', '-50']

20100116

('Minimum Temperature:', '20100108', -328)

('Maximum Temperature:', '20100114', 39)

['20100117', '33', '-44']

20100117

('Minimum Temperature:', '20100108', -328)

('Maximum Temperature:', '20100114', 39)

['20100118', '6', '-172']

20100118

('Minimum Temperature:', '20100108', -328)

('Maximum Temperature:', '20100114', 39)

['20100119', '-56', '-183']

20100119

('Minimum Temperature:', '20100108', -328)

('Maximum Temperature:', '20100114', 39)

['20100120', '-67', '-139']

20100120

('Minimum Temperature:', '20100108', -328)

('Maximum Temperature:', '20100114', 39)

['20100121', '-67', '-94']

20100121

('Minimum Temperature:', '20100108', -328)

('Maximum Temperature:', '20100114', 39)

['20100122', '-44', '-67']

20100122

('Minimum Temperature:', '20100108', -328)

('Maximum Temperature:', '20100114', 39)

['20100123', '-6', '-44']

20100123

('Minimum Temperature:', '20100108', -328)

('Maximum Temperature:', '20100114', 39)

['20100124', '0', '-11']

20100124

('Minimum Temperature:', '20100108', -328)

('Maximum Temperature:', '20100114', 39)

['20100125', '-11', '-161']

20100125

('Minimum Temperature:', '20100108', -328)

('Maximum Temperature:', '20100114', 39)

['20100126', '-161', '-233']

20100126

('Minimum Temperature:', '20100108', -328)

('Maximum Temperature:', '20100114', 39)

['20100127', '-167', '-222']

20100127

('Minimum Temperature:', '20100108', -328)

('Maximum Temperature:', '20100114', 39)

['20100128', '-167', '-283']

20100128

('Minimum Temperature:', '20100108', -328)

('Maximum Temperature:', '20100114', 39)

['20100129', '-189', '-283']

20100129

('Minimum Temperature:', '20100108', -328)

('Maximum Temperature:', '20100114', 39)

['20100130', '-156', '-267']

20100130

('Minimum Temperature:', '20100108', -328)

('Maximum Temperature:', '20100114', 39)

['20100131', '-150', '-272']

20100131

('Minimum Temperature:', '20100108', -328)

('Maximum Temperature:', '20100114', 39)

[cloudera@quickstart weather]$ ^C

[cloudera@quickstart weather]$