STAT 480 Homework 5 kakilai2

Command line for copying data to the distributed file system:
#hadoop fs -mkdir -p input/ncdc/all
#hadoop fs -ls /user/cloudera
#hadoop fs -ls input/ncdc

Exercise 1:
Command :
#hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
#-files /home/host-data/hb-workspace/ch02-mr-intro/src/main/python/data_map.py,\
/home/host-data/hb-workspace/ch02-mr-intro/src/main/python/min_temperature_reduce.py
\
#-input input/ncdc/all \
#-output p1outputpy \
#-mapper "/home/host-data/hb-workspace/ch02-mr-intro/src/main/python/data_map.py" \
#-reducer "/home/host-data/hb-workspace/ch02-mr-
intro/src/main/python/min_temperature_reduce.py"

The Minimum temperature for each year is as the screen cap below:

```
[cloudera@quickstart python]$ hadoop fs -cat p1outputpy/part*
1901    -333
1902    -328
1903    -306
1904    -294
1905    -328
1906    -250
1907    -350
1908    -378
1909    -378
1910    -372
```

The first column shows the years while the second column show the minimum temperature of that year (which is scaled by a factor of 10 due to the data format of the temperature value in the source file). Refer to min_temperature_reduce.py for the script.

Exercise 2:
Command:
#hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
-files /home/host-data/hb-workspace/ch02-mr-intro/src/main/python/data_map.py,\
/home/host-data/hb-workspace/ch02-mr-intro/src/main/python/count.py \
#-input input/ncdc/all \

```
 #-output p2outputpy \
 #-mapper "/home/host-data/hb-workspace/ch02-mr-intro/src/main/python/data_map.py" \
 #-reducer "/home/host-data/hb-workspace/ch02-mr-intro/src/main/python/count.py"
```

The total number of trusted temperature observations (temperature observations that are not missing and that have acceptable quality codes) for each year from 1901 to 1910 is as the screen cap below:

```
[cloudera@quickstart python]$ hadoop fs -cat p2outputpy/part*
1901    6564
1902    6565
1903    6511
1904    6582
1905    6561
1906    5474
1907    5461
1908    6584
1909    7534
1910    7645
```

The first column shows the years of the observations, while the second column is the total number of trusted temperature. Refer to count.py for the script.

Exercise 3:

Command:

```
#hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
 #-files /home/host-data/hb-workspace/ch02-mr-intro/src/main/python/data_map.py,\
/home/host-data/hb-workspace/ch02-mr-intro/src/main/python/combine.py \
 #-input input/ncdc/all \
 #-output p3outputpy \
 #-mapper "/home/host-data/hb-workspace/ch02-mr-intro/src/main/python/data_map.py" \
 #-reducer "/home/host-data/hb-workspace/ch02-mr-intro/src/main/python/combine.py"
```

The number of trusted temperature observations and the minimum and maximum temperatures and for each year from 1901 to 1910 is as the screen cap below:

```
[cloudera@quickstart python]$ hadoop fs -cat p3outputpy/part*
1901    317     -333    6564
1902    244     -328    6565
1903    289     -306    6511
1904    256     -294    6582
1905    283     -328    6561
1906    294     -250    5474
1907    283     -350    5461
1908    289     -378    6584
1909    278     -378    7534
1910    294     -372    7645
```

The first column shows the year, the second column shows the maximum temperature (which is scaled by a factor of 10 due to the data format of the temperature value in the source file),

the third column shows the minimum temperature (which is 10 times larger due to the format of the data). Refer to combine.py for the script.

Exercise 4:
Command :

```
#hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
 #-files /home/host-data/hb-workspace/ch02-mr-intro/src/main/python/data_map.py,\
/home/host-data/hb-workspace/ch02-mr-intro/src/main/python/findmean.py \
 #-input input/ncdc/all \
 #-output p4outputpy \
 #-mapper "/home/host-data/hb-workspace/ch02-mr-intro/src/main/python/data_map.py" \
 #-reducer "/home/host-data/hb-workspace/ch02-mr-intro/src/main/python/findmean.py"
```

The mean temperature for each year from 1901 to 1910 is as the screen cap below:

```
1901    46.698507
1902    21.659558
1903    48.241745
1904    33.322242
1905    43.332266
1906    47.083486
1907    31.764146
1908    28.836574
1909    26.565304
1910    35.558666
```

The first column shows the year, the second column shows the total number of record for each year while the second column shows mean temperature of each year (which is scaled by a factor of 10 due to the data format of the temperature value in the source file). Refer to findmean.py for the script.
To calculate the mean, the following formula is used to compute the rolling mean:

```
mean * (count/(count+1)) + val/(count + 1)
```

Where *mean* stores the rolling mean and *val* stores the temperature of the record being read. This avoids the storage of the large accumulate sum.

Files Description:
data_map.py : mapper for Exercise 1 – 4
min_temperature_reduce.py : reducer for Exercise 1
count.py : reducer for Exercise 2
combine.py : reducer for Exercise 3
findmean.py : reducer for Exercise 4
STAT 480 Homework 5 kakilai2_Report : Report for the Assignment
Exercise (1 – 4) Result : Output Generated
STAT 480 command.sh : command script for running Exercise 1 to 4