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

Section 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 \

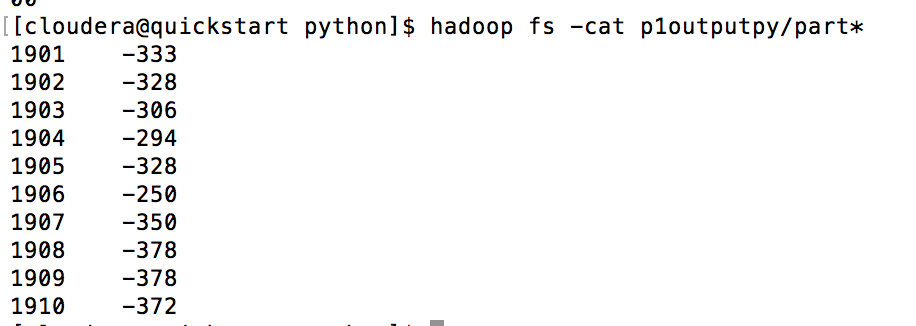
#-input input/ncdc/all \

#-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:



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.

Section 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 \

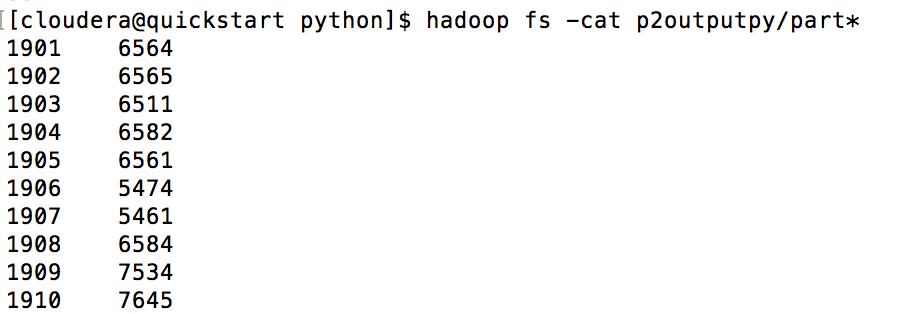
#-input input/ncdc/all \

#-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:



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.

Section 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 \

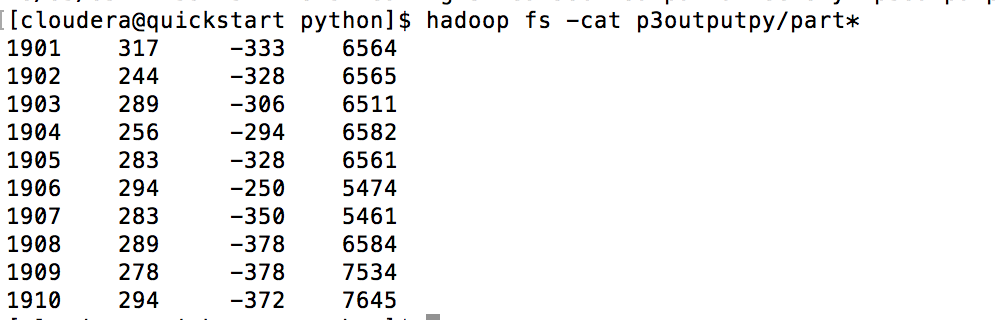
#-input input/ncdc/all \

#-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:



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.

Section 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 \

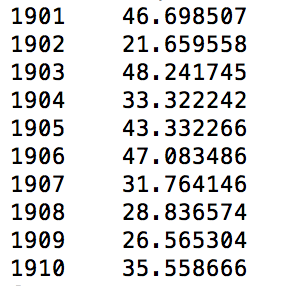
#-input input/ncdc/all \

#-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:



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.